José María
Matías Fernández

Tesis doctoral

Redes neuronales para regresión y clasificación. Nuevos algoritmos y aplicaciones

Santiago de Compostela, 2004

Facultad de Matemáticas. Departamento de Estadística e Investigación Operativa

UNIVERSIDADE DE
SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE
ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

REDES NEURONALES PARA REGRESIÓN Y CLASIFICACIÓN. NUEVOS ALGORITMOS Y APLICACIONES

JOSÉ MARÍA MATÍAS FERNÁNDEZ

12 DE DICIEMBRE DE 2003

Tesis Doctoral

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

I.S.B.N: 84-688-8028-0
Depósito Legal: C-2056-2004

Imprime:



NINO
Centro de Impresión Digital
Teléfono 981 59 03 02
Rosalía de Castro, 58
Santiago de Compostela

AGRADECIMIENTOS

Son numerosas las personas a las que tengo que agradecer la ayuda y el apoyo prestado durante el desarrollo de este trabajo de investigación.

En general, quisiera expresar mi agradecimiento a la Universidad de Vigo por haber creado el marco institucional que me ha permitido desarrollar este trabajo, y a todos los miembros del Departamento de Estadística e Investigación Operativa de dicha universidad, por su compañerismo y comprensión en muchos momentos de este período.

En especial, quiero expresar mi profundo agradecimiento a los directores de este trabajo, Wenceslao González-Manteiga y Antonio Vaamonde Liste, por la confianza depositada desde el principio, y por su apoyo técnico y personal. Sin ellos, este trabajo no se habría realizado.

Asimismo, deseo manifestar mi agradecimiento expreso a Ángeles Saavedra González por haberme introducido en las técnicas de estadística espacial, y a Manuel Febrero Bande y Juan Carlos Reboredo Nogueira, de la Universidad de Santiago de Compostela, por las didácticas y entretenidas sesiones de trabajo mantenidas sobre modelos de predicción de series financieras.

Finalmente, no podré olvidar el apoyo personal e incluso financiero de Javier Taboada Castro, así como la relación mantenida con Celestino Ordóñez Galán, ambos del Departamento de Ingeniería de los Recursos Naturales y Medio Ambiente de la Universidad de Vigo, que han sido esenciales para la realización de esta memoria.

Este trabajo ha sido financiado en parte por el proyecto BFM2002-03213 del Ministerio de Ciencia y Tecnología y el programa ERDF de la Unión Europea.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

A Cati

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Índice

Introducción **1**

Motivación y Objetivos 1

Aportaciones 2

Futuras Líneas de Investigación y Desarrollo 3

Estructura de la Memoria 4

I Marco General de la Teoría de Aprendizaje **7**

I.1 Planteamiento del Problema de Aprendizaje 7

I.1.1 Elementos del Problema de Aprendizaje 8

I.1.2 Problema de Aprendizaje Bien Planteado 11

I.2 Ejemplos de Problemas de Aprendizaje 12

I.2.1 Problema de Regresión 12

I.2.2 Problema de Clasificación 13

I.2.3 Problema de Estimación de la Densidad 16

I.3 Condiciones para el Éxito del Aprendizaje 17

I.3.1 Consistencia del Proceso de Aprendizaje 18

I.3.2 Capacidad de la Familia de Hipótesis 20

I.3.3 Velocidad de Convergencia 22

I.3.4 Complejidad Computacional 23

I.4 Control de la Complejidad 24

I.4.1 Regularización 25

I.5 Selección del Modelo 27

I.5.1 Descomposición del Riesgo Esperado en Sesgo y Varianza 29

I.5.2 Criterios Asintóticos Clásicos 33

I.5.3 Validación Cruzada: Expresiones Analíticas 39

I.5.4 Comparación de distintos criterios 42

I.5.5 Técnicas de Remuestreo 43

I.5.6 Técnicas No Asintóticas 45

I.6 Consistencia de los Algoritmos 56

I.6.1 Optimalidad de los Criterios Clásicos 57

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

I.6.2 Consistencia de los Criterios No Asintóticos 59

II Redes Neuronales para Clasificación y Regresión 67

II.1 Las Redes Neuronales 67
II.1.1 Breve Reseña Histórica 67
II.1.2 La Red Neuronal 69
II.1.3 Redes Neuronales para la Resolución de Problemas Estadísticos 72
II.1.4 Modelo Funcional de la Red Neuronal 73
II.1.5 Relación Con Otros Modelos Estadísticos 75
II.1.6 Requisitos de una Red Neuronal 76
II.2 Redes Neuronales de Una Capa 77
II.2.1 El Perceptrón 79
II.2.2 Adaline 82
II.2.3 Adaline con Activación Sigmoide 86
II.2.4 Clasificación con Varias Clases 89
II.2.5 Funciones de Pérdida y Salidas como Probabilidades a Posteriori 92
II.2.6 Redes Neuronales de Una Capa para el Problema de Regresión 94
II.3 Arquitecturas Multicapa Hacia Adelante (MLF) 94
II.3.1 Arquitecturas MLF Lineales 95
II.3.2 Arquitecturas MLF No Lineales. Visión General 100
II.4 Arquitecturas de Tipo Proyección 106
II.4.1 Redes Perceptrón Multicapa (MLP) 107
II.5 Arquitecturas MLF de Base Radial 111
II.5.1 Redes RBF 112
II.5.2 Redes RBF de Regularización (RRBF) 126
II.5.3 Support Vector Machines (SVM) 138
II.5.4 Consistencia de las Arquitecturas de Tipo Radial 152
II.6 Algoritmos para redes MLF 156
II.6.1 Algoritmos *Back-Propagation* 158
II.6.2 Algoritmos de Regresión Lineal Generalizada 171
II.6.3 Algoritmos Mixtos 182
II.6.4 Algoritmos para SVM 183
II.6.5 Aprendizaje Bayesiano 185
II.6.6 Algoritmos Greedy 198

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

II.6.7 Otros Algoritmos con Filosofía Diccionario 209
II.6.8 Análisis de Parsimonia de los Distintos Algoritmos 213
II.7 Técnicas de Refinamiento del Aprendizaje. *Boosting* 216
II.7.1 Motivación 216
II.7.2 Tipología 220
II.7.3 *Boosting*. Reseña Histórica 221
II.7.4 *Boosting* como Algoritmo de Máxima Verosimilitud 228
II.7.5 Relación entre *Boosting* y Otros Algoritmos 235
II.7.6 Análisis del Comportamiento del *Boosting* con Redes RBF-OLS en un Problema de Clasificación 238

III Relaciones con Otras Técnicas Estadísticas, Nuevos Algoritmos y Aplicaciones 245
III.1 Regresión Bayesiana, Procesos Gausianos y Redes Neuronales 245
III.1.1 Regresión Bayesiana y Procesos Gausianos 245
III.1.2 Regresión Mediante Procesos Gausianos y Redes Neuronales RBF y RRBF . 249
III.2 Kriging y Redes Neuronales 250
III.2.1 Kriging 250
III.2.2 Kriging y Redes RRBF 254
III.2.3 Kriging y SVM 257
III.2.4 Kriging Regularizado 260
III.2.5 Conclusiones: Dualidad Tendencia-Dependencia y Enfoque Bayesiano 273
III.3 Estimación de la Explotabilidad en un Yacimiento de Pizarra 279
III.3.1 Objetivos y Metodología 280
III.3.2 Modelización de la Variabilidad Espacial 282
III.3.3 Algoritmos Utilizados 283
III.3.4 Resultados 285
III.3.5 Conclusiones 288
III.4 Planificación de la Reforestación 292
III.4.1 Motivación 292
III.4.2 Planteamiento del Problema y Metodología 292
III.4.3 Algoritmos Utilizados 294
III.4.4 Resultados 295
III.5 Nuevos Algoritmos para la Predicción de Series de Tiempo en Finanzas 299
III.5.1 Técnicas de Modelización de la Volatilidad 299

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

III.5.2 Nuevas Técnicas *Boosting* para la Modelización de la Volatilidad 302
III.5.3 Evaluación de los Nuevos Algoritmos 308

ANEXOS **333**

A Clasificación y Discriminante **335**
Discriminante Lineal de Fisher 335
Relación entre Diversos Criterios de Clasificación 336

B Espacios de Hilbert con Núcleo Reproductor **339**
Preliminares 339
Operadores en Espacios de Hilbert 341
Descomposición Espectral en Espacios de Hilbert 344
Espacios de Hilbert con Núcleo Reproductor 345
Espacios de Hilbert con Seminúcleo Reproductor 350
Caracterización de las Funciones Condicionalmente Definidas Positivas 352

C Propiedades Aproximadoras de las Redes Neuronales **357**
Conceptos Básicos de la Teoría de Aproximación de Funciones 357
Aproximación Mediante Redes Neuronales 365

D Optimización **389**
Algoritmos Basados en el Gradiente 389
Algoritmos Basados en el Hesiano 391
Optimización Estocástica. Algoritmo de Robbins-Monro 393
Problema de Optimización con Restricciones 395
Algoritmos de Punto Interior 398

E Consistencia del Algoritmo *Back-Propagation* **401**

F *Boosting* como Algoritmo de Máxima Verosimilitud **405**

G Otros Resultados Útiles **411**
Algebra Matricial, Diferenciación y Formas Cuadráticas 411
Momentos 415
Cálculo de Probabilidades 416

USC DE COMPOSTELA

Capacidad de las Redes Neuronales 416
Regularización 418

H Notación e Índice de Símbolos 421

REFERENCIAS 429

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Introducción

Motivación y Objetivos

Las técnicas de redes neuronales, como otras técnicas estadísticas con una gran componente computacional, han experimentado un gran desarrollo en los últimos años, debido al enorme crecimiento de la capacidad computacional de los ordenadores.

Este auge, sigue produciéndose en la actualidad hasta el punto de que, resulta muy difícil seguir los últimos desarrollos, en el contexto de un marco teórico compacto e interrelacionado.

Todo ello, se ve agravado por la aparición de nuevas publicaciones especializadas, produciéndose una gran dispersión de literatura entre diversos canales, asociados tradicionalmente a distintas disciplinas.

Esta situación, dificulta el acceso a estas técnicas por parte de investigadores inicialmente ajenos a las mismas, impide intercambios conceptuales y metodológicos entre áreas, y limita sus posibilidades de aplicación en muchos campos de interés.

Así, existen técnicas de redes neuronales que, a primera vista, parecen poseer un estrecho parentesco con otras técnicas estadísticas, pero cuyas relaciones no han sido suficientemente clarificadas. Asimismo, existen muchas áreas de aplicación, donde estas nuevas técnicas no han sido suficientemente aprovechadas, en las que las redes neuronales tienden a considerarse como *cajas negras* con buenos resultados, pero imposibles de comprender debido a su complejidad.

En lo que respecta a la ciencia estadística, las redes neuronales son una parte de la misma, pero su origen en áreas de tipo tecnológico, las ha mantenido alejadas de aquélla.

En los últimos años, se ha producido un doble acercamiento en ambos sentidos: por una parte, la teoría del aprendizaje (PAC) surgida en el seno de la inteligencia artificial ha vuelto a crear conceptos ya existentes en la estadística desde hace décadas; por otra, la ciencia estadística, ha comenzado a reconocerse a sí misma como una ciencia sobre el proceso de adquisición de conocimiento; sobre todo, después de la aparición de la teoría del aprendizaje estadístico con origen en [Vapnik y Chervonenkis 1971] y [Vapnik y Chervonenkis 1981].

El encuentro entre estos dos movimientos, ha supuesto una síntesis conceptual y de esfuerzos, que está sentando las bases de lo que será esta ciencia en el nuevo siglo.

En este contexto, la presente memoria estudia las técnicas de redes neuronales para clasificación y regresión, persiguiendo los siguientes objetivos:

1. Resumir de forma sistemática e integrada, el cuerpo de conocimiento que constituyen las técnicas de redes neuronales para regresión y clasificación en su estado más reciente, y en el marco de la teoría estadística de aprendizaje.

2. Mostrar las posibles relaciones teóricas o metodológicas, entre dichas técnicas y otros métodos estadísticos.

3. Proponer los nuevos algoritmos que dichas interrelaciones puedan sugerir.

4. Aplicar los algoritmos existentes y los de nuevo desarrollo a diferentes problemas de aplicación, diseñando metodologías al efecto y analizando su comportamiento.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Aportaciones

Las aportaciones más significativas de esta memoria son las siguientes:

1. Se clarifican las relaciones entre el kriging, y las redes neuronales de base radial de regularización (RRBF) [Matías, Vaamonde, Taboada y González-Manteiga 2004*a*] y las *support vector machines* (SVM) [Matías, Vaamonde, Taboada y González-Manteiga 2004*b*], y, en dicho marco, se propone una fundamentación para la utilización de funciones de covarianza en dichas redes y en las redes neuronales de base radial (RBF).

2. Se muestra la aplicabilidad de los resultados recientes de [Audrino y Bühlmann 2003] sobre el algoritmo LS_Boost, a las redes RBF y RRBF como suavizadores lineales, éstas últimas como generalización de la regresión mediante *smoothing splines*.

3. Se evalúa [Matías 2003] el comportamiento del algoritmo AdaBoost aplicado a redes RBF en un problema de clasificación, utilizando la descomposición sesgo-varianza propuesta recientemente por [Domingos 2000].

4. Se desarrolla un algoritmo de entrenamiento de redes RBF con distintos centros (posiblemente repetidos), diferentes parámetros de escala, y varios regularizadores, combinando técnicas de proceso de señales (*orthogonal matching pursuit*) y regularización bayesiana, bajo una filosofía de diccionario.

5. Se desarrolla un algoritmo LS_Boost con *back-fitting*, utilizando redes neuronales RBF sobre proyecciones en el espacio de entrada [Matías, Vaamonde, Taboada y González-Manteiga 2004*b*].

6. Se desarrolla el algoritmo kriging regularizado [Matías y González-Manteiga 2003*b*], como resultado de la aplicación de la metodología de los vectores soporte para relajar la condición de insesgadez del kriging, se evalúa su comportamiento en simulaciones y se muestra su relación con la regresión bayesiana.

7. Se desarrolla una bateria de nuevos algoritmos para la predicción series de tiempo en finanzas evaluando su comportamiento en diferentes simulaciones:
   (a) Para la predicción de precios: RBF-GARCH y RBF-GARCHRBF, y versiones análogas para redes *multilayer perceptron* (MLP), MLP-GARCH y MLP-GARCHMLP.
   (b) Para la predicción de volatilidades: BoostMLP (y su versión para RBF), BoostARX, BoostGARCH, y WildBoostGARCH.

8. Se aplican distintos algoritmos existentes y los nuevos desarrollados, a diversos problemas reales de aplicación:
   (a) La predicción de precios y volatilidades sobre una serie de datos reales del índice Standard & Poor's 500.
   (b) La estimación de la explotabilidad de un yacimiento de pizarra situado en la provincia de Zamora, [Matías, Vaamonde, Taboada y González-Manteiga 2004*a*], [Matías, Taboada, Vaamonde y González-Manteiga 2004], [Matías, Vaamonde, Taboada y González-Manteiga 2004*b*].
   (c) La elaboración de un plan de reforestación de la comarca del río Liévana en la Comunidad Autónoma de Cantabria, [Ordóñez *et al.* 2003*b*], [Ordóñez *et al.* 2003*a*].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Futuras Líneas de Investigación y Desarrollo

A lo largo de esta memoria, se identifican las siguientes líneas de investigación y desarrollo, de interés:

1. La implementación y evaluación empírica de los criterios de selección del modelo basados en penalizaciones adaptadas a los datos, y basados en los métodos de la estabilidad del algoritmo.

2. El desarrollo de un algoritmo bayesiano de selección del parámetro de escala en las redes RBF.

3. La fundamentación teórica y el desarrollo de algoritmos para la incorporación de parámetros de escala variables en los núcleos de las SVM, y el desarrollo de algoritmos para la utilización de pérdidas ε-insensible con $\varepsilon = \varepsilon(\mathbf{x})$, para problemas de regresión heterocedásticos.

4. El desarrollo de algoritmos que permitan la incorporación de información a priori en las redes MLP (por ejemplo, fijando direcciones de proyección). Dicha generalización, permitiría la incorporación, a estas redes, de información sobre direcciones de anisotropía en el ámbito de la estadística espacial.

5. El desarrollo de algoritmos para la estimación de la estructura de dependencia en espacios de gran dimensionalidad, para su utilización como funciones básicas en las redes neuronales con arquitectura radial.

6. El desarrollo de algoritmos *boosting* para series financieras utilizando modelos FIGARCH.

7. La aplicación del método del núcleo (*kernel trick*) y pérdida ε-insensible, a modelos lineales existentes para la predicción de series de tiempo homocedásticas y heterocedásticas.

8. El desarrollo de nuevos algoritmos basados en redes neuronales y *boosting* para la estimación de la densidad condicionada en el contexto de series de tiempo heterocedásticas.

9. El desarrollo de metodologías y algoritmos de regresión funcional basados en redes neuronales, y su comparación con los existentes para series de tiempo homocedásticas y heterocedásticas.

10. El refinamiento de los algoritmos *boosting* para series de tiempo, mediante la selección de los algoritmos de optimización de redes neuronales más adecuados, y la incorporación de técnicas de regularización a los mismos.

11. El desarrollo teórico y algorítmico de tests de bondad de ajuste con modelos basados en redes neuronales.

12. El estudio de consistencia y tasas de convergencia de la red neuronal RBF entrenada mediante OLS con selección del modelo mediante análisis espectral y con posible regularización.

13. El estudio de consistencia y tasas de convergencia de los nuevos algoritmos basados en redes neuronales y *boosting*, para series heterocedásticas, bajo diferentes hipótesis de dependencia para la serie.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Estructura de la Memoria

La presente memoria se estructura en tres capítulos:

1. El primero, tiene por objeto sentar las bases del problema de aprendizaje estadístico, en cuyo marco se desenvuelven las técnicas de redes neuronales que se tratarán en el documento.

   Con el fin de lograr la máxima integración, la presentación de dicho problema posee un carácter ecléctico, aglutinando de forma sistematizada, conceptos y técnicas relacionados, procedan éstos de la estadística o del ámbito de la complejidad computacional.

2. En el marco conceptual establecido en el capítulo primero, el segundo capítulo describe y analiza, los modelos y algoritmos recientes de redes neuronales más relevantes para los problemas de regresión y clasificación. En él, se detallan en mayor medida los diferentes elementos que vayan a ser utilizados en el capítulo siguiente.

   La exposición comienza con una rápida revisión ilustrativa de las redes neuronales de una sóla capa, y de las redes multicapa de carácter lineal, con el fin de aprovechar sus propiedades en el análisis posterior de las redes multicapa hacia adelante (*multilayer feed-forward*, MLF).

   Dado que diversos algoritmos pueden ser utilizados con distintas arquitecturas MLF, hemos optado por presentar en primer lugar los diversos tipos de arquitectura, su fundamentación y sus propiedades estadísticas, para, posteriormente, presentar todos los algoritmos más importantes, susceptibles de utilización con aquéllas. Algunos, son exclusivos de un grupo de arquitecturas pero, otros, poseen una aplicabilidad más universal.

   Como colofón a este capítulo, realizamos una rápida introducción de las técnicas de combinación de modelos, para centrarnos seguidamente en las técnicas *boosting*, que serán utilizadas en el capítulo siguiente.

3. El tercer capítulo recoge las aportaciones de la presente memoria.

   La primera parte del capítulo (las dos primeras secciones) analiza las relaciones entre las redes neuronales RBF, RRBF y SVM, y otras técnicas estadísticas como la regresión bayesiana, la regresión mediante procesos gausianos y el kriging. Asimismo, dada su estrecha relación con las redes SVM, se presenta, en esta primera parte, el algoritmo kriging regularizado.

   En una segunda parte, (tercera y cuarta sección) se aplican diversos algoritmos de redes neuronales a dos problemas de aplicación: la estimación de la explotabilidad de un yacimiento de pizarra y la elaboración del plan de reforestación de una comarca española. En el primer problema nos centramos en la comparación de los distintos algoritmos evaluando el alcance de las conclusiones obtenidas en la primera parte del capítulo. El segundo problema posee una mayor dimensionalidad, lo que aconseja restringir los algoritmos utilizados. En él, nos centramos en la comparación entre estos algoritmos y las técnicas clásicas utilizadas tradicionalmente.

   Finalmente, en una última sección, presentamos un nuevo conjunto de algoritmos para la predicción de series de tiempo heterocedásticas, evaluamos su comportamiento en diversos juegos de simulación y aplicamos dichos algoritmos al problema real de predicción de precios y rentabilidades de la serie Standard & Poor's.

Los anexos incluyen una breve revisión de diferentes áreas conceptuales, necesarias para el seguimiento del texto. En el último, se incluye un índice de los símbolos utilizados con carácter general en el documento, indicando la página donde dichos símbolos se delimitan o definen. En

USC DE COMPOSTELA

los casos en los que ello es posible, se acompaña cada entrada con una breve descripción de la misma.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Capítulo I
Marco General de la Teoría de Aprendizaje

Las redes neuronales son estructuras matemáticas con capacidad para generalizar a partir de un conjunto finito de observaciones. En el ámbito de la estadística, esta propiedad se denomina de manera clásica, *inferencia*, pero, en los últimos años, como resultado de la interacción entre diversas disciplinas como la inteligencia artificial, la ciencia de la computación, la psicología y la propia estadística, se ha ido consolidando el uso de una palabra que incorpora connotaciones adicionales útiles a todas ellas: *aprendizaje*.

Por tanto, es imposible entender y utilizar las redes neuronales fuera del marco de la teoría del aprendizaje, sobre la cuál, este capítulo, a modo de referencia, pretende aportar tanto los conceptos básicos como una visión general de sus líneas de desarrollo más recientes.

A semejanza de su referente psico-biológico, el **problema de aprendizaje** tiene por objeto fundamental la determinación de las condiciones de adquisición de conocimiento (capacidad de generalización). Así, una formulación adecuada del mismo tiene que ser lo suficientemente general como para englobar una amplia gama de problemas:

1. Según sus objetivos: suavizado (pasado), predicción (futuro), representación, explicación, aproximación, filtrado, análisis estructural, compresión de información, etc.

2. Según sus enfoques: clasificación, regresión, análisis temporal, diseño de experimentos, reducción de dimensionalidad, reconocimiento de patrones, etc.

3. Según las circunstancias del experimentador ante el problema, entre otras:
   - Con conocimiento, o sin él, de una supuesta función (o familia de funciones) verdadera, característica del proceso generador de los datos, o disponiendo, o no, de información a priori sobre ella.
   - Existencia o no, de ruido en los datos disponibles.
   - Existencia de restricciones tanto en la toma de datos como en el modelo, por razones de coste, viabilidad física o computacional, etc.

I.1 Planteamiento del Problema de Aprendizaje

Como ya hemos apuntado, el planteamiento actual del problema de aprendizaje artificial es producto de diversas contribuciones procedentes de diversas disciplinas y, en especial, de la teoría de la computación y de la teoría de la probabilidad y la estadística.

El primer planteamiento reseñable, sin pretensión de exhaustividad, en esta reciente pero intensa historia, proviene de [Valiant 1984] que afronta el problema de aprendizaje con una formulación inicial simplificada e incidiendo especialmente en la problemática de la complejidad computacional.

Sin embargo, esta iniciativa consigue enmarcar teorías existentes con anterioridad y anima aportaciones posteriores, aglutinando conocimientos existentes en otras disciplinas y atrayendo sus esfuerzos.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En el primer grupo, destacan la teoría de los procesos empíricos ([Vapnik y Chervonenkis 1971], [Vapnik y Chervonenkis 1981], [Vapnik 1982], [Pollard 1984], [Pollard 1990]) y la teoría de la decisión, (p. ej. [Berger 1985]) y en el segundo, las aportaciones de [Haussler 1992] que generaliza las formulaciones iniciales del problema (*Probably Approximately Correct Learning,* –PAC *Learning)* incorporando la teoría de la decisión y considerando situaciones en las que el objeto de aprendizaje es desconocido y los datos pueden estar sujetos a ruido.

Trabajos posteriores procedentes del campo computacional (p. ej. [Vidyasagar 1997], [Anthony y Bartlett 1999]), y la estadística, ([Vapnik 1995], [Vapnik 1998]), y más recientemente, [Cucker y Smale 2001], [Bousquet y Elisseeff 2001], [Bousquet y Elisseeff 2002], han ido perfilando la teoría y proporcionando enfoques alternativos.

I.1.1 Elementos del Problema de Aprendizaje

La formulación del problema de aprendizaje que presentamos a continuación pretende ser una breve síntesis actualizada recogiendo elementos de distintas contribuciones. Los elementos básicos del problema suelen ser comunes a todas ellas si bien las definiciones concretas pueden estar sujetas a modificaciones adecuadas al contexto.

De acuerdo con ello, los elementos básicos del problema de aprendizaje son los siguientes:

1. Un **generador** de vectores aleatorios (*input*) X con valores en un conjunto $\mathcal{X} \subset \mathbb{R}^d$, extraídos independientemente de una distribución P_{X}.

2. Un **supervisor** (*oracle*) que, para cada $\mathbf{x}$, produce una respuesta Y con valores en un conjunto $\mathcal{Y}$, de acuerdo con la distribución $\mathsf{P}_{\mathsf{Y}/\mathbf{x}}$. La respuesta Y puede ser vectorial en general, pero en lo que sigue, salvo indicación en contrario, se considera escalar por simplicidad.

   Este supervisor puede ser determinista o aleatorio. Casos particulares de lo anterior se tienen cuando la respuesta del supervisor viene dada a través de una función $\mathsf{f}(\mathbf{x})$ denominada *target*, cuyo resultado puede estar sujeto a ruido aleatorio. Un caso particular, es que este ruido posea carácter aditivo $\varepsilon(\mathbf{x}) = \mathsf{Y} - \mathsf{f}(\mathbf{x})$ de tal forma que, $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}(y) = \mathsf{p}_{\varepsilon(\mathbf{x})}(y - \mathsf{f}(\mathbf{x}))$ es la función de masa de probabilidad o densidad condicionada y $\mathsf{p}_{\varepsilon(\mathbf{x})}$ es la correspondiente al ruido. En particular, si $\mathsf{p}_{\varepsilon(\mathbf{x})}(y - \mathsf{f}(\mathbf{x})) = \delta[y - \mathsf{f}(\mathbf{x})]$ siendo δ la delta de Kronecker, la respuesta del supervisor es determinística y viene dada por la función $\mathsf{f}(\mathbf{x})$.

   Al espacio $\mathcal{T}$ de funciones al que se considera pertenece f se le denomina **espacio objetivo** (*target space*).

   La presencia del supervisor en esta formulación es opcional. Cuando está presente, el problema de aprendizaje se denomina de tipo **supervisado**, y cuando no, de tipo **no supervisado**[1].

3. Un **dispositivo**[2] (*learning machine*) capaz de implementar una familia de funciones medibles $\mathcal{F} = \{f : \mathcal{X} \to \widehat{\mathcal{Y}}\}$, denominada **espacio de hipótesis**, (*Hypotesis space*), donde $\widehat{\mathcal{Y}} \subset \mathcal{Y}$. En general, $\forall f \in \mathcal{F}$ escribiremos $f : \mathcal{X} \to \mathcal{Y}$ salvo que sea necesario lo contrario.

[1] Existen problemas en esencia no supervisados que por conveniencia se formulan como supervisados: por ejemplo el aprendizaje autoasociativo utilizado fundamentalmente en problemas de reducción de dimensionalidad. En ellos, el vector aleatorio $\mathbf{x}$ desempeña también el papel de respuesta del supervisor.

[2] En la teoría psicológica del aprendizaje se conoce que éste no puede producirse en el vacío, si no que es necesaria una estructura inicial. La familia de funciones implementable por este dispositivo es dicha estructura. Las limitaciones (a priori) de esta estructura determinan las limitaciones del proceso de aprendizaje.

Normalmente, la familia $\mathcal{F}$ es una familia dependiente de un conjunto de parámetros Θ de tal forma que:

$$\mathcal{F} = \{f_{\boldsymbol{\theta}} = f(\cdot, \boldsymbol{\theta}) : \mathcal{X} \to \widehat{\mathcal{Y}}, \boldsymbol{\theta} \in \Theta\}$$

donde se supone que $\mathcal{F}$ es simétrica con respecto a Θ, es decir, que cualquier permutación de las componentes del vector $\boldsymbol{\theta}$ determina la misma función $f_{\boldsymbol{\theta}} \in \mathcal{F}$.

Con frecuencia denotaremos por $\mathsf{Z} = (\mathsf{X}, \mathsf{Y})$ a la v. a. con valores en $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ y distribución $\mathsf{P}_{\mathsf{Z}} = \mathsf{P}_{(\mathsf{X},\mathsf{Y})}$.

El **problema de aprendizaje** consiste en seleccionar una función del espacio de hipótesis $\mathcal{F}$ según algún criterio de interés. Si el problema es supervisado, este criterio suele reflejar el grado de *compatibilidad o ajuste* entre el valor de la respuesta y, y el valor proporcionado por el dispositivo.

Para el caso supervisado, este criterio se formula a través de una **función de pérdida** $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{F} \to \mathbb{R}$ que en el caso no supervisado adopta la forma $\ell : \mathcal{X} \times \mathcal{F} \to \mathbb{R}$. Sin embargo, en el caso supervisado, para simplificar la exposición, utilizaremos la expresión más común $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}$ tal que $\ell(y, f(\mathbf{x}))$ determina el coste en que se incurre cuando el dispositivo produce $f(\mathbf{x})$ como valor de respuesta cuando el verdadero valor de la misma es y. En lo que sigue, salvo indicación en contrario, consideramos por defecto el caso supervisado.

Un conjunto de funciones de pérdida que será de interés, es el de las denominadas γ**-admisibles**, que verifican la siguiente condición de Lipschitz:

$$\exists \mu > 0 : |\ell(y, u_1) - \ell(y, u_2)| \leq \gamma \, |u_1 - u_2| \, , \forall u_1, u_2 \in \widehat{\mathcal{Y}}, \forall y \in \mathcal{Y}$$

Si $\mathcal{Y}$ es acotado, puede demostrarse que todas las funciones de pérdida de la forma $\ell(y, u) = |y - u|^p$ con $p \in [1, \infty)$ son admisibles.

Con todo ello, se define el **riesgo** (o riesgo teórico) como:

$$R(f) = \mathbb{E}_{(\mathsf{X},\mathsf{Y})}\{\ell(\mathsf{Y}, f(\mathsf{X}))\} = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, f(\mathbf{x})) d\mathsf{P}(\mathbf{x}, y)]$$

donde $\mathsf{P}(\mathbf{x}, y) = \mathsf{P}(y|\mathbf{x})\mathsf{P}(\mathbf{x})$ es la distribución conjunta de (X, Y), de tal manera, que el objetivo del proceso de aprendizaje es determinar la función $f^* = \arg\inf_{f\in\mathcal{F}} R(f)$.

En general, utilizaremos la siguiente notación para los valores óptimos de R:

$$\mathrm{R} = \inf_{f} R(f) \leq R^* = \inf_{f\in\mathcal{F}} R(f)$$

donde el primer ínfimo se toma sobre todas las hipótesis factibles para el problema, algunas de las cuáles no están necesariamente en $\mathcal{F}$. Si se postula la existencia de una función *verdadera* f, entonces $\mathrm{R} = R(\mathsf{f})$ y si dicha función está en $\mathcal{F}$, entonces $R^* = R(\mathsf{f}) = \mathrm{R}$.

Dada una muestra aleatoria de datos $\mathbf{z}^n = (\mathbf{z}_1, ..., \mathbf{z}_n) \in \mathcal{Z}^n$ extraída de Z, un principio inductivo frecuentemente utilizado es el denominado **minimización del riesgo empírico**, (ERM), consistente en elegir la hipótesis que minimiza el **riesgo empírico**:

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i))$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En general, salvo que se se especifique lo contrario, supondremos que la muestra aleatoria está compuesta por observaciones independientes, si bien gran parte de los resultados de la teoría del aprendizaje se están generalizando al caso de observaciones dependientes con diferentes hipótesis sobre dicha dependencia, (p. ej. [Vidyasagar 2002]).

El principio inductivo anterior es lo suficientemente general pues, como veremos a continuación, engloba principios frecuentemente utilizados en los problemas de clasificación, regresión y estimación de la densidad. Sin embargo, el principio ERM puede resultar problemático al ser excesivamente fiel a los datos, lo que, en escenarios sujetos a ruido, puede producir resultados contraproducentes.

Los elementos anteriores pueden describirse con mayor detalle considerando los siguientes conceptos adicionales:

1. El **modelo** o **arquitectura** M_p del dispositivo (con p parámetros), es decir: una función $M_p : \Theta_p \subset \mathbb{R}^p \rightarrow \mathcal{F}$ con $\mathcal{F}_p = M_p(\Theta_p) \subset \mathcal{F}$.

2. El **algoritmo**[3] A utilizado para adaptar el dispositivo (los parámetros de su arquitectura) a los datos, es decir: una función $A : \mathcal{D} \rightarrow \Theta$ donde $\mathcal{D} = \cup_{n=1}^{\infty} \mathcal{Z}^n$, con $\mathcal{Z}^n = \mathcal{Z} \times \overset{n}{\cdots} \times \mathcal{Z}$, es el conjunto de posibles muestras extraídas de Z, de tal forma que $M_p(A(\mathbf{z}^n)) \in \mathcal{F}$ es la hipótesis producida por el modelo M_p con el algoritmo A a partir de los datos $\mathbf{z}^n$.

   Se denomina **algoritmo minimizador del riesgo empírico** (*sample error minimization algorithm* –SEM, [Anthony y Bartlett 1999]), a un algoritmo que verifique:

$$A(\mathbf{z}^n) = \arg \min_{\boldsymbol{\theta} \in \Theta_p} \hat{R}_n(f_{\boldsymbol{\theta}})$$

   es decir, que implemente el principio ERM.

La formulación anterior del problema responde básicamente al enfoque PAC completado. La formulación PAC inicial consideraba únicamente el problema de clasificación con una función (concepto) *target* explícitamente considerada dentro de la familia de hipótesis (**realizable** por la familia), con ausencia de ruido en la respuesta del supervisor y pérdida 0-1. Generalizaciones posteriores al enfoque PAC ya incluidas en la formulación anterior, fueron[4]:

1. La consideración del aprendizaje de funciones reales.

2. La consideración de escenarios de ruido en la respuesta del supervisor.

3. La suposición de que la función *target* puede no estar en la familia de hipótesis considerada (**no realizable**) e, incluso, el cuestionamiento de su propia existencia. Este caso extremo se denomina aprendizaje **PAC agnóstico**, (p. ej. [Kearns *et al.* 1994]).

Finalmente, la formulación anterior puede verse como una aplicación del problema estadístico de decisión en el que: la distribución $\mathbb{P}_{(\mathsf{X},\mathsf{Y})}$ es un estado de la naturaleza dentro del conjunto

[3] En la teoría de aproximación de funciones ([DeVore *et al.* 1989], [Lorentz 1986]) se define algoritmo como una función $A : \mathcal{T} \rightarrow \Lambda$ de tal forma que $M_p(A(\mathsf{f}))$ es la aproximación a f producida por el modelo M_p con el algoritmo A. Aquí la función f se transmite a través de un conjunto de pares $(\mathbf{x}_i, y_i)$ a través de cuyos y_i el *oracle* indica el valor $\mathsf{f}(\mathbf{x}_i)$ (posiblemente con ruido superpuesto).

[4] Existen diversas extensiones a esta formulación, entre las que destacamos la posibilidad de que los algoritmos de aprendizaje permitan realizar experimentos o preguntas al oracle durante el aprendizaje, (.Active Learning"). Un ejemplo de este tipo de preguntas es el *membership query* en la que, el algoritmo, propone ejemplos en $\mathcal{X}$ para que sean catalogados por el oracle con su valor de y. Este mecanismo permite incrementar sensiblemente la eficacia del aprendizaje, (p. ej. [Angluin 1988]).

de estados posibles (distribuciones posibles $\mathcal{P}$), que produce un dato $(\mathbf{x}, y)$ en el que la respuesta y del supervisor es desconocida por el decisor (dispositivo) e independiente de la acción $f(\mathbf{x})$ adoptada por éste siguiendo una regla de decisión f perteneciente a un espacio de reglas $\mathcal{F}$, la cuál, produce una pérdida $\ell(y, f(\mathbf{x}))$.

I.1.2 Problema de Aprendizaje Bien Planteado

Por otra parte, en lo anterior, hemos supuesto implícitamente que el problema de minimización del principio ERM, es resoluble, lo cual, no tiene por qué suceder en la práctica. Por ello, conviene traer a colación la definición de *problema bien planteado* de la teoría del problema inverso, (p. ej. [Tikhonov y Arsenin 1977] y [Bertero 1986]), e incorporarla al planteamiento del problema de aprendizaje.

Definición 1 ***Problema bien planteado***. *Sean $\mathcal{F}$, $\mathbb{Y}$ espacios métricos denominados dominio y espacio de datos del problema, con métricas $\rho_{\mathcal{F}}$ y $\rho_{\mathbb{Y}}$, respectivamente. Sea un operador $\mathcal{O} : \mathcal{F} \to \mathbb{Y}$. De la ecuación $\mathcal{O}(f) = \mathbf{y} \in \mathbb{Y}$ con $f \in \mathcal{F}$ se dice que es un problema bien planteado si $\mathcal{O}$ es biyectivo y el operador inverso $\mathcal{O}^{-1} : \mathbb{Y} \to \mathcal{F}$ es continuo. En caso contrario se dice que el problema está mal planteado.*

En un problema de aprendizaje, $\mathcal{F}$ es el espacio de hipótesis[5], $\mathbb{Y} \equiv \mathcal{Y}^n$ es el espacio de datos con $\mathbf{y} = (y_1, ..., y_n)$, el operador inverso $\mathcal{O}^{-1}$ a determinar (de ahí el nombre de **teoría del problema inverso**) es el algoritmo $A \equiv \mathcal{O}^{-1} : \mathcal{Y}^n \to \mathcal{F}$ generador de hipótesis[6] $f = A(\mathbf{y})$, y, por último, la métrica $\rho_{\mathbb{Y}}$ es el riesgo empírico:

$$\rho_{\mathcal{Y}^n}(\mathbf{y}^1, \mathbf{y}^2) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i^1, y_i^2)$$

donde ℓ es la función de pérdida considerada[7].

La condición de continuidad (**estabilidad**) del operador inverso está inspirada en el hecho de que los fenómenos que podemos aprender son los que poseen algún grado de regularidad: observaciones con cierto grado de similitud deben estar producidos por procesos reales también similares (el éxito de la inferencia (aprendizaje) depende de que esta regularidad supuesta sea realmente cierta). Por tanto, la condición de estabilidad se formula como:

$$\forall f_1, f_2 \in \mathcal{F}, \forall \varepsilon > 0, \exists \delta > 0 : \rho_{\mathbb{Y}}(\mathcal{O}(f_1), \mathcal{O}(f_2)) < \delta \Rightarrow \rho_{\mathcal{F}}(f_1, f_2) < \varepsilon$$

Por el lema de continuidad de un operador inverso [Tikhonov y Arsenin 1977, p. 29], si el operador $\mathcal{O}$ es continuo y biyectivo, y está definido sobre un espacio métrico compacto $\mathcal{F}_0 \subset \mathcal{F}$, entonces el operador inverso $\mathcal{O}^{-1}$ es continuo en su dominio $\mathcal{O}(\mathcal{F}_0) \subset \mathbb{Y}$.

Si $\mathcal{F}$ es una familia parametrizada por un espacio de parámetros Θ, el operador $\mathcal{O}$ es continuo, si lo es (con respecto a $\boldsymbol{\theta}$) la expresión general $f_{\boldsymbol{\theta}}$ de las funciones de $\mathcal{F}$ (lo que suele ser el

[5] O bien, un espacio de parámetros Θ sobre el que se define un modelo o arquitectura $M : \Theta \to \mathcal{F}_\Theta$ donde $\mathcal{F} \equiv \mathcal{F}_\Theta$ es el espacio de hipótesis.

[6] Si los datos hubiesen sido generados por una red neuronal (lo que usualmente se denomina aprendizaje con profesor –*teacher*), dicha red sería precisamente la determinada por el algoritmo $A \equiv \mathcal{O}^{-1}$ que se trata de recuperar.

[7] Como veremos, en los modelos de redes neuronales: la **existencia** de solución depende de la capacidad de representación de la familia para representar el proceso generador de los datos, y la **unicidad** no está garantizada, pues, puede haber varias funciones $f \in \mathcal{F}$ implementables por una misma arquitectura de red dada, que se ajusten a los datos, (al margen de las clases de equivalencia definidas por las permutaciones de sus parámetros).

USC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

caso en las redes neuronales). Por tanto, la estabilidad se logra restringiendo el problema a algún subespacio compacto $\mathcal{F}_0 \subset \mathcal{F}$.

Este es el camino seguido por el método de los *Sièves*, [Grenander 1981] que considera una sucesión creciente $\{\Theta_v\}$ de conjuntos compactos en el espacio Θ de los parámetros tal que $\cup_v \Theta_v$ es denso en Θ. Una vía alternativa es la del método de **regularización** que pretende aproximar el operador inverso $\mathcal{O}^{-1}$ mediante una sucesión de operadores estables, mediante el concepto de operador regularizador $\mathcal{Q}$, que presentaremos con detalle más adelante. La idea básica es considerar como principio inductivo el **principio ERM regularizado**, de la siguiente forma:

$$\hat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)) + \lambda \mathcal{E}(f) \tag{I.1}$$

donde $\mathcal{E}$ es un funcional definido en $\mathcal{F}$.

I.2 Ejemplos de Problemas de Aprendizaje

La formulación anterior del problema de aprendizaje incluye como casos particulares los siguientes más relevantes, en los que se postula la existencia de una función *verdadera* f:

I.2.1 Problema de Regresión

Si el supervisor (*oracle*) produce un valor real, $\mathcal{F}$ es una familia de funciones con valores reales, (clase de funciones), se adopta la pérdida cuadrática $\ell(y, f(\mathbf{x})) = [y - f(\mathbf{x})]^2$, y se postula una función de regresión $\mathsf{f}(\mathsf{X}) = \mathbb{E}_{\mathsf{Y}/\mathsf{X}}(\mathsf{Y})$, el riesgo puede descomponerse como:

$$\begin{aligned} R(f) &= \mathbb{E}_{\mathsf{Z}}\{[\mathsf{Y} - f(\mathsf{X})]^2\} = \mathbb{E}_{\mathsf{Z}}[(\mathsf{Y} - \mathsf{f}(\mathsf{X}) + \mathsf{f}(\mathsf{X}) - f(\mathsf{X}))^2] \\ &= \mathbb{E}_{\mathsf{X}}\mathbb{E}_{\mathsf{Y}/\mathsf{X}}[(\mathsf{Y} - \mathsf{f}(\mathsf{X}))^2] + \mathbb{E}_{\mathsf{X}}[(f(\mathsf{X}) - \mathsf{f}(\mathsf{X}))^2] \\ &\quad -2\mathbb{E}_{\mathsf{X}}\{(f(\mathsf{X}) - \mathsf{f}(\mathsf{X}))\mathbb{E}_{\mathsf{Y}/\mathsf{X}}[(\mathsf{Y} - \mathsf{f}(\mathsf{X}))]\} \\ &= R(\mathsf{f}) + \mathbb{E}_{\mathsf{X}}[(f(\mathsf{X}) - \mathsf{f}(\mathsf{X}))^2] \end{aligned} \tag{I.2}$$

donde el primer término no depende de la hipótesis seleccionada y el segundo se minimiza cuando $f(\mathbf{x}) = \mathsf{f}(\mathbf{x})$. Con ello, si se utiliza la pérdida cuadrática, la formulación general del problema de aprendizaje da lugar al problema de regresión.

El riesgo empírico para este problema es:

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)) = \frac{1}{n}\sum_{i=1}^{n} [y_i - f(\mathbf{x}_i)]^2$$

con lo que, con la pérdida cuadrática, el principio inductivo ERM recomienda utilizar el criterio de mínimos cuadrados para seleccionar la mejor hipótesis.

Si el ruido participa de forma aditiva en la respuesta $\mathsf{Y} = \mathsf{f}(\mathbf{x}) + \varepsilon(\mathbf{x})$ con $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}(y) = \mathsf{p}_{\varepsilon(\mathbf{x})}(y - \mathsf{f}(\mathbf{x}))$, entonces una función de pérdida conveniente por sus buenas propiedades asintóticas es el menos logaritmo de la verosimilitud:

$$\ell(y, f(\mathbf{x})) = -\ln \mathsf{p}_{\varepsilon(\mathbf{x})}(y - f(\mathbf{x}))$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

	Densidad	**Pérdida**
Gausiana	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}r^2)$	$\frac{1}{2}r^2$
Laplaciana	$\frac{1}{2}\exp(-\lvert r\rvert)$	$\lvert r\rvert$
Huber	$\propto \begin{cases} \exp(-\frac{r^2}{2\sigma}) & \lvert r\rvert \le \sigma \\ \exp(\frac{\sigma}{2}-\lvert r\rvert) & \text{resto} \end{cases}$	$\begin{cases} \frac{1}{2\sigma}r^2 & \lvert r\rvert \le \sigma \\ \lvert r\rvert - \frac{\sigma}{2} & \text{resto} \end{cases}$
ε-insensible	$\frac{1}{2(1+\varepsilon)}\exp(-\max(\lvert r\rvert-\varepsilon,0))$	$\max(\lvert r\rvert-\varepsilon,0)$

Tabla 1. Funciones de densidad y funciones de pérdida a las que dan lugar en el problema de regresión

Si además p_ε no depende de $\mathbf{x}$ y se modeliza como una densidad gausiana de media cero, se obtiene como caso particular la pérdida de mínimos cuadrados. En este caso, si las observaciones son independientes, el riesgo empírico coincide con el anterior.

La consideración de otros modelos para p_ε da lugar a diferentes funciones de pérdida (véase la Tabla 1 para el problema de regresión con distribuciones uniformemente continuas).

I.2.2 Problema de Clasificación

Si el supervisor asigna valores respuesta $\mathsf{Y} \in \{0,1\}$ y $\mathsf{f} = 1_B$ con $B \in \mathcal{A}_\mathsf{X}$, σ-álgebra de X, en la formulación PAC del aprendizaje, (p. ej. [Vidyasagar 1997], [Haussler 1992]), la familia de hipótesis $\mathcal{F} = \{1_S : S \in \mathcal{A}_\mathsf{X}\}$ suele denominarse **clase de conceptos**, denominándose el problema, **aprendizaje de conceptos** o de conjuntos (conjuntos de elementos de $\mathcal{X}$ que satisfacen una propiedad o concepto), pues aprender un conjunto $B \subset \mathcal{A}_\mathsf{X}$ (o la función $1_B \in \mathcal{T}$) equivale a un problema de clasificación en el que el supervisor asigna $y = 1$ si $\mathbf{x} \in B$.

Si se considera la función de pérdida 0-1: $\ell(y, g(\mathbf{x})) = 1_{\{y \neq g(\mathbf{x})\}}$, el riesgo $R(g)$ representa la probabilidad de error en la clasificación realizada por la regla g :

$$\begin{aligned} R(g) &= \int \ell(y, g(\mathbf{x}))d\mathsf{P}(\mathbf{x}, y) \\ &= \int_{\{y \neq g(\mathbf{x})\}} 1_{\{g(\mathbf{x}) \neq y\}} d\mathsf{P}(\mathbf{x}, y) + \int_{\{y = g(\mathbf{x})\}} 1_{\{g(\mathbf{x}) \neq y\}} d\mathsf{P}(\mathbf{x}, y) \\ &= \mathbb{P}[\mathsf{Y} \neq g(\mathsf{X})] \end{aligned}$$

con lo que el objetivo del aprendizaje es determinar el clasificador g que minimiza la probabilidad de error de clasificación. Dada una muestra de datos $\mathbf{z}^n$, el riesgo empírico es:

$$\hat{R}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, g(\mathbf{x}_i)) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i \neq g(\mathbf{x}_i)\}}$$

con lo que el principio ERM recomienda seleccionar la hipótesis que minimiza el número de errores de clasificación en la muestra.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Por otra parte, si se utiliza la pérdida cuadrática $\ell(y, g(\mathbf{x})) = [y - g(\mathbf{x})]^2$ con hipótesis dicotómicas $\mathcal{F} = \{g_G = 1_G : G \in \mathcal{A}_\mathsf{X}\}$, entonces $[y - g(\mathbf{x})]^2 = 1$ si $g(\mathbf{x}) \neq y$, y cero en caso

contrario, con lo que para $g \in \mathcal{F}$ se tiene que:

$$\begin{aligned} R(g) &= \mathbb{E}_{\mathsf{Z}}[(\mathsf{Y} - g(\mathsf{X}))^2] = \int_{\mathcal{Z}} (y - g(\mathbf{x}))^2 d\mathsf{P}(\mathbf{x}, y) \\ &= \int_{\{y \neq g(\mathbf{x})\}} d\mathsf{P}(\mathbf{x}, y) = \mathbb{P}[\mathsf{Y} \neq g(\mathsf{X})] \end{aligned}$$

con lo que, el riesgo empírico también reflejaría la tasa de error de clasificación en los datos de la muestra. Por tanto, con hipótesis dicotómicas con valores en $\{0, 1\}$, la pérdida 0-1 equivale a la pérdida cuadrática.

Con las dos pérdidas anteriores (con la cuadrática cuando se utilizan hipótesis dicotómicas) la hipótesis óptima es la regla de Bayes, pero este hecho también se verifica para la pérdida absoluta $\ell(y, g(\mathbf{x})) = |y - g(\mathbf{x})|$ sin que $\mathcal{F}$ deba restringirse a hipótesis dicotómicas [Devroye *et al.* 1996, p. 20].

Un método usual para obtener reglas de clasificación dicotómicas g_f es obtener una estimación previa f de la probabilidad a posteriori $\mathbb{P}(\mathsf{Y} = 1|\mathbf{x})$ y utilizar posteriormente el denominado método *plug-in*:

$$g_f(\mathbf{x}) = \begin{cases} 0 & \text{si } f(\mathbf{x}) \leq \frac{1}{2} \\ 1 & \text{si } f(\mathbf{x}) > \frac{1}{2} \end{cases}$$

Ahora el espacio $\mathcal{F}$ es el conjunto de posibles hipótesis $f(\mathbf{x})$ sobre $\mathbb{P}(\mathsf{Y} = 1|\mathbf{x})$. Al igual que en el problema de regresión, una función de pérdida natural en este caso es el menos logaritmo de la verosimilitud. Ahora, la distribución a posteriori es un modelo de Bernouilli:

$$\mathsf{p}_{\mathsf{Y}/\mathbf{x}}(y) = \mathsf{f}(\mathbf{x})^y [1 - \mathsf{f}(\mathbf{x})]^{1-y}$$

con lo que, si $q(y; \mathbf{x}) = f(\mathbf{x})^y [1 - f(\mathbf{x})]^{1-y}$ es la función de masa o de densidad postulada para $\mathsf{Y}/\mathbf{x}$ determinada por f, la pérdida asociada es:

$$\ell(y, f(\mathbf{x})) = -\ln q(y; \mathbf{x}) = -[y \ln f(\mathbf{x}) + (1 - y) \ln(1 - f(\mathbf{x}))] \tag{I.3}$$

conocida como **función de error de entropía cruzada** (p. ej. [Bishop 1995], [Ripley 1996]). El riesgo empírico resulta:

$$\hat{R}_n(f) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \ln f(\mathbf{x}_i) + (1 - y_i) \ln(1 - f(\mathbf{x}_i))]$$

Para el caso de una codificación $\mathsf{Y} \in \{-1, 1\}$, se tiene:

$$q(y; \mathbf{x}) = f(\mathbf{x})^{(1+y)/2} [1 - f(\mathbf{x})]^{(1-y)/2}$$

lo que da lugar a la pérdida:

$$\ell(y, f(\mathbf{x})) = -\ln q(y; \mathbf{x}) = -[\frac{1 + y}{2} \ln f(\mathbf{x}) + \frac{1 - y}{2} \ln(1 - f(\mathbf{x}))]$$

Finalmente, si para esta misma codificación $\mathcal{Y} = \{-1, 1\}$, la familia $\mathcal{F}$ es el conjunto de posibles hipótesis $h(\mathbf{x})$ para[8] $\mathsf{h}(\mathbf{x}) = \ln \frac{\mathbb{P}(\mathsf{Y}=1|\mathbf{x})}{\mathbb{P}(\mathsf{Y}=-1|\mathbf{x})}$, entonces ambas familias están relacionadas

[8] Suponiendo iguales probabilidades a priori para las clases.

Denominación	**Pérdida** $\ell(y, h(\mathbf{x}))$
Logística	$\ln(1 + e^{-yh(\mathbf{x})})$
Cuadrática en $h(\mathbf{x})$	$[y - h(\mathbf{x})]^2 = y^2[y - h(\mathbf{x})]^2 = [1 - yh(\mathbf{x})]^2$
Cuadrática en $f(\mathbf{x})$	$[y^* - f(\mathbf{x})]^2 = [1 + e^{yh(\mathbf{x})}]^{-2}$
Exponencial	$e^{-yh(\mathbf{x})}$

Tabla 2. Funciones de pérdida en el problema de regresión logística con codificación $\{-1, 1\}$. En la tercera, $y^* = (1 + y)/2 \in \{0, 1\}$, $f(\mathbf{x}) = (1 + e^{-h(\mathbf{x})})^{-1}$ es la hipótesis para $\mathbb{P}(\mathsf{Y} = 1|\mathbf{x})$ y la expresión final se obtiene fácilmente.

pues para cada $f(\mathbf{x})$ anterior existe una única $h(\mathbf{x})$ que verifica:

$$e^{h(\mathbf{x})} = \frac{f(\mathbf{x})}{1 - f(\mathbf{x})} \Rightarrow \begin{cases} f(\mathbf{x}) = \dfrac{e^{h(\mathbf{x})}}{1 + e^{h(\mathbf{x})}} = \dfrac{1}{1 + e^{-h(\mathbf{x})}} \\ 1 - f(\mathbf{x}) = 1 - \dfrac{1}{1 + e^{-h(\mathbf{x})}} = \dfrac{1}{1 + e^{h(\mathbf{x})}} \end{cases}$$
$$\Rightarrow \quad q(y; \mathbf{x}) = \frac{1}{1 + e^{-yh(\mathbf{x})}} = q(y^*; \mathbf{x}) \tag{I.4}$$

con $y^* = \frac{1}{2}(y + 1) \in \{0, 1\}$. Con ello, se obtiene la pérdida:

$$\ell(y, h(\mathbf{x})) = -\ln q(y; \mathbf{x}) = \ln(1 + e^{-yh(\mathbf{x})}) \equiv \ell(y^*, h(\mathbf{x})) \tag{I.5}$$

que se denomina función de **pérdida logística**. Si $\mathcal{F}$ es la familia de funciones $h(\mathbf{x})$ lineales en $\mathbf{x}$ se tiene el problema de regresión logística.

La tabla 2 muestra otras pérdidas utilizadas en el problema de regresión logística (figura 1). Como puede observarse en dicha tabla, la pérdida cuadrática en $h(\mathbf{x})$ penaliza los ejemplos demasiado bien clasificados.

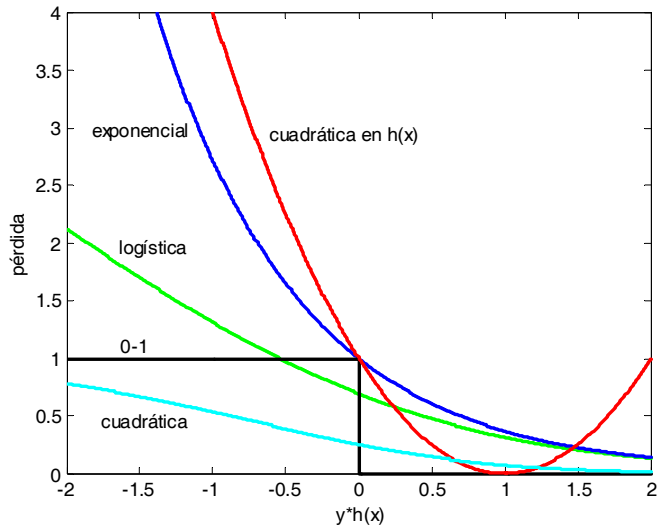


Figura 1. Diferentes funciones de pérdida (tabla 2) para el problema de clasificación con codificación $\{-1, +1\}$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

I.2.3 Problema de Estimación de la Densidad

Si se considera un problema de aprendizaje no supervisado en el que el espacio de hipótesis $\mathcal{F}$ es una familia de funciones de densidad (denotadas mediante la letra q), y se utiliza la pérdida $\ell(\mathbf{x}, q) = -\ln q(\mathbf{x})$, el riesgo es la denominada **entropía cruzada** (*cross entropy*):

$$\begin{aligned} R(q) &= -\int \ln q(\mathbf{x}) \mathsf{p}_{\mathsf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \int \ln \frac{\mathsf{p}_{\mathsf{X}}(\mathbf{x})}{q(\mathbf{x})} \mathsf{p}_{\mathsf{X}}(\mathbf{x}) d\mathbf{x} - \int \ln \mathsf{p}_{\mathsf{X}}(\mathbf{x}) \mathsf{p}_{\mathsf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \mathbb{E}_{\mathsf{X}}[\ln \mathsf{p}_{\mathsf{X}}(\mathsf{X}) - \ln q(\mathsf{X})] + R(\mathsf{p}_{\mathsf{X}}) \\ &= D_{q,\mathsf{p}_{\mathsf{X}}} + R(\mathsf{p}_{\mathsf{X}}) \end{aligned}$$

donde $D_{q,\mathsf{p}_{\mathsf{X}}}$ es la denominada **información (o divergencia) de Kullback-Leibler** [Kullback 1968] de q sobre la verdadera densidad p_{X} y $R(\mathsf{p}_{\mathsf{X}})$ es la entropía (constante) de p_{X}. Como $R(\mathsf{p}_{\mathsf{X}})$ no depende de la hipótesis elegida, $D_{q,\mathsf{p}_{\mathsf{X}}}$ puede elegirse también como riesgo, (aunque no posee propiedades de métrica).

El mínimo de la expresión anterior (decisión óptima de Bayes), se alcanza para la verdadera densidad p_{X} : igualando a cero la derivada funcional respecto a q de la función objetivo $R(q) + \lambda[\int q(\mathbf{x})d\mathbf{x} - 1]$ siendo λ el multiplicador de Lagrange, se tiene: $-\frac{\mathsf{p}_{\mathsf{X}}(\mathbf{x})}{q(\mathbf{x})} + \lambda = 0 \Rightarrow q(\mathbf{x}) = \mathsf{p}_{\mathsf{X}}(\mathbf{x})/\lambda$ con lo que sustituyendo en la restricción:

$$1 = \int q(\mathbf{x})d\mathbf{x} = \int \frac{\mathsf{p}_{\mathsf{X}}(\mathbf{x})}{\lambda} d\mathbf{x} = \frac{1}{\lambda} \Rightarrow \lambda = 1 \Rightarrow q(\mathbf{x}) = \mathsf{p}_{\mathsf{X}}(\mathbf{x})$$

De esta forma, el problema de la estimación de la densidad encaja también en la formulación general del problema de aprendizaje, en este caso no supervisado, utilizando la pérdida mencionada.

El principio inductivo ERM pretenden minimizar[9]:

$$\hat{R}_n(q) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i, q) = -\frac{1}{n} \sum_{i=1}^{n} \ln q(\mathbf{x}_i)$$

que coincide con el principio de máxima verosimilitud cuando las observaciones son independientes.

Un caso particular de este problema es el de estimación de los parámetros de una familia de distribuciones dada, $\{q(\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, en cuyo caso el riesgo es $R(q) \equiv R(\boldsymbol{\theta}) = -\int \ln q(\mathbf{x}, \boldsymbol{\theta}) \mathsf{p}(\mathbf{x}) d\mathbf{x}$ y el riesgo empírico: $\hat{R}_n(\boldsymbol{\theta}) = -1/n \sum_{i=1}^{n} \ln q(\mathbf{x}, \boldsymbol{\theta})$.

Como puede observarse, tanto el problema de regresión como el de clasificación con el método *plug-in*, pueden formularse también como problema de aprendizaje de la función de densidad condicionada $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$: si como espacio de hipótesis sobre dicha densidad se considera la familia de distribuciones gausianas, el planteamiento anterior da lugar a la pérdida cuadrática y el problema de regresión por mínimos cuadrados. Si como espacio de hipótesis se considera la familia de distribuciones de Bernouilli, se obtiene como pérdida la función de error de entropía cruzada (I.3).

[9] El término $2\sum_{i=1}^{n}[\ln \mathsf{p}(\mathbf{x}_i) - \ln f(\mathbf{x}_i)] = -2\sum_{i=1}^{n} \ln \frac{f(\mathbf{x}_i)}{\mathsf{p}(\mathbf{x}_i)} = 2n\hat{R}_n(f) + 2n \ln \mathsf{p}(\mathbf{x}_i)$ suele denominarse *deviance*, (p. ej. [Ripley 1996]).

Todo ello, al margen de que, realmente, $p_{Y/\mathbf{x}} \in \mathcal{F}$, o no, en cuyo caso, el riesgo representa igual que antes la discrepancia de Kullback-Leibler, (método de la quasi-verosimilitud, p. ej. [White 1982], [White 1992], [White 1994]).

Por último, en los ejemplos anteriores, se ha supuesto la utilización del principio ERM pero, como ya hemos indicado, puede ser conveniente o necesario utilizar otros principios, como el principio ERM regularizado, ya mencionado.

I.3 Condiciones para el Éxito del Aprendizaje

El objetivo del aprendizaje es obtener una hipótesis con la máxima capacidad de generalización, es decir, que permita reproducir, a partir de un número finito de datos, el proceso generador de los mismos de la mejor forma posible. Siempre que el problema esté bien planteado, ello requiere que el proceso de aprendizaje sea consistente, es decir, que, a medida de que se dispone de nuevos datos, la hipótesis seleccionada converja a la mejor hipótesis posible de la familia $\mathcal{F}$ con el criterio de la función de pérdida elegida. Si existiese una función verdadera de interés y fuese realizable por la familia $\mathcal{F}$, la consistencia consistiría en la convergencia a dicha función.

Sin embargo, esta condición no es suficiente. En primer lugar, dicha convergencia puede ser lo suficientemente lenta en la práctica como para que no sea de utilidad. Por otra parte, la familia puede no poseer la complejidad suficiente como para reproducir adecuadamente el proceso de interés. Finalmente, el algoritmo utilizado para seleccionar la hipótesis puede requerir una carga computacional insalvable.

Por tanto, las cuatro condiciones fundamentales para el éxito del aprendizaje, además de la condición previa ya expuesta de que el problema esté **bien planteado**, son las siguientes:

1. La **consistencia** del proceso de aprendizaje, en su vertiente puramente estadística, que no tiene en cuenta su viabilidad computacional.

2. La **velocidad de convergencia** teórica, que puede ser lo suficientemente lenta como para que la consistencia no sea de utilidad.

   La velocidad de convergencia es importante pues permite conocer, dado un tamaño muestral n, lo cerca que la solución proporcionada por el algoritmo está de la solución óptima en la familia $\mathcal{F}$. Y a la inversa, qué tamaño muestral se requiere para que la hipótesis seleccionada se encuentre a determinada distancia (en términos del riesgo) de la solución óptima en $\mathcal{F}$, (**complejidad muestral** en terminología PAC).

3. La **capacidad** de la familia $\mathcal{F}$ de hipótesis considerada, para reproducir el proceso real de interés. En principio, la familia ha de ser suficientemente compleja, sin embargo, en contrapartida, ello puede originar problemas de estabilidad (problema bien planteado) y, en definitiva, de consistencia.

   Por tanto, esta condición puede entrar en conflicto con las anteriores. Para evitarlo, el algoritmo de aprendizaje debe incorporar mecanismos de **control de la complejidad** en consonancia con la cantidad de datos disponibles.

4. La viabilidad algorítmica del proceso de aprendizaje. Esta condición es estudiada por el aprendizaje PAC bajo la denominación de **complejidad computacional** del aprendizaje, cuyo elemento esencial es el concepto de **eficiencia** del algoritmo, (p. ej. [Anthony y Bartlett 1999], [Kearns y Schapire 1994]).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En esta dimensión computacional, el enfoque PAC persigue obtener resultados generales sobre la imposibilidad algorítmica de la consistencia, o condiciones necesarias y suficientes para su posibilidad que sirvan de marco para el estudio de problemas particulares.

Los tres primeros aspectos constituyen la dimensión estadística del proceso de aprendizaje. Su estudio permite evaluar el éxito potencial del mismo, conocer su comportamiento a medida que se dispone de nuevos datos, obtener criterios para seleccionar la hipótesis con mayor capacidad de generalización (**selección del modelo**) y determinar su fiabilidad.

I.3.1 Consistencia del Proceso de Aprendizaje

La palabra *consistencia* posee diferentes significados en el campo del aprendizaje de dispositivos (*machine learning*), además de entre diferentes enfoques seguidos por distintos grupos de investigadores.

En esencia, el proceso de aprendizaje es **consistente** si la hipótesis seleccionada por el dispositivo a partir de un conjunto de datos $\mathsf{Z}^n = (\mathsf{Z}_1, ..., \mathsf{Z}_n)$, converge con el criterio especificado, a la hipótesis óptima de la familia de hipótesis considerada a medida que se incrementa el número de datos disponible, $(n \to \infty)$.

Esta idea general se concreta de forma ligeramente diferente en las dos principales teorías del aprendizaje que, debido a su pretensión de universalidad, pueden calificarse como tales: la Teoría Estadística del Aprendizaje (*Statistical Learning Theory* –SLT) de Vapnik, ([Vapnik y Chervonenkis 1971], [Vapnik y Chervonenkis 1981], [Vapnik 1982], [Vapnik 1995], [Vapnik 1998]), y la teoría del Aprendizaje Probablemente Aproximadamente Correcto (PAC *learning*), ([Valiant 1984], [Angluin 1988], [Haussler 1992], [Anthony y Bartlett 1999]).

Los dos enfoques consideran el principio de minimización del riesgo empírico, (ERM):

1. En la SLT, la consistencia se define a través del concepto de **consistencia no trivial del principio de minimización del riesgo empírico**, como (véase notación en Anexo H):

$$\lim_{n\to\infty} \mathbb{P}_{\mathsf{Z}^n}\{\mathbf{z}^n \in \mathcal{Z}^n : \left| \inf_{f\in\mathcal{F}} R(f) - \inf_{f\in\mathcal{F}} \hat{R}_n(f) \right| > \varepsilon\} = 0$$

lo que denotamos brevemente por $\operatorname{plim}_{n\to\infty} \inf_{f\in\mathcal{F}} \hat{R}_n(f) = \inf_{f\in\mathcal{F}} R(f)$, donde $\hat{R}_n$ es el riesgo empírico obtenido a partir de Z^n, y R es el riesgo.

Si dicha condición se verifica $\forall \mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$, se dice que el principio ERM es consistente (no trivialmente) para cualquier distribución, (*distribution free*).

Nótese que en condiciones poco restrictivas en la práctica (p. ej. R acotado), para cada $f \in \mathcal{F}$, las leyes de los grandes números afirman $\operatorname{plim}_{n\to\infty} \hat{R}_n(f) = R(f)$ pues $\mathbb{E}_{\mathsf{Z}^n}[\hat{R}_n(f)] = R(f)$, y parecería que la condición no aporta nada nuevo. Sin embargo, la convergencia para cada $f \in \mathcal{F}$ no garantiza que el ínfimo de $\hat{R}_n$ converja al ínfimo de R.

La satisfacción de esta propiedad requiere que *todas* las funciones de la familia, a la vez, se comporten *bien* (uniformemente): **condición necesaria y suficiente** para la consistencia anterior bajo cualquier medida de probabilidad (análogamente bajo una medida de probabilidad $\mathbb{P}_{\mathsf{Z}}$ prefijada) es que $\mathcal{F}$ verifique la propiedad de **convergencia uniforme de medias a las**

esperanzas, uniformemente en distribución, (CUMEUD):[10] $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\mathbb{P}_Z \in \mathcal{P}} \mathbb{P}_{Z^n}\{\mathbf{z}^n \in \mathcal{Z}^n : \sup_{f\in\mathcal{F}} \left| R(f) - \hat{R}_n(f) \right| > \varepsilon\} = 0$$

2. En el enfoque PAC, el concepto de consistencia se expresa en términos de un algoritmo que produce hipótesis pertenecientes a la familia $\mathcal{F}$.

   Así, se denomina **algoritmo PAC** (o simplemente algoritmo de aprendizaje, p. ej. [Anthony y Bartlett 1999]) a un algoritmo $A : \cup_{n=1}^{\infty} \mathcal{Z}^n \to \mathcal{F}$ tal que, $\forall \varepsilon > 0$,

   $$\lim_{n\to\infty} \sup_{\mathbb{P}_Z \in \mathcal{P}} \mathbb{P}_{Z^n}\{\mathbf{z}^n \in \mathcal{Z}^n : R(A(\mathbf{z}^n)) - \inf_{f\in\mathcal{F}} R(f) > \varepsilon\} = 0 \tag{I.6}$$

   donde $A(\mathbf{z}^n) \in \mathcal{F}$ es la hipótesis proporcionada por el algoritmo A cuando se entrena con la muestra $\mathbf{z}^n$ (no se especifica el principio inductivo pues se entiende implícito en el algoritmo). Asimismo, se dice que $\mathcal{F}$ es **aprendible**[11], (*learnable*), si existe un algoritmo PAC para ella.

   Este concepto de consistencia es el usual en estadística (para toda distribución) y se denomina **consistencia universal débil** (o consistencia universal en probabilidad o, simplemente, consistencia universal, p. ej. [Devroye *et al.* 1996]). Su contrapartida **fuerte** sería:

   $$\sup_{\mathbb{P}_Z \in \mathcal{P}} \mathbb{P}_Z^{\infty}\{\mathbf{z}^{\infty} \in \mathcal{Z}^{\infty} : \lim_{n\to\infty} R(A(\mathbf{z}^n)) - \inf_{f\in\mathcal{F}} R(f) > \varepsilon\} = 0 \tag{I.7}$$

   donde, $\mathbb{P}_Z^{\infty}$ denota la única medida de probabilidad definida en el espacio probabilizable $(\mathcal{Z}^{\infty}, \mathcal{A}_Z^{\infty})$ con $\mathcal{Z}^{\infty} = \Pi_{n=1}^{\infty} \mathcal{Z}_n$ producto cartesiano infinito con $\mathcal{Z}_n = \mathcal{Z}\ \forall n \in \mathbb{N}$ y $\mathcal{A}_Z^{\infty} = \{B = \Pi_{n=1}^{\infty} B_n : B_n \in \mathcal{Z}_n\}$, que verifica $\mathbb{P}_Z^{\infty}(B) = \Pi_{n=1}^{\infty} \mathbb{P}_Z(B_n)$ (teorema de extensión de Kolmogorov [Vidyasagar 1997, p. 25]).

   Por analogía con los conceptos anteriores, una sucesión de estimadores $\{\hat{f}_n = A(Z^n)\}_{n=1}^{\infty}$ se dice **universalmente consistente** si se verifica la condición (I.6) y **universalmente fuertemente consistente** si verifica la condición[12] (I.7).

   La relación entre este concepto de consistencia y el anterior de la SLT se basa en el concepto de algoritmo SEM definido con anterioridad y equivalente al principio ERM a través del siguiente resultado [Vidyasagar 1997]: si $\mathcal{F}$ verifica la propiedad CUMEUD, y A es un algoritmo SEM para $\mathcal{F}$, entonces $\mathcal{F}$ es PAC-aprendible (A es PAC)[13].

   Este resultado indica que, si la familia verifica la propiedad CUMEUD, entonces el principio ERM (implementado por un algoritmo SEM) es consistente en el sentido PAC. Como la propiedad CUMEUD era necesaria y suficiente para la consistencia no trivial de la SLT, ésta es más fuerte [Vapnik 1998, pág. 82] que la consistencia PAC, (aprendibilidad PAC).

Como puede observarse, el enfoque PAC y el enfoque SLT son enfoques basados en el peor caso, (*worst-case* o minimax), tanto con respecto a la distribución (*distribution free*), como con

[10] Esta propiedad posee como caso particular la convergencia uniforme ($\forall B \in \mathcal{S}_X$ σ-álgebra de X) de las frecuencias a las probabilidades uniformemente en distribución, lo que representa la generalización del teorema de Glivenko-Cantelli para cualquier distribución. (Basta considerar que $\mathcal{F}$ es una familia de funciones indicadoras).

[11] La costumbre de denominar *aprendible* a $\mathcal{F}$ proviene del modelo inicial de PAC en el que se suponía que la función verdadera estaba en $\mathcal{F}$.

[12] Si dichas condiciones se verifican para una $\mathbb{P}_Z$ determinada, se tienen las denominaciones clásicas de estimador (sucesión de estimadores) consistente (en probabilidad) o fuertemente consistente (consistente c.s.), respectivamente.

[13] Este resultado se verifica incluso para algoritmos que çasi minimizan el riesgo empírico", ([Vidyasagar 1997, th. 3.2], [Anthony y Bartlett 1999]).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

respecto a las hipótesis de la familia. Sin embargo, esta última condición es, a pesar de todo, necesaria para la consistencia no trivial en la SLT.

En las definiciones anteriores, no se ha supuesto para nada una posible función verdadera f (*target*) o una familia objetivo $\mathcal{T}$, (aunque en su origen, la formulación PAC era más simple y suponía que la función verdadera era realizable por $\mathcal{F}$). Evidentemente, si se verifica $\mathsf{f} \in \mathcal{F}$, la consistencia anterior supone la convergencia a f del proceso de aprendizaje (suponiendo que el $\inf_{\mathcal{F}} R(f)$ sea único).

I.3.2 Capacidad de la Familia de Hipótesis

El siguiente paso en el estudio de la consistencia es determinar en qué condiciones una familia $\mathcal{F}$ de hipótesis, satisfará la propiedad CUMEUD.

La respuesta a dicho problema vino dada por Vapnik que, tanto para familias de funciones indicadoras [Vapnik y Chervonenkis 1971], como de funciones reales [Vapnik y Chervonenkis 1981] aportó condiciones necesarias y suficientes para una $\mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$ prefijada y, más tarde, condiciones necesarias y suficientes $\forall \mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$ (*distribution free*) para funciones indicadoras [Vapnik 1982] y condiciones suficientes $\forall \mathbb{P}_{\mathsf{Z}}$ para funciones reales [Vapnik 1998]. Finalmente, [Alon *et al.* 1997] culminaron la solución del problema con condiciones necesarias y suficientes $\forall \mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$ para funciones reales.

Dichos resultados se basan en una medida de complejidad de $\mathcal{F}$: la dimensión de Vapnik-Chervonenkis (VCdim) para familias de conceptos (funciones indicadoras), y una generalización de la anterior, la dimensión *fat-shattering* ($V_\gamma \dim$), para funciones reales.

Sea $\mathcal{F}$ una familia de funciones indicadoras y $\mathbf{z}^n = (\mathbf{z}_1, ..., \mathbf{z}_n)$ una muestra aleatoria. Denotando por $\#$ el cardinal de un conjunto, realizamos las siguientes definiciones:

Definición 2 *Se definen el* ***coeficiente de disección empírico*** *(empirical shatter coefficient) y el* ***coeficiente de disección*** *(shatter coefficient), respectivamente, como:*

$$\begin{aligned} N_{\mathcal{F}}(\mathbf{z}^n) &= \#\{(f(\mathbf{z}_1), ..., f(\mathbf{z}_n)) : f \in \mathcal{F}\} \leq 2^n \\ N_{\mathcal{F}}(n) &= \max_{\mathbf{z}^n \in \mathcal{Z}^n} N_{\mathcal{F}}(\mathbf{z}^n) \end{aligned}$$

Como puede observarse, el primero, indica el número de posibles clasificaciones dicotómicas de los datos de $\mathbf{z}^n$ que pueden realizarse con funciones de $\mathcal{F}$ y, el segundo, es el máximo de dicha cantidad en todas las muestras posibles de tamaño n.

Definición 3 *[Vapnik 1998]* ***Dimensión VC****. Dada una familia $\mathcal{F}$ de funciones indicadoras, su dimensión VC se define como:*

$$\mathrm{VCdim}(\mathcal{F}) = \max\{n \in \mathbb{N}^+ : N_{\mathcal{F}}(n) = 2^n\}$$

y $\mathrm{VCdim}(\mathcal{F}) = \infty$ *si dicho máximo no existe.*

Por tanto, la VCdim es el número máximo de cualesquiera puntos que pueden diseccionarse mediante funciones de $\mathcal{F}$. Puede demostrarse, (p. ej. [Vapnik 1998], [Vidyasagar 1997]) que,

denotando $v = \mathrm{VCdim}(\mathcal{F})$, se verifica que[14]:

$$G_{\mathcal{F}}(n) \doteq \ln N_{\mathcal{F}}(n) \begin{cases} = n \ln 2 & \text{si } n \leq v \\ \leq v(\ln \frac{n}{v} + 1) & \text{si } n > v \end{cases} \tag{I.8}$$

donde $G_{\mathcal{F}}(n)$ es la denominada *growth function*. Con ello, esta función crece de forma lineal con n hasta $n = v$ (posiblemente infinito) y después, si es el caso, crece de forma logarítmica. Como veremos a continuación, este régimen es el que garantiza la consistencia.

Para funciones reales, existen distintas generalizaciones de la VCdim: la VCdim para funciones reales [Vapnik 1998] y la *pseudodimension* o dimensión P ([Pollard 1990], [Haussler 1992], [Vidyasagar 1997]), que coinciden con la VCdim para el caso de funciones indicadoras. Esta última, se define de la siguiente forma, donde denotamos $\{1 : n\} \doteq \{1, ..., n\}$ y $\{a_i\}_{i=1}^n = \{a_i : i \in \{1 : n\}\}$:

Definición 4 *[Anthony y Bartlett 1999, p. 152]* ***Dimensión P****. Sea $\mathcal{F}$ una clase de funciones definidas en $\mathcal{Z}$ y con valores en $\mathbb{R}$. Se dice que $\mathcal{F}$ pseudo-disecciona un conjunto $S = \{\mathbf{z}_i\}_{i=1}^n \subset \mathcal{Z}$ si $\exists \{t_i\}_{i=1}^n \subset \mathbb{R}$ tal que $\forall \mathbf{b} \in \{0,1\}^n$, $\exists f_{\mathbf{b}} \in \mathcal{F}$ tal que:*

$$\left\{ \begin{array}{l} f_{\mathbf{b}}(\mathbf{z}_i) \geq t_i \text{ si } b_i = 1 \\ f_{\mathbf{b}}(\mathbf{z}_i) < t_i \text{ si } b_i = 0 \end{array} \right|, \forall i \in \{1 : n\}$$

Se define la dimensión P de $\mathcal{F}$ como:

$$\mathrm{Pdim}(\mathcal{F}) = \max\{\#S : S \subset \mathcal{Z} \text{ pseudo-diseccionado por } \mathcal{F}\}$$

y $\mathrm{Pdim}(\mathcal{F}) = \infty$ si dicho máximo no existe.

La ***dimensión VC*** *para funciones reales, se define [Vapnik 1998, p. 191] como un caso particular del anterior tomando $t_i = t \in \mathbb{R}$, $\forall i = 1 : n$.*

La dimensión P puede definirse también [Anthony y Bartlett 1999, p. 153] a partir de la región inferior del grafo de las funciones de $\mathcal{F}$. Para ello, nótese que la definición anterior equivale a $\psi(f_{\mathbf{b}}(\mathbf{z}_i) - t_i) = b_i \ \forall i \in \{1 : n\}$ donde ψ es la función de Heaviside $\psi(r) = 1$ si $r \geq 0$ y $\psi(r) = 0$ si $r < 0$. Entonces, si $h_f(\mathbf{z}, t) = \psi(f(\mathbf{z}) - t)$ es la función indicadora de dicha región para la función f, la dimensión P viene a ser la dimensión VC de la familia $\{h_f : f \in \mathcal{F}\}$.

Sin embargo, aquí son de interés sus respectivas versiones escaladas: la dimensión *fat-shattering* ($V_\gamma \dim$) de [Alon *et al.* 1997], que generaliza la dimensión VC, y la dimensión P_γ (P_γ-dim) de [Kearns y Schapire 1994], que generaliza la dimensión P, incluyendo un factor de escala. Ambas son equivalentes en lo que respecta al problema de la consistencia [Alon *et al.* 1997], pero aquí presentamos la segunda, más utilizada, que ha adquirido la notación $V_\gamma \dim$ de la primera.

Definición 5 *[Anthony y Bartlett 1999, p. 159]* ***Dimensión*** *V_γ. Sea $\mathcal{F}$ una clase de funciones definidas en $\mathcal{Z}$ y con valores en $\mathbb{R}$, y sea $\gamma \in \mathbb{R}^+$. Se dice que $\mathcal{F}$ γ-**disecciona** un conjunto $S = \{\mathbf{z}_i\}_{i=1}^n \subset \mathcal{Z}$ si $\exists \{t_i\}_{i=1}^n \subset \mathbb{R}$ $\forall \mathbf{b} \in \{0,1\}^n$, $\exists f_{\mathbf{b}} \in \mathcal{F}$ tal que*

$$\left\{ \begin{array}{l} f_{\mathbf{b}}(\mathbf{z}_i) \geq t_i + \gamma \text{ si } b_i = 1 \\ f_{\mathbf{b}}(\mathbf{z}_i) \leq t_i + \gamma \text{ si } b_i = 0 \end{array} \right|, \forall i \in \{1 : n\}$$

[14] En términos del coefficiente de disección, dicha relación se escribe como: $N_{\mathcal{F}}(n) = 2^n$ si $n \leq v$ y $N_{\mathcal{F}}(n) \leq (en/v)^v$ si $n > v$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Se define la dimensión V_γ de $\mathcal{F}$ como:

$$V_\gamma\dim(\mathcal{F}) = \text{máx}\{\#S : S \subset \mathcal{Z} \textit{ diseccionado por } \mathcal{F}\}$$

y $V_\gamma \dim(\mathcal{F}) = \infty$ si dicho máximo no existe. (La dimensión V_γ original se define con $t_i = t$, $\forall t \in \mathbb{R}$, generalizando análogamente la VCdim*).*

La utilización de un valor $\gamma > 0$ en la definición anterior exige una mayor complejidad a la familia $\mathcal{F}$ para diseccionar un conjunto S que un valor $\gamma = 0$. Con ello,

$$\gamma > 0 \Rightarrow V_\gamma \dim(\mathcal{F}) \leq V_0 \dim(\mathcal{F}) = \mathrm{Pdim}(\mathcal{F})$$

Las relaciones entre la dimensión P y la dimensión V_γ pueden verse en [Anthony y Bartlett 1999, cap. 11].

Pues bien, las condiciones necesarias y suficientes para la propiedad CUMEUD basados en estas medidas de complejidad, se resumen en lo siguiente: una familia $\mathcal{F}$ de funciones indicadoras (reales) verifica la condición CUMEUD si y sólo sí $\mathrm{VCdim}(\mathcal{F}) < \infty$ [Vapnik 1998] ($V_\gamma \dim(\mathcal{F}) < \infty$ [Alon *et al.* 1997]). (En el caso de funciones reales, la finitud de la dimensión P es sólo condición suficiente).

Estos resultados ponen en relación la complejidad de la familia y la consistencia del problema de aprendizaje. Una familia excesivamente compleja puede provocar una inestabilidad suficiente en el proceso como para impedir la convergencia[15].

Sin embargo, como veremos a continuación, los resultados mencionados se expresan en términos de la familia de funciones $\mathcal{L} = \{\ell_f(\mathbf{z}) = \ell(y, f(\mathbf{x})) : f \in \mathcal{F}\}$, en lugar de en términos de $\mathcal{F}$. La razón es que las funciones $\ell_f(\mathbf{z})$ son las que participan directamente en la definición del proceso empírico ([Pollard 1984], [Vapnik 1982], [Pollard 1990]) $\{\xi_n : n \in \mathbb{N}^+\}$ siendo:

$$\xi_n = \left|\hat{R}_n(f) - R(f)\right| = \left|\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)) - \int \ell(y, f(\mathbf{x}))d\mathsf{P}_{\mathsf{Z}}(\mathbf{x}, y)\right|$$

Ello no contradice las conclusiones anteriores pues, si la función de pérdida ℓ es γ-admisible, se verifica (en general para hipótesis con valores reales [Vidyasagar 1997, ths. 5.11 y 7.3], y p. ej. para $\ell(y, u) = |y - u|$ con $\mathcal{Y} = \{0, 1\}$, [Vidyasagar 1997, ths. 5.12 y 7.4], que:

$$\mathrm{VCdim}(\mathcal{F}) < \infty \Rightarrow \mathrm{VCdim}(\mathcal{L}) < \infty$$

y, análogamente, en términos de la $V_\gamma \dim$ (p. ej. [Anthony y Bartlett 1999, th. 12.11 y lema 17.6]). De esta forma, la familia $\mathcal{L}$ hereda el grado de complejidad de la familia $\mathcal{F}$.

I.3.3 Velocidad de Convergencia

Los resultados mencionados anteriormente (mejorados posteriormente con cotas más ajustadas), poseen la forma general [Evgeniou 2000] que presentamos a continuación para el caso de funciones reales.

[15] Resulta ilustrativo mencionar la relación que [Vapnik 1998] destaca entre el significado de estos resultados y el principio de Falsabilidad de [Popper 1962]. Una familia excesivamente compleja que contenga hipótesis capaces de reproducir perfectamente cualquier conjunto de datos, no contendría conocimiento alguno. En términos de K. Popper, la teoría científica implícita en $\mathcal{F}$ no sería falsable, con lo que no sería válida.

Teorema 6 *[Alon* et al. *1997]. Sea $\mathcal{F}$ familia de funciones tal que $a \leq R(f) \leq b, \forall f \in \mathcal{F}$ con $a, b \in \mathbb{R}$. $\forall \varepsilon \geq 0$, $\forall n \geq 2/\varepsilon^2$ y $v = V_\gamma \dim(\mathcal{L}) < \infty$ con $\mathcal{L} = \{\ell_f(\mathbf{z}) = \ell(y, f(\mathbf{x})) : f \in \mathcal{F}\}$, entonces, para $\gamma = \alpha\varepsilon$ con $\alpha \geq 1/48$, y $\forall \mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$, se verifica:*

$$\mathbb{P}_{\mathsf{Z}^n}\{\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| > \varepsilon\} \leq \eta(n, \varepsilon, v) \tag{I.9}$$

donde η es una función creciente en v y decreciente en ε y n, con $\lim_{n\to\infty} \eta = 0$.

La función η da idea de la **velocidad de convergencia** uniforme del riesgo empírico al riesgo teórico. El resultado anterior implica en particular que, para cualesquiera n y $\eta \in (0, 1)$ prefijados, existe un $\varepsilon(n, \eta, v) \geq 0$ tal que, con probabilidad al menos $1 - \eta$ sobre $\mathbf{z}^n \in \mathcal{Z}^n$ para cualquier medida de probabilidad $\mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$, se verifica:

$$R(f) \leq \hat{R}_n(f) + \varepsilon(n, \eta, v) \tag{I.10}$$

donde $\varepsilon(n, \eta, v)$ es creciente en v y $1 - \eta$ y decreciente en n, con $\varepsilon(n, \eta, v) \to 0$ cuando $n \to \infty$. De esta forma, $\varepsilon(n, \eta, v)$ actúa como **cota al error de estimación** de $R(f)$ mediante $\hat{R}_n(f)$, bajo cualquier medida de probabilidad $\mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$.

Igualmente, dados $\varepsilon \geq 0, \eta \in (0, 1)$, existe un $n(\varepsilon, \eta, v)$ tal que la acotación anterior se verifica para ε con probabilidad al menos $1 - \eta$ para cualquier $\mathbb{P}_{\mathsf{Z}} \in \mathcal{P}$. En el contexto PAC, el valor $n(\varepsilon, \eta, v)$ recibe el nombre de **complejidad muestral**[16].

I.3.4 Complejidad Computacional

Mientras que puede no haber barreras estadísticas a la posibilidad del aprendizaje, es posible que las haya computacionales, en el sentido de que el algoritmo pueda no ser implementable en la práctica.

El objetivo de la teoría estadística del aprendizaje es preguntarse por las condiciones para el aprendizaje (o para su imposibilidad) en su dimensión de información o estadística, pero queda por estudiar las condiciones en que el algoritmo es, o no, realmente computable.

Esta computabilidad se estudia en términos del tiempo que requiere el algoritmo para proporcionar una solución. Este tiempo se define de forma normalizada, para que sea independiente de la velocidad de la máquina, en términos del número de operaciones básicas necesarias entre dos bits, (de tal forma que cada operación se realiza en una unidad de tiempo), donde se supone que toda cantidad, sea entera o real, se representa por el mismo número de bits.

La rama de la informática conocida como **complejidad computacional** estudia la viabilidad computacional del aprendizaje en términos del tiempo requerido por el algoritmo, (el número de operaciones requeridas) en función de la dimensión de los datos de entrada.

Para ello, en la notación de todos los elementos básicos del problema de aprendizaje, se introduce la dimensión d del espacio de entrada. Por ejemplo, en un problema de clasificación: X_d es el vector de entrada, $\mathbf{z}_d^n$ es una muestra de datos extraídos de una población $\mathsf{Z}_d^n \in \mathcal{Z}_d^n$, $\mathcal{F}_d$ es el espacio de hipótesis, etc.

Ahora, en lugar de considerar un algoritmo concreto $A : \cup_{n=1}^{\infty} \mathcal{Z}_d^n \to \mathcal{F}_d$, conviene estudiar una clase de algoritmos (denominada también algoritmo) representada por $\mathbb{A} : \cup_{d=1}^{\infty} \cup_{n=1}^{\infty} \mathcal{Z}_d^n \to \cup_{d=1}^{\infty} \mathcal{F}_d$, que es realmente un método de estimación que no depende de la dimensión d del espacio de entrada. A la familia $\mathcal{F} = \cup_{d=1}^{\infty} \mathcal{F}_d$ se le denomina **familia escalonada** (*graded class*).

[16] Nótese que las funciones $\varepsilon(n, \eta, v)$ y $n(\varepsilon, \eta, v)$ no dependen para nada de la distribución P_{Z}.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Se dice que un algoritmo $\mathbb{A}$ así definido, posee un **tiempo polinómico** de ejecución si existe una constante α tal que para todos los datos de dimensión d, el algoritmo requiere como máximo, un número $O(d^\alpha)$ de operaciones para producir su salida.

Sin embargo, la dimensión d del espacio de entrada no es la única variable a considerar, pues el tiempo de ejecución del algoritmo está determinado también por la precisión ε y la significación η que se pretende conseguir, (véase la discusión del Apartado I.3.3). Con ello, la cualidad de algoritmo deseable para un problema de aprendizaje desde el punto de vista computacional, es la siguiente:

Definición 7 *Dada una familia escalonada $\mathcal{F} = \cup_{d=1}^{\infty} \mathcal{F}_d$, un algoritmo $\mathbb{A}$ es **eficiente** si:*

1. *El tiempo de ejecución de $\mathbb{A}$ en muestras $\mathbf{z}_d^n$ es polinómico en n y en d.*
2. *La complejidad muestral $n(d, \varepsilon, \eta)$ de $\mathbb{A}$ en $\mathcal{F}_d$ es polinómica en d, $1/\varepsilon$ y $\ln(1/\eta)$.*

El comportamiento polinómico de la complejidad muestral se deriva de los resultados de la teoría estadística del aprendizaje o del aprendizaje PAC, mediante el siguiente teorema para funciones indicadoras que posee uno análogo para funciones reales [Anthony y Bartlett 1999]:

Teorema 8 *Se verifican los siguientes resultados:*

1. *Si $\mathrm{VCdim}(\mathcal{F}_d)$ es polinómico en d, entonces un algoritmo que realice el principio ERM en $\mathcal{F}$ produce una complejidad muestral $n(d, \varepsilon, \eta)$ polinómica en d, $1/\varepsilon$ y $\ln(1/\eta)$.*
2. *Si existe un algoritmo eficiente para $\mathcal{F}$ entonces $\mathrm{VCdim}(\mathcal{F}_d)$ es polinómica en d.*

El estudio de la complejidad computacional del aprendizaje gira alrededor del concepto de algoritmo eficiente, con el objetivo de determinar las condiciones para su existencia o imposibilidad. En lo que sigue no volveremos a tratar esta dimensión del aprendizaje. Una visión general puede obtenerse por ejemplo, en [Anthony y Bartlett 1999], [Vidyasagar 1997] o [Kearns y Vazirani 1997].

I.4 Control de la Complejidad

El teorema 6 admite un análisis asintótico: con un tamaño muestral suficientemente grande y siempre que no consideremos una familia con complejidad ilimitada ($v < \infty$), el valor $\hat{R}_n(f)$ es una buena representación del riesgo $R(f)$ en que estamos incurriendo. Ello nos llevaría a elegir como familia $\mathcal{F}$ más adecuada la que produjese en el límite (n suficientemente grande) un valor mínimo para $\hat{R}_n(\hat{f}_n)$ siendo[17] $\hat{f}_n = \arg\inf_{\mathcal{F}} \hat{R}_n(f)$.

Pero, en la práctica, lo usual es que tengamos que enfrentarnos a una muestra de tamaño limitado, en cuyo caso, se plantea el problema de la elección de la familia $\mathcal{F}$ más adecuada.

Si $\mathcal{F}$ posee poca complejidad, el término $\varepsilon(n, \eta, v)$ disminuye, lo que hace pensar que la mejor familia es la de menor complejidad. Sin embargo, esta limitación provocaría un incremento del riesgo empírico $\hat{R}_n(f)$ y la esperanza de reducir dicho riesgo, ampliando si es posible el

[17] Nótese que a esta conclusión no podría llegarse esgrimiendo $\mathrm{plim}_{n\to\infty} \hat{R}_n(\hat{f}_n) = R(\hat{f}_n)$ por las leyes de los grandes números, pues, al ser $\hat{f}_n = A(\mathbf{z}^n)$, las variables aleatorias $l(y_i, \hat{f}_n(\mathbf{x}_i))$ no son independientes y tales teoremas no son aplicables.

tamaño muestral, estaría limitada por la capacidad de representación de la familia. Las limitaciones impuestas sobre $\mathcal{F}$ se traducirían en una incapacidad estructural ineludible para reproducir el proceso generador de los datos.

Por tanto, la mejor solución es considerar a priori una familia $\mathcal{F}$ lo suficientemente compleja y, fijado n, controlar su complejidad eligiendo alguna subfamilia $\mathcal{F}_{v(n)} \subset \mathcal{F}$ que minimice el término de la derecha en (I.10), (véase la discusión de [Geman *et al.* 1992]). La determinación de la regla $v = v(n)$ más adecuada para elegir la familia en función del tamaño muestral es el objeto del problema de **selección del modelo**.

Pero la selección del modelo, sólo puede plantearse si se dispone de algún parámetro v indicador de la complejidad de la familia. La construcción de tal indicador constituye el objeto principal del problema de **control de la complejidad**. En el enfoque SLT y PAC, dicho parámetro de complejidad es la VCdim, $V_\gamma \dim$ o similar, pero existen otras posibilidades, algunas de las cuáles, se exponen más adelante.

En la literatura estadística, las técnicas clásicas (con origen en modelos lineales) de control de la complejidad suelen encuadrarse en dos grandes grupos (p. ej. [Breiman 1997*a*]):

1. Técnicas basadas en el control de las variables que se consideran en el problema, (selección *forward*, *backward, stepwise*), (p. ej. [Hocking 1976], [Seber 1977], [Miller 1990]).

2. Técnicas basadas en parámetros, (métodos de regularización), en las que se controla el tamaño de los parámetros. Un ejemplo es la regresión *ridge* ([Hoerl y Kennard 1970*b*], [Hoerl y Kennard 1970*a*], [Seber 1977]) destinada a disminuir la varianza de los estimadores de los parámetros de una regresión lineal (o a evitar el mal condicionamiento del sistema), mediante una pequeña inyección de sesgo resultado de la penalización del tamaño de los parámetros.

Sin embargo, ambos métodos están relacionados pues, la eliminación de variables, puede verse como la máxima penalización de los parámetros correspondientes.

I.4.1 Regularización

El enfoque de regularización (Teoría del Problema Inverso, –p. ej. [Bertero 1986]), tiene su origen realmente en la necesidad de convertir un problema mal planteado (sin solución, o con solución no única o no estable), en uno bien planteado, (con solución única y estable), para lo cuál se restringe el dominio del problema imponiendo condiciones adicionales.

Ya hemos mencionado anteriormente (Apartado I.1.2) que, por el lema de continuidad del operador inverso (p. ej. [Tikhonov y Arsenin 1977]), una forma de lograr la estabilidad es restringir el espacio de hipótesis a algún subespacio compacto del espacio métrico $\mathcal{F}$.

Este es el camino seguido por el método de los *Sièves* [Grenander 1981], que considera una sucesión creciente $\{\Theta_v\}$ de subconjuntos compactos en el espacio Θ de los parámetros, tal que, $\cup_v \Theta_v$ es denso en Θ. Pero, una vía alternativa es la del método de regularización que pretende aproximar el operador inverso $\mathcal{O}^{-1}$ mediante una sucesión de operadores estables, mediante el concepto de operador regularizador $\mathcal{Q}$, tal como describimos a continuación.

Definición 9 *[Tikhonov y Arsenin 1977]. Sea f_0 solución, que por simplicidad suponemos única, de la ecuación $\mathcal{O}(f) = \mathbf{y}_0$, un operador $\mathcal{Q}(\mathbf{y}, \lambda)$ se dice **regularizador** (u operador de regularización) para la ecuación $\mathcal{O}(f) = \mathbf{y}$ en un entorno de $\mathbf{y} = \mathbf{y}_0$ si:*

1. Existe $\delta_1 > 0$ tal que $\mathcal{Q}$ está definido $\forall \lambda > 0$ y $\forall \mathbf{y} \in \mathbb{Y}$ tal que $\rho_{\mathbb{Y}}(\mathbf{y}, \mathbf{y}_0) \leq \delta_1$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2. *Existe una función* $\lambda = \lambda(\delta, \mathbf{y}_0)$ *tal que,* $\forall \varepsilon > 0$, $\exists \delta(\varepsilon) \leq \delta_1$ *verificando:*

$$\mathbf{y}_\delta \in \mathbb{Y}, \rho_{\mathbb{Y}}(\mathbf{y}_0, \mathbf{y}_\delta) \leq \delta(\varepsilon) \Rightarrow f_\lambda = \mathcal{Q}(\mathbf{y}_\delta, \lambda(\delta, \mathbf{y}_\delta)) \text{ satisface } \rho_{\mathcal{F}}(f_0, f_\lambda) \leq \varepsilon$$

Definición 10 *Un regularizador* $\mathcal{Q}(\mathbf{y}, \lambda)$ *definido* $\forall \mathbf{y} \in \mathbb{Y}$ *y* $\forall \lambda > 0$*, continuo respecto de* $\mathbf{y}$*, que verifica que* $\lim_{\lambda \to 0} \mathcal{Q}(\mathcal{O}(f), \lambda) = \mathbf{y}$*, se denomina* ***regularizador para la ecuación*** $\mathcal{O}(f) = \mathbf{y}$.

Definición 11 *Se denomina* ***estabilizador*** *a toda aplicación* $\mathcal{E} : \mathcal{F} \to \mathbb{R}$ *que satisface que,* $\forall s > 0$, $\{f : \mathcal{E}(f) \leq s\}$ *es un subconjunto compacto de* $\mathcal{F}$.

Sea $\mathbf{y}_\delta$ tal que $\rho_{\mathbb{Y}}(\mathbf{y}_\delta, \mathbf{y}_0) \leq \delta$. Se verifica que [Tikhonov y Arsenin 1977], si un operador $\mathcal{Q}$ satisface $\mathcal{Q}(\mathbf{y}_\delta, \delta) = f_\delta$ siendo f_δ la solución del problema de optimización con restricciones:

$$\begin{cases} \min_f \mathcal{E}(f) \\ \rho_{\mathbb{Y}}(\mathbf{y}_\delta, \mathcal{O}(f)) \leq \delta \end{cases} \tag{I.11}$$

entonces, es un regularizador. Con ello, por definición, f_δ es una aproximación de f_0. Lo anterior también es válido si consideramos restricciones de igualdad $\rho_{\mathbb{Y}}(\mathbf{y}_\delta, \mathcal{O}(f)) = \delta$, con lo que, el problema anterior, puede plantearse mediante el método de Lagrange, como[18]:

$$\min_f \{\mathcal{R}(f, \mathbf{y}) = \rho_{\mathbb{Y}}(\mathbf{y}_\delta, \mathcal{O}(f)) + \lambda \mathcal{E}(f)\} \tag{I.12}$$

siendo λ es el multiplicador de Lagrange.

[Tikhonov y Arsenin 1977, p. 65-67] denominan **suavizador** al funcional $\mathcal{R}$ y demuestran que para una elección apropiada de λ, el operador $\mathcal{Q}_1$, tal que $f_\lambda = \mathcal{Q}_1(\mathbf{y}_\delta, \lambda(\delta, \mathbf{y}_\delta))$ es la solución del problema anterior, es un operador de regularización.

Puede demostrarse que [Vasin 1970], si las distintas funciones objetivo son convexas, los dos problemas anteriores son análogos al siguiente:

$$\begin{cases} \min_f \rho_{\mathbb{Y}}(\mathbf{y}_\delta, \mathcal{O}(f)) \\ \mathcal{E}(f) \leq v \end{cases} \tag{I.13}$$

adoptando como parámetros de control los parámetros δ, λ y v, respectivamente. En cada caso, es equivalente especificar a priori: el subconjunto $\mathcal{F}_v = \{f : \mathcal{E}(f) \leq v\}$, el valor máximo de $\rho_{\mathbb{Y}}(\mathbf{y}_\delta, \mathcal{O}(f))$ o el balance λ entre $\rho_{\mathbb{Y}}(\mathbf{y}_\delta, \mathcal{O}(f))$ y $\mathcal{E}(f)$.

Normalmente, el problema más sencillo de abordar numéricamente es el problema (I.12), un ejemplo del cuál es la regresión *ridge* ya citada.

Por su parte, la formulación (I.13) es la base del método de los *Sièves*[19]. La dificultad de la aplicación de este método reside en la determinación de una regla $v = v(n)$ que aporte el valor más adecuado de v ante una muestra de tamaño n determinada, y que depende, además, de la distribución desconocida de los datos. (Análogamente, una regla $\delta = \delta(n)$ en (I.11) o una regla $\lambda = \lambda(n)$ en (I.12)).

La determinación de tal regla es el objetivo primordial del **problema de selección del modelo** que veremos en el siguiente apartado y que, realmente, debe incluirse también en el algoritmo de aprendizaje, pues es la última fase de la producción de la hipótesis a partir de los datos.

[18] Su correlato en un problema de aprendizaje sería el principio ERM regularizado (I.1): $\min_f \{\hat{\mathcal{R}}_n(f) = \hat{R}_n(f) + \lambda \mathcal{E}(f)\}$.

[19] Este método se formula normalmente en el marco de un espacio de parámetros Θ que parametriza la familia de hipótesis $\mathcal{F}$.

Finalmente, mencionar, en esta introducción a las técnicas de control de la complejidad, los métodos bayesianos (p. ej. [Berger 1985], [O'Hagan 1994], [MacKay 1992*a*]), para los que una formulación del tipo (I.12), responde a la obtención del estimador *maximum a posteriori* (MAP) de la distribución a posteriori.

Si suponemos un modelo de distribución a priori p_f definido sobre la familia $\mathcal{F}$, y un conjunto de datos $\mathbf{y}^n = (y_1, ..., y_n)$ con y_i extraído de $\mathsf{Y}/\mathbf{x}_i$, $i = 1 : n$, la distribución a posteriori $\mathsf{p}_{f/\mathbf{y}^n}$ viene dada por:

$$\mathsf{p}(f|\mathbf{y}^n) = \frac{\mathsf{p}(\mathbf{y}^n|f)\mathsf{p}(f)}{\mathsf{p}(\mathbf{y}^n)} \propto \mathsf{p}(\mathbf{y}^n|f)\mathsf{p}(f)$$

donde hemos denotado la verosimilitud por $\mathsf{p}(\mathbf{y}^n|f) \equiv \mathsf{p}_{\mathbf{y}^n/f}(\mathbf{y}^n)$ y $\mathsf{p}(\mathbf{y}^n) \equiv \mathsf{p}_{\mathsf{Y}^n/\mathbf{x}^n}(\mathbf{y}^n)$ con $\mathbf{x}^n = (\mathbf{x}_1, ..., \mathbf{x}_n)$.

La maximización de dicha distribución a posteriori en términos de f equivale al problema:

$$\min_{f\in\mathcal{F}}\{-\ln \mathsf{p}(f|\mathbf{y}^n) = -[\ln \mathsf{p}(\mathbf{y}^n|f) + \ln \mathsf{p}(f)]\}$$

que es análogo a (I.12) para el caso en que:

$$\begin{aligned} \mathsf{p}(\mathbf{y}^n|f) &= \exp\{-\rho_{\mathbb{Y}}(\mathbf{y}^n, \mathcal{O}(f))\} \\ \mathsf{p}(f) &= \exp\{-\lambda\mathcal{E}(f)\} \end{aligned}$$

I.5 Selección del Modelo

Como ya hemos indicado, el objetivo del problema de control de la complejidad es la parametrización de la complejidad en el contexto de una familia marco de hipótesis. Disponiendo de estimadores del verdadero riesgo o cotas para la discrepancia entre éste y el riesgo empírico en función de dicho parámetro (al modo de (I.9) y (I.10)), puede realizarse la selección del modelo que posea mejores perspectivas en cuanto a su capacidad de generalización.

De la importancia de este problema es testigo el gran número de enfoques y técnicas, propuestos para resolverlo, de los que aquí, sólo podemos hacer una resumida reseña. Prácticamente, todos ellos proceden del ámbito de la estadística y, de esta forma, pueden reinterpretarse bajo el marco de la teoría de la aprendizaje.

En una primera aproximación a los métodos existentes, el factor diferenciador principal reside en el tamaño muestral requerido para su utilización. Así, es preciso distinguir los métodos con buenas propiedades cuando el tamaño muestral es suficientemente grande y cuyo uso ante muestras pequeñas sería más bien aventurado, de los métodos cuyas propiedades no dependen, al menos en teoría, del tamaño muestral. En este sentido, una propuesta de clasificación podría ser la siguiente:

1. Criterios clásicos basados en la estimación asintótica del riesgo, tradicionales en modelos lineales, pero generalizables a modelos no lineales (como las redes neuronales), utilizando normalmente aproximaciones asintóticas. Este grupo a su vez podría dividirse en los siguientes:
   (a) Criterios basados en la estimación del error cuadrático de predicción, por ejemplo: FPE de [Akaike 1969], SMS de [Shibata 1976], y técnicas del tipo validación cruzada como LOO, (*Leave-one-out*), (p. ej. [Shao 1993], [Shao 1997]) y GCV (*General cross-validation*), (p. ej. [Wahba 1990*b*]).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Normalmente, estos métodos están orientados a pérdida cuadrática y modelos lineales, para los que poseen versiones analíticas, aunque existen versiones analíticas para pérdidas más generales y modelos no lineales, mediante aproximaciones de primer o segundo orden de la función objetivo, (p. ej. [Hansen y Larsen 1996]).

(b) Criterios basados en medidas de discrepancia, (modelos no realizables o mal identificados), que generalizan los anteriores. Este grupo puede a su vez dividirse entre los que consideran o no, la posibilidad de funciones objetivo regularizadas:

I. Sin regularización: AIC de Akaike, ([Akaike 1973], [Akaike 1974]), BIC de [Schwarz 1978], AIC_c de [Sugiura 1978] y [Hurvich y Tsai 1989] y, en general [Linhart y Volkers 1984].

II. Con regularización: GPE de [Moody 1992], NIC de [Murata *et al.* 1991], y otros (p. ej. [Liu 1995]), admitiendo todos posible mala especificación.

(c) Métodos *Plug-in.* Estos métodos utilizan aproximaciones asintóticas de los estimadores para determinar los parámetros que minimizan determinadas medidas de la capacidad de generalización. Por ello, han de desarrollarse específicamente para cada tipo de algoritmo (p. ej. regresión tipo núcleo y regresión polinómico local).

(d) Métodos basados en técnicas de Remuestreo.

I. Validación cruzada: LKO (*Leave-k-out*) (p. ej. [Shao 1993]) y sus variantes.

II. Técnicas bootstrap: *Naïve* bootstrap y su versión corregida [Efron y Tibshirani 1993], y LOO bootstrap, Bootstrap .632 y .632+ [Efron y Tibshirani 1997].
En este grupo se incluyen asimismo correcciones a los métodos clásicos para el caso de muestras pequeñas, si bien sus propiedades no asintóticas siguen estando en cuestión: EIC de [Ishiguro *et al.* 1997], Bootstrap CV y .632CV de [Pan 1999], etc.

2. Técnicas no asintóticas, válidas para muestras pequeñas.

(a) Métodos bayesianos, que serán tratados con detalle en el capítulo siguiente, en su aplicación a redes neuronales.

(b) Métodos basados en la teoría de la información: MDL (*minimum description length*) ([Rissanen 1978], [Rissanen 1983]), y su generalización bajo la denominación de control de la complejidad estocástica, ([Rissanen 1987], [Barron 1991], [Barron y Cover 1991]).

(c) Criterios basados en cotas *worst-case*: Minimización del riesgo estructural, (SRM) de [Vapnik 1982] y [Vapnik 1998], y técnicas similares en el ámbito del aprendizaje PAC o SLT.

(d) Criterios basados en cotas que incluyen penalizaciones dependientes de los datos en lugar de en medidas de complejidad independientes de la distribución, como en [Bartlett *et al.* 2002], introducidas para evitar la excesiva holgura de las cotas *worst-case* anteriores.

(e) Criterios basados en cotas construidas a partir del concepto de estabilidad del algoritmo, como en [Bousquet y Elisseeff 2002], donde dicha estabilidad se define en función del comportamiento del algoritmo específico ante cambios en la muestra tales, como la eliminación o sustitución de una de sus observaciones, siguiendo la filosofía de LOO.

Al margen de estas técnicas, es preciso destacar un conjunto de técnicas recientes que abandonan la filosofía de la selección entre modelos alternativos, sustituyéndola por una más general de combinación de modelos. Entre éstas es preciso destacar: *Bagging*, (p. ej. [Breiman 1996*a*]),

USC DE COMPOSTELA

Boosting, (p. ej. [Freund y Schapire 1997]) y, en general, *Stacking* ([Wolpert 1992], [Breiman 1993]).

Finalmente, es necesario mencionar las técnicas basadas en tests de bondad de ajuste para la comparación entre dos modelos específicos. Estas técnicas, a pesar de sus buenos resultados reseñados en la literatura, (p. ej. [Refenes y Holt 2001], [Anders y Korn 1999], [Teräsvirta *et al.* 1993]) no han sido utilizadas más que someramente en el campo de las redes neuronales. Ello se debe, por una parte, a la frecuente existencia de óptimos locales (y/o diversos óptimos globales) que cuestionan las aproximaciones asintóticas necesarias para determinar estadísticos con distribución conocida (p. ej. [White 1989], [Fine y Mukherjee 1999]). Pero, también, a la práctica común de entrenar, un gran número de arquitecturas entre las que es preciso elegir, para lo que se prefieren métodos de selección del modelo masivos y lo más automáticos posibles.

Teniendo en cuenta estas razones, y dadas las limitaciones de espacio, estos métodos serán objeto de futuros trabajos.

En lo que sigue, presentamos someramente los distintos métodos que son de interés para el presente documento. Antes de ello, realizamos un análisis de las componentes del riesgo que contribuye a ilustrar el problema de selección del modelo y, al mismo tiempo, es útil para el análisis de los métodos clásicos.

Dicho análisis refleja, bajo un punto de vista diferente al de la ecuación (I.10), el necesario equilibro entre la complejidad de la familia $\mathcal{F}$ y el grado de fidelidad a los datos.

I.5.1 Descomposición del Riesgo Esperado en Sesgo y Varianza

Pérdida Cuadrática

Otra forma, ya clásica (p. ej. [Geman *et al.* 1992], [Breiman 1997*a*]), de ilustrar el equilibro necesario entre dotar de suficiente complejidad al modelo y evitar que asimile el ruido de los datos, es la descomposición del riesgo esperado en dos términos que reflejan ambos lados de la balanza. Con pérdida cuadrática $\ell(y, f(\mathbf{x})) = [y - f(\mathbf{x})]^2$, el riesgo suele denominarse **error cuadrático de predicción** (*prediction squared error* –PSE) que, como vimos anteriormente, (I.2), verifica:

$$\mathrm{PSE}(f) = R(f) = \mathbb{E}_{(\mathsf{X},\mathsf{Y})}\{[\mathsf{Y} - f(\mathsf{X})]^2\} = R(\mathsf{f}) + \mathbb{E}_{\mathsf{X}}\{[f(\mathsf{X}) - \mathsf{f}(\mathsf{X})]^2\} \qquad \text{(I.14)}$$

donde $R(\mathsf{f}) = \mathbb{E}_{\mathsf{X}}[\mathrm{Var}_{\mathsf{Y}/\mathsf{X}}(\mathsf{Y})] = \mathbb{E}_{\mathsf{X}}[\sigma^2(\mathsf{X})]$.

El criterio PSE proporciona una medida de calidad para una hipótesis determinada. Si, dada una muestra aleatoria simple Z^n, denotamos por $\hat{f}_n$ la hipótesis producida por el algoritmo a partir de Z^n, tomando esperanzas en la muestra, se tiene la **esperanza del error cuadrático de predicción** (*mean prediction squared error* –MPSE) o **riesgo esperado**, que proporciona una medida de calidad para el algoritmo:

$$\begin{aligned} \mathrm{MPSE}(\hat{f}_n) &= \mathbb{E}_{\mathsf{Z}^n}[\mathrm{PSE}(\hat{f}_n)] = \mathbb{E}_{\mathsf{X}}[\sigma^2(\mathsf{X})] + \mathbb{E}_{\mathsf{Z}^n}\mathbb{E}_{\mathsf{X}}\{[\hat{f}_n(\mathsf{X}) - \mathsf{f}(\mathsf{X})]^2\} \qquad \text{(I.15)} \\ &= \mathbb{E}_{\mathsf{X}}[\sigma^2(\mathsf{X})] + \mathbb{E}_{\mathsf{X}}\{\mathrm{MSE}[\hat{f}_n(\mathsf{X})]\} \\ &= \mathbb{E}_{\mathsf{X}}\{\sigma^2(\mathsf{X}) + \mathrm{MSE}[\hat{f}_n(\mathsf{X})]\} \end{aligned}$$

donde definimos la **esperanza del error cuadrático** *(mean squared error* –MSE) en un punto $\mathbf{x} \in \mathcal{X}$ y, análogamente, su versión integrada (*integrated mean squared error* –IMSE), respecti-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

vamente, como:

$$\begin{aligned} \mathrm{MSE}[\hat{f}_n(\mathbf{x})] &= \mathbb{E}_{\mathsf{Z}^n}[(\hat{f}_n(\mathbf{x}) - \mathsf{f}(\mathbf{x}))^2] \\ \mathrm{IMSE}(\hat{f}_n) &= \mathbb{E}_{\mathsf{X}}\{\mathrm{MSE}[\hat{f}_n(\mathsf{X})]\} \end{aligned}$$

Como el valor $\mathbb{E}_{\mathsf{X}}[\sigma^2(\mathsf{X})]$ no depende de la hipótesis seleccionada, el MSE suele ser objeto de estudio y puede descomponerse como:

$$\begin{aligned} \mathrm{MSE}[\hat{f}_n(\mathbf{x})] &= \mathbb{E}_{\mathsf{Z}^n}\{[\hat{f}_n(\mathbf{x}) - \mathsf{f}(\mathbf{x})]^2\} \\ &= \mathbb{E}_{\mathsf{Z}^n}\{[\hat{f}_n(\mathbf{x}) - \mathbb{E}_{\mathsf{Z}^n}(\hat{f}_n(\mathbf{x})) + \mathbb{E}_{\mathsf{Z}^n}(\hat{f}_n(\mathbf{x})) - \mathsf{f}(\mathbf{x})]^2\} \\ &= \mathrm{Var}_{\mathsf{Z}^n}[\hat{f}_n(\mathbf{x})] + \mathrm{Sesgo}^2_{\mathsf{Z}^n}[\hat{f}_n(\mathbf{x})] \end{aligned}$$

donde $\mathrm{Sesgo}_{\mathsf{Z}^n}[\hat{f}_n(\mathbf{x})] = \mathbb{E}_{\mathsf{Z}^n}[\hat{f}_n(\mathbf{x})] - \mathsf{f}(\mathbf{x})$ es el sesgo de la hipótesis $\hat{f}_n$. Con ello, el MPSE puede escribirse como:

$$\mathrm{MPSE}(\hat{f}_n) = \mathbb{E}_{\mathsf{X}}\{\sigma^2(\mathsf{X}) + \mathrm{Var}_{\mathsf{Z}^n}[\hat{f}_n(\mathsf{X})] + \mathrm{Sesgo}^2_{\mathsf{Z}^n}[\hat{f}_n(\mathsf{X})]\}$$

La expresión anterior muestra las tres fuentes de error en la esperanza del error cuadrático de predicción: el **ruido** de los datos, la variabilidad muestral reflejada en la **varianza** de la hipótesis $\hat{f}_n$ producida por el algoritmo y, por último, las limitaciones intrínsecas de la hipótesis para representar la verdadera función bajo la distribución P_{Z}, aspecto que recoge el **cuadrado del sesgo**.

De ello, se deduce que, un modelo excesivamente complejo será muy sensible al ruido de los datos, con lo que poseerá mayor varianza. Por otra parte, un modelo con mucho sesgo indica una complejidad insuficiente para reproducir el proceso generador de los datos. La descomposición anterior es otra forma de poner de manifiesto el necesario equilibrio entre estabilidad y complejidad con el fin de minimizar el riesgo esperado.

Descomposición Sesgo-Varianza para Pérdidas Generales

La descomposición anterior del riesgo esperado se obtiene suponiendo la pérdida cuadrática. Esta pérdida se prescribe sobre todo para problemas de regresión debido a sus buenas propiedades y, sobre todo, en el caso de ruido gausiano.

Para el problema de clasificación u otras situaciones en las que se utilice otra pérdida, en los últimos años se han propuesto diversas descomposiciones del riesgo ([Kong y Dietterich 1995], [Tibshirani 1996*a*], [Kohavi y Wolpert 1996], [Breiman 1996*b*], [James y Hastie 1997], [Heskes 1997], [Friedman 1997], [Breiman 1998] y [Domingos 2000]) definiendo componentes similares de ruido, sesgo y varianza, y manteniendo en lo posible las propiedades de la descomposición para el caso de pérdida cuadrática.

Salvo las propuestas de [Heskes 1997], que se basa en la discrepancia de Kulback-Leibler, y la de [James y Hastie 1997] y la de [Domingos 2000] que consideran pérdidas generales persiguiendo que la descomposición anterior sea un caso particular, las demás son específicas del problema de clasificación con pérdida 0-1. Sin embargo, siguiendo a [Domingos 2000]:

1. [Kong y Dietterich 1995] ignora la componente del ruido y permite que la varianza sea negativa.
2. La descomposición de [Breiman 1996*b*] define el sesgo de una hipótesis a partir del conjunto total de puntos en el que dicha hipótesis es sesgada, y no permite definir el sesgo en

una observación específica. Igualmente, permite que la varianza sea cero o indeterminada en situaciones en que el algoritmo puede fluctuar en función de la muestra de entrenamiento.

3. [Tibshirani 1996*a*] descompone el riesgo en sesgo y una cantidad denominada *efecto de agregación* que no está relacionada con su concepto de varianza.

4. [James y Hastie 1997] generalizan el enfoque anterior definiendo sesgo y varianza, pero descomponiendo el riesgo en dos cantidades denominadas *efecto sistemático* y *efecto de varianza*.

5. La propuesta de [Kohavi y Wolpert 1996] permite que el sesgo de la regla de Bayes sea distinto de cero.

6. La descomposición de [Friedman 1997] relaciona la pérdida 0-1 con el sesgo y la varianza según la pérdida cuadrática del estimador de las probabilidades a posteriori.

Como dice [Domingos 2000], dichas descomposiciones, o bien no se realizan en términos del sesgo y la varianza de la pérdida 0-1, o bien, se desarrollan de forma independiente a la original de la pérdida cuadrática sin una clara relación con ésta.

A nuestro juicio, la propuesta de [Domingos 2000] es la que mejor capta la esencia de la descomposición del riesgo y el significado del sesgo y la varianza de un estimador, en los términos que resumimos a continuación[20].

Definición 12 *Se definen los siguientes conceptos:*

1. *Se define la* ***hipótesis óptima*** f *como la que para cada* $\mathbf{x} \in \mathcal{X}$ *verifica:*

$$\mathsf{f}(\mathbf{x}) = \arg\min_{f} \mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\ell(\mathsf{Y}, f(\mathbf{x}))]$$

   y $\mathsf{f}(\mathbf{x})$ *se denomina* ***predicción óptima*** *en* $\mathbf{x} \in \mathcal{X}$.

2. $\mathcal{D} = \{\mathbf{z}^n : n \in N \subset \mathbb{N}^+\}$ *es un conjunto de muestras de entrenamiento, no necesariamente todas las posibles. En este contexto,* Z^n *es la variable aleatoria cuyas realizaciones son muestras de* $\mathcal{D}$ *y* $\mathbb{E}_{\mathcal{D}} \equiv \mathbb{E}_{\mathsf{Z}^n}$ *denota la esperanza con respecto a dicha variable.*

3. *Dado* $\mathbf{x} \in \mathcal{X}$, *la* ***predicción principal*** *(main prediction)* $\hat{f}^*_{\mathcal{D}}(\mathbf{x})$ *con respecto al conjunto* $\mathcal{D}$ *es la que verifica:*

$$\hat{f}^*_{\mathcal{D}}(\mathbf{x}) = \arg\min_{f(\mathbf{x})} \mathbb{E}_{\mathsf{Z}^n}[\ell(\hat{f}_{\mathsf{Z}^n}(\mathbf{x}), f(\mathbf{x}))]$$

   donde $\hat{f}_{\mathsf{Z}^n} \in \mathcal{F}$ *es cualquier hipótesis producida por el algoritmo cuando se entrena con cualquier muestra* $\mathsf{Z}^n \in \mathcal{D}$.

Por tanto, la predicción principal es la que menos *difiere* del conjunto de las posibles predicciones realizadas por el algoritmo a partir de muestras de entrenamiento de $\mathcal{D}$. Esta predicción principal no tiene por que pertenecer al conjunto de las predicciones $\{\hat{f}_{\mathbf{z}^n}(\mathbf{x}) : \mathbf{z}^n \in \mathcal{D}\}$ producidas por el algoritmo en base a muestras $\mathbf{z}^n \in \mathcal{D}$.

De las definiciones anteriores se deduce que, en el caso de pérdida cuadrática, la hipótesis óptima f es la función de regresión y $\hat{f}^*_{\mathcal{D}}(\mathbf{x})$ es la media de las posibles predicciones $\hat{f}_{\mathbf{z}^n}(\mathbf{x})$ con

[20] En algunos conceptos, la notación es específica de este apartado.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$\mathbf{z}^n \in \mathcal{D}$. En el caso de pérdida 0-1, f es la regla de Bayes y $\hat{f}^*_{\mathcal{D}}(\mathbf{x})$ es la moda de las posibles predicciones $\hat{f}_{\mathbf{z}^n}(\mathbf{x})$ con $\mathbf{z}^n \in \mathcal{D}$. En el caso de pérdida absoluta, $\ell(y, y') = |y - y'|$, $\mathsf{f}(\mathbf{x})$ es la mediana de la distribución de $\mathsf{Y}/\mathbf{x}$ (la regla de Bayes en un problema de clasificación dicotómica) y $\hat{f}^*_{\mathcal{D}}(\mathbf{x})$ es la mediana de las posibles predicciones $\hat{f}_{\mathbf{z}^n}(\mathbf{x})$ con $\mathbf{z}^n \in \mathcal{D}$.

Definición 13 *Se definen los siguientes conceptos:*

1. *El **sesgo** de un algoritmo en un punto $\mathbf{x} \in \mathcal{X}$, se define como la pérdida incurrida por la predicción principal con respecto a la predicción óptima:*

$$\text{Sesgo}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x})] = \ell(\mathsf{f}(\mathbf{x}), \hat{f}^*_{\mathcal{D}}(\mathbf{x}))$$

2. *La **varianza** de un algoritmo en $\mathbf{x} \in \mathcal{X}$ se define como la pérdida media incurrida por el algoritmo con respecto a la predicción principal:*

$$\text{Var}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x})] = \mathbb{E}_{\mathsf{Z}^n}[\ell(\hat{f}^*_{\mathcal{D}}(\mathbf{x}), \hat{f}_{\mathsf{Z}^n}(\mathbf{x}))]$$

3. *El **sesgo esperado** y la **varianza esperada** se definen, respectivamente, como $\mathbb{E}_{\mathsf{X}}\{\text{Sesgo}[\hat{f}_{\mathsf{Z}^n}(\mathsf{X})]\}$ y $\mathbb{E}_{\mathsf{X}}\{\text{Var}[\hat{f}_{\mathsf{Z}^n}(\mathsf{X})]\}$.*

4. *El **ruido** en un punto $\mathbf{x} \in \mathcal{X}$ se define como:*

$$\sigma^2(\mathbf{x}) = \mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\ell(\mathsf{Y}, \mathsf{f}(\mathbf{x}))]$$

De las definiciones anteriores, se deduce que:

1. El ruido es una magnitud inevitable que, en el caso de pérdida cuadrática, se corresponde con la varianza de $\mathsf{Y}/\mathbf{x}$.

2. $\hat{f}^*_{\mathcal{D}}(\mathbf{x})$ es una **medida de tendencia central del algoritmo**, (el significado de tendencia central depende de la pérdida considerada).

3. El sesgo mide la **pérdida sistemática** incurrida por el algoritmo. Es independiente de la muestra de entrenamiento y es nulo para un algoritmo que siempre produce la predicción óptima.

4. La varianza mide la **pérdida debida a las fluctuaciones** de la predicción $\hat{f}_{\mathsf{Z}^n}(\mathbf{x})$ del algoritmo en las distintas muestras de $\mathcal{D}$. Es independiente del verdadero valor de $\mathsf{Y}/\mathbf{x}$ y es nula para un algoritmo que siempre realice la misma predicción.

Además, si la pérdida es no negativa, entonces, el sesgo y la varianza son siempre no negativos. Pero la única restricción exigida para la pérdida en todo lo anterior, es que su esperanza pueda calcularse en cada caso.

Con las definiciones anteriores, se verifica el siguiente teorema:

Teorema 14 *[Domingos 2000, th. 1] En un problema de clasificación con dos clases, para cualquier pérdida ℓ que satisfaga $\ell(y, y) = 0\ \forall y \in \mathcal{Y}$ y $\ell(y, y') \neq 0\ \forall y, y' \in \mathcal{Y}$, entonces el riesgo esperado satisface la siguiente descomposición*[21]*:*

$$\text{MPE}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x})] \doteq \mathbb{E}_{\mathsf{Z}^n, \mathsf{Y}/\mathbf{x}}[\ell(\mathsf{Y}, \hat{f}_{\mathsf{Z}^n}(\mathbf{x}))] = c_1\sigma^2(\mathbf{x}) + \text{Sesgo}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x})] + c_2\,\text{Var}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x})] \quad \text{(I.16)}$$

[21] Utilizamos el mismo acrónimo que en el caso de pérdida cuadrática, pero eliminando la "s"de "*squared*". Igualmente, denominamos esperanza del error (ME) a la contrapartida de la esperanza del error cuadrático (MSE).

siendo:

$$
\begin{aligned}
c_1 &= \mathbb{P}_{\mathsf{Z}^n}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x}) = \mathsf{f}(\mathbf{x})] - \frac{\ell(\mathsf{f}(\mathbf{x}), \hat{f}_{\mathsf{Z}^n}(\mathbf{x}))}{\ell(\hat{f}_{\mathsf{Z}^n}(\mathbf{x}), \mathsf{f}(\mathbf{x}))}\mathbb{P}_{\mathsf{Z}^n}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x}) \neq \mathsf{f}(\mathbf{x})] \\
c_2 &= \begin{cases} 1 & \textit{si } \hat{f}^*_{\mathcal{D}}(\mathbf{x}) = \mathsf{f}(\mathbf{x}),\ \textit{es decir, } \mathrm{Sesgo}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x})] = 0 \\ -\dfrac{\ell(\mathsf{f}(\mathbf{x}), \hat{f}^*_{\mathcal{D}}(\mathbf{x}))}{\ell(\hat{f}^*_{\mathcal{D}}(\mathbf{x}), \mathsf{f}(\mathbf{x}))} & \textit{en otro caso} \end{cases}
\end{aligned}
$$

En particular, siguiendo con el problema de clasificación con dos clases, si la función de pérdida es simétrica (por ejemplo, la pérdida 0-1), es decir, $\ell(y, y') = \ell(y', y)$, $\forall y, y' \in \mathcal{Y}$, se tiene:

$$
\begin{aligned}
c_1 &= 2\mathbb{P}_{\mathsf{Z}^n}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x}) = \mathsf{f}(\mathbf{x})] - 1 \\
c_2 &= \begin{cases} 1 & \text{si } \hat{f}^*_{\mathcal{D}}(\mathbf{x}) = \mathsf{f}(\mathbf{x}) \\ -1 & \text{en otro caso} \end{cases}
\end{aligned}
$$

Cuando se utiliza una pérdida simétrica en un problema de clasificación, es importante destacar dos circunstancias cruciales que no se producen con la pérdida cuadrática en el problema de regresión:

1. Cuando la hipótesis producida por el algoritmo es insesgada ($\ell(\mathsf{f}(\mathbf{x}), \hat{f}^*_{\mathcal{D}}(\mathbf{x})) = 0$), la varianza posee un efecto aditivo en el riesgo esperado, sin embargo, cuando dicha hipótesis es sesgada ($\ell(\mathsf{f}(\mathbf{x}), \hat{f}^*_{\mathcal{D}}(\mathbf{x})) \neq 0$), su efecto es beneficioso para el riesgo esperado pues, en ese caso, $c_2 < 0$. Este comportamiento, ayuda a explicar que clasificadores altamente inestables puedan producir buenos resultados incluso con pocos datos.

2. En muestras de entrenamiento y puntos $\mathbf{x}$ de predicción para los cuáles $\hat{f}_{\mathsf{Z}^n}(\mathbf{x}) \neq \mathsf{f}(\mathbf{x})$, el incremento del ruido hace decrecer el riesgo esperado, con lo que un alto nivel de ruido puede ser, en principio, beneficioso para el estimador.

Se verifica asimismo un teorema análogo para el problema de clasificación con varias clases:

Teorema 15 *[Domingos 2000, th. 2] Para un problema de clasificación con varias clases y pérdida 0-1, la ecuación (I.16) es válida con:*

$$
\begin{aligned}
c_1 &= \mathbb{P}_{\mathsf{Z}^n}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x}) = \mathsf{f}(\mathbf{x})] - \mathbb{P}_{\mathsf{Z}^n}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x}) \neq \mathsf{f}(\mathbf{x})]\mathbb{P}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x}) = \mathsf{Y} | \mathsf{f}(\mathbf{x}) \neq \mathsf{Y}] \\
c_2 &= \begin{cases} 1 & \textit{si } \hat{f}^*_{\mathcal{D}}(\mathbf{x}) = \mathsf{f}(\mathbf{x}) \\ -\mathbb{P}_{\mathsf{Z}^n}[\hat{f}_{\mathsf{Z}^n}(\mathbf{x}) = \mathsf{f}(\mathbf{x}) | \hat{f}_{\mathsf{Z}^n}(\mathbf{x}) \neq \hat{f}^*_{\mathcal{D}}(\mathbf{x})] & \textit{en otro caso} \end{cases}
\end{aligned}
$$

I.5.2 Criterios Asintóticos Clásicos

Desgraciadamente, el análisis anterior es puramente conceptual al desconocerse la distribución de Z. Por ello, la mayor parte de los criterios asintóticos se basan en estimadores de la capacidad de generalización, con buenas propiedades asintóticas. Estos criterios, se han desarrollado básicamente para el problema de regresión.

USAL UNIVERSIDAD DE SALAMANCA DE CÓRDOBA

Pérdida Cuadrática y Modelos Realizables, Lineales o Aproximados Linealmente

Los criterios clásicos de selección del modelo se obtienen en el marco del problema de regresión suponiendo pérdida cuadrática (o distribución gausiana), que el modelo está bien especificado (realizabilidad o insesgadez) y que el modelo es de tipo lineal o aproximable linealmente.

Sea una muestra aleatoria $\mathsf{Y}^n/\mathbf{x}^n \doteq (\mathsf{Y}_1/\mathbf{x}_1, ..., \mathsf{Y}_n/\mathbf{x}_n)$ obtenida del modelo $\mathsf{Y}_i = \mathsf{f}(\mathbf{x}_i)+\varepsilon_i$, $i = 1 : n$, con $\{\varepsilon_i\}_{i=1}^n$ independientes e idénticamente distribuídas de media cero y $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1 : n$.

Sea $\mathsf{Y}_*^n/\mathbf{x}^n \doteq (\mathsf{Y}_1^*/\mathbf{x}_1, ..., \mathsf{Y}_n^*/\mathbf{x}_n)$ una nueva muestra aleatoria independiente de la anterior. Se definen los siguientes criterios de error:

1. **Media del error cuadrático**, (*averaged squared error*):

$$\mathrm{ASE}(\hat{f}_n; \mathbf{x}^n) = \frac{1}{n}\sum_{i=1}^{n}[\hat{f}_n(\mathbf{x}_i) - \mathsf{f}(\mathbf{x}_i)]^2 = \frac{1}{n}\left\|\hat{\mathbf{f}}_n - \mathbf{f}\right\|_2^2$$

   donde $(\hat{\mathbf{f}}_n)_i = \hat{f}_n(\mathbf{x}_i)$ y $(\mathbf{f})_i = \mathsf{f}(\mathbf{x}_i)$, $i = 1 : n$ y $\|\cdot\|_2$ es la norma euclídea.

2. **Media del error cuadrático medio**, (*averaged mean squared error*):

$$\mathrm{AMSE}(\hat{f}_n; \mathbf{x}^n) = \mathbb{E}_{\mathbf{Y}^n/\mathbf{x}^n}[\mathrm{ASE}(\hat{f}_n; \mathbf{x}^n)]$$

3. **Media del error cuadrático de predicción**, (*averaged mean prediction squared error*) que es la contrapartida del anterior en predicción, y suele tomarse como aproximación al MPSE reemplazando $\mathbb{E}_{\mathsf{Z}}$ por la media en una muestra de predicción:

$$\begin{aligned}\mathrm{AMPSE}(\hat{f}_n; \mathbf{x}^n) &= \mathbb{E}_{\mathbf{Y}_*^n/\mathbf{x}^n}\mathbb{E}_{\mathbf{Y}^n/\mathbf{x}^n}[\frac{1}{n}\sum_{i=1}^{n}(\mathsf{Y}_i^* - \hat{f}_n(\mathbf{x}_i))^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{Y}^n/\mathbf{x}^n}\mathbb{E}_{\mathbf{Y}_i^*/\mathbf{x}_i}[(\mathsf{Y}_i^* - \hat{f}_n(\mathbf{x}_i))^2] \\ &= \sigma^2 + \mathrm{AMSE}(\hat{f}_n; \mathbf{x}^n)\end{aligned}$$

Veamos la calidad del riesgo empírico $\hat{R}_n(\hat{f}_n)$ como estimador del AMPSE. Denotando por $\mathbf{y}$ una realización $\mathbf{y}^n = (y_1, ..., y_n)^T$ de $\mathsf{Y}^n/\mathbf{x}^n$, se verifica que:

$$\begin{aligned}\hat{R}_n(\hat{f}_n) &= \frac{1}{n}\left\|\mathbf{y} - \hat{\mathbf{f}}_n\right\|^2 = \frac{1}{n}\left\|\mathbf{y} - \mathbf{f} + \mathbf{f} - \hat{\mathbf{f}}_n\right\|^2 \\ &= \frac{1}{n}\|\mathbf{y} - \mathbf{f}\|^2 + \frac{1}{n}\left\|\mathbf{f} - \hat{\mathbf{f}}_n\right\|^2 + \frac{2}{n}(\mathbf{y} - \mathbf{f})^T(\mathbf{f} - \hat{\mathbf{f}}_n)]\end{aligned}$$

Denotando $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$ con $\varepsilon_i \overset{d}{=} \varepsilon$, $i = 1 : n$, es decir distribuidas igual que ε, y $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$ donde $\mathbf{I}_n$ es la matriz identidad de tamaño $n \times n$, se obtiene:

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}^n/\mathbf{x}^n}[\hat{R}_n(\hat{f}_n)] &= \sigma^2 + \mathrm{AMSE}(\hat{f}_n; \mathbf{x}^n) - \frac{2}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^T(\hat{\mathbf{f}}_n - \mathbf{f})] \\ &= \mathrm{AMPSE}(\hat{f}_n; \mathbf{x}^n) - \frac{2}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^T(\hat{\mathbf{f}}_n - \mathbf{f})] \qquad \text{(I.17)}\end{aligned}$$

Para conocer la calidad de $\hat{R}_n(\hat{f}_n)$ como estimador de AMPSE, es necesario evaluar el segundo término de (I.17):

1. Si se utilizan suavizadores lineales[22] de la forma $\hat{\mathbf{f}}_n = \mathbf{S}\mathbf{y}$ donde $\mathbf{S}$ es una matriz *hat* ([Eubank 1984], [Hastie y Tibshirani 1990]) $n \times n$ estimada a partir de $\mathbf{y}$, se tiene:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^T(\hat{\mathbf{f}}_n - \mathbf{f})] &= \mathbb{E}_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}^T\hat{\mathbf{f}}_n) = \mathbb{E}_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}^T\mathbf{S}\mathbf{y}) = \mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^T\mathbf{S}(\mathbf{f}+\boldsymbol{\varepsilon})] \\ &= \mathbb{E}_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}^T\mathbf{S}\boldsymbol{\varepsilon}) = \sigma^2\,\mathrm{tr}(\mathbf{S}) = \sigma^2 p\end{aligned}$$

   con $p = \mathrm{tr}(\mathbf{S})$ donde tr denota la traza.

2. Para modelos no lineales $\mathcal{F} = \{f(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$ realizables, $\mathsf{f}(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}^*)$, si se aproxima linealmente mediante Taylor:

$$\hat{f}_n(\mathbf{x}) \approx \mathsf{f}(\mathbf{x}) + \dot{\mathbf{f}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \Rightarrow (\hat{f}_n(\mathbf{x}) - \mathsf{f}(\mathbf{x})) = \dot{\mathbf{f}}^T(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$$

   donde denotamos por $\dot{\mathbf{f}} = \nabla_{\boldsymbol{\theta}} f(\mathbf{x},\boldsymbol{\theta}^*) = (\partial f(\mathbf{x},\boldsymbol{\theta}^*)/\partial\theta_1, ..., \partial f(\mathbf{x},\boldsymbol{\theta}^*)/\partial\theta_p)^T$ el gradiente respecto de $\boldsymbol{\theta}$ en $\boldsymbol{\theta}^*$, con componentes $(\dot{\mathbf{f}})_j = \partial f(\mathbf{x};\boldsymbol{\theta}^*)/\partial\theta_j$, $j = 1:p$, entonces se verifica que:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^T(\hat{\mathbf{f}}_n - \mathbf{f})] &\approx \mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^T\dot{\mathbf{F}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)] = \mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}^T\dot{\mathbf{F}}(\dot{\mathbf{F}}^T\dot{\mathbf{F}})^{-1}\dot{\mathbf{F}}^T\boldsymbol{\varepsilon}] \\ &= \mathrm{tr}[\dot{\mathbf{F}}(\dot{\mathbf{F}}^T\dot{\mathbf{F}})^{-1}\dot{\mathbf{F}}^T\sigma^2\mathbf{I}_n] = \sigma^2\,\mathrm{tr}[\dot{\mathbf{F}}^T\dot{\mathbf{F}}(\dot{\mathbf{F}}^T\dot{\mathbf{F}})^{-1}] = \sigma^2 p\end{aligned}$$

   donde $\dot{\mathbf{F}}$ es la matriz con elementos $(\dot{\mathbf{F}})_{ij} = \partial f(\mathbf{x}_i)/\partial\theta_j$ $i = 1:n$, $j = 1:p$, y se supone $\dot{\mathbf{F}}^T\dot{\mathbf{F}}$ de máximo rango.

Así pues, en estos casos, $\hat{R}_n(\hat{f}_n)$ es sesgado negativamente como estimador de AMPSE debido al término $\mathbb{E}[\boldsymbol{\varepsilon}^T\hat{\mathbf{f}}_n]$ en (I.17), el cuál, recoge el efecto de evaluar el modelo en los mismos datos utilizados para la estimación. De esta forma, es necesario corregirlo y una corrección natural es:

$$\widehat{\mathrm{AMPSE}} = \hat{R}_n + \frac{2p}{n}\sigma^2$$

estimador que sirve de criterio de selección del modelo como indicador de la capacidad de generalización.

Pérdida General Diferenciable, Modelos No Lineales y Posiblemente no Realizables

Sin embargo, la utilización de este criterio clásico no está justificada cuando: la pérdida no es cuadrática, no se puede asumir la realizabilidad del modelo o se utiliza un término estabilizador en la función objetivo.

Para cualquier función de pérdida y modelos no realizables o regularizados, se tiene el siguiente resultado inspirado en [Stone 1977*b*], ([White 1982], [Linhart y Volkers 1984], [Murata *et al.* 1991], [Amari y Murata 1993], [Murata *et al.* 1994], [Ripley 1996], [Ljung 1999]), conocido como de la **curva de aprendizaje**, (*learning curve*) debido a que, estudia la evolución de la bondad del aprendizaje a medida que crece el tamaño muestral.

Dicho resultado, proporciona resultados similares a los anteriores utilizando el enfoque de la quasi-verosimilitud mediante una aproximación cuadrática del riesgo (que, para pérdida cuadrática, equivale realmente a una aproximación lineal del modelo).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[22] Incluyen como caso particular el modelo de regresión lineal múltiple $\hat{f}_n(\mathbf{x}) = f(\mathbf{x};\hat{\boldsymbol{\theta}}_n) = \mathbf{x}^T\hat{\boldsymbol{\theta}}_n = \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ con $\mathbf{X}$ matriz de diseño con filas $\mathbf{x}_i$, $i = 1:n$.

Supongamos el problema de estimar el modelo $\mathsf{Y} = \mathsf{f}(\mathsf{X}) + \varepsilon$ con $\mathrm{Var}(\varepsilon) = \sigma^2$, a partir de una muestra aleatoria simple de tamaño n, utilizando la familia $\mathcal{F} = \{f(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ compacto$\}$, la pérdida $\ell(y, f(\mathbf{x};\boldsymbol{\theta}))$ y el riesgo $R(\boldsymbol{\theta}) = \mathbb{E}_{\mathsf{Z}}[\ell(\mathsf{Y}, f(\mathsf{X};\boldsymbol{\theta}))]$.

En lo que sigue, $\partial\ell/\partial\boldsymbol{\theta}$ denota el gradiente de ℓ como función de $\boldsymbol{\theta}$ y $\partial^2\ell/\partial\boldsymbol{\theta}^2$ su Hesiano. Suponiendo suficientes condiciones de regularidad que permitan intercambiar esperanza y derivadas, igualmente denotamos por:

$$\frac{\partial R}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathsf{Z}}[\nabla_{\boldsymbol{\theta}}\ell(\mathsf{Y}, f(\mathsf{X};\boldsymbol{\theta}))];\ \frac{\partial^2 R}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\theta}) = \mathbb{E}_{\mathsf{Z}}[\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\ell(\mathsf{Y}, f(\mathsf{X};\boldsymbol{\theta})]$$

el gradiente y el Hesiano, respectivamente, de R como función de $\boldsymbol{\theta}$, y por $\mathrm{Var}_{\mathsf{Z}}[\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\mathsf{Y}, f(\mathsf{X};\boldsymbol{\theta})]$ la correspondiente matriz de varianzas-covarianzas con elementos:

$$\mathrm{Cov}_{\mathsf{Z}}[\frac{\partial}{\partial\theta_i}\ell(\mathsf{Y}, f(\mathsf{X};\boldsymbol{\theta})), \frac{\partial}{\partial\theta_j}\ell(\mathsf{Y}, f(\mathsf{X};\boldsymbol{\theta}))]$$

Teorema 16 *(Learning Curve). Sea* $\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} \hat{R}_n(\boldsymbol{\theta})$ *y* $\boldsymbol{\theta}^* = \arg\min_{\Theta} R(\boldsymbol{\theta})$. *Bajo condiciones suficientes de regularidad, (p. ej. [White 1982]):*

1. *Si* $\ddot{\mathbf{R}} = \frac{\partial^2 R}{\partial\boldsymbol{\theta}^2}(\boldsymbol{\theta}^*)$, $\mathbf{G} = \mathrm{Var}_{\mathsf{Z}}[\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\mathsf{Y}, f(\mathsf{X};\boldsymbol{\theta}^*)]$ *y* $\mathbf{V} = \ddot{\mathbf{R}}^{-1}\mathbf{G}\ddot{\mathbf{R}}^{-1}$, *se verifica que:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}_p, \mathbf{V}) \tag{I.18}$$

   donde $\mathbf{0}_p \in \mathbb{R}^p$ *es el vector nulo.*

2. *Si* $O(\cdot)$ *es la función de orden de Landau (Anexo H):*

$$\begin{aligned}\mathbb{E}_{\mathsf{Z}^n}[\hat{R}_n(\hat{\boldsymbol{\theta}}_n)] &= R(\boldsymbol{\theta}^*) - \frac{1}{2n}\,\mathrm{tr}[\ddot{\mathbf{R}}\mathbf{V}] + O(n^{-3/2})\\ \mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n)] &= R(\boldsymbol{\theta}^*) + \frac{1}{2n}\,\mathrm{tr}[\ddot{\mathbf{R}}\mathbf{V}] + O(n^{-3/2})\end{aligned}$$

   con lo que:

$$\mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n)] = \mathbb{E}_{\mathsf{Z}^n}[\hat{R}_n(\hat{\boldsymbol{\theta}}_n)] + \frac{1}{n}\,\mathrm{tr}[\ddot{\mathbf{R}}\mathbf{V}] + O(n^{-3/2}) \tag{I.19}$$

Situaciones contempladas en el teorema anterior como casos particulares son:

1. **El método de la quasi-verosimilitud** (posible mala especificación), en el que la pérdida es $\ell(y, f(\mathbf{x};\boldsymbol{\theta})) = -\ln q(y;\mathbf{x},\boldsymbol{\theta})$ siendo $\mathcal{Q} = \{q(y;\mathbf{x},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ una familia de modelos de densidad para $\mathsf{Y}/\mathbf{x}$ que no incluye necesariamente a la verdadera. En este caso,

$$\begin{aligned}\ddot{\mathbf{R}} &= \mathbb{E}_{\mathsf{Y}/\mathbf{x}}[-\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\ln q(\mathsf{Y};\mathbf{x},\boldsymbol{\theta}^*)]\\ \mathbf{G} &= \mathrm{Var}_{\mathsf{Y}/\mathbf{x}}[\frac{\partial}{\partial\boldsymbol{\theta}}\ln q(\mathsf{Y};\mathbf{x},\boldsymbol{\theta}^*)]\end{aligned}$$

2. **La estimación por la máxima verosimilitud**, donde ahora $\mathsf{p}(y|\mathbf{x}) = q(y;\mathbf{x};\boldsymbol{\theta}^*) \in \mathcal{Q}$ para algún $\boldsymbol{\theta}^*$, con lo que $\ddot{\mathbf{R}} = \mathcal{I}(\boldsymbol{\theta}^*) = \mathbf{G}$ es la matriz de información de Fisher, $\mathrm{Var}(\hat{\boldsymbol{\theta}}_n) \approx$

$\mathbf{V} = \mathcal{I}(\boldsymbol{\theta}^*)^{-1}$ y $\ddot{\mathbf{R}}\mathbf{V} = \mathbf{I}_p$, por lo que el teorema expresa:

$$\left\{ \begin{array}{c} \mathbb{E}_{\mathsf{Z}^n}[\hat{R}_n(\hat{\boldsymbol{\theta}}_n)] \approx R(\boldsymbol{\theta}^*) - \dfrac{1}{2n}p \\ \mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n)] \approx R(\boldsymbol{\theta}^*) + \dfrac{1}{2n}p \end{array} \right| \Rightarrow \mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n)] \approx \mathbb{E}_{\mathsf{Z}^n}[\hat{R}_n(\hat{\boldsymbol{\theta}}_n)] + \frac{p}{n} \tag{I.20}$$

3. **La estimación por mínimos cuadrados**, caso particular del anterior cuando la distribución condicional es gausiana, en cuyo caso, $\ell(y, f(\mathbf{x})) = -\ln q(y; \mathbf{x}, \boldsymbol{\theta}) \propto [y_i - f(\mathbf{x}_i; \boldsymbol{\theta})]^2$ con lo que:

$$\begin{aligned} \ddot{\mathbf{R}} &= \mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\frac{\partial^2}{\partial \boldsymbol{\theta}^2}(\mathsf{Y} - f(\mathbf{x}; \boldsymbol{\theta}^*))^2] \\ &= 2E_{\mathsf{Y}/\mathbf{x}}[\dot{\mathbf{f}}\dot{\mathbf{f}}^T - (\mathsf{Y} - f(\mathbf{x}; \boldsymbol{\theta}^*)\ddot{\mathbf{F}}] = 2\dot{\mathbf{f}}\dot{\mathbf{f}}^T \end{aligned}$$

siendo $(\dot{\mathbf{f}})_i = \partial f(\mathbf{x}; \boldsymbol{\theta}^*)/\partial\theta_i$ y $(\ddot{\mathbf{F}})_{ij} = \partial^2 f(\mathbf{x}_i; \boldsymbol{\theta}^*)/\partial\theta_i\partial\theta_j$, con $i, j = 1 : p$. Como $\mathbf{V} = \sigma^2(\dot{\mathbf{f}}\dot{\mathbf{f}}^T)^{-1}$, se tiene que $\ddot{\mathbf{R}}\mathbf{V} = 2\sigma^2\dot{\mathbf{f}}\dot{\mathbf{f}}^T(\dot{\mathbf{f}}\dot{\mathbf{f}}^T)^{-1} = 2\sigma^2\mathbf{I}_p$, con lo que:

$$\mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n)] \approx \mathbb{E}_{\mathsf{Z}^n}[\hat{R}_n(\hat{\boldsymbol{\theta}}_n)] + \frac{2p}{n}\sigma^2 \tag{I.21}$$

4. **El principio ERM regularizado**:

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) = \hat{R}_n(\boldsymbol{\theta}) + \lambda\mathcal{E}(\boldsymbol{\theta})$$

donde ℓ es la pérdida cuadrática, $\mathcal{E}$ es un funcional estabilizador y λ una constante regularizadora. Si por ejemplo, $\mathcal{E}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2 = \boldsymbol{\theta}^T\boldsymbol{\theta}$, conocido como *weigth decay*, se tiene: $\mathbf{V} = \ddot{\mathcal{R}}^{-1}G\ddot{\mathcal{R}}^{-1}$ con $\ddot{\mathcal{R}} = \ddot{\mathbf{R}} + \lambda\mathbf{I}_p$ y $\mathbf{G} = \sigma^2\ddot{\mathbf{R}}$, con lo que, haciendo $\ddot{\mathbf{R}} = \ddot{\mathcal{R}} - \lambda\mathbf{I}_p$, se tiene:

$$\begin{aligned} \operatorname{tr}[\ddot{\mathcal{R}}\mathbf{V}] &= \sigma^2 \operatorname{tr}[\ddot{\mathcal{R}}\ddot{\mathcal{R}}^{-1}\ddot{\mathbf{R}}\ddot{\mathcal{R}}^{-1}] = \operatorname{tr}[\ddot{\mathbf{R}}\ddot{\mathcal{R}}^{-1}] \\ &= \operatorname{tr}[(\ddot{\mathcal{R}} - \lambda\mathbf{I}_p)\ddot{\mathcal{R}}^{-1}] \\ &= \operatorname{tr}[\mathbf{I}_p - \lambda\ddot{\mathcal{R}}^{-1}] = p - \lambda\operatorname{tr}[\ddot{\mathcal{R}}^{-1}] \end{aligned}$$

donde $\operatorname{tr}[\ddot{\mathcal{R}}\mathbf{V}] = p_{eff}$ es el **número efectivo de parámetros** del modelo y $\lambda\operatorname{tr}[\ddot{\mathcal{R}}^{-1}]$ puede entenderse como la reducción efectiva de los parámetros después de la regularización. Si los autovalores de la matriz $\ddot{\mathbf{R}}$ son μ_j, $j = 1, ..., p$, los autovalores de $\ddot{\mathcal{R}}^{-1}$ son $1/(\lambda + \mu_j)$ con lo que el número efectivo de parámetros es:

$$\begin{aligned} p_\lambda &= p - \lambda\operatorname{tr}[\ddot{\mathcal{R}}^{-1}] = p - \lambda\sum_{j=1}^{p}\frac{1}{\lambda + \mu_j} \\ &= \sum_{j=1}^{p}[1 - \frac{\lambda}{\lambda + \mu_j}] = \sum_{j=1}^{p}\frac{\mu_j}{\lambda + \mu_j} \end{aligned} \tag{I.22}$$

En el caso de ausencia de regularización, ($\lambda = 0$), el número efectivo de parámetros p_λ coincide con el número de parámetros p del modelo.

Las expresiones anteriores muestran el *coste* de los parámetros en términos del riesgo esperado y, además, vuelven a poner de manifiesto el necesario equilibrio entre complejidad y estabilidad: cada parámetro adicional en pos de una mayor complejidad en el modelo supone, por

ejemplo en (I.21), una penalización de $2\sigma^2/n$ en el riesgo esperado pero, al mismo tiempo, tiende a producir un menor valor en el primer término de la suma.

De los resultados anteriores pueden deducirse diversos criterios clásicos de selección del modelo, todos ellos similares en espíritu a la ecuación (I.10) pero, ahora, obtenidos mediante un análisis asintótico:

1. **C_p de Mallows**. En el caso de mínimos cuadrados, la expresión (I.21) sugiere estimar el riesgo esperado mediante el riesgo empírico penalizado:

$$\widehat{\mathbb{E}_{\mathbf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n)]} = \widehat{\text{MPSE}}(\hat{\boldsymbol{\theta}}_n) = \hat{R}_n(\hat{\boldsymbol{\theta}}_n) + \frac{2p}{n}\sigma^2 \tag{I.23}$$

   Esta expresión es una forma del criterio C_p de Mallows, ([Gorman y Toman 1966], [Mallows 1973]), que da origen a otros criterios conocidos para la selección del modelo en el caso de pérdida cuadrática:

   (a) **FPE de Akaike**. El FPE (*Final Prediction Error*) de [Akaike 1969], que se obtiene cuando se estima σ^2 mediante el estimador insesgado $\hat{\sigma}_f^2 = \frac{n}{n-p}\hat{R}_n(f)$:

$$\text{FPE}(\hat{f}_n) = \hat{R}_n \frac{1+p/n}{1-p/n} \tag{I.24}$$

   (b) **SMS de Shibata**. El SMS (*Shibata Model Selection*) de [Shibata 1976], que se obtiene cuando se estima σ^2 mediante el método de la máxima verosimilitud $\hat{\sigma}_f^2 = \hat{R}_n(f)$:

$$\text{SMS}(\hat{f}_n) = \hat{R}_n(1 + \frac{2p}{n})$$

2. **AIC de Akaike**. En el caso de máxima verosimilitud, la expresión (I.20) sugiere estimar el riesgo esperado mediante:

$$\widehat{\text{MPSE}} = \hat{R}_n(\hat{f}_n) + \frac{p}{n}$$

   es decir:

$$\frac{1}{2n}\text{AIC}(\hat{f}_n) = -\frac{1}{n}\ln q_f(\mathbf{y}^n; \mathbf{x}^n, \hat{\boldsymbol{\theta}}_n) + \frac{p}{n}$$

   donde $\text{AIC}(\hat{f}_n) = -2\ln q_f(\mathbf{y}^n; \mathbf{x}^n, \hat{\boldsymbol{\theta}}_n) + 2p$ es el criterio de [Akaike 1970], [Akaike 1973].

   El criterio AIC es válido para espacios paramétricos de dimensión infinita, pero presenta problemas para espacios de dimensión finita con muestras de pequeño tamaño. Para corregir AIC en estas situaciones, [Sugiura 1978] y [Hurvich y Tsai 1989] proponen el AIC_c (AIC corregido):

$$\text{AIC}_c(\hat{f}_n) = \text{AIC} + \frac{2(p+1)(p+2)}{n-p-2}$$

3. **NIC de Murata** y **GPE de Moody**. En el caso de una función objetivo penalizada, el teorema 16 motiva el *network information criterion* (NIC) de [Murata *et al.* 1991], [Murata *et al.* 1994]:

$$\text{NIC}(\hat{f}_n) = \hat{R}_n(\hat{f}_n) + \frac{1}{n}\,\text{tr}[\hat{\mathbf{H}}_n^{-1}\hat{\mathbf{G}}_n]$$

   siendo:
$$\begin{cases} \hat{\mathbf{G}}_n = \frac{1}{n}\sum_{i=1}^{n}[\frac{\partial}{\partial\boldsymbol{\theta}}\ell(y_i, \hat{f}_n(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n)]^2 - (\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(y_i, \hat{f}_n(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n))^2 \\ \hat{\mathbf{H}}_n^{-1} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\ell(y_i, \hat{f}_n(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n) \end{cases}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

que generaliza los criterios anteriores. Criterios similares, son el **FPER** de [Larsen y Hansen 1994] y el **GPE** (*generalized prediction error*) de [Moody 1991], [Moody 1992], todos ellos, propuestos en el ámbito de las redes neuronales. Por ejemplo, este último parte de una función objetivo de la forma:

$$\hat{L}_n(\boldsymbol{\theta}) = n\hat{R}_n(\boldsymbol{\theta}) + \lambda \sum_{j=1}^{p} \mathcal{E}(\theta_j)$$

donde $\mathcal{E}$ es una función estabilizadora. El criterio de selección es:

$$\text{GPE}(\hat{f}_n) = \hat{R}_n(\hat{f}_n) + 2\hat{\sigma}^2_{eff}\frac{\hat{p}_{eff}}{n}$$

donde $\hat{p}_{eff} = \text{tr}(\hat{G})$ es el número efectivo de parámetros donde:

$$\hat{G} = \frac{n}{2}\frac{\partial^2 \hat{R}_n}{\partial \mathbf{y} \partial \boldsymbol{\theta}^T}[\frac{\partial^2 \hat{L}_n}{\partial \boldsymbol{\theta}^2}]^{-1}\frac{\partial^2 \hat{R}_n}{\partial \boldsymbol{\theta} \partial \mathbf{y}^T}$$

valorada en $\hat{\boldsymbol{\theta}}_n = \arg\,\text{mín}_{\Theta}\, \hat{L}_n(\boldsymbol{\theta})$, y $\sigma^2_{eff} = \frac{n}{n-p_{eff}}\hat{R}_n$ es el análogo al estimador insesgado de la varianza. Así, el criterio resultante es:

$$\text{GPE}(\hat{f}_n) = \hat{L}_n\frac{n + p_{eff}}{n - p_{eff}}$$

que equivale al NIC para los casos analizados anteriormente y al FPE si no existe regularización, en cuyo caso: $p_{eff} = p$.

Cuando el problema es regularizado, en lugar de realizar la selección del modelo sobre el parámetro regularizador, se han desarrollado métodos para estimar dicho parámetro a partir de una muestra de validación durante el entrenamiento, (véase por ejemplo [Larsen *et al.* 1996], [Larsen *et al.* 1998], [Chen y Hagan 1999]). Una comparación de diversos métodos específicos puede verse en [Sigurdsson *et al.* 2000]. Un estudio sobre los efectos del entrenamiento controlado por una muestra de validación es [Amari *et al.* 1997] (véase capítulo siguiente).

I.5.3 Validación Cruzada: Expresiones Analíticas

La correción a $\mathbb{E}_{\mathbf{Z}^n}[\hat{R}_n(f)]$ realizada anteriormente, es una de las alternativas para conseguir un estimador insesgado de AMPSE a partir del riesgo empírico. Otra de las alternativas es eliminar el término del sesgo buscando la independencia entre $\boldsymbol{\varepsilon}$ y el vector de estimadores $\hat{\mathbf{f}}_n$, de tal forma que, por ejemplo, en (I.17):

$$\mathbb{E}[\boldsymbol{\varepsilon}^T\hat{\mathbf{f}}_n] = \mathbb{E}(\boldsymbol{\varepsilon}^T)\mathbb{E}(\hat{\mathbf{f}}_n) = 0$$

Pero $\hat{\mathbf{f}}_n$ depende de $\boldsymbol{\varepsilon}$ al haberse estimado con los datos de la muestra. Los métodos de **validación cruzada** (*cross validation* –CV), ([Allen 1974], [Stone 1974]), persiguen evitar el optimismo del riesgo empírico. En concreto, el método *s-fold-cross-validation* (CV_s) (p. ej. [Kohavi 1995]) divide los datos en s grupos aproximadamente del mismo tamaño, estimando el modelo s veces con todos los datos menos los de un grupo cada vez y el calculando el error de predicción sobre el grupo no utilizado. Concretamente:

$$\text{CV}_s = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_n^{(i)}(\mathbf{x}_i))^2 \tag{I.25}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde $\hat{f}_n^{(i)}(\mathbf{x}_i)$ es la estimación en $\mathbf{x}_i$ del modelo estimado con todos los datos menos el grupo al que pertenece la observación i-ésima. Si $s = n$, se tiene el método de validación cruzada *leave-one-out* (LOO), ([Allen 1974] lo denomina PRESS –*prediction sum of squares*), que estima el modelo n veces utilizando cada vez $n - 1$ observaciones.

En general, otra forma de identificar los métodos de validación cruzada es mediante el número de observaciones que se dejan fuera en cada entrenamiento, y se denomina *leave-r-out* (LrO) el método en el que cada conjunto de entrenamiento deja fuera r observaciones de la muestra total. Con ello, el método LOO equivale al método CV_n.

En general, es fácil comprobar que el criterio LrO es un estimador insesgado del riesgo esperado para un estimador $\hat{f}_{n-r}$ entrenado con $n - r$ datos independientes, (en particular para el caso $r = 1$ del LOO), con lo que la única fuente de sesgo es la reducción del número de datos (no la dependencia entre el estimador y el ruido).

En paralelo con [Efron 1982, p. 6] en su análisis del *jackknife*[23], (también [Ripley 1996, p. 72]), podemos escribir:

$$\mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n)] \equiv \bar{R}(n) = \bar{R}(\infty) + \frac{a_1(\mathsf{Pz})}{n} + \frac{a_2(\mathsf{Pz})}{n^2} + \cdots$$

con a_i constantes $i = 1, 2, \ldots$, y $\bar{R}(\infty) = R^* = \arg\min_\Theta R(\boldsymbol{\theta})$. En base a ello, el sesgo del criterio LOO como estimador de $\bar{R}(n)$ es:

$$\bar{R}(n-1) - \bar{R}(n) = \frac{a_1}{n(n-1)} \approx \frac{a_1}{n^2}$$

con lo que dicho sesgo es de orden $O(n^{-2})$ mejor que el orden $O(n^{-3/2})$ proporcionado por el teorema 16 de la curva de aprendizaje.

Modelos Lineales

Para modelos lineales de la forma $\hat{\mathbf{f}}_n = \mathbf{S}\mathbf{y}$, puede demostrarse que (p. ej. [Wahba 1990*b*], [Hastie y Tibshirani 1990]):

$$\begin{aligned} \text{LOO}(\hat{f}_n) &= \frac{1}{n}\sum_{i=1}^{n}\frac{[y_i - \hat{f}_n(\mathbf{x}_i)]^2}{(1 - S_{ii})^2} \\ \mathbb{E}_{\mathsf{Z}^n}(\text{LOO}(\hat{f}_n)) &\approx \text{AMPSE}(\hat{f}_n) + \frac{2}{n}\sum_{i=1}^{n} S_{ii} b_i^2 \end{aligned} \tag{I.26}$$

donde $S_{ii} \doteq (\mathbf{S})_i$ $i = 1 : n$, y b_i es el sesgo de $\hat{f}_n(\mathbf{x}_i)$. Una variación al LOO para modelos lineales, es la validación cruzada generalizada, (*generalized cross-validation*, –GCV) de [Craven y Wahba 1979], que evita el cálculo de S_{ii} sustituyéndolo por la media de los términos de la diagonal, $\text{tr}(\mathbf{S})/n$:

$$\text{GCV}(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{f}_n(\mathbf{x}_i))^2}{(1 - \text{tr}(\mathbf{S})/n)^2} = \hat{R}_n(\hat{f}_n)(1 - \frac{p}{n})^{-2} \tag{I.27}$$

donde $p = \text{tr}(S)$ representa los grados de libertad del modelo ([Wahba 1990*b*], [Hastie y Tibshirani 1990]).

[23] A diferencia de LOO, el *jackknife* produciría como estimador del Riesgo esperado: $n\hat{R}_n + (n-1)\text{LOO}$, donde $(n-1)\text{LOO}$ es el estimador *jackknife* del sesgo (optimismo) esperado: $\mathbb{E}_{\mathsf{Z}^n}(R - \hat{R}_n)$ [Efron y Gong 1983, p. 46].

Asimismo, se han desarrollado expresiones analíticas aproximadas de CV_s para modelos lineales, entre las que citamos la de [Shao 1993], en la que n_s es el número de datos utilizado en cada estimación (si $n \bmod(s) = 0$, $n_s = n - n/s$):

$$\widehat{\mathrm{CV}}_s(\hat{f}_n) = \hat{R}_n + \frac{n + n_s}{n_s(n-1)} p\sigma^2$$

Modelos No Lineales

Sin embargo, las expresiones analíticas anteriores no son válidas para modelos no lineales, por lo que, para estos casos se han propuesto expresiones analíticas aproximadas. Por ejemplo, [Liu 1995] y más tarde [Hansen y Larsen 1996], repiten la argumentación dada años atrás por [Stone 1977*b*] aplicándola a problemas regularizados.

Sea $\hat{L}_n(\boldsymbol{\theta}) = n\hat{R}_n(\boldsymbol{\theta}) + \mathcal{E}(\boldsymbol{\theta})$ la función objetivo regularizada que se supone diferenciable infinitas veces con derivadas acotadas en un entorno de $\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} \hat{L}_n(\boldsymbol{\theta})$. Sea $\hat{f}_n^{(i)} = f(\cdot, \hat{\boldsymbol{\theta}}_n^{(i)})$ el estimador obtenido sin utilizar la observación i-ésima, $i = 1 : n$, y $\hat{L}_n^{(j)}(\boldsymbol{\theta}) = \sum_{j \neq i = 1}^{n} \ell(y_i, f(\mathbf{x}_i; \boldsymbol{\theta})) + \mathcal{E}(\boldsymbol{\theta})$, $j = 1 : n$. Si se expande mediante Taylor la condición de óptimo $\nabla_{\boldsymbol{\theta}} \hat{L}_n^{(i)}(\hat{\boldsymbol{\theta}}_n^{(i)}) = 0$ en un entorno de $\hat{\boldsymbol{\theta}}_n$, se obtiene el siguiente resultado:

Teorema 17 *[Liu 1995], [Hansen y Larsen 1996] En las condiciones anteriores, si n es suficientemente grande tal que $\hat{\boldsymbol{\theta}}_n^{(i)}$ y $\hat{\boldsymbol{\theta}}_n$ son suficientemente próximos, el estimador $\hat{\boldsymbol{\theta}}_n^{(i)}$ puede aproximarse mediante:*

$$\hat{\boldsymbol{\theta}}_n^{(i)} \approx \hat{\boldsymbol{\theta}}_n - \mathbf{H}_{(i)}^{-1} \dot{\ell}_i$$

donde $\mathbf{H}_{(j)} = \partial^2 \hat{L}_n^{(j)}(\hat{\boldsymbol{\theta}}_n)/\partial \boldsymbol{\theta}^2$ es el Hesiano que se supone invertible y $\dot{\ell}_i = \nabla_{\boldsymbol{\theta}} \ell(y_i, f(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n))$.

El teorema anterior permite obtener una aproximación a $\hat{\boldsymbol{\theta}}_n^{(i)}$ $i = 1, \ldots, n$, evitando el entrenamiento del modelo n veces, para después utilizarlo en el criterio (I.25) ($r = n$) de selección del modelo, pero, a su vez, permite obtener una aproximación asintótica del propio criterio LOO, si se expande cada término $\ell(y_i, f(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n^{(i)}))$ en un entorno de $\hat{\boldsymbol{\theta}}_n$:

$$\ell(y_i, f(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n^{(i)})) \approx \ell(y_i, f(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)) + (\hat{\boldsymbol{\theta}}_n^{(i)} - \hat{\boldsymbol{\theta}}_n)^T \nabla_{\boldsymbol{\theta}} \ell(y_i, f(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n))$$

$$\Rightarrow \mathrm{LOO}(\hat{\boldsymbol{\theta}}_n) \approx \hat{R}_n(\hat{\boldsymbol{\theta}}_n) + \frac{1}{n} \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_n^{(i)} - \hat{\boldsymbol{\theta}}_n)^T \nabla_{\boldsymbol{\theta}} \ell(y_i, f(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n))$$

Teorema 18 *[Hansen y Larsen 1996] En las mismas condiciones anteriores:*

$$\mathrm{LOO}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)) + \frac{1}{n} \sum_{i=1 i}^{n} \dot{\ell}_i^T \mathbf{H}_{(i)}^{-1} \dot{\ell}_i + o(\frac{1}{n})$$

que, si ℓ es la pérdida cuadrática y denotando $\dot{\mathbf{f}}_i = \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)$, toma la forma:

$$\mathrm{LOO}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_n(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n))^2 \frac{1 + 2\dot{\mathbf{f}}_i^T \mathbf{H}_{(i)}^{-1} \dot{\mathbf{f}}_i}{1 - 2\dot{\mathbf{f}}_i^T \mathbf{H}_{(i)}^{-1} \dot{\mathbf{f}}_i} + o(\frac{1}{n}) \qquad \text{(I.28)}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

El resultado anterior es similar al basado en el teorema 16: si n es suficientemente grande, $\mathbf{H}_{(i)} \approx \mathbf{H}_n = \partial^2 \hat{L}_n(\hat{\boldsymbol{\theta}}_n)/\partial \boldsymbol{\theta}^2$, con lo que $\frac{1}{n}\sum_{i=1}^n \dot{\ell}_i^T \mathbf{H}_{(i)}^{-1} \dot{\ell}_i \approx \mathrm{tr}(\mathbf{H}_n^{-1} G_n)$ con $\mathbf{G}_n = \frac{1}{n}\sum_{i=1}^n \dot{\ell}_i \dot{\ell}_i^T$. Si además, $\ell(y, f(\mathbf{x};\boldsymbol{\theta})) = -\ln q(y;\mathbf{x},\boldsymbol{\theta})$ para alguna familia de densidades $\{q(y;\mathbf{x},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ y el modelo es realizable, entonces $\mathrm{tr}(\mathbf{H}_n^{-1}\mathbf{G}_n) \approx p$ con lo que se tiene el criterio AIC visto anteriormente. Asimismo, puede demostrarse [Liu 1995] que, con pérdida cuadrática regularizada, el criterio anterior equivale asintóticamente al criterio GPE de [Moody 1992].

Sin embargo, la aproximación (I.28) para el caso de pérdida cuadrática no incluye como caso particular la expresión (I.26) del caso lineal debido a que la expresión anterior se deriva utilizando una aproximación lineal de ℓ. [Rivals y Personnaz 2000] proponen otro estimador similar en una aproximación lineal de $f(\mathbf{x};\boldsymbol{\theta})$ que sí lo hace, pero no considera la posibilidad de regularización. Sin embargo, utilizando una aproximación segundo orden de ℓ, se obtiene como aproximación:

$$\widehat{\mathrm{LOO}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}_n(\mathbf{x}_i;\hat{\boldsymbol{\theta}}_n))^2 \frac{1}{(1 - 2\dot{\mathbf{f}}_i^T H_{(i)}^{-1}\dot{\mathbf{f}}_i)^2}$$

que sí posee como caso particular la expresión LOO del modelo lineal.

I.5.4 Comparación de distintos criterios

En los distintos criterios anteriores puede detectarse la estructura:

$$[\textbf{Riesgo Empírico} + \textbf{Penalización Complejidad}] \tag{I.29}$$

que, como vimos, pretende contrarrestar el sesgo optimista del riesgo empírico. Esta estructura se asemeja a la ecuación (I.10) con la única diferencia de que aquí el riesgo empírico corresponde ya al óptimo en alguna familia con p parámetros. Asimismo, se asemeja al problema (I.12) pero éste todavía está en la fase de parametrización de la complejidad de las soluciones óptimas (con el parámetro $1/\lambda$, cuyo papel aquí ya está jugando p): posteriormente, será necesario recurrir a criterios de selección del modelo como los anteriores, para elegir entre las diversas soluciones (λ).

Por tanto, los criterios de selección del modelo actúan realmente como fase final de la estabilización del problema de aprendizaje, persiguiendo su solución única.

Alternativamente, los criterios anteriores (o expresiones equivalentes) admiten la siguiente forma general (p. ej. [Eubank 1988, p. 40], [Härdle *et al.* 1988], [Härdle 1990, p. 167]):

$$\Upsilon(\hat{f}_n) = \hat{R}_n(\hat{f}_n)\kappa(\frac{p}{n})$$

donde κ es una función con valores reales.

En el marco de dicha expresión, las diferencias entre los diferentes criterios residen en la función κ que expresa la fuerza de la penalización del riesgo empírico, en función de p/n, (complejidad relativa al tamaño muestral o número de parámetros por observación).

Una plataforma común para la comparación de dichos criterios la proporciona el desarrollo en serie de Maclaurin de cada función κ produciendo los resultados de la Tabla 3, donde añadimos como referencia el estimador insesgado de la varianza (*unbiased estimator of variance* –UEV): $n\hat{R}_n/(n-p)$ y el criterio BIC (*Bayesian Information Criterion*), que se tratará más adelante, propuesto por [Schwarz 1978] con objeto de paliar la no optimalidad del AIC:

$$\frac{1}{n}\mathrm{BIC}(\hat{f}_n) = -\frac{1}{n}\ln q(\mathbf{y}^n;\mathbf{x}^n,\hat{\boldsymbol{\theta}}_n) + \frac{p}{2n}\ln n \tag{I.30}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Penalización $\kappa(\frac{p}{n})$	**Desarrollo en Serie**
$\kappa_{\text{UEV}}(x) = \dfrac{1}{1-x}$	$= 1 + x + x^2 + x^3 + O(x^4)$
$\kappa_{\text{FPE}}(x) = \dfrac{1+x}{1-x}$	$= 1 + 2x + 2x^2 + 2x^3 + O(x^4)$
$\kappa_{\text{SMS}}(x) = 1 + 2x$	$= 1 + 2x$
$\kappa_{\text{AIC}}(x) = e^{2x}$	$= 1 + 2x + 2x^2 + \frac{4}{3}x^3 + O(x^4)$
$\kappa_{\text{AIC}_c}(x) = \exp(\dfrac{1+x}{1-x})$	$= 1 + 2x + 4x^2 + \frac{22}{3}x^3 + O(x^4)$
$\kappa_{\text{GCV}}(x) = (1-x)^{-2}$	$= 1 + 2x + 3x^2 + 4x^3 + O(x^4)$
$\kappa_{\text{BIC}}(x) = (1 + \dfrac{x}{1-x}\ln x)$	$= \ln n[1 + x + x^2 + x^3 + O(x^4)]$

Tabla 3. Desarrollos asintóticos de la penalización del riesgo empírico para distintos criterios clásicos

Como puede observarse, para un mismo modelo con p parámetros, estimado con una muestra de tamaño n suficientemente grande y $p << n$, se tiene la siguiente jerarquía[24]:

$$\kappa_{\text{UEV}} \leq \kappa_{\text{SMS}} \leq \kappa_{\text{AIC}} \approx \kappa_{\text{FPE}} \approx \kappa_{\text{GCV}} \leq \kappa_{\text{AIC}_c} \leq \kappa_{\text{BIC}}$$

lo que concuerda con las apreciaciones de diversos autores en relación con modelos lineales. Algunas de ellas son las siguientes:

1. La equivalencia entre AIC y FPE, y que ambos tienden a sobreestimar el tamaño del modelo [Shibata 1976].

2. La equivalencia asintótica de LOO (aquí su aproximación, GCV) con AIC [Stone 1977*b*], y con el criterio GPE [Liu 1993] (si no existe regularización, GPE = FPE).

3. La equivalencia [Zhang 1992] entre el método LrO y un criterio de la forma $\hat{R}_n + \lambda p/n$ con $\lambda = \frac{2-p}{1-p}$ donde $p = r/n$ es la fracción de datos que se deja fuera. En ella, LOO equivaldría a un $\lambda = \frac{2n-1}{n-1} \to 2$ cuando n tiende a infinito[25].

4. La equivalencia asintótica entre LOO y AIC, y entre LrO y BIC [Shao 1997].

5. El mejor comportamiento de BIC que AIC y LrO, con muestras pequeñas [Hurvich y Tsai 1989].

I.5.5 Técnicas de Remuestreo

El problema de los criterios asintóticos clásicos es que dejan de ser fiables con muestras pequeñas al estar basados en aproximaciones asintóticas del riesgo esperado. Lo mismo sucede con las expresiones analíticas aproximadas de la validación cruzada para modelos no lineales.

En estas circunstancias, los dos enfoques que se plantean como alternativa para crear independencia entre el ruido y el estimador $\hat{f}_n$ y así reducir el sesgo del riesgo empírico, son las técnicas de validación cruzada no aproximadas y las técnicas de remuestreo.

[24] En lo anterior, debe tenerse en cuenta que, salvo en el caso del GCV, todos los criterios en la forma dada, asumen implícita (mínimos cuadrados) o explícitamente, una distribución condicionada gausiana. En el caso de AIC o BIC ello no es necesario, pero las expresiones dadas se obtienen de dicha hipótesis.

[25] Al respecto, [Zhang 1992] recomienda $p \leq 0{,}75 \equiv (1{,}5 \leq \lambda \leq 5)$ y, en términos de s-fold-CV, recomienda $s \in [10, 15]$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Al respecto de estas últimas, (p. ej. [Politis *et al.* 1999]), y, en particular, de las técnicas bootstrap, ([Efron 1979], [Efron 1982], [Hall 1992], [Efron y Tibshirani 1993], [Davison y Hinkley 1997]), existe cierta controversia sobre el efecto de su utilización, teniendo en cuenta que descansan en la estimación de la distribución del estimador del riesgo esperado, p. ej. el MPSE, (y de los estimadores de los parámetros).

La razón es, la gran posibilidad de que, cuando son necesarios algoritmos de optimización no lineal, éstos no alcancen el mínimo global del riesgo empírico o de su versiones regularizadas, y se queden atrapados en alguno de sus múltiples mínimos locales dependientes de las condiciones iniciales aleatorias del algoritmo[26]. En estos casos, cualquier alteración en la muestra de entrenamiento, puede producir un mínimo local radicalmente diferente al anterior, debido, no sólo a la aleatoriedad muestral, sino también a la propia incertidumbre del proceso de optimización.

Este hecho, pone en cuestión la insesgadez del estimador del riesgo esperado, y produce un incremento artificial de su varianza, lo que heredan las técnicas bootstrap y de validación cruzada que, en su lugar, tienden a producir una media de distintos riesgos empíricos sin una clara vinculación con un parámetro poblacional definido.

Para el caso de LrO, se han propuesto métodos heurísticos destinados a evitar el problema, como por ejemplo el de [Moody 1993], consistente en utilizar en cada subconjunto de entrenamiento, la misma inicialización del algoritmo que se utilizó con la muestra total. Sin embargo, muchos autores (p. ej. [Rivals y Personnaz 2000], [Fine 1999]) siguen considerando que el uso de estas técnicas en redes neuronales, especialmente el bootstrap, no está fundamentada cuando se dan las circunstancias anteriores, –aunque esta opinión no está generalizada, ([Tibshirani 1996*c*], [Baxt y White 1995], [Weigend y LeBaron 1994], [Paass 1993]).

Tal crítica puede aceptarse si el objetivo es estimar el riesgo esperado o construir intervalos de confianza para los parámetros, en cuyo caso, como hemos apuntado, tienden a producirse estimaciones por exceso de la varianza, ([Rivals y Personnaz 2000, p. 480]). Sin embargo, cuando el objetivo es únicamente la selección del modelo, las técnicas de remuestreo podrían tener sentido en sí mismas para comparar algoritmos iguales con arquitecturas diferentes, si se considera la incertidumbre del proceso de optimización como parte de la aleatoriedad del riesgo empírico del modelo[27].

De cualquier forma, la frecuente gran dimensionalidad de los problemas tratados con redes neuronales provoca que las estimaciones bootstrap de la distribución empírica requieran, para ser fiables, muchos más datos de los normalmente disponibles. Por otra parte, si la muestra es grande, la carga computacional de estas técnicas tiende a hacerlas prohibitivas y, en todo caso, en dicha circunstancia, es más recomendable apartar una fracción aleatoria de los datos como muestra de validación.

Por estas razones, el uso de las técnicas bootstrap en redes neuronales, con su finalidad original, se ha restringido notablemente. Sin embargo, ha resurgido recientemente como:

1. Método exploratorio inicial de los algoritmos con el fin de incrementar las posibilidades de obtener mínimos globales: *Bumping* de [Tibshirani y Knight 1999].

[26] Incluso, pueden existir varios mínimos globales, complicando todavía más la situación.

[27] Ya, en otro plano, se encuentra la crítica de los que cuestionan el método mismo de validación cruzada, sobre todo LOO: es interesante la discusión iniciada por [Zhu y Rohwer 1996], la respuesta [Goutte 1997] y, finalmente, la respuesta [Rivals y Personnaz 1999] a este último, que concluye que LOO no se comporta bien seleccionando entre modelos insesgados, incluso de tipo lineal, (en cuyo caso, la varianza es lo importante) en comparación con los test estadísticos de bondad de ajuste cuando éstos pueden aplicarse, y que LOO sólo parece útil cuando alguno de los modelos sujetos a comparación posee sesgo de orden mayor que la varianza de los estimadores.

2. Método de reducción de la varianza, mediante la combinación de las hipótesis obtenidas con las muestras bootstrap: *Bagging* de [Breiman 1996*a*].

De cualquier forma, son razones de espacio las que impiden el tratamiento de estas técnicas en el presente trabajo, y serán objeto de futuras investigaciones.

I.5.6 Técnicas No Asintóticas

De alguna u otra manera, el buen comportamiento de las técnicas asintóticas de selección del modelo descansa sobre un gran tamaño muestral, bien porque los criterios se obtienen a partir de aproximaciones asintóticas del riesgo esperado, como sucede con las técnicas clásicas, o bien, porque sus buenas propiedades sólo se pueden garantizar a nivel asintótico, como es el caso de las técnicas de remuestreo.

Por otra parte, es imposible conocer el tamaño muestral mínimo necesario para obtener un buen comportamiento, pues aquél depende de la verdadera distribución subyacente, que es desconocida.

El conjunto de criterios presentados en este apartado no poseen estas debilidades, al menos teóricamente, y poseen la misma estructura general (I.29) con el fin de contrarrestar el sesgo del riesgo empírico.

Los métodos más relevantes son los siguientes:

1. Métodos bayesianos, ([MacKay 1992*c*], [MacKay 1992*b*], [MacKay 1994], [Bernardo y Smith 1994]). Aunque, en sí, la idea de selección del modelo es contraria a la filosofía bayesiana pura, existen métodos que permiten obtener criterios de selección del modelo.

2. Métodos denominados de control de la complejidad estocástica, basados en la teoría de la información: MDL (*minimum description length*) ([Rissanen 1978], [Rissanen 1983]) y su generalización ([Rissanen 1987], [Barron 1991], [Barron y Cover 1991]).

3. El método de minimización del riesgo estructural, (*structural risk minimization* –SRM), de [Vapnik 1982] y [Vapnik 1998], que utiliza las cotas de la forma (I.10) obtenidas en el contexto de la SLT, el aprendizaje PAC o sus variantes.

4. Métodos de penalización del riesgo empírico basada en datos, ([Bartlett *et al.* 2002], [Lugosi y Nobel 1999]), destinadas a paliar la excesiva holgura de los métodos basados en cotas del tipo *whorst-case* de la SLT o el aprendizaje PAC.

5. Métodos basados en la estabilidad del algoritmo, ([Bousquet y Elisseeff 2002], [Bousquet y Elisseeff 2001], [Kearns y Ron 1999]).

En lo que sigue, presentamos las ideas básicas de dichos métodos, salvo los basados en la complejidad estocástica, ya que, por su especificidad, requerirían un mayor espacio y, por otra parte, poseen una filosofía similar a los métodos bayesianos.

Métodos Bayesianos

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Las técnicas bayesianas de selección del modelo han de analizarse bajo el método bayesiano general, que se estructura en sucesivos niveles de inferencia: el primer nivel o de estimación del

modelo, y el segundo nivel de selección del modelo (que, a su vez, puede estar compuesto de sucesivos niveles).

Con objeto de concretar estas ideas, suponemos el problema en el que se pretende estimar la función de masa de probabilidad o de densidad $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$ a partir de un conjunto de observaciones $\mathbf{y}^n = \{y_i\}_{i=1}^n$ donde y_i es una realización de $\mathsf{Y}/\mathbf{x}_i$, $i = 1 : n$.

1. **Nivel 1. Estimación**. Si se postula que la función de densidad $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$ es de la forma $\mathsf{p}(\cdot|\boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$ donde en Θ está definida una función de masa o de densidad a priori $\mathsf{p}(\boldsymbol{\theta})$ (por brevedad *prior*), el método bayesiano[28] persigue la distribución *a posteriori* (por brevedad, *posterior*) de $\boldsymbol{\theta}$ que, por el teorema de Bayes, viene dada por:

$$\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n) = \frac{\mathsf{p}(\mathbf{y}^n, \boldsymbol{\theta})\mathsf{p}(\boldsymbol{\theta})}{\mathsf{p}(\mathbf{y}^n)} \propto \mathsf{p}(\mathbf{y}^n, \boldsymbol{\theta})\mathsf{p}(\boldsymbol{\theta}) \tag{I.31}$$

donde $\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta}) \equiv \mathsf{p}(\mathbf{y}^n|\mathbf{x}^n, \boldsymbol{\theta}) \equiv \mathsf{p}_{\mathsf{Y}^n/\mathbf{x}^n,\boldsymbol{\theta}}(\mathbf{y}^n)$ es la verosimilitud y $\mathsf{p}(\mathbf{y}^n) \equiv \mathsf{p}(\mathbf{y}^n|\mathbf{x}^n)$ es la función de masa de probabilidad o de densidad de la muestra que, en esta fase de inferencia, puede considerarse como un factor normalizador sin efecto en el problema de estimación.

Ya hemos indicado en el Apartado I.4.1 que, en este marco, tiene su encaje el método de regularización, en el que el estabilizador forma parte de la distribución a priori de los parámetros: con ello, el método bayesiano acepta la necesidad de un método para incorporar información a priori en el modelo y sistematiza su tratamiento.

Con la distribución a posteriori, se puede determinar su máximo, el estimador *maximum a posteriori* –MAP, pero ello responde más a un enfoque frecuentista, (el método de la máxima verosimilitud ponderada, en este caso por la prior). El enfoque bayesiano puro considera los parámetros como variables aleatorias por lo que su verdadera respuesta al problema es la posterior $\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n)$ y, en última instancia, la distribución a posteriori de los datos, (**distribución predictiva**)[29]:

$$\mathsf{p}(y|\mathbf{y}^n) = \int \mathsf{p}(y, \boldsymbol{\theta}|\mathbf{y}^n)d\boldsymbol{\theta} = \int \mathsf{p}(y|\boldsymbol{\theta}, \mathbf{y}^n)\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n)d\boldsymbol{\theta}$$

donde por simplicidad, siguiendo el criterio anterior, hemos obviado las observaciones de la variable independiente implícitas en $\mathbf{y}^n$ que, realmente, están condicionando (p. ej. $\mathsf{p}(y|\mathbf{y}^n) \equiv \mathsf{p}(y|\mathbf{y}^n, \mathbf{x}^n)$).

El procedimiento anterior se denomina **marginalización** y es esencial en el enfoque bayesiano: la respuesta bayesiana al problema de predicción se obtiene como media de las respuestas de diferentes estimadores según la distribución a posteriori de los parámetros.

Con ello, el problema de selección del modelo se diluye en el enfoque bayesiano puro: no se pretende obtener la *mejor* de las estimaciones de los parámetros sino *integrar* todas ellas con la distribución a posteriori.

Si lo que se requiere es una estimación puntual de la variable dependiente, la respuesta del enfoque bayesiano depende de la pérdida asociada: si la pérdida es cuadrática, la solución es la esperanza de la distribución predictiva, pues es la que minimiza dicha expresión; si la

[28] En términos frecuentistas, el planteamiento anterior se interpretaría como la consideración de una familia paramétrica $\mathcal{F} = \{q(\cdot/\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ de hipótesis sobre $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$ entre las que es preciso seleccionar una hipótesis óptima. Sin embargo, la filosofía bayesiana no admite la existencia de una hipótesis mejor que las demás y consideraría el proceso de aprendizaje como un proceso en el cuál, son los propios datos los que van concretando la verdadera densidad $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$ mediante las sucesivas distribuciones predictivas $\mathsf{p}(y/\mathbf{y}^n, \mathbf{x}^n) = \int_\Theta \mathsf{p}(y, \boldsymbol{\theta}/\mathbf{y}^n, \mathbf{x}^n)\mathsf{p}(\boldsymbol{\theta}/\mathbf{y}^n, \mathbf{x}^n)d\boldsymbol{\theta}$

[29] Se utiliza la integral de forma laxa, representando un sumatorio en caso de variables aleatorias discretas.

pérdida es la desviación absoluta, la solución es la mediana. En un problema de clasificación, si la pérdida es 0-1, la solución es la moda.

2. **Nivel 2. Selección del Modelo**. En esta fase, es necesario ampliar la notación para considerar:
   (a) Los **hiperparámetros**, denominados así porque son parámetros que controlan la distribución a priori de los parámetros $\boldsymbol{\theta}$, (que denotamos por $\boldsymbol{\lambda}$), pero también de los datos, (denotados por $\boldsymbol{\beta}$).
   (b) El **modelo** en sí: por una parte, la familia $\mathcal{F}$ postulada para estimar $\mathsf{p}_{\mathbf{Y}/\mathbf{x}}$ y, por otra, una familia $\mathcal{G}$ postulada para la distribución a priori de los parámetros, (que, como vimos en el Apartado I.4.1, determina la forma del estabilizador).

   Con ello, reescribimos (I.31) como:

$$\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathcal{F}, \mathcal{G}) = \frac{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathcal{F})\mathsf{p}(\boldsymbol{\theta}|\boldsymbol{\lambda}, \mathcal{G})}{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathcal{F}, \mathcal{G})} \propto \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathcal{F})\mathsf{p}(\boldsymbol{\theta}|\boldsymbol{\lambda}, \mathcal{G}) \tag{I.32}$$

   En este marco, este nivel de inferencia posee dos subniveles (pero su número dependería de la complejidad de la jerarquía de conceptos):
   (a) La selección de los hiperparámetros $\boldsymbol{\lambda}$ y $\boldsymbol{\beta}$, para cada modelo $\mathcal{F}$ y regularizador $\mathcal{G}$, (nótese que en realidad, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\mathcal{G}}$ y $\boldsymbol{\beta} = \boldsymbol{\beta}_{\mathcal{F}}$ pero simplificamos la notación).
   (b) La selección del modelo: $\mathcal{F}$ para los datos y $\mathcal{G}$ para el estabilizador.

   En lo que respecta a los hiperparámetros, el método de la marginalización produciría para la distribución a posteriori de los parámetros:

$$\begin{aligned} \mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n, \mathcal{F}, \mathcal{G}) &= \int \mathsf{p}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}^n, \mathcal{F}, \mathcal{G}) d\boldsymbol{\beta} d\boldsymbol{\lambda} \\ &= \int \mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathcal{F}, \mathcal{G}) \mathsf{p}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}^n, \mathcal{F}, \mathcal{G}) d\boldsymbol{\beta} d\boldsymbol{\lambda} \end{aligned} \tag{I.33}$$

   Sin embargo, la integral anterior puede ser intratable. Para resolver el problema, existen dos corrientes con origen más computacional que conceptual, que todavía producen controversia:
   (a) La corriente de la **marginalización**, partidaria de seguir el enfoque bayesiano puro con origen en [Buntine y Weigend 1991], pero que, al final, se ve obligada a realizar una aproximación gausiana de la posterior resultante, que puede no considerar la mayor parte de su masa de probabilidad.
   (b) La corriente de la **evidencia**, con origen en [MacKay 1992*a*], [MacKay 1994], [Neal 1996], partidaria de aproximar la integral mediante la determinación de las posibles modas de la posterior de los hiperparámetros $\mathsf{p}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}^n, \mathcal{F}, \mathcal{G})$, basando en ellas el siguiente subnivel de inferencia.

Para una discusión de las ventajas e inconvenientes de ambos enfoques puede verse [MacKay 1994]. Dado que el enfoque de la evidencia es un verdadero método de selección del modelo, lo presentamos brevemente a continuación.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Si suponemos una sóla moda, $(\boldsymbol{\beta}_M, \boldsymbol{\lambda}_M)$, la aproximación a (I.33) es:

$$\begin{aligned} \mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n, \mathcal{F}, \mathcal{G}) &\approx \mathsf{p}(\boldsymbol{\theta}, |\mathbf{y}^n, \boldsymbol{\beta}_M, \boldsymbol{\lambda}_M, \mathcal{F}, \mathcal{G}) \int \mathsf{p}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}^n, \mathcal{F}, \mathcal{G}) d\boldsymbol{\beta} d\boldsymbol{\lambda} \\ &= \mathsf{p}(\boldsymbol{\theta}, |\mathbf{y}^n, \boldsymbol{\beta}_M, \boldsymbol{\lambda}_M, \mathcal{F}, \mathcal{G}) \end{aligned}$$

Además, se tiene:

$$\mathsf{p}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}^n, \mathcal{F}, \mathcal{G}) = \frac{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathcal{F}, \mathcal{G})\mathsf{p}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathcal{F}, \mathcal{G})}{\mathsf{p}(\mathbf{y}^n|\mathcal{F}, \mathcal{G})} \propto \mathsf{p}(\mathbf{y}^n|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathcal{F}, \mathcal{G})\mathsf{p}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathcal{F}, \mathcal{G}) \quad \text{(I.34)}$$

donde el denominador no tiene utilidad en este subnivel de inferencia y $\mathsf{p}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathcal{F}, \mathcal{G})$ es la distribución a priori de los hiperparámetros que puede considerarse no informativa. De esta forma, la maximización de $\mathsf{p}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}^n, \mathcal{F}, \mathcal{G})$ equivale a la maximización de $\mathsf{p}(\mathbf{y}^n|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathcal{F}, \mathcal{G})$, lo que se puede interpretar como que se seleccionan los valores de los hiperparámetros más verosímiles a la vista de los datos.

La cantidad $\mathsf{p}(\mathbf{y}^n|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathcal{F}, \mathcal{G})$ recibe el nombre de **evidencia** de los hiperparámetros y, como puede observarse, es precisamente el factor obviado en el primer nivel de inferencia (I.32).

El cociente entre las evidencias de dos modelos, se conoce como el **factor de Bayes**, (*Bayes Factor*), (p. ej. [Gelfand y Dey 1994], [Bernardo y Smith 1994], [Berger 1985]), del que se deduce el BIC (*Bayesian Information Criterion*) de [Schwarz 1978], avanzado con anterioridad en (I.30), aunque fué derivado por Schwarz en el contexto de una familia exponencial, aproximando asintóticamente la distribución a posteriori, (cuando la distribución a priori no posee ya influencia).

Este enfoque de la evidencia puede continuarse de forma jerarquizada en posteriores subniveles de inferencia, tanto con otros posibles hiperparámetros superiores a los anteriores, como con respecto al modelo de datos $\mathcal{F}$ y de estabilizador $\mathcal{G}$: en el siguiente subnivel, el denominador del teorema de Bayes del subnivel actual, que ahora hemos obviado como factor, pasará a ser la evidencia de $\mathcal{F}$ y $\mathcal{G}$ en el siguiente subnivel y servirá de base para su selección final. Como siempre, para ello, se estudia la distribución a posteriori:

$$\mathsf{p}(\mathcal{F}, \mathcal{G}|\mathbf{y}^n) = \frac{\mathsf{p}(\mathbf{y}^n|\mathcal{F}, \mathcal{G})\mathsf{p}(\mathcal{F}, \mathcal{G})}{\mathsf{p}(\mathbf{y}^n)} \propto \mathsf{p}(\mathbf{y}^n|\mathcal{F}, \mathcal{G})\mathsf{p}(\mathcal{F}, \mathcal{G})$$

donde $\mathsf{p}(\mathcal{F}, \mathcal{G})$ es la distribución a priori del modelo y $\mathsf{p}(\mathbf{y}^n|\mathcal{F}, \mathcal{G})$ es la evidencia (denominador en el subnivel anterior (I.34)).

Una ventaja del enfoque de la evidencia, a diferencia de los criterios clásicos derivados del teorema de la curva de aprendizaje, [Murata *et al.* 1994], es que, por propia construcción, su uso no requiere que los modelos estén anidados, pudiéndose comparar entre sí modelos heterogéneos. El inconveniente es la necesidad de determinar distribuciones a priori que, ante muestras pequeñas, pueden influenciar el resultado de la selección del modelo.

Validación Cruzada Bayesiana

Unido a los métodos anteriores, es preciso citar también las posturas bayesianas que aceptan desde el principio un enfoque de selección del modelo tipo frecuentista, aunque con metodología bayesiana. En este sentido, por ejemplo [Bernardo y Smith 1994] identifica tres posibles enfoques a la selección del modelo, según que: dispongamos de una familia de modelos al que pertenece el *verdadero*, (situación cerrada –*closed*), dispongamos de un conjunto de modelos que queremos contrastar con uno que creemos verdadero, (situación completada –*completed*), o simplemente

disponemos de un conjunto de modelos sin uno que creamos verdadero, (situación abierta), que más se parecería a una posición no paramétrica, y para la que [Bernardo y Smith 1994] indica que los modelos alternativos han de competir en términos de **validación cruzada**.

Siguiendo a [Bernardo y Smith 1994], considérese el problema de selección del modelo como un problema de decisión, en el que, entre un conjunto de modelos $\{\mathcal{F}_j : j \in J\}$, se selecciona el que maximiza la utilidad esperada $\bar{u}$:

$$\mathcal{F}_0 = \arg \max_{j \in J} \bar{u}(\mathcal{F}_j|\mathbf{y}^n)$$

siendo ésta:

$$\bar{u}(\mathcal{F}_j|\mathbf{y}^n) = \int u(\mathcal{F}_j, y|\mathbf{y}^n)\mathsf{p}(y|\mathbf{y}^n)dy$$

donde u es una función de utilidad[30], y donde, si el objetivo son predicciones puntuales, la utilidad u puede elegirse como:

$$u(\mathcal{F}_j, y|\mathbf{y}^n) = -[y - \mathbb{E}(\mathsf{Y}|\mathbf{x}, \mathcal{F}_j, \mathbf{y}^n)]^2$$

donde la esperanza se toma con respecto a la distribución predictiva $(\mathsf{Y}|\mathbf{x}, \mathcal{F}_j, \mathbf{y}^n)$; y si se trata de comparar distribuciones, puede elegirse como:

$$u(\mathcal{F}_j, y|\mathbf{y}^n) = \ln \mathsf{p}(y|\mathcal{F}_j, \mathbf{y}^n)$$

Entonces, si $\mathbf{y}^{n(i)}$ es la muestra $\mathbf{y}^n$ sin la observación y_i, por la ley fuerte de los grandes números se tiene que:

$$\frac{1}{n}\sum_{i=1}^{n} u(\mathcal{F}_j, y|\mathbf{y}^{n(i)}) \underset{n\to\infty}{\overset{\text{c.s.}}{\to}} \int u(\mathcal{F}_j, y|\mathbf{y}^n)\mathsf{p}(y|\mathbf{y}^n)dy$$

donde c.s. denota convergencia casi seguro (Anexo H). Debido a ello, las utilidades esperadas de los modelos $\mathcal{F}_j$ pueden compararse [Bernardo y Smith 1994, p. 404] mediante[31]:

$$U_j = \frac{1}{n}\sum_{i=1}^{n} u(\mathcal{F}_j, y|\mathbf{y}^{n(i)})$$

Nótese que si $u(\mathcal{F}_j, y|\mathbf{y}^n) = \ln \mathsf{p}(y|\mathcal{F}_j, \mathbf{y}^n)$ este método puede considerarse como una generalización del LOO con pérdida cuadrática, a otras pérdidas (el logaritmo de la distribución predictiva $\mathsf{p}(\mathsf{Y}|\mathbf{x}, \mathcal{F}_j, \mathbf{y}^n)$). Una comparación entre ambos, favorable a la versión bayesiana puede verse en [Vila *et al.* 2000], que puede calcular la posterior gracias a la utilización de familias conjugadas.

Minimización del Riesgo Estructural

El método de minimización estructural del riesgo (*structural risk minimization* –SRM) de [Vapnik 1982], sigue la filosofía del método de los *Sièves* y, de hecho, éste último puede encuadrarse dentro de la formulación general de aquél [Vapnik 1998, p. 222]. Al analizar la ecuación (I.10), hemos motivado el problema de selección del modelo, cuyo objetivo, es determinar una regla $v(n)$ que permita acompasar complejidad con tamaño muestral.

[30] La función de utilidad puede definirse como el concepto opuesto a la función de pérdida del problema de aprendizaje. Para más detalles, puede verse p. ej. [Bernardo y Smith 1994] y [Berger 1985].

[31] Este método da lugar al *Pseudo-Bayes Factor* (PSBF), cuya expresión asintótica es [Gelfand y Dey 1994]: $\ln(\text{PSBF}) \approx \ln \mathsf{p}(\mathbf{y}^n|\hat{\boldsymbol{\theta}}_n) + \frac{p}{2}$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En los criterios asintóticos clásicos, el parámetro de complejidad es la dimensión p del espacio paramétrico Θ (o el número equivalente de parámetros), y la regla $v(n)$ está implícita en cada uno de los criterios de tal forma que, la aplicación de éstos, supone seleccionar la dimensión p (complejidad) del modelo estimado que, asintóticamente, produzca la mayor capacidad de generalización.

El procedimiento SRM es análogo: supone una familia marco $\mathcal{F}$ de hipótesis, lo suficientemente compleja, como para contener a la función *verdadera*, y la dota de una estructura anidada de subfamilias[32]:

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_t \subset \cdots$$

que da lugar a una estructura igualmente anidada:

$$\mathcal{L}_1 \subset \mathcal{L}_2 \subset \cdots \subset \mathcal{L}_t \subset \cdots \subset \bigcup_{t=1}^{\infty} \mathcal{L}_h$$

con $\mathcal{L}_t = \{\ell_f(\mathbf{z}) = \ell(y, f(\mathbf{x})) : f \in \mathcal{F}_t\}$ tal que verifique las condiciones de **estructura admisible** para el principio SRM[33] [Vapnik 1998, p. 221]:

$$\begin{array}{l} 1)\ \mathrm{VCdim}(\mathcal{L}_t) = v_t < \infty \\ 2)\ \forall \ell_f \in \mathcal{L}_t :\ 0 \leq \ell_f(\mathbf{z}) \leq b_t, \forall \mathbf{z} \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \\ 3)\ \bigcup_{t=1}^{\infty} \mathcal{L}_h \text{ es densa en } \mathcal{L}_{\mathcal{T}} = \{\ell_{\mathsf{f}}(\mathbf{z}) = \ell(y, \mathsf{f}(\mathbf{x})) : \mathsf{f} \in \mathcal{T}\} \end{array} \tag{I.35}$$

donde la tercera condición persigue garantizar la consistencia del método, al ser $\mathcal{L}_{\mathcal{T}}$ la familia correspondiente a $\mathcal{T}$ a través de ℓ.

Fijado un tamaño muestral n, se aplica el algoritmo a las primeras $N(n)$ clases $\mathcal{F}_t$, $t = 1 : N(n)$ obteniéndose las hipótesis $\hat{f}_{n,1}, ..., \hat{f}_{n,N(n)}$. Posteriormente, se selecciona la hipótesis con menor riesgo garantizado (I.10):

$$\hat{f}_n = \hat{f}_{n,t^*} \text{ con } t^* = \arg \min_{t \in \{1:N(n)\}} \{\hat{R}_n(\hat{f}_{n,t}) + \varepsilon(n, \eta, v_t)\} \tag{I.36}$$

Como puede observarse, el riesgo garantizado juega el mismo papel que los criterios vistos con anterioridad y posee su misma estructura (I.29), donde ahora la complejidad v_t viene dada por la VCdim o la $V_\gamma \dim$ introducidas anteriormente. Sin embargo, el método SRM basa sus diferencias en las propiedades fundamentales del riesgo garantizado:

1. Su carácter no asintótico.
2. Su validez para cualquier distribución que genere los datos.

Una de los pocos análisis comparativos entre este criterio (denominado también *Vapnik measure*) y diferentes criterios clásicos es [Cherkassky *et al.* 1999].

Sin embargo, este método no está exento de dificultades: por una parte, la determinación práctica de la VCdim o $V_\gamma \dim$ de $\mathcal{F}_t$ resulta complicada en ciertos casos, y, por otra, las cotas proporcionadas por el riesgo garantizado suelen ser con frecuencia excesivamente holgadas. Ello, por una parte, no permite estimar el riesgo realmente cometido con la elección y, por otra, no garantiza la elección del modelo con mejor capacidad de generalización.

[32] Esta estructura es similar en espíritu por ejemplo, al método de la multirresolución de la teoría de los wavelets.

[33] Formulamos las condiciones en términos de la dimensión VC, pero vale cualquier otra medida análoga de complejidad (Pdim, $V_\gamma$dim, etc.).

Para resolver el primero de los problemas, se ha propuesto el método de la VCdim efectiva, [Vapnik *et al.* 1994] que estima la VCdim a partir de los datos, mediante el ajuste de $\varepsilon(n, \eta, v)$ con una expresión paramétrica $\Phi(\frac{v}{n})$ a partir de la estimación de la máxima diferencia en la tasa de error:

$$\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right|$$

en dos muestras de tamaño n, para varios tamaños muestrales.

Sin embargo, este procedimiento presenta dificultades, sobre todo en problemas de regresión, ante algoritmos que han de enfrentarse a mínimos locales pues, en ese caso, el área de búsqueda no abarca toda la familia $\mathcal{F}_t$ con lo que la VCdim es menor que la teórica y, además, cambia con el tamaño muestral, lo que hace inestable y poco fiable el ajuste.

La otra dificultad, la holgura de las cotas por excesivamente conservadoras, se debe a su carácter *worst-case*. Este carácter se produce en dos dimensiones diferentes: en términos de la distribución, (el que las cotas sean válidas para cualquier distribución), y en términos de las funciones de la familia $\mathcal{F}$ considerada. Los dos últimos enfoques que presentamos a continuación persiguen liberarse, respectivamente, de estas dos restricciones.

Penalizaciones del Riesgo Empírico Adaptadas a los Datos

En esta línea de trabajo, destacamos las contribuciones de [Lugosi y Nobel 1999] y [Bartlett *et al.* 2002] que se encuadran en el marco de la filosofía SRM: si una familia $\mathcal{F}$ posee suficiente complejidad para representar la *verdadera* función f, lo normal es que produzca sobreajuste con el principio ERM.

Por ello, se controla la complejidad considerando una estructura anidada de subfamilias $\mathcal{F}_t$ con $\cup_t \mathcal{F}_t = \mathcal{F}$ y, a partir de un conjunto de datos $\mathbf{z}^n$, se obtienen diversas hipótesis que minimizan el riesgo empírico en distintas subfamilias de la estructura: $\hat{f}_{n,t} = \arg\min_{\mathcal{F}_t} \hat{R}_n(f)$.

Entonces, se trata de seleccionar la hipótesis $\hat{f}_n \in \mathcal{F}$ que minimiza el siguiente criterio (tipo (I.29)) de selección del modelo:

$$\hat{\mathcal{R}}_n(\hat{f}_{n,t}) = \hat{R}_n(\hat{f}_{n,t}) + \mathcal{E}_n(\hat{f}_{n,t})$$

es decir:

$$\hat{f}_n = \hat{f}_{n,t^*} \text{ con } t^* = \arg\min_t \{\hat{\mathcal{R}}_n(\hat{f}_{n,t})\} \tag{I.37}$$

donde $\mathcal{E}_n(f)$ es un funcional que penaliza la complejidad de f y se define suponiendo disponible un *buen* estimador $\hat{R}_n^E(\hat{f}_t)$ del verdadero riesgo $R(\hat{f}_t)$, mediante:

$$\mathcal{E}_n(\hat{f}_t) = \hat{R}_n^E(\hat{f}_t) - \hat{R}_n(\hat{f}_t) + \sqrt{\frac{\log t}{m}}$$

donde la cantidad $\hat{R}_n^E(\hat{f}_t) - \hat{R}_n(\hat{f}_t)$ es una estimación del *optimismo* de $\hat{R}_n(\hat{f}_t)$ que se introduce para contrarrestar su posible sesgo negativo, y el último término con $m > 0$ aparece por razones puramente técnicas.

En lo anterior, la condición de *buen* estimador $\hat{R}_n^E(\hat{f}_t)$ de $R(\hat{f}_t)$ es una de las siguientes, (la segunda más fuerte que la primera):

$$1)\ \exists c, m > 0 : \mathbb{P}_{\mathbf{Z}^n}[R(\hat{f}_t) > \hat{R}_n^E(\hat{f}_t) + \varepsilon] \leq c e^{-2m\varepsilon^2}\ \forall \varepsilon > 0 \tag{I.38}$$

$$2)\ \exists c, m > 0 : \mathbb{P}_{\mathbf{Z}^n}[R(\hat{f}_t) > \hat{R}_n^E(\hat{f}_t) + \varepsilon] \leq c e^{-m\varepsilon}\ \forall \varepsilon > 0 \tag{I.39}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Entonces, se verifica el siguiente teorema que expresa la bondad del criterio de selección del modelo:

Teorema 19 *[Bartlett* et al. *2002, ths. 1 y 2] Sea* $\mathrm{R} = \inf_f R(f)$ *donde el ínfimo se toma para todas las hipótesis factibles para el problema y sea* $\hat{f}_n$ *la hipótesis seleccionada por el método (I.37). La utilización en dicho método de un estimador* $\hat{R}_n^E$ *que satisface la condición (I.38), o la condición (I.39) anteriores, produce, respectivamente:*

$$\begin{aligned} &1)\ \mathbb{P}_{\mathsf{Z}^n}[R(\hat{f}_n) - \hat{\mathcal{R}}_n(\hat{f}_n) \quad > \quad \varepsilon] \leq 2ce^{-2m\varepsilon^2} \\ &2)\ \mathbb{P}_{\mathsf{Z}^n}[R(\hat{f}_n) - \hat{\mathcal{R}}_n(\hat{f}_n) \quad > \quad \varepsilon] \leq 2ce^{-m\varepsilon} \end{aligned}$$

Además, si se denota por $\Delta_t = [\inf_{f\in\mathcal{F}_t} R(f) - \mathrm{R}]$ *el denominado* ***error de aproximación*** *(término de sesgo) asociado a la utilización de la familia* $\mathcal{F}_t$*, entonces se verifica, respectivamente:*

$$\begin{aligned} &1)\ \mathbb{E}_{\mathsf{Z}^n}[R(\hat{f}_n)] - \mathrm{R} \quad \leq \quad \min_t\{\mathbb{E}_{\mathsf{Z}^n}[\mathcal{E}_n(t)] + \Delta_t\} + \sqrt{\tfrac{\log(ce)}{2m}} \\ &2)\ \mathbb{E}_{\mathsf{Z}^n}[R(\hat{f}_n)] - \mathrm{R} \quad \leq \quad \min_t\{\mathbb{E}_{\mathsf{Z}^n}[\mathcal{E}_n(t)] + \Delta_t\} + \sqrt{\tfrac{\log(2ce)}{m}} \end{aligned}$$

En el marco anterior, encajan muy diversas cotas a la velocidad de convergencia:

1. **Selección del modelo con una muestra de validación**, (*hold-out sample*). Si se supone una muestra de m observaciones independientes $\mathbf{z}'_i = (\mathbf{x}'_i, y'_i)$, $i = 1 : m$ extraídas de Z independientemente de las anteriores, el estimador $\hat{R}_n^E(\hat{f}_t) = \hat{R}'_{n,t}(\hat{f}_t) = \frac{1}{m}\sum_{i=1}^m \ell(y'_i, \hat{f}_t(\mathbf{x}'_i)$ satisface la condición (I.38) con $c = 1$, (de hecho, $\hat{R}'_{n,t}(\hat{f}_t)$ converge en el sentido fuerte a $R(\hat{f}_t)$ [Plutowski *et al.* 1994]), por lo que, aplicando el teorema 19, se obtiene [Bartlett *et al.* 2002, cor. 1]:

$$\mathbb{E}_{\mathsf{Z}^n}[R(\hat{f}_n)] - \mathrm{R} \leq \min_t\{\mathbb{E}_{\mathsf{Z}^n}[R(\hat{f}_t) - \hat{R}_n(\hat{f}_t)] + \Delta_t + \sqrt{\frac{\log t}{m}}\} + \frac{1}{\sqrt{2m}} \tag{I.40}$$

   con lo que el estimador $\hat{f}_n$ seleccionado con el método (I.37) alcanza un equilibrio óptimo entre el error de aproximación Δ_t (término de sesgo) y la cantidad $\mathbb{E}_{\mathsf{Z}^n}[R(\hat{f}_t) - \hat{R}_n(\hat{f}_t)]$ que mide el grado de sobreajuste (término de varianza).

2. **Selección del modelo con un estimador basado en el coeficiente** *shatter* **empírico.** En el caso concreto del problema de clasificación, con pérdida 0-1 y $\mathcal{Y} = \{0, 1\}$, una cota fundamental de la SLT [Vapnik y Chervonenkis 1971] expresa que:

$$\mathbb{P}_{\mathsf{Z}^n}[\sup_{f\in\mathcal{F}_t}(R(f) - \hat{R}_n(f)) > \varepsilon] \leq 4E_{\mathsf{X}^{2n}}[N_{\mathcal{F}_t}(\mathsf{X}^{2n})]e^{-n\varepsilon^2}$$

   donde X^{2n} es una muestra aleatoria simple de tamaño $2n$ de la variable independiente y $N_{\mathcal{F}_t}(\mathsf{X}^{2n})$ es el coeficiente de disección empírico de $\mathcal{F}_t$ para una muestra de tamaño $2n$ (número de posibles formas de clasificar los $2n$ puntos de X^{2n} con las funciones de $\mathcal{F}_t$). Entonces, puede demostrarse [Bartlett *et al.* 2002], que el estimador:

$$\hat{R}_n^E(\hat{f}_t) = \hat{R}_n(\hat{f}_n) + \sqrt{\frac{\ln \mathbb{E}_{\mathsf{X}^{2n}}[N_{\mathcal{F}_t}(\mathsf{X}^{2n})] + \ln 4}{n}} \tag{I.41}$$

   verifica la condición (I.38) con $m = n/2$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

El segundo término de la expresión anterior (es decir $\hat{R}_n^E(\hat{f}_t) - \hat{R}_n(\hat{f}_n)$) puede acotarse para cualquier distribución teniendo en cuenta que, debido a (I.8), se verifica que $\mathbb{E}_{\mathsf{X}^{2n}}[N_{\mathcal{F}_t}(\mathsf{X}^{2n})] \leq (e2n/v_t)^{v_t}$ donde $v_t = \mathrm{VCdim}(\mathcal{F}_t)$. Pero este tipo de cotas *worst-case* presentan los problemas de excesiva holgura mencionados en el apartado anterior. Si, en su lugar, se estima $\mathbb{E}_{\mathsf{X}^{2n}}[N_{\mathcal{F}_t}(\mathsf{X}^{2n})]$ mediante $N_{\mathcal{F}_t}(\mathsf{X}^n)$, y el estimador resultante en (I.41) se utiliza en el criterio de selección del modelo, aplicando el teorema 19 resulta [Bartlett *et al.* 2002, th. 6]:

$$\begin{aligned} \mathbb{E}_{\mathsf{Z}^n}[R(\hat{f}_n)] - \mathrm{R} \;\leq\; & \min_t \{ \sqrt{\frac{12 E_{\mathsf{X}^n}[\ln N_{\mathcal{F}_t}(\mathsf{X}^n)] + \ln 4}{n}} \\ & + \Delta_t + 8{,}95 \sqrt{\frac{\ln t}{n}} \} + \frac{8{,}23}{\sqrt{n}} \end{aligned} \tag{I.42}$$

3. **Selección del modelo con la máxima discrepancia**. En el mismo problema de clasificación anterior, puede adoptarse como estimador:

$$\hat{R}_n^E(\hat{f}_t) = \hat{R}_n(\hat{f}_t) + \max_{f \in \mathcal{F}_t} \{ \hat{R}_n^{(1)}(f) - \hat{R}_n^{(2)}(f) \}$$

donde, suponiendo que n es par, $\hat{R}_n^{(1)}(f) = \frac{2}{n} \sum_{i=1}^{n/2} \ell(y_i, f(\mathbf{x}_i))$ y análogamente $\hat{R}_n^{(2)}$ en la segunda mitad de la muestra. El segundo término de la expresión anterior se denomina **la máxima discrepancia** y se estima cambiando las clases y_i de una de las mitades de la muestra, (técnica utilizada también en la estimación de la VCdim efectiva de [Vapnik *et al.* 1994]). La utilización del estimador anterior en el teorema 19 produce, con $m = n/21$, [Bartlett *et al.* 2002, th. 8]:

$$\begin{aligned} \mathbb{E}_{\mathsf{Z}^n}[R(f_n)] - \mathrm{R} \;\leq\; & \min_t \{ \mathbb{E}_{\mathsf{Z}^n} \max_{f \in \mathcal{F}_t} [\hat{R}_n^{(1)}(f) - \hat{R}_n^{(2)}(f)] \\ & + \Delta_t + 4{,}59 \sqrt{\frac{\ln t}{n}} \} + \frac{4{,}7}{\sqrt{n}} \end{aligned} \tag{I.43}$$

Como puede observarse, las cotas anteriores se construyen en base a penalizaciones obtenidas de los datos, con lo que se espera sean más ajustadas que las cotas para cualquier distribución de la SLT o el aprendizaje PAC.

Métodos Basados en la Estabilidad del Algoritmo

El enfoque de las penalizaciones empíricas descrito anteriormente persigue liberarse de las desventajas del carácter *worst-case* en términos de la distribución.

El otro elemento determinante del carácter *worst-case* de las cotas obtenidas en la SLT o el aprendizaje PAC es la consideración de todas las funciones de la familia $\mathcal{F}$ elegida. Aunque, como hemos indicado en el Apartado I.3.1, esta circunstancia es una condición necesaria [Vapnik 1998] para la consistencia del aprendizaje en la SLT, puede ser que la familia *efectiva* considerada por el algoritmo de entrenamiento sea mucho más reducida que la contemplada a priori.

Por ejemplo, si se utiliza el principio ERM regularizado con $\mathcal{F}$, la familia efectiva utilizada será un subconjunto $\mathcal{F}_r \subset \mathcal{F}$ verificando seguramente: $\mathrm{VCdim}(\mathcal{F}_r) < \mathrm{VCdim}(\mathcal{F})$. Sin embargo, en las cotas a la velocidad de convergencia del tipo (I.10) seguiremos utilizando $\mathrm{VCdim}(\mathcal{F})$ pues la estimación de $\mathrm{VCdim}(\mathcal{F}_r)$, si es posible, es difícil y/o costosa.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En apartados anteriores, hemos analizado la capacidad de generalización de una hipótesis descomponiéndola en sus dos factores fundamentales: el sesgo, determinado por la insuficiente

capacidad aproximadora de la familia, y la varianza, determinada precisamente por el exceso en dicha capacidad que hace al estimador demasiado sensible a la aleatoriedad los datos.

Ello indica, que el sesgo, queda determinado fundamentalmente por la elección de la familia, y la causa fundamental de la varianza reside en el grado de **estabilidad** del algoritmo ante la aleatoriedad de los datos. Por esta razón, el enfoque de la estabilidad se centra en el estudio de la estabilidad del algoritmo. Esta sugestiva idea resume este prometedor enfoque con origen en [Rogers y Wagner 1978], [Devroye y Wagner 1979*a*], [Devroye y Wagner 1979*b*], abordado posteriormente por [Lugosi y Pawlak 1994], [Kohavi 1995] y [Kearns y Ron 1999], y retomado recientemente por [Bousquet y Elisseeff 2002], [Bousquet y Elisseeff 2001].

Con objeto de poner el énfasis en el algoritmo, seguimos la notación de [Bousquet y Elisseeff 2002]: si $A : \mathcal{D} \rightarrow \mathcal{F}$ con $\mathcal{D} = \cup_{n=1}^{\infty} \mathcal{Z}^n$ es un algoritmo, denotamos por: $A_n \equiv A(\mathbf{z}^n) = \hat{f}_n$ la hipótesis producida por A a partir de una muestra aleatoria simple $\mathbf{z}^n$,por $A_n^{(i)} = \hat{f}_n^{(i)}$ la hipótesis obtenida sin contar con la observación $\mathbf{z}_i$, y por y A_n^i la hipótesis obtenida cuando dicha observación se sustituye por otra $\mathbf{z}_i'$ extraída independientemente de P_{Z}.

En consonancia con lo anterior, con un pequeño abuso de notación para no perder la expresividad de los símbolos, denotamos igualmente: $\ell(A_n, \mathbf{z}) \equiv \ell(y, \hat{f}_n(\mathbf{x}))$ con $\mathbf{z} = (\mathbf{x}, y)$, $R(A_n) = R(\hat{f}_n)$ y $\hat{R}_n(A_n) = \hat{R}_n(\hat{f}_n)$.

En todo lo que sigue, se supone un problema de regresión[34] y que el algoritmo es simétrico respecto de $\mathbf{z}^n$, es decir, su resultado no depende del orden de los elementos de $\mathbf{z}^n$.

Definición 20 *[Bousquet y Elisseeff 2002] De un algoritmo A se dice que posee las siguientes propiedades con respecto a la pérdida ℓ:*

1. ***Estabilidad punto a punto*** β_n*, si,* $\forall i \in \{1 : n\}$*:*

$$\mathbb{E}_{\mathsf{Z}^n}[\left|\ell(A_n, \mathbf{z}_i) - \ell(A_n^{(i)}, \mathbf{z}_i)\right|] \leq \beta_n$$

2. ***Estabilidad*** β_n*, si:*

$$\forall i \in \{1 : n\},\ \mathbb{E}_{\mathsf{Z}^n,\mathsf{Z}}[\left|\ell(A_n, \mathsf{Z}) - \ell(A_n^{(i)}, \mathsf{Z})\right|] = \mathbb{E}_{\mathsf{Z}^n}[\left\|\ell(A_n, \cdot) - \ell(A_n^{(i)}, \cdot)\right\|_{L_1(\mathsf{P}_{\mathsf{Z}})}] \leq \beta_n$$

3. ***Estabilidad uniforme*** β_n*, si,* $\forall \mathbf{z}^n \in \mathcal{Z}^n, \forall i \in \{1 : n\}$*:*

$$\mathbb{E}_{\mathsf{Z}^n} \sup_{\mathbf{z} \in \mathcal{Z}}[\left|\ell(A_n, \mathbf{z}) - \ell(A_n^{(i)}, \mathbf{z})\right|] = \mathbb{E}_{\mathsf{Z}^n}[\left\|\ell(A_n, \cdot) - \ell(A_n^{(i)}, \cdot)\right\|_{L_\infty(\mathsf{P}_{\mathsf{Z}})}] \leq \beta_n$$

4. ***Estabilidad en riesgo*** β_n*, si,* $\forall \mathbf{z}^n \in \mathcal{Z}^n, \forall i \in \{1 : n\}$*:*

$$\left|\mathbb{E}_{\mathsf{Z}}[\ell(A_n, \mathsf{Z})] - \mathbb{E}_{\mathsf{Z}}[\ell(A_n^{(i)}, \mathsf{Z})]\right| = \left|R(A_n) - R(A_n^{(i)})\right| \leq \beta_n$$

*En cada uno de los casos anteriores, un algoritmo se dice **estable**, si* $\beta_n = O(\frac{1}{n})$.

En lo que sigue se supone que la estabilidad β_n no es creciente en n. Es claro que la estabilidad uniforme β_n implica que, $\forall \mathbf{z}^n \in \mathcal{Z}^n$ y $\forall \mathbf{z}_i', \mathbf{z}$ independientes extraídas de P_{Z}:

$$\left|\ell(A_n, \mathbf{z}) - \ell(A_n^i, \mathbf{z})\right| \leq \left|\ell(A_n, \mathbf{z}) - \ell(A_n^{(i)}, \mathbf{z})\right| + \left|\ell(A_n^i, \mathbf{z}) - \ell(A_n^{(i)}, \mathbf{z})\right| \leq 2\beta_n$$

[34] El problema de clasificación sigue la misma filosofía. El caso de hipótesis con valores reales es tratado por [Bousquet y Elisseeff 2002].

Se verifican los siguientes resultados, donde $\text{LOO}(A_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(A_n^{(i)}, \mathbf{z}_i)$:

Lema 21 *[Bousquet y Elisseeff 2002, lema 7] Para cualquier algoritmo A simétrico, $\forall i \in \{1 : n\}$, se cumple que:*

$$\begin{aligned} &1) \quad \mathbb{E}_{\mathsf{Z}^n}[R(A_n) - \hat{R}_n(A_n)] = \mathbb{E}_{\mathsf{Z}^n,\mathsf{Z}_i'}[\ell(A_n, \mathsf{Z}_i') - \ell(A_n^i, \mathsf{Z}_i')] \\ &2) \quad \mathbb{E}_{\mathsf{Z}^n}[R(A_n^{(i)}) - LOO(A_n)] = 0 \\ &3) \quad \mathbb{E}_{\mathsf{Z}^n}[R(A_n) - LOO(A_n)] = \mathbb{E}_{\mathsf{Z}^n,\mathsf{Z}}[\ell(A_n, \mathsf{Z}) - \ell(A_n^{(i)}, \mathsf{Z})] \end{aligned}$$

Teorema 22 *[Bousquet y Elisseeff 2002, th. 11] Sea A un algoritmo con estabilidad $\beta_{n,1}$ y estabilidad punto a punto $\beta_{n,2}$, ambas con respecto a una pérdida ℓ acotada en $[0, C]$, entonces, con probabilidad al menos $1 - \eta$:*

$$\begin{aligned} &1) \quad R(A_n) \leq \hat{R}_n(A_n) + \sqrt{\frac{C^2 + 12Cn\beta_{n,2}}{2n\eta}} \\ &2) \quad R(A_n) \leq LOO(A_n) + \sqrt{\frac{C^2 + 6Cn\beta_{n,1}}{2n\eta}} \end{aligned}$$

De este teorema se deduce que, para algoritmos estables, el riesgo empírico y el estimador LOO del riesgo se comportan de forma muy similar como estimadores del verdadero riesgo, lo que lleva a pensar que estos algoritmos tienden a evitar que el riesgo empírico posea un optimismo excesivo.

Para el problema de regresión se tiene el siguiente resultado:

Teorema 23 *[Bousquet y Elisseeff 2002, th. 12] En un problema de regresión, si A es un algoritmo con estabilidad uniforme β con respecto a la pérdida ℓ acotada en $[0, C]$. Entonces, $\forall n \geq 1$ y $\forall \eta \in (0, 1)$, se cumple con probabilidad al menos $1 - \eta$:*

$$\begin{aligned} &1) \quad R(A_n) \leq \hat{R}_n(A_n) + 2\beta + (4n\beta + C)\sqrt{\frac{\ln 1/\eta)}{2n}} \\ &2) \quad R(A_n) \leq LOO(A_n) + \beta + (4n\beta + C)\sqrt{\frac{\ln 1/\eta}{2n}}) \end{aligned} \tag{I.44}$$

Las cotas anteriores trasladan el problema de la determinación de la complejidad de la familia $\mathcal{F}$ (de la $\text{VCdim}(\mathcal{F})$) en las cotas de la SLT o el aprendizaje PAC, a la determinación de la estabilidad β del algoritmo concreto que estemos utilizando.

Para un problema de regresión regularizado, se verifica el teorema general que presentamos a continuación y que resultará de interés en el capítulo siguiente. Para ello, definimos antes los importantes conceptos siguientes (Anexo B):

Definición 24 ***Núcleo Reproductor y Espacio de Hilbert con Núcleo Reproductor (RKHS)**. Sea $\mathcal{H}$ un espacio de Hilbert de funciones $f : \mathcal{U} \to \mathbb{R}$ con producto interno $< \cdot, \cdot >$. Una función $k : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ se denomina núcleo reproductor de $\mathcal{H}$, y a $\mathcal{H}$ se le denomina espacio de Hilbert con Núcleo Reproductor (RKHS) k, si:*

$$\begin{aligned} &1) \quad \forall s \in \mathcal{U},\ k(\cdot, s) \doteq k_s \in \mathcal{H} \\ &2) \quad \forall t \in \mathcal{U}, \forall f \in \mathcal{H},\ f(t) =< f, k_t > \end{aligned}$$

En lo que sigue, $\|\cdot\|_{\mathcal{F}}$ denota la norma de un espacio normado $\mathcal{F}$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Teorema 25 *[Bousquet y Elisseeff 2002, th. 22] Sea $\mathcal{F} = \mathcal{H}$ un espacio de Hilbert con núcleo reproductor k, (RKHS), tal que $\forall \mathbf{x} \in \mathcal{X}$, $k(\mathbf{x}, \mathbf{x}) \leq \kappa^2 < \infty$ y sea ℓ una pérdida γ-**admisible**. Entonces el algoritmo A definido por:*

$$A_n = A(\mathbf{z}^n) = \arg \min_{f \in \mathcal{H}} \{ \frac{1}{n} \sum_{i=1}^{n} \ell(f, \mathbf{z}_i) + \lambda \|f\|_{\mathcal{H}}^2 \} \tag{I.45}$$

tiene estabilidad uniforme con respecto a ℓ:

$$\beta \leq \frac{\gamma^2 \kappa^2}{2\lambda n}$$

Para un problema de regularización en un RKHS del tipo (I.45), el lema siguiente aporta una condición suficiente para que una pérdida sea γ-admisible.

Lema 26 *[Bousquet y Elisseeff 2002, lema 23] Sea $\ell(y, \hat{f}_n(\mathbf{x})) \equiv \ell(A_n, \mathbf{z})$ una función de pérdida convexa como función de $(y, f(\mathbf{x}))$ y con valores positivos. Sea ξ una función positiva y no decreciente con valores reales tal que:*

$$\forall \hat{y} \in \widehat{\mathcal{Y}} : \forall y \in \mathcal{Y},\ \ell(y, \hat{y}) \leq \xi(\hat{y})$$

Entonces, en el marco del problema (I.45) con $k(\mathbf{x}, \mathbf{x}) \leq \kappa^2 < \infty$, $\forall \mathbf{x} \in \mathcal{X}$, para cualquier conjunto de entrenamiento $\{\mathbf{z}_i\}_{i=1}^n$ con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, se verifica que $\|f\|_{\mathcal{H}}^2 \leq \xi(0)/\lambda$ y que:

$$\forall \mathbf{z} = (\mathbf{x}, y) \in \mathcal{Z} : 0 \leq \ell(y, \hat{f}_n(\mathbf{x})) \leq \xi_{\kappa,\lambda} \doteq \xi(\kappa \sqrt{\xi(0)/\lambda}) \tag{I.46}$$

Además, ℓ es γ-admisible donde γ puede tomarse como:

$$\gamma = \sup_{y \in \mathcal{Y}} \sup_{|u| \leq \xi_{\kappa,\lambda}} \left| \frac{\partial \ell}{\partial u}(y, u) \right|$$

Si se utiliza por ejemplo la pérdida cuadrática $\ell(y, f(\mathbf{x})) = [y - f(\mathbf{x})]^2$, y si $\widehat{\mathcal{Y}} \subset \mathcal{Y} \subset \mathbb{R}$ está acotado, utilizando (I.46) en las condiciones anteriores se obtiene que $\ell(y, \hat{f}_n(\mathbf{x})) \leq C$ para algún C. Con ello, puede aplicarse el teorema 23 para seleccionar el parámetro regularizador λ (selección del modelo) con mejor riesgo garantizado en cualquiera de las dos formas de (I.44).

I.6 Consistencia de los Algoritmos

En principio, para una familia fija $\mathcal{F}$, si se verifica que:

$$\sum_{n=1}^{\infty} \eta(n, \varepsilon, v) < \infty,\ \forall \varepsilon > 0 \tag{I.47}$$

en la ecuación (I.9), entonces, la aplicación del lema de Borel-Cantelli (lemas 214 y 215 del Anexo G) en el espacio probabilístico[35] $(\mathcal{Z}^\infty, \mathcal{A}_{\mathsf{Z}}^\infty, \mathbb{P}_{\mathsf{Z}}^\infty)$, proporciona la consistencia universal casi seguro del proceso de aprendizaje.

Si la familia $\mathcal{F}$ incluye la verdadera función $\mathtt{f}$, o es densa en alguna familia $\mathcal{T}$ a la que esta función pertenece (realizabilidad), obtendríamos la convergencia a la verdadera función (si el mínimo del riesgo es único) según el criterio determinado por la pérdida ℓ. Y, en caso de no

[35] Véase el significado de $\mathbb{P}_{\mathsf{Z}}^\infty$ en el Anexo H.

realizabilidad (mala especificación), obtendríamos al menos la convergencia a la función óptima en $\mathcal{F}$, (en cuyo caso deberíamos elegir una familia lo más compleja posible, aunque siempre con la restricción $\mathrm{VCdim}(\mathcal{F}) < \infty$ (o medida equivalente) y, a ser posible, (I.47).

El problema es que, para lograr la realizabilidad, en muchos problemas de aplicación se requiere una familia $\mathcal{F}$ suficientemente compleja, (a veces, incluso, con $\mathrm{VCdim}(\mathcal{F}) = \infty$), que, con una muestra finita de tamaño n, probablemente producirá sobreajuste.

Este hecho, motivó la estrategia de partir de una familia marco $\mathcal{F}$ con suficiente complejidad y, utilizando una estructura anidada de subfamilias y con técnicas de selección del modelo, acompasar el nivel de complejidad al tamaño de la muestra disponible [Geman *et al.* 1992]. Ello provoca que la familia realmente utilizada no sea fija, sino variable $\mathcal{F}_{t(n)}$ en función del tamaño muestral, a través de una regla $t(n)$ a determinar por el algoritmo.

Por tanto, en la práctica, no son útiles los resultados de consistencia para una familia de hipótesis $\mathcal{F}$ prefijada (suponiendo disponible un algoritmo que minimiza el riesgo empírico en $\mathcal{F}$), sino que es necesario estudiar la consistencia del *algoritmo completo*, entendiendo éste como:

[minimización del riesgo empírico + selección del modelo]

Así, suponiendo que se dispone de un algoritmo que minimiza el riesgo empírico (u otro principio inductivo elegido) y que la familia $\mathcal{F}$ es suficientemente compleja, la esencia de la consistencia descansa en el método de selección del modelo implementado por aquél.

En lo que sigue, exponemos someramente las propiedades de consistencia de los métodos de selección del modelo presentados anteriormente.

I.6.1 Optimalidad de los Criterios Clásicos

Básicamente, los estudios sobre el comportamiento de los criterios clásicos se han realizado suponiendo el modelo de regresión lineal, la mayor parte de las veces con ruido gausiano. Dichos estudios han seguido diferentes baremos de optimalidad y, en general, puede clasificarse en: los que evalúan el grado de aproximación asintótica a algún funcional de riesgo y los que evalúan la probabilidad asintótica de seleccionar el modelo correcto.

Entre los resultados más relevantes, citamos los siguientes en los que, tratando con modelos paramétricos, identificamos el estimador de los parámetros $\hat{\boldsymbol{\theta}}_n$ con la hipótesis seleccionada $\hat{f}_n$ a partir de una muestra aleatoria de datos $\mathbf{z}^n$:

1. La optimalidad asintótica [Shibata 1981] del criterio SMS para modelos lineales, observaciones independientes y errores gausianos en el sentido:

$$\operatorname*{plim}_{n\to\infty} \frac{\mathrm{ASE}(\hat{\boldsymbol{\theta}}_n; \mathbf{x}^n)}{\mathrm{AMSE}(\boldsymbol{\theta}_n^*)} = 1$$

   donde $\boldsymbol{\theta}_n^* = \arg\min_{\Theta} \mathrm{AMSE}_n(\boldsymbol{\theta})$ y $\hat{\boldsymbol{\theta}}_n = \arg\min_{\Theta_n} \mathrm{SMS}(\boldsymbol{\theta})$.

2. La optimalidad asintótica [Li 1987] de los criterios C_p, LOO y GCV para modelos lineales ($\hat{\mathbf{f}}_n = \mathbf{S}\mathbf{y}$), con observaciones independientes y ruido de media cero, con el criterio:

$$\operatorname*{plim}_{n\to\infty} \frac{\mathrm{ASE}(\hat{\boldsymbol{\theta}}_n; \mathbf{x}^n)}{\inf_{\boldsymbol{\theta}\in\Theta} \mathrm{ASE}(\boldsymbol{\theta}; \mathbf{x}^n)} = 1$$

   donde $\hat{\boldsymbol{\theta}}_n$ es el óptimo según dichos métodos de selección del modelo.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3. La consistencia del criterio BIC al estar basado en la distribución a posteriori asintótica (para familias exponenciales), ([Schwarz 1978], [Gelfand y Dey 1994]).

4. Considerando el modelo de regresión lineal con observaciones independientes y ruido homocedástico con varianza σ^2 y sin otras hipótesis sobre la distribución, si en el criterio general de selección:

$$\mathrm{RW}(p) = n\hat{R}_n + p\sigma^2 C_n \equiv \hat{R}_n + \frac{p}{n}\sigma^2 C_n$$

la penalización C_n se elige tal que $n^{-1}C_n \to 0$ y $(\ln \ln n)^{-1}C_n \to \infty$, entonces, el número de parámetros del modelo seleccionado por dicho criterio, es estimador fuertemente consistente del número de parámetros del modelo verdadero [Rao y Wu 1989].

Sin embargo, los resultados no son tan favorables para algunos criterios si se considera la probabilidad de elegir el modelo verdadero:

1. LOO es estimador consistente de AMPSE sin embargo, no selecciona el modelo óptimo con probabilidad 1 astintóticamente. Por el contrario, el criterio LrO además de ser consistente de AMPSE, sí selecciona el modelo óptimo con probabilidad 1 asintóticamente, ([Shao 1993], [Shao 1997]).

2. [Zhang 1992] muestra que, para modelos lineales con ruido gausiano y observaciones independientes, los criterios de la forma $\hat{R}_n + \lambda p/n$ (se supone $\sigma^2 = 1$) con $\lambda = 2$, que es el caso de casi todos los clásicos (AIC, FPE, SMS, LOO, etc.), sobreajustan. Además, su probabilidad asintótica de elegir el modelo correcto es 0.7117 y la esperanza del número de parámetros elegido es, asintóticamente, el número de parámetros del modelo verdadero, más uno.

3. [Shibata 1976] deriva para modelos autorregresivos la distribución asintótica del criterio AIC mostrando que este criterio tiende a sobreestimar el número de parámetros. Sin embargo, la probabilidad de seleccionar asintóticamente un modelo demasiado grande decrece rápidamente con el tamaño del modelo verdadero.

Otros resultados, válidos para modelos no lineales en general o técnicas no lineales específicas, son los siguientes:

1. La consistencia fuerte del criterio LOO para modelos no lineales [Plutowski *et al.* 1994], en el sentido:

$$\max_{1\leq 1\leq n} d(\hat{\boldsymbol{\theta}}_n^{(i)}, \Theta^*) \overset{\text{c.s.}}{\underset{n\to\infty}{\to}} 0$$

donde Θ^* es el conjunto de posibles óptimos del riesgo en Θ y $d(\boldsymbol{\theta}, \Theta^*) = \inf_{\boldsymbol{\theta}^* \in \Theta^*} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|$. Además, el criterio LOO no subestima nunca el riesgo óptimo.

2. La optimalidad asintótica de los criterios C_p, LOO y GCV para la regresión regularizada y *smoothing splines* [Li 1986].

3. La optimalidad asintótica del criterio GCV para *smoothing splines*, con el criterio AMSE [Wahba 1990*b*, p. 55-59].

USC

4. La optimalidad asintótica de la validación cruzada [Härdle y Marron 1985] para la selección del parámetro de suavizado en la regresión tipo núcleo[36] y de los criterios AIC, FPE y SMS [Härdle *et al.* 1988] a partir de su equivalencia asintótica con aquélla.

I.6.2 Consistencia de los Criterios No Asintóticos

Métodos Bayesianos

Para analizar la consistencia de los métodos bayesianos es preciso distinguir entre la selección de los hiperparámetros bajo un mismo modelo y la selección entre modelos diferentes.

Con respecto al primer caso, la consistencia del método de la evidencia está estrechamente relacionada con la consistencia de la distribución a posteriori de los hiperparámetros. Por ejemplo, denotando $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda})$ y obviando $\mathcal{F}$ y $\mathcal{G}$ en la notación, la expresión (I.34) se escribe como:

$$\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n) = \frac{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\mathsf{p}(\mathbf{y}^n)} \propto \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta})\mathsf{p}(\boldsymbol{\theta})$$

donde la evidencia, aplicando la ley fuerte de los grandes números, satisface, (p. ej. [O'Hagan 1994], [Diaconis y Freedman 1986]):

$$\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathsf{p}(y_i|\boldsymbol{\theta}) = \exp\{n\frac{1}{n}\sum_{i=1}^{n} \ln \mathsf{p}(y_i|\boldsymbol{\theta})\} \underset{n\to\infty}{\overset{\text{c.s.}}{\to}} \exp\{nE_{\mathsf{Y}/\mathbf{x}}[\ln \mathsf{p}(y|\boldsymbol{\theta})]\}$$

Pero, si $\boldsymbol{\theta}^*$ es el verdadero valor del parámetro y escribiendo $\frac{1}{n}\ln \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ln \mathsf{p}(y_i|\boldsymbol{\theta})$, se tiene:

$$\frac{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta}^*)}{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta})} = \frac{\exp\{n\frac{1}{n}\ln \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta}^*)\}}{\exp\{n\frac{1}{n}\ln \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta})\}} = \left[\frac{\exp\{\frac{1}{n}\ln \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta}^*)\}}{\exp\{\frac{1}{n}\ln \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta})\}}\right]^n \underset{n\to\infty}{\overset{\text{c.s.}}{\to}} \infty$$

ya que $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ maximiza la entropía cruzada $\mathbb{E}_{\mathsf{Z}}[\ln \mathsf{p}(\mathbf{z}|\boldsymbol{\theta})] = \int_{\mathcal{Z}} \ln \mathsf{p}(\mathbf{z}|\boldsymbol{\theta})\mathsf{p}(\mathbf{z}|\boldsymbol{\theta}^*)d\mathbf{z}$. Por tanto, la probabilidad de seleccionar $\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}^*$ es asintóticamente uno.

Para realizar la comparación entre modelos, (en un número finito), la consistencia del método de la evidencia se reduce a la consistencia del factor de Bayes (*Bayes Factor*), (p. ej. [Gelfand y Dey 1994], [O'Hagan 1994]). Por ejemplo, haciendo $\mathcal{M} = (\mathcal{F}, \mathcal{G})$ en la sección anterior (pág. 47):

$$\frac{\mathsf{p}(\mathcal{M}_1|\mathbf{y}^n)}{\mathsf{p}(\mathcal{M}_2|\mathbf{y}^n)} = \frac{\mathsf{p}(\mathbf{y}^n|\mathcal{M}_1)}{\mathsf{p}(\mathbf{y}^n|\mathcal{M}_2)} \times \frac{\mathsf{p}(\mathcal{M}_1)}{\mathsf{p}(\mathcal{M}_2)} = \text{Bayes Factor} \times \frac{\mathsf{p}(\mathcal{M}_1)}{\mathsf{p}(\mathcal{M}_2)}$$

En el caso de situación inicial no informativa, el cociente de las distribuciones a priori $\mathsf{p}(\mathcal{M}_i)$, $i = 1, 2$, es 1, y el método de la evidencia se reduce al del factor de Bayes, que es consistente, (p. ej. [Gelfand y Dey 1994], [O'Hagan 1994]), aunque sólo válido para un número finito de modelos.

[36] En el sentido de [Shibata 1981]: $\lim_{n\to\infty} \frac{\rho(\hat{f},m)}{\inf_{\mathcal{F}_n} \rho(\hat{f},m)} = 1$, donde $\rho(\hat{f}, m)$ puede ser: el MSE, $\mathbb{E}_{\mathsf{X}}[(\hat{f}(\mathsf{X}) - m(\mathsf{X}))^2]$ y el *average squared error* $\frac{1}{n}\sum_{i=1}^{n}[\hat{f}_n(\mathbf{x}_i) - m(\mathbf{x}_i)]^2 w(x_i)$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Minimización del Riesgo Estructural

Dado n, la aplicación de los criterios no asintóticos proporciona una regla $t = t(n)$ que determina una estructura anidada de familias $\mathcal{F}_{t(n)}$. Sean:

$$\begin{aligned} f^*_{t(n)} &= \arg \inf_{\mathcal{F}_{t(n)}} R(f) \\ \hat{f}_{t(n)} &= \arg \inf_{\mathcal{F}_{t(n)}} \hat{R}_n(f) \end{aligned}$$

En este marco, la discrepancia entre el riesgo cometido y el riesgo óptimo puede descomponerse como:

$$\begin{aligned} R(\hat{f}_{t(n)}) - R(\mathsf{f}) &= [R(\hat{f}_{t(n)}) - R(f^*_{t(n)})] + [R(f^*_{t(n)}) - R(\mathsf{f})] \\ &\quad \textbf{error estimación} + \textbf{error aproximación} \end{aligned} \tag{I.48}$$

donde el término $R(\mathsf{f})$ puede leerse como el riesgo inevitable en el problema.

En un problema de regresión con pérdida cuadrática, $R(\mathsf{f}) = \mathbb{E}_{\mathsf{Z}}[\ell(\mathsf{Y}, \mathsf{f}(\mathsf{X}))] = \mathbb{E}_{\mathsf{X}}[\sigma^2(\mathsf{X})]$ y, si el ruido es homocedástico, $R(\mathsf{f}) = \sigma^2$ sería la varianza del ruido. Asimismo, debido a la descomposición (I.14), el error de aproximación sería:

$$R(f^*_{t(n)}) - R(\mathsf{f}) = \mathbb{E}_{\mathsf{X}}\{[f^*_{t(n)}(\mathsf{X}) - \mathsf{f}(\mathsf{X})]^2\} = \left\| f^*_{t(n)} - \mathsf{f} \right\|^2_{L_2(\mathsf{P}_{\mathsf{X}})}$$

donde $L_p(\mathsf{P}_{\mathsf{X}})$ con $p = 2$ es el espacio de las funciones de cuadrado integrable con la medida de Lebesgue-Stieljes inducida por P_{X} (Anexo C).

La expresión (I.48) refleja de nuevo las dos componentes del riesgo: las limitaciones de la familia elegida para representar la verdadera función representadas por el error de aproximación, (análogo al sesgo), y la sensibilidad de la familia ante la aleatoriedad de los datos, (la segunda fuente de error, análoga a la varianza del estimador).

El error de estimación refleja el coste asociado a elegir $\hat{f}_{t(n)}$ cuando la mejor función en la familia $\mathcal{F}_{t(n)}$ es $f^*_{t(n)}$, debido al limitado número de datos. El segundo término recoge el coste asociado a la aproximación $f^*_{t(n)}$ a la verdadera función f debido a la utilización de un número finito de parámetros (teniendo en cuenta que f está determinada por un número infinito de parámetros).

En la teoría estadística de tipo paramétrico, (p. ej. [Seber y Wild 1989]), los estudios de consistencia se centran en la convergencia a cero del error de estimación, es decir, de la diferencia entre el riesgo del estimador que minimiza un criterio empírico (con una complejidad predeterminada –cierto número de parámetros) y el riesgo asociado al mejor modelo de igual complejidad (que suele suponerse verdadero).

En contraste, la teoría estadística no paramétrica trata la totalidad del error de generalización, descomponiéndolo en un término de sesgo y un término de varianza, para clases de funciones con determinadas condiciones de suavidad (y, en general, de moderada dimensionalidad), (p. ej. [Eubank 1988], [Härdle 1990]).

En la teoría estadística del aprendizaje, el enfoque seguido en la mayoría de los resultados de consistencia, se basa en demostrar que, tanto el error de aproximación como el error de estimación, tienden a cero.

Con respecto al error de estimación, tenemos el siguiente lema, para el que, por su brevedad, aportamos la demostración:

USC UNIVERSIDADE DE COMPOSTELA

Lema 27 *[Devroye 1988], [Lugosi y Zeger 1995], [Niyogi y Girosi 1996], [Devroye* et al. *1996, th. 8.4], Para el error de estimación se cumple que:*

$$R(\hat{f}_{t(n)}) - R(f^*_{t(n)}) \leq 2 \sup_{f \in \mathcal{F}_{t(n)}} \left| R(f) - \hat{R}_n(f) \right|$$

Demostración.

$$\begin{aligned} R(\hat{f}_{t(n)}) - R(f^*_{t(n)}) &= [R(\hat{f}_{t(n)}) - \hat{R}_n(\hat{f}_{t(n)})] + [\hat{R}_n(\hat{f}_{t(n)}) - R(f^*_{t(n)})] \\ &\leq [R(\hat{f}_{t(n)}) - \hat{R}_n(\hat{f}_{t(n)})] + \sup_{f \in \mathcal{F}_{t(n)}} \left| R(f) - \hat{R}_n(f) \right| \\ &\leq 2 \sup_{f \in \mathcal{F}_{t(n)}} \left| R(f) - \hat{R}_n(f) \right| \end{aligned}$$

■

Con respecto al error de aproximación, si por ejemplo, suponemos[37] $R(f) = \|\mathsf{Y} - f(\mathsf{X})\|_{L_p(\mathsf{P_Z})}$ y denotamos $f_{t(n)} = \arg\min_{\mathcal{F}_{t(n)}} \|f - \mathsf{f}\|_{L_p(\mathsf{P_Z})}$, entonces, aplicando la desigualdad triangular [Lugosi y Zeger 1995], se tiene:

$$\begin{aligned} R(f^*_{t(n)}) - R(\mathsf{f}) &\leq R(f_{t(n)}) - R(\mathsf{f}) = \left\| y - f_{t(n)} \right\|_{L_p(\mathsf{P_Z})} - \|y - \mathsf{f}\|_{L_p(\mathsf{P_Z})} \\ &\leq \left\| f_{t(n)} - \mathsf{f} \right\|_{L_p(\mathsf{P_Z})} \end{aligned}$$

De esta forma, la descomposición (I.48) puede acotarse como:

$$R(\hat{f}_{t(n)}) - R(\mathsf{f}) \leq 2 \sup_{f \in \mathcal{F}_{t(n)}} \left| R(f) - \hat{R}_n(f) \right| + \left\| f_{t(n)} - \mathsf{f} \right\|_{L_p(\mathsf{P_Z})} \tag{I.49}$$

con lo que, para obtener la consistencia fuerte universal (*distribution-free*) con el riesgo indicado, es suficiente conseguir las dos siguientes condiciones $\forall \mathbb{P}_\mathsf{Z} \in \mathcal{P}$:

$$\begin{cases} 1) & \mathbb{P}^\infty_\mathsf{Z}\{ \lim_{n\to\infty} \sup_{f \in \mathcal{F}_{t(n)}} \left| R(f) - \hat{R}_n(f) \right| = 0\} = 1 \\ 2) & \lim_{n\to\infty} \inf_{\mathcal{F}_{t(n)}} \|f - \mathsf{f}\|_{L_p(\mathsf{P_Z})} = 0 \end{cases} \tag{I.50}$$

Las condiciones (I.50) reflejan un esquema general a seguir en los estudios de consistencia de cualquier estimador obtenido mediante el principio de la minimización del riesgo estructural [Vapnik 1998, ths. 6.2 y 6.3] y, en particular, de las redes neuronales ([Devroye 1988], [Faragó y Lugosi 1993], [Lugosi y Zeger 1995], [Niyogi y Girosi 1996], [Lugosi y Zeger 1996]).

La segunda condición en (I.50), garantiza que el modelo va adquiriendo a medida que crece el tamaño muestral, la necesaria complejidad para representar la verdadera función. Ello se consigue por ejemplo si las familias $\mathcal{F}_t$ están anidadas: $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_t \subset \cdots$ tal que $\cup_t \mathcal{F}_t$ es denso en $L_p(\mathsf{P_Z})$, $\forall \mathsf{P_Z}$.

Por su parte, la primera condición, garantiza que el algoritmo realiza un efectivo control de la complejidad limitando la sensibilidad del estimador ante la aleatoriedad muestral, lo cuál, ha de ser comprobado con cada algoritmo específico.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[37] Si $p = 2$, $\mathsf{f} = \mathbb{E}[\mathsf{Y}|\mathbf{x}]$ es la función de regresión (la probabilidad a posteriori $\mathbb{P}[\mathsf{Y} = 1|\mathbf{x}]$ en un problema de clasificación), y si $p = 1$ en un problema de clasificación la regla de Bayes es hipótesis óptima.

De cualquier forma, se tienen los siguientes resultados generales sobre la consistencia del principio SRM, que utilizan el esquema (I.48) de descomposición del riesgo, bajo las condiciones (I.35) de estructura admisible vistas anteriormente.

El primer resultado, aporta condiciones para la existencia de reglas $t(n)$ de selección del modelo que resultan consistentes al margen del principio de selección del modelo empleado, es decir, que provocan que $R(\hat{f}_{t(n)}) \to \inf_{f\in\mathcal{F}} R(f)$ cuando $n \to \infty$. En lo que sigue, $\mathcal{F}_t$ son familias de funciones reales y $\mathrm{VCdim}(\mathcal{F}_t)$ es su dimensión VC que generaliza el mismo concepto para funciones indicadoras.

Teorema 28 *[Vapnik 1998, th. 6.2] Sea $t(n)$ una regla de selección de la clase $\mathcal{F}_t$ ($\longleftrightarrow \mathcal{L}_t$) para una muestra aleatoria simple de tamaño n y sean $f^*_{t(n)}$ y $\hat{f}_{t(n)}$ los óptimos del riesgo y del riesgo empírico, respectivamente, sobre la familia $\mathcal{F}_{t(n)}$ seleccionada por la regla $t(n)$. Sea $v_t = \mathrm{VCdim}(\mathcal{F}_t)$. Si se verifica que:*

$$\lim_{n\to\infty} t(n) = \infty \; y \; \lim_{n\to\infty} \frac{b^2_{t(n)} v_{t(n)} \ln n}{n} \to 0 \tag{I.51}$$

siendo b_t la cota de la segunda condición de estructura admisible (I.35), entonces, para cualquier distribución $\mathsf{P_Z}$, se verifica que:

$$\lim_{n\to\infty} R(\hat{f}_{t(n)}) = \inf_{f\in\mathcal{F}} R(f) \; c.s.$$

con tasa de convergencia $\alpha(n)$, es decir, $\mathbb{P}_{\mathsf{Z}^\infty}\{\alpha(n)^{-1}[R(\hat{f}_{t(n)}) - \inf_{f\in\mathcal{F}} R(f)]\} = 1$, dada por:

$$\alpha(n) = R(f_{t(n)}) - \inf_{f\in\mathcal{F}} R(f) + \sqrt{\frac{b^2_{t(n)} v_{t(n)} \ln n}{n}}$$

El método SRM se concibió [Vapnik 1982] con el criterio (I.36), pero el teorema anterior indica que las condiciones (I.50) garantizan la consistencia del método independientemente del método de selección del modelo que se utilice (que denominaremos igualmente método SRM), siempre que verifique las condiciones (I.51).

El siguiente teorema de [Vapnik 1998], que generaliza otro de Lugosi y Zeger [Devroye *et al.* 1996, ths. 18.2 y 18.3] para funciones indicadoras, establece la consistencia fuerte universal del principio SRM.

Teorema 29 *[Vapnik 1998, th. 6.3] Sea $t(n)$ la regla proporcionada por el principio SRM. Si las cotas superiores b_t de las funciones $\ell_f \in \mathcal{L}_t$ en (I.35), satisfacen $b_t^2 \le t^{1-\delta}$ para algún $\delta > 0$, entonces, para cualquier distribución $\mathsf{P_Z}$:*

$$\lim_{n\to\infty} R(\hat{f}_{t(n)}) = \inf_{f\in\mathcal{F}} R(f) \; c.s.$$

Además, si el óptimo $\ell^*_f \in \bigcup_{t=1}^{\infty} \mathcal{L}_t$ y $b^2_{t(n)} \le \rho(n) \le n^{1-\delta}$, entonces se obtiene la tasa de convergencia:

$$\alpha(n) = O(\sqrt{\frac{\rho(n)\ln n}{n}})$$

El teorema anterior puede aplicarse a funciones indicadoras tomando $b_t = 1 \; \forall t$, en cuyo caso, $\rho(n) = 1$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En el contexto de las condiciones (I.35) de estructura admisible, los teoremas anteriores exigen que $\mathrm{VCdim}(\mathcal{F}_t) = v_t < \infty$, lo que no se verifica para todas las redes neuronales, y, por ejemplo, que las funciones $\ell_f(\mathbf{z}) = \ell(y, f(\mathbf{x}))$ estén acotadas.

La primera condición requiere un adecuado control de la complejidad y, en el capítulo siguiente, veremos que ello se logra, por ejemplo, restringiendo el tamaño de los parámetros de las funciones de $\mathcal{F}_t$.

La segunda condición, en el caso frecuente $R(f) = \|\mathsf{Y} - f(\mathsf{X})\|_{L_p(\mathsf{P_Z})}$, requiere que

$$\ell(\mathsf{Y}, f(\mathsf{X})) = |\mathsf{Y} - f(\mathsf{X})|^p$$

sea acotada con probabilidad 1 $\forall f \in \mathcal{F}_t$, o lo que es lo mismo:

$$\mathbb{P}_{\mathsf{Z}}[|\mathsf{Y} - f(\mathsf{X})| \leq b_n] = 1, \forall f \in \mathcal{F}_t$$

Pero esta condición pueden ser muy restrictiva, pues es frecuente que Y sea no acotada (por ejemplo, gausiana). El siguiente teorema de [Lugosi y Zeger 1995] aporta una solución a este problema exigiendo únicamente que $\mathbb{E}_{\mathsf{Z}}(|\mathsf{Y}|^p) < \infty$, lo que es una condición perfectamente asumible en la práctica.

Teorema 30 *[Lugosi y Zeger 1995, th. 1] Sea* $R(f) = \|\mathsf{Y} - f(\mathsf{X})\|_{L_p(\mathsf{P_Z})}$. *Si* $\forall \mathsf{P_Z}$ *tal que* $\mathbb{P}_{\mathsf{Z}}[|\mathsf{Y}| \leq b] = 1$ *se verifica que:*

$$\sup_{f \in \mathcal{F}_{t(n)}} \left| R(f) - \hat{R}_n(f) \right| \overset{c.s.}{\underset{n\to\infty}{\to}} 0$$

entonces:

$$R(\hat{f}_n) - \inf_{f \in \mathcal{F}_{t(n)}} R(f_n) \overset{c.s.}{\underset{n\to\infty}{\to}} 0$$

$\forall \mathsf{P_Z}$ *tal que* $\mathbb{E}_{\mathsf{Z}}(|\mathsf{Y}|^p) < \infty$.

El teorema anterior permite demostrar los resultados de consistencia para las distribuciones $\mathsf{P_Z}$ que verifican la hipótesis restrictiva $\mathbb{P}_{\mathsf{Z}}[|\mathsf{Y}| \leq b] = 1$ y obtener como consecuencia la propiedad de consistencia para todas las $\mathsf{P_Z}$ que satisfacen la condición más general $\mathbb{E}_{\mathsf{Z}}(|\mathsf{Y}|^p) < \infty$.

Métodos de Penalización Adaptable a los Datos

Los resultados (I.40), (I.42) y (I.43) obtenidos del teorema 19, permiten deducir la consistencia del método de selección del modelo para cada uno de los estimadores del riesgo allí indicados. En lo que sigue, mostramos el esquema de la demostración en cada caso.

Para concretar, suponemos un problema de clasificación con funciones dicotómicas, (el razonamiento es análogo para funciones reales sustituyendo VCdim por $V_\gamma \dim$), y supongamos una estructura $\{\mathcal{F}_t\}$ de subfamilias tal que $\cup_t \mathcal{F}_t = \mathcal{F}$ con $v_t = \mathrm{VCdim}(\mathcal{F}_t) < \infty \; \forall t$ (o $V_\gamma \dim$ si se trata de funciones reales) y $v_t \leq v_{t'}$ si $t \leq t'$, y $\mathcal{F}$ con complejidad suficiente como para contener la verdadera función f.

En primer lugar, debido a las propiedades de la estructura $\{\mathcal{F}_t\}$, el error de aproximación satisface: $\Delta_t = (\inf_{f \in \mathcal{F}_t} R(f) - \mathrm{R}) \to 0$ para t suficientemente grande. Además:

1. Si se utiliza el estimador basado en una muestra de validación (*hold-out*), con m observaciones $\mathbf{z}'_i = (\mathbf{x}'_i, y'_i)$ extraídas independientemente, $\hat{R}'_n(\hat{f}_t) = \frac{1}{m} \sum_{i=1}^m \ell(y'_i, \hat{f}_t(\mathbf{x}'_i))$, se

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

obtenía la cota (I.40):

$$\mathbb{E}_{\mathsf{Z}^n}[R(f_n)] - \mathrm{R} \leq \min_t\{\mathbb{E}_{\mathsf{Z}^n}[R(\hat{f}_t) - \hat{R}_n(\hat{f}_t)] + \Delta_t + \sqrt{\frac{\ln t}{m}}\} + \frac{1}{\sqrt{2m}}$$

El término $\mathbb{E}_{\mathsf{Z}^n}[R(\hat{f}_t) - \hat{R}_n(\hat{f}_t)]$ tiende a cero $\forall t$ cuando $n \to \infty$ debido a que $\mathrm{VCdim}(\mathcal{F}_t) < \infty\ \forall t$, con ello, con n suficientemente grande basta tomar un t lo suficientemente grande para que Δ_t sea suficientemente pequeño. De esta forma, si $m \to \infty$, se tiene la consistencia universal.

2. Para el estimador basado en el coeficiente de disección empírico, se obtenía la cota (I.42):

$$\begin{aligned}\mathbb{E}_{\mathsf{Z}^n}[R(f_n)] - \mathrm{R} \quad \leq \quad & \min_t\{\sqrt{\frac{12E_{\mathsf{X}^n}\ln N_t(\mathsf{X}^n) + \ln 4}{n}} \\ & + \Delta_t + 8{,}95\sqrt{\frac{\ln t}{n}}\} + \frac{8{,}23}{\sqrt{n}}\end{aligned}$$

Pero, utilizando la ecuación (I.8), el término $\frac{1}{n}\mathbb{E}_{\mathsf{X}^n}\ln N_{\mathcal{F}_t}(\mathsf{X}^n)$ puede acotarse como sigue:

$$\frac{1}{n}\mathbb{E}_{\mathsf{X}^n}\ln N_{\mathcal{F}_t}(\mathsf{X}^n) \leq \frac{1}{n}\mathbb{E}_{\mathsf{X}^n}\ln\sup N_{\mathcal{F}_t}(\mathsf{X}^n) \leq \frac{1}{n}G_{\mathcal{F}_t}(n) \leq \frac{1}{n}v_t(1 + \ln\frac{n}{v_t}) \underset{n\to\infty}{\to} 0$$

donde $G_{\mathcal{F}_t}(n)$ es la *growth function* de $\mathcal{F}_t$. Con ello, cuando $n \to \infty$, se tiene la consistencia universal.

3. Para el estimador basado en la máxima discrepancia en una muestra, se obtenía la cota (I.43):

$$\begin{aligned}\mathbb{E}_{\mathsf{Z}^n}[R(f_n)] - \mathrm{R} \quad \leq \quad & \min_t\{\mathbb{E}_{\mathsf{Z}^n}[\max_{f\in\mathcal{F}_t}[\hat{R}_n^{(1)}(f) - \hat{R}_n^{(2)}(f)] \\ & + \Delta_t + 4{,}59\sqrt{\frac{\ln t}{n}}\} + \frac{4{,}7}{\sqrt{n}}\end{aligned}$$

donde $\hat{R}_n^{(1)}(f) = \frac{2}{n}\sum_{i=1}^{n/2}\ell(y_i, f(\mathbf{x}_i))$ y análogamente $\hat{R}_n^{(2)}(f)$ en la segunda mitad de la muestra. Pero se verifica que [Vapnik *et al.* 1994, p. 858]:

$$\mathbb{E}_{\mathsf{Z}^n}\{\max_{f\in\mathcal{F}_t}[\hat{R}_n^{(1)}(f) - \hat{R}_n^{(2)}(f)]\} \leq C\sqrt{\frac{2\ln(n/v_t)}{n/v_t}}$$

donde $C > 0$ es una constante. Con ello, al igual que en los casos anteriores, cuando $n \to \infty$ se obtiene la consistencia universal del correspondiente método de selección del modelo.

Métodos de Estabilidad Algorítmica

Tomemos como estimador el riesgo empírico $\hat{R}_n(A_n)$, (el argumento para $\mathrm{LOO}(A_n)$ es análogo). El teorema 23 producía con probabilidad $1-\eta$ y $\forall n \geq 1$:

$$\begin{aligned}&1)\quad R(A_n) \leq \hat{R}_n(A_n) + 2\beta + (4n\beta + C)\sqrt{\tfrac{\ln 1/\eta}{2n}}) \\ &2)\quad R(A_n) \leq \mathrm{LOO}(A_n) + \beta + (4n\beta + C)\sqrt{\tfrac{\ln 1/\eta}{2n}})\end{aligned} \tag{I.52}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

para un algoritmo A con estabilidad uniforme β y pérdida acotada en $[0, C]$. Además, el teorema 25 acotaba la estabilidad uniforme para la regresión regularizada:

$$\beta \leq \frac{\gamma^2 \kappa^2}{2\lambda n} \tag{I.53}$$

donde $\forall \mathbf{x} \in \mathcal{X}$, $k(\mathbf{x}, \mathbf{x}) \leq \kappa^2 < \infty$ y γ era la constante de la pérdida γ-admisible.

Combinando los resultados (I.52) y (I.53), se obtiene, por ejemplo para el error de estimación utilizando el riesgo empírico, (el argumento es análogo si se utiliza LOO(A_n)):

$$R(A_n) - \hat{R}_n(A_n) \leq \frac{\gamma^2 \kappa^2}{\lambda n} + (\frac{2\gamma^2 \kappa^2}{\lambda} + C)\sqrt{\frac{\ln 1/\eta}{2n}} \underset{n\to\infty}{\to} 0 \tag{I.54}$$

siempre que se elija $\lambda = \lambda(n)$ tal que $\lim \sqrt{n}\lambda(n) = \infty$.

Sin embargo, la velocidad óptima para el método es un problema abierto todavía, (véase discusión en [Bousquet y Elisseeff 2001] y [Evgeniou *et al.* 2000, p. 40] que sugiere $\lambda(n) \approx \ln(\ln n)$). Suponiendo que tal condición se satisface, queda por satisfacer la convergencia a cero del error de aproximación, pero esto se consigue si $\mathcal{F}$ es una familia suficientemente compleja como para contener la verdadera función y, además, $\lambda = \lambda(n)$ es tal que $\lambda(n) \to 0$ cuando $n \to \infty$, pues, con ello, en el límite, se estaría utilizando la totalidad de la familia[38] $\mathcal{F}$. En todo caso, dicha condición debe satisfacerse simultáneamente con la condición $\lim \sqrt{n}\lambda(n) = \infty$ mencionada.

Otros Enfoques sin Criterio de Selección del Modelo

Además de los enfoques anteriores, se han desarrollado otros programas de obtención de consistencia, pero que no han dado lugar a métodos asociados para la selección del modelo, por lo que pueden considerarse únicamente como resultados de interés teórico. Entre éstos, destacamos los siguientes:

1. Los basados en el método de los *Sièves* ([White 1990], [Lugosi y Zeger 1995] y más recientemente [Chen y White 1999]) que se describirán en el capítulo siguiente. Al igual que los dos métodos anteriores, estos métodos están basados en técnicas de la teoría Estadística del Aprendizaje y del aprendizaje PAC, y utilizan cotas basadas en métodos combinatorios, mediante *covering numbers* que dependen de la muestra y que se traducen en cotas basadas en la dimensión VC (o dimensión V_γ) de la familia $\mathcal{F}$, que no está dotada de ninguna propiedad específica.

2. El enfoque de [Cucker y Smale 2001] que consideran el espacio de hipótesis $\mathcal{F}$ dotado de una topología, en concreto, $\mathcal{F} \subset \mathcal{C}(\mathcal{X})$ compacto, donde $\mathcal{C}(\mathcal{X})$ es el conjunto de las funciones continuas dotado con la norma $\|f\|_{L_\infty(\mathcal{X})} = \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|$. Con ello, pueden utilizarse todas las herramientas del análisis funcional clásico y sustituir la dimensión VC como medida de complejidad de $\mathcal{F}$ por el radio de una bola centrada en el origen, que ahora constituye el espacio de hipótesis $\mathcal{F}$.

   Un completo análisis puede verse en [Cucker y Smale 2001], donde en este contexto se dan condiciones para la consistencia universal. Sin embargo, lo reciente de este enfoque no se ha traducido en métodos de selección del modelo que sean aplicables en la práctica.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[38] Nótese que, por el análisis realizado al tratar el problema de regularización, la sucesión $\lambda(n) \to 0$ construye una estructura anidada $\mathcal{F}_{\lambda(n)}$ de creciente complejidad al modo del método SRM.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Capítulo II
Redes Neuronales para Clasificación y Regresión

II.1 Las Redes Neuronales

II.1.1 Breve Reseña Histórica

Las redes neuronales tienen su origen en el ámbito de las ciencias del cerebro, (neuroanatomía, psiquiatría, psicología), como resultado del programa de modelización de la capacidad representativa del cerebro.

Diversas reseñas históricas, (p. ej. [Cowan 1990], [Haykin 1999], y para una recopilación de gran valor histórico de los trabajos originales más relevantes, [Anderson y Rosenfeld 1988]), sitúan en los años cuarenta, con los trabajos del psiquiatra McCulloch y el matemático Pitts [MacCulloch y Pitts 1943], los primeros modelos lógico-matemáticos del funcionamiento de las neuronas que conjugaban la neurofisiología con la lógica matemática. En ellos, se demostraba que con un número suficiente de neuronas y un conjunto de conexiones sinápticas apropiadamente establecidas, una red neuronal podría en principio representar cualquier función computable.

El siguiente avance relevante se produce con [Hebb 1949] que, siguiendo las intuiciones de Ramón y Cajal, emite su *postulado del aprendizaje*, estableciendo la esencia del aprendizaje en la repetición de la activación de las neuronas que acaba produciendo el reforzamiento de sus correspondientes conexiones sinápticas con las demás.

Durante los años cincuenta estas ideas se fueron trasladando al terreno de la computación y los autómatas, teniendo como figuras relevantes, entre otros, a Von Neumann, con sus trabajos sobre la robustez de los modelos de MacCulloch y Pitts, y a Minsky, uno de los primeros en acuñar los términos inteligencia artificial y redes neuronales artificiales.

Una figura especialmente importante en el desarrollo de las redes neuronales fué Rosemblatt quién, con un nuevo algoritmo de aprendizaje supervisado, concibió el denominado *Perceptron* [Rosemblatt 1958] y sentó las bases de la demostración de su teorema de convergencia. Rosemblatt y su equipo trabajaron en redes de perceptrones pero no concibieron algoritmos de aprendizaje eficaces para redes complejas.

Un poco después [Widrow y Hoff 1988] desarrollaron *Adaline* (*Adaptive linear element*), cuya única diferencia con el *Perceptron*, residía en el algoritmo de aprendizaje, y, posteriormente, la red *Madaline*, (varias unidades *Adaline*).

Sin embargo, ya a mediados de los cincuenta, [Gabor 1954] había anticipado este tipo de redes al desarrollar el denominado *Learning filter* que utilizaba el concepto entrada-salida y la idea de presentar sucesivamente muestras al modelo con el fin de que el sistema se adaptase convenientemente para reproducir la salida deseada.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

El *Learning filter* de Gabor ya poseía un nivel intermedio (*oculto*) de unidades en su estructura y se adaptaba mediante el algoritmo del gradiente minimizando el error cuadrático medio. Un poco después, [Aizerman *et al.* 1964*b*] introdujeron estructuras con funciones de base radial.

Pero, tales circunstancias, pasaron desapercibidas en la época cuando [Minsky y Papert 1969], mostraron los límites del Perceptron y de Adaline, y conjeturaron que dichos límites afectaban también a redes con varias capas de neuronas, (problema del *Credit asignment* a las unidades ocultas).

La influencia del trabajo de Minsky y Pappert fué tan grande, que durante los setenta, pocos investigadores siguieron trabajando en el campo de las redes neuronales. Sin embargo, en contra de lo que generalmente se cree, en estos años se desarrollaron algoritmos para el entrenamiento de redes de varias capas (véase [Barron y Barron 1988, p. 196]) o basados en aprendizaje competitivo [Grossberg 1980].

Los años ochenta fueron protagonistas del resurgimiento definitivo de las redes neuronales en una gran cantidad de direcciones, de las que citamos, entre las más relevantes:

1. La introducción de la retroalimentación en la arquitectura de las redes neuronales, abriéndose un inmenso campo sobre la estabilidad de su comportamiento dinámico, (*Hopfield networks*, [Hopfield 1982]).

2. Las *self-organizing networks* de [Kohonen 1982].

3. Las *Boltzmann machines* de [Ackley *et al.* 1985], que fueron la primera implementación de una red con varios niveles ocultos y cuyo entrenamiento se basaba en el recién creado algoritmo de *simulated annealing* [Kirkpatrick *et al.* 1983] para problemas de optimización combinatorios. Con ello, se rompía la barrera psicológica creada por el libro de Minsky y Pappert.

4. El desarrollo del algoritmo *back-propagation* ([Rumelhart *et al.* 1986]) para el entrenamiento de redes con más de un nivel oculto, lo que fomentó la utilización de funciones de activación diferenciables.

5. El desarrollo de las redes neuronales de base radial ([Broomhead y Lowe 1988]) como alternativa a las redes *multilayer perceptron*.

La gran explosión de los años ochenta, se vió catalizada por el rápido crecimiento de la capacidad computacional de los ordenadores, que sirvió de laboratorio para el desarrollo y depuración de nuevas técnicas.

Los años noventa fueron testigos de la utilización de las redes neuronales en diferentes campos de aplicación y de un creciente interés por estas técnicas por parte de otras áreas de la ciencia, lo que produjo beneficios en las dos direcciones.

En lo que a la estadística se refiere, los investigadores de las redes neuronales descubrieron que muchas de las técnicas consideradas inicialmente como cajas negras (*black box*) podían ser fundamentadas sobre el cálculo de probabilidades y la estadística inferencial, y que, técnicas similares y, en general, más fundamentadas teóricamente, eran ya clásicas en la ciencia estadística.

Esta influencia provocó que en los primeros años noventa proliferaran los intentos destinados a fundamentar y explicar teóricamente el buen comportamiento de las redes neuronales. Sin embargo, la influencia en el sentido inverso sobre la comunidad estadística no fué tan intensa hasta

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

la segunda mitad de los años noventa, cuando ya se hizo manifiesto el éxito de la aplicación de las redes neuronales a algunos problemas estadísticos.

Los desarrollos de estos últimos años no fueron ya patrimonio de la neurofisiología, la ingeniería o la física, si no producto de investigadores procedentes de diversas áreas y, en particular, del campo de la estadística. Entre los más relevantes para este trabajo, cabe mencionar:

1. Las redes neuronales de regularización ([Girosi *et al.* 1995], [Poggio y Girosi 1990]) producto de la introducción en el campo de las redes neuronales de la teoría de la regularización de Tikhonov [Tikhonov y Arsenin 1977] y que sirvieron para establecer un puente conceptual con los *smoothing splines* [Wahba 1990*b*].

2. La aplicación del enfoque bayesiano a las redes neuronales, dando lugar a métodos específicos de entrenamiento y selección del modelo ([MacKay 1992*c*], [Williams 1998], [Neal 1996]).

3. Las *Support Vector Machines* desarrolladas por Vapnik y colaboradores, ([Boser *et al.* 1992], [Cortes y Vapnik 1995]), en el marco de la *Statistical Learning Theory*, ([Vapnik y Chervonenkis 1971], [Vapnik y Chervonenkis 1981], [Vapnik 1982], [Vapnik 1998]).

4. Las combinaciones de redes, (*Committee Machines*) ideadas con diversos fines: como medio de repartir la tarea de aprendizaje entre varias redes especializadas (mixturas de expertos, ([Jacobs *et al.* 1991], [Jordan y Jacobs 1994]), o entre varios mecanismos poco complejos (*weak learners*) con el fin de: obtener uno combinado más potente, (*boosting*, [Freund y Schapire 1997]), extraer la máxima información contenida en la muestra mejorando las propiedades de la red, (*bagging*, [Breiman 1996*a*]) o, simplemente, aprovechar al máximo las diversas redes resultantes de los procesos de entrenamiento y evitar la simplificación que lleva consigo la táctica de seleccionar el supuesto *mejor* modelo: combinaciones lineales de redes, [Hashem 1997] y, en general, *stacking* ([Breiman 1996*c*], [Wolpert 1992]).

En este nuevo período, resulta cada vez más difícil encontrar la motivación neurobiológica que inspiraron los modelos iniciales y, por tanto, delimitar el concepto diferenciador de las redes neuronales. En esta idea redunda el hecho de que algoritmos nacidos en el ámbito de las redes neuronales existían ya, a veces en versiones menos eficientes, o fueron redescubiertos posteriormente en otras áreas.

En la actualidad puede decirse que, como técnicas estadísticas, las redes neuronales forman parte de un gran conjunto de técnicas paramétricas con filosofía no paramétrica, que comparten un paradigma común e intercambian métodos y algoritmos. Ello produce el continuo enriquecimiento del conjunto y una continua fuente de inspiración para la aparición de nuevas técnicas.

II.1.2 La Red Neuronal

El modelo de red neuronal, inspirado en su contrapartida biológica, es un conjunto de dos elementos básicos, (una propuesta reciente de formalización como estructura de datos es [Atencia *et al.* 2000]):

1. Una **estructura** constituida por un conjunto de unidades básicas de proceso, nodos o **neuronas**, organizadas en niveles o capas (*layers*). A veces, se considera también neuronas a las señales de entrada, constituyendo la capa de entrada (*input layer*). Sin embargo, para mayor claridad no seguiremos dicha convención en el presente documento. La última capa de

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

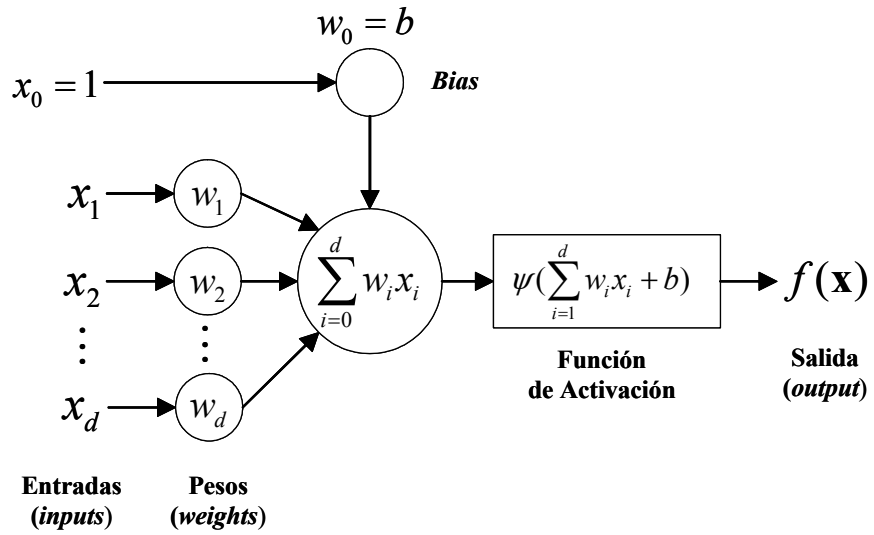


Figura 2. Estructura de una neurona en la que se muestran sus distintos elementos y se trata el *bias* o parámetro independiente como peso de una entrada constante.

unidades de proceso se denomina nivel de salida (*output layer*) y las capas intermedias se denominan **ocultas**.

Cada **neurona** (figura 2) se compone de los siguientes elementos:

(a) Un conjunto de conexiones de entrada o *sinapsis*, cuya intensidad está caracterizada por **pesos** o parámetros que refuerzan o inhiben la intensidad de las señales recibidas.

(b) Un valor que se detrae de la agregación de las señales de entrada transmitidas por las *dendritas* a partir de las sinapsis, reflejando la idea de **umbral de activación**. El valor opuesto a este umbral se denomina tradicionalmente *bias*, aunque sin el significado tradicional de este término en estadística.

(c) Una **función de activación** que actúa sobre las señales de entrada alteradas por los pesos, menos el umbral.

(d) Una salida resultado de la función de activación que, a través del *axón*, se transmite mediante sinapsis a las neuronas posteriores, (o forma parte de la salida de la red neuronal). Esta salida de la neurona como función de las señales entrantes suele denominarse **función de transferencia** de la neurona.

La estructura anterior puede representarse como un grafo dirigido: una colección de nodos o vértices, que representan las unidades de proceso, conectados mediante arcos dirigidos y ponderados por los pesos. En este marco, se define una **arquitectura** de red neuronal, como una familia de redes neuronales que pueden representarse mediante el mismo grafo dirigido con iguales funciones de activación en los nodos, pero con posiblemente diferentes pesos asociados.

Como tipos de arquitectura más relevantes, destacamos los siguientes:

(a) Según el número de capas: redes de una sóla capa o redes multicapa (*multilayer*). Entre estas últimas, las más utilizadas poseen dos capas: una primera capa denominada oculta y una capa de salida.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

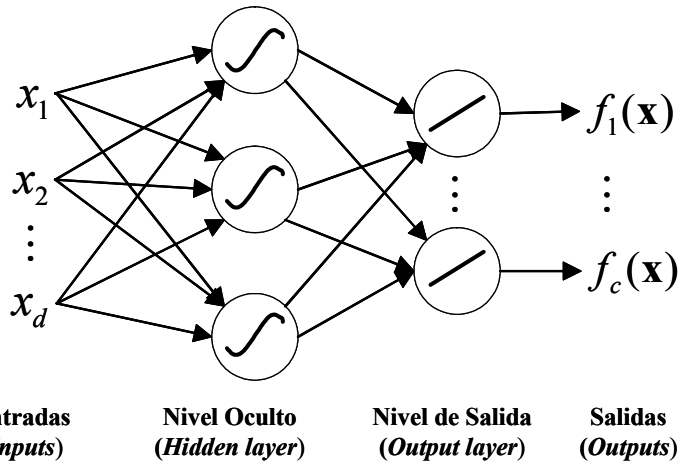


Figura 3. Estructura de una red neuronal *feed-forward* con una capa oculta y dos neuronas en el nivel de salida.

(b) Según el grado de interconexión de la estructura: redes totalmente conectadas, si todos los nodos son elementos iniciales o finales de un camino del grafo, y parcialmente conectadas, en caso contrario.

(c) Según el carácter de las conexiones: redes *feed-forward* (figura 3), cuando el grafo de su arquitectura es de carácter acíclico, es decir ningún camino dirigido puede iniciarse y terminar en el mismo nodo, y redes retroalimentadas o recurrentes cuando el grafo admite ciclos (*feedback*).

2. Un **algoritmo de entrenamiento** o aprendizaje, que, a partir de un conjunto de datos, adapta los pesos de la red neuronal según algún criterio que mide la discrepancia entre la salida de la red y la realidad que se pretende representar.

   En este punto, es importante resaltar que, en la terminología de redes neuronales, el término algoritmo no suele incluir mecanismos de selección del modelo según lo tratado en el capítulo anterior, sino, sólo, los procedimientos de minimización del riesgo empírico u otro principio inductivo utilizado[39].

   En una primera aproximación, los tipos de algoritmo de entrenamiento, pueden clasificarse en algoritmos de tipo *on-line* cuando la red se adapta en tiempo real a medida que se presentan nuevos datos, y de tipo *batch*, cuando la red se adapta después de la presentación de todos o algún subconjunto de los datos disponibles.

   Igualmente, dependiendo de las características del problema de aplicación, puede hablarse de dos tipos de aprendizaje:

   (a) **Supervisado**, cuando la red debe representar la posible relación entre un conjunto de variables independientes (*inputs*) y un conjunto de variables dependientes (*targets*).

   (b) **No supervisado**, cuando la red debe autoorganizarse con el fin de que su estructura reproduzca la estructura intrínseca de la realidad representada por un conjunto de variables de igual jerarquía.

[39] Sin embargo, esta aseveración es relativa al grado de desarrollo de estas técnicas, pues el objetivo del algoritmo de aprendizaje es conseguir la máxima capacidad de generalización en la red neuronal (aprendizaje). Un ejemplo, de ello es el aprendizaje bayesiano que trataremos más adelante.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

A medio camino entre estos dos tipos de aprendizaje se encuentra el aprendizaje **autoasociativo**, que es realmente un aprendizaje supervisado en el que la variable dependiente coincide con la independiente, por lo que se utiliza para los mismos fines que el aprendizaje no supervisado.

El problema de aplicación determina el objetivo del aprendizaje (criterio de utilidad o función de pérdida), así como la arquitectura necesaria para la red neuronal: estructura y complejidad de la red (número de capas y neuronas), y funciones de activación necesarias para representar adecuadamente la realidad del problema.

II.1.3 Redes Neuronales para la Resolución de Problemas Estadísticos

Con estos conceptos generales se han desarrollado multitud de redes neuronales específicas para diferentes problemas de aplicación. Excelentes monografías al respecto son: [Haykin 1999] por su exhaustividad, [Ripley 1996] por su profundidad y [Bishop 1995] y [Cherkassky y Mulier 1999] por su carácter didáctico. Entre estas arquitecturas, destacamos las siguientes para problemas de tipo estadístico:

1. Aprendizaje Supervisado.
   (a) **Clasificación**. Para poder adaptarse a variables dependientes (*target*) de tipo dicotómico, este tipo de redes utilizan funciones de activación en salida de tipo dicotómico, (*Heaviside* o cualquier función tipo salto (*step*), o bien, versiones suavizadas de las anteriores (funciones tipo sigmoide como la logística o la tangente hiperbólica, etc., e incluso funciones lineales), que producen una regla de decisión mediante el método *plug-in*.
   Ejemplos de este tipo de redes son, con una sóla capa, el perceptrón, y la red adaline, y con varias capas, las redes probabilísticas, la red *multilayer perceptron* (MLP), y las *support vector machines* (SVM).

   (b) **Regresión**. Las redes de este grupo son similares a las utilizadas para el problema de clasificación, con la salvedad de que, las funciones de activación de la capa de salida, poseen valores reales con el fin de adaptarse al rango de la variable dependiente. Dichas funciones de activación poseen en muchos casos carácter lineal, debido a su simplicidad y a que, la capa oculta, suministra la posiblemente necesaria no linealidad.
   Ejemplos de este grupo son las redes *multilayer perceptron* (aunque las unidades de salida no son ahora perceptrones), las redes neuronales de base radial (*radial basis function networks*, redes RBF), las support vector machines, etc.

   (c) **Series de tiempo**: Estos problemas pueden tratarse mediante redes *estáticas* diseñadas para el problema de regresión incorporando variables retardadas en la entrada, en cuyo caso, se denominan redes *time lagged feed-forward*. Al margen de ellas, se han desarrollado redes específicas entre las que cabe destacar las redes Elman, que poseen carácter retroalimentado y se entrenan también con carácter supervisado.

2. Aprendizaje no supervisado y autoasociativo, para reducción de la dimensionalidad y reconocimiento de patrones. Ejemplos estándar de este grupo de redes son (p. ej. [Kohonen 2001], [Diamantaras y Kung 1996], [Cherkassky y Mulier 1999]):
   (a) **Análisis de componentes principales**: redes lineales y redes MLF de cuatro capas con aprendizaje autoasociativo, *hebbian learning*, *kernel principal component analysis*, etc.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

(b) **Reconocimiento de patrones**: *self-organizing maps*, *vector quantization*, *competitive networks*.

En lo que sigue, nos centraremos en las redes neuronales *multilayer feed-forward* (MLF) totalmente conectadas, con dos capas (una capa oculta), y con aprendizaje supervisado (y, en algún caso, autoasociativo). Ello es debido a que, como veremos, su arquitectura es suficiente para los problemas de clasificación y regresión, objeto de este trabajo. Sólo trataremos someramente las redes *feed-forward* de una sóla capa, a modo de referencia y por razones de completitud conceptual.

II.1.4 Modelo Funcional de la Red Neuronal

Como hemos apuntado, las funciones de activación de las distintas neuronas deben adaptarse al problema de aplicación. Si se pretende que la red neuronal capture la relación existente entre un conjunto de variables independientes y una variable dependiente, la función de activación de la unidad de salida debe adaptarse a las características de este última. Por su parte, la capa oculta, si existe, realiza una transformación previa de los datos, cuya complejidad está determinada por la propia complejidad de las funciones de activación de las unidades ocultas, y por su número.

Un estudio taxonómico de los diferentes tipos de función de activación utilizados con frecuencia para diferentes problemas puede verse por ejemplo en [Duch y Jankowski 1999].

Como consecuencia de la acción sucesiva de las neuronas de las diferentes capas, una red neuronal *feed-forward* definida como anteriormente, con funciones de activación determinadas y pesos con valores fijos, implementa una función $\mathbf{f} : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathcal{Y} \subset \mathbb{R}^c$, siendo $\mathcal{X}$ e $\mathcal{Y}$ su dominio e imagen, respectivamente. (Sin embargo, diferentes redes pueden implementar la misma función pues, por ejemplo, cualquier permutación de los nodos de una capa con sus pesos respectivos no afecta a la salida de la red).

En particular, las funciones implementables por una arquitectura de red neuronal *feed-forward* pueden formularse mediante el siguiente modelo general:

$$\begin{array}{c} \mathbf{f}(\mathbf{x}) = \boldsymbol{\varphi}(\boldsymbol{\psi}(\mathbf{x})) = (\boldsymbol{\varphi} \circ \boldsymbol{\psi})(\mathbf{x}) \\ \boldsymbol{\psi} : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathcal{U} \subset \mathbb{R}^h \\ \boldsymbol{\varphi} : \mathcal{U} \subset \mathbb{R}^h \rightarrow \mathcal{Y} \subset \mathbb{R}^c \end{array} \tag{II.1}$$

donde $\boldsymbol{\psi} = (\psi_1, ..., \psi_h)^T$ y $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_c)^T$, $\mathcal{X}$ es el espacio de entrada, $\mathcal{Y}$ el espacio de salida y $\mathcal{U}$ el espacio de variables ocultas denominado frecuentemente espacio de características (*features*).

Las siguientes arquitecturas son casos particulares de lo anterior:

1. Redes multilayer perceptron (MLP):
   (a) $\psi_j(\mathbf{x}) = \psi(\mathbf{w}_j^T\mathbf{x} + w_{j0})$ con $\mathbf{w}_j \in \mathbb{R}^d$, $j = 1 : h$, siendo ψ la función de activación de las unidades del nivel oculto, $\mathbf{w}_j$ el vector de parámetros de dicha unidad y $w_{j0} \in \mathbb{R}$ su valor umbral. Normalmente, la función ψ es de tipo sigmoide, (logística, tangente hiperbólica, etc.), pero pueden utilizarse otras que no sean polinomios (Anexo **??**).
   (b) $\varphi_j(\mathbf{u}) = \varphi(\mathbf{c}_j^T\mathbf{u} + c_{j0})$, $j = 1 : c$, donde $\mathbf{u} = \boldsymbol{\psi}(\mathbf{x}) \in \mathcal{U}$, φ es la función de activación de la unidad j-ésima del nivel de salida, $\mathbf{c}_j \in \mathbb{R}^h$, $j = 1 : h$ el vector de pesos de dicha unidad y $c_{j0} \in \mathbb{R}$ su valor umbral. La función φ depende del problema en cuestión: para problemas de regresión suele utilizarse la función identidad , y para problemas de clasificación, la función *Heaviside* o cualquier función dicotómica.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

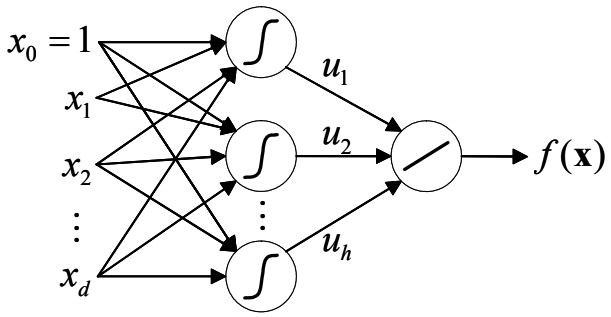


Figura 4. Arquitectura *multilayer perceptron* (MLP), (ecuación (II.2)), para regresión. Las funciones de activación del nivel oculto suelen ser sigmoides y la del nivel de salida, lineal.

En particular, las redes MLP con una sóla salida ($c = 1$), responden al siguiente modelo específico (figura 4):

$$f_{\mathbf{c},\mathbf{W}}(\mathbf{x}) = \varphi(\sum_{j=1}^{h} c_j \psi(\mathbf{w}_j^T \mathbf{x} + w_{j0}) + c_0) \tag{II.2}$$

donde hemos denotado $\mathbf{c} = (c_0, ..., c_h)^T$ y $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_h)$ con $\mathbf{w}_j = (w_{j0}, w_{j1}, ..., w_{jh})^T$, $j = 1 : h$.

2. Redes de base radial, (RBF), para las que:
   (a) $\psi_j(\mathbf{x}) = \psi(\|\mathbf{x} - \mathbf{w}_j\| / \sigma_j)$ siendo ψ la función de activación de las unidades del nivel oculto, $\mathbf{w}_j$ el vector de pesos de dicha unidad y σ_j un parámetro umbral que determina el rango de influencia de la función de activación, (este parámetro puede generalizarse a una matriz de covarianza reflejando diferentes rangos de influencia en las distintas direcciones del espacio).
   (b) $\varphi_j(\mathbf{u}) = \varphi(\mathbf{c}_j^T \mathbf{u} + c_{j0})$ donde $\mathbf{u} = \boldsymbol{\psi}(\mathbf{x}) \in \mathcal{U}$, φ es la función de activación de la unidad j-ésima del nivel de salida, $\mathbf{c}_j$ el vector de parámetros de dicha unidad y c_{j0} su valor umbral. Normalmente, las redes RBF se utilizan para el problema de regresión por lo que la función φ es la función identidad.

   Con ello, la red neuronal RBF con una sóla salida ($c = 1$) responde al siguiente modelo (figura 5):

$$f_{\mathbf{c},\mathbf{W},\boldsymbol{\sigma}}(\mathbf{x}) = \sum_{j=1}^{h} c_j \psi(\frac{\|\mathbf{x} - \mathbf{w}_j\|}{\sigma_j}) + c_0 \tag{II.3}$$

   donde hemos denotado $\mathbf{c} = (c_0, ..., c_h)^T$, $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_h)$, $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_h)^T$.

En general, la consideración de un espacio de salida de varias dimensiones no supone pérdida de generalidad pues, puede implementarse mediante otras tantas redes simultáneas con $c = 1$ – una por cada variable de salida. Por ello, para simplificar la exposición, supondremos $c = 1$ salvo que se indique lo contrario.

Haciendo explícitos los parámetros, la expresión general de las redes *feed-forward* con una sóla capa oculta y una sóla salida, sería la siguiente, (nótese que h es también un parámetro de la arquitectura):

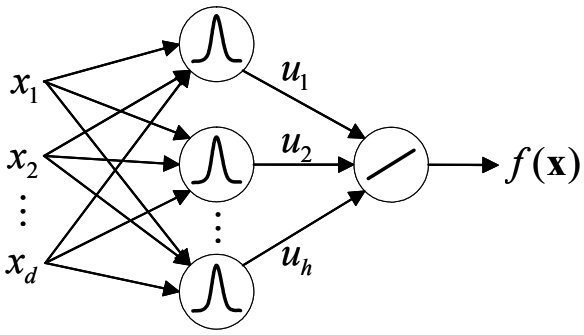


Figura 5. Arquitectura de función de base radial (RBF) (ecuación (II.3)) para regresión. Las funciones de activación del nivel oculto son funciones radiales y la del nivel de salida, lineal.

$$f_{\mathbf{\Gamma},\mathbf{c}}(\mathbf{x}) = \varphi(\textstyle\sum_{j=1}^{h} c_j \psi_{\boldsymbol{\gamma}_j}(\mathbf{x}) + c_0) \\ \mathbf{c} \in \mathbb{R}^h, \boldsymbol{\gamma}_j \in \mathbb{R}^{p_j}, j = 1:h \tag{II.4}$$

donde $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_h)$, $\{\psi_{\boldsymbol{\gamma}} : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R} : \boldsymbol{\gamma} \in \mathbb{R}^p\}$ es una familia de funciones parametrizada por $\boldsymbol{\gamma}$ y $\varphi : \mathbb{R} \to \mathcal{Y} \subset \mathbb{R}$.

II.1.5 Relación Con Otros Modelos Estadísticos

Es interesante observar que, además de los modelos paramétricos de regresión (lineal, polinómica, etc., p. ej. [Seber 1977], [Seber y Wild 1989]), muchos modelos de regresión y clasificación son también casos particulares del modelo general anterior. Por ejemplo:

1. Los modelos multivariantes de regresión lineal generalizada (con transformación previa de variables), y los modelos aditivos (p. ej. [Hastie y Tibshirani 1990]), con $\varphi(u) = u$, que pueden verse como modelos de regresión lineal sobre un espacio de características (*features*) definidas por funciones de las variables de entrada.

2. Los modelos lineales generalizados y aditivos generalizados (p. ej. [McCullagh y Nelder 1983], [Hastie y Tibshirani 1990]), para los que la función *link* se correspondería con la función inversa de φ.

3. La regresión tipo núcleo (regresión polinómico local mediante constantes), (p. ej. [Härdle 1990], [Wand y Jones 1995]), en la que $\psi_{\boldsymbol{\gamma}_j}$ es el núcleo, $\boldsymbol{\gamma}_j = (\mathbf{x}_j, \sigma_j)$ y $c_j = y_j, j = 1:n$.

4. La regresión polinómico local, (p. ej. [Loader 1999], [Fan y Gijbels 1996]). Por ejemplo, si $d = 1$ para polinomios lineales se tiene $c_j = y_j$ y $\psi_{\boldsymbol{\gamma}_j}(x) = w_j(x) / \sum_{j=1}^{n} w_j(x)$ con:

$$w_j(x) = k_{\sigma_j}(x_j - x)[s_{n,2} - (x_j - x)s_{n,1}],\ j = 1:n$$
$$\text{siendo:} \qquad s_{n,j} = \sum_{j=1}^{n} k_{\sigma_j}(x_j - x)(x_j - x)^j, j = 1, 2$$

   donde $k_{\sigma_j}(\cdot) = k(\cdot/\sigma_j)/\sigma_j$ siendo k alguna función de densidad y $\boldsymbol{\gamma}_j = (x_j, \sigma_j)$, $j = 1:n$.

5. Los *splines,* (p. ej. [Green y Silverman 1994], [Wahba 1990*b*], [Eubank 1988]), en los que por ejemplo, para $d = 1$: $\varphi_j(u) = u$ y las funciones $\psi_{\boldsymbol{\gamma}_j}$ son, o bien monomios x^r, $r = 0, ..., p$, y $(x - t_j)_+^p$ con t_j los *nudos* (*knots*), o bien las funciones básicas *B-splines*. Igualmente, los

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

smoothing splines donde las funciones $\psi_{\boldsymbol{\gamma}_j}$ son expresiones en función del núcleo reproductor k de un espacio de Hilbert con Núcleo Reproductor (RKHS), y términos polinómicos.

6. El modelo de regresión *projection pursuit*, ([Friedman 1987], [Friedman y Stuetzle 1981]), en el que $\varphi(u) = u$ y las funciones $\psi_{\boldsymbol{\gamma}_j}$ son suavizadores en cada proyección del espacio de entrada, donde $\boldsymbol{\gamma}_j = (\mathbf{w}_j, \boldsymbol{\alpha}_j)$, con $\mathbf{w}_j$ vector de proyección y $\boldsymbol{\alpha}_j$ vector de parámetros del suavizador correspondiente.

7. Expansiones en serie, (p. ej. [Efromovich 1999]), en las que $\varphi(u) = u$, las funciones $\psi_{\boldsymbol{\gamma}_j}$ son elementos de una base del espacio de funciones considerado y c_j son los coeficientes estimados de la expansión en términos de dicha base.

   Caso particular de lo anterior, son las *wavelets* (p. ej. [Vidakovic 1999], [Mallat 1989], [Daubechies 1992]), (así como sus generalizaciones [Candes 1998]), donde $\varphi(u) = u$ y las funciones $\psi_{\boldsymbol{\gamma}_j}$ son los elementos de la base ortonormal de $L_2(\mathbb{R})$ constituida por las traslaciones de la función de escala (*father wavelet*), y escalamientos y traslaciones de la función *mother wavelet*, y c_j son los coeficientes estimados de la expansión.

8. Modelos basados en diccionarios constituidos por bases o estructuras (*frames*) redundantes en algún espacio de funciones, (Matching Pursuit [Mallat y Zhang 1992] y Basis Pursuit [Chen *et al.* 1999], en el que las funciones $\psi_{\boldsymbol{\gamma}_j}$ son funciones del diccionario.

9. Combinaciones de los modelos anteriores: *bagging*, (p. ej. [Breiman 1996*a*]), *boosting*, (p. ej. [Freund y Schapire 1997]) y, en general, *stacking* ([Wolpert 1992], [Breiman 1993]).

Por tanto, todas estas técnicas admiten una representación conexionista como la mostrada en la figura 3, lo que permite analizarlas desde una nueva perspectiva. La consideración o no de dichas técnicas como redes neuronales depende de la convención terminológica que más se adapte a nuestra finalidad.

Por tanto, el amplio abanico de modelos que pueden encuadrarse en la familia (II.4) sugiere que, al margen de las formas funcionales específicas de cada uno, la diferencia más profunda entre unas técnicas y otras, son los algoritmos empleados para la estimación o aprendizaje de sus parámetros.

II.1.6 Requisitos de una Red Neuronal

Sin embargo, el algoritmo de entrenamiento (aprendizaje, adaptación o estimación) no es el único aspecto relevante: la capacidad de una red neuronal (así como de las demás técnicas anteriores) para capturar una relación o función subyacente en la realidad de unos datos depende de las siguientes cuestiones fundamentales:

1. **Capacidad de representación**. En primer lugar, interesa la capacidad *expresiva* de la arquitectura (**modelo**), es decir, la determinación de la familia de funciones que puede implementar (o aproximar) dado su grado de complejidad. Y a la inversa, dada una familia de funciones determinada, cuál es la complejidad necesaria en la arquitectura para que dicha familia sea implementable (aproximable).

   Las figuras 6 y 7 muestran respectivamente la evolución de la capacidad de representación de las redes MLP y RBF, a medida que se incrementa su complejidad (se añaden unidades ocultas). Obsérvese como la mayor sencillez de la función logística $f(r) = (1 + e^{-r})^{-1}$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

en comparación con la función gausiana $f(r) = \exp(-r^2)$, hace que la red MLP requiera dos neuronas para representar una función que una red RBF gausiana representa con tan sólo una neurona. En contrapartida, ello proporciona a la MLP mayor flexibilidad en algunos problemas[40].

2. **Capacidad de adaptación**. La segunda cuestión transcendental es cómo (**algoritmo**) adaptar la arquitectura (número de nodos y parámetros de la red), a un conjunto de datos.

3. **Capacidad de generalización**. Las dos cuestiones anteriores confluyen en esta tercera. Cuando se conoce la función que la red neuronal debe aprender (p. ej. en un problema de cálculo aproximado de una función o funcional) el éxito de la red se reduce a los dos primeros aspectos. Sin embargo, cuando en lugar de una función conocida, se dispone de un conjunto limitado de datos, –circunstancia intrínseca a los problemas estadísticos–, la bondad de la red descansa en su capacidad de generalización, es decir, su capacidad para inferir con éxito a partir de la particularidad del conjunto de datos y, al mismo tiempo, evitar la asimilación del posible ruido a que está sujeto el proceso generador de los mismos.

   Por tanto, el algoritmo de aprendizaje no debe limitarse a conseguir la máxima adaptación a los datos disponibles, sino que, debe incorporar mecanismos de selección del modelo en los términos expuestos en el capítulo anterior.

   Una arquitectura excesivamente compleja con un algoritmo de gran adaptabilidad puede producir una mala capacidad de generalización si la red propone una representación gratuitamente sofisticada de los pocos datos disponibles y/o si no es capaz de obviar el posible ruido de los datos.

En lo que resta de capítulo, presentamos bajo una visión estadística y en el marco de la teoría de aprendizaje, los modelos y algoritmos de redes neuronales más importantes para clasificación y regresión.

Como hemos indicado, nuestro objeto principal son las redes MLF totalmente conectadas, con dos capas (una capa oculta), y con aprendizaje supervisado (y, en algún caso, autoasociativo). Sin embargo, a modo de referencia, comenzamos con una breve exposición de los modelos y algoritmos para redes neuronales de una sóla capa en los problemas de clasificación y regresión, con el fin de utilizar su significado estadístico en el análisis de las redes más complejas[41].

II.2 Redes Neuronales de Una Capa

Las redes neuronales de una sóla capa comúnmente tratadas, están diseñadas para resolver problemas de clasificación de ejemplos identificados por una variable X con valores en $\mathcal{X} \subset \mathbb{R}^d$ y con distribución P_{X}.

Consideraremos en primer lugar el problema de clasificación con dos posibles clases $\mathcal{C}_1, \mathcal{C}_2$, exhaustivas y excluyentes, en las que pueden clasificarse los puntos $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$, identificadas por una variable respuesta $\mathsf{Y} \in \mathcal{Y}$ mediante $\mathcal{C}_j \equiv \{\mathsf{Y} = y_j\}$, $j = 1, 2$, cuya codificación se concretará en cada caso. Para este problema es suficiente una neurona.

[40] Sin embargo, como veremos y como puede deducirse del comentario anterior, la capacidad potencial de representación de estas arquitecturas es similar.

[41] Normalmente, tiende a responsabilizarse a la capa oculta, del significado estadístico de las redes neuronales. Sin embargo, dicho significado está presente ya en las redes de una sóla capa.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

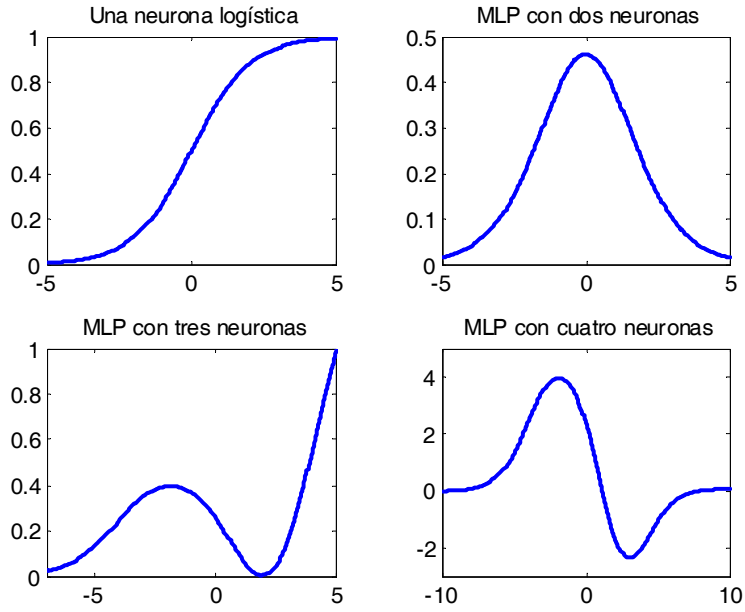


Figura 6. Redes MLP con diferente complejidad. De izquierda a derecha y de arriba a abajo: una neurona logística $(1+e^{-x})^{-1}$, y MLP con dos, tres y cuatro neuronas, respectivamente.

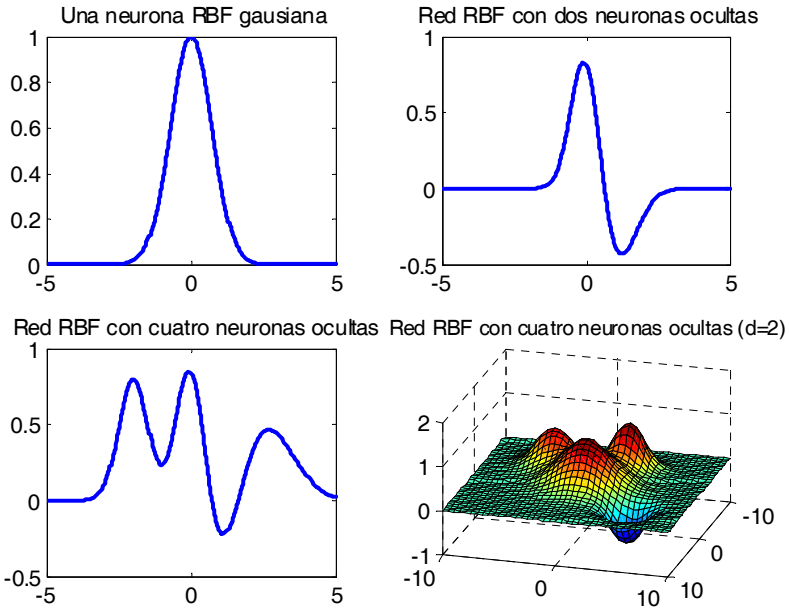


Figura 7. Redes RBF con diferente complejidad. De izquierda a derecha y de arriba a abajo: una neurona gausiana $\exp(-x^2)$, redes RBF con dos y cuatro neuronas, respectivamente, y red RBF con cuatro neuronas en $\mathbb{R}^2$.

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Para fijar terminología, denominamos:

1. **Regla de decisión, de clasificación o regla discriminante,** a una función $g : \mathcal{X} \to \mathcal{Y}$ que, a cada $\mathbf{x} \in \mathcal{X}$, asigna una clase $\mathcal{C}_i$, $i \in \{1, 2\}$.

2. **Función discriminante,** a la función $f : \mathcal{X} \to \mathbb{R}$ que, para cada $\mathbf{x}$, produce un valor $f(\mathbf{x})$ que será utilizado posteriormente por una regla de decisión g_f para asignar $\mathbf{x}$ a la clase $\mathcal{C}_i$, $i \in \{1, 2\}$.

3. **Frontera de decisión o frontera discriminante,** a la región de $\mathcal{X}$ utilizada por la regla g_f para intentar separar las clases entre sí, (si es lineal, se denominará **hiperplano separador**).

Con posterioridad, trataremos el problema de varias clases, $\mathcal{C}_j$, $j = 1 : c$, tanto en el caso de atributos excluyentes como en el caso de atributos independientes, para el cual, ya son necesarias varias neuronas constituyendo una red neuronal de una sóla capa.

En lo que sigue, utilizaremos indistintamente las expresiones $f_{\mathbf{w}}(\mathbf{x}) \equiv f(\mathbf{x}; \mathbf{w})$ y obviaremos el vector de parámetros siempre que no exista riesgo de ambigüedad.

II.2.1 El Perceptrón

Uno de los primeros modelos computacionales de la neurona fué el **perceptrón** propuesto por [Rosemblatt 1958], con la siguiente forma general (figura 2):

$$\begin{array}{c} g_{\bar{\mathbf{w}}}(\mathbf{x}) = \psi(\mathbf{w}^T\mathbf{x} + w_0) = \psi(\bar{\mathbf{w}}^T\bar{\mathbf{x}}) \\ \text{con: } \left\{ \begin{array}{l} \bar{\mathbf{w}} = (w_0, \mathbf{w}^T)^T;\ \bar{\mathbf{x}} = (1, \mathbf{x}^T)^T \\ \psi(r) = \text{signo}(r) \end{array} \right. \end{array} \tag{II.5}$$

donde se define la función signo como $\text{signo}(r) = 1$ si $r \geq 0$ y $\text{signo}(r) = -1$ en caso contrario.

De esta forma, para $\mathcal{Y} = \{-1, 1\}$, el perceptrón implementa una regla de clasificación basada en una función lineal: $f_{\bar{\mathbf{w}}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$.

Dada una muestra aleatoria $\mathbf{z}^n = (\mathbf{z}_1, ..., \mathbf{z}_n)$, con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i = 1 : n$, el problema de aprendizaje del perceptrón utiliza como principio inductivo, la minimización de la siguiente función de error:

$$\hat{R}_n(\bar{\mathbf{w}}) = -\sum_{y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i) < 0} y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i) = -\sum_{y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i) < 0} y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \tag{II.6}$$

que es lineal a trozos, continua y con discontinuidades en su gradiente. Este principio no encaja en la formulación general del problema de aprendizaje realizada en el capítulo anterior, pues no está basado en un función de pérdida de la forma $\ell(y, g(\mathbf{x}))$ donde g es la hipótesis (en este caso, una regla de clasificación). Sin embargo, la elección de este principio es puramente instrumental como función objetivo del algoritmo de entrenamiento con el fin de conseguir el menor número de errores de clasificación. Así, en esencia, la pérdida realmente considerada es la pérdida 0-1.

Se tiene que:

$$\bar{\mathbf{w}}^T\bar{\mathbf{x}}_i = \mathbf{w}^T\mathbf{x}_i + w_0 \Rightarrow 0 = \mathbf{w}^T\mathbf{x}_i + (w_0 - \bar{\mathbf{w}}^T\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i + \alpha_i$$

donde $\alpha_i = w_0 - \bar{\mathbf{w}}^T\mathbf{x}_i$ con lo que el hiperplano $\{\mathbf{w}^T\mathbf{x} + \alpha_i = 0\}$ pasa por $\mathbf{x}_i$. Con ello, $|w_0 - \alpha_i| = \left|\mathbf{w}^T\mathbf{x}_i + w_0\right|$ es la *distancia*, salvo norma $\|\mathbf{w}\|$, de $\mathbf{x}_i$ al hiperplano $\{\mathbf{w}^T\mathbf{x} + w_0 = 0\}$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

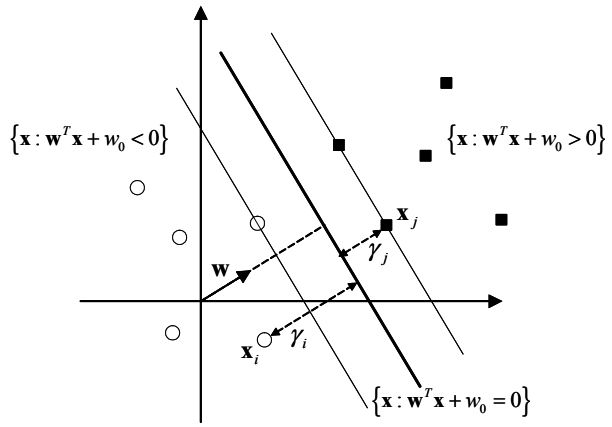


Figura 8. Muestra separable, hiperplano separador y margen geométrico γ_i de una observación $(\mathbf{x}_i, y_i)$

Definición 31 *Una muestra* $\mathbf{z}^n$ *se dice* ***linealmente separable*** *si verifica:* $y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i) \geq 0, \forall i \in \{1 : n\}$ *siendo* $f_{\bar{\mathbf{w}}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$*, en cuyo caso, el hiperplano* $\{\mathbf{w}^T\mathbf{x} + w_0 = 0\}$ *se denomina* ***hiperplano separador****. En este marco, se denomina:*

1. ***Margen funcional*** *de la observación* $(\mathbf{x}_i, y_i)$ *con respecto al hiperplano* $\{\mathbf{w}^T\mathbf{x}_i + w_0 = 0\}$*, a la cantidad:*

$$\chi_i = y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i)$$

2. ***Margen geométrico*** *de la observación* $(\mathbf{x}_i, y_i)$ *con respecto al hiperplano* $\{\mathbf{w}^T\mathbf{x}_i + w_0 = 0\}$*, a la distancia euclídea. entre* $\mathbf{x}_i$ *y el hiperplano, es decir:*

$$\gamma_i = \frac{1}{\|\mathbf{w}\|} y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = \frac{1}{\|\mathbf{w}\|} y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i) = \frac{1}{\|\mathbf{w}\|}\chi_i$$

Así, si el vector $\mathbf{w}$ *está normalizado,* $\|\mathbf{w}\| = 1$*, el margen funcional coincide con el margen geométrico.*

Algoritmo

El algoritmo de aprendizaje del perceptrón resulta de la aplicación del algoritmo del gradiente al problema de minimización de $\hat{R}_n(\bar{\mathbf{w}})$. La versión interactiva (*on-line*) del aprendizaje es exactamente análoga debido a la aditividad del error. En cada iteración $(s+1)$-ésima, sólo los ejemplos mal clasificados modifican el hiperplano actual, mediante[42]:

$$\bar{\mathbf{w}}(s+1) = \bar{\mathbf{w}}(s) + \rho y_i \bar{\mathbf{x}}_i \tag{II.7}$$

donde $\rho > 0$ es un parámetro que controla la velocidad de convergencia. Ello provoca que, en cada iteración, se reduzca el error de los ejemplos mal clasificados:

$$-y_i\bar{\mathbf{w}}(s+1)^T\bar{\mathbf{x}}_i = -[y_i\bar{\mathbf{w}}(s)^T\bar{\mathbf{x}}_i + \rho y_i^2\bar{\mathbf{x}}_i^T\bar{\mathbf{x}}_i] < -y_i\bar{\mathbf{w}}(s)^T\bar{\mathbf{x}}_i$$

[42] Recuérdese que $(\bar{\mathbf{x}}_i)_0 = 1$, por lo que $\partial\hat{R}_n(\bar{\mathbf{w}})/\partial w_0 = -y_i$, y así, para el término independiente, se tiene: $w_0(s) = w_0(s-1) + \rho y_i$

Sin embargo, si la muestra no es linealmente separable, el algoritmo no finalizaría pues, la posible clasificación correcta de algún punto mal clasificado, provocaría la mala clasificación de alguno de los puntos correctos, provocando un comportamiento oscilatorio.

Para muestras linealmente separables, se tiene el siguiente teorema:

Teorema 32 *Novikoff (p. ej. [Vapnik 1998]). Sea* $\mathbf{z}^n$ *una muestra aleatoria linealmente separable y sea* $r = \max_{1\leq i\leq n} \|\mathbf{x}_i\|$ *. Entonces:*

1. *El algoritmo anterior converge en un número finito de iteraciones a alguna solución* $\bar{\mathbf{w}}_* = (\mathbf{w}_*^T, w_{0*})^T$ *(que suponemos normalizado con* $\|\mathbf{w}_*\| = 1$*).*

2. *Si el hiperplano definido por* $\mathbf{w}_*$ *separa la muestra con margen al menos* γ*, es decir,* $y_i(\mathbf{w}_*^T\mathbf{x}_i + w_0) \geq \gamma, \forall i = \{1 : n\}$*, entonces, el número de iteraciones* τ *coincide con el número de errores que comete el perceptrón durante el aprendizaje, y verifica:*

$$\tau \leq (\frac{r}{\gamma})^2$$

Al ser $\hat{R}_n$ aditiva, el gradiente es la suma de los gradientes de las contribuciones de cada ejemplo $(\mathbf{x}_i, y_i)$ mal clasificado: $\nabla_{\bar{\mathbf{w}}}\hat{R}_n(\bar{\mathbf{w}}) = -\sum_{y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i)<0} y_i\bar{\mathbf{x}}_i$, con lo que este algoritmo puede verse como un algoritmo interactivo (*on-line*) en el que, cada iteración, (II.7) es el resultado de la llegada de un nuevo ejemplo mal clasificado.

Si cada ejemplo $\mathbf{x}_i$ ha sido mal clasificado τ_i veces durante el entrenamiento y si, sin pérdida de generalidad, suponemos que el valor inicial de algoritmo es $\bar{\mathbf{w}}(0) = \mathbf{0}_{d+1}$, debido a (II.7) y obviando ρ, el vector óptimo toma la forma:

$$\mathbf{w}_* = \sum_{i=1}^{n} \tau_i y_i \mathbf{x}_i$$

lo que indica que los vectores más difíciles de clasificar participan más decisivamente en la frontera de decisión. Además, debido a ello, se tiene:

$$f_{\bar{\mathbf{w}}_*}(\mathbf{x}) = \mathbf{w}_*^T\mathbf{x} + w_{0*} = \sum_{i=1}^{n} \tau_i y_i \mathbf{x}_i^T\mathbf{x} + w_{0*} \tag{II.8}$$

expresión que incluye el producto interno $\langle \mathbf{x}_i, \mathbf{x}\rangle = \mathbf{x}_i^T\mathbf{x}$ con los puntos de la muestra.

Consistencia

Debido al teorema 32, dada una muestra aleatoria $\mathbf{z}^n$ de una población $\mathsf{Z} = (\mathsf{X}, \mathsf{Y})$, el algoritmo del perceptrón consigue, después de un número finito de iteraciones, la clasificación perfecta de la muestra. Ello significa que, en términos de la pérdida 0-1 $\ell(y, g(\mathbf{x})) = 1_{\{y\neq g(\mathbf{x})\}}$, el riesgo empírico se hace cero $\hat{R}_n(g_*) = 0$, donde $g_*(\mathbf{x}) = \text{signo}(\bar{\mathbf{w}}_*^T\bar{\mathbf{x}})$ siendo $\bar{\mathbf{w}}_*$ la solución $\bar{\mathbf{w}}_*$ obtenida por el algoritmo.

Como vimos en el capítulo anterior, condición necesaria y suficiente para la consistencia del principio de minimización del riesgo empírico era la finitud de la $\text{VCdim}(\mathcal{L})$ donde $\mathcal{L} = \{\ell_f(\mathbf{z}) = \ell(y, f(\mathbf{x})) : f \in \mathcal{F}\}$ y que dicha condición se verificaba si $\text{VCdim}(\mathcal{F}) < \infty$, donde $\mathcal{F}$ es la familia de hipótesis. En este caso, $\mathcal{F}$ es la familia de perceptrones, de la forma:

$$\mathcal{F} = \{g(\cdot) = \text{signo}(\bar{\mathbf{w}}^T\cdot) : \bar{\mathbf{w}} \in \mathbb{R}^{d+1}\}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Puede demostrarse (teorema 218, Anexo G) que $\mathrm{VCdim}(\mathcal{F}) = v = d+1 < \infty$, lo que garantiza la consistencia del problema de aprendizaje del perceptrón. Asimismo, dado que $\hat{R}_n(g_*) = 0$, la cota genérica (I.10) toma la forma siguiente con probabilidad al menos $1-\eta$:

$$R(g_*) \leq \hat{R}_n(g_*) + \varepsilon(n, \eta, v) = \varepsilon(n, \eta, v)$$

Para funciones indicadoras como la implementada por el perceptrón, se tiene [Vapnik 1998, p. 148-149]:

$$\varepsilon(n, \eta, v) = \frac{4}{n}(v \ln \frac{2en}{v} - \ln \frac{\eta}{4})$$

expresión de la que pueden obtenerse, con probabilidad $1-\eta$: fijado n, la cota ε al riesgo incurrido, y, fijado ε, el tamaño muestral n necesario para que el riesgo no supere dicha cantidad.

El Método del Núcleo (*Kernel Trick)*

La debilidad del perceptrón es que su algoritmo no puede enfrentarse a un conjunto de datos no linealmente separable. Una de las formas de resolver este problema es generalizar el perceptrón mediante el denominado método del núcleo (*kernel trick*), que consiste en transformar previamente los vectores de entrada en un espacio de mayor dimensión (**espacio de características** o *features*): $\boldsymbol{\phi} : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathbb{R}^h$ en el que la muestra transformada $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\mathbf{x}_i)$, $i = 1 : n$ sea ya linealmente separable.

El problema en este caso es el cómputo del producto interno implícito en (II.8). Pero esta dificultad se resuelve definiéndolo [Boser *et al.* 1992] mediante alguna función definida positiva $k : \mathbb{R}^h \times \mathbb{R}^h \rightarrow \mathbb{R}$, con $\boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}')$ que garantiza que está bien definido (demostración del teorema 117, Anexo B).

La aplicación del algoritmo en este nuevo espacio (ecuación (II.8) con $\boldsymbol{\phi}(\mathbf{x})$ y $\boldsymbol{\phi}(\mathbf{x}_i)$ en lugar de $\mathbf{x}$ y $\mathbf{x}_i$), produce una solución de la forma:

$$f_{\bar{\mathbf{w}}_*}(\mathbf{x}) = \sum_{i=1}^{n} \tau_i y_i \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}) + w_{0*} = \sum_{i=1}^{n} \tau_i y_i k(\mathbf{x}_i, \mathbf{x}) + w_{0*}$$

que permite resolver el problema exactamente mediante una frontera no lineal (y en el caso particular $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$ se tiene el modelo original). (Nótese que el análisis anterior sobre la consistencia es aplicable igualmente en este caso).

Finalmente, destacamos que el perceptrón se limita a encontrar un hiperplano separador, independientemente de la rotundidad con la que separe los ejemplos de la muestra. Es obvio que el problema no posee solución única, y la regla proporcionada por el algoritmo depende de las condiciones iniciales. Más adelante veremos la técnica de las *support vector machines* que proporciona el hiperplano que separa los ejemplos con un máximo margen.

II.2.2 Adaline

En el planteamiento del perceptrón no interesa la distribución de los datos, pues se persigue la separación perfecta de una muestra. Este planteamiento puede no ser una buena idea si, por ejemplo, los datos están sujetos a ruido. En este caso, o en el caso de que la muestra no sea linealmente separable, una alternativa al perceptrón es la neurona **adaline**, que proporciona una frontera lineal de compromiso para separar las clases.

Con la misma codificación $\mathcal{Y} = \{-1, 1\}$ anterior, la arquitectura original de adaline es la misma que la del perceptrón (II.5):

$$g_{\bar{\mathbf{w}}}(\mathbf{x}) = \psi(\mathbf{w}^T\mathbf{x} + w_0) = \psi(\bar{\mathbf{w}}^T\bar{\mathbf{x}})$$
$$\text{con: } \begin{cases} \bar{\mathbf{w}} = (w_0, \mathbf{w}^T)^T;\ \bar{\mathbf{x}} = (1, \mathbf{x}^T)^T \\ \psi(r) = \begin{cases} +1 & \text{si } r \geq 0 \\ -1 & \text{si } r < 0 \end{cases} \end{cases} \tag{II.9}$$

pero se diferencia en el principio inductivo que, ahora, es el de mínimos cuadrados sobre el argumento $f_{\bar{\mathbf{w}}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ de la función de activación ψ:

$$\hat{R}_n(\bar{\mathbf{w}}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f_{\bar{\mathbf{w}}}(\mathbf{x}))^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})^2 = \frac{1}{2}\left\|\mathbf{y} - \bar{\mathbf{X}}\bar{\mathbf{w}}\right\|^2$$

donde $\bar{\mathbf{X}}$ es la matriz que tiene por filas los vectores $\bar{\mathbf{x}}_i^T$, $i = 1 : n$, e $\mathbf{y} = (y_1, ..., y_n)^T$.

Como puede observarse, la pérdida anterior no encaja en la formulación general del problema de aprendizaje, en la que es de la forma $\ell(y, g(\mathbf{x}))$, y lo mismo sucedería con los demás clasificadores basados en una función discriminante (p. ej. las reglas de clasificación tipo *plug-in*) cuando utilizan dicha pérdida. Cuando se utilizan redes con una capa oculta para ello, este enfoque suele presentarse como un problema de aprendizaje de una función real (regresión), en lugar de un problema de clasificación. En las próximas secciones, seguiremos esta costumbre, pues permite simplificar la exposición.

Algoritmo

El algoritmo de entrenamiento puede ser de tipo *on-line* o *batch*. En este último caso, se obtiene analíticamente la solución de mínimos cuadrados:

$$\bar{\mathbf{w}} = (\bar{\mathbf{X}}^T\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}^T\mathbf{y} \Rightarrow f_{\bar{\mathbf{w}}}(\mathbf{x}) = \bar{\mathbf{x}}^T(\bar{\mathbf{X}}^T\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}^T\mathbf{y}$$

donde se supone que $\bar{\mathbf{X}}^T\bar{\mathbf{X}}$ es de rango máximo. En caso contrario, se utilizaría $\bar{\mathbf{w}} = (\bar{\mathbf{X}}^T\bar{\mathbf{X}})^-\bar{\mathbf{X}}^T\mathbf{y}$ donde $(\bar{\mathbf{X}}^T\bar{\mathbf{X}})^-$ es una inversa generalizada (Anexo G) de $\bar{\mathbf{X}}^T\bar{\mathbf{X}}$, que provoca que:

$$\mathbf{P}_{\bar{\mathbf{X}}} = \bar{\mathbf{X}}(\bar{\mathbf{X}}^T\bar{\mathbf{X}})^-\bar{\mathbf{X}}^T$$

sea la única matriz (simétrica e idempotente) que proyecta $\mathbf{y}$ sobre $\langle\{\mathbf{x}_1, ..., \mathbf{x}_n\}\rangle$, el espacio lineal generado por $\{\mathbf{x}_i\}_{i=1}^n$.

En el caso interactivo, se utiliza el algoritmo del gradiente (Anexo D), que da lugar a la denominada **regla delta** o de **Widrow-Hoff**, [Widrow y Hoff 1988]. Si denotamos $\ell_i(\bar{\mathbf{w}}) = \ell(y_i, f_{\bar{\mathbf{w}}}(\mathbf{x}_i)) = (y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})^2$, entonces, para cada nuevo ejemplo $(\mathbf{x}_i, y_i)$ se tiene:

$$\begin{aligned} \bar{\mathbf{w}}(s+1) &= \bar{\mathbf{w}}(s) - \rho(s)\nabla_{\bar{\mathbf{w}}}\ell_i(\bar{\mathbf{w}}(s)) \\ &= \bar{\mathbf{w}}(s) - \rho(s)(y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})\bar{\mathbf{x}}_i \end{aligned} \tag{II.10}$$

La cantidad $\delta_i = (y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})$ da nombre al algoritmo y si $\rho(s) \to 0$, (por ejemplo, $\rho(s) = \rho_0/s$), el algoritmo es un caso particular del algoritmo de Robbins-Monro ([Robbins y Monro 1951] y Anexo D), que converge a la solución única, $\bar{\mathbf{w}} = (\bar{\mathbf{X}}^T\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}^T\mathbf{y}$.

Teniendo en cuenta el significado de la matriz $\mathbf{P}_{\bar{\mathbf{X}}}$ como operador proyección ortogonal sobre el espacio generado por los vectores $\bar{\mathbf{x}}_i$, $i = 1 : n$:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$$\mathbf{0}_{d+1} = \nabla_{\bar{\mathbf{w}}}\hat{R}_n(\bar{\mathbf{w}}) = \bar{\mathbf{X}}^T\bar{\mathbf{X}}\bar{\mathbf{w}} - \bar{\mathbf{X}}^T\mathbf{y} = \bar{\mathbf{X}}^T(\bar{\mathbf{X}}\bar{\mathbf{w}} - \mathbf{y})$$

el aprendizaje con la regla delta consiste en conseguir que el vector $\bar{\mathbf{X}}\bar{\mathbf{w}} - \mathbf{y}$ se vaya haciendo ortogonal a $\langle\{\mathbf{x}_1, ..., \mathbf{x}_n\}\rangle$.

Consistencia

El riesgo empírico utilizado por adaline no está basado en la pérdida 0-1 como en el caso del perceptrón, por lo que no puede hacerse un análisis de consistencia en los mismos términos. Una solución, es considerar el problema como un aprendizaje de una función real, en el que el riesgo se define como $R(f) = \mathbb{E}_{\mathsf{Z}}\{[\mathsf{Y} - f(\mathsf{X})]^2\}$, en cuyo caso la solución adaline responde al principio de minimización del riesgo empírico.

En este punto, están disponibles los resultados de consistencia y normalidad asintótica del estimador de mínimos cuadrados pero, al mismo tiempo, si se considera $\mathcal{X} \subset \mathbb{R}^d$ compacto, las funciones de $\mathcal{F}$ están totalmente acotadas y, con ello, la pérdida cuadrática es γ-admisible. Con ello, se verifica que $\mathrm{Pdim}(\mathcal{L}) < \infty$ si se cumple que $\mathrm{Pdim}(\mathcal{F}) < \infty$. Pero esta última condición se cumple debido al teorema 222 del Anexo G pues $\mathcal{F}$ es un espacio vectorial de dimensión $d+1$, con lo que $\mathrm{Pdim}(\mathcal{F}) = d + 1 < \infty$.

De esta forma, el problema de aprendizaje de adaline es consistente a la hora de determinar, en términos de la pérdida cuadrática, la mejor función discriminante en la familia de funciones lineales $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$.

Discriminante Lineal de Fisher

En lo anterior, hemos supuesto una codificación $\mathcal{Y} = \{-1, 1\}$. La modificación de dicha codificación permite dar mayor peso a las observaciones con mayores valores de $|y_i|$ en la determinación del hiperplano óptimo.

Una codificación de interés es $y = n/n_1$ si $\mathbf{x}$ está en la clase $\mathcal{C}_1$ e $y = -n/n_2$ si $\mathbf{x}$ está en $\mathcal{C}_2$, donde n_i es el número de ejemplos de la muestra en la clase $\mathcal{C}_i$, $i = 1, 2$, pues, en ese caso, la neurona adaline (con función de activación modificada en consecuencia) implementa (p. ej. [Duda *et al.* 2001, p. 242], [Bishop 1995, p. 109]) la regla de clasificación generada por el discriminante lineal de Fisher $f_{\bar{\mathbf{w}}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ con $w_0 = -\mathbf{w}^T\mathbf{m}$ donde $\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$.

Dicha función discriminante suele obtenerse (Anexo A, [Seber 1984, p. 283 y 293]), sin hipótesis sobre la distribución de los datos, como resultado de proyectar los datos en la dirección, definida por $\mathbf{w} \in \mathbb{R}^d$, que maximiza la distancia entre las medias de las clases y minimiza la varianza dentro de ellas, es decir, de maximizar, con $\|\mathbf{w}\| = a$ constante, el criterio:

$$L(\mathbf{w}) = \frac{(\bar{u}_2 - \bar{u}_1)^2}{s_1^2 + s_2^2} \tag{II.11}$$

donde, para cada clase $\mathcal{C}_j$, $j = 1 : 2$, $\bar{u}_j = \frac{1}{n}\sum_{i\in\mathcal{C}_j}\mathbf{w}^T\mathbf{x}_i$ y s_j^2 son la media y la varianza, respectivamente, de las proyecciones de los datos sobre el correspondiente hiperplano.

Estimación de las Probabilidades a Posteriori

En hipótesis de normalidad e igualdad de covarianzas para $\mathsf{p}_{\mathsf{X}/\mathcal{C}_j}$, $j = 1, 2$, y suponiendo desconocimiento a priori sobre la probabilidad de cada clase, la regla discriminante correspondiente al discriminante lineal de Fisher minimiza la probabilidad de error de clasificación $\mathsf{P}_{\mathsf{Z}}\{\mathsf{Y} \neq g(\mathsf{X})\}$ (Anexo A y p. ej. [Johnson y Wichern 1992]). Sin embargo, cuando estas hipótesis no se veri-

fican, son recomendables otros métodos, (un análisis detallado de robustez puede verse p. ej. en [Seber 1984, p. 297 y 317]).

Una alternativa, consiste en transformar las variables de entrada para que se cumplan las hipótesis anteriores y se mejore la discriminación. Es lo que se conoce en la literatura como discriminante lineal generalizado, (p. ej. [Duda *et al.* 2001, p. 219], [Ripley 1996, p. 121]) y, como se verá más adelante, se consigue con la inclusión de una capa adicional de neuronas: el nivel oculto.

Otro camino, consiste en la utilización de la arquitectura adaline con una función de activación ψ no lineal. Pudiera parecer, que el objetivo de esta modificación, fuese la obtención de una frontera discriminante no lineal con mayor potencial para separar datos no linealmente separables. Sin embargo, las funciones de activación utilizadas suelen ser monótonas, con lo que la frontera resultante sigue siendo lineal.

En realidad, la mejora que se persigue es otra: la estimación de las probabilidades a posteriori $\mathbb{P}(\mathsf{Y} = y|\mathbf{x})$ para aproximar, mediante el método *plug-in*, la regla de Bayes, en contextos más generales que la normalidad y la igualdad de varianzas. Además, este enfoque, aporta una medida de la seguridad con la que se clasifica un nuevo ejemplo.

La estimación de las probabilidades a posteriori puede realizarse mediante regresión por mínimos cuadrados con codificación $\mathcal{Y} = \{0, 1\}$ pues, en ese caso:

$$\mathbb{E}_{\mathsf{Y}/\mathbf{x}}(\mathsf{Y}) = 1 \cdot \mathbb{P}(\mathsf{Y} = 1|\mathbf{x}) + 0 \cdot \mathbb{P}(\mathsf{Y} = 0|\mathbf{x}) = \mathbb{P}(\mathsf{Y} = 1|\mathbf{x})$$

Así, mediante una estructura tipo adaline con activación Heaviside[43] $\psi(r) = 1$ si $r > 0$ y $\psi(r) = 0$ si $r \leq 0$, podría implementarse la denominada regla *plug-in*:

$$g_{\hat{f}_n}(\mathbf{x}) = \begin{cases} 0 & \text{si } \hat{f}_n(\mathbf{x}) \leq \frac{1}{2} \\ 1 & \text{si } \hat{f}_n(\mathbf{x}) > \frac{1}{2} \end{cases}$$

donde $\hat{f}_n(\mathbf{x}) = \hat{\mathbb{P}}(\mathsf{Y} = 1|\mathbf{x})$ es la estimación de las probabilidades a posteriori a partir de los datos. Al tratarse de modelos lineales, esta estructura resulta de adaptar el adaline original $g_{\hat{f}'_n}$ a la nueva codificación:

$$g_{\hat{f}_n}(\mathbf{x}) = \frac{1}{2} g_{\hat{f}'_n}(\mathbf{x}) + \frac{1}{2} \Leftrightarrow g_{\hat{f}'_n}(\mathbf{x}) = 2g_{\hat{f}_n}(\mathbf{x}) - 1$$

Por otra parte, este enfoque está justificado por el siguiente teorema:

Teorema 33 *(p. ej. [Devroye* et al. *1996, th. 2.2]) Sea* $\mathsf{f}(\mathbf{x}) = \mathbb{P}(\mathsf{Y} = 1|\mathbf{x})$ *y* g_{f} *la regla de Bayes que satisface* $R(g_{\mathsf{f}}) = \inf_g R(g)$ *con* $R(g) = \mathbb{P}_{\mathsf{Z}}[g(\mathsf{X}) \neq \mathsf{Y}]$. *La regla plug-in* g_f *construida con* f, *verifica:*

$$\begin{aligned} &1) \quad R(g_f) - R(g_{\mathsf{f}}) = 2 \int_{\mathcal{X}} \left| f(\mathbf{x}) - \tfrac{1}{2} \right| 1_{\{g_f(\mathbf{x}) \neq g_{\mathsf{f}}(\mathbf{x})\}} d\mathsf{P}(\mathbf{x}) \\ &2) \quad R(g_f) - R(g_{\mathsf{f}}) \leq 2E_{\mathsf{X}}[|f(\mathbf{x}) - \mathsf{f}(\mathbf{x})|] \leq 2(\mathbb{E}_{\mathsf{X}}[|\mathsf{f}(\mathbf{x}) - f(\mathbf{x})|^2])^{1/2} \end{aligned}$$

Es decir, si $\hat{f}_n$ es buen estimador de las verdaderas probabilidades a posteriori en el sentido $L_1(\mathsf{P}_{\mathsf{X}})$, (y, por la desigualdad de Cauchy-Schwartz, en el sentido $L_2(\mathsf{P}_{\mathsf{X}})$), entonces la regla *plug-in* $g_{\hat{f}_n}$ es un buen estimador de la regla de Bayes[44].

[43] Esta función resulta de la transformación de la función de activación signo utilizada en el perceptrón, mediante: $\psi(x) = [\text{signo}(x) + 1]/2$ donde ψ es la función de Heaviside.

[44] Téngase en cuenta de todas formas el resultado de [Devroye *et al.* 1996, th. 25.9], que indica que la utilización de

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Sin embargo, el enfoque de mínimos cuadrados presenta otros problemas bien conocidos:

1. La distribución de $\mathsf{Y}/\mathbf{x}$ es de Bernouilli, con lo que el criterio de mínimos cuadrados no es equivalente a la máxima verosimilitud, lo que no garantiza la eficiencia de los estimadores ni que $\hat{f}_n(\mathbf{x}) = \hat{\mathbb{P}}(\mathsf{Y}=1|\mathbf{x}) \in [0,1]$.

2. Las perturbaciones son heterocedásticas, lo que exigiría mínimos cuadrados ponderados.

Sólamente, en el caso de observaciones repetidas $\{(t_{ij}, \mathbf{x}_i),\ j = 1, ...n_i, i = 1 : n\}$ con $t_{ij} \in \{0,1\}$, si n_i es suficientemente grande, se toma como codificación $y_i = \frac{1}{n_i}\sum_{j=1}^{n_i} t_{ij}$ y las clases son igualmente probables, podría utilizarse adaline de forma aproximada ya que, entonces, $\mathsf{Y}/\mathbf{x}$ es asintóticamente normal.

II.2.3 Adaline con Activación Sigmoide

Con el fin de construir funciones discriminantes que representasen probabilidades a posteriori y, al mismo tiempo, dotar de mejores propiedades analíticas al modelo perceptrón-adaline, se sustituyó su función de activación discreta por una función de activación de tipo **sigmoide**, es decir, monótona con derivada continua y con:

$$-\infty < a = \lim_{a\to-\infty} \psi(r) < \lim_{a\to\infty} \psi(r) = a' < \infty$$

entre las que, la más frecuentemente utilizada es la logística:

$$\psi(r) = \frac{1}{1+e^{-r}} = \frac{e^r}{1+e^r}$$

Ahora, la neurona no representa ya una regla de decisión sino un modelo discriminante que, una vez entrenado contra una codificación $\mathcal{Y} = \{0,1\}$, implementa una regresión logística (p. ej. [McCullagh y Nelder 1983], [Seber 1984, p. 308]). Ello es debido a que:

$$\ln \frac{\mathsf{p}(\mathcal{C}_1|\mathbf{x})}{1-\mathsf{p}(\mathcal{C}_1|\mathbf{x})} = \bar{\mathbf{w}}^T\bar{\mathbf{x}} \Rightarrow \mathsf{p}(\mathcal{C}_1|\mathbf{x}) = \frac{1}{1+e^{-\bar{\mathbf{w}}^T\bar{\mathbf{x}}}}$$

donde $\mathsf{p}(\mathcal{C}_1|\mathbf{x}) \equiv \mathsf{p}_{\mathcal{C}_1/\mathbf{x}}(\mathbf{x})$ y $\ln\frac{a}{1-a}$ es la función *link*.

Como es conocido, el modelo logístico es válido para la familia exponencial:

$$\mathsf{p}(\mathbf{x}|\mathcal{C}_j) = \exp\{A(\boldsymbol{\theta}_j) + B(\mathbf{x},\boldsymbol{\alpha}) + \boldsymbol{\theta}_j^T\mathbf{x}\},\ j = 1,2$$

que, sustituyéndola en el teorema de Bayes:

$$\mathsf{p}(\mathcal{C}_1|\mathbf{x}) = \frac{\mathsf{p}(\mathbf{x}|\mathcal{C}_1)\mathsf{p}(\mathcal{C}_1)}{\mathsf{p}(\mathbf{x}|\mathcal{C}_1)\mathsf{p}(\mathcal{C}_1)+\mathsf{p}(\mathbf{x}|\mathcal{C}_2)\mathsf{p}(\mathcal{C}_2)} = \frac{1}{1+e^{-\bar{\mathbf{w}}^T\bar{\mathbf{x}}}}$$

produce como resultado (p. ej. [Bishop 1995, p. 234]): $\mathbf{w} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ y $w_0 = A(\boldsymbol{\theta}_1) - A(\boldsymbol{\theta}_2) + \ln\frac{\mathsf{p}(\mathcal{C}_1)}{\mathsf{p}(\mathcal{C}_2)}$. En particular, en hipótesis de normalidad:

$$\mathsf{p}(\mathbf{x}|\mathcal{C}_j) = (2\pi)^{-d/2}\,|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\},\ j = 1,2$$

mínimos cuadrados puede no ser la mejor idea para seleccionar el parámetro de suavizado h de una regla de clasificación basada en estimadores tipo núcleo.

se obtiene:

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{\mathsf{p}(\mathcal{C}_1)}{\mathsf{p}(\mathcal{C}_2)} \end{aligned}$$

con lo que la red, una vez entrenada, proporcionará estimadores de dichos parámetros[45].

La metodología anterior puede generalizarse a otros modelos lineales generalizados, utilizando como función de activación la inversa de la función *link*. Por ejemplo, si $\psi(r) = \Phi(r) = \frac{2}{\sqrt{\pi}}\int_0^r e^{-t^2}dt$, función de distribución gausiana, (*error function*) se tiene el modelo **probit**:

$$f_{\bar{\mathbf{w}}}(\mathbf{x}) = \Phi(\bar{\mathbf{w}}^T\bar{\mathbf{x}})$$

orientado a la estimación de frecuencias binomiales que dependen del nivel $\mathbf{x}$ de un conjunto de factores, (p. ej. [McCullagh y Nelder 1983]).

Función de Pérdida

En lo anterior no se indicó la función de pérdida a utilizar en el riesgo empírico durante el entrenamiento. En principio, con el apoyo del teorema 33, puede utilizarse la pérdida cuadrática. Además, la función de activación logística, garantiza que $\hat{f}_n(\mathbf{x}) \in [0,1]$, aunque se pierde la ventaja de un problema de optimización cuadrático. Por otra parte, la pérdida cuadrática ignora la distribución de los datos, con lo que queda en cuestión la eficiencia de los estimadores.

Una alternativa más fundamentada consiste en utilizar la función de verosimilitud basada en la distribución condicionada:

$$\mathsf{p}_{\mathbf{Y}^n/\mathbf{z}^n}(\mathbf{y}^n;\mathbf{x}^n) = \prod_{i=1}^n \mathsf{f}(\mathbf{x}_i)^{y_i}[1-\mathsf{f}(\mathbf{x}_i)]^{1-y_i}$$

que da lugar a la función de pérdida de entropía cruzada (I.3):

$$\ell(y_i, f(\mathbf{x}_i)) = -[y_i \ln f_{\bar{\mathbf{w}}}(\mathbf{x}_i) + (1-y_i)\ln(1-f_{\bar{\mathbf{w}}}(\mathbf{x}_i))]$$

y a la función objetivo de entropía cruzada a minimizar:

$$\hat{R}'_n(\bar{\mathbf{w}}) = -\frac{1}{n}\sum_{i=1}^n [y_i \ln f_{\bar{\mathbf{w}}}(\mathbf{x}_i) + (1-y_i)\ln(1-f_{\bar{\mathbf{w}}}(\mathbf{x}_i))]$$

Este riesgo empírico se corrige usualmente restándole el valor óptimo, para que posea un valor mínimo nulo. La condición de óptimo es:

$$\frac{\partial \hat{R}'_n}{\partial f_{\bar{\mathbf{w}}}(\mathbf{x}_i)} = \frac{f_{\bar{\mathbf{w}}}(\mathbf{x}_i) - y_i}{f_{\bar{\mathbf{w}}}(\mathbf{x}_i)(1-f_{\bar{\mathbf{w}}}(\mathbf{x}_i))} = 0 \Rightarrow f_{\bar{\mathbf{w}}}(\mathbf{x}_i) = y_i,\ i = 1:n$$

que es mínimo global y para una codificación $\mathcal{Y} = \{0,1\}$, produce ya valor mínimo nulo. Sin embargo, este no es el caso en un problema con observaciones repetidas donde se utiliza una codificación continua que refleje, para cada $\mathbf{x}_i$, las frecuencias relativas de pertenencia a la clase

[45] A pesar de que la expresión anterior coincide con el discriminante de Fisher, en la práctica, no produce los mismos resultados, pues mientras aquél se estima mediante la distribución conjunta, éste se estima con la distribución condicionada, por lo que es menos eficiente al utilizar una menor cantidad de información, ([Seber 1984, p. 317], [Ripley 1996, p. 45] citando a [Efron 1975]).

USC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

$\mathcal{C}_1$, en cuyo caso, el mínimo sería distinto para cada muestra. De esta forma, suele utilizarse la función objetivo corregida:

$$\hat{R}_n(\bar{\mathbf{w}}) = \hat{R}'_n(\bar{\mathbf{w}}) - \hat{R}'^*_n(\bar{\mathbf{w}}) = \frac{1}{n}\sum_{i=1}^{n}[y_i \ln \frac{y_i}{f_{\bar{\mathbf{w}}}(\mathbf{x}_i)} + (1-y_i)\ln \frac{1-y_i}{1-f_{\bar{\mathbf{w}}}(\mathbf{x}_i)}] \tag{II.12}$$

que es la versión empírica de la distancia de Kullback-Leibler ya vista con anterioridad.

La utilización de esta función objetivo basada en la verosimilitud asegura bajo condiciones de regularidad, que el estimador es eficiente y función de un estimador suficiente si éste existe, además de consistente y asintóticamente normal, como ya garantizaba el estimador de mínimos cuadrados.

Para esta pérdida y otras γ-admisibles (en particular la cuadrática en un dominio compacto), puede llegarse a la consistencia del principio de minimización del riesgo empírico utilizando la finitud de la dimensión P de la familia de neuronas con activación sigmoide, (teorema 225 del Anexo G, o combinando los teoremas 220 y 224 de dicho anexo).

Por otra parte, esta función objetivo es más robusta que la de mínimos cuadrados, ya que denotando $f_{\bar{\mathbf{w}}}(\mathbf{x}_i) = y_i + e_i$ siendo e_i el error cometido por $f_{\bar{\mathbf{w}}}(\mathbf{x}_i)$, $i = 1 : n$, la expresión (II.12) resulta:

$$\hat{R}_n(\bar{\mathbf{w}}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \ln \frac{e_i}{y_i} + (1-y_i)\ln \frac{1-e_i}{1-y_i}]$$

que depende de los errores relativos e_i/y_i cometidos por la red, en lugar de los errores *absolutos* $[y_i - f_{\bar{\mathbf{w}}}(\mathbf{x}_i)]^2 = e_i^2$ como en el caso de la pérdida cuadrática, la cuál, recibe mayor contribución de los casos con mayores errores, independientemente del tamaño de las posteriores $\mathsf{P}(\mathcal{C}_j|\mathbf{x})$ $j = 1, 2$, a estimar. Por el contrario, la entropía cruzada equilibra el peso de los errores dando mayor peso a los errores pequeños y estimando así mejor las probabilidades pequeñas.

Algoritmo

Al incluirse ahora en la función objetivo funciones de activación ψ no lineales, el problema de optimización ya no es cuadrático (incluso con pérdida cuadrática), con lo que se hacen necesarios algoritmos de optimización no lineal.

Por ejemplo, el algoritmo del gradiente (Anexo D), produce la regla delta ya vista con anterioridad. En entrenamiento interactivo:

$$\bar{\mathbf{w}}(s+1) = \bar{\mathbf{w}}(s) - \rho(s)\nabla_{\bar{\mathbf{w}}}\ell_i(\bar{\mathbf{w}}(s))$$

y en entrenamiento *batch*:

$$\begin{aligned} \bar{\mathbf{w}}(s+1) &= \bar{\mathbf{w}}(s) - \rho(s)\nabla_{\bar{\mathbf{w}}}\hat{R}_n(\bar{\mathbf{w}}(s)) \\ &= \bar{\mathbf{w}}(s) - \rho(s)\sum_{i=1}^{n}\nabla_{\bar{\mathbf{w}}}\ell_i(\bar{\mathbf{w}}(s)) \end{aligned} \tag{II.13}$$

1. En el caso de mínimos cuadrados, $\ell_i(\bar{\mathbf{w}}) = \ell(y_i, f_{\bar{\mathbf{w}}}(\mathbf{x}_i)) = [y_i - \psi(\bar{\mathbf{x}}_i^T\bar{\mathbf{w}})]^2$, el gradiente es:

$$\nabla_{\bar{\mathbf{w}}}\ell_i(\bar{\mathbf{w}}(s)) = [y_i - \psi(\bar{\mathbf{x}}_i^T\bar{\mathbf{w}})]\dot{\psi}(\bar{\mathbf{x}}_i^T\bar{\mathbf{w}})\bar{\mathbf{x}}_i = \delta_i\bar{\mathbf{x}}_i, \text{ con } \delta_i = [y_i - \psi(\bar{\mathbf{x}}_i^T\bar{\mathbf{w}})]\dot{\psi}(\bar{\mathbf{x}}_i^T\bar{\mathbf{w}})$$

donde, si ψ es la función logística, $\dot{\psi}(r) = \psi(r)(1-\psi(r))$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2. En el caso de la entropía cruzada:

$$\ell_i(\bar{\mathbf{w}}) = \ell(y_i, f_{\bar{\mathbf{w}}}(\mathbf{x}_i)) = y_i \ln \frac{y_i}{f_{\bar{\mathbf{w}}}(\mathbf{x}_i)} + (1-y_i) \ln \frac{1-y_i}{1-f_{\bar{\mathbf{w}}}(\mathbf{x}_i)}$$

y, con activación logística y haciendo $r_i = \bar{\mathbf{x}}_i^T \bar{\mathbf{w}}$, el gradiente es:

$$\begin{aligned} \nabla_{\bar{\mathbf{w}}} \ell_i(\bar{\mathbf{w}}(s)) &= \frac{\partial \ell_i(\bar{\mathbf{w}}(s))}{\partial f_{\bar{\mathbf{w}}}(\mathbf{x}_i)} \frac{\partial f_{\bar{\mathbf{w}}}(\mathbf{x}_i)}{\partial r_i} \frac{\partial r_i}{\partial \bar{\mathbf{w}}} \\ &= \frac{f_{\bar{\mathbf{w}}}(\mathbf{x}_i) - y_i}{f_{\bar{\mathbf{w}}}(\mathbf{x}_i)[1 - f_{\bar{\mathbf{w}}}(\mathbf{x}_i)]} f_{\bar{\mathbf{w}}}(\mathbf{x}_i)[1 - f_{\bar{\mathbf{w}}}(\mathbf{x}_i)] \bar{\mathbf{x}}_i \\ &= [f_{\bar{\mathbf{w}}}(\mathbf{x}_i) - y_i] \bar{\mathbf{x}}_i = \delta_i \bar{\mathbf{x}}_i, \text{ con } \delta_i = [f_{\bar{\mathbf{w}}}(\mathbf{x}_i) - y_i] \end{aligned}$$

Una manera fundamentada de elegir el parámetro $\rho(s)$ (*learning rate*) es (Anexo D):

$$\rho(s) = \frac{\left\| \nabla_{\bar{\mathbf{w}}} \hat{R}_n \right\|^2}{\nabla_{\bar{\mathbf{w}}} \hat{R}_n^T \mathbf{H} \nabla_{\bar{\mathbf{w}}} \hat{R}_n}$$

II.2.4 Clasificación con Varias Clases

Clases Mútuamente Excluyentes

El método implementado por la red adaline con activación logística puede generalizarse a un problema de clasificación con c clases $\mathcal{C}_1, ..., \mathcal{C}_c$, $c \geq 2$ exhaustivas y excluyentes, considerando una red con tantas salidas como clases, (ahora $\mathsf{Y} = (\mathsf{Y}_1, ..., \mathsf{Y}_c)^T \in \mathcal{Y}^c$), con la codificación $y_j = 1$ si $\mathbf{x}$ está en la clase $\mathcal{C}_j$, y cero en caso contrario, $j = 1, 2, ..., c$, y construyendo una regla de decisión emuladora de la de Bayes (enfoque *one-against-all*):

$$g_{\mathbf{f}}(\mathbf{x}) = \arg \max_{j \in \{1:c\}} f_j(\mathbf{x})$$

donde $\mathbf{f} = (f_1, ..., f_c)$ y cada f_j es una función discriminante para la clase $\mathcal{C}_j$.

Se podría intentar resolver el problema mediante c problemas paralelos de regresión logística, utilizando c neuronas adaline con activación logística (figura 9), pues ello produciría otras tantas estimaciones de las probabilidades a posteriori de las clases:

$$\hat{\mathsf{p}}(\mathcal{C}_j|\mathbf{x}) = \hat{f}_{n,j}(\mathbf{x}; \bar{\mathbf{w}}_j) = e^{\bar{\mathbf{w}}_j^T \bar{\mathbf{x}}} / (1 + e^{\bar{\mathbf{w}}_j^T \bar{\mathbf{x}}}) \in [0,1], j = 1, ..., c$$

Sin embargo, la resolución independiente de estos c problemas es incompatible con la dependencia existente entre las variables Y_j, lo que puede violar la propiedad $\sum_{j=1}^c \hat{f}_{n,j}(\mathbf{x}; \bar{\mathbf{w}}_j) = 1$. Ello se evita sustituyendo la función de activación logística de cada neurona, por la función de activación denominada **softmax** en la literatura de redes neuronales, [Bridle 1990]:

$$\psi(\bar{\mathbf{w}}_j^T \bar{\mathbf{x}}) = \frac{\exp\{\bar{\mathbf{w}}_j^T \bar{\mathbf{x}}\}}{\sum_{i=1}^c \exp\{\bar{\mathbf{w}}_i^T \bar{\mathbf{x}}\}}, \; j = 1, ..., c \tag{II.14}$$

pero ya utilizada antes en el modelo de regresión logística múltiple que generaliza el modelo de dos clases, (p. ej. [Seber 1984, p. 335], [McCullagh y Nelder 1983]).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

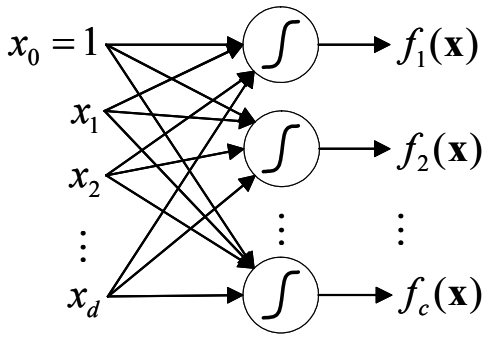


Figura 9. Red neuronal de una sóla capa con neuronas adaline con activación logística.

Pérdida de Entropía Cruzada

Este modelo es válido para la familia exponencial, $\mathsf{p}(\mathbf{x}|\mathcal{C}_j) = \exp\{A(\boldsymbol{\theta}_j) + B(\mathbf{x}, \boldsymbol{\alpha}) + \boldsymbol{\theta}_j^T \mathbf{x}\}$, $j = 1, ..., c$, pues, sustituyendo dicha densidad en el teorema de Bayes, se obtiene (II.14) [Bishop 1995, p. 239], con:

$$\begin{aligned} \mathbf{w}_j &= \boldsymbol{\theta}_j \\ w_{j0} &= A(\boldsymbol{\theta}_j) + \ln \mathsf{p}(\mathcal{C}_j) \end{aligned}$$

La red $\mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x})$ posee ahora c salidas $(\mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x}))_j = f(\mathbf{x}; \bar{\mathbf{w}}_j) = f_{\bar{\mathbf{w}}_j}(\mathbf{x})$, $j = 1, ..., c$, con $\bar{\mathbf{W}} = (\bar{\mathbf{w}}_1, ..., \bar{\mathbf{w}}_c)$. En este caso, la función de verosimilitud para una observación $\mathbf{y}$ en $\mathbf{x}$ puede escribirse como:

$$q(\mathbf{y}; \mathbf{x}, \bar{\mathbf{W}}) = \prod_{j=1}^{c} f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i)^{y_j}$$

que es igual a $f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i)$ si $\mathbf{x} \in \mathcal{C}_j$. Para n observaciones independientes $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, denotando por $\mathbf{Y}^n = (\mathbf{y}_1, ..., \mathbf{y}_n)^T$ la matriz $n \times c$ con elementos $y_{ij} \doteq (\mathbf{Y})_{ij} = 1$ si el dato $\mathbf{x}_i$ está en la clase j y cero en caso contrario, se tiene:

$$q(\mathbf{Y}^n; \mathbf{x}^n, \bar{\mathbf{W}}) = \prod_{i=1}^{n} q(\mathbf{y}_i; \mathbf{x}_i, \bar{\mathbf{W}}) = \prod_{i=1}^{n} \prod_{j=1}^{c} f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i)^{y_{ij}}$$

que da lugar a la función objetivo:

$$\hat{R}'_n(\bar{\mathbf{W}}) = \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} \ln f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i)$$

y, utilizando el mismo argumento que en el apartado anterior, a la función objetivo corregida cuyo mínimo es siempre cero:

$$\hat{R}_n(\bar{\mathbf{W}}) = \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} \ln \frac{y_{ij}}{f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i)}$$

Pérdida Cuadrática

La utilización de mínimos cuadrados no aporta beneficios en caso de utilizar la activación logística en cada neurona de la arquitectura anterior, pues sigue siendo imposible su resolución analítica.

Otra alternativa es abordar el problema mediante c problemas paralelos de regresión tipo adaline con activación $\psi(r) = \text{signo}(r)$, estimándose con cada uno las probabilidades a posteriori $\mathsf{p}(\mathcal{C}_j|\mathbf{x})$. Ahora, se obtendría la matriz óptima de coeficientes:

$$\bar{\mathbf{W}} = (\bar{\mathbf{X}}^T\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}^T\mathbf{Y}$$

El problema de este enfoque es que, como ya hemos indicado al tratar este modelo para dos clases, la pérdida cuadrática no garantiza $\hat{f}_{n,j}(\mathbf{x}) \in [0,1]$. Sin embargo, sí se obtiene en este caso la propiedad $\sum_{j=1}^{c} \hat{f}_{n,j}(\mathbf{x}) = 1$ sin necesidad de utilizar la función softmax o equivalente. Ello es debido al siguiente teorema:

Teorema 34 *[Lowe y Webb 1991] Sea una red* $\mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x}) = \mathbf{W}^T\mathbf{x} + \mathbf{w}_0 = \bar{\mathbf{W}}\bar{\mathbf{x}}$ *entrenada con mínimos cuadrados a partir de una muestra* $\mathbf{z}^n$ *con* $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$, *con* $\mathbf{x}_i \in \mathbb{R}^d$, $\mathbf{y}_i \in \mathbb{R}^c$, $i = 1:n$. *Si los datos de la muestra satisfacen la restricción* $\boldsymbol{\alpha}^T\mathbf{y}_i = \boldsymbol{\alpha}^T\bar{\mathbf{y}}$, $i = 1:n$, *con* $\boldsymbol{\alpha} \in \mathbb{R}^c$ *siendo* $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$, *entonces, la salida de la red satisface la misma restricción:* $\boldsymbol{\alpha}^T\mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x}_i) = \boldsymbol{\alpha}^T\bar{\mathbf{y}}$, $i = 1:n$.

En el problema de clasificación de c clases con la codificación que venimos utilizando, se tiene:

$$\sum_{j=1}^{c} y_{ij} = \mathbf{1}_c^T\mathbf{y}_i = 1,\ \forall i = 1:n$$

donde $\mathbf{1}_c = (1,...,1)^T \in \mathbb{R}^c$ con lo que, utilizando el teorema anterior, resulta:

$$\mathbf{1}^T\mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x}_i) = \sum_{j=1}^{c}(\mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x}_i))_j = 1,\ \forall i = 1:n$$

Este teorema es también aplicable a una codificación en forma de frecuencias relativas en el caso de observaciones repetidas: $y_j(\mathbf{x}) = n_j(\mathbf{x})/n(\mathbf{x})$ donde $n_j(\mathbf{x})$ es el número de casos en la clase $\mathcal{C}_j$, pues también se cumple $\sum_{j=1}^{c} y_j(\mathbf{x}) = 1$.

Atributos Independientes

El problema de clasificación anterior suponía clases mútuamente excluyentes. La consideración de dos clases no excluyentes e independientes, lleva directamente al planteamiento de sendos problemas de clasificación con dos clases excluyentes: $\mathcal{C}_j$ y su complementario, $j = 1, 2$, análogos al tratado anteriormente.

La misma generalización puede realizarse a c clases $\mathcal{C}_j$ no excluyentes e independientes. Utilizando la misma codificación que en el epígrafe anterior, ahora la función de verosimilitud para una observación $(\mathbf{x}, \mathbf{y})$ es:

$$q(\mathbf{y}; \mathbf{x}, \bar{\mathbf{W}}) = \prod_{j=1}^{c} q(y_j; \mathbf{x}, \bar{\mathbf{w}}_j) = \prod_{j=1}^{c} f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i)^{y_j}(1 - f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i))^{1-y_j}$$

donde hemos utilizado la independencia entre $\mathsf{Y}_1, ..., \mathsf{Y}_c$. Análogamente, para n observaciones independientes $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ y utilizando la misma notación que en el epígrafe anterior:

$$q(\mathbf{Y}^n; \mathbf{x}^n, \bar{\mathbf{W}}) = \prod_{i=1}^{n} q(\mathbf{y}_i; \mathbf{x}_i, \bar{\mathbf{W}}) = \prod_{i=1}^{n}\prod_{j=1}^{c} f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i)^{y_{ij}}(1 - f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i))^{1-y_{ij}}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

que da lugar a la función de pérdida:

$$\hat{R}'_n(\bar{\mathbf{W}}) = \sum_{i=1}^{n}\sum_{j=1}^{c}[y_{ij}\ln f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i) + (1-y_{ij})\ln(1-f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i))]$$

y, utilizando el mismo argumento que en el epígrafe anterior, a la siguiente función objetivo corregida, cuyo mínimo es siempre cero:

$$\hat{R}_n(\bar{\mathbf{W}}) = \sum_{i=1}^{n}\sum_{j=1}^{c}[y_{ij}\ln\frac{y_{ij}}{f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i)} + (1-y_{ij})\ln\frac{1-y_{ij}}{1-f_{\bar{\mathbf{w}}_j}(\mathbf{x}_i)}]$$

Como puede observarse, la minimizacion de esta función objetivo equivale a la minimización, para cada clase j-ésima, de la suma en i, con lo que el problema se reduce a c problemas de dos clases excluyentes cada uno.

II.2.5 Funciones de Pérdida y Salidas como Probabilidades a Posteriori

Es interesante el siguiente análisis de [Hampshire y Pearlmutter 1990], sobre las condiciones a satisfacer por una función de pérdida para que las salidas de una red entrenada puedan interpretarse como probabilidades a posteriori.

En lo que sigue, seguimos utilizando la notación del epígrafe anterior, es decir $y_j = 1$ si $\mathbf{x}$ está en la clase $\mathcal{C}_j$ y cero en caso contrario, así como una red neuronal con salida $\mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x}) \in [0,1]^c$, donde denotamos $f_{ij} = (\mathbf{f}_{\bar{W}}(\mathbf{x}_i))_j$.

El análisis es válido para problemas de clasificación con principio inductivo basado en la minimización de una función objetivo aditiva en las observaciones y en las variables de salida:

$$\hat{R}_n(\mathbf{f}_{\bar{W}}) = \sum_{i=1}^{n}\sum_{j=1}^{c}\ell_j(y_{ij}, f_{ij}) = \sum_{i=1}^{n} L(\mathbf{y}_i, \mathbf{f}_{\bar{W}}(\mathbf{x}_i))$$

donde $L(\mathbf{y}_i, \mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x}_i)) \doteq \sum_{j=1}^{c}\ell_j(y_{ij}, f_{ij})$ es la contribución de la observación i-ésima. En el marco de la verosimilitud, esto equivale a observaciones y atributos independientes, por lo que la función softmax no se admite aquí.

Se trata de establecer condiciones para que una función de pérdida de la forma: $\ell(y_{ij}, f_{ij}) = l(|y_{ij} - f_{ij}|)$ permita interpretar las salidas f_{ij} como estimaciones de las probabilidades a posteriori.

El teorema de la probabilidad total permite escribir:

$$\mathsf{p}(y_j|\mathbf{x}) = \sum_{l=1}^{c}\mathsf{p}(y_j|\mathcal{C}_l)\mathsf{p}(\mathcal{C}_l|\mathbf{x}) = \sum_{l=1}^{c}\delta(y_j - \delta_{lj})\mathsf{p}(\mathcal{C}_l|\mathbf{x})$$

donde $\mathsf{p}(y_j|\mathcal{C}_l) \doteq \mathsf{p}_{\mathsf{Y}_j/\mathcal{C}_l}(y_j)$, δ es la delta de Kronecker, y $\delta_{lj} = 1$ si $l = j$ y cero en caso contrario, ya que $\mathsf{p}(1|\mathcal{C}_l) = 1$ si $j = l$ y $\mathsf{p}(0|\mathcal{C}_l) = 1$ si $j \neq l$.

Con ello, debido a la independencia de los atributos, podemos escribir:

$$\mathsf{p}(\mathbf{y}|\mathbf{x}) = \prod_{j=1}^{c}\sum_{l=1}^{c}\delta(y_j - \delta_{lj})\mathsf{p}(\mathcal{C}_l|\mathbf{x})$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Utilizando este resultado y denotando $f_j(\mathbf{x}) = (\mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x}))_j$, se obtiene:

$$\begin{aligned}\mathbb{E}_{\mathsf{Z}}[L(\mathbf{y}, \mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x}))] &= \sum_{j=1}^{c} \int\int l(|y_j - f_j(\mathbf{x})|)\mathsf{p}(\mathbf{y}|\mathbf{x})\mathsf{p}(\mathbf{x})d\mathbf{y}d\mathbf{x} \\ &= \sum_{j=1}^{c} \int \{l(1 - f_j(\mathbf{x}))\mathsf{p}(\mathcal{C}_j|\mathbf{x}) + l(f_j(\mathbf{x}))[1 - \mathsf{p}(\mathcal{C}_j|\mathbf{x})]\}\mathsf{p}(\mathbf{x})d\mathbf{x}\end{aligned}$$

con lo que, la condición de óptimo para $\mathbf{f}_{\bar{\mathbf{W}}}(\mathbf{x})$ resulta, para $j = 1 : c$:

$$\begin{aligned}\frac{\partial}{\partial f_j(\mathbf{x})}\mathbb{E}_{\mathsf{Z}}[L(\mathbf{y}, \mathbf{f}_{\bar{W}}(\mathbf{x}))] &= -\dot{l}(1 - f_j(\mathbf{x}))\mathsf{p}(\mathcal{C}_j|\mathbf{x}) + \dot{l}(f_j(\mathbf{x}))[1 - \mathsf{p}(\mathcal{C}_j|\mathbf{x})] = 0 \\ \Rightarrow \quad \frac{\dot{l}(1 - f_j(\mathbf{x}))}{\dot{l}(f_j(\mathbf{x}))} &= \frac{1 - \mathsf{p}(\mathcal{C}_j|\mathbf{x})}{\mathsf{p}(\mathcal{C}_j|\mathbf{x})} \qquad \text{(II.15)}\end{aligned}$$

donde $\dot{l}(r) = dl(r)/dr$.

Por tanto, para que las salidas de la red representen probabilidades a posteriori, es decir, $f_j(\mathbf{x}) = \mathsf{p}(\mathcal{C}_j|\mathbf{x})$, $j = 1 : c$, es condición necesaria que:

$$\frac{\dot{l}(1 - f_j(\mathbf{x}))}{\dot{l}(f_j(\mathbf{x}))} = \frac{1 - f_j(\mathbf{x})}{f_j(\mathbf{x})} \qquad \text{(II.16)}$$

Una clase de funciones de pérdida que satisface esta condición es:

$$l(r) = \int r^s(1 - r)^{s-1}dr, \; s \in \mathbb{R}$$

que incluye como casos particulares relevantes, los casos:

$$\begin{cases} s = 0, \text{ entropía cruzada:} & l(r) = -\ln(1 - r) \\ s = 1, \text{ pérdida cuadrática:} & l(r) = \frac{r^2}{2} \end{cases}$$

pues, en el primer caso:

$$\begin{aligned}&\text{si } y = 1\text{:} \quad l(|y - f_j(\mathbf{x})|) = l(1 - f_j(\mathbf{x})) = -\ln f_j(\mathbf{x}) = -y \ln f_j(\mathbf{x}) \\ &\text{si } y = 0\text{:} \quad l(|y - f_j(\mathbf{x})|) = l(f_j(\mathbf{x})) = -\ln(1 - f_j(\mathbf{x})) = -(1 - y)\ln(1 - f_j(\mathbf{x}))\end{aligned}$$

Para el caso general de la pérdida de Minkowski: $\ell(r) = r^s$, si ésta se sustituye en (II.16), se obtiene: $r^{s-2} = (1 - r)^{s-2}$, lo que se verifica sólo si $r = 2$, correspondiente a la pérdida cuadrática. De cualquier forma, la utilización de otros valores de s, produce reglas que minimizan la probabilidad de error, pues, sustituyendo dicha pérdida en (II.15), se obtiene:

$$f_j(\mathbf{x}) = \frac{\mathsf{p}(\mathcal{C}_j|\mathbf{x})^{1/(s-1)}}{\mathsf{p}(\mathcal{C}_j|\mathbf{x})^{1/(s-1)} + [1 - \mathsf{p}(\mathcal{C}_j|\mathbf{x})]^{1/(s-1)}}$$

donde, para $s = 2$, se tiene la probabilidad a posteriori pero, al ser $f_j(\mathbf{x})$ función monótona de $\mathsf{p}(\mathcal{C}_j|\mathbf{x})^{1/(s-1)}$, las fronteras de decisión se corresponden con la regla que minimiza la probabilidad de error de clasificación para cualquier valor de s.

Finalmente, el análisis anterior supuso c atributos independientes, con lo que, es válido para el caso de un sólo atributo, que se corresponde con el caso de dos clases mútuamente excluyentes.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

II.2.6 Redes Neuronales de Una Capa para el Problema de Regresión

La estimación de las probabilidades a posteriori es un enfoque al problema de clasificación basado en el método de regresión. Salvo la red adaline con activación sigmoide, históricamente no se han desarrollado redes de una sóla capa con otras activaciones no lineales, probablemente, porque las redes que podían resultar de interés producían modelos paramétricos de sobra conocidos en estadística. Por ejemplo, una neurona adaline con una función de activación identidad ψ produce el modelo lineal de regresión:

$$f(\mathbf{x}) = \psi(\bar{\mathbf{w}}^T\bar{\mathbf{x}}) = \bar{\mathbf{w}}^T\bar{\mathbf{x}}$$

y con una función de activación ψ polinómica produce un modelo polinómico:

$$f(\mathbf{x}) = \psi(\bar{\mathbf{w}}^T\bar{\mathbf{x}}) = (\bar{\mathbf{w}}^T\bar{\mathbf{x}})^2 = w_0^2 + 2w_0w_1x_1 + 2w_0w_2x_2 + 2w_1w_2x_1x_2 + w_1^2x_1^2 + w_2^2x_2^2$$

Una mayor flexibilidad se obtiene previa transformación no lineal de las variables de entrada, lo que requiere un nivel oculto adicional. Esta técnica era también conocida de la estadística, pero se utilizaba más bien para el pre-proceso de datos con objetivos muy específicos, (normalidad, homocedasticidad, incorrelación, etc.), o para reducir el problema a uno con interacciones aditivas en las covariables, pero no como una técnica para resolver problemas generales.

II.3 Arquitecturas Multicapa Hacia Adelante (MLF)

Si a una red (neurona) adaline con activación sigmoide, se le añade una nueva capa con una neurona con activación dicotómica para implementar la regla de decisión como lo hace el perceptrón, aquélla pasaría a ser una red neuronal con una capa oculta, implementando una regla de decisión lineal.

Es claro que, no es equitativa la comparación realizada entre las redes que implementan una regla de clasificación (perceptrón y adaline con activación dicotómica), y las que implementan funciones discriminantes que sirven de base para una regla posterior, (adaline con activación sigmoide). Si la comparación se realizase en igualdad de condiciones, éstas últimas deberían incorporar una nueva capa con una neurona de activación dicotómica que implementase la regla de clasificación.

En la literatura, se denomina **perceptrón multicapa** (*multilayer perceptron*), a las redes con funciones de activación sigmoide en las unidades ocultas y activaciones dicotómicas, sigmoides o lineales en el nivel de salida. Sin embargo, igual que antes, es justo distinguir entre las que poseen activaciones dicotómicas en salida de las que poseen salida continua pues, estas últimas, en un problema de clasificación, requerirían una capa adicional para implementar la regla de clasificación.

Con esta consideración en mente, en lo que sigue nos centramos en las redes **redes multicapa** *feed-forward* (MLF) con un nivel oculto y un nivel de salida con activación continua, (sigmoide o lineal), (que denominaremos igualmente MLF, figura 10), pues, en general, para problemas de regresión y clasificación, (al contrario que en problemas de reducción de dimensionalidad), un nivel oculto suele ser suficiente para los problemas usuales de aplicación y capas adicionales no aportan beneficios añadidos[46].

[46] Aunque, para dichos tipos de problemas, queda mucho por conocer sobre las redes con un mayor número de capas.

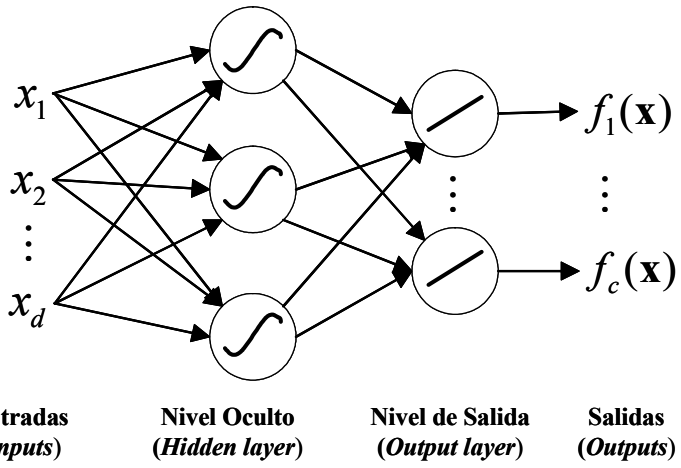


Figura 10. Arquitectura multicapa *feed-forward* con una capa oculta y dos neuronas en el nivel de salida.

Así pues, el modelo general de la **red neuronal multicapa** *feed-forward*, (MLF), es el siguiente (una sóla salida):

$$f(\mathbf{x};\boldsymbol{\theta}) = \varphi(\sum_{j=1}^{h} c_j \psi_j(\mathbf{x};\mathbf{w}_j) + c_0) = \varphi(\mathbf{c}^T\boldsymbol{\psi} + c_0) = \varphi(\bar{\mathbf{c}}^T\bar{\boldsymbol{\psi}}) \tag{II.17}$$

donde φ es una función continua, $\boldsymbol{\psi} = (\psi_1, ..., \psi_h)^T$ con $\psi_j = \psi_j(\mathbf{x};\mathbf{w}_j)$, $\mathbf{c} = (c_1, ..., c_h)^T$, $\bar{\boldsymbol{\psi}} = (1, \boldsymbol{\psi}^T)^T$, $\bar{\mathbf{c}} = (c_0, \mathbf{c}^T)^T$ y $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ es el conjunto de todos los parámetros de la red incluyendo el número h de unidades ocultas.

En esta sección y las dos siguientes analizaremos las arquitecturas MLF más frecuentes. En una sección posterior, presentaremos los algoritmos más apropiados para ellas.

II.3.1 Arquitecturas MLF Lineales

Antes de analizar arquitecturas MLF más generales, presentamos brevemente un caso particular de (II.17) en el que las funciones de activación son todas lineales, debido a que, al no incluir transformación no lineal alguna en el modelo, su problemática puede servir de antesala a las demás arquitecturas con un nivel oculto, en varios sentidos:

1. Muchas redes con activación no lineal en el nivel oculto poseen activación lineal en salida, lo que permite entender mejor el papel de cada nivel.

2. La problemática y funcionamiento de las redes más complejas se investiga muchas veces mediante aproximación lineal.

3. El algoritmo de aprendizaje de las redes no lineales se inicia normalmente con valores iniciales pequeños en los parámetros, lo que suele suponer que, en las primeras etapas del proceso, la red opera en un régimen cuasi-lineal.

   Por ejemplo, con función de activación logística $\psi(\mathbf{w}^T\mathbf{x}+w_0)$ (figura 11), si las componentes de $\mathbf{w}$ y w_0 son pequeñas, el valor resultante de $\mathbf{w}^T\mathbf{x} + w_0$ es pequeño, con lo que la logística opera en la zona próxima a cero donde es aproximadamente lineal.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

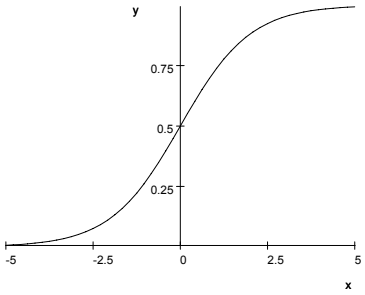


Figura 11. Función Logística: $\psi(r) = (1 + \exp(-r))^{-1}$

Incluso, finalizado el entrenamiento, este régimen puede existir para parámetros con valor muy pequeño (resultado de la poca influencia de los factores correspondientes). De hecho, la complejidad del modelo está determinada por el número de parámetros no nulos.

Excepcionalmente, para esta rápida exposición, suponemos varias salidas $\mathbf{y} = (y_1, ..., y_c)^T$, $\mathbf{f} = (f_1, ..., f_c)^T$, con lo que, utilizando notación estadística estándar[47], formulamos esta arquitectura como:

$$\mathbf{f}(\mathbf{x}; \mathbf{W}, \mathbf{C})^T = \mathbf{x}^T\mathbf{W}\mathbf{C} = \mathbf{x}^T\bar{\mathbf{B}} \tag{II.18}$$

donde $\mathbf{W}_{d\times h}$, $\mathbf{C}_{h\times c}$ y $\bar{\mathbf{B}}_{d\times c}$ son matrices. Estudiamos específicamente el caso en que $h < c \leq d$, es decir, el nivel oculto produce reducción de dimensionalidad y, así, posible pérdida de información.

Sobre esta arquitectura es preciso notar que:

1. A pesar de que el modelo es lineal en $\mathbf{x}$, es no lineal en los parámetros.

2. Las matrices $\mathbf{W}, \mathbf{C}$ no son únicas, ya que, si $\mathbf{G}$ es una matriz no singular, $\bar{\mathbf{B}} = \mathbf{W}\mathbf{C} = (\mathbf{W}\mathbf{G}^{-1})(\mathbf{G}\mathbf{C})$.

3. El problema no es, en general, equivalente al de regresión lineal $\mathbf{y}^T = \mathbf{x}^T\mathbf{B}$, con solución $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, donde $\mathbf{X}$ e $\mathbf{Y}$ son matrices $n \times d$ y $n \times c$ respectivamente, con las variables en columna. Ello se debe a que, en las mejores condiciones, $\text{rango}(\hat{\mathbf{B}}) = c$, mientras que, si $h < c$, $\text{rango}(\bar{\mathbf{B}}) = \text{rango}(\mathbf{W}\mathbf{C}) \leq h < c$, con lo que $\hat{\mathbf{B}}$ no puede ser solución del problema planteado.

   De hecho, la reducción de dimensionalidad hace perder información sobre $\mathbf{x}$ y, así, el ajuste es subóptimo en relación con el que puede conseguir el problema lineal sin restricciones (respecto a mínimos cuadrados).

USC

[47] En la literatura de redes neuronales es usual plantear el modelo como $\mathbf{f} = \bar{\mathbf{B}}^T\mathbf{x}$, tal como hemos venido haciendo, pero, su análisis, produciría expresiones menos familiares para el estadístico.

Análisis de la Solución

Ante una muestra aleatoria simple, distribución condicionada gausiana y variables respuesta independientes, podríamos considerar la pérdida cuadrática, que daría lugar al riesgo empírico:

$$\hat{R}_n(\bar{\mathbf{B}}) = \hat{R}_n(\mathbf{W}, \mathbf{C}) \propto \sum_{i=1}^{n} \left\|\mathbf{y}_i^T - \mathbf{x}_i^T\mathbf{W}\mathbf{C}\right\|^2 = \mathrm{tr}\{(\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{C})^T(\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{C})\} \quad \text{(II.19)}$$

cuya minimización, al ser un problema lineal, puede resolverse por métodos analíticos.

En lo que sigue, suponemos sin pérdida de generalidad variables centradas ($\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \mathbf{0}_d$, $\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i = \mathbf{0}_c$) y omitimos el factor $1/n$ en la notación $\mathbf{S_{xx}} = \mathbf{X}^T\mathbf{X}$, $\mathbf{S_{xy}} = \mathbf{X}^T\mathbf{Y} = \mathbf{S_{yx}}^T$ al igual que con las demás variables utilizadas. Asimismo, salvo indicación en contrario, suponemos $\mathbf{W}$ y $\mathbf{C}$ de rango máximo.

Siguiendo a [Baldi y Hornik 1989], se tienen los siguientes resultados preliminares inmediatos:

1. Para cualquier $\mathbf{C}_{h\times c}$ prefijada en (II.18), se obtiene un modelo de regresión lineal en función de $\mathbf{W}_{d\times h}$:

$$\mathbf{y}^T = \mathbf{x}^T\mathbf{W}\mathbf{C} \Leftrightarrow \mathbf{y}^T\mathbf{C}^T = \mathbf{x}^T\mathbf{W}\mathbf{C}\mathbf{C}^T \Leftrightarrow \mathbf{y}^T\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} = \mathbf{x}^T\mathbf{W} \Leftrightarrow \mathbf{v}^T = \mathbf{x}^T\mathbf{W} \quad \text{(II.20)}$$

   con $\mathbf{v}_{h\times 1} = (\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C}\mathbf{y}$. Las ecuaciones normales del problema final anterior son $\mathbf{X}^T\mathbf{X}\mathbf{W} = \mathbf{X}^T\mathbf{V}$, es decir:

$$\begin{aligned} \mathbf{S_{xx}}\mathbf{W} &= \mathbf{S_{xv}} \Leftrightarrow \mathbf{S_{xx}}\mathbf{W} = \mathbf{S_{xy}}\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \Leftrightarrow \mathbf{S_{xx}}\mathbf{W}(\mathbf{C}\mathbf{C}^T) = \mathbf{S_{xy}}\mathbf{C}^T \\ &\Rightarrow \mathbf{W} = \mathbf{S_{xx}}^{-1}\mathbf{S_{xy}}\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \end{aligned} \quad \text{(II.21)}$$

2. Para cualquier $\mathbf{W}_{d\times h}$ prefijada en (II.18), se obtiene un modelo de regresión lineal en función de $\mathbf{C}_{h\times c}$:

$$\mathbf{y}^T = \mathbf{x}^T\mathbf{W}\mathbf{C} \Leftrightarrow \mathbf{y}^T = (\mathbf{x}^T\mathbf{W})\mathbf{C} \Leftrightarrow \mathbf{y}^T = \mathbf{u}^T\mathbf{C} \quad \text{(II.22)}$$

   con $(\mathbf{u}_{h\times 1})^T = \mathbf{x}^T\mathbf{W}$, cuyas ecuaciones normales en términos de $\mathbf{W}$, son $(\mathbf{U}^T\mathbf{U})\mathbf{C} = \mathbf{U}^T\mathbf{Y}$, es decir:

$$\mathbf{S_{uu}}\mathbf{C} = \mathbf{S_{uy}} \Leftrightarrow \mathbf{W}^T\mathbf{S_{xx}}\mathbf{W}\mathbf{C} = \mathbf{W}^T\mathbf{S_{xy}} \Rightarrow \mathbf{C} = (\mathbf{W}^T\mathbf{S_{xx}}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S_{xy}} \quad \text{(II.23)}$$

3. Considerando el problema global, si dos matrices $\mathbf{W}, \mathbf{C}$ definen un punto crítico de $\hat{R}_n$ ($\frac{\partial \hat{R}_n}{\partial(\mathbf{W})_{ij}} = 0$, $\frac{\partial \hat{R}_n}{\partial(\mathbf{C})_{ij}} = 0$), entonces, $\mathbf{W}, \mathbf{C}$ son puntos críticos de los problemas lineales anteriores. Entonces, por ser $\mathbf{W}$ punto crítico del problema lineal (II.20), utilizando (II.21) para cualquier $\mathbf{C}$, se tiene:

$$\bar{\mathbf{B}} = \mathbf{W}\mathbf{C} = \mathbf{S_{xx}}^{-1}\mathbf{S_{xy}}\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C} = \mathbf{S_{xx}}^{-1}\mathbf{S_{xy}}\mathbf{P}_{\mathbf{C}}^T = \hat{\mathbf{B}}\mathbf{P}_{\mathbf{C}}^T \Rightarrow \bar{\mathbf{B}}^T = \mathbf{P_C}\hat{\mathbf{B}}^T \quad \text{(II.24)}$$

   donde $\mathbf{P_C} = \mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C}$ es la matriz proyección sobre el espacio de los vectores columna de $\mathbf{C}$, y $\hat{\mathbf{B}}$ es la solución del problema de regresión lineal $\mathbf{y}^T = \mathbf{x}^T\mathbf{B}$.

Con estos resultados, puede demostrarse el siguiente teorema:

Teorema 35 *[Baldi y Hornik 1989] Sea $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$ la estimación del problema de regresión lineal en la muestra y sea $\mathbf{S_{\hat{y}\hat{y}}} = \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$ de rango máximo con autovalores $\mu_1 > \cdots > \mu_c > 0$ diferentes. Entonces:*

USE UNIVERSIDAD DE SALAMANCA DE CONSULTA

1. $\mathbf{W}, \mathbf{C}$ *definen un punto crítico* $\bar{\mathbf{B}} = \mathbf{WC}$ *de* $\hat{R}_n$ *(definido por (II.19)), si y sólo sí:*

$$\mathbf{W} = \hat{\mathbf{B}}\mathbf{\Gamma}_J\mathbf{G}^{-1} \quad \text{(II.25a)}$$
$$\mathbf{C} = \mathbf{G}\mathbf{\Gamma}_J^T \quad \text{(II.25b)}$$

donde $\mathbf{G}_{h\times h}$ *es cualquier matriz no singular y* $\mathbf{\Gamma}_J$ *es una matriz cuyas columnas son* $r < h$ *autovectores* $\boldsymbol{\xi}_j$, $j \in J \subset \{1 : c\}$, *de* $S_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$. *(Si* $\mathbf{W}, \mathbf{C}$ *son de rango máximo* $r = h$*).*

2. *Para un punto crítico, se cumple que, si* $\mathbf{P}_{\mathbf{\Gamma}_J}$ *es la matriz proyección sobre el subespacio* $\langle\{\boldsymbol{\xi}_j : j \in J\}\rangle$*:*

$$\bar{\mathbf{B}}^T = \mathbf{P}_{\mathbf{\Gamma}_J}\hat{\mathbf{B}}^T \text{ y } \bar{\mathbf{B}}^T = \mathbf{P}_{\mathbf{C}}\hat{\mathbf{B}}^T \quad \text{(II.26)}$$

con lo que $\mathbf{P}_{\mathbf{\Gamma}_J}$ *y* $\mathbf{P}_{\mathbf{C}}$ *proyectan sobre el mismo espacio generado por los autovectores de* $\mathbf{S}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$*. Además:*

$$\hat{R}_n(\mathbf{W}, \mathbf{C}) = \mathrm{tr}(\mathbf{Y}^T\mathbf{Y}) - \sum_{j\in J}\mu_j \quad \text{(II.27)}$$

Debido a este teorema, se obtienen las siguientes conclusiones:

1. Debido a (II.26), las filas de todo punto crítico $\bar{\mathbf{B}} = \mathbf{WC}$ de $\hat{R}_n$ son una proyección ortogonal de las filas de la matriz de coeficientes de la regresión ordinaria $\hat{\mathbf{B}} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xy}}$ sobre un subespacio generado por autovectores de $\mathbf{S}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$.

2. Debido a (II.27), $\hat{R}_n(\mathbf{W}, \mathbf{C})$ posee un mínimo absoluto para los h mayores autovalores de $\mathbf{S}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$, con lo que, formando la matriz Γ_* con los correspondientes autovectores y combinando (II.25a) con (II.25b), resulta que:

$$\bar{\mathbf{B}}_*^T = \mathbf{C}_*^T\mathbf{W}_*^T = \mathbf{\Gamma}_*\mathbf{\Gamma}_*^T\hat{\mathbf{B}}^T = \mathbf{P}_{\mathbf{\Gamma}_*}\hat{\mathbf{B}}^T \quad \text{(II.28)}$$

es un mínimo global, mientras que los demás puntos críticos $\bar{\mathbf{B}} = \mathbf{WC}$, con o sin rango máximo, son puntos de silla.

3. Con el óptimo $\mathbf{W}_*$ se realiza una reducción de dimensionalidad de $\mathbf{x}$ a partir de las h direcciones de mayor variabilidad de $\hat{\mathbf{y}}$, ya que, utilizando (II.25a) por ejemplo con $\mathbf{G} = \mathbf{I}_h$:

$$\mathbf{f} = \bar{\mathbf{B}}_*^T\mathbf{x} = \mathbf{C}_*^T[\mathbf{W}_*^T\mathbf{x}] = \mathbf{C}_*^T[\mathbf{\Gamma}_*^T\hat{\mathbf{B}}^T\mathbf{x}] = \mathbf{C}_*[\mathbf{\Gamma}_*^T\hat{\mathbf{y}}] = \mathbf{\Gamma}_*[\mathbf{\Gamma}_*^T\hat{\mathbf{y}}]$$

En esta expresión, $\mathbf{\Gamma}_*^T\cdot$ es la operación de transformación de coordenadas a dicho espacio. Posteriormente, $\mathbf{C}_*$ a través de $\mathbf{\Gamma}_*$ vuelve al espacio de c dimensiones de $\mathbf{y}$. ($\mathbf{P}_{\mathbf{\Gamma}_*} = \mathbf{\Gamma}_*\mathbf{\Gamma}_*^T$ es una matriz $c \times c$ con $c - h$ columnas nulas).

4. Si $h < c$, la reducción de dimensionalidad sobre el espacio de salida de las unidades ocultas supone una pérdida de información sobre la variable $\hat{\mathbf{y}}$ que la transformación $\mathbf{C}$ ya no puede recuperar. Por ello, si $h < c$, la solución siempre será subóptima con respecto a la del modelo lineal clásico, $\hat{\mathbf{y}}^T = \mathbf{x}^T\hat{\mathbf{B}}$.

Componentes Principales con Aprendizaje Autoasociativo

Si se toma $\mathbf{y} = \mathbf{x}$ como variable respuesta en el problema anterior, el aprendizaje persigue que la red implemente la aplicación identidad $\mathbf{I}_d$ lo mejor posible, mediante: una transformación intermedia en el nivel oculto que reduzca el número de variables de d a h, y una transformación inversa en el nivel de salida destinada a reconstruir las variables originales.

Este problema de aprendizaje es un ejemplo del denominado **aprendizaje autoasociativo**: un **aprendizaje supervisado** típico de los problemas de regresión y clasificación, pero con funcionalidad típica del **aprendizaje no supervisado** típico de los problemas de reducción de dimensionalidad y *clustering*.

Para aplicar los resultados anteriores a este caso $\mathbf{y} = \mathbf{x}$, sean $\mathbf{W}$ y $\mathbf{C}$ puntos críticos como función de $\mathbf{C}$ y $\mathbf{W}$, respectivamente (ecuaciones (II.21) y (II.23)), y sea $\bar{\mathbf{B}} = \mathbf{WC}$ que es punto crítico del problema global. Entonces, aplicando la ecuación (II.21), resulta que:

$$\mathbf{W} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xy}}\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} = \mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \Rightarrow \bar{\mathbf{B}} = \mathbf{WC} = \mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C} = \mathbf{P}_{\mathbf{C}}$$

donde hemos utilizado que, ahora, $\mathbf{S}_{\mathbf{xx}} = \mathbf{S}_{\mathbf{xy}}$.

Asimismo, aplicando el teorema 35 suponiendo $\mathbf{S}_{\mathbf{xx}}$ de rango máximo con autovalores $\mu_1 > \cdots > \mu_d > 0$, se obtiene (ahora $\hat{\mathbf{B}} = \mathbf{I}_d$) que:

1. $\mathbf{W}, \mathbf{C}$ definen un punto crítico de $\hat{R}_n$ si y sólo sí:

   $$\mathbf{W} = \mathbf{\Gamma}_J\mathbf{G}^{-1}; \mathbf{C} = \mathbf{G}\mathbf{\Gamma}_J^T$$

   donde $\mathbf{G}_{h\times h}$ es no singular y $\mathbf{\Gamma}_J$ es una matriz cuyas columnas son r autovectores $\boldsymbol{\xi}_j$, $j \in J \subset \{1:d\}$ de $S_{\mathbf{xx}}$.

2. Para dicho punto crítico $\mathbf{W}, \mathbf{C}$ se cumple que:

   $$\begin{aligned} \bar{\mathbf{B}}^T &= \mathbf{P}_{\mathbf{\Gamma}_J} \\ \hat{R}_n(\mathbf{W}, \mathbf{C}) &= \operatorname{tr}(\mathbf{S}_{\mathbf{xx}}) - \sum_{j\in J}\mu_j \end{aligned}$$

Como consecuencia, y en paralelo con las conclusiones resultantes del teorema 35:

1. Todo punto crítico $\bar{\mathbf{B}} = \mathbf{WC}$ de $\hat{R}_n$ es siempre una proyección ortogonal sobre el subespacio generado por r autovectores $\boldsymbol{\xi}_j$ de $\mathbf{S}_{\mathbf{xx}}$.

2. $\hat{R}_n(\mathbf{W}, \mathbf{C})$ posee un mínimo absoluto para los h mayores autovalores de $\mathbf{S}_{\mathbf{xx}}$, con lo que, formando la matriz Γ_* con los correspondientes autovectores:

   $$\bar{\mathbf{B}}_* = \mathbf{W}_*\mathbf{C}_* = \mathbf{\Gamma}_*\mathbf{G}^{-1}\mathbf{G}\mathbf{\Gamma}_*^T = \mathbf{\Gamma}_*\mathbf{\Gamma}_*^T = \mathbf{P}_{\mathbf{\Gamma}_*}$$

3. En este mínimo global, si $\mathbf{G} = \mathbf{I}_h$, la salida de las unidades del nivel oculto es:

   $$\mathbf{W}_*^T\mathbf{x} = \mathbf{\Gamma}_*^T\mathbf{x} = [\boldsymbol{\xi}_1^T\mathbf{x}, ..., \boldsymbol{\xi}_h^T\mathbf{x}]$$

   que son las coordenadas de $\mathbf{x}$ en el espacio generado por los h autovectores $\{\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_h\}$ correspondientes a los h mayores autovalores de $\mathbf{S}_{\mathbf{xx}}$.

Así pues, la red neuronal lineal bajo régimen autoasociativo implementa la reducción de dimensionalidad mediante componentes principales, (transformación de Karhunen-Loève, p. ej. [Diamantaras y Kung 1996, th. 3.4], teorema 120 Anexo B).

En la práctica sin embargo, si las condiciones iniciales del algoritmo son aleatorias, el algoritmo de optimización no lineal utilizado no producirá una matriz $\mathbf{G} = \mathbf{I}_h$, con lo que no debe esperarse que $\mathbf{W}_* = \mathbf{\Gamma}_*\mathbf{G}^{-1} = \mathbf{\Gamma}_* = [\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_h]$, es decir, que las componentes sea ortogonales.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Aunque esto puede parecer negativo, produce sin embargo estructuras más robustas, al evitar que una neurona concentre mucha mayor varianza que las demás, (la de mayor autovalor con respecto a las de los menores), equilibrando así el papel que desempeñan todas las unidades ocultas en la reducción de la dimensionalidad.

Sobre lo anterior, es importante destacar que la red obtiene las componentes principales mediante el criterio de mínimos cuadrados:

$$\min_{\mathbf{W},\mathbf{C}} \left\|\mathbf{x}^T - \mathbf{x}^T\mathbf{W}\mathbf{C}\right\|^2 = \min_{\mathbf{W},\mathbf{C}} \|\mathbf{x} - \mathbf{P}_{\mathbf{\Gamma}}\mathbf{x}\|^2$$

pero, siendo este último valor, la norma del vector proyección sobre el espacio ortogonal al generado por los autovectores, dicho criterio equivale a $\max_{\mathbf{W},\mathbf{C}} \|\mathbf{P}_{\mathbf{\Gamma}}\mathbf{x}\|^2$, es decir a maximizar la varianza del vector proyección, como es tradicional.

Finalmente, mencionar que el análisis de componentes principales implementado por la red neuronal lineal con una capa oculta puede generalizarse al caso no lineal.

Sin embargo, en este caso, no es suficiente con considerar activaciones no lineales en el nivel oculto [Bourlard y Kamp 1988], debido a que la no linealidad ha de producirse en la reducción de la dimensionalidad y, al mismo tiempo, en la transformación inversa. Ello requiere dos redes con un nivel oculto cada una (activación de salida lineal), conectadas en serie repartiéndose las dos tareas anteriores, lo que produce un total de tres niveles ocultos.

Otras opciones en este sentido son, por ejemplo, [Kramer 1992], [Kramer 1991], [Tan y Mavrovouniotis 1995], éste último implementando la técnica de curvas principales de [Hastie y Stuetzle 1989]. Una visión general de otros enfoques puede verse por ejemplo en [Diamantaras y Kung 1996].

II.3.2 Arquitecturas MLF No Lineales. Visión General

En lo que resta de sección, nos centramos en su objeto principal: las redes MLF no lineales, es decir, de la forma (II.17):

$$f(\mathbf{x};\boldsymbol{\theta}) = \varphi(\sum_{j=1}^{h} c_j\psi_j(\mathbf{x};\mathbf{w}_j) + c_0) = \varphi(\mathbf{c}^T\boldsymbol{\psi} + c_0) = \varphi(\bar{\mathbf{c}}^T\bar{\boldsymbol{\psi}}) \tag{II.30}$$

donde ahora las funciones de activación $\psi_j(r)$, $r \in \mathbb{R}$, $j = 1, ..., r$, se suponen no lineales.

En el presente apartado nos limitamos a realizar una rápida revisión de los distintos tipos de arquitectura MLF según su aplicación a problemas de clasificación o regresión, los cuáles, condicionan fundamentalmente las funciones de transferencia a utilizar.

En las secciones siguientes, profundizaremos en dichos tipos de arquitectura e identificaremos los correspondientes problemas de estimación o entrenamiento.

En una sección posterior, estudiaremos los distintos tipos de algoritmo utilizados para el entrenamiento.

Funciones de Transferencia del Nivel de Salida

El nivel de salida constituye realmente una red neuronal de una sóla capa (o una neurona si se trata de una sóla salida) cuya entrada es ahora la salida del nivel oculto. Así pues, el análisis realizado

en la sección anterior para este tipo de redes es aplicable aquí al nivel de salida y las funciones de activación de este nivel están determinadas fundamentalmente por el problema de aplicación:

1. **Clasificación**. Si se trata de un problema de clasificación en el que se pretende un estimador *plug-in* de la regla de Bayes, del análisis de la sección anterior se deduce que para estimar las probabilidades a posteriori pueden utilizarse funciones de activación sigmoides o lineales, o softmax para el caso de varias clases, con función de pérdida cuadrática o de entropía cruzada. La segunda es la más natural pero la primera produce buenos resultados y es la más utilizada.

2. **Regresión**. Si se trata de un problema de regresión, es necesario utilizar funciones de activación compatibles con el rango de los datos. En este sentido, las funciones sigmoides restringen la salida al intervalo $[0, 1]$ y, a pesar de que los datos podrían ser transformados previamente para ser compatibles con ello, suelen utilizarse activaciones lineales (la identidad), debido a su mayor facilidad de entrenamiento con pérdida cuadrática. Ello, como veremos, no penaliza la capacidad de la red pues su naturaleza no lineal se realiza previamente en el nivel oculto.

Por tanto, dado que el problema de clasificación mencionado utiliza un enfoque de regresión, en lo que sigue, salvo que se especifique lo contrario, trataremos el problema de regresión como problema marco. Únicamente, haremos referencia explícita al problema de clasificación cuando una técnica específica lo requiera o cuando puedan obtenerse conclusiones específicas para dicho problema.

Funciones de Transferencia del Nivel Oculto

En el nivel oculto suelen utilizarse funciones de transferencia no lineales, pues, como vimos en el apartado anterior, la transformación lineal, si reduce la dimensión, hace perder información, y si la incrementa, lo hace de forma redundante.

Dichas funciones de transferencia son el resultado de la composición, de alguna transformación no acotada de las variables de entrada, con una función de activación no lineal, acotada, para limitar los valores finales de la función de transferencia.

Estas funciones de transferencia pueden clasificarse en dos grandes grupos (aunque pueden producir conjuntos de arquitecturas no disjuntos) según su motivación inicial:

1. La implementación de algún criterio de **proximidad o similitud** entre dos ejemplos de la muestra, de tal forma, que sus respectivas respuestas sean similares.

   Este tipo de enfoque puede dividirse a su vez en dos grandes grupos, según el carácter de la transformación de las variables de entrada previa a la función de activación. Estos grupos, se corresponden básicamente con los dos grandes métodos de clasificación en estadística: discriminación y *clustering*:

   (a) Transformaciones basadas en **proyecciones** o en un producto interno de la forma $\langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^T \mathbf{x}$, en las que, la proximidad entre dos ejemplos, se define, según la proximidad de sus proyecciones sobre la dirección determinada por el vector $\mathbf{w}$.
   Como resultado, las funciones de transferencia son funciones *ridge* $\psi_j(\mathbf{w}_j^T \mathbf{x} + w_0)$ verificando que, para dos ejemplos $\mathbf{x}_1, \mathbf{x}_2$ con $\mathbf{w}_j^T \mathbf{x}_1 \approx \mathbf{w}_j^T \mathbf{x}_2$ la salida $\psi_j(\mathbf{w}_j^T \mathbf{x}_i + w_0)$ de la neurona j-ésima del nivel oculto será también similar, contribuyendo de forma muy parecida a la salida final de la red.
   En lo que sigue, denominaremos a las arquitecturas con funciones de transferencia de este tipo, arquitecturas MLF de **tipo proyección**. Ejemplos de este tipo de arqui-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

tectura son: las redes MLF con activaciones lineales del apartado anterior, las redes MLP (*multilayer perceptron*), cuyas activaciones ocultas son siempre sigmoides y todas iguales, y las redes **proyección pursuit** (PP) de [Friedman y Tukey 1974] y [Friedman y Stuetzle 1981]. Todas las redes de una capa tratadas en la sección anterior, definen también la similitud entre dos ejemplos mediante proyecciones.

(b) Transformaciones basadas en **diferencias** $\mathbf{x} - \mathbf{w}$ con respecto a un vector referencia $\mathbf{w}$, y, en particular, muy frecuentemente, **distancias** $\|\mathbf{x} - \mathbf{w}\|$, donde la norma no tiene por qué ser necesariamente euclídea. Las primeras son invariantes ante traslaciones y las segundas ante traslaciones y rotaciones, y su elección depende del problema de aplicación (que, en algún caso, puede requerir otras opciones a medida).
Como resultado, las funciones de transferencia son funciones del tipo $\psi_j(\mathbf{x} - \mathbf{w}_j)$ o funciones radiales $\psi_j(\|\mathbf{x} - \mathbf{w}_j\|)$ de tal forma que, ejemplos similares por referencia a $\mathbf{w}$ (*próximos* a $\mathbf{w}$ en el caso radial), producirán también contribuciones similares a la salida de la red.
En lo que sigue denominaremos a las arquitecturas con funciones de transferencia de este tipo, arquitecturas de **tipo radial** pues la distancia es el criterio de similitud más utilizado. Ejemplos de este tipo son: las redes neuronales de base radial (RBF) con un número de neuronas ocultas menor que el tamaño muestral, ([Broomhead y Lowe 1988, sec. 2], [Moody y Darken 1988]), las **redes de base radial de regularización** (RRBF), de [Poggio y Girosi 1990] y las **support vector machines** (SVM), de [Boser *et al.* 1992] y [Cortes y Vapnik 1995].

Es interesante la siguiente relación entre el criterio de similitud de producto interno y el criterio de distancia:

$$\|\mathbf{x} - \mathbf{w}\|^2 = \langle (\mathbf{x} - \mathbf{w}), (\mathbf{x} - \mathbf{w}) \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle - 2 \langle \mathbf{x}, \mathbf{w} \rangle \tag{II.31}$$

con lo que, si los vectores están normalizados, se obtiene: $\|\mathbf{x} - \mathbf{w}\|^2 = 2 - 2 \langle \mathbf{x}, \mathbf{w} \rangle$.

De esta forma, a mayor distancia, menor producto interno, y, así, puede pensarse en $\langle \cdot, \cdot \rangle$ como una medida de similitud y en $\|\cdot - \cdot\|^2$ como una medida de disimilitud.

2. La aproximación de una función de interés mediante funciones básicas de un espacio funcional, al modo de los estimadores clásicos basados en expansiones en serie de Fourier.

   En este enfoque, las funciones de transferencia coinciden con las funciones básicas, y su función de activación es una misma función escalada y trasladada convenientemente, con el fin de capturar la información que, la función original, posee en diferentes puntos del espacio de entrada y a diferentes escalas, (dominio tiempo-escala).

   Ejemplos de este tipo de redes son las **redes wavelet** (WN) que utilizan bases ortonormales o estructuras (*frames* –bases redundantes) de wavelets multidimensionales, en sus vertientes **monoescala** ([Boubez 1993], [Kugarajah y Zhang 1995], [Zhang 1997]) y **multiescala** ([Pati y Krishnaprasad 1990], [Zhang y Benveniste 1992], [Pati y Krishnaprasad 1993]), según que utilicen la misma escala o escalas diferentes en las distintas dimensiones del espacio de entrada.

   En esta línea, se han propuesto recientemente nuevas funciones de transferencia que siguen la misma filosofía, pero, utilizando **wavelets de tipo proyección** denominadas *ridgelets* ([Candes 1998], [Candes 1999]) que implementan una transformación previa de las variables de entrada según proyecciones al modo de las redes MLP, (**redes ridgelets**).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

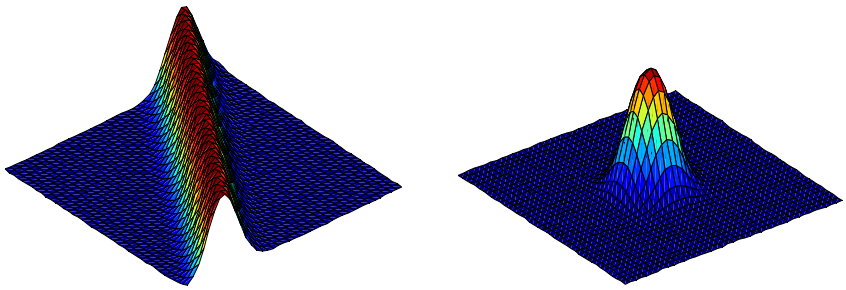


Figura 12. Izquierda: activación gausiana con proyección: $f(\mathbf{x}) = \exp(-0{,}1(2x_1 + 3x_2)^2)$. Derecha: activación logística aplicada a una transformación distancia: $f(\mathbf{x}) = -(1 + e^{-(x_1^2+x_2^2)})^{-1}$

A pesar de estas diferencias en motivación, los distintos tipos de función de transferencia suelen poder interpretarse en términos de algún criterio de *similitud* entre los ejemplos de la muestra, y sus relaciones son más fuertes de lo que, en principio, parecen:

1. Las redes SVM son, en origen, redes lineales de tipo proyección al modo del perceptrón generalizado: $f(\mathbf{x}; \mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\phi} \rangle + w_0$ donde $\mathbf{w} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i$ y $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$ es alguna transformación no lineal de las variables de entrada en algún espacio de gran dimensión. Sin embargo, estas redes, implementan el producto interno mediante $\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}_i) \rangle = k(\mathbf{x}, \mathbf{x}_i)$, donde k es una función definida positiva (Anexo B), (radial como caso particular), de tal forma que $f(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{n} \alpha_i \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}_i) \rangle = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$.

2. Las redes wavelet monoescala dan lugar a arquitecturas radiales, y las redes multiescala poseen un configuración de tipo proyección.

Realmente, puede decirse que la diferencia más importante entre los diferentes tipos de funciones de transferencia es el carácter **global** de las de tipo proyección, y el carácter **local** de las demás. En este sentido, la transformación de proyección provoca que puntos muy alejados (globalidad) pero situados en un mismo hiperplano de nivel $\{\mathbf{w}^T\mathbf{x} = b\}$ de la función de activación, posean respuestas muy similares independientemente de que la función de activación sea local o no (figura 12, izquierda). A la inversa: la aplicación de una función de activación típicamente asociada a una red de proyección, por ejemplo logística, aplicada a una transformación distancia da lugar también a una función de transferencia local (figura 12, derecha).

Sin embargo, la suma de dos funciones de transferencia (modelo aditivo) puede adquirir ya carácter global, como era de esperar, para poder aproximar cualquier función regular (figura 13).

Un criterio general perseguido para la elección de la función de transferencia es que aporte la suficiente complejidad a la red neuronal (II.30), para poder aproximar cualquier función del espacio de interés y así poder reducir a cero el error de aproximación tratado en el capítulo anterior.

Estas propiedades aproximadoras se describirán en las secciones siguientes a medida que vayamos presentando las diferentes arquitecturas. Para las redes del segundo grupo, esta capacidad de aproximación se obtiene mediante la condición de que la familia de funciones básicas sea una base para el espacio $\mathcal{F}$ de hipótesis (II.30) y que este espacio sea denso en algún espacio de interés.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

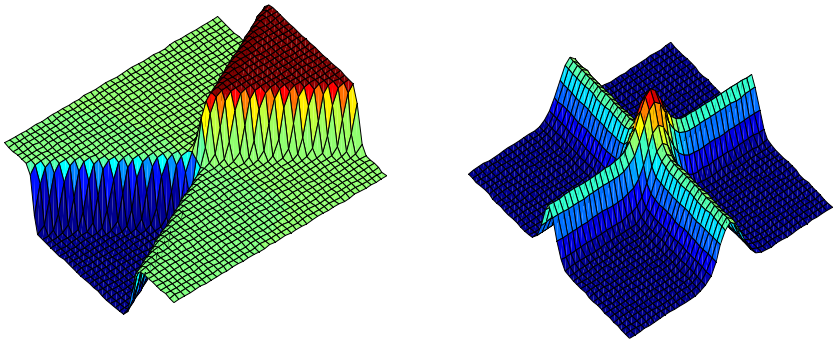


Figura 13. Izquierda: dos neuronas con activación logística con transformación proyección: $f(\mathbf{x}) = (1 + e^{-(2x_1+3x_2)})^{-1} + (1 + e^{-(x_1-2x_2)})^{-1}$. Derecha: dos neuronas con activación gausiana y transformación distancia: $f(\mathbf{x}) = e^{-x^2} + e^{-2y^2}$.

Con respecto al primer grupo de arquitecturas, un requisito adicional es una regularidad suficiente para permitir la utilización de algoritmos de optimización no lineal en caso necesario (normalmente, que posean segunda derivada continua). Una revisión general de las comúnmente utilizadas en este grupo es [Duch y Jankowski 1999].

Significado del Nivel Oculto en una Arquitectura MLF

En este apartado, ilustramos el papel que juega la capa oculta de una arquitectura MLF en un problema de clasificación siguiendo un análisis debido a [Webb y Lowe 1990].

Para ello, suponemos un problema de s clases $\mathcal{C}_j$, $j = 1 : s$, excluyentes con la codificación $y_j(\mathbf{x}) = 1$ si $\mathbf{x}$ está en $\mathcal{C}_j$ y cero en caso contrario, en el que se desea estimar las probabilidades a posteriori con pérdida cuadrática y activación lineal en salida.

Suponiendo prefijado la matriz de parámetros del nivel oculto $\mathbf{W}$, el nivel de salida efectúa una regresión lineal sobre las variables de salida $\mathbf{u}$ del nivel oculto:

$$\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})^T = \mathbf{u}^T\mathbf{C} + \mathbf{c}_0^T$$

donde $\boldsymbol{\theta}$ representa todos los parámetros de la red, $\mathbf{u} = (u_1, ..., u_h)^T$ con $u_j = \psi_j(\mathbf{x}; \mathbf{w}_j)$, $j = 1 : h$, $\mathbf{C}$ es la matriz $h \times s$ de coeficientes de las covariables y $\mathbf{c}_0 = (c_{01}, ..., c_{0s})^T$ es el vector de coeficientes independientes.

Dada una muestra aleatoria simple $\mathbf{z}^n$ con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, resolviendo el problema para las variables $\mathbf{u}$ e $\mathbf{y}$ centradas, la solución para $\mathbf{C}$ es[48]:

$$\mathbf{C}(\mathbf{W}) = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Y}$$

donde $\mathbf{Y}$ es una matriz $n \times s$, $\mathbf{U}_{n\times h} = \mathbf{U}(\mathbf{W})$ con $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_n]^T$ matriz de diseño con las variables ocultas centradas y hemos supuesto $(\mathbf{U}^T\mathbf{U})$ de rango máximo (de lo contrario, se sustituye por la inversa generalizada –Anexo G).

[48] La solución para $\mathbf{c}_0$ sería $\mathbf{c}_0^T = \tilde{\mathbf{y}}^T - \tilde{\mathbf{u}}^T\mathbf{C}$ donde $\tilde{\mathbf{y}}$, $\tilde{\mathbf{u}}$ son las medias muestrales respectivas de $\mathbf{y}$ y $\mathbf{u}$.

Sustituyendo la estimación anterior en el riesgo empírico y utilizando las propiedades de la traza (Anexo G):

$$\begin{aligned}\hat{R}_n(\mathbf{W}) &\propto \mathrm{tr}[(\mathbf{Y}-\mathbf{UC})(\mathbf{Y}-\mathbf{UC})^T] = \mathrm{tr}[(\mathbf{Y}-\mathbf{UC})^T(\mathbf{Y}-\mathbf{UC})] \\ &= \mathrm{tr}[\mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Y}] \\ &= \mathrm{tr}[\mathbf{Y}\mathbf{Y}^T - \mathbf{U}^T\mathbf{Y}\mathbf{Y}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}] \\ &= \mathrm{tr}[\mathbf{Y}^T\mathbf{Y} - \mathbf{S}_E\mathbf{S}_T^{-1}] \end{aligned} \tag{II.32}$$

donde hemos denotado:

1. $\mathbf{S}_T = \mathbf{U}^T\mathbf{U}$, que es la covarianza muestral, salvo factor, de las salidas del nivel oculto (que habíamos supuesto centradas).

2. $\mathbf{S}_E = \mathbf{U}^T\mathbf{Y}\mathbf{Y}^T\mathbf{U}$, que es una especie de covarianza entre clases en la salida del nivel oculto, ya que con la codificación indicada, puede demostrarse que:

$$\mathbf{S}_E = \sum_{j=1}^{s} n_j^2(\tilde{\mathbf{u}}_j - \tilde{\mathbf{u}})(\tilde{\mathbf{u}}_j - \tilde{\mathbf{u}})^T \tag{II.33}$$

   donde n_j es el número de casos en $\mathcal{C}_j$, $\tilde{\mathbf{u}}_j = \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\bar{\mathbf{u}}_i$ es la media en la clase $\mathcal{C}_j$ y $\tilde{\mathbf{u}}$ es la media de dichas variables para todas las observaciones. Con ello, $\mathbf{S}_E$ difiere de la varianza entre clases (ecuación (A.2) del Anexo A) en que posee factores n_j^2 en lugar de n_j, con lo que pondera más fuertemente a favor de las clases con mayor número de casos.

Como el primer término de (II.32) dentro de la traza sólo depende de los datos, el problema de minimización del riesgo empírico equivale a:

$$\max_{\mathbf{W}} \mathrm{tr}[\mathbf{S}_E\mathbf{S}_T^{-1}]$$

que es similar al que daría lugar a la función discriminante de Fisher en el espacio de variables ocultas (Anexo A, donde ahora es la varianza total $\mathbf{S}_T$ de dichas variables la que actúa en el denominador en lugar de la varianza dentro de las clases $\mathbf{S}_I$).

Con ello, en la estimación de las probabilidades a posteriori con una arquitectura MLF, el problema de discriminación alcanza al espacio de variables ocultas, persiguiendo, el aprendizaje, la transformación no lineal que maximiza la discriminación en dicho espacio.

Sin embargo, el hecho de que (II.33) favorezca las clases con mayor número de casos, hace que, cuando haya discrepancias relevantes entre las probabilidades a priori $\mathsf{p}(\mathcal{C}_j)$ y las proporciones muestrales $q_j = n_j/n$ de las clases, [Lowe y Webb 1990] recomienden utilizar, en el caso de redes MLF con salida lineal, en lugar de mínimos cuadrados ordinarios, mínimos cuadrados ponderados mediante $\hat{\mathsf{p}}(\mathcal{C}_j)/q_j$ en cada ejemplo de la clase $\mathcal{C}_j$, donde $\hat{\mathsf{p}}(\mathcal{C}_j)$ es una estimación de $\mathsf{p}(\mathcal{C}_j)$ en una muestra de validación.

De cualquier forma, del análisis anterior se deduce que, durante el aprendizaje de estas redes, se están resolviendo simultáneamente dos problemas:

1. Un problema de regresión en la capa oculta que da lugar a la estimación de las probabilidades a posteriori.

2. Un problema de extracción de las características (*features*) de las variables de entrada (pudiendo dar lugar o no, a reducción de dimensionalidad) que resultan relevantes para el problema anterior, mediante un criterio similar al discriminante de Fisher.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Tradicionalmente, estos dos problemas se realizaban por separado lo que, al carácter muchas veces heurístico de la búsqueda de las mejores transformaciones, se unía su suboptimalidad y carga computacional. Del análisis anterior, se deduce, la potencia de estas redes para resolver este tipo de problemas.

II.4 Arquitecturas de Tipo Proyección

Entre la gran cantidad de posibles arquitecturas MLF de proyección para regresión, la más utilizada es la red neuronal perceptrón multicapa (MLP) que presentamos seguidamente, y se caracteriza por poseer la misma función de activación sigmoide en todas las neuronas del nivel oculto y activaciones también lineales o sigmoides en salida.

Pero, además de ella, son importantes las siguientes que, por cuestiones de espacio, no se tratan aquí en profundidad:

1. Las redes *wavelet* de tipo proyección, que incluyen: las redes *wavelet* multiescala, que utilizan funciones de activación resultado del producto tensorial de *wavelet* unidimensionales con diferentes escalas en diferentes dimensiones del espacio, y las *ridgelets* de [Candes 1998] (redes con activaciones *wavelet* aplicadas a proyecciones).

   En particular, las redes MLP pueden considerarse una expansión *wavelet* [Pati y Krishnaprasad 1993]. Trabajos representativos al respecto son [Pati y Krishnaprasad 1990], [Zhang y Benveniste 1992], [Pati y Krishnaprasad 1993], [Candes 1998] y [Candes 1999].

2. Las redes *projection pursuit* (PP). Estas redes nacen originalmente en el campo de la estadística [Friedman y Tukey 1974], como medio de seleccionar las mejores proyecciones en el espacio de entrada para un problema de reconocimiento de patrones (*clustering*).

   Posteriormente, estas técnicas se extienden al problema de regresión, ([Friedman y Stuetzle 1981], [Diaconis y Shahshahani 1984], [Huber 1985], [Friedman 1987], [Jones 1987], [Donoho *et al.* 1992]) y son anexionadas por el campo de las redes neuronales ([Barron y Barron 1988], [Jones 1992], [Barron 1993], [Hwang *et al.* 1994], [Zhao y Atkeson 1996], [Jones 2000]), pues resultan en una generalización de la arquitectura MLP al admitir funciones de activación diferentes en el nivel oculto, (en su origen, suavizadores no paramétricos, diferentes en cada proyección y elegidos a partir de los datos).

   Sin embargo, la diferencia más importante entre ambos tipos de red y que mantiene su diferente denominación, reside en el algoritmo tradicionalmente utilizado con cada una. El algoritmo original de entrenamiento de las redes MLP (*back-propagation*) estima todos los parámetros de la red simultáneamente (el número de neuronas está prefijado de antemano y es preciso después realizar la selección del modelo), mientras que el algoritmo típico de *projection pursuit* (algoritmo *greedy*) va incorporando neuronas adicionales progresivamente hasta que el error es menor que un umbral prefijado (la selección del modelo se traslada a la selección de dicho umbral).

   A pesar de ello, la arquitectura típica MLP también puede entrenarse con el algoritmo *greedy*, lo que no justifica el mantenimiento de ambas denominaciones. La única motivación final para ello, sería la utilización de suavizadores no paramétricos como función de activación en las redes PP, técnica que no es usual en el ámbito de las redes neuronales.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

II.4.1 Redes Perceptrón Multicapa (MLP)

La red neuronal perceptrón multicapa (MLP), se caracteriza por utilizar la misma función de activación para todas las neuronas del nivel oculto y activaciones lineales en salida. El significado estadístico de estas redes ha sido ya analizado. Para el problema de regresión, posee como forma general:

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{h} c_j \psi(\mathbf{w}_j^T \mathbf{x} + w_{j0}) + c_0 \tag{II.34}$$

Como hemos indicado anteriormente, en el problema de clasificación si se desea implementar el estimador *plug-in*, las activaciones de salida pueden ser igualmente lineales o también sigmoides, en ambos casos bajo pérdida cuadrática o de entropía cruzada.

En el nivel oculto, se han venido utilizando las funciones sigmoides mencionadas en la sección anterior, entre las que la más conocida es la logística, pero también transformaciones de ella como, por ejemplo, la función $\psi^1(r) = \tanh(r) \in [-1, 1]$, cuya relación con la logística $\psi^0(r)$ es:

$$\psi^1(r) = \frac{e^r - e^{-r}}{e^r + e^{-r}} = \frac{1 - e^{-2r}}{1 + e^{-2r}} = 2\frac{1}{1 + e^{-2r}} - 1 = 2\psi^0(2r) - 1$$

con lo que:

$$f(\mathbf{x};\boldsymbol{\theta}) = c_0^1 + \sum_{j=1}^{h} c_j^1 \psi^1((\bar{\mathbf{w}}_j^1)^T \bar{\mathbf{x}}) = c_0^1 - h + \sum_{j=1}^{h} 2c_j^1 \psi^0(2(\bar{\mathbf{w}}_j^0)^T \bar{\mathbf{x}})$$

De esta forma, con la función $\tanh$ se obtendría, después del entrenamiento, la misma hipótesis que produce la logística con $\bar{\mathbf{w}}_j^1 = 2\bar{\mathbf{w}}_j^0$ en los parámetros del nivel oculto y $c_j^1 = 2c_j^0$, $c_0^1 = c_0^0 - h$ en los del nivel de salida.

Otras funciones sigmoides frecuentemente utilizadas son:

1. La función de distribución de la gausiana (*error function*):

$$\psi(r) \propto \int_0^r e^{-t^2} dt$$

2. Aproximaciones de la logística, como puede ser la función *ramp* (*ramp function*):

$$\psi(r) = \begin{cases} 0 & r \leq a \\ \dfrac{r-a}{b-a} & a \leq r \leq b \\ 1 & b \leq r \end{cases} \tag{II.35}$$

Pero, en general, puede utilizarse cualquier función con buenas propiedades aproximadoras, según los resultados que presentamos en el siguiente apartado.

Consistencia de las Redes MLP

A pesar de que los algoritmos que pueden utilizarse con las redes MLP se analizan en una sección posterior, presentamos ahora algunos resultados sobre la consistencia de estas redes pues suponen la existencia de un algoritmo, sin especificar, que minimiza el riesgo empírico.

USAL UNIVERSIDAD DE SALAMANCA DE CÓRDOBA

Siendo claro su interés teórico, estos resultados poseen una aplicación práctica limitada pues, la suposición anterior es una gran idealización del entrenamiento de estas redes, al no considerar la posibilidad de que el algoritmo desemboque en mínimos locales del problema de optimización.

A diferencia de las redes lineales cuyas limitadas posibilidades de representación están claras de antemano, la arquitectura MLP implementa funciones no lineales con mucha mayor complejidad. Debido a ello, el estudio de la consistencia de esta arquitectura no se limita a conocer si el proceso de aprendizaje converge a la mejor hipótesis de la familia $\mathcal{F}$ de redes MLP (II.34), si no también, en qué condiciones para una supuesta función verdadera f se obtiene la convergencia a dicha función.

Dicha convergencia lleva implícito no sólamente que f pueda ser aproximada por funciones de $\mathcal{F}$ si no también que el proceso de aprendizaje en sí, en el contexto de un número limitado de datos sujetos a ruido, produzca hipótesis que converjan a dicha función a medida que aumenta el tamaño muestral.

En este marco, son oportunos los comentarios sobre el problema de control de la complejidad y selección del modelo realizados en el primer capítulo. El método usual de selección del modelo en el ámbito de las redes MLP (y, en general, de las redes MLF) consiste en entrenar de distintas redes con diferente número de unidades ocultas y seleccionar la que minimiza algún criterio de selección del modelo de los presentados en el primer capítulo.

Entre este tipo de métodos, uno de los mejor fundamentados teóricamente es el método de minimización del riesgo estructural (SRM), que, además, es válido para cualquier distribución[49]. Debido a ello, se han obtenido resultados de consistencia con este método, específicos para este tipo de arquitectura.

En el Apartado I.6.2 se describe el procedimiento más usual para analizar la consistencia de las redes neuronales en el marco del método SRM. Dicho procedimiento se basa en la descomposición de la discrepancia entre el riesgo cometido y el riesgo óptimo, en los denominados error de aproximación y error de estimación:

$$R(\hat{f}_{t(n)}) - R(\mathsf{f}) = [R(\hat{f}_{t(n)}) - R(f^*_{t(n)})] + [R(f^*_{t(n)}) - R(\mathsf{f})]$$

donde $t(n)$ define una regla de control de la complejidad con el fin de acompasar ésta al tamaño muestral determinando una subfamilia $\mathcal{F}_{t(n)} \subset \mathcal{F}$, y donde $\hat{f}_{t(n)}$ y $f^*_{t(n)}$minimizan el riesgo empírico y el riesgo, respectivamente, en $\mathcal{F}_{t(n)}$.

Si se adopta como riesgo $R(f) = \|\mathsf{Y} - f(\mathsf{X})\|_{L_p(\mathsf{P_Z})}$, la descomposición anterior puede acotarse como (ecuación (I.49)):

$$R(\hat{f}_{t(n)}) - R(\mathsf{f}) \leq 2 \sup_{f \in \mathcal{F}_{t(n)}} \left| R(f) - \hat{R}_n(f) \right| + \left\| f_{t(n)} - \mathsf{f} \right\|_{L_p(\mathsf{P_Z})}$$

con lo que, las condiciones básicas para lograr la consistencia del método SRM (ecuaciones (I.50), pág. 61) son:

$$\begin{cases} 1) & \mathbb{P}^{\infty}_{\mathsf{Z}}\{ \lim_{n\to\infty} \sup_{f \in \mathcal{F}_{t(n)}} \left| R(f) - \hat{R}_n(f) \right| = 0\} = 1 \\ 2) & \lim_{n\to\infty} \inf_{\mathcal{F}_{t(n)}} \|f - \mathsf{f}\|_{L_p(\mathsf{P_Z})} = 0 \end{cases}$$

Sobre estas condiciones, destacamos lo siguiente:

[49] Aunque, en la práctica, en ocasiones, es problemática la utilización de su criterio de selección del modelo, al estar basado en la dimensión VC o medida similar de capacidad, que suele ser difícil de determinar, por lo que suelen utilizarse otros criterios.

1. La segunda condición garantiza que el modelo va adquiriendo, a medida que crece el tamaño muestral, la necesaria complejidad para representar la verdadera función.

   Esta condición se consigue por ejemplo si las familias $\mathcal{F}_{t(n)}$ están anidadas: $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_t \subset \cdots$ y $\cup_t \mathcal{F}_t$ es denso en $L_p(\mathsf{P_Z})$, $\forall \mathsf{P_Z}$, es decir, que cualquier función de $L_p(\mathsf{P_Z})$ $\forall \mathsf{P_Z}$ pueda ser aproximada mediante funciones implementadas por la arquitectura.

   Por ello, una línea de trabajo fundamental en los estudios de consistencia, va dirigida a determinar las familias de funciones en que dicha condición se verifica. La tabla 4 presenta los resultados más importantes sobre las propiedades aproximadoras de la arquitectura MLP, cuyos términos y símbolos de nueva aparición se describen en el Anexo C y, en todo caso, en el Anexo H.

2. Por su parte, la primera condición garantiza que el algoritmo realiza un efectivo control de la complejidad limitando la sensibilidad del estimador ante la aleatoriedad muestral. En general, los resultados existentes, consiguen este objetivo limitando el tamaño de los parámetros de la red al modo del método de los *sièves* (lo que determina una secuencia $\{\mathcal{F}_{t(n)}\}$ de complejidad creciente) e imponiendo condiciones adecuadas sobre el comportamiento de dichas cotas en función del tamaño muestral.

El teorema que presentamos a continuación para el problemas de regresión es el ejemplo más ilustrativo. Para ello, consideramos el riesgo $R(f) = \|\mathsf{Y} - f(\mathsf{X})\|_{L_p(\mathsf{P_Z})}$ para algún $p \in [1, \infty)$ y su correspondiente riesgo empírico:

$$\hat{R}_n(f) = \frac{1}{n}[\sum_{i=1}^{n} |y_i - f(\mathbf{x}_i)|^p]^{1/p}$$

así como las familias de la forma:

$$\mathcal{F}_n = \{f(\mathbf{x}) = \sum_{j=1}^{h_n} c_j \psi(\mathbf{w}_j^T \mathbf{x} + w_{j0}) + c_0 : \mathbf{c} \in \mathbb{R}^{h_n}, c_0 \in \mathbb{R}, \sum_{j=1}^{h_n} |c_j| \leq \beta_n\}$$

con $\mathcal{F}_n \subset \mathcal{F} = \{f(\mathbf{x}) = \sum_{j=1}^{h} c_j \psi(\mathbf{w}_j^T \mathbf{x} + w_{j0}) + c_0 : \mathbf{c} \in \mathbb{R}^h, c_0 \in \mathbb{R}, h \in \mathbb{N}^+\}$, donde $\psi : \mathbb{R} \to [0, 1]$ es una función sigmoide monótona y no decreciente.

Teorema 36 *[Lugosi y Zeger 1995, th. 3] Sea $p \in [1, \infty)$ con R, $\hat{R}_n$ definidos como anteriormente y $\mathcal{F}_n$ con $n \in \mathbb{N}^+$ las familias anteriores. Si h_n y β_n satisfacen:*

$$h_n \to \infty, \beta_n \to \infty \; y \; \frac{h_n \beta_n^{2p} \log(h_n \beta_n)}{n} \to 0$$

entonces, $\forall \mathsf{P_Z}$ con $\mathbb{E}_{\mathsf{Z}}(|\mathsf{Y}|^p) < \infty$, $\{\hat{f}_n\}$ es consistente, es decir:

$$\text{plim}[R(\hat{f}_n) - R(\mathsf{f})] = 0$$

Si, además, se satisface que $\beta_n^{2p} = o(n^{1-\delta})$ para algún $\delta > 0$, entonces $R(\hat{f}_n) - R(\mathsf{f}) \to 0$ c.s., es decir $\{\hat{f}_n\}$ es fuertemente consistente[50] *$\forall \mathsf{P_Z}$ con $\mathbb{E}_{\mathsf{Z}}(|\mathsf{Y}|^p) < \infty$.*

[50] Recuérdese el teorema 30 tratado cuando expusimos los condiciones generales para la consistencia (pág. 63).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Propiedades Aproximadoras de la red MLP: $f(\mathbf{x}) = \sum_{i=1}^{h} c_i \psi(\mathbf{w}_i^T \mathbf{x} + w_{i0})$			
Espacio Funcional	$\|\cdot\|$	**Condiciones**	**Enfoque**
$\mathcal{C}(\mathcal{X})$, $\mathcal{X}$ compacto [Cybenko 1989]	L_∞	ψ continua, sigmoide	Existencial Th. Hahn-Banach
$\mathcal{T}_C(\mathbb{B}_{d,\tau})$ μ med. prob. [Barron 1993]	L_2	ψ sigmoide medible, acotada	A.- Existencial: 1) Lema Maurey-Jones 2) f $\rightarrow$ Trigonom. $\rightarrow$ Func. Salto $\rightarrow$ Sigmoides B.- Constructivo: *greedy*
$\mathcal{C}(\mathcal{X})$, $\mathcal{X}$ compacto [Funahashi 1989]	L_∞	ψ sigmoide continua, monótona	Existencial
$\mathcal{C}(\mathcal{X}, \mu)$, $\mathcal{X}$ compacto [Hornik 1991]	L_∞	ψ continua, acotada μ medida finita,	Existencial Th. Hahn-Banach
$\mathcal{C}(\mathbb{R}^d)$ [Leshno *et al.* 1993]	L_∞	$\psi \in L_\infty^{\text{loc}}(\mathbb{R}^d)$ $\mu_{\text{L}}(\text{cl}(\Upsilon_\psi)) = 0$ con Υ_ψ cto. discont. ψ no polinomio c.p.d	Existencial
$L_1(\mathcal{X})$, $\mathcal{X}$ compacto [Cybenko 1989]	L_1	$\psi \in L^1(\mathcal{X})$ $\int \psi(t)dt \neq 0$	Existencial Th. Hahn-Banach
$L_2(\mathcal{X})$, $\mathcal{X}$ compacto [Carroll y Dickinson 1989]	L_1	ψ continua, sigmoide	Constructiva Transf. de Radon
$L_p(\mathcal{X}, \mu)$, $\mathcal{X}$ compacto, $p \in [1, \infty)$, μ medida finita [Hornik *et al.* 1989]	L_p	ψ sigmoide medible, monótona	Aprox. Trigonométrica
$L_p(\mathbb{R}^d, \mu)$, $p \in [1, \infty)$ [Hornik 1991]	L_p	ψ acotada μ medida finita	Existencial Th. Hahn-Banach
$L_p(\mu)$, $p \in [1, \infty)$ $\mu \in [0, \infty)$ sop. comp. abs. cont. resp. μ_{L} [Leshno *et al.* 1993]	L_p	$\psi \in L_\infty^{\text{loc}}(\mathbb{R}^d)$, $\mu_{\text{Leb.}}(\text{cl}(\Upsilon_\psi)) = 0$ con Υ_ψ cto. discont. ψ no polinomio c.p.d	Existencial
$\mathcal{C}^r(\mathcal{X}, \mu)$, $\mathcal{X}$ compacto [Hornik 1991]	$L_{p,r}$ $L_{\infty,r}$	$\psi \in \mathcal{C}^r(\mathcal{X})$ acotada μ medida finita	Existencial Th. Hahn-Banach
$\mathcal{C}^r(\mathbb{R}^d, \mu)$ [Hornik 1991]	$L_{p,r}$	$\psi \in \mathcal{C}^r(\mathbb{R}^d)$ $D^{\boldsymbol{\alpha}}\psi$ acotada, $\|\boldsymbol{\alpha}\| \leq r$ μ medida finita	Existencial Th. Hahn-Banach

Tabla 4. Resultados más relevantes sobre las posibilidades aproximadoras de las redes MLP. Para el significado de los términos y nuevos símbolos , véase el anexo C. En todos los casos se supone que ψ es no constante.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En la demostración del teorema anterior, el error de aproximación tiende a cero debido a la densidad en $L_p(\mathsf{P}_\mathsf{X})$ de las redes MLP con activación sigmoide monótona no decreciente, (p. ej. [Hornik 1991, th. 1], tabla 4).

El teorema anterior formulado en términos del problema de regularización, podría aplicarse para establecer la consistencia universal en $L_p(\mathsf{P}_\mathsf{Z})$ de las redes MLP con activación sigmoide monótona y no decreciente, entrenadas mediante regularización (*weight decay* –véase el algoritmo *back-propagation* más adelante) siempre que la sucesión de parámetros regularizadores verifique $\{\lambda_n\} \to 0$ cuando $n \to \infty$ (aunque no conocemos resultados sobre la tasa necesaria de convergencia de la sucesión $\{\lambda_n\}$).

Otros resultados al respecto utilizan ya un algoritmo concreto (*greedy*, p. ej. [Barron 1994, th. 3]) que permite obtener tasas de convergencia, y que presentaremos al tratar dicho algoritmo.

Por otra parte, si en un problema de clasificación, las probabilidades a posteriori se estiman mediante una red neuronal MLP $\hat{f}_n$ utilizando el riesgo $R(f) = \|\mathsf{Y} - f(\mathsf{X})\|_{L_p(\mathsf{P}_\mathsf{Z})}$, con $p = 1, 2$, entonces, combinando el teorema anterior con el teorema 33 se obtiene que la regla de clasificación *plug-in* correspondiente $g_{\hat{f}_n}$ es estimador consistente de la regla de Bayes.

Sin embargo, para muestras de tamaño moderado, el riesgo empírico basado en la pérdida $\ell(y, f(\mathbf{x})) = |y - f(\mathbf{x})|^p$ con $p = 1, 2$, no produce tan buenos resultados como el basado en la pérdida 0-1.

Intuitivamente, ello es debido a que las primeras están orientadas a un problema más ambicioso que la mera clasificación (la estimación de las probabilidades a posteriori) y, ante pocos datos, es mejor aprovechar la información disponible para abordar bien problemas sencillos en lugar de otros más complejos con menor fiabilidad (véase discusión al respecto en [Vapnik 1995, sec. 1.9]).

Para estos casos, por ejemplo [Faragó y Lugosi 1993, th. 1] prueban la consistencia universal de las reglas de clasificación construidas mediante redes MLP con activaciones dicotómicas en el nivel oculto, cuando se utiliza la pérdida 0-1.

II.5 Arquitecturas MLF de Base Radial

Las primeras manifestaciones de este tipo de arquitectura son las redes neuronales de base radial (RBF) que surgen por primera vez en la literatura en los años sesenta, en problemas de clasificación, con el método de las funciones potenciales de [Aizerman *et al.* 1964*b*] y [Aizerman *et al.* 1964*a*] inspirado en la estimación no paramétrica de densidades [Parzen 1962].

Sin embargo, no fué hasta finales de los años ochenta cuando este tipo de redes comenzaron a estudiarse sistemáticamente ([Broomhead y Lowe 1988], [Moody y Darken 1988], [Moody y Darken 1989], [Specht 1990]) con el fin de explorar alternativas a las funciones de activación del nivel oculto de las redes MLP, que evitasen la problemática de múltiples mínimos locales de estas redes.

Durante este período, las arquitecturas de base radial se consolidan debido a su facilidad de entrenamiento y a su parentesco con técnicas de otras áreas, que las convierten en un nexo de unión entre ellas:

1. Las teorías de interpolación y de aproximación de funciones, [Powell 1985], [Powell 1987] y [Powell 1992], ([Broomhead y Lowe 1988]).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2. La estadística no paramétrica y, en particular, la estimación tipo núcleo iniciada en [Parzen 1962] y [Cacoullos 1966], ([Moody y Darken 1988], [Moody y Darken 1989], [Specht 1990], [Schioler y Hartmann 1992]).

3. Las técnicas de regularización con enfoque [Tikhonov y Arsenin 1977], ([Poggio y Girosi 1990]).

4. Las técnicas *wavelet* con quiénes comparten el concepto de escala y que han inspirado la construcción de arquitecturas de este tipo: redes *wavelet* monoescala ([Boubez 1993], [Kugarajah y Zhang 1995], [Zhang 1997]).

Finalmente, la aparición a mediados de los años noventa de las *support vector machines* (SVM) ([Boser *et al.* 1992], [Cortes y Vapnik 1995]), apuntaló de forma definitiva el uso de este tipo de arquitecturas.

A diferencia de las redes MLP, para la que hemos realizado una presentación unificada bajo el marco del problema de regresión por cuanto, con las arquitecturas tratadas, los problemas de clasificación podrían reducirse a aquél[51], las redes RBF incluyen algún modelo específico para clasificación por lo que, ambos problemas, serán presentados por separado.

II.5.1 Redes RBF

Redes RBF para Clasificación

Los comentarios realizados para la red MLP en la construcción del estimador *plug-in* de la Regla de Bayes, son igualmente aplicables a una red con funciones de transferencia radiales en el nivel oculto, (denominadas usualmente **funciones básicas**).

En teoría, el nivel de salida admite funciones logísticas (o softmax para varias clases), o funciones lineales, y el entrenamiento puede realizarse con función de pérdida cuadrática o de entropía cruzada, con las propiedades analizadas anteriormente para cada caso.

Para ciertos modelos de red con activaciones radiales prefijadas en el nivel oculto, suele utilizarse pérdida cuadrática y activaciones lineales en salida debido a que se obtiene un problema de regresión generalizado, fácilmente resoluble.

Si las activaciones ocultas poseen parámetros a estimar (modelo no lineal en parámetros), la ventaja anterior no se produce, por lo que sería más recomendable utilizar la pérdida de entropía cruzada (máxima verosimilitud) e incluso activaciones sigmoides en salida. Sin embargo, es costumbre utilizar en todos los casos activaciones lineales y función de pérdida cuadrática, teniendo en cuenta que ello también produce buenos resultados.

La **red neuronal RBF** para clasificación puede fundamentarse en el siguiente análisis de [Bishop 1995, sec. 5.7]: supongamos un problema de clasificación con s clases para el que se desea estimar las probabilidades a posteriori $\mathsf{p}(\mathcal{C}_j|\mathbf{x})$, $j = 1, ..., s$ mediante una red con una neurona oculta por cada clase, con activación $\psi_j(\mathbf{x})$. Por el teorema de Bayes:

$$\mathsf{p}(\mathcal{C}_j|\mathbf{x}) = \frac{\mathsf{p}(\mathbf{x}|\mathcal{C}_j)\mathsf{p}(\mathcal{C}_j)}{\mathsf{p}(\mathbf{x})} \tag{II.36}$$

[51] Como ya hemos indicado anteriormente, estudiamos los modelos MLF con una sóla capa oculta, es decir, sin la capa de salida con activación dicotómica, necesaria para implementar la regla de decisión.

pero esta expresión puede verse como la salida j-ésima de una red con $f_j(\mathbf{x}) = c_j\psi_j(\mathbf{x})$, siendo:

$$\begin{aligned} \psi_j(\mathbf{x}) &= \frac{\mathsf{p}(\mathbf{x}|\mathcal{C}_j)}{\mathsf{p}(\mathbf{x})} \\ c_j &= \mathsf{p}(\mathcal{C}_j) \end{aligned}$$

Con ello, una vez entrenada la red neuronal, sus salidas serían estimaciones de dichas probabilidades a posteriori. Sin embargo, las probabilidades $\mathsf{p}(\mathbf{x}|\mathcal{C}_j)$ son desconocidas, con lo que es necesario estimarlas. Una posibilidad sería utilizar, para cada clase $\mathcal{C}_j$, una mixtura de la forma:

$$\mathsf{p}(\mathbf{x}|\mathcal{C}_j) \approx \sum_{i=1}^{h_j} \mathsf{p}(\mathbf{x}|i)\mathsf{p}(i|\mathcal{C}_j)$$

donde $\mathsf{p}(\mathbf{x}|i)$ son densidades con buenas propiedades aproximadoras. Ello equivaldría a configurar h_j neuronas en el nivel oculto y, teniendo en cuenta todas las clases, un total de $h_1 + \cdots + h_s$ neuronas ocultas. Sin embargo, es esperable que, si su número es suficientemente grande, algunas de las componentes de la mixtura sean comunes para todas las clases, (ello no representa un problema pues cada clase $\mathcal{C}_j$ utiliza un parámetro diferente $\mathsf{p}(i|\mathcal{C}_j)$ para cada componente i). Con ello:

$$\mathsf{p}(\mathbf{x}|\mathcal{C}_j) \approx \sum_{i=1}^{h} \mathsf{p}(\mathbf{x}|i)\mathsf{p}(i|\mathcal{C}_j),\ j = 1, ..., s \tag{II.37}$$

Sustituyendo esta expresión en (II.36), se obtiene:

$$\mathsf{p}(\mathcal{C}_j|\mathbf{x}) = \frac{\mathsf{p}(\mathbf{x}|\mathcal{C}_j)\mathsf{p}(\mathcal{C}_j)}{\mathsf{p}(\mathbf{x})} \approx \sum_{i=1}^{h} \frac{\mathsf{p}(\mathbf{x}|i)\mathsf{p}(i|\mathcal{C}_j)\mathsf{p}(\mathcal{C}_j)}{\mathsf{p}(\mathbf{x})}\frac{\mathsf{p}(i)}{\mathsf{p}(i)} = \sum_{i=1}^{h} c_{ji}\psi_i(\mathbf{x}) \tag{38a}$$

$$\text{siendo: } \begin{cases} \psi_i(\mathbf{x}) = \dfrac{\mathsf{p}(\mathbf{x}|i)\mathsf{p}(i)}{\mathsf{p}(\mathbf{x})} = \mathsf{p}(i|\mathbf{x}) \\ c_{ji} = \dfrac{\mathsf{p}(i|\mathcal{C}_j)\mathsf{p}(\mathcal{C}_j)}{\mathsf{p}(i)} = \mathsf{p}(\mathcal{C}_j|i) \end{cases} \tag{38b}$$

expresión, que puede verse como la salida j-ésima de la red, donde $\mathsf{p}(i|\mathbf{x})$ es la probabilidad de que un punto $\mathbf{x}$ prefijado, provenga de la *partición* i-ésima de la mixtura.

El análisis anterior es válido en general aunque las funciones básicas ψ_i no sean **radiales**, pero este es el caso más utilizado. Si estas probabilidades *a posteriori* se modelizan por ejemplo como gausianas:

$$\psi_i(\mathbf{x}) = \mathsf{p}(i|\mathbf{x}) \propto \exp\{-\frac{1}{2}\left\|\mathbf{x} - \mathbf{w}_i\right\|^2_{\mathbf{\Sigma}_i}\}$$

donde $\|\mathbf{x} - \mathbf{w}_i\|^2_{\mathbf{\Sigma}_i} = (\mathbf{x} - \mathbf{w}_i)^T\Sigma_i^{-1}(\mathbf{x} - \mathbf{w}_i)$, y Σ_i son matrices simétricas definidas positivas, $i = 1 : h$, entonces los parámetros de posición $\mathbf{w}_i$ (a estimar durante el entrenamiento) pueden interpretarse como ejemplos-prototipo del espacio de entrada, relevantes para la clasificación, que actuarían a modo de **centros** de *clusters* de puntos *similares* a los efectos de su clasificación en la categorías $\mathcal{C}_j$ (y que recibirían, por tanto, un tratamiento similar $c_{ji} = \mathsf{p}(\mathcal{C}_j|i)$ en el nivel de salida).

En este sentido, el entrenamiento de la red consiste en estructurar el espacio de entrada en *clusters* adecuados al problema. La matriz $\mathbf{\Sigma}_i$ actuaría como parámetro de control de la **distancia** implementada por $\psi_i(\mathbf{x})$, determinando el *grado* de pertenencia de cada $\mathbf{x}$ al *cluster* representado por $\mathbf{w}_i$.

USC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

El número de términos h de la expansión (II.38a) puede ser menor que el tamaño muestral n (y, de hecho, lo es de forma usual). Para referirnos a este tipo de arquitecturas radiales en las que $h \leq n$ y diferenciarlas de otras que imponen $h = n$, utilizaremos el acrónimo RBF tanto en el problema de clasificación como en el de regresión. La única excepción, es la arquitectura tratada a continuación, agrupada tradicionalmente en este grupo, pero que se conoce también como red probabilística, (PN).

Redes Probabilísticas

Un caso particular de lo anterior en el que $h = n$, es la denominada **red probabilistica** (PN), [Specht 1990] que se obtiene cuando las probabilidades condicionadas (II.37) se estiman mediante estimadores tipo núcleo, (p. ej. *Parzen windows*, [Parzen 1962], [Cacoullos 1966]).

Con una muestra $\mathbf{z}^n$ con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i = 1 : n$, siendo n_j el número de casos en la clase $\mathcal{C}_j$, estas estimaciones serían:

$$\mathsf{p}(\mathbf{x}|\mathcal{C}_j) \approx \frac{1}{n_j} \sum_{i \in \mathcal{C}_j} k_{\mathbf{\Sigma}_{ij}}(\mathbf{x} - \mathbf{x}_i) \tag{II.39}$$

que puede verse como la media de los valores $k_{\mathbf{\Sigma}_{ij}}(\mathbf{x} - \mathbf{x}_i)$, $i \in \mathcal{C}_j$, donde $k_{\mathbf{\Sigma}_{ij}}$ son funciones de densidad multivariante (radiales como caso particular[52]: $k_{\mathbf{\Sigma}_{ij}}(\mathbf{x} - \mathbf{x}_i) \equiv k(\|\mathbf{x} - \mathbf{x}_i\|_{\mathbf{\Sigma}_{ij}})$ para alguna densidad univariante k satisfaciendo ciertas condiciones para lograr buenas propiedades asintóticas (p. ej. [Wand y Jones 1995, p. 95]). Por simplicidad, en lo que sigue suponemos una misma matriz $\mathbf{\Sigma} = \mathbf{\Sigma}_{ij}\ \forall i, j$.

Por el teorema de Bayes:

$$\mathsf{p}(\mathcal{C}_j|\mathbf{x}) = \frac{\mathsf{p}(\mathbf{x}|\mathcal{C}_j)\mathsf{p}(\mathcal{C}_j)}{\sum_{l=1}^{s} \mathsf{p}(\mathbf{x}|\mathcal{C}_l)\mathsf{p}(\mathcal{C}_l)} \approx \frac{\mathsf{p}(\mathcal{C}_j)\frac{1}{n_j}\sum_{i \in \mathcal{C}_j} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)}{\sum_{l=1}^{s} \mathsf{p}(\mathcal{C}_l)\frac{1}{n_l}\sum_{i \in \mathcal{C}_l} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)}$$

expresión que, si las probabilidades a priori pueden estimarse mediante $\hat{\mathsf{p}}(\mathcal{C}_j) = n_j/n$, se convierte en:

$$\mathsf{p}(\mathcal{C}_j|\mathbf{x}) \approx \frac{\sum_{i \in \mathcal{C}_j} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)}{\sum_{i=1}^{n} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)} \propto \sum_{i \in \mathcal{C}_j} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i) \tag{II.40}$$

Si el objetivo fuese sólo clasificar (no estimar probabilidades a posteriori), la red neuronal no requeriría funciones de activación normalizadas en el nivel de salida, con lo que la estimación *plug-in* de la regla de Bayes podría construirse mediante:

$$\text{asignar } \mathbf{x}_i \text{ a } \mathcal{C}_{j_0} \text{ si } j_0 = \arg \max_{j \in \{1:s\}} \sum_{i \in \mathcal{C}_j} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)$$

El encaje de la red probabilística en el esquema general presentado con anterioridad es claro: si se compara (II.39) con (II.37) con la codificación estándar $y_{ij} = 1_{\{i \in \mathcal{C}_j\}}$, se tiene:

$$\mathsf{p}(\mathbf{x}|\mathcal{C}_j) \approx \frac{1}{n_j} \sum_{i \in \mathcal{C}_j} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i) = \sum_{i=1}^{n} \frac{y_{ij}}{n_j} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i) \equiv \sum_{i=1}^{h} \mathsf{p}(\mathbf{x}|i)\mathsf{p}(i|\mathcal{C}_j)$$

[52] Otra forma común de generar núcleos multivariantes es el producto tensorial de núcleos univariantes (p. ej. [Wand y Jones 1995, cap. 4]): $k(x - x_i) = \prod_{j=1}^{d} k_{\sigma_{ji}}(x_j - x_{ji})$.

con lo que se tienen las equivalencias: $h = n$ y $\mathsf{p}(i|\mathcal{C}_j) = \frac{y_{ij}}{n_j}$, que es $1/n_j$ si $y_{ij} = 1$ ($i \in \mathcal{C}_j$) y cero en caso contrario, (nótese que $\sum_i y_{ij} = \frac{n_j}{n_j} = 1$). Por tanto, cada $\mathbf{x}_i$ de la muestra actúa como centro del *cluster* i-ésimo, con $\mathsf{p}(\mathbf{x}|i) = k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)$ y $\mathsf{p}(i) = \frac{1}{n}$ (los datos de la muestra son ahora valores prefijados). Además, como las clases son excluyentes:

$$\mathsf{p}(\mathbf{x}) = \sum_{j=1}^{s} \mathsf{p}(\mathbf{x}|\mathcal{C}_j)\mathsf{p}(\mathcal{C}_j) = \sum_{j=1}^{s} \frac{1}{n_j} \sum_{l \in \mathcal{C}_j} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_l)\frac{n_j}{n} = \frac{1}{n}\sum_{l=1}^{n} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_l)$$

De esta forma, utilizando (II.38b):

$$\begin{aligned} \psi_i(\mathbf{x}) &= \mathsf{p}(i|\mathbf{x}) = \frac{\mathsf{p}(\mathbf{x}|i)\mathsf{p}(i)}{\mathsf{p}(\mathbf{x})} \approx \frac{\frac{1}{n}k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)}{\frac{1}{n}\sum_{l=1}^{n} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_l)} \\ c_{ji} &= \mathsf{p}(\mathcal{C}_j|i) = \frac{\mathsf{p}(i|\mathcal{C}_j)\mathsf{p}(\mathcal{C}_j)}{\mathsf{p}(i)} \approx \frac{\frac{y_{ij}}{n_j}\frac{n_j}{n}}{\frac{1}{n}} = y_{ij} \end{aligned}$$

con lo que, a través de (II.38a) se obtiene la **red probabilística**:

$$f_j(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=1}^{n} c_{ji}\psi_i(\mathbf{x}) = \sum_{i=1}^{n} \frac{y_{ij}k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)}{\sum_{l=1}^{n} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_l)} \propto \sum_{i=1}^{n} y_{ij}k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i),\ j = 1:s \qquad \text{(II.41)}$$

donde la última expresión se utiliza si sólo interesa clasificar, no estimar las probabilidades a posteriori.

La expresión (II.41) indica que, con la codificación $y_{ij} = 1_{\{i \in \mathcal{C}_j\}}$, la red probabilística construye las funciones discriminantes mediante la realización en paralelo de s regresiones tipo núcleo ([Nadaraya 1964], [Watson 1964], [Priestley y Chao 1972], [Stone 1977*a*], [Gasser y Müller 1979], [Cleveland 1979] y p. ej. [Härdle 1990], [Wand y Jones 1995]).

El atractivo de estas redes reside en su inmediato entrenamiento, pues $c_{ji} = y_{ij}$, y en sus buenas propiedades estadísticas, que se heredan de las del estimador tipo núcleo utilizado. Otra cuestión es la selección del parámetro ventana $\mathbf{\Sigma}$, que requiere métodos de selección del modelo.

En contrapartida, su representación no puede considerarse idónea pues utiliza tantos centros como datos en la muestra. Esto puede suponer una debilidad importante[53] en problemas con espacio de entrada de gran dimensión, en los que se requieren grandes tamaños muestrales.

Bajo el esquema general de las ecuaciones (II.37) y (II.38), es posible realizar una estimación más parsimoniosa de las probabilidades condicionadas, aunque ello debe pagarse con la utilización de algún algoritmo de entrenamiento para estimar los parámetros de la red en (II.38). Ejemplos de ello, son las redes RBF para regresión del siguiente apartado, cuando se aplican a la estimación de las probabilidades a posteriori.

Red RBF para Regresión

Si el enfoque anterior seguido por la red probabilística se aplica a una variable respuesta continua $\mathcal{Y} \subset \mathbb{R}$ (suponemos ahora salida unidimensional), se tiene la regresión tipo núcleo tradicional

[53] [Moody y Darken 1989, p. 282] catalogan un modelo de este tipo como *no neural* pues *no puede ser fácilmente implementable mediante una red adaptable con arquitectura y tamaño fijos.*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

que, así, puede verse como una red neuronal RBF con $h = n$:

$$f_{k,\mathbf{\Sigma}}(\mathbf{x}) = \sum_{i=1}^{n} c_i \frac{k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)}{\sum_{l=1}^{n} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_l)} \tag{II.42}$$

y con entrenamiento inmediato, $c_i = y_i$, $i = 1 : n$, pero con los mismos problemas de representación mencionados anteriormente.

Una circunstancia similar se presenta en el **problema de interpolación con funciones de base radial**, (p. ej. [Micchelli 1986], [Dyn *et al.* 1986], [Powell 1987], [Dyn 1987], [Madych y Nelson 1988], [Dyn 1989], [Madych y Nelson 1990]; una revisión puede verse en [Powell 1992]), en el que, dado un conjunto de puntos $\mathbf{z}^n = (\mathbf{z}_1, ..., \mathbf{z}_n)$ con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, se pretende determinar una función $f : \mathbb{R}^d \to \mathbb{R}$ que verifique las condiciones de interpolación $f(\mathbf{x}_i) = y_i$, $i = 1 : n$.

La técnica de interpolación con funciones de base radial, utiliza también n funciones básicas radiales prefijadas, de la forma $\zeta(\|\mathbf{x} - \mathbf{x}_i\|_{\mathbf{\Sigma}_i})$, y propone soluciones de la forma:

$$f_{\zeta,\mathbf{\Sigma}_i}(\mathbf{x}; \mathbf{c}) = \sum_{i=1}^{n} c_i \zeta(\|\mathbf{x} - \mathbf{x}_i\|_{\mathbf{\Sigma}_i})$$

donde los parámetros c_i se obtienen de su sustitución en las condiciones de interpolación: $\mathbf{Kc} = \mathbf{y}$, donde $\mathbf{c} = (c_1, ..., c_n)^T$, $\mathbf{y} = (y_1, ..., y_n)^T$ y $\mathbf{K}$ es la matriz de *distancias* entre los puntos con $(\mathbf{K})_{ij} = \zeta(\|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{\Sigma}_j})$. La generalización a problemas con variable respuesta s-dimensional es inmediata, dando lugar a s problemas simultáneos de interpolación.

Si los puntos $\{\mathbf{x}_i\}_{i=1}^n$ son todos distintos, bajo determinadas condiciones para la función $k(\mathbf{x}_1, \mathbf{x}_2) = \zeta(\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbf{\Sigma}_i})$, (p. ej. estrictamente definida positiva, [Micchelli 1986], Anexo B), se garantiza la invertibilidad de $\mathbf{K}$, de tal manera que el problema de interpolación posee solución única[54]: $\mathbf{c} = \mathbf{K}^{-1}\mathbf{y}$.

Es claro que el problema de interpolación es un problema de regresión en el que los datos no están sujetos a ruido y, de hecho, la solución a dicho problema pasa exactamente por los puntos interpolados. Sin embargo, si los datos están sujetos a perturbaciones aleatorias es necesario evitar que dichas perturbaciones sean tratadas como información y sean incorporadas al modelo estimado.

Las redes neuronales de base radial para regresión surgen de la adaptación [Broomhead y Lowe 1988, sec. 2] del método de interpolación, a situaciones con ruido en los datos. Para ello, estos autores introducen las siguientes modificaciones:

1. Se reduce el número de funciones básicas (el número de centros) produciendo una solución que sólamente puede aproximarse a los datos en lugar de interpolarlos. Además, los centros pueden constituir un subconjunto de los datos de la muestra o bien, ser elegidos a partir de alguna información a priori.

2. Puesto que ahora el ajuste no es exacto, se incluye un parámetro adicional (*bias*), para compensar la diferencia entre la media del valor de las funciones básicas sobre los datos y la media de la variable respuesta.

De esta forma, se obtiene un modelo de regresión lineal generalizada donde las variables regresoras son transformaciones no lineales de las variables de entrada.

USC Universidade de Santiago de Compostela

[54] Dicha unicidad no está garantizada, si sólo se exige que k sea definida positiva ($\mathbf{K}$ semidefinida positiva).

En ese mismo período, [Moody y Darken 1988] proponen independientemente una red análoga, pero sin utilizar funciones de transferencia prefijadas en el nivel oculto, sino haciéndolas dependientes de un parámetro ventana variable.

En general, la arquitectura **RBF para regresión** posee la forma general:

$$\begin{aligned} f(\mathbf{x};\boldsymbol{\theta}) &= \sum_{j=1}^{h} c_j \psi_j(\mathbf{x}) + c_0 \\ \text{con: } \psi_j(\mathbf{x}) &= \zeta(\|\mathbf{x}-\mathbf{w}_j\|_{\boldsymbol{\Sigma}_j}) \end{aligned} \tag{II.43}$$

donde, en general, $h \leq n$ (aunque frecuentemente $h << n$), ζ es alguna función válida para el problema de interpolación y $\boldsymbol{\theta}$ incluye los parámetros de la red: h, $\bar{\mathbf{c}}$, $\mathbf{w}_j$, $\boldsymbol{\Sigma}_j$, $j = 1:h$.

Siendo su motivación inicial la obtención de un problema de regresión lineal generalizado, la función de pérdida utilizada suele ser la pérdida cuadrática, si bien, cuando los centros $\mathbf{w}_j$ y las matrices $\boldsymbol{\Sigma}_j$ se incluyen en el problema de optimización, éste posee múltiples mínimos locales y dicha pérdida no aporta las ventajas usuales.

Cuando los centros $\mathbf{w}_j$ se eligen como un subconjunto de la muestra, (normalmente en este caso, los parámetros $\boldsymbol{\Sigma}_j$ están también prefijados), se emplea algún método de selección de variables en la regresión lineal generalizada resultante.

En la arquitectura (II.43) no se requiere que las funciones de transferencia ψ sean funciones de densidad, como en la regresión tipo núcleo, pues cualquier factor necesario es absorbido perfectamente por los coeficientes c_j durante el entrenamiento.

La función más utilizada como función de activación es la gausiana, con lo que las funciones de transferencia son:

$$\psi_j(\mathbf{x}) \propto \exp\{-\frac{1}{2}\|\mathbf{x}-\mathbf{w}_j\|^2_{\boldsymbol{\Sigma}_j}\} = \exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{w}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\mathbf{w}_j)\}$$

donde $\boldsymbol{\Sigma}_j$ es alguna matriz simétrica y definida positiva. Simplificaciones importantes de este modelo son:

1. Cuando $\boldsymbol{\Sigma}_j = \sigma_j^2\mathbf{I}_d$, $j \in \{1:h\}$, lo que significa que la función de transferencia es esférica con amplitud determinada por σ_j^2.

2. Cuando $\boldsymbol{\Sigma}_j = \text{diag}(\sigma_{j1}^2, ..., \sigma_{jd}^2)$, $j \in \{1:h\}$, con lo que la función de transferencia es elíptica con ejes paralelos a los ejes de coordenadas, y con amplitudes determinadas por σ_{ji}^2, en cada dimensión i-ésima, $i = 1:d$.

Si $\boldsymbol{\Sigma}_j$ no es diagonal pero simétrica, admite una descomposición de la forma (Anexo G) $\boldsymbol{\Sigma}_j = \mathbf{V}_j\mathbf{D}_j\mathbf{V}_j^T$ donde $\mathbf{D}_j$ es una matriz diagonal que determina, como antes, la forma y el tamaño del campo de atracción de la función básica, y $\mathbf{V}_j$ es una matriz de rotación que determina su orientación.

Un caso particular es el caso $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$, $j = 1:h$, en el que, el número de parámetros de cada función ψ_j, es $d(d+3)/2$, en lugar de los $d+1$ del caso anterior (d componentes de $\mathbf{w}_j$ más el parámetro de escala).

Lo normal es que, a mayor complejidad de $\boldsymbol{\Sigma}_j$ menor sea el número de funciones básicas necesarias, sin embargo, en la práctica, es necesario mantener un equilibrio entre muchos parámetros con pocas funciones más flexibles y menos parámetros con más funciones menos flexibles.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Estrategias de Entrenamiento de la red RBF

Al margen de que con posterioridad se traten los algoritmos de entrenamiento, conviene realizar ahora una breve descripción de las diferentes estrategias que pueden seguirse con las redes RBF.

Básicamente, estas estrategias pueden clasificarse en dos grandes grupos, según que: se disponga de conjunto prefijado de funciones básicas (parámetros fijos), lo que produce un **modelo lineal** en parámetros, o, por el contrario, las funciones básicas dependan de parámetros de posición y escala a estimar, lo que produce un **modelo no lineal**.

En el primer caso, suelen utilizarse como centros potenciales todos los datos de la muestra y valores prefijados como parámetros de escala. Con ello, el problema se reduce a uno de regresión lineal generalizada, con lo que su resolución se realiza mediante algún método de selección de variables (ocultas): $u_j = \psi_j(\mathbf{x}), j = 1 : h$.

El segundo caso es totalmente análogo al de las redes MLP vistas anteriormente y admite dos enfoques:

1. Aprendizaje en una sóla fase, consistente en la estimación global de todos los parámetros de la red mediante la resolución de un problema de optimización no lineal con múltiples mínimos locales, mediante algún algoritmo del tipo *back-propagation* o sus variantes, como en el caso de las redes MLP.

   Esta estrategia es la más adecuada si se utilizan parámetros de escala matriciales o diferentes para las distintas funciones de transferencia, pues, en caso contrario, es necesario utilizar métodos de selección del modelo en relación con dichos parámetros, lo que resulta inabordable con parámetros multidimensionales.

2. Aprendizaje en dos fases:
   (a) Una primera fase de tipo no supervisado, dedicada a la estimación de los parámetros del nivel oculto, normalmente mediante técnicas de *clustering*, (k-medias, etc.) o de estimación de la densidad de X mediante mixturas con el método de la máxima verosimilitud utilizando el algoritmo E-M (*expectation-minimization*).
   (b) Una segunda fase dedicada a la estimación de los parámetros del nivel de salida, que ahora resulta un problema de regresión lineal que se resuelve mediante los mismos algoritmos empleados para el caso de funciones básicas prefijadas.

La debilidad principal del aprendizaje en dos fases, es que, la primera fase no supervisada no tiene en cuenta la variable respuesta dando lugar a una estructuración del espacio de entrada que no tiene por qué ser la idónea para el problema de regresión particular, (recuérdese el análisis sobre el significado del nivel oculto de una red MLF, pág. 104). Para corregir esta debilidad, se han propuesto métodos híbridos de entrenamiento utilizando dos fases que tienen en cuenta la variable respuesta, ([Uykan *et al.* 2000] y referencias allí citadas). A pesar de ello, el método más utilizado en este caso, sigue siendo el entrenamiento en una sóla fase, aunque también produce soluciones subóptimas.

RBF para Regresión y Regresión Tipo Núcleo

Anteriormente, mencionamos que la regresión tipo núcleo podría verse como una red neuronal RBF con entrenamiento inmediato y mencionamos las diferencias de su arquitectura en relación con la RBF, ($h = n$ y utilización de funciones básicas normalizadas). Asimismo, vimos la cercanía

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

entre las redes probabilísticas y la regresión no paramétrica tipo núcleo aplicada al problema de clasificación.

En este apartado, analizamos la red RBF como estimador de la función de regresión y lo comparamos con el estimador tipo núcleo. Para ello, suponemos una red RBF de tipo lineal, es decir, con un conjunto de funciones de transferencia del nivel oculto, prefijadas, $\psi_1, ..., \psi_h$, dando lugar a un problema de regresión lineal generalizada sobre las variables ocultas: $\psi_j = \psi_j(\mathbf{x})$, $j = 1 : h$.

La red RBF como Estimador No Paramétrico Lineal

En las hipótesis de este apartado, supuesto un problema de regresión $\mathsf{Y} = \mathsf{f}(\mathsf{X}) + \varepsilon$ con $\mathcal{Y} = \mathbb{R}$ y $\mathcal{X} \subset \mathbb{R}^d$, con ε ruido aleatorio de media cero y una muestra aleatoria simple $\mathbf{z}^n = (\mathbf{z}_1, ..., \mathbf{z}_n)$ con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, la red RBF entrenada posee la forma:

$$\hat{\mathbf{f}}_n = \mathbf{\Psi}\bar{\mathbf{c}} = \mathbf{\Psi}(\mathbf{\Psi}^T\mathbf{\Psi})^{-1}\mathbf{\Psi}^T\mathbf{y} = \mathbf{S}\mathbf{y}$$

donde $\mathbf{\Psi}$ es la matriz de diseño (extendida con una primera columna de unos), $\bar{\mathbf{c}} = (c_0, \mathbf{c}^T)^T$, $\mathbf{y} = (y_1, ..., y_n)^T$ y $\mathbf{S}$ es la matriz *hat*, (p. ej. [Hoaglin y Welsch 1980], [Eubank 1984]), que es simétrica e idempotente, como corresponde a un estimador no paramétrico lineal de tipo proyección (p. ej. [Hastie y Tibshirani 1990]): la estimación $\hat{\mathbf{f}}_n$ es la proyección de $\mathbf{y}$ sobre el espacio $\langle\{\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_n\}\rangle$, siendo $\boldsymbol{\psi}_i^T = (\psi_0(\mathbf{x}_i), \psi_1(\mathbf{x}_i), ..., \psi_h(\mathbf{x}_i))$ con $\psi_0(\mathbf{x}_i) = 1$, $i = 1 : n$, las filas de $\mathbf{\Psi}$.

De esta forma, el grado de suavizado coincide[55] con la traza de la matriz $\mathrm{tr}(\mathbf{S}) = \mathrm{rango}(\mathbf{S})$ que coincide con el número de autovalores unidad de $\mathbf{S}$. (Al ser $\mathbf{S}$ una matriz proyección sus autovalores son unos o ceros, correspondientes a los autovectores que residen en dicho espacio o en su espacio ortogonal, respectivamente).

Por tanto, este rango sirve para determinar el número adecuado de neuronas que debe contener el nivel oculto: si se utilizan h neuronas y la matriz $\mathbf{S}$ resulta de rango $p < h$, ello significa que $h - p$ neuronas son redundantes y pueden eliminarse para conseguir el mismo grado de suavizado.

Si para un nuevo valor de $\mathbf{x}$, denotamos los puntos del espacio oculto (espacio de características o *features*) por:

$$\boldsymbol{\psi}(\mathbf{x}) = (\psi_0(\mathbf{x}), \psi_1(\mathbf{x}), ..., \psi_h(\mathbf{x}))^T$$

la estimación de la red RBF resulta:

$$\hat{f}_n(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})^T\bar{\mathbf{c}} = \boldsymbol{\psi}(\mathbf{x})^T(\mathbf{\Psi}^T\mathbf{\Psi})^{-1}\mathbf{\Psi}^T\mathbf{y} = \boldsymbol{\varphi}(\mathbf{x})^T\mathbf{y}$$

con $\boldsymbol{\varphi}(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})^T(\mathbf{\Psi}^T\mathbf{\Psi})^{-1}\mathbf{\Psi}^T = (\varphi_1(\mathbf{x}), ..., \varphi_n(\mathbf{x}))^T$, donde $\varphi_i(\mathbf{x}) \equiv \varphi_{\mathbf{x}}(\mathbf{x}_i)$, $i = 1 : n$, hacen el papel de **núcleos efectivos o equivalentes** ([Silverman 1984] y p. ej. [Wand y Jones 1995, sec. 5.6.2], [Hastie y Tibshirani 1990, sec. 2.8]) del estimador en el punto $\mathbf{x}$. Así, la red RBF entrenada mediante regresión lineal generalizada puede verse como un estimador tipo núcleo con las funciones $\varphi_{\mathbf{x}}(\cdot)$ como núcleos variables en cada punto $\mathbf{x}$.

Esta formulación de la regresión en términos de la variable respuesta y los núcleos, recibe el nombre de **formulación dual**, mientras que la formulación en términos de las covariables ψ_j se denomina **formulación primal**.

[55] Si $\mathbf{S}$ es idempotente como es el caso, se verifica que $\mathrm{tr}(\mathbf{S}) = \mathrm{tr}(2\mathbf{S} - \mathbf{S}\mathbf{S}^t) = \mathrm{tr}(\mathbf{S}\mathbf{S}^t)$, con lo que las tres medidas de suavizado coinciden, [Hastie y Tibshirani 1990, sec. 3.5].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

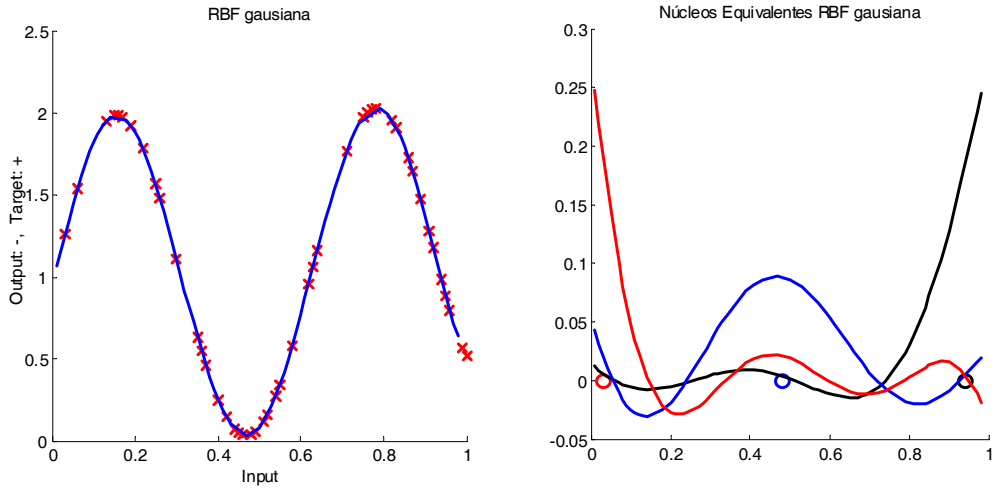


Figura 14. Izquierda: ajuste de una red RBF gausiana a una muestra de test procedente de una función sigmoide perturbada con ruido gausiano. Derecha: núcleos equivalentes en tres puntos de test señalados con un círculo. A cada punto le corresponde el núcleo con moda más cercana.

Con ello, el estimador RBF es un estimador esencialmente local a pesar de haber sido entrenado mediante una regresión de tipo global. Ello es debido a que la transformación que produce el nivel oculto es, implícitamente, de carácter local.

En la figura 14 (izquierda) puede verse el comportamiento, sobre una muestra de test, de una red RBF gausiana entrenada con una muestra de $n = 54$ datos procedentes de una función sinusoide perturbada con ruido gausiano. En el gráfico de la derecha de la misma figura se puede observar la forma de los núcleos equivalentes en tres puntos de test diferentes.

A pesar de que se utilizó un parámetro de escala $\sigma = 0{,}5$ constante para las funciones básicas de la expansión, los núcleos resultantes $\phi_{\mathbf{x}}$ poseen amplitud y forma diferentes para cada $\mathbf{x}$ (señalados con un círculo y ubicados próximos a la moda de sus respectivos núcleos) implementando así núcleos variables (p. ej. [Wand y Jones 1995, sec. 2.10.2]). En particular, puede observarse la forma de los núcleos en la frontera, con grandes valores en su proximidad y valores negativos en puntos más lejanos, al modo de lo que sucede con la regresión polinómico local ([Wand y Jones 1995, p. 135], [Fan y Gijbels 1996, sec. 3.2.5], [Loader 1999, p. 29]).

En la figura 15, se presenta la función de influencia del suavizado (elementos S_{ii} de la diagonal de la matriz *hat* $\mathbf{S}$) mostrando el mayor peso asignado por los núcleos a los puntos de la frontera, lo que muestra como las redes RBF corrigen el problema de sesgo que los estimadores tipo núcleo presentan en dichos puntos [Hart 1997, sec. 2.5] debido al truncamiento de sus núcleos.

En las figuras 16 a 17, se presenta el resultado de la utilización de redes RBF con otras funciones básicas radiales admisibles: la función multicuádrica inversa y el núcleo Epanechnikov, donde puede verse que los ajustes son menos suaves, al igual que las correspondientes funciones núcleo.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

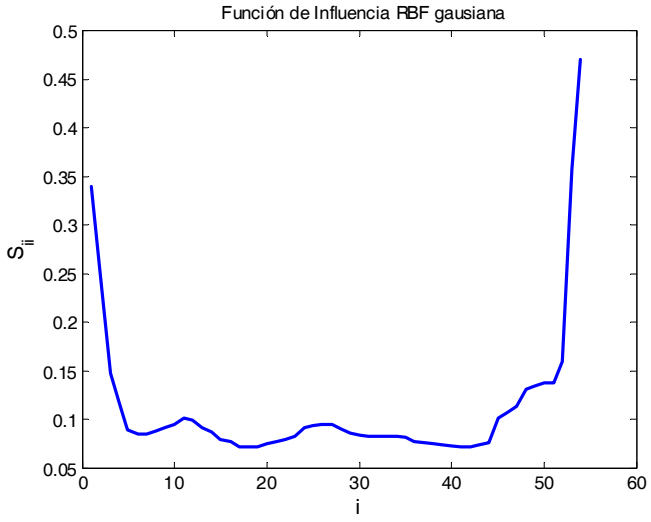


Figura 15. Función de influencia (i, S_{ii}) de una red RBF gausiana en una muestra de $n = 50$ datos de entrenamiento

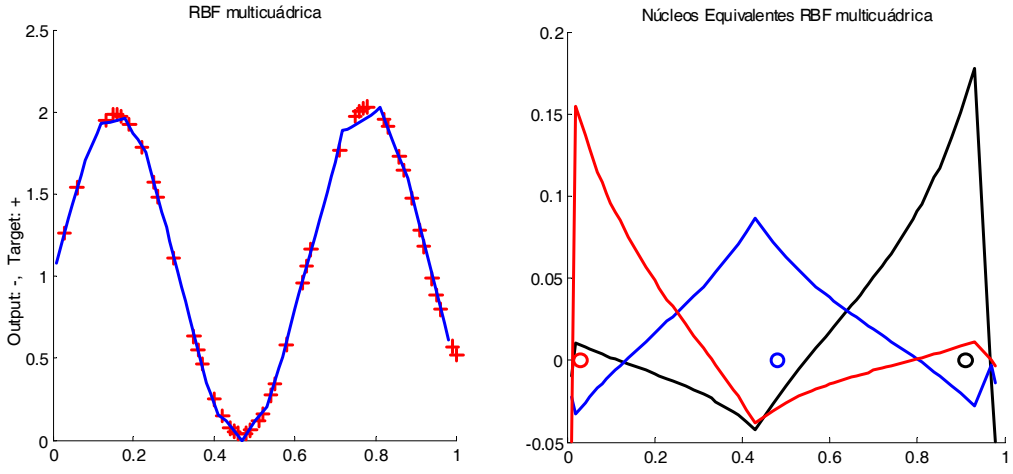


Figura 16. Izquierda: ajuste de una red RBF con funciones básicas multicuádrica inversa $\psi_j(\mathbf{x}) = (b^2 + \|\mathbf{x} - \mathbf{w}_j\|)^{-1/2}$, a una muestra de test procedente de una función sigmoide perturbada con ruido gausiano. Derecha: núcleos equivalentes en tres puntos de test señalados con un círculo. A cada punto le corresponde el núcleo con moda más cercana.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

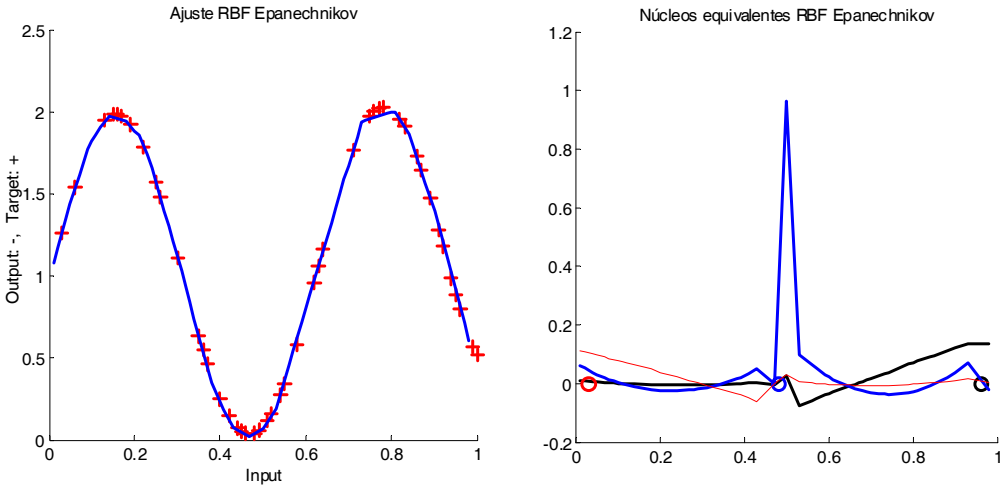


Figura 17. Izquierda: ajuste de una red RBF con funciones básicas Epanechnikov $\psi_j(\mathbf{x}) = 0{,}75(1 - \|\mathbf{x}-\mathbf{w}_j\|^2)1_{\{\|\mathbf{x}-\mathbf{w}_j\|\leq 1\}}$ a una muestra de test procedente de una función sigmoide perturbada con ruido gausiano. Derecha: núcleos equivalentes en tres puntos de test señalados con un círculo. A cada punto le corresponde el núcleo con moda más cercana.

El Estimador Tipo Núcleo como Red RBF

Un análisis en el sentido inverso puede realizarse para el estimador tipo núcleo de la regresión pues, si el núcleo utilizado en (II.42):

$$\psi_{k,\mathbf{\Sigma}}(\mathbf{x}-\mathbf{x}_i) = \frac{k_{\mathbf{\Sigma}}(\mathbf{x}-\mathbf{x}_i)}{\sum_{l=1}^{n} k_{\mathbf{\Sigma}}(\mathbf{x}-\mathbf{x}_l)} \equiv \psi(\mathbf{x},\mathbf{x}_i)$$

es continuo y simétrico en $L_2(\mathcal{X})$, con $\mathcal{X} \subset \mathbb{R}^d$ compacto, entonces ψ admite una descomposición de la forma (teorema de Mercer, Anexo B):

$$\psi(\mathbf{x},\mathbf{x}') = \sum_{l=1}^{\infty} \mu_l \phi_l(\mathbf{x})\phi_l(\mathbf{x}')$$

donde $\{\phi_l\}$ es una base ortogonal de $L_2(\mathcal{X})$ compuesta por los autovectores del operador integral $\int \psi(\mathbf{x},\mathbf{x}')\phi_l(\mathbf{x}')d\mathbf{x}' = \lambda_l\phi_l(\mathbf{x})$, siendo μ_l sus correspondientes autovalores, todos positivos, con la propiedad $\sum_{l=1}^{\infty}\lambda_l < \infty$.

Aplicando dicha propiedad a la expresión (II.42) con $c_i = y_i$, se obtiene:

$$f_{k,\mathbf{\Sigma}}(\mathbf{x}) = \sum_{i=1}^{n} y_i\psi(\mathbf{x},\mathbf{x}_i) = \sum_{i=1}^{n} y_i \sum_{l=1}^{\infty}\lambda_l\phi_l(\mathbf{x})\phi_l(\mathbf{x}_i) = \sum_{l=1}^{\infty}[\sum_{i=1}^{n} y_i\lambda_l\phi_l(\mathbf{x})]\phi_l(\mathbf{x}_i) = \sum_{l=1}^{\infty} c'_l\phi_l(\mathbf{x}_i) \tag{II.44}$$

donde hemos definido $c'_l = \sum_{i=1}^{n} y_i\lambda_l\phi_l(\mathbf{x})$, $l = 1:h$. De esta forma, si el número h de autovalores verifica $h < \infty$, se tiene la expresión de una red RBF con **funciones básicas equivalentes** ϕ_l (no necesariamente radiales) y, en caso contrario, el estimador tipo núcleo admite una expresión aproximada en términos de una red RBF, teniendo en cuenta que los autovalores tienden rápidamente a cero.

El desarrollo (II.44) muestra el paso desde el plano dual de la regresión, típico del estimador tipo núcleo, al plano primal, típico de la red RBF.

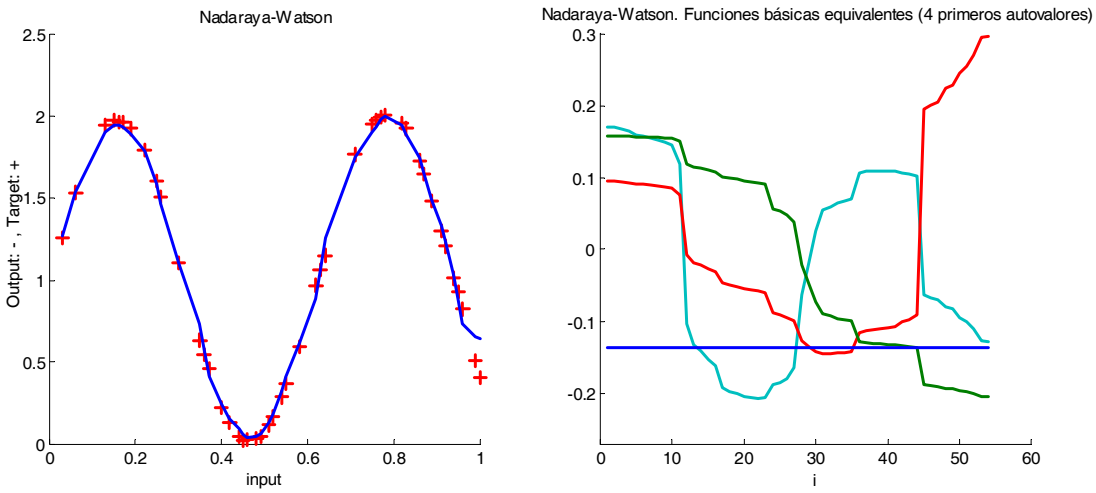


Figura 18. Izquierda: ajueste de un estimador Nadaraya-Watson a una muestra de test procedente de una función sigmoide perturbada con ruido gausiano. Derecha: funciones básicas equivalentes (autovectores) correspondientes a los cuatro mayores autovalores de la matriz *hat* $\mathbf{S} = \mathbf{\Psi}$.

En la práctica, las funciones básicas $\phi_l(\mathbf{x})$ correspondientes a los n primeros autovalores, pueden obtenerse a través de sus valores en los puntos de la muestra $\phi_l(\mathbf{x}_i)$ $i, l = 1 : n$, mediante los autovectores de la matriz *hat* $\mathbf{S} = \mathbf{\Psi}$ [Hastie y Tibshirani 1990, sec. 3.7] con $S_{ij} = (\mathbf{\Psi})_{ij} = \psi(\mathbf{x}_i - \mathbf{x}_j) \equiv \psi(\mathbf{x}_i, \mathbf{x}_j)$:

$$\lambda_l \phi_l(\mathbf{x}_i) = \int \psi(\mathbf{x}_i, \mathbf{x}')\phi_l(\mathbf{x}')d\mathbf{x}' \approx \sum_{j=1}^{n} \psi(\mathbf{x}_i, \mathbf{x}_j)\phi_l(\mathbf{x}_j) = \sum_{j=1}^{n} S_{ij}\phi_l(\mathbf{x}_j) = \mathbf{s}_i \boldsymbol{\phi}_l$$

es decir: $\lambda_l \boldsymbol{\phi}_l = \mathbf{S}\boldsymbol{\phi}_l$, $l = 1 : n$, donde $\mathbf{s}_i$ es la fila i-ésima de $\mathbf{S}$ y $\boldsymbol{\phi}_l = (\phi_l(\mathbf{x}_1), ..., \phi_l(\mathbf{x}_n))^T$ $l = 1 : n$, son los autovectores de la matriz $\mathbf{S}$ en $\mathbf{S} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T = \sum_{l=1}^{h} \lambda_l \boldsymbol{\phi}_l \boldsymbol{\phi}_l^T$.

La figura 18 (izquierda) muestra el comportamiento ante una muestra de test de un estimador Nadaraya-Watson con núcleo gausiano estimado con la misma muestra de los casos anteriores. La figura de la derecha muestra las funciones básicas equivalentes ϕ_l (autovectores) correspondientes a los cuatro mayores autovalores de la matriz *hat* $\mathbf{S} = \mathbf{\Psi}$. La figura 19 muestra el rápido decrecimiento a cero de los autovalores de dicha matriz.

Comparación RBF y Estimador Tipo Núcleo

De la comparación entre la red RBF y el modelo de regresión tipo núcleo, resultan las siguientes similitudes y diferencias más importantes:

1. **Carácter no paramétrico**. El estimador tipo núcleo:

$$\hat{f}_n(\mathbf{x}) = \sum_{i=1}^{n} c_i \frac{k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_i)}{\sum_{l=1}^{n} k_{\mathbf{\Sigma}}(\mathbf{x} - \mathbf{x}_l)} \text{ con } c_i = y_i$$

posee un claro carácter no paramétrico al no realizar hipótesis alguna sobre la función de regresión. Por otra parte, las redes RBF poseen una forma paramétrica con parámetros libres a estimar. Sin embargo, como veremos más adelante, la propiedad aproximadora univer-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

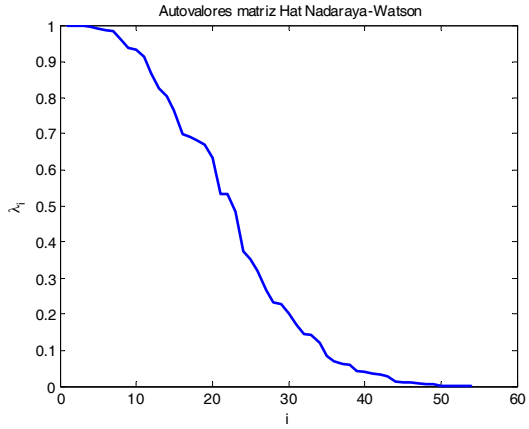


Figura 19. Espectro de autovalores de la matriz *hat* $\mathbf{S} = \mathbf{\Psi}$ de un estimador Nadaraya-Watson en una muestra de $n = 54$ datos. Se observa su valor prácticamente nulo a partir del quincuagésimo.

sal de estas redes (Anexo C) le convierten en un estimador con filosofía auténticamente no paramétrica.

Por otra parte, el estimador tipo núcleo depende más fuertemente de la muestra disponible pues sus *centros* coinciden con los datos de la muestra. Ello también es posible en la red RBF pero, el hecho de que, en ella, los centros se elijan con criterios más estructurales (lo que produce una mayor parsimonia), provoca que la función estimada esté sujeta a variaciones menos radicales que el estimador tipo núcleo. (Este hecho es más contundente si la red se entrena en una sóla fase, en cuyo caso, los centros no tienen porque ser los datos de la muestra).

En este sentido, podría decirse que la red RBF aporta más estructura a la solución, aunque dicha estructura es suficientemente flexible como para aproximarse al proceso generador de los datos.

2. **Problema de dimensionalidad**. Como consecuencia de lo anterior la red RBF resiste mejor los problemas de dimensionalidad que la regresión tipo núcleo que se prescribe para problemas con dos o tres dimensiones como máximo. En espacios de mayores dimensiones, se requiere una muestra de tamaño mucho mayor, ante la cual, la red RBF está mejor preparada para extraer los centros relevantes y producir un modelo más parsimonioso.

3. **Entrenamiento**. En contrapartida, la regresión tipo núcleo es de entrenamiento inmediato mientras la red RBF ha de enfrentarse con la inversión de la matriz $\mathbf{\Psi}^T\mathbf{\Psi}$ de la regresión que, para funciones complejas que requieran un gran número de centros, puede suponer un carga computacional importante.

4. **Carácter Local**. Desde el momento en que la red RBF se entrena mediante una regresión de tipo global, podría parecer que el estimador resultante no posee carácter local como la regresión tipo núcleo o, en general, la regresión polinómico local. Sin embargo, este no es el caso al producir un estimador no paramétrico lineal con sus propios **núcleos efectivos o equivalentes**, no prefijados como en el caso de la regresión tipo núcleo.

   Estos núcleos equivalentes suelen ser distintos en amplitud, implementando una especie de núcleo variable con una mayor amplitud en áreas con menos datos para un mejor suavizado, y distintos en forma, normalmente asimétricos, mostrando su adaptación a los datos.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5. **Comportamiento en los puntos frontera**. Debido a la asimetría y variabilidad de sus núcleos, la red RBF posee un buen comportamiento en los puntos frontera, al contrario que la regresión tipo núcleo. De hecho, sus núcleos se comportan al modo de la regresión polinómico local, deformándose cerca de la frontera hasta el punto de asignar pesos negativos a las observaciones más lejanas a la misma, lo que contribuye a reducir el sesgo.

6. **Dualidad**. La red RBF y la regresión tipo núcleo poseen carácter dual con respecto a la regresión: mientras la red RBF es un modelo concebido en el plano primal como expansión en funciones básicas centradas en los valores $\mathbf{x}_i$ de las variables regresoras, la regresión tipo núcleo reside en el plano dual definiéndose en términos de los valores y_i de la variable respuesta.

La Red RBF como Solución de un Problema de Regularización

En los apartados I.1.2 y I.4.1 del capítulo anterior, presentamos el enfoque de regularización como medio de asegurar que el problema de aprendizaje era un problema bien planteado, e introdujimos el principio ERM regularizado ((I.1)) como principio inductivo:

$$\hat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)) + \lambda \mathcal{E}(f) \tag{II.45}$$

donde ℓ es la función de pérdida, λ es un parámetro regularizador y $\mathcal{E} : \mathcal{F} \rightarrow \mathbb{R}$ es un estabilizador que penaliza la complejidad de las funciones del espacio de hipótesis $\mathcal{F}$.

Si se considera la pérdida cuadrática, se verifica que, todo suavizador lineal $\hat{\mathbf{f}}_n = \mathbf{S}\mathbf{y}$, es solución del siguiente problema de regularización, ([Buja *et al.* 1989], [Hastie y Tibshirani 1990, sec. 3.11.3]):

$$\min_f \hat{\mathcal{R}}_n(f) = \min_f \{\|\mathbf{y} - \mathbf{f}\|_2^2 + \mathbf{f}^T(\mathbf{S}^- - \mathbf{I}_n)\mathbf{f}\}$$

donde $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^T$ y $\mathbf{S}^-$ es la inversa generalizada de $\mathbf{S}$ (Anexo G). Ello es debido a que, la condición de óptimo produce:

$$\frac{\partial \hat{\mathcal{R}}_n(\mathbf{f})}{\partial \mathbf{f}} = -\mathbf{y} + \mathbf{f} + (\mathbf{S}^- - \mathbf{I}_n)\mathbf{f} = 0 \Leftrightarrow \mathbf{S}^-\mathbf{f} = \mathbf{y} \Leftrightarrow \mathbf{f} = \mathbf{S}\mathbf{y}$$

Así pues, por su carácter de suavizador lineal, la red RBF es solución del principio ERM regularizado con estabilizador $\mathcal{E}_n(f) = \mathbf{f}^T(\mathbf{S}^- - \mathbf{I}_n)\mathbf{f}$ y parámetro regularizador $\lambda = 1$.

La utilidad que podría tener este resultado es que alguna información a priori permitiese definir el estabilizador $\mathcal{E}_n$ pues, ello, determinaría unívocamente la matriz *hat* $\mathbf{S}$. Sin embargo, es difícil trasladar en la práctica, a una matriz $\mathbf{S}$, la información a priori que pueda tenerse sobre la suavidad de la verdadera función. De cualquier forma, como veremos en el siguiente apartado, existen estrategias más interpretables para incorporar dicha información a priori en un estabilizador adecuado.

USAL UNIVERSIDAD DE SALAMANCA DE CONOCIMIENTO

II.5.2 Redes RBF de Regularización (RRBF)

Problema de Regularización según Tikhonov

El enfoque más común para convertir el problema de aprendizaje en un problema bien planteado es el enfoque de [Tikhonov y Arsenin 1977], que define estabilizadores de la forma:

$$\mathcal{E}(f) = \|Pf\|^2 \tag{II.46}$$

donde P es un operador diferencial lineal[56] definido en un espacio de Hilbert de funciones $\mathcal{H}_P$ (usualmente, $\mathcal{H}_P \subset L_2(\mathcal{X})$) compuesto por todas las funciones compatibles con P y sus condiciones de contorno, y cuya imagen es un espacio de funciones en el que el producto interno $\langle Pf, Pg\rangle$ de L_2 está bien definido. En lo que sigue, consideramos que el espacio de hipótesis $\mathcal{F}$ es un subconjunto de $L_2(\mathcal{X})$ con $\mathcal{X} \subset \mathbb{R}^d$.

Con el estabilizador (II.46), el concepto de suavidad de una función se define ahora en términos del *tamaño* de las derivadas de algún orden, de las funciones del dominio. Con ello, el principio ERM regularizado se formula como:

$$\min_{f\in\mathcal{H}_P} \hat{\mathcal{R}}_n(f) = \min_{f\in\mathcal{H}_P} \{\hat{R}_n(f) + \lambda\mathcal{E}(f)\} \equiv \min_{f\in\mathcal{H}_P} \{\|\mathbf{y} - \mathbf{f}\|_2^2 + \lambda \|Pf\|_{L_2}^2\} \tag{II.47}$$

Lo anterior es un problema variacional cuya solución puede obtenerse por métodos clásicos a través del siguiente teorema cuya demostración se presenta en el Anexo G:

Teorema 37 *[Poggio y Girosi 1990] La solución al problema (II.47) posee la forma general:*

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i) + g(\mathbf{x}) \tag{II.48}$$

donde k es la función de Green (Anexo B) del operador diferencial P^+P donde P^+ es el operador adjunto de P y g es una función perteneciente al espacio nulo del mismo.

Si la función de Green k es definida positiva (Anexo B), el correspondiente operador integral es positivo, con lo que el espacio nulo de P^+P es el espacio trivial y la solución del problema (II.47) es de la forma:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i) \tag{II.49}$$

Si P es invariante por traslaciones, (una traslación del espacio de entrada no altera la penalización $\mathcal{E}(f)$), puede demostrarse [Poggio y Girosi 1990] que k es de la forma $k(\mathbf{x}, \mathbf{x}') \equiv \psi(\mathbf{x} - \mathbf{x}')$ para alguna función ψ y si, además, es invariante ante rotaciones (todas las direcciones del dominio de definición de las funciones son igualmente penalizadas), k es de carácter radial: $k(\mathbf{x}, \mathbf{x}') \equiv \zeta(\|\mathbf{x} - \mathbf{x}'\|)$.

Las redes neuronales de la forma (II.48) y su caso particular (II.49) han sido propuestas por [Poggio y Girosi 1990] y [Girosi *et al.* 1995] bajo el nombre de **redes de regularización** (*regularization networks*), que aquí denominaremos redes RBF de regularización (RRBF). Obsérvese que el número de neuronas ocultas en dicha arquitectura coincide con el número de datos de la muestra.

[56] En general, es válido cualquier operador lineal, no necesariamente diferencial, para el que las funciones suaves resulten *transparentes* según algún criterio.

Solución al Problema de Regularización en un RKHS

Dado un estabilizador definido en base a información a priori sobre la verdadera función, la determinación de la función de Green del operador diferencial P^+P puede resultar complicada en la práctica.

Otro enfoque para determinar la forma general de la solución del principio ERM regularizado (II.47), es el basado en **espacios de Hilbert con núcleo reproductor**, (RKHS), (Anexo B), considerando como dominio del problema un RKHS: $\mathcal{F} = \mathcal{H}_k$ con núcleo k, lo que equivale a decir que k es una función definida positiva (Anexo B). Para una mayor generalidad, escribimos ahora el problema (II.47), de la forma:

$$\min_{f\in\mathcal{H}_k} \hat{\mathcal{R}}_n(f) = \min_{f\in\mathcal{H}_k} \{\hat{R}_n(f) + \varrho(\|f\|^2_{\mathcal{H}_k})\} \tag{II.50}$$

donde ahora $\lambda\mathcal{E}(f) = \varrho(\|f\|^2_{\mathcal{H}_k})$ siendo $\|\cdot\|_{\mathcal{H}_k}$ la norma en el RKHS y $\varrho : \mathbb{R} \to \mathbb{R}$ alguna función.

En un RKHS, la forma de la solución viene dada por el siguiente teorema que generaliza el originalmente demostrado[57] por [Kimeldorf y Wahba 1971, lema 3.1] (también [Wahba 1990*b*, th. 1.3.1]), para pérdida cuadrática y $\varrho(\|f\|^2_{\mathcal{H}_k}) = \lambda\|f\|^2_{\mathcal{H}_k}$, que demostramos por su relevancia y brevedad:

Teorema 38 *[Schölkopf, Herbrich, Smola y Williamson 2000], [Schölkopff y Smola 2002] Sea $\varrho : [0,\infty) \to \mathbb{R}$ una función estrictamente monótona, $\mathcal{X}$ un conjunto y $\ell : \mathcal{X}\times\mathcal{Y} \to [0,\infty)$ una función de pérdida que verifica $\min_{t\in\mathcal{Y}} \ell(y,t) = \ell(y,y)$. Entonces, cualquier óptimo $f \in \mathcal{H}_k$ del problema (II.50) admite una representación de la forma:*

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x},\mathbf{x}_i) \tag{II.51}$$

Demostración. Al ser el conjunto de sumas finitas, $\sum_{i=1}^n c_i k_{\mathbf{x}_i}$ con $k_{\mathbf{x}_i} = k(\cdot,\mathbf{x}_i)$ subespacio lineal cerrado en $\mathcal{H}_k$, entonces $\forall f \in \mathcal{H}_k$:

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} c_i k(\mathbf{x},\mathbf{x}_i) = \sum_{i=1}^{n} c_i k(\mathbf{x},\mathbf{x}_i) + f_\perp(\mathbf{x})$$

donde $f_\perp$ es ortogonal a $k(\cdot,\mathbf{x}_i)$, es decir, $f_\perp \perp k(\cdot,\mathbf{x}_i) \equiv \psi_{\mathbf{x}_i}, \forall i \in \{1:n\}$. Con ello,

$$\|f\|^2_{\mathcal{H}_k} = \left\|\sum_{i=1}^{n} c_i\psi(\mathbf{x},\mathbf{x}_i)\right\|^2_{\mathcal{H}_k} + \|f_\perp\|^2_{\mathcal{H}_k}$$

Además, debido a la propiedad reproductora de k (Anexo B), $\forall f \in \mathcal{H}_k$ y $\forall i \in \{1:n\}$:

$$f(\mathbf{x}_i) = \langle f, k_{\mathbf{x}_i}\rangle = \left\langle \sum_{j=1}^{n} c_j k_{\mathbf{x}_j} + f_\perp, k_{\mathbf{x}_i}\right\rangle = \sum_{j=1}^{n} c_j \langle k_{\mathbf{x}_j}, k_{\mathbf{x}_i}\rangle + \langle f_\perp, k_{\mathbf{x}_i}\rangle = \sum_{j=1}^{n} c_j k(\mathbf{x}_i,\mathbf{x}_j)$$

con lo que el mínimo de $\hat{R}_n$ siempre se alcanza para una expresión del tipo (II.51). Por otra parte,

$$\varrho(\|f\|^2_{\mathcal{H}_k}) = \varrho\left(\left\|\sum_{i=1}^{n} c_i k_{\mathbf{x}_i}\right\|^2_{\mathcal{H}_k} + \|f_\perp\|^2_{\mathcal{H}_k}\right) \geq \varrho\left(\left\|\sum_{i=1}^{n} c_i k_{\mathbf{x}_i}\right\|^2_{\mathcal{H}_k}\right)$$

[57] Para el caso más general de espacios de Hilbert con seminúcleo reproductor, que veremos a continuación.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

luego cualquier óptimo de $\hat{\mathcal{R}}_n(f)$ en $\mathcal{H}_k$ se alcanzará para $f = \sum_{i=1}^n c_i k_{\mathbf{x}_i}$ para algún conjunto de valores c_i, $i = 1, .., n$.

■

Generalizaciones

La resolución del principio ERM regularizado (II.45) bajo el marco de un RKHS posee como ventaja respecto al enfoque variacional que, la relación biunívoca existente entre los RKHS y sus núcleos respectivos, (las funciones definidas positivas, Anexo B), facilita la labor de modelización del dominio de soluciones factibles.

Pero, además, el enfoque RKHS permite una solución más sencilla de problemas de aprendizaje más generales que el de regresión, como por ejemplo:

$$\min_{f \in \mathcal{H}_k} \{\sum_{i=1}^n (y_i - L_i(f))^2 + \lambda \|f\|^2_{\mathcal{H}_k}\} \tag{II.52}$$

donde L_i, $i = 1 : n$ son operadores lineales acotados en $\mathcal{H}_k$, (el problema de regresión es un caso particular con $L_i(f) = f(\mathbf{x}_i)$). En este caso, se satisface una versión generalizada del teorema anterior, pues, al igual que en la demostración anterior, ahora:

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i \eta_i(\mathbf{x}) + f_\perp(\mathbf{x})$$

con lo que la norma se minimiza para el primer término de la derecha. Igualmente:

$$L_i(f) = \langle \eta_i, f \rangle_{\mathcal{H}_k} = \left\langle \eta_i, \sum_{j=1}^n \alpha_j \eta_j + f_\perp \right\rangle_{\mathcal{H}_k} = \left\langle \eta_i, \sum_{j=1}^n \alpha_j \eta_j \right\rangle_{\mathcal{H}_k} = L_i(\sum_{j=1}^n \alpha_j \eta_j) \tag{II.53}$$

con lo que $L_i(f)$ es el resultado de aplicar L_i a una expansión en η_j con n términos. De esta forma, se minimiza (II.52) con una expansión del tipo [Wahba 1990*b*, th. 1.3.1]:

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i \eta_i(\mathbf{x}) \tag{II.54}$$

Pero, una de las propiedades útiles de los núcleos reproductores, es que, a partir de ellos, puede obtenerse el representante en el sentido de Riestz, (Anexo B, teorema 89), de cualquier operador acotado: si η_i es el representante de L_i, es decir $L_i f = \langle \eta_i, f \rangle_{\mathcal{H}_k}$, entonces:

$$\eta_i(\mathbf{x}) = \langle \eta_i, \psi_{\mathbf{x}} \rangle_{\mathcal{H}_k} = L_i(\psi_{\mathbf{x}}(\cdot)) = L_i(\psi(\cdot, \mathbf{x})) \tag{II.55}$$

donde la primera igualdad se deduce de la propiedad de núcleo de reproductor de ψ.

Así pues, $\eta_i(\mathbf{x})$ se obtiene aplicando L_i a $k_{\mathbf{x}}(\cdot)$ lo que permite abordar problemas donde los datos y_i poseen significados más complejos que el mero valor $\mathtt{f}(\mathbf{x}_i)$ de la función $\mathtt{f}$ a estimar. Por ejemplo, si $\mathcal{X} \subset \mathbb{R}$, si $y_i = \frac{d}{dx}\mathtt{f}(x_i) = \dot{f}(x_i)$, entonces $L_i(f) = \dot{f}(x_i)$ en (II.52):

$$\eta_i(\bar{x}) = \langle \eta_i, k_{\bar{x}} \rangle_{\mathcal{H}_k} = L_i(k_{\bar{x}}) = \dot{k}_{\bar{x}}(x_i), \ \forall \bar{x} \in \mathcal{X}$$

con lo que, utilizando lo anterior, si $f \in \mathcal{H}_\psi$, entonces la solución general es de la forma:

$$\hat{y}_i = L_i(f) = \sum_{i=1}^n \alpha_i \frac{\partial^2}{\partial x_i \partial x_j} k(x_i, x_j)$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ya que, (II.53) también puede escribirse como:

$$L_i(f) = \sum_{j=1}^{n} \alpha_j \left\langle \eta_i, \eta_j \right\rangle_{\mathcal{H}_k} = \sum_{j=1}^{n} \alpha_j L_i(\eta_j) = \sum_{j=1}^{n} \alpha_j L_i(L_j(k(\cdot,\cdot)))$$

Otro ejemplo resulta cuando la variable respuesta es el resultado de la aplicación de un operador integral a la función de interés, en cuyo caso puede definirse $L_i(f) = \int_{\mathcal{X}} \xi_i(\mathbf{x}')f(\mathbf{x}')d\mathbf{x}'$. Entonces, la solución posee también la forma (II.54) donde, debido a (II.55) se tiene: $\eta_i(\mathbf{x}) = L_i(\psi(\cdot,\mathbf{x})) = \int \xi_i(\mathbf{x}')\psi(\mathbf{x},\mathbf{x}')d\mathbf{x}'$, con lo que puede seguirse un enfoque análogo.

Solución al Problema de Regularización en un SRKHS

El teorema 38 puede generalizarse al caso en que k es una función condicionalmente definida positiva de orden m (Anexo B) que determina un espacio de Hilbert con seminúcleo reproductor k (*semireproducing kernel Hilbert space*, –SRKHS, Anexo B, [Madych y Nelson 1988, th. 1.1], [Dyn 1987]), sobre el que se define una seminorma $\|\cdot\|_{\mathcal{H}_k}$ y cuyo espacio nulo no trivial está constituido por el espacio $\Pi_m(\mathcal{X})$ de polinomios de grado menor que m.

Dicha generalización viene dada por el siguiente teorema, cuya demostración aportamos por breve e ilustrativa, y que generaliza también el resultado ya mencionado de [Kimeldorf y Wahba 1971, lema 3.1].

Teorema 39 *Sea $\mathcal{H}_k$ un espacio de Hilbert con seminúcleo reproductor k función condicionalmente definida positiva de orden m con $m \leq n$. Entonces, cualquier óptimo en $\mathcal{H}_k$ del principio ERM regularizado $\hat{\mathcal{R}}_n(f) = \hat{R}_n(f) + \varrho(\|f\|^2_{\mathcal{H}_k})$ admite una representación de la forma:*

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x},\mathbf{x}_i) + \sum_{j=1}^{r} b_j g_j(\mathbf{x}) \tag{II.56}$$

con $\{g_j\}_{j=1}^{r} \subset \Pi_m(\mathcal{X})$ es una base[58] *de $\Pi_m(\mathcal{X})$ y $c_i, b_j \in \mathbb{R}$, $i = 1:n$, $j = 1:r$.*

Demostración. El conjunto de sumas finitas $\sum_{i=1}^{n} c_i k_{\mathbf{x}_i}$ con $k_{\mathbf{x}_i} = k(\cdot,\mathbf{x}_i)$ es un subespacio lineal cerrado $\mathbb{H}$ en el espacio de Hilbert $\mathcal{H}_k$ con $\mathcal{H}_k \oplus \Pi_m(\mathcal{X})$ suma directa ortogonal, (teorema 127, Anexo B). Por tanto, toda función $f \in \mathcal{H}_k$ posee la forma:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x},\mathbf{x}_i) + f_\perp(\mathbf{x}) + \sum_{j=1}^{r} b_j g_j(\mathbf{x})$$

donde $\{g_j\}_{j=1}^{r} \subset \Pi_m(\mathcal{X})$ es base de $\Pi_m(\mathcal{X})$ y $f_\perp$ es la componente ortogonal a $\sum_{i=1}^{n} c_i k(\mathbf{x},\mathbf{x}_i)$ en $\mathbb{H}$. Con ello, el mínimo de $\hat{R}_n$ en $\mathcal{H}_k$ se alcanza necesariamente en funciones así. Pero al ser $\Pi_m(\mathcal{X})$ el espacio nulo de $\|\cdot\|_{\mathcal{H}_k}$ entonces

$$\varrho(\|f\|^2_{\mathcal{H}_k}) = \varrho(\left\|\sum_{i=1}^{n} c_i k_{\mathbf{x}_i}\right\|^2_{\mathcal{H}_k} + \|f_\perp\|^2_{\mathcal{H}_k}) \geq \varrho(\left\|\sum_{i=1}^{n} c_i k_{\mathbf{x}_i}\right\|^2_{\mathcal{H}_k})$$

luego cualquier óptimo de $\hat{\mathcal{R}}_n(f)$ en $\mathcal{H}_k$ será de la forma $\sum_{i=1}^{n} c_i k(\mathbf{x},\mathbf{x}_i)$ salvo términos de la forma $\sum_{j=1}^{r} b_j g_j(\mathbf{x})$ que son transparentes para $\|\cdot\|^2_{\mathcal{H}_k}$, es decir, poseerá la forma (II.56).

■

Como resultado, se obtiene una red RBF regularizada *semiparamétrica*, en el sentido de que debe añadirse a la expresión general (II.51) un término adicional. Como vamos a ver a conti-

[58] Nótese que $r = \dim(\Pi_m(\mathbb{R}^d)) = \binom{d+m-1}{d}$

nuación, esto se corresponde con el caso en que el operador P^+P del enfoque Tikhonov, posea un espacio nulo no trivial.

Relación entre el Enfoque de Tikhonov y el Enfoque RKHS

A pesar de las ventajas del enfoque RKHS para determinar la forma de la solución al problema de regularización, no está claro en qué términos la norma del RKHS $\mathcal{E}(f) = \|f\|^2_{\mathcal{H}_k} = \langle f, f\rangle_{\mathcal{H}_k}$ restringe la suavidad de la solución como lo hacía el operador diferencial lineal P de $\mathcal{E}(f) = \|Pf\|^2_2 = \langle Pf, Pf\rangle_2$ en el enfoque de Tikhonov.

El siguiente teorema aclara este asunto relacionando los productos internos $\langle\cdot,\cdot\rangle_{\mathcal{H}_k}$ y $\langle\cdot,\cdot\rangle_2$ (el caso de SRKHS es análogo). En él, podemos suponer sin pérdida de generalidad [Schölkopff y Smola 2002, p. 93] estabilizadores lineales de la forma $\mathcal{E}(f) = \left\|\bar{P}f\right\|^2_2$ con $\bar{P}$ operador lineal positivo pues lo que importa es la positividad del operador P^+P y esto se consigue por ejemplo con[59] $\bar{P} = (P^+P)^{1/2}$ ya que entonces $\langle f, P^+Pf\rangle_2 = \left\langle \bar{P}f, \bar{P}f\right\rangle_2 \geq 0$ y $\bar{P}$ posee las mismas propiedades regularizadoras que P :

$$\left\|\bar{P}f\right\|^2 = \left\langle \bar{P}f, \bar{P}f\right\rangle = \left\langle f, P^+Pf\right\rangle = \langle Pf, Pf\rangle = \|Pf\|^2$$

Teorema 40 *[Smola* et al. *1998], [Schölkopff y Smola 2002] Para todo estabilizador positivo $P : \mathcal{F} \to \mathcal{G}$ con $\mathcal{F}, \mathcal{G}$ espacios de funciones con producto interno, existe un RKHS $\mathcal{H}_k$ con núcleo k satisfaciendo:*

$$f(\mathbf{x}) = \langle Pk_{\mathbf{x}}, Pf\rangle_{\mathcal{G}}\,, \forall f \in \mathcal{H}_k \quad \text{(II.57)}$$

$$k(\mathbf{x}, \mathbf{x}') = \langle Pk_{\mathbf{x}}, Pk_{\mathbf{x}'}\rangle_{\mathcal{G}} \quad \text{(II.58)}$$

con producto interno $\langle f_1, f_2\rangle_{\mathcal{H}_k} = \langle Pf_1, Pf_2\rangle_{\mathcal{G}}$. Y la inversa, para todo $\mathcal{H}_k$ RKHS con núcleo k, existe un operador lineal $P : \mathcal{H}_k \to \mathcal{G}$ tal que en $\mathcal{G}$ está definido un producto interno y $\forall f \in \mathcal{H}_k$ se verifican las condiciones anteriores.

Demostración. La segunda afirmación se cumple tomando como P el operador identidad. Para la primera, por definición de función de Green $k_{\mathbf{x}}$ del operador P^+P, (solución de la ecuación $P^+Pk_{\mathbf{x}} = \delta_{\mathbf{x}}$ en $\mathcal{G}$), (Anexo B), $\forall g \in P^+P\mathcal{F} \subset \mathcal{F}$ tal que $P^+Pf = g$ para algún $f \in \mathcal{F}$ entonces:

$$f(\mathbf{x}) = \langle k_{\mathbf{x}}, g\rangle_{\mathcal{G}} = \left\langle k_{\mathbf{x}}, P^+Pf\right\rangle_{\mathcal{G}} = \langle Pk_{\mathbf{x}}, Pf\rangle_{\mathcal{G}} = \langle k_{\mathbf{x}}, g\rangle_{\mathcal{H}_k}$$

con lo que $k_{\mathbf{x}}$ verifica la propiedad reproductora en P^+Pf con $\langle\cdot,\cdot\rangle_{\mathcal{H}_k}$. Lo anterior se verifica también para $f = k_{\mathbf{x}}$ (pues $k_{\mathbf{x}}$ es solución de la ecuación $P^+Pk_{\mathbf{x}} = \delta_{\mathbf{x}}$ en $\mathcal{G}$):

$$k_{\mathbf{x}}(\mathbf{x}') = \left\langle k_{\mathbf{x}'}, P^+Pk_{\mathbf{x}}\right\rangle_{\mathcal{G}} = \langle Pk_{\mathbf{x}'}, Pk_{\mathbf{x}}\rangle_{\mathcal{G}} = \langle Pk_{\mathbf{x}}, Pk_{\mathbf{x}'}\rangle_{\mathcal{G}} = k_{\mathbf{x}'}(\mathbf{x})$$

(con lo que, además, $k_{\mathbf{x}}$ es simétrica). Con todo ello, $k_{\mathbf{x}} \in P^+P\mathcal{F}\ \forall \mathbf{x} \in \mathcal{X}$. Por tanto, el RKHS buscado es la clausura de $\{f \in P^+P\mathcal{F} : \|Pf\|^2 < \infty\}$ que posee núcleo k pues $k_{\mathbf{x}}$ está en dicho espacio $\forall \mathbf{x} \in \mathcal{X}$ y en él verifica la propiedad reproductora.

■

En el enfoque de Tikhonov, habíamos supuesto $\mathcal{G} = L_2(\mathcal{X})$. De esta forma, el estabilizador $\|\cdot\|_{\mathcal{H}_k}$ en el marco de un RKHS $\mathcal{H}_k$ puede interpretarse en términos del estabilizador correspondiente a algún operador lineal P positivo (por lo menos, el operador identidad):

$$\mathcal{E}(f) = \|f\|^2_{\mathcal{H}_k} = \langle f, f\rangle_{\mathcal{H}_k} = \langle \textstyle\sum_i \alpha_i k_{\mathbf{x}_i}, \sum_i \alpha_i k_{\mathbf{x}_i}\rangle_{\mathcal{H}_k} = \sum_{i,j} \alpha_i\alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$= \textstyle\sum_{i,j} \alpha_i\alpha_j \langle Pk_{\mathbf{x}}, Pk_{\mathbf{x}'}\rangle_{\mathcal{G}} = \langle P\sum_i \alpha_i k_{\mathbf{x}_i}, P\sum_i \alpha_i k_{\mathbf{x}_i}\rangle_{\mathcal{G}} = \langle Pf, Pf\rangle_{\mathcal{G}} = \|Pf\|^2_{\mathcal{G}}$$

[59] Dado un operador Q, se denota por $Q^{1/2}$ un operador que verifique $Q^{1/2}Q^{1/2} = Q$.

Y a la inversa, dado un estabilizador $\mathcal{E}$ con operador P, el dominio del problema es un RKHS $\mathcal{F} = \mathcal{H}_k$ siendo su núcleo k la función de Green de P^+P.

El siguiente teorema permite construir este RKHS a partir de la descomposición espectral del correspondiente operador P.

Teorema 41 *[Schölkopff y Smola 2002, prop. 4.10] Sea un operador P lineal positivo de un estabilizador $\mathcal{E}$, tal que P^+P posee un conjunto finito numerable de autovectores $\{\phi_j\}$ con correspondientes autovalores $\{\mu_j\}$. Si se define:*

$$k(\mathbf{x}, \mathbf{x}') = \sum_{\mu_j \neq 0} \frac{\alpha_j}{\mu_j} \phi_j(\mathbf{x}) \phi_j(\mathbf{x}')$$

con $\alpha_j \in \{0, 1\}$ para todo j y $\sum_j \frac{\alpha_j}{\mu_j}$ convergente, entonces el correspondiente RKHS viene dado por $\mathcal{H}_k = \left\langle \{\phi_j : \alpha_j = 1, j \in \mathbb{N}\} \right\rangle$.

Demostración. Debido a la ortonormalidad de $\{\phi_j\}$:

$$\begin{aligned} \left\langle k_{\mathbf{x}_i}, P^+ P k_{\mathbf{x}_j} \right\rangle &= \left\langle \sum_{\mu_i \neq 0} \frac{\alpha_i}{\mu_i} \phi_i(\mathbf{x}_i) \phi_i(\cdot), P^+P \sum_{\mu_j \neq 0} \frac{\alpha_j}{\mu_j} \phi_j(\cdot) \phi_j(\mathbf{x}_j) \right\rangle \\ &= \sum_{i,j} \frac{\alpha_i}{\mu_i} \frac{\alpha_j}{\mu_j} \phi_i(\mathbf{x}_i) \phi_j(\mathbf{x}_j) \left\langle \phi_i(\cdot), P^+ P \phi_j(\cdot) \right\rangle = \sum_i \frac{\alpha_i}{\mu_i} \phi_i(\mathbf{x}_i) \phi_i(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

■

El teorema anterior permite construir una gran variedad de núcleos asociados a un operador P, y a la inversa, mediante la selección de un determinado subconjunto de autofunciones. La exclusión de las autofunciones más complejas supone reducir la complejidad del espacio de hipótesis $\mathcal{F} = \mathcal{H}_k$ elegido para el problema, mientras la exclusión de las más suaves provocará un efecto inverso.

Medida de Suavidad de una Función

Un caso particular de la relación anterior, interpretable en términos de suavidad de una función, viene dado por el siguiente operador P propuesto por [Girosi *et al.* 1995]:

$$\mathcal{E}_{\check{\psi}}(f) = P(f) = \int_{\mathbb{R}^d} \frac{\left| \tilde{f}(\boldsymbol{\omega}) \right|^2}{\check{\psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} \tag{II.59}$$

donde $\tilde{f}$ es la transformada de Fourier de f, $\check{\psi}$ es una función positiva tal que $\lim_{\|\boldsymbol{\omega}\| \to \infty} \check{\psi}(\boldsymbol{\omega}) = 0$, dando lugar a una familia de funciones $\mathcal{F}_{\check{\psi}}$ para la que dicha expresión esté bien definida. La conexión entre este estabilizador y la suavidad de una función $f \in \mathcal{F}_{\check{\psi}}$ se debe a lo siguiente:

1. Por la identidad de Parseval:

$$\int_{\mathbb{R}^d} \left| \tilde{f}(\boldsymbol{\omega}) \right|^2 d\boldsymbol{\omega} = \int_{\mathbb{R}^d} |f(\mathbf{x})|^2 \, d\mathbf{x}$$

que se define como la cantidad de energía de la función f, que da una medida de su comportamiento oscilatorio. Por tanto, refleja el grado de suavidad de la función aunque en el dominio de las frecuencias: una función es más *suave* que otra si *oscila* menos que ella, es decir, si posee menos energía a altas frecuencias.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$\mathcal{E}_P(f) = \|Pf\|_2^2$	$\mathcal{E}_{\breve{\psi}}(f) = \int_{\mathbb{R}^d} \frac{\left\|\tilde{f}(\boldsymbol{\omega})\right\|^2}{\breve{\psi}(\boldsymbol{\omega})} d\boldsymbol{\omega}$	$\breve{\psi}(\boldsymbol{\omega})$
$\int_{\mathbb{R}} \dot{f}(x)^2 dx$	$\int_{\mathbb{R}} \omega^2 \left\|\tilde{f}(\omega)\right\|^2 d\omega$	$\frac{1}{\omega^2}$
$\int_{\mathbb{R}} \ddot{f}(x)^2 dx$	$\int_{\mathbb{R}} \omega^4 \left\|\tilde{f}(\omega)\right\|^2 d\omega$	$\frac{1}{\omega^4}$
$\int_{\mathbb{R}} f(x)^2 e^{\frac{1}{2\sigma^2}x^2} dx$	$\int_{\mathbb{R}} \left\|\tilde{f}(\omega)\right\|^2 e^{\frac{\sigma^2}{2}\omega^2} d\omega$	$e^{-\sigma^2\omega/2}$

Tabla 5. Ejemplos de equivalencias entre estabilizadores diferenciales y estabilizadores en el dominio de la frecuencia

2. El contenido de energía a altas frecuencias puede determinarse midiendo su cantidad de energía después de haber filtrado previamente sus frecuencias más bajas: el cociente $1/\breve{\psi}(\boldsymbol{\omega})$ representa este filtrado previo (*high pass filter*).

De la relación entre esta formulación del estabilizador y su formulación en términos de un operador positivo P, dan idea los ejemplos en $\mathbb{R}$ incluidos en la tabla 5 donde se ha eliminado el factor 2π que no afecta a la función objetivo:

Con ello, el principio ERM regularizado posee la forma general:

$$\min_{f \in \mathcal{F}_{\breve{\psi}}} \hat{\mathcal{R}}_n(f) = \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)) + \lambda \int_{\mathbb{R}^d} \frac{\left|\tilde{f}(\boldsymbol{\omega})\right|^2}{\breve{\psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} \tag{II.60}$$

Redes RBF de Regularización

La conexión entre el estabilizador (II.59) y los resultados presentados anteriormente en el marco de un RKHS la proporciona el siguiente teorema, caso particular de uno más general incluido en el Anexo B (teorema 129):

Teorema 42 *Sea $\breve{\psi}$ una función positiva, simétrica e integrable en $\mathbb{R}^d$ verificando $\lim_{\|\boldsymbol{\omega}\|\to\infty} \breve{\psi}(\boldsymbol{\omega}) = 0$. Sea $k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x} - \mathbf{x}') = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \breve{\psi}(\boldsymbol{\omega}) e^{-i(\mathbf{x}-\mathbf{x}')^T\boldsymbol{\omega}} d\boldsymbol{\omega}$ la transformada de Fourier de $\breve{\psi}$, $(\tilde{\breve{\psi}} = \psi)$. Entonces, la familia de funciones:*

$$\mathcal{H}_k \equiv \mathcal{H}_\psi = \{f \in L_2(\mathcal{X}) : \int_{\mathbb{R}^d} \frac{\left|\tilde{f}(\boldsymbol{\omega})\right|^2}{\breve{\psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} < \infty\}$$

es un espacio de Hilbert con núcleo reproductor $k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x} - \mathbf{x}')$, con producto interno:

$$\langle f, g \rangle = (2\pi)^{d/2} \int_{\mathbb{R}^d} \frac{\tilde{f}(\boldsymbol{\omega})\overline{\tilde{g}}(\boldsymbol{\omega})}{\breve{\psi}(\boldsymbol{\omega})} d\boldsymbol{\omega}$$

Demostración. Basta demostrar que $k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x} - \mathbf{x}')$ es la función de Green del operador P^+P asociado a $P(f) = \tilde{f}(\boldsymbol{\omega})$ definido en $L_2(1/\breve{\psi})$, pues por el teorema 40, $\mathcal{H}_k$ es RKHS. Pero ello se cumple pues siendo $k_{\mathbf{x}} \equiv \psi_{\mathbf{x}} = \psi(\cdot - \mathbf{x})$:

$$\langle k_{\mathbf{x}}, f \rangle_{\mathcal{H}_k} = \langle \psi_{\mathbf{x}}, f \rangle_{\mathcal{H}_k} = \int_{\mathbb{R}^d} \frac{\tilde{f}(\boldsymbol{\omega})\overline{\tilde{\psi}_{\mathbf{x}}}(\boldsymbol{\omega})}{\breve{\psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} = \int_{\mathbb{R}^d} \frac{\overline{\tilde{\psi}}(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}^T\mathbf{x}} \tilde{f}(\boldsymbol{\omega})}{\breve{\psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T\mathbf{x}} \tilde{f}(\boldsymbol{\omega}) d\boldsymbol{\omega} = f(\mathbf{x})$$

■

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Por el teorema de Bochner (Anexo B), las hipótesis del teorema son condición necesaria y suficiente para que la función ψ sea definida positiva (en el sentido de Bochner, teorema 130 Anexo B), por lo que podíamos haber concluido antes que k es definida positiva y así es núcleo reproductor de un RKHS. Sin embargo, el teorema anterior determina concretamente dicho espacio y su norma.

Aplicando el teorema 38 al RKHS $\mathcal{H}_k = \mathcal{F}_{\check{\psi}}$, se tiene que la solución del problema (II.60) posee la forma general:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i) = \sum_{i=1}^{n} c_i \psi(\mathbf{x} - \mathbf{x}_i) \tag{II.61}$$

que es la expresión de una red RBF de regularización (RBFR) sin término constante.

El teorema anterior puede generalizarse a funciones condicionalmente definidas positivas k de orden m, mediante el teorema 129 Anexo B ya mencionado, dando lugar a un espacio de Hilbert $\mathcal{H}_k$ con núcleo semirreproductor k verificando $\mathcal{H}_k = \mathbb{H} \oplus \Pi_m(\mathcal{X})$ suma directa ortogonal donde $\mathbb{H}$ es un espacio de Hilbert.

Con ello, las funciones de $\mathcal{H}_k$ son de la forma $f = f_1 + g$ con $g \in \Pi_m(\mathcal{X})$ y $f_1 \in \mathbb{H}$. Además, $\mathcal{H}_k$ está dotado de una seminorma $\|\cdot\|_{\mathcal{H}_k}$ cuyo espacio nulo no trivial es $\Pi_m(\mathcal{X})$.

Por tanto, aplicando el teorema 39 al SRKHS $\mathcal{H}_k$ se tiene que la solución general del problema (II.60) es de la forma:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{j=1}^{r} b_j g_j(\mathbf{x}) = \sum_{i=1}^{n} c_i \psi(\mathbf{x} - \mathbf{x}_i) + \sum_{j=1}^{r} b_j g_j(\mathbf{x}) \tag{II.62}$$

donde $\{g_j\}_{j=1}^{r}$ es una base de $\Pi_m(\mathcal{X})$ espacio de polinomios de grado menor que m. Dicha expresión es una red RBF de regularización (RRBF) de carácter *semiparamétrico*, (con término polinómico adicional).

Redes RBF de Regularización con Parámetro Constante

Si la función de transferencia $k(\mathbf{x}, \mathbf{x}_i) = \psi(\mathbf{x} - \mathbf{x}_i)$ es definida positiva, la arquitectura RRBF (ecuación (II.48)) no incluye parámetro constante pero, por lo anterior, este puede conseguirse mediante una función k' condicionalmente definida positiva de orden 1.

Sin embargo, también con un núcleo definido positivo puede construirse también una red con parámetro constante, tal como mostramos a continuación. Si k es un núcleo definido positivo entonces el nuevo núcleo: $k'(\mathbf{x}, \mathbf{x}_i) = k(\mathbf{x}, \mathbf{x}_i) + a$, resulta condicionalmente definido positivo de orden 1 pues:

$$\sum_{i,j=1}^{r} c_i c_j k'(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j=1}^{r} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) + a \sum_{i,j=1}^{r} c_i c_j = \sum_{i,j=1}^{r} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) + a(\sum_{i=1}^{n} c_i)^2$$

expresión que es $\geq 0\ \forall \mathbf{c} \in \mathbb{R}^r, \forall \mathbf{x}_i \in \mathbb{R}^d, i = 1, ..., r$ si y sólo sí $\sum_{i=1}^{n} c_i = 0$.

Por tanto, $\mathcal{H}_{k'}$ es un SRKHS constituido por las funciones de la forma $f = \sum_{i=1}^{r} c_i k'(\cdot, \mathbf{x}_i) + b$ (y sus límites) con seminorma:

$$\|f\|_{\mathcal{H}_{k'}} = \langle f, f \rangle_{\mathcal{H}_{k'}} = \left\| \sum_{i=1}^{n} c_i k'(\cdot, \mathbf{x}_i) \right\|_{\mathcal{H}_{k'}} + \|b\|_{\mathcal{H}_{k'}} = \mathbf{c}^T \mathbf{K} \mathbf{c}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

con $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ que, por lo anterior, sólo está bien definida si $\mathbf{c}^T\mathbf{1}_r = 0$.

Así, por el teorema 39, k' da lugar a arquitecturas de la forma:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i k'(\mathbf{x}, \mathbf{x}_i) + c_0 = \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i) + a\sum_{i=1}^{n} c_i + c_0 = \sum_{i=1}^{n} c_i \psi(\mathbf{x} - \mathbf{x}_i) + c_0$$

ya en términos del núcleo definido positivo.

La inclusión del parámetro constante en la red aporta una mayor flexibilidad a la arquitectura permitiendo agilizar el entrenamiento de la red.

La red RRBF como Suavizador Lineal

Al poseer una arquitectura análoga, la red RRBF admite los mismos comentarios realizados para la red RBF para regresión (pág. 119), sobre las estrategias de entrenamiento en una y dos fases.

Sin embargo, cuando las funciones de transferencia $\psi_i(\mathbf{x}) = \psi(\mathbf{x} - \mathbf{x}_i)$, $i = 1 : n$, están prefijadas (no dependen de ningún parámetro libre adicional), el hecho de que el número de unidades ocultas sea igual al tamaño muestral, $h = n$, introduce algunas diferencias. Para el caso más general, la arquitectura es (II.62):

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \psi(\mathbf{x} - \mathbf{x}_i) + \sum_{j=1}^{r} b_j g_j(\mathbf{x})$$

con $\{g_j\}_{j=1}^{r}$ base del espacio $\Pi_m(\mathcal{X})$ y $k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x} - \mathbf{x}')$ es condicionalmente definida positiva de orden m (si $m = 0$, k es definida positiva y no hay término adicional). Si, además, se utiliza pérdida cuadrática, entonces, el principio ERM regularizado resulta:

$$\hat{\mathcal{R}}_n(f) = \|\mathbf{y} - \mathbf{f}\|_2^2 + \lambda \|f\|_{\mathcal{H}_k}^2 = \|\mathbf{y} - (\mathbf{K}\mathbf{c} + \mathbf{G}\mathbf{b})\|_2^2 + \lambda \mathbf{c}^T\mathbf{K}\mathbf{c} \tag{II.63}$$

donde $\mathbf{y} = (y_1, ..., y_n)^T$, $\mathbf{c} = (c_1, ..., c_n)^T$, $\mathbf{b} = (b_1, ..., b_r)^T$ y $\mathbf{G}$ es una matriz $n \times r$ con $\text{rango}(\mathbf{G}) = r$ y $(\mathbf{G})_{ij} = g_j(\mathbf{x}_i)$, $i = 1 : n$, $j = 1 : r$.

De esta forma, la minimización de dicha expresión es un problema de regresión *ridge* ([Hoerl y Kennard 1970*b*], [Hoerl y Kennard 1970*a*]), cuyas condiciones de óptimo para $\mathbf{c}$ son $\mathbf{K}\mathbf{K}\mathbf{c} + \mathbf{K}\mathbf{G}\mathbf{b} + \lambda\mathbf{K}\mathbf{c} = \mathbf{K}\mathbf{y}$ donde hemos utilizado que $\mathbf{K}$ es simétrica. Dicha ecuación, constituye un sistema de n ecuaciones con $n+r$ incógnitas, pero, al ser k condicionalmente definida positiva de orden m, han de verificarse las r condiciones adicionales $\mathbf{G}^T\mathbf{c} = \mathbf{0}_r$. Por tanto, las condiciones de óptimo son:

$$\begin{cases} \mathbf{K}\mathbf{K}\mathbf{c} + \mathbf{K}\mathbf{G}\mathbf{b} + \lambda\mathbf{K}\mathbf{c} = \mathbf{K}\mathbf{y} \\ \mathbf{G}^T\mathbf{c} = \mathbf{0}_r \end{cases} \tag{II.64}$$

Si, además, k es **estrictamente condicionalmente definida positiva** de orden m (Anexo B) entonces $\mathbf{c}^T\mathbf{K}\mathbf{c} > 0\ \forall \mathbf{c} \in \mathbb{R}^n \backslash \{\mathbf{0}_n\}$ (para que $\|\cdot\|_{\mathcal{H}_k}$ esté bien definida fuera del espacio nulo del SRKHS $\mathcal{H}_k$) lo que equivale a que $\mathbf{c}^T\mathbf{K}\mathbf{K}\mathbf{c} > 0$ (de lo contrario $\mathbf{c}^T\mathbf{K}\mathbf{K}\mathbf{c} = 0 \Leftrightarrow \mathbf{K}\mathbf{c} = 0 \Leftrightarrow \mathbf{c}^T\mathbf{K}\mathbf{c} = 0$) y a que $\mathbf{c}^T(\mathbf{K}\mathbf{K} + \lambda\mathbf{I}_n)(\mathbf{K}\mathbf{K} + \lambda\mathbf{I}_n)\mathbf{c} > 0$ sobre los mismos $\mathbf{c}$. Con ello:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$$\begin{cases} (\mathbf{K}\mathbf{K} + \lambda\mathbf{K})\mathbf{c} = \mathbf{K}\mathbf{y} - \mathbf{K}\mathbf{G}\mathbf{b} \\ \mathbf{G}^T\mathbf{c} = \mathbf{0}_r \end{cases}$$

posee solución única (véase también [Madych y Nelson 1988, p. 79]) y es la misma solución que proporciona:

$$\left\{ \begin{array}{l} (\mathbf{K}+\lambda\mathbf{I}_n)\mathbf{c} = \mathbf{y} - \mathbf{G}\mathbf{b} \\ \mathbf{G}^T\mathbf{c} = \mathbf{0}_r \end{array} \right. \equiv \left\{ \begin{array}{l} (\mathbf{K}+\lambda\mathbf{I}_n)\mathbf{c} + \mathbf{G}\mathbf{b} = \mathbf{y} \\ \mathbf{G}^T\mathbf{c} = \mathbf{0}_r \end{array} \right. \tag{II.65}$$

Aplicando el lema de inversión de matrices particionadas (pág. 413, Anexo G) se obtiene:

$$\begin{array}{l} \hat{\mathbf{c}} = (\mathbf{K}+\lambda\mathbf{I}_n)^{-1}[\mathbf{I}_n - \mathbf{G}[\mathbf{G}^T(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{G}]^{-1}\mathbf{G}^T(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}]\mathbf{y} \\ \hat{\mathbf{b}} = [\mathbf{G}^T(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{G}]^{-1}\mathbf{G}^T(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{y} \end{array} \tag{II.66}$$

con lo que la red RRBF es un suavizador lineal de la forma:

$$\begin{array}{rcl} \hat{\mathbf{f}}_n & = & \mathbf{K}\hat{\mathbf{c}} + \mathbf{G}\hat{\mathbf{b}} \\ & = & \{\mathbf{K}(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}[\mathbf{I}_n - \mathbf{G}[\mathbf{G}^T(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{G}]^{-1}\mathbf{G}^T(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}] \\ & & +\mathbf{G}[\mathbf{G}^T(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{G}]^{-1}\mathbf{G}^T(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\}\mathbf{y} \\ & = & \mathbf{S}\mathbf{y} \end{array} \tag{II.67}$$

donde la última ecuación es la formulación dual de la solución y $\mathbf{S}$ es la matriz *hat*.

Además, en el caso $\lambda = 0$, se tiene (véase el caso general en la pág. 414):

$$\hat{\mathbf{f}}_n = \{\mathbf{K}\mathbf{K}^{-1}[\mathbf{I}_n - \mathbf{G}[\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{K}^{-1}] + \mathbf{G}[\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{K}^{-1}\}\mathbf{y} = \mathbf{y}$$

con lo que, si se cumple dicha condición, la red RRBF es **interpoladora** de los datos. Ello parece lógico teniendo en cuenta que $\lambda = 0$ supone no penalizar la complejidad ($\|f\|^2_{\mathcal{H}_k}$) de la solución en el problema regularizado (II.63).

Cuando la función k es definida positiva, las ecuaciones que se obtienen son un caso particular de lo anterior sin términos polinómicos, es decir:

$$(\mathbf{K}+\lambda\mathbf{I}_n)\mathbf{c} = \mathbf{y} \Rightarrow \hat{\mathbf{c}} = (\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{y}$$

donde el término $\lambda\mathbf{I}_n$ garantiza la no singularidad de la matriz $(\mathbf{K}+\lambda\mathbf{I}_n)$ (aunque no su buen condicionamiento si λ es excesivamente pequeño). Con ello, la solución puede escribirse como:

$$\hat{\mathbf{f}}_n = \mathbf{K}\hat{\mathbf{c}} = \mathbf{K}(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{y} = \mathbf{S}\mathbf{y}$$

con matriz *hat* $\mathbf{S} = \mathbf{K}(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}$.

Por tanto, las redes RRBF admiten un análisis similar al realizado para las redes RBF, sólo que ahora, su formulación primal utiliza tantos términos como datos en la muestra.

Estabilizadores Radiales y Redes RRBF Resultantes

Debido a los resultados anteriores, cualquier función $k(\mathbf{x},\mathbf{x}') = \psi(\mathbf{x}-\mathbf{x}')$ definida positiva o condicionalmente definida positiva permite construir una red RRBF a partir de la elección del estabilizador:

$$\mathcal{E}_{\check{\psi}}(f) = \|f\|^2_{\mathcal{H}_k} = \int_{\mathbb{R}^d} \frac{\left|\tilde{f}(\boldsymbol{\omega})\right|^2}{\check{\psi}(\boldsymbol{\omega})} d\boldsymbol{\omega}$$

cuyo correspondiente RKHS o SRKHS, actuaría como espacio de hipótesis. Por construcción, la función de transferencia es invariante por traslaciones $\psi(\mathbf{x}-\mathbf{x}')$, pero a través de ella, es posible incorporar incorporar diferentes aspectos conocidos a priori sobre la verdadera función: invariancia ante rotaciones, grado de suavidad, etc.

USAL UNIVERSIDAD DE SALAMANCA DE CÓRDOBA

$\check{\psi}(\boldsymbol{\omega})$	$\mathcal{E}_{\check{\psi}}(f) = \|f\|^2_{\mathcal{H}_k}$	$k(\mathbf{x},\mathbf{x}') = \zeta(\|\mathbf{x}-\mathbf{x}'\|)$	m	**RRBF**
$e^{-\sigma^2\|\boldsymbol{\omega}\|^2/2}$	$\int_{\mathbb{R}^d} \left\|\tilde{f}(\boldsymbol{\omega})\right\|^2 e^{\sigma^2\|\boldsymbol{\omega}\|^2/2} d\boldsymbol{\omega}$	$\zeta(r) = e^{-r^2/(2\sigma^2)}$	0	Gausiana
$\frac{1}{1+\omega^2}$	$\int_{\mathbb{R}} (1+\omega^2) \left\|\tilde{f}(\omega)\right\|^2 d\omega$	$\zeta(r) = e^{-r}$	0	Laplaciana
$\frac{1}{\omega^2}$	$\int_{\mathbb{R}} \omega^2 \left\|\tilde{f}(\omega)\right\|^2 d\omega$	$\zeta(r) = r$	1	Spline lineal
$\frac{1}{\omega^4}$	$\int_{\mathbb{R}} \omega^4 \left\|\tilde{f}(\omega)\right\|^2 d\omega$	$\zeta(r) = r^{3/2}$	2	Spline cúbico
$\frac{1}{\|\boldsymbol{\omega}\|^{2p}}$	$\int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|^4 \left\|\tilde{f}(\boldsymbol{\omega})\right\|^2 d\boldsymbol{\omega}$	$\zeta(r) = \begin{cases} r^{2p-d}\ln r & 2p > d \text{ par} \\ r^{2p-d} & \text{resto} \end{cases}$	p	Thin Plate Splines

Tabla 6. Ejemplos de estabilizadores y redes RRBF a las que dan lugar

En general, todas las funciones definidas positivas o condicionalmente definidas positivas típicamente utilizadas en el problema de interpolación con base radial (Anexo B) son admisibles como funciones de transferencia de una RBF regularizada, por ejemplo: gausiana, multicuádrica inversa, multicuádrica, etc.

La tabla 6 muestra casos particulares bien conocidos que generan redes RRBF de regularización con funciones de transferencia radial. La potencia generalizadora del enfoque basado en RKHS y en núcleos reproductores que hemos seguido en los apartados anteriores, puede apreciarse si se tiene en cuenta el laborioso análisis requerido en [Wahba 1990*b*] para la obtención de los modelos spline.

Pero, hay tantas alternativas a elegir para la función de transferencia $\psi(\mathbf{x}-\mathbf{x}') = k(\mathbf{x},\mathbf{x}')$ como funciones definidas positivas k, (p. ej. [Berg *et al.* 1984, p. 79], Anexo B). En particular:

1. Cualquier función de densidad $\check{\psi}$ satisface el teorema de Bochner, por lo que la función característica correspondiente $\psi = \check{\check{\psi}}$ es definida positiva y puede ser utilizada como núcleo.

2. Funciones de covarianza de procesos estocásticos.

3. Núcleos reproductores conocidos de algún RKHS. Uno bien conocido es el **núcleo de Bergman** (*Bergman kernel*), definido en algún subespacio completo $\mathcal{G} \subset L_2(\mathbb{R}^d)$ como: $k(\mathbf{x},\mathbf{x}') = \sum_i \gamma_i(\mathbf{x})\gamma_i(\mathbf{x}')$ donde $\{\gamma_i\}$ es una base ortonormal de $\mathcal{G}$.

   Casos particulares relevantes de este tipo de núcleo, son los núcleos de los subespacios de un **análisis de multirresolución** mediante **wavelets**, (p. ej. [Walter 1994, sec. 3.3], [Vidakovic 1999, sec. 6.2.1 y p. 175], [Härdle *et al.* 1998, cap. 8 y p. 127]), con lo que una combinación de ellos permite implementar dicho análisis mediante redes RBF de regularización.

4. Núcleos obtenidos a partir de núcleos mediante operaciones permitidas (Anexo B y p. ej. [Aronszajn 1950], [Cristianini y Shawe-Taylor 2000, sec. 3.3.2], [Schölkopff y Smola 2002, cap. 2 y sec. 4.4]), tales como la suma, el producto tensorial, funciones polinómicas de núcleos, exponenciación, etc.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Estabilizadores de Producto Tensorial

Si el espacio de variables regresoras posee una dimensión elevada, una alternativa es elegir estabilizadores de producto tensorial:

$$\|f\|^2_{\mathcal{H}_k} = \int_{\mathbb{R}^d} \frac{\left|\tilde{f}(\boldsymbol{\omega})\right|^2}{\prod_{j=1}^d \check{\psi}_j(\omega_j)} d\boldsymbol{\omega}$$

con $\check{\psi}_j$, $j = 1, ..., d$, funciones positivas, simétricas e integrables en $\mathbb{R}$, con $\lim_{\omega\to\infty} \check{\psi}_j(\omega) = 0$. Entonces, si $\psi_j(x - x')$ $(= k_j(x, x'))$ es la transformada de Fourier de $\check{\psi}_j(\omega)$, k_j es definida positiva y por tanto núcleo reproductor de un RKHS en la dimensión j. Con ello, la función de transferencia resulta del producto tensorial de los núcleos:

$$\psi(\mathbf{x} - \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^d k_j(x_j, x'_j) = \prod_{j=1}^d \psi_j(x_j - x'_j)$$

que es núcleo reproductor de $\mathcal{H}_k = \bigotimes_{j=1}^d \mathcal{H}_{k_j}$ (teorema 125, Anexo B). Con ello, la forma general de correspondiente RBF regularizada, es de la forma:

$$f(\mathbf{x}) = \sum_{i=1}^n c_i \psi(\mathbf{x} - \mathbf{x}_i) = \sum_{i=1}^n c_i \prod_{j=1}^d \psi_j(x_j - x_{i,j})$$

Los núcleos gausiano y laplaciano multidimensionales vistos anteriormente (tabla 6) pueden obtenerse también mediante este enfoque tensorial:

$$\check{\psi}_j(\omega_j) = e^{-\omega_j^2\sigma^2/2} \rightarrow \psi(\mathbf{x} - \mathbf{x}') = \prod_{j=1}^d \psi_j(x_j - x'_j) = \prod_{j=1}^d e^{-(x_j - x'_j)^2/(2\sigma^2)} = e^{-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{x}'\|_2^2}$$

$$\check{\psi}_j(\omega_j) = e^{-\omega_j^2\sigma^2/2} \rightarrow \psi(\mathbf{x} - \mathbf{x}') = \prod_{j=1}^d \psi_j(x_j - x'_j) = \prod_{j=1}^d e^{-|x_j - x'_j|} = e^{-\|\mathbf{x}-\mathbf{x}'\|_1}$$

Estabilizadores Aditivos

El estabilizador de producto tensorial, permite introducir en el problema información a priori sobre el comportamiento de la verdadera función en cada dimensión o, con igual filosofía, grupos de dimensiones del espacio de entrada, (grupos de variables independientes).

Sin embargo, esta información puede no estar disponible o no ser relevante, y en su lugar puede interesar considerar la verdadera función como suma de componentes, con la filosofía de los modelos aditivos[60], (p. ej. [Hastie y Tibshirani 1990]), por ejemplo, superposiciones de diverso grado de suavidad, componentes que reflejen interacciones entre ciertos grupos de variables, etc.:

$$f(\mathbf{x}) = \sum_{j=1}^h f_j(\mathbf{x})$$

[60] Ello, al margen de que la propia red neuronal MLF es un modelo aditivo en sí mismo.

donde una de las funciones puede ser constante. Si la información a priori sobre cada componente se modeliza mediante un estabilizador:

$$\mathcal{E}_j(f_j) = \int_{\mathbb{R}^d} \frac{\left|\tilde{f}_j(\boldsymbol{\omega})\right|^2}{\breve{\psi}_j(\boldsymbol{\omega})} d\boldsymbol{\omega}$$

con las condiciones anteriores para $\breve{\psi}_j$, entonces se determina unívocamente un RKHS con núcleo k_j, y su suma:

$$\psi(\mathbf{x}-\mathbf{x}') = k(\mathbf{x},\mathbf{x}') = \sum_{j=1}^{h} k_j(\mathbf{x},\mathbf{x}') = \sum_{j=1}^{h} \psi_j(\mathbf{x}-\mathbf{x}')$$

es núcleo (teorema 121, Anexo B) del RKHS $\mathcal{H}_k = \bigoplus_{j=1}^{h} \mathcal{H}_{k_j}$, con lo que la correspondiente RRBF pose la forma general:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \psi(\mathbf{x}-\mathbf{x}_i) = \sum_{i=1}^{n} c_i \sum_{j=1}^{h} \psi_j(\mathbf{x}-\mathbf{x}')$$

Pero esta opción, supone valores c_i comunes para los RKHS de cada componente aditiva. Una alternativa es considerar el principio ERM regularizado con factores para cada componente:

$$\begin{aligned} \hat{\mathcal{R}}_n(f) &= \hat{R}_n(f) + \lambda \sum_{j=1}^{h} \frac{1}{\lambda_j} \|f_j\|^2_{\mathcal{H}_{k_j}} = \hat{R}_n(f) + \lambda \sum_{j=1}^{h} \int_{\mathbb{R}^d} \frac{\left|\tilde{f}_j(\boldsymbol{\omega})\right|^2}{\lambda_j \breve{\psi}_j(\boldsymbol{\omega})} d\boldsymbol{\omega} \\ \Rightarrow \quad f(\mathbf{x}) &= \sum_{i=1}^{n} \sum_{j=1}^{h} c_i \lambda_j \psi_j(\mathbf{x}-\mathbf{x}') \end{aligned} \tag{II.68}$$

donde $\lambda_j > 0$, $j = 1:h$, y, por la linealidad de la transformada de Fourier, $\widetilde{(\lambda_j \breve{\psi}_j)} = \lambda_j \widetilde{\breve{\psi}_j} = \lambda_j \psi_j$, lo cuál, provoca que el núcleo reproductor de cada RKHS sea $\lambda_j k_j(\mathbf{x},\mathbf{x}')$ y, así, el núcleo de $\mathcal{H}_k$ sea $k(\mathbf{x},\mathbf{x}') = \sum_{j=1}^{h} \lambda_j k_j(\mathbf{x},\mathbf{x}')$.

La formulación (II.68) permite estimar independientemente los parámetros $c'_i = c_i \lambda_j$, evitando el problema anterior. Nótese, además, que, en este caso, la red resultante posee, en general, funciones de transferencia diferentes en las unidades de la capa oculta.

II.5.3 Support Vector Machines (SVM)

El origen de las **support vector machines**[61] (SVM), reside en el concepto de hiperplano separador óptimo de una muestra de datos en un problema de clasificación, [Vapnik 1982]. La consideración de la familia de hiperplanos separadores como fronteras de decisión, permite la aplicación del método SRM de selección del modelo (pág. 49), ([Vapnik 1995], [Vapnik 1998]), dando lugar al hiperplano separador que separa la muestra con un margen máximo. Pero los hiperplanos separadores sufren de dos debilidades en muchos problemas de aplicación: por una parte, exigen la separabilidad perfecta de la muestra y, por otra, poseen carácter lineal.

Las redes **support vector machines** se desarrollan para evitar estas limitaciones, siendo sus hitos históricos fundamentales, los siguientes:

USC

[61] No utilizamos la traducción española de esta denominación, por ser ya un término totalmente acuñado.

1. La transformación previa del espacio de entrada $\mathcal{X} \subset \mathbb{R}^d$ en un **espacio de características** (*feature espace*) de mayor dimensión, $\mathcal{U} = \phi(\mathcal{X}) \subset \mathbb{R}^s$, (donde s puede ser infinito), en el que las fronteras lineales de los hiperplanos separadores dan lugar, mediante la transformación inversa, a fronteras no lineales en el espacio de entrada, (*Kernel Trick* [Boser *et al.* 1992], véase introducción en la pág. 82 al generalizar el perceptrón).

2. La creación del algoritmo *Soft-Margin* [Cortes y Vapnik 1995], para enfrentarse a muestras no separables o a problemas donde no interesa la separabilidad perfecta (p. ej. ruido en las observaciones).

3. La generalización de las SVM a problemas de regresión mediante la utilización de la pérdida ε**-insensible** de Vapnik, [Drucker *et al.* 1997].

Dos problemas importantes es necesario resolver para poder utilizar de forma efectiva las SVM: la selección de un algoritmo de entrenamiento eficiente, y la selección de los hiperparámetros del modelo.

Con respecto al primero, diversos trabajos ([Osuna *et al.* 1997], [Joachims 1999], [Collobert y Bengio 2001], [Vanderbei 1998], [Smola y Schölkopf 1998], etc.), han puesto a disposición algoritmos que pueden ser utilizados con éxito.

Respecto al segundo, en el problema de clasificación, existen enfoques que pueden ser utilizados basados en el concepto de dimensión VC, (p. ej. [Vapnik 1998], [Anguita *et al.* 2002]), pero en el caso de regresión, el problema de selección del modelo en la práctica, no puede considerarse todavía resuelto con satisfacción: cuando no se dispone de muestra de validación, lo más seguro es recurrir a técnicas de validación cruzada, (la literatura al respecto es ingente; para una visión general puede verse por ejemplo, [Schölkopff y Smola 2002], [Cristianini y Shawe-Taylor 2000], además de [Vapnik 1998].

SVM para Clasificación

Hiperplano Separador de Máximo Margen

Sea un problema de clasificación con dos clases y una muestra de datos $\mathbf{z}^n$ con $z_i = (\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, $y_i \in \mathcal{Y} = \{-1, 1\}$, $i = 1 : n$ **linealmente separable** (pág. 79), es decir, que separarse sin errores mediante una función de decisión de la forma[62] (arquitectura perceptrón) $f_{\mathbf{w},b}(\mathbf{x}) = \text{signo}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$, con $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$, es decir: $y_i f_{\mathbf{w},b}(\mathbf{x}_i) \geq 0$, $i = 1 : n$.

En este contexto, se define el hiperplano separador óptimo, como el **hiperplano separador de máximo margen**, es decir, que maximiza su distancia a las clases (figura 20).

Definición 43 *Dado que la distancia entre un punto* $\mathbf{x}$ *y un hiperplano* $\tau_{\mathbf{w},b} \equiv \{\mathbf{x} \in \mathcal{X} : \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$ *es* $\frac{1}{\|\mathbf{w}\|} |\langle \mathbf{w}, \mathbf{x} \rangle + b|$, *el* ***margen*** *(o* ***margen geométrico****) del hiperplano* $\tau_{\mathbf{w},b}$ *con respecto a la muestra* $\mathbf{z}^n$*, se define como (véase definición 31):*

$$\gamma(\tau_{\mathbf{w},b}) = \min_{i \in \{1:n\}} \frac{1}{\|\mathbf{w}\|} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = \frac{1}{\|\mathbf{w}\|} \min_{i \in \{1:n\}} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = \frac{1}{\|\mathbf{w}\|} \chi(\tau_{\mathbf{w},b}) \qquad \text{(II.69)}$$

donde $\chi(\tau_{\mathbf{w},b})$ *se denomina* ***margen funcional****, que depende del tamaño de* $\mathbf{w}$.

[62] En este apartado, adoptamos la notación $b = w_0$ para el parámetro independiente del perceptrón, por cuanto puede considerarse estándar en la literatura al respecto.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

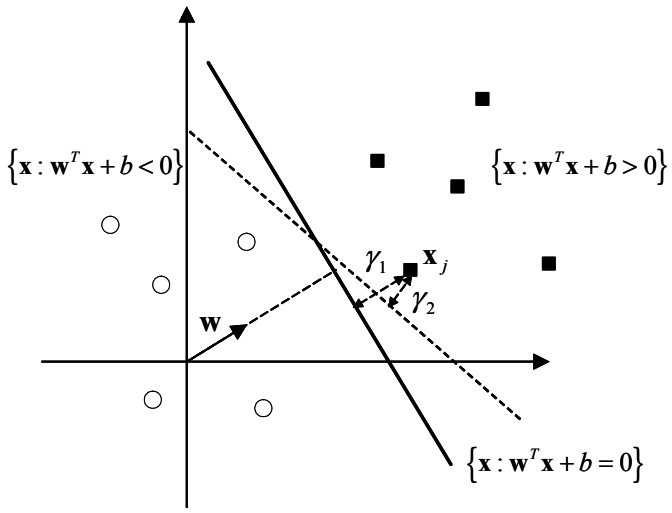


Figura 20. El hiperplano en línea continua produce un mayor margen γ_1 en la muestra que el representado con linea a trozos con margen γ_2.

Con ello, el hiperplano separador óptimo se obtiene como solución del siguiente programa con restricciones:

$$\max_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \{\gamma(\tau_{\mathbf{w},b}) = \min_{i\in\{1:n\}} |\langle \mathbf{w}, \mathbf{x}_i\rangle + b|\}$$
$$\text{sujeto a: } \|\mathbf{w}\| = 1$$

Alternativamente, en lugar de normalizar el vector $\mathbf{w}$ en la expresión (II.69), puede normalizarse el margen funcional, $\chi(\tau_{\mathbf{w},b}) = 1$ (en cuyo caso se dice que el hiperplano está en **forma canónica**), con lo que el problema de optimización persigue maximizar $1/\|\mathbf{w}\|$ con la restricción $\chi(\tau_{\mathbf{w},b}) = 1$, es decir:

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 \tag{70}$$
$$\text{sujeto a: } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1, i = 1, \ldots, n$$

El problema anterior, es cuadrático con restricciones lineales con lo que, las condiciones de óptimo de Kuhn-Tucker (teorema 206, Anexo D) son necesarias y suficientes.

La solución puede obtenerse a través del problema dual, y resulta combinación lineal de un subconjunto de puntos de la muestra, denominados **vectores soporte** (*support vectors*) pues la solución *descansa* en dichos puntos, (la deducción es similar a la que se muestra más adelante para el problema de regresión y puede verse p. ej. en [Schölkopff y Smola 2002], [Cristianini y Shawe-Taylor 2000] y [Vapnik 1998]):

$$\mathbf{w} = \sum_{\text{S.V.}} \beta_i \mathbf{x}_i \Rightarrow f_{\mathbf{w},b}(\mathbf{x}) = \sum_{\text{S.V.}} \beta_i \langle \mathbf{x}_i, \mathbf{x}\rangle + b \tag{II.71}$$

Con ello, dado un nuevo punto $\mathbf{x}$ a clasificar, la solución se obtiene mediante el producto interno con los puntos de la muestra.

Esta propiedad no es en absoluto extraña, pues por ejemplo, en la regresión *ridge* que se deduce del principio ERM con pérdida cuadrática y modelo lineal:

$$\min_{\bar{\mathbf{w}}} \hat{\mathcal{R}}_n(f_{\bar{\mathbf{w}}}) = \min_{\mathbf{w},b}\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})^2 + \frac{\lambda}{2}\bar{\mathbf{w}}^T\bar{\mathbf{w}}\} \tag{II.72}$$

con $\bar{\mathbf{w}} = (b, \mathbf{w}^T)^T$, $\bar{\mathbf{x}} = (1, \mathbf{x}^T)^T$ y $f_{\bar{\mathbf{w}}}(\mathbf{x}) = \bar{\mathbf{x}}^T\bar{\mathbf{w}} = \mathbf{x}^T\mathbf{w} + b$, la condición de óptimo es de la forma:

$$\frac{\partial\hat{\mathcal{R}}_n}{\partial\bar{\mathbf{w}}} = -\sum_{i=1}^{n}(y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})\bar{\mathbf{x}}_i + \lambda\bar{\mathbf{w}} = \mathbf{0}_{d+1} \Rightarrow \bar{\mathbf{w}} = \frac{1}{\lambda}\sum_{i=1}^{n}(y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})\bar{\mathbf{x}}_i = \sum_{i=1}^{n}\beta_i\bar{\mathbf{x}}_i$$

con $\beta_i = \frac{1}{\lambda}(y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})$.

Como hemos anticipado al principio, la relevancia del hiperplano de máximo margen reside en que es la solución óptima para el método SRM de selección del modelo, como puede verse en lo siguiente.

Si se considera un conjunto de familias de funciones de decisión basadas en hiperplanos separadores normalizados en margen funcional:

$$\mathcal{F}_\rho = \{g_{\mathbf{w},b}(\mathbf{x}) = \text{signo}(\langle\mathbf{w},\mathbf{x}\rangle + b)), \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\| \leq \rho, b \in \mathbb{R}, \min_{i=1:n}|\langle\mathbf{w},\mathbf{x}\rangle + b)| = 1\}$$

entonces, tomando una sucesión $\{\rho_t\}_{t=1}^{\infty}$ creciente de valores, puede construirse una estructura anidada de familias:

$$\mathcal{F}_{\rho_1} \subset \mathcal{F}_{\rho_2} \subset \cdots \subset \mathcal{F}_{\rho_t} \subset \cdots \subset \bigcup_{t=1}^{\infty}\mathcal{F}_{\rho_t} = \mathcal{F}$$

en la que puede aplicarse el principio SRM.

Nótese que la cota superior $\|\mathbf{w}\| \leq \rho$ equivale a una cota inferior para el margen $1/\|\mathbf{w}\| \geq 1/\rho$ de los correspondientes hiperplanos $f_{\mathbf{w},b}(\mathbf{x}) = \langle\mathbf{w},\mathbf{x}\rangle + b$, con lo que las familias más pequeñas (de menor complejidad) exigen un mayor margen a sus hiperplanos.

Ello lo confirma el siguiente teorema:

Teorema 44 *[Vapnik 1982], [Vapnik 1998] Sea D el radio de la menor bola centrada en el origen que contiene a los datos $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X} \subset \mathbb{R}^d$. Entonces, la dimensión VC de $\mathcal{F}_\rho$ satisface la siguiente cota:*

$$\text{VCdim}(\mathcal{F}_\rho) = v_{\mathcal{F}_\rho} \leq \min\{D^2\rho^2, d\} + 1$$

Dicho teorema, expresa que, fijado D, la cota a la capacidad de la familia $\mathcal{F}_t$ crece con el valor máximo ρ permitido para $\|\mathbf{w}\|$ (a medida que el margen mínimo $1/\rho \leq 1/\|\mathbf{w}\|$ exigido en ella, decrece).

En la suposición anterior de separabilidad de la muestra, los hiperplanos separadores producirán un riesgo empírico nulo $\hat{R}_n(g_{\mathbf{w},b}) = 0$, con lo que el riesgo garantizado en cada familia verifica con probabilidad $1 - \eta$ (ecuación (I.10) o (I.36)):

$$R_{\mathcal{F}_\rho}(g) \leq \hat{R}_{n,\mathcal{F}_\rho}(g) + \varepsilon(n,\eta,v_{\mathcal{F}_\rho}) = \varepsilon(n,\eta,v_{\mathcal{F}_\rho})$$

Por tanto, al ser $\varepsilon(n,\eta,v)$ creciente en v interesa (dada una muestra, es decir, fijados d y D) elegir la familia $\mathcal{F}_{\rho_t}$ con mínima complejidad $v_{\mathcal{F}_{\rho_t}}$, es decir, la familia con un margen mínimo $1/\rho_t$ exigido en ella, que sea máximo entre todas las familias.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Así, **elegir el hiperplano de máximo margen supone implementar el principio SRM**.

Hiperplano Separador Soft-Margin

Cuando la muestra no es linealmente separable, o no interesa su separación perfecta, (lo que es frecuente en la práctica, p. ej. debido al ruido de los datos), es preciso admitir un conjunto de observaciones mal clasificadas, (enfoque *soft margin* en contraposición del anterior, denominado *hard margin*), lo que puede formularse mediante variables *slack* (Anexo D), $\xi_i \geq 0$, $i = 1 : n$, que permiten la violación de las restricciones:

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \xi_i \geq 1,\ i = 1 : n$$

En este punto, aparece un conflicto de intereses entre un número mínimo de observaciones mal clasificadas y una clasificación lo más robusta posible. Este conflicto puede formularse como [Cortes y Vapnik 1995]:

$$\begin{array}{c} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} \{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i\} \\ \left\{ \begin{array}{ll} y_i(\langle \mathbf{w}, \mathbf{x}\rangle + b) \geq 1 - \xi_i & i = 1, \ldots, n \\ \xi_i \geq 0 & \end{array} \right. \end{array} \tag{II.73}$$

donde el parámetro[63] C expresa la importancia asignada a los casos mal clasificados. Nótese que el problema (II.73) no es una formulación extraña pues equivale al problema:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\ell(y_i, f_{\mathbf{w},b}(\mathbf{x}_i))\}$$

utilizando como pérdida:

$$\ell(y, f(\mathbf{x})) = [1 - yf(\mathbf{x})]_+ \doteq \left\{ \begin{array}{ll} 1 - yf(\mathbf{x}) & \text{si } 1 - yf(\mathbf{x}) \geq 0 \\ 0 & \text{en caso contrario} \end{array} \right. \tag{II.74}$$

De nuevo, el programa (II.73) es un programa cuadrático con restricciones lineales, cuya solución puede obtenerse mediante el programa dual, y es de forma análoga a (II.71).

En dicho problema, C es una constante que determina el equilibrio a realizar entre la maximización del margen y la minimización del error de clasificación en la muestra. Por tanto, la determinación a priori de C no puede realizarse de forma intuitiva. [Schölkopf, Smola, Williamson y Bartlett 2000] proponen una modificación del programa anterior sustituyendo dicho parámetro por otro parámetro ν, que representa (cota superior) la fracción de datos en la frontera o fuera del margen y el número de vectores soporte (cota inferior). El nuevo programa se formula como:

$$\begin{array}{c} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} \{\frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{n}\sum\limits_{i=1}^{n}\xi_i\} \\ \left\{ \begin{array}{ll} y_i(\langle \mathbf{w}, \mathbf{x}\rangle + b) \geq \kappa - \xi_i & i = 1, \ldots, n \\ \xi_i \geq 0,\ \rho \geq 0 & \end{array} \right. \end{array}$$

donde κ es una nueva variable que sirve para controlar el margen (ahora éste es $2\kappa/\|\mathbf{w}\|$). Por tanto, la maximización del margen se representa ahora mediante el término $\frac{1}{2}\|\mathbf{w}\|^2 - \nu\kappa$, mientras que la bondad de la clasificación sigue representándose mediante la suma de las variables *slack*.

[63] Adoptamos C como parámetro regularizador en lugar de λ, como hemos venido haciendo, pues tal notación es muy estándar.

El Método del Núcleo (*Kernel Trick*): Support Vector Machines

Con el enfoque *Soft Margin*, se resuelve el problema cuando la muestra no es linealmente separable o no interesa lograr esa separabilidad perfecta. Sin embargo, seguimos restringidos a soluciones de tipo lineal, de limitada utilidad en muchos problemas prácticos.

Una táctica para resolver esta dificultad, es utilizar una transformación no lineal del espacio de entrada (al modo de la regresión polinómica o del perceptrón generalizado –pág. 82), de tal forma que, las soluciones lineales en el espacio transformado (**espacio de características** o *features)*, den lugar, mediante la transformación inversa, a soluciones no lineales en el espacio original.

Las redes MLF vistas con anterioridad basan su potencial en la misma táctica, pero no se plantean objetivo alguno sobre la dimensión del nuevo espacio. Por ejemplo, en el caso de las MLP y las RBF, la determinación de la dimensión más adecuada para el nuevo espacio descansa en métodos de selección del modelo. En el caso de las RRBF, dicha dimensión es el tamaño muestral. Sin embargo, aquí, la particularidad de dicha transformación es que, el nuevo espacio, posea una mayor dimensión que el original, (posiblemente infinita).

Pero, en el caso de los hiperplanos separadores (II.71), esta estrategia plantea el problema de cómo gestionar el producto interno en un espacio de gran dimensión.

La solución a este problema fué propuesta por [Boser *et al.* 1992], y es la misma que la expuesta en la generalización del perceptrón: la representación del producto interno en el nuevo espacio, mediante una función definida positiva, de tal forma que, si la transformación elegida es $\boldsymbol{\phi} : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathcal{U}$ con $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x}) \in \mathcal{U}$ posiblemente de dimensión infinita, entonces:

$$\left\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \right\rangle = \sum_i \phi_i(\mathbf{x})\phi_i(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}') \tag{II.75}$$

es un producto interno bien definido, $\phi_i(\mathbf{x})$ son las componentes de $\boldsymbol{\phi}(\mathbf{x})$.

Para avalar esta solución, se tiene el siguiente teorema de [Parthasarathy y Schmidt 1972] (más general que el teorema 117 del Anexo B):

Teorema 45 *Si $\mathcal{X}$ es un espacio topológico y k es una función continua definida positiva en $\mathcal{X} \times \mathcal{X}$, entonces existe un espacio de Hilbert $\mathcal{H}$ y una aplicación continua $\boldsymbol{\phi} : \mathcal{X} \rightarrow \mathcal{H}$, tal que $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$, se verifica $k(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle$.*

Si, utilizando el método anterior, se plantea el problema de clasificación en el nuevo espacio de características, donde la muestra es $\{(y_i, \boldsymbol{\phi}_i)\}_{i=1}^n$ con $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\mathbf{x}_i)$, $i = 1 : n$, entonces la expresión (II.71) del hiperplano separador de máximo margen en dicho espacio, es[64]:

$$\begin{aligned} \mathbf{w} &= \sum_{\text{S.V.}} \beta_i \boldsymbol{\phi}(\mathbf{x}_i) \\ &\Rightarrow \quad f_{\mathbf{w},b}(\boldsymbol{\phi}(\mathbf{x})) = \sum_{\text{S.V.}} \beta_i \left\langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}) \right\rangle + b = \sum_{\text{S.V.}} \beta_i k(\mathbf{x}_i, \mathbf{x}) + b \end{aligned} \tag{II.76}$$

Dicha expresión es la fórmula general de las **support vector machines** (SVM) para clasificación tanto en el caso *hard margin,* como en el caso *soft margin*, que proviene de la resolución

[64] En este apartado, adoptamos la notación $\boldsymbol{\beta} = \mathbf{c}$, $b = c_0$ para los parámetros del nivel de salida de la red y $\psi = k$ para la función de transferencia, ya que, para las SVM, puede considerarse notación estándar.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

de los problemas (II.70) y (II.73) en el espacio de características. Por ejemplo, el programa del caso *soft margin* posee ahora la forma:

$$\begin{array}{c} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} \{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i\} \\ \left\{\begin{array}{ll} y_i(\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b) \geq 1 - \xi_i & i = 1,\ldots,n \\ \xi_i \geq 0 \end{array}\right. \end{array} \tag{II.77}$$

Con el método del núcleo, se resuelve el problema de la dimensionalidad del nuevo espacio. Pero, además, no es necesario conocer específicamente dicho espacio, si no que basta el conocimiento de k para construir la solución.

De todas formas, la tarea de modelización puede realizarse incorporando la información a priori a través de k o bien, a través del espacio de características; aunque el primer enfoque es el más usual, pues un mismo núcleo k puede resultar de diferentes transformaciones correspondientes a diferentes espacios de características verificando la condición (II.75).

El espacio de características es relevante tan sólo en términos de interpretación de la solución, pues ésta sólo depende del núcleo utilizado. Entre estas transformaciones, poseen interés, al menos teórico, las siguientes, (p. ej. [Schölkopff y Smola 2002]):

1. La **transformación de Mercer**[65], (*Mercer map*) que da lugar a las **características de Mercer**, definidas a través del teorema homónimo (teorema 112 Anexo B), mediante las autofunciones ortogonales $\{\varphi_j\}_{j=1}^{\infty}$ del operador integral de núcleo k :

$$\begin{array}{rcl} k(\mathbf{x},\mathbf{x}') & = & \displaystyle\sum_{j=1}^{\infty}\mu_j\varphi_j(\mathbf{x})\varphi_j(\mathbf{x}') = \left\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\right\rangle \\ \text{donde p. ej.:} & & \boldsymbol{\phi}(\mathbf{x}) = (\sqrt{\mu_1}\varphi_1(\mathbf{x}), ..., \sqrt{\mu_j}\varphi_j(\mathbf{x}), ...)^T \end{array}$$

   En la transformación de Mercer, si alguno de los autovalores μ_j es cero o muy pequeño, las componentes de las frecuencias correspondientes a φ_j no participan en el espacio de características, con lo que el efecto del núcleo es evitar dichas frecuencias en la solución resultante.

   Un caso particular muy utilizado es el núcleo gausiano[66] cuyos autovalores son todos distintos de cero (con lo que el espacio de características posee dimensión infinita), pero tendiendo a cero muy rápidamente. Con ello, realmente, se están filtrando progresivamente las altas frecuencias.

2. La **transformación núcleo** (*kernel map*), definida por $\boldsymbol{\phi}(\mathbf{x}) = k_{\mathbf{x}} = k(\cdot,\mathbf{x}) \in L_2(\mathcal{X})$ donde en el RKHS de núcleo k, $k_{\mathbf{x}}$ verifica:

$$\langle f, k_{\mathbf{x}}\rangle = f(\mathbf{x})$$

   es decir, $k_{\mathbf{x}}$, $\mathbf{x} \in \mathcal{X}$ son los representantes de Riestz de los operadores evaluación en el RKHS de núcleo k. Así, el espacio de características es un subespacio del espacio funcional consti-

[65] Nótese que esta transformación es diferente de la transformación de Karhunen-Loève (teorema 120 Anexo B), construida a partir de las autofunciones $\{\varphi_j\}$ de la función de covarianza k de un proceso estacionario de segundo orden $\mathsf{Z}(\mathbf{x}_i)$, definiéndose sus características principales U_j mediante el producto interno de $L_2(\mathcal{X})$, mediante $\mathsf{U}_j = \int \mathsf{Z}(\mathbf{x})\varphi_j(\mathbf{x})d\mathbf{x}$.

[66] Este núcleo transforma el espacio de entrada en la intersección entre la esfera de radio uno y un cuadrante del espacio de carácterísticas ya que, (sin factor de normalización): $\|\boldsymbol{\phi}(\mathbf{x})\|^2 = \langle\boldsymbol{\phi}(\mathbf{x}),\boldsymbol{\phi}(\mathbf{x})\rangle = k(\mathbf{x},\mathbf{x}) = 1$ y además, $\cos(\measuredangle(\phi(\mathbf{x}),\phi(\mathbf{x}'))) = \langle\boldsymbol{\phi}(\mathbf{x}),\boldsymbol{\phi}(\mathbf{x}')\rangle = k(\mathbf{x},\mathbf{x}') > 0$, con lo que el ángulo entre dos puntos es menor que $\pi/2$.

tuido por todas las combinaciones lineales de representantes $k_{\mathbf{x}} : f = \sum_j c_j k_{\mathbf{x}_j}$ (el RKHS asociado a k), con lo que su producto interno es:

$$\left\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \right\rangle = \langle k_{\mathbf{x}}, k_{\mathbf{x}'} \rangle = k(\mathbf{x}, \mathbf{x}')$$

3. La **transformación núcleo empírica**, (*empirical kernel map*), que consiste una aproximación de la anterior, utilizando la transformación anterior $\boldsymbol{\phi}(\mathbf{x}) = k_{\mathbf{x}}$ valorada únicamente en los datos de la muestra $\{\mathbf{x}_i\}_{i=1}^n$:

$$\boldsymbol{\phi}(\mathbf{x}) = k_{\mathbf{x}} \mid_{\{\mathbf{x}_1,...,\mathbf{x}_n\}} \doteq (k(\mathbf{x}, \mathbf{x}_1), ..., k(\mathbf{x}, \mathbf{x}_n))^T$$

   En la práctica, esta transformación no pierde información con respecto a la anterior pues en la solución del problema, la función $k_{\mathbf{x}}$ sólo se evalúa en los puntos de la muestra.

   Sin embargo, ahora no está garantizado que $\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle = k(\mathbf{x}, \mathbf{x}')$ pues falta definir un producto interno $\langle \cdot, \cdot \rangle$ en $\mathbb{R}^n$ para ello. Si se impone: $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle$, $i, j = 1 : n$, entonces, escribiendo el producto interno en la forma $< \cdot, \mathbf{M} \cdot >$, donde $\mathbf{M}$ es una matriz definida positiva, se obtiene la condición: $\mathbf{K} = \mathbf{KMK}$, siendo $\mathbf{K}$ la denominada matriz **Gram** con $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

Esta última transformación permite relajar la condición de definida positiva para la función k, gracias al siguiente teorema:

Teorema 46 *[Schölkopff y Smola 2002, prop. 2.16] Si los datos $\mathbf{x}_1, ..., \mathbf{x}_n$ y la función k son tales que la **matriz de Gram** $\mathbf{K}$, con $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, es definida positiva, entonces es posible construir una transformación $\boldsymbol{\phi}$ en un espacio $\mathcal{U}$ con $\dim(\mathcal{U}) = n$, tal que $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle$, $i, j = 1 : n$.*

Y, a la inversa: dada una transformación arbitraria $\boldsymbol{\phi}$, en algún espacio de características $\mathcal{U}$ dotado de producto interno, la matriz $\mathbf{K}$ con $(\mathbf{K})_{ij} = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle$ es definida positiva.

Demostración. Como $\mathbf{K}$ es definida positiva es simétrica, entonces podemos escribir $\mathbf{K} = \mathbf{SDS}^T$ con $\mathbf{D}$ diagonal y $\mathbf{D}$ matriz con columnas ortogonales. Entonces:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{SDS}^T)_{ij} = \langle \mathbf{s}_i, \mathbf{D}\mathbf{s}_j \rangle = \left\langle \sqrt{\mathbf{D}}\mathbf{s}_i, \sqrt{\mathbf{D}}\mathbf{s}_j \right\rangle \geq 0, i, j = 1 : n$$

donde $\mathbf{s}_i$ son las filas de $\mathbf{S}$ y $\sqrt{\mathbf{D}}$ es la matriz diagonal con elementos diagonales $\sqrt{(\mathbf{D})_{ii}}$. Es decir, $\mathbf{K}$ es la matriz de Gram de los vectores $\sqrt{(\mathbf{D})_{ii}}\mathbf{s}_i$. Así, basta definir la transformación:

$$\boldsymbol{\phi} : \mathbf{x}_i \rightarrow \sqrt{(\mathbf{D})_{ii}}\mathbf{s}_i$$

A la inversa: si $\boldsymbol{\alpha} \in \mathbb{R}^n$:

$$\sum_{i,j} \alpha_i \alpha_j (\mathbf{K})_{ij} = \left\langle \sum_{i=1}^n \alpha_i \boldsymbol{\phi}(\mathbf{x}_i), \sum_{j=1}^n \alpha_j \boldsymbol{\phi}(\mathbf{x}_j) \right\rangle = \left\| \sum_{i=1}^n \alpha_i \boldsymbol{\phi}(\mathbf{x}_i) \right\| \geq 0$$

y la igualdad sólo se verifica si $\alpha_i = 0 \; \forall i$ al ser $\langle \cdot, \cdot \rangle$ producto interno en $\mathcal{U}$.

■

El resultado anterior implica que, dados los datos $\{\mathbf{x}_i\}_{i=1}^n$ y una función k, que dan lugar a una matriz $\mathbf{K}$ definida positiva, entonces, es siempre posible construir un espacio de características $\mathcal{U}$ con $\dim(\mathcal{U}) \leq n$. Con ello, aunque k no sea definida positiva, siempre es posible utilizar k en un algoritmo basado en un producto interno definido por una función definida positiva, (en particular, para la resolución del programa cuadrático (II.77)).

US UNIVERSIDAD DE SALAMANCA DE CÓRDOBA

$k(\mathbf{x},\mathbf{x}')$	**Observaciones**
$\cos(x-x')$	
$\cos(x^2-x')$	
$\exp(-(x+x')^p), 0<p\leq 1$	
$\exp(-\left\|\mathbf{x}-\mathbf{x}'\right\|^p), 0<p\leq 2$	Gausiana, Laplaciana,...
$(c^2+\left\|\mathbf{x}-\mathbf{x}'\right\|^2)^{-p}, p>0$	Multicuádrica inversa
$\langle\mathbf{x},\mathbf{x}'\rangle^p, p\in\mathbb{N}$	Polinómico homogéneo
$(\langle\mathbf{x},\mathbf{x}'\rangle+c)^p, p\in\mathbb{N}$	Polinómico no homogéneo
$B_{2p+1}(\left\|\mathbf{x}-\mathbf{x}'\right\|)$ con $p\in\mathbb{N}, B_s=\circledast_{j=1}^{s}1_{[-\frac{1}{2},\frac{1}{2}]}$	B-splines
$\cos(\langle\mathbf{x},\mathbf{x}'\rangle)$	Coseno con producto interno
$\sin[(2p+1)\frac{x}{2}]/\sin\frac{x}{2}, p\in\mathbb{N}$	Dirichlet
$k(B,D)=\mathbb{P}(B\cap D)-\mathbb{P}(B)\mathbb{P}(D)$	En $\mathcal{A}\times\mathcal{A}$ con $(\Omega,\mathcal{A},\mathbb{P})$ esp. prob.

Tabla 7. Diversas funciones definidas positivas que pueden utilizarse como núcleos

Incluso, si $\mathbf{K}$ poseyese algunos autovalores negativos de pequeño valor absoluto, la nueva matriz $\mathbf{K}'=\mathbf{K}+\delta\mathbf{I}_n$ con δ tal que $\mu_{\text{mín}}+\delta\geq 0$ con $\mu_{\text{mín}}$ el menor autovalor de $\mathbf{K}$, bastaría para hacer $\mathbf{K}'$ definida positiva [Schölkopff y Smola 2002, p. 45].

Al margen de lo anterior, son válidas todas las funciones definidas positivas, y, en este sentido, son admisibles las mencionadas en el caso de las redes RRBF, aunque allí la función de transferencia, por construcción, era invariante por traslaciones: $k(\mathbf{x},\mathbf{x}')=\psi(\mathbf{x}-\mathbf{x}')$.

La tabla 7 muestra algunos ejemplos de núcleos utilizados en la literatura, algunos de los cuáles son resultado de métodos de construcción de núcleos descritos en el Anexo B, (por ejemplo los del tipo $k(\mathbf{x},\mathbf{x}')=\xi(\langle\mathbf{x},\mathbf{x}'\rangle)$ donde $\xi:[0,\infty)\to\mathbb{R}$ satisfaciendo algunas condiciones), (véase también [Berg *et al.* 1984, p. 79], [Schölkopff y Smola 2002] y [Cristianini y Shawe-Taylor 2000]).

En este punto, resulta de interés trasladar a este ámbito la relación (II.31) entre producto interno y distancia realizado en la pág. 102. Si k es una función definida positiva, en el espacio de características dicha relación resulta:

$$\left\|\boldsymbol{\phi}(\mathbf{x})-\boldsymbol{\phi}(\mathbf{x}')\right\|^2=k(\mathbf{x},\mathbf{x})+k(\mathbf{x}',\mathbf{x}')-2k(\mathbf{x},\mathbf{x}') \tag{II.78}$$

Además, se verifica el siguiente resultado de [Berg *et al.* 1984, lema p. 74] (teorema 134 Anexo B; también [Schölkopff y Smola 2002, Prop. 2.22]) que afirma que fijado $\mathbf{x}_0\in\mathcal{X}$:

$$k(\mathbf{x},\mathbf{x}')=\frac{1}{2}[k_1(\mathbf{x},\mathbf{x}')-k_1(\mathbf{x},\mathbf{x}_0)-k_1(\mathbf{x}_0,\mathbf{x}')]+k_1(\mathbf{x}_0,\mathbf{x}_0) \tag{II.79}$$

es definida positiva si y sólo si k_1 es condicionalmente definida positiva de orden 1 ([Berg *et al.* 1984, lema p. 74] formula este resultado en términos de una función $k_2=-k_1$ condicionalmente definida negativa[67], condición que define como equivalente a que k_1 sea condicionalmente definida positiva de orden $m=1$).

Sustituyendo k de la ecuación (II.79) en (II.78), se obtiene

$$\left\|\boldsymbol{\phi}(\mathbf{x})-\boldsymbol{\phi}(\mathbf{x}')\right\|^2=-k_1(\mathbf{x},\mathbf{x}')+\frac{1}{2}[k_1(\mathbf{x},\mathbf{x})+k_1(\mathbf{x}',\mathbf{x}')]$$

USC

[67] [Berg *et al.* 1984, lema p. 74] utiliza el término definida negativa.

y además, si[68] $k_1(\mathbf{x},\mathbf{x}) = 0\ \forall \mathbf{x} \in \mathcal{X}$, entonces:

$$\left\|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x}')\right\|^2 = -k_1(\mathbf{x},\mathbf{x}') = k_2(\mathbf{x},\mathbf{x}')$$

Con ello, $\|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x}')\|^2$ puede interpretarse como una medida de disimilitud en términos de una función definida negativa k_2.

Lo anterior permite, a los efectos de modelización, relacionar las propiedades del espacio transformado $\mathcal{U}$ con las de la función c.d.p. de orden 1 (o función definida negativa) elegida.

SVM para Regresión

Utilizando el método del núcleo expuesto en el apartado anterior, podemos plantear ya el problema de regresión en el espacio de características. En dicho espacio, la extensión al problema de regresión de la idea de *soft margin* equivale a permitir soluciones que no se ajusten exactamente a todos los puntos $\{(\boldsymbol{\phi}(\mathbf{x}_i), y_i)\}_{i=1}^n$ de la muestra transformada. Esto puede formularse generalizando el problema (II.77) de la siguiente forma [Drucker *et al.* 1997]:

$$\begin{array}{c} \min\limits_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}'} \{\frac{1}{2}\|\mathbf{w}\|_{\mathcal{U}}^2 + C\sum_{i=1}^n(\xi_i + \xi_i')\} \\ \left\{ \begin{array}{c} \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + b - y_i \le \varepsilon + \xi_i \\ y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + b) \le \varepsilon + \xi_i' \\ \xi_i, \xi_i' \ge 0 \end{array} \right| \ i = 1,\ldots,n \end{array} \qquad \text{(II.80)}$$

donde $\mathbf{w} \in \mathcal{U}$, y $\xi_i, \xi_i' \in \mathbb{R}$ son variables *slack* que evitan que la solución tenga que contener en la banda de radio ε todos los puntos $(\mathbf{x}_i, y_i)$ de la muestra, defendiéndose así de posibles *outliers* y evitando el sobreajuste.

Esta formulación equivale a la utilización de la **función de pérdida ε-insensible** ideada por Vapnik ([Vapnik 1995], [Vapnik 1998]) con $p = 1, 2$:

$$\ell(y, f(\mathbf{x})) = |y - f(\mathbf{x})|_\varepsilon^p = \text{máx}\{0, (|y - f(\mathbf{x})| - \varepsilon)^p\} \qquad \text{(II.81)}$$

que, con $p = 1$, da lugar al problema equivalente:

$$\min_{\mathbf{w},b}\{\frac{1}{2}\|\mathbf{w}\|_{\mathcal{U}}^2 + C\sum_{i=1}^n |y - f_{\mathbf{w},b}(\mathbf{x})|_\varepsilon\} \qquad \text{(II.82)}$$

donde $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + b$. Pero esta formulación es más problemática debido a la función $|\cdot|_\varepsilon$. La opción $p = 2$, equivale a considerar variables *slack* en la función objetivo anterior, pero el argumento que sigue es análogo, (p. ej. [Cristianini y Shawe-Taylor 2000], [Vapnik 1998]).

La solución al problema (II.80) se obtiene formulando el Lagrangiano (Anexo D) con respecto a las variables primales $\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}'$ y duales $\boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\rho}, \boldsymbol{\rho}'$ (en adelante, damos por supuesto el subíndice $\mathcal{U}$ en el producto interno):

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}', \boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\rho}, \boldsymbol{\rho}') \ = \ & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n(\xi_i + \xi_i') + \sum_{i=1}^n \alpha_i[\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b - y_i - \varepsilon - \xi_i] \\ & + \sum_{i=1}^n \alpha_i'[y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b) - \varepsilon - \xi_i'] - \sum_{i=1}^n \rho_i\xi_i - \sum_{i=1}^n \rho_i'\xi_i' \end{aligned}$$

[68] Esta condición no es restrictiva pues si k_1 es una función c.d.p. de orden $m = 1$, entonces $k_1 + b\ \forall b \in \mathbb{R}$ también lo es al ser toda función constante c.d.p. de orden $m = 1$ (Anexo B).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Las condiciones de óptimo para las variables primales son:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n}(\alpha_i' - \alpha_i)\boldsymbol{\phi}(\mathbf{x}_i) = 0 \\
\frac{\partial L}{\partial b} &= -\sum_{i=1}^{n}(\alpha_i' - \alpha_i) = 0 \\
\frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \rho_i = 0,\ i = 1:n \\
\frac{\partial L}{\partial \xi_i'} &= C - \alpha_i' - \rho_i' = 0,\ i = 1:n
\end{aligned}$$

que, sustituyéndolas en el Lagrangiano, producen el siguiente problema dual:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathbb{R}^n} \{-\tfrac{1}{2}\sum_{i,j=1}^{n}(\alpha_i' - \alpha_j)(\alpha_j' - \alpha_j)k(\mathbf{x}_i,\mathbf{x}_j) - \varepsilon\sum_{i=1}^{n}(\alpha_i' + \alpha_i) + \textstyle\sum_{i=1}^{n}(\alpha_i' - \alpha_i)y_i \\
&\qquad \begin{cases} \displaystyle\sum_{i=1}^{n}(\alpha_i' - \alpha_i) = 0 \\ 0 \le \alpha_i, \alpha_i' \le C,\ i = 1:n \end{cases}
\end{aligned} \tag{II.83}$$

con las siguientes condiciones adicionales de óptimo de Kuhn-Tucker, (complementariedad), (Anexo D):

$$\left.\begin{cases} \alpha_i[\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b - y_i - \varepsilon - \xi_i] = 0 \\ \alpha_i'[y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b) - \varepsilon - \xi_i'] = 0 \\ \xi_i\xi_i' = 0,\ \alpha_i\alpha_i' = 0 \\ (C - \alpha_i)\xi_i = 0,\ (C - \alpha_i')\xi_i' = 0 \end{cases}\right| , i = 1:n \tag{II.84}$$

pues $0 = \xi_i\rho_i = \xi_i(C - \alpha_i)$ y $0 = \xi_i'\rho_i' = \xi_i'(C - \alpha_i')$.

Sobre la solución del programa (II.83)-(II.84), destacamos lo siguiente haciendo $\beta_i = \alpha_i' - \alpha_i$, (el problema de clasificación en sus versiones *hard* y *soft margin* admite observaciones similares):

1. El óptimo $\mathbf{w}$ posee la forma: $\mathbf{w} = \sum_{i=1}^{n}\beta_i\boldsymbol{\phi}_i$ con $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\mathbf{x}_i)$, $i = 1:n$, con lo que la solución se expresa en términos del producto interno con los puntos de la muestra en el espacio de características.

2. Los vectores que satisfacen $\beta_i = \alpha_i' - \alpha_i \neq 0$, ($\alpha_i > 0$ o $\alpha_i' > 0$), es decir, los que saturan las dos primeras restricciones en (II.84), son los únicos que participan en la expresión de la solución. Por ello, se denominan **vectores soporte** (*support vectors*). De ellos, los que verifican $0 < \alpha_i < C$ o $0 < \alpha_i' < C$ están ubicados justo en la frontera de la banda de radio ε, y el resto, $\alpha_i = C$ o $\alpha_i' = C$, quedan fuera de esa banda contabilizados como errores por la función de pérdida.

3. Con ello, el óptimo es combinación lineal de los vectores soporte $\mathbf{w} = \sum_{\text{S.V.}}\beta_i\boldsymbol{\phi}_i$ y la solución al problema (II.83) viene dada por:

$$f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x})\rangle + b = \sum_{\text{S.V.}}\beta_i\,\langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x})\rangle_{\mathcal{U}} + b = \sum_{\text{S.V.}}\beta_i k(\mathbf{x}_i,\mathbf{x}) + b \tag{II.85}$$

que es la forma general de las redes SVM para regresión.

Con todo ello, la resolución del problema primal mediante la formulación dual permite reducir el número de términos de la solución al basarse ésta en las restricciones activas del primal

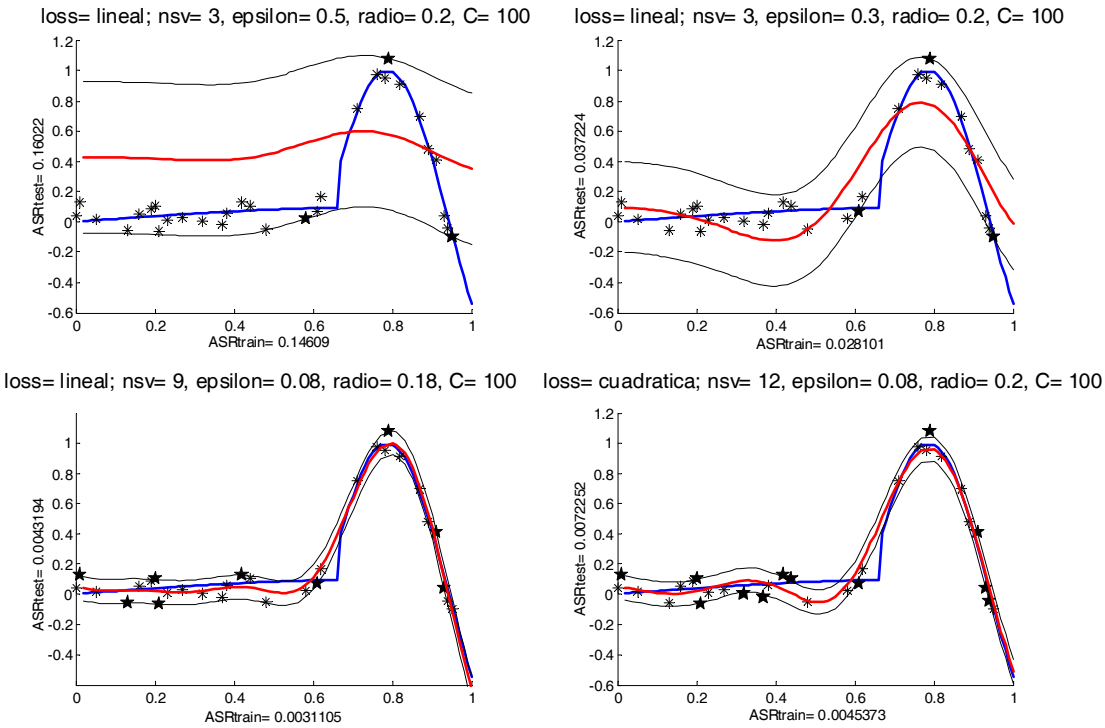


Figura 21. De izquierda a derecha y de arriba a abajo: redes SVM entrenadas con pérdida ε-insensible lineal para $\varepsilon = 0{,}5$, $0{,}3$ y $0{,}08$, respectivamente, y red SVM entrenada con pérdida ε-insensible cuadrática con $\varepsilon = 0{,}08$ y demás parámetros iguales a la tercera de las anteriores.

(correspondientes a los vectores realmente relevantes o *difíciles* para el problema), en lugar de utilizar tantos términos como datos en la muestra.

De esta forma, la solución obtenida depende únicamente de la dificultad intrínseca del problema a resolver y no, por ejemplo, del tamaño de la muestra o de la dimensión del espacio de entrada. Una ventaja de lo anterior, es que la solución sería la misma si se eliminan de la muestra todos los puntos que no resultan relevantes, lo que permite adoptar tácticas de entrenamiento progresivo añadiendo datos sucesivamente y conservando únicamente los vectores soporte de las etapas anteriores.

La figura 21 muestra el ajuste de cuatro redes SVM gausianas en un problema de regresión. En cada subfigura puede observarse la banda $\pm\varepsilon$ centrada en la estimación proporcionada por cada red, así como los vectores soporte de la misma, identificados con estrellas.

Puede observarse la influencia del parámetro ε en la complejidad del estimador final: las tres primeras redes se entrenaron utilizando la pérdida ε-insensible lineal con $\varepsilon = 0{,}5$, $0{,}3$ y $0{,}08$ (este último fué el valor óptimo sobre una muestra de validación), obteniéndose respectivamente 3, 3 y 9 vectores soporte, lo que muestra la influencia inversa de ε sobre dicho número.

La última red se entrenó con pérdida ε-insensible cuadrática y los demás parámetros iguales a la última de las anteriores ($\varepsilon = 0{,}08$), obteniéndose 12 vectores soporte. Por tanto, esta red resultó menos parsimoniosa que la correspondiente obtenida con pérdida ε-insensible lineal, lo que suele ser el caso en general.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

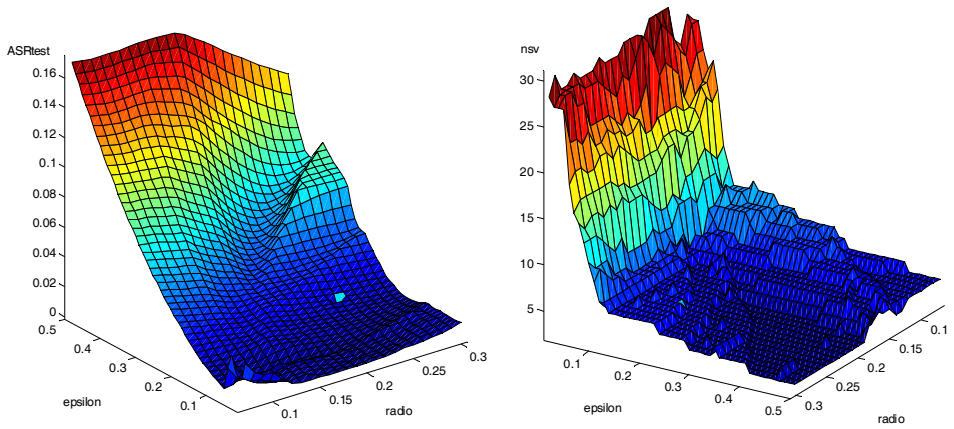


Figura 22. Izquierda: error en una muestra de validación en función del parámetro ε y del parámetro de escala del núcleo. Los valores extremos, especialmente de ε, empeoran la capacidad de generalización. Derecha: número de vectores soporte de la solución en función de dichos parámetros. Ahora, es sobre todo el incremento de ε el que provoca el incremento de dicho valor. (Nótese que la escala de los ejes del plano horizontal de la segunda figura, está invertida con respecto a la primera (para mejor visualización).

Finalmente, la figura 22 muestra la influencia de los hiperparámetros ε y σ (escala del núcleo), sobre el error en una muestra de validación y el número de vectores soporte de la solución.

Al respecto del gráfico de la izquierda, la zona de mayores errores es la zona de infraajuste de la red que se corresponde con valores grandes de ε. La zona de sobreajuste no se muestra debido a su rápido crecimiento y se obtiene para valores muy pequeños de ambos parámetros, sobre todo de ε. En un rango de valores moderados de ambos parámetros, el parámetro ε posee mucha mayor influencia en el grado de complejidad de la red.

El gráfico de la derecha de la figura 22 (nótese que la escala de los ejes del plano horizontal está invertida con respecto al gráfico de la derecha), muestra como los valores pequeños de ε producen un gran incremento en el número de vectores soporte debido al estrechamiento de la banda $\pm\varepsilon$. Si este parámetro es suficientemente grande, no habría vectores soporte.

ν-SVM

La formulación de las SVM para regresión admiten, como en el caso de clasificación, una formulación alternativa que, en determinados escenarios puede ser más fácil de interpretar en la modelización.

En lo anterior, el parámetro ε representa el ancho de la banda de tolerancia de la solución. Este parámetro es útil cuando existe información a priori sobre dicho valor. Sin embargo, cuando éste no es el caso, puede ser más interesante la formulación del problema en términos de la máxima fracción de datos que se admite fuera de la banda [Schölkopf, Smola, Williamson y Bartlett 2000]

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

de la siguiente forma:

$$\begin{min}_{\mathbf{w},\mathbf{d},\boldsymbol{\xi},\boldsymbol{\xi}',\varepsilon} \{\tfrac{1}{2}\|\mathbf{w}\|_{\mathcal{U}}^2 + C[\nu\varepsilon + \textstyle\sum_{i=1}^n(\xi_i+\xi_i')]\} \\ \left\{ \begin{array}{c} \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + b - y_i \leq \varepsilon + \xi_i \\ y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + b) \leq \varepsilon + \xi_i' \\ \xi_i, \xi_i' \geq 0, \varepsilon \geq 0 \end{array} \right| \; i = 1, \ldots, n$$

donde $\nu \geq 0$ es una cota superior para la fracción de errores y una cota inferior para la fracción de vectores soporte [Schölkopf, Smola, Williamson y Bartlett 2000, prop. 1].

Como puede observarse, la complejidad de la solución sigue representándose mediante el término $\frac{1}{2}\|\mathbf{w}\|_{\mathcal{U}}^2$, pero, ahora, el término (sesgo) que representa la calidad del ajuste, incluye el parámetro ε, que se penaliza al igual que las variables *slack*. Como resultado, ε se obtiene como parte de la solución dependiendo del parámetro ν. Al igual que antes, el problema se resuelve a través del programa dual.

SVM como Solución a un Problema de Regularización en un RKHS

En el espacio de características, el cuadrado de la norma de un vector $\mathbf{w} = \sum_{i=1}^n \alpha_i \boldsymbol{\phi}(\mathbf{x}_i)$ combinación lineal de n puntos $\{\boldsymbol{\phi}(\mathbf{x}_i)\}_{i=1}^n \in \mathcal{U}$, es:

$$\|\mathbf{w}\|_{\mathcal{U}}^2 = \langle \mathbf{w}, \mathbf{w}\rangle_{\mathcal{U}} = \sum_{i,j=1}^n \alpha_i\alpha_j k(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$$

donde $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^T$. Pero, $\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$ es también la norma $\left\|f_{\boldsymbol{\alpha}(\mathbf{w})}\right\|_{\mathcal{H}_k}^2$ de una función $f_{\boldsymbol{\alpha}(\mathbf{w})} = \sum_{i=1}^n \alpha_i k(\cdot, \mathbf{x}_i) \in \mathcal{H}_k$ siendo $\mathcal{H}_k$ el RKHS asociado a k (donde n puede ser ∞ –entendido como límite). Si consideramos por un momento, como hiperplanos factibles, sólo los que pasan por el origen, ($b = 0$): $f_{\mathbf{w}} = \sum_{i=1}^n \alpha_i k(\cdot, \mathbf{x}_i) = f_{\boldsymbol{\alpha}(\mathbf{w})}$, el problema (II.82) puede escribirse como [Evgeniou *et al.* 2000]:

$$\begin{min}_{\mathbf{w}}\{\frac{1}{2}\|\mathbf{w}\|_{\mathcal{U}}^2 + C\sum_{i=1}^n |y - f_{\mathbf{w}}(\mathbf{x})|_{\varepsilon}\} \equiv \begin{min}_{f\in\mathcal{H}_k}\{\sum_{i=1}^n \left|y - f_{\boldsymbol{\alpha}(\mathbf{w})}(\mathbf{x})\right|_{\varepsilon} + \frac{1}{2C}\left\|f_{\boldsymbol{\alpha}(\mathbf{w})}\right\|_{\mathcal{H}_k}^2\}$$

y, así, la SVM con $b = 0$ puede verse como la solución a un problema de regularización en el RKHS $\mathcal{H}_k$ de núcleo k, con parámetro regularizador $\lambda = \frac{1}{2C}$. De hecho, por el teorema 38, su solución es una combinación lineal finita de funciones $k_{\mathbf{x}_i}$, lo que concuerda con (II.85) con $b = 0$.

Lo anterior es válido asimismo para la SVM con término constante pues, la incorporación de dicho término con un núcleo definido positivo puede realizarse argumentando de igual manera que se hizo al tratar las redes RRBF (pág. 133) (Véase también el corolario 133 del Anexo B).

Redes RRBF y SVM para Regresión

El análisis anterior permite clarificar las relaciones entre las redes SVM para regresión y las redes RRBF solución también de un problema de regularización en un RKHS, ya que, el problema de regularización que da lugar a las RRBF:

$$\begin{min}_{f\in\mathcal{F}_{\tilde{\psi}}} \hat{\mathcal{R}}_n(f) = \sum_{i=1}^n \ell(y_i, f(\mathbf{x}_i)) + \lambda\|f\|_{\mathcal{H}_k}^2$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde $\ell(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$ es la pérdida cuadrática, equivale al siguiente programa cuadrático con restricciones lineales en un espacio de características $\mathcal{U}$ utilizando la metodología SVM [Evgeniou *et al.* 2000] (téngase en cuenta que $\xi_i = y_i - f(\mathbf{x}_i)$):

$$\begin{gathered}\text{mín}\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i^2\} \equiv \text{mín}\{\sum_{i=1}^{n}\xi_i^2 + \frac{1}{2C}\|\mathbf{w}\|^2\} \\ y_i - \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle - b = \xi_i,\ i = 1, \ldots, n\end{gathered}$$

siendo $\lambda = \frac{1}{2C}$ el parámetro regularizador.

La solución es, como en el caso de las SVM, una combinación lineal de los puntos de la muestra en el espacio de características (Anexo G), $\mathbf{w} = \sum_{i=1}^{n}\beta_i\boldsymbol{\phi}(\mathbf{x}_i)$ con lo que:

$$f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x})\rangle + b = \sum_{i=1}^{n}\beta_i k(\mathbf{x}_i, \mathbf{x}) + b$$

Sin embargo, con ello se pierde la parsimonia propia de las SVM basada en la utilización de los puntos verdaderamente relevantes para el problema: los vectores soporte. A diferencia de la pérdida cuadrática, muy sensible a los posibles datos atípicos de la muestra, la pérdida ε-insensible dota a las SVM de un mecanismo de gestión de atípicos a través del parámetro ε que refleja el grado de fiabilidad que los datos poseen para el experimentador.

II.5.4 Consistencia de las Arquitecturas de Tipo Radial

Propiedades Aproximadoras

La tabla 8 recoge los resultados más relevantes sobre las propiedades aproximadoras de las arquitecturas de tipo radial (que incluye modelos no específicamente radiales). En particular, el resultado antepenúltimo de la tabla se refiere a las redes RBF de regularización y los dos últimos a las redes SVM para regresión y clasificación, respectivamente. Los detalles de estos resultados pueden verse en el anexo C.

Al respecto, es importante destacar la denominada propiedad de la mejor aproximación de las arquitecturas RBF con funciones básicas prefijadas (teorema 170, Anexo C), que no cumplen las demás arquitecturas radiales ni las redes MLP (teorema 169, Anexo C) al concretarse sus funciones básicas mediante algoritmos de optimización no lineal.

Consistencia

Redes RBF

De esta forma, adoptando como riesgo $R(f) = \|\mathsf{Y} - f(\mathsf{X})\|_{L_p(\mathsf{P_Z})}$, el error de aproximación de la descomposición dada por la ecuación (I.49), puede hacerse tan pequeño como se quiera, cumpliéndose la primera condición para la consistencia de las ecuaciones (I.50) (pág. 61).

Un riesgo de este tipo (p. ej. con pérdida cuadrática) se utiliza en las redes RBF presentadas en primer lugar en esta sección, con lo que suponiendo que el algoritmo utilizado proporciona el óptimo del riesgo empírico, se obtiene la consistencia.

Propiedades Aproximadoras de la red RBF: $f(\mathbf{x}) = \sum_{j=1}^{h} c_j k_j(\mathbf{x}, \mathbf{w}_j)$			
Espacio Funcional	$\|\cdot\|$	**Condiciones**	**Enfoque**
$\mathcal{C}(\mathcal{X})$, $\mathcal{X}$ abierto, acotado [Jackson 1988]	L_1	$k_i(\mathbf{x}, \mathbf{w}_i) = \zeta(\|\mathbf{x} - \mathbf{w}_i\|)$ $\zeta(az) = a^\beta \zeta(z)$ $a, z \in \mathbb{R}^+, \beta \in \mathbb{R}$ $\{\mathbf{w}_i\}_{i=1}^{\infty}$ denso en $\mathrm{cl}(\mathcal{X})$	Existencial
$\mathcal{C}(\mathcal{X})$, $\mathcal{X}$ compacto y convexo [Hartman *et al.* 1990]	L_∞	$k_i(\mathbf{x}, \mathbf{w}_i) = e^{-a_i \|\mathbf{x} - \mathbf{x}_i\|^2}$	Existencial Stone-Weieres.
$\mathcal{C}(\mathbb{R}^d)$ [Park y Sandberg 1991]	L_∞	$k_j(\mathbf{x}, \mathbf{w}_i) = \psi(\mathbf{x} - \mathbf{w}_i)$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ ψ acotada, continua c.p.d	Existencial
$\mathcal{C}(\mathcal{X})$, $\mathcal{X}$ compacto [Park y Sandberg 1993]	L_∞	$k_j(\mathbf{x}, \mathbf{w}_i) = \psi(\frac{\mathbf{x} - \mathbf{w}_j}{\sigma_j})$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ ψ continua	Existencial
$L_p(\mathbb{R}^d)$ y $L_p^{\mathrm{loc}}(\mathbb{R}^d)$, $p \in [1, \infty)$ [Park y Sandberg 1991]	L_p	$k_i(\mathbf{x}, \mathbf{w}_i) = \psi(\mathbf{x} - \mathbf{w}_i), \forall i$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ ψ acotada, continua c.p.d	Existencial Convolución
$L_p(\mathbb{R}^d)$, $p \in [1, \infty)$ [Park y Sandberg 1993]	L_p	$k_i(\mathbf{x}, \mathbf{x}_i) = \psi(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma_i})$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ $\int_{\mathbb{R}^d} \|\psi(\mathbf{x})\|^p d\mathbf{x} < \infty$	Existencial Convolución
$L_p(\mathbb{R}^d, \mu)$, $p \in (0, \infty)$ μ medida de probabilidad [Krzyzak *et al.* 1996]	L_p	$k_i(\mathbf{x}, \mathbf{x}_i) = \psi(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma_i})$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ $\psi \in L_1(\mathbb{R}^d) \cap L_p(\mathbb{R}^d)$	Existencial Convolución
$\mathcal{T} = \{f = \psi * \lambda : \lambda \in L^1(\mathbb{R}^d)\}$ [Girosi y Anzellotti 1993]	L_∞	$\psi(\mathbf{x}) = e^{-\|\mathbf{x}\|^2}$ $k_i(\mathbf{x}, \mathbf{w}_i) = \psi(\mathbf{x} - \mathbf{w}_i)$	Lema Maurey
$\mathcal{W}_1^{2r}(\mathbb{R}^d), 2r > d$ [Girosi y Anzellotti 1993]	L_∞	$k_i(\mathbf{x}, \mathbf{w}_i) = g_r(\|\mathbf{x} - \mathbf{w}_i\|^2)$ $g_r = \tilde{g}$ potencial de Bessel i.e. $g(\boldsymbol{\omega}) = (1 + \|\boldsymbol{\omega}\|^2)^{-r/2}$	Lema Maurey
$\mathcal{W}_1^{2r}(\mathbb{R}^d), 2r > d$ [Girosi 1994]	L_2	$k_i(\mathbf{x}, \mathbf{w}_i) = e^{-\|\mathbf{x} - \mathbf{w}_i\|^2 / \sigma_i^2}$	Lema Maurey
$\mathcal{C}(\mathcal{X}), \mathcal{X} \subset \mathbb{R}^d$ compacto [Girosi y Poggio 1990]	L_∞	$\min_{\mathcal{H}_P} \{\hat{R}_n(f) + \lambda \|Pf\|_{L_2}^2\}$ $k_i(\mathbf{x}, \mathbf{w}_i) = k(\mathbf{x}, \mathbf{x}_i)$ k función de Green de P^+P k continua a trozos	Existencial Propieades de la F. Green
$\mathcal{C}(\mathcal{X}), \mathcal{X} \subset \mathbb{R}^d$ compacto [Hammer y Gersmann 2003]	L_∞	1) $k(\mathbf{x}, \mathbf{x}') = \psi(\frac{1}{\sigma}\mathbf{x} - \mathbf{x}')$ con: $\int \psi(\mathbf{x}) d\mathbf{x} \neq 0$ y k acotada y continua 2) $k(\mathbf{x}, \mathbf{x}') = \kappa(\langle \mathbf{x}, \mathbf{x}' \rangle + \alpha)$ κ analítica, no polinomio continua y $\exists j : \kappa^{(j)}(\alpha) \neq 0$	1) Aprox. RBF 2) Aprox. MLP
$\mathcal{M}(\mathcal{X}, \mathbb{P}_{\mathsf{X}}) = \{\mathsf{g}\ \mathbb{P}_{\mathsf{X}}\text{-medibles}\}$ con $\mathsf{g} : \mathbb{R}^d \to \{-1, 1\}$ $\mathcal{X} \subset \mathbb{R}^d$ compacto [Hammer y Gersmann 2003]	$\mathbb{P}_{\mathsf{X}}[g \neq \mathsf{g}]$	1) $k(\mathbf{x}, \mathbf{x}') = \psi(\frac{1}{\sigma}\mathbf{x} - \mathbf{x}')$ con: $\int \psi(\mathbf{x}) d\mathbf{x} \neq 0$ 2) $k(\mathbf{x}, \mathbf{x}') = \kappa(\langle \mathbf{x}, \mathbf{x}' \rangle + \alpha)$ κ analítica y no polinomio y $\exists j : \kappa^{(j)}(\alpha) \neq 0$	1) Aprox. RBF 2) Aprox. MLP

Tabla 8. Propiedades aproximadoras más relevantes de las redes RBF.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Resultados de este tipo son los que mostramos a continuación que, al igual que el presentado para las arquitecturas MLP, utiliza el método SRM para lograr el control de la complejidad.

Por tanto, sea el riesgo $R(f) = \|\mathsf{Y} - f(\mathsf{X})\|_{L_p(\mathsf{P_Z})}$ para algún $p \in [1, \infty)$ y su correspondiente riesgo empírico en una muestra aleatoria simple:

$$\hat{R}_n(f) = \frac{1}{n}[\sum_{i=1}^{n} |y_i - f(\mathbf{x}_i)|^p]^{1/p}$$

Sean las familias $\mathcal{F}_n$ cuyas funciones son combinaciones lineales de un conjunto de funciones básicas prefijadas $\{\psi_j\}$, es decir, **modelos lineales en parámetros**:

$$\mathcal{F}_n = \{f = \sum_{i=1}^{h_n} c_i \psi_i : \mathbf{c} \in \mathbb{R}^{h(n)}, \sum_{i=1}^{h_n} |c_i| \leq \beta_n\}$$

y sea $\mathcal{F} = \{f = \sum_{i=1}^{h} c_i \psi_i : \mathbf{c} \in \mathbb{R}^h, h \in \mathbb{N}^+\}$.

Se tiene el siguiente teorema válido para cualquier modelo de regresión lineal generalizada:

Teorema 47 *[Lugosi y Zeger 1995, th. 2] Sea $p \in [1, \infty)$ con R y $\hat{R}_n$ definidos anteriormente. Sea $\{\psi_j\}$ una sucesión de funciones acotadas uniformemente y sean las familias $\mathcal{F}$ y $\mathcal{F}_n$ definidas anteriormente para $n = 1, 2, ...$ tal que $\mathcal{F}$ es densa en $L_p(\mathsf{P_X})$. Si cuando $n \to \infty$, h_n y β_n satisfacen:*

$$h_n \to \infty, \beta_n \to \infty \; y \; \frac{h_n \beta_n \ln(\beta_n)}{n} \to 0$$

entonces, $\forall \mathsf{P_Z}$ con $\mathbb{E}_{\mathsf{Z}}(|\mathsf{Y}|^p) < \infty$, $\{\hat{f}_n\}$ es consistente, es decir:

$$\text{plim}[R(\hat{f}_n) - R(\mathsf{f})] = 0$$

Si, además, se satisface que $\beta_n = o(n^{1-\delta})$ para algún $\delta > 0$, entonces $R(\hat{f}_n) - R(\mathsf{f}) \to 0$ c.s., es decir $\{\hat{f}_n\}$ es fuertemente consistente $\forall \mathsf{P_Z}$ con $\mathbb{E}_{\mathsf{Z}}(|\mathsf{Y}|^p) < \infty$.

El siguiente teorema es válido para arquitecturas radiales **no lineales en parámetros** (adaptables por ejemplo mediante algoritmos de optimización lineal en una sóla fase).

Supongamos que las familias $\mathcal{F}_n$ son de la forma:

$$\begin{aligned} \mathcal{F}_n &= \{f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=1}^{h_n} c_i \psi([\mathbf{x} - \mathbf{w}_i]^T \Sigma_i [\mathbf{x} - \mathbf{w}_i]) + c_0, \; \boldsymbol{\theta} \in \Theta_h, \; h_n \in \mathbb{N}^+\} \\ \text{con} \quad : \quad \Theta_h &= \{\boldsymbol{\theta} : \sum_{i=0}^{h_n} |c_i| \leq \beta_n\} \end{aligned}$$

donde $\boldsymbol{\theta} = \{c_0, c_1, ..., c_{h_n}, \mathbf{w}_1, ..., \mathbf{w}_{h_n}, \Sigma_1, ..., \Sigma_{h_n}\}$ con $\Sigma_i \in \mathbb{R}^{d^2}$ matriz definida positiva, $\mathbf{w}_i \in \mathbb{R}^d$, $c_i \in \mathbb{R}^d, i = 1, ..., h_n$, $c_0 \in \mathbb{R}$ y $\psi : [0, \infty) \to \mathbb{R}$.

Si se supone sin pérdida de generalidad que $|\psi(r)| \leq 1 \; \forall r \geq 0$, la restricción $\sum_{i=0}^{h_n} |c_i| \leq b_n$ y la desigualdad de Cauchy-Schwartz producen:

$$|f(\mathbf{x}, \boldsymbol{\theta})|^2 = \left| \sum_{i=1}^{h_n} c_i \psi([\mathbf{x} - \mathbf{w}_i]^T \Sigma_i [\mathbf{x} - \mathbf{w}_i]) + c_0 \right|^2 \leq \beta_n (h_n + 1)$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

con lo que las funciones están acotadas.

El siguiente teorema se verifica para la pérdida cuadrática, es decir, $R(f) = \mathbb{E}_{\mathsf{Z}}[(\mathsf{Y} - f(\mathsf{X}))^2]$ y $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2$.

Teorema 48 *[Krzyzak* et al. *1996, th. 2] Sea ψ una función integrable Lebesgue, monótona creciente, continua por la izquierda y acotada. Si cuando $n \to \infty$, h_n y β_n satisfacen:*

$$h_n \to \infty, b_n \to \infty \; y \; \frac{h_n^3 \beta_n^2 \ln(h_n^3 \beta_n^2)}{n} \to 0$$

entonces, $\forall \mathsf{P}_{\mathsf{Z}}$ con $\mathbb{E}_{\mathsf{Z}}(|\mathsf{Y}|^p) < \infty$, $\{\hat{f}_n\}$ es consistente, es decir:

$$\operatorname{plim}[R(\hat{f}_n) - R(\mathsf{f})] = 0$$

Si, además, $\exists \delta > 0$ tal que $h_n/n^{\delta} \to \infty$ cuando $n \to \infty$, entonces $R(\hat{f}_n) - R(\mathsf{f}) \to 0$ c.s., es decir $\{\hat{f}_n\}$ es fuertemente consistente $\forall \mathsf{P}_{\mathsf{Z}}$ con $\mathbb{E}_{\mathsf{Z}}(|\mathsf{Y}|^p) < \infty$.

En la demostración del teorema anterior el error de aproximación tiende a cero gracias a las propiedades aproximadoras universales de las redes RBF en $L_2(\mathsf{P}_{\mathsf{X}})$ ([Krzyzak *et al.* 1996, th. 1], teorema 164 y tabla 42 del anexo C).

Otros resultados al respecto (p. ej. [Niyogi y Girosi 1996, th. 4.1]) utilizan un algoritmo concreto (*greedy*) que permite obtener tasas de convergencia, y que presentaremos al tratar dicho algoritmo.

Redes RRBF y SVM

Para las redes RBF de regularización y las Redes SVM en el problema de regresión, ambas solución de un problema de regularización en un RKHS (pág. 151), es aplicable el método de la estabilidad algorítmica de [Bousquet y Elisseeff 2002] expuesto en el capítulo anterior (teorema 25).

Utilizando una pérdida γ-admisible y acotada en $[0, C]$, el teorema 23 producía con probabilidad $1 - \eta$ y $\forall n \geq 1$:

$$\begin{aligned} &1. \quad R(A_n) \leq \hat{R}_n(A_n) + 2\beta + (4n\beta + C)\sqrt{\tfrac{\ln 1/\eta)}{2n}} \\ &2. \quad R(A_n) \leq \mathrm{LOO}(A_n) + \beta + (4n\beta + C)\sqrt{\tfrac{\ln 1/\eta}{2n}}) \end{aligned} \tag{II.86}$$

para un algoritmo A con estabilidad uniforme β. Por otra parte, el teorema 25 acotaba la estabilidad uniforme para la regresión regularizada en un RKHS:

$$\beta \leq \frac{\gamma^2 \kappa^2}{2\lambda n} \tag{II.87}$$

donde $\forall \mathbf{x} \in \mathcal{X}$, $k(\mathbf{x}, \mathbf{x}) \leq \kappa^2 < \infty$ y γ era la constante de la pérdida γ-admisible.

Combinando los resultados (I.52) y (I.53), se obtiene, por ejemplo para el error de estimación utilizando el riesgo empírico, (el argumento es análogo si se utiliza $\mathrm{LOO}(A_n)$):

$$R(A_n) - \hat{R}_n(A_n) \leq \frac{2\gamma^2\kappa^2}{\lambda n} + (\frac{2\gamma^2\kappa^2}{\lambda} + C)\sqrt{\frac{\ln 1/\eta}{2n}} \underset{n\to\infty}{\to} 0 \tag{II.88}$$

siempre que se elija $\lambda = \lambda(n)$ tal que $\lim \sqrt{n}\lambda(n) = \infty$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Esta propiedad, combinada con las propiedades aproximadoras de las redes RRBF y SVM, permite probar la consistencia del proceso de aprendizaje con estas redes. Para ello, sólo falta demostrar que las correspondientes funciones de pérdida son γ-admisibles y están acotadas. Ello se consigue imponiendo que $\mathcal{Y}$ está acotado y utilizando el lema 26. Por ejemplo, suponiendo $\mathcal{Y} = [0, B]$, se tienen los siguientes resultados:

1. Redes RRBF. Para estas redes se utiliza la pérdida cuadrática. Basta tomar $\xi(\hat{y}) = B^2$ constante en el lema 26 para obtener, en el marco de un problema de regularización en un RKHS, que:

   $$\forall \mathbf{z} = (\mathbf{x}, y) \in \mathcal{Z} : 0 \leq \ell(y, \hat{f}_n(\mathbf{x})) \leq \xi(\kappa\sqrt{\xi(0)/\lambda}) = B^2$$

   y que dicha pérdida es γ-admisible con $\gamma = 2B$. Con ello, con el algoritmo posee estabilidad:

   $$\beta \leq \frac{2B^2\kappa^2}{\lambda n}$$

   Combinando los resultados (I.52) y (I.53) para este caso, con $C = B^2$, se obtiene, por ejemplo para el error de estimación utilizando el riesgo empírico (el argumento es análogo si se utiliza LOO(A_n)) :

   $$R(A_n) - \hat{R}_n(A_n) \leq \frac{4\kappa^2 B^2}{\lambda n} + (\frac{8\kappa^2 B^2}{\lambda} + B^2)\sqrt{\frac{\ln 1/\eta}{2n}} \underset{n\to\infty}{\to} 0$$

   siempre que se elija $\lambda = \lambda(n)$ tal que $\lim \sqrt{n}\lambda(n) = \infty$.

2. SVM para regresión. Basta tomar $\xi(\hat{y}) = B$ constante en el lema 26 para obtener que la pérdida ε-insensible de Vapnik (expresión (II.81)) es γ-admisible con $\gamma = 1$ y que $\ell(y, \hat{f}_n(\mathbf{x})) \leq B\ \forall \mathbf{z} \in \mathcal{Z} : 0$.

   Con ello, el algoritmo SVM para regresión posee estabilidad $\beta \leq \kappa^2/(2\lambda n)$ y, de la misma forma que antes, para este caso se obtiene con $C = B$:

   $$R(A_n) - \hat{R}_n(A_n) \leq \frac{4\kappa^2}{\lambda n} + (\frac{2\kappa^2}{\lambda} + B)\sqrt{\frac{\ln 1/\eta}{2n}} \underset{n\to\infty}{\to} 0$$

Un resultado similar se verifica para la red SVM para clasificación (*soft margin*) [Bousquet y Elisseeff 2002, th. 17 y lema 20]. Ahora, la función de pérdida (pág. 142) es $\ell(y, f(\mathbf{x}) = [1 - yf(\mathbf{x})]_+$ que es γ-admisible con $\gamma = 1$ (ahora $\mathcal{Y} = \{-1, 1\}$).

Para el caso de las redes SVM para clasificación *hard margin*, puede aplicarse el teorema 29 del capítulo anterior válido también para funciones indicadoras (pág. 62) sobre la consistencia del principio SRM que, como vimos, se utiliza implícitamente con los hiperplanos de máximo margen (pág. 141). Otros resultados sobre la consistencia de las SVM son: [Steinwart 2001], [Steinwart 2002*a*] y [Steinwart 2002*b*].

II.6 Algoritmos para redes MLF

En esta sección se presentan los algoritmos utilizados para el entrenamiento de las redes *multilayer feed-forward* (MLF) cuyas diferentes arquitecturas se motivaron en las secciones anteriores. Estos algoritmos pueden clasificarse en los siguientes grandes grupos:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1. **Algoritmos de optimización no lineal sin restricciones**. Estos algoritmos utilizan la información contenida en las derivadas de la función objetivo a minimizar, y se utilizan fundamentalmente con redes no lineales en parámetros, como son:
   (a) Las redes que implementan proyecciones en el espacio de entrada, como las redes MLP y las redes wavelet multiescala, con la excepción de las redes projection pursuit, cuyo número de neuronas ocultas no está fijado de antemano.
   (b) Las redes RBF que no parten de un conjunto prefijado de funciones de transferencia sino que, éstas, dependen de un conjunto de parámetros de escala y posición que, así, influyen de forma no lineal en aquéllas.

   En este conjunto de algoritmos, incluimos los que utilizan el enfoque *back-propagation* para un cálculo eficiente de las derivadas, aunque este término se ha extendido y suele utilizarse para referirse a todo este grupo.

2. **Algoritmos SVM**. Un caso especial, es la red SVM, que posee carácter no lineal en los parámetros de posición (centros) al ser resultado de la resolución de un programa cuadrático con restricciones, y, por tanto, se entrena mediante algoritmos específicos de optimización con restricciones y estrategias específicas para este tipo de arquitecturas.

3. **Algoritmos de Regresión Lineal Generalizada**. Estos algoritmos se utilizan con arquitecturas lineales en parámetros, para lo que se requiere que las funciones de transferencia del nivel oculto estén prefijadas. En este caso, los únicos parámetros libres son los coeficientes del nivel de salida. En estas situaciones, suele utilizarse la pérdida cuadrática con objeto de dar lugar a un problema de regresión generalizado que se resuelve de forma estándar.

   Por tanto, en general, estos algoritmos son aplicables a las redes RBF (y a las redes wavelet monoescala).

   Un caso particular de lo anterior es cuando el número de parámetros coincide con el número de datos, en cuyo caso se tiene un algoritmo procedente del ámbito de la interpolación de funciones, cuya solución ya hemos obtenido al presentar las redes RRBF (pág. 135). Por tanto, lo que queda por tratar de estos algoritmos son métodos numéricos de cálculo matricial y obtención de inversas, que no son objeto de este trabajo (p. ej. [Press *et al.* 1992]), y algún método de selección del parámetro regularizador además de los genéricos ya presentados en el capítulo primero.

4. **Algoritmos** *Greedy*, utilizado casi exclusivamente con las redes projection pursuit aunque, como veremos, muy empleado a nivel teórico para obtener tasas de consistencia con las distintas arquitecturas. Este algoritmo surge realmente en el ámbito estadístico aunque se incorpora muy rápidamente al mundo de las redes neuronales al ser de los pocos asequibles para obtener tasas de convergencia.

   En realidad, más que un algoritmo, el algoritmo *greedy* constituye una batería de algoritmos cuya característica común es el entrenamiento progresivo de nuevas neuronas ocultas a partir de los residuos resultantes de la inclusión de las anteriores, y que se diferencian en el *subalgoritmo* utilizado para el entrenamiento de cada una de estas neuronas ocultas.

Asimismo, a la vista de las relaciones existentes entre los tres últimos grupos de algoritmos y diversas técnicas de proceso de señales, se utilizan estas últimas para su utilización con arquitecturas radiales considerando parámetros de escala diferentes y regularización con la norma absoluta.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Además de los grupos anteriores, es necesario mencionar por su importancia, otros tipos de algoritmo que no serán tratados aquí por razones de espacio, aunque también son menos utilizados en comparación con los anteriores: por un lado, los algoritmos de optimización global como los algoritmos genéticos (p. ej. [Goldberg 1989]) y los algoritmos de optimización combinatoria (*Simulated Annealing* [Kirkpatrick *et al.* 1983], y p. ej. [Press *et al.* 1992]) utilizados fundamentalmente en escenarios bayesianos o como algoritmo de entrenamiento de redes no lineales, por otro, el algoritmo E-M (*Expectation-Minimization*) utilizado para la estimación por la máxima verosimilitud sobre todo con redes RBF con entrenamiento en dos fases.

En general, mientras que con una arquitectura de proyección, por su naturaleza, sólo pueden utilizarse algoritmos de optimización no lineal y algoritmos *greedy*, con las arquitectura radiales pueden utilizarse estos dos tipos de algoritmo además de los algoritmos diseñados específicamente para ellas. Sin embargo, en ese caso, las redes neuronales resultantes poseerán propiedades diferentes a las tratadas anteriormente en cada uno de los casos.

En lo que sigue, no se pretende realizar una exposición ni mucho menos exhaustiva de cada uno de los algoritmos, pues la literatura al respecto en algunos casos es ingente. Únicamente se persigue resaltar los aspectos más importantes, incidiendo en lo posible en su vertiente estadística.

II.6.1 Algoritmos *Back-Propagation*

El enfoque de **propagación hacia atrás de los errores** (*error back-propagation*) propuesto por [Rumelhart *et al.* 1986], tiene por objeto el cálculo de las derivadas de la función objetivo con respecto a los parámetros de una forma eficiente. Dichas derivadas pueden ser utilizadas en cualquier algoritmo de optimización no lineal como el del gradiente, gradiente conjugado o métodos basados en el Hesiano (Anexo D).

La literatura al respecto es ingente (véase por ejemplo, [Battiti 1992], [Bishop 1995], [Hagan *et al.* 1996], [Haykin 1999], [Fine 1999], [Demuth y Beale 2000], [Duda *et al.* 2001], y referencias allí citadas), por lo que nos limitaremos aquí a presentarlo brevemente en su implementación original con el algoritmo del gradiente.

El método es válido para cualquier función de pérdida diferenciable siempre que la contribución de cada observación en el riesgo empírico sea aditiva:

$$\hat{R}_n(\boldsymbol{\theta}) \propto \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i; \boldsymbol{\theta}))$$

Para la presentación del método suponemos el entrenamiento de una red del tipo MLP con activación diferenciable en el nivel de salida:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \varphi[\sum_{j=1}^{h} c_j \psi(\mathbf{w}_j^T \mathbf{x} + w_{j0}) + c_0]$$

Las ecuaciones básicas se deducen de la simple aplicación de la regla de la cadena. Para mayor claridad, reflejamos la cadena de composiciones funcionales implícita en la expresión anterior:

$$\begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} \Rrightarrow \begin{bmatrix} v_1 = \bar{\mathbf{w}}_1^T \bar{\mathbf{x}} \\ \vdots \\ v_h = \bar{\mathbf{w}}_h^T \mathbf{x} \end{bmatrix} \xrightarrow{\psi} \begin{bmatrix} u_0 = 1 \\ u_1 = \psi(v_1) \\ \vdots \\ u_h = \psi(v_h) \end{bmatrix} \Rrightarrow a = \bar{\mathbf{c}}^T \bar{\mathbf{u}} \xrightarrow{g} \varphi(a) = f(\mathbf{x}; \boldsymbol{\theta})$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde hemos agrupado todos los parámetros un el vector $\boldsymbol{\theta}$, y hemos utilizado la notación extendida, $\bar{\mathbf{x}} = (1, x_1, ..., x_d)^T$, $\bar{\mathbf{w}}_j = (w_{j0}, w_{j1}, ..., w_{jd})^T$, y lo mismo con $\bar{\mathbf{u}}$, $\bar{\mathbf{c}}$.

Con este esquema, las derivadas de la función objetivo con respecto a los parámetros son:

$$\begin{aligned}
\frac{\partial \ell}{\partial \bar{c}_j} &= \frac{\partial \ell}{\partial f}\frac{\partial f}{\partial a}\frac{\partial a}{\partial \bar{c}_j} = \frac{\partial \ell}{\partial f}\dot{\varphi}\bar{u}_j,\ j = 0:h \\
\frac{\partial \ell}{\partial \bar{w}_{ji}} &= \frac{\partial \ell}{\partial f}\frac{\partial f}{\partial a}\frac{\partial a}{\partial u_j}\frac{\partial u_j}{\partial v_j}\frac{\partial v_j}{\partial w_{ji}} = \frac{\partial \ell}{\partial f}\dot{\varphi}c_j\dot{\psi}x_i,\ i = 0:d,\ j = 1:h
\end{aligned}$$

La utilización de *back-propagation* por ejemplo con el algoritmo del gradiente es inmediata. En el caso de pérdida cuadrática, $\ell(y, f(\mathbf{x}, \boldsymbol{\theta})) = \frac{1}{2}[y - f(\mathbf{x}, \boldsymbol{\theta})]^2$, si la iteración $(s+1)$ es el resultado de presentar al algoritmo una nueva observación $(\mathbf{x},y)$, se tiene[69]:

$$\begin{aligned}
\bar{c}_j^{(s+1)} &= \bar{c}_j^{(s)} - \rho\frac{\partial \ell}{\partial \bar{c}_j} = \bar{c}_j^{(s)} + \rho(y - \hat{y})\dot{\varphi}\bar{u}_j = \bar{c}_j^{(s)} + \rho\delta\bar{u}_j \\
\text{con} &: \quad \delta = (y - \hat{y})\dot{\varphi},\ j = 0:h \\
\bar{w}_{ji}^{(s+1)} &= \bar{w}_{ji}^{(s)} - \rho\frac{\partial \ell}{\partial \bar{w}_{ji}} = \bar{w}_{ji}^{(s)} + \rho(y - \hat{y})\dot{\varphi}c_j\dot{\psi}x_i = \bar{w}_{ji}^{(s)} - \rho\delta_j x_i \\
\text{con} &: \quad \delta_j = \delta c_j\dot{\psi},\ i = 0:d,\ j = 1:h
\end{aligned}$$

donde hemos denotado $\hat{y} = f(\mathbf{x}, \boldsymbol{\theta}_s)$ con $\boldsymbol{\theta}_s$ el vector de parámetros resultado de la iteración anterior.

En lo anterior, puede observarse como el *error* $\delta = (y - \hat{y})\dot{\varphi}$ obtenido en el nivel de salida *se propaga hacia atrás* para obtener los *errores* que se producen en la capa oculta $\delta_j = \delta c_j\dot{\psi}$, $j = 1:h$. Este hecho da nombre al algoritmo, el cuál se resume en los siguientes pasos:

1. Etapa **forward**. Se presenta al algoritmo cada ejemplo $(\mathbf{x}, y)$, cuya información *se propaga hacia adelante* a través de la arquitectura de la red, produciendo la salida $\bar{\mathbf{u}}$ en el nivel oculto y la salida $f(\mathbf{x}; \boldsymbol{\theta})$ de la red.

2. Etapa **backward**. Se determinan los *errores* $\delta = (y - f(\mathbf{x}; \boldsymbol{\theta}))\dot{\varphi}(a)$ del nivel de salida y se *propagan hacia atrás* para obtener los *errores* de la capa anterior $\delta_j = \delta c_j\dot{\psi}(v_j)$.

3. Actualización de los parámetros. Los *errores* δ, δ_j, $j = 1:h$, anteriores, permiten el cálculo del gradiente $\nabla_{\boldsymbol{\theta}}\ell(y, f(\mathbf{x}; \boldsymbol{\theta}))$ para cada observación $(\mathbf{x}, y)$. Entonces,
   (a) Si el entrenamiento es de tipo interactivo (**on line**), se actualizan los parámetros de las capas oculta y de salida, según las ecuaciones anteriores.
   (b) Si el entrenamiento es por lotes (**batch**), el gradiente total se obtiene después de presentar todos los datos al algoritmo, mediante:

$$\nabla_{\boldsymbol{\theta}}\hat{R}_n(\boldsymbol{\theta}) \propto \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}\ell(y_i, f(\mathbf{x}_i; \boldsymbol{\theta}))$$

[69] Nótese que las ecuaciones correspondientes a los parámetros del nivel de salida coinciden con las de la regla delta mostrada en el apartado II.2.3, pues dicho nivel puede verse como una red de una sóla capa.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

y después se actualizan los parámetros. Con el algoritmo del gradiente, por ejemplo los parámetros del nivel de salida se actualizarían mediante:

$$\begin{aligned}\bar{c}_j^{(s+1)} &= \bar{c}_j^{(s)} - \rho\frac{\partial \hat{R}_n}{\partial \bar{c}_j} = \bar{c}_j^{(s)} - \rho\sum_{i=1}^{n}\frac{\partial \ell}{\partial f}(\mathbf{x}_i)\frac{\partial f}{\partial a}(a_i)\frac{\partial a}{\partial \bar{c}_j}(\bar{\mathbf{u}}_i)\\ &= \bar{c}_j^{(s)} - \rho\sum_{i=1}^{n}\frac{\partial \ell}{\partial f}(\mathbf{x}_i)\dot{\varphi}(a_i)\bar{u}_{j,i},\ j=0:h\end{aligned}$$

Aprendizaje Interactivo y Por Lotes

En general, el aprendizaje interactivo se prescribe cuando los datos se obtienen de uno en uno y es necesaria una respuesta inmediata. En estas condiciones, el algoritmo posee naturaleza estocástica y es un caso particular del algoritmo de aproximación estocástica de [Robbins y Monro 1951] (Anexo D).

Por el contrario, si la muestra de entrenamiento se obtiene de una sóla vez, se trata de calcular el óptimo de un problema de optimización completamente determinado por dicha muestra. En este caso, se actualizan los parámetros después de cada presentación de la muestra completa al algoritmo, (*epoch*), y esta acción se repite hasta la detención del algoritmo según algún criterio, por ejemplo, que el riesgo empírico alcance un umbral definido a priori, (véase más adelante, pág. 163).

De cualquier forma, el algoritmo *on-line* puede utilizarse como alternativa al aprendizaje *batch*, y existen posiciones a favor de esta estrategia argumentando que, la aleatorización que introduce, ayuda a evitar mínimos locales. Sin embargo, a ello puede objetarse también que, dicha aleatorización, puede impedir *buenos* mínimos locales, o que técnicas de aleatorización pueden utilizarse también con la versión *batch*.

En general, es difícil determinar cuál de las dos estrategias es la mejor pues sería necesario tener en cuenta el tiempo que es necesario en cada una para la monitorización de sus parámetros (tasa de aprendizaje, etc.). Una discusión al respecto puede verse en [Battiti 1992, p. 144].

Variaciones de *Back-propagation*

La dimensión algorítmica asociada al enfoque *back-propagation* exigiría por sí sólo un tratamiento específico de las dimensiones de este trabajo, (referencias anteriores). Las siguientes variaciones al algoritmo *back-propagation* original, representan sólo un esbozo de sus posibles implementaciones:

1. Modificaciones en el propio método del gradiente, como por ejemplo:
   (a) La consideración de una tasa de aprendizaje variable $\rho = \rho(s)$ para el control de la velocidad de convergencia, pues salvo en el caso interactivo, no hay garantía de convergencia con una tasa fija ([Luo 1991], [Battiti 1992, p. 144]).
   (b) La adición de un suavizado exponencial, denominado **momentum** en la literatura, destinado a evitar que el algoritmo se detenga en un mínimo local sin explorar un entorno del mismo y a evitar oscilaciones ineficientes del algoritmo.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

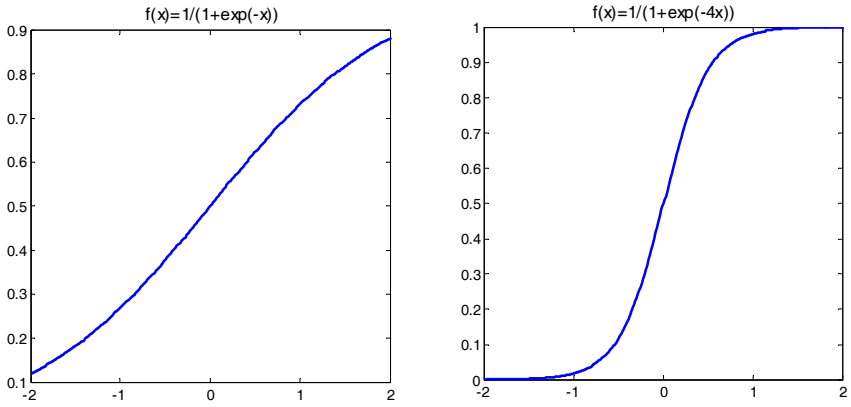


Figura 23. Derecha: $\psi(x) = (1 + e^{-x})^{-1}$. Izquierda: $\psi(x) = (1 + e^{-4x})^{-1}$

2. La utilización de algoritmos de búsqueda en una dirección (*line search*), como el algoritmo del **gradiente conjugado** (Anexo D) que evitan la determinación de la tasa de aprendizaje por métodos heurísticos.

3. La utilización de **algoritmos basados en el Hesiano**: algoritmo de Newton o algoritmos Quasi-Newton basados en aproximaciones del Hessiano (p. ej. [Press *et al.* 1992], [Bishop 1995]). Para problemas con pérdida cuadrática, el método más eficiente en nuestra experiencia es el método de Levenberg-Marquardt que utiliza también una aproximación del Hesiano ([Marquardt 1963], [Hagan y Menhaj 1994] y Anexo D).

Una amplia visión de las posibles variaciones de *back-propagation* puede verse en [Demuth y Beale 2000]. Una modificación reciente de interés que, en el caso de pérdida cuadrática, utiliza métodos lineales para agilizar la convergencia de *back-propagation* es [Fontenla-Romero *et al.* 2003]. Dicho trabajo, aporta también un método de inicialización del algoritmo.

Inicialización del Algoritmo

Cuando *back-propagation* se utiliza con redes MLP, el método más usual de inicialización del algoritmo es el propuesto por [Nguyen y Widrow 1990], que puede visualizarse mejor suponiendo un espacio de entrada $\mathcal{X} \subset \mathbb{R}$. El método se basa en que la función logística $\psi(x) = (1 + e^{-x})^{-1}$ (las demás funciones sigmoides admiten un análisis similar) es prácticamente lineal en el intervalo $[-1, 1]$, (figura 23, izquierda).

Si suponemos la variable x normalizada en $\mathcal{X} = [-1, 1]$ y la función implementada por la red es $f(x) = \sum_{j=1}^{h} c_j \psi(w_j x + w_{j0})$ con ψ la función logística, entonces el término j-ésimo de la suma $f_j(x) = c_j \psi(w_j x + w_{j0})$, es aproximadamente lineal en algún intervalo cuyo tamaño está determinado por w_j (figura 23, derecha) de tal forma que a mayor w_j menor intervalo. Además, la posición del intervalo depende de w_{j0} estando localizado en $x_0 = -w_{j0}/w_j$, y la pendiente de la curva en el intervalo es aproximadamente $c_j w_j$.

Así pues, la función implementada por la red es una función aproximadamente lineal a trozos, como suma de las funciones f_j en sus correspondientes intervalos. El método de [Nguyen y Widrow 1990] propone elegir los valores iniciales de w_j y w_{j0}, $j = 1 : h$, de tal forma que los intervalos correspondientes a las funciones f_j se repartan el dominio $\mathcal{X}$. Por tanto, si $\mathcal{X} = [-1, 1]$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

y hay h unidades ocultas, la longitud inicial de cada intervalo será $2/h$. Como f_j es aproximadamente lineal en $-1 < w_j x + w_{j0} < 1$, es decir:

$$-\frac{1}{w_j} - \frac{w_{j0}}{w_j} < x < \frac{1}{w_j} - \frac{w_{j0}}{w_j} \tag{II.89}$$

cada intervalo es de longitud $2/w_j$, se tiene:

$$\frac{2}{h} = \frac{2}{w_j} \Rightarrow w_j = h$$

De todas formas, [Nguyen y Widrow 1990] recomiendan un pequeño solapamiento de esos intervalos y proponen $w_j = 0{,}7h$.

El valor w_{j0}, se inicializa de tal forma que esos intervalos recubran en lo posible el espacio de entrada $\mathcal{X}$, con lo que, cada centro $x_0 \doteq -w_{j0}/w_j$ en (II.89), se elige aleatoriamente en $(-1, 1)$. O, lo que es lo mismo, w_{j0} se elige aleatoriamente en:

$$w_{j0} \in (-w_j, w_j)$$

Este análisis se extiende a un espacio $\mathcal{X} \subset \mathbb{R}^d$, con $d > 1$ aplicando el mismo procedimiento en diferentes direcciones $\mathbf{w} = (w_{j1}, ..., w_{jd})^T$ elegidas al azar.

Consistencia de *Back-propagation*

En este apartado presentamos resultados sobre la consistencia del algoritmo *back-propagation* a los parámetros óptimos. Con ello, estos resultados inciden en el error de estimación, no en el de aproximación (pág. 60), por ello, en este apartado, utilizaremos la palabra consistencia en este sentido. La consistencia del proceso de aprendizaje, en su auténtico sentido, podría obtenerse combinando las propiedades aproximadoras de las redes MLP o RBF, con los resultados que aquí presentamos.

El mayor problema para obtener resultados de consistencia y, sobre todo, normalidad asintótica y tasas de convergencia, para el algoritmo *back-propagation* es la propia geografía de la función objetivo, con posibles múltiples mínimos locales e incluso globales.

En nuestro conocimiento, los resultados más completos hasta la fecha son los de [Fine y Mukherjee 1999] (Anexo E), que consideran tal posibilidad no sólo para demostrar la consistencia del algoritmo (ya lo hacía [White 1989, th. 3.1]), sino también, para aportar tasas de convergencia y mostrar su normalidad asintótica.

Además, a diferencia de resultados anteriores, dichos resultados son válidos para cualquier distribución al utilizar la dimensión VC, no de la familia de funciones $\mathcal{L} = \{\ell_f(\mathbf{z}) = \ell(y, f(\mathbf{x})) : f \in \mathcal{F}\}$ sino de sus gradientes, que se supone finita[70].

Con ello, en esencia, se tiene la consistencia del proceso de aprendizaje de los gradientes, con lo cuál, sus resultados son válidos para cualquier algoritmo basado en el gradiente, incluyendo los métodos que aproximan el Hesiano (p. ej., Quasi-Newton [Press *et al.* 1992] y Levemberg-Marquardt, Anexo D), tanto en aprendizaje por lotes como interactivo.

[70] Aunque no conocemos resultados que permitan comprobar esta suposición. Para la función logística, se verifica $\dot{\psi} = \psi \cdot (1 - \psi)$ con lo que bastaría disponer de cotas a la dimensión VC de una familia $\mathcal{F}$ cuyas funciones son productos de funciones de otra, a partir de cotas a la dimensión VC de ésta.

Sobre dicha consistencia se tiene el resultado que presentamos a continuación, cuyos detalles pueden verse en el Anexo E. En él, se considera la siguiente familia de funciones implementadas por redes neuronales:

$$\mathcal{F} = \{f_{\boldsymbol{\theta}} = \varphi(\sum_{j=1}^{h} \psi_j) : \mathcal{X} \subset \mathbb{R}^d \to \widehat{\mathcal{Y}} \subset \mathbb{R} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$$

donde $\boldsymbol{\theta}$ recoge todos los parámetros de la red, como familia de hipótesis de un problema de aprendizaje con pérdida cuadrática $\ell(y, f(\mathbf{x})) = [y - f(\mathbf{x})]^2$, en el que se dispone de una muestra aleatoria simple $\{\mathbf{z}_i\}_{i=1}^n$ con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i = 1 : n$ obtenida de $\mathbb{P}_{\mathsf{Z}}$.

Teorema 49 *[Fine y Mukherjee 1999, th. 2] Sea M el conjunto de mínimos locales y globales del riesgo con $\#M < \infty$. Entonces, bajo determinadas condiciones (Anexo E), el estimador $\hat{\boldsymbol{\theta}}_n$ proporcionado por un algoritmo converge casi seguro a su mínimo más cercano $\boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n} \in M$, es decir:*

$$\mathbb{P}^{\infty}_{\mathsf{Z}}\{\lim_{n\to\infty} \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n} \right\| = 0\} = 1$$

El resultado anterior es válido para algoritmos con inicialización aleatoria, pues no se especifica a qué óptimo converge la solución y, de hecho, en cada entrenamiento, es de esperar que dicho óptimo sea distinto.

[Fine 1999] obtienen, entre otros, los siguientes resultados adicionales:

1. La tasa de convergencia [Fine 1999, th. 4] es de orden:

$$\left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n} \right\| = O_p(1/\sqrt{n})$$

2. La distribución de $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n})$, condicionada a que $\boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n}$ sea un óptimo particular $\boldsymbol{\theta}^*_j \in M$, es [Fine 1999, th. 5] asintóticamente normal con media $\mathbf{0}$ y matriz de covarianzas:

$$\mathbf{V} = \mathbf{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*_j)^{-1}\mathbf{C}_j\mathbf{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*_j)^{-1}$$
$$\text{con: } \mathbf{C} = 4\mathbb{E}_{\mathsf{Z}}[\ell(\mathsf{Y}, f_{\boldsymbol{\theta}^*_j}(\mathsf{X}))\nabla f_{\boldsymbol{\theta}^*_j}\nabla f^T_{\boldsymbol{\theta}^*_j}]$$

Control de la Complejidad

El problema de los algoritmos que persiguen la minimización del riesgo empírico, es que, si la red neuronal posee excesiva complejidad, el algoritmo puede producir una solución que de lugar a sobreajuste. Con ello, son necesarias técnicas de control de la complejidad que permitan una adecuada selección del modelo.

Para los algoritmos del tipo *back-propagation*, se han propuesto diferentes técnicas, algunas de las cuáles, no son exclusivas de este grupo y pueden utilizarse con otros tipos de algoritmo. Dichas técnicas pueden clasificarse en dos grandes grupos:

1. **Métodos directos**, que, a su vez, son de dos tipos:
   (a) *Pruning* ([Cun *et al.* 1990], [Hassibi y Stork 1993], [Reed 1993]) y *Growing* ([Fahlman 1990], [Littman y Ritter 1992]). Estas técnicas utilizan una filosofía de control de la complejidad basada en selección de variables. De hecho, sus métodos van dirigidos a controlar el número de unidades ocultas de la red, lo que significa controlar el número de variables ocultas.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

(b) *Early Stopping*. Este tipo de técnicas está basada en la parada prematura del algoritmo para evitar la saturación de la función de error. Más adelante veremos que, en cierto modo, su efecto equivale a una acción de regularización sobre los parámetros.

2. **Métodos indirectos**. Este tipo de métodos aborda el control de la complejidad en la fase de modelización, canalizando la información a priori que pueda disponerse sobre la verdadera función a través de la función objetivo del problema de optimización. En este grupo, se incluyen:
   (a) Técnicas de regularización, consistentes en penalizar el tamaño de los parámetros.
   (b) El enfoque bayesiano, que, además de aportar a las anteriores un fundamento probabilístico, ofrece un método integrado para todo el proceso de control de la complejidad y selección del modelo.

A continuación, presentamos los métodos más relevantes de los indicados, dejando el método bayesiano para un apartado posterior aglutinando técnicas de este tipo utilizadas con otros algoritmos (Apartado II.6.5).

Pruning: OBD y OBS

La denominación *pruning* de este grupo de técnicas se deriva de su acción de *podar* la red una vez entrenada, con objeto de obtener la mejor capacidad de generalización. Dicha poda afecta a las unidades del nivel oculto.

Los métodos más conocidos son los denominados imaginativamente *Optimal Brain Damage* (OBD) de [Cun *et al.* 1990] y *Optimal Brain Surgeon* (OBS) de [Hassibi y Stork 1993]. Ambos son métodos basados en el Hesiano de la función objetivo, aunque OBD utiliza una aproximación diagonal no muy realista. Además, OBD es un caso particular de OBS, por lo que presentamos únicamente un esquema de este último.

El método OBS supone una red neuronal de complejidad suficiente, ya entrenada, a riesgo de haberse producido sobreajuste. En este punto, el método procede a la eliminación progresiva de parámetros según la influencia de éstos (*saliency*) en el valor alcanzado para la función objetivo.

Para mayor claridad, suponemos una sóla salida, aunque el método se generaliza fácilmente.

Si denotamos por $\boldsymbol{\theta} \in \mathbb{R}^p$ el vector de todos los parámetros de la red neuronal, utilizando un desarrollo de Taylor de segundo orden en términos de dichos parámetros, se obtiene:

$$\Delta \hat{R}_n = \dot{\mathbf{r}}_n^T \Delta\boldsymbol{\theta} + \frac{1}{2} \Delta\boldsymbol{\theta}^T \mathbf{H}_n \Delta\boldsymbol{\theta} + O(\|\Delta\boldsymbol{\theta}\|^3)$$

donde $\dot{\mathbf{r}}_n = \nabla_{\boldsymbol{\theta}} \hat{R}_n$ es el gradiente y $\mathbf{H}_n = \partial^2 \hat{R}_n / \partial \boldsymbol{\theta}^2$ es el Hesiano de la función objetivo. Una vez entrenada la red, el término del gradiente es cero. Se trata entonces de eliminar el parámetro θ_s, $s \in \{1 : p\}$ que minimice el incremento del error derivado de su eliminación. Si $\Delta\boldsymbol{\theta}$ es la actualización del vector $\boldsymbol{\theta}$ que produce esta eliminación, entonces, $\theta_s + (\Delta\boldsymbol{\theta})_s = 0$, o lo que es lo mismo:

$$\theta_j + \mathbf{v}_j^T \Delta\boldsymbol{\theta} = 0$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde $\mathbf{v}_s$ es el vector con $(\mathbf{v}_s)_i = \delta_{is}$ delta de Kronecker.

Con ello, el objetivo es elegir θ_s, tal que:

$$s = \arg \min_{j=1:p} \left\{ \begin{array}{c} \min_{\Delta\boldsymbol{\theta}} \{\Delta\hat{R}_n = \frac{1}{2}\Delta\boldsymbol{\theta}^T \mathbf{H}_n \Delta\boldsymbol{\theta}\} \\ \text{sujeto a: } \theta_j + \mathbf{v}_j^T \Delta\boldsymbol{\theta} = 0 \end{array} \right\}$$

El Lagrangiano (Anexo D) de cada uno de estos p problemas de optimización con restricciones, es:

$$L(\theta_j, \lambda_j) = \frac{1}{2}\Delta\boldsymbol{\theta}^T \mathbf{H}_n \Delta\boldsymbol{\theta} + \lambda_j(\theta_j + \mathbf{v}_j^T \Delta\boldsymbol{\theta}_j)$$

donde $\{\lambda_j\}_{j=1}^p$ son los multiplicadores de Lagrange, con lo que el óptimo es:

$$\Delta\boldsymbol{\theta} = -\frac{\theta_j}{[\mathbf{H}_n^{-1}]_{jj}} \mathbf{H}_n^{-1} \mathbf{v}_j \Rightarrow L_j^* = \frac{1}{2}\frac{\theta_j^2}{[\mathbf{H}_n^{-1}]_{jj}} \tag{II.90}$$

El valor óptimo L_j^* refleja la variación en la función objetivo al eliminar el parámetro θ_j, por lo que se denomina **influencia** (*saliency*) de dicho parámetro.

De esta forma, el algoritmo OBS va eliminando sucesivamente los parámetros de la red con menor influencia en la función objetivo, previa actualización de los mismos mediante (II.90) después de cada eliminación. Como criterio de parada puede utilizarse un umbral de máximo incremento en la función objetivo después de las eliminaciones, o bien la utilización de una muestra de validación para seleccionar con antelación el número de parámetros que produce la mayor capacidad de generalización.

El problema del método anterior es el cálculo de la inversa del Hesiano. Para soslayar esta dificultad, [Hassibi y Stork 1993] proponen la siguiente aproximación del Hesiano: si la salida de la red es $\hat{y} = f(\mathbf{x}; \boldsymbol{\theta})$, el gradiente y el Hesiano son:

$$\begin{aligned} \frac{\partial \hat{R}_n}{\partial \boldsymbol{\theta}} &= -\frac{1}{n}\sum_{i=1}^n \dot{\mathbf{f}}_i (y_i - \hat{y}_i) \\ \frac{\partial^2 \hat{R}_n}{\partial \boldsymbol{\theta}^2} &= -\frac{1}{n}\sum_{i=1}^n [\dot{\mathbf{f}}_i \dot{\mathbf{f}}_i^T - \ddot{\mathbf{F}}_i (y_i - \hat{y}_i)] \end{aligned}$$

donde $\dot{\mathbf{f}}_i = \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i; \boldsymbol{\theta})$ y $\ddot{\mathbf{F}}_i = \partial^2 f(\mathbf{x}_i; \boldsymbol{\theta})/\partial\boldsymbol{\theta}^2$ son, respectivamente, el gradiente y el Hesiano de f en $\mathbf{x}_i$. Sin embargo, si la red ha producido sobreajuste, pueden obviarse los errores $y_i - \hat{y}_i$, con lo que la aproximación del Hesiano es:

$$\mathbf{H} = -\frac{1}{n}\sum_{i=1}^n \dot{\mathbf{f}}_i \dot{\mathbf{f}}_i^T$$

que coincide con la aproximación de Gauss-Newton (Anexo D, pág. 392). Pero, incluso, esta aproximación puede justificarse por dos razones:

1. Por razones computacionales, pues es probable que, al finalizar la labor de *pruning*, el Hesiano tienda a estar mal condicionado, con lo que, su cálculo directo, al margen de costoso, podría producir resultados poco precisos[71]. De hecho, la aproximación anterior se utiliza para lograr mayor estabilidad, al ser la esperanza del Hesiano (algoritmo *Fisher scoring*, Anexo D).

[71] De hecho, en el algoritmo *Fisher scoring*, esta aproximación se utiliza para lograr mayor estabilidad, al ser la esperanza del hesiano [Seber y Wild 1989].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2. Por razones conceptuales, pues, realmente, el proceso de *pruning* consiste en ir extrayendo del modelo toda componente de ruido con el fin de que la red alcance capacidad de generalización. Teniendo en cuenta que la red original entrenada poseía suficiente complejidad, estos residuos formarían parte de la perturbación, y no desecharlos a la hora de calcular el Hesiano, supondría introducir aquélla en el modelo.

Finalmente, es preciso notar que, el enfoque *pruning* no evita la selección posterior del modelo ya que, los resultados que produce, están condicionados por el modelo elegido inicialmente. En este sentido, el efecto del método es depurar el modelo estimado, eliminando los parámetros que aportan varianza sin beneficio alguno, pero ello no equivale a conseguir el óptimo de un modelo más reducido.

Técnicas de Regularización y *Early Stopping*

Las tres técnicas de regularización más relevantes con *back-propagation* son las siguientes:

1. *Weight Decay*.
2. *Weight Elimination* de [Weigend *et al.* 1991].
3. *Aproximate Smoother* de [Moody y Rögnvaldsson 1997].

Todas ellas se distinguen por el tipo de estabilizador utilizado. La última emplea como estabilizador una medida funcional de complejidad adaptada a la forma específica de las funciones del espacio de hipótesis $\mathcal{F}$, mientras que los dos primeros utilizan estabilizadores sobre el tamaño de los parámetros (en contraste con las técnicas directas orientadas a controlar su número).

Weight Decay

Si $\mathcal{F} = \{f(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, es el espacio de hipótesis, el estabilizador denominado *weight decay* es $\mathcal{E}(\boldsymbol{\theta}) = \sum_{j=1}^{p} \theta_j^2$ donde p es el número de parámetros de la red. Con ello, la función objetivo resulta:

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) = \hat{R}_n(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta}$$

Su denominación proviene de que $\frac{\partial}{\partial \boldsymbol{\theta}} \hat{\mathcal{R}}_n = \frac{\partial}{\partial \boldsymbol{\theta}} \hat{R}_n + 2\lambda \boldsymbol{\theta}$, con lo que sólo en términos del estabilizador, la variación *instantánea* (en las iteraciones) de $\boldsymbol{\theta}$ con el algoritmo del gradiente (Anexo D), sería:

$$\frac{d\boldsymbol{\theta}}{ds} \approx \Delta\boldsymbol{\theta} = -\rho \frac{\partial \mathcal{R}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\rho\lambda\boldsymbol{\theta}$$

lo que se corresponde con una solución $\boldsymbol{\theta}(s) = \boldsymbol{\theta}(0)\exp(-2\rho\lambda s)$ con decaimiento exponencial.

El efecto del *weight decay* puede ilustrarse [Bishop 1995], con una función objetivo cuadrática (situación análoga a la de un modelo lineal o aproximado linealmente, con pérdida cuadrática). Si se considera el problema no regularizado, esta función objetivo sería de la forma:

$$\hat{R}_n(\boldsymbol{\theta}) = a + \boldsymbol{\theta}^T \mathbf{b} + \frac{1}{2}\boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta}$$

con $\mathbf{H}$ (el Hesiano) simétrica y definida positiva, y donde el óptimo $\hat{\boldsymbol{\theta}}$ satisface: $\mathbf{b} + \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{0}$. Por otra parte, si se considera el problema regularizado con el estabilizador $\mathcal{E}(\boldsymbol{\theta}) = \lambda \boldsymbol{\theta}^T \boldsymbol{\theta}$, la nueva

solución óptima $\bar{\boldsymbol{\theta}}$ satisface $\mathbf{b} + (\mathbf{H} + \lambda \mathbf{I}_p)\bar{\boldsymbol{\theta}} = \mathbf{0}$. Con ello:

$$\mathbf{b} + \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{0} = \mathbf{b} + (\mathbf{H} + \lambda \mathbf{I}_p)\bar{\boldsymbol{\theta}}$$
$$\Rightarrow \mathbf{H}\hat{\boldsymbol{\theta}} = (\mathbf{H} + \lambda \mathbf{I}_p)\bar{\boldsymbol{\theta}}$$

Si $\{\mu_j\}_{j=1}^p$ son los autovalores de $\mathbf{H}$, utilizando como base los correspondientes autovectores $\{\boldsymbol{\xi}_j\}_{j=1}^p$, puede escribirse $\hat{\boldsymbol{\theta}} = \sum_{j=1}^p \hat{\theta}_j \boldsymbol{\xi}_j$ y $\bar{\boldsymbol{\theta}} = \sum_{j=1}^p \bar{\theta}_j \boldsymbol{\xi}_j$. Con ello, cada lado de la igualdad verifica, respectivamente:

$$\mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H}\sum_{j=1}^p \hat{\theta}_j \boldsymbol{\xi}_j = \sum_{j=1}^p \hat{\theta}_j \mu_j \boldsymbol{\xi}_j$$
$$(\mathbf{H} + \lambda \mathbf{I}_p)\bar{\boldsymbol{\theta}} = (\mathbf{H} + \lambda \mathbf{I}_p)\sum_{j=1}^p \bar{\theta}_j \boldsymbol{\xi}_j = \sum_{j=1}^p \bar{\theta}_j (\mu_j + \lambda)\boldsymbol{\xi}_j$$

con lo que, al ser $\mathbf{H}$ simétrica y $\{\boldsymbol{\xi}_j\}_{j=1}^p$ ortogonales, igualando componentes en ambos términos, se tiene:

$$\hat{\theta}_j \mu_j = \bar{\theta}_j(\mu_j + \lambda) \Rightarrow \bar{\theta}_j = \frac{\mu_j}{\lambda + \mu_j}\hat{\theta}_j \tag{II.91}$$

De esta forma, dependiendo del valor de λ, si $\mu_j >> \lambda$ entonces $\bar{\theta}_j \approx \hat{\theta}_j$, pero si $\mu_j << \lambda$, la componente correspondiente $\bar{\theta}_j$ del óptimo regularizado prácticamente desaparece produciendo una disminución en el número efectivo de parámetros, el cuál, ahora resulta:

$$p' = \sum_{j=1}^p \frac{\mu_j}{\lambda + \mu_j} \tag{II.92}$$

Así, el coeficiente regularizador λ actúa como un listón en el proceso de eliminación de parámetros. La hipótesis anterior de función objetivo cuadrática equivale a la utilización de un modelo lineal con pérdida cuadrática que, con el estabilizador *weight decay* se conoce como regresión *ridge*, en cuyo marco se obtiene formalmente el mismo resultado (pág. 37).

El estabilizador *weight decay*, aunque sencillo y poco costoso computacionalmente posee defectos:

1. Produce resultados diferentes ante cambios de escala en el espacio de entrada, o ante diferentes parametrizaciones de la función (por ejemplo $\theta'_j = \theta_j^{-1}$).

2. Supone una gran simplificación en el problema al tratar todos los pesos por igual. Un siguiente paso sería la consideración de regularizadores del tipo $\mathcal{E}_\Lambda(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \Lambda \boldsymbol{\theta}$ con Λ matriz diagonal. Por ejemplo, el problema de la inconsistencia ante cambios de escala podría resolverse considerando dos regularizadores λ_1 y λ_2 para los parámetros del nivel oculto y del nivel de salida respectivamente [Bishop 1995], (al margen de los parámetros independientes).

3. Finalmente, este estabilizador, al promover parámetros de pequeño tamaño, puede favorecer la existencia de muchos parámetros pequeños, cuando quizás resultase más conveniente, tanto desde el punto de vista computacional, como para la interpretación del modelo, obtener un número menor aunque los parámetros tuviesen mayor tamaño.

A modo ilustrativo [Weigend *et al.* 1991], supónganse dos parámetros w_1, w_2 que afectan a entradas idénticas de una misma neurona (son redundantes). En lo que respecta al riesgo

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

empírico $\hat{R}_n(w)$, resulta indiferente utilizar dos pesos iguales $w_1 = w_2 = w/2$, o bien, uno grande $w_1 = w$ y otro nulo $w_2 = 0$. Sin embargo, *weight decay* tenderá a producir la primera solución ya que los costes respectivos para el estabilizador verifican: $2w^2/4 < w^2$.

Weight Elimination

Con objeto de evitar el efecto anterior, [Weigend *et al.* 1991] proponen como estabilizador:

$$\mathcal{E}(\boldsymbol{\theta}) = \sum_{j=1}^{p} \frac{\theta_j^2/\theta_0^2}{1+\theta_j^2/\theta_0^2} = \sum_{j=1}^{p} \frac{\theta_j^2}{\theta_0^2+\theta_j^2}$$

donde θ_0 es un valor de referencia que define el significado de *grande* y *pequeño* al referirse a los parámetros. Si θ_0 es suficientemente grande, se tiene el caso *weight decay* pero, en otro caso, se evita el efecto antes indicado.

Approximate Smoother

Sin embargo, el regularizador *weight elimination* trata también a todos los parámetros por igual sin considerar la forma de las funciones del espacio de hipótesis, lo que puede producir resultados no óptimos. Un enfoque más preciso sería, por ejemplo, imponer restricciones de suavidad a través de las derivadas de orden superior de las funciones candidatas.

Esto es lo que persiguen [Moody y Rögnvaldsson 1997], que proponen, para las redes MLP:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{h} c_j \psi(\mathbf{w}_j^T\mathbf{x} + w_0) + c_0$$

donde $\boldsymbol{\theta}$ aglutina todos los parámetros de la red, los siguientes estabilizadores, denominados *approximate smoother*, en sus versiones de suavizado global y local, respectivamente:

$$\begin{aligned} \mathcal{E}_{\text{global}}(\boldsymbol{\theta}) &= \sum_{j=1}^{h} c_j^2 \left\|\mathbf{w}_j\right\|^{2r-1} \\ \mathcal{E}_{\text{local}}(\boldsymbol{\theta}) &= \sum_{j=1}^{h} c_j^2 \left\|\mathbf{w}_j\right\|^{2r} \end{aligned}$$

donde r es el orden de la derivada $d^r f_{\boldsymbol{\theta}}/d\mathbf{x}^r$ que se pretende penalizar. Estos estabilizadores ya distinguen entre los parámetros del nivel oculto y del nivel de salida, algo que no hacían los estabilizadores anteriores.

Early Stopping

Al principio (pág. 164), anticipamos el hecho de que *early stopping*, siendo una técnica de parada anticipada del algoritmo, producía un efecto de regularización sobre los parámetros. Este hecho se ilustra de la siguiente forma ([Sjöberg y Ljung 1995], [Bishop 1995]): supongamos que $\hat{R}_n$ es una función cuadrática con óptimo global en $\hat{\boldsymbol{\theta}}$:

$$\hat{R}_n(\boldsymbol{\theta}) = \hat{R}_n(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\mathbf{H}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

con $\mathbf{H}$ matriz simétrica y definida positiva, y se utiliza el algoritmo del gradiente (Anexo D). El valor de los parámetros en cada iteración es:

$$\boldsymbol{\theta}_{s+1} = \boldsymbol{\theta}_s - \rho \nabla_{\boldsymbol{\theta}} \hat{R}_n(\boldsymbol{\theta}_s)$$

Aproximando el gradiente en $\hat{\boldsymbol{\theta}}$ en un entorno de $\boldsymbol{\theta}_s$, se tiene:

$$0 = \nabla_{\boldsymbol{\theta}} \hat{R}_n(\hat{\boldsymbol{\theta}}) \approx \nabla_{\boldsymbol{\theta}} \hat{R}_n(\boldsymbol{\theta}_s) + \mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_s) \Rightarrow \nabla_{\boldsymbol{\theta}} \hat{R}_n(\boldsymbol{\theta}_s) \approx \mathbf{H}(\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}})$$

Con ello:

$$\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \rho \nabla_{\boldsymbol{\theta}} \hat{R}_n(\boldsymbol{\theta}_s) - \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \rho \mathbf{H}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}} = (\mathbf{I}_p - \rho \mathbf{H})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})$$

y repitiendo este proceso, se consigue:

$$\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}} = (\mathbf{I}_p - \rho \mathbf{H})^s (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})$$

Para mayor claridad, supongamos que se toma como valor inicial $\boldsymbol{\theta}_0 = \mathbf{0}$ y que el algoritmo converge desde este punto. Entonces, se tiene $\boldsymbol{\theta}_s = [\mathbf{I}_p - (\mathbf{I}_p - \rho \mathbf{H})^s]\hat{\boldsymbol{\theta}}$. Si ahora, se toma como sistema de coordenadas los autovectores ortogonales $\{\boldsymbol{\xi}_j\}_{j=1}^p$ de $\mathbf{H}$ siendo $\{\mu_j\}_{j=1}^p$ sus correspondientes autovalores, y según dicho sistema: $\hat{\boldsymbol{\theta}} = \sum_j \hat{\theta}_j \boldsymbol{\xi}_j$ y $\boldsymbol{\theta}_s = \sum_j \theta_j^{(s)} \boldsymbol{\xi}_j$, entonces:

$$\boldsymbol{\theta}_s = \sum_{j=1}^p \theta_j^{(s)} \boldsymbol{\xi}_j = [\mathbf{I}_p - (\mathbf{I}_p - \rho \mathbf{H})^s]\hat{\boldsymbol{\theta}} = [\mathbf{I}_p - (\mathbf{I}_p - \rho \mathbf{H})^s] \sum_{j=1}^p \hat{\theta}_j \boldsymbol{\xi}_j$$

con lo que, igualando coordenadas:

$$\theta_j^{(s)} = [1 - (1 - \rho \mu_j)^s]\hat{\theta}_j \tag{II.93}$$

Por tanto, para que el algoritmo converja es suficiente con que $|1 - \rho\mu_1| < 1$, siendo μ_1 el mayor autovalor de $\mathbf{H}$. Sin embargo, si se para el algoritmo en la iteración s-ésima (s fijo), como $(1-a)^s \approx 1 - sa + O(a^2)$, en ese momento:

$$\begin{cases} (1 - \rho\mu_j)^s \approx 1 - s\rho\mu_j \approx 0, & \text{si } s \approx (\rho\mu_j)^{-1} \Leftrightarrow \mu_j \approx (\rho s)^{-1} \\ (1 - \rho\mu_j)^s \approx 1 - s\rho\mu_j \approx 1, & \text{si } \mu_j << (\rho s)^{-1} \end{cases}$$

y, así, en el primer caso, $\theta_j^{(s)} \approx \hat{\theta}_j$, mientras que, en el segundo: $\theta_j^{(s)} \approx 0$.

Con ello, el término $(\rho s)^{-1}$ desempeña el mismo papel de filtro que desempeñaba el regularizador λ en *weight decay*, (ecuación (II.91)), con respecto a las componentes de $\boldsymbol{\theta}_s$ asociadas a los autovalores muy pequeños de $\mathbf{H}$ (cuya –falta de– información se debe sobre todo del ruido de los datos).

De lo anterior, y de la ecuación (II.92) con $\lambda = 1/(\rho s)$, se deduce también que la continuación del algoritmo produce un incremento progresivo del número de parámetros.

Rendimiento de *Early Stopping*

La gran dificultad práctica de las técnicas *early stopping* reside en la determinación del momento óptimo de parada del algoritmo. El único método posible para ello es utilizar una muestra de validación.

USAL UNIVERSIDAD DE SALAMANCA

En cualquier caso, es conveniente evaluar los posibles beneficios de una parada prematura. A este respecto, son ilustrativos los resultados de [Amari *et al.* 1997] que describimos a continuación, si bien, suponen que existe un único valor óptimo para el problema y no consideran la posibilidad de mínimos locales.

Sea R y $\hat{R}_n$ el riesgo y el riesgo empírico, respectivamente, construidos a partir de la máxima verosimilitud:

$$R(\boldsymbol{\theta}) = -\mathbb{E}_{\mathsf{Z}}[q(\mathsf{Z}, \boldsymbol{\theta})]; \; \hat{R}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \ln q(\mathbf{z}_i, \boldsymbol{\theta})$$

donde $\mathbf{z}^n = (\mathbf{z}_1, ..., \mathbf{z}_n)$ es una muestra aleatoria simple y q es un modelo realizable de la función de probabilidad o densidad de Z : $q(\mathbf{z}, \boldsymbol{\theta}^*) = \mathsf{p}_{\mathsf{Z}}(\mathbf{z})$ para algún $\boldsymbol{\theta}^* \in \Theta \subset \mathbb{R}^p$. Además sean:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_n &= \arg\min_{\boldsymbol{\theta}\in\Theta} \hat{R}_n(\boldsymbol{\theta}) \\ \hat{\boldsymbol{\theta}}_n^{(s)} &= \arg\min_{t\in\mathbb{N}} R(\hat{\boldsymbol{\theta}}_n^{(t)}) \end{aligned}$$

donde $\hat{\boldsymbol{\theta}}_n^{(t)}$ es el valor obtenido al parar el algoritmo en la iteración t-ésima. Es decir: $\boldsymbol{\theta}^*$ es el valor óptimo en la familia de hipótesis, $\hat{\boldsymbol{\theta}}_n$ es el estimador de máxima verosimilitud y $\hat{\boldsymbol{\theta}}_n^{(s)}$ es el vector óptimo de parámetros que se obtiene con la parada óptima s.

Teorema 50 *[Amari* et al. *1997] Si s es la parada óptima del algoritmo, entonces con n suficientemente grande:*

$$\mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n^{(s)})] \approx R(\boldsymbol{\theta}^*) + \frac{p - \frac{1}{2}}{2n} \tag{II.94}$$

Por otra parte, en el primer capítulo obtuvimos (ecuación (I.20), pág. 37) que:

$$\mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n)] \approx R(\boldsymbol{\theta}^*) + \frac{p}{2n} \tag{II.95}$$

de donde se concluye que la parada óptima equivale a disminuir la dimensión efectiva de los parámetros en $1/2$. El beneficio de ello es:

$$\mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n)] - \mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_n^{(s)}] \approx \frac{p}{2n} - \frac{1}{2n}(p - \frac{1}{2}) = \frac{1}{4n}$$

Sin embargo, el valor (II.94) es esperado, y en la demostración de este teorema, [Amari *et al.* 1997] prueban que existe la posibilidad de que no se gane nada con la parada prematura.

Teorema 51 *Supongamos que se desea reservar un subconjunto de datos como muestra de validación y r es la proporción de datos dedicados a entrenamiento. Si $\hat{\boldsymbol{\theta}}_{rn}$ es la estimación obtenida con la parada óptima según esa muestra de validación, entonces, con n suficientemente grande:*

$$\mathbb{E}_{\mathsf{Z}^n}[R(\hat{\boldsymbol{\theta}}_{rn})] \approx R(\boldsymbol{\theta}^*) + \frac{2p-1}{4rn} + \frac{1}{4(1-r)n}$$

A partir de este teorema, se puede obtener la fracción óptima de datos a destinar a entrenamiento:

Teorema 52 *La fracción óptima de datos a destinar a entrenamiento es:*

$$r^* = 1 - \frac{\sqrt{2p-1} - 1}{2(p-1)}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

expresión que, si p es grande, puede escribirse como:

$$r^* \approx 1 - \frac{1}{\sqrt{2p}}$$

El teorema anterior produce de forma inmediata el siguiente:

Teorema 53 *Sea $\hat{\boldsymbol{\theta}}_{r^*n}$ la estimación obtenida utilizando la fracción óptima anterior, entonces el riesgo esperado es asintóticamente:*

$$E_{\mathbf{Z}^n}[R(\hat{\boldsymbol{\theta}}_{r^*n})] \approx R(\boldsymbol{\theta}^*) + \frac{1}{4n}(\sqrt{2p-1})+1)^2$$

expresión que, si p es grande, puede escribirse como:

$$E_{\mathbf{Z}^n}[R(\hat{\boldsymbol{\theta}}_{r^*n})] \approx R(\boldsymbol{\theta}^*) + \frac{p}{2n}(1+\sqrt{2/c}) \tag{II.96}$$

De los teoremas anteriores puede concluirse que:

1. El teorema 52 indica que la proporción óptima de observaciones que deben dedicarse a validación es:
$$r_{val} = \frac{1}{\sqrt{2p}}$$
lo que, por ejemplo, significa que si $p = 50$ deberíamos dedicar a estimación el 90 % de los datos y el 10 % a validación. (Cuanto mayor el número p de parámetros, mayor proporción se requiere para entrenamiento).

2. Si se compara (II.96) con (II.95), se concluye que, para muestras suficientemente grandes, el riesgo esperado obtenido mediante *early stopping* mediante una muestra de validación construida con la fracción óptima de los datos, es mayor que el que se obtiene si se utilizan todos los datos para entrenamiento.

Sin embargo, los resultados anteriores no pueden aplicarse a muestras de tamaño moderado pues la fracción óptima de validación y el verdadero beneficio de *early stopping*, dependerá del nivel de ruido de los datos y de la representatividad de la muestra.

II.6.2 Algoritmos de Regresión Lineal Generalizada

Cuando la red puede verse como un modelo lineal en parámetros, es decir, es de la forma:

$$f(\mathbf{x}) = \sum_{j=1}^{h} c_j \psi_j(\mathbf{x}) + c_0$$

donde las funciones $\psi_j(\mathbf{x})$ son fijas y conocidas, entonces, suele utilizarse pérdida cuadrática debido a que, el problema de aprendizaje se reduce a la resolución de un problema de regresión lineal generalizada, (las covariables son transformaciones no lineales de las variables originales: $\boldsymbol{\psi}(\mathbf{x}) = (\psi_1(\mathbf{x}), ..., \psi_h(\mathbf{x}))^T$). En este marco, se obtiene un suavizador lineal que hemos analizado en un apartado anterior para el caso de las RBF (pág. 119).

La suposición anterior requiere una estimación previa de los parámetros $\boldsymbol{\theta}_j$ de cada función de transferencia del nivel oculto, $\psi_j(\mathbf{x}) = \psi(\mathbf{x}; \boldsymbol{\theta}_j)$, además de la estimación del número h de dichas

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

funciones, con lo que, considerado todas estas operaciones en el algoritmo, éste es, realmente, de carácter no lineal.

A pesar de esta no linealidad, los algoritmos que presentamos a continuación están basados en técnicas de regresión lineal y suelen emplearse con las redes RBF cuando los centros $\mathbf{w}_j$ se eligen de entre los datos de la muestra $\{\mathbf{x}_i\}_{i=1}^n$, y los parámetros de escala se suponen reales $\sigma_j \in \mathbb{R}$, $j = 1 : h$, (siendo por tanto: $\boldsymbol{\theta}_j = (\mathbf{w}_j^T, \sigma_j)^T$).

El procedimiento seguido, consiste en fijar valores a priori para h (u otro parámetro que determine éste último –p. ej. un criterio de parada en la adición de nuevas funciones) y $\sigma_j : j = 1 : h$, con el fin de convertir el problema en uno de regresión lineal con selección de variables (ocultas) $\psi_j = \psi_{\sigma_j}(\cdot; \mathbf{w}_j)$ para, posteriormente, abordar el problema de selección del modelo para distintos valores de h y σ_j, $j = 1 : h$.

En este punto, conviene notar que, la selección del modelo en h, puede ser perfectamente abordable, pues se trata de una sóla variable, sin embargo, con respecto a los parámetros de escala, el problema de selección del modelo es exponencial en h pues si se fijan a priori r valores para cada σ_j, se requiere la comprobación de r^h posibles vectores $(\sigma_1, ..., \sigma_h)$, ($r^{hd}$ si hubiese un parámetro por cada dimensión y centro). Por esta razón, es frecuente que estos métodos utilicen un valor único σ, con $\sigma_j = \sigma$, $j = 1 : h$, teniendo en cuenta que, en este caso, no se pierden sus propiedades aproximadoras (tabla 8).

Como ya hemos apuntado en el capítulo primero (pág. 25), el problema de selección de variables puede abordarse con técnicas propiamente diseñadas para tal fin (selección *forward* o selección *backward*), o bien, mediante técnicas de regularización aplicada a los parámetros asociados a dichas variables.

Regularización

Si, en el marco del problema de regresión, consideramos el caso general donde el espacio de hipótesis es:

$$\mathcal{F} = \{f(\mathbf{x}) = \sum_{j=1}^{h} c_j \psi_j(\mathbf{x}) : \mathbf{c} \in \mathbb{R}^h, h \in \mathbb{N}\} \tag{II.97}$$

el cuál, engloba el caso de las RBF mencionadas, sin más que considerar $\psi_j = \psi(\mathbf{x}; \mathbf{w}_j)$ con σ prefijado y, posiblemente, alguna ψ con valor constante, el enfoque de regularización plantea el riesgo empírico regularizado:

$$\min_{\mathbf{c} \in \mathbb{R}^h} \{\hat{\mathcal{R}}_n(\mathbf{c}) = \|\mathbf{y} - \boldsymbol{\Psi}\mathbf{c}\|^2 + \sum_{j=1}^{h} \lambda_j c_j^2\} \tag{II.98}$$

donde $\mathbf{y} = (y_1, ..., y_n)^T$ y $\boldsymbol{\Psi}$ es la matriz $n \times h$ con fila i-ésima $(\psi_1(\mathbf{x}_i), ..., \psi_h(\mathbf{x}_i))$.

Esta formulación coincide con la del problema de regresión *ridge*, con la única salvedad de que aquí se plantean h parámetros regularizadores λ_j. Las condiciones de óptimo y la solución del problema anterior son:

$$\begin{aligned} \boldsymbol{\Psi}^T\boldsymbol{\Psi}\hat{\mathbf{c}} + \boldsymbol{\Lambda}\hat{\mathbf{c}} &= \boldsymbol{\Psi}^T\mathbf{y} \Rightarrow \hat{\mathbf{c}} = (\boldsymbol{\Psi}^T\boldsymbol{\Psi} + \boldsymbol{\Lambda})^{-1}\boldsymbol{\Psi}^T\mathbf{y} \\ &\Rightarrow \hat{\mathbf{y}} = \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\boldsymbol{\Psi} + \boldsymbol{\Lambda})^{-1}\boldsymbol{\Psi}^T\mathbf{y} = \mathbf{S}\mathbf{y} \end{aligned}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En el caso de un sólo regularizador $\lambda_j = \lambda$, $j = 1 : h$, la solución es $\hat{\mathbf{c}} = (\mathbf{\Psi}^T\mathbf{\Psi} + \lambda\mathbf{I}_h)^{-1}\mathbf{\Psi}^T\mathbf{y}$.

Selección de los Regularizadores

Para la selección de los parámetros regularizadores, [Orr 1995] aprovecha la forma lineal de la red para optimizar en términos de aquéllos los criterios analíticos de selección del modelo y obtener los regularizadores óptimos.

Para el caso de un sólo regularizador se tiene el siguiente resultado, que se obtiene sin excesiva complicación igualando a cero la derivada respecto de λ de la expresión de los correspondientes criterios de selección de modelo, (ecuaciones (I.24), (I.27) y (I.30)).

Teorema 54 *[Orr 1996] Para los siguientes criterios de selección del modelo, el valor óptimo de λ es:*

$$\begin{aligned}
&\textit{FPE:} \quad \hat{\lambda} = \frac{\mathbf{y}^T(\mathbf{I}_n - \mathbf{S})^2\mathbf{y}\,\mathrm{tr}(\mathbf{A}^{-1} - \hat{\lambda}\mathbf{A}^{-2})}{\hat{\mathbf{c}}^T\mathbf{A}^{-1}\hat{\mathbf{c}}}\frac{n}{(n-p)(n+p)}\\
&\textit{GCV:} \quad \hat{\lambda} = \frac{\mathbf{y}^T(\mathbf{I}_n - \mathbf{S})^2\mathbf{y}\,\mathrm{tr}(\mathbf{A}^{-1} - \hat{\lambda}\mathbf{A}^{-2})}{\hat{\mathbf{c}}^T\mathbf{A}^{-1}\hat{\mathbf{c}}}\frac{1}{(n-p)}\\
&\textit{BIC:} \quad \hat{\lambda} = \frac{\mathbf{y}^T(\mathbf{I}_n - \mathbf{S})^2\mathbf{y}\,\mathrm{tr}(\mathbf{A}^{-1} - \hat{\lambda}\mathbf{A}^{-2})}{\hat{\mathbf{c}}^T\mathbf{A}^{-1}\hat{\mathbf{c}}}\frac{n\ln n}{2(n-p)[n + (\ln n - 1)p]}
\end{aligned}$$

donde hemos denotado $\mathbf{A} = (\mathbf{\Psi}^T\mathbf{\Psi} + \lambda\mathbf{I}_h)^{-1}$ y $n - \mathrm{tr}(\mathbf{S}) = n - p$ es la medida de suavizado y p los grados de libertad del modelo.

Las ecuaciones anteriores son de tipo iterativo, es decir, se parte de un valor inicial para $\hat{\lambda}$ con el que se calcula la expresión de la derecha que, a su vez, produce un nuevo valor de $\hat{\lambda}$; y así sucesivamente hasta la convergencia.

Para poder abordar este enfoque en el caso de varios regularizadores, es necesario conocer como depende de cada regularizador λ_j, $j = 1 : h$, el criterio de selección del modelo utilizado.

Por ejemplo, en el caso concreto de GCV, denotando $\mathbf{P} = \mathbf{I}_n - \mathbf{S}$ y $\mathbf{P}_j = \mathbf{I}_n - \mathbf{S}_j$ siendo $\mathbf{S}_j$ la matriz *hat* producto de la regresión sin la variable ψ_j, el lema de inversión de una matriz particionada produce [Orr 1996], (pág. 413, Anexo G):

$$\mathbf{P} = \mathbf{P}_j - \frac{\mathbf{P}_j\boldsymbol{\psi}_j\boldsymbol{\psi}_j^T\mathbf{P}_j}{\lambda_j + \boldsymbol{\psi}_j^T\mathbf{P}_j\boldsymbol{\psi}_j} \tag{II.99}$$

donde $\boldsymbol{\psi}_j = (\psi_j(\mathbf{x}_1), ..., \psi_j(\mathbf{x}_n))^T$.

Puede demostrarse que, la expresión (I.27) de GCV en términos de $\mathbf{P}$, es:

$$\mathrm{GCV}(\hat{\mathbf{c}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - f(\mathbf{x}_i;\hat{\mathbf{c}}))^2}{(1 - \mathrm{tr}(\mathbf{S})/n)^2} = \frac{n\mathbf{y}^T\mathbf{P}^2\mathbf{y}}{[\mathrm{tr}(\mathbf{P})]^2}$$

con lo que, sustituyendo la expresión (II.99) en la ecuación anterior, se obtiene la dependencia de GCV en términos de λ_j [Orr 1996]:

$$\mathrm{GCV}(\widehat{\mathbf{c}}) = \frac{n(a\Delta_j^2 - 2b\Delta_j + d)}{(\alpha\Delta_j - \beta)^2}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

con: $\Delta_j = \lambda_j + \boldsymbol{\psi}_j^T \mathbf{P}_j \boldsymbol{\psi}_j$, $a = \mathbf{y}^T \mathbf{P}_j^2 \mathbf{y}$, $b = \mathbf{y}^T \mathbf{P}_j^2 \boldsymbol{\psi}_j \mathbf{y}^T \mathbf{P}_j \boldsymbol{\psi}_j$, $d = \boldsymbol{\psi}_j \mathbf{P}_j^2 \boldsymbol{\psi}_j (\mathbf{y}^T \mathbf{P}_j \boldsymbol{\psi}_j)^2$, $\alpha = \mathrm{tr}(\mathbf{P}_j)$ y $\beta = \boldsymbol{\psi}_j^T \mathbf{P}^2 \boldsymbol{\psi}_j$.

Derivando la expresión anterior con respecto a λ_j e igualando a cero, se obtiene el siguiente resultado:

Teorema 55 *[Orr 1996] Los valores óptimos de los regularizadores λ_j, $j = 1 : h$, para el criterio GCV, son:*

$$\hat{\lambda}_j = \begin{cases} \infty & si\ a\beta = \alpha b \\ \infty & si\ \boldsymbol{\psi}_j^T \mathbf{P}_j \boldsymbol{\psi}_j > \frac{d\alpha - b\beta}{b\alpha - a\beta}\ y\ a\beta > \alpha b \\ 0 & si\ \boldsymbol{\psi}_j^T \mathbf{P}_j \boldsymbol{\psi}_j > \frac{d\alpha - b\beta}{b\alpha - a\beta}\ y\ a\beta < \alpha b \\ \frac{d\alpha - b\beta}{b\alpha - a\beta} - \boldsymbol{\psi}_j^T \mathbf{P}_j \boldsymbol{\psi}_j & si\ \boldsymbol{\psi}_j^T \mathbf{P}_j \boldsymbol{\psi}_j \leq \frac{d\alpha - b\beta}{b\alpha - a\beta} \end{cases}$$

Estas ecuaciones son, de nuevo, de tipo iterativo pues, las matrices $\mathbf{P}_j$ se calculan a partir de $\mathbf{P}$ invirtiendo la fórmula (II.99) para obtener (pág. 412, Anexo G):

$$\mathbf{P}_j = \mathbf{P} + \frac{\mathbf{P} \boldsymbol{\psi}_j \boldsymbol{\psi}_j^T \mathbf{P}}{\lambda_j - \boldsymbol{\psi}_j^T \mathbf{P} \boldsymbol{\psi}_j} \tag{II.100}$$

Esta vez, han de elegirse h valores iniciales $\hat{\lambda}_j$,(una alternativa es resolver el problema para un sólo λ y utilizar dicho valor como valor inicial para todos los $\hat{\lambda}_j$), determinar uno a uno sus nuevos valores y repetir el proceso hasta que el criterio GCV no decrezca de forma significativa.

En el caso típico de las RBF, este procedimiento ha de realizarse para cada h, σ elegidos a priori. El valor de h puede elegirse suficiententemente grande pues, en teoría, el resultado de la regularización eliminará las variables ocultas ψ_j asociadas a los regularizadores λ_j grandes. Con respecto a σ, el hecho de que las funciones básicas sean radiales, implica que los regularizadores λ_j afectan a un entorno local de cada centro, lo que supone realizar una especie de suavizado local. Como consecuencia, [Orr 1996] indica que, el valor de dicho parámetro, no resulta tan relevante en la práctica, (aunque ello sólo es cierto en un entorno adecuado del valor óptimo pues, si el valor elegido es excesivo, la complejidad resultará demasiado limitada a priori, y la regularización nunca podrá incorporar complejidad adicional sino todo lo contrario).

De cualquier forma, lo cierto es que el procedimiento anterior requiere una gran carga computacional aunque sólo sea para muestras moderadas ($n \approx 100$), pues supone partir de un valor h lo suficientemente grande, lo que introduce otros tantos parámetros λ_j a estimar.

Selección de Variables mediante OLS

Una alternativa al algoritmo de regularización anterior es la selección de las variables más relevantes para el problema, en términos de la reducción que cada una produce en el riesgo empírico, (p. ej. [Miller 1990, cap. 3], [Seber 1977]).

El esquema general del procedimiento consiste en partir de ninguna variable oculta e ir añadiendo variables hasta que se satisface algún criterio de parada.

La función objetivo puede escribirse como:

$$\hat{R}(\mathbf{c}) = (\mathbf{y} - \boldsymbol{\Psi}\mathbf{c})^T (\mathbf{y} - \boldsymbol{\Psi}\mathbf{c}) = \mathbf{y}^T (\mathbf{I}_n - \mathbf{S})^2 \mathbf{y} = \mathbf{y}^T \mathbf{P} \mathbf{y}$$

USC

donde hemos denotado de nuevo $\mathbf{P} = \mathbf{I}_n - \mathbf{S}$ que es idempotente, y, por simplicidad, hemos eliminado el subíndice n del riesgo empírico.

Proceso de Ortogonalización de Gram-Schmidt

1. Inicio: $\boldsymbol{\psi}_j^{(0)} = \boldsymbol{\psi}_j, v(j) = j, \forall j \in \mathcal{N}_0 = \{1:h\}, \mathbf{U} = \mathbf{I}_h$
2. Iteración $s = 1, ..., h-1$
   - 2.1 Selección del nuevo vector ortogonal a los anteriores $\bar{\boldsymbol{\psi}}_j, \forall j \in \{1:h\} - \mathcal{N}_{s-1}$:
     $\bar{\boldsymbol{\psi}}_s = \boldsymbol{\psi}_{j_s}^{(s-1)}$ para algún $j_s \in \mathcal{N}_{s-1}$
   - 2.2 Reordenamiento de las no seleccionadas:
     - a) Si $s > 1$:
       - i) $(\mathbf{u})_i = (\mathbf{U})_{i,v(j_s)}\ i = 1:s-1; (\mathbf{u})_s = 1; (\mathbf{u})_i = 0, i = s+1:h$
       - ii) Mover columnas $v(j)$ a la derecha si $v(j) < v(j_s)$: $l = v(j_s) :_{\downarrow} s$
         $(\mathbf{U})_{i,l} = (\mathbf{U})_{i,l-1}\ i = 1:s-1; (\mathbf{U})_{i,l} = 0\ i = s:l; (\mathbf{U})_{l,l} = 1$
         $v(j) = v(j) + 1, \forall j \in \mathcal{N}_{s-1} \cap \{j : v(j) < v(j_s)$
       - iii) $(\mathbf{U})_{is} = (\mathbf{u})_i, i = 1:h$
     - a) $v(j_s) = s$
     - b) $\mathcal{N}_s = \mathcal{N}_{s-1} \backslash \{j_s\} = \{j \in 1:h$ tal que $v(j) > s\}$
   - 2.3 Actualización de los no seleccionados, (se detrae de cada $\boldsymbol{\psi}_j^{(s)}$ su componente según $\bar{\boldsymbol{\psi}}_s$):
     $\boldsymbol{\psi}_j^{(s)} = \boldsymbol{\psi}_j^{(s-1)} - (\mathbf{U})_{s,v(j)}\bar{\boldsymbol{\psi}}_s, \forall j \in \mathcal{N}_s$ siendo:
     $$(\mathbf{U})_{s,v(j)} = \frac{\left\langle \boldsymbol{\psi}_j^{(s-1)}, \bar{\boldsymbol{\psi}}_s \right\rangle}{\left\langle \bar{\boldsymbol{\psi}}_s, \bar{\boldsymbol{\psi}}_s \right\rangle},\ \forall j \in \mathcal{N}_s$$
3. Iteración h : Último vector ortogonal: $\bar{\boldsymbol{\psi}}_h = \boldsymbol{\psi}_{j_s}^{(h-1)}$ con $\{j_s\} = \mathcal{N}_s$

Tabla 9. Proceso de ortogonalización de Gram-Schmidt. $v(j)$ representa la columna que la variable ψ_j ocupa, en cada momento, en la matriz $\mathbf{U}$.

Si $\boldsymbol{\Psi}_s$ es la matriz cuyas columnas son los s vectores $\boldsymbol{\psi}_j$, $j = 1:s$ correspondientes a las s variables seleccionadas hasta el momento y:

$$\mathbf{P}_s = \mathbf{I}_n - \mathbf{S}_s = \mathbf{I}_n - \boldsymbol{\Psi}_s(\boldsymbol{\Psi}_s^T\boldsymbol{\Psi}_s)^{-1}\boldsymbol{\Psi}_s^T \tag{II.101}$$

es la matriz análoga a $\mathbf{P}$, después de introducir la variable s-ésima, entonces, el cambio en la función objetivo producido al introducir la $(s+1)$-ésima variable ψ_j será:

$$\begin{aligned} \hat{R}_s - \hat{R}_{s+1} &= \mathbf{y}^T\mathbf{P}_s\mathbf{y} - \mathbf{y}^T\mathbf{P}_{s+1}\mathbf{y} = \mathbf{y}^T(\mathbf{P}_s - \mathbf{P}_{s+1})\mathbf{y} \\ &= \mathbf{y}^T\frac{(\mathbf{P}_s\boldsymbol{\psi}_j\boldsymbol{\psi}_j^T\mathbf{P}_s)^2}{\boldsymbol{\psi}_j^T\mathbf{P}_s\boldsymbol{\psi}_j}\mathbf{y} \end{aligned} \tag{II.102}$$

donde hemos utilizado la ecuación (II.99). Por tanto, dicha expresión puede utilizarse como criterio para seleccionar la variable $(s+1)$-ésima.

Sin embargo, las variables ocultas pueden estar correladas con lo que, al considerar la introducción de una nueva variable, sólo resulta relevante su componente ortogonal al espacio generado por las ya introducidas. La consideración de estas aportaciones netas individuales permite reducir considerablemente la carga computacional del proceso.

Ello se consigue, mediante el algoritmo de **mínimos cuadrados ortogonales** (*orthogonal least squares* –OLS, [Chen *et al.* 1989], [Chen *et al.* 1991]) utilizado en regresión lineal para agilizar la resolución de un problema con un conjunto prefijado de variables regresoras.

Dicho algoritmo, se basa en el **proceso de ortogonalización de Gram-Schmidt** (tabla 9, [Press *et al.* 1992]) o alguna de sus variantes [Chen *et al.* 1989], que, después de la selección de cada variable, detrae, de las no seleccionadas, sus respectivas componentes en el subespacio lineal generado por aquélla, de tal forma que, las variables resultantes se mantienen ortogonales a las seleccionadas hasta el momento.

US UNIVERSIDAD DE SALAMANCA DE CO

Al finalizar el proceso, se obtiene una transformación ortogonal de la matriz de variables originales de la forma (QR *decomposition* –p. ej. [Press *et al.* 1992, p. 98]):

$$\mathbf{\Psi} = \bar{\mathbf{\Psi}}\mathbf{U} \tag{II.103}$$

donde h es el número de variables, $\mathbf{\Psi}$ es la matriz con las columnas originales $\boldsymbol{\psi}_j$, $j = 1 : h$, $\bar{\mathbf{\Psi}}$ es una matriz cuyas columnas son las nuevas variables ortogonalizadas $\bar{\boldsymbol{\psi}}_j$, $j = 1 : h$ de tal forma que $\bar{\mathbf{\Psi}}_h^T\bar{\mathbf{\Psi}}_h = \mathrm{diag}(\bar{\boldsymbol{\psi}}_1^T\bar{\boldsymbol{\psi}}_1, ..., \bar{\boldsymbol{\psi}}_h^T\bar{\boldsymbol{\psi}}_h)$ y $\mathbf{U}$ es una matriz $h \times h$ triangular superior con unos en la diagonal cuyos coeficientes no nulos son:

$$(\mathbf{U})_{sj} = \frac{\langle \boldsymbol{\psi}_j, \bar{\boldsymbol{\psi}}_s \rangle}{\langle \bar{\boldsymbol{\psi}}_s, \bar{\boldsymbol{\psi}}_s \rangle},\ s \leq j;\ j, s \in \{1 : h\}$$

es decir, las componentes de cada variable original $\boldsymbol{\psi}_j$ en términos de las nuevas variables ortogonales $\bar{\boldsymbol{\psi}}_s$.

El algoritmo OLS es el resultado de utilizar en el proceso de ortogonalización de Gram-Schmidt, el criterio de selección de variables indicado en (II.102), que, ahora, adopta la forma que describimos a continuación.

Denotando por $\mathbf{U}_s$ con $s < h$ el estado de la matriz $\mathbf{U}$ después de la iteración s-ésima (es decir, después de haber seleccionado s variables ortogonales), y utilizando las propiedades de la inversión de matrices, la expresión (II.101) puede escribirse como:

$$\begin{aligned}
\mathbf{P}_s &= \mathbf{I}_n - \mathbf{\Psi}_s(\mathbf{\Psi}_s^T\mathbf{\Psi}_s)^{-1}\mathbf{\Psi}_s^T = \mathbf{I}_n - \bar{\mathbf{\Psi}}_s\mathbf{U}_s(\mathbf{U}_s^T\bar{\mathbf{\Psi}}_s^T\bar{\mathbf{\Psi}}_s\mathbf{U}_s)^{-1}\mathbf{U}_s^T\bar{\mathbf{\Psi}}_s^T \\
&= \mathbf{I}_n - \bar{\mathbf{\Psi}}_s(\bar{\mathbf{\Psi}}_s^T\bar{\mathbf{\Psi}}_s)^{-1}\bar{\mathbf{\Psi}}_s^T = \mathbf{I}_n - \sum_{j=1}^{s}\frac{\bar{\boldsymbol{\psi}}_j\bar{\boldsymbol{\psi}}_j^T}{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j} \\
\Rightarrow\ \mathbf{P}_{s-1} - \mathbf{P}_s &= \frac{\bar{\boldsymbol{\psi}}_s\bar{\boldsymbol{\psi}}_s^T}{\bar{\boldsymbol{\psi}}_s^T\bar{\boldsymbol{\psi}}_s}
\end{aligned}$$

Sustituyendo esta expresión en (II.102), se obtiene como criterio:

$$\hat{R}_{s-1} - \hat{R}_s = \mathbf{y}^T(\mathbf{P}_{s-1} - \mathbf{P}_s)\mathbf{y} = \frac{\mathbf{y}^T\bar{\boldsymbol{\psi}}_s\bar{\boldsymbol{\psi}}_s^T\mathbf{y}}{\bar{\boldsymbol{\psi}}_s^T\bar{\boldsymbol{\psi}}_s} = \frac{\langle \mathbf{y}, \bar{\boldsymbol{\psi}}_s \rangle^2}{\langle \bar{\boldsymbol{\psi}}_s, \bar{\boldsymbol{\psi}}_s \rangle} \tag{II.104}$$

Por otra parte, los coeficientes del problema ortogonalizado pueden obtenerse mucho más fácilmente que en el problema original, mediante:

$$\hat{\mathbf{d}} = (\bar{\mathbf{\Psi}}^T\bar{\mathbf{\Psi}})^{-1}\bar{\mathbf{\Psi}}^T\mathbf{y} \Rightarrow \hat{d}_s = \frac{\bar{\boldsymbol{\psi}}_s^T\mathbf{y}}{\bar{\boldsymbol{\psi}}_s^T\bar{\boldsymbol{\psi}}_s},\ s = 1 : h \tag{II.105}$$

con lo que pueden calcularse en el mismo momento de la introducción de la variable s-ésima. Al finalizar el algoritmo, pueden obtenerse los coeficientes de las variables originales mediante la relación:

$$\hat{\mathbf{y}} = \mathbf{\Psi}\hat{\mathbf{c}} = \bar{\mathbf{\Psi}}\mathbf{U}\hat{\mathbf{c}} = \bar{\mathbf{\Psi}}\hat{\mathbf{d}} \Rightarrow \hat{\mathbf{d}} = \mathbf{U}\hat{\mathbf{c}} \Rightarrow \hat{\mathbf{c}} = \mathbf{U}^{-1}\hat{\mathbf{d}} \tag{II.106}$$

Criterio de Parada en OLS

USC UNIVERSIDADE

Los criterios utilizados para finalizar el algoritmo OLS suelen consistir en el establecimiento de un umbral mínimo prefijado para el error mínimo alcanzado o para la disminución del error

Algoritmo de Entrenamiento RBF

1. Definir A, conjunto de valores σ. Para cada σ, estimar f_σ como sigue:
   a. Calcular $\mathbf{\Psi}$ y centrar las variables y, ψ_j, $j = 1 : h$
   b. Número de unidades ocultas:
      b.1 Calcular el espectro de Ψ y ordenar los autovalores $\{\mu_j\}$ en orden ascendente:
      b.2 Fijar un máximo número de condición para Ψ, cond_0, y calcular:
      $$l = \min\{j \in \{1, ..., n\} : \text{cond}(\tilde{\Psi}) = \tfrac{\mu_n}{\mu_j} < \text{cond}_0\}$$
      b.3 Calcular el número de funciones básicas requerido: $h_\sigma = n - l$
   c. Regresión lineal mediante OLS con las variables y, ψ_j, $j = 1 : h_\sigma$ centradas
      c.1 $\mathcal{N}_0 = \{1 : n\}$, $\boldsymbol{\psi}_j^{(0)} = \boldsymbol{\psi}_j$, $j = 1 : h_\sigma$
      c.2 Para $s = 1, 2, ..., h_\sigma$, calcular:
      i) $\bar{\boldsymbol{\psi}}_s = \bar{\boldsymbol{\psi}}_{j_s}$ con $j_s = \arg \min_{j \in \mathcal{N}_{s-1}} \frac{\left\langle \mathbf{y}, \boldsymbol{\psi}_j^{(s-1)} \right\rangle^2}{\left\langle \boldsymbol{\psi}_j^{(s-1)}, \boldsymbol{\psi}_j^{(s-1)} \right\rangle}$ y $\hat{d}_s = \frac{\langle \mathbf{y}, \bar{\boldsymbol{\psi}}_s \rangle}{\langle \bar{\boldsymbol{\psi}}_s, \bar{\boldsymbol{\psi}}_s \rangle}$
      ii) $\mathcal{N}_s = \mathcal{N}_{s-1} - \{j_s\}$
      iii) $\boldsymbol{\psi}_j^{(s)} = \boldsymbol{\psi}_j^{(s-1)} - \frac{\left\langle \boldsymbol{\psi}_j^{(s-1)}, \bar{\boldsymbol{\psi}}_s \right\rangle}{\langle \bar{\boldsymbol{\psi}}_s, \bar{\boldsymbol{\psi}}_s \rangle}$, $\forall j \in \mathcal{N}_s$
      c.3 Calcular: $\hat{\mathbf{c}} = \mathbf{U}^{-1}\hat{\mathbf{d}}$, con $(\mathbf{U})_{sj} = \frac{\langle \boldsymbol{\psi}_j, \bar{\boldsymbol{\psi}}_s \rangle}{\langle \bar{\boldsymbol{\psi}}_s, \bar{\boldsymbol{\psi}}_s \rangle}$, $s \leq j$, $(\mathbf{U})_{sj} = 0$, $s > j$, $s, j \in \{1 : h_\sigma\}$
2. Seleccionar $\hat{f} \in \{\hat{f}_\sigma : \sigma \in A\}$ mediante algún criterio analítico

Tabla 10. Algoritmo de entrenamiento de la red RBF

obtenido en cada iteración. Por ejemplo, un criterio propuesto por [Chen *et al.* 1996] basándose en la expresión (II.104), consiste en parar el algoritmo cuando, para un $\Delta > 0$ se satisfaga:

$$\frac{1}{\mathbf{y}^T\mathbf{y}}(\mathbf{y}^T\mathbf{y} - \sum_{j=1}^{h} \frac{(\mathbf{y}^T\bar{\boldsymbol{\psi}}_j)^2}{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j}) = 1 - \sum_{j=1}^{h} \frac{(\mathbf{y}^T\bar{\boldsymbol{\psi}}_j)^2}{(\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j)(\mathbf{y}^T\mathbf{y})} < \Delta \qquad \text{(II.107)}$$

donde el sumatorio recoge las contribuciones de las variables introducidas.

Sin embargo, la elección de estos valores umbral es de carácter heurístico, y puede producir efectos no deseados: por ejemplo, si las columnas de la matriz de diseño $\mathbf{\Psi}$ poseen un alto grado de redundancia, un valor pequeño para Δ (excesivas funciones básicas) puede acabar produciendo singularidades en las contribuciones

$$\frac{\left\langle \mathbf{y}, \boldsymbol{\psi}_j^{(s-1)} \right\rangle^2}{\left\langle \boldsymbol{\psi}_j^{(s-1)}, \boldsymbol{\psi}_j^{(s-1)} \right\rangle}$$

de las variables no introducidas, pues sus componentes ortogonales (residuos con respecto a las introducidas) pueden llegar a ser insignificantes y, así, el denominador anterior.

Esto es lo que sucede, por ejemplo, en el caso en las redes RBF si el parámetro de escala σ de las funciones básicas es excesivamente grande.

Un procedimiento más fundamentado se basa en la idea de que el número de funciones básicas necesarias dependa del valor de σ elegido a priori, por cuanto dicho valor determina el grado de redundancia de dichas funciones. En teoría, a mayor número de funciones básicas, mayor complejidad en el modelo. Sin embargo, en la práctica, fijado un σ, el incremento de la complejidad es significativo sólo hasta un cierto número h_σ de funciones básicas, pues, a partir de él, las nuevas variables que se introduzcan pasan a ser prácticamente redundantes con las anteriores.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Este grado de redundancia puede determinarse realizando un análisis espectral de la matriz de diseño (o de $\mathbf{\Psi}^T\mathbf{\Psi}$) y, con los autovalores ordenados, determinar cuántos de los más pequeños

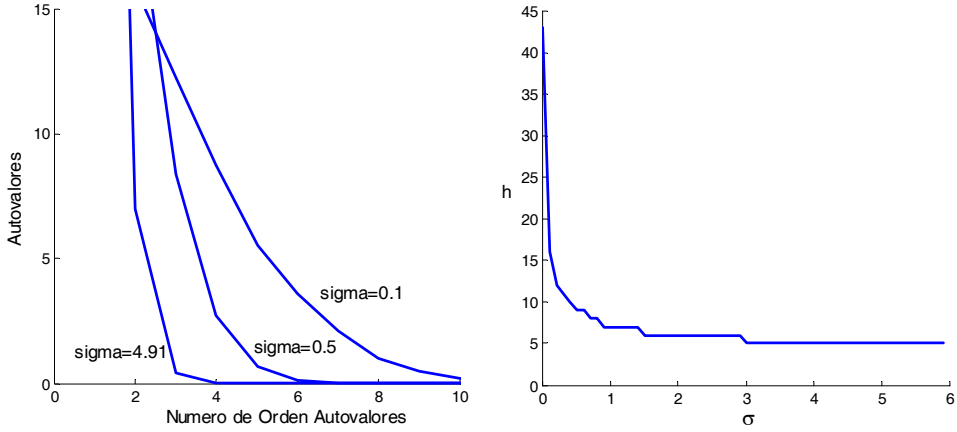


Figura 24. Izquierda: influencia de σ en el número de autovalores significativos de la matriz $\mathbf{\Psi}$, indicando el grado de redundancia de las variables ocultas, en cada caso. Derecha: número de funciones básicas no redundantes (a nivel de precisión del software) determinado por el valor de σ. Para un σ dado, la adición de nuevas variables a partir de un número h no aporta capacidad de representación (p. ej. si $\sigma = 5$, la utilización de más de 4 neuronas no aporta beneficios y produce un mal condicionamiento de la matriz $\mathbf{\Psi}$.

(columnas) hay que eliminar para que la matriz esté bien condicionada al nivel de precisión del software, (véase tabla 10).

La figura 24 (izquierda) refleja el espectro de la matriz $\mathbf{\Psi}$ para varios valores de σ y la derecha, el número de centros a que da lugar cada uno según el método anterior.

El hecho de que el valor de σ determine el número de funciones básicas necesarias, provoca que aquél sea menos sensible que el parámetro ventana de la regresión tipo núcleo o polinómico local, a la hora de seleccionar el modelo, y no sea necesaria una precisión tan alta a la hora de estimarlo: valores no tan próximos de σ pueden determinar un número similar de funciones básicas y, así, producir una capacidad de generalización equivalente.

Suboptimalidad y Grado de Parsimonia del Método OLS

El método OLS selecciona en cada paso la variable que produce mayor reducción en el error. Sin embargo, ello no tiene por qué producir la mejor selección de h variables [Sherstinsky y Picard 1996]. Ello es debido a que OLS persigue el óptimo a corto plazo (en cada iteración) en contraposición a una estrategia de largo plazo que evaluase el efecto de cada selección individual en las selecciones posteriores.

Como resultado, es posible que el método OLS no produzca la solución más parsimoniosa para un nivel de error mínimo prefijado, o, alternativamente, la solución con mínimo error para un número dado de funciones básicas.

Cualquier intento de evitar esta debilidad requeriría realizar un análisis combinatorio de todas las posibilidades ($n!$), pero ello es inviable computacionalmente con los tamaños muestrales usuales.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

OLS con Regularización

En algunas situaciones, puede ser conveniente combinar la técnica de selección de variables mediante OLS con técnicas de regularización. Por ejemplo, [Orr 1995] y [Chen *et al.* 1996] combinan el método anterior con la regularización *weight decay* planteando el problema:

$$\min_{\mathbf{c}\in\mathbb{R}^h}\{\hat{\mathcal{R}}(\mathbf{c}) = \hat{R}(\mathbf{c}) + \lambda\mathbf{c}^T\mathbf{c} = \|\mathbf{y} - \mathbf{\Psi c}\|^2 + \lambda\sum_{j=1}^{h} c_j^2\}$$

Para dicho problema, de forma análoga a lo anterior, se obtienen las siguientes ecuaciones en la iteración s-ésima:

$$\mathbf{P}_s = \mathbf{I}_n - \sum_{j=1}^{s}\frac{\bar{\boldsymbol{\psi}}_j\bar{\boldsymbol{\psi}}_j^T}{\lambda + \bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j};\ \hat{\mathcal{R}}_{s-1} - \hat{\mathcal{R}}_s = \frac{(\mathbf{y}^T\bar{\boldsymbol{\psi}}_s)^2}{\lambda + \bar{\boldsymbol{\psi}}_s^T\bar{\boldsymbol{\psi}}_s};\ \hat{d}_s = \frac{\mathbf{y}^T\bar{\boldsymbol{\psi}}_s}{\lambda + \bar{\boldsymbol{\psi}}_s^T\bar{\boldsymbol{\psi}}_s} \qquad \text{(II.108)}$$

En un principio, podría parecer que esta combinación de métodos de control de la complejidad (selección de variables y de su número, y regularización) podría ser redundante y, en muchas ocasiones, lo es.

Dado que el parámetro de escala σ determina como vimos el nivel de complejidad de la red RBF, la regularización sólo tendría sentido como nuevo control de complejidad en el marco del nivel de complejidad permitido por aquél.

La cuestión es, si un segundo nivel de control puede aportar beneficios, sobre todo teniendo en cuenta que, se introduce un nuevo parámetro λ que es necesario estimar. Además, como vimos en la página 166 cuando tratamos el método *weight decay*, este enfoque tiende a producir redes menos parsimoniosas debido a que, la limitación del tamaño de los parámetros, tiende a requerir un mayor número de funciones básicas.

Este último hecho posee un doble efecto negativo pues, al haber más funciones básicas, tiende a diluirse el significado de los centros, lo que limita una de los atractivos de las redes RBF: su interpretabilidad en términos de aquéllos.

En general, puede decirse que la respuesta depende del problema en cuestión. Cuando la función a estimar posee diferentes grados de complejidad en distintas zonas de su dominio, la formulación de una red compleja puede producir sobreajuste en las zonas donde dicha complejidad no es necesaria. Y, viceversa, la formulación de una red menos compleja puede producir infraajuste en la zonas donde la complejidad se necesita.

En estos casos, la regularización combinada con el método OLS permite partir de una red de mayor complejidad en la confianza de que la regularización limitará la magnitud de los parámetros en las zonas (funciones básicas) en que ello sea necesario. Sin embargo, para ello debería formularse una regularización diferenciada de cada uno de los parámetros con un vector de regularizadores $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_h)^T$, en lugar de un sólo regularizador λ, planteándose a continuación el problema de la selección del vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_h)^T$. Este problema será tratado más adelante en el contexto del entrenamiento bayesiano.

En otro caso, es decir, si se presume que la función a estimar posee un nivel de complejidad similar en las distintas zonas del espacio, entonces, la regularización no suele aportar beneficios. Por el contrario, además de perjudicar la parsimonia e interpretabilidad del modelo, es necesario estimar dos parámetros de complejidad (σ y λ) que probablemente resulten redundantes complicando innecesariamente la tarea de selección (véase figura 25).

USE UNIVERSIDAD DE SALAMANCA DE CÓRDOBA

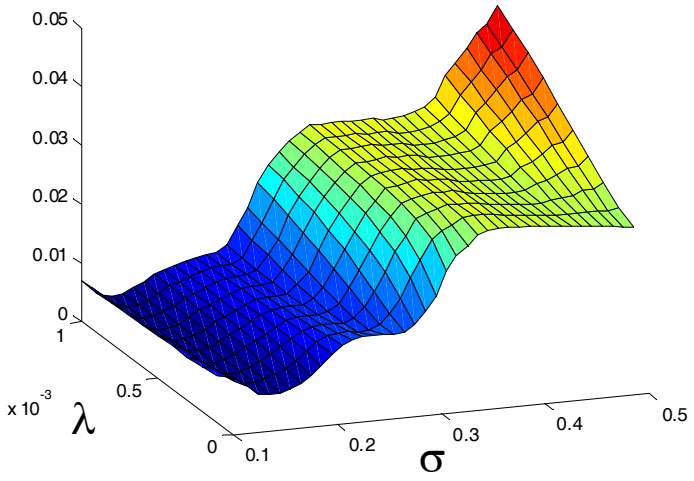


Figura 25. Influencia del par (σ, λ) de una red RBF, en el riesgo empírico de una muestra de validación, en un problema típico de aplicación. Como puede observarse, para cada σ, un regularizador $\lambda > 0$ no aporta beneficios al control de la complejidad que realiza σ

En estos casos, es mejor utilizar el análisis espectral de la matriz $\mathbf{\Psi}$, descrito en el epígrafe anterior, que, en sí mismo, es una regularización selectiva sobre las variables ocultas menos relevantes. Si por las características del problema se desease explorar el efecto de un nivel adicional de regularización, podría utilizarse un método mixto consistente en determinar primero primero el número de funciones básicas según el procedimiento indicado en el apartado anterior y después proceder a la regularización, de la siguiente forma:

1. Para cada valor $\sigma \in A$ donde A es un conjunto fijado a priori, determinar el número de funciones básicas h_σ necesarias y entrenar la red mediante OLS sin regularización.

2. Realizar la selección del modelo $\hat{f}_\sigma \in \{f_\sigma : \sigma \in A\}$ mediante algún criterio analítico presentado en el capítulo primero, teniendo en cuenta que la red entrenada es un suavizador lineal.

3. Explorar los posibles beneficios de la regularización mediante una regresión *ridge* para varios valores de (σ, λ) pero utilizando h_σ en cada caso como número de funciones básicas.

Algoritmos de Interpolación en Ambiente de Ruido

Como vimos en el Apartado II.5.2, las redes RBF de regularización se obtienen ya como solución de un problema de regularización:

$$\min_{f \in \mathcal{H}_k} \hat{\mathcal{R}}_n(f) = \min_{f \in \mathcal{H}_k} \{\hat{R}_n(f) + \lambda \|f\|^2_{\mathcal{H}_k})\}$$

donde $\mathcal{H}_k$ es el RKHS o SRKHS determinado por la función $k(\cdot, \mathbf{x}) = \psi(\cdot - \mathbf{x})$ donde k es definida positiva o condicionalmente definida positiva de orden m, respectivamente.

Como consecuencia, los parámetros de dichas redes quedan completamente determinados (pág. 135). Por tanto, para este tipo de redes, sólo resta realizar la tarea de selección del modelo, es

decir, la estimación del parámetro o parámetros del núcleo utilizado, y el parámetro regularizador λ.

A primera vista, podría pensarse en una posible redundancia entre estos dos parámetros de control de complejidad, tal como argumentamos en el apartado anterior para el caso de las redes RBF. Sin embargo, con las redes RRBF la ausencia de regularización ($\lambda = 0$) produce (Anexo G, pág. 414) una solución interpoladora de los datos, lo que es claramente pernicioso en ambiente de ruido. Por tanto, aquí sí tiene sentido la regularización, cuyo significado se verá con mayor profundidad en el capítulo siguiente.

Los métodos típicamente utilizados para la selección de ambos parámetros son los tratados en el capítulo primero. Sin embargo, en ausencia de ruido ($\lambda = 0$), si se desea utilizar el criterio CV con una menor carga computacional, interesa el método que presentamos a continuación, que es de utilidad para las técnicas analizadas en el capítulo siguiente.

Selección del Parámetro de Escala en la RRBF Interpoladora

El método que presentamos a continuación, debido a [Rippa 1999], es válido para una red RRBF sin término polinómico adicional y con el mismo parámetro de escala σ en todas las funciones básicas.

El método va dirigido a reducir la carga computacional derivada de la utilización del criterio de validación cruzada (CV). Suponiendo que se dispone de una muestra de datos $\mathbf{z}^n = (\mathbf{z}_1, ..., \mathbf{z}_n)$ con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, denotamos por $\mathbf{z}^{(i)}$ la muestra resultado de eliminar la observación $\mathbf{z}_i$ y por $\hat{f}_\sigma^{(i)}(\mathbf{x})$ la estimación RRBF a partir de la muestra $\mathbf{z}^{(i)}$ utilizando σ.

Si consideramos funciones de pérdida que son función de la discrepancia $y_i - \hat{f}_\sigma^{(i)}(\mathbf{x}_i)$, de la forma: $\ell(y_i, \hat{f}_\sigma^{(i)}(\mathbf{x}_i)) \equiv \ell(y_i - \hat{f}_\sigma^{(i)}(\mathbf{x}_i))$, entonces, el criterio CV es:

$$\mathrm{CV}(\hat{f}_\sigma) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \hat{f}_\sigma^{(i)}(\mathbf{x}_i)) \equiv \frac{1}{n}\sum_{i=1}^{n} \ell(y_i - \hat{f}_\sigma^{(i)}(\mathbf{x}_i))$$

Para cada muestra $\mathbf{z}^{(i)}$, el problema de regresión no regularizado con una red RRBF sin término polinómico, posee como solución:

$$\mathbf{\Psi}_{(i)}\mathbf{c}_{(i)} = \mathbf{y}_{(i)} \Rightarrow \hat{\mathbf{c}}_{(i)} = \mathbf{\Psi}_{(i)}^{-1}\mathbf{y}_{(i)}$$

donde el subíndice (i) denota que la observación $\mathbf{z}_i$ no está presente en la estimación, es decir $\mathbf{\Psi}_{(i)}$ es la matriz $\mathbf{\Psi}$ con $(\mathbf{\Psi})_{ij} = \psi(\mathbf{x}_i - \mathbf{x}_j)$, $i, j = 1 : n$, de la que se elimina la fila y la columna i-ésimas, (por simplicidad, no denotamos la dependencia de $\hat{\mathbf{c}}_{(i)}$ y $\mathbf{\Psi}_{(i)}$ del parámetro σ).

Por tanto, la aplicación de CV supone resolver n sistemas como el anterior, lo cuál, para muestras de tamaño relevante supone una carga computacional considerable. Sin embargo, cada término $y_i - \hat{f}_\sigma^{(i)}(\mathbf{x}_i)$ puede evaluarse de una forma más sencilla utilizando el siguiente teorema:

Teorema 56 *[Rippa 1999] Para cada $i \in \{1 : n\}$, sean $\mathbf{e}_i = (0, ..., 1^{(i}, ..., 0) \in \mathbb{R}^n$ y $\mathbf{u}^{(i)}$ la solución del sistema $\mathbf{\Psi}\mathbf{u}^{(i)} = \mathbf{e}_i$. Entonces, se verifica que:*

$$y_i - \hat{f}_\sigma^{(i)}(\mathbf{x}_i) = \frac{\hat{c}_i}{u_i^{(i)}}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

II.6.3 Algoritmos Mixtos

Dado el carácter lineal de las redes RBF con funciones básicas prefijadas, se han propuesto algoritmos mixtos que combinan esta propiedad con técnicas de optimización no lineal basadas en las derivadas. La idea es utilizar los métodos de regresión lineal en cada iteración del algoritmo no lineal, en las direcciones de búsqueda tipo gradiente en las que los parámetros del nivel oculto pueden considerarse constantes (y así, las funciones básicas).

Un ejemplo es el algoritmo de entrenamiento propuesto por [Müller *et al.* 1999] para redes RBF:

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{h} c_j \psi_j(\mathbf{x};\mathbf{w}_j)$$

donde $\boldsymbol{\theta}$ incluye todos los centros $\mathbf{w}_j$ y parámetros de escala σ_j, y donde suponemos funciones de activación gausiana (trasladable fácilmente a otras funciones de activación válidas para estas redes):

$$\psi_j(\mathbf{x}) = \exp\{-\frac{\|\mathbf{x}-\mathbf{w}_j\|^2}{2\sigma_j^2}\}$$

En general, se considera el problema regularizado con *weight decay*:

$$\hat{\mathcal{R}}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i;\boldsymbol{\theta})]^2 + \frac{\lambda}{2}\sum_{j=1}^{h} c_j^2$$

Sea $\boldsymbol{\theta}_h$ el subvector que incluye los parámetros correspondientes a los centros $\mathbf{w}_1, ..., \mathbf{w}_h$ y los radios $\sigma_1, ..., \sigma_h$. Entonces, podemos escribir $\boldsymbol{\theta} = [\boldsymbol{\theta}_h^T\ \mathbf{c}^T]^T$.

La idea básica del algoritmo se basa en que, para cada valor $\boldsymbol{\theta}_h$ fijo, (funciones básicas prefijadas), el vector óptimo $\mathbf{c}$ puede obtenerse como:

$$\mathbf{c}(\boldsymbol{\theta}_h) = (\boldsymbol{\Psi}_{\boldsymbol{\theta}_h}^T \boldsymbol{\Psi}_{\boldsymbol{\theta}_h} + \frac{\lambda}{2}\mathbf{I}_h)^{-1}\boldsymbol{\Psi}_{\boldsymbol{\theta}_h}^T \mathbf{y} \tag{II.109}$$

con lo que el problema puede reducirse a la optimización de:

$$\hat{\mathcal{R}}(\boldsymbol{\theta}) = \hat{\mathcal{R}}((\boldsymbol{\theta}_h\ \mathbf{c}(\boldsymbol{\theta}_h))) = \hat{\mathcal{R}}(\boldsymbol{\theta}_h)$$

Esta optimización se lleva a cabo mediante el método del gradiente conjugado versión Polak-Rivière (Anexo D) de la siguiente forma:

1. En una primera iteración, los centros $\mathbf{w}_j$ se inicializan mediante *clustering* con el método de las *k-medias* (p. ej. [Seber 1984]) y los parámetros de escala σ_j se inicializan como la distancia entre cada $\mathbf{w}_j$ y el centro $\mathbf{w}_i$, $i \neq j$ más próximo. Con estos valores se obtienen los valores iniciales $\boldsymbol{\theta}_h^{(1)}$ de $\boldsymbol{\theta}_h$ y, mediante (II.109), de $\mathbf{c}^{(1)}$.
2. Posteriormente, se aborda la optimización de $\hat{\mathcal{R}}(\boldsymbol{\theta}_h)$ mediante el método del gradiente conjugado en su versión de Polak-Ribière. Las derivadas de dicha función respecto a los parámetros

son para $j = 1 : h$:

$$\frac{\partial \hat{\mathcal{R}}}{\partial \mathbf{w}_j} = \sum_{i=1}^{n}(y_i - f(\mathbf{x}_i;\boldsymbol{\theta})\frac{\partial}{\partial \mathbf{w}_j}f(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - f(\mathbf{x}_i;\boldsymbol{\theta})c_j\frac{\mathbf{x}_i - \mathbf{w}_j}{\sigma_j^2}\psi_j(\mathbf{x};\mathbf{w}_j)$$
$$\frac{\partial \hat{\mathcal{R}}}{\partial \sigma_j} = \sum_{i=1}^{n}(y_i - f(\mathbf{x}_i;\boldsymbol{\theta})\frac{\partial}{\partial \sigma_j}f(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - f(\mathbf{x}_i;\boldsymbol{\theta})c_j\frac{\|\mathbf{x}_i - \mathbf{w}_j\|^2}{\sigma_j^3}\psi_j(\mathbf{x};\mathbf{w}_j)$$

En cada iteración $(s+1)$-ésima, $s \geq 1$, se construye el gradiente con estas derivadas en el punto $\boldsymbol{\theta}_h^{(s)}$:

$$\frac{\partial \hat{\mathcal{R}}}{\partial \boldsymbol{\theta}_h} = \mathbf{g}_s = [(\frac{\partial \hat{\mathcal{R}}}{\partial \mathbf{w}_1})^T, ..., (\frac{\partial \hat{\mathcal{R}}}{\partial \mathbf{w}_h})^T, \frac{\partial \hat{\mathcal{R}}}{\partial \sigma_1}, ..., \frac{\partial \hat{\mathcal{R}}}{\partial \sigma_h}]^T$$

se determina la dirección conjugada $\mathbf{d}_s$ y se realiza una búsqueda en dicha dirección hasta encontrar el valor óptimo:

$$\rho^* = \arg\min_{\rho \in \mathbb{R}} \hat{\mathcal{R}}(\boldsymbol{\theta}_h^{(s)} + \rho \mathbf{d}_s)$$

Al finalizar, se actualizan los parámetros: $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \rho^* \mathbf{d}_s$.

El algoritmo se cita en [Müller *et al.* 1999] para una red sin término independiente, pero dicho término mejora los resultados.

La ventaja del algoritmo anterior es que proporciona un método fiable de entrenar una red RBF con varios centros y, al mismo tiempo, éstos no están restringidos al conjunto de datos, permitiendo una mayor adaptación a la información contenida en la muestra. En contrapartida, las desventajas del algoritmo son:

1. La posibilidad de producir mínimos locales, en su calidad de algoritmo de optimización no lineal. Ello implica que, con los mismos datos diferentes pasadas del algoritmo pueden producir soluciones diferentes. A ello contribuye también el algoritmo inicial de las *k-medias*, que se resuelve también mediante optimización no lineal.

2. Su importante carga computacional, incluso con muestras de moderado tamaño, pues se requiere la resolución de un gran número de problemas de regresión durante la búsqueda en cada dirección conjugada para recalcular $\mathbf{c}$ mediante (II.109) y así poder evaluar $\hat{\mathcal{R}}(\boldsymbol{\theta}_h)$.

3. Finalmente, teniendo en cuenta lo anterior, la tarea de selección del modelo es compleja pues requiere la estimación del número de funciones básicas y del parámetro regularizador. Este último parámetro conviene incluirlo al margen de las razones mencionadas en el apartado anterior, pues ayuda a evitar situaciones de mal condicionamiento de la matriz $\boldsymbol{\Psi}_{\boldsymbol{\theta}_h}^T \boldsymbol{\Psi}_{\boldsymbol{\theta}_h}$ durante la búsqueda en cada dirección conjugada.

II.6.4 Algoritmos para SVM

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En el epígrafe de la página 147, dedujimos que para el entrenamiento de las redes SVM en el problema de regresión (el problema de clasificación produce un programa similiar a los efectos

de lo tratado en este apartado), requería la resolución del siguiente problema cuadrático:

$$\begin{gathered}\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathbb{R}^n}\{-\tfrac{1}{2}\sum_{i,j=1}^{n}(\alpha_i'-\alpha_j)(\alpha_j'-\alpha_j)k(\mathbf{x}_i,\mathbf{x}_j)-\varepsilon\sum_{i=1}^{n}(\alpha_i'+\alpha_i)+\textstyle\sum_{i=1}^{n}(\alpha_i'-\alpha_i)y_i \\ \begin{cases}\displaystyle\sum_{i=1}^{n}(\alpha_i'-\alpha_i)=0 \\ 0\leq\alpha_i,\alpha_i'\leq C,\ i=1:n\end{cases}\end{gathered} \tag{II.110}$$

Para poder utilizar algoritmos estándar de optimización cuadrática, el programa (II.110) puede escribirse de la siguiente forma [Drucker *et al.* 1997]:

$$\begin{gathered}\min_{\boldsymbol{\gamma}}\{\tfrac{1}{2}\boldsymbol{\gamma}^T\bar{\mathbf{K}}\boldsymbol{\gamma}+\boldsymbol{\gamma}^T\mathbf{c}\} \\ \boldsymbol{\gamma}^T\bar{\mathbf{G}}=0,\ \mathbf{0}_{2n}\leq\boldsymbol{\gamma}\leq C\mathbf{1}_{2n}\end{gathered} \tag{II.111}$$

donde:

$$\boldsymbol{\gamma}=\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\alpha}'\end{bmatrix},\ \bar{\mathbf{K}}=\begin{bmatrix}\mathbf{K} & -\mathbf{K}\\ -\mathbf{K} & \mathbf{K}\end{bmatrix},\bar{\mathbf{G}}=\begin{bmatrix}\mathbf{1}_n\\ -\mathbf{1}_n\end{bmatrix},$$
$$\mathbf{c}=(\varepsilon+y_1,\ldots,\varepsilon+y_n,\varepsilon-y_1,\ldots,\varepsilon-y_n)^T$$

Algoritmos especialmente eficientes para resolver programas del tipo (II.111) son los que utilizan el método del punto interior (*interior point*), por ejemplo, el algoritmo *primal-dual path-following* de [Vanderbei 1998] y su adaptación a SVM en [Smola y Schölkopf 1998], (Anexo D).

Sin embargo, los algoritmos del tipo punto interior requieren el almacenamiento en memoria de la matriz $\mathbf{K}$ con $(\mathbf{K})_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)$ lo que para muestras grandes puede resultar problemático. Para evitar este inconveniente, se han propuesto diversas alternativas orientadas a la descomposición del problema.

Algoritmos de Descomposición

Este enfoque tiene su origen en [Vapnik 1982] (*chunking*) y consiste en presentar sucesivamente al algoritmo subconjuntos de la muestra, aprovechando que la solución se expresa exclusivamente en términos de los vectores soporte. Con ello, en cada etapa, el algoritmo se aplica al conjunto de datos unión del nuevo subconjunto muestral con el conjunto de los vectores soporte de la pasada anterior. Ello permite procesar fragmentadamente grandes cantidades de datos.

Los algoritmos de descomposición (*decomposition algorithms*), implementan la filosofía anterior en base a una idea, sugerida por [Osuna *et al.* 1997], consistente en resolver el problema para un grupo de variables (*working set*) mientras se mantienen fijas las restantes.

Un ejemplo especialmente relevante de este grupo de algoritmos lo constituye el *sequential minimal optimization* (SMO), creado originalmente para el problema de clasificación por [Platt 1999] y adaptado al problema de regresión por [Smola y Schölkopf 1998], que elige cada subconjunto muestral con tamaño 2, con la ventaja de que cada problema parcial posee solución analítica agilizando significativamente el algoritmo.

Otra implementación de esta idea es el algoritmo SVM^{light} de [Joachims 1999] para el problema de clasificación que [Collobert y Bengio 2001] extienden (SVMTorch) al problema de regresión combinándola con la de Platt para obtener una solución analítica de cada problema parcial de dos variables.

La convergencia de los algoritmos de descomposición para SVM ha comenzado a estudiarse recientemente. Al respecto, puede verse por ejemplo [Chang *et al.* 2000], [Lin 2001] y [Keerthi y Gilbert 2002].

II.6.5 Aprendizaje Bayesiano

En el Apartado I.5.6 (pág. 45), hemos presentado el método bayesiano de selección del modelo basado en la maximización de la **evidencia**, el cuál, aproximaba la posterior de los hiperparámetros mediante la ecuación (I.34):

$$\mathsf{p}(\boldsymbol{\beta},\boldsymbol{\lambda}|\mathbf{y}^n,\mathcal{F},\mathcal{G}) = \frac{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\beta},\boldsymbol{\lambda},\mathcal{F},\mathcal{G})\mathsf{p}(\boldsymbol{\beta},\boldsymbol{\lambda}|\mathcal{F},\mathcal{G})}{\mathsf{p}(\mathbf{y}^n|\mathcal{F},\mathcal{G})} \propto \mathsf{p}(\mathbf{y}^n|\boldsymbol{\beta},\boldsymbol{\lambda},\mathcal{F},\mathcal{G})\mathsf{p}(\boldsymbol{\beta},\boldsymbol{\lambda}|\mathcal{F},\mathcal{G}) \quad \text{(II.112)}$$

donde, con objeto de facilitar el seguimiento, utilizamos la notación que hemos venido utilizando en el contexto bayesiano: $\mathsf{p}(\mathbf{u}|\mathbf{v},\mathbf{t}) \doteq \mathsf{p}_{\mathsf{U}/(\mathbf{v},\mathbf{t})}(\mathbf{u})$ e igualmente obviamos las observaciones de las variables independientes que siempre aparecen condicionando los valores de $\mathbf{y}^n$.

Así, en la expresión anterior: $\mathsf{p}(\mathbf{y}^n|\boldsymbol{\beta},\boldsymbol{\lambda},\mathcal{F},\mathcal{G}) \equiv \mathsf{p}(\mathbf{y}^n|\mathbf{x}^n,\boldsymbol{\beta},\boldsymbol{\lambda},\mathcal{F},\mathcal{G})$ es la evidencia siendo: $\mathbf{y}^n = (y_1,...,y_n)$ un conjunto de observaciones independientes de $\mathsf{Y}/\mathbf{x}_i$, $i = 1:n$,

$$\mathsf{p}(\mathbf{y}^n|\boldsymbol{\beta},\boldsymbol{\lambda},\mathcal{F},\mathcal{G}) \equiv \mathsf{p}(\mathbf{y}^n|\mathbf{x}^n,\boldsymbol{\beta},\boldsymbol{\lambda},\mathcal{F},\mathcal{G})$$

es la verosimilitud, $\mathcal{F} = \{\mathsf{p}(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p\}$ es una familia de hipótesis sobre la distribución $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$ de los datos, $\mathcal{G}$ es una familia de hipótesis sobre la distribución a priori $\mathsf{p}_{\boldsymbol{\theta}}$ de los parámetros, y $\boldsymbol{\beta}$ y $\boldsymbol{\lambda}$ son los hiperparámetros que controlan las distribuciones $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$ y $\mathsf{p}_{\boldsymbol{\theta}}$, respectivamente.

Asimismo, obviamos los términos $\mathcal{F}$ y $\mathcal{G}$ pues, los algoritmos bayesianos que aquí presentamos, tienen como objetivo la selección de los $\boldsymbol{\beta}$ y $\boldsymbol{\lambda}$ en el marco de dichas familias.

Así pues, supongamos hiperparámetros escalares β, λ y distribuciones $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$ y $\mathsf{p}_{\boldsymbol{\theta}}$ de la forma [MacKay 1992*a*]:

$$\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta},\beta) \equiv \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta},\beta) = \frac{1}{C_\beta}\exp\{-\beta\hat{R}_n(\mathbf{y}^n,\boldsymbol{\theta})\} \quad \text{(II.113a)}$$

$$\mathsf{p}(\boldsymbol{\theta}|\lambda) = \frac{1}{C_\lambda}\exp\{-\lambda\mathcal{E}(\boldsymbol{\theta})\} \quad \text{(II.113b)}$$

donde $\hat{R}_n(\mathbf{y}^n,\boldsymbol{\theta}) \equiv \hat{R}_n(\mathbf{y}^n,\mathbf{x}^n,\boldsymbol{\theta})$ es el riesgo empírico y $\mathcal{E}$ es un estabilizador.

Nótese que, si la distribución de los datos es gausiana, entonces:

$$\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i;\boldsymbol{\theta}))^2$$

$\beta = 1/\sigma^2$ y $C_\beta = (2\pi/\beta)^{n/2}$ es el factor normalizador de la gausiana. Igualmente, si la distribución a priori $\mathsf{p}_{\boldsymbol{\theta}}$ es gausiana, se obtiene: $\mathcal{E}(\boldsymbol{\theta}) = 1/2\,\|\boldsymbol{\theta}\|_2^2$ además de $\lambda = 1/\sigma_{\boldsymbol{\theta}}^2$ y $C_\lambda = (2\pi/\lambda)^{p/2}$.

Esta situación equivale a la regularización con estabilizador *early stopping* pues, como apuntamos en el Apartado I.4.1, la distribución a posteriori es (ecuación (I.31)):

$$\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n,\beta,\lambda) = \frac{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta};\beta)\mathsf{p}(\boldsymbol{\theta}|\lambda)}{\mathsf{p}(\mathbf{y}^n)} \propto \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta};\beta)\mathsf{p}(\boldsymbol{\theta}|\lambda) \quad \text{(II.114)}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde suponemos que la distribución de los datos no está afectada por λ ni la de los parámetros por β, con lo que la determinación de sus modas equivale a la minimización de la función objetivo:

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) = \beta \hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}) + \lambda \mathcal{E}(\boldsymbol{\theta}) = \frac{\beta}{2} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i;\boldsymbol{\theta}))^2 + \frac{\lambda}{2} \sum_{j=1}^{p} \theta_j^2$$

En el caso general de (II.113), la **evidencia** es:

$$\begin{aligned} \mathsf{p}(\mathbf{y}^n|\beta,\lambda) &= \int_\Theta \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta},\beta,\lambda)\mathsf{p}(\boldsymbol{\theta}|\beta,\lambda)d\boldsymbol{\theta} = \int_\Theta \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta},\beta)\mathsf{p}(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta} \\ &= \frac{1}{C_\beta}\frac{1}{C_\lambda}\int_\Theta \exp\{-\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}) - \lambda\mathcal{E}(\boldsymbol{\theta})\}d\boldsymbol{\theta} \\ &= \frac{1}{C_\beta C_\lambda}\int_\Theta \exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta})\}d\boldsymbol{\theta} = \frac{C_{\beta,\lambda}}{C_\beta C_\lambda} \end{aligned} \tag{II.115}$$

donde hemos definido:

$$\begin{cases} \hat{\mathcal{R}}_n(\boldsymbol{\theta}) \doteq \beta\hat{R}(\mathbf{y}^n|\boldsymbol{\theta}) + \lambda\mathcal{E}(\boldsymbol{\theta}) \\ C_{\beta,\lambda} \doteq \int_\Theta \exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta})\}d\boldsymbol{\theta} \end{cases}$$

Por tanto, para obtener la evidencia, es necesario calcular dicho factor normalizador $C_{\beta,\lambda}$. Pero éste se obtiene abordando el cálculo de la distribución a posteriori $\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n,\beta,\lambda)$ pues $C_{\beta,\lambda}$ es, precisamente, el factor normalizador de esta distribución. Desarrollando (II.114) se obtiene:

$$\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n,\beta,\lambda) = \frac{1}{C_{\beta,\lambda}}\exp\{-\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}) - \lambda\mathcal{E}(\boldsymbol{\theta})\} = \frac{1}{C_{\beta,\lambda}}\exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta})\}$$

donde $C_{\beta,\lambda}$ definido como antes, incluye los factores C_β, C_λ y $\mathsf{p}(\mathbf{y}^n)$.

Aproximación Gausiana de la Posterior

Cuando las distribuciones $\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$ y $\mathsf{p}_{\boldsymbol{\theta}}$ son gausianas, las funciones $\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta})$ y $\mathcal{E}(\boldsymbol{\theta})$ son cuadráticas, con lo que la posterior resultante es gausiana. Sin embargo, el factor de normalización $C_{\beta,\lambda}$ no posee forma analítica y es necesario algún método de aproximación. En general, se requiere algún método de aproximación o de Monte Carlo, para evaluar la integral implícita en $C_{\beta,\lambda}$.

El enfoque más sencillo, es el de la aproximación gausiana a la posterior en un entorno de su máximo. Esta aproximación puede justificarse teóricamente para un n grande pues la distribución a posteriori es asintóticamente gausiana (p. ej. [Berger 1985, p. 224]), aunque esta fundamentación no es bien vista en el contexto bayesiano, pues supone aceptar que el número de datos disponibles influye en el proceso de selección del modelo [Neal 1996, p. 9]. Por ello, la razón fundamental, es puramente computacional.

Así, si suponemos [MacKay 1992*a*] que $\hat{\mathcal{R}}_n(\boldsymbol{\theta})$ posee sólo una moda $\boldsymbol{\theta}_{MP}$ (máximo a posteriori), y denotamos por $\mathbf{H} = \partial^2\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP})/\partial\boldsymbol{\theta}^2$ el Hesiano de $\hat{\mathcal{R}}_n$ en $\boldsymbol{\theta}_{MP}$, entonces, desarrollando $\hat{\mathcal{R}}_n(\boldsymbol{\theta})$ en serie de Taylor en un entorno de $\boldsymbol{\theta}_{MP}$, se tiene:

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) \approx \hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MP})^T\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MP})$$

Con ello:

$$\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n,\beta,\lambda) \approx \frac{1}{\hat{C}_{\beta,\lambda}}\exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MP})^T\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MP})\} \tag{II.116}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

verificándose que:

$$\begin{aligned}\hat{C}_{\beta,\lambda} &= \int_{\Theta} \exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MP})^T \mathbf{H}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MP})\} d\boldsymbol{\theta} \\ &= \exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP})\}(2\pi)^{p/2}\,|\mathbf{H}|^{-1/2} \end{aligned} \tag{II.117}$$

De esta forma, mediante la aproximación anterior, se obtiene una distribución gausiana cuya matriz de covarianza es la inversa del Hesiano del logaritmo de la posterior.

El enfoque anterior es válido si la posterior posee una sóla moda, como es el caso de los modelos lineales, pero si dicha posterior posee varias modas relevantes, (caso típico de las redes MLF), el método puede ser problemático. Soluciones alternativas en este caso, son las siguientes:

1. Determinar las distintas modas relevantes y realizar una aproximación gausiana en la vecindad de cada una, (con covarianza, la inversa del Hesiano del logaritmo de la posterior), y aproximar la posterior mediante una mixtura gausiana ponderando cada aproximación mediante la masa de probabilidad de la posterior que reside en el entorno de cada una.

2. Establecer una **combinación de redes** (véase sección II.7) construidas, cada una, con los valores de los parámetros de cada una de las modas.

Pero, en general, si el número de datos es del mismo orden que el número de parámetros, las aproximaciones gausianas pueden producir resultados cuestionables. Por ejemplo, [MacKay 1992*c*] apunta que si $n/p < 3 \pm 1$, debe abandonarse esta técnica para la resolución de la integral en beneficio de otras basadas en Monte Carlo. Un tratamiento detallado de estos métodos puede verse por ejemplo en [Neal 1996].

Maximización de la Evidencia

En el caso concreto de distribuciones gausianas para los datos y para la prior de los parámetros, si utilizamos el valor aproximado (II.117) obtenido mediante la aproximación gausiana anterior, podemos escribir el logaritmo de la evidencia (II.115), a maximizar en β y λ, como:

$$\begin{aligned}\ln \mathsf{p}(\mathbf{y}^n|\beta,\lambda) &= \ln \frac{C_{\beta,\lambda}}{C_\beta C_\lambda} = \ln \frac{\exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP})\}(2\pi)^{p/2}\,|\mathbf{H}|^{-1/2}}{(2\pi/\beta)^{n/2}(2\pi/\lambda)^{p/2}} \\ &= -\beta \hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) - \lambda \mathcal{E}(\boldsymbol{\theta}_{MP}) \\ &\quad -\frac{1}{2}\ln|\mathbf{H}| + \frac{p}{2}\ln\lambda + \frac{n}{2}\ln\beta + \frac{n}{2}\ln(2\pi) \end{aligned} \tag{II.118}$$

donde $\boldsymbol{\theta}_{MP} = \boldsymbol{\theta}_{MP}(\lambda,\beta)$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Maximización de la Evidencia en la Varianza de la Prior de los Parámetros

Para obtener el óptimo λ_{MP} de la evidencia, la condición de óptimo en λ es:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\lambda}\ln \mathsf{p}(\mathbf{y}^n|\beta,\lambda) \\
&= -\frac{\partial}{\partial\lambda}[\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP})] - \frac{\partial}{\partial\lambda}[\lambda\mathcal{E}(\boldsymbol{\theta}_{MP})] - \frac{1}{2}\frac{\partial}{\partial\lambda}\ln|\mathbf{H}| + \frac{p}{2\lambda} \\
&= -[\beta\frac{\partial}{\partial\lambda}\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) + \mathcal{E}(\boldsymbol{\theta}_{MP}) + \lambda\frac{\partial}{\partial\lambda}\mathcal{E}(\boldsymbol{\theta}_{MP})] - \frac{1}{2}\frac{\partial}{\partial\lambda}\ln|\mathbf{H}| + \frac{p}{2\lambda} \\
&= -\mathcal{E}(\boldsymbol{\theta}_{MP}) - [\beta\frac{\partial\hat{R}_n}{\partial\boldsymbol{\theta}}(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) + \lambda\frac{\partial\mathcal{E}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_{MP})]\frac{\partial\boldsymbol{\theta}_{MP}}{\partial\lambda} - \frac{1}{2}\frac{\partial}{\partial\lambda}\ln|\mathbf{H}| + \frac{p}{2\lambda} \\
&= -\mathcal{E}(\boldsymbol{\theta}_{MP}) - \frac{1}{2}\frac{\partial}{\partial\lambda}\ln|\mathbf{H}| + \frac{p}{2\lambda}
\end{aligned}$$

donde hemos utilizado que $\mathbf{0} = \frac{\partial\hat{\mathcal{R}}_n}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_{MP}) = \lambda\frac{\partial\mathcal{E}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_{MP}) + \beta\frac{\partial\hat{R}_n}{\partial\boldsymbol{\theta}}(\mathbf{y}^n|\boldsymbol{\theta}_{MP})$ por ser $\boldsymbol{\theta}_{MP}(\lambda,\beta)$ óptimo para $\hat{\mathcal{R}}_n$.

Por otra parte, si escribimos el Hesiano como:

$$\mathbf{H} = \frac{\partial^2\hat{\mathcal{R}}_n}{\partial\boldsymbol{\theta}^2} = \beta\frac{\partial^2\hat{R}_n}{\partial\boldsymbol{\theta}^2} + \lambda\frac{\partial^2\mathcal{E}}{\partial\boldsymbol{\theta}^2} = \mathbf{B} + \lambda\mathbf{D}$$

con $\mathbf{B} = \beta\frac{\partial^2\hat{R}_n}{\partial\boldsymbol{\theta}^2}$ y $\mathbf{D} = \frac{\partial^2\mathcal{E}}{\partial\boldsymbol{\theta}^2}$, y suponemos[72] un estabilizador tipo *weight decay*, $\mathcal{E}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_2^2$ entonces $\mathbf{D} = \mathbf{I}_p$ y $\mathbf{H} = \mathbf{B} + \lambda\mathbf{I}_p$, con lo que:

$$\frac{\partial}{\partial\lambda}\ln|\mathbf{H}| = \frac{\partial}{\partial\lambda}\ln\prod_{j=1}^{p}(\mu_j+\lambda) = \frac{\partial}{\partial\lambda}\sum_{j=1}^{p}\ln(\mu_j+\lambda) = \sum_{j=1}^{p}\frac{1}{\mu_j+\lambda} = \operatorname{tr}\mathbf{H}^{-1} \qquad \text{(II.119)}$$

donde μ_j son los autovalores de $\mathbf{B}$ y hemos obviado su dependencia de λ a través de $\boldsymbol{\theta}_{MP}(\lambda,\beta)$ pues $\hat{R}_n$ es una función cuadrática. Con ello, la condición de óptimo para λ resulta:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\lambda}\ln \mathsf{p}(\mathbf{y}^n|\beta,\lambda) = -\mathcal{E}(\boldsymbol{\theta}_{MP}) - \frac{1}{2}\operatorname{tr}\mathbf{H}^{-1} + \frac{p}{2\lambda} \\
&\Rightarrow \quad 2\lambda\mathcal{E}(\boldsymbol{\theta}_{MP}) = p - \lambda\operatorname{tr}\mathbf{H}^{-1} = \sum_{j=1}^{p}\frac{\mu_j}{\mu_j+\lambda} = p' \qquad \text{(II.120)}
\end{aligned}$$

donde p' puede interpretarse como el número efectivo de parámetros (compárese con las ecuaciones (I.22) –pág. 37– y (II.92) –pág. 167), de la siguiente forma:

1. La cantidad de la izquierda en (II.120) es la cantidad de estructura que los datos introducen en el modelo, pues $\mathcal{E}(\boldsymbol{\theta}_{MP})$ es la medida en que los parámetros, después del entrenamiento, son distintos de cero.

2. La cantidad de la derecha en (II.120) refleja lo mismo, pero en función del número inicial de parámetros p. Por ello, la cantidad $\lambda\operatorname{tr}\mathbf{H}^{-1}$ es la reducción que se produce en dicho número después del entrenamiento, dando lugar al número efectivo final $p' = p - \lambda\operatorname{tr}\mathbf{H}^{-1}$.

[72] Esta suposición no es necesaria [MacKay 1992*a*, p. 432] y se realiza para simplificar la exposición. En el caso de que $\mathcal{E}$ sea cuadrática, esta transformación consiste un cambio a la base de autovectores de $\partial^2\mathcal{E}/\partial\boldsymbol{\theta}^2$ y realizar un cambio de escala de tal forma que la forma cuadrática $\mathcal{E}$ sea homogénea.

3. La expresión $\sum_{j=1}^{p} \frac{\mu_j}{\mu_j+\lambda}$ admite el mismo análisis realizado para el regularizador *weight decay* en cuanto a que, las direcciones del Hesiano $\mathbf{B}$ con $\mu_j >> \lambda$ contribuyen con una unidad al número efectivo de parámetros (los parámetros que son realmente significativos para los datos).

   Por el contrario, las direcciones no relevantes para los datos (a través de $\mathbf{B}$) quedan a merced de la distribución a priori de los parámetros (gausiana de media cero), y acaban poseyendo una magnitud irrelevante.

Maximización de la Evidencia en la Varianza de los Datos

Como μ_j son los autovalores de $\mathbf{B}$, por definición de esta matriz, han de ser proporcionales a β: $\mu_j = \alpha\beta \Rightarrow \frac{d\mu_j}{d\beta} = \alpha = \frac{\mu_j}{\beta}$. Así, al igual que en (II.119), se tiene:

$$\frac{\partial}{\partial\beta} \ln|\mathbf{H}| = \frac{\partial}{\partial\beta} \sum_{j=1}^{p} \ln(\mu_j + \lambda) = \frac{1}{\beta} \sum_{j=1}^{p} \frac{\mu_j}{\mu_j + \lambda}$$

con lo que las condiciones de máximo para β en (II.118), son:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\beta} \ln \mathsf{p}(\mathbf{y}^n|\beta,\lambda) \\
&= -\frac{\partial}{\partial\beta}[\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP})] - \frac{\partial}{\partial\beta}[\lambda\mathcal{E}(\boldsymbol{\theta}_{MP})] - \frac{1}{2}\frac{\partial}{\partial\beta}\ln|\mathbf{H}| + \frac{n}{2\beta} \\
&= -[\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) + \beta\frac{\partial\hat{R}_n}{\partial\boldsymbol{\theta}}(\mathbf{y}^n|\boldsymbol{\theta}_{MP})\frac{\partial\boldsymbol{\theta}_{MP}}{\partial\beta} + \lambda\frac{\partial\mathcal{E}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_{MP})\frac{\partial\boldsymbol{\theta}_{MP}}{\partial\beta}] - \frac{1}{2}\frac{\partial}{\partial\beta}\ln|\mathbf{H}| + \frac{n}{2\beta} \\
&= -\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) - [\beta\frac{\partial\hat{R}_n}{\partial\boldsymbol{\theta}}(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) + \lambda\frac{\partial\mathcal{E}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_{MP})]\frac{\partial\boldsymbol{\theta}_{MP}}{\partial\beta} - \frac{1}{2}\frac{\partial}{\partial\beta}\ln|\mathbf{H}| + \frac{n}{2\beta} \\
&= -\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) - \frac{1}{2}\frac{\partial}{\partial\beta}\ln|\mathbf{H}| + \frac{n}{2\beta}
\end{aligned}$$

obteniéndose:

$$2\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) = n - \sum_{j=1}^{p} \frac{\mu_j}{\mu_j+\lambda} = n - p' \tag{II.121}$$

Analizando conjuntamente (II.120) y (II.121), se ve que, por cada parámetro que los datos hacen significativo (cada uno de los contabilizados en p'), se reduce el término de error de los datos y se incrementa el *error* de los parámetros, en una unidad. Con ello, el número de grados de libertad de los datos, n, podría descomponerse como:

$$n = p' + (n - p') = \text{g.l.}[2\lambda\mathcal{E}(\boldsymbol{\theta}_{MP})] + \text{g.l.}[2\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP})]$$

es decir, el número de grados de libertad del modelo más el número de grados de libertad del error no explicado por el mismo.

Aprendizaje Bayesiano con Algoritmos tipo *Back-propagation*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

La implementación del enfoque anterior de la evidencia, debe determinar los valores óptimos λ_{ME} y β_{ME}, así como el valor más probable de los parámetro $\boldsymbol{\theta}_{MP}(\lambda_{ME}, \beta_{ME})$, para dichos valores.

Sin embargo, las expresiones (II.120) y (II.121) anteriores constituyen fórmulas circulares, pues la determinación de λ_{ME} y β_{ME} requiere previamente conocer el valor $\boldsymbol{\theta}_{MP}$ en dichas ecuaciones.

La solución a este problema reside en un algoritmo iterativo que avance alternativamente en λ, β y en $\boldsymbol{\theta}$. En esta táctica se basa la propuesta de [Foresee y Hagan 1997], siguiendo las sugerencias de [MacKay 1994], basada en el algoritmo de Levenberg-Marquardt, ([Marquardt 1963], Anexo D), cuya estructura es la siguiente:

1. Inicializar $\lambda_0 = 0$, $\beta_0 = 1$, y $\boldsymbol{\theta}$ mediante el método de [Nguyen y Widrow 1990] visto en el epígrafe de la página 161.

2. Iteración s-ésima, $s \geq 1$:
   (a) Iterar una vez el algoritmo de Levenberg-Marquardt en la minimización de $\hat{\mathcal{R}}_n(\boldsymbol{\theta})$, obteniendo $\boldsymbol{\theta}_s$. Este algoritmo utiliza como Hesiano:
   $$\mathbf{H} = \frac{\partial^2}{\partial \boldsymbol{\theta}^2}\hat{\mathcal{R}}_n \approx \beta[(\nabla_{\boldsymbol{\theta}}\hat{R}_n)(\nabla_{\boldsymbol{\theta}}\hat{R}_n)^T] + \lambda \mathbf{I}_p$$
   es decir, utiliza la aproximación de Gauss-Newton al Hesiano $\partial^2 \hat{R}_n / \partial \boldsymbol{\theta}^2$ del error correspondiente a los datos.
   (b) Calcular el número efectivo de parámetros mediante la ecuación (II.120). Si $\mathbf{H}_s = \mathbf{H}(\boldsymbol{\theta}_s)$:
   $$p'_s = p - \mathrm{tr}(\mathbf{H}_s^{-1})$$
   (c) Calcular nuevas estimaciones de los hiperparámetros a partir de (II.120) y (II.121):
   $$\lambda_s = \frac{p'_s}{2\mathcal{E}(\boldsymbol{\theta}_s)} \text{ y } \beta_s = \frac{n - p'_s}{2\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_s)}$$

3. Repetir el paso 2 hasta la convergencia.

Después de finalizar el algoritmo, si el número efectivo de parámetros final, p', es similar a p, puede ser que la red se haya saturado debido a una insuficiente complejidad (pocas unidades ocultas). En este caso, conviene repetir el algoritmo añadiendo más unidades ocultas. Si el número efectivo de parámetros no cambia es que la complejidad de la red anterior era suficiente.

Es importante notar, que no es necesario preocuparse por la complejidad de la red pues el algoritmo la selecciona automáticamente utilizando sólo el número de parámetros que realmente requieren los datos y penalizando los restantes. Ello es debido a que, el algoritmo, selecciona los hiperparámetros automáticamente y, en particular, el parámetro regularizador λ.

Nótese que, a diferencia de otros métodos como validación cruzada, este método produce resultados óptimos en una sóla pasada del algoritmo, sin necesidad de entrenar diversas redes con distintos valores del regularizador λ, (teniendo en cuenta que si este parámetro fuese vectorial, dichos métodos serían inviables). Adicionalmente, se obtiene una estimación de la varianza de los datos utilizando que $\sigma^2 = 1/(2\beta)$.

Aprendizaje Bayesiano con Algoritmos de Regresión Lineal Generalizada OLS

USC UNIVERSIDADE

En este apartado, utilizamos el método bayesiano para entrenar redes lineales en parámetros, conjuntamente con los algoritmos tratados en el Apartado II.6.2 que resuelven el problema regular-

izado (ecuación (II.98)):

$$\min_{\mathbf{c}\in\mathbb{R}^h}\{\hat{\mathcal{R}}_n(\mathbf{c}) = \|\mathbf{y}-\mathbf{\Psi c}\|^2 + \sum_{j=1}^{p}\lambda_j c_j^2\}$$

Para ello, consideramos redes de la forma (expresión (II.97)):

$$f(\mathbf{x};\mathbf{c}) = \sum_{j=1}^{h} c_j\psi_j(\mathbf{x})$$

donde $\{\psi_j\}$ es un conjunto prefijado de funciones básicas.

El objetivo es estimar los hiperparámetros λ_j, $j = 1,...,p$ mediante un método alternativo al proporcionado por el teorema 55 de [Orr 1996]. La estrategia consiste en aplicar el método bayesiano al problema ortogonalizado equivalente.

Realizando la descomposición QR (pág. 176) de la matriz de diseño, se obtiene:

$$\mathbf{\Psi} = \bar{\mathbf{\Psi}}\mathbf{U}$$

donde $\bar{\mathbf{\Psi}}$ es una matriz $n \times h$ cuyas columnas $\bar{\boldsymbol{\psi}}_j$ son vectores ortogonales que generan el mismo espacio que las columnas de $\mathbf{\Psi}$, verificando: $\bar{\mathbf{\Psi}}^T\bar{\mathbf{\Psi}} = \mathrm{diag}(\bar{\boldsymbol{\psi}}_1^T\bar{\boldsymbol{\psi}}_1,...,\bar{\boldsymbol{\psi}}_h^T\bar{\boldsymbol{\psi}}_h)^T$. Con ello, la regresión puede escribirse como (ecuación (II.106)):

$$\hat{\mathbf{y}} = \mathbf{\Psi}\hat{\mathbf{c}} = \bar{\mathbf{\Psi}}\mathbf{U}\hat{\mathbf{c}} = \bar{\mathbf{\Psi}}\hat{\mathbf{d}}$$

expresión que relaciona la solución del problema ortogonalizado con la solución del original mediante $\hat{\mathbf{d}} = \mathbf{U}\hat{\mathbf{c}}$. En relación con el Apartado II.6.2, $p = h$ es el número de parámetros y $\boldsymbol{\theta} = \mathbf{d}$ es el vector de parámetros del modelo ortogonalizado (notación que mantenemos por continuidad con el apartado anterior).

Considerando este problema ortogonalizado, escribimos la función objetivo en la forma:

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) = \left\|\mathbf{y}-\bar{\mathbf{\Psi}}\boldsymbol{\theta}\right\|^2 + \sum_{j=1}^{p}\bar{\lambda}_j\theta_j^2 \equiv \frac{\beta}{2}\left\|\mathbf{y}-\bar{\mathbf{\Psi}}\boldsymbol{\theta}\right\|^2 + \frac{1}{2}\sum_{j=1}^{p}\tau_j\theta_j^2 \qquad \text{(II.122)}$$

donde hemos multiplicado por $\beta/2$, donde β es una constante con el mismo cometido que en el apartado anterior, y hemos definido:

$$\tau_j \doteq \beta\bar{\lambda}_j \qquad \text{(II.123)}$$

Definiendo el estabilizador $\mathcal{E}_{\mathbf{\Upsilon}}(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\mathbf{\Upsilon}\boldsymbol{\theta}$ con $\mathbf{\Upsilon} = \mathrm{diag}(\boldsymbol{\tau})$ con $\boldsymbol{\tau} = (\tau_1,...,\tau_p)^T$, podemos escribir la ecuación anterior como:

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) = \frac{\beta}{2}\left\|\mathbf{y}-\bar{\mathbf{\Psi}}\boldsymbol{\theta}\right\|^2 + \frac{1}{2}\boldsymbol{\theta}^T\mathbf{\Upsilon}\boldsymbol{\theta} = \beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}) + \mathcal{E}_{\mathbf{\Upsilon}}(\boldsymbol{\theta}) \qquad \text{(II.124)}$$

En la notación bayesiana del apartado anterior, suponemos hiperparámetros $\beta, \boldsymbol{\tau} = (\tau_1,...,\tau_p)^T$ y distribuciones $\mathsf{p}_{\mathsf{Y}/\mathsf{X}}$ y $\mathsf{p}_{\boldsymbol{\theta}}$ de la forma:

$$\begin{aligned}\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta};\beta) &= \frac{1}{C_\beta}\exp\{-\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta})\}\\ \mathsf{p}(\boldsymbol{\theta}|\boldsymbol{\tau}) &= \frac{1}{C_{\boldsymbol{\tau}}}\exp\{-\boldsymbol{\theta}^T\mathbf{\Upsilon}\boldsymbol{\theta}\}\end{aligned}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Por tanto, suponemos implícitamente una distribución gausiana en los errores con parámetro de varianza $\sigma^2 = 1/\beta$ donde $C_\beta = (2\pi/\beta)^{n/2}$ es el factor normalizador de la gausiana. Igualmente, la distribución a priori de los parámetros $\mathsf{p}_{\boldsymbol{\theta}}$ es gausiana con matriz de covarianza $\boldsymbol{\Upsilon}^{-1}$ y factor normalizador:

$$C_{\boldsymbol{\tau}} = \left|\boldsymbol{\Upsilon}^{-1}\right|^{1/2}(2\pi)^{p/2} = |\boldsymbol{\Upsilon}|^{-1/2}(2\pi)^{p/2} = (\prod_{j=1}^{p}\tau_j)^{-1/2}(2\pi)^{p/2}$$

Con ello, la minimización de (II.124) equivale a maximizar la distribución a posteriori (ecuación (II.114)):

$$\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n,\beta,\boldsymbol{\tau}) = \frac{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta};\beta)\mathsf{p}(\boldsymbol{\theta}|\boldsymbol{\tau})}{\mathsf{p}(\mathbf{y}^n)} \propto \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta};\beta)\mathsf{p}(\boldsymbol{\theta}|\boldsymbol{\tau}) \tag{II.126}$$

La evidencia de β y $\boldsymbol{\tau}$ es (como en (II.115)):

$$\begin{aligned}
\mathsf{p}(\mathbf{y}^n|\beta,\boldsymbol{\tau}) &= \int \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta},\beta,\boldsymbol{\tau})\mathsf{p}(\boldsymbol{\theta}|\beta,\boldsymbol{\tau})d\boldsymbol{\theta} = \int \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta},\beta)\mathsf{p}(\boldsymbol{\theta}|\boldsymbol{\tau})d\boldsymbol{\theta} \\
&= \frac{1}{C_\beta}\frac{1}{C_{\boldsymbol{\tau}}}\int \exp\{-\beta\hat{R}(\mathbf{y}^n|\boldsymbol{\theta}) - \mathcal{E}_{\boldsymbol{\Upsilon}}(\boldsymbol{\theta})\}d\boldsymbol{\theta} \\
&= \frac{1}{C_\beta C_{\boldsymbol{\tau}}}\int \exp\{-\hat{\mathcal{R}}(\boldsymbol{\theta})\}d\boldsymbol{\theta} = \frac{C_{\beta,\boldsymbol{\tau}}}{C_\beta C_{\boldsymbol{\tau}}}
\end{aligned}$$

donde:

$$C_{\beta,\boldsymbol{\tau}} = \int \exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta})\}d\boldsymbol{\theta}$$

Suponiendo una sóla moda $\boldsymbol{\theta}_{MP}$ en $\hat{\mathcal{R}}_n(\boldsymbol{\theta})$ y desarrollando en serie de Taylor en un entorno de la misma, se obtiene una aproximación gausiana como en (II.116) que permite estimar el factor normalizador $C_{\beta,\boldsymbol{\tau}}$ mediante (ecuación (II.117)):

$$\begin{aligned}
\hat{C}_{\beta,\boldsymbol{\tau}} &= \exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP})\}\int \exp\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MP})^T\mathbf{H}((\boldsymbol{\theta}-\boldsymbol{\theta}_{MP})\}d\boldsymbol{\theta} \\
&= \exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP})\}(2\pi)^{p/2}\,|\mathbf{H}|^{-1/2}
\end{aligned}$$

donde $\mathbf{H} = \partial^2\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP})/\partial\boldsymbol{\theta}^2$ es el Hesiano. Con ello, el neperiano de la evidencia a maximizar es (ecuación (II.118)):

$$\begin{aligned}
\ln \mathsf{p}(\mathbf{y}^n|\beta,\boldsymbol{\tau}) &= \ln\frac{C_{\beta,\boldsymbol{\tau}}}{C_\beta C_{\boldsymbol{\tau}}} = \ln\frac{\exp\{-\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP})\}(2\pi)^{p/2}\,|\mathbf{H}|^{-1/2}}{(2\pi/\beta)^{n/2}(\prod_{j=1}^{p}\tau_j)^{-1/2}(2\pi)^{p/2}} \\
&= -\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) - \mathcal{E}_{\boldsymbol{\Upsilon}}(\boldsymbol{\theta}_{MP}) - \frac{1}{2}\ln|\mathbf{H}| - \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln\beta + \frac{1}{2}\sum_{j=1}^{p}\ln\tau_j
\end{aligned}$$

donde $\boldsymbol{\theta}_{MP} = \boldsymbol{\theta}_{MP}(\beta,\boldsymbol{\tau})$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Maximización de la Evidencia en la Varianza de los Datos

La condición de óptimo para β es:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\beta}\ln \mathsf{p}(\mathbf{y}^n|\beta,\lambda) \\
&= -\frac{\partial}{\partial\beta}[\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP})] - \frac{\partial}{\partial\beta}\mathcal{E}_{\boldsymbol{\Upsilon}}(\boldsymbol{\theta}_{MP})] - \frac{1}{2}\frac{\partial}{\partial\beta}\ln|\mathbf{H}| + \frac{n}{2\beta} \\
&= -[\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) + \beta\frac{\partial\hat{R}_n}{\partial\boldsymbol{\theta}}(\mathbf{y}^n|\boldsymbol{\theta}_{MP})\frac{\partial\boldsymbol{\theta}_{MP}}{\partial\beta} + \frac{\partial\mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_{MP})\frac{\partial\boldsymbol{\theta}_{MP}}{\partial\beta}] - \frac{1}{2}\frac{\partial}{\partial\beta}\ln|\mathbf{H}| + \frac{n}{2\beta} \\
&= -\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) - [\beta\frac{\partial\hat{R}_n}{\partial\boldsymbol{\theta}}(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) + \frac{\partial\mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_{MP})]\frac{\partial\boldsymbol{\theta}_{MP}}{\partial\beta} - \frac{1}{2}\frac{\partial}{\partial\beta}\ln|\mathbf{H}| + \frac{n}{2\beta} \\
&= -\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) - \frac{1}{2}\frac{\partial}{\partial\beta}\ln|\mathbf{H}| + \frac{n}{2\beta}
\end{aligned}$$

Pero, el Hesiano es:

$$\begin{aligned}
\mathbf{H} &= \frac{\partial^2\hat{\mathcal{R}}_n}{\partial\boldsymbol{\theta}^2}(\boldsymbol{\theta}_{MP}) = \beta\frac{\partial^2\hat{R}_n}{\partial\boldsymbol{\theta}^2}(\boldsymbol{\theta}_{MP}) + \frac{\partial^2\mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial\boldsymbol{\theta}^2}(\boldsymbol{\theta}_{MP}) = \beta\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}} + \Upsilon \\
&= \mathrm{diag}(\beta\bar{\boldsymbol{\psi}}_1^T\bar{\boldsymbol{\psi}}_1 + \tau_1, ..., \beta\bar{\boldsymbol{\psi}}_p^T\bar{\boldsymbol{\psi}}_p + \tau_p)
\end{aligned}$$

con lo que, utilizando (II.123), se tiene:

$$\ln|\mathbf{H}| = \sum_{j=1}^{p}\ln(\beta\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \tau_j) \Rightarrow \frac{\partial}{\partial\beta}\ln|\mathbf{H}| = \sum_{j=1}^{p}\frac{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j}{\beta\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \tau_j}$$

y así la condición de óptimo respecto a β es:

$$\begin{aligned}
0 &= -\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) - \frac{1}{2}\sum_{j=1}^{p}\frac{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j}{\beta\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \tau_j} + \frac{n}{2\beta} \\
&\Rightarrow \quad 2\beta\hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) = n - \sum_{j=1}^{p}\frac{\beta\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j}{\beta\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \tau_j} = n - p'
\end{aligned} \tag{II.127}$$

Esta ecuación es análoga a la ecuación (II.121) pues los autovalores de $\beta\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}}$ son los elementos de su diagonal $\mu_j = \beta\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j$, con lo que admite la misma interpretación que aquélla donde p' es el número equivalente de parámetros.

Maximización de la Evidencia en la Varianza de la Prior de los Parámetros

Teniendo en cuenta que $\mathcal{E}_{\boldsymbol{\Upsilon}}(\boldsymbol{\theta}_{MP}) = \frac{1}{2}\sum_{j=1}^{p}\tau_j(\boldsymbol{\theta}_{MP})_j^2 = \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\Upsilon}\boldsymbol{\theta}$, y $\boldsymbol{\theta}_{MP}$ verifica:

$$\mathbf{0} = \frac{\partial}{\partial\boldsymbol{\theta}}\hat{\mathcal{R}}_n(\boldsymbol{\theta}_{MP}) = \frac{\partial\hat{R}_n}{\partial\boldsymbol{\theta}}(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) + \frac{\partial\mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_{MP}) = 0$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

la condición de óptimo para cada τ_j es:

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\tau}} \ln \mathsf{p}(\mathbf{y}^n|\beta,\boldsymbol{\tau}) \\
&= -\frac{\partial}{\partial \boldsymbol{\tau}}[\beta \hat{R}_n(\mathbf{y}^n|\boldsymbol{\theta}_{MP})] - \frac{\partial}{\partial \boldsymbol{\tau}}[\mathcal{E}_{\boldsymbol{\Upsilon}}(\boldsymbol{\theta}_{MP})] - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\tau}} \ln|\mathbf{H}| + \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\tau}} \ln|\boldsymbol{\Upsilon}| \\
&= -\beta \frac{\partial \hat{R}_n}{\partial \boldsymbol{\theta}}(\mathbf{y}^n|\boldsymbol{\theta}_{MP})\frac{\partial \boldsymbol{\theta}_{MP}}{\partial \boldsymbol{\tau}} - [\frac{\partial \mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\theta}_{MP}) + \frac{\partial \mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_{MP})\frac{\partial \boldsymbol{\theta}_{MP}}{\partial \boldsymbol{\tau}}] - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\tau}} \ln|\mathbf{H}| + \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\tau}} \ln|\boldsymbol{\Upsilon}| \\
&= -\frac{\partial \mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\theta}_{MP}) - [\beta \frac{\partial \hat{R}_n}{\partial \boldsymbol{\theta}}(\mathbf{y}^n|\boldsymbol{\theta}_{MP}) + \frac{\partial \mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_{MP})]\frac{\partial \boldsymbol{\theta}_{MP}}{\partial \boldsymbol{\tau}} - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\tau}} \ln|\mathbf{H}| + \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\tau}} \ln|\boldsymbol{\Upsilon}| \\
&= -\frac{\partial \mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial \boldsymbol{\tau}}(\boldsymbol{\theta}_{MP}) - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\tau}} \ln|\mathbf{H}| + \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\tau}} \ln|\boldsymbol{\Upsilon}|
\end{aligned}$$

Pero:

$$\begin{cases}
\dfrac{\partial \mathcal{E}_{\boldsymbol{\Upsilon}}}{\partial \tau_j}(\boldsymbol{\theta}_{MP}) = \frac{1}{2}(\boldsymbol{\theta}_{MP})_j^2 \\
\dfrac{\partial}{\partial \tau_j} \ln|\mathbf{H}| = \dfrac{1}{(\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \tau_j)} = \dfrac{1}{\beta(\bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \lambda_j)} \\
\dfrac{\partial}{\partial \tau_j} \ln|\boldsymbol{\Upsilon}| = \dfrac{\partial}{\partial \tau_j} \sum_{j=1}^{p} \ln \tau_j = \dfrac{1}{\tau_j}
\end{cases}$$

Por tanto, para cada τ_j, la condición anterior es:

$$\begin{aligned}
0 &= -\frac{1}{2}(\boldsymbol{\theta}_{MP})_j^2 - \frac{1}{2(\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \tau_j)} + \frac{1}{2\tau_j} \\
\Rightarrow\ (\boldsymbol{\theta}_{MP})_j^2 \tau_j &= -\frac{\tau_j}{\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \tau_j} + 1 = \frac{-\tau_j + \beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \tau_j}{\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \tau_j} = \frac{\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j}{\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \tau_j}
\end{aligned} \tag{II.128}$$

condición que, por la derecha, puede interpretarse como la aportación del parámetro $(\boldsymbol{\theta}_{MP})_j$ al número efectivo de parámetros p' de la ecuación (II.127) y, por la izquierda, el efecto de su parámetro regularizador específico τ_j, (compárese con la ecuación (II.120) donde allí se tiene un único regularizador que afecta a todos los parámetros).

Para obtener las ecuaciones en términos de los regularizadores iniciales $\bar{\lambda}_1, ..., \bar{\lambda}_p$ en (II.122), basta utilizar las ecuaciones (II.127) y (II.128) en la relación (II.123), teniendo en cuenta que $\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \tau_j = \beta(\bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j)$:

$$\begin{aligned}
\bar{\lambda}_j &= \frac{\tau_j}{\beta} = \frac{2\hat{R}(\mathbf{y}^n|\boldsymbol{\theta}_{MP})}{(\boldsymbol{\theta}_{MP})_j^2} \frac{\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j / (\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \tau_j)}{n - \sum_{l=1}^{p} \beta \bar{\boldsymbol{\psi}}_l^T \bar{\boldsymbol{\psi}}_l / (\beta \bar{\boldsymbol{\psi}}_l^T \bar{\boldsymbol{\psi}}_l + \tau_l)} \\
&= \frac{\mathbf{e}^T \mathbf{e}}{(\boldsymbol{\theta}_{MP})_j^2} \frac{\gamma_j}{n - \sum_{l=1}^{p} \gamma_l},\ j = 1:p
\end{aligned} \tag{II.129}$$

donde $\mathbf{e} \doteq \mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\theta}_{MP}$ son los errores en la muestra y:

$$\gamma_j \doteq \frac{\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j}{\beta \bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \tau_j} = \frac{\bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j}{\bar{\boldsymbol{\psi}}_j^T \bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j} \tag{II.130}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Un caso particular de lo anterior, es cuando se considera un sólo regularizador $\bar{\lambda}$, en cuyo caso, se obtienen las siguientes ecuaciones:

$$\bar{\lambda} = \frac{\mathbf{e}^T\mathbf{e}}{\boldsymbol{\theta}_{MP}^T\boldsymbol{\theta}_{MP}}\frac{\gamma}{n-\gamma} \text{ con } \gamma = \sum_{j=1}^{p}\frac{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j}{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \bar{\lambda}}$$

Número Efectivo de Parámetros y Medida de Suavizado

La cantidad (II.130) es el valor efectivo de cada parámetro, es decir, la contribución de cada parámetro al número efectivo de parámetros del modelo: $p' = \sum_{j=1}^{p}\gamma_j$.

Este número equivalente de parámetros p', coincide con el concepto homónimo correspondiente al carácter de suavizador lineal de la solución (p. ej. [Hastie y Tibshirani 1990]), pues, al ser $\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}} = \mathrm{diag}(\bar{\boldsymbol{\psi}}_1^T\bar{\boldsymbol{\psi}}_1, ..., \bar{\boldsymbol{\psi}}_p^T\bar{\boldsymbol{\psi}}_p)$, resulta:

$$\mathrm{tr}(\mathbf{S}_{\bar{\boldsymbol{\Psi}}}) = \mathrm{tr}[\bar{\boldsymbol{\Psi}}(\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}} + \bar{\boldsymbol{\Lambda}})^{-1}\bar{\boldsymbol{\Psi}}^T] = \mathrm{tr}[\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}}(\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}} + \bar{\boldsymbol{\Lambda}})^{-1}] = \sum_{j=1}^{p}\frac{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j}{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j} = p'$$

Además, se tiene el siguiente teorema:

Teorema 57 *Sea* $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ *con* $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p)^T$ *la matriz regularizadora del problema original (sin ortogonalizar):*

$$\hat{\mathcal{R}}_n(\mathbf{c}) = \|\mathbf{y} - \boldsymbol{\Psi}\mathbf{c}\|^2 + \sum_{j=1}^{p}\lambda_j c_j^2 = \|\mathbf{y} - \boldsymbol{\Psi}\mathbf{c}\|^2 + \frac{1}{2}\mathbf{c}^T\boldsymbol{\Lambda}\mathbf{c}$$

y sea $\bar{\boldsymbol{\Lambda}} = \mathrm{diag}(\bar{\boldsymbol{\lambda}})$ *la matriz del problema ortogonalizado. Si* $\boldsymbol{\Lambda}$ *verifica* $\boldsymbol{\Lambda} = \mathbf{U}^T\bar{\boldsymbol{\Lambda}}\mathbf{U}$ *siendo* $\mathbf{U}$ *la matriz triangular superior de la descomposición QR* $\boldsymbol{\Psi} = \bar{\boldsymbol{\Psi}}\mathbf{U}$. *Entonces, el número equivalente de parámetros de ambos problemas es el mismo:*

$$p' = \sum_{j=1}^{p}\frac{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j}{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j}$$

Demostración. Si $\boldsymbol{\Psi} = \bar{\boldsymbol{\Psi}}\mathbf{U}$ la relación entre el problema original y el ortogonalizado es:

$$\begin{aligned}\mathbf{y} &= \bar{\boldsymbol{\Psi}}\boldsymbol{\theta} + \boldsymbol{\varepsilon} = \boldsymbol{\Psi}\mathbf{U}^{-1}\boldsymbol{\theta} + \boldsymbol{\varepsilon} = \boldsymbol{\Psi}\mathbf{c} + \boldsymbol{\varepsilon}, \text{ con } \mathbf{c} = \mathbf{U}^{-1}\boldsymbol{\theta} \Rightarrow \boldsymbol{\theta} = \mathbf{U}\mathbf{c} \\ \Rightarrow &\quad \left\|\mathbf{y} - \bar{\boldsymbol{\Psi}}\boldsymbol{\theta}\right\|^2 + \frac{1}{2}\boldsymbol{\theta}^T\bar{\boldsymbol{\Lambda}}\boldsymbol{\theta} = \|\mathbf{y} - \boldsymbol{\Psi}\mathbf{c}\|^2 + \frac{1}{2}\mathbf{c}^T\mathbf{U}^T\bar{\boldsymbol{\Lambda}}\mathbf{U}\mathbf{c}\end{aligned}$$

que son el mismo problema si $\boldsymbol{\Lambda} = \mathbf{U}^T\bar{\boldsymbol{\Lambda}}\mathbf{U}$. Con ello, las matrices *hat* son las mismas:

$$\begin{aligned}S_{\boldsymbol{\Psi}} &= \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\boldsymbol{\Psi} + \boldsymbol{\Lambda})^{-1}\boldsymbol{\Psi}^T = \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\boldsymbol{\Psi} + \mathbf{U}^T\bar{\boldsymbol{\Lambda}}\mathbf{U})^{-1}\boldsymbol{\Psi}^T \\ &= \bar{\boldsymbol{\Psi}}\mathbf{U}(\mathbf{U}^T\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}}\mathbf{U} + \mathbf{U}^T\bar{\boldsymbol{\Lambda}}\mathbf{U})^{-1}\mathbf{U}^T\bar{\boldsymbol{\Psi}}^T \\ &= \bar{\boldsymbol{\Psi}}\mathbf{U}[\mathbf{U}^T(\boldsymbol{\Psi}^T\bar{\boldsymbol{\Psi}} + \bar{\boldsymbol{\Lambda}})^{-1}\mathbf{U}]^{-1}\mathbf{U}^T\bar{\boldsymbol{\Psi}}^T \\ &= \bar{\boldsymbol{\Psi}}\mathbf{U}\mathbf{U}^{-1}(\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}} + \bar{\boldsymbol{\Lambda}})^{-1}(\mathbf{U}^T)^{-1}\mathbf{U}^T\bar{\boldsymbol{\Psi}}^T \\ &= \bar{\boldsymbol{\Psi}}(\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}} + \bar{\boldsymbol{\Lambda}})^{-1}\bar{\boldsymbol{\Psi}}^T = \mathbf{S}_{\bar{\boldsymbol{\Psi}}}\end{aligned}$$

■

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Algoritmo OLS Bayesiano

Al igual que en el algoritmo bayesiano tipo *back-propagation*, las ecuaciones (II.127) y (II.128) son circulares pues dependen del vector $\boldsymbol{\theta}_{MP} = \boldsymbol{\theta}_{MP}(\beta, \boldsymbol{\tau})$ que, a su vez, depende de los hiperparámetros. Lo mismo sucede con la ecuación (II.129) que aglutina β y τ_j en términos de $\bar{\lambda}_j$.

Como en aquél caso, la solución se consigue mediante un algoritmo que avance alternativamente en $\bar{\boldsymbol{\lambda}} = (\bar{\lambda}_1, ..., \bar{\lambda}_p)^T$ y en $\boldsymbol{\theta}_{MP}$ (o en $(\beta, \boldsymbol{\tau})$ y en $\boldsymbol{\theta}_{MP}$) de la siguiente forma [Chen 2002*a*]:

1. Inicializar $\bar{\boldsymbol{\lambda}}_0 = (\bar{\lambda}_1^{(0)}, ..., \bar{\lambda}_p^{(0)})^T$ con valores positivos pequeños en cada componente.
2. Para $s = 1, 2, ...$:
   (a) Resolver la regresión OLS regularizada con los valores de $\bar{\boldsymbol{\lambda}}^{(s-1)}$ con el criterio de selección de variables que se describe a más adelante (ecuación (II.131)).
   (b) Obtener $\bar{\boldsymbol{\lambda}}_s$ mediante las ecuaciones (II.129), $j = 1, ..., h_s$ donde h_s es el número de funciones básicas introducidas en 2.a.
   (c) Finalizar cuando los valores de $\bar{\boldsymbol{\lambda}}$ se estabilicen.

En el óptimo $\hat{\boldsymbol{\theta}} = (\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}})^{-1}\bar{\boldsymbol{\Psi}}^T\mathbf{y}$, la función objetivo del problema de regresión ortogonal regularizado es:

$$\begin{aligned}\hat{\mathcal{R}}_n(\hat{\boldsymbol{\theta}}) &= (\mathbf{y} - \bar{\boldsymbol{\Psi}}\hat{\boldsymbol{\theta}})^T(\mathbf{y} - \bar{\boldsymbol{\Psi}}\hat{\boldsymbol{\theta}}) + \hat{\boldsymbol{\theta}}^T\boldsymbol{\Upsilon}\hat{\boldsymbol{\theta}} = \mathbf{y}^T(\mathbf{y} - \bar{\boldsymbol{\Psi}}\hat{\boldsymbol{\theta}}) + \hat{\boldsymbol{\theta}}^T\boldsymbol{\Upsilon}\hat{\boldsymbol{\theta}} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\bar{\boldsymbol{\Psi}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}^T\boldsymbol{\Upsilon}\hat{\boldsymbol{\theta}} \\ &= \mathbf{y}^T\mathbf{y} - (\bar{\boldsymbol{\Psi}}\hat{\boldsymbol{\theta}} + \mathbf{e})^T\bar{\boldsymbol{\Psi}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}^T\boldsymbol{\Upsilon}\hat{\boldsymbol{\theta}} = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}^T\boldsymbol{\Upsilon}\hat{\boldsymbol{\theta}} \\ &= \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T(\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}} + \boldsymbol{\Upsilon})\hat{\boldsymbol{\theta}} = \mathbf{y}^T\mathbf{y} - \sum_{j=1}^{p}(\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j)\hat{\theta}_j^2\end{aligned}$$

con lo que la disminución en el error producida por la introducción de la variable j-ésima es (compárese con la ecuación (II.108)):

$$\Delta_j = \hat{\mathcal{R}}_{j-1} - \hat{\mathcal{R}}_j = (\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j)\hat{\theta}_j^2 = \frac{(\mathbf{y}^T\bar{\boldsymbol{\psi}}_j)^2}{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j}$$

pues $\hat{\boldsymbol{\theta}} = (\bar{\boldsymbol{\Psi}}^T\bar{\boldsymbol{\Psi}})^{-1}\bar{\boldsymbol{\Psi}}^T\mathbf{y} \Rightarrow \hat{\theta}_j = \mathbf{y}^T\bar{\boldsymbol{\psi}}_j/(\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j)$, (ecuación (II.105)). El criterio de parada propuesto por [Chen 2002*a*] es análogo a (II.107):

$$\frac{1}{\mathbf{y}^T\mathbf{y}}(\mathbf{y}^T\mathbf{y} - \sum_{j=1}^{p}\frac{(\mathbf{y}^T\bar{\boldsymbol{\psi}}_j)^2}{\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j}) = 1 - \sum_{j=1}^{p}\frac{(\mathbf{y}^T\bar{\boldsymbol{\psi}}_j)^2}{(\bar{\boldsymbol{\psi}}_j^T\bar{\boldsymbol{\psi}}_j + \bar{\lambda}_j)(\mathbf{y}^T\mathbf{y})} < \delta \qquad \text{(II.131)}$$

para algún $\delta > 0$ pequeño.

El esquema del algoritmo se muestra en la tabla 11. En las primeras iteraciones, [Chen 2002*a*] recomienda que el paso 2.1.b.iv se realice minimizando el error al máximo con el fin de obtener un modelo con más variables de las necesarias y, así, no excluir a priori variables que puedan ser representativas. Más adelante, cuando $\bar{\boldsymbol{\lambda}}$ se estabiliza, conviene utilizar un criterio de parada más holgado.

En nuestra experiencia, el algoritmo es excesivamente sensible a los valores iniciales del vector regularizador $\bar{\boldsymbol{\lambda}}$. Esto no suele suceder cuando se utiliza un mismo regularizador para todas las funciones básicas, en cuyo caso, se obtienen resultados similares a los obtenidos en el apartado

Algoritmo de Entrenamiento RBF Bayesiano

1. Inicio: $\bar{\boldsymbol{\lambda}}_0 = (\bar{\lambda}_1^{(0)}, ..., \bar{\lambda}_p^{(0)})^T$ y centrar las variables $y, \psi_j, j = 1 : h$
2. Para $t = 1, 2, ..$ hasta que $\bar{\boldsymbol{\lambda}}$ se estabilice:
   2.1 Regresión lineal mediante OLS con las variables $y, \psi_j, j = 1 : h_\sigma$ centradas
   a) $\mathcal{N}_0 = \{1 : n\}$, $\boldsymbol{\psi}_j^{(0)} = \boldsymbol{\psi}_j, j = 1 : h_\sigma$
   b) Para $s = 1, 2, ..., h_\sigma$, calcular:
   i) $\bar{\boldsymbol{\psi}}_s = \bar{\boldsymbol{\psi}}_{j_s}$ con $j_s = \arg \min_{j \in \mathcal{N}_{s-1}} \frac{\left\langle \mathbf{y}, \boldsymbol{\psi}_j^{(s-1)} \right\rangle^2}{\left\langle \boldsymbol{\psi}_j^{(s-1)}, \boldsymbol{\psi}_j^{(s-1)} \right\rangle}$ y $\hat{d}_s = \frac{\langle \mathbf{y}, \bar{\boldsymbol{\psi}}_s \rangle}{\langle \bar{\boldsymbol{\psi}}_s, \bar{\boldsymbol{\psi}}_s \rangle}$
   ii) $\mathcal{N}_s = \mathcal{N}_{s-1} - \{j_s\}$
   iii) $\boldsymbol{\psi}_j^{(s)} = \boldsymbol{\psi}_j^{(s-1)} - \frac{\left\langle \boldsymbol{\psi}_j^{(s-1)}, \bar{\boldsymbol{\psi}}_s \right\rangle}{\langle \bar{\boldsymbol{\psi}}_s, \bar{\boldsymbol{\psi}}_s \rangle}, \forall j \in \mathcal{N}_s$
   iv) Finalizar si $1 - \sum_{j=1}^{s} \frac{\langle \mathbf{y}, \bar{\boldsymbol{\psi}}_j \rangle^2}{[\langle \bar{\boldsymbol{\psi}}_j, \bar{\boldsymbol{\psi}}_j \rangle + \bar{\lambda}_j^{(t-1)}] \langle \mathbf{y}, \mathbf{y} \rangle} < \delta$ con $\delta > 0$
   c) Calcular: $\hat{\mathbf{c}} = \mathbf{U}^{-1}\hat{\mathbf{d}}$, con $(\mathbf{U})_{sj} = \frac{\langle \boldsymbol{\psi}_j, \bar{\boldsymbol{\psi}}_s \rangle}{\langle \bar{\boldsymbol{\psi}}_s, \bar{\boldsymbol{\psi}}_s \rangle}$, $s \leq j$, $(\mathbf{U})_{sj} = 0$, $s > j$, $s, j \in \{1 : h_\sigma\}$
   2.2 Calcular: $\bar{\lambda}_j^{(t)} = \frac{\mathbf{e}^T\mathbf{e}}{(\boldsymbol{\theta}_{MP}^{(t)})_j^2} \frac{\gamma_j}{n - \sum_{t=1}^{p} \gamma_t}$ con $\mathbf{e} = \mathbf{y} - \bar{\boldsymbol{\Psi}}\boldsymbol{\theta}^{(t)}$, $\gamma_j = \frac{\langle \bar{\boldsymbol{\psi}}_j, \bar{\boldsymbol{\psi}}_j \rangle}{\langle \bar{\boldsymbol{\psi}}_j, \bar{\boldsymbol{\psi}}_j \rangle + \bar{\lambda}_j^{(t-1)}}$, $j = 1, ..., p$

Tabla 11. Algoritmo de entrenamiento de la red RBF mediante el método bayesiano. Por simplicidad, se obvia el subíndice t de los vectores en cada etapa OLS

cuando la selección se realiza mediante GCV según el teorema 54, sin embargo, la red resultante es mucho más parsimoniosa.

La regularización diferenciada con redes RBF produce una regularización de tipo *local*, por el propio carácter de las funciones básicas. Al igual que en el algoritmo bayesiano del tipo *back-propagation*, el método bayesiano elimina la necesidad de realizar la selección del modelo en relación con el número de términos de la expansión (número de unidades ocultas), pues el valor óptimo del regularizador correspondiente a cada función básica, juega el papel de barrera a la entrada efectiva de cada una. Cuanto mayor sea el regularizador menos relevante será su función básica (su centro) para el modelo.

La aplicación del método bayesiano a la regresión OLS con un sólo regularizador se menciona primero en [Orr 1995] que produce los mismos resultados que los que se obtienen aplicando regularización a OLS (epígrafe de la pág. 179) y resulta un caso particular de lo anterior cuando todos los regularizadores son iguales. Sin embargo, como se ha indicado allí, al ser una regularización no selectiva, el número de funciones básicas tiende a incrementarse perdiéndose parsimonia e interpretabilidad en el modelo.

La aplicación del método bayesiano a la regresión OLS con diferentes regularizadores se realiza por primera vez por [Chen 2002*a*] y [Chen 2002*b*], aunque sin el análisis realizado anteriormente sobre el grado de suavizado. Nótese que la técnica de [Orr 1996] para el mismo problema, descrita en el Apartado II.6.2, se realiza con la regresión original en lugar de con la ortogonalizada. Además, dicha técnica es excesivamente costosa computacionalmente y no resulta muy eficaz en la consecución de la parsimonia deseada.

Finalmente, es importante destacar que, la técnica bayesiana descrita, no resuelve el problema de la selección del hiperparámetro de escala de las funciones básicas en las redes RBF, suponiéndose éste prefijado en todo el desarrollo anterior. Realmente, puede decirse que, la aplicación del enfoque bayesiano a la selección de este parámetro, es todavía un problema abierto sujeto a investigación.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

II.6.6 Algoritmos Greedy

Otro conjunto de algoritmos para el entrenamiento de las redes neuronales *feed-forward* lo constituyen los agrupados bajo la denominación de algoritmos *greedy* (*greedy algorithms*).

Estos algoritmos surgen al albor de los métodos de regresión *projection pursuit* ([Friedman y Stuetzle 1981], [Diaconis y Shahshahani 1984], [Huber 1985], [Friedman 1987]), se fundamentan en las propiedades de convergencia aportadas por [Jones 1987], [Jones 1992], [Chen 1991] y [Jones 2000], entre otros, y su aplicación a las redes neuronales MLP puede tener su origen en [Barron y Barron 1988] y, posteriormente, en [Jones 1992] motivado por [Barron 1993].

Dichos algoritmos se basan en demostraciones constructivas al lema que mostramos a continuación, atribuido a Maurey [Barron 1993], que determina un orden de convergencia $O(h^{-1/2})$ para la aproximación mediante combinaciones convexas de funciones de una familia $\mathcal{G}$ acotada ($\|\varphi\|_{\mathcal{K}} \leq M < \infty, \forall \varphi \in \mathcal{G}$) en un espacio de Hilbert $\mathcal{K}$.

Sea $\mathcal{T}_{\mathcal{G}} = \mathrm{cl}(\mathrm{co}(\mathcal{G})) \subset \mathcal{K}$ la clausura del conjunto de las combinaciones convexas de elementos de $\mathcal{G}$ y por $\mathcal{F}_h$, $h \in \mathbb{N}^+$ el conjunto de las combinaciones convexas de h elementos de $\mathcal{G}$.

Lema 58 *(Maurey) [Barron 1993] Sea $\mathcal{K}$ un espacio de Hilbert con producto interno $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ y norma $\|\cdot\|_{\mathcal{K}}$. Sea $\mathcal{G} \subset \mathcal{K}$ con $\|\varphi\|_{\mathcal{K}} \leq b \; \forall \varphi \in \mathcal{G}$ y sean $\mathcal{F}$, $\mathcal{F}_h \; \forall h \in \mathbb{N}^+$ definidos como anteriormente. Entonces, dada $\mathsf{f} \in \mathcal{T}_{\mathcal{G}}$, y fijado $h \in \mathbb{N}^+$ se verifica que, para toda constante $C > b_{\mathsf{f}}^2 \doteq b^2 - \|\mathsf{f}\|_{\mathcal{K}}$, existe un $f_h \in \mathcal{F}_h$ tal que:*

$$\|f_h - \mathsf{f}\|_{\mathcal{F}}^2 \leq \frac{C}{h}$$

[Jones 1987] aporta la primera demostración constructiva a este resultado, que se mejoran o amplían posteriormente en [Jones 1992], [Barron 1993], [Dingankar y Sandberg 1996], [Lee *et al.* 1996, th. 2], [Docampo *et al.* 1997], [Fine 1999, th. 4.11.1], [Zhang 2002]:

Teorema 59 *[Docampo* et al. *1997, th. 2.1] En el marco del teorema anterior, sea $\mathsf{f} \in \mathcal{T}_{\mathcal{G}}$ y sea:*

$$\gamma = \inf_{g \in \mathcal{K}} \sup_{\varphi \in \mathcal{G}} \{\|\varphi - g\|_{\mathcal{K}}^2 - \|\mathsf{f} - g\|_{\mathcal{K}}^2\} \tag{II.132}$$

Sea C una constante tal que $C > \gamma$. Entonces, dado $\mathsf{f} \in \mathcal{T}_{\mathcal{G}}$, es posible construir de forma iterativa una sucesión $\{f_h\}$ donde $f_h = (1 - \beta_h) f_{h-1} + \beta_h \varphi_h$ con $\beta_h \in [0, 1]$, $\varphi_h \in \mathcal{G}$, $h \in \mathbb{N}^+$ tal que[73]*:*

$$\|f_h - \mathsf{f}\|_{\mathcal{K}}^2 \leq \inf_{\varphi \in \mathcal{G}} \|(1 - \beta_h) f_{h-1} + \beta_h \varphi - \mathsf{f}\|_{\mathcal{K}}^2 + \varepsilon_h \tag{II.133}$$

de tal forma que:

$$\|f_h - \mathsf{f}\|_{\mathcal{K}}^2 \leq \frac{C}{h}$$

Como parámetro γ para elegir C basta tomar el parámetro b_{f}^2 del lema de Maurey pues, tomando $g = 0 \in \mathcal{K}$, la condición (C.11) resulta:

$$\gamma \leq \sup_{\varphi \in \mathcal{G}} \{\|\varphi\|_{\mathcal{K}}^2 - \|\mathsf{f}\|_{\mathcal{K}}^2\} \leq b^2 - \|\mathsf{f}\|_{\mathcal{K}} = b_{\mathsf{f}}^2$$

[73] Esta condición supone que $f_h \in \mathrm{co}(\{f_{h-1}, \psi_h\})$ y, por tanto, es más restrictiva que la original $f_{h-1} \in \mathrm{co}(\{\psi_j\}_{j=1}^{h-1})$ y $f_h \in \mathrm{co}(\{\psi_j\}_{j=1}^{h})$ del teorema [Docampo *et al.* 1997, th. 2.1] y otros similares [Jones 1992], pero para ella se obtienen las principales conclusiones y es frecuente en la práctica.

El teorema anterior es la versión *relajada* [Jones 1992] del algoritmo *greedy*. Casos particulares son las versiones de [Barron 1993] con $\beta_h = b_{\mathsf{f}}^2/(b_{\mathsf{f}}^2 + \|f_{h-1} - \mathsf{f}\|_{\mathcal{K}}^2)$, [Dingankar y Sandberg 1996] con $\beta_h = \frac{1}{h}$, y [Lee *et al.* 1996, th. 2] y [Fine 1999, th. 4.11.1] con $\beta_h = \frac{2}{h}$.

[Docampo *et al.* 1997] analizan extensamente el algoritmo para cotas más generales del tipo: $\|f_h - \mathsf{f}\|_{\mathcal{K}}^2 \leq C/b(h)$ obteniendo resultados sobre la limitación en su tasa de convergencia y el error máximo ε_h que puede permitirse en cada iteración (II.133), (véase pág. 376 y sgtes. en Anexo C).

[Lee *et al.* 1996, th. 2] y [Fine 1999, th. 4.11.1] extienden el teorema al caso más general en el que $\mathsf{f} \in \mathcal{K}$, es decir, que f pueda no ser no aproximable por funciones de $\cup_{h=1}^{\infty}\mathcal{F}_h$. En este caso, se obtiene fácilmente una cota del tipo:

$$\|f_h - \mathsf{f}\|_{\mathcal{K}}^2 - d_{\mathsf{f}}^2 \leq \frac{C}{h}$$
$$\text{con: } d_{\mathsf{f}} = \inf_{f \in \mathcal{T}_{\mathcal{G}}} \|f - \mathsf{f}\|_{\mathcal{K}}$$

[Donahue *et al.* 1997] obtienen resultados del tipo anterior para espacios generales de Banach y, en particular, para $L_p(\mathcal{X})$ con $1 \leq p < \infty$, con el caso $p = 2$, como caso particular (teorema 177, Anexo C). Finalmente, [Zhang 2002] extiende los resultados a espacios lineales (no necesariamente de Hilbert) utilizando como medida de la aproximación una función de riesgo R convexa, derivando una cota a la tasa de aproximación basada en el Hesiano de $R(\cdot)$.

Las versiones iniciales del teorema anterior estaban destinadas a suministrar pruebas de convergencia de los métodos *projection pursuit* y, posteriormente, de los algoritmos *greedy* aplicados a las redes MLP, por lo que consideraban conjuntos $\mathcal{G} = \{\varphi_j\}$ con funciones de la forma $\varphi_{\mathbf{w}}(\mathbf{x}) \equiv \psi(\mathbf{w}^T\mathbf{x})$, ([Jones 1987], [Jones 1992]). Pero la formulación del teorema anterior muestra su aplicación general a todas las arquitecturas MLF vistas anteriormente pues, cualquier función implementable mediante una red MLF, puede ser expresada en términos de una combinación convexa de funciones de interés:

$$f(\mathbf{x}) = \sum_{j=1}^{h} c_j\phi_j(\mathbf{x}) + c_0 = \sum_{j=0}^{h} c_j\phi_j(\mathbf{x}) = \sum_{j=0}^{h} \beta_j\varphi_j(\mathbf{x})$$

con $\phi_0(\mathbf{x}) = 1$, y donde $\beta_j = |c_j|\,/C$ con $C = \sum_{j=0}^{h} c_j$, y $\varphi_j(\mathbf{x}) = C\,\text{signo}(c_j)\phi_j(\mathbf{x})$, $\forall \mathbf{x} \in \mathcal{X}$.

Con ello, basta considerar la familia $\mathcal{G}_C = \{\varphi_{c,\mathbf{w}} = c\psi(\mathbf{x};\mathbf{w}) : |c| \leq C\}$ en algún espacio de Hilbert, por ejemplo $L_2(\mathsf{P}_{\mathsf{X}})$ (pérdida cuadrática) donde P_{X} es la distribución del vector X de variables regresoras, para cuyos elementos se verifica $\left\|\varphi_{c,\mathbf{w}}\right\|_{\mathcal{F}} \leq \gamma$ para algún γ, si, por ejemplo, se considera un espacio de entrada compacto $\mathcal{X} \subset \mathbb{R}^d$.

A veces (p. ej. [Chen 1991]), se denominan *greedy* a ciertos algoritmos que determinan todas las funciones $\{\varphi_s\}$ a la vez realizando una búsqueda simultánea (global) de todos los parámetros respectivos (al modo de *back-propagation*), sin embargo, como ya hemos indicado, la denominación *greedy* se asigna tradicionalmente a los algoritmos basados en el lema 58 de Maurey que, con una filosofía *forward* de selección de variables, introducen una nueva $\varphi_s \in \mathcal{G}$ en cada iteración, sin considerar sus efectos posteriores en la optimalidad del conjunto.

De los diferentes algoritmos *greedy* recogidos en la literatura, destacamos los siguientes, ([Docampo *et al.* 1997], [DeVore y Temlyakov 1996], [Chen 1991]):

1. Algoritmo **greedy puro** [Jones 1987] cuya sucesión aproximadora se construye de la forma $f_j = f_{j-1} + \varphi_j$. Este algoritmo se corresponde con el método *stage-wise* de la regresión por

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Algoritmo Matching Pursuit Ortogonal

1. Inicio: $\mathsf{f}_0 = 0$, $R_0\mathsf{f} = \mathsf{f}$
2. Para $s = 1, 2, ...$ hasta que $|\langle R_{s-1}\mathsf{f}, \psi_s\rangle| < \delta > 0$
   - 2.1 Determinar $\psi_s \in \mathcal{G}$ tal que:
     $$|\langle R_{s-1}\mathsf{f}, \psi_s\rangle| \geq \alpha \sup_{\boldsymbol{\gamma}\in\Gamma} \left|\left\langle R_{s-1}\mathsf{f}, \psi_{\boldsymbol{\gamma}}\right\rangle\right| \text{ con } 0 < \alpha \leq 1$$
   - 2.2 Calcular $\bar{\psi}_s$, $\{\tau_j\}_{j=1}^{s-1}$, resolviendo:
     $$\psi_s = \sum_{j=1}^{s-1} \tau_j \psi_j + \bar{\psi}_s \text{ con } \left\langle \bar{\psi}_s, \psi_j \right\rangle = 0,\ j = 1, ..., s-1$$
   - 2.3 Calcular:
     $$\alpha_s^{(s)} = \frac{\langle R_{s-1}\mathsf{f}, \psi_s\rangle}{\left\|\bar{\psi}_s\right\|^2}$$
     $$\alpha_j^{(s)} = a_j^{(s-1)} - \frac{\langle R_s\mathsf{f}, \psi_s\rangle}{\left\|\bar{\psi}_s\right\|^2}\tau_j,\ j = 1, ..., s-1$$
   - 2.4 Actualizar:
     $$f_s = \sum_{j=1}^{s} \alpha_j^{(s)} \psi_j$$
     $$R_s\mathsf{f} = \mathsf{f} - f_s$$

Tabla 12. Algoritmo matching pursuit ortogonal.

pasos y es utilizado por ejemplo en [Friedman y Stuetzle 1981]. En el ámbito del proceso de señales, este algoritmo es el denominado **Matching Pursuit** (MP) de [Mallat y Zhang 1992], y se utiliza con familias $\mathcal{G}$ denominadas diccionarios (pág. 387, Anexo C).

2. Algoritmo(s) **greedy relajado(s)** (*relaxed*) como el del teorema 59, que mejoran la tasa de convergencia de los anteriores [Jones 1992]. Se utilizan por ejemplo en [Friedman *et al.* 1983].

3. Algoritmos **greedy con back-fitting**. El problema de los algoritmos *greedy puro* y *greedy* relajado, es que los residuos obtenidos al finalizar el algoritmo (y en cada iteración) $\mathbf{e}_j$ no son ortogonales al espacio $\left\langle \{\boldsymbol{\varphi}_s\}_{s=1}^{j} \right\rangle$. Con ello, el estimador f_j resultante no es la mejor combinación lineal que puede obtenerse con el conjunto $\{\varphi_s\}_{s=1}^{j}$ de funciones seleccionadas.

   El algoritmo *greedy* puro o relajado, con *back-fitting*, persigue lograr dicha ortogonalidad. Esto puede hacerse en cada iteración del algoritmo, mediante la ortogonalización sucesiva de las funciones que se van seleccionando, o bien, al final del mismo, ajustando todos los coeficientes conjuntamente.

Algoritmos *Greedy* con *Back-fitting*. Relaciones con Otros Algoritmos

En el ámbito del proceso de señales, el algoritmo *greedy* puro con *back-fitting*, es el denominado **Matching Pursuit Ortogonal** (*Orthogonal Matching Pursuit* –OMP) de [Pati *et al.* 1993] y [Davis *et al.* 1994], que, al igual que el MP, se aplica a un diccionario redundante de funciones básicas (**átomos**) [74] $\mathcal{G} = \{\psi_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \Gamma\}$ en un espacio de Hilbert $\mathcal{K}$, donde $\Gamma \in \mathbb{R}^p$ es un espacio de parámetros.

Dicho algoritmo (tabla 12 y Anexo C), sigue exactamente la misma filosofía que el proceso de ortogonalización de Gram-Schmidt (tabla 9, pág. 175) sólo que en un espacio de funciones.

[74] Generalmente se supone que los átomos son de norma unidad, pero para mayor generalidad, evitamos esta restricción en lo que sigue.

Si después de la iteración $(s-1)$-ésima, denotamos el residuo por:

$$R_{s-1}\mathsf{f} = \mathsf{f} - \sum_{j=1}^{s-1} \frac{\langle R_{s-1}\mathsf{f}, \bar{\psi}_j \rangle}{\|\bar{\psi}_j\|^2} \bar{\psi}_j$$

donde $\bar{\psi}_j$ $j = 1 : s-1$ son las funciones ortogonales seleccionadas, en la iteración s-ésima se elige el vector ψ_s según el criterio:

$$|\langle R_{s-1}\mathsf{f}, \psi_s \rangle| \geq \alpha \sup_{\boldsymbol{\gamma} \in \Gamma} |\langle R_{s-1}\mathsf{f}, \psi_{\boldsymbol{\gamma}} \rangle| \text{ con } 0 < \alpha \leq 1 \tag{II.134}$$

y se determina su componente $\bar{\psi}_s$ no redundante (ortogonal) con los vectores ya introducidos:

$$\bar{\psi}_s = \psi_s - \sum_{j=1}^{s-1} \frac{\langle \psi_s, \bar{\psi}_j \rangle}{\|\bar{\psi}_j\|^2} \bar{\psi}_j$$

Con ello, se realiza la descomposición:

$$R_{s-1}\mathsf{f} = \frac{\langle R_{s-1}\mathsf{f}, \bar{\psi}_s \rangle}{\|\bar{\psi}_s\|^2} \bar{\psi}_s + R_s\mathsf{f} \Rightarrow \mathsf{f} = \sum_{j=1}^{s} \frac{\langle R_{s-1}\mathsf{f}, \bar{\psi}_j \rangle}{\|\bar{\psi}_j\|^2} \bar{\psi}_j + R_s\mathsf{f}$$

en la que, por construcción, las funciones $\bar{\psi}_j$, $j = 1, ..., s$ son ortogonales y $R_s\mathsf{f} \perp \mathcal{H}_s = \left\langle \{\bar{\psi}_j\}_{j=1}^s \right\rangle = \left\langle \{\psi_j\}_{j=1}^s \right\rangle$.

La expresión de la solución en términos de los átomos originales se obtiene mediante (Anexo C):

$$\mathsf{f} = \sum_{j=1}^{s} c_j^{(s)} \psi_j + R_s\mathsf{f}, \text{ siendo:}$$

$$\begin{cases} c_j^{(s)} = c_j^{(s-1)} - \dfrac{\langle R_1\mathsf{f}, \psi_s \rangle}{\|\bar{\psi}_s\|^2} \dfrac{\langle \psi_s, \bar{\psi}_j \rangle}{\|\bar{\psi}_j\|^2}, \ j = 1 : s-1 \\ c_s^{(s)} = \dfrac{\langle R_{s-1}\mathsf{f}, \psi_s \rangle}{\|\bar{\psi}_s\|^2} \end{cases}$$

que utiliza una sustitución análoga a la obtenida de la ecuación (II.103): al finalizar el algoritmo, (iteración h-ésima):

$$\bar{\boldsymbol{\Psi}}_{\infty \times h}\mathbf{d} = \boldsymbol{\Psi}_{\infty \times h}\mathbf{U}^{-1}\mathbf{d} = \boldsymbol{\Psi}_{\infty \times h}\mathbf{c}, \text{ con } \mathbf{c} = \mathbf{U}^{-1}\mathbf{d}$$

donde hemos enfatizado la posible dimensión infinita del espacio funcional, en lugar de la dimensión n (tamaño muestral) del espacio de vectores $\boldsymbol{\psi}_j$ $j = 1 : h$, del algoritmo OLS.

Sobre la convergencia del algoritmo OMP, se tiene el siguiente teorema:

Teorema 60 *[Pati* et al. *1993], [Davis* et al. *1994] Sea* $\mathsf{f} \in \mathcal{T}_{\mathcal{G}} = \mathrm{cl}(\mathrm{co}(\mathcal{G}))$ *y sea* $r = \dim(\mathcal{T}_{\mathcal{G}})$ *(posiblemente infinita). El algoritmo OMP converge en* $h \leq r$ *iteraciones y se verifica que:*

$$\lim_{h \to r} \|R_h\mathsf{f}\| = 0; \ \mathsf{f} = \sum_{j=1}^{h} \frac{\langle R_{j-1}\mathsf{f}, \bar{\psi}_j \rangle}{\|\bar{\psi}_j\|^2} \bar{\psi}_j; \ \ \|\mathsf{f}\|^2 = \sum_{j=1}^{h} \frac{|\langle R_{j-1}\mathsf{f}, \bar{\psi}_j \rangle|^2}{\|\bar{\psi}_j\|^2}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Implementación del Algoritmo *Greedy* Relajado con *Back-fitting*
1. Elegir $f_0(\mathbf{x}) = 0, \forall \mathbf{x} \in \mathcal{X}$
2. Para $j = 1, 2, ..., h$ con h tal que $\hat{\mathcal{R}}_n(f_{h-1}) - \hat{\mathcal{R}}_n(f_h) \leq \Delta$ prefijado:
a. Calcular $\mathbf{e}_{j-1} = \mathbf{y} - \mathbf{f}_{j-1}$ con $(\mathbf{e}_{j-1})_i = y_i - f_{j-1}(\mathbf{x}_i), i = 1 : n$
b. Determinar $\varphi_j = \arg\min_{\varphi \in \mathcal{G}} \|\mathbf{e}_{j-1} - \boldsymbol{\varphi}\|$ con $(\boldsymbol{\varphi})_i = \varphi(\mathbf{x}_i), i = 1 : n$
c. Determinar $(\alpha_j, \beta_j) = \arg\min_{\alpha,\beta} \left\|\mathbf{y} - \alpha \mathbf{f}_{j-1} + \beta \boldsymbol{\varphi}_j\right\|$
d. Hacer $f_j = \alpha_j f_{j-1} + \beta_j \varphi_j \doteq \sum_{s=1}^{j} c_s \varphi_s$
3. *Back-fitting*: Para $j = 1, 2, .., h$:
a. Calcular $f_j = \sum_{j \neq s=1}^{h} c_s \varphi_s$
b Calcular $\mathbf{e}_j = \mathbf{y} - \mathbf{f}_j$
c. Determinar $\varphi_j = \arg\min_{\varphi \in \mathcal{G}} \|\mathbf{e}_j - \boldsymbol{\varphi}\|$ con $(\varphi)_i = \varphi(\mathbf{x}_i)$
d. Determinar $\mathbf{c} = \arg\min_{\boldsymbol{\alpha}} \left\|\mathbf{y} - \sum_{s=1}^{h} \alpha_s \boldsymbol{\varphi}_s\right\|$
e. Actualizar: $f_h = \sum_{s=1}^{h} c_s \varphi_s$

Tabla 13. Esquema general del algoritmo greedy relajado con re-fitting (paso 4). A veces, se refiere a este algoritmo como back-fitting pero este término admite varias interpretaciones.

Una alternativa al criterio (II.134), es la versión debida a [Vincent y Bengio 2002] denominada **pre-fitting**, (aunque ya [Friedman y Stuetzle 1981, p. 819] se refería a ella con el calificativo *back-fitting*) consistente en elegir simultáneamente, en cada iteración, el par $(\psi_s, \boldsymbol{\alpha}^{(s)})$ con $\boldsymbol{\alpha}^{(s)} = (\alpha_1^{(s)}, ..., \alpha_s^{(s)})^T$, tal que:

$$(\psi_s, \boldsymbol{\alpha}_s^{(s)}) = \arg \inf_{\psi_{\boldsymbol{\gamma}} \in \mathcal{G}, \boldsymbol{\alpha} \in \mathbb{R}^s} \left\| f - \sum_{j=1}^{s-1} \alpha_j \psi_j + \alpha_s \psi_s \right\|$$

Finalmente, el algoritmo presentado en la tabla 13) es un algoritmo *greedy* relajado con *back-fitting*, pero con una prestación adicional [Docampo *et al.* 1997]: se aprovecha el *back-fitting* final para afinar la selección de las funciones φ_s, $s = 1 : j$ introducidas durante el algoritmo.

Aplicación a las Arquitecturas Radiales

Algoritmo OLS como Algoritmo *Greedy* con *Back-fitting*

Por todo lo anterior, el algoritmo OLS presentado en el epígrafe anterior (pág. 174), puede considerarse un algoritmo *greedy* puro con *back-fitting* en un espacio de dimensión finita n (el tamaño muestral), tanto en la versión presentada (pág. 177), que ortogonaliza las variables en cada iteración, (regresión *stepwise*, [Friedman y Stuetzle 1981]), como en la versión también usual pero computacionalmente más costosa, en la que la ortogonalización se utiliza sólo para seleccionar las variables (criterio (II.104)), mientras que los coeficientes de la combinación se obtienen al final, de forma conjunta, mediante $\hat{\mathbf{c}} = (\boldsymbol{\Psi}^T \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^T \mathbf{y}$ donde $\boldsymbol{\Psi}$ es la matriz cuyas columnas son las variables ocultas.

Algoritmo *Greedy* con *back-fitting* para RBF con Filosofía Diccionario

El algoritmo OMP nace independientemente en el contexto del proceso de señales ([Pati *et al.* 1993], [Davis *et al.* 1994]), en cuyo ámbito, son relativamente frecuentes los problemas de reconstrucción de señales en ausencia de ruido. Por ello, dicho algoritmo no se formuló en versión regularizada.

Sin embargo, de lo anterior, se deduce también que, el algoritmo OLS bayesiano (tabla de pág. 197) es un algoritmo *greedy* con *back-fitting* al que se le ha aplicado **regularización**.

Al tratar el método bayesiano (pág. 197) con técnicas de regresión lineal generalizada, destacamos que, dicho método, no resuelve todavía el problema de selección del hiperparámetro parámetro de escala de las funciones básicas, de una forma fácilmente implementable. Por otra parte, la selección de dicho parámetro mediante técnicas clásicas resulta prohibitiva si se consideran parámetros diferentes en cada función básica (vector $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_h)^T$), pues, si $r = \#\Lambda$ es la cardinalidad del conjunto de valores de prueba, sería necesario entrenar r^h redes RBF para realizar posteriormente la selección.

Sin embargo, inspirándose en el concepto de diccionario redundante del proceso de señales (Anexo C), puede aplicarse el algoritmo OMP (su homólogo OLS), con o sin regularización, a un diccionario $\mathcal{G}$ con n funciones básicas (por ejemplo, centradas en los datos), cada una repetida tantas veces como posibles valores de prueba $\sigma \in \Lambda$ se deseen considerar. Es decir:

$$\mathcal{G} = \{\psi_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} = (\mathbf{w}^T, \sigma)^T, \mathbf{w} \in \Omega \subset \mathbb{R}^d, \sigma \in \Lambda \subset \mathbb{R}\}$$

donde Ω puede ser $\{\mathbf{x}_i\}_{i=1}^n$. Con ello, la cardinalidad del diccionario sería $\#\mathcal{G} = n \cdot r$.

Esta estrategia, produciría implícitamente la selección de los parámetros de escala más apropiados para cada función básica, y se podría integrar en un sólo algoritmo, el entrenamiento y la selección del modelo de una red RBF con un conjunto tan complejo de hiperparámetros (varios parámetros de escala y varios regularizadores).

En el marco de un diccionario así definido, las funciones básicas de la RBF recogerían la información *tiempo-frecuencia* de la función de regresión. Incluso podría ocurrir que un mismo dato fuese seleccionado varias veces como centro, con diferentes parámetros de escasa cada vez (con filosofía multirresolución). Por su parte, la regularización bayesiana, tendría por objeto evitar la asimilación del ruido de los datos.

Es claro que el diccionario así construido contiene al que resulta de considerar un sólo parámetro σ y sólo n funciones básicas. Con ello, se obtienen, como mínimo, los resultados que se obtenían con el algoritmo OLS usual y es de esperar que mejoren. Por otra parte, un mismo algoritmo diseñado para diccionarios generales permite resolver la situación tradicional sin más que tomar $\#\Lambda = r = 1$.

Por tanto, el algoritmo sigue el mismo esquema que el mostrado en la tabla 11, sólo que, ahora, el número de funciones básicas potenciales no es el tamaño muestral n, sino la cardinalidad del diccionario $\mathcal{G}$ construido.

No conocemos aplicaciones de este tipo en la literatura pero, nuestros ensayos, han producido muy buenos resultados obteniéndose redes RBF que tienden a ser más parsimoniosas que las que utilizan el OLS con el diccionario tradicional (véase una comparativa más adelante).

Aplicación a las Arquitecturas de Proyección

La metodología descrita en el epígrafe anterior, basada en diccionarios, puede aplicarse igualmente a las arquitecturas de tipo Proyección, sin más que considerar conjuntos de funciones:

$$\mathcal{G} = \{\psi_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} = \bar{\mathbf{w}} = (\mathbf{w}^T, w_0)^T, \bar{\mathbf{w}} \in \Omega \subset \mathbb{R}^{d+1}\}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde, salvo que exista información a priori muy concreta sobre un conjunto discreto de valores $\bar{\mathbf{w}}$, lo normal ahora es que Ω sea un subconjunto compacto de $\mathbb{R}^{d+1}$. Por ello, la selección de una

Algoritmo de Entrenamiento *Projection Pursuit*

1. Elegir un suavizador no paramétrico $\varphi(u;\boldsymbol{\tau})$, $u \in \mathbb{R}$, $\boldsymbol{\tau} \in \mathbb{R}^r$, y $f_0(\mathbf{x}) = 0$, $\forall \mathbf{x} \in \mathcal{X}$
2. Para $j = 1, 2, ..., h$ con h tal que $\hat{\mathcal{R}}_n(f_{h-1}) - \hat{\mathcal{R}}_n(f_h) > \Delta$ prefijado:
   a. Calcular $\mathbf{e}_{j-1} = \mathbf{y} - \mathbf{f}_{j-1}$ con $(\mathbf{e}_{j-1})_i = y_i - f_{j-1}(\mathbf{x}_i)$, $i = 1:n$
   b. Determinar: $(\mathbf{w}_j, \boldsymbol{\tau}_j) = \arg\min_{\mathbf{w},\boldsymbol{\tau}} \left\| \mathbf{e}_{j-1} - \boldsymbol{\varphi}_{\boldsymbol{\tau},\mathbf{w}} \right\|$, con $(\boldsymbol{\varphi}_{\boldsymbol{\tau},\mathbf{w}})_i = \varphi(\mathbf{w}^T\mathbf{x}_i;\boldsymbol{\tau})$, $i = 1:n$
      i) Elegir valor inicial $\mathbf{w}^{(0)}$
      ii) Iterar algún algoritmo de optimización hasta la estabilidad, $s = 1, 2, ...$:
         $\mathbf{w}^{(s+1)} = \mathbf{w}^{(s)} + \Delta\mathbf{w}$ con $\Delta\mathbf{w}$ óptimo obtenido utilizando:
         $\boldsymbol{\tau}^{(s)} = \arg\min_{\boldsymbol{\tau}\in\mathbb{R}^{r_s}} \left\| \mathbf{e}_{j-1} - \boldsymbol{\varphi}^{(s)} \right\|$, con $(\boldsymbol{\varphi}^{(s)})_i = \varphi(\mathbf{w}^{(s)T}\mathbf{x}_i;\boldsymbol{\tau})$, $i = 1:n$
   c. Determinar $(\alpha_j, \beta_j) = \arg\min_{\alpha,\beta} \left\| \mathbf{y} - \alpha\mathbf{f}_{j-1} + \beta\boldsymbol{\varphi}_j \right\|$, con $(\boldsymbol{\varphi}_j)_i = \varphi(\mathbf{w}_j^T\mathbf{x}_i;\boldsymbol{\tau}_j)$, $i = 1:n$
   d. Hacer: $f_j = \alpha_j f_{j-1} + \beta_j \varphi(\mathbf{w}_j^T\mathbf{x};\boldsymbol{\tau}_j) \equiv \sum_{s=1}^{j} \gamma_s \varphi(\mathbf{w}_s^T\mathbf{x};\boldsymbol{\tau}_s)$
3. *Back-fitting*:
   a. Determinar $\mathbf{c} = \arg\min_{\boldsymbol{\gamma}} \left\| \mathbf{y} - \sum_{s=1}^{h} \gamma_s \boldsymbol{\varphi}_s \right\|$
   b. Actualizar: $f_h = \sum_{s=1}^{h} c_s \varphi_s$

Tabla 14. Esquema general del algoritmo greedy relajado con back-fitting (paso 4) aplicado al entrenamiento de una red projection pursuit.

nueva función $\psi_{\bar{\mathbf{w}}}$ en cada iteración del algoritmo *greedy* debería realizarse mediante un algoritmo de optimización no lineal.

La tabla 14 presenta una implementación de esta filosofía para el entrenamiento de las redes *projection pursuit* con suavizador $\varphi(\cdot;\boldsymbol{\tau}_j)$, $\boldsymbol{\tau}_j \in \mathbb{R}^{r_j}$ en cada proyección. El paso fundamental del algoritmo es el paso 2.b. en el que se realiza la búsqueda de la dirección $\mathbf{w}_j$ más informativa en la que proyectar los datos.

Dicha búsqueda se realiza mediante algún algoritmo de optimización no lineal (gradiente conjugado, Levenberg-Marquardt o algún método basado en el Hesiano), en cuyas iteraciones, es necesario intercalar la estimación del suavizador en la dirección obtenida hasta ese momento, lo que supone una carga computacional importante. [Friedman y Stuetzle 1981] utilizaron el denominado *supersmoother,* un suavizador de ventana variable con buenas propiedades computacionales, pero puede utilizarse cualquier método no paramétrico.

El algoritmo de la tabla mencionada, puede ser utilizado con las redes MLP sin más que elegir, por ejemplo, $\varphi(\bar{\mathbf{w}}^T\bar{\mathbf{x}};\alpha) = \alpha\psi(\mathbf{w}^T\mathbf{x} + w_0)$ con $\bar{\mathbf{w}}^T = (\mathbf{w}^T, w_0)^T$, donde g es una función sigmoide como la logística, la tangente hiperbólica o cualquiera de las mencionadas con anterioridad. En este caso, puede resolverse el problema de optimización del paso 2.b conjuntamente en $(\bar{\mathbf{w}}, \alpha)$ al no existir suavizadores no paramétricos.

En relación con los algoritmos del tipo *back-propagation*, la ventaja de este enfoque es que se sustituye un problema de optimización global en términos $d \times h$ parámetros por h problemas de optimización menos complejos con d parámetros cada uno, lo que reduce el impacto asociado a la obtención de mínimos locales.

En contrapartida, los algoritmos greedy progresan buscando el beneficio a corto plazo, sin considerar la mejor combinación global de funciones a introducir (aquí es aplicable lo comentado en la pág. 178) en relación con el algoritmo OLS). Con ello, estos algoritmos no garantizan mejores resultados que *back-propagation*.

En la figura 26 (izquierda) se muestra una red MLP entrenada con el algoritmo *projection-pursuit* en su versión *greedy* pura con *back-fitting*. Con estas redes, el algoritmo es muy sensible a las condiciones iniciales que, además, dependen de la función desconocida a estimar. En la figura indicada, se ha utilizado el método de inicialización de Nguyen-Widrow adaptado al intervalo

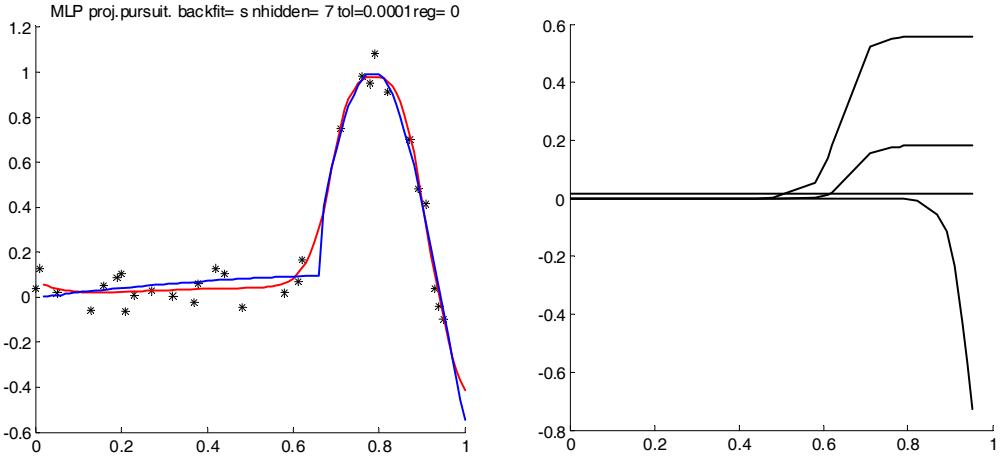


Figura 26. Izquierda: ajuste de una red MLP (función más suave) con activaciones logísticas, entrenada con el algoritmo *greedy* con *back-fitting*. Derecha: cuatro primeras funciones básicas de la red anterior, de la forma $\alpha g(wx + w_0)$ con g la función logística

$[0, 1]$, lo cuál requiere aportar un número a priori de funciones básicas, (para lo que se ha realizado una exploración previa).

En el gráfico derecho de la misma figura se muestran las cuatro primeras funciones básicas de la red MLP anterior, de la forma $\alpha g(wx + w_0)$ con g la función logística.

Otras implementaciones relevantes son: [Hwang *et al.* 1994] que utilizan combinaciones lineales de polinomios de Hermite (véase también [Kwok y Yeung 1996]), [Zhao y Atkeson 1996] que utilizan splines como suavizadores y [Hush y Horne 1998], que, al utilizar funciones de activación *ramp* (ecuación (II.35)) que dan lugar a un número finito de soluciones factibles, proporcionan algoritmos que convergen en un número finito de iteraciones.

Consistencia de las Redes Neuronales Entrenadas con Algoritmos *Greedy*

Casi todos los resultados que proporcionan órdenes de aproximación de las redes neuronales suponen que éstas son entrenadas con un algoritmo *greedy* (no necesariamente con *back-fitting*). Con ello, aprovechando la descomposición del riesgo en error de aproximación y error de estimación, (pág. 60), pueden obtenerse, a su vez, cotas para la consistencia de los correspondientes estimadores.

La tabla 15 presenta los resultados más importantes sobre estos órdenes de aproximación para las redes neuronales RBF y MLP en diferentes espacios funcionales. Dichos resultados pueden verse con más detalle en el Anexo C.

Como puede observarse todas los resultados se verifican en la norma L_2. Para lograr dichas tasas, es necesario concretar el grado de suavidad de las funciones del espacio funcional objetivo $\mathcal{T}$. En la tabla anterior, se denota (véase el Anexo C):

$$\mathcal{V}_p^r(\mathbb{B}_d) = \{f \in \mathcal{W}_p^r(\mathbb{B}_d) : \|f\|_{\mathcal{W}_p^r} \leq 1\} \text{ con } \mathbb{B}_d = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\}$$

donde $\mathcal{W}_p^r(\mathbb{B}_d)$ es el espacio de Sobolev compleción del espacio de funciones con r derivadas continuas con norma L_p finita.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Tasa de Convergencia de las Redes Neuronales con Algoritmos *Greedy* **MLP** $\hat{f}(\mathbf{x}) = \sum_{i=1}^{h} c_i \sigma(\mathbf{w}_i^T \mathbf{x} + w_{i0})$ **y RBF** $\hat{f}(\mathbf{x}) = \sum_{i=1}^{h} c_i \psi_i(\mathbf{x} - \mathbf{w}_i)$		
Espacio Funcional	**Condiciones**	**Tasa**
$\mathcal{T}_C(\mathbb{B}_{d,\tau})$, μ med. prob. [Barron 1993]	σ sigmoide medible, acotada	$\left\|\mathsf{f} - \hat{f}\right\|^2_{L_2(\mathbb{B}_{d,\tau},\mu)} \leq (2\tau C_\mathsf{f})^2 h^{-1}$
$\mathcal{T}_C(\mathbb{B}_{d,\tau})$, μ med. prob. [Makovoz 1996]	σ sigmoide medible, acotada	$\left\|\mathsf{f} - \hat{f}\right\|^2_{L_2(\mathbb{B}_{d,\tau},\mu)} \leq C h^{-(d+1)/2d}$
$\mathcal{V}_p^r(\mathbb{B}_d)$ [Pinkus 1999, cor. 6.11]	σ sigmoide de grado s $r \in \{1, ..., s + 1 + \frac{d-1}{2}\}$	$\left\|\mathsf{f} - \hat{f}\right\|_{L_2(\mathbb{B}_{d,\tau})} \leq C h^{-r/d}$
$\mathcal{T}_\psi = \{\mathsf{f} = \psi * \mu\}$ $\{\mathsf{f} : \tilde{\mathsf{f}} \in C_c^r(\mathbb{R}^d)\} \subset \mathcal{T}_\psi$ [Girosi y Anzellotti 1993]	μ medida Radon, $\|\mu\| \leq \infty$ $\psi_i(\mathbf{x} - \mathbf{w}_i) = e^{-\|\mathbf{x}-\mathbf{w}_i\|^2}$ $2r > d$	$\left\|\mathsf{f} - \hat{f}\right\|^2_{L_2} \leq C_\mathsf{f} h^{-1}$
$\mathcal{W}_1^{2r}(\mathbb{R}^d)$ [Girosi 1994]	$\psi_i(\mathbf{x} - \mathbf{w}_i) = e^{-\|\mathbf{x}-\mathbf{w}_i\|^2/\sigma_i^2}$ $2r > d$	$\left\|\mathsf{f} - \hat{f}\right\|^2_{L_2} \leq C_\mathsf{f} h^{-1}$

Tabla 15. Algunas tasas de convergencia con redes neuronales. Todos los resultados se basan en el algoritmo greedy. Para la notación véase el texto o el anexo C. En todos los casos, se supone que σ es una funcion no constante

Asimismo, $\mathcal{T}_C(\mathbb{B}_{d,\tau})$ es la restricción a $\mathbb{B}_{d,\tau} = \mathbb{B}_{d,\tau} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \tau\}$ de la siguiente familia estudiada en profundidad[75] por [Barron 1993]:

$$\mathcal{T}_C(\mathbb{R}^d) = \{\mathsf{f} : \mathbb{R}^d \to \mathbb{R} : C_\mathsf{f} = \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_1 \left|\tilde{\mathsf{f}}(\boldsymbol{\omega})\right| d\boldsymbol{\omega} < \infty\} \quad \text{con: } \mathbb{B}_{d,\tau} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \tau\} \tag{II.135}$$

Precisamente, el orden de aproximación dado por [Barron 1993] para las redes MLP (tabla anterior), como puede observarse, no depende de la dimensión d del espacio de entrada, lo que en principio haría pensar que las redes MLP están a salvo de la *maldición de la dimensionalidad*.

Sin embargo, $\mathcal{T}_C(\mathbb{R})$ es una familia que disminuye su complejidad a medida que aumenta dicha dimensión (la integral n-dimensional C_f puede crecer con la dimensión), lo que permite una tasa independiente de d. Por ejemplo [Barron 1993], si $\mathsf{f}(\mathbf{x}) = e^{-\|\mathbf{x}\|^2/2}$ entonces, $C_\mathsf{f} \leq d^{1/2}$ (un orden similar se tiene para las funciones definidas positivas).

Los resultados de consistencia de este apartado suponen todos pérdida cuadrática, es decir:

$$R(f) = \|\mathsf{Y} - f(\mathsf{X})\|^2_{L_2(\mathsf{P_Z})} = R(\mathsf{f}) + \|f - \mathsf{f}\|^2_{L_2(\mathsf{P_X})}$$

Consistencia *Greedy* de las Redes RBF

El resultado más importante sobre las tasas de consistencia para las redes RBF entrenadas con algoritmos *greedy* es el de [Niyogi y Girosi 1996, th. 4.1], que se obtiene combinando tasas al error de estimación con las tasas proporcionadas por [Girosi y Anzellotti 1993] (tabla 15).

El espacio objetivo es el espacio (teorema 184, Anexo C):

$$\mathcal{T} = \{\mathsf{f} \in L_2(\mathbb{R}^d) : \mathsf{f} = \psi * \mu, \ \|\mu\| \leq b\}$$

[75] Las funciones de esta familia poseen derivadas continuas y la transformada de su gradiente es integrable [Barron 1993]: $\nabla f(\mathbf{x}) = \int e^{i\boldsymbol{\omega}\mathbf{x}} \widetilde{\nabla f}(\boldsymbol{\omega}) d\boldsymbol{\omega}$. En [Pinkus 1999, p. 181] se indica la siguiente relación entre la familia y los espacios de Sobolev en $\mathbb{B}_d$: $\mathcal{W}_\infty^{d/2+2} \subseteq \mathcal{T}_C(\mathbb{B}_d) \subseteq \mathcal{W}_\infty^1 \subseteq \mathcal{W}_2^1$.

donde μ es una medida de Radon con signo y cuya variación total[76] $|\mu|$ es acotada, $\|\mu\| < \infty$.

Se consideran las siguientes familias de funciones:

$$\mathcal{F}_h = \{f \in L_2(\mathbb{R}^d) : f(\mathbf{x}) = \sum_{i=1}^{h} c_i \zeta(\|\mathbf{x} - \mathbf{w}_i\|) : \mathbf{c} \in \mathbb{R}^{h_n}, \mathbf{w}_i \in \mathbb{R}^d, \sum_{i=1}^{h} |c_i| \leq b\}$$

donde $\zeta(r)$ es una función gausiana restringida al conjunto $[0, a] \subset \mathbb{R}$.

Se verifica el siguiente lema que aporta una cota para el error de estimación:

Lema 61 *[Niyogi y Girosi 1996, asev. D.10] Con probabilidad* $1 - \eta$, *y* $\forall f \in \mathcal{F}_h$, *se verifica que:*

$$\left|\hat{R}_n(f) - R(f)\right| \leq O\left(\left[\frac{dh\ln(nh) + \ln(1/\eta)}{n}\right]^{1/2}\right)$$

Combinando este resultado con el del teorema [Girosi y Anzellotti 1993, th. 8.2], se tiene el siguiente resultado, en el que se denota $\hat{f}_{h,n} = \arg\inf_{\mathcal{F}_h}(\hat{R}_n)$:

Teorema 62 *[Niyogi y Girosi 1996, th. 4.1]* $\forall \eta \in (0,1)$, $h \in \mathbb{N}^+$ *se verifica con probabilidad* $1 - \eta$, *que:*

$$R(\hat{f}_{h,n}) - R(\mathsf{f}) = \left\|\hat{f}_{h,n} - \mathsf{f}\right\|^2_{L_2(\mathsf{P_X})} \leq O\left(\frac{1}{h}\right) + O\left(\left[\frac{hd\ln(nh) + \ln(1/\eta)}{n}\right]^{1/2}\right)$$

Del teorema anterior, se deduce que el error de generalización tiende a cero si $h/n \to 0$ cuando $n \to \infty$.

Con objeto de determinar la regla óptima $h_n = h(n)$ de selección del número de unidades ocultas de la red en función del tamaño muestral, basta obtener fijado n, el óptimo en h. Es decir:

$$\frac{\partial}{\partial h}[R(\hat{f}_{h,n}) - R(\mathsf{f})] = 0$$

Utilizando la cota del teorema anterior, resulta [Niyogi y Girosi 1996] que, suponiendo $1/\eta << (nh)^{hd}$, el número de parámetros debe ser con probabilidad $1 - \eta$:

$$p(n) = h \times (d+1) \propto (\frac{1}{d\ln n})^{1/3}$$

Consistencia *Greedy* de las Redes MLP

Con anterioridad al resultado anterior y abriendo el camino a este tipo de resultados, [Barron 1994] obtiene cotas a la esperanza del error de generalización de las redes MLP utilizando cotas al error de estimación obtenidos por él mismo [Barron 1991, p. 569], en el marco del método del control de la complejidad estocástica, e incorporando los resultados sobre órdenes de aproximación de las redes MLP obtenidos en [Barron 1993] mediante el lema de Maurey para la familia (II.135).

[76] [Hoffmann-Jorgensen 1994, p. 23]: **medida interior** es la medida ν_* definida $\forall B \in \mathcal{B}$ como $\nu_*(B) = \sup\{\sum_{j=1}^{t} B_j : B_j \subset \mathcal{B},\ B_j \cap B_i = \varnothing,\ i \neq j,\ \cup_{j=1}^{t} B_j \subseteq B\}$. [Hoffmann-Jorgensen 1994, p. 79] [Rudin 1988, p. 133]: dada una medida μ en un espacio $(\mathcal{X}, \mathcal{B})$ con $\mathcal{B}$ σ-álgebra en $\mathcal{X}$, se denomina **variación total** $|\mu|$ de μ a la medida interior correspondiente a la función $\nu(B) = |\mu(B)|$, $B \in \mathcal{B}$ y se denota $\|\mu\| = \nu_*(\mathcal{X}) = \int_{\mathcal{X}} d\,|\mu|\,\mathbf{x}$.

USC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Por tanto, se supone que la función de regresión pertenece a la familia (II.135). Asimismo, se utiliza la pérdida cuadrática, se supone que f es acotada, es decir, $\mathsf{f}(\mathbf{x}) \in [\xi_1, \xi_2] \subset \mathbb{R}, \forall \mathbf{x} \in \mathcal{X})$ y que se dispone de una muestra aleatoria simple $\mathbf{z}^n$ de tamaño $n \geq 2$. Las familias se consideran de la forma:

$$\mathcal{F}_h = \{f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=1}^{h} c_i \psi(\mathbf{w}_i^T \mathbf{x} + w_{i0}) + c_0 : \boldsymbol{\theta} \in \Theta_{h,\tau,C}, h \in \mathbb{N}^+\} \tag{II.136}$$

donde $\Theta_{h,\tau,C}$ está constituido por los vectores de parámetros:

$$\boldsymbol{\theta} = (c_0, c_1, ..., c_h, \mathbf{w}_1^T, ..., \mathbf{w}_h^T, w_{01}, ..., w_{0h}) \in \mathbb{R}^{h(2+d)+1}$$

con $c_i, w_{0i} \in \mathbb{R}$ y $\mathbf{w}_i \in \mathbb{R}^d$ que verifican $\|\mathbf{w}_i\|_1 \leq \tau$, $|w_{0i}| \leq \tau$, $\sum_{i=1}^{h} |c_i| \leq C$ y $c_0 \in [\xi_1 - C, \xi_2 + C]$.

Además, ψ es una función acotada y sigmoide con valores límites $\psi(-\infty) = -1$ y $\psi(\infty) = 1$, que satisface una condición de Lipschitz del tipo $|\psi(r) - \psi(r')| \leq a\,|r - r'|$ para algún $a > 0$, y que, cuando $r \to \infty$ $(r \to -\infty)$ se aproxima a los límites anteriores con una velocidad al menos polinómica [Barron 1994, p. 119 y 123].

El conjunto de parámetros se discretiza mediante un conjunto $\Theta_{h,\tau,C,\varepsilon}$ de cardinalidad finita mediante recubrimientos de tamaño ε de tal forma que $\forall \boldsymbol{\theta} \in \Theta_{h,\tau,C}$ existe un $\boldsymbol{\theta}' \in \Theta_{h,\tau,C,\varepsilon}$ tal que: $\|\mathbf{w}_i - \mathbf{w}_i'\|_1 \leq \varepsilon$, $|w_{0i} - w_{0i}'| \leq \varepsilon$, $\sum_{i=1}^{h} |c_i - c_i'| \leq C\varepsilon$ y $|c_0 - c_0'| \leq C\varepsilon$.

[Barron 1994, cor. 2] extiende las propiedades aproximadoras de las redes MLP del tipo (II.136) a las funciones con parámetros discretizados y probando que si $\varepsilon_h = O(1/\sqrt{h})$ entonces, existe $f_{\boldsymbol{\theta}} \in \Theta_{h,\tau,C,\varepsilon}$ tal que:

$$\|f_{\boldsymbol{\theta}} - \mathsf{f}\|_{L_2(\mathsf{P_X})} = O(\frac{C}{\sqrt{h}}) \tag{II.137}$$

La minimización del riesgo empírico en $\Theta_{h,\tau,C,\varepsilon}$ equivale [Barron 1994, cor. 2] a un problema de mínimos cuadrados regularizado con estabilizador constante $\mathcal{E}_h(f_{\boldsymbol{\theta}}) = \ln \#\Theta_{h,\tau,C,\varepsilon}$:

$$\min_{\Theta_{h,\tau,C,\varepsilon}} \hat{R}_n(f_{\boldsymbol{\theta}}) \equiv \min_{\Theta_{h,\tau,C}} \{\hat{R}_n(f_{\boldsymbol{\theta}}) + \frac{\lambda}{n} \ln \#\Theta_{h,\tau,C,\varepsilon}\}$$

y, bajo ciertas condiciones para ε, [Barron 1994, p. 126] prueba que $\#\Theta_{h,\tau,C,\varepsilon} = O(hd \ln(Chdn))$. Con ello, y utilizando (II.137), puede acotar el denominado índice de resolubilidad (*index of resolvability*):

$$\begin{aligned} \min_{\Theta_{h,\tau,C,\varepsilon}} \{R(f_{\boldsymbol{\theta}}) + \frac{\lambda}{n} \ln \#\Theta_{h,\tau,C,\varepsilon}\} &= \min_{\Theta_{h,\tau,C,\varepsilon}} \{\|f_{\boldsymbol{\theta}} - \mathsf{f}\|^2_{L_2(\mathsf{P_X})} + \frac{\lambda}{n} \ln \#\Theta_{h,\tau,C,\varepsilon}\} \\ &\leq O(\frac{C^2}{h}) + O(\frac{hd}{n} \ln(Chdn)) \end{aligned}$$

Entonces, utilizando que las hipótesis $f \in \mathcal{F}_h$ están acotadas, se obtiene el siguiente resultado:

Teorema 63 *[Barron 1994, th. 3] Sea* $\hat{f}_{h,n,C} = \arg\min_{\Theta_{h,\tau,C,\varepsilon}} \hat{R}_n(f)$. *Si la función de regresión* f *satisface* $C_{\mathsf{f}} \leq C$, *con* C_{f} *la constante de la expresión (II.135), entonces:*

$$E_{\mathsf{Z}^n}[R(\hat{f}_{h,n,C}) - R(\mathsf{f})] = E_{\mathsf{Z}^n}[\left\|\hat{f}_{h,n,C} - \mathsf{f}\right\|^2_{L_2(\mathsf{P_X})}] \leq O(\frac{C^2}{h}) + O(\frac{hd}{n} \ln n)$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Además, para $h \approx C(\frac{n}{d \ln n})^{1/2}$, *dicha tasa es de orden:*

$$E_{\mathsf{Z}^n}[\left\| \hat{f}_{h,n,C} - \mathsf{f} \right\|^2_{L_2(\mathsf{P_X})}] \leq O(C\frac{d}{n}\ln n)^{1/2}$$

II.6.7 Otros Algoritmos con Filosofía Diccionario

Regularización Cuadrática o Absoluta: MOF y BP

La regularización bayesiana es una penalización cuadrática de los coeficientes de la red. En proceso de señales, es conocido que una penalización cuadrática tiende a producir soluciones poco parsimoniosas.

Este tipo de penalización se utiliza por ejemplo en el método de los *frames*, (*method of frames* –MOF, [Daubechies 1988], [Chen *et al.* 1999]), el cuál, entre todos los posibles vectores $\mathbf{c} \in \mathbb{R}^h$ solución al problema de análisis de una señal $\mathbf{f} \in \mathbb{R}^n$ en ausencia de ruido, utilizando $h \leq n$ elementos ψ_j de un diccionario $\mathcal{G}$ (pág. 383, Anexo C):

$$\mathbf{f} = \mathbf{\Psi c}$$

elige el que posee mínima longitud. Es decir, realiza el proceso de análisis resolviendo el problema:

$$\begin{gathered} \min_{\boldsymbol{\alpha}} \|\mathbf{c}\|_2^2 \\ \text{sujeto a: } \mathbf{\Psi c} = \mathbf{f} \end{gathered}$$

el cuál, posee solución única dada por la inversa generalizada [Chen *et al.* 1999]:

$$\mathbf{c} = \mathbf{\Psi}^* \mathbf{f} = \mathbf{\Psi}^T (\mathbf{\Psi \Psi}^T)^- \mathbf{f}$$

Persiguiendo la máxima parsimonia de la representación, [Chen *et al.* 1999] propusieron el algoritmo *basis pursuit* (BP), que elige el vector con mínima norma en ℓ_1, resolviendo el siguiente problema:

$$\begin{gathered} \min_{\boldsymbol{\alpha}} \|\mathbf{c}\|_1 \\ \text{sujeto a: } \Psi \mathbf{c} = \mathbf{f} \end{gathered}$$

Sin embargo, el objetivo de parsimonia posee como coste una mayor dificultad de resolución pues, el problema anterior no admite solución analítica y ha de resolverse mediante algoritmos de programación lineal de la siguiente forma:

$$\begin{gathered} \min_{\boldsymbol{\beta}} \mathbf{1}_h^T \boldsymbol{\beta} \\ \text{sujeto a: } \begin{cases} [\mathbf{\Psi}, -\mathbf{\Psi}]\boldsymbol{\beta} = \mathbf{f} \\ \boldsymbol{\beta} \geq \mathbf{0} \end{cases} \end{gathered}$$

con $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ con $\boldsymbol{\beta}_i \in \mathbb{R}^\tau$ con $\tau = \#\mathcal{G}$, y donde, al final, $\mathbf{c} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$.

Con objeto de evaluar este incremento de parsimonia, hemos aplicado los métodos MOF y BP en el ajuste de un conjunto de datos sujetos a ruido, utilizando un diccionario $\mathcal{G}$ construido con funciones básicas gausianas centradas en los datos con parámetros de escala $\sigma \in [0.01, 0.2]$ a

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

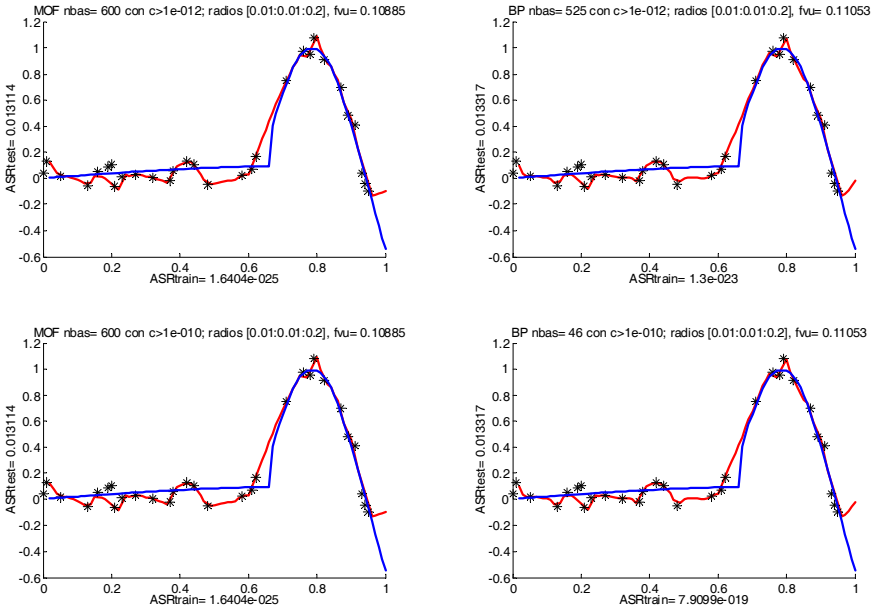


Figura 27. Comparación del grado de parsimonia de los métodos MOF y BP. En la primera fila, se filtraron los coeficientes menores que 10^{-12} : a pesar del corte MOF sigue utilizando todos los átomos del diccionario, mientras que BP utiliza 525. En la segunda fila, se filtraron los coeficientes menores que 10^{-10}, en cuyo caso MOF todavía requiere todos los átomos mientras que BP utiliza sólo 46. En contrapartida, BP produce un ligero mayor error.

intervalos de 0.01 con el fin de que el problema $\Psi\boldsymbol{\alpha} = \mathbf{f}$ posea solución. Por tanto, la cardinalidad de $\mathcal{G}$ es 600.

Los resultados obtenidos (figuras 27 y 28) muestran que, efectivamente, el método BP es claramente más parsimonioso. Cuando se filtran los coeficientes menores que 10^{-10}, el método BP requiere sólamente 46 átomos del diccionario para representar la función mientras que el método MOF sigue utilizando los 600 átomos del diccionario. El histograma de los coeficientes corrobora este hecho al producir para el caso BP una mayor concentración en el cero.

La poca parsimonia del método MOF se corresponde con el mismo efecto que en el problema de regresión, produce la regularización cuadrática con un único parámetro regularizador, (*weight decay*).

Basis Pursuit Denoising y Regularización con *Lasso*

Basis Pursuit Denoising

Como es lógico, por su propia naturaleza, los métodos MOF y BP presentados en el epígrafe anterior, no son adecuados para un problema de regresión en ambiente de ruido pues imponen como condición la interpolación de los datos. Para el caso de señales sujetas a perturbaciones aleatorias, [Chen *et al.* 1999] proponen el método denominado *basis pursuit denoising*, (BPDN)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

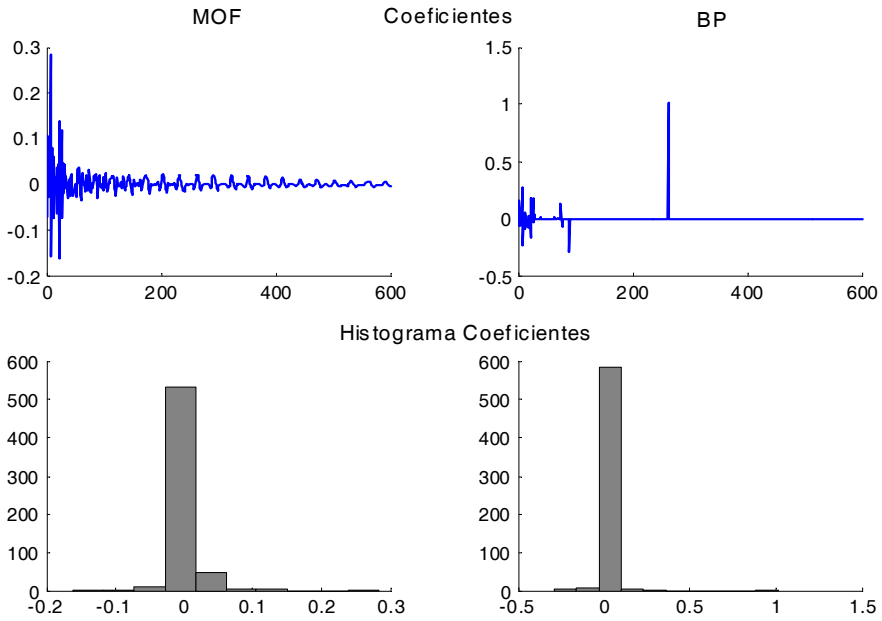


Figura 28. Serie e histograma de los coeficientes de las descomposiciones MOF y BP de la figura 27.

consistente en la resolución del siguiente problema regularizado:

$$\min_{\boldsymbol{\alpha}}\{\|\mathbf{y}-\Psi\mathbf{c}\|_2^2+\lambda\|\mathbf{c}\|_1\} \tag{II.138}$$

Como puede observarse, este problema equivale a un problema de regresión regularizado con penalización absoluta en lugar de cuadrática. (Por su parte, la correspondiente formulación de MOF en un ambiente de ruido produce el problema de regresión regularizado tratado varias veces con anterioridad).

Bajo un enfoque bayesiano, la función objetivo de (II.138) puede verse como la distribución a posteriori de los parámetros, cuando la distribución a priori de éstos es de la forma:

$$\mathsf{p}_{\mathsf{C}}(\mathbf{c}) \propto \exp\{\lambda\sum_{j=1}^{\tau}|c_j|\}$$

lo que se corresponde con una distribución a priori de Laplace. Sin embargo, el método BPDN no persigue la esperanza de la distribución a posteriori $\mathsf{P}_{\mathsf{C}/\mathbf{y}}$ como en el método bayesiano, sino su moda. En el caso de penalización cuadrática (prior gausiana) y ruido gausiano, ambas coinciden, pero no en este caso.

La resolución del problema anterior puede realizarse mediante el siguiente problema equivalente [Chen *et al.* 1999]:

$$\begin{array}{c}\min_{\boldsymbol{\beta}}\{\lambda\mathbf{1}_h^T\boldsymbol{\beta}+\|\gamma\boldsymbol{\beta}\|_2^2+\|\boldsymbol{\varepsilon}\|_2^2\}\\ \text{sujeto a: } \left\{\begin{array}{c}[\boldsymbol{\Psi},-\boldsymbol{\Psi}]\boldsymbol{\beta}+\delta\boldsymbol{\varepsilon}=\mathbf{y}\\ \boldsymbol{\beta}\geq\mathbf{0}\end{array}\right.\end{array} \tag{II.139}$$

con $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^T,\boldsymbol{\beta}_2^T)^T$, $\boldsymbol{\beta}_i\in\mathbb{R}^h$, $i=1,2$, tal que $\mathbf{c}=\boldsymbol{\beta}_1-\boldsymbol{\beta}_2$ y donde $\boldsymbol{\varepsilon}$ es un vector de holgura que *relaja* las restricciones interpoladoras, controlado por el parámetro δ que, con el parámetro γ actúan como hiperparámetros. De ellos, el parámetro δ es el relevante pudiéndose elegir γ entre

un amplio abanico de valores en $(0, \delta]$, con lo que el término penalizador cuadrático $\|\gamma\boldsymbol{\beta}\|_2^2$ no rompe la filosofía de penalización con norma absoluta.

El problema anterior es un programa cuadrático que puede resolverse mediante algoritmos de punto interior (Anexo D) como los utilizados con las redes SVM, lo cuál induce a preguntarse sobre las relación entre las soluciones proporcionadas por ambos métodos.

Es interesante comprobar cómo esta relación efectivamente existe [Girosi 1998], aunque sólo se ha podido demostrar analíticamente entre la formulación SVM y la formulación BP (ausencia de ruido en los datos).

Regularización con *Lasso*

Como hemos visto en el capítulo primero, (Apartado I.4.1), la función objetivo (ecuación (II.138)) del método BPDN puede escribirse de forma equivalente, como:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \boldsymbol{\Psi}\mathbf{c}\|_2^2 \\ &\text{sujeto a: } \|\mathbf{c}\|_1 \leq \kappa \end{aligned} \tag{140}$$

Inspirándose en una propuesta previa de [Breiman 1995] (método del *garrote*, véase también [Hastie *et al.* 2001]), esta formulación del problema regularizado con penalización absoluta fué propuesta independientemente por [Tibshirani 1996*b*], quién la denominó *lasso* en referencia al papel que juega el parámetro κ como limitador del tamaño de los coeficientes.

La formulación (II.140) es un problema de regresión con restricciones en los coeficientes que puede escribirse como:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \|\mathbf{y} - [\boldsymbol{\Psi}, -\boldsymbol{\Psi}]\boldsymbol{\beta}\|_2^2 \\ &\text{sujeto a: } \mathbf{1}^T\boldsymbol{\beta} \leq \kappa \end{aligned}$$

con $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, $\boldsymbol{\beta}_i \in \mathbb{R}^h$, $i = 1, 2$, tal que $\mathbf{c} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$. El problema anterior es un programa cuadrático con restricciones lineales que determina un método alternativo de solución al planteado en (II.139) que, además, no requiere más que un hiperparámetro κ de control de la complejidad, en contraste con los tres parámetros allí requeridos.

Aplicación a Arquitecturas Radiales

Los métodos BPDN y *lasso*, pueden aplicarse al entrenamiento de las arquitecturas radiales de la misma forma que se utilizó el método OMP anteriormente, es decir, construyendo un diccionario $\mathcal{G}$ con tantas funciones básicas, como datos de la muestra y diferentes valores para el parámetro de escala.

Nótese que, en este caso, no se utiliza regularización bayesiana ya que ambos métodos utilizan ya regularización con la norma absoluta.

En nuestras pruebas (véase a continuación), los dos métodos produjeron peores resultados que el algoritmo OLS con regularización local bayesiana, tanto en términos de parsimonia como de error conseguido en una muestra de validación.

A nuestro juicio, la razón de este hecho es la utilización de un sólo regularizador λ que trata a todas los átomos del diccionario por igual anulando las posibles ventajas de la norma absoluta frente a la cuadrática (ahora acompañada por regularización local).

La introducción en el algoritmo de diferentes regularizadores mediante un vector $\boldsymbol{\lambda}$ siguiendo el enfoque bayesiano no parece un problema trivial pues el estabilizador absoluto no es una función diferenciable, pero sus posibles beneficios hacen de ello una línea de investigación de interés.

II.6.8 Análisis de Parsimonia de los Distintos Algoritmos

En este apartado evaluamos la parsimonia de los diferentes algoritmos para redes neuronales analizados en los apartados anteriores.

Dicho análisis, sólo pretende evaluar *grosso modo* el comportamiento de cada algoritmo en este aspecto, pues una comparación precisa entre ellos requeriría considerar diferentes funciones de regresión en diferentes circunstancias y utilizando diferentes métodos de selección del modelo, lo cuál, no es objeto del presente trabajo y exigiría mucho mayor espacio.

Este tipo de análisis no tiene sentido cuando se utilizan técnicas de regresión no paramétricas como la regresión tipo núcleo, pues el número de funciones básicas está prefijado en el número de datos. Las redes neuronales, a pesar de que utilizan una filosofía no paramétrica debido a sus propiedades aproximadoras, son modelos paramétricos que sí permiten evaluar la fidelidad de su representación de una función de regresión, así como su parsimonia.

La función de regresión se eligió al azar como la siguiente combinación lineal de gausianas:

$$\mathtt{f}(x) = 0{,}1\exp\{-\frac{|x-0{,}1|^2}{2\cdot 0{,}1}\} + 0{,}7\exp\{-\frac{|x-0{,}5|^2}{2\cdot 2}\} + 0{,}5\exp\{-\frac{|x-0{,}9|^2}{2\cdot 1}\}$$

y la muestra aleatoria de entrenamiento, de tamaño $n = 30$, se generó perturbando dicha función con ruido independiente gausiano de media cero, utilizando un *signal-to-noise rate* (SNR) de valor 4, siendo:

$$\text{SNR} = \frac{\sum_{i=1}^{30}(\mathtt{f}_i - \bar{\mathtt{f}})^2}{\sum_{i=1}^{30}(e_i - \bar{e})^2}$$

donde $\mathtt{f}_i = \mathtt{f}(x_i)$ y e_i es una realización del ruido, $i = 1:30$.

La selección del modelo se realizó en cada caso, en base al error en una muestra de validación sin tener en cuenta el número de funciones básicas utilizadas.

La tabla 16 muestra para cada algoritmo utilizado, la media del error cuadrático sobre una muestra de test (ASRtest) y el número de funciones básicas obtenido (NFB). Todos los modelos son redes RBF con funciones básicas gausianas, salvo una red RBF con funciones básicas *mexican hat* ($\psi(x) = (1-x^2)e^{-x^2/2}$), y dos redes MLP entrenadas respectivamente, con *back-propagation* y algoritmo bayesiano, que se incluyen a efectos comparativos.

A pesar de que el algoritmo OMP coincide con el algoritmo OLS, hemos denominado OMP a la utilización del mismo algoritmo pero utilizando un diccionario de gausianas centradas en los datos con diferentes radios, al modo de lo propuesto en el epígrafe de la página 202, y OLS a la técnica tradicional en que seleccionan las funciones básicas gausianas de entre sólo las n posibles centradas en los datos.

Sobre dichos resultados, cabe realizar los siguientes comentarios:

1. Es interesante observar como, en este problema, la red RBF con funciones básicas *mexican hat* presenta una gran parsimonia coincidiendo su número de funciones básicas con el de la verdadera función de regresión (esta particularidad se ha observado en otros problemas simulados).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Parsimonia de Distintos Algoritmos		
Algoritmo	**ASRtest**	**NFB**
OLS	0,0017958	13
OLS (*mexican hat*)	0,0022657	3
Mixto - Gradiente Conjugado	0,0011508	3
OLS bayesiano local	0,0014635	5
OMP bayesiano global	0,0009152	5
OMP bayesiano local	0,0006435	5
BPDN	0,0012846	7
Lasso	0,0009929	8
SVM pérdida absoluta	0,0020858	27
SVM pérdida cuadrática	0,0017823	25
Back-propagation MLP	0,0011529	2
Bayesiano MLP	0,0012942	2

Tabla 16. Comparación de la parsimonia de diferentes algoritmos para RBF. Se muestra el error en una muestra de test (ASRtest) y el número de funciones básicas (NFB). Todas las redes RBF son gausianas salvo la que utiliza mexican hat. El gran número de vectores soporte en las SVM es producto de utilizar el valor óptimo de epsilon, pero un ligero incremento de su valor disminuye radicalmente dicho número sin incrementar significativamente el error

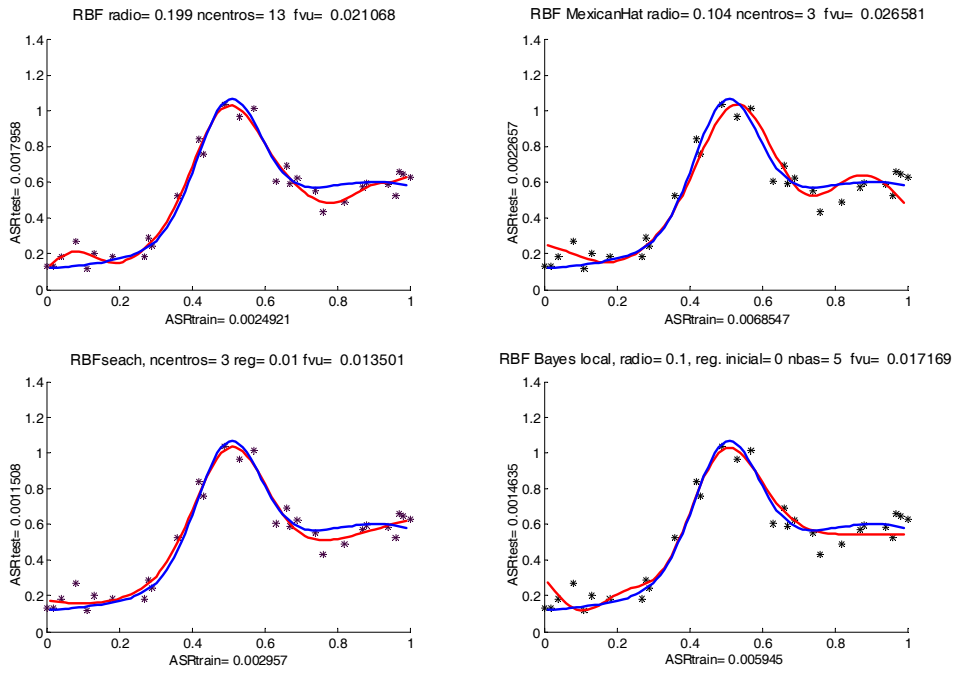


Figura 29. De izquierda a derecha y de arriba a abajo: ajustes realizados con los algoritmos OLS, OLS con *mexican hat*, algoritmo mixto y OLS con regularización bayesiana local.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

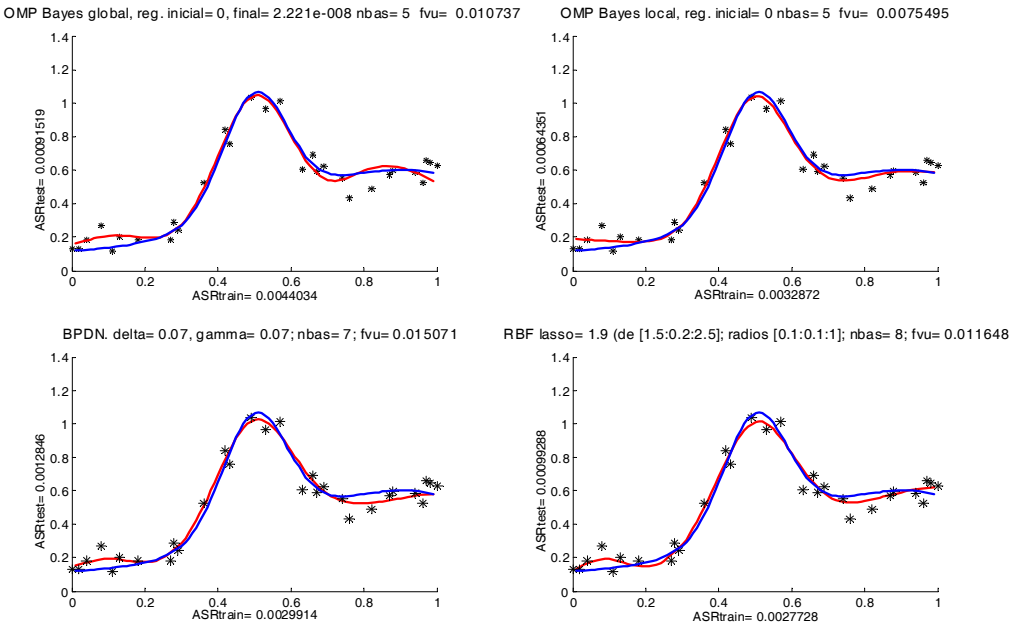


Figura 30. De izquierda a derecha y de arriba a abajo: ajustes realizados con los algoritmos OMP con regularización bayesiana global, OMP con regularización bayesiana local, BPDN y método *lasso*.

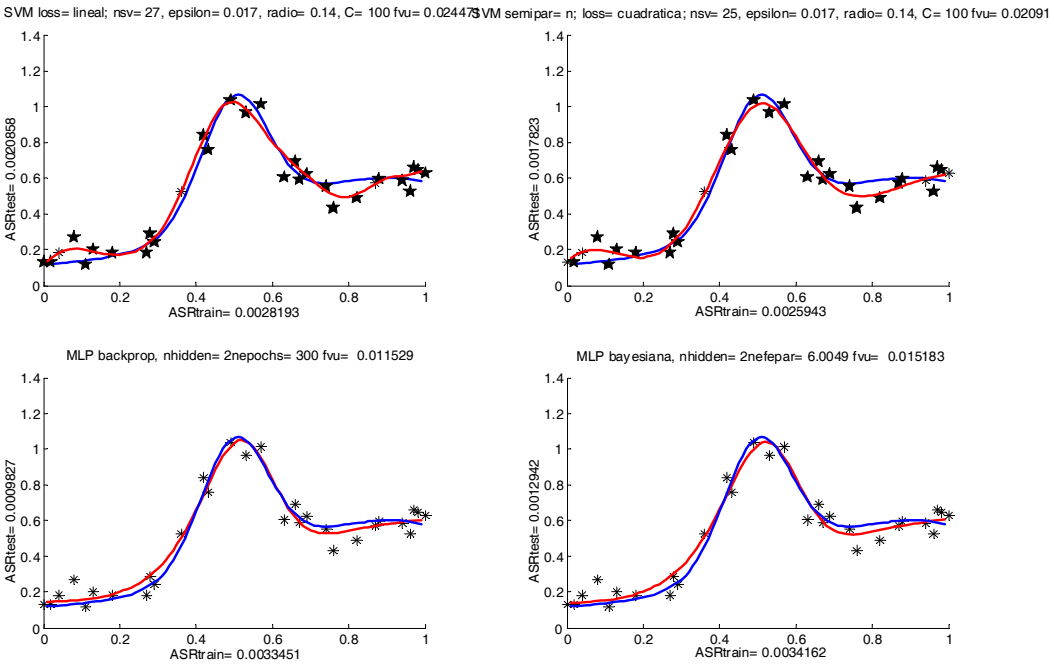


Figura 31. De izquierda a derecha y de arriba a abajo: ajustes realizados con los algoritmos SVM con pérdida absoluta, SVM con pérdida cuadrática, MLP con *back-propagation* y MLP con entrenamiento bayesiano (regularización).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2. Con muy buenos resultados en términos de error, el algoritmo mixto resultó el más parsimonioso de todos los empleados con una red RBF gausiana, coincidiendo su número de funciones básicas, con el de la verdadera función de regresión. La red RBF con funciones básicas *mexican hat* consigue el mismo grado de parsimonia, propiedad que también hemos observado en otras situaciones.

3. Se observan las ventajas del algoritmo OMP bayesiano con respecto al algoritmo OLS bayesiano: la única diferencia es la consideración de un diccionario mucho más flexible que incluye funciones básicas centradas en los datos con diferentes valores para el parámetro de escala, pero los beneficios son claros: con una misma parsimonia se obtiene un mejor comportamiento.

4. Los algoritmos *lasso* y BPDN no resultaron excesivamente parsimoniosos a pesar de utilizar la norma absoluta. Anteriormente, mencionamos que una posible razón para ello es la utilización de un parámetro regularizador único para todas las funciones básicas, a diferencia del enfoque bayesiano local. Otra razón es la utilización simultánea de todos los átomos del diccionario en el programa cuadrático correspondiente: ello provoca que muchas funciones básicas posean coeficientes no nulos suficientemente pequeños, cuya eliminación no afectaría de forma excesiva al error. El método de selección del modelo basado exclusivamente en la magnitud absoluta del error cometido, tiende a recoger estas funciones menos relevantes incrementando el número final sin un excesivo beneficio.

5. Sorprenden los resultados obtenidos con las redes SVM teniendo en cuenta las muestras de parsimonia referidas en la literatura para estas redes. En las pruebas realizadas, el incremento del margen permitido ε supuso una rápida reducción del número de vectores soporte, sin un coste excesivo en el error cometido. Para este tipo de algoritmos, al igual que para *lasso* y BPDN, es necesario elegir entre una extrema precisión y parsimonia.

   Por otra parte, sorprende también que, en contra de lo que suele suceder, la pérdida cuadrática produzca un menor número de funciones básicas que la pérdida absoluta.

6. En cualquier caso, ninguna de las redes RBF gausianas obtenidas reproduce los mismos centros y coeficientes de la verdadera función de regresión. Ello es debido al ruido de los datos y a la gran redundancia y flexibilidad de las mixturas gausianas que pueden aproximar distintas funciones muy parecidas utilizando diferentes parámetros.

7. Finalmente, destaca la enorme parsimonia de las redes MLP, sobre todo, teniendo en cuenta que sus funciones básicas son logísticas en contraste con las gausianas de la función de regresión.

II.7 Técnicas de Refinamiento del Aprendizaje. *Boosting*

II.7.1 Motivación

Los algoritmos presentados anteriormente están orientados al entrenamiento de una arquitectura específica postulada de antemano por muy diversas razones: rapidez de entrenamiento, inter-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

pretabilidad, buen comportamiento en determinados tipos de problemas, viabilidad ante problemas de gran dimensionalidad, etc.

Sin embargo, en muchos problemas de aplicación, no es fácil determinar cuál de todos ellos produce la mejor representación de la realidad que está detrás de un número limitado de datos, sobre todo, teniendo en cuenta que, la información que cada uno puede extraer de dichos datos es *esclava* de su arquitectura. Así, es posible que, en muchos problemas, la mejor representación se obtenga considerando conjuntamente las proporcionadas por diversos modelos alternativos.

Ello, unido a que, muchas veces, se dispone de modelos entrenados con conjuntos de datos distintos provinientes de la misma fuente, plantea la conveniencia de disponer de técnicas capaces de combinar la información proporcionada por cada uno, con la esperanza que los resultados mejoren los obtenidos con cada uno de los modelos particulares.

En los últimos años, las técnicas de combinación de estimadores (*ensembles*, *committes, voting algorithms*, *leveraging*, etc.), ha experimentado un gran desarrollo, apoyándose en el gran crecimiento de la potencia de los ordenadores.

A pesar de ello, según [Hoeting *et al.* 1999], la primera referencia sobre la utilización de este tipo de técnica se remonta a los años sesenta: [Barnard 1963] y una temprana revisión general sobre estas iniciativas puede verse en [Clemen 1989].

Las motivaciones que impulsaron el desarrollo y utilización de estas técnicas en problemas de clasificación y regresión, son diversas según las áreas de procedencia, aunque, todas ellas, poseen como objetivo común mejorar el comportamiento de los estimadores individuales.

Con ánimo de síntesis, podría decirse que esta gama de motivaciones determina los grandes grupos de técnicas existentes:

1. **Bagging**. Procedente del campo de la estadística es la siguiente argumentación de [Breiman 1996*b*] en su motivación del **bagging** [Breiman 1996*a*]: en un problema de regresión $\mathsf{Y} = \mathsf{f}(\mathsf{X}) + \varepsilon$ con ε ruido aleatorio independiente y de media cero, si $\mathsf{Z}^n = (\mathsf{Z}_1, ..., \mathsf{Z}_n)$ es una muestra aleatoria con $\mathsf{Z}_i = (\mathsf{X}_i, \mathsf{Y}_i)$, $i = 1 : n$, y $\hat{f}_n$ es una estimación de f a partir de los datos, el riesgo esperado puede descomponerse como:

$$\text{MPSE}[\hat{f}_n(\mathbf{x})] = \sigma^2(\mathbf{x}) + \text{Sesgo}^2_{\mathsf{Z}^n}[\hat{f}_n(\mathbf{x})] + \text{Var}_{\mathsf{Z}^n}[\hat{f}_n(\mathbf{x})]$$

   es decir, en el término ineludible debido al ruido de los datos, el cuadrado del sesgo y la varianza del estimador. Denominando $\hat{f}_A(\mathsf{X}) = \mathbb{E}_{\mathsf{Z}^n}(\hat{f}_n(\mathsf{X}))$ a la media del estimador, es claro que $\text{Var}_{\mathsf{Z}^n}(\hat{f}_A) = 0$ y, sin embargo el sesgo es el mismo:

$$\text{Sesgo}_{\mathsf{Z}^n}[\hat{f}_A(\mathbf{x})] = \mathbb{E}_{\mathsf{Z}^n}[\hat{f}_n(\mathbf{x})] - \mathsf{f}(\mathbf{x}) = \text{Sesgo}_{\mathsf{Z}^n}[\hat{f}_n(\mathbf{x})]$$

   Por tanto, $\text{MPSE}[\hat{f}_A(\mathbf{x})] < \text{MPSE}[\hat{f}_n(\mathbf{x})]$, con lo que una aproximación de $\hat{f}_A$ reduciría el riesgo esperado debido a una reducción en varianza. Una aproximación de este tipo, puede conseguirse mediante la media de un conjunto de estimadores obtenidos de otras tantas muestras.

   Por esta razón, este tipo de métodos se prescribe con estimadores *inestables* (con varianza relevante) que, además, son los que suelen poseer un menor sesgo.

2. **Stacking y Combinaciones Lineales Óptimas**. Su motivación se basa en que, diversos modelos estimados a partir de un conjunto de datos pueden recoger información diferenciada y relevante sobre éstos, que, sin embargo, se pierde después del filtro que resulta de un proceso

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

de selección del modelo. Esta consideración es aceptada generalmente, pero es aducida especialmente por los defensores de las **combinaciones lineales óptimas** (p. ej. [Hashem 1997]).

En el extremo de esta posición, se sitúan los autores (p. ej. [Wolpert 1992] motivando el denominado **stacking**), que rechazan *per se* la metodología de selección del modelo, al considerar que no existe tal *mejor modelo* y que dicha metodología supone elegir, sin justificación alguna, una combinación muy específica de estimadores: una combinación lineal en la que al modelo elegido se le asigna un coeficiente unidad y al resto, ceros.

Pero las combinaciones lineales de estimadores poseen también una justificación en términos de la capacidad de generalización. La siguiente argumentación de [Krogh y Vedelsby 1995] muestra que existen otros caminos, diferentes al cálculo de la media de estimadores obtenidos a partir de un conjunto finito de muestras, que permiten obtener combinaciones beneficiosas: si se dispone de r estimadores $\hat{f}_j$, $j = 1, ..., r$ y se considera su combinación convexa:

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{r} \alpha_j \hat{f}_j(\mathbf{x})$$

donde $\sum_{j=1}^{r} \alpha_j = 1$, podemos definir la denominada **ambigüedad** de la combinación como una especie de varianza interna ponderada:

$$\begin{aligned} a(\mathbf{x}) &= \sum_{j=1}^{r} \alpha_j [\hat{f}_j(\mathbf{x}) - \hat{f}(\mathbf{x})]^2 = \sum_{j=1}^{r} \alpha_j [\hat{f}_j(\mathbf{x}) - \mathsf{f}(\mathbf{x}) + \mathsf{f}(\mathbf{x}) - \hat{f}(\mathbf{x})]^2 \\ &= \sum_{j=1}^{r} \alpha_j [\hat{f}_j(\mathbf{x}) - \mathsf{f}(\mathbf{x})]^2 - 2[\hat{f}(\mathbf{x}) - \mathsf{f}(\mathbf{x})] \sum_{j=1}^{r} \alpha_j [\hat{f}_j(\mathbf{x}) - \mathsf{f}(\mathbf{x})] + \sum_{j=1}^{r} \alpha_j [\hat{f}(\mathbf{x}) - \mathsf{f}(\mathbf{x})]^2 \\ &= \sum_{j=1}^{r} \alpha_j [\hat{f}_j(\mathbf{x}) - \mathsf{f}(\mathbf{x})]^2 - [\hat{f}(\mathbf{x}) - \mathsf{f}(\mathbf{x})]^2 \Rightarrow \bar{e}(\mathbf{x})^2 = \sum_{j=1}^{r} \alpha_j e_j(\mathbf{x})^2 - a(\mathbf{x}) \end{aligned}$$

donde $\bar{e}_j(\mathbf{x})$ es el error de la combinación y $e_j(\mathbf{x})$ es el error de cada componente. Tomando esperanzas en X, se obtiene:

$$\mathbb{E}_{\mathsf{X}}[\bar{e}(\mathsf{X})^2] = \sum_{j=1}^{r} \alpha_j \mathbb{E}_{\mathsf{X}}[e_j(\mathsf{X})^2] - \mathbb{E}_{\mathsf{X}}[a(\mathsf{X})] \tag{II.141}$$

Esta expresión, es una forma de descomposición sesgo-varianza donde el primer término es la combinación de los cuadrados de los sesgos de los elementos y el segundo es la varianza interna esperada en el conjunto. Sin embargo, al contrario que la descomposición usual, esta descomposición muestra que la esperanza del error cuadrático de la combinación será menor cuanto mayor sea la variabilidad de las predicciones de los elementos con respecto a la predicción combinada.

3. **Bayesian Model Averaging**, (BMA). Con origen igualmente estadístico, esta motivación clásica procede de la estadística bayesiana. De manera similar a la filosofía del primer grupo, pero de una manera más fundamentada, esta postura considera que la metodología de selección del modelo ignora la incertidumbre asociada a dicho proceso produciendo una confianza excesiva e injustificada en el modelo resultante.

Como hemos analizado en capítulos anteriores, el enfoque bayesiano pretende obtener la distribución a posteriori de los parámetros $\boldsymbol{\theta}$ (donde $f(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta})$):

$$\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n) = \frac{\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta})\mathsf{p}(\boldsymbol{\theta})}{\mathsf{p}(\mathbf{y}^n)} \propto \mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta})\mathsf{p}(\boldsymbol{\theta})$$

donde $\mathsf{p}(\mathbf{y}^n|\boldsymbol{\theta})$ es la verosimilitud y $\mathsf{p}(\boldsymbol{\theta})$ la distribución a priori; y, a partir de él, predecir utilizando la distribución predictiva obtenida mediante marginalización:

$$\mathsf{p}(y|\mathbf{y}^n) = \int \mathsf{p}(y,\boldsymbol{\theta}|\mathbf{y}^n)d\boldsymbol{\theta} = \int \mathsf{p}(y|\mathbf{y}^n,\boldsymbol{\theta})\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n)d\boldsymbol{\theta} \tag{II.142}$$

Pero el procedimiento de marginalización es, esencialmente, un proceso de combinación lineal de los diferentes estimadores $\mathsf{p}(y|\mathbf{y}^n,\boldsymbol{\theta})$ ponderada por las probabilidades a posteriori $\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n)$.

En este marco bayesiano, la metodología clásica de selección del modelo ignora la incertidumbre representada por $\mathsf{p}(\boldsymbol{\theta}|\mathbf{y}^n)$ al elegir como aproximación a dicha esperanza la que resulta de asignar toda la probabilidad a un sólo modelo[77].

Este enfoque da lugar a las técnicas bayesianas de combinación de modelos conocidas bajo la denominación de *Bayesian Model Averaging* (BMA). Una visión general reciente puede verse en [Hoeting *et al.* 1999].

4. **Mixturas de Expertos**. En el marco de los esfuerzos por diseñar nuevas y más eficientes arquitecturas de redes neuronales, otro enfoque director de combinaciones de estimadores es el basado en la máxima *divide y vencerás* consistente en repartir la tarea de aprendizaje entre diversas redes especializadas.

   Este enfoque da lugar a las **mixturas de expertos** (*mixtures of experts*) de [Jacobs *et al.* 1991] y [Jordan y Jacobs 1994], que dividen un problema complejo en varios más sencillos, mediante la asignación de diferentes regiones del espacio de entrada a cada una de las redes componentes de la combinación (componentes de una mixtura aproximadora de la densidad de Y/X). Los pesos de la misma, se estiman mediante una red adicional (*gating network*) que, así, actúa como distribuidora de los datos de entrada entre las diversas redes componentes.

5. **Boosting**. Una motivación diferente surge del campo de la complejidad computacional del aprendizaje, a través del concepto de Aprendizaje PAC débil (*weak probably approximately correct learning*) [Kearns y Vazirani 1997], contrapunto del aprendizaje PAC visto en el primer capítulo (que, en este contexto, denominamos aprendizaje PAC fuerte).

   En su ya histórico artículo, [Schapire 1990] demuestra la equivalencia entre ambos tipo de aprendizaje mediante la construcción de un algoritmo que convierte (*boosts*), un algoritmo PAC débil (cuyo comportamiento es ligeramente mejor que la predicción aleatoria[78]) en un algoritmo PAC fuerte (que puede alcanzar una precisión tan grande como se quiera con alta probabilidad, pág. 19).

   Sucesivos refinamientos de este algoritmo ([Drucker *et al.* 1993], [Freund 1995], [Freund y Schapire 1997]), han dado lugar a una familia de técnicas de combinación de estimadores conocidas como *boosting*, que ha supuesto una gran fuente de nuevas ideas y puede con-

[77] Si se trata de aproximar, el método bayesiano utilizaría en todo caso la moda de dicha distribución (*maximum a posteriori* –MAP).

[78] En un problema de clasificación, basta que produzca hipótesis con tasas de error ligeramente menor que $1/2$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

siderarse, junto con el *bagging*, la principal responsable del auge actual de las técnicas de combinación de estimadores.

En cierto modo, la filosofía boosting está emparentada con la de las mixturas de expertos, pues dedican diversos modelos débiles a tareas difíciles y muy específicas.

II.7.2 Tipología

La expresión (II.141) muestra que si se combinan estimadores heterogéneos, es posible obtener beneficios de la combinación. Pero puede realizarse también un análisis en contrario: estimadores con predicciones similares o con errores correlados producirán menores ventajas en la combinación. La idea posee un fundamento intuitivo pues, la combinación de multitud de estimadores, todos iguales, no producirá beneficio alguno.

De cualquier forma, es relevante la cuestión planteada por [Sharkey 1999] sobre si es mejor combinar estimadores que produzcan errores poco correlados en un conjunto de validación, o combinar otros que no produzcan errores en dicho conjunto (pero quizás produzcan errores muy correlados en nuevos datos).

En consecuencia, un análisis general de la tipología de las combinaciones de estimadores descansa en dos cuestiones fundamentales [Sharkey 1999]: cómo generar miembros candidatos de combinaciones y cómo crear las combinaciones mismas:

1. **Técnicas de generación de candidatos**. En consonancia con lo anterior, el objetivo de estas técnicas, es conseguir miembros que produzcan predicciones o errores poco correlados. Métodos que pueden utilizarse para ello, son los siguientes:
   (a) Variación de las condiciones iniciales de los respectivos algoritmos.
   (b) Utilización de diferentes algoritmos.
   (c) Utilización de diferentes modelos o arquitecturas.
   (d) Utilización de diferentes conjuntos de datos en el entrenamiento. Este método es uno de los más utilizados, por ejemplo, por las siguientes técnicas:
      I. Remuestreo *bootstrap* y utilización de técnicas de validación cruzada entrenando diferentes estimadores con los distintos subconjuntos resultantes. Sería el caso de las técnicas que [Breiman 1996*b*], [Breiman 1998] denomina de *perturbar y combinar* (*perturbe and combine*), refiriéndose a las técnicas que, de alguna manera, introducen alguna forma de ruido en los datos. Las técnicas **bagging** ([Breiman 1996*a*], [Breiman 2001]) y **stacking** ([Wolpert 1992], [Breiman 1993]) serían también ejemplos de este grupo.
      II. Utilización de conjuntos disjuntos de entrenamiento, por ejemplo, mediante remuestreo sin reemplazamiento.
      III. Remuestreo adaptado según los resultados producidos por miembros anteriores de una combinación, con objeto de concentrar la actividad de los nuevos en los datos *difíciles* para los anteriores. Las técnicas más relevantes de este grupo serían las técnicas **boosting** denominadas **arcing** (*adaptive resampling and combine*) por [Breiman 1996*b*], [Breiman 1998].
      IV. Utilización de diferentes fuentes de datos o diferentes métodos de pre-proceso de los mismos.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2. **Técnicas de combinación**. Una vez elegidos los candidatos, el segundo problema es determinar la mejor manera de combinarlos. Los métodos más relevantes son los siguientes.
   (a) Métodos basados en media simple y media ponderada (**averaging and weighted averaging**, –*voting* en el problema de clasificación). Este grupo incluye:
      I. Las técnicas **bagging**, cuyos pesos se calculan *a priori* con valores $1/N$ donde N es el número de muestras *bootstrap* generadas, antes de la generación de los elementos de la combinación.
      II. Las técnicas **boosting**, cuyos pesos se calculan durante el proceso de generación de los miembros de la combinación. Como consecuencia, a diferencia del *bagging*, no pueden implementarse en paralelo.
      III. Las técnicas **bayesian model averaging** (BMA) esbozadas anteriormente. Las mayores dificultades de este tipo de métodos, son de tipo computacional en la estimación de las expresiones del tipo (II.142), pues requieren técnicas de aproximación tipo Montecarlo con cadenas de Markov o aproximaciones mediante la(s) moda(s) de la posterior. Una visión general y referencias se incluyen, por ejemplo, en [Hoeting *et al.* 1999] y [Neal 1996].
   (b) **Combinaciones lineales óptimas**. En algunos casos, estos métodos construyen combinaciones lineales ponderadas, pero también producen combinaciones mediante regresión lineal sin restricciones evitando la posible colinealidad entre los miembros. Técnicas empleadas para ello, son la regresión mediante componentes principales [Merz y Pazzani 1999], y métodos de eliminación de variables, ([Hashem 1997], [Belsley 1991]).
   (c) Las técnicas **stacking**, cuyos pesos se calculan *a posteriori*, a partir de versiones ya entrenadas de los miembros de la combinación. Realmente, las técnicas *stacking* pueden abarcar, incluso, combinaciones no lineales, pero no parecen aportar beneficios añadidos debido a su tendencia al sobreajuste.

A medio camino en complejidad entre los métodos basados en combinaciones lineales (sensibles a la presencia de estimadores poco precisos) y métodos más complicados como el *stacking* que constituye una especie de meta-aprendizaje sobre los resultados de estimadores de una fase anterior, se encuentran las técnicas de combinación basadas en estadísticos de orden, como la mediana, [Tumer y Ghosh 1999].

De entre las técnicas anteriores de combinación, el presente trabajo se centra en las técnicas *boosting*, que serán utilizadas en el capítulo siguiente. A continuación, realizamos un breve análisis sobre este tipo de técnicas.

II.7.3 *Boosting*. Reseña Histórica

Las técnicas *boosting* nacen como método para convertir un algoritmo **PAC débil** en un algoritmo **PAC fuerte** (ecuación (I.6), pág. 19). Básicamente, un algoritmo *débil* en el problema de clasificación es un algoritmo cuyas predicción es ligeramente mejor que una predicción aleatoria ($50\,\%$ de error en un problema de dos clases), es decir con riesgo $R(g) = 1/2$.

De esta forma, la condición PAC *débil* consiste en satisfacer la condición PAC para un determinado valor $\varepsilon \geq 1/2 - \gamma(n, v)$ con $\gamma(n, v) \in (0, 1/2)$ donde v es el alguna medida del tamaño o la complejidad de $\mathcal{F}$, y n es el tamaño muestral.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Algoritmo AdaBoost [Freund y Schapire 1997]

0. $B \in \mathbb{N}^+$, $\mathbf{d}^{(0)} \in [0,1]^n$ con $d_i^{(0)} = 1/n$, $i = 1:n$
1. Para $t = 1, ..., B$
   1.1 Determinar la nueva distribución: $\mathbf{d}^{(t)} = \mathbf{d}^{(t-1)} / \sum_{i=1}^{n} d_i^{(t-1)}$
   1.2 Ejecutar el algoritmo débil A con una muestra obtenida de $\mathbf{z}^n$ bajo la distribución $\mathsf{P}_{\mathbf{d}^{(t)}}$ obteniendo la hipótesis $\psi_t : \mathcal{X} \to [0,1]$
   1.3 Obtener el error de ψ_t: $\varepsilon_t = \sum_{i=1}^{n} d_i^{(t)} \, |\psi_t(\mathbf{x}_i) - y_i|$ y determinar $\beta_t = \dfrac{\varepsilon_t}{1-\varepsilon_t}$
   1.4 Terminar si $\varepsilon_t = 0$ o $\varepsilon_t = 1/2$
   1.5 Actualizar los pesos: $d_i^{(t+1)} = d_i^{(t)} \beta_t^{1-|\psi_t(\mathbf{x}_i)-y_i|}$
2. Hipótesis final: $g_B(\mathbf{x}) = \begin{cases} 1 & \text{si } \sum_{t=1}^{B} \log \frac{1}{\beta_t} (\psi_t(\mathbf{x}) - \frac{1}{2}) \geq 0 \\ 0 & \text{resto} \end{cases}$

Tabla 17. Algoritmo AdaBoost como adaptación del algoritmo Hedge al problema de clasificación.

Durante un tiempo, se conjeturó sobre si a partir de un algoritmo PAC débil, podría obtenerse un algoritmo PAC fuerte con un tiempo de ejecución y un tamaño muestral polinómicos (véase pág. 24).

La respuesta afirmativa al problema vino dada por [Schapire 1990] al construir un algoritmo que, mediante sucesivas utilizaciones de un algoritmo PAC débil, incrementaba sus prestaciones (*boosting*) dando lugar finalmente a un algoritmo PAC fuerte. La esencia de dicho algoritmo, consistía en aplicar repetidamente el algoritmo débil, a los datos en los que las ejecuciones anteriores del algoritmo habían producido error.

La historia reciente, pero intensa, de este tipo de técnicas, hace que los siguientes apuntes sobre el desarrollo del *boosting*, sean sólo una mero esbozo de los hitos más ilustrativos (un análisis más pausado que refleja el estado del arte de estas técnicas es [Meir y Rätsch 2003]):

1. **AdaBoost**. La siguiente versión de algoritmo *boosting* digna de mencionar en este breve resumen es el algoritmo **AdaBoost** (o AdaBoost discreto) de [Freund y Schapire 1997] para el problema de clasificación con codificación $\mathcal{Y} = \{0,1\}$. Dicho algoritmo (tabla 17) es el resultado de la adaptación al problema de aprendizaje de un algoritmo de distribución de recursos (algoritmo *Hedge*) entre diferentes estrategias como medio de combinarlas de manera óptima.

   AdaBoost produce una combinación lineal de hipótesis débiles, cada una de las cuáles se obtiene, en cada iteración t-ésima, como resultado de aplicar el algoritmo débil a los datos cuya probabilidad de aparición se ha alterado mediante unos pesos $\mathbf{d}^{(t)} \in [0,1]^n$.

   Finalmente, la regla de clasificación proporcionada por AdaBoost es el resultado del voto de los componentes de la combinación, ponderado por la importancia $\log(1/\beta_t)$ de cada hipótesis en la misma, proporcional a su acierto al clasificar los datos con la distribución $\mathbf{d}^{(t)}$.

   El algoritmo exige la utilización de algoritmos débiles cuyas hipótesis resultantes produzcan un error de clasificación $\varepsilon_t \in (0, 1/2)$. Cuando esta restricción no se cumple, el algoritmo se detiene.

   [Freund y Schapire 1997, th. 6] mostraron que AdaBoost era capaz de reducir rápidamente a cero el error en la muestra de entrenamiento. Asimismo, [Freund y Schapire 1997, th. 8] estudiaron su capacidad de generalización demostrando que la dimensión VC de la combi-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

nación resultante era finita en el caso de clasificadores base[79] dicotómicos pero proporcional al número de componentes de la combinación.

2. **Margen**. Debido a lo anterior, era de esperar que la hipótesis resultante de AdaBoost produjese sobreajuste después de un número suficiente de iteraciones. Sin embargo, durante un tiempo no se entendió suficientemente bien el magnífico comportamiento de AdaBoost en muchos experimentos realizados[80]. Véase la figura 32.

   La circunstancia más chocante de AdaBoost era su continua mejora de la capacidad de generalización a medida que aumentaban las iteraciones, incluso, después de haber conseguido clasificar correctamente todos los datos de la muestra de entrenamiento. Este hecho contradecía la experiencia acumulada sobre el significado de sobreajuste.

   Durante algún tiempo no se entendió suficientemente el buen comportamiento del *boosting*, pero un salto cualitativo en este entendimiento lo produjeron [Schapire *et al.* 1998], quiénes formularon las razones de dicho comportamiento en términos del denominado **margen** de un clasificador en una observación $(\mathbf{x}_i, y_i)$:

$$\rho_i \equiv \rho(\mathbf{x}_i) = y_i f(\mathbf{x}_i)$$

   donde f era la hipótesis final del algoritmo, que ahora poseía salida real. Con ello, el margen proporcionaba una medida de la confianza en la clasificación realizada, siendo ésta correcta si el margen es positivo, e incorrecta en caso contrario.

   [Schapire *et al.* 1998] estudiaron la evolución del margen de los ejemplos de la muestra durante las iteraciones del *boosting* y concluyeron que su buen comportamiento se debía a la maximización del margen mínimo en la muestra.

   Además, ratificaron sus expectativas aportando la siguiente cota al error de generalización:

Teorema 64 *[Schapire* et al. *1998, th. 2] Sean* P_{Z} *una distribución en* $\mathcal{X} \times \mathcal{Y}$ *con* $\mathcal{Y} = \{-1, 1\}$ *y una muestra aleatoria simple* $\mathbf{z}^n$ *con* $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i = 1 : n$ *extraída de* P_{Z}. *Sea* $\mathcal{F}$ *un espacio de hipótesis dicotómicas en* $\{-1, 1\}$ *con* $1 \leq v = \mathrm{VCdim}(\mathcal{F}) \leq n$, *y sea* f_B *una combinación obtenida mediante media ponderada de funciones de* $\mathcal{F}$ *produciendo márgenes* $\rho_i = y_i f_B(\mathbf{x}_i)$ *sobre los datos de la muestra. Entonces, para cualquier distribución* P_{Z}, *se verifica, con probabilidad al menos* $1 - \eta$ *en* $\mathsf{P}_{\mathsf{Z}^n}$ *que,* $\forall \theta \in (0, 1]$*:*

$$R(f_B) = \mathsf{P}_{\mathsf{Z}}[\mathsf{Y} f_B(\mathsf{X}) \leq 0] \leq \frac{1}{n} \sum_{i=1}^{n} 1_{\{\rho_i \leq \theta\}} + O\left(\frac{1}{\sqrt{n}} \sqrt{\frac{v \log^2(n/v)}{\theta^2} + \log(1/\eta)}\right) \qquad \text{(II.143)}$$

   Esta ecuación produce una cota similar a las cotas PAC o de la teoría del aprendizaje estadístico[81], pero, en lugar de utilizar el riesgo empírico en la derecha, utiliza un término empírico definido en términos del margen: la fracción de márgenes menor que un umbral θ.

   El segundo término de la cota, que disminuye al aumentar θ (para η, n constantes), es una medida del *tamaño* o *capacidad* de la familia de las combinaciones, en función de la capacidad de la familia de los estimadores base.

[79] Denominamos hipótesis débiles o hipótesis base a las producidas por el algoritmo débil en cada iteración del *boosting* que, así, son los elementos de la combinación.

[80] Debido a sus características, muchos de estos experimentos se realizaron aplicando *boosting* a árboles de clasificación y regresión (*classification and regression trees*, p. ej. [Breiman *et al.* 1984]), de poca complejidad (*stumps*), con el fin de obtener estimadores débiles.

[81] En las que no influye explícitamente la dimensionalidad del espacio de entrada, pero sí puede hacerlo a través de la medida de complejidad v.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

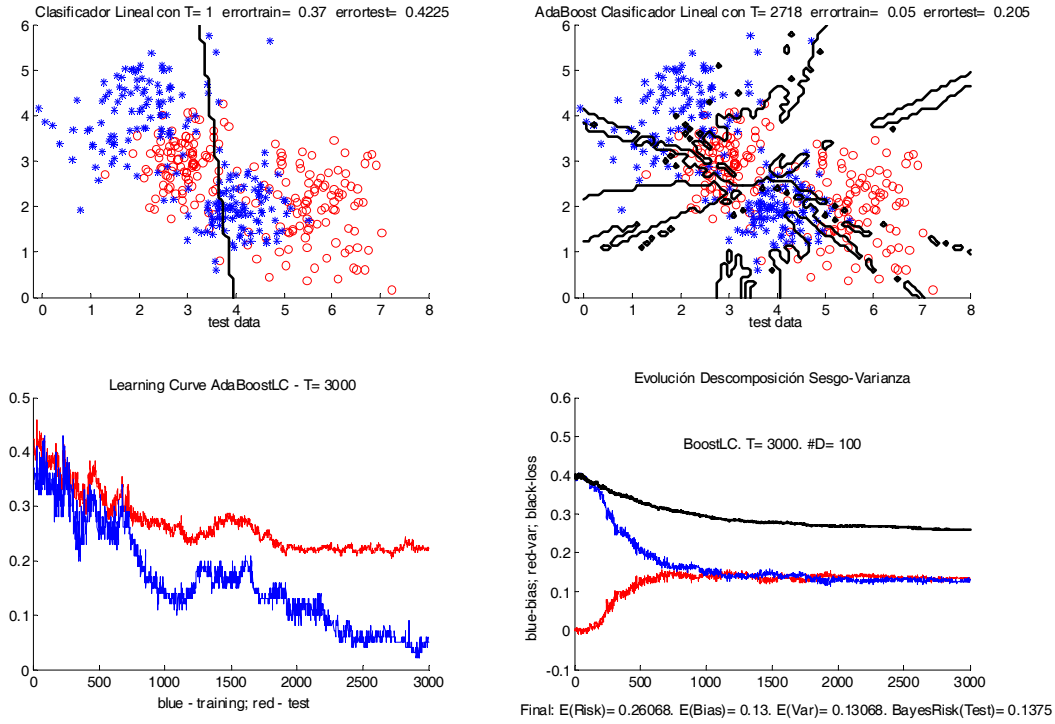


Figura 32. AdaBoost aplicado a una neurona adaline $y = \text{signo}(\mathbf{w}^T\mathbf{x} + w_0)$. La figura 32 muestra el efecto de AdaBoost sobre un clasificador débil como una neurona adaline utilizada para construir un clasificador *plug in* a partir de una estimación de las probabilidades a posteriori (pág. 85). En la fila superior, la figura de la izquierda muestra el sencillo clasificador producido por una neurona adaline, mientras que la figura de la derecha muestra la compleja regla de clasificación producida por AdaBoost después de 2728 iteraciones (otras tantas adaline en la combinación).

En la fila inferior, la figura de la izquierda muestra la evolución del error en la muestra de entrenamiento (línea inferior) y en una muestra independiente de test (línea superior). La figura de la derecha de la fila inferior muestra la evolución de la esperanza del riesgo ME (linea superior) y su descomposición sesgo-varianza vista en el capítulo primero para la pérdida 0-1 (teorema 14, pág. 32), estimados mediante un conjunto de 100 muestras de tamaño 100. La línea inferior es la varianza y la intermedia el sesgo. La figura indica que la mejora de AdaBoost se debe fundamentalmente a una disminución del sesgo (a costa de un incremento de la varianza que se estabiliza a partir de un punto).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Algoritmo AdaBoost [Schapire *et al.* 1998]
0. $B \in \mathbb{N}^+$, $\mathbf{d}^{(0)} \in [0,1]^n$ con $d_i^{(0)} = 1/n$, $i = 1, ..., n$
1. Para $t = 1, ..., B$
1.1 Ejecutar el algoritmo débil A con una muestra obtenida de $\mathbf{z}^n$ bajo la distribución $\mathsf{P}_{\mathbf{d}^{(t)}}$ obteniendo la hipótesis $\psi_t : \mathcal{X} \to \{0,1\}$
1.2 Obtener el error de ψ_t: $\varepsilon_t = \sum_{i=1}^{n} \alpha_i^t 1_{\{y_i \neq \psi_t(\mathbf{x}_i)\}}$ y determinar $\beta_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$
1.3 Terminar si $\varepsilon_t = 0$ o $\varepsilon_t = 1/2$
1.4 Actualizar los pesos: $d_i^{(t+1)} = d_i^{(t)} \exp\{-\beta_t y_i \psi_t(\mathbf{x}_i)\}/Z_t$ con Z_t tal que $\sum_{t=1}^{B} d_i^{(t+1)} = 1$. Se verifica que: $Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$
2. Hipótesis final: $g_B(\mathbf{x}) = \text{signo}[f_B(\mathbf{x})]$ con $f_B(\mathbf{x}) = \sum_{t=1}^{B} \beta_t \psi_t(\mathbf{x}) / \sum_{t=1}^{B} \beta_t$

Tabla 18. Algoritmo AdaBoost en términos del margen para clasificadores base dicotómicos.

En base a este resultado, [Schapire *et al.* 1998] interpretaron que la clave del éxito de AdaBoost residía en la maximización del margen mínimo alcanzado en los datos de la muestra pues, si f_B logra clasificar los ejemplos con un margen mínimo θ, cuanto mayor sea éste, menor será la cota anterior al riesgo del clasificador: el primer término es cero y el segundo disminuye con θ. Con ello, a mayor margen mínimo obtenido por la hipótesis final de *boosting*, menor cota al error de generalización.

Debido a ello, [Schapire *et al.* 1998] modificaron el algoritmo AdaBoost original formulándolo en términos del margen utilizando la codificación $\mathcal{Y} = \{-1, 1\}$, (tabla 18). Como puede observarse, en la iteración t-ésima el algoritmo asigna mayores pesos a los datos $(\mathbf{x}_i, y_i)$ para los que el clasificador obtenido en dicha iteración obtuvo menor margen:

$$d_i^{(t+1)} = \frac{d_i^{(t)} \exp\{-\beta_t y_i \psi_t(\mathbf{x}_i)\}}{Z_s} = \frac{\exp\{-y_i \sum_{s=1}^{t} \beta_s \varphi_s(\mathbf{x}_i)\}}{\prod_{s=1}^{t} Z_s}$$

con la hipótesis $\sum_{s=1}^{t} \beta_s \varphi_s(\mathbf{x}_i)$ combinada hasta la iteración t-ésima.

3. **Boosting Regularizado. Soft Margin**. Sin embargo, esta explicación no fué definitiva: diversos experimentos (p. ej. [Breiman 1998], [Grove y Schuurmans 1998]) mostraron que, en determinadas circunstancias, AdaBoost producía sobreajuste mientras progresaba en el incremento del margen (véase figura 33). Véase figura 33.

   De hecho, se diseñaron algoritmos específicamente dirigidos a maximizar directamente el margen (p. ej. el LP Boosting de [Grove y Schuurmans 1998] lo hacía mediante un programa lineal), y resultó que producían incluso peores resultados que AdaBoost. Por tanto, era claro que la maximización del margen era beneficiosa, pero no garantizaba, por si sólo, un buen comportamiento para AdaBoost en cualquier situación.

   Una explicación clarificadora vino dada por [Rätsch *et al.* 1998], los cuáles, concluyeron que, con el transcurso de las iteraciones, los ejemplos difíciles se convertían en los únicos puntos relevantes para el algoritmo, concentrando toda la ponderación de los datos. Con ello, el algoritmo se centraba en la clasificación correcta de dichos puntos, sacrificando la robustez de la clasificación.

   Con ello, estos puntos difíciles acababan determinando la frontera de clasificación y actuaban como vectores soporte de la misma, al modo de sus homónimos en las *support vector machines*.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Por tanto, [Rätsch *et al.* 1998] concluyó que AdaBoost realizaba la clasificación con una filosofía *hard margin* (pág. 142) intentando la separación perfecta de los datos de la muestra. Pero, este enfoque, ya se había revelado inadecuado en las SVM en escenarios con un nivel importante de ruido en los datos. Ello había provocado la introducción de la filosofía *soft margin* introduciendo variables *slack* en el programa cuadrático de las SVM, con objeto de introducir un mecanismo de desconfianza en los datos.

En consecuencia, [Rätsch *et al.* 1998] y [Rätsch *et al.* 1999] desarrollaron algoritmos *boosting* incorporando técnicas de regularización, con el fin de controlar la complejidad de la hipótesis final.

4. **Boosting con hipótesis fuertes**. Hasta el momento, el problema de AdaBoost y otros algoritmos *boosting* era la necesidad de utilizar hipótesis débiles pues, si en alguna iteración la hipótesis base obtenida poseía error $\varepsilon_t = 0$ 0 $\varepsilon_t = 0{,}5$, el algoritmo finalizaba. Para solventar este problema, [Onoda *et al.* 1998] propusieron una generalización de AdaBoost para tratar con hipótesis de mayor complejidad.

   Este algoritmo es análogo a AdaBoost (tabla 18), con la salvedad de que los coeficientes β_t en cada iteración se calculan mediante:

$$\beta_t = \ln(\frac{1-\varepsilon_t}{\varepsilon_t}\frac{\phi}{1-\phi}) = \ln\frac{1-\varepsilon_t}{\varepsilon_t} + \ln\frac{\phi}{1-\phi}$$

   y el algoritmo se para si $\varepsilon_t = 0$ o $\varepsilon_t = \phi$.

   Claramente, para $\phi = \frac{1}{2}$ se tiene el algoritmo AdaBoost como caso particular. El valor ϕ es una cota del error que puede cometer el algoritmo base utilizado, y puede estimarse a partir del error obtenido por una hipótesis base individual.

   El efecto de ϕ es penalizar en menor medida los datos mal clasificados, teniendo en cuenta que, las hipótesis complejas poseen capacidad suficiente para sobreajustarse a ellos, y ello tendería a producir sobreajuste. (Como normalmente ϕ será pequeño para hipótesis complejas, $\phi/(1-\phi)$ actúa reduciendo el argumento del logaritmo y así produciendo un valor positivo menor para β_t, con lo que la penalización de los datos mal clasificados por la hipótesis base correspondiente, será menor).

   Más tarde, [Rätsch *et al.* 2001, cor. 5] demostró que este algoritmo produce asintóticamente un margen mínimo ρ verificando:

$$\rho \geq \frac{\ln(\phi\varepsilon^{-1}) + \ln((1-\phi)(1-\varepsilon)^{-1})}{\ln(\phi\varepsilon^{-1}) - \ln((1-\phi)(1-\varepsilon)^{-1})}$$

   si $\varepsilon = \text{máx}_t\, \varepsilon_t \leq (1-\rho)/2$.

   Puede comprobarse que, de la expresión anterior se deduce que, para alcanzar un margen mínimo que sea grande, es necesario que ϕ y ε no sea muy parecidos y que, además, ϕ sea lo más pequeña posible. Con ello, a mayor complejidad de las hipótesis base, mayores posibilidades de conseguir un mayor margen mínimo asintótico de AdaBoost.

   Es decir, el *boosting* de hipótesis complejas con este tipo de algoritmos, tiende a producir un mayor margen mínimo que cuando se aplica a hipótesis débiles.

   Sin embargo, este margen mínimo se refiere a la muestra de entrenamiento, con lo que, con niveles importantes de ruido, un gran margen mínimo no tiene porque significar una buena capacidad de generalización.

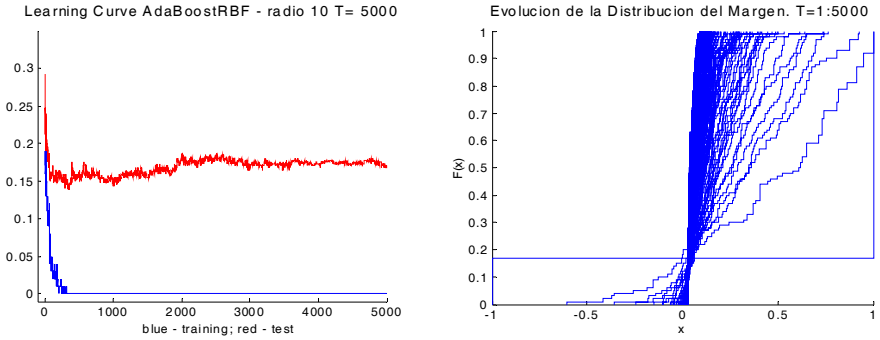


Figura 33. AdaBoost aplicado a una red RBF con un radio grande $\sigma = 10$, con el fin de limitar su complejidad. Izquierda: evolución de la curva de aprendizaje en entrenamiento (inferior) y en una muestra independiente de test (superior). Puede observarse como AdaBoost ha iniciado el camino del sobreajuste. Derecha: evolución de la distribución muestral del margen. La línea horizontal es el margen de la primera hipótesis que clasifica incorrectamente un $17\,\%$ de los datos. Durante el transcurso de las iteraciones, el margen mínimo se desplaza a la derecha terminando en un valor mayor que cero (todos los ejemplos se clasifican correctamente) a costa de que los ejemplos mejor clasificados (con margen elevado) vayan disminuyendo su margen, con lo que la frontera de clasificación va perdiendo robustez y se produce sobreajuste provocado por los puntos difíciles (de menor margen).

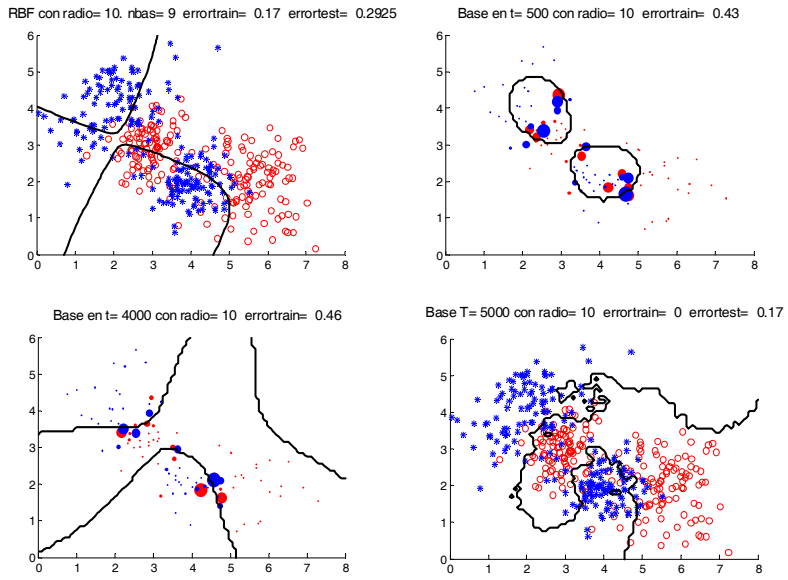


Figura 34. AdaBoost aplicado a una red RBF con $\sigma = 10$ (poca complejidad). Las tres primeras figuras (de izquierda a derecha y de arriba a abajo) muestran las hipótesis base correspondientes a las iteraciones $t = 1$, 500 y 4000. La correspondiente a $t = 1$ es el clasificador que produciría una RBF de este tipo con los datos originales. La segunda y tercera figuras muestra los puntos difíciles representados con puntos de grosor tanto más grande como mayor es su ponderación después de $t - 1$ iteraciones. El clasificador base correspondiente ha de basarse en dichos puntos (*vectores soporte*) pues la ponderación de los fáciles de clasificar tiende a cero. Con ello, la combinación corre el riesgo de sobreajuste. La última figura es la frontera de clasificación de AdaBoost en la iteración $t = 5000$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Algoritmo Real AdaBoost [Schapire y Singer 1999]
0. $B \in \mathbb{N}^+$, $\mathbf{d}^{(0)} \in [0,1]^n$ con $d_i^{(0)} = 1/n$, $i = 1, ..., n$ 1. Para $t = 1, ..., B$ 1.1 Obtener $(\psi_t, \alpha_t) = \arg \min_{\psi \in \mathcal{F}, \alpha_t \in \mathbb{R}} \sum_{i=1}^{n} d_i^{(t)} \exp\{-y_i \alpha_t \psi_t(\mathbf{x}_i)\}$ 1.2 Actualizar los pesos: $d_i^{(t+1)} = d_i^{(t)} \exp\{-\alpha_t y_i \psi_t(\mathbf{x}_i)\}/Z_t$ con Z_t tal que $\sum_{t=1}^{B} d_i^{(t+1)} = 1$. 2. Hipótesis final: $g_B(\mathbf{x}) = \text{signo}[f_B(\mathbf{x})]$ con $f_B(\mathbf{x}) = \sum_{t=1}^{B} \alpha_t \psi_t(\mathbf{x})$

Tabla 19. Algoritmo Real AdaBoost con hipótesis base con salida real. La justificación de la tarea 1.1 se trata en el anexo.

5. **Real AdaBoost**. Entretanto, [Schapire y Singer 1999] generalizaron el algoritmo AdaBoost a hipótesis base con salida real, dando lugar al algoritmo **Real AdaBoost** (tabla 19). En este algoritmo, la hipótesis base en cada iteración, se selecciona mediante la resolución del siguiente problema de optimización:

$$\psi_t = \arg \min_{\psi \in \mathcal{F}} \sum_{i=1}^{n} d_i^{(t)} \exp\{-y_i \psi_t(\mathbf{x}_i)\} \tag{II.144}$$

cuya expresión a minimizar (que incluye el margen $y_i \psi_t(\mathbf{x}_i)$ de la hipótesis base a seleccionar), se justifica mediante el siguiente resultado:

Teorema 65 *[Schapire y Singer 1999, th. 1] Si Z_t es la constante normalizadora de la tarea* 1.2 *de la tabla 19, se verifica la siguiente cota para el error de g_B sobre la muestra de entrenamiento:*

$$\frac{1}{n} \sum_{i=1}^{n} 1_{\{y_i \neq g_B(\mathbf{x}_i)\}} \leq \prod_{t=1}^{B} Z_t \text{ siendo } Z_t = \sum_{i=1}^{n} d_i^{(t)} \exp\{-\alpha_t y_i \psi_t(\mathbf{x}_i)\}$$

De la expresión (II.144), se deduce que el algoritmo Real AdaBoost puede verse como un algoritmo de tipo *greedy* que, en cada iteración, incorpora a la combinación la nueva componente que minimiza el criterio anterior.

II.7.4 *Boosting* como Algoritmo de Máxima Verosimilitud

Después de los resultados de [Rätsch *et al.* 1998] sobre las razones *hard margin* del sobreajuste de AdaBoost, el gran salto en el entendimiento estadístico del *boosting*[82] tuvo lugar con el artículo de [Friedman *et al.* 2000]. Su resultado más relevante fué el descubrimiento de que los algoritmos *boosting* construían, con una filosofía *greedy*, un modelo aditivo de regresión logística de la forma:

$$\frac{1}{2} \ln \frac{\mathbb{P}(\mathsf{Y} = 1|\mathbf{x})}{1 - \mathbb{P}(\mathsf{Y} = 1|\mathbf{x})} = \sum_{t=1}^{B} \psi_t(\mathbf{x})$$

que es proporcional en $1/2$ al modelo logístico aditivo usual.

[82] No se han mencionado aquí por falta de espacio, las aportaciones de [Breiman 1997*b*] y [Breiman 1999], ésta última sobre la dualidad entre los programas de maximización del margen mínimo y de minimización del denominado *edge* de una hipótesis base: $\gamma(\psi, \mathbf{d}) = \sum_{i=1}^{n} d_i \psi(\mathbf{x}_i)$.

Algoritmo LogitBoost [Friedman *et al.* 2000]

0. $B \in \mathbb{N}^+$, $\mathbf{d}^{(0)} \in [0,1]^n$ con $d_i^{(0)} = 1/n$, $i = 1:n$
   $\pi(\mathbf{x}_i) = \mathbb{P}(\mathsf{Y}^* = 1/\mathbf{x}_i) = \frac{1}{2}$, $i = 1:n$ con $\mathsf{Y}^* = (\mathsf{Y}+1)/2 \in \{0,1\}$
1. Para $t = 1, ..., B$
   1.1 Calcular para $i = 1, ..., n$:
   $$r_i = \frac{y_i^* - \pi(\mathbf{x}_i)}{\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)]};\ d_i^{(t)} = \pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)]$$
   1.2 Determinar:
   $$\psi_t = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} d_i^{(t)}[r_i - \psi(\mathbf{x}_i)]^2$$
   1.3 Actualizar para $i = 1, ..., n$:
   $$f_t(\mathbf{x}_i) = f_{t-1}(\mathbf{x}_i) + \tfrac{1}{2}\psi_t(\mathbf{x}_i);\ \pi(\mathbf{x}_i) = \frac{1}{1+\exp[-f_t(\mathbf{x}_i)]}$$
2. Hipótesis final: $g_B(\mathbf{x}) = \text{signo}[f_B(\mathbf{x})]$ con $f_B(\mathbf{x}) = \sum_{t=1}^{B} \psi_t(\mathbf{x})$

Tabla 20. Algoritmo LogitBoost

[Friedman *et al.* 2000, lema 1] demuestran que dicho modelo, minimiza la siguiente expresión del riesgo:

$$R(f) = \mathbb{E}_{\mathsf{Z}}[e^{-\mathsf{Y}f(\mathsf{X})}]$$

en el que se utiliza la pérdida exponencial $\ell(y, f(\mathbf{x})) = e^{-yf(\mathbf{x})}$.

Los resultados fundamentales de dicho trabajo son los siguientes que, por su interés, se incluyen con demostración en el Anexo F:

1. [Friedman *et al.* 2000, rtdo. 1] (teorema 211, Anexo F). En su versión poblacional, el algoritmo AdaBoost (discreto) construye un modelo aditivo de regresión logística mediante la minimización *greedy* de $R(f) = \mathbb{E}_{\mathsf{Z}}[e^{-\mathsf{Y}f(\mathsf{X})}]$, incorporando en cada iteración nuevos términos ψ_t, α_t a la solución $f_B(\mathbf{x}) = \sum_{t=1}^{B} \alpha_t \psi_t(\mathbf{x})$ mediante iteraciones similares a las del algoritmo de Newton (basadas en información de segundo orden).

   Además, en cada iteración, después de la actualización de las ponderaciones, el error de clasificación de la nueva hipótesis débil sobre la distribución ponderada de la muestra es del $50\,\%$.

2. [Friedman *et al.* 2000, rtdo. 2] (teorema 212, Anexo F). En su versión poblacional, el algoritmo Real AdaBoost construye un modelo aditivo de regresión logística mediante la minimización *greedy* aproximada de $R(f) = \mathbb{E}_{\mathsf{Z}}[e^{-\mathsf{Y}f(\mathsf{X})}]$.

   Además, en el óptimo f^*, la media condicionada de Y con la ponderación resultante para los datos, es cero.

3. **Algoritmo LogitBoost**. Teniendo en cuenta que la pérdida exponencial $\ell(y, f(\mathbf{x})) = e^{-yf(\mathbf{x})}$ no procede de ninguna verosimilitud pues no es el logaritmo de ninguna función de masa de probabilidad en $\{-1, 1\}$, [Friedman *et al.* 2000] proponen utilizar en su lugar la pérdida logística asociada a la verosimilitud binomial (ecuación (I.5), pág. 15):

   $$\ell(y, f(\mathbf{x}))) = \ln(1 + e^{-yf(\mathbf{x})})$$

   cuyo óptimo coincide con la pérdida logística.

   Con esta modificación, [Friedman *et al.* 2000, rtdo. 3] proponen el algoritmo **LogitBoost** (tabla 20), con la misma filosofía *greedy* que los anteriores, como resultado del teorema 213 (Anexo F).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Dicho teorema, establece además, que, dicho algoritmo, construye un modelo aditivo de regresión logística mediante la minimización *greedy* de la pérdida logística (maximización del logaritmo de la verosimilitud binomial), según iteraciones del algoritmo de Newton.

En su artículo, [Friedman *et al.* 2000] modifican el algoritmo AdaBoost siguiendo la misma metodología (algoritmo **Gentle AdaBoost**, tabla 47, Anexo F), y obtienen resultados similares para distintos algoritmos *boosting* diseñados para el problema de clasificación con varias clases.

Finalmente, proponen la técnica denominada *weigth trimming* consistente en eliminar de la muestra en cada iteración los datos con ponderaciones muy pequeñas, dado que no influyen en el entrenamiento de ψ_t y ello permite agilizar los cálculos (aunque el recálculo de dichas ponderaciones debe realizarse cada vez con todos los datos).

Algoritmo *Gradient_Boost*

Una de las mayores aportaciones de [Friedman *et al.* 2000] es la aportación de una metodología de construcción de algoritmos *boosting* para cualquier función de verosimilitud.

Una primera explotación de esta nueva vía vino dada por [Friedman 2001] que la aplicó al problema de regresión y la interpretó como un algoritmo de optimización en espacios funcionales (véase también [Mason *et al.* 2000]).

Antes de ello, se había intentado aplicar el *boosting* al problema de regresión adaptando los algoritmos diseñados para el problema de clasificación, (p. ej. [Freund y Schapire 1997], [Drucker 1997], [Avnimelech y Intrator 1999], [Ridgeway *et al.* 1999]), pero los algoritmos resultantes poseían excesivas reminiscencias del problema de clasificación lo que los hacía especialmente difíciles de implementar.

Siguiendo a [Friedman 2001], la determinación no paramétrica de la solución a un problema de aprendizaje:

$$\min_{f\in\mathcal{F}}\{R(f)=\mathbb{E}_{\mathsf{X},\mathsf{Y}}[\ell(\mathsf{Y},f(\mathsf{X}))]=\mathbb{E}_{\mathsf{X}}\mathbb{E}_{\mathsf{Y}/\mathsf{X}}[\ell(\mathsf{Y},f(\mathsf{X}))]\} \tag{II.145}$$

puede considerarse como un problema parámetrico con infinitos parámetros desconocidos $f(\mathbf{x})$ con $\mathbf{x}\in\mathcal{X}$, que puede plantearse para cada $\mathbf{x}\in\mathcal{X}$, como:

$$\min_{f(\mathbf{x})}\{r[f(\mathbf{x})]=\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\ell(\mathsf{Y},f(\mathbf{x}))]\}$$

En este marco, podemos utilizar el paradigma de la optimización numérica tradicional para obtener una solución del tipo:

$$f_B(\mathbf{x})=\sum_{t=1}^{B}\psi_t(\mathbf{x})$$

en la que cada $\psi_t(\mathbf{x})$ representa el avance (*boost*) en la minimización de la función objetivo, correspondiente a la iteración t-ésima del algoritmo. Si, por ejemplo, se utiliza el algoritmo del gradiente con *line search* (Anexo D), se tiene que:

$$\text{con:}\begin{cases} f_t(\mathbf{x})=f_{t-1}(\mathbf{x})+\psi_t(\mathbf{x}) \\ \psi_t(\mathbf{x})=-\rho_t\xi_{t-1}(\mathbf{x}) \\ \xi_{t-1}(\mathbf{x})=\nabla_{f(\mathbf{x})}r(f_{t-1}(\mathbf{x})) \\ \rho_t=\arg\min_{\rho} r[f_{t-1}(\mathbf{x})-\rho\xi_{t-1}(\mathbf{x})] \end{cases} \tag{II.146}$$

con lo que $-\xi_{t-1}(\mathbf{x})$ es la mejor dirección de avance en la iteración t-ésima.

USC UNIVERSIDADE

Suponiendo suficientes condiciones de regularidad, puede intercambiarse diferencial y esperanza, con lo que, en la iteración t-ésima:

$$\xi_{t-1}(\mathbf{x}) = \frac{\partial r}{\partial f(\mathbf{x})}[f_{t-1}(\mathbf{x})] = \mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\left.\frac{\partial \ell(\mathsf{Y}, f(\mathbf{x}))}{\partial f(\mathbf{x})}\right|_{f(\mathbf{x})=f_{t-1}(\mathbf{x})}] \tag{II.147}$$

Sin embargo, en la práctica la distribución de $\mathsf{Y}/\mathbf{x}$ no es conocida y sólo se dispone de un conjunto de datos $\mathbf{z}^n$ con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i = 1:n$ con lo que la versión empírica del algoritmo anterior tendría que determinar infinitas incógnitas $f^*(\mathbf{x})$ con sólo un número finito de componentes $\xi_{t-1}(\mathbf{x})$ en las direcciones de avance.

Una posible solución podría consistir en estimar dicha distribución para cada $\mathbf{x}$, apoyándose en los puntos $\mathbf{x}_i$ de la muestra *cercanos* al punto $\mathbf{x}$ de interés, pero, una alternativa usual, es aprovechar la información proporcionada por dichos puntos, imponiendo cierto grado de suavidad a la solución, por ejemplo, postulando un modelo del tipo $f(\mathbf{x}) = \sum_{j=1}^{h} \alpha_j \psi(\mathbf{x}; \mathbf{w}_j)$.

Con ello, la versión empírica del problema (II.145) adoptaría la forma:

$$\{(\alpha_j, \mathbf{w}_j)\}_{j=1}^{h} = \arg \min_{\{(\alpha'_j, \mathbf{w}'_j)\}_{j=1}^{h}} \sum_{i=1}^{n} \ell(y_i, \sum_{j=1}^{h} \alpha'_j \psi(\mathbf{x}_i; \mathbf{w}'))$$

ante el cuál, puede utilizarse un enfoque *greedy*, con o sin *back-fitting*, en el que, para cada $t = 1:h$,

$$\begin{aligned} (\alpha_t, \mathbf{w}_t) &= \arg \min_{\alpha, \mathbf{w}} \sum_{i=1}^{n} \ell(y_i, f_{t-1}(\mathbf{x}_i) + \alpha \psi(\mathbf{x}_i; \mathbf{w})) \\ \Rightarrow \quad f_t(\mathbf{x}) &= f_{t-1}(\mathbf{x}) + \alpha_t \psi(\mathbf{x}; \mathbf{w}_t) \end{aligned} \tag{II.148}$$

Con ello, en la línea mostrada por [Friedman *et al.* 2000], los algoritmos *boosting* serían casos particulares sin *back-fitting* (*stage-wise*) de este enfoque *greedy* general, utilizando diferentes pérdidas (exponencial, logística, etc.), y donde las componentes $\alpha_t \psi(\mathbf{x}; \mathbf{w}_t)$ serían las hipótesis base producidas por el algoritmo débil utilizado.

Bajo el esquema conceptual de optimización funcional introducido anteriormente, cada componente $\alpha_t \psi(\mathbf{x}; \mathbf{w}_t)$ podría verse como un paso del algoritmo de optimización (II.146) en su versión empírica, con la restricción de que la dirección funcional $-\xi_{t-1}$ de optimización debe ser un miembro de la familia considerada:

$$-\xi_{t-1} = -\nabla_{f(\mathbf{x})} r(f_{t-1}(\mathbf{x})) \in \mathcal{F}_\psi = \{\alpha \psi(\mathbf{x}; \mathbf{w}) : \alpha \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^s\} \tag{II.149}$$

De acuerdo con la versión empírica de (II.147), cada iteración del algoritmo avanzaría en un espacio n-dimensional según la dirección definida por el vector $-\boldsymbol{\xi}_t \in \mathbb{R}^n$ con componentes:

$$-(\boldsymbol{\xi}_t)_i = -\xi_t(\mathbf{x}_i) = -\left.\frac{\partial \ell(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\right|_{f(\mathbf{x}_i)=f_{t-1}(\mathbf{x}_i)}, i = 1:n$$

y una forma de salvar en lo posible la restricción (II.149) de su pertenencia a la familia $\mathcal{F}_\psi$, sería elegir la dirección de optimización $\alpha_t \psi(\mathbf{x}; \mathbf{w}_t)$ más correlada con aquélla. Es decir, mediante:

$$\mathbf{w}_t = \arg \max_{\mathbf{w}, \alpha} \langle -\boldsymbol{\xi}_t, \alpha \boldsymbol{\psi}_{\mathbf{w}} \rangle_2 = \arg \min_{\mathbf{w}, \alpha} \sum_{i=1}^{n} [-\xi_t(\mathbf{x}_i) - \alpha \psi(\mathbf{x}_i; \mathbf{w})]^2 \tag{II.150}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Algoritmo Gradient_Boost [Friedman 2001]
0. $f_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{n} \ell(y_i, \rho)$, ρ constante
1. Para $t = 1, ..., B$, determinar:
$-\xi_t(\mathbf{x}_i) = -\left.\dfrac{\partial \ell(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\right\|_{f(\mathbf{x}_i)=f_{t-1}(\mathbf{x}_i)}, i = 1:n$
$\mathbf{w}_t = \arg\min_{\mathbf{w},\alpha} \sum_{i=1}^{n} [-\xi_t(\mathbf{x}_i) - \alpha\psi(\mathbf{x}_i;\mathbf{w})]^2$
$\rho_t = \arg \min_{\{\rho:\rho\psi\in\mathcal{F}_\psi\}} \sum_{i=1}^{n} \ell(y_i, f_{t-1}(\mathbf{x}_i) + \rho\psi(\mathbf{x}_i;\mathbf{w}_t))$
2. Hipótesis final: $f_B(\mathbf{x}) = \sum_{t=1}^{B} \rho_t \psi(\mathbf{x};\mathbf{w}_t)$

Tabla 21. Algoritmo Gradient Boost.

Algoritmo LS_Boost [Friedman 2001]
0. $f_0 = 0$
1. Para $t = 1, ..., B$, determinar:
1.1 $\xi_i(\mathbf{x}_i) = y_i - f_{t-1}(\mathbf{x}_i)$, $i = 1:n$
1.2 $(\rho_t, \mathbf{w}_t) = \arg\min_{\mathbf{w},\rho} \sum_{i=1}^{n} [\xi_i(\mathbf{x}_i) - \rho\psi(\mathbf{x}_i;\mathbf{w})]^2$
2. Hipótesis final: $f_B(\mathbf{x}) = \sum_{t=1}^{B} \rho_t \psi(\mathbf{x};\mathbf{w}_t)$

Tabla 22. Algoritmo LS Boost

donde $(\boldsymbol{\psi}_{\mathbf{w}})_i = \psi(\mathbf{x}_i;\mathbf{w})$, $i = 1, : n$. Con ello, la búsqueda en la dirección $\boldsymbol{\psi}_{\mathbf{w}}$ (*line search*) y la actualización, se realizarían respectivamente, mediante:

$$\begin{aligned} \rho_t &= \arg \min_{\{\rho:\rho\psi\in\mathcal{F}_\psi\}} \sum_{i=1}^{n} \ell(y_i, f_{t-1}(\mathbf{x}_i) + \rho\psi(\mathbf{x}_i;\mathbf{w}_t)) \\ f_t(\mathbf{x}) &= f_{t-1}(\mathbf{x}) + \rho_t\psi(\mathbf{x};\mathbf{w}_t) \end{aligned} \tag{II.151}$$

Con ello, como indica [Friedman 2001], en lugar de obtener la solución bajo una restricción de suavidad definida por el modelo elegido $f(\mathbf{x}) = \sum_{j=1}^{h} \alpha_j \psi(\mathbf{x};\mathbf{w}_j)$, dicha restricción se aplica a la dirección de avance (al menos *gradiente* efectivamente utilizado). Ello permite sustituir una optimización (II.148) posiblemente complicada, por un problema de mínimos cuadrados (II.150) (u otro criterio que permita estimar la esperanza condicionada por $\mathbf{x}$) para estimar (suavizado) el menos gradiente, más un problema de optimización en términos de un sólo parámetro (II.151).

El esquema anterior da lugar a un algoritmo marco general que [Friedman 2001] denomina **Gradient_Boost** y que se presenta en la tabla 21.

En el marco general del Gradient_Boost, pueden utilizarse diferentes pérdidas que darían lugar a otros tantos algoritmos. [Friedman 2001] formula el correspondiente a la pérdida cuadrática en su forma general, denominado **LS_Boost** (tabla 22), que implementa con árboles de regresión (*regression trees*, [Breiman *et al.* 1984]), así como los correspondientes a la pérdida absoluta $\ell(y, f(\mathbf{x})) = |y - f(\mathbf{x})|$ y a la pérdida robusta de Huber, [Huber 1964]:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$$\ell(y, f(\mathbf{x})) = \begin{cases} \frac{1}{2}(y - f(\mathbf{x}))^2 & |y - f(\mathbf{x})| \le \varepsilon \\ \varepsilon(|y - f(\mathbf{x})| - \frac{\varepsilon}{2}) & |y - f(\mathbf{x})| > \varepsilon \end{cases}$$

El potencial del enfoque combinado de [Friedman *et al.* 2000] y [Friedman 2001] reside en la posibilidad de poder utilizar diferentes pérdidas en el algoritmo, sin mayores modificaciones. Por ejemplo, [Ridgeway 1999] utiliza el marco del Gradient_Boosting para formular un algoritmo genérico con funciones de pérdida basadas en la verosimilitud de la familia exponencial.

En términos de lo tratado (pág. 220) sobre las condiciones a verificar por los candidatos a una combinación de estimadores, el proceso de combinación de hipótesis seguido por Gradient_Boost y sus casos particulares, asegura la baja correlación entre las diferentes hipótesis base mediante la descomposición sucesiva de la magnitud de los datos, en lugar de, por ejemplo, mediante la partición del espacio de entrada en diferentes regiones.

LS_Boost con Hipótesis Lineales

En un reciente trabajo [Bühlmann y Yu 2001] explota las características del algoritmo LS_Boost con hipótesis base lineales, para producir un conjunto de resultados de relevancia tanto computacionales como estadísticos.

Sea el modelo $\mathsf{Y}_i = \mathsf{f}(\mathbf{x}_i) + \varepsilon_i$, $i = 1 : n$, con $\{\varepsilon_i\}_{i=1}^n$ independientes e idénticamente distribuidas de media cero y $\mathrm{Var}(\varepsilon_i) = \sigma^2$, y sea una muestra aleatoria $\mathbf{y} = (y_1, ..., y_n)^T$ donde y_i son las respuestas correspondientes a los valores de $\mathbf{x}_i$, $i = 1 : n$.

En este marco, el problema de estimación de una hipótesis puede considerarse como la determinación del operador inverso (pág. 11) $\mathcal{S} = \mathcal{O}^{-1} : \mathbb{Y} \subset \mathbb{R}^n \rightarrow \mathbb{F} \subset \mathbb{R}^n$ donde $\mathbb{F}$ es la restricción del espacio de hipótesis $\mathcal{F}$ a sus valores en los puntos $\{\mathbf{x}_i\}_{i=1}^n$, que transforma un vector $\mathbf{y}$ en su estimación $\hat{\mathbf{f}} = \mathcal{S}(\mathbf{y})$ (donde el término $\mathcal{S} = \mathcal{S}_{\mathbf{x}^n}$ incluye su dependencia de los valores $\mathbf{x}^n = (\mathbf{x}_1, ..., \mathbf{x}_n)$).

El algoritmo LS_Boost, determina la siguiente hipótesis base en cada iteración, ajustando por mínimos cuadrados el gradiente negativo en la iteración t-ésima, es decir, el vector de residuos $\mathbf{e}_t = \mathbf{y} - \hat{\mathbf{f}}_t$ donde $\hat{\mathbf{f}}_t$ es la estimación de LS_Boost después de dicha iteración. Con ello, dicho algoritmo realiza (tabla 22):

$$\begin{aligned} \mathbf{e}_t &= \mathbf{y} - \hat{\mathbf{f}}_t = \mathbf{e}_{t-1} - \mathcal{S}(\mathbf{e}_{t-1}) \\ &= (I - \mathcal{S})\mathbf{e}_{t-1} = (I - \mathcal{S})^2\mathbf{e}_{t-2} = \cdots = (I - \mathcal{S})^t\mathbf{e}_0 = (I - \mathcal{S})^t\mathbf{y} \\ \Rightarrow\ & \hat{\mathbf{f}}_t = \mathbf{y} - (I - \mathcal{S})^t\mathbf{y} = [I - (I - \mathcal{S})^t]\mathbf{y}, t = 1 : B \end{aligned}$$

Así, si $\mathcal{S}$ es el operador utilizado para estimar cada hipótesis base, el algoritmo LS_Boost después de B iteraciones puede asociarse al operador $\mathcal{B}_B$ definido por:

$$\mathcal{B}_B(\mathbf{y}) \doteq [I - (I - \mathcal{S})^B]\mathbf{y} \tag{II.152}$$

Un ejemplo relevante de lo anterior, es cuando $\mathcal{S}$ es un operador lineal, que suponemos en lo que sigue, es decir, $\mathcal{S}(\mathbf{y}) = \mathbf{S}\mathbf{y}$ para alguna matriz *hat* $\mathbf{S}$, que incluye el caso de las redes RBF entrenadas mediante algoritmos de regresión lineal con $\mathbf{S} = \mathbf{\Psi}(\mathbf{\Psi}^T\mathbf{\Psi})^{-1}\mathbf{\Psi}$ con $(\mathbf{\Psi})_{ij} = \psi_j(\mathbf{x}_i)$, (pág. 119) y las redes RBF de regularización (pág. 135).

Si $\{\mu_j\}_{j=1}^n$ son los autovalores de $\mathcal{S}$, entonces, debido a (II.152), los autovalores de $\mathcal{B}_B$ son $\{1 - (1 - \mu_j)^t, j = 1 : n\}$. Con ello, si $\mathcal{S}$ es un operador proyección, también lo es $\mathcal{B}_B$, lo que sucede con las redes RBF mencionadas.

En el caso general en que $\mathcal{S}$ sea un operador lineal autoadjunto (es decir, $\mathcal{S} = \mathcal{S}^+$, que poseen autovalores reales y, por analogía con la matriz $\mathbf{S}$ asociada, denominamos simétricos en adelante),

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

también lo es $\mathcal{B}_B$ con lo que éste puede descomponerse como:

$$\mathcal{B}_B = \mathbf{U}\mathbf{\Lambda}_B\mathbf{U}^T \text{ con } \mathbf{\Lambda}_B = \text{diag}[\{1-(1-\mu_j)^B\}_{j=1}^n]$$

donde $\mathbf{U}$ es la matriz ortonormal con columnas los correspondientes autovectores normalizados de $\mathcal{S}$.

Basándose en lo anterior, [Bühlmann y Yu 2001, prop. 3] analiza la evolución de la descomposición sesgo-varianza de LS_Boost en función del número de iteraciones, produciendo los siguientes resultados, en los que denotamos (pág. 34):

$$\begin{aligned}
\text{AMSE}(\hat{f}_B;\mathbf{x}^n) &= \mathbb{E}_{\mathbf{Y}^n/\mathbf{x}^n}[\text{ASE}(\hat{f}_B;\mathbf{x}^n)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{Y}^n/\mathbf{x}^n}\{[\hat{f}_B(\mathbf{x}_i)-\mathsf{f}(\mathbf{x}_i)]^2\} \\
&= \frac{1}{n}\sum_{i=1}^n [\mathbb{E}_{\mathbf{Y}^n/\mathbf{x}^n}(\hat{f}_B(\mathbf{x}_i))-\mathsf{f}(\mathbf{x}_i)]^2 \\
&\quad + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{Y}^n/\mathbf{x}^n}\{[\hat{f}_B(\mathbf{x}_i)-\mathbb{E}_{\mathbf{Y}^n/\mathbf{x}^n}(\hat{f}_B(\mathbf{x}_i))]^2\} \\
&= \frac{1}{n}\sum_{i=1}^n \text{Sesgo}^2_{\mathbf{Y}^n/\mathbf{x}^n}[\hat{f}_B(\mathbf{x}_i)] + \frac{1}{n}\sum_{i=1}^n \text{Var}_{\mathbf{Y}^n/\mathbf{x}^n}[\hat{f}_B(\mathbf{x}_i)]
\end{aligned}$$

siendo $\hat{f}_B$ la hipótesis producida por LS_Boost después de la iteración B-ésima.

Teorema 66 *[Bühlmann y Yu 2001, prop. 3] En el marco del modelo de regresión anterior, sea $\mathcal{S}$ un algoritmo lineal y simétrico con autovalores $\{\mu_j\}_{j=1}^n$ y correspondiente matriz de autovectores $\mathbf{U}$. Sea $\mathbf{f} = (\mathsf{f}(\mathbf{x}_1),...,\mathsf{f}(\mathbf{x}_n))^T$. Entonces:*

$$\begin{aligned}
\text{Sesgo}^2(B,\mathcal{S},\mathbf{f}) &\doteq \frac{1}{n}\sum_{i=1}^n \text{Sesgo}^2_{\mathbf{Y}^n/\mathbf{x}^n}[\hat{f}_t(\mathbf{x}_i)] = \frac{1}{n}\mathbf{f}^T\mathbf{U}\,\text{diag}[\{(1-\mu_j)^{2B}\}_{j=1}^n)\mathbf{U}^T\mathbf{f} \\
\text{Var}(B,\mathcal{S},\sigma^2) &\doteq \frac{1}{n}\sum_{i=1}^n \text{Var}_{\mathbf{Y}^n/\mathbf{x}^n}[\hat{f}_t(\mathbf{x}_i)] = \frac{\sigma^2}{n}\sum_{j=1}^n [1-(1-\mu_j)^B]^2 \\
\text{AMSE}(B,\mathcal{S};\mathbf{f},\sigma^2) &\doteq \text{AMSE}(\hat{f}_B;\mathbf{x}^n) = \text{Sesgo}^2(B,\mathcal{S},\mathbf{f}) + \text{Var}(B,\mathcal{S},\sigma^2)
\end{aligned}$$

Como puede observarse, el número de iteraciones B del algoritmo LS_Boost, actúa como parámetro de control de la complejidad que permite realizar un balance entre los términos de sesgo y de varianza. El siguiente resultado ilustra la evolución de dichos términos en función del número de iteraciones de LS_Boost:

Teorema 67 *[Bühlmann y Yu 2001, th. 1] En las condiciones del teorema anterior, si además se satisface que $0 < \mu_j \leq 1$, $j = 1:n$, entonces, cuando $B \to \infty$:*

1) $\text{Sesgo}^2(B,\mathcal{S},\mathbf{f})$ *decrece exponencialmente*
2) $\text{Var}(B,\mathcal{S},\sigma^2)$ *crece exponencialmente con menor tasa*
3) $\lim_{B\to\infty} \text{AMSE}(B,\mathcal{S};\mathbf{f},\sigma^2) = \sigma^2$

Además, si $\mathbf{f}_\mathbf{U} = \mathbf{U}^T\mathbf{f} = (\mathsf{f}_1,...,\mathsf{f}_n)^T$ es el vector $\mathbf{f}$ expresado en términos de los autovectores, entonces:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1. *Si* $\mathsf{f}_j^2/\sigma^2 > 1/(1-\mu_j)^2 - 1$ *para todo* j *con* $\mu_j < 1$, *entonces LS_Boost mejora el* AMSE *con respecto a la hipótesis que produciría* $\mathcal{S}$.

2. *Si* $\mu_j < 1$ *para al menos un* $j \in \{1:n\}$ *(*$\mathcal{S}$ *no es el operador identidad), existe un* B *tal que, en la iteración* B*-ésima,* $\mathrm{AMSE}(B, \mathcal{S}; \mathbf{f}, \sigma^2) < \sigma^2$.

La primera condición anterior, puede interpretarse como que, si la complejidad relativa de f en la dirección j con respecto a σ^2 (f_j^2/σ^2), es mayor que el grado de suavizado $1/(1-\mu_j)^2-1$ que $\mathcal{S}$ emplea en dicha dirección, entonces, LS_Boost mejora los resultados del algoritmo individual.

Es decir, para que LS_Boost produzca mejoras, es necesario que, o bien f_j^2 sea complejo con respecto al nivel de ruido σ^2 o bien que $\mathcal{S}$ sea una hipótesis débil en esa dirección. Ello, sugiere un buen comportamiento de LS_Boost con algoritmos débiles.

[Bühlmann y Yu 2001] analiza el caso particular de los *smoothing splines*, definidos como la solución del problema de regresión regularizado:

$$\min_{f \in \mathcal{W}_2^r(\mathcal{X})} \{\hat{\mathcal{R}}_n(f) = \hat{R}_n(f) + \lambda \|f\|^2_{\mathcal{W}_2^r(\mathcal{X})}\}$$

donde $\mathcal{X} = [a,b]$ y $\mathcal{W}_2^r(\mathcal{X})$ con $\mathcal{X} = [a,b] \subset \mathbb{R}$ es el espacio de Sobolev[83] (págs. 363 y 382, Anexo C) y $\|f\|^2_{\mathcal{W}_2^r(\mathcal{X})} = \int_a^b [f^{(r)}(x)]^2 dx$, que es seminorma en $\mathcal{W}_2^r(\mathcal{X})$.

Para este caso particular de redes RBF de regularización, [Bühlmann y Yu 2001, th. 3] prueba que, para cualquier λ fijo, existe un $B = B(n) = O(n^{2r/(2s+1)}) \to \infty$ cuando $n \to \infty$, tal que, $\hat{f}_{B(n)}$ alcanza la tasa $n^{-2s/(2s+1)}$ de convergencia a $\mathsf{f} \in \mathcal{W}_2^s(\mathcal{X})$ con $s \leq r$ ($r, s \in \mathbb{N}$) arbitrario[84], en términos del AMSE.

Finalmente, [Bühlmann y Yu 2001] obtiene un resultado similar para el problema de clasificación, que trata como un problema de regresión heterocedástico: $\mathsf{Y}_i = \mathbb{E}_{\mathbf{Y}/\mathbf{x}_i}(\mathsf{Y}) + \varepsilon_i$, $i = 1:n$, donde $\mathsf{Y}_i \in \{-1, 1\}$, $\mathbb{E}_{\mathbf{Y}/\mathbf{x}}(\mathsf{Y}) = 2\pi(\mathbf{x}) - 1$ con $\pi(\mathbf{x}) = \mathbb{P}(\mathsf{Y} = 1|\mathbf{x})$, y ε_i son variables aleatorias independientes de media cero y varianza $\mathrm{Var}(\varepsilon_i) = 4\pi(\mathbf{x}_i)[1 - \pi(\mathbf{x}_i)]$, considerando reglas de decisión del tipo $g_B(\mathbf{x}) = \mathrm{signo}[f_B(\mathbf{x})]$.

Como resultado, obtiene que para un algoritmo *smoothing spline*, se verifica que:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}[g_B(\mathbf{x}_i) \neq \mathsf{Y}_i] - R(\mathsf{g}) = O(n^{-s/(2s+1)})$$

donde $R(\mathsf{g})$ es el riesgo de la regla de Bayes con la pérdida 0-1.

II.7.5 Relación entre *Boosting* y Otros Algoritmos

Boosting y Algoritmos *Greedy*

Como hemos ya hemos observado con anterioridad, el algoritmo Gradient_Boost, es un caso particular de algoritmo *greedy* puro. En concreto, el algoritmo LS_Boost es un algoritmo *greedy* puro con pérdida cuadrática (*matching pursuit*, pág. 199).

[83] Este espacio es el espacio de Hilbert con seminúcleo reproductor k núcleo de Bessel-MacDonalds, (pág. 382).

[84] Para $r = s$, el *smoothing spline* alcanza dicha tasa, pero LS_Boost permite cualquier función verdadera más suave ($s \leq r$).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Por lo tratado en el Apartado II.6.6, sabemos que estos algoritmos son subóptimos pues no producen la mejor combinación lineal de las hipótesis base utilizadas en la combinación (*stage-wise*).

Versiones optimizadas de estos algoritmos son las que emplean *back-fitting* persiguiendo que los residuos finales sean ortogonales al espacio lineal $\left\langle \{\boldsymbol{\psi}_t\}_{t=1}^h \right\rangle \subset \mathbb{R}^n$ donde $(\boldsymbol{\psi}_t)_i = \psi_t(\mathbf{x}_i)$, $i = 1:n$, $t = 1:h$. Ejemplos de estos algoritmos son el algoritmo *orthogonal least squares* (OLS, pág. 174) y el algoritmo *orthogonal matching pursuit* (OMP, pág. 200), vistos en secciones anteriores como algoritmos de entrenamiento de las redes neuronales, que así, son implementaciones con *back-fitting* del algoritmo LS_Boost, en el que, en cada iteración se introduce un nueva función básica.

De hecho, la única diferencia entre las redes neuronales entrenadas con los algoritmos *greedy* considerados en los apartados anteriores, y las hipótesis producidas con el *boosting*, reside en la pérdida utilizada y en que, éste último, no utiliza *back-fitting*. Si se utiliza la pérdida cuadrática, los algoritmos mencionados son equivalentes y, por tanto, lo son las redes que aquéllos determinan.

El análisis anterior de [Bühlmann y Yu 2001] generaliza esta situación (sin *back-fitting*) al caso en el que, cada *función básica* es un estimador lineal, por ejemplo, una red RBF completa. En este caso, se obtendría también una red RBF aunque con un entrenamiento diferente.

De esta forma, es aplicable al algoritmo LS_Boost todo lo tratado en la presentación de dichos algoritmos y, en particular, las técnicas de regularización con enfoque bayesiano (que podría extenderse al algoritmo Gradient_Boost más general), y los resultados de consistencia utilizando redes MLP o RBF, presentados para este tipo de algoritmos[85].

Boosting y SVM

Centrándonos por ejemplo en el algoritmo AdaBoost de [Schapire *et al.* 1998] (tabla 18), la hipótesis combinada final del algoritmo era:

$$f_B(\mathbf{x}) = \frac{\sum_{t=1}^B \beta_t \psi_t(\mathbf{x})}{\sum_{t=1}^B \beta_t} = \frac{\langle \boldsymbol{\beta}, \boldsymbol{\psi}(\mathbf{x}) \rangle}{\|\boldsymbol{\beta}\|_1}$$

donde $\boldsymbol{\psi}(\mathbf{x})$ tiene por componentes las hipótesis base $\psi_t(\mathbf{x})$, $t = 1:B$ y $\boldsymbol{\beta} = (\beta_1, ..., \beta_B)^T$ cuyas componentes $\beta_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$ son todas positivas al ser $\varepsilon_t \in (0, 1/2)$.

Por tanto, el margen que dicho clasificador produce en un punto $\mathbf{x}$ determinado, perteneciente a la clase $y \in \{-1, 1\}$ es:

$$\gamma(\mathbf{x}) = y f_B(\mathbf{x}) = \frac{y \langle \boldsymbol{\beta}, \boldsymbol{\psi}(\mathbf{x}) \rangle}{\|\boldsymbol{\beta}\|_1} \tag{II.153}$$

Por otra parte, cuando presentamos las *support vector machines* (pág. 139) (y el perceptrón generalizado, pág. 79), vimos que la distancia (margen geométrico) entre un punto y un hiperplano (en el espacio de características) $\tau_{\mathbf{w},b} \equiv \{\mathbf{x} \in \mathcal{X} : \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle + b = 0\}$, venía dado por:

$$\gamma(\mathbf{x}) = \frac{|\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle + b|}{\|\mathbf{w}\|_2} = \frac{y(\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle + b)}{\|\mathbf{w}\|_2} \tag{II.154}$$

donde $\mathbf{w}$ era el vector normal al hiperplano y $\boldsymbol{\phi}$ era la transformación del espacio de entrada en el espacio de características, de tal forma que $\langle \boldsymbol{\phi}(\mathbf{x}'), \boldsymbol{\phi}(\mathbf{x}) \rangle = k(\mathbf{x}', \mathbf{x})$. Igualmente, definíamos

[85] Otros resultados recientes sobre la consistencia del *boosting* son: [Lugosi y Vayatis 2002], [Lugosi y Vayatis 2001], [Lugosi y Vayatis 2003], [Zhang y Yu 2003].

el margen (geométrico) de un hiperplano con respecto a una muestra, como el mínimo margen geométrico con respecto a cada uno de los puntos de la muestra.

Comparando las expresiones (II.153) y (II.154), y asociando $\mathbf{w}$ con $\boldsymbol{\beta}$ y $\boldsymbol{\psi}(\mathbf{x})$ con $\boldsymbol{\phi}(\mathbf{x})$, puede concluirse que:

1. Las dos técnicas (*boosting* y SVM) transforman el espacio de entrada en un espacio de gran dimensionalidad[86], (espacio de características o *features*) y construyen en él un clasificador lineal mediante un hiperplano que trata de separar los datos de la muestra con un gran margen: las SVM *hard margin* son solución a un problema de maximización del margen y la pérdida de AdaBoost es la pérdida exponencial $\ell(y, f(\mathbf{x})) = e^{-yf(\mathbf{x})}$.

   Dicha similitud ilustra la necesidad de controlar la complejidad de AdaBoost (y en general de las técnicas *boosting*), al igual que en las SVM se motivó la filosofía *soft margin* en un ambiente de ruido en los datos.

2. Mientras en las SVM, se utiliza la norma euclídea. para la normalización el *boosting* utiliza la norma absoluta.

Esta última circunstancia puede interpretarse a la luz del siguiente teorema:

Teorema 68 *[Mangasarian 1999] Sea $\mathbf{x}_0 \in \mathbb{R}^d$ cualquier punto no situado en el plano $\tau_{\mathbf{w},b} = \{\mathbf{x} : |\langle \mathbf{w}, \mathbf{x}\rangle + b| = 0\}$. Entonces, para $p \in [1, \infty]$ y $1/p + 1/q = 1$:*

$$\frac{|\langle \mathbf{w}, \mathbf{x}\rangle + b|}{\|\mathbf{w}\|_p} = \min_{\mathbf{x}\in\tau_{\mathbf{w},b}} \|\mathbf{x} - \mathbf{x}_0\|_q$$

donde el término de la derecha es la distancia entre $\mathbf{x}$ y el hiperplano medida según la norma dual de ℓ_q.

Por tanto, el teorema anterior indica que, mientras las SVM utilizan la distancia euclídea. para calcular el margen, el *boosting* utiliza la norma infinita.

Este contraste, posee su correlato en las correspondientes estrategias de ambas técnicas para gestionar la enorme dimensionalidad de los respectivos espacios de características: mientras el *boosting* utiliza un enfoque *greedy* explorando coordenada a coordenada dicho espacio (lo que puede verse como un programa lineal aproximado p. ej. [Grove y Schuurmans 1998]), las SVM utilizan el método del núcleo (*kernel trick*) con $\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\rangle = k(\mathbf{x}, \mathbf{x}')$ para gestionar dicha dimensionalidad y resolver el problema de la determinación del hiperplano de máximo margen en dicho espacio.

Finalmente, procede preguntarse por qué no utilizar la metodología SVM utilizando como espacio de características el construido por AdaBoost, $\boldsymbol{\phi}(\mathbf{x}) \equiv \boldsymbol{\psi}(\mathbf{x})$, y utilizar con él un núcleo de la forma:

$$k(\mathbf{x}, \mathbf{x}') = \big\langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}')\big\rangle = \sum_{t=1}^{B} \psi_t(\mathbf{x})\psi_t(\mathbf{x}')$$

Una razón [Rätsch 2002], es que, teniendo en cuenta que las hipótesis base de AdaBoost verifican $\psi_t(\mathbf{x}) \in \{-1, 1\}$, sería imposible en espacios de dimensión infinita controlar la dimensión VC del clasificador resultante pues, no existiría una bola de radio D con la norma euclídea., que contuviese a los datos transformados, para así poder aplicar el teorema 44 (pág. 141). Sin embargo, si se utiliza la norma infinita, esas distancias se mantienen finitas.

[86] Nótese que el *boosting* puede poseer miles de hipótesis en la combinación.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

II.7.6 Análisis del Comportamiento del *Boosting* con Redes RBF-OLS en un Problema de Clasificación

Finalizamos este capítulo con un análisis del comportamiento del *boosting* en un problema de clasificación con pérdida 0-1, utilizando como hipótesis base redes RBF entrenadas con el algoritmo OLS.

Para ello, analizaremos la evolución de la descomposición sesgo-varianza de las hipótesis resultantes, en un problema simulado.

En nuestro conocimiento, los únicos experimentos de aplicación del *boosting* a las redes RBF son [Rätsch *et al.* 1999] y [Rätsch *et al.* 2001], en los que se entrenan dichas redes mediante el algoritmo del gradiente conjugado (algoritmo mixto) y se limita su complejidad restringiendo su número de funciones básicas. En dichos experimentos, las redes RBF individuales presentaron un comportamiento sólo ligeramente inferior al *boosting*.

Las redes RBF así entrenadas poseen, para un mismo número de funciones básicas, un mayor nivel de complejidad que las redes RBF entrenadas mediante OLS, pues poseen radios diferentes en las distintas funciones básicas y sus centros no están restringidos a los datos de la muestra (Apartado II.6.3).

Como consecuencia, las redes RBF con algoritmo mixto, tienden a poseer una mayor componente de varianza, amplificada, además, por la propia aleatoriedad del algoritmo de optimización no lineal empleado (su inicialización se realiza mediante el algoritmo de *k-medias*, también no lineal).

Por ello, es interesante conocer, si un algoritmo en teoría más estable como RBF con OLS, podría competir mejor con el *boosting*, o si éste puede aportar mejoras adicionales a aquél, y, en todo caso, cuál es el grado de complejidad más adecuado en este tipo de algoritmos, para que el *boosting* produzca sus mejores resultados, suponiendo que algoritmo es detenido a tiempo antes del sobreajuste (por ejemplo con una muestra de validación).

Descomposición Sesgo-Varianza con Pérdida 0-1

Dado que será la base del análisis a realizar, a continuación reproducimos el siguiente resultado de [Domingos 2000] que concreta el resultado del teorema 14 (pág. 32) para el caso particular de pérdida 0-1.

Corolario 69 *[Domingos 2000] En el problema de clasificación anterior, si $\mathcal{D}$ es un conjunto de muestras aleatorias de entrenamiento obtenidas de la distribución $\mathsf{P}_{\mathsf{X},\mathsf{Y}}$, y $\hat{g}_n$ es el estimador obtenido a partir de una muestra aleatoria $\mathsf{Z}^n = (\mathsf{Z}_1, ..., \mathsf{Z}_n)$, con $\mathsf{Z}_i = (\mathsf{X}_i, \mathsf{Y}_i)$, $i = 1 : n$, se verifica la siguiente descomposición de la esperanza del error de predicción:*

$$MPE(\hat{g}_n(\mathbf{x})) = E_{\mathbf{Z}^n,\mathbf{Y}/\mathbf{x}}[\ell(\mathsf{Y}, \hat{g}_n(\mathbf{x}))] = c_1\sigma^2(\mathbf{x}) + \text{Sesgo}(\hat{g}_n(\mathbf{x})) + c_2\,\text{Var}(\hat{g}_n(\mathbf{x})) \quad \text{(II.155)}$$

donde:

$$c_1 = c_1(\mathbf{x}) = 2\mathbb{P}_{\mathbf{Z}^n}[\hat{g}_n(\mathbf{x}) = \mathsf{g}(\mathbf{x})] - 1$$

$$c_2 = c_2(\mathbf{x}) = \begin{cases} +1 & \textit{si } \hat{g}^*_{\mathcal{D}}(\mathbf{x}) = \mathsf{g}(\mathbf{x}) \\ -1 & \textit{en otro caso} \end{cases}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

son constantes dependientes de $\mathbf{x}$ *(y también del algoritmo), y:*

$$\begin{aligned} \mathrm{Sesgo}(\hat{g}_n(\mathbf{x})) &\doteq \ell(\mathsf{g}(\mathbf{x}), \hat{g}^*_{\mathcal{D}}(\mathbf{x})) \\ \mathrm{Var}(\hat{g}_n(\mathbf{x})) &\doteq \mathbb{E}_{\mathsf{Z}^n}[\ell(\hat{g}^*_{\mathcal{D}}(\mathbf{x}), \hat{g}_n(\mathbf{x})] \\ \sigma^2(\mathbf{x}) &\doteq \mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\ell(\mathsf{Y}, \mathsf{g}(\mathbf{x}))] \end{aligned}$$

donde $\hat{g}^*_{\mathcal{D}}$ *es la* ***predicción principal*** *en* $\mathcal{D}$ *definida como:*

$$\hat{g}^*_{\mathcal{D}}(\mathbf{x}) \doteq \arg\min_{g(\mathbf{x})} E_{\mathsf{Z}^n}[\ell(g(\mathbf{x}), \hat{g}_n(\mathbf{x}))]$$

que, en el problema de clasificación, es la moda de las predicciones de las hipótesis entrenadas con muestras de $\mathcal{D}$*.*

Análogamente, se define:

$$\mathrm{ME}(\hat{g}_n(\mathbf{x})) = \mathrm{Sesgo}(\hat{g}_n(\mathbf{x})) + c_2 \,\mathrm{Var}(\hat{g}_n(\mathbf{x})) \tag{II.157}$$

y su correspondiente valor medio (y análogamente para el MPE):

$$\mathrm{ME}(\hat{g}_n) = \mathbb{E}_{\mathsf{X}}[\mathrm{ME}(\hat{g}_n(\mathsf{X}))] = \mathbb{E}_{\mathsf{X}}[\mathrm{Sesgo}(\hat{g}_n(\mathsf{X}))] + \mathbb{E}_{\mathsf{X}}[c_2(\mathsf{X})\,\mathrm{Var}(\hat{g}_n(\mathsf{X}))] \tag{II.158}$$

A diferencia de la descomposición en regresión, el término de varianza, puede resultar negativo debido a que $c_2(\mathbf{x}) = -1$ en los puntos donde existe sesgo: $\mathrm{Sesgo}(\hat{g}_n(\mathbf{x})) = \ell(\mathsf{g}(\mathbf{x}), \hat{g}^*_{\mathcal{D}}(\mathbf{x})) = 1$. Ello significa que en estos puntos, cuanto más varianza menor error de generalización. Claramente, esto no ocurre en el caso de regresión con pérdida cuadrática.

Por otra parte, es peculiar el término del ruido, pero su expresión sigue rigurosamente el paralelismo con la descomposición con pérdida cuadrática. Para la regla de Bayes, su valor resulta intuitivo al ser el riesgo de Bayes en $\mathbf{x}$:

$$c_1(\mathbf{x})\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\ell(\mathsf{Y}, \mathsf{g}(\mathbf{x}))] = (2\mathbb{P}_{\mathsf{Z}^n}[\mathsf{g}(\mathbf{x}) = \mathsf{g}(\mathbf{x})] - 1)\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\ell(\mathsf{Y}, \mathsf{g}(\mathbf{x}))] = \mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\ell(\mathsf{Y}, \mathsf{g}(\mathbf{x}))]$$

Sin embargo, para cualquier otro algoritmo entrenado con muestras de $\mathcal{D}$, dicho valor resulta menor que dicho riesgo si $\mathbb{P}_{\mathsf{Z}^n}[\hat{g}_n(\mathbf{x}) = \mathsf{g}(\mathbf{x})] < 1$ e, incluso negativo, si $\mathbb{P}_{\mathsf{Z}^n}[\hat{g}_n(\mathbf{x}) = \mathsf{g}(\mathbf{x})] < 0{,}5$ con lo que, además, depende del algoritmo utilizado. Así, si el estimador está lejos de ser óptimo, cuanto mayor ruido, mejor es su comportamiento.

[Domingos 2000] estudia el comportamiento de su propuesta de descomposición con árboles de regresión (*regression trees*) y con clasificadores *k-nearest neighbors*, y anima a estudiar dicho comportamiento con otros estimadores.

El problema de clasificación que hemos simulado, es un problema de dos clases, cada una de ellas constituida por dos *clusters* gausianos con la siguiente distribución de mixtura en cada clase (figura 35):

$$\begin{aligned} \mathsf{p}_{\mathsf{X}/y=-1}(\mathbf{x}) &= \tfrac{1}{2}\mathsf{p}(\mathbf{x}; \boldsymbol{\mu}_{11}, \Sigma_1) + \tfrac{1}{2}\mathsf{p}(\mathbf{x}; \boldsymbol{\mu}_{12}, \Sigma_2) \\ \mathsf{p}_{\mathsf{X}/y=+1}(\mathbf{x}) &= \tfrac{1}{2}\mathsf{p}(\mathbf{x}; \boldsymbol{\mu}_{21}, \Sigma_2) + \tfrac{1}{2}\mathsf{p}(\mathbf{x}; \boldsymbol{\mu}_{22}, \Sigma_1) \\ \boldsymbol{\mu}_{11} = (2,4)^t,\ \boldsymbol{\mu}_{12} &= (4,2)^t; \qquad \Sigma_1 = \mathrm{diag}(0{,}75, 0{,}75) \\ \boldsymbol{\mu}_{21} = (3,3)^t,\ \boldsymbol{\mu}_{22} &= (5{,}5, 2)^t; \quad \Sigma_2 = \mathrm{diag}(0{,}5, 0{,}5) \end{aligned} \tag{II.159}$$

donde $\mathsf{p}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ es la densidad normal de media $\boldsymbol{\mu}$ y matriz de varianzas-covarianzas $\boldsymbol{\Sigma}, \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Para la simulación, hemos considerado el conjunto $\mathcal{D} = \{(\mathbf{z}^n)_j,\ j = 1, ..., N\}$ de $N = 100$ muestras de entrenamiento, de tamaño $n = 100$ cada una, con igual número de datos en cada clase generados aleatoriamente según (II.159).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

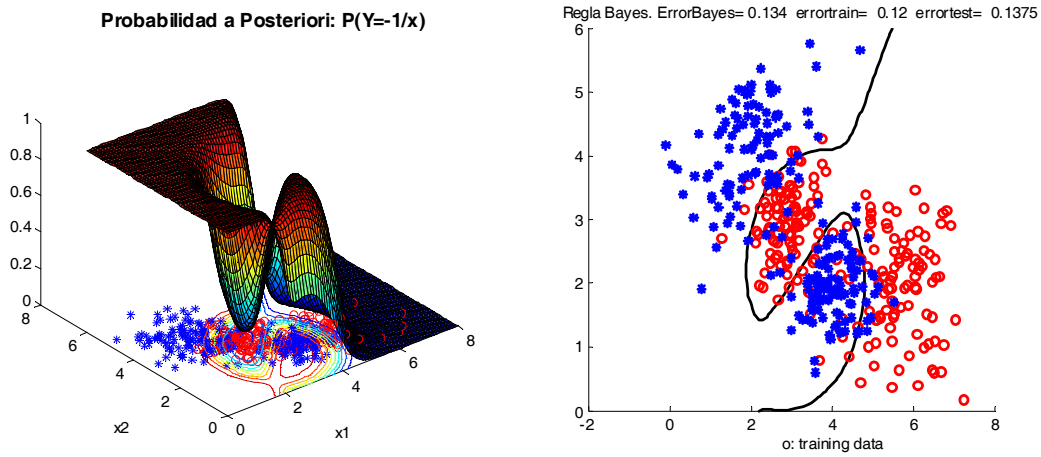


Figura 35. Izquierda: superficie $\mathbb{P}(\mathsf{Y} = -1/\mathbf{x})$. Derecha: frontera de Bayes $\mathbb{P}(\mathsf{Y} = -1/\mathbf{x}) = 1/2$.

Para este conjunto de muestras, hemos estimado la expresión (II.158), mediante sus correspondientes valores empíricos en una muestra de test de tamaño $n_t = 400$, de la siguiente forma:

$$\widehat{\mathrm{ME}}(\hat{g}_n) = \widehat{\mathbb{E}}_{\mathsf{X}}[\mathrm{ME}(\hat{g}_n(\mathsf{X}))] = \frac{1}{n_t}\sum_{i=1}^{n_t}\mathrm{ME}[\hat{g}_n(\mathbf{x}_i)] \tag{II.160}$$

$$\text{donde} \quad : \quad \mathrm{ME}[\hat{g}_n(\mathbf{x}_i)] = \mathbb{E}_{\mathsf{Z}^n}[1_{\{\hat{g}_n(\mathbf{x}_i)\neq \mathbf{g}(\mathbf{x}_i)\}}] = \frac{1}{N}\sum_{j=1}^{N} 1_{\{\hat{g}_{n,j}(\mathbf{x}_i)\neq \mathbf{g}(\mathbf{x}_i)\}}$$

donde $\mathbb{E}_{\mathsf{Z}^n}$ denota la esperanza en las posibles muestras aleatorias de $\mathcal{D}$ y $\hat{g}_{n,j}(\mathbf{x}_i)$ es la estimación de la clase del punto de test $\mathbf{x}_i$, obtenida utilizando la j-ésima muestra $(\mathbf{z}^n)_j$ de tamaño $n = 100$.

De la misma forma, el cálculo de los términos de sesgo y varianza para cada $\mathbf{x}_i$ se obtienen mediante:

$$\mathrm{Sesgo}_{\mathsf{Z}^n}[\hat{g}_n(\mathbf{x}_i)] = \ell(\mathbf{g}(\mathbf{x}_i), \hat{g}^*_{\mathcal{D}}(\mathbf{x}_i)) = 1_{\{\hat{g}^*_{\mathcal{D}}(\mathbf{x}_i)\neq \mathbf{g}(\mathbf{x}_i)\}} = 1_{\{\mathrm{signo}[\frac{1}{N}\sum_{j=1}^{N}\hat{g}_{n,j}(\mathbf{x}_i)]\neq \mathbf{g}(\mathbf{x}_i)\}}$$

$$\mathrm{Var}_{\mathsf{Z}^n}[\hat{g}_n(\mathbf{x}_i)] = \mathbb{E}_{\mathsf{Z}^n}[\ell(\hat{g}^*_{\mathcal{D}}(\mathbf{x}_i), \hat{g}_n(\mathbf{x}_i)] = \frac{1}{N}\sum_{j=1}^{N} 1_{\{\hat{g}_{n,j}(\mathbf{x}_i)\neq \mathrm{signo}[\frac{1}{N}\sum_{j=1}^{N}\hat{g}_{n,j}(\mathbf{x}_i)]\}}$$

donde hemos utilizado que la **predicción principal** $\hat{g}^*_{\mathcal{D}}(\mathbf{x})$ es la moda de las predicciones $\hat{g}_{n.j}(\mathbf{x}_i) \in \{-1, 1\}$, es decir:

$$\hat{g}^*_{\mathcal{D}}(\mathbf{x}) = \mathrm{signo}[\frac{1}{N}\sum_{j=1}^{N}\hat{g}_{n,j}(\mathbf{x}_i)]$$

es el voto mayoritario de las predicciones $\hat{g}_{n,j}(\mathbf{x}_i)$. Con ello, la esperanza del error en $\mathbf{x}_i$ es la suma de los términos de sesgo y varianza según (II.157):

$$\mathrm{ME}[\hat{g}_n(\mathbf{x}_i)] = c_2(\mathbf{x}_i)\,\mathrm{Var}_{\mathsf{Z}^n}[\hat{g}_n(\mathbf{x}_i)] + \mathrm{Sesgo}_{\mathsf{Z}^n}[\hat{g}_n(\mathbf{x}_i)]$$

y la media sobre la muestra de test, da lugar a la estimación del $\widehat{\mathrm{MSE}}(\hat{g}_n)$ según (II.160):

$$\widehat{\mathrm{ME}}(\hat{g}_n) = \widehat{\mathbb{E}}_{\mathsf{X}}[\mathrm{ME}(\hat{g}_n(\mathsf{X}))] = \frac{1}{n_t}\sum_{i=1}^{n_t}[c_2(\mathbf{x}_i)\,\mathrm{Var}_{\mathsf{Z}^n}[\hat{g}_n(\mathbf{x}_i)] + \mathrm{Sesgo}_{\mathsf{Z}^n}[\hat{g}_n(\mathbf{x}_i)] \tag{II.161}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En dichas expresiones el coeficiente: $c_2(\mathbf{x}) = 1$ si $\hat{g}^*_{\mathcal{D}}(\mathbf{x}) = \mathsf{g}(\mathbf{x})$ y -1 en otro caso, actúa como indicador de los ejemplos donde existe sesgo, en los cuáles la varianza reduce la esperanza del error.

Finalmente, a partir de $\widehat{\mathrm{ME}}(\hat{g}_n)$ puede calcularse la estimación $\widehat{\mathrm{MPE}}(\hat{g}_n)$ del error de predicción (esperanza del riesgo) estimando previamente el término del ruido, mediante:

$$\hat{\sigma}^2(\mathbf{x}_i) = \widehat{\mathbb{E}}_{\mathsf{Y}/\mathbf{x}_i}[\ell(\mathsf{Y}, \mathsf{g}(\mathbf{x}_i))] = \ell(1, \mathsf{g}(\mathbf{x}_i)) \cdot \mathbb{P}(\mathsf{Y} = 1|\mathbf{x}_i) + \ell(-1, \mathsf{g}(\mathbf{x}_i)) \cdot \mathbb{P}(\mathsf{Y} = -1|\mathbf{x}_i)$$

$$c_1(\mathbf{x}_i) = 2\mathbb{P}_{\mathsf{Z}^n}[\hat{g}_n(\mathbf{x}_i) = \mathsf{g}(\mathbf{x}_i)] - 1 = 2\frac{1}{N}\sum_{j=1}^{N} 1_{\{\hat{g}_{n,j}(\mathbf{x}_i) = \mathsf{g}(\mathbf{x}_i)\}} - 1$$

$$\widehat{\mathbb{E}}_{\mathsf{X}}[c_1(\mathsf{X})\hat{\sigma}^2(\mathsf{X})] = \frac{1}{n_t}\sum_{i=1}^{n_t} c_1(\mathbf{x}_i)\hat{\sigma}^2(\mathbf{x}_i)$$

El término de ruido para la regla de Bayes ($c_1(\mathbf{x}) = 1$, $\forall \mathbf{x}$) es la estimación del error de Bayes en la muestra de test que, en nuestro ejemplo ($n_t = 400$), tuvo un valor de $0{,}1375$. Una vez estimado el término de ruido, se obtendría una estimación del MPE, mediante:

$$\widehat{\mathrm{MPE}}(\hat{g}_n) = \widehat{\mathbb{E}}_{\mathsf{X}}[c_1(\mathsf{X})\hat{\sigma}^2(\mathsf{X})] + \widehat{\mathrm{ME}}(\hat{g}_n) \tag{II.162}$$

Una vía alternativa para la estimación del término de ruido, es aprovechar la estimación de la esperanza del MPE y estimar el término del ruido como la diferencia:

$$\widehat{\mathbb{E}}_{\mathsf{X}}[c_1(\mathsf{X})\hat{\sigma}^2(\mathsf{X})] = \widehat{\mathrm{MPE}}(\hat{g}_n) - \widehat{\mathrm{ME}}(\hat{g}_n)$$

donde $\widehat{\mathrm{ME}}(\hat{g}_n)$ se obtiene según (II.161).

AdaBoost con RBF-OLS

La Tabla 23 muestra las estimaciones de la esperanza del riesgo y su descomposición sesgo-varianza, correspondientes a las hipótesis producidas por AdaBoost (Gradient_Boost con pérdida exponencial) aplicado a diferentes redes RBF-OLS con un mismo radio en todas las funciones básicas, seleccionando los centros de entre los puntos de la muestra.

Dicha tabla incluye, como comparación, los resultados obtenidos por cada estimador individual (H. Base), así como de dos redes RBF-OLS con radios con radios $\sigma = 2$ y 5, que utilizan el máximo número de funciones básicas permitido por la precisión del ordenador basado en el análisis espectral de la matriz de diseño (pág. 177). (en concreto, de AdaBoost)

La figura 36 muestra la evolución del riesgo esperado, y de su descomposición sesgo-varianza, con AdaBoost aplicado a las redes RBF-OLS con $h = 1$ y 8 función básicas, para valores $\sigma = 0.3$, 2.1 y 5.

Sobre los resultados anteriores, puede concluirse lo siguiente en relación con el problema estudiado [Matías 2003]:

1. AdaBoost mejora siempre, sensiblemente, los resultados de las hipótesis base individuales. Sin embargo, la mejora con respecto a la mejor red RBF-OLS (última fila de la tabla 23), es pequeña.

2. Dicha mejora es debida, en todos los casos, a una disminución radical del sesgo, lo que concuerda con los resultados de [Bühlmann y Yu 2001] para el caso de LS_Boost, presentados

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

AdaBoost-RBF	σ	**NB**	**Tópt**	$\widehat{\mathbb{E}}$(**Sesgo**)	$\widehat{\mathbb{E}}(c_2$**Var**$)$	$\widehat{\textbf{ME}}$	$\widehat{\mathbb{E}}(c_1$**Ru.**$)$	$\widehat{\textbf{MPE}}$
H. Base	0,3	1		0,2950	0,0609	0,3559	0,0335	0,3894
			118	0,0150	0,1212	0,1362	0,0707	0,2069
H. Base		3		0,1200	0,1509	0,2709	0,0460	0,3169
			57	0,0150	0,1218	0,1368	0,0653	0,2020
H. Base		5		0,0725	0,1471	0,2196	0,0575	0,2771
			24	0,0225	0,1168	0,1393	0,0635	0,2027
H. Base		8		0,0325	0,1435	0,1760	0,0681	0,2441
			7	0,0100	0,1348	0,1448	0,0697	0,2145
H. Base	2	1		0,3975	−0,0056	0,3919	0,0043	0,3962
			+876	0,1825	0,1117	0,2942	0,0176	0,3118
H. Base		3		0,3225	−0,0198	0,3027	0,0257	0,3283
			38	**0,0250**	**0,0783**	**0,1033**	**0,0736**	**0,1770**
H. Base		5		0,1175	0,0388	0,1563	0,0612	0,2176
			16	**0,0200**	**0,0815**	**0,1015**	**0,0753**	**0,1768**
H. Base		8		0,0825	0,0281	0,1106	0,0657	0,1763
			8	**0,0350**	**0,0604**	**0,0954**	**0,0772**	**0,1725**
H. Base	5	1		0,4875	−0,0011	0,4764	−0,0178	0,4586
			+997	0,1475	0,1004	0,2479	0,0334	0,2813
H. Base		3		0,3875	−0,0023	0,3852	0,0106	0,3958
			200	0,0950	0,1030	0,1980	0,0452	0,2432
H. Base		5		0,3325	−0,0050	0,3275	0,0225	0,3501
			152	0,0325	0,0915	0,1240	0,0698	0,1939
H. Base		8		0,1750	0,0215	0,1965	0,0448	0,2412
			9	0,0275	0,0854	0,1129	0,0708	0,1837
H. Base	10	(9)		0,1050	0,0492	0,1542	0,0512	0,2055
			20	0,0150	0,0966	0,1116	0,0692	0,1808
RBF	2	(30)		0,0125	0,1064	0,1189	0,0794	0,1982
RBF	**5**	(**15**)		**0,0650**	**0,0444**	**0,1094**	**0,0679**	**0,1774**

Tabla 23. Descomposición sesgo-varianza de AdaBoost aplicado a distinto número de funciones básicas radiales y comparación con las hipótesis base. Sesgo-varianza de AdaBoost con redes RBF-OLS. Se muestra: σ, el número de funciones básicas (NB), la iteración óptima (Topt), y las componentes del riesgo. El número de iteraciones marcado con un signo más indica que el algoritmo sigue mejorando lentamente. El número de básicas entre paréntesis indica que es el número máximo de funciones básicas admitidas para el radio correspondiente.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

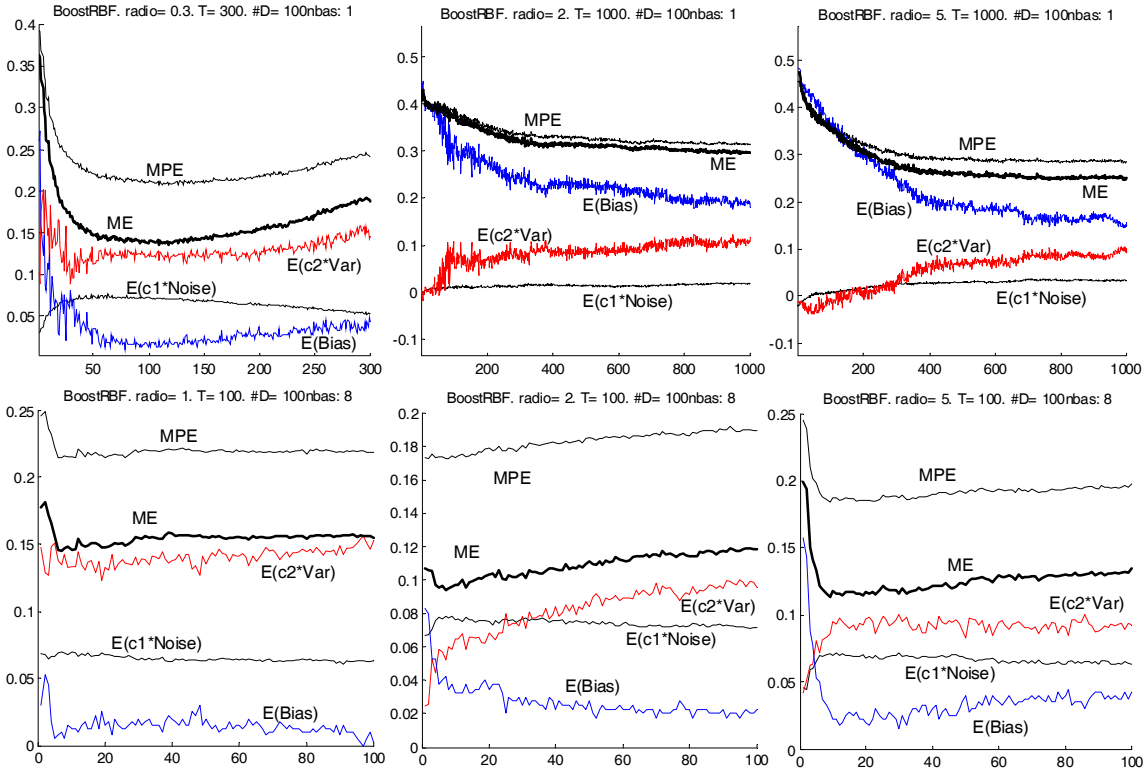


Figura 36. Evolución de la descomposición sesgo-varianza de AdaBoost aplicado a una función basica radial gausiana (arriba) y a 8 funciones básicas (abajo), para $\sigma = 0.3$, 2 y 5, cuyas magnitudes se muestran en la tabla 23.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

anteriormente. Dicha disminución va acompañada de un incremento de varianza si el nivel de varianza del estimador base lo admite. Pero, si éste posee ya alta varianza, AdaBoost la reduce también aunque más ligeramente que el sesgo (tabla 23: casos con $\sigma = 0.3$ y número de básicas igual a 3, 5 y 8).

En todo caso, si la varianza es pequeña, AdaBoost suele encontrar espacio para alguna mejora con respecto a la hipótesis base, siempre que, a su vez, el sesgo no sea excesivamente pequeño. Así, **la mejor combinación de circunstancias parece ser una varianza pequeña y un sesgo intermedio-bajo**.

3. Por tanto, para un mismo valor del parámetro de escala, salvo para valores muy pequeños, es preferible un número mayor de funciones básicas, pues se acelera la iteración óptima sin producir un sobreajuste repentino posterior.

4. El número óptimo de iteraciones de AdaBoost es más pequeño cuanto mayor es la complejidad de la hipótesis. Si dicha complejidad se debe a un radio pequeño, el sobreajuste se produce más rápidamente (figura 36 $h = 1$, $\sigma = 0.3$) que si la complejidad se debe al número de funciones básicas (misma figura $h = 8$, $\sigma = 2$ y $\sigma = 5$).

   Por tanto, para un mismo valor del parámetro de escala, salvo para valores pequeños, es preferible un número mayor de funciones básicas pues se acelera la iteración óptima sin producir un sobreajuste repentino posterior.

   El máximo número de funciones básicas para el que AdaBoost puede producir beneficios es menor que el correspondiente a la red RBF-OLS óptima.

5. Con una o dos funciones básicas, en ningún caso AdaBoost supera la mejor red RBF-OLS.

Por tanto, parece que AdaBoost con RBF-OLS produce sus mejores resultados con un nivel medio-bajo de complejidad (relativo a la del problema) en la hipótesis base. Una varianza pequeña le permite a AdaBoost encontrar recursos para una reducción del sesgo cuando éste no es excesivamente bajo; y, si éste no es excesivamente alto, se tienen los mejores resultados.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Capítulo III
Relaciones con Otras Técnicas Estadísticas, Nuevos Algoritmos y Aplicaciones

Este capítulo tiene por objeto:

1. Analizar las relaciones existentes entre diversos algoritmos presentados en el capítulo anterior y otras técnicas estadísticas. En concreto, expondremos las relaciones existentes entre las redes neuronales RBF de regularización, las redes SVM, el kriging y la regresión mediante procesos gausianos.

2. Proponer nuevos algoritmos basados en redes neuronales. En concreto, presentaremos un nuevo algoritmo kriging resultado de la incorporación de regularización, un algoritmo *projection pursuit* basado en suavizadores RBF en cada proyección, y una batería de nuevos algoritmos para la predicción de series de tiempo financieras.

3. Aplicar y comparar las técnicas anteriores y otras tratadas en el capítulo anterior, en tres grupos de problemas reales: la estimación de la explotabilidad de un yacimiento de pizarra, el diseño de un plan de reforestación y la predicción de precios y rentabilidades en finanzas.

En las secciones siguientes, por conveniencia, mantenemos la notación de los capítulos anteriores sólo en los conceptos comunes a todo el documento que sean utilizados en cada una. Los conceptos específicos, seguirán su notación propia detallada en cada momento.

III.1 Regresión Bayesiana, Procesos Gausianos y Redes Neuronales

III.1.1 Regresión Bayesiana y Procesos Gausianos

Sea el modelo de regresión lineal generalizada:

$$\mathsf{Y}_i = \mathsf{Y}(\mathbf{x}_i) = \mathsf{f}(\mathbf{x}_i) + \varepsilon_i = \sum_{i=1}^{r} \mu_j g_j(\mathbf{x}_i) + \varepsilon_i,\ i = 1, .., n \tag{III.1}$$

donde $g_j : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$, $j = 1 : r$ son funciones conocidas cuyos valores actúan a modo de covariables, $\boldsymbol{\mu} = (\mu_1, ..., \mu_r)^T \in \mathbb{R}^r$ es un vector de coeficientes desconocidos, y ε_i es ruido gausiano de media cero y $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \boldsymbol{\Sigma}_{ij}$, $i, j = 1 : n$, para alguna matriz $\boldsymbol{\Sigma}$.

En el marco bayesiano, supongamos que se dispone de información a priori sobre los parámetros, que puede codificarse mediante una distribución a priori gausiana $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}})$ tal que, $\boldsymbol{\mu}$ es independiente de $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$. Con ello, el modelo anterior puede escribirse como:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$$\mathsf{Y}^n / \boldsymbol{\mu} \sim \mathcal{N}(\mathbf{G}^T \boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ con } \boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}_r, \boldsymbol{\Sigma}_{\boldsymbol{\mu}})$$

donde $\mathsf{Y}^n = (\mathsf{Y}_1, ..., \mathsf{Y}_n)^T$ con valores $\mathbf{y} = (y_1, ..., y_n)^T$, $\mathbf{G}$ es una matriz $n \times r$ con $(\mathbf{G})_{ij} = g_j(\mathbf{x}_i)$, $i = 1 : n$, $j = 1 : r$, y $\mathbf{\Sigma} = \mathrm{Cov}(\boldsymbol{\varepsilon})$. En lo que sigue, denotamos f mediante F reflejando su aleatoriedad implícita a través de $\boldsymbol{\mu}$.

El enfoque bayesiano determina la distribución a posteriori:

$$\mathsf{p}(\boldsymbol{\mu}|\mathbf{y}) = \frac{\mathsf{p}(\mathbf{y}|\boldsymbol{\mu})\mathsf{p}(\boldsymbol{\mu})}{\mathsf{p}(\mathbf{y})}$$

donde $\mathsf{p}(\mathbf{y}|\boldsymbol{\mu})$ es la verosimilitud $\mathcal{N}(\mathbf{G}^T\boldsymbol{\mu}, \mathbf{\Sigma})$, $\mathsf{p}(\boldsymbol{\mu})$ es la densidad $\mathcal{N}(\mathbf{0}_r, \mathbf{\Sigma}_{\boldsymbol{\mu}})$, y

$$\mathsf{p}(\mathbf{y}) = \int \mathsf{p}(\mathbf{y}|\boldsymbol{\mu})\mathsf{p}(\boldsymbol{\mu})d\boldsymbol{\mu}$$

es un factor normalizador. Como el numerador es un producto de gausianas, la distribución a posteriori también lo es y su esperanza minimiza:

$$\begin{aligned} L(\boldsymbol{\mu}) &= (\mathbf{y} - \mathbf{G}\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{G}\boldsymbol{\mu}) + \boldsymbol{\mu}^T\mathbf{\Sigma}_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu} \\ &= 2\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} - 2\boldsymbol{\mu}^T\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{y} + 2\boldsymbol{\mu}^T\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G}\boldsymbol{\mu} + \boldsymbol{\mu}^T\mathbf{\Sigma}_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu} \end{aligned}$$

que, así, viene dada por:

$$\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G}\boldsymbol{\mu} + \mathbf{\Sigma}_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu} = \mathbf{0}_r \Rightarrow \boldsymbol{\mu}_{MP} = \mathbb{E}_{\boldsymbol{\mu}/\mathbf{y}}(\boldsymbol{\mu}) = (\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G} + \mathbf{\Sigma}_{\boldsymbol{\mu}}^{-1})^{-1}\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{y} \tag{III.2}$$

Igualmente, puede demostrarse (p. ej. [Pratt *et al.* 1995, sec. 24.3]), que la matriz de varianzas-covarianzas es $\mathrm{Var}(\boldsymbol{\mu}/\mathbf{y}) = (\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G} + \mathbf{\Sigma}_{\boldsymbol{\mu}}^{-1})^{-1}$. De esta forma, la estimación bayesiana de $\mathsf{F}_0 \doteq \mathsf{F}(\mathbf{x}_0)$ y su varianza, son:

$$\mathbb{E}_{\mathsf{F}_0/\mathbf{y}}[\mathsf{F}_0] = \mathbb{E}_{\boldsymbol{\mu}/\mathbf{y}}(\mathbf{g}_0^T\boldsymbol{\mu}) = \mathbf{g}_0^T\boldsymbol{\mu}_{MP} = \mathbf{g}_0^T(\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G} + \mathbf{\Sigma}_{\boldsymbol{\mu}}^{-1})^{-1}\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{y} \tag{III.3}$$

$$\begin{aligned} \mathrm{Var}_{\mathsf{F}_0/\mathbf{y}}[\mathsf{F}_0] &= \mathbb{E}_{\boldsymbol{\mu}/\mathbf{y}}\{[\mathsf{F}(\mathbf{x}_0) - \mathbb{E}_{\mathsf{F}/\boldsymbol{\mu}}(\mathsf{F}(\mathbf{x}_0))]^2\} = \mathbb{E}_{\boldsymbol{\mu}/\mathbf{y}}\{(\mathbf{g}_0^T\boldsymbol{\mu} - \mathbf{g}_0^T\boldsymbol{\mu}_{MP})^2\} \\ &= \mathbf{g}_0^T\mathbb{E}_{\boldsymbol{\mu}/\mathbf{y}}[(\boldsymbol{\mu} - \boldsymbol{\mu}_{MP})(\boldsymbol{\mu} - \boldsymbol{\mu}_{MP})^T]\mathbf{g}_0 \\ &= \mathbf{g}_0^T(\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G} + \mathbf{\Sigma}_{\boldsymbol{\mu}}^{-1})^{-1}\mathbf{g}_0 \end{aligned} \tag{III.4}$$

donde $\mathbf{g}_0 = \mathbf{g}(\mathbf{x}_0) = (g_1(\mathbf{x}_0), ..., g_r(\mathbf{x}_0))^T$. Para obtener la estimación bayesiana de $\mathsf{Y}(\mathbf{x}_0)$ basta añadir $\mathrm{Var}(\varepsilon_{\mathbf{x}_0})$ a la expresión de la varianza pues ambas fuentes de aleatoriedad ($\boldsymbol{\mu}$ y $\varepsilon_{\mathbf{x}_0}$) se suponen independientes.

Del Espacio de Parámetros al Espacio de Funciones: Procesos Gausianos

La incertidumbre inicial que hemos supuesto en el espacio de los parámetros $\boldsymbol{\mu}$ a través de su distribución a priori gausiana, puede trasladarse a un espacio de funciones $\mathcal{H} = \{f_{\boldsymbol{\mu}}(\mathbf{x}) = \sum_{j=1}^r \mu_j g_j(\mathbf{x}) : \boldsymbol{\mu} \in \mathbb{R}^r\}$ cuyos elementos $f_{\boldsymbol{\mu}}$ pueden verse como realizaciones de una función aleatoria gausiana $\mathsf{F} = \sum_{j=1}^r \mu_j g_j$ indexada en $\mathbf{x} \in \mathcal{X}$. Además, al ser $\mathbb{E}_{\boldsymbol{\mu}}(\boldsymbol{\mu}) = \mathbf{0}_r$, se verifica que $\mathbb{E}[\mathsf{F}(\mathbf{x})] = 0$ y que:

$$k(\mathbf{x}, \mathbf{x}') \doteq \mathrm{Cov}[\mathsf{F}(\mathbf{x}), \mathsf{F}(\mathbf{x}')] = \mathbb{E}[\mathsf{F}(\mathbf{x})\mathsf{F}(\mathbf{x}')] = \mathbb{E}_{\boldsymbol{\mu}}[\mathbf{g}(\mathbf{x})^T\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{g}(\mathbf{x})] = \mathbf{g}(\mathbf{x})^T\mathbf{\Sigma}_{\boldsymbol{\mu}}\mathbf{g}(\mathbf{x})$$

donde k, así definida, es una función definida positiva.

Contando con un proceso gausiano, podemos aprovechar sus facilidades para la predicción. Si disponemos de n observaciones $\mathsf{F}^n = (\mathsf{F}_1, ..., \mathsf{F}_n)^T$ con $\mathsf{F}_i = \mathsf{F}(\mathbf{x}_i)$, $i = 1 : n$, la predicción $\hat{\mathsf{F}}(\mathbf{x}_0)$ de F en un punto de predicción $\mathbf{x}_0$ se basa en que la distribución conjunta de $\mathsf{F}_+ = (\mathsf{F}^n, \mathsf{F}_0)$

es gausiana, con lo que basta determinar la distribución condicionada de $\mathsf{F}_0/\mathsf{F}^n$:

$$\mathsf{p}(\mathsf{F}_0|\mathsf{F}^n) = \frac{\mathsf{p}(\mathsf{F}_0, \mathsf{F}^n)}{\mathsf{p}(\mathsf{F}^n)}$$

Si la matriz de covarianza de $(\mathsf{F}_0, \mathsf{F}^n)$ es:

$$\mathbf{K}_+ = \begin{bmatrix} \mathbf{K} & \mathbf{k}_0 \\ \mathbf{k}_0^T & k_0 \end{bmatrix}$$

donde $\mathbf{K} \doteq \mathrm{Var}(\mathsf{F}^n) = \mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T$, $\mathbf{k}_0 \doteq (k(\mathbf{x}_0, \mathbf{x}_1), ..., k(\mathbf{x}_0, \mathbf{x}_n))^T = \mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{g}_0$ y $k_0 \doteq k(\mathbf{x}_0, \mathbf{x}_0) = \mathbf{g}_0^T\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{g}_0$, siendo $\mathbf{g}_0 = \mathbf{g}(\mathbf{x}_0)$, entonces, por las fórmulas de los momentos de las distribuciones condicionadas de una distribución gausiana (ecuaciones (G.10) y (G.11), Anexo G), se verifica que:

$$\begin{aligned} \mathbb{E}_{\mathsf{F}_0/\mathbf{f}}(\mathsf{F}_0) &= \mathbf{k}_0^T\mathbf{K}^{-1}\mathbf{f} \\ \mathrm{Var}_{\mathsf{F}_0/\mathbf{f}}(\mathsf{F}_0) &= k_0 - \mathbf{k}_0^T\mathbf{K}^{-1}\mathbf{k}_0 \end{aligned}$$

con lo que la esperanza condicionada es un estimador lineal de las observaciones.

Sin embargo, la variable F^n no es observable al estar sujeta al ruido $\boldsymbol{\varepsilon}$, y en su lugar, se dispone de n observaciones $\mathsf{Y}^n = (\mathsf{Y}_1, ..., \mathsf{Y}_n)^T$. Siguiendo el mecanismo anterior, teniendo en cuenta que $\mathsf{F}_+ \sim \mathcal{N}(\mathbf{0}_{n+1}, \mathbf{K}_+)$ y que $\mathsf{Y}_i = \mathsf{F}_i + \varepsilon_i$, $i = 1 : n$, se tiene que [Williams 1998]: $(\mathsf{Y}^n, \mathsf{F}_0) \sim \mathcal{N}(\mathbf{0}_{n+1}, \mathbf{K}_+ + \boldsymbol{\Sigma}_+)$ donde $\boldsymbol{\Sigma}_+$ es la matriz $\boldsymbol{\Sigma}$ con una fila y columna adicional de ceros al ser $\boldsymbol{\varepsilon}$ y F_0 incorreladas. Entonces, siguiendo el mismo procedimiento que antes al ser el proceso, gausiano, resulta:

$$\mathbb{E}_{\mathsf{F}_0/\mathbf{y}}(\mathsf{F}_0) = \mathbf{k}_0^T(\mathbf{K} + \boldsymbol{\Sigma})^{-1}\mathbf{y} = \mathbf{g}_0^T\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T\mathbf{A}^{-1}\mathbf{y} \quad \text{(III.5)}$$

$$\mathrm{Var}_{\mathsf{F}_0/\mathbf{y}}(\mathsf{F}_0) = k_0 - \mathbf{k}_0^T(\mathbf{K} + \boldsymbol{\Sigma})^{-1}\mathbf{k}_0 = \mathbf{g}_0^T\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{g}_0 - \mathbf{g}_0^T\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T\mathbf{A}^{-1}\mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{g}_0 \quad \text{(III.6)}$$

donde $\mathbf{A} = (\mathbf{K} + \boldsymbol{\Sigma}) = (\mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T + \boldsymbol{\Sigma})$.

Para obtener la equivalencia entre ambos resultados y las ecuaciones (III.3) y (III.4) basta utilizar que [Williams 1998], siendo $\mathbf{H} = (\mathbf{G}^T\boldsymbol{\Sigma}^{-1}\mathbf{G} + \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1})$:

$$\mathbf{H}\Sigma_{\boldsymbol{\mu}}\mathbf{G}^T = \mathbf{G}^T\boldsymbol{\Sigma}^{-1}\mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T + \mathbf{G}^T = \mathbf{G}^T\boldsymbol{\Sigma}^{-1}(\mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T + \boldsymbol{\Sigma}) = \mathbf{G}^T\boldsymbol{\Sigma}^{-1}\mathbf{A}$$

y multiplicar por la izquierda por $\mathbf{H}^{-1}$ y por la derecha por $\mathbf{A}^{-1}$ para obtener $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T\mathbf{A}^{-1} = \mathbf{H}^{-1}\mathbf{G}^T\boldsymbol{\Sigma}^{-1}$. Si esta igualdad se sustituye en (III.3) se obtiene el resultado (III.5).

Igualmente, utilizando la fórmula de la inversa de un binomio de matrices (ecuación (G.1), pág. 412), con la expresión $\mathbf{H}^{-1} = (\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} + \mathbf{G}^T\boldsymbol{\Sigma}^{-1}\mathbf{G})^{-1}$, se obtiene:

$$\begin{aligned} \mathbf{H}^{-1} &= \boldsymbol{\Sigma}_{\boldsymbol{\mu}} - \boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T(\mathbf{I}_n + \boldsymbol{\Sigma}^{-1}\mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}} \\ &= \boldsymbol{\Sigma}_{\boldsymbol{\mu}} - \boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T[\boldsymbol{\Sigma}(\mathbf{I}_n + \boldsymbol{\Sigma}^{-1}\mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T)]^{-1}\mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}} \\ &= \boldsymbol{\Sigma}_{\boldsymbol{\mu}} - \boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T(\boldsymbol{\Sigma} + \mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T)^{-1}\mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{\mu}} \end{aligned}$$

que sustituyéndolo en (III.4) produce el resultado (III.6).

Esta equivalencia entre la estimación bayesiana y la proporcionada mediante procesos gausianos fué demostrada [Williams 1998] para el caso de ruido independiente con $\mathrm{Var}(\boldsymbol{\varepsilon}) = \frac{1}{\beta}\mathbf{I}_n$ y aquí se ha generalizado al caso $\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

La comparación entre las ecuaciones (III.3) y (III.5) de la solución de la regresión bayesiana y de la correspondiente a la regresión mediante procesos gausianos, pone de manifiesto una segunda dualidad de las ecuaciones de la regresión, diferente a la tratada por ejemplo cuando analizamos la red RBF como suavizador lineal, y ayuda a clarificar la relación entre ambos enfoques.

En concreto, la primera ecuación, aplicada a los puntos de la muestra $\mathbf{g}_i = (g_1(\mathbf{x}_i), ..., g_r(\mathbf{x}))^T$, $i = 1 : n$, (ahora las covariables son g_j, $j = 1 : r$), coincide con la forma general de la solución de una regresión por mínimos cuadrados regularizada (*ridge*) y ponderada por $\mathbf{\Sigma}$:

$$\mathbb{E}_{\mathsf{F}^n/\mathbf{y}}(\mathsf{F}^n) = (\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G} + \mathbf{\Sigma}_{\boldsymbol{\mu}}^{-1})^{-1}\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{y} \tag{III.7}$$

Por su parte, la segunda ecuación, aporta la misma solución, pero expresada en términos de la matriz Gram (pág. 419, Anexo[87] G), formada por los productos internos cruzados de los vectores de datos. En los puntos de la muestra, esta segunda ecuación (III.5), se escribe como:

$$\mathbb{E}_{\mathsf{F}^n/\mathbf{y}}(\mathsf{F}^n) = \mathbf{G}\mathbf{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T(\mathbf{G}\mathbf{\Sigma}_{\boldsymbol{\mu}}\mathbf{G}^T + \mathbf{\Sigma})^{-1}\mathbf{y} \tag{III.8}$$

y, ambas ecuaciones, como es lógico, producen la misma solución.

Del Espacio de Funciones al Espacio de Parámetros: Regresión Bayesiana

En lo anterior hemos visto como la distribución a priori de los parámetros determina una función de covarianza en el proceso gausiano correspondiente. Sin embargo, es posible comenzar codificando la información a priori mediante una función de covarianza de una función aleatoria gausiana:

$$\mathbb{E}[\mathsf{F}(\mathbf{x})\mathsf{F}(\mathbf{x}')] = k(\mathbf{x}, \mathbf{x}')$$

incluso, sin necesidad de plantearse cuál puede ser el espacio de funciones correspondiente ni su distribución a priori.

De cualquier forma, la elección de la función de covarianza permite determinar un espacio de funciones y una distribución a priori de la siguiente forma. Al ser la función k definida positiva, el teorema de Mercer (teorema 112, pág. 344) permite la siguiente representación:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x})\phi_i(\mathbf{x}')$$

donde $\{\lambda_i\}$ son los autovalores positivos del correspondiente operador de Hilbert-Schmidt (Anexo B) y $\{\phi_i\}$ son sus autofunciones ortonormales que son base de $L_2(\mathcal{X})$ con $\mathcal{X} \subset \mathbb{R}^d$ compacto.

Si ahora, consideramos el problema de regresión generalizada (III.1) utilizando como funciones g_j las autofunciones ϕ_j, es decir consideramos un espacio funcional $\mathcal{H}$ de posibles funciones de regresión de la forma $f(\mathbf{x}) = \sum_{j=1}^{r} \mu_j \phi_j(\mathbf{x})$, para cualquier $r \in \mathbb{N}^+ \cup \{\infty\}$ (entendido como límite de sumas finitas), donde los parámetros $\{\mu_j\}_{j=1}^r$ son tales que las correspondientes funciones están bien definidas en $L_2(\mathcal{X})$, y, asimismo, consideramos una distribución a priori para el vector de parámetros $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}_r, \mathbf{\Sigma}_{\boldsymbol{\mu}})$ tal que, $\mathbf{\Sigma}_{\boldsymbol{\mu}}$ es una matriz diagonal con elementos diagonales $(\mathbf{\Sigma}_{\boldsymbol{\mu}})_{ii} = \mathrm{Var}(\mu_i) = \mathbb{E}(\mu_i^2) = \lambda_i$, $i = 1 : r$, entonces se obtiene un problema de

[87] Aunque en dicho anexo la solución se deriva para el caso $\mathbf{\Sigma}_{\boldsymbol{\mu}} = \lambda \mathbf{I}_r$.

regresión bayesiana, con función de covarianza asociada:

$$\begin{aligned}\mathbb{E}[\mathsf{F}(\mathbf{x})\mathsf{F}(\mathbf{x}')] &= \mathbb{E}_{\boldsymbol{\mu}}[\sum_{i=1}^{r}\mu_i\phi_i(\mathbf{x})\sum_{j=1}^{r}\mu_j\phi_j(\mathbf{x}')] = \sum_{i,j=1}^{r}\phi_i(\mathbf{x})\phi_j(\mathbf{x}')\mathbb{E}_{\boldsymbol{\mu}}(\mu_i\mu_j)\\ &= \sum_{i=1}^{r}\lambda_i\phi_i(\mathbf{x})\phi_i(\mathbf{x}') = k(\mathbf{x},\mathbf{x}')\end{aligned}$$

Con ello, la regresión mediante procesos gausianos con una función de covarianza elegida a priori, determina una regresión bayesiana en el espacio de los parámetros cuyas funciones básicas (en número posiblemente infinito) son las autofunciones de la función de covarianza.

III.1.2 Regresión Mediante Procesos Gausianos y Redes Neuronales RBF y RRBF

Cuando comparamos las soluciones de la regresión bayesiana y la regresión mediante procesos gausianos (ecuaciones (III.7) y (III.8)), dedujimos que, ambos métodos, producen una solución equivalente a la del problema de regresión regularizado.

En dichas ecuaciones, pueden reconocerse dos tipos de redes neuronales RBF tratadas en el capítulo anterior:

1. En el caso de ruido independiente ($\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$) y matriz de covarianzas de los parámetros (regularizador) $\boldsymbol{\Sigma}_{\boldsymbol{\mu}} = \lambda\mathbf{I}_r$, la primera ecuación mencionada toma la forma:

$$\mathbf{f} = \mathbf{G}(\mathbf{G}^T\mathbf{G} + \lambda\mathbf{I}_n)^{-1}\mathbf{G}^T\mathbf{y}$$

   con $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^T$, la cuál, coincide con la solución de una red RBF $f(\mathbf{x}) = \sum_{j=1}^{r} g_j(\mathbf{x})$ (con constante si alguna g_j es constante), entrenada mediante mínimos cuadrados regularizados. (Nótese que, en general, no hemos puesto condiciones en el capítulo anterior para la elección de dichas funciones básicas, más que el correspondiente problema sea resoluble, aunque tradicionalmente suelen emplearse las funciones radiales heredadas del problema de interpolación).

2. Por otra parte, si en el mismo caso de ruido independiente ($\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$), se considera la ecuación original (III.5), que dió lugar a la segunda ecuación mencionada, y se utiliza su expresión en términos de la función de covarianza del proceso gausiano, se obtiene en cualquier $\mathbf{x} = \mathbf{x}_0$:

$$\begin{gathered}f(\mathbf{x}) = \mathbf{k}_0^T(\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} = \mathbf{k}^T\mathbf{c} = \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i)\\ \text{donde: } \mathbf{c} \doteq (\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}\end{gathered} \tag{III.9}$$

   ecuaciones que coinciden con la expresión funcional de una red RBF de regularización (k es definida positiva), que es solución del siguiente problema de regularización con $\lambda = \sigma^2$:

$$\min_{f\in\mathcal{H}_k}\{\|\mathbf{y} - \mathbf{f}\|_2^2 + \lambda\,\|f\|_{\mathcal{H}_k}^2$$

   donde $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^T$ y $\mathcal{H}_k$ es el RKHS con núcleo la función de covarianza k del proceso.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

III.2 Kriging y Redes Neuronales

III.2.1 Kriging

Puede decirse que el kriging es la técnica más utilizada para la predicción espacial de funciones aleatorias. Como referencia para posteriores epígrafes, presentamos a continuación una concisa descripción del estimador kriging universal.

Sea una función aleatoria ([Matheron 1973], [Journel y Huijbregts 1978], [Ripley 1981], [Cressie 1993], [Stein 1999]):

$$\mathsf{Y}(\mathbf{x}) = \sum_{j=1}^{r} \mu_j g_j(\mathbf{x}) + \mathsf{U}(\mathbf{x}) \tag{III.10}$$

donde $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_m) \in \mathbb{R}^m$ es un vector desconocido; $g_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1 : m$, y $\mathsf{U}(\mathbf{x})$ es un proceso estocástico de media cero, estacionario de segundo orden con función de covarianza:

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[\mathsf{U}(\mathbf{x}), \mathsf{U}(\mathbf{x}')] = \begin{cases} k(\mathbf{x}, \mathbf{x}) + \sigma^2 & \text{si } \mathbf{x} = \mathbf{x}' \\ k(\mathbf{x}, \mathbf{x}') & \text{si } \mathbf{x} \neq \mathbf{x}' \end{cases} \tag{III.11}$$

donde k es una función definida positiva y donde σ^2 puede ser cero. Este último término se denomina **pepita** (*nugget effect*) y suele poseer dos componentes: la variabilidad de un proceso denominado **variación en la microescala** (recogido ya en U), que opera a distancias menores que la mínima distancia entre los puntos de control $\mathbf{x} \in \mathcal{X}$, y el posible ruido en la medición del proceso (p. ej. [Cressie 1993, p. 113 y 128]).

El objetivo, es obtener un estimador $\hat{\mathsf{Y}}(\mathbf{x})$ para el proceso $\mathsf{Y}(\mathbf{x})$ a partir de n observaciones $\{\mathsf{Y}(\mathbf{x}_i) = y_i\}_{i=1}^n$ (en lo que sigue suponemos diseño fijo para $\mathbf{x}$).

Nótese que, el problema anterior, puede formularse también, considerando un proceso subyacente $\mathsf{F}(\mathbf{x})$ que se desea predecir, de la forma (III.10) con función de covarianza k, cuyas observaciones $\mathsf{Y}(\mathbf{x})$ están sujetas a ruido aleatorio independiente, de media cero y varianza[88] $\mathrm{Var}(\nu) = \sigma^2$, independiente también de $\mathsf{F}(\mathbf{x})$. Es decir, $\mathsf{Y}(\mathbf{x}) = \mathsf{F}(\mathbf{x}) + \nu$, con lo cuál, $\mathrm{Cov}[\mathsf{Y}(\mathbf{x}), \mathsf{Y}(\mathbf{x}')] = \kappa(\mathbf{x}, \mathbf{x}')$ y, en el caso en que $\sigma^2 = 0$, los procesos $\mathsf{Y}(\mathbf{x})$ y $\mathsf{F}(\mathbf{x})$ coinciden.

En el contexto de la formulación más clásica planteada inicialmente, el **estimador kriging universal** $\hat{\mathsf{Y}}(\mathbf{x})$ de Y en $\mathbf{x}_0 \in \mathcal{X}$ se define como el estimador lineal insesgado de mínima varianza, es decir:

$$\hat{\mathsf{Y}}(\mathbf{x}_0) = \arg \min_{\mathsf{T}(\mathbf{x}_0) \in \mathcal{F}} \mathbb{E}(\mathsf{T}(\mathbf{x}_0) - \mathsf{Y}(\mathbf{x}_0))^2$$

$$\text{sujeto a: } \begin{cases} \mathbb{E}[\mathsf{T}(\mathbf{x})] = \mathbb{E}[\mathsf{Y}(\mathbf{x})], \forall \mathbf{x} \in S \\ \mathsf{T}(\mathbf{x}_0) \in \mathcal{F} = \{\boldsymbol{\lambda}(\mathbf{x}_0)^T \mathsf{Y}^n : \boldsymbol{\lambda}(\mathbf{x}_0) \in \mathbb{R}^n\} \end{cases}$$

donde hemos denotado $\mathsf{Y}^n = (\mathsf{Y}_1, ..., \mathsf{Y}_n)^T$ con $\mathsf{Y}_i = \mathsf{Y}(\mathbf{x}_i)$, $i = 1 : n$.

[88] Suponemos que el proceso *variación en la microescala* está incluida en lo que hemos denominado ruido, con varianza σ^2. Alternativamente, podía suponerse que su varianza está incluída en la propia función k (la suma de funciones definidas positivas es definida positiva) dejando que σ^2 refleje sólo la parte de la pepita correspondiente al error de medición, [Cressie 1993, p. 113].

Siendo $\boldsymbol{\mu} = (\mu_1, ..., \mu_r)^T \in \mathbb{R}^r$ cualquiera desconocido con $\mu_j \neq 0$, $j = 1 : r$, la condición de insesgadez es:

$$\mathbb{E}(\hat{\mathsf{Y}}_0 - \mathsf{Y}_0) = \sum_{j=1}^{r} \mu_j \left[\sum_{i=1}^{n} \lambda_i g_j(\mathbf{x}_i) \right] - g_j(\mathbf{x}_0) = \boldsymbol{\mu}^T(\mathbf{G}^T\boldsymbol{\lambda} - \mathbf{g}_0) = 0 \tag{III.12}$$

que, se cumple, si $\mathbf{G}^T\boldsymbol{\lambda} - \mathbf{g}_0 = \mathbf{0}_r$, donde $\mathbf{G}$ es una matriz $n \times r$ con $(\mathbf{G})_{ij} = g_j(\mathbf{x}_i)$ y $\mathbf{g}_0 \in \mathbb{R}^r$ es un vector con componentes $(\mathbf{g}_0)_j = g_j(\mathbf{x}_0)$, $i = 1 : n$, $j = 1 : r$. Asimismo, teniendo en cuenta que $\mathsf{R}_0 = \hat{\mathsf{Y}}_0 - \mathsf{Y}_0 = \boldsymbol{\lambda}^T\mathsf{Y}^n - \mathsf{Y}_0$, su varianza es:

$$\begin{aligned} \mathrm{Var}(\mathsf{R}_0) &= \mathrm{Var}(\mathsf{Y}_0) + \mathrm{Var}(\boldsymbol{\lambda}^T\mathsf{Y}^n) - 2\,\mathrm{Cov}(\mathsf{Y}_0, \boldsymbol{\lambda}^T\mathsf{Y}^n) \\ &= k(\mathbf{x}_0, \mathbf{x}_0) + \boldsymbol{\lambda}^T(\mathbf{K} + \sigma^2\mathbf{I}_n)\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T\mathbf{k}_0 \end{aligned} \tag{III.13}$$

donde $\mathbf{k}_0 = (k(\mathbf{x}_0, \mathbf{x}_1), ..., k(\mathbf{x}_0, \mathbf{x}_n))^T$ y $\mathbf{K}$ es la matriz con $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1 : n$. De esta forma, el problema de Lagrange se formula como:

$$\min_{\boldsymbol{\lambda},\boldsymbol{\beta}} \{ \frac{1}{2}[k(\mathbf{x}_0, \mathbf{x}_0) + \boldsymbol{\lambda}^T(\mathbf{K} + \sigma^2\mathbf{I}_n)\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T\mathbf{k_0}] + \boldsymbol{\beta}^T(\mathbf{G}^T\boldsymbol{\lambda} - \mathbf{g}_0) \} \tag{III.14}$$

siendo $\boldsymbol{\beta} \in \mathbb{R}^r$ el vector de multiplicadores de Lagrange (Anexo D). Las condiciones de óptimo son:

$$\left\{ \begin{array}{l} (\mathbf{K} + \sigma^2\mathbf{I}_n)\boldsymbol{\lambda} + \mathbf{G}\boldsymbol{\beta} = \mathbf{k}_0 \\ \mathbf{G}^T\boldsymbol{\lambda} - \mathbf{g}_0 = \mathbf{0}_r \end{array} \right. \equiv \left[\begin{array}{cc} \mathbf{K} + \sigma^2\mathbf{I}_n & \mathbf{G} \\ \mathbf{G}^T & \mathbf{0}_{r\times r} \end{array} \right] \left[\begin{array}{c} \boldsymbol{\lambda} \\ \boldsymbol{\beta} \end{array} \right] = \left[\begin{array}{c} \mathbf{k}_0 \\ \mathbf{g}_0 \end{array} \right] \tag{III.15}$$

lo que denotamos brevemente por $\bar{\mathbf{K}}\bar{\boldsymbol{\lambda}} = \bar{\mathbf{k}}_0$. La solución del problema (III.14) es (pág. 413, Anexo G):

$$\bar{\boldsymbol{\lambda}} = \bar{\mathbf{K}}^{-1}\bar{\mathbf{k}}_0 \tag{III.16}$$

y el $\boldsymbol{\lambda}$ óptimo en función del vector de multiplicadores $\boldsymbol{\beta}$, es:

$$(\mathbf{K} + \sigma^2\mathbf{I}_n)\boldsymbol{\lambda} + \mathbf{G}\boldsymbol{\beta} = \mathbf{k}_0 \Rightarrow \boldsymbol{\lambda} = (\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}(\mathbf{k}_0 - \mathbf{G}\boldsymbol{\beta}) \tag{III.17}$$

Además, sustituyendo (III.16) en (III.13), y utilizando (III.15), se obtiene inmediatamente la varianza del error del kriging:

$$\mathrm{Var}(\mathsf{R}_0) = k(\mathbf{x}_0, \mathbf{x}_0) - \bar{\boldsymbol{\lambda}}^T\bar{\mathbf{k}}_0 \tag{III.18}$$

Denotando $\bar{\mathbf{c}} = \bar{\mathbf{K}}^{-1}\bar{\mathbf{y}}$, con $\bar{\mathbf{y}} = (\mathbf{y}^T, \mathbf{0}_{1\times r})^T$ donde $\mathbf{y} = (y_1, ..., y_n)^T$, y $\bar{\mathbf{c}}^T = (\mathbf{c}^T, \mathbf{b}^T)$ con $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^r$, la estimación kriging universal puede escribirse alternativamente como:

$$\begin{aligned} \hat{y}_0 = \boldsymbol{\lambda}^T\mathbf{y} = \bar{\boldsymbol{\lambda}}^T\bar{\mathbf{y}} = \bar{\mathbf{k}}_0^T\bar{\mathbf{K}}^{-1}\bar{\mathbf{y}} = \sum_{i=1}^{n} \lambda_i y_i \\ = \bar{\mathbf{k}}_0^T\bar{\mathbf{c}} = \sum_{i=1}^{n} c_i k(\mathbf{x}_0, \mathbf{x}_i) + \sum_{j=1}^{r} b_j g_j(\mathbf{x}_0) \end{aligned} \tag{III.19}$$

La expresión del estimador (III.19) en términos de $\bar{\boldsymbol{\lambda}}$ con las condiciones de óptimo anteriores se conoce como la formulación **primal** del Kriging, mientras que su expresión en términos de $\bar{\mathbf{c}} = \bar{\mathbf{K}}^{-1}\bar{\mathbf{y}}$, se conoce como su formulación **dual**, siendo $\bar{\mathbf{c}}$ la solución del sistema $\bar{\mathbf{K}}\bar{\mathbf{c}} = \bar{\mathbf{y}}$:

$$\left\{ \begin{array}{l} (\mathbf{K} + \sigma^2\mathbf{I}_n)\mathbf{c} + \mathbf{G}\mathbf{b} = \mathbf{y} \\ \mathbf{G}^T\mathbf{c} = \mathbf{0}_r \end{array} \right. \equiv \left[\begin{array}{cc} \mathbf{K} + \sigma^2\mathbf{I}_n & \mathbf{G} \\ \mathbf{G}^T & \mathbf{0}_{r\times r} \end{array} \right] \left[\begin{array}{c} \mathbf{c} \\ \mathbf{b} \end{array} \right] = \left[\begin{array}{c} \mathbf{y} \\ \mathbf{0}_r \end{array} \right] \tag{III.20}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Casos particulares del modelo descrito son:

1. El denominado **Kriging Ordinario**, cuando $g_1(\mathbf{x}) = 1\ \forall \mathbf{x} \in \mathcal{X}$, $\mu_1 = \mu$, $\mu_j = 0$, $j = 2 : r$, $\mathbf{G} = (1, ..., 1)^T = \mathbf{1}_n^T$, con lo que el $\boldsymbol{\lambda}$ óptimo en función del multiplicador $\boldsymbol{\beta} \equiv \beta$, es:

$$\boldsymbol{\lambda} = (\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}(\mathbf{k}_0 - \beta\mathbf{1}_n)$$

lo que produce como estimación dual:

$$\begin{aligned} \hat{y}_0 &= \boldsymbol{\lambda}^T\mathbf{y} = (\mathbf{k}_0 - \beta\mathbf{1}_n)^T(\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} \\ &= \mathbf{k}_0(\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} - \beta\mathbf{1}^t(\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} \\ &= \sum_{i=1}^{n} c_i k(\mathbf{x}_0, \mathbf{x}_i) + b \end{aligned}$$

con $b = -\beta\mathbf{1}^T(\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}$.

2. Y, en particular, el **Kriging Simple**, ($\mu = 0$), cuyo $\boldsymbol{\lambda}$ óptimo es:

$$\boldsymbol{\lambda} = (\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{k}_0 \tag{III.21}$$

lo que produce como estimación dual:

$$\hat{y}_0 = \boldsymbol{\lambda}^T\mathbf{y} = \mathbf{k}_0^T(\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} = \mathbf{k}_0^T\mathbf{c}$$

con $\mathbf{c} = (\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}$, análoga a la obtenida mediante procesos gausianos (ecuación (III.5) con $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$) y a la proporcionada por una red RBF de regularización (ecuación (III.9) con $\boldsymbol{\Sigma} = \lambda\mathbf{I}_n$), siempre que la función k sea la misma.

Kriging de una Función Aleatoria Intrínseca de Orden m

El estimador kriging universal puede generalizarse bajo el concepto de función aleatoria intrínseca de orden m [Matheron 1973], según las siguientes definiciones ([Journel 1989], [Wahba 1990*b*, p. 32], [Cressie 1993, sec. 5.4 y p. 363]):

Definición 70 *Dada una función aleatoria* $\mathsf{Y}(\mathbf{x}) = \sum_{j=1}^{r} \mu_j g_j(\mathbf{x}) + \mathsf{U}(\mathbf{x})$ *con* U *estacionario de segundo orden de media cero, con* $\mathbb{E}[\mathsf{Y}(\mathbf{x})] = \sum_{j=1}^{r} \mu_j g_j(\mathbf{x})$ *donde* $\{g_j\}_{j=1}^{r}$ *es una base para el espacio* $\Pi_m(\mathcal{X})$ *de polinomios*[89] *de grado menor que* $m \in \mathbb{Z}^+$*, se denomina* ***incremento generalizado de orden***[90] m *a toda combinación lineal* $\sum_{i=1}^{s} \alpha_i \mathsf{Y}(\mathbf{x}_i)$ *con* $\boldsymbol{\alpha} = (\alpha_1, ...\alpha_s)^T \neq \mathbf{0}_s$*, cuyos coeficientes verifican la condición:*

$$\mathbf{G}^T\boldsymbol{\alpha} = \mathbf{0}_r \equiv \sum_{i=1}^{s} \alpha_i g_j(\mathbf{x}_i) = 0,\ i = 1 : r \tag{III.22}$$

siendo $(\mathbf{G})_{ij} = g_j(\mathbf{x}_i), i = 1 : s, j = 1 : r$*, y* $\operatorname{rango}(\mathbf{G}) = r$*. En particular, un incremento generalizado de orden* 1 *verifica* $\sum_{i=1}^{s} \alpha_i = 0$.

[89] Esta definición puede generalizarse a cualquier combinación lineal de funciones conocidas g_j, $j = 1 : r$, no necesariamente polinomios [Cressie 1993].

[90] Este incremento generalizado se denomina de orden $m - 1$ en [Matheron 1973] (véase también [Cressie 1993, sec. 5.4]), pero la definición anterior permite establecer mejor la analogía con los procesos integrados de orden m (véase definición siguiente). Así un incremento generalizado de orden cero según [Matheron 1973] se denomina aquí de orden 1.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Un incremento generalizado de orden m del proceso $\mathsf{Y}(\mathbf{x})$ *filtra* su tendencia polinómica de grado menor que m pues, debido a (III.22):

$$\sum_{i=1}^{s}\alpha_i\mathsf{Y}(\mathbf{x}_i)=\boldsymbol{\alpha}^T\mathsf{Y}^s=\boldsymbol{\alpha}^T\mathbf{G}\boldsymbol{\mu}+\boldsymbol{\alpha}^T\mathsf{U}^s=\boldsymbol{\alpha}^T\mathsf{U}^s=\sum_{i=1}^{s}\alpha_i\mathsf{U}(\mathbf{x}_i)$$

donde U es proceso estacionario de segundo orden y $\mathsf{U}^s=(\mathsf{U}(\mathbf{x}_1),...,\mathsf{U}(\mathbf{x}_s))^T$.

Definición 71 *Se denomina* ***función aleatoria intrínseca***[91] ***de orden*** m *(IRF$_m$) con* $m\in\mathbb{Z}^+$, *a todo proceso aleatorio* Y *para el que se verifica que todo incremento generalizado de orden* m *es estacionario de segundo orden. Por tanto, una IRF$_m$ es una IRF$_{m+1}$.*

Con ello, una IRF$_m$ es una clase de equivalencia de procesos aleatorios, definida salvo media polinómica de grado menor que m.

[Matheron 1973] demostró que para una IRF$_m$ existe una única función k (salvo polinomios de grado $2m$), denominada **función de covarianza generalizada**, que verifica:

$$\mathrm{Var}(\boldsymbol{\alpha}^T\mathsf{Y}^n)=\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}>0$$
$$\text{con: }\mathbf{G}^T\boldsymbol{\alpha}=\mathbf{0}_r$$

donde $(\mathbf{K})_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)$, $i,j=1:s$. Es decir, k es una función estrictamente condicionalmente definida positiva de orden m (Anexo B), o equivalentemente, $-k$ es una función condicionalmente definida negativa de orden m.

Con estas definiciones, si en el planteamiento del kriging universal, se supone que el proceso $\mathsf{Y}(\mathbf{x})$ es una función aleatoria intrínseca de orden m, entonces, $\forall\boldsymbol{\lambda}\in\mathbb{R}^n$, el proceso $\mathsf{R}_0=\mathsf{R}(\mathbf{x}_0)=\sum_{i=1}^{n}\lambda_i\mathsf{Y}(\mathbf{x}_i)-\mathsf{Y}(\mathbf{x}_0)$ es estacionario de segundo orden y verifica:

$$\mathbf{G}^T\boldsymbol{\lambda}-\mathbf{g}_0=\bar{\mathbf{G}}^T\boldsymbol{\alpha}=\mathbf{0}_r$$
$$\text{con: }\bar{\mathbf{G}}^T\doteq(\mathbf{G}^T,\mathbf{g}_0)\text{ y }\boldsymbol{\alpha}\doteq(\boldsymbol{\lambda}^T,-1)^T$$

es decir, $\hat{\mathsf{Y}}(\mathbf{x}_0)=\sum_{i=1}^{n}\lambda_i\mathsf{Y}(\mathbf{x}_i)$ verifica la condición de insesgadez (III.12) (R_0 es de media cero), es decir, las restricciones que antes representaban la condición de insesgadez, se obtienen mediante la condición de incremento generalizado de orden m). Asimismo, existe una función k estrictamente condicionalmente definida positiva de orden m, que actúa como función de covarianza generalizada de $\mathsf{Y}(\mathbf{x})$.

Con ello, el estimador $\hat{\mathsf{Y}}(\mathbf{x}_0)$ de $\mathsf{Y}(\mathbf{x}_0)$, insesgado y de mínima varianza, se obtiene del mismo problema de optimización (III.14) que da lugar al kriging universal, sin necesidad de imponer una función definida positiva como función de covarianza del proceso[92] pues, basta utilizar la función de covarianza generalizada k de la IRF$_m$ $\mathsf{Y}(\mathbf{x})$, para que las ecuaciones (III.15) posean solución (única si k es estrictamente condicionalmente definida positiva).

Con ello, todas las ecuaciones anteriores (III.15), (III.19) y (III.20) del kriging universal son análogas a las del denominado **kriging-IRF$_m$** [Cressie 1993, p. 303]. Además, si una IRF$_m$ $\mathsf{Y}(\mathbf{x})$

[91] Con esta definición, una IRF$_m$ es equivalente [Cressie 1993, p. 301] a un proceso $\mathsf{Y}(t)$ integrado de orden m con $m\in\mathbb{N}^+$, es decir: $(1-\mathcal{B})^q\mathsf{Y}(t)=\varepsilon(t)$ donde $\varepsilon(t)$ es un proceso de ruido blanco y $\mathcal{B}$ tal que $\mathcal{B}\mathsf{Y}(t)=\mathsf{Y}(t-1)$, es el operador retardo. Utilizando la definición de [Matheron 1973], una IRF$_m$ es equivalente a un proceso integrado de orden $m+1$. Una IRF$_m$ con $m=0$ ($m=1$ según nuestra definición) sería un proceso integrado de orden 1).

[92] O una función de semivariograma (definida negativa de orden $m=1$) $\gamma(\mathbf{x},\mathbf{x}')=\frac{1}{2}\mathrm{Var}[\mathsf{Y}(\mathbf{x})-\mathsf{Y}(\mathbf{x}')]$ [Cressie 1993, p. 86 y 364]. Al no exigirse la existencia del variograma, se generaliza el estimador kriging universal a cualquier *variograma generalizado*, función condicionalmente negativa de orden $m\geq 1$, (serlo de orden $m=1$ implica serlo de orden $m>1$).

USC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

posee una auténtica función de covarianza, entonces el kriging con dicha función de covarianza, produce los mismos resultados que el kriging utilizando la función de covarianza generalizada (kriging-IRF$_m$), [Cressie 1993, p. 303].

III.2.2 Kriging y Redes RRBF

De lo anterior, y de la dualidad expresada en la ecuación (III.19) puede ya entreverse la estrecha relación entre el kriging universal y las redes RBF de regularización (RRBF). Esta red, se obtiene en el marco de un SRKHS (Anexo B) con núcleo una función k condicionalmente definida positiva de orden m y posee la forma general:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \psi(\mathbf{x} - \mathbf{x}_i) + \sum_{j=1}^{r} b_j g_j(\mathbf{x})$$

donde $\psi(\mathbf{x}-\mathbf{x}') \equiv k(\mathbf{x}, \mathbf{x}')$ y $\{g_j\} \subset \Pi_m(\mathcal{X})$ espacio de polinomios de grado menor que $m \in \mathbb{Z}^+$ (si $m = 0$, la función es definida positiva y no hay término adicional).

Los coeficientes de la solución se obtienen mediante la resolución del siguiente problema de regresión regularizado:

$$\hat{\mathcal{R}}_n(f) = \|\mathbf{y} - \mathbf{f}\|_2^2 + \lambda \|f\|_{\mathcal{H}_k} = \|\mathbf{y} - (\mathbf{Kc} + \mathbf{Gb})\|_2^2 + \lambda \mathbf{c}^T \mathbf{Kc} \qquad \text{(III.23)}$$

donde $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^T$.

La solución de dicho problema se obtiene mediante las ecuaciones (II.65) (pág. 135), las cuáles, coinciden con las ecuaciones (III.20) duales del kriging si [Matías, Vaamonde, Taboada y González-Manteiga 2004*a*]:

1. La función condicionalmente definida positiva de orden m, utilizada por la red RRBF coincide con la función de covarianza generalizada de la IRF$_m$.

2. El término *semiparamétrico* de la red RRBF utiliza las mismas funciones polinómicas g_j, $j = 1 : r$ de grado menor que m, que las utilizadas en la hipótesis de la media de la IRF$_m$.

3. El valor del **regularizador** λ utilizado en el problema (III.23), coincide con la **pepita** σ^2 de la función de covarianza generalizada de la IRF$_m$.

   En particular, la consideración de una IRF$_m$ sin pepita, equivale a considerar un regularizador nulo en el problema (III.23), dando prioridad al ajuste a los datos, lo que produce una solución interpoladora.

   En caso contrario, puede perderse la propiedad interpoladora (pág. 414). Esta circunstancia, a veces, no queda suficientemente clara en la literatura sobre estadística espacial, en la que el kriging suele presentarse como una técnica de interpolación.

   Esta circunstancia, se entiende mejor *en clave* de regresión: la consideración de un término de pepita $\sigma^2 > 0$, equivale a suponer que el proceso $\mathsf{Y}(\mathbf{x})$ representa observaciones sujetas a ruido[93] de un proceso subyacente $\mathsf{F}(\mathbf{x})$ que realmente interesa predecir (pág. 250). Es decir, $\mathsf{Y}(\mathbf{x}) = \mathsf{F}(\mathbf{x}) + \nu$ con ν ruido independiente con $\mathrm{Var}(\nu) = \sigma^2$ e independiente también de $\mathsf{F}(\mathbf{x})$. Con ello, $\kappa(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + \sigma^2 1_{\{\mathbf{x} \neq \mathbf{x}'\}}$, donde κ y k son las funciones de covarianza (generalizada) de $\mathsf{Y}(\mathbf{x})$ y $\mathsf{F}(\mathbf{x})$, respectivamente.

[93] Como dijimos en la nota al pie de la pág. 250, venimos considerando que la variabilidad de la variación de *microescala* está incluida en σ^2, (pero podría suponerse también que dicha variabilidad está incluida en k y que σ^2 recoge sólo la parte de pepita debido a errores de medición) [Cressie 1993, p. 113 y 128].

En este marco[94], conviene que el kriging no sea interpolante pues, de lo contrario, asimilaría el ruido de los datos $\{y_i\}_{i=1}^n$.

De lo anterior, se deduce que, en estas condiciones, **la forma dual de las RRBF coincide con la forma primal del kriging**. Utilizando (III.17) se obtiene su expresión como suavizador no paramétrico lineal (ecuación (II.67)):

$$\hat{\mathbf{f}}_n = \boldsymbol{\lambda}^T\mathbf{y} = (\mathbf{k}_0 - \mathbf{G}\boldsymbol{\beta})^T(\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} = \mathbf{S}\mathbf{y}$$

con $\mathbf{S} = (\mathbf{k}_0 - \mathbf{G}\boldsymbol{\beta})^T(\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}$ y $\boldsymbol{\beta}$ dado por la ecuación (G.8) de la pág. 413.

En particular, en el caso en que k sea una función condicionalmente definida positiva de orden $m = 1$, se obtiene una red RRBF con término constante:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i) + b$$

que, así, equivale al kriging ordinario. Y, en el caso en que k sea una función de covarianza (definida positiva), la red RRBF no posee término constante y equivale al kriging simple.

Dualidad entre Tendencia y Dependencia

La equivalencia formal entre ambas técnicas, presentada anteriormente, permite interpretar la solución:

$$\hat{\mathrm{Y}}(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{j=1}^{r} b_j g_j(\mathbf{x}) = \hat{f}_n(\mathbf{x}) \tag{III.24}$$

bajo dos enfoques duales entre sí (véase Apartado III.2.5), los cuáles, se diferencian fundamentalmente, en la proporción de variabilidad que asignan a una componente de tendencia (determinista) y a una componente estocástica [Matías, Vaamonde, Taboada y González-Manteiga 2004*a*]:

1. El primer enfoque (regresión con observaciones independientes) asigna toda la variabilidad a la función de tendencia y postula, para la misma, un modelo flexible, para adaptarse a los datos, y suave, para reproducir el grado de regularidad del fenómeno.

   Además, tratando de evitar el aprendizaje del posible ruido en los datos, pueden imponerse condiciones de suavidad adicionales a dicho modelo, mediante técnicas de regularización, cuya intensidad (parámetro regularizador), se estima a posteriori mediante técnicas de selección del modelo.

2. El segundo enfoque (predicción de una función aleatoria) distribuye la variabilidad del fenómeno entre la componente de tendencia y la componente estocástica.

   Posteriormente, modeliza primero esta última a partir de los datos (variograma o covariograma), y supone un modelo parámetrico para la primera, que, en general, es más sencillo que el segundo pues, una parte importante de la variabilidad del fenómeno se asigna a la componente estocástica.

   Sin embargo, el *trade-off* final entre ambas componentes depende de información a priori sobre el problema en cuestión.

[94] Como puede observarse, el planteamiento anterior es realmente un problema de regresión con observaciones dependientes: exactamente el mismo que el de un problema de predicción de una función aleatoria.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Bajo la dualidad anterior, el término $\sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i)$ de la expresión (III.24) **puede verse como un término debido a la tendencia (RRBF) o como un término debido a la estructura de dependencia (kriging)**, y en ambos casos, las propiedades analíticas de k determinan el grado de regularidad de la estimación.

Por su parte, el término $\sum_{j=1}^{r} d_j g_j(\mathbf{x})$ de la expresión (III.24) constituye, en ambos enfoques, un término de tendencia, (si bien no es frecuente su utilización en las RRBF debido a que no es imprescindible para que dichas redes posean la capacidad de aproximación a cualquier función continua).

Diferencias entre Kriging y RRBF

De cualquier forma, existen tres grandes diferencias entre el kriging y las RRBF en su práctica habitual [Matías, Vaamonde, Taboada y González-Manteiga 2004*a*], similares a las existentes entre el kriging y los *splines*, y que han originado cierta controversia ([Cressie 1989], [Wahba 1990*a*], [Cressie 1990]):

1. La primera diferencia reside en la forma de la elegir la función k. Mientras en el kriging, esta función (o el variograma, –usualmente anisotrópico) se estima a partir de los datos utilizando métodos bien fundamentados, por el contrario, en el ámbito de las RRBF el núcleo se selecciona suponiendo una determinada hipótesis de suavidad para la función de regresión (pág. 131).

   Sin embargo, en la práctica de las RRBF, resulta difícil muchas veces codificar la información a priori en estos términos. Por ello, con el fin de garantizar la solución única al problema, es frecuente utilizar núcleos estándar como el gausiano o funciones polinómicas (*splines*) [Wahba 1990*b*], sin excesiva fundamentación empírica (frecuentemente isotrópicos), en la creencia de que, la elección del núcleo, influye menos en los resultados que la estimación de los parámetros de la familia correspondiente.

   Sin embargo, como veremos en las aplicaciones, esta práctica es, cuando menos, arriesgada, y no aprovecha la información contenida en los datos para formular hipótesis sobre el núcleo.

   Una de las circunstancias que ha podido originar esta costumbre es la frecuente alta dimensionalidad de los problemas de aplicación típicos de las redes neuronales, en los que la tarea anterior resulta muy difícil.

2. La segunda gran diferencia reside en el hecho de que, debido a las hipótesis estadísticas que gobiernan el kriging, éste, al contrario que las RRBF, suministra la varianza del error (III.18) (aunque los *splines* pueden verse formalmente como un caso particular del kriging, p. ej. [Laslett 1994]).

   Las RRBF (y las RBF) son suavizadores de tipo lineal, lo que permite estimar la varianza del estimador de la función de regresión. Sin embargo, no puede estimarse la varianza de la predicción puntual salvo que se estime por algún método adicional la varianza del ruido.

3. Finalmente, es preciso destacar la capacidad del kriging, no compartida por las RRBF, para predecir determinadas funciones del proceso aleatorio en el marco de diferentes soportes [Cressie 1990].

A priori, estas circunstancias, inclinan la balanza en favor de la metodología de la estadística espacial cuando la dimensionalidad del espacio de entrada no es mayor que 3. En problemas

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

de mayor dimensionalidad, las técnicas de estimación de la estructura de dependencia, aunque deseables, resultan más problemáticas.

Las características de nuestro problema de aplicación nos impedirán evaluar el impacto de los dos últimos aspectos anteriores, pero nos permitirán comprobar experimentalmente los efectos del primero.

III.2.3 Kriging y SVM

En el capítulo anterior (pág. 151), vimos que las redes SVM con término independiente $b = 0$, podrían verse como solución de un problema de regularización en un RKHS $\mathcal{H}_k$ asociado a una función definida positiva k [Evgeniou *et al.* 2000]:

$$\min_{\mathbf{w}}\{\frac{1}{2}\left\|\mathbf{w}\right\|_{\mathcal{U}}^2 + C\sum_{i=1}^{n}\left|y - f_{\mathbf{w}}(\mathbf{x})\right|_{\varepsilon}\} \equiv \min_{f\in\mathcal{H}_k}\{\sum_{i=1}^{n}\left|y - f_{\boldsymbol{\alpha}(\mathbf{w})}(\mathbf{x})\right|_{\varepsilon} + \frac{1}{2C}\left\|f_{\boldsymbol{\alpha}(\mathbf{w})}\right\|_{\mathcal{H}_k}^2\} \quad \text{(III.25)}$$

donde $\mathbf{w} = \sum_{i=1}^{n}\alpha_i\boldsymbol{\phi}(\mathbf{x}_i)$ y $f_{\mathbf{w}} = \sum_{i=1}^{n}\alpha_i k(\cdot,\mathbf{x}_i) = f_{\boldsymbol{\alpha}(\mathbf{w})} \in \mathcal{H}_k$ (donde n puede ser ∞ –entendido como límite), puesto que, en el espacio de características $\mathcal{U}$ con producto interno definido por el núcleo k, se verifica que:

$$\left\|\mathbf{w}\right\|_{\mathcal{U}}^2 = \langle\mathbf{w},\mathbf{w}\rangle_{\mathcal{U}} = \sum_{i,j=1}^{n}\alpha_i\alpha_j k(\mathbf{x}_i,\mathbf{x}_j) = \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$$

que es también la norma $\left\|f_{\boldsymbol{\alpha}(\mathbf{w})}\right\|_{\mathcal{H}_k}^2$ de la función $f_{\boldsymbol{\alpha}(\mathbf{w})}$.

De hecho, en este marco (pág. 151), vimos que la única diferencia entre estas SVM y las RRBF sin término independiente, estaba producida por la pérdida utilizada en el problema de regularización: pérdida cuadrática en las RRBF y la pérdida ε-insensible de Vapnik en las SVM.

SVM Semiparamétrica

En el contexto anterior, si en lugar de suponer k definida positiva, se supone condicionalmente definida positiva de orden m, entonces $\mathcal{H}_k$ es un SRKHS y, por el teorema 39 (pág. 129), la solución al problema de regularización (III.25) en $\mathcal{H}_k$ posee una forma general análoga a la de las RRBF:

$$f_{\boldsymbol{\alpha}(\mathbf{w}),\mathbf{b}}(\mathbf{x}) = \sum_{i=1}^{n}\alpha_i k(\mathbf{x},\mathbf{x}_i) + \sum_{j=1}^{r}b_j g_j(\mathbf{x}) \quad \text{(III.26)}$$

donde $\{g_j\}_{i=1}^{r}$ es una base de polinomios de $\Pi_m(\mathcal{X})$.

Sustituyendo dicha forma general en (III.25), se obtiene del siguiente programa cuadrático que da lugar a la denominada **SVM semiparamétrica** con término polinómico ([Smola *et al.* 1999], [Matías, Vaamonde, Taboada y González-Manteiga 2004*b*]):

$$\min_{f\in\mathcal{H}_k}\{\sum_{i=1}^{n}\left|y - f_{\boldsymbol{\alpha}(\mathbf{w}),\mathbf{b}}(\mathbf{x})\right|_{\varepsilon} + \frac{1}{2C}\left\|f_{\boldsymbol{\alpha}(\mathbf{w}),\mathbf{b}}\right\|_{\mathcal{H}_k}^2\} \equiv \min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi},\boldsymbol{\xi}'}\{\frac{1}{2}\left\|\mathbf{w}\right\|_{\mathcal{U}}^2 + C\sum_{i=1}^{n}(\xi_i+\xi_i')\}$$
$$\left\{\begin{array}{c} \langle\mathbf{w},\boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + \sum_{j=1}^{r}b_j g_j(\mathbf{x}_i) - y_i \le \varepsilon + \xi_i \\ y_i - [\langle\mathbf{w},\boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + \sum_{j=1}^{r}b_j g_j(\mathbf{x}_i)] \le \varepsilon + \xi_i' \\ \xi_i,\xi_i' \ge 0 \end{array}\right| \; i = 1:n \quad \text{(III.27)}$$

Un caso particular, es cuando $m = 1$, en el que se obtiene la SVM con término constante b.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Por tanto, dada la analogía entre la formulación dual del kriging y las redes RRBF, lo anterior muestra que la diferencia entre el kriging dual y las SVM semiparamétricas, reside en la utilización de la pérdida cuadrática o la pérdida ε-insensible de Vapnik, respectivamente, en el correspondiente problema de regularización [Matías, Vaamonde, Taboada y González-Manteiga 2004*b*], siempre que el parámetro regularizador de las SVM $\frac{1}{2C}$ coincida con el la pepita σ^2 considerada en el kriging : $\lambda = \frac{1}{2C} = \sigma^2$.

Esta comparación puede llevarse asimismo al terreno de las SVM teniendo en cuenta que, como vimos en el capítulo anterior (pág. 151), el problema de regularización que da lugar a las RRBF (y a su equivalente, el kriging dual):

$$\min_{f \in \mathcal{H}_k} \hat{\mathcal{R}}_n(f) = \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i)]^2 + \lambda \left\| f \right\|^2_{\mathcal{H}_k}$$

puede formularse también como un programa cuadrático con restricciones lineales en el espacio de características $\mathcal{U}$ utilizando la metodología SVM:

$$\begin{array}{c} \min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi},\boldsymbol{\xi}'} \{ \sum_{i=1}^{n} \xi_i^2 + \frac{1}{2C} \left\| \mathbf{w} \right\|^2_{\mathcal{U}} \} \equiv \min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi},\boldsymbol{\xi}'} \{ \frac{1}{2} \left\| \mathbf{w} \right\|^2_{\mathcal{U}} + C \sum_{i=1}^{n} \xi_i^2 \} \\ y_i - [\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i) \rangle_{\mathcal{U}} + \sum_{j=1}^{r} b_j g_j(\mathbf{x}_i)] = \xi_i, \; i = 1:n \end{array}$$

donde $\xi_i = y_i - f(\mathbf{x}_i)$ y $\frac{1}{2C} = \lambda = \sigma^2$ es el parámetro regularizador.

Sin embargo, la solución resultante no posee la parsimonia propia de las SVM basada en la utilización de los puntos verdaderamente relevantes para el problema: los vectores soporte. A diferencia de la pérdida cuadrática, muy sensible a los posibles datos atípicos de la muestra, la pérdida ε-insensible dota a las SVM de un mecanismo de gestión de estos datos a través del parámetro ε que refleja el grado de fiabilidad que los datos poseen para el experimentador.

Dada una base de polinomios $\{g_j\} \in \Pi_m(\mathcal{X})$, $j = 1:r$, definidas a priori, el problema (III.27) es un programa cuadrático convexo (Anexo D) que, por tanto posee solución única. Dicha solución[95], se obtiene mediante la optimización del Lagrangiano con respecto a las variables primales $\mathbf{w}, \mathbf{b}, \boldsymbol{\xi}, \boldsymbol{\xi}'$ y duales $\boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\rho}, \boldsymbol{\rho}'$ (damos por supuesto el subíndice $\mathcal{U}$ en el producto interno):

$$\begin{aligned} L(\mathbf{w}, \mathbf{b}, \boldsymbol{\xi}, \boldsymbol{\xi}', \boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\rho}, \boldsymbol{\rho}') \;=\; & \frac{1}{2} \left\| \mathbf{w} \right\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i') \\ & + \sum_{i=1}^{n} \alpha_i [\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i) \rangle + \sum_{j=1}^{r} b_j g_j(\mathbf{x}_i) - y_i - \varepsilon - \xi_i] \\ & + \sum_{i=1}^{n} \alpha_i' [y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i) \rangle + \sum_{j=1}^{r} b_j g_j(\mathbf{x}_i)) - \varepsilon - \xi_i'] \\ & - \sum_{i=1}^{n} \rho_i \xi_i - \sum_{i=1}^{n} \rho_i' \xi_i' \end{aligned}$$

[95] Por tanto, esta deducción engloba, como casos particulares, los casos $b = 0$ y $b \neq 0$ vistos en el capítulo anterior. Los problemas *hard* y *soft margin* del problema de clasificación siguen la misma metodología.

Las condiciones necesarias y suficientes de óptimo para las variables primales son:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n}(\alpha_i' - \alpha_i)\boldsymbol{\phi}(\mathbf{x}_i) = 0 \\
\frac{\partial L}{\partial b_j} &= -\sum_{i=1}^{n}(\alpha_i' - \alpha_i)g_j(\mathbf{x}_i) = 0,\ j = 1:r \\
\frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \rho_i = 0,\ i = 1:n \\
\frac{\partial L}{\partial \xi_i'} &= C - \alpha_i' - \rho_i' = 0,\ i = 1:n
\end{aligned}$$

que, sustituyéndolas en el Lagrangiano, producen el siguiente problema dual:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathbb{R}^n} \{-\tfrac{1}{2}\sum_{i,j=1}^{n}(\alpha_i' - \alpha_j)(\alpha_j' - \alpha_j)k(\mathbf{x}_i,\mathbf{x}_j) - \varepsilon\sum_{i=1}^{n}(\alpha_i' + \alpha_i) + \textstyle\sum_{i=1}^{n}(\alpha_i' - \alpha_i)y_i \\
&\qquad \left\{ \begin{array}{l} \sum_{i=1}^{n}(\alpha_i' - \alpha_i)g_j(\mathbf{x}_i) = 0,\ j = 1:r \\ 0 \le \alpha_i, \alpha_i' \le C,\ i = 1,...,n \end{array} \right.
\end{aligned} \tag{III.28}$$

con las siguientes condiciones adicionales de óptimo de Kuhn-Tucker, (complementariedad), (Anexo D):

$$\left\{ \begin{array}{l} \alpha_i[\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + \sum_{j=1}^{q} b_j g_j(\mathbf{x}_i) - y_i - \varepsilon - \xi_i] = 0 \\ \alpha_i'[y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + \sum_{j=1}^{q} b_j g_j(\mathbf{x}_i)) - \varepsilon - \xi_i'] = 0 \\ \xi_i\xi_i' = 0,\ \alpha_i\alpha_i' = 0 \\ (C - \alpha_i)\xi_i = 0,\ (C - \alpha_i')\xi_i' = 0 \end{array} \right| , i = 1:n \tag{III.29}$$

pues $0 = \xi_i\rho_i = \xi_i(C - \alpha_i)$ y $0 = \xi_i'\rho_i' = \xi_i'(C - \alpha_i')$.

Sobre la solución del programa (III.28)-(III.29), destacamos lo siguiente haciendo $\beta_i = \alpha_i' - \alpha_i$, pueden hacerse comentarios similares a los realizados para la SVM usual (pág. 148), (el problema de clasificación en sus versiones *hard* y *soft margin* admite observaciones análogas):

1. El óptimo $\mathbf{w}$ posee la forma: $\mathbf{w} = \sum_{i=1}^{n}\beta_i\boldsymbol{\phi}_i$ con $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\mathbf{x}_i)$, $i = 1:n$, con lo que la solución se expresa en términos del producto interno con los puntos de la muestra en el espacio de características.

2. Los vectores que satisfacen $\beta_i = \alpha_i' - \alpha_i \neq 0$, ($\alpha_i > 0$ o $\alpha_i' > 0$), es decir, los que saturan las dos primeras restricciones en (III.29), son los únicos que participan en la expresión de la solución. Por ello, se denominan **vectores soporte** (*support vectors*). De ellos, los que verifican $0 < \alpha_i < C$ o $0 < \alpha_i' < C$ están ubicados justo en la frontera de la banda de radio ε, y el resto, $\alpha_i = C$ o $\alpha_i' = C$, quedan fuera de esa banda contabilizados como errores por la función de pérdida (figura 37).

3. Con ello, el óptimo es combinación lineal de los vectores soporte $\mathbf{w} = \sum_{\text{S.V.}}\beta_i\boldsymbol{\phi}_i$ y la solución al problema (III.28) viene dada por:

$$\begin{aligned}
f_{\mathbf{w},\mathbf{b}}(\mathbf{x}) &= \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x})\rangle + \sum_{j=1}^{q} b_j g_j(\mathbf{x}) = \sum_{\text{S.V.}} \beta_i \langle \boldsymbol{\phi}_i, \boldsymbol{\phi}(\mathbf{x})\rangle + \sum_{j=1}^{q} b_j g_j(\mathbf{x}) \qquad \text{(III.30)} \\
&= \sum_{\text{S.V.}} \beta_i k(\mathbf{x}_i, \mathbf{x}) + \sum_{j=1}^{q} b_j g_j(\mathbf{x})
\end{aligned}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

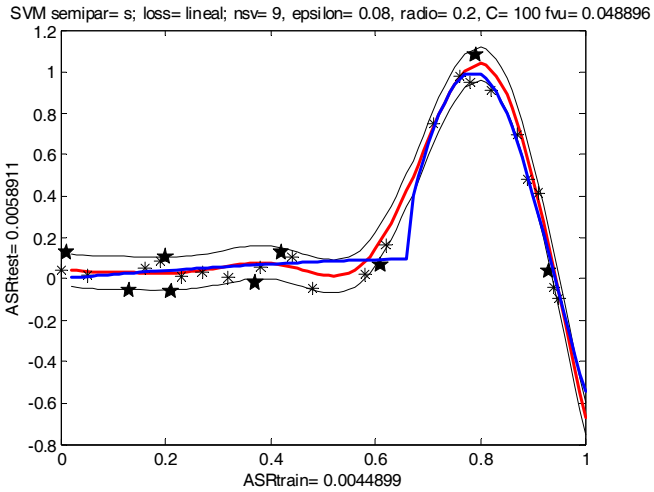


Figura 37. Ajuste de una SVM semiparamétrica con términos constante y lineal, entrenada con la pérdida ε-insensible lineal de Vapnik. En el ejemplo, se obtiene el mismo número de vectores soporte que para la SVM usual con la misma pérdida (figura 21).

con lo que la resolución del problema primal mediante la formulación dual permite reducir el número de términos de la solución al basarse ésta en las restricciones activas del primal (correspondientes a los vectores realmente relevantes o *difíciles* para el problema), en lugar de utilizar tantos términos como datos en la muestra.

4. Los valores óptimos para el vector de coeficientes $\mathbf{b} = (b_1, \ldots, b_q)^T$ del término paramétrico se obtienen de las dos primeras ecuaciones de (III.29) correspondientes a puntos (vectores soporte) que están sobre la banda de radio ε, es decir, que verifican $0 < \beta_i = \alpha_i' - \alpha_i < C$ o $-C < \beta_i = \alpha_i' - \alpha_i < 0$, (ya que $\alpha_i' \alpha_i = 0$).

   Pero, también es posible obtenerlos como producto del algoritmo de optimización cuando se utiliza el método del punto interior (Anexo D) pues son las variables duales del problema dual.

III.2.4 Kriging Regularizado

En la formulación del kriging universal, cuando las observaciones están sujetas a ruido, el término $\sigma^2 \mathbf{I}_n$ de las ecuaciones (III.15) y (III.20), juega el mismo papel que el término $\lambda \mathbf{I}_n$ en la regresión *ridge* generalizada asociada a las redes RRBF. Dicho término, produce el suavizado de los datos impidiendo la interpolación con la esperanza de conseguir una mayor capacidad de generalización del estimador resultante.

En el caso de las SVM, se dispone de un mecanismo similar, en el que el inverso de C juega el mismo papel que el regularizador λ, pero más potente, debido a la utilización de la pérdida ε-insensitiva que define una banda $\pm\varepsilon$ entorno a los datos, dentro de la cuál, la pérdida no percibe error alguno (figura 37).

Con ambos mecanismos, se pretende admitir un pequeño nivel de sesgo en la solución si ello incrementa la capacidad de generalización del estimador resultante.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Sin embargo, en el marco de las hipótesis estocásticas de la estadística espacial, el kriging universal, que resulta como contrapartida de las RRBF (con ruido en los datos), es el mejor estimador lineal insesgado de mínima varianza, con lo que, el suavizado de los datos, en su caso, va destinado a contrarrestar el ruido pero no a admitir relajación alguna de las hipótesis de insesgadez [Matías y González-Manteiga 2003*b*]. (Recuérdese, que el contenido de la media es diferente en los dos ámbitos: en el kriging la media es el término polinómico, mientras que, en las RRBF, la media es una combinación lineal del núcleo).

Esta relajación de la hipótesis de insesgadez, podría resultar beneficiosa en general, cuando el proceso de interés Y no es gausiano[96] [Cressie 1993, p. 110], pero también cuando, aún siendo Y gausiano, existen serias dudas de que la hipótesis paramétrica sobre la media del proceso se encuentre mal identificada. En ambos casos, el kriging universal no minimizaría el error cuadrático medio:

$$\begin{aligned} \mathrm{MSE}(\hat{\mathsf{Y}}_0) &= \mathbb{E}[(\hat{\mathsf{Y}}_0 - \mathsf{Y}_0)^2] = \mathbb{E}[\mathsf{R}_0^2] = \mathrm{Var}(\mathsf{R}_0) + [\mathbb{E}(\mathsf{R}_0)]^2 \\ &= \mathrm{Var}(\hat{\mathsf{Y}}_0 - \mathsf{Y}_0) + [\mathrm{Sesgo}(\hat{\mathsf{Y}}_0)]^2 \end{aligned}$$

donde $\hat{\mathsf{Y}}_0 = \mathsf{Y}(\mathbf{x}_0) = \sum_{i=1}^{n} \lambda_i \mathsf{Y}(\mathbf{x}_i)$ y $\mathsf{R}_0 = \mathsf{R}(\mathbf{x}_0) = \hat{\mathsf{Y}}_0 - \mathsf{Y}_0$ es el error. Esta circunstancia es destacada ya por [Cressie 1993, p. 174] y [Gotway y Cressie 1993] que analizan diferentes estimadores resultado de permitir un cierto nivel de sesgo.

Como vimos anteriormente, el estimador kriging universal impone $\mathrm{Sesgo}(\hat{\mathsf{Y}}_0) = 0$ (o se supone una IRF_m), y en este marco, minimiza la varianza. Con objeto de evaluar los posibles beneficios de admitir un cierto grado de sesgo en el kriging, formulamos el siguiente problema de estimación regularizado [Matías y González-Manteiga 2003*b*]:

$$\hat{\mathsf{Y}}(\mathbf{x}_0) = \arg \min_{\mathsf{T}(\mathbf{x}_0) \in \mathcal{F}} \{\mathrm{Var}[\mathsf{T}(\mathbf{x}_0) - \mathsf{Y}_0] + C \cdot [\mathrm{Sesgo}(\mathsf{T}(\mathbf{x}_0))]^2\} \tag{III.31}$$

donde $\mathcal{F}$ es el conjunto de los estimadores lineales y C es una constante regularizadora tal que:

1. Si $C = 1$, se tiene como caso particular el error cuadrático medio del estimador.
2. Si $C = 0$, el problema se reduce a la determinación del estimador lineal de mínima varianza (sin restricciones en cuanto al sesgo).
3. Si $C \to \infty$, la prioridad es la obtención de un estimador lineal insesgado.

Supongamos las mismas hipótesis formuladas en el marco del kriging universal, siendo Y el proceso de interés (III.10) con función de covarianza (III.11) y media:

$$\mathbb{E}[\mathsf{Y}(\mathbf{x})] = \sum_{j=1}^{r} \mu_j g_j(\mathbf{x})$$

donde $\mu_j \neq 0$ son constantes desconocidas y g_j son funciones conocidas, $j = 1 : r$. La condición de insesgadez (III.12), se consigue con:

$$\boldsymbol{\mu}_j[\langle \boldsymbol{\lambda}, \mathbf{g}_j \rangle - g_j(\mathbf{x}_0)] = 0,\ j = 1 : r \tag{III.32}$$

Con objeto de formular el equilibrio reflejado en (III.31) entre el cuadrado del sesgo y la varianza, planteamos el siguiente problema de optimización basado en el enfoque de las variables

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[96] Si el proceso es gausiano y la media está correctamente identificada (es conocida, –kriging simple), es fácil comprobar que el kriging universal minimiza el error cuadrático medio [Cressie 1993, p. 109]

de holgura de las SVM (compárese con el problema (III.14) correspondiente al kriging):

$$\begin{array}{c} \min\limits_{\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\xi}'} \frac{1}{2}\{k(\mathbf{x}_0,\mathbf{x}_0)+\boldsymbol{\lambda}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)\boldsymbol{\lambda}-2\boldsymbol{\lambda}^T\mathbf{k}_0+C\sum\limits_{j=1}^{r}(\xi_j^2+\xi_j'^2)\} \\ \text{sujeto a: } \left\{\begin{array}{c} \mu_j[g_j(\mathbf{x}_0)-\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle]\leq\varepsilon+\xi_j \\ \mu_j[\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle-g_j(\mathbf{x}_0)]\leq\varepsilon+\xi_j' \\ \xi_j,\xi_j'\geq 0 \end{array}\right| , j=1:r \end{array} \tag{III.33}$$

donde $\varepsilon \geq 0$ es una constante, $\boldsymbol{\xi},\boldsymbol{\xi}' \in (\mathbb{R}^+ \cup \{0\})^r$ cuyas componentes ξ_j, ξ_j', $j = 1 : r$, son variables *slack* y $\mathbf{g}_j = (g_j(\mathbf{x}_1), ..., g_j(\mathbf{x}_n))^T$ es la j-ésima componente conocida de la media, valorada en los puntos de la muestra.

Al respecto de esta formulación, realizamos las siguientes observaciones:

1. Los tres primeros términos de la función objetivo en (III.33), recogen la varianza del estimador $\hat{\mathrm{Y}}(\mathbf{x}_0) = \sum_{i=1}^{n} \lambda_i \mathrm{Y}(\mathbf{x}_i)$, de los cuáles, el primero no depende de $\boldsymbol{\lambda}$ y no es relevante para la función objetivo.

2. Las restricciones anteriores permiten que las condiciones de insesgadez (III.32) puedan ser violadas en cada componente de la media, si ello resulta en alguna mejora para la función objetivo. Las variables slack recogen el sesgo, por exceso o por defecto, en cada componente, con lo que su suma recoge el sesgo del estimador, y la suma de sus cuadrados, $\sum_{j=1}^{r}(\xi_j^2+\xi_j'^2)$, es una aproximación al cuadrado del sesgo pues no se consideran los productos cruzados. (Con la pérdida absoluta, el desarrollo es similar a lo que sigue).

3. La constante $\varepsilon \geq 0$ se mantiene para mayor generalidad, y permite contrarrestar la aproximación al cuadrado del sesgo debido a la ausencia de los productos cruzados. Al mismo tiempo, podría reflejar el grado de desconfianza en las hipótesis sobre la media del proceso, de tal forma que a mayor ε, mayor permisibilidad con el sesgo del estimador.

4. Finalmente, las restricciones $\xi_j, \xi_j' \geq 0$ no son necesarias, pues, si las restricciones principales se verificasen para valores negativos de ξ_j y ξ_j', también se verificarían para valores cero, siendo éstos más favorables para la función objetivo.

Como el vector $\boldsymbol{\mu}$ desconocido aparece ahora en las restricciones, realizamos las siguientes transformaciones. Si $\mu_j > 0$:

$$\left\{\begin{array}{l} \mu_j[g_j(\mathbf{x}_0)-\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle]\leq\varepsilon+\xi_j \\ \mu_j[\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle-g_j(\mathbf{x}_0)]\leq\varepsilon+\xi_j' \end{array}\right. \equiv \left\{\begin{array}{l} [g_j(\mathbf{x}_0)-\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle]\leq\varepsilon/\mu_j+\xi_j/\mu_j=\varepsilon_j+\bar{\xi}_j \\ [\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle-g_j(\mathbf{x}_0)]\leq\varepsilon/\mu_j+\xi_j'/\mu_j=\varepsilon_j+\bar{\xi}_j' \end{array}\right.$$

con $\varepsilon_j = \varepsilon/\mu_j, \bar{\xi}_j = \xi_j/\mu_j, \bar{\xi}_j' = \xi_j'/\mu_j \geq 0$. Análogamente, si $\mu_j < 0$:

$$\left\{\begin{array}{l} [g_j(\mathbf{x}_0)-\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle]\geq\varepsilon_j+\bar{\xi}_j \\ [\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle-g_j(\mathbf{x}_0)]\geq\varepsilon_j+\bar{\xi}_j' \end{array}\right. \equiv \left\{\begin{array}{l} -[g_j(\mathbf{x}_0)-\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle]\leq-\varepsilon_j-\bar{\xi}_j \\ -[\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle-g_j(\mathbf{x}_0)]\leq-\varepsilon_j-\bar{\xi}_j' \end{array}\right.$$

donde ahora $\varepsilon_j = \varepsilon/\mu_j, \bar{\xi}_j = \xi_j/\mu_j, \bar{\xi}_j' = \xi_j'/\mu_j \leq 0$, por lo que, considerando como constante $-\varepsilon_j \geq 0$,y como variables slack $-\bar{\xi}_j, -\bar{\xi}_j' \geq 0$ se tienen restricciones de la misma forma que en el caso anterior. Con todo ello, aproximamos el término del sesgo en la función objetivo, de la

siguiente forma:

$$C\sum_{j=1}^{r}(\xi_j^2+\xi_j'^2)=C\sum_{j=1}^{r}\mu_j^2(\bar{\xi}_j^2+\bar{\xi}_j'^2)\leq C\max_{j\in\{1,r\}}\{\mu_j^2\}\sum_{j=1}^{r}(\bar{\xi}_j^2+\bar{\xi}_j'^2)=\bar{C}\sum_{j=1}^{r}(\bar{\xi}_j^2+\bar{\xi}_j'^2)$$

Esta aproximación resulta exacta en el caso $r=1$ (kriging ordinario), en el que $\mathbb{E}[\mathsf{Y}(\mathbf{x})]=\mu g_1(\mathbf{x})=\mu$ y $\max_{j\in\{1,r\}}\{\mu_j^2\}=\mu^2$, en cuyo caso, si desea que la función objetivo represente el error cuadrático medio, $\bar{C}$ puede estimarse mediante $\bar{y}^2$ donde:

$$\bar{y}=\hat{\mu}_{OLS}=(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

al ser $\mathbf{G}=\mathbf{1}_n$, o bien, mediante su versión generalizada[97]:

$$\bar{y}=\hat{\mu}_{GLS}=(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{K}^{-1}\mathbf{y}$$

De esta forma, denominando C a $\bar{C}$ y $\varepsilon_j,\xi_j,\xi_j'$ a las constantes ≥ 0 y a las variables *slack* ≥ 0, $j=1:r$, planteamos el siguiente problema de optimización[98]:

$$\min_{\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\xi}'}\frac{1}{2}\{\boldsymbol{\lambda}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)\boldsymbol{\lambda}-2\boldsymbol{\lambda}^T\mathbf{k}_0+C\sum_{j=1}^{r}(\xi_j^2+\xi_j'^2)\} \tag{34}$$

$$\text{sujeto a: } \left\{\begin{array}{l} g_j(\mathbf{x}_0)-\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle\leq\varepsilon_j+\xi_j \\ \langle\boldsymbol{\lambda},\mathbf{g}_j\rangle-g_j(\mathbf{x}_0)\leq\varepsilon_j+\xi_j'\end{array}\right|,\ j=1:r \tag{35}$$

Nótese al respecto que una y sólo una de cada par de restricciones j puede saturarse, (si $\varepsilon_j>0$ pues, si $\varepsilon_j=0$, cabe la posibilidad de saturación simultánea en cero), lo que significa que las variables *slack* correspondientes verifican $\xi_j\xi_j'=0$.

El problema de optimización (III.34) es un problema cuadrático con restricciones lineales, con lo que las condiciones de Kuhn-Tucker son condiciones necesarias y suficientes de óptimo (Anexo D). La solución se obtiene mediante la optimización del Lagrangiano respecto a las variables primales y duales:

$$\begin{aligned}\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}\min_{\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\xi}'}\{L(\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\xi}',\boldsymbol{\alpha},\boldsymbol{\alpha}') &= \frac{1}{2}[\boldsymbol{\lambda}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)\boldsymbol{\lambda}-2\boldsymbol{\lambda}^T\mathbf{k}_0]+\frac{C}{2}\sum_{j=1}^{r}(\xi_j^2+\xi_j'^2)\\ &+\sum_{j=1}^{r}\alpha_j[g_j(\mathbf{x}_0)-\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle-\varepsilon_j-\xi_j]\\ &+\sum_{j=1}^{r}\alpha_j'[\langle\boldsymbol{\lambda},\mathbf{g}_j\rangle-g_j(\mathbf{x}_0)-\varepsilon_j-\xi_j']\end{aligned}$$

97 Que se obtiene del problema de minimos cuadrados generalizados: $\min_{\boldsymbol{\mu}}(\mathbf{y}-\mathbf{G}\boldsymbol{\mu})^T\mathbf{K}^{-1}(\mathbf{y}-\mathbf{G}\boldsymbol{\mu})$.

98 Esta formulación sugiere una generalización de las SVM utilizando un parámetro ε heterocedástico: $\varepsilon=\varepsilon(x)$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

cuyas condiciones de óptimo para las variables primales son, (compárese la primera con (III.17)):

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\lambda}} &= (\mathbf{K}+\sigma^2\mathbf{I}_n)\boldsymbol{\lambda}-\mathbf{k}_0+\sum_{j=1}^{r}(\alpha'_j-\alpha_j)\mathbf{g}_j=0 \\
&\Rightarrow \quad \boldsymbol{\lambda}=(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}[\mathbf{k}_0-\mathbf{G}(\boldsymbol{\alpha}'-\boldsymbol{\alpha})] \qquad \text{(III.36)} \\
\frac{\partial L}{\partial \xi_j} &= C\xi_j-\alpha_j=0\Rightarrow \xi_j=\frac{\alpha_j}{C},\ j=1:r \qquad \text{(III.37)} \\
\frac{\partial L}{\partial \xi'_j} &= C\xi'_j-\alpha'_j=0\Rightarrow \xi'_j=\frac{\alpha'_j}{C},\ j=1:r \qquad \text{(III.38)}
\end{aligned}$$

Sustituyendo estas condiciones en el Lagrangiano se obtiene el siguiente problema dual:

$$\begin{aligned}
\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}'}\{&-\tfrac{1}{2}(\boldsymbol{\alpha}'-\boldsymbol{\alpha})^T\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G}(\boldsymbol{\alpha}'-\boldsymbol{\alpha})+(\boldsymbol{\alpha}'-\boldsymbol{\alpha})^T\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k}_0 \\
&-\sum_{j=1}^{r}\varepsilon_j(\alpha_j+\alpha'_j)-(\boldsymbol{\alpha}'-\boldsymbol{\alpha})^T\mathbf{g}(\mathbf{x}_0)-\tfrac{1}{2C}(\boldsymbol{\alpha}'-\boldsymbol{\alpha})^T(\boldsymbol{\alpha}'-\boldsymbol{\alpha})\} \\
&\alpha_j,\alpha'_j\geq 0,\ j=1:r
\end{aligned}$$

Además, debido a (III.37) y (III.38), se verifica que $\alpha_j\alpha'_j=0$, con lo que, si se define $\boldsymbol{\beta}=\boldsymbol{\alpha}'-\boldsymbol{\alpha}$, el problema anterior puede escribirse como:

$$\begin{aligned}
\max_{\boldsymbol{\beta}}\{W(\boldsymbol{\beta})=&-\frac{1}{2}\boldsymbol{\beta}^T\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k_0} \\
&-\sum_{j=1}^{r}\varepsilon_j\left|\beta_j\right|-\boldsymbol{\beta}^T\mathbf{g}(\mathbf{x}_0)-\frac{1}{2C}\boldsymbol{\beta}^T\boldsymbol{\beta}\}
\end{aligned} \qquad \text{(III.39)}$$

donde ahora $\boldsymbol{\beta}$ no posee restricciones. (Puede comprobarse que, en el caso de utilizar la norma 1 en las variables *slack*, habría que añadir las restricciones $-C\mathbf{1}\leq\boldsymbol{\beta}\leq C\mathbf{1}$).

Este problema posee cierto parecido al planteado en la formulación semiparamétrica de las SVM [Smola *et al.* 1999], sin embargo mientras el enfoque aquí presentado se plantea en el terreno del primal del kriging, aquél se plantea en el terreno del dual dando lugar a las SVM con componente paramétrica.

Ambos enfoques poseen objetivos similares: la pérdida ε-insensible que da lugar a las SVM permite dotar de sesgo al estimador, y, aquí, se pretende relajar la condición de insesgadez. Pero ambas no son equivalentes, ya que, el marco de hipótesis en el que se desarrollan las SVM semiparamétricas, supone un estimador de la función de regresión que pretende recoger toda la variabilidad de los datos mientras, aquí, sólo una parte (usualmente pequeña) de dicha variabilidad, se recoge bajo la hipótesis para la media, postulándose, el resto, bajo la estructura de dependencia del proceso [Matías y González-Manteiga 2003*b*].

En el marco de las hipótesis del kriging, la expresión general (III.26) de las SVM semiparamétricas:

$$f(\mathbf{x})=\sum_{i=1}^{n}\alpha_i k(\mathbf{x},\mathbf{x}_i)+\sum_{j=1}^{r}b_j g_j(\mathbf{x})$$

puede considerarse compuesta de un término que recoge la estructura de dependencia más otro que expresa la tendencia. Con ello, podría decirse que el problema (III.27) cuestiona, no sólo la hipótesis de la media como aquí, sino también, simultáneamente, la hipótesis sobre la función de covarianza.

Las condiciones de óptimo para $\boldsymbol{\beta}$ en (III.39), son:

$$\frac{\partial W}{\partial \boldsymbol{\beta}} = -\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G}\boldsymbol{\beta} + \mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k_0} - \mathbf{g}(\mathbf{x}_0) - \boldsymbol{\varepsilon}\cdot\text{signo}(\boldsymbol{\beta}) - \tfrac{1}{C}\boldsymbol{\beta} = 0$$

$$\Rightarrow \quad \boldsymbol{\beta} = (\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G} + \tfrac{1}{C}\mathbf{I}_r)^{-1}[\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k_0} - \mathbf{g}(\mathbf{x}_0) - \boldsymbol{\varepsilon}\cdot\text{signo}(\boldsymbol{\beta})] \tag{III.40}$$

donde $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_r)^T$, y denotamos por $\boldsymbol{\varepsilon}\cdot\text{signo}(\boldsymbol{\beta})$ el vector con componentes $\varepsilon_j\,\text{signo}(\beta_j)$, $j = 1:r$. Para $r = 1$, la ecuación puede resolverse mediante una regla sencilla. Para $r > 1$, la solución puede determinarse por ejemplo mediante la evaluación en la función objetivo de las posibles soluciones que se obtienen para los valores $\pm\varepsilon_j$ (téngase en cuenta que, en el kriging, r es usualmente pequeño).

En la expresión (III.40) puede observarse el efecto de dos acciones de regularización: una típicamente utilizada para contrarrestar el ruido ($\sigma^2\mathbf{I}_n$) como hemos visto al comparar las RRBF con el kriging, y una nueva, resultado de la relajación de la condición de insesgadez ($\frac{1}{C}\mathbf{I}_r$).

Nótese que, en todo lo anterior, se tiene un problema de optimización para cada punto $\mathbf{x}_0$ en el que se desea predecir, con lo que el vector $\boldsymbol{\beta}$ debería denotarse realmente $\boldsymbol{\beta}_0 = \boldsymbol{\beta}(\mathbf{x}_0)$ lo que hemos evitado por simplicidad. Esto es algo estándar en la derivación del kriging que no genera problemas en la práctica al existir una fórmula exacta para calcular $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}(\mathbf{x}_0)$.

Kriging Simple y Universal como Casos Particulares

Para el caso $\varepsilon_j = \varepsilon = 0$, $j = 1:r$, se tienen los siguientes casos particulares:

1. **Kriging Simple**. Si $C = 0$, es decir, si el sesgo es irrelevante para nuestros intereses, sustituyendo en la fórmula anterior se tiene $\boldsymbol{\beta} = \mathbf{0}_r$, con lo que la solución (III.36) con $(\boldsymbol{\alpha}' - \boldsymbol{\alpha}) = \boldsymbol{\beta}$ resulta:

   $$\boldsymbol{\lambda}^0 = (\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}[\mathbf{k}_0 - \mathbf{G}\boldsymbol{\beta}] = (\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k}_0$$

   que es la solución (III.21) proporcionada por el kriging simple (SK).

   Nótese que, el kriging simple resulta de la hipótesis de media cero (o media conocida que previamente se detrae del proceso) que produce insesgadez a partir de la relación $\hat{\mathsf{Y}}(\mathbf{x}_0) = \sum_{i=1}^n \lambda_i \mathsf{Y}(\mathbf{x}_i)$. Sin embargo, si se utiliza el kriging simple con un proceso de media no nula, en realidad se está ignorando el sesgo que efectivamente puede producirse, con lo que, con este uso indiscriminado, el kriging simple puede considerarse como un método que obvia el sesgo.

2. **Kriging Universal**. Si $C \to \infty$, es decir, si es prioritario conseguir un estimador insesgado, se tiene:

   $$\boldsymbol{\beta} = (\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G})^{-1}[\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k_0} - \mathbf{g}(\mathbf{x}_0)]$$

   que, como puede comprobarse aplicando en (III.16) la fórmula de inversión de una matriz particionada, coincide (ecuación (G.8), pág. 413) con el multiplicador de Lagrange del problema del kriging universal. Con ello:

   $$\begin{aligned}\boldsymbol{\lambda}^\infty &= (\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}(\mathbf{k}_0 - \mathbf{G}\boldsymbol{\beta}) \\ &= (\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\{\mathbf{k}_0 - \mathbf{G}(\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G})^{-1}[\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k_0} - \mathbf{g}(\mathbf{x}_0)]\}\end{aligned}$$

   coincide con la solución del kriging universal (UK), (ecuación (G.7), pág. 413).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Caso $r = 1$. Kriging Simple y Kriging Ordinario

Para ilustrar mejor *el continuo* de soluciones que produce el Kriging Regularizado en el *intervalo* [SK, UK], planteamos el caso en que la media del proceso se supone constante, es decir: $r = 1$, $g_1(\mathbf{x}) = 1$ y $\mathbf{G} = \mathbf{1}_n \in \mathbb{R}^n$. En este caso, se obtiene (para cualquier $\varepsilon \geq 0$) [Matías y González-Manteiga 2003*a*]:

$$\begin{aligned}\beta &= (\mathbf{1}_n^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{1}_n + \frac{1}{C})^{-1}[\mathbf{1}_n^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k_0} - g(\mathbf{x}_0) - \varepsilon\,\mathrm{signo}(\beta)] \\ &= [\mathbf{1}_n^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{1}_n + \frac{1}{C}]^{-1}[\mathbf{1}_n^T\boldsymbol{\lambda}^0 - 1 - \varepsilon\,\mathrm{signo}(\beta)] \\ &= \frac{[\sum_{i=1}^n \lambda_i^0 - 1] - \varepsilon\,\mathrm{signo}(\beta)}{\frac{1}{C} + \sum_{i,j=1}^n (k_\sigma^{-1})_{ij}} \qquad \text{(III.41)}\end{aligned}$$

donde hemos denotado $(k_\sigma^{-1})_{ij} \doteq [(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}]_{ij} =, i, j = 1 : n$.

De esta forma, si $C = 0$, sustituyendo en (III.41) se tiene $\beta = 0$ (para cualquier valor de ε) con lo que se obtiene el kriging simple. Por el contrario, si $C = \infty$, (con $\varepsilon = 0$) se tiene:

$$\beta = \frac{[\sum_{i=1}^n \lambda_i^0 - 1]}{\sum_{i,j=1}^n (k_\sigma^{-1})_{ij}}$$

cuyo efecto en el valor de $\boldsymbol{\lambda}$ a través de (III.36), es conseguir la condición $\sum_{i=1}^n \lambda_i = 1$, suficiente para la insesgadez en el kriging ordinario:

$$\begin{aligned}\sum_{i=1}^n \lambda_i^\infty &= \sum_{i=1}^n [\lambda_i^0 - \beta \sum_{j=1}^n (k_\sigma^{-1})_{ij}] = \sum_{i=1}^n \lambda_i^0 - \beta \sum_{i,j=1}^n (k_\sigma^{-1})_{ij} \\ &= \sum_{i=1}^n \lambda_i^0 - \frac{[\sum_{i=1}^n \lambda_i^0 - 1]}{\sum_{i,j=1}^n (k_\sigma^{-1})_{ij}} \sum_{i,j=1}^n (k_\sigma^{-1})_{ij} \\ &= \sum_{i=1}^n \lambda_i^0 - [\sum_{i=1}^n \lambda_i^0 - 1] = 1\end{aligned}$$

Ello permite comprender mejor la transformación que el kriging ordinario produce en el kriging simple para conseguir la insesgadez. En lo anterior puede observarse también el papel de ε a la hora de facilitar el sesgo aunque C sea infinito.

Support Vectors respecto a la Media del Proceso

Para simplificar la exposición supongamos $\varepsilon > 0$. Los puntos $\mathbf{x}_0 \in S \subset \mathbb{R}^d$ pueden clasificarse en dos grupos según su comportamiento en las restricciones de sus problemas de optimización respectivos:

1. Puntos que saturan las restricciones. Son puntos para los que:

$$g(\mathbf{x}_0) - \langle \boldsymbol{\lambda}, \mathbf{g} \rangle = 1 - \sum_{i=1}^n \lambda_i = \varepsilon + \xi_j$$

(o, alternativamente, $\langle \boldsymbol{\lambda}, \mathbf{g} \rangle - g(\mathbf{x}_0) = \sum_{i=1}^n \lambda_i - 1 = \varepsilon + \xi_j'$). Por las condiciones de complementariedad de Kuhn-Tucker, estos puntos verifican $\alpha_j = C\xi_j > 0$ (o, alternativamente,

$\alpha'_j = C\xi'_j > 0$), es decir, admiten un pequeño sesgo (por encima del umbral ε) que resulta beneficioso para la función objetivo.

2. Puntos que no saturan las restricciones. En este caso, $\alpha_j = \xi_j = 0$ y $\alpha'_j = \xi'_j = 0$, con lo que estos puntos no admiten una relajación de las hipótesis de insesgadez (mayor o igual que el umbral ε) que mejore la función objetivo.

El primer grupo de puntos actuaría a modo de **vectores soporte de la hipótesis para la media** del proceso. Sin embargo, al contrario que en las SVM, ello no incrementa la parsimonia de la solución: el kriging regularizado hereda la falta de parsimonia que sufre el kriging, al expresarse su solución en función de todos los puntos de la muestra.

Formulación Dual del Kriging Regularizado

De la misma forma que el kriging admite una formulación dual, puede obtenerse la del kriging regularizado. Por ejemplo, para el caso $\varepsilon = 0$ y utilizando (III.40), se tiene:

$$\begin{aligned}\hat{\mathsf{Y}}_0 &= \boldsymbol{\lambda}^T \mathsf{Y}^n = [\mathbf{k}_0 - \mathbf{G}\boldsymbol{\beta}]^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} \\ &= \mathbf{k}_0^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} [\mathbf{I}_n - \mathbf{G}(\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G} + \tfrac{1}{C}\mathbf{I}_r)^{-1} \mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}] \mathbf{y} \\ &\quad + \mathbf{g}(\mathbf{x}_0)^T [\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G} + \tfrac{1}{C}\mathbf{I}_r]^{-1} \mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} \\ &= \sum_{i=1}^{n} c_i k(\mathbf{x}_0, \mathbf{x}_i) + \sum_{j=1}^{r} b_j g_j(\mathbf{x}_0)\end{aligned}$$

con:

$$\begin{cases} \mathbf{c} = (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \{\mathbf{I}_n - \mathbf{G}[\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G} + \tfrac{1}{C}\mathbf{I}_r]^{-1} \mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\} \mathbf{y} \\ \mathbf{b} = [\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G} + \tfrac{1}{C}\mathbf{I}_r]^{-1} \mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} \end{cases}$$

En el caso $C = 0$, estas ecuaciones producen la solución dual del kriging simple, $\mathbf{c}^0 = (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{Y}$, $\mathbf{b}^0 = \mathbf{0}_r$, y en el caso $C = \infty$, se obtiene la solución del kriging universal (ecuaciones (G.5) y (G.6), pág. 413):

$$\begin{cases} \mathbf{c}^\infty = (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \{\mathbf{I}_n - \mathbf{G}[\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\} \mathbf{y} \\ \mathbf{b}^\infty = [\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} \end{cases}$$

Para el caso de media constante, ($r = 1$), la solución dual del kriging regularizado es de la forma:

$$\hat{\mathsf{Y}}(\mathbf{x}_0) = \sum_{i=1}^{n} c_i k(\mathbf{x}_0, \mathbf{x}_i) + b \text{ con } b = \frac{\sum_{i=1}^{n} c_i^0}{\frac{1}{C} + \sum_{i=1}^{n} (k_\sigma^{-1})_{ij}}$$

en lugar de con $b = \sum_{i=1}^{n} c_i^0 / \sum_{i=1}^{n} (k_\sigma^{-1})_{ij}$ del kriging ordinario o $b = 0$ del kriging simple.

Relación entre el Kriging Regularizado y la Regresión Lineal *Ridge* (Regresión Bayesiana-Procesos Gausianos)

Sea el modelo lineal general de regresión:

$$\mathsf{Y}^n = \mathbf{G}\boldsymbol{\mu} + \mathsf{U}^n \tag{III.42}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde[99] $\mathsf{Y}^n = (\mathsf{Y}_1, ..., \mathsf{Y}_n)^T$, $\mathbf{G}$ es la matriz de diseño con $(\mathbf{G})_{ij} = g_{ij}$ observación i-ésima de la variable g_j donde g_j, $j = 1 : r$ actúan como covariables, $\boldsymbol{\mu} = (\mu_1, ..., \mu_r)^T$ es un vector desconocido y $\mathsf{U}^n = (\mathsf{U}_1, ..., \mathsf{U}_n)^T$ con $\mathrm{Var}(\mathsf{U}^n) = \mathbf{K}$ con $(\mathbf{K})_{ij} = \mathrm{Cov}(\mathsf{U}_i, \mathsf{U}_j)$.

Si el vector $\boldsymbol{\mu}$ es conocido, puede demostrarse fácilmente que el estimador que minimiza el MSE es un estimador insesgado que viene dado por:

$$\hat{\mathsf{Y}} = \hat{\mathsf{Y}}_0 = \mathbf{k}_0^T \mathbf{K}^{-1} \mathsf{Y}^n + (\mathbf{g}_0^T - \mathbf{k}_0^T \mathbf{K}^{-1} \mathbf{G}) \boldsymbol{\mu} \tag{III.43}$$

donde $\mathbf{k}_0$ es el vector de componentes $(\mathbf{k}_0)_i = \mathrm{Cov}(\mathsf{U}_0, \mathsf{U}_i)$, $i = 1 : n$.

Teniendo este estimador como referencia, [Gotway y Cressie 1993] consideran, en el caso usual en que $\boldsymbol{\mu}$ es desconocido, la familia de estimadores generada al sustituir $\boldsymbol{\mu}$ por algún estimador $\hat{\boldsymbol{\mu}}$ del mismo. En particular, cuando se utiliza el estimador de mínimos cuadrados generalizados obtenido mediante:

$$\min_{\boldsymbol{\mu}} (\mathbf{y} - \mathbf{G}\boldsymbol{\mu})^T \mathbf{K}^{-1} (\mathbf{y} - \mathbf{G}\boldsymbol{\mu}) \rightarrow \hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_{GLS} = (\mathbf{G}^T \mathbf{K}^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{K}^{-1} \mathsf{Y}^n \tag{III.44}$$

entonces, puede demostrarse fácilmente, que, el estimador resultante, coincide con el kriging universal:

$$\hat{\mathsf{Y}}_{UK} = \mathbf{k}_0^T \mathbf{K}^{-1} \mathsf{Y}^n + (\mathbf{g}_0^T - \mathbf{k}_0^T \mathbf{K}^{-1} \mathbf{G}) \hat{\boldsymbol{\mu}}_{GLS} \tag{III.45}$$

Si, en su lugar [Gotway y Cressie 1993], se considera el estimador obtenido mediante mínimos cuadrados generalizados y regularizados con parámetro regularizador τ :

$$\min_{\boldsymbol{\mu}} \{(\mathbf{y} - \mathbf{G}\boldsymbol{\mu})^T \mathbf{K}^{-1} (\mathbf{y} - \mathbf{G}\boldsymbol{\mu}) + \tau \left\|\boldsymbol{\mu}\right\|_2^2\} \rightarrow \hat{\boldsymbol{\mu}}_{RGLS} = (\mathbf{G}^T \mathbf{K}^{-1} \mathbf{G} + \tau \mathbf{I}_r)^{-1} \mathbf{G}^T \mathbf{K}^{-1} \mathsf{Y}^n \tag{III.46}$$

entonces, el estimador resultante es:

$$\begin{aligned}
\hat{\mathsf{Y}}_{RGLS} &= \mathbf{k}_0^T \mathbf{K}^{-1} \mathsf{Y}^n + (\mathbf{g}_0^T - \mathbf{k}_0^T \mathbf{K}^{-1} \mathbf{G}) \hat{\boldsymbol{\mu}}_{RGLS} \\
&= \mathbf{k}_0^T \mathbf{K}^{-1} \mathsf{Y}^n + (\mathbf{g}_0^T - \mathbf{k}_0^T \mathbf{K}^{-1} \mathbf{G}) (\mathbf{G}^T \mathbf{K}^{-1} \mathbf{G} + \tau \mathbf{I}_r)^{-1} \mathbf{G}^T \mathbf{K}^{-1} \mathsf{Y}^n \\
&= \mathbf{k}_0^T [\mathbf{K}^{-1} - \mathbf{K}^{-1} \mathbf{G} (\mathbf{G}^T \mathbf{K}^{-1} \mathbf{G} + \tau \mathbf{I}_r)^{-1} \mathbf{G}^T \mathbf{K}^{-1}] \mathbf{F} \\
&\quad + \mathbf{g}_0^T (\mathbf{G}^T \mathbf{K}^{-1} \mathbf{G} + \tau \mathbf{I}_r)^{-1} \mathbf{G}^T \mathbf{K}^{-1} \mathsf{Y}^n \\
&= [\mathbf{k}_0 - Q\boldsymbol{\beta}]^T \mathbf{K}^{-1} \mathsf{Y}^n \\
\text{con } \boldsymbol{\beta} &= (\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G} + \tau \mathbf{I}_r)^{-1} [\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{k_0} - \mathbf{g}_0]
\end{aligned}$$

que es, precisamente, el estimador proporcionado por el **kriging regularizado** para el caso particular de $\varepsilon = 0$ y $C = 1/\tau$.

Dada la relación entre la regresión regularizada (*ridge*) y la regresión bayesiana, el kriging regularizado puede verse como resultado de la aplicación del enfoque bayesiano al problema de predicción espacial. Si se consideran las hipótesis:

$$\mathsf{Y}^n / \boldsymbol{\mu} \sim \mathcal{N}(\mathbf{G}\boldsymbol{\mu}, \mathbf{K}) \text{ y } \boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}_r, \Sigma_{\boldsymbol{\mu}})$$

donde $\boldsymbol{\Sigma}_{\boldsymbol{\mu}} = \mathrm{Var}(\boldsymbol{\mu})$ y la distribución de $\boldsymbol{\mu}$ refleja la información a priori disponible sobre dicho vector, la solución del enfoque bayesiano es (ecuación (III.2) con $\boldsymbol{\Sigma} = \mathbf{K}$):

$$\hat{\boldsymbol{\mu}}_B = \mathbb{E}_{\boldsymbol{\mu}/\mathbf{y}}(\boldsymbol{\mu}) = (\mathbf{G}^T \mathbf{K}^{-1} \mathbf{G} + \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1})^{-1} \mathbf{G}^T \mathbf{K}^{-1} \mathsf{Y}^n$$

[99] Este es el marco conceptual del kriging si las observaciones Y_i se consideran tomadas en algún punto espacial $\mathbf{x}_i$ de tal forma que $\mathsf{Y}_i = \mathsf{Y}(\mathbf{x}_i)$, $i = 1 : n$.

que coincide con la proporcionada por la regresión mediante procesos gausianos y que, para el caso $\Sigma_{\boldsymbol{\mu}} = \frac{1}{\tau}\mathbf{I}_r$ produce la solución[100] (III.46).

En lo anterior, si representamos la ausencia de información a priori sobre $\boldsymbol{\mu}$ mediante $\Sigma_{\boldsymbol{\mu}}^{-1} = \mathbf{0}_{r\times r}$, entonces el estimador anterior se reduce al kriging universal. En el marco del kriging regularizado, ello equivale a $C = 1/\tau = 0$, es decir, a la no relajación de la hipótesis de insesgadez, lo que da lugar al kriging universal.

Simulaciones

Con objeto de evaluar la eficacia de la regularización en el kriging, se compararon los siguientes estimadores: kriging simple (SK), kriging regularizado (RK) con $r = 1$ ($g_1(\mathbf{x}) = 1$), kriging ordinario (OK), kriging universal (UK) con hipótesis de media lineal ($r = 3$: $g_1(\mathbf{x}) = 1$, $g_2(\mathbf{x}) = x_1$, $g_3(\mathbf{x}) = x_2$) y kriging regularizado (LRK) con igual componente de tendencia.

Se consideraron 4 procesos gausianos con diferentes hipótesis para la media:

1. Constante:
$$m(\mathbf{x}) = 100$$
2. Lineal (con marcada pendiente):
$$m(\mathbf{x}) = 100x_1 + 100x_2 + 100$$
3. Cuadrática:
$$m(\mathbf{x}) = 0{,}5x_1^2 + 0{,}5x_2^2 + 2x_1x_2 + 10$$
4. Aditiva:
$$\begin{aligned} m(\mathbf{x}) &= 100 + 0{,}25[1{,}5(1-x_1) + e^{2x_1-1}\sin(3\pi(x_1-0{,}6)^2) \\ &\quad + e^{3(x_2-0{,}5)}\sin(4\pi(x_2-0{,}9)^2)] \end{aligned}$$

Para cada uno de los procesos se generaron 200 realizaciones, en los que los cinco estimadores anteriores se compararon en diferentes circunstancias comunes a todos ellos:

1. Cada realización se compuso de 400 observaciones del proceso en otros tantos puntos elegidos uniformemente en $[0,1]\times[0,1]$. Las observaciones estuvieron sujetas a un ruido gausiano con desviación $\sigma = 1$. Con ellas, se eligieron muestras de entrenamiento de tamaño 25, 75 y 100 seleccionando los datos con igual o desigual probabilidad para obtener muestras regular o irregularmente distribuidas, respectivamente. Las observaciones restantes, (375, 325 o 300), se reservaron como muestra de test para comparar la capacidad predictiva de los estimadores en cada situación.

2. La estimación se realizó en diferentes circunstancias, iguales para todos los estimadores: bien aportando el covariograma real de varios tipos, (gausiano, esférico, exponencial, etc.), o bien estimando dicho covariograma mediante diferentes modelos.

[100] La obtención de este caso particular se debe a la utilización de $\sum_{j=1}^{r}(\xi_j^2 + \xi_j'^2)$ como estimación del cuadrado del sesgo en la función objetivo (III.33) (sin considerar los productos cruzados), en lugar del auténtico cuadrado del sesgo $[\sum_{j=1}^{r}(\xi_j + \xi_j')]^2$. El parámetro ε pretende corregir la aproximación anterior y no hemos explorado si esta otra alternativa llega a producir finalmente el caso general bayesiano con $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\mu}})$. Por otra parte, la utilización de la norma 1 en las variables *slack*, produciría un estimador diferente.

UNIVERSIDAD DE SALAMANCA DE CONSULTA

Como criterios de bondad de la predicción en la muestra de test, se utilizaron la media de los errores cuadráticos (ASR) y la desviación absoluta relativa (ARD):

$$\text{ASR} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} [f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)]^2; \ \text{ARD} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left| \frac{f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)}{f(\mathbf{x}_i)} \right|$$

Dado que los resultados obtenidos fueron muy similares para todas las pruebas realizadas, se muestran los resultados correspondientes la comparación entre el KO y el KR con procesos gausianos con covariograma gausiano con $\sigma_k = 0.3$ cuando se utilizó este covariograma *real* en la estimación y muestras de tamaño 100 y 300 para entrenamiento y test, respectivamente. La selección del parámetro C en el RK se realizó mediante el cuadrado de la media de los datos, tal como se sugirió anteriormente, y la de ε, mediante una muestra de validación independiente entre los valores: 0, 10^{-5}, 10^{-4} y 10^{-3}.

Las figuras 38 y 39 muestran, respectivamente, los diagramas de caja de las diferencias en ASR y ARD del kriging ordinario y el kriging regularizado obtenidas en cada una de las 200 realizaciones de cada proceso.

Como puede observarse en dichas figuras, cuando el estimador asume la media correcta o esta media puede representarse bastante bien mediante una media constante, la diferencia entre el kriging ordinario y su versión regularizada es aleatoria y de pequeña cuantía. Sin embargo, cuando la media está mal identificada, por ejemplo, cuando se utiliza kriging ordinario y la media es lineal, la versión regularizada produce mejoras más frecuentes. En todo caso, la discrepancia entre la verdadera media y el modelo de kriging utilizado debe ser grande, para que la mejora sea perceptible, (por ejemplo, un proceso claramente monótono, p. ej. con la media lineal considerada, contra una hipótesis de media constante).

En las distintas pruebas realizadas, se comprobó que las mejoras del kriging regularizado eran más perceptibles en ARD que en ASR, sugiriendo una mayor robustez del RK frente al OK. Debido a ello, se diseñaron nuevas pruebas en las que se incluyeron outliers artificiales en la muestra de entrenamiento, sustituyendo una proporción prefijada de las observaciones (ahora no sujetas a ruido) por valores generados mediante:

$$y_i \leftarrow \bar{y} + u_i$$

siendo $\bar{y}$ la media de las observaciones iniciales y $u_i \sim \mathcal{N}(0, 1)$. Situaciones similares a ésta, en que un porcentaje de los datos pueden no ser suficientemente fiables, pueden ser relativamente frecuentes en determinados escenarios de aplicación.

Con el mismo número de realizaciones (200) para cada tipo de proceso e iguales tamaños muestrales para entrenamiento y test que anteriormente, se realizaron dos pruebas, la primera con un $10\,\%$ de outliers y la segunda con un $20\,\%$, generados como se ha indicado antes, ambas sin ruido en los datos[101] $(\sigma = 0)$. Con ellas, se compararon el kriging ordinario y el kriging regularizado con componente constante. A la vista de los valores de epsilon seleccionados en las pruebas anteriores, se eligió $\varepsilon = 10^{-3}$ y $C = \bar{y}^2$.

Las figuras 40 y 41 muestran, respectivamente, los diagramas de caja de las diferencias en ASR y ARD del kriging ordinario y el kriging regularizado obtenidas para el caso de un $10\,\%$ de valores atípicos en los datos de entrenamiento.

[101] Esta situación puede verse como un modelo de ruido heterocedástico (el porcentaje de outliers actuando como factor de heterocedasticidad).

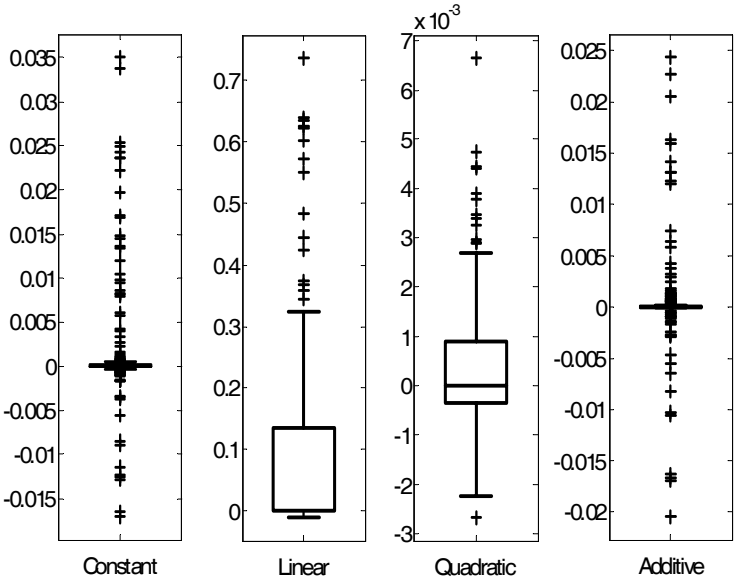


Figura 38. Diferencias $ASR_{KO} - ASR_{KR}$ sin valores atípicos en la muestra de entrenamiento para cada una de las hipótesis consideradas para la tendencia. De izquierda a derecha: constante, lineal, cuadrática y aditiva.

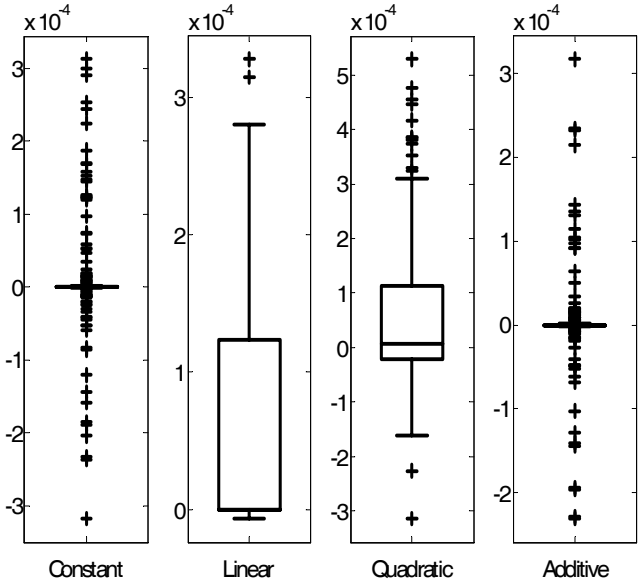


Figura 39. Diferencias $ARD_{KO} - ARD_{KR}$ sin valores atípicos en la muestra de entrenamiento para cada una de las hipótesis consideradas para la tendencia. De izquierda a derecha: constante, lineal, cuadrática y aditiva.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

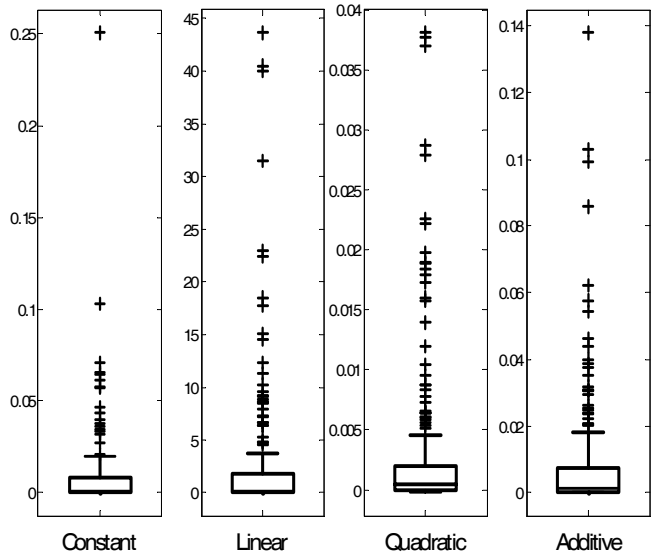


Figura 40. Diferencias $ASR_{KO} - ASR_{KR}$ con un 10 % de valores atípicos en la muestra de entrenamiento para cada una de las hipótesis consideradas para la tendencia. De izquierda a derecha: constante, lineal, cuadrática y aditiva.

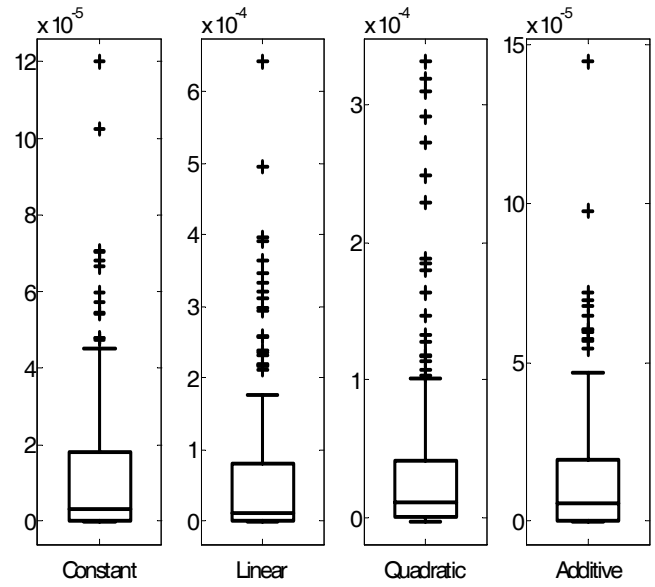


Figura 41. Diferencias $ARD_{KO} - ARD_{KR}$ con un 10 % de valores atípicos en la muestra de entrenamiento para cada una de las hipótesis consideradas para la tendencia. De izquierda a derecha: constante, lineal, cuadrática y aditiva.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Los resultados indican que el kriging regularizado produce consistentemente mejores resultados que el kriging ordinario, (y estos son mejores a medida que aumenta el porcentaje de datos atípicos en la muestra). Ello confirma la mayor robustez de RK con un ε elegido adecuadamente.

Sin embargo, como puede observarse, las diferencias son de pequeña magnitud en todos los casos. [Gotway y Cressie 1993, sec. 4.1] obtienen también pequeñas mejoras en MSE para el caso de media bien identificada y valores pequeños de $\boldsymbol{\mu}$ en consonancia con la distribución a priori bayesiana $\mathcal{N}(\mathbf{0}, \frac{1}{C}\mathbf{I}_r)$. Los resultados aquí expuestos extienden estos resultados al caso de media mal identificada y valores arbitrarios de $\boldsymbol{\mu}$ debido a la utilización de los parámetros $\varepsilon_j \geq 0$, $j = 1 : r$, para los que el kriging regularizado ya no posee equivalente bayesiano.

III.2.5 Conclusiones: Dualidad Tendencia-Dependencia y Enfoque Bayesiano

De las comparaciones realizadas anteriormente entre las redes neuronales y las técnicas kriging, se deduce que, el aspecto diferenciador más importante entre ambos grupos de técnicas, es la cantidad de variabilidad de los datos que, las distintas técnicas, modelizan bajo la componente de tendencia y la componente de dependencia.

La equivalencia entre el problema de predicción de una función aleatoria, en el que se enmarca el kriging, y el modelo lineal general (III.42):

$$\mathsf{Y}^n = \mathbf{G}\boldsymbol{\mu} + \mathsf{U}^n \tag{III.47}$$

ayuda a entender este hecho, pues permite situar el kriging en el terreno de un problema de regresión, como es el caso de las redes neuronales.

La equivalencia entre la forma dual del kriging y la forma primal de las redes RRBF produce confusión pues hace pensar que los dos modelos son conceptualmente el mismo[102].

Sin embargo, a la vista de (III.47), se deduce que, la regresión que realiza el kriging, utiliza como covariables, las componentes g_j, $j = 1 : r$ de la hipótesis de la tendencia de la función aleatoria[103] y, además, se formula en el marco de observaciones dependientes. Dichas funciones, suelen ser poco complejas (polinómicas de grado no excesivo) lo que limita su capacidad para reproducir la variabilidad de los datos, tarea que, en su mayor parte, se deja a la estructura de dependencia.

Sin embargo, en el problema de regresión en cuyo marco se formulan las redes neuronales:

$$\mathsf{Y}^n = \boldsymbol{\Psi}\mathbf{c} + \boldsymbol{\varepsilon}^n \tag{III.48}$$

con $\boldsymbol{\varepsilon}^n = (\varepsilon_1, ..., \varepsilon_n)^T$ y $(\boldsymbol{\Psi})_{ij} = \psi_j(\mathbf{x}_i)$ (en el caso de las RRBF y SVM se tiene $\psi_j(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}_j)$), las covariables son el resultado de una transformación realizada mediante funciones básicas mucho más flexibles, cuyas propiedades de aproximación universal[104] (Anexo C), les permiten recoger toda la variabilidad de los datos (salvo el ruido), siempre que la función de regresión posea las necesarias condiciones de regularidad.

---

[102] Para producir más confusión, los conceptos primal y dual poseen significados opuestos en ambas áreas (la forma primal en el kriging es su expresión en términos de las observaciones Y^n, lo que, en el ámbito de las redes neuronales se denomina forma dual [Girosi *et al.* 1995]).

[103] Expresado con propiedad: utiliza como covariables las variables transformadas mediante estas funciones.

[104] Los polinomios poseen esta propiedad pero sin restricciones de grado. El espacio generado por un conjunto de polinomios de grado menor que m, ya no posee la propiedad de aproximación univesal.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Como consecuencia, este segundo problema de regresión, se plantea con una función de tendencia a priori mucho más compleja. Debido a ello, no necesita plantearse una estructura de dependencia y se formula en el marco de observaciones independientes, $\mathrm{Var}(\boldsymbol{\varepsilon}^n) = \sigma^2 \mathbf{I}_n$.

En este contexto, la equivalencia entre las redes RRBF y el kriging, expresa que, si se plantean los problemas de regresión (III.47) y (III.48), cumpliéndose que la función de covarianza k del *ruido* $\mathsf{U}(\mathbf{x})$ determina las funciones básicas $\psi_j(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}_j)$ y viceversa, entonces las soluciones de ambos problemas son la misma.

Este hecho subyace en el análisis realizado al comparar, por ejemplo, las redes RRBF y el kriging (pág. 255), y es análogo para las SVM (y para cualquier red neuronal que se plantee cumpliendo esas condiciones).

Por tanto, la hipótesis de dependencia en los datos, es complementaria con la hipótesis de complejidad de la función de tendencia[105].

Con objeto de contrastar esta afirmación, hemos simulado un problema de predicción de una función aleatoria gausiana (regresión con datos dependientes):

$$\mathsf{F}(\mathbf{x}) = \sum_{j=1}^{3} \mu_j g_j(\mathbf{x}) + \mathsf{U}(\mathbf{x})$$

con $\mathbf{x} \in [0,1]\times[0,1]$, $g_1(\mathbf{x}) = 1$, $g_2(\mathbf{x}) = x_1$ y $g_3(\mathbf{x}) = x_2$ (media lineal) y función de covarianza k de $\mathsf{U}(\mathbf{x})$ gausiana con parámetro de escala $\sigma_k = 0.2$, en el que, la función aleatoria $\mathsf{F}(\mathbf{x})$ de interés, no es observable y, en su lugar, se observa un nuevo proceso:

$$\mathsf{Y}(\mathbf{x}) = \mathsf{F}(\mathbf{x}) + \varepsilon$$

sujeto a ruido aleatorio independiente ε de media cero y desviación $\sigma = 0.3$, independiente también del proceso $\mathsf{F}(\mathbf{x})$. Esta situación, equivale a considerar que la función de covarianza κ (ecuación (III.11)) del proceso $\mathsf{Y}(\mathbf{x})$ posee una pepita debida sólo a errores de medición ε: $\kappa(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + \sigma^2 1_{\{\mathbf{x}\neq\mathbf{x}'\}}$.

Se generaron 400 datos de $\mathsf{Y}(\mathbf{x})$ de los cuáles se extrajeron aleatoriamente $n = 100$ datos para entrenamiento dejando el resto como datos de test.

A dicho problema, hemos aplicado una red neuronal RBF entrenada mediante OLS (propia del problema (III.42)), además del kriging universal con igual componente de tendencia (lineal), con objeto de evaluar tanto su capacidad predictiva como la estructura de dependencia de sus residuos.

Los resultados se ilustran en la figura 42, donde se muestra de izquierda a derecha y de arriba a abajo: el proceso $\mathsf{F}(\mathbf{x})$, el proceso $\mathsf{Y}(\mathbf{x})$, el ajuste mediante el kriging universal y el ajuste mediante la red RBF.

Como puede observarse, la red RBF hace un trabajo perfectamente comparable con el del kriging universal a pesar de poseer un parámetro de escala constante (al igual que el kriging universal). Ello indica que la red RBF resultante recoge a través de su estimación de la función de regresión, tanto la información de tendencia (determinista) como la información de dependencia, postuladas por las hipótesis del kriging universal.

La prueba de esta afirmación se muestra en la figura 43, en la que puede observarse el covariograma muestral antes del ajuste de la RBF y el covariograma muestral prácticamente plano de

USC

[105] Esta apreciación, no tiene en cuenta las propiedades estimadoras de los algoritmos utilizados (eficiencia o tasa de convergencia) ante muestras pequeñas, lo que, en la práctica, puede marcar las diferencias entre ambos enfoques.

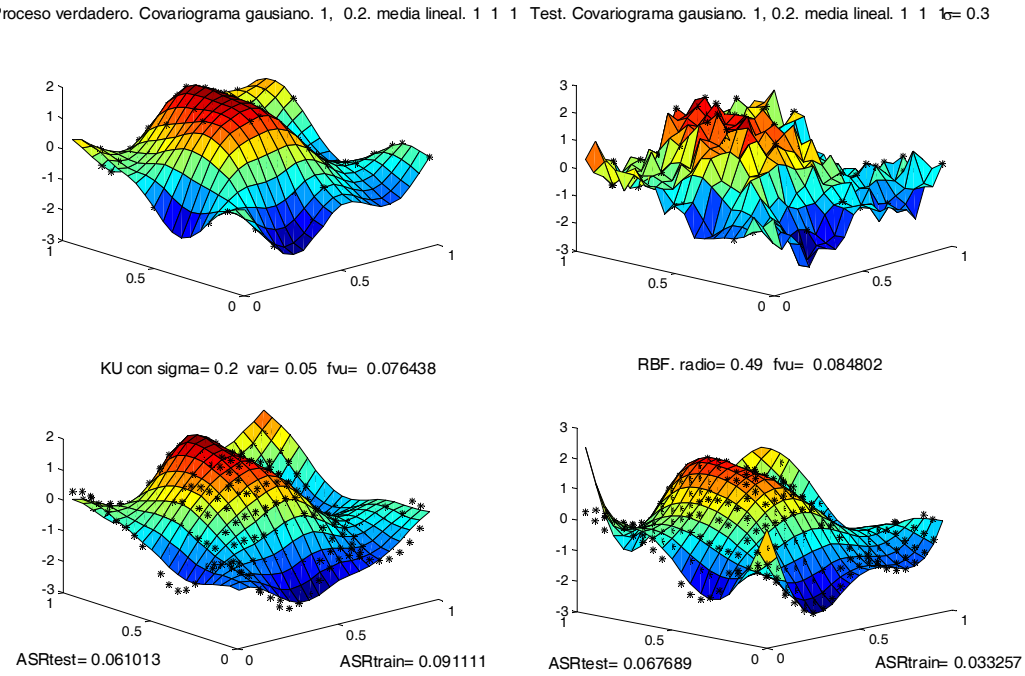


Figura 42. De izquierda a derecha y de arriba a abajo: 1) proceso $\mathsf{F}(\mathbf{x})$ gausiano con media lineal, 2) proceso $\mathsf{Y}(\mathbf{x}) = \mathsf{F}(\mathbf{x}) + \varepsilon$ con ε ruido independiente con $\mathbb{E}(\varepsilon) = 0$ y $\mathrm{Var}(\varepsilon) = 0.3^2$ independiente también de $\mathsf{F}(\mathbf{x})$, 3) ajuste mediante kriging universal, 4) ajuste mediante una red RBF con parámetro de escala constante entrenada mediante OLS.

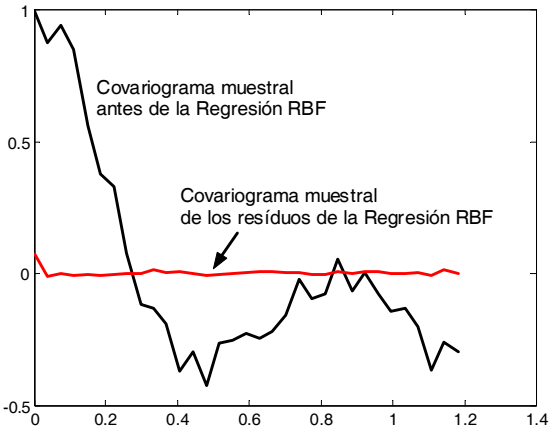


Figura 43. En relación con la figura 42, se muestra el covariograma muestral del proceso $\mathsf{Y}(\mathbf{x})$ antes del ajuste de una red RBF y el covariograma de los resíduos del ajuste (línea prácticamente horizontal). Ello muestra que la red RBF recoge prácticamente toda la información de los datos bajo la forma de una función de regresión, sin necesidad de postular dependencia.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

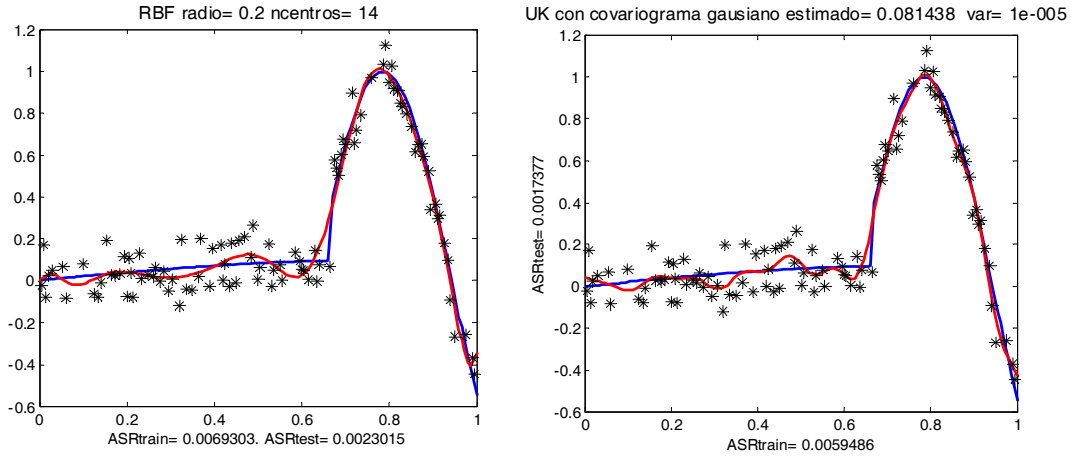


Figura 44. Estimación de la función de regresión mediante una red RBF con ventana constante entrenada con OLS (izquierda), y mediante kriging universal con componente lineal de tendencia (derecha) y covariograma estimado a partir de los datos. Se observa una mayor irregularidad del kriging debido al bajo nivel de suavizado producido por el pequeño valor estimado para la pepita $\sigma^2 = 10^{-5}$.

los residuos de dicho ajuste. Ello muestra que la red RBF recoge bajo la forma de una función de regresión, prácticamente toda la información contenida en los datos, sin necesidad de postular dependencia.

Por otra parte, un estimador lineal producido por una red neuronal entrenada mediante mínimos cuadrados (por ejemplo una red RBF entrenada mediante OLS como la anterior) $\hat{f}_n(\mathbf{x}) = \sum_{j=1}^{n} c_i g(\mathbf{x}, \mathbf{x}_j)$ (n unidades ocultas) puede verse trivialmente como un estimador lineal insesgado de mínima varianza (kriging) de una función aleatoria $\mathrm{Y}(\mathbf{x})$ con función de tendencia $\sum_{j=1}^{n} \mu_j g(\mathbf{x}, \mathbf{x}_j)$ y función de covarianza $k(\mathbf{x}, \mathbf{x}') = \sigma^2 1_{\{\mathbf{x}=\mathbf{x}'\}}$.

En este caso, el kriging universal se obtiene según (III.44) y (III.45), dando lugar a la siguiente predicción en un nuevo punto $\mathbf{x}_0$:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{GLS} &= (\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{K}^{-1}\mathrm{Y}^n = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathrm{Y}^n \Rightarrow \\ \hat{\mathrm{Y}}_{UK}(\mathbf{x}_0) &= \mathbf{k}_0^T\mathbf{K}^{-1}\mathrm{Y}^n + [\mathbf{g}_0^T - \mathbf{k}_0^T\mathbf{K}^{-1}\mathbf{G}](\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathrm{Y}^n = \mathbf{g}_0^T(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathrm{Y}^n \end{aligned}$$

con $(\mathbf{G})_{ij} = g(\mathbf{x}_i, \mathbf{x}_j)\ i, j = 1 : n$, ya que, $\mathbf{k}_0 = \mathbf{0}$. Por otra parte, en los puntos de la muestra ($\mathbf{K} = \sigma^2\mathbf{I}_n$):

$$\hat{\mathrm{Y}}_{UK}^n = \mathbf{K}\mathbf{K}^{-1}\mathrm{Y}^n + (\mathbf{G} - \mathbf{K}\mathbf{K}^{-1}\mathbf{G})(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathrm{Y}^n = \mathrm{Y}^n \qquad \text{(III.49)}$$

con lo que, en este caso, la RBF sería interpoladora.

Para cerrar el círculo de la argumentación anterior, hemos realizado el camino inverso: hemos simulado un problema de regresión con observaciones independientes y hemos utilizado una red RBF (con menos centros que datos) y el kriging universal lineal para realizar el ajuste.

El problema con éste último, era la estimación del covariograma, (teniendo en cuenta que las observaciones eran independientes). Sin embargo, si era cierta la aseveración de que la variabilidad de los datos podía recogerse mediante una función de tendencia o una estructura de dependencia, no deberíamos tener problemas para recopilar dicha variabilidad mediante una función de covarianza (aunque hubiese sido generada como tendencia).

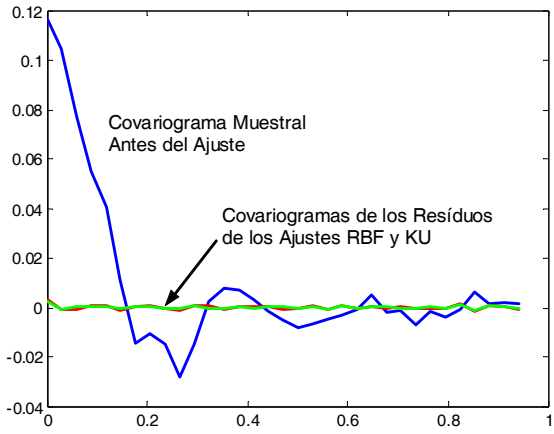


Figura 45. En relación con la figura 44, se muestra el supuesto covariograma muestral y los covariogramas de los resíduos (lineas horizontales prácticamente superpuestas), después del ajuste mediante la red RBF y mediante el kriging universal.

La figura 44 muestra el ajuste de la red RBF (izquierda) y del kriging universal lineal (derecha), donde puede observarse el comportamiento similar de ambas técnicas.

En el caso del kriging, se obtuvo, efectivamente, un supuesto covariograma muestral que reflejaba asociación, en base al cuál, se estimó el supuesto covariograma poblacional mediante mínimos cuadrados utilizando un modelo gausiano. Con él, el kriging universal lineal produjo unos resultados excelentes sobre la muestra de test.

La figura 45 muestra el supuesto covariograma muestral y el covariograma muestral de los residuos después del ajuste de la RBF y del kriging universal (líneas horizontales prácticamente superpuestas). Puede observarse el mismo efecto que en la figura 43: después del ajuste, los residuos, ahora sí, son prácticamente independientes.

Síntesis: El Enfoque Bayesiano

La causa de la estructura de dependencia observada en los datos *independientes* del ejemplo anterior, reside en la suavidad de la función de regresión: una función suave, aún sujeta a ruido independiente, provoca que los valores de la variable dependiente en puntos próximos estén de alguna forma relacionados y no puedan tomar valores totalmente desvinculados unos de otros.

La hipótesis de suavidad que suele imponerse en los modelos de regresión (continuidad, limitación de la norma de alguna derivada, etc.) significa precisamente ésto: dado un valor $\mathsf{f}(\mathbf{x})$ donde f es la función de regresión, el valor $\mathsf{f}(\mathbf{x}')$, donde $\mathbf{x}'$ es un punto próximo, está *condicionado* por aquél (tiene sentido tratar de predecir el valor de $\mathsf{f}(\mathbf{x}+\mathbf{h})$ con $\|\mathbf{h}\|$ pequeño, a partir de un valor $\mathsf{f}(\mathbf{x})$ conocido[106]) y, como consecuencia, también lo están los posibles valores de la variable aleatoria $\mathsf{Y}(\mathbf{x}') = \mathsf{f}(\mathbf{x}') + \varepsilon$ con $\mathbb{E}(\varepsilon) = 0$.

Esta *asociación* entre valores próximos de f no suele modelizarse, en el problema de regresión, mediante hipótesis estocásticas, sino que se recurre a conceptos externos a la estadística,

[106] En esta suposición descansan, precisamente, las técnicas no paramétricas y, en particular, las basadas en redes neuronales.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

tales como la suavidad de la función. Ello provoca que el analista tienda a no considerar esta *dependencia* como una asociación entre los datos.

Una forma de modelizar este hecho de forma estocástica, sin recurrir a conceptos externos, lo constituye el marco bayesiano (la regresión mediante procesos gausianos constituiría una metodología particular, entre otras posibles, de dicho enfoque). En dicho marco, los parámetros de la distribución de la variable aleatoria $\mathsf{Y}(\mathbf{x})$ son también variables aleatorias entre las cuáles cabe la posibilidad de relaciones de asociación. En concreto, el parámetro $\mathbb{E}_{\mathbf{Y}/\mathbf{x}}(\mathsf{Y}) = \mathsf{f}(\mathbf{x})$ puede modelizarse como una variable aleatoria (que ahora denotamos $\mathsf{F}(\mathbf{x})$) y, en este contexto, puede postularse una relación de dependencia entre ellos:

$$\mathrm{Cov}(\mathsf{F}(\mathbf{x}), \mathsf{F}(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}')$$

Con ello, si el ruido de los datos se postula independiente, entonces:

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[\mathsf{Y}(\mathbf{x}), \mathsf{Y}(\mathbf{x}')] = k(\mathbf{x}, \mathbf{x}') + \sigma^2 1_{\{\mathbf{x}=\mathbf{x}'\}}$$

lo que da razón teórica a los resultados experimentales anteriores. Por tanto, la estimación $\hat{\kappa}$ de la función de covarianza obtenida anteriormente en el marco de observaciones independientes, podría verse como la *asociación* implícita entre los valores $\mathsf{Y}(\mathbf{x})$ e, igualmente, $\hat{k}(\mathbf{x}, \mathbf{x}') = \hat{\kappa}(\mathbf{x}, \mathbf{x}') - \sigma^2 1_{\{\mathbf{x}=\mathbf{x}'\}}$ como la correspondiente a los valores de la función de regresión $\mathsf{F}(\mathbf{x})$.

En este marco, el grado de suavidad de la función de regresión descansa ahora en el grado de suavidad de la función de covarianza, y todo el conocimiento existente al respecto en el campo de los procesos estocásticos y la estadística espacial es aplicable aquí[107] (p. ej. [Banerjee y Gelfand 2003], [Stein 1999], [Kent 1989], [Loève 1976, sec. 34]).

En particular, ello permite interpretar las expresiones de las redes neuronales que incluyen funciones definidas positivas o condicionalmente definidas positivas[108], como términos que representan la asociación entre los datos (y no sólamente una función de tendencia compleja), lo que proporciona un soporte conceptual para la utilización, en este tipo de redes, del covariograma estimado a partir de los datos, al modo de la estadística espacial.

Este enfoque bayesiano no es nuevo en el ámbito de la estadística espacial, (kriging bayesiano [Omre 1987], [Omre y Halvorsen 1989], [Cressie 1993, sec. 3.4.4]), pero la discusión anterior muestra que, estas técnicas corren el riesgo de una muy probable redundancia entre la asociación modelizada para la tendencia y la modelizada para el proceso estacionario: los dos términos han de repartirse la información contenida en los datos.

La discusión anterior ilustra también el éxito del kriging ordinario (y la poco frecuente consideración de hipótesis complejas para la tendencia en el kriging universal[109]), cuya media constante se corresponde con una función de covarianza nula de la función (aleatoria) de tendencia. Dicha media constante no detrae información sobre la asociación existente entre los datos que, así, se recoge toda ella en la función de covarianza del proceso estacionario.

---

[107] Por ejemplo, [Stein 1999] prueba que un proceso definido en $\mathbb{R}$ es L_2-diferenciable si $d^2k(0)/dx^2$ existe y es finita (lo que no sucede por ejemplo con el modelo exponencial) y [Kent 1989] prueba que un proceso definido en $\mathbb{R}^d$ posee realizaciones continuas si k es d veces continuamente diferenciable y el resto de la expansión de Taylor de k es de orden $O(\mathbf{h}^{d+\beta})$ con $\beta > 0$.

[108] Este tipo de funciones jugaría el papel de funciones de covarianza generalizadas de la función de regresión, considerada ésta como función aleatoria intrínseca.

[109] Del mismo modo que, en las redes neuronales, se consideran estructuras de dependencia sencillas (observaciones independientes).

La complementariedad entre tendencia y asociación sugiere también la no necesidad de entrenar las redes neuronales mediante mínimos cuadrados generalizados[110]: si consideramos la forma general de la solución (III.43) del problema de predicción de una función aleatoria (regresión con datos dependientes):

$$\hat{\mathsf{Y}}_0 = \mathbf{k}_0^T \mathbf{K}^{-1} \mathsf{Y}^n + (\mathbf{g}_0^T - \mathbf{k}_0^T \mathbf{K}^{-1} \mathbf{G}) \boldsymbol{\mu} \tag{III.50}$$

donde $\boldsymbol{\mu}$ es desconocida y se estima mediante redes neuronales a través del estimador $\hat{\boldsymbol{\mu}}_{NN}$, entonces, en principio, puede postularse ausencia de asociación en los datos[111] ($\mathbf{K} = \sigma^2 \mathbf{I}_n$), con lo que se obtiene la predicción ($\mathbf{k}_0 = \mathbf{0}_n$):

$$\hat{\mathsf{Y}}_0 = \mathbf{g}_0^T \hat{\boldsymbol{\mu}}_{NN}$$

Si, por el contrario, se postulase una función de tendencia más sencilla (una función paramétrica clásica, p. ej., un polinomio) para que la estructura de dependencia poseyese algún contenido informativo y poder calcular así el estimador de tendencia mediante:

$$\hat{\boldsymbol{\mu}}_{NN} = (\mathbf{G}^T \mathbf{K}^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{K}^{-1} \mathbf{F}$$

se estaría desperdiciando la capacidad aproximadora de la red neuronal (lo que no suele producir buenos resultados) y serían más adecuados modelos paramétricos al modo del kriging.

Por último, la utilización de un estimador de red neuronal $\hat{\boldsymbol{\mu}}_{NN}$ (con NN $\in$ {RBF, MLP, SVM, etc.}) en la ecuación (III.50) determina, al modo de [Gotway y Cressie 1993], un nuevo conjunto de estimadores generales que ilustra, desde otro punto de vista, la relación entre el kriging (tendencia sencilla y estructura de asociación compleja) y las redes neuronales (tendencia compleja y nula estructura de asociación).

III.3 Estimación de la Explotabilidad en un Yacimiento de Pizarra

La predicción espacial de funciones aleatorias es de gran utilidad en disciplinas como la geología, hidrogeología, meteorología, minería y otras ciencias de los recursos naturales, (p. ej. [Taboada *et al.* 2001], [Taboada *et al.* 1997]). Este problema se ha abordado tradicionalmente utilizando técnicas tales como ([Laslett 1994], [Weber y Englund 1994], [Weber y Englund 1992], [Yakowitz y Szidarovszky 1985], [Watson 1984]): kriging, distancias inversas ponderadas (*inverse distance weighting*), polinomios interpoladores, regresión tipo núcleo, *splines* y, más recientemente, mediante redes neuronales, ([Koike *et al.* 2001], [Chakraborty *et al.* 1992]).

En esta sección ([Matías, Vaamonde, Taboada y González-Manteiga 2004*a*], [Matías, Vaamonde, Taboada y González-Manteiga 2004*b*]), aplicamos diversos algoritmos de redes neuronales a la predicción de la explotabilidad de un yacimiento de pizarra, ([Taboada *et al.* 1997], [Taboada *et al.* 1998], [Taboada *et al.* 2001]), comparándolos con el kriging.

[110] Aunque esta apreciación no es necesariamente aplicable a muestras pequeñas y no tiene en cuenta la eficacia del algoritmo en cuestión.

[111] No hacerlo supondría que la red neuronal no puede aproximar la función de tendencia lo que es falso si se trata por ejemplo de una función continua.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

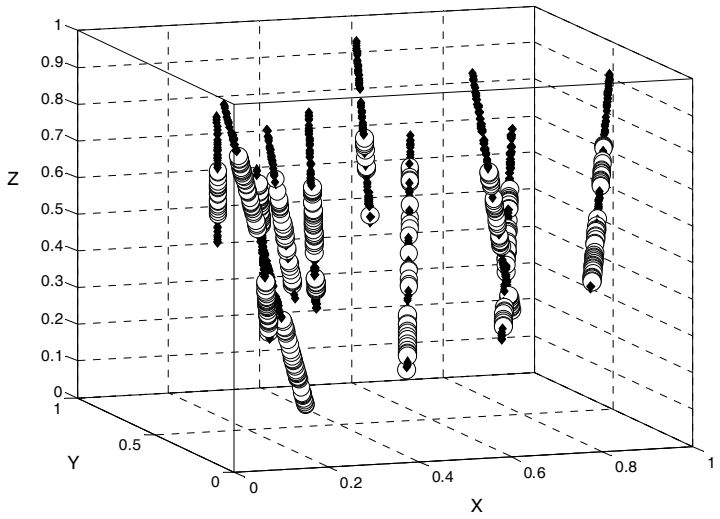


Figura 46. Muestra de entrenamiento sobre la explotabilidad de la pizarra a partir de 10 sondeos. Los círculos indican explotabilidad ($\mathsf{Y}(\mathbf{x}) = 1$). Los ejes (escalados en $[0, 1]$) representan: X Este, Y Norte, and Z, profundidad del depósito.

III.3.1 Objetivos y Metodología

Para determinar la explotabilidad, se analizó la calidad geotécnica del macizo rocoso [García-Guinea *et al.* 1998], es decir, la posibilidad de obtener bloques de suficiente tamaño para que el proceso de elaboración de pizarra de techar sea rentable. Este condicionante es común a todas las rocas ornamentales [Taboada *et al.* 1997]. Las variables que consideramos en este caso son [Taboada *et al.* 1998]: la calidad de la roca (*rock quality designation*, RQD), fracturas, intercalaciones arenosas, alteración superficial, esquistosidad de crenulación, *kink bands* y venas de cuarzo.

Dicho análisis se realizó a sobre la base de un conjunto de diez sondeos de testigo continuo, (figura 46), que permitieron el descubrimiento de los niveles de pizarra explotable [Taboada *et al.* 2001]. Los testigos tuvieron un diámetro de 63 mm, y se testificaron cada 0.5 m, valorando las variables de calidad del macizo rocoso.

Como resultado, se dispuso de una muestra $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ con $N = 1932$ observaciones procedentes de los diez sondeos, donde $\mathsf{Y}(\mathbf{x}) \in \{0, 1\}$ indica si el mineral es $(\mathsf{Y}(\mathbf{x}) = 1)$ o no es explotable, y donde $\mathbf{x} \in \mathbb{R}^3$ identifica la posición geográfica de la observación.

Los datos de cada sondeo se dividieron en dos grupos elegidos aleatoriamente, el primero destinado a conformar la muestra de entrenamiento de tamaño $n = 1000$ y el segundo la de test de tamaño $n_{\text{test}} = 932$ que sirvió para evaluar la calidad de la estimación proporcionada por cada modelo.

En este contexto, los objetivos principales del estudio fueron los siguientes:

1. Comparar las diferentes estimaciones del depósito producidas por el kriging y diversos algoritmos de redes neuronales, tanto en términos de su comportamiento en la muestra de test, como en términos de morfología.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2. Comparar las estimaciones de la pepita σ^2 en el kriging y del regularizador λ en las redes RRBF, dada su equivalencia, y teniendo en cuenta los diferentes métodos de estimación utilizados en cada caso: estimación paramétrica de la estructura de dependencia en el caso del kriging, y validación cruzada en el caso de las RRBF.

3. Evaluar los beneficios de la utilización, en las arquitecturas RBF, de núcleos estimados mediante técnicas de estadística espacial.

Dado el carácter discreto de la variable $\mathsf{Y}(\mathbf{x}) \in \{0,1\}$, los estimadores $\hat{f}_n(\mathbf{x})$ obtenidos mediante las redes neuronales se utilizaron para estimar las probabilidades a posteriori:

$$\mathbb{P}[\mathsf{Y}(\mathbf{x}) = 1|\mathbf{x}]$$

de que cada observación fuese explotable. Posteriormente, se adoptó la siguiente regla de decisión:

$$\hat{\mathsf{Y}}(\mathbf{x}) = \begin{cases} 1 & \text{si } \hat{f}_n(\mathbf{x}) > \frac{1}{2} \\ 0 & \text{en otro caso} \end{cases}$$

donde $\hat{f}_n(\mathbf{x}) = \hat{\mathbb{P}}[\mathsf{Y}(\mathbf{x}) = 1|\mathbf{x}]$ es la estimación de dichas probabilidades por cada algoritmo.

Como criterio de bondad de la estimación, se utilizó el error de clasificación cometido sobre la muestra de test:

$$\text{Error Test} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} 1_{\{\hat{\mathsf{Y}}(\mathbf{x}_i) = \mathsf{Y}(\mathbf{x}_i)\}}$$

En el caso del kriging, esta misma metodología puede fundamentarse como sigue en el contexto del kriging indicador [Cressie 1993, p. 281-283]: la explotabilidad o no de la pizarra se determina evaluando 12 variables $\{\xi_i(\mathbf{x})\}_{i=1}^{12}$ que reflejan un conjunto de características del mineral, perjudiciales para la explotabilidad del mismo (la pizarra no se caracteriza por leyes u otras variables continuas típicas de otros escenarios mineros).

De dichas variables, 11 son variables discretas (muchas de ellas indicadoras), que reflejan la existencia o no de determinadas propiedades macroscópicas, y la última es una variable continua (RQD) que refleja el grado de fragilidad de la pizarra. Estas 12 variables son evaluadas conjuntamente por el experto con el fin de determinar visualmente o mediante pruebas físicas, la explotabilidad de la pizarra correspondiente.

Por tanto, puede postularse la existencia de un proceso subyacente $\mathsf{Z}(\mathbf{x})$ continuo pero desconocido, función de este conjunto de variables:

$$\mathsf{Z}(\mathbf{x}) = \varphi[\boldsymbol{\xi}(\mathbf{x})]$$

con $\boldsymbol{\xi}(\mathbf{x}) = (\xi_1(\mathbf{x}), ..., \xi_{12}(\mathbf{x}))^T$, en el cuál, la función φ reflejaría el proceso de decisión del experto dependiendo de los valores del vector $\boldsymbol{\xi}(\mathbf{x})$.

No hemos considerado conveniente abordar en este trabajo la estimación de la función φ puesto que las variables ξ_j, $j = 1:12$ anteriores, no fueron definidas específicamente para ello. Dicho problema de clasificación, se abordará en un futuro trabajo, una vez redefinidas cuidadosamente dichas variables.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En el contexto anterior, el proceso observado $\mathsf{Y}(\mathbf{x})$ se define como:

$$\mathsf{Y}(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathsf{Z}(\mathbf{x}) \leq z_0 \\ 0 & \text{si } \mathsf{Z}(\mathbf{x}) > z_0 \end{cases}$$

de tal forma que $\mathbb{E}[\mathsf{Y}(\mathbf{x})] = \mathbb{P}[\mathsf{Y}(\mathbf{x}) = 1|\mathbf{x}]$ es desconocida.

De esta forma, la aplicación del kriging indicador al proceso $\mathsf{Y}(\mathbf{x})$, produce [Cressie 1993, p. 282], estimaciones $\hat{\mathsf{Y}}(\mathbf{x})$ de las probabilidades:

$$\mathbb{P}[\mathsf{Z}(\mathbf{x}) \leq z_0|\mathsf{Y}(\mathbf{x}_1), ..., \mathsf{Y}(\mathbf{x}_n)] = \mathbb{P}[\mathsf{Y}(\mathbf{x}) = 1|\mathsf{Y}(\mathbf{x}_1), ..., \mathsf{Y}(\mathbf{x}_n)]$$

donde hemos aplicado la definición de la variable $\mathsf{Y}(\mathbf{x})$. Con ello, un criterio disponible para la toma de decisiones sobre la explotabilidad, es el estimador:

$$\hat{\mathsf{Y}}(\mathbf{x}) = \hat{\mathbb{P}}[\mathsf{Y}(\mathbf{x}) = 1|\mathsf{Y}(\mathbf{x}_1), ..., \mathsf{Y}(\mathbf{x}_n)]$$

a través de la regla de decisión *plug-in*:

$$\widetilde{\mathsf{Y}}(\mathbf{x}) = \begin{cases} 1 \text{ si } \hat{\mathsf{Y}}(\mathbf{x}) \geq \frac{1}{2} \\ 0 \text{ si } \hat{\mathsf{Y}}(\mathbf{x}) < \frac{1}{2} \end{cases} \tag{III.51}$$

En este marco, si el covariograma estimado para $\mathsf{Y}(\mathbf{x})$ no posee discontinuidad en el origen (pepita), el predictor kriging $\hat{\mathsf{Y}}(\mathbf{x})$ resulta interpolador en los puntos de la muestra. Si, por el contrario, dicho covariograma estimado es discontinuo en el origen, $\hat{\mathsf{Y}}(\mathbf{x})$ podría perder dicha propiedad interpoladora, aunque todavía podría conservarla el estimador *plug-in* resultante $\widetilde{\mathsf{Y}}(\mathbf{x})$.

III.3.2 Modelización de la Variabilidad Espacial

Para estimar la estructura de dependencia y poder comparar los beneficios de una adecuada estimación de la misma, se utilizaron dos modelos, uno isotrópico, y otro anisotrópico.

Para el primero, se postuló un modelo gausiano exponencial, obteniéndose el siguiente covariograma estimado[112]:

$$\zeta(h) = \begin{cases} 0{,}23 - 0{,}104[1 - \exp(-\dfrac{h}{0{,}0083})] - 0{,}126[1 - \exp(-\dfrac{h^2}{0{,}043^2})] & \text{si } h > 0 \\ 0{,}26 & \text{si } h = 0 \end{cases} \tag{III.52}$$

que, como puede observarse, posee una discontinuidad en el origen de magnitud 0.03.

Sin embargo, la pizarra es una roca especialmente anisótropa, ya que debe su origen a un proceso de metamorfismo regional. La utilización de la pizarra para techar, se debe a la existencia de un plano de esquistosidad, que es el de máxima debilidad. En la cuenca estudiada, dicho plano coincide con el azimut N 120^o E y con una inclinación de 60^o S, siendo la dirección perpendicular a este plano la de mínima debilidad de la roca.

La Figura 47 recoge los variogramas experimentales obtenidos en las tres direcciones principales de anisotropía de la pizarra: dirección *longrain* (L1) (azimut N 120^o E, dip 0^o), dirección de buzamiento del plano de equistosidad (N 210^o, dip 60^o) y dirección perpendicular al plano de

[112] Quisiéramos expresar nuestro agradecimiento a Ángeles Saavedra, por su ayuda en la obtención de este covariograma.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

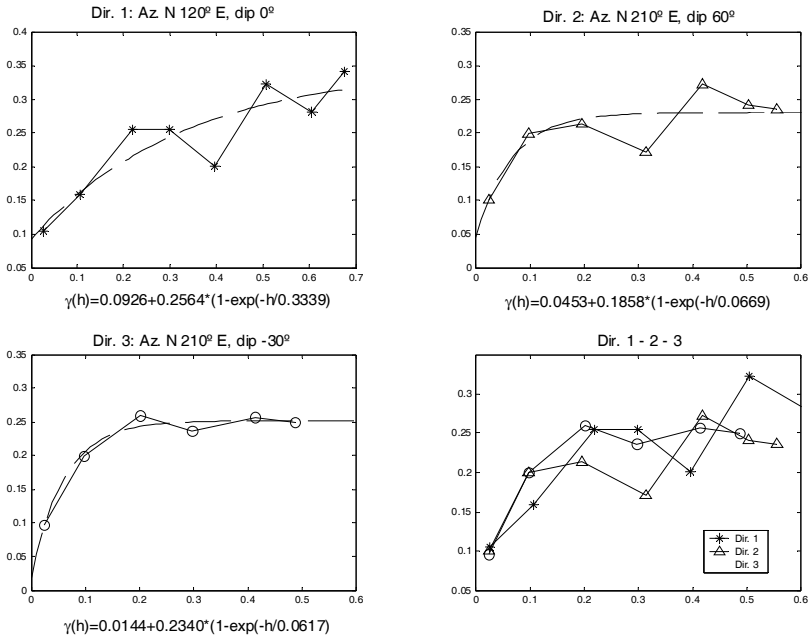


Figura 47. Variogramas experimentales en las direcciones principales de anisotropía de la pizarra. De izquierda a derecha y de arriba a abajo, semivariogramas correspondientes a las direcciones: [N 120^o E, dip 0^o], [N 210^o, dip 60^o] y [N 210^o E, dip -30^o), y representación conjunta de los mismos.

esquistosidad (N 210^o E, dip -30^o). El alcance del variograma es, como puede observarse, mayor en la dirección N-120^o-E, dirección del *longrain* (L1).

En base a estos variogramas experimentales, se consideró el siguiente modelo anidado:

$$\gamma(\mathbf{h}) = \gamma_0(|\mathbf{h}|)+\gamma_1(|\mathbf{h}_1|) \tag{III.53}$$

donde $\gamma_0(|\mathbf{h}|) = 0.07$ es el nugget effect y $\gamma_1(|\mathbf{h}_1|)$ es un modelo con anisotropía geométrica de tipo exponencial con meseta 0.25 y alcances 0.3339, 0.0669 y 0.0617 en las direcciones [N 120^o E, dip 0^o], [N 210^o, dip 60^o] y [(N 210^o E, dip -30^o], respectivamente.

III.3.3 Algoritmos Utilizados

Los algoritmos utilizados en el estudio, fueron los siguientes[113] ([Matías, Vaamonde, Taboada y González-Manteiga 2004*a*], [Matías, Taboada, Vaamonde y González-Manteiga 2004], [Matías, Vaamonde, Taboada y González-Manteiga 2004*b*]):

1. Kriging ordinario (OK) con el variograma anisotrópico estimado (III.53).

2. Kriging ordinario con el covariograma isotrópico con pepita (III.52).

[113] Se utilizó en este problema el algoritmo kriging regularizado presentado en un apartado anterior, pero sus resultados eran análogos a los del kriging. El carácter discreto de la variable respuesta es el marco más adecuado para evaluar sus posibilidades.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3. Red neuronal RRBF utilizando como núcleo la función:

$$\zeta'(h) = 0{,}23 - 0{,}104[1 - \exp(-\frac{h}{0{,}0083})] - 0{,}126[1 - \exp(-\frac{h^2}{0{,}043^2})] \qquad \text{(III.54)}$$

es decir, el covariograma isotrópico de la Ecuación (III.52) sin nugget effect $\sigma^2 = 0.03$. No se incluyó aquí el nugget effect $\sigma^2 = 0.03$, con el fin de comparar el valor de este parámetro con el del regularizador λ estimado mediante 10-*fold cross-validation* (Apartado I.5.3) que, finalmente, resultó $\lambda = 0.04$. Esta RRBF es equivalente a un kriging ordinario con el mismo covariograma.

4. Red neuronal RRBF con núcleo gausiano con parámetro de escala $\sigma_r = 0.022$ y regularizador $\lambda = 0.9$, seleccionados ambos mediante 10-*fold cross-validation*. Este modelo es equivalente al kriging ordinario con el mismo covariograma gausiano más una pepita $\sigma^2 = \lambda = 0.9$ (pág. 254).

5. Red neuronal RBF con funciones básicas gausianas de igual parámetro de escala $\sigma_r = 0.022$, entrenadas mediante OLS y utilizando el número de funciones básicas determinado por el método del espectro (tabla 10, pág. 177), seleccionando dicho parámetro de escala mediante 10-fold cross-validation.

6. Red neuronal RBF utilizando como función básica el covariograma isotrópico de la ecuación (III.52).

7. *Support Vector Machines* (SVM) para clasificación (con respuestas transformadas previamente con valores en $\{-1, +1\}$), utilizando, bien un núcleo gausiano, o bien, el covariograma isotrópico de la ecuación (III.52). Con ambas redes, se resolvió el programa cuadrático correspondiente mediante el algoritmo SVM^{light} de [Joachims 1999], seleccionando el parámetro de escala, mediante 10-*fold cross-validation*.

   Este modelo fué el único que no se utilizó para estimar las probabilidades a posteriori, pues aporta directamente la clasificación mediante la función signo .

8. *Support Vector Machines* para regresión, con término constante utlizando, bien un núcleo gausiano, o bien, el covariograma isotrópico de la ecuación (III.52). Para estas redes, se resolvió el programa cuadrático mediante el algoritmo SVMTorch [Collobert y Bengio 2001], seleccionándose el valor ε de la pérdida ε-insensible y, en su caso, el parámetro de escala del núcleo gausiano, mediante 10-*fold cross-validation*. La constante regularizadora C no resultó muy sensible sobre un amplio rango de valores y se tomó el valor $C = 1000$.

9. *Support Vector Machines* semiparamétricas para regresión, con término lineal, utlizando, bien un núcleo gausiano, o bien, el covariograma isotrópico de la ecuación (III.52). Para estas redes, se resolvió el programa cuadrático mediante el algoritmo Interior Point (Algoritmo D) de [Vanderbei 1998] y [Smola y Schölkopf 1998], obteniendo los parámetros del término semiparamétrico como subproducto de dicho algoritmo, como variables duales del programa dual. La selección del modelo (ε, el parámetro de escala del núcleo gausiano, en su caso, y la constante C) se seleccionaron igual que en el caso anterior.

10. Red neuronal MLP con 14 unidades ocultas con activación tangente hiperbólica y activación lineal en salida, utilizando entrenamiento bayesiano, (pág. 189).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

RBFPP: **LS_Boost con RBF sobre Proyecciones**
Para cada proyección $t = 1, ...$
0. Asignar valores iniciales a $\boldsymbol{\theta}_t \equiv (\mathbf{w}_t, \boldsymbol{\mu}_t, \mathbf{c}_t, \sigma_t)$
1. Calcular: $\hat{\mathbf{w}}_t = \arg \min_{\|\mathbf{w}\|=1} L_t(\mathbf{w}_t) = \sum_{i=1}^{n} [r_{i,t} - f_t(\mathbf{w}_t^T \mathbf{x}_i)]^2\}$, con $r_{i,t} = [y_i - \sum_{l<t} f_l(\mathbf{w}_l^T \mathbf{x}_i)]$
2. Definir Λ_t, conjunto de valores para σ_t. Para cada σ_t, estimar f_{t,σ_t} mediante OLS.
3. Seleccionar $f_t \in \{f_{\sigma_t} : \sigma_t \in \Lambda_t\}$ mediante un criterio analítico.
4. Repetir los pasos 1-2-3 utilizando f_t, hasta la convergencia (dos iteraciones bastan).
5. Parar cuando se alcanza un nivel de error.
6. *Back-fitting*: determinar $\mathbf{c} = \arg \min_{\boldsymbol{\alpha}} \left\|\mathbf{y} - \sum_{s=1}^{t} \alpha_s \mathbf{f}_s\right\|$ y actualizar: $f = \sum_{s=1}^{t} c_s f_s$

Tabla 24. LS-Boosting algorithm for the RBFPP network

11. Algoritmo LS-Boost (pág. 232) utilizando como hipótesis base, redes neuronales RBF gausianas aplicadas en diferentes proyecciones del espacio de entrada. Como vimos en el capítulo anterior (pág. 200), este algoritmo (tabla 24), es un algoritmo *greedy projection pursuit* (*matching pursuit* ortogonal) con *back-fitting*, en el que las redes RBF gausianas, actúan como suavizadores en cada proyección (RBFPP).

    De esta forma, el modelo posee la siguiente forma general [Matías, Vaamonde, Taboada y González-Manteiga 2004*b*]:

$$f(\mathbf{x}) = \sum_{t=1}^{m} f_t(\mathbf{w}_t^T \mathbf{x}) = \sum_{t=1}^{m} [\sum_{j_t=1}^{m_t} c_{j_t} k_{j_t}(\left|\mathbf{w}_t^T \mathbf{x} - \mu_{j_t}\right|) + c_{t0}]$$

Este último algoritmo, se concibió en una segunda fase del estudio, con objeto de disponer de una nueva técnica de tipo proyección, además de la red MLP, que permitiese obtener nuevas propuestas morfológicas para el yacimiento, (véase el siguiente epígrafe).

Como puede observarse, los modelos anteriores basados en funciones básicas gausianas o tangente hiperbólica, no utilizan información a priori ni la que pueda estar contenida en los datos, sobre la variabilidad espacial del yacimiento (en todo caso, utilizan información errónea). Los modelos que utilizan el covariograma isotrópico con pepita, utilizan dicha información sólo de forma parcial.

III.3.4 Resultados

Los resultados obtenidos se muestran en la tabla 25, en la que, para cada uno de los modelos, se muestra: el tipo de funciones básicas o hiperparámetros utilizados (**Hyperparam**), la tasa de error de clasificación en la muestra de test (**Test Err**), el número de funciones básicas (**NBF**), el número efectivo de parámetros (**ENP**) y el carácter interpolador del estimador *plug-in* $\widetilde{\mathrm{Y}}(\mathbf{x})$ de la ecuación (III.51) (o del clasificador correspondiente en el caso de las SVM para clasificación).

La columna **ENP** proporciona el número de parámetros equivalentes, que da una idea del grado de suavizado de los estimadores lineales, y que se obtiene mediante la traza de la matriz *hat*. En el caso de la red MLP, el grado de suavizado se obtiene mediante la ecuación (II.120) (pág. 188).

En relación con los objetivos planteados en este estudio, sobre dichos resultados cabe realizar los siguientes comentarios:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Model	Hiperparam.	Test Err	NBF	ENP	Int.
OK, variogr. anisotrópico	ec. (III.53)	0.108	1000	1000.0	sí
OK covariogr. isotrópico	ec. (III.52)	0.117	1000	1000.0	sí
UK lineal covariogr. isotrópico	ec. (III.52)	0.114	1000	1000.0	sí
UK lineal con ruido	ec. (III.54), $\sigma = 0{,}04$	0.115	1000	643.3	sí
RRBF gausiana (OK con ruido)	$\sigma_k = 0{,}022$, $\lambda = 0{,}9$	0.121	1000	311.3	no
RRBF covariograma (OK con ruido)	ec. (III.54), $\lambda = 0{,}04$	0.114	1000	143.2	sí
RBF gausiana	$\sigma_k = 0{,}09$	0.127	163	164.0	no
RBF covariograma	ec. (III.54)	0.119	1000	1000.0	sí
MLP	14 uds. ocultas tanh	0.118	14	64.5	no
SVM clasif. – covariograma	ec. (III.54)	0.117	625	–	sí
SVM clasif. – gausiano	$\sigma_k = 0{,}007$, $C = 0{,}1$	0.130	747	–	no
SVM regr. con cte. – covariogr.	ec. (III.54), $\varepsilon = 0{,}1$	0.117	1000	–	sí
SVM regr. con cte. – gausiana	$\sigma_k = 0{,}01$, $\varepsilon = 0{,}4$	0.135	390	–	sí
SVM regr. semipar. lineal – covariogr.	ec. (III.54), $\varepsilon = 0{,}1$	0.114	1000	–	sí
SVM regr. semipar. lineal – gausiana	$\sigma_k = 0{,}007$, $\varepsilon = 0{,}1$	0.160	995	–	sí
RBF Projection Pursuit – gausiana	4 RBFs, $\sigma_k = 0{,}02$	0.119	264	262.9	no

Tabla 25. Comportamiento de los distintos algoritmos. Las columnas representan, en este orden: los hiperparámetros del modelo, el error de clasificación en la muestra de test, el número de funciones básicas, el número efectivo de parámetros, y el carácter interpolador del estimador resultante sobre la muestra de entrenamiento (explotable o no explotable).

1. En primer lugar, destacan los resultados del kriging ordinario con variograma anisotrópico, mostrando las claras ventajas de una adecuada estimación de la estructura de dependencia espacial. Las mejoras se producen con respecto al kriging sin anisotropía y, sobre todo, en relación con las redes neuronales de tipo radial cuyos núcleos son elegidos a priori sin una razón aparente.

2. En general, la utilización del covariograma como núcleo mejoró significativamente los resultados de todos los modelos de tipo radial (RBF, RRBF y SVM), aunque a costa de una menor parsimonia en los modelos resultantes.

   Esta menor parsimonia, se debe al hecho de que el covariograma isotrópico es lineal en el origen, con un radio de influencia que desciende más bruscamente que el núcleo gausiano. Con ello, además, las soluciones son menos suaves. Por otra parte, la utilización del covariograma estimado simplifica significativamente el proceso de selección del modelo, pues se estima una sóla vez para todos los modelos.

   En el caso específico de las redes RBF, los buenos resultados obtenidos con el uso del covariograma omnidireccional resultaron al principio sorprendentes dado que, todos los datos resultaron seleccionados como centros por el algoritmo OLS, lo cuál, viene normalmente acompañado de un fuerte efecto de sobreajuste. De hecho, esta red RBF con covariograma resultó interpolador de los datos, al contrario de lo que suele suceder con esta clase de redes.

   Sin embargo, con tantos centros como datos, esta RBF produce exactamente el mismo estimador que el kriging simple sin ruido en los datos:

$$\hat{\mathbf{y}} = \mathbf{K}\mathbf{c} \text{ cuando } \mathbf{c} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{y} = \mathbf{K}^{-1}\mathbf{y}$$

   donde $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \zeta(\|\mathbf{x}_i - \mathbf{x}_j\|)$, $i, j = 1 : n$ (véase lo apuntado en la pág. 276 donde ahora $\mathbf{K} = \mathbf{I}$ y $\mathbf{G} = \mathbf{K}$).

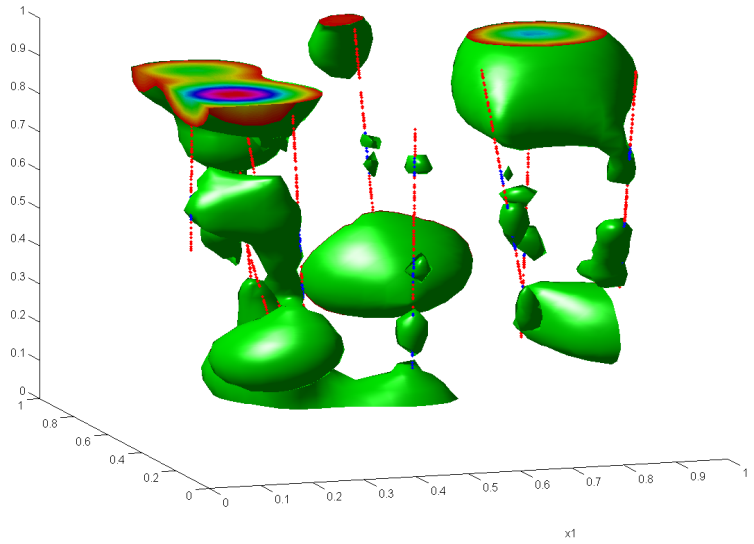


Figura 48. Estimación del yacimiento mediante una RBF con funciones básicas gausianas de radio constante.

3. La red neuronal MLP produjo resultados satisfactorios en términos de error en test, velocidad de entrenamiento y parsimonia.

   No conocemos la existencia de métodos para incorporar información a priori en este tipo de redes. Una línea de investigación interesante sería el entrenamiento de este tipo de redes fijando direcciones principales de antemano.

   Por otra parte, esta red neuronal puede quedar atrapada en un óptimo local del problema de optimización no lineal correspondiente, y si se utilizan condiciones iniciales aleatorias, cada nuevo entrenamiento puede producir una solución diferente.

4. Es interesante observar como la red RRBF con covariograma isotrópico y $\lambda = 0.04$ (equivalente al kriging isotrópico con pepita $\sigma = 0.04$) mejora los resultados del kriging con pepita $\sigma = 0.03$, sugiriendo que habíamos infravalorado inicialmente dicho parámetro. Ello indica que, en ocasiones, puede ser una buena idea contrastar la estimación original de este importante parámetro mediante técnicas a posteriori de selección del modelo como por ejemplo validación cruzada.

5. Con estos datos, los modelos SVM resultaron muy sensibles al núcleo empleado. En este sentido, la utilización del núcleo gausiano en estas redes, empeoró mucho más los resultados que en el resto de los modelos radiales. (Se utilizaron otros tipos de núcleo, como el polinómico y la tangente hiperbólica, pero producían un frecuente mal condicionamiento de los algoritmos).

6. Analizando los gráficos de explotabilidad del yacimiento, (figuras 48 a 52), pueden observarse dos tipos de reconstrucción: por una parte las de los estimadores de tipo radial basados en alguna función de distancia (núcleo o covariograma), y por otra, los modelos de tipo proyección como la MLP y la RBFPP, que producen estimadores basados en proyecciones $\mathbf{w}^T\mathbf{x}$ en el espacio de entrada.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Los primeros tienden a representar las zonas explotables como *bolsas* de mineral, mientras los segundos proponen formas más *estratificadas*.

El caso del kriging anisotrópico (figura 49), es especialmente interesante al incorporar la anisotropía en su representación del yacimiento. Podría considerarse a este tipo de estimador como un caso intermedio entre la configuración radial y la de proyección, produciendo esta última en el límite.

7. El grado de suavizado de la SVM con núcleo gausiano fué menor que el de las RBF (con igual núcleo) y las MLP, y comparable al del kriging o RN. Esto puede observarse en el grado de aprendizaje de la clases en la muestra de entrenamiento (columna **Int**. en la tabla), pero también en las reconstrucciones del yacimiento producidas: las de las SVM (figura 50) tendían a formar zonas más pequeñas de mineral que las RBF o las MLP, fragmentando en mayor medida el yacimiento, (en este sentido, sería interesante disponer de observaciones adicionales en zonas intermedias).

8. Por construcción, el kriging y las RRBF producen modelos menos parsimoniosos que los otros estimadores (salvo en el caso ya comentado en que se utiliza el covariograma en las SVM y las RBF). El grado de suavizado de los primeros puede incrementarse introduciendo una hipótesis de ruido en los datos (pepita), en cuyo caso, el número equivalente de parámetros disminuye, aunque no aumenta el grado de parsimonia.

9. Con estos datos, los modelos SVM de regresión mostraron cierta tendencia a producir un mayor número de vectores soporte que las SVM para clasificación, debido a que, para estimar $\mathbb{P}(\mathsf{Y}(\mathbf{x}) = 1|\mathbf{x})$, han de ajustarse a valores discretos ($\mathsf{Y}(\mathbf{x}) \in \{0, 1\}$) con posibles frecuentes oscilaciones en distancias cortas.

   Si realmente se persigue la parsimonia de la solución, la utilización de modelos de regresión en problemas de clasificación puede no ser la mejor estrategia. En contrapartida, la selección de ε es mucho menos problemática que en un problema usual de regresión.

10. Por último, produciendo resultados relativamente satisfactorios, el algoritmo RBFPP permite realizar una prospección más profunda del problema en determinadas direcciones de interés, con lo que resulta potencialmente más interpretable.

III.3.5 Conclusiones

De los resultados obtenidos, puede concluirse lo siguiente:

1. En primer lugar, siempre que sea posible (problemas de dimensionalidad menor o igual que 3 como es el caso de los problemas espaciales), conviene modelizar cuidadosamente la estructura de dependencia del proceso de interés y, en este sentido, las técnicas de estadística espacial pueden aportar ventajas significativas sobre las redes neuronales cuando las funciones básicas de éstas se eligen arbitrariamente.

2. En segundo lugar, resalta la diferente morfología de las estimaciones del yacimiento producidas por los estimadores comparados. La razón de estas diferencias se debe a la arquitectura específica de los estimadores cuyas funciones básicas poseen diferentes superficies de nivel: tipo *ridge* como en las redes MLP y RBFPP, o hiperesferas (hiperelipses en el caso anisotrópico) como en el caso del kriging, las RRBF, las RBF y las SVM.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

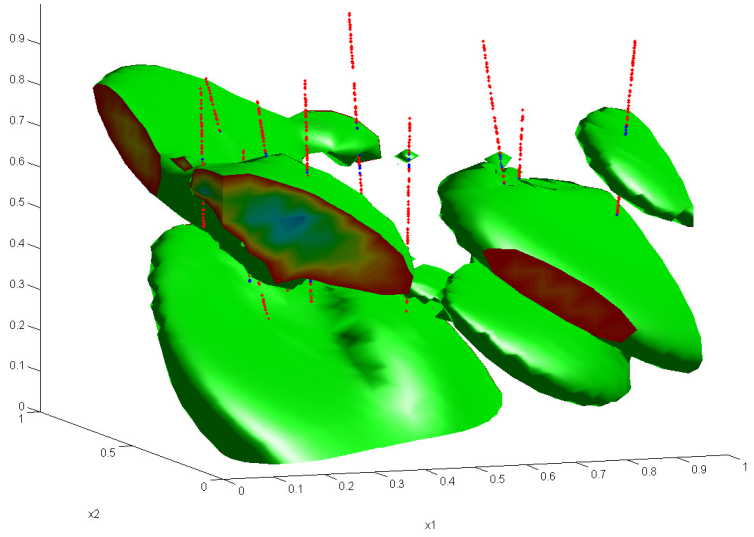


Figura 49. Estimación del yacimiento mediante el kriging ordinario con el semivariograma anisotrópico de la ecuación (III.53).

Este tipo de distinción no es nueva: véase por ejemplo [Donoho y Johnstone 1989] donde se comparan los estimadores tipo núcleo con la regresión *projection pursuit*.

La idoneidad de cada tipo de reconstrucción dependerá del problema concreto de aplicación. En nuestro caso, los modelos que mejor se ajustaron a las características del yacimiento fueron los modelos radiales, y, en particular, el que utilizó el semivariograma anisotrópico, que recogió la información relevante sobre el yacimiento.

3. El tercer aspecto de relevancia en las pruebas realizadas, fué el grado de interpolación-suavizado de los estimadores.

   En este sentido, con estos datos, la red MLP no fué capaz de interpolar la muestra sin producir sobreajuste. En varias pruebas realizadas, sin controlar la complejidad del modelo (sin entrenamiento bayesiano), el error de test se incrementaba sensiblemente. Por tanto, con estos datos, la MLP actúa como suavizador.

   Por el contrario, en el kriging y las RRBF puede conseguirse la interpolación, sin más que utilizar un regularizador $\lambda = \sigma^2 = 0$ (hipótesis de ausencia de ruido en los datos). Con estos estimadores, se ve claramente la relación grado de interpolación-número efectivo de parámetros. Sin embargo, su falta de parsimonia se mantiene en todo caso, por construcción.

   Con las redes RBF, raramente interpoladoras sin sobreajuste, la interpolación puede conseguirse con buena capacidad de generalización, mediante la utilización de un covariograma estimado a partir de los datos pues, en este caso, se transforman en kriging simple.

   Por tanto, el logro de la interpolación debe pagarse en términos grado de suavizado (la interpolación produce restricciones que sólo pueden satisfacerse con un mayor número (efectivo) de parámetros).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Finalmente, es preciso destacar dos aspectos importantes que, por sus características, nuestro problema de aplicación no pudo comparar, fueron:

1. El comportamiento de las diferentes técnicas ante un espacio de entrada de gran dimensión, en el que, previsiblemente, las SVM, las MLP y la RBFPP deberían competir entre sí y en el que, las demás técnicas radiales, (kriging, RRBF y RBF) se enfrentarían con dificultades debidas al problema de la dimensionalidad.

2. Las dificultades con que deben enfrentarse los algoritmos de entrenamiento ante mayores tamaños muestrales. Con tamaños moderados como los de nuestro problema, la implementación de los diferentes algoritmos no fué problemática, destacando la rapidez del entrenamiento bayesiano de las MLP.

Aspectos que requieren investigaciones adicionales en el futuro son, por ejemplo, la utilización en las redes MLP o RBFPP, de direcciones prefijadas como medio de incorporar información a priori al modelo, y, en las redes RBF, el uso de diferentes radios o normas del tipo $\|\mathbf{x}-\mathbf{w}_j\|_{\Sigma_j} = (\mathbf{x}-\mathbf{w}_j)^T\Sigma_j(\mathbf{x}-\mathbf{w}_j)$ en las funciones básicas (algo que no puede realizarse con el kriging-RRBF o las SVM), cuyos parámetros sea estimados durante el entrenamiento.

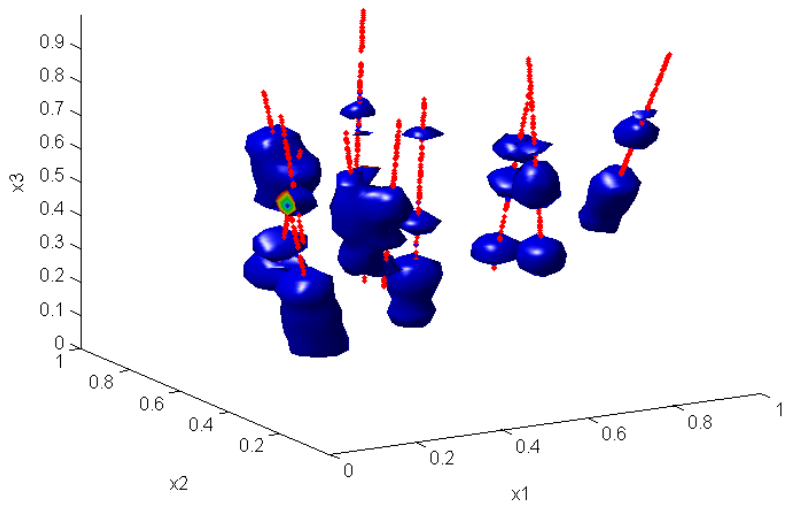


Figura 50. Estimación del yacimiento mediante una red SVM para clasificación utilizando como núcleo, el covariograma isotrópico (III.54).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

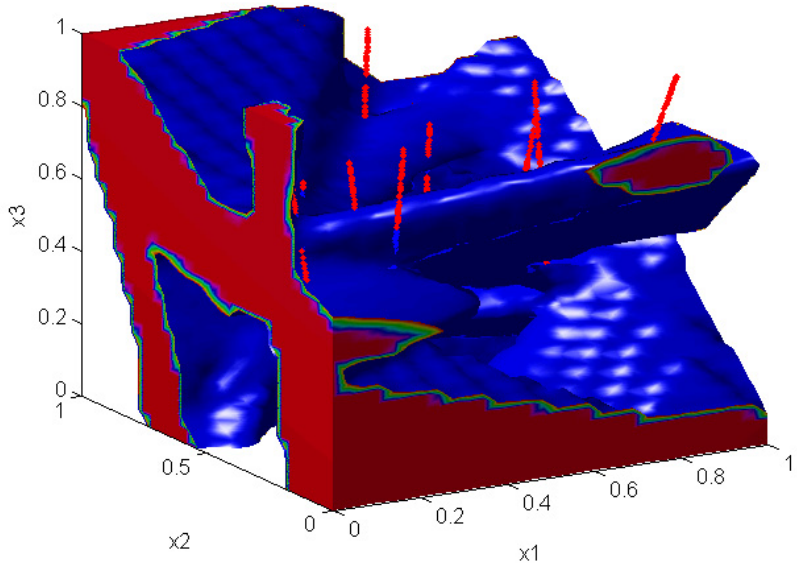


Figura 51. Estimación del yacimiento de pizarra mediante una red MLP con 14 unidades ocultas con activaciones tanh y activación lineal en el nivel de salida.

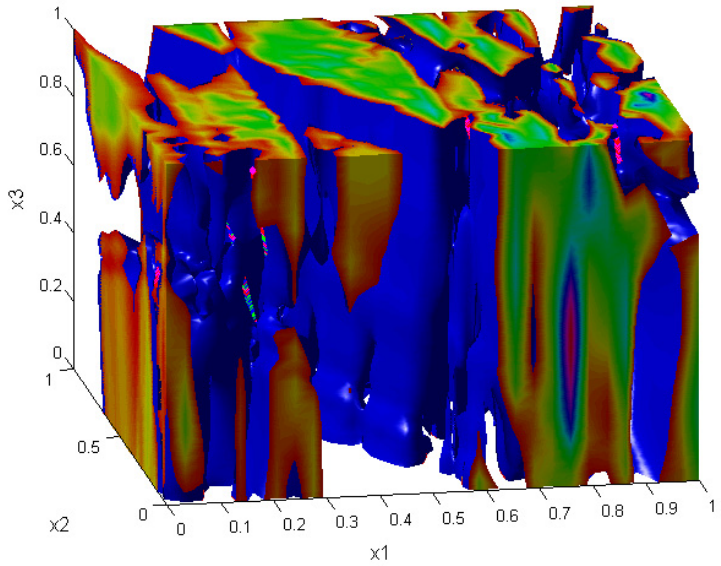


Figura 52. Estimación del yacimiento mediante el algoritmo LS_Boost utilizando hipótesis base RBF aplicadas a diferentes proyecciones del espacio de entrada (RBFPP).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

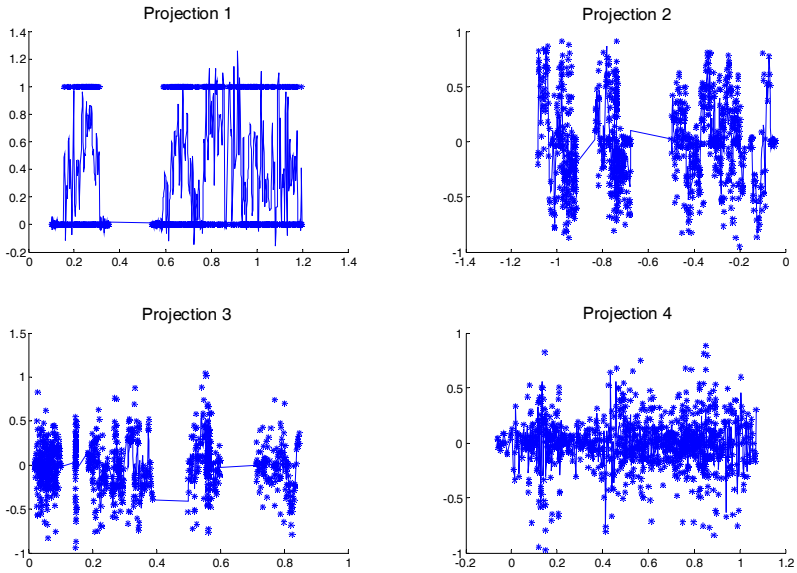


Figura 53. De arriba a abajo y de izquierda a derecha, progreso del entrenamiento del algoritmo RBFPP desde la adición de la primera RBF hasta la cuarta. Se observa como el algoritmo va extrayendo información de los resíduos y convirtiéndolos progresivamente en ruido.

III.4 Planificación de la Reforestación

III.4.1 Motivación

Cuando se elabora un plan de reforestación con el fin de reducir la fragmentación de los bosques y preservar la biodiversidad, se toma como punto de partida la situación existente. Sin embargo, cuando se aborda un proyecto de este tipo, la superficie boscosa ya se ha reducido considerablemente, lo que dota al proceso de diseño del modelo de vegetación, de un importante carácter subjetivo, basado en el criterio y la experiencia concreta de cada experto.

Con el fin de evitar esta subjetividad, algunos autores ([Guisan *et al.* 1998], [Felicísimo *et al.* 2002]), han propuesto metodologías que permiten construir modelos potenciales de vegetación teniendo en cuenta una serie de factores físicos y biológicos que influyen en la distribución de los bosques: la litología del terreno, el gradiente del mismo, la altitud o la temperatura.

III.4.2 Planteamiento del Problema y Metodología

Dichos factores, pueden ser considerados como variables estadísticas con influencia en los tipos de bosque, la cuál, puede ser modelizada a partir de la distribución actual de los bosques en un área determinada.

Un enfoque posible al respecto, es considerar el problema como un problema de clasificación, en el que los factores mencionados actúan como atributos, y el tipo de bosque es la variable respuesta. Con ello, ante un zona de terreno a reforestar, los valores de dichos atributos en la zona determinarán el tipo o tipos de bosque más viables en su contexto.

USC DE SANTIAGO DE COMPOSTELA

Z	Bosque	km^2
1	Bf. Fagus sylvatica, Blechno spicanti-Fagetum	89.6
2	Cf. Fagus sylvatica, Carici sylvaticae-Fageto)	11.9
3	Cq. Quercus rotundifolia, Cephalanthero longifoliae-Querceto petraeae	53.2
4	Liqpe. Quercus petraea, Linario triornithophorae-Querceto petraeae	44.6
5	Mf. Quercus petraea, Mercuri-alidi perennis-Fraxineto excelsioris	12.6
6	Reb. Quercus petraea, Luzulo henriquesii-Querceto petraeae	59.5

Tabla 26. Valores de la variable dependiente Z, su especificación en términos del tipo de bosque, y la extensión actual de cada tipo.

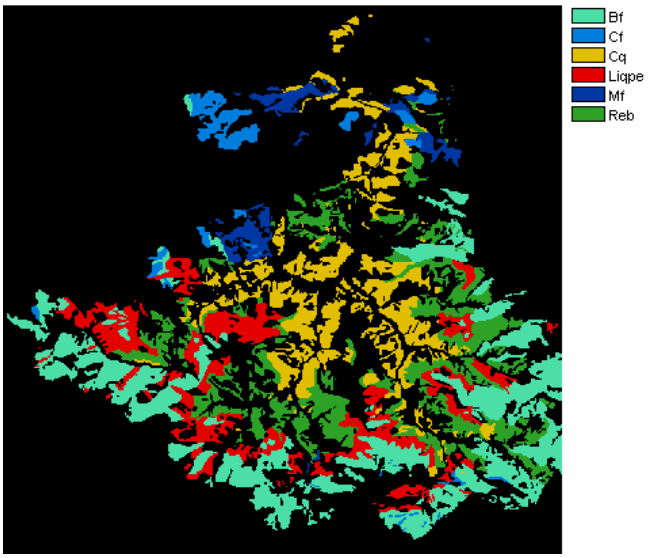


Figura 54. Distribución actual de los bosques en la comarca del Liébana.

Este enfoque, se utilizó para elaborar un modelo potencial de vegetación en el valle del Río Liébana en la comunidad autónoma de Cantabria [Ordóñez *et al.* 2003*b*]. Esta comarca, posee una superficie de 629 km^2 y, a pesar de que la zona boscosa ha disminuido considerablemente en los últimos años, en ella todavía existen áreas con bosques autóctonos suficientemente amplios como para obtener muestras representativas, y con los necesarios contrastes fitogeográficos y de topografía, como para poder desarrollar un modelo de dichas características.

Como valores de la variable dependiente $\mathsf{Z} \in \{1 : c\}$, se adoptaron los distintos tipos de bosque existentes actualmente en la comarca. Esta información fué obtenida de un mapa de vegetación con una rejilla de 50 metros por celda, que identificó 18 tipos de bosque, aunque, de éllos, sólo $c = 6$ poseían la suficiente extensión como para obtener muestras representativas con los necesarios contrastes para los atributos. La especificación de estos tipos de bosque se recoge en la tabla 26, junto con la extensión existente para cada uno, cuya distribución actual se muestra en la figura 54.

Como atributos, se identificaron las 7 variables siguientes $\mathbf{x} = (x_1, ..., x_7)$: la altitud, el gradiente del terreno, el grado potencial de insolación en los solsticios de verano y de invierno, la pluviometría, la distancia al mar y la litología. Todas estas variables son variables continuas, excepto la litología, que es cualitativa y refleja 19 posibles clases litológicas codificadas del 1 al 19.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Dado que los tipos de bosque son excluyentes, el problema de clasificación resultante se abordó mediante un enfoque *one-against-all* (pág. 89), utilizando la codificación:

$$\mathsf{Y}(\mathbf{x}) = (\mathsf{Y}_1(\mathbf{x}), ..., \mathsf{Y}_6(\mathbf{x}))^T$$
$$\text{con } \mathsf{Y}_i(\mathbf{x}) \in \{0, 1\},\ i = 1:6$$

donde $\mathsf{Y}_i(\mathbf{x}) = 1$ indica la presencia del tipo de bosque i, e $\mathsf{Y}_i(\mathbf{x}) = 0$ su ausencia, cuando se dan los valores $\mathbf{x} = (x_1, ..., x_6)^T \in \mathbb{R}^6 \times \{1:19\}$ en el vector de atributos de una zona determinada del terreno.

Con objeto de estimar el modelo, se tomó una muestra $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ con $n = 3222$ observaciones, extraídas de las 103615 celdillas en las que se había dividido la comarca, de las que $n = 2418$ se utilizaron para entrenamiento y $n_{\text{test}} = 804$, se reservaron como muestra de test (25 %).

La estimación $\hat{\mathbf{f}}(\mathbf{x}) = (\hat{f}_1(\mathbf{x}), ..., \hat{f}_6(\mathbf{x}))^T$ producida en cada combinación de atributos $\mathbf{x} = (x_1, ..., x_7)$, por cada uno de los algoritmos utilizados, se empleó para estimar las probabilidades a posteriori de los tipos de bosque:

$$\hat{\mathbb{P}}[\mathsf{Y}_i(\mathbf{x}) = 1|\mathbf{x}] = \hat{f}_i(\mathbf{x}), i = 1:6$$

y, a partir de ellas, se determinó el tipo de bosque propuesto por cada técnica (estimación de $\mathsf{Z}(\mathbf{x}) \in \{1:6\}$):

$$\hat{g}(\mathbf{x}) = \arg \max_{i \in \{1:6\}} \{\hat{f}_1(\mathbf{x}), ..., \hat{f}_6(\mathbf{x})\}$$

Dicha propuesta se aplicó posteriormente al valor de $\mathbf{x}$ en cada una de las 103615 celdillas en que se había dividido la zona, produciendo el mapa potencial de vegetación final.

III.4.3 Algoritmos Utilizados

Como puede observarse, el problema anterior posee ya una dimensionalidad respetable para las redes RBF y, sobre todo, para el kriging-RRBF cuya representación se expresa en términos de los datos de la muestra (lo mismo puede decirse de las técnicas no paramétricas (Nadaraya-Watson o regresión polinómico local).

En el caso de las RBF, en este problema, *la maldición* de la dimensionalidad no es tan intensa como para los otros algoritmos mencionados, pues el número de centros es mucho menor que el número de datos de la muestra. Sin embargo, la mayor dispersión de los datos en un espacio de mayor dimensionalidad también afecta a las RBF pues, hace necesario un mayor número de centros (número de funciones básicas). Con ello, la inversión de la matriz $(\mathbf{\Psi}^T\mathbf{\Psi})^{-1}$, donde $\mathbf{\Psi}$ es la matriz de diseño $n \times h$ con h número de funciones básicas, y su análisis espectral para la selección del modelo (varios valores del radio), también supone una carga computacional importante.

Este coste, podría pagarse, si se dispusiese de métodos consolidados para estimar la estructura de dependencia en espacios de gran dimensionalidad, lo cuál, como vimos en la sección anterior, podría aportar beneficios. Sin embargo, no conocemos de la existencia de tales técnicas, y, en cualquier caso, su desarrollo podría ser de gran utilidad.

Por tanto, los algoritmos seleccionados para este problema fueron las redes MLP y SVM.

Asimismo, con objeto de contrastar los resultados de estos algoritmos con las técnicas tradicionalmente utilizadas en este tipo de problemas, se utilizaron un modelo de análisis discriminante

lineal de Fisher (pág. 84 Apartado II.2.2, y Anexo A), y un modelo de regresión logística (adaline con activación logística, Apartado II.2.3) para varias clases (pág. 89).

El carácter categórico de la variable litología provocó que los resultados del análisis discriminante no se viesen prácticamente alterados si se eliminaba del modelo. Por ello, en lugar de dicha variable, se introdujeron en el modelo sus puntuaciones según un análisis de correspondencias ([Ordóñez *et al.* 2003*a*], [Taboada *et al.* 2002], [Greenacre 1984]).

Asimismo, con el modelo de regresión logística, se utilizó el método de método de los **pesos de evidencia**[114] ([Good 1977], [Bonham-Carter *et al.* 1988], [Agterberg *et al.* 1990]) para la variable litología, que describimos brevemente a continuación.

Este método, está basado en una versión logística de la regla de Bayes, y sirve para cuantificar la asociación existente entre variables categóricas u ordinales. Aplicado a la regresión logística, el método pretende mejorar los resultados que se obtienen cuando se utiliza una covariable nominal u ordinal en dicha regresión. El método consiste en realizar la regresión sin dicha covariable y, posteriormente, corregir (ponderar) sus resultados mediante los pesos de evidencia.

Esta metodología, fué utilizada en primer lugar por [Agterberg *et al.* 1990] (también [Agterberg y Bonham-Carter 1999]) para la investigación de depósitos minerales y,posteriormente, por [Felicísimo *et al.* 2002], para elaborar modelos potenciales de vegetación.

El **peso de evidencia** $W_i^+(j)$ del tipo de bosque i en la clase litológica j, se estima mediante la siguiente expresión[115]:

$$\hat{W}_i^+(j) = \ln \frac{\hat{O}_i(j)}{\hat{O}_i(\bullet)} = \ln \frac{\gamma_{ij}/(1-\gamma_{ij})}{\gamma_{i\bullet}/(1-\gamma_{i\bullet})}$$

donde γ_{ij} es la frecuencia relativa del tipo de bosque i con litología j, $\gamma_{i\bullet}$ es la frecuencia relativa del bosque i en toda la muestra, $\hat{O}_i(\bullet)$ es la estimación (*a priori*) de $\mathbb{P}(\mathsf{Z}=i)/[1-\mathbb{P}(\mathsf{Z}=i)]$ (*odd*) en la muestra para el tipo de bosque i, es decir, el ratio $\gamma_{i\bullet}/(1-\gamma_{i\bullet})$, y $\hat{O}_i(j)$ es el mismo concepto (*a posteriori*), es decir, dentro de la clase litológica j.

Por tanto, cuanto mayor sea la presencia de un tipo particular i de bosque, en la clase litológica j, mayor será el valor de $W_i^+(j)$. Si las probabilidades a priori y a posteriori, son iguales, el peso de evidencia es cero, indicando ausencia total de asociación entre un tipo de bosque y un tipo de litología.

Los pesos obtenidos para cada tipo de bosque, se utilizan para ponderar las estimaciones de probabilidad a posteriori resultantes de la regresión logística (sin la variable litología) para cada tipo de bosque, mediante $\omega_i \cdot \hat{\mathbb{P}}[\mathsf{Y}_i(\mathbf{x})=1|\mathbf{x}]$, $i=1:6$, donde ω_i es un factor de ponderación que depende de W_i^+, elegido por el analista. Aquí, se han elegido los mismos factores utilizados por [Felicísimo *et al.* 2002].

III.4.4 Resultados

Los resultados de la aplicación de las técnicas anteriores se muestran en la tabla 27, y los modelos potenciales de vegetación correspondientes, en la figura 55.

[114] Este método, es muy popular en el ámbito de los sistemas de información geográfica ([Bonham-Carter 1994], [Bonham-Carter y Agterberg 1999]), pero el concepto de peso de evidencia se extiende a los modelos de razonamiento bayesiano [Haddawy *et al.* 1997].

[115] Puede definirse el concepto opuesto $W_i^-(j)$ (peso de no evidencia).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Con respecto a la tabla, se muestra la matriz de errores y el **índice kappa** (*kappa index of agreement*, KIA, véase más abajo), obtenido con cada técnica.

Las columnas de la matriz de error se refieren a la distribución actual de los bosques y las filas a la distribución estimada por cada técnica. Cada elemento a_{ij} de la matriz indica el número de observaciones en las que el bosque potencial es i y el real es j. Con ello, los elementos no diagonales, se corresponden con puntos mal clasificadas. La última columna, muestra los errores de comisión (*comission errors*, EC), es decir, el porcentaje de puntos clasificados según un tipo de bosque cuando, de hecho, otro tipo de bosque está presente en la realidad. La última fila representa los errores de omisión (*errors of omission*, EO), es decir, el porcentaje de puntos que muestran un tipo de bosque en la realidad, pero han sido clasificados según un tipo diferente.

Para cada matriz de errores, el coeficiente *kappa*, se calcula mediante la ecuación, [Rosenfield y Fitzpatric-Lins 1986]:

$$\kappa = \frac{N \sum_{i=1}^{r} a_{ii} - \sum_{i=1}^{r} a_{i\bullet} \times a_{\bullet i}}{N^2 - \sum_{i=1}^{r} a_{i\bullet} \times a_{\bullet i}}$$

donde $N = 36$ es el número total de celdas de la matriz, r es el número de filas, a_{ij}, $i, j = 1 : 6$, son los elementos de dicha matriz, y $a_{i\bullet}$ y $a_{\bullet i}$, $i = 1 : 6$, son el número de observaciones en cada fija y en cada columna, respectivamente.

Por tanto, dicho coeficiente, pretende reflejar el grado de acierto de cada técnica excluyendo las posibilidades de acierto por azar:

$$\kappa = \frac{\text{tasa observada de acierto} - \text{tasa de acierto por azar}}{1 - \text{tasa de acierto por azar}}$$

y, así, vale cero si la coincidencia es aleatoria, y uno si es perfecta.

En base a estos resultados, se deduce que:

1. Los modelos de red neuronal MLP y SVM recogen en mayor medida la distribución actual de los bosques, mejorando la capacidad de representación de las técnicas clásicas, aunque frecuentemente utilizadas, en este área de aplicación.

   En ambas redes, el índice *kappa* resultó muy similar, y no se puede afirmar cuál de las dos produce mejores resultados ya que, ninguna de las dos, mejora la clasificación en todos los tipos de bosque a la vez.

   Observando la figura 55, se concluye que las redes neuronales proponen un modelo de vegetación más detallado y realista, para la reforestación.

2. En comparación con el método de regresión logística ponderada propuesto anteriormente por algunos investigadores, el método de análisis discriminante combinado con el análisis de correspondencias, proporciona un mapa de vegetación potencial más acorde con la distribución actual de los bosques.

3. En algunos bosques, los errores son bastante elevados. Ello es debido a que, una variable fundamental como la temperatura no se ha incluido en el modelo al no disponerse de datos al respecto. Por otra parte, la pluviometría, aunque si se ha tenido en cuenta, sólo ha podido estimarse de forma aproximada, por cuanto, sólo existen cuatro estaciones de control en toda la zona de estudio.

   Es muy probable que la aplicación de estos modelos de redes neuronales a zonas donde se pueda disponer de estos datos, proporcione resultados aún más satisfactorios.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Logística-Pesos Evidencia

	1	2	3	4	5	6	E.C.
1	23117	394	480	5510	35	3132	0.292
2	1870	2888	356	113	333	589	0.530
3	21	7	12004	93	209	2291	0.179
4	5185	343	709	8324	49	5116	0.578
5	333	1024	3311	312	4225	1565	0.608
6	2367	18	3668	3071	38	10572	0.464
E.O.	0.298	0.382	0.415	0.522	0.136	0.546	

Kappa = 0.4802

Discriminante-Correspondencias

	1	2	3	4	5	6	E.C.
1	24119	362	5	2957	0	1098	0.155
2	1123	3252	16	19	238	133	0.319
3	9	57	13423	319	1036	3382	0.263
4	5339	409	305	10357	56	4143	0.497
5	70	546	3213	76	3069	700	0.600
6	2275	48	3570	3695	490	13809	0.422
E.0.	0.268	0.305	0.346	0.406	0.372	0.406	

Kappa = 0.5629

MLP

	1	2	3	4	5	6	E.C.
1	24681	341	64	2172	0	1312	0.136
2	1157	3715	66	206	173	95	0.314
3	284	26	17568	443	539	6072	0.295
4	4818	96	244	12937	20	4599	0.430
5	243	450	1235	318	4068	895	0.436
6	1752	46	1355	1347	89	10292	0.308
E.O.	0.251	0.205	0.144	0.257	0.168	0.557	

Kappa = 0.6285

SVM

	1	2	3	4	5	6	E.C.
1	23403	242	251	1929	29	946	0.127
2	1980	3799	158	516	190	276	0.451
3	401	51	17048	541	413	4801	0.267
4	5494	259	339	11987	47	3901	0.456
5	102	301	898	255	4067	640	0.351
6	1555	22	1838	2195	143	18454	0.312
E.O.	0.289	0.187	0.169	0.312	0.168	0.454	

Kappa = 0.6262

Tabla 27. Matriz de errores e índice *kappa* para la regresión logística con pesos de evidencia (izquierda) y el análisis discriminante con análisis de correspondencias (derecha). Las columnas de la matriz de error se refieren a la distribución actual de los bosques y las filas a la distribución estimada por cada técnica. Cada elemento (i, j) de la matriz indica el número de puntos en los que el bosque potencial es i y el real es j. Así, los elementos no diagonales, se corresponden con puntos mal clasificados (véase el texto). De izquierda a derecha y de arriba a abajo, regresión logística con pesos de evidencia para la variable litología, análisis discriminante de Fisher con análisis de correspondencia con dicha variable, red neuronal MLP y SVM.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

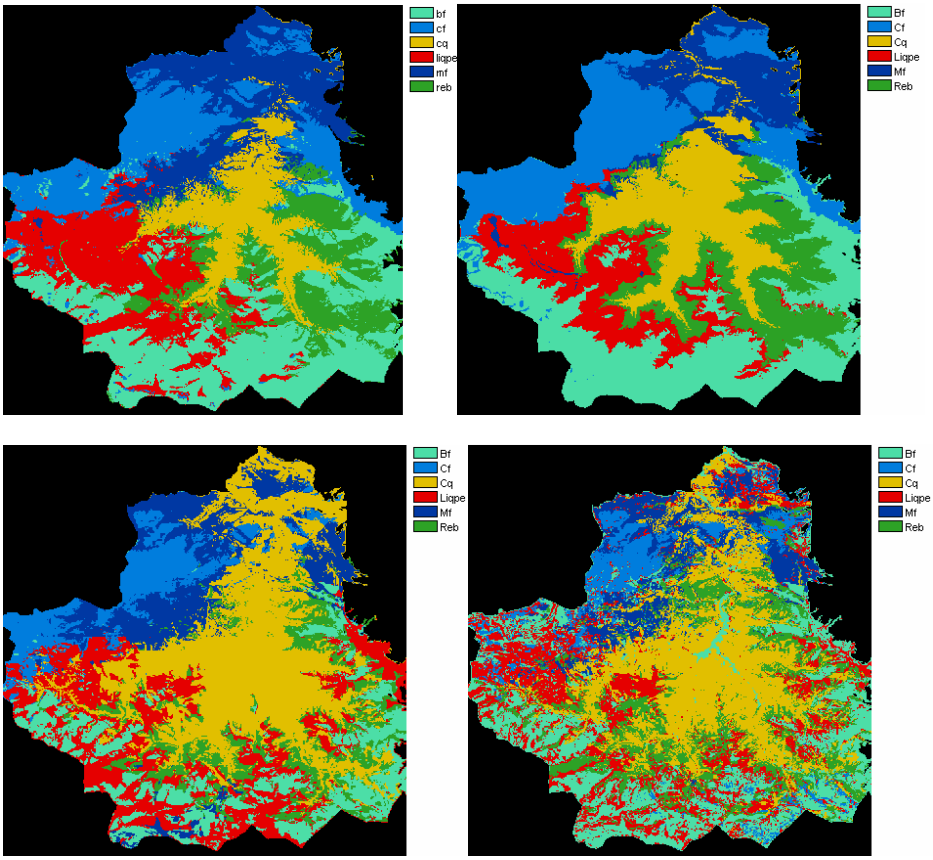


Figura 55. De izquierda a derecha y de arriba a abajo, modelos potenciales de vegetación producidos por la regresión logística con pesos de evidencia, el análisis discriminante de Fisher combinado con análisis de correspondencias, la red MLP y la red SVM.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En conclusión, los resultados anteriores muestran los beneficios de la aplicación de redes neuronales a la construcción objetiva de un modelo potencial de vegetación, que sirva de base para un plan de reforestación.

La facilidad con que las redes neuronales gestionan la posible gran dimensionalidad de estos problemas, además de su carácter no lineal, permiten recoger más fielmente la información sobre la relación entre los tipos de vegetación y los factores que la determinan, con más fidelidad que las técnicas clásicas tradicionalmente utilizadas en este ámbito de aplicación.

III.5 Nuevos Algoritmos para la Predicción de Series de Tiempo en Finanzas

Una de las cuestiones fundamentales de la economía financiera es la predictibilidad del precio de los activos financieros, [González-Manteiga *et al.* 2002]. Dado el carácter aleatorio de los precios de dichos activos, éstos se caracterizan como un proceso estocástico $\{\mathsf{Y}_t, t \in \mathbb{T}\}$ (donde, en lo que sigue, $\mathbb{T} = \mathbb{N}$ o $\mathbb{Z}$, naturales o enteros, respectivamente), para el que interesa conocer la función de densidad condicionada $\mathsf{p}_{\mathsf{Y}_t/\mathcal{A}_{t-1}}$ donde $\mathcal{A}_{t-1}$ es la σ-álgebra generada por $(\mathsf{Y}_{t-1}, \mathsf{Y}_{t-2}, ...)$, que representa toda la información disponible hasta el momento t por el mercado.

Con este objetivo, se han desarrollado gran cantidad de técnicas estadísticas, paramétricas y no paramétricas, (p. ej. [Priestley 1981], [Granger y Newbold 1986], [Box *et al.* 1994], [Granger y Teräsvirta 1993], [Ramsay y Silverman 1997], [Bosq 1998]), a las que, en la última década, se han incorporado las redes neuronales (p. ej. [Weigend y Gershenfeld 1994], [Dorffner 1996], [Refenes *et al.* 1997], [Stock y Watson 1998], [Medeiros *et al.* 2002]).

Un objetivo primordial en el uso de estas técnicas es la predicción. Las técnicas lineales tradicionales (modelos ARMA) que suponen varianza constante para el proceso $\{\mathsf{Y}_t\}$, han ido perdiendo protagonismo en la predicción de series financieras pues, éstas, se caracterizan frecuentemente por su heterocedasticidad.

Desde el punto de vista financiero [González-Manteiga *et al.* 2002], los inversores exigen una mayor compensación en términos de rentabilidad esperada cuando soportan un nivel de riesgo más elevado, por lo que variaciones en dicho nivel, pueden alterar las pautas inversoras de los individuos. Por otra parte, desde el punto de vista estadístico, los cambios en la varianza a lo largo del tiempo tienen implicaciones en la validez y eficiencia de los estimadores que suponen lo contrario.

III.5.1 Técnicas de Modelización de la Volatilidad

Con objeto de evitar los problemas mencionados, se han desarrollado nuevas técnicas de predicción que consideran diferentes hipótesis para la varianza condicionada (**volatilidad**) $\mathrm{Var}_{\mathsf{Y}_t/\mathcal{A}_{t-1}}(\mathsf{Y}_t)$, tratando de modelizar su dinámica temporal.

En general, estas técnicas pueden agruparse en dos grandes grupos, según su enfoque:

1. Las que persiguen la modelización directa de la dinámica de la volatilidad, mediante algún método parámetrico o no paramétrico. Este es el caso de las técnicas ARCH y GARCH, y sus variantes.

2. Las que persiguen la modelización de la distribución condicionada $\mathsf{p}_{\mathsf{Y}_t/\mathcal{A}_{t-1}}(y)$, a partir de la cuál se obtienen modelos dinámicos para la tendencia y para la volatilidad de los precios.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Dicha modelización se realiza usualmente mediante redes neuronales como las tratadas en el capítulo anterior.

A continuación describimos brevemente estos dos enfoques.

Técnicas GARCH

Este tipo de técnicas consideran diferentes hipótesis para la varianza condicionada $\mathrm{Var}_{\mathsf{Y}_t/\mathcal{A}_{t-1}}(\mathsf{Y}_t)$.

En concreto, los **modelos ARCH** (*autoregresive conditional heteroskedasticity process*, [Engle 1982]), combinan un modelo para la esperanza condicionada de la serie $\mathbb{E}_{\mathsf{Y}_t/\mathcal{A}_{t-1}}(\mathsf{Y}_t)$, (usualmente ARMA), con un modelo AR(s) estacionario para las varianzas condicionadas[116]:

$$\begin{cases} \mathsf{Y}_t = \mu_t + \varepsilon_t = f(\mathsf{Y}_{t-1}, \mathsf{Y}_{t-2}, ...) + \varepsilon_t \\ \varepsilon_t^2 = h_t + \omega_t = b_0 + \sum_{j=1}^{s} b_j \varepsilon_{t-j}^2 + \omega_t \end{cases}$$

donde $\mu_t = f(\mathsf{Y}_{t-1}, \mathsf{Y}_{t-2}, ...)$, $h_t = b_0 + \sum_{j=1}^{s} b_j \varepsilon_{t-j}^2$, $\{\varepsilon_t, t \in \mathbb{T}\}$ es un proceso estocástico denominado **proceso de innovación**, cuyas variables poseen media cero, y $\{\omega_t, t \in \mathbb{T}\}$ es un proceso cuyas variables verifican $\mathbb{E}_{\omega_t/\mathcal{A}_{t-1}}(\omega_t) = 0$ y $\mathrm{Var}(\omega_t) = \sigma_\omega^2$ constante $\forall t \in \mathbb{T}$. De esta forma:

$$\sigma_t^2 \doteq \mathrm{Var}_{\mathsf{Y}_t/\mathcal{A}_{t-1}}(\mathsf{Y}_t) = \mathbb{E}_{\mathsf{Y}_t/\mathcal{A}_{t-1}}[(\mathsf{Y}_t - \mu_t)^2] = \mathbb{E}_{\varepsilon_t/\mathcal{A}_{t-1}}(\varepsilon_t^2) = b_0 + \sum_{j=1}^{s} b_j \varepsilon_{t-j}^2 = h_t$$

Para que el modelo anterior para la varianza sea válido, es necesario que ε_t^2 posea valores positivos, lo que se consigue con $b_j \geq 0$, $j = 1 : s$, y $\omega_t \geq -b_0$. Además, la condición de estacionariedad exige que: $\sum_{j=1}^{s} b_j < 1$.

Una representación alternativa de los modelos ARCH, es $\varepsilon_t = \sqrt{h_t}\vartheta_t$ donde $\{\vartheta_t, t \in \mathbb{T}\}$ verifica $\mathbb{E}_{\omega_t/\mathcal{A}_{t-1}}(\vartheta_t) = 0$ y $\mathrm{Var}(\vartheta_t) = 1\ \forall t \in \mathbb{T}$, y $h_t = b_0 + \sum_{j=1}^{s} b_j \varepsilon_{t-j}^2$. Con ello, $\varepsilon_t^2 = h_t \vartheta_t^2$ o, utilizando la representación anterior, $\varepsilon_t^2 = h_t + \omega_t$, y en ambos casos, $\sigma_t^2 = \mathrm{Var}_{\varepsilon_t/\mathcal{A}_{t-1}}(\varepsilon_t) = h_t$.

Un problema que puede surgir con los modelos ARCH es que sea necesario un elevado número de retardos para representar la dinámica de la volatilidad. Para evitar este problema, [Bollerslev 1986] propuso una generalización de los mismos, los modelos GARCH(r, s), (*generalized autoregresive conditional heteroskedasticity process*) definidos como:

$$\begin{gathered} \varepsilon_t = \sqrt{h_t}\vartheta_t \\ h_t = a_0 + \sum_{j=1}^{r} a_j h_{t-j} + \sum_{j=1}^{s} b_j \varepsilon_{t-j}^2 \end{gathered}$$

pretendiendo recoger, de la varianza h_t en momentos cercanos, la información del cuadrado de las innovaciones más lejanas. Puede demostrarse, que la no negatividad de h_t está garantizada si $a_0 > 0$ y $a_i, b_j \geq 0$, $i = 1 : r$, $j = 1 : s$.

El efecto de la modelización de un modelo GARCH puede verse [González-Manteiga *et al.* 2002] en las siguientes propiedades de un modelo GARCH(1,1):

$$h_t = a_0 + a_1 h_{t-1} + b_1 \varepsilon_{t-1}^2 \text{ con } 0 \leq a_1, b_1 \leq 1 \text{ y } a_1 + b_1 < 1$$

[116] En esta sección, seguimos la notación estándar en la literatura, que representa el proceso de varianzas por $\{h_t\}$ (en minúsculas), por estar totalmente acuñada.

En primer lugar, valores altos de h_{t-1} y de ε_{t-1}^2 dan lugar a valores altos de h_t (lo que se conoce por el fenómeno de la *clusterización* de la volatilidad en las series financieras). En segundo lugar, puede demostrarse que la curtosis verifica:

$$\frac{\mathbb{E}(\varepsilon_t^4)}{[\mathbb{E}(\varepsilon_t^2)]^2} = \frac{3[(1-a_1)^2-b_1]}{(1-a_1)^2-3b_1^2} > 3 \text{ si } (1-a_1)^2 - 3b_1^2 > 0$$

lo que indica que, si se satisface esa condición, la distribución condicionada posee colas más pesadas que las de una normal.

La estimación de los modelos GARCH se realiza mediante el método de la máxima verosimilitud (con restricciones en los parámetros para garantizar la positividad de la varianza). Por ejemplo, en el caso de una distribución condicionada gausiana: $\mathsf{Y}_t/\mathcal{A}_{t-1} \sim \mathcal{N}(\mu_t, \sigma_t^2)$, el menos logaritmo de la verosimilitud es:

$$\ell(\varepsilon_t, h_t) = -\ln q(\varepsilon_t; h_t, \varepsilon_{t-1}, \varepsilon_{t-2}, ..) = -\ln[\frac{1}{h_t\sqrt{2\pi}}\exp(-\frac{1}{2h_t}\varepsilon_t^2) = \frac{1}{2}(\ln h_t + \frac{\varepsilon_t^2}{h_t} + \ln 2\pi)$$

cuyo mínimo se alcanza para:

$$0 = \frac{\partial \ell(\varepsilon_t, h_t)}{\partial h_t} = \frac{1}{2}[\frac{1}{h_t} - \frac{\varepsilon_t^2}{h_t^2}] \Leftrightarrow h_t = \varepsilon_t^2 \tag{III.55}$$

Es interesante observar que el mismo mínimo se obtiene mediante regresión por mínimos cuadrados sobre ε_t^2 (utilizando por ejemplo, $\varepsilon_{t-1}^2, \varepsilon_{t-2,}^2 ...$ como covariables), ya que: $\mathbb{E}_{\varepsilon_t^2/\mathcal{A}_{t-1}}(\varepsilon_t^2) = \sigma_t^2$:

$$\begin{aligned} \ell(\varepsilon_t, h_t) &= (\varepsilon_t^2 - h_t)^2 \Rightarrow \\ 0 &= \frac{\partial \ell(\varepsilon_t, h_t)}{\partial h_t} = 2(\varepsilon_t^2 - h_t) \Leftrightarrow h_t = \varepsilon_t^2 \end{aligned}$$

Otras alternativas son la consideración de distribuciones con colas más pesadas para ε_t, como la distribución de Student con ν grados de libertad, en cuyo caso el menos logaritmo de la verosimilitud es:

$$\ell(\varepsilon_t, h_t) = \frac{1}{2}\ln(h_t) + \frac{\nu+1}{2}\ln[1 + \frac{\varepsilon_t^2}{h_t(\nu-2)}] - \ln\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})[\pi(\nu-2)]^{1/2}}$$

Se han propuesto diferentes variaciones al modelo GARCH [González-Manteiga *et al.* 2002], desde, por ejemplo, los modelos log-GARCH(s) de [Geweke 1986], GARCH integrado (IGARCH) de [Engle y Bollerslev 1986], GARCH en media de [Engle *et al.* 1987] y NARCH (no lineal) de [Higgins y Bera 1992]:

$$h_t^\tau = a_0 + \sum_{j=1}^{r} a_j h_t^\tau + \sum_{j=1}^{s} b_j \left|\varepsilon_{t-j}\right|^{2\tau}$$

hasta los modelos *Markov-Switching*-ARCH de [Hamilton y Susmel 1992] que consideran diferentes regímenes de volatilidad según un modelo de Markov, GARCH fraccionariamente integrado (FIGARCH) de [Baillie *et al.* 1996] y los *Switching* ARCH de [Fornai y Mele 1997], éstos últimos, dirigidos a capturar efectos asimétricos y los efectos de las correlaciones positivas entre rentabilidad y volatilidad.

Otras técnicas recientes que han producido muy buenos resultados se basan en la aplicación de técnicas no lineales más flexibles o no paramétricas para la modelización de la dinámica de

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

la volatilidad, por ejemplo, [Bühlmann y McNeil 2002] que utiliza regresión polinómico local, [Dettling y Bühlmann 2001] que utilizan redes *projection pursuit* y, más recientemente, [Audrino y Bühlmann 2003] que utilizan el algoritmo Gradient_Boost aplicado a árboles de regresión.

Redes Neuronales de Mixtura para la Estimación de la Densidad Condicionada

Con el fin de estimar la densidad condicionada $\mathsf{p}_{\mathsf{Y}_t/\mathcal{A}_{t-1}}(y)$, este tipo de técnica modeliza dicha densidad como una mixtura gausiana:

$$\begin{gathered} q_{\mathsf{Y}_t/\mathcal{A}_{t-1}}(y) = \sum_{j=1}^{m} c_{t,j}\phi_{t,j}(y_t) \\ \text{con: } \phi_{t,j}(y_t) = \phi(y_t;\mu_{t,j},h_{t,j}) = \frac{1}{\sqrt{2\pi h_{t,j}}}\exp\{-\frac{(y_t-\mu_{t,j})^2}{2h_{t,j}}\} \end{gathered} \tag{III.56}$$

estimando sus parámetros mediante una **red neuronal de densidad de mixtura** (*mixture density networks* –MDN, [Bishop 1995, sec. 6.4]), la cuál, es una red[117] MLF (usualmente MLP):

$$\boldsymbol{\kappa}_t = \mathbf{f}(y_{t-1}, y_{t-2}, ..., ; \boldsymbol{\theta}_t)$$

cuya salida recoge los $3m$ parámetros de la mixtura $\boldsymbol{\kappa}_t = (c_{t,1}, \mu_{t,1}, h_{t,1}, ..., c_{t,m}, \mu_{t,m}, h_{t,m})$, de las cuáles, las correspondientes a los coeficientes $c_{t,j}$ utilizan activación *softmax* para garantizar $\sum_{j=1}^{m} c_{t,j} = 1$ de tal forma que q sea una función de densidad.

La red se entrena con cualquiera de los algoritmos vistos en el capítulo anterior, utilizando la pérdida asociada a la máxima verosimilitud (Apartado I.2.3):

$$\ell(y_t, q_t) = -\ln q(y_t;\boldsymbol{\kappa}_t) = -\ln \sum_{j=1}^{m} c_{t,j}\phi_{t,j}(y_t)$$

teniendo en cuenta que el modelo (III.56) puede aproximar cualquier función de densidad con arbitrario grado de precisión [McLachlan y Basford 1988].

Como consecuencia, la dinámica de la tendencia μ_t y de la volatilidad h_t acaban recogidas en la estimación de la red para dichos parámetros: [Schittenkopf *et al.* 1998], [Schittenkopf *et al.* 1999], [Schittenkopf *et al.* 2000], [Tino *et al.* 2001].

III.5.2 Nuevas Técnicas *Boosting* para la Modelización de la Volatilidad

En este trabajo, adoptamos el primero de los enfoques mencionados, con el fin de proponer y evaluar las posibilidades de las redes neuronales y de las técnicas *boosting* en la modelización de series de tiempo heterocedásticas.

La utilización de las redes neuronales en la predicción de series financieras no es nueva, sin embargo, salvo su aplicación a la estimación de la densidad condicionada mencionada anteriormente, dicha aplicación se ha centrado en la modelización de series de precios bajo un enfoque de regresión bien, en la modelización de comportamientos dinámicos no lineales deterministas

[117] O tres redes, cada una de las cuáles está especializada en un grupo de parámetros (medias, varianzas y coeficientes de la mixtura).

(caóticos), sujetos o no a ruido aleatorio, o bien en la modelización no lineal de procesos estocásticos autorregresivos con mayor capacidad explicativa que los modelos clásicos lineales.

Entre la gran literatura generada sobre la materia, [Weigend y Gershenfeld 1994] constituye una de las primeras referencias colectivas, ilustrativa de la problemática asociada a la utilización de redes neuronales para la predicción de series de tiempo, [Dorffner 1996] es una sucinta revisión de los diferentes modelos de red neuronal típicamente aplicados a este tipo de problemas, y [Refenes *et al.* 1997], y más recientemente [Medeiros *et al.* 2002], constituyen una rigurosa exposición de la metodología estadística aplicable.

De las distintas arquitecturas de red neuronal aplicables a la modelización de la dinámica temporal (p. ej. [Haykin 1999]), nos centramos en las de tipo *feed-forward* con activación lineal en la capa de salida, (RBF y MLP).

En un contexto de observaciones dependientes, la consistencia de estas redes ha sido probada [White 1990], [Hornik *et al.* 1994] y más recientemente, por [Chen y White 1999] para diferentes hipótesis de dependencia sobre la serie[118]. Por otra parte, en hipótesis de realizabilidad del modelo, es decir, cuando la serie es solución del modelo autorregresivo:

$$\mathsf{Y}_t = f(\mathsf{X}_t; \boldsymbol{\theta}) + \varepsilon_t \tag{III.57}$$

donde $\mathsf{X}_t = (\mathsf{Y}_{t-1}, \mathsf{Y}_{t-2}, ..., \mathsf{Y}_{t-p})^T$, $f(\mathsf{X}_t; \boldsymbol{\theta})$ es una red neuronal del tipo RBF o MLP y ε_t es por ejemplo ruido blanco, [Trapletti *et al.* 2000] demuestran la estacionariedad de la serie $\{\mathsf{Y}_t\}$, así como la consistencia y normalidad asintótica del estimador de mínimos cuadrados bajo hipótesis que aseguran la identificabilidad de los parámetros de la red [Hwang y Ding 1997].

En cuanto a las técnicas *boosting*, [Audrino y Bühlmann 2003] constituye la primera aplicación de estas técnicas, (en concreto el algoritmo Gradient_Boost) a la predicción de series de tiempo heterocedásticas. En ella, se utilizan como hipótesis base árboles de regresión y técnicas projection pursuit clásicas, considerando como pérdida el menos logaritmo de la verosimilitud gausiana.

En el marco de lo anterior, los principales aspectos novedosos de los algoritmos desarrollados en este trabajo, son:

1. La utilización de redes neuronales para la modelización de la volatilidad bajo un enfoque de regresión.

2. La modelización simultánea de la tendencia y la volatilidad por el método de la máxima verosimilitud, mediante redes neuronales.

3. La utilización del *boosting* (algoritmo Gradient_Boost) para la modelización simultánea de la tendencia y volatilidad, mediante la máxima verosimilitud generalizando el algoritmo de [Audrino y Bühlmann 2003]. Como hipótesis base para el ajuste de las dos componentes (tendencia y volatilidad) del gradiente, se utilizan:
   (a) Redes neuronales RBF y MLP, con ajuste mediante mínimos cuadrados.
   (b) Técnicas de predicción de series de tiempo heterocedásticas (ARMA-GARCH) con ajuste del gradiente mediante máxima verosimilitud (a diferencia del enfoque usual del Gradient_Boost), con objeto de capturar mejor la posible estructura heterocedástica de la serie del gradiente heredada de la serie original.

[118] Recientemente, se han generalizado gran parte de los resultados de la teoría estadística del aprendizaje al contexto de observaciones dependientes [Vidyasagar 2002].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En particular, los algoritmos desarrollados son los siguientes:

1. **RBF**(m,p) y **MLP**(m,p). Modelos ARCH basados en redes neuronales RBF y MLP con m unidades ocultas y activación lineal en salida que utilizan p retardos para estimar la volatilidad mediante un enfoque de regresión, es decir, estimando la esperanza condicionada $\mathbb{E}_{\varepsilon_t^2/\mathcal{A}_{t-1}}(\varepsilon_t^2) = \sigma_t^2$, a partir de $\boldsymbol{\tau}_t = (\varepsilon_{t-1}^2, ..., \varepsilon_{t-p}^2)^T$, utilizando pérdida cuadrática, mediante:

$$\text{MLP}(m,p) : h_t = f(\boldsymbol{\tau}_t;\boldsymbol{\theta}) = \sum_{j=1}^{m} c_j \psi(\mathbf{w}_j^T \boldsymbol{\tau}_t + w_{j0}) + c_0$$
$$\text{RBF}(m,p) : h_t = f(\boldsymbol{\tau}_t;\boldsymbol{\theta}) = \sum_{j=1}^{m} c_j \psi(|\boldsymbol{\tau}_t - \mathbf{w}_j| / w_{j0}) + c_0$$

   Si $p = 1$, ambos modelos se identifican simplemente por MLP(m) o RBF(m).

2. **RBF**(m,p) y **MLP**(m,p) **iteradas**. Son extensiones GARCH de los modelos anteriores, en las que las estimaciones iniciales $\{\hat{h}_t\}$ producidas por la red, se incorporan como covariables en las siguientes iteraciones junto con $\boldsymbol{\tau}_t = (\varepsilon_{t-1}^2, ..., \varepsilon_{t-p}^2)^T$.

3. **Nadaraya-Watson (NW)** y **Regresión Polinómico Local (LP)**. Se utilizan con la misma metodología que las redes neuronales de los puntos 1. y 2.

4. **RBF**(m,p)-**GARCH**(r,s) y **MLP**(m,p)-**GARCH**(r,s). Combinación de: a) una red neuronal, RBF o MLP, para modelizar la tendencia de la serie de precios con p retardos y m neuronas, y b) de un modelo GARCH para modelizar la dinámica de la volatilidad. En el caso $p = r = s = 1$ (la generalización a un mayor número de retardos es inmediata):

$$\begin{cases} \mathsf{Y}_t = \sum_{j=1}^{m} c_j \psi(w_j \mathsf{Y}_{t-1} + w_{j0}) + c_0 + \varepsilon_t \\ h_t = \beta h_{t-1} + \gamma \varepsilon_{t-1}^2 \end{cases} \text{ ó } \begin{cases} \mathsf{Y}_t = \sum_{j=1}^{m} c_j \psi(|\mathsf{Y}_{t-1} - w_j| / w_{j0}) + c_0 + \varepsilon_t \\ h_t = \beta h_{t-1} + \gamma \varepsilon_{t-1}^2 \end{cases}$$

   Ambos modelos se estiman conjuntamente mediante máxima verosimilitud gausiana, de forma iterativa, utilizando el algoritmo del gradiente conjugado. El algoritmo es similar al indicado a continuación, con la única diferencia de que, allí, las ecuaciones GARCH se modelizan también mediante redes neuronales.

5. **RBF**(m_1,p)-**GARCHRBF**(m_2,r,s) y **MLP**(m_1,p)-**GARCHMLP**(m_2,r,s). Algoritmos análogos a los anteriores pero, modelizando la volatilidad mediante una red RBF o MLP con m_2 neuronas, r retardos en h_t y s retardos en ε_t^2.

   Ambos modelos se estiman conjuntamente mediante máxima verosimilitud gausiana de forma iterativa, es decir, si $\{\mathsf{Y}_t\}_{t=1}^n$ es el conjunto de observaciones, se minimiza el riesgo empírico:

$$\hat{R}_n(\bar{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{t=1}^{n} \ell(\varepsilon_t, h_t; \bar{\boldsymbol{\theta}})$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Algoritmo MLP$(m_1,1)$**-GARCHMLP**$(m_2,1,1)$
1. Inicio: a. Estimar la tendencia μ_t mediante MLP$(m_1,1)$: $\hat{\mu}_t = f(y_{t-1};\boldsymbol{\theta}_1)$ b. Obtener los resíduos $e_t = y_t - f(y_{t-1};\boldsymbol{\theta}_1)$, $t = 2:n$ b. Estimar la componente de varianza $h_t = g(h_{t-1},\varepsilon^2_{t-1};\boldsymbol{\theta}_2)$ mediante MLP$(m_2,1)$, utilizando los resíduos $\{e_t\}$: $\hat{h}_t = g(h_{t-1},\varepsilon^2_{t-1};\boldsymbol{\theta}_2)$. 2. Optimizar por el método del gradiente conjugado, donde $\boldsymbol{\theta}$ es el vector de parámetros de MLP$(m_1,1)$ y MLP$(m_2,1)$: a. Inicio: $\bar{\boldsymbol{\theta}}_0 = (\boldsymbol{\theta}_1^T,\boldsymbol{\theta}_2^T)^T$ i) Calcular el menos gradiente $\boldsymbol{\xi}_0 = -\nabla_{\boldsymbol{\theta}}\hat{R}_n(\bar{\boldsymbol{\theta}}_0)$ y elegir $\boldsymbol{\zeta}_0 = \boldsymbol{\xi}_0$ ii) Calcular: $\rho_0 = \min_{\rho} \hat{R}_n(\bar{\boldsymbol{\theta}}_0 + \rho\boldsymbol{\zeta}_0)$ iii) Actualizar $\bar{\boldsymbol{\theta}}_1 = \bar{\boldsymbol{\theta}}_0 + \rho_0\boldsymbol{\zeta}_0$ b. Para $j = 1:l$ con l número de iteraciones prefijado: i) Calcular el menos gradiente $\boldsymbol{\xi}_j = -\nabla_{\boldsymbol{\theta}}\hat{R}_n(\bar{\boldsymbol{\theta}}_{j-1})$ ii) Calcular $\boldsymbol{\zeta}_j$, la dirección conjugada con $\boldsymbol{\xi}_j$ iii) Calcular: $\rho_0 = \min_{\rho} \hat{R}_n(\bar{\boldsymbol{\theta}}_{j-1} + \rho\boldsymbol{\zeta}_j)$ iv) Actualizar $\bar{\boldsymbol{\theta}}_j = \bar{\boldsymbol{\theta}}_{j-1} + \rho_0\boldsymbol{\zeta}_j$

Tabla 28. Algoritmo de entrenamiento del modelo MLP(m_1,1)-GARCHMLP(m_2,1,1) mediante la máxima verosimilitud, utilizando el método del gradiente conjugado

siendo:

$$\ell(\varepsilon_t,h_t;\bar{\boldsymbol{\theta}}) = -\ln q(\varepsilon_t,h_t;\bar{\boldsymbol{\theta}}) = \frac{1}{2}[\ln h_t(\boldsymbol{\theta}_2) + \frac{\varepsilon_t^2(\boldsymbol{\theta}_1)}{h_t(\boldsymbol{\theta}_2)} + \ln 2\pi]$$
$$= \frac{1}{2}[\ln h_t(\boldsymbol{\theta}_2) + \frac{[\mathsf{Y}_t - \mu_t(\boldsymbol{\theta}_1)]^2}{h_t(\boldsymbol{\theta}_2)} + \ln 2\pi]$$

con $\bar{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1^T,\boldsymbol{\theta}_2^T)^T$, y donde la tendencia y la volatilidad se modelizan mediante redes MLF (MLP o RBF):

$$\mu_t = \mu_t(\boldsymbol{\theta}_1) = f(\mathsf{Y}_{t-1},...,\mathsf{Y}_{t-p};\boldsymbol{\theta}_1)$$
$$h_t = h_t(\boldsymbol{\theta}_2) = g(h_{t-1},...,h_{t-r},\varepsilon^2_{t-1},...,\varepsilon^2_{t-s};\boldsymbol{\theta}_2)$$

En el caso $p = r = s = 1$, estos modelos son:

$$\begin{cases} \mathsf{Y}_t = \sum_{j=1}^{m_1} c_j\psi(w_j\mathsf{Y}_{t-1} + w_{j0}) + c_0 + \varepsilon_t \\ h_t = \sum_{j=1}^{m_2} c'_j\psi(\mathbf{w}_j^T\mathbf{h}_{t-1} + w'_{j0}) + c'_0 \end{cases} \text{ ó } \begin{cases} \mathsf{Y}_t = \sum_{j=1}^{m_1} c_j\psi(|\mathsf{Y}_{t-1} - w_j|\,/w_{j0}) + c_0 + \varepsilon_t \\ h_t = \sum_{j=1}^{m_2} c'_j\psi(|\mathbf{h}_{t-1} - \mathbf{w}_j|\,w'_{j0}) + c'_0 \end{cases}$$

según se utilicen redes MLP o RBF, y donde $\mathbf{h}_{t-1} = (h_{t-1},\varepsilon^2_{t-1})^T$. La tabla 28 muestra un esquema del algoritmo para este caso $p = r = s = 1$. Su generalización al caso de más retardos es inmediata.

6. **BoostRBF**(m,p) y **BoostMLP**(m,p). Algoritmo Gradient_Boost (pág. 232 cuyo esquema volvemos a reproducir en la tabla 29) con verosimilitud gausiana para la modelización de la volatilidad, utilizando como hipótesis base redes RBF o MLP con p retardos. En el caso $p = 1$:

$$h_t = \sum_{j=1}^{B} c_j g_j(h_{t-1},\varepsilon^2_{t-1};\boldsymbol{\theta}_j)$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Algoritmo Gradient_Boost [Friedman 2001]
0. $f_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{n} \ell(y_i, \rho)$, ρ constante
1. Para $t = 1 : B$, determinar:
$-\xi_t(\mathbf{x}_i) = -\left.\dfrac{\partial \ell(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\right\|_{f(\mathbf{x}_i)=f_{t-1}(\mathbf{x}_i)}$, $i = 1 : n$
$\mathbf{w}_t = \arg\min_{\mathbf{w},\alpha} \sum_{i=1}^{n} [-\xi_t(\mathbf{x}_i) - \alpha\psi(\mathbf{x}_i; \mathbf{w})]^2$
$\rho_t = \arg \min_{\{\rho:\rho\psi\in\mathcal{F}_\psi\}} \sum_{i=1}^{n} \ell(y_i, f_{t-1}(\mathbf{x}_i) + \rho\psi(\mathbf{x}_i; \mathbf{w}_t))$
2. Hipótesis final: $f_B(\mathbf{x}) = \sum_{t=1}^{B} \rho_t \psi(\mathbf{x}; \mathbf{w}_t)$

Tabla 29. Algoritmo Gradient Boost.

donde g_j son redes RBF o MLP y donde, a veces, se omite h_{t-1} como covariable si la primera estimación para ésta no es de calidad, pues, en ese caso, se obtienen peores resultados.

Cada iteración del algoritmo ajusta por mínimos cuadrados el menos gradiente del logaritmo de la verosimilitud:

$$\xi_t = \frac{1}{2}[\frac{\varepsilon_t^2}{h_t^2} - \frac{1}{h_t}] \tag{III.58}$$

con restricciones con el fin de que las varianzas sean positivas. El algoritmo es similar al utilizado por [Audrino y Bühlmann 2003] con árboles de regresión (*regression trees*).

7. **BoostARX**(u, r). Algoritmo análogo al anterior, pero utilizando como hipótesis base, en lugar de redes neuronales, modelos autorregresivos con u retardos para la serie $\{\xi_t\}$ del menos gradiente (ecuación (III.58)) y covariables $\varepsilon_{t-1}^2, \varepsilon_{t-2}^2, ..., \varepsilon_{t-r}^2$ (componente X del ARX), con el fin de tratar de recoger mejor la posible heterocedasticidad de la serie del menos gradiente y obtener un mejor ajuste de éste en cada iteración del *boosting*.

8. **BoostARMA**(u, v)**_GARCH**(r, s). Algoritmo análogo al anterior, en el que se sustituye el modelo ARX por un modelo ARMA-GARCH para ajustar la serie del menos gradiente mediante máxima verosimilitud gausiana.

   Teóricamente, esta técnica *boosting* se separa de la filosofía Gradient_Boost, al utilizar el criterio de máxima verosimilitud como medida de similitud en lugar del criterio de mínimos cuadrados (norma euclídea en $\mathbb{R}^n$) para el ajuste del menos gradiente (III.58). Esta medida de discrepancia es más correcta que mínimos cuadrados, si el gradiente satisface las hipótesis correspondientes a un modelo GARCH gausiano.

   La utilización de modelos ARMA sin restricciones puede producir en cada iteración valores negativos en la volatilidad. Dichas situaciones se corrigen sustituyendo dichos valores negativos por el menor valor positivo permitido por la precisión del ordenador.

9. **BoostARMA**(u_1, v_1)**_GARCH**(r_1, s_1)-**BoostARMA**(u_1, v_1)**_GARCH**(r_1, s_1). Algoritmo Gradient_Boost que generaliza los anteriores (6. a 8.) a series de precios, estimando conjuntamente los términos de tendencia de varianza.

Ahora, el menos gradiente es un vector con dos componentes[119]: una asociada a la varianza como en los casos anteriores, y otra asociada a la tendencia. Si $\ell(\varepsilon_t, h_t)$ es la verosimilitud gausiana:

$$\ell(\varepsilon_t, h_t) = \frac{1}{2}(\ln h_t + \frac{\varepsilon_t^2}{h_t} + \ln 2\pi) = \frac{1}{2}(\ln h_t + \frac{(y_t - \mu_t)^2}{h_t} + \ln 2\pi) \tag{III.59}$$

el vector menos gradiente es:

$$\boldsymbol{\xi}_t = -\left(\frac{\partial \ell}{\partial \mu_t}, \frac{\partial \ell}{\partial h_t}\right)^T = \left(\frac{y_t - \mu_t}{h_t},\ \frac{1}{2}[\frac{(y_t - \mu_t)^2}{h_t^2} - \frac{1}{h_t}]\right)^T = \left(\varepsilon_t, \frac{1}{2}(\frac{\varepsilon_t^2}{h_t^2} - \frac{1}{h_t})\right)^T$$

con ello, la hipótesis final que produce el algoritmo en la iteración B-ésima, suponiendo $u_i = v_i = r_i = s_i = 1$, $i = 1, 2$, es un vector:

$$\mathbf{f}_{B,t} = (\mu_t, h_t)^T = \sum_{j=1}^{B} \boldsymbol{\psi}_j(y_{t-1}, h_{t-1}, \varepsilon_{t-1}^2) = \left(\sum_{j=1}^{B} f_j(y_{t-1}), \sum_{j=1}^{B} g_j(h_{t-1}, \varepsilon_{t-1}^2)\right)$$

donde $\boldsymbol{\psi}_j = (f_j, g_j)^T$ es la hipótesis vectorial producida por cada iteración de Gradient_Boost.

10. **BoostMLP**(m_1)**-MLP**(m_2) y **BoostMLP**(m_1)**-MLP**(m_2). Se utiliza el algoritmo anterior con la misma finalidad, pero empleando redes RBF o MLP, respectivamente, como componentes de la hipótesis base vectorial $\boldsymbol{\psi}_j$ $j = 1 : B$, que ajustan el vector menos gradiente en el marco del algoritmo Gradient_Boost.

11. **WildBoostGARCH**(r, s). Este algoritmo se utiliza para estimar volatilidades y sigue una filosofía *boosting* diferente a los anteriores. Pretende corregir situaciones en los que los datos observados pertenecen a un período con volatilidad inusualmente baja. Dicho algoritmo se presenta con detalle en el epígrafe siguiente.

Algoritmo WildBoostGARCH

El algoritmo WildBoostGARCH(r, s) se presenta en la tabla 30 para el caso $r = s = 1$ (en otro caso, basta sustituir el GARCH$(1, 1)$ de las tareas 1.a y 2.c de dicha tabla por modelos GARCH(r, s)).

La filosofía del algoritmo es recoger, en cada iteración, la parte de *varianza* no explicada en cada momento por el correspondiente modelo GARCH, e inyectarla de nuevo al modelo (GARCH de la iteración siguiente), mediante el cuadrado de números aleatorios gausianos cuya varianza es la no explicada en cada momento.

[119] En el algoritmo Gradient_Boost de [Audrino y Bühlmann 2003], el espacio teórico de búsqueda es $\mathbb{R}^{\infty}$ (funciones de varianza) y ahora es $\mathbb{R}^{2\times\infty}$ (funciones de tendencia y de varianza). Con n datos, el espacio de búsqueda es $\mathbb{R}^{2\times n}$ en lugar de $\mathbb{R}^n$.

USC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Algoritmo WildBoostGARCH

1. Primera iteración:
   a. Estimar un GARCH$(1,1)$ con la serie $\{\mathsf{Y}_t\}$, obteniéndose:
   $$\begin{cases} \mathsf{Y}_t = \alpha + e_{t,1} \\ h_t^{(1)} = a_1 h_{t-1}^{(1)} + b_1 e_{t-1,1}^2 \end{cases} \quad t \in \mathbb{T}$$
   b. Hipótesis del *boosting* para la varianza:
   $$H_t^{(1)} = h_t^{(1)}, t \in \mathbb{T}$$
2. Iteración $j = 2, ...$:
   a. Calcular las varianzas *no explicadas* por el modelo anterior:
   $$n_{t,j}^2 = [e_{t,j-1}^2 - H_t^{(j-1)}]_+$$
   b. Generar ruido aleatorio gausiano con dichas varianzas:
   $$e_{t,j} = \sqrt{n_{t,j}^2} \cdot \mathcal{N}(0,1)$$
   c. Estimar un GARCH$(1,1)$ con la serie $\{e_{t,j}\}$:
   $$\begin{cases} e_{t,j} = 0 + e_{t,j} \\ h_t^{(j)} = a_1 h_{t-1}^{(j)} + b_1 e_{t-1,j}^2 \end{cases} \quad t \in \mathbb{T}$$
   d. Hipótesis del *boosting* para la varianza:
   $$H_t^{(j)} = H_t^{(j-1)} + h_t^{(j)}, t \in \mathbb{T}$$
   e. Detener si empeora la predicción en una muestra de validación.
3. Hipótesis final del *boosting*: $H_t^{(B)} = \sum_{j=1}^{B} h_t^{(j)}$

Tabla 30. Algoritmo WildBoostGARCH.

Una vez estimado el modelo con las observaciones $\{y_t\}_{i=1}^n$, con $n > 1$, la predicción en $t = n+1$ se realizaría mediante la siguiente secuencia[120] $(r = s = 1)$:

1) $e_{n,1} = \mathsf{Y}_n - \alpha$
2) $h_n^{(1)} = a_1 h_{n-1}^{(1)} + b_1 e_{n-1,1}^2$
3) Predicción:
$$\begin{cases} h_{n+1}^{(1)} = a_1 h_n^{(1)} + b_1 e_{n,1}^2 \\ \vdots \\ h_{n+1}^{(B)} = a_B h_n^{(B)} + b_B e_{n,B}^2 \end{cases} \Rightarrow \begin{cases} H_{n+1}^{(B)} = \sum_{j=1}^{B} h_{n+1}^{(j)} \\ \hat{\mathsf{Y}}_{n+1} = \alpha \end{cases}$$

III.5.3 Evaluación de los Nuevos Algoritmos

En este apartado, se presenta los resultados de la evaluación de los algoritmos propuestos, primero en un conjunto de datos simulados y, después, en una serie de datos reales. Una tarea pendiente es el estudio de las propiedades asintóticas de estos algoritmos.

Salvo que se especifique lo contrario, todos los modelos basados en redes neuronales utilizan un sólo retardo, en cuyo caso, se omite dicho parámetro en la simbología que identifica el modelo. Por ejemplo, un modelo MLP(m)-GARCH(1,1) indica un modelo MLP$(m,1)$-GARCH(1,1), es decir, un modelo MLP con m unidades ocultas y un sólo retardo para la tendencia, y modelo GARCH(1,1) para la volatilidad.

[120] Este proceso produce sobreajuste en los puntos $t \leq n$, lo que se evita si, en dichos puntos las componentes $h_t^{(j)}$ participan con signo negativo en la estimación $H_t^{(j)}$ cuando la diferencia $e_{t,j}^2 = e_{t,j-1}^2 - H_t^{(j-1)}$ sea negativa. Nótese que esto no puede hacerse en los $t > n$ pues no se conocen las innovaciones $e_{t,1}^2, ..., e_{t,j-1}^2$.

Aplicación a Datos Simulados

Los modelos anteriores se evaluaron con los siguientes juegos de datos simulados, todos ellos con innovaciones gausianas:

1. Modelos de Volatilidad (rentabilidades):
   (a) El modelo aditivo de volatilidades utilizado por [Audrino y Bühlmann 2003]:

$$\begin{aligned} \mathsf{Y}_t &= \sqrt{h_t}\varepsilon_t && \text{(III.60a)} \\ h_t &= \mathsf{f}(y_{t-1}, h_{t-1}) && \text{(III.60b)} \end{aligned}$$

   donde:

$$\mathsf{f}(y,h) = (0{,}1 + 0{,}2\,|y| + 0{,}9y^2) * (0{,}8\exp(-1{,}5\,|y|\,|h|)) + (0{,}4y^2 + 0{,}5h)^{3/4} \qquad \text{(III.61)}$$

   (b) El siguiente modelo logarítmico:

$$\begin{aligned} \mathsf{Y}_t &= \sqrt{h_t}\varepsilon_t && \text{(III.62)} \\ \ln h_t &= 0{,}6\ln h_{t-1} + 0{,}3\ln(0{,}5 + 0{,}2y_{t-1}^2) \end{aligned}$$

   (c) El siguiente modelo *threshold*:

$$\begin{aligned} \mathsf{Y}_t &= \sqrt{h_t}\varepsilon_t && \text{(III.63)} \\ h_t &= \mathsf{f}(y_{t-1}, h_{t-1}) \\ \mathsf{f}(y,h) &= \begin{cases} 0{,}2 + 0{,}3y^2 & \text{si } y < 0 \\ 0{,}3 + 0{,}2y^2 + 0{,}6h & \text{si } y \geq 0 \text{ y } h < 0{,}5 \\ 0{,}7 + 0{,}8h & \text{si } y \geq 0 \text{ y } h \geq 0{,}5 \end{cases} \end{aligned}$$

2. Para precios, se utilizó el siguiente modelo con tendencia no lineal y varianza GARCH$(1,1)$:

$$\begin{aligned} \mathsf{Y}_t &= \mathsf{Y}_{t-1}\sin(\mathsf{Y}_{t-1}) + \varepsilon_t && \text{(III.64)} \\ h_t &= 0{,}1 + 0{,}85h_{t-1} + 0{,}1\varepsilon_{t-1}^2 && \text{(III.65)} \end{aligned}$$

Como referencia en la evaluación de los distintos algoritmos presentados anteriormente, se utilizan modelos ARMA-GARCH estimados mediante máxima verosimilitud gausiana. En algún caso, estos modelos se estimaron mediante verosimilitud t-Student con diversos grados de libertad con el fin de comparar los modelos altamente no lineales anteriores, con modelos no gausianos (véase [Schittenkopf *et al.* 1999]).

A continuación, presentamos los resultados obtenidos en cada caso, en los que se utilizan los siguientes criterios de error, para una muestra de tamaño N (que puede ser n si es la muestra de entrenamiento y n_{test}, si se trata de una muestra independiente de test):

1. La media de los errores al cuadrado para la serie de precios. Si $\hat{f}_n(y_{t-1}, y_{t-2}, ...)$ es la predicción de la serie:

$$\text{ASRy} \doteq \frac{1}{N}\sum_{t=1}^{N}[y_t - \hat{f}_n(y_{t-1}, y_{t-2}, ...)]^2 \qquad \text{(III.66)}$$

2. La media de los errores al cuadrado para las varianzas cuando éstas han sido simuladas y se conocen. Si σ_t^2 son las varianzas verdaderas y $\hat{g}_n(\varepsilon_{t-1}^2, \varepsilon_{t-2}^2, ..., h_{t-1}, h_{t-2}, ...)$ su esti-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

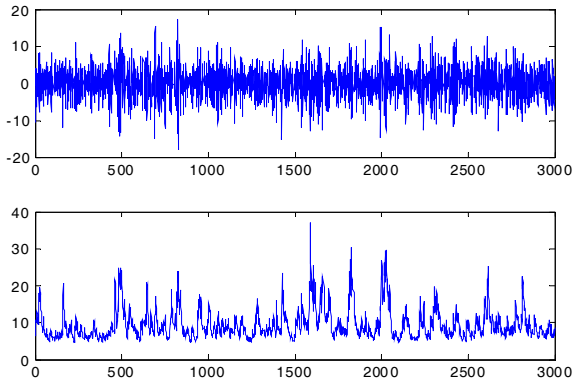


Figura 56. Arriba: serie $Y_t \sin(Y_t)$ generada según la ecuación (III.64). Abajo: serie de varianzas generadas el modelo GARCH de la ecuación (III.65).

mación:

$$\text{ASRh} \doteq \frac{1}{N}\sum_{t=1}^{N}[\sigma_t^2 - \hat{g}_n(\varepsilon_{t-1}^2, \varepsilon_{t-2}^2, ..., h_{t-1}, h_{t-2}, ...)]^2 \tag{III.67}$$

3. El menos logaritmo neperiano de la verosimilitud gausiana (salvo que se especifique lo contrario):

$$\text{-lnLkhd} \doteq \sum_{t=1}^{N} \ell(y_t; \hat{\mu}_t, \hat{h}_t)$$

$$\text{con: } \begin{cases} \ell(y_t, \mu_t, h_t) = \frac{1}{2}(\ln h_t + \dfrac{(y_t - \mu_t)^2}{h_t} + \ln 2\pi) \\ \hat{\mu}_t = \hat{f}_n(y_{t-1}, y_{t-2}, ...) \\ \hat{h}_t = \hat{g}_n(\varepsilon_{t-1}^2, \varepsilon_{t-2}^2, ..., h_{t-1}, h_{t-2}, ...) \end{cases} \tag{III.68}$$

donde $\hat{\mu}_t$ y $\hat{h}_t$ son las estimaciones para la media y la varianza respectivamente.

Serie $Y_t sin(Y_t)$ de Precios

La figura 56 muestra el aspecto de esta serie (gráfico superior) y de la serie de sus varianzas (gráfico inferior).

La figura 57 muestra las funciones de autocorrelación y autocorrelación parcial de la serie y de su cuadrado. Como puede observarse respecto a la primera, las funciones de autocorrelación no reflejan asociación temporal alguna, pues estas funciones, por definición, sólo detectan asociaciones de tipo lineal. Por el contrario, las funciones de autocorrelación del cuadrado de la serie muestran una clara dependencia, lo que sugiere heterocedasticidad.

Para esta serie se generaron 3000 observaciones de las cuáles las $n = 1000$ primeras se utilizaron para el entrenamiento de los modelos y las $n_{\text{test}} = 2000$ siguientes como muestra de test para la evaluación de los mismos.

La tabla 31 muestra los resultados obtenidos con algunos de los algoritmos descritos anteriormente. En ella se refleja, para cada modelo: los valores ASRh y ASRy (ecuaciones (III.67) y (III.66), respectivamente), en la muestra de test, los valores de la verosimilitud (III.68) en la

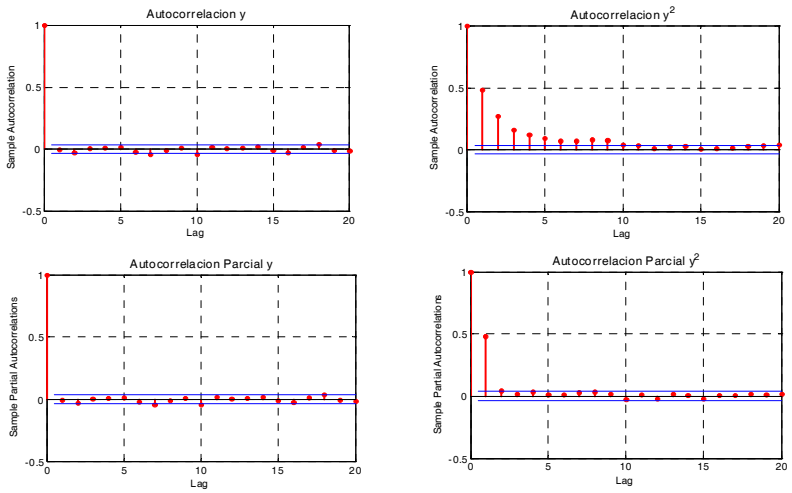


Figura 57. Izquierda: autocorrelación (arriba) y autocorrelación parcial (abajo) de la serie definda por las ecuaciones (III.64) y (III.65). Derecha: autocorrelación y autocorrelación parcial del cuadrado de la serie.

Modelo	ASRhtest	ASRytest	-lnLkhdtrain	-lnLkhdtest	K	GARCH	ARCH
ARMA(0,0)-GARCH(1,1)	312.6492	17.2874	2735.1013	5534.5503	4.9194	0.1901	0.5629
ARMA(1,0)-GARCH(1,1)	313.5288	17.3006	2731.8485	5534.3346	4.9794	0.1827	0.5651
RBF(20)-GARCH(1,1)	4.0412	10.2725	2487.0357	5112.2975	0.8759	0.7631	0.1427
MLP(13)-GARCH(1,1)	1.2296	9.7356	2446.0942	5057.4413	0.4992	0.8668	0.0723
MLP(13)-MLP(1)	27.8006	10.2420	2507.9589	5158.4596			

Tabla 31. Resultados de diferentes algoritmos con la serie $Y_t \sin(Y_t)$

muestra de entrenamiento y de test, y las estimaciones de los parámetros del modelo GARCH de la ecuación (III.65).

Al respecto de dicha tabla, destacan los magníficos resultados de los algoritmos basados en redes neuronales, especialmente el MLP(13)-GARCH(1,1), que propone un modelo de varianza muy parecido al verdadero (ecuación (III.65)).

Para valorar correctamente los resultados obtenidos en términos de verosimilitud (menos logaritmo), es preciso tener en cuenta la dificultad que posee producir mejoras en dicho criterio, debido a que éste mide la calidad de la predicción de rentabilidades futuras y es mucho más volátil que el error cuadrático, el cuál, mide la calidad de la predicción de las volatilidades [Audrino y Bühlmann 2003].

Como puede observarse, la estimación de los parámetros del modelo GARCH verdadero son totalmente erróneas en los modelos ARMA(0,0)-GARCH(1,1) y ARMA(1,0)-GARCH(1,1), mostrando que han confundido la asociación debida a la tendencia, con la originada por la varianza.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

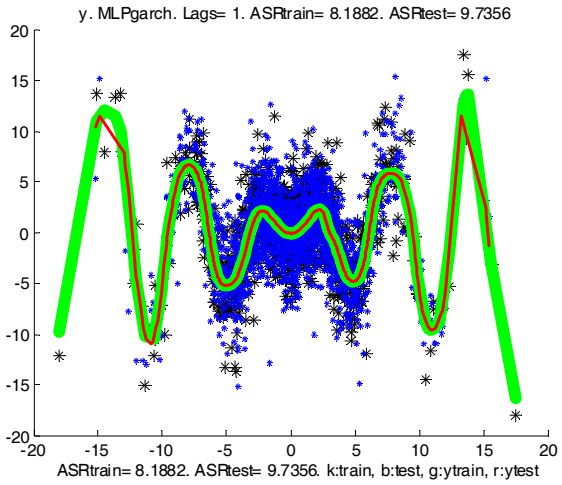


Figura 58. Ajuste de la tendencia por parte del modelo MLP(13)-GARCH(1,1) en la serie (III.64).

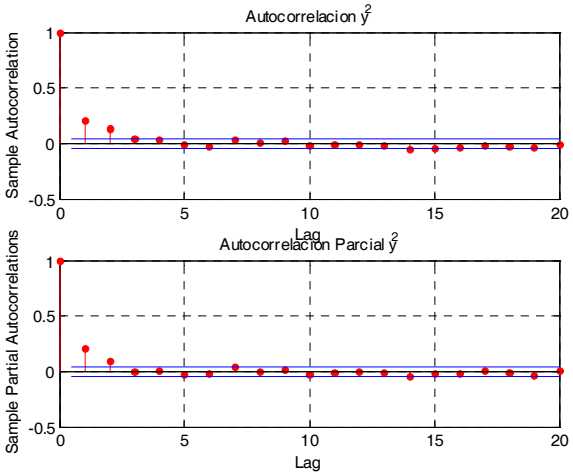


Figura 59. Función de autocorrelación y autocorrelación parcial para el cuadrado de la serie (III.60a)

Modelo Aditivo de Volatilidad Audrino-Bühlmann'03

La figura 59 muestra las funciones de autocorrelación y autocorrelación parcial del cuadrado de la serie (III.60a), las cuáles sugieren la heterocedasticidad que posee la serie, aunque la no linealidad de la función (III.61), provoca una asociación lineal de dos retardos en lugar de uno.

La tabla 32 muestra los resultados de diferentes algoritmos presentados en el epígrafe anterior, en relación con esta serie de rentabilidades. Dicha tabla, además, de los conceptos análogos a la tabla mostrada para la serie anterior, incluye además, para los algoritmos *boosting*, la indicación de si se inicia con un modelo GARCH(1,1) o no, (IniGH) y la iteración en la que produjo su mejor resultado (BestT).

Sobre dichos resultados, destacamos lo siguiente:

1. El mejor resultado en verosimilitud (menos logaritmo) lo produjo el modelo MLP(3) utilizando 3 retardos en la estimación de la función de regresión h_t de la serie ε_t^2. Dado el modelo

Modelo	IniGH	BestT	ASRhtrain	ASRhtest	-lnLkhdtrain	-lnLkhdtest
ARMA(1,0) sobre y^2			0.57228	0.56725	3.6457E+05	3.2779E+05
ARMA(3,0) sobre y^2			0.37940	0.38728	2721.6763	2614.0159
ARMA(1,1) sobre y^2			0.33840	0.31131	1505.4146	1443.9350
ARMA(3,1) sobre y^2			0.14276	0.16257	1460.4261	1419.1002
ARMA(0,0)-GARCH(1,1)			0.01379	0.01050	1438.1846	1394.9637
KR - 1 delay sobre y^2			0.06900	0.04718	1442.4608	1397.0376
LPR - 1 delay sobre y^2			0.05981	0.04143	1442.7515	1395.1033
MLP(5) - 1 delay sobre y^2			0.03614	0.02817	1445.7921	1393.4111
MLP(3) - 3 delay sobre y^2			0.03245	0.02278	1431.4373	1389.8469
RBF(1)-GARCH(1,1) sin cte.			0.01274	0.01175	1438.4027	1394.6164
MLP(1)-GARCH(1,1)			0.01406	0.01406	1438.1995	1395.0281
MLP(0)-GARCHMLP(1) sin h			0.04113	0.03251	1445.0503	1394.7193
MLP(0)-GARCHMLP(2) sin h			0.03961	0.03069	1444.9565	1393.9235
MLP(0)-GARCHMLP(3) sin h			0.03951	0.03056	1444.9451	1393.7686
MLP(0)-GARCHMLP(4) sin h			0.03830	0.02976	1444.9358	1393.7970
MLP(1)-GARCHMLP(3) sin h			0.03724	0.02881	1444.9451	1393.4319
BostMLP(1) sin h	no	5	0.03724	0.02941	1445.8222	1394.1309
BostMLP(2) sin h	no	2	0.03615	0.02858	1445.7836	1393.9043
BostMLP(3) sin h	no	4	0.03616	0.02805	1445.7688	1393.8758
BostMLP(4) sin h	no	2	0.03519	0.02766	1445.7617	1393.9121
BostMLP(1) con h	no	2	0.02704	0.01749	1436.6535	1394.1750
BostMLP(2) con h	no	2	0.02045	0.02109	1436.2065	1393.1753
BostMLP(3) con h	no	3	0.01663	0.01679	1434.8903	1392.7571
BostMLP(4) con h	no	3	0.04078	0.02802	1444.7673	1393.0451
BoostMLP(3) sin h	si	1	= GARCH	= GARCH	= GARCH	= GARCH
BoostMLP(3) con h	si	1	= GARCH	= GARCH	= GARCH	= GARCH
BoostARX(1,0) (Grad,Innov^2)	no	26	0.03112	0.02599	1446.0340	1393.8932
BoostARX(1,1) (Grad,Innov^2)	no	3	0.03118	0.02682	1446.0547	1393.7485
BoostARX(2,1) (Grad,Innov^2)	no	6	0.01751	0.01366	1435.8167	1392.0045
BoostARX(2,1) (Grad,Innov^2)	si	43	0.01083	0.00956	1439.0777	1393.5385
BoostARMA(4,1)-GARCH(1,1) real	si	2	0.01328	0.00943	1437.9186	1394.3541
BoostGARCH-AR(1)GARCH	no	20	0.03369	0.02654	1446.1254	1394.7119
BoostMLP(1) - MLP(3)	no	3	0.02625	0.01914	1434.9496	1394.6956
WildBoostGARCH(1,1)	--	1	= GARCH	= GARCH	= GARCH	= GARCH

Tabla 32. Resultados de diferentes algoritmos con la serie de rentabilidades aditiva Audrino-Buhlmann 2003.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

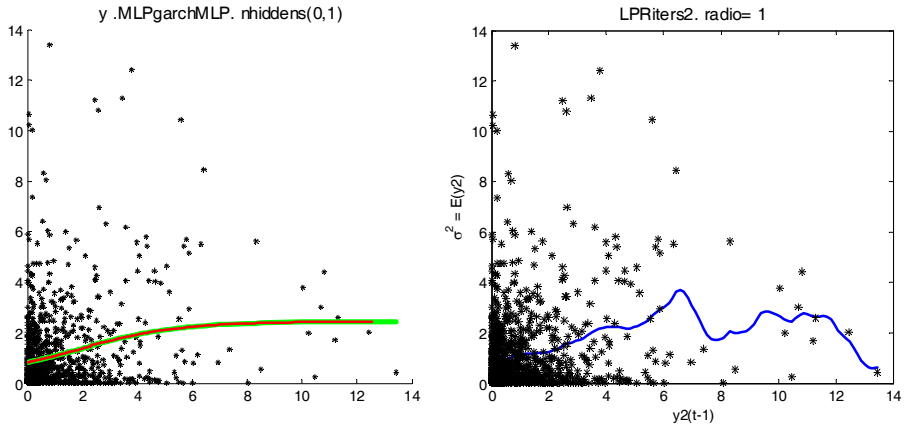


Figura 60. Ajuste de ε_t^2 mediante el modelo MLP(0)-GARCHMLP(1) (izquierda) y mediante regresión polinómico local (derecha). Se observa como este último es excesivamente sensible a la aleatoriedad de los datos, en esta serie.

generador de la varianza, puede sorprender este resultado pero, puede entenderse, a la vista de las funciones de autocorrelación de la figura 59, teniendo en cuenta la utilización de la pérdida de mínimos cuadrados en el entrenamiento de dicha red.

2. En general, el modelo GARCH(1,1) se ve superado por casi todos los modelos no lineales. Después de la MLP(3) destaca el BoostARX(2,1) (Gradient_Boost con verosimilitud, ajustando el gradiente con un modelo AR más ε_t^2 como covariable), y el BoostMLP(3) donde la inclusión de h_{t-1} como covariable ('con h' en la tabla) produjo mejores resultados.

   Los modelos MLP(0)-GARCHMLP(m) son estructuras GARCH no lineales modelizadas con una red MLP. Puede observarse como, con esta serie, mejoran claramente los resultados del GARCH.

3. De cualquier forma, no se perciben grandes diferencias entre los grupos de algoritmos basados en redes neuronales individuales y los basados en Gradient_Boost.

4. Con los algoritmos *boosting*, el inicio con un GARCH(1,1) no suele producir mejoras en esta serie y, en todo caso, lleva al sobreajuste rápidamente (al contrario de lo detectado por [Audrino y Bühlmann 2003] con esta serie en el *boosting* de árboles de regresión).

5. Los modelos no paramétricos Nadaraya-Watson (KR) y de regresión polinómico local (lineal) (LPR), no produjeron los mejores resultados con esta serie, no pudiendo mejorar siquiera los resultados del GARCH(1,1).

   A nuestro juicio, la razón de ello, es su excesiva sensibilidad a la aleatoriedad de los datos, al menos en esta serie. La figura 60 muestra el ajuste del modelo LPR en contraste con la MLP(0)-GARCHMLP(1). A primera vista podría parecer que el modelo LPR requiere una ventana más amplia, pero ello producía peores resultados (el modelo Nadaraya-Watson se comporta de forma similar).

6. El algoritmo WildBoostGARCH(1,1) no produce beneficios con esta serie en relación con el GARCH(1,1). Ello es debido a que los datos de entrenamiento son perfectamente representativos de su estructura de varianza.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

	ASRtest	vs. GARCH	-lnLkhdtest	vs. GARCH
Modelos [Audrino y Bühlmann 2003]				
GARCH(1,1)	0,1115	1,0000	1656,3630	1,0000
BoostTree$(1, 0{,}1, 3)$	0,0895	0,8027	1654,3610	0,9988
BoostTree$(2, 0{,}1, 5)$	0,0932	0,8359	1654,5410	0,9989
BoostPPR$(2, 0{,}05, 2)$	0,1098	0,9848	1656,1780	0,9999
Modelos Desarrollados				
GARCH(1,1)	0,0596	1,0000	1437,0522	1,0000
BoostMLP(3) con h	0,0921	1,5453	1435,4498	0,9989
BoostARX(2,1)	0,0376	0,6309	1435,9747	0,9993
BoostGARCH$(1, 1)$-ARMA$(2, 1)$_GARCH$(1, 1)$	0,0523	0,8775	1436,4960	0,9996
BoostARMA$(4, 1)$-GARCH$(1, 1)$	0,0527	0,8842	1436,5656	0,9997

Tabla 33. Comparación de resultados entre la simuación realizada por Audrino y Buhlmann, 2003 y la realizada en este trabajo.

7. Finalmente, es curioso observar que el modelo GARCH(1,1) haya producido uno de los mejores resultados en media cuadrática (junto con los modelos BoostARX(2,1) y BoostARMA(4,1)-GARCH(1,1)), no siendo de los mejores en verosimilitud. Parece que los modelos no lineales pagan un precio por un mejor ajuste a la verosimilitud.

[Audrino y Bühlmann 2003] prueban sus algoritmos en 50 muestras de tamaño 2000 generadas mediante la estructura (III.60a), dedicando $n = 1000$ a entrenamiento y reservando $n_{\text{test}} = 1000$ como muestra de test.

Para comparar algunos de estos algoritmos con los propuestos por [Audrino y Bühlmann 2003], dado que los juegos de muestras pueden producir diferencias importantes, incluso con los mismos modelos, (por ejemplo, la varianza inicial de cada realización de la serie es muy importante y no conocemos el valor de ese parámetro utilizado por [Audrino y Bühlmann 2003]), hemos tomado como referencia los resultados del modelo GARCH(1,1) utilizado por cada uno en sus propios juegos de muestras.

La tabla 33 recoge los resultados de dicha comparación, mostrando, para el error cuadrático y para la verosimilitud, la tasa de mejora de cada algoritmo con respecto al modelo GARCH(1,1) utilizado en su mismo conjunto de muestras.

Las grandes diferencias en verosimilitud entre las pruebas de [Audrino y Bühlmann 2003] y las nuestras, se deben probablemente a que, los primeros, hayan utilizado el logaritmo en base dos para el cálculo de la verosimilitud en lugar del logaritmo natural que se ha utilizado aquí. No hemos modificado el criterio porque, de todas formas, las cifras no eran comparables: con logaritmo base dos, los resultados en nuestro juego de muestras eran del orden[121] 1800.

Los modelos utilizados por [Audrino y Bühlmann 2003] son: un algoritmo Gradient_Boost con árboles de regresión (BoostTree(r, λ, m) donde r es el número de retardos utilizados en el modelo, λ es un parámetro regularizador de cada hipótesis base introducida, y m es el número de nodos finales de cada hipótesis base, y un algoritmo Gradient_Boost sobre redes *projection pursuit*, BoostPPR(r, λ, m) donde r y λ poseen el mismo significado que en el anterior, y m es el número de funciones básicas de la red (no se especifica su tipo).

Con respecto a dicha tabla, puede observarse que, las mejoras en verosimilitud son del mismo orden a las conseguidas por [Audrino y Bühlmann 2003], y las correspondientes a error cuadrático

[121] Y la no inclusión de la constante $\log_2(2\pi)$ en la verosimilitud produce valores menores que los de [Audrino y Bühlmann 2003].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

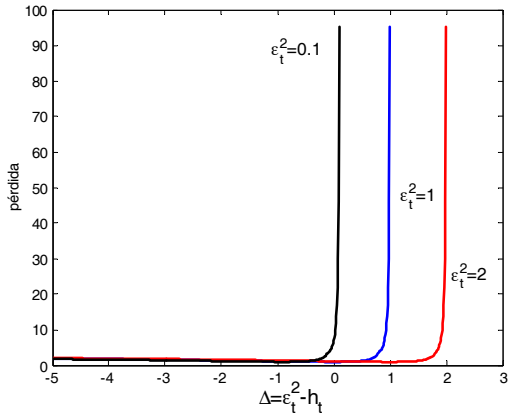


Figura 61. Pérdida correspondiente al menos logaritmo de la verosimilitud gausiana: $\ell(\varepsilon_t^2, \Delta_t) \equiv \ell(\varepsilon_t^2, h_t) = \ln(\varepsilon_t^2 - \Delta_t) + \varepsilon_t^2/(\varepsilon_t^2 - \Delta_t)$, donde $\Delta_t = \varepsilon_t^2 - h_t \in (-\infty, \varepsilon_t^2)$. Se observa como, un mismo error Δ_t, se penaliza con más fuerza cuando ε_t^2 es pequeño, salvo que $h_t > 0$ sea grande con respecto a ε_t^2.

medio son más heterogéneas. Con este último criterio, destaca, a favor, el modelo BoostARX(2,1), que utiliza la estructura temporal del gradiente hasta con dos retardos, y, en contra, el modelo BoostMLP(3), que puede verse afectado por una mayor inestabilidad debido al algoritmo de optimización no lineal.

En relación con este último modelo, debe tenerse en cuenta que, este tipo de pruebas automáticas y masivas como la anterior, suelen impedir una correcta evaluación de su potencial dado que, impiden la intervención del usuario en el filtrado de posibles soluciones menos favorecidas por valores iniciales del algoritmo de optimización (lo que contrarrestaría la mayor varianza del estimador).

El relativamente bajo grado de correlación en el *ranking* de los estimadores, en términos de verosimilitud y error cuadrático, se ha manifestado con diferente intensidad en los distintos juegos de datos que hemos utilizado, y, muchas veces, el mejor estimador en verosimilitud no es el mejor en media cuadrática y viceversa.

Esta circunstancia, no es sorprendente, si tenemos en cuenta las propiedades de la pérdida basada en la verosimilitud gausiana:

$$\ell(\varepsilon_t^2, h_t) = \frac{1}{2}(\ln h_t + \frac{\varepsilon_t^2}{h_t} + \ln 2\pi)$$

en la que, haciendo $\Delta_t = \varepsilon_t^2 - h_t \in (-\infty, \varepsilon_t^2)$, se tiene:

$$\ell(\varepsilon_t^2, \Delta_t) = \ln(\varepsilon_t^2 - \Delta_t) + \frac{\varepsilon_t^2}{\varepsilon_t^2 - \Delta_t}$$

con lo que (figura 61), un mismo error Δ_t se ve mucho más penalizado cuando ε_t^2 es pequeño, (se penaliza el error relativo en lugar del absoluto), salvo que $h_t > 0$ sea grande con respecto a ε_t^2.

No hemos llegado a evaluar los efectos de la aplicación de regularización en los algoritmos aquí propuestos pero, creemos que pueden llegar a mejorar los resultados a costa de mayor número de iteraciones. Esta evaluación se llevará a cabo en un futuro, junto con la evaluación de otros tipos de métodos de optimización de las hipótesis base basadas en redes neuronales.

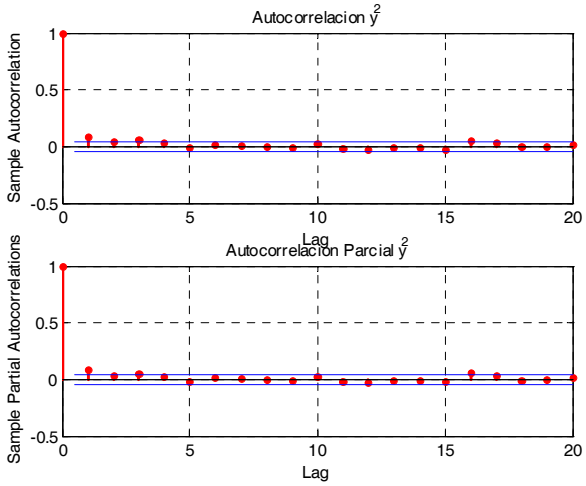


Figura 62. Función de autocorrelación (arriba) y de autocorrelación parcial (abajo) del cuadrado de la serie logarítmica (III.62).

Serie de Rentabilidades con Volatilidad Logarítmica

La figura 62 muestra el aspecto de las funciones de autocorrelación y autocorrelación parcial de los cuadrados de esta serie, mostrando una casi nula asociación temporal de tipo lineal.

La tabla 34 muestra los resultados de diferentes algoritmos en dicha serie, mostrando los mismos conceptos que en la tabla del epígrafe anterior. Al respecto, destacan las siguientes circunstancias:

1. En general, todos los algoritmos producen resultados bastante similares. Quizás su estructura de varianza posea una configuración no lineal excesivamente débil.

2. De cualquier forma, todos los no lineales superan al GARCH(1,1) en verosimilitud y gran parte en media cuadrática. (Con esta serie, los resultados en media cuadrática van más parejos con los correspondientes en verosimilitud).

3. En esta serie, destaca el buen resultado de la regresión polinómico-local, (segundo después del BoostARMA(2,1)-GARCH(1,1)), aunque no produce tan buenos resultados en media cuadrática.

La figura 63 muestra el grado de reproducción por el modelo MLP(0)-MLPGARCH(1) (derecha) de la estructura teórica de la varianza (izquierda):

$$\ln h_t = 0{,}6 \ln h_{t-1} + 0{,}3 \ln(0{,}5 + 0{,}2\varepsilon_{t-1}^2) \qquad \text{(III.69)}$$

Serie de Rentabilidades con Volatilidad *Threshold*

La figura 64 muestra la estructura de correlación de la serie de rentabilidades con estructura *threshold* de varianzas, definida en (III.63).

La tabla 35 muestra los resultados de diferentes algoritmos aplicados a esta serie. Sobre ella, destacamos lo siguiente:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Modelo	IniGH	BestT	ASRhtrain	ASRhtest	-lnLkhdtrain	-lnLkhdtest
ARMA(1,0) sobre y^2			0.3053	0.3094	1.2780E+06	2.1964E+06
ARMA(3,0) sobre y^2			0.1913	0.1937	2791.6343	2721.4980
ARMA(1,1) sobre y^2			0.0453	0.0541	1324.0161	1303.3490
ARMA(3,1) sobre y^2			0.0349	0.0453	1307.6621	1292.9485
ARMA(0,0)-GARCH(1,1)			0.0081	0.0070	1265.1735	1246.2287
KR - 1 delay sobre y^2			0.0138	0.0096	1267.5032	1245.6727
LPR - 1 delay sobre y^2			0.0111	0.0100	1268.4531	1244.7191
MLP(4) iterada - 1 delay y^2		2	0.0065	0.0062	1268.2422	1245.4585
MLP(1) - 2 delay sobre y^2			0.0051	0.0050	1268.1405	1245.3521
RBF(1)-GARCH(1,1) sin cte.			0.0078	0.0068	1265.1941	1246.0689
MLP(1)-GARCH(1,1)			0.0080	0.0069	1265.1813	1246.1249
MLP(0)-GARCHMLP(1) sin h			0.0072	0.0068	1268.8563	1245.6150
MLP(0)-GARCHMLP(1) con h			0.0054	0.0052	1267.6847	1245.1837
BostMLP(1) sin h	no	10	0.0070	0.0068	1268.8391	1245.6241
BostMLP(1) sin h	si	8	0.0079	0.0070	1265.1732	1246.2197
BostMLP(2) sin h	no	9	0.0070	0.0068	1268.8273	1245.6726
BoostARX(1,0) (Grad,Innov^2)	no	60	0.0077	0.0075	1268.4021	1245.7670
BoostARX(1,1) (Grad,Innov^2)	no	2	0.0084	0.0084	1268.4143	1245.8939
BoostARX(2,1) (Grad,Innov^2)	no	2	0.0071	0.0070	1266.2630	1245.6203
BoostARX(2,1) (Grad,Innov^2)	si	5	0.0077	0.0067	1262.7971	1246.1387
BoostARMA(1,0)-GARCH(1,1) real	no	2	0.0051	0.0050	1268.7107	1245.2785
BoostARMA(1,0)-GARCH(1,1) real	si	1	=GARCH	=GARCH	=GARCH	=GARCH
BoostARMA(2,1)-GARCH(1,1) real	no	4	0.0026	0.0024	1267.0710	1244.6539
BoostGARCH-AR(1)GARCH	no	3	0.0056	0.0052	1260.6675	1245.4049
BoostGARCH-AR(1)GARCH	si	2	0.0086	0.0073	1264.8575	1246.2240
BoostMLP(1) - MLP(2)	no	3	0.0066	0.0064	1269.2528	1246.1576
BoostMLP(1) - MLP(2)	si	1	=GARCH	=GARCH	=GARCH	=GARCH
WildBoostGARCH(1,1)	--	1	=GARCH	=GARCH	=GARCH	=GARCH

Tabla 34. Resultados de diferentes algoritmos con la serie de rentabilidades con volatilidad logaritmica.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

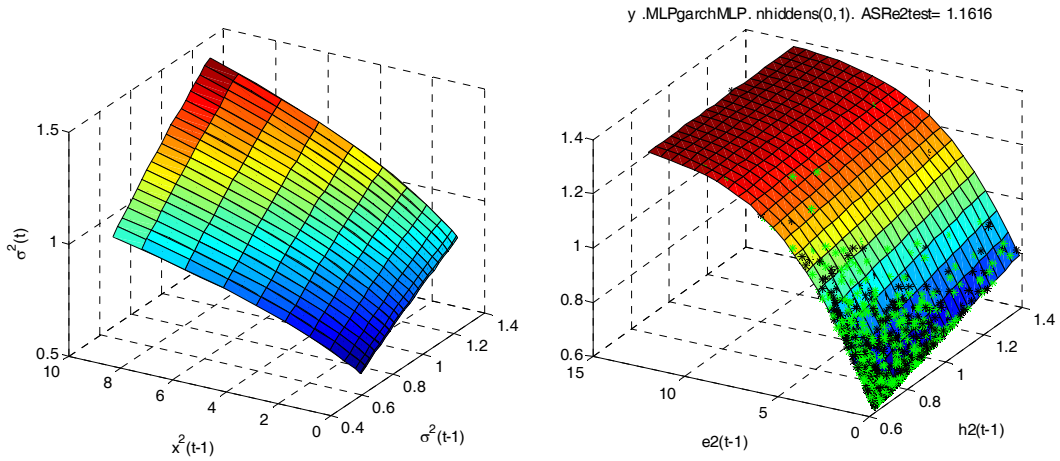


Figura 63. Izquierda: estructura de la varianza de la serie con volatilidades según la función (III.69). Derecha: ajuste de dicha estructura mediante un modelo MLP(0)-GARCHMLP(1).

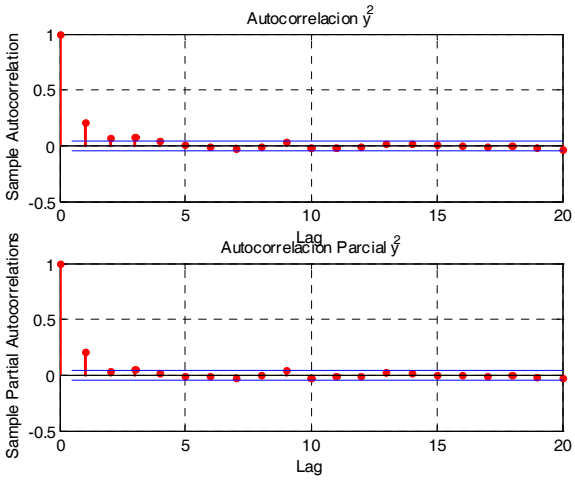


Figura 64. Funciones de autocorrelación (izquierda) y de autocorrelación parcial (derecha), del cuadrado de la serie de rentabilidades (III.63).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Modelo	IniGH	BestT	ASRhtrain	ASRhtest	-lnLkhdtrain	-lnLkhdtest
ARMA(1,0) sobre y^2			0.7452	0.6916	2.3950E+06	6.6833E+06
ARMA(1,1) sobre y^2			0.5984	0.6130	1533.8653	1615.9590
ARMA(2,1) sobre y^2			0.5148	0.4881	1286.5948	1365.3863
ARMA(0,0)-GARCH(1,1)			0.3505	0.3532	1202.4471	1273.2961
KR - 1 delay sobre y^2			0.4195	0.3938	1214.1310	1279.1391
LPR - 1 delay sobre y^2			0.3885	0.3515	1220.6305	1284.5577
MLP(4) sin h - 1 delay y^2			0.3672	0.3306	1212.0721	1271.3006
MLP(4) iterada - 1 delay y^3		3	0.3441	0.3179	1208.8329	1268.5263
MLP(2) - 2 delay sobre y^2			0.3380	0.3236	1207.6881	1271.2977
RBF(1)-GARCH(1,1) sin cte.			0.3422	0.3303	1203.8232	1274.3164
MLP(1)-GARCH(1,1)			0.3421	0.3306	1203.6727	1274.2907
MLP(0)-GARCHMLP(1) sin h			0.3561	0.3198	1209.1506	1273.6081
MLP(0)-GARCHMLP(1) con h			0.3403	0.3165	1207.5920	1268.3193
MLP(0)-GARCHMLP(2) con h			0.3405	0.3132	1204.8237	1267.6430
BostMLP(1) sin h	n	4	0.3650	0.3286	1212.9630	1272.0321
BostMLP(2) sin h	n	3	0.3658	0.3283	1211.8407	1271.5125
BostMLP(3) sin h	n	3	0.3645	0.3270	1211.9749	1271.5048
BostMLP(1) con h	n	3	0.3787	0.3220	1212.4533	1272.3057
BostMLP(2) con h	n	3	0.3674	0.3261	1213.1490	1272.7768
BostMLP(3) con h	n	3	0.3661	0.3393	1213.8407	1272.0114
BostMLP(3) sin h	s	1	= GARCH	= GARCH	= GARCH	= GARCH
BostMLP(3) con h	s	1	= GARCH	= GARCH	= GARCH	= GARCH
BoostARX(1,0) (Grad,Innov^2)	n	2	0.3671	0.3286	1214.9189	1274.0842
BoostARX(1,1) (Grad,Innov^2)	n	2	0.3670	0.3285	1214.9397	1274.0449
BoostARX(1,1) (Grad,Innov^2)	s	1	= GARCH	= GARCH	= GARCH	= GARCH
BoostARMA(1,0)GARCH(1,1) real	n	6	0.3576	0.3293	1212.4091	1272.9840
BoostARMA(1,1)GARCH(1,1) real	n	64	0.3364	0.3256	1207.8976	1272.9273
BoostARMA(2,1)GARCH(1,1) real	n	15	0.3506	0.3257	1211.3820	1274.5138
BoostGARCH-AR(1)GARCH	n	9	0.4986	0.3383	1212.3332	1273.2042
BoostGARCH-AR(1)GARCH	s	2	0.3526	0.3548	1196.3490	1263.4975
BoostMLP(1) - MLP(2)	n	2	0.3622	0.3273	1213.5563	1271.0019
BoostMLP(1) - MLP(3)	s	2	0.3505	0.3532	1196.7159	1263.4881
WildBoostGARCH(1,1)	--	1	= GARCH	= GARCH	= GARCH	= GARCH

Tabla 35. Resultados de diferentes algoritmos con la serie de rentabilidades con volatilidad threshold.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1. En general, esta serie provoca mayores diferencias entre los diferentes algoritmos. Todos los algoritmos no lineales superan al modelo GARCH(1,1) tanto en verosimilitud como en media cuadrática, salvo los modelos no paramétricos Naradaya-Watson y polinómico local lineal (KR y LPR).

2. En particular, los algoritmos Gradient_Boost simultáneos para tendencia y varianza producen los mejores resultados con diferencia. En este sentido, son muy significativas las mejoras del modelo[122] BoostGARCH-AR(1)GARCH y del modelo BoostMLP(1)-MLP(3).

   Ambos algoritmos se iniciaron con un GARCH(1,1), lo que indica que la conveniencia de esta práctica depende de la serie en cuestión (y quizás de la propia muestra) y también de la hipótesis base utilizada (véase el caso de los modelos BoostMLP(m) y BoostARX(1,1) que sobreajustan rápidamente).

3. Destaca asimismo, el algoritmo MLP(4) iterado, el cuál retroalimenta en el modelo, la estimación de la varianza de la iteración anterior. (Con tres iteraciones adicionales, mejora los resultados del modelo original de forma significativa).

4. En esta serie, sorprende el mal resultado del modelo polinómico local (LPR) a pesar del cuidado puesto en la selección del modelo.

Aplicación a Datos Reales. Indice Standard & Poor's 500

Como conjunto de datos reales hemos tomado la serie diaria del índice Standard & Poor's 500 (S&P). Dicha serie es un índice ponderado de 500 valores estadounidenses elegidos según tamaño de mercado, liquidez y representación en su sector, ponderados según su valor de mercado.

Los datos corresponden al período que va desde el 2 de enero de 1992 hasta el 31 de diciembre de 2002 (2772 datos), a partir de los cuáles, se obtiene la correspondiente serie de rentabilidades, mediante:

$$\mathsf{X}_t = \ln \frac{\mathsf{Y}_t}{\mathsf{Y}_{t-1}}$$

donde $\{\mathsf{Y}_t\}$ es la serie S&P.

A continuación analizamos el comportamiento de los distintos algoritmos en la serie de rentabilidades y, posteriormente, lo haremos con la serie de precios.

Serie de Rentabilidades S&P

La figura 65 muestra las funciones de autocorrelación y de autocorrelación parcial para la serie de rentabilidades S&P.

Como puede observarse, la serie no presenta asociación lineal (figura superior izquierda), sin embargo, la serie de sus cuadrados muestra una asociación persistente en el tiempo (figura superior derecha), lo que sugiere heterocedasticidad.

Esta circunstancia se manifiesta también en la serie de las 2000 primeras observaciones (primeros de diciembre de 1999, figura inferior izquierda), aunque más atemperada. Sin embar-

[122] Denotado al inicio como BoostARMA(0,0)_GARCH(1,1)-ARMA(1,0)_GARCH(1,1), donde la primera combinación ajusta la componente del gradiente relativa a la tendencia, y la segunda, la relativa a la varianza.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

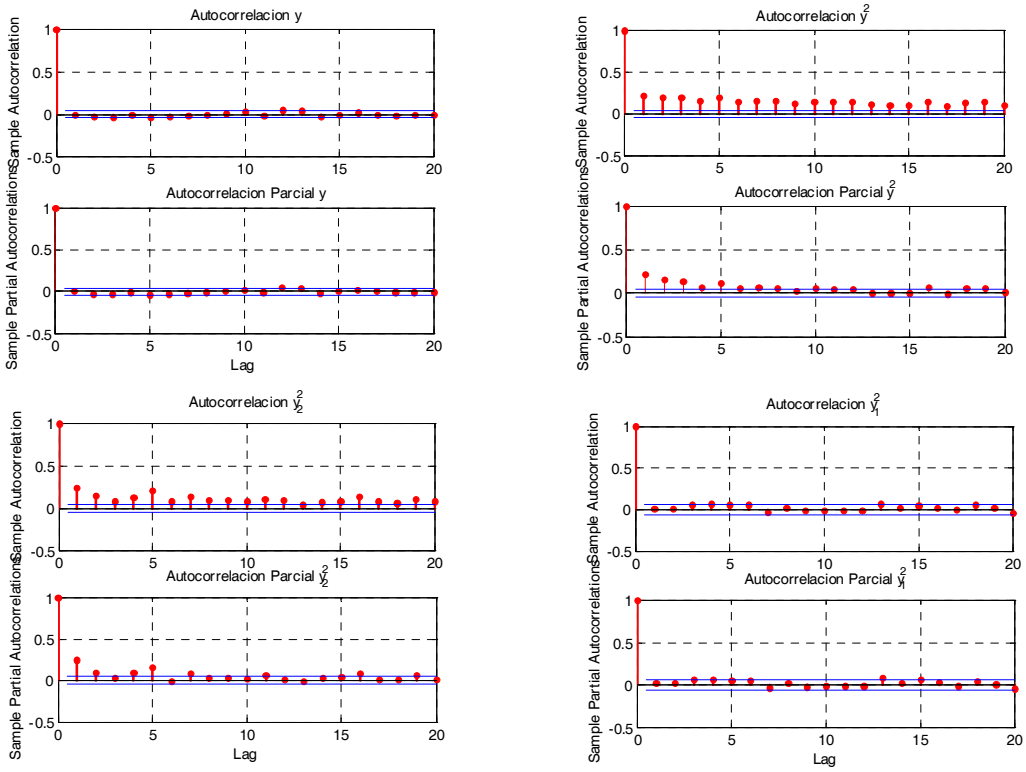


Figura 65. De izquierda a derecha y de arriba a abajo: funciones de autocorrelación y autocorrelación parcial de: la serie original de volatilidades S&P, la serie de sus cuadrados con todas las observaciones, con las 2000 primeras y con las 1000 primeras. Puede observarse como en el último caso, se pierde la asociación lineal entre el cuadrado de las innovaciones, que existe en toda la serie y en sus primeros 2000 datos.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

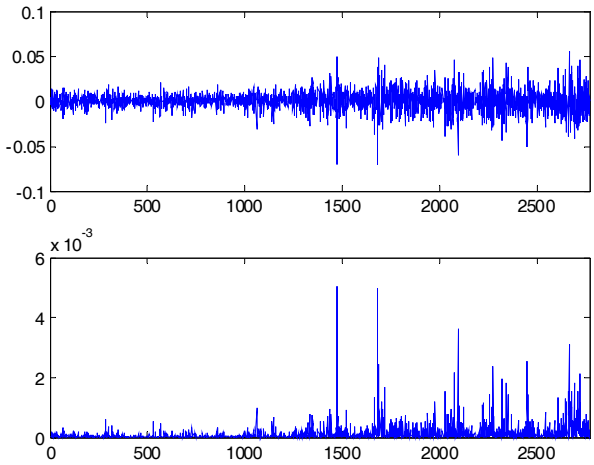


Figura 66. Arriba: serie de rentabilidades S&P. Abajo: cuadrados de dicha serie.

go, sorprendentemente, no se manifiesta durante las primeras 1000 observaciones (mediados de diciembre de 1995, figura inferior derecha).

Este hecho supone una dificultad importante a la hora de predecir la volatilidad de la serie con este primer conjunto de datos.

Por otra parte, la figura 66 muestra la serie de rentabilidades (figura superior) y la serie de sus cuadrados (figura inferior). Como puede observarse, parece existir un primer período que finaliza alrededor de la observación 1250 (primeros de diciembre de 1996), con un nivel de varianza inferior a la del período que va, desde dicha observación, hasta el final.

Por tanto, la serie parece poseer dos regímenes de volatilidad en sendos períodos diferentes, tanto en el nivel de la misma, como en su grado de asociación.

A la vista de estas particularidades, antes de evaluar los diferentes algoritmos con esta serie, hemos decidido comprobar la validez de la hipótesis de normalidad.

La tabla 36 muestra los resultados de la utilización de diferentes modelos ARMA-GARCH, con esta serie de rentabilidades utilizando, como muestra de entrenamiento, las primeras 1000, 2000 y 2500 observaciones, respectivamente, y muestra de test las restantes en cada caso.

En dichos modelos, se han considerado diferentes hipótesis para la distribución de las innovaciones: gausiana y t-Student con diferentes grados de libertad.

Dicha tabla muestra asimismo, las estimaciones de los parámetros del modelo GARCH (cuando hay más de dos parámetros en alguna componente de este modelo, se muestran los de magnitud más significativa), proporcionadas por cada modelo, y el menos logaritmo de la verosimilitud en la muestra de entrenamiento y en la de test.

Como puede observarse, la consideración de la hipótesis t-Student con 3 grados de libertad, produce mejores resultados en los dos primeros períodos ($n = 1000$ y $n = 2000$), con grandes diferencias en el primero, pero dichas mejoras se diluyen a medida que se incorporan observaciones de la zona de mayor varianza.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Modelo	Distrib.	n_train	-lnLkhdtrain	-lnLkhdtest	K	GARCH	ARCH
ARMA(0,0)-GARCH(1,1)	N(0,1)	1000	-3758.2562	-5123.2969	1.5732E-06	0.9173	0.0340
ARMA(0,0)-GARCH(1,1)	t-3		-3712.9712	-5340.2240	1.9766E-06	0.9298	0.0490
ARMA(0,0)-GARCH(1,1)	t-4		-3750.5777	-5226.1110	1.2791E-06	0.9368	0.0349
ARMA(0,0)-GARCH(1,1)	t-5		-3756.4744	-5214.8570	1.1474E-06	0.9371	0.0314
ARMA(0,0)-GARCH(1,1)	t-10		-3757.6999	-5155.7150	1.0311E-06	0.9369	0.0289
ARMA(0,0)-GARCH(1,1)	N(0,1)	2000	-6896.0795	-2192.0593	5.8554E-07	0.9305	0.0635
ARMA(0,0)-GARCH(2,5)	N(0,1)		-6899.3051	-2190.7520	1.3164E-06	[0.18;0.68]	[0.03;0.08]
ARMA(0,0)-GARCH(2,5)	t-3		-6821.2632	-2193.3340	3.4378E-06	[5E-6;0.88]	[0.03;0.05;
ARMA(0,0)-GARCH(1,1)	t-3		-6819.9892	-2193.6030	1.4226E-06	0.9488	0.0513
ARMA(0,0)-GARCH(1,1)	t-4		-6883.1855	-2191.5810	4.9852E-07	0.9544	0.0456
ARMA(0,0)-GARCH(1,1)	N(0,1)	2500	-8334.1866	-755.5590	6.2165E-07	0.9248	0.0716
ARMA(0,0)-GARCH(2,5)	N(0,1)		-8335.9562	-756.1882	1.1302E-06	[0.23;0.64]	[0.05;0.08]
ARMA(0,0)-GARCH(2,5)	t-3		-8245.4853	-755.6050	2.9863E-06	[0.32;0.58]	[0.03;0.06]
ARMA(0,0)-GARCH(1,1)	t-3		-8244.0060	-755.0030	1.6856E-06	0.9460	0.0540
ARMA(0,0)-GARCH(1,1)	t-4		-8320.6504	-755.0980	6.5908E-07	0.9467	0.0533

Tabla 36. Resultados de diferentes modelos ARMA-GARCH con diferentes hipótesis sobre la distribución (Gausiana o t-Student con diferentes grados de libertad), ante la serie de rentabilidades S&P. Se muestra: el modelo empleado, la distribución, el número de observaciones de entrenamiento, y los criterios de error en la muestra de entrenamiento y de test.

Asimismo, salvo con todos los datos, no parece que produzca grandes beneficios considerar un número de retardos mayor que dos en la estructura de varianzas como podría sugerir la estructura de correlación de ε_t^2 para $n = 2000$ y 2500.

Con esta serie, hemos utilizado el algoritmo WildBoostGARCH mencionado al principio de esta sección, con los resultados que se muestran en la tabla 37.

Dicha tabla muestra que el algoritmo aporta beneficios en el primer período, en el que las observaciones no sugieren el incremento futuro de varianza. En el resto, sus resultados coinciden con el modelo GARCH correspondiente. Si la muestra de validación (p. ej. de un período posterior) presenta una mayor varianza que la muestra de entrenamiento, el algoritmo la recogerá sugiriendo mayores varianzas en períodos futuros.

Es interesante observar la similitud de resultados, en el primer período, entre el WildBoostGARCH y el modelo GARCH con distribución t-Student con 3 grados de libertad. Ello sugiere que, la verosimilitud t-Student produce un efecto similar a la acción del *wild boosting*.

El problema de la utilización del WildBoostGARCH, es la necesidad de disponer de tal muestra de validación, de tal forma que, el algoritmo mejore los resultados que produciría un modelo GARCH estimado utilizando la totalidad de dichos datos. Esto es cierto también para cualquier algoritmo *boosting*, (y para el propio modelo GARCH, aunque los modelos más sencillos como el GARCH(1,1) suelen producir buenos resultados).

Una solución para el *boosting*, es la introducción de mecanismos de regularización que ralenticen su posible sobreajuste. Esta vía se explorará con este algoritmo en el futuro. De cualquier forma, cuando se utilizaron las primeras $n = 1000$ observaciones y, como muestra de validación, las siguientes $n_{\text{val}} = 1000$, el algoritmo mejoró los resultados de un modelo GARCH(1,1) esti-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Modelo	Distribución	n_train	BestT	-lnLkhdtrain	-lnLkhdtest
WildBoostGARCH(1,1)	N(0,1)	1000	4	-3518.5306	-5344.7509
WildBoostGARCH(1,1)	t-3		1	-3227.7863	-5340.2247
WildBoostGARCH(1,1)	t-4		2	-3442.3589	-5316.3130
WildBoostGARCH(1,1)	t-5		2	-3417.7742	-5336.3145
WildBoostGARCH(1,1)	t-10		3	-3597.8675	-5335.7415
WildBoostGARCH(1,1)	N(0,1)	2000	1	-6665.9478	-2192.0593
WildBoostGARCH(1,1)	t-3		1	-6819.9892	-2193.6030
WildBoostGARCH(1,1)	t-4		1	-6883.1855	-2191.5810
WildBoostGARCH(1,1)	N(0,1)	2500	1	-8334.1866	-755.5590
WildBoostGARCH(1,1)	t-3		1	-8244.0060	-755.0030
WildBoostGARCH(1,1)	t-4		1	-8320.6504	-755.0980

Tabla 37. Resultados del algoritmo WildBoostGARCH con la serie de rentabilidades S&P. El algoritmo no produce beneficios cuando se utilizan observaciones de la zona de mayor varianza.

mado con los $n + n_{\text{val}}$ datos, sobre una muestra de test constituida por los últimos $n_{\text{test}} = 772$ datos.

La tabla 38 muestra los resultados de diferentes algoritmos presentados al principio de la sección, con esta serie de rentabilidades, distinguiendo entre los tres períodos analizados anteriormente. En cada caso, se utiliza como muestra de test las observaciones posteriores hasta 2772, no utilizadas en el entrenamiento. Todos los modelos utilizan un sólo retardo.

Sobre dicha tabla, resaltamos lo siguiente:

1. En el primer período, destaca el algoritmo BoostARMA(4,1)-GARCH(1,1) con inicio GARCH(1,1) que mejora de forma muy significativa los resultados de los demás modelos.

2. Los algoritmos que utilizan redes neuronales, presentan grandes problemas con esta serie pues, aparecen frecuentes problemas de singularidad o mal condicionamiento durante el entrenamiento, y gran inestabilidad en función de las condiciones iniciales aleatorias. Estas situaciones son típicas cuando los datos no poseen información suficiente para la estructura del estimador. Los modelos de este tipo que se muestran, son los que produjeron resultados al menos comparables.

3. En general, con esta serie, el inicio de los algoritmos *boosting* con un GARCH(1,1) produjo beneficios. El tipo de inicio es fundamental con esta serie.

En general, habría que concluir con esta serie, que los modelos utilizados no son los más adecuados. Será necesario evaluar otros modelos que consideren mayor horizonte temporal de dependencia, como los GARCH fraccionarios (el gráfico superior derecha de la figura 65, de la autocorrelación de la serie completa ya lo sugería).

La figura 67 muestra las dificultades con las que tienen que enfrentarse los algoritmos para detectar alguna estructura en la varianza (con un sólo retardo): con $n = 1000$ (primer período de la serie) por que no parece existir ninguna definida, y con $n = 2500$, porque parece existir más de una.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Modelo	n_train	IniGH	BestT	-lnLkhdtrain	-lnLkhdtest
ARMA(5,2) sobre y^2	1000			-3678.6228	-5085.9550
KR - 1 delay sobre y^2				-3749.6556	-3139.4448
ARMA(0,0)-GARCH(1,1)				-3758.2562	-5123.2969
ARMA(0,0)-GARCH(5,1)				-3758.2634	-5119.7196
MLP(1)-GARCH(1,1)				-3757.6171	-5123.4121
MLP(0)-GARCHMLP(1)				-3747.1732	-3122.8300
MLP(0)-GARCHMLP(2)				-3747.6759	-3297.4872
MLP(1)-GARCHMLP(1)				-3747.0794	-3124.1670
BoostARMA(1,0)-GARCH(1,1)		si	3	-3758.3761	-5124.4690
BoostARMA(4,1)-GARCH(1,1)		si	8	-3747.6664	-5194.5630
ARMA(20,2) sobre y^2	2000			-6721.6468	-2094.7527
KR - 1 delay sobre y^2				-6680.9046	-1923.6697
ARMA(0,0)-GARCH(1,1)				-6896.0795	-2192.0593
MLP(1)-GARCH(1,1)				-6895.9389	-2192.4465
MLP(0)-GARCHMLP(1)				-6684.1626	-1954.4320
MLP(0)-GARCHMLP(3)				-6696.7214	-1956.6060
MLP(1)-GARCHMLP(1)				-6683.7621	-1948.9246
BoostARMA(1,0)-GARCH(1,1)		si	3	-6896.0533	-2192.4088
BoostARMA(4,1)-GARCH(1,1)		si	2	-6896.0645	-2192.4259
ARMA(20,2) sobre y^2	2500			-7974.0617	-696.5552
ARMA(0,0)-GARCH(1,1)				-8334.1866	-755.5590
MLP(3) - 1 delay sobre y^2				-8058.0094	-685.0106
KR - 1 delay sobre y^2				-8042.3015	-678.1963
RBF(1)-GARCH(1,1)				-8333.6211	-755.8356
MLP(1)-GARCH(1,1)				-8332.6821	-755.9893
MLP(0)-GARCHMLP(1)				-8034.1210	-680.6063
MLP(0)-GARCHMLP(3)				-8053.2855	-684.7160
MLP(1)-GARCHMLP(1)				-8031.8760	-680.3145
BoostMLP(1) - 1 delay - y^2		no	4	-7954.1485	-733.2399
BoostMLP(1) - 1 delay - y^2		si	15	-8334.1866	-755.5591
BoostARMA(1,0)-GARCH(1,1)		si	3	-8334.1527	-755.6701
BoostARMA(4,1)-GARCH(1,1)		si	8	-8283.8592	-756.6817

Tabla 38. Resultados de diferentes algoritmos ante la serie de rentabilidades S&P. Se muestra: el modelo empleado, el número de observaciones de la muestra de entrenamiento, la mejor iteración de los algoritmos boosting, y el menos logaritmo de la verosimilitud en la muestra de entrenamiento y de test. Todos los modelos utilizan un sólo retardo.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

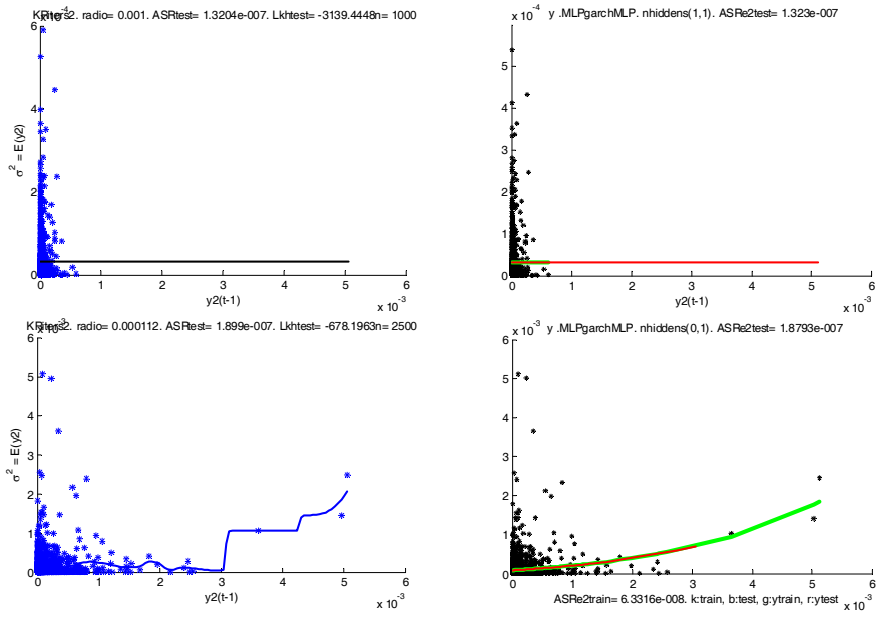


Figura 67. Arriba: ajuste de la estructura de la varianza (con un retardo) del modelo Nadaraya-Watson (izquierda) y del modelo MLP(0)-GARCHMLP(1) (derecha), con $n = 1000$. Abajo: lo mismo para $n = 2500$.

Lo anterior se manifiesta igualmente en la influencia de la iteración inicial de las técnicas *boosting*. Si por ejemplo, con $n = 1000$, se toma la varianza de toda la serie $(1{,}14 \times 10^{-4})$ como estimación inicial en lugar de la varianza de la muestra de entrenamiento $(3{,}22 \times 10^{-5})$, el BoostARMA(4,1)-GARCH(1,1) sin inicio GARCH, puede pasar de una pérdida en test de $-3121{,}4937$ a otra de $-5257{,}4178$. Teniendo en cuenta que el GARCH(1,1) en esta muestra produce una pérdida de $-5123{,}2969$, la mejora potencial del *boosting* en esta serie con un inicio adecuado puede ser importante.

Serie de Precios S&P

Finalmente, presentamos los resultados relativos a la serie de precios Standard & Poor's. La figura 68 muestra las funciones de autocorrelación y autocorrelación parcial de la serie (izquierda), y el aspecto de la propia serie (derecha).

Como puede observarse, la serie presenta una asociación lineal clara de un sólo retardo.

Los resultados de los diferentes algoritmos evaluados se presentan en la tabla 39. En ella, la verosimilitud de los modelos ARMA, RBF y MLP individuales se calculó estimando la varianza como la media de los errores al cuadrado en la muestra de entrenamiento, y no mediante el cuadrado del error en cada t, pues, para dichos algoritmos, la serie se supone homocedástica.

En relación con esta tabla, destacamos lo siguiente:

1. Las dificultades, mencionadas para el caso de rentabilidades, de los algoritmos basados en redes neuronales, se mantienen con esta serie. Las razones esta vez, se deben a la excesiva linealidad de la asociación existente con un retardo.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

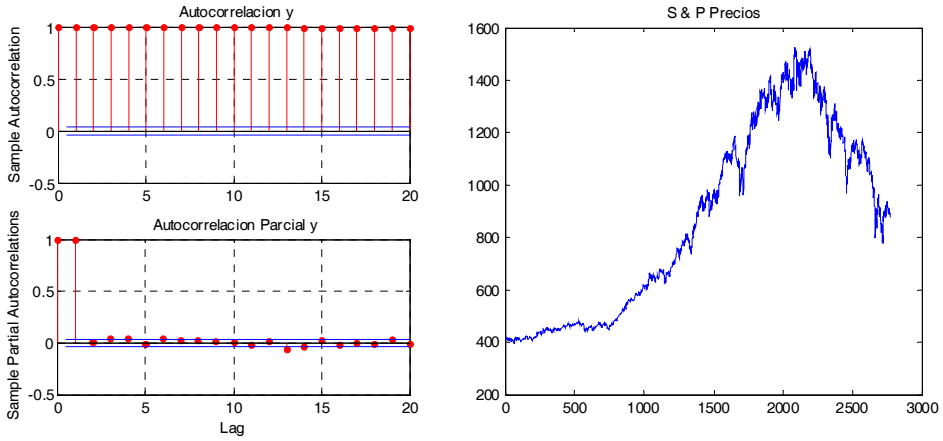


Figura 68. Izquierda: funciones de autocorrelación y autocorrelación parcial de la serie de precios S&P. Derecha: serie S&P.

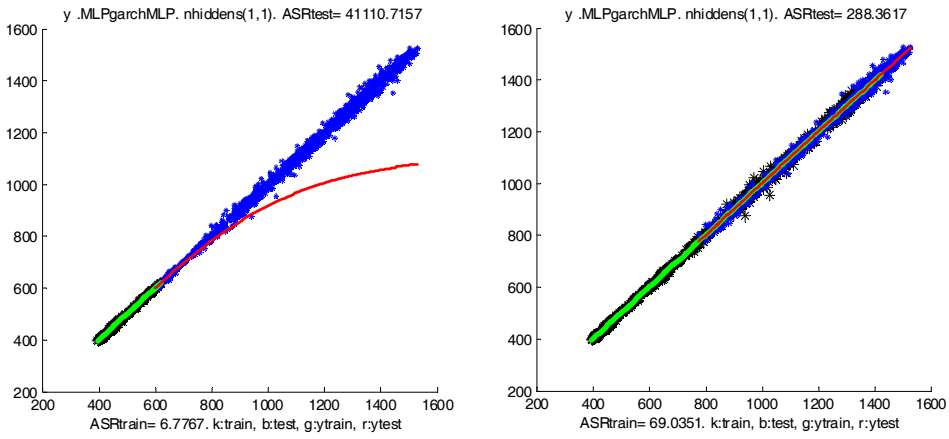


Figura 69. Ajuste de la tendencia con un retardo del algoritmo MLP(1)-GARCHMLP(1) en la serie S&P precios para $n = 1000$ (izquierda) y $n = 2000$ (derecha).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

La figura 69 ilustra esta circunstancia: la relación entre Y_t e Y_{t-1} es completamente lineal con correlación unidad como muestra también la autocorrelación parcial (figura 68 izquierda abajo), (el modelo AR(1) produce un coeficiente de $1{,}0003$).

Esta situación no es la ideal para una red neuronal: cuando la muestra de entrenamiento está constituida por los $n = 1000$ primeros datos, la red no posee información sobre el futuro (muestra de test y predicción de la red, en otro color) y adopta la forma que le permite su estructura no lineal. En la figura de la derecha, la muestra de entrenamiento ($n = 2000$) ya cubre todo el rango de la serie completa y revela a la red su estructura lineal en la totalidad de dicho rango, con lo que el modelo se ajusta perfectamente.

2. En general, hay pocas diferencias entre los mejores modelos. Para los tres tamaños muestrales, destaca el ARMA(1,0)_GARCH(1,1)-ARMA(1,0)_GARCH(1,1) siempre con inicio AR(1)-GARCH(1,1). Estos algoritmos *boosting* de corte lineal y menor complejidad son una buena alternativa cuando los modelos más complejos basados en redes neuronales son contraproducentes.

   Salvo para $n = 1000$, destacan también los algoritmos MLP(m)-GARCH(1,1) (tendencia MLP con modelo GARCH para la varianza, simultáneamente estimados mediante verosimilitud).

Conclusiones

En los epígrafes anteriores, se ha presentado y evaluado en un conjunto de datos simulados y reales, una batería de nuevos algoritmos para la predicción de series de precios y de rentabilidades: unos basados en la combinación de redes neuronales con modelos GARCH (clásicos o, a su vez, basados en redes neuronales), y otros, basados en técnicas *boosting* aplicadas a redes neuronales y a modelos autorregresivos y heterocedásticos.

Como resultado de la evaluación de dichos algoritmos, podríamos extraer las siguientes conclusiones más importantes:

1. En general, no existe un algoritmo idóneo para todas las situaciones. Esta pretensión de un algoritmo universal debería sustituirse por la utilización de los algoritmos como ayuda al problema de identificación y al contraste de hipótesis.

   En este sentido, un algoritmo fuertemente no lineal con buenos resultados *testifica* la presencia de una serie no lineal y viceversa.

2. Los modelos MLP-GARCH o RBF-GARCH que combinan redes neuronales para la tendencia con un modelo GARCH mejoran radicalmente los resultados de los modelos ARMA-GARCH cuando se presente una función de tendencia no lineal. En situaciones en las que, aparentemente, no existe tendencia global, pero están presentes mecanismos de tendencia locales, este tipo de algoritmos puede producir grandes beneficios. Es el caso de la serie $y_t \sin y_t$ tratada en primer lugar.

3. Los modelos que estiman la tendencia y la estructura de varianza simultáneamente mediante la máxima verosimilitud (MLP-GARCHMLP), son apropiados cuando se presuma un gran nivel de no linealidad en la serie (caso de la serie *threshold*). Cuando no se requiera tal complejidad o cuando la serie sea muy heterogénea (caso de la serie S&P), estos algoritmos resultan excesivamente inestables. El éxito de estos modelos se basa en una adecuada esti-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Modelo	n_train	Inicio	BestT	ASRtrain	ASRtest	-lnLkhdtrain	-lnLkhdtest
ARMA(1,0)	1000			6.8094	199.8895	2378.0921	29336.3723
ARMA(1,1)				6.8103	199.8905	2378.1582	29333.1781
ARMA(1,0)-GARCH(1,1)				6.7930	204.7442	2387.0034	7840.6258
RBF(2)				6.7795	210.2556	2375.8904	30801.9758
MLP(1)-GARCH(1,1)				6.7772	41106.7771	2517.1972	11352.4955
MLP(1)-GARCHMLP(1)				6.7767	41110.7157	2374.6316	5.2362E+06
MLP(1)-GARCHMLP(2)				6.7772	41106.7767	2374.1657	1.4463E+07
BoostAR_GARCH-AR_GARCH		--	3	6.7929	204.9775	4832.2132	8660.4864
BoostAR_GARCH-AR_GARCH		AR-GARCH		6.7862	204.5942	2385.9359	7835.3738
BoostMLP(1)-GARCHMLP(1)		--	2	6.7748	8596.5812	4832.2107	11599.2216
BoostMLP(1)-GARCHMLP(1)		AR-GARCH	1	=AR-GARCH	=AR-GARCH	=AR-GARCH	=AR-GARCH
ARMA(1,0)	2000			69.2304	289.1551	7075.3173	3957.2809
ARMA(1,1)				69.2297	289.3331	7075.3066	3958.2869
ARMA(1,0)-GARCH(1,1)				69.2941	290.5677	6101.8132	3268.1690
RBF(4)				68.9496	289.1844	7071.2524	3962.4428
MLP(1)				69.0300	288.4663	7072.4182	3956.9913
MLP(1)-GARCH(1,1)				69.0349	288.5923	6100.1135	3266.3067
MLP(1)-GARCHMLP(1)				69.0351	288.3617	6777.1800	3817.7284
MLP(1)-GARCHMLP(2)				69.0373	288.3440	6769.2180	3807.1284
BoostAR_GARCH-AR_GARCH		--		69.2524	289.2858	13317.0147	5141.2506
BoostAR_GARCH-AR_GARCH		AR-GARCH		69.2606	290.3997	6097.3996	3267.7400
BoostMLP(1)-GARCHMLP(1)		--	3	69.0377	287.5732	13609.2047	5253.8050
BoostMLP(1)-GARCHMLP(1)		AR-GARCH	1	=AR-GARCH	=AR-GARCH	=AR-GARCH	=AR-GARCH
ARMA(1,0)	2500			119.6704	226.6968	9528.2729	1158.3067
ARMA(1,1)				119.6870	226.6120	9528.4464	1158.1936
ARMA(1,0)-GARCH(1,1)				119.7364	227.1694	8250.1961	1117.3174
RBF(4)				118.9801	228.5083	9521.0416	1161.0854
MLP(2)				118.9284	227.7860	9520.4985	1160.3138
MLP(1)-GARCH(1,1)				119.0785	230.3028	8246.5795	1119.1604
MLP(2)-GARCH(1,1)				118.9375	228.2078	8247.7593	1117.7611
MLP(1)-GARCHMLP(1)				119.1619	231.5667	9209.9536	1173.8890
MLP(1)-GARCHMLP(2)				119.1492	231.4399	9185.4115	1174.5820
BoostAR_GARCH-AR_GARCH		--	2	119.6969	227.0656	17103.6948	1860.9865
BoostAR_GARCH-AR_GARCH		AR-GARCH	4	119.6933	227.1407	8245.6188	1117.2943
BoostMLP(2)-GARCHMLP(2)		--	3	119.0766	225.2870	17103.6909	1860.9842
BoostMLP(2)-GARCHMLP(2)		AR-GARCH	1	=AR-GARCH	=AR-GARCH	=AR-GARCH	=AR-GARCH

Tabla 39. Resultados de diferentes algoritmos ante la serie de precios S&P. Se muestra: el modelo empleado, el número de observaciones de la muestra de entrenamiento, la mejor iteración de los algoritmos boosting, y el error cuadrático y el menos logaritmo de la verosimilitud, en la muestra de entrenamiento y de test. Todos los modelos utilizan un sólo retardo.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

mación de la tendencia para que la estimación de la varianza con los residuos, sea efectiva. Es necesario seguir evaluando estos algoritmos con series no lineales en tendencia y en varianza.

4. Los algoritmos *boosting* suelen aportar beneficios en cualquier situación: cuando la estructura de la serie es excesivamente sencilla (o demasiado heterogénea) como para utilizar modelos complejos de excesiva varianza como hipótesis base, pueden utilizarse hipótesis base de corte lineal como BoostARMA-GARCH, para evitar la inestabilidad del modelo.

5. El Gradient-Boost con técnicas de ajuste del gradiente que consideran la estructura temporal de dependencia (p. ej. ARMA-GARCH) producen buenos resultados. Ello sugiere que hay que tratar al gradiente no como un simple vector de dimensión n sino como una serie temporal en sí misma, que hereda la relación de dependencia temporal que posee la serie original y/o sus varianzas. Las técnicas clásicas de regresión, si no poseen flexibilidad suficiente (árboles sencillos, etc.), pueden no recoger dicha estructura.

6. No puede afirmarse categóricamente [Audrino y Bühlmann 2003], que el inicio de los algoritmos *boosting* con un modelo GARCH, mejorará los resultados de este tipo de técnicas. Ello depende de la serie en cuestión y de la técnica utilizada. Con la serie de rentabilidades S&P, fué beneficioso. En el resto de las series, el resultado dependió de la hipótesis base utilizada en el *boosting*, pero, en general, sin producir los mejores resultados.

7. El modelo de distribución utilizado en los modelos GARCH, se reveló muy importante en la serie S&P. El WildBoostGARCH realiza un efecto similar al de la distribución t-Student en dicha serie pues, sus colas más pesadas que las de la gausiana, hacen que el modelo postule mayor varianza para la serie.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

ANEXOS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Anexo A

Clasificación y Discriminante

Discriminante Lineal de Fisher

Dos Clases

Dado un problema de delimitar las dos posibles clases $\mathcal{C}_j$, $j = 1, 2$, en los que se clasifican los datos de una muestra $\{\mathbf{x}_i\}_{i=1}^n$ con $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, $i = 1 : n$, el discriminante lineal de Fisher $f_{\bar{\mathbf{w}}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ con $\bar{\mathbf{w}} = (\mathbf{w}^T, w_0)^T \in \mathbb{R}^{d+1}$, sin hipótesis sobre la distribución, se obtiene ([Seber 1984, p. 283 y 293], [Duda *et al.* 2001, p. 117]) como resultado de proyectar los datos en la dirección, definida por $\mathbf{w} \in \mathbb{R}^d$, que maximiza la distancia entre las medias de las clases y minimiza la varianza dentro de ellas, es decir, de maximizar, con $\|\mathbf{w}\| = a$ constante, el criterio:

$$\begin{aligned} L(\mathbf{w}) &= \frac{(\bar{u}_2 - \bar{u}_1)^2}{s_1^2 + s_2^2} = \frac{[\sum_{i\in\mathcal{C}_2} \mathbf{w}^T\mathbf{x}_i - \sum_{i\in\mathcal{C}_1} \mathbf{w}^T\mathbf{x}_i]^2}{\sum_{i\in\mathcal{C}_1}[\mathbf{w}^T(\mathbf{x}_i - \mathbf{m}_1)]^2 + \sum_{i\in\mathcal{C}_1}[\mathbf{w}^T(\mathbf{x}_i - \mathbf{m}_2)]^2} \\ &= \frac{[\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1)]^2}{\mathbf{w}^T(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{w}} = \frac{\mathbf{w}^T\mathbf{S}_E\mathbf{w}}{\mathbf{w}^T\mathbf{S}_I\mathbf{w}} \end{aligned} \tag{A.1}$$

donde:

- para cada clase $\mathcal{C}_j$, $j = 1 : 2$, $\bar{u}_j = \frac{1}{n}\sum_{i\in\mathcal{C}_j} \mathbf{w}^T\mathbf{x}_i$ y s_j^2 son la media y la varianza, respectivamente, de las proyecciones de los datos sobre el correspondiente hiperplano.
- $\mathbf{S}_j = \sum_{i\in\mathcal{C}_j}(\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x} - \mathbf{m}_j)^T$ es, salvo factor, la matriz de covarianza de los datos en cada clase, siendo $\mathbf{m}_j$ su media, $j = 1, 2$, y $\mathbf{S}_I = \mathbf{S}_1 + \mathbf{S}_2$ es la denominada varianza dentro de las clases.
- $\mathbf{S}_E = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \propto \sum_{j=1}^2(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T$ es la denominada varianza entre clases, donde $\mathbf{m} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$.

Derivando la ecuación (A.1) respecto de $\mathbf{w}$ e igualando a cero, se obtiene inmediatamente la condición de óptimo, de la que se deduce que:

$$(\mathbf{w}^T\mathbf{S}_E\mathbf{w})\mathbf{S}_I\mathbf{w} = (\mathbf{w}^T\mathbf{S}_I\mathbf{w})\mathbf{S}_E\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1) \Rightarrow \mathbf{w} \propto \mathbf{S}_I^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

resultado que, también, se obtiene mediante la ecuación (G.9) del Anexo G que, además, indica que el óptimo $\mathbf{w}$ es el autovector asociado al máximo autovalor de $\mathbf{S}_I^{-1}\mathbf{S}_E$.

Una nueva observación $\mathbf{x}$, se asigna a la clase $\mathcal{C}_j$ cuya media es más cercana a la proyección $\mathbf{w}^T\mathbf{x}$ del nuevo dato, donde el punto medio $\frac{1}{2}\mathbf{w}^T(\mathbf{m}_1 + \mathbf{m}_2) = \mathbf{w}^T\mathbf{m}$ actúa como frontera. Por tanto, la regla discriminante, asigna a $\mathbf{x}$ la clase $\mathcal{C}_1$ si $\mathbf{w}^T(\mathbf{x} - \mathbf{m}) = \mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m} \geq 0$, y $\mathcal{C}_2$ en caso contrario, con lo que $w_0 = -\mathbf{w}^T\mathbf{m}$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Varias Clases

La solución anterior se generaliza a c clases [Duda *et al.* 2001, p. 121] con $d \geq c$ definiendo $\mathbf{m}_j$ y $\mathbf{S}_j$, $j = 1 : c$, como antes y $\mathbf{S}_I = \sum_{j=1}^{c} \mathbf{S}_j$. Además, teniendo en cuenta que la varianza total (salvo factor) puede descomponerse como:

$$\begin{aligned}
\mathbf{S}_T &= \sum_{j=1}^{c} \sum_{i \in \mathcal{C}_j} (\mathbf{x}_i - \mathbf{m}_j + \mathbf{m}_j - \mathbf{m})(\mathbf{x}_i - \mathbf{m}_j + \mathbf{m}_j - \mathbf{m})^T \\
&= \sum_{j=1}^{c} \sum_{i \in \mathcal{C}_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T + \sum_{j=1}^{c} \sum_{i \in \mathcal{C}_j} (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \\
&= \mathbf{S}_I + \sum_{j=1}^{c} n_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T
\end{aligned}$$

donde n_j es el número de ejemplos de la muestra en la clase $\mathcal{C}_j$, es natural definir como varianza entre clases el segundo término de la expresión anterior:

$$\mathbf{S}_E = \sum_{j=1}^{c} n_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \tag{A.2}$$

de tal forma que $\mathbf{S}_T = \mathbf{S}_I + \mathbf{S}_E$.

Ahora, la proyección de los datos se realiza sobre un hiperplano de dimensión $c - 1$ determinado por la matriz $\mathbf{W}_{d \times (c-1)}$ con columnas $\mathbf{w}_j$, $j = 1, ..., c - 1$, de tal forma que los datos proyectados son de la forma $\mathbf{u}_i = \mathbf{W}^T \mathbf{x}_i$, $i = 1 : n$.

Teniendo en cuenta que los autovalores de una matriz de covarianza dan una idea de la magnitud de la varianza en sus direcciones principales y que su determinante es el producto de dichos autovalores, de forma análoga a la primera expresión de la ecuación (A.1), uno entre los posibles criterios a maximizar es [Duda *et al.* 2001, p. 123], (véase también [Johnson y Wichern 1992, p. 561] y [Fukunaga 1990]):

$$L(\mathbf{W}) = \frac{\left|\mathbf{S}_E^P\right|}{\left|\mathbf{S}_I^P\right|} = \frac{\left|\mathbf{W}^T \mathbf{S}_E \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{S}_I \mathbf{W}\right|}$$

donde $\mathbf{S}_E^P = \sum_{j=1}^{c} n_j (\bar{\mathbf{u}}_j - \bar{\mathbf{u}})(\bar{\mathbf{u}}_j - \bar{\mathbf{u}})^T$ y $\mathbf{S}_I^P = \sum_{j=1}^{c} \sum_{i \in \mathcal{C}_j} (\mathbf{u} - \bar{\mathbf{u}}_j)(\mathbf{u} - \bar{\mathbf{u}}_j)^T$ con $\bar{\mathbf{u}}_j = \mathbf{W}^T \mathbf{m}_j$ y $\bar{\mathbf{u}} = \mathbf{W}^T \mathbf{m}$, son los análogos a $\mathbf{S}_E$ y $\mathbf{S}_I$ para los datos proyectados, y la segunda igualdad se prueba fácilmente.

Puede demostrarse ([Johnson y Wichern 1992, p. 561], [Duda *et al.* 2001], [Fukunaga 1990]), que las columnas $\mathbf{w}_j$ de la matriz óptima $\mathbf{W}$ son los $c - 1$ autovectores correspondientes a los $c - 1$ mayores autovalores de $\mathbf{S}_I^{-1} \mathbf{S}_E$, y definen $c - 1$ fronteras discriminantes entre las c clases.

Relación entre Diversos Criterios de Clasificación

Sea una población X con valores en $\mathcal{X} \subset \mathbb{R}^d$, cuyos elementos están asignados a dos posibles clases $\mathcal{C}_j$, $j = 1, 2$, siendo $\mathbb{P}(\mathcal{C}_j)$ la proporción de ejemplos asignados a la clase $\mathcal{C}_j$, $j = 1, 2$, con $\mathbb{P}(\mathcal{C}_1) + \mathbb{P}(\mathcal{C}_2) = 1$, y sea $\mathsf{p}_j(\mathbf{x}) = \mathsf{p}_{\mathsf{X}/\mathcal{C}_j}(\mathbf{x})$ la función de masa de probabilidad o de densidad de X si X procede de la clase $\mathcal{C}_j$.

Sean $\mathcal{X}_1, \mathcal{X}_2 \subset \mathcal{X}$ dos regiones de $\mathcal{X}$ exhaustivas ($\mathcal{X}_1 \cup \mathcal{X}_2 = \mathcal{X}$) y excluyentes ($\mathcal{X}_1 \cap \mathcal{X}_2 = \varnothing$), y supongamos una regla de clasificación que asigna cada ejemplo $\mathsf{X} = \mathbf{x}$ a la clase $\mathcal{C}_1$ si $\mathbf{x} \in \mathcal{X}_1$ y a la clase $\mathcal{C}_2$ si $\mathbf{x} \in \mathcal{X}_2$.

Minimización de la Probabilidad de Error

Para una regla de clasificación, la probabilidad de error es:

$$\mathbb{P}(\text{error}) = \mathbb{P}(\text{asignar } \mathbf{x} \text{ a } \mathcal{C}_2 | \mathbf{x} \text{ está en } \mathcal{C}_1)\mathbb{P}(\mathcal{C}_1) + \mathbb{P}(\text{asignar } \mathbf{x} \text{ a } \mathcal{C}_1 | \mathbf{x} \text{ está en } \mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)$$

Se verifica el siguiente lema donde denotamos la diferencia de conjuntos mediante $\backslash$, es decir, dados A, B conjuntos: $A \backslash B = \{u \in A : u \notin B\}$.

Lema 72 *[Seber 1984, lema 6.1] Para cualquier función $f : \mathcal{X} \to \mathbb{R}$, se verifica que $\mathcal{X}_{01} = \{\mathbf{x} : f(\mathbf{x}) < 0\}$ minimiza $\int_{\mathcal{X}_1} f(\mathbf{x})d\mathbf{x}$ con respecto a $\mathcal{X}_1$.*

Demostración. Sea $\mathcal{X}_{02} = \{\mathbf{x} : f(\mathbf{x}) \geq 0\}$. $\forall \mathcal{X}_1 \subset \mathcal{X}$ se verifica $\mathcal{X}_1 = (\mathcal{X}_1 \cap \mathcal{X}_{01}) \cup (\mathcal{X}_1 \cap \mathcal{X}_{02})$. Así:

$$\begin{aligned}\int_{\mathcal{X}_1} f(\mathbf{x})d\mathbf{x} &= \int_{\mathcal{X}_1 \cap \mathcal{X}_{01}} f(\mathbf{x})d\mathbf{x} + \int_{\mathcal{X}_1 \backslash (\mathcal{X}_1 \cap \mathcal{X}_{01})} f(\mathbf{x})d\mathbf{x} \geq \int_{\mathcal{X}_1 \cap \mathcal{X}_{01}} f(\mathbf{x})d\mathbf{x} \\ &= \int_{\mathcal{X}_{01}} f(\mathbf{x})d\mathbf{x} - \int_{\mathcal{X}_{01} \backslash (\mathcal{X}_1 \cap \mathcal{X}_{01})} f(\mathbf{x})d\mathbf{x} \geq \int_{\mathcal{X}_{01}} f(\mathbf{x})d\mathbf{x}\end{aligned}$$

■

Además, si $B = \{f(\mathbf{x}) = 0\}$, entonces $\forall C \subset B$, $\int_{\mathcal{X}_{01} \cup A} f(\mathbf{x})d\mathbf{x} = \int_{\mathcal{X}_{01}} f(\mathbf{x})d\mathbf{x}$, con lo que $\mathcal{X}_{01}$ no es único y la frontera puede asignarse a $\mathcal{X}_{01}$ o a $\mathcal{X}_{02}$ (por supuesto, si X es continua $\mathbb{P}_{\mathsf{X}}(B) = 0$).

Para la regla de clasificación mencionada anteriormente, la probabilidad de error es:

$$\begin{aligned}\mathbb{P}(\text{error}) &= \mathbb{P}(\mathcal{C}_1)[1 - \int_{\mathcal{X}_1} \mathsf{p}_1(\mathbf{x})d\mathbf{x}] + \mathbb{P}(\mathcal{C}_2)[\int_{\mathcal{X}_1} \mathsf{p}_2(\mathbf{x})d\mathbf{x}] \\ &= \mathbb{P}(\mathcal{C}_1) + \int_{\mathcal{X}_1} [\mathbb{P}(\mathcal{C}_2)\mathsf{p}_2(\mathbf{x}) - \mathbb{P}(\mathcal{C}_1)\mathsf{p}_1(\mathbf{x})]d\mathbf{x}\end{aligned}$$

Aplicando el lema anterior, dicha probabilidad se minimiza si se elige $\mathcal{X}_1 = \mathcal{X}_{01} = \{\mathbf{x} : [\mathbb{P}(\mathcal{C}_2)\mathsf{p}_2(\mathbf{x}) - \mathbb{P}(\mathcal{C}_1)\mathsf{p}_1(\mathbf{x})] < 0\}$, es decir, se utiliza la regla de clasificación que asigna $\mathbf{x}$ a la clase $\mathcal{C}_1$ si:

$$\frac{\mathsf{p}_1(\mathbf{x})}{\mathsf{p}_2(\mathbf{x})} > \frac{\mathbb{P}(\mathcal{C}_2)}{\mathbb{P}(\mathcal{C}_1)}$$

y a la clase $\mathcal{C}_2$ en caso contrario, (y, en caso de igualdad, vale una asignación arbitraria).

Maximización de la Probabilidad A Posteriori. Regla de Bayes

La regla de Bayes maximiza la probabilidad a posteriori, es decir, asigna $\mathbf{x} \in \mathcal{X}$ a la clase $\mathcal{C}_1$ si $\mathbb{P}(\mathcal{C}_1|\mathbf{x}) > \mathbb{P}(\mathcal{C}_2|\mathbf{x})$, y a la clase $\mathcal{C}_2$ en caso contrario. Pero:

$$\mathbb{P}(\mathcal{C}_1|\mathbf{x}) > \mathbb{P}(\mathcal{C}_2|\mathbf{x}) \Leftrightarrow \frac{\mathbb{P}(\mathcal{C}_1)\mathsf{p}_1(\mathbf{x})}{\mathsf{p}_{\mathsf{X}}(\mathbf{x})} > \frac{\mathbb{P}(\mathcal{C}_2)\mathsf{p}_2(\mathbf{x})}{\mathsf{p}_{\mathsf{X}}(\mathbf{x})} \Leftrightarrow \frac{\mathsf{p}_1(\mathbf{x})}{\mathsf{p}_2(\mathbf{x})} > \frac{\mathbb{P}(\mathcal{C}_2)}{\mathbb{P}(\mathcal{C}_1)}$$

con lo que dicha regla es equivalente a la que minimiza la probabilidad de error.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Maximización del Ratio de Verosimilitud

La regla que maximiza el ratio de verosimilitud asigna $\mathbf{x} \in \mathcal{X}$ a la clase $\mathcal{C}_1$ si $\mathsf{p}_1(\mathbf{x})/\mathsf{p}_2(\mathbf{x}) > 1$, que resulta un caso especial de las anteriores si $\mathbb{P}(\mathcal{C}_1|\mathbf{x}) = \mathbb{P}(\mathcal{C}_2|\mathbf{x}) = 1/2$.

En hipótesis de normalidad en ambas clases con igualdad de matriz de covarianzas:

$$\mathsf{p}_j(\mathbf{x}) = (2\pi)^{-d/2}\,|\Sigma|^{-1/2}\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T\Sigma(\mathbf{x}-\boldsymbol{\mu}_j)],\ j=1,2$$

se tiene:

$$\begin{aligned}\frac{\mathsf{p}_1(\mathbf{x})}{\mathsf{p}_2(\mathbf{x})} &= \exp\{\frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-(\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)]\}\\ \Rightarrow\ \ln\frac{\mathsf{p}_1(\mathbf{x})}{\mathsf{p}_2(\mathbf{x})} &= \frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-(\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)]\end{aligned}$$

que es la diferencia de las distancias de Mahalanobis del vector $\mathbf{x}$ a clasificar, respecto a las medias de las clases, y puede escribirse:

$$\begin{aligned}&= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\Sigma^{-1}\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\\ &= \mathbf{w}^T[\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)]\\ &= \mathbf{w}^T(\mathbf{x}-\bar{\boldsymbol{\mu}})\end{aligned}$$

donde $\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ es la dirección discriminante de Fisher (en su versión poblacional) y:

$$\bar{\mu} = \frac{1}{2}\mathbf{w}^T(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) = \frac{1}{2}(\bar{\mu}_1+\bar{\mu}_2)$$

siendo $\bar{\mu}_j = \mathbf{w}^T\boldsymbol{\mu}_j$, $j=1,2$, las medias de cada clase en la dirección discriminante.

Con ello, la regla de clasificación asigna un vector $\mathbf{x} \in \mathcal{X}$ a la clase $\mathcal{C}_j$ si su proyección sobre la dirección discriminante cae a un lado u otro del punto medio $\bar{\mu}$ de los puntos proyectados en dicha dirección.

La regla que minimiza el error de clasificación y su equivalente, la regla de Bayes, producen como función discriminante:

$$\mathbf{w}^T[\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)] = \ln\frac{\mathbb{P}(\mathcal{C}_2)}{\mathbb{P}(\mathcal{C}_1)}$$

que, si $\mathbb{P}(\mathcal{C}_1|\mathbf{x}) = \mathbb{P}(\mathcal{C}_2|\mathbf{x}) = 1/2$, coincide con el discriminante de Fisher poblacional.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anexo B

Espacios de Hilbert con Núcleo Reproductor

Por conveniencia, la notación del presente anexo puede ser diferente de la del resto del documento, en cuyo caso se detalla en cada momento.

Preliminares

En este apartado se resumen conceptos básicos utilizados en este anexo, así como otros conceptos de álgebra lineal utilizados en el documento.

Definición 73 ***Operador Lineal***. *Sean $\mathbb{E}_1, \mathbb{E}_2$ espacios vectoriales (o lineales) definidos en un conjunto de escalares $\mathbb{T}$ ($\mathbb{R}$ o $\mathbb{C}$). Un operador $L : \mathbb{E}_1 \to \mathbb{E}_2$ se dice lineal, $L \in \mathcal{L}(\mathbb{E}_1, \mathbb{E}_2)$, si $\forall f, g \in \mathbb{E}_1, \forall s, t \in \mathbb{T}$, $L(s \cdot f + t \cdot g) = s \cdot Lf + t \cdot Lg$*

Definición 74 *Si $L \in \mathcal{L}(\mathbb{E}_1, \mathbb{E}_2)$, entonces $\exists L^{-1} \in \mathcal{L}(L(\mathbb{E}_1), \mathbb{E}_1) \Leftrightarrow [Lf = 0 \Rightarrow f = 0]$, [Abellanas y Galindo 1991, p. 18].*

Definición 75 ***Espacio nulo de un operador***. *Si $L \in \mathcal{L}(\mathbb{E}_1, \mathbb{E}_2)$, se denomina espacio nulo de L al conjunto $\ker(L) \doteq \{f \in \mathbb{E} : Lf = 0\}$.*

Definición 76 ***Producto escalar*** *o* ***producto interno***. *Sea $\mathbb{E}$ un espacio lineal definido sobre $\mathbb{T}$. Un producto escalar o producto interno es una aplicación $\langle \cdot, \cdot \rangle : L \times L \to \mathbb{T}$ tal que:*

1) $\langle f, f \rangle \geq 0$ *y* $\langle f, f \rangle = 0 \Leftrightarrow f = 0$
2) $\langle f, \lambda_1 g_1 + \lambda_2 g_2 \rangle = \lambda_1 \langle f, g_1 \rangle + \lambda_2 \langle f, g_2 \rangle$
3) $\langle f, g \rangle = \overline{\langle g, f \rangle}$

Definición 77 ***Vectores ortogonales y ortonormales***. *Sea $\mathbb{E}$ espacio lineal. Dos vectores $\mathbf{x}, \mathbf{y} \in \mathbb{E}$ se dice* ***ortogonales*** *respecto al producto interno $\langle \cdot, \cdot \rangle$ si $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ y se denotan como $\mathbf{x} \perp \mathbf{y}$. Un conjunto $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{E}$ es ortogonal si $\mathbf{x}_i \perp \mathbf{x}_j$ $\forall i, j \in \{1 : n\}$, $i \neq j$. Si $\langle \mathbf{x}_i, \mathbf{x}_i \rangle = 1$ $\forall i, j \in \{1 : n\}$ el conjunto se dice* ***ortonormal***.

Definición 78 ***Norma y métrica asociada a un producto escalar***. *Dado un producto interno $\langle \cdot, \cdot \rangle$ definido en un espacio lineal, la norma definida como [Abellanas y Galindo 1991, th. 4.6] $\|f\| = \langle f, f \rangle^{1/2}$ se denomina la norma asociada al producto interno. Asimismo, la métrica definida como $d(f, g) = \|f - g\|$ se denomina métrica asociada a la norma correspondiente.*

Teorema 79 *[Abellanas y Galindo 1991, p. 80] Condición necesaria y suficiente para que un espacio normado $(\mathbb{E}, \|\cdot\|)$ admita un producto escalar tal que $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$ $\forall \mathbf{x} \in \mathbb{E}$ es que se satisfaga la ley del paralelogramo: $\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$ $\forall \mathbf{x}, \mathbf{y} \in \mathbb{E}$.*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Teorema 80 ***Pitágoras*** *[Abellanas y Galindo 1991, p. 77] Sea $\mathbb{E}$ espacio lineal con producto interno (Espacio Pre-Hilbert) Si $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{E}$ es conjunto ortonormal, entonces $\forall \mathbf{x} \in \mathbb{E}$, $\|\mathbf{x}\|^2 = \sum_{i=1}^n |\langle \mathbf{x}_i, \mathbf{x}\rangle|^2 + \|\mathbf{x} - \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{x}\rangle \mathbf{x}_i\|^2$.*

Corolario 81 *[Abellanas y Galindo 1991, cor. 4.5] $\forall \mathbf{x}, \mathbf{y} \in \mathbb{E}$, se verifica que $\mathbf{x} \perp \mathbf{y} \Rightarrow \|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$, y además:*

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\| \, \|\mathbf{y}\| \text{ Desigualdad de Schwartz} \tag{B.1}$$

Definición 82 ***Espacio de Hilbert****. Un espacio de Hilbert es un espacio lineal dotado de producto interno, y completo (toda sucesión de Cauchy es convergente) con la norma asociada a su producto interno $\|f\|_{\mathcal{H}} = \langle f, f\rangle_{\mathcal{H}}^{1/2}$ y la métrica asociada a dicha norma $d(f, g) = \|f - g\|_{\mathcal{H}}$.*

En lo que sigue, consideramos espacios de Hilbert, con la norma asociada a su producto interno, la métrica asociada a dicha norma y la topología asociada a dicha métrica. Asimismo, siempre que no haya riesgo de confusión obviaremos el subíndice $\mathcal{H}$ en la norma y el producto interno.

Definición 83 *[Kolmogorov y Fomin 1978, p. 171] Sea un conjunto numerable $\{\mathbb{E}_i\}_{i\in I}$ de espacios lineales definidos en $\mathbb{T}$.*

1. *El espacio lineal **suma directa** se define como:*

$$\oplus_{i\in I}\mathbb{E}_i = \mathbb{E}_1 \oplus \cdots \oplus \mathbb{E}_i \oplus \cdots \doteq \{\eta = \sum_{i\in I} f_i : f_i \in \mathbb{E}_i, i \in I\}$$

2. *Una aplicación $P_{\mathbb{E}_i} : \oplus_{i\in I}\mathbb{E}_i \to \mathbb{E}_i$ se denomina proyector o **proyección** en la dirección de $\mathbb{E}_i$.*

3. *Si $\mathbb{E}_1, \mathbb{E}_2$ son subespacios lineales de un espacio de Hilbert $\mathcal{H}$ verificando $\mathbb{E}_1 \perp \mathbb{E}_2$. Al espacio suma directa se le denomina **suma directa ortogonal**. Este concepto puede extenderse a un conjunto numerable de subespacios ortogonales entre sí.*

Definición 84 *[Bagget 1992, p. 3] Sea un conjunto numerable $\{\mathbb{E}_i\}_{i\in I}$ de espacios lineales definidos en $\mathbb{T}$. El espacio lineal **producto directo** denotado por $\otimes_{i\in I}\mathbb{E}_i = \mathbb{E}_1 \otimes \cdots \otimes \mathbb{E}_i \otimes \cdots$ es el producto cartesiano $\prod_{i\in I} \mathbb{E}_i = \{\mathbf{f} = (f_1, f_2, ...) : f_i \in \mathbb{E}_i \, \forall i \in I\}$ dotado de las siguientes propiedades:*

1) $\forall \mathbf{f}, \mathbf{g} \in \prod_{i\in I} \mathbb{E}_i$, $\mathbf{f} + \mathbf{g} = \mathbf{h}$ *con* $h_i = f_i + g_i \forall i \in I$
2) $\forall a \in \mathbb{T}, \forall \mathbf{f} \in \prod_{i\in I} \mathbb{E}_i$, $a\mathbf{f} = (af_1, af_2, ...)$

Definición 85 *Dado $\mathcal{H}$ espacio de Hilbert y $B \subset \mathcal{H}$, se denomina **complemento ortogonal** de B al conjunto: $B^\perp = \{f \in \mathcal{H} : f \perp B\}$.*

Teorema 86 *[Abellanas y Galindo 1991, p. 85-87] Dado $\mathcal{H}$ espacio de Hilbert, se verifican las siguientes propiedades:*

1. *$\forall B \subset \mathcal{H}$, $B^\perp$ es subespacio lineal cerrado (y por tanto subespacio de Hilbert con el producto interno de $\mathcal{H}$).*

USC UNIVERSIDADE

2. *Proyección ortogonal. Si* $\mathbb{H} \subset \mathcal{H}$ *es subespacio lineal cerrado entonces:* $\forall f \in \mathcal{H} : f = f_1 + f_2$ *con* $f_1 \in \mathbb{H}$ *y* $f_2 \in \mathbb{H}^\perp$*, de forma única. Se dice que* f_1 *es la* ***proyección ortogonal*** *de* f *sobre* $\mathbb{H}$*. Como consecuencia,* $\mathcal{H} = \mathbb{H} \oplus \mathbb{H}^\perp$ *suma directa ortogonal.*

3. $\forall B \subset \mathcal{H}$, $B \neq \varnothing$, $(B^\perp)^\perp = \mathrm{cl}(\langle B \rangle)$, *donde* cl *denota la clausura y* $\langle B \rangle$ *el espacio lineal generado por los elementos de* B.

4. *Un subespacio lineal* $\mathbb{E} \subset \mathcal{H}$ *es denso en* $\mathcal{H}$ *si y sólo sí* $\mathbb{E}^\perp = \{\mathbf{0}\}$.

Operadores en Espacios de Hilbert

Definición 87 ***Operador acotado***. *Sean* $\mathcal{H}_1, \mathcal{H}_2$ *espacios de Hilbert. Un operador lineal* $L : \mathcal{H}_1 \to \mathcal{H}_2$ *se dice acotado, y se denota* $L \in \mathcal{A}(\mathcal{H}_1, \mathcal{H}_2)$*, si:*

$$\|L\| \doteq \sup_{f \in \mathcal{H}} \frac{\|Lf\|}{\|f\|} < \infty$$

Definición 88 ***Operador continuo***. *Sean* $\mathcal{H}_1, \mathcal{H}_2$ *espacios de Hilbert. Un operador* $L : \mathcal{H}_1 \to \mathcal{H}_2$ *se dice continuo en* $f_0 \in \mathcal{H}_1$ *si* $\forall \varepsilon > 0, \exists \delta > 0$ *tal que* $\|f - f_0\|_{\mathcal{H}_1} < \delta \Rightarrow \|Lf - Lf_0\|_{\mathcal{H}_2} < \varepsilon$*. Un operador se dice continuo en* $\mathcal{H}_1$ *si lo es* $\forall f \in \mathcal{H}_1$.

Teorema 89 ***Representación de Riesz***. *[Abellanas y Galindo 1991, p. 144]. Sea* $\mathcal{H}$ *espacio de Hilbert, entonces* $\forall L : \mathcal{H} \to \mathbb{T}$ *funcional lineal continuo,* $\exists$ *un único* $f_L \in H$*, tal que:* $\forall g \in \mathcal{H}, L(g) = \langle g, f_L \rangle$ *. Al vector* $f_L \in \mathcal{H}$ *del teorema anterior, se le denomina* ***representante de Riestz*** *del operador* L.

Teorema 90 *Sea* $L \in \mathcal{L}(H_1, H_2)$*. Entonces:* L *acotado* $\Leftrightarrow$ L *continuo* $\Leftrightarrow$ L *continuo en algún* $f \in \mathcal{H}_1$.

Definición 91 ***Operador compacto***. *Un operador* $A \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ *se dice compacto, y se denota* $A \in \mathcal{C}(\mathcal{H}_1, \mathcal{H}_2)$*, si para cualquier sucesión acotada* $\{f_n\}_{n=1}^\infty \subset \mathcal{H}_1$*, su imagen* $\{Af_n\}_{n=1}^\infty$ *posee alguna subsucesión convergente.*

Todo operador compacto es acotado, de lo contrario, para alguna $\{f_n\}_{n=1}^\infty \subset \mathcal{H}_1$ acotada, $\|Af_n\| \to \infty$ cuando $n \to \infty$. Además, $\mathcal{C}(\mathcal{H}_1, \mathcal{H}_2)$ es subespacio lineal de $\mathcal{A}(H_1, H_2)$, [Abellanas y Galindo 1991, p. 199].

En lo que sigue si $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$, las familias de operadores lineales, acotados y compactos se denotan mediante $\mathcal{L}(\mathcal{H})$, $\mathcal{A}(\mathcal{H})$ y $\mathcal{C}(\mathcal{H})$, respectivamente.

Definición 92 ***Operador adjunto***. *Sea* $A \in \mathcal{A}(\mathcal{H})$*, con* $\mathcal{H}$ *espacio de Hilbert, un operador* A^+ *se denomina adjunto de* A *si verifica:* $\langle Af, g \rangle = \langle f, A^+ g \rangle$, $\forall f, g \in \mathcal{H}$.

El operador adjunto A *siempre existe, es único,* $A^{++} = A$, $\|A^+\| = \|A\|$ *y* $A^+ \in \mathcal{A}(\mathcal{H})$ *[Abellanas y Galindo 1991, p. 167].*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Definición 93 ***Operador autoadjunto***. $A \in \mathcal{A}(\mathcal{H})$ *se dice autoadjunto si* $A^+ = A$.

Si $A \in \mathcal{A}(\mathcal{H})$, entonces, [Abellanas y Galindo 1991, p. 170]: A autoadjunto $\Leftrightarrow [\langle Af, g\rangle = \langle f, Ag\rangle\,, \forall f, g \in \mathcal{H}] \Leftrightarrow \langle Af, f\rangle \in \mathbb{R}, \forall f \in \mathcal{H}$. Además, el conjunto de los operadores autoadjuntos es un subespacio lineal cerrado de $\mathcal{A}(H)$, [Pryce 1973, p. 281].

Definición 94 ***Operador positivo***. *Un operador $A \in \mathcal{A}(\mathcal{H})$ se dice positivo y se denota $A \geq 0$, si $\langle Af, f\rangle \geq 0$, $\forall f \in \mathcal{H}$.*

Se verifica que $A \geq 0 \Rightarrow A$ autoadjunto, pues, $\langle f, Af\rangle \in \mathbb{R}$.

Definición 95 ***Operador proyección ortogonal***. *$P \in \mathcal{A}(\mathcal{H})$ se denomina operador proyección ortogonal (o proyector ortogonal) si es autoadjunto e idempotente, es decir, $P^+ = P = PP \doteq P^2$.*

Teorema 96 *[Abellanas y Galindo 1991, p. 179] Si $P \in \mathcal{A}(\mathcal{H})$ es un proyector ortogonal, entonces $\exists \mathbb{H} \subset \mathcal{H}$, subespacio lineal cerrado, tal que P es el proyector ortogonal sobre $\mathbb{H}$.*

Definición 97 ***Operador normal***. *$A \in \mathcal{A}(\mathcal{H})$ se dice normal si $AA^+ = A^+A$.*

Teorema 98 *[Abellanas y Galindo 1991, p. 194] $A \in \mathcal{A}(\mathcal{H})$ es normal $\Leftrightarrow \|Af\| = \|A^+f\|\,,$ $\forall f \in \mathcal{H}$.*

Por consiguiente, $A \in \mathcal{A}(\mathcal{H})$ autoadjunto $\Rightarrow$ A normal. En particular, $A \in \mathcal{C}(\mathcal{H})$ autoadjunto $\Rightarrow$ A normal.

En lo que sigue, suponemos que $\mathcal{H}$ es **espacio de Hilbert separable**, es decir, que con la topología asociada a la métrica definida con su norma, posee un subconjunto denso y numerable.

Definición 99 ***Operador Hilbert-Schmidt***. *Un operador $K \in \mathcal{A}(\mathcal{H})$ se dice de tipo Hilbert-Schmidt y se denota $K \in \mathcal{C}_2(\mathcal{H})$, si $\sum_{j\in J} \|Ke_j\|^2 < \infty$ con $\{e_i\}_{i\in J} \subset \mathcal{H}$ base ortonormal de $\mathcal{H}$.*

En lo que resta de apartado, salvo indicación en contrario, nos referiremos al espacio de funciones:

$$\mathcal{H} = L_2(\mathcal{X}) = \{f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}\}$$
$$\text{con: } \|f\|_{\mathcal{H}} = \|f\|_{L_2(\mathcal{X})} = \left[\int_{\mathcal{X}} |f(\mathbf{x})|^2\, d\mathbf{x}\right]^{1/2} < \infty\}$$

Definición 100 ***Núcleo integral admisible***. *Una función $k(\mathbf{x}, \mathbf{y})$, medible Borel en $\mathcal{X} \times \mathcal{X}$, se dice núcleo integral admisible en $L_2(\mathcal{X})$ (o simplemente núcleo), si $f \in L_2(\mathcal{X}) \Rightarrow \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y} \in L_2(\mathcal{X})$.*

Definición 101 ***Núcleo de un operador integral***. *Sea el operador K que a cada $f \in L_2(\mathcal{X})$ le hace corresponder Kf definida como: $(Kf)(\mathbf{x}) = \int_T k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y}$. Si $k(\mathbf{x}, \mathbf{y})$ es núcleo integral admisible, se dice que $k(\mathbf{x}, \mathbf{y})$ es el núcleo del operador K, que es claramente lineal: $K \in \mathcal{L}(L_2(\mathcal{X}))$.*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Teorema 102 ***Propiedades de los operadores integrales con núcleo***. *Se verifica que [Abellanas y Galindo 1991, p. 241-242]*[123]*:*

1. *Si K operador con núcleo, entonces, $K \in \mathcal{A}(L_2(\mathcal{X}))$.*
2. *Dados K_1, K_2 definidos por los núcleos admisibles k_1, k_2, entonces: $K_1 = K_2 \Leftrightarrow k_1 = k_2$ c.p.d. en $\mathcal{X} \times \mathcal{X}$.*
3. *Si K tiene núcleo $k(\mathbf{x}, \mathbf{y})$, entonces K^+ tiene por núcleo k^+, siendo $k^+(\mathbf{y}, \mathbf{x}) = \overline{k(\mathbf{x}, \mathbf{y})}$, [Abellanas y Galindo 1991, p. 242]. Con ello, K autoadjunto $\Leftrightarrow k(\mathbf{x}, \mathbf{y}) = \overline{k(\mathbf{y}, \mathbf{x})}$, ($k$ es simétrico). En particular, si $k(\mathbf{x}, \mathbf{y})$ es real, K autoadjunto $\Leftrightarrow k$ simétrica.*
4. *Si $K \geq 0$ y autoadjunto entonces $\int_{\mathcal{X}' \times \mathcal{X}'} k(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0 \; \forall$ conjunto acotado $\mathcal{X}' \subset \mathcal{X}$.*
5. *Si $K \geq 0$ y autoadjunto, y k continua en todo $(\mathbf{x}, \mathbf{x}) \in \mathcal{X} \times \mathcal{X}$, entonces $k(\mathbf{x}, \mathbf{x}) \geq 0 \; \forall \mathbf{x} \in \mathcal{X}$.*

Teorema 103 ***Caracterización de los operadores Hilbert-Schmidt***. *[Abellanas y Galindo 1991, p. 204 y 245] Sea $K \in \mathcal{L}(L_2(\mathcal{X}))$. Entonces, K es Hilbert-Schmidt $\Leftrightarrow$ K es operador con núcleo $k \in L_2(\mathcal{X} \times \mathcal{X})$ que, entonces, se denomina* ***núcleo Hilbert-Schmidt****.*

Teorema 104 *[Abellanas y Galindo 1991, th. 9.11] $\mathcal{C}_2(\mathcal{H}) \subset \mathcal{C}(\mathcal{H})$, es decir, un operador tipo Hilbert-Schmidt es compacto.*

Definición 105 ***Función de Green***. *[Bagget 1992, p. 108]. Sea $A \in \mathcal{L}(L_2(\mathcal{X}))$. Se denomina función de Green para A, a una función k medible Borel, definida en $\mathcal{X} \times \mathcal{X}$ y núcleo admisible de algún operador integral K tal que, se verifica $A(Kg) = g$, es decir, K es operador inverso por la derecha de A.*

Por el teorema 103, si $k \in L_2(\mathcal{X} \times \mathcal{X})$, la función de Green k es núcleo de un operador Hilbert-Schimidt K (compacto).

Una función de Green k es útil para resolver ecuaciones del tipo $L(f) = g$ pues si se conoce k, entonces Kg es solución de la ecuación. Una aplicación típica de las funciones de Green es la resolución de ecuaciones diferenciales: si P es un operador diferencial lineal, y se plantea la ecuación diferencial $Pf = g$, la función de Green k es tal que $k(\cdot, \mathbf{y})$ es solución de la ecuación $P_{\mathbf{x}} f = \delta_{\mathbf{y}}$ donde $\delta_{\mathbf{y}}(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{y})$ es la función generalizada denominada delta de Dirac centrada en $\mathbf{y}$ [Kolmogorov y Fomin 1978, p. 137 y 219-220] y donde el subíndice $\mathbf{x}$ refleja el argumento de la función f. Ello es debido a que (p. ej. [Holland 1990, p. 63-64]):

$$[P(Kg)](\mathbf{x}) = P[\int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) d\mathbf{y}] = \int_{\mathcal{X}} Pk(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) d\mathbf{y} = \int_{\mathcal{X}} \delta_{\mathbf{y}}(\mathbf{x}) g(\mathbf{y}) d\mathbf{y} = g(\mathbf{x})$$

En general, si $\{\phi_j\}_{j=1}^s$ es una base para el espacio nulo de P, la solución general de la ecuación $Pf = g$ posee la forma: $f(\mathbf{x}) = k(\mathbf{x}, \mathbf{y}) + \sum_{j=1}^s c_j \phi_j(\mathbf{x})$.

[123] Los resultados que en lo que sigue se refieren a [Abellanas y Galindo 1991] se formulan allí para funciones definidas en $\mathcal{X} \subset \mathbb{R}$ pero las demostraciones pueden adaptarse fácilmente al caso $\mathcal{X} \subset \mathbb{R}^d$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Descomposición Espectral en Espacios de Hilbert

Definición 106 ***Autovalores y autovectores propios de un operador lineal***. *Sea $A \in \mathcal{L}(\mathcal{H})$ definido en un dominio denso en $\mathcal{H}$. $\lambda \in \mathbb{C}$ se dice autovalor o valor propio de A si $\exists f \in \mathcal{H}$, $f \neq 0$, tal que $Af = \lambda f$. A f se le denomina autovector o vector propio de A asociado a λ.*

El conjunto de los autovalores propios de A se denota por $\sigma_p(A)$.

Se verifica que $\mathcal{H}_\lambda(A) \doteq \{g \in \mathcal{H} : Ag = \lambda g\}$ es un subespacio lineal de $\mathcal{H}$ y a $\dim(\mathcal{H}_\lambda)$ se le denomina la multiplicidad de λ.

Teorema 107 *Se verifica que:*

1. $\dim(\mathcal{H}) < \infty \Rightarrow [\nexists (A - \lambda I)^{-1} \Rightarrow \lambda \in \sigma_p(A)]$ *donde I es el operador identidad [Abellanas y Galindo 1991, p. 214].*
2. $P \in \mathcal{A}(\mathcal{H})$ *proyector ortogonal* $\Rightarrow \sigma_p(P) = \{0, 1\}$ *[Abellanas y Galindo 1991, p. 227].*
3. $A \in \mathcal{C}(H) \Rightarrow \sigma_p(A)$ *es a lo sumo numerable, con un único punto de acumulación posible en el cero, (la sucesión de sus autovalores converge o termina en el cero), ([Abellanas y Galindo 1991, p. 231], [Pryce 1973, p. 300]).*

Teorema 108 *[Abellanas y Galindo 1991, p. 223] Sea $A \in \mathcal{A}(\mathcal{H})$ normal. Entonces se verifica que:*

1) $Af = \lambda f \Leftrightarrow A^+ f = \overline{\lambda} f$
2) $Af_1 = \lambda_1 f_1, Af_2 = \lambda_2 f_2,$ *con* $\lambda_1 \neq \lambda_2, \Rightarrow f_1 \perp f_2 \equiv H_{\lambda_1} \perp H_{\lambda_2}$

Teorema 109 ***Forma canónica de los operadores compactos normales***. *([Abellanas y Galindo 1991, p. 234], [Pryce 1973, p. 303]). Sea $A \in \mathcal{C}(\mathcal{H})$ normal. Si $\{\phi_j\}_{j \in J}$ son los autovectores de A repetidos de acuerdo con la multiplicidad de $\lambda_j \in \sigma_p(A)$ con $\lambda_j \neq 0$, entonces:*

1. $\forall f \in \mathcal{H}$, $f = \sum_{j \in J} \langle f, \phi_j \rangle \phi_j + f_0$, *donde* $f_0 \in \ker(A)$.
2. $Af = \sum_{j \in J} \lambda_j \langle f, \phi_j \rangle \phi_j$

Corolario 110 *[Pryce 1973, p. 303] Sea $A \in \mathcal{C}(\mathcal{H})$ normal. Entonces: $0 \notin \sigma_p(A) \Leftrightarrow A(\mathcal{H})$ es denso en $\mathcal{H} \Leftrightarrow \mathcal{H}$ posee una base ortonormal numerable $\{\phi_j\}_{j \in J}$ de vectores propios de A, asociados a autovalores $\lambda_j \neq 0$. (En este caso, $\ker(A) = \{0\}$).*

Teorema 111 *Sea $A \in \mathcal{C}(H)$ normal. Entonces, $A \geq 0 \Leftrightarrow \lambda_j \geq 0$, $\forall \lambda_j \in \sigma_p(A)$.*

Los resultados anteriores relativos a los operadores compactos y normales son aplicables a un operador K Hilbert-Schmidt, ($\Rightarrow$ compacto), con núcleo simétrico $k(\mathbf{x}, \mathbf{y})$ ($\Rightarrow$ autoadjunto $\Rightarrow$ normal) y/o positivo.

Teorema 112 ***Mercer***. *[Abellanas y Galindo 1991, p. 261] Sea K operador Hilbert-Schmidt con núcleo $k(\mathbf{x}, \mathbf{y})$ simétrico y continuo, con $\{\lambda_j\}_{j=1}^{\infty}$ autovalores de igual signo, excepto un número finito, y autovectores $\{\phi_j\}_{j=1}^{\infty}$ de K. Si $\mathcal{X} \subset \mathbb{R}^d$ compacto, entonces:*

1. $\sum_{j\in J}\lambda_j < \infty$.

2. $k(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{\infty}\lambda_j\phi_j(\mathbf{x})\overline{\phi_j(\mathbf{y})}$, *con convergencia uniforme (en* $\mathcal{X}\times\mathcal{X}$*) y absoluta (para cada* $(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{X}$*) de la serie.*

Si K es Hilbert-Schmidt positivo, el teorema anterior es aplicable con $\lambda_j \geq 0, \forall j = 1, ..., \infty$.

Espacios de Hilbert con Núcleo Reproductor

Definición 113 *[Wahba 1990b, p. 1]* ***Función definida positiva****. Sea* $\mathcal{X}\subset\mathbb{R}^d$*. Una función simétrica,* $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ *es definida positiva (d.p.), si dado* $n, \forall\mathbf{a} = (a_1, a_2, ..., a_n)^T \in \mathbb{R}^n$ *y* $\forall\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \in \mathcal{X}$ *se verifica que:*

$$\mathbf{a}^T\mathbf{K}\mathbf{a} \equiv \sum_{i,j=1}^{n} a_i a_j k(\mathbf{x}_i,\mathbf{x}_j) \geq 0 \tag{B.2}$$

siendo $(\mathbf{K})_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$*), es decir si la matriz* $\mathbf{K}$ *es semidefinida positiva. Si la desigualdad anterior es estricta* $\forall\mathbf{a}\neq\mathbf{0}$ *se dice que* k *es* ***estrictamente definida positiva*** *(e.d.p.)*[124].

De lo anterior, tomando $n = 1$ y $a = 1$, se deduce que $k(\mathbf{x},\mathbf{x}) \geq 0\ \forall\mathbf{x}\in\mathcal{X}$. Asimismo, si k es d.p. y $c\in\mathbb{R}^+$ entonces ck es d.p.

Asimismo, puede demostrarse [Stewart 1976] que la condición anterior es equivalente a la condición:

$$\int_{\mathcal{X}\times\mathcal{X}} k(\mathbf{x},\mathbf{y})f(\mathbf{x})f(\mathbf{y})d\mathbf{x}d\mathbf{y} \geq 0 \tag{B.3}$$

para toda f continua no idénticamente cero.

Ejemplos importantes de funciones son las de la forma $k(\mathbf{x},\mathbf{y}) = \psi(\mathbf{x}-\mathbf{y})$ para alguna función ψ definida en $\mathbb{R}^d$, y de la forma $k(\mathbf{x},\mathbf{y}) = \zeta(\|\mathbf{x}-\mathbf{y}\|)$ donde $\|\cdot\|$ es alguna norma en $\mathbb{R}^d$. En estos casos, la denominación de función definida positiva (estrictamente) se extiende a las funciones ψ y ζ, respectivamente (véase [Schoemberg 1938, p. 814-815] y para una visión histórica [Stewart 1976]).

Definición 114 *[Aronszajn 1950] Sea* $\mathcal{H}$ *un espacio de Hilbert de funciones* $f:\mathcal{X}\to\mathbb{R}$*. La función* $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ *se denomina* ***núcleo reproductor*** *de* $\mathcal{H}$*, y a* $\mathcal{H}$ *se le denomina* ***espacio de Hilbert con núcleo reproductor (RKHS)*** k*, si:*

$$\begin{array}{l} 1)\ \forall\mathbf{y}\in\mathcal{X}, k_{\mathbf{y}} \doteq k(\cdot,\mathbf{y}) \in \mathcal{H} \\ 2)\ \forall\mathbf{y}\in\mathcal{X}, \forall f\in\mathcal{H}, f(\mathbf{y}) = \langle f(\mathbf{x}), k_{\mathbf{y}}(\mathbf{x})\rangle \end{array} \tag{B.4}$$

La segunda condición anterior equivale a decir que $k_{\mathbf{y}}$ *es el representante de Riestz (teorema 89) del operador evaluación* $L_{\mathbf{y}}$ *definido como* $L_{\mathbf{y}}(f) = f(\mathbf{y})\ \forall\mathbf{y}\in\mathcal{X}$.

Teorema 115 *[Aronszajn 1950] Si* $\mathcal{H}$ *posee un núcleo reproductor* k*, éste es único.*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[124] Nótese que, cuando $\mathbf{K}$ es estrictamente definida positiva, k se dice definida positiva, y cuando $\mathbf{K}$ es definida positiva, k se dice semidefinida positiva.

Demostración. Si existiese otro núcleo reproductor $k'(\mathbf{x},\mathbf{y})$, entonces, para cualquier $\mathbf{y}\in\mathcal{X}$, por las propiedades reproductoras de k y k':

$$\begin{aligned} 0 &\leq \left\|k_{\mathbf{y}}-k'_{\mathbf{y}}\right\|^2 = \left\langle k_{\mathbf{y}}-k'_{\mathbf{y}}, k_{\mathbf{y}}-k'_{\mathbf{y}}\right\rangle \\ &= \langle k_{\mathbf{y}}, k_{\mathbf{y}}\rangle > + \left\langle k'_{\mathbf{y}}, k'_{\mathbf{y}}\right\rangle - \left\langle k'_{\mathbf{y}}, k_{\mathbf{y}}\right\rangle - \left\langle k_{\mathbf{y}}, k'_{\mathbf{y}}\right\rangle \\ &= k(\mathbf{y},\mathbf{y}) + k'(\mathbf{y},\mathbf{y}) - k'(\mathbf{y},\mathbf{y}) - k(\mathbf{y},\mathbf{y}) = 0 \end{aligned}$$

■

Teorema 116 *[Aronszajn 1950] $\mathcal{H}$ es espacio de Hilbert con núcleo reproductor $k \Leftrightarrow \forall\mathbf{x}\in\mathcal{X}$, el operador evaluación $L_{\mathbf{x}}:\mathcal{H}\to\mathbb{R}$ con $L_{\mathbf{x}}(f)=f(\mathbf{x})$ es un funcional lineal acotado en $\mathcal{H}$.*

Demostración. $\Rightarrow$) Se verifica que:

$$|f(\mathbf{y})| = |\langle f(\mathbf{x}), k(\mathbf{x},\mathbf{y})\rangle| \leq \|f\| \langle k(\mathbf{x},\mathbf{y}), k(\mathbf{x},\mathbf{y})\rangle^{1/2} = k(\mathbf{y},\mathbf{y})^{1/2}\|f\|$$

$\Leftarrow$) Si el operador evaluación, $L_{\mathbf{y}}:\mathcal{H}\to\mathbb{R}$ tal que $L_{\mathbf{y}}(f)=f(\mathbf{y})$, es acotado, entonces, por el teorema de representación de Riesz, posee un representante $k_{\mathbf{y}}\in\mathcal{H}$ tal que: $L_{\mathbf{y}}(f)=\langle f,k_{\mathbf{y}}\rangle = f(\mathbf{y})$, $\forall f\in\mathcal{H}$. Si tomamos $k(\mathbf{x},\mathbf{y})=k_{\mathbf{y}}(\mathbf{x})$, se verifican las condiciones de núcleo reproductor: $k_{\mathbf{y}}\in\mathcal{H}$ y $\langle f(\cdot), k(\cdot,\mathbf{y})\rangle = \langle f,k_{\mathbf{y}}\rangle = f(\mathbf{y})$.

■

Es decir, el núcleo reproductor, k, determina y está determinado por, los operadores evaluación del RHKS.

Del teorema anterior, se deduce claramente que $L_2(\mathcal{X})$ no es RKHS. Una función que verifica la propiedad reproductora en $L_2(\mathcal{X})$ es la función delta de Dirac, pero no satisface la condición 1) en (B.4).

Teorema 117 *[Aronszajn 1950], [Wahba 1990b] k es definida positiva $\Leftrightarrow$ k es núcleo reproductor (único) de algún espacio de Hilbert con Núcleo Reproductor (único). Con ello, existe una correspondencia biunívoca entre las funciones definidas positivas y los espacios de Hilbert de funciones, con núcleo reproductor*[125].

Demostración. $\Leftarrow$) Si $\mathcal{H}$ es espacio de Hilbert con Núcleo Reproductor $k(\mathbf{x},\mathbf{y})$, entonces, dados $a_1,...,a_n\in\mathbb{R}$ y $\mathbf{x}_1,...,\mathbf{x}_n\in\mathcal{X}$, entonces $k(\cdot,\mathbf{x}_i)\in\mathcal{H}$, y, aplicando la propiedad reproductora:

$$0\leq\left\|\sum_{i=1}^n a_i k(\mathbf{x},\mathbf{x}_i)\right\|^2 = \left\langle\sum_{i=1}^n a_i k(\mathbf{x},\mathbf{x}_i), \sum_{i=1}^n a_i k(\mathbf{x},\mathbf{x}_i)\right\rangle = \sum_{i,j=1}^n a_i a_j k(\mathbf{x}_i,\mathbf{x}_j)$$

$\Rightarrow$) Si k definida positiva, se construye el RKHS $\mathcal{H}=\mathcal{H}_k$ como sigue: fijado $\mathbf{x}\in\mathcal{X}$, se define la función real $k_{\mathbf{x}}=k(\mathbf{x},\cdot)$ y se construye el espacio formado por todas las combinaciones lineales finitas de la forma $\sum_{i=1}^n a_i k_{\mathbf{x}_i}$ con $a_i\in\mathbb{R}$ y $\mathbf{x}_i\in\mathcal{X}$, $n\in\mathbb{N}^+$, con el producto interno:

$$\left\langle\sum_i a_i k_{\mathbf{x}_i}, \sum_j b_j k_{\mathbf{z}_j}\right\rangle = \sum_{i,j} a_i b_j < k_{\mathbf{x}_i}, k_{\mathbf{z}_j} >= \sum_{i,j} a_i b_j k(\mathbf{x}_i,\mathbf{z}_j)$$

que está bien definido al ser k definida positiva pues utilizando la desigualdad de Schwartz:

$$|f(\mathbf{x})|^2 = \langle k_{\mathbf{x}}, f\rangle^2 \leq \|k_{\mathbf{x}}\|^2\|f\|^2 = \langle k_{\mathbf{x}},k_{\mathbf{x}}\rangle\|f\|^2 = k(\mathbf{x},\mathbf{x})\|f\|^2 \Rightarrow \|f\|^2 \geq \frac{|f(\mathbf{x})|^2}{k(\mathbf{x},\mathbf{x})} \qquad \text{(B.5)}$$

[125] Aquí, hablar de espacio de Hilbert con núcleo reproductor supone concretar una norma, y es a todo ello, a lo que corresponde un función definida positiva, no a la familia de funciones. Porque, una misma familia de funciones con normas diferentes dan lugar a diferentes espacios de Hilbert con núcleo reproductor[Aronszajn 1950, p. 382].

con lo que, teniendo en cuenta que $k(\mathbf{x},\mathbf{x}) \geq 0$, se cumple que $\langle f, f\rangle = \|f\|^2 = 0$ si y sólo sí $f = 0$. Además, se verifica que $k_{\mathbf{x}} \in \mathcal{H}$ y que $\forall f = \sum_{i=1}^{n} a_i k_{\mathbf{x}_i}$:

$$\langle f, k_x\rangle = \left\langle \sum_{i=1}^{n} a_i k_{\mathbf{x}_i}, k_{\mathbf{x}} \right\rangle = \sum_{i=1}^{n} a_i \langle k_{\mathbf{x}_i}, k_{\mathbf{x}}\rangle = \sum_{i=1}^{n} a_i k_{\mathbf{x}_i}(\mathbf{x}) = f(\mathbf{x})$$

con lo que k verifica las dos condiciones de núcleo reproductor[126].
Por último, la convergencia en la norma derivada del producto interno, $\|f\| =< f, f >^{1/2}$, implica convergencia punto a punto, ya que, por la desigualdad de Schwartz:

$$|f_n(\mathbf{x}) - f(\mathbf{x})| = |< (f_n - f), k_{\mathbf{x}} >| \leq \|f_n - f\| \, \|k_{\mathbf{x}}\|$$

De esta forma, podemos añadir al espacio así construido todos los límites de sucesiones de Cauchy, que estarán bien definidos como límites punto a punto[127]. Con ello, el espacio construido es completo y, por tanto, de Hilbert con Núcleo Reproductor, $k(\mathbf{x},\mathbf{y})$.

■

En la demostración anterior se muestra (ecuación (B.5)) que, aunque k sea definida positiva (no estrictamente) el producto interno está bien definido. Esta propiedad motiva la denominación definida positiva para k aunque la matriz $\mathbf{K}$ correspondiente sea semidefinida positiva.

Sin embargo, a pesar de que en la literatura no se concede excesivo espacio a las funciones estrictamente definidas positivas, es necesario destacar que si $\mathbf{a}^T\mathbf{K}\mathbf{a} = 0$ para algún $\{\mathbf{x}_i\}_{i=1}^n$ y $\mathbf{a} \in \mathbb{R}^n$ $\mathbf{a} \neq 0$, entonces:

$$\begin{aligned} 0 &= \sum_{i,j=1}^{n} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j=1}^{n} a_i a_j \left\langle k_{\mathbf{x}_i}, k_{\mathbf{x}_j} \right\rangle \\ &= \left\langle \sum_{i=1}^{n} a_i k_{\mathbf{x}_i}, \sum_{j=1}^{n} a_j k_{\mathbf{x}_j} \right\rangle = \left\| \sum_{i=1}^{n} a_i k_{\mathbf{x}_i} \right\|^2 \end{aligned}$$

(y con ello $\langle f, \sum_{i=1}^{n} a_i k_{\mathbf{x}_i}\rangle = 0 \; \forall f \in \mathcal{H}_k$) lo que no es deseable en problemas de estimación. Además, [Pinkus 2003] indica que si k es definida positiva, la condición de estrictamente definida positiva es necesaria para conseguir la densidad del espacio $\langle\{k_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}\rangle$ en $\mathcal{C}(\mathcal{X})$ con $\mathcal{X}$ compacto y $\mathcal{X} \subset \mathbb{R}^d$ (aunque no suficiente).

Antes de continuar, es preciso notar que, si consideramos el operador Hilbert-Schmidt K definido en $L_2(\mathcal{X})$ por $K(f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x},\mathbf{y}) f(\mathbf{y}) d\mathbf{y}$, cuyo núcleo k es definido positivo, entonces $K \in \mathcal{C}(L_2(\mathcal{X}))$ es normal (autoadjunto), compacto (teorema 104) y positivo (condición (B.3)), con lo que pueden aplicarse los resultados de la sección anterior (teoremas 107, 108 y 109).

Asimismo, si k es continua, por el teorema 102 $k(\mathbf{x},\mathbf{x}) \geq 0$ e, igualmente, puede aplicarse el teorema 112 de Mercer. Con todo ello, resulta que [Wahba 1990*b*]:

1. Existe un conjunto ortonormal numerable de autovectores $\{\phi_i\} \subset \mathcal{H} = L_2(\mathcal{X})$ y de autovalores $\{\lambda_j : \lambda_j \geq 0\}$ de K.

2. Para cada autovalor λ_j y correspondiente autovector ϕ_j de K, se verifica que:

$$(K\phi_j)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x},\mathbf{y}) \phi_j(\mathbf{y}) d\mathbf{y} = \lambda_j \phi_j(\mathbf{x})$$

[126] Nótese que el comportamiento del núcleo reproductor k es similar al de la función delta de Dirac $\delta(x)$ en L^2, salvo que L^2 con la norma usual no es espacio de Hilbert.

[127] La demostración de que esta ampliación es única puede verse por ejemplo en [Aronszajn 1950, p. 347].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3. Finalmente, k puede descomponerse como:

$$k(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\mathbf{x})\phi_j(\mathbf{y})$$

Nótese que en el caso $\mathcal{X} = \{1, 2, ..., N\}$ estamos ante la diagonalización de una matriz mediante sus autovectores: $\mathbf{K}\boldsymbol{\phi}_j = \lambda_j \boldsymbol{\phi}_j$, $\mathbf{K} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$ con $\mathbf{V}$ matriz cuyas columnas son los autovectores y $\boldsymbol{\Lambda}$ es una matriz diagonal con los autovalores en la diagonal.

Teorema 118 *[Wahba 1990b] Sea $k \in L_2(\mathcal{X}\times\mathcal{X})$ definida positiva y continua, sea:*

$$f_j = \int_{\mathcal{X}} f(\mathbf{x})\phi_j(\mathbf{x})d\mathbf{x}$$

con $\{\phi_j\}_{j=1}^{\infty}$ autovectores ortonormales del operador Hilbert-Schmidt con núcleo k, entonces:

$$f = \sum_{j=1}^{\infty} f_j\phi_j \in \mathcal{H}_k \Leftrightarrow \sum_{j=1}^{\infty} \frac{f_j^2}{\lambda_j} < \infty \; y \; \|f\|_k^2 \doteq \sum_{j=1}^{\infty} \frac{f_j^2}{\lambda_j} \tag{B.6}$$

donde $\mathcal{H}_k$ es el RHKS asociado a k con la norma $\|\cdot\|_{\mathcal{H}_k} \equiv \|\cdot\|_k$. Además, todas las funciones de $\mathcal{H}_k$ son continuas [Cucker y Smale 2001, sec. III.3, th.2].

Demostración. Veamos que $\mathcal{H}_k = \mathcal{M}$ con $\mathcal{M} = \{f : \sum_{j=1}^{\infty} \frac{f_j^2}{\lambda_j} < \infty\}$. En primer lugar, se verifica fácilmente que $\mathcal{M}$ es un espacio de Hilbert con la norma $\|f\| = \sum_{j=1}^{\infty} \frac{f_j^2}{\lambda_j}$.
Veamos que $k_{\mathbf{x}} \in \mathcal{H}_k$ y $\langle f, k_{\mathbf{x}}\rangle = f(\mathbf{x})$ con $f \in \mathcal{H}_k$ y $\mathbf{x} \in \mathcal{X}$: por ser $\{\phi_j\}$ sistema ortornormal en $L_2(\mathcal{X})$, con $\{\lambda_j\}$ sus correspondientes autovalores, se verifica:

$$k_{\mathbf{x}} = k(\mathbf{x},\cdot) = \sum_{j=1}^{\infty} \lambda_j\phi_j(\mathbf{x})\phi_j(\cdot) = \sum_{j=1}^{\infty} [\lambda_j\phi_j(\mathbf{x})]\phi_j(\cdot)$$

pero:

$$\sum_{j=1}^{\infty} [\lambda_j\phi_j(\mathbf{x})]^2/\lambda_j = \sum_{j=1}^{\infty} \lambda_j\phi_j^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) < \infty \Rightarrow k_{\mathbf{x}} \in \mathcal{H}_k$$

Asimismo:

$$f = \sum_{j=1}^{\infty} f_j\phi_j \Rightarrow \langle f, k_x\rangle = \sum_{j=1}^{\infty} \frac{f_j[\lambda_j\phi_j(\mathbf{x})]}{\sqrt{\lambda_j}\sqrt{\lambda_j}} = \sum_{j=1}^{\infty} f_j\phi_j(\mathbf{x}) = f(\mathbf{x})$$

donde $\langle\cdot,\cdot\rangle$ es el producto interno asociado [Abellanas y Galindo 1991, p. 80] a la norma $\|\cdot\|_k$ es decir, $\langle f, g\rangle = \sum_{j=1}^{\infty} f_j g_j/\sqrt{\lambda_j}$.

■

Definición 119 ***Representación primal y dual de las funciones de*** *$\mathcal{H}_k$. Sea una función $k \in L_2(\mathcal{X}\times\mathcal{X})$ simétrica y definida positiva, y sea $\mathcal{H}_k$ el espacio de Hilbert con núcleo reproductor construido en el teorema anterior. Las funciones de $\mathcal{H}_k$ admiten una doble representación:*

1. *En función del Núcleo Reproductor k : $f = \sum_{i=1}^{n} a_i k_{\mathbf{x}_i} = \sum_{i=1}^{n} a_i k(\mathbf{x}_i,\cdot)$, (o límite de sucesiones de Cauchy de funciones de ese tipo), que se denomina **representación dual** de las funciones del espacio. En este caso, la norma de $\mathcal{H}_k$ es:*

$$\|f\|_{\mathcal{H}_k} = \langle f, f\rangle_{\mathcal{H}_k} = \left\langle \sum_{i=1}^{n} a_i k_{\mathbf{x}_i}, \sum_{i=1}^{n} a_i k_{\mathbf{x}_i} \right\rangle = \mathbf{a}^T\mathbf{K}\mathbf{a}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

o sus límites en n.

2. *En función de la descomposición en serie de Fourier del núcleo reproductor $k : f = \sum_{j=1}^{\infty} f_j \phi_j$, que se denomina **representación primal** en el espacio de las características $\phi_j(\mathbf{x})$ (features), obtenido mediante las transformaciones $\phi_j(\mathbf{x})$ de las variables originales $\mathbf{x}$. En este caso $\|f\|_{\mathcal{H}_k}$ viene dada por la correspondiente expresión en (B.6).*

Una extensión de lo anterior al ámbito de las funciones aleatorias [Bosq 2000, p. 15-16], es la transformación de Karhunen-Loève.

Teorema 120 *[Bosq 2000, th. 1.5] **Transformación de Karhunen-Loève**. Sea $\{\mathsf{F}(x) : x \in \mathcal{X}\}$ un proceso estocástico en tiempo continuo de media cero, y sea $k(x_1, x_2) = \mathbb{E}[\mathsf{F}(x_1)\mathsf{F}(x_2)]$ su función de covarianza que suponemos continua. Entonces $\forall x \in \mathcal{X}$, $\mathsf{F}(x)$ admite una representación de forma única en media cuadrática de la forma:*

$$\mathsf{F}(x) = \sum_{j=1}^{\infty} \mathsf{F}_j \phi_j(x)$$

donde

$$\mathsf{F}_j = \left\langle \mathsf{F}, \phi_j \right\rangle_{L_2(\mathcal{X})} = \int_{\mathcal{X}} \mathsf{F}(x)\phi_j(x)dx$$

*son las denominadas **características** (features o componentes principales) del proceso que verifican $\mathbb{E}(\mathsf{F}_i\mathsf{F}_j) = \lambda_i\delta_{ij}$ con $\delta_{ij} = 1$ si $i = j$ y cero en caso contrario, y $\{\lambda_j\}_{j=1}^{\infty}$ y $\{\phi_j\}_{j=1}^{\infty}$ son los autovalores y respectivos autovectores del operador K de Hilbert-Schmidt con núcleo k que permiten descomponer k según el teorema 112 de Mercer.*

Sin embargo, si K tiene un número infinito de autovalores distintos de cero, las realizaciones f del proceso F no están en el RKHS $\mathcal{H}_k$ [Wahba 1990*b*, p. 5].

Finalizamos este anexo con un conjunto de resultados que ilustran propiedades importantes de los núcleos reproductores:

Teorema 121 *[Aronszajn 1950, p. 353] **Suma de núcleos reproductores**. Sean $\mathcal{H}_i$ $i = 1, 2$, subespacios de Hilbert ortogonales con núcleo reproductor k_i. Entonces, $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{x}, \mathbf{y})$ es el núcleo reproductor de la clase $\mathcal{H}$ de todas las funciones $f = f_1 + f_2$ con $f_i \in \mathcal{H}_i$ y con la norma $\|f\|^2 = \min[\|f_1\|^2_{\mathcal{H}_1} + \|f_2\|^2_{\mathcal{H}_2}]$ donde el mínimo se toma sobre todas las posibles descomposiciones $f = f_1 + f_2$.*

Como caso particular, si $\mathcal{H}_1 \cap \mathcal{H}_2 = \{0\}$, entonces $\|f\|^2 = \|f_1\|^2_{\mathcal{H}_1} + \|f_2\|^2_{\mathcal{H}_2}$, y sólo entonces $\mathcal{H}_1, \mathcal{H}_2$ son subespacios cerrados.

Teorema 122 *[Aronszajn 1950, p. 358] Sean $\mathcal{H}_i$ espacios de Hilbert con núcleo reproductor k_i, de funciones reales definidas en $\mathcal{X}$, con productos escalares $\langle \cdot, \cdot \rangle_{\mathcal{H}_i}$, $i = 1, 2$. Entonces, el producto directo $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ coincide con el espacio de Hilbert de funciones definidas en $\mathcal{X} \times \mathcal{X}$ de la forma:*

$$f(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{r} f_1^j(\mathbf{x}_1) f_2^j(\mathbf{x}_2)$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

o sus límites, ($\mathcal{H}$ es completo), donde $f_i^j \in \mathcal{H}_i$ $i = 1, 2$, $j = 1, ..., r$, y con el producto escalar:

$$\langle f, g \rangle = \sum_{j=1}^{r} \sum_{l=1}^{s} \left\langle f_1^j, g_1^l \right\rangle_{\mathcal{H}_1} \left\langle f_2^j, g_2^l \right\rangle_{\mathcal{H}_2}$$

Teorema 123 *[Aronszajn 1950, p. 361]* ***Producto de núcleos reproductores****. En las condiciones del teorema anterior, el producto directo $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ es un espacio de Hilbert con núcleo reproductor $k(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2) = k_1(\mathbf{x}_1, \mathbf{y}_1) k_2(\mathbf{x}_2, \mathbf{y}_2)$.*

Demostración. $k^*(\cdot, \cdot, \mathbf{y}_1, \mathbf{y}_2)$ es de la forma indicada en el teorema anterior con $r = 1$, con lo que $k^* \in \mathcal{H}$. Además, $\forall f \in \mathcal{H}$, se verifica que:

$$\begin{aligned} \langle f(\cdot, \cdot), k^*(\cdot, \cdot, \mathbf{y}_1, \mathbf{y}_2) \rangle &= \sum_{j=1}^{n} \left\langle f_1^j(\cdot), k_1(\cdot, \mathbf{y}_1) \right\rangle \left\langle f_2^j(\cdot), k_2(\cdot, \mathbf{y}_2) \right\rangle \\ &= \sum_{j=1}^{n} f_1^j(\mathbf{y}_1) f_2^j(\mathbf{y}_2) = f(\mathbf{y}_1, \mathbf{y}_2) \end{aligned}$$

■

Teorema 124 *[Aronszajn 1950, p. 351]* ***Restricción de un núcleo reproductor****. Si k es el núcleo reproductor de un RKHS $\mathcal{H}$ con norma $\|\cdot\|_k$, entonces la restricción de k a un subconjunto del dominio de definición de las funciones de $\mathcal{H}$, es núcleo reproductor del espacio de las funciones restringidas a dicho subconjunto, cuya norma verifica: $\|f\|_r = \min_g \|g\|_k$ con $g \in \mathcal{H}$ y cuya restricción es f.*

Teorema 125 *[Aronszajn 1950, p. 361].* ***Producto de núcleos reproductores (2)****. En las condiciones del teorema anterior, el núcleo $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) k_2(\mathbf{x}, \mathbf{y})$ es el núcleo reproductor del espacio $\mathcal{H}_k$ de todas las funciones de $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ restringidas al conjunto diagonal $\{(\mathbf{x}, \mathbf{x}) : \mathbf{x} \in \mathcal{X}\} \subset \mathcal{X} \times \mathcal{X}$, cuya norma es $\|f\| = \min_g \|g\|$ con $g \in \mathcal{H}$ y cuya restricción es f.*

Demostración. $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) k_2(\mathbf{x}, \mathbf{y})$ es la restricción de $k^*(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2)$ a la diagonal de $\mathcal{X} \times \mathcal{X}$, y así $k \in \mathcal{H}_k$. Aplicando el teorema anterior, se tiene el resultado.

■

Espacios de Hilbert con Seminúcleo Reproductor

Definición 126 *[Madych y Nelson 1988, p. 79] Sea $\mathcal{X} \subset \mathbb{R}^d$. Una función simétrica, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ se dice* ***condicionalmente definida positiva de orden*** *m (en $\mathbb{R}^d$) (c.d.p.) si dado $n \in \mathbb{N}^+, \forall \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \in \mathcal{X}$ distintos y $\forall \mathbf{a} = (a_1, a_2, ..., a_n)^T \in \mathbb{R}^n$ tal que $\sum_{i=1}^{n} a_i g(\mathbf{x}_i) = 0$ con $g \in \Pi_m(\mathbb{R}^d)$ espacio de polinomios de grado menor que m, se verifica que:*

$$\mathbf{a}^T \mathbf{K} \mathbf{a} \equiv \sum_{i,j=1}^{n} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

siendo $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$). Si en la condición anterior se verifica la desigualdad estricta $\forall \mathbf{a} \neq \mathbf{0}$, se dice que es ***estrictamente condicionalmente definida positiva*** *de orden m (e.c.d.p.)*

Como caso particular, las funciones (estrictamente) condicionalmente definidas positivas de orden $m = 0$ son las funciones (estrictamente) definidas positivas.

Al igual que para las funciones definidas positivas, la definición anterior puede aplicarse también a funciones ψ y ζ tal que $k(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$ o $k(\mathbf{x}, \mathbf{y}) = \zeta(\|\mathbf{x} - \mathbf{y}\|)$.

Teorema 127 *[Madych y Nelson 1988, th. 1.1] Sea $m \geq 0$ y k condicionalmente definida positiva de orden m. Entonces, existe un subespacio $\mathcal{H}_k \subset \mathcal{C}(\mathbb{R}^d)$ con un semiproducto interno $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ tal que:*

1. *El espacio nulo de correspondiente seminorma es el espacio de polinomios en $\mathbb{R}^d$ de grado menor que m : $\Pi_m(\mathbb{R}^d)$.*
2. *$\mathcal{H}_k = \mathbb{H} \oplus \Pi_m(\mathbb{R}^d)$ suma directa ortogonal, donde $\mathbb{H}$ es un espacio de Hilbert .*
3. *Si $\mathbf{c} \in \mathbb{R}^n$ tal que $\forall g \in \Pi_m(\mathbb{R}^d)$ y $\forall \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ se verifica $\sum_{i=1}^n c_i g(\mathbf{x}_i) = 0$, entonces $\sum_{i=1}^n c_i k_{\mathbf{x}_i} \in \mathcal{H}_k$ y se satisface la propiedad reproductora:*

$$\left\langle \sum_{i=1}^n c_i k_{\mathbf{x}_i}, f \right\rangle_{\mathcal{H}_k} = \sum_{i=1}^n c_i f(\mathbf{x}_i) \; \forall f \in \mathcal{H}_k \tag{B.7}$$

Además, las propiedades anteriores determinan $\mathcal{H}_k$ y $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ de forma única.

Nótese que, por la condición (B.7), si $g \in \Pi_m(\mathbb{R}^d)$ se tiene que $\forall \mathbf{c} \in \mathbb{R}^n$ tal que $\forall g \in \Pi_m(\mathbb{R}^d)$ y $\forall \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$:

$$\left\langle \sum_{i=1}^n c_i k_{\mathbf{x}_i}, g \right\rangle_{\mathcal{H}_k} = \sum_{i=1}^n c_i g(\mathbf{x}_i) = 0$$

con lo que:

$$\begin{array}{l} 1)\ \mathbb{H} \perp \not> _m(\mathbb{R}^d) \\ 2)\ \sum_{i=1}^n c_i k_{\mathbf{x}_i} \in \mathbb{H} \end{array} \tag{B.8}$$

donde la última condición se verifica $\forall \mathbf{c} \in \mathbb{R}^n$ tal que $\forall g \in \Pi_m(\mathbb{R}^d)$ y $\forall \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$, $\sum_{i=1}^n c_i g(\mathbf{x}_i) = 0$.

Definición 128 *[Dyn 1987, p. 54] El espacio $\mathcal{H}_k$ resultado del teorema anterior se denomina **espacio de Hilbert con seminúcleo reproductor** k (semireproducing kernel Hilbert space, –SRKHS) y a esta función el seminúcleo reproductor de $\mathcal{H}_k$.*

Un caso particular de lo anterior es cuando $m = 0$, en cuyo caso, k es definida positiva y $\mathcal{H}_k = \mathcal{H}_k$ es RKHS con núcleo k.

Las funciones condicionalmente definidas positivas y sus respectivos SRKHS $\mathcal{H}_k$ pueden caracterizarse utilizando técnicas de Fourier con el teorema que presentamos a continuación.

Antes de ello, ([Madych y Nelson 1990] y p. ej. [Dyn 1987], [Dyn 1989]) sea $\mathcal{S}_m$ la clase de funciones de test $\xi \in \mathcal{C}^\infty(\mathbb{R}^d)$ rápidamente decrecientes[128] (espacio de Schwartz –p. ej. [Bagget 1992, p. 53]) que poseen soporte compacto y son ortogonales a $\Pi_m(\mathbb{R}^d)$ en el sentido de L_2. Si una función $\check{\psi} \in \mathcal{C}(\mathbb{R}^d)$ está acotada por polinomios, se define su **transformada de Fourier**

[128] Es decir que ξ y sus parciales tienden a cero en el infinito más rápido que la inversa de un polinomio.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

generalizada de orden m, $\widetilde{\check{\psi}}_m$, con respecto al espacio $\mathcal{S}_m$, como una función generalizada[129] tal que:

$$\int_{\mathbb{R}^d} \check{\psi}(\mathbf{x})\xi(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^d} \widetilde{\check{\psi}}_m(\boldsymbol{\omega})\tilde{\xi}(\boldsymbol{\omega})d\boldsymbol{\omega},\ \xi \in \mathcal{S}_m$$

Teorema 129 *[Madych y Nelson 1990], [Dyn 1987]. Si* $\psi(\mathbf{x}-\mathbf{x}') = k(\mathbf{x},\mathbf{x}')$ *con* $\psi : \mathbb{R}^d \to \mathbb{R}$, *es la transformada de Fourier generalizada de orden* m *de una medida positiva* $\check{\psi}$ *en* $\mathbb{R}^d - \{0\}$, *(es decir:* $\psi = \widetilde{\check{\psi}}_m$*), satisfaciendo las condiciones:*

$$1)\ \int_{\|\mathbf{x}\|\leq 1} \|\mathbf{x}\|^{2m}\,\psi(\mathbf{x})d\mathbf{x} < \infty$$
$$2)\ \int_{\|\mathbf{x}\|\leq 1} \psi(\mathbf{x})d\mathbf{x} < \infty$$

entonces $\psi(\mathbf{x}-\mathbf{x}')$ *es condicionalmente definida positiva de orden* m *y así seminúcleo reproductor de un SRKHS* $\mathcal{H}_k$. *Dicho espacio está caracterizado por:*

$$\mathcal{H}_k = \{f : \|f\|^2_{\mathcal{H}_k} = \int_{\mathbb{R}^d} \frac{\left|\tilde{f}_m(\boldsymbol{\omega})\right|^2}{\check{\psi}(\boldsymbol{\omega})}d\boldsymbol{\omega} < \infty\}$$
$$\langle f,g\rangle_{\mathcal{H}_k} = \int_{\mathbb{R}^d} \frac{\tilde{f}(\boldsymbol{\omega})\overline{\tilde{g}}(\boldsymbol{\omega})}{\check{\psi}(\boldsymbol{\omega})}d\boldsymbol{\omega},\ f,g \in \mathcal{H}_k$$

y su espacio nulo es el espacio $\Pi_m(\mathbb{R}^d)$ *de polinomios de grado menor que* m, *es decir:* $\langle g,g\rangle = 0 \Leftrightarrow g \in \Pi_m(\mathbb{R}^d)$.

Caracterización de las Funciones Condicionalmente Definidas Positivas

Una caracterización clásica de las funciones definidas positivas la proporciona el siguiente teorema de Bochner:

Teorema 130 ***Bochner*** *([Dyn 1987], [Stewart 1976]). Sea* $\psi \in \mathcal{C}(\mathbb{R}^d)$ *acotada. Entonces* ψ *es definida positiva si y solo si* $\psi(\mathbf{x})$ *es la transformada de Fourier-Stieljes de una medida de Borel finita* μ *en* $\mathbb{R}^d$:

$$\psi(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-i\mathbf{x}^t\omega}d\mu \equiv \int_{\mathbb{R}^d} \dot{\mu}(\boldsymbol{\omega})e^{-i\mathbf{x}^t\boldsymbol{\omega}}d\boldsymbol{\omega}$$

donde $\dot{\mu}(\boldsymbol{\omega}) = \nabla_{\boldsymbol{\omega}}\mu$ *es positiva e integrable.*

Sin embargo, dicho teorema no resulta cómodo en la práctica para identificar funciones definidas positivas, por lo que se han desarrollado caracterizaciones alternativas, las más importantes debidas a [Micchelli 1986].

Definición 131 *[Micchelli 1986, p. 13]. Una función* $\zeta \in \mathcal{C}^\infty((0,\infty))$ *es* ***completamente monótona*** *en* $(0,\infty)$ *si:*

$$\forall r \in (0,\infty), \forall j \in \mathbb{N} : (-1)^j\zeta^{(j)}(r) \geq 0$$

[129] Es decir, un elemento del espacio dual de $\mathcal{S}$, (espacio de los funcionales lineales de $\mathcal{S}$ en $\mathbb{T}$) [Kolmogorov y Fomin 1978, p. 219]. Se denominan también distribuciones [Bagget 1992, p. 86], [Richards y Youn 1990].

USC

Teorema 132 *[Micchelli 1986], [Dyn 1989, th. 1] Una función $k(\mathbf{x},\mathbf{y})$ es condicionalmente definida positiva de orden m en $\mathbb{R}^d$ si y sólo sí k es de la forma $k(\mathbf{x}_1,\mathbf{x}_2)=\zeta(\|\mathbf{x}_1-\mathbf{x}_2\|)$ con $\zeta\in\mathcal{C}^\infty(\mathbb{R})$ verificando:*

$$(-1)^j\frac{d^j}{dr^j}\zeta(\sqrt{r})\geq 0,\ r>0,\ j\geq m$$

es decir, si y sólo si $(-1)^m\frac{d^j}{dr^j}\zeta(\sqrt{r})$ es completamente monótona.

Si además de la condición anterior, $\frac{d^m}{dr^m}\zeta(\sqrt{r})$ no es constante, entonces k es estrictamente condicionalmente definida positiva de orden m.

Funciones que cumplen lo anterior son: $\zeta(r)=e^{-r^2}$ y $\zeta(r)=(1+r^2)^{-\alpha}$ con $\alpha>0$, que, al ser completamente monótonas, las funciones $k(\mathbf{x}_1,\mathbf{x}_2)=\zeta(\|\mathbf{x}_1-\mathbf{x}_2\|)=e^{-\|\mathbf{x}_1-\mathbf{x}_2\|^2}$ y $k(\mathbf{x}_1,\mathbf{x}_2)=(1+\|\mathbf{x}_1-\mathbf{x}_2\|^2)^{-\alpha}$ son estrictamente definidas positivas. Otros ejemplos son los siguientes [Dyn 1989, th. 1] donde $\beta,\sigma^2\in\mathbb{R}$ y m es el orden de su definición:

1) $\zeta(r)=(-1)^m r^\beta$ con $2m-2<\beta<2m$
2) $\zeta(r)=(-1)^m r^{2m-2}\ln r$
3) $\zeta(r)=(-1)^m(r^2+\sigma^2)^{\beta/2}$ con $2m-2<\beta<2m$
4) $\zeta(r)=(r^2+\sigma^2)^{\beta/2}$ con $\beta<0,\ m=0$
5) $\zeta(r)=(-1)^m(r^2+\sigma^2)^{m-1}\ln(r^2+\sigma^2)^{1/2}$
6) $\zeta(r)=b$ constante con $m=1$

Del teorema anterior, se deduce también que la suma de funciones condicionalmente definidas positivas de orden m es definida positiva de orden m. En particular, si k es s.d.p de orden $m=1$, $k+c$ también lo es. De esta forma, como una función d.p. es c.d.p., si k es d.p. entonces $k+c$ siempre será al menos c.d.p. de orden $m=1$.

Además de las condiciones anteriores, para la resolución del problema de interpolación en el que, dado un conjunto de puntos $\mathbf{z}^n=(\mathbf{z}_1,...,\mathbf{z}_n)$ con $\mathbf{z}_i=(\mathbf{x}_i,y_i)$, se pretende determinar una función $f:\mathbb{R}^d\to\mathbb{R}$ de la forma:

$$f(\mathbf{x})=\sum_{i=1}^n c_i k(\mathbf{x},\mathbf{x}_i)=\sum_{i=1}^n c_i\zeta(\|\mathbf{x}-\mathbf{x}_i\|)$$

que verifique las condiciones de interpolación $f(\mathbf{x}_i)=y_i$, $i=1:n$, es decir $\mathbf{Kc}=\mathbf{y}$ donde $\mathbf{c}=(c_1,...,c_n)^T$, $\mathbf{y}=(y_1,...,y_n)^T$ y $\mathbf{K}$ es la matriz con $(\mathbf{K})_{ij}=\zeta(\|\mathbf{x}_i-\mathbf{x}_j\|)$, se tiene el siguiente resultado que permite utilizar funciones estrictamente condicionalmente definidas positivas de orden $m=1$ (no necesariamente de orden $m=0$).

Corolario 133 *[Micchelli 1986], [Dyn 1989, cor. 2] Sea $k(\mathbf{x}_1,\mathbf{x}_2)=\zeta(\|\mathbf{x}_1-\mathbf{x}_2\|)$ es estrictamente condicionalmente definida positiva de orden $m=1$, entonces $|\mathbf{K}|\neq 0$ si y sólo sí $\zeta(r)>0$ $\forall r>0$.*

Ejemplos de funciones a las que dicho resultado es aplicable son:

1) $\zeta(r)=(r^2+\sigma^2)^\alpha$ con $0<\alpha<1,\ \sigma\geq 0$
2) $\zeta(r)=\ln(r^2+\sigma^2)^{1/2}$ con $\sigma^2\geq 1$

Caso particular del segundo tipo es la función multicuádrica $\zeta(r)=(r^2+\sigma^2)^{1/2}$.

El siguiente resultado aporta una interesante relación entre las funciones definidas positivas y las condicionalmente definidas positivas de orden $m=1$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Teorema 134 *[Schölkopff y Smola 2002, Prop. 2.22], [Berg* et al. *1984, lema p. 74] Sea* $\mathbf{x}_0 \in \mathcal{X}$ *y sea* $k_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. *Entonces:*

$$k(\mathbf{x}, \mathbf{x}') = \frac{1}{2}[k_1(\mathbf{x}, \mathbf{x}') - k_1(\mathbf{x}, \mathbf{x}_0) - k_1(\mathbf{x}_0, \mathbf{x}')] + k_1(\mathbf{x}_0, \mathbf{x}_0) \tag{B.9}$$

es definida positiva si y sólo si k_1 *es condicionalmente definida positiva de orden* 1.

En [Berg *et al.* 1984, lema p. 74], el resultado anterior se establece en términos de la función $k_2 = -k_1$ definida negativa, condición que define como equivalente a que $k_1 = -k_2$ sea condicionalmente definida positiva de orden $m = 1$. Con ello, la relación (B.9) se escribe como:

$$k(\mathbf{x}, \mathbf{x}') = \frac{1}{2}[k_2(\mathbf{x}, \mathbf{x}_0) + k_2(\mathbf{x}_0, \mathbf{x}') - k_2(\mathbf{x}, \mathbf{x}')] - k_2(\mathbf{x}_0, \mathbf{x}_0)$$

Los siguientes resultados aportan condiciones suficientes para que una función de la forma $k(\mathbf{x}, \mathbf{y}) = \kappa(\langle \mathbf{x}, \mathbf{y} \rangle)$ sea definida positiva y estrictamente definida positiva, respectivamente.

Teorema 135 *[Berg* et al. *1984, p. 159], [Pinkus 2003] Si* $\mathbf{A}$ *es definida positiva y* $\mathbb{H}$ *es un espacio de Hilbert de dimensión infinita. Para que la matriz* $\kappa(\mathbf{A})$ *donde* $\kappa : \mathbb{R} \to \mathbb{R}$ *se aplica a cada elemento de* $\mathbf{A}$, *sea definida positiva, es necesario que* κ *sea de la forma:*

$$\kappa(r) = \sum_{j=0}^{\infty} c_j r^j$$

con $c_j \geq 0 \; \forall j$.

Teniendo en cuenta que, dados $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X} \subset \mathbb{H}$, la matriz $\mathbf{K}$ con $(\mathbf{K})_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ es definida positiva, entonces, para que la función k definida como $k(\mathbf{x}, \mathbf{y}) = \kappa(\langle \mathbf{x}, \mathbf{y} \rangle)$ sea definida positiva en un espacio de dimensión infinita, es necesario que κ satisfaga la condición anterior.

Teorema 136 *[Pinkus 2003] Sea* $\kappa(r) = \sum_{j=0}^{\infty} c_j r^j$ *con* $c_j \geq 0 \; \forall j$, *convergente* $\forall r \in \mathbb{R}$. *Sea* $J = \{j : c_j > 0\}$. *Sea* $\mathbb{H}$ *un espacio real con producto interno y sea* $B \subset \mathbb{H}$ *tal que* B *contiene un intervalo simétrico centrado en el origen, es decir,* $\{\alpha \mathbf{x} : |\alpha| \leq b\} \subset B$ *para algún* $b > 0$ *y* $\mathbf{x} \in \mathbb{H}$ *con* $\|\mathbf{x}\| = 1$. *Entonces,* $k(\mathbf{x}, \mathbf{y}) = \kappa(\langle \mathbf{x}, \mathbf{y} \rangle)$ *es estrictamente definida en* B *si y sólo sí* J *contiene a* $j = 0$, *a un conjunto infinito de enteros pares y a un conjunto infinito de enteros impares.*

Finalmente, el siguiente teorema resume un conjunto de operaciones permisibles para obtener funciones definidas positivas:

Teorema 137 *[Cristianini y Shawe-Taylor 2000, prop. 3.12] Sean* k_1, k_2 *funciones definidas positivas definidas en* $\mathcal{X} \times \mathcal{X}$ *con* $\mathcal{X} \in \mathbb{R}^d$, k_3 *definida positiva definida en* $\mathbb{R}^r \times \mathbb{R}^r$, $\mathbf{B}$ *una matriz semidefinida positiva y* π *un polinomio con coeficientes positivos. Entonces:*

1) $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{x}, \mathbf{y})$
2) $k(\mathbf{x}, \mathbf{y}) = ck_1(\mathbf{x}, \mathbf{y})$ *con* $c \in \mathbb{R}^+$
3) $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y})k_2(\mathbf{x}, \mathbf{y})$
4) $k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})f(\mathbf{y})$ *con* $f : \mathcal{X} \to \mathbb{R}$
5) $k(\mathbf{x}, \mathbf{y}) = k_1(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{y}))$ *con* $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^r$
6) $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{B} \mathbf{y}$
7) $k(\mathbf{x}, \mathbf{y}) = \pi(k_1(\mathbf{x}, \mathbf{y}))$
8) $k(\mathbf{x}, \mathbf{y}) = \exp(k_1(\mathbf{x}, \mathbf{y}))$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Demostración. Para un conjunto $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$, sean $\mathbf{a} \in \mathbb{R}^d$ cualquiera, y $\mathbf{K}$ y $\mathbf{K}_i$, las matrices Gram correspondientes a k y k_i, $i = 1, 2$, respectivamente.

1. Resulta por el teorema 121, pero también: $\mathbf{a}^T(\mathbf{K}_1 + \mathbf{K}_2)\mathbf{a} = \mathbf{a}^T\mathbf{K}_1\mathbf{a} + \mathbf{a}^T\mathbf{K}_2\mathbf{a} \geq 0$.
2. $\mathbf{a}^T\mathbf{K}\mathbf{a} = c\mathbf{a}^T\mathbf{K}_1\mathbf{a} \geq 0$.
3. Por el teorema 125.
4. $\mathbf{a}^T\mathbf{K}\mathbf{a} = \sum_{i,j=1}^n a_i a_j f(\mathbf{x}_i) f(\mathbf{x}_j) = [\sum_{i=1}^n a_i f(\mathbf{x}_i)]^2 \geq 0$.
5. k es restricción de una función d.f.
6. $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{B}\mathbf{y} = \mathbf{x}^T\mathbf{V}^T\mathbf{D}\mathbf{V}\mathbf{y} = \mathbf{x}^T\mathbf{V}^T\sqrt{\mathbf{D}}\sqrt{\mathbf{D}}\mathbf{V}\mathbf{y} = \left\langle \sqrt{\mathbf{D}}\mathbf{V}\mathbf{x}, \sqrt{\mathbf{D}}\mathbf{V}\mathbf{y} \right\rangle \geq 0$ producto interno en el espacio transformado por $\boldsymbol{\phi}(\mathbf{x}) = \sqrt{\mathbf{D}}\mathbf{V}\mathbf{x}$.
7. Utilizando 1. y 2. y para las constantes, 4.
8. La función exponencial se puede aproximar mediante polinomios con coeficientes positivos y los límites punto a punto de funciones d.f. son d.f. El resultado se obtiene utilizando 7.

■

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Anexo C

Propiedades Aproximadoras de las Redes Neuronales

Este apartado tiene por objeto aportar una somera visión del problema de aproximación de funciones y presentar las propiedades aproximadoras más relevantes de las redes neuronales. Una visión general sobre la teoría clásica de aproximación (lineal) puede verse en [DeVore y Lorentz 1993], [Lorentz 1986] y [Davis 1975]. Una revisión detallada y reciente sobre el problema de aproximación mediante esquemas no lineales puede verse en [DeVore 1998].

Conceptos Básicos de la Teoría de Aproximación de Funciones

El Problema de Aproximación de Funciones

Sea $(\mathcal{K}, \|\cdot\|_{\mathcal{K}})$ un espacio lineal normado y completo[130] (Banach) y $\mathcal{T}, \mathcal{F} \subset \mathcal{K}$. En lo que sigue, salvo indicación en contrario, suponemos que $\mathcal{K}$ es un espacio de funciones definidas en $\mathbb{R}^d$ con valores reales. En la teoría de aproximación de funciones se plantean, entre otros, los siguientes problemas de interés para nuestra exposición:

1. **Posibilidad de aproximación**: la posibilidad de aproximar los elementos de $\mathcal{T}$ mediante elementos $\mathcal{F}$ con la precisión deseada.

2. La existencia de la **mejor aproximación** mediante funciones de $\mathcal{F}$ a un elemento dado de $\mathcal{T}$.

3. Problema de **complejidad** (velocidad de convergencia): qué complejidad se requiere en la función aproximadora de $\mathcal{F}$ para aproximar una función de $\mathcal{T}$ con una precisión dada. Y a la inversa: con una complejidad dada, qué precisión puede conseguirse con la aproximación.

4. Problema de los **límites a la mejor aproximación**. La determinación de la velocidad máxima que es posible conseguir en la convergencia.

En lo que sigue exponemos los conceptos y resultados básicos relativos a los tres primeros problemas, el último puede verse en [Pinkus 1985].

Para abordar la respuesta a dichas preguntas es necesario en primer lugar definir un criterio que permita cuantificar la discrepancia entre el elemento a aproximar y los posibles elementos aproximadores. Dicho criterio es una métrica o distancia definida en términos de la norma del

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[130] Toda sucesión de Cauchy es convergente.

espacio marco $\mathcal{K}$:

$$\begin{aligned} \delta(f,g) &= \|f-g\|_{\mathcal{K}} \\ \delta(f,\mathcal{F}) &= \inf_{g\in\mathcal{F}} \|f-g\|_{\mathcal{K}} \end{aligned} \tag{C.1}$$

Definición 138 *Una familia $\mathcal{F}$ de funciones puede aproximar arbitrariamente bien un conjunto $\mathcal{T}$, o que $\mathcal{F}$ posee la **propiedad de aproximación universal** en $\mathcal{T}$, si $\forall \mathsf{f} \in \mathcal{T}$, $\forall \varepsilon > 0, \exists f \in \mathcal{F}$ tal que: $\delta(\mathsf{f}, f) = \|\mathsf{f} - f\|_{\mathcal{K}} < \varepsilon$, es decir, $\mathcal{F}$ es **denso** en $\mathcal{T}$ o, lo que es lo mismo, $\mathrm{cl}(\mathcal{F}) = \mathcal{T}$, donde $\mathrm{cl}(\mathcal{F}) = \{g \in \mathcal{K} : \forall \varepsilon > 0, \exists f \in \mathcal{F}, \|f-g\|_{\mathcal{K}} < \varepsilon\}$ es la clausura de $\mathcal{F}$.*

Definición 139 *Una familia $\mathcal{F}$ de funciones posee la **propiedad de la mejor aproximación** con respecto a una función $\mathsf{f} \in \mathcal{T}$ si existe alguna función f^* denominada **mejor aproximación** a f mediante $\mathcal{F}$, verificando:*

$$f^* = \inf_{f\in\mathcal{F}} \|\mathsf{f} - f\|_{\mathcal{K}}$$

En muchos problemas de interés, $\mathcal{F}$ es de la forma $\mathcal{F}_{\mathcal{G}} = \{\sum_{i=1}^{h} c_i\psi_i : \psi_i \in \mathcal{G}, \mathbf{c} \in \mathbb{R}^h, h \in \mathbb{Z}\}$ con $\mathcal{G} = \{\psi_1, ..., \psi_h, ...\} \subset \mathcal{K}$. En esta situación, se denomina **orden** h **de aproximación** a una función $\mathsf{f} \in \mathcal{T}$, al valor:

$$\delta_h(\mathsf{f}, \mathcal{F}_{\mathcal{G}}) \doteq \inf_{\mathbf{c}\in\mathbb{R}^h} \left\| \mathsf{f} - \sum_{i=1}^{h} c_i\psi_i \right\|_{\mathcal{K}}$$

La teoría clásica de aproximación funciones se refiere a esquemas de **aproximación lineal**, es decir, a los casos en los que $\mathcal{F}$ es un espacio lineal generado por un conjunto de funciones $\mathcal{G}$ (polinomios[131] algebraicos y trigonométricos, etc.). Con objeto de permitir el estudio del problema de complejidad, dicha familia se supone estructurada en una secuencia de subfamilias:

$$\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_h \subset \cdots$$

donde $\mathcal{F}_h$ es un subespacio lineal con[132] $\dim(\mathcal{F}_h) = h = \dim(\langle\{\psi_i\}_{i=1}^{h}\rangle)$ para alguna base $\{\psi_i\}_{i=1}^{h}$. El problema de la posibilidad de la aproximación consiste entonces en determinar si $\mathcal{F} = \cup_{h=1}^{\infty}\mathcal{F}_h$ es denso en el conjunto de interés $\mathcal{T}$. Por su parte, el problema de complejidad se formula como la determinación de la familia (complejidad) $\mathcal{F}_h$ asociada a una precisión determinada (o la precisión que aporta la familia $\mathcal{F}_h$).

Sin embargo, las redes neuronales son esquemas de **aproximación no lineal** en el sentido de que las familias $\mathcal{F}_h$ son variedades no lineales dependientes de un conjunto de parámetros. Incluso, expansiones aparentemente lineales en términos de funciones básicas pueden constituir aproximaciones de tipo no lineal. Por ejemplo, en el contexto del problema de aproximación en un espacio de Hilbert con base ortonormal $\{\psi_j\}_{j=1}^{\infty}$, [DeVore y Temlyakov 1996] indica que mientras la aproximación lineal utiliza un subespacio lineal $\mathcal{F}_h = \langle\{\psi_1, ..., \psi_h\}\rangle$, la aproximación no lineal se realiza mediante un espacio $\mathcal{F}'_h = \{g : g = \sum_{j\in\mathcal{J}} c_j\psi_j\}$ donde $\mathcal{J} \subset \mathbb{N}$ y $\#\mathcal{J} \leq h$. Así, en contraste con la aproximación lineal, la suma de dos elementos de $\mathcal{F}'_h$ puede requerir en general $2h$ términos para su representación. Un caso particular de lo anterior son las expansiones *wavelet* resultado de ordenar de mayor a menor los términos según el valor absoluto de sus coeficientes y construir una aproximación de grado h utilizando los h primeros.

[131] En estos casos, el orden h de aproximación suele referirse al grado del polinomio $f_h = \sum_{i=1}^{h} c_i\psi_i$ más que al número de términos de la combinación.

[132] Parametrizable por tanto en $\mathbb{R}^h$.

Posibilidad de Aproximación

Antes de presentar, a modo de referencia, teoremas clásicos de densidad, adoptamos la siguiente notación, válida en todo el documento:

1. $\mathcal{C}(\mathcal{X}, \mu)$ es el espacio de funciones continuas $f : \mathcal{X} \to \mathbb{R}$ con $\mathcal{X} \subset \mathbb{R}^d$, donde μ es una medida definida en $\mathcal{X}$.

2. $L_p(\mathcal{X}, \mu)$ con $p \in [1, \infty]$, es el espacio de las funciones $f : \mathcal{X} \to \mathbb{R}$ con $\mathcal{X} \subset \mathbb{R}^d$, donde μ es una medida definida en $\mathcal{X}$, que verifican $\|f\|_{L_p(\mathcal{X},\mu)} < \infty$, donde:

$$\|f\|_{L_p(\mathcal{X},\mu)} \doteq \begin{cases} [\int_{\mathcal{X}} |f(\mathbf{x})|^p \, \mu(d\mathbf{x})]^{1/p} & \text{si } p \in [1, \infty) \\ \operatorname{ess\,sup}_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})| & \text{si } p = \infty \end{cases}$$

donde $\operatorname{ess\,sup}$ denota el supremo esencial de $|f(\mathbf{x})|$ definido como (p. ej. [Pryce 1973, p. 73]) el mínimo valor a tal que $|f(\mathbf{x})| \leq a$ c.p.d., es decir, salvo en un conjunto $A \subset \mathcal{X}$ con $\mu(A) = 0$.

Si $\mu = \mu_{\mathrm{L}}$ donde μ_{L} es la medida de Lebesgue, denotamos dichos espacios por $\mathcal{C}(\mathcal{X})$ y $L_p(\mathcal{X}, \mu)$ respectivamente. Asimismo, cuando esté claro en el contexto, puede denotarse $\|f\|_p \equiv \|f\|_{L_p} \equiv \|f\|_{L_p(\mathcal{X},\mu)}$.

Además, si $\mathsf{P} : \mathbb{R}^d \to \mathbb{R}$ es una función continua por la derecha y con incrementos positivos, denotamos por μ_{P} a la única [Hoffmann-Jorgensen 1994, p. 38] medida de Lebesgue-Stieltjes inducida por P y denotamos, respectivamente [Hoffmann-Jorgensen 1994, p. 202]:

$$\begin{gathered} L_p(\mathcal{X}, \mathsf{P}) \doteq L_p(\mathcal{X}, \mu_{\mathsf{P}}) \\ \int_{\mathcal{X}} f(\mathbf{x}) d\mathsf{P}(\mathbf{x}) \doteq \int_{\mathcal{X}} f(\mathbf{x}) \mu_{\mathsf{P}}(d\mathbf{x}) \end{gathered}$$

donde $\mathcal{X}$ se supone que está en la σ-álgebra de Borel de $\mathbb{R}^d$.

Análogamente, en todo el documento se definen los espacios ℓ_p como el conjunto de las sucesiones $\mathbf{a} = \{a_t\}_{t \in T}$ con $a_t \in \mathbb{R}$ y T numerable, tal que $\|\mathbf{a}\|_p < \infty$, siendo:

$$\|\mathbf{a}\|_p = \begin{cases} [\sum_{t \in T} |a_t|^p]^{1/p} & \text{si } p \in [1, \infty) \\ \sup_{t \in T} |a_t| & \text{si } p = \infty \end{cases}$$

Si $p = 2$ y está claro en el contexto, $\|\cdot\|_2$ puede denotarse abreviadamente por $\|\cdot\|$.

Se tienen los siguientes resultados clásicos de densidad para familias importantes:

Teorema 140 *[Rudin 1988, th. 3.13] Sea $\mathcal{X}$ un espacio de medida con medida positiva μ. Sea $\mathcal{M}_s$ la familia de las funciones medibles simples, es decir, que verifican $\#\{f(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\} < \infty$, definidas en $\mathcal{X}$ y cuyo soporte $\operatorname{sop}(f) = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) \neq 0\}$ tiene medida finita. Si $1 \leq p < \infty$, entonces $\mathcal{M}_s$ es denso en $L_p(\mathcal{X}, \mu)$.*

Antes de presentar el siguiente resultado, denotamos por $\mathcal{C}_c(\mathcal{X})$ la familia de funciones continuas con soporte compacto.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Teorema 141 *[Rudin 1988, th. 3.1.4] Sea $\mathcal{X}$ un espacio Haussdorf*[133] *localmente compacto*[134] *con medida positiva μ. Sea . Se verifica que $\mathcal{C}_c(\mathcal{X})$ es denso en $L_p(\mathcal{X},\mu)$ con $p \in [1,\infty)$.*

Como consecuencia del teorema anterior, en sus mismas condiciones se tiene que $\mathcal{C}_c(\mathbb{R}^d)$ es denso en $L_p(\mathbb{R}^d,\mu)$ y que ($\mathcal{C}_c(\mathcal{X},\mu) \subset$) $\mathcal{C}(\mathcal{X},\mu)$ es denso en $L_p(\mathcal{X},\mu)$, con la norma $\|\cdot\|_{L_p}$, $p \in [1,\infty)$.

El resultado anterior no se cumple para $p=\infty$ [Rudin 1988, p. 79]: p. ej., la complección de $C_c(\mathcal{X})$ con la norma $\|\cdot\|_{L_\infty}$ es $C_0(\mathcal{X})$, la familia de las funciones continuas que *se anulan en el infinito*, es decir, que verifican que, $\forall \varepsilon > 0$, $\exists B \subset \mathcal{X}$ compacto tal que $|f(\mathbf{x})| < \varepsilon \; \forall \mathbf{x} \notin B$.

El resultado quizás más conocido de la teoría clásica de aproximación de funciones es el siguiente resultado de Weierestrass:

Teorema 142 *(Weierestrass), ([Davis 1975, th. 6.1.1], [Rudin 1988, th. 7.26]) Si $\mathsf{f} \in \mathcal{C}([a,b])$ y $\varepsilon > 0$, existe un polinomio algebraico $\xi_h(x)$ de grado h, tal que:* $\lim_{h\to\infty} \xi_h(x) = \mathsf{f}(x)$, $\forall x \in [a,b]$.

El teorema anterior indica que si $\mathcal{F}_h$ es el espacio de polinomios de grado $\leq h$, dado ε basta elegir la familia $\mathcal{F}_h$ con h suficientemente grande para aproximar cualquier función continua en $[a,b]$.

La siguiente definición y el siguiente teorema son herramientas básicas en la teoría de aproximación de funciones:

Definición 143 *[Davis 1975, p. 124] Una subfamilia $\mathcal{F} \subset \mathcal{T}$ con $\mathcal{T}$ conjunto de funciones reales es una **subálgebra** si es cerrada bajo el producto y las combinaciones lineales:*

$$\begin{array}{l} \textit{1) } f,g \in \mathcal{F} \Rightarrow fg \in \mathcal{F} \\ \textit{2) } f,g \in \mathcal{F} \Rightarrow \alpha f + \beta g \in \mathcal{F},\ \forall \alpha,\beta \in \mathbb{R} \end{array}$$

Teorema 144 *(Stone), [Davis 1975, th. 6.6.3], [Lorentz 1986, th. 7] [Rudin 1988, th. 7.32] Sea $\mathcal{X}$ un espacio topológico compacto Hausdorff y sea $\mathcal{C}(\mathcal{X})$ el espacio de funciones continuas definidas en $\mathcal{X}$. Si $\mathcal{F} \subset \mathcal{C}(\mathcal{X})$ verifica:*

$$\begin{array}{l} \textit{1) } \mathcal{F} \subset \mathcal{C}(\mathcal{X}) \textit{ es subálgebra} \\ \textit{2) } \exists g(\mathbf{x}) = \textit{cte con } g \in \mathcal{F} \\ \textit{3) } \forall \mathbf{x}_1,\mathbf{x}_2 \in \mathcal{X} \textit{ con } \mathbf{x}_1 \neq \mathbf{x}_2, \exists f \in \mathcal{F} : f(\mathbf{x}_1) \neq f(\mathbf{x}_2) \end{array}$$

entonces $\mathrm{cl}(\mathcal{F}) = \mathcal{C}(\mathcal{X})$ *uniformemente (con la norma* $\|\cdot\|_\infty$*).*

La tercera condición anterior suele expresarse como que $\mathcal{F}$ **separa puntos** de $\mathcal{X}$ y es necesaria pues, de no satisfacerse, dados dos puntos distintos en $\mathcal{X}$ siempre existe una función continua tal que $g(\mathbf{x}_1) \neq g(\mathbf{x}_2)$ que no sería aproximable mediante funciones de $\mathcal{F}$.

Dado que los polinomios algebraicos y los trigonométricos constituyen sendas subálgebras de $\mathcal{C}(\mathcal{X})$, se obtiene la propiedad aproximadora de dichas familias con respecto a $\mathcal{C}(\mathcal{X})$ con la norma uniforme $\|\cdot\|_\infty$.

[133] Es decir, $\forall \mathbf{x}_1,\mathbf{x}_2 \in \mathcal{X}$ con $\mathbf{x}_1 \neq \mathbf{x}_2$, $\exists B_1, B_2$ abiertos tal que $\mathbf{x}_i \in B_i$, $i=1,2$, entonces $B_1 \cap B_2 = \varnothing$), [Rudin 1988, p. 40].

[134] Es decir, si todo punto de $\mathcal{X}$ pertenece a un abierto con clausura compacta. Un espacio localmente compacto es compacto [Rudin 1988, p. 40].

Como veremos más adelante, una de las propiedades aproximadoras de las redes neuronales es su capacidad para aproximar cualquier función continua. Esta propiedad puede extenderse a las funciones medibles sin más que utilizar el siguiente teorema de Lusin:

Teorema 145 *(Lusin) [Rudin 1988, th. 2.24] Sea $\mathcal{X}$ un espacio de Haussdorf localmente compacto*[135] *y μ una medida regular positiva y finita en todo compacto incluido en $\mathcal{X}$. Entonces, $\forall K \subset \mathcal{X}$ con $\mu(K) < \infty$, con $\mathtt{f}(x) = 0$ si $x \notin K$ y $\forall \varepsilon > 0$, existe una función $f \in \mathcal{C}_c(\mathcal{X})$ tal que:*

$$\mu(\{x : \mathtt{f}(x) \neq f(x)\}) < \varepsilon$$

El teorema anterior indica que, en cualquier conjunto de medida finita, toda función medible puede ser aproximada con precisión arbitraria mediante una función continua, salvo en un conjunto de medida arbitrariamente pequeña. Así pues, todos los resultados sobre la posibilidad de aproximación de las redes neuronales a funciones continuas podrán extenderse a funciones medibles y, con ello, a cualquier posible función de decisión medible de un problema de clasificación.

Existencia de la Mejor Aproximación

En la teoría de aproximación lineal, existe la mejor aproximación debido al siguiente teorema:

Teorema 146 *[Davis 1975, th. 7.4.1] Sea $\mathcal{K}$ espacio de Banach, dado $\mathtt{f} \in \mathcal{T}$ y $\psi_1, ..., \psi_h \in \mathcal{F}$ elementos linealmente independientes. entonces el siguiente problema posee solución:*

$$\min_{f \in \mathcal{F}_h} \|\mathtt{f} - f\|_{\mathcal{K}} = \min_{\mathbf{c} \in \mathbb{R}^h} \left\| \mathtt{f} - \sum_{i=1}^{h} c_i \psi_i \right\|_{\mathcal{K}} \tag{C.2}$$

Como corolario del teorema anterior, el problema de aproximación de funciones continuas $\mathcal{C}(\mathcal{X})$ con $\mathcal{X} \subset \mathbb{R}^d$ compacto y Haussdorf, mediante polinomios algebraicos y trigonométricos, posee solución [Davis 1975, th. 7.4.1]. Además, si $\mathcal{K} = \mathcal{T}$ es estrictamente convexo[136], el problema anterior posee solución única. Casos particulares son: $\mathcal{K} = \mathcal{T} = L_p(\mathcal{X})$, $p \in (1, \infty)$. Sin embargo, $L_1(\mathcal{X})$ y $\mathcal{C}(\mathcal{X})$ no son estrictamente convexos (este último con la norma uniforme). Para este último caso, se demuestra que [Lorentz 1986, th. 6] si $\mathcal{G}$ es un sistema de Tchebycheff[137], el problema (C.2) posee solución única. Con ello, el problema de aproximación de funciones de $\mathcal{C}(\mathcal{X})$ mediante el conjunto de polinomios algebraicos o trigonométricos, posee la propiedad de la mejor aproximación.

En general, cuando el problema (C.2) posee solución (única) en una familia $\mathcal{F}$, se dice [Girosi y Poggio 1990] que $\mathcal{F}$ es un **conjunto de existencia (unicidad)**.

Se verifican las siguientes propiedades sencillas de demostrar para un conjunto de existencia:

Teorema 147 *[Girosi y Poggio 1990, Prop. 3.1 y th. 3.1] Sea $\mathcal{F} \subset \mathcal{T}$ familias aproximadora y de interés, respectivamente. Se verifica que:*

[135] Un espacio topológico $\mathcal{X}$ es localmente compacto si todo punto de $\mathcal{X}$ posee un entorno con clausura compacta.

[136] Un espacio es estrictamente convexo [Lorentz 1986, p. 17] si $\forall \mathbf{x}_1, \mathbf{x}_2$ con $\mathbf{x}_1 \neq \mathbf{x}_2$ y $\|\mathbf{x}_1\| = \|\mathbf{x}_2\| = 1$, y $\alpha_1, \alpha_2 > 0$ con $\alpha_1 + \alpha_2 = 1$, entonces $\|\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2\| < 1$ (la esfera unidad no contiene en su superficie ningún segmento convexo).

[137] Un conjunto $\mathcal{G}$ definido en $\mathcal{X}$ compacto y Haussdorf, es un **sistema de Tchebycheff** si verifica las dos condiciones siguientes [Lorentz 1986, p. 24]: 1) $\mathcal{X}$ contiene al menos h puntos, 2) Cada expresión $\sum_{i=1}^{h} c_i \psi_i$ con $c_i \neq 0$ para algún i y $\psi_i \in \mathcal{G}$, $i = 1 : h$, tiene como máximo $h - 1$ ceros en $\mathcal{X}$.

USC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

1. *Todo conjunto de existencia $\mathcal{F}$ es cerrado. La inversa no se verifica, pero sí la siguiente condición suficiente más fuerte:*

2. *Si $\mathcal{F} \subset \mathcal{T}$ es un conjunto compacto en un espacio métrico $\mathcal{T}$, entonces $\mathcal{F}$ es un conjunto de existencia.*

Demostración. Para demostrar 1: si $\mathcal{F}$ no es cerrado, entonces $\exists\{f_h\} \subset \mathcal{F}$ tal que $\lim_{h\to\infty} f_h = \mathsf{f} \notin \mathcal{F}$. Ello significa que $\lim_{h\to\infty} \delta_h(\mathsf{f}, f_h) = 0$ con $\mathsf{f} \in \mathcal{T}\backslash\mathcal{F}$, lo que implica que $\delta_h(\mathsf{f}, \mathcal{F}) = 0$ y como $\mathcal{F}$ es conjunto de existencia $\exists f_0 \in \mathcal{F} : \delta_h(\mathsf{f}, f_0) = 0$ absurdo.
Para demostrar 2, sólo es necesario utilizar que $\forall \mathsf{f} \in \mathcal{T}$, $\delta_h(\mathsf{f}, f) = \|\mathsf{f} - f\|$ es una función continua definida en un conjunto compacto $\mathcal{F}$ de un espacio métrico, con lo que alcanza su valor mínimo (y máximo) en $\mathcal{F}$. Una de las funciones $f \in \mathcal{F}$ donde se alcanza dicho mínimo es solución a (C.2).
■

Problema de la Complejidad. Velocidad de Convergencia

El teorema 142 de Berstein indica que el error de aproximación $\delta_h(\mathsf{f}) \equiv \delta(\mathsf{f}, \mathcal{F}_h)$ es una función monótona decreciente como función de h, pero no dice nada acerca de su velocidad de convergencia a medida que se incrementa la complejidad de la aproximación[138]. Esta información [Lorentz 1986], puede conseguirse si se suponen hipótesis adicionales sobre la suavidad (p. ej. en términos del módulo de continuidad, condiciones de Lipschitz o el número de derivadas) de las funciones de la familia de interés $\mathcal{T}$. Como es lógico, cuanto menos sea la familia $\mathcal{T}$ más rápida será la convergencia.

En teoría de aproximación se denominan teoremas *directos* los que proporcionan tasas de convergencia en función de las condiciones de suavidad de $\mathcal{T}$ (tasas tipo Jackson) y teoremas *inversos* los que determinan las condiciones de suavidad de $\mathcal{T}$ en función de una tasa de convergencia dada (tasas tipo Bernstein).

Resultados del segundo tipo pueden verse en [Lorentz 1986, sec. 4.4 y 5.4]. Ejemplos del primer tipo se presentan a continuación, aunque antes de ello, es preciso definir ciertos espacios de funciones.

En todo el documento, denotamos por $\mathcal{C}^r(\mathcal{X})$ el conjunto de funciones $f : \mathcal{X} \to \mathbb{R}$ con $\mathcal{X} \subset \mathbb{R}^d$, el espacio de funciones cuyas derivadas:

$$D^{\boldsymbol{\alpha}} f(\mathbf{x}) \doteq \frac{\partial^{\alpha_1,\ldots,\alpha_d} f(\mathbf{x})}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \text{ con } |\boldsymbol{\alpha}| \doteq \sum_{i=1}^{d} \alpha_i \leq r \in \mathbb{N}$$

son continuas. Se verifica [DeVore y Lorentz 1993, p. 17] que $\|D^{\boldsymbol{\alpha}} f\|_{L_\infty(\mathcal{X},\mu)}$ es seminorma y

$$\|f\|_{\mathcal{C}^r(\mathcal{X})} \doteq \|f\|_{L_\infty(\mathcal{X})} + \|D^{\boldsymbol{\alpha}} f\|_{L_\infty(\mathcal{X})}$$

es norma en $\mathcal{C}^r(\mathbb{R}^d)$.

[138] Nótese que si $\mathcal{K}$ es un espacio con producto interno la cantidad $\rho_h(\mathsf{f})$ es fácil de calcular si la familia Ψ es una base ortonormal [Davis 1975, p. 328]: $\rho_h(\mathsf{f})^2 = \|\mathsf{f}\|^2 - \sum_{i=1}^{h} |\langle \mathsf{f}, \psi_i \rangle|^2$. Y si $\mathcal{K}$ es un espacio de Hilbert $\rho_h(\mathsf{f})^2 = \sum_{i=h+1}^{\infty} |\langle \mathsf{f}, \psi_i \rangle|^2$.

Asimismo, se denota por $\mathcal{C}_p^r(\mathcal{X}) \subset \mathcal{C}^r(\mathcal{X})$ con $p \in [1, \infty]$ el espacio de funciones que verifican $\|f\|_{\mathcal{C}_p^r(\mathcal{X},\mu)} < \infty$, siendo:

$$\|f\|_{\mathcal{C}_p^r(\mathcal{X})} = [\sum_{|\boldsymbol{\alpha}|\leq r} \int_{\mathcal{X}} |D^{\boldsymbol{\alpha}} f(\mathbf{x})|^p \, d\mathbf{x}]^{1/p} \quad \text{si } p \in [1, \infty)$$
$$\|f\|_{\mathcal{C}_p^r(\mathcal{X})} = \max_{|\boldsymbol{\alpha}|\leq r} \|D^{\boldsymbol{\alpha}} f(\mathbf{x})\|_{L_\infty(\mathcal{X})} \quad \text{si } p = \infty$$

Si en $\mathcal{X}$ está definida una medida μ diferente a la medida usual de Lebesgue, denotamos los espacios anteriores mediante $\mathcal{C}^r(\mathcal{X}, \mu)$ y $\mathcal{C}_p^r(\mathcal{X}, \mu)$, y esta convención se sigue con todos los demás espacios de funciones.

Definición 148 *[DeVore 1998] Se define el **espacio de Sobolev** $\mathcal{W}_p^r(\mathcal{X}) \subset L_p(\mathcal{X})$ como la complección del espacio $\mathcal{C}_p^r(\mathcal{X})$ con la norma anterior.*

Se verifica que $\mathcal{W}_p^0(\mathbb{R}^d) = L_p(\mathbb{R}^d)$ y la siguiente caracterización de $\mathcal{W}_2^r(\mathbb{R}^d)$ [Girosi y Anzellotti 1993, p. 112]:

$$\mathcal{W}_2^r(\mathbb{R}^d) = \{f \in L_2(\mathbb{R}^d) : (1 + \|\boldsymbol{\omega}\|^2)^{r/2} \tilde{f}(\boldsymbol{\omega}) \in L_2(\mathbb{R}^d)\}$$

que puede utilizarse para definir los espacios de Sobolev para valores de r no enteros. Asimismo, se verifica que [Girosi y Anzellotti 1993, th. 8B.1] si $r > s + d/2$ con $s \in \mathbb{N}^+$, entonces:

$$\mathcal{W}_2^r(\mathbb{R}^d) \subset \mathcal{C}^s(\mathbb{R}^d) \text{ y } \exists C_1 > 0 : \|\cdot\|_{\mathcal{C}_\infty^s(\mathcal{X})} \leq C_1 \|\cdot\|_{\mathcal{W}_2^r(\mathbb{R}^d)} \tag{C.3}$$

Para una función $f : \mathcal{X} \to \mathbb{R}$, denotamos la **primera diferencia** de f como: $\Delta_{\mathbf{u}}(f, \mathbf{x}) = f(\mathbf{x} + \mathbf{u}) - f(\mathbf{x})$ donde dicha operación esté bien definida. Asimismo, por inducción, definimos la **diferencia de orden** $r \in \mathbb{N}$ con $r \geq 2$ como [DeVore 1998]:

$$\Delta_{\mathbf{u}}^r(f, \mathbf{x}) = \Delta_{\mathbf{u}}[\Delta_{\mathbf{u}}^{r-1}(f, \mathbf{x})] = \sum_{j=0}^{r} (-1)^{r-j} \binom{r}{j} f(\mathbf{x} + \mathbf{u}) - f(\mathbf{x})$$

y seguimos la convención [DeVore 1998] de que $\Delta_{\mathbf{u}}^r(f, \mathbf{x}) \doteq 0$ si cualquiera de los puntos $\mathbf{x}, ..., \mathbf{x} + r\mathbf{u}$ no está en $\mathcal{X}$.

Definición 149 *(p. ej. [DeVore 1998], [DeVore y Lorentz 1993]) Sea una función $f \in L_p(\mathcal{X})$ con $\mathcal{X} \subset \mathbb{R}^d$. Sea $r \in \mathbb{N}^+$ y $0 < p \leq \infty$. Se denomina **módulo de suavidad** (modulus o moduli of smoothness) de f a:*

$$\kappa_r(f, t)_p = \sup_{\|\mathbf{u}\| \leq t} \|\Delta_{\mathbf{u}}^r(f, \cdot)\|_{L_p(\mathcal{X})}$$

Si $p = \infty$, $L_p(\mathcal{X})$ se reemplaza por $\bar{\mathcal{C}}(\mathcal{X})$ el conjunto de funciones uniformemente continuas en $\mathcal{X}$.

*Si $r = 1$, el modulo de suavidad se denomina **modulo de continuidad**.*

El módulo de continuidad es una medida de suavidad de una función, que siempre verifica $\kappa_r(f, t)_p \to 0$ cuando $t \to 0$ monótonamente. Cuanto más rápida sea esta convergencia más suave es la función.

Definición 150 *[DeVore 1998]. Se define el espacio de Besov $\mathcal{B}_p^{\alpha q}(\mathcal{X})$ con $\mathcal{X} \subset \mathbb{R}^d$ abierto, $0 < \alpha < r$, $p, q \in (0, \infty]$ y $\alpha < r$, como el espacio de funciones que verifican que $|f|_{\mathcal{B}_p^{\alpha q}(\mathcal{X})} < \infty$*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

siendo:

$$|f|_{\mathcal{B}_p^{\alpha q}(\mathcal{X})} \doteq \begin{cases} (\int_0^\infty [s^{-\alpha}\kappa_r(f,s)_p]^q \frac{ds}{s})^{1/q} & si\ q \in (0,\infty) \\ \sup_{t>0} t^{-\alpha}\kappa_r(f,t)_p & si\ q = \infty \end{cases}$$

El parámetro α determina el grado de suavidad de las funciones y el parámetro p indica el espacio L_p en el que se mide dicha suavidad. Se verifica que $|\cdot|_{\mathcal{B}_p^{\alpha q}(\mathcal{X})}$ define una seminorma en $\mathcal{B}_p^{\alpha q}(\mathcal{X})$ cuya norma usual asociada es:

$$\|f\|_{\mathcal{B}_p^{\alpha q}(\mathcal{X})} = |f|_{\mathcal{B}_p^{\alpha q}(\mathcal{X})} + \|f\|_{L_p(\mathcal{X})}$$

Se verifican los siguientes teoremas que aportan tasas de convergencia:

Teorema 151 *(Jackson), [Davis 1975, th. 13.3.7] Sea* $\mathsf{f} \in \mathcal{C}([-1,1])$ *y* $\delta(\mathsf{f},\mathcal{F}_h)$ *el orden* h*-ésimo de aproximación a* f *mediante polinomios de grado* $\leq h$*. Si* $\kappa(t)$ *es el módulo de continuidad de* f *y* $C = 1 + \pi^2/2$*, se verifica:*

$$\delta(\mathsf{f},\mathcal{F}_h) \leq C\kappa(\frac{1}{h}) \tag{C.4}$$

El siguiente teorema proporciona una cota tipo Jackson a la aproximación mediante funciones constantes a trozos.

Teorema 152 *[Popov 1989, p. 525] Sea* $\mathcal{T} = \mathcal{B}_p^{\alpha q}(\mathcal{X})$ *el espacio de Besov con* $\mathcal{X}$ *abierto y* $1/q = \alpha/d + 1/p\ y\ p \in (0,\infty)$*. Sea* $\mathcal{G} = \{\psi_B : B \in \mathbb{D}\}$ *con* $\mathbb{D} = \cup_{j=0}^\infty \mathbb{D}_j$ *con* $\mathbb{D}_j$ *el conjunto de todos los hipercubos diádicos en* $\mathcal{X} \subset \mathbb{R}^d$ *con aristas de longitud* 2^{-j} *y sea* $\mathcal{F}_h = \{\sum_{i\in\mathcal{J}} c_i\psi_i : \psi_i \in \mathcal{D}, \#\mathcal{J} \leq h\}$*. Entonces, se verifica:*

$$\delta(\mathsf{f},\mathcal{F}_h)_{L_p} \leq Ch^{-\alpha/d}\,\|\mathsf{f}\|_{\mathcal{B}_p^{\alpha q}} \tag{C.5}$$

La tasa anterior también se verifica [Popov 1989, p. 533] para las aproximaciones polinómicas a trozos de grado total s si $\alpha < s$ y para aproximaciones mediante funciones racionales (para $p < \infty$) [Popov 1989, th. 11].

Sea $\mathcal{T} = \mathrm{Lip}_{r,\alpha}^d(\mathcal{X})$ el espacio de funciones en d variables con derivadas parciales de orden $j = 0, 1, ..., r$ continuas y acotadas en la norma uniforme en $\mathcal{X} = [0,1]^d$ y sus parciales de orden r satisfacen la condición de Lipschitz $\left\| D_j^r f(\mathbf{x}) - D_j^r f(\mathbf{x}') \right\|_{L_\infty(\mathcal{X})} \leq \gamma \left\| \mathbf{x} - \mathbf{x}' \right\|^\alpha$, $j = 1, ..., d$.

Teorema 153 *[Lorentz 1986, p. 51 y th. 8], [Girosi 1994] Sean* $\mathcal{F}_h$ *los espacios de polinomios algebraicos de grado* $s-1$ *en en cada una de las* d *variables con* $\dim(\mathcal{F}_h) = s^d = h$*. Entonces, con la estructura* $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_h \subset \cdots$*, se verifica la siguiente tasa de aproximación:*

$$\delta(\mathsf{f},\mathcal{F}_h) \leq Cdh^{-(r+\alpha)/d} \tag{C.6}$$

Resultados como el anterior permiten determinar el número h de términos en la aproximación necesarios para lograr un error máximo de aproximación ε (sobre todas las funciones de la familia):

$$\varepsilon = Cdh^{-(r+\alpha)/d} \Rightarrow h = O(\varepsilon^{-d/(r+\alpha)}) \tag{C.7}$$

donde puede observarse el problema de la dimensionalidad, que sólo puede contrarrestarse si la suavidad de la función se incrementa con la dimensión del espacio de entrada. En particular, si $r = 0$, $\alpha = 1$, se tiene $h = O(\varepsilon^{-d})$.

Espacios Funcionales Caracterizados por Técnicas de Aproximación [Popov 1989, p. 533]		
Familia Aproximadora $\mathcal{F}$	Espacio Aproximado	Condiciones
Polinomios trigonométricos	$\mathcal{A}^{\alpha,q}(L_p,\mathcal{F}) = \mathcal{B}_p^{\alpha q}$	$\forall p \in (0,\infty]$
Splines con nodos equidistantes	$\mathcal{A}^{\alpha,q}(L_p,\mathcal{F}) = \mathcal{B}_p^{\alpha q}$	$\forall p \in (0,\infty]$
Splines de grado r con nodos libres	$\mathcal{A}^{\alpha,q}(L_p,\mathcal{F}) = \mathcal{B}_p^{\alpha q}$	$\forall p \in (0,\infty), \alpha < r$ $1/q = \alpha + 1/p$
Comb. lin. de func. constantes a trozos	$\mathcal{A}^{\alpha/d,q}(L_p,\mathcal{F}) = \mathcal{B}_p^{\alpha q}$	$\forall p \in (0,\infty),\ \alpha > 0$ $\alpha < \min\{1, \frac{d}{(d-1)p}\}$

Tabla 40. Espacios funcionales caracterizados por diversas ténicas de aproximación. Las dos últimas técnicas son de tipo no lineal.

Las cotas (C.6), (C.5) y (C.4) muestran que la tasa de convergencia depende del grado de suavidad de las funciones de interés. Este hecho sugiere una caracterización de las funciones por el orden de convergencia que puede conseguirse con una técnica determinada. De hecho, la sucesión $\{\delta(\mathtt{f},\mathcal{F}_h)\}_{h=1}^{\infty}$ donde $\mathcal{F}_h$ son familias aproximadoras de creciente complejidad, (y su velocidad de convergencia a cero) es fundamental en la teoría de aproximación de funciones pues caracteriza a la función $\mathtt{f}$. En concreto, [Davis 1975, th. 13.1.6] dado $\mathcal{K}$ espacio de Banach y $\{\varepsilon_h\}_{h=0}^{\infty}$ sucesión de valores $\varepsilon_h \geq 0$, decreciente monótamente a cero, entonces $\exists \mathtt{f} \in \mathcal{K}$ tal que $\delta(\mathtt{f},\mathcal{K}_h) = \varepsilon_h$, $h = 0, 1, ...$ siendo $\mathcal{K}_h$ subespacios aproximadores de dimensión h.

Uno de los problemas fundamentales ([Popov 1989], [DeVore 1989]) de la teoría de aproximación es la determinación de la familia de funciones que poseen un mismo orden de aproximación. El ejemplo más sencillo de tales clases es la colección de funciones con orden de aproximación $\delta(\mathtt{f},\mathcal{F}_h) = O(h^{-\alpha})$, es decir $\{h^{\alpha}\delta_h(\mathtt{f})\} \in \ell_{\infty}$, siendo $0 < \alpha < \infty$, h el número de términos de la aproximación y $\mathcal{F}_h$ la familia aproximadora. Pero, en general, son de interés [DeVore 1989] los espacios $\mathcal{A}_q^{\alpha}(\mathcal{K},\mathcal{F}_h)$ con $q \in (0,\infty]$ constituidos por las funciones que verifican $|f|_{\mathcal{A}_q^{\alpha}} < \infty$, siendo $|\cdot|_{\mathcal{A}_q^{\alpha}}$ la seminorma:

$$|f|_{\mathcal{A}_q^{\alpha}} \doteq \begin{cases} [\sum_{h=1}^{\infty} (h^{\alpha}\delta(\mathtt{f},\mathcal{F}_h))^q \frac{1}{h})^{1/q} & \text{si } q \in (0,\infty) \\ \sup_{h\geq 1} h^{\alpha}\delta(\mathtt{f},\mathcal{F}_h) & \text{si } q = \infty \end{cases}$$

donde la norma $\|\cdot\|_{\mathcal{A}_q^{\alpha}}$ en $\mathcal{A}_q^{\alpha}$ se define mediante $\|\cdot\|_{\mathcal{A}_q^{\alpha}} = \|\cdot\|_{\mathcal{K}} + |\cdot|_{\mathcal{A}_q^{\alpha}}$. De esta definición, se deduce que los espacios $\mathcal{A}_q^{\alpha}$ recopilan todas las funciones con un orden de convergencia determinado o superior, donde el parámetro α indica el grado de suavidad de las funciones. La tabla 40 muestra algunos de los resultados clásicos conocidos [Popov 1989] para esquemas lineales y no lineales.

Aproximación Mediante Redes Neuronales

En este apartado, presentamos brevemente los resultados existentes relativos a esquemas de aproximación mediante redes neuronales en lo que respecta a la posibilidad de aproximación y a la existencia de la mejor aproximación.

Para las redes MLP, [Scarselli y Tsoi 1998] y [Fine 1999] proporcionan una buena visión general y [Pinkus 1999] realiza una detallada visión histórica y del estado del arte del problema de aproximación con estas redes. En el caso de las redes RBF, los resultados se encuentran mucho más dispersos en la literatura.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Posibilidad de Aproximación mediante Redes Neuronales

En general, los enfoques más frecuentemente utilizados en la literatura para probar la posibilidad de aproximación mediante esquemas basados en redes neuronales, son, entre otros, los siguientes:

1. La utilización del teorema 144 de Stone que generaliza el resultado de Weierestrass (teorema 142), vistos anteriormente. Fundamentalmente, este enfoque se utiliza con las redes RBF, [Hartman *et al.* 1990].

2. La utilización del teorema de Hahn-Banach aplicado a funciones discriminativas, según se describe a continuación. Este enfoque se utiliza con las redes MLP ([Cybenko 1989], [Hornik 1991]).

3. Enfoques basados en la posibilidad de aproximación de las redes neuronales con respecto a otras familias densas (p. ej. polinomios trigonométricos, funciones constantes a trozos, etc.) en el espacio de interés, ([Barron 1993], [Leshno *et al.* 1993]).

4. La utilización del lema de Maurey visto en el texto que, en el contexto de un espacio de Hilbert, muestra la posibilidad aproximadora de combinaciones convexas de una familia de funciones de norma finita. Dicho lema admite una demostración existencial [Barron 1993] y una constructiva [Jones 1990].

5. Utilización de teoremas de representación integral basados en las propiedades de la convolución ([Park y Sandberg 1991], [Park y Sandberg 1993]) u otros [Funahashi 1989].

6. Enfoques basados en el teorema de Kolmogorov de representación de funciones de varias variables mediante composición de funciones de una sóla variable. Este tipo de enfoque suele utilizarse para probar propiedades aproximadoras de redes con más de una capa oculta, que no se tratan en este documento, ([Kurková 1991], [Kurková 1992]).

7. Otros enfoques, como la utilización de la transformada de Radon [Carroll y Dickinson 1989].

De los resultados existentes, describimos brevemente los resultados más sencillos y referimos el contenido de los demás, distinguiendo entre los relativos a las redes MLP y los que tratan de las redes RBF.

Redes MLP

A continuación describimos brevemente los resultados más significativos sobre la posibilidad de aproximación de las redes neuronales MLP. Estos son el de [Cybenko 1989] por su elegancia y el de [Leshno *et al.* 1993] por haber el resuelto el problema de la posibilidad de aproximación imponiendo las mínimas condiciones sobre la función sigmoide.

Definición 154 *[Cybenko 1989] Una función ψ se dice* ***discriminativa*** *(discriminatory) si para toda medida de Borel μ con signo (real), finita y regular*[139] *en $[0,1]^d$, se verifica que:*

$$\int_{[0,1]^d} \psi(\mathbf{w}^T\mathbf{x} + w_0)\mu(d\mathbf{x}) = 0, \forall \mathbf{w} \in \mathbb{R}^d, \forall w_0 \in \mathbb{R} \Rightarrow \mu = 0$$

[139] Se dice que una medida es **con signo** si puede tomar valores reales (positivos o negativos) [Rudin 1988, p. 136]. Una medida es **regular** si $\forall B \in \mathcal{B}$ σ-álgebra de Borel con $\mu(B) < \infty$, se verifica que: $\mu(B) = \inf\{\mu(U) : B \subset U, U \text{ abierto}\} = \mu(B) = \sup\{\mu(V) : V \subset B, V \text{ compacto}\}$.

Para una función ψ discriminativa, denotamos:

$$\mathcal{F}_\psi = \{\sum_{i=1}^{h} c_i \psi(\mathbf{w}_i^T \mathbf{x} + w_{i0}) : \mathbf{w}_i \in \mathbb{R}^d, w_{i0} \in \mathbb{R}\} \tag{C.8}$$

El siguiente teorema es el resultado principal de [Cybenko 1989], que por su brevedad demostramos:

Teorema 155 *[Cybenko 1989, th. 1] Sea ψ una función continua y discriminativa. Entonces, la familia $\mathcal{F}_\psi$ es densa en $\mathcal{C}([0,1]^d)$ con la norma $\|\cdot\|_\infty$.*

Demostración. Es fácil demostrar que $\mathcal{F}_\psi \subset \mathcal{C}([0,1]^d)$ es un subespacio lineal. Supóngase que $\mathrm{cl}(\mathcal{F}_\psi) \neq \mathcal{C}([0,1]^d)$. Entonces, por un corolario[140] al teorema de Hahn-Banach, [Rudin 1988, th. 5.19], existe un funcional lineal acotado[141] $L \neq 0$ en $\mathcal{C}([0,1]^d)$ tal que $L(\mathrm{cl}(\mathcal{F})) = 0$. Por el teorema de representación de Riestz [Rudin 1988, th. 2.14], L admite la representación:

$$L(g) = \int_{[0,1]^d} g(\mathbf{x}) \mu(d\mathbf{x}),\ \forall g \in \mathcal{C}([0,1]^d)$$

para alguna medida μ de Borel con signo, finita y regular. Pero por $L(\mathrm{cl}(\mathcal{F})) = 0$, ha de verificarse que $L(\psi(\mathbf{w}^T\mathbf{x} + w_0)) = 0\ \forall \mathbf{w} \in \mathbb{R}^d, \forall w_0 \in \mathbb{R}$ lo que, por ser ψ discriminativa, implica que $\mu = 0$, lo que es una contradicción.

■

Los enfoques basados en el teorema de Hahn-Banach según el teorema anterior se basan en probar que una función sigmoide ψ es discriminativa. A este respecto, [Cybenko 1989, lema 1] prueba que toda función sigmoide acotada y medible es discriminativa y, en particular, toda sigmoide continua, y [Hornik 1991] extiende este resultado probando que basta ser continua y acotada para ser discriminativa en un contexto de medida finita en $\mathcal{X}$. Este último resultado muestra que **el éxito de las redes MLP se basa más en la potencia de la composición de funciones que en la especificidad sigmoide de la función de activación**.

El siguiente teorema es consecuencia de la dualidad (p. ej. [Pryce 1973, p. 213], [Bagget 1992, p. 81]) entre los espacios L_∞ y L_1:

Teorema 156 *[Cybenko 1989, th. 4] Sea ψ una función sigmoide, acotada y medible. Entonces, la familia $\mathcal{F}_\psi$ es densa en $L_1([0,1]^d)$ con su propia norma.*

Por los teoremas anteriores, la familia constituida por las redes neuronales MLP es densa en $\mathcal{C}([0,1]^d)$ y en $L_1([0,1]^d)$.

Sea $L_\infty^{\mathrm{loc}}(\mathbb{R}^d)$ el espacio de las funciones $f \in L_\infty(B)\ \forall B \subset \mathbb{R}^d$ compacto. El siguiente teorema determina las condiciones más generales que debe satisfacer la función sigmoide ψ para obtener la propiedad de aproximación de las redes neuronales MLP.

Teorema 157 *[Leshno* et al. *1993, th. 1] Sea $\psi \in L_\infty^{loc}(\mathbb{R}^d)$ una función cuyo conjunto de discontinuidad Υ_ψ satisface $\mu_L(cl(\Upsilon_\psi)) = 0$ donde μ_L es la medida de Lebesgue, y sea $\mathcal{F}_\psi$ definida como en (C.8). Entonces, $\mathcal{F}_\psi$ es denso en $\mathcal{C}(\mathbb{R}^d)$ si y sólo si ψ no es un polinomio casi por doquier (c.p.d.).*

[140] [Rudin 1988, th. 5.19] Si $\mathcal{M}$ es un subespacio vectorial de un espacio normado $\mathcal{X}$. Entonces $\mathbf{x}_0 \in \mathrm{cl}(\mathcal{M}) \Leftrightarrow \nexists L$ funcional lineal acotado en $\mathcal{X}$ tal que $L(\mathbf{x}) = 0\ \forall \mathbf{x} \in \mathcal{M}$ pero $L(\mathbf{x}_0) \neq 0$.

[141] En $\mathbb{R}^d$, el representante de Riestz de este funcional sería un vector ortogonal al subespacio considerado.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En su discusión sobre los resultados de su artículo, [Leshno *et al.* 1993] ilustran la necesidad de incluir el término w_0 en las funciones básicas $\psi(\mathbf{w}^T\mathbf{x} + w_0)$ de la familia $\mathcal{F}_\psi$ mencionando el caso de la función seno que verifica las condiciones del teorema pero es una función impar cuya correspondiente familia $\mathcal{F}_\psi$ no podría aproximar las funciones pares (como la función coseno) si no se utiliza el término umbral w_0 en la aproximación.

La tabla 41 resume los resultados anteriores y otros relevantes sobre las posibilidades aproximadoras de las redes MLP. En ella, $\tilde{f}$ denota la transformada de Fourier de f definida como: $\tilde{f}(\boldsymbol{\omega}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(\mathbf{x}) e^{-i\boldsymbol{\omega}^T\mathbf{x}} d\mathbf{x}$.

Aproximación de Derivadas mediante redes MLP

Las posibilidades de aproximación anteriores no implican que las derivadas de la función de interés puedan ser aproximadas en las mismas condiciones por las derivadas de las funciones de $\mathcal{F}_\psi$. Un caso claro, es la aproximación mediante funciones constantes a trozos, cuyas derivadas son nulas c.p.d., y no pueden poseer la propiedad de aproximación a las derivadas.

El resultado más relevante al respecto para las redes MLP es el siguiente (véase también [Hornik *et al.* 1994]).

Teorema 158 *[Hornik 1991, th. 3] Si $\psi \in \mathcal{C}^r(\mathbb{R}^d)$ es acotada y no constante, entonces $\forall \mu$ medida finita en un compacto $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{F}_\psi$ es denso en $\mathcal{C}_p^r(\mathcal{X})$, $p \in [1, \infty]$ con respecto a su propia norma.*

Redes RBF

En la demostración de la posibilidad aproximadora de las redes neuronales RBF se utilizan fundamentalmente tres enfoques de los descritos anteriormente, basados respectivamente en:

1. El teorema de Stone (Weierestrass) ([Girosi y Poggio 1990], [Hartman *et al.* 1990]).
2. Teoremas de representación integral basados en las propiedades de la convolución ([Park y Sandberg 1991], [Park y Sandberg 1993], [Girosi y Anzellotti 1993]).
3. El lema de Maurey, ([Girosi y Anzellotti 1993], [Girosi 1994]).

El primero de los enfoques se refiere a las redes RBF con activación gausiana y se basa en el siguiente lema que permite la aplicación inmediata del teorema de Stone. Al respecto, denotamos:

$$\mathcal{F} = \{\sum_{i=1}^{h} c_i \psi_{a_i}(\mathbf{x}, \mathbf{x}_i) : \mathbf{c} \in \mathbb{R}^h, \psi_{a_i}(\mathbf{x}, \mathbf{x}_i) = e^{-a_i \|\mathbf{x}-\mathbf{x}_i\|^2}, \mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d, a_i \in \mathbb{R}^+, h \in \mathbb{N}^+\}$$

Lema 159 *[Hartman* et al. *1990], [Girosi y Poggio 1990, prop. B1] Si $\mathcal{X} \subset \mathbb{R}^d$ es convexo y compacto, la familia $\mathcal{F}$ anterior es subálgebra.*

Demostración. Hay que probar que si $f, g \in \mathcal{F}$ entonces, $fg \in \mathcal{F}$ y $\alpha f + \beta g \in \mathcal{F}$, $\forall \alpha, \beta \in \mathbb{R}$. La primera condición es trivial. Para la segunda condición, se verifica que:

$$a_1 \|\mathbf{x} - \mathbf{x}_1\|^2 + a_2 \|\mathbf{x} - \mathbf{x}_2\|^2 = (a_1 + a_2) \|\mathbf{x} - \mathbf{x}_0\|^2 - a_1 a_2 (a_1 + a_2)^{-1} \|\mathbf{x}_1 - \mathbf{x}_2\|^2$$

donde $\mathbf{x}_0 = \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2$ con $\lambda_1 = a_1(a_1 + a_2)^{-1}$, $\lambda_2 = a_2(a_1 + a_2)^{-1}$ y $\lambda_1 + \lambda_2 = 1$, con lo que

Propiedades Aproximadoras de la red MLP: $f(\mathbf{x}) = \sum_{i=1}^{h} c_i \psi(\mathbf{w}_i^T \mathbf{x} + w_{i0})$			
Espacio Funcional	$\|\cdot\|$	**Condiciones**	**Enfoque**
$\mathcal{C}(\mathcal{X}),\ \mathcal{X}$ compacto [Cybenko 1989]	L_∞	ψ continua, sigmoide	Existencial Th. Hahn-Banach
$\mathcal{T}_C(\mathbb{B}_{d,\tau})$ μ med. prob. [Barron 1993]	L_2	ψ sigmoide medible, acotada	A.- Existencial: 1) Lema Maurey-Jones 2) f $\rightarrow$ Trigonom. $\rightarrow$ Func. Salto $\rightarrow$ Sigmoides B.- Constructivo: *greedy*
$\mathcal{C}(\mathcal{X}),\ \mathcal{X}$ compacto [Funahashi 1989]	L_∞	ψ sigmoide continua, monótona	Existencial
$\mathcal{C}(\mathcal{X}, \mu),\ \mathcal{X}$ compacto [Hornik 1991]	L_∞	ψ continua, acotada μ medida finita,	Existencial Th. Hahn-Banach
$\mathcal{C}(\mathbb{R}^d)$ [Leshno *et al.* 1993]	L_∞	$\psi \in L_\infty^{\text{loc}}(\mathbb{R}^d)$ $\mu_{\text{L}}(\text{cl}(\Upsilon_\psi)) = 0$ con Υ_ψ cto. discont. ψ no polinomio c.p.d	Existencial
$L_1(\mathcal{X}),\ \mathcal{X}$ compacto [Cybenko 1989]	L_1	$\psi \in L^1(\mathcal{X})$ $\int \psi(t)dt \neq 0$	Existencial Th. Hahn-Banach
$L_2(\mathcal{X}),\ \mathcal{X}$ compacto [Carroll y Dickinson 1989]	L_1	ψ continua, sigmoide	Constructiva Transf. de Radon
$L_p(\mathcal{X}, \mu),\ \mathcal{X}$ compacto, $p \in [1, \infty)$, μ medida finita [Hornik *et al.* 1989]	L_p	ψ sigmoide medible, monótona	Aprox. Trigonométrica
$L_p(\mathbb{R}^d, \mu),\ p \in [1, \infty)$ [Hornik 1991]	L_p	ψ acotada μ medida finita	Existencial Th. Hahn-Banach
$L_p(\mu), p \in [1, \infty)$ $\mu \in [0, \infty)$ sop. comp. abs. cont. resp. μ_{L} [Leshno *et al.* 1993]	L_p	$\psi \in L_\infty^{\text{loc}}(\mathbb{R}^d)$, $\mu_{\text{Leb.}}(\text{cl}(\Upsilon_\psi)) = 0$ con Υ_ψ cto. discont. ψ no polinomio c.p.d	Existencial
$\mathcal{C}^r(\mathcal{X}, \mu),\ \mathcal{X}$ compacto [Hornik 1991]	$L_{p,r}$ $L_{\infty,r}$	$\psi \in \mathcal{C}^r(\mathcal{X})$ acotada μ medida finita	Existencial Th. Hahn-Banach
$\mathcal{C}^r(\mathbb{R}^d, \mu)$ [Hornik 1991]	$L_{p,r}$	$\psi \in \mathcal{C}^r(\mathbb{R}^d)$ $D^{\boldsymbol{\alpha}}\psi$ acotada, $\|\boldsymbol{\alpha}\| \leq r$ μ medida finita	Existencial Th. Hahn-Banach

Tabla 41. Resultados más relevantes sobre las posibilidades aproximadoras de las redes MLP. En todos los casos se supone que ψ es no constante.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$\mathbf{x}_0 \in \mathcal{X}$. Entonces:

$$\psi_{a_1}(\mathbf{x}, \mathbf{x}_1)\psi_{a_2}(\mathbf{x}, \mathbf{x}_2) = C\psi_{a_1+a_2}(\mathbf{x}, \mathbf{x}_0) \text{ con } C = \exp\{-(a_1 a_2 (a_1 + a_2)^{-1} \|\mathbf{x}_1 - \mathbf{x}_2\|^2\}$$

■

Teorema 160 *[Hartman* et al. *1990], [Girosi y Poggio 1990, prop. B1] La familia $\mathcal{F}$ es uniformemente densa en $\mathcal{C}(\mathcal{X})$ con $\mathcal{X} \subset \mathbb{R}^d$ convexo y compacto.*

Demostración. Como las funciones ψ_{a_i} no se anulan en ningún punto y $\mathcal{F}$ separa cualesquiera dos puntos de $\mathcal{X}$ debido a la propiedad monótona de las exponenciales. basta aplicar el teorema 144 de Stone.

■

El segundo enfoque es válido para funciones de transferencia más generales y utiliza el siguiente lema fundamental (véase también [DeVore y Lorentz 1993, th. 2.1 y nota 6.1.]) en el que se denota la convolución mediante $(\phi * f)(\mathbf{x}) = \int f(\mathbf{x} - \mathbf{t})\phi(\mathbf{t})d\mathbf{t}$.

Lema 161 *[Park y Sandberg 1991, lema 1] Sea $f \in L_p(\mathbb{R}^d)$, $p \in [1, \infty)$ y sea $\phi : \mathbb{R}^d \to \mathbb{R}$ una función integrable con $\int_{\mathbb{R}^d} \phi(\mathbf{x})d\mathbf{x} = 1$. Sea $\phi_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ con $\phi_\varepsilon(\mathbf{x}) = \frac{1}{\varepsilon^d}\phi(\frac{\mathbf{x}}{\varepsilon})$ con $\varepsilon > 0$. Entonces: $\|\phi_\varepsilon * f - f\|_{L_p} \to 0$ cuando $\varepsilon \to 0$.*

Para el siguiente teorema denotamos:

$$\mathcal{F}_\psi = \{\sum_{i=1}^{h} c_i \psi(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma_i}) : \mathbf{c} \in \mathbb{R}^h, \mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d, \sigma_i \in \mathbb{R}^+, i = 1 : h \in \mathbb{N}^+\}$$

y por $\mathcal{F}_\psi^0$ la familia análoga en la que $\sigma_i = \sigma$ para todo $i = 1 : h$.

Teorema 162 *Sea $\psi : \mathbb{R}^d \to \mathbb{R}$ integrable. Se verifican los siguientes resultados:*

1. *[Park y Sandberg 1991, th. 1] Si ψ es acotada y continua c.p.d con $\int_{\mathbb{R}^d} \psi(\mathbf{x})d\mathbf{x} \neq 0$, entonces $\mathcal{F}_\psi^0$ es densa en $L_p(\mathbb{R}^d)$ $\forall p \in [1, \infty)$.*

2. *[Park y Sandberg 1993, th. 1] $\mathcal{F}_\psi^0$ es densa en $L_1(\mathbb{R}^d) \iff \int_{\mathbb{R}^d} \psi(\mathbf{x})d\mathbf{x} \neq 0$ ([142]).*

3. *[Park y Sandberg 1993, th. 2] $\mathcal{F}_\psi$ es densa en $L_1(\mathbb{R}^d) \iff \int_{\mathbb{R}^d} \psi(\mathbf{x})d\mathbf{x} \neq 0$.*

4. *[Park y Sandberg 1993, th. 3] $\mathcal{F}_\psi$ es densa en $L_2(\mathbb{R}^d) \iff \forall B \subset \mathbb{R}^d : \mu_L(B) > 0$, entonces $\tilde{\psi}(\sigma\cdot) \neq 0$ c.p.d en B.*

5. *[Park y Sandberg 1993, prop. 1] Si $\int_{\mathbb{R}^d} \psi(\mathbf{x})d\mathbf{x} \neq 0$ y $\int_{\mathbb{R}^d} |\psi(\mathbf{x})|^p d\mathbf{x} < \infty$, entonces $\mathcal{F}_\psi$ es densa en $L_p(\mathbb{R}^d)$, $p \in (1, \infty)$.*

Por ejemplo, el resultado 1 se obtiene de la siguiente forma: teniendo en cuenta que las funciones continuas con soporte compacto son densas en $L_p(\mathbb{R}^d)$, basta obtener la densidad respecto a dichas funciones: si ψ es una función continua y acotada, se aproximan dichas funciones utilizando la aproximación integral del lema 161 con la función $\phi(\mathbf{x}) = \frac{1}{C}\psi(\mathbf{x})$ con $C = \int_{\mathbb{R}^d} \psi(\mathbf{x})d\mathbf{x}$. Posteriormente, se obtiene el resultado aproximando dicha representación integral mediante sumas de Riemann.

[142] En este y en los tres siguientes resultados de este teorema, no se requiere que ψ sea continua o acotada.

Teorema 163 *[Park y Sandberg 1993, th. 4 y 5] Sea $\psi : \mathbb{R}^d \to \mathbb{R}$ integrable. Entonces, $\mathcal{F}_\psi$ es uniformemente denso en $\mathcal{C}(\mathcal{X})$ $\forall \mathcal{X} \subset \mathbb{R}^d$ compacto, si se verifica una de las tres siguientes condiciones (las dos últimas son casos particulares de la primera):*

1. *ψ continua y $\tilde{\psi}^{-1}(0)$ no contiene un subespacio propio.*
2. *ψ continua y satisface $\int_{\mathbb{R}^d} \psi(\mathbf{x})d\mathbf{x} \neq 0$.*
3. *$\psi \neq 0$ continua y radial[143] con respecto a la norma euclídea.*

Nótese que en los resultados de los dos teoremas anteriores no se requiere que ψ sea radial salvo en la tercera condición del teorema anterior.

Por último se tiene el siguiente resultado de [Krzyzak *et al.* 1996] que extiende la propiedad de aproximación universal de las redes RBF a los espacios $L_p(\mathcal{X}, \mu)$ $p \in (0, \infty)$ donde μ es una medida de probabilidad.

Teorema 164 *[Krzyzak* et al. *1996, th. 1] Sea $\psi : \mathbb{R}^d \to \mathbb{R}$ acotada con $\psi \in L_1(\mathbb{R}^d) \cap L_p(\mathbb{R}^d)$ para algún $p \in [1, \infty)$, e integrable con $\int_{\mathbb{R}^d} \psi(\mathbf{x})d\mathbf{x} \neq 0$. Sea μ una medida de probabilidad arbitraria en $\mathbb{R}^d$ y sea $q \in (0, \infty)$. Entonces, $\mathcal{F}_\psi$ es densa simultáneamente en $L_q(\mu)$ y $L_p(\mathbb{R}^d)$.*

La tabla 42 refleja los resultados anteriores y otros relevantes sobre las posibilidades aproximadoras de las redes RBF, así como de las redes RRBF y SVM (los tres últimos resultados) cuyas propiedades aproximadoras presentamos a continuación.

Redes de Regularización (RRBF)

Dado un conjunto de datos $\mathbf{z}^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ tal que $y_i = \mathsf{f}(\mathbf{x}_i)$ siendo $\mathsf{f} \in \mathcal{C}(\mathcal{X})$ una función que se desea aproximar, el siguiente teorema establece las propiedades aproximadoras de las redes RBF de Regularización (RRBF) de la forma:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \psi(\mathbf{x}, \mathbf{x}_i) + g(\mathbf{x}) \tag{C.9}$$

donde g es una función polinómica que pertenece al espacio nulo de un operador diferencial lineal P y ψ es la función de Green del operador diferencial P^+P con P^+ el operador adjunto de P, es decir, solución de la ecuación diferencial $P^+P\psi(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}')$ con δ la función delta de Dirac.

Teorema 165 *[Girosi y Poggio 1990, prop. C.1] Sea $\mathcal{T} = \mathcal{C}(\mathcal{X})$ con $\mathcal{X} \subset \mathbb{R}^d$ compacto y sea ψ una función continua a trozos función de Green de algún operador diferencial P^+P (autoadjunto) siendo P un operador diferencial lineal definido en algún espacio de funciones $\mathcal{H}_P$ compatibles con P y sus condiciones de contorno. Entonces, $\forall \mathsf{f} \in \mathcal{T} = \mathcal{C}(\mathcal{X})$ se verifica que $\forall \varepsilon > 0, \exists f_h(\mathbf{x}) = \sum_{i=1}^{h} c_i \psi(\mathbf{x}, \mathbf{x}_i) \in \mathcal{H}_P$ tal que $\forall \mathbf{x} \in \mathcal{X}$:*

$$|\mathsf{f}(\mathbf{x}) - f_h(\mathbf{x})| < \varepsilon$$

Es importante destacar que del teorema anterior se deduce que, aunque el espacio nulo del operador P no sea el espacio trivial, el término polinómico $g(\mathbf{x})$ de la expresión (C.9) no es necesario para conseguir la propiedad de aproximación universal de las redes RRBF en $\mathcal{T} = \mathcal{C}(\mathcal{X})$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[143] Radial en el sentido de que $\|\mathbf{x}\| = \|\mathbf{x}'\| \Rightarrow \psi(\mathbf{x}) = \psi(\mathbf{x}')$ [Park y Sandberg 1993, p. 307].

Propiedades Aproximadoras de la red RBF: $f(\mathbf{x}) = \sum_{j=1}^{h} c_j k_j(\mathbf{x}, \mathbf{w}_j)$			
Espacio Funcional	$\|\cdot\|$	**Condiciones**	**Enfoque**
$\mathcal{C}(\mathcal{X})$, $\mathcal{X}$ abierto, acotado [Jackson 1988]	L_1	$k_i(\mathbf{x}, \mathbf{w}_i) = \zeta(\|\mathbf{x} - \mathbf{w}_i\|)$ $\zeta(az) = a^\beta \zeta(z)$ $a, z \in \mathbb{R}^+, \beta \in \mathbb{R}$ $\{\mathbf{w}_i\}_{i=1}^{\infty}$ denso en $\mathrm{cl}(\mathcal{X})$	Existencial
$\mathcal{C}(\mathcal{X})$, $\mathcal{X}$ compacto y convexo [Hartman *et al.* 1990]	L_∞	$k_i(\mathbf{x}, \mathbf{w}_i) = e^{-a_i \|\mathbf{x} - \mathbf{x}_i\|^2}$	Existencial Stone-Weieres.
$\mathcal{C}(\mathbb{R}^d)$ [Park y Sandberg 1991]	L_∞	$k_j(\mathbf{x}, \mathbf{w}_i) = \psi(\mathbf{x} - \mathbf{w}_i)$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ ψ acotada, continua c.p.d	Existencial
$\mathcal{C}(\mathcal{X})$, $\mathcal{X}$ compacto [Park y Sandberg 1993]	L_∞	$k_j(\mathbf{x}, \mathbf{w}_i) = \psi(\frac{\mathbf{x} - \mathbf{w}_j}{\sigma_j})$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ ψ continua	Existencial
$L_p(\mathbb{R}^d)$ y $L_p^{\mathrm{loc}}(\mathbb{R}^d)$, $p \in [1, \infty)$ [Park y Sandberg 1991]	L_p	$k_i(\mathbf{x}, \mathbf{w}_i) = \psi(\mathbf{x} - \mathbf{w}_i), \forall i$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ ψ acotada, continua c.p.d	Existencial Convolución
$L_p(\mathbb{R}^d)$, $p \in [1, \infty)$ [Park y Sandberg 1993]	L_p	$k_i(\mathbf{x}, \mathbf{x}_i) = \psi(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma_i})$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ $\int_{\mathbb{R}^d} \|\psi(\mathbf{x})\|^p d\mathbf{x} < \infty$	Existencial Convolución
$L_p(\mathbb{R}^d, \mu)$, $p \in (0, \infty)$ μ medida de probabilidad [Krzyzak *et al.* 1996]	L_p	$k_i(\mathbf{x}, \mathbf{x}_i) = \psi(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma_i})$ $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \in (0, \infty)$ $\psi \in L_1(\mathbb{R}^d) \cap L_p(\mathbb{R}^d)$	Existencial Convolución
$\mathcal{T} = \{f = \psi * \lambda : \lambda \in L^1(\mathbb{R}^d)\}$ [Girosi y Anzellotti 1993]	L_∞	$\psi(\mathbf{x}) = e^{-\|\mathbf{x}\|^2}$ $k_i(\mathbf{x}, \mathbf{w}_i) = \psi(\mathbf{x} - \mathbf{w}_i)$	Lema Maurey
$\mathcal{W}_1^{2r}(\mathbb{R}^d), 2r > d$ [Girosi y Anzellotti 1993]	L_∞	$k_i(\mathbf{x}, \mathbf{w}_i) = g_r(\|\mathbf{x} - \mathbf{w}_i\|^2)$ $g_r = \tilde{g}$ potencial de Bessel i.e. $g(\boldsymbol{\omega}) = (1 + \|\boldsymbol{\omega}\|^2)^{-r/2}$	Lema Maurey
$\mathcal{W}_1^{2r}(\mathbb{R}^d), 2r > d$ [Girosi 1994]	L_2	$k_i(\mathbf{x}, \mathbf{w}_i) = e^{-\|\mathbf{x} - \mathbf{w}_i\|^2 / \sigma_i^2}$	Lema Maurey
$\mathcal{C}(\mathcal{X}), \mathcal{X} \subset \mathbb{R}^d$ compacto [Girosi y Poggio 1990]	L_∞	$\min_{\mathcal{H}_P} \{\hat{R}_n(f) + \lambda \|Pf\|_{L_2}^2\}$ $k_i(\mathbf{x}, \mathbf{w}_i) = k(\mathbf{x}, \mathbf{x}_i)$ k función de Green de P^+P k continua a trozos	Existencial Propieades de la F. Green
$\mathcal{C}(\mathcal{X}), \mathcal{X} \subset \mathbb{R}^d$ compacto [Hammer y Gersmann 2003]	L_∞	1) $k(\mathbf{x}, \mathbf{x}') = \psi(\frac{1}{\sigma}\mathbf{x} - \mathbf{x}')$ con: $\int \psi(\mathbf{x}) d\mathbf{x} \neq 0$ y k acotada y continua 2) $k(\mathbf{x}, \mathbf{x}') = \kappa(\langle \mathbf{x}, \mathbf{x}' \rangle + \alpha)$ κ analítica, no polinomio continua y $\exists j : \kappa^{(j)}(\alpha) \neq 0$	1) Aprox. RBF 2) Aprox. MLP
$\mathcal{M}(\mathcal{X}, \mathbb{P}_{\mathsf{X}}) = \{\mathsf{g}\ \mathbb{P}_{\mathsf{X}}\text{-medibles}\}$ con $\mathsf{g} : \mathbb{R}^d \to \{-1, 1\}$ $\mathcal{X} \subset \mathbb{R}^d$ compacto [Hammer y Gersmann 2003]	$\mathbb{P}_{\mathsf{X}}[g \neq \mathsf{g}]$	1) $k(\mathbf{x}, \mathbf{x}') = \psi(\frac{1}{\sigma}\mathbf{x} - \mathbf{x}')$ con: $\int \psi(\mathbf{x}) d\mathbf{x} \neq 0$ 2) $k(\mathbf{x}, \mathbf{x}') = \kappa(\langle \mathbf{x}, \mathbf{x}' \rangle + \alpha)$ κ analítica y no polinomio y $\exists j : \kappa^{(j)}(\alpha) \neq 0$	1) Aprox. RBF 2) Aprox. MLP

Tabla 42. Propiedades aproximadoras más relevantes de las redes RBF.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Redes SVM

Sea $\mathcal{F}_k$ la familia de las redes SVM para regresión construidas con núcleo k mediante la resolución del correspondiente programa cuadrático con restricciones, utilizando la pérdida ε-insensible de Vapnik, y sea $\mathcal{F}_k^c$ la familia análoga en el problema de clasificación, es decir, compuesta por las soluciones al problema cuadrático con restricciones que resulta de maximizar el margen con un enfoque *soft margin* (con variables *slack*).

Los siguientes resultados expresan las propiedades aproximadoras de las redes SVM, correspondientes a los tipos de núcleo más frecuentemente utilizados (núcleos de producto interno y núcleos invariantes por traslaciones). En ellos, se define las siguientes propiedades de aproximación universal:

1. Una familia $\mathcal{F}$ de funciones $f : \mathbb{R}^d \to \mathbb{R}$ (problema de regresión) posee la propiedad de aproximación universal si $\mathcal{F}$ es uniformemente densa en $\mathcal{C}(\mathcal{X})$ con $\mathcal{X}$ subconjunto compacto de $\mathbb{R}^d$.

2. Una familia $\mathcal{G}$ de funciones $f : \mathbb{R}^d \to \{-1, 1\}$ (problema de clasificación) posee la propiedad de aproximación universal con respecto a una medida de probabilidad $\mathbb{P}_{\mathsf{X}}$ si cualquier función medible con valores en $\{-1, 1\}$ puede ser aproximada mediante funciones de $\mathcal{G}$, es decir: para toda función $\mathsf{g} : \mathbb{R}^d \to \{-1, 1\}$ medible respecto a $\mathbb{P}_{\mathsf{X}}$, y $\forall \eta \in (0, 1)$, existe una función $g \in \mathcal{G}$ tal que $\mathbb{P}_{\mathsf{X}}[\mathsf{g}(\mathsf{X}) \neq g(\mathsf{X})] < \eta$.

Sean $\mathcal{F}_k$ y $\mathcal{G}_k$ familias de funciones implementadas por redes SVM con núcleo k, para regresión y clasificación, respectivamente. Se verifican los siguientes resultados (véanse los teoremas 135 y **??** del Anexo B):

Teorema 166 *[Hammer y Gersmann 2003, th. 1] Sea k un núcleo medible de la forma $k(\mathbf{x}, \mathbf{x}') = \psi(\langle \mathbf{x}, \mathbf{x}' \rangle + \alpha)$ donde ψ es una función no polinómica y analítica en un intervalo abierto B con $\alpha \in B$ y tal que todas las derivadas de ψ no son nulas en α. Entonces:*

1. *$\mathcal{G}_k$ posee la propiedad de aproximación universal.*

2. *Si además, k es continua, $\mathcal{F}_k$ posee la propiedad de aproximación universal.*

Teorema 167 *[Hammer y Gersmann 2003, th. 2] Sea k un núcleo de la forma $k(\mathbf{x}, \mathbf{x}') = \psi(\frac{1}{\sigma}(\mathbf{x} - \mathbf{w}))$ con $\psi : \mathbb{R}^d \to \mathbb{R}$ integrable con $\int_{\mathbb{R}^d} \psi(\mathbf{x}) d\mathbf{x} \neq 0$. Entonces:*

1. *$\mathcal{G}_k$ posee la propiedad de aproximación universal.*

2. *Si además, k es acotada y continua, $\mathcal{F}_k$ posee al propiedad de aproximación universal.*

Los dos teoremas anteriores se demuestran basándose en las propiedades aproximadoras de las redes neuronales MLP y de las redes RBF, respectivamente.

Para el caso de núcleos del tipo polinómico, se tiene el siguiente teorema:

Teorema 168 *[Hammer y Gersmann 2003, th. 3] Sea $\mathcal{F} = \bigcup_{r=1}^{\infty} \mathcal{F}_{k_r}$ y $\mathcal{G} = \bigcup_{r=1}^{\infty} \mathcal{G}_{k_r}$ donde $k_r(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + 1)^r$ con $r \in \mathbb{N}^+$. Entonces $\mathcal{F}$ y $\mathcal{G}$ poseen la propiedad de aproximación universal.*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Nótese que si no se utiliza la unión infinita de las familias $\mathcal{F}_{k_r}$ y $\mathcal{G}_{k_r}$, el núcleo polinómico k_r produciría sólo expresiones polinómicas de grado r que no garantizarían las propiedades de

aproximación universal. Pero estas propiedades se consiguen admitiendo polinomios de grado ilimitado.

Propiedad de la Mejor Aproximación con Redes Neuronales

Sea $\mathcal{T} = \mathcal{C}(\mathcal{X})$ con $\mathcal{X} \subset \mathbb{R}^d$ el espacio de las funciones continuas. Respecto a este espacio, las redes MLP no poseen la propiedad de la mejor aproximación:

Teorema 169 *[Girosi y Poggio 1990, prop. 4.1] Sea* $\mathcal{F}_h = \{\sum_{i=1}^h c_i \psi(\mathbf{w}_i^T \mathbf{x} + w_{i0}) : \mathbf{c}, \mathbf{w}_i \in \mathbb{R}^d, w_{i0} \in \mathbb{R}, i = 1 : h\}$ *donde* ψ *es una función continua no lineal. Entonces,* $\mathcal{F}_h$ *no posee la propiedad de la mejor aproximación.*

Sin embargo, dicha propiedad si la verifica cualquier esquema de aproximación basado en funciones básicas prefijadas: $\mathcal{F}_h = \{\sum_{i=1}^h c_i \psi_i(\mathbf{x}) : \mathbf{c} \in \mathbb{R}^d, \psi_i \in \mathcal{T} = \mathcal{C}(\mathcal{X}),\ i = 1 : h,$ independientes$\}$.

Teorema 170 *[Girosi y Poggio 1990, prop. 4.2] Sea* $\mathcal{F}_h$ *como anteriormente, donde* $\psi \in \mathcal{T} = \mathcal{C}(\mathcal{X})$. *Entonces,* $\mathcal{F}_h$ *con* $h \geq 1$ *es un conjunto de existencia para el problema de la mejora aproximación (C.2).*

Demostración. Sea $\mathsf{f} \in \mathcal{T} = \mathcal{C}(\mathcal{X})$. Si $f_0 \in \mathcal{F}_h$ arbitraria, el conjunto $\mathcal{F}_0 = \{f \in \mathcal{F}_h : \|f - \mathsf{f}\| \leq \|f_0 - \mathsf{f}\|\} \subset \mathcal{F}_h$ es cerrado y acotado, luego compacto, con lo que aplicando el teorema 147.2 a $\mathcal{F}_0$ se tiene que $\mathcal{F}_h \supset \mathcal{F}_0$ es conjunto de existencia.

■

El siguiente teorema aporta una condición suficiente para conjunto de unicidad cuando el esquema de aproximación es una combinación lineal de funciones básicas: $\mathcal{F}_h = \{\sum_{i=1}^h c_i \psi_i : \mathbf{c} \in \mathbb{R}^h, \psi_i \in \mathcal{T}, i = 1 : h,$ independientes$\}$.

Teorema 171 *[Girosi y Poggio 1990, Prop. 4.3] Si el espacio normado* $\mathcal{T} = \mathcal{C}(\mathcal{X})$ *es convexo*[144]*, entonces* $\mathcal{F}_h$ *con* $h \geq 1$, *definida como anteriormente es un conjunto de unicidad.*

Los resultados anteriores determinan que, cuando la red RBF posee centros $\{\mathbf{w}_j\}_{j=1}^h$ preseleccionados de algún conjunto de datos $\{\mathbf{w}_j\}_{j=1}^h \subset \{\mathbf{x}_i\}_{i=1}^n$ y los parámetros de escala están también prefijados, el esquema de aproximación resultante posee la propiedad de la mejor aproximación.

Sin embargo, si los centros $\{\mathbf{w}_j\}_{j=1}^h$ son resultado de un proceso de optimización no lineal de adaptación a los datos (algoritmo *back-propagation* o gradiente conjugado), la familia aproximadora resultante, pierde dicha propiedad al igual que las redes MLP (la familia aproximadora deja de ser cerrada [Girosi y Poggio 1990, p. 173] y no puede aplicarse el teorema 147.2).

Tasas de Convergencia con Redes Neuronales

Los resultados anteriores muestran las posibilidades aproximadoras de las redes neuronales RBF y MLP pero no aportan el orden de aproximación. En consecuencia, sin estas tasas de convergencia no es posible obtener tasas para la consistencia estadística de la estimación mediante redes neuronales.

[144] Un caso particular es $\mathcal{T} = \mathcal{C}(\mathcal{X})$ con las normas $\|\cdot\|_{L_p}$, $p \in (1, \infty)$.

Los enfoques más relevantes utilizados para abordar el problema son de tres tipos:

1. El primer enfoque consiste en fijar una familia de interés $\mathcal{T}$ y determinar el orden de aproximación a dicha familia con un esquema de aproximación determinado.

2. El segundo enfoque, más ambicioso, consiste en recopilar bajo una misma familia todas las funciones que poseen un mismo orden de aproximación con un esquema de aproximación determinado.

3. Pero la única forma de evitar la dependencia del orden de aproximación en términos de la dimensión d del espacio de entrada consiste en suponer familias con un grado de regularidad creciente con la dimensión del espacio de entrada (ecuación (C.7) y discusión al respecto).

De cualquier forma, los resultados son escasos, y ésta es un área de investigación que no puede considerarse ni mucho menos culminada. En lo que sigue presentamos los resultados más relevantes al respecto.

Con respecto al primer enfoque destacamos los resultados que presentamos a continuación. El primero, debido a [Mhaskar 1996], proporciona órdenes de aproximación para redes neuronales que denomina **redes de traslación** (*translation newtorks*) constituidas según la siguiente forma general:

$$f_h(\mathbf{x}) = \sum_{i=1}^{h} c_i \psi(\mathbf{W}_i \mathbf{x} + \mathbf{w}_{0i})$$

donde $f_h : \mathbb{R}^d \to \mathbb{R}$, $\psi : \mathbb{R}^s \to \mathbb{R}$, $\mathbf{W}_i$ son matrices $s \times d$, $\mathbf{w}_{0i} \in \mathbb{R}^s$ y $c_i \in \mathbb{R}$, $i = 1 : h$.

Es claro que estas redes poseen como caso particular las redes MLP (con $s = 1, \mathbf{W}_i = \mathbf{w}_i \in \mathbb{R}^d$ y $\mathbf{w}_{0i} = w_{0i} \in \mathbb{R}$) y las redes RBF (con $\mathbf{W}_i = \mathbf{I}_d$). Al igual que hasta ahora, denotamos el conjunto de tales funciones mediante $\mathcal{F}_h$.

Teorema 172 *[Mhaskar 1996, th. 2.1] [Pinkus 1999, th. 6.8] Sea $\psi : \mathbb{R}^d \to \mathbb{R}$, $\psi \in \mathcal{C}^\infty(\mathbb{B})$ con $\mathbb{B} \subset \mathbb{R}^d$ abierto. Si existen $\mathbf{x}_0 \in \mathbb{B}$, $\boldsymbol{\alpha} \in \mathbb{N}^d$ tal que $D^{\boldsymbol{\alpha}}\psi(\mathbf{x}_0) \neq 0$, entonces, existen $\{\mathbf{W}_j\}_{j=1}^h$ matrices $d \times s$ con $d \leq s$ tal que, $\forall \mathsf{f} \in \mathcal{T} = \mathcal{W}_p^r([-1,1]^d)$, $p \in [1, \infty]$, existen $c_i(\mathsf{f}) \in \mathbb{R}$ funcionales continuos, $i = 1 : h$, tal que:*

$$\|\mathsf{f} - f_h\|_{L_p} \leq C h^{-r/d} \|\mathsf{f}\|_{\mathcal{W}_p^r}$$

[Mhaskar 1996] indica que si la función a aproximar se considera en L_p en lugar de en $\mathcal{W}_p^r$ se obtienen cotas similares pero en términos del módulo de continuidad como es estándar en teoría de aproximación.

Para el espacio normalizado $\mathcal{V}_p^r(\mathcal{X}) = \{f \in \mathcal{W}_p^r(\mathcal{X}) : \|f\|_{\mathcal{W}_p^r} \leq 1\}$ la cota anterior posee la forma $Ch^{-r/d}$. Esta tasa es la mejor que puede obtenerse con esta familia (véase [DeVore *et al.* 1989, th. 4.2] que basándose en el concepto de n-width no lineal, demuestra que dicha cota no puede superarse).

De cualquier forma, como indica [Mhaskar 1996] el orden del n-width supone que los parámetros se determinan mediante algún operador continuo (algoritmo de *selección continua*), lo cual no está garantizado en la práctica [Kainen *et al.* 1999]. Por tanto, la comparación anterior debe considerarse ilustrativa y no como una propiedad matemática firme de los algoritmos utilizados en la práctica.

USC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Para el siguiente resultado, se define una función **sigmoide de grado** $s \in \mathbb{Z}^+$ como aquella función $\psi : \mathbb{R} \to \mathbb{R}$ continua o continua a trozos, que verifica la condición [Mhaskar y Micchelli 1992, def. 4.1]:

$$\lim_{t\to-\infty} \frac{\psi(t)}{t^s} = 0, \quad \lim_{t\to\infty} \frac{\psi(t)}{t^s} = 1 \tag{C.10}$$

Como puede observarse, para $s = 0$, se tiene una función sigmoide usual. En el siguiente teorema denotamos $\mathcal{V}_p^r(\mathbb{B}_d) = \{f \in \mathcal{W}_p^r(\mathbb{B}_d) : \|f\|_{\mathcal{W}_p^r} \leq 1\}$ con $\mathbb{B}_d = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\}$.

Teorema 173 *[Pinkus 1999, cor. 6.11] Sea ψ una función sigmoide de grado s. Entonces, $\forall \mathsf{f} \in \mathcal{T} = \mathcal{V}_p^r(\mathbb{B}_d)$ con $r \in \{1, ..., s + 1 + \frac{d-1}{2}\}$ se verifica que:*

$$\delta(\mathsf{f}, \mathcal{F}_h)_{L_2} \leq Ch^{-r/d}$$

El siguiente teorema obtiene órdenes de convergencia para funciones definidas en todo $\mathbb{R}^d$ (no sólo en subconjuntos propios) en el contexto del espacio $\mathcal{K} = L_p(\mathbb{R}^d, \mu)$ con μ medida que se especifica a continuación.

Teorema 174 *[Maiorov y Meir 1998, th. 3.2] Sea $\mu(\mathbf{x}) = \prod_{i=1}^d e^{-|x_i|^q}$, $q \geq 2$ y $p \in [1, \infty]$. Entonces, $\forall \mathsf{f} \in \mathcal{T} = \{f \in \mathcal{W}_p^r(\mathbb{R}^d, \mu) : \|f\|_{L_p(\mathbb{R}^d,\mu)} \leq M\}$, se verifica que:*

$$\delta(\mathsf{f}, \mathcal{F}_h)_{L_{p,\mu}} \leq Ch^{-(1-1/q)r/d}$$

Ordenes *Greedy* de Aproximación

En relación con el tercer tipo de enfoque, gran parte de los resultados existentes siguen el camino iniciado por [Barron 1993] basado en el lema de Maurey y el algoritmo *greedy* de [Jones 1987] que determina un orden de convergencia $O(h^{-1/2})$ para la aproximación mediante combinaciones convexas de funciones de una familia $\mathcal{G}$ acotada en un espacio de Hilbert $\mathcal{K}$, ($\|\psi\|_{\mathcal{K}} \leq M < \infty$, $\forall \psi \in \mathcal{G}$).

Para el siguiente resultado, dada la familia $\mathcal{G}$, sea $\mathcal{T}_{\mathcal{G}} = \mathrm{cl}(\mathrm{co}(\mathcal{G})) \subset \mathcal{K}$ la clausura del conjunto de las combinaciones convexas de elementos de $\mathcal{G}$ y sea $\mathcal{F}_h$, $h \in \mathbb{N}^+$ el conjunto de las combinaciones convexas de h elementos de $\mathcal{G}$.

Lema 175 *(Maurey) [Barron 1993] Sea $\mathcal{K}$ un espacio de Hilbert con producto interno $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ y norma $\|\cdot\|_{\mathcal{K}}$. Sea $\mathcal{G} \subset \mathcal{K}$ con $\|\psi\|_{\mathcal{K}} \leq b$ $\forall \psi \in \mathcal{G}$ y sean $\mathcal{F}$, $\mathcal{F}_h$ $\forall h \in \mathbb{N}^+$ definidos como anteriormente. Entonces, dada $\mathsf{f} \in \mathcal{T}_{\mathcal{G}}$, y fijado $h \in \mathbb{N}^+$ se verifica que, para toda constante $C > b_{\mathsf{f}}^2 \doteq b^2 - \|\mathsf{f}\|_{\mathcal{K}}$, existe un $f_h \in \mathcal{F}_h$ tal que:*

$$\|f_h - \mathsf{f}\|_{\mathcal{F}}^2 \leq \frac{C}{h}$$

[Jones 1987] aporta la primera demostración constructiva a este resultado, que se mejoran o amplían posteriormente en [Jones 1992], [Barron 1993], [Dingankar y Sandberg 1996], [Lee *et al.* 1996, th. 2], [Docampo *et al.* 1997], [Fine 1999, th. 4.11.1], [Zhang 2002]:

Teorema 176 *[Docampo* et al. *1997, th. 2.1] En el marco del teorema anterior, sea $\mathsf{f} \in \mathcal{T}_{\mathcal{G}}$ y sea:*

$$\gamma = \inf_{g\in\mathcal{K}} \sup_{\psi\in\mathcal{G}} \{\|\psi - g\|_{\mathcal{K}}^2 - \|\mathsf{f} - g\|_{\mathcal{K}}^2\} \tag{C.11}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Sea C una constante tal que $C > \gamma$. Entonces, dado $\mathsf{f} \in \mathcal{T}_{\mathcal{G}}$, es posible construir de forma iterativa una sucesión $\{f_h\}$ donde $f_h = (1-\beta_h)f_{h-1} + \beta_h\psi_h$ con $\beta_h \in [0,1]$, $\psi_h \in \mathcal{G}$, $h \in \mathbb{N}^+$ tal que[145]*:*

$$\|f_h - \mathsf{f}\|_{\mathcal{K}}^2 \leq \inf_{\psi\in\mathcal{G}} \|(1-\beta_h)f_{h-1} + \beta_h\psi - \mathsf{f}\|_{\mathcal{K}}^2 + \varepsilon_h \tag{C.12}$$

de tal forma que:

$$\|f_h - \mathsf{f}\|_{\mathcal{K}}^2 \leq \frac{C}{h}$$

Como parámetro γ para elegir C basta tomar el parámetro b_{f}^2 del lema de Maurey pues, tomando $g = 0 \in \mathcal{K}$, la condición (C.11) resulta:

$$\gamma \leq \sup_{\psi\in\mathcal{G}}\{\|\psi\|_{\mathcal{K}}^2 - \|\mathsf{f}\|_{\mathcal{K}}^2\} \leq b^2 - \|\mathsf{f}\|_{\mathcal{K}} = b_{\mathsf{f}}^2$$

El algoritmo *greedy* cuya sucesión aproximadora se construye de la forma $f_j = f_{j-1} + \varphi_j$ se denomina algoritmo *greedy* puro [Jones 1987]. El teorema anterior es la versión *relajada* [Jones 1992] del algoritmo *greedy*. Casos particulares son las versiones de [Barron 1993] con $\beta_h = b_{\mathsf{f}}^2/(b_{\mathsf{f}}^2 + \|f_{h-1} - \mathsf{f}\|_{\mathcal{K}}^2)$, [Dingankar y Sandberg 1996] con $\beta_h = \frac{1}{h}$, y [Lee *et al.* 1996, th. 2] y [Fine 1999, th. 4.11.1] con $\beta_h = \frac{2}{h}$.

[Docampo *et al.* 1997] analizan extensamente el algoritmo para cotas más generales del tipo:

$$\|f_h - \mathsf{f}\|_{\mathcal{K}}^2 \leq \frac{C}{b(h)}$$

con los siguientes resultados:

1. Si el error permitido en cada iteración (C.12) es de la forma $\varepsilon_h = \beta_h^2(C-\gamma)$ entonces para un β_h dado, la tasa máxima de convergencia es $C/b(h)$, que se obtiene para $b(h) = h$. Además, la combinación óptima es $\beta_h = 1/h$.

2. Si se optimiza en β_h la cota al error $\|f_h - \mathsf{f}\|_{\mathcal{K}}^2$ en (C.12), la cota óptima satisface [Docampo *et al.* 1997, ec. 1.11]:
$$\frac{C}{1+\alpha+b(h-1)} + \varepsilon_h = \frac{C}{b(h)}$$
con $\alpha = (C-\gamma)/\gamma$, y se alcanza cuando:
$$\beta_h = \frac{1+\alpha}{1+\alpha+b(h-1)}$$
de donde se deduce la siguiente limitación fundamental en la tasa de convergencia:
$$b(h) - b(h-1) \leq 1 + \alpha = \frac{C}{\gamma}$$

3. Con ello, si puede alcanzarse el error $\varepsilon_h = 0$ en cada iteración y se elige f_1 tal que $b(1) = 1+\alpha$, se obtiene:
$$b(h) = h(1+\alpha) \Rightarrow \frac{C}{b(h)} = \frac{C}{h(1+\alpha)} = \frac{\gamma}{h}$$

[145] Esta condición supone que $f_h \in \mathrm{co}(\{f_{h-1}, \psi_h\})$ y, por tanto, es más restrictiva que la original $f_{h-1} \in \mathrm{co}(\{\psi_j\}_{j=1}^{h-1})$ y $f_h \in \mathrm{co}(\{\psi_j\}_{j=1}^{h})$ del teorema [Docampo *et al.* 1997, th. 2.1] y otros similares [Jones 1992], pero para ella se obtienen las principales conclusiones y es frecuente en la práctica.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

es la mejor cota que puede alcanzarse y, en este caso, el β_h óptimo se corresponde con la media de los h elementos de $\mathcal{G}$ introducidos:

$$\beta_h = \frac{1+\alpha}{1+\alpha+(h-1)(1+\alpha)} = \frac{1+\alpha}{(1+\alpha)h} = \frac{1}{h}$$

4. Si $b(h) = h$, de lo anterior se obtiene que:

$$\beta_h = \frac{1+\alpha}{h+\alpha} \text{ y } \varepsilon_h = \frac{\alpha C}{h(h+\alpha)} = O(h^{-2})$$

siendo este valor de ε_h el más grande que puede permitirse en cada iteración para alcanzar la tasa $b(h) = h$.

[Lee *et al.* 1996, th. 2] y [Fine 1999, th. 4.11.1] extienden el teorema al caso más general en el que $\mathsf{f} \in \mathcal{K}$, es decir, que f pueda no ser no aproximable por funciones de $\cup_{h=1}^{\infty}\mathcal{F}_h$. En este caso, se obtiene fácilmente una cota del tipo:

$$\|f_h - \mathsf{f}\|_{\mathcal{K}}^2 - d_{\mathsf{f}}^2 \leq \frac{C}{h}$$
$$\text{con: } d_{\mathsf{f}} = \inf_{f \in \mathcal{T}_{\mathcal{G}}} \|f - \mathsf{f}\|_{\mathcal{K}}$$

Los resultados anteriores son válidos en espacios de Hilbert. Resultados sobre órdenes de convergencia en espacios generales de Banach pueden verse en [Donahue *et al.* 1997], del que extraemos el siguiente resultado general:

Teorema 177 *[Donahue* et al. *1997, cor. 2.6] Sea $\mathcal{K}$ un espacio L_p con $p \in [1,\infty)$ y $\mathcal{G} \subset \mathcal{K}$. Si dado $\mathsf{f} \in \mathcal{T}_{\mathcal{G}}$ se satisface que $\|\psi - \mathsf{f}\|_{L_p} \leq C_{\mathsf{f}}\ \forall \psi \in \mathcal{G}$ con $C_{\mathsf{f}} > 0$, y $f_h \in \mathcal{F}_h$, entonces se verifica que:*

$$\|\mathsf{f} - f_h\|_{\mathcal{K}} \leq C_{\mathsf{f}} C_p h^{-(1-1/t)}$$

donde $t = \min\{p, 2\}$ y $C_p = 1$ si $p \in [1,2]$ y $C_p = \sqrt{2}[\Gamma((p+1)/2)/\sqrt{h}]^{1/p}$ si $p \in (2,\infty)$, siendo Γ la función gamma.

Como puede observarse, si $p = 2$, el teorema anterior aporta una tasa de orden $O(h^{-1/2})$ como en el lema de Maurey.

Órdenes *Greedy* de Aproximación Mediante Redes MLP

En el contexto del espacio de Hilbert $\mathcal{K} = L_2(\mathcal{X}, \mu)$ donde μ es una medida de probabilidad y $\mathcal{X} = \mathbb{B}_{d,\tau} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \tau\}$ [Barron 1993] demuestra el siguiente teorema aplicable a la familia[146]:

$$\mathcal{T}_C(\mathbb{R}^d) = \{\mathsf{f} : \mathbb{R}^d \to \mathbb{R} : C_{\mathsf{f}} = \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_1 \left|\tilde{\mathsf{f}}(\boldsymbol{\omega})\right| d\boldsymbol{\omega} < \infty\} \tag{C.13}$$

restringida a $\mathcal{X} = \mathbb{B}_{d,\tau}$, lo que denotamos por $\mathcal{T}_C(\mathbb{B}_{d,\tau})$.

[146] Las funciones de esta familia poseen derivadas continuas y la transformada de su gradiente es integrable [Barron 1993]: $\nabla f(\mathbf{x}) = \int e^{i\boldsymbol{\omega}\mathbf{x}} \widetilde{\nabla f}(\boldsymbol{\omega}) d\boldsymbol{\omega}$. En [Pinkus 1999, p. 181] se indica la siguiente relación entre la familia y los espacios de Sobolev en $\mathbb{B}_d$: $\mathcal{W}_\infty^{d/2+2} \subseteq \mathcal{T}_C(\mathbb{B}_d) \subseteq \mathcal{W}_\infty^1 \subseteq \mathcal{W}_2^1$.

Teorema 178 *[Barron 1993, th. 2] Sea $\mathcal{K} = L_2(\mathcal{X}, \mu)$ con μ medida de probabilidad y $\mathcal{X} = \mathbb{B}_{d,\tau}$. Sea $\mathcal{G} = \{c\psi(\mathbf{w}^T\mathbf{x} + w_0) : |c| \leq 2C, \mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}\}$ donde ψ es una función sigmoide de grado 0. Entonces, $\forall \mathtt{f} \in \mathcal{T} = \mathcal{T}_C(\mathbb{B}_{d,\tau})$ se verifica que $\mathtt{f}(\mathbf{x}) - \mathtt{f}(\mathbf{0}) \in \mathcal{T}_{\mathcal{G}} = \mathrm{cl}(\mathrm{co}(\mathcal{G}))$.*

Utilizando el teorema anterior en el lema de Maurey, [Barron 1993, th. 1] obtiene sobre la familia $\mathcal{T} = \mathcal{T}_C(\mathbb{B}_{d,\tau})$, mediante redes MLP, la siguiente tasa de aproximación:

$$\|f_h - \mathtt{f}\|_{\mathcal{K}}^2 \leq \frac{(2\tau C_{\mathtt{f}})^2}{h}$$

donde $C_{\mathtt{f}} = \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\| \left|\widetilde{\mathtt{f}}(\boldsymbol{\omega})\right| d\boldsymbol{\omega}$. Además, si $f_h(\mathbf{x}) = \sum_{i=1}^h c_i \psi(\mathbf{w}_i^T\mathbf{x} + w_{i0}) + c_0$, los coeficientes de esta expresión pueden restringirse mediante ([Barron 1993, th. 1], [Barron 1994, th. 1]) $\sum_{i=1}^h |c_i| \leq 2\tau C_{\mathtt{f}} = C$ y $|c_0 - \mathtt{f}(\mathbf{0})| \leq C$ y $|w_{i0}| \leq \|\mathbf{w}_i\|_1$ manteniéndose la tasa de convergencia anterior.

La cota anterior no depende aparentemente de la dimensionalidad d, lo cuál originó inicialmente una importante controversia. Sin embargo, [Barron 1993, p. 932] indica que la dimensionalidad aparece indirectamente en la constante $C_{\mathtt{f}}$ al ser una integral d-dimensional, con lo que puede crecer con la dimensionalidad. Por ejemplo, si $\mathtt{f}(\mathbf{x}) = e^{-\|\mathbf{x}\|^2/2}$ entonces, $C_{\mathtt{f}} \leq d^{1/2}$ (un orden similar se tiene para las funciones definidas positivas).

Un resultado en la misma línea que el anterior, pero incluyendo la aproximación simultánea a las derivadas de la función de interés, es el de [Hornik *et al.* 1994, th. 2.1].

Para ello, se considera el espacio $\bar{\mathcal{C}}^r(\mathbb{R}^d, \mu) \equiv \mathcal{W}_\infty^r(\mathbb{R}^d, \mu)$, complección (espacio de Hilbert) del espacio $\mathcal{C}^r(\mathbb{R}^d, \mu)$ de las funciones reales definidas en $\mathbb{R}^d$ con derivadas $D^{\boldsymbol{\alpha}} f(\mathbf{x})$ continuas y acotadas uniformemente de orden $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i \leq r$ pero donde se considera la norma $\|f\|_{\mathcal{W}_2^r(\mathcal{X})}$.

Para dicho resultado, sea $\psi \in \bar{\mathcal{C}}^r(\mathbb{R}, \mu)$ y sea $\mathcal{M}^r(\mathbb{R}^d)$ el espacio de las medidas ν con signo en $\mathbb{R}^d \times \mathbb{R}$ tal que:

$$\|\nu\|_{\mathcal{M}^r(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \tau(\mathbf{w})^r d\,|\nu|\,(\mathbf{w}, w_0) < \infty$$

donde $|\nu|$ es la medida variación total[147] de ν y $\tau(\mathbf{w}) = \max\{\|\mathbf{w}\|_2, 1\}$. Denotamos:

$$\mathcal{T}_\psi = \{f = \psi * \nu = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \psi(\mathbf{w}^T \cdot + w_0) d\nu(\mathbf{w}, w_0) : \nu \in \mathcal{M}^r(\mathbb{R}^d)\} \subset \bar{\mathcal{C}}^r(\mathbb{R}^d, \mu)\}$$

$$\mathcal{F}_{h,B} = \{f_{h,B}(\mathbf{x}) = \sum_{i=1}^h \beta_i \tau(\mathbf{w}_i)^{-r} \psi(\mathbf{w}_i^T\mathbf{x} - w_{i0}) : \mathbf{w}_i \in \mathbb{R}^d, \beta_i, w_{i0} \in \mathbb{R}, \sum_{i=1}^h |\beta_i| \leq B\} \quad (3.14)$$

Las funciones de $\mathcal{T}_\psi$ pueden verse como funciones implementadas por una red neuronal con un continuo de unidades ocultas todas con función de activación ψ, donde la medida ν determina la distribución de los pesos de las salidas ocultas en la capa de salida.

El siguiente teorema describe el orden de aproximación a estas redes *continuas* (y a sus derivadas) mediante redes *discretas* (y sus derivadas) de las familias $\mathcal{F}_{h,B}$.

[147] [Hoffmann-Jorgensen 1994, p. 23]: **medida interior** es la medida λ_* definida $\forall B \in \mathcal{B}$ como $\lambda_*(B) = \sup\{\sum_{j=1}^t B_j : B_j \subset \mathcal{B},\ B_j \cap B_i = \varnothing,\ i \neq j,\ \cup_{j=1}^t B_j \subseteq B\}$. [Hoffmann-Jorgensen 1994, p. 79] [Rudin 1988, p. 133]: dada una medida ν en un espacio $(\mathcal{X}, \mathcal{B})$ con $\mathcal{B}$ σ-álgebra en $\mathcal{X}$, se denomina **variación total** $|\nu|$ de ν a la medida interior correspondiente a la función $\nu(B) = |\nu(B)|$, $B \in \mathcal{B}$ y se denota $\|\nu\| = \lambda_*(\mathcal{X}) = \int_{\mathcal{X}} d\,|\nu|\,\mathbf{x}$.

Teorema 179 *[Hornik* et al. *1994, th. 2.1] Sea $\psi \in \bar{\mathcal{C}}^r(\mathbb{R},\mu)$ y $\nu \in \mathcal{M}^r(\mathbb{R}^d)$. Si $B \geq \|\nu\|_{\mathcal{M}^r(\mathbb{R}^d)}$, entonces $\forall \mathsf{f} \in \mathcal{T}_\psi$:*

$$\delta(\mathsf{f}, \mathcal{F}_{h,B}) < \frac{C_{r,d}\,\|\psi\|^2_{\bar{\mathcal{C}}^r(\mathbb{R},\mu)}\,\|\nu\|_{\mathcal{M}^r(\mathbb{R}^d)}}{\sqrt{n}}$$

El teorema anterior se demuestra utilizando los siguientes dos lemas que incluimos por su interés:

Lema 180 *[Hornik* et al. *1994, lema A.1] Sea $\mathcal{K}$ un espacio de Banach separable y $\mathcal{G} \subset \mathcal{K}$ acotado no vacío verificando que $\mathcal{G} = -\mathcal{G}$. Sea ν una medida finita con signo en $\mathcal{G}$ y sea $\mathsf{f} = \int f\nu(df)$ la integral de ν en sentido débil*[148]*. Entonces, $\forall B \geq |\nu|(\mathcal{G})$, se verifica que $\mathsf{f} \in \mathcal{T}_\mathcal{G} = \mathrm{cl}(\mathrm{co}(\mathcal{G}))$.*

La utilización del teorema 176 y del lema anterior con la familia:

$$\mathcal{G} = \{\psi_{\mathbf{w},w_0} = \beta\tau(\mathbf{w})^{-r}\psi(\mathbf{w}^T \cdot + w_0) : \mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}, |\beta| \leq B\} \tag{C.15}$$

produce el resultado del teorema 179.

Entre otros resultados de la aplicación del teorema anterior, (véase [Hornik *et al.* 1994]) destacamos el siguiente donde se denota:

$$\mathcal{T}_r = \{\mathsf{f} = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T\mathbf{x}}\nu_\mathsf{f}(d\boldsymbol{\omega})\}$$

donde ν_f es una medida compleja en $\mathbb{R}^d$ satisfaciendo: $\|\nu_\mathsf{f}\|_{(r)} = \int_{\mathbb{R}^d}\tau(\boldsymbol{\omega})^r\,|\nu_\mathsf{f}|\,(d\boldsymbol{\omega}) < \infty$. Casos particulares de funciones de $\mathcal{T}_r$ son las funciones f tal que ν_f es absolutamente continua respecto a la medida de Lebesgue y su densidad es la transformada de Fourier $\tilde{\mathsf{f}}$.

Teorema 181 *[Hornik* et al. *1994, cor. 2.6] Sea μ medida de probabilidad con soporte compacto y sea $\psi \in \bar{\mathcal{C}}^r(\mathbb{R},\mu)$ tal que*[149] *$\||t|^\gamma D^\alpha\psi(t)\|_{L_1(\mathbb{R})} < \infty$ para algún $\gamma > 0$ para todo $0 \leq \alpha \leq r$. Entonces, $\forall \mathsf{f} \in \mathcal{T}_r$, existen constantes B_0, C que dependen sólo de r, d, μ y ψ tal que, si $B \geq B_0\,\|\nu_\mathsf{f}\|_{(r+1)}\,h^{\frac{1}{2}(\gamma+1)}$, se verifica que:*

$$\delta(\mathsf{f}, \mathcal{F}_{h,B}) \leq C\,\|\nu_\mathsf{f}\|_{(r+1)}\,h^{-\frac{1}{2}\frac{\gamma}{\gamma+1}}$$

Nótese que, cuanto mayor se elija γ (mayor suavidad en ψ) más se aproxima la tasa de convergencia en $\mathcal{T}_r$ a la proporcionada por el teorema 176.

[Makovoz 1996] consigue mejorar las cotas que se obtienen a través del lema de Maurey y del teorema 176 incorporando en la cota el valor:

$$\varepsilon_h(\mathcal{G}) = \inf\{\varepsilon > 0 : \mathcal{G} \text{ puede ser recubierto con } h \text{ conjuntos de diámetro } \leq \varepsilon\}$$

tal como se muestra en el siguiente teorema:

Teorema 182 *[Makovoz 1996] [Pinkus 1999, th. 6.12] Sea $\mathcal{G} = \{\psi_1, \psi_2, ...\} \subset \mathcal{K}$ en las condiciones anteriores. Sea $\mathsf{f} \in \mathcal{T}_\mathcal{G} = \mathrm{cl}(\mathrm{co}(\mathcal{G}))$. Entonces, existe $f_h = \sum_{i=1}^h c_i\psi_i$ con $\psi_i \in \mathcal{G}$, $c_i \geq 0$,*

[148] [Hornik *et al.* 1994, lema A.1]: en las condiciones anteriores para $\mathcal{K}$ y μ, se define la **integral de μ en sentido débil** y se denota por $\int f\mu(df)$, como el elemento $\mathsf{f} \in \mathcal{K}$, si existe, tal que, $\forall g^* \in \mathcal{K}^*$ dual de $\mathcal{K}$, $\int g^*(f)\mu(df) = g^*(f)$.

[149] Esta condición determina una velocidad de decaimiento a cero de las colas de las derivadas de ψ.

$i = 1 : h$ *con* $\sum_{i=1}^{h} c_i \leq 1$, *tal que:*

$$\|f_h - \mathsf{f}\|_{\mathcal{K}} \leq \frac{2\varepsilon_h(\mathcal{G})}{\sqrt{h}}$$

Nótese que si $\mathcal{G}$ posee un cierto grado de compacidad, entonces $\varepsilon_h(\mathcal{G}) \to 0$ cuando $h \to \infty$, lo que supone una mejora de las cotas del teorema 176.

Inspirándose en [Barron 1993], en el espacio de Hilbert $\mathcal{K} = L_2(\mathbb{B}_{d,\tau})$ [Pinkus 1999, p. 180] considera el conjunto $\mathcal{G} = \{\pm\psi(\mathbf{w}^T\mathbf{x} + w_0) : \|\mathbf{w}\|_2 = 1 \text{ y } |w_0| \leq 1\}$ donde ψ es una función sigmoide de grado 0 (pág. 376) y obtiene que que $\varepsilon_h(\mathcal{G}) \leq Ch^{-1/2d}$. Con ello, aplica el teorema anterior al espacio $\mathcal{T} = \mathcal{T}_C(\mathbb{B}_{d,\tau}) \subset \mathcal{T}_{\mathcal{G}} = \mathrm{cl}(\mathrm{co}(\mathcal{G}))$, de lo que resulta la cota $Ch^{-(d+1)/2d}$.

[Chen y White 1999] aplican la mejora del teorema 182 a los resultados de [Hornik *et al.* 1994] obteniendo el siguiente resultado:

Teorema 183 *[Chen y White 1999, th. 2.1] Sea* $\mathsf{f} \in \mathcal{T}_{r+1}$. *Sean* μ *y* $\psi \in \bar{\mathcal{C}}^r(\mathbb{R}, \mu)$ *ambas de soporte compacto y* ψ *satisfaciendo:*

$$\left\|\psi_{\mathbf{w},w_0} - \psi_{\mathbf{w}',w_0'}\right\| \leq C[(\left\|\mathbf{w} - \mathbf{w}'\right\|_2 + \left|w_0 - w_0'\right|]^q$$

para algún $q \in (0, 1]$, *donde* $\psi_{\mathbf{w},w_0}$ *es como en (C.15). Entonces, si* $C \geq B_0 \|\nu_{\mathsf{f}}\|_{(r)}$ *para alguna constante* B_0, *se verifica que:*

$$\delta(\mathsf{f}, \mathcal{F}_{h,B}) \leq C \|\nu_{\mathsf{f}}\|_{(r+1)} h^{-\frac{1}{2}-\frac{q}{d^*}}$$

donde $d^* = d$ *si* ψ *es homogénea*[150]*, y* $d^* = d + 1$ *en otro caso.*

Órdenes *Greedy* de Aproximación Mediante Redes RBF

El enfoque basado en el lema de Maurey seguido por [Hornik *et al.* 1994] en el teorema 179 fué aplicado previamente por [Girosi y Anzellotti 1993] a las redes RBF en el siguiente teorema en el que, para $\psi \in L_2(\mathbb{R}^d)$ con $\|\psi\|_{L_2} = 1$, denotamos:

$$\begin{aligned}
\mathcal{T}_\psi &= \{f = (\psi * \mu)(\mathbf{x}) = \int_{\mathbb{R}^d} \psi(\mathbf{x} - \mathbf{w})\mu(d\mathbf{w}) : \mu \in \mathcal{M}(\mathbb{R}^d)\} \subset L_2(\mathbb{R}^d) \\
\mathcal{G} &= \{c\psi(\mathbf{x} - \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d, |c| \leq \|\mu\|\} \\
\mathcal{F}_h &= \{f_h(\mathbf{x}) = \sum_{i=1}^{h} c_i\psi(\mathbf{x} - \mathbf{w}_i) : \mathbf{c} \in \mathbb{R}^h, \mathbf{w}_i \in \mathbb{R}^d, i = 1 : h\}
\end{aligned}$$

donde $\mathcal{M}(\mathbb{R}^d) \supset L_1(\mathbb{R}^d)$ es el conjunto de las medidas μ de Radon[151] con signo cuya variación total $|\mu|$ es acotada con $\|\mu\| < \infty$.

Teorema 184 *[Girosi y Anzellotti 1993, th. 8.2] Sea* $\mathsf{f} \in \mathcal{T}_\psi$ *con* $\psi \in L_2(\mathbb{R}^d)$ *con* $\|\psi\|_{L_2} = 1$ *y* μ *medida de Radon con signo de variación total acotada* $\|\mu\| < \infty$. *Sea* $\mathcal{G}$ *la familia definida anteriormente. Entonces:* $\mathsf{f} \in \mathcal{T}_{\mathcal{G}} = \mathrm{cl}(\mathrm{co}(\mathcal{G}))$ *y además,* $\forall C_f > \|\mu\|^2 - \|\mathsf{f}\|_{L_2}^2$, $\exists c_i \in \mathbb{R}, \mathbf{w}_i \in \mathbb{R}^d$, $i = 1 : h$, *tal que:*

$$\|\mathsf{f} - f_h\|_{L_2}^2 \leq C_f h^{-1} \tag{C.16}$$

[150] Es decir, si cumple $\psi(bx) = b\psi(x)\ \forall x \in \mathbb{R}$.

[151] [Hoffmann-Jorgensen 1994, p. 39]: si $\mathcal{X}$ es un espacio métrico y μ una medida de Borel finita en $\mathcal{X}$, se dice que μ es medida de Radon si $\forall B \in \mathcal{B}(\mathcal{X})$ σ-álgebra de Borel, se verifica $\mu(B) = \sup\{\mu(K) : K \subseteq B \text{ compacto}\}$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Un teorema exactamente igual que el anterior se verifica en el espacio de Sobolev $\mathcal{W}_2^r(\mathbb{R}^d)$ con su norma tradicional:

Teorema 185 *[Girosi y Anzellotti 1993, th. 8.3], Sea* $\mathsf{f} \in \mathcal{T}_\psi$ *con* $\psi \in \mathcal{W}_2^r(\mathbb{R}^d)$ *con* $\|\psi\|_{\mathcal{W}_2^r} = 1$ *y* μ *medida de Radon con signo de variación total acotada* $\|\mu\| < \infty$. *Entonces,* $\mathsf{f} \in \mathcal{T}_\mathcal{G} = \mathrm{cl}(\mathrm{co}(\mathcal{G}))$ *con la norma de* $\mathcal{W}_2^r(\mathbb{R}^d)$ *y existe* $f_h \in \mathcal{F}_h$ *tal que* $\|\mathsf{f} - f_h\|^2_{\mathcal{W}_2^r(\mathbb{R}^d)} \leq C_f h^{-1}$ *para todo* $C_f > \|\mu\|^2 - \|\mathsf{f}\|^2_{\mathcal{W}_2^r(\mathbb{R}^d)}$.

Utilizando el resultado (C.3), se tiene que si $r > d/2$, entonces $\mathcal{W}_2^r(\mathbb{R}^d) \subset \mathcal{C}(\mathbb{R}^d)$ y existe C_1 tal que $\|\cdot\|_\infty \leq C_1 \|\cdot\|_{\mathcal{W}_2^r(\mathbb{R}^d)}$. Con ello, del teorema anterior se obtiene el siguiente resultado:

Corolario 186 *[Girosi y Anzellotti 1993, cor. 8.4] Bajo las condiciones del teorema anterior, si* $r > d/2$, *existe* $f_h \in \mathcal{F}_h$ *tal que:* $\|\mathsf{f} - f_h\|^2_{L_\infty(\mathbb{R}^d)} \leq C_1 C_f h^{-1}$.

Las cotas anteriores pueden expresarse en términos de la dimensión de la variedad no lineal $\mathcal{F}_h$ que es $a = h(d+1)$. Por ejemplo, la cota (C.16) resulta:

$$\delta(\mathsf{f}, \mathcal{F}_a)_{L_2} \leq C_f \sqrt{\frac{d+1}{a}}$$

lo que muestra que la familia $\mathcal{T}_\psi$ no incrementa su complejidad dramáticamente con la dimensión.

[Girosi y Anzellotti 1993] caracteriza este tipo de familias $\mathcal{T}_\psi$ para el núcleo gausiano y para los núcleos conocidos como de Bessel-MacDonald, de la siguiente forma:

1. Si $\psi(\mathbf{x}) = e^{-\|\mathbf{x}\|^2} \in \mathcal{W}_2^r(\mathbb{R}^d)\ \forall r > 0$, puede aplicarse el corolario anterior. En este caso, el espacio $\mathcal{T}_\psi$ es más bien reducido debido a la suavidad de la gausiana y a su rápido decaimento a cero. Este espacio no posee una caracterización precisa [Niyogi y Girosi 1999], sin embargo, contiene al espacio $\tilde{\mathcal{C}}_c^r(\mathbb{R}^d) = \{f : \tilde{f} \in \mathcal{C}_c^r(\mathbb{R}^d)\}$ si $r > d/2$, donde $\mathcal{C}_c^r(\mathbb{R}^d)$ es el espacio de funciones con soporte compacto y r derivadas continuas. De hecho, si $f \in \tilde{\mathcal{C}}_c^r(\mathbb{R}^d)$, se tiene con α constante [Girosi y Anzellotti 1993]:

$$\frac{\tilde{f}(\boldsymbol{\omega})}{\tilde{\psi}(\boldsymbol{\omega})} = \alpha \frac{\tilde{f}(\boldsymbol{\omega})}{e^{-\|\boldsymbol{\omega}\|^2}} = \alpha e^{\|\boldsymbol{\omega}\|^2} \tilde{f}(\boldsymbol{\omega}) \equiv \tilde{\mu} \in \mathcal{C}_c^r(\mathbb{R}^d) \subset L_1(\mathbb{R}^d)$$

   con lo que $\tilde{f} = \tilde{\psi}\tilde{\mu}$, es decir, $f = \psi * \mu$ donde μ es la transformada de Fourier de $\tilde{\mu} = \tilde{f}/\tilde{\psi}$.

2. Los núcleos de Bessel-MacDonalds pueden definirse a través de sus transformadas de Fourier:

$$\tilde{\psi}_s(\boldsymbol{\omega}) = \frac{1}{(1 + 4\pi^2 \|\boldsymbol{\omega}\|_2^2)^s}, s > 0$$

   que son positivas y, debido al teorema de Bochner (Anexo B), sus transformadas inversas son definidas positivas (y radiales). Estos núcleos decaen exponencialmente en el infinito y pueden sufrir discontinuidades en el origen. Sin embargo, si $s > d$ son diferenciables de cualquier orden $q < s - d$. Para estos casos, [Girosi y Anzellotti 1993] muestra que:

$$\mathcal{W}_1^s(\mathbb{R}^d) \subset \mathcal{T}_{\psi_s}^1 \subset \mathcal{T}_{\psi_s}$$

   donde $\mathcal{T}_{\psi_s}^1$ es el espacio de funciones $f = \psi_s * \mu$ con $\mu \in L_1(\mathbb{R}^d)$.

Estos núcleos son interesantes pues dan lugar a un ejemplo relevante de espacio de Hilbert $\mathcal{H}_{\psi_s}$ con Núcleo Reproductor ψ_s, al verificarse que [Schaback 1993, th. 1]:

$$\mathcal{H}_{\psi_s} = \mathcal{W}_2^s(\mathbb{R}^d)$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Tasa de Convergencia de las Redes Neuronales $\textbf{MLP } \hat{f}(\mathbf{x}) = \sum_{i=1}^{h} c_i \sigma(\mathbf{w}_i^T \mathbf{x} + w_{i0})$ **y RBF** $\hat{f}(\mathbf{x}) = \sum_{i=1}^{h} c_i \psi_i(\mathbf{x})$		
Espacio Funcional	**Condiciones**	**Tasa**
$\mathcal{V}_p^r(\mathbb{B}_d)$ [Pinkus 1999, cor. 6.11]	σ sigmoide de grado s $r \in \{1, ..., s+1+\frac{d-1}{2}\}$	$\left\|\mathsf{f} - \hat{f}\right\|_{L_2(\mathbb{B}_{d,\tau})} \leq Ch^{-r/d}$
$\mathcal{T}_C(\mathbb{B}_{d,\tau}),\ \mu$ med. prob. [Barron 1993]	σ sigmoide medible, acotada	$\left\|\mathsf{f} - \hat{f}\right\|^2_{L_2(\mathbb{B}_{d,\tau},\mu)} \leq (2\tau C_\mathsf{f})^2 h^{-1}$
$\mathcal{T}_C(\mathbb{B}_{d,\tau}),\ \mu$ med. prob. [Makovoz 1996]	σ sigmoide medible, acotada	$\left\|\mathsf{f} - \hat{f}\right\|^2_{L_2(\mathbb{B}_{d,\tau},\mu)} \leq Ch^{-(d+1)/2d}$
$\mathcal{T}_\psi = \{\mathsf{f} = \psi * \mu\}$ $\{\mathsf{f} : \tilde{\mathsf{f}} \in \mathcal{C}_c^r(\mathbb{R}^d)\} \subset \mathcal{T}_\psi$ [Girosi y Anzellotti 1993]	μ medida Radon, $\|\mu\| \leq \infty$ $\psi_i(\mathbf{x}) = e^{-\|\mathbf{x}-\mathbf{w}_i\|^2}$ $2r > d$	$\left\|\mathsf{f} - \hat{f}\right\|^2_{L_2} \leq C_\mathsf{f} h^{-1}$
$\mathcal{W}_1^{2r}(\mathbb{R}^d)$ [Girosi 1994]	$\psi_i(\mathbf{x}) = e^{-\|\mathbf{x}-\mathbf{w}_i\|^2/\sigma_i^2}$ $2r > d$	$\left\|\mathsf{f} - \hat{f}\right\|^2_{L_2} \leq C_\mathsf{f} h^{-1}$
$\mathcal{W}_p^r([-1,1]^d), p \in [1,\infty]$ [Mhaskar 1996]	$\psi_i(\mathbf{x}) \equiv \psi(\mathbf{W}_i \mathbf{x} + \mathbf{w}_{0,i})$ $\psi \in \mathcal{C}^\infty(B),\ B$ abierto $D^{\boldsymbol{\alpha}}\psi(\mathbf{x}_0) \neq 0$	$\|\mathsf{f} - f_h\|_{L_p} \leq Ch^{-r/d} \|\mathsf{f}\|_{\mathcal{W}_p^r}$

Tabla 43. Algunas tasas de convergencia con redes neuronales. Todos los resultados se basan en el algoritmo greedy salvo el último de Mhaskar. Se supone que sigma es una funcion no constante

La forma de estos núcleos cambia con la dimensión d y no poseen una expresión analítica [Girosi y Anzellotti 1993, p. 110], lo que complica su utilización en la práctica, sin embargo, poseen una representación integral en términos de gausianas con diferentes parámetros de escala, lo que permite obtener [Girosi 1994, th. 2.1] con estas funciones básicas, las mismas posibilidades aproximadoras que con los núcleos de Bessel-MacDonalds.

La tabla 43, resume los resultados sobre tasas de convergencia presentados hasta ahora. De ellos, todos se obtienen a través de algoritmos *greedy* salvo el último de [Mhaskar 1996], presentado al principio de este apartado.

Ordenes de Aproximación con Diccionarios

Con respecto al segundo tipo de enfoque mencionado (pág. 362) para tratar el problema de la velocidad de convergencia, (recopilar bajo una misma familia todas las funciones que poseen un mismo orden de aproximación con un esquema de aproximación determinado), no conocemos ningún resultado relativo a redes neuronales salvo los obtenidos por [DeVore y Temlyakov 1996] para diccionarios utilizando algoritmos *greedy* en el contexto de un espacio de Hilbert $\mathcal{K}$, y que presentamos en este apartado.

Diccionarios

Análisis y Síntesis

El proceso de señales tiene por objeto representar una señal en términos de sus características más relevantes. El proceso de **análisis** consiste en la descomposición de la señal para identificar dichas características, mientras que el proceso de **síntesis** va dirigido a la reconstrucción de la señal a partir de dichos elementos de información.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Si $\mathcal{G} = \{\psi_j\}_{j \in J}$ es un conjunto completo de elementos en el espacio en el que se supone reside la señal f, y $\mathbf{f} \in \mathbb{R}^n$ es el vector de datos que se observa sobre la misma, el proceso de análisis persigue la determinación de un vector $\boldsymbol{\alpha}$ tal que:

$$\mathbf{f} = \Psi \boldsymbol{\alpha}$$

donde $\boldsymbol{\Psi}$ es una matriz cuyas columnas son los elementos ψ_j valorados en los datos. Si $\mathcal{G}$ es una base, la solución es única: $\boldsymbol{\alpha} = \boldsymbol{\Psi}^{-1}\mathbf{f}$ y si es una base ortogonal, $\boldsymbol{\Psi}^{-1} = \boldsymbol{\Psi}^T$. Sin embargo, si $\mathcal{G}$ es redundante, puede haber diferentes soluciones $\hat{\boldsymbol{\alpha}}_i$ al problema, proporcionando diferentes representaciones de $\mathbf{f}$.

Un objetivo primordial en el proceso de señales, es la parsimonia de la representación debido, entre otras razones, a un menor error de aproximación, una mayor interpretabilidad y a razones computacionales y de almacenamiento.

Diccionarios

El problema de las descomposiciones en términos de una base (ortogonal), como por ejemplo una base de Fourier o una base *wavelet* (p. ej. [Härdle *et al.* 1998], [Vidakovic 1999]), es que utilizan un conjunto de elementos lo más reducido posible para expresar las características intrínsecas de la señal. Esta simplicidad posee beneficios computacionales pero, en contrapartida, la calidad de la descomposición depende excesivamente de la adecuación de los elementos de la base a las características de la señal. Por ejemplo, una base de Fourier no produce buenas representaciones de señales muy localizadas en el tiempo y las bases *wavelet* no son adecuadas para representar señales cuyas transformadas de Fourier poseen un soporte de altas frecuencias muy concentrado.

En estas situaciones, es difícil obtener información sobre la señal a partir de los coeficientes de la correspondiente expansión, pues la información tiende a diluirse entre todos los elementos de la base utilizada. Parafraseando a [Mallat y Zhang 1992]: *Aunque este vocabulario (base) pueda ser suficiente para expresar todas las ideas (características de la señal), obliga a utilizar largas frases para sustituir las palabras no disponibles en el diccionario.*

Por esta razón, tienden a preferirse **diccionarios** redundantes cuyos elementos poseen *significados* parecidos pero que aportan una mayor flexibilidad a la hora de elegir la descomposición más apropiada de una determinada señal. Las estructuras *wavelet* (*frames*) (p. ej. [Kaiser 1994]) son un paso en esta persecución de una mayor flexibilidad.

Pero, además, es necesario considerar diccionarios cuyos elementos (**átomos**) posean la máxima adaptabilidad. Cuando la señal posee características irregularmente localizadas en el tiempo y en la frecuencia, es necesario utilizar descomposiciones adaptables a estas características. Por ejemplo, señales tipo impulso requieren descomposiciones con elementos concentrados en el dominio del tiempo, mientras que las líneas espectrales se representan mejor mediante elementos con un soporte concentrado en el dominio de las frecuencias. Si la señal posee ambos tipos de características, son necesarios **átomos** de ambos tipos.

Por esta razón, se consideran los **diccionarios de átomos de tiempo-frecuencia** que generalizan los diccionarios indexados en frecuencia (Fourier) o en tiempo y escala (*wavelet*). El progreso en la dotación de flexibilidad se ve en la tabla 44:

Una vez construido el diccionario $\mathcal{G}$, es necesario un procedimiento para seleccionar los átomos más adecuados para descomponer una señal dada. Entre los procedimientos más relevantes, (una revisión general puede verse p. ej. en [Chen *et al.* 1999]), destacamos los siguientes:

1. El *método de las estructuras* (*method of frames*, MOF), [Daubechies 1988].

Diccionario	Transformada	Átomos
Fourier	$f(\omega) = \int f(x) e^{-i\omega x} dx$	$\psi_\omega(x) = e^{-i\omega x}$
Fourier Local	$f(\omega, w) = \int f(x)\psi(x-w) e^{-i\omega x} dx$	$\psi_{\omega,w}(x) = \psi(x-w) e^{-i\omega x}$
Wavelet	$f(w, v) = \lvert v\rvert^{-1/2} \int f(x)\psi(\frac{x-w}{v}) dw dv$	$\psi_{w,v}(x) = \lvert v\rvert^{-1/2} \psi(\frac{x-w}{v})$
Tiempo-Frecuencia	$f(\omega, w, v) = \int f(x)\psi(\frac{x-w}{v}) e^{-i\omega x} dx$	$\psi_{\omega,w,v}(x) = \psi(\frac{x-w}{v}) e^{-i\omega x}$

Tabla 44. Diccionarios, transformadas y átomos correspondientes.

Algoritmo Matching Pursuit Ortogonal

1. Inicio: $\mathsf{f}_0 = 0$, $R_0\mathsf{f} = \mathsf{f}$
2. Para $s = 1, 2, ...$ hasta que $\lvert\langle R_{s-1}\mathsf{f}, \psi_s\rangle\rvert < \delta > 0$
   - 2.1 Determinar $\psi_s \in \mathcal{G}$ tal que:
     $$\lvert\langle R_{s-1}\mathsf{f}, \psi_s\rangle\rvert \geq \alpha \sup_{\boldsymbol{\gamma}\in\Gamma} \lvert\langle R_{s-1}\mathsf{f}, \psi_{\boldsymbol{\gamma}}\rangle\rvert \text{ con } 0 < \alpha \leq 1$$
   - 2.2 Calcular $\bar{\psi}_s$, $\{\tau_j\}_{j=1}^{s-1}$, resolviendo:
     $$\psi_s = \sum_{j=1}^{s-1} \tau_j \psi_j + \bar{\psi}_s \text{ con } \langle \bar{\psi}_s, \psi_j \rangle = 0,\ j = 1, ..., s-1$$
   - 2.3 Calcular:
     $$\alpha_s^{(s)} = \frac{\langle R_{s-1}\mathsf{f}, \psi_s\rangle}{\lVert\bar{\psi}_s\rVert^2}$$
     $$\alpha_j^{(s)} = a_j^{(s-1)} - \frac{\langle R_s\mathsf{f}, \psi_s\rangle}{\lVert\bar{\psi}_s\rVert^2}\tau_j,\ j = 1, ..., s-1$$
   - 2.4 Actualizar:
     $$f_s = \sum_{j=1}^{s} \alpha_j^{(s)} \psi_j$$
     $$R_s\mathsf{f} = \mathsf{f} - f_s$$

Tabla 45. Algoritmo matching pursuit ortogonal.

2. *Best Orthogonal Basis*, [Coifman y Vickerhauser 1992].

3. *Matching pursuit* (MP) [Mallat y Zhang 1992], y su versión ortogonal *(orthogonal matching pursuit*, OMP), ([Pati y Krishnaprasad 1990], [Davis *et al.* 1994]).

4. *Basis Pursuit* (BP) y su versión en ambiente de ruido, *Basis Pursuit Denoising*, (BPDN), [Chen *et al.* 1999].

A continuación se presenta el tercero de estos métodos en su versión ortogonal (OMP).

Algoritmo *Greedy* con *Back-fitting*. *Matching Pursuit* Ortogonal

El algoritmo *greedy* puro con *back-fitting*, coincide con el que, en el ámbito del proceso de señales se denomina **Matching Pursuit Ortogonal** (*Orthogonal Matching Pursuit* –OMP), [Pati *et al.* 1993] y [Davis *et al.* 1994]. Este algoritmo, se aplica a un diccionario redundante de funciones básicas (**átomos**) [152] $\mathcal{G} = \{\psi_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \Gamma\}$ en un espacio de Hilbert $\mathcal{K}$, donde $\Gamma \in \mathbb{R}^p$ es un espacio de parámetros.

Dicho algoritmo (tabla 45), sigue exactamente la misma filosofía que el proceso de ortogonalización de Gram-Schmidt (tabla 9, pág. 175) sólo que en un espacio de funciones.

[152] Generalmente se supone que los átomos son de norma unidad, pero para mayor generalidad, evitamos esta restricción en lo que sigue.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En la primer iteración, se elige la función ψ_1 tal que:

$$|\langle \mathsf{f}, \psi_1 \rangle| \geq \alpha \sup_{\boldsymbol{\gamma} \in \Gamma} \left| \langle \mathsf{f}, \psi_{\boldsymbol{\gamma}} \rangle \right| \text{ con } 0 < \alpha \leq 1$$

y se descompone el primer residuo $R_0\mathsf{f} = \mathsf{f}$ en términos de $\bar{\psi}_1 = \psi_1$:

$$\mathsf{f} = R_0\mathsf{f} = P_{\mathcal{H}_1}\mathsf{f} + R_1\mathsf{f} = \frac{\langle R_0\mathsf{f}, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\bar{\psi}_1 + R_1\mathsf{f} \text{ con } \langle R_1\mathsf{f}, \psi_1 \rangle = 0$$

donde $\mathcal{H}_1 = \langle \{\psi_1\} \rangle = \langle \{\bar{\psi}_1\} \rangle$.

En la segunda iteración, ha de conseguirse $\langle R_2\mathsf{f}, \psi_2 \rangle = 0$ pero también $\langle R_2\mathsf{f}, \psi_1 \rangle = 0$. Se elige ψ_2 según el mismo criterio:

$$|\langle R_1\mathsf{f}, \psi_2 \rangle| \geq \alpha \sup_{\boldsymbol{\gamma} \in \Gamma} \left| \langle R_1\mathsf{f}, \psi_{\boldsymbol{\gamma}} \rangle \right| \text{ con } 0 < \alpha \leq 1$$

y se determina su componente $\bar{\psi}_2$ no redundante con $\bar{\psi}_1$, que resulta de proyectar ψ_2 sobre $\mathcal{H}_1$:

$$\psi_2 = P_{\mathcal{H}_1} + \bar{\psi}_2 = \frac{\langle \psi_2, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\bar{\psi}_1 + \bar{\psi}_2 \Rightarrow \bar{\psi}_2 = \psi_2 - \frac{\langle \psi_2, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\bar{\psi}_1$$

Con ello, $\langle \bar{\psi}_2, \bar{\psi}_1 \rangle = 0$ y $\{\bar{\psi}_1, \bar{\psi}_2\}$ es base ortogonal de $\mathcal{H}_2 =< \{\psi_1, \psi_2\} >$. Ahora, se ajusta el residuo anterior:

$$R_1\mathsf{f} = \frac{\langle R_1\mathsf{f}, \bar{\psi}_2 \rangle}{\|\bar{\psi}_2\|^2}\bar{\psi}_2 + R_2\mathsf{f} = \frac{\left\langle \mathsf{f} - \frac{\langle \mathsf{f}, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\bar{\psi}_1, \bar{\psi}_2 \right\rangle}{\|\bar{\psi}_2\|^2}\bar{\psi}_2 + R_2\mathsf{f} = \frac{\langle \mathsf{f}, \bar{\psi}_2 \rangle}{\|\bar{\psi}_2\|^2}\bar{\psi}_2 + R_2\mathsf{f}$$

produciendo la descomposición:

$$\mathsf{f} = \frac{\langle R_0\mathsf{f}, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\bar{\psi}_1 + R_1\mathsf{f} = \frac{\langle \mathsf{f}, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\bar{\psi}_1 + \frac{\langle \mathsf{f}, \bar{\psi}_2 \rangle}{\|\bar{\psi}_2\|^2}\bar{\psi}_2 + R_2\mathsf{f}$$

con lo que $R_2\mathsf{f} \perp \mathcal{H}_2 = \langle \{\bar{\psi}_1, \bar{\psi}_2\} \rangle =< \langle \{\psi_1, \psi_2\} \rangle$. En términos de los átomos originales, ψ_1, ψ_2, esta descomposición resulta:

$$\begin{aligned}
\mathsf{f} &= \frac{\langle R_0\mathsf{f}, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\bar{\psi}_1 + \frac{\langle R_1\mathsf{f}, \bar{\psi}_2 \rangle}{\|\bar{\psi}_2\|^2}\bar{\psi}_2 + R_2\mathsf{f} \\
&= \frac{\langle R_0\mathsf{f}, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\psi_1 + \frac{\left\langle R_1\mathsf{f}, \psi_2 - \frac{\langle \psi_2, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\psi_1 \right\rangle}{\|\bar{\psi}_2\|^2}(\psi_2 - \frac{\langle \psi_2, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2}\psi_1) + R_2\mathsf{f} \\
&= (\frac{\langle R_0\mathsf{f}, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2} - \frac{\langle R_1\mathsf{f}, \psi_2 \rangle}{\|\bar{\psi}_2\|^2}\frac{\langle \psi_2, \bar{\psi}_1 \rangle}{\|\bar{\psi}_1\|^2})\psi_1 + \frac{\langle R_1\mathsf{f}, \psi_2 \rangle}{\|\bar{\psi}_2\|^2}\psi_2 + R_2\mathsf{f} \\
&= c_1^{(2)}\psi_1 + c_2^{(2)}\psi_2 + R_2\mathsf{f}, \text{ siendo:} \\
&\quad \begin{cases} c_1^{(2)} = c_1^{(1)} - \dfrac{\langle R_1\mathsf{f}, \psi_2 \rangle}{\|\bar{\psi}_2\|^2}\dfrac{\langle \bar{\psi}_1, \psi_2 \rangle}{\|\bar{\psi}_1\|^2} \\ c_2^{(2)} = \dfrac{\langle R_1\mathsf{f}, \psi_2 \rangle}{\|\bar{\psi}_2\|^2} \end{cases}
\end{aligned}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde hemos utilizado que $R_1\mathtt{f} \perp \psi_1$.

En la iteración s-ésima, se elige el vector ψ_s según el criterio:

$$|\langle R_{s-1}\mathtt{f}, \psi_s\rangle| \geq \alpha \sup_{\boldsymbol{\gamma}\in\Gamma} \left|\left\langle R_{s-1}\mathtt{f}, \psi_{\boldsymbol{\gamma}}\right\rangle\right| \text{ con } 0 < \alpha \leq 1$$

se determina su componente $\bar{\psi}_s$ no redundante (ortogonal) con los vectores ya introducidos:

$$\bar{\psi}_s = \psi_s - \sum_{j=1}^{s-1} \frac{\langle \psi_s, \bar{\psi}_j\rangle}{\left\|\bar{\psi}_j\right\|^2}\bar{\psi}_j$$

y se realiza la descomposición:

$$R_{s-1}\mathtt{f} = \frac{\left\langle R_{s-1}\mathtt{f}, \bar{\psi}_s\right\rangle}{\left\|\bar{\psi}_s\right\|^2}\bar{\psi}_s + R_s\mathtt{f} \Rightarrow \mathtt{f} = \sum_{j=1}^{s} \frac{\left\langle R_{s-1}\mathtt{f}, \bar{\psi}_j\right\rangle}{\left\|\bar{\psi}_j\right\|^2}\bar{\psi}_j + R_s\mathtt{f}$$

donde, por construcción, los vectores $\bar{\psi}_j$, $j = 1, ..., s$ son todos ortogonales y $R_s\mathtt{f} \perp \mathcal{H}_s = \left\langle \{\bar{\psi}_j\}_{j=1}^s \right\rangle = \left\langle \{\psi_j\}_{j=1}^s \right\rangle$.

La expresión de la solución en términos de los átomos originales se obtiene mediante:

$$\mathtt{f} = \sum_{j=1}^{s} \alpha_j^{(s)}\psi_j + R_s\mathtt{f}, \text{ siendo:}$$

$$\begin{cases} c_j^{(s)} = c_j^{(s-1)} - \dfrac{\langle R_1\mathtt{f}, \psi_s\rangle}{\left\|\bar{\psi}_s\right\|^2}\dfrac{\langle \psi_s, \bar{\psi}_j\rangle}{\left\|\bar{\psi}_j\right\|^2}, \ j = 1, ..., s-1 \\ c_s^{(s)} = \dfrac{\langle R_{s-1}\mathtt{f}, \psi_s\rangle}{\left\|\bar{\psi}_s\right\|^2} \end{cases}$$

Sobre la convergencia del algoritmo OMP, se tiene el siguiente teorema:

Teorema 187 *[Pati* et al. *1993], [Davis* et al. *1994] Sea* $\mathtt{f} \in \mathcal{T}_{\mathcal{G}} = \mathrm{cl}(\mathrm{co}(\mathcal{G}))$ *y sea* $r = \dim(\mathcal{T}_{\mathcal{G}})$ *(posiblemente infinita). Sea* $\{\bar{\psi}_s\}$ *las funciones ortogonalizadas introducidas en el algoritmo y sea* $R_j\mathtt{f} = \mathtt{f} - \sum_{s=1}^{j} c_s\bar{\psi}_s$ *con* $\varphi_s \in \mathcal{G}$, $s = 1 : j$, *el residuo después de la iteración* j*-ésima. El algoritmo OMP converge en* $h \leq r$ *iteraciones y se verifica que:*

$$\lim_{h\to r} \|R_h\mathtt{f}\| = 0; \ \mathtt{f} = \sum_{j=1}^{h} \frac{\left\langle R_{j-1}\mathtt{f}, \bar{\psi}_j\right\rangle}{\left\|\bar{\psi}_j\right\|^2}\bar{\psi}_j; \ \ \|\mathtt{f}\|^2 = \sum_{j=1}^{h} \frac{\left|\left\langle R_{j-1}\mathtt{f}, \bar{\psi}_j\right\rangle\right|^2}{\left\|\bar{\psi}_j\right\|^2}$$

Ordenes *Greedy* de Aproximación con Diccionarios

[DeVore y Temlyakov 1996] define **diccionario** como un conjunto de funciones $\mathcal{G}$ en un espacio de Hilbert $\mathcal{G} \subset \mathcal{K}$ que verifica las dos condiciones siguientes:

1) $\|\psi\| = 1, \forall\psi \in \mathcal{G}$
2) $\psi \in \mathcal{G} \Rightarrow -\psi \in \mathcal{G}$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Si $\mathcal{G}$ está compuesto por una base ortonormal $\{\psi_i\}_{i=1}^{\infty}$ de $\mathcal{K}$, se supone que $\pm\psi_i \in \mathcal{G}$. Dado un diccionario $\mathcal{G}$, se definen las siguientes familias:

$$\mathcal{A}_{\tau,M}^0 = \{f \in \mathcal{K} : f = \textstyle\sum_{i\in\Lambda} c_i\psi_i \text{ con } \psi_i \in \mathcal{G},\ \#\Lambda < \infty,\ \sum_{i\in\Lambda} |c_i|^\tau \leq M^\tau\}$$
$$\mathcal{A}_{\tau,M} = \mathrm{cl}(\mathcal{A}_\tau^0) \text{ y } \mathcal{A}_\tau = \textstyle\bigcup_{M>0} \mathcal{A}_{\tau,M}$$

En $\mathcal{A}_\tau$ se define la seminorma [DeVore y Temlyakov 1996]: $|f|_{\mathcal{A}_\tau} = \inf\{M : f \in \mathcal{A}_{\tau,M}\}$. Si $\mathcal{G}$ es una base ortonormal, $\{\psi_i\}_{i=1}^{\infty}$, entonces $f \in \mathcal{A}_\tau \Leftrightarrow \sum_i |\langle f, \psi_i\rangle|^\tau < \infty$ y entonces $|f|_{\mathcal{A}_\tau} = \sum_i |\langle f, \psi_i\rangle|^\tau$.

Teorema 188 *[DeVore y Temlyakov 1996, th. 1.1] Si $\mathcal{G}$ es una base ortonormal, entonces para cada $\alpha > 0$ y $\tau = (\alpha + 1/2)^{-1}$, se verifica:*

$$\sum_{h=1}^{\infty} [h^\alpha \delta_h(\mathsf{f}, \mathcal{K})]^\tau \frac{1}{h} < \infty \Longleftrightarrow \mathsf{f} \in \mathcal{A}_\tau$$

El teorema anterior proporciona una caracterización de las funciones con un orden de aproximación $\delta_h(\mathsf{f}, \mathcal{K}) = O(h^{-\alpha})$. Dos ejemplos relevantes son:

1. Si $\mathcal{G}$ es una base wavelet ortonormal en $L_2(\mathbb{R})$, entonces $\mathcal{A}_\tau = \mathcal{B}_p^{\alpha p}(\mathbb{R})$.

2. Si $\mathcal{G}$ está definido mediante una base de Fourier $\{e^{ikx}, k \in \mathbb{Z}\}$, entonces el teorema anterior caracteriza el orden de aproximación a una función f por la sumabilidad absoluta de una potencia de sus coeficientes de Fourier. Por ejemplo, si $\alpha = 1/2$ y $\tau = 1$, $\mathcal{A}_1$ es la clase de las funciones cuya serie de Fourier converge absolutamente. Como caso especial del teorema anterior se obtiene [DeVore y Temlyakov 1996, p. 176], para $\tau = (\alpha + 1/2)^{-1}$:

$$\delta_h(\mathsf{f}, \mathcal{K}) \leq C\, |f|_{\mathcal{A}_\tau}\, h^{-\alpha} \tag{C.17}$$

   que para $\alpha = 1/2$ ($\tau = 1$) coincide con la tasa del lema de Maurey.

Teorema 189 *Si se utiliza un esquema de aproximación basado en un diccionario $\mathcal{G}$, se verifican las siguientes tasas de aproximación con algoritmos greedy:*

1. *Para el algoritmo greedy relajado con $\beta_h = 1/h$, $\forall \mathsf{f} \in \mathcal{A}_{1,1}$, se verifica que [DeVore y Temlyakov 1996, th. 3.2]:* $\|\mathsf{f} - f_h\|_{\mathcal{K}} \leq 2h^{-1/2}$

2. *Para el algoritmo greedy (puro) se verifica que [DeVore y Temlyakov 1996, th. 3.6],* $\forall \mathsf{f} \in \mathcal{A}_1 : \|\mathsf{f} - f_h\|_{\mathcal{K}} \leq |\mathsf{f}|_{\mathcal{A}_1}\, h^{-1/6}$.

3. *Para el algoritmo greedy (puro) se verifica que $\exists \mathsf{f} \in \mathcal{A}_\tau$ con $0 < \tau \leq 2$ tal que [DeVore y Temlyakov 1996, th. 4.1]:* $\|\mathsf{f} - f_h\|_{\mathcal{K}} \geq h^{-1/2}$.

4. *Para el algoritmo greedy ortogonal (con back-fitting) se verifica que [DeVore y Temlyakov 1996, th. 3.7],* $\forall \mathsf{f} \in \mathcal{A}_{1,M} : \|\mathsf{f} - f_h\|_{\mathcal{K}} \leq Mh^{-1/2}$.

Del teorema anterior, se concluye que para $\alpha = 1/2$, el algoritmo *greedy* (puro) produce una cota peor que (C.17), pero dicha cota se alcanza con el *greedy* relajado y con el *greedy* con *back-fitting*.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anexo D

Optimización

Algoritmos Basados en el Gradiente

Algoritmo del Gradiente

Dado el problema:

$$\min_{\mathbf{w}} L(\mathbf{w})$$

donde $\mathbf{w} \in W \subset \mathbb{R}^p$ y $L : \mathbb{R}^p \to \mathbb{R}$ es una función continuamente diferenciable en W, el **algoritmo del gradiente** es un método iterativo que se basa en la actualización del valor de los parámetros en cada iteración, en la dirección opuesta al gradiente $\nabla L(\mathbf{w}) \equiv \nabla_{\mathbf{w}} L(\mathbf{w})$ de la función objetivo.

En la iteración $(s+1)$-ésima, actualiza los parámetros según la ecuación:

$$\mathbf{w}_{s+1} = \mathbf{w}_s - \rho \nabla L(\mathbf{w}_s) \tag{D.1}$$

donde $\rho > 0$ es un parámetro denominado **tasa de aprendizaje** (*learning rate*). Dicho parámetro puede ser fijo o variable $\rho_s \equiv \rho(s)$, condicionando su convergencia (véase [Luo 1991], [Battiti 1992], [Bishop 1995]).

Si L es dos veces continuamente diferenciable, un método fundamentado (p. ej. [Duda *et al.* 2001, p. 225]) para seleccionar la tasa de aprendizaje ρ_s es partir de un desarrollo de Taylor de segundo orden de $L(\mathbf{w})$ en un entorno de $\mathbf{w}_s$:

$$L(\mathbf{w}) \approx L(\mathbf{w}_s) + \nabla L(\mathbf{w}_s)(\mathbf{w} - \mathbf{w}_s) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_s)^T \mathbf{H}_s (\mathbf{w} - \mathbf{w}_s) \tag{D.2}$$

donde $\mathbf{H}_s$ es el Hesiano de L en $\mathbf{w}_s$. Sustituyendo $\mathbf{w}$ en dicha expresión por el valor $\mathbf{w}_{s+1}$ de (D.1), se obtiene:

$$L(\mathbf{w}_{s+1}) \approx \nabla L(\mathbf{w}_s) - \rho_s \left\| \nabla L(\mathbf{w}_s) \right\|^2 + \frac{1}{2}\rho_s^2 \nabla L(\mathbf{w}_s) \mathbf{H}_s \nabla L(\mathbf{w}_s)$$

cuyo valor mínimo se obtiene para:

$$\rho_s = \frac{\left\| \nabla L(\mathbf{w}_s) \right\|^2}{\nabla L(\mathbf{w}_s) \mathbf{H}_s \nabla L(\mathbf{w}_s)}$$

Line Search. Algoritmo del Gradiente Conjugado

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Otra alternativa es determinar, en la dirección del gradiente $\nabla L(\mathbf{w}_s)$ el punto en el que se minimiza la función objetivo, sustituyendo así un problema de optimización p-dimensional, por una secuencia de problemas unidimensionales. Si $\mathbf{d}_s$ es la dirección elegida (en particular el gradiente)

en la iteración s-ésima, el problema es:

$$\min_{\rho} L(\mathbf{w}_s - \rho \mathbf{d}_s)$$

Las condiciones de óptimo para $\rho \equiv \rho_s$, son:

$$\frac{\partial}{\partial \rho} L(\mathbf{w}_s - \rho \mathbf{d}_s) = \frac{\partial}{\partial \rho} L(\mathbf{w}_{s+1}) = -\nabla L(\mathbf{w}_{s+1})^T \mathbf{d}_s = 0 \tag{D.3}$$

con lo que ambas direcciones son ortogonales. De esta forma, si la dirección siguiente $\mathbf{d}_{s+1}$ se elige también según el gradiente $\nabla_{\mathbf{w}} L(\mathbf{w}_{s+1})$ en el nuevo punto, se produce un efecto oscilotario perjudicial para la rapidez de convergencia, haciendo que lo conseguido en la iteración anterior se desperdicie en gran parte en la siguiente (es el efecto *greedy* de optimizar a corto plazo). Por tanto, elegir la mejor dirección en cada iteración puede no ser la mejor idea en términos de su efecto en el conjunto.

Se denominan algoritmos de *line search*, los algoritmos que utilizan la misma metodología que el algoritmo del gradiente:

$$\begin{aligned} \mathbf{w}_{s+1} &= \mathbf{w}_s + \rho_* \mathbf{d}_s \\ \rho_* &= \arg\min_{\rho} L(\mathbf{w}_s + \rho \mathbf{d}_s) \end{aligned}$$

pero eligen sucesivas direcciones $\mathbf{d}_s$ más adecuadas, con el objetivo de que, en la iteración siguiente, no se deshaga el avance conseguido en las anteriores. Ello se consigue si la componente del gradiente según la dirección anterior (que se acaba de hacer cero) se mantiene igualmente nula en el nuevo punto $\mathbf{w}_{s+2}$. Es decir:

$$\nabla L(\mathbf{w}_{s+1} + \rho \mathbf{d}_{s+1})^T \mathbf{d}_s = 0$$

Desarrollando en Taylor en $\rho = 0$, se tiene que:

$$\nabla L(\mathbf{w}_{s+1} + \rho \mathbf{d}_{s+1}) = \nabla L(\mathbf{w}_{s+1}) + \rho \mathbf{H}_{s+1}^T \mathbf{d}_{s+1}$$

con lo que la condición de óptimo, utilizando (D.3), produce:

$$\begin{aligned} 0 &= (\nabla L(\mathbf{w}_{s+1}) + \rho \mathbf{H}_{s+1}^T \mathbf{d}_{s+1})^T \mathbf{d}_s = \nabla L(\mathbf{w}_{s+1})^T \mathbf{d}_s + \rho \mathbf{d}_{s+1}^T \mathbf{H}_{s+1} \mathbf{d}_s \\ &\Rightarrow \quad \mathbf{d}_{s+1}^T \mathbf{H}_{s+1} \mathbf{d}_s = 0 \end{aligned}$$

Definición 190 *Dos direcciones $\mathbf{d}_i, \mathbf{d}_j$ distintas, se dicen mútuamente conjugadas con respecto a la matriz $\mathbf{H}$ si:*

$$\mathbf{d}_i^T \mathbf{H} \mathbf{d}_j = 0$$

Puede demostrarse (p. ej. [Bishop 1995, p. 279]) que para una función objetivo cuadrática, un algoritmo *line search* que actualice sucesivamente los parámetros mediante direcciones conjugadas, alcanza el mínimo en como máximo p iteraciones (la dimensión del espacio paramétrico). Dichas direcciones pueden obtenerse mediante las siguientes ecuaciones (versión Polak-Ribière –recomendada en [Press *et al.* 1992, p. 423]):

$$\begin{aligned} \mathbf{d}_{s+1} &= -\mathbf{g}_{s+1} + \beta_s \mathbf{d}_s \\ \beta_s &= \frac{\mathbf{g}_{s+1}^T (\mathbf{g}_{s+1} - \mathbf{g}_s)}{\mathbf{g}_s^T \mathbf{g}_s} \end{aligned}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde $\mathbf{g}_s \equiv \nabla L(\mathbf{w}_s)$ y $\mathbf{d}_1 = \mathbf{g}_0$ para algún valor inicial $\mathbf{w}_0$.

Algoritmos Basados en el Hesiano

Algoritmo de Newton

Si en lugar de utilizar el esquema (D.1) se minimiza directamente (D.2) en términos de $\mathbf{w}$, se obtiene el denominado **algoritmo de Newton**:

$$\begin{aligned} \nabla_{\mathbf{w}} L(\mathbf{w}_s) + \mathbf{H}_s(\mathbf{w} - \mathbf{w}_s) &= 0 \Rightarrow \mathbf{w} - \mathbf{w}_s = \mathbf{H}_s^{-1} \nabla_{\mathbf{w}} L(\mathbf{w}_s) \\ &\Rightarrow \mathbf{w}_{s+1} = \mathbf{w}_s - \mathbf{H}_s^{-1} \nabla_{\mathbf{w}} L(\mathbf{w}_s) \end{aligned}$$

que utiliza más información sobre L pero, en contrapartida, es más costoso computacionalmente pues es necesario calcular el Hessiano en cada iteración.

Algoritmo Fisher *scoring*

Si $\mathbf{z}^n$ es una muestra aleatoria simple, y si $L(\mathbf{w}) \equiv L(\mathbf{z}^n; \mathbf{w}) = -\ln q(\mathbf{z}^n; \mathbf{w})$ donde $q(\mathbf{z}^n; \mathbf{w})$ es la función de verosimilitud, el **algoritmo Fisher scoring**, utiliza la siguiente ecuación para la actualización de los parámetros:

$$\begin{aligned} \mathbf{w}_{s+1} &= \mathbf{w}_s + \left\{ \mathbb{E}_{\mathsf{Z}^n} \left[\frac{\partial^2 L(\mathsf{Z}^n, \mathbf{w}_s)}{\partial \mathbf{w}^2} \right] \right\}^{-1} \frac{\partial L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}} \\ &= \mathbf{w}_s - \left\{ \mathbb{E}_{\mathsf{Z}^n} \left[\frac{\partial L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}} \frac{\partial L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}} \right] \right\}^{-1} \frac{\partial L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}} \\ &= \mathbf{w}_s - \left\{ \mathbb{E}_{\mathsf{Z}^n} \left[\frac{\partial L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}} \frac{\partial L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}^T} \right] \right\}^{-1} \frac{\partial L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}} \end{aligned} \tag{D.4}$$

donde:

$$\mathcal{I}(\mathbf{w}) = -\left\{ \mathbb{E}_{\mathsf{Z}^n} \left[\frac{\partial^2 L(\mathsf{Z}^n, \mathbf{w}_s)}{\partial \mathbf{w}^2} \right] \right\} = \left\{ \mathbb{E}_{\mathsf{Z}^n} \left[\frac{\partial L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}} \frac{\partial L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}^T} \right] \right\}$$

es la matriz de información de Fisher, que es definida positiva.

Por tanto, el algoritmo Fisher *scoring* (ecuación (D.4)), puede verse como un algoritmo de Newton para minimizar L que utiliza la matriz de información de Fisher. Como dicha matriz usualmente no se conoce, se utiliza su versión empírica, la cuál, puede calcularse fácilmente por métodos iterativos (pág. 393) y resulta más estable que el Hesiano de L.

Si se aplica el algoritmo anterior al modelo no lineal: $\mathbf{y} = \mathbf{f}(\mathbf{w}) + \boldsymbol{\varepsilon}$ con $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{V})$, con $\mathbf{V}$ conocida, donde $\mathbf{y}, \boldsymbol{\varepsilon}, \mathbf{f}(\mathbf{w}) \in \mathbb{R}^n$, se tiene que:

$$\frac{\partial L}{\partial \mathbf{w}} = \frac{1}{\sigma^2} \dot{\mathbf{F}}^T \mathbf{V}^{-1} [\mathbf{y} - \mathbf{f}(\mathbf{w})]$$

donde $\dot{\mathbf{F}} = \partial \mathbf{f} / \partial \mathbf{w}^T$ es la matriz Jacobiana, y:

$$\begin{aligned} \mathbb{E}_{\mathsf{Z}^n} \left[\frac{\partial^2 L(\mathsf{Z}^n, \mathbf{w})}{\partial \mathbf{w}^2} \right] &= -\frac{1}{\sigma^2} \mathbb{E}_{\mathsf{Z}^n} \{ \dot{\mathbf{F}}^T \mathbf{V}^{-1} [\mathbf{y} - \mathbf{f}(\mathbf{w})] [\mathbf{y} - \mathbf{f}(\mathbf{w})]^T \mathbf{V}^{-1} \dot{\mathbf{F}} \} \\ &= -\frac{1}{\sigma^4} \mathbb{E}_{\mathsf{Z}^n} \{ \dot{\mathbf{F}}^T \mathbf{V}^{-1} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^T \mathbf{V}^{-1} \dot{\mathbf{F}} \} = -\frac{1}{\sigma^2} \dot{\mathbf{F}}^T \mathbf{V}^{-1} \dot{\mathbf{F}} \end{aligned}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

con lo que (D.4) se convierte en:

$$\begin{aligned}\mathbf{w}_{s+1} &= \mathbf{w}_s - \sigma^2(\dot{\mathbf{F}}_s^T\mathbf{V}^{-1}\dot{\mathbf{F}}_s)^{-1}\frac{1}{\sigma^2}\dot{\mathbf{F}}_s^T\mathbf{V}^{-1}\mathbf{e}_s \\ &= \mathbf{w}_s - (\dot{\mathbf{F}}_s^T\mathbf{V}^{-1}\dot{\mathbf{F}}_s)^{-1}\dot{\mathbf{F}}_s^T\mathbf{V}^{-1}\mathbf{e}_s\end{aligned}$$

donde $(\mathbf{e}_s)_i = y_i - [\mathbf{f}(\mathbf{w}_s)]_i$, $i = 1:n$, y el subíndice s indica valoración en $\mathbf{w} = \mathbf{w}_s$. En el caso $\mathbf{V} = \mathbf{I}_n$, se tiene:

$$\mathbf{w}_{s+1} = \mathbf{w}_s - (\dot{\mathbf{F}}_s^T\dot{\mathbf{F}}_s)^{-1}\dot{\mathbf{F}}^T\mathbf{e}_s$$

Algoritmo de Levenberg-Marquardt

El Algoritmo de Levenberg-Marquardt, desarrollado por Marquardt a partir de una idea de Levenberg, ([Chong y Zak 2001], [Bishop 1995], [Press *et al.* 1992]), pretende evitar la carga computacional asociada al algoritmo de Newton para el caso particular de mínimos cuadrados:

$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n} e_i(\mathbf{w})^2 = \frac{1}{2}\left\|\mathbf{e}(\mathbf{w})\right\|^2 \tag{D.5}$$

donde $e_i = \ell(y_i, f(\mathbf{x}_i))$ es la función de pérdida valorada en el dato i-ésimo $(y_i, \mathbf{x}_i)$. Supongamos que, en la iteración s, se aproxima linealmente la función de pérdida en un entorno de $\mathbf{w}_s$:

$$\mathbf{e}(\mathbf{w}) \approx \mathbf{e}_s + \dot{\mathbf{E}}_s(\mathbf{w} - \mathbf{w}_s) \tag{D.6}$$

donde $\mathbf{e}_s = \mathbf{e}(\mathbf{w}_s)$ y $\dot{\mathbf{E}}_s = \dot{\mathbf{E}}(\mathbf{w}_s) = \partial\mathbf{e}(\mathbf{w}_s)/\partial\mathbf{w}$, es la matriz Jacobiana. Sustituyendo en $L(\mathbf{w})$:

$$L(\mathbf{w}) \approx \frac{1}{2}\left\|\mathbf{e}_s + \dot{\mathbf{E}}_s(\mathbf{w} - \mathbf{w}_s)\right\|^2 \tag{D.7}$$

Si se desea minimizar esta expresión para determinar el siguiente punto $\mathbf{w}_{s+1}$, podemos considerarlo como un problema de mínimos cuadrados con matriz de diseño $-\dot{\mathbf{E}}$ y coeficientes $\mathbf{w} - \mathbf{w}_s$. Por tanto su solución es:

$$\mathbf{w} - \mathbf{w}_s = -(\dot{\mathbf{E}}_s^T\dot{\mathbf{E}}_s)^{-1}\dot{\mathbf{E}}_s^T\mathbf{e}_s$$

de donde:

$$\mathbf{w}_{s+1} = \mathbf{w}_s - (\dot{\mathbf{E}}_s^T\dot{\mathbf{E}}_s)^{-1}\dot{\mathbf{E}}_s^T\mathbf{e}_s \tag{D.8}$$

Por otra parte, el Hessiano $\mathbf{H}$ de (D.5) posee como elementos:

$$(\mathbf{H})_{ij} = \frac{\partial^2 L}{\partial w_i \partial w_j} = \sum_{l=1}^{n}\frac{\partial e_l}{\partial w_i}\frac{\partial e_l}{\partial w_j} + \sum_{l=1}^{n} e_l\frac{\partial^2 e_i}{\partial w_i \partial w_j} \approx \sum_{l=1}^{n}\frac{\partial e_l}{\partial w_i}\frac{\partial e_l}{\partial w_j},\ i,j = 1:p$$

que, para n no demasiado pequeño, puede aproximarse obviando los errores e_i (aproximación de Gauss-Newton[153]), es decir:

$$(\mathbf{H})_{ij} \approx \sum_{l=1}^{n}\frac{\partial e_l}{\partial w_i}\frac{\partial e_l}{\partial w_j},\ i,j = 1:p,\ \text{es decir, } \mathbf{H} \approx \sum_{l=1}^{n}\dot{\mathbf{e}}_l\dot{\mathbf{e}}_l^T = \dot{\mathbf{E}}^T\dot{\mathbf{E}} \tag{D.9}$$

[153] Puede observarse que ello resulta en un caso particular del algoritmo *Fisher scoring* para la optimización del menos logaritmo de la verosimilitud en el caso de ruido gausiano. Si el modelo incluye el modelo verdadero, el ruido es independiente de media cero, y los errores están incorrelados con las segundas derivadas, los errores e_i tienden a su esperanza, cero, cuando $n \to \infty$.

donde $\dot{\mathbf{e}}_i = \partial e_i / \partial \mathbf{w}$.

Con ello, como $\nabla L(\mathbf{w}_s) = \dot{\mathbf{E}}_s^T \mathbf{e}_s$, el algoritmo (D.8) es del tipo Newton: $\mathbf{w}_{s+1} = \mathbf{w}_s - \mathbf{H}^{-1}\nabla L$. Pero la aproximación lineal (D.6) no es válida si el tamaño de $\mathbf{w}_{s+1} - \mathbf{w}_s$ es demasiado grande, por lo que se controla regularizando el problema (D.7):

$$L_r(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{e}_s + \dot{\mathbf{E}}_s(\mathbf{w} - \mathbf{w}_s)\right\|^2 + \lambda \left\|\mathbf{w} - \mathbf{w}_s\right\|^2$$

que es un problema de regresión *ridge,* cuya solución es:

$$\begin{aligned} \mathbf{w} - \mathbf{w}_s &= -(\dot{\mathbf{E}}_s^T \dot{\mathbf{E}}_s + \lambda \mathbf{I}_m)^{-1} \dot{\mathbf{E}}_s^T \mathbf{e}_s \\ &\Rightarrow \mathbf{w}_{s+1} = \mathbf{w}_s - (\dot{\mathbf{E}}_s^T \dot{\mathbf{E}}_s + \lambda \mathbf{I}_m)^{-1} \dot{\mathbf{E}}_s^T \mathbf{e}_s \end{aligned}$$

En la práctica, el regularizador λ se monitoriza durante el ejecución del algoritmo: se inicia por ejemplo con un valor $\lambda = 0{,}1$ y, si en cada iteración, $L(\mathbf{w})$ disminuye, se mantiene su valor; en caso contrario, se incrementa en algún factor.

Finalmente, la aproximación (D.9) puede obtenerse de forma iterativa, mediante:

$$\mathbf{H}_{l+1} = \mathbf{H}_l + \frac{1}{n}\dot{\mathbf{e}}_{l+1}\dot{\mathbf{e}}_{l+1}^T, l = 1 : n-1$$

obteniéndose finalmente $\mathbf{H} = \mathbf{H}_n$. Con ello, $\mathbf{H}^{-1}$puede obtenerse mediante el lema de inversión de matrices:

$$\mathbf{H}_{l+1}^{-1} = \mathbf{H}_l^{-1} - \frac{\mathbf{H}_l^{-1}\dot{\mathbf{e}}_{l+1}\dot{\mathbf{e}}_{l+1}^T\mathbf{H}_l^{-1}}{n + \dot{\mathbf{e}}_{l+1}^T\mathbf{H}_l^{-1}\dot{\mathbf{e}}_{l+1}}, l = 1 : n-1$$

Optimización Estocástica. Algoritmo de Robbins-Monro

Sean G, W dos variables aleatorias correladas, donde $\mathsf{G}(w) \doteq \mathsf{G}/(\mathsf{W} = w)$ posee varianza finita para todo w. Consideramos el problema de determinar la raíz w^* de la ecuación:

$$L(w) = \mathbb{E}_{\mathsf{G}/w}[\mathsf{G}(w)] = b$$

con $b < \infty$, suponiendo que se verifica:

$$L(w) > 0 \text{ si } w < w^* \text{ y } L(w) < 0 \text{ si } w > w^* \tag{D.10}$$

El algoritmo de Robbins-Monro produce sucesivos estimadores de w^* mediante la ecuación:

$$\mathsf{W}_{n+1} = \mathsf{W}_n + a_n \mathsf{G}(\mathsf{W}_n) \tag{D.11}$$

donde los coeficientes a_n son de *tipo* $1/n$, es decir $C_1/n \leq a_n \leq C_2/n$, con C_i constantes $i = 1, 2$. Sin embargo, en general, son válidas las siguientes condiciones, [Bishop 1995, sec. 2.4]:

$$\begin{cases} 1) \lim\limits_{n\to\infty} a_n = 0 \\ 2) \sum\limits_{n=1}^{\infty} a_n = \infty \\ 3) \sum\limits_{n=1}^{\infty} a_n^2 < \infty \end{cases}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

donde la primera condición asegura que las sucesivas correcciones de w_n tienden a cero, la segunda asegura que la corrección es suficientemente grande como para que se encuentre la raíz w^*, y la tercera garantiza que la aleatoriedad acumulada por el algoritmo (D.11) posea varianza finita de tal forma que no impida la convergencia.

Se verifica el siguiente teorema, cuya demostración puede verse en [Robbins y Monro 1951]:

Teorema 191 *Sea* $\mathsf{G}(w) = \mathsf{G}/(\mathsf{W} = w)$ *variable aleatoria acotada casi seguro, es decir,* $\mathbb{P}_{\mathsf{G}/w}[|\mathsf{G}(w)| \leq C] = 1$, *para todo* w. *Si* $\{a_n\}$ *es de tipo* $1/n$ *y* $L(w)$ *satisface (D.10), entonces el proceso estocástico no estacionario* $\{\mathsf{W}_n\}$ *obtenido de (D.11), converge en media cuadrática a la raíz* w^*, *es decir:*

$$\lim_{n\to\infty} \mathbb{E}[(\mathsf{W}_n - w^*)^2] = 0$$

Por tanto, el algoritmo de Robbins-Monro es consistente.

El algoritmo de Robbins-Monro posee especial aplicación en problemas de estimación secuencial, (*on-line*), es decir, en los que cada elemento de la muestra se obtiene en diferentes momentos en el tiempo, y se requiere una estimación instantánea de un determinado parámetro de la población.

Aplicación a la Estimación *On-line* mediante la Máxima Verosimilitud

Supongamos el problema de estimación mediante la máxima verosimilitud de un parámetro w, cuya solución verifica la condición de óptimo:

$$\frac{\partial}{\partial w}[\frac{1}{n}L(\mathbf{y};\mathbf{x},w)] = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}\ln q(y_i;x_i,w) = 0$$

donde q es la función de masa o de densidad y hemos añadido por conveniencia el factor $1/n$, para tomar el límite en el tamaño muestral:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w}\ln p(y_i;x_i,w) = \mathbb{E}_{\mathsf{Y}/x}[\frac{\partial}{\partial w}\ln q(\mathsf{Y}|x;w)]$$

Con ello, asintóticamente, la estimación de máxima verosimilitud equivale a la solución de la ecuación:

$$\mathbb{E}_{\mathsf{Y}/x}[\frac{\partial}{\partial w}\ln q(\mathsf{Y}|x,w)] = 0$$

Aplicando el algoritmo de Robbins-Monro, con $b = 0$, $\mathsf{G}(w) = \frac{\partial}{\partial w}\ln q(\mathsf{Y}|x,w)$ y W la variable aleatoria que toma valores w, se obtiene la expresión:

$$\mathsf{W}_{n+1} = \mathsf{W}_n + a_n\frac{\partial}{\partial w}\ln q(\mathsf{Y}_{n+1}|x_{n+1};\mathsf{W}_n)$$

que se utiliza para la estimación *on-line* del valor óptimo w^*, para cada observación Y/x disponible. Nótese que este algoritmo no requiere el almacenamiento de toda la muestra para determinar la estimación.

El algoritmo *back-propagation* es un caso particular del algoritmo de Robbins-Monro cuando se utiliza en su versión *on line*, es decir, se presenta cada observación al algoritmo de forma

secuencial, para producir sucesivos estimadores del vector óptimo de parámetros w de la red neuronal. El problema de las redes MLP es que las condiciones de óptimo del problema de optimización poseen más de un punto crítico, (óptimos locales), con lo que no puede garantizarse la consistencia al óptimo global, (aunque sí a un óptimo local).

Problema de Optimización con Restricciones

Lo que sigue puede verse de forma detallada, p. ej. en [Chong y Zak 2001] y [Luenberger 1984].

Definición 192 *Un problema (**Primal**) de optimización con restricciones se formula de la siguiente forma: dadas funciones* $f, \mathbf{g} = (g_1, ..., g_r), \mathbf{h} = (h_1, ..., h_m)^T$, *con* f, g_i, h_j, $i = 1, ..., r$, $j = 1, ..., m$, *definidas en* $\Omega \subseteq \mathbb{R}^p$, *determinar:*

$$\begin{cases} \min\limits_{w \in \Omega} f(\mathbf{w}) \text{ sujeto a:} \\ \quad \mathbf{g}(\mathbf{w}) \leq 0 \\ \quad \mathbf{h}(\mathbf{w}) = 0 \end{cases} \tag{D.12}$$

Definición 193 *La función* $f(\mathbf{w})$ *se denomina **función objetivo o de coste**. Se denomina **región factible** la región en la que el problema está definido:* $R = \{\mathbf{w} \in \Omega : \mathbf{g}(\mathbf{w}) \leq 0, \mathbf{h}(\mathbf{w}) = 0\}$. *Una restricción de desigualdad* $g_i(\mathbf{w}) \leq 0$ *se dice **activa** si, en el óptimo* $\mathbf{w}^*$, *supuesto existente, verifica:* $g_i(\mathbf{w}^*) = 0$. *Se denomina **slack**, a una variable* $\xi_i \geq 0$ *asociada a la restricción* $g_i(\mathbf{w}) \leq 0$ *que la convierte en activa:* $g_i(\mathbf{w}) + \xi_i = 0$. *Si la restricción es activa,* $\xi_i = 0$.

Definición 194 *Una función* $f : \mathbb{R}^p \to \mathbb{R}$ *se dice **convexa** si* $\forall \mathbf{u}, \mathbf{w} \in \mathbb{R}^p$ *y* $\forall \lambda \in (0, 1)$:

$$f(\lambda \mathbf{u} + (1 - \lambda)\mathbf{w}) \leq \lambda f(\mathbf{u}) + (1 - \lambda) f(\mathbf{w})$$

*y **estrictamente convexa** si la desigualdad es estricta. Un conjunto* $\Omega \subseteq \mathbb{R}^p$ *se dice **convexo** si* $\forall \mathbf{u}, \mathbf{w} \in \Omega$ *y* $\forall \lambda \in (0, 1) : \lambda \mathbf{u} + (1 - \lambda)\mathbf{w} \in \Omega$.

Lema 195 *Si una función* f *es convexa, cualquier mínimo local* $\mathbf{w}^*$ *del problema sin restricciones,* $\min_{\mathbf{w} \in \Omega} f(\mathbf{w})$, *es también mínimo global.*

Demostración. $\forall \mathbf{u} \neq \mathbf{w}^*$, $\exists \lambda \in (0, 1)$, tal que:

$$f(\mathbf{w}^*) \leq f(\lambda \mathbf{w}^* + (1 - \lambda)\mathbf{u}) \leq \lambda f(\mathbf{w}^*) + (1 - \lambda) f(\mathbf{u}) \Rightarrow f(\mathbf{w}^*) \leq f(\mathbf{u})$$

■

Definición 196 *Un problema de optimización con restricciones se dice **convexo** si* Ω, *la función objetivo y todas las restricciones son convexas.*

Teorema 197 ***Fermat**. Una condición necesaria para que* $\mathbf{w}^*$ *sea mínimo de* $f(\mathbf{w})$, *con* f *función diferenciable en* $\mathbf{w}^*$, *es que* $\nabla_{\mathbf{w}} f(\mathbf{w}^*) = \mathbf{0}$. *Si, además,* f *es convexa, dicha condición es suficiente.*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Definición 198 *Dado un problema de optimización con función objetivo $f(\mathbf{w})$ y restricciones de igualdad $\mathbf{h}(\mathbf{w}) = \mathbf{0}$, se define la **Función de Lagrange (Lagrangiano)** como la función:*

$$L(\mathbf{w}, \lambda) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{w})$$

*donde los λ_i se denominan **coeficientes de Lagrange**.*

Teorema 199 ***Lagrange**. Una condición necesaria para que $\mathbf{w}^*$ sea solución de $\min_{\mathbf{w}\in\Omega} f(\mathbf{w})$ sujeto a $\mathbf{h}(\mathbf{w}) = \mathbf{0}$, con $f, \mathbf{h}$ diferenciables en un entorno de $\mathbf{w}^*$ es que, exista $\boldsymbol{\lambda}^* = (\lambda_1^*, ..., \lambda_m^*)$ tal que:*

$$\begin{cases} \nabla_{\mathbf{w}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \\ \nabla_{\boldsymbol{\lambda}} f(\mathbf{w}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \end{cases}$$

Si, además, $L(\mathbf{w}, \boldsymbol{\lambda}^*)$ es convexa como función de $\mathbf{w}$, entonces dichas condiciones son suficientes.

Los coeficientes de Lagrange expresan la tasa de variación del valor óptimo $f(\mathbf{w}^*)$ ante cambios en el valor de las restricciones, pues:

$$\frac{\partial L(\mathbf{w}^*, \boldsymbol{\lambda}^*)}{\partial h_i} = \lambda_i^*, \ i = 1 : m$$

Definición 200 *Dado un problema de optimización con restricciones de igualdad y desigualdad, (D.12), se define la **función de Lagrange (Lagrangiano) generalizada** como la función, con $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_r)$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)$:*

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{r} \alpha_i g_i(w) + \sum_{i=1}^{m} \beta_i h_i(w)$$

Definición 201 *Se define el problema de optimización **Dual**, asociado al problema (**Primal**) (D.12), como el problema:*

$$\begin{cases} \max \rho(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{sujeto a: } \boldsymbol{\alpha} \geq 0 \end{cases}$$

*donde $\rho(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\mathbf{w}\in\Omega} L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Si $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ es la solución óptima, se denomina **valor del problema** a: $\rho(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$.*

Teorema 202 *Sea $\mathbf{w} \in \Omega$ una solución factible del problema primal y $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ solución factible del problema dual. Entonces: $f(\mathbf{w}) \geq \rho(\boldsymbol{\alpha}, \boldsymbol{\beta})$.*

Demostración. Se verifica que:

$$\rho(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \inf_{\mathbf{u}\in\Omega} L(\mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{w}) + \boldsymbol{\alpha}^T \mathbf{g}(\mathbf{w}) + \boldsymbol{\beta}^T \mathbf{h}(\mathbf{w}) \leq f(\mathbf{w})$$

pues, al ser $\mathbf{w}$ factible, $\mathbf{g}(\mathbf{w}) \leq \mathbf{0}$, $\mathbf{h}(\mathbf{w}) = \mathbf{0}$ y al ser $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ factible se tiene $\boldsymbol{\alpha} \geq \mathbf{0}$.

■

De esta forma, el valor del problema dual es siempre menor o igual que el valor del problema primal.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Corolario 203 *Si se verifica que $f(\mathbf{w}^*) = \rho(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, siendo $\mathbf{w}^*$ y $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ soluciones factibles para los problemas primal y dual respectivamente, entonces $\mathbf{w}^*$ y $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ son soluciones del problema primal y dual, respectivamente, en cuyo caso: $\alpha_i^* g_i(\mathbf{w}^*) = 0, i = 1 : r$.*

Demostración. La condición de óptimos para los respectivos problemas está clara pues alcanzan sus cotas inferior y superior. Además, como en la demostración del teorema anterior, por hipótesis, las desigualdades son igualdades, ha de ser $\boldsymbol{\alpha}^{*T}\mathbf{g}(\mathbf{w}^*) = 0$.

■

Los problemas primal y dual pueden poseer soluciones óptimas sin que se alcance la igualdad anterior, en cuyo caso, se denomina **gap de dualidad** a la diferencia $f(\mathbf{w}^*) - \rho(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$.

Definición 204 *Un punto $(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ se dice **Punto de Silla** para el Lagrangiano generalizado si $\mathbf{w}^* \in \Omega, \boldsymbol{\alpha}^* \geq 0$ y:*

$$L(\mathbf{w}^*, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq L(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \leq L(\mathbf{w}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$

Teorema 205 *$(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ es punto de silla del Lagrangiano generalizado si y sólo si $\mathbf{w}^*$ y $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ son soluciones del problema primal y dual, respectivamente y no existe gap de dualidad.*

Teorema 206 ***Teorema de Dualidad Fuerte**. Si el problema primal (D.12), verifica que Ω es convexo y las funciones $\mathbf{g}$, $\mathbf{h}$ son funciones lineales, ($\mathbf{g}(\mathbf{w}) = \mathbf{A}\mathbf{w} - \mathbf{b}$, $\mathbf{h}(\mathbf{w}) = \mathbf{B}\mathbf{w} - \mathbf{c}$), entonces, el gap de dualidad es cero.*

Teorema 207 ***Condiciones de óptimo de Kuhn-Tucker**. Dado un problema primal (D.12) con dominio Ω convexo, f diferenciable y $\mathbf{g}, \mathbf{h}$ lineales como en el teorema anterior, entonces, $\mathbf{w}^*$ es óptimo para el problema si y sólo si existen $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$, tales que:*

$$\begin{aligned}
\nabla_{\mathbf{w}} f(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) &= \mathbf{0} \\
\nabla_{\boldsymbol{\beta}} f(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) &= \mathbf{0} \\
\alpha_i^* g_i(\mathbf{w}^*) &= 0, i = 1 : r \\
g_i(\mathbf{w}^*) &\leq 0, i = 1 : r \\
\alpha_i^* &\geq 0, i = 1 : r
\end{aligned}$$

La relación $\alpha_i^* g_i(\mathbf{w}^*) = 0, i = 1 : r$ se denomina **relación de complementariedad de Kuhn-Tucker**, e implica que, para restricciones activas, se verifica $\alpha_i^* \geq 0$, mientras que, para restricciones inactivas, $\alpha_i^* = 0$. Asimismo, al igual que con los multiplicadores de Lagrange, α_i^* refleja la tasa de cambio en la función objetivo al variar el término independiente de la restricción $g_i(\mathbf{w}^*)$.

Aplicación a un Programa Cuadrático

Sea el problema (primal):

$$\begin{cases} \min\limits_{w} \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} - \mathbf{b}^T\mathbf{w} \text{ sujeto a:} \\ \mathbf{C}\mathbf{w} \leq \mathbf{d} \end{cases}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

con $\mathbf{A}_{p\times p}, \mathbf{b}_{p\times 1}, \mathbf{C}_{m\times p}, \mathbf{d}_{m\times 1}$. Suponiendo que la región factible no es vacía, planteamos el problema dual:

$$\max_{\boldsymbol{\alpha}\geq 0}\{\min_{w}\{\tfrac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}-\mathbf{b}^T\mathbf{w}+\boldsymbol{\alpha}^T(\mathbf{C}\mathbf{w}-\mathbf{d})\}\}$$

donde $\boldsymbol{\alpha}\in\mathbb{R}^m$ y $\boldsymbol{\alpha}\geq 0$ significa $\alpha_i\geq 0\ \forall i\in\{1:m\}$.

El mínimo se obtiene para $\mathbf{w}=\mathbf{A}^{-1}(\mathbf{b}-\mathbf{C}^T\boldsymbol{\alpha})$ y sustituyendo esta expresión en el primal, se tiene el dual:

$$\begin{cases}\max\limits_{\boldsymbol{\alpha}} -\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{P}\boldsymbol{\alpha}-\boldsymbol{\alpha}^T\mathbf{u}-\frac{1}{2}\mathbf{b}^T\mathbf{Q}\mathbf{b}\\ \text{sujeto a: } \boldsymbol{\alpha}\geq 0\end{cases}$$

donde $\mathbf{P}=\mathbf{C}\mathbf{A}^{-1}\mathbf{C}^T$, $\mathbf{u}=\mathbf{d}-\mathbf{C}\mathbf{A}^{-1}\mathbf{b}$. De esta forma, el dual de un programa cuadrático es otro programa cuadrático pero con restricciones más sencillas.

Algoritmos de Punto Interior

Sea el siguiente programa cuadrático:

$$\begin{gathered}\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathbb{R}^n}\{-\tfrac{1}{2}\sum_{i,j=1}^{n}(\alpha_i'-\alpha_j)(\alpha_j'-\alpha_j)k(\mathbf{x}_i,\mathbf{x}_j)-\varepsilon\sum_{i=1}^{n}(\alpha_i'+\alpha_i)+\textstyle\sum_{i=1}^{n}(\alpha_i'-\alpha_i)y_i\\ \begin{cases}\sum\limits_{i=1}^{n}(\alpha_i'-\alpha_i)g_j(\mathbf{x}_i)=0,\ j=1:r\\ 0\leq\alpha_i,\alpha_i'\leq C,\ i=1:n\end{cases}\end{gathered}\tag{D.13}$$

Dicho programa incluye el caso de la SVM con término constante sin más que considerar $j=1$ y $g_1(\mathbf{x})=1,\ \forall\mathbf{x}\in\mathcal{X}$ (y sin término constante si $g=0$). El problema de clasificación produce un programa similiar a los efectos de lo tratado en este anexo.

Para poder utilizar algoritmos estándar de optimización cuadrática, el programa (D.13) puede escribirse de la siguiente forma [Drucker *et al.* 1997]:

$$\begin{gathered}\min_{\boldsymbol{\gamma}}\{\tfrac{1}{2}\boldsymbol{\gamma}^T\bar{\mathbf{K}}\boldsymbol{\gamma}+\boldsymbol{\gamma}^T\mathbf{c}\}\\ \boldsymbol{\gamma}^T\bar{\mathbf{G}}=\mathbf{0}_r^T,\ \mathbf{0}_{2n}\leq\boldsymbol{\gamma}\leq C\mathbf{1}_{2n}\end{gathered}\tag{D.14}$$

donde:

$$\boldsymbol{\gamma}=\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\alpha}'\end{bmatrix},\ \bar{\mathbf{K}}=\begin{bmatrix}\mathbf{K} & -\mathbf{K}\\ -\mathbf{K} & \mathbf{K}\end{bmatrix},\bar{\mathbf{G}}=\begin{bmatrix}\mathbf{G}\\ -\mathbf{G}\end{bmatrix},$$
$$(\mathbf{G})_{ij}=g_j(\mathbf{x}_i),j=1,\ldots,r,\ i=1,\ldots.n$$
$$\mathbf{c}=(\varepsilon+y_1,\ldots,\varepsilon+y_n,\varepsilon-y_1,\ldots,\varepsilon-y_n)^T$$

Algoritmos especialmente eficientes para resolver programas del tipo (D.14) son los que utilizan el **método del punto interior** (*interior point*), (por ejemplo, el algoritmo *primal-dual path-following* de [Vanderbei 1998] y su adaptación a SVM en [Smola y Schölkopf 1998]), que resuelven simultáneamente (D.14) (que ahora actúa como primal) y su dual.

Si las restricciones $\mathbf{0}\leq\boldsymbol{\gamma}\leq C\mathbf{1}$ de (D.14) se escriben como $\mathbf{0}\leq\boldsymbol{\gamma}$ y $\boldsymbol{\gamma}+\boldsymbol{\delta}=C\mathbf{1}$ con $\boldsymbol{\delta}\geq\mathbf{0}$ nueva variable primal), el Lagrangiano es:

$$L(\boldsymbol{\gamma},\boldsymbol{\delta},\mathbf{h},\mathbf{s},\mathbf{z},\mathbf{v})=\frac{1}{2}\boldsymbol{\gamma}^T\bar{K}\boldsymbol{\gamma}+\mathbf{c}^T\boldsymbol{\gamma}+\mathbf{h}^T(\bar{G}^T\boldsymbol{\gamma}-\mathbf{d})+\mathbf{s}^T(\boldsymbol{\gamma}+\boldsymbol{\delta}-C\mathbf{1})-\mathbf{z}^T\boldsymbol{\gamma}-\mathbf{v}^T\boldsymbol{\delta}$$

USC DE SANTIAGO

con $\mathbf{z}, \mathbf{v} \geq 0$ y $\mathbf{h}, \mathbf{s}$ sin restricciones. Las condiciones de optimalidad son:

$$\left.\begin{array}{l} \dfrac{\partial L}{\partial \boldsymbol{\gamma}} = \bar{K}\boldsymbol{\gamma} + \mathbf{c} + \bar{G}\mathbf{h} + \mathbf{s} - \mathbf{z} = \mathbf{0} \text{ (factibilidad dual)} \\ \dfrac{\partial L}{\partial \mathbf{h}} = \bar{G}^T\boldsymbol{\gamma} - \mathbf{d} = 0 \\ \dfrac{\partial L}{\partial \mathbf{s}} = \boldsymbol{\gamma} + \boldsymbol{\delta} - C\mathbf{1} = 0 \end{array}\right\} \text{(factibilidad primal)}$$

además de $\mathbf{s} = \mathbf{v} \geq 0$. Con ello, el dual es:

$$\begin{array}{c} \max\limits_{\mathbf{h},\mathbf{s},\mathbf{z}}\{-\dfrac{1}{2}\boldsymbol{\gamma}^T\bar{K}\boldsymbol{\gamma} - \mathbf{d}^T\mathbf{h} - C\mathbf{1}^T\mathbf{s}\} \\ \bar{K}\boldsymbol{\gamma} + \mathbf{c} + \bar{G}\mathbf{h} + \mathbf{s} - \mathbf{z} = \mathbf{0} \\ \mathbf{s} \geq \mathbf{0}, \mathbf{z} \geq \mathbf{0} \text{ con } \mathbf{h} \text{ libre} \end{array}$$

Debido a que las condiciones de Kuhn-Tucker son necesarias y suficientes, basta encontrar una solución factible para ambos problemas que satisfaga las condiciones de complementariedad $\boldsymbol{\gamma}^T\mathbf{z} = 0 = \boldsymbol{\delta}^T\mathbf{s}$. A partir de unos valores iniciales, esto se consigue obteniendo en cada iteración nuevos valores para las variables primales $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, y duales $\mathbf{s}$, $\mathbf{z}$, $\mathbf{h}$, resolviendo las condiciones de optimalidad:

$$\left.\begin{array}{l} \dfrac{\partial L}{\partial \boldsymbol{\gamma}} = \bar{K}\boldsymbol{\gamma} + \mathbf{c} + \bar{G}\mathbf{h} + \mathbf{s} - \mathbf{z} = \mathbf{0} \text{ (factibilidad dual)} \\ \dfrac{\partial L}{\partial \mathbf{h}} = \bar{G}^T\boldsymbol{\gamma} - \mathbf{d} = \mathbf{0} \\ \dfrac{\partial L}{\partial \mathbf{s}} = \boldsymbol{\gamma} + \boldsymbol{\delta} - C\mathbf{1} = \mathbf{0} \end{array}\right\} \text{(factibilidad primal)}$$

siendo L el Lagrangiano, y satisfaciendo progresivamente las condiciones de complementariedad: $\boldsymbol{\gamma}^T\mathbf{z} = \mu$, $\boldsymbol{\delta}^T\mathbf{s} = \mu$, donde μ se va reduciendo en cada iteración. Dado que, en el óptimo, el máximo del dual coincide con el mínimo del primal, el algoritmo finaliza cuando el *gap* primal-dual (*duality gap*) se hace suficientemente pequeño.

La variación en las variables primales y duales en cada iteración se obtiene como solución de la linealización [Smola y Schölkopf 1998] de las condiciones anteriores de factibilidad y complementariedad (etapa predictora):

$$\begin{array}{rcccc} \bar{G}^T\Delta\boldsymbol{\gamma} & = & \mathbf{d} - \bar{G}^T\boldsymbol{\gamma} & \doteq & \boldsymbol{\rho}_{p1} \\ \Delta\boldsymbol{\gamma}+\Delta\boldsymbol{\delta} & = & C\mathbf{1} - \boldsymbol{\gamma} - \boldsymbol{\delta} & \doteq & \boldsymbol{\rho}_{p2} \\ \bar{K}\Delta\boldsymbol{\gamma}+\bar{G}^T\Delta\mathbf{h} + \Delta\mathbf{s} - \Delta\mathbf{z} & = & -\bar{K}\boldsymbol{\gamma} - \mathbf{c} - \bar{G}\mathbf{h} - \mathbf{s} + \mathbf{z} & \doteq & \boldsymbol{\rho}_d \\ \gamma_i^{-1}z_i\Delta\gamma_i + \Delta z_i & = & \mu\gamma_i^{-1} - z_i - \gamma_i^{-1}\Delta z_i\Delta\gamma_i & \doteq & (\rho_{KT1})_i \\ t_i^{-1}s_i\Delta t_i + \Delta s_i & = & \mu t_i^{-1} - s_i - t_i^{-1}\Delta s_i\Delta t_i & \doteq & (\rho_{KT2})_i \end{array}$$

y su corrección posterior en las condiciones no lineales (etapa correctora). La solución a este sistema produce las ecuaciones:

$$\begin{array}{l} \Delta\boldsymbol{\delta} = \boldsymbol{\rho}_{p2} - \Delta\boldsymbol{\gamma} \\ \Delta z_i = (\rho_{KT1})_i - \gamma_i^{-1}z_i\Delta\gamma_i \\ \Delta s_i = (\rho_{KT2})_i - t_i^{-1}s_i\Delta\gamma_i \end{array}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

así como el sistema reducido de Kuhn-Tucker:

$$\begin{bmatrix} \bar{K} + \text{diag}(\boldsymbol{\delta}^{-1}\mathbf{s} + \boldsymbol{\gamma}^{-1}\mathbf{z}) & \bar{G} \\ \bar{G}^T & \mathbf{0}_{r\times r} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\gamma} \\ \Delta\mathbf{h} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\rho}_d + \boldsymbol{\rho}_{KT1} - \boldsymbol{\rho}_{KT2} \\ \boldsymbol{\rho}_{p1} \end{bmatrix} \tag{D.15}$$

donde $\boldsymbol{\delta}^{-1}$ denota inversión de cada componente de $\boldsymbol{\delta}$ y $\boldsymbol{\delta}^{-1}\mathbf{s}$ denota producto componente a componente (y lo mismo con $\boldsymbol{\gamma}^{-1}\mathbf{z}$). Este último sistema puede resolverse mediante:

$$\begin{aligned} \Delta\mathbf{h} &= (\bar{G}^T\widetilde{K}\bar{G})^{-1}\left[\bar{G}^T\widetilde{K}^{-1}\mathbf{a}_1 - \mathbf{a}_2\right] \\ \Delta\boldsymbol{\gamma} &= \widetilde{K}^{-1}\left[\mathbf{a}_1 - \bar{G}(\bar{G}^T\widetilde{K}\bar{G})^{-1}(\bar{G}^T\widetilde{K}^{-1}\mathbf{a}_1 - \mathbf{a}_2)\right] \end{aligned}$$

donde $\widetilde{K} = \bar{K} + \text{diag}(\boldsymbol{\delta}^{-1}\mathbf{s} + \boldsymbol{\gamma}^{-1}\mathbf{z})$, $\mathbf{a}_1 = \boldsymbol{\rho}_d + \boldsymbol{\rho}_{KT1} - \boldsymbol{\rho}_{KT2}$ y $\mathbf{a}_2 = \boldsymbol{\rho}_{p1}$.

Vanderbei ([Vanderbei 1998], [Smola y Schölkopf 1998]) recomienda elegir las condiciones iniciales mediante regularización de (D.15) mediante:

$$\begin{bmatrix} \bar{K} + I_{2n} & \bar{G} \\ \bar{G}^T & -I_r \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}_0 \\ \mathbf{h}_0 \end{bmatrix} = \begin{bmatrix} -\mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

asegurando posteriormente la factibilidad de la solución en las demás variables. [Smola y Schölkopf 1998] recomienda actualizar μ mediante:

$$\mu = \frac{\boldsymbol{\gamma}^T\mathbf{z} + \boldsymbol{\delta}^T\mathbf{s}^T}{4n}\left(\frac{1-\kappa}{10+\kappa}\right)^2$$

donde κ es un factor de contracción con el fin de asegurar que las variables satisfacen las restricciones de positividad. El criterio de parada del algoritmo se basa en el tamaño relativo del *gap* entre la función objetivo primal y la dual, [Schölkopf, Herbrich, Smola y Williamson 2000].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anexo E

Consistencia del Algoritmo *Back-Propagation*

Sea $\mathcal{F}$ la familia de funciones implementadas por redes neuronales:

$$\mathcal{F} = \{f(\cdot;\boldsymbol{\theta}) = \varphi[\sum_{j=1}^{h} \psi_j(\cdot)] : \mathcal{X} \subset \mathbb{R}^d \to \widehat{\mathcal{Y}} \subset \mathbb{R} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$$

como familia de hipótesis de un problema de aprendizaje con pérdida cuadrática $\ell(y, f(\mathbf{x})) = [y - f(\mathbf{x})]^2$ y sea $\{\mathbf{z}_i\}_{i=1}^n$ con $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i = 1 : n$ una muestra aleatoria simple obtenida de $\mathbb{P}_{\mathsf{Z}}$.

Hipótesis 1. $\mathcal{X}$ es compacto, Θ es convexo y compacto, $\widehat{\mathcal{Y}}$ es acotado. Además, las funciones de activación φ y ψ_j, $j = 1 : h$, son dos veces continuamente diferenciables (y, con ello, las funciones $f_{\boldsymbol{\theta}}(\mathbf{x}) \equiv f(\mathbf{x};\boldsymbol{\theta})$ con respecto a $\mathcal{X}$ y Θ, que, así, hasta sus segundas parciales son uniformemente acotadas).

La convexidad de Θ se exige para poder utilizar desarrollos de Taylor dicho conjunto. Al ser $\widehat{\mathcal{Y}}$ acotado, están acotadas las funciones $\ell(\mathbf{z}, f_{\boldsymbol{\theta}}) = \ell(y, f_{\boldsymbol{\theta}}(\mathbf{x}))$ con $f_{\boldsymbol{\theta}} \in \mathcal{F}$:

$$\exists b < \infty : \forall \mathbf{x} \in \mathcal{X}, \forall \boldsymbol{\theta} \in \Theta, \forall \hat{y} \in \widehat{\mathcal{Y}}, \forall f_{\boldsymbol{\theta}} \in \mathcal{F} : \ell(\mathbf{z}, f_{\boldsymbol{\theta}}) = \ell(y, f_{\boldsymbol{\theta}}(\mathbf{x})) = [y - f_{\boldsymbol{\theta}}(\mathbf{x})]^2 \leq b$$

Con ello, se tiene $\hat{R}_n(f_{\boldsymbol{\theta}}) \equiv \hat{R}_n(\boldsymbol{\theta}) \leq b$ y $R(f_{\boldsymbol{\theta}}) \equiv R(\boldsymbol{\theta}) \leq b \; \forall f_{\boldsymbol{\theta}} \in \mathcal{F}$, además, existen los gradientes $\nabla_{\boldsymbol{\theta}} \hat{R}_n(\boldsymbol{\theta})$, $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})$ y los Hesianos $\hat{\mathbf{H}}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, $\mathbf{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ con $(\hat{\mathbf{H}}_{\boldsymbol{\theta}}(\boldsymbol{\theta}))_{ij} = \frac{\partial^2 \hat{R}_n}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta})$, $i, j = 1 : p$, y análogamente con $\mathbf{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\partial^2 R}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\theta})$, y poseen elementos acotados y uniformemente continuos.

Sean $E_n = \{\boldsymbol{\theta} : \nabla_{\boldsymbol{\theta}} \hat{R}_n(\boldsymbol{\theta}) = \mathbf{0}\}$ y M_n el conjunto de mínimos locales y globales de $\hat{R}_n$ y E, M, los correspondientes a $R(\boldsymbol{\theta})$.

Hipótesis 2. Criterio de parada del algoritmo. Sea $\{\delta_n\} \to 0$ con $\delta_n > 0$. El algoritmo de entrenamiento no se para hasta que se encuentra un $\hat{\boldsymbol{\theta}}_n$ tal que:

1) $\left\| \nabla_{\boldsymbol{\theta}} \hat{R}_n(\hat{\boldsymbol{\theta}}_n) \right\| < \delta_n$
2) $\lim_{n\to\infty} \sqrt{n}\delta_n = 0$
3) $\hat{\mathbf{H}}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ es definida positiva

Este criterio es válido también para algoritmos *on line* si se verifican dichas condiciones ocasionalmente.

La hipótesis anterior, indica que los resultados que se obtengan son válidos para cualquier algoritmo basado en el gradiente, incluidos los que aproximan el Hesiano mediante el gradiente (p. ej., Quasi-Newton y Levenberg-Marquardt).

Hipótesis 3. $R(\boldsymbol{\theta})$ tiene un número finito de mínimos, $\#M < \infty$, tal que $\exists B \subset \mathbb{R}^p$ abierto con $M \subset B \subset \Theta$. Es decir, $M = \{\boldsymbol{\theta}_j^*\}_{j=1}^m$ con $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}_j^*) = 0$ y $\mathbf{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_j^*)$ es definida positiva, $j = 1 : m$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Hipótesis 4. Sea $\rho_j(\boldsymbol{\theta}) = \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_j^*\right\|$, $\boldsymbol{\theta}_j^* \in M$ la distancia entre $\boldsymbol{\theta}$ y el mínimo $\boldsymbol{\theta}_j^*$ y sea $\rho(\boldsymbol{\theta}) = \min \rho_j(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}^*\|$ con $\boldsymbol{\theta}_{\boldsymbol{\theta}}^* = \boldsymbol{\theta}_j^*$ para algún j el mínimo más cercano a $\boldsymbol{\theta}$. Entonces, existe $\delta > 0$, $\tau < \infty$ tal que:

$$\begin{array}{l} 1)\ \mathbf{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \text{ es definida positiva} \\ 2)\ \|\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})\| < \delta \Rightarrow \rho(\boldsymbol{\theta}) < \tau \|\nabla R(\boldsymbol{\theta})\| \end{array}$$

Esta última condición relaciona el tamaño del gradiente con la distancia de $\boldsymbol{\theta}$ a su mínimo más próximo y, *grosso modo* está inspirada en que, si se desarrolla en Taylor $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) = \mathbf{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{\boldsymbol{\theta}}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}^*) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}^*\|)$ con $\boldsymbol{\theta}_{\boldsymbol{\theta}}^*$ el mínimo más cercano a $\boldsymbol{\theta}$, entonces, por la convexidad de Θ y por el teorema del valor medio aplicado a cada componente $g_i = [\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})]_i$, $\exists \bar{\boldsymbol{\theta}}_i \in \Theta$ en el segmento que une $\boldsymbol{\theta}$ y $\boldsymbol{\theta}_j^*$ tal que $g_i = \mathbf{h}_i(\bar{\boldsymbol{\theta}}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}^*)$, donde $\mathbf{h}_i(\bar{\boldsymbol{\theta}}_i)$ es la i-ésima fila de $\mathbf{H}_{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}_i)$, $i = 1 : p$.

Además, sea $\bar{\mathbf{H}}$ es la matriz con filas $\mathbf{h}_i(\bar{\boldsymbol{\theta}}_i)$ $i = 1 : p$, entonces, si $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}^*\|$ es suficientemente pequeño, $\mathbf{H}_{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}_i)$ es definida positiva por la continuidad uniforme de las segundas derivadas. Con ello, $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) = \bar{\mathbf{H}}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}^*) \Rightarrow (\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}^*) = \bar{\mathbf{H}}^{-1} \nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})$, y así, si $\lambda_{\text{mín}}, \lambda_{\text{máx}}$ son los autovalores mínimo y máximo de $\bar{\mathbf{H}}$ se tiene:

$$\frac{1}{\lambda_{\text{máx}}^2} \|\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})\|^2 \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}^*\| \leq \frac{1}{\lambda_{\text{mín}}^2} \|\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})\|^2$$

relacionando el tamaño del gradiente con la proximidad entre $\boldsymbol{\theta}$ y su mínimo más próximo.

Dado que el análisis se concentra en los algoritmos que se basan en determinar los puntos en los que $\nabla_{\boldsymbol{\theta}} \hat{R}_n(\boldsymbol{\theta}) = \mathbf{0}$, y aportan puntos próximos a dichos puntos, interesa aplicar la teoría estadística del aprendizaje para acotar la discrepancia entre $\nabla_{\boldsymbol{\theta}} \hat{R}_n(\boldsymbol{\theta})$ y $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})$, en lugar de, como es tradicional, entre $R(\boldsymbol{\theta})$ y $\hat{R}_n(\boldsymbol{\theta})$.

Hipótesis 5. Dimensión VC finita para la familia de las derivadas. Sean $\dot{\ell}_{\theta_j}(\cdot; \boldsymbol{\theta}) = \partial \ell(\cdot; \boldsymbol{\theta}) / \partial \theta_j$ que están acotadas uniformemente por algún B_j, $j = 1 : p$. Entonces, las familias de funciones de $(\mathbf{x}, y)$ indicadoras:

$$\mathcal{G}_j = \{1_{\{\dot{\ell}_{\theta_j}(\cdot; \boldsymbol{\theta}) > a\}} : \boldsymbol{\theta} \in \Theta, |a| \leq B_j\}, j = 1 : p \tag{E.1}$$

satisfacen $\text{VCdim}(\mathcal{G}_j) < \infty$.

Para demostrar el siguiente teorema de consistencia, se necesita el siguiente resultado:

Teorema 208 *[Talagrand 1994] Sea $\xi(\mathbf{z}; \boldsymbol{\theta})$ no negativa y uniformemente acotada por una constante b. Sea la familia:*

$$\mathcal{J} = \{1_{\{\xi(\cdot; \boldsymbol{\theta}) > a\}} : \boldsymbol{\theta} \in \Theta, a \in [0, b]$$

tal que $\text{VCdim}(\mathcal{J}) = v < \infty$. Si $\mathbf{z}^n$ es una muestra aleatoria simple extraída de P_{Z}, entonces:

$$\mathbb{P}_{\mathsf{Z}}^{\infty}\{\sup_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \xi(\mathbf{z}_i; \boldsymbol{\theta}) - \mathbb{E}_{\mathsf{Z}}[\xi(\mathsf{Z}; \boldsymbol{\theta})] \right| > b\varepsilon\} \leq \frac{C}{\varepsilon \sqrt{n}} (\frac{Cn\varepsilon^2}{v})^v e^{-2nv} = \eta_n$$

En el teorema anterior, la cota η_n depende de n y ε a través de $n\varepsilon^2$. Entonces, se puede elegir ε_n tendiendo a cero cuando $n \to \infty$ (p. ej. $0 < \mu < 1/2$, $\varepsilon_n = n^{\mu - 1/2}$) tal que no sólo η_n tiende a cero sino también $\sum_{n=1}^{\infty} \eta_n < \infty$. Con ello, por el lema de Borel-Cantelli (Anexo G), se concluye que:

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \xi(\mathbf{z}_i; \boldsymbol{\theta}) - \mathbb{E}_{\mathsf{Z}}[\xi(\mathsf{Z}; \boldsymbol{\theta})] \right| \underset{n \to \infty}{\to} 0 \text{ c.s.}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Teorema 209 *[Fine y Mukherjee 1999, th. 2] En las hipótesis 1 a 5, el estimador $\hat{\boldsymbol{\theta}}_n$ proporcionado por un algoritmo converge casi seguro a su mínimo más cercano $\boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n} \in M$, es decir:*

$$\mathbb{P}^\infty_{\mathsf{Z}}\{\lim_{n\to\infty} \rho(\hat{\boldsymbol{\theta}}_n) = 0\} = \mathbb{P}^\infty_{\mathsf{Z}}\{\lim_{n\to\infty} \left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n}\right\| = 0\} = 1$$

Demostración. Aplicando el teorema 208 a las familias $\mathcal{G}_j$ definidas en (E.1), se tiene que para ε_n, η_n tendiendo a cero cuando $n \to \infty$:

$$\mathbb{P}^\infty_{\mathsf{Z}}\{\sup_{\boldsymbol{\theta}\in\Theta} \left|\frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_{\theta_j}(\mathbf{z}_i;\boldsymbol{\theta}) - \mathbb{E}_{\mathsf{Z}}[\dot{\ell}_{\theta_j}(\mathsf{Z};\boldsymbol{\theta})]\right| > B_j\varepsilon_n\} = \mathbb{P}^\infty_{\mathsf{Z}}\{\sup_{\boldsymbol{\theta}\in\Theta} \left|\frac{\partial \hat{R}_n}{\partial\theta_j}(\boldsymbol{\theta}) - \mathbb{E}_{\mathsf{Z}}[\frac{\partial R}{\partial\theta_j}(\boldsymbol{\theta})]\right| > B_j\varepsilon_n\}$$

pero tomando $B = \max_j B_j$ se tiene:

$$\mathbb{P}^\infty_{\mathsf{Z}}\{\sup_{\boldsymbol{\theta}\in\Theta} \left\|\nabla_{\boldsymbol{\theta}}\hat{R}_n(\boldsymbol{\theta}) - \mathbb{E}_{\mathsf{Z}}[\nabla_{\boldsymbol{\theta}}R(\boldsymbol{\theta})]\right\| > B\varepsilon_n\} \leq \sum_{j=1}^{p} \mathbb{P}^\infty_{\mathsf{Z}}\{\sup_{\boldsymbol{\theta}\in\Theta} \left|\frac{\partial \hat{R}_n}{\partial\theta_j}(\boldsymbol{\theta}) - \mathbb{E}_{\mathsf{Z}}[\frac{\partial R}{\partial\theta_j}(\boldsymbol{\theta})]\right| > B_j\varepsilon_n\} \leq \eta_n$$

resultado que indica la consistencia del proceso de aprendizaje de los gradientes.
Combinando la Hipótesis 2.1 con la ecuación anterior y utilizando la desigualdad triangular, se obtiene con probabilidad al menos $1-\eta_n$:

$$\left\|\nabla_{\boldsymbol{\theta}}\hat{R}_n(\hat{\boldsymbol{\theta}}_n)\right\| \leq \left\|\nabla_{\boldsymbol{\theta}}\hat{R}_n(\hat{\boldsymbol{\theta}}_n) - \mathbb{E}_{\mathsf{Z}}[\nabla_{\boldsymbol{\theta}}R(\boldsymbol{\theta})]\right\| + \left\|\nabla_{\boldsymbol{\theta}}\hat{R}_n(\hat{\boldsymbol{\theta}}_n)\right\| \leq B\varepsilon_n + \delta_n$$

con lo que se satisface la condición $\left\|\nabla_{\boldsymbol{\theta}}\hat{R}_n(\hat{\boldsymbol{\theta}}_n)\right\| < \delta$ de la Hipótesis 4.2. De esta forma, tan pronto como el algoritmo entre en un entorno de un mínimo, entonces con probabilidad al menos $1-\eta_n$:

$$\rho(\hat{\boldsymbol{\theta}}_n) = \left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n}\right\| < \rho(B\varepsilon_n + \delta_n)$$

Si en la expresión anterior, se elige ε_n tendiendo a cero tal que $\sum_{n=1}^{\infty}\eta_n < \infty$, se tiene el resultado. ■

El resultado anterior es válido para algoritmos con inicialización aleatoria, pues no se especifica a qué óptimo converge la solución y, de hecho, en cada entrenamiento con un n diferente, es de esperar que dicho óptimo sea diferente.

Asimismo, [Fine 1999] obtienen, entre otros, los siguientes resultados adicionales que, a diferencia de otros trabajos anteriores (p. ej. [White 1989], [Amari y Murata 1993]), consideran la existencia de distintos posibles mínimos:

1. La tasa de convergencia [Fine 1999, th. 4] es de orden:

   $$\left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n}\right\| = O_p(1/\sqrt{n})$$

   donde, dada una sucesión $\{\mathsf{A}_n\}$ de variables aleatorias y una sucesión $\{a_n\} \subset \mathbb{R}^+$, el orden O_p se define mediante la condición:

   $$\forall\varepsilon > 0, \exists\delta(\varepsilon) \in \mathbb{R} : \mathbb{P}[\left|\frac{\mathsf{A}_n}{a_n}\right| \geq \delta] < \varepsilon, \forall n$$

2. La distribución de $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n})$ condicionada a que $\boldsymbol{\theta}^*_{\hat{\boldsymbol{\theta}}_n}$ sea un óptimo particular $\boldsymbol{\theta}^*_j \in M$, es [Fine 1999, th. 5] asintóticamente normal con media $\mathbf{0}$ y matriz de covarianzas:

   $$\mathbf{V} = \mathbf{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*_j)^{-1}\mathbf{C}_j\mathbf{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*_j)^{-1}$$
   $$\text{con: } \mathbf{C} = 4\mathbb{E}_{\mathsf{Z}}[\ell(\mathsf{Y}, f_{\boldsymbol{\theta}^*_j}(\mathsf{X}))\nabla f_{\boldsymbol{\theta}^*_j}\nabla f^T_{\boldsymbol{\theta}^*_j}]$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Anexo F

Boosting como Algoritmo de Máxima Verosimilitud

La gran aportación para la comprensión del *boosting* desde el punto de vista estadístico vino dada por [Friedman *et al.* 2000]. El resultado más relevante fué el descubrimiento de que los algoritmos *boosting* construían con una filosofía *greedy* un modelo aditivo de regresión logística de la forma:

$$\frac{1}{2}\ln\frac{\mathbb{P}(\mathsf{Y}=1|\mathbf{x})}{1-\mathbb{P}(\mathsf{Y}=1|\mathbf{x})}=\sum_{t=1}^{B}\psi_t(\mathbf{x})$$

que es proporcional en $1/2$ al modelo logístico aditivo usual.

En lo que sigue, salvo indicación en contrario, la codificación es $\mathcal{Y}=\{-1,1\}$. El siguiente lema es de utilidad en adelante:

Lema 210 *[Friedman* et al. *2000, lema 1] Sea la pérdida $\ell(y,f(\mathbf{x}))=e^{-yf(\mathbf{x})}$. La función que minimiza $R(f)=\mathbb{E}_{\mathsf{Z}}[e^{-\mathsf{Y}f(\mathsf{X})}]$ es:*

$$\mathsf{f}_e(\mathbf{x})=\frac{1}{2}\ln\frac{\mathbb{P}(\mathsf{Y}=1|\mathbf{x})}{1-\mathbb{P}(\mathsf{Y}=1|\mathbf{x})}$$

Demostración. Considerando cada $\mathbf{x}$:

$$\begin{aligned}\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f(\mathbf{x})}] &= \mathbb{P}(\mathsf{Y}=1|\mathbf{x})e^{-f(\mathbf{x})}+\mathbb{P}(\mathsf{Y}=-1|\mathbf{x})e^{f(\mathbf{x})}\\ &\Rightarrow\ 0=\frac{\partial\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f(\mathbf{x})}]}{\partial f(\mathbf{x})}=-\mathbb{P}(\mathsf{Y}=1|\mathbf{x})e^{-f(\mathbf{x})}+\mathbb{P}(\mathsf{Y}=-1|\mathbf{x})e^{f(\mathbf{x})}\\ &\Rightarrow\ \frac{\mathbb{P}(\mathsf{Y}=1|\mathbf{x})}{\mathbb{P}(\mathsf{Y}=-1|\mathbf{x})}=e^{2f(\mathbf{x})}\Rightarrow f(\mathbf{x})=\frac{1}{2}\ln\frac{\mathsf{P}(\mathsf{Y}=1|\mathbf{x})}{1-\mathsf{P}(\mathsf{Y}=1|\mathbf{x})}\end{aligned}$$

■

A continuación, reproducimos la demostración de algunos resultados de [Friedman *et al.* 2000]. En ellos resulta especialmente importante el concepto de **distribución condicionada de $\mathsf{Y}/\mathbf{x}$ ponderada** mediante alguna función de ponderaciones $d(\mathbf{x},y)$, que da lugar a la variable aleatoria $\mathsf{Y}_d/\mathbf{x}$, definida mediante la función de probabilidad (o densidad):

$$\mathsf{p}_{\mathsf{Y}_d/\mathbf{x}}(y)=\frac{d(\mathbf{x},y)}{\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[d(\mathbf{x},\mathsf{Y})]}\mathsf{p}_{\mathsf{Y}/\mathbf{x}}(y) \tag{F.1}$$

donde el denominador $\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[d(\mathbf{x},\mathsf{Y}]$ hace el papel de factor normalizador de tal forma que, $\mathsf{p}_{\mathsf{Y}_d/\mathbf{x}}$ es una función de probabilidad:

$$\sum_{y\in\mathcal{Y}}\mathsf{p}_{\mathsf{Y}_d/\mathbf{x}}(y)=\sum_{y\in\mathcal{Y}}\frac{d(\mathbf{x},y)}{\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[d(\mathbf{x},\mathsf{Y})]}\mathsf{p}_{\mathsf{Y}/\mathbf{x}}(y)=\frac{1}{\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[d(\mathbf{x},\mathsf{Y})]}\sum_{y\in\mathcal{Y}}d(\mathbf{x},y)\mathsf{p}_{\mathsf{Y}/\mathbf{x}}(y)=1$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Con ello, la esperanza de alguna función $a(\mathbf{x}, y)$ bajo esta distribución sería:

$$\mathbb{E}_{\mathsf{Y}_d/\mathbf{x}}[a(\mathbf{x},\mathsf{Y})] = \frac{\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[d(\mathbf{x},\mathsf{Y})a(\mathbf{x},\mathsf{Y})]}{\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[d(\mathbf{x},\mathsf{Y})]} \tag{F.2}$$

Teorema 211 *[Friedman* et al. *2000, rtdo. 1] En su versión poblacional, el algoritmo AdaBoost (discreto) construye un modelo aditivo de regresión logística mediante la minimización greedy de* $R(f) = \mathbb{E}_{\mathsf{Z}}[e^{-\mathsf{Y}f(\mathsf{X})}]$ *incorporando en cada iteración nuevos términos* ψ_t, α_t *a la solución* $f_B(\mathbf{x}) = \sum_{t=1}^{B} \alpha_t \psi_t(\mathbf{x})$ *mediante iteraciones similares a las del algoritmo de Newton.*

Además, en cada iteración, después de la actualización de las ponderaciones, el error de clasificación de la nueva hipótesis débil sobre la distribución ponderada de la muestra es del $50\,\%$.

Demostración. Se considera AdaBoost con $\mathcal{Y} = \{-1, 1\}$ e hipótesis base $\psi(\mathbf{x}) \in \{-1, 1\}$. Antes de iniciar la iteración t-ésima se considera la hipótesis combinada $f_{t-1}(\mathbf{x}) = \sum_{s=1}^{t-1} \alpha_s \psi_s(\mathbf{x})$ y se persigue una mejora del tipo: $f_{t-1} + \alpha_t \psi_t$. Si $\ell(y, f(\mathbf{x}))$ es la pérdida exponencial, y se expande $\ell(f_{t-1}(\mathbf{x}) + \alpha\psi(\mathbf{x}))$ con α y $\mathbf{x}$ fijos mediante Taylor en $\psi(\mathbf{x}) = 0$ se tiene:

$$\begin{aligned} e^{-y[f_{t-1}(\mathbf{x})+\alpha\psi(\mathbf{x})]} &\approx e^{-yf_{t-1}(\mathbf{x})}(1 - y\alpha\psi(\mathbf{x}) + \alpha^2 y^2 \psi(\mathbf{x})^2/2) \\ &= e^{-yf_{t-1}(\mathbf{x})}(1 - y\alpha\psi(\mathbf{x}) + \alpha^2/2) \end{aligned}$$

Para cada $\mathbf{x}$, el óptimo $\psi(\mathbf{x}) \in \{-1, 1\}$ se obtiene mediante:

$$\psi_t(\mathbf{x}) = \arg\min_{\psi} \mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f_{t-1}(\mathbf{x})}(1 - \mathsf{Y}\alpha\psi(\mathbf{x}) + \alpha^2/2)] \tag{F.3}$$

que es una esperanza condicionada a $\mathbf{x}$ pero ponderada por pesos $d^{(t)}(\mathbf{x}, y) = e^{-yf_{t-1}(\mathbf{x})}$. Dividiendo la expresión a minimizar por $\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[d^{(t)}(\mathsf{Y}, \mathbf{x})]$, se obtiene una media ponderada como en (F.2), con lo que la ecuación (F.3) es equivalente a (α es fijo):

$$\psi_t(\mathbf{x}) = \arg\max_{\psi} \mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[\mathsf{Y}\psi_t(\mathbf{x})]$$

donde $\mathsf{Y}_{d^t}/\mathbf{x}$ se define según (F.1).
La solución es $\psi(\mathbf{x}) = 1$ si $\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}(\mathsf{Y}) = \mathbb{P}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}(\mathsf{Y} = 1) - \mathbb{P}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}(\mathsf{Y} = -1) \geq 0$ y cero en caso contrario donde $\mathbb{P}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}$ es la correspondiente probabilidad condicionada a $\mathbf{x}$ y ponderada por $d^{(t)}(\mathbf{x}, y) = e^{-yf_{t-1}(\mathbf{x})}$.
Si se tiene en cuenta que, dado que $\psi(\mathbf{x})^2 = y^2 = 1$:

$$-\mathbb{E}_{\mathsf{Y}_{d^t}/\mathbf{x}}[\mathsf{Y}\psi_t(\mathbf{x})] = \mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[\mathsf{Y} - \psi(\mathbf{x})]^2/2 - 1$$

entonces, la elección de $\psi_t(\mathbf{x})$ en (F.3), equivale a una regresión por mínimos cuadrados ponderados mediante $d^{(t)}(\mathbf{x}, y) = e^{-yf_{t-1}(\mathbf{x})}$.
Finalmente, dado el óptimo ψ_t, teniendo en cuenta que $e^{-y[f_{t-1}(\mathbf{x})+\alpha\psi(\mathbf{x})]} = e^{-yf_{t-1}(\mathbf{x})}e^{-y\alpha\psi(\mathbf{x})}$ se determina α_t mediante:

$$\begin{aligned} \alpha_t &= \arg\min_{\alpha}\{R(f_{t-1} + \alpha\psi_t) = \mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[e^{-\alpha\mathsf{Y}\psi_t(\mathbf{x})}]\} \\ &= \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t} \text{ con } \varepsilon_t = \mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[1_{\{\mathsf{Y}\neq\psi_t(\mathbf{x})\}}] \end{aligned}$$

donde se ha utilizado el lema 210, y donde ε_t puede ser negativo si el error del clasificador débil es mayor que el $50\,\%$, en cuyo caso, se invierte su polaridad. Con todo ello, la iteración t-ésima produce:

$$\begin{aligned} f_t(\mathbf{x}) &= f_{t-1}(\mathbf{x}) + \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}\psi_t(\mathbf{x}) \\ d^{(t+1)}(\mathbf{x}, y) &= d^{(t)}(\mathbf{x}, y)e^{-\alpha_t y\psi_t(\mathbf{x})} \end{aligned}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Por último, para comprobar que la ponderación resultado de cada iteración, es tal, que provoca un error del $50\,\%$ en la nueva hipótesis débil, basta observar que:

$$\begin{aligned} 0 &= \frac{\partial R(f_{t-1}+\alpha\psi)}{\partial\alpha}(\alpha_t) = -\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[e^{-\alpha\mathsf{Y}\psi_t(\mathbf{x})}\mathsf{Y}\psi(\mathbf{x})] \\ &\Rightarrow \mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[\mathsf{Y}\psi(\mathbf{x})] = 0 \end{aligned}$$

■

Un resultado similar se obtiene para el algoritmo Real AdaBoost:

Teorema 212 *[Friedman* et al. *2000, rtdo. 2] En su versión poblacional, el algoritmo Real AdaBoost construye un modelo aditivo de regresión logística mediante la minimización greedy aproximada de* $R(f) = \mathbb{E}_{\mathsf{Z}}[e^{-\mathsf{Y}f(\mathsf{X})}]$.

Además, en el óptimo f^**, la media condicionada ponderada de* Y *es cero:* $\mathbb{E}_{\mathsf{Y}_{d^*}/\mathbf{x}}(\mathsf{Y}) = 0$.

Demostración. Dada la hipótesis $f_{t-1}(\mathbf{x})$, la incorporación de $\psi_t(\mathbf{x})$ en la iteración t-ésima se produce de la siguiente forma. En primer lugar, la expresión a minimizar en $\psi \in \mathcal{F}$, es:

$$\begin{aligned} R(f_{t-1}(\mathbf{x})+\psi(\mathbf{x})) &= \mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f_{t-1}(\mathbf{x})}e^{-\mathsf{Y}\psi(\mathbf{x})}] \\ &= e^{-\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f_{t-1}(\mathbf{x})}1_{\{\mathsf{Y}/\mathbf{x}=1\}}] + e^{\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f_{t-1}(\mathbf{x})}1_{\{\mathsf{Y}/\mathbf{x}=-1\}}] \\ &= e^{-\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\frac{e^{-\mathsf{Y}f_{t-1}(\mathbf{x})}}{\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f_{t-1}(\mathbf{x})}]}1_{\{\mathsf{Y}/\mathbf{x}=1\}}] + e^{\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[\frac{e^{-\mathsf{Y}f_{t-1}(\mathbf{x})}}{\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f_{t-1}(\mathbf{x})}]}1_{\{\mathsf{Y}/\mathbf{x}=-1\}}] \\ &\equiv e^{-\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[1_{\{\mathsf{Y}/\mathbf{x}=1\}}] + e^{\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[1_{\{\mathsf{Y}/\mathbf{x}=-1\}}] \end{aligned}$$

donde, sin alterar el óptimo del problema, hemos dividido por $\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f_{t-1}(\mathbf{x})}]$ para obtener obtener esperanzas condicionadas ponderadas. Haciendo la derivada cero, se obtiene:

$$\begin{aligned} -\psi(\mathbf{x})e^{-\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[1_{\{\mathsf{Y}/\mathbf{x}=1\}}] + \psi(\mathbf{x})e^{\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[1_{\{\mathsf{Y}/\mathbf{x}=-1\}}] &= 0 \\ -e^{-\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[1_{\{\mathsf{Y}/\mathbf{x}=1\}}] + e^{\psi(\mathbf{x})}\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[1_{\{\mathsf{Y}/\mathbf{x}=-1\}}] &= 0 \end{aligned}$$

con lo que:

$$\psi_t(\mathbf{x}) = \frac{1}{2}\ln\frac{\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[1_{\{\mathsf{Y}/\mathbf{x}=1\}}]}{\mathbb{E}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[1_{\{\mathsf{Y}/\mathbf{x}=-1\}}]} = \frac{1}{2}\ln\frac{\mathbb{P}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[\mathsf{Y}=1]}{1-\mathbb{P}_{\mathsf{Y}_{d^{(t)}}/\mathbf{x}}[\mathsf{Y}=1]}$$

Después de ello, las nuevas ponderaciones se actualizan mediante:

$$d^{t+1}(\mathbf{x},y) = e^{-y(f_t(\mathbf{x})+\psi_t(\mathbf{x}))} = d^{(t)}(\mathbf{x},y)e^{-y\psi_t(\mathbf{x})}$$

Por otra parte, si f^* es el óptimo, se tiene:

$$0 = \frac{\partial R}{\partial f(\mathbf{x})}(f^*(\mathbf{x})) = -\mathbb{E}_{\mathsf{Y}/\mathbf{x}}[e^{-\mathsf{Y}f^*(\mathbf{x})}\mathsf{Y}] = \mathbb{E}_{\mathsf{Y}_{d^*}/\mathbf{x}}(\mathsf{Y})$$

■

Como puede observarse, este algoritmo poblacional finalizaría en una iteración, pero la utilización de algoritmos débiles produce aproximaciones pobres de las probabilidades a posteriori, con lo que se requieren muchas iteraciones.

El hecho de que, $E_{\mathsf{Y}_{d^*}/\mathbf{x}}[\mathsf{Y}] = 0$, indica que, en el óptimo f^*, la distribución de $\mathsf{Y}/\mathbf{x}$ ponderada mediante $d^*(\mathbf{x},\mathsf{Y})$ no aporta información adicional alguna, con lo que las ponderaciones $d^t(\mathbf{x},\mathsf{Y})$ juegan el mismo papel en el problema de clasificación que los residuos de la pérdida cuadrática en el problema de regresión.

USAL UNIVERSIDAD DE SALAMANCA DE CO

Algoritmo LogitBoost [Friedman *et al.* 2000]
0. $B \in \mathbb{N}^+$, $\mathbf{d}^{(0)} \in [0,1]^n$ con $d_i^{(0)} = 1/n$, $i = 1:n$ $\pi(\mathbf{x}_i) = \mathbb{P}(\mathsf{Y}^* = 1 \mid \mathbf{x}_i) = \frac{1}{2}$, $i = 1:n$ con $\mathsf{Y}^* = (\mathsf{Y}+1)/2 \in \{0,1\}$ 1. Para $t = 1, ..., B$ 1.1 Calcular para $i = 1, ..., n$: $r_i = \dfrac{y_i^* - \pi(\mathbf{x}_i)}{\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)]}$; $d_i^{(t)} = \pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)]$ 1.2 Determinar: $\psi_t = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} d_i^{(t)} [r_i - \psi(\mathbf{x}_i)]^2$ 1.3 Actualizar para $i = 1, ..., n$: $f_t(\mathbf{x}_i) = f_{t-1}(\mathbf{x}_i) + \frac{1}{2}\psi_t(\mathbf{x}_i)$; $\pi(\mathbf{x}_i) = \dfrac{1}{1+\exp[-f_t(\mathbf{x}_i)]}$ 2. Hipótesis final: $g_B(\mathbf{x}) = \text{signo}[f_B(\mathbf{x})]$ con $f_B(\mathbf{x}) = \sum_{t=1}^{B} \psi_t(\mathbf{x})$

Tabla 46. Algoritmo LogitBoost

[Friedman *et al.* 2000] analizan las posibles ventajas de la pérdida $\ell(y, f(\mathbf{x})) = e^{-yf(\mathbf{x})}$ en relación con otras pérdidas usuales como la del logaritmo de la verosimilitud binomial o la pérdida cuadrática (figura 1).

Si $y^* = \frac{y+1}{2} \in \{0,1\}$ y denotamos[154] $\mathbb{P}(\mathsf{Y}^* = 1|\mathbf{x}) = \pi(\mathbf{x}) = \frac{1}{1-e^{-2f(\mathbf{x})}}$, entonces, la pérdida originada por el logaritmo de la verosimilitud binomial es la pérdida de entropía cruzada o pérdida logística (ecuación (I.5), pág. 15):

$$\ell_1(y^*, f(\mathbf{x})) = -[y^* \ln \pi(\mathbf{x}) + (1-y^*)\ln(1-\pi(\mathbf{x}))] = \ln(1+e^{-2yf(\mathbf{x})})$$

cuya función óptima coincide con la de la pérdida exponencial. Sin embargo, esta pérdida no procede de ninguna verosimilitud pues no es el logaritmo de ninguna función de masa de probabilidad en ± 1.

En base a ello, [Friedman *et al.* 2000] proponen un nuevo algoritmo *boosting* que denominan **LogitBoost** (tabla 46), utilizando la misma filosofía *greedy* que los anteriores pero utilizando la pérdida logística construida mediante la verosimilitud.

Teorema 213 *[Friedman* et al. *2000, rtdo. 3] En su versión poblacional, el algoritmo LogitBoost (tabla 46) construye un modelo aditivo de regresión logística mediante la minimización greedy de la pérdida logística (maximización del logaritmo de la verosimilitud), según el algoritmo de Newton.*

Demostración. La justificación de las tareas 1.1 y 1.2 de la tabla 46 es como sigue. En la iteración t-ésima, se pretende resolver:

$$\psi_t = \arg\min_{\psi} \mathbb{E}_{\mathsf{Y}^*,\mathsf{X}}[\ell_1(\mathsf{Y}^*, f_{t-1}(\mathsf{X}) + \psi(\mathsf{X}))]$$

Teniendo en cuenta que $\pi(\mathbf{x}) = [1 + e^{-2[f_{t-1}(\mathbf{x})+\psi(\mathbf{x})]}]^{-1}$ y $1 - \pi(\mathbf{x}) = [1 + e^{2(f_{t-1}(\mathbf{x})+\psi(\mathbf{x}))}]^{-1}$,

[154] Nótese que en la página 15, esta probabilidad se denotaba por f y sus estimaciones por f, mientras que el modelo logístico se denotaba por h y sus estimaciones por h. Ahora, por conveniencia, f es la hipótesis producida por el *boosting*, con lo que $2f$ es la estimación del modelo logístico que sustituye a h en aquél lugar, y π es la estimación de la probabilidad a posteriori.

Algoritmo Gentle AdaBoost [Friedman *et al.* 2000]
0. $B \in \mathbb{N}^+$, $\mathbf{d}^{(0)} \in [0,1]^n$ con $d_i^{(0)} = 1/n$, $i = 1:n$, $f_0 = 0$.
1. Para $t = 1, ..., B$
1.1 Determinar:
$\psi_t = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} d_i^{(t)} [y_i - \psi(\mathbf{x}_i)]^2$
1.2 Actualizar para $i = 1:n$:
$f_t(\mathbf{x}_i) = f_{t-1}(\mathbf{x}_i) + \frac{1}{2}\psi_t(\mathbf{x}_i)$
$d^{(t+1)}(\mathbf{x}_i) = d^{(t)}(\mathbf{x}_i) e^{-y_i f_t(\mathbf{x}_i)}/Z_t$ con Z_t constante normalizadora
2. Hipótesis final: $g_B(\mathbf{x}) = \text{signo}[f_B(\mathbf{x})]$ con $f_B(\mathbf{x}) = \sum_{t=1}^{B} \psi_t(\mathbf{x})$

Tabla 47. Algoritmo Gentle AdaBoost

entonces, para cada $\mathbf{x}$:

$$\begin{aligned}
\ell_1(y^*, f_{t-1}(\mathbf{x}) + \psi(\mathbf{x})) &= -[y^* \ln \pi(\mathbf{x}) + (1-y^*)\ln(1-\pi(\mathbf{x}))] \\
&= -[y^* \ln(1 + e^{-2[f_{t-1}(\mathbf{x})+\psi(\mathbf{x})]}) - (1-y^*)\ln(1+e^{2(f_{t-1}(\mathbf{x})+\psi(\mathbf{x}))})] \\
&= -[y^*[\ln(1+e^{2[f_{t-1}(\mathbf{x})+\psi(\mathbf{x})]}) - \ln(1+e^{-2[f_{t-1}(\mathbf{x})+\psi(\mathbf{x})]})] \\
&\quad - \ln(1+e^{2[f_{t-1}(\mathbf{x})+\psi(\mathbf{x})]})] \\
&= -[y^* \ln \frac{1+e^{2[f_{t-1}(\mathbf{x})+\psi(\mathbf{x})]}}{1+e^{-2[f_{t-1}(\mathbf{x})+f(\mathbf{x})]}} - \ln(1+e^{2[f_{t-1}(\mathbf{x})+\psi(\mathbf{x})]})] \\
&= -[y^* \ln \frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} - \ln(1+e^{2[f_{t-1}(\mathbf{x})+\psi(\mathbf{x})]})] \\
&= -[2y^*(f_{t-1}(\mathbf{x})+\psi(\mathbf{x})) - \ln(1+e^{2[f_{t-1}(\mathbf{x})+\psi(\mathbf{x})]})]
\end{aligned}$$

Resolviendo el problema para cada $\mathbf{x}$, la primera y segunda derivadas de $\mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\ell_1(\mathsf{Y}^*, f_{t-1}(\mathbf{x})+\psi(\mathbf{x}))]$ son:

$$\begin{aligned}
s(\mathbf{x}) &= \left.\frac{\partial \mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\ell_1(\mathsf{Y}^*, f_{t-1}(\mathbf{x})+\psi(\mathbf{x}))]}{\partial \psi(\mathbf{x})}\right|_{\psi(\mathbf{x})=0} = -2\mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\mathsf{Y}^* - \pi(\mathbf{x})] \\
H(\mathbf{x}) &= \left.\frac{\partial^2 \mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\ell_1(\mathsf{Y}^*, f_{t-1}(\mathbf{x})+\psi(\mathbf{x}))]}{\partial \psi(\mathbf{x})^2}\right|_{\psi(\mathbf{x})=0} = 4\mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\pi(\mathbf{x})(1-\pi(\mathbf{x})]
\end{aligned}$$

con lo que algoritmo de Newton (D) produce la actualización:

$$\begin{aligned}
f_t(\mathbf{x}) &= f_{t-1}(\mathbf{x}) - H(\mathbf{x})^{-1} s(\mathbf{x}) \\
&= f_{t-1}(\mathbf{x}) + \frac{1}{2}\frac{\mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\mathsf{Y}^* - \pi(\mathbf{x})]}{\mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\pi(\mathbf{x})(1-\pi(\mathbf{x})]} \\
&= f_{t-1}(\mathbf{x}) + \frac{1}{2}\mathbb{E}_{\mathsf{Y}^*_{d^{(t)}}/\mathbf{x}}[\frac{\mathsf{Y}^* - \pi(\mathbf{x})}{\pi(\mathbf{x})(1-\pi(\mathbf{x})}]
\end{aligned}$$

donde se definen las ponderaciones $d^{(t)}(\mathbf{x}, y^*) = \pi(\mathbf{x})(1-\pi(\mathbf{x}))$. Pero esta iteración del método de Newton equivale a resolver la siguiente regresión por mínimos cuadrados ponderados:

$$\min_{\psi(\mathbf{x})} \mathbb{E}_{\mathsf{Y}^*_{d^{(t)}}/\mathbf{x}}\{[f_{t-1}(\mathbf{x}) + \frac{1}{2}\frac{\mathsf{Y}^* - \pi(\mathbf{x})}{\pi(\mathbf{x})(1-\pi(\mathbf{x})} - (f_{t-1}(\mathbf{x}) + \psi(\mathbf{x}))]^2\}$$

■

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Con el enfoque anterior, [Friedman *et al.* 2000] proponen una modificación al algoritmo Real AdaBoost utilizando la pérdida exponencial bajo la denominación de **Gentle AdaBoost** (tabla

47). De igual manera que en la demostración anterior, se demuestra que:

$$\begin{aligned} s(\mathbf{x}) &= \left.\frac{\partial \mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\ell(\mathsf{Y}^*, f_{t-1}(\mathbf{x}) + \psi(\mathbf{x}))]}{\partial \psi(\mathbf{x})}\right|_{\psi(\mathbf{x})=0} = -\mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\mathsf{Y} e^{-\mathsf{Y}^* f_{t-1}(\mathbf{x})}] \\ H(\mathbf{x}) &= \left.\frac{\partial^2 \mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\ell(y^*, f_{t-1}(\mathbf{x}) + f(\mathbf{x}))]}{\partial \psi(\mathbf{x})^2}\right|_{\psi(\mathbf{x})=0} = \mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[e^{-\mathsf{Y}^* f_{t-1}(\mathbf{x})}] \end{aligned}$$

con lo que la iteración del método de Newton produce:

$$\begin{aligned} f_t(\mathbf{x}) &= f_{t-1}(\mathbf{x}) - H(\mathbf{x})^{-1} s(\mathbf{x}) \\ &= f_{t-1}(\mathbf{x}) + \frac{\mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[\mathsf{Y}^* e^{-\mathsf{Y}^* f_{t-1}(\mathbf{x})}]}{\mathbb{E}_{\mathsf{Y}^*/\mathbf{x}}[e^{-\mathsf{Y}^* f_{t-1}(\mathbf{x})}]} = f_{t-1}(\mathbf{x}) + E_{\mathsf{Y}^*_{d^{(t)}}/\mathbf{x}}(\mathsf{Y}^*) \end{aligned}$$

Con ello, cada iteración del algoritmo de Newton (iteración del *boosting*) equivale a realizar una regresión ponderada por los pesos $d^{(t)}$, (tarea 1.1 de la tabla 47).

Asimismo, en su artículo, [Friedman *et al.* 2000] obtienen resultados similares para los distintos algoritmos *boosting* diseñados para el problema de clasificación con varias clases. Finalmente, proponen la técnica denominada *weigth trimming* consistente en eliminar de la muestra en cada iteración los datos con ponderaciones muy pequeñas, dado que no influyen en el entrenamiento de ψ_t y ello permite agilizar los cálculos. De cualquier forma, el recálculo de dichas ponderaciones debe realizarse de nuevo con todos los datos.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anexo G

Otros Resultados Útiles

Algebra Matricial, Diferenciación y Formas Cuadráticas

Álgebra Matricial

1. Si $\mathbf{A}$ y $\mathbf{C}$ son matrices $p \times q$ y $q \times p$ respectivamente, se verifica que:
   (a) $\mathrm{tr}(\mathbf{AC}) = \mathrm{tr}(\mathbf{CA})$.
   (b) $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}\,\mathbf{A} + \mathrm{tr}\,\mathbf{B}$.

2. Si $\mathbf{A}$ es una matriz $n \times n$, con autovalores $\{\lambda_i\}_{i=1}^n$, entonces:
   (a) $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i$.
   (b) $|\mathbf{A}| = \prod_{i=1}^n \lambda_i$.

3. **Descomposición espectral** (*Singular Value Decomposition*). [Press *et al.* 1992, p. 59] Si $\mathbf{A}$ es una matriz $n \times n$ simétrica, entonces, $\exists \mathbf{V}$ matriz ortogonal $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$ tal que $\mathbf{V}^T\mathbf{AV} = \mathbf{D}$ con $\mathbf{D}$ matriz diagonal con elementos $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ los autovalores ordenados de $\mathbf{A}$. Con ello, $\mathbf{A} = \mathbf{VDV}^T$, es decir $\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ donde $\mathbf{v}_i$ son las columnas de $\mathbf{V}$. Además:
   (a) Si $\mathbf{A}$ es semidefinida positiva, es decir, $\mathbf{x}^T\mathbf{Ax} \geq 0\ \forall \mathbf{x}$, entonces $\lambda_i \geq 0\ i = 1 : n$.
   (b) Si $\mathbf{A}$ es definida positiva, es decir, $\mathbf{x}^T\mathbf{Ax} > 0\ \forall \mathbf{x}$, entonces $\lambda_i > 0\ i = 1 : n$.

4. Si $\mathbf{X}$ es una matriz $n \times p$ de rango p, entonces $\mathbf{X}^T\mathbf{X}$ es definida positiva.

5. $\mathbf{A}$ matriz $n \times n$ es definida positiva si y sólo sí los determinantes de los menores diagonales $\mathbf{A}_m = (a_{ij})_{i,j=1,m}\ m = 1 : n$, de $\mathbf{A}$ son positivos, [Seber 1977, p. 387].

6. Si $\mathbf{A}$ es una matriz $n \times n$ definida positiva y $\mathbf{B}$ es una matriz $n \times n$ simétrica, entonces $\mathbf{A} - \lambda\mathbf{B}$ es definida positiva para $|\lambda|$ suficientemente pequeño [Seber 1977, p. 388].

7. Una matriz $\mathbf{A}^-$ es **inversa generalizada** de una matriz $\mathbf{A}$ si satisface $\mathbf{AA}^-\mathbf{A} = \mathbf{A}$. Si además, satisface: 1) $\mathbf{A}^-\mathbf{AA}^- = \mathbf{A}^-$, 2) $(\mathbf{AA}^-)^T = \mathbf{AA}^-$ y 3) $(\mathbf{A}^-\mathbf{A})^T = \mathbf{A}^-\mathbf{A}$, entonces $\mathbf{A}^-$ es única y se denomina inversa generalizada de Moore-Penrose. Se verifica que:
   (a) Si $\mathbf{A}^-$ es inversa generalizada de $\mathbf{A}$, entonces $\mathbf{w} = \mathbf{A}^-\mathbf{y}$ es solución del problema $\mathbf{Aw} = \mathbf{y}$ donde $\mathbf{A}$ es $n \times d$, $\mathbf{y} \in \mathbb{R}^n$ y $\mathbf{w} \in \mathbb{R}^d$.
   (b) En particular, si $(\mathbf{X}^T\mathbf{X})^-$ es inversa generalizada de $\mathbf{X}^T\mathbf{X}$, entonces $\mathbf{w} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$ es solución del sistema de ecuaciones normales $\mathbf{X}^T\mathbf{Xw} = \mathbf{X}^T\mathbf{y}$ [Seber 1977, p. 76].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Además, la matriz $(\mathbf{X}^T\mathbf{X})^-$ provoca que la matriz proyección $\mathbf{P_X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ sea la única matriz (simétrica e idempotente) que proyecta $\mathbf{y}$ sobre el espacio lineal $\langle\{\mathbf{x}_1, ..., \mathbf{x}_n\}\rangle$ generado por los vectores fila $\mathbf{x}_i$, $i = 1 : n$, de $\mathbf{X}$ [Seber 1977, p. 78].

(c) $\mathbf{X}$ es una inversa generalizada de $\mathbf{X}$ que verifica las condiciones 2) y 3) de Moore-Penrose sí y sólo sí puede expresarse en la forma $\mathbf{X}^- = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}$ para alguna inversa generalizada $(\mathbf{X}^T\mathbf{X})^-$ de $\mathbf{X}^T\mathbf{X}$ [Seber 1977, p. 78].

(d) Si $\mathbf{X}^-$ es la inversa generalizada de Moore-Penrose de $\mathbf{X}$, entonces $\mathbf{w}^- = \mathbf{X}^-\mathbf{y}$ es la única solución del sistema de ecuaciones normales $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$ que minimiza $\mathbf{w}^T\mathbf{w}$ [Seber 1977, p. 78].

8. **Lema de inversión de matrices**. Si las todas las inversas existen, entonces [Seber 1984, p. 519]:

$$(\mathbf{A} + \mathbf{UBV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{UB}(\mathbf{B} + \mathbf{BVA}^{-1}\mathbf{UB})^{-1}\mathbf{BVA}^{-1} \tag{G.1}$$

y, en particular, para $\mathbf{B} = 1$, $\mathbf{U} = \mathbf{u}$ y $\mathbf{V} = \mathbf{v}^T$:

$$(\mathbf{A} + \mathbf{uv}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{uv}^T}{(1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u})}$$

y además:

$$\mathbf{x}^T(\mathbf{A} + \mathbf{xx}^T)^{-1}\mathbf{x} = \frac{\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}{1 + \mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}$$

9. **Inversa de una matriz particionada**. Si todas las inversas existen, entonces [Seber y Wild 1989, p. 678]:

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21} & -\mathbf{B}_{21}\mathbf{B}_{22}^{-1} \\ -\mathbf{B}_{22}^{-1}\mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \tag{G.2}$$

donde: $\mathbf{B}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, $\mathbf{B}_{12} = \mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, $\mathbf{B}_{21} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}$, $\mathbf{C}_{11} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$, $\mathbf{C}_{12} = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}$, $\mathbf{C}_{21} = \mathbf{A}_{22}^{-1}\mathbf{A}_{21}$.

10. **Matriz proyección al añadir una nueva función básica**. Sea $\{\boldsymbol{\psi}_j\}$ un conjunto de vectores columna y sean las matrices [Orr 1996]:

$$\mathbf{A}_j = \boldsymbol{\Psi}_j^T\boldsymbol{\Psi}_j + \boldsymbol{\Lambda}_j \text{ con } \boldsymbol{\Lambda}_j = \text{diag}(\lambda_1, ..., \lambda_j), \text{ y } \boldsymbol{\Psi}_{j+1} = \begin{bmatrix} \boldsymbol{\Psi}_j & \boldsymbol{\psi}_{j+1} \end{bmatrix},\ j = 1, 2, ...$$

Entonces:

$$\mathbf{A}_{j+1} = \begin{bmatrix} \boldsymbol{\Psi}_j^T \\ \boldsymbol{\psi}_{j+1}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\Psi}_j & \boldsymbol{\psi}_{j+1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Lambda}_j & \mathbf{0} \\ \mathbf{0}^T & \lambda_{j+1} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_j & \boldsymbol{\Psi}_j^T\boldsymbol{\psi}_{j+1} \\ \boldsymbol{\psi}_{j+1}^T\boldsymbol{\Psi}_j & \lambda_{j+1} + \boldsymbol{\psi}_{j+1}^T\boldsymbol{\psi}_{j+1} \end{bmatrix}$$

Aplicando la fórmula de la inversa de una matriz particionada, se obtiene:

$$\mathbf{A}_{j+1}^{-1} = \begin{bmatrix} \mathbf{A}_j^{-1} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix} + \frac{1}{\lambda_{j+1} + \boldsymbol{\psi}_{j+1}^T\mathbf{P}_j\boldsymbol{\psi}_{j+1}} \begin{bmatrix} \mathbf{A}^{-1}\boldsymbol{\Psi}_j^T\boldsymbol{\psi}_{j+1} \\ -1 \end{bmatrix} \begin{bmatrix} \mathbf{A}^{-1}\boldsymbol{\Psi}_j^T\boldsymbol{\psi}_{j+1} \\ -1 \end{bmatrix}^T$$

donde $\mathbf{P}_j = \mathbf{I} - \boldsymbol{\Psi}_j\mathbf{A}^{-1}\boldsymbol{\Psi}_j^T$. Utilizando el lema de inversión de matrices, se obtiene:

$$\mathbf{P}_{j+1} = \mathbf{I} - \boldsymbol{\Psi}_{j+1}\mathbf{A}_{j+1}^{-1}\boldsymbol{\Psi}_{j+1}^T = \mathbf{P}_j - \frac{\mathbf{P}_j\boldsymbol{\psi}_{j+1}\boldsymbol{\psi}_{j+1}^T\mathbf{P}_j}{\lambda_{j+1} + \boldsymbol{\psi}_{j+1}^T\mathbf{P}_j\boldsymbol{\psi}_{j+1}} \tag{G.3}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

11. **Matriz proyección al eliminar una función básica**. La matriz $\mathbf{P}_j = \mathbf{I} - \mathbf{\Psi}_j \mathbf{A}^{-1} \mathbf{\Psi}_j^T$ del punto anterior es invariante ante permutaciones de las columnas de la matriz $\mathbf{\Psi}_j$ con lo cuál es indiferente la posición del vector columna que se va a eliminar, y puede suponerse en la última columna. Sea $\mathbf{P}_{j-1} = \mathbf{I} - \mathbf{\Psi}_{j-1} \mathbf{A}^{-1} \mathbf{\Psi}_{j-1}^T$ el resultado de eliminar a la matriz $\mathbf{\Psi}_j$ su última columna $\boldsymbol{\psi}_j$. Entonces, se obtiene [Orr 1996]:

$$\mathbf{P}_{j-1} = \mathbf{P}_j + \frac{\mathbf{P}_j \boldsymbol{\psi}_j \boldsymbol{\psi}_j^T \mathbf{P}_j}{\lambda_j + \boldsymbol{\psi}_j^T \mathbf{P}_j \boldsymbol{\psi}_j}$$

12. Sea la matriz:

$$\mathbf{A} = \begin{bmatrix} \mathbf{K} + \lambda \mathbf{I}_n & \mathbf{G} \\ \mathbf{G}^T & \mathbf{0}_{r \times r} \end{bmatrix} \tag{G.4}$$

con $\mathbf{K}$ matriz $n \times n$ simétrica y semidefinida positiva y $\mathbf{G}$ matriz $n \times r$. La inversa $\mathbf{A}^{-1}$ existe pues todos sus menores son definidos positivos ya que $(\mathbf{K} + \lambda \mathbf{I}_n)$ con $\lambda > 0$ es definida positiva. Entonces, aplicando la fórmula (G.2) a $\mathbf{A}$, se obtiene:

$$\begin{aligned} \mathbf{B}_{22} &= -\mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{G} \\ \mathbf{B}_{12} &= (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{G} \\ \mathbf{B}_{21} &= \mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \end{aligned}$$

con lo que que $\mathbf{A}^{-1}$ es tal que:

$$\begin{aligned} (\mathbf{A})_{11}^{-1} &= (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} [\mathbf{I}_n - \mathbf{G} [\mathbf{G}^T (K + \lambda \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}] \\ (\mathbf{A})_{21}^{-1} &= [\mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \\ (\mathbf{A})_{12}^{-1} &= (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{G} [\mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \\ (\mathbf{A})_{22}^{-1} &= -[\mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \end{aligned}$$

(a) Sea el sistema $\mathbf{A}\bar{\mathbf{c}} = \bar{\mathbf{y}}$ con $\mathbf{A}$ definida en (G.4), $\bar{\mathbf{c}} = (\mathbf{c}^T, \mathbf{b}^T)^T$ e $\bar{\mathbf{y}} = (\mathbf{y}^T, \mathbf{0}_r^T)^T$, siendo $\mathbf{y}, \mathbf{c} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^r$. Entonces, $\bar{\mathbf{c}} = \mathbf{A}^{-1} \bar{\mathbf{y}}$, siendo:

$$\mathbf{c} = (\mathbf{A})_{11}^{-1} \mathbf{y} = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} [\mathbf{I}_n - \mathbf{G} [\mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}] \mathbf{y} \tag{G.5}$$

$$\mathbf{b} = (\mathbf{A})_{21}^{-1} \mathbf{y} = [\mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \mathbf{G}^T (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{y} \tag{G.6}$$

(b) Sea el sistema $\mathbf{A}\bar{\boldsymbol{\lambda}} = \bar{\mathbf{k}}$ con $\mathbf{A}$ definida en (G.4) con $\lambda = \sigma^2 > 0$, $\bar{\boldsymbol{\lambda}} = (\boldsymbol{\lambda}^T, \boldsymbol{\beta}^T)^T$ y $\bar{\mathbf{k}} = (\mathbf{k}^T, \mathbf{g}^T)^T$, siendo $\mathbf{k}, \boldsymbol{\lambda} \in \mathbb{R}^n$, $\mathbf{g}, \boldsymbol{\beta} \in \mathbb{R}^r$, Entonces, $\bar{\boldsymbol{\lambda}} = \mathbf{A}^{-1} \bar{\mathbf{k}}$, siendo:

$$\begin{aligned} \boldsymbol{\lambda} &= (\mathbf{A})_{11}^{-1} \mathbf{k} + (\mathbf{A})_{12}^{-1} \mathbf{g} = (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} [\mathbf{I}_n - \mathbf{G} [\mathbf{G}^T (K + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \\ &\quad \mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}] \mathbf{k} + (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G} [\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \mathbf{g} \end{aligned} \tag{G.7}$$

$$\begin{aligned} \boldsymbol{\beta} &= (\mathbf{A})_{21}^{-1} \mathbf{k} + (\mathbf{A})_{22}^{-1} \mathbf{g} = [\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{k} \\ &\quad -[\mathbf{G}^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{G}]^{-1} \mathbf{g} \end{aligned} \tag{G.8}$$

(c) Sea el sistema $\mathbf{A}\bar{\mathbf{\Lambda}} = \bar{\mathbf{K}}$ con $\mathbf{A}$ definido en (G.4), de la forma:

$$\mathbf{A}(\bar{\boldsymbol{\lambda}}_1, ..., \bar{\boldsymbol{\lambda}}_n) = (\bar{\mathbf{k}}_1, ..., \bar{\mathbf{k}}_n)$$

con $\bar{\mathbf{\Lambda}} = (\bar{\boldsymbol{\lambda}}_1, ..., \bar{\boldsymbol{\lambda}}_n)$ y $\bar{\mathbf{K}} = (\bar{\mathbf{k}}_1, ..., \bar{\mathbf{k}}_n)$, donde $\bar{\boldsymbol{\lambda}}_i = (\boldsymbol{\lambda}_i^T, \boldsymbol{\beta}_i^T)^T$, $\bar{\mathbf{k}}_1 = (\mathbf{k}_i^T, \mathbf{g}_i^T)^T$, siendo $\boldsymbol{\lambda}_i, \mathbf{k}_i \in \mathbb{R}^n$ y $\mathbf{g}_i, \boldsymbol{\beta}_i \in \mathbb{R}^r$, $i = 1 : n$, o equivalentemente:

$$\mathbf{A} \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{\Omega} \end{bmatrix} = \begin{bmatrix} \mathbf{K} \\ \mathbf{G}^T \end{bmatrix}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

con $\mathbf{K}_{n\times n}$ con columnas $\mathbf{k}_i$ y $\mathbf{G}_{n\times r}$ con filas $\mathbf{g}_i^T$, $\mathbf{\Lambda}_{n\times n}$ con columnas $\boldsymbol{\lambda}_i$ y $\mathbf{\Omega}_{r\times n}$ con columnas $\boldsymbol{\beta}_i$.
Entonces, $\bar{\mathbf{\Lambda}} = \mathbf{A}^{-1}\bar{\mathbf{K}}$, donde, de igual forma que antes:

$$\begin{aligned}\mathbf{\Lambda} &= (\mathbf{A})_{11}^{-1}\mathbf{K} + (\mathbf{A})_{12}^{-1}\mathbf{G}^T = (\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}[\mathbf{I}_n - \mathbf{G}[\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G}]^{-1}\\ &\quad \mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}]\mathbf{K} + (\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G}[\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G}]^{-1}\mathbf{G}^T\\ \mathbf{\Omega} &= (\mathbf{A})_{21}^{-1}\mathbf{K} + (\mathbf{A})_{22}^{-1}\mathbf{G}^T = [\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G}]^{-1}\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{K}\\ &\quad -[\mathbf{G}^T(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{G}]^{-1}\mathbf{G}^T\end{aligned}$$

En el caso particular en el que $\sigma^2 = 0$ y $\mathbf{K}$ sea definida positiva, se obtiene:

$$\begin{aligned}\mathbf{\Lambda} &= (\mathbf{A})_{11}^{-1}\mathbf{K} + (\mathbf{A})_{12}^{-1}\mathbf{G}^T\\ &= \mathbf{K}^{-1}[\mathbf{I}_n - \mathbf{G}[\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{K}^{-1}]\mathbf{K} + \mathbf{K}^{-1}\mathbf{G}[\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G}]^{-1}\mathbf{G}^T = \mathbf{I}_n\\ \mathbf{\Omega} &= (\mathbf{A})_{21}^{-1}\mathbf{K} + (\mathbf{A})_{22}^{-1}\mathbf{G}^T\\ &= [\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{K}^{-1}\mathbf{K} - [\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G}]^{-1}\mathbf{G}^T = \mathbf{0}_{r\times n}\end{aligned}$$

Con ello, si $\sigma^2 = 0$, el kriging es interpolador de los datos.

13. [Seber 1984, p. 526] Sean $\mathbf{A}, \mathbf{C}$ matrices $d \times d$, $\mathbf{A}$ simétrica y $\mathbf{C}$ definida positiva. Sean $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_d$ los autovalores de $\mathbf{C}^{-1}\mathbf{A}$ con correspondientes autovectores $\mathbf{v}_1, ..., \mathbf{v}_d$. Entonces:

$$\sup_{\mathbf{w}}\{\frac{\mathbf{w}^T\mathbf{A}\mathbf{w}}{\mathbf{w}^T\mathbf{C}\mathbf{w}}\} = \gamma_1 \tag{G.9}$$

donde el supremo se alcanza para $\mathbf{w} = \mathbf{v}_1$.

Diferenciación Vectorial y Matricial

Sea $L = L(\boldsymbol{\theta}) \in \mathbb{R}$ y sean: $\mathbf{f} = \mathbf{f}(\boldsymbol{\theta})$, $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{w})$, con $\mathbf{f} \in \mathbb{R}^n$, $\boldsymbol{\theta} \in \mathbb{R}^p$ y $\mathbf{w} \in \mathbb{R}^h$, con componentes f_i, θ_i, w_i, respectivamente. En lo que sigue, todos los vectores son vectores columna y seguimos la siguiente notación:

$$1)\ \dot{L}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \doteq \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_1}, ..., \frac{\partial L(\boldsymbol{\theta})}{\partial \theta_p}\right)^T = \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)^T$$

$$2)\ \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \theta_j} \doteq \left(\frac{\partial f_1(\boldsymbol{\theta})}{\partial \theta_j}, ..., \frac{\partial f_n(\boldsymbol{\theta})}{\partial \theta_j}\right)^T,\ j = 1:p$$

$$3)\ \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \doteq \left(\frac{\partial \mathbf{f}}{\partial \theta_1}, ..., \frac{\partial \mathbf{f}}{\partial \theta_m}\right) = \begin{pmatrix} \frac{\partial f_1}{\partial \boldsymbol{\theta}^T} \\ \vdots \\ \frac{\partial f_n}{\partial \boldsymbol{\theta}^T} \end{pmatrix} = \left[\left(\frac{\partial f_i}{\partial \theta_j}\right)_{ij}\right]_{\substack{i=1:n \\ j=1:p}} = \left(\frac{\partial \mathbf{f}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}\right)^T$$

$$4)\ \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \doteq \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \left[\left(\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right)\right]_{i,j=1:p} \text{ donde } \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_i}\left(\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j}\right)$$

Además, con el fin de simplificar las expresiones correspondientes, siempre que no produzca ambigüedad, se realiza un cierto abuso de notación, denotando:

$$\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \doteq \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\ |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

y lo mismo para los demás casos anteriores, y derivadas de cualquier orden.

Se verifican los siguientes resultados:

1. $\frac{\partial L}{\partial \theta_r} = \sum_{j=1}^{n} \frac{\partial L}{\partial f_j}\frac{\partial f_j}{\partial \theta_r}\ r = 1:p$, es decir: $\frac{\partial L}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T}\right)^T \left(\frac{\partial L}{\partial \mathbf{f}}\right) \equiv \frac{\partial L}{\partial \boldsymbol{\theta}^T} = \frac{\partial L}{\partial \mathbf{f}^T}\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T}$

2. $\frac{\partial^2 L}{\partial \theta_r \partial \theta_s} = \frac{\partial}{\partial \theta_r}\left[\sum_{j=1}^{n} \frac{\partial L}{\partial f_j}\frac{\partial f_j}{\partial \theta_s}\right]_{\substack{j=1:n \\ s=1:p}} = \sum_{j=1}^{n}\sum_{i=1}^{n} \frac{\partial^2 L}{\partial f_i \partial f_j}\frac{\partial f_i}{\partial \theta_r}\frac{\partial f_j}{\partial \theta_s} + \sum_{j=1}^{n} \frac{\partial L}{\partial f_j}\frac{\partial^2 f_j}{\partial \theta_r \partial \theta_s}, r, s = 1:$ p, es decir: $\frac{\partial^2 L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T}\right)^T \left(\frac{\partial^2 L}{\partial \mathbf{f}\partial \mathbf{f}^T}\right)\left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T}\right) + \sum_{j=1}^{n} \frac{\partial L}{\partial f_j}\frac{\partial^2 f_j}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}$

3. $\frac{\partial \mathbf{f}}{\partial \mathbf{w}^T} = \left[\left(\frac{\partial f_r}{\partial w_s}\right)\right] = \left[\left(\sum_{j=1}^{p} \frac{\partial f_r}{\partial \theta_j}\frac{\partial \theta_j}{\partial w_s}\right)\right] = \frac{\partial f}{\partial \boldsymbol{\theta}^T}\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{w}^T}$

4. Si $\boldsymbol{\theta} = \mathbf{A}\mathbf{w}$:
   (a) Debido a 1.:
   $$\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{w}^T} = \mathbf{A} \Rightarrow \frac{\partial L}{\partial \mathbf{w}} = \mathbf{A}^T\frac{\partial L}{\partial \boldsymbol{\theta}}$$
   (b) Debido a 2.:
   $$\frac{\partial^2 L}{\partial \mathbf{w}\partial \mathbf{w}^T} = \left(\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{w}^T}\right)^T \left(\frac{\partial^2 L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right)\frac{\partial \theta}{\partial \mathbf{w}^T} = \mathbf{A}^T\frac{\partial^2 L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\mathbf{A}$$

5. Si $\mathbf{a}$, $\mathbf{A}$ son un vector y una matriz simétrica, ambos constantes, entonces:
   (a) $\frac{\partial\left(\boldsymbol{\theta}^T\mathbf{a}\right)}{\partial \boldsymbol{\theta}} = \mathbf{a}$
   (b) $\frac{\partial\left(\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = 2\mathbf{A}\boldsymbol{\theta}$

Momentos

1. **Momentos de Formas Cuadráticas**. Sean $\mathsf{X}_{n\times 1}$ un vector aleatorio y $\mathbf{A}_{n\times n}$ una matriz simétrica. Si $\mathbb{E}_{\mathsf{X}}(\mathsf{X}) = \boldsymbol{\mu}$ y $\mathrm{Var}(\mathsf{X}) = \boldsymbol{\Sigma}$, entonces:
   $$\mathbb{E}(\mathsf{X}^T\mathbf{A}\mathsf{X}) = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$$

2. **Momentos de las Distribuciones Condicionadas de una Gausiana**. Sea $\mathsf{Z} = (\mathsf{Z}_1, \mathsf{Z}_2)^T \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Gamma})$ con $\mathbf{m} = (\mathbf{m}_1^T, \mathbf{m}_2^T)^T$, y con $\mathbb{E}(\mathsf{Z}_i) = \mathbf{m}_i, \mathrm{Var}(\mathsf{Z}_i) = \boldsymbol{\Gamma}_{ii}, i = 1, 2$, y $\mathrm{Cov}(\mathsf{Z}_1, \mathsf{Z}_2) = \boldsymbol{\Gamma}_{12} = \boldsymbol{\Gamma}_{21}^T = \mathrm{Cov}(\mathsf{Z}_2, \mathsf{Z}_1)$. Entonces (p. ej. [Pratt *et al.* 1995, sec. 22.3.1]):
   $$\mathbb{E}(\mathsf{Z}_i/\mathbf{z}_j) = \mathbf{m}_i + \boldsymbol{\Gamma}_{ij}\boldsymbol{\Gamma}_{jj}^{-1}(\mathbf{z}_j - \mathbf{m}_j),\ i, j = 1, 2 \quad \text{(G.10)}$$
   $$\mathrm{Var}(\mathsf{Z}_i/\mathbf{z}_j) = \boldsymbol{\Gamma}_{ii} - \boldsymbol{\Gamma}_{ij}\boldsymbol{\Gamma}_{jj}^{-1}\boldsymbol{\Gamma}_{ji},\ i, j = 1, 2 \quad \text{(G.11)}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Cálculo de Probabilidades

Lema 214 *[Vidyasagar 1997, lema 2.5]* ***Borel-Cantelli****. Sea $(\mathcal{Z}, \mathcal{A}, \mathbb{P})$ un espacio probabilístico y sea $\{B_n\}$ una sucesión de sucesos $B_n \in \mathcal{A}$. Sea $B = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} B_m$. Entonces,*

$$\sum_{n=1}^{\infty} \mathbb{P}(B_n) < \infty \Rightarrow \mathbb{P}(B) = 0$$

Utilizando el lema de Borel-Cantelli, se demuestra la siguiente condición suficiente para la convergencia casi seguro.

Lema 215 *[Vidyasagar 1997, lema 2.6] Sea $(\mathcal{Z}, \mathcal{A}, \mathbb{P})$ un espacio probabilístico, $\{R_n\}_{n\geq 1}$ un proceso estocástico definido en dicho espacio y $R : \mathcal{Z} \rightarrow \mathbb{R}$ una función medible. Entonces, $\{R_n\}$ converge casi seguro a R si se verifica que:*

$$\sum_{n=1}^{\infty} \mathbb{P}\{z \in \mathcal{Z} : |R_n(z) - R(z)| > \varepsilon\} < \infty, \ \forall \varepsilon > 0$$

Capacidad de las Redes Neuronales

Este anexo tiene por objeto presentar algunos resultados utilizados en el texto, sobre la capacidad de las redes neuronales. Referencias al respecto son, por ejemplo [Vidyasagar 1997] y [Anthony y Bartlett 1999].

Redes Neuronales con Salida Heaviside

Perceptrón-Adaline

Teorema 216 *[Anthony 1997, prop. 4.1]* ***Perceptrón sin término independiente*** *(bias). Sea $\mathcal{F}$ la familia de funciones de la forma $f(\mathbf{x}; \mathbf{w}) = \psi(\mathbf{w}^T\mathbf{x})$, con $\mathbf{w} \in \mathbb{R}^d$, donde ψ es la función de Heaviside. Un conjunto $S = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbb{R}^d$ puede ser diseccionado por funciones de $\mathcal{F}$ si y sólo sí, los vectores de S son linealmente independientes.*

Teorema 217 *[Anthony 1997, th. 4.2] Sea $\mathcal{F}'$ un espacio vectorial de funciones reales y sea $\mathcal{F} = \{\psi \circ f : f \in \mathcal{F}'\}$. Entonces:*

$$\mathrm{VCdim}(\mathcal{F}) = \dim(\mathcal{F}')$$

Teorema 218 *[Anthony 1997, th. 4.3] Sea $\mathcal{F}$ la familia de funciones de la forma $g(\mathbf{x}) = \psi(\mathbf{w}^T\mathbf{x} + w_0)$ con $\mathbf{w} \in \mathbb{R}^d$ y $w_0 \in \mathbb{R}$, donde ψ es la función de Heaviside. Entonces:*

$$\mathrm{VCdim}(\mathcal{F}) = d + 1$$

USC UNIVERSIDADE DE COMPOSTELA

En realidad, se verifica incluso un resultado más general:

Teorema 219 *Sea $\mathcal{F}'$ espacio vectorial de funciones con valores reales y sea f_0 una función real. Sea $\mathcal{F} = \{\psi \circ (f + g) : f \in \mathcal{F}'\}$. Entonces:*

$$\mathrm{VCdim}(\mathcal{F}) = \dim(\mathcal{F}')$$

Otras Redes con Salida Heaviside

Teorema 220 *[Anthony y Bartlett 1999, th. 8.13]* ***Activaciones Sigmoide-Heaviside.*** *Sea $\mathcal{F}$ una familia de funciones implementadas por una red neuronal con p parámetros y m neuronas, de las cuáles, las ocultas poseen activación sigmoide y la de salida Heaviside. Entonces:*

$$\mathrm{VCdim}(\mathcal{F}) \leq (pm)^2 + 11pm \log_2(18pm^2)$$

Teorema 221 *[Anthony y Bartlett 1999, th. 8.14]* ***Cualquier Activación.*** *Sea $f : \mathbb{R}^d \times \mathbb{R}^p \to \{0,1\}$ y $\mathcal{F} = \{f(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^p\}$. Si dado $(\mathbf{x}, \boldsymbol{\theta})$, $f(\mathbf{x};\boldsymbol{\theta})$ puede ser determinada por un algoritmo después de t operaciones de tipo exponencial ($\exp\{\cdot\}$), aritmético ($+, -, \times, \div$) o lógico ($>, \geq, <, \leq, =, \neq$), más la operación de producir la salida 0 ó 1, entonces:*

$$\mathrm{VCdim}(\mathcal{F}) \leq t^2 d[d + 19 \log_2(9d)]$$

Además, si no hay más que r operaciones exponenciales, entonces:

$$\mathrm{VCdim}(\mathcal{F}) \leq [d(r+1)]^2 + 11d(r+1)[t + \log_2(9d(r+1))]$$

Redes Neuronales con Salida Real

Con familias de funciones reales, es necesario utilizar otras medidas de complejidad que generalizan la dimensión VC a este tipo de familias. La dimensión P (Pdim) coincide con la dimensión VC en el caso de funciones indicadoras.

El siguiente teorema generaliza el concepto algebraico de dimensión de un espacio vectorial de funciones reales:

Teorema 222 *[Haussler 1992, th. 4] Sea $\mathcal{F}$ una clase de funciones con valores reales. Se verifica que:*

$$\mathrm{Pdim}(\mathcal{F}) = \dim(\mathcal{F})$$

Corolario 223 ***Neurona lineal.*** *Sea $\mathcal{F}$ una familia de funciones de la forma $f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + w_0$ con $\mathcal{X} = \mathbb{R}^d$ o $\mathcal{X} = \{0,1\}^d$. Entonces:*

$$\mathrm{Pdim}(\mathcal{F}) = d + 1$$

El siguiente teorema proporciona una técnica útil para calcular la dimensión P de redes neuronales con salida real a partir de la dimensión VC de una familia de funciones indicadoras.

Teorema 224 *[Vidyasagar 1997, lema 10.1] Sea $\mathcal{F}$ una familia de funciones definidas en $\mathbb{R}^d$, implementadas por una arquitectura, verificando $f(\mathbf{x};\boldsymbol{\theta}) \in [0,1]$, $\forall \mathbf{x} \in \mathbb{R}^d$, $\forall \boldsymbol{\theta} \in \mathbb{R}^p$. Supóngase que a partir de dicha arquitectura se construye una nueva resultado de añadir una nueva variable de entrada $x_0 \in [0,1]$ y una nueva capa en salida compuesta por un perceptrón al que se*

USE

conecta la nueva entrada y la salida de la arquitectura original $f(\mathbf{x};\boldsymbol{\theta})$, produciendo como salida $\psi[f(\mathbf{x};\boldsymbol{\theta})-x_0]\in\{0,1\}$ donde ψ es la función de Heaviside.

Sea $\mathcal{G}$ la familia de funciones obtenida con la nueva arquitectura. Entonces:

$$\mathrm{Pdim}(\mathcal{F})=\mathrm{VCdim}(\mathcal{G})$$

Teorema 225 *[Anthony y Bartlett 1999, th. 14.2]* ***MLF con activaciones sigmoides****. Sea $\mathcal{F}$ la familia de funciones implementada por una arquitectura MLF con p parámetros y m neuronas todas ellas con activación sigmoide. Entonces:*

$$\mathrm{Pdim}(\mathcal{F})\leq(pm)^2+11pm\log_2(18pm^2)$$

Regularización

Demostración del Teorema 37

Sea el problema de minimización del riesgo empírico regularizado:

$$\min_{f\in\mathcal{H}_P}\hat{\mathcal{R}}_n(f)=\min_{f\in\mathcal{H}_P}\{\hat{R}_n(f)+\lambda\mathcal{E}(f)\}\equiv\min_{f\in\mathcal{H}_P}\{\|\mathbf{y}-\mathbf{f}\|_2^2+\lambda\|Pf\|_{L_2}^2\}\tag{G.12}$$

donde P es un operador diferencial lineal definido en un espacio de Hilbert de funciones $\mathcal{H}_P\subset L_2(\mathcal{X})$ definidas en $\mathcal{X}\subset\mathbb{R}^d$ y con valores en $\mathbb{R}$, compuesto por todas las funciones compatibles con P y sus condiciones de contorno, y cuya imagen es un espacio de funciones en el que el producto interno $\langle Pf,Pg\rangle$ de $L_2(\mathcal{X})$ está bien definido.

Se verifica el siguiente teorema [Haykin 1999, sec. 5.5]:

Teorema 226 *[Poggio y Girosi 1990] La solución al problema (G.12) posee la forma general:*

$$f(\mathbf{x})=\sum_{i=1}^{n}c_ik(\mathbf{x},\mathbf{x}_i)+g(\mathbf{x})$$

donde k es la función de Green (Anexo B) del operador diferencial P^+P y g es una función perteneciente al espacio nulo del mismo.

Demostración. Sea la diferencial de Frèchet (p. ej. [Ioffe y Tihomirov 1979, sec. 0.2]) del funcional $\hat{\mathcal{R}}_n(f)$ en la dirección $\xi\in\mathcal{H}_P$ definida como:

$$d_\xi\hat{\mathcal{R}}_n(f)=\frac{d}{dt}\hat{\mathcal{R}}_n(f+t\xi)\mid_{t=0}$$

Entonces, la condición necesaria de óptimo del problema es $\forall\xi\in\mathcal{H}_P$:

$$d_\xi\hat{\mathcal{R}}_n(f)=d_\xi\hat{R}_n(f)+\lambda d_\xi\mathcal{E}(f)=0$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Las diferenciales de ambos términos son:

$$\begin{aligned} d_\xi \hat{R}_n(f) &= \frac{d}{dt}\hat{R}_n(f+t\xi)\mid_{t=0} = \frac{d}{dt}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i) - t\xi(\mathbf{x}_i))^2 \mid_{t=0} \\ &= -\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))\xi(\mathbf{x}_i) = -\left\langle \xi, \sum_{i=1}^{n}(y_i - f)\delta_{\mathbf{x}_i}\right\rangle \\ d_\xi \mathcal{E}(f) &= \frac{d}{dt}\mathcal{E}(f+t\xi)\mid_{t=0} = \frac{d}{dt}\int_{\mathcal{X}}(P(f+t\xi))^2 d\mathbf{x}\mid_{t=0} = \int_{\mathcal{X}} P(f+t\xi)P\xi d\mathbf{x}\mid_{t=0} \\ &= \int_{\mathcal{X}} Pf P\xi d\mathbf{x} = \langle P\xi, Pf\rangle = \left\langle \xi, P^+Pf\right\rangle \end{aligned}$$

donde P^+ es el operador adjunto de P y $\delta_{\mathbf{x}_i}$ es la función generalizada delta de Dirac (p. ej. [Kolmogorov y Fomin 1978, p. 137 y 219-220]) centrada en $\mathbf{x}_i$, $\delta_{\mathbf{x}_i} = \delta(\mathbf{x} - \mathbf{x}_i)$ para la que utilizamos el producto interno de L_2 como una extensión notacional: $\langle \xi, \delta_{\mathbf{x}}\rangle = \int_{\mathcal{X}} \xi(\mathbf{x}')\delta_{\mathbf{x}}(\mathbf{x}')d\mathbf{x}' = \xi(\mathbf{x})$.
Con ello, la condición de óptimo resulta:

$$\begin{aligned} 0 &= d_\xi \hat{\mathcal{R}}_n(f) = \frac{1}{\lambda} d_\xi \hat{R}_n(f) + d_\xi \mathcal{E}(f) = \left\langle \xi, P^+Pf - \frac{1}{\lambda}\sum_{i=1}^{n}(y_i - f)\delta_{\mathbf{x}_i}\right\rangle \\ \Leftrightarrow\quad & P^+Pf = \frac{1}{\lambda}\sum_{i=1}^{n}(y_i - f)\delta_{\mathbf{x}_i} \end{aligned}$$

que es la ecuación de Euler-Lagrange del problema variacional (p. ej. [Ioffe y Tihomirov 1979, sec. 0.2], [Elsgoltz 1977]). Dicha condición, es una ecuación diferencial en derivadas parciales, $P^+Pf = \xi$ cuyas soluciones pueden obtenerse mediante la función de Green $k(\mathbf{x}, \mathbf{x}')$ del operador P^+P, solución de la ecuación diferencial $P^+Pk(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}')$ en el sentido de las distribuciones (p. ej. [Kolmogorov y Fomin 1978, p. 137 y 219]), mediante el operador integral inverso:

$$\begin{aligned} f(\mathbf{x}) &= \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}')\xi(\mathbf{x}')d\mathbf{x}' + g(\mathbf{x}) \\ &= \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}')[\frac{1}{\lambda}\sum_{i=1}^{n}(y_i - f(\mathbf{x}'))]\delta(\mathbf{x}' - \mathbf{x}_i)d\mathbf{x}' + g(\mathbf{x}) \\ &= \frac{1}{\lambda}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))k(\mathbf{x}, \mathbf{x}_i) + g(\mathbf{x}) \\ &= \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i) + g(\mathbf{x}) \end{aligned}$$

donde $c_i = \frac{1}{\lambda}(y_i - f(\mathbf{x}_i))$, $i = 1 : n$ y g es una función perteneciente al espacio nulo de P^+P.

■

Regresión *Ridge* en Términos de la Matriz Gram

Sea el problema regularizado de regresión lineal (regresión *ridge*):

$$\min_{\bar{\mathbf{w}}} \hat{\mathcal{R}}_n(f_{\bar{\mathbf{w}}}) = \min_{\mathbf{w},b}\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})^2 + \frac{\lambda}{2}\bar{\mathbf{w}}^T\bar{\mathbf{w}}\} \tag{G.13}$$

con $\bar{\mathbf{w}} = (b, \mathbf{w}^T)^T$, $\bar{\mathbf{x}} = (1, \mathbf{x}^T)^T$ y $f_{\bar{\mathbf{w}}}(\mathbf{x}) = \bar{\mathbf{x}}^T\bar{\mathbf{w}} = \mathbf{x}^T\mathbf{w} + b$.

La condición de óptimo es de la forma:

$$\frac{\partial \hat{\mathcal{R}}_n}{\partial \bar{\mathbf{w}}} = -\sum_{i=1}^{n}(y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})\bar{\mathbf{x}}_i + \lambda\bar{\mathbf{w}} = \mathbf{0} \Rightarrow \bar{\mathbf{w}} = \frac{1}{\lambda}\sum_{i=1}^{n}(y_i - \bar{\mathbf{x}}_i^T\bar{\mathbf{w}})\bar{\mathbf{x}}_i = \sum_{i=1}^{n}\beta_i\bar{\mathbf{x}}_i$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

con $\beta_i = \frac{1}{\lambda}(y_i - \bar{\mathbf{x}}_i^T \bar{\mathbf{w}})$, lo que equivale a $\bar{\mathbf{w}} = \bar{\mathbf{X}}^T \boldsymbol{\beta}$ donde $(\bar{\mathbf{X}})_{ij} = (\bar{\mathbf{x}}_i)_j$, $i = 1:n$, $j = 1:d$. Sustituyendo dicha expresión en (G.13), se obtiene una función objetivo en términos de nuevas incógnitas $\boldsymbol{\beta}$:

$$\begin{aligned} \hat{\mathcal{R}}_n(\boldsymbol{\beta}) &= \hat{\mathcal{R}}_n(f_{\bar{\mathbf{w}}}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{\mathbf{x}}_i^T \bar{\mathbf{X}}^T \boldsymbol{\beta})^2 + \frac{\lambda}{2}\boldsymbol{\beta}^T \bar{\mathbf{X}}\bar{\mathbf{X}}^T \boldsymbol{\beta} \\ &= \frac{1}{2}(\mathbf{y} - \mathbf{K}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{K}\boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^T \mathbf{K}\boldsymbol{\beta} \end{aligned}$$

donde hemos denotado $\mathbf{K} = \bar{\mathbf{X}}\bar{\mathbf{X}}^T = \mathbf{K}^T$ a la matriz de productos internos cruzados entre los puntos de la muestra $(\mathbf{K})_{ij} = \bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_j = \langle \bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j \rangle$ $i, j = 1:n$, que se denomina **matriz Gram** .

La condición de óptimo en términos de $\boldsymbol{\beta}$ produce:

$$\begin{aligned} \frac{\partial \hat{\mathcal{R}}_n}{\partial \boldsymbol{\beta}} &= \mathbf{K}\mathbf{K}\boldsymbol{\beta} - \mathbf{K}\mathbf{y} + \lambda\mathbf{K}\boldsymbol{\beta} = \mathbf{K}(\mathbf{K}\boldsymbol{\beta} + \lambda\boldsymbol{\beta} - \mathbf{y}) = \mathbf{0} \\ \Rightarrow \quad & (\mathbf{K} + \lambda\mathbf{I}_n)\boldsymbol{\beta} = \mathbf{y} \Rightarrow \boldsymbol{\beta} = (\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{y} \end{aligned}$$

Con ello, la solución es:

$$f_{\bar{\mathbf{w}}}(\mathbf{x}) = \bar{\mathbf{x}}^T \bar{\mathbf{w}} = \bar{\mathbf{x}}^T \bar{\mathbf{X}}^T (\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{y} = \mathbf{k}_{\bar{\mathbf{x}}}(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{y}$$

donde hemos denotado $\mathbf{k}_{\bar{\mathbf{x}}} \in \mathbb{R}^n$ con $(\mathbf{k}_{\bar{\mathbf{x}}})_i = \bar{\mathbf{x}}^T \bar{\mathbf{x}}_i = \langle \bar{\mathbf{x}}, \bar{\mathbf{x}}_i \rangle$, $i = 1:n$.

Lo anterior pone de manifiesto las ventajas de poseer un método rápido de cálculo del producto interno en el espacio de entrada ya que permite obtener rápidamente la solución $\bar{\mathbf{w}}$ y la predicción en un nuevo punto $\bar{\mathbf{x}}$. Ello motiva el método del núcleo (*Kernel Trick*).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anexo H

Notación e Índice de Símbolos

A continuación se relacionan los símbolos y acrónimos utilizados con carácter general ordenados con un criterio pseudo-alfabético. Para cada símbolo, se aporta una breve descripción y, al final de la misma, se indica el número de la página donde el símbolo se define.

0-1. Pérdida $\ell(y, g(\mathbf{x})) = 1_{\{y \neq g(\mathbf{x})\}}$, 13
$\mathbf{0}_p = (0, ..., 0, ..., 0) \in \mathbb{R}^p$, 36
$\mathbf{1}_d = (1, ..., 1)^T \in \mathbb{R}^d$, 91
1_S. Función indicadora del conjunto S: $1_S(a) = 1$ si $a \in S$ y cero en caso contrario, 13
$1 : n$ Conjunto de valores $\{1, ..., n\}$ que toma una variable índice. Así, $i = 1 : n$ significa $i = 1, ..., n$, 27
$\{1 : n\} = \{1, ..., n\}$, 21
$\{a_i\}_{i=1}^n$. Conjunto $\{a_1, ..., a_n\}$, 21
$\langle\{\mathbf{x}_1, ..., \mathbf{x}_n\}\rangle = \{\sum_{i=1}^n \lambda_i \mathbf{x}_i : \lambda_i \in \mathbb{R}, i = 1 : n\}$ Espacio lineal generado por $\{\mathbf{x}_i\}_{i=1}^n$, 83

$\doteq$ Por definición, 21
$\equiv$ Equivalencia. Si $a \equiv b$ donde a y b son expresiones matemáticas, ambas juegan el mismo papel en el contexto. Por ejemplo, se utiliza para equivalencias de notación, 11
$\propto$ Proporcional. $a \propto b \Leftrightarrow \exists c \in \mathbb{R}$ constante, tal que $a = cb$, 27
$\stackrel{d}{=}$ Igualdad en distribución, 34
$\xrightarrow{d}$ Convergencia en distribución. $\mathsf{X}_n \xrightarrow{d} \mathsf{X}$ si $\lim_{n\to\infty} \mathbb{P}(\mathsf{X}_n \leq \mathbf{x}) = \mathbb{P}(\mathsf{X} \leq \mathbf{x}) = \mathsf{P}_{\mathsf{X}}(\mathbf{x})$ en cada punto $\mathbf{x}$ en que P_{X} es continua, 36
$\circ$ Composición de funciones: $f \circ g(\mathbf{x}) = f[g(\mathbf{x})]$, 416
$*$ Operación convolución definida por $(\phi * f)(\mathbf{x}) = \int f(\mathbf{x} - \mathbf{t})\phi(\mathbf{t})d\mathbf{t}$, 370
$\circledast_{j=1}^s = * \overset{s \text{ veces}}{\cdots} *$ donde $*$ es el operador convolución, 146
$\oplus$ Suma directa, 340
$\otimes$ Producto directo, 340
$[\cdot]_+$. $[u]_+ = u$ si $u \geq 0$ y 0 en caso contrario, 142
$\#A$. Cardinal del conjunto A, 20
$\setminus$ Operación diferencia de conjuntos. $A \setminus B = \{u \in A : u \notin B\}$, 337
$\langle \cdot, \cdot \rangle$. Producto interno definido en un espacio vectorial, 81
$|\cdot|$ Si $\mathbf{A}$ es una matriz $n \times n$, $|\mathbf{A}|$ es su determinante. Si $a \in \mathbb{R}$ o $\mathbb{C}$, $|a|$ es su valor absoluto o su módulo, respectivamente, 86
$\|\cdot\|_{\mathcal{F}}$. Norma de un espacio normado $\mathcal{F}$, 55
$\|\cdot\|_p$ Norma del espacio ℓ_p con $p \in [1, \infty]$. Si está claro en el contexto, la norma Euclídea se denota por $\|\cdot\| \equiv \|\cdot\|_2$, 359
$\|f\|_{L_p} \equiv \|f\|_{L_p(\mathcal{X},\mu)}$. Norma del espacio $L_p(\mathcal{X}, \mu)$, 359
$(\cdot)^T$ Traspuesta de una matriz. S^T es la traspuesta de S, 34
$(\cdot)^-$ Si $\mathbf{A}$ es una matriz, $\mathbf{A}^-$ es una inversa generalizada de $\mathbf{A}$, es decir que satisface: $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$, 83
$(\cdot)^+$ Si P es un operador, P^+ es el operador adjunto de P, 341
$\partial f(\boldsymbol{\theta})/\partial\theta_j$ Componente j-ésima del gradiente $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$, 414

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$\partial f(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = (\partial f(\mathbf{x}, \boldsymbol{\theta}_0)/\partial \theta_1, ..., \partial f(\mathbf{x}, \boldsymbol{\theta}_0)/\partial \theta_p)^T$ Gradiente de f respecto de $\boldsymbol{\theta}$. También se denota por $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$ o $\dot{\mathbf{f}}$, 414
$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0) = \partial f(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$. Gradiente de f respecto de $\boldsymbol{\theta}$, 414
$\partial^2 f(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^2 = \partial^2 f(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ Hesiano de f con $(\partial^2 f(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^2)_{ij} = \partial^2 f/\partial \theta_i \partial \theta_j$. Puede denotarse también por $\ddot{\mathbf{F}}$, 414
$\int_{\mathcal{X}} f(\mathbf{x}) d\mathsf{P}(\mathbf{x})$, 359

$A : \mathcal{D} \rightarrow \Theta$. En un problema de aprendizaje, algoritmo de entrenamiento del dispositivo, donde $\mathcal{D} = \cup_{n=1}^{\infty} \mathcal{Z}^n$ es el conjunto de posibles muestras extraídas de Z y Θ es el espacio paramétrico que indexa las hipótesis., 10, 54
$\mathbb{A} : \cup_{d=1}^{\infty} \cup_{n=1}^{\infty} \mathcal{Z}_d^n \rightarrow \cup_{d=1}^{\infty} \mathcal{F}_d$. Clase de algoritmos (también denominado algoritmo) en la teoría de la complejidad computacional, 23
$\mathcal{A}_{\mathsf{X}}$. σ-álgebra del espacio de probabilidad $(\mathcal{X}, \mathcal{A}_{\mathsf{X}}, \mathbb{P}_{\mathsf{X}})$ inducido por una v. a. X, 13
AIC. Criterio AIC de Akaike para la selección del modelo, 38
AIC_c. Criterio de AIC de Akaike corregido para muestras pequeñas, 38
Algoritmo PAC, 19
$\text{AMPSE}(\hat{f}_n; \mathbf{x}^n) = \sigma^2 + \text{AMSE}(\hat{f}_n; \mathbf{x}^n)$. Media del error cuadrático de predicción (*averaged mean prediction squared error*), donde σ^2 es la varianza del ruido, 34
$\text{AMSE}(\hat{f}_n; \mathbf{x}^n) = \mathbb{E}_{\mathsf{Y}^n/\mathbf{x}^n}[\text{ASE}(\hat{f}_n; \mathbf{x}^n)]$. Media del error cuadrático medio (*averaged mean squared error*) en el conjunto de puntos $\mathbf{x}^n$, donde $\hat{f}_n$ es un estimador de la regresión, 34
ARCH, 300
ARD. Media de los errores relativos en valor absoluto, 270
$\text{ASE}(\hat{f}_n; \mathbf{x}^n) = \frac{1}{n} \sum_{i=1}^{n} [\hat{f}_n(\mathbf{x}_i) - m(\mathbf{x}_i)]^2$. Media del error cuadrático (*averaged squared error*) en los puntos de $\mathbf{x}^n$, donde m es la función de regresión y $\hat{f}_n$ un estimador, 34

Banach. Espacio lineal normado y completo, 357
$\mathbb{B}_{d,\tau} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \tau\}$. Si $\tau = 1$, se denota por $\mathbb{B}_d$, 378
BIC. Criterio BIC (*Bayesian Information Criterion*) de Schwartz de selección del modelo, 42, 48
BP. Algoritmo *Basis Pursuit* para el análisis de señales en ausencia de ruido, 209
$\mathcal{B}_p^{\alpha q}(\mathcal{X})$. Espacio de Besov, 363
BPDN. Algoritmo *Basis Pursuit Denoising* para el análisis de señales en ambiente de ruido, 210

c.d.p. Condicionalmente definida positiva, 350
$\text{cl}(\mathcal{F}) = \mathcal{T}$. Conjunto clausura. $\mathcal{T}$ es la clausura de $\mathcal{F}$, $\mathcal{T} = \text{cl}(\mathcal{F})$ si $\mathcal{F}$ es denso en $\mathcal{T}$, 358
$\text{co}(\mathcal{G})$ Conjunto de las combinaciones convexas de elementos de $\mathcal{G}$, 376
$C_0(\mathcal{X})$. Familia de las funciones continuas que *se anulan en el infinito*, es decir, $\forall \varepsilon > 0$, $\exists B \subset \mathcal{X}$ compacto tal que $|f(\mathbf{x})| < \varepsilon \; \forall \mathbf{x} \notin B$, 360
$\mathcal{C}_c(\mathcal{X})$. Espacio de funciones continuas con soporte compacto, 359
$\mathcal{C}^r(\mathcal{X})$. Espacio de funciones cuyas derivadas de orden menor que r son continuas, 362
$\mathcal{C}_p^r(\mathcal{X}) \subset \mathcal{C}^r(\mathcal{X})$, 363
$\text{Cov}_{(\mathsf{X},\mathsf{Y})}(\mathsf{X}, \mathsf{Y}) = \mathbb{E}_{(\mathsf{X},\mathsf{Y})}\{[\mathsf{X} - \mathbb{E}_{\mathsf{X}}(\mathsf{X})][\mathsf{Y} - \mathbb{E}_{\mathsf{Y}}(\mathsf{Y})]\}$. Covarianza entre las v. a. X e Y. Si se utiliza con argumentos puede omitirse el subíndice $\text{Cov}(\mathsf{X}, \mathsf{Y}) \equiv \text{Cov}_{(\mathsf{X},\mathsf{Y})}(\mathsf{X}, \mathsf{Y})$, 36
C_p. Criterio C_p de Mallows para la selección del modelo, 38
c.p.d. Casi por doquier: salvo conjunto de medida nula, 359
$\bar{\mathcal{C}}(\mathcal{X})$. Conjunto de funciones uniformemente continuas, 363
c.s. Convergencia casi seguro. $\mathsf{X}_n \xrightarrow{\text{c.s.}} \mathsf{X}$ si $\mathbb{P}[\lim_{n\to\infty} \mathsf{X}_n = \mathsf{X}] = 1$, donde X, X_n, $n \in \mathbb{N}$ son v. a., 49

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CUMEUD. Propiedad de convergencia uniforme de medias a las esperanzas, uniformemente en distribución, 19
CV. Validación cruzada (*cross validation*), 39
CV_s. *S-fold cross-validation*. La muestra se divide en s grupos de aprox. igual tamaño y el modelo se estima s veces dejando un grupo fuera cada vez, 39
$\mathcal{C}(\mathcal{X}, \mu)$. Conjunto de las funciones $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ continuas con μ medida sobre $\mathcal{X}$, 359

δ. Delta de Kronecker: $\delta(a) = 1$ si $a = 0$ y $\delta(a) = 0$ si $a \neq 0$, 8
$D^{\boldsymbol{\alpha}} f(\mathbf{x})$ Derivada de f de orden $\boldsymbol{\alpha}$, 362
$\text{diag}(\cdot)$. Si $\mathbf{A}$ es una matriz, $\text{diag}(\mathbf{A})$ es el vector compuesto por sus elementos de la diagonal. Si $\mathbf{a}$ es un vector, $\text{diag}(\mathbf{a})$ es una matriz diagonal con diagonal los elementos de $\mathbf{a}$, 117
$\dim(\mathcal{F})$. Dimensión del espacio vectorial $\mathcal{F}$, 416
d.p. Definida positiva, 345

$\mathcal{E} : \mathcal{F} \to \mathbb{R}$. Estabilizador de un problema de regularización, 12, 26
$\mathbb{E}_{\mathsf{X}}(\mathsf{X}) = \int_{\mathcal{X}} x\mathbb{P}(dx) = \int_{\mathcal{X}} x d\mathsf{P}(x)$. Esperanza de la v. a. X. Se utiliza la integral sea la variable continua o discreta. En éste último caso, representa un sumatario., 9
$\mathbb{E}_{\mathsf{Y}/\mathsf{X}}(\mathsf{Y})$. V. a. que toma valores $\mathbb{E}_{\mathsf{Y}/\mathbf{x}}(\mathsf{Y}) = \int_{\mathcal{Y}} y d\mathsf{P}_{\mathsf{Y}/\mathbf{x}}(y)$, 12
$\mathbb{E}_{\mathsf{Y}/\mathbf{x}}(\mathsf{Y}) = \int_{\mathcal{Y}} y d\mathsf{P}_{\mathsf{Y}/\mathbf{x}}(y)$. Esperanza de la v. a. $\mathsf{Y}/\mathbf{x}$, 12
e.c.d.p. Estrictamente condicionalmente definida positiva, 350
e.d.p. Estrictamente definida positiva, 345
$\mathcal{D}$. Conjunto de muestras de una v. a. Z. Puede representar todas las posibles: $\mathcal{D} = \cup_{n=1}^{\infty} \mathcal{Z}^n$, 10
$\varepsilon(\mathbf{x})$: V. a. condicionada a $\mathbf{x}$. Suele representar ruido aleatorio al que está sujeta otra v. a. $\mathsf{Y}/\mathbf{x}$, 8
ERM. *(Empirical Risk Minimization)* Principio inductivo basado en la minimización del riesgo empírico $\hat{R}_n(f)$, 9
ess sup . Supremo esencial, 359

$\mathcal{F} = \{f : \mathcal{X} \to \widehat{\mathcal{Y}}\}$. Espacio de hipótesis (*hypothesis space*) utilizado para estimar una función de interés f normalmente desconocida. Las funciones de $\mathcal{F}$ pueden estar parametrizadas mediante un vector de parámetros $\boldsymbol{\theta} \in \Theta$: $\mathcal{F} = \{f_{\boldsymbol{\theta}} = f(\cdot, \boldsymbol{\theta}) : \mathcal{X} \to \widehat{\mathcal{Y}}, \boldsymbol{\theta} \in \Theta\}$, 8
$\mathsf{f}(\mathbf{x})$. Función definida en un dominio $\mathcal{X} \subset \mathbb{R}^d$. Normalmente, es una función desconocida que se quiere estimar (*target*), 8
$f^* = \arg\inf_{f \in \mathcal{F}} R(f)$. Si $\mathcal{F}$ está parametrizada por $\boldsymbol{\theta}$ y $f^* = f(\cdot, \boldsymbol{\theta}^*)$ se denota $\boldsymbol{\theta}^* \equiv f^*$, 9
$\hat{f}_n^{(i)}$. Estimador estimado sin incluir el dato i-ésimo (LOO) o el grupo al que pertenece el dato i-ésimo (CV_s), 40
$\hat{f}_n$. Estimación de una función f en base a una muestra Z^n, 29
$\hat{f}_n$. Estimación de una función f en base a una muestra Z^n. Si $f(\mathbf{x}, \boldsymbol{\theta})$ es un modelo paramétrico, $\hat{f}_n \equiv f(\cdot, \hat{\boldsymbol{\theta}}_n)$ para algún $\hat{\boldsymbol{\theta}}_n$ estimado a partir de Z^n, 24
$\tilde{f}$ Transformada de Fourier de f, 131
$\tilde{f}$. Transformada de Fourier de f: $\tilde{f}(\boldsymbol{\omega}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(\mathbf{x}) e^{-i\boldsymbol{\omega}^T \mathbf{x}} d\mathbf{x}$ donde $f : \mathbb{R}^d \to \mathbb{R}$, 368
FPE. Criterio FPE (*final prediction error*) de Akaike para la selección del modelo, 38
$f(\cdot, \boldsymbol{\theta})$. Función del primer argumento para cada valor $\boldsymbol{\theta}$ del segundo. También puede denotarse como $f_{\boldsymbol{\theta}} \equiv f(\cdot, \boldsymbol{\theta})$, 9
$f_{\boldsymbol{\theta}} \in \mathcal{F}$. Hipótesis de un espacio de hipótesis $\mathcal{F}$ parametrizado mediante espacio de parámetros Θ, 9

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$f_{t(n)} = \arg\inf_{\mathcal{F}_{t(n)}} \|f - \mathsf{f}\|_{L_p(\mathsf{P}_Z)}$ donde $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_t \subset \cdots \subset \mathcal{F}$ es una estructura anidada de familias de hipótesis, 61

γ-admisible. Función de pérdida $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ que satisface la condición: $\exists \mu > 0 : |\ell(y, u_1) - \ell(y, u_2)| \leq \gamma |u_1 - u_2|, \forall y, u_1, u_2 \in \mathcal{Y}$, 9
GARCH, 300
GCV. Validación cruzada generalizada (*generalized cross-validation*), 40
$G_{\mathcal{F}}(n)$. *Growth function* de una familia $\mathcal{F}$ de funciones, 21
GPE. Criterio GPE (*generalized prediction error*) de Moody para la selección del modelo, 39
Gram. Matriz de productos internos cruzados entre los vectores de una muestra, 420

Haussdorf. Espacio topológico Haussdorf, 360
Heaviside. Función $\psi(r) = 1$ si $r \geq 0$ y $\psi(r) = 0$ si $r < 0$, 21, 85
$\mathcal{H}_k$ RKHS o SRKHS con núcleo k, 127

$\mathrm{IMSE}(\hat{f}_n) = \mathbb{E}_{\mathsf{X}}\{\mathrm{MSE}(\hat{f}_n(\mathsf{X})]\}$. Esperanza integrada del error cuadrático (*integrated mean squared error*), 29
$\mathbf{I}_n$. Matriz identidad de tamaño $n \times n$, 34
$\mathcal{I}(\boldsymbol{\theta})$. Matriz de información de Fisher, 36
IRF_m. Función aleatoria intrínseca de orden m (*Intrinsic random function*), 253

jackknife, 40

KIA. *Kappa index of agreement*, 296

$\ell : \mathcal{Y} \times \widetilde{\mathcal{Y}} \to \mathbb{R}$. En un problema de aprendizaje supervisado, función de pérdida tal que $\ell(y, f(\mathbf{x}))$ es el coste incurrido al estimar y mediante $f(\mathbf{x})$. En un problema no supervisado, se define como $\ell : \mathcal{X} \times \mathcal{F} \to \mathbb{R}$, 9
$L_\infty^{\mathrm{loc}}(\mathbb{R}^d)$. Espacio de las funciones $f \in L_\infty(B)\ \forall B \subset \mathbb{R}^d$ compacto, 367
LOO. Validación cruzada (*leave-one-out*). Equivale a CV_n (n-*fold cross-validation*), 40
ℓ_p, 359
$L_p(\mathcal{X}, \mu)$ con $p \in [1, \infty]$, 359
$L_p(\mathcal{X}, \mathsf{P})$, 359
LrO. Validación cruzada (*leave-r-out*). En cada entrenamiento se dejan fuera r datos de la muestra, 40

MAP. En el contexto bayesiano, estimador *maximum a posteriori*, 27
MLF. Red neuronal *multilayer feed-forward*, 73
MLP. *Multilayer perceptron*, 107
MOF. Algoritmo *Method of Frames* para el proceso de señales en ausencia de ruido, 209
MP. Algorimto *Matching Pursuit*, 200
$M_p : \Theta_p \to \mathcal{F}$. En un problema de aprendizaje, modelo o arquitectura del dispositivo, donde $\mathcal{F}_p = M_p(\Theta_p) \subset \mathcal{F}$ espacio de hipótesis y p es la dimensión de un espacio paramétrico Θ_p, 10
$\mathrm{MPSE}(\hat{f}_n) = \mathbb{E}_{\mathsf{Z}^n}[\mathrm{PSE}(\hat{f}_n)]$. Esperanza del error cuadrático de predicción (*mean prediction squared error*), 29
$\mathrm{MSE}[\hat{f}_n(\mathbf{x})] = \mathbb{E}_{\mathsf{Z}^n}\{[\hat{f}_n(\mathbf{x}) - \mathsf{f}(\mathbf{x})]^2\}$. Esperanza del error cuadrático (*mean squared error*) en $\mathbf{x} \in \mathcal{X}$, 29
μ_{L}. Medida de Lebesgue, 367

n. Tamaño muestral, 9
$N_{\mathcal{F}}(n)$. Coeficiente de disección (*shatter coefficient*). $N_{\mathcal{F}}(n) = \max_{\mathbf{z}^n \in \mathcal{Z}^n} N_{\mathcal{F}}(\mathbf{z}^n)$, 20

$N_{\mathcal{F}}(\mathbf{z}^n)$. Coeficiente de disección empírico (*empirical shatter coefficient*). Número de clasificaciones dicotómicas de la muestra $\mathbf{z}^n$ que pueden realizarse con funciones de la familia $\mathcal{F}$., 20
NIC. Criterio NIC (*network information criterion*) de Murata para la selección del modelo, 38
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 239

$O(\cdot)$. Si $\{a_n\}, \{b_n\} \subset \mathbb{R}$ son sucesiones, $a_n = O(b_n) \Leftrightarrow \lim_{n\to\infty} |a_n/b_n| < \infty$ (Función de orden de Landau), 24
$o(\cdot)$. Si $\{a_n\}, \{b_n\} \subset \mathbb{R}$ son dos sucesiones, $a_n = o(b_n)$ si $\lim_{n\to\infty} a_n/b_n = 0$, 41
OBD. *Optimal Brain Damage*, 164
OBS. *Optimal Brain Surgeon*, 164
OK. Kriging ordinario, 269
OLS. Algoritmo de mínimos cuadrados ortogonales (*orthogonal least squares*), 175
OMP. Algoritmo *Orthogonal Matching Pursuit*, 385
$O_p(\cdot)$. Dada una sucesión $\{\mathsf{A}_n\}$ de v. a. y una sucesión $\{a_n\} \subset \mathbb{R}^+$, $\mathsf{A}_n = O_p(a_n)$ si $\forall \varepsilon > 0, \exists \delta(\varepsilon) \in \mathbb{R} : \mathbb{P}[|\mathsf{A}_n/a_n| \geq \delta] < \varepsilon, \forall n$, 403

PAC. *Probably Approximately Correct*. Enfoque al problema de aprendizaje, 10
$\mathbb{P}_{\mathsf{Z}}$. Función de probabilidad de la v. a. Z. Si se utiliza con argumento, puede omitirse el subíndice, 10
$\mathbb{P}^{\infty}_{\mathsf{Z}}$. Única medida de probabilidad definida en $(\mathcal{Z}^{\infty}, \mathcal{A}^{\infty}_{\mathsf{Z}})$ con $\mathcal{A}^{\infty}_{\mathsf{Z}} = \{B = \Pi^{\infty}_{n=1} B_n : B_n \in \mathcal{Z}\}$ que verifica $\mathbb{P}^{\infty}_{\mathsf{Z}}(B) = \Pi^{\infty}_{n=1} \mathbb{P}_{\mathsf{Z}}(B_n)$ según el teorema de extensión de Kolmogorov, 19
$\mathbb{P}^n_{\mathsf{Z}}$. Sinónimo de $\mathbb{P}_{\mathsf{Z}^n}$. Función de probabilidad de la v. a. Z^n en el espacio probabilístico inducido $(\mathcal{Z}^n, \mathcal{S}^n, \mathbb{P}^n_{\mathsf{Z}})$, 19
Pdim. Dimensión P, 21
p_f. Función de masa de probabilidad o densidad definida en alguna familia de funciones f. Si se usa con argumento, puede omitirse el subíndice, 27
$\mathcal{P}$. Conjunto de medidas de probabilidad., 11
plim. Límite en probabilidad. $\operatorname{plim}_{n\to\infty} \mathsf{X}_n = \mathsf{X}$ si $\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\{|\mathsf{X}_n - \mathsf{X}| > \varepsilon\} = 0$, 18
plug-in. En un problema de clasificación con dos clases, método de construcción de reglas de decisión g_f a partir de funciones reales f que estiman la probabilidad a posteriori. $g_f(\mathbf{x}) = 0$ si $f(\mathbf{x}) \leq 1/2$ y $g_f(\mathbf{x}) = 1$ en caso contrario, 14
$\Pi_m(\mathcal{X})$. Espacio de polinomios algebraicos de grado menor que $m \in \mathbb{N}$ definidos en $\mathcal{X} \subset \mathbb{R}^d$, 129
PRESS. Suma de cuadrados de predicción (*prediction sum of squares*). LOO, 40
$\mathrm{PSE}(f) = \mathbb{E}_{\mathsf{Z}}\{[\mathsf{Y} - f(\mathsf{x})]^2\}$. Error cuadrático de predicción (*prediction squared error*), 29
ψ. En general, función de activación del nivel oculto de una red neuronal, 73
P_{X}. Función de distribución de la v. a. X. Si se utiliza con su argumento x éste identifica la variable y puede omitirse el subíndice: $\mathsf{P}(\mathbf{x}) \equiv \mathsf{P}_{\mathsf{X}}(\mathbf{x})$, 8
p_{X}. Función de masa de probabilidad o de densidad de una v. a. X. Cuando se utiliza con su argumento, la letra utilizada en éste determina la v. a. y puede no utilizarse el subíndice $\mathsf{p}(x) = \mathsf{p}_{\mathsf{X}}(x)$, 8
$\mathsf{P}_{(\mathsf{X},\mathsf{Y})}(\mathbf{x}, y)$. Función de distribución de la v. a. (X, Y). Si se utiliza con argumentos, puede omitirse el subíndice $\mathsf{P}(\mathbf{x}, y) \doteq \mathsf{P}_{(\mathsf{X},\mathsf{Y})}(\mathbf{x}, y)$, 9
$\mathsf{P}_{\mathsf{Y}/\mathbf{x}}$. Función de distribución de la v. a. $\mathsf{Y}/\mathbf{x}$, 8
$\mathsf{p}_{\mathsf{Y}/\mathbf{x}}$. Función de masa de probabilidad o densidad de la variable aleatoria $\mathsf{Y}/\mathbf{x}$. En el contexto del teorema de Bayes, suele denotarse $\mathsf{p}(y/\mathbf{x}) \doteq \mathsf{p}_{\mathsf{Y}/\mathbf{x}}(y)$, 8

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

q. Hipótesis sobre la función de densidad (o masa de probabilidad) en un problema de estimación de la densidad (juega el papel de f en un problema general de aprendizaje), 16

$R(f) = \mathbb{E}_{(\mathsf{X},\mathsf{Y})}\{\ell(\mathsf{Y}, f(\mathsf{X}))\}$. En un problema de aprendizaje, es el riesgo (o riesgo teórico como contraposición al riesgo empírico) incurrido al utilizar f como estimador de la respuesta Y, 9

Radon. Medida de, 381

$\mathrm{rango}(\mathbf{A})$ Rango de una matriz $\mathbf{A}$, es decir, la dimensión algebraica del espacio imagen de la transformación $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$, 96

RBF. Función de base radial (*radial basis function*). Por extensión, red neuronal de base radial, 114

RBFPP. Algoritmo LS_Boost con redes RBF sobre direcciones del espacio de entrada, 285

$R^* = \inf_{f\in\mathcal{F}} R(f)$, donde $\mathcal{F}$ es la familia de hipótesis considerada. Si dicha familia está parametrizada por $\boldsymbol{\theta}$, entonces $R^* = \inf_{\Theta} R(\boldsymbol{\theta})$ con $\boldsymbol{\theta} \equiv f(\cdot, \boldsymbol{\theta})$, 9

$\mathrm{R} = \inf_f R(f)$, donde el ínfimo se toma sobre todas las hipótesis factibles para el problema de aprendizaje, 9

RK. Kriging regularizado, 269

RKHS. Espacio de Hilbert con núcleo reproductor (*reproducing kernel Hilbert space*), 345

$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(\mathbf{x}_i))$. En un problema de aprendizaje, riego empírico incurrido al estimar las respuestas y_i mediante $f(\mathbf{x}_i)$, donde $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ es una muestra aleatoria simple, 9

$\hat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(\mathbf{x}_i)) + \mathcal{E}(f)$. En un problema de aprendizaje, riesgo empírico regularizado, donde ℓ es la pérdida y $\mathcal{E}$ es un estabilizador, 12

RRBF. Redes RBF de regularización, 126, 133

SEM. Algoritmo que minimiza el riesgo empírico (*sample error minimization algorithm*), 10

$\mathrm{Sesgo}_{\mathsf{Z}^n}[\hat{f}_n(\mathbf{x})]$. Sesgo del estimador $\hat{f}_n(\mathbf{x})$, 30

Sigmoide. Función de tipo sigmoide, 86

sigmoide de grado s, 376

signo . Función definida como $\mathrm{signo}(r) = 1$ si $r \geq 0$ y $\mathrm{signo}(r) = -1$ en caso contrario, 79

SK. Kriging simple, 265

SLT. Teoría Estadística del Aprendizaje (*Statistical Learning Theory*) iniciada por V. Vapnik, 18

SMS. Criterio SMS de Shibata para la selección del modelo, 38

$\mathrm{sop}(f) = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) \neq 0\}$. Soporte de una función definida en $\mathcal{X}$, 359

SRKHS. Espacio de Hilbert con seminúcleo reproductor (*semireproducing kernel Hilbert space*), 351

$\sum_{j\neq i=1}^n a_i$. Suma de los términos a_i sin incluir el término a_j, 41

SVM. *Support vector machines*, 143, 148

$\mathcal{T}$. Espacio objetivo (*target space*) de funciones al que pertenece una función de interés f normalmente desconocida, 8

$\mathcal{T}_C(\mathbb{R}^d) = \{\mathsf{f} : \mathbb{R}^d \to \mathbb{R}$, con $C_{\mathsf{f}} = \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_1 \left|\tilde{\mathsf{f}}(\boldsymbol{\omega})\right| d\boldsymbol{\omega} < \infty\}$, 378

$\mathcal{T}_{\mathcal{G}} = \mathrm{cl}(\mathrm{co}(\mathcal{G}))$ para una familia de funciones $\mathcal{G}$, 376

$\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. Vector de parámetros en un espacio paramétrico Θ, 9

$\Theta \subset \mathbb{R}^p$. Espacio de parámetros con elementos $\boldsymbol{\theta} \in \Theta$. Si es conveniente, puede incluirse el subíndice p en la notación: Θ_p, 9

$\mathrm{tr}(\mathbf{A})$. Traza de una matriz $\mathbf{A}$, 35

$\mathcal{U} \subset \mathbb{R}^h$. Espacio de características (*features*) donde toman valores las variables ocultas de una red neuronal, 73
UK. Kriging universal, 265

v. Párametro de complejidad de una familia de funciones. Usualmente, la VCdim o $\mathrm{V}_\gamma$dim, 21
v. a. Variable aleatoria o variables aleatorias, 9
$\mathrm{Var}(\mathsf{X}) = \int_{\mathcal{X}}[\mathbf{x} - \mathbb{E}_{\mathsf{X}}(\mathsf{X})]^2 d\mathsf{P}(\mathbf{x})$. Varianza de la v. a. X. Si X es vectorial, denota la matriz $\mathbf{V}$ de varianzas covarianzas con $(\mathbf{V})_{ij} = \mathrm{Cov}((\mathsf{X})_i, (\mathsf{X})_j)$, 29
VCdim. Dimensión de Vapnik-Chervonenkis de una familia de funciones, 20
$\mathrm{V}_\gamma$dim. Dimensión *fat-shattering*, 22
$\mathcal{V}_p^r(\mathcal{X}) = \{f \in \mathcal{W}_p^r(\mathcal{X}) : \|f\|_{\mathcal{W}_p^r} \leq 1\}$, 375

WN. Redes wavelet (*wavelet networks*), 102
$\mathcal{W}_p^r(\mathcal{X})$. Espacio de Sobolev, 363

X. V. a. con valores en $\mathbb{R}^d$ con $d \in \mathbb{N}^+$. Normalmente, juega el papel de variable independiente o vector de covariables, 8
$\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$. Vector de valores que puede tomar una v. a. (vectorial) X, 8
$\mathcal{X} \subset \mathbb{R}^d$. Espacio muestral del espacio de probabilidad $(\mathcal{X}, \mathcal{A}_{\mathsf{X}}, \mathbb{P}_{\mathsf{X}})$ inducido por una v. a. X, 8

Y. V. a. con valores en $\mathbb{R}^c$ con $c \in \mathbb{N}^+$ Normalmente, juega el papel de variable respuesta o dependiente, 8
$\widehat{\mathcal{Y}} \subset \mathcal{Y} \subset \mathbb{R}^c$. Rango de valores de las funciones aleatorias $f(\mathsf{X})$ con $f \in \mathcal{F}$ espacio de hipótesis, 8
$\mathsf{Y}^n/\mathbf{x}^n = (\mathsf{Y}_1/\mathbf{x}_1, ..., \mathsf{Y}_n/\mathbf{x}_n)$, 34
$\mathcal{Y} \subset \mathbb{R}^c$. Espacio muestral del espacio de probabilidad $(\mathcal{Y}, \mathcal{A}_{\mathsf{Y}}, \mathbb{P}_{\mathsf{Y}})$ inducido por una v. a. Y, 8

$\mathsf{Z} = (\mathsf{X}, \mathsf{Y})$. En un problema de aprendizaje, v. a. conjunta con componentes la entrada X y la respuesta Y, 9
$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Rango de valores de la v. a. $\mathsf{Z} = (\mathsf{X}, \mathsf{Y})$, 9
$\mathcal{Z}^n = \mathcal{Z} \times \overset{n}{\cdots} \times \mathcal{Z}$. Producto cartesiano, 10
$\mathsf{Z}^n = (\mathsf{Z}_1, ..., \mathsf{Z}_n)$. V. a. donde Z_i, $i = 1, .., n$ son v. a. independientes e idénticamente distribuidas con espacio de probabilidad inducido $(\mathcal{Z}^n, \mathcal{A}^n, \mathbb{P}_{\mathsf{Z}^n})$, 18
$\mathbf{z}^n = (\mathbf{z}_1, ..., \mathbf{z}_n) \in \mathcal{Z}^n$. Muestra de tamaño n extraída de la v. a. Z, 9

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Referencias

Abellanas L y Galindo A [1991]. *Espacios de Hilbert*, Eudema Universidad.

Ackley D, Hinton G y Sejnowski T [1985]. A Learning Algorithm for Boltzmann Machines, *Cognitive Science* **9**, 147–169.

Agterberg F P, Bonham-Carter G F y Wright D F [1990]. Statistical Pattern Integration for Mineral Exploration, *en* G Gaal y D F Merriam, eds, *Computer Applications in Resource Estimation: Prediction and Assessment for Metals and Petroleum*, Pergamon, pp. 1–21.

Agterberg F P y Bonham-Carter G F [1999]. Logistic Regression and Weights of Evidence Modeling in Mineral Exploration, *en APCOM'99, Golden, Colorado*, pp. 20–22.

Aizerman M A, Braverman E M y Rozonoer L I [1964*a*]. The Probability Problem of Pattern Recognition Learning and the Method of Potential Functions, *Automation and Remote Control* **25**, 1175–1193.

Aizerman M A, Braverman E M y Rozonoer L I [1964*b*]. Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning, *Automation and remote control* **25**, 821–837.

Akaike H [1969]. Fitting Autoregressive Models for Prediction, *Annals of the Institute of Statistical Mathematics* **21**, 243–347.

Akaike H [1970]. Statistical Predictor Identification, *Annals of the Institute of Statistical Mathematics* **22**, 203–217.

Akaike H [1973]. Information Theory and an Extension of the Maximum Likelihood Principle, *en Proceedings of the 2nd International Symposium on Information Theory*, Akademiai Kaido, Budapest.

Akaike H [1974]. A New Look at the Statistical Model Idenfication, *IEEE Transactions on Automatic Control* **19**, 716–723.

Allen D [1974]. The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction, *Technometrics* **16**, 125–127.

Alon N, Ben-David S, Cesa-Bianchi N y Haussler D [1997]. Scale-Sensitive Dimensions, Uniform Convergence, and Learnability, *Journal of the ACM* **44**(4), 615–631.

Amari S, Murata N, Müller K R, Finke M y Yang H H [1997]. Asymptotic Statistical Theory of Overtraining and Cross-Validation, *IEEE Transactions on Neural Networks* **8**, 985–996.

Amari S y Murata N [1993]. Statistical Theory of Learning Curves under Entropic Loss Criterion, *Neural Computation* **5**, 140–153.

Anders U y Korn O [1999]. Model Selection in Neural Networks, *Neural Networks* **12**, 309–323.

Anderson J A y Rosenfeld E [1988]. *Neurocomputing: Foundations of Research*, MIT Press, Cambridge.

Angluin D [1988]. Queries and Concept Learning, *Machine Learning* **2**, 319–342.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anguita D, Ridella S, Rivieccio F y Zunino R [2002]. Automatic Hyperparameters Tunning for Support Vector Machines, *en* J R Dorronsoro, ed., *Proceedings of the International Conference on Artificial Neural Networks, ICANN 2002*, Springer, pp. 1345–1350.

Anthony M [1997]. Probabilistic Analysis of Learning in Artificial Neural Networks: The PAC Model and its Variants, *Neural Computing Surveys* **1**, 1:47.

Anthony M y Bartlett P [1999]. *Neural Network Learning: Theoretical Foundations*, Cambridge University Press.

Aronszajn N [1950]. Theory of Reproducing Kernels, *Transactions of the American Mathematical Society* **68**, 337–404.

Atencia M A, Joya G y Sandoval F [2000]. A Formal Model for Definition and Simulation of Generic Neural Networks, *Neural Processing Letters* **11**, 87–105.

Audrino F y Bühlmann P [2003]. Volatility Estimation with Functional Gradient Descent for Very High Dimensional Financial Time Series, *Journal of Computational Finance* **6**, 65–89.

Avnimelech R y Intrator N [1999]. Boosting Regression Estimators, *Neural Computation* **11**, 491–513.

Bagget L W [1992]. *Functional Analysis*, Marcel Dekker.

Baillie R T, Bollerslev T y Mikelsen H O [1996]. Fractionally Integrated Generalized Autorregressive Conditional Heteroskedasticity, *Journal of Econometrics* **74**, 3–30.

Baldi P y Hornik K [1989]. Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima, *Neural Networks* **2**, 53–58.

Banerjee S y Gelfand A E [2003]. On Smoothness Properties of Spatial Processes, *Journal of Multivariate Analysis* **84**, 85–100.

Barnard G A [1963]. New Methods of Quality Control, *Journal of the Royal Statistical Society, Ser. A* **126**, 255.

Barron A R [1991]. Complexity Regularization with Application to Artificial Neural Networks, *en* G Roussas, ed., *Nonparametric Functional Estimation and Related Topics*, Kluwer Academic Publishers, pp. 561–576.

Barron A R [1993]. Universal Approximation Bounds for Superpositions of a Sigmoidal Function, *IEEE Transactions on Information Theory* **39**, 930–945.

Barron A R [1994]. Approximation and Estimation Bounds for Artificial Neural Networks, *Machine Learning* **14**, 115–133.

Barron A R y Barron R L [1988]. Statistical Learning Networks: A Unifying View, *en* E Wegman, D Gantz y J Miller, eds, *Proceedings of the 20th Symposium on the Interface*, American Statistical Association, pp. 192–203.

Barron A R y Cover T M [1991]. Minimum Complexity Density Estimation, *IEEE Transactions on Information Theory* **IT-37**, 1034–1053.

Bartlett P, Boucheron S y Lugosi G [2002]. Model Selection and Error Estimation, *Machine Learning* **48**, 85–113.

USC DE SANTIAGO DE COMPOSTELA

Battiti R [1992]. First and Second-Order Methods for Learning: Between Steepest Descent and Newton's Method, *Neural Computation* **4**, 141–166.

Baxt W G y White H [1995]. Bootstrapping Confidence Intervals for Clinical Input Variable Effects in a Network Trained to Identify the Presence of Acute Myocardial Infraction, *Neural Computation* **7**, 624–638.

Belsley D A [1991]. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, John Wiley.

Berg C, Christensen J P R y Ressel P [1984]. *Harmonic Analysis on Semigroups. Theory of Positive Definite and Related Functions*, Springer-Verlag.

Berger J O [1985]. *Statistical Decision Theory and Bayesian Analysis*, Springer, Berlin.

Bernardo J M y Smith A F M [1994]. *Bayesian Theory*, John Wiley.

Bertero M [1986]. Regularization Methods for Linear Inverse Problems, *en* G Talenti, ed., *Inverse Problems*, Springer.

Bishop C M [1995]. *Neural Networks and Pattern Recognition*, Oxford University Press.

Bollerslev T [1986]. Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics* **31**, 27–37.

Bonham-Carter G F [1994]. *Geographic Information Systems for Geoscientists. Modelling with GIS*, Pergamon.

Bonham-Carter G F, Agterberg F P y Wright D F [1988]. Integration of Geological Datasets for Gold Exploration in Nova Scotia, *Photogrammetry and Remote Sensing* **54**, 1585–1592.

Bonham-Carter G F y Agterberg F P [1999]. Arc-WofE: A GIS Tool for Statistical Integrationi of Mineral Exploration Datasets, *en Proceedings of the International Statistical Institute, Helsinki*, pp. 497–450.

Boser B E, Guyon I M y Vapnik V [1992]. A Training Algorithm for Optimal Margin Classifiers, *en* D Haussler, ed., *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, ACM Press, Pittsburgh, PA, pp. 144–152.

Bosq D [1998]. *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction*, Springer.

Bosq D [2000]. *Linear Processes in Function Spaces*, Springer.

Boubez T I [1993]. Receptive Field Partitioning for Wavelet Networks, *en* R J Mammone, ed., *Artificial Neural Networks for Speech and Vision*, Chapman and Hall, pp. 79–96.

Bourlard H y Kamp Y [1988]. Autoassociation by Multilayer Perceptrons and Singular Value Decomposition, *Biological Cybernetics* **59**, 291–294.

Bousquet G y Elisseeff A [2002]. Stability and Generalizatoin, *Journal of Machine Learning Research* **2**, 499–526.

Bousquet O y Elisseeff A [2001]. Algorithmic Stability and Generalization Performance, *en* T K Leen, T G Dietterich y V Tresp, eds, *Advances in Neural Information Processing Systems*, MIT Press.

Box G E P, Jenkins G M y Reinsel G C [1994]. *Time Series Analysis. Forecasting and Control*, Prentice-Hall.

Breiman L [1993]. Stacked Regression, Technical report, University of California, Berkeley.

Breiman L [1995]. Better Subset Regression Using the Nonnegative Garrote, *Technometrics* **37**, 373–384.

Breiman L [1996*a*]. Bagging Predictors, *Machine Learning* **24**, 123–140.

Breiman L [1996*b*]. Bias, Variance and Arcing Classifiers, Technical Report TR 460, Statistics Department, University of California, Berkeley.

Breiman L [1996*c*]. Stacked Regressions, *Machine Learning* **24**, 49–64.

Breiman L [1997*a*]. Bias-Variance, Regularization, Instability and Stabilization, *en* C M Bishop, ed., *Neural Networks and Machine Learning. Proceedings of the NATO Advanced Study Institute on Generalization in Neural Networks and Machine Learning*, Springer, pp. 27–56.

Breiman L [1997*b*]. Prediction Games and Arcing Algorithms, Technical Report TR 504, Statistics Dpt., University of California, Berkeley.

Breiman L [1998]. Arcing Classifiers, *Annals of Statistics* **26**, 801–849.

Breiman L [1999]. Prediction Games and Arcing Algorithms, *Neural Computation* **11**, 1493–1518.

Breiman L [2001]. Using Iterated Bagging to Debias Regressions, *Machine Learning* **45**, 261–277.

Breiman L, Friedman J, Olshen R y Stone C [1984]. *Classification and Regression Trees*, Wadsworth.

Bridle J S [1990]. Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationship to Statistical Pattern Recognition, *en* F Fogleman y J Hérault, eds, *Neurocomputing: Algorithms, Architectures and Applications*, Springer, pp. 227–236.

Broomhead D S y Lowe D [1988]. Multivariable functional interpolation and adaptive networks, *Complex Systems* **2**, 321–355.

Bühlmann P y McNeil A J [2002]. An Algorithm for Nonparametric GARCH Modelling, *Computational Statistics and Data Analysis* **40**, 665–683.

Bühlmann P y Yu B [2001]. Boosting with the L2-Loss: Regression and Classification, Technical Report 605, Department of Statistics. UC Berkeley.

Buja A, Hastie T y Tibshirani R [1989]. Linear Smoothers and Additive Models, *The Annals of Statistics* **17**, 453–555.

Buntine W L y Weigend A S [1991]. Bayesian Back-Propagation, *Complex Systems* **5**, 603–643.

Cacoullos T [1966]. Estimation of a Multivariate Density, *Annals of the Institute of Statistical Mathematics* **18**, 179–189.

Candes E J [1998], Ridgelets: Theory and Applications, PhD thesis, Department of Statistics, Stanford University.

Candes E J [1999]. Ridgelets: Estimating with Ridge Functions, Technical report, Department of Stastistics, Stanford University. Aceptado para publicación por The Annals of Statistics.

USC DE SANTIAGO DE COMPOSTELA

Carroll S M y Dickinson B W [1989]. Construction of Neural Nets Using the Radon Transform, *en Proceedings of the International Conference on Neural Networks*, IEEE, pp. I–607–I–611.

Chakraborty K, Mehrotra K, Mohan C K y Ranka S [1992]. Forecasting the Behaviour of Multivariate Time Series Using Neural Networks, *Neural Networks* **5**, 961–970.

Chang C C, Hsu C W y Lin C J [2000]. The Analysis of Decomposition Methods for Support Vector Machines, *IEEE Transactions on Neural Networks* **11**, 1003–1008.

Chen D y Hagan M [1999]. Optimal Use of Regularization and Cross-Validation in Neural Network Modeling, *en Proceedings of the International Joint Conference on Neural Networks*, pp. 1275–1289.

Chen H [1991]. Estimation of a Projection Pursuit Type Regression Model, *The Annals of Statistics* **19**, 142–157.

Chen S [2002*a*]. Locally Regularised Orthogonal Least Squares Algorithm for the Construction of Sparse Kernel Regression Models, *en Proceedings of the 6th International Conference on Signal Processing*, pp. 1229–1232.

Chen S [2002*b*]. Multi-Output Regression Using a Locally Regularised Orthogonal Least Squares Algorithm, *IEEE Proceedings on Vision Image and Signal Processing* **149**, 185–195.

Chen S, Billings S A y Luo W [1989]. Orthogonal Least Squares Methods and their Application to Non-Linear System Identification, *International Journal of Control* **50**, 1873–1896.

Chen S, Chng E S y Alkadhimi K [1996]. Regularization Orthogonal Least Squares Algorithm for Constructing Radial Basis Function Networks, *International Journal of Control* **64**, 829–837.

Chen S, Cowan C FÑ y Grant P M [1991]. Orthogonal Least Squares Learning for Radial Basis Function Networks, *IEEE Transactions on Neural Networks* **2**, 302–309.

Chen S S, Donoho D L y Saunders M A [1999]. Atomic Decomposition by Basis Pursuit, *SIAM Journal of Scientific Computing* **20**, 33–61.

Chen X y White H [1999]. Improved Rates and Asymptotic Normality for Nonparametric Neural Networks Estimators, *IEEE Transactions on Information Theory* **45**, 682–691.

Cherkassky V, Shao X, Mulier F M y Vapnik V [1999]. Model Complexity Control for Regression Using VC Generalization Bounds, *IEEE Transactions on Neural Networks* **10**, 1075–1088.

Cherkassky V y Mulier F [1999]. *Learning from Data*, John Wiley.

Chong E K P y Zak S H [2001]. *An Introduction to Optimization*, John Wiley.

Clemen R T [1989]. Combining Forecasts: A Review and Annotated Bibliography, *International Journal of Forecasting* **5**, 559–583.

Cleveland W [1979]. Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

Coifman R R y Vickerhauser M V [1992]. Entropy-Based Algorithms for Best Basis Selection, *IEEE Transactions on Information Theory* **38**, 713–718.

Collobert R y Bengio S [2001]. SVMTorch: Support Vector Machines for Large-Scale Regression Problems, *Journal of Machine Learning* **1**, 143–160.

Cortes C y Vapnik V [1995]. Support Vector Networks, *Machine Learning* **20**, 273–297.

Cowan J D [1990]. Neural Networks: The Early Days, *en Advances in Neural Information Processing Systems*, Vol. 2, Morgan Kauffman, pp. 828–842.

Craven P y Wahba G [1979]. Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation, *Numerische Matematik* **31**, 377–403.

Cressie N [1989]. Geostatistics, *The American Statistician* **43**, 197–202.

Cressie N [1990]. Reply to Wahba's Letter, *The American Statistician* **43**, 256–258.

Cressie N [1993]. *Statistics for Spatial Data*, John Wiley.

Cristianini N y Shawe-Taylor J [2000]. *Support Vector Machines*, Cambridge University Press.

Cucker F y Smale S [2001]. On the Mathematical Foundations of Learning, *Bulletin ofthe American Mathematical Society* **30**, 1–49.

Cun Y L, Denker J S y Solla S A [1990]. Optimal Brain Damage, *en* D S Touretzky, ed., *Advances in Neural Information Processing Systems*, Morgan Kaufmann, pp. 598–605.

Cybenko G [1989]. Approximation by Superpositions of a Sigmoidal Function, *Mathematics of Control, Signals and Systems* **2**, 303–314.

Daubechies I [1988]. Time-Frecuency Localization Operators: A Geometric Phase-Space Approach, *IEEE Transactions on Information Theory* **34**, 605–612.

Daubechies I [1992]. *Ten Lectures on Wavelets*, SIAM.

Davis G, Mallat S y Zhang Z [1994]. Adaptive Time-Frequency Decompositions, *Optical Engineering* **33**, 2183–2191.

Davis P J [1975]. *Interpolation and Approximation*, Dover.

Davison A C y Hinkley D V [1997]. *Bootstrap Methods and their Application*, Cambridge University Press.

Demuth H y Beale M [2000]. *Matlab Neural Network Toolbox User's Guide*, Math Works Inc.

Dettling M y Bühlmann P [2001]. Volatility and Risk Estimation with Linear and Nonlinear Methods Based on High Frequency Data, Technical report, Department of Statistics. UC Berkeley.

DeVore R A [1989]. Degree of Nonlinear Approximation, *en* C K Chui, L L Schumaker y J D Ward, eds, *Approximation Theory VI: Volume*, Academic Press, pp. 175–201.

DeVore R A [1998]. Nonlinear Approximation, *Acta Numerica* pp. 51–150.

DeVore R A, Howard R y Micchelli C [1989]. Optimal Nonlinear Approximation, *Manuscripta Mathematica* **63**, 469–478.

DeVore R A y Lorentz G G [1993]. *Constructive Approximation*, Springer-Verlag.

DeVore R A y Temlyakov VÑ [1996]. Some Remarks on Greedy Algorithms, *Advances in Computational Mathematics* **5**, 173–187.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Devroye L [1988]. Automatic Pattern Recognition: A Study of the Probability of Error, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**, 530–543.

Devroye L, Györfi L y Lugosi G [1996]. *A Probabilistic Theory of Pattern Recognition*, Springer.

Devroye L P y Wagner T J [1979*a*]. Distribution-Free Inequalities for the Deleted and Holdout Error Estimates, *IEEE Transactions on Information Theory* **IT-25**, 202–207.

Devroye L P y Wagner T J [1979*b*]. Distribution-Free Performance Bounds for Potential Function Rules, *IEEE Transactions on Information Theory* **IT-25**, 601–604.

Diaconis P y Freedman D [1986]. On the Consistence of Bayes Estimates, *The Annals of Statistics* **14**, 1–67.

Diaconis P y Shahshahani M [1984]. On Nonlinear Functions of Linear Combinations, *SIAM J. Sci. Statist. Comput.* **5**, 175–191.

Diamantaras K I y Kung S Y [1996]. *Principal Component Neural Networks*, John Wiley.

Dingankar A T y Sandberg I W [1996]. A Note on Error Bounds for Approximation in Inner Product Spaces, *Circuits, Systems and Signal Processing* **15**, 519–522.

Docampo D, Hush D R y Abdallah C T [1997]. Constructive Function Approximation: Theory and Practice, *en Intelligent Methods in Signal Processing and Communications*, Birkhäuser, pp. 199–220.

Domingos P [2000]. A Unified Bias-Variance Decomposition and its Applications, *en Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, pp. 231–238.

Donahue M, Gurvits L, Darken C y Sontag E [1997]. Rates of Convex Approximation in Non-Hilbert Spaces, *Constructive Approximation* **13**, 187–220.

Donoho D L, Johnstone I M, Hoch J C y Stern A S [1992]. Maximum Entropy and the Nearly Black Object, *Journal of the Royal Statistical Society, B* **54**, 41–81.

Donoho D L y Johnstone I M [1989]. Projection-Based Approximation and a Duality with Kernel Methods, *The Annals of Statistics* **17**, 58–106.

Dorffner G [1996]. Neural Networks for Time Series Processing, *Neural Network World* **6**, 447–468.

Drucker H [1997]. Improving Regressors Using Boosting Techniques, *en Proceedings of the Fourteenth International Conference on Machine Learning*, Morgan Kaufmann, pp. 107–115.

Drucker H, Burges C, Kaufman L, Smola A y Vapnik V [1997]. Support Vector Regression Machines, *en* M Mozer, M Jordan y T Petsche, eds, *Advances in Neural Information Processing Systems,*, MIT Press, pp. 155–161.

Drucker H, Schapire R y Simard P [1993]. Boosting Performance in Neural Networks, *International Journal of Pattern Recognition and Artificial Intelligence* **7**, 705–719.

Duch W y Jankowski N [1999]. Survey of Neural Transfer Functions, Technical report, Dpt. of Computer Methods, Nicholas Copernicus University, Torun, Poland.

Duda R O, Hart P E y Stork D G [2001]. *Pattern Classification*, John Wiley.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Dyn N [1987]. Interpolation of Scattered Data by Radial Functions, *en* C K Chui, L L Schumaker y F Utreras, eds, *Topics in Multivariate Approximation*, Academic Press, pp. 47–61.

Dyn N [1989]. Interpolation and Approximation by Radial and Related Functions, *en* C K Chui, L L Schumaker y J D Ward, eds, *Approximation Theory VI: Vol.*, Academic Press, pp. 211–234.

Dyn N, Levin D y Rippa S [1986]. Numerical Procedures for Surface Fitting of Scattered Data by Radial Functions, *SIAM J. of Sci. Stat. Comput.* **7**, 639–659.

Efromovich S [1999]. *Nonparametric Curve Estimation*, Springer.

Efron B [1975]. The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis, *Journal of the American Statistical Association* **70**, 892–898.

Efron B [1979]. Boostrap Methods: Another Look at the Jackknife, *Annals of Statistics* **7**, 1–26.

Efron B [1982]. *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM.

Efron B y Gong G [1983]. A Leisurely Look at the Bootstrap, the Jackknife and Cross-Validation, *The American Statistician* **37**, 36–48.

Efron B y Tibshirani R [1993]. *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Efron B y Tibshirani R J [1997]. Improvements on Cross-Validation: The .632+ Bootstrap Method, *Journal of the American Statistical Association* **92**, 648–560.

Elsgoltz L [1977]. *Ecuaciones Diferenciales Y Cálculo Variacional*, Editorial Mir.

Engle R F [1982]. Autoregressive Conditional Heteroskedasticiy with Estimates of the Variance of United Kingdom Inflation, *Econometrica* **50**, 987–1007.

Engle R F, Lilien D M y Robins R P [1987]. Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model, *Econometrica* **55**, 391–407.

Engle R F y Bollerslev P [1986]. Modelling Persistence of Conditional Variances, *Econometric Reviews* **5**, 1–50.

Eubank R L [1984]. The Hat Matrix for Smoothing Splines, *Statistics and Probability Letters* **2**, 9–14.

Eubank R L [1988]. *Spline Smoothing and Nonparametric Regression*, Marcel Dekker.

Evgeniou T [2000], Learning with Kernel Machine Architectures, Master's thesis, Massachusetts Institute of Technology.

Evgeniou T, Pontil M y Poggio T [2000]. Regularization Networks and Support Vector Machines, *Advances in Computational Mathematics* **13**, 1–50.

Fahlman S [1990]. The Cascade-Correlation Learning Architecture, *en* D Touretzky, ed., *Advances in Neural Information Processing Systems 2*, Morgan Kaufmann Publishers, San Mateo, CA, pp. 524–532.

Fan J y Gijbels I [1996]. *Local Polynomial Modelling and Its Applications*, Chapman and Hall.

Faragó A y Lugosi G [1993]. Strong Universal Consistency of Neural Network Classifiers, *IEEE Transactions on Information Theory* **39**, 1146–1151.

USC DE SANTIAGO DE COMPOSTELA

Felicísimo A M, Francés E, Fernández J M, González-Díez A y Varas J [2002]. Modelling the Potential Distribution of Forests with a GIS, *Photogrammetric Engineering and Remote Sensing* **68**, 457–461.

Fine T [1999]. *Feedforward Neural Network Methodology*, Springer.

Fine T y Mukherjee S [1999]. Parameter Convergence and Learning Curves for Neural Networks, *Neural Computation* **11**, 747–769.

Fontenla-Romero O, Erdogmus D, Príncipe J C, Alonso-Betanzos A y Castillo E [2003]. Linear Least-Squared Based Methods for Neural Networks Learning, *en* O Kaynak, E Alpaydin, E Oja y L Xu, eds, *Artificial Neural Networks and Neural Information Processing, ICANN/ICONIP 2003, Istambul.*, Springer, pp. 84–91.

Foresee F D y Hagan T [1997]. Gauss-Newton Approximation to Bayesian Regularization, *en Proceedings of the 1997 International Joint Conference on Neural Networks*, pp. 1930–1935.

Fornai F y Mele A [1997]. Sign and Volatility Switching ARCH Models: Theory and Applications to International Stock Markets, *Journal of Applied Econometrics* **12**, 49–65.

Freund Y [1995]. Boosting a Weak Learning Algorithm by Majority, *Information and Computation* **121**, 256–285.

Freund Y y Schapire R E [1997]. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting, *Journal of Computer and Systems Sciences* **55**, 119–139.

Friedman J [2001]. Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics* **39**.

Friedman J H [1987]. Exploratory Projection Pursuit, *Journal of the American Statistical Association* **82**, 249–266.

Friedman J H [1997]. On Bias, Variance, 0/1-Loss and the Curse-of-Dimensionality, *Data Mining and Knowledge Discovery* **1**, 55–77.

Friedman J H, Grosse E y Stuetzle W [1983]. Multidimensional Additive Spline Approximation, *SIAM J. Sci. Statist. Comput* **4**, 291–301.

Friedman J H y Stuetzle W [1981]. Projection Pursuit Regression, *Journal of the American Statistical Association* **76**, 817–823.

Friedman J H y Tukey J W [1974]. A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions on Computers* **C-23**, 881–890.

Friedman J, Hastie T y Tibshirani R [2000]. Additive Logistic Regression: A Statistical View of Boosting, *The Annals of Statistics* **28**, 337–407.

Fukunaga K [1990]. *Introduction to Statistical Pattern Recognition*, Morgan Kaufmann.

Funahashi K I [1989]. On the Approximate Realization of Continuous Mappings by Neural Networks, *Neural Networks* **2**, 183–192.

Gabor D [1954]. Communication Theory and Cybernetics, *IRE Transactions on Circuit Theory* **CT-1**, 19–31.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

García-Guinea J, Lombardero M, Roberts B, Taboada J y Peto A [1998]. Mineralogía Y Microestructura de la Pizarra de Techar: Comportamiento Termoóptico Y Fisilidad, *Materiales de Construcción* **48**, 37–48.

Gasser T y Müller H G [1979]. Kernel Estimation of Regression Functions, *en* T Gasser y M Rosenblatt, eds, *Smoothing Techniques for Curve Estimation*, Springer-Verlag, pp. 23–68.

Gelfand A E y Dey D K [1994]. Bayesian Model Choice: Asymptotics and Exact Calculations, *Journal of the Royal Statistical Society, B* **56**, 501–514.

Geman S, Bienenstock E y Doursat R [1992]. Neural Networks and the Bias/Variance Dilemma, *Neural Computation* **4**, 1–58.

Geweke J [1986]. Comment to "Modelling Persistence of Conditional Variance", R. F. Engle and T. Bollerslev, *Econometric Reviews* **5**, 57–61.

Girosi F [1994]. Regularization Theory, Radial Basis Functions and Networks, *en* V Cherkassky, J H Friedman y H Wechsler, eds, *From Statistics to Neural Networks*, Springer, pp. 166–187.

Girosi F [1998]. An Equivalence Between Sparse Approximation and Support Vector Machines, *Neural Computation* **10**, 1455–1480.

Girosi F, Jones M y Poggio T [1995]. Regularization Theory and Neural Networks Architectures, *Neural Computation* **7**, 219–269.

Girosi F y Anzellotti G [1993]. Rates of Convergence for Radial Basis Functions and Neural Networks, *en* R J Mammone, ed., *Artificial Neural Networks for Speech and Vision*, Chapman and Hall, pp. 97–114.

Girosi F y Poggio T [1990]. Networks and the Best Approximation Property, *Biological Cybernetics* **63**, 169–176.

Goldberg D E [1989]. *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.

González-Manteiga W, Febrero M y Reboredo J C [2002]. Análisis Y Predicción Del Precio de Los Activos Financieros, Technical report, Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela.

Good I J [1977]. Explicativity: A Mathematical Theory of Explanation with Statistical Applications, *Proceedings of the Royal Society of London, series A* **354**, 303–330.

Gorman J W y Toman R J [1966]. Selection of Variables for Fitting Equations to Data, *Technometrics* **8**, 27–51.

Gotway C A y Cressie N [1993]. Improved Multivariate Prediction under a General Linear Model, *Journal of Multivariate Analysis* **45**, 56–72.

Goutte C [1997]. Note on Free Lunches and Cross-Validation, *Neural Computation* **9**, 1211–1215.

Granger C W J y Newbold P [1986]. *Forecasting Economic Time Series*, Academic Press.

Granger C W J y Teräsvirta T [1993]. *Modelling Nonlinear Economic Relationships*, Oxford University Press.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Green P J y Silverman B W [1994]. *Nonparametric Regression and Generalized Lineal Models*, Chapman and Hall.

Greenacre M J [1984]. *Theory and Applications of Correspondonce Analysis*, Academic Press.

Grenander U [1981]. *Abstract Inference*, John Wiley.

Grossberg S [1980]. How Does a Brain Build a Cognitive Code?, *Psychological Review* **87**, 1–51.

Grove A J y Schuurmans D [1998]. Boosting in the Limit: Maximizing the Margin of Learned Ensembles, *en Proceedings of the Fifthteen National Conference on Artificial Intelligence*.

Guisan A, Theurillat J P y Kiesnat F [1998]. Predicting the Potential Distribution of Plant Species in an Alpine Environment, *Journal of Vegetation Science* **9**, 65–74.

Haddawy P, Jacobson J y Kahn C E [1997]. Banter: A Bayesian Network Tutoring Shell, *Artificial Intelligence in Medicine* **10**, 177–200.

Hagan M T, Demuth H B y Beale M H [1996]. *Neural Network Design*, PWS Publishing.

Hagan M T y Menhaj M [1994]. Training Feedforward Networks with the Marquardt Algorithm, *IEEE Transactionson Neural Networks* **5**, 989–993.

Hall P [1992]. *The Bootstrap and Edgeworth Expansion*, Springer.

Hamilton J D y Susmel R [1992]. Autoregressive Conditional Heteroskedasticity and Changes in Regime, *Journal of Econometrics* **64**, 307–333.

Hammer B y Gersmann K [2003]. A Note on the Universal Approximation Capability of Support Vector Machines, *Neural Processing Letters* **17**, 43–53.

Hampshire J B y Pearlmutter B A [1990]. Equivalence Proofs for Multi-Layer Perceptron Classifiers and the Bayesian Discriminant Function, *en* Touretzky, Elman, Sejnowski y Hinton, eds, *Proceedings of the 1990 Connectionist Models Summer School*, Morgan Kaufmann.

Hansen L K y Larsen J [1996]. Linear Unlearning for Cross-Validation, *Advances in Computational Mathematics* **5**, 269–280.

Härdle W [1990]. *Applied Nonparametric Regression*, Cambridge University Press.

Härdle W, Hall P y Marron J S [1988]. How Far are Automatically Chosen Regression Smoothing Parameters from their Optimum?, *Journal of the American Statistical Association* **83**, 86–101.

Härdle W, Kerkyacharian G, Picard D y Tsybakov A [1998]. *Wavelets, Approximation and Statistical Applications*, Springer.

Härdle W y Marron J S [1985]. Optimal Bandwidth Selection in Nonparametric Regression Function Estimation, *The Annals of Statistics* **13**, 1465–1481.

Hart J D [1997]. *Nonparametric Smoothing and Lack-of-Fit Tests*, Springer.

Hartman E J, Keeler J D y Kowalski J M [1990]. Layered Neural Networks with Gaussian Hidden Units as Universal Approximations, *Neural Computation* **2**, 210–215.

Hashem S [1997]. Optimal Linear Combinations of Neural Networks, *Neural Networks* **10**, 599–614.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Hassibi B y Stork D G [1993]. Second order derivatives for network pruning: Optimal Brain Surgeon, *en* C L Giles, S J Hanson y J D Cowan, eds, *Advances in Neural Information Processing Systems*, Vol. 5, Morgan Kaufmann, San Mateo, CA.

Hastie T J y Tibshirani R J [1990]. *Generalized Additive Models*, Chapman and Hall.

Hastie T, Tibshirani R J y Friedman J [2001]. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer.

Hastie T y Stuetzle W [1989]. Principal Curves, *Journal of the American Statistical Association* **84**, 502–516.

Haussler D [1992]. Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications, *Information and Computation* **100**.

Haykin S [1999]. *Neural Networks. A Comprehensive Foundation*, Prentice Hall.

Hebb D [1949]. *The Organization of Behaviour*, Wiley.

Heskes T [1997]. Bias/Variance Decompositions for Likelihood-Based Estimators, *Neural Computation* **10**, 1425–1433.

Higgins M L y Bera A K [1992]. A Class of Non Linear ARCH Models, *International Economic Review* **33**, 137–158.

Hoaglin D C y Welsch R E [1980]. The Hat Matrix in Regression and ANOVA, *American Statistician* **32**, 17–22.

Hocking R R [1976]. The Analysis and Selection of Variables in Linear Regression, *Biometrics* **32**, 1–49.

Hoerl A E y Kennard R W [1970*a*]. Ridge Regression: Applications to Non-Orthogonal Problems, *Technometrics* **12**, 69–82.

Hoerl A E y Kennard R W [1970*b*]. Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics* **12**, 55–67.

Hoeting J A, Madigan D, Raftery A E y Volinsky C T [1999]. Bayesian Model Averaging: A Tutorial, *Statistical Science* **14**, 382–417.

Hoffmann-Jorgensen J [1994]. *Probability with a View Toward Statistics, I-II*, Chapman and Hall.

Holland S S [1990]. *Applied Analysis by the Hilbert Space Method*, Marcel Dekker.

Hopfield J [1982]. Neural Networks and Physical Systems with Emergent Collective Computational Abilities, *en Proceedings of the National Academy of Sciences - USA*, Vol. 79, p. 2554.

Hornik K [1991]. Approximation Capabilities of Multilayer Feedforward Networks, *Neural Networks* **4**, 251–257.

Hornik K, Stinchcombe M, White H y Auer P [1994]. Degree of Approximation Results for Feedforward Networks Approximating Unknown Mappings and their Derivatives, *Neural Computation* **6**, 1262–1275.

Hornik K, Stinchcombe M y White H [1989]. Multilayer Feedforward Networks are Universal Approximators, *Neural Networks* **2**, 359–366.

USC DE COMPOSTELA

Huber P [1964]. Robust Estimation of a Location Parameter, *Annals of Mathematical Statistics* **35**, 73–101.

Huber P J [1985]. Projection Pursuit., *The Annals of Statistics* **13**, 435–525.

Hurvich C M y Tsai C L [1989]. Regression and Time Series Model Selection in Small Samples, *Biometrika* **76**, 297–307.

Hush D R y Horne B [1998]. Efficient Algorithms for Function Approximation with Piecewise Linear Sigmoidal Networks, *IEEE Transactions on Neural Networks* **9**, 1129–1141.

Hwang JÑ, Lay S R, Maechler M, Martin R D y Schimert J [1994]. Regression Modeling in Back-Propagation and Projection Pursuit Learning, *IEEE Transactions on Neural Networks* **5**, 342–353.

Hwang J T G y Ding A A [1997]. Prediction Intervals for Artificial Neural Networks, *Journal of the American Statistical Association* **92**, 748–757.

Ioffe A D y Tihomirov V M [1979]. *Theory of Extremal Problems*, North-Holland.

Ishiguro M, Sakamoto Y y Kitagawa G [1997]. Bootstrapping Log Likelihood and EIC, an Extension of AIC, *Annals of the Institute of Statistical Mathematics* **49**, 411–434.

Jackson I R H [1988]. Convergence Properties of Radial Basis Functions, *Constructive Approximation* **4**, 243–264.

Jacobs R A, Jordan M I, Nowlan S J y Hinton G E [1991]. Adaptive Mixtures of Local Experts, *Neural Computation* **3**, 79–87.

James G y Hastie T [1997]. Generalizations of the Bias/Variance Decomposition for Prediction Error, Technical report, Department of Statistics, Stanford University.

Joachims T [1999]. Making Large-Scale SVM Learning Practical, *en* B Schölkopf, C J C Burges y A J Smola, eds, *Advances in Kernel Methods - Support Vector Learning*, MIT Press, pp. 169–184.

Johnson R A y Wichern D W [1992]. *Applied Multivariate Statistical Analysis*, Prentice-Hall.

Jones L K [1987]. On a Conjecture of Huber Concerning the Convergence of Projection Pursuit Regression, *The Annals of Statistics* **15**, 880–882.

Jones L K [1990]. Constructive Approximations for Neural Networks by Sigmoidal Functions, *Proceedings of the IEEE* **78**, 1586–1589.

Jones L K [1992]. A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training, *The Annals of Statistics* **20**, 608–613.

Jones L K [2000]. Local Greedy Approximation for Nonlinear Regression and Neural Network Training, *The Annals of Statistics* **28**, 1379–1389.

Jordan M I y Jacobs R A [1994]. Hierarchical Mixtures of Experts and the EM Algorithm, *Neural Computation* **6**, 181–214.

Journel A G [1989]. *Fundamentals of Geostatistics in Five Lessons*, American Geophysical Union.

Journel A G y Huijbregts C J [1978]. *Mining Geostatistics*, Academic Press.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Kainen P C, Kurkova V y Vogt A [1999]. Approximation by Neural Networks is Not Continuous, *Neurocomputing* **29**, 47–56.

Kaiser G [1994]. *A Friendly Guide to Wavelets*, Birkhäuser.

Kearns M J, Schapire R B y Sellie L M [1994]. Toward Efficient Agnostic Learning, *Machine Learning* **17**, 115–142.

Kearns M J y Schapire R E [1994]. Efficient Distribution-Free Learning of Probabilistic Concepts, *Journal of Computer and System Sciences* **48**, 464–497.

Kearns M J y Vazirani U V [1997]. *An Introduction to Computational Learning Theory*, MIT Press.

Kearns M y Ron D [1999]. Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation, *Neural Computation* **11**, 1427–1453.

Keerthi S y Gilbert E G [2002]. Convergence of a Generalized SMO Algorithm for SVM Classifier Design., *Machine Learning* **46**, 351–360.

Kent J T [1989]. Continuity Properties for Random Fields, *Annals of Probability* **17**, 1432–1440.

Kimeldorf G y Wahba G [1971]. Some Results on Tchebycheffian Spline Functions, *Journal of Mathematical Analysis and Applications* **33**, 82–95.

Kirkpatrick S, Gelatt C D y Vecchi M P [1983]. Optimization by Simulated Annealing, *Science* **220**, 671–680.

Kohavi R [1995]. A Study of Cross-Validation and Bootstrap for Acuracy Estimation and Model Selection, *en Proceedings of the Joint Conference on Artificial Intelligence*, pp. 1137–1145.

Kohavi R y Wolpert D H [1996]. Bias Plus Variance Decomposition for Zero-One Loss Functions, *en Proceedings of the Thirteenth International Conference on Machine Learning*, Morgan Kaufmann, pp. 275–283.

Kohonen T [1982]. Self-Organized Formation of Topologically Correct Feature Maps, *Biological Cybernetics* **43**, 59–69.

Kohonen T [2001]. *Self-Organizing Maps*, Springer.

Koike K, Matsuda S y Gu B [2001]. Evaluation of Interpolation Accuracy of Neural Kriging with Application to Temperature-Distribution Analysis, *Mathematical Geology* **33**, 421–448.

Kolmogorov AÑ y Fomin S V [1978]. *Elementos de la Teoría de Funciones Y Del Análisis Funcional*, 1975 edn, Editorial Mir.

Kong E B y Dietterich T G [1995]. Error-Correcting Output Coding Corrects Bias and Variance, *en Proceedings of the Twelfth International Conference on Machine Learning*, Morgan Kaufmann, pp. 313–321.

Kramer M A [1991]. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks, *AIChe Journal* **37**, 233–243.

Kramer M A [1992]. Autoassociative Neural Networks, *Computers and Chemical Engineering* **16**, 313–328.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Krogh A y Vedelsby J [1995]. Neural Network Ensembles, Cross Validation and Active Learning, *en* G Tesauro, D S Touretzky y T K Leen, eds, *Advances in Neural Information Processing Systems*, MIT Press.

Krzyzak A, Linder T y Lugosi G [1996]. Nonparametric Estimation and Classification Using Radial Basis Function Nets and Empirical Risk Minimization, *IEEE Transactions on Neural Networks* **7**, 475–487.

Kugarajah T y Zhang Q [1995]. Multidimensional Wavelet Frames, *IEEE Transactions on Neural Networks* **5**, 1552–1556.

Kullback S [1968]. *Information Theory and Statistics*, Peter Smith.

Kurková V [1991]. Kolmogorov's Theorem is Relevant, *Neural Computation* **3**, 617–622.

Kurková V [1992]. Kolmogorov's Theorem and Multilayer Neural Networks, *Neural Networks* **5**, 501–506.

Kwok T Y y Yeung D Y [1996]. Use of Bias Term in Projection Pursuit Learning Improves Approximation and Convergence Properties, *IEEE Transactions on Neural Networks* **7**, 1168–1183.

Larsen J, Hansen L K, Svarer C y Ohlsson M [1996]. Design and Regularization of Neural Networks. The Optimal Use of a Validation Set, *en* S Usui, Y Tohkura, S Katagiri y E Wilson, eds, *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, IEEE, pp. 62–71.

Larsen J, Svarer C, Andersen LÑ y Hansen L K [1998]. Adaptive Regularization in Neural Network Modeling, *en* G B Orr y K Müller, eds, *Neural Networks: Tricks of the Trade*, Springer, pp. 113–132.

Larsen J y Hansen L K [1994]. Generalization Performance of Regularized Neural Network Models, *en* J Vlontzos, JÑ Hwang y E Wilson, eds, *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, IEEE, pp. 42–51.

Laslett G M [1994]. Kriging and Splines: An Empirical Comparison of their Predictive Performance in some Applications, *Journal of the American Statistical Association* **89**, 391–409.

Lee W, Bartlett P y Williamson R [1996]. Efficient Agnostic Learning of Neural Networks with Bounded Fan-in, *IEEE Transactions on Information Theory* **42**, 2118–2132.

Leshno M, Lin V Y, Pinkus A y Schocken S [1993]. Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function, *Neural Networks* **6**, 861–867.

Li K C [1986]. Asymptotic Optimality of CL and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing, *The Annals of Statistics* **14**, 1101–1112.

Li K C [1987]. Asymptotic Optimality for Cp, CL, Cross-Validation and Generalized Cross-Validation: Discrete Index Set, *The Annals of Statistics* **15**, 958–975.

Lin C J [2001]. On the Convergence of the Decomposition Method for Support Vector Machines, *IEEE Transactions on Neural Networks* **12**, 1288–1298.

Linhart H y Volkers P [1984]. Asymptotic Criteria for Model Selection, *OR Spektrum* **6**, 161–165.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Littman E y Ritter H [1992]. Generalization Abilities of Cascade Network Architectures, *en* S J Hanson, J D Cowan y C L Giles, eds, *Advances in Neural Information Processing Systems*, Morgan Kaufmann.

Liu Y [1993]. Neural Network Model Selection Using Asymptotic Jackknife Estimator and Cross-Validation Method, *en* C L Giles, S J Hanson y J D Cowan, eds, *Advances in Neural Information Processing Systems*, Vol. 5, Morgan Kauffman.

Liu Y [1995]. Unbiased Estimate of Generalization Error and Model Selection in Neural Networks, *Neural Networks* **8**, 215–219.

Ljung L [1999]. *System Identification*, Prentice Hall.

Loader C [1999]. *Local Regression and Likelihood*, Springer.

Loève M [1976]. *Teoría de la Probabilidad*, Tecnos.

Lorentz G G [1986]. *Approximation of Functions*, Chelsea Publishing Co.

Lowe D y Webb A R [1990]. Exploiting Prior Knowledge in Network Optimization: An Illustration from Medical Prognosis, *Network: Computation in Neural Systems* **1**, 299–323.

Lowe D y Webb A R [1991]. Optimized Feature Extraction and the Bayes Decision in Feed-Forward Classifier Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 355–364.

Luenberger D E [1984]. *Programación Lineal Y No Lineal*, Addison-Wesley Iberoamericana.

Lugosi G y Nobel A B [1999]. Adaptive Model Selection Using Empirical Complexities, *The Annals of Statistics* **27**, 1830–1864.

Lugosi G y Pawlak M [1994]. On the Posterior-Probability Estimate of the Error of Nonparametric Classification Rules, *IEEE Transactions on Information Theory* **40**, 475–481.

Lugosi G y Vayatis N [2001]. On the Bayes Risk Consistency of Boosting Methods, Technical report, Departamento de Economía, Universidad Pompeu Fabra, Barcelona.

Lugosi G y Vayatis N [2002]. A Consistent Strategy for Boosting Algorithms, *en Proceedings of the Annual Conference on Computational Learning Theory*, Springer, pp. 303–318.

Lugosi G y Vayatis N [2003]. On the Bayes-Risk Consistency of Regularized Boosting Methods, Technical report, Laboratoire de Probabilités et Modèles Aléatoires. Université Pierre et Marie Curie, Paris 6, et Denis Diderot, Paris 7.

Lugosi G y Zeger K [1995]. Nonparametric Estimation Via Empirical Risk Minimization, *IEEE Transactions on Information Theory* **41**, 677–687.

Lugosi G y Zeger K [1996]. Concept Learning Using Complexity Regularization, *IEEE Transactions on Information Theory* **42**, 48–54.

Luo Z Q [1991]. On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks, *Neural Computation* **3**, 227–245.

MacCulloch W S y Pitts W S [1943]. A Logical Calculus of the Ideas Inmanent in Nervous Activity, *Bulletin of Mathematical Biophysics* **5**, 115–133.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

MacKay D J C [1992*a*]. Bayesian Interpolation, *Neural Computation* **4**, 415–447.

MacKay D J C [1992*b*]. The Evidence Framework Applied to Classification Networks, *Neural Computation* **4**, 720–736.

MacKay D J C [1992*c*]. A Practical Bayesian Framework for Back-Propagation Networks, *Neural Computation* **4**, 448–472.

MacKay D J C [1994]. Hyperparameters: Optimize or Integrate Out, *en* G Heidbreder, ed., *Maximum Entropy and Bayesian Methods*, Kluwer.

Madych W R y Nelson S A [1988]. Multivariate Interpolation and Conditionally Positive Definite Functions, *Approximation Theory and its Applications* **4**, 77–89.

Madych W R y Nelson S A [1990]. Multivariate Interpolation and Conditionally Positive Definite Functions, II, *Mathematics of Computation* **54**, 211–230.

Maiorov V y Meir R [1998]. Approximation Bounds for Smooth Functions in C(Rd) by Neural and Mixture Networks, *IEEE Transactions on Neural Networks* **9**, 969–978.

Makovoz Y [1996]. Random Approximants and Neural Networks, *Journal of Approximation Theory* **85**, 98–109.

Mallat S [1989]. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

Mallat S y Zhang Z [1992]. Matching Pursuit in a Time-Frequency Dictionary, *IEEE Transactions on Information Theory* **38**, 617–643.

Mallows C L [1973]. Some Comments on Cp, *Technometrics* **15**, 661–675.

Mangasarian O L [1999]. Arbitrary-Norm Separating Plane, *Operations Research Letters* **24**, 15–23.

Marquardt D W [1963]. An Algorithm for Least-Squares Estimation of Nonlinear Parameters, *Journal of the Society of Industrial and Applied Mathematics* **11**, 431–441.

Mason L, Baxter J, Bartlett P y Frean M [2000]. Boosting Algorithms as Gradient Descent, *en* S Solla, T Leen y K Müller, eds, *Advances in Neural Information Processing Systems*, MIT Press.

Matheron G [1973]. The Intrinsic Random Functions and their Applications, *Advances in Applied Probability* **5**, 439–468.

Matías J M [2003]. Boosting con Redes Neuronales RBF. Análisis Sesgo-Varianza En un Problema de Clasificación, *en VI Congreso Galego de Estatística e Investigación Operativa, Vigo*, Sociedad Galega de Estatítica e Investigación de Operacións.

Matías J M, Taboada J, Vaamonde A y González-Manteiga W [2004]. Comparison Between Kriging, MLP and RBF in a Slate Mine, *en* X Sánchez-Vila, J Carrera y J Gómez-Hernández, eds, *Geostatistics for Environmental Applications. GeoEnV IV.*, Kluwer Academic Publishers, pp. 518–519.

Matías J M, Vaamonde A, Taboada J y González-Manteiga W [2004*a*]. Comparison of Kriging and Neural Networks with Application to the Exploitation of a Slate Mine, *Mathematical Geology* **36**(4), 485–508.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Matías J M, Vaamonde A, Taboada J y González-Manteiga W [2004*b*]. Support Vector Machines and Gradient Boosting for Graphical Estimation of a Slate Deposit, *Journal of Stochastic Environmental Research and Risk Assessment* **18**, 1–15.

Matías J M y González-Manteiga W [2003*a*]. Regularized Kriging as a Generalization of Simple and Universal Kriging, *en Actas Del 27 Congreso Nacional de Estadística e Investigación Operativa, Lleida, Spain*.

Matías J M y González-Manteiga W [2003*b*]. Regularized Kriging: The Support Vectors Method Applied to Kriging, *en Joint 13th International Conference on Artificial Neural Networks and 10th International Conference on Neural Information Processing. Istambul, Turkey. Lecture Notes in Computer Science (LNCS 2714)*, Springer.

McCullagh P y Nelder J A [1983]. *Generalized Linear Models*, 1989 edn, Chapman and Hall.

McLachlan G J y Basford K E [1988]. *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker.

Medeiros M C, Teräsvirta T y Rech G [2002]. Building Neural Networks Models for Time Series: A Statistical Approach, Technical Report 461, Departamento de Economia. PUC-Rio, Department of Economic Statistics, Stockholm School of Economics, Sweden.

Meir R y Rätsch G [2003]. An Introduction to Boosting and Leveraging, *en* S Mendelson y A Smola, eds, *Advanced Lectures on Machine Learning*, Springer, pp. 119–184.

Merz C J y Pazzani M J [1999]. A Principal Components Approach to Combining Regression Estimates, *Machine Learning* **36**, 9–32.

Mhaskar HÑ [1996]. Neural Networks for Optimal Approximation of Smooth and Analytic Functions, *Neural Computation* **8**, 164–177.

Mhaskar HÑ y Micchelli C A [1992]. Approximation by Superposition of Sigmoidal and Radial Basis Functions, *Advances in Applied Mathematics* **13**, 350–373.

Micchelli C A [1986]. Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions, *Constructive Approximation* **2**, 11–22.

Miller A J [1990]. *Subset Selection in Regression*, Chapman and Hall.

Minsky M y Papert S [1969]. *Perceptrons: An Introduction to Computational Geometry*, MIT Press.

Moody J E [1991]. Note on Generalization, Regularization and Architecture Selection in Nonlinear Learning Systems, *en* B H Juang, S Y Kang y C A Kamm, eds, *Proceedings Fo the First IEEE Workshop on Neural Networks for Signal Processing*, IEEE, pp. 1–10.

Moody J E [1992]. The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems, *en* J E Moody, S J Hanson y R P Lippmann, eds, *Advances in Neural Information Processing Systems*, Morgan Kaufmann, pp. 847–854.

Moody J E [1993]. Prediction Risk and Architecture Selection for Neural Networks, *en* V Cherkassky, J H Friedman y H Wechsler, eds, *From Statistics to Neural Networks*, Springer-Verlag, pp. 147–165.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Moody J E y Darken C J [1988]. Learning with Localized Receptive Fields, *en* Touretzky, Hinton y Sejnowski, eds, *Proceedings of the 1988 Connectionist Models Summer School*, Morgan Kaufmann.

Moody J E y Darken C J [1989]. Fast Learning in Networks of Locally-Tuned Processing Units, *Neural Computation* **1**, 281–294.

Moody J E y Rögnvaldsson T [1997]. Smoothing Regularizers for Projective Basis Function Networks, *en* M C Moozer, M Jordan y T Petsche, eds, *Advances in Neural Information Processing Systems*, Morgan Kaufmann.

Müller K R, Smola A J, Rätsch G, Schölkopf B, Kohlmorgen J y Vapnik V [1999]. Using Support Vector Machines for Time Series Prediction, *en* B Schölkopf, C J C Burges y A J Smola, eds, *Advances in Kernel Methods. Support Vector Learning*, MIT Press, pp. 243–253.

Murata N, Yoshizawa S y Amari S [1991]. A Criterion for Determining the Number of Parameters in an Artificial Neural Network Model, *en* T Kohonen y K Mäkisara, eds, *Proceedings of the International Conference on Neural Networks*, North Holland, pp. 9–14.

Murata N, Yoshizawa S y Amari S [1994]. Network Information Criterion -Determining the Number of Hidden Units for Artificial Neural Network Models, *IEEE Transaction on Neural Networks* **5**, 865–872.

Nadaraya E A [1964]. On Estimating Regression, *Theory of Probability and its Applications* **10**, 186–190.

Neal R M [1996]. *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics. Springer.

Nguyen D y Widrow B [1990]. Improving the Learning Speed of 2-Layer Neural Networks by Choosing Initial Values of the Adaptive Weigths, *Proceedings of the Joint Conference on Neural Networks* **3**, III.21–III.26.

Niyogi P y Girosi F [1996]. On the Relationship Between Generalization Error, Hypothesis Complexity, and Sample Complexity for Radial Basis Functions, *Neural Computation* **8**, 819–842.

Niyogi P y Girosi F [1999]. Generalization Bounds for Function Approximation from Scattered Noisy Data, *Advances in Computational Mathematics* **10**, 51–80.

O'Hagan A [1994]. *Bayesian Inference. Kendall's Advanced Tehroy of Statistics, Volumen 2B*, John Wiley.

Omre H [1987]. Bayesian Kriging -Merging Observations and Qualified Guesses in Kriging-, *Mathematical Geology* **19**, 25–39.

Omre H y Halvorsen K B [1989]. The Bayesian Bridge Between Simple and Universal Kriging, *Mathematical Geology* **21**, 767–786.

Onoda T, Rätsch G y Müller K R [1998]. An Asymptotic Analysis of AdaBoost in the Binary Classification Case, *en* L Niklasson, M Bodén y T Ziemke, eds, *Proceedings of the International Conference on Artificial Neural Networks*, ICSC Academic Press, pp. 195–200.

Ordóñez C, Matías J M, Felicísimo A M y Taboada J [2003*a*]. Learning Machines Applied to Potential Distribution of Forests, Technical report, Departamento de Ingeniería de los Recursos Naturales y Medio Ambiente. Universidad de Vigo. En proceso de revisión por Journal of Environmental Management.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Ordóñez C, Matías J M, Felicísimo A M y Taboada J [2003*b*]. Support Vector Machines Applied to Reforestation, *en Conference Book of the ISI International Conference on Environmental Statistics and Health*, Universidad de Santiago de Compostela.

Orr M J L [1995]. Regularization in the Selection of Radial Basis Function Centers, *Neural Computation* **7**, 606–623.

Orr M J L [1996]. Introduction to Radial Basis Function Networks, Technical report, Centre for Cognitive Science, University of Edimburg, http:
www.anc.ed.ac.uk.

Osuna E, Freund R y Girosi F [1997]. An Improved Training Algorithm for Support Vector Machines, *en* J Príncipe, L Gile, N Morgan y E Wilson, eds, *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Workshop*, IEEE, pp. 276–285.

Paass G [1993]. Assesing and Improving Neural Network Predictions by the Bootstrap Algorithm, *en* S Hanson, J Cowan y C Giles, eds, *Advances in Neural Information Processing Systems*, Morgan Kaufmann, pp. 196–203.

Pan W [1999]. Bootstrapping Likelihood for Model Selection with Small Samples, *Journal of Computational and Graphical Statistics* **8**, 687–698.

Park J y Sandberg I W [1991]. Universal Approximation Using Radial-Basis-Function Networks, *Neural Computation* **3**, 246–257.

Park J y Sandberg I W [1993]. Approximation and Radial-Basis-Function Networks, *Neural Computation* **5**, 305–316.

Parthasarathy K R y Schmidt K [1972]. *Positive Definite Kernels, Continuous Tensor Products and Central Limit Theorems of Probability Theory*, Springer-Verlag.

Parzen E [1962]. On Estimation of a Probability Density Function and Mode, *Annals of Mathematical Statistics* **33**, 1065–1076.

Pati Y C, Rezaiifar R y Krishnaprasad P S [1993]. Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition, *en* A Singh, ed., *Proceedings 27th Asilomar Conference on Signals, Systems and Computers*, IEEE Comput. Soc. Press.

Pati Y C y Krishnaprasad P S [1990]. Discrete Affine Wavelet Transforms for Analysis and Synthesis Feedforward Neural Networks, *en* R Lippman, J Moody y D Touretzky, eds, *Advances in Neural Information Processing Systems,*, Morgan Kaufmann, pp. 743–749.

Pati Y C y Krishnaprasad P S [1993]. Analysis and Synthesis of Feedforward Neural Networks Using Discrete Affine Wavelet Forms, *IEEE Transactions on Neural Networks* **4**, 73–85.

Pinkus A [1985]. *N-Widths in Approximation Theory*, Springer-Verlag.

Pinkus A [1999]. Approximation Theory of the MLP Model in Neural Networks, *Acta Numerica* pp. 143–195.

Pinkus A [2003]. Strictly Positive Definite Functions on a Real Inner Product Space, *Advances in Computational Mathematics (pendiente de aparición)* .

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Platt J [1999]. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, *en* B Schölkopf, C J C Burges y A J Smola, eds, *Advances in Kernel Methods - Support Vector Learning*, MIT Press, pp. 185–208.

Plutowski M, Sakata S y White H [1994]. Cross-Validation Estimates IMSE, *en* J Cowan, G Tesauro y J Alspector, eds, *Advances in Neural Information Processing Systems*, Morgan Kaufmann, pp. 1135–1142.

Poggio T y Girosi F [1990]. Networks for Approximation and Learning, *Proceedings of the IEEE* **78**, 1481–1497.

Politis DÑ, Romano J P y Wolf M [1999]. *Subsampling*, Springer.

Pollard D [1984]. *Convergence of Stochastic Processes*, Springer.

Pollard D [1990]. *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics - American Statistical Association.

Popov V A [1989]. Nonlinear Multivariate Approximation, *en* C K Chui, L L Schumaker y J D Ward, eds, *Approximation Theory VI, Vol.*, Academic Press, pp. 519–535.

Popper K R [1962]. *La Lógica de la Investigación Científica*, Tecnos.

Powell M J D [1985]. Radial Basis Functions for Multivariable Interpolation: A Review, *en IMA Conference on Algorithms for the Approximation of Functions and Data*, RMCS Shrivenham.

Powell M J D [1987]. Radial Basis Functions for Multivariable Interpolation: A Review, *en* J C Mason y M G Cox, eds, *Algorithms for Approximation*, Clarendon Press, pp. 143–167.

Powell M J D [1992]. The Theory of Radial Basis Function Approximation in 1990, *en* W A Light, ed., *Advances in Numerical Analysis III. Wavelets, Subdivision Algorithms and Radial Basis Functions*, Clarendon Press, pp. 105–210.

Pratt J W, Raiffa H y Schlaifer R [1995]. *Introduction to Statistical Decision Theory*, MIT Press.

Press W H, Teukolsky S A, Vetterling W T y Flannery B P [1992]. *Numerical Recipes in*, Cambridge University Press.

Priestley M B [1981]. *Spectral Analysis and Time Series*, Academic Press.

Priestley M B y Chao M T [1972]. Non-Parametric Function Fitting, *Journal of the Royal Statistical Society, Ser. B* **34**, 385–392.

Pryce J D [1973]. *Basic Methods of Linear Functional Analysis*, Hutchinson University Library.

Ramsay J O y Silverman B W [1997]. *Functional Data Analysis*, Springer.

Rao C R y Wu Y [1989]. A Strongly Consistent Procedure for Model Selection in a Regression Problem, *Biometrika* **76**, 369–374.

Rätsch G [2002], Robust Boosting Via Convex Optimization: Theory and Applications, PhD thesis, Mathematisch-Naturwissenschaftlichen Fackultät der Universität Postdam.

Rätsch G, Onoda T y Müller K R [1998]. Soft Margin for AdaBoost, Technical Report NC-TR-1998-021, NeuroCOLT.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Rätsch G, Onoda T y Müller K R [1999]. Regularizing AdaBoost, *en* M S Kearns, S A Solla y D A Cohn, eds, *Advances in Neural Information Processing Systems, 11*, MIT Press, pp. 564–570.

Rätsch G, Onoda T y Müller K R [2001]. Soft Margins for AdaBoost, *Machine Learning* **42**, 287–320.

Reed R [1993]. Pruning Algorithms –A Survey, *IEEE Transactions on Neural Networks* **4**, 740–747.

Refenes AÑ y Holt W T [2001]. Forecasting Volatility with Neural Regression: A Contribution to Model Adequacy, *IEEE Transactions on Neural Networks* **12**, 850–864.

Refenes A PÑ, Burgess AÑ y Bentz Y [1997]. Neural Networks in Financial Engineering: A Study in Methodology, *IEEE Transactions on Neural Networks* **8**, 1222–1267.

Richards J I y Youn H K [1990]. *Theory of Distributions*, Cambridge University Press.

Ridgeway G [1999]. The State of Boosting, *Computing Science and Statistics* **31**, 172–181.

Ridgeway G, Madigan D y Richardson T [1999]. Boosting Methodology for Regression Problems, *en* D Heckerman y J Whittaker, eds, *Proceedings of Artificial Intelligence and Statistics*, pp. 152–161.

Ripley B D [1981]. *Spatial Statistics*, John Wiley.

Ripley B D [1996]. *Pattern Recognition and Neural Networks*, Cambridge University Press.

Rippa S [1999]. An Algorithm for Selecting a Good Value for the Parameter C in Radial Basis Function Interpolation, *Advances in Computational Mathematics* **11**, 193–210.

Rissanen J [1978]. Modelling by Shortest Data Description, *Automatica* **14**, 465–471.

Rissanen J [1983]. A Universal Prior for Integers and Estimation by Minimum Description Length, *Annals of Statistics* **11**, 416–431.

Rissanen J [1987]. Stochastic Complexity, *Journal of the Royal Statistical Society, B* **49**, 223–239.

Rivals I y Personnaz L [1999]. On Cross-Validation for Model Selection, *Neural Computation* **11**, 863–870.

Rivals I y Personnaz L [2000]. Construction of Confidence Intervals for Neural Networks Based on Least Squares Estimation, *Neural Networks* **13**, 143–145.

Robbins H y Monro S [1951]. A Stochastic Approximation Method, *Annals of Mathematical Statistics* **22**, 400–407.

Rogers W H y Wagner T J [1978]. A Fine Sample Distribution-Free Performance Bound for Local Discrimination Rules, *The Annals of Statistics* **6**, 506–514.

Rosemblatt F [1958]. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, *Psychological Review* **65**, 386–408.

Rosenfield G H y Fitzpatric-Lins K [1986]. A Coefficient of Agreement as a Measure of Thematic Classification Accuracy, *Photogrammetric Engineering and Remote Sensing* **52**, 223–227.

Rudin W [1988]. *Análisis Real Y Complejo*, McGraw-Hill.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Rumelhart D E, Hinton G E y Williams R J [1986]. Learning Internal Representations by Error Propagation, *en* T W Anderson y E Rosenfeld, eds, *Neurocomputing: Foundations of Research*, MIT Press.

Scarselli F y Tsoi A C [1998]. Universal Approximation Using Feedforward Neural Networks: A Survey of some Existing Methods and some New Results, *Neural Networks* **11**, 15–37.

Schaback R [1993]. Comparison of Radial Basis Function Interpolants, *en* K Jetter y F I Utreras, eds, *Multivariate Approximations*, World Scientific, pp. 293–305.

Schapire R E [1990]. The Strength of Weak Learnability, *Machine Learning* **5**, 197–227.

Schapire R E y Singer Y [1999]. Improved Boosting Algorithms Using Confidence-Rated Predictions, *Machine Learning* **37**, 297–336.

Schapire R, Freund Y, Bartlett Y y Lee W [1998]. Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods, *The Annals of Statistics* **26**, 1651–1686.

Schioler H y Hartmann U [1992]. Mapping Neural Network Derived from the Parzen Window Estimator, *Neural Networks* **5**, 903–909.

Schittenkopf C, Dorffner G y Dockner E [1998]. Volatility Prediction with Mixture Density Networks, *en Proc. Of the International Conference on Artificial Neural Networks (ICANN-98)*, pp. 929–934.

Schittenkopf C, Dorffner G y Dockner E [1999]. Fat Tails and Non-Linearity in Volatility Models: What is More Important?, *en Proceedings of the 1999 Conference on Computational Intelligence for Financial Engineering (CIFEr99), New York City*, pp. 259–266.

Schittenkopf C, Dorffner G y Dockner E J [2000]. Forecasting Time-Dependent Conditional Densities: A Seminonparametric Neural Network Approach, *Journal of Forecasting* **19**, 355–374.

Schoemberg I J [1938]. Metric Spaces and Completely Monotone Functions, *Annals of Mathematics* **39**, 811–841.

Schölkopf B, Herbrich R, Smola A J y Williamson R [2000]. A Generalized Representer Theorem, Technical Report NC2-TR-2000-81, NeuroCOLT2, http://www.neurocolt.com.

Schölkopf B, Smola A J, Williamson R C y Bartlett P L [2000]. New Support Vector Algorithm, *Neural Computation* **12**, 1207–1245.

Schölkopff B y Smola A J [2002]. *Learning with Kernels*, The MIT Press.

Schwarz G [1978]. Estimating the Dimension of a Model, *Annals of Statistics* **6**, 461–464.

Seber G A F [1977]. *Linear Regression Analysis*, John Wiley.

Seber G A F [1984]. *Multivariate Observations*, John Wiley.

Seber G A F y Wild C J [1989]. *Nonlinear Regression*, John Wiley.

Shao J [1993]. Linear Model Selection by Cross-Validation, *Journal of the American Statistical Association* **88**, 486–494.

Shao J [1997]. An Asymptotic Theory for Linear Model Selection, *Statistica Sinica* **7**, 221–264.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Sharkey A J C [1999]. Multi-Net Systems, *en* A J C Sharkey, ed., *Combining Artificial Neural Nets. Ensemble and Modular Multi-Net Systems*, Springer, pp. 1–30.

Sherstinsky A y Picard R W [1996]. On the Efficiency of the Orthogonal Least Squares Training Method for Radial Basis Function Networks, *IEEE Transactions on Neural Networks* **7**, 195–200.

Shibata R [1976]. Selection of the Order of an Autoregressive Model by Akaike's Information Criterion, *Biometrika* **63**, 117–126.

Shibata R [1981]. An Optimal Selection of Regression Variables, *Biometrika* **68**, 45–54.

Sigurdsson S, Larsen J y Hansen L K [2000]. On Comparison of Adaptive Regularization Methods, *en* B Widrow, L Guan, K Paliwa y T Adali, eds, *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, IEEE, pp. 221–230.

Silverman B W [1984]. Spline Smoothing: The Equivalent Variable Kernel Method, *The Annals of Statistics* **12**, 898–916.

Sjöberg J y Ljung L [1995]. Overtraining, Regularization and Searching for a Minimum with Application to Neural Networks, *International Journal of Control* **62**, 1391–1407.

Smola A, Frieß J y Scholköpf B [1999]. Semiparametric Support Vector and Linear Programming Machines, *en* M S Kearns, S A Solla y D A Cohn, eds, *Advances in Neural Information Processing Systems*, MIT Press, pp. 585–591.

Smola A, Schölkopf B y Müller K R [1998]. The Connection Between Regularization Operators and Support Vector Kernels, *Neural Networks* **11**, 637–649.

Smola A y Schölkopf B [1998]. A Tutorial on Support Vector Regression, Technical Report NC2-TR-1998-030, NeuroCOLT2, http://www.neurocolt.com.

Specht D F [1990]. Probabilistic Neural Networks, *Neural Networks* **3**, 109–118.

Stein M [1999]. *Interpolation of Spatial Data: Some Theory for Kriging*, Springer.

Steinwart I [2001]. On the Influence of the Kernel on the Consistency of Support Vector Machines, *Journal of Machine Learning Research* **2**, 67–93.

Steinwart I [2002*a*]. Consistency of Support Vector Machines and Other Regularized Kernel Machines, *remitido a IEEE Transactions on Information Theory* .

Steinwart I [2002*b*]. Support Vector Machines are Universally Consistent, *Journal of Complexity* **18**, 768–791.

Stewart J [1976]. Positive Definite Functions and its Generalizations, a Historical Survey, *Rocky Mountain Journal of Mathematics* **6**, 409–434.

Stock J H y Watson M W [1998]. A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series, Technical Report 6607, National Bureau of Economic Research, USA.

Stone C J [1977*a*]. Consistent Nonparametric Regression, *The Annals of Statistics* **5**, 595–620.

Stone M [1974]. Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society, Series B* **36**, 111–133.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Stone M [1977*b*]. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion, *Journal of the Royal Statistical Society, Ser. B* **39**, 44–47.

Sugiura N [1978]. Further Analysis of the Data by Akaike's Information Criteron and the Finite Corrections, *Communications in Statistics, Theory and Methods* **A7**, 13–26.

Taboada J, Vaamonde A, Saavedra A y Alejano L [1997]. Application of Geostatistical Techniques to Exploitation Planning in Slate Quarries, *Engineering Geology* **47**, 269–277.

Taboada J, Vaamonde A, Saavedra A y Argüelles A [1998]. Quality Index for Ornamental Slate Deposits, *Engineering Geology* **30**, 203–210.

Taboada J, Vaamonde A, Saavedra A y Ordóñez C [2002]. Geostatistical Study of the Feldspar Content and Quality of a Granite Deposit, *Engineering Geology* **65**, 285–292.

Taboada J, Vaamonde A y Saavedra A [2001]. Evaluation of a Slate Extraction Bank, *Transactions of the Institution of Mining and Metallurgy* **A 110**, 40–46.

Talagrand M [1994]. Sharper Bounds for Gaussian and Empirical Processes, *The Annals of Probability* **22**, 28–76.

Tan S y Mavrovouniotis M L [1995]. Reducing Data Dimensionality Through Optimizing Neural Networks Inputs, *AIChe Journal* **41**, 1471–1480.

Teräsvirta T, Lin C F y Granger C W J [1993]. Power of the Neural Network Linearity Test, *Journal of Time Series Analysis* **14**, 209–220.

Tibshirani R [1996*a*]. Bias, Variance and Prediction Error for Classification Rules, Technical report, Department of Statistics, University of Toronto.

Tibshirani R [1996*b*]. Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society, B* **58**, 267–288.

Tibshirani R J [1996*c*]. A Comparison of some Error Estimates for Neural Models, *Neural Computation* **8**, 152–163.

Tibshirani R J y Knight K [1999]. Model Search and Inference by Bootstrap "bumping", *Journal of Computational and Graphical Statistics* **8**, 671–686.

Tikhonov AÑ y Arsenin V Y [1977]. *Solutions of Ill-Posed Problems*, John Wiley.

Tino P, Schittenkopf C y Dorffner G [2001]. Financial Volatility Trading Using Recurrent Neural Networks, *IEEE Transactions on Neural Networks* **12**, 865–874.

Trapletti A, Leisch F y Hornik K [2000]. Stationarity and Integrated Autoregressive Neural Network Processes, *Neural Computation* **12**, 2427–2450.

Tumer K y Ghosh J G [1999]. Linear and Order Statistics Combiners for Pattern Classification, *en* A J C Sharkey, ed., *Combining Artificial Neural Nets. Ensemble and Modular Multi-Net Systems*, Springer, pp. 127–161.

Uykan Z, Güzelis C, Celebi M E y Koivo HÑ [2000]. Analysis of Input-Output Clustering for Determining Centers of RBFN, *IEEE Transactions on Neural Networks* **11**, 851–857.

Valiant L G [1984]. A Theory of the Learnable, *Communications of the ACM* **27**, 1134–1142.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Vanderbei R J [1998]. LOQO: An Interior Point Code for Quadratic Programming, Technical Report SOR-94-15, Statistics and Operations Research, Princeton University.

Vapnik V [1982]. *Estimation of Dependences Based on Empirical Data*, Springer.

Vapnik V [1995]. *The Nature of Statistical Learning Theory*, Springer, N.Y.

Vapnik V [1998]. *Statistical Learning Theory*, John Wiley.

Vapnik V, Levin E y Cun Y L [1994]. Measuring the VC-Dimension of a Learning Machine, *Neural Computation* **6**, 851–876.

Vapnik V y Chervonenkis A Y [1971]. On the Uniform Convergence of Relative Frequencies of Events to their Probabilities, *Theory of Probability and Its Applications* **XVI**, 264–280.

Vapnik V y Chervonenkis A Y [1981]. Necessary and Sufficient Conditions for the Uniform Convergence of Means to their Expectations, *Theory of Probability and Its Applications* **XXVI**, 532–553.

Vasin V V [1970]. Relationship of Several Variational Methods for the Approximate Solution of Ill-Posed Problems, *Mathematical Notes* **7**, 161–165.

Vidakovic B [1999]. *Statistical Modeling by Wavelets*, John Wiley.

Vidyasagar M [1997]. *A Theory of Learning and Generalization with Applications to Neural Networks and Control Systems*, Springer.

Vidyasagar M [2002]. *Learning and Generalisation: With Applications to Neural Networks*, Springer-Verlag.

Vila J P, Wagner V y Neveu P [2000]. Bayesian Nonlinear Model Selection and Neural Networks: A Conjugate Prior Approach, *IEEE Transactions on Neural Networks* **11**, 265–277.

Vincent P y Bengio Y [2002]. Kernel Matching Pursuit, *Machine Learning* **48**, 165–187.

Wahba G [1990*a*]. Reply to Cressie, *The American Statistician* **44**, 255–256.

Wahba G [1990*b*]. *Spline Models for Observational Data*, SIAM.

Walter G G [1994]. *Wavelets and Other Orthogonal Systems with Applications*, CRC Press.

Wand M P y Jones M C [1995]. *Kernel Smoothing*, Chapman and Hall.

Watson G S [1964]. Smooth Regression Analysis, *Sankhya, Ser. A* **26**, 101–116.

Watson G S [1984]. Smoothing and Interpolation by Kriging and with Splines, *Mathematical Geology* **16**, 601–615.

Webb A R y Lowe D [1990]. The Optimised Internal Representation of Multilayer Classifier Networks Performs Nonlinear Discriminant Analysis, *Neural Networks* **3**, 367–375.

Weber D D y Englund E J [1992]. Evaluation and Comparison of Spatial Interpolators, *Mathematical Geology* **24**, 381–391.

Weber D D y Englund E J [1994]. Evaluation and Comparison of Spatial Interpolators, II., *Mathematical Geology* **26**, 589–603.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Weigend A S, Rumelhart D E y Huberman B A [1991]. Generalization by Weight-Elimination with Application to Forecasting, *en Advances in Neural Information Processing Systems*, Vol. 3, Morgan Kaufmann.

Weigend A S y Gershenfeld N A [1994]. *Time Series Prediction: Forecasting the Future and Understanding the Past. Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis*, Perseus Books.

Weigend A S y LeBaron B [1994]. Evaluating Neural Network Predictors by Bootstrapping, Technical Report CU-CS-725-94, Computer Science Department, University of Colorado.

White H [1982]. Maximum Likelihood Estimation of Mispecified Models, *Econometrica* **50**, 1–25.

White H [1989]. Some Asymptotic Results for Learning in Single Hidden Layer Feedforward Networks, *Journal of the American Statistical Association* **84**, 1003–1013.

White H [1990]. Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings, *Neural Networks* **3**, 535–549.

White H [1992]. *Artificial Neural Networks. Approximation and Learning Theory*, Blackwell.

White H [1994]. *Estimation, Inference and Specification Analysis*, Cambridge University Press.

Widrow B y Hoff M E [1988]. Adaptive Switching Circuits, *en* J A Anderson y E Rosenfeld, eds, *Neurocomputing: Foundations of Research*, MIT Press.

Williams C K I [1998]. Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond, *en* M I Jordan, ed., *Learning and Inference in Graphical Models*, Kluwer.

Wolpert D H [1992]. Stacked Generalization, *Neural Networks* **5**, 241–259.

Yakowitz S J y Szidarovszky F [1985]. A Comparison of Kriging with Nonparametric Regression Methods, *Multivariate Analysis* **16**, 21–53.

Zhang P [1992]. On the Distributional Properties of Model Selection Criteria, *Journal of the American Statistical Association* **87**, 732–737.

Zhang Q [1997]. Using Wavelet Networks in Nonparametric Estimation, *IEEE Transactions on Neural Networks* **8**, 227–236.

Zhang Q y Benveniste A [1992]. Wavelet Networks, *IEEE Transactions on Neural Networks* **3**, 889–898.

Zhang T [2002]. A General Greedy Approximation Algorithm with Applications, *en Advances in Neural Information Processing Systems*, MIT Press.

Zhang T y Yu B [2003]. Boosting with Early Stopping: Convergence and Consistency, Technical Report 635, Department of Statistics, UC Berkeley.

Zhao Y y Atkeson C G [1996]. Implementing Projection Pursuit Learning, *IEEE Transactions on Neural Networks* **7**, 362–373.

Zhu H y Rohwer R [1996]. No Free Lunch for Cross-Validation, *Neural Computation* **8**, 1421–1426.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA