Traballo Fin de Grao

Métodos de Clasificación e Ensamblado de Clasificadores en Aprendizaxe Supervisada

Antón Gómez López

2023/2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Métodos de Clasificación e Ensamblado de Clasificadores en Aprendizaxe Supervisada

Antón Gómez López

Xullo, 2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Métodos de Clasificación e Ensamblado de Clasificadores en Aprendizaxe Supervisada
Breve descrición do contido
Dentro da aprendizaxe supervisada, a clasificación é unha das técnicas máis empregadas para predicir a clase ou etiqueta dun obxecto a partir das súas características. Nos últimos anos, a investigación en clasificación centrouse no desenvolvemento de métodos de ensamblado de clasificadores, que buscan mellorar a exactitude das predicións combinando a saída de varios clasificadores. Existen diferentes metodoloxías para levar a cabo a combinación de predicións de clasificadores base, dende métodos de ponderación e votación ata os chamados métodos de meta-aprendizaxe nos que, en lugar de combinar directamente as predicións de clasificadores base, utilízanse ditas predicións como entradas para outro clasificador, chamado meta-clasificador. O obxectivo deste TFG é facer unha revisión dos principais métodos de clasificación e ensamblado e discutir os enfoques máis recentes no deseño de métodos de ensamblado.
Recomendacións
Outras observacións

Índice

Resumo

Neste traballo analízanse diversas técnicas de ensamblado en aprendizaxe supervisada, enfocándose en bagging, bosques aleatorios e adaBoost. Inicialmente, explícanse os fundamentos da clasificación estatística e da aprendizaxe supervisada. Seguidamente, examínanse as diferentes estratexias para combinar saídas de clasificadores cando estas consisten en predicións e valores continuos. Finalmente, detállanse os métodos de ensamblado, subliñando as características que os diferencian.

Abstract

In this work, various ensemble techniques in supervised learning are analyzed, focusing on bagging, random forests, and adaBoost. Initially, the fundamentals of statistical classification and supervised learning are explained. Then, the different strategies for combining classifier outputs, consisting of predictions and continuous values, are examined. Finally, the ensemble methods are detailed, highlighting the characteristics that differentiate them.

Introdución

A necesidade de clasificar obxectos está presente en numerosos campos. Por exemplo, os xestores de correo electrónico clasifican as mensaxes entrantes en "spam" ou "non spam", os médicos clasifican as células cancerígenas en "benignas" ou "malignas" e os xestores de risco clasifican as operacións financeiras en "fraude" ou "non fraude".

Deste xeito, no campo da aprendizaxe automática, un problema de clasificación consiste en asignar etiquetas a obxectos descritos mediante características. Os obxectos poden ser mensaxes de correo electrónico, células cancerígenas ou operacións financeiras, entre outros. As características poden ser o contido do correo electrónico, as características das células ou os datos da operación financeira, respectivamente.

A aprendizaxe supervisada é unha subdisciplica da aprendizaxe automática que adoita ser o enfoque predominante para abordar problemas de clasificación. Neste paradigma, adéstrase un modelo utilizando un conxunto de datos etiquetado, onde cada obxecto vén acompañado da súa respectiva etiqueta. O obxectivo do modelo é aprender a xeralizar a partir deste conxunto de datos, de modo que poida predicir con exactitude as etiquetas de novos obxectos. Con outras palabras, a aprendizaxe supervisada busca descubrir a relación subxacente entre as características presentes no conxunto de datos e as etiquetas de saída a través dun proceso de adestramento.

Durante o século XX, propuxéronse diversos métodos de clasificación que hoxe en día seguen a utilizarse en problemas de clasificación e que son a base doutros métodos máis complexos. Son exemplos a análise discriminante linear, proposta por Fisher en 1936, o método dos k veciños máis próximos, proposto por Fix e Hodges en 1951, o perceptrón, proposto por Rosenblatt en 1958, e as árbores de decisión, propostas por Quinlan en 1986.

A finais do século XX, propuxéronse os primeiros métodos de ensamblado. O obxectivo destes métodos é combinar as predicións de múltiples modelos de clasificación base para mellorar a precisión e a robustez do modelo final. En 1996, Leo Breiman propuxo o método de ensamblado denominado *Bootstrap Aggregating* (bagging). Nese mesmo ano, Yoav Freund e Robert Schapire propuxeron o método *Adaptive Bosting* (adaBoost). Posteriormente, en 2001, Leo Breiman presentou o método bosques aleatorios, que é un caso particular do método bagging, onde se utilizan

árbores de decisión como clasificadores base.

Á hora de deseñar un ensemble, hai diversas cuestións que afectan ao seu rendemento. Unha das máis importantes é a forma na que se combinan as saídas dos clasificadores base. Algúns métodos de combinación son intuitivos e amplamente utilizados, como o voto por maioría. Outros métodos son máis complexos e requiren o cálculo de parámetros específicos, chegando ao punto de que o propio método de combinación se converte nun clasificador adicional. Desta maneira, neste traballo estudaranse os distintos métodos de combinación de saídas de clasificadores e presentaranse tres métodos de ensamblado: bagging, bosques aleatorios e AdaBoost.

A continuación preséntase a estrutura do traballo. No Capítulo 1 presentaranse os conceptos básicos necesarios para a comprensión dos métodos de combinación de saídas de clasificadores. Introdúcense os conceptos básicos da aprendizaxe supervisada, a definición de clasificador, a definición de erro de clasificación e, por último, preséntase o clasificador de Bayes.

Os Capítulos 2 e 3 teñen unha estrutura similar e describen os métodos existentes para combinar as saídas dos clasificadores que forman parte dun ensemble cando estes consisten en predicións e en valores continuos, respectivamente.

No Capítulo 4 preséntanse tres métodos de ensamblado: bagging, bosques aleatorios e adaBoost e describirase o seu funcionamento. Por último, no Capítulo 5 presentaranse as conclusións do traballo. Tanto a estrutura como o contido dos capítulos seguen a referencia [8].

Adicionalmente, no Apéndice A preséntase un exemplo que ilustra a optimalidade dos métodos de combinación de predicións abordados no Capítulo 2.

Capítulo 1

Preliminares

Ao longo deste capítulo introdúcense algúns conceptos básicos necesarios para a comprensión dos métodos de ensamblado, que se utilizaran recorrentemente nos seguintes capítulos. En particular, presentaranse os conceptos básicos da aprendizaxe supervisada, a definición de clasificador, a definición de erro de clasificación e, por último, preséntase o clasificador de Bayes.

1.1. Conceptos básicos: clase, características, conxunto de datos

Cada obxecto ten asociada unha etiqueta. Se dous obxectos teñen a mesma etiqueta pertencen á mesma clase. Deste modo, as etiquetas relacionan os obxectos e agrúpanos en clases.

Nun problema de clasificación con c clases, denótase por $\Omega = \{\omega_1, \dots, \omega_c\}$ ao conxunto de etiquetas. Aínda que existen problemas de clasificación nos que un obxecto pode ter asignada máis dunha etiqueta, neste traballo asumiremos os obxectos soamente pertencen a unha clase.

Os obxectos descríbense mediante características. Son exemplos de características a velocidade do vento, a temperatura ou a altura dunha persoa. Podemos modelar as características dun obxecto mediante un vector númerico $\mathbf{x} = [x_1, \dots, x_n]^T \in \mathbb{R}^n$. Denomínase espazo de características ao espazo real $\mathbb{R}^n$, onde cada eixo se corresponde con unha característica.

Nun problema real as características dos obxectos poden ser categóricas ou, incluso, non numéricas. Por exemplo, a talla de camiseta dunha persoa pode ser unha característica non numérica. Neste traballo consideraremos que as características son numéricas e continuas.

A información necesaria para adestrar un clasificador agrúpase nun conxunto de datos que se denota por $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}, \mathbf{z}_j \in \mathbb{R}^n$, sendo n o número de características que describen os obxectos. A etiqueta de cada obxecto $\mathbf{z}_j$ denótase por $y_j \in \Omega$, $j = 1, \dots, N$. Tipicamente, o

conxunto de datos represéntase como unha matriz onde as filas almacenan os obxectos e as columnas, as características, cunha columna extra para as etiquetas.

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{Nn} \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

Así, o valor z_{ij} é o valor da característica j para o obxecto i, e y_i, a etiqueda do obxecto i.

1.2. Clasificadores, funcións discriminantes e rexións de clasificación

Un clasificador é unha función que asigna unha clase a un obxecto $\mathbf{x}$:

$$D : \mathbb{R}^n \to \Omega.$$

Polo xeral, a forma que ten un clasificador de realizar esta tarefa consiste en asignar un valor numérico por cada clase a cada obxecto. Deste modo, internamente un clasificador calcula c funcións discriminantes:

$$g_i : \mathbb{R}^n \to \mathbb{R}, \quad i = 1, \dots, c.$$

Para decidir que clase asignar a cada obxecto, o clasificador calcula o índice da función discriminante cun valor máis alto. Este xeito de asignar etiquetas particiona o espazo de características en c rexións de clasificación:

$$\mathcal{R}_i = \left\{ \mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) = \max_{k=1,\dots,c} g_k(\mathbf{x}) \right\}.$$

Cada rexión $\mathcal{R}_i$ consiste nun conxunto de puntos do espazo de características $\mathbb{R}^n$, para os cales a i-ésima función discriminante obtén un valor máis alto que todas as demais. Deste xeito, o clasificador e, en particular, as funcións discriminantes, determinan as rexións de clasificación. Na Figura 1.1 amósase un esquema dun clasificador.

As fronteiras das rexións de clasificación denomínanse fronteiras de decisión e conteñen os puntos do espazo de características para os cales, polo menos dúas funcións discriminantes obteñen o mesmo valor. Un punto situado nunha fronteira de decisión pode ser clasificado a calquera das clases que comparten a fronteira.

O obxectivo dun clasificador consiste en dividir o espazo de características en rexións de clasificación de modo que cada unha soamente conteña obxectos que pertencen á mesma clase. Se isto non sucede, é dicir, se dados $i, j, i \neq j$, a rexión $\mathcal{R}_i$ contén obxectos do dataset $\mathbf{Z}$ etiquetados coa clase ω_j, dicimos que as clases ω_i e ω_j se superpoñen. Deste modo, as superposicións veñen dadas polo clasificador e non polo conxunto de datos.

Polo xeral, calquera conxunto de funcións $\{g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})\}$ é un conxunto de funcións discriminantes. Porén, a forma na que estas dividen o espazo pode ser máis ou menos axeitada.

Sexa $G = \{g_1, \ldots, g_c\}$ un conxunto de funcións discriminantes. Pódense obter infinitos conxuntos de funcións discriminantes que dividen o espazo de características da mesma maneira. Por exemplo, se aplicamos unha transformación monótona f ás funcións discriminantes, $G' = \{f(g_1), \ldots, f(g_c)\}$, o conxunto G' será un conxunto de funcións que preservará a orde das funcións discriminantes para calquera obxecto $\mathbf{x} \in \mathbb{R}^n$ e, polo tanto, dividirá o espazo de características da mesma maneira. Nos seguintes capítulos estudaremos diferentes métodos de combinación de clasificadores e seranos de utilidade transformar as funcións discriminantes para chegar a expresións máis sinxelas ou para comparar dous métodos de combinación.

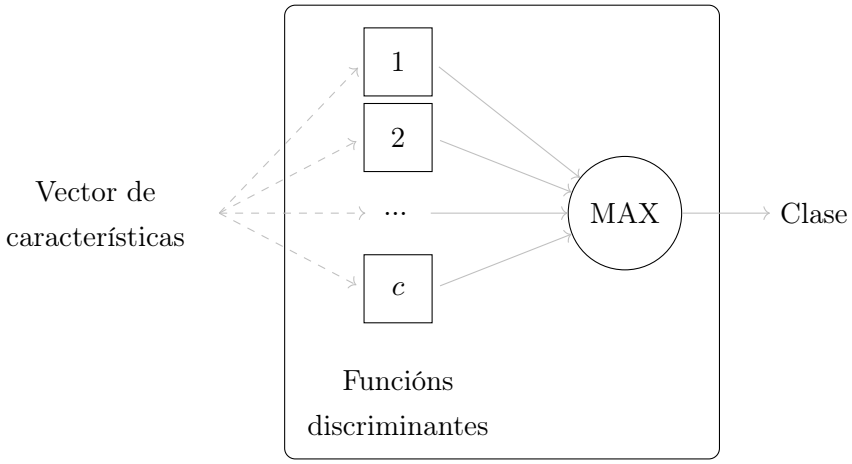


Figura 1.1: Esquema dun clasificador. Un vector de características n-dimensional, $\mathbf{x}$, pásase a través das c funcións discriminantes, e a función discriminante co valor maior determina a clase.

1.3. Erro de clasificación e exactitude

Para determinar o rendemento dun clasificador é necesario definir unha medida de erro. Unha medida común é o erro de clasificación, que se define a partir da exactitude. Defínese a exactitude como a fracción de obxectos ben clasificados:

$$\text{Exactitude} = \frac{\text{Número de obxectos ben clasificados}}{\text{Número total de obxectos}}.$$

Así, o erro de clasificación defínese como

$$\text{Erro de clasificación} = 1 - \text{Exactitude}.$$

No caso ideal de que coñecésemos a saída do clasificador para calquera posible $\mathbf{x} \in \mathbb{R}^n$, poderíamos calcular a súa exactitude. Non obstante, isto na práctica non é posible, polo que debemos utilizar un conxunto de datos para realizar unha estimación. É habitual utilizar un subconxunto do conxunto de datos que conteña obxectos non vistos polo clasificador durante o adestramento para estimar o erro de clasificación. A este subconxunto chámaselle conxunto de test e denotarémolo por $\mathbf{Z}_{\text{test}}$. O erro de clasificación estimado, $\hat{P_E}$, é a proporción de obxectos mal clasificados no conxunto de test:

$$\hat{P_E} = \frac{\text{Número de obxectos do conxunto de test mal clasificados}}{\text{Número total de obxectos do conxunto de test}}.$$

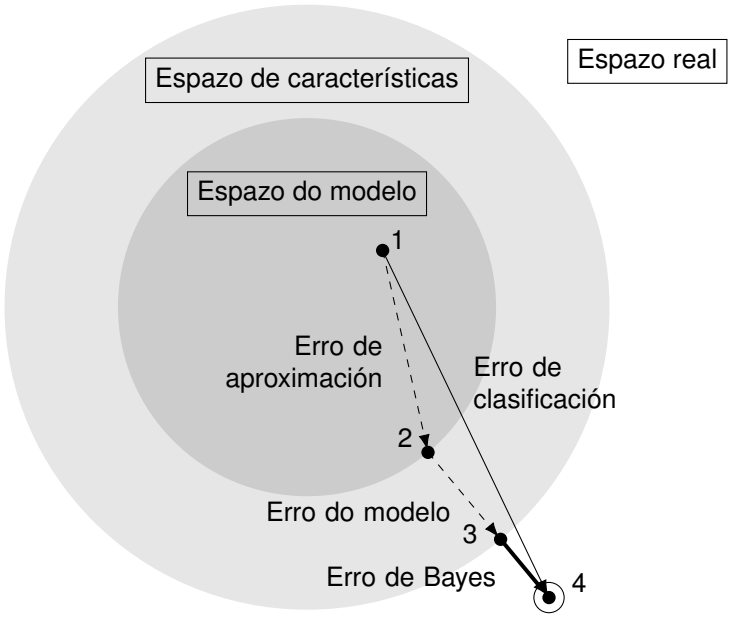
A estimación do erro depende en grande medida do tamaño do conxunto de test. Canto maior sexa o conxunto de test, máis fiable será a estimación. Non obstante, o tamaño do conxunto de test non é o único factor que inflúe na fiabilidade da estimación. Outro factor que pode afectar á estimación é a distribución dos obxectos no conxunto de test.

1.3.1. Descomposición do erro de clasificación

A Figura 1.2 amosa as distintas fontes posibles de erro dun clasificador.

- Erro de aproximación. Aínda que utilicemos un algoritmo moi bo, a nosa solución (marcada como 1 na Figura 1.2) pode ser distinta da mellor solución posible co modelo escollido (marcada como 2 na Figura 1.2). Este erro provén de que non dispoñemos dun conxunto de datos infinito para adestrar o modelo e, polo tanto, depende do conxunto de datos.
- Erro do modelo. Pode ser que o modelo escollido non sexa o máis axeitado para o problema, ou que exista outro modelo que sexa capaz de chegar a unha solución mellor a partir do espazo de características do problema. O punto 3 na Figura 1.2 representa a mellor solución posible coas características dispoñibles.
- Erro de Bayes. Finalmente, existe un erro irreducible. Este erro provén de que poden existir outras características que non temos modeladas que afectan ao noso problema e que non podemos medir. Coas características dispoñibles, pode darse o caso de que dous obxectos coas mesmas características pertenzan a clases distintas.

Desta maneira, o erro de clasificación, que denotaremos por $P_E(D, \mathbf{Z})$ pódese descompoñer



1. A nosa solución.
2. A mellor solución posible co modelo escollido.
3. A mellor solución posible coas características dispoñibles.
4. A solución real.

Figura 1.2: Descomposición do erro de clasificación.

en tres sumandos:

$$P_E(D, \mathbf{Z}) = P_A(\mathbf{Z}) + P_M + P_B,$$

onde $P_A(\mathbf{Z})$ é o erro de aproximación, P_M é o erro do modelo e P_B é o erro de Bayes.

1.3.2. Matriz de confusión

Ata agora, supuxemos que o clasificador comete un erro con probabilidade P_E para calquera obxecto $\mathbf{x}$. Non obstante, nun problema real é habitual que isto non sexa así. Por exemplo, ante un problema de clasificación de díxitos manuscritos, é máis probable que se cometa un erro ao clasificar un 4 como un 9 que como un 1.

A matriz de confusión permite analizar como se distribúen os erros a través das clases. Esta constrúese a partir das etiquetas reais e das preditas polo clasificador de $\mathbf{Z}_{\text{test}}$. Consiste nunha matriz cadrada de tamaño $c \times c$ onde a entrada i, j almacena o número de obxectos que pertencen á clase ω_i e foron clasificados como ω_j. A diagonal da matriz contén o número de obxectos ben clasificados, mentres que as entradas fóra da diagonal conteñen os obxectos mal clasificados. Desta forma, a matriz de confusión proporciona información de onde se cometen os erros.

1.4. Clasificador de Bayes

A teoría de decisión de Bayes permite construír un clasificador teórico que pode ser utilizado para comparar a exactitude doutros clasificadores. Asumamos que a etiqueta dun obxecto ω é unha variable aleatoria que toma valores en $\Omega = \{\omega_1, \dots, \omega_c\}$. As probabilidades a priori, $P(\omega_i)$, $i = 1, \dots, c$ constitúen a función de masa de probabilidade da variable ω:

$$0 \leq P(\omega_i) \leq 1, \qquad \sum_{i=1}^{c} P(\omega_i) = 1.$$

Asumimos que os obxectos de cada clase ω_i se distribúen en $\mathbb{R}^n$ segundo a función de densidade de probabilidade $p(\mathbf{x}|\omega_i)$, onde

$$p(\mathbf{x}|\omega_i) \geq 0, \, \forall \mathbf{x} \in \mathbb{R}^n, \qquad \int p(\mathbf{x}|\omega_i) d\mathbf{x} = 1, \, \forall i = 1, \dots, c.$$

A verosimilitude de $\mathbf{x} \in \mathbb{R}^n$ vén dada pola función de densidade de probabilidade incondicional

$$p(\mathbf{x}) = \sum_{i=1}^{c} P(\omega_i) p(\mathbf{x}|\omega_i).$$

Coñecidas as funcións de densidade condicionais e as probabilidades a priori, podemos calcular a probabilidade a posteriori de que un obxecto con vector de características $\mathbf{x}$ pertenza á clase ω_i utilizando o Teorema de Bayes:

$$P(\omega_i|\mathbf{x}) = \frac{P(\omega_i) p(\mathbf{x}|\omega_i)}{p(\mathbf{x})} = \frac{P(\omega_i) p(\mathbf{x}|\omega_i)}{\sum_{j=1}^{c} P(\omega_j) p(\mathbf{x}|\omega_j)}. \tag{1.1}$$

A Ecuación (1.1) proporciona a probabilidade a posteriori de que un obxecto pertenza á clase ω_i coñecido o seu vector de características, $\mathbf{x}$. Podemos utilizar as probabilidades a posteriori como funcións discriminantes dun clasificador.

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}), \; i = 1, \dots, c. \tag{1.2}$$

Deste xeito, podemos construír un clasificador D^* que, dado un obxecto $\mathbf{x}$ asigne a clase ω_i tal que

$$D^*(\mathbf{x}) = \omega_{i^*} \iff P(\omega_{i^*}|\mathbf{x}) = \max_{i=1,\dots,c} P(\omega_i|\mathbf{x}).$$

Este clasificador denomínanse clasificador de Bayes e, polo xeral, utilízase como clasificador

óptimo para comparar o rendemento con outros clasificadores. Nun probelma real non se coñecen as funcións de densidade de probabilidade condicionais nin as probabilidades a priori, polo que, é un clasificador que non se pode empregar na práctica.

Na Sección 1.2 xustificamos que se pode obter un conxunto de funcións discriminantes equivalentes a través dunha transformación monótona. Posto que o denominador da Ecuación (1.1) non depende de i, podemos simplificar a expresión das funcións discriminantes do seguinte xeito:

$$g_i(\mathbf{x}) = P(\omega_i)p(\mathbf{x}|\omega_i),\ i = 1, \dots, c. \tag{1.3}$$

Para un obxecto con vector de características $\mathbf{x}$, a orde das funcións discriminantes definidas en (1.2) e (1.3) é a mesma. Polo tanto, as rexións de clasificación definidas por ambas son as mesmas.

Exemplo 1.1. Sexa un problema de clasificación con 3 clases, $c = 3$, e supoñamos que coñecemos as probabilidades a priori e as funcións de densidade de probabilidade condicionais.

$$\begin{aligned} P(\omega_1) &= 0.45, \qquad p(x|\omega_1) \sim N(4, 2.0^2). \\ P(\omega_2) &= 0.35, \qquad p(x|\omega_2) \sim N(5, 1.2^2). \\ P(\omega_3) &= 0.25, \qquad p(x|\omega_3) \sim N(7, 1.0^2). \end{aligned}$$

Na Figura 1.3a amósanse as funcións discriminantes definidas en (1.3) e na Figura 1.3b as definidas en (1.2). En ambos casos, as fronteiras de decisión coinciden.

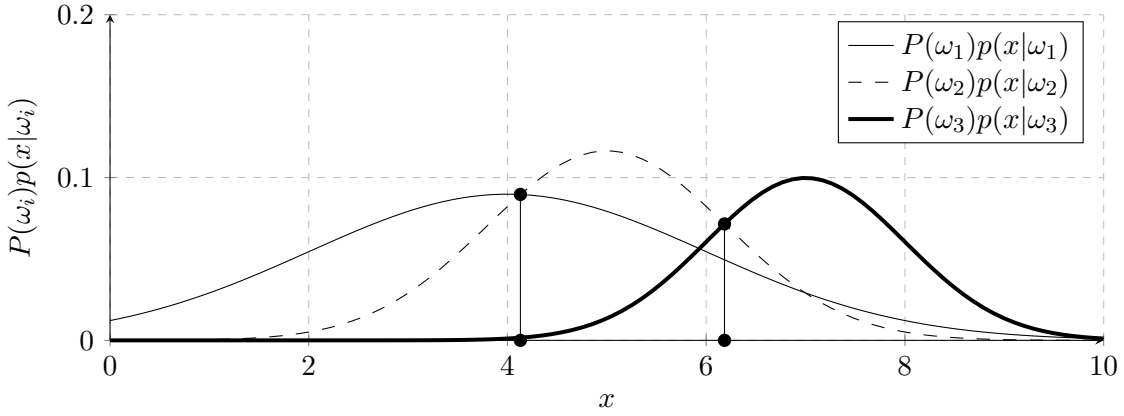
1.4.1. Erro de Bayes

Sexa D^* o clasificador de Bayes. Dado $\mathbf{x}$, a probabilidade de que o clasificador acerte é

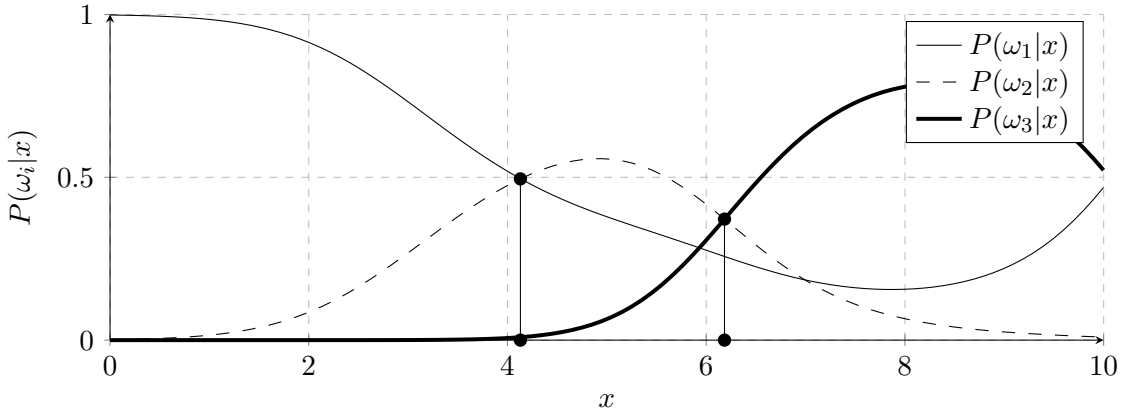
$$P(\omega_{i^*}|\mathbf{x}) = \max_{i=1,\dots,c} P(\omega_i|\mathbf{x}).$$

Polo tanto, a probabilidade de que o clasificador cometa un erro é $1 - P(\omega_{i^*}|\mathbf{x})$. Así, a probabilidade total de que o clasificador cometa un erro é a suma da probabilidade de erro para cada $\mathbf{x} \in \mathbb{R}^n$ ponderada pola función de densidade incondicional de $\mathbf{x}$

$$P_e(D^*) = \int_{\mathbb{R}^n} \left(1 - P(\omega_{i^*}|\mathbf{x})\right) p(\mathbf{x}) d\mathbf{x}.$$



(a) Funcións discriminantes definidas mediante $g_i = P(\omega_1)p(x|\omega_i)$.



(b) Funcións discriminantes definidas mediante $g_i = P(\omega_i|x)$.

Figura 1.3: Dous conxuntos de funcións discriminantes equivalentes.

Podemos separar a anterior integral en c integrais, unha para cada rexión de clasificación:

$$P_e(D^*) = \sum_{i=1}^{c} \int_{\mathcal{R}_i^*} \left(1 - P(\omega_i|\mathbf{x})\right) p(\mathbf{x}) d\mathbf{x}, \tag{1.4}$$

onde $\mathcal{R}_i^*$ é a rexión de clasificación asociada á clase ω_i, $\mathcal{R}_i^* \cap \mathcal{R}_j^* = \emptyset$, $i \neq j$ e $\bigcup_{i=1}^{c} \mathcal{R}_i^* = \mathbb{R}^n$. Desta forma,

$$\begin{aligned} P_e(D^*) &= \sum_{i=1}^{c} \int_{\mathcal{R}_i^*} \left(1 - \frac{P(\omega_i)p(\mathbf{x}|\omega_i)}{p(\mathbf{x})}\right) p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^n} p(\mathbf{x}) d\mathbf{x} \sum_{i=1}^{c} \int_{\mathcal{R}_i^*} P(\omega_i)p(\mathbf{x}|\omega_i) d\mathbf{x} \\ &= 1 - \sum_{i=1}^{c} P(\omega_i) \int_{\mathcal{R}_i^*} p(\mathbf{x}|\omega_i) d\mathbf{x}. \end{aligned}$$

Este erro denomínase erro de Bayes e é o mínimo erro de clasificación posible. Calquera outro clasificador que produza rexións de clasificación distintas, terá un erro de clasificación maior. Se consideramos outro clasificador distinto, D, con rexións de clasificación $\mathcal{R}_i$, $i = 1, \dots, c$, podemos expresar o erro dunha forma similar:

$$P_e(D) = 1 - \sum_{i=1}^{c} P(\omega_i) \int_{\mathcal{R}_i} p(\mathbf{x}|\omega_i) d\mathbf{x}.$$

No seguinte exemplo ilustramos o erro de Bayes e o erro adicional ao considerar un clasificador que produce rexións de clasificación distintas ás do clasificador de Bayes.

Exemplo 1.2. Consideremos un caso simple, onde o vector de características soamente ten unha dimensión, $x \in \mathbb{R}$, e o problema de clasificación soamente ten dúas clases, $\Omega = \{\omega_1, \omega_2\}$. Supoñamos que as funcións de densidade de probabilidade condicionais son normais e que as funcións discriminantes teñen a forma $g_i(x) = P(\omega_i)p(x|\omega_i)$, $i = 1, 2$. Na Figura 1.4 amósanse as funcións discriminantes e márcanse as fronteiras de decisión que delimitan as rexións de clasificación.

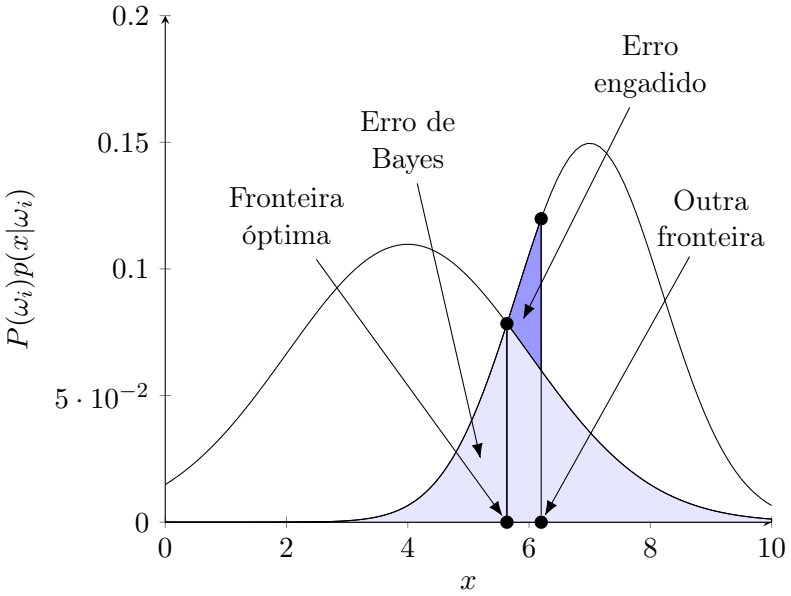


Figura 1.4: Gráfica de dúas funcións discriminantes $P(\omega_1)p(x|\omega_1)$ á dereita, e $P(\omega_2)p(x|\omega_2)$ á esquerda con $x \in [0, 10]$. En cor clara represéntase o erro se se establece a fronteira de decisión óptima (erro de Bayes). En cor máis escura represéntase o erro engadido se se establece outra fronteira de decisión distinta.

Para un problema de clasificación con dúas clases, temos que $P(\omega_1|x) = 1 - P(\omega_2|x)$. Así,

partindo da Ecuación (1.4) podemos reescribir o erro de Bayes como

$$\begin{aligned} P_e(D^*) &= \int_{\mathcal{R}_1^*} \left(1 - P(\omega_1|x)\right) p(x)dx + \int_{\mathcal{R}_2^*} \left(1 - P(\omega_2|x)\right) p(x)dx \\ &= \int_{\mathcal{R}_1^*} P(\omega_2|x)p(x)dx + \int_{\mathcal{R}_2^*} P(\omega_1|x)p(x)dx \\ &= \int_{\mathcal{R}_1^*} P(\omega_2)p(x|\omega_2)dx + \int_{\mathcal{R}_2^*} P(\omega_1)p(x|\omega_1)dx. \end{aligned}$$

Así, nun problema de clasificación binaria, o erro de Bayes pode interpretarse como a área baixo a curva de $P(\omega_2)p(x|\omega_2)$ na rexión de clasificación $\mathcal{R}_1^*$ máis a área baixo a curva de $P(\omega_1)p(x|\omega_1)$ na rexión de clasificación $\mathcal{R}_2^*$.

Se consideramos un clasificador distinto, D, as rexións de clasificación e, polo tanto, as fronteiras de decisión, serán distintas. Na Figura 1.4 represéntase unha segunda fronteira de decisión que delimitaría as rexións de clasificación asociadas a este outro clasificador, D. Do mesmo modo que para o clasificador de Bayes, o erro de clasificación para este clasificador D consiste na suma das áreas baixo as funcións discriminantes $P(\omega_2)p(x|\omega_2)$ e $P(\omega_1)p(x|\omega_1)$ nas rexións de clasificación $\mathcal{R}_1$ e $\mathcal{R}_2$, respectivamente. Polo tanto, ao considerar o clasificador D estamos desprazando a fronteira de decisión. Este desprazamento ten asociado un erro adicional, que se representa nunha cor máis escura na Figura 1.4.

A Figura 1.4 ilustra como empregando as probabilidades a posteriori reais como funcións discriminantes conseguimos alcanzar o mínimo erro de clasificación posible. Calquera outra fronteira de decisión, como a representada na figura, incrementará o erro de clasificación.

Capítulo 2

Métodos para combinar predicións de clasificadores

Un ensemble $\mathcal{D} = \{D_1, \ldots, D_L\}$ consiste nun conxunto de L clasificadores base que se combinan para realizar unha única predición, co obxectivo de mellorar a exactitude das predicións individuais. Á hora de deseñar un ensemble, Kuncheva [8] formula 4 cuestións ás que atender:

Manipulación do conxunto de datos. O conxunto de datos empregado para adestrar os clasificadores pode modificarse para asegurar que diversidade entre os clasificadores e garantir que non proporcionen as mesmas saídas para cada obxecto de entrada $\mathbf{x}$.

Selección de características. Cada clasificador pódese adestrar con todas as características dos obxectos ou soamente cun subconxunto delas. De novo, isto pode axudar a introducir diversidade no ensemble.

Clasificadores base. O ensemble pode estar formado por clasificadores iguais ou distintos. Ademais, estes pódense adestrar á vez ou un a continuación do outro.

Método de combinación. Por último, o método de combinación determina como se combinan as saídas dos clasificadores base.

Todas estas cuestións afectan ao rendemento do ensemble. Por exemplo, se as saídas dos clasificadores base son moi similares, será difícil que o método de combinación mellore a exactitude das predicións individuais. Por exemplo, os bosques aleatorios son un método de ensamblado que utiliza árbores de decisión como clasificadores base. Para introducir diversidade entre as árbores de decisión, cada unha adéstrase cun subconxunto de características diferente. Por outro lado, adaBoost adestra os clasificadores base de forma secuencial, utilizando os erros cometidos en cada etapa para condicionar o adestramento dos clasificadores das etapas posteriores. No Capítulo 4

estudaremos estes métodos de ensmablado en maior profundidade.

Neste capítulo e no seguinte, enfocarémonos en estudar os distintos métodos de combinación que se poden empregar para combinar as saídas dos clasificadores dun ensemble. En función da forma que tomen as saídas dos clasificadores, diferenciamos entre métodos de combinación de predicións e métodos de combinación de valores continuos. Dicimos que a saída dun clasificador é unha predición cando se trata dunha etiqueta de clase, mentres que é un valor continuo cando se trata dun número real, que pode ser interpretado como unha probabilidade ou unha medida de confianza na clase proposta.

Neste capítulo centrarémonos nos métodos de combinación de predicións. Comezaremos estudando o método máis sinxelo, o voto por maioría simple, e iremos avanzando cara métodos máis complexos, ata chegar ao método de combinación multinomial.

2.1. Introdución

Sexa $\mathbf{s} = [s_1, \dots, s_L]^T \in \Omega^L$ o vector que contén as predicións dos clasificadores base para un determinado $\mathbf{x} \in \mathbb{R}^n$. Dado $\mathbf{s}$, interésanos minimizar o erro de clasificación. Noutras palabras, queremos asignar a clase ω_k que maximice a probabilidade a posteriori $P(\omega_k|\mathbf{s})$, $k = 1, \dots, c$. Asumiremos que os clasificadores toman as súas decisións de forma independente condicionada á clase verdadeira, é dicir,

$$P(\mathbf{s}|\omega_k) = \prod_{i=1}^{L} P(s_i|\omega_k).$$

Deste xeito, podemos calcular a probabilidade a posteriori da clase ω_k a partir da probabilidade a priori $P(\omega_k)$ e da probabilidade condicional de s_i dado ω_k, $P(s_i|\omega_k)$, mediante a regra de Bayes

$$P(\omega_k|\mathbf{s}) = \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i=1}^{L} P(s_i|\omega_k). \tag{2.1}$$

Será de utilidade dividir o produto da expresión anterior en dúas partes en función do que propón cada un dos clasificadores. Para un obxecto dado, $\mathbf{x}$, enténdese que un clasificador propón a clase ω_k se a probabilidade de que a clase verdadeira sexa ω_k é máxima. Denotamos por I_+^k o conxunto de índices dos clasificadores que propoñen a clase ω_k e por I_-^k o conxunto de índices dos clasificadores que non propoñen a clase ω_k. Así, podemos escribir a probabilidade a posteriori da clase ω_k como

$$P(\omega_k|\mathbf{s}) = \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i \in I_+^k} P(s_i|\omega_k) \prod_{i \in I_-^k} P(s_i|\omega_k). \tag{2.2}$$

Esta descomposición seranos útil para definir as condicións de optimalidade para tres dos catro métodos de combinación de clasificadores que imos describir neste capítulo: voto por maioría, voto por maioría ponderado, Naïve Bayes e o método de combinación multinomial. Os tres primeiros baséanse na asunción da independenza condicional, mais diferéncianse na suposición sobre a exactitude individual dos clasificadores. O último, é o óptimo cando as saídas dos clasificadores son condicionalmente dependentes en base á clase verdadeira, ω_k. Se se cumpre a suposición, o respectivo combinador é óptimo no sentido de que garante o mínimo erro de Bayes.

Exactitude individual igual. Se $P(s_i = \omega_k|\omega_k) = p$ e $P(s_i = \omega_j|\omega_k) = \frac{1-p}{c-1}$, para calquera clasificador $i \in \{1, ..., L\}$, con $k, j \in \{1, ..., c\}, j \neq k$, entón a votación por maioría é o método de combinación óptimo.

Exactitude individual diferente. Se $P(s_i = \omega_k|\omega_k) = p_i$ e $P(s_i = \omega_j|\omega_k) = \frac{1-p_i}{c-1}$, para calquera $i \in \{1, ..., L\}$, $k, j \in \{1, ..., c\}, j \neq k$, entón o voto por maioría ponderado é o método de combinación óptimo.

Matriz de confusión diferente. Se $P(s_i = \omega_k|\omega_k) = p_{ijk}$, entón, o método de combinación óptimo é o de Naïve Bayes.

Dependencia condicional. No caso de que as saídas dos clasificadores sexan condicionalmente dependentes en base á clase verdadeira, ω_k, o combinador óptimo é o método de combinación multinomial.

Cada método de combinación obtense a partir do anterior cando se relaxa ou se elimina unha certa suposición. Non obstante, o prezo a pagar consiste en que se deben estimar máis parámetros. Na Sección 2.6 compararemos os métodos de combinación de clasificadores base que acabamos de mencionar e que explicaremos con máis detalle nas sucesivas seccións.

2.2. Voto por maioría

O método de voto por maioría é un dos métodos máis sinxelos para combinar as saídas dos clasificadores base. Existen tres variantes que se diferencian na forma de chegar a un consenso: por unanimidade, por maioría absoluta e por maioría simple. Polo xeral, cando se fala de voto por maioría, refírese ao voto por maioría simple, é dicir, a clase que recibe máis votos é a que se asigna ao obxecto.

Para resolver os empates, existen dúas posibilidades. A primeira, e máis simple, consiste en resolver o empate de forma arbitraria, por exemplo, asignando a clase coa menor etiqueta (no caso de que estas se poidan ordear). A segunda, consiste en engadir unha clase extra, ω_{c+1}, ao conxunto de clases Ω, e asignar esta clase ao obxecto en caso de empate.

Se as saídas dos clasificadores consisten nun vector c-dimensional $[d_{i,1}, \dots, d_{i,c}]^T,\ i = 1, \dots, L$, onde $d_{i,j} = 1$ se D_i propón a clase ω_j para $\mathbf{x}$ e $d_{i,j} = 0$ en caso contrario, podemos expresar a función de soporte para a clase ω_k do método de voto por maioría como

$$\mu_k(\mathbf{x}) = \sum_{i=1}^{L} d_{i,k}. \tag{2.3}$$

O valor da función de soporte de cada clase en (2.3) será o número de clasificadores que propoñen a clase ω_k para o obxecto $\mathbf{x}$.

2.2.1. Exactitude do voto por maioría

A simplicidade do voto por maioría fai que sexa doado calcular a exactitude do método a partir da exactitude dos clasificadores base. Como veremos na Sección 2.2.2, unha das condicións de optimalidade do voto por maioría é que a exactitude individual dos clasificadores base sexa igual. Asumimos que:

- O número de clasificadores L, é impar.
- A probabilidade de que un clasificador propoña a clase verdadeira é p para calquera $\mathbf{x} \in \mathbb{R}^n$.
- As decisións dos clasificadores son independentes.

Partindo disto, imos fundamentar a idea intuitiva de que podemos esperar unha mellora na exactitude cantos máis clasificadores utilicemos no ensemble. Para simplificar o razoamento, imos considerar un problema de clasificación con dúas clases, $c = 2$. Neste caso, a maioría simple coincide coa absoluta e non hai posibilidade de empate. A partir da expresión (2.3), dado $\mathbf{x}$, o ensemble asignará a clase ω_k se

$$\sum_{i=1}^{L} d_{i,k} > \left\lfloor \frac{L}{2} \right\rfloor.$$

Así, o método de voto por maioría asigna correctamente a clase cando, polo menos, $\lfloor L/2 \rfloor + 1$ clasificadores propoñen a clase verdadeira. Polo tanto, a exactitude do ensemble será

$$p_{\text{ens}} = \sum_{m=\lfloor L/2 \rfloor + 1}^{L} \binom{L}{m} p^m (1-p)^{L-m}. \tag{2.4}$$

Na Táboa 2.1 amósanse os valores da exactitude do ensemble para diferentes valores de L e p. Nela, pódese observar o comportamento que se mencionaba anteriormente: a exactitude do ensemble aumenta cantos máis clasificadores se engaden ao ensemble.

Táboa 2.1: Exactitude dun ensemble con voto por maioría con L clasificadores base independentes e exactitude individual p.

	$L = 3$	$L = 5$	$L = 7$	$L = 9$
$p = 0.6$	0.6480	0.6826	0.7102	0.7334
$p = 0.7$	0.7840	0.8369	0.8740	0.9012
$p = 0.8$	0.8960	0.9421	0.9667	0.9804
$p = 0.9$	0.9720	0.9914	0.9973	0.9991

Non obstante este comportamento depende da exactitude individual dos clasificadores. Se $p < 0.5$, entón $p_{\text{ens}} < p$ para $L > 1$. O seguinte teorema, coñecido como Teorema do Xurado de Condorcet, que pode consultarse en [8], formaliza esta idea no caso de que a asignación da clase se realice mediante maioría absoluta.

Teorema 2.1. *Sexa $\mathcal{D}$ un ensemble de L clasificadores. Asumindo,*

1. *A probabilidade de que un clasificador propoña a clase axeitada é p para calquera $\mathbf{x} \in \mathbb{R}^n$.*

2. *As decisións dos clasificadores son independentes.*

Entón:

1. *Se $p > 0.5$, a exactitude do ensemble con mecanismo de voto por maioría absoluta é monótona crecente con L:*
$$p_{ens} \xrightarrow{L\to\infty} 1$$

2. *Se $p < 0.5$, a exactitude do ensemble con mecanismo de voto por maioría absoluta é monótona decrecente con L:*
$$p_{ens} \xrightarrow{L\to\infty} 0$$

3. *Se $p = 0.5$, a exactitude do ensemble con mecanismo de voto por maioría absoluta é constante, $p_{ens} = 0.5$ para calquera L.*

Lam e Suen [9] analizaron o efecto de engadir e eliminar clasificadores cando L é par. Observaron que, no caso de que L sexa par, é necesario que $p > \frac{n/2}{n+1}$ para que p_{ens} sexa monótona crecente con L. Por outro lado, Shapley e Grofman [14] probaron que o resultado é válido incluso cando a exactitude individual dos clasificadores é distinta, sempre e cando a distribución destas sexa simétrica sobre a media.

2.2.2. Optimalidade do voto por maioría

O voto por maioría é o método de combinación óptimo cando a exactitude individual dos clasificadores é igual, a probabilidade de erro restante se distribúe uniformemente entre as clases restantes e a probabilidade a priori de cada clase é igual. O seguinte teorema proba isto de maneira formal.

Teorema 2.2. *Sexa $\mathcal{D}$ un ensemble de L clasificadores. Asumindo,*

1. *Os clasificadores asignan as etiquetas de forma independente condicionada cada clase.*
2. *A exactitude individual dos clasificadores é igual, $P(s_i = \omega_k|\omega_k) = p$ para calquera clasificador $i = 1, \dots, L$ e calquera clase ω_k.*
3. *A probabilidade de clasificar incorrectamente un obxecto distribúese uniformemente entre as clases restantes, $P(s_i = \omega_j|\omega_k) = \frac{1-p}{c-1}$ para calquera $i \in \{1, \dots, L\}$, $k, j \in \{1, \dots, c\}, j \neq k$.*
4. *A probabilidade a priori de cada clase, $P(\omega_k)$, é igual para todas as clases.*

Entón, o método de voto por maioría simple é o método de combinación óptimo.

Demostración. Imos demostrar que o método de voto por maioría maximiza a probabilidade a posteriori da clase verdadeira, $P(\omega_k|\mathbf{s})$. Para iso, partimos da expresión (2.2).

$$
\begin{aligned}
P(\omega_k|\mathbf{s}) &= \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i \in I_+^k} P(s_i|\omega_k) \prod_{i \in I_-^k} P(s_i|\omega_k) \\
&= \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i \in I_+^k} p \prod_{i \in I_-^k} \frac{1-p}{c-1} \qquad (2.5) \\
&= \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i \in I_+^k} p \prod_{i \in I_-^k} \frac{1-p}{c-1} \frac{\prod_{i \in I_+^k} \frac{1-p}{c-1}}{\prod_{i \in I_+^k} \frac{1-p}{c-1}} \\
&= \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i \in I_+^k} \frac{p(c-1)}{1-p} \prod_{i=1}^{L} \frac{1-p}{c-1} \qquad (2.6) \\
&= \frac{P(\omega_k)}{P(\mathbf{s})} \left(\frac{p(c-1)}{(1-p)} \right)^{|I_+^k|} \left(\frac{1-p}{c-1} \right)^{L}.
\end{aligned}
$$

En (2.5) multiplicamos $\prod_{i \in I_+^k} \frac{1-p}{c-1}$ no numerador e no denominador co obxectivo de que en (2.6) o segundo termo non dependa da clase. Tomando logaritmos

$$
\log(P(\omega_k|\mathbf{s})) = \log \left(\frac{(1-p)^L}{P(\mathbf{s})(c-1)^L} \right) + \log(P(\omega_k)) + |I_+^k| \log \left(\frac{p(c-1)}{1-p} \right). \qquad (2.7)
$$

Os termos que non dependen do método de combinación non son relevantes para maximizar a

probabilidade a posteriori. Polo tanto, dividindo por $\log(\frac{p(c-1)}{1-p})$ e eliminando o primeiro termo de (2.7), que non depende da clase nin do número de votos, obtemos

$$\mu_k(\mathbf{x}) = \frac{\log(P(\omega_k))}{\log(\frac{p(c-1)}{1-p})} + |I_+^k|.$$

Tendo en conta que $|I_+^k|$ é o número de votos para a clase ω_k e que a probabilidade a priori de cada clase é igual para todas as clases, escoller a clase que máis votos recibe é equivalente a escoller a clase que maximiza a probabilidade a posteriori. □

Se relaxamos a hipótese de que a probabilidade a priori de cada clase, $P(\omega_k)$, é igual para todas as clases, esta pode influír na decisión do clasificador. Para solucionar isto, pódese engadir unha constante de clase $\frac{\log(P(\omega_k))}{\log(\frac{p(c-1)}{1-p})}$ á función de decisión, mais isto engadiría c parámetros ao método, xa que sería necesario estimar $P(\omega_k)$. Non obstante, isto convertiría ao método de votación por maioría nun método con parámetros adestrables, algo que debilita unha das súas principais vantaxes.

2.3. Voto por maioría ponderado

É razoable pensar que os clasificadores do ensemble non teñan todos a mesma exactitude. Neste caso, sería conveniente darlle máis peso na decisión final a aqueles clasificadores con mellor exactitude. Así, se as saídas dos clasificadores consisten nun vector c-dimensional $[d_{i,1}, \ldots, d_{i,c}]$, $i = 1, \ldots, L$, onde $d_{i,j} = 1$ se D_i porpón a clase ω_j para $\mathbf{x}$ e $d_{i,j} = 0$ en caso contrario, a función de soporte para a clase ω_k do método de voto ponderado por maioría é

$$\mu_k(\mathbf{x}) = \sum_{i=1}^{L} b_i d_{i,k} + \ln(c-1) \sum_{i=1}^{L} d_{i,k}, \tag{2.8}$$

onde b_i é o coeficiente para o clasificador D_i. O valor da función de soporte de cada clase en (2.8) será a suma dos pesos dos clasificadores que propoñen a clase ω_k para o obxecto $\mathbf{x}$.

Exemplo 2.3. Consideremos un problema de clasificación con dúas clases ($c = 2$) e un ensemble de 5 clasificadores $D_1, \ldots, D_5$, con exactitudes (0.9, 0.9, 0.6, 0.6, 0.6). Supoñamos que os clasificadores son independentes. Para que o método de voto por maioría ponderado propoña a clase correcta, precisamos que tres ou máis clasificadores propoñan a clase correcta. A probabilidade de que isto suceda pódese descompoñer da seguinte maneira:

- Polo menos os tres clasificadores con exactitude 0.6 propoñen a clase correcta:

$$\begin{aligned} & 0.6^3 \cdot 0.9^2 + 0.6^3 \cdot \binom{2}{1} \cdot 0.9 \cdot 0.1 + 0.6^3 \cdot 0.1^2 \\ &= 0.6^3 \cdot (0.9^2 + 2 \cdot 0.9 \cdot 0.1 + 0.1^2) \\ &= 0.6^3 \cdot (0.9 + 0.1)^2 = 0.6^3. \end{aligned}$$

- Dous clasificadores con exactitude 0.9 e un ou dous dos tres clasificadores con exactitude 0.6 propoñen a clase correcta:

$$\begin{aligned} & 0.9^2 \cdot \binom{3}{2} \cdot 0.6^2 \cdot 0.4 + 0.9^2 \cdot \binom{3}{1} \cdot 0.6 \cdot 0.4^2 \\ &= 3 \cdot 0.9^2 \cdot 0.6 \cdot 0.4 \cdot (0.6 + 0.4) \\ &= 3 \cdot 0.9^2 \cdot 0.6 \cdot 0.4. \end{aligned}$$

- Un dos clasificadores con exactitude 0.9 e os dous clasificadores con exactitude 0.6 propoñen a clase correcta:

$$\binom{2}{1} 0.9 \cdot 0.1 \cdot \binom{3}{2} \cdot 0.6^2 \cdot 0.4 = 6 \cdot 0.9 \cdot 0.1 \cdot 0.6^2 \cdot 0.4.$$

Desta forma, se sumamos as probabilidades dos tres casos obtemos que a probabilidade de que o método de voto por maioría propoña a clase correcta é

$$p_{\text{ens}} = 0.6^3 + 3 \cdot 0.9^2 \cdot 0.4 \cdot 0.6 + 6 \cdot 0.9 \cdot 0.1 \cdot 0.6^2 \cdot 0.4 \approx 0.877.$$

Claramente, é preferible quedarse con D_1 ou D_2, antes que co ensemble cos 5 clasificadores. Non obstante, supoñamos que asignamos os pesos $\mathbf{b} = (1/3, 1/3, 1/9, 1/9, 1/9)$ aos clasificadores $D_1, \ldots, D_5$. Nun problema de clasificación binaria a expresión das funcións de soporte para o método de voto por maioría ponderado simplifícase a

$$\mu_k(\mathbf{x}) = \sum_{i=1}^{L} b_i d_{i,k},$$

xa que $\log(2-1) = 0$. Deste modo, para que o método de voto por maioría ponderado propoña a clase correcta chega con que a suma dos pesos das dos clasificadores que propoñen a clase correcta sexa maior que 0.5. Podemos descompoñer a probabilidade de que isto suceda dun xeito semellante ao caso anterior:

- Polo menos os dous clasificadores con exactitude 0.9 propoñen a clase correcta:

$$0.9^2 \cdot 0.6^3 + 0.9^2 \cdot \binom{3}{1} \cdot 0.6 \cdot 0.4^2 + 0.9^2 \cdot \binom{3}{2} \cdot 0.6^2 \cdot 0.4 + 0.9^2 \cdot 0.4^3$$
$$= 0.9^2 \cdot (0.6 + 0.4)^3 = 0.9^2.$$

- Un dos clasificadores con exactitude 0.9 e os tres clasificadores con exactitude 0.6 propoñen a clase correcta:

$$\binom{2}{1} \cdot 0.9 \cdot 0.1 \cdot 0.6^3 = 2 \cdot 0.9 \cdot 0.1 \cdot 0.6^3.$$

- Un dos clasificadores con exactitude 0.9 e dous dos tres clasificadores con exactitude 0.6 propoñen a clase correcta:

$$\binom{2}{1} \cdot 0.9 \cdot 0.1 \cdot \binom{3}{2} \cdot 0.6^2 \cdot 0.4 = 6 \cdot 0.9 \cdot 0.1 \cdot 0.6^2 \cdot 0.4.$$

Sumando os tres casos, obtemos:

$$p_{\text{ens}}^{\mathbf{b}} = 0.9^2 + 6 \cdot 0.9 \cdot 0.1 \cdot 0.6^2 \cdot 0.4 + 2 \cdot 0.9 \cdot 0.1 \cdot 0.6^3 \approx 0.927. \tag{2.9}$$

Desta maneira, asignando un maior peso aos clasificadores con mellor exactitude conseguimos que o método de voto por maioría ponderado supere a exactitude dos clasificadores individuais. Cambiando os pesos, poderíamos obter unha exactitude superior ou inferior, polo que a elección dos pesos é crucial para o rendemento do ensemble. Na seguinte sección, veremos como obter os pesos óptimos.

2.3.1. Optimalidade do voto por maioría ponderado

O voto por maioría ponderado obtense ao relaxar a hipótese de que a exactitude individual dos clasificadores é igual. Desta maneira, o método de voto por maioría é un caso particular do voto por maioría ponderado cando as exactitudes dos clasificadores base son iguais.

Teorema 2.4. *Sexa $\mathcal{D}$ un ensemble de L clasificadores. Asumindo,*

1. *Os clasificadores asignan as etiquetas de forma independente condicionada cada clase.*
2. *A exactitude individual dos clasificadores é diferente, $P(s_i = \omega_k|\omega_k) = p_i$ para calquera clasificador $i = 1, \dots, L$ e calquera clase ω_k.*
3. *A probabilidade de clasificar incorrectamente un obxecto distribúese uniformemente entre as clases restantes, $P(s_i = \omega_j|\omega_k) = \frac{1-p_i}{c-1}$ para calquera $i \in \{1, \dots, L\}$, $k, j \in \{1, \dots, c\}, j \neq k$.*

4. A probabilidade a priori de cada clase, $P(\omega_k)$, é igual para todas as clases.

Entón, o método de voto por maioría ponderado é o método de combinación óptimo con pesos

$$b_i = \log\left(\frac{p_i}{1-p_i}\right).$$

Demostración. Seguindo o mesmo razoamento que co método de voto por maioría, imos demostrar que o método de voto por maioría ponderado maximiza a probabilidade a posteriori da clase verdadeira, $P(\omega_k|\mathbf{s})$. Partindo da expresión (2.2),

$$\begin{aligned}
P(\omega_k|\mathbf{s}) &= \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i\in I_+^k} P(s_i|\omega_k) \prod_{i\in I_-^k} P(s_i|\omega_k) \\
&= \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i\in I_+^k} p_i \prod_{i\in I_-^k} \frac{1-p_i}{c-1} \\
&= \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i\in I_+^k} \frac{p_i(c-1)}{1-p_i} \prod_{i=1}^{L} \frac{1-p_i}{c-1} \\
&= \frac{1}{P(\mathbf{s})} \prod_{i=1}^{L} \frac{1-p_i}{c-1} P(\omega_k) \prod_{i\in I_+^k} \frac{p_i(c-1)}{1-p_i}.
\end{aligned}$$

Tomando logaritmos

$$log(P(\omega_k|\mathbf{s})) = log\left(\frac{\prod_{i=1}^{L}(1-p_i)}{P(\mathbf{s})(c-1)^L}\right) + \log(P(\omega_k)) + \sum_{i\in I_+^k} log\left(\frac{p_i}{1-p_i}\right) + |I_+^k|\,\log(c-1).$$

O primeiro termo non influe na decisión da clase, polo que pódese ignorar. Se denotamos por $b_i = \log(\frac{p_i}{1-p_i})$, obtemos

$$\mu_k(\mathbf{x}) = \log(P(\omega_k)) + \sum_{i\in I_+^k} b_i + |I_+^k|\,\log(c-1) \tag{2.10}$$

Posto que a probabilidade a priori de cada clase é igual para todas as clases, o primeiro termo de (2.10) non influe na decisión da clase. Polo tanto, a seguinte función maximiza a probabilidade a posteriori de ω_k

$$\mu_k(\mathbf{x}) = \sum_{i\in I_+^k} b_i + |I_+^k|\,\log(c-1). \tag{2.11}$$

□

Na literatura, con frecuencia simplifícase a expresión das funcións de soporte para o método

de voto por maioría ponderado a

$$\mu_k^*(\mathbf{x}) = \sum_{i=1}^{L} b_i d_{i,k}, \tag{2.12}$$

independente do número de clases. Na práctica, non é habitual que se cumpran as condicións de optimalidade do método de voto por maioría ponderada, polo que non se garante que o ensemble teña maior exactitude utilizando como función de soporte a expresión (2.11) en lugar de (2.12).

2.4. Naïve Bayes

Dado un clasificador do ensemble, D_i, para que o método de voto por maioría ponderado sexa óptimo é necesario asumir que $P(s_i = \omega_k|\omega_k) = p_i$ para calquera clase ω_k. Noutras palabras, é necesario asumir que a exactitude individual dos clasificadores é igual para todas as clases. Na práctica, estimar a exactitude do clasificador sen atender á exactitude por clase pode levar a unha sobreestimación da exactitude do clasificador [1, 12]. O método de voto por maioría ponderado é sensible a esta sobreestimación, xa que os pesos que se asignan aos clasificadores dependen das exactitudes individuais estimadas.

O método de Naïve Bayes obtense a partir de eliminar a suposición de que a exactitude individual dos clasificadores é igual no contexto que definimos na Sección 2.1. Deste xeito, a función de soporte para a clase ω_k do método de Naïve Bayes é

$$\mu_k(\mathbf{x}) = P(\omega_k) \prod_{i=1}^{L} P(s_i|\omega_k). \tag{2.13}$$

As probabilidades $P(s_i|\omega_k)$ pódense estimar a partir das matrices de confusión dos clasificadores base.

2.4.1. Optimalidade de Naïve Bayes

Teorema 2.5. *Sexa $\mathcal{D}$ un ensemble de L clasificadores. Asumindo, que os clasificadores asignan as etiquetas de forma independente condicionada cada clase e que as clases son equiprobables, entón, o método de Naïve Bayes é o método de combinación óptimo.*

Demostración. Partimos directamente de (2.2)

$$P(\omega_k|\mathbf{s}) = \frac{P(\omega_k)}{P(\mathbf{s})} \prod_{i=1}^{L} P(s_i|\omega_k).$$

Eliminando $P(\mathbf{s})$ da expresión, obtense a función de soporte do método de Naïve Bayes

$$\mu_k(\mathbf{x}) = P(\omega_k) \prod_{i=1}^{L} P(s_i|\omega_k).$$

□

2.4.2. Estimación dos parámetros de Naïve Bayes

Para implementar o método de Naïve Bayes, é necesario calcular unha matriz de confusión de dimensión $c \times c$ para cada clasificador base, D_i, a partir do conxunto de datos. Denotaremos por cm^i a matriz de confusión do clasificador D_i. O elemento $cm^i_{k,s}$ da matriz de confusión do clasificador D_i é o número de obxectos do conxunto de datos da clase ω_k que foron clasificados como ω_s polo clasificador D_i. Se N_k é o número de obxectos da clase ω_k no conxunto de datos, podemos empregar $cm^i_{k,s_i}/N_k$ como estimación da probabilidade $P(s_i|\omega_k)$ e N_k/N como estimación da probabilidade a priori $P(\omega_k)$. Así, podemos calcular a función de soporte para a clase ω_k do método de Naïve Bayes como

$$\mu_k(\mathbf{x}) = \frac{N_k}{N} \prod_{i=1}^{L} \frac{cm^i_{k,s_i}}{N_k} = \frac{1}{N} \frac{1}{N_k^{L-1}} \prod_{i=1}^{L} cm^i_{k,s_i}. \tag{2.14}$$

Podemos eliminar o termo $1/N$ da expresión (2.14), xa que é igual para todas as clases e non inflúe na decisión da clase. Desta maneira, chegamos a

$$\mu_k(\mathbf{x}) = \frac{1}{N_k^{L-1}} \prod_{i=1}^{L} cm^i_{k,s_i}. \tag{2.15}$$

Exemplo 2.6. Consideremos un problema con $L = 2$ clasificadores, D_1 e D_2, e $c = 3$ clases. Sexa $N = 20$ o número de obxectos do conxunto de datos, dos cales 8 pertencen á clase ω_1, 9 á clase ω_2 e 3 á clase ω_3. As matrices de confusión dos clasificadores son

$$cm^1 = \begin{bmatrix} 6 & 2 & 0 \\ 1 & 8 & 0 \\ 1 & 0 & 2 \end{bmatrix}, \qquad cm^2 = \begin{bmatrix} 4 & 3 & 1 \\ 3 & 5 & 1 \\ 0 & 0 & 3 \end{bmatrix}.$$

Supoñamos que $D_1(\mathbf{x}) = s_1 = \omega_2$ e $D_2(\mathbf{x}) = s_2 = \omega_1$, con $\mathbf{x} \in \mathbb{R}^n$. Utilizando (2.15),

$$\mu_1(\mathbf{x}) = \frac{1}{8} \cdot 2 \cdot 4 = 1,\ \mu_2(\mathbf{x}) = \frac{1}{9} \cdot 8 \cdot 3 = \frac{8}{3},\ \mu_3(\mathbf{x}) = \frac{1}{3} \cdot 0 \cdot 0 = 0.$$

Posto que $\mu_2(\mathbf{x})$ é o maior valor dos tres, o ensemble asignará a clase ω_2 ao obxecto $\mathbf{x}$.

No momento no que a estimación de $P(s_i|\omega_k)$ é 0, automaticamente $\mu_k(\mathbf{x}) = 0$. Titterington et al. [15] realizaron un estudo comparativo con varios métodos de clasificación con variables categóricas entre os que incluíron o clasificador de Naïve Bayes. O conxunto de datos que empregaron contiña numerosos valores ausentes, polo que adaptaron o clasificador de Naïve Bayes para solucionar este problema.

Os clasificadores base pódense pensar como variables categóricas e os valores nulos das matrices de confusión, como valores ausentes. Así, poderíamos utilizar a técnica proposta por Titterington et al. para estimar as probabilidades $P(s_i|\omega_k)$:

$$\hat{P}(s_i|\omega_k) = \frac{cm^i_{k,s_i} + \frac{1}{c}}{N_k + 1}. \tag{2.16}$$

Deste xeito, cando $cm^i_{k,s_i} = 0$, o numerador de $\hat{P}(s_i|\omega_k)$ é distinto de 0 e xa non anula automaticamente a función de soporte $\mu_k(\mathbf{x})$. A continuación, vexamos como inflúe esta modificación no exemplo 2.6.

Exemplo 2.7. Consideremos as mesmas matrices de confusión do exemplo 2.6 e que $D_1(\mathbf{x}) = s_1 = \omega_2$ e $D_2(\mathbf{x}) = s_2 = \omega_1$. Utilizando a Ecuación (2.16) para estimar as probabilidades $P(s_i|\omega_k)$,

$$\mu_1(\mathbf{x}) = \frac{N_1}{N}\left(\frac{cm^1_{1,2} + \frac{1}{3}}{N_1 + 1}\right)\left(\frac{cm^2_{1,1} + \frac{1}{3}}{N_1 + 1}\right) = \frac{8}{20}\left(\frac{2 + \frac{1}{3}}{8 + 1}\right)\left(\frac{4 + \frac{1}{3}}{8 + 1}\right) \approx 0.050.$$
$$\mu_2(\mathbf{x}) = \frac{N_2}{N}\left(\frac{cm^1_{2,2} + \frac{1}{3}}{N_2 + 1}\right)\left(\frac{cm^2_{2,1} + \frac{1}{3}}{N_2 + 1}\right) = \frac{9}{20}\left(\frac{8 + \frac{1}{3}}{9 + 1}\right)\left(\frac{3 + \frac{1}{3}}{9 + 1}\right) \approx 0.125.$$
$$\mu_3(\mathbf{x}) = \frac{N_3}{N}\left(\frac{cm^1_{3,2} + \frac{1}{3}}{N_3 + 1}\right)\left(\frac{cm^2_{3,1} + \frac{1}{3}}{N_3 + 1}\right) = \frac{3}{20}\left(\frac{0 + \frac{1}{3}}{3 + 1}\right)\left(\frac{0 + \frac{1}{3}}{3 + 1}\right) \approx 0.001.$$

De novo, o método asigna a clase ω_2 ao obxecto $\mathbf{x}$, pero a probabilidade de que o obxecto pertenza á clase ω_3 xa non é 0.

2.5. Método de combinación multinomial (BKS)

Ao relaxar a suposición de que os clasificadores asignan as etiquetas de forma independente condicionada cada clase, xa non podemos expresar a probabilidade a posteriori da clase ω_k como o produto das probabilidades condicionadas $P(s_i|\omega_k)$.

O método de combinación multinomial, ou BKS polas súas siglas en inlés "Behaviour Knowledge Space", consiste nun método que estima directamente $P(\omega_k|\mathbf{s})$, sen necesidade de asumir que os clasificadores asignan as etiquetas de forma independente condicionada cada clase.

Para implementar o método de combinación BKS, constrúese unha táboa de busca a partir

dun conxunto de datos $\mathbf{Z}$. A táboa ten c^L filas, unha por cada posible combinación de saídas dos clasificadores. Para cada obxecto do conxunto de datos $\mathbf{z}_j \in \mathbf{Z}$, calcúlase a saída dos clasificadores, $\mathcal{D}(\mathbf{z}_j) = \mathbf{s} = [s_1, ..., s_L]^T$. O obxecto $\mathbf{z}_j$ colócase na fila da táboa que corresponde a $\mathbf{s}$. Unha vez colocados na táboa todos os obxectos do conxunto de datos, asígnase a clase maioritaria a cada fila dos obxectos que a compoñen.

Deste modo, dado un novo obxecto $\mathbf{x}$, calcúlase a saída dos clasificadores, $\mathcal{D}(\mathbf{x}) = \mathbf{s}$, e búscase a fila da táboa que corresponde a $\mathbf{s}$. A clase asignada ao obxecto $\mathbf{x}$ será a clase maioritaria dos obxectos que compoñen a fila da táboa.

Os empates resólvense de maneira arbitraria e, no caso de que existan celas baleiras, aígnase unha clase aleatoria ou utilízase outro método de ensamblado, como o voto por maioría. A táboa 2.2 amosa un exemplo dunha táboa BKS.

Táboa 2.2: A táboa BKS é unha táboa de busca. Os elementos $\mathbf{z}_j$ colócanse na nela indexados por $\mathcal{D}(\mathbf{z_j}) = \mathbf{s}$.

	s_1	s_2	...	s_{L-1}	s_L	$\{\mathbf{z}_j \in \mathbf{Z}/\mathcal{D}(\mathbf{z}_j) = \mathbf{s}\}$
$\mathbf{s}^1$	ω_1	ω_1	...	ω_1	ω_1	$\{\mathbf{z}_j \in \mathbf{Z}/\mathcal{D}(\mathbf{z}_j) = \mathbf{s}^1\}$
$\mathbf{s}^2$	ω_1	ω_1	...	ω_1	ω_2	$\{\mathbf{z}_j \in \mathbf{Z}/\mathcal{D}(\mathbf{z}_j) = \mathbf{s}^2\}$
$\mathbf{s}^3$	ω_1	ω_1	...	ω_2	ω_1	$\{\mathbf{z}_j \in \mathbf{Z}/\mathcal{D}(\mathbf{z}_j) = \mathbf{s}^3\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{s}^{c^L-1}$	ω_c	ω_c	...	ω_c	ω_{c-1}	$\{\mathbf{z}_j \in \mathbf{Z}/\mathcal{D}(\mathbf{z}_j) = \mathbf{s}^{c^L-1}\}$
$\mathbf{s}^{c^L}$	ω_c	ω_c	...	ω_c	ω_c	$\{\mathbf{z}_j \in \mathbf{Z}/\mathcal{D}(\mathbf{z}_j) = \mathbf{s}^{c^L}\}$

Exemplo 2.8. Consideremos un problema con 3 clasificadores e 2 clases, polo que a táboa terá $2^3 = 8$ entradas. Supoñamos que o conxunto de datos, $\mathbf{Z}$ ten 20 obxectos, dos cales os 10 primeiros serán da clase ω_1 e os 10 últimos, da clase ω_2. Sexa a táboa 2.3 a táboa BKS para este problema.

Táboa 2.3: Exemplo dunha táboa BKS con 3 clasificadores 2 clases. Para determinar a clase das celas baleiras utilízase o método do voto por maioría.

$\mathbf{s}$				Número de elementos da clase:		
s_1	s_2	s_3		ω_1	ω_2	Clase
ω_1	ω_1	ω_1	$\{z_1, z_2, z_8, z_9, z_{10}\}$	5	0	ω_1
ω_1	ω_1	ω_2	$\{z_3, z_4, z_{13}, z_{14}, z_{15}\}$	2	3	ω_2
ω_1	ω_2	ω_1	$\{z_7\}$	1	0	ω_1
ω_1	ω_2	ω_2	$\{\}$	0	0	ω_2
ω_2	ω_1	ω_1	$\{\}$	0	0	ω_1
ω_2	ω_1	ω_2	$\{z_5, z_6, z_{20}\}$	2	1	ω_1
ω_2	ω_2	ω_1	$\{z_{18}, z_{19}\}$	0	2	ω_2
ω_2	ω_2	ω_2	$\{z_{11}, z_{12}, z_{16}, z_{17}\}$	0	4	ω_2

Deste modo, se os clasificadores base producen a saída $\mathbf{s} = [s_1, s_2, s_3]^T = [\omega_1, \omega_1, \omega_2]^T$, o

método de BKS asignará a clase ω_2 ao obxecto, sen importar o número de clasificadores que propuxeron cada clase. Por outro lado, posto que no conxunto de datos non existen elementos que produzan a saída $\mathbf{s} = [\omega_1, \omega_2, \omega_2]^T$, o método de BKS pode utilizar o método de voto por maioría para asignar a clase ω_2.

Dende o punto de vista da implementación, o método BKS é equivalente ao método dos k veciños máis próximos no espazo das saídas dos clasificadores, Ω^L. Neste caso, o concepto de distancia substitúese pola coincidencia exacta. Dada unha saída dos clasificadores base, $\mathbf{s}$, para un obxecto $\mathbf{x}$, a clase asignada será a clase máis representativa dos elementos de $\mathbf{Z}$ que produzan a mesma saída dos clasificadores base.

O método de combinación BKS é óptimo cando existe algún tipo de dependencia entre os clasificadores base. Porén, polo xeral é complicado obter unha estimación real das probabilidades a posteriori para as c^L combinacións de $\mathbf{s}$. Deste xeito, no caso de non dispoñer un conxunto de datos suficientemente grande, o método de combinación BKS pode non ser o máis axeitado.

2.6. Comparación dos métodos de combinación

A relaxación progresiva das asuncións dá lugar a que os métodos de combinación teñan un alcance de optimalidade anidado. A ampliación do alcance de optimalidade págase adquirindo máis parámetros axustables. A Táboa 2.4 amosa os alcances de optimalidade e o número de parámetros axustables para cada combinador.

Táboa 2.4: Expectativas de optimalidade (denotado por cadrados negros) e número de parámetros adestrables dos catro métodos de combinación para un problema con L clasificadores e c clases.

Método	1	2	3	4	Número de parámetros
Voto por maioría	■	-	-	-	-
Voto por maioría ponderado	■	■	-	-	L
Naïve Bayes	■	■	■	-	$L \cdot c^2$
BKS	■	■	■	■	c^L

Cabeceiras das columnas:
1. p igual.
2. p_i específica para cada clasificador.
3. Matriz de confusión enteira.
4. Non se require independencia.

O feito de que un método de combinación teña un alcance de optimalidade maior non implica que sexa o mellor método de combinación. Na práctica, o éxito dun método de combinación dependerá tanto dos supostos como da dispoñibilidade de datos suficientes para facer estimacións

fiables dos parámetros. Combinadores non óptimos pero máis robustos poden funcionar mellor que o combinador óptimo.

No Apéndice A utilízase un exemplo práctico para exemplificar a influencia dos supostos e da fiabilidade da estimación dos parámetros.

Capítulo 3

Métodos para combinar saídas continuas de clasificadores

3.1. Introdución

Neste capítulo consideraremos que as saídas dos clasificadores son continuas. Estas pódense interpretar como os graos de soporte que os clasificadores dan a cada clase para unha determinada entrada $\mathbf{x}$. As dúas formas máis comúns de interpretar estes graos de soporte son como confianza nas clases propostas ou como estimacións das probabilidades a posteriori das clases.

Sexa $\mathbf{x} \in \mathbb{R}^n$ un vector de características e $\Omega = \{\omega_1, \ldots, \omega_c\}$ o conxunto de clases. Cada clasificador D_i do ensemble $\mathcal{D} = \{D_1, \ldots, D_L\}$ proporciona como saída c graos de soporte, un para cada clase. Podemos supoñer que os c graos de soporte pertencen ao intervalo $[0, 1]$,

$$D_i : \mathbb{R}^n \to [0, 1]^c.$$

Denotaremos por $d_{i,j}(\mathbf{x})$ o grao de soporte que o clasificador D_i proporciona á hipótese de que o obxecto $\mathbf{x}$ pertenza á clase ω_j. Deste xeito, canto maior sexa o valor de $d_{i,j}(\mathbf{x})$, máis confianza ten o clasificador D_i en que o obxecto $\mathbf{x}$ pertenza á clase ω_j. As saídas dos L clasificadores do ensemble pódense organizar nunha matriz que denominaremos perfil de decisión, $DP(\mathbf{x})$, amosada na Figura 3.1.

Os métodos descritos neste capítulo empregan a matriz $DP(\mathbf{x})$ para combinar as saídas do ensemble e asignar unha clase ou proporcionar os graos de soporte para cada clase para o obxecto $\mathbf{x}$. Non obstante, esta matriz pódese empregar de dúas formas diferentes. En primeiro lugar, podemos construír un grao de soporte para cada clase a partir das columnas de $DP(\mathbf{x})$

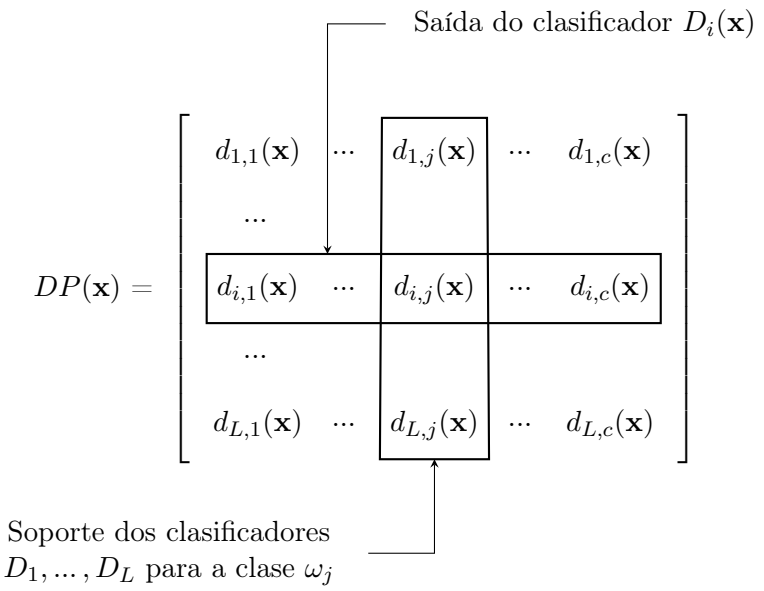


Figura 3.1: Perfil de decisión para unha entrada $\mathbf{x}$.

mediante expresións alxebraicas como unha media ou un produto. Alternativamente, podemos considerar a matriz $DP(\mathbf{x})$ como un novo conxunto de datos. Deste xeito, cada $d_{i,j}(\mathbf{x})$ sería unha característica dun novo espazo de características que denominaremos espazo intermedio de características. A decisión final realizaríaa outro clasificador que utiliza este espazo intermedio de características como entrada.

3.2. Métodos de combinación non adestrables

Os métodos de combinación non adestrables calculan os graos de soporte de cada clase ω_j a partir das L entradas da j-ésima columna da matriz $DP(\mathbf{x})$. A función de soporte pódese expresar de forma xenérica como:

$$\mu_j(\mathbf{x}) = \mathcal{F}(d_{1,j}(\mathbf{x}), \ldots, d_{L,j}(\mathbf{x})),$$

onde $\mathcal{F}$ é unha función que combina os graos de soporte dos clasificadores. A clase de $\mathbf{x}$ vén determinada polo índice j que maximiza $\mu_j(\mathbf{x})$. $\mathcal{F}$ pódese escoller de varias maneiras:

- **Media**:

$$\mu_j(\mathbf{x}) = \frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x}).$$

- **Mínimo/Máximo/Mediana**, por exemplo, para o mínimo:

$$\mu_j(\mathbf{x}) = \min_{i=1,\dots,L}\{d_{i,j}(\mathbf{x})\}.$$

- **Media recortada**: dado un determinado K, ordéanse os L graos de soporte para cada clase e calcúlase unha media eliminando os $K/2\,\%$ valores máis altos e os $K/2\,\%$ valores máis baixos.
- **Produto**:

$$\mu_j(\mathbf{x}) = \prod_{i=1}^{L} d_{i,j}(\mathbf{x}).$$

Estes métodos de combinación denomínanse non adestrables debido a que non requiren a estimación de ningún parámetro. Na Figura 3.2 amósase un exemplo do cálculo dos graos de soporte para un obxecto $\mathbf{x}$ utilizando a media.

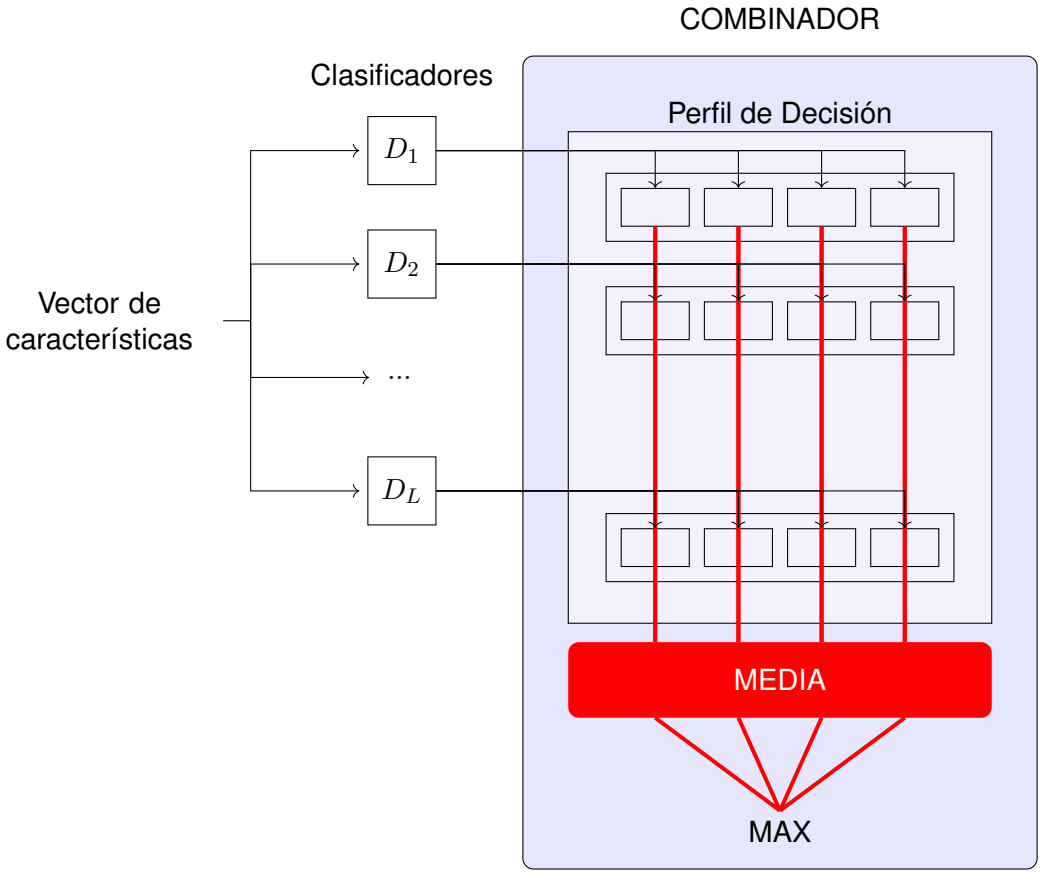


Figura 3.2: Cálculo dos graos de soporte utilizando a media.

Exemplo 3.1. Consideremos un problema de clasificación con 3 clases e sexa un ensemble con 5 clasificadores base que para un determinado obxecto $\mathbf{x}$ proporcionan o perfil de decisión $DP(\mathbf{x})$

amosado a continuación:

$$DP(\mathbf{x}) = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.4 & 0.3 & 0.3 \\ 0.0 & 0.1 & 0.9 \\ 0.8 & 0.0 & 0.2 \\ 0.6 & 0.3 & 0.1 \end{bmatrix}. \tag{3.1}$$

Os graos de soporte obtidos cos métodos de combinación non adestrables amósanse na Táboa 3.1.

Táboa 3.1: Graos de soporte para cada clase dado o perfil de decisión (3.1).

Método de combinación	$\mu_1(\mathbf{x})$	$\mu_2(\mathbf{x})$	$\mu_3(\mathbf{x})$
Media	**0.52**	0.16	0.32
Mínimo	0.00	0.00	**0.10**
Máximo	0.80	0.30	**0.90**
Mediana	**0.60**	0.10	0.20
Media recortada ($K = 40\,\%$)	**0.33**	0.17	0.20
Produto	0.00	0.00	**0.0005**

Neste caso, con $K = 40\,\%$, a media recortada elimina a saída do clasificador que proporciona un grao de soporte máis alto e o que proporciona un grao de soporte máis baixo para cada clase. Na Táboa 3.1 apréciase que o produto é moi sensible a valores próximos a 0.

Cabe resaltar que non se require que as saídas dos clasificadores sumen 1. Soamente asumimos que están medidas nas mesmas unidades. Se seleccionamos a clase cun maior grao de soporte polo método de combinación, a media, a mediana e a media recortada propoñen a clase ω_1 como a clase máis probable. O método do mínimo, do máximo e do produto propoñen a clase ω_3.

3.2.1. Equivalencias dos métodos de combinación non adestrables

Sexa $\mathcal{D} = \{D_1, \ldots, D_L\}$ un ensemble de clasificadores base e $\Omega = \{\omega_1, \omega_2\}$ o conxunto de clases. Se as entradas do perfil de decisión son estimacións das probabilidades a posteriori, $P(y = \omega_j|\mathbf{x})$, o método de combinación do mínimo e o máximo son equivalentes e o método de combinación da mediana é equivalente ao voto por maioría. As seguintes proposicións demostran estas equivalencias.

Proposición 3.2. *Sexa $\mathcal{D} = \{D_1, \ldots, D_L\}$ e $\Omega = \{\omega_1, \omega_2\}$. Sexan $a_1, \ldots, a_L$ as estimacións dos clasificadores do ensemble para a clase ω_1 e $1 - a_1, \ldots, 1 - a_L$, as estimacións para a clase ω_2, $a_i \in [0, 1]$. Entón, a clase asignada polo método do mínimo é a mesma que a asignada polo método do máximo.*

Demostración. Sen perda de xeralidade supoñamos que $a_1 = \text{mín}_i\, a_i$ e que $a_L = \text{máx}_i\, a_i$. Deste modo, os graos de soporte para cada clase proporcionados polo mínimo serán

$$\mu_1(\mathbf{x}) = a_1, \qquad \mu_2(\mathbf{x}) = 1 - a_L,$$

e os proporcionados polo máximo serán

$$\mu_1(\mathbf{x}) = a_L, \qquad \mu_2(\mathbf{x}) = 1 - a_1.$$

Consideremos as tres posibilidades para a_1 e a_L:

- Se $a_1 > 1 - a_L$, entón $1 - a_1 < a_L$ e ω_1 será a clase asignada por ambos métodos.
- Se $a_1 < 1 - a_L$, entón $1 - a_1 > a_L$ e ω_2 será a clase asignada por ambos métodos.
- Se $a_1 = 1 - a_L$, entón $1 - a_1 = a_L$ e ambos métodos asignarán unha clase segundo o método establecido para desempatar.

□

Na práctica, coas condicións da Proposición 3.2, se o método de desempate é aleatorio, o método do mínimo e o máximo poden asignar clases diferentes nalgúns casos. Non obstante, se empregan outros métodos de desempate como o voto por maioría, ou asignar sempre a clase ω_1 en caso de empate, os métodos do mínimo e do máximo tamén serán equivalentes na práctica.

Proposición 3.3. *Sexa $\mathcal{D} = \{D_1, \ldots, D_L\}$ con L impar e $\Omega = \{\omega_1, \omega_2\}$. Sexan $a_1, \ldots, a_L$ as estimacións dos clasificadores do ensemble para a clase ω_1 e $1 - a_1, \ldots, 1 - a_L$, as estimacións para a clase ω_2, $a_i \in [0, 1]$. Entón, a clase asignada polo método da mediana é a mesma que a asignada polo voto por maioría.*

Demostración. Sen perda de xeralidade supoñamos que $a_1 < \ldots < a_L$. Deste xeito, a mediana das saídas dos clasificadores será $\frac{a_{(L+1)/2}}{2}$.

- Se $\frac{a_{(L+1)/2}}{2} > 0.5$, entón o método da mediana asignará a clase ω_1. Por outra banda, as estimacións de probabilidade para a clase ω_1 de, polo menos $\frac{L+1}{2}$ clasificadores, é maior que 0.5. Polo tanto o voto por maioría asignará a clase ω_1.
- Se $\frac{a_{(L+1)/2}}{2} < 0.5$, entón o método da mediana asignará a clase ω_2. Neste caso, polo menos $\frac{L+1}{2}$ clasificadores propoñen a clase ω_2 polo que o voto por maioría asignará tamén esta clase.
- Se $\frac{a_{(L+1)/2}}{2} = 0.5$, entón, se ambos métodos resolven os empates do mesmo xeito, asignarán a mesma clase.

□

De novo, os empates poden levar a que o método da mediana e o voto por maioría asignen clases diferentes. Se asignamos unha clase aleatoria nos casos de empate a asignación de clases pode diferir. Non obstante, se aplicamos outro método de combinación, como a media para desempatar, a equivalencia manterase.

3.2.2. Formulación xeral

A media xeralizada aplicada neste contexto [8] permite expresar os métodos de combinación non adestrables de forma xeral. Deste xeito, dado $\mathbf{x}$, a saída do ensemble para a clase ω_j é

$$\mu_j(\mathbf{x}, \alpha) = \left(\sum_{i=1}^{L} d_{i,j}(\mathbf{x})^{\alpha} \right)^{\frac{1}{\alpha}},$$

onde α é un parámetro que determina que método de combiandión se aplica. Na Táboa 3.2 amósanse algúns casos especiais desta media xeralizada.

Táboa 3.2: Casos especiais da media xeralizada.

$\alpha \to -\infty$	$\mu_j(\mathbf{x}) = \text{mín}_i\, d_{i,j}(\mathbf{x})$	mínimo
$\alpha = -1$	$\mu_j(\mathbf{x}) = \left(\frac{1}{L} \sum_{i=1}^{L} \frac{1}{d_{i,j}(\mathbf{x})} \right)^{-1}$	media armónica
$\alpha \to 0$	$\mu_j(\mathbf{x}) = \left(\prod_{i=1}^{L} d_{i,j}(\mathbf{x}) \right)^{1/L}$	media xeométrica
$\alpha = 1$	$\mu_j(\mathbf{x}) = \frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})$	media aritmética
$\alpha \to +\infty$	$\mu_j(\mathbf{x}) = \text{máx}_i\, d_{i,j}(\mathbf{x})$	máximo

Para que estes métodos sigan sendo non adestrables o valor de α debe ser fixo e non depender dos datos. Este parámetro pode interpretarse como o nivel de optimismo do método de combinación. Dete modo, o mínimo é o método de combinación máis pesimista, pois require que todos os clasificadores proporcionen un grao de soporte para a clase ω_j maior ou igual que $\mu_j(\mathbf{x})$. No lado oposto, o máximo é o método de combinación máis optimista, pois require que polo menos un clasificador propoña un grao de soporte para a clase ω_j menor ou igual que $\mu_j(\mathbf{x})$.

Vexamos que efectivamento cando $\alpha \to 0$ a media xeralizada é equivalente ao produto.

Calculando o límite e tomando logaritmos obtemos a seguinte indeterminación:

$$l = \lim_{\alpha \to 0} \left(\frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})^{\alpha} \right)^{\frac{1}{\alpha}}.$$

$$\ln(l) = \ln \left(\lim_{\alpha \to 0} \left(\frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})^{\alpha} \right)^{\frac{1}{\alpha}} \right) = \lim_{\alpha \to 0} \frac{\ln \left(\frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})^{\alpha} \right)}{\alpha}.$$

Aplicando a regra de L'Hôpital obtemos

$$\begin{aligned} \ln(l) &= \lim_{\alpha \to 0} \frac{\frac{d}{d\alpha} \ln \left(\frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})^{\alpha} \right)}{\frac{d}{d\alpha}\alpha} = \lim_{\alpha \to 0} \frac{\frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})^{\alpha} \ln(d_{i,j}(\mathbf{x}))}{\frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})^{\alpha}} \\ &= \frac{1}{L} \sum_{i=1}^{L} \ln(d_{i,j}(\mathbf{x})). \end{aligned}$$

Por último, aplicando a exponencial obtemos

$$l = \exp \left(\frac{1}{L} \sum_{i=1}^{L} \ln(d_{i,j}(\mathbf{x})) \right) = \exp \left(\sum_{i=1}^{L} \ln(d_{i,j}(\mathbf{x})) \right)^{1/L} = \left(\prod_{i=1}^{L} d_{i,j}(\mathbf{x}) \right)^{1/L}.$$

A media xeométrica é equivalente ao produto, pois esta obtense elevando a $1/L$ as funcións de soporte do produto. Ao tratarse dunha transformación monótona que non depende da clase, a asignación final da clase non se ve afectada.

O razoamento para obter o método do máximo e o mínimo a partir da media xeralizada é moi semellante. No caso do máximo, calculando límites e aplicando logaritmos chegamos de novo a unha indeterminación:

$$l = \lim_{\alpha \to \infty} \left(\frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})^{\alpha} \right)^{\frac{1}{\alpha}}.$$

$$\ln(l) = \lim_{\alpha \to \infty} \frac{\ln \left(\frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})^{\alpha} \right)}{\alpha}.$$

Aplicando outra vez a regra de L'Hôpital,

$$\ln(l) = \lim_{\alpha\to\infty} \frac{\frac{d}{d\alpha}\ln\left(\frac{1}{L}\sum_{i=1}^{L} d_{i,j}(\mathbf{x})^\alpha\right)}{\frac{d}{d\alpha}\alpha} = \lim_{\alpha\to\infty} \frac{\frac{1}{L}\sum_{i=1}^{L} d_{i,j}(\mathbf{x})^\alpha \ln(d_{i,j}(\mathbf{x}))}{\frac{1}{L}\sum_{i=1}^{L} d_{i,j}(\mathbf{x})^\alpha} \tag{3.2}$$

$$= \lim_{\alpha\to\infty} \frac{\ln(d_{1,j}(\mathbf{x}))d_{1,j}(\mathbf{x})^\alpha}{\sum_{i=1}^{L} d_{i,j}(\mathbf{x})^\alpha} + \cdots + \frac{\ln(d_{L,j}(\mathbf{x}))d_{L,j}(\mathbf{x})^\alpha}{\sum_{i=1}^{L} d_{i,j}(\mathbf{x})^\alpha},$$

onde, se $M_j = \{d_{i,j}(\mathbf{x}), i = 1, \ldots, L : d_{i,j}(\mathbf{x}) = \text{máx}_i\{d_{i,j}(\mathbf{x})\}\}$ é o conxunto dos $d_{i,j}(\mathbf{x})$ que son máximos e $|M_j|$ é o número de elementos de M_j, obtemos

$$\lim_{\alpha\to\infty} \frac{\ln(d_{k,j}(\mathbf{x}))d_{k,j}(\mathbf{x})^\alpha}{\sum_{i=1}^{L} d_{i,j}(\mathbf{x})^\alpha} = \begin{cases} \ln(d_{k,j}(\mathbf{x}))/|M_j| & \text{se } d_{k,j}(\mathbf{x}) = \text{máx}_i\{d_{i,j}(\mathbf{x})\}, \\ 0 & \text{noutro caso.} \end{cases}$$

Polo tanto, na Ecuación (3.2) temos $|M|$ sumandos cuxo límite é $\ln\left(\text{máx}_i\{d_{i,j}(\mathbf{x})\}\right)/|M|$, polo que o logaritmo do límite é

$$\ln(l) = \ln(\max_i\{d_{i,j}(\mathbf{x})\}),$$

e, finalmente,

$$l = \max_i\{d_{i,j}(\mathbf{x})\}.$$

3.2.3. Xustificación dos métodos non adestrables: Diverxencia de Kullback-Leiber

Algúns dos métodos non adestrables teñen a súa orixe no sentido común. Por exemplo, a media aritmética é unha forma natural de responder á pregunta "Cal é o consenso xeral dos clasificadores?". Por outra banda, o método do máximo é unha forma de responder á pregunta "Hai algún clasificador que propoña unha clase con moita confianza?" e o método do mínimo é unha forma de responder á pregunta "Hai algunha clase que ningún clasificador descarte completamente?". Deste modo, o método do mínimo asigna a clase "menos mala", mentres que o método do máximo basea a súa asignación no clasificador "máis optimista".

Non obstante, algúns destes métodos de combinación tamén teñen unha xustificación teórica. A diverxencia de Kullback-Leiber (KL) é unha medida da diferencia entre dúas distribucións de probabilidade, unha distribución a priori, Q, e unha distribución a posteriori, P. Dado un valor discreto x no espazo mostral de P e Q, a diverxencia de KL defínese como

$$D_{KL}(P||Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

A diverxencia de KL pode ser interpretada como a cantidade de información necesaria para transformar a distribución de probabilidade a priori Q na distribución a posteriori, P. Aínda que, en certo modo, mide a distancia entre dúas distribucións, non é unha métrica no sentido formal, pois non é simétrica nin cumpre a desigualdade triangular. O feito de que non sexa simétrica ímolo empregar para derivar dous dos métodos non adestrables: o produto e a media.

En primeiro lugar, imos empregar as filas de $DP(\mathbf{x})$ como estimacións da distribución de probabilidade a priori. Deste modo, $d_{i,j}$ é unha estimación de $P(\omega_j|\mathbf{x}, D_i)$. Denotaremos por $P_{(i)}$ a distribución de probabilidade en Ω do clasificador D_i para o obxecto $\mathbf{x}$. Así, $P_{(i)} = (d_{i,1}, \ldots, d_{i,c})$. Por exemplo, dado o seguinte $DP(\mathbf{x})$,

$$DP(\mathbf{x}) = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.4 & 0.3 & 0.3 \\ 0.0 & 0.1 & 0.9 \\ 0.8 & 0.0 & 0.2 \\ 0.6 & 0.3 & 0.1 \end{bmatrix},$$

teríamos $P_{(1)} = (0.8, 0.1, 0.1)$, $P_{(2)} = (0.4, 0.3, 0.3)$, etc.

Por outro lado, asumiremos que as funcións de soporte do ensemble toman os valores das probabilidades a posteriori reais, $P(\omega_i|\mathbf{x})$. Denotaremos as funcións de soporte do ensemble por $P_{\text{ens}} = (\mu_1(\mathbf{x}), \ldots, \mu_c(\mathbf{x}))$. Desta maneira, imos considerar cada $P_{(i)}$ como a distribución de probabilidade a priori na diverxencia de KL e P_{ens}, como a distribución a posteriori. Polo tanto, a media das diverxencias de KL vén dada por

$$KL_{\text{media}} = \frac{1}{L} \sum_{i=1}^{L} KL(P_{\text{ens}}||P_{(i)}). \tag{3.3}$$

Buscamos P_{ens} que minimice KL_{media}. Para iso, consideramos o seguinte problema de optimización

$$\begin{aligned} \text{mín} \quad & KL_{\text{media}} = \frac{1}{L} \sum_{i=1}^{L} \sum_{k=1}^{c} \mu_k \log_2\left(\frac{\mu_k}{d_{i,k}}\right) \\ \text{s.a.} \quad & \sum_{k=1}^{c} \mu_k = 1. \end{aligned}$$

Para atopar os μ_k que minimizan KL_{media} incluímos o termo de Lagrange asociado á restrición

e formulamos un sistema coas derivadas parciais igualadas a 0. Deste xeito,

$$\frac{\partial}{\partial \mu_j}\left[KL_{\text{media}} + \lambda\left(1 - \sum_{k=1}^{c}\mu_k\right)\right]$$

$$= \frac{1}{L}\sum_{i=1}^{L}\frac{\partial}{\partial \mu_k}\left[\sum_{k=1}^{c}\mu_j \log_2\left(\frac{\mu_k}{d_{i,k}}\right)\right] + \lambda\frac{\partial}{\partial \mu_j}\left[1 - \sum_{k=1}^{c}\mu_k\right]$$

$$= \frac{1}{L}\sum_{i=1}^{L}\sum_{k=1}^{c}\frac{\partial \mu_k}{\partial \mu_j}\log_2\left(\frac{\mu_k}{d_{i,k}}\right) + \mu_k\frac{\partial}{\partial \mu_j}\left[\log_2\left(\frac{\mu_k}{d_{i,k}}\right)\right] - \lambda$$

$$= \frac{1}{L}\sum_{i=1}^{L}\left(\log_2\left(\frac{\mu_k}{d_{i,k}}\right) + C\right) - \lambda = 0,$$

onde $C = \log_2(e) = \frac{1}{\ln(2)}$. Resolvendo para μ_j obtemos

$$\mu_j = 2^{(\lambda - C)}\prod_{i=1}^{L}(d_{i,j})^{1/L}. \tag{3.4}$$

Substituíndo a Ecuación (3.4) en $\sum_{i=1}^{c}\mu_k = 1$ e resolvendo para λ obtemos

$$\lambda = C - \log_2\left(\sum_{k=1}^{c}\prod_{i=1}^{L}(d_{i,k})^{1/L}\right).$$

Substituíndo λ na Ecuación (3.4) obtemos unha expresión final para as funcións de soporte do ensemble que consiste na media xeométrica normalizada:

$$\mu_j = \frac{\prod_{i=1}^{L}(d_{i,j})^{1/L}}{\sum_{k=1}^{c}\prod_{i=1}^{L}(d_{i,k})^{1/L}}.$$

Se eliminamos o denominador (que non depende de j) e a potencia $1/L$ do numerador, obtemos a fórmula para as funcións de soporte do produto. Ambas son transformacións monótonas que non alteran a orde dos μ_j, polo que a asignación de clases non se ve afectada. Polo tanto as funcións de soporte resultan

$$\mu_j = \prod_{i=1}^{L} d_{i,j}.$$

Se intercambiamos as funcións de P_{ens} e $P_{(i)}$ na Ecuación (3.3) e buscamos os μ_j que mini-

mizan KL_{media} suxeito á restrición $\sum_{k=1}^{c} \mu_k = 1$, obtemos

$$\frac{\partial}{\partial \mu_j}\left[KL_{\text{media}} + \lambda\left(1 - \sum_{k=1}^{c}\mu_k\right)\right]$$

$$= \frac{1}{L}\sum_{i=1}^{L}\frac{\partial}{\partial \mu_k}\left[\sum_{k=1}^{c} d_{i,k}\log_2\left(\frac{d_{i,k}}{\mu_k}\right)\right] + \lambda\frac{\partial}{\partial \mu_j}\left[1 - \sum_{k=1}^{c}\mu_k\right]$$

$$= \frac{1}{L}\sum_{i=1}^{L}\frac{d_{i,j}}{\mu_j}\frac{1}{C} - \lambda = \frac{1}{CL\mu_j}\sum_{i=1}^{L} d_{i,j} - \lambda = 0,$$

onde, de novo, $C = \frac{1}{\log_2(e)}$. Resolvendo para μ_j obtemos

$$\mu_j = -\frac{1}{\lambda CL}\sum_{i=1}^{L} d_{i,j}. \tag{3.5}$$

Substituíndo a Ecuación (3.5) en $\sum_{i=1}^{c} \mu_k = 1$ e resolvendo para λ,

$$\lambda = -\frac{1}{\lambda CL}\sum_{k=1}^{c}\sum_{i=1}^{L} d_{i,j} = -\frac{L}{CL} = -\frac{1}{C}.$$

Por último, substituíndo λ na Ecuación (3.5) obtemos unha expresión final para as funcións de soporte do ensemble que consiste na media aritmética:

$$\mu_j = -\frac{1}{L}\sum_{i=1}^{L} d_{i,j}.$$

Para derivar o produto, asumimos que as funcións de soporte do ensemble conforman a distribución a posteriori real e tratamos de buscar un método de combinación que aproxime a distribución do ensemble ás distribucións dos L clasificadores. Por outro lado, para derivar a media, asumimos que as funcións de soporte dos clasificadores conforman a distribución a posteriori real e tratamos de buscar un método de combinación que minimice a información necesaria para transformar a distribución do ensemble na distribución dos clasificadores.

Miller e Yan [10] engaden un clasificador binario por cada clasificador do ensemble, que denominan *crítico*. O papel deste clasificador é determinar se a saída do clasificador base é fiable ou non, e utilízano para determinar a exactitude de cada clasificador do ensemble. Xustifican o papel deste crítico en contextos onde os clasificadores do ensemble deben distinguir entre máis

de dúas clases, ou a exactitude dalgún dos clasificadores é inferior a 0.5. No seu traballo, en lugar de minimizar a diverxencia de KL media dos clasificadores, minimizan a diverxencia de KL media ponderada cos pesos proporcionados polos críticos. Deste modo, o caso considerado neste traballo pode ser interpretado como un caso particular do traballo de Miller e Yan, onde os pesos de todos os clasificadores do ensemble son iguais a $1/L$.

3.3. Media ponderada

O método de combinación da media ponderada combina os elementos de $DP(\mathbf{x})$ mediante unha media ponderada. En función de que elementos de $DP(\mathbf{x})$ se empreguen, podemos distinguir tres modelos:

- L **pesos**. Este modelo considera un peso para cada clasificador base. A función de soporte para a clase ω_j vén dada por

$$\mu_j(\mathbf{x}) = \sum_{i=1}^{L} w_i d_{i,j}(\mathbf{x}),$$

  onde w_i é o peso asociado clasificador D_i.

- $c \cdot L$ **pesos**. Este modelo emprega un peso para cada clasificador base e para cada clase. A función de soporte para a clase ω_j vén dada por

$$\mu_j(\mathbf{x}) = \sum_{i=1}^{L} \sum_{k=1}^{c} w_{i,k} d_{i,j}(\mathbf{x}),$$

  onde $w_{i,k}$ é o peso asociado ao clasificador D_i para a clase ω_k.

- $c^2 \cdot L$ **pesos**. Este modelo calcula o soporte de cada clase mediante unha combinación linear do perfil de decisión $DP(\mathbf{x})$ enteiro.

$$\mu_j(\mathbf{x}) = \sum_{i=1}^{L} \sum_{k=1}^{c} w_{i,k,j} d_{i,j}(\mathbf{x}), \tag{3.6}$$

  onde $w_{i,k,j}$ é o peso asociado ao elemento $d_{i,k}$ de $DP(\mathbf{x})$ para a clase ω_k.

Nas seguintes seccións preséntanse tres formas distintas de asignar pesos aos clasificadores base e analízase a súa influencia na exactitude do ensemble.

3.3.1. Cálculo dos pesos mediate críticos

Miller e Yan [10] propoñen un método para calcular os pesos dos clasificadores base mediante un clasificador binario adicional, denominado crítico. Os críticos adéstranse despois de adestrar os clasificadores base para o que utilizan os obxectos do dataset cos que se adestran os clasificadores e as predicións dos clasificadores.

Desta maneira, un crítico consiste nun clasificador binario que recibe como entrada un obxecto $\mathbf{x}$ e a predición dun clasificador do ensemble. Como saída, o crítico determina se o clasificador acertou na súa clasificación ou se, pola contra, cometeu un erro. Así, as dúas posibles clases son "obxecto ben clasificado" e "obxecto mal clasificado".

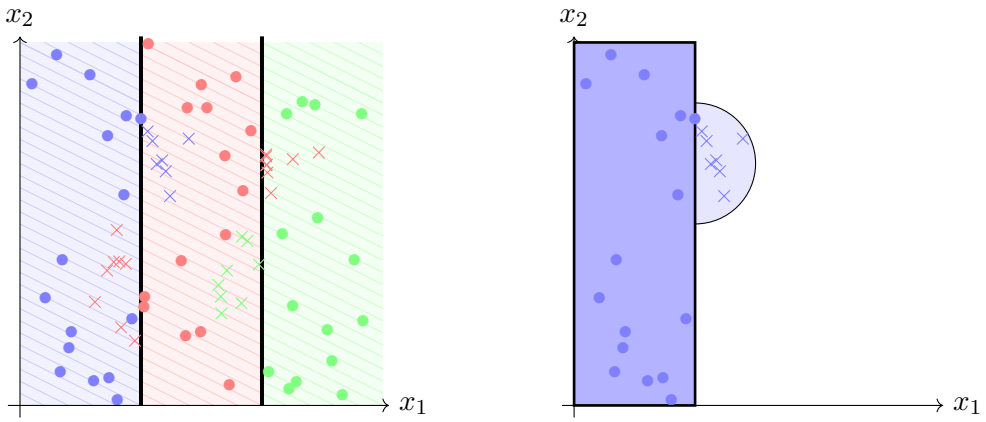
Dunha maneira máis visual, o obxectivo durante o adestramento dun crítico é detectar as rexións do espazo de características nas que o clasificador proporciona unha clase incorrecta, e as rexións nas proporciona a clase correcta. É dicir, o crítico ten como tarefa separar o espazo de características en dúas rexións para cada clase: unha rexión onde o clasificador base acerta e outra onde falla.

Supoñamos que temos un problema con $\Omega = \{\omega_1, \omega_2, \omega_3\}$ e que $\mathbf{x} \in \mathbb{R}^2$. Na Figura 3.3 represéntanse as rexións de clasificación dun clasificador moi simple. Os obxectos do conxunto de datos represéntanse cun círculo se o clasificador suxire a clase correcta, e cunha cruz se suxire unha clase incorrecta. A cor dos puntos representa á clase á que pertencen. Nas Figuras 3.3b, 3.3c e 3.3d represéntanse as rexións de clasificación que se espera que o crítico consiga esbozar durante o adestramento para as clases ω_1, ω_2 e ω_3, respectivamente.

Para calcular o peso dun clasificador para un determinado obxecto $\mathbf{x}$, utilízase o valor da función de soporte do crítico asociado á clase "obxecto ben clasificado". Desta forma, se o clasificador cometeu un erro, a función de soporte do crítico asociada a esta clase proporcionará un valor baixo e, se o clasificador acertou, proporcionará un valor alto. Así, se por exemplo o clasificador da Figura 3.3a prediciu a clase ω_1 para o obxecto $\mathbf{x}$, esperamos que o crítico proporcione un peso máis baixo para o clasificador canto máis próximo esté da rexión azul clara da Figura 3.3b.

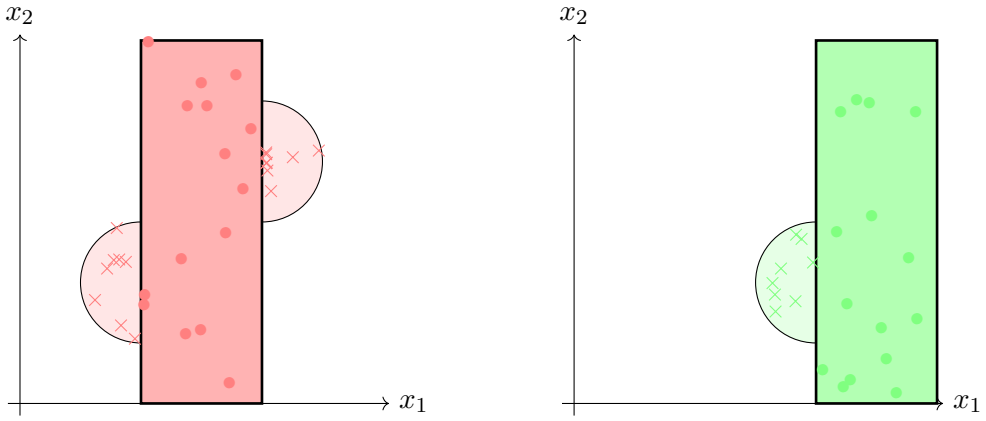
Miller e Yan xustifican este método de asignar pesos aos clasificadores en problemas de clasificación orixinal con máis de dúas clases. Argumentan que as predicións do crítico son máis confiables que as do clasificador, pois o primeiro soamente debe resolver un problema de dúas clases, mentres que o segundo, resolve un problema máis complexo. Ademais, argumentan que o crítico permite manter a efectividade dos métodos de combinación cando a exactitude dos clasificadores é inferior a 0.5.

Cabe resaltar que, a diferenza do resto dos métodos desta sección, neste caso para cada $\mathbf{x}$ obtemos un peso distinto do crítico asociado ao clasificador.



(a) Rexións de clasificación dun clasificador simple. Os obxectos mal clasificados represéntanse cunha cruz e os obxectos ben clasificados cun punto.

(b) Rexións de clasificación asociadas á clase ω_1 que se esperan do crítico.



(c) Rexións de clasificación asociadas á clase ω_2 que se esperan do crítico.

(d) Rexións de clasificación asociadas á clase ω_3 que se esperan do crítico.

Figura 3.3: Rexións de clasificación dentro do espazo de características $\mathbb{R}^2$ dun clasificador simple e do crítico. Nas Subfiguras (b), (c) e (d), as zonas máis escuras correspóndense cos puntos que o clasificador clasifica correctamente e, as máis claras, cos puntos que clasifica incorrectamente. O obxectivo do crítico é detectar estas rexións durante o adestramento.

3.3.2. Cálculo dos pesos a partir do erro engadido

Seguindo o traballo de Tumer e Ghosh [16], Fumera e Roli [5] derivaron unha forma de calcular os pesos para o método de combinación da media ponderada. En primeiro lugar, supoñen que as estimacións de $P(\omega_j|\mathbf{x})$ son

$$\hat{P}(\omega_j|\mathbf{x}) = \sum_{i=1}^{L} w_i d_{i,j}, \qquad i = 1, \dots, c,$$

onde $d_{i,j}$ é o respectivo elemento de $DP(\mathbf{x})$ e w_i, $i = 1, \dots, c$ son os pesos dos clasificadores do ensemble $\mathcal{D} = \{D_1, \dots, D_L\}$ cumprindo

$$\sum_{i=1}^{L} w_i = 1, \qquad w_i \geq 0.$$

Fumera e Roli reutilizan a hipótese principal do traballo de Tumer e Ghost, que consiste en que as decisións obtidas a partir das probabilidades a posteriori aproximadas son próximas ás decisións de Bayes, ou o que é o mesmo, que podemos utilizar o erro aproximado obtido das estimacións de $P(\omega_j|\mathbf{x})$ como estimación do erro engadido. Baixo certas asuncións, pódese calcular o conxunto de pesos óptimos para L clasificadores independentes a partir do erro engadido de cada clasificador do ensemble, E^i_{add}, $i = 1, \dots, L$. O erro engadido defínese como a diferenza entre o erro total e o erro de Bayes para un problema concreto de clasificación. No seu traballo, Fumera e Roli chegan á seguinte expresión para os pesos óptimos:

$$w_i = \frac{1/E^i_{\text{add}}}{\sum_{k=1}^{L} 1/E^i_{\text{add}}}, \qquad i = 1, \dots, L. \tag{3.7}$$

A Ecuación (3.7) amosa que os pesos óptimos son inversamente proporcionais ao erro engadido esperado dos clasificadores do ensemble. Polo tanto, para valores iguais do erro engadido esperado, os pesos óptimos resultan $w_i = \frac{1}{L}$.

Fumera e Roli realizaron un estudo computacional comparando o método da media ponderada con esta forma de calcular os pesos coa media [6]. Os resultados amosaron que esta forma de calcular os pesos non mellora como esperaban ao método da media. Un dos motivos débese a que as hipóteses baixo as cales esta forma de calcular os pesos é óptima son demasiado restritivas e pouco realistas. Os seus experimentos amosan esta forma de calcular os pesos consegue un mellor rendemento cando o número de clasificadores do enesemble é pequeno, os clasificadores teñen taxas de erro significativamente distintas ou as saídas dos clasificadores están altamente correlacionadas.

No seu traballo, Fumera e Roli non utilizaron un conxunto de validación para calcular o erro engadido, senón que utilizaron o conxunto de test para calculalos. O seu obxectivo era comprobar cal era a mellora ideal que podían conseguir calculando os pesos desta forma en comparación co método da media. Nun contexto de traballo real, a estimación dos pesos pode ser unha fonte de erro que cancele a pequena vantaxe que se poida obter coa forma de calcular os pesos proposta por Fumera e Roli.

3.3.3. Cáculo dos pesos mediante regresión linear

Outro xeito de calcular os pesos dos clasificadores base é mediante regresión linear. Para cada clase, adéstrase unha regresión linear da forma da expresión (3.6), onde os valores das variables independentes son os elementos de $DP(\mathbf{x})$ e a variable dependente é $P(\omega_j|\mathbf{x})$. Posto que non coñecemos os valores reais de $P(\omega_j|\mathbf{x})$, para cada obxecto $\mathbf{x}$, utilizamos como variable dependente unha variable que toma o valor 1 se a clase do obxecto é ω_j e 0 en caso contrario.

Ergodan e Sen [3] propoñen unha formulación para a función de erro a minimizar durante o adestramento da regresión linear. Dado un conxunto de datos $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$, con etiquetas $\{y_1, \dots, y_N\}$, $y_i \in \Omega$, se denotamos por μ_j a función de soporte para a clase ω_j e por $\mathbf{w}$ os coeficientes da regresión, a función de erro a minimizar vén dada por

$$\Psi(\mathbf{w}) = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{c} \mathcal{L}(\mu_i(\mathbf{z}_j), y_i, \omega_i, \mathbf{w}) + \mathcal{R}(\mathbf{w}), \tag{3.8}$$

onde $\mathcal{L}$ é termo que mide a discrepancia entre a predición do obxecto $\mathbf{z}_j$ e a clase ω_j, e $R(\mathbf{w})$ é un termo de regularización que serve para penalizar pesos moi grandes. Ante este escenario, é necesario tomar dúas decisións: a elección da función de erro e a elección do termo de regularización.

Existen diferentes posibilidades para a función de erro $\mathcal{L}$ e para o termo de regularización $R(\mathbf{w})$. O erro cadrático medio é a función de optimización utilizada tradicionalmente para regresión. Porén, existen outras posibilidades para a función de erro, como a función de perda de Hinge e a función de perda loxística. Ergodan e Sen compararon o rendemento do erro cadrático medio coa función de perda de Hinge en problemas de clasificación multiclase con tres conxuntos de datos diferentes e en varios escenarios [3]. Observaron que, en xeral, a función de perda de Hinge proporciona mellores resultados.

Se simplificamos a notación da función de perda na expresión (3.8) a $\mathcal{L}(a, b)$, onde $a \in \{-1, 1\}$ toma o valor 1 se o ensemble suxire a clase incorrecta, e -1, se suxire unha clase incorrecta, e b é o valor da función de soporte asociada á clase predita polo ensemble, podemos definir as funcións de erro da seguinte maneira [13]:

- Erro cadrático medio:
$$\mathcal{L}(a, b) = (a - b)^2 = (1 - ab)^2.$$

- Función de perda de Hinge (función de perda do clasificador SVM):
$$\mathcal{L}(a, b) = \text{máx}(0, 1 - ab).$$

- Función de perda loxística:

$$\mathcal{L}(a, b) = \log(1 + \exp(-ab)).$$

Segundo empregamos un método de combinación máis sofisticado, o risco de sobreaxuste aumenta, especialmente, cando o ensemble está composto por modelos precisos e altamente correlacionados. A regularización, axuda a evitar o sobreaxuste e a mellorar a exactitude da predición. Reid e Grudic [11] estudaron o efecto de diferentes funcións de regularización e compararon a exactitude obtida mediante unha regresión fronte a outros métodos máis simples, como a media e o voto por maioría. As funcións de regularización consideradas foron as seguintes:

- Regularización L_2. Utilizada coa función de erro cadrático medio denomínase regresión de Ridge:

$$R(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2.$$

- Regularización L_1 (LASSO):

$$R(\mathbf{w}) = \lambda\|\mathbf{w}\|_1.$$

- Regurización Elastic Net. Consiste nunha combianción das dúas anteriores:

$$R(\mathbf{w}) = \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2.$$

A regresión Ridge produce modelos densos. Isto significa que todos os clasificadores do ensemble son utilizados para facer a predición. Porén, LASSO produce modelos dispersos, ou o que é o mesmo, só un subconxunto dos clasificadores do ensemble son utilizados para facer a predición. Reid e Grudic [11] realizaron o seu estudo utilizando como función de erro o erro cadrático medio e chegaron ás seguintes conclusións. A regresión Ridge supera á regresión non regularizada e mellora o rendemento do mellor clasificador individual do ensemble, en xeral. LASSO non foi tan exitosa como a regresión Ridge, polo que os autores concluíron que os modelos densos son mellores que os modelos dispersos.

3.4. Un clasificador como método de combinación

Nesta sección consideraremos o espazo intermedio de características que estará formado por $DP(\mathbf{x})$ (considerado como un vector, construído concatenando as L filas). Sobre este espazo, consideraremos diferentes clasificadores que se encargarán de realizar a clasificación. O obxectivo de sofisticar o método de combinación consiste en lograr mellores resultados a base de recoñecer patróns dentro do espazo intermedio (o que se coñece como meta-aprendizaxe).

3.4.1. Modelos de Decisión

A idea dos modelos de decisión (*DT*, *Decision Templates* en inglés) consiste en "lembrar" o perfil de decisión máis habitual para cada clase ω_j para despois poder comparalo con novos perfís de decisión mediante algunha medida de similitude ou proximidade, $\mathcal{S}$. Durante o adestramento, para construír o modelo de decisión j, DT_j, promédianse os perfís de decisión de todos os obxectos do conxunto de datos da clase ω_j. Para un novo obxecto $\mathbf{x}$, calcúlase o seu $DP(\mathbf{x})$ e compárase con $DT_1, \ldots, DT_c$ mediante a medida de similitude $\mathcal{S}$. A clase predita para $\mathbf{x}$ será a clase asociada ao DT_j máis similar a $DP(\mathbf{x})$. Na Figura 3.4 amósase un esquema do proceso de clasificación mediante modelos de decisión.

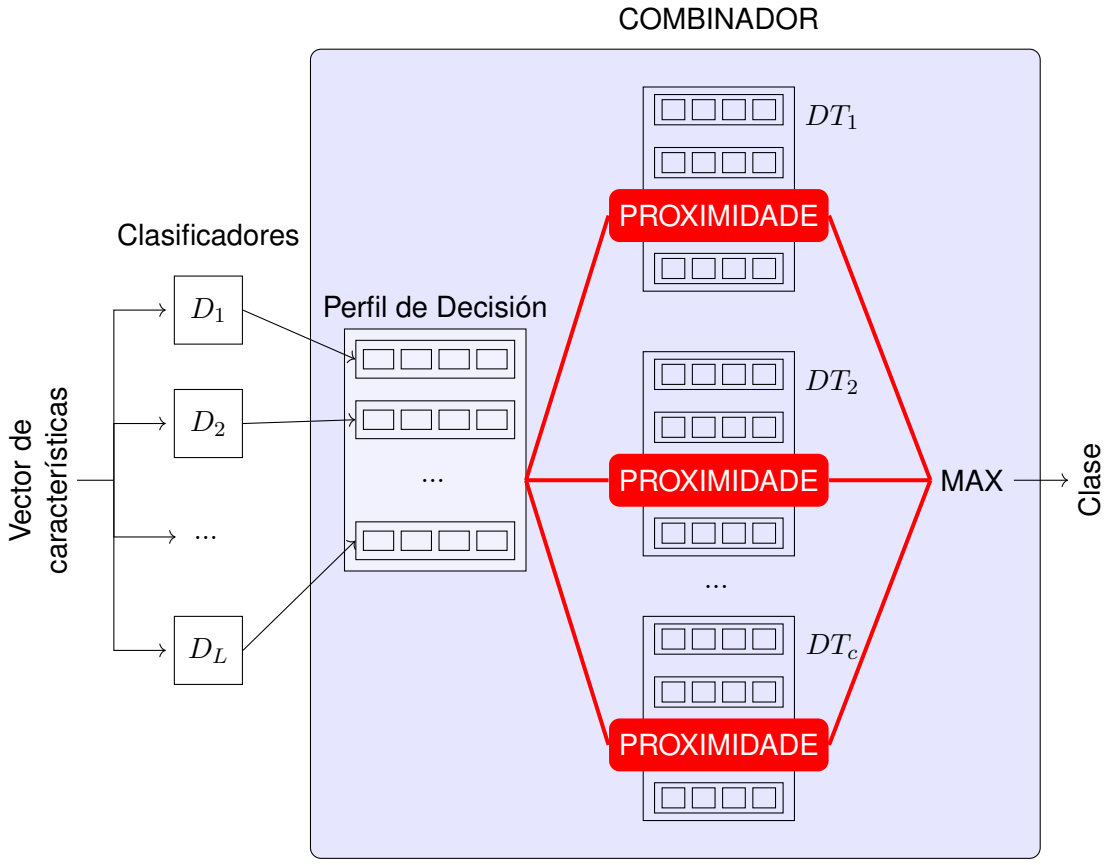


Figura 3.4: Predición mediante modelos de decisión. O vector de características $\mathbf{x}$ pásase a través dos clasificadores $D_1, \ldots, D_L$ e calcúlase a proximidade cos modelos de decisión $DT_1, \ldots, DT_c$. O modelo de decisión con maior proximidade determina a clase.

Existen varias formas de medir a similitude entre $DP(\mathbf{x})$ e DT_j. Dúas medidas de similitude comúns están baseadas na distancia euclídea e na diferencia simétrica da teoría de conxuntos difusos [7].

- Distancia euclídea. Podemos definir as funcións de soporte do ensemble da seguinte forma:

$$\mu_j(\mathbf{x}) = 1 - \frac{1}{L\,c}\sum_{i=1}^{L}\sum_{k=1}^{c}(DT_j(i,k) - d_{i,k}(\mathbf{x}))^2, \tag{3.9}$$

  onde $DT_j(i,k)$ é o elemento (i,k) do modelo de decisión DT_j e $d_{i,k}(\mathbf{x})$ é o elemento (i,k) do perfil de decisión $DP(\mathbf{x})$. A clase predita para $\mathbf{x}$ será a clase asociada ao modelo de decisión DT_j que maximice $\mu_j(\mathbf{x})$.

  Buscar a función de soporte μ_j, $j = 1, \ldots, c$ que maximiza a Ecuación (3.9) é equivalente a buscar a media máis próxima no espazo intermedio de características.

  En lugar de empregar a distancia euclídea, tamén se pode empregar calquera outra distancia definida en $\mathbb{R}^n$, como a distancia de Mahalanobis ou a distancia de Minkowski.

- Diferencia simétrica. Se consideramos que DT_j, $j = 1, \ldots, c$ e $DP(\mathbf{x})$ son subconxuntos difusos, podemos medir a similitude entre DT_j e $DP(\mathbf{x})$ e expresar as funcións de soporte do ensemble como

$$\mu_j(\mathbf{x}) = 1 - \frac{1}{L\,c}\sum_{i=1}^{L}\sum_{k=1}^{c}\text{máx}\{\text{mín}\{DT_j(i,k), (1 - d_{i,k}(\mathbf{x}))\}, \text{mín}\{(1 - DT_j(i,k)), d_{i,k}(\mathbf{x})\}\}.$$

Exemplo 3.4. Sexan $c = 2$, $L = 3$ e os modelos de decisión DT_1 e DT_2, asociados a ω_1 e ω_2 respectivamente,

$$DT_1 = \begin{bmatrix} 0.6 & 0.4 \\ 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}, \qquad DT_2 = \begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \\ 0.1 & 0.9 \end{bmatrix}.$$

Asumimos que o perfil de decisión do obxecto $\mathbf{x}$ é

$$DP(\mathbf{x}) = \begin{bmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \\ 0.5 & 0.5 \end{bmatrix}.$$

As funcións de soporte do ensemble que miden a similitude entre $DP(\mathbf{x})$ e DT_1 e DT_2 son as seguintes:

Modelo de decisión	$\mu_1(\mathbf{x})$	$\mu_2(\mathbf{x})$	Clase
Distancia euclídea	0.9567	0.9333	ω_1
Diferencia simétrica	0.5000	0.5333	ω_2

3.5. Clasificación apilada

Sobre o espazo de características intermedio pódense empregar calquera tipo de clasificador (árbores de decisión, redes neuronais, máquinas de vectores soporte, etc.). Por exemplo, as máquinas vectores soporte (SVM) pode ser unha opción máis axeitada que adestrar unha regresión, xa que estes método busca minimizar directamente o erro de clasificación en lugar de optimizar as estimacións das probabilidades a posteriori.

Cando se utiliza un método sofisticado de combinación, existen varias estratexias á hora de adestrar os clasficadores base e o combinador:

1. Utilizar un só conxunto de datos para adestrar os clasificadores base e o combinador. Neste caso, é conveniente deixar os clasificadores base subadestrados e completar o adestramento mediante o combinador. Isto permítelle ao combinador corrixir os erros dos clasificadores base.

2. Utilizar dous conxuntos de datos para adestrar os clasificadores base e o combinador. Neste caso, os clasificadores base pódense sobreadestrar e utilizar o combinador para corrixir o sesgo.

Dietrich et al. [2] propoñen unha división do conxunto de adestramento en dous conxuntos parcialmente superpostos. Sexa $\mathbf{Z}$ o conxunto de adestramento, R, o conxunto de adestramento e T, o conxunto de test, de modo que $R \bigcup T = \mathbf{Z}$. O conxunto de adestramento R pódese dividir en dous subconxuntos R_1 e R_2 de tal xeito que $R_1 \bigcup R_2 = R$. Os clasificadores base adestraranse con R_1 e o combinador con R_2. Se $\mathbf{Z}$ é pequeno e $R_1 \cup R_2 = \emptyset$, pode dar lugar a que o adestramento dos clasificadores base e do combinador sexa demasiado pobre e a que as estimacións mediante o conxunto de test non sexan boas. Para remediar isto, podemos introducir un parámetro ρ que controla a superposición entre R_1 e R_2:

$$\rho = \frac{|R_1 \cap R_2|}{|R_1|},$$

onde $|.|$ denota a cardinalidade do conxunto. Para $\rho = 0$, R_1 e R_2 son disxuntos; para $\rho = 1$, R_1 e R_2 son idénticos. Os autores atoparon que un valor intermedio de $\rho = 0.5$ ofrece mellores resultados, o que suxire a necesidade dun compromiso cando o conxunto de datos $\mathbf{Z}$ é relativamente pequeno.

A clave á hora de escoller un método de combinación dos clasificadores base está en entender e analizar as condicións específicas de cada problema. O tipo de datos, o tamaño do ensemble, a homoxeneidade, diversidade e estratexia de construción son algúns dos factores que debemos ter en conta á hora de escoller o combinador axeitado. No seguinte exemplo poñemos de manifesto

un caso no que é máis axeitado un combinador non adestrable, como o é o método da media, que un clasificador sofisticado como método de combinación.

Exemplo 3.5. Consideremos o conxunto de datos *Wine Quality Red* obtido do repositorio *UCI Machine Learning Repository*. O conxunto de datos contén 1599 obxectos con 11 variables de entrada continuas e unha variable de saída que indica a calidade do viño. Para o noso caso, quedarémonos cos obxectos das clases 5, 6 e 7, dando un total de 1518 obxectos. O obxectivo é reducir o desbalanceo das clases e simplificar o problema.

Para executar as probas utilizaremos o paquete de R *caretEnsemble*. Este paquete permítenos adestrar L clasificadores base que poidan ser adestrados con *caret* e adestrar un clasificador sobre as saídas dos clasificadores base. Posto que o paquete non permite utilizar conxuntos de datos distintos para adestrar os clasificadores base e o combinador, limitarémonos a este caso. Como clasificadores base utilizaremos os seguintes:

- *rpart*: unha árbore de decisión cunha profundidade máxima de 4.
- *knn*: un clasificador k-NN con $k = 5$.
- *lda*: un clasificador LDC.
- *qda*: un clasificador QDC.

Imos comparar 4 combinadores distintos sobre o ensemble: 3 adestrables e un non adestrable. Para os combinadores adestrables, imos axustar os hiperparámetros mediante o método de validación cruzada *K-fold* con $K = 10$. Os combinadores que imos comparar son os seguintes:

- *nnet*. Unha rede neuronal, onde o número de capas ocultas e o número de neuróns por capa serán hiperparámetros a axustar.
- *glmnet*. Unha regresión loxística regularizada, onde o parámetro de regularización será o hiperparámetro a axustar.
- *svmLinear*. Unha máquina de vectores soporte lineal, onde o parámetro de regularización será o hiperparámetro a axustar.
- *media*. Como combinador non adestrable, utilizaremos o método da media.

Dividiremos o conxunto de datos en dúas partes, un conxunto de adestramento e un conxunto de test. O conxunto de adestramento será o 60 % do conxunto de datos e o conxunto de proba será o 40 % restante. O obxectivo é comparar a exactitude dos combinadores sobre o conxunto de proba. Realizaremos 10 experimentos e calcularemos a media da exactitude obtida en cada experimento. Ademais, na comparación incluiremos o clasificador base que obtén unha maior exactitude media no conxunto de proba. Na Táboa 3.3 amósanse os resultados obtidos.

Combinador	Exactitude media
nnet	0.6267
glmnet	0.6257
svmLinear	0.6181
media	**0.6311**
Mellor clasificador base: *lda*	0.6193

Táboa 3.3: Exactitude media obtida con 4 combinadores sobre un ensemble con 4 clasificadores base.

Observamos que o combinador non adestrable, o método da media, obtén unha maior exactitude media no conxunto de proba. Este exemplo pon de manifesto que non sempre é necesario empregar un combinador sofisticado para obter unha maior exactitude.

Á hora de deseñar un ensemble, o combinador non é a única decisión que se debe tomar. Neste caso, adestráronse todos os clasificadores base e o combinador co mesmo conxunto de datos. Unha estratexia que se podería empregar para mellorar os resultados consiste en sesgar os clasificadores base adestrándoos cunha parte do conxunto de datos e tratar de corrixir o sesgo mediante o combinador. Outra posibilidade sería dividir o conxunto de datos en dúas partes, unha para adestrar os clasificadores base e outra para adestrar o combinador. Isto permitiría que os clasificadores base se especializasen en diferentes partes do espazo de características e o combinador corrixise os erros dos clasificadores base.

Capítulo 4

Métodos de ensamblado

Nos Capítulos 2 e 3 analizamos as distintas posibilidades para establecer o combinador dun ensemble. Porén, como xa apuntamos, á hora de deseñar un ensemble existen outras cuestións que abordar. Cantos clasificadores deben formar parte do ensemble? Como introducimos diversidade? Os clasificadores deben adestrarse á vez ou secuencialmente? Neste capítulo abordaremos estas cuestións e presentaremos tres dos métodos de ensamblado maís coñecidos: bagging, bosques aleatorios e adaBoost.

4.1. bagging

O método de bagging (acrónimo de *Bootstrap AGGregatING*) consiste en combinar L clasificadores base adestrados con diferentes subconxuntos do conxunto de datos. Como combinador utilízase o voto por maioría.

Nun ensemble, se todos os clasificadores se adestran co mesmo conxunto de datos, córrese o risco de que os clasificadores base estean altamente correlacionados. Ante este escenario, por moi sofisticado que sexa o combinador, a exactitude do ensemble será limitada. Desta maneira, o método de bagging busca introducir diversidade utilizando conxuntos de adestramento diferentes para os clasificadores do ensemble.

Idealmente, os clasificadores deberían ser adestrados en mostras independentes da distribución orixinal dos datos. Posto que na práctica, isto non é posible, os conxuntos de adestramento deben ser construídos mediante mostraxe con reemprazamento a partir conxunto de datos $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$. Para conseguir a diversidade deste xeito é fundamental que os clasificadores base sexan inestables, é dicir, as saídas dos clasificadores deben ser sensibles a pequenos cambios no conxunto de datos.

Como vimos no Capítulo 2, o método do voto por maioría é óptimo cando os clasificadores base son independentes e teñen a mesma exactitude. O método de bagging constrúe L clasificadores independentes adestrando os clasificadores con subconxuntos distintos do conxunto de datos. Na práctica, é difícil conseguir que os clasificadores sexan independentes, pois os conxuntos de adestramento de cada clasificador obtéñense a partir do mesmo conxunto de datos. Porén, incluso no caso de adestrar os clasificadores mediante mostras independentes da distribución real dos datos, as saídas dos clasificadores estarían correlacionadas igualmente. No seguinte exemplo poñemos de manifesto isto último.

Exemplo 4.1. Supoñamos que temos o conxunto de datos, $\mathbf{Z}$, amosado na Figura 4.1. Este conxunto de datos está formado por 500 obxectos con dúas características, é dicir, $\mathbf{z}_j \in \mathbb{R}^2$, $j = 1, \ldots, N$. Utilizaremos 100 obxectos para o conxunto de adestramento e os outros 400, como conxunto de test. Imos asumir que coñecemos a distribución dos datos, de modo que os puntos da clase vermella son aqueles que $x_1 < 1$ e $x_2 < 1$ ou $x_1 > 1$ e $x_2 > 1$, mentres que os puntos da clase azul son aqueles que $x_1 < 1$ e $x_2 > 1$ ou $x_1 > 1$ e $x_2 < 1$.

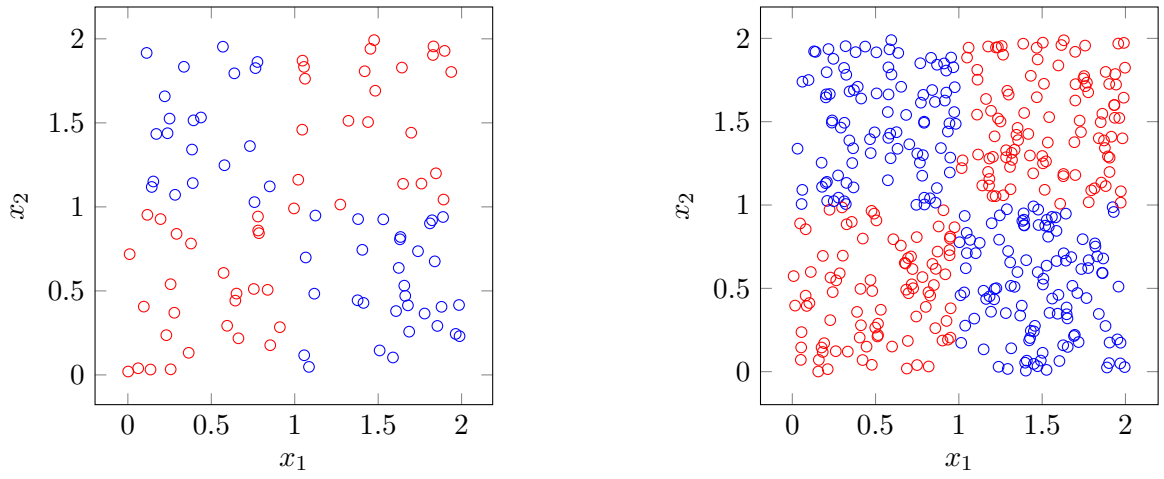


Figura 4.1: Conxunto de datos de adestramento (esquerda) e conxunto de datos de proba (dereita) para o exemplo de bagging.

Consideraremos un ensemble co método de ensamblado bagging con $L = 5, 10, \ldots, 30$ clasificadores base e co voto por maioría como método de combinación. Como clasificadores base empregamos LDC. O propósito deste exemplo consiste en comparar o aumento da correlación entre as saídas dos clasificadores do ensemble ao utilizar mostras con reemprazo do conxunto de datos de adestramento en lugar de mostras independentes da propia distribución dos datos. Ademais, veremos como a correlación entre as saídas dos clasificadores afecta ao erro de clasificación do ensemble.

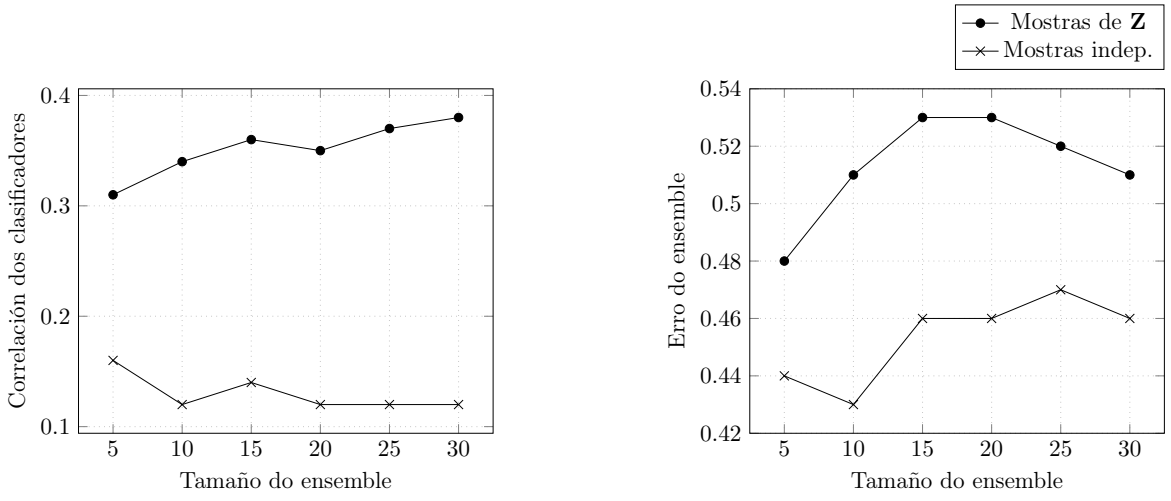
A correlación entre dous clasificadores do ensemble, D_i e D_j pódese medir do seguinte xeito:

$$\rho_{i,j} = \frac{N^{11}N^{00} - N^{10}N^{01}}{\sqrt{(N^{11} + N^{10})(N^{11} + N^{01})(N^{00} + N^{10})(N^{00} + N^{01})}},$$

onde N^{11} é o número de obxectos clasificados correctamente polos dous clasificadores, N^{00} é o número de obxectos clasificados incorrectamente polos dous clasificadores, N^{10} é o número de obxectos clasificados correctamente polo primeiro clasificador e incorrectamente polo segundo clasificador e N^{01} é o número de obxectos clasificados incorrectamente polo primeiro clasificador e correctamente polo segundo clasificador. Para obter un só valor para cada ensemble, calculamos a media da correlación entre todos os pares de clasificadores.

Na Figura 4.2 amósanse os resultados obtidos. Para cada tamaño de ensemble, xeramos aleatoriamente 10 conxuntos de datos con 500 obxectos. Para cada experimento, calculamos a media da correlación entre os clasificadores do ensemble e do erro de clasificación. Na Figura 4.2a, obsérvase o comportamento esperado: a correlación entre os clasificadores é maior cando se utilizan mostras con reemprazo do conxunto de datos de adestramento, en lugar de mostras independentes da distribución real dos datos. Cabe resaltar que a correlación dos clasificadores adestrados con mostras independentes non é nula, o que pon de manifesto que os clasificadores tamén poden ser dependentes neste caso.

Podemos interpretar a correlación entre as saídas dos clasificadores como unha medida inversamente proporcional á diversidade do ensemble. Así, na Figura 4.2b obsérvase que o erro de clasificación do ensemble é menor cando o ensemble ten máis diversidade.



(a) (b)

Figura 4.2: Correlación e erro de clasificación para un ensemble con método de ensamblado bagging utilizando mostras con reemprazamento do conxunto de datos e mostras independentes.

4.1.1. Bosques aleatorios

Os bosques aleatorios son un caso particular de bagging no que os clasificadores base son árbores de decisión. A diferenza do bagging, tal e como o presentamos na sección 4.1, a diversidade nos clasificadores base dun bosque aleatorio introdúcese mediante a selección aleatoria dun subconxunto das características. Desta maneira para cada árbore de decisión que forma parte do ensemble escóllese un subconxunto S de m características de xeito aleatorio. Un valor típico para m é $\sqrt{n}$, onde n é o número de características.

Desta forma, cada árbore de decisión do ensemble atende soamente a un subconxutno das características. Isto axuda a reducir a correlación entre os clasificadores base e a aumentar a diversidade do ensemble.

4.2. AdaBoost

A idea principal detrás de adaBoost (acrónimo de *ADAptive BOOSTing*) é engadir clasificadores base secuencialmente ao ensemble, de xeito que cada clasificador k se adestre de forma que corrixa os erros dos $k-1$ clasificadores anteriores. Para isto, o clasificador que se une ao ensemble na iteración k adéstrase nunha mostra do conxunto de adestramento seleccionada cunha distribución onde os elementos clasificados erróneamente nas iteracións anteriores do adestramento teñen maior probabilidade. Na primeira iteración é habitual que a distribución sexa uniforme. A medida que se engaden clasificadores base, a distribución vai cambiando de xeito que os obxectos mal clasificados teñen unha maior probabilidade de ser seleccionados para formar parte do conxunto de adestramento dos seguintes clasificadores.

Na Figura 4.3 amósase o algoritmo de adestramento de *AdaBoost.M1*. A distribución coa que se constrúe o conxunto de datos en cada iteración vén determinada por $\mathbf{w}^k$. Inicialmente, para que todos os obxectos do conxunto de datos teñan a mesma probabildiade de ser seleccionados, establécese $w_j^1 = 1/N$, $j = 1, \ldots, N$. En cada iteración, a distribución $\mathbf{w}^k$ actualízase en función dos erros cometidos polo clasificador D_k.

Unha vez adestrado o ensemble, o valor da función de soporte para un obxecto $\mathbf{x}$ ven determinado por:

$$\mu_t(\mathbf{x}) = \sum_{k=1}^{L} \frac{1}{\ln \beta_k} I(D_k(\mathbf{x}) = \omega_t),$$

onde β_k é o peso asociado ao clasificador D_k obtido na iteración k e $I(D_k(\mathbf{x}) = \omega_t)$ toma o valor 1 se o clasificador k propón a clase ω_t e 0, noutro caso. Noutras palabras, súmase o logaritmo da inversa dos pesos dos clasificadores que predicen correctamente a clase ω_t para o obxecto $\mathbf{x}$.

4.2.1. Cota superior do erro de adaBoost

Freud e Schapire [4] demostraron que o erro de clasificación do ensemble de adaBoost pode ser acotado en función do erro de clasificación dos clasificadores base e do número de clasificadores que forman parte do ensemble.

Teorema 4.2. *Sexa* $\Omega = \{\omega_1, \dots, \omega_c\}$. *Sexa* ϵ *o erro do ensemble durante o adestramento e* ϵ_i, $i = 1, \dots, L$ *o erro do clasificador* D_i *obtido a través da Ecuación* (4.1), *con* $\epsilon_i < 0.5$. *Entón,*

$$\epsilon < 2^L \prod_{i=1}^{L} \sqrt{\epsilon_i(1-\epsilon_i)}.$$

Segundo aumentamos o tamaño do ensemble mediante clasificadores cun erro inferior a 0.5, o erro do ensemble de adaBoost aproxímase a cero. Un dos feitos que fan que adaBoost sexa un algoritmo popular para problemas de clasificación é precisamente a rapidez coa que se reduce o erro do ensemble no conxunto de adestramento.

ADABOOST.M1

Sexa $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ o conxunto de datos de adestramento, o algoritmo ADABOOST.M1 constrúe un ensemble de clasificadores mediante a iteración dos seguintes pasos:

1. Escoller o tamaño do ensemble, L e o tipo de clasificadores que formarán parte del.
2. Establecer os pesos $\mathbf{w} = \{w_1^1, \dots, w_N^1\}$, $w_j^1 \in [0,1]$ e $\sum_{j=1}^N w_j^1 = 1$. Habitualmente $w_j^1 = 1/N$, $j = 1, \dots, N$.
3. Para $k = 1, \dots, L$:
   1. Obter a mostra S_k a partir da distribución $\mathbf{w}^k$.
   2. Adestrar o clasificador D_k coa mostra S_k.
   3. Calcular o erro do ensemble ponderado do paso k como:
   $$\epsilon_k = \sum_{i=1}^{N} w_i^k l_j^k, \tag{4.1}$$
   onde $l_j^k = 1$ se $\mathbf{z}_j$ é clasificado incorrectamente polo clasificador D_k e $l_j^k = 0$ en caso contrario.
   4. En función do valor de ϵ_k, actualizar os pesos $\mathbf{w}^{k+1}$:
      1. Se $\epsilon_k = 0$, entón reinicializar os pesos w_j^k a $1/N$ e continuar.
      2. Se $\epsilon \geq 0.5$, ignorar D_k, reinicializar os pesos w_j^k a $1/N$ e continuar.
      3. Noutro caso, Calcular
      $$\beta_k = \frac{\epsilon_k}{1-\epsilon_k}, \quad \epsilon_k \in (0, 0.5),$$
      e actualizar os pesos individuais:
      $$w_j^{k+1} = \frac{w_j^k \beta_k^{(1-l_j^k)}}{\sum_{i=1}^N w_i^k \beta_k^{(1-l_j^k)}}, \quad j = 1, \dots, N.$$
5. Devolver $\mathcal{D} = \{D_1, \dots, D_L\}$ e $\{\beta_1, \dots, \beta_L\}$.

Figura 4.3: Algoritmo de adestramento do ADABOOST.M1.

Capítulo 5

Conclusións

Neste traballo estúdanse distintos métodos para combinar saídas de clasificadores, xa sexan predicións ou valores continuos, e preséntanse tres métodos de ensamblado: bagging, bosques aleatorios e adaBoost.

Con respecto aos métodos de combinación de predicións, cada un deles é óptimo baixo determinadas asuncións. Aínda que na práctica estas condicións de optimalidade non sempre se cumpren, un método non óptimo pode ser preferible a outro, dependendo da calidade das estimacións dos parámetros. Isto depende, principalmente, do tamaño do conxunto de datos e de se o conxunto de datos segue ou non a distribución real dos datos.

Cando as saídas dos clasificadores do ensemble son valores continuos, existen numerosas formas de combinalas, que van desde métodos non adestrables, como calcular a media ou a mediana, ata métodos máis complexos, como empregar outro clasificador para combinar as saídas. De novo, o número de parámetros a estimar do método de combinación é un factor importante a ter en conta á hora de elixir un método de combinación.

A pesar de que o método de combinación é unha das eleccións máis importantes á hora de construír un ensemble, existen outros factores que poden determinar o rendemento do ensemble. Por exemplo, os bosques aleatorios son un dos métodos máis utilizados nos últimos anos en problemas de clasificación de diferentes ámbitos. Estes métodos empregan clasificadores base sinxelos, como son as árbores de decisión, e utilizan como método de combinación o voto por maioría. Por outra banda, adaBoost adestra os clasificadores do ensemble de forma secuencial, e utiliza como método de combinación un método non adestrable.

En resumo, cada método presenta vantaxes e desvantaxes que dependen do problema de clasificación específico a resolver. Os bosques aleatorios destacan pola súa simplicidade e robustez, mentres que adaBoost ofrece a vantaxe de adaptar o peso dos clasificadores en función dos erros

previos. Polo tanto, a elección do método de combinación e do resto das cuestións que afectan á construción do ensemble deben facerse considerando as características específicas do problema a resolver.

Anexo A

Comparación de métodos de combinación de predicións de clasificadores

Neste exemplo veremos como afectan as suposicións sobre os clasificadores base na exactitude e optimalidade dos métodos de combinación vistos no Capítulo 2. Inicialmente, construirase un ensemble no que clasificadores base cumpran os requisitos para que o método de voto por maioría sexa o óptimo. Para simplificar o exemplo, as clases serán equiprobables, polo que o método de voto por maioría non terá ningún parámetro adestrable. Comprobaremos que, neste caso, a exactitude empírica do ensemble coincide coa exactitude teórica do método de voto por maioría calculada mediante a Ecuación (2.4).

Iremos facendo variacións sobre os clasificadores de modo que se deixen de cumprir as condicións de optimalidade. Definiremos 5 situacións diferentes:

1. Os clasificadores base son independentes e teñen a mesma exactitude. O método de voto por maioría é o óptimo.
2. Os clasificadores base son independentes, teñen exactitudes distintas. O método de voto por maioría ponderado é o óptimo.
3. Os clasificadores base son idependentes. O método de Naïve Bayes é o óptimo.
4. Os clasificadores base son dependentes. O método de BKS é o óptimo.
5. Os clasificadores base son dependentes e hai escaseza de datos. O método de BKS é o óptimo, mais veremos como afecta a escaseza de datos na exactitude empírica.

En cada unha das situacións calcularemos a exactitude empírica dos métodos. Para isto,

realizaremos 100 experimentos nos que xeraremos un conxunto de datos con 2 clases e 3 variables. O conxunto de datos terá 1000 obxectos de cada clase que serán xerados mediante distribucións normais. Utilizaremos 500 obxectos de cada clase para construír un conxunto de adestramento, e os 500 obxectos restantes para construír un conxunto de test. A exactitude empírica será a media da exactitude dos 100 experimentos.

A.1. Situación 1: Clasificadores base independentes e con igual exactitude

Inicialmente, xeraremos os datos de modo que o método de voto por maioría sexa o óptimo. Para iso, empregaremos dúas distribucións normais cos seguintes parámetros:

$$\mu_1 = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix}, \qquad \Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
$$\mu_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \Sigma_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Imos empregar tres clasificadores base independentes condicionado a cada clase ω_k e con exactitudes p idénticas. Para isto, empregaremos clasificadores lineais que separen o espazo en dúas rexións. Os puntos que queden nunha rexión serán clasificados como ω_1 e os que queden na outra rexión serán clasificados como ω_2.

Para conseguir que os clasificadores sexan independentes, cada clasificador soamente utilizará unha variable para realizar a clasificación. Como as variables x_i, $i = 1, 2, 3$ son independentes, os clasificadores serán independentes.

Por outro lado, para lograr que teñan exactitudes idénticas, situaremos o hiperplano de decisión no punto medio entre as medias das clases. Vexamos que deste modo, efectivamente se cumpre que $P(s_i = \omega_k|\omega_k) = p$, $i = 1, 2, 3$ para cada clasificador. Vexámolo para D_1 o clasificador que atende á primeira das variables, x_1. Para D_2 e D_3, o razoamento é análogo. Posto que as primeiras compoñentes (as asociadas a x_1) de μ_1 e μ_2 son -1 e 1, respectivamente, o hiperplano de decisión situarémolo en $x_1 = 0$. Así, o clasificador D_1 clasificará os obxectos como ω_1 se $x_1 < 0$

e como ω_2 se $x_1 > 0$. Así,

$$\begin{aligned} P(s_1 = \omega_1|\omega_1) &= P[X \leq 0] \approx 0.841, \quad X \sim N(-1, 1), \\ P(s_1 = \omega_2|\omega_1) &= P[X > 0] \approx 0.159, \quad X \sim N(-1, 1), \\ P(s_1 = \omega_2|\omega_2) &= P[Y \leq 0] \approx 0.159, \quad Y \sim N(1, 1), \\ P(s_1 = \omega_1|\omega_2) &= P[Y > 0] \approx 0.841, \quad Y \sim N(1, 1). \end{aligned} \tag{A.1}$$

Polo tanto, a exactitude do clasificador D_1 é do 84.1 % independentemente da clase a que pertenza o obxecto. Na Figura A.1 amosamos os hiperplanos de decisión dos clasificadores D_1 e D_2.

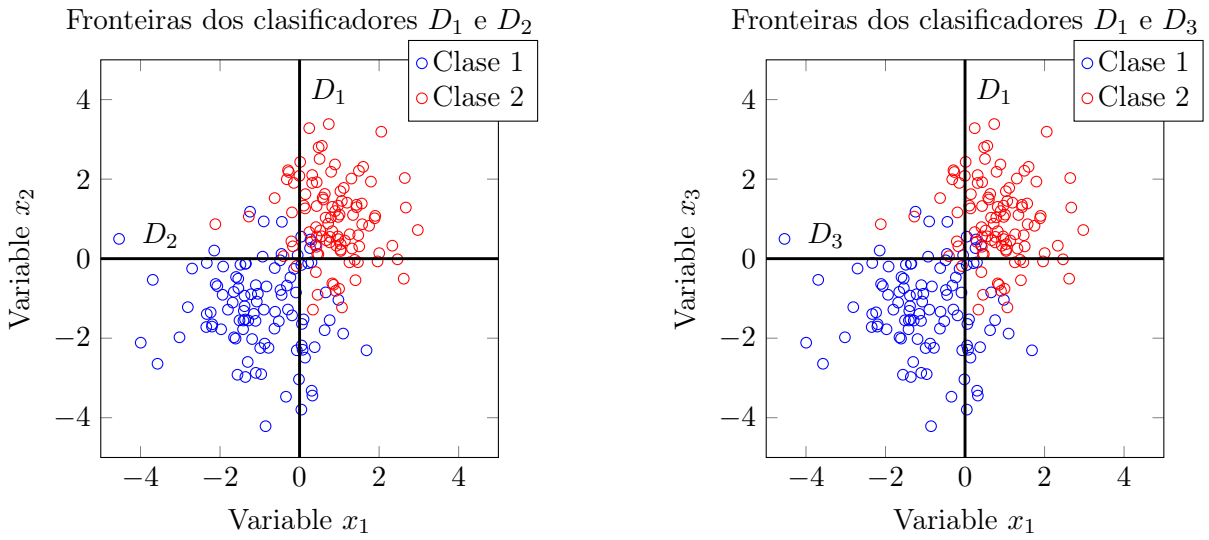


Figura A.1: Hiperplanos de decisión dos clasificadores D_1 e D_2 (esquerda) e D_1 e D_3 (dereita) para a situación 1.

Nun contexto de clasificación binaria e, baixo as condicións que fan que o método de voto por maioría sexa óptimo, pódese calcular a exactitude do ensemble mediante a Ecuación (2.4). Así, dado $p = 0.841$, $L = 3$ e $\left\lfloor \frac{L}{2} \right\rfloor + 1 = 2$, a exactitude do ensemble empregando o método de voto por maioría é

$$p_{ens} = \sum_{m=2}^{3} \binom{3}{m} p^m (1-p)^{3-m} = 3 \cdot p^2(1-p) + p^3 = 0.841^2 \cdot 0.159 + 0.841^3 \approx 0.932.$$

A Táboa A.1 amosa as exactitudes medias obtidas mediante os diferentes métodos ante as 5 situacións con diferentes condicións de optimalidade. A primeira columna correspónde á situación na que o método de voto por maioría é o óptimo. Obsérvase que a diferencia entre a exactitude empírica do ensemble e a exactitude teórica do método de voto por maioría é do 0.1 %. Ademais,

posto que o conxunto de adestramento ten suficientes datos[1], a estimación dos parámetros dos métodos adestrables é fiable e a exactitude empírica destes métodos é similar á exactitude teórica. Na situación 5 veremos como a escaseza de datos de adestramento afecta en maior medida aos métodos con un maior número de parámetros.

A.2. Situación 2: Clasificadores base independentes e con distinta exactitude

Neste caso empregaremos os mesmos 3 clasificadores que no caso anterior, mais modificaremos a matriz de covarianzas de modo que a varianza das variables 2 e 3 sexan maiores. As novas matrices de covarianzas son

$$\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}.$$

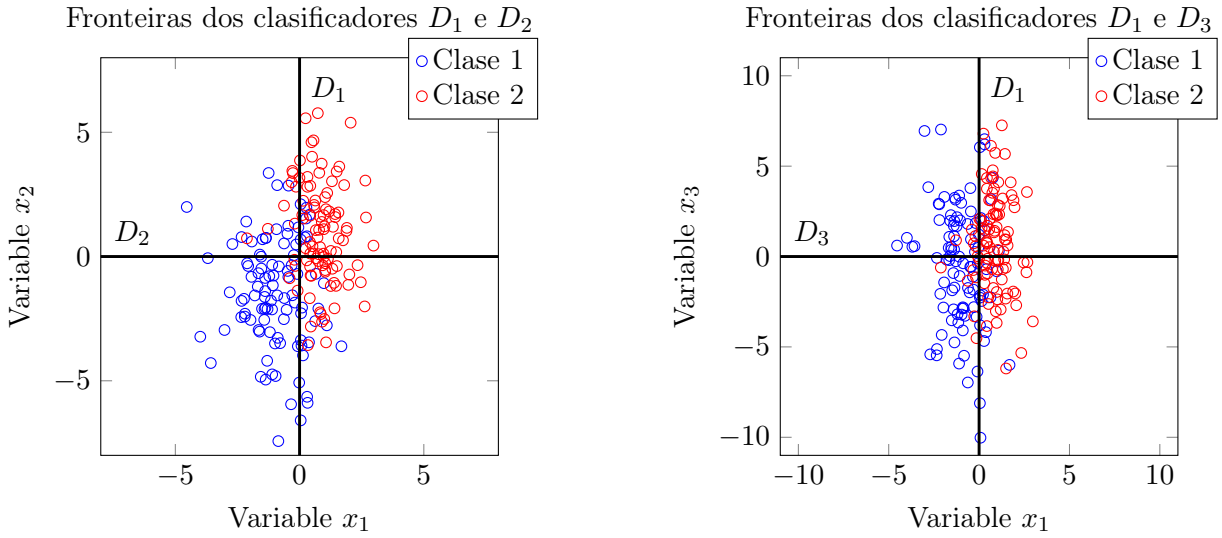


Figura A.2: Hiperplanos de decisión dos clasificadores D_1 e D_2 (esquerda) e D_1 e D_3 (dereita) da situación 2.

Na Figura A.2 amosamos os hiperplanos de decisión dos clasificadores D_1 e D_2 e D_3. Observamos que agora a dispersión das variables 2 e 3 é maior, polo que a exactitude dos clasificadores D_2 e D_3 diminúe. Calculando de forma análoga á Ecuación (A.1), obtemos que as exactitudes dos clasificadores D_1, D_2, D_3 son $p_1 = 0.841$, $p_2 = 0.691$, $p_3 = 0.631$. Na segunda columna da

[1] O conxunto de adestramento contén 500 obxectos de cada clase. Nun problema máis complexo, podería ser necesario un conxunto de adestramento máis grande.

Táboa A.1 observamos que a exactitude do método de voto por maioría é inferior á exactitude do resto dos métodos. Isto débese a que non se cumpre unha das asuncións deste método, o que inflúe negativamente na exactitude do ensemble.

O feito de que as exactitudes dos métodos na segunda columna da Táboa A.1 sexan inferiores ás da primeira columna débese a que agora D_2 e D_3 teñen unha exactitude inferior. Polo tanto, é razoable que a exactitude do ensemble se vexa afectada.

A.3. Situación 3: Clasificadores base independentes

Neste caso, xeramos os datos coas mesmas distribucións normais que na situación 2, mais agora modificamos os hiperplanos de decisión dos clasificadores. Para iso, situaremos o hiperplano de decisión de D_1 en $x_1 = 0.75$, o de D_2 en $x_2 = -0.25$ e o de D_3 en $x_3 = 2$. Así, por exemplo, D_3 será un clasificador que predí case sempre a clase ω_1. Na Figura A.3 amosamos os hiperplanos de decisión dos clasificadores D_1 e D_2 e D_3.

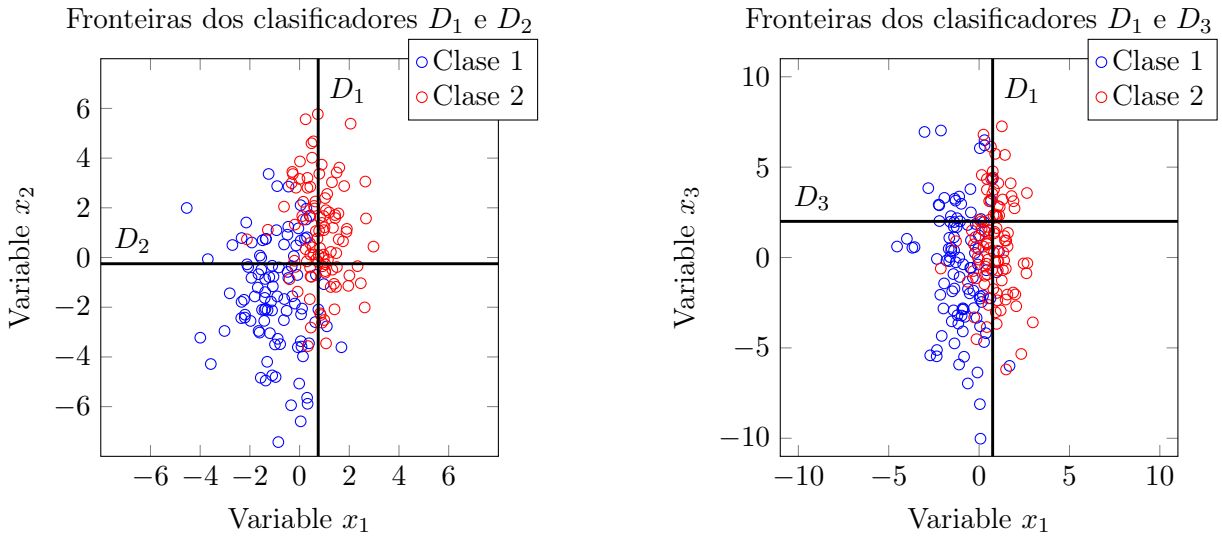


Figura A.3: Hiperplanos de decisión dos clasificadores D_1 e D_2 (esquerda) e D_1 e D_3 (dereita) da situación 3.

Na terceira columna da Táboa A.1 obsérvase que a exactitude do método de voto por maioría e de voto por maioría ponderado é inferior á exactitude do método de Naïve Bayes. Isto débese a que estes asumen que as exactitudes de cada clasificador base para cada clase son iguais, algo que nesta situación non se cumpre.

A.4. Situación 4: Clasificadores base dependentes

Neste caso, xeraremos os datos coas mesmas distribucións normais que na situación 2 e 3, mais agora introduciremos dous clasificadores máis. D_4 dependerá da variable x_1, igual que D_1, e D_5 dependerá da variable x_2, igual que D_2. Establecerase o hiperplano de decisión de D_4 en $x_1 = 2$ e o de D_5 en $x_2 = -1$. Deste xeito, existirá unha relación de dependencia entre D_1 e D_4 e entre D_2 e D_5. Na Figura A.4 amosamos os hiperplanos de decisión dos clasificadores D_1, D_2, D_3, D_4 e D_5.

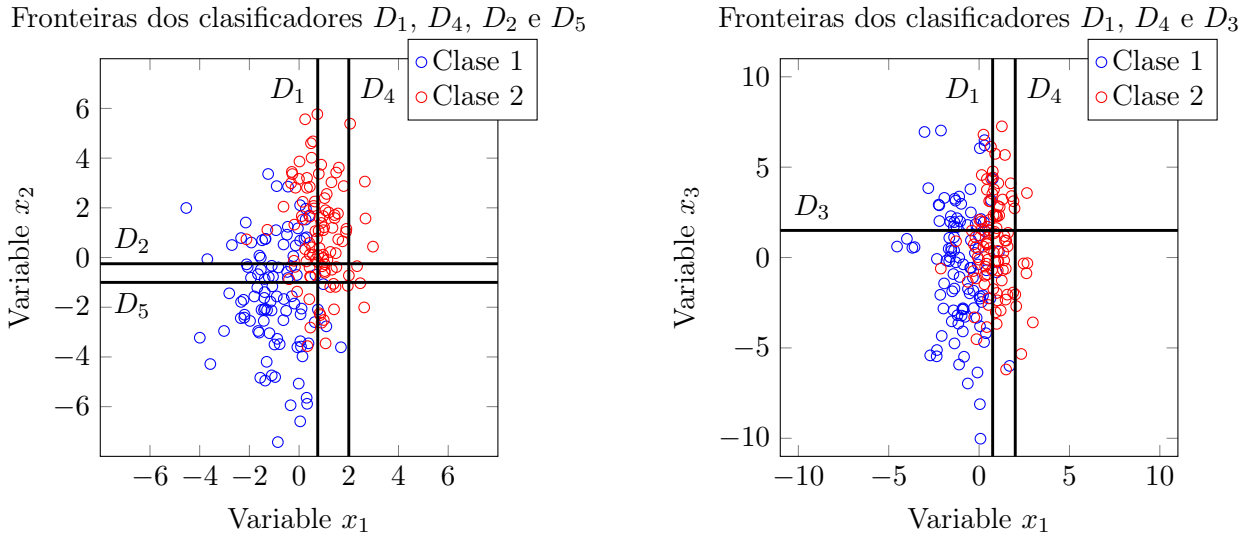


Figura A.4: Hiperplanos de decisión dos clasificadores D_1 e D_2 (esquerda) e D_1 e D_3 (dereita) da situación 4.

Na cuarta columna da Táboa A.1 obsérvase que a exactitude do método BKS é maior que a exactitude de todos os demais métodos. Ademais, a exactitude deste método non se ve afectada ao engadir dous clasificadores pouco precisos e que apenas aportan información. Isto amosa a robustez do método BKS ante a presenza de clasificadores dependentes con suficientes datos para estimar os parámetros.

A.5. Situación 5: Clasificadores base dependentes e escaseza de datos

Por último, imos reducir substancialmente o tamaño do conxunto de adestramento. Soamente imos empregar 50 obxectos de cada clase para adestrar os clasificadores co obxectivo de analizar como inflúe a escaseza de datos na exactitude dos métodos.

Na quinta columna da Táboa A.1 obsérvase que a exactitude do método BKS é a que máis se ve afectada pola escaseza de datos e que é inferior á exactitude dos demais métodos. Isto débese a que o método BKS ten un maior número de parámetros que os demais métodos, polo que necesita un conxunto de adestramento máis grande para estimar os parámetros con exactitude. Ademais, o único método que non se ve afectado pola escaseza de datos é o método de voto por maioría, xa que é un método non adestrable.

Esta situación exemplifica que, ante un problema con poucos datos, a estimación dos parámetros debe ser a principal preocupación e que un método con menos parámetros, como Naïve Bayes, pode ser preferible aínda que a súa asunción de optimalidade non se cumpra.

Táboa A.1: Estimación da exactitude dos métodos de combinación predicións de clasificadores ante 5 situacións con condicións distintas de optimalidade.

Método	1	2	3	4	5
Voto por maioría	0.933	0.816	0.768	0.767	0.767
Voto por maioría ponderado	0.933	0.843	0.774	0.781	0.763
Naïve Bayes	0.933	0.843	0.808	0.797	0.779
BKS	0.933	0.841	0.808	0.809	0.757

Cabeceiras das columnas:
1. p igual.
2. p_i específica para cada clasificador.
3. Matrices de confusión distintas.
4. Clasificadores dependentes.
5. Escaseza de datos.

Bibliografía

[1] Max Bramer. *Measuring the Performance of a Classifier*, pages 173–185. Springer London, London, 2007.

[2] Christian Dietrich, Günther Palm, and Friedhelm Schwenker. Decision templates for the classification of bioacoustic time series. *Information Fusion*, 4:101–109, 06 2003.

[3] Hakan Erdogan and Mehmet Umut Sen. A unifying framework for learning the linear combiners for classifier ensembles. In *2010 20th International Conference on Pattern Recognition*, pages 2985–2988, 2010.

[4] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[5] Giorgio Fumera and Fabio Roli. Performance analysis and comparison of linear combiners for classifier fusion. In Terry Caelli, Adnan Amin, Robert P. W. Duin, Dick de Ridder, and Mohamed Kamel, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, pages 424–432, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[6] Giorgio Fumera and Fabio Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:942–956, 2005.

[7] Ludmila I. Kuncheva. "Fuzzy" versus "nonfuzzy" in combining classifiers designed by Boosting. *IEEE Transactions on Fuzzy Systems*, 11(6):729–741, 2003.

[8] Ludmila I. Kuncheva. *Combining Pattern Classifiers. Methods and Algorithms.* Wiley, 2nd edition, 2014.

[9] L. Lam and S.Y. Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 27(5):553–568, 1997.

[10] D.J. Miller and Lian Yan. Critic-driven ensemble classification. *IEEE Transactions on Signal Processing*, 47(10):2833–2844, 1999.

[11] Sam Reid and Greg Grudic. Regularized linear models in stacked generalization. In Jón Atli Benediktsson, Josef Kittler, and Fabio Roli, editors, *Multiple Classifier Systems*, pages 112–121, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[12] John A. Richards. Classifier performance and map accuracy. *Remote Sensing of Environment*, 57(3):161–166, 1996.

[13] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[14] Lloyd Shapley and Bernard Grofman. Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 43(3):329–343, 1984.

[15] D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, J. D. F. Habbema, and G. J. Gelpke. Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society. Series A (General)*, 144(2):145–175, 1981.

[16] Kagan Tumer and Joydeep Ghosh. Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition*, 29(2):341–348, 1996.