FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Técnicas de clasificación no contexto de Big Data

Josefa Arán Paredes

2018/2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Técnicas de clasificación no contexto de Big Data

Josefa Arán Paredes

Xullo, 2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística
Título: Técnicas de clasificación no contexto de Big Data
Breve descrición do contido
Trátase de estudar as técnicas de clasificación ou análise discriminante máis importantes e de referencia, e o papel que xogan ditas técnicas no moderno contexto do Big Data.
Recomendacións
Guión aproximado: 1) Técnicas de clasificación lineal ou cuadrática. 2) Técnicas de clasificación non paramétrica. 3) Algunhas técnicas de clasificación adaptadas ao contexto de alta dimensión en Big Data. 4) Ilustración en bases de datos reais.
Outras observacións
Preténdese que o alumno dedique aproximadamente tres meses ao estudo metodolóxico das técnicas correspondentes aos puntos 1), 2) e 3) do guión. Un mes para a avaliación do software dispoñible en R e outro para o desenvolvemento da aplicación con datos simulados ou reais.

Índice xeral

Resumo

O principal obxectivo deste traballo é introducir a análise discriminante e algunhas técnicas de clasificación, centrándose no seu papel no contexto de Big Data. No primeiro capítulo preséntanse os conceptos xerais da clasificación, as regras e a idea de optimalidade. Os principais métodos paramétricos de clasificación (regras lineais e cuadráticas) e algúns non paramétricos (regra dos k veciños máis cercanos e regras tipo kernel) son explicados no segundo capítulo. O terceiro trata sobre a avaliación das técnicas anteriormente definidas cando son aplicadas na práctica. Os dous últimos capítulos introducen os problemas que se poden achar nun contexto de Big Data, particularmente o caso en que a dimensión é maior que o número de observacións. Preséntanse dúas técnicas adaptadas a esta situación e finalmente ilústrase o seu funcionamento cun exercicio de simulación e cunha aplicación a unha base de datos real.

Abstract

This project aims to introduce discriminant analysis and some classification techniques, focusing on their role in the Big Data context. In the first chapter, the general concepts of classification, rules and the idea of optimality are presented. The principal parametric methods (linear and quadratic rules) and some non-parametric ones (k nearest neighbors and kernel rules) are introduced in the second one. The third chapter verses on rule evaluation when applied in practice. The last two chapters introduce the problems encountered in the Big Data context, when the dimension is greater than the sample size in particular. Finally, two techniques adapted to this situation are explained and their performance is demonstrated in a simulation exercise and an application to a real database.

Introdución

A análise discriminante é a disciplina estatística que trata de separar os individuos dunha poboación en distintos grupos ou clases, segundo as súas características. Estes grupos son coñecidos de antemán e tratarase de atopar a regra discriminante óptima segundo certos criterios.

Dependendo das características da poboación coa que tratemos, as regras óptimas construiranse de distinto xeito, partindo do contexto poblacional para despois estimalas e aplicalas a unha mostra. A clase á que pertence cada dato da mostra é coñecida, por iso, dentro do campo do Machine Learning estaríamos a falar de Supervised Learning. Esta terminoloxía máis moderna está asociada tamén ao Big Data. Cara ao final do traballo, comentaranse algunhas técnicas que permitan construir regras cando a dimensión é moi grande, suplindo as deficiencias que xorden ao recorrer aos clasificadores clásicos. Por último, empregaranse as técnicas de clasificación estudadas sobre unha base de datos real ou simulada, ilustrando as vantaxes e carencias dalgúns métodos.

Algunhas notas sobre o Big Data

Nas últimas décadas, os avances tecnolóxicos e o estendido uso dos mesmos fixeron de cada persoa unha productora de datos sociais, que unidos aos xa recollidos de actividades industriais, comerciais e de servizos, conforman unha gran cantidade de información aos que se lles dá o nome de Big Data. Estámolos producindo constantemente, en abundancia e prácticamente gratis, polo que do seu tratamento pódese obter un grande beneficio económico. Isto supuxo o máis recente interese de importantes entidades, o cal sitúa o Big Data asiduamente nos titulares.

Este contexto é moi distinto daquel no que se creou a estatística, onde os datos eran escasos e había que facer suposicións sobre eles para extraer información e chegar a conclusións. O método clásico para facer inferencia sobre unha mostra de certas variables aleatorias non se pode empregar na situación actual. Cada vez os datos son máis hetero-

xéneos e complicados de tratar, como se comenta en Galeano e Peña (2019). Poden ser imaxes, textos ou sons, e non simplemente medidas dunha variable aleatoria real, provintes de moitas poboacións, incluso con máis variables que observacións. Precísanse novos métodos para tratalos, e non basta coa estatística para facelo.

Antes era a estatística a disciplina encargada case por completo da análise de datos: dende o deseño de experimentos ou enquisas, a descripción dos datos e selección dun modelo, a estimación deste e a súa validación, ata a interpretación dos resultados. Agora o papel da estatística vese menguado pola abundancia de datos. Son necesarios novos plantexamentos computacionais cunha filosofía moi distinta. Non só no almacenamento e procesamento de datos para que sexan accesibles e analizables, senón novas técnicas de análise. Os avances computacionais e a rapidez do procesamento dos datos supón tamén un cambio na metodoloxía actual. Búscase reducir a dimensión e moitas veces trátase de achar modelos dispersos, con poucos parámetros que expliquen ou fagan predicción acertadas sobre os datos. Xorde tamén unha importante colaboración entre a estatística e a investigación operativa arredor da optimización. Todo isto expandiuse ata formar a ciencia de datos.

A intelixencia artificial é outro novo campo que mecaniza o proceso do pensamento humano. Unha parte desta, o machine learning trata de entrenar ordenadores con procesos automáticos a partir dos datos. Dende o perceptrón de Rosenblatt, que non é máis que unha regra discriminante linear, pero a primeira e máis simple rede neuronal que imita o cerebro humano nunha computadora, esta área investiga métodos de supervised ou unsupervised learning, que en estadística se corresponden ca análise discriminante e o clustering, respectivamente.

Este traballo fala algo do chamado aprendizaxe supervisado (supervised learning). Algúns dos métodos máis importantes son as redes neuronais, a regra dos k veciños máis cercanos, as árbores de clasificación, as support vector machines (SVM), os random forests ou o clasificador naíf de Bayes. Este último é pouco apreciado pero moi útil na práctica pola súa simplicidade, con mellores resultados que outros métodos de maior interese teórico.

A estatística aporta procesos rigorosos que a comunidade do Machine Learning pasa moitas veces por alto, como a mostraxe, a análise exploratoria e descriptiva, a inferencia, a predicción, a medida da incertidume e a interpretación dos métodos. Baixo o nome de ciencia de datos converxen entón distintos enfoques, con disciplinas coma a intelixencia artificial e a aprendizaxe automática aportando novas maneiras de predicir outputs a partir de inputs, cuxa relación non acaba de entenderse, pero coa estatística como sustento.

Capítulo 1

Introdución á análise discriminante

A análise discriminante é unha técnica estatística que busca construir unha regra que permita asignar un vector aleatorio provinte dunha poboación heteroxénea na que hai unhas clases predefinidas e coñecidas a unha destas clases. Esta disciplina encárgase tamén de establecer e estudar criterios para avaliar a adecuación das regras aos distintos tipos de poboacións, segundo as suposicións que fagamos sobre elas, e tamén o seu comportamento. Quereremos medir a calidade dunha regra minimizando o seu erro de clasificación.

Este estudo farase para o contexto poboacional, considerando un vector aleatorio e as probabilidades de erro de clasificación das regras, todo baseado na distribución deste. Pero na realidade non coñeceremos a poboación, co cal presentarase tamén o caso mostral, partindo dun conxunto de observacións cuxa pertenza a certas clases predeterminadas é coñecida. Será preciso estimar a partir da mostra as regras definidas e tamén o erro.

1.1. Clases, etiquetas, regras e funcións de decisión.

Unha **clase** $\mathcal{C}$, obxecto fundamental da análise discriminante, é a natureza descoñecida dunha observación, ben sexa a especie dunha planta ou o carácter benigno ou maligno dun tumor. Cada clase $\mathcal{C}$ pódese ver como unha poboación que está determinada por unha función de densidade $f(\mathbf{X})$. Son de especial interese as familias de distribucións normais $\mathcal{N}(\boldsymbol{\mu}_\nu, \Sigma_\nu)$ que se diferencian no seu vector de medias ou na matriz de covarianzas. Posto que partimos dun vector aleatorio d-dimensional, as regras de clasificación pódense ver como particións en conxuntos disxuntos dun espazo d-dimensional de vectores aleatorios.

No que segue, o enteiro $\kappa \geq 2$ denotará o número de clases, este número é fixo e coñecido. Chamarémoslle $\mathcal{C}_1, ..., \mathcal{C}_\kappa$ ás κ clases.

Definición 1.1. Consideramos as clases $\mathcal{C}_1, ..., \mathcal{C}_\kappa$, determinadas polas funcións de desidade $f_1, \dots, f_\kappa$. Dicimos que un vector aleatorio d-dimensional $\mathbf{X}$ **pertence á clase** $\mathcal{C}_\nu$ para algún $\nu \leq \kappa$, se $\mathbf{X}$ ten como densidade de probabilidade f_ν, é dicir, satisfai as propiedades que caracterizan $\mathcal{C}_\nu$. Esta pertenza denotámola como

$$\mathbf{X} \in \mathcal{C}_\nu \quad \text{ou} \quad \mathbf{X}^{[\nu]}.$$

Para vectores aleatorios $\mathbf{X}$ pertencentes a unha das κ clases, a **etiqueta** Y de $\mathbf{X}$ é unha variable aleatoria que toma valores discretos 1,...,κ, tal que

$$Y = \nu \quad \text{se} \quad \mathbf{X} \in \mathcal{C}_\nu.$$

Consideramos $\begin{bmatrix} \mathbf{X} \\ Y \end{bmatrix}$ como un vector aleatorio $(d+1)$-dimensional e chamámolo vector aleatorio etiquetado.

Ás clases ás veces chámaselles **poboacións**, e usaremos ambos termos indistintamente. Se as κ clases están caracterizadas polas súas medias $\boldsymbol{\mu}_\nu$ e as súas matrices de covarianzas Σ_ν de xeito que $\mathcal{C}_\nu \equiv (\boldsymbol{\mu}_\nu, \Sigma_\nu)$, escribiremos

$$\mathbf{X} \sim (\boldsymbol{\mu}_\nu, \Sigma_\nu) \quad \text{ou} \quad \mathbf{X} \in \mathcal{C}_\nu.$$

De non dicir o contrario, os vectores aleatorios $\mathbf{X}$ pertencentes ás clases $\mathcal{C}_1, ..., \mathcal{C}_\kappa$ teñen a mesma dimensión d. Posto que se os vectores de distintas clases tivesen dimensións diferentes ou medisen distintas variables, esta información podería empregarse para clasificalos e non precisaríamos as técnicas máis sofisticadas da análise discriminante.

Unha regra é o mecanismo que nos permite asignar un vector aleatorio a unha clase, matemáticamente definímola da seguinte maneira:

Definición 1.2. Sexan κ clases $\mathcal{C}_1, ..., \mathcal{C}_\kappa$ e $\mathbf{X}$ un vector aleatorio que pertence a unha delas. Unha **regra discriminante** ou **clasificador** para $\mathbf{X}$ é unha aplicación $\mathfrak{r}$ que asigna a $\mathbf{X}$ un número $l \in \{1, \dots, \kappa\}$. Escribimos

$$\mathfrak{r}(\mathbf{X}) = l \quad \text{con} \quad 1 \leq l \leq \kappa.$$

A regra $\mathfrak{r}$ **asigna X á clase correcta** ou **clasifica X correcamente** se

$$\mathfrak{r}(\mathbf{X}) = \nu \quad \text{cando} \quad \mathbf{X} \in \mathcal{C}_\nu$$

e **clasifica X incorrectamente** noutro caso.

Son de especial interese as situacións nas que hai dúas clases, nas cales unha función de decisión dá unha expresión alternativa á da regra en cuestión.

Definición 1.3. Sexan $\mathbf{X}$ un vector aleatorio que pertence a unha das dúas clases $\mathcal{C}_1$ ou $\mathcal{C}_2$ e $\mathfrak{r}$ unha regra de discriminante para $\mathbf{X}$. Unha **función de decisión** para $\mathbf{X}$, asociada a $\mathfrak{r}$, é unha función real h definida do seguinte xeito

$$h(\mathbf{X}) > 0 \quad \text{se} \quad \mathfrak{r}(\mathbf{X}) = 1$$

$$h(\mathbf{X}) < 0 \quad \text{se} \quad \mathfrak{r}(\mathbf{X}) = 2.$$

Observación 1.4. Unha función de decisión correspondente a unha regra non é única. Por exemplo, calquera múltiplo dunha función de decisión por un escalar positivo tamén é unha función de decisión para a mesma regra discriminante.

1.1.1. Fronteiras e rexións discriminantes

Posto que unha observación non é máis que un vector aleatorio d-dimensional, construir unha regra pódese ver como definir unha partición en rexións $G_1, \ldots, G_\kappa$ disxuntas de $\mathbb{R}^d$, intentando que os vectores que caian na rexión G_ν sexan da clase $\mathcal{C}_\nu$.

Para un problema de dúas clases, a decisión da pertenza a unha clase pódese basar nunha función de decisión que separa en rexións. As fronteiras entre as rexións son interesantes como axuda visual cando temos dúas clases e dimensión d pequena.

Definición 1.5. Sexa $\mathbf{X}$ un vector aleatorio pertencente a $\mathcal{C}_1$ ou $\mathcal{C}_2$. Se $\mathfrak{r}$ é unha regra discriminante para $\mathbf{X}$ e h é unha función de decisión asociada, definimos a **fronteira de decisión** B da regra $\mathfrak{r}$ como o conxunto formado por tódolos vectores aleatorios $\mathbf{X}$ tales que $h(\mathbf{X}) = 0$.

Cando a regra é linear veremos que a fronteira de decisión é un hiperplano, posto que h é unha función linear. Para regras non lineais, téñense fronteiras máis complexas. Pódese tamén estender a definición de fronteira de decisión para máis de dúas clases, pero estas complícanse ao medrar o número de clases e deixan de ser tan útiles.

Tamén podemos definir unha regra de decisión a partir da fronteira, diferenciando os vectores aleatorios que quedan a un e outro lado de B.

Definición 1.6. Se temos κ clases e $\mathbf{X}$ un vector aletorio, para $\nu \leq \kappa$, a **rexión discriminante** G_ν da regra $\mathfrak{r}$ defínese como

$$G_\nu = \{\mathbf{X} : \mathfrak{r}(\mathbf{X}) = \nu\}.$$

As rexións discriminantes G_ν son disxuntas, xa que cada $\mathbf{X}$ se asigna a unha única clase. Podemos interpretalas como as clases definidas pola regra, isto é, as determinadas por $\mathfrak{r}$. Recíprocamente, dadas as rexións disxuntas, podemos definir unha regra de xeito que $\mathfrak{r}(\mathbf{X}) = \nu$ se $\mathbf{X} \in G_\nu$.

Polo tanto os conceptos de regra e de rexións discriminantes defínense un ao outro, e a fronteira separa as rexións G_ν. Se a regra fose perfecta, entón tódolos puntos na rexión G_ν serían observacións da clase $\mathcal{C}_\nu$ e viceversa.

1.2. Avaliación das regras e probabilidade de clasificación errónea

Nun caso ideal, a nosa regra asignaría a $\mathbf{X}$ a súa etiqueta Y. Pero isto non sempre ocorre, polo que precisamos criterios que axuden a medir a calidade das nosas regras. Para iso, é necesario desenvolver un marco teórico contra o cal contrastar as regras, xa que se só nos fixásemos nos datos dos que dispoñemos poderiamos atopar unha regra que clasificase á perfección as observacións dispoñibles pero fallase ao empregala sobre novos datos. Cando dispoñemos de máis dunha regra, é importante entender cal delas se comporta mellor, e baixo que condicións. Non hai unha única medida do comportamento dunha regra nin unha regra universal que funcione mellor en tódalas situacións.

Cantas máis clases temos, máis erros pode cometer unha regra á hora de clasificar, como se pode ver na seguinte táboa. Nela indícase o que pode ocorrer ao clasificar unha observación en κ clases.

Cadro 1.1: Etiquetas, regras e erros de clasificación

		Asignación da regra			
		1	2	$\cdots$	κ
	1	(1,1)	(1,2)	$\cdots$	(1,κ)
Valor da	2	(2,1)	(2,2)	$\cdots$	(2,κ
etiqueta	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	κ	(κ,1)	(κ,2)	$\cdots$	(κ,κ)

Un dato só estaría ben clasificado cando a asignación pola regra coincide coa súa etiqueta, que serían os casos da diagonal. Así, a probabilidade de acertar de maneira azarosa

sería $\frac{1}{\kappa}$. No caso de ter dúas clases, asignar ao azar ten un 50 % de probabilidades de acerto, pero clasificar correctamente deste xeito vólvese máis difícil cantas máis clases hai, pois a probabilidade de acerto diminúe ao medrar κ. A probabilidade de cometer un erro sería entón $\frac{\kappa-1}{\kappa}$, que tende ao 100 % a medida que aumenta o número de clases κ.

Cando construamos unha regra, sempre quereremos que a proporción de observacións mal clasificadas fronte ao total sexa menor que $\frac{\kappa-1}{\kappa}$, pois de non ser así, daríanos peores resultados que asignando cada observación a unha clase ao azar.

Non só iso, debemos averiguar como de boa pode chegar a ser a nosa regra, ou cal será a mellor regra entre unha clase de regras específica como poden ser as lineais. Haberá problemas para os que esa regra óptima nunca sexa mellor que decidir ao azar, é dicir, cuxo erro mínimo sexa maior que $\frac{\kappa-1}{\kappa}$ para κ clases. Os criterios que permiten determinar cal é a regra óptima deben medirse tanto en termos da distribución das variables aleatorias como da mostra, co cal é preciso achar maneiras de estimar o erro cometido.

Unha maneira natural de cuantificar o bo comportamento dunha regra sería contar o número de clasificacións erróneas fronte ao total de observacións que intentamos clasificar, pero serán necesarias medidas máis complexas para medir se unha regra se comporta como é desexable, que introduciremos ao longo desta sección.

1.2.1. O problema de Bayes

Comezamos introducindo os conceptos para o caso de dúas clases, aínda que se poden estender facilmente a problemas con máis clases.

Definición 1.7. Sexa $\mathbf{X}$ un vector aleatorio que pertence a unha das clases $\mathcal{C}_1$ ou $\mathcal{C}_2$ e sexa Y a etiqueta de $\mathbf{X}$. Consideramos $\mathfrak{r}$ unha regra para $\mathbf{X}$, e definimos a **probabilidade de clasificación errónea** ou **probabildade de erro** de $\mathfrak{r}$ como

$$L(\mathfrak{r}) = \mathbb{P}\{\mathfrak{r}(\mathbf{X}) \neq Y\}.$$

Quereremos entón atopar regras para as que a probabilidade de erro sexa o máis pequena posible. Definimos a regra que minimiza a probabilidade de erro de clasificación entre tódalas posibles regras como segue:

$$\mathfrak{r}^*(\mathbf{X}) = \underset{\mathfrak{r}:\mathbb{R}^d \to \{1,\dots,\kappa\}}{\arg\min}\ \mathbb{P}(\mathfrak{r}(\mathbf{X}) \neq Y).$$

Esta regra depende da distribución do vector etiquetado $\begin{bmatrix} \mathbf{X} \\ Y \end{bmatrix}$, e se a coñecemos, poderemos calcular a regra $\mathfrak{r}^*$ explícitamente. O máis habitual é que descoñezamos a súa distribución, e con ela tamén a regra $\mathfrak{r}^*$. Ao problema de achar esta regra chamarémoslle

problema de Bayes. No seguinte resultado veremos unha expresión alternativa para $\mathfrak{r}^*$, denominada **regra de Bayes**.

Proposición 1.8. *Sexa $\boldsymbol{X}$ un vector aleatorio pertencente a unha das clases $\mathcal{C}_1$ ou $\mathcal{C}_2$ coa súa etiqueta Y. Se*

$$p(\boldsymbol{X}) = \mathbb{P}\{Y = 1|\boldsymbol{X}\}$$

é a probabilidade condicional de $Y = 1$ dado $\boldsymbol{X}$ e definimos a regra discriminante $\mathfrak{r}^*$ como

$$\mathfrak{r}^*(\boldsymbol{X}) = \begin{cases} 1 & se \quad p(\boldsymbol{X}) > 1/2 \\ 2 & noutro\ caso. \end{cases}$$

Dada $\mathfrak{r}$ outra regra discriminante para $\boldsymbol{X}$, entón

$$\mathbb{P}\{\mathfrak{r}^*(\boldsymbol{X}) \neq Y\} \leq \mathbb{P}\{\mathfrak{r}(\boldsymbol{X}) \neq Y\}$$

Demostración. Considerada a regra $\mathfrak{r}$ para $\mathbf{X}$, tense

$$\mathbb{P}\{\mathfrak{r}(\mathbf{X}) \neq Y|\mathbf{X}\} = 1 - \mathbb{P}\{\mathfrak{r}(\mathbf{X}) = Y|\mathbf{X}\}$$

$$= 1 - (\mathbb{P}\{Y = 1, \mathfrak{r}(\mathbf{X}) = 1|\mathbf{X}\} + \mathbb{P}\{Y = 2, \mathfrak{r}(\mathbf{X}) = 2|\mathbf{X}\})$$

$$= 1 - \left(I_{\{\mathfrak{r}(\mathbf{X})=1\}}\mathbb{P}\{Y = 1|\mathbf{X}\} + I_{\{\mathfrak{r}(\mathbf{X})=2\}}\mathbb{P}\{Y = 2|\mathbf{X}\}\right)$$

$$= 1 - (I_{\{\mathfrak{r}(\mathbf{X})=1\}}p(\mathbf{X}) + I_{\{\mathfrak{r}(\mathbf{X})=2\}}(1 - p(\mathbf{X}))$$

onde I_G é a función indicador do conxunto G. Empregando os cálculos anteriores, obtemos

$$\delta^* := \mathbb{P}\{\mathfrak{r}^*(\mathbf{X}) \neq Y\} - \mathbb{P}\{\mathfrak{r}(\mathbf{X}) \neq Y\}$$

$$= p(\mathbf{X})(I_{\{\mathfrak{r}^*(\mathbf{X})=1\}} - I_{\{\mathfrak{r}(\mathbf{X})=1\}}) + (1 - p(\mathbf{X}))(I_{\{\mathfrak{r}^*(\mathbf{X})=2\}} - I_{\{\mathfrak{r}(\mathbf{X})=2\}})$$

$$= (2p(\mathbf{X}) - 1)(I_{\{\mathfrak{r}^*(\mathbf{X})=1\}} - I_{\{\mathfrak{r}(\mathbf{X})=1\}})$$

posto que $I_{\{\mathfrak{r}(\mathbf{X})=2\}} = 1 - I_{\{\mathfrak{r}(\mathbf{X})=1\}}$. Finalmente, pola definición de $\mathfrak{r}^*$, chegamos a que $\delta^* \geq 0$. □

Esta proposición proba a optimalidade da regra $\mathfrak{r}^*$. A probabilidade de erro desta regra é o que denominamos **erro de Bayes**

$$L^* = L(\mathfrak{r}^*) = \mathbb{P}(\mathfrak{r}^*(\mathbf{X}) \neq Y)$$

e é o mínimo erro que podemos esperar cometer ao clasificar. Podemos calculalo cando a distribución do vector etiquetado é coñecida.

Observación 1.9. Como temos dúas clases, dada $\mathbf{X}$, $\mathbb{P}(Y=2|\mathbf{X}) = 1 - p(\mathbf{X})$, polo tanto

$$\mathbb{P}(Y=1|\mathbf{X}) > \mathbb{P}(Y=2|\mathbf{X}) \quad \Leftrightarrow \quad \mathbb{P}(Y=1|\mathbf{X}) - \mathbb{P}(Y=2|\mathbf{X}) > 0$$

$$\Leftrightarrow \quad p(\mathbf{X}) - (1-p(\mathbf{X})) = 2p(\mathbf{X}) - 1 > 0 \quad \Leftrightarrow \quad p(\mathbf{X}) > 1/2.$$

Co cal podemos reescribir $\mathfrak{r}^*$ como

$$\mathfrak{r}^*(\mathbf{X}) = \begin{cases} 1 & \text{se} \quad \mathbb{P}(Y=1|\mathbf{X}) > \mathbb{P}(Y=2|\mathbf{X}) \\ 2 & \text{noutro caso.} \end{cases}$$

É dicir, a regra de Bayes asigna unha observación $\mathbf{X}$ á clase á que é máis probable que pertenza, como podía suxerir a nosa intuición.

1.2.2. A regra de Bayes

Introducimos un entorno probabilístico de xeito que $\begin{bmatrix} \mathbf{X} \\ Y \end{bmatrix}$ é un vector aleatorio etiquetado que toma valores en $\mathbb{R}^d \times \{1, \dots, \kappa\}$, cunha certa distribución e que representa a probabilidade de atoparnos certos pares vector-etiqueta na práctica. Cando $Y = \nu$, a distribución de $\mathbf{X}$ virá dada por f_ν. Para avaliar as regras dende un punto de vista teórico é importante definir o concepto de probabilidade do erro, no que se fai fincapé en Devroye et al. (1996), o cal quereremos acotar co obxectivo de ter un indicador da propensión dos nosos datos a ser clasificados correctamente.

Nun contexto bayesiano con clases $\mathcal{C}_1, ..., \mathcal{C}_\kappa$ supoñemos que coñecemos a probabilidade de que unha observación pertenza á clase $\mathcal{C}_l$ e denotámola π_l, é dicir, $\pi_l = \mathbb{P}(\mathbf{X} \in \mathcal{C}_l)$. As probabilidades $\pi_1, \dots, \pi_\kappa$ chámanse **probabilidades a priori**. Se dispoñemos dunha mostra, tomamos como probabilidades a priori a proporción de observacións que pertencen a cada clase.

Incluir as probabilidades a priori pode enriquecer a clasificación cando son substancialmente distintas entre as clases. Por exemplo, se temos dúas clases e sabemos que dous tercios dos datos proveñen da primeira clase e o tercio restante da segunda, unha regla debería asignar máis probablemente unha observación á clase 1. Se dispoñemos desta información, incorporada de xeito coidadoso no problema de decisión pode mellorar as nosas regras.

Definición 1.10. Sexan $\mathcal{C}_1, ..., \mathcal{C}_\kappa$ clases que se diferencian nas súas medias e matrices de covarianzas. Sexa $\mathbf{X}$ un vector aleatorio pertencente a unha das clases. Consideramos as probabilidades a priori $\pi_1, \dots, \pi_\kappa$ asociadas ás clases. Sexa $\mathfrak{r}$ unha regra discriminante derivada das rexións G_ν.

1. A **probabilidade condicional** de que $\mathfrak{r}$ asigne $\mathbf{X}$ á clase $\mathcal{C}_\nu$, cando $\mathbf{X}$ é da clase $\mathcal{C}_l$ é
$$p(\nu|l) = \mathbb{P}\{\mathfrak{r}(\mathbf{X}) = \nu|\mathbf{X} \in \mathcal{C}_l\}.$$

2. A **probabilidade a posteriori** de que unha observación $\mathbf{X}$ pertenza á clase $\mathcal{C}_l$ cando a regra lle asigna o valor ν é
$$\mathbb{P}\{\mathbf{X} \in \mathcal{C}_l|\mathfrak{r}(\mathbf{X}) = \nu\} = \frac{\mathbb{P}\{\mathfrak{r}(\mathbf{X}) = \nu|\mathbf{X} \in \mathcal{C}_l\}\pi_l}{\mathbb{P}\{\mathfrak{r}(\mathbf{X}) = \nu\}} = \frac{p(\nu|l)\pi_l}{\mathbb{P}\{\mathfrak{r}(\mathbf{X}) = \nu\}}.$$

3. A **probabilidade de clasificación errónea** ou **probabildade de erro** de $\mathfrak{r}$ defínese como
$$L(\mathfrak{r}) = \mathbb{P}\{\mathfrak{r}(\mathbf{X}) \neq Y\},$$
e usando as probabilidades condicionais e a priori ten a forma
$$L(\mathfrak{r}) = \sum_{\nu \neq l} p(\nu|l)\pi_l.$$

Posto que as rexións discriminantes e as regras discriminantes se determinan unhas a outras, a probabilidade de asignar $\mathbf{X}$ a G_ν cando $\mathbf{X} \in \mathcal{C}_l$ é
$$p(\nu|l) = \int_{G_\nu} f_l(\mathbf{x})d\mathbf{x} = \int I_{G_\nu} f_l,$$
sendo I_{G_ν} a función característica da rexión ν-ésima e f_l a función de densidade da clase l-ésima. Tense entón que a probabilidade de acertar ao asignar $\mathbf{X}$ á clase $\mathcal{C}_\nu$ é
$$\mathbb{P}\{\mathbf{X} \in \mathcal{C}_\nu|\mathfrak{r}(\mathbf{X}) = \nu\} = \frac{p(\nu|\nu)\pi_\nu}{\mathbb{P}\{\mathfrak{r}(\mathbf{X}) = \nu\}},$$
mentres que a probabilidade de clasificar correctamente unha observación da clase ν-ésima é $p(\nu|\nu)$ e a de cometer un erro é $1 - p(\nu|\nu)$.

O seguinte Teorema define a regra de Bayes e da a súa expresión.

Teorema 1.11. *Sexan $\mathcal{C}_1, ..., \mathcal{C}_\kappa$ clases con distintas medias e $\boldsymbol{\pi} = [\pi_1, \dots, \pi_\kappa]$ as probabilidades a priori asociadas. Sexa $\boldsymbol{X}$ un vector aleatorio dunha das κ clases e f_ν a función de densidade da clase $\mathcal{C}_\nu$. Definimos as rexións*
$$G_\nu = \{\boldsymbol{X} : f_\nu(\boldsymbol{X})\pi_\nu = \max_{1 \leq l \leq \kappa} [f_l(\boldsymbol{X})\pi_l]\},$$
e consideramos $\mathfrak{r}_{Bayes}$ a regra definida a partir das G_ν de tal xeito que $\mathfrak{r}_{Bayes}(\boldsymbol{X}) = \nu$ se $\boldsymbol{X} \in G_\nu$.

A regra $\mathfrak{r}_{Bayes}$ asigna $\boldsymbol{X}$ á clase $\mathcal{C}_\nu$ en preferencia a $\mathcal{C}_l$ cando $\frac{f_\nu(\boldsymbol{X})}{f_l(\boldsymbol{X})} > \frac{\pi_l}{\pi_\nu}$.

Demostración. A proba séguese directamente da definición da regra $\mathfrak{r}_{Bayes}$, xa que se ten

$$f_\nu(\mathbf{X})\pi_\nu = \max_{1\leq l\leq \kappa}[f_l(\mathbf{X})\pi_l] \quad \Leftrightarrow \quad f_\nu(\mathbf{X})\pi_\nu \geq f_l(\mathbf{X})\pi_l \quad \forall l \quad \Leftrightarrow \quad \frac{f_\nu(\mathbf{X})}{f_l(\mathbf{X})} \geq \frac{\pi_l}{\pi_\nu} \quad \forall l.$$

□

Chamamos a $\mathfrak{r}_{Bayes}$ **regra (discriminante) de Bayes**. Empregando esta regra, a probabilidade de asignar $\mathbf{X}$ á clase correcta é

$$\mathfrak{p} = \sum_{\nu=1}^{\kappa} \mathbb{P}\{\mathfrak{r}_{Bayes}(\mathbf{X}) = \nu | \mathbf{X} \in \mathcal{C}_\nu\}\pi_\nu = \sum_{\nu=1}^{\kappa} \int I_{G_\nu} f_\nu \pi_\nu,$$

e a función característica I_{G_ν} cumpre que

$$I_{G_\nu}(\mathbf{X}) = \begin{cases} 1 & \text{se} \quad f_\nu(\mathbf{X})\pi_\nu \geq f_l(\mathbf{X})\pi_l \quad \forall l \\ 0 & \text{noutro caso.} \end{cases}$$

Se só tiveramos dúas clases, esta sería a regra $\mathfrak{r}^*$ definida na sección anterior.

1.2.3. Pérdida e risco de Bayes

Cando traballamos nun contexto bayesiano, a pérdida e o risco adoitan usarse para avaliar o comportamento dun método e tamén comparalo con outros diferentes. Estas ideas pertencen ao campo da teoría de decisión e nos adaptarémolas ás regras discriminantes.

Á hora de clasificar, podemos acertar, obter un resultado 'non totalmente correcto' ou chegar a unha clasificación moi incorrecta. O grao de acerto está explicado por unha función de pérdida K que asigna un coste ou pérdida a unha decisión incorrecta.

Definición 1.12. Sexan $\mathfrak{C}$ unha colección de clases $\mathcal{C}_1, ..., \mathcal{C}_\kappa$ e $\mathbf{X}$ un vector aleatorio dunha destas clases. Sexa $\mathfrak{r}$ unha regra discriminante para $\mathbf{X}$.

1. Unha **función de pérdida** K é unha aplicación que leva $\mathbf{X}$ e a regra $\mathfrak{r}$ nun número non negativo chamado **pérdida** ou **coste**. Se $\mathbf{X} \in \mathcal{C}_l$ e $\mathfrak{r}(\mathbf{X}) = \nu$, entón a pérdida $c_{l,\nu}$ producida por tomar a decisión ν cando a clase real era l é

$$\mathsf{K}(\mathbf{X}, \mathfrak{r}) = c_{l,\nu} \quad \text{con} \quad \begin{cases} c_{l,\nu} = 0 & \text{se} \quad l = \nu, \\ c_{l,\nu} > 0 & \text{noutro caso.} \end{cases}$$

2. A **función de risco** R é a pérdida esperada producida ao usar a regra $\mathfrak{r}$. Escribimos

$$\mathsf{R}(\nu, \mathfrak{r}) = \mathbb{E}[\mathsf{K}(\mathbf{X}, \mathfrak{r})],$$

onde a esperanza se toma respecto á distribución de $\mathbf{X}$, dada por f_ν.

3. Sexan $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_\kappa]$ as probabilidades a priori para as clases de $\mathfrak{C}$. O **risco de Bayes** B dunha regra discriminante $\mathfrak{r}$ respecto ás probabilidades a priori $\boldsymbol{\pi}$ é

$$\mathsf{B}(\boldsymbol{\pi}, \mathfrak{r}) = \mathbb{E}_\pi[\mathsf{R}(\mathfrak{C}, \mathfrak{r})],$$

onde a media se toma respecto $\boldsymbol{\pi}$.

O risco considera a pérdida en tódalas clases para unha regra específica. Agora temos unha ferramenta para escoller entre dúas regras, eliximos a que teña o menor risco. É difícil atopar unha regra que funciona ben para tódalas clases e a información adicional proporcionada polas probabilidades a priori pode facilitar a toma de decisións.

Observación 1.13. O criterio que nós estamos tomando para decidir cal é a mellor regra, aquela que minimiza o risco de Bayes, é o criterio de Bayes. Se temos dúas clases, tense que

$$\mathsf{R}(l, \mathbf{X}) = \mathbb{E}[\mathsf{K}(\mathbf{X}, \mathfrak{r})] = c_{l,1}\mathbb{P}(\mathfrak{r}(\mathbf{X}) = 1) + c_{l,2}\mathbb{P}(\mathfrak{r}(\mathbf{X}), 2)$$

$$\text{e} \quad \mathsf{B}(\boldsymbol{\pi}, \mathfrak{r}) = \pi_1\mathsf{R}(1, \mathfrak{r}) + \pi_2\mathsf{R}(2, \mathfrak{r}).$$

Polo que o criterio de Bayes supón minimizar

$$\min_{\mathfrak{r}}\{\pi_1\mathsf{R}(1, \mathfrak{r}) + \pi_2\mathsf{R}(2, \mathfrak{r})\}.$$

Este criterio non é único, podemos ter unha postura pesimista e intentar minimizar o risco máximo que poida cometerse nalgunha das clases. O criterio minimax para dúas clases consistiría en

$$\min_{\mathfrak{r}}\left\{\max\{\mathsf{R}(1, \mathfrak{r}), \mathsf{R}(2, \mathfrak{r})\}\right\}.$$

O tipo máis común de pérdida é a pérdida *cero-un*, que toma $c_{l,\nu} = 0$ cando $l = \nu$ e $c_{l,\nu} = 1$ cando non coinciden. Estos costes din soamente cando non asignamos o vector á clase á que realmente pertence. En xeral, se queremos graduar como de incorrecta foi a clasificación, os costes $c_{l,\nu}$ e $c_{\nu,l}$ poderían ser distintos. Nós empregaremos a pérdida cero-un, respecto á cal a regra de Bayes ten unha interpretación en función das probabilidades a posteriori. Se $\mathbf{X}$ é da clase l-ésima,

$$\mathsf{K}(\mathbf{X}, \mathfrak{r}) = \begin{cases} 0 & \text{se} \quad f_l(\mathbf{X})\pi_l \geq f_\nu(\mathbf{X})\pi_\nu \quad \forall \nu \leq \kappa, \\ 1 & \text{noutro caso.} \end{cases}$$

Co seguinte Teorema relacionaremos o erro de Bayes e a regra de Bayes definidas na sección anterior coas ideas de pérdida e risco de Bayes.

Teorema 1.14. *Sexa $\mathfrak{C}$ unha colección de clases $\mathcal{C}_1, ..., \mathcal{C}_\kappa$ e $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_\kappa]$ o vector de probabilidades a priori asociadas a $\mathfrak{C}$. Sexa $\boldsymbol{X}$ un vector aleatorio pertencente a unha das clases $\mathcal{C}_\nu$, $\mathfrak{r}_{Bayes}$ a regra de Bayes e G_ν as rexións discriminantes dadas por*

$$G_\nu = \{\boldsymbol{X} : f_\nu(\boldsymbol{X})\pi_\nu = \max_{1\leq l\leq \kappa}[f_l(\boldsymbol{X})\pi_l]\}.$$

Para a pérdida cero-un, a regra de Bayes é óptima entre todas as regras discriminantes no sentido de que ten:

1. *a maior probabilidade de asignar $\boldsymbol{X}$ á clase correcta, e*

2. *o menor risco de Bayes para a función de pérdida cero-un.*

Demostración. Pola definición da función característica de G_ν, tense que a función de pérdida vale cero $\mathsf{K}(\mathbf{X}, \mathfrak{r}) = 0 \quad \Leftrightarrow \quad \mathbf{X} \in G_\nu \quad \Leftrightarrow \quad I_{G_\nu}(\mathbf{X}) = 1$.

Sexa $\mathfrak{r}_{Bayes}$ a regra de Bayes e supoñamos que existe outra regra $\mathfrak{r}'$ baseada na mesma función de pérdida e que a súa probabilidade de asignar $\mathbf{X}$ á clase correcta é maior que a de $\mathfrak{r}_{Bayes}$. Denotamos por $p'(\nu|\nu)$ a probabilidade de que $\mathfrak{r}'$ clasifique correctamente unha observación da clase $\mathcal{C}_\nu$, e G'_ν son as rexións discriminantes asociadas a $\mathfrak{r}'$. Se p' é a probabilidade de clasificar correctamente $\mathbf{X}$ usando $\mathfrak{r}'$, entón

$$\mathfrak{p}' = \sum_{\nu=1}^{\kappa} p'(\nu|\nu)\pi_\nu = \sum_{\nu=1}^{\kappa} \int I_{G'_\nu} f_\nu \pi_\nu$$

$$\leq \sum_{\nu=1}^{\kappa} \int I_{G'_\nu} \max_\nu\{f_\nu\pi_\nu\} = \sum_{\nu=1}^{\kappa} \int_{G'_\nu} \max_\nu\{f_\nu\pi_\nu\}$$

$$= \int \max_\nu\{f_\nu\pi_\nu\} = \sum_{\nu=1}^{\kappa} \int_{G_\nu} \max_\nu\{f_\nu\pi_\nu\} = \sum_{\nu=1}^{\kappa} \int I_{G_\nu} f_\nu \pi_\nu$$

$$= \sum_{\nu=1}^{\kappa} p(\nu|\nu)\pi_\nu = \mathfrak{p}.$$

Este cálculo contradí que $\mathfrak{r}'$ leve a unha probabilidade de asignar correctamente mellor, é dicir, que $\mathfrak{p}' > \mathfrak{p}$. Polo que $\mathfrak{r}_{Bayes}$ é óptima.

Para a segunda parte do Teorema, partimos de que,

$$\mathsf{B}(\boldsymbol{\pi}, \mathfrak{r}) = \sum_{\nu=1}^{\kappa} \pi_\nu \mathsf{R}(\nu, \mathfrak{r}) = \sum_{\nu=1}^{\kappa} \pi_\nu \int \mathsf{K}(\mathbf{x}, \mathfrak{r}) f_\nu(\mathbf{x}) d\mathbf{x} =$$

$$= \sum_{\nu=1}^{\kappa} \pi_\nu \int_{\mathbb{R}^d - G_\nu} f_\nu \quad = \sum_{\nu=1}^{\kappa} \pi_\nu \left(\int_{\mathbb{R}^d} f_\nu - \int_{G_\nu} f_\nu \right) =$$

$$= \sum_{\nu=1}^{\kappa} \pi_\nu \left(1 - \int_{G_\nu} f_\nu\right) \quad = \sum_{\nu=1}^{\kappa} \pi_\nu - \sum_{\nu=1}^{\kappa} \pi_\nu \int_{G_\nu} f_\nu =$$

$$= 1 - \sum_{\nu=1}^{\kappa} \int_{G_\nu} f_\nu \pi_\nu.$$

Sexa $\mathfrak{r}'$ unha regra calquera, entón

$$\mathsf{B}(\boldsymbol{\pi}, \mathfrak{r}') = 1 - \sum_{\nu=1}^{\kappa} \int_{G'_\nu} f_\nu \pi_\nu \geq 1 - \sum_{\nu=1}^{\kappa} \int_{G'_\nu} \max_\nu f_\nu \pi_\nu =$$

$$= 1 - \int \max_\nu f_\nu \pi_\nu = 1 - \sum_{\nu=1}^{\kappa} \int_{G_\nu} f_\nu \pi_\nu = \mathsf{B}(\boldsymbol{\pi}, \mathfrak{r}_{Bayes})$$

Polo tanto, $\mathfrak{r}_{Bayes}$ ten o menor risco de Bayes de tódalas regras discriminantes. □

A optimalidade da regra de Bayes ten moito interés teórico, permítenos saber como de ben se desempeña unha regra e se o fai tan ben como a de Bayes, asintóticamente.

1.3. Análise discriminante no caso mostral

Agora ocupámonos do caso no que temos unha mostra. Na práctica, non coñeceremos a poboación, senón que disporemos dun conxunto de datos ou mostra desta poboación. Sexa $\mathbb{X} = [\mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_n]$ unha matriz de datos $d \times n$. Supoñemos que cada $\mathbf{X}_i$ pertence a unha das clases $\mathcal{C}_\nu$. Se a orde dos $\mathbf{X}_i$ non importa, por comodidade agruparemos as observacións e escribiremos

$$\mathbb{X}^{[\nu]} = \left\{ \mathbf{X}_i^{[\nu]} : i = 1, \ldots, n_\nu \right\},$$

onde $\mathbf{X}_i^{[\nu]}$ é o vector aleatorio etiquetado, que pertence á clase $\mathcal{C}_\nu$. Reagrupando, tense

$$\mathbb{X} = [\mathbb{X}^{[1]} \mathbb{X}^{[2]} \ldots \mathbb{X}^{[\kappa]}].$$

Así, o número de observacións da clase ν-ésima $\mathcal{C}_\nu$ é n_ν. Este número varía dunha clase a outra, e o total será $n = \sum_{\nu=1}^{\kappa} n_\nu$.

Precisamos un mecanismo que nos permita asignar unha observación á clase á que pertence. Igual que un médico diagnostica unha enfermidade analizando os síntomas que presenta un paciente, as nosas regras clasificarán os datos en distintos grupos. E do mesmo xeito que un experto debe estudar e ver moitos casos antes de saber de que doenza se trata, nós debemos entrenar as regras empregando os datos dispoñibles para que nos leven á mellor decisión posible.

Podemos partir da regra de Bayes e estimar as funcións de densidade de cada unha das clases f_ν a partir das observacións pertencentes a elas mediante $\hat{f}_\nu$. Así, a regra quedaría

$$\hat{\mathfrak{r}}(\mathbf{X}) = \nu \quad \text{se} \quad \frac{\hat{f}_\nu(\mathbf{X})}{\hat{f}_l(\mathbf{X})} \geq \frac{\pi_l}{\pi_\nu} \quad \forall l,$$

sendo $\pi_l = \frac{n_\nu}{n}$ a proporción de observacións pertencentes a cada clase.

Podemos facer suposicións acerca da poboación para estimar a densidade. Se lle pedimos aos datos que sexan normais, para estimar f_ν non teremos máis que estimar os parámetros $\boldsymbol{\mu}_\nu$ e Σ_ν, xa que estamos nun contexto paramétrico. En cambio, se non facemos ningunha hipótese sobre os datos, teremos que recorrer a técnicas non paramétricas.

Unha vez cosntruidas as regras, é necesario avalialas. Hai moitas maneiras de estimar o erro dunha regra a partir dos datos, como veremos no Capítulo 3.

Capítulo 2

Técnicas de clasificación

Agora que xa presentamos o problema de clasificación nun contexto poboacional e mostral, pasamos á construción das regras discriminantes. Estas regras e as súas deduccións poden verse en textos clásicos coma Mardia et al. (1979) e noutros máis modernos coma Koch (2014), Hastie et al. (2001) ou James et al. (2013). En primeiro lugar presentaremos as regras lineais e cuadráticas, que son resultado de supoñer a normalidade das clases. A continuación introduciremos un par de técnicas de clasificación non paramétrica: a regra dos k veciños máis cercanos e as regras de tipo kernel. Existen moitas máis regras que por unha cuestión de brevidade non se inclúen neste traballo.

2.1. Regras discriminantes lineais

As regras discriminantes lineais son as primeiras que aparecen, presentadas nun inicio por Fisher. Trátase de regras cuxa función de decisión asociada é lineal, é dicir

$$h(\mathbf{X}) = a^\intercal \mathbf{X} + a_0 = \sum_{i=1}^{d} a_i X_i + a_0,$$

onde $a \in \mathbb{R}^d$ é un vector de pesos e a_0 un escalar, e X_i son as compoñentes do vector aleatorio $\mathbf{X}$.

2.1.1. A regra discriminante de Fisher

A principios do s.XX, Fisher tivo a idea de particionar os datos de modo que a variabilidade dentro de cada clase fose pequena e a variabilidade entre as distintas clases fose grande. Así, plantexou un dos primeiros métodos de análise discriminante. Trátasede atopar a dirección $\mathbf{e} \in \mathbb{R}^d$ que minimiza a varianza dentro de cada clase e maximiza a varianza entre clases de $\mathbf{e}^\intercal \mathbf{X}$, para poder traballar con cantidades unidimensionais. Fisher

considerou clases identificadas pola súa media e matriz de covarianzas $\mathcal{C}_\nu = \mathcal{N}(\boldsymbol{\mu}_\nu, \Sigma_\nu)$ e vectores aleatorios $\mathbf{X}^{[\nu]}$ de cada unha delas. Presentamos a continuación os elementos necesarios para a súa construcción.

Definición 2.1. Sexan $\mathbf{X}^{[\nu]} \sim (\boldsymbol{\mu}_\nu, \Sigma_\nu)$ vectores aleatorios etiquetados con $\nu \leq \kappa$. Se $\bar{\boldsymbol{\mu}} = \frac{1}{\kappa} \sum_{\nu=1}^{\kappa} \boldsymbol{\mu}_\nu$ é a media das medias de cada clase, e $\mathbf{e}$ un vector unitario d-dimensional, definimos:

1. A **between-class variability** ou variabilidade entre as clases $\mathfrak{b}$ é

$$\mathfrak{b}(\mathbf{e}) = \sum_{\nu=1}^{\kappa} |\mathbf{e}^\intercal(\boldsymbol{\mu}_\nu - \bar{\boldsymbol{\mu}})|^2$$

2. A **within-class variability** ou variabilidade dentro das clases $\mathfrak{m}$ é

$$\mathfrak{w}(\mathbf{e}) = \sum_{\nu=1}^{\kappa} \operatorname{var}(\mathbf{e}^\intercal \mathbf{X}^{[\nu]})$$

3. Para $\mathfrak{q}(\mathbf{e}) = \frac{\mathfrak{b}(\mathbf{e})}{\mathfrak{w}(\mathbf{e})}$, o **discriminante de Fisher** $\mathfrak{d}$ é

$$\mathfrak{d} = \max_{\{\mathbf{e}/\|\mathbf{e}\|=1\}} \mathfrak{q}(\mathbf{e})$$

Tanto $\mathfrak{b}$ como $\mathfrak{w}$ son funcións de $\mathbf{e}$, mentres que $\mathfrak{d}$ non é unha función, senón un escalar que busca maximizar a variabilidade entre clases con $\mathfrak{b}$ e minimizala dentro de cada clase con $\mathfrak{w}$. Podemos preguntarnos como atopar a dirección $\mathbf{e}$ na que se acada este máximo de $\mathfrak{q}$ e qué significa dito máximo, $\mathfrak{d}$.

Teorema 2.2. *Sexan $\boldsymbol{X}^{[\nu]} \sim (\boldsymbol{\mu}_\nu, \Sigma_\nu)$ vectores aleatorios etiquetados e $\bar{\boldsymbol{\mu}} = \frac{1}{\kappa} \sum_{\nu=1}^{\kappa} \boldsymbol{\mu}_\nu$. Definimos as matrices*

$$B = \sum_{\nu=1}^{\kappa} (\boldsymbol{\mu}_\nu - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_\nu - \bar{\boldsymbol{\mu}})^\intercal \quad e \quad W = \sum_{\nu=1}^{\kappa} \Sigma_\nu$$

e supoñemos que W é invertible. Sexa $\mathbf{e}$ un vector unitario d-dimensional, entón cúmprese:

1. *A between-class variability relaciónase con B mediante*

$$\mathfrak{b}(\mathbf{e}) = \mathbf{e}^\intercal B \mathbf{e}$$

2. *A within-class variability relaciónase con W mediante*

$$\mathfrak{w}(\mathbf{e}) = \mathbf{e}^\intercal W \mathbf{e}$$

3. *O maior autovalor de* $W^{-1}B$ *é o discriminante de Fisher* $\mathfrak{d}$.

4. *Se* $\boldsymbol{\eta}$ *maximiza o cociente* $\mathfrak{q}(\mathbf{e})$ *sobre tódolos vectores unitarios* $\mathbf{e}$*, entón* $\boldsymbol{\eta}$ *é o autovector de* $W^{-1}B$ *correspondente a* $\mathfrak{d}$.

Demostración. Probaremos a relación entre as variabilidades e as matrices B e W simplemente empregando a definición dos conceptos e que $\mathbf{e}$ non depende do índice das clases ν.

$$\mathfrak{b}(\mathbf{e}) = \sum_{\nu=1}^{\kappa} (\mathbf{e}^\intercal(\boldsymbol{\mu}_\nu - \bar{\boldsymbol{\mu}}))(\mathbf{e}^\intercal(\boldsymbol{\mu}_\nu - \bar{\boldsymbol{\mu}}))^\intercal = \sum_{\nu=1}^{\kappa} \mathbf{e}^\intercal(\boldsymbol{\mu}_\nu - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_\nu - \bar{\boldsymbol{\mu}})^\intercal \mathbf{e} =$$

$$= \mathbf{e}^\intercal(\sum_{\nu=1}^{\kappa} (\boldsymbol{\mu}_\nu - \bar{\boldsymbol{\mu}}_\nu)(\boldsymbol{\mu}_\nu - \bar{\boldsymbol{\mu}}_\nu)^\intercal)\mathbf{e} = \mathbf{e}^\intercal B \mathbf{e}.$$

Agora, empregando a bilinealidade da matriz de covarianzas, obtemos:

$$\mathfrak{w}(\mathbf{e}) = \sum_{\nu=1}^{\kappa} \mathrm{var}(\mathbf{e}^\intercal \mathbf{X}^{[\nu]}) = \sum_{\nu=1}^{\kappa} \mathbf{e}^\intercal \mathrm{var}(\mathbf{X}^{[\nu]})\mathbf{e} = \mathbf{e}^\intercal(\sum_{\nu=1}^{\kappa} \Sigma_\nu)\mathbf{e} = \mathbf{e}^\intercal W \mathbf{e}.$$

Por último, para probar 3. e 4. partimos de que a derivada de $\mathfrak{q}(\mathbf{e}) = \frac{\mathbf{e}^\intercal B \mathbf{e}}{\mathbf{e}^\intercal W \mathbf{e}}$ é

$$\frac{d\mathfrak{q}}{d\mathbf{e}} = \frac{2}{\mathbf{e}^\intercal W \mathbf{e}}[B\mathbf{e} - \mathfrak{q}(\mathbf{e})W\mathbf{e}].$$

O máximo valor de $\mathfrak{q}$ é $\mathfrak{q}(\boldsymbol{\eta}) = \mathfrak{d}$, no cal se debe anular a derivada de $\mathfrak{q}$ por ser un extremo, logo

$$B\boldsymbol{\eta} = \mathfrak{d} W \boldsymbol{\eta}$$

e posto que supuxemos W invertible, tense que

$$W^{-1}B\boldsymbol{\eta} = \mathfrak{d}\boldsymbol{\eta}.$$

Con isto, chegamos a que $\mathfrak{d}$ é un autovalor de $W^{-1}B$, o maior deles, e además $\boldsymbol{\eta}$ é o autovector asociado a $\mathfrak{d}$. □

Definición 2.3. Para $\nu \leq \kappa$, sexan $\mathcal{C}_\nu$ clases caracterizadas por $(\boldsymbol{\mu}_\nu, \Sigma_\nu)$ e $\mathbf{X}$ un vector aleatorio dunha delas. Definimos B e W como no teorema anterior e supoñemos W invertible.

Sexa $\boldsymbol{\eta}$ o autovector de $W^{-1}B$ asociado ao autovalor $\mathfrak{d}$, chamámoslle a $\boldsymbol{\eta}$ **dirección discriminante** .

A **regra discriminante (linear) de Fisher** $\mathfrak{r}_F$ defínese como:

$$\mathfrak{r}_F(\mathbf{X}) = l \quad \text{se} \quad |\boldsymbol{\eta}^\intercal \mathbf{X} - \boldsymbol{\eta}^\intercal \boldsymbol{\mu}_l| < |\boldsymbol{\eta}^\intercal \mathbf{X} - \boldsymbol{\eta}^\intercal \boldsymbol{\mu}_\nu| \quad \forall \nu \neq l.$$

Así, a regra de Fisher asigna a $\mathbf{X}$ o número l se $\boldsymbol{\eta}^\intercal\boldsymbol{\mu}_l$ é a media escalar máis cercana ao escalar $\boldsymbol{\eta}^\intercal\mathbf{X}$. Empregamos as cantidades escalares por simplicidade, en vez de buscar a media $\boldsymbol{\mu}_l$ que se achegue máis a $\mathbf{X}$. Ademáis, co uso de $\boldsymbol{\eta}$ destámoslle dando máis peso a variables importantes de $\mathbf{X}$, e reducimos o efecto daquelas variables que non contribúen moito a $W^{-1}B$.

Para o caso de dúas clases $\mathcal{C}_1$ e $\mathcal{C}_2$, podemos deducir unha función de decisión h para a regra linear de Fisher $\mathfrak{r}_F$.

$$|\boldsymbol{\eta}^\intercal\mathbf{X} - \boldsymbol{\eta}^\intercal\boldsymbol{\mu}_1| < |\boldsymbol{\eta}^\intercal\mathbf{X} - \boldsymbol{\eta}^\intercal\boldsymbol{\mu}_2|$$

$$\iff (\boldsymbol{\eta}^\intercal\mathbf{X} - \boldsymbol{\eta}^\intercal\boldsymbol{\mu}_1)^\intercal(\boldsymbol{\eta}^\intercal\mathbf{X} - \boldsymbol{\eta}^\intercal\boldsymbol{\mu}_1) < (\boldsymbol{\eta}^\intercal\mathbf{X} - \boldsymbol{\eta}^\intercal\boldsymbol{\mu}_2)^\intercal(\boldsymbol{\eta}^\intercal\mathbf{X} - \boldsymbol{\eta}^\intercal\boldsymbol{\mu}_2)$$

$$\iff \mathbf{X}^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal\mathbf{X} - 2\mathbf{X}^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal\boldsymbol{\mu}_1 < \mathbf{X}^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal\mathbf{X} - 2\mathbf{X}^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal\boldsymbol{\mu}_2$$

$$\iff 2\mathbf{X}^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \boldsymbol{\mu}_1^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal\boldsymbol{\mu}_2 = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\intercal\boldsymbol{\eta}\boldsymbol{\eta}^\intercal(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$\iff \left[\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]^\intercal \boldsymbol{\eta}\boldsymbol{\eta}^\intercal(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0$$

Como supoñemos $\boldsymbol{\eta}^\intercal(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0$ e non depende de $\mathbf{X}$, podemos definir a función de decisión h como segue:

$$h(\mathbf{X}) = \left[\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]^\intercal \boldsymbol{\eta}.$$

Obtendo así que $h(\mathbf{X}) > 0$ se $\mathbf{X} \in \mathcal{C}_1$ e $h(\mathbf{X}) < 0$ cando $\mathbf{X} \in \mathcal{C}_2$.

Cando non se coñecen os parámetros poboacionais, empregando os datos modificaremos a regra discriminante linear de Fisher presentada para a poboación cambiando o vector de medias $\boldsymbol{\mu}_\nu$ e a matriz de covarianzas Σ_ν polas súas cantidades mostrais. Para $\nu \leq \kappa$, a **media mostral da clase ν-ésima** é

$$\overline{\mathbf{X}}_\nu = \frac{1}{n_\nu}\sum_{i=1}^{n_\nu} \mathbf{X}_i^{[\nu]}$$

e a media das medias mostrais das clases é

$$\overline{\overline{\mathbf{X}}} = \frac{1}{\kappa}\sum_{\nu=1}^{\kappa} \overline{\mathbf{X}}_\nu.$$

Nótese que $\overline{\overline{\mathbf{X}}}$ non fai unha media ponderada das medias de cada clase dependendo do número de observacións dentro delas. Cando tódalas clases teñen o mesmo número de observacións, $\overline{\overline{\mathbf{X}}}$ será a media mostral habitual, resultante de considerar tódolos datos xuntos. Na definición da regra, cambiaremos $\bar{\boldsymbol{\mu}}$ por $\overline{\overline{\mathbf{X}}}$ e cada $\boldsymbol{\mu}_\nu$ por $\overline{\mathbf{X}}_\nu$.

A matriz de covarianzas mostral, empregada en lugar de Σ_ν para cada clase vén dada por

$$S_\nu = \frac{1}{n_\nu - 1} \sum_{i=1}^{n_\nu} (\mathbf{X}_i^{[\nu]} - \overline{\mathbf{X}}_\nu)(\mathbf{X}_i^{[\nu]} - \overline{\mathbf{X}}_\nu)^\intercal.$$

En xeral, as regras de Fisher para a poboación e para a mostra non teñen por qué coincidir. Mostraremos no seguinte exemplo, con datos simulados que proveñen de distribucións normais, como estas regras lineais poden levarnos a resultados distintos.

Exemplo 2.4. Consideramos un problema de dúas clases cuns datos simulados bidimensionais. Collemos as medias e a matriz de covarianzas seguintes

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{e} \quad \Sigma = \begin{bmatrix} 0{,}5 & 0 \\ 0 & 0{,}5 \end{bmatrix}.$$

Simularemos unha mostra de tamaño $n = 50$, con 20 observacións da primeira clase e 30 da segunda, ambas gaussianas cos parámetros elixidos.

Construimos a regra de Fisher empregando os parámetros poboacionais a partir da función de decisión $h(\mathbf{X}) = \left[\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]^\intercal \boldsymbol{\eta}$, del tal modo que $\mathfrak{r}_F(\mathbf{X}) = 1$ se $h(\mathbf{X}) > 0$. Ademáis, facendo isto obteremos a fronteira de decisión dada polos $\mathbf{X}$ tales que $h(\mathbf{X}) = 0$.

O autovector que precisamos é $\boldsymbol{\eta} = \begin{bmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}$ e como un múltiplo de h por un número positivo segue sendo función de decisión para a mesma regra, podemos calcular

$$h(\mathbf{X}) = \left\{ \mathbf{X} - \frac{1}{2} \left(\begin{bmatrix} 0 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \right\}^\intercal \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$= \left\{ \mathbf{X} - \begin{bmatrix} 1/2 \\ 3/2 \end{bmatrix} \right\}^\intercal \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} X_1 - \frac{1}{2} & X_2 - \frac{3}{2} \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$= \frac{1}{2} - X_1 + X_2 - \frac{3}{2} = X_2 - X_1 - 1$$

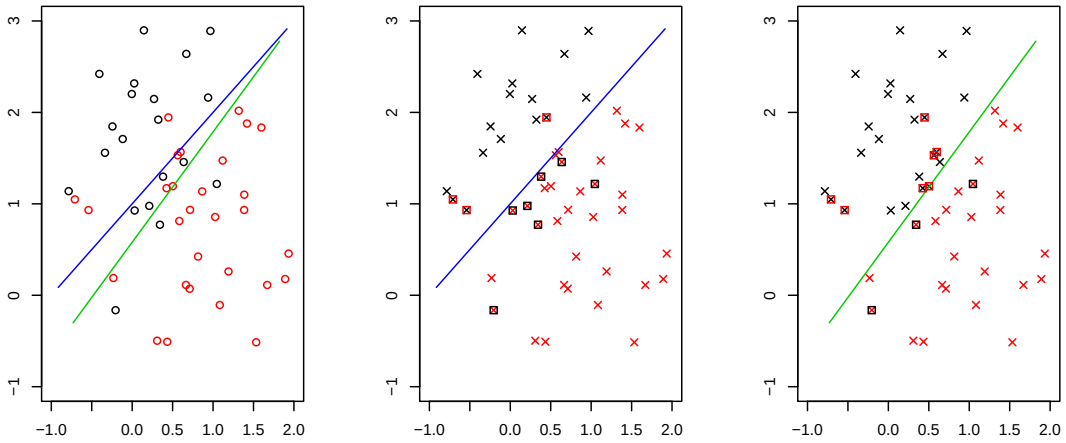
.

A fronteira virá dada entón polos vectores $\left\{ \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} : X_2 - X_1 = 1 \right\}$.

Por outro lado, calcularemos a regra no caso mostral, estimando o vector de medias por $\overline{\mathbf{X}}_\nu$ e a matriz de covarianzas por S_ν para $\nu = 1, 2$, que nos dan

$$\overline{\mathbf{X}}_1 = \begin{bmatrix} 0{,}1945 \\ 1{,}7168 \end{bmatrix}, \quad \overline{\mathbf{X}}_2 = \begin{bmatrix} 0{,}8979 \\ 0{,}7606 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 0{,}2382 & 0{,}0928 \\ 0{,}0928 & 0{,}6109 \end{bmatrix} \quad \text{e} \quad S_2 = \begin{bmatrix} 0{,}4808 & -0{,}0263 \\ -0{,}0263 & 0{,}5579 \end{bmatrix}.$$

Figura 2.1: Datos simulados de dúas clases normais e as fronteiras de decisión para a regra de Fisher poblacional e mostral



Observemos a Figura 2.1. Na gráfica da esquerda móstranse os datos da primera clase en negro e os da segunda en vermello. A liña azul que cruza o gráfico é a fronteira entre as rexións discriminantes, para a regra de Fisher empregando os datos poboacionais, mentres que a verde é a fronteira asociada á regra con datos calculados a partir da mostra simulada. Estas dúas fronteiras non coinciden, co cal tampouco han coincidir as regras que definen.

Nas gráficas do centro e da dereita móstranse con cruces os datos coloreados segundo a clase á que foron asignados, pola regra de Fisher para a poboación e para a mostra respectivamente, e aqueles que foron clasificados erróneamente están rodeados na cor da clase á que realmente pertencen. Vemos entón que ademáis de que as funcións de decisión para estas dúas regras non coinciden, tampouco clasifican os datos do mesmo xeito.

Podemos ademáis contabilizar os erros cometidos cunha das medidas introducidas para a avaliación das regras. O erro de clasificación é dun 20 % para as dúas regras xa que se clasificaron mal 10 observacións, ainda que as imaxes mostran como a clasificación que efectúan é distinta. A primeira regra clasifica erróneamente 3 observacións da primeira clase e 7 da segunda, mentres que a regra mostral falla en 7 datos da clase 1 e en 3 da clase 2. Hai 3 datos de cada clase que clasifican mal ambas regras pois están na rexión discriminante contraria, o cal se debe a que hai certa superposición das clases. O resto de datos mal clasificados están moi preto das rexións de decisión, de feito, entre as dúas rectas, co que cada regra os asigna a un grupo distinto.

2.2. Clasificación baixo hipótese de normalidade

Canto maior coñecemento teñamos dos datos, mellores decisións poderemos tomar. Se sabemos que os datos pertencen a unha poboación normal, deberíamos incorporar este coñecemento no noso proceso de decisión.

2.2.1. Dúas clases normais univariantes coa mesma varianza

Comezamos co caso máis sinxelo, cando sabemos que as nosas observacións unidimensionais proveñen de dúas clases de distribucións normais coa mesma covarianza pero distintas medias. É dicir, $\mathcal{C}_1 = \mathcal{N}(\mu_1, \sigma^2)$ e $\mathcal{C}_2 = \mathcal{N}(\mu_2, \sigma^2)$, e supoñemos que $\mu_1 > \mu_2$.

Para construir a nosa regra, basearémonos en que se a observación X pertence a unha das clases, asignarémola á primeira clase cando sexa máis verosímil que proveña de $\mathcal{C}_1$. Deste xeito, a regra que definimos será a regra de Bayes, que minimiza a probabilidade de erro de clasificación. Para isto, empregaremos a función de verosimilitude dunha distribución normal $\mathcal{N}(\mu, \sigma^2)$, que non é máis que a súa función de densidade f, dada por

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}\right],$$

que depende só do parámetro μ, pois a varianza non cambia entre as clases. A regra discriminante definirase entón como

$$\mathfrak{r}(X) = 1 \quad \text{se} \quad f_1(X) > f_2(X).$$

Para chegar a unha expresión explícita para esta regra, desenvolvemos unha serie de desigualdades equivalentes partindo da fórmula da verosimilitude para unha observación X.

$$\begin{aligned}
& f_1(X) > f_2(X) \\
\Leftrightarrow\ & \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(X-\mu_1)^2}{\sigma^2}\right] > \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(X-\mu_2)^2}{\sigma^2}\right] \\
\Leftrightarrow\ & \exp\left[-\frac{1}{2}\frac{(X-\mu_1)^2}{\sigma^2}\right] > \exp\left[-\frac{1}{2}\frac{(X-\mu_2)^2}{\sigma^2}\right] \\
\Leftrightarrow\ & -\frac{1}{2\sigma^2}(X-\mu_1)^2 > -\frac{1}{2\sigma^2}(X-\mu_2)^2 \\
\Leftrightarrow\ & (X-\mu_1)^2 < (X-\mu_2)^2 \\
\Leftrightarrow\ & X^2 - 2X\mu_1 + \mu_1^2 < X^2 - 2X\mu_2 + \mu_2^2 \\
\Leftrightarrow\ & 2X(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2 = (\mu_1 + \mu_2)(\mu_1 - \mu_2) \\
\Leftrightarrow\ & 2X > \mu_1 + \mu_2
\end{aligned}$$

$$\Leftrightarrow X > \frac{\mu_1 + \mu_2}{2}.$$

Nas últimas liñas, usamos a hipótese de que $\mu_1 > \mu_2$. O resultado final ten sentido, pois se a media da primeira clase é maior que a da segunda, cando a nosa observación excede a media de μ_1 e μ_2 será que está máis cerca de μ_1 e será máis probable que pertenza a $\mathcal{C}_1$.

Esta regra pode estenderse de maneira fácil e natural ao caso d-dimensional, considerando a función de verosimilitude para unha distribución normal multivariante.

2.2.2. Dúas ou máis clases normais multivariantes coa mesma matriz de covarianzas

Se é sabido que os vectores aleatorios nun problema con dúas clases veñen dunha distribución normal que comparte a matriz de covarianzas, podemos presentar o seguinte Teorema:

Teorema 2.5. *Sexa $\boldsymbol{X}$ un vector aleatorio gaussiano que pertence a unha das clases $\mathcal{C}_\nu = \mathcal{N}(\boldsymbol{\mu}_\nu, \Sigma)$, con $\nu = 1, 2$ e supoñamos que $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. A función de densidade da clase ν-ésima vén dada por*

$$f(\boldsymbol{X}) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_\nu)^\intercal \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_\nu)\right],$$

onde $\det(\Sigma)$ é o determinante da matriz de covarianzas común. Sexa $\mathfrak{r}_{norm}$ a regra que asigna $\boldsymbol{X}$ á clase $\mathcal{C}_1$ se $f_1(\boldsymbol{X}) > f_2(\boldsymbol{X})$. A función h definida por

$$h(\boldsymbol{X}) = \left[\boldsymbol{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]^\intercal \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

será entón unha función de decisión para a regra $\mathfrak{r}_{norm}$, e $h(\boldsymbol{X}) > 0$ se e só se $f_1(\boldsymbol{X}) > f_2(\boldsymbol{X})$.

Chamamos a $\mathfrak{r}_{norm}$ a **regra discriminante (linear) normal**, basada en clases normais coa mesma matriz de covarianzas. Normalmente, regra e función de decisión considéranse unha mesma cousa.

Demostración. Cunha serie de desigualdades equivalentes chegaremos á expresión de h, partindo de que $\mathfrak{r}_{norm}(\mathbf{X}) = 1 \quad \text{se} \quad f_1(\mathbf{X}) > f_2(\mathbf{X})$. Empregaremos tamén que Σ é unha matriz simétrica.

$$f_1(\mathbf{X}) > f_2(\mathbf{X})$$
$$\Leftrightarrow (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_1)^\intercal \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}_1)\right] >$$
$$> (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_2)^\intercal \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}_2)\right]$$

$$\Leftrightarrow \exp\left[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_1)^\intercal\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_1)\right] > \exp\left[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_2)^\intercal\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_2)\right]$$

$$\Leftrightarrow -\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_1)^\intercal\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_1) > -\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_2)^\intercal\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_2)$$

$$\Leftrightarrow (\mathbf{X}-\boldsymbol{\mu}_1)^\intercal\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_1) < (\mathbf{X}-\boldsymbol{\mu}_2)^\intercal\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_2)$$

$$\Leftrightarrow \mathbf{X}^\intercal\Sigma^{-1}\mathbf{X}-\boldsymbol{\mu}_1^\intercal\Sigma^{-1}\mathbf{X}-\mathbf{X}^\intercal\Sigma^{-1}\boldsymbol{\mu}_1+\boldsymbol{\mu}_1^\intercal\Sigma^{-1}\boldsymbol{\mu}_1 < \mathbf{X}^\intercal\Sigma^{-1}\mathbf{X}-\boldsymbol{\mu}_1^\intercal\Sigma^{-1}\mathbf{X}-\mathbf{X}^\intercal\Sigma^{-1}\boldsymbol{\mu}_2+\boldsymbol{\mu}_2^\intercal\Sigma^{-1}\boldsymbol{\mu}_2$$

$$\Leftrightarrow -2\mathbf{X}^\intercal\Sigma^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^\intercal\Sigma^{-1}\boldsymbol{\mu}_1 < -2\mathbf{X}^\intercal\Sigma^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^\intercal\Sigma^{-1}\boldsymbol{\mu}_2$$

$$\Leftrightarrow 2\mathbf{X}^\intercal\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \boldsymbol{\mu}_2^\intercal\Sigma^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 > 0$$

$$\Leftrightarrow 2\mathbf{X}^\intercal\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \boldsymbol{\mu}_1^\intercal\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \boldsymbol{\mu}_1^\intercal\Sigma^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^\intercal\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \boldsymbol{\mu}_2^\intercal\Sigma^{-1}\boldsymbol{\mu}_1 > 0$$

$$\Leftrightarrow 2\mathbf{X}^\intercal\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) > 0$$

$$\Leftrightarrow \left[\mathbf{X}-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right]^\intercal\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) > 0.$$

Esta última desigualdade coincide con $h(\mathbf{X}) > 0$. □

Para definir a versión mostral desta regra, simplemente temos que cambiar as cantidades poboacionais polas correspondentes mostrais obtidas a partir de $\mathbb{X}$ na fórmula dada pola función de decisión. Sexa $\mathbf{X}_i$ un dos datos de $\mathbb{X}$,

$$\mathfrak{r}_{norm}(\mathbf{X}) = \begin{cases} 1 & \text{se} \quad \left[\mathbf{X}_i - \frac{1}{2}(\overline{\mathbf{X}}_1+\overline{\mathbf{X}}_2)\right]^\intercal S^{-1}(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2) > 0 \\ 2 & \text{noutro caso.} \end{cases}$$

Mentres que se temos unha nova observación $\mathbf{X}_{new}$, tomaremos $\mathfrak{r}_{norm}(\mathbf{X}_{new}) = 1$ se

$$h(\mathbf{X}_{new}) = \left[\mathbf{X}_{new} - \frac{1}{2}(\overline{\mathbf{X}}_1+\overline{\mathbf{X}}_2)\right]^\intercal S^{-1}(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2) > 0,$$

onde S e $\overline{\mathbf{X}}_\nu$, $\nu = 1, 2$ se calculan a partir de $\mathbb{X}$ sen ter en conta $\mathbf{X}_{new}$.

Observación 2.6. Algunhas veces refírese á regra discriminante linear normal como regra de Fisher, aínda que, como en Koch (2014), as definimos de maneiras distintas. A primera baséase na función de verosimilitude mentres que a segunda emprega autovectores dunha matriz. As formas das funcións de decisión asociadas son moi parecidas, e esta comparación motiva unha clase máis xeral de funcións de decisión lineais para problemas con dúas clases. Consideramos

$$h_\beta(\mathbf{X}) = \left[\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right]^\intercal\boldsymbol{\beta},$$

onde $\boldsymbol{\beta}$ é un vector adecuado, á nosa elección. Serían $\boldsymbol{\beta} = \boldsymbol{\eta}$ para a regra de Fisher e $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ para a regra linear normal.

Definindo W a matriz de variabilidade within-class e usando o Teorema 2.2 temos que

$$W = \sum_{\nu=1}^{2} \Sigma_\nu = 2\Sigma,$$

$$B = \sum_{\nu=1}^{2} (\boldsymbol{\mu}_\nu - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})(\boldsymbol{\mu}_\nu - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})^\intercal = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal$$

e $\boldsymbol{\eta} = \frac{W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\|W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|}$ é o autovector asociado ao máximo autovalor de $W^{-1}B$.

Hai certas situacións nas que non se cumplen as hipóteses do Teorema 2.5 e a regra linear normal non é a adecuada. Isto ocorre cando sabemos que a distribución dos vectores non é normal, cando as matrices de covarianzas das clases non coinciden ou cando as descoñecemos.

No caso de descoñecer Σ_ν pero poder asegurar que son a mesma para ambas clases, temos a opción de empregar a **matriz de covarianzas mostral agrupada** (pooled)

$$S_{pool} = \sum_{\nu=1}^{2} \frac{n_\nu - 1}{n - 2} S_\nu,$$

sendo S_ν a matriz de covarianzas mostral de $\mathcal{C}_\nu$ e n_ν o número de observacións desta clase. Antes de facer uso da matriz agrupada, teremos que asegurarnos de que é un procedemento apropiado.

A regra linear normal pode non comportarse tan ben para vectores aleatorios con distribucións non coñecidas ou non normais. Aínda así, cando non se desvían "demasiado" da distribución gaussiana, podemos obter bos resultados con $\mathfrak{r}_{norm}$. Determinar canto é ese "demasiado" é difícil e para datos de dimensión moderada poderemos levar a cabo tests de normalidade ou tirar de axudas visuais. De todos modos, o proceder correcto é aplicar varias regras e avaliar o seu comportamento.

Todo o introducido neste apartado pode estenderse a máis de dúas clases, baseándose igualmente na verosimilitude.

Supoñamos que un vector aleatorio $\mathbf{X}$ pertence a unha das clases $\mathcal{C}_\nu = \mathcal{N}(\boldsymbol{\mu}_\nu, \Sigma)$, con $\nu \leq \kappa$ só con medias distintas. Trataremos de atopar o parámetro $\boldsymbol{\mu}_k$ que maximice a verosimilitude para asignar $\mathbf{X}$ a $\mathcal{C}_k$. Definimos as **funcións de decisión preferenciais**

$$h_{(l,\nu)}(\mathbf{X}) = \left[\mathbf{X} - \frac{\boldsymbol{\mu}_l + \boldsymbol{\mu}_\nu}{2}\right]^\intercal \Sigma^{-1}(\boldsymbol{\mu}_l - \boldsymbol{\mu}_\nu) \quad \text{para} \quad l, \nu = 1, 2, \ldots, \kappa \quad \text{e} \quad l \neq \nu.$$

Nótese ademáis que $h_{(l,\nu)} = -h_{(\nu,l)}$. Poñamos entón

$$h_{norm}(\mathbf{X}) = \max_{(l,\nu)=1,2,\ldots,\kappa} h_{(l,\nu)}(\mathbf{X}) \quad \text{e} \quad \mathfrak{r}_{norm}(\mathbf{X}) = k,$$

sendo k o primeiro índice do par (l, ν) no que se alcanza o máximo $h_{norm}(\mathbf{X})$.

Alternativamente, podemos considerar as densidades $f_\nu(\mathbf{X})$ para $\nu \leq \kappa$, e definir a regra

$$\tilde{\mathfrak{r}}_{norm}(\mathbf{X}) = k \quad \text{se} \quad f_k(\mathbf{X}) = \max_{1\leq\nu\leq\kappa} f_\nu(\mathbf{X}).$$

Para a poboación, as dúas regras asignarán $\mathbf{X}$ á mesma clase. Ao levalas ao caso da mostra pode que den lugar a dúas regras distintas, dependendo da variabilidade dentro das clases e a elección do estimador S para a matriz de covarianzas común Σ.

2.2.3. Regras discriminantes cuadráticas

Ata agora consideramos únicamente regras lineais da forma $\boldsymbol{h(X) = a^\intercal X + c}$, para un vector a e un escalar c que non dependen do vector aleatorio $\mathbf{X}$.

Na regra linear normal a linealidade é unha consecuencia de que a mesma matriz de covarianzas se usa para tódalas clases. Mentres que, se sabemos que ditas matrices son diferentes (tanto que non nos serve agrupalas en S_{pool}) teremos que considerar matrices distintas. Isto levaranos a unha función de decisión non linear e, consecuentemente, a unha regra non linear.

Consideremos entón un vector aleatorio $\mathbf{X}$ que pertence a unha das clases $\mathcal{C}_\nu = \mathcal{N}(\boldsymbol{\mu}_\nu, \Sigma_\nu)$, para $\nu \leq \kappa$, que se diferencian tanto no vector de medias como na matriz de covarianzas. Partimos outra vez da regra de Bayes e a densidade f para definir a regra, pero neste caso non só dependerá do parámetro $\boldsymbol{\mu}_\nu$, senón que nos interesa $\theta_\nu = (\boldsymbol{\mu}_\nu, \Sigma_\nu)$. Asignaremos $\mathbf{X}$ á clase $\mathcal{C}_l$ se

$$f_l(\mathbf{X}) > f_\nu(\mathbf{X}) \quad \text{para} \quad l \neq \nu.$$

Teorema 2.7. *Para $\nu \leq \kappa$ e un vector aleatorio $\boldsymbol{X}$ pertencente a unha das clases $\mathcal{C}_\nu = \mathcal{N}(\boldsymbol{\mu}_\nu, \Sigma_\nu)$. A regra discriminante $\mathfrak{r}_{quad}$ baseada nas funcións de densidade das κ clases asigna $\boldsymbol{X}$ a $\mathcal{C}_l$ se*

$$\|\boldsymbol{X}_{\Sigma_l}\|^2 + \log[\det(\Sigma_l)] = \min_{1\leq\nu\leq\kappa} \left\{\|\boldsymbol{X}_{\Sigma_\nu}\|^2 + \log[\det(\Sigma_\nu)]\right\},$$

onde $\boldsymbol{X}_\Sigma = \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ é o vector $\boldsymbol{X}$ estandarizado, é dicir, $\boldsymbol{X}_\Sigma \sim \mathcal{N}_d(0, I)$.

Chamaremos a $\mathfrak{r}_{quad}$ **regra discriminante cuadrática (normal)** porque é cuadrática en $\mathbf{X}$. Posto que derivamos esta regra de asumir unha distribución normal, non podemos esperar que o seu comportamento para vectores non gaussianos sexa o óptimo.

Demostración. Obtemos coa definición da función de verosimilitude da normal multivariante que

$$\log f(\mathbf{X}) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\left[\log[\det(\Sigma)] + (\mathbf{X} - \boldsymbol{\mu})^\intercal \Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu})\right]$$

$$= -\frac{1}{2}\left[\|\mathbf{X}_\Sigma\|^2 + \log[\det(\Sigma)]\right] + c,$$

sendo $c = -\frac{d}{2}\log(2\pi)$ independente do parámetro $\theta = (\boldsymbol{\mu}, \Sigma)$. Asignaremos $\mathbf{X}$ preferentemente a $\mathcal{C}_l$ antes que a $\mathcal{C}_\nu$ cando

$$f_l(\mathbf{X}) > f_\nu(\mathbf{X})$$
$$\Leftrightarrow \log f_l(\mathbf{X}) > \log f_\nu(\mathbf{X})$$
$$\Leftrightarrow -\frac{1}{2}\left[\|\mathbf{X}_{\Sigma_l}\|^2 + \log[\det(\Sigma_l)]\right] + c > -\frac{1}{2}\left[\|\mathbf{X}_{\Sigma_\nu}\|^2 + \log[\det(\Sigma_\nu)]\right] + c$$
$$\Leftrightarrow -\frac{1}{2}\left[\|\mathbf{X}_{\Sigma_l}\|^2 + \log[\det(\Sigma_l)]\right] > -\frac{1}{2}\left[\|\mathbf{X}_{\Sigma_\nu}\|^2 + \log[\det(\Sigma_\nu)]\right]$$
$$\Leftrightarrow \|\mathbf{X}_{\Sigma_l}\|^2 + \log[\det(\Sigma_l)] < \|\mathbf{X}_{\Sigma_\nu}\|^2 + \log[\det(\Sigma_\nu)]$$

E o resultado séguese desta desigualdade, tendo en conta tódalas clases e non só un par. □

Para a situación de dúas clases pódese obter unha expresión explícita máis sinxela para a regra $\mathfrak{r}_{quad}$.

Corolario 2.8. *Na mesma situación que a do Teorema 2.7 con $\kappa = 2$, supoñemos a maiores que ambas matrices de covarianzas son de rango r. Sexa $\Sigma_\nu = \Gamma_\nu \Lambda_\nu \Gamma_\nu^\intercal$ a descomposición espectral de Σ_ν, con $\lambda_{\nu,j}$ o autovalor j-ésimo de Λ_ν.*

A regra cuadrática $\mathfrak{r}_{quad}$ asigna $\boldsymbol{X}$ á clase $\mathcal{C}_1$ cando

$$\|\Lambda_1^{-1/2}\Gamma_1^\intercal(\boldsymbol{X} - \boldsymbol{\mu}_1)\|^2 - \|\Lambda_2^{-1/2}\Gamma_2^\intercal(\boldsymbol{X} - \boldsymbol{\mu}_2)\|^2 + \sum_{j=1}^{r} \log\frac{\lambda_{1,j}}{\lambda_{2,j}} < 0.$$

Nótese que se o rango r da matriz é estrictamente menor que a súa dimensión d, entón Σ ten a descomposición espectral $\Sigma = \Gamma_r \Lambda_r \Gamma_r^\intercal$, onde Γ_r é de tamaño $d \times r$ e Λ_r unha matriz diagonal de tamaño $r \times r$. Ademáis, Γ_r non é ortogonal, e tense unha relación máis débil, $\Gamma_r^\intercal \Gamma_r = I_{r\times r}$, pero $\Gamma_r \Gamma_r^\intercal \neq I_{d\times d}$. Ás matrices que cumplen isto chámaselles **r-ortogonais**.

Tense así, para $k, m \in \mathbb{Z}$

$$\Sigma^{k/m} = \Gamma \Lambda^{k/m} \Gamma^\intercal,$$

e en particular, $\Sigma^{-1/2} = \Gamma \Lambda^{-1/2} \Gamma^\intercal$.

Demostración. Probamos o resultado para o caso en que $r = d$, i.e., Σ_ν son invertibles. Partimos dos cálculos feitos para a demostración do Teorema 2.7, pois $\mathfrak{r}_{quad}(\mathbf{X}) = 1$ se

$$\|\mathbf{X}_{\Sigma_1}\|^2 + \log[\det(\Sigma_1)] < \|\mathbf{X}_{\Sigma_2}\|^2 + \log[\det(\Sigma_2)]$$

Tense que

$$\|\mathbf{X}_\Sigma\|^2 = \mathbf{X}_\Sigma^\intercal \mathbf{X}_\Sigma = [\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})]^\intercal [\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})] =$$

$$[\Gamma\Lambda^{-1/2}\Gamma^\intercal(\mathbf{X}-\boldsymbol{\mu})]^\intercal\Gamma\Lambda^{-1/2}\Gamma^\intercal(\mathbf{X}-\boldsymbol{\mu}) = (\mathbf{X}-\boldsymbol{\mu})^\intercal\Gamma\Lambda^{-1/2}\Gamma^\intercal\Gamma\Lambda^{-1/2}\Gamma^\intercal(\mathbf{X}-\boldsymbol{\mu})$$

$$= (\mathbf{X}-\boldsymbol{\mu})^\intercal\Gamma\Lambda^{-1/2}\Lambda^{-1/2}\Gamma^\intercal(\mathbf{X}-\boldsymbol{\mu}) = [\Lambda^{-1/2}\Gamma^\intercal(\mathbf{X}-\boldsymbol{\mu})]^\intercal[\Lambda^{-1/2}\Gamma^\intercal(\mathbf{X}-\boldsymbol{\mu})] = \|\Lambda^{-1/2}\Gamma^\intercal(\mathbf{X}-\boldsymbol{\mu})\|^2,$$

e posto que o determinante dunha matriz é o produto dos seus autovalores, temos que

$$\det\Sigma = \prod_{j=1}^{d}\lambda_j.$$

Gracias a estos dous cálculos, desenvolvemos a desigualdade

$$\|\mathbf{X}_{\Sigma_1}\|^2 + \log[\det(\Sigma_1)] < \|\mathbf{X}_{\Sigma_2}\|^2 + \log[\det(\Sigma_2)]$$

$$\Leftrightarrow \|\mathbf{X}_{\Sigma_1}\|^2 - \|\mathbf{X}_{\Sigma_2}\|^2 + \log[\det(\Sigma_1)] - \log[\det(\Sigma_2)] < 0$$

$$\Leftrightarrow \|\Lambda_1^{-1/2}\Gamma_1^\intercal(\mathbf{X}-\boldsymbol{\mu}_1)\|^2 - \|\Lambda_2^{-1/2}\Gamma_2^\intercal(\mathbf{X}-\boldsymbol{\mu}_2)\|^2 + \log\left[\prod_{j=1}^{d}\lambda_{1,j}\right] - \log\left[\prod_{j=1}^{d}\lambda_{2,j}\right] < 0$$

$$\Leftrightarrow \|\Lambda_1^{-1/2}\Gamma_1^\intercal(\mathbf{X}-\boldsymbol{\mu}_1)\|^2 - \|\Lambda_2^{-1/2}\Gamma_2^\intercal(\mathbf{X}-\boldsymbol{\mu}_2)\|^2 + \sum_{j=1}^{d}\log\lambda_{1,j} - \sum_{j=1}^{d}\log\lambda_{2,j} < 0$$

$$\Leftrightarrow \|\Lambda_1^{-1/2}\Gamma_1^\intercal(\mathbf{X}-\boldsymbol{\mu}_1)\|^2 - \|\Lambda_2^{-1/2}\Gamma_2^\intercal(\mathbf{X}-\boldsymbol{\mu}_2)\|^2 + \sum_{j=1}^{d}[\log\lambda_{1,j} - \log\lambda_{2,j}] < 0$$

$$\Leftrightarrow \|\Lambda_1^{-1/2}\Gamma_1^\intercal(\mathbf{X}-\boldsymbol{\mu}_1)\|^2 - \|\Lambda_2^{-1/2}\Gamma_2^\intercal(\mathbf{X}-\boldsymbol{\mu}_2)\|^2 + \sum_{j=1}^{r}\log\frac{\lambda_{1,j}}{\lambda_{2,j}} < 0.$$

□

Exemplo 2.9. Neste exemplo imos considerar simulacións por pares. En cada unha delas, haberá dúas clases de datos tridimensionais con vectores de medias distintos. En cada par, unha simulación terá matrices de covarianzas iguais e na outra serán distintas. Deste xeito poderemos ver se supón unha mellora empregar a regra cuadrática.

Consideramos os vectores e matrices seguintes.

$$\boldsymbol{\mu}_1 = \begin{bmatrix}0\\0\\0\end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix}2\\1\\-0{,}5\end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix}1 & 0 & 0\\0 & 0{,}25 & 0\\0 & 0 & 0{,}25\end{bmatrix} \quad \text{e} \quad \Sigma_2 = \begin{bmatrix}1 & 0{,}25 & 0{,}125\\0{,}25 & 0{,}5 & 0\\0{,}125 & 0 & 0{,}25\end{bmatrix}.$$

En primeiro lugar, simulamos datos de dúas clases normais. Na primeira gráfica da Figura 2.2 aparecen 200 observacións da clase $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ e 300 de $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_1)$, mentres que na segunda a matriz cambia dunha clase a outra, $\mathcal{C}_1 = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ e $\mathcal{C}_2 = \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$. Os puntos da clase $\mathcal{C}_1$ están en negro e os da clase $\mathcal{C}_2$ en vermello. Vemos que na da esquerda as nubes de puntos teñen a mesma forma, pero centradas en distintas medias e na da

dereita, a aparencia rotada da nube de puntos vermella mostra a matriz de covarianzas non diagonal.

Ademáis, simularemos dúas mostras de datos que non son normais, partindo da distribución de Poisson. Para $\mathcal{P}1$, simularemos 500 observacións dun vector aleatorio con media e varianza $\lambda = 10$. Os primeiros 200 datos, trasladámolos polo vector $(5, -5, 2)$ para obter unha clase con media $\boldsymbol{\mu}_1 = (15, 5, 12)$ e os seguintes 300 datos non os modificamos, así $\boldsymbol{\mu}_2 = (10, 10, 10)$. Procedendo do mesmo xeito pero cambiando λ por $\lambda = (10, 20, 30)$ e restándolle $(0, 10, 20)$ aos datos da segunda clase para ter $\mathcal{C}_1$ igual ca antes e $\mathcal{C}_2$ coa mesma media pero $\Sigma_2 = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 30 \end{pmatrix}$, construimos os datos $\mathcal{P}2$.

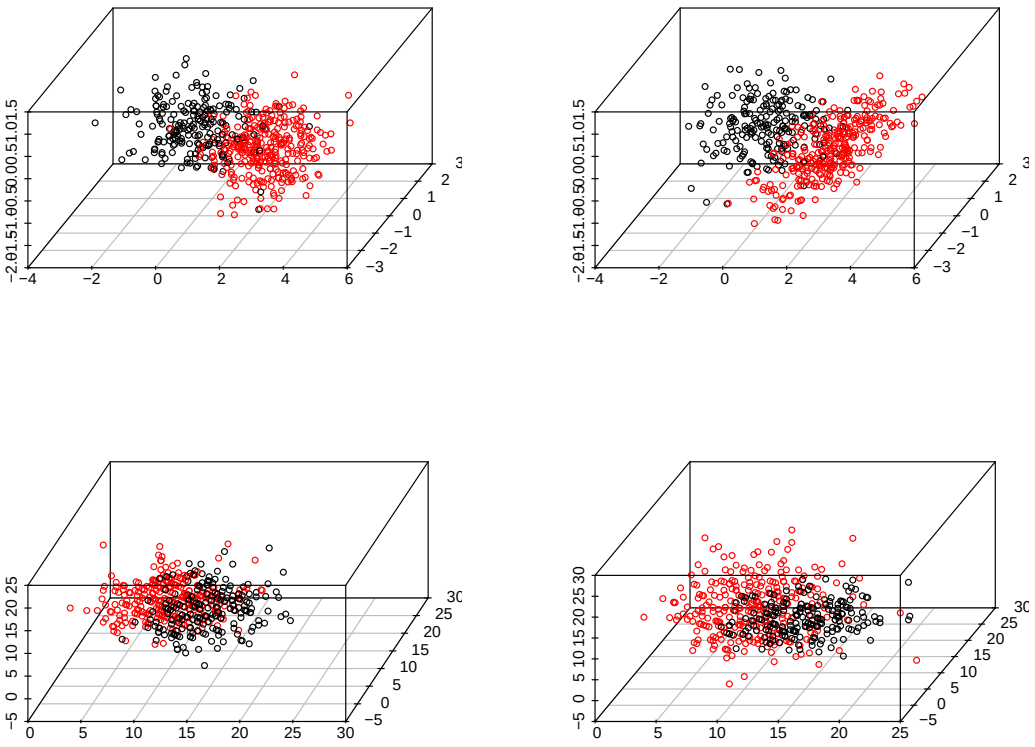
Calculamos a regra de Fisher, a regra normal linear e tamén a regra cuadrática para estos datos simulados, empregando os parámetros mostrais. Para avaliar como se comportan os métodos, repítese a simulación 100 veces, e así obtemos un erro de clasificación medio e tamén a desviación típica desta medida nas 100 repeticións. A Táboa 2.1 recolle os resultados para cada caso e cada regra discriminante.

Cadro 2.1: Erro de clasificación medio e desviación típica (entre parénteses) para as tres regras e os catro conxuntos de datos simulados, en tanto por cento.

	Datos			
	$\mathcal{N}1$	$\mathcal{N}2$	$\mathcal{P}1$	$\mathcal{P}2$
Fisher	12.914	5.698	12.142	14.772
	(1.50)	(0.97)	(1.52)	(1.63)
Normal	12.884	5.674	12.122	14.770
	(1.49)	(0.98)	(1.52)	(1.58)
Cuadrática	12.856	4.634	12.024	13.040
	(1.53)	(1.04)	(1.55)	(1.42)

Este cadro mostra que a regra cuadrática ten un mellor comportamento que as regras lineais, sobre todo nos datos que veñen de clases con matrices de covarianzas distintas. As dúas regras lineais compórtanse de xeito similar, tanto para o caso normal como para o derivado dunha distribución de Poisson. Isto pode deberse a que, para os parámetros λ considerados, a Poisson aseméllase bastante a unha campana de Gauss.

Figura 2.2: Nubes de puntos das catro situacións simuladas.



2.3. Técnicas de clasificación non paramétrica

A diferencia dos clasificadores introducidos no capítulo anterior, os clasificadores non paramétricos non se basean na forma das funcións de decisión ou regras. Estes son útiles cando non coñecemos os datos e non queremos facer suposicións acerca deles, como por exemplo a normalidade. Algúns dos métodos máis coñecidos son o dos k veciños máis cercanos, as regras tipo kernel, as árbores de decisión ou as support vector machines (SVM), entre outros. Neste traballo presentaranse só as dúas primeiras.

As técnicas de clasificación non paramétrica teñen a vantaxe de adaptarse a datos con distribucións moi diversas e a veces poden levar a mellores resultados que aquelas deducidas a partir da distribución dos datos. De todos modos, moitas veces precisaremos máis observacións para chegar a unha función que prediga de maneira satisfactoria a pertenza a unha clase, ademáis de que estos métodos poden ser máis costosos computacionalmente, especialmente ao aumentar a dimensión ou o tamaño da nosa mostra.

En Devroye et al. (1996) faise un estudo en profundidade das regras non paramétricas. Danse cotas para o seu erro independentes da distribución de $\mathbf{X}$, explóranse as súas propiedades asintóticas e a súa posible adaptación a contextos de alta dimensión.

2.3.1. Regra dos k veciños máis cercanos

O primeiro clasificador non paramétrico que presentamos xorde de maneira intuitiva, baseado en propiedades dos veciños. Foi proposto inicialmente en Fix e Hodges (1951).

Partindo dos datos

$$\begin{bmatrix} \mathbb{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_n \\ Y_1 & Y_2 & \cdots & Y_n \end{bmatrix},$$

que pertencen a κ clases distintas. Consideramos Δ unha distancia entre dous vectores, por exemplo a distancia euclídea, así

$$\Delta(\mathbf{X}_i, \mathbf{X}_j) \quad \begin{cases} > 0 & \text{se} \quad i \neq j, \\ = 0 & \text{se} \quad i = j. \end{cases}$$

Fixado un vector aleatorio $\mathbf{X}_0$ coas mesmas propiedades distributivas que as observacións $\mathbb{X} = [\mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_n]$, ordenámolas segundo a crecente distancia ao vector $\mathbf{X}_0$, sendo $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \ldots, \mathbf{X}_{(n)}$ tal que

$$\Delta(\mathbf{X}_{(1)}, \mathbf{X}_0) \leq \Delta(\mathbf{X}_{(2)}, \mathbf{X}_0) \leq \cdots \leq \Delta(\mathbf{X}_{(n)}, \mathbf{X}_0).$$

Se $\mathbf{X}_0$ é un dos datos, a primeira distancia é nula, e empezamos a contar pola primeira distancia distinta de cero.

Definición 2.10. Sexan $\begin{bmatrix} \mathbb{X} \\ \mathbf{Y} \end{bmatrix}$ n vectores aleatorios etiquetados pertencentes a κ clases distintas. Consideramos un vector aleatorio $\mathbf{X}$ da mesma poboación que $\mathbb{X}$ e unha distancia Δ. Ordenamos as observacións segundo a súa distancia a $\mathbf{X}$. Sexa $k \geq 1$ un enteiro fixado.

1. Os k **veciños máis cercanos** respecto a $\mathbf{X}$ é o conxunto

$$N(\mathbf{X}, k) = \{\mathbf{X}_i \in \mathbb{X} : \Delta(\mathbf{X}_i, \mathbf{X}) \leq \Delta(\mathbf{X}_{(k)}, \mathbf{X})\},$$

   formado polos k vectores a menor distancia de $\mathbf{X}$.

2. A **regra dos k veciños máis cercanos** ou **regra kNN** vén dada por

   $\mathfrak{r}_{kNN}(\mathbf{X}) = l$ cando $Y = l$ é a etiqueta que máis se repite entre os vectores da veciñanza $N(\mathbf{X}, k)$.

A regra kNN asigna $\mathbf{X}$ á clase máis común entre os seus puntos veciños, de xeito que se basea únicamente en información local. Cando hai un empate, é dicir, hai dúas clases distintas que aparecen as mesmas veces preto de $\mathbf{X}$, a asignación é arbitraria entre estos dous candidatos. Debido a isto, cando estamos ante un problema con dúas clases, terá sentido escoller un k impar, para evitar que haxa empates. A moda das etiquetas en $N(\mathbf{X}, k)$ dependerá entón tanto da distancia Δ escollida como do número de veciños k que consideremos. O cal nos leva a preguntarnos cal é o mellor k.

A escolla da distancia afectará á forma da veciñanza $N(\mathbf{X}, k)$ pero non é a clave da regra kNN, esta é k. O valor máis simple será $k = 1$, de xeito que asignamos $\mathbf{X}$ á clase do dato máis cercano a él. Existen distribucións para as que $\mathfrak{r}_{1NN}$ se comporta mellor que calquera $\mathfrak{r}_{kNN}$ con $k > 1$, como indican en Devroye et al. (1996). En xeral, tomar valores máis grandes para k levaranos a regras que clasifiquen mellor, con menos erros.

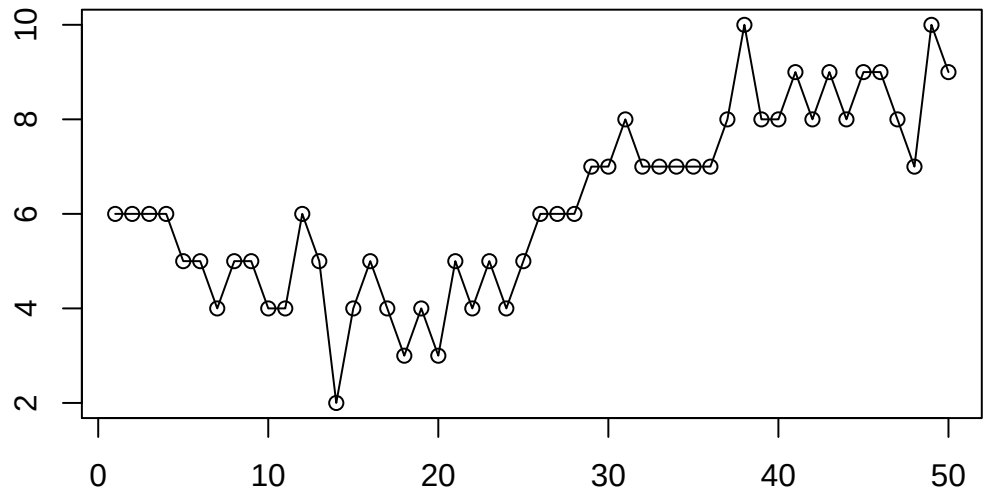
Para escoller o k óptimo, un enfoque práctico sería calcular o erro de clasificación para a regra con distintos valores de k, con $1 \leq k \leq n$, xa que non imos tomar veciñanzas con máis puntos que as observacións da nosa mostra, e decantarnos por aquel que leve ao menor número de observacións mal clasificadas. Empregamos o erro ϵ_{mis}, entrenando a regra para tódolos datos excepto o que quitamos en cada repetición, respecto o cal medimos a distancia.

Este método ten a vantaxa de ser simple e fácilmente interpretable, aínda que pode ser costoso computacionalmente, ao ter que tomar todo par de observacións posible e calcular a distancia entre elas.

Exemplo 2.11. Para ilustrar como facer a búsqueda do k óptimo, consideramos os datos **iris** de Fisher, dispoñibles en R, con catro variables que miden distintas características de pétalos e sépalos en tres especies distintas de flores. Posto que hai 50 observacións de cada clase, consideramos $k \leq 50$. Na Figura 2.3 represéntase o número de erros na clasificación para os distintos valores de k. A tendencia xeral é que os erros decrecen nun primeiro momento ata chegar a un mínimo de 2, para despois aumentar á par que k. Observamos que considerar só o veciño máis cercano non da lugar á mellor regra, xa que hai 6 observacións que asigna a unha clase á que non pertencen.

Así, a regra óptima será a correspondente a $k = 14$. E só falla para os datos 84 e 107, os cales se clasifican mal para case todos os valores de k. Estas observacións pertencen ás clases *versicolor* e *virginica*, cuxas nubes de puntos se superpoñen e están separadas da clase restante *setosa*. Podemos calcular a regra normal linear para estos datos e comprobamos que clasifica ben tódolos datos excepto 3 deles, entre os cales está tamén a observación 84. Polo tanto a regra dos k veciños máis cercanos é lixeiramente mellor que a lineal.

Figura 2.3: Regra kNN para os datos **iris** co número de observacións clasificadas erróneamente no eixo y e k no eixo x.



Observación 2.12. Debemos ter en conta que ao empregar a distancia euclídea, aquelas variables cuxa escala sexa significativamente maior que as das demáis terán un peso maior á hora de medir a distancia. Debido a isto, influirán desproporcionadamente na regra discriminante. Para evitar que isto ocorra, é recomendable estandarizar as variables, de xeito que todas estean na mesma escala.

O método kNN é atractivo pola súa simplicidade conceptual e por non estar suxeito a hipóteses de normalidade. A noción de distancia está ben definida tanto para baixa como alta dimensión, polo que este método é un bo candidato a aplicarse a situacións con moitas variables, posiblemente conectada con técnicas de reducción da dimensión. Na aprendizaxe supervisada, o valor óptimo de k pode acharse empregando cross-validation, aínda que medidas máis simples como ϵ_{mis} e ϵ_{loo} poden ser útiles para buscar entre os posibles valores de k, como fixemos no exemplo.

Cando queremos predecir a clase á que pertencerá un novo dato $\mathbf{X}_{new}$ mediante $\mathfrak{r}_{kNN}$, determinamos o número de veciños óptimo igual ca antes e despois aplicamos a regra con k^*, determinando a moda das etiquetas das observacións na veciñanza $N(\mathbf{X}_{new}, k^*)$.

Unha posible modificación desta regra é considerar pesos $w_1, w_2, \ldots, w_k$ de xeito que se asigne $\mathbf{X}$ á clase C_l se

$$\sum_{i:\mathbf{X}_{(i)}\in\mathcal{C}_l} w_i > \sum_{i:\mathbf{X}_{(i)}\in\mathcal{C}_\nu} w_i \qquad \text{para} \quad \nu \leq \kappa.$$

Esto pode ser útil para darlle máis importancia aos datos máis cercanos a $\mathbf{X}$ á hora de determinar a clase á que se asigna. Tomaríamos $w_1 > w_2 > \cdots > w_k \geq 0$, de xeito que a etiqueta de $\mathbf{X}_{(1)}$ sería a máis determinante na clasificación. Pode que considerando esta regra derivada da dos k veciños máis cercanos obteñamos mellores resultados que coa de pesos uniformes. Ademáis, os pesos tamén poden servirnos para evitar situacións de empate cando hai máis dunha moda na veciñanza $N(\mathbf{X}, k)$.

Observación 2.13. Cando temos unha matriz de datos moi longa, é dicir, con moitas observacións en comparación co número de variables ou n moi grande, seguramente atopemos moitos puntos cerca do que queremos clasificar, co que as veciñanzas serán máis pequenas. En cambio, cando temos d moi grande, moitas máis variables que observacións, os datos están moi dispersos, e a distancia aos puntos máis cercanos pode ser moi grande, tendo así veciñanzas cada vez de maior tamaño, o cal pode supor unha complicación. Precisamos facer adpatacións do método dos k veciños máis cercanos cando estamos nestas situacións.

2.3.2. Regras tipo kernel

Os métodos tipo kernel empregan pesos que diminúen de forma suave cara cero a medida que aumenta a distancia ao punto que queremos clasificar, en vez dos pesos cero-un que emprega a regra kNN. Para casos de alta dimensión, os kernels poden ser modificados para enfatizar algunha variable máis que outras.

Esta regra pode introducirse, como fan en Devroye et al. (1996), dun xeito parecido á motivación da estimación non paramétrica dunha función de densidade. Moitas regras discriminantes fan unha partición de $\mathbb{R}^d$ en celas disxuntas $A_1, A_2, \ldots$ e clasifican dentro de cada cela segundo a maioría das etiquetas das observacións $\mathbf{X}_i$ que caen dentro dela. As celas poden depender dos puntos $\mathbf{X}_1, \ldots, \mathbf{X}_n$, pero as etiquetas non poden influir na súa construción. As celas deberían ser suficientemente pequenas para captar os cambios locais da distribución de $\mathbf{X}$ pero tamén suficientemente grandes para conter moitos puntos da mostra. Este tipo de regras chámanse **regras de particionamento**.

Un tipo particular destas regras é a **regra do histograma**, a cal divide $\mathbb{R}^d$ en hipercubos do mesmo tamaño, e toma unha decisión segundo a etiqueta maioritaria entre as observacións que caen no cubo ao que pertence o $\mathbf{X}$ que queremos clasificar. Neste caso, a

partición de $\mathbb{R}^d$ sería con $A_i = \prod_{i=1}^{n}[k_i h, (k_i+1)h)$, onde $h > 0$ é o tamaño dos cubos e $k_i \in \mathbb{Z}$. Podemos escribir a regra de forma matemática como:

$$\mathfrak{r}(\mathbf{X}) = \underset{1\leq\nu\leq\kappa}{\arg\max} \sum_{i=1}^{n_\nu} I_{\{Y_i=\nu\}} I_{\{\mathbf{X}_i \in A(\mathbf{X})\}},$$

onde $A(\mathbf{X}) = A_j$ cando $\mathbf{X} \in A_j$.

A regra do histograma presenta o problema de que a regra é menos precisa cando $\mathbf{X}$ está preto do borde das celas que cando está no medio. Polo tanto os puntos cercanos aos bordes deberían ter menos peso na decisión da clase asignada á cela. Para solventalo podemos introducir a **regra da ventana móbil**, que é máis suave que a do histograma porque toma os datos a certa distancia do punto que queremos clasificar e asígnalle a etiqueta maioritaria entre estos datos. Formalmente defínese como:

$$\mathfrak{r}(\mathbf{X}) = \underset{1\leq\nu\leq\kappa}{\arg\max} \sum_{i=1}^{n_\nu} I_{\{Y_i=\nu, \mathbf{X}_i \in S_{x,h}\}}$$

onde $h > 0$ e $S_{x,h}$ denota a bóla pechada de centro $\mathbf{X}$ e radio h. A regra kNN é un caso particular desta, na que h é a distancia ao k-ésimo dato máis cercano, que cambia para cada $\mathbf{X}$ que queiramos clasificar e depende da mostra, co que non é unha constante.

Podemos facer unha regra aínda máis suave se damos máis peso a aqueles puntos máis cercanos a $\mathbf{X}$ que aos máis distantes. Sexa $K : \mathbb{R}^d \to \mathbb{R}$ unha **función kernel**, normalmente non negativa e monótonamente decrecente sobre os radios partindo da orixe. A **regra discriminante kernel** vén dada por

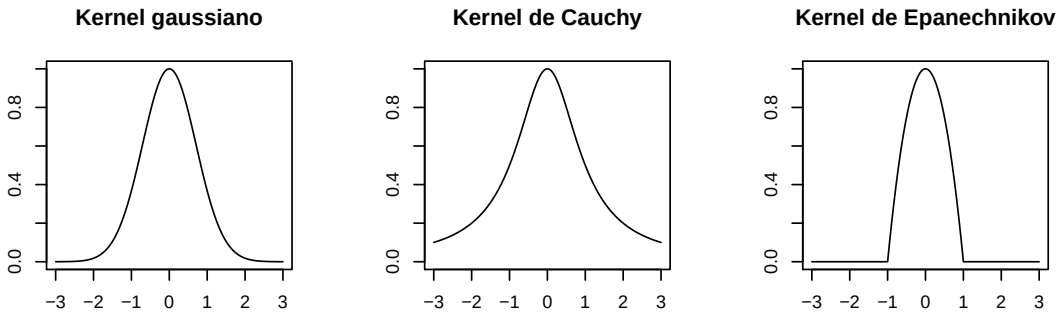
$$\mathfrak{r}(\mathbf{X}) = \underset{1\leq\nu\leq\kappa}{\arg\max} \sum_{i=1}^{n_\nu} I_{\{Y_i=\nu\}} K\left(\frac{\mathbf{X}-\mathbf{X}_i}{h}\right).$$

Ao parámetro h chámaselle **ancho de banda**, e proporciona unha especie de ponderación da distancia. Canto maior é h, máis contan as observacións que non están tan cerca de $\mathbf{X}$. Claramente, a regra kernel é unha xeneralización da da ventana móbil, co kernel naive $K(x) = I_{\{x\in S_{0,1}\}}$. Algunhas das funcións kernel máis empregadas son:

- Kernel gaussiano $K(x) = e^{-\|x\|^2}$,
- Kernel de Cauchy $K(x) = \frac{1}{1+\|x\|^{d+1}}$,
- Kernel de Epanechnikov $K(x) = (1-\|x\|^2) I_{\{\|x\|\leq 1\}}$,

sendo $\|\cdot\|$ a distancia euclídea. Na Figura 2.4 vese a forma destas funcións para unha dimensión. Os dous primeiros teñen como soporte todo $\mathbb{R}$, mentres que o de Epanechnikov soamente é non nulo no intervalo $[-1, 1]$.

Figura 2.4: Gráficas das funcións kernel máis usuais.



Observación 2.14. As regras tipo kernel tamén se poden definir a partir da regra de Bayes, estimando a función de densidade de cada clase empregando métodos non paramétricos,

$$\hat{f}_\nu(\mathbf{X}) = \frac{1}{n_\nu} \sum_{i=1}^{n_\nu} I_{\{Y_i=\nu\}} K\left(\frac{\mathbf{X}-\mathbf{X}_i}{h}\right),$$

onde K é unha función kernel coma as anteriores e h é un parámetro de suavización. Esta aproximación é a que se toma en Klamelä (2014).

Polo que, coñecidas as probabilidades a priori $\pi_1, \ldots, \pi_\kappa$, teriamos a expresión seguinte para a regra.

$$\mathfrak{r}(\mathbf{X}) = \underset{1\leq\nu\leq\kappa}{\arg\max} \left\{ \pi_\nu \frac{1}{n_\nu} \sum_{i=1}^{n_\nu} I_{\{Y_i=\nu\}} K\left(\frac{\mathbf{X}-\mathbf{X}_i}{h}\right) \right\}.$$

Se empregamos as proporcións mostrais como probabilidades a priori $\pi_\nu = \frac{n_\nu}{n}$, esta regra é equivalente á que definimos como regra tipo kernel.

Capítulo 3

Estimación do erro e avaliación de regras discriminantes

Tras definir unha regra discriminante e estimala a partir da mostra, é preciso ter medidas da súa calidade. Neste capítulo definiremos algunhas medidas do erro de clasificación e ilustraremos os procedementos presentados ata agora a través dalgún exemplo.

Pretendemos adiviñar Y a través da regra e a partir dos datos etiquetados, xa que normalmente non coñecemos a distribución do vector etiquetado. Polo tanto é esencial estimar a probabilidade de erro dunha regra $L(\mathfrak{r})$ para saber que esperar da calidade desta. É de especial interese estimar a probabilidade de erro óptima L^*, xa que se é grande, sabemos que calquera regra empregada fará unha clasificación bastante pobre, además de que comparar $L(\mathfrak{r})$ con L^* dinos canto podemos chegar a mellorar a regra $\mathfrak{r}$.

Exemplo 3.1. Consideramos que temos dúas poboacións normais tridimensionais con distintos vectores de medias pero a mesma matriz de covarianzas. Simularemos datos de situacións en que cambie a proporción de datos que pertencen a cada unha das clases e tamén a dificultade do problema de clasificación, determinada pola distancia entre as medias das clases. Co segundo punto do Teorema 1.11 poderemos calcular a regra de Bayes baixo suposicións de normalidade, considerando as probabilidades a priori ou non para así comparar o seu comportamento, entre elas e respecto ao erro de Bayes.

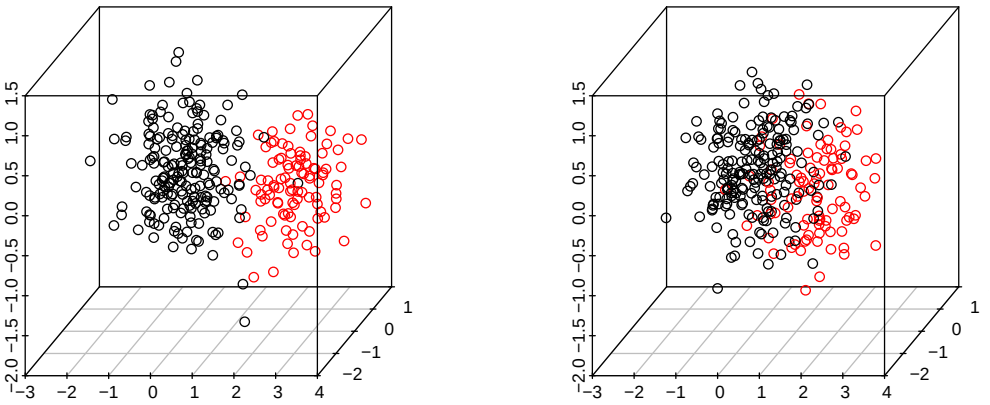
Tomamos os seguintes parámetros para definir as clases:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 2 \\ 1 \\ -0{,}5 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 0{,}5 & 0 & 0 \\ 0 & 0{,}25 & 0 \\ 0 & 0 & 0{,}25 \end{bmatrix} \quad \text{e} \quad \boldsymbol{\mu}_3 = \begin{bmatrix} 1 \\ 1 \\ -0{,}5 \end{bmatrix},$$

de xeito que $\mathcal{C}_1 = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma)$ e $\mathcal{C}_2 = \mathcal{N}(\boldsymbol{\mu}_2, \Sigma)$ para o caso 'fácil' e o mesmo pero cambiando

$\boldsymbol{\mu}_2$ por $\boldsymbol{\mu}_3$ na segunda clase para o caso 'difícil'. Na Figura 3.1 móstranse simulacións para o caso fácil e o difícil, con 200 datos da clase $\mathcal{C}_1$ en negro e 100 da clase $\mathcal{C}_2$ en vermello.

Figura 3.1: Datos simulados de dous problemas de distinta dificultade con dúas clases normais.



Podemos calcular o erro de Bayes para estas dúas situacións, no suposto de coñecer as probabilidades a priori (2/3 dos datos son da primeira clase e 1/3 da segunda) e tamén de que nos falte esta información. Como notación, $\mathfrak{r}^*$ representa a regra sen usar $\boldsymbol{\pi}$ e $\mathfrak{r}_{Bayes}$ a que sí as emprega. Faremos as contas só para o problema 'fácil' coa regra $\mathfrak{r}^*$ xa que as do resto de casos son análogas. A regra óptima é lineal, como vimos no Teorema 2.5, polo tanto

$$L^* = \mathbb{P}(\mathfrak{r}^*(\mathbf{X}) \neq Y) = \mathbb{P}(\mathfrak{r}^*(\mathbf{X}) = 1|Y = 2)\pi_2 + \mathbb{P}(\mathfrak{r}^*(\mathbf{X}) = 2|Y = 1)\pi_1$$

$$= \mathbb{P}(h(\mathbf{X}) > 0|Y = 2)\pi_2 + \mathbb{P}(h(\mathbf{X}) < 0|Y = 1)\pi_1,$$

sendo $h(\mathbf{X}) = -4X_1 - 4X_2 + 2X_3 + 6{,}5$.

Debido a que a matriz de covarianzas Σ é diagonal, tense que as compoñentes X_i do vector aleatorio son normais unidimensionais independentes. Co cal, cando $\mathbf{X} \in \mathcal{C}_1$, $h(\mathbf{X}) \in \mathcal{N}(6{,}5, 13)$ e cando pertence á segunda clase, $h(\mathbf{X}) \in \mathcal{N}(-6{,}5, 13)$. Os cálculos polo tanto darán que

$$L^* = \frac{1}{3}\mathbb{P}(Z > \frac{6{,}5}{\sqrt{13}}) + \frac{2}{3}\mathbb{P}(Z < \frac{-6{,}5}{\sqrt{13}} = \mathbb{P}(Z > \frac{6{,}5}{\sqrt{13}}) = 0{,}0357.$$

En cambio, para o caso difícil, tense que $L^* = 0{,}1217$, é dicir, espérase que clasifiquemos mal o 12 % das observacións. O erro de Bayes podémolo calcular cando coñecemos os parámetros poboacionais, pero de non ser así, debe estimarse a partir das regras lineais estimadas.

Para ver o comportamento das regras con e sen probabilidades a priori, repetiremos a simulación dos datos e cálculo da regra 100 veces. Así podemos calcular a media e tamén a desviación típica da porcentaxe de observacións mal clasificadas. Os resultados móstranse na seguinte táboa, para todos os casos considerados e cambiando o tamaño das mostras das poboacións, n_1 e n_2.

Cadro 3.1: Erro de clasificación medio e desviación típica (entre parénteses) para as dúas regras e os catro conxuntos de datos simulados para o caso fácil e o difícil, en tanto por cento.

Tamaño da mostra		Caso fácil		Caso difícil	
n_1	n_2	$\mathfrak{r}^*$	$\mathfrak{r}_{Bayes}$	$\mathfrak{r}^*$	$\mathfrak{r}_{Bayes}$
30	10	2.6	2.175	7.425	6.150
		(2.796)	(2.427)	(4.955)	(3.915)
300	100	3.6675	3.11	9.06	7.2425
		(0.983)	(0.779)	(1.61)	(1.257)
20	30	2.8	2.72	8.86	8.7
		(2.412)	(2.503)	(4.151)	(4.113)
200	300	3.422	3.306	9.174	8.896
		(0.799)	(0.769)	(1.219)	(1.146)

Observando os resultados do cadro, os erros cometidos no caso difícil son maiores para todas as regras e tamaños mostrais. Ademáis, canto menos se parecen as probabilidades a priori (que neste caso son a proporción de datos de cada clase), a vantaxa que supón empregar $\mathfrak{r}_{Bayes}$ é maior respecto a usar $\mathfrak{r}^*$, sobre todo nos problemas difíciles. Podemos salientar tamén que para as mostras de maior tamaño n cométese un porcentaxe maior de erros de clasificación (aínda que a desviación típica é máis pequena) que para o mesmo problema pero con menos observacións.

No exemplo supuxemos que as probabilidades a priori son correctas, pero de non selo poden empeorar a toma de decisións.

Posto que en xeral quen constrúe a regra non coñece a distribución dos datos, son necesarios métodos de estimación ou cotas da probabilidade de erro que non dependan da distribución de $\mathbf{X}$. Mediremos a calidade dunha regra construida a partir de $\mathbb{X} = [\mathbb{X}^{[1]}\mathbb{X}^{[2]} \dots \mathbb{X}^{[\kappa]}]$ coa probabilidade de erro condicional $\mathbb{P}(\mathfrak{r}(\mathbf{X}; \mathbb{X}) \neq Y|\mathbb{X})$. O problema disto é que avaliar a regra sobre os mesmos datos usados para construíla, lévanos a un sesgo optimista.

Igual que debemos comprobar o funcionamento dunha regra con distintas medidas do erro, tamén debemos achar as limitacións das distintas medidas do erro empregándoas sobre varias regras.

Definimos a continuación un par de medidas de erro de clasificación, estimadores naturais da probabilidade de erro que se calculan a partir dos datos.

Definición 3.2. Sexan $\mathbb{X} = [\mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_n]$ datos etiquetados e $\mathfrak{r}$ unha regra. Un **factor de costes** $\mathbf{c} = [c_1 \cdots c_n]$ asociado a $\mathfrak{r}$ defínese como

$$c_i \begin{cases} = 0 & \text{se } \mathfrak{r} \text{ clasificou } \mathbf{X}_i \text{ correctamente} \\ > 0 & \text{se a clasificou incorretamente.} \end{cases}$$

O **erro de clasificación** ϵ_{mis} da regra $\mathfrak{r}$ e factor de costes $\mathbf{c}$ é o erro dado por

$$\epsilon_{mis} = \frac{1}{n}\sum_{i=1}^{n} c_i.$$

Cando $\mathbf{X}_i$ é clasificado incorrectamente, adoitamos usar $c_i = 1$. Un erro de clasificación con factor de costes 0 e 1 é natural pois, sen importar o número de clases que haxa, o que fai é contar o número de observacións que foron mal clasificadas.

De todos modos, outros factores de costes, positivos e non constantes, son útiles. Por exemplo, cando queremos distinguir entre falsos positivos e falsos negativos nunha clasificación binaria, como sería identificar clientes morosos. Pode que nos interese penalizar máis aqueles erros nos que se predí que un cliente pagará cando non o vai facer que a situación en que un cliente non moroso se clasifique como moroso, pois a primeira pode poñer nun risco maior á empresa. Outro exemplo sería que as clases indiquen o grao de severidade dunha característica. Se temos un cancro en estado IV, que xa se estendeu a partes distantes do corpo, podemos dar pesos maiores ao erro de clasificalo nun estado 0, no que só hai presenza dalgunhas células anormais, que ao de clasificalo como estado II, que xa hai cancro presente e estendido a tecidos cercanos. Así reflexaríamos que o primeiro erro é máis grave que o segundo.

A fase de aprendizaxe consiste na construción dunha regra empregando os datos etiquetados $\mathbb{X}_0$ e a fase de predicción en aplicar dita regra a $\mathbb{X}_{new}$ para predicir as respostas $\mathbf{Y}_{new}$.

Normalmente, quereremos empregar tódolos datos dipoñibles para entrenar o clasificador, co cal só temos unha mostra de entrenamento, pero non de testeo. Primeiro derivamos unha regra $\mathfrak{r}_0$ da submostra $\mathbb{X}_0$ de $\mathbb{X}$, e calculamos o erro de clasificar as observacións $\mathbb{X}_p$ que nos quedan de $\mathbb{X}$ empregando $\mathfrak{r}_0$. $\mathbb{X}_0$ é a **mostra de entrenamento** e $\mathbb{X}_p$ a **mostra de testeo**. Este proceso é a base da idea de entrenamento e testeo da Aprendizaxe Supervisada. Podemos escoller $\mathbb{X}_0$ de moitas maneiras, a máis simple sería usar tódalas observacións excepto unha.

Definición 3.3. Consideramos os datos etiquetados $\mathbb{X} = [\mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_n]$ e a regra $\mathfrak{r}$. Para cada $i \leq n$ tomamos $\mathbb{X}_{0,(-i)} = [\mathbf{X}_1 \cdots \mathbf{X}_{i-1}\mathbf{X}_{i+1} \cdots \mathbf{X}_n]$, ao que chamamos **conxunto de entrenamento leave-one-out** (deixar un fóra) i-ésimo, e $\mathbb{X}_{p,(-i)} = \mathbf{X}_i$. Sexa $\mathfrak{r}_{(-i)}$ a regra construida igual que $\mathfrak{r}$ pero baseada en $\mathbb{X}_{0,(-i)}$, considerando así $\mathbf{X}_i$ unha "nova" observación.

Un factor de costes $\mathbf{k} = [k_{-1}k_{-2} \cdots k_{-n}]$ asociado coas regras $\mathfrak{r}_{(-i)}$ é

$$k_{-i} \begin{cases} = 0 & \text{se } \mathbf{X}_i \text{ foi clasificada correctamente por } \mathfrak{r}_{(-i)} \\ > 0 & \text{se a clasificou incorretamente.} \end{cases}$$

Co cal o **erro leave-one-out**, ϵ_{loo}, baseado nas n regras $\mathfrak{r}_{(-i)}$, os valores asignados $\mathfrak{r}_{(-i)}(\mathbf{X}_i)$ e o factor de costes $\mathbf{k}$ será

$$\epsilon_{loo} = \frac{1}{n}\sum_{i=1}^{n} k_{-i}.$$

Así, o **método leave-one-out** consistirá na elección dos conxuntos de entrenamento $\mathbb{X}_{0,(-i)}$, a construción das regras $\mathfrak{r}_{(-i)}$ e o cálculo do erro ϵ_{loo}. Deste xeito cada observación déixase fóra exactamente unha vez. Adóitase tomar costes 0 e 1, xa que así contamos o número de erros de clasificación empregando $\mathfrak{r}_{(-i)}$.

Polo xeral, tódalas regras $\mathfrak{r}_{(-i)}$ serán distintas, e difiren tamén de $\mathfrak{r}$. Un bo clasificador caracterizarase porque esta diferencia se faga pequena a medida medra o tamaño da mostra. Se comparamos ϵ_{mis} e ϵ_{loo}, ϵ_{mis} debería ser máis pequeno pois cada punto é parte do conxunto de entrenamento da regra usada para clasificalo, pero ϵ_{loo} introduce a maiores unha medida do erro da predicción.

Moitas veces, será preferible tomar un conxunto $\mathbb{X}_p$ con máis dun elemento, para analizar mellor o comportamento da regra "fóra da mostra". Outro modo de dividir os datos dispoñibles entre mostra de entrenamento e de testeo é a seguinte:

Definición 3.4. Sexan $\mathbb{X} = [\mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_n]$ un conxunto de datos etiquetados e $\mathfrak{r}$ unha regra. Sexan k e m enteiros tales que $km = n$. Para $j \leq m$, definimos o **conxunto de**

entrenamento j-ésimo $\mathbb{X}_{0,j}$ e o **conxunto de testeo j-ésimo** $\mathbb{X}_{p,j}$ como

$$\mathbb{X}_{0,j} = [\mathbf{X}_1 \cdots \mathbf{X}_{(j-1)k} \quad \mathbf{X}_{jk+1} \cdots \mathbf{X}_n]$$

$$\mathbb{X}_{p,j} = [\mathbf{X}_{(j-1)k+1} \cdots \mathbf{X}_{jk}].$$

Sexa $\mathfrak{r}_j$ a regra derivada soamente de $\mathbb{X}_{0,j}$ e usámola para calcular o erro de clasificación para cada $\mathbf{X}_i$ do conxunto $\mathbb{X}_{p,j}$. O erro ϵ_j sobre $\mathbb{X}_{p,j}$ é

$$\epsilon_j = \sum_{i=(j-1)k+1}^{jk} c_i$$

sendo $c_i = 0$ cando $\mathfrak{r}_j(\mathbf{X}_i = Y_i$ e $c_i > 0$ noutro caso.

O **erro de validación cruzada m veces** (m-fold cross-validation), ϵ_{cv} é

$$\epsilon_{cv} = \frac{1}{n} \sum_{j=1}^{m} \epsilon_j.$$

Chámaselle entón **validación cruzada m veces** á partición dos datos en $\{\mathbb{X}_{0,j}, \mathbb{X}_{p,j}\}$ m veces xunto coas regras $\mathfrak{r}_j$ e os erros ϵ_j e ϵ_{cv}. Neste proceso, as mostras de testeo sempre teñen k vectores e son disxuntas, de xeito que cada observación $\mathbf{X}_i$ dos datos orixinais $\mathbb{X}$ é contada unha única vez de cara ao cálculo de ϵ_{cv}. Pódese ver fácilmente que o erro leave-one-out é un caso particular de cross-validation con $m = n$ e $k = 1$.

Para este método, hai que dividir os datos e entrenar a regra m veces, o cal supón un alto coste computacional, que medra tamén canto máis grande é m. Se k é demasiado grande, entón o conxunto de entrenamento pode ser demasiado pequeno, o cal pode afectar á precisión da regra. Debemos achar valores para m e k de tal xeito que non cheguemos a ter estos problemas, por exemplo tomar m entre 10 e 20 mentres que entre o 10 e 20 % dos datos sexan de testeo en cada partición. A validación cruzada é moi costosa computacionalmente, polo que moitas veces emprégase unha única partición dos datos. Este proceder non é o mellor cando queremos comparar distintos métodos ou regras.

Na clasificación, o número de clases e o número de observacións que pertencen a cada unha delas tamén xogan un papel importante. Se temos dúas clases, unha delas con moitos máis vectores que a outra, debemos escoller coidadosamente o tamaño das mostras de entrenamento e testeo.

Estas dúas últimas medidas do erro son de especial interese, pois achegan información sobre a capacidade predictora da regra. En moitas ocasións, unha regra estímase a partir dunha mostra de xeito que se minimiza o erro de clasificación destos datos coñecidos, pero ao aplicala a novos datos comete un erro moito maior. Isto coñécese como *overfitting* e

supón un problema ao construir unha regra discriminante, sobre todo cando o número de variables é moi grande, como veremos no seguinte capítulo.

Exemplo 3.5. Volvemos a considerar os catro conxuntos de datos simulados do Exemplo 2.9, dous deles normais, e outros dous non normais, con igualdade de matriz de covarianzas e sen ela.

Para clasificalos usaremos regras tipo kernel, considerando distintas funcións kernel (normal, o de Epanechnikov e uniforme) para poder comparar os resultados obtidos. Empregarase a función `classif.np` do paquete `fda.usc`. Esta función calcula de maneira empírica o parámetro h que leva á clasificación óptima, empregando unha modificación do criterio de validación cruzada, e devolve tamén unha estimación da probabilidade de clasificación correcta para cada clase, indicada na seguinte táboa.

Cadro 3.2: Probabilidades de clasificación correcta por clase, para cada función kernel considerada.

	N1		N2		P1		P2	
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$
Normal	0.835	0.943	0.93	0.9867	0.825	0.92	0.895	0.903
Epanechnikov	0.83	0.94	0.9	0.9967	0.84	0.92	0.91	0.8767
Uniforme	0.815	0.95	0.875	0.9967	0.77	0.933	0.885	0.8967

Pódese apreciar que a probabilidade de clasificar correctamente observacións da segunda clase é maior en tódolos casos, isto débese a que esta clase é a maioritaria nos nosos conxuntos de datos. En canto a determinar cal dos tres kernel é máis adecuado para cada unha das situacións, os resultados non son concluíntes, pois todos teñen un poder predictivo similar e bastante elevado.

Para o primeiro caso, podemos calcular o erro de Bayes partido da función de densidade da normal multivariante. Para N1 sería $L^* = 0{,}097$, co que se clasificarían correctamente o 90,3 % dos datos, que é máis ou menos o que se obten cas regras tipo kernel. L^* é o valor ao que debería converxer a probabilidade de erro ao medrar o tamaño mostral se usásemos unha estimación da regra de Bayes. Este é o caso das regras tipo kernel, nas que a estimación se fai con técnicas non paramétricas, a diferenza da regra linear.

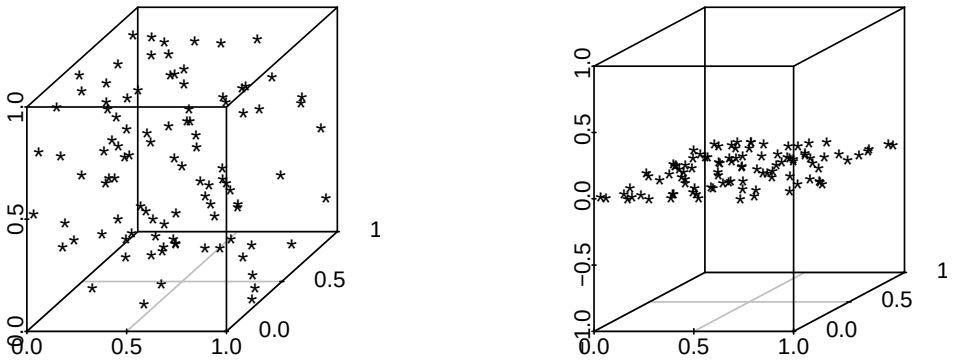
Capítulo 4

Consideracións sobre a alta dimensión e Big Data

Habitualmente tratamos con matrices de datos $n \times d$ nas que o número de observacións é moito maior que o número de variables. Trátase da situación clásica na que $n > d$ e os teoremas do límite clásicos funcionan. Cando falamos de alta dimensión, podemos atoparnos con distintas situacións e problemas.

A medida que aumentan as dimensións, o espazo está cada vez máis baleiro e as distancias entre as observacións aumentan, desaparecendo así as nubes de puntos e pasando a haber puntos máis solitarios. Pódese ver na Figura 4.1 como ao simular datos tridimensionais con distribución uniforme en $[0,1]^3$, están moito máis separados no espazo que proxectados sobre o plano XY. Esta situación pode presentar problemas en métodos como o dos k veciños máis cercanos, ao aumentar a distancia entre os puntos, tamén o fai o tamaño das veciñanzas.

Figura 4.1: Distribución de 100 datos simulados en 3D e en 2D.



A alta dimensión non está definida con precisión. Hai situacións en que ter máis dunhas cuantas variables se considera alta dimensión, mentres que noutras 30 é unha cantidade moderada de variables. Pero máis importante aínda é a relación entre d e n. Os conxuntos de datos en moitas dimensións chámanse **datos de alta dimensión** ou high-dimensional data **HDD**. Distinguimos entre varios tipos de HDD segundo:

1. d é grande pero menor que n,
2. d é grande, máis grande que n, é dicir **datos de alta dimensión e mostra de baixo tamaño** ou high-dimensional low sample size data **HDLSS**, e
3. os datos son funcións dunha variable continua d, chamados **datos funcionais** (as observacións son curvas, en lugar de valores de variables individuais).

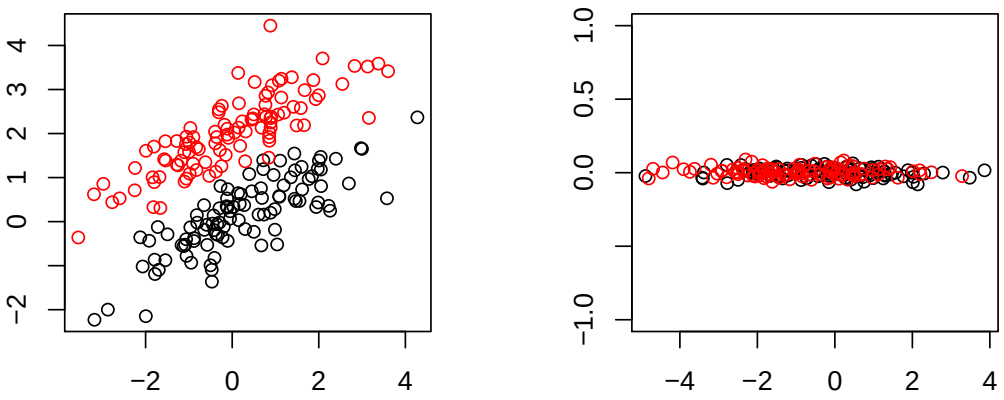
Un dos problemas cos que nos atopamos en Big Data é a gran cantidade de información da que dispoñemos. Moitas veces o simple cálculo da media é inabarcable, debido ao seu coste computacional, por iso intentaremos na medida do posible reducir o número de parámetros a estimar no noso método. Isto pode facerse impoñendo estruturas máis sinxelas sobre os modelos, como por exemplo que as matrices de covarianzas sexan diagonais ou incluso que sexan múltiplos da matriz identidade. Aínda que ás veces non se axuste á realidade, pode levarnos a mellores resultados que un modelo máis complexo e difícil de calcular.

Ademáis, cando analizamos datos de alta dimensión, é de gran importancia a reducción do número de variables. Para HDLSS, unha cota superior para o número de variables pode ser o rango da matriz de covarianzas mostral, aínda que normalmente este número é bastante máis grande que o número de variables útiles, é dicir, que conteñan información significativa.

Unha maneira de reducir o número de variables é empregar a análise de compoñentes principais (ACP). Esta pretende achar combinacións de variables que expliquen a máxima proporción de variabilidade dos datos posible, é dicir, as direccións con maior varianza. Para iso atópanse os autovectores asociados aos autovalores máis grandes da matriz de covarianzas do noso vector aleatorio ou conxunto de datos.

Empregar este método para a clasificación pode levar a situacións moi engañosas. Por exemplo, na Figura 4.2 móstranse dúas clases normais ben separadas e fáciles de clasificar, que ao ser proxectadas sobre a dirección de maior variación dos datos (a correspondente á primeira compoñente principal) se solapan, facéndoos moi complicados de clasificar. Este problema pode ser fácil de identificar cando a dimensión é pequena, pero para d grande posiblemente sexa difícil decatarnos da perda de información producida ao proxectar os datos.

Figura 4.2: Nubes de puntos de dúas clases normais paralelas e a súa proxección sobre a primeira compoñente principal.



Por isto, o enfoque de Fisher de maximizar a variación entre as clases mentres que a minimiza dentro de cada unha delas e achar a dirección discriminante $\boldsymbol{\eta}$ que faga isto,

pode ter máis sentido cando queremos reducir o número de variables nun problema de clasificación. Inténtase deste xeito diferenciar as clases e non resumir o conxunto total dos datos.

Outra situación na que debemos ter moito coidado é cando tratamos con dimensións grandes, especialmente se $d > n$, xa que a matriz de covarianzas S ten rango $r \leq \text{mín}\{n, d\}$, e polo tanto pode non ser invertible, estropeando así a estrutura matemática de moitas das regras que definimos. O mesmo ocorrería para a estimación da matriz W empregada na regra de Fisher, que coincide coa suma das matrices de covarianzas de cada clase. Este será o principal problema no que se centra este capítulo, solventándoo a través da regularización. A modo de exemplo dos avances feitos no campo da clasificación en alta dimensión das últimas décadas, presentaranse un par de métodos que empregan estas técnicas.

4.1. Reducción da dimensión e regularización

Para estimar a regra linear en contextos de alta dimensión, pódense atopar moitos métodos de regularización centrados en que Σ ou $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ sexan dispersos. Outro enfoque máis flexible sería estimar a cantidade $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, pedíndolle a esta que sexa dispersa. Esta idea é a base do método, totalmente baseado nos datos e que non precisa de parámetros de axuste, proposto no recente artigo de Cai e Zhai (2019), onde fan un detallado estudo teórico do seu comportamento. Isto danos unha idea da gran producción científica actual que se está a desenvolver na área que se introduce neste capítulo.

Presentamos a continuación un par de técnicas ás que se pode recurrir ao fallar a invertibilidade da matriz de covarianzas ou a matriz de variabilidade within-class, W.

4.1.1. Análise discriminante de Fisher

Este método combina unha reducción da dimensión e un método de análise discriminante e pode levar a bos resultados en alta dimensión.

O primeiro paso é calcular a dirección $\boldsymbol{\eta}$ da Definición 2.3 que maximiza a varianza entre as clases e a minimiza dentro de cada clase. Despois de $\boldsymbol{\eta}_1$, áchase a seguinte dirección $\boldsymbol{\eta}_2$ ortogonal á primeira respecto da matriz W que maximice o cociente $\mathfrak{q}(\mathbf{e})$. Pódense calcular así sucesivas direccións $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_k$ sobre as cales se poden proxectar os datos para reducir a dimensión. Ás variables proxectadas resultantes chámaselles ás veces variables discriminantes, e son útiles tanto para a clasificación como para a visualización. Finalmente empregamos unha regra discriminante sobre estas novas variables, como a kNN ou a linear.

Un dos problemas deste método é que pode que as variables discriminantes sigan sendo redundantes e que chegue un subconxunto delas para facer a clasificación. Isto é unha mostra de que incrementar o número de variables non leva necesariamente a aumentar o poder discriminante.

Además, como estas direccións son autovectores da matriz $W^{-1}B$ do Teorema 2.2, cando a estimación de W a partir da mostra non é invertible por non ter rango máximo, non se poderían facer estos cálculos.

En Qiao et al. (2008) proponse unha modificación desta matriz que permita usar esta técnica de reducción da dimensión no contexto HDLSS. Tomado un parámetro de control $\gamma > 0$, defínese a matriz

$$\tilde{W} = W + \frac{\gamma}{d}\mathrm{tr}(W)I,$$

e substitúese onde apareza W no método empregado.

4.1.2. Análise discriminante regularizado

En Friedman (1989) introdúcese un método que lidia co problema de non poder invertir as matrices de covarianzas. Empréganse dous parámetros para diseñar un clasificador a medias entre a regra linear normal e a regra cuadrática.

Para $\alpha, \gamma \in [0, 1]$, considérase unha modificación da matriz de covarianzas mostral

$$S_\nu(\alpha, \gamma) = \alpha S_\nu + (1 - \alpha)[\gamma S_{pool} + (1 - \gamma)s^2 I],$$

onde S_ν é a matriz de covarianzas mostral da clase $\mathcal{C}_\nu$, S_{pool} é a matriz agrupada e s é un escalar escollido adecuadamente.

Obtense así unha familia biparamétrica de matrices $S_\nu(\alpha, \gamma)$ e tamén de regras $\mathfrak{r}_{\alpha,\gamma}$ definidas formalmente igual que a regra cuadrática, cambiando a matriz de covarianzas de cada clase Σ_ν pola matriz regularizada correspondente $S_\nu(\alpha, \gamma)$. Por último escolleríanse os parámetros óptimos que combinen a estabilidade da solución e o mellor comportamento da regra.

Estos procedementos menciónanse tamén en Qin (2018), como técnicas para adaptar a regra cuadrática clásica $\mathfrak{r}_{quad}$ á alta dimensión. Comentan que a literatura se centra en xeral en variacións da regra linear e a cuadrática recibe moita menos atención.

4.2. Análise dicriminante linear regularizado disperso: sparse rLDA

O *Sparse Regularized Linear Discriminant Analysis* (sparse rLDA) presentado por Qiao et al. (2008) parte da regra de Fisher e introduce a selección de variables impoñendo que o autovector $\boldsymbol{\eta}$ sexa disperso, é dicir, que case todas as súas compoñentes sexan nulas. Con isto pretenden identificar aquelas variables que son determinantes para diferenciar as clases entre a estrutura de covarianzas que as variables do problema poidan presentar e así descartar información redundante que perxudique a clasificación.

Cando a matriz de variabilidade within-class W é singular, substitúese no problema pola matriz regularizada $\tilde{W}$ presentada ao comezo deste capítulo. Dise neste artigo que a elección do parámetro γ non inflúe nos resultados obtidos, sempre que este sexa pequeno.

O rLDA consiste en empregar esta matriz regularizada, achar a primeira dirección discriminante da mesma, $\boldsymbol{\eta}$ e proxectar os datos sobre ela. Estas proxeccións considéranse como novas variables discriminantes unidimensionais, ás que se lle aplican regras discriminantes como a do centroide máis cercano, SVM ou unha regra kNN, por exemplo. Chegando así a unha clasificación dos datos orixinais.

Este vector $\boldsymbol{\eta}$ pode ter moitas compoñentes non nulas, de xeito que tódalas variables dos datos influirán na clasificación. Ao introducir o sparse rLDA, impoñen que $\boldsymbol{\eta}$ teña só unhas poucas entradas non nulas e están facendo unha selección de variables e eliminando a información redundante.

Para obter variables discriminantes dispersas, primeiro relacionan o vector $\boldsymbol{\eta}$ co vector de coeficientes dunha regresión transformando o problema de autovalores do Teorema 2.2 nun problema de regresión, e despois plantéxano como un problema de mínimos cadrados no que introducen unha penalización da norma $L1$ do vector de coeficientes na función obxectivo igual que no problema LASSO de Tibshirani (1996).

4.3. Análise discriminante en alta dimensión: HDDA

Nesta sección, presentaremos o método introducido por Bouveyron et al. (2007), ao que chaman *High Dimensional Discriminant Analysis* (HDDA). Baséase en que cando a dimensión é moi grande ocurre o fenómeno do espacio vacío e podemos supoñer que os datos viven en subespacios de dimensión menor. O que fai é reducir a dimensión para cada clase $\mathcal{C}_\nu$ de forma independente e impón unha estrutura determinada sobre as matrices de covarianzas Σ_ν para adaptar un contexto gaussiano á alta dimensión, reducindo o número

de parámetros a estimar. Suponse que as clases son esféricas (é dicir, a matriz de covarianzas é un múltiplo da identidade) nestes subespazos ou o que é o mesmo, que Σ_ν teñen só dous autovalores distintos.

Igual que nos métodos de análise discriminante clásicos, supoñemos a normalidade das clases $\mathcal{C}_\nu = \mathcal{N}(\boldsymbol{\mu}_\nu, \Sigma_\nu)$ para $\nu \in \{1, \dots \kappa\}$. Posto que Σ_ν son simétricas, gracias ao Teorema espectral obtemos unha descomposición matricial $\Sigma_\nu = Q_\nu \Delta_\nu Q_\nu^\intercal$, onde Q_ν é unha matriz ortogonal cuxas columnas son unha base de autovectores de Σ_ν e Δ_ν é unha matriz diagonal formada polos autovalores de Σ_ν. Supoñemos que Δ_ν ten dous autovalores diferentes, $a_\nu > b_\nu$. Chamamos E_ν ao subespacio afín de dimensión d_ν xerado polos autovectores asociados ao autovalor a_ν con $\boldsymbol{\mu}_\nu \in E_\nu$, e sexa $E_\nu^\perp$ tal que $E_\nu \oplus E_\nu^\perp = \mathbb{R}^d$ con $\boldsymbol{\mu}_\nu \in E_\nu^\perp$. Consideramos $P_\nu(\mathbf{x}) = \tilde{Q_\nu}\tilde{Q_\nu}^\intercal(\mathbf{x} - \boldsymbol{\mu}_\nu) + \boldsymbol{\mu}_\nu$ a proxección de $\mathbf{x}$ sobre E_ν, onde $\tilde{Q_\nu}$ é a matriz formada polas d_ν primeiras columnas de Q_ν e o resto ceros. Análogamente, $P_\nu^\perp(\mathbf{x}) = (Q_\nu - \tilde{Q_\nu})(Q_\nu - \tilde{Q_\nu})^\intercal(\mathbf{x} - \boldsymbol{\mu}_\nu) + \boldsymbol{\mu}_\nu$ é a proxección de $\mathbf{x}$ sobre $E_\nu^\perp$.

Partindo da regra de Bayes e impoñendo a forma descrita da matriz de covarianzas, a regra discriminante asignará un vector aleatorio $\mathbf{X}$ á clase $\mathcal{C}_\nu$ que minimice a expresión

$$h_\nu(\mathbf{X}) = \frac{\|\boldsymbol{\mu}_\nu - P_\nu(\mathbf{X})\|^2}{a_\nu} + \frac{\|\mathbf{X} - P_\nu(\mathbf{X})\|^2}{b_\nu} + d_\nu \log a_\nu + (d - d_\nu)\log b_\nu - 2\log \pi_\nu.$$

Vexamos como se pode chegar a ela a partir da función de densidade dunha distribución normal. Asignarase $\mathbf{X}$ á clase $\mathcal{C}_\nu$ que maximice $\pi_\nu f_\nu(\mathbf{X})$, ou o que é o mesmo, que minimice $-2\log \pi_\nu f_\nu(\mathbf{X})$.

$$-2\log \pi_\nu f_\nu(\mathbf{X}) = -2\log\left(\pi_\nu (2\pi)^{-d/2}|\Sigma_\nu|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_\nu)^\intercal \Sigma_\nu^{-1}(\mathbf{X} - \boldsymbol{\mu}_\nu)\right\}\right)$$

$$= -2\log \pi_\nu + d\log(2\pi) + \log|\Sigma_\nu| + (\mathbf{X} - \boldsymbol{\mu}_\nu)^\intercal \Sigma_\nu^{-1}(\mathbf{X} - \boldsymbol{\mu}_\nu)$$

$$= -2\log \pi_\nu + d\log(2\pi) + \log\left(a_\nu^{d_\nu} b_\nu^{(d-d_\nu)}\right) + (\mathbf{X} - \boldsymbol{\mu}_\nu)^\intercal Q_\nu \Delta_\nu^{-1} Q_\nu^\intercal (\mathbf{X} - \boldsymbol{\mu}_\nu)$$

$$= -2\log \pi_\nu + d\log(2\pi) + d_\nu \log a_\nu + (d - d_\nu)\log b_\nu$$

$$+(\mathbf{X} - \boldsymbol{\mu}_\nu)^\intercal (\tilde{Q_\nu} + Q_\nu - \tilde{Q_\nu})(\frac{1}{a_\nu}\tilde{Q_\nu} + \frac{1}{b_\nu}(Q_\nu - \tilde{Q_\nu}))^\intercal (\mathbf{X} - \boldsymbol{\mu}_\nu)$$

$$= -2\log \pi_\nu + d\log(2\pi) + d_\nu \log a_\nu + (d - d_\nu)\log b_\nu$$

$$+\frac{1}{a_\nu}(\mathbf{X} - \boldsymbol{\mu}_\nu)^\intercal (\tilde{Q_\nu} + Q_\nu - \tilde{Q_\nu})\tilde{Q_\nu}^\intercal (\mathbf{X} - \boldsymbol{\mu}_\nu) + \frac{1}{b_\nu}(\mathbf{X} - \boldsymbol{\mu}_\nu)^\intercal (\tilde{Q_\nu} + Q_\nu - \tilde{Q_\nu})(Q_\nu - \tilde{Q_\nu})^\intercal (\mathbf{X} - \boldsymbol{\mu}_\nu)$$

$$= -2\log \pi_\nu + d\log(2\pi) + d_\nu \log a_\nu + (d - d_\nu)\log b_\nu + \frac{1}{a_\nu}\|\boldsymbol{\mu}_\nu - P_\nu(\mathbf{X})\|^2 + \frac{1}{b_\nu}\|\mathbf{X} - P_\nu(\mathbf{X})\|^2.$$

Como $d\log(2\pi)$ é constante para tódalas clases, elimínase e chegamos á expresión de h_ν.

Introducimos a seguinte notación para simplificar a interpretación da regra: $a_\nu = \frac{\sigma_\nu^2}{\alpha_\nu}$ e $b_\nu = \frac{\sigma_\nu^2}{1-\alpha_\nu}$, con $\alpha_\nu \in [0,1]$ e $\sigma_\nu^2 > 0$. Así a expresión anterior pódese reescribir como

$$h_\nu(\mathbf{X}) = \frac{1}{\sigma_\nu^2}\left(\alpha_\nu \|\boldsymbol{\mu}_\nu - P_\nu(\mathbf{X})\|^2 + (1-\alpha_\nu)\|\mathbf{X} - P_\nu(\mathbf{X})\|^2\right)$$

$$+2d\log(\sigma_\nu) + d_\nu \log\frac{1-\alpha_\nu}{\alpha_\nu} - d\log(1-\alpha_\nu) - 2\log\pi_\nu.$$

Para certos valores dos parámetros α_ν e σ_ν^2, téñense regras particulares das que xa falamos anteriorimente.

Se $\alpha_\nu = 1/2 \quad \forall\nu$, esta regra non é máis que a regra cuadrática coa suposición adicional de que a matriz de covarianzas é un múltiplo da identidade, $\Sigma_\nu = \sigma_\nu^2 I$. Se a maiores o parámetro $\sigma_\nu^2 = \sigma^2$ é o mesmo para todas as clases, é a regra linear esférica, con $\Sigma = \sigma^2 I$.

Deixando fixos algúns, pero non todos os parámetros involucrados na regra HDDA, pódense obter moitos modelos distintos con claras interpretacións xeométricas. Presentamos dúas opcións:

- **Regra discriminante isométrica** (HDDAi): Fanse as seguintes suposicións:

  $$\alpha_\nu = \alpha, \quad \sigma_\nu = \sigma, \quad d_\nu = d^* \quad \text{e} \quad \pi_\nu = \pi^* \quad \forall\nu \leq \kappa,$$

  de xeito que a expresión que queremos minimizar será

  $$h_\nu(\mathbf{X}) = \alpha\|\boldsymbol{\mu}_\nu - P_\nu(\mathbf{X})\|^2 + (1-\alpha)\|\mathbf{X} - P_\nu(\mathbf{X})\|^2.$$

  Se $\alpha = 0$, HDDAi asignará $\mathbf{X}$ a $\mathcal{C}_l$ se $d(\mathbf{X}, E_l) < d(\mathbf{X}, E_\nu) \quad \forall\nu \neq l$. É dicir, levará $\mathbf{X}$ á clase asociada ao subespazo E_ν máis cercano.

  Se $\alpha = 1$, $\mathfrak{r}(\mathbf{X}) = l$ cando $d(\mu_l, P_l(\mathbf{X})) < d(\boldsymbol{\mu}_\nu, P_\nu(\mathbf{X})) \quad \forall\nu \neq l$, é dicir, leva $\mathbf{X}$ á clase cuxa media está máis cerca da proxección de $\mathbf{X}$ sobre o subespazo.

  Cando $0 < \alpha < 1$, a regra será unha mestura destas dúas. Será necesario tamén estimar α, pero discutirémolo máis adiante.

- **Regra discriminante homotécica** (HDDAh): A diferencia deste método co anterior é que non se impón que σ_ν sexa constante para tódalas clases. Así a expresión queda

  $$h_\nu(\mathbf{X}) = \frac{1}{\sigma_\nu^2}\left(\alpha\|\boldsymbol{\mu}_\nu - P_\nu(\mathbf{X})\|^2 + (1-\alpha)\|\mathbf{X} - P_\nu(\mathbf{X})\|^2\right) + 2d\log\sigma_\nu.$$

  Posto que σ_ν está no denominador, se $\mathbf{X}$ está á mesma distancia de dúas clases, a regra favorecerá aquela cuxa varianza sexa maior.

Obviamente, pode haber situacións en que as hipóteses anteriores sexan demasiado restrictivas, co cal optaremos por empregar a expresión xeral.

4.3.1. Estimación dos parámetros

A estimación de parámetros é necesaria para calquera regra, pois dispoñemos dunha mostra e non da poboación total. Neste caso empregaremos os estimadores de máxima verosimilitude calculados a partir dunha mostra $\mathbb{X}$ de tamaño n.

Igual que fixemos para as regras lineais e cuadráticas, estimaremos as probabilidades a priori polas proporcións mostrais, as medias de cada clase polas medias mostrais e as matrices Σ_ν polas matrices de covarianzas mostrais.

Podemos supoñer nun principio que coñecemos a dimensión dos subespazos d_ν, e así obtéñense os estimadores

$$\hat{a_\nu} = \frac{1}{d_\nu}\sum_{j=1}^{d_\nu}\lambda_{\nu,j} \quad \text{e} \quad b_\nu = \frac{1}{d-d_\nu}\sum_{j=d_\nu+1}^{d}\lambda_{\nu,j},$$

onde $\lambda_{\nu,j}$ son os autovalores de S_ν. A columna j-ésima de Q_ν estímase polo autovector de S_ν asociado ao autovalor $\lambda_{\nu,j}$. Así a_ν e b_ν son estimados polas varianzas empíricas da clase $\mathcal{C}_\nu$ nos subespazos E_ν e $E_\nu^\perp$ respectivamente. Pódense entón deducir os seguintes estimadores:

$$\hat{\alpha_\nu} = \frac{\hat{b_\nu}}{\hat{a_\nu}+\hat{b_\nu}} \quad \text{e} \quad \hat{\sigma_\nu}^2 = \frac{\hat{a_\nu}\hat{b_\nu}}{\hat{a_\nu}+\hat{b_\nu}}.$$

Por último, queda achar os parámetros d_ν. A proposta de Bouveyron et al. (2007) é empregar o método empírico dos *scree plots*, que analiza a diferencia entre os autovalores para atopar unha ruptura na gráfica. Este método baséase en que o autovalor $\lambda_{\nu,j}$ representa a fracción da varianza total correspondente ao j-ésimo autovector de Σ_ν. Escollemos a dimensión a partir da cal as diferencias son moi pequenas respecto á máxima das diferencias. Os *scree plots* empréganse tamén na ACP para decidir o número de compoñentes principais que se consideran para representar os datos, como se explica no segundo capítulo de Koch (2014). Un dos problemas co que nos atopamos ao facer isto é que a medida que d aumenta, os autovalores aportan menos información ao total da varianza, e é máis difícil identificar o índice d_ν a partir do cal o poder explicativo decrece significativamente.

No último capítulo ilustraremos con datos simulados como obter a estimación dos parámetros d_ν, ademáis de comparar os resultados obtidos co HDDA respecto aos obtidos con métodos clásicos cando a dimensión medra.

Capítulo 5

Ilustración sobre datos simulados e reais

Neste último capítulo farase unha simulación de datos empregando R e tamén se tratará unha base de datos real, que mostre como para unha dimensión d grande comezan a fallar os métodos clásicos e os introducidos para o contexto de Big Data funcionan mellor.

Implementarase o método HDDA de Bouveyron et al. (2007) no software R. Neste artigo móstranse os resultados obtidos ao aplicalo ao recoñecemento de obxectos en imaxes reales en comparación con métodos de clasificación clásicos. Agora será aplicado sobre os dous exemplos estudados en Qiao et al. (2008) para poder comparar o comportamento das dúas propostas. Empregaremos o erro de clasificación ϵ_{mis} como medida da calidade das regras, aplicadas sobre o conxunto de entrenamento e tamén sobre o de testeo. Deste xeito mediremos o poder clasificador das regras fóra da mostra.

5.1. Datos simulados

Simularemos un conxunto de datos de entrenamento de tamaño 25 para cada unha das dúas clases e un conxunto de datos de proba de tamaño 100 para cada clase tamén. Os datos serán de dimensión $d = 100$, polo que nos atopamos nun caso HDLSS. Das 100 variables que forman o vector aleatorio $\mathbf{X}$, só as dúas primeiras serán diferentes entre as clases $\mathcal{C}_1$ e $\mathcal{C}_2$. Os datos seguirán unha distribución normal cos seguintes parámetros:

$$\mathbf{X} \sim \begin{pmatrix} \mathcal{N}_2(\boldsymbol{\mu}_\nu, \Sigma) \\ \mathcal{N}_{d-2}(0, \boldsymbol{I}_{p-2}) \end{pmatrix}, \quad \text{para} \quad \nu = 1, 2,$$

$$\text{onde} \quad \boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0{,}9 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ -0{,}9 \end{pmatrix} \quad \text{e} \quad \Sigma = \begin{pmatrix} 1 & 0{,}7 \\ 0{,}7 & 1 \end{pmatrix}.$$

Claramente, a clasificación depende únicamente das dúas primeiras variables, e tódalas demáis aportan información redundante. Identificar aquelas variables suficientes para a clasificación axudará a evitar o *overfitting* da regra sobre a mostra de entrenamento e levará a mellores resultados sobre datos fóra desta mostra.

O enfoque de Qiao et al. (2008) céntrase especialmente no fenómeno chamado *data piling*. Ao empregar unha estimación da matriz de covarianzas ou da variabilidade within-class, a cal non é nada fiable cando $d > n$, as direccións sobre as que se proxectan os datos para reducir a dimensión poden ser moi distintas das teóricas, levando a un sobreaxuste da regra aos datos de entrenamento. O que ocorre é que as proxeccións dos datos de cada clase están moi separados, pero ao proxectar o conxunto de testeo hai un solapamento moito maior ca usando as direccións teóricas. Eles introducen a dispersión no modelo para evitar que isto pase e ilustran como as direccións discriminantes dispersas estimadas se parecen máis ás teóricas, levando a unha pérdida de precisión sobre os datos mostrais pero aumentándoa fóra da mostra.

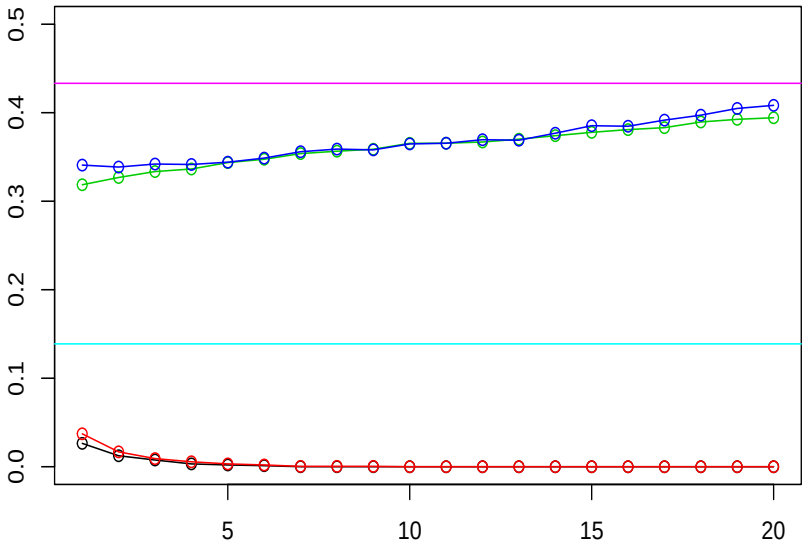
Ao implementar o método HDDA, atopámonos co problema de determinar a dimensión intrínseca d_ν dos subespazos de autovectores de S. Non nos podemos axudar dos *scree plots* ao estar nun contexto HDLSS, co cal para identificar o mellor d_ν, estimaremos a regra para valores de d_ν entre 1 e 20 e escolleremos aquel que nos leve ao menor erro de clasificación. Repetirase a simulación dos datos e o cálculo da regra 50 veces para poder promediar os erros de clasificación.

Tomarase d_ν igual para as dúas clases por simplicidade. As suposicións da regra discriminante isométrica teñen sentido neste caso, xa que ambas clases aparecen na mesma proporción na mostra e teñen a mesma matriz de covarianzas teórica. Tamén podemos esperar que a regra isométrica leve a mellores resultados que a homotécica, ao supoñer que $\sigma_1 = \sigma_2$. Implementaranse soamente estas dúas versións particulares do HDDA.

Na Figura 5.1 observamos como o erro de clasificación no conxunto de entrenamento, en negro para HDDAi e vermello para HDDAh, decrece ao aumentar d_ν, pero fóra da mostra, en verde para HDDAi e azul para HDDAh, aumenta. Tomaremos polo tanto o menor valor posible $d_\nu = 1$, para evitar na medida do posible o *overfitting*.

As liñas horizontais debuxadas mostran o erro promedio ao aplicar a regra linear, que non depende de d_ν. Estos valores foron calculados coa función `lda` do paquete `MASS` de R, que emprega unha pseudoinversa para lidiar coa singularidade da matriz de covarianzas mostral. Está claro que empregando calquera das versións de HDDA obtemos mellores resultados que con esta regra clásica. Para o valor óptimo de d_ν, obtemos un erro do 14 %

Figura 5.1: Erro de clasificación medio para as distintas regras (HDDAi, HDDAh e regra linear) sobre os conxuntos de entrenamento e testeo simulados, fronte aos distintos valores considerados para a dimensión intrínseca d_ν.



sobre a mostra e do 43 % sobre o conxunto de testeo, en azul claro e rosa respectivamente. Como era de esperar, HDDAi da mellores resultados que HDDAh.

Na táboa 5.1 resúmense os resultados obtidos con estes métodos e os que aparecen en Qiao et al. (2008). As tres primeiras columnas correspóndense co `lda`, e as dúas variantes do HDDA implementadas, e mostran os datos obtidos facendo unha simulación que imita a feita por Qiao et al. (2008). As dúas últimas columnas conteñen os valores proporcionados no seu artigo, pois non se implementou o método nin se repetiu a simulación neste traballo por cuestións de tempo. Pódese facer entón unha comparación dos métodos, pero sobre todo é unha ilustración das técnicas presentadas e un exercicio computacional.

Podemos observar que o método implementado de Bouveyron et al. (2007) chega a uns resultados similares aos da regra linear regularizada rLDA, pero non tan bos coma os da sparse rLDA (calculado para 5 coeficientes non nulos, é dicir, 5 variables significativas).

Cadro 5.1: Erro cometido por cada método considerado no conxunto de entrenamento e no de testeo (en tanto por cento).

	LDA	HDDAi	HDDAh	rLDA	sparse rLDA
Entrenamento	13,9	2,6	3,7	0	12
Testeo	43,3	31,8	34	32	13.5

Tanto o HDDA coma o rLDA fan unha reducción da dimensión e solvéntase o problema de que as matrices de covarianzas sexan singulares regularizándoas, pero non hai unha selección de variables implícita. O método disperso busca reducir explícitamente o número de variables explicativas, e por iso obtén a mellor clasificación fóra da mostra. Esta simulación era de tal xeito que só dúas variables diferenciaban as clases, ten sentido que unha regra que dependa de moi poucas variables funcione mellor.

5.2. Datos de medidas de expresión xénica

Consideramos o conxunto de datos Colon de Alon et al. (1999), que contén 42 mostras de tecido tumoral e 20 de tecido de colon normal. Para cada mostra hai 2000 medidas do nivel de expresión de xens. O obxectivo é clasificar as mostras normais e tumorais en función das medidas de expresión xénica.

Estos son os datos empregados en Qiao et al. (2008) para ilustrar o sparse rLDA nun caso real. Fixeron unha análise de dúas etapas, repetida 50 veces para promediar a proporción de erros obtida.

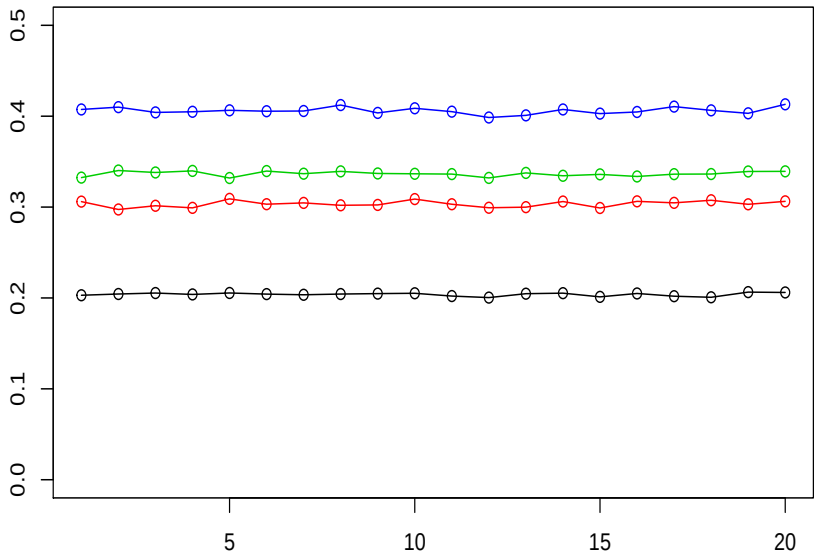
Posto que 2000 variables suporían un esforzo computacional moi grande, reducen a dimensión seleccionando as 200 máis significativas para axilizar os cálculos. Só se dispón dunha mostra de 62 casos en total, que se debe particionar para obter conxuntos de entrenamento e de testeo. Divídese en 2/3 e 1/3 das observacións, respectivamente, de xeito que a proporción de datos tumorais e normais sexa equilibrada. Aos datos resultantes con $d = 200$, aínda nunha situación HDLSS, aplícanselle os métodos rLDA e sparse rLDA usando o centroide máis cercano, SVM e o veciño máis cercano sobre as proxeccións obtidas para distintos números de coeficientes non nulos da dirección discriminante. Pódese ver unha gráfica cos resultados sobre o conxunto de testeo no seu artigo, onde os mellores foron obtidos usando a regra do centroide máis cercano con entre 10 e 20 xens significativos, con erros de clasificación preto do 15 %. Considerando tódalas variantes do sparse rLDA

mencionadas, o erro mantense case sempre por debaixo do 20 %.

Para ilustrar o comportamento da regra de Bouveyron et al. (2007) sobre un conxunto de datos real, procederemos dun xeito parecido ao que se acaba de explicar empregando a implementación do HDDA feita para este traballo. Tamén se repetirá o proceso 50 veces para achar o erro promedio, e de novo outras 20 para ver a desviación deste.

Tomaranse 50 das 2000 variables ao chou cada vez, para poder levar a cabo os cálculos nun ordenador de sobremesa, e empregaranse tanto HDDAi como HDDAh para dimensións intrínsecas entre 1 e 20. A selección de variables previa non pretende quedarse ca información significativa e rexeitar a redundante, polo que non podemos esperar que os resultados que se obteñan sexan moi bos, será máis ben un exercicio computacional para ilustrar o método.

Figura 5.2: Erro de clasificación medio para as variantes do HDDA sobre os conxuntos de entrenamento e testeo sacados dos datos Colon, fronte aos distintos valores considerados para a dimensión intrínseca d_ν.



A Figura 5.2 mostra os resultados obtidos para a versión isométrica e a homotécica, en negro e verde sobre os datos de entrenamento, e en vermello e azul sobre os de testeo, respectivamente. O erro de clasificación medio calculado para ambas regras sobre as dúas mostras ten unha desviación típica entre 0.01 e 0.02. Vemos que independentemente do valor tomado para d_ν, HDDAi leva a unha mellor clasificación en ambas mostras, e aínda así, é peor que os resultados aos que chegan en Qiao et al. (2008). Nin sequera sobre a mostra de entrenamento se clasifican mal menos do 20 % dos datos.

Se asignásemos as observacións á clase de tecido tumoral, $\mathcal{C}_1$, ou a de tecido normal, $\mathcal{C}_2$, basándonos simplemente nas proporcións mostrais, a probabilidade de erro sería

$$L(\mathfrak{r}) = \mathbb{P}(\mathfrak{r}(\mathbf{X}) = 1|\mathbf{X} \in \mathcal{C}_2)\mathbb{P}(\mathbf{X} \in \mathcal{C}_2) + \mathbb{P}(\mathfrak{r}(\mathbf{X}) = 2|\mathbf{X} \in \mathcal{C}_1)\mathbb{P}(\mathbf{X} \in \mathcal{C}_1) =$$

$$= \mathbb{P}(\mathfrak{r}(\mathbf{X}) = 1)\mathbb{P}(\mathbf{X} \in \mathcal{C}_2) + \mathbb{P}(\mathfrak{r}(\mathbf{X}) = 2)\mathbb{P}(\mathbf{X} \in \mathcal{C}_1)$$

$$= \frac{42}{62}\frac{20}{62} + \frac{20}{62}\frac{42}{62} = 0{,}437$$

xa que a asignación a unha clase sería independente da clase á que realmente pertence. É dicir, que clasificaríamos erróneamente o 43.7 % dos datos, co que usar HDDAi si supón unha mellora, pero HDDAh fóra da mostra compórtase máis ou menos igual que o azar.

Estos casos de estudo foron sacados de Qiao et al. (2008) para poder comparar, en certa medida, a proposta de Bouveyron et al. (2007) co rLDA e a súa versión dispersa. Agora que xa vimos cómo actúan os métodos de clasificación introducidos para o contexto do Big Data sobre un caso simulado e outro real particulares, podemos concluír que o que supón unha verdadeira mellora nestos dous casos é a introdución da selección de variables mediante un modelo disperso.

Apéndice A

Scripts utilizados para os exemplos e implementación do HDDA

Neste apéndice recóllese o código de R empregado para xerar as gráficas e táboas que aparecen nos exemplos ao longo do traballo, o usado para a implementación do método HDDA e tamén a súa aplicación a un conxunto de datos simulado e outro real.

A.1. Exemplos

O seguinte trozo de código é a simulación dos datos da Figura 2.1, co cálculo das fronteiras correspondentes á regra de Fisher calculada a partir dos parámetros poboacionais e os mostrais.

```
set.seed(1997)

# Definición dos parámetros de cada clase
mu1<-c(0,2)
mu2<-c(1,1)
E<-cbind(c(0.5,0),c(0,0.5))
n=50 # tamaño da mostra

# Simulación dos datos
X1<-cbind(mvrnorm(20,mu1,E), rep(1,20))
X2<-cbind(mvrnorm(30,mu2,E),rep(2,30))
X<-data.frame(rbind(X1,X2))
```

```
# Cálculo da regra poboacional
mubarra<- 0.5*(mu1+mu2)
B<- (mu1-mubarra)%*%t(mu1-mubarra)+(mu2-mubarra)%*%t(mu2-mubarra)
W<-2.*E
eta<-eigen(solve(W)%*%B)$vectors[,1]
etaT<-eigen(solve(W)%*%B)$vectors[,2]
hX<-(X[,1]-mubarra[1])*eta[1]+ (X[,2]-mubarra[2])*eta[2]
rX<-2-1*(hX>0)

# Puntos da fronteira
h01<-c(mubarra[1]+eta[2]*(-2:2))
h02<-c(mubarra[2]-eta[1]*(-2:2))

# Cálculo regra mostral
m1<-rapply(X[1:20,1:2],mean)
m2<-rapply(X[21:50,1:2],mean)
mbarra<-(m1+m2)/2
S1<-cov(X[1:20,1:2])
S2<-cov(X[21:50,1:2])
Wm<-S1+S2
Bm<- (m1-mbarra)%*%t(m1-mbarra)+(m2-mbarra)%*%t(m2-mbarra)
etam<-eigen(solve(Wm)%*%Bm)$vectors[,1]

#t(etam)%*%(m1-m2) # é negativo, polo que a regra será ao revés

hXm<-(X[,1]-mbarra[1])*etam[1]+ (X[,2]-mbarra[2])*etam[2]
rXm<- 1+1*(hXm>0)

# Puntos da fronteira
h01m<-c(mbarra[1]+etam[2]*(-2:2))
h02m<-c(mbarra[2]-etam[1]*(-2:2))

# Gráfica
par(mfrow=c(1,3))
plot(X[,1:2],col=X[,3],xlab='',ylab='',xlim=c(-1,2),ylim=c(-1,3))
lines(h01,h02,col=4)
```

```
lines(h01m,h02m,col=3)

plot(X[,1:2],col=rX,xlab='',ylab='',xlim=c(-1,2),ylim=c(-1,3),pch=4)
lines(h01,h02,col=4)
points(X[X[,3]!=rX,1:2],col=X[X[,3]!=rX,3],pch=0)

plot(X[,1:2],col=rXm,xlab='',ylab='',xlim=c(-1,2),ylim=c(-1,3),pch=4)
lines(h01m,h02m,col=3)
points(X[X[,3]!=rXm,1:2],col=X[X[,3]!=rXm,3],pch=0)
```

A regra linear mostral e a cuadrática están implementadas en R no paquete `MASS` nas funcións `lda` e `qda`. Definiuse para este traballo unha función para o cálculo da regra de Fisher empregando os parámetros mostrais de dúas clases.

```
regraFisher_mostra<-function(X,d,n1){
m1<-rapply(X[1:n1,1:d],mean)
m2<-rapply(X[(n1+1):nrow(X),1:d],mean)
m<-(m1+m2)/2
W<-cov(X[1:n1,1:d])+cov(X[(n1+1):nrow(X),1:d])
B<- (m1-m)%*%t(m1-m)+(m2-m)%*%t(m2-m)
eta<-Re(eigen(solve(W)%*%B)$vectors[,1])
hX<-(1:nrow(X))*0
for(j in 1:d){
hX<-hX+(X[,j]-m[j])*eta[j]
}
if(t(eta)%*%(m1-m2)>0){
rX<-2-1*(hX>0)
} else{
rX<- 1+1*(hX>0)
}
return(rX)
}
```

Estas liñas xeran as nubes de puntos de datos dimulados da Figura 2.2 e os resultados de clasificalos mostrados no Cadro 2.1.

```
# Parámetros
mu1<-c(0,0,0)
```

```
mu2<-c(2,0,-0.5)
E1<-cbind(c(1,0,0),c(0,0.25,0),c(0,0,0.25))
E2<-cbind(c(0.5,0.5,0.125),c(0.5,1,0),c(0.125,0,0.25))

# Simulación dos datos normais
set.seed(1997)
N1<-cbind(rbind(mvrnorm(200,mu1,E1),mvrnorm(300,mu2,E1)), c(rep(1,200),rep(2,300)))
N2<-cbind(rbind(mvrnorm(200,mu1,E1),mvrnorm(300,mu2,E2)), c(rep(1,200),rep(2,300)))

par(mfrow=c(2,2))
scatterplot3d(N1[,1:3],color=N1[,4],xlab='',ylab='',zlab='',
xlim=c(-4,6),ylim=c(-3,3),zlim=c(-2,1.5),angle=70)
scatterplot3d(N2[,1:3],color=N2[,4],xlab='',ylab='',zlab='',
xlim=c(-4,6),ylim=c(-3,3),zlim=c(-2,1.5),angle=70)

# Simulación dos datos a partir da Poisson
P1<-rbind(cbind(rpois(200,10)+5,rpois(200,10)-5,rpois(200,10)+2,rep(1,200)),
cbind(rpois(300,10),rpois(300,10),rpois(300,10),rep(2,300)))
P2<-rbind(cbind(rpois(200,10)+5,rpois(200,10)-5,rpois(200,10)+2,rep(1,200)),
cbind(rpois(300,10),rpois(300,20)-10,rpois(300,30)-20,rep(2,300)))

scatterplot3d(P1[,1:3],color=P1[,4],xlab='',ylab='',zlab='',angle=70,
zlim=c(-5,25),ylim=c(-5,30),xlim=c(0,30))
scatterplot3d(P2[,1:3],color=P2[,4],xlab='',ylab='',zlab='',angle=70)

# Repetición 100 veces da simulación dos datos anteriores e avaliación das regras
set.seed(1997)
efN1<-efN2<-efP1<-efP2<-enN1<-enN2<-enP1<-enP2<-eqN1<-eqN2<-eqP1<-eqP2<-c()

for(i in 1:100){
# Simulación dos datos
N1<-data.frame(cbind(rbind(mvrnorm(200,mu1,E1),mvrnorm(300,mu2,E1)),
c(rep(1,200),rep(2,300))))
N2<-data.frame(cbind(rbind(mvrnorm(200,mu1,E1),mvrnorm(300,mu2,E2)),
c(rep(1,200),rep(2,300))))
P1<-data.frame(rbind(cbind(rpois(200,10)+5,rpois(200,10)-5,rpois(200,10)+2,rep(1,200)),
```

```
cbind(rpois(300,10),rpois(300,10),rpois(300,10),rep(2,300))))
P2<-data.frame(rbind(cbind(rpois(200,10)+5,rpois(200,10)-5,rpois(200,10)+2,rep(1,200)),
cbind(rpois(300,10),rpois(300,20)-10,rpois(300,30)-20,rep(2,300))))

# Cálculo das clasificacións pola regra de Fisher mostral e do erro aparente
rN1<-regraFisher_mostra(N1,3,200)
efN1<-c(efN1, sum(rX!=N1[,4])/5)
rN2<-regraFisher_mostra(N2,3,200)
efN2<-c(efN2, sum(rN2!=N2[,4])/5)
rP1<-regraFisher_mostra(P1,3,200)
efP1<-c(efP1, sum(rP1!=P1[,4])/5)
rP2<-regraFisher_mostra(P2,3,200)
efP2<-c(efP2, sum(rP2!=P2[,4])/5)

# Cálculo das clasificacións pola regra linear normal e do erro aparente
nN1<-lda(N1$X4~.,data=N1,prior=c(0.5,0.5))
nN2<-lda(N2$X4~.,data=N2,prior=c(0.5,0.5))
nP1<-lda(P1$X4~.,data=P1,prior=c(0.5,0.5))
nP2<-lda(P2$X4~.,data=P2,prior=c(0.5,0.5))

enN1<-c(enN1,sum(N1[,4]!=predict(nN1)$class)/5)
enN2<-c(enN2,sum(N2[,4]!=predict(nN2)$class)/5)
enP1<-c(enP1,sum(P1[,4]!=predict(nP1)$class)/5)
enP2<-c(enP2,sum(P2[,4]!=predict(nP2)$class)/5)

# Cálculo das clasificacións pola regra cuadrática e do erro aparente
qN1<-qda(N1$X4~.,data=N1,prior=c(0.5,0.5))
qN2<-qda(N2$X4~.,data=N2,prior=c(0.5,0.5))
qP1<-qda(P1$X4~.,data=P1,prior=c(0.5,0.5))
qP2<-qda(P2$X4~.,data=P2,prior=c(0.5,0.5))

eqN1<-c(eqN1,sum(N1[,4]!=predict(qN1)$class)/5)
eqN2<-c(eqN2,sum(N2[,4]!=predict(qN2)$class)/5)
eqP1<-c(eqP1,sum(P1[,4]!=predict(qP1)$class)/5)
eqP2<-c(eqP2,sum(P2[,4]!=predict(qP2)$class)/5)
}
```

```
# Media e desviación típica dos erros nas 100 repeticións
erros<-list(efN1,efN2,efP1,efP2,enN1,enN2,enP1,enP2,eqN1,eqN2,eqP1,eqP2)
em<-rapply(erros,mean)
edt<-rapply(erros,sd)
```

Na ilustración da regra dos k veciños máis cercanos, a gráfica da Figura 2.3 obtense co seguinte código. A función `knn` é da librería `class` e é necesario fixar unha semente porque cando hai empate nas votacións entre os k veciños máis cercanos, a asignación a unha clase ou outra faina ao azar.

```
set.seed(1997)
ek<-c()
erros<-list()
for(j in 1:50){
rX<-c()
for(i in 1:150){
rX<-c(rX,knn(iris[-i,1:4],iris[i,1:4],iris[-i,5],k=j))
}
ek<-c(ek,sum(rX!=as.integer(iris[,5])))
erros[[j]]<-which(rX!=as.integer(iris[,5]))
}
plot(1:50,ek,xlab='',ylab='',type='o')
```

A representación das distintas funcións kernel da Figura 2.4 é resultado destas liñas.

```
par(mfrow=c(1,3))
x<-seq(-3,3,by=0.01)
plot(x,exp(-x^2),main='Kernel gaussiano',type='l',ylim=c(0,1),xlab='',ylab='')
plot(x,1/(1+x^2),main='Kernel de Cauchy',type='l',ylim=c(0,1),xlab='',ylab='')
plot(x, (1-x^2)*(abs(x)<=1),main='Kernel de Epanechnikov',type='l',ylim=c(0,1),xlab='',ylab
```

E a táboa do exemplo dos métodos kernel obtense co seguinte código.

```
mu1<-c(0,0,0)
mu2<-c(2,0,-0.5)
E1<-cbind(c(1,0,0),c(0,0.25,0),c(0,0,0.25))
```

```
E2<-cbind(c(0.5,0.5,0.125),c(0.5,1,0),c(0.125,0,0.25))
set.seed(1997)
N1<-rbind(mvrnorm(200,mu1,E1),mvrnorm(300,mu2,E1))
N2<-rbind(mvrnorm(200,mu1,E1),mvrnorm(300,mu2,E2))
P1<-rbind(cbind(rpois(200,10)+5,rpois(200,10)-5,rpois(200,10)+2),
cbind(rpois(300,10),rpois(300,10),rpois(300,10)))
P2<-rbind(cbind(rpois(200,10)+5,rpois(200,10)-5,rpois(200,10)+2),
cbind(rpois(300,10),rpois(300,20)-10,rpois(300,30)-20))
y=factor(c(rep(1,200),rep(2,300)))

N1.norm<-classif.np(y,N1,Ker =Ker.norm,metric=metric.dist)
N1.epa<-classif.np(y,N1,Ker =Ker.epa,metric=metric.dist,h=seq(1.22,2,by=0.02))
N1.uni<-classif.np(y,N1,Ker =Ker.unif,metric=metric.dist,h=seq(1.22,2,by=0.02))
N1.norm$prob.classification
N1.epa$prob.classification
N1.uni$prob.classification

N2.norm<-classif.np(y,N2,Ker =Ker.norm)
N2.epa<-classif.np(y,N2,Ker =Ker.epa,metric=metric.dist,h=seq(1.1,2,by=0.02))
N2.uni<-classif.np(y,N2,Ker =Ker.unif,metric=metric.dist,h=seq(1.1,2,by=0.02))
N2.norm$prob.classification
N2.epa$prob.classification
N2.uni$prob.classification

P1.norm<-classif.np(y,P1,Ker =Ker.norm)
P1.epa<-classif.np(y,P1,Ker =Ker.epa,metric=metric.dist,h=seq(6,10,by=0.02))
P1.uni<-classif.np(y,P1,Ker =Ker.unif,metric=metric.dist,h=seq(7,10,by=0.02))
P1.norm$prob.classification
P1.epa$prob.classification
P1.uni$prob.classification

P2.norm<-classif.np(y,P2,Ker =Ker.norm)
P2.epa<-classif.np(y,P2,Ker =Ker.epa,metric=metric.dist,h=seq(8,10,by=0.02))
P2.uni<-classif.np(y,P2,Ker =Ker.unif,metric=metric.dist,h=seq(8,10,by=0.02))
P2.norm$prob.classification
P2.epa$prob.classification
```

```
P2.uni$prob.classification
```

A Figura 4.2 é resultado do que segue.

```
set.seed(1997)
mu1<-c(0,0); mu2<-c(0,2)
E<-matrix(c(2,1,1,0.75),2,2)
N<-rbind(mvrnorm(100,mu1,E),mvrnorm(100,mu2,E))
PC1<-eigen(E)$vector[,1]
prox<-N[,1]*PC1[1]+N[,2]*PC1[2]

par(mfrow=c(1,2))
plot(N,col=c(rep(1,100),rep(2,100)),xlab='',ylab='')
plot(prox,0.03*rnorm(200),xlab='',ylab='',col=c(rep(1,100),rep(2,100)),ylim=c(-1,1))
```

A.2. Implementación do HDDA e aplicación a datos simulados e reais

As seguintes liñas definen unha función de R que calcula as clases ás que se asigna unha mostra de entrenamento e outra de testeo cos métodos HDDAi e HDDAh a partir da primeira, e tamén os erros de clasificación correspondentes. Esta foi a función empregada para obter os resultados mostrados no último capítulo.

```
hdda<-function(train,test,ctrain,ctest,k,d,dnu){

alfa<-sigma<-c()
mu<-S<-list()
lambda<-Q<-list()
classi<-classh<-ei<-eh<-list()
classi$train<-classi$test<-classh$train<-classh$test<-1

# Cálculo dos estimadores dos parámetros para cada unha das k clases
for(i in 1:k){
ni<-sum(ctrain==i)
mu[[i]]<-colMeans(train[ctrain==i,])
S[[i]]<-(ni-1)/ni*var(train[ctrain==i,])
```

```
#Autovalores e autovectores de S
lambda[[i]]<-eigen(S[[i]])$values
Qnu<-eigen(S[[i]])$vectors
# Matriz das proxeccións
Q[[i]]<-cbind(Qnu[,1:dnu],matrix(0,d,d-dnu))
#Parámetros do HDDA
a<-sum(lambda[[i]][1:dnu])/dnu
b<-sum(lambda[[i]][(dnu+1):d])/(d-dnu)
alfa[i]<-b/(a+b)
sigma[i]<-a*b/(a+b)
}
alfa<-mean(alfa)

#Clasificación da mostra de entrenamento con HDDAi e HDDAh
HDDAi<-HDDAh<-c()
for(j in 1:length(ctrain)){
for(i in 1:k){
prox<-Q[[i]]%*%t(Q[[i]])%*%matrix(as.matrix(train[j,]-mu[[i]]),d,1) +mu[[i]]
HDDAi[i]<-alfa*sum((mu[[i]]- prox)^2)+(1-alfa)*sum((train[j,]-prox)^2)
HDDAh[i]<-1/sigma[i]*HDDAi[i]+2*d*log(sigma[i])
}
classi$train[j]<-which(HDDAi==min(HDDAi))
classh$train[j]<-which(HDDAh==min(HDDAh))
}
ei$train<-sum(classi$train!=ctrain)/length(ctrain)
eh$train<-sum(classh$train!=ctrain)/length(ctrain)

#Clasificación da mostra de testeo con HDDAi e HDDAh
HDDAi<-HDDAh<-c()
for(j in 1:length(ctest)){
for(i in 1:k){
prox<-Q[[i]]%*%t(Q[[i]])%*%matrix(as.matrix(test[j,]-mu[[i]]),d,1)+mu[[i]]
HDDAi[i]<-alfa*sum((mu[[i]]- prox)^2)+(1-alfa)*sum((test[j,]-prox)^2)
HDDAh[i]<-1/sigma[i]*HDDAi[i]+2*d*log(sigma[i])
}
classi$test[j]<-which(HDDAi==min(HDDAi))
```

```
classh$test[j]<-which(HDDAh==min(HDDAh))
}
ei$test<-sum(classi$test!=ctest)/length(ctest)
eh$test<-sum(classh$test!=ctest)/length(ctest)

return(list(classi=classi,classh=classh,ei=ei,eh=eh))
}
```

A continuación móstrase a simulación dos datos e os cálculos feitos sobre eles. Están tamén as liñas que xeran a Figura 5.1.

```
set.seed(1997)
dnu<-1:20
d<-100; train<-25; test<-100
mu1<-c(0,0.9); mu2<-c(0,-0.9); E<-matrix(c(1,0.7,0.7,1),2,2)
cero<-rep(0,d-2); I<-diag(d-2)
ctrain<-c(rep(1,train),rep(2,train)); ctest<-c(rep(1,test),rep(2,test))
err_test_l<-err_train_l<-err_test_h<-err_train_h<-err_test_i<-err_train_i<-matrix(0,50,20)
for(veces in 1:50){
Xtrain<-rbind(cbind( mvrnorm(train,mu1,E), mvrnorm(train,cero,I)),
cbind( mvrnorm(train,mu2,E), mvrnorm(train,cero,I)))
Xtest<-rbind(cbind( mvrnorm(test,mu1,E), mvrnorm(test,cero,I)),
cbind( mvrnorm(test,mu2,E), mvrnorm(test,cero,I)))
for(j in dnu){
res<-hdda(Xtrain,Xtest,ctrain,ctest,2,d,j)
#Almacenamento dos erros para promediar
err_train_i[veces,j]<-res$ei$train
err_train_h[veces,j]<-res$eh$train
err_test_i[veces,j]<-res$ei$test
err_test_h[veces,j]<-res$eh$test
#Erros da regra linear
l<-lda(Xtrain,ctrain)
err_train_l[veces,j]<-sum(predict(l)$class!=ctrain)/(2*train)
err_test_l[veces,j]<-sum(predict(l,Xtest)$class!=ctest)/(2*test)
}
}
plot(dnu,colMeans(err_test_i),col=3,type='o',ylim=c(0,0.5),xlab='',ylab='')
```

```
lines(dnu,colMeans(err_train_i),col=1,type='o')
lines(dnu,colMeans(err_train_h),col=2,type='o')
lines(dnu,colMeans(err_test_h),col=4,type='o')
abline(h=colMeans(err_train_l),col=5)
abline(h=colMeans(err_test_l),col=6)
```

Por último, o código empregado para tratar os datos de expresión xénica de Alon et al. (1999), que están dispoñibles no paquete `dprep`, usando o método HDDA de Bouveyron et al. (2007).

```
data(colon)
dnu<-1:20

HDDAi<-HDDAh<-list()
HDDAi$meantrain<-HDDAi$meantest<-HDDAi$sdtrain<-HDDAi$sdtest<-1
HDDAh$meantrain<-HDDAh$meantest<-HDDAh$sdtrain<-HDDAh$sdtest<-1
for(k in dnu){
# Clasificación tomando menos variables, repetida moitas veces
eHDDAitrain<-eHDDAitest<-eHDDAhtrain<-eHDDAhtest<-matrix(0,50,20)
for(j in 1:20){
for(i in 1:50){
# Subconxuntos de 50 expresións xénicas extraídos ao azar
datos<-colon[,c(sample(1:2000,50),2001)]
# 2/3 dos datos para entrenamento e 1/3 para testeo
part<-sample(1:62,20)
train<-datos[-part,1:50]; ctrain<-datos[-part,51]
test<-datos[part,1:50]; ctest<-datos[part,51]
# Chamada á función e resultados de aplicar HDDAi e HDDAh
res<-hdda(train,test,ctrain,ctest,2,50,dnu)
eHDDAitrain[i,j]<-res$ei$train
eHDDAitest[i,j]<-res$ei$test
eHDDAhtrain[i,j]<-res$eh$train
eHDDAhtest[i,j]<-res$eh$test
}
}
# Almacenamento dos resultados para cada dimensión intrínseca considerada
eHDDAitrain<-colMeans(eHDDAitrain)
```

```
HDDAi$meantrain[k]<-mean(eHDDAitrain); HDDAi$sdtrain[k]<-sd(eHDDAitrain)
eHDDAitest<-colMeans(eHDDAitest)
HDDAi$meantest[k]<-mean(eHDDAitest); HDDAi$sdtest[k]<-sd(eHDDAitest)
eHDDAhtrain<-colMeans(eHDDAhtrain)
HDDAh$meantrain[k]<-mean(eHDDAhtrain); HDDAh$sdtrain[k]<-sd(eHDDAhtrain)
eHDDAhtest<-colMeans(eHDDAhtest)
HDDAh$meantest[k]<-mean(eHDDAhtest); HDDAh$sdtest[k]<-sd(eHDDAhtest)
}
# Gráfica do erro promedio cometido sobre as mostras usando HDDA
plot(dnu,HDDAi$meantrain,col=1,type='o',ylim=c(0,0.5),xlab='',ylab='')
lines(dnu,HDDAi$meantest,col=2,type='o')
lines(dnu,HDDAh$meantrain,col=3,type='o')
lines(dnu,HDDAh$meantest,col=4,type='o')
```

Bibliografía

[1] Koch, I., *Analysis of Multivariate and High-Dimensional Data*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambride University Press, 2014.

[2] Devroye, L., Györfi, L. e Lugosi, G., *A Probabilistic Theory of Pattern Recognition*, Applications of Mathematics, Stochastic Modelling and Applied Probability, 31, Springer-Verlag, New York, 1996.

[3] James, G., Witten, D., Hastie, T. e Tibshirani, R., *An Introduction to Statistical Learning*, Springer Texts in Statistics, 103, Springer-Verlag, New York, 2013.

[4] Fix, E. e Hodges J., *Discrminatory analysis, nonparametric discrimination: Consistency properties*, Tecnical report, Randolph Field, TX, USAF School of Aviation Medicine, 1951.

[5] Hastie, T., Tibshirani, R. e Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlag, New York, 2001.

[6] Galeano, P. e Peña, D., *Data science, big data and statistics*, TEST, 28: 289, 2019.

[7] Klamelä, J., *Multivariate Nonparametric Regression and Visualization: with R and applications to finance*, Wiley Series in Computational Statistics, Wiley, 2014.

[8] Mardia, K., Kent, J. e Bibby, J., *Multivariate Analysis*, Probability and Mathematical Statistics, Academic Press, London, 1979.

[9] Bouveyron, C., Girard, S. e Schmid, C., *High Dimensional Discriminant Analysis*, Communication in Statistics-Theory and Methods, 36, 2007.

[10] Qiao, Z., Zhou, L. e Huang, J. Z., *Effective Linear Discriminant Analysis for High Dimensional, Low Sample Size Data*, Proceedings of the World Congress on Engineering 2008 Vol II WCE 2008, July 2 - 4, 2008, London, U.K.

[11] Friedman, J.H., *Regularized Discriminant Analysis*, Journal of the American Statistical Association, 84, 1989.

[12] Qin, Y., *A review of quadratic discriminant analysis for high-dimensional data*, Wiley Interdisciplinary Reviews: Computational Statistics. 10. e1434. 10.1002/wics.1434., 2018.

[13] Tibshirani, R., *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, 1996

[14] Alon, U., Barkai, N. et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proceedings of the National Academy of Sciences of the United States of America, 96 (12), 1999.

[15] Cai, T. e Zhang, L, *High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data*, Journal of the Royal Statistical Society , Series B, Statistical Methodology, 2019.

[16] Manuel Febrero-Bande, Manuel Oviedo de la Fuente, *Statistical Computing in Functional Data Analysis: The R Package fda.usc.* Journal of Statistical Software, 51(4), 1-28, 2012.

[17] Venables, W. N. e Ripley, B. D. *Modern Applied Statistics with S*, Fourth Edition. Springer, New York, 2002.