Análisis Discriminante Lineal y Cuadrático: Tutorial

Benyamin Ghojogh BGHOJOGH@UWATERLOO.California
Departamento de Ingeniería Eléctrica e Informática,
Laboratorio de aprendizaje automático, Universidad de Waterloo, Waterloo, ON, Canadá

marca crowley MCROWLEY@UWATERLOO.California
Departamento de Ingeniería Eléctrica e Informática,
Laboratorio de aprendizaje automático, Universidad de Waterloo, Waterloo, ON, Canadá

Abstracto

Este tutorial explica el análisis discriminante lineal (LDA) y el análisis discriminante cuadrático (QDA) como dos métodos de clasificación fundamentales en el aprendizaje estadístico y probabilístico. Comenzamos con la optimización del límite de decisión en el que los posteriores son iguales. Luego, se derivan LDA y QDA para clases binarias y múltiples. También se cubre la estimación de parámetros en LDA y QDA. Luego, explicamos cómo LDA y QDA están relacionados con el aprendizaje de métricas, el análisis de componentes principales del núcleo, la distancia de Mahalanobis, la regresión logística, el clasificador óptimo de Bayes, el Bayes ingenuo de Gauss y la prueba de razón de verosimilitud. También probamos que LDA y el análisis discriminante de Fisher son equivalentes. Finalmente aclaramos algunos de los conceptos teóricos con simulaciones que proporcionamos.

1. Introducción

Supongamos que tenemos un conjunto de datos de *instancias* $\{(x_i, y_i)\}_{i=1}^{n}$ con tamaño de la muestra n y dimensionalidad $x_i \in \mathbb{R}^d$ y $y_i \in \mathbb{R}$. El y_i's son las etiquetas de clase. Nos gustaría *clasificar* el espacio de datos utilizando estas instancias. Análisis Discriminante Lineal (LDA) y Análisis Discriminante Cuadrático (QDA) (Friedman et al., 2009) son dos conocidas *clasificación supervisada* métodos en el aprendizaje estadístico y probabilístico. Este documento es un tutorial para estos dos clasificadores donde se detalla la teoría de la clasificación binaria y multiclase. Luego, se explican las relaciones de LDA y QDA con el aprendizaje métrico, el análisis de componentes principales (PCA) del núcleo, el análisis discriminante de Fisher (FDA), la regresión logística, el clasificador óptimo de Bayes, el bayesiano ingenuo gaussiano y la prueba de razón de verosimilitud (LRT) para una mejor comprensión de estos dos métodos fundamentales. Finalmente, se reportan y analizan algunos experimentos en conjuntos de datos sintéticos para ilustración.

2. Optimización para el Límite de Clases

Primero suponga que los datos son unidimensionales, $x \in \mathbb{R}$. Supongamos que tenemos dos clases con las funciones de distribución acumulativa (CDF) $F_1(x)$ y $F_2(x)$, respectivamente. Sean las funciones de densidad de probabilidad (PDF) de estas CDF:

$$F_1(x) = \frac{\partial F_1(x)}{\partial x}, \tag{1}$$

$$F_2(x) = \frac{\partial F_2(x)}{\partial x}, \tag{2}$$

respectivamente.

Suponemos que las dos clases tienen una distribución normal (gaussiana), que es la distribución predeterminada más común en las aplicaciones del mundo real. La media de una de las dos clases es mayor que la otra; asumimos $\mu_1 < \mu_2$. Una instancia $x \in \mathbb{R}$ pertenece a una de estas dos clases:

$$x \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2), & \text{si } x \in \mathcal{C}_1, \\ \mathcal{N}(\mu_2, \sigma_2^2), & \text{si } x \in \mathcal{C}_2, \end{cases} \tag{3}$$

donde $\mathcal{C}_1$ y $\mathcal{C}_2$ denotan la primera y segunda clase, respectivamente.

por una instancia x, podemos tener un error en la estimación de la clase a la que pertenece. En un punto, que denotaremos por x_*, la probabilidad de que las dos clases sean iguales; por lo tanto, el punto x_* está en el límite de las dos clases. como tenemos $\mu_1 < \mu_2$, podemos decir $\mu_1 < x_* < \mu_2$ como se muestra en la figura. 1. Por lo tanto, si $x < x_*$ o $x > x_*$ la instancia x pertenece a la primera y segunda clase, respectivamente. Por lo tanto, estimar $x < x_*$ o $x > x_*$ pertenecer respectivamente a la segunda y primera clase es un error en la estimación de la clase. Esta probabilidad de error puede expresarse como:

$$\text{PAG(error)} = \text{PAG}(x > x_*, x \in \mathcal{C}_1) + \text{PAG}(x < x_*, x \in \mathcal{C}_2). \tag{4}$$

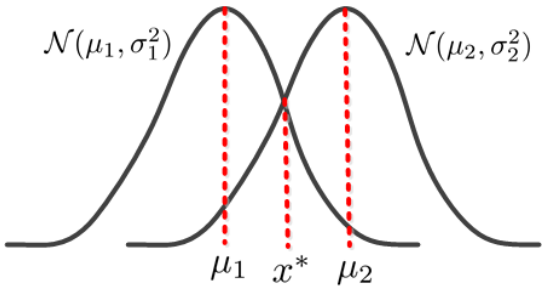

*Figura 1.*Dos funciones de densidad gaussianas donde son iguales en el puntoX_*.

como tenemosPAG(A,B) =PAG(A/B)PAG(B), podemos decir:

$$\begin{aligned}\text{PAG(error)} &= \text{PAG}(x > x_* / X \in C_1)\text{PAG}(X \in C_1) \\ &+ \text{PAG}(x < x_* / X \in C_2)\text{PAG}(X \in C_2),\end{aligned} \tag{5}$$

que queremos minimizar:

$$\underset{X_*}{\text{minimizar}} \text{PAG(error)},$$

encontrando el mejor límite de clases, es decir, X_*. Según la definición de CDF, tenemos:

$$\text{PAG}(x < c,\ x \in C_1) = F_1(C),$$

$$\Rightarrow \text{PAG}(x > x_*,\ X \in C_1) = 1 - F_1(X_*), \text{ PAG}(\tag{7}$$

$$x < x_*,\ X \in C_2) = F_2(X_*). \tag{8}$$

Según la definición de PDF, tenemos:

$$\text{PAG}(X \in C_1) = F_1(X) = \pi_1,$$

$$\text{PAG}(X \in C_2) = F_2(X) = \pi_2,$$

donde denotamos los anteriores$F_1(X)$y$F_2(X)$porπ_1yπ_2, respectivamente.

Por lo tanto, las Ecs. (5) y (6) convertirse en:

$$\underset{X_*}{\text{minimizar}} \left(1 - F_1(X_*)\pi_1 + F_2(X_*)\pi_2\right). \tag{11}$$

Derivamos en aras de la minimización:

$$\frac{\partial \text{PAG(error)}}{\partial x_*} = -F_1(X_*)\pi_1 + F_2(X_*)\pi \overset{\text{se}}{=} 0,$$

$$\Rightarrow F_1(X_*)\pi_1 = F_2(X_*)\pi_2. \tag{12}$$

Otra forma de obtener esta expresión es igualando las probabilidades posteriores de tener la ecuación de la frontera de clases:

$$\text{PAG}(X \in C_1 / X = X) = \overset{\text{colocar}}{=} \text{PAG}(X \in C_2 / X = X). \tag{13}$$

De acuerdo con la regla de Bayes, la*posterior*es:

$$\begin{aligned}\text{PAG}(X \in C_1 / X = X) &= \frac{\text{PAG}(X = x / x \in C_1)\text{PAG}(X \in C_1)}{\text{PAG}(X = X)} \\ &= \frac{F_1(X)\pi_1}{\sum_{k=1}^{/C/} \text{PAG}(X = x / x \in C_k)\pi_k},\end{aligned} \tag{14}$$

dónde$/C/$es el número de clases que es dos aquí. El $F_1(X)$y$\pi_1$son los*probabilidad (clase condicional)*y *previo* probabilidades, respectivamente, y el denominador es la probabilidad marginal.

Por lo tanto, la ecuación. (13) se convierte en:

$$\begin{aligned}&\frac{F_1(X)\pi_1}{\sum_{i=1}^{/C/} \text{PAG}(X = x / x \in C_i)\pi_i} \\ &\overset{\text{colocar}}{=} \frac{F_2(X)\pi_2}{\sum_{i=1}^{/C/} \text{PAG}(X = x / x \in C_i)\pi_i} \\ &\Rightarrow F_1(X)\pi_1 = F_2(X)\pi_2.\end{aligned} \tag{15}$$

Ahora pensemos en los datos como*multivariado*datos con dimensionalidad$d$. El PDF para la distribución gaussiana multivariada (6) ción,$X \sim$ *norte*(μ,Σ)es:

$$F(X) = \sqrt{\frac{1}{(2\pi)^d /\Sigma/}} \text{Exp}\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu))}{2}\right), \tag{dieciséis}$$

(9) donde$X \in \mathbb{R}^d$,$\mu \in \mathbb{R}^d$es la media, $\Sigma \in \mathbb{R}^{d \times d}$es el (10) matriz de covarianza, y$/.$/es el determinante de la matriz. El $\pi \approx 3.14$ en esta ecuación no debe confundirse con la π_k(anterior) en la ecuación. (12) o (15). Por lo tanto, la Ec. (12) o (15) se convierte en:

$$\begin{aligned}&\sqrt{\frac{1}{(2\pi)^d /\Sigma_1/}} \text{Exp}\left(-\frac{(x-\mu)^\top \Sigma_1^{-1}(x-\mu))}{2}\right)\pi_1 \\ &= \sqrt{\frac{1}{(2\pi)^d /\Sigma_2/}} \text{Exp}\left(-\frac{(x-\mu_2)^\top \Sigma_2^{-1}(x-\mu_2))}{2}\right)\pi_2,\end{aligned} \tag{17}$$

donde las distribuciones de primera y segunda clase son *norte* (μ_1,Σ_1)y*norte*(μ_2,Σ_2), respectivamente.

3. Análisis Discriminante Lineal para Binario Clasificación

En el análisis discriminante lineal (LDA), asumimos que las dos clases tienen matrices de covarianza iguales:

$$\Sigma_1 = \Sigma_2 = \Sigma. \tag{18}$$

Por lo tanto, la Ec. (17) se convierte en:

$$\begin{aligned}
&\sqrt{\frac{1}{(2\pi)^d|\Sigma|}}\exp\left(-\frac{(x-\mu)^\top\Sigma_1^{-1}(x-\mu_1)}{2}\right)\pi_1 \\
&=\sqrt{\frac{1}{(2\pi)^d|\Sigma|}}\exp\left(-\frac{(x-\mu)^\top\Sigma_2^{-1}(x-\mu_2)}{2}\right)\pi_2, \\
&\Rightarrow -\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1)+\ln(\pi_1) \\
&\quad = -\frac{1}{2}(x-\mu_2)^\top\Sigma^{-1}(x-\mu_2)+\ln(\pi_2),
\end{aligned}$$

dónde (a)toma el logaritmo natural de los lados de la ecuación.

Podemos simplificar este término como:

$$\begin{aligned}
&(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1) = (x-\mu_1)^\top\Sigma^{-1}(x-\mu_1) \\
&= x^\top\Sigma^{-1}x - x^\top\Sigma^{-1}\mu_1 - \mu_1^\top\Sigma^{-1}x + \mu_1^\top\Sigma^{-1}\mu_1 \\
&\overset{(a)}{=} x^\top\Sigma^{-1}x + \mu_1^\top\Sigma^{-1}\mu_1 - 2\mu_1^\top\Sigma^{-1}x,
\end{aligned} \tag{19}$$

dónde (a)es porque $x^\top\Sigma^{-1}\mu_1 = \mu_1^\top\Sigma^{-1}x$como es un escalar y Σ^{-1}es simétrica entonces $\Sigma^{-\top} = \Sigma^{-1}$. Así, tenemos:

$$\begin{aligned}
&-\frac{1}{2}x^\top\Sigma^{-1}x + \frac{1}{2}\mu_1^\top\Sigma^{-1}\mu_1 + \mu_1^\top\Sigma^{-1}x + \ln(\pi_1) \\
&= -\frac{1}{2}x^\top\Sigma^{-1}x + \frac{1}{2}\mu_2^\top\Sigma^{-1}\mu_2 + \mu_2^\top\Sigma^{-1}x + \ln(\pi_2).
\end{aligned}$$

Por lo tanto, si multiplicamos los lados de la ecuación por2, tenemos:

$$2\left(\Sigma^{-1}(\mu_2-\mu_1)\right)^\top x + \left(\mu_1-\mu_2\right)^\top\Sigma^{-1}(\mu_1-\mu_2) + 2\ln\left(\frac{\pi_2}{\pi_1}\right) = 0,$$

que es la ecuación de una recta en forma de$a^\top x + b = 0$. Por lo tanto, si consideramos distribuciones gaussianas para las dos clases donde se supone que las matrices de covarianza son iguales, el límite de decisión de la clasificación es una línea. Debido a la linealidad del límite de decisión que discrimina las dos clases, este método se denomina*discriminante lineal*an álisis.

Para obtener la Ec. (20), trajimos al lado derecho las expresiones que correspondían a la segunda clase; por lo tanto, si usamos$d(x) : \mathbb{R}^d \to \mathbb{R}$como la expresión del lado izquierdo (función) en la ecuación. (20):

$$d(x) := 2\left(\Sigma^{-1}(\mu_2 - \mu_1)\right)^\top x + \left(\mu_1-\mu_2\right)^\top\Sigma^{-1}(\mu_1-\mu_2) + 2\ln\left(\frac{\pi_2}{\pi_1}\right), \tag{21}$$

la clase de una instancia$x$se estima como:

$$\hat{C}(x) = \begin{cases} 1, \text{si } d(x) < 0, \\ 2, \text{si } d(x) > 0. \end{cases} \tag{22}$$

Si los priores de dos clases son iguales, es decir,$\pi_1 = \pi_2$, la ecuación. (20) se convierte en:

$$2\left(\Sigma^{-1}(\mu_2-\mu_1)\right)^\top x + \left(\mu_1-\mu_2\right)^\top\Sigma^{-1}(\mu_1-\mu_2) = 0, \tag{23}$$

cuya expresión del lado izquierdo se puede considerar como$d(x)$en la ecuación (22).

4. Análisis Discriminante Cuadrático para Clasificación Binaria

En el Análisis Discriminante Cuadrático (QDA), relajamos el supuesto de igualdad de las matrices de covarianza:

$$\Sigma_1 \neq \Sigma_2, \tag{24}$$

lo que significa que las covarianzas no son*necesariamente* iguales (si son realmente iguales, el límite de decisión será lineal y QDA se reduce a LDA).

Por lo tanto, la Ec. (17) se convierte en:

$$\begin{aligned}
&\sqrt{\frac{1}{(2\pi)^d|\Sigma_1|}}\exp\left(-\frac{(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)}{2}\right)\pi_1 \\
&=\sqrt{\frac{1}{(2\pi)^d|\Sigma_2|}}\exp\left(-\frac{(x-\mu)^\top\Sigma_2^{-1}(x-\mu)}{2}\right)\pi_2, \\
&\Rightarrow -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_1|) \\
&\quad -\frac{1}{2}(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) + \ln(\pi_1) \\
&\quad = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_2|) \\
&\quad -\frac{1}{2}(x-\mu_2)^\top\Sigma_2^{-1}(x-\mu_2) + \ln(\pi_2),
\end{aligned}$$

dónde (a)toma el logaritmo natural de los lados de la ecuación (20) ción. De acuerdo con la Ec. (19), tenemos:

$$\begin{aligned}
&-\frac{1}{2}\ln(|\Sigma_1|) - \frac{1}{2}x^\top\Sigma_1^{-1}x - \frac{1}{2}\mu_1^\top\Sigma_1^{-1}\mu_1 \\
&+ \mu_1^\top\Sigma_1^{-1}x + \ln(\pi_1) \\
&= -\frac{1}{2}\ln(|\Sigma_2|) - \frac{1}{2}x^\top\Sigma_2^{-1}x - \frac{1}{2}\mu_2^\top\Sigma_2^{-1}\mu_2 \\
&+ \mu_2^\top\Sigma_2^{-1}x + \ln(\pi_2).
\end{aligned}$$

Por lo tanto, si multiplicamos los lados de la ecuación por2, tenemos:

$$\begin{aligned}
&x^\top(\Sigma_1^{-1}-\Sigma_2^{-1})x + 2\left(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1\right)^\top x \\
&+ (\mu_1^\top\Sigma_1^{-1}\mu_1 - \mu_2^\top\Sigma_2^{-1}\mu_2) + \ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) \\
&+ 2\ln\left(\frac{\pi_2}{\pi_1}\right) = 0,
\end{aligned} \tag{25}$$

que está en la forma cuadrática $X^{\top}Hacha+b^{\top}X+C=0$. Por lo tanto, si consideramos distribuciones gaussianas para las dos clases, el límite de decisión de la clasificación es cuadrático. Debido al límite de decisión cuadrático que discrimina las dos clases, este método se denomina *discriminante cuadrático* análisis.

Para obtener la Ec. (25), trajimos al lado derecho las expresiones que correspondían a la segunda clase; por lo tanto, si usamos $d(X) : \mathrm{R}^{d} \rightarrow \mathrm{R}$como la expresión del lado izquierdo (función) en la ecuación. (25):

$$d(X) := X^{\top}(\Sigma_1^{-1} - \Sigma_2^{-1})X + 2\,(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1)^{\top}X + (\mu_1^{\top}\Sigma_1^{-1}\mu_1 - \mu_2^{\top}\Sigma_2^{-1}\mu_2) + \ln\frac{(|\Sigma_1|)}{|\Sigma_2|} + 2\ln\left(\frac{\pi}{\pi_1}\right)$$

(26) Eliminamos los términos constantes $-(1/2)\ln(|\Sigma|)$y $-(1/2)X^{\top}\Sigma^{-1}X$que son los mismos para todas las clases (nótese que antes de tomar el logaritmo, el término $-(1/2)\ln(|\Sigma|)$ se multiplica y el término $-(1/2)X^{\top}\Sigma^{-1}X$se multiplica como un término exponencial). Por lo tanto, la escala posterior de la k-ésima clase se convierte en:

$$d_k(X) := \mu_k^{\top}\Sigma^{-1}x - \frac{1}{2}\mu_k^{\top}\Sigma^{-1}\mu_k + \ln(\pi_k). \qquad (31)$$

En LDA, la clase de la instancia Xestá determinada por la ecuación. (29), dónde $d(X)$es la ecuación (31), porque maximiza el posterior de esa clase.

En conclusión, QDA y LDA tratan de maximizar la *posterior* de clases pero trabajar con el *probabilidades (clase condicional) y anteriores*.

la clase de una instancia Xse estima como la Ec. (22).

Si los priores de dos clases son iguales, es decir, $\pi_1 = \pi_2$, la ecuación. (20) se convierte en:

$$X^{\top}(\Sigma_1^{-1} - \Sigma_2^{-1})X + 2\,(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1)^{\top}X + (\mu_1^{\top}\Sigma_1^{-1}\mu_1 - \mu_2^{\top}\Sigma_2^{-1}\mu_2) + \ln\frac{(|\Sigma_1|)}{|\Sigma_2|} = 0, \qquad (27)$$

cuya expresión del lado izquierdo se puede considerar como $d(X)$ en la ecuación (22).

5. LDA y QDA para clasificación multiclase

Ahora consideramos varias clases, que pueden ser más de dos, indexadas por $k \in \{1, \ldots, |C|\}$. Recuerde la ecuación. (12) o (15) donde estamos usando la escala posterior, es decir, $F_k(X)\pi_k$. De acuerdo con la Ec. (dieciséis), tenemos:

$$F_k(X)\pi_k = \sqrt{\frac{1}{(2\pi)^d|\Sigma_k|}}\,\mathrm{Exp}\left(-\frac{(x-\mu)^{\top}\Sigma_k^{-1}(x-\mu_k)}{2}\right)\pi_k.$$

Tomando logaritmo natural da:

$$\ln(F_k(X)\pi_k) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_k|) - \frac{1}{2}(x-\mu_k)^{\top}\Sigma_k^{-1}(x-\mu_k) + \ln(\pi_k).$$

Eliminamos el término constante $-(d/2)\ln(2\pi)$que es igual para todas las clases (nótese que este término se multiplica antes de sacar el logaritmo). Por lo tanto, la escala posterior de la k-ésima clase se convierte en:

$$d_k(X) := -\frac{1}{2}\ln(|\Sigma_k|) - \frac{1}{2}(x-\mu)^{\top}\Sigma_k^{-1}(x-\mu_k) + \ln(\pi_k).$$

En QDA, la clase de la instancia Xse estima como:

$$C(X) = \underset{k}{\text{argumento máximo}}\, d_k(X), \qquad (29)$$

porque maximiza la parte posterior de esa clase. expresi En esto ón, $d(X)$es la ecuación (28).

En LDA, asumimos que las matrices de covarianza de los k las clases son iguales:

$$\Sigma_1 = \cdots = \Sigma_{|C|} = \Sigma. \qquad (30)$$

Por lo tanto, la Ec. (28) se convierte en:

$$d_k(X) = -\ln(|\Sigma|) - \frac{1}{2}(x-\mu)^{\top}\Sigma_k^{-1}(x-\mu_k) + \ln(\pi_k) = -\frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}X^{\top}\Sigma^{-1}X + \frac{1}{2}X^{\top}\Sigma^{-1}\mu_k + \frac{1}{2}\mu_k^{\top}\Sigma^{-1}X - \frac{1}{2}\mu_k^{\top}\Sigma^{-1}\mu_k + \ln(\pi_k).$$

6. Estimación de Parámetros en LDA y QDA

En LDA y QDA, tenemos varios parámetros que se requieren para calcular los posteriores. Estos parámetros son las medias y las matrices de covarianza de las clases y los priores de las clases.

Los priores de las clases son muy difíciles de calcular. Es algo así como un problema del huevo y la gallina porque queremos conocer las probabilidades de clase (previas) para estimar la clase de una instancia, pero no tenemos las previas y debemos estimarlas. Por lo general, el anterior de la k-ésima clase se estima de acuerdo con el tamaño de la muestra de la k-ésima clase:

$$\hat{\pi_k} = \frac{n_k}{n}, \qquad (32)$$

dónde n_ky $n$son el número de instancias de entrenamiento en el k-ésima clase y en total, respectivamente. Esta estimación considera la distribución de Bernoulli para elegir cada instancia del conjunto de entrenamiento general para estar en el k-ésima clase.

(28) La media de la k-ésima clase se puede estimar usando el Max-Estimación de verosimilitud mínima (MLE), o método de momentos (MOM), para la media de una distribución gaussiana:

$$\hat{\mu_k} = \frac{1}{n_k}\sum_{i=1}^{n} X_i\,\mathbb{1}(C(X_i) = k), \qquad (33)$$

dóndeI(.)es la función indicadora que es uno y cero si su condición se cumple y no se cumple, respectivamente.

En QDA, la matriz de covarianza de la$k$-ésima clase se estima usando MLE:

$$\mathbb{R}^{d\times d}\ni\hat{\Sigma}_k=\frac{1}{norte_k}\sum_{i=1}^{norte}(x_i-\hat{\mu}_k)(X_i-\hat{\mu}_k)^\top \mathbb{I}(C(X_i)=k) \quad (34)$$

O podemos usar el*imparcial*estimación de la matriz de covarianza:

$$\mathbb{R}^{d\times d}\ni\hat{\Sigma}_k=\frac{1}{norte_k-1}\sum_{i=1}^{norte}(x_i-\hat{\mu}_k)(x-\hat{\mu}_k)^\top \mathbb{I}(C(X_i)=k) \quad (35)$$

En LDA, suponemos que las matrices de covarianza de las clases son iguales; por lo tanto, usamos el promedio ponderado de las matrices de covarianza estimadas como la matriz de covarianza común en LDA:

$$\mathbb{R}^{d\times d}\ni\hat{\Sigma}=\sum_{k=1}^{|C|}\frac{norte_k}{\sum_{r=1}^{|C|}norte_r}\hat{\Sigma}_k=\frac{\sum_{k=1}^{|C|}norte_k\hat{\Sigma}_k}{norte} \quad (36)$$

donde los pesos son la cardinalidad de las clases.

7. ¡LDA y QDA son aprendizaje métrico!

Recuerde la ecuación. (28) que es el posterior escalado para el QDA. Primero, suponga que las matrices de covarianza son todas iguales (como tenemos en LDA) y todas son la matriz identidad:

$$\Sigma_1=\cdots=\Sigma_{|C|}=I, \quad (37)$$

lo que significa que se supone que todas las clases están distribuidas esféricamente en el$d$espacio dimensional. Después de esta suposición, la Ec. (28) se convierte en:

$$d_k(X)=-\frac{1}{2}(x-\mu_k)^\top(x-\mu_k)+\ln(\pi_k), \quad (38)$$

porque$|I|=1$, ln(1) = 0,y$I^{-1}=I$. Si asumimos que los anteriores son todos iguales, el términoen(π_k)es constante y se puede descartar:

$$d_k(X)=-\frac{1}{2}(x-\mu_k)^\top(x-\mu_k)=-\frac{1}{2}d_k^2 \quad (39)$$

dónde$d_k$es la distancia euclidiana desde la media de la$k$-ésima clase:

$$d_k=||x-\mu_k||_2=\sqrt{(x-\mu_k)^\top(x-\mu_k)}. \quad (40)$$

Por lo tanto, QDA o LDA se reducen a una simple distancia euclidiana de las medias de las clases si las matrices de covarianza son todas matrices identidad y las priores son iguales. Simple

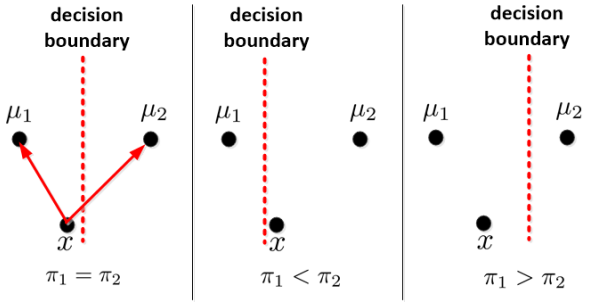

*Figura 2.*El QDA y LDA donde las matrices de covarianza son matrices de identidad. Para anteriores iguales, QDA y LDA se reducen a una clasificación simple utilizando la distancia euclidiana de las medias de las clases. Cambiar el anterior modifica la ubicación del límite de decisión donde incluso un punto puede clasificarse de manera diferente para diferentes anteriores.

La distancia desde la media de las clases es uno de los métodos de clasificación más simples donde la métrica utilizada es la distancia euclidiana.

La ecuación. (39) tiene un mensaje muy interesante. Sabemos que en escala multidimensional métrica (MDS) (cox y cox, 2000) y análisis de componentes principales (PCA) del kernel, tenemos (ver (jamón et al.,2004) y el Capítulo 2 en (Extraño y Zwiggelaar,2014)):

$$k=-\frac{1}{2}HDH, \quad (41)$$

dónde$D\in\mathbb{R}^{norte\times norte}$es la matriz de distancia cuyos elementos son las distancias entre las instancias de datos,$k\in\mathbb{R}^{norte\times norte}$es la matriz del núcleo sobre las instancias de datos,$\mathbb{R}^{norte\times norte}\ni H=I-(1/norte)11^\top$es la matriz de centrado, y$\mathbb{R}^{norte}\ni 1:=[1,1,\ldots,1]^\top$. Si los elementos de la matriz de distancia$D$se obtienen usando la distancia Euclidiana, el MDS es equivalente al An álisis de Componentes Principales (PCA) (Jolliffe,2011).

Comparando Ecs. (39) y (41) muestra una conexión interesante entre la parte posterior de una clase en QDA y el núcleo sobre las instancias de datos de la clase. En esta comparación, la Ec. (41) debe considerarse para una clase y no para todos los datos, por lo que$k\in\mathbb{R}^{norte_k\times norte_k}$,$D\in\mathbb{R}^{norte_k\times norte_k}$,y $H\in\mathbb{R}^{norte_k\times norte_k}$.

Ahora, considere el caso en el que aún las matrices de covarianza son todas matrices de identidad pero las priores no son iguales. En este caso, tenemos la Ec. (38). Si tomamos un exponencial (inverso del logaritmo) de esta expresión, el$\pi_k$se convierte en un factor de escala (peso). Esto significa que todavía estamos usando la métrica de distancia para medir la distancia de una instancia desde la media de las clases, pero estamos escalando las distancias por el antecedentes de clases. Si una clase ocurre más, es decir, su anterior es más grande, debe tener un posterior más grande por lo que reducimos la distancia de la media de su clase. En otras palabras, movemos el límite de decisión de acuerdo con el prior de clases (ver Fig.2).

Como siguiente paso, considere un caso más general donde las matrices de covarianza no son iguales como las que tenemos en QDA. Aplicamos Descomposición de valores singulares (SVD) a la matriz de covarianza de la$k$-ésima clase:

$$\Sigma_k = tu_k \Lambda_k tu_k^\top,$$

donde las matrices izquierda y derecha de vectores singulares son iguales porque la matriz de covarianza es simétrica. Por lo tanto:

$$\Sigma_k^{-1} = tu_k \Lambda_k^{-1} tu_k^\top,$$

dónde$tu_k^{-1} = tu_k^\top$porque es una matriz ortogonal. Por lo tanto, podemos simplificar el siguiente término:

$$\begin{aligned}
&(x-\mu_k)^\top \Sigma_k^{-1} (x-\mu_k) \\
&= (x-\mu_k)^\top tu_k \Lambda_k^{-1} tu_k^\top (x-\mu_k) \\
&= (tu_k^\top x - tu_k^\top \mu_k)^\top \Lambda_k^{-1} (tu_k^\top x - tu_k^\top \mu_k).
\end{aligned}$$

como$\Lambda_k$es una matriz diagonal con elementos no negativos (porque es covarianza), podemos descomponerlo como:

$$\Lambda_k^{-1} = \Lambda_k^{-1/2} \Lambda_k^{-1/2}.$$

Por lo tanto:

$$\begin{aligned}
&(tu_k^\top x - tu_k^\top \mu_k)^\top \Lambda_k^{-1} (tu_k^\top x - tu_k^\top \mu_k) \\
&= (tu_k^\top x - tu_k^\top \mu_k)^\top \Lambda_k^{-1/2} \Lambda_k^{-1/2} (tu_k^\top x - tu_k^\top \mu_k) \\
&\overset{(a)}{=} (\Lambda_k^{-1/2} tu_k^\top x - \Lambda_k^{-1/2} tu_k^\top \mu_k)^\top \\
&\quad (\Lambda_k^{-1/2} tu_k^\top x - \Lambda_k^{-1/2} tu_k^\top \mu_k),
\end{aligned}$$

dónde(a)es porque$\Lambda_k^\top = \Lambda_k^{-1/2}$porque es diagonal Definimos la siguiente transformación:

$$\varphi_k : X \mapsto \Lambda_k^{-1/2} tu_k^\top X, \tag{42}$$

lo que también resulta en la transformación de la media:$\varphi_k : \mu \mapsto \Lambda_k^{-1/2} tu_k \mu$. Por lo tanto, la Ec. (28) se puede reformular como:

$$\begin{aligned}
d_k(X) &= -\frac{1}{2} \log(|\Sigma_k|) \\
&\quad - \frac{1}{2} (\varphi_k(X) - \varphi_k(\mu_k))^\top (\varphi_k(X) - \varphi_k(\mu_k)) + \log(\pi_k).
\end{aligned} \tag{43}$$

Ignorando los términos$-(1/2) \log(|\Sigma_k|)$y$\log(\pi_k)$, podemos ver que la transformación ha cambiado la matriz de covarianza de la clase a matriz identidad. Por lo tanto, el QDA (y también el LDA) puede verse como una simple comparación de distancias desde las medias de las clases después de aplicar una transformación a los datos de cada clase. En otras palabras, estamos aprendiendo la métrica usando el SVD de la matriz de covarianza de cada clase. Por lo tanto, LDA y QDA pueden verse como*aprendizaje métrico*(Yang y jin,2006;Kulís,2013) en perspectiva. Tenga en cuenta que en el aprendizaje de métricas, una métrica de distancia válida se define como (yang y jin,2006):

$$d_A^2(x, \mu_k) := ||x-\mu_k||_A^2 = (x-\mu_k)^\top A (x-\mu_k), \tag{44}$$

dónde$A$es una matriz semidefinida positiva, es decir,$A \succeq 0$. En QDA, también estamos usando$(x-\mu_k)^\top \Sigma_k^{-1} (x-\mu_k)$. El La matriz de covarianza es semidefinida positiva según las características de la matriz de covarianza. Además, según las características de una matriz semidefinida positiva, la inversa de una matriz semidefinida positiva es semidefinida positiva por lo que$\Sigma_k^{-1} \succeq 0$.Por lo tanto, QDA está usando métrica aprendizaje (y como se discutirá en la siguiente sección, puede verse como un*aprendizaje múltiple*método también).

También es digno de mención que QDA y LDA también pueden verse como*distancia de Mahalanobis*(McLachlan,1999; De Maesschalck et al.,2000) que también es una métrica:

$$d_{METRO}^2(x, \mu) := ||x-\mu||_{METRO}^2 = (x-\mu)^\top \Sigma^{-1} (x-\mu), \tag{45}$$

dondeΣes la matriz de covarianza de la nube de datos cuya media esμ. La intuición de la distancia de Mahalanobis es que si tenemos varias nubes de datos (p. ej., clases), la distancia de la clase con mayor variación debe reducirse porque esa clase ocupa más espacio, por lo que es más probable que suceda. La reducción de escala se muestra a la inversa.
de matriz de covarianza. Comparando$(x-\mu_k)^\top \Sigma_k^{-1} (x-\mu_k)$en QDA o LDA con Eq. (45) muestra que QDA y LDA están usando la distancia de Mahalanobis.

8. LDA≡?FDA

En la sección anterior, vimos que LDA y QDA pueden verse como aprendizaje de métricas. Sabemos que el aprendizaje métrico puede verse como una familia de múltiples métodos de aprendizaje. Explicamos brevemente el por qué de esta afirmación: Como$A \succeq 0$,nosotros puedo decir$A = EE^\top$.Por lo tanto, la ecuación. (44) se convierte en:

$$\begin{aligned}
||x-\mu_k||_A^2 &= (x-\mu_k)^\top EE^\top (x-\mu_k) \\
&= (tu_k^\top x - U_k^\top \mu_k)^\top (tu_k^\top x - U_k^\top \mu_k),
\end{aligned}$$

lo que significa que el aprendizaje métrico puede verse como una comparación de distancias euclidianas simples después de la transformación$\varphi : X \mapsto tu_k^\top X$que es una proyección en un subespacio con matriz de proyeccióntu_k. Por lo tanto, el aprendizaje métrico es un enfoque de aprendizaje múltiple. Esto da una pista de que el Análisis Discriminante de Fisher (FDA) (Pescador,1936; brotando,2005), que es un enfoque de aprendizaje múltiple (Tharwat et al., 2017), podría tener una conexión con LDA; especialmente, porque los nombres FDA y LDA a menudo se usan indistintamente en la literatura. En realidad, otros nombres de FDA son Fisher LDA (FLDA) e incluso LDA.

Sabemos que si proyectamos (transformamos) los datos de una clase usando un vector de proyección $tu \in \mathbb{R}_{pag}$ a un pag subespacio dimensional ($pag \leq d$), es decir:

$$X7 \rightarrow ux^{>},$$

para todas las instancias de datos de la clase, la media y la covariable matriz de ance de la clase se transforman como:

$$\mu 7 \rightarrow tu^{>}\mu, \tag{47}$$
$$\Sigma 7 \rightarrow tu^{>}\Sigma tu, \tag{48}$$

debido a las características de la media y la varianza.

El criterio de Fisher (Xu y Lu,2006) es la relación de varianza entre clases, $\sigma^2_{b,}$ y la varianza dentro de la clase, σ^2_{w}:

$$F := b\ \frac{\sigma^2}{\sigma^2_w} = \frac{(tu^{>}\mu_2 - tu^{>}\mu_1)^2}{tu^{>}\Sigma_2 tu + tu^{>}\Sigma_1 uu^{>}(\Sigma_2 + \Sigma_1)tu} = \frac{(tu^{>}(\mu_2 - \mu_1))_2}{}. \tag{49}$$

La FDA maximiza el criterio de Fisher:

$$\underset{tu}{\text{maximizar}} \quad \frac{(tu^{>}(\mu_2 - \mu_1)^2)}{tu^{>}(\Sigma_2 + \Sigma_1)tu}, \tag{50}$$

que se puede reformular como:

$$\underset{tu}{\text{maximizar}} \quad (tu^{>}(\mu_2 - \mu_1)_2,)$$
$$\text{sujeto a} \quad tu^{>}(\Sigma_2 + \Sigma_1)tu = 1,$$

según el método del cociente de Rayleigh-Ritz (Croot,2005). El lagrangiano (Boyd y Vandenberghe,2004) es:

$$L = tu^{>}(\mu_2 - \mu_1)_2 - \lambda tu^{>}(\Sigma_2 + \Sigma_1)tu - 1,$$

dónde λ es el multiplicador de Lagrange. Igualando la derivada de L a cero da:

$$\frac{\partial L}{\partial tu} = 2(\mu_2 - \mu_1)tu - 2\lambda(\Sigma_2 + \Sigma_1)tu = 0^{\text{colocar}}$$
$$\Rightarrow (\mu_2 - \mu_1)_2 tu = \lambda(\Sigma_2 + \Sigma_1)tu,$$

que es un problema de valor propio generalizado $((\mu_2 - \mu_2),(\Sigma_2 + \Sigma_1))$ de acuerdo a (Ghojogh et al.,2019b).

El vector de proyección es el vector propio de $(\Sigma_2 + \Sigma_1)^{-1}(\mu_2 - \mu_1)_2$; por lo tanto, podemos decir:

$$tu \propto (\Sigma_2 + \Sigma_1)^{-1}(\mu_2 - \mu_2 \quad {}_1).$$

En LDA, se asume la igualdad de matrices de covarianza. Así, según la Ec. (18), podemos decir:

$$tu \propto (2\Sigma)^{-1}(\mu_2 - \mu_1) \quad {}^2 \propto \Sigma^{-1}(\mu_2 - \mu_2 \quad {}_1). \tag{52}$$

De acuerdo con la Ec. (46), tenemos:

$$tu^{>}X \propto \Sigma^{-1}(\mu \quad {}_2 - \mu \ {}_1)^{2^{>}}) \ X. \tag{53}$$

Comparando la Ec. (53) con la ecuación. (23) muestra que (t LDA y FDA son equivalentes hasta un factor de escala $\mu_1 - \mu_2)^{>}\Sigma^{-1}(\mu_1 - \mu_2)$ (tenga en cuenta que este término se multiplica como un factor exponencial antes de tomar el logaritmo para obtener la ecuación. (46) (23), por lo que este término es un factor de escala). Por lo tanto, podemos decir:

$$\text{LDA} \equiv \text{FDA}. \tag{54}$$

En otras palabras, FDA se proyecta en un subespacio. Por otra parte, según el artículo7, LDA puede ser visto como un met- (48) aprendizaje rico con un subespacio donde la distancia euclidiana se utiliza después de proyectar en ese subespacio. *Los dos subespacios de FDA y LDA son el mismo subespacio.* Cabe señalar que en el aprendizaje múltiple (subespacial), la escala no importa porque todas las distancias escalan de manera similar.

Tenga en cuenta que LDA asume *uno* (y no varios) Gaussiano para cada clase y también lo hace la FDA. Por eso la FDA se enfrenta problema para datos multimodales (Sugiyama,2007).

9. Relación con la regresión logística

De acuerdo con las Ecs. (dieciséis) y (32), se utilizan distribuciones gaussianas y de Bernoulli para verosimilitud (clase condicional) y previa, respectivamente, en LDA y QDA. Por lo tanto, estamos haciendo suposiciones para la verosimilitud y la previa, aunque finalmente trabajamos con la posterior en LDA y QDA de acuerdo con la ecuación. (15). *Regresión logística* (Kleinbaum et al.,2002) dice (51) ¿Por qué hacemos suposiciones sobre la probabilidad y los antecedentes cuando queremos trabajar en posterior finalmente. Hagamos una suposición directamente para el posterior.

En la regresión logística, primero se aplica una función lineal a los datos para tener $\beta^{>}X$ dónde $\mathbb{R}^{d+1} 3 X = [X^{>}, 1]^{>}$ y $\beta \in \mathbb{R}^{d+1}$ incluir el intercepto. Luego, se usa la función logística para tener un valor en el rango $(0,1)$ para simular probabilidad. Por lo tanto, en la regresión logística, se supone que el posterior es:

$$\text{PAG}(C(X)/X = X) = \frac{(\text{Exp}(\beta^{>}X))^{C(X)}(}{1 + \exp(\beta^{>}X)} \qquad \frac{1}{1 + \exp(\beta^{>}X)})^{1 - C(X)}, \tag{55}$$

dónde $C(X) \in \{-1, +1\}$ para las dos clases. La regresión logística considera el coeficiente β como parámetro a optimizar y utiliza el método de Newton (Boyd y Vandenberghe,2004) para la optimización. Por lo tanto, en resumen, la regresión logística hace suposiciones sobre el posterior, mientras que LDA y QDA hacen suposiciones sobre la probabilidad y la previa.

10. Relación con el clasificador óptimo bayesiano y el bayesiano ingenuo gaussiano

El clasificador de Bayes maximiza los posteriores de las clases (murphy,2012):

$$C(X) = \text{argumento ma}\underset{k}{} \text{XPAG}(X \in C_k / X = X). \tag{56}$$

De acuerdo con la Ec. (14) y la regla de Bayes, tenemos:

$$\text{PAG}(X\in C_k/X=X)\propto\text{PAG}(X=x/x\in C_k)\underbrace{\text{PAG}(X\in C_k)}_{\pi_k}, \tag{57}$$

donde el denominador de posterior (el marginal) que es:

$$\text{PAG}(X=X)=\sum_{r=1}^{|C|}\text{PAG}(X=x/x\in C_r)\pi_r,$$

se ignora porque no depende de las clases C_1 a $C_{|C|}$.

De acuerdo con la Ec. (57), el posterior se puede escribir en términos de verosimilitud y previo; por lo tanto, la ecuación. (56) puede reformularse como:

$$C(X)=\underset{k}{\text{argumento máximo}}\,\pi_k\text{PAG}(X=x/x\in C_k). \tag{58}$$

Tenga en cuenta que el clasificador de Bayes no hace ninguna suposición sobre el posterior, el anterior y la probabilidad, a diferencia de LDA y QDA que asumen la distribución gaussiana unimodal para la probabilidad (y nosotros *mayo* asumir la distribución de Bernoulli para el anterior en LDA y QDA de acuerdo con la ecuación. (32)). Por lo tanto, podemos decir que la diferencia de Bayes y QDA está en la suposición de *unimodal* Distribución gaussiana para la verosimilitud (clase condicional); por lo tanto, si las probabilidades ya son gaussianas unimodales, el clasificador de Bayes se reduce a QDA. Asimismo, la diferencia de Bayes y LDA está en el supuesto de distribución gaussiana para la probabilidad (clase condicional) y la igualdad de matrices de covarianza de clases; así, si las verosimilitudes ya son gaussianas y las matrices de covarianza ya son iguales, el clasificador de Bayes se reduce a LDA.

Cabe señalar que el clasificador de Bayes es un clasificador óptimo porque puede verse como un conjunto de hipótesis (modelos) en el espacio de hipótesis (modelo) y ningún otro conjunto de hipótesis puede superarlo (consulte el Capítulo 6, página 175 en (Mitchell,1997)). En la literatura se denomina *clasificador óptimo bayesiano*. Para formular mejor las declaraciones explicadas, el clasificador óptimo de Bayes estima la clase como:

$$C(X)=\underset{C_k\in C}{\text{argumento máximo}}\sum_{h_j\in A}\text{PAG}(C_k/h_j)\text{PAG}(D/h_j)\text{PAG}(h_j), \tag{59}$$

dónde $C:=\{C_1,\ldots,C_{|C|}\}$, $D:=\{X_i\}_{i=1}^{norte}$ es el entrenamiento colocar, h_j es una hipótesis para estimar la clase de instancias, y H es el espacio de hipótesis que incluye todas las hipótesis posibles.

De acuerdo con la regla de Bayes, similar a lo que teníamos para la Ec. (57), tenemos:

$$\text{PAG}(h_j/D)\propto\text{PAG}(D/h_j)\text{PAG}(h_j).$$

Por lo tanto, la ecuación. (59) se convierte en (Mitchell,1997):

$$C(X)=\underset{C_k\in C}{\text{argumento máximo}}\sum_{h_j\in A}\text{PAG}(C_k/h_j)\text{PAG}(h_j/D), \tag{60}$$

En conclusión, el clasificador de Bayes es óptimo. Por lo tanto, si las verosimilitudes de las clases son gaussianas, QDA es un clasificador óptimo y si las verosimilitudes son gaussianas y las matrices de covarianza son iguales, el LDA es un clasificador óptimo. A menudo, las distribuciones en la vida natural son gaussianas; especialmente, debido al teorema del límite central (Hazewinkel, 2001), la suma de variables independientes e idénticamente distribuidas (iid) es gaussiana y las señales generalmente se suman en el mundo real. Esto explica por qué LDA y QDA son clasificadores muy efectivos en el aprendizaje automático. También vimos que FDA es equivalente a LDA. Por lo tanto, la razón de la efectividad del poderoso clasificador FDA se vuelve clara.

Hemos visto el desempeño muy exitoso de FDA y LDA en diferentes aplicaciones, como el reconocimiento facial (Belhumeur et al.,1997;Etemad y Chellappa,1997; Zhao et al.,1999), reconocimiento de acción (Ghojogh et al., 2017; Mokari et al.,2018), y clasificación EEG (Malekmohammadi et al.,2019).

La implementación del clasificador de Bayes es difícil en la práctica, por lo que lo aproximamos por *Bayes ingenuo* (zhang,2004). Si X_j denota el j-ésima dimensión (característica) de $X=[X_1,\ldots,X_d]^\top$, ecuación (58) se reformula como:

$$C(X)=\underset{k}{\text{argumento máximo}}\,\pi_k\text{PAG}(X_1,X_2,\ldots,X_d/X\in C_k). \tag{61}$$

El término $\text{PAG}(X_1,X_2,\ldots,X_d/X\in C_k)$ es muy difícil de calcular ya que las características posiblemente estén correlacionadas. Naive Bayes relaja esta posibilidad e ingenuamente asume que las características son condicionalmente independientes ($\perp\!\!\!\perp$) cuando están condicionadas a la clase:

$$\begin{aligned}&\text{PAG}(X_1,X_2,\ldots,X_d/X\in C_k)\\ &\overset{\perp\!\!\!\perp}{=}\text{PAG}(X_1/C_k)\text{PAG}(X_2/C_k)\cdots\text{PAG}(X_d/C_k)=\prod_{j=1}^{d}\text{PAG}(X_j/C_k).\end{aligned}$$

Por lo tanto, la ecuación. (61) se convierte en:

$$C(X)=\underset{k}{\text{argumento máximo}}\,\pi_k\prod_{j=1}^{d}\text{PAG}(X_j/C_k). \tag{62}$$

En *Bayes ingenuo gaussiano*, se supone una distribución gaussiana univariada para la probabilidad (clase condicional) de cada característica:

$$\text{PAG}(X_j/C_k)=\frac{1}{\sqrt{2\pi\sigma_k^2}}\text{Exp}\left(-\frac{(x_j-\mu_k)^2}{2\sigma_k^2}\right), \tag{63}$$

donde la media y la varianza no sesgada se estiman como:

$$\hat{\mu_k} = \frac{1}{n_k}\sum_{i=1}^{n} X_i \mathbb{I}(C(X_i) = k), \quad (64)$$

$$\hat{\sigma}_k^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \hat{\mu_k})^2 \mathbb{I}(C(X_i) = k), \quad (65)$$

dónde $X_{i,j}$ denota el j-ésima característica de la i-ésima instancia de entrenamiento. El anterior se puede estimar nuevamente usando la Ec. (32).

De acuerdo con las Ecs. (62) y (63), Gaussian naive Bayes es equivalente a QDA donde las matrices de covarianza son *diagonal*, es decir, se ignoran las matrices de covarianza fuera de la diagonal. Por lo tanto, podemos decir que QDA es más poderoso que Gaussian Naive Bayes porque Gaussian Naive Bayes es una versión simplificada de QDA. Además, es obvio que Gaussian naive Bayes y QDA son equivalentes para *uno* datos dimensionales. En comparación con LDA, el bayesiano ingenuo gaussiano es equivalente a LDA si las matrices de covarianza son diagonales y son todos iguales, es decir, $\sigma_1 = \cdots = \sigma_C^2$; por lo tanto, LDA y Gaussian Naive Bayes tienen sus propios supuestos, uno sobre la diagonal fuera de la diagonal de las matrices de covarianza y el otro sobre la igualdad de las matrices de covarianza. Como Gaussian naive Bayes tiene cierto nivel de optimalidad (zhang,2004), queda claro por qué LDA y QDA son clasificadores tan efectivos.

11. Prueba de relación con la razón de verosimilitud

considerar r dos hipótesis para estimar algún parámetro, una hipótesis nula H_0 y una hipótesis alternativa H_A. La probabilidad PAG(rechazar $H_0 | H_0$) se denomina error de tipo 1, error de falso positivo o error de falsa alarma. La probabilidad PAG(aceptar $H_0 | H_A$) se denomina error tipo 2 o error falso negativo. El PAG(rechazar $H_0 | H_0$) es tambien llamado *Nivel significativo*, mientras 1 −PAG(aceptar $H_0 | H_A$) =PAG(rechazar $H_0 | H_A$) se llama *fuerza*.

Si $L(\theta_A)$ y $L(\theta_0)$ son las verosimilitudes (probabilidades) para las hipótesis alternativa y nula, la razón de verosimilitud es:

$$\Lambda = \frac{L(\theta_A)}{L(\theta_0)} = \frac{F(X;\theta_A)}{F(X;\theta_0)}. \quad (66)$$

La prueba de la razón de verosimilitud (LRT) (Casella y Berger, 2002) rechaza la H_0 a favor de H_A si la razón de verosimilitud es mayor que un umbral, es decir, $\Lambda \geq t$. La LRT es una prueba estadística muy efectiva porque de acuerdo con el lema de Neyman-Pearson (neyman y pearson,1933), tiene la mayor potencia entre todas las pruebas estadísticas con el mismo nivel de significación.

Si el tamaño de la muestra es grande, $n \to \infty$, y el θ_A se estima utilizando MLE, el logaritmo de la razón de verosimilitud tiene asintóticamente la distribución de χ^2 bajo la hipótesis nula (Blanco,1984;Casella y Berger,2002):

$$2\ln(\Lambda)_{H_0} \sim \chi^2_{(d.f.)}, \quad (67)$$

donde el grado de libertad de χ^2 la distribución es $d.f.$:= oscuro(H_A) −oscuro(H_0) y tenue(.) es el número de fijó parámetros en la hipótesis.

Hay una conexión entre LDA o QDA y el LRT (Lachenbruch y Goldstein,1979). Recuerde la ecuación. (12) o (15) que se puede reformular como:

$$\frac{F_2(X)\pi_2}{F_1(X)\pi_1} = 1, \quad (68)$$

que es para el límite de decisión. La ecuación. (22) se ocupó de la diferencia de $F_2(X)\pi_2$ y $F_1(X)\pi_1$; sin embargo, aquí estamos tratando con su proporción. Recuerde la figura 1 donde si nos movemos X_* a la derecha y a la izquierda, la relación $F_2(X_*)\pi_2 / F_1(X_*)\pi_1$ disminuye y aumenta, respectivamente, porque las probabilidades de que ocurra la primera y la segunda clase cambian. En otras palabras, mover el X_* cambia el nivel de significación y el poder. Por lo tanto, la ecuación. (68) se puede usar para tener una prueba estadística donde los posteriores se usan en la proporción, ya que también usamos posteriores en LDA y QDA. La hipótesis nula/alternativa se puede considerar como la media y la covarianza de la primera/segunda clase. En otras palabras, las dos hipótesis dicen que el punto pertenece a una clase específica. Por lo tanto, si la relación es mayor que un valor t, la instancia X se estima que pertenece a la segunda clase; de lo contrario, se elige la primera clase. De acuerdo con la Ec. (dieciséis), la ecuación. (68) se convierte en:

$$\frac{(|\Sigma_2|)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right)\pi_2}{(|\Sigma_1|)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)\pi_1} \geq yo, \quad (69)$$

para QDA. En LDA, las matrices de covarianza son iguales, por lo que:

$$\frac{\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right)\pi_2}{\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)\pi_1} \geq t. \quad (70)$$

Como se puede ver, cambiar el cambio a priori afecta la relación como se esperaba. Además, el valor de t se puede elegir de acuerdo con el nivel de significancia deseado en el χ^2 distribución usando el χ^2 mesa. Las ecuaciones (69) y (70) muestran la relación de LDA y QDA con LRT. Como el LRT tiene la mayor potencia (neyman y pearson,1933), la efectividad de LDA y QDA en la clasificación se explica desde el punto de vista de la prueba de hipótesis.

12. Simulaciones

En esta sección, presentamos algunas simulaciones que aclaran los conceptos del tutorial mediante ilustraciones.

12.1. Experimentos con tamaños de muestra de clase iguales

Creamos un conjunto de datos sintéticos de tres clases, cada una de las cuales es una distribución gaussiana bidimensional. El medias y matrices de covarianza de las tres gaussianas de

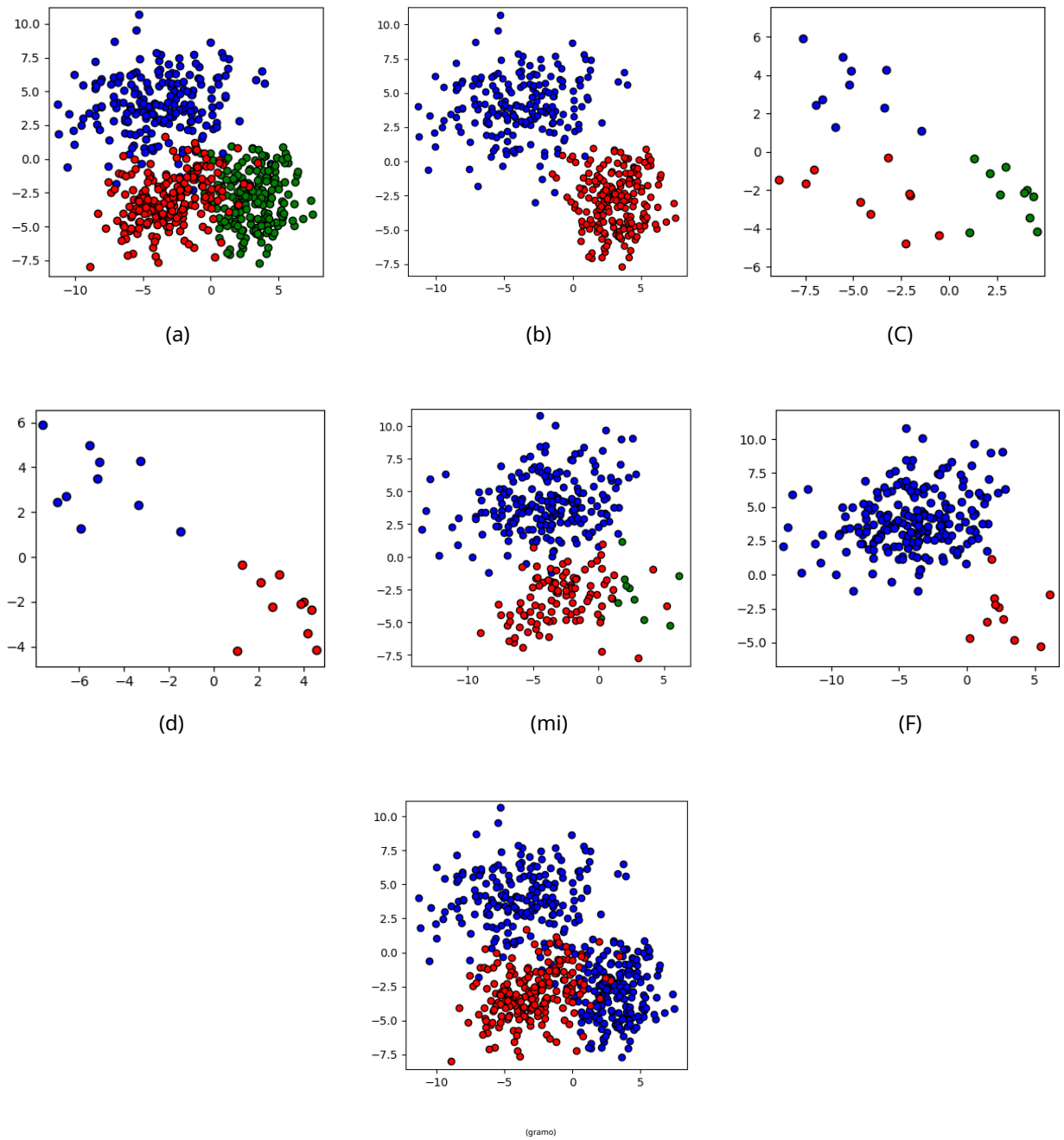

*Figura 3.*El conjunto de datos sintético: (a) tres clases cada una con tamaño200, (b) dos clases cada una con tamaño200, (c) tres clases cada una con tamaño 10, (d) dos clases cada una con tamaño10, (e) tres clases con tamaños200, 100,y10, (f) dos clases con tamaños200y10,y (g) dos clases con tamaños400y200donde la clase más grande tiene dos modas.

cuales las muestras de clase fueron extraídas al azar son:

$$\mu_1 = [-4,4]^\top, \mu_2 = [3,-3]^\top, \mu_1 = [-3,3]^\top,$$
$$\Sigma_1 = \begin{bmatrix} 10 & 1 \\ 1 & 5 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 6 & 1.5 \\ 1.5 & 4 \end{bmatrix}.$$

Las tres clases se muestran en la Fig.3-a donde cada uno tiene un tamaño de muestra200.Los experimentos se realizaron en las tres clases. También realizamos experimentos en dos de las tres clases para probar una clasificación binaria. Las dos clases se muestran en la Fig.3-b. LDA, QDA, Naive Bayes y

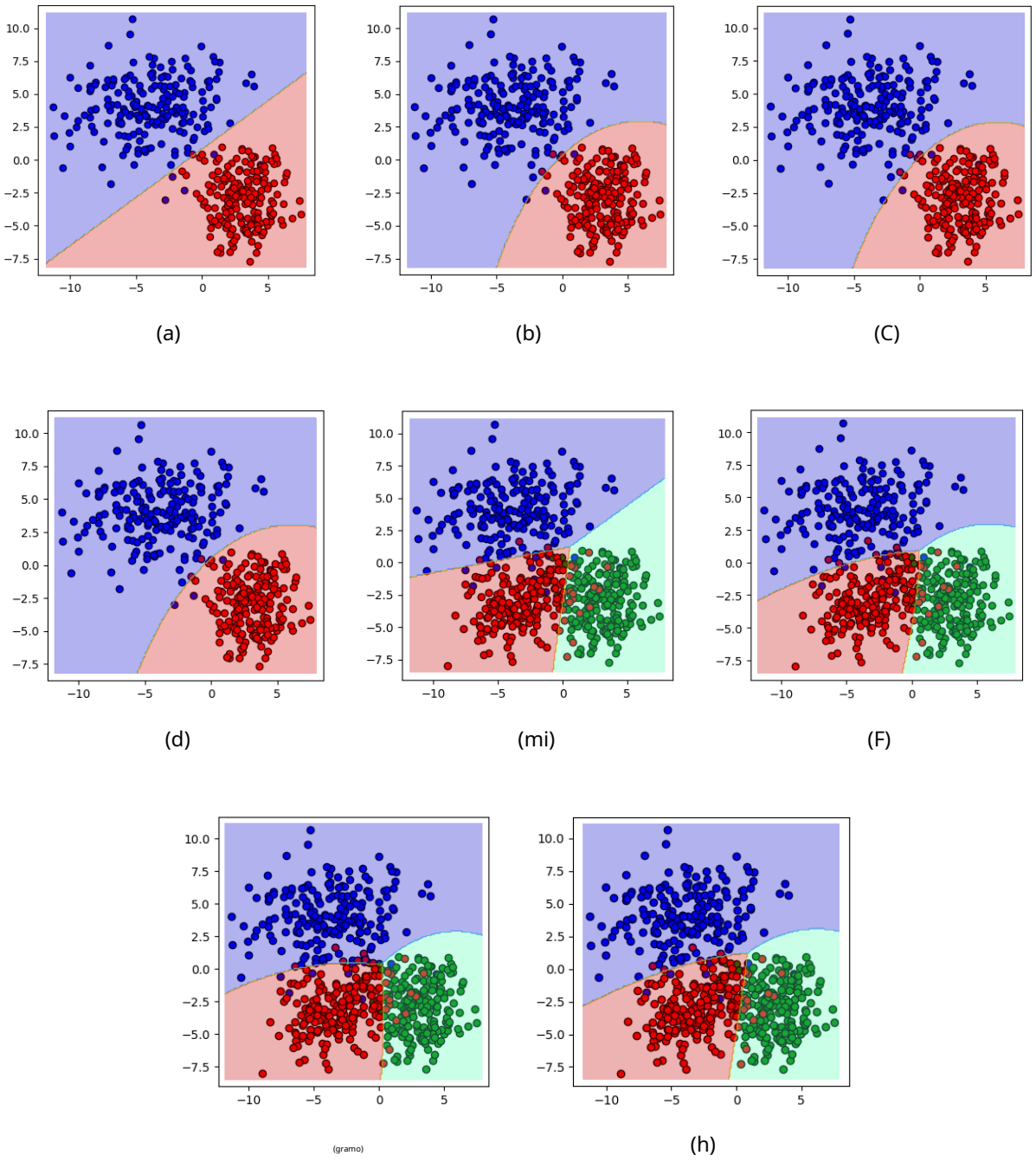

*Figura 4.*Experimentos con tamaños de muestra de clase iguales: (a) LDA para dos clases, (b) QDA para dos clases, (c) Gaussian naive Bayes para dos clases, (d) Bayes para dos clases, (e) LDA para tres clases, (f) QDA para tres clases, (g) bayesiana ingenua gaussiana para tres clases y (h) bayesiana para tres clases.

Las clasificaciones de Bayes de las dos y tres clases se muestran en la Fig.4. Para la clasificación binaria y ternaria con LDA y QDA, usamos las Ecs. (31) y (28), respectivamente, con la Ec. (29). Tambi én estimamos la media y la covarianza utilizando las Ecs. (33), (35), y (36). Para el ingenuo bayesiano gaussiano, usamos las Ecs. (62) y (63) y estimó los parámetros utilizando las Ecs. (64) y (sesenta y cinco). Para el clasificador de Bayes, usamos la ecuaci ón. (58) con la ecuación. (63) pero no estimamos la media y la varianza; excepto, con el fin de utilizar las probabilidades exactas en Eq. (58), utilizamos la media exacta y las matrices de covarianza

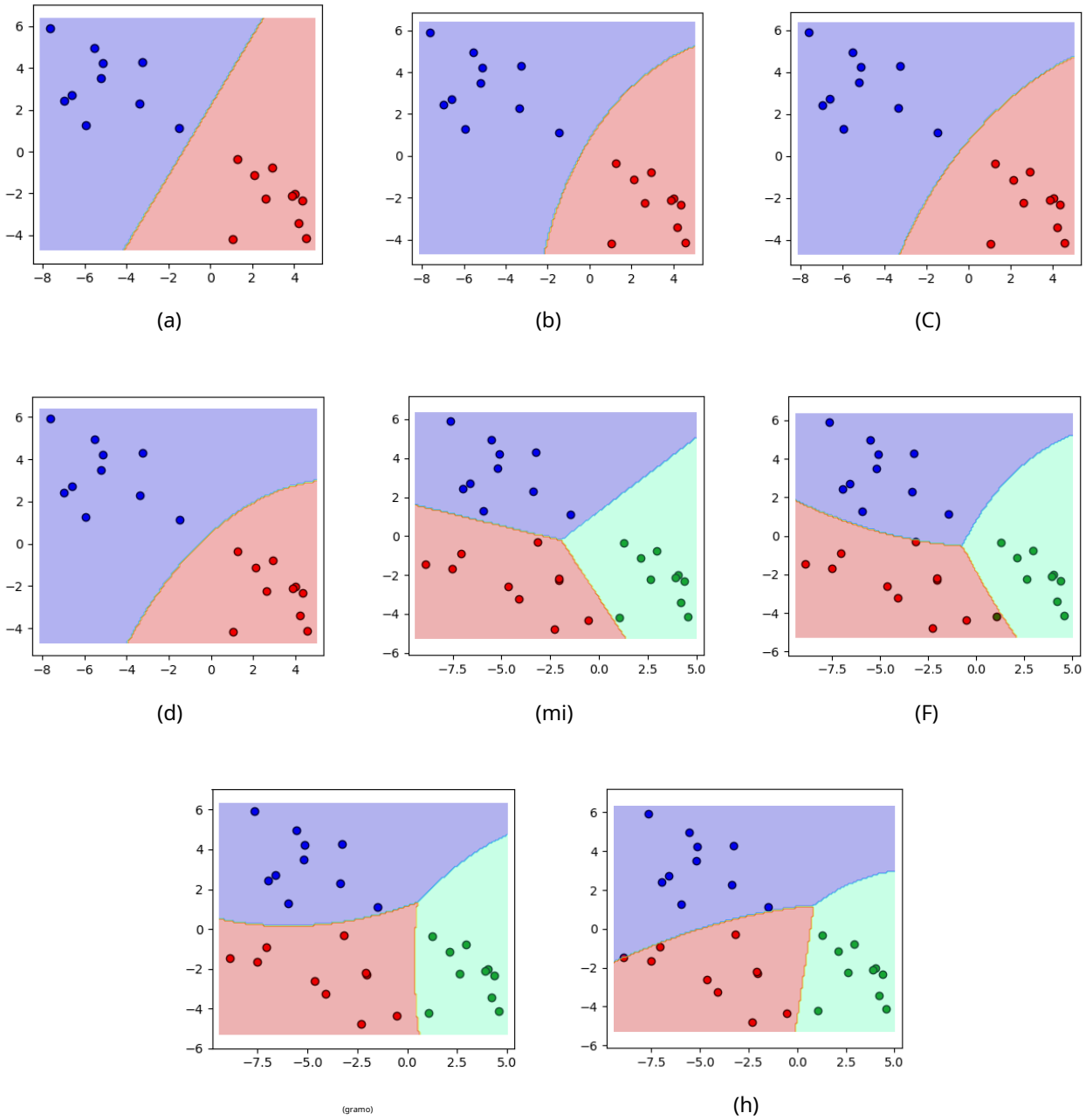

*Figura 5.*Experimentos con tamaños de muestra de clase pequeños: (a) LDA para dos clases, (b) QDA para dos clases, (c) Gaussian naive Bayes para dos clases, (d) Bayes para dos clases, (e) LDA para tres clases, (f) QDA para tres clases, (g) bayesiana ingenua gaussiana para tres clases y (h) bayesiana para tres clases.

de las distribuciones de las que tomamos muestras. Nosotros, sin embargo, estimamos los anteriores. Los priores se estimaron utilizando la Ec. (32) para todos los clasificadores.

Como se puede ver en la Fig.4, el espacio se divide en dos o tres partes y esto valida la afirmación de que LDA y QDA pueden considerarse métodos de aprendizaje de métricas, como se explica en la Sección7. Como era de esperar, los límites de LDA y QDA son lineales y curvos (cuadráticos), respectivamente. Los resultados de QDA, Gaussian naive Bayes y Bayes son muy similares aunque tienen ligeras diferencias. Esto se debe a que las clases ya son gaussianas, por lo que si las estimaciones de las medias y las matrices de covarianza son lo suficientemente precisas, QDA y Bayes son equivalentes. Las clases son gaussianas y los elementos de covarianza fuera de la diagonal

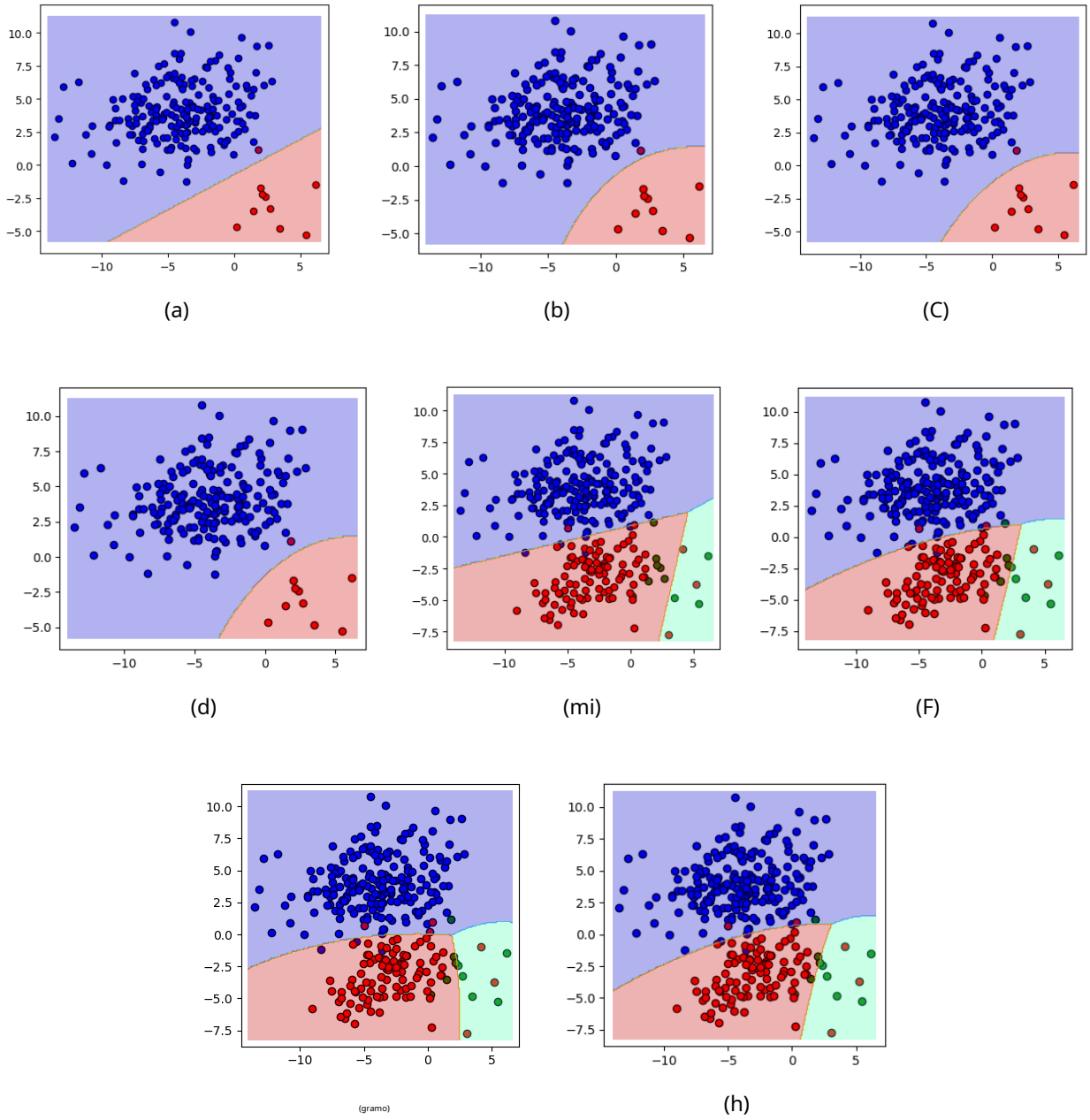

*Figura 6.*Experimentos con diferentes tamaños de muestra de clase: (a) LDA para dos clases, (b) QDA para dos clases, (c) Gaussian naive Bayes para dos clases, (d) Bayes para dos clases, (e) LDA para tres clases, (f) QDA para tres clases, (g) bayesiana ingenua gaussiana para tres clases y (h) bayesiana para tres clases.

las matrices también son pequeñas en comparación con la diagonal; por lo tanto, Naive Bayes también se comporta de manera similar.

12.2. Experimentos con tamaños de muestra de clase pequeños

Según la aproximación de Montecarlo (Roberto y Casela,2013), las estimaciones en las Ecs. (33), (35), (64) y (sesenta y cinco) son más precisos si el tamaño de la muestra tiende a infinito, es decir,$norte\to \infty$. Por lo tanto, si el tamaño de la muestra es pequeño, esperamos una diferencia de modo entre los clasificadores QDA y Bayes. Hicimos un conjunto de datos sintético con tres o dos clases con las mismas medias y matrices de covarianza mencionadas. El tamaño de la muestra de cada clase fue10Cifras3-c y 3-d muestra estos conjuntos de datos. Los resultados de los clasificadores LDA, QDA, Gaussian Naive Bayes y Bayes para este conjunto de datos son

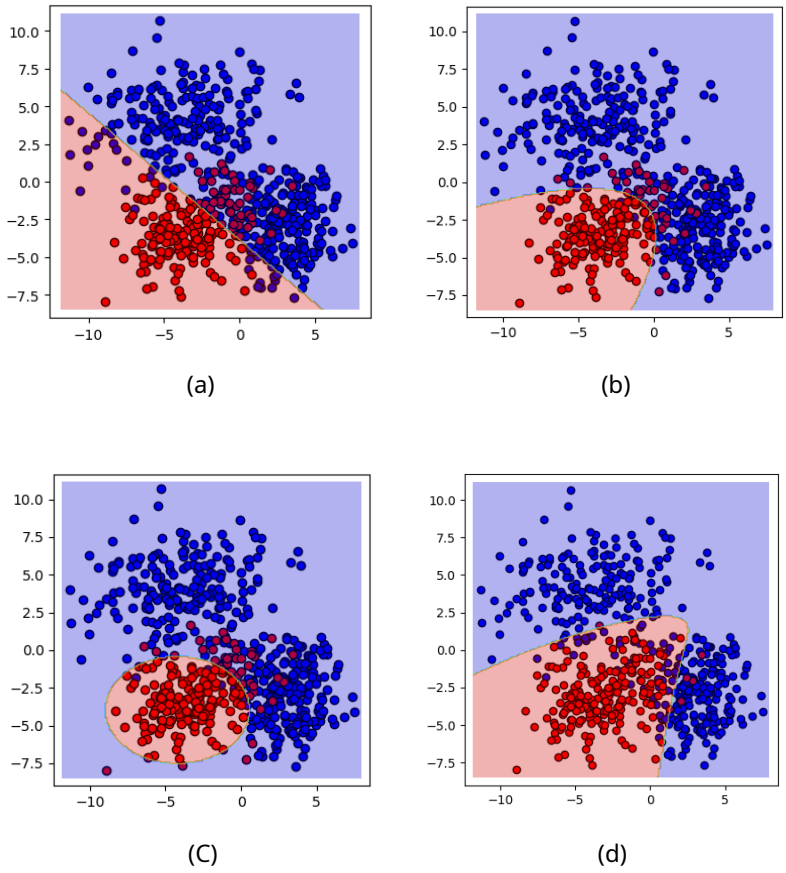

*Figura 7.*Experimentos con datos multimodales: (a) LDA, (b) QDA, (c) bayesiano ingenuo gaussiano y (d) bayesiano.

se muestra en la figura.5. Como se puede ver ahora, los resultados de QDA, Gaussian naive Bayes y Bayes son diferentes por la razón explicada.

12.3. Experimentos con diferentes tamaños de muestra de clase

De acuerdo con la Ec. (32) utilizado en las ecuaciones. (28), (31), (58), y (62), el anterior de una clase cambia según el tamaño de la muestra de la clase. Para ver el efecto del tamaño de la muestra, hicimos un conjunto de datos sintéticos con diferentes tamaños de clase, es decir, 200, 100,y10,mostrado en las Figs.3-mi,3-F. Utilizamos las mismas medias y matrices de covarianza mencionadas. Los resultados se muestran en la fig.6. Como puede verse, la clase con tamaño de muestra pequeño ha cubierto una pequeña porción de espacio en la discriminación que se esperaba porque su anterior es pequeño de acuerdo con la ecuación. (32); por lo tanto, su parte posterior es peque ña. Por otro lado, la clase con un tamaño de muestra grande ha cubierto una porción más grande debido a un previo más grande.

12.4. Experimentos con datos multimodales

Como se menciona en la Sección8, LDA y QDA asumen una distribución gaussiana unimodal para cada clase y, por lo tanto, FDA o LDA enfrenta un problema para datos multimodales (Sugiyama, 2007). Para probar esto, creamos un conjunto de datos sintético con dos clases, una con tamaño de muestra400teniendo dos modas de gaussianas y la otra con tamaño de muestra200tener un modo. Nuevamente usamos las mismas medias y matrices de covarianza mencionadas. El conjunto de datos se muestra en la Fig.3-gramo.

Los resultados de los clasificadores LDA, QDA, Gaussian Naive Bayes y Bayes para este conjunto de datos se muestran en la Fig.7. La media y la matriz de covarianza de la clase más grande, aunque tiene dos modas, se estimaron utilizando las Ecs. (33), (35), (64) y (sesenta y cinco) en LDA, QDA y Gaussian Naive Bayes. Sin embargo, para la verosimilitud utilizada en el clasificador de Bayes, es decir, en la Ec. (58), necesitamos saber la distribución multimodal exacta. Por lo tanto, ajustamos una mezcla de dos gaussianas (Ghojogh et al.,2019a) a los datos de la clase más grande:

$$\text{PAG}(X=x/x\in C_k) = \sum_{k=1}^{2} w_k P(X;\mu_k,\Sigma_k), \tag{71}$$

dónde$P(X;\mu_k,\Sigma_k)$es la ecuación (dieciséis) y tenemos el parámetro ajustado

ters fueron:

$$\mu_1 = [-3.88, 4]^\top, \mu_2 = [3.04, -2.92]^\top,$$
$$\Sigma_1 = \begin{bmatrix} 9.27 & 0.79 \\ 0.79 & 4.82 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2.87 & 0.03 \\ 0.03 & 3.78 \end{bmatrix},$$
$$w_1 = 0.49, w_2 = 0.502.$$

Como la fig.7espectáculos, LDA no se ha desempeñado lo suficientemente bien como se esperaba. El rendimiento de QDA es más aceptable que el de LDA pero aún no es lo suficientemente bueno porque QDA también asume un Gaussiano unimodal para cada clase. El resultado de Gaussian Naive Bayes es muy diferente del Bayes aquí porque Gaussian Naive Bayes asume Gaussiano unimodal con covarianza diagonal para cada clase. Finalmente, el Bayes tiene el mejor resultado ya que tiene en cuenta la multimodalidad de los datos y es óptimo (Mitchell, 1997).

13. Conclusión y trabajo futuro

Este documento fue un documento tutorial para LDA y QDA como dos métodos de clasificación fundamentales. Explicamos las relaciones de estos dos métodos con algunos otros métodos de aprendizaje autom ático, aprendizaje múltiple (subespacial), aprendizaje métrico, estad ísticas y pruebas estadísticas. También se proporcionaron algunas simulaciones para una mejor aclaración.

Este artículo se centró en LDA y QDA, que son discriminadores con uno y dos polinomios de grados de libertad, respectivamente. Como trabajo futuro, trabajaremos en un documento tutorial para el análisis discriminante no lineal utilizando kernels (Baudat & Anouar,2000;Li et al.,2003;Lu et al.,2003), Lo que es llamado*análisis discriminante del kernel*, tener discriminadores con más de dos grados de libertad.

Reconocimiento

Los autores agradecen enormemente al Prof. Ali Ghodsi (ver sus excelentes cursos relacionados en línea (Ghodsi,2015;2017)), Prof. Mu Zhu, Prof. Hoda Mohammadzade y otros profesores cuyos cursos han cubierto parcialmente los materiales mencionados en este tutorial.

Referencias

Baudat, Gaston y Anouar, Fatiha. discriminación generalizada análisis nant utilizando un enfoque kernel.*Computación neuronal*, 12(10):2385–2404, 2000.

Belhumeur, Peter N, Hespanha, João P y Kriegman, David J. Eigenfaces vs. Fisherfaces: Reconocimiento mediante proyección lineal específica de clase.*Transacciones IEEE sobre análisis de patrones e inteligencia artificial*, (7):711–720, 1997.

Boyd, Stephen y Vandenberghe, Lieven.*Optica convexa mización*. Prensa de la Universidad de Cambridge, 2004.

Casella, George y Berger, Roger L.*Inferencia estadística*, volumen 2. Duxbury Pacific Grove, CA, 2002.

Cox, Trevor F y Cox, Michael AA.*Multidimensional escalada*. Chapman and hall/CRC, 2000.

Croot, Ernie. El principio de Rayleigh para encontrar valores propios. Informe técnico, Instituto de Georgia de Tecnología, Escuela de Matemáticas, 2005. http:// En línea: people.math.gatech.edu/ ~ecroot/notes_linear.pdf, consultado: marzo de 2019.

De Maesschalck, Roy, Jouan-Rimbaud, Delphine y Massart, Désiré L. La distancia de Mahalanobis.*Quimiometr ía y sistemas de laboratorio inteligentes*, 50(1):1–18, 2000.

Etemad, Kamran y Chellappa, Rama. Anal discriminante ysis para el reconocimiento de imágenes de rostros humanos. *Revista de la Sociedad Óptica de América A*, 14(8):1724–1733, 1997.

Fisher, Ronald A. El uso de múltiples medidas en problemas taxonómicos.*Anales de eugenesia*, 7(2):179–188, 1936.

Friedman, Jerome, Hastie, Trevor y Tibshirani, Robert. *Los elementos del aprendizaje estadístico*, volumen 2. Springer series in statistics, Nueva York, NY, EE. UU., 2009.

Ghodsi, Ali. Curso de clasificación, departamento de estadística tics y ciencia actuarial, universidad de Waterloo. Vídeos de YouTube en línea, 2015. Consultado: enero de 2019.

Ghodsi, Ali. Curso de visualización de datos, departamento de estadística y ciencia actuarial, universidad de Waterloo. Vídeos de Youtube en línea, 2017. Consultado: enero de 2019.

Ghojogh, Benyamin, Mohammadzade, Hoda y Mokari, Mozhgan. Fisherposes para el reconocimiento de la acción humana utilizando los datos del sensor Kinect.*Revista de sensores IEEE*, 18(4): 1612–1627, 2017.

Ghojogh, Benyamin, Ghojogh, Aydin, Crowley, Mark y Karray, Fakhri. Ajuste de una distribución de mezcla a los datos: Tutorial.*preimpresión de arXiv arXiv:1901.06708*, 2019a.

Ghojogh, Benyamin, Karray, Fakhri y Crowley, Mark. Problemas de valores propios y valores propios generalizados: Tutorial.*preimpresión de arXiv arXiv:1903.11240*, 2019b.

Ham, Ji Hun, Lee, Daniel D, Mika, Sebastian y Scholkopf, Bernhard. Una vista del núcleo de la reducción de la dimensionalidad de las variedades. En*Conferencia Internacional sobre Aprendizaje Automático*, 2004.

Hazewinkel, Michiel. Teorema del límite central.*Enciclopedia de Matemáticas, Springer*, 2001.

Jolliffe, Ian.*Análisis de componentes principales*. Springer, 2011.

Kleinbaum, David G, Dietz, K, Gail, M, Klein, Mitchell, y Klein, Mitchell. *Regresión logística*. Springer, 2002.

Kulis, Brian. Aprendizaje métrico: una encuesta. *Fundaciones y Tendencias® en aprendizaje automático*, 5(4):287–364, 2013.

Lachenbruch, Peter A and Goldstein, M. Discriminante análisis. *Biometría*, págs. 69–85, 1979.

Li, Yongmin, Gong, Shaogang y Liddell, Heather. Reconocimiento de trayectorias de identidades faciales mediante an álisis discriminante del kernel. *Computación de imagen y visión*, 21 (13-14):1077–1086, 2003.

Lu, Juwei, Plataniotis, Konstantinos N y Venetsanopoulos, Anastasios N. Reconocimiento facial mediante algoritmos de análisis discriminante directo del kernel. *Transacciones IEEE en redes neuronales*, 14(1):117–126, 2003.

Malekmohammadi, Alireza, Mohammadzade, Hoda, Chamanzar, Alireza, Shabany, Mahdi y Ghojogh, Benyamin. Una implementación de hardware eficiente para un sistema de interfaz de computadora cerebro de imágenes motoras. *Ciencia Iranica*, 26:72–94, 2019.

McLachlan, Goeffrey J. Mahalanobis distancia. *Resonancia*, 4(6):20–26, 1999.

Mitchell, Tomás. *Aprendizaje automático*. McGraw colina superior Educación, 1997.

Mokari, Mozhgan, Mohammadzade, Hoda y Ghojogh, Benyamin. Reconocimiento de acciones involuntarias a partir de datos esqueléticos en 3D utilizando estados corporales. *Ciencia Iranica*, 2018.

MurphyKevin P. *Aprendizaje automático: un per- prospectivo*. Prensa del MIT, 2012.

Neyman, Jerzy y Pearson, Egon Sharpe. IX. sobre el prolema de las pruebas más eficientes de hipótesis estad ísticas. *Transacciones filosóficas de la Royal Society de Londres. Serie A, que contiene papeles de carácter matemático o físico*, 231(694-706):289–337, 1933.

Robert, Christian y Casella, George. *Estación de Montecarlo métodos estadísticos*. Medios de comunicación de ciencia y negocios de Springer, 2013.

Strange, Harry y Zwiggelaar, Reyer. *Problemas abiertos en Reducción de dimensionalidad espectral*. Springer, 2014.

Sugiyama, Masashi. Reducción de la dimensionalidad de multi- datos etiquetados modales por análisis discriminante de pescadores locales. *Diario de investigación de aprendizaje automático*, 8 (mayo): 1027 – 1061, 2007.

Tharwat, Alaa, Gaber, Tarek, Ibrahim, Abdelhameed y Hassanien, Aboul Ella. Análisis discriminante lineal: un tutorial detallado. *comunicaciones de IA*, 30(2):169–190, 2017.

Welling, Max. Análisis discriminante lineal de Fisher. tecnología- Informe técnico, Universidad de Toronto, Toronto, Ontario, Canadá, 2005.

Blanco, Halbert. *Teoría asintótica para econometristas*. Prensa académica, 1984.

Xu, Yong y Lu, Guangming. Análisis sobre la discriminación de los pescadores criterio inant y separabilidad lineal del espacio de caracter ísticas. En *2006 Conferencia Internacional sobre Inteligencia Computacional y Seguridad*, volumen 2, págs. 1671–1676. IEEE, 2006.

Yang, Liu y Jin, Rong. Aprendizaje métrico a distancia: encuesta completa. Informe técnico, Departamento de Ciencias de la Computación e Ingeniería, Universidad Estatal de Michigan, 2006.

Zhang, Harry. La optimalidad de naive Bayes. En *Americano Asociación para la Inteligencia Artificial (AAAI)*, 2004.

Zhao, Wenyi, Chellappa, Rama y Phillips, P Jonathon. *Análisis discriminante lineal subespacial para reconocimiento facial*. Citaseer, 1999.