Discriminación

Graciela Boente

Tenemos k poblaciones o grupos diferentes $\mathcal{P}_1, \ldots, \mathcal{P}_k$ y un vector $\mathbf{x} \in \mathbb{R}^p$ que puede pertenecer a cualquiera de esas poblaciones.

a) G una variable aleatoria que indica la pertenencia al grupo,

$$\mathbb{P}(G = j) = \pi_j$$

Enfoque bayesiano,

- $\Theta = \{\mathcal{P}_1, \ldots, \mathcal{P}_k\}$ o en forma simplificada tenemos una variable aleatoria G que toma valores $\{1, \ldots, k\}$.
- Sobre Θ definimos una probabilidad a priori τ que es discreta y es tal que $\mathbb{P}(G = j) = \pi_j$

b) La distribución de $\mathbf{x}$ varía según el grupo de pertenencia. Si $\mathbf{x}$ pertenece a la población $\mathcal{P}_j$ entonces $\mathbf{x}$ tiene densidad f_j,

$$\mathbf{x} | G = j \sim f_j$$

c) Para dar una regla de clasificación daremos una partición de $\mathbb{R}^p$ en k conjuntos disjuntos

$$\mathbb{R}^p = \cup_{i=1}^k \mathcal{G}_j \qquad \mathcal{G}_j \cap \mathcal{G}_s = \emptyset$$

Definiciones

La densidad marginal de $\mathbf{x}$ está dada por

$$f_{\mathbf{x}}(\mathbf{x}) = \sum_{j=1}^{k} \pi_j f_j(\mathbf{x})$$

y la probabilidad condicional de que una observación pertenezca a $\mathcal{P}_j$ dado que $\mathbf{x} = \mathbf{x}_0$, está dada por

$$q_j(\mathbf{x}_0) = \mathbb{P}(G = j | \mathbf{x} = \mathbf{x}_0) = \frac{\pi_j f_j(\mathbf{x}_0)}{\sum_{\ell=1}^{k} \pi_\ell f_\ell(\mathbf{x}_0)}$$

La cantidad $q_j(\mathbf{x}_0)$ es la probabilidad a posteriori.

Definición 1.

Una regla de clasificación es una variable aleatoria $G^\star(\mathbf{x})$ tal que

$$G^\star(\mathbf{x}) = j \qquad \text{si} \quad \mathbf{x} \in \mathcal{G}_j$$

donde $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$ es una partición de $\mathbb{R}^p$.

Definiciones

Podemos ver a $G^\star$ como la pertenencia predicha mientras que G es la pertenencia real.

La teoría de clasificación trata de encontrar reglas de clasificación óptimas en algún sentido. Lo ideal sería que $\mathbb{P}(G^\star = G) = 1$, pero esto no es posible.

Definición 2. Para una regla de clasificación $G^\star$ con regiones de clasificación $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$, la probabilidad de asignar la observación $\mathbf{x}$ a $\mathcal{P}_i$ cuando en realidad, $\mathbf{x} \in \mathcal{P}_j$ es

$$p_{i|j} = \mathbb{P}(G^\star = i | G = j) = \mathbb{P}(\mathbf{x} \in \mathcal{G}_i | G = j) = \int_{\mathcal{G}_i} f_j(\mathbf{x}) d\mathbf{x}$$

Observemos que $\sum_{i=1}^{k} p_{i|j} = 1$.

Definiciones

A veces es posible asignar un costo $c_{i|j} \geq 0$ a la clasificación de una observación del grupo j en el grupo i. En muchos casos se elige $c_{i|j} = 1$ si $i \neq j$.

Definimos

a) La función de pérdida como

$$L(\mathcal{P}_j, i) = \begin{cases} c_{i|j} & \text{si} \quad i \neq j \\ 0 & \text{si} \quad i = j \end{cases}$$

$$L(\mathcal{P}_j, G^\star) = L(j, G^\star) = \sum_{i=1}^{k} c_{i|j} \mathbb{I}(\mathbf{x} \in \mathcal{G}_i | \mathbf{x} \in \mathcal{P}_j)$$

donde $c_{j|j} = 0$.

b) El riesgo de $G^\star$ es

$$R(\mathcal{P}_j, G^\star) = \mathbb{E}\ L(\mathcal{P}_j, G^\star) = \sum_{i=1}^{k} c_{i|j} p_{i|j} = \sum_{i=1}^{k} c_{i|j} \int_{\mathcal{G}_i} f_j(\mathbf{x}) d\mathbf{x}$$

Definiciones

El Riesgo de Bayes de una regla de clasificación $G^\star$ será

$$r(\tau, G^\star) = \mathbb{E}R(\Theta, G^\star) = \sum_{j=1}^{k}\sum_{i\neq j}\pi_j c_{i|j} p_{i|j} = \sum_{j=1}^{k}\sum_{i\neq j}\pi_j c_{i|j}\int_{\mathcal{G}_i} f_j(\mathbf{x})d\mathbf{x}$$

En particular, si $c_{i|j} = 1$ si $i \neq j$ tenemos que

$$r(\tau, G^\star) = \sum_{j=1}^{k}\sum_{i\neq j}\pi_j p_{i|j} = 1 - \sum_{j=1}^{k}\pi_j p_{j|j} = 1 - \sum_{j=1}^{k}\pi_j\int_{\mathcal{G}_j} f_j(\mathbf{x})d\mathbf{x}$$

que se llama la probabilidad total de mala clasificación ya que coincide con $\mathbb{P}(G^\star \neq G)$.

Si $k = 2$ y $c_{i|j} = 1$ si $i \neq j$ tenemos que

$$r(\tau, G^\star) = \pi_1 + \int_{\mathcal{G}_1}\left[\pi_2 f_2(\mathbf{x}) - \pi_1 f_1(\mathbf{x})\right]d\mathbf{x}$$

Definiciones

1. Diremos que una regla de clasificación $G_0^\star$ es Bayes respecto de la distribución a priori τ si

$$r(\tau, G_0^\star) = \min_{G^\star} r(\tau, G^\star)$$

2. Diremos que una regla de clasificación $G_0^\star$ es minimax si

$$\max_{1\leq j\leq k} R(\mathcal{P}_j, G_0^\star) = \min_{G^\star} \max_{1\leq j\leq k} R(\mathcal{P}_j, G^\star)$$

Propiedad

La regla Bayes respecto de τ clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

$$\mathcal{G}_{i,0} = \{\mathbf{x} \in \mathbb{R}^p : \quad \sum_{j=1}^{k} \pi_j c_{i|j} f_j(\mathbf{x}) < \sum_{j=1}^{k} \pi_j c_{\ell|j} f_j(\mathbf{x}) \quad \forall \ell \neq i\}$$

o sea, clasifico $\mathbf{x} \in \mathcal{P}_i$ si

$$\sum_{j=1}^{k} \pi_j c_{i|j} f_j(\mathbf{x}) = \min_{\ell} \sum_{j=1}^{k} \pi_j c_{\ell|j} f_j(\mathbf{x})$$

siendo la asignación en la frontera de $\mathcal{G}_{i,0}$ arbitraria.

Casos particulares

a) Supongamos que $c_{i|j} = 1$ si $i \neq j$ entonces la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

$$\begin{aligned} \mathcal{G}_{i,0} &= \{\mathbf{x} \in \mathbb{R}^p : \quad \pi_\ell\, f_\ell(\mathbf{x}) < \pi_i\, f_i(\mathbf{x}) \quad \forall \ell \neq i\} \\ &= \{\mathbf{x} \in \mathbb{R}^p : \quad q_\ell(\mathbf{x}) < q_i(\mathbf{x}) \quad \forall \ell \neq i\} \end{aligned}$$

es decir, clasifico $\mathbf{x} \in \mathcal{P}_i$ si $q_i(\mathbf{x}) = \max_{1 \leq \ell \leq k} q_\ell(\mathbf{x})$ siendo la asignación en la frontera de $\mathcal{G}_{i,0}$ arbitraria. Por lo tanto,

i) la regla Bayes coincide con el criterio de minimizar la probabilidad total de mala clasificación.

ii) la regla Bayes coincide con el criterio de maximizar la probabilidad a posteriori.

iii) si además, $\pi_j = 1/k$, $1 \leq j \leq k$, la regla Bayes coincide con el criterio de máxima verosimilitud, que asigna $\mathbf{x}$ a la población que máximiza la verosimilitud de $\mathbf{x}$.

Casos particulares

b) Supongamos que $k = 2$ entonces la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_1$ si $\mathbf{x} \in \mathcal{G}_{1,0}$ donde

$$\mathcal{G}_{1,0} = \left\{ \mathbf{x} \in \mathbb{R}^p : \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2 c_{1|2}}{\pi_1 c_{2|1}} \right\}$$

i) Si $\pi_2 c_{1|2} = \pi_1 c_{2|1}$ la regla Bayes da el criterio de máxima verosimilitud. En particular, si $c_{i|j} = 1$ si $i \neq j$, el criterio de máxima verosimilitud es la regla Bayes asociada a $\pi_1 = \pi_2 = \frac{1}{2}$

ii) si $c_{i|j} = 1$ si $i \neq j$ y $\pi_1 = 1 - \alpha$, $\pi_2 = \alpha$, $0 < \alpha < 1$, entonces la regla Bayes clasifica

- $\mathbf{x} \in \mathcal{P}_1$ si

$$\mathbf{x} \in \mathcal{G}_{1,0} = \{ \mathbf{x} \in \mathbb{R}^p : \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > a = \frac{\alpha}{1-\alpha} \}$$

- $\mathbf{x} \in \mathcal{P}_2$ si

$$\mathbf{x} \in \mathcal{G}_{2,0} = \{ \mathbf{x} \in \mathbb{R}^p : \quad \frac{f_2(\mathbf{x})}{f_1(\mathbf{x})} > \frac{1}{a} = \frac{1-\alpha}{\alpha} \}$$

Casos particulares

b) ii) Sea a_0 tal que

$$\int_{\mathcal{G}_{1,0}} f_2(\mathbf{x})d\mathbf{x} = \int_{\mathcal{G}_{2,0}} f_1(\mathbf{x})d\mathbf{x}$$

Luego, la regla Bayes respecto de $\tau = (\alpha_0, 1-\alpha_0)$ con $\alpha_0 = 1/(1+a_0)$ iguala riesgos y es la regla minimax.

c) Supongamos que $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)^{\mathrm{T}}$ con $\mathbf{x}_1 \in \mathbb{R}^q$ y que $\mathbf{x}_1$ y $\mathbf{x}_2$ son independientes en todas las poblaciones, o sea, $f_j(\mathbf{x}) = h_j(\mathbf{x}_1)\ell_j(\mathbf{x}_2)$. Más aún, supongamos que $\ell_j(\mathbf{x}_2) = \ell(\mathbf{x}_2)$ para todo j. Entonces, la regla de clasificación se basa solamente en $\mathbf{x}_1$, es decir, la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

$$\mathcal{G}_{i,0} = \{\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^p : \sum_{j=1}^{k} \pi_j c_{i|j} h_j(\mathbf{x}_1) < \sum_{j=1}^{k} \pi_j c_{\ell|j} h_j(\mathbf{x}_1) \quad \forall \ell \neq i\}$$

o sea, clasifico $\mathbf{x} \in \mathcal{P}_i$ si $\sum_{j=1}^{k} \pi_j c_{i|j} h_j(\mathbf{x}_1) = \min_\ell \sum_{j=1}^{k} \pi_j c_{\ell|j} h_j(\mathbf{x}_1)$.

Problema

Hasta ahora supusimos que la distribución de $\mathbf{x}$ en cada población es conocida. En la mayoría de los casos esto no ocurre y tenemos alguna de las siguientes situaciones

a) la distribución es conocida salvo por algunos parámetros que deberemos estimar, $f_j = f_j(\cdot, \boldsymbol{\theta}_j)$

b) la distribución es parcialmente desconocida, o sea, sabemos por ejemplo que

$$\log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}$$

c) la distribución es desconocida

En a) y b) estimamos los parámetros. La regla en este caso se estima por $\widehat{G}_0^\star$ reemplazando los parámetros desconocidos por sus estimadores.

Entonces, necesitamos conocer las probabilidades de error cometido, o sea, aproximar $r(\tau, G_0^\star)$ y $R(\mathcal{P}_j, G_0^\star)$.

Error óptimo de clasificación

El error de mala clasificación de la población j

$$R(\mathcal{P}_j, G_0^\star) = \sum_{i \neq j}^{k} p_{i|j} = \sum_{i \neq j}^{k} \int_{\mathcal{G}_i} f_j(\mathbf{x}) d\mathbf{x}$$

Si $k = 2$, llamaremos

$$\begin{aligned}
e_{1,opt} = R(\mathcal{P}_1, G_0^\star) &= \int_{\mathcal{G}_{2,0}} f_1(\mathbf{x}) d\mathbf{x} = 1 - \int_{\mathcal{G}_{1,0}} f_1(\mathbf{x}) d\mathbf{x} \\
e_{2,opt} = R(\mathcal{P}_2, G_0^\star) &= \int_{\mathcal{G}_{1,0}} f_2(\mathbf{x}) d\mathbf{x} = 1 - \int_{\mathcal{G}_{2,0}} f_2(\mathbf{x}) d\mathbf{x} \\
e_{opt} &= \pi_1 e_{1,opt} + \pi_2 e_{2,opt}
\end{aligned}$$

Error óptimo de clasificación si $f_j = f_j(\cdot, \boldsymbol{\theta}_j)$

El error de mala clasificación de la población j

$$R(\mathcal{P}_j, G_0^\star) = \sum_{i \neq j}^{k} p_{i|j} = \sum_{i \neq j}^{k} \int_{\mathcal{G}_i} f_j(\mathbf{x}, \boldsymbol{\theta}_j) d\mathbf{x}$$

Si $k = 2$, llamaremos

$$\begin{aligned}
e_{1,opt} = R(\mathcal{P}_1, G_0^\star) &= \int_{\mathcal{G}_{2,0}} f_1(\mathbf{x}, \boldsymbol{\theta}_1) d\mathbf{x} = 1 - \int_{\mathcal{G}_{1,0}} f_1(\mathbf{x}, \boldsymbol{\theta}_1) d\mathbf{x} \\
e_{2,opt} = R(\mathcal{P}_2, G_0^\star) &= \int_{\mathcal{G}_{1,0}} f_2(\mathbf{x}, \boldsymbol{\theta}_2) d\mathbf{x} = 1 - \int_{\mathcal{G}_{2,0}} f_2(\mathbf{x}, \boldsymbol{\theta}_2) d\mathbf{x} \\
e_{opt} &= \pi_1 e_{1,opt} + \pi_2 e_{2,opt}
\end{aligned}$$

$f_j = f_j(\cdot, \boldsymbol{\theta}_j)$, $\boldsymbol{\theta}_j$ desconocido

Hasta ahora supusimos que la distribución de $\mathbf{x}$ en cada población es conocida. Supongamos que la distribución es conocida salvo por algunos parámetros que deberemos estimar, $f_j = f_j(\cdot, \boldsymbol{\theta}_j)$ y sea

- $\widehat{\boldsymbol{\theta}}_j$ un estimador de $\boldsymbol{\theta}_j$ basado en la muestra $\mathbf{x}_{j,1}, \dots, \mathbf{x}_{j,n_j}$.

- $\widehat{f}_j(\mathbf{x}) = f_j(\cdot, \widehat{\boldsymbol{\theta}}_j)$

entonces la regla Bayes, con $c_{i|j} = 1$ si $i \neq j$, se estima por la regla $\widehat{G}_0^\star$ que clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \widehat{\mathcal{G}}_{i,0}$ donde

$$\widehat{\mathcal{G}}_{i,0} = \{\mathbf{x} \in \mathbb{R}^p : \quad \pi_\ell \widehat{f}_\ell(\mathbf{x}) < \pi_i \widehat{f}_i(\mathbf{x}) \quad \forall \ell \neq i\}$$

Se sugiere que n_j sea tres veces por lo menos la cantidad de parámetros $\boldsymbol{\theta}_j$ a estimar y el número puede ser mayor si los grupos no están bien separados.

Cálculo errores de clasificación f_j, $j = 1, 2$

Se definen varios tipos de errores

a) El error actual

$$
\begin{aligned}
e_{1,act} = R(\mathcal{P}_1, \widehat{G}_0^\star) &= \int_{\widehat{\mathcal{G}}_{2,0}} f_1(\mathbf{x})d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{1,0}} f_1(\mathbf{x})d\mathbf{x} \\
e_{2,act} = R(\mathcal{P}_2, \widehat{G}_0^\star) &= \int_{\widehat{\mathcal{G}}_{1,0}} f_2(\mathbf{x})d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{2,0}} f_2(\mathbf{x})d\mathbf{x} \\
e_{act} &= \pi_1 e_{1,act} + \pi_2 e_{2,act}
\end{aligned}
$$

Claramente, $e_{opt} \leq e_{act}$

b) La tasa de error actual esperada

$$
\mathbb{E}e_{act} = \pi_1 \mathbb{E}e_{1,act} + \pi_2 \mathbb{E}e_{2,act}
$$

Estimación de los errores de clasificación

1) <u>El estimador *plug–in*</u>

$$\begin{aligned}
\widehat{e}_{j,act} &= \sum_{\ell \neq j} \int_{\widehat{\mathcal{G}}_{\ell,0}} \widehat{f}_j(\mathbf{x}) d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{j,0}} \widehat{f}_j(\mathbf{x}) d\mathbf{x} \\
\widehat{e}_{act} &= \sum_{j=1}^{k} \pi_j \widehat{e}_{k,act}
\end{aligned}$$

Este error se basa en la corrrecta especificación del modelo pero además en muchos casos, como veremos, subestima el error real e_{opt}.
Si $k = 2$

$$\begin{aligned}
\widehat{e}_{1,act} &= \int_{\widehat{\mathcal{G}}_{2,0}} \widehat{f}_1(\mathbf{x}) d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{1,0}} \widehat{f}_1(\mathbf{x}) d\mathbf{x} \\
\widehat{e}_{2,act} &= \int_{\widehat{\mathcal{G}}_{1,0}} \widehat{f}_2(\mathbf{x}) d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{2,0}} \widehat{f}_2(\mathbf{x}) d\mathbf{x} \qquad \widehat{e}_{act} = \pi_1 \widehat{e}_{1,act} + \pi_2 \widehat{e}_{2,act}
\end{aligned}$$

Estimación de los errores de clasificación

Sea

$$\widehat{r}_i(\mathbf{x}) = \pi_i f(\mathbf{x}, \widehat{\boldsymbol{\theta}}_i)$$

un estimador **insesgado** de

$$r_i(\mathbf{x}) = \pi_i f(\mathbf{x}, \boldsymbol{\theta}_i)$$

es decir,

$$\mathbb{E}_{\boldsymbol{\theta}} \widehat{r}_i(\mathbf{x}) = r_i(\mathbf{x}) \quad \text{para casi todo } \mathbf{x}$$

Entonces, si $c_{i|j} = 1$ para $i \neq j$ se cumple que

$$\mathbb{E}_{\boldsymbol{\theta}} \widehat{e}_{act} \leq e_{opt} \leq \mathbb{E}_{\boldsymbol{\theta}} e_{act}$$

Estimación de los errores de clasificación

2) La tasa de error aparente.

Consideremos la regla basada en las regiones $\widehat{\mathcal{G}}_{j,0}$, $1 \le j \le k$ y sean

$$n_{i,j} = \#\{\mathbf{x}_{i\ell} \text{ clasificadas en la población } \mathcal{P}_j\} = \#\{\mathbf{x}_{i\ell} \in \widehat{\mathcal{G}}_{j,0}\}$$

$$\begin{aligned} n_i &= \sum_{j=1}^{k} n_{ij} \text{ el total de observaciones de la población } i\text{−ésima,} \\ \widehat{\pi}_i &= \frac{n_i}{n} \quad \text{con} \quad n = \sum_{i=1}^{k} n_i \\ m_i &= \#\{\mathbf{x}_{i\ell} \text{ mal clasificadas }\} = \sum_{j \neq i} n_{ij} \end{aligned}$$

Estimación de los errores de clasificación

2) La tasa de error aparente es

$$e_{i,app} = \frac{m_i}{n_i} \qquad e_{app} = \sum_{j=1}^{k} \pi_i e_{i,app}$$

$$\widehat{e}_{app} = \sum_{j=1}^{k} \widehat{\pi}_i e_{i,app} = \frac{\sum_{i=1}^{k} m_i}{n}$$

El método basado en $\widehat{e}_{app}$ se llama también de resustitución.

Este estimador del error es muy optimista ya que tiende a subestimar la probabilidad real de error, pues los mismos datos se usan para armar la regla (estimar los parámetros) y para evaluar la regla resultante. Los estimadores de los parámetros obtenidos son los que mejor ajustan a los datos y por ello tiendo a clasificar mejor.

Estimación de los errores de clasificación

2) Veamos un ejemplo de como la tasa de error aparente subestima.

Sea $X \in \mathbb{R}$ y supongamos que $n_1 = n_2 = 1$, o sea, tenemos las observaciones x_1 y x_2 de la población 1 y 2, respectivamente. Supongamos $x_1 > x_2$.

Consideremos la regla de clasificación que asigna x a $\mathcal{P}_1$ si $x \geq (x_1 + x_2)/2$ y al grupo 2 en otro caso.

Entonces, $\widehat{e}_{app} = 0$, lo cual es demasiado optimista.

Estimación de los errores de clasificación

3) El estimador de convalidación cruzada.
En este método se sacan las observaciones de a una. Con los $n-1$ datos restantes se arma la regla y se clasifica la observación extraída. Sea

$$a_i = \#\{\mathbf{x}_{i\ell} \text{ mal clasificadas }, 1 \le \ell \le n_i\}$$

$$e_{i,cv} = \frac{a_i}{n_i} \qquad e_{cv} = \sum_{i=1}^{k} \pi_i e_{i,cv}$$

$$\widehat{e}_{cv} = \sum_{i=1}^{k} \widehat{\pi}_i e_{i,cv} = \frac{\sum_{i=1}^{k} a_i}{n}$$

Estimación de los errores de clasificación

3) El estimador de convalidación cruzada.

   Este método da estimadores consistentes del error pero con varianza grande.
   Obviamente, es más costoso computacionalmente pero da resultados más honestos y debería ser usado si es posible.
   En el caso normal, las fórmulas para los estimadores de los parámetros evitan efectuar el cálculo de la regla en cada paso.

4) Otra opción es usar el método de $M-$fold que divide la muestra total en M grupos y construye la regla con $M-1$ grupos mientras clasifica el grupo restante, sucesivamente.

Estimación de los errores de clasificación

5) El estimador bootstrap.

Como $e_{i,app}$ es sesgado, Efron (1979) sugiere estimar su sesgo usando boostrap.

a) Para cada $1 \le i \le k$, tomamos una muestra $\mathbf{x}^*_{i\ell}$ con reemplazo de la muestra original de la población $\mathcal{P}_i$, de tamaño n_i.

b) Construyamos la regla de clasificación basada en esta muestras que llamaremos $\widehat{G}_0^{\star,*}$ con regiones $\widehat{\mathcal{G}}^*_{i,0}$. Sean

$$\begin{aligned}
m_i^* &= \#\{\mathbf{x}^*_{i\ell} \text{ mal clasificadas } 1 \le \ell \le n_i\} = \#\{\mathbf{x}^*_{i\ell} \notin \widehat{\mathcal{G}}^*_{j,0}\ 1 \le \ell \le n_i\} \\
m_i^{**} &= \#\{\mathbf{x}_{i\ell} \text{ mal clasificadas } 1 \le \ell \le n_i\} = \#\{\mathbf{x}_{i\ell} \notin \widehat{\mathcal{G}}^*_{j,0}\ 1 \le \ell \le n_i\} \\
d_i &= \frac{m_i^{**} - m_i^*}{n_i}
\end{aligned}$$

c) Repitase a) y b) un número B grande de veces. Sea $d_{i,s}$ el valor de d_i en la replicación s y defina $\overline{d}_i = \sum_{s=1}^{B} d_{i,s}$

Estimación de los errores de clasificación

5) El estimador bootstrap.
El estimador bootstrap se define como

$$e_{i,boot} = \frac{m_i}{n_i} + \overline{d}_i \qquad e_{boot} = \sum_{i=1}^{k} \pi_i e_{i,boot}$$

$$\widehat{e}_{boot} = \sum_{i=1}^{k} \widehat{\pi}_i e_{i,boot}$$

$$f_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

Supongamos que $c_{i|j} = 1$ si $i \neq j$ entonces la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

$$\begin{aligned}
\mathcal{G}_{i,0} &= \{\mathbf{x} \in \mathbb{R}^p : \quad \frac{\pi_\ell}{\pi_i} < \frac{f_i(\mathbf{x})}{f_\ell(\mathbf{x})} \quad \forall \ell \neq i\} \\
&= \{\mathbf{x} \in \mathbb{R}^p : \quad \log\left(\frac{f_i(\mathbf{x})}{f_\ell(\mathbf{x})}\right) > \log\left(\frac{\pi_\ell}{\pi_i}\right) \quad \forall \ell \neq i\}
\end{aligned}$$

donde

$$\begin{aligned}
\log\left(\frac{f_i(\mathbf{x})}{f_\ell(\mathbf{x})}\right) &= \frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_\ell)}{\det(\boldsymbol{\Sigma}_i)} - \frac{1}{2}\left(\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^{\mathrm{T}}\boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\mu}_\ell\right) \\
&\quad -\frac{1}{2}\left\{\mathbf{x}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_\ell^{-1}\right)\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\mu}_\ell\right)\right\}
\end{aligned}$$

o sea, obtenemos una forma cuadrática en $\mathbf{x}$.

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

En este caso,

$$\log\left(\frac{f_i(\mathbf{x})}{f_\ell(\mathbf{x})}\right) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_\ell}{2})$$
$$= \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i) - \boldsymbol{\mu}_\ell^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_\ell)$$

Por lo tanto, si llamamos

$$L_i(\mathbf{x}) = \log\pi_i + \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i)$$

asigno $\mathbf{x}$ al grupo con mayor $L_i(\mathbf{x})$, o sea, clasifico $\mathbf{x} \in \mathcal{P}_i$ si

$$L_i(\mathbf{x}) = \max_{1\leq\ell\leq k} L_\ell(\mathbf{x})$$

Si $\pi_j = 1/k$ para todo j, esta regla de clasificación es la obtenida antes con las coordenadas discriminantes.

Caso $\mathbf{\Sigma}_\ell = \mathbf{\Sigma}$

En este caso, asigno $\mathbf{x}$ al grupo con mayor $L_i(\mathbf{x})$ donde llamamos

$$L_i(\mathbf{x}) = \log \pi_i + \boldsymbol{\mu}_i^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i)$$

Las funciones

$$d_{i\ell}(\mathbf{x}) = L_i(\mathbf{x}) - L_\ell(\mathbf{x}) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)^{\mathrm{T}} \mathbf{\Sigma}^{-1} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_i + \boldsymbol{\mu}_\ell)}{2} \right) + \log \pi_i - \log \pi_\ell$$

se llaman funciones discriminantes y $d_{i\ell}(\mathbf{x}) = -d_{\ell i}(\mathbf{x})$.

Sea $\boldsymbol{\alpha}_{i,\ell} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)$. Luego

$$d_{i\ell}(\mathbf{x}) = \boldsymbol{\alpha}_{i,\ell}^{\mathrm{T}} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_i + \boldsymbol{\mu}_\ell)}{2} \right) + \log \pi_i - \log \pi_\ell \ .$$

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

Si tenemos k poblaciones, sólo necesitamos encontrar $r = \min(p, k-1)$ direcciones de proyección en lugar de $\binom{k}{2}$.

Efectivamente, basta conocer $\boldsymbol{\alpha}_{i,i+1}$, $1 \leq i \leq k-1$ ya que

$$\boldsymbol{\alpha}_{i,i+2} = \boldsymbol{\alpha}_{i,i+1} - \boldsymbol{\alpha}_{i+1,i+2}$$

Ejemplo: Si $k = 3$ y obtenemos que
Si $L_1(\mathbf{x}) > L_2(\mathbf{x})$ y $L_2(\mathbf{x}) > L_3(\mathbf{x}) \Longrightarrow L_1(\mathbf{x}) > L_3(\mathbf{x})$.

Si además $p = 2$ cada ecuación $d_{i\ell}(\mathbf{x}) = 0$ es una recta y las tres rectas se cortan en el mismo punto ya que

$$d_{13}(\mathbf{x}) = L_1(\mathbf{x}) - L_3(\mathbf{x}) = L_1(\mathbf{x}) - L_2(\mathbf{x}) + (L_2(\mathbf{x}) - L_3(\mathbf{x})) = d_{12}(\mathbf{x}) + d_{23}(\mathbf{x})$$

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

Si $k = 2$, como $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

$$d_{12}(\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{2} \right) + \log \pi_1 - \log \pi_2$$

es la regla discriminante lineal de Fisher que clasifica en el grupo 1 si $d_{12}(\mathbf{x}) > 0$.

El hiperplano $d_{12}(\mathbf{x}) = 0$ determina un hiperplano que separa los dos grupos.

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

Veamos que si $\pi_j = 1/k$ para todo j, la regla de clasificación Bayes es la obtenida antes con las coordenadas discriminantes.

- $\boldsymbol{\Sigma}_{\mathrm{B}} = \sum_{i=1}^{k} \pi_i(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^{\mathrm{T}}$, $s = \text{rango}(\boldsymbol{\Sigma}_{\mathrm{B}})$
- $\mathbf{z} = \mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x} = (\mathbf{z}_1^{\mathrm{T}}, \mathbf{z}_2^{\mathrm{T}})^{\mathrm{T}}$ el vector de variables discriminantes con

$$\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p) = (\mathbf{A}_1, \mathbf{A}_2) \quad \text{donde} \quad \mathbf{A}_1 = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_s)$$

- $\mathbf{z}^{(1)} = \mathbf{A}_1^{\mathrm{T}}\mathbf{x}$, $\mathbf{z}^{(2)} = \mathbf{A}_2^{\mathrm{T}}\mathbf{x}$
- $\boldsymbol{\nu}_i = \mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}_i$, entonces $\boldsymbol{\nu}_i^{(2)} = \boldsymbol{\nu}^{(2)}$

Si $\mathbf{x}|G = i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ vimos que

$$\mathbf{z}|G = i \sim N(\boldsymbol{\nu}_i, \mathbf{I}_p)$$

o sea,

$$\mathbf{z}^{(1)}|G = i \sim N(\boldsymbol{\nu}_i^{(1)}, \mathbf{I}_s) \qquad \mathbf{z}^{(2)} \sim N(\boldsymbol{\nu}^{(2)}, \mathbf{I}_{p-s})$$

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

La regla que vimos asignaba $\mathbf{x}_0$ al grupo i si $\mathbf{v}_0 = \mathbf{A}_1^{\mathrm{T}}\mathbf{x}_0 \in \mathcal{G}_i$ donde

$$
\begin{aligned}
\mathcal{G}_i &= \{\mathbf{v} \in \mathbb{R}^s : \quad \|\mathbf{v} - \boldsymbol{\nu}_i^{(1)}\| < \|\mathbf{v} - \boldsymbol{\nu}_\ell^{(1)}\| \quad \forall \ell \neq i\} \\
&= \{\mathbf{v} \in \mathbb{R}^s : \quad (\boldsymbol{\nu}_i^{(1)})^{\mathrm{T}}(\mathbf{v} - \frac{1}{2}\boldsymbol{\nu}_i^{(1)}) > (\boldsymbol{\nu}_\ell^{(1)})^{\mathrm{T}}(\mathbf{v} - \frac{1}{2}\boldsymbol{\nu}_\ell^{(1)}) \quad \forall \ell \neq i\} \\
&= \{\mathbf{v} \in \mathbb{R}^s : \quad (\boldsymbol{\nu}_i^{(1)} - \boldsymbol{\nu}_\ell^{(1)})^{\mathrm{T}} \left(\mathbf{v} - \frac{\boldsymbol{\nu}_i^{(1)} + \boldsymbol{\nu}_\ell^{(1)}}{2}\right) > 0 \quad \forall \ell \neq i\}
\end{aligned}
$$

Esta regla es Bayes cuando $\pi_j = \frac{1}{k}$ para todo j.

Para una probabilidad a priori τ general tenemos que modificar $\mathcal{G}_i$ por $\mathcal{G}_{i,\tau}$, o sea, asigno $\mathbf{x}_0$ al grupo i si $\mathbf{v}_0 = \mathbf{A}_1^{\mathrm{T}}\mathbf{x}_0 \in \mathcal{G}_{i,\tau}$ donde

$$
\begin{aligned}
\mathcal{G}_{i,\tau} &= \{\mathbf{v} \in \mathbb{R}^s : \|\mathbf{v} - \boldsymbol{\nu}_i^{(1)}\|^2 - 2\log(\pi_i) < \|\mathbf{v} - \boldsymbol{\nu}_\ell^{(1)}\|^2 - 2\log(\pi_\ell) \ \forall \ell \neq i\} \\
&= \{\mathbf{v} \in \mathbb{R}^s : (\boldsymbol{\nu}_i^{(1)} - \boldsymbol{\nu}_\ell^{(1)})^{\mathrm{T}} \left(\mathbf{v} - \frac{\boldsymbol{\nu}_i^{(1)} + \boldsymbol{\nu}_\ell^{(1)}}{2}\right) > \log(\frac{\pi_\ell}{\pi_i}) \ \forall \ell \neq i\}
\end{aligned}
$$

$f_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),\ j = 1, 2$

Definición. Supongamos que $\mathbf{x}|\mathcal{P}_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2$.

Se dice que $\mathbf{x}$ está dada en forma canónica si $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$ y $\boldsymbol{\Sigma}_2$ es diagonal.

Llamaremos $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_2 = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ y $\boldsymbol{\mu}_2 = \boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^{\mathrm{T}}$.

Propiedad. Supongamos que $\mathbf{x}|\mathcal{P}_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2$ donde $\boldsymbol{\Sigma}_1 > 0$, $\boldsymbol{\Sigma}_2 > 0$. Entonces existe $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times p}$ no singular tal que $\mathbf{z} = \boldsymbol{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1)$ está en forma canónica.

$$\boldsymbol{\Gamma} = \left(\mathbf{C}^{-1}\right)^{\mathrm{T}} \beta$$

- $\boldsymbol{\Sigma}_1 = \mathbf{C}\,\mathbf{C}^{\mathrm{T}}$
- β son autovectores de $\mathbf{C}^{-1}\boldsymbol{\Sigma}_2\left(\mathbf{C}^{-1}\right)^{\mathrm{T}}$

Por lo tanto, si $c_{i|j} = 1$ si $i \neq j$ entonces la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_1$ si $\mathbf{z} = \mathbf{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1) \in \mathcal{G}_{1,0}$ donde

$$\mathcal{G}_{1,0} = \{\mathbf{z} \in \mathbb{R}^p : \quad \log\left(\frac{f_1(\mathbf{z})}{f_2(\mathbf{z})}\right) > \log\left(\frac{\pi_2}{\pi_1}\right)\} = \{\mathbf{z} \in \mathbb{R}^p : \ Q(\mathbf{z}) > 0\}$$

con

$$\begin{aligned}
Q(\mathbf{z}) &= \sum_{i=1}^{p} a_{ii} z_i^2 + \sum_{i=1}^{p} b_i z_i + c \\
a_{ii} &= \frac{1}{2}\left(\frac{1}{\lambda_i} - 1\right) \qquad\qquad b_i = -\frac{\delta_i}{\lambda_i} \\
c &= \log\left(\frac{\pi_1}{\pi_2}\right) + \frac{1}{2}\sum_{s=1}^{p} \log \lambda_s + \frac{1}{2}\sum_{s=1}^{p} \frac{\delta_s^2}{\lambda_s}
\end{aligned}$$

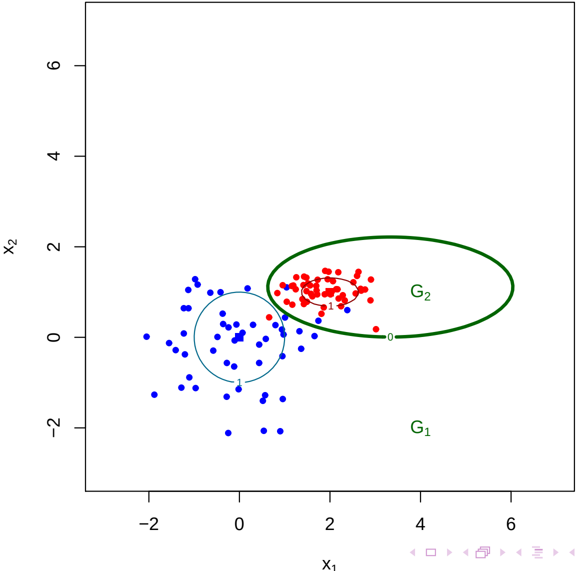
La ventaja de la forma canónica es que los términos de la forma $a_{ij}x_ix_j$ desaparecen lo que hace más fácil de entender la regla de clasificación.

Ejemplo $p = 2$

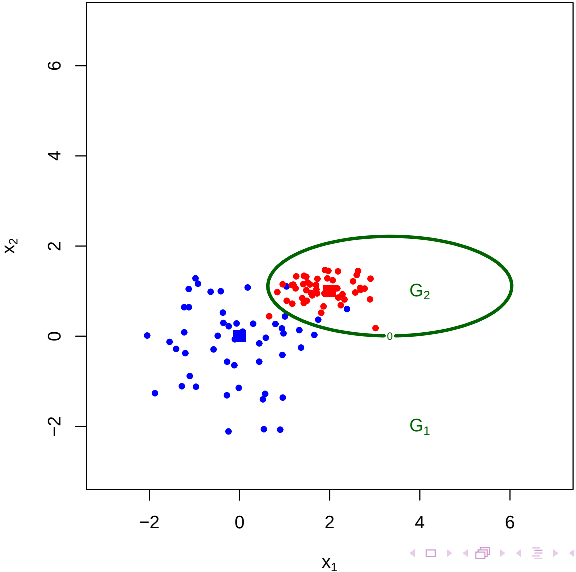
Tomemos $\pi_1 = \pi_2 = 1/2$ y los siguientes valores para $\boldsymbol{\delta}$ y $\boldsymbol{\Lambda}$

Caso	δ_1	δ_2	λ_1	λ_2	a_{11}	a_{22}	b_1	b_2	c
A	2	1	0.4	0.1	0.75	4.5	-5	-10	8.391
B	2	-1	4	0.25	-0.375	1.5	-0.5	4	2.5
C	3	1	4	1	-0.375	0	-0.75	-1	2.318
D	2	0	10	1	-0.45	0	-0.2	0	1.351

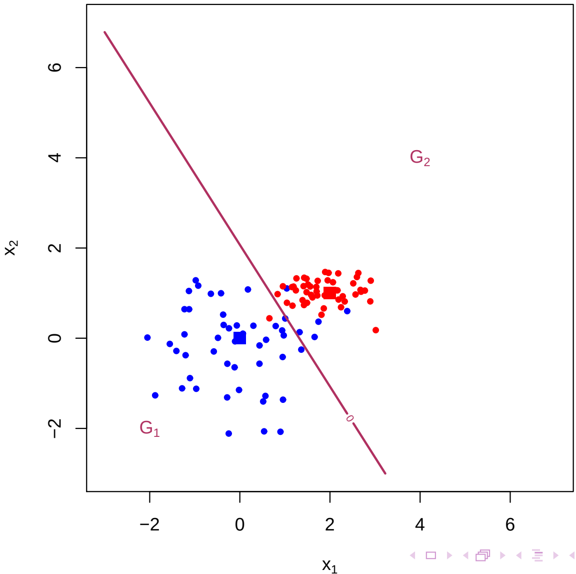
Ejemplo $p = 2$, Caso A



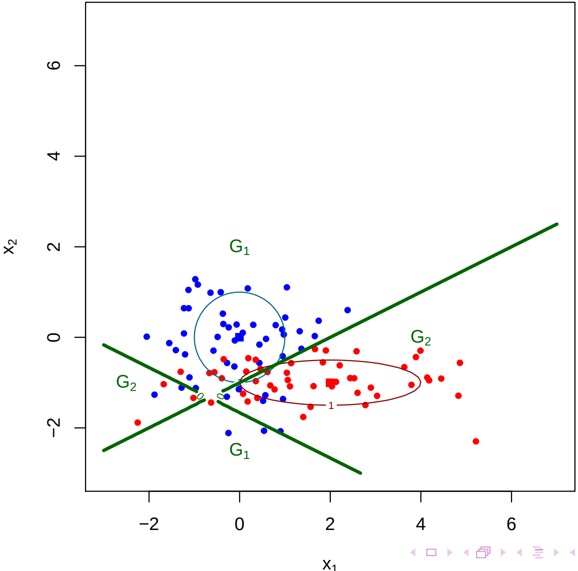
Ejemplo $p = 2$, Caso A



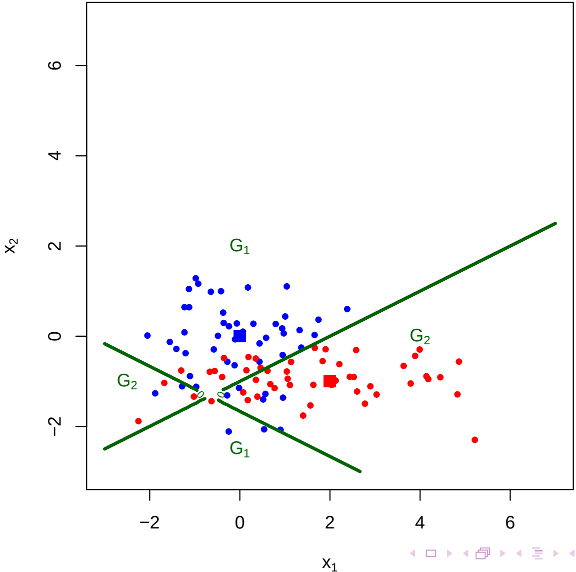
Ejemplo $p = 2$, Caso A: Regla Lineal como si $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ lo cual es FALSO



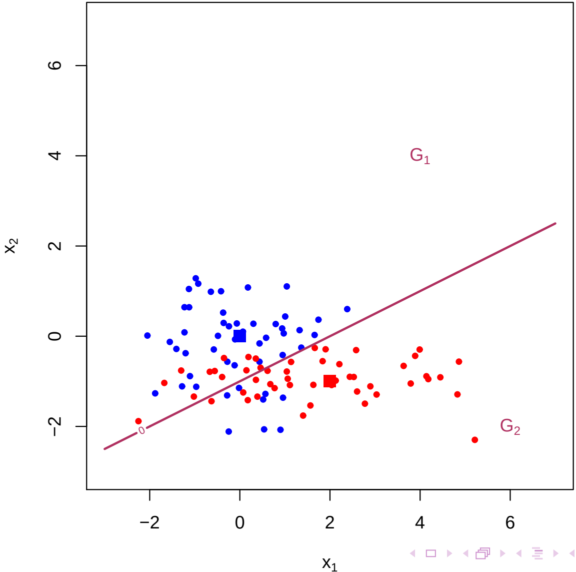
Ejemplo $p = 2$, Caso B



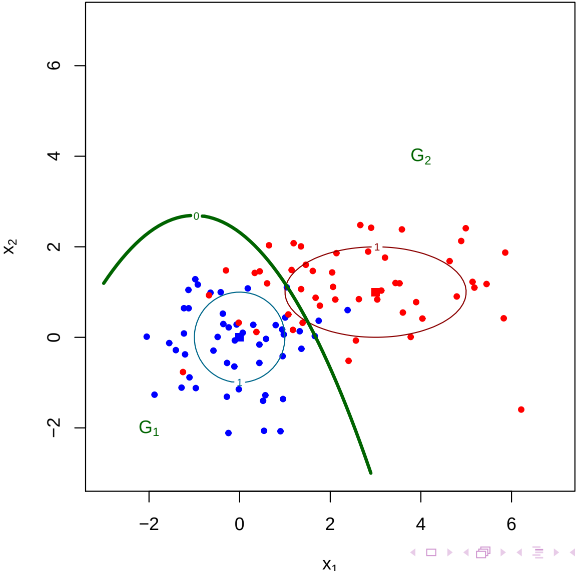
Ejemplo $p = 2$, Caso B



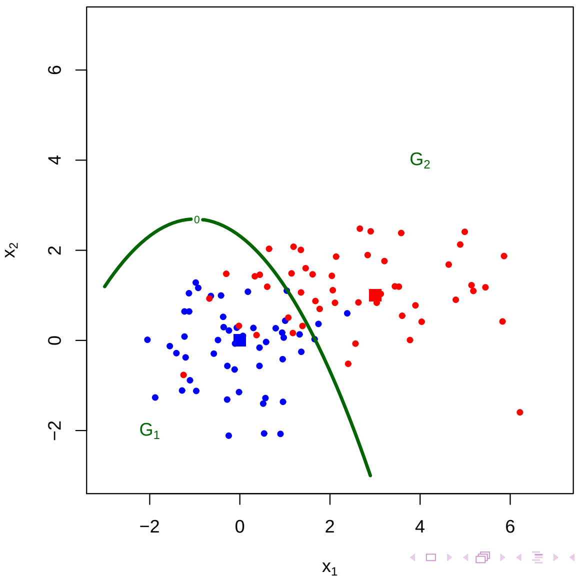
Ejemplo $p = 2$, Caso B: Regla Lineal como si $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ lo cual es FALSO



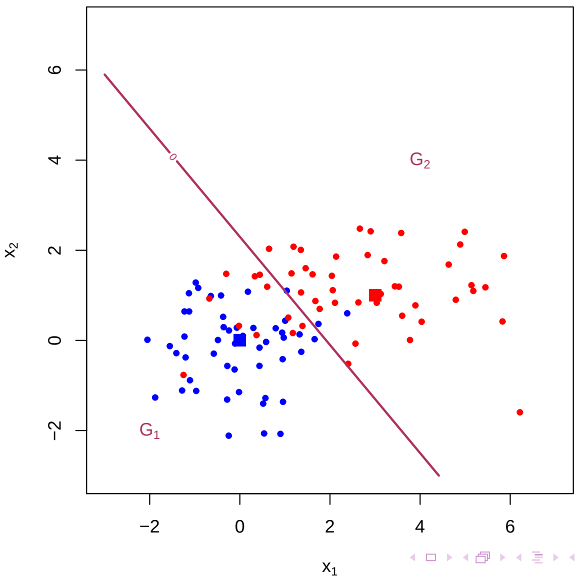
Ejemplo $p = 2$, Caso C, $a_{22} = 0$



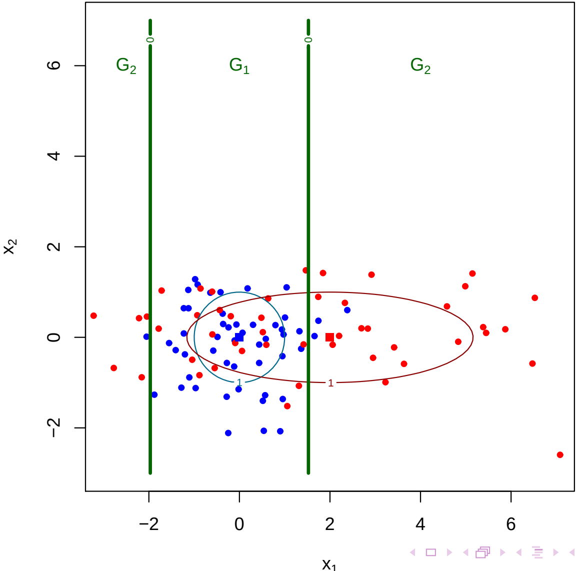
Ejemplo $p = 2$, Caso C, $a_{22} = 0$



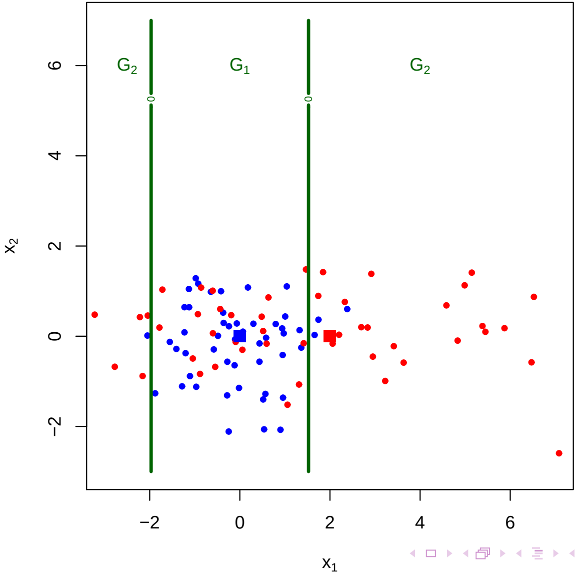
Ejemplo $p = 2$, Caso C: Regla Lineal como si $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ lo cual es FALSO



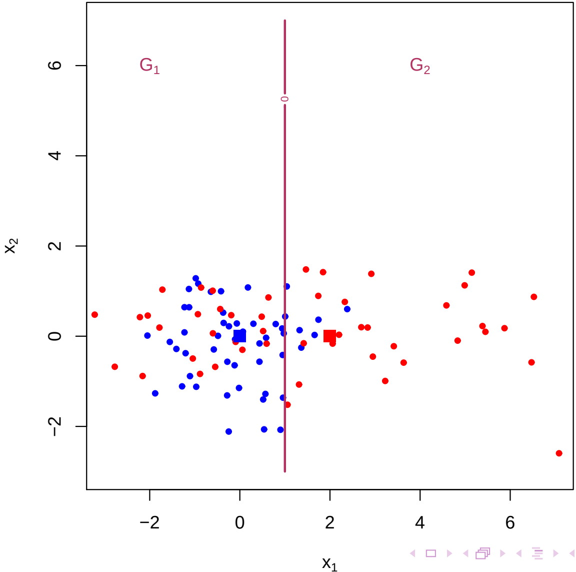
Ejemplo $p = 2$, Caso D, $a_{22} = 0$, $b_2 = 0$



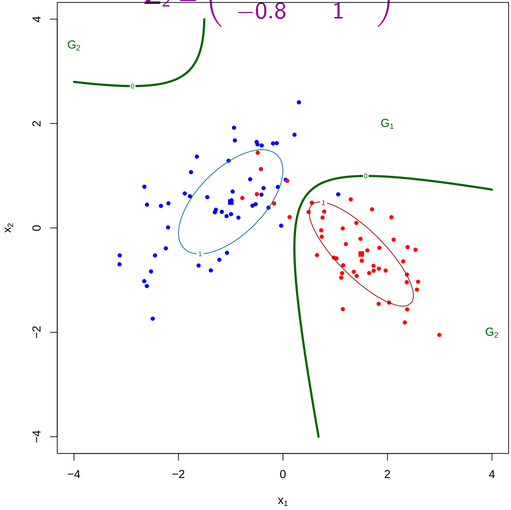
Ejemplo $p = 2$, Caso D, $a_{22} = 0$, $b_2 = 0$



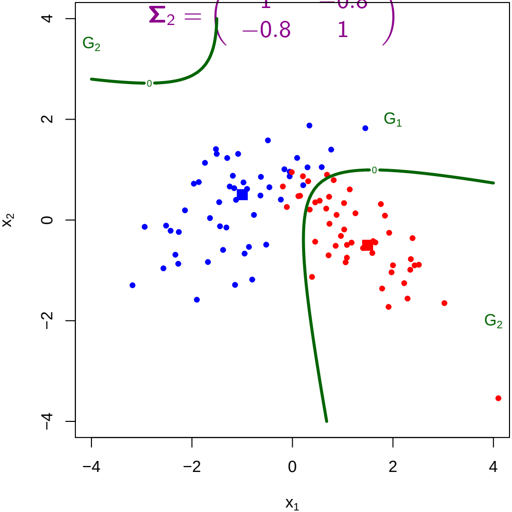
Ejemplo $p = 2$, Caso D: Regla Lineal como si $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ lo cual es FALSO



Ejemplo $\boldsymbol{\mu}_1 = (-1, 0.5)$, $\boldsymbol{\mu}_2 = (1.5, -0.5)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$



Ejemplo $\boldsymbol{\mu}_1 = (-1, 0.5)$, $\boldsymbol{\mu}_2 = (1.5, -0.5)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$

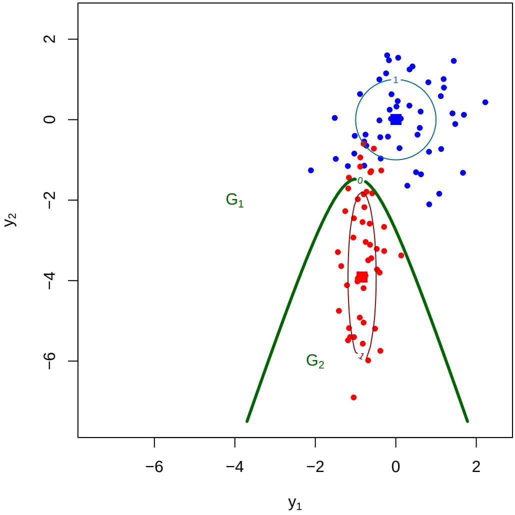


Ejemplo $\boldsymbol{\mu}_1 = (-1, 0.5)$, $\boldsymbol{\mu}_2 = (1.5, -0.5)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$,
$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$

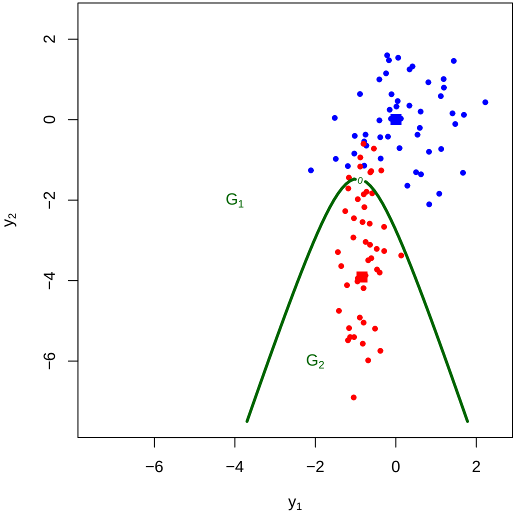
Datos transformados $\mathbf{z} = \boldsymbol{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1)$.

$$\begin{aligned}
\boldsymbol{\Gamma} &= \begin{pmatrix} -0.5590 & -1.1180 \\ -0.5590 & 1.1180 \end{pmatrix} \\
\lambda_1 &= 0.125 \qquad \lambda_2 = 4.5 \\
a_{11} &= 3.500 \qquad a_{22} = -0.3889 \\
b_1 &= 6.708 \qquad b_2 = 0.8696 \qquad c = 5.325
\end{aligned}$$

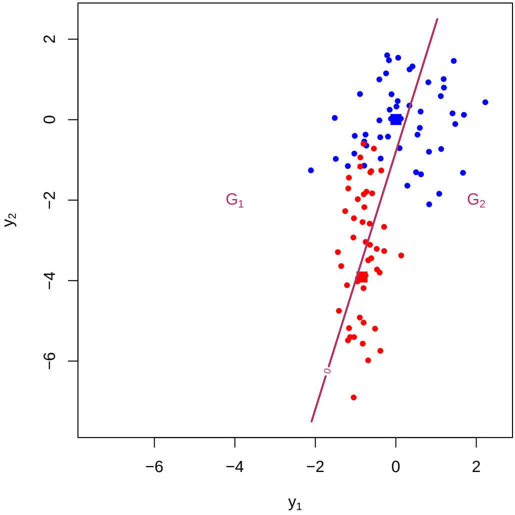
Datos transformados $\mathbf{z} = \mathbf{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1)$



Datos transformados $\mathbf{z} = \mathbf{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1)$



Datos transformados $\mathbf{z} = \mathbf{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1)$, Regla Lineal como si $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ lo cual es FALSO



Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, 2$

- $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$,
- la regla $G_0^\star$ clasifica en el grupo 1 si $\mathbf{x} \in \mathcal{G}_{1,0}$, con $\mathcal{G}_{1,0} = \{\mathbf{x} : d_{12}(\mathbf{x}) > 0\}$

$$d_{12}(\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{2}\right) + \log \pi_1 - \log \pi_2$$

$$\Delta_p^2 = \boldsymbol{\alpha}^{\mathrm{T}} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\alpha}$$

Luego

$$R(\mathcal{P}_1, G_0^\star) = \int_{\mathcal{G}_{2,0}} f_1(\mathbf{x}) d\mathbf{x} = 1 - \int_{\mathcal{G}_{1,0}} f_1(\mathbf{x}) d\mathbf{x}$$

Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, 2$

En $\mathcal{P}_1$, $\mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ luego

$$\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} \sim N(\boldsymbol{\alpha}^{\mathrm{T}}\mu_1, \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}) = N(\boldsymbol{\alpha}^{\mathrm{T}}\mu_1, \Delta_p^2)$$

de donde

$$\begin{aligned}
R(\mathcal{P}_1, G_0^\star) &= \Phi\left(\frac{\log\left(\frac{\pi_2}{\pi_1}\right) - \frac{1}{2}\Delta_p^2}{\Delta_p}\right) \\
R(\mathcal{P}_2, G_0^\star) &= \Phi\left(-\frac{\log\left(\frac{\pi_2}{\pi_1}\right) + \frac{1}{2}\Delta_p^2}{\Delta_p}\right) \\
e_{opt} = r(\tau, G_0^\star) &= \pi_1\Phi\left(\frac{\log\left(\frac{\pi_2}{\pi_1}\right) - \frac{1}{2}\Delta_p^2}{\Delta_p}\right) + \pi_2\Phi\left(-\frac{\log\left(\frac{\pi_2}{\pi_1}\right) + \frac{1}{2}\Delta_p^2}{\Delta_p}\right)
\end{aligned}$$

Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, 2$

Si $\pi_1 = \pi_2 = 1/2$ entonces

$$R(\mathcal{P}_1, G_0^\star) = R(\mathcal{P}_2, G_0^\star) = \Phi\left(-\frac{1}{2}\Delta_p\right)$$

y $G_0^\star$ es minimax.

En general, la regla minimax asigna al grupo 1 si $D(\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}}\left(\mathbf{x} - \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{2}\right) > \log(c)$ donde c se elige de modo que

$$\Phi\left(\frac{\log(c) - \frac{1}{2}\Delta_p^2}{\Delta_p}\right) = \Phi\left(\frac{-\log(c) - \frac{1}{2}\Delta_p^2}{\Delta_p}\right)$$

que tiene como solution $c = 1$ coincidiendo con el método de cociente de verosimilitud.

Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}),\ j = 1, \ldots, k$

- la regla $G_0^\star$ clasifica en el grupo i si $\mathbf{x} \in \mathcal{G}_{i,0}$, con

$$\mathcal{G}_{i,0} = \{\mathbf{x} : d_{i\ell}(\mathbf{x}) > 0 \forall \ell \neq i\} = \{\mathbf{x} : L_i(\mathbf{x}) = \max_{\ell} L_\ell(\mathbf{x})\}$$

$$d_{i\ell}(\mathbf{x}) = L_i(\mathbf{x}) - L_\ell(\mathbf{x}) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_i + \boldsymbol{\mu}_\ell)}{2}\right) + \log \pi_i - \log \pi_\ell$$

- $\Delta_{i\ell}^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)^{\mathrm{T}}\, \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)$

Luego

$$R(\mathcal{P}_i, G_0^\star) = \sum_{\ell \neq i} \int_{\mathcal{G}_{\ell,0}} f_i(\mathbf{x}) d\mathbf{x} = 1 - \int_{\mathcal{G}_{i,0}} f_i(\mathbf{x}) d\mathbf{x}$$

Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, \ldots, k$

En $\mathcal{P}_i$, $\mathbf{x} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ luego

$$d_{i\ell}(\mathbf{x}) \sim N\left(\frac{1}{2}\Delta^2_{i\ell} + \log \pi_i - \log \pi_\ell, \Delta^2_{i\ell}\right)$$

Más aún, el vector $\mathbf{d}_i(\mathbf{x}) = (d_{i\ell}(\mathbf{x}))_{\ell \neq i}$ tiene distribución normal $(k-1)-$variada y

$$\mathrm{Cov}(d_{i\ell}(\mathbf{x}), d_{ij}(\mathbf{x})) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

Si $k = 3$ se pueden calcular, fácilmente.

$f_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ con $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ desconocidos

Supongamos tener $\mathbf{x}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $1 \le j \le n_i$, $1 \le i \le k$ entonces estimamos $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}_i$ por

$$\widehat{\boldsymbol{\mu}}_i = \overline{\mathbf{x}}_i \qquad \mathbf{S}_i = \frac{\mathbf{Q}_i}{n_i - 1} = \frac{1}{n_i} \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}}$$

Clasificamos $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \widehat{\mathcal{G}}_{i,0}$ con

$$\widehat{\mathcal{G}}_{i,0} = \left\{\mathbf{x} \in \mathbb{R}^p : \quad \log\left(\frac{\widehat{f}_i(\mathbf{x})}{\widehat{f}_\ell(\mathbf{x})}\right) > \log\left(\frac{\pi_\ell}{\pi_i}\right) \quad \forall \ell \neq i\right\}$$

donde $\widehat{f}_i(\mathbf{x}) = f(\mathbf{x}, \widehat{\boldsymbol{\mu}}_i, \mathbf{S}_i)$ con $f(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\begin{aligned} \log\left(\frac{\widehat{f}_i(\mathbf{x})}{\widehat{f}_\ell(\mathbf{x})}\right) &= \frac{1}{2}\log\frac{\det(\mathbf{S}_\ell)}{\det(\mathbf{S}_j)} - \frac{1}{2}\left(\widehat{\boldsymbol{\mu}}_i^{\mathrm{T}}\mathbf{S}_i^{-1}\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_\ell^{\mathrm{T}}\mathbf{S}_\ell^{-1}\widehat{\boldsymbol{\mu}}_\ell\right) \\ &\quad -\frac{1}{2}\left\{\mathbf{x}^{\mathrm{T}}\left(\mathbf{S}_i^{-1} - \mathbf{S}_\ell^{-1}\right)\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\left(\mathbf{S}_i^{-1}\widehat{\boldsymbol{\mu}}_i - \mathbf{S}_\ell^{-1}\widehat{\boldsymbol{\mu}}_\ell\right)\right\} \end{aligned}$$

$f_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ con $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}$ desconocidos

Supongamos tener $\mathbf{x}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $1 \le j \le n_i$, $1 \le i \le k$ entonces estimamos $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}$ por

$$\widehat{\boldsymbol{\mu}}_i = \overline{\mathbf{x}}_i \qquad \mathbf{S} = \frac{1}{n-k}\sum_{i=1}^{k}\mathbf{Q}_i = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{i,j}-\overline{\mathbf{x}}_i)(\mathbf{x}_{i,j}-\overline{\mathbf{x}}_i)^{\mathrm{T}}$$

Clasificamos $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \widehat{\mathcal{G}}_{i,0}$ con

$$\widehat{\mathcal{G}}_{i,0} = \{\mathbf{x} \in \mathbb{R}^p : \quad \widehat{L}_i(\mathbf{x}) > \widehat{L}_\ell(\mathbf{x}) \quad \forall \ell \neq i\} = \{\mathbf{x} \in \mathbb{R}^p : \quad \widehat{d}_{i\ell}(\mathbf{x}) > 0\}$$

donde

$$\widehat{L}_i(\mathbf{x}) = \log \pi_i + \widehat{\boldsymbol{\mu}}_i^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_i)$$

$$\begin{aligned}\widehat{d}_{i\ell}(\mathbf{x}) &= \widehat{L}_i(\mathbf{x}) - \widehat{L}_\ell(\mathbf{x}) \\ &= \log\left(\frac{\pi_i}{\pi_\ell}\right) + (\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_\ell)^{\mathrm{T}}\mathbf{S}^{-1}\left(\mathbf{x} - \frac{\widehat{\boldsymbol{\mu}}_i + \widehat{\boldsymbol{\mu}}_\ell}{2}\right)\end{aligned}$$

Estimación de los errores de clasificación $k = 2$, $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$

En el caso normal clasificamos $\mathbf{x} \in \mathcal{P}_1$ si $\widehat{d}_{12}(\mathbf{x}) > 0$,

$$\begin{aligned}\widehat{d}_{12}(\mathbf{x}) &= \log\left(\frac{\pi_1}{\pi_2}\right) + (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}} \mathbf{S}^{-1}\left(\mathbf{x} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) \\ &= \log\left(\frac{\pi_1}{\pi_2}\right) + \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\mathbf{x} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right)\end{aligned}$$

Por lo tanto, como si $\mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y si llamamos $\sigma^2 = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{\Sigma} \widehat{\boldsymbol{\alpha}}$, $\mathbf{X}_1 = (\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n_1})$ y $\mathbf{X}_2 = (\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,n_2})$, entonces

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\mathbf{x} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) |(\mathbf{X}_1, \mathbf{X}_2) &\sim N\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\boldsymbol{\mu}_1 - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right), \sigma^2\right) \\ \widehat{d}_{12}(\mathbf{x}) &\sim N\left(\widehat{d}_{12}(\boldsymbol{\mu}_1), \sigma^2\right)\end{aligned}$$

Estimación de los errores de clasificación $k = 2$, $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$

$$
\begin{aligned}
e_{1,act} &= \int_{\widehat{\mathcal{G}}_{2,0}} f_1(\mathbf{x})d\mathbf{x} = \int_{\widehat{d}_{12}(\mathbf{x})<0} f_1(\mathbf{x})d\mathbf{x} \\
&= \mathbb{P}\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\mathbf{x} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) + \log\left(\frac{\pi_1}{\pi_2}\right) < 0 | \mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right) \\
&= \Phi\left(-\frac{\widehat{d}_{12}(\boldsymbol{\mu}_1)}{\sigma}\right) = \Phi\left(\frac{\widehat{d}_{21}(\boldsymbol{\mu}_1)}{\sigma}\right)
\end{aligned}
$$

Por lo tanto,

$$
e_{act} = \pi_1 \Phi\left(\frac{\widehat{d}_{21}(\boldsymbol{\mu}_1)}{\sigma}\right) + \pi_2 \Phi\left(\frac{\widehat{d}_{12}(\boldsymbol{\mu}_2)}{\sigma}\right)
$$

Estimación de los errores de clasificación $k = 2$, $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$

$$\widehat{e}_{act} = \pi_1 \Phi\left(\frac{\log\left(\frac{\pi_2}{\pi_1}\right) - \frac{1}{2}D_p^2}{D_p}\right) + \pi_2 \Phi\left(-\frac{\log\left(\frac{\pi_2}{\pi_1}\right) + \frac{1}{2}D_p^2}{D_p}\right)$$

donde

$$D_p^2 = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right) = \left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right)^{\mathrm{T}} \mathbf{S}^{-1}\left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right)$$

Si π_1 y π_2 son desconocidos se estiman por n_1/n y n_2/n, con $n = n_1 + n_2$. Si $\pi_1 = \pi_2 = 0.5$, entonces

$$\widehat{e}_{act} = \Phi\left(-\frac{1}{2}D_p\right)$$

Estimación de los errores de clasificación $k = 2$, $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$

Como $\mathbb{E}\mathcal{F}_{\nu_1,\nu_2}(\lambda) = \nu_2(\nu_1 + \lambda)/[\nu_1(\nu_2 - 2)]$ si $\nu_2 > 2$ y

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T_0^2 \sim \mathcal{F}_{p,n_1+n_2-p-1}(\lambda^2)$$

con $\lambda^2 = \frac{n_1\, n_2}{n_1+n_2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ y

$$T_0^2 = \frac{n_1\, n_2}{n_1 + n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}}\mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

tenemos que

$$\mathbb{E}D_p^2 = \frac{n-2}{n-p-3}\left(\Delta_p^2 + \frac{pn}{n_1 n_2}\right)$$

D_p^2 sobreestima a Δ_p^2 y por lo tanto, $\widehat{e}_{act}$ subestima el error real e_{act}.

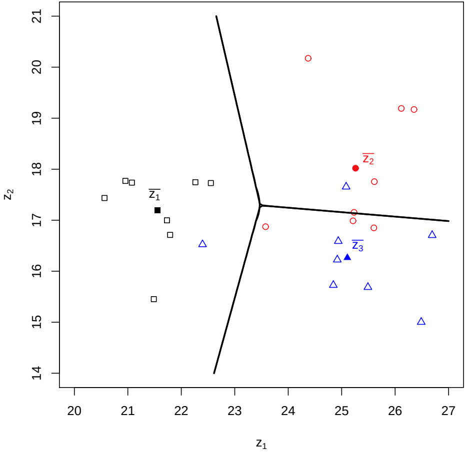
Estimación de los errores de clasificación $k = 2$, $\mathbf{x}|\mathbf{x} \in \mathcal{P}_j \sim f_j = f_j(\mathbf{x}, \boldsymbol{\theta}_j)$

Esto se explica también pues recordemos que probamos el siguiente resultado

Lema. Sea $\widehat{r}_i(\mathbf{x}) = f_i(\mathbf{x}, \widehat{\boldsymbol{\theta}}_i)\pi_i$ un estimador insesgado de $r_i(\mathbf{x})$, o sea, $\mathbb{E}_{\widehat{\boldsymbol{\theta}}}\widehat{r}_i(\mathbf{x})|\mathbf{x} = r_i(\mathbf{x})$ para casi todo $\mathbf{x}$. Luego

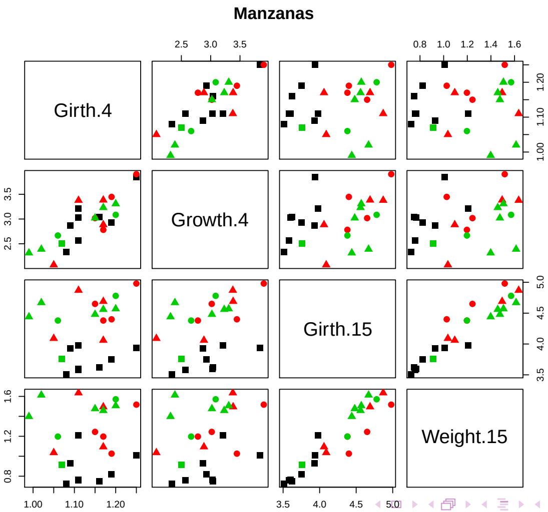
$$\mathbb{E}\widehat{e}_{act} \leq e_{opt} < \mathbb{E}e_{act}$$

Ejemplo Manzanos

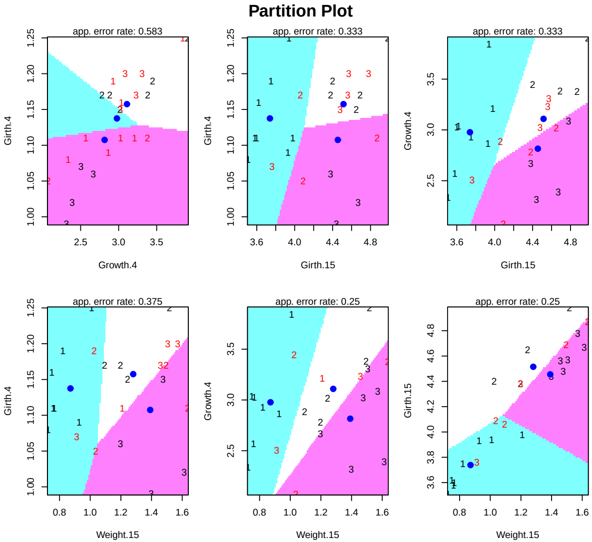


Ejemplo Manzanos

■: Asignado al grupo 1, • asignado al grupo 2, ▲ asignado al grupo 3



Ejemplo Manzanos: Regla LDA: Bien clasificados, Mal clasificados



Comparación entre LDF y QDF, $k = 2$

- En general la decisión de elegir entre la regla lineal (LDF) y cuadrática (QDF) se hace en base al resultado del test para $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Si el test rechaza se usa QDF.

- A pesar de que esta decisión es razonable ya que LDF es óptima si H_0 es cierta, hay un número importante de trabajos que muestran que aunque no lo sea, LDF es tan buena como QDF.

- Uno podría basar su decisión en elegir el método que da menor error aparente, $\widehat{e}_{app}$, lo cual es peligroso ya que este estimador del error subestima el error real, e_{act}. El cálculo del error actual esperado sólo puede hacerse por simulación.

- Una opción es utilizar el e_{cv} para elegir entre ambas reglas.

Comparación entre LDF y QDF, $k = 2$

En general, LDF es buena para pequeños alejamientos de $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. El mejor comportamiento de QDF depende del tamaño de las muestras y de la dimensión.

- Para n_1 y n_2 pequeñas y $p \leq 6$ hay poca pérdida al elegir LDF.
- Para $n_1, n_2 \leq 25$ y p grande y/o diferencias entre $\mathbf{\Sigma}_1$ y $\mathbf{\Sigma}_2$, LDF es preferible.
- Sin embargo, cuando p grande y $\mathbf{\Sigma}_1$ y $\mathbf{\Sigma}_2$ son muy distintas, la probabilidades de mala clasificación $e_{1,act}$ y $e_{2,act}$ pueden ser muy grandes para un uso práctico

Comparación entre LDF y QDF, $k = 2$

- Si $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ y la diferencia es grande y $p > 6$, QDF es mucho mejor que LDF si el tamaño de muestra es grande.
  - Se recomienda para $p = 4$, $n_1 = n_2 = 25$
  - 25 observaciones adicionales cada dos dimensiones, o sea, para $p = 6, 8, 10$ se necesitan $n_1 = n_2 = 50, 75, 100$
- Para $n_i \geq 100$ y p moderado los resultados asintóticos que favorecen QDF se alcanzan bastante rápido.
- QDF se deteriora rápidamente si p crece porque $\mathbf{S}_i$ no provee una estimación confiable de $\mathbf{\Sigma}_i$ si p es una fracción moderada de n_i.

Modelos parsimoniosos: Ejemplo

Sean $\mathbf{y}_{ij}$, $1 \leq j \leq n_i$, $1 = 1, 2$ medidas correspondientes a hijos varones mellizos

- $n_1 = 49$ gemelos (monocigota)
- $n_2 = 40$ mellizos (heterocigota)

Se midieron

- altura,
- ancho de cadera,
- Circunferencia del pecho

del primer y segundo nacido, dando origen a un vector de dimensión 6.

Modelos parsimoniosos: Ejemplo

Las coordenadas del vector $\mathbf{y}_{ij}$ son

- Las tres primeras mediciones corresponden al primer hijo, o sea, $y_{i,j,1}, y_{i,j,2}, y_{i,j,3}$ son las medidas de altura, ancho de cadera y circunferencia de pecho del primer hijo en el par de mellizos j del grupo i.
- Las tres últimas mediciones corresponden al segundo hijo, o sea, $y_{i,j,4}, y_{i,j,5}, y_{i,j,6}$ son las medidas de altura, ancho de cadera y circunferencia de pecho del segundo hijo en el par de mellizos j del grupo i.

Es natural suponer que tanto para el primer como para el segundo hijo las medidas tomadas tienen igual media es decir

$$\mathbb{E}y_{i,j,\ell} = \mathbb{E}y_{i,j,\ell+3}\,, \qquad \ell = 1, 2, 3$$

Modelos parsimoniosos: Ejemplo

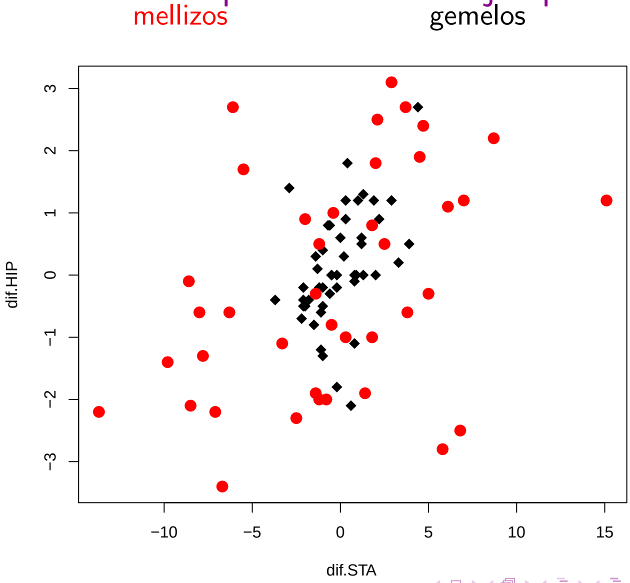
Definamos

- $x_{ij,1} = y_{i,j,1} - y_{i,j,4}$
  $= \text{diferencia de altura entre los } j\text{−ésimos gemelos del grupo } i$
- $x_{ij,2} = y_{i,j,2} - y_{i,j,5}$
  $= \text{diferencia de ancho de cadera entre los } j\text{−ésimos gemelos}$ del grupo i

Luego,

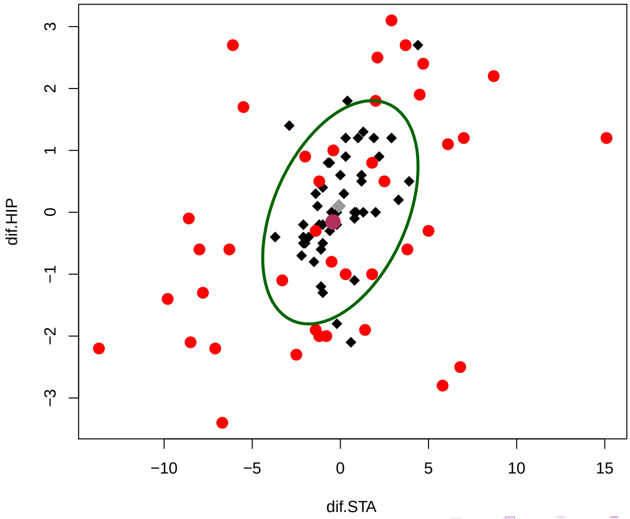
$$\mathbb{E}\mathbf{x}_{ij} = \mathbf{0} \qquad i = 1, 2, \quad 1 \leq j \leq n_i$$

Modelos parsimoniosos: Ejemplo



Modelos parsimoniosos: Ejemplo

mellizos gemelos



Modelos parsimoniosos

- $\mathbf{x}_i \sim N(\mathbf{0}_p, \rho_i \boldsymbol{\Sigma})$ con $\rho_1 = 1$, $1 = 1, 2$

La regla de discriminación en este caso es:

$$\mathcal{G}_{1,0} = \{\mathbf{x} : Q(\mathbf{x}) > 0\}$$

con $Q(\mathbf{z}) = \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x} + c$

$$\begin{aligned} \mathbf{A} &= \frac{1-\rho}{2\rho} \boldsymbol{\Sigma}^{-1} \\ c &= \log\left(\frac{\pi_1}{\pi_2}\right) + \frac{p}{2} \log(\rho) \end{aligned}$$

Modelos parsimoniosos

Sean $\mathbf{x}_{ij} \sim N(\mathbf{0}_p, \rho_i \boldsymbol{\Sigma})$ con $\rho_1 = 1$, $1 \leq j \leq n_i$, $1 = 1, 2$. Definamos

$$\mathbf{M}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}_{ij}^{\mathrm{T}}$$

Los EMV de ρ y $\boldsymbol{\Sigma}$ resuelven

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n} \left(n_1 \mathbf{M}_1 + \frac{1}{\widehat{\rho}} n_2 \mathbf{M}_2 \right) \\ \widehat{\rho} &= \frac{1}{p} \operatorname{traza}(\widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{M}_2) \end{aligned}$$

con lo cual

$$\widehat{\mathbf{A}} = \frac{1 - \widehat{\rho}}{2\widehat{\rho}} \widehat{\boldsymbol{\Sigma}}^{-1} \qquad \widehat{c} = \log\left(\frac{\pi_1}{\pi_2}\right) + \frac{p}{2} \log(\widehat{\rho})$$

Modelos parsimoniosos: Ejemplo

Sean $\mathbf{y}_{ij}$, $1 \leq j \leq n_i$, $1 = 1, 2$. $n_1 = 49$ son los gemelos y $n_2 = 40$ los mellizos. Definimos:

- $x_{ij,1} = y_{i,j,1} - y_{i,j,4}$
- $x_{ij,2} = y_{i,j,2} - y_{i,j,5}$

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 3.8166 & 0.6462 \\ 0.6462 & 0.6473 \end{pmatrix}$$

$$\widehat{\rho} = 7.070$$

Modelos parsimoniosos: Ejemplo

mellizos gemelos