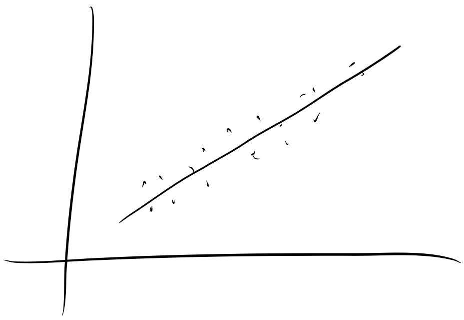
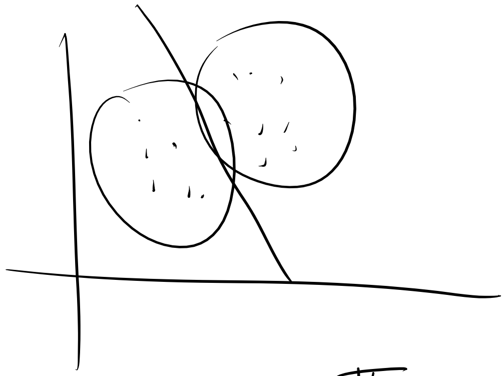
		Clasif. Π_1	Clasif. Π_2
Verdadero	Π_1	0	$C(2\|1)$
	Π_2	$C(1\|2)$	0

El costo esperado o promedio de clasif

$$ECM = C(2|1)\cdot p(2|1)p_1 + C(1|2)\, p(1|2)\, p_2 \to \text{pepe}$$

Buscamos minimizar ese costo.





$p(2|1)$ prob. de clasif. un objeto en Π_2 siendo Π_1

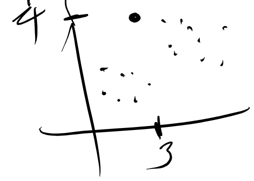
$p(1|2)$ " " " " " " Π_1 siendo Π_2

$$= \int_{R_1} f_2(x)\,dx$$

Nuestro objetivo es minimizar a pepe

Teorema:

Las regiones R_1 y R_2 que minimizan a pepe se definen por los x que satisfacen



$x = (3,4$

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{C(1|2)}{C(2|1)}\right)\cdot\left(\frac{p_2}{p_1}\right)$$

$f_1(x) = 3x_1 + 2x_2$

$3\cdot 3 + 2\cdot 4$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{C(1|2)}{C(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$f_2(x) \quad \backslash\ C(2|1) \quad \backslash\ p_1$

Ejercicio: 11.3 libro. (Tarea)

Para usar esta técnica necesitamos:

1) Las funciones de densidad (por lo menos evaluadas en los puntos de interes x)

2) Los costos de clasificación errónea

3) Las probabilidades previas.

Casos especiales

si $\frac{p_2}{p_1} = 1 \qquad (p_2 = p_1)$

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)}$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)}$$

2) $c(1|2) = c(2|1) \Longrightarrow \frac{c(1|2)}{c(2|1)} = 1$

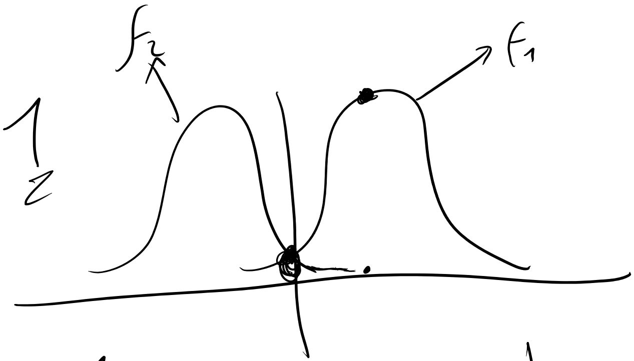
$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1}$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}$$

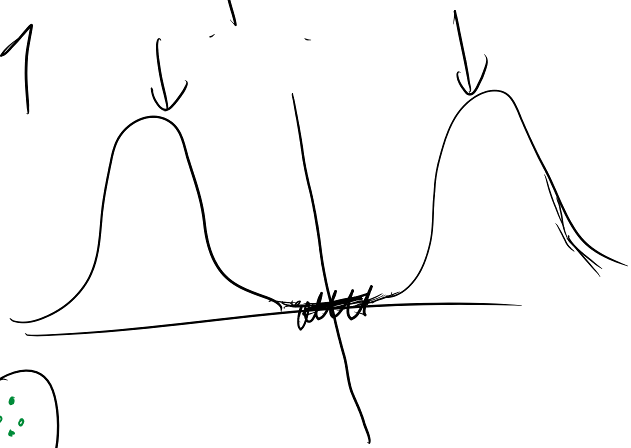
$f_2(x) \qquad p_1$

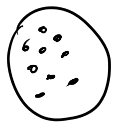
3) costos iguales y p_1 y p_2 también

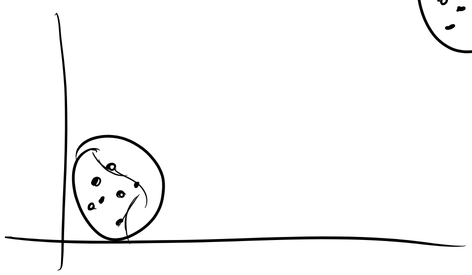
$$R_1 : \frac{f_1(x)}{f_2(x)} \geq 1$$

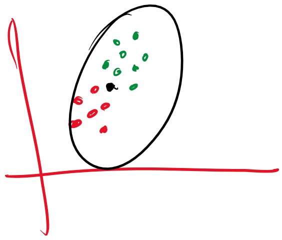


$$R_2 : \frac{f_1(x)}{f_2(x)} < 1$$









Clasificar en 2 pobl. normales multivariadas

Bajo la supo. de normalidad la clasif. es mucho más sencilla y resulta más práctica

supongamos $f_1(x)$, $f_2(x)$ PDF normales multiv. con medias μ_1 y μ_2, cov Σ_1, Σ_2 respect.

Caso 1 $\quad \Sigma_1 = \Sigma_2 = \Sigma$

$$f_i(x) = \frac{1}{(2\pi)^{p/2} \, |\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2} (x-\mu_i)' \Sigma^{-1} (x-\mu_i)\right)$$

supongamos μ_1, μ_2 y Σ conocidos

Teorema

$$R_1: \exp\left(-\frac{1}{2}(\mathbb{X}-\mu_1)'\Sigma^{-1}(\mathbb{X}-\mu_1)+\frac{1}{2}(\mathbb{X}-\mu_2)'\cdot\Sigma^{-1}(\mathbb{X}-\mu_2)\right) \geq \frac{c(1|2)}{c(2|1)}\cdot\frac{p_2}{p_1}$$

$$R_2: \boxed{} < \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$

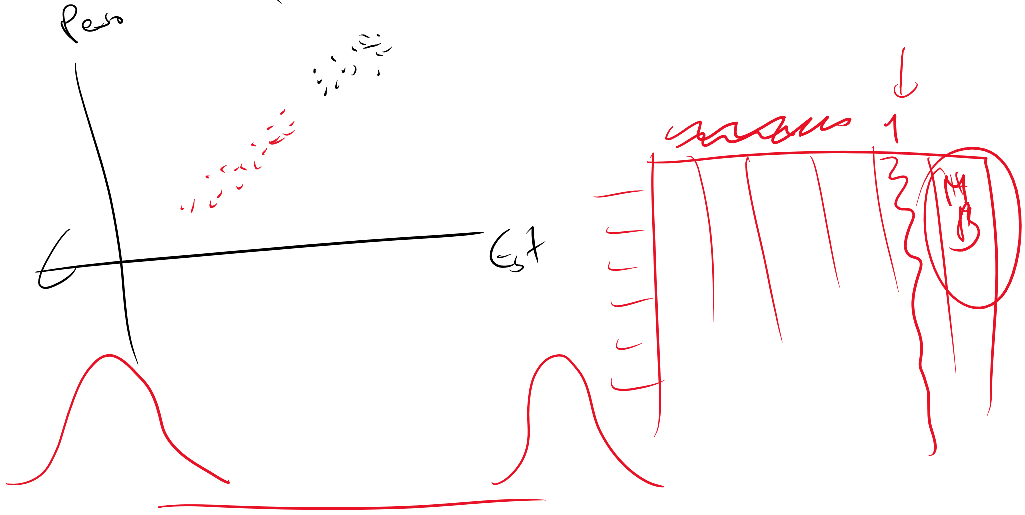
Teorema:
Sean Π_1, Π_2 pobl. con PDF normal multiv., entonces la regla de clasif. que minimiza el error es:

Clasificar $\mathbb{X}_0$ como Π_1 si:

$$(\mu_1-\mu_2)'\Sigma^{-1}\mathbb{X}_0-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)$$

$$\geq \ln\left(\frac{c(1|2)}{c(2|1)}\cdot\frac{p_2}{p_1}\right)$$

Demostración (libro - TAREA)



En la mayoría de los casos μ_1, μ_2 y Σ son

En la mayoría de los casos μ_1, μ_2 y Σ son desconocidos y debemos estimarlos con los posibles errores que eso implica.

Supongamos que hay n_1 obs. de $\mathbb{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ de Π_1

y que hay n_2 obs de $\mathbb{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ de Π_2

de variables

con $n_1 + n_2 - 2 \geq \hat{p}$.

tenemos $\mathbb{X}_1 = \begin{pmatrix} \mathbb{X}_{11}' \\ \vdots \\ \mathbb{X}_{1n_1}' \end{pmatrix}$, $\mathbb{X}_2 = \begin{pmatrix} \mathbb{X}_{21}' \\ \vdots \\ \mathbb{X}_{2n_2}' \end{pmatrix}$

con medias $\bar{\mathbb{X}}_1$ y $\bar{\mathbb{X}}_2$

y cov. muestrales

$$S_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\mathbb{X}_{1i} - \bar{\mathbb{X}}_1)(\mathbb{X}_{1i} - \bar{\mathbb{X}}_1)'$$

S_2 = igual pero con $\mathbb{X}_2$

$$S_{pooled} = \left(\frac{n_1 - 1}{n_1 + n_2 - 2} \right) S_1 + \left(\frac{n_2 - 1}{n_1 + n_2 - 2} \right) S_2$$

→ estimador insesgado de Σ

Entonces

Clasificar $\mathbb{X}_0$ como Π_1 si:

$$\boxed{(\bar{\mathbb{X}}_1 - \bar{\mathbb{X}}_2)' S_{pooled}^{-1} \mathbb{X}_0 - \frac{1}{2} (\bar{\mathbb{X}}_1 - \bar{\mathbb{X}}_2)' S_{pooled}^{-1} (\bar{\mathbb{X}}_1 + \bar{\mathbb{X}}_2)} \geq \ln \left(\frac{c(1|2)}{\ldots} \right) \frac{p_2}{\ldots}$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\boxed{\text{Jose}}} \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\frac{p_2}{p_1}$$

Si $\dfrac{c(1|2)}{c(2|1)}\dfrac{p_2}{p_1} = 1 \Longrightarrow \ln(1) = 0$

Entonces, clasificamos como π_1 si $\boxed{\text{Jose}} \geq 0$

Sea $\hat{y}_0 = (\bar{X}_1 - \bar{X}_2)\,S_{pooled}^{-1}\,X_0$

$= \hat{a}'X_0$

y también $\hat{m} = \frac{1}{2}(\bar{X}_1 - \bar{X}_2)'S_{pooled}^{-1}(\bar{X}_1 + \bar{X}_2) = \frac{1}{2}(\underbrace{\bar{Y}_1}_{\hat{a}'\bar{X}_1} + \underbrace{\bar{Y}_2}_{\hat{a}'\bar{X}_2})$

Usaremos $\hat{Y}$ y $\hat{m}$ para clasificar

(Si $\hat{Y}_0$ es mayor o menor al promed. de medias $\bar{Y}_1$ y $\bar{Y}_2$)

costs) This example is adapted from a study [4] concerned with the detection of hemophilia A carriers. (See also Exercise 11.32.)

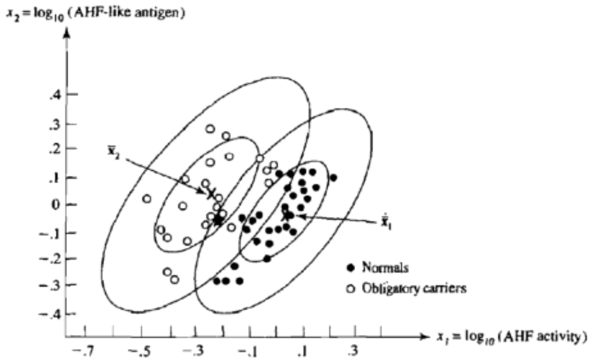
To construct a procedure for detecting potential hemophilia A carriers, blood samples were assayed for two groups of women and measurements on the two variables,

$$X_1 = \log_{10}(\text{AHF activity})$$
$$X_2 = \log_{10}(\text{AHF-like antigen})$$

recorded. ("AHF" denotes antihemophilic factor.) The first group of $n_1 = 30$ women were selected from a population of women who did not carry the hemophilia gene. This group was called the *normal* group. The second group of $n_2 = 22$ women was selected from known hemophilia A carriers (daughters of hemophiliacs, mothers with more than one hemophilic son, and mothers with one hemophilic son and other hemophilic relatives). This group was called the *obligatory carriers*. The pairs of observations (x_1, x_2) for the two groups are plotted in Figure 11.4. Also shown are estimated contours containing 50% and 95% of the probability for bivariate normal distributions centered at $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, respectively. Their common covariance matrix was taken as the pooled sample covariance matrix $\mathbf{S}_{\text{pooled}}$. In this example, bivariate normal distributions seem to fit the data fairly well.

The investigators (see [4]) provide the information

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} -.2483 \\ .0262 \end{bmatrix}$$



and

$$\mathbf{S}_{\text{pooled}}^{-1} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Therefore, the equal costs and equal priors discriminant function [see (11-19)] is

$$\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}$$

$$= [.2418 \quad -.0652] \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\boxed{= 37.61x_1 - 28.92x_2}$$

Moreover,

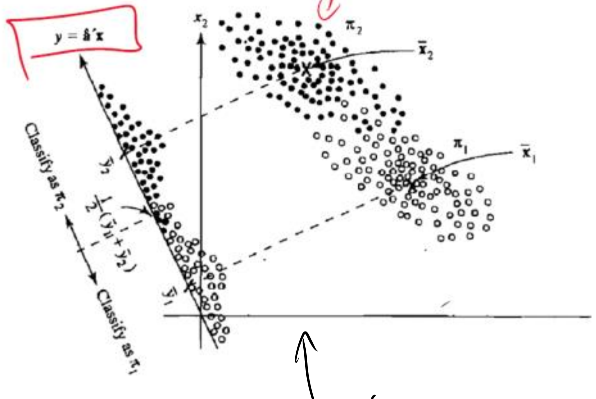
$$\bar{y}_1 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_1 = [37.61 \quad -28.92] \begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix} = .88$$

$$\bar{y}_2 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_2 = [37.61 \quad -28.92] \begin{bmatrix} -.2483 \\ .0262 \end{bmatrix} = -10.10$$

and the midpoint between these means [see (11-20)] is

and the midpoint between these means [see (11-20)] is

$$\hat{m} = \tfrac{1}{2}(\bar{y}_1 + \bar{y}_2) = \tfrac{1}{2}(.88 - 10.10) = -4.61$$



Método de discriminante de Fisher.