

# Análisis Multivariante

## Tema 8. Análisis Discriminante y Clasificación

**uc3m** | Universidad Carlos III de Madrid

Aurea Grané  
Dpto. Estadística  
[aurea.grane@uc3m.es](mailto:aurea.grane@uc3m.es)

## Análisis Discriminante y Clasificación

1. Introducción
2. Clasificación en dos poblaciones
3. Clasificación en poblaciones normales
4. Clasificación en más de dos poblaciones
5. Clasificadores no paramétricos

Como ya comentamos en el Tema 4 (MANOVA), el estadístico británico Ronald Aylmer Fisher (1890-1962) fue el inventor del análisis discriminante.

Dio la primera solución al problema de la clasificación, inventando un método general, basado en el análisis de la varianza.

En 1937 visitó la India invitado por Mahalanobis (1893-1972), estadístico indio inventor de la distancia de Mahalanobis. En la India, Fisher descubrió la relación entre la distancia de Mahalanobis y sus resultados en análisis discriminante. Consiguió unificar estas ideas y relacionarlas con los trabajos de Hotelling (1885-1973) sobre el contraste de medias de poblaciones multivariantes.

C. R. Rao, estudiante de Mahalanobis, extendió el análisis discriminante de Fisher a más de dos poblaciones.

## 1. Introducción

Supongamos que tenemos  $g$  poblaciones conocidas  $\Omega_1, \dots, \Omega_g$  y en cada una de ellas observamos una muestra de cierto vector de variables cuantitativas de interés  $\mathbf{X} = (X_1, \dots, X_p)'$ .

El **análisis discriminante** se ocupa de describir, mediante las variables  $X_i$ , los rasgos diferenciales entre las poblaciones.

Se trata de encontrar *funciones discriminantes o reglas de decisión*  $h = h(x_1, \dots, x_p)$  cuyos valores en los distintos grupos (o poblaciones) estén lo más separados posible. Es decir, buscamos funciones  $h$  sencillas que permitan asignar cada uno de los individuos a una población concreta  $\Omega_\alpha$ ,  $\alpha = 1, \dots, g$ , minimizando la tasa de error en dicha asignación.

La más conocida, es la regla discriminante lineal de Fisher, donde  $h$  es una función lineal de  $\mathbf{x} = (x_1, \dots, x_p)'$ .

## El problema de la clasificación (**typicality**)

Consideremos un nuevo individuo  $\omega$  sobre el cual se pueden medir las variables,  $X_1, \dots, X_p$ , es decir, que se conocen los valores que toma el individuo  $\omega$  para las variables  $X_1, \dots, X_p$ ,  $\mathbf{x} = (x_1, \dots, x_p)'$ , donde  $x_i = X_i(\omega)$ , para  $i = 1, \dots, p$ .

Sin embargo, se desconoce la población de la cual procede el individuo  $\omega$ .

El **problema de clasificación** trata de asignar este individuo a alguna de las poblaciones  $\Omega_\alpha$  conocidas, para  $\alpha = 1, \dots, g$ .

Para ello se utilizan las funciones discriminantes construidas a partir de la muestra.

## 2. Clasificación en $g = 2$ poblaciones

### 2.1. Discriminador lineal

Sean  $\mu_1, \mu_2$  los vectores de medias de las poblaciones  $\Omega_1, \Omega_2$ , respectivamente. Sea  $\Sigma$  la matriz de covarianzas común para ambas poblaciones. Sea  $\omega$  el individuo a clasificar, para el cual se ha observado  $\mathbf{x} = (x_1, \dots, x_p)'$

El *criterio geométrico*, consiste en asignar el individuo  $\omega$  a la población más próxima, utilizando la distancia de Mahalanobis:

$$\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_i) = (\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i), \quad i = 1, 2.$$

La regla de decisión es la siguiente:

- $\omega$  se asigna a  $\Omega_1$  si  $\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_1) < \delta_M^2(\mathbf{x}, \boldsymbol{\mu}_2)$ ,
- $\omega$  se asigna a  $\Omega_2$  en caso contrario.

A partir de la diferencia  $\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_2) - \delta_M^2(\mathbf{x}, \boldsymbol{\mu}_1)$ , se construye la función discriminante lineal

$$L(\mathbf{x}) = \left( \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right)' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

y se expresa la regla de decisión en función de ésta:

- $\omega$  se asigna a  $\Omega_1$  si  $L(\mathbf{x}) > 0$ ,
- en caso contrario, se asigna  $\omega$  a  $\Omega_2$ .

Esta función discriminante que acabamos de construir es el **discriminador lineal de Fisher**.

**Ejercicio 1:** Obtener  $L(\mathbf{x})$  como una diferencia entre  $\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_2)$  y  $\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_1)$ .

**Ejercicio 1:** Obtener  $L(\mathbf{x})$  como una diferencia entre  $\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_2)$  y  $\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_1)$ .

**Solución:** Empezamos escribiendo la diferencia entre las dos distancias de Mahalanobis:

$$\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_2) - \delta_M^2(\mathbf{x}, \boldsymbol{\mu}_1) = (\mathbf{x} - \boldsymbol{\mu}_2)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) - (\mathbf{x} - \boldsymbol{\mu}_1)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1)$$

$$= \mathbf{x}' \Sigma^{-1} \mathbf{x} - \mathbf{x}' \Sigma^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}'_2 \Sigma^{-1} \mathbf{x} + \boldsymbol{\mu}'_2 \Sigma^{-1} \boldsymbol{\mu}_2 - \mathbf{x}' \Sigma^{-1} \mathbf{x} + \mathbf{x}' \Sigma^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}'_1 \Sigma^{-1} \mathbf{x} - \boldsymbol{\mu}'_1 \Sigma^{-1} \boldsymbol{\mu}_1$$

Simplificando y usando que  $\mathbf{x}' \Sigma^{-1} \boldsymbol{\mu}_i = \boldsymbol{\mu}'_i \Sigma^{-1} \mathbf{x}$ , para  $i = 1, 2$ , al ser  $\Sigma$  una matriz simétrica, tenemos que:

$$\begin{aligned} &= -2\mathbf{x}' \Sigma^{-1} \boldsymbol{\mu}_2 + \boldsymbol{\mu}'_2 \Sigma^{-1} \boldsymbol{\mu}_2 + 2\mathbf{x}' \Sigma^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}'_1 \Sigma^{-1} \boldsymbol{\mu}_1 \\ &= 2\mathbf{x}' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \underbrace{\boldsymbol{\mu}'_2 \Sigma^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}'_1 \Sigma^{-1} \boldsymbol{\mu}_1}_{=(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}. \end{aligned}$$

Por tanto:

$$\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_2) - \delta_M^2(\mathbf{x}, \boldsymbol{\mu}_1) = (2\mathbf{x}' + (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)') \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

y multiplicando a ambos lados de la igualdad por  $1/2$ :

$$\frac{1}{2} (\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_2) - \delta_M^2(\mathbf{x}, \boldsymbol{\mu}_1)) = \left( \mathbf{x} + \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right)' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = L(\mathbf{x}).$$

## Clasificación cuando los parámetros son estimados

En las aplicaciones prácticas,  $\mu_1$ ,  $\mu_2$  y  $\Sigma$  son desconocidas y se deberán estimar a partir de muestras de tamaños  $n_1$ ,  $n_2$  de las dos poblaciones  $\Omega_1$  y  $\Omega_2$ .

Sean  $\bar{x}_1$ ,  $\bar{x}_2$  y  $S_1$ ,  $S_2$  los vectores de medias y las matrices de covarianzas muestrales.

La versión muestral del discriminador lineal de Fisher es

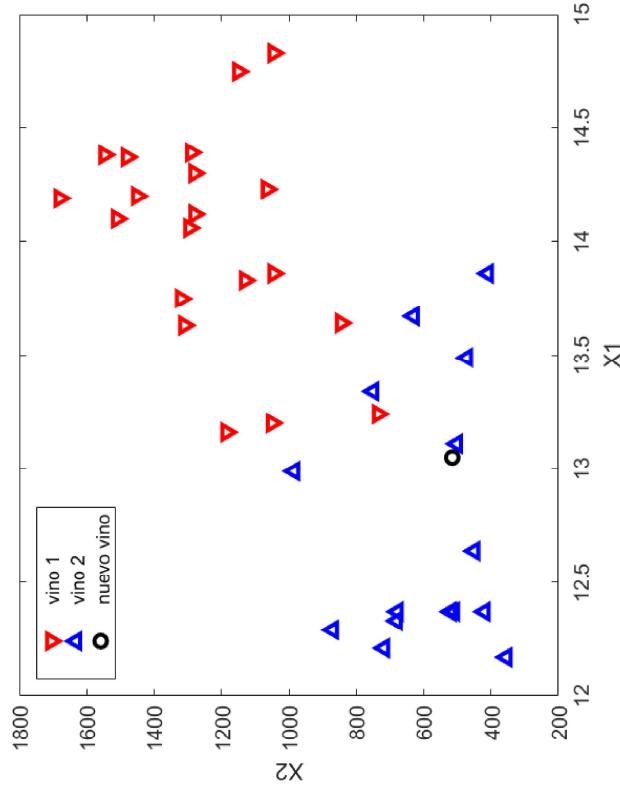
$$\hat{L}(\mathbf{x}) = \left( \mathbf{x} - \frac{1}{2}(\bar{x}_1 + \bar{x}_2) \right)' S_P^{-1} (\bar{x}_1 - \bar{x}_2),$$

donde  $S_P = (n_1 S_1 + n_2 S_2)/(n_1 + n_2 - 2)$ , es la *pooled within matrix*.

De forma análoga,  $\omega$  se asigna a  $\Omega_1$  si  $\hat{L}(\mathbf{x}) > 0$ ; se asignará a  $\Omega_2$  en caso contrario.

**Ejemplo 1.** Un enólogo analiza dos componentes  $X_1$  y  $X_2$  en sendas muestras de dos tipos de vinos (Datos de Newman *et al.*, 1998).

Vino 1	Vino 2	
	$X_1$	$X_2$
14.23	1065	12.37
13.20	1050	12.33
13.16	1185	12.64
14.37	1480	13.67
13.24	735	12.37
14.20	1450	12.17
14.39	1290	12.37
14.06	1295	13.11
14.83	1045	12.37
13.86	1045	13.34
14.10	1510	12.21
14.12	1280	12.29
13.75	1320	13.86
14.75	1150	13.49
14.38	1547	12.99
13.63	1310	
14.30	1280	
13.83	1130	
14.19	1680	
13.64	845	



Clasificar mediante el discriminador lineal de Fisher la nueva observación  $\mathbf{x} = (13.05, 515)'$ .

Empezamos calculando los vectores de medias y las matrices de covarianzas muestrales. La primera muestra consta de  $n_1 = 20$  observaciones del primer tipo de vino:

$$\bar{\mathbf{x}}_1 = (14.0115, 1234.6000)', \quad \mathbf{S}_1 = \begin{pmatrix} 0.2115 & 42.6801 \\ 42.6801 & 52947.0400 \end{pmatrix},$$

y la segunda muestra tiene  $n_2 = 15$  observaciones del segundo tipo de vino:

$$\bar{\mathbf{x}}_2 = (12.7720, 596.6667)', \quad \mathbf{S}_2 = \begin{pmatrix} 0.3400 & -6.4900 \\ -6.4900 & 33019.9524 \end{pmatrix}.$$

La matriz de covarianzas ponderada es:

$$\mathbf{S}_p = (20\mathbf{S}_1 + 15\mathbf{S}_2)/33 = \begin{pmatrix} 0.2827 & 22.9167 \\ 22.9167 & 47098.1844 \end{pmatrix}$$

La estimación del discriminador lineal de Fisher para la nueva observación  $\mathbf{x} = (13.05, 515)'$  es:

$$\begin{aligned}\hat{L}(\mathbf{x}) &= \left( \mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \right)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \\ &\left( \begin{pmatrix} 13.05 \\ 515 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 26.7835 \\ 1831.2667 \end{pmatrix} \right)' \begin{pmatrix} 0.2827 & 22.9167 \\ 22.9167 & 47098.1844 \end{pmatrix}^{-1} \begin{pmatrix} 1.2395 \\ 637.9333 \end{pmatrix} \\ &= -5.9288.\end{aligned}$$

Puesto que  $\hat{L}(\mathbf{x}) < 0$ , asignaremos la nueva observación al segundo tipo de vino.

Observad que nos hemos creído una hipótesis que no hemos comprobado ¿cuál es? ¿qué contraste deberíamos realizar?

**Criterio geométrico:** La regla del discriminador lineal de Fisher equivale a asignar el nuevo individuo a la población más próxima según la distancia de Mahalanobis:

$$\delta_M^2(\mathbf{x}, \bar{\mathbf{x}}_i) = (\mathbf{x} - \bar{\mathbf{x}}_i)' \mathbf{S}_p^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i), \quad i = 1, 2.$$

En nuestro caso, tenemos que:

$$\delta_M^2(\mathbf{x}, \bar{\mathbf{x}}_1) = 12.3709, \quad \delta_M^2(\mathbf{x}, \bar{\mathbf{x}}_2) = 0.5134.$$

Puesto que  $\delta_M^2(\mathbf{x}, \bar{\mathbf{x}}_2) < \delta_M^2(\mathbf{x}, \bar{\mathbf{x}}_1)$ , la nueva observación se asigna al segundo tipo de vino.

*Relación con el MANOVA:* La regla del discriminador linear de Fisher equivale a usar el primer eje canónico como función clasificadora.

El fichero vinos.txt contiene los datos de este ejercicio, junto con la nueva observación a clasificar (última final de fichero de datos). Empezamos realizando el MANOVA mediante la función canp.m al conjunto de datos formado por los primeros 35 individuos:

```
data=load('vino.txt');
X=data(1:35,:);
n=[20 15];
[mY,V,B,W,percent,Test1,texto1,Test2,texto2]=canp(X,n);
V =
    1.0148
    0.0035
Test2= 4.0096   3.0000  0.2604 % no se rechaza la igualdad de covarianzas
```

El primer (y, en este caso, único) eje canónico es  $Y = 1.0148X_1 + 0.0035X_2$ .

El vector mY contiene las coordenadas de los centroides de cada grupo expresadas en función de este eje:

```
mY =
    18.5089 % centroide del primer grupo
    15.0346 % centroide del segundo grupo
```

Ahora sólo queda representar el nuevo individuo a clasificar y asignarlo al centroide más cercano (según la distancia euclídea):

```
new=data(36,:);
new*V=15.0330
```

Por tanto, el nuevo individuo se asigna al grupo 2.

## 2.2. Regla de máxima verosimilitud

Sean  $\Omega_1$  y  $\Omega_2$  dos poblaciones y  $\mathbf{X} = (X_1, \dots, X_p)'$  un vector con distribución de probabilidad conocida, dependiente de un parámetro  $\theta$  que toma el valor  $\theta_1$  si  $\mathbf{X} \in \Omega_1$  y  $\theta_2$  si  $\mathbf{X} \in \Omega_2$ .

Sea  $\mathbf{x} = (x_1, \dots, x_p)'$  el vector de observaciones de  $\mathbf{X}$  sobre un individuo  $\omega$ . La probabilidad o verosimilitud de la observación  $\mathbf{x}$  en  $\Omega_i$  es  $\mathcal{L}_i(\mathbf{x}) = f(x_1, \dots, x_p; \theta_i)$ .

La regla discriminante de máxima verosimilitud consiste en asignar  $\omega$  a la población  $\Omega_i$  para la cual la verosimilitud de la observación es mayor. Esta regla tiene asociada la siguiente función discriminante

$$V(\mathbf{x}) = \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}).$$

Si  $V(\mathbf{x}) > 0$ ,  $\omega$  se asigna a  $\Omega_1$ ; En caso contrario,  $\omega$  se asignará a  $\Omega_2$ .

## 2.3. Regla de Bayes

Con la misma notación del apartado 2.2., se consideran dos poblaciones  $\Omega_1$ ,  $\Omega_2$ ,  $\mathbf{X} = (X_1, \dots, X_p)'$  un vector con distribución de probabilidad conocida y  $\mathbf{x} = (x_1, \dots, x_p)'$  el vector de observaciones de  $\mathbf{X}$  sobre un individuo  $\omega$ . La probabilidad o verosimilitud de la observación  $\mathbf{x}$  en  $\Omega_i$  es  $P(\mathbf{x}|\omega \in \Omega_i) = \mathcal{L}_i(\mathbf{x}) = f(x_1, \dots, x_p; \theta_i)$ ,  $i = 1, 2$ .

Si además se conocen las probabilidades a priori de que  $\omega$  pertenezca a cada una de las poblaciones:

$$q_1 = P(\omega \in \Omega_1), \quad q_2 = P(\omega \in \Omega_2), \quad q_1 + q_2 = 1,$$

aplicando el teorema de Bayes se puede calcular la probabilidad a posteriori de que  $\omega$  pertenezca a  $\Omega_i$ , ( $i = 1, 2$ ):

$$P(\omega \in \Omega_i | \mathbf{x}) = \frac{P(\mathbf{x} | \omega \in \Omega_i) P(\omega \in \Omega_i)}{P(\mathbf{x})} = \frac{q_i \mathcal{L}_i(\mathbf{x})}{q_1 \mathcal{L}_1(\mathbf{x}) + q_2 \mathcal{L}_2(\mathbf{x})}.$$

La regla discriminante de Bayes consiste en asignar  $\omega$  a la población  $\Omega_i$  para la que la probabilidad a posteriori  $P(\omega \in \Omega_i | \mathbf{x})$  es mayor.

La regla de Bayes tiene asociada la siguiente función discriminante, que se conoce como **discriminador de Bayes**:

$$\begin{aligned} B(\mathbf{x}) &= \log P(\omega \in \Omega_1 | \mathbf{x}) - \log P(\omega \in \Omega_2 | \mathbf{x}) = \dots \\ &= \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}) + \log(q_1/q_2). \end{aligned}$$

Si  $B(\mathbf{x}) > 0$ ,  $\omega$  se asigna a  $\Omega_1$ ; en caso contrario,  $\omega$  se asignará a  $\Omega_2$ .

### Propiedades:

1. Cuando  $q_1 = q_2 = 1/2$ , entonces  $B(\mathbf{x}) = V(\mathbf{x})$ .
2. La regla de Bayes minimiza la probabilidad de clasificación errónea.

**Observación:** Para estimar  $q_1$  y  $q_2$  se utiliza la evidencia empírica, es decir,  $\hat{q}_1 = n_1/(n_1 + n_2)$ ,  $\hat{q}_2 = n_2/(n_1 + n_2)$ , donde  $n_i$  es el tamaño muestral de la muestra que proviene de la población  $\Omega_i$ , para  $i = 1, 2$ .

**Ejercicio 2:** Obtener el discriminador de Bayes como la diferencia (en logaritmos) de las dos probabilidades a posteriori  $P(\omega \in \Omega_1 | \mathbf{x})$  y  $P(\omega \in \Omega_2 | \mathbf{x})$ .

**Solución:** Empezamos escribiendo la diferencia:

$$\begin{aligned} &\log P(\omega \in \Omega_1 | \mathbf{x}) - \log P(\omega \in \Omega_2 | \mathbf{x}) \\ &= \log \left( \frac{q_1 \mathcal{L}_1(\mathbf{x})}{q_1 \mathcal{L}_1(\mathbf{x}) + q_2 \mathcal{L}_2(\mathbf{x})} \right) - \log \left( \frac{q_2 \mathcal{L}_2(\mathbf{x})}{q_1 \mathcal{L}_1(\mathbf{x}) + q_2 \mathcal{L}_2(\mathbf{x})} \right) \\ &= \log(q_1 \mathcal{L}_1(\mathbf{x})) - \log(q_1 \mathcal{L}_1(\mathbf{x}) + q_2 \mathcal{L}_2(\mathbf{x})) - \log(q_2 \mathcal{L}_2(\mathbf{x})) + \log(q_1 \mathcal{L}_1(\mathbf{x}) + q_2 \mathcal{L}_2(\mathbf{x})) \\ &= \log q_1 + \log \mathcal{L}_1(\mathbf{x}) - \log q_2 - \log \mathcal{L}_2(\mathbf{x}) = \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}) = B(\mathbf{x}). \end{aligned}$$

### 3. Clasificación en poblaciones normales

Supongamos ahora que:

$$\mathbf{X} = (X_1, \dots, X_p)' \sim N_p(\boldsymbol{\mu}_1, \Sigma_1) \text{ en } \Omega_1$$

$$\mathbf{X} = (X_1, \dots, X_p)' \sim N_p(\boldsymbol{\mu}_2, \Sigma_2) \text{ en } \Omega_2,$$

entonces, la función de verosimilitud de  $\mathbf{x} = (x_1, \dots, x_p)$  es:

$$\mathcal{L}_i(\mathbf{x}) = \frac{|\Sigma_i|^{-1/2}}{(2\pi)^{p/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right\}, \quad i = 1, 2.$$

#### 3.1. Matrices de covarianzas poblacionales iguales

Si  $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$  y  $\Sigma_1 = \Sigma_2 = \Sigma$ , entonces:

- a) Los clasificadores de máxima verosimilitud y lineal de Fisher coinciden:

$$\begin{aligned} V(\mathbf{x}) &= \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}) \\ &= \frac{1}{2} ((\mathbf{x} - \boldsymbol{\mu}_2)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) - (\mathbf{x} - \boldsymbol{\mu}_1)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1)) \\ &= L(\mathbf{x}). \end{aligned}$$

- b) Si  $\mathbf{x} \in \mathbb{R}^p$  es el vector de observaciones de  $\mathbf{X}$  sobre un individuo  $\omega$ , que proviene de alguna de las poblaciones  $\Omega_i$ , para  $i = 1, 2$ , entonces el discriminador lineal de Fisher tiene distribución normal:

$$L(\mathbf{x}) = \left( \mathbf{x} - \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right)' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\mathbf{x} - \boldsymbol{\mu})' \mathbf{a} = \mathbf{a}' (\mathbf{x} - \boldsymbol{\mu}),$$

donde  $\mathbf{a} = \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$  y  $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ .

Su varianza y esperanza son:  $\text{var}(L(\mathbf{x})) = \text{var}(\mathbf{a}' (\mathbf{x} - \boldsymbol{\mu})) = \mathbf{a}' \Sigma \mathbf{a} = M^2$ ,

$$\mathbb{E}(L(\mathbf{x})) = \mathbf{a}' \mathbb{E}(\mathbf{x} - \boldsymbol{\mu}) = \begin{cases} \frac{1}{2} \mathbf{a}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2} M^2, & \text{si } \omega \in \Omega_1, \\ -\frac{1}{2} \mathbf{a}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = -\frac{1}{2} M^2, & \text{si } \omega \in \Omega_2, \end{cases}$$

Por tanto,

$$L(\mathbf{x}) \sim N\left(\frac{1}{2}M^2, M^2\right) \text{ si } \omega \in \Omega_1, L(\mathbf{x}) \sim N\left(-\frac{1}{2}M^2, M^2\right) \text{ si } \omega \in \Omega_2.$$

**Ejercicio 3:** Demostrar que si  $\mathbf{x} \in \mathbb{R}^p$  es el vector de observaciones de  $\mathbf{X}$  sobre un individuo  $\omega$ , que proviene de alguna de las poblaciones  $\Omega_i$ , donde  $\mathbf{X} = (X_1, \dots, X_p)' \sim N_p(\boldsymbol{\mu}_i, \Sigma)$ , para  $i = 1, 2$ , entonces

$$L(\mathbf{x}) \sim N\left(\frac{1}{2}M^2, M^2\right) \text{ si } \omega \in \Omega_1, L(\mathbf{x}) \sim N\left(-\frac{1}{2}M^2, M^2\right) \text{ si } \omega \in \Omega_2.$$

**Solución:** Consideremos  $L(\mathbf{x}) = \mathbf{a}'(\mathbf{x} - \boldsymbol{\mu})$ , donde  $\mathbf{a} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$  y  $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ .

Empezamos calculando la varianza de  $L(\mathbf{x})$ :

$$\begin{aligned} \text{var}(L(\mathbf{x})) &= \text{var}(\mathbf{a}'(\mathbf{x} - \boldsymbol{\mu})) = \mathbf{a}' \text{var}(\mathbf{x} - \boldsymbol{\mu}) \mathbf{a} = \mathbf{a}' \text{var}(\mathbf{x}) \mathbf{a} = \mathbf{a}' \Sigma \mathbf{a} \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \Sigma^{-1} \Sigma \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \delta_M^2(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = M^2. \end{aligned}$$

La esperanza de  $L(\mathbf{x})$  es:

$$E(L(\mathbf{x})) = E(\mathbf{a}'(\mathbf{x} - \boldsymbol{\mu})) = \mathbf{a}' E(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{a}'(E(\mathbf{x}) - \boldsymbol{\mu})$$

Si  $\omega \in \Omega_1$ ,  $\mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \Sigma)$ , por tanto:

$$\mathbf{a}'(E(\mathbf{x}) - \boldsymbol{\mu}) = \mathbf{a}'(\boldsymbol{\mu}_1 - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)) = \frac{1}{2}\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}M^2,$$

y si  $\omega \in \Omega_2$ ,  $\mathbf{x} \sim N_p(\boldsymbol{\mu}_2, \Sigma)$ , por tanto:

$$\mathbf{a}'(E(\mathbf{x}) - \boldsymbol{\mu}) = \mathbf{a}'(\boldsymbol{\mu}_2 - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)) = \frac{1}{2}\mathbf{a}'(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = -\frac{1}{2}M^2.$$

Puesto que  $L(\mathbf{x})$  tiene distribución de probabilidad conocida, puede calcularse la **probabilidad de clasificación errónea**.

Se dice que el individuo  $\omega$  se clasifica erróneamente cuando se asigna a la población  $\Omega_1$  y en realidad proviene de  $\Omega_2$ , o bien, cuando se asigna a la población  $\Omega_2$  y en realidad proviene de  $\Omega_1$ .

Recordemos que  $\omega$  se asigna a  $\Omega_1$  cuando  $L(\mathbf{x}) > 0$ ; en caso contrario  $\omega$  se asigna a  $\Omega_2$ .

Luego la probabilidad de clasificación errónea es:

$$pce = \frac{1}{2} P(L(\mathbf{x}) > 0 | \omega \in \Omega_2) + \frac{1}{2} P(L(\mathbf{x}) \leq 0 | \omega \in \Omega_1) = \Phi\left(-\frac{M}{2}\right),$$

donde  $\Phi$  es la función de distribución de la ley  $N(0, 1)$ .

**Ejercicio 4:** Demostrar que  $pce = \Phi\left(-\frac{M}{2}\right)$ .

**Ejercicio 4:** Demostrar que  $pce = \Phi\left(-\frac{M}{2}\right)$ .

**Solución:** Empezamos escribiendo la expresión de la probabilidad de clasificación errónea, teniendo en cuenta que cuando  $\omega \in \Omega_1$  entonces  $L(\mathbf{x}) \sim N\left(\frac{1}{2}M^2, M^2\right)$  y cuando  $\omega \in \Omega_2$  entonces  $L(\mathbf{x}) \sim N\left(-\frac{1}{2}M^2, M^2\right)$ :

$$\begin{aligned} pce &= \frac{1}{2} P(L(\mathbf{x}) > 0 | \omega \in \Omega_2) + \frac{1}{2} P(L(\mathbf{x}) \leq 0 | \omega \in \Omega_1) \\ &= \frac{1}{2} P\left(\frac{L(\mathbf{x}) + M^2/2}{M} > \frac{0 + M^2/2}{M}\right) + \frac{1}{2} P\left(\frac{L(\mathbf{x}) - M^2/2}{M} \leq \frac{0 - M^2/2}{M}\right) \\ &= \frac{1}{2} P(Z > M/2) + \frac{1}{2} P(Z \leq -M/2) = P(Z \leq -M/2) = \Phi(-M/2), \end{aligned}$$

donde  $Z \sim N(0, 1)$  y hemos usado que  $P(Z > k) = P(Z \leq -k)$ .

### uc3m

- c) Si conocemos las probabilidades *a priori*  $q_1 = P(\omega \in \Omega_1)$ ,  $q_2 = P(\omega \in \Omega_2)$ , con  $q_1 + q_2 = 1$ , entonces el discriminador de Bayes es:  $B(\mathbf{x}) = L(\mathbf{x}) + \log(q_1/q_2)$ .

Recordemos que  $B(\mathbf{x})$  se había definido como  $B(\mathbf{x}) = V(\mathbf{x}) + \log(q_1/q_2)$ . En el caso que  $\Sigma_1 = \Sigma_2 = \Sigma$ , ya hemos visto en a) que  $V(\mathbf{x}) = L(\mathbf{x})$ .

Observad que el término  $\log(q_1/q_2)$  puede cambiar la decisión incluso cuando  $L(\mathbf{x}) > 0$ . Por ejemplo:

$$\begin{aligned} B(\mathbf{x}) &= \underbrace{L(\mathbf{x})}_{>0} + \log(q_1/q_2) < 0 \Leftrightarrow \underbrace{L(\mathbf{x})}_{>0} + \log q_1 - \log q_2 < 0 \\ &\Leftrightarrow 0 < L(\mathbf{x}) - \log q_2 - \log q_1 = \log(q_2/q_1) \\ &\Leftrightarrow q_2 > q_1 \text{ y } \log(q_2/q_1) > L(\mathbf{x}), \end{aligned}$$

lo que significa que hay mucha evidencia *a priori* de que  $\omega \in \Omega_2$ .

## 3.2. Distintas covarianzas poblacionales

Si  $\mu_1 \neq \mu_2$  y  $\Sigma_1 \neq \Sigma_2$ , entonces:

- a) La regla de máxima verosimilitud proporciona el **discriminador cuadrático**:

$$\begin{aligned} V(\mathbf{x}) &= \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}) = \dots \\ &= \frac{1}{2} \mathbf{x}' (\Sigma_2^{-1} - \Sigma_1^{-1}) \mathbf{x} + \mathbf{x}' (\Sigma_1^{-1} \boldsymbol{\mu}_1 - \Sigma_2^{-1} \boldsymbol{\mu}_2) \\ &\quad + \frac{1}{2} \boldsymbol{\mu}_2' \Sigma_2^{-1} \boldsymbol{\mu}_2 - \frac{1}{2} \boldsymbol{\mu}_1' \Sigma_1^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \log |\Sigma_2| - \frac{1}{2} \log |\Sigma_1| \\ &= Q(\mathbf{x}), \end{aligned}$$

- b) Si conocemos las probabilidades *a priori*  $q_1 = P(\omega \in \Omega_1)$ ,  $q_2 = P(\omega \in \Omega_2)$ , con  $q_1 + q_2 = 1$ , entonces el discriminador de Bayes es:  $B(\mathbf{x}) = Q(\mathbf{x}) + \log(q_1/q_2)$ .

## 4. Clasificación en $g \geq 2$ poblaciones

- **Criterio geométrico.** Se asigna el individuo  $\omega$  a la población más próxima según la distancia de Mahalanobis,

$$\min_{1 \leq i \leq g} \delta_M^2(\mathbf{x}, \boldsymbol{\mu}_i).$$

- **Máxima verosimilitud.** Se asigna el individuo  $\omega$  a la población con mayor verosimilitud,

$$\max_{1 \leq i \leq g} \mathcal{L}_i(\mathbf{x}) = \max_{1 \leq i \leq g} f(x_1, \dots, x_p; \boldsymbol{\theta}_i).$$

- **Regla de Bayes.** Se asigna el individuo  $\omega$  a la población con mayor probabilidad *a posteriori*,

$$\max_{1 \leq i \leq g} P(\omega \in \Omega_i | \mathbf{x}) = \max_{1 \leq i \leq g} \frac{q_i \mathcal{L}_i(\mathbf{x})}{\sum_{i=1}^g q_i \mathcal{L}_i(\mathbf{x})},$$

dónde  $q_i = P(\omega \in \Omega_i)$  son las *a priori* con  $\sum_{i=1}^g q_i = 1$ .

## 5. Clasificadores no-paramétricos: $k$ -NN

Uno de los clasificadores más utilizados, por su sencillez y buenos resultados en poblaciones no normales, es la regla  $k$ -NN ( $k$  nearest neighbours) o de los  $k$  vecinos más cercanos.

Supongamos que tenemos  $g$  poblaciones conocidas  $\Omega_1, \dots, \Omega_g$  y en cada una de ellas observamos una muestra de cierto vector de variables cuantitativas de interés  $\mathbf{X} = (X_1, \dots, X_p)'$ .

Dada una nueva observación  $\mathbf{x} = (x_1, \dots, x_p)'$  a clasificar, se toman las  $k$  observaciones  $\mathbf{x}_i$  de la muestra más cercanas a  $\mathbf{x}$  y se clasifica  $\mathbf{x}$  según el “voto de la mayoría”.

En concreto, los pasos de este algoritmo son:

1. Definir una medida de distancia entre las observaciones, habitualmente la distancia de Mahalanobis,
2. Calcular las distancias de la observación a clasificar,  $\mathbf{x}$ , a todas las observaciones de la muestra,
3. Seleccionar las  $k$  observaciones más cercanas a  $\mathbf{x}$  (los  $k$  vecinos). Calcular la proporción de estos  $k$  vecinos que pertenecen a cada una de las poblaciones  $\Omega_1, \dots, \Omega_g$ . Clasificar  $\mathbf{x}$  en la población con mayor frecuencia de vecinos.

Un elemento clave de este método es la selección de  $k$ . Una práctica habitual es tomar  $k = \sqrt{\frac{1}{g} \sum_{i=1}^g n_i}$ , donde  $n_i$  es el tamaño de la submuestra de la población  $\Omega_i$ ,  $i = 1, \dots, g$ . Otra posibilidad es probar para distintos valores de  $k$ : Se aplica el método a los puntos de la muestra cuya población es conocida obteniendo el error de clasificación (por ejemplo, por validación cruzada) en función de  $k$  y se escoger aquel valor de  $k$  con menor error observado.

## uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)

**Ejemplo 2.** El fichero iris.txt contiene los datos de Fisher sobre tres especies de flores del género Iris (*I. setosa*, *I. versicolor*, *I. virginica*). Las variables que se miden son  $X_1 =$  longitud del sépalo,  $X_2 =$  anchura del sépalo,  $X_3 =$  longitud del pétalo,  $X_4 =$  anchura del pétalo. El fichero contiene 150 flores, las 50 primeras pertenecen a la primera especie, las 50 siguientes corresponden a la segunda especie y las 50 últimas son de la tercera especie. Usar la regla  $k$ -NN para clasificar tres nuevas flores:

ind.	$X_1$	$X_2$	$X_3$	$X_4$
1	4.6	3.5	1.0	0.2
2	6.8	2.7	4.8	1.4
3	7.1	3.2	6.0	1.8

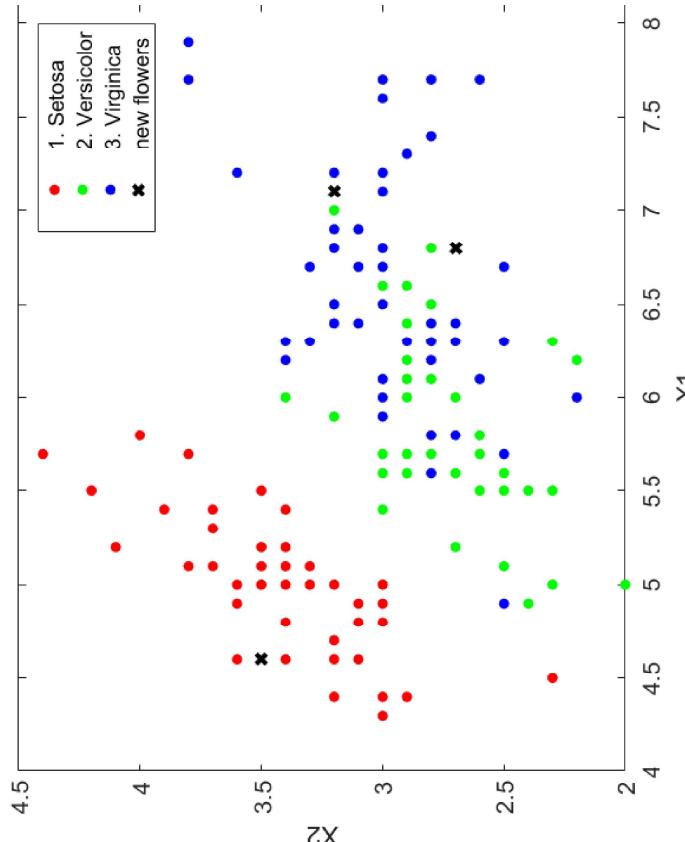
Empezamos cargando los datos y preparando los grupos a los que sabemos que pertenece cada flor:

```
data=load('iris.txt');
group=ones(150,1); 2*ones(50,1); 3*ones(50,1);
new=[4.6 3.5 1 0.2; 6.8 2.7 4.8 1.4; 7.1 3.2 6 1.8];
% representacion de los datos
gscatter(data(:,1),data(:,2),group)
hold on;
plot(new(:,1),new(:,2),'xk','MarkerSize',6,'LineWidth',2);
legend('1. Setosa','2. Versicolor','3. Virginica','new flowers','Location','NE')
xlabel('X1')
ylabel('X2')
```

## uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)



## uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)

Aplicamos el clasificador  $k$ -NN, usando la distancia de Mahalanobis:

```
[Idx,D] = knnsearch(data,new,'K',5,'Distance','mahalanobis');
```

```
for i=1:size(Idx,1)
    class(i)=mode(group(Idx(i,:)));
end
```

```
Idx =
      23    27     7    41    22
      77    59    75    76    55
     126   108   130   106   117
```

```
D =
0.3325 0.6407 0.6607 0.8143 0.8273
0.3325 0.9287 0.9922 1.0454 1.0467
0.3212 1.0354 1.1179 1.3270 1.3999
```

```
class=
```

```
    1    2    3
```

y dibujamos el resultado

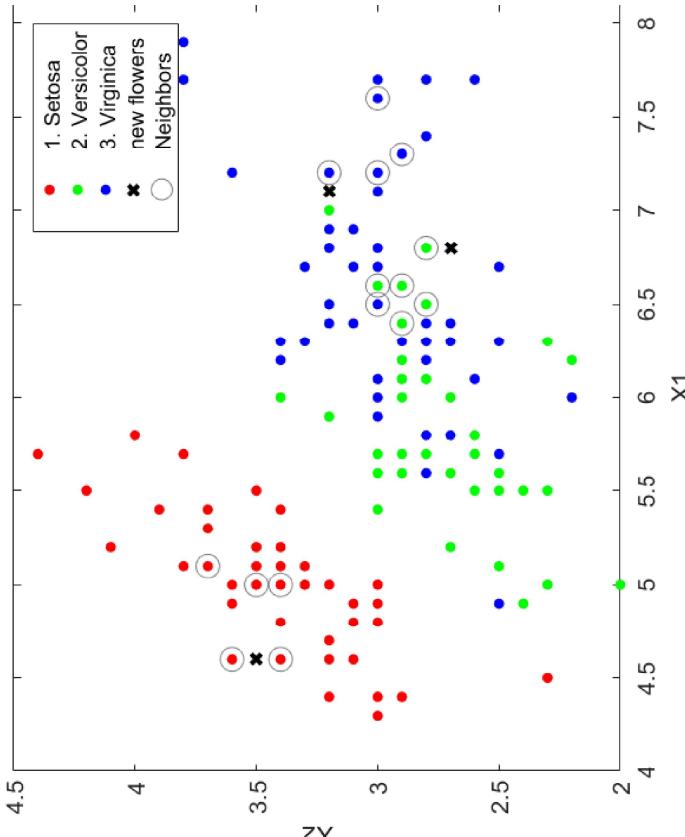
```
plot(data(Idx,1),data(Idx,2),'Color',[.5 .5 .5],'Marker','o',...
...,'Markersize',10,'LineStyle','none')
legend('1. Setosa','2. Versicolor','3. Virginica',...
...'new flowers','Neighbors')
```

```
...,'Neigbors','Location','NE')
```

## uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)



Podemos estimar la probabilidad de clasificación errónea por validación cruzada, mediante una pequeña función Matlab:

```
function error=pce(data,group,k)
[n,p]=size(data);
for i=1:n
    new=data(i,:);
    X=data([1:i-1 i+1:n],:);
    Idx = knnsearch(X,new,'K',k,'Distance','mahalanobis');
    class(i,1)=mode(group(Idx));
end
error=1-sum(class==group)/n;
end
```

Para los datos de iris.txt y  $k = 5$ , obtenemos:

```
error=pce(data,group,5)
error=0.1000
```

Para escoger el mejor  $k$  probaríamos para distintos valores de  $k$  y nos quedaríamos con aquel que llevara a una menor probabilidad de clasificación errónea.

## uc3m

### Algunas funciones Matlab para Discriminante (Statistics and Machine Learning Toolbox)

#### Clasificar observaciones

<code>predict</code>	Predict labels using discriminant analysis classification model
<code>resubPredict</code>	Predict resubstitution labels of discriminant analysis classification model
<code>classify</code>	Discriminant analysis

#### Crear un modelo de análisis discriminatorio

<code>fitcdiscr</code>	Fit discriminant analysis classifier
<code>makecdiscr</code>	Construct discriminant analysis classifier from parameters
<code>compact</code>	Compact discriminant analysis classifier

#### Validación cruzada

<code>crossval</code>	Cross-validated discriminant analysis classifier
<code>kfoldEdge</code>	Classification edge for observations not used for training
<code>kfoldLoss</code>	Classification loss for observations not used for training
<code>kfoldfun</code>	Cross validate function
<code>kfoldMargin</code>	Classification margins for observations not used for training
<code>kfoldPredict</code>	Predict response for observations not used for training