

# **ANÁLISIS DISCRIMINANTE**

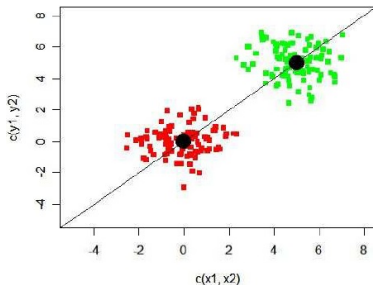
ALAN REYES-FIGUEROA  
APRENDIZAJE ESTADÍSTICO

(AULA 22) 29.ABRIL.2024

# Análisis discriminante

Queremos separar o clasificar un conjunto de datos en 2 clases.

Punto de partida: ¿en cuál dirección proyectar los datos para separar los puntos de diferentes clases lo mayor posible?



Después usar un clasificador lineal del tipo  $\hat{y}(\mathbf{x}) = \mathbf{1}(\mathbf{l}^T \mathbf{x} > l_0)$ .

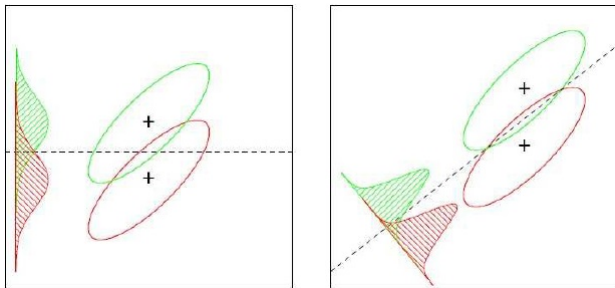
# Análisis discriminante

¿Cómo definir “separar las clases lo mayor posible”?

Primera idea:

Considerar los centroides de cada clase como representantes de cada clase. Separamos estos lo más posible.

Problema: hay que tomar en cuenta la estructura de la covarianza.



# Análisis discriminante

Supongamos que tenemos un conjunto de datos  $(\mathbf{x}_i, y_i)_{i=1}^n$  divididos en dos clases: 0 y 1. La variabilidad o dispersión de los datos puede escribirse como

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n \sum_{i=1}^n \bar{x}^2 = \mathbf{x}^T \mathbf{x} - n \mathbf{x}^T \mathbf{1} \mathbf{1}^T \bar{x},$$

la cual es una expresión de la forma  $\mathbf{x}^T (I - \mathbf{1} \mathbf{1}^T) \mathbf{x} = \mathbf{x}^T (I - J) \mathbf{x} = \mathbf{x}^T P \mathbf{x}$ , donde  $P = I - J = I - \mathbf{1} \mathbf{1}^T$  funciona como una matriz de covarianza.

Queremos maximizar la separación entre ambas clases. Si

$$\begin{aligned} \mathbf{x}^T A \mathbf{x} &= \text{dispersión entre clases,} \\ \mathbf{x}^T B \mathbf{x} &= \text{discrepancia dentro de cada clase,} \end{aligned}$$

estamos interesados en calcular los valores extremos del cociente  $\frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T B \mathbf{x}}$ .

## Teorema (Cociente de Rayleigh)

Si  $A \in \mathbb{R}^{d \times d}$  es simétrica,  $B \in \mathbb{R}^{d \times d}$  es simétrica y positiva definida, entonces los valores extremos  $\sup_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T B \mathbf{x}} = \lambda_1$ ,  $\inf_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T B \mathbf{x}} = \lambda_n$ , corresponden a los autovalores máximo y mínimo, respectivamente, de  $B^{-1}A$ , y los valores extremos se alcanzan en  $\mathbf{q}_1$  y  $\mathbf{q}_n$ , los autovectores correspondientes.  $\square$

## Corolario

Si  $\mathbf{a} \in \mathbb{R}^d$ ,  $B \in \mathbb{R}^{d \times d}$  es simétrica y positiva definida, entonces

$$\sup_{\mathbf{x} \neq 0} \frac{(\mathbf{a}^T \mathbf{x})}{\mathbf{x}^T B \mathbf{x}} = \sup_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T (\mathbf{a} \mathbf{a}^T) \mathbf{x}}{\mathbf{x}^T B \mathbf{x}} = \mathbf{a}^T B^{-1} \mathbf{a},$$

se alcanza en  $\mathbf{x} = k B^{-1} \mathbf{a}$ .  $\square$

# Análisis discriminante

Eso fue lo que Fisher (1936) estudió en



## THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

By R. A. FISHER, Sc.D., F.R.S.

Table I shows measurements of the flowers of fifty plants each of the two species *Iris setosa* and *I. versicolor*, found growing together in the same colony and measured by Dr E. Anderson, to whom I am indebted for the use of the data. Four flower measurements are given. We shall first consider the question: What linear function of the four measurements

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$$

will maximize the ratio of the difference between the specific means to the standard deviations within species? The observed means and their differences are shown in Table II.

# Análisis discriminante

Trabajó con el conjunto de datos IRIS (colectado por Edgar Anderson): 150 datos, 4 atributos, 3 clases  $\Rightarrow X \in \mathbb{R}^{150 \times 4}$ ,  $\mathbf{y} \in \mathbb{R}^{150}$ .

Dada una observación  $\mathbf{x} \in \mathbb{R}^4$ , Fisher se preguntó si existe alguna función lineal

$$\lambda^T \bar{\mathbf{x}} = \lambda_1 \bar{x}_1 + \lambda_2 \bar{x}_2 + \lambda_3 \bar{x}_3 + \lambda_4 \bar{x}_4$$

de las medias, tal que maximice el cociente de las diferencias entre las medias y la dispersión (varianza) dentro de cada grupo.

Consideremos el caso de dos poblaciones, con medias  $\mu_1, \mu_2 \in \mathbb{R}^4$ . Denotamos la diferencia de ambas por

$$\mathbf{d} = \mu_1 - \mu_2 = (d_1, d_2, d_3, d_4)^T.$$

# Análisis discriminante

Consideramos la expresión lineal

$$D = \lambda^T \mathbf{d} = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3 + \lambda_4 d_4 = \lambda^T (\mu_1 - \mu_2),$$

y

$$S = \text{Var}(\lambda^T \bar{\mathbf{x}}_1 - \lambda^T \bar{\mathbf{x}}_2) = \lambda^T \text{Var}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \lambda = \lambda^T \Sigma \lambda.$$

Queremos maximizar la cantidad

$$\frac{D^2}{S} = \frac{(\lambda^T \mathbf{d})^2}{\lambda^T \Sigma \lambda} = \frac{(\mathbf{d}^T \lambda)^2}{\lambda^T \Sigma \lambda}. \quad (1)$$

Por el corolario anterior, el óptimo está dado por el múltiplo

$$\lambda^* = k \Sigma^{-1} \mathbf{d} = k \Sigma^{-1} (\mu_1 - \mu_2).$$

# Análisis discriminante

Fisher obtuvo el mismo resultado sin usar propiedades espectrales, sino derivando la expresión (1):

$$\nabla_{\lambda} \frac{D^2}{S} = \nabla_{\lambda} \frac{(\mathbf{d}^T \lambda)^2}{\lambda^T \Sigma \lambda} = \frac{2(\mathbf{d}^T \lambda)(\lambda^T \Sigma \lambda) \mathbf{d} - (\mathbf{d}^T \lambda)^2 \cdot 2 \Sigma \lambda}{(\lambda^T \Sigma \lambda)^2} = 0,$$

$$\Rightarrow (\mathbf{d}^T \lambda)[(\lambda^T \Sigma \lambda) \mathbf{d} - (\lambda^T \mathbf{d}) \Sigma \lambda] = 0 \Rightarrow (\lambda^T \Sigma \lambda) \mathbf{d} = (\lambda^T \mathbf{d}) \Sigma \lambda$$

De ahí que

$$\Sigma \lambda = \frac{\lambda^T \Sigma \lambda}{\lambda^T \mathbf{d}} \mathbf{d} = t \mathbf{d},$$

y se tiene que  $\lambda^* = t \Sigma^{-1} \mathbf{d} = t \Sigma^{-1} (\mu_1 - \mu_2)$ .

# Análisis discriminante

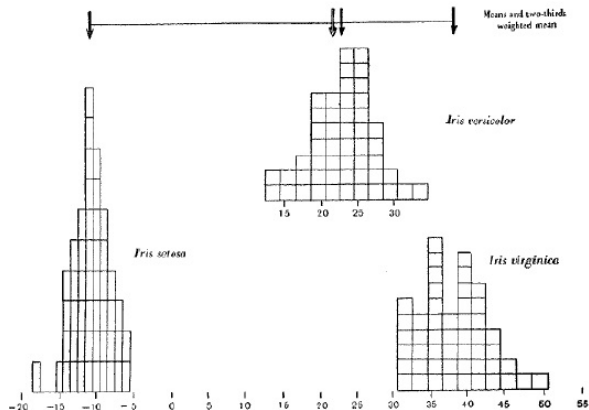


Fig. 1. Frequency histograms of the discriminating linear function, for three species of *Iris*.

Resultados obtenidos por Fisher para el conjunto de datos IRIS.

# Criterios de separabilidad

La **dispersión dentro de clases** (*variance within groups*)  $S_W$  está determinada por la suma de las varianzas de cada grupo:

$$S_W = S_1 + S_2 + \dots, S_K,$$

donde  $S_i = \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)$ ,  $\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$ . Por otro lado, la **dispersión entre clases** (*variance between groups*)  $S_B$  es

$$S_B = \sum_{i=1}^k n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}),$$

con  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} \mathbf{x}_{ij}$ ,  $n = n_1 + n_2 + \dots + n_k$ .

La **dispersión total** es  $S_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}})^T (\mathbf{x}_{ij} - \bar{\mathbf{x}})$ .

# Criterios de separabilidad

## Propiedad

$$S_T = S_W + S_B. \square$$

### Obs!

- La varianza total (de datos  $\mathbf{x}_{ij}$  agrupados en  $k$  clases) puede descomponerse en dos factores:
  - variabilidad dentro de cada grupo  $S_W$ ,
  - variabilidad entre grupos  $S_B$ .
- Relación con ANOVA.
- Podemos combinar las técnicas para determinar la mayor separación entre grupos, con técnicas de reducción de dimensión: Aplicamos una transformación lineal  $W : \mathbb{R}^d \rightarrow \mathbb{R}^p$ ,  $p < r$ , a los datos  $\tilde{\mathbf{X}} = W\mathbf{X}$ .  
Definimos en ese caso  $\tilde{S}_W = W^T S_W W$ ,  $\tilde{S}_B = W^T S_B W$ , y  $\tilde{S}_T = W^T S_T W$ .

# Criterios de separabilidad

Tenemos los siguientes criterios de separabilidad:

- El criterio estándar: la variabilidad entre grupos, sobre la variabilidad dentro de cada grupo

$$J(W) = \frac{\tilde{S}_B}{\tilde{S}_W},$$

- $J(W) = \text{tr}(\tilde{S}_W^{-1} \tilde{S}_B),$
- $J(W) = \text{tr}(\tilde{S}_B) - \lambda \text{tr}(\tilde{S}_W),$
- $J(W) = \frac{\text{tr} \tilde{S}_B}{\text{tr} \tilde{S}_W},$
- $J(W) = \log \det(\tilde{S}_W^{-1} \tilde{S}_B) = \sum_{i=1}^r \log \lambda_i$

# Enfoque probabilístico

Considera el clasificador bayesiano óptimo

$$\hat{y}(\mathbf{x}) = \mathbf{1}\left(\frac{\mathbb{P}(X = \mathbf{x} \mid Y = 1)}{\mathbb{P}(X = \mathbf{x} \mid Y = 0)} > c_1\right), \quad \text{con } c_1 = \frac{\lambda_{01} \mathbb{P}(Y = 0)}{\lambda_{10} \mathbb{P}(Y = 1)}.$$

Vamos a suponer, por ahora, que las dos poblaciones siguen distribuciones normales multivariadas:

$$f_j(\mathbf{x}) = \mathbb{P}(X = \mathbf{x} \mid Y = j) = \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} e^{-\frac{1}{2}(\mathbf{x} - \mu_j)^T \Sigma_j^{-1} (\mathbf{x} - \mu_j)}.$$

Y así

$$\log \frac{\mathbb{P}(X = \mathbf{x} \mid Y = 1)}{\mathbb{P}(X = \mathbf{x} \mid Y = 0)} = -\frac{(\mathbf{x} - \mu_1)^T \Sigma_1^{-1} (\mathbf{x} - \mu_1) - (\mathbf{x} - \mu_0)^T \Sigma_0^{-1} (\mathbf{x} - \mu_0)}{2} + \frac{1}{2} \log \frac{|\Sigma_0|}{|\Sigma_1|}.$$

# Enfoque probabilístico

Entonces

$$\begin{aligned}\hat{y}(\mathbf{x}) = 1 &\Leftrightarrow -\frac{(\mathbf{x} - \mu_1)^T \Sigma_1^{-1} (\mathbf{x} - \mu_1) - (\mathbf{x} - \mu_0)^T \Sigma_0^{-1} (\mathbf{x} - \mu_0)}{2} > \log \frac{\lambda_{01} \pi_0 |\Sigma_1|^{1/2}}{\lambda_{10} \pi_1 |\Sigma_0|^{1/2}} \\ &\Leftrightarrow \mathbf{x}^T (\Sigma_1^{-1} - \Sigma_0^{-1}) \mathbf{x} - 2(\mu_1 \Sigma_1^{-1} - \mu_0 \Sigma_0^{-1}) \mathbf{x} + \mu_1 \Sigma_1^{-1} \mu_1 - \mu_0 \Sigma_0^{-1} \mu_0 > -2 \log \frac{\lambda_{01} \pi_0 |\Sigma_1|^{1/2}}{\lambda_{10} \pi_1 |\Sigma_0|^{1/2}}\end{aligned}$$

Esta es una ecuación de la forma  $\mathbf{b}_0 + \mathbf{b}_1^T \mathbf{x} + \mathbf{x}^T \mathbf{b}_2 \mathbf{x} > 0$ .

## Definición

El término anterior se llama el **discriminante cuadrático** (QDA).

# Enfoque probabilístico

En el caso en que las covarianzas  $\Sigma_0$  y  $\Sigma_1$  coinciden, entonces la ecuación anterior se reduce a

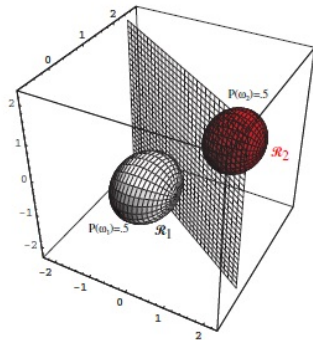
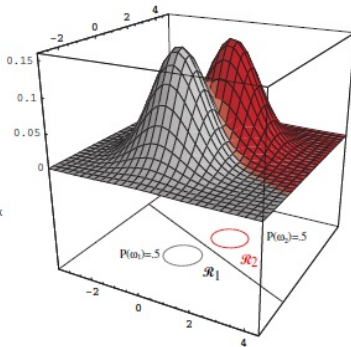
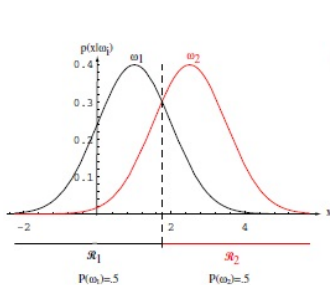
$$\begin{aligned}\hat{y}(\mathbf{x}) = 1 &\Leftrightarrow \mathbf{x}^T(\Sigma_1^{-1} - \Sigma_0^{-1})\mathbf{x} - 2(\mu_1\Sigma_1^{-1} - \mu_0\Sigma_0^{-1})\mathbf{x} + \mu_1\Sigma_1^{-1}\mu_1 - \mu_0\Sigma_0^{-1}\mu_0 > -2 \log \frac{\lambda_{01}\pi_0|\Sigma_1|^{1/2}}{\lambda_{10}\pi_1|\Sigma_0|^{1/2}} \\ &\Leftrightarrow -2(\mu_1\Sigma_1^{-1} - \mu_0\Sigma_0^{-1})\mathbf{x} + \mu_1\Sigma_1^{-1}\mu_1 - \mu_0\Sigma_0^{-1}\mu_0 > -2 \log \frac{\lambda_{01}\pi_0|\Sigma_1|^{1/2}}{\lambda_{10}\pi_1|\Sigma_0|^{1/2}}.\end{aligned}$$

Esta es una ecuación de la forma  $\mathbf{b}_0 + \mathbf{b}_1^T\mathbf{x} > 0$ .

## Definición

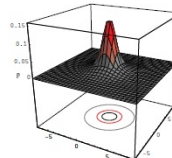
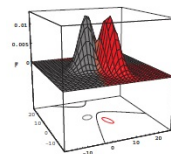
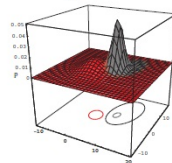
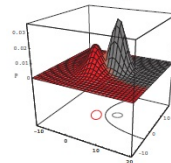
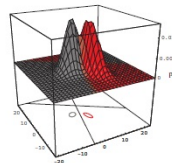
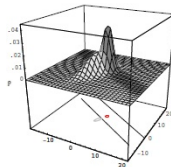
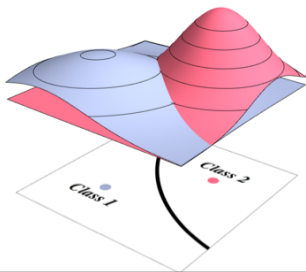
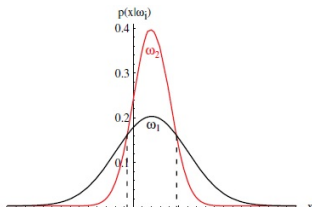
El término anterior se llama el **discriminante lineal** (LDA).

# Enfoque probabilístico



Análisis discriminante lineal, para el caso de dos normales  $f_i(\mathbf{x})$ .

# Enfoque probabilístico



# Enfoque probabilístico

