Tema 5.- Análisis discriminante (AD)

Asignatura: ESTADÍSTICA MULTIVARIANTE (Prácticas)

Grados en: Física y Matemáticas; Ingeniería Informática y Matemáticas; Matemáticas

(4º Curso - 1er semestre 2023-2024)

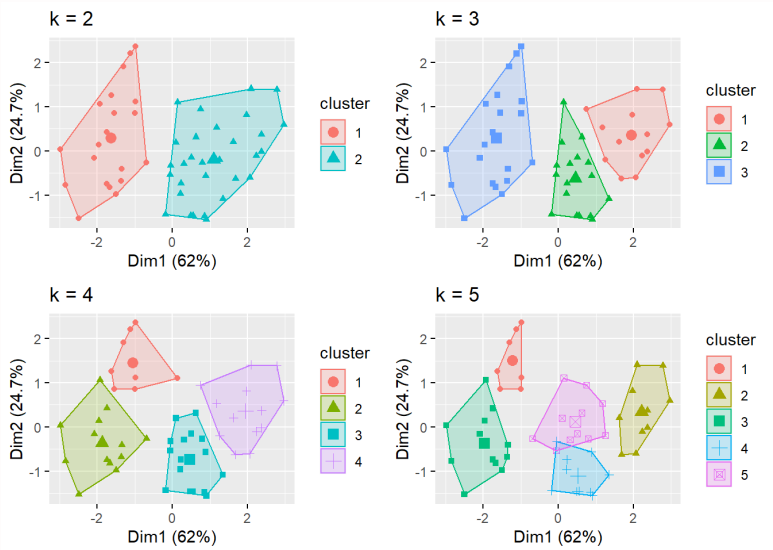
©Prof. Dr. José Luis Romero Béjar

Departamento de Estadística e Investigación Operativa
Facultad de Ciencias (Despacho 10)

Periodo de docencia: 11/09/2023 a 22/12/2023

Análisis exploratorio de datos (recordatorio de temas anteriores)

- Reducción de la dimensión:
  - ACP - Análisis de componentes principales (variables observables).
  - AF - Análisis factorial (variables latentes).
- Análisis cluster (aprendizaje no supervisado):
  - Busca agrupamientos de forma objetiva.
  - Define variables respuesta para clasificación.
  - Con frecuencia es el **punto de partida para aprendizaje supervisado**.



Descripción general del aprendizaje supervisado

- Objetivo: **clasificar** nuevos registros, según sus características (predictores), en los diferentes niveles de una variable de respuesta cualitativa.
- Elementos:
  - Variable respuesta definida en niveles (cualitativa ordinal o nominal).
  - Variables explicativas o predictivas (preferiblemente que conformen un vector aleatorio continuo).
- Procedimiento:
  1. **Estima la probabilidad** de que una observación, dado el valor de los predictores, **pertenezca a cada uno de los niveles** de la variable respuesta.
  2. **Asigna** la observación a la modalidad con la **mayor probabilidad**.
- Modelos y algoritmos:
  - Support Vector Machine (SVM).
  - Árboles de decisión.
  - Regresión logística.
  - **Análisis discriminante**.

Notación

- Y es una **variable de respuesta categórica** con $k \geq 2$ niveles.
- $X = (X_1, \ldots, X_n), n \in \mathbb{N}$ es un **vector aleatorio continuo** de variables explicativas.
- π_k es la **probabilidad a priori**, $\mathrm{P}(Y = k)$, de cada nivel de la variable respuesta.
- $f_k(x)$ es la **función de densidad** de probabilidad condicionada, $P(X = x | Y = k)$

Supuestos

$X = (X_1, \ldots, X_n), n \in \mathbb{N}$ es un vector aleatorio continuo con distribución multivariante **Gaussiana**, con

- varianza **homogénea** -> Análisis Discriminante Lineal (LDA).
- varianza **heterogénea** -> Análisis Discriminante Cuadrático (QDA).

Definición del modelo

Existen diferentes enfoques para la definición del modelo discriminante (Fisher, Bayes, etc.). Para la formulación siguiente, **se considera el enfoque de Bayes**.

LDA con un único predictor o variable explicativa

Dada Y una **variable aleatoria de respuesta categórica** con $k \geq 2$ niveles y X una **variable aleatoria continua**, se pretende **clasificar** en los diferentes niveles de Y para valores específicos de X.

- Se necesita **estimar**, $\frac{P(Y=i|X=x)}{P(Y=j|X=x)} = \frac{P(Y=i,X=x)}{P(Y=j,X=x)}; i, j \in 1, \ldots, k.$
- De acuerdo con el **Teorema de Bayes** y la notación anterior (diapositiva 8) se tiene que, $\frac{P(Y=i|X=x)}{P(Y=j|X=x)} = \frac{\pi_i P(X=x|Y=i)}{\pi_j P(X=x|Y=j)} = \frac{\pi_i f_i(x)}{\pi_j f_j(x)}$.
- **Regla de decisión**: si $\frac{\pi_i f_i(x)}{\pi_j f_j(x)} > 1$, o $\frac{f_i(x)}{f_j(x)} > \frac{\pi_j}{\pi_i}$ entonces, el registro se asigna a la clase o nivel de respuesta i.
- Teniendo en cuenta que $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$ es una **densidad Gaussiana** con media μ_k y **varianza homogénea**, σ^2, en los k niveles y aplicando logaritmo para linealizar se tiene que, **el registro es asignado al nivel** i, si y solo si,

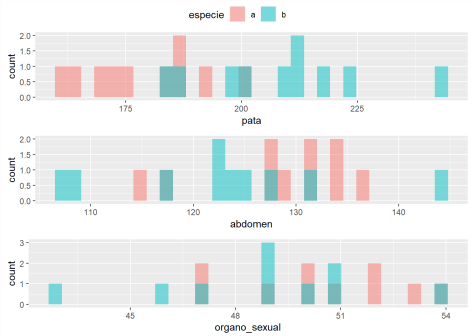
$$log\left(\frac{f_i(x)}{f_j(x)}\right) > log\left(\frac{\pi_j}{\pi_i}\right) \Leftrightarrow \frac{\mu_i - \mu_j}{\sigma^2}x - \frac{\mu_i^2 - \mu_j^2}{2\sigma^2} - log(\frac{\pi_j}{\pi_i}) > 0 \qquad (1)$$

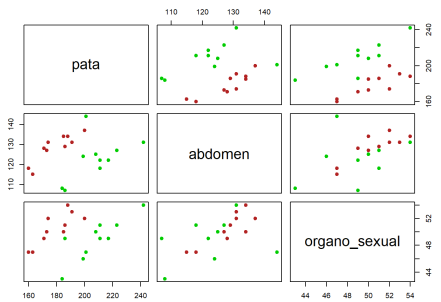
LDA con un único predictor o variable explicativa (comentarios)

- La ecuación (1) se dice que es un **clasificador discriminante lineal**.
- **Regla de decisión como relación de probabilidades**. Si la variable respuesta Y tiene $k = 2$ niveles, entonces:
  - Si $\frac{P(Y=1|X=x)}{P(Y=2|X=x)} > 1$, el registro es **asignado al primer nivel** de Y.
  - Si $\frac{P(Y=1|X=x)}{P(Y=2|X=x)} < 1$, el registro es **asignado al segundo nivel** de Y.
- **Varianza heterogénea**. La ecuación incluirá un **término cuadrático** derivado de la estructura de covarianza (**clasificador discriminante cuadrático**).
- Para **más de un regresor** simplemente se considera la expresión general del teorema de Bayes.

El procedimiento AD se puede resumir en seis pasos

0. **Recomendación previa**. Análisis exploratorio gráfico.





1. Elegir un **conjunto de entrenamiento**. Es un conjunto de registros con nivel conocido para la variable de respuesta.
2. Estimar las **probabilidades a priori**, π_k, o la proporción esperada de registros para cada nivel de Y.
3. Discutir entre varianza **homogénea** (LDA) o **heterogénea** (QDA).
4. **Estimar parámetros**.
5. Construir el **clasificador discriminante**.
6. **Validación cruzada**. Elija un **conjunto de prueba** (test) para estimar la tasa de clasificación correcta.

Aspectos tratados

- Aspectos generales relacionados con el Aprendizaje Supervisado.
- Fundamento matemático del Análisis Discriminante Lineal.
- Enfoque metodológico del Análisis Discriminante en la práctica.

Tareas voluntarias

- Deducir la ecuación del clasificador discriminante lineal con $n > 1$ predictores.
- Deducir la ecuación de un clasificador discriminante si la varianza es heterogénea (clasificador discriminante cuadrático).

Práctica 4 de AD

En esta práctica se ilustra un ejemplo de clasificación con un **modelo discriminante lineal** y otro ejemplo con un **clasificador cuadrático**.
Para realizar esta práctica se debe **descargar y ejecutar** el siguiente archivo, **AD_4_esp.Rmd** disponible en la plataforma PRADO del curso.

Aspectos tratados:

- Paquetes de R necesarios.
- Exploración gráfica de los datos.
- Supuestos: normalidad y homogeneidad de la varianza.
- Funciones discriminante.
- Visualización de las clasificaciones.

[1] Anderson, T.W. (2003, 3ª ed.). An Introduction to Multivariate Statistical Analysis. John Wiley & Sons.

[2] Gutiérrez, R. y González, A. (1991). Estadística Multivariable. Introducción al Análisis Multivariante. Servicio de Reprografía de la Facultad de Ciencias. Universidad de Granada.

[3] Härdle, W.K. y Simar, L. (2015, 4ª ed.). Applied Multivariate Statistical Analysis. Springer.

[4] Johnson, R.A. y Wichern, D.W. (1988). Applied Multivariate Analysis. Prentice Hall International, Inc.

[5] Rencher, A.C. y Christensen, W.F. (2012, 3ª ed.). Methods of Multivariate Analysis. John Wiley & Sons.

[6] Salvador Figueras, M. y Gargallo, P. (2003). Análisis Exploratorio de Datos. Online en http://www.5campus.com/leccion/aed.

[7] Timm, N.H. (2002). Applied Multivariate Analysis. Springer.

[8] Vera, J.F. (2004). Análisis Exploratorio de Datos. ISBN: 84-688-8173-2.