# Unit 5.- Discriminant analysis (DA)

## Course: MULTIVARIATE STATISTICS

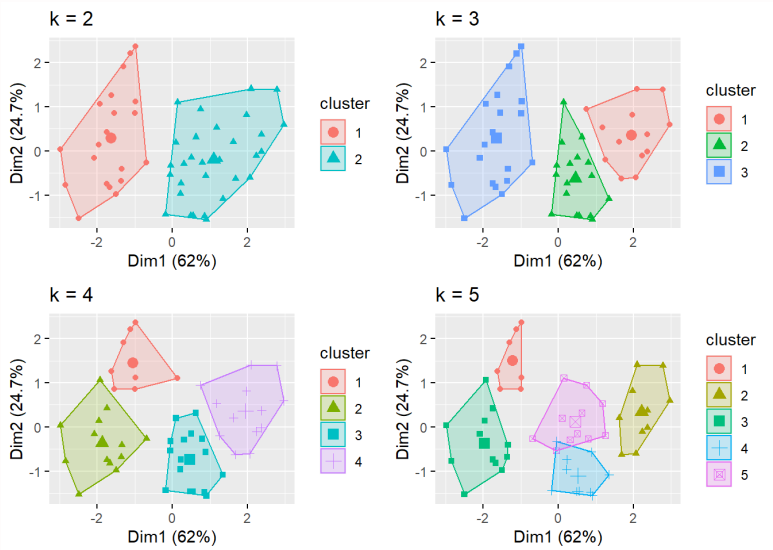
© Prof. Dr. José Luis Romero Béjar - Guillermo Arturo Cañadas de la Fuente




Junio, 2024

## Exploratory data analysis (recall from previous lessons)

- Dimensionality reduction:
  - PCA - Principal component analysis (observable variables).
  - FA - Factorial analysis (latent variables).
- Cluster analysis (unsupervised learning):
  - Looking for groupings.
  - Defining response variables for classification.
  - Often the **starting point for supervised learning**.



## Supervised learning overview

- Aim: **to classify** new records, according to their characteristics (predictors), into the different levels of a qualitative response variable.
- Elements:
  - Level-defined response variable (qualitative).
  - Explanatory or predictor variables (continuous random vector desirable).
- Procedure:
  1. **Estimates the probability** that one observation, given the value of the predictors, **belongs to each of the levels** of the response variable.
  2. **Assigns** the observation to the modality with the **highest probability**.
- Models and algorithms:
  - Support Vector Machine (SVM).
  - Decision Trees.
  - Logistic Regression.
  - **Discriminant Analysis**.

## Technical notation

- $Y$ is a **categorical response variable** with $k \geq 2$ levels.
- $X = (X_1, \ldots, X_n), n \in \mathbb{N}$ is a **continuous random vector** of explanatory variables.
- $\pi_k$ is the **prior probability**, $\mathrm{P}(Y = k)$.
- $f_k(x)$ is the **density function** of the conditional probability, $P(X = x | Y = k)$

## Assumptions

$X = (X_1, \ldots, X_n), n \in \mathbb{N}$ is a **multivariate Gaussian** continuous random vector with,

- **homogeneous** variance -> Linear Discriminant Analysis (LDA).
- **heterogeneous** variance -> Quadratic Discriminant Analysis (QDA).

## Model definition

There are different DA model definition approaches (Fisher, Bayes, etc.) For the formulation bellow the **Bayes approach is considered**.

## LDA with a single predictor

Given $Y$ a **categorical response random variable** with $k \geq 2$ levels and $X$ a **single continuous random variable**, it is intended to **classify** in the different levels of Y for specific values of X.

- Need to **estimate**, $\frac{P(Y=i|X=x)}{P(Y=j|X=x)} = \frac{P(Y=i,X=x)}{P(Y=j,X=x)}; i, j \in 1, \ldots, k$
- According to **Bayes' Theorem** and previous notation (slide 8),

$$\frac{P(Y=i|X=x)}{P(Y=j|X=x)} = \frac{\pi_i P(X=x|Y=i)}{\pi_j P(X=x|Y=j)} = \frac{\pi_i f_i(x)}{\pi_j f_j(x)}$$

- **Decision rule**: if $\frac{\pi_i f_i(x)}{\pi_j f_j(x)} > 1$, or $\frac{f_i(x)}{f_j(x)} > \frac{\pi_j}{\pi_i}$ then, the record is assigned to class $i$.
- Assuming $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$ is a **Gaussian density** with mean $\mu_k$ and **homogeneous variance**, $\sigma^2$, in the $k$ levels and applying logarithm to linearise then, **the record is assigned to class** $i$ if an only if,

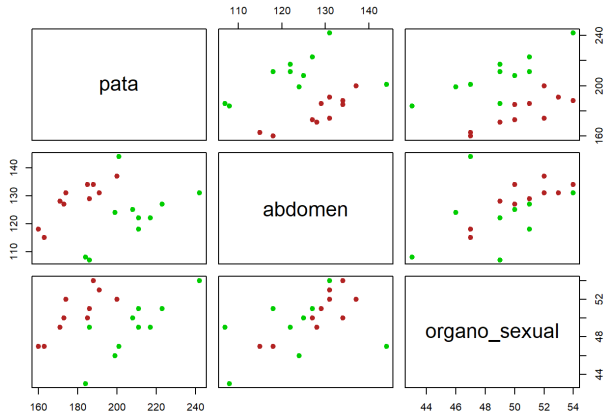
$$log\left(\frac{f_i(x)}{f_j(x)}\right) > log\left(\frac{\pi_j}{\pi_i}\right) \Leftrightarrow \frac{\mu_i - \mu_j}{\sigma^2}x - \frac{\mu_i^2 - \mu_j^2}{2\sigma^2} - log(\frac{\pi_j}{\pi_i}) > 0 \qquad (1)$$

## LDA with a single preditor (remarks)

- Equation (1) is called **Linear Discriminant classifier**.
- **Decision rule as a probabilities ratio**. If the response $Y$ has $k = 2$ levels, then:
  - If $\frac{P(Y=1|X=x)}{P(Y=2|X=x)} > 1$, the record is **assigned to the first level** of $Y$.
  - If $\frac{P(Y=1|X=x)}{P(Y=2|X=x)} < 1$, the record is **assigned to the second level** of $Y$.
- **Heterogeneous variance**. The equation will include a **quadratic term** derived from covariance structure (**Quadratic Discriminant classifier**.)
- **More than one regressor** simply by considering the general expression of Bayes' theorem.

DA procedure can be summarized in six steps.

0. **Prior recommendation**. Graphical exploratory analysis.


1. Choose a **training set**. It is a record set with known level for the response variable.
2. Estimate **prior probabilities**, $\pi_k$, or expected ratio of records for every level of $Y$.
3. Discuss between **homogeneous** (LDA) or **heterogeneous** variance (QDA).
4. **Parameter estimate**.
5. Build the **discriminant classifier**.
6. **Cross-validation**. Choose a **test set** to estimate the correct classification rate.

## Aspects covered in this class

- General aspects related to Supervised Learning.
- Mathematical foundation of Linear Discriminant Analysis.
- Methodological approach for Discriminant Analysis in practice.

## Elective homework

- Deduce the equation of Linear Discriminant Classifier with $n > 1$ predictors.
- Probe the equation of a Discriminant Classifier due to a heterogeneous variance (Quadratic Discriminant Classifier).

## DA Practice 4

In this practice an example of classification with a **linear discriminant model** and another example with a **quadratic classifier** is illustrated.
To carry out this practice you must **download and execute** the file **DA_4_en.Rmd** available on the PRADO platform.

**Topics covered:**

- R packages required.
- Graphical exploration of data.
- Assumptions: normality and homogeneity of variance.
- Model validation.
- Visualization of the classifications.

[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.