

# Técnicas multivariadas

## Análisis de conglomerados (Cluster)

Mario J. P. López

Departamento de Matemáticas  
Programa de Estadística  
Universidad El Bosque

2020

## Objetivo

- Agrupar conjuntos de datos/individuos multivariados
- Los grupos obtenidos deben ser lo más homogéneos dentro y lo más heterogéneos entre sí

## Pasos

- 1 Selección de una medida de proximidad entre individuos
- 2 Selección de un algoritmo de agrupamiento

## Proximidad entre individuos

- Considere una matriz de datos,  $\mathbf{X}$  ( $n \times p$ ), con  $n$  individuos y  $p$  variables
- Una medida de distancias entre individuos,  $D$ , es una matriz de la forma:

$$D_{(n \times n)} = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix}$$

si  $d_{ij'}$  son distancias entonces  $D$  mide disimilaridad y si  $d_{ij'}$  son proximidades entonces  $D$  mide similaridad.

## Ejemplo

$$d_{ii'} = \begin{cases} \|x_i - x_{i'}\|_2, & \text{disimilaridad} \\ \max_{i, i'} \{d_{ii'}\} - d_{ii'}, & \text{similaridad} \end{cases}$$

- En el primer caso, valores grandes de  $d_{ii'}$  indican mayor heterogeneidad
- En el segundo caso, valores grandes de  $d_{ii'}$  indican mayor homogeneidad

## Ejemplo

En el caso de variables continuas se define la norma  $L_r$ ,  $r \geq 1$

$$d_{ii'} = ||x_i - x_{i'}||_r = \left( \sum_{j=1}^p |x_{ij} - x_{i'j}|^r \right)^{1/r}$$

- $r = 2$ : norma euclídea,
- $r = 1$ : Manhattan.

## Ejemplo

La norma  $L_2$  con métrica  $A$  ( $A > 0$ ) es

$$d_{ii'}^2 = \|x_i - x_{i'}\|_A = (x_i - x_{i'})^\top A (x_i - x_{i'})$$

- si  $A = I_p \rightarrow L_2$
- si  $A = \text{diag}(s_1^2, s_2^2, \dots, s_p^2) \rightarrow L_2$  no depende de las unidades de medida

## Ejemplo

El coeficiente de correlación entre individuos, denominado correlación  $Q$ , es una medida de proximidad.

$$d_{ij'} = \frac{\sum_{j=1}^p (x_{ij} - \bar{x}_i) (x_{i'j} - \bar{x}_{i'})}{\left[ \sum_{j=1}^p (x_{ij} - \bar{x}_i)^2 \sum_{j=1}^p (x_{i'j} - \bar{x}_{i'})^2 \right]^{1/2}}$$

donde

$$\bar{x}_i = \frac{\sum_{j=1}^p x_{ij}}{p}$$

es el promedio sobre todas las variables de un individuo.

## Datos binarios

Sea  $x_i^\top = (x_{i1}, x_{i2}, \dots, x_{ip})$  con  $x_{ij} \in \{0, 1\}$ , definiendo

$$a_1 = \sum_{j=1}^p I(x_{ij} = x_{i'j} = 1), \quad a_2 = \sum_{j=1}^p I(x_{ij} = 0, x_{i'j} = 1)$$

$$a_3 = \sum_{j=1}^p I(x_{ij} = 1, x_{i'j} = 0), \quad a_4 = \sum_{j=1}^p I(x_{ij} = x_{i'j} = 0)$$

$$\Rightarrow d_{ii'} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda (a_2 + a_3)} \quad (\text{similaridad})$$



## Similaridad

- Concordancia simple: si  $\delta = \lambda = 1$

$$d_{ij'} = \frac{(a_1 + a_4)}{p}$$

- Jaccard: si  $\delta = 0, \lambda = 1$

$$d_{ij'} = \frac{a_1}{(a_1 + a_2 + a_3)}$$

## Disimilaridad

- Binary (usada por R):

$$1 - J = \frac{a_2 + a_3}{a_1 + a_2 + a_3}$$

## Algoritmos de agrupamiento

- Jerárquicos:
  - aglomerativos: vecino más cercano, ward,...
  - divisivos
- No jerárquicos:
  - particionamiento: k-medias, k-medoides,...
  - mixtura de distribuciones
  - estimación de densidades

## Algoritmo aglomerativo

- 1 construir la partición más fina
- 2 calcular la matriz de distancias
- 3 encontrar los grupos con menor distancia
- 4 unir esos dos grupos en un solo grupo
- 5 calcular la distancia entre los nuevos grupos y obtener la matriz de distancias reducida

Repetir hasta tener un solo grupo con todos los individuos

## Medidas de distancia

La medida de distancia entre los nuevos grupos define el algoritmo. Para  $x_i \in A$  y  $x_{i'} \in B$ :

- Single Linkage (Nearest Neighbor)

$$D(A, B) = \min \{d(x_i, x_{i'})\}$$

- Complete Linkage (Farthest Neighbor)

$$D(A, B) = \max \{d(x_i, x_{i'})\}$$

- Average Linkage

$$D(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{i'=1}^{n_B} d(x_i, x_{i'})$$

## Medidas de distancia

- Centroid

$$D(A, B) = d(\bar{x}_A, \bar{x}_B)$$

Teniendo en cuenta que luego de que dos cluster, A y B, se unen

$$\bar{x}_{AB} = \frac{n_A \bar{x}_A + n_B \bar{x}_B}{n_A + n_B}$$

- Median: en lugar de usar  $\bar{x}_{AB}$  en el método Centroid usa

$$m_{AB} = \frac{1}{2} (\bar{x}_A + \bar{x}_B)$$

## Medidas de distancia

- Ward: une los grupos  $A$  y  $B$  con mínimo incremento de

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$$

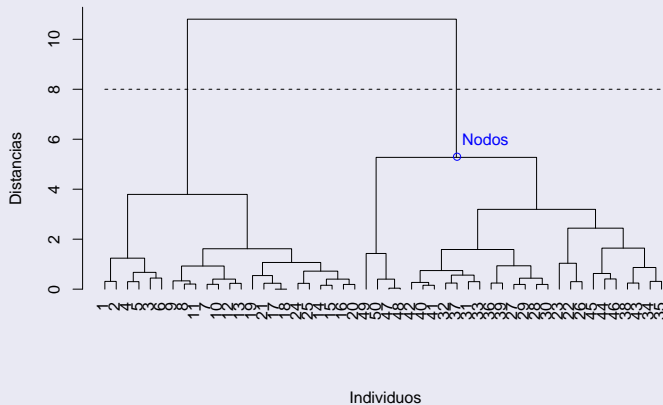
donde

$$SSE_A = \sum_{i=1}^{n_A} (x_i - \bar{x}_A)^T (x_i - \bar{x}_A)$$

$$SSE_B = \sum_{i=1}^{n_B} (x_i - \bar{x}_B)^T (x_i - \bar{x}_B)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (x_i - \bar{x}_{AB})^T (x_i - \bar{x}_{AB})$$

## Dendogramas



## Método $k$ -Means

- El objetivo es minimizar la variabilidad dentro de grupos
- El número de grupos es preestablecido
- La variabilidad dentro de grupos se puede medir como

$$W(C_k) = \sum_{x_i \in C_k} \|x_i - \bar{x}_k\|^2, \quad k = 1, \dots, G$$




donde

- $x_i$  es cada individuo que pertenece a un grupo  $C_k$
- $\bar{x}_k$  es el centroide o vector de promedios de cada grupo



## Algoritmo

- 1 Indicar el número de grupos,  $G$ , a formar,
- 2 Seleccionar  $G$  individuos de forma aleatoria (centroide inicial),
- 3 Asignar los individuos al centroide más cercano de acuerdo a la distancia euclídea,  $W(\cdot)$ ,
- 4 Calcular los nuevos centroides como el vector de promedios de cada grupo.  
Repetir los pasos 3 y 4 hasta convergencia.

-  Johnson, R., Wichern, D.  
*Applied Multivariate Statistical Analysis*.  
6th ed. Pearson, 2007.
-  Mardia, K., Kent, J., Bibby, J.  
*Multivariate Analysis*.  
Academic Press, 1995.
-  Rencher, A.  
*Methods of Multivariate Analysis*.  
2nd ed. Willey, 2002.