



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

TAREA 12

EJERCICIOS

1. Considera los datos ya manejados en anteriores tareas en el fichero `coches.dat` (también en forma de dataframe, en `coches.frame`). Sobre ellos puedes emplear la distancia que consideres oportuna.
 - a) Haz un análisis jerárquico de agrupamientos. Constata las diferencias en el dendrograma resultante consecuencia de emplear ligadura simple y alguno de los otros métodos (ligadura completa o promedio).
 - b) Decide el número de grupos o tipos de coche de acuerdo con la valoración que de ellos hacen los usuarios.
 - c) Haz ahora un análisis de reasignación empleando el algoritmo k -medias o el (bastante análogo) de partición en torno a medoides.
2. Repite el ejercicio anterior con los datos sobre elecciones en Vizcaya (datos en `elec96.dat` o en la dataframe `Bizkaia.frame`). Compara la tipología de municipios que obtengas con las pautas visibles en los análisis previos que has hecho (en componentes principales, factorial correspondencias, reescalado multidimensional. . .).

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

- **Bibliografía.** Los libros [3], [7], [8], [2], [5], y [1] son manuales de Análisis Multivariante en general, que en alguna medida hablan también de análisis cluster.

Tienes manuales específicos como [4] (muy antiguo, pero aún de interés), [6] (que describe las funciones en el paquete de R `cluster`) o [9].

Todo está en Biblioteca.

- **Sugerencias sobre disimilaridades a emplear.** Valen las observaciones hechas anteriormente. Podrías, dado que todas las variables son numéricas en los dos casos, emplear simplemente distancia euclídea (en el ejemplo de las elecciones, sobre los datos expresados en porcentajes) No obstante, dado que las varianzas son diferentes —sobre todo en el segundo caso— y hay una muy acusada correlación entre las variables, parece más indicado emplear una distancia como la de Mahalanobis definida como:

$$d^2(i, j) = (\mathbf{x}_i - \mathbf{x}_j)' C^{-1} (\mathbf{x}_i - \mathbf{x}_j) \quad (1)$$

en que C es la matriz de covarianzas muestral.

- **Funciones en R.** En el paquete `mva` dispones de algunas funciones para análisis cluster jerárquico convencional y k -medias. En el paquete `cluster` dispones de las funciones descritas en [6].

Referencias

- [1] C. Chatfield and A.J. Collins. *Introduction to Multivariate Analysis*. Chapman & Hall, London, 1980.
- [2] C.M. Cuadras. *Métodos de Análisis Multivariante*. Eunibar, Barcelona, 1981.
- [3] W.R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications*. Wiley, New York, 1984.
- [4] J.A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.
- [5] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.
- [6] L. Kaufman and P.J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. Wiley, 1990. Signatura: 519.237.8 KAU.
- [7] D. Peña. *Análisis de Datos Multivariantes*. McGraw-Hill, 2002.
- [8] G.A.F. Seber. *Multivariate Observations*. Wiley, New York, 1984.
- [9] H. Späth. *Cluster Analysis Algorithms*. Ellis Norwood Ltd., Chichester, England., 1980.