QMM: Exercise Sheet 5 - Clustering

Fabien Baeriswyl, Jérôme Reboulleau, Tom Ruszkiewicz

Exercise 1. For this exercise, use the `Cars.csv` datafile. This dataset consists of 32 cars from different manufacturers, for which we record information about their prices, fuel consumption, and power notably. We want to cluster the data.

a) First, provide a quick exploratory analysis of the data. Use visuals.

b) Create a function that computes the Euclidean distance between two vectors (of same size, but this size can be anything). Use it to find the distance between the Ford Mondeo LX and the Ford Galaxy LX (the first two observations of the dataset).

c) Compute the distance matrix between all the 32 observations.

d) Run a single-linkage hierarchical clustering on the data. Plot the result. Discuss any specific feature.

e) Run a complete-linkage hierarchical clustering on the data. Plot the result. Discuss any specific feature.

f) Run an average-linkage hierarchical clustering on the data. Plot the result. Discuss any specific feature. Comparing it to the two previous models, which one is best? Briefly explain.

g) Under the complete-linkage and the average-linkage cluster models, using 4 clusters, provide the repartition of the observations in these 4 clusters.

h) Rescale your dataset. Then, run the average-linkage model again and discuss any change.

Exercise 2. For this exercise, use the `Country.csv` datafile. This dataset consists of 109 countries for which we have indicators about their gdp, literacy and urban population among others.

a) Rescale the data.

b) Use a visual approach to determine how many clusters you should use in a k-means partitioning method.

c) Run k-means clustering without setting the `nstart` argument to anything. Observe the changes while running this function over and over again. Comment briefly.

d) Run k-means clustering setting `nstart` to 25. Plot the result and discuss briefly.

e) Run k-means clustering using 3 clusters. Discuss on the changes.

f) Plot the gdp versus the literacy and emphasise the clusters from your 3-cluster model from point e).

Exercise 3. For this exercise, use the `South.csv` datafile. This consists of 16 US states and their characteristics in 1965, among which their mean temperature and precipitation, income, share of african americans and GOP vote share (for Grand Old Party, formally the Republican Party) for various elections. Run clustering methods of your choice to cluster the states using the tools from the previous exercises. Use the resulting clusters to discuss some cluster particularities.