Statistical Methods 2

Portfolio 4 – Cluster analysis

Complete the following tasks and submit your work on Blackboard by 4pm on Friday 17/02/2023

Task 1 (30 marks)

Using `R`, apply agglomerative clustering and K-means clustering on a dataset of your choice (e.g. on one of the dataset you used in Portfolio 1-3). Remark that agglomerative clustering can e.g. be implemented in `R` using the command `hclust` while K-means clustering can be implemented using the command `kmeans`.

The following constraints must be satisfied:

- For agglomerative clustering,
  - You need to justify the measure of distance between observations that you use.
  - You need to mention which method is used to measure the distance $d_{t,(r,s)}$ between an old cluster t and a new cluster obtained by merging cluster r and cluster s.
  - A dendrogram should be used to summarize the agglomerative clustering process.
  - You need to choose the number of clusters to use, and discuss your results.
  - Use a two dimensional reduction of the data to visualize the different clusters.
  - Discuss whether or not agglomerative clustering is a suitable clustering method for your dataset.
- For K-means clustering,
  - You need to justify the measure of distance between observations that you use.
  - Lloyd's algorithm should be used.
  - You need to justify your choice for K, and discuss your results.
  - Use a two dimensional reduction of the data to visualize the different clusters.
  - Discuss whether or not K-means clustering is a suitable clustering method for your dataset.
  - Perform K-means clustering on the first $q < p$ principal components of your dataset. Justify your choice for q and discuss the results.

Task 2 (30 marks)

Using `R` and a dataset of your choice, illustrate the benefit of spectral clustering over K-means clustering. For this task you can use a simulated dataset if you want (e.g. the one you used in Portfolio 3 on kernel PCA) and you can perform spectral clustering using e.g. the `R` package `kernlab`.

The following constraints must be satisfied:

- Your example should illustrate the fact that spectral clustering can be useful in situations where K-means clustering does not work well. This means that you should perform your analysis using both spectral and K-means clustering, and that the results obtained with the former method should be significantly "better" than those obtained using the latter.
- You must perform the analysis for different measures of similarities, and assess how the results obtained with spectral clustering are sensitive to the similarity measure that you use.
- Perform spectral clustering with the number of cluster M chosen as discussed in the lecture notes, and assess if this choice for M works well for your dataset.
- You must comment all your results.

Task 3 (40 marks)

We saw in Chapter 5 that K-means clustering performs poorly when the different groups of observations are not linearly separable, and that for such datasets spectral clustering is a more suitable clustering method.

Kernel K-means clustering is another clustering method that can be used when the different groups of observations are not linearly separable. As we saw in Chapter 4, projecting the data points $\{x_i^0\}_{i=1}^n$ in a high dimensional space using a mapping $\Phi : \mathbb{R}^p \to \mathbb{R}^m$ (with $m \in \mathbb{N} \cup \{\infty\}$) often makes them more likely to be linearly separable. Based on this observation, kernel K-means clustering simply amounts to performing K-means clustering on the transformed dataset $\{\Phi(x_i^0)\}_{i=1}^n$.

- Letting $k(x, x') = \Phi(x)^\top \Phi(x')$, show that to perform K-means clustering on the transformed observations $\{\Phi(x_i^0)\}_{i=1}^n$ it is enough to be able to evaluate $k(\cdot, \cdot)$ pointwise (and thus that we do not need to explicitly compute the transformed observations $\{\Phi(x_i^0)\}_{i=1}^n$).
- Considering the same dataset that you used in Task 2, on which K-means clustering perform poorly,
  - Perform kernel K-means clustering in `R`, e.g. using the function `kkmeans`. (Remark: it is unclear which K-means algorithm is used by this function.)
  - Justify your choice for the kernel $k(\cdot, \cdot)$. If $k(x, x') = f(\|x - x'\|/c)$ for some $f : \mathbb{R} \to \mathbb{R}$, only justify your choice for the bandwidth parameter $c > 0$.
  - Justify your choice for the number K of clusters to use.
  - Compare the results with those obtained in Task 2 with spectral clustering. Which method seems the most appropriate for your dataset?