TMA4268 Statistical Learning

Chapter 10: Solution sketches

Thiago G. Martins, Department of Mathematical Sciences, NTNU

Spring 2021

Recommended exercise 1

- For the New York Times stories dataset:
  - Create a biplot and explain the type of information that you can extract from the plot.
  - Create plots for the PVE and Cumulative PVE. Describe what type of information you can extract from the plots.

The pca-examples.rdata can be downloaded from the Blackboard.

Recommended exercise 2

Show that the algorithm below is guaranteed to decrease the value of the objective

$$\underset{C_1,\ldots,C_k}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

at each step.

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Recommended exercise 3

Perform k-means clustering in the New York Times stories dataset.

The pca-examples.rdata can be downloaded from the Blackboard.

Recommended exercise 4

Perform hierarchical clustering in the New York Times stories dataset.

The pca-examples.rdata can be downloaded from the Blackboard.