Statistical Learning – Data science - 2022/23 – Exercise 6 – 18/05/2023

Exercise 6: Analysis of Human Tumor Microarray dataset – unsupervised learning, clustering with k-means

Please, execute the following tasks and provide answers to the proposed questions.

1. Download the "14-cancer microarray data" from the book website https://web.stanford.edu/~hastie/ElemStatLearn/

- Get information about the dataset in file 14cancer.info and in Chapter 1 (page 5) and Chapter 14 (page 512) of the book (Hastie et al., 2009)

2. Generate a new Kernel SLDatascience_EX5_HumanTumorMicro_Clustering_Surname in Kaggle

3. Load the data in Kaggle

- Use, for instance, the training set gene expression data
- Load also the labels

4. Use the sklearn.cluster module to perform clustering analysis on the dataset. In particular, repeat the analysis proposed in section 14.3.8 of the book (Hastie et al., 2009)

- Preprocess data, if needed (e.g., if there are missing values then remove the related columns and rows)
- Start using K-means and then test some other clustering algorithms at your choice
- Cluster the samples (i.e., columns). Each sample has a label (tumor type)
- Do not use the labels in the learning phase but examine them posthoc to interpret the clusters
- Run k-means with K from 2 to 10 and compare the clusterings in terms of within-sum of squares
- Show the chart of the performance depending on K
- Select some K and analyze the clusters producing tables such as those displayed in the book

Reference

[Hastie et al., 2009] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second edition). Springer. 2009. https://web.stanford.edu/~hastie/ElemStatLearn/