Assignment 3

Deadline: 26/6/2023

Instructor: Panagiotis Papastamoulis
papastamoulis@aueb.gr
Department of Statistics
Athens University of Economics and Business

Exercise 1 (Gene expression measurements). Consider the Ramaswamy dataset available from http://www-stat.stanford.edu/~hastie/glmnet/glmnetData/, which contains 16063 gene expression measurements and 198 samples belonging to 14 distinct cancer subtypes. The aim is to construct a rule for predicting **cancer types** based on gene expression measurements. Evaluate the predictive performance of your proposed classifier by cross-validation.

Exercise 2 (Clustering coffee samples). The coffee dataset available on the pgmm package contains data measuring the chemical composition of coffee samples collected from around the world, comprising 43 samples from 29 countries. Each sample is either of the Arabica or Robusta variety. The first two columns contain "Variety" and "Country", respectively. The purpose is to cluster the dataset (without of course taking the ground-truth classification into account) based on the chemical properties measured in the last 12 columns of the data frame. The number of clusters is assumed unknown. For this purpose you may use distance-based methods, partition methods as well as model-based clustering methods. For the model-based clustering methods use standard mixture of normals (mclust) as well as factor analytic ones (pgmm, fabMix) considering that the number of factors is at most equal to 2 (for pgmm use both random starting values as well as k-means starting values with the option: zstart = 1 and zstart = 2, respectively). Compare the estimated clustering and the ground-truth partition of the data in terms of the adjusted Rand Index. Obtain the dataset using the commands:

```
> library(pgmm)
> data(coffee)
> x <- coffee[,-c(1,2)]
> x <- scale(x)
```

(the input data for clustering is x).

Exercise 3 (Social network). Consider a sample (minimum: 20 persons) from your personal social environment (e.g. colleagues, family friends, college friends, etc...). Construct the corresponding friendship network (that is: the adjacency matrix takes the value 1 whenever two persons are friends and 0 otherwise). Visualize, analyze and cluster the data using appropriate method(s) and discuss the results.