Statistical Learning with High-Dimensional Data

Duration: ~ 3 hours - All documents are allowed - The exam should be done independently
Reports may be a combination of handwritten notes, a Rmarkdown document and a PDF version
The final reports have to be uploaded on Moodle (PDF + RMD).

1 Exercise 1: general questions

1. Describe some techniques allowing to select the number of clusters when clustering with Gaussian mixture models.
2. Describe the general setup and the goal of double cross-validation.

2 Exercise 2: k-means clustering

We have the following data points in $\mathbb{R}^2$:

Indiv.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Var. 1	0	0	1	3	3.5	1	3	4
Var. 2	1	2	1	1	1	5	4	5

1. Explain in a few sentences how we can relate the quality of a clustering with the notions of variance.
2. We propose to use the k-means clusteriong technique (with the Euclidean distance). Apply this algorithm to the above data set to cluster them into 2 groups (all calculations have to be detailed).

3 Exercise 3: the Vélib data

The objective of this study is to analyze a data set coming from the Vélib system in Paris (a bike sharing system). The data are loading profiles of the bike stations over one week. The data were collected every hour during the period Sunday 1st Sept. - Sunday 7th Sept., 2014.

3.1 Loading the data

The data can be loaded within R as follows:

```
load('path/to/the/data/velib.Rdata')
```

3.2 Pretreatment et descriptive analysis

We consider the 1189 Vélib stations as the individuals of this study. First, do all required pretreatments and the usual descriptive analysis of the data. A selection of the most useful data will be probably necessary at first.

3.3 Data visualization

Use PCA to visualize the data. Choose the number of PCA axes to retain for the visualization and interpret the results. In particular, the PCA axes should be explained regarding the original variables.

3.4 Clustering

3.4.1 Hierarchical clustering

Apply the hierarchical clustering with the appropriate distance, choose the right number of cluster and comment the results. It will be important to explain why the retained distance is the appropriate one for these data.

A map of the results may be obtained using the GPS coordinates of the stations, thanks to the leaflet package:

```
palette = colorFactor("RdYlBu", domain = NULL)
leaflet(X) %>% addTiles() %>%
       addCircleMarkers(radius = 3,
       color = palette(clusters),
       stroke = FALSE, fillOpacity = 0.9)
```

3.4.2 k-means

Apply now the k-means clustering on the same data. Choose also the right number of clusters using the appropriate technique. Comment and compare with the result obtained with the hierarchical clustering.

3.5 Summary

It is expected a final summary of all information extracted during the analysis.