Machine learning

M. Mougeot

ENSIIE, December 8th 2020

Instructions

- `Nature of the work:`

-A unique jupyter notebook is delivered for this project. The notebook contains the python instructions to answer to the questions but also the comments or any remarks you want to share (in comment or in markdown)

-The **name of the jupyter notebook** file is **your lastname** (without space or accentuated letter).

-Moreover, **the following information should be written in comment in the first line of the notebook : NAME, FORNAME, University (ENSIIE or UEVE), Double diploma if any (M2DS, M2IF, M2IA or other)**

- `Delivery of the work:`

-This work should first be started and accomplished between 9h-12h30 Tuesday 8th December.

-The work done should be delivered before 12H30.

-ENSIIE students (even if they follow a double diploma in UEVE) should download their work on the ENSIIE exam website, repository **MAL2020proj2**, (exam.ensiie.fr).

-Other Students should send by mail their notebook file at **mal.ensiie@gmail.com**

- `Warning:`

No grade will be delivered if the previous instructions are not strictly followed.

The three exercices are independant.

Don't forget to mention the exercice and the question number before to answer.

Exercice I

The data

The file ‘ cluster1.txt" contains $n = 100$ two-dimentional points $\{(y_i, x_i),\ 1 \leq i \leq n\}$.

The first colum corresponds to the horizontal axis (x), the second colum to the vertical axis (y)

The data can be uploaded in a python environment using the following instructions:

```
import pandas as pd
tab=pd.read_csv('cluster1.txt',sep=';',header=None)
X=np.array(tab);
N,p=np.shape(X)
```

Following Figure displays the $n = 100$ points of the `cluster1.txt` data set.

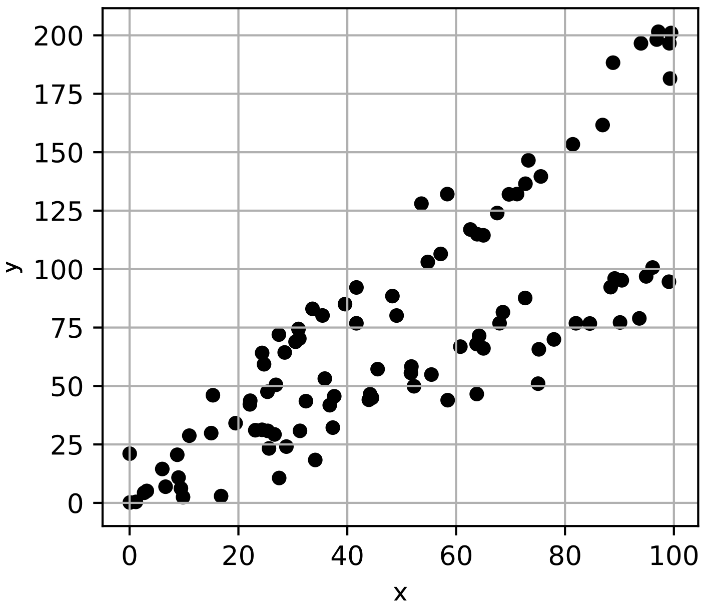

Figure 1: Raw data set

II.A Clustering pre-processed data

The goal of this section is to propose and implement a first clustering model on the observations of the data set `cluster1.txt`. We assume that there exits $K = 2$ clusters. This first clustering is not performed on the raw data (observations of `cluster1.txt`) but on pre-processed data.

A.1. Implement a Principal Component Analysis (PCA) on the raw data.

A.2. Visualize the histogram on the projected observation on the two principal axes. Which axis seems to be more appropriate to cluster easily the data into two clusters? justify your answer.

A.3a Use the Gaussian Mixture Models (GMM) clustering algorithm to cluster the projected observations on the previous chosen PCA axis. Recall the GMM model used and provide the estimated parameters of the model.

A.3b We aim now at model the observations in each group. What model do you advice to propose? Estimate the parameters of the models in each group and provide the final estimated model in each group.

A.4a Use the K-means algorithm to cluster the projected data on the previous chosen PCA axis.

A.4b We aim now at model the observations in each group. What model do you advice to propose? Estimate the parameters of the models in each group and provide the final estimated model in each group.

A.5. Conclusion.

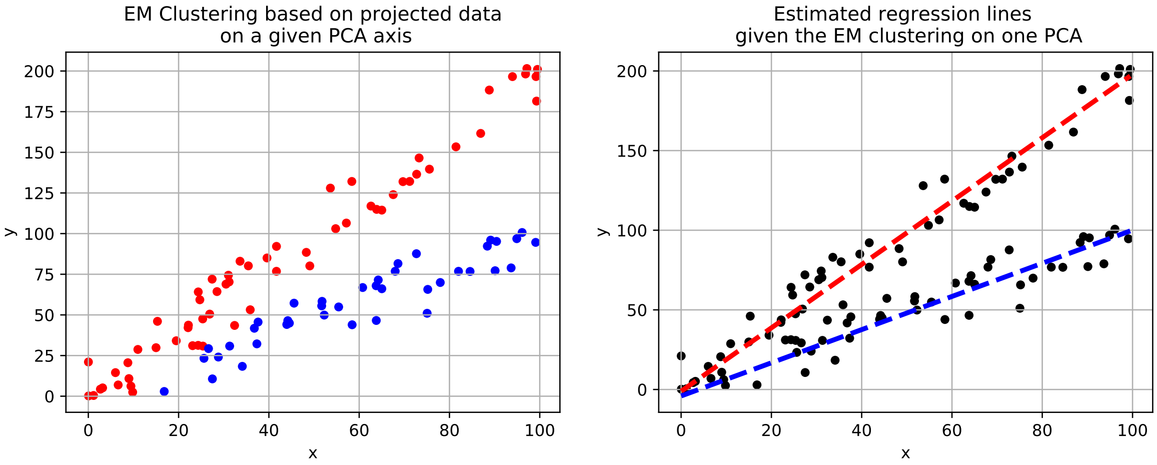

Figure 2: EM Clustering using projected observations on a given PCA axis and estimated models.

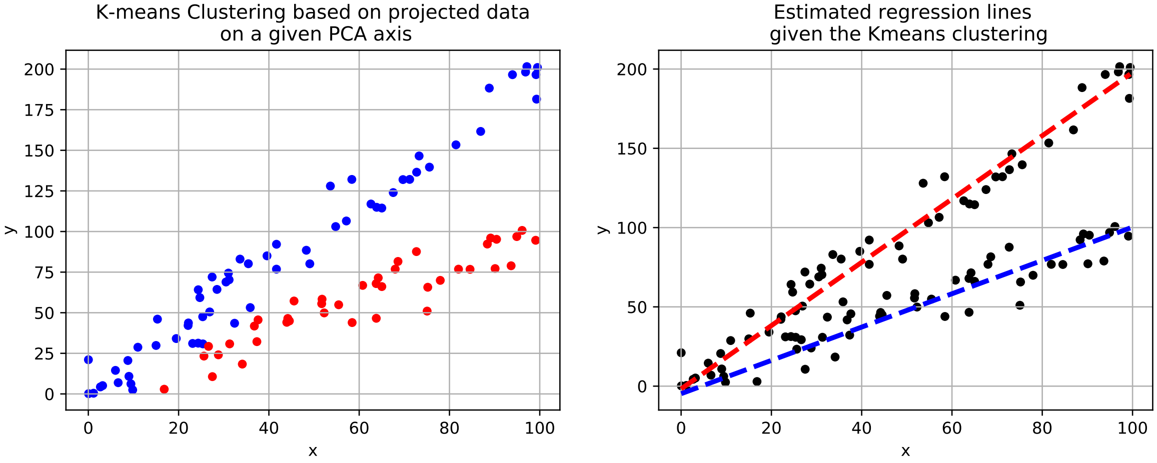

Figure 3: Kmeans Clustering using projected observations on a given PCA axis and estimated models.

I.B. Clustering raw data

The goal of this section is to propose and implement an appropriate clustering model directly on the 2D raw observations of the data set `cluster1.txt` assuming two main informations:

- The observations are assumed to be split in $K = 2$ clusters.
- The observations of each cluster can be modeled using a linear model.

B.1 Propose and justify the use of Gaussian Mixture Modelling (GMM) to cluster the observations.

B.2 What parameters do you propose to consider for this GMM?

B.3 Explain how it is possible to estimate the parameters using the Expectation Maximization algorithm

We consider in this following that the intercepts of the two underlaying regression models (for each cluster) are equaled to zero.

B.4 Implement the EM algorithm to automatically compute the two underlaying clusters.

B.5 Use your algorithm to compute for each cluster, the parameters of the underlaying linear model.

Following figures display the clustering results using the EM algorithm implemented with an appropriate model of each cluster (left) and the underlaying regression models in each clsuter

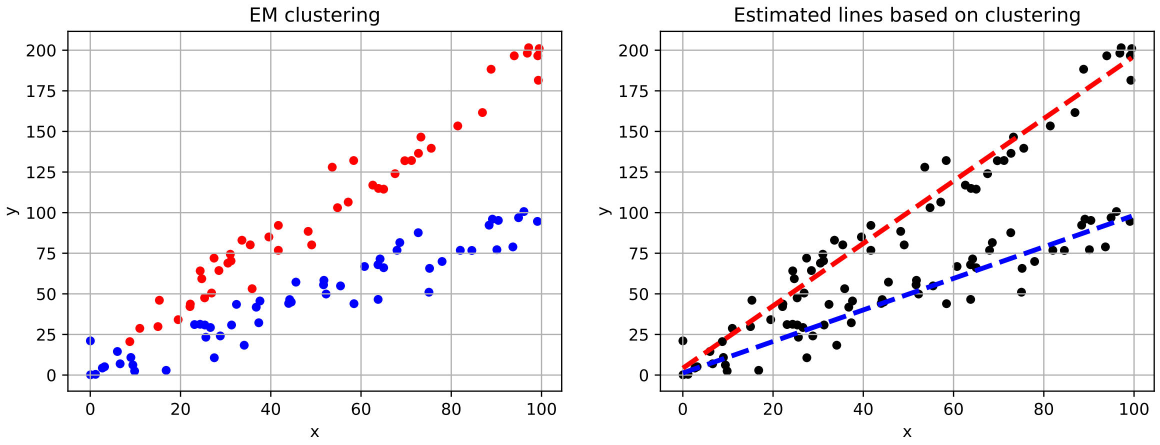

Figure 4: EM Clustering of the data set using an EM algorithm (left). Visualization of the estimated regression lines in each cluster

II Signals representation for data mining.

We are given two sets of signals, denoted by A and B, recorded at 101 equidistant time points.

```
import csv
tabA = pd.read_csv("signalA.csv",header=1);
tabA=np.array(tabA); tabA=np.transpose(tabA[:,1:55]);
n1,p1=np.shape(tabA)
tabB = pd.read_csv("signalB.csv",header=1);
tabB=np.array(tabB);  tabB=np.transpose(tabB[:,1:40]);
n2,p2=np.shape(tabB)
```

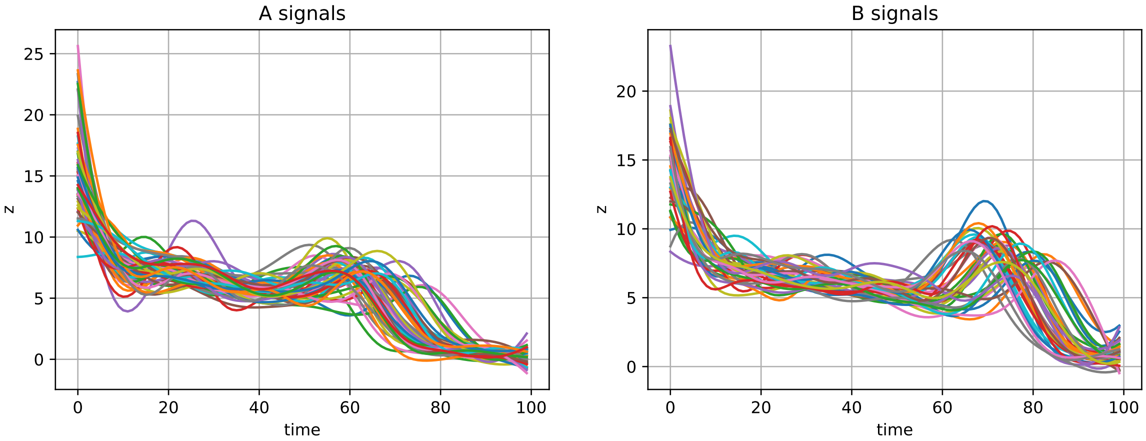

Figure 5: A and B signal families.

It is not clear whether these two signals are the subsets of the same population group. We will denote the combined dataset by C.

C.1 Compute the sample mean of each dataset, A, B and C, and compare them. Do you think they could be assumed to have the same mean pattern?

For the purpose of the following analysis, we will use the combined dataset C.

C.2 Perform principal component analysis to find a parsimonious representation of the signals. Display the first 4 components and describe characteristics of those components, by commenting on similarities and differences in the shape of the components.

C.3 Find an approximate representation of the signal using the finite number of components that can explain at least (1) 85% and (2) 95% of the total variability in the data. Is there any difference between these two representations? Select two signals from the dataset and visualize the approximation as the number of the components increases. Comment on the quality of the approximations.

C.4 For the first 4 components you have found, examine the distribution of the projection coefficients for each component. Assess whether a normal distribution is a good approximation.

C.5 For the first 4 component you have found above, make a scatter plot of the projection coefficients for each possible pairs: (1,2), (1,3), (1,4), (2,3), (2,4), (3,4). Use two different symbols to represent points from A and B. Examine the plots to see if there are any patterns associated with group labels (symbols). Based on your findings, do you think the datasets A and B could have come from the same population?