Practical work 08 – 6th of November 2018
Clustering algorithms

Summary for the organisation :

— Submit the solutions of the practical work before Monday 12h00 next week in Moodle.
— Preferred modality : **one**[1] pdf generated from your iPython notebook(s) and an archive with your notebook(s).
— Alternative modality : one pdf report with annotated code and outputs.
— The file name must contain the number of the practical work, followed by the names of the team members by alphabetical order, for example `02_dupont_muller_smith.zip`.
— Put also the name of the team members in the body of the notebook (or report).
— Only one submission per team.

Context

The goal of this practical work is the implement by yourself the k-means algorithm and to experiment with the different parameters of this algorithm.

Exercice 1 Getting the data

a) Load the two given datasets :

```
X1,label1 = pickle.load(open("dataset_1.pkl","rb")
X2,label2 = pickle.load(open("dataset_2.pkl","rb")
```

b) Visualize the data using various color for each unique labels like in figure 1 :

[1] And only one so use a tool to merge your pdfs if more than one.

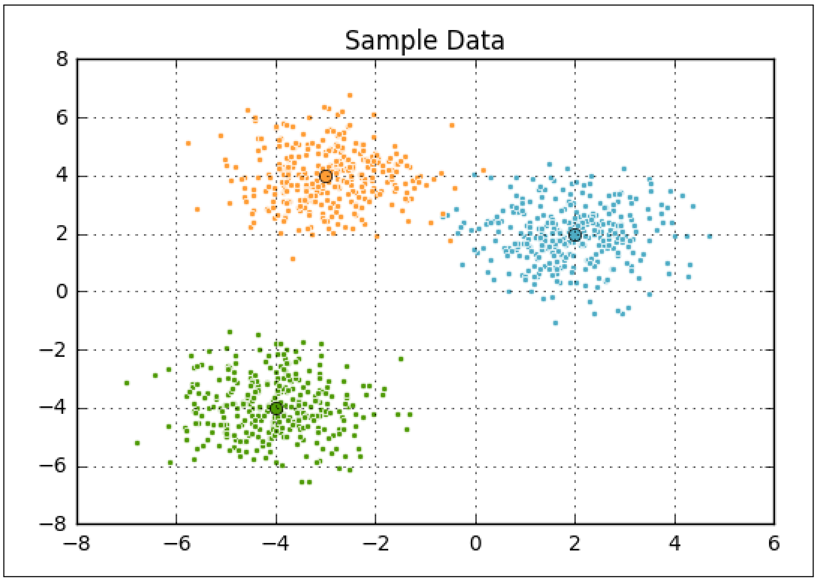

FIGURE 1 – Data visualization

Exercice 2 The k-means algorithm

Using numpy, implements the k-means algorithm as follow :

a) Initialise the centroids $\mu_1, \mu_2, \ldots, \mu_K$.

b) Until convergence :

   i) Find the closest centroid for each point

   ii) Reevaluate the centroids

c) Return the centroids and the label predicted.

We also ask you to define and implement strategies for the :

- Initialisation of the centroids.
- Convergence criteria.

Exercice 3 Evaluate your model

At this point, your k-means algorithm is working :

- Visualize your convergence criteria over the epochs [2] using the dataset 1.
- Visualize the output of your k-means on the dataset 1.
- Do you experience sensitivity to the initial values of the centroids ? Is your strategy for initialization working well in most cases ?
- Document your convergence criteria. Could you think about other convergence criteria ?
- Visualize your convergence criteria over time using the dataset 2.
- Visualize the output of your k-means on the dataset 2 and comment your results.

[2] One epoch is a complete visit of the training set.