Machine Learning: Fifth Home Work

Clustering with K-means and Gaussian Mixture Models

Edoardo Ghini

December 17, 2016

Dipartimento di Ingegneria dell'Università di Roma La Sapienza

Contents

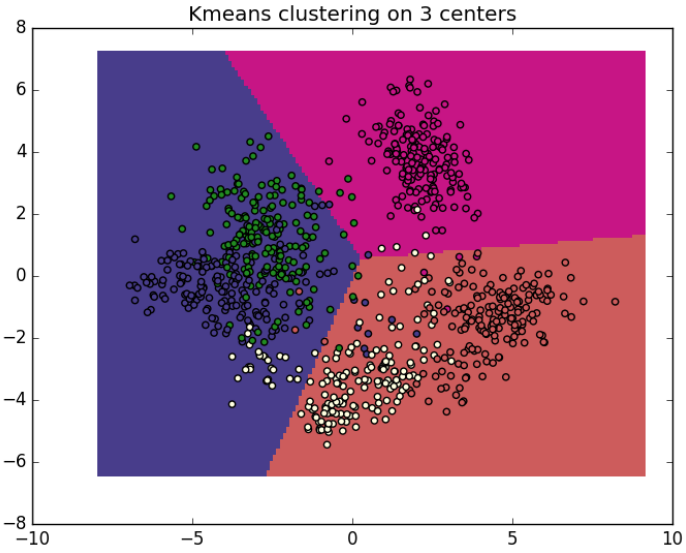


Figure 1

Part I

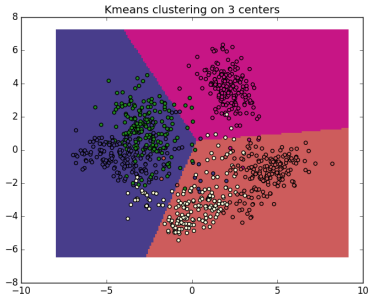
Introduction

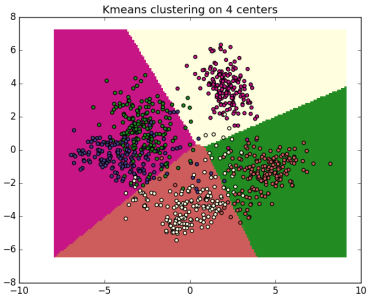
1 Scope

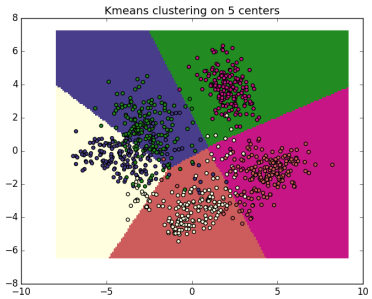
Familiarise with unsupervised techniques to analyse datas.

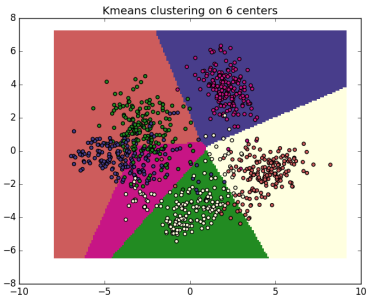
2 Objectives

The assignment consider the usage of two effective approaches to clustering problems and an analysis of the results with three different metrics.









Part II

Development

3 Data Manipulations

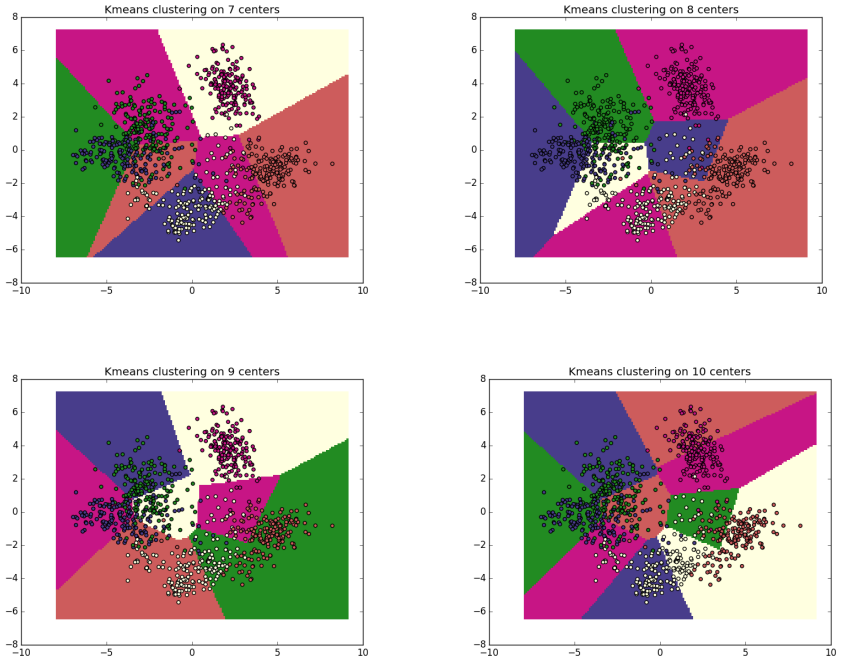
At the beginning I had to load a dataset and applying a principal component analysis extracting the first two most significative vectors of data in order to being able to visualise them on a two-dimensional space.

4 Clusterization

4.1 Kmeans

4.1.1 Execution

At first, I iterated the Kmeans algorithm changing the number of clusters to define, as can be seen from the plots.



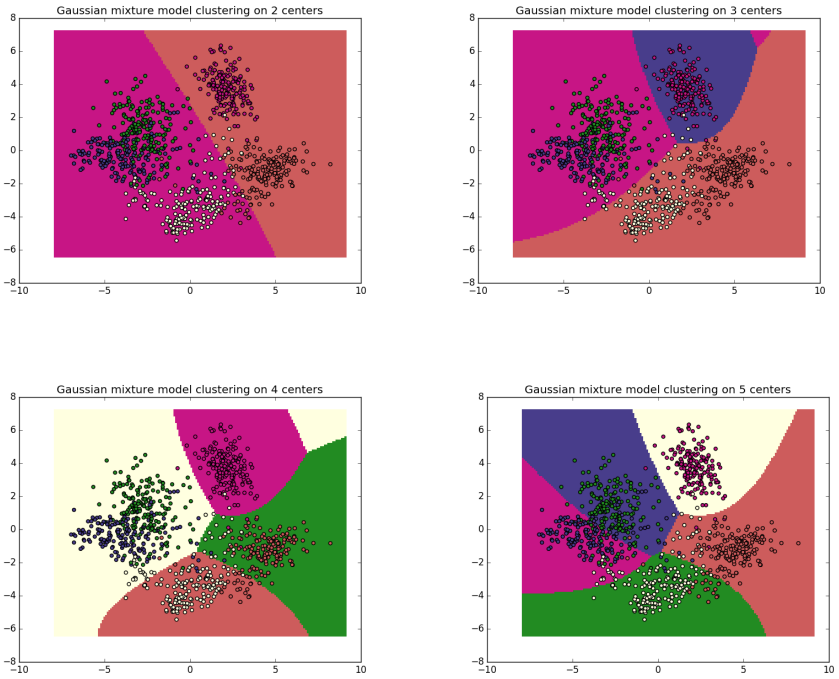
4.1.2 Score Analysis

After the iteration was completed, I applied three different metrics to evaluate the effectiveness of the model feeding the evaluators with the predictions coming from the model and the true labels of the data. In fig(5) there are the results.

4.2 GMM

4.2.1 Execution

In this other case, I made the same iteration as before, but this time I adopted an approach based on a Gaussian Mixture Model. This strategy allows to define non-linear boundaries of the clusters, therefore it brings more flexibility finding the best separators between clusters. The following plots show how the GMM recognise different numbers of clusters within data.


Gaussian mixture model clustering on 2 centers
Gaussian mixture model clustering on 3 centers
Gaussian mixture model clustering on 4 centers
Gaussian mixture model clustering on 5 centers

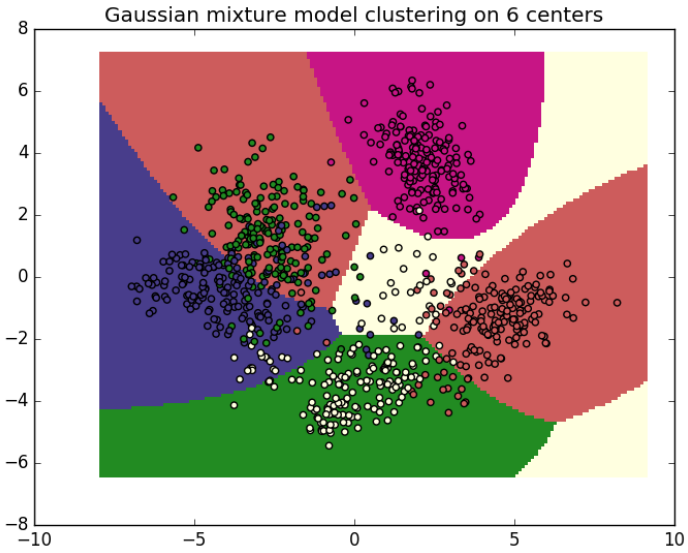
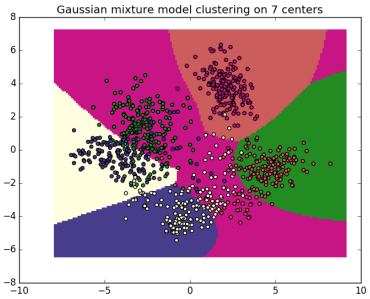
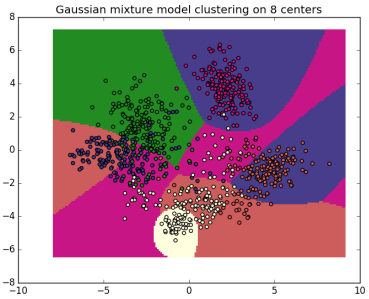
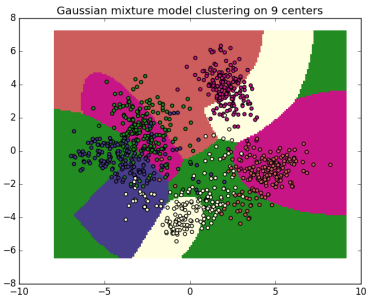
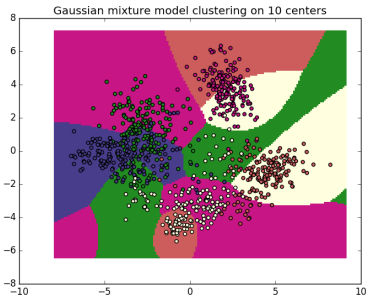

Gaussian mixture model clustering on 6 centers

Figure 8









4.2.2 Score Analysis

In the same fashion as before, in fig(10) there are results about three different metric evaluations according to predicted and real labels.

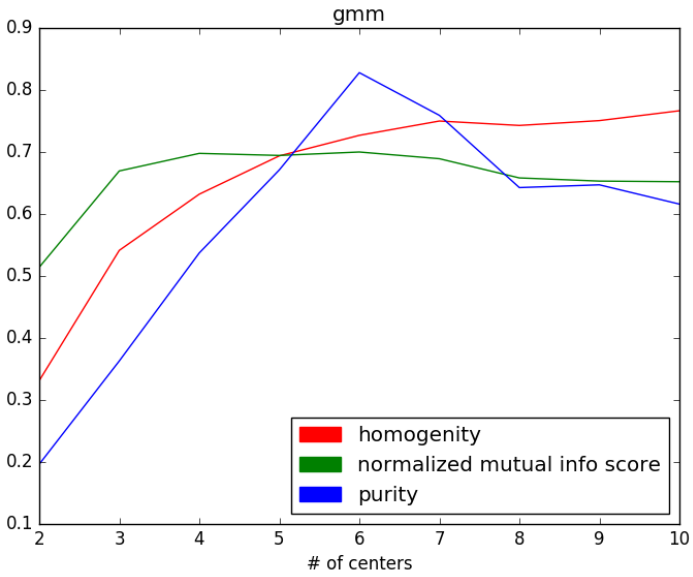


Figure 11

Part III

Conclusions

In the final analysis the best score of the various experiments is extremely metric-dependent. In fact, the metric choice is a major concern when we are dealing with unsupervised learning. For this particular case, the purity seems to be the highest considering only six clusters.
Finally, the three metric scores have similar behaviour with the variation of the number of cluster in the respect of both the approaches (Kmeans and GMM).