Report Machine Learning, Homework 5

Domenico Alfano

The report describes the results obtained during the development of Homework 1. The main goal of the assignment has been to evaluate the capabilities and functionalities of Clustering with K-Means and GMM algorithms.

1 Introduction

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering that is an example of unsupervised learning: it is not based on predefined classes and on samples of training labeled, clustering is a form of learning by observation rather that of learning for example. K-means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The algorithm takes as input a parameter K and partitions a set of n objects into K clusters in so that the intra-cluster similarity is high while the inter-cluster similarity is low. The main idea is to define K centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. In addition to K-means method there is another way to deal with clustering problems: a model-based approach, which consists in using certain models for clusters and attempting to optimize the fit between the data and the model. In practice, each cluster can be mathematically represented by a parametric distribution, like a Gaussian. The entire data set is therefore modelled by a mixture of these distributions. So, the GMM (Gaussian Mixture Model) is one of the most widely used clustering method.

2 Clustering with K-Means

For this section the clustering was perform by using the K-means algorithm. Only the first 5 classes $\{0, 1, 2, 3, 4\}$ of the dataset was used to perform the clustering. In the first test 5 clusters was requested from KMeans by using the function KMeans(). Figure 1 shows the centroids of the 5 clusters defined by KMeans.

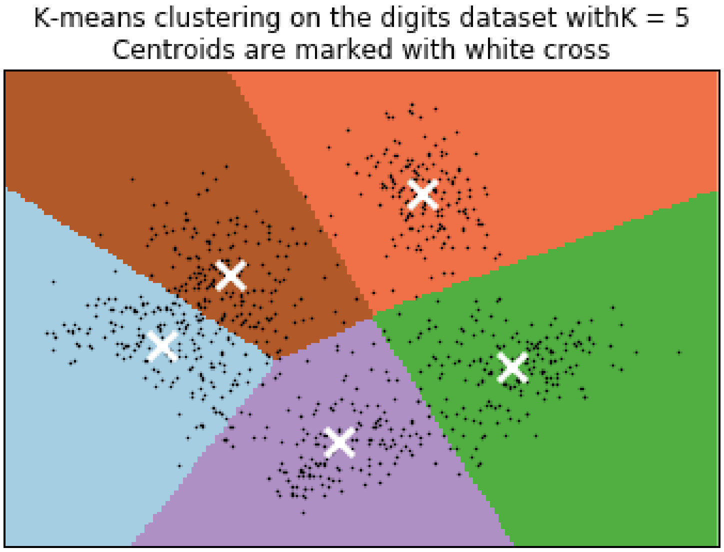


Figure 1

In the second test the clustering process was perform from 3 to 10 clusters by using the same 5 classes of X.

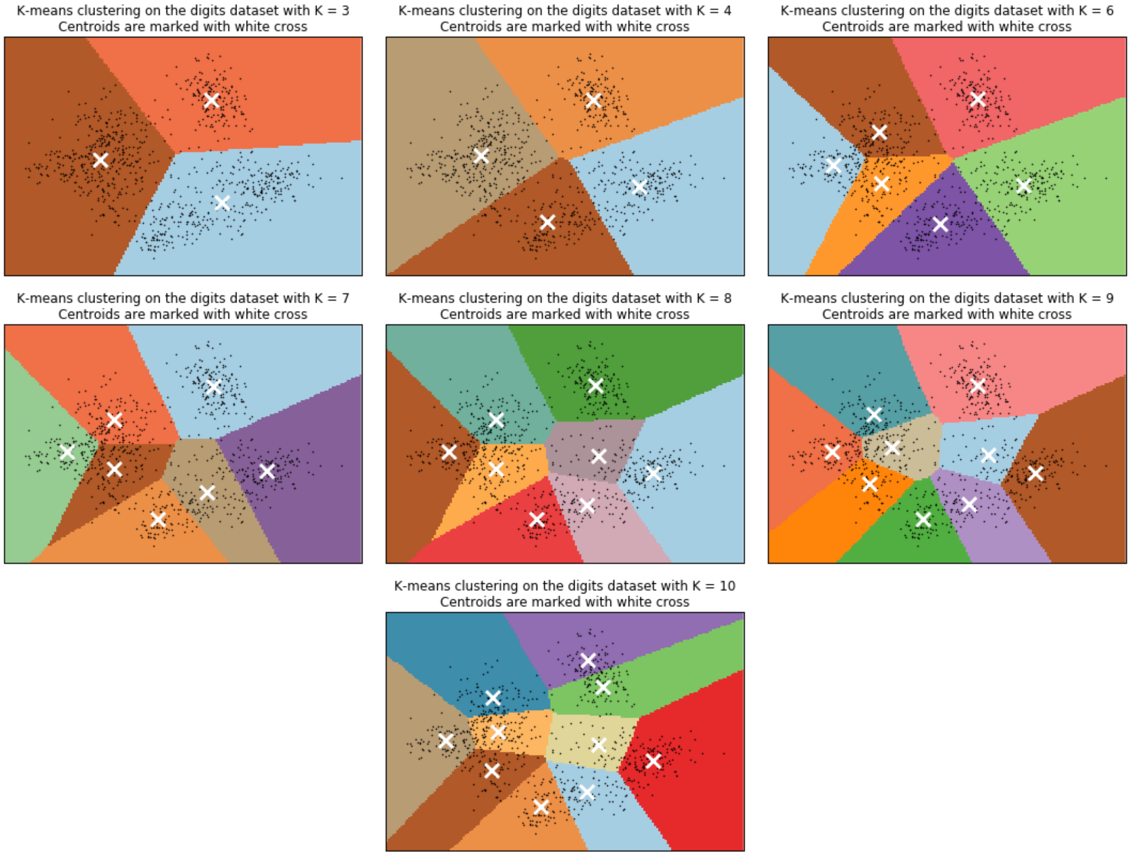


Figure 2

3 Clustering with GMM

For this section the clustering was perform by using the GMM algorithm. As requested in this part of Homework, only the first 5 classes $\{0, 1, 2, 3, 4\}$ of the dataset was used to perform the clustering. As we can see in Figure 3, the clustering process was perform from 2 to 10 clusters.

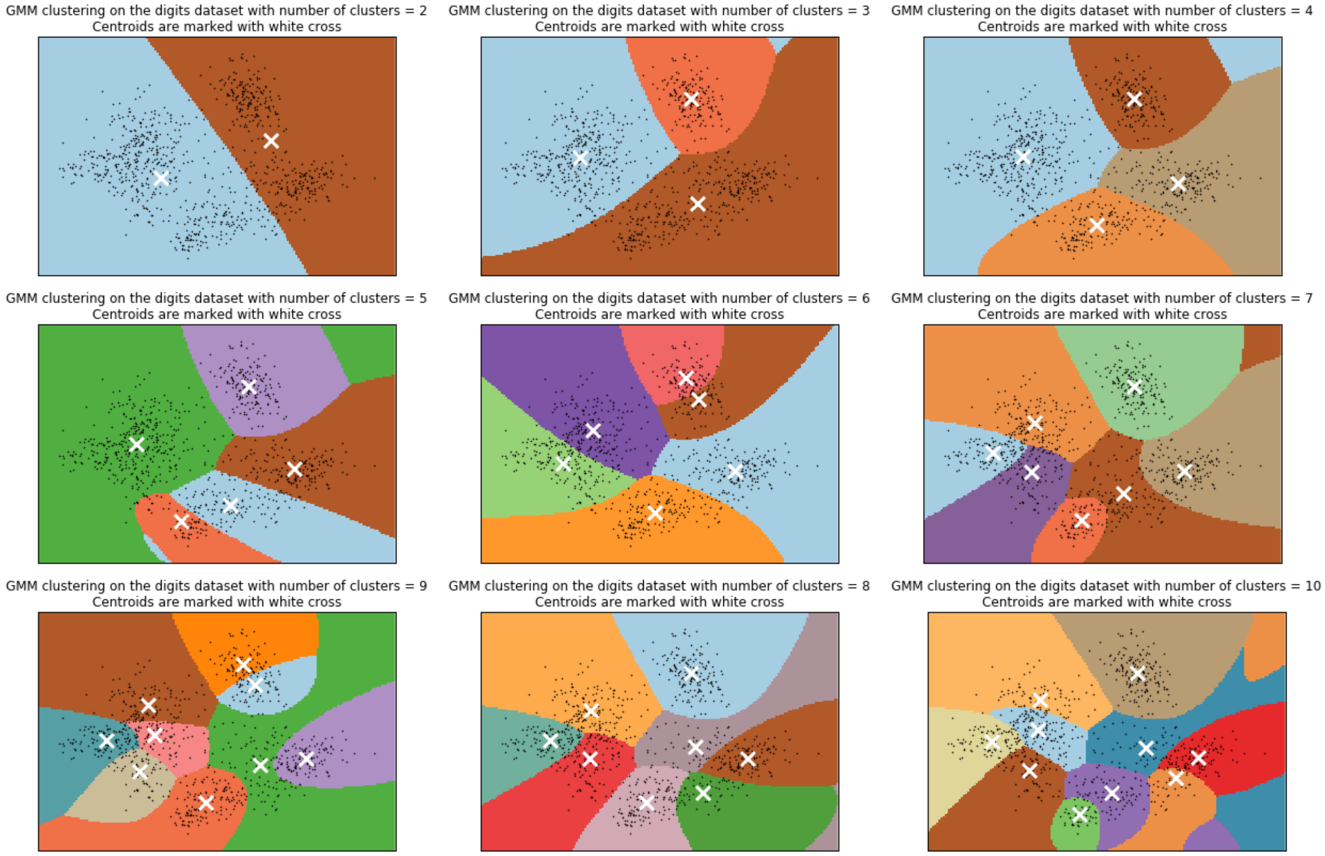
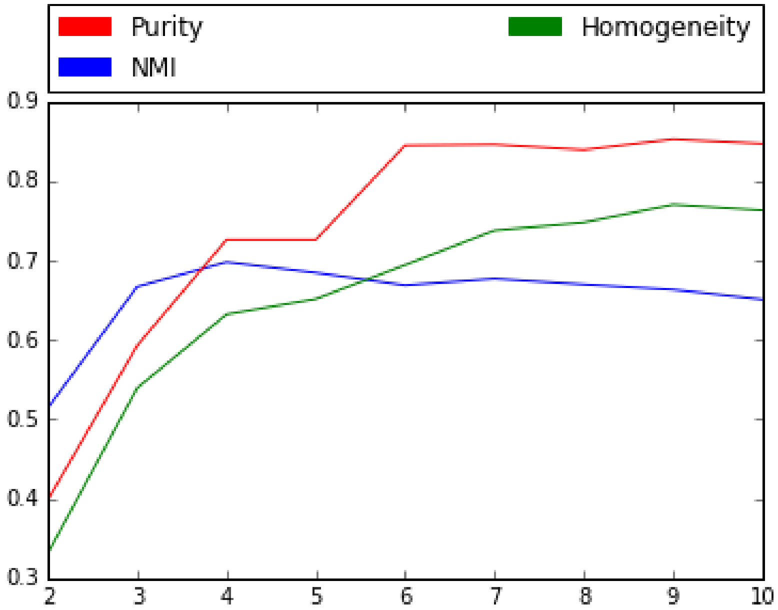


Figure 3

Moreover, was computed the cluster Purity, the Normalized Mutual Information and the Homogeneity.

Normalized Mutual Information (NMI) is an normalization of the Mutual Information (MI) score to scale the results between 0 (no mutual information) and 1 (perfect correlation). Homogeneity, each cluster contains only members of a single class The Normalized Mutual Information score and the Homogeneity score was obtained by `normalized_mutual_info_score`() and `homogeneity_score`() functions. Instead the Purity score was obtained by programming a new function. The Purity is a simple and transparent evaluation measure: each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned images and dividing by the total number of classified images. The Figure 4 is the plot of number of cluster against the Purity, the Normalized Mutual Information and the Homogeneity, with a number of cluster that varies in the range $\{2, .., 10\}$



The purity of the classification, however, can be further improved by using number of clusters much greater than the number of classes of the data set. Moreover, this property represent a limit for the Normalized Mutual Information, in fact, using number of clusters much greater than the number of classes of the data set, the score decreases. The Homogeneity score, instead, in the same situation levels off.