Homework 5: Clustering
class in "Machine Learning", Fall 2016/17

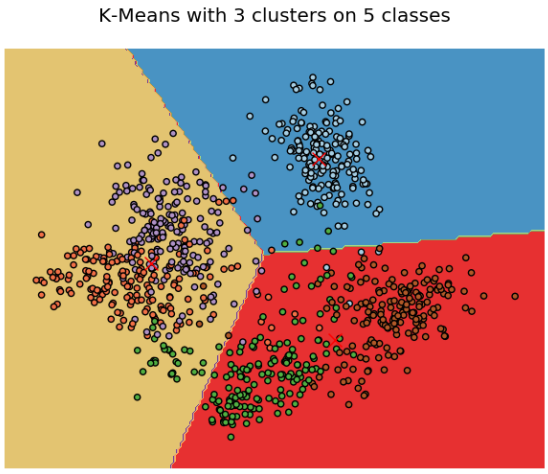
Marco Favorito
Master of Science in Engineering in Computer Science
Department of Computer, Control, and Management Engineering
University of Rome "La Sapienza'
favorito.1609890@studenti.uniroma1.it

28 November 2016

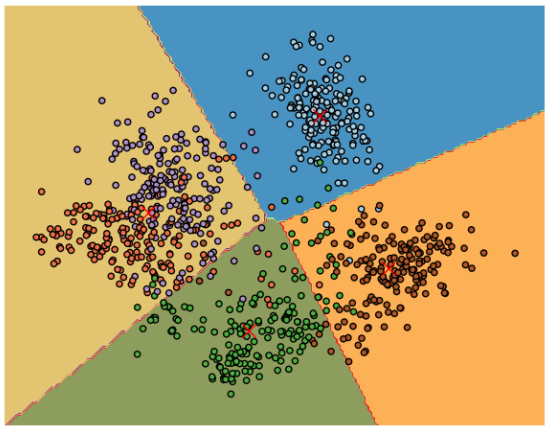
Contents

1 K-Means

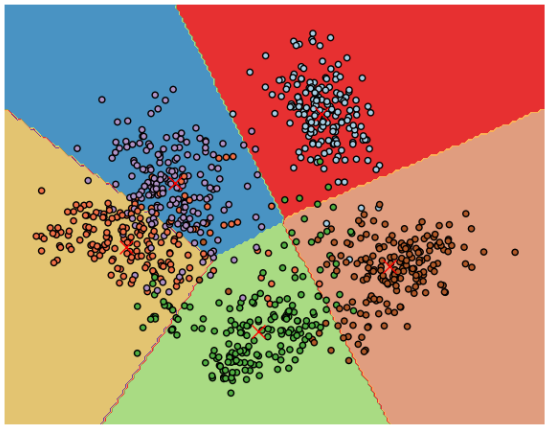
After loaded Digit dataset, I filter it for only 5 classes (i.e. from digit 0 to digit 4). Then, I performed K-Means iteratively, from $k = 3$ to $k = 10$. In the following you can see the plots:



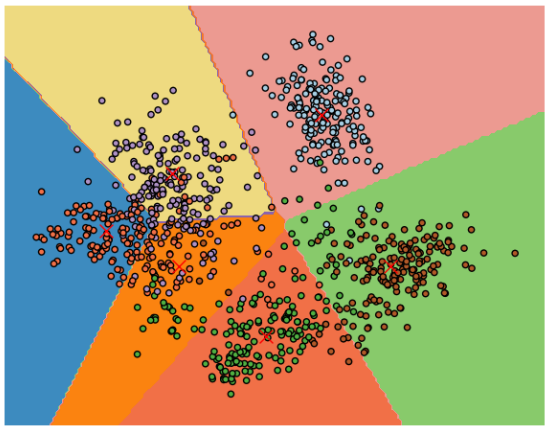
K-Means with 4 clusters on 5 classes



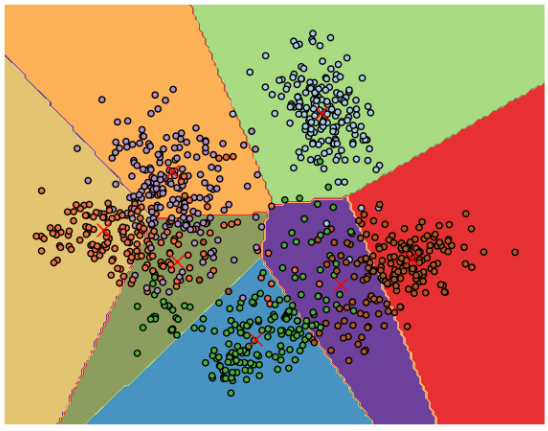
K-Means with 5 clusters on 5 classes



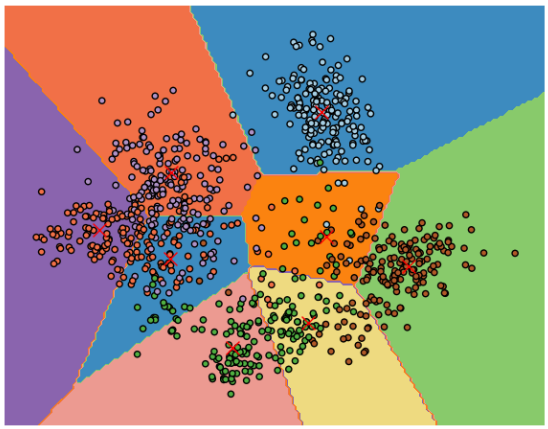
K-Means with 6 clusters on 5 classes



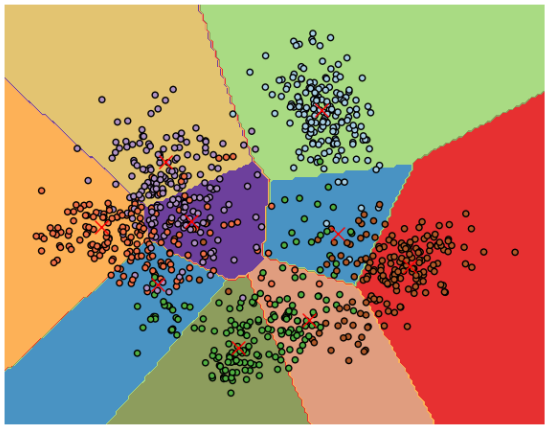
K-Means with 7 clusters on 5 classes



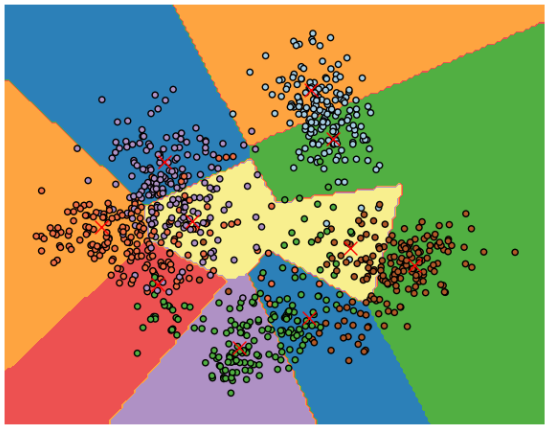
K-Means with 8 clusters on 5 classes



K-Means with 9 clusters on 5 classes

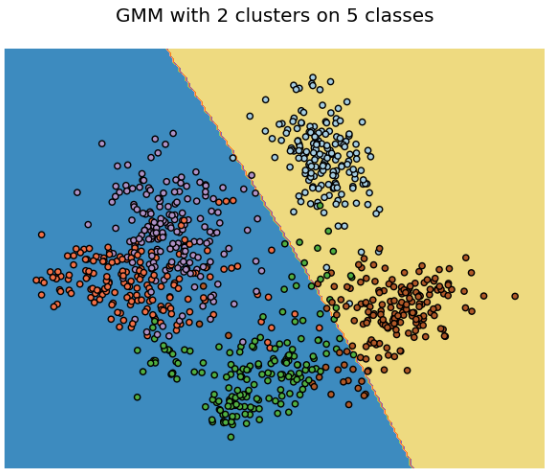


K-Means with 10 clusters on 5 classes

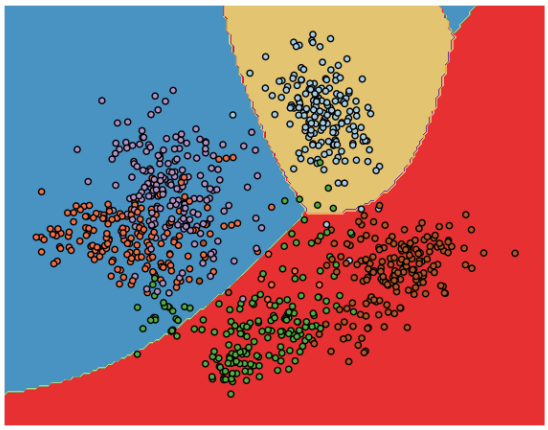


2 GMM

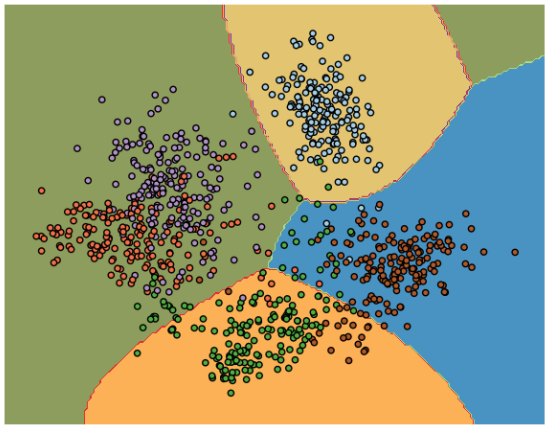
Now, instead, I show th same operations but using a Gaussian Mixture Model and choosing k from 2 to 10:



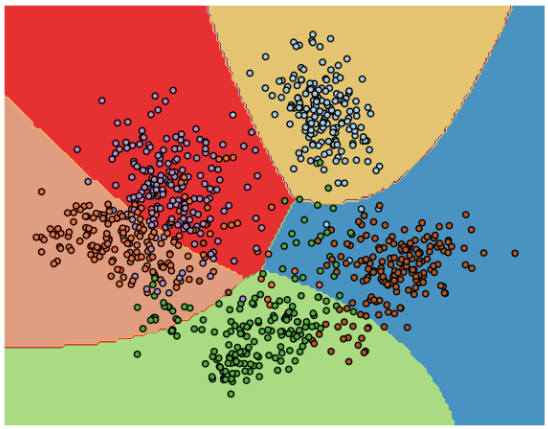
GMM with 3 clusters on 5 classes



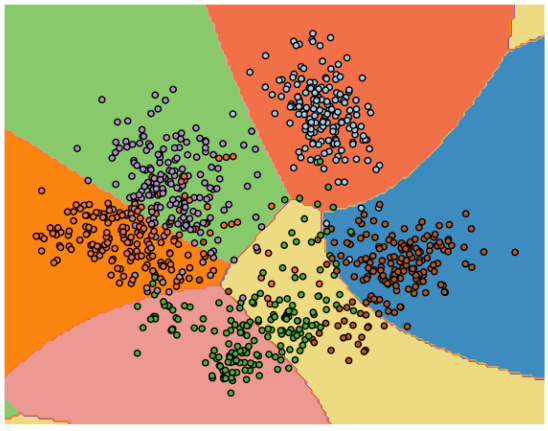
GMM with 4 clusters on 5 classes



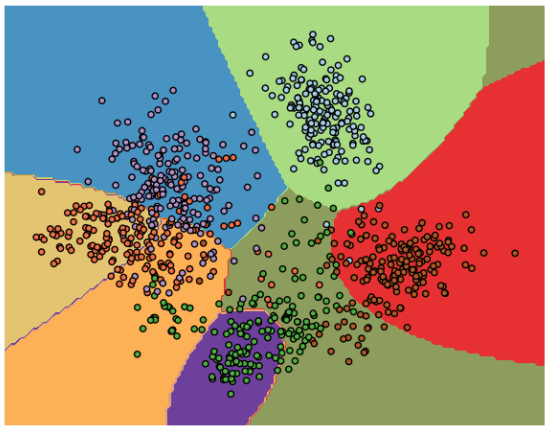
GMM with 5 clusters on 5 classes



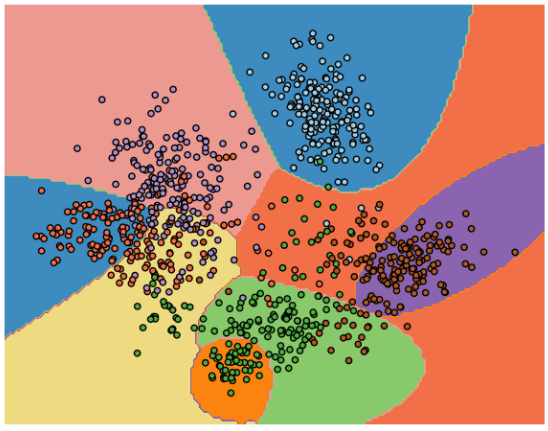
GMM with 6 clusters on 5 classes



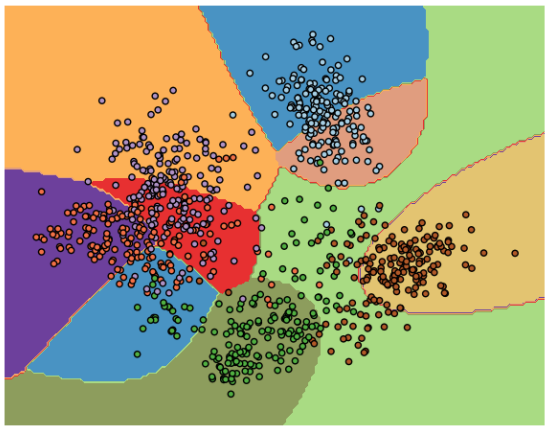
GMM with 7 clusters on 5 classes



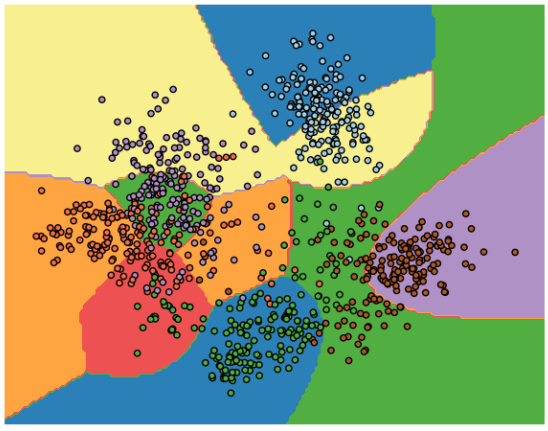
GMM with 8 clusters on 5 classes



GMM with 9 clusters on 5 classes



GMM with 10 clusters on 5 classes



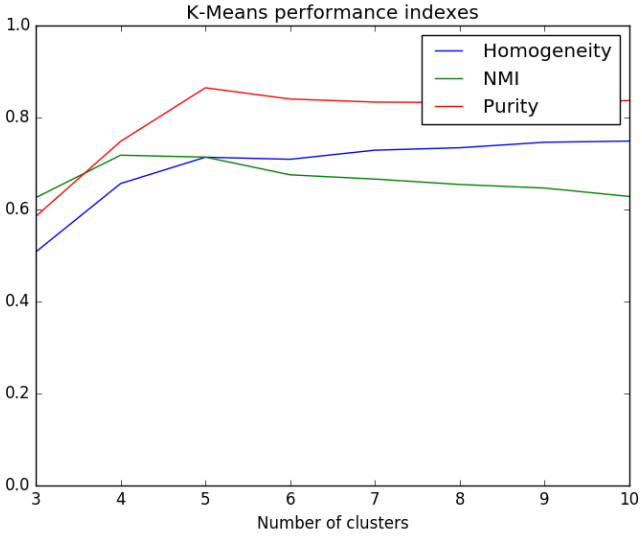
3 Performance evaluation

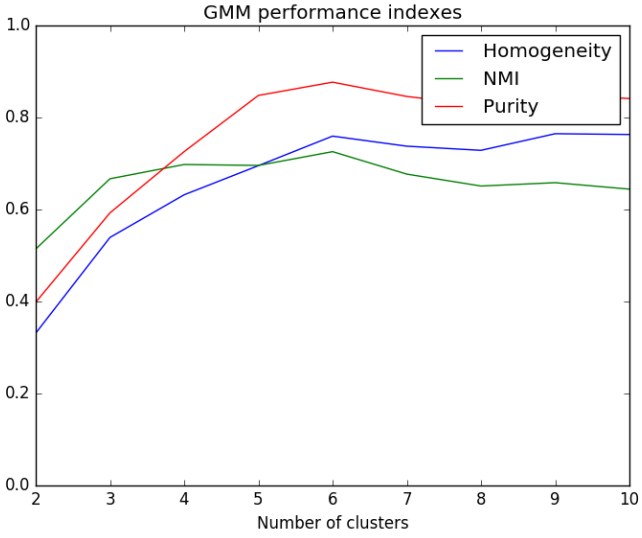
The performance evaluation strategies of a cluster classifier are many; as the homework requires, I evaluated three parameters:

1. **Homogeneity**: each cluster contains only members of a single class;
2. **Normalized Mutual Information (NMI)**: is an *internal evaluation*, in the sense that it is used for trade off the quality of the clustering against the number of clusters;
3. **Purity**: is an *external evaluation* in the sense that takes into account only if members of the cluster belong to the same class.

For the last one I implemented a function that, given y_{true} i.e. true label of the dataset, and $y_{predicted}$ i.e. predicted label by the model, return the purity. In short, I build a confusion matrix ($\#Clusters \times \#Classes$) and then I get the max of each row and sum all the maximum, then I normalized to the number of samples. Once found these values, I plotted them. In the following, it will be shown these values in function of the number of clusters used, both on K-Means and GMM. From the plots we can see that:

- as number of clusters increases, homogeneity increases, because it is more probable that for each cluster we have member in that cluster that belongs to the same class.
- as number of clusters increases, purity first increases, then remains stable, since probably in dataset there are some noisy data that prevent the algorithm to find a correct cluster configuration to maximize purity;
- as number of clusters increases, NMI decreases, since we cannot establish wheter a member of a certain cluster belongs to a class. It seems like that when we have several clusters, the meaning of cluster itself loses meaning.


K-Means performance indexes
Homogeneity
NMI
Purity
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
9
10
Number of clusters


GMM performance indexes
Homogeneity
NMI
Purity
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
7
8
9
10
Number of clusters