Homework 5 Machine Learning

Marco Treglia

1 Clustering

Clustering is a technique for finding similarity groups in a data, called clusters. It attempts to group individuals in a population together by similarity, but not driven by a specific purpose. Clustering is often called an unsupervised learning, as you dont have prescribed labels in the data and no class values denoting a priori grouping of the data instances are given. The goal is to assign a cluster to each data point.

1.1 K-Means

K-means is a clustering method that aims to find the positions $\mu_i \;\; , \; i \; = \; 1 \; ... \; k$ centroid of the clusters that minimize the distance from the data points to the cluster. K-means clustering solves:

$$\arg \; \min_c \sum_{i=1}^{k} \sum_{x \, \in \, c_i} d\,(x, \; \mu) \; = \arg \; \min_c \sum_{i=1}^{k} \sum_{x \, \in \, c_i} ||x \; - \; \mu||^2$$

with c_i et of points that belong to cluster i

Algorithm

1) Initialize the center of the clusters :

$$\mu_i = somevalue \;\; , \; i \; = \; 1 \; ... \; k$$

2) Attribute the closest cluster to each data point :

$$c_i \; = \; \{ j \; : \; d\,(x_j \; , \; \mu_i) \; \leq \; d\,(x_j \; , \; \mu_k) \;\; , \; k \; \neq \; i \; , \; j \; = \; 1, \; . \; . \; ., \; n \}$$

3) Set the position of each cluster to the mean of all data points belonging to that cluster:

$$\mu_i \;\; = \;\; \frac{1}{|c_i|} \sum_{j \, \in \, c_i} x_j \; , \; \forall i$$

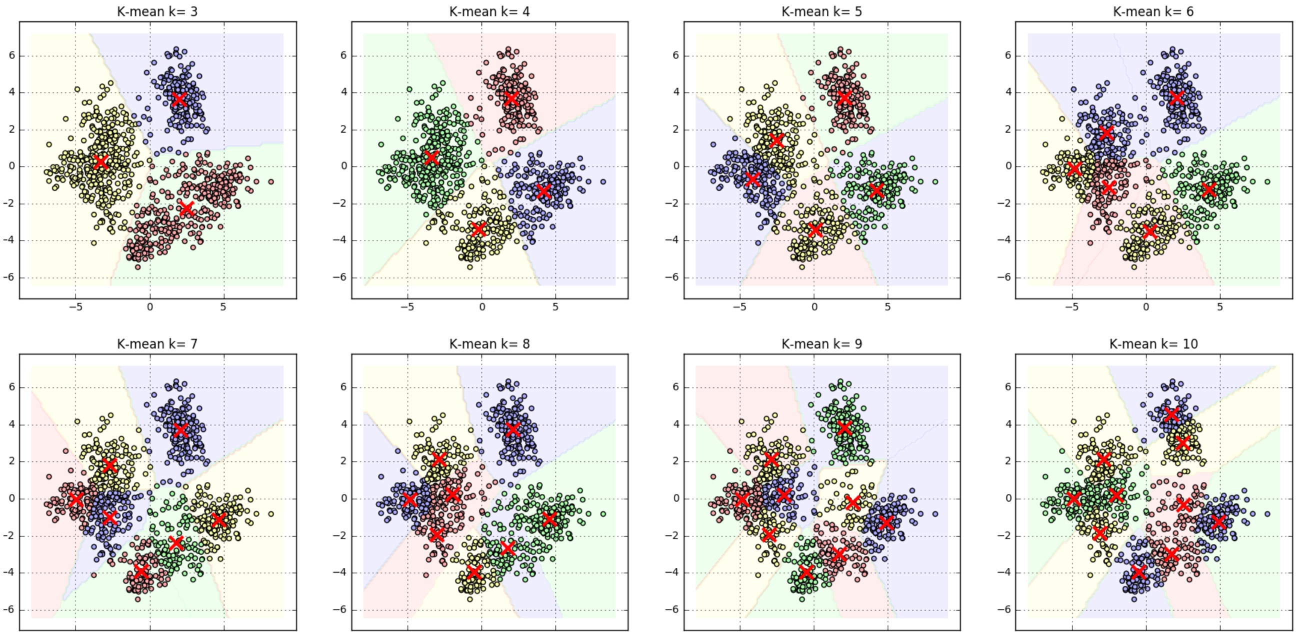
4) Repeat steps 2-3 until convergence.

1.2 Gaussian Mixture Model (GMM)

GMM A Gaussian Mixture Model (GMM) is an extension of the K-means model , wich cluster are modolled with gaussian distribuition. So we have not ony their mean but also the covariance that describe their ellipsoidal shape , then we can fit the model by maximizing the likelihood of the observed data. This estimation is given by using the iterative Expectation-Maximization (EM), that will assign data to each cluster with some soft probability.

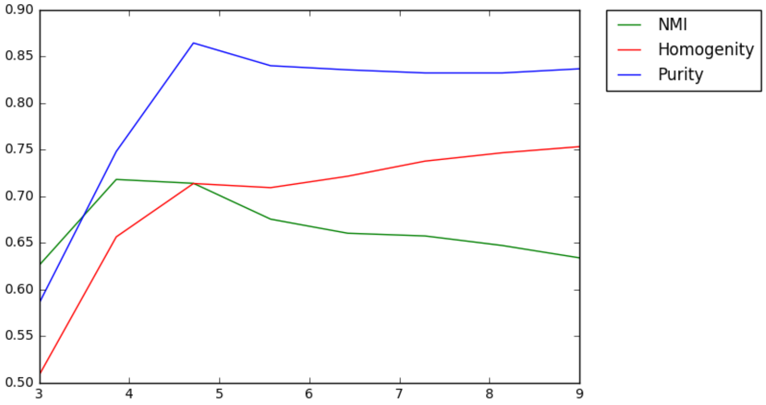
1.3 Visualization Clustering

K-Means

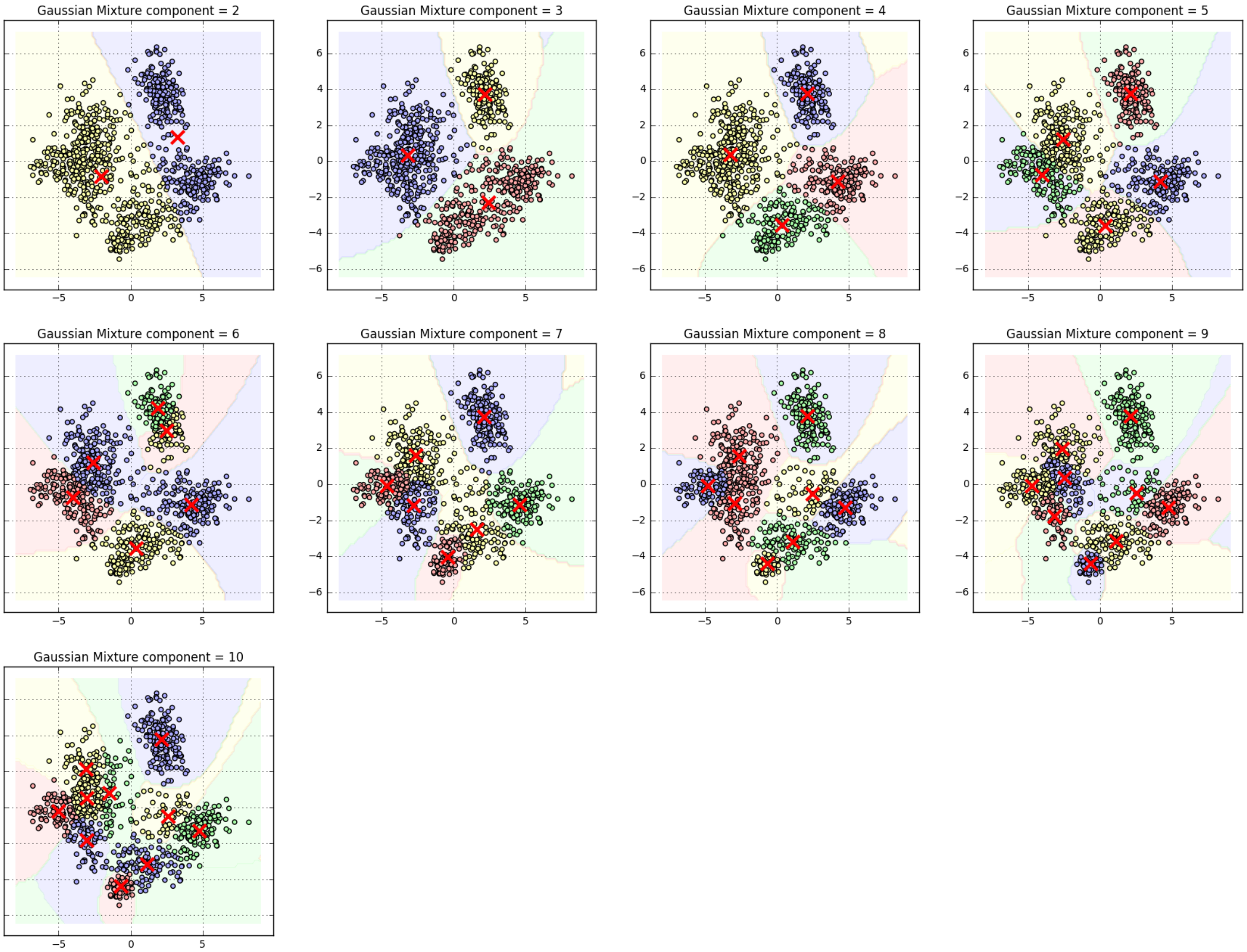


Here the result of clissified our data with K-means from 3 to 10 cluster.

K	Normalized mutual	Homogeneity	Purity
3	0.625751	0.50771	0.584906
4	0.718065	0.656366	0.748058
5	0.712958	0.712744	0.863485
6	0.679237	0.713051	0.842397
7	0.659002	0.720104	0.834628
8	0.658886	0.740584	0.843507
9	0.649282	0.748835	0.833518
10	0.635634	0.755502	0.835738

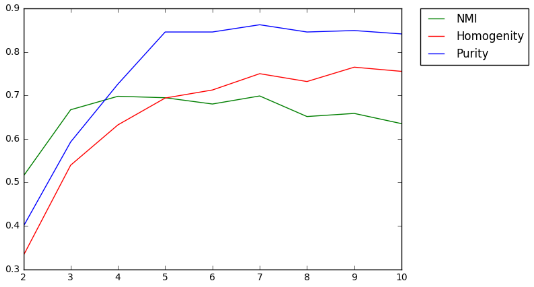


GMM



Here the result of clissified our data with GMM from 2 to 10 cluster.

K	Normalized mutual	Homogeneity	Purity
2	0.499168	0.3	
3	0.669425	0.5	
4	0.696318	0.6	
5	0.695807	0.6	
6	0.694674	0.7	
7	0.692433	0.7	
8	0.651789	0.7	
9	0.674586	0.7	
10	0.640276	0.751036	0.804661



To compute purity , 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 documents and dividing by N. This can gives a measure of the quality of the cluster.