Introduction to Machine Learning
Homework 10: Clustering: K-means and EM-GMM algorithm

Prof. Yao Wang

1. Figure 1 shows a set of samples to be clustered. Show the results from K-means algorithm in successive iterations, starting with the initial centroids indicated in the figure. You can do nearest neighbor partition and centroid update approximately by "eyeballing".

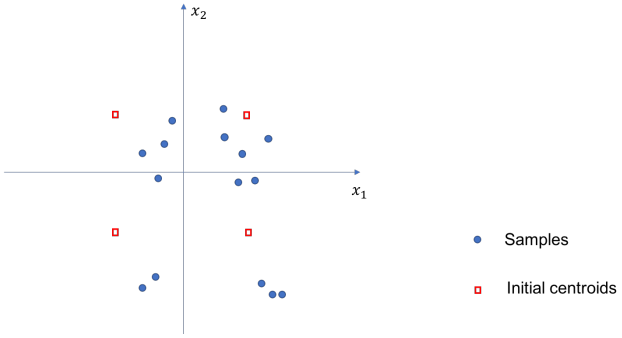

Figure 1: The set of samples to be clustered

2. Suppose you have conducted a clustering analysis for a dataset with each sample described by D features, and you used K-means algorithm to derive K clusters and determined the cluster model parameters (including the centroids of the K clusters). Given a test dataset containing N samples, you want to classify each sample into one of the cluster using the nearest neighbor rule. How many computations are needed? For simplicity, for this and all following problems, only count multiplications (consider the square operation as multiplication).

3. Suppose you are given N samples each described by D features, and you are asked to cluster them into K clusters using the K-means algorithm. Suppose you run the K-means iteration T times. How many computations are needed?

4. (Optional) Suppose you have conducted a clustering analysis for a dataset with each sample described by D features, and you used EM-GMM algorithm to derive K clusters and determined the cluster model parameters (including the prior probabilities, centroids and covariance matrices of the K clusters). Given a test dataset containing N samples, you want to classify each sample into the cluster that has the highest posterior probability. How many computations are needed?

5. (Optional) Suppose you are given N samples each described by D features, and you are asked to cluster them into K clusters using the EM-GMM algorithm. Suppose you run the EM iteration T times. How many computations are needed?