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

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".

   Solution: See the figure below. After step 3a, the partition is the same as in step 2a. So the algorithm converged.

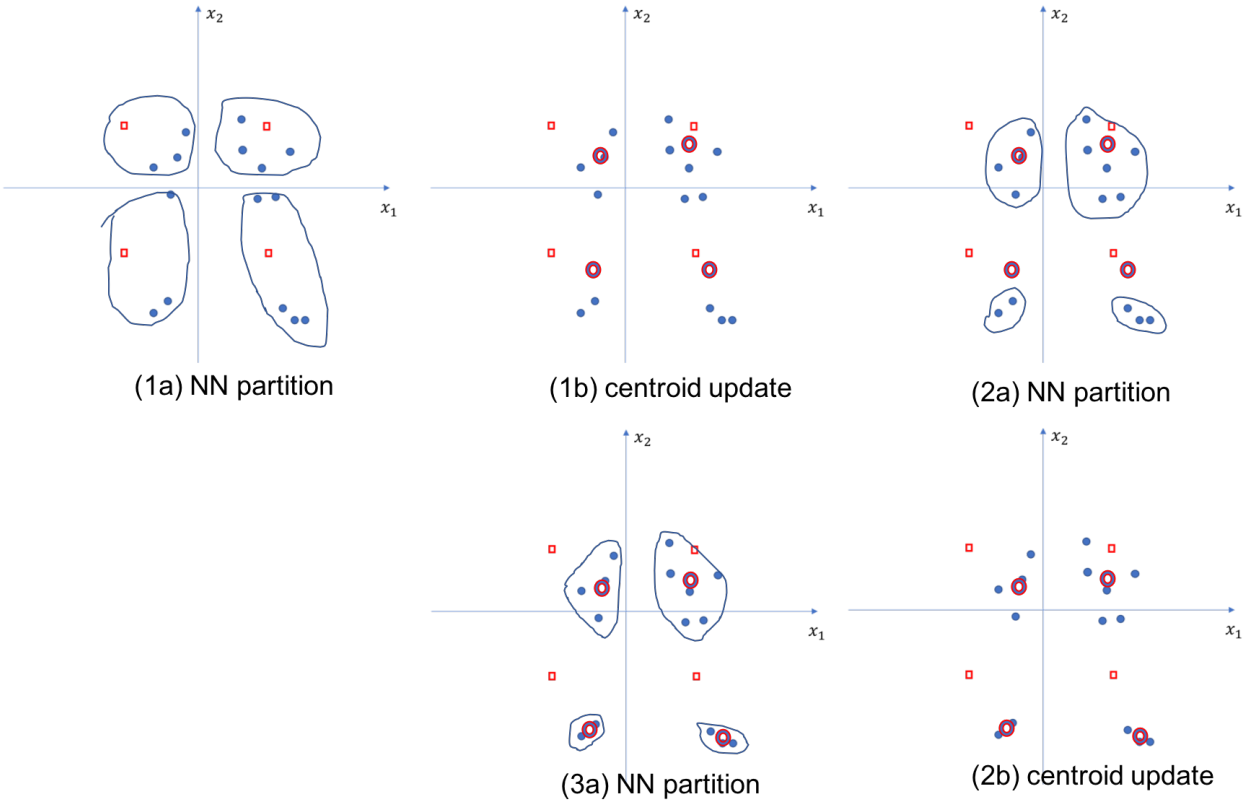

Figure 1: Illustration of successive K-means iterations.

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).

Solution: For each sample x, we need to compare it with each of the K cluster centroids μ_k by computing the distance square $d(x, \mu_k)^2 = \|x - \mu_k\|^2 = \sum_{d=1}^{D}(x_d - \mu_{k,d})^2$. We need to compute the distance with all K centroids and finding the one with the minimal distance. Each distance calculation needs D multiplications and $D-1$ additions. Comparing each pixel with K centroids requires KD multiplications and $K(D-1)$ additions. Repeating this for N pixels requires NKD multiplications and $NK(D-1)$ additions. Overall, we can say the complexity is $O(NKD)$.

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?

   Solution: In each iteration, we first need to do a nearest neighbor search for each sample. This requires the same amount of computation as in the previous problem: NKD multiplications and $NK(D-1)$ additions. Then we need to recompute the centroid of each cluster. If a cluster has N_k samples assigned to it, computing the mean requires $D(N_k - 1)$ additions and D divisions. Computing the centroids for all K clusters requires $D(N-K)$ additions and DK divisions. This is negligible compared to the first step for nearest neighbor search. Therefore, the total computation in each iteration is $O(NKD)$. T iterations would take $O(TNKD)$.

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?

   Solution: For each sample x_n, we need to compute the posterior probability $\gamma_{n,i}$ that it belongs to cluster i for all K clusters, and find the cluster that has the highest probability. Recall that this probability can be expressed as

$$\gamma_{n,i} = \frac{q_i \mathcal{N}(x_n|\mu_i, P_i)}{\sum_{k=1}^{K} q_k \mathcal{N}(x_n|\mu_k, P_k)}$$

   Therefore, we need to compute $q_k \mathcal{N}(x_n|\mu_k, P_k), \forall k \in 1, 2, \ldots, K$. Recall that

$$\mathcal{N}(x_n|\mu_k, P_k) = \frac{1}{(2\pi)^{D/2}|P_k|^{1/2}} \exp\left\{-\frac{1}{2}(x_n - \mu_k)^T P_k^{-1}(x_n - \mu_k)\right\}.$$

   Given the GMM model, q_k, μ_k, P_k are constants, so we mainly need to compute the distance in the exponent $(x_n - \mu_k)^T P_k^{-1}(x_n - \mu_k)$. If the covariance matrix P_k and hence P_k^{-1} is a full matrix (of dimension $D \times D$), we first need to compute $z = P_k^{-1}(x_n - \mu_k)$. This requires D^2 multiplications (we will ignore the additions). Then we need to compute $(x_n - \mu_k)^T z$, which requires D multiplications. Ignore all other operations, computing $q_k \mathcal{N}(x_n|\mu_k, P_k)$ for each k

requires about $D^2 + D$ multiplications. Computing this for all K clusters and consequently $\gamma_{n,i}$ for all $i \in 1, 2, \ldots, K$ requires $K(D^2 + D)$ multiplications. Repeating this all for samples will require $NK(D^2 + D)$ multiplications. Overall, we can say the complexity is $O(NKD^2)$. Compared to k-means (Prob. 2), it is at least D times more computations.

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?

   Solution: In each iteration, we first need to compute the posterior probability $\gamma_{n,i}$ based on the model parameters from the previous iteration. This will require $O(NKD^2)$ computations, as shown in the previous problem. Then we need to update the q_k, μ_k, and P_k. Computing q_k requires only addition. Computing $\mu_k = \sum_n \gamma_{n,k} x_n$ requires N multiplications. Computing $P_k = \sum_n \gamma_{n,k}(x_n - \mu_k)(x_n - \mu_k)^T$ requires $N(D^2 + 1)$ multiplications. Computing these parameters for all K clusters take on the order of $KN(D^2 + 2)$ operations. So each iteration takes $O(NKD^2)$ operations. T iterations will thus take $O(TNKD^2)$ operations. Again, this is about D times more than the k-means algorithms (Prob. 3).