DS5230 : Unsupervised Data Mining : HWK2

Name: Jenish Sunil Kothari

1. (15 points) PROBLEM 1: KMeans Theory

Given the KMeans objective discussed in class with Euclidean distance:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} \pi_{ik} ||x_i - \mu_k||^2$$

where:

- x_i is the i-th data point,
- μ_k is the centroid of cluster k,
- π_{ik} is the membership indicator for data point i to cluster k,
- N is the total number of data points,
- K is the number of clusters.

A) Prove that the E-step update on membership π achieves the minimum objective given the current centroids μ.

B) Prove that the M-step update on centroids μ achieves the minimum objective given the current memberships π.

C) Explain why KMeans has to stop (converge), but not necessarily to the global minimum objective value.

Solution:

Part A : E - Step

- In the E-step we are already sure about the centroids (μ_k) for each of the K clusters. We need to identify the memberships for each of the data points π_{ik}. We can achieve this by calculating the distance of the points with each of the clusters and then assigning the point to the closest cluster.

$$\pi_{ik} = \begin{cases} 1 & \text{if } k = \arg\min_k ||x_i - \mu_k||^2 \\ 0 & \text{otherwise} \end{cases}$$

- From the above equation, we are pretty sure that the E-step update achieves the minimum objective given the fixed centroids since it takes into account the minimal distance from each of the clusters.

Part B : M - Step

- In the M-step we are already sure about the memberships π_{ik} for each of the datapoints. We need to identify the centroids (μ_k) . We know that average of the points is the best centroid. We can prove this by differentiating the objective function. Lets only consider 1 cluster as of now, we get,

$$J_k(\mu_k) = \sum_{i=1}^{N} \pi_{ik} \|x_i - \mu_k\|^2$$

$$\frac{d}{d\mu_k} \|x_i - \mu_k\|^2 = -2(x_i - \mu_k)$$

$$\frac{dJ_k(\mu_k)}{d\mu_k} = -2 \sum_{i=1}^{N} \pi_{ik}(x_i - \mu_k)$$

To minimize the objective function, we set the derivative equal to zero:

$$-2 \sum_{i=1}^{N} \pi_{ik}(x_i - \mu_k) = 0$$

Thus, the update rule for the centroid μ_k is:

$$\mu_k = \frac{1}{\sum_{i=1}^{N} \pi_{ik}} \sum_{i=1}^{N} \pi_{ik} x_i$$

- From the above equation, we know that the centroid is the weighted average of the points assigned to that cluster. Hence we can say that M step achieves the minimum objective given the current memberships.

Part C : Convergence

- From the above steps it is pretty clear that, the objective function J decreases or remains the same after each E and M steps.
- The KMeans algorithm, has finite assignments of points to clusters i.e. let's say we have 4 points and 2 clusters, there 16 ways for assignment. Similar thing goes for given datasets, since the assignment states are finite, KMeans has to stop.
- However the point would be the local optimum since different randomized initializations of the centroids or memberships can lead to more local optimums better than the current optimum. Hence, KMeans has to stop (converge), but not necessarily to the global minimum objective value.