Naman Jain Vimal Kumar
1001670153

Homework 9

Explain why the K-means objective function decreases in each of the two steps in K-mean algorithm:
(a) re-assign every data points to their nearest cluster centroids.
(b) Given the grouping (or clustering), re-computer the cluster centroids.

Solution

About K-means:

- K-means algorithm is a clustering algorithm.
- It partitions 'n' datapoints to 'k' clusters.
- It is a KNN algorithm if 'k' is equal to 'n'.
- The objective function of K-means is to minimize the within- cluster sum of squares distance.

$$Z = \arg_{\min=} \sum_{i=1}^{k} \sum_{x}^{N_k} \|x - \mu_i\|^2$$

$$where\ x - Data\ point\ for\ ith\ cluster$$
$$and\ \mu_i - Cluster\ Point$$

Algorithm:

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j \|x^{(i)} - \mu_j\|^2.$$

   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

Re-Assign every data point to their nearest cluster centroids:

- We know that, we need to reduce the distance to reduce the 'Z' as it is directly proportional.
$$Z \propto \|x - \mu_i\|^2$$
- Thus, when we assign data points to its closest centroid, this reduces the 'Z' and helps to achieve a better clustering.
- Say following data with centroids [1 , 1] and [-1,-1]

Data Point	Distance with [1,1]	Distance with [0,0]	Nearest Distance
[-2, -2]	4.24	1.41	1.41
[-1, -1]	2.82	0	0

Naman Jain Vimal Kumar
1001670153

[0, 0]	1.41	1.41	1.41
[1, 1]	0	2.82	0
[2, 2]	1.41	4.24	1.41
Z	8.47	8.47	**4.23**

- The above example shows that we get the lowest 'Z' with nearest centroid.

Given the grouping (or clustering), re-computer the cluster centroids:

- The grouping is done based on a given cluster centroid.
- But then, its 'Z' could generate a larger output when compared to a centroid which is at the center of the grouped points.
- For example,

Data Point	**Centroid [1,1]**	**Centroid [-1,-1]**	**Mean Centroid [0, 0]**
[-2, -2]	4.24	1.41	2.82
[-1, -1]	2.82	0	1.41
[0, 0]	1.41	1.41	0
[1, 1]	0	2.82	1.41
[2, 2]	1.41	4.24	2.82
Mean Centroid = [0, 0]	8.47	8.47	**8.46**

- Cluster Centroid can be computed to get the mean of the grouped points.
- Thus mean helps us to minimize 'Z' cost generating better output.

References,

- www.wikipedia.com