k Means Algorithm

Aaryan
CO21BTECH11001

k-Means is an unsupervised learning algorithm which is used to group the data into few(k) clusters.

Training set - $\{X^{(1)}, X^{(2)}, \dots, X^{(m)}\}$

The k-means clustering algorithm is as follows:

1. Initialize cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \ \in \ R^n$ randomly. Generally, we choose k random training examples as cluster centroids.
   Initialize $conv = \infty, j_0 = 0$
   Tolerance $\epsilon = 1.0e - 10$
2. We define a cost function as follows -
   $$J(c, \mu) = \sum_{i=1}^{m} \left|\left|X^{(i)} - \mu_{c^{(i)}}\right|\right|^2$$
   where $\mu_{c^{(i)}}$ is the cluster centroid assigned to $X^{(i)}$
3. while $|conv| \geq \epsilon$ {
   a. To each $X^{(i)}$, assign the cluster centroid nearest to it -
   $$c^{(i)} = argmin_j \left|\left|X^{(i)} - \mu_j\right|\right|^2$$
   b. $j_1 = J(c, \mu)$
   $conv = j_1 - j_0$
   $j_0 = j_1$
   c. To each μ_j, assign the average of points assigned to j^{th} cluster
   $$\mu_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)} = \mu_j\} X^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = \mu_j\}}$$

   }

The visualization for what k-Means algorithm do is here .

Questions –

1. What type of algorithm is k-Means algorithm?
   Ans. It is an unsupervised learning algorithm.

2. Where is k-Means algorithm generally used?
   Ans. It is used for clustering of dataset in fields of market clustering, campaigning etc.

3. How do we choose the value of k in k-Means algorithm?
   Ans. Value of k is generally dependent on need of the problem i.e., the motive of using the algorithm. For example, a company wants to cluster the market in atmost 10 clusters, therefore k=10.

4. If a cluster has no point assigned to it, we can't calculate the mean for that cluster, then what will you do in that situation?
   Ans. In such situation, we generally eliminate that cluster and we now make just (k-1) clusters of dataset.
   Another approach maybe to re-initialize cluster centroid of that cluster, which is less often used.

5. What are the advantages of using k-Means Algorithm?
   Ans. 1. It can easily scale to large datasets.
   2. It guarantees convergence.
   3. It easily adapts to new examples.

6. What are the disadvantages of using k-Means Algorithm?
   Ans. 1. We have to choose k manually.

2. Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored.