**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 -*

The k-means clustering algorithm is as follows:

1. Initialize cluster centroids randomly. Generally, we choose k random training examples as cluster centroids.  
   Initialize   
   Tolerance
2. We define a cost function as follows -  
   where is the cluster centroid assigned to
3. while {
   1. To each , assign the cluster centroid nearest to it -
   3. To each assign the average of points assigned to cluster

}

The visualization for what k-Means algorithm do is [here](https://en.wikipedia.org/wiki/File:K-means_convergence.gif) .

**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.