k-Mean Clustering

Learning Type: Unsupervised, Task: Clustering, Algorithm: k-mean Clustering

Definition:

It is a clustering algorithm where data points are assigned into k groups, based on the closest distance between each group's centroid(mean) to the data point. All data points assigned to the same centroid form a group which is called a cluster. For k groups, we will have k centroids and the corresponding k clusters.

Applications: k-means Clustering is commonly used in pattern recognition, market segmentation, and image compression.

Algorithm of k-mean Clustering:

Step1: Specify k centroids corresponding to k clusters and randomly initialize them.
Step2: Assign each data point to the closest centroid.
Step3: Computer the new centroid by taking the average (mean) of each data point assigned to the same cluster.
Continue until the centroids do not change or the maximum number of iterations is reached.

Step1: k (k=2) centroid random initialization
Step2: Centroid assignment
Step3: Centroid update
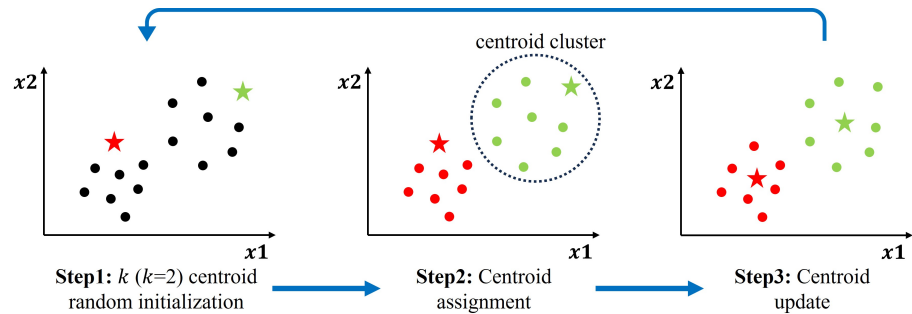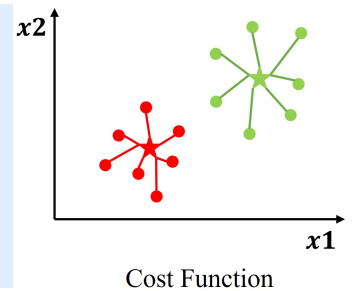
Evaluation Criteria 1: Cost Function

It is the sum of the distance of each data point from the centroid in the assigned cluster.
$J = \sum_{j=1}^{k} \sum_{i=1}^{m} a_{ij} \|x_i - \mu_j\|_2^2$
where
$a_{ij} = 1$ if x_i belongs to j cluster
$a_{ij} = 0$ if x_i does not belong to j cluster

Cost Function

Evaluation Criteria 2: Silhouette Coefficient

It is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation). The silhouette coefficient is between [-1, 1]. 1 means distinct clusters, 0 denotes overlapping and -1 means worst case.

$$S(i) = (b(i) - a(i))/max(a(i), b(i))$$
and for each cluster
$$AverageSilhouette = mean(S(i))$$

$S(i)$ is a silhouette score for datapoint i, $a(i)$ is the average distance between i and all data points belong to a cluster of i, $b(i)$ is the average distance between i and all other data points of neighboring clusters.
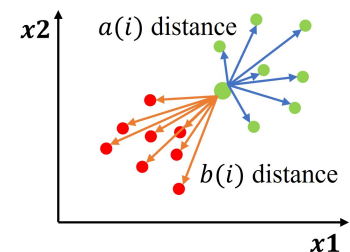

How to select the value of k:

1. Based on the Application: Number of clusters k can be decided based on an application like dress size will be small, medium, and large.

2. Based on Elbow Method: In this method, observe a sudden fall of cost function in the plot of cost function vs k and no significant change is observed afterward, that point is selected for k value and point is called elbow point.

Clothes Size Application

Elbow Method
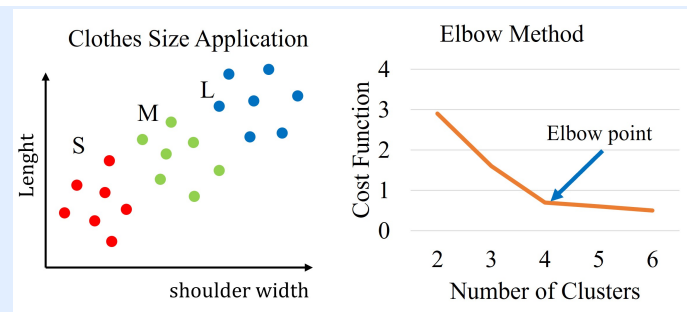
How to initialize the value of centroids:

1. Random data points: Initialize the centroids by choosing random k data points. This method is highly volatile.
2. Least Cost Value Cost function is computed with multiple random initializations. The specific initialization is selected which gives the lowest cost function value. This method is computationally expensive.
3. k-means++: The first centroid is selected as a randomly chosen data point, and the other centroids are placed as far away from each other as possible, based on the squared distance from the existing centroids. k-means++ original paper can be found here.