ASSIGNMENT NO. 4

Title: Write a program to do following:

We have given a collection of 8 points. P1=[0.1,0.6] P2=[0.15,0.71] P3=[0.08,0.9] P4=[0.16,

0.85] P5=[0.2,0.3] P6=[0.25,0.5] P7=[0.24,0.1] P8=[0.3,0.2]. Perform the k-mean clustering

with initial centroids as m1=P1=Cluster#1=C1 and m2=P8=cluster#2=C2.

Answer the following:

a) Which cluster does P6 belong to?

b) What is the population of a cluster around m2?

c) What is the updated value of m1 and m2?

Theory/Methodology:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

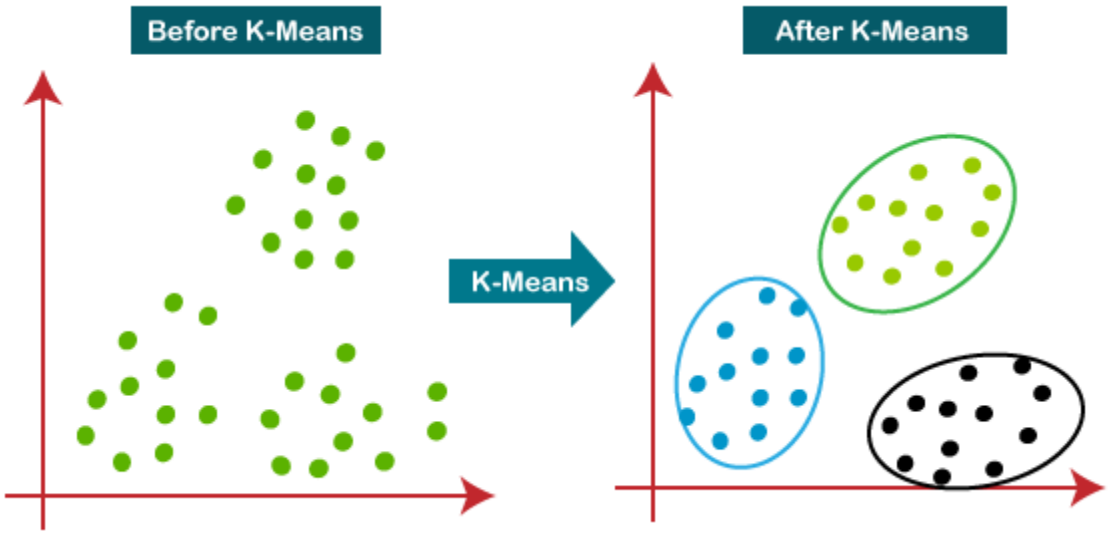
The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



How does the K-Means Algorithm Work?

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

Advantages:

- Simple and easy to implement.
- Efficient for large datasets.
- Works well with data that is well-separated into clusters.

Limitations/Examples:

- Requires the number of clusters (k) to be specified in advance.
- Sensitive to the initial placement of centroids, which can affect the final clustering result.
- May not perform well with clusters of different sizes or densities.

Working steps:

1. Initialize centroids (m1 and m2) based on given points P1 and P8.
2. Assign each point to the nearest centroid (cluster) based on Euclidean distance.
3. Calculate the mean of the points in each cluster to update the centroids.
4. Repeat steps 2 and 3 until convergence (centroids do not change significantly).
5. After convergence, determine:
   - Which cluster P6 belongs to.
   - Population of the cluster around m2.
   - Updated values of m1 and m2.

Conclusion:

K-Means clustering is a powerful algorithm for partitioning data into clusters based on similarity. In this practical, we implemented K-Means clustering on a collection of points and answered specific questions regarding cluster assignments, cluster populations, and centroid updates. This algorithm provides a straightforward approach to cluster analysis, but it's essential to consider its limitations and understand the impact of parameter choices on the clustering results.