Assignment No: - 4

K-means Clustering

Problem Statement:-

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?

Objective:

The objective of this program is to implement the K-Means clustering algorithm on a collection of 8 points.

Specifically, we aim to:

a) Determine which cluster point P6 belongs to.

b) Calculate the population of the cluster around centroid m2.

c) Update the values of centroids m1 and m2 based on the mean of points assigned to each cluster.

S/W Packages and H/W apparatus used: OS: Windows, Kernel: Python 3, Tools: Google Colab

Libraries and Packages Used: Numpy, Matplotlib, Scikit-learn (Clusters)

Theory: -

Methodology:

- **K-Means Clustering:** K-Means Clustering is an unsupervised learning algorithm that is used to solve clustering problems in machine learning or data science

- **What is the K-Means Algorithm?** K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined 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 allows us to cluster the data into different groups and is 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 data points with some commonalities, and it is away from other clusters.

Advantages and Disadvantages & Limitation/Example:

1. Advantages:
   - Simple and Intuitive: K-means clustering is easy to understand and implement.
   - Efficient: It works well for large datasets and can handle high-dimensional data efficiently.
   - Scalability: K-means scales well with increasing dataset sizes.
   - Interpretability: Results are straightforward and easy to interpret.

2. Disadvantages & Limitations/Example:

- Sensitivity to Initial Centroids: Results can vary depending on the initial centroid selection.
- Assumption of Spherical Clusters: K-means assumes that clusters are spherical, which may not always be the case.
- Impact of Outliers: Outliers can significantly affect the cluster centroids and result in suboptimal clustering.
- Determining Number of Clusters: The number of clusters needs to be specified beforehand, which can be subjective and challenging to determine.

Applications with example:

1. Customer Segmentation: In marketing, K-means clustering can be used to segment customers based on their purchasing behavior. For example, a retail company can cluster customers into groups such as high-value customers, frequent buyers, and occasional shoppers.

2. Anomaly Detection: In cybersecurity, K-means clustering can be utilized to detect anomalies or unusual patterns in network traffic. For example, network administrators can cluster network traffic data and identify clusters with significantly different characteristics, indicating potential security threats or anomalies.

3. Document Clustering: In natural language processing, K-means clustering can be employed to cluster similar documents together. For instance, news articles can be clustered into groups based on their topics, allowing users to explore related articles more efficiently.

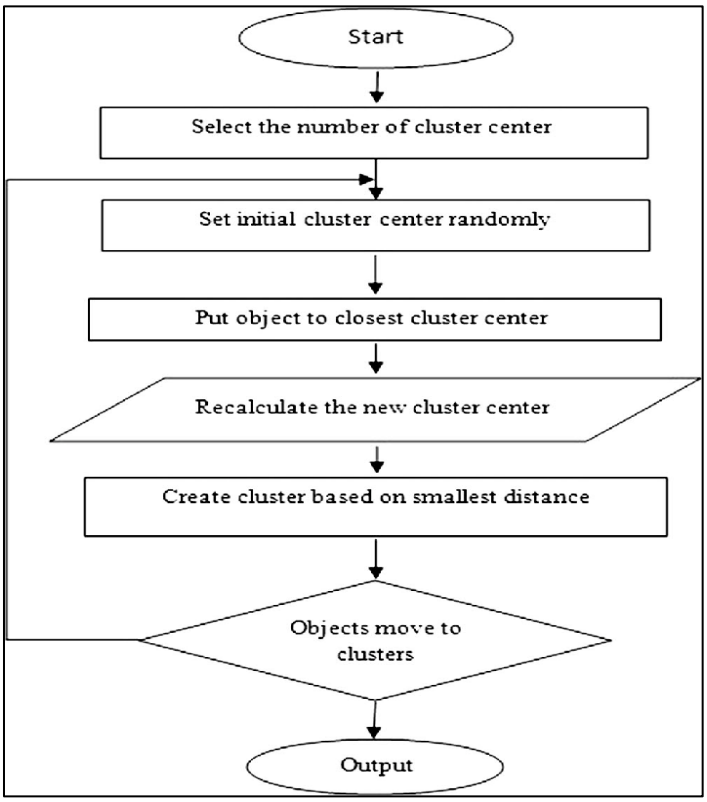
Working / Algorithm:

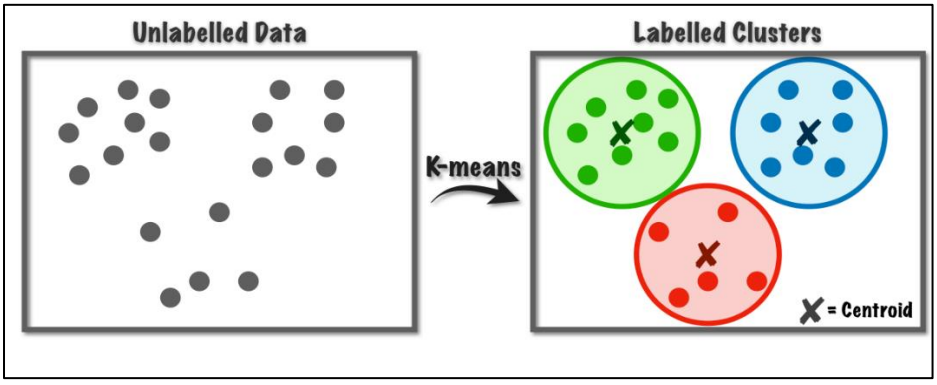
1. Initialization: Start by defining the initial centroids, in this case, m1=P1 and m2=P8.
2. Assign Points to Clusters: Calculate the Euclidean distance between each point and both centroids. Assign each point to the cluster corresponding to the nearest centroid.
3. Update Centroids: Calculate the mean of all points belonging to each cluster. Update the centroids to the new mean values.
4. Repeat: Iterate steps 2 and 3 until the centroids no longer change significantly or until a specified number of iterations is reached.

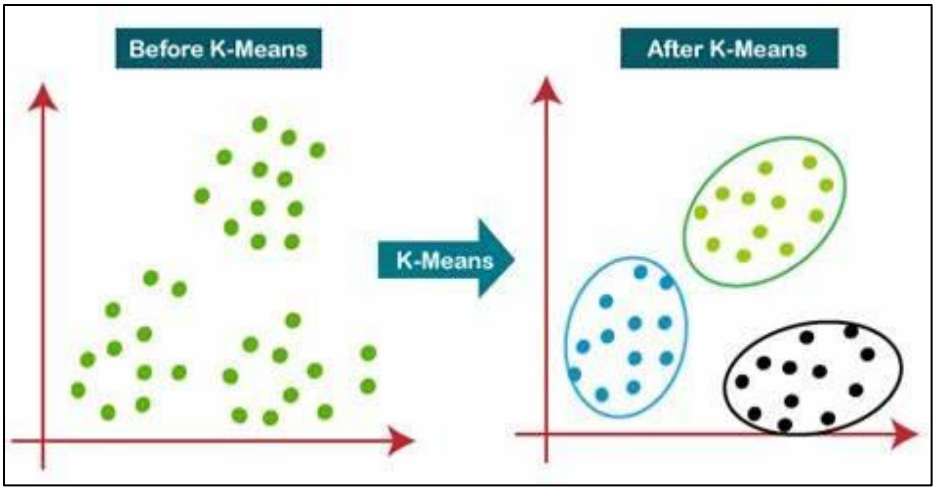
5. Answering Questions:

- To determine which cluster P6 belongs to, calculate its distance from both centroids and assign it to the cluster with the nearest centroid.
- To find the population of the cluster around m2, count the number of points assigned to cluster C2.
- To update the values of m1 and m2, calculate the mean of points in each cluster and set m1 and m2 to these new mean values.

Diagram:







Conclusion

In conclusion, K-means clustering is a simple yet effective method for grouping data points into clusters. While it's straightforward to implement and scales well with large datasets, it requires careful consideration of initial centroids and assumes spherical clusters. Despite its limitations, K-means remains a popular choice for clustering tasks due to its efficiency and interpretability.