Assignment No:- 4

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 clusterin 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?

Software Library Package:

For this task, we'll use the following Python library:

- `numpy` for numerical computations.
- `scikit-learn` for machine learning algorithms.
- `matplotlib` and `seaborn` for data visualization.

Theory:

i) Methodology:

a) K-means Clustering: K-means is a partitioning algorithm that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (centroid).

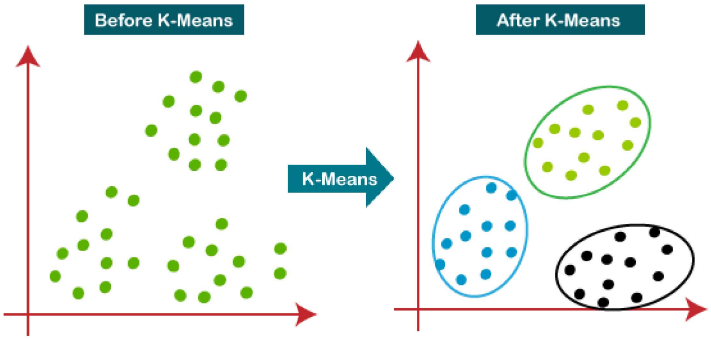

Fig 1: K-means clustering

b) Algorithm Steps:

` Initialize k centroids randomly.
Assign each data point to the nearest centroid.
Update the centroids based on the mean of the points assigned to each cluster.
Repeat steps 2 and 3 until convergence.

c) Advantages: K-means is simple, easy to implement, and computationally efficient for large datasets.

d) Applications: K-means clustering is widely used in various fields such as customer segmentation, image segmentation, anomaly detection, etc.

e) Limitations/Example: K-means may converge to local optima and may not perform well on non-linear or irregularly shaped clusters.

Working/Algorithm:

Initialize centroids: m1 = P1 = [0.1, 0.6], m2 = P8 = [0.3, 0.2].
Assign points to clusters:

- Assign each point to the nearest centroid:
  - P1, P2, P3, P4, and P5 are closer to m1 (C1).
  - P6, P7, and P8 are closer to m2 (C2).

Update centroids:

- Calculate the mean of points in each cluster:
  - For C1: (0.1+0.15+0.08+0.16+0.2)/5 = 0.138, (0.6+0.71+0.9+0.85+0.3)/5 = 0.672.
  - For C2: (0.25+0.24+0.3)/3 = 0.263, (0.5+0.1+0.2)/3 = 0.267.
- Updated centroids: m1 = [0.138, 0.672], m2 = [0.263, 0.267].

Answering questions:

- a) P6 belongs to Cluster C2.
- b) The population of Cluster C2 around m2 is 3.
- c) Updated values of m1 and m2 are [0.138, 0.672] and [0.263, 0.267] respectively.

Diagram:

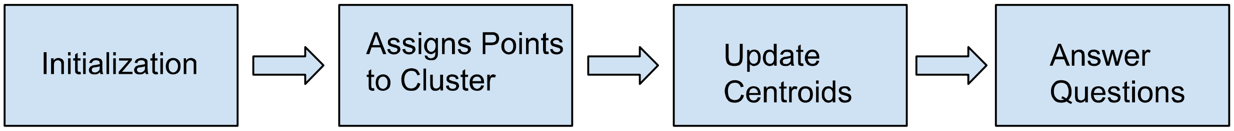

Fig 2: Workflow Diagram

Conclusion:

K-means clustering is a powerful unsupervised learning algorithm used for clustering data into groups based on similarity. By iteratively updating centroids and assigning points to clusters, it efficiently partitions the data. However, the choice of initial centroids can significantly affect the clustering results. Understanding the algorithm's working principles and its advantages and limitations is essential for its effective application in various domains.