Practical No 9(d)

Aim: Data Clustering using K means

Theory:

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into K distinct, non-overlapping clusters. The objective of K-means is to group data points into clusters in such a way that points within the same cluster are similar to each other, while points in different clusters are dissimilar.

1. Initialization:

- The algorithm starts by randomly initializing K cluster centroids. These centroids represent the center of each cluster.

2. Assignment:

- Each data point is assigned to the nearest cluster centroid based on a distance metric, typically Euclidean distance. The data points are grouped into clusters according to which centroid they are closest to.

3. Update:

- After the initial assignment, the centroids of the clusters are recalculated. Each centroid is updated to be the mean of all data points assigned to its cluster. This step ensures that the centroids better represent the center of their respective clusters.

4. Iteration:

- Steps 2 and 3 are repeated iteratively until convergence, which occurs when the centroids no longer change significantly or when a specified number of iterations is reached.

5. Convergence:

- The algorithm converges when the centroids no longer change or the change is within a predefined threshold.

6. Final Result:

- The final result of K-means clustering is a partitioning of the dataset into K clusters, where each data point belongs to the cluster with the nearest centroid.

Key Points:

- K-means is sensitive to the initial placement of cluster centroids. Different initializations can lead to different final cluster assignments.
- It is important to choose an appropriate value of K, the number of clusters, which can be determined using techniques such as the elbow method or silhouette analysis.
- K-means is computationally efficient and scales well to large datasets, making it suitable for a wide range of applications.
- The algorithm can converge to local optima, meaning that the final clustering may not always be the globally optimal solution.

- K-means assumes that clusters are spherical and have similar sizes, which may not always hold true for real-world datasets. As a result, the algorithm may not perform well on datasets with complex cluster shapes or varying cluster sizes.

Overall, K-means clustering is a simple yet powerful algorithm for data clustering, widely used in various domains such as image segmentation, customer segmentation, and document clustering. Its simplicity and efficiency make it a popular choice for exploratory data analysis and clustering tasks.

Program code:

```
# Step 1: Import necessary libraries
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Step 2: Generate random data for demonstration
np.random.seed(42)
data = np.random.rand(100, 2)

# Step 3: Number of clusters (you can adjust this)
num_clusters = 3

# Step 4: Create KMeans instance
kmeans = KMeans(n_clusters=num_clusters, random_state=42)

# Step 5: Fit the data to the KMeans model
kmeans.fit(data)

# Step 6: Get the cluster assignments and centroids
labels = kmeans.labels_
centroids = kmeans.cluster_centers_

# Step 7: Visualize the clusters
plt.scatter(data[:, 0], data[:, 1], c=labels, cmap='viridis', edgecolor='k')
plt.scatter(centroids[:, 0], centroids[:, 1], marker='X', s=200, color='red')
plt.title('K-means Clustering')
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.show()
```

Result:

In conclusion, the provided Python code demonstrates the implementation of K-means clustering using the scikit-learn library for a synthetic dataset. Overall, this program demonstrates a simple yet powerful technique for unsupervised learning—K-means clustering—which is widely used for various applications such as customer segmentation, anomaly detection, and image compression. The visualization provides insight into how the

algorithm partitions the data into distinct clusters based on their similarities, showcasing its effectiveness in exploratory data analysis and pattern recognition tasks.