Experiment 4

Objective:

To implement k means clustering.

Theory

K-means clustering is a popular unsupervised machine learning algorithm used to identify and group similar data points into clusters. The algorithm is commonly applied in exploratory data analysis, customer segmentation, pattern recognition, and image compression. Unlike supervised learning algorithms that rely on labeled data, K-means clustering operates without labels, identifying inherent patterns in the data based solely on the similarities among data points.

The K-means Algorithm aims to partition data points into k distinct clusters, where k is a user-defined parameter. Each cluster is represented by its centroid (the average of all points in the cluster). The algorithm works as follows:

Initialization: Choose k initial centroids, either by selecting random data points or by using methods like the K-means++ algorithm, which ensures initial centroids are spread out.
Assignment Step: Each data point is assigned to the nearest centroid based on a distance metric (typically Euclidean distance). This step forms k clusters, where each cluster contains the points closest to a specific centroid.

Update Step: After assigning data points to clusters, the centroids are recalculated by taking the mean of all points in each cluster. This updated centroid represents the new center of the cluster.
Iterate: The assignment and update steps are repeated until the centroids stabilize (i.e., they no longer change significantly) or until a maximum number of iterations is reached.
The objective of K-means is to minimize the within-cluster sum of squares (WCSS), which measures the variance within each cluster. By minimizing this, the algorithm ensures that data points within each cluster are as close as possible to their respective centroid, creating compact clusters.

Choosing the Number of Clusters (k) is a critical step in K-means. One commonly used technique is the elbow method, which involves running the algorithm for a range of k values and plotting the WCSS for each. The "elbow point," or the point where the WCSS reduction slows significantly, often suggests an optimal k. Another method is the silhouette score, which measures how similar points are within clusters compared to points in other clusters.

Advantages and Limitations: K-means is computationally efficient, especially for large datasets, and is relatively easy to implement. It performs well with spherical or well-separated clusters. However, K-means has limitations: it requires the number of clusters to be specified in advance, which is not always straightforward in unsupervised tasks. Additionally, it is sensitive to outliers and may perform poorly with non-spherical or overlapping clusters.
Image Compression: By clustering pixel colors, K-means can reduce the number of colors in an
In practical implementation, libraries like scikit-learn in Python provide an efficient K-means function that automates many aspects of the process, including initialization and convergence monitoring. By adjusting the parameter k and applying techniques like the elbow method, K-means can help reveal valuable insights from complex datasets.

Result

As a result of this Experiment, we successfully wrote and executed the To implement k means clustering.

Learning Outcomes

Understand and apply the K-means clustering algorithm to group data points into clusters, selecting optimal k and interpreting results effectively.