SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)
(Established under section 3 of the UGC Act 1956)
Re - accredited by NAAC with 'A' Grade
Founder: Prof.Dr. S. B. Mujumdar, M.Sc.,Ph.D. (Awarded Padma Bhushan and Padma Shri by President of India)

Name: Naivedya Rai

PRN: 20070122083

Assignment 8:

Write a program to cluster a set of points using K-means.

Theory:

K-means clustering is a popular unsupervised machine learning algorithm used to partition a dataset into distinct, non-overlapping groups or clusters based on the similarity of data points. It is widely employed for tasks such as customer segmentation, image compression, and anomaly detection. The primary objective of K-means is to group data points into clusters in such a way that data points within the same cluster are more similar to each other than to those in other clusters.

Here's a step-by-step explanation of how K-means clustering works:

Initialization:

Choose the number of clusters (K) you want to create. This is often determined based on prior knowledge or through techniques like the elbow method.

Initialize K cluster centroids randomly. These centroids are the initial points around which the clusters will be formed.

Assignment Step:

For each data point in your dataset, calculate its distance (usually Euclidean distance) to each of the K centroids.

Assign the data point to the cluster associated with the nearest centroid. In other words, each data point is placed in the cluster whose centroid is closest to it.

Update Step:

Recalculate the centroids for each of the K clusters. This is done by finding the mean (average) of all data points within each cluster.

The new centroids represent the centre of the clusters and may have shifted based on the data points assigned to them.

Repeat:

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

Final Clustering:

Once the algorithm converges, the data points are divided into K clusters, and each cluster is characterised by its centroid.

The Sum of Squared Errors (SSE), also known as the Sum of Squares Within (SSW), is a commonly used metric in data analysis, particularly in the context of clustering and optimization. It quantifies the overall "error" or "dissimilarity" of data points within a cluster or group. SSE is frequently associated with clustering algorithms like K-means, where it serves as an objective function to measure the quality of cluster assignments.

Here's how the SSE is calculated and what it represents:

- Calculate the distance between each data point and the centroid of its assigned cluster. Typically, Euclidean distance is used, but other distance metrics can also be employed.
- Square each of these distances.
- Sum up all the squared distances for all data points within the same cluster.
- Repeat the above steps for each cluster in your dataset.
- Finally, sum up the squared distances for all clusters. This total is the SSE.

Code:

https://github.com/Naivedya-Rai/ML-Algo-Implement/blob/main/Lab8_kmeans.ipynb