**Assignment – 4**

**K-Means Clustering**

**Problem statement:** The objective of this project is to conduct k-means clustering on a collection of 8 points in a two-dimensional space. The clustering process begins with the definition of initial centroids, set as P1 = [0.1,0.6] and P8 = [0.3,0.2]. After clustering, we determine the cluster to which point P6 = [0.25,0.5] belongs, thereby revealing its cluster affiliation. Additionally, the population of the cluster around centroid P8 is ascertained. Finally, the centroids' positions, denoted as m1 and m2, are updated based on the newly formed clusters, refining the clustering solution. The goal is to help us understand how to do clustering using the K- Means Clustering algorithm.

**Software used:**

1. Python 3.x
2. Google Colab

**Libraries and packages used:** NumPy, Matplotlib, scikit-learn

**Theory:**

**Methodology:**

* K-means Clustering is a popular unsupervised machine learning algorithm used for partitioning data into distinct clusters. It 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.
* The algorithm aims to minimize the variance within each cluster while maximizing the variance between clusters. The process involves iteratively assigning data points to the nearest cluster centroid and updating the centroids based on the mean of the points assigned to each cluster.
* The k-means clustering algorithm mainly performs two tasks:

1. Determines the best value for K center points or centroids by an iterative process.
2. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

**Advantages:**

1. Simplicity: K-means is straightforward to implement and easy to understand.
2. Efficiency: It is computationally efficient and scales well to large datasets.
3. Versatility: Suitable for a wide range of applications and data types.
4. Scalability: Performs well even with a large number of dimensions.
5. Interpretability: Results are easily interpretable, especially with low-dimensional data.

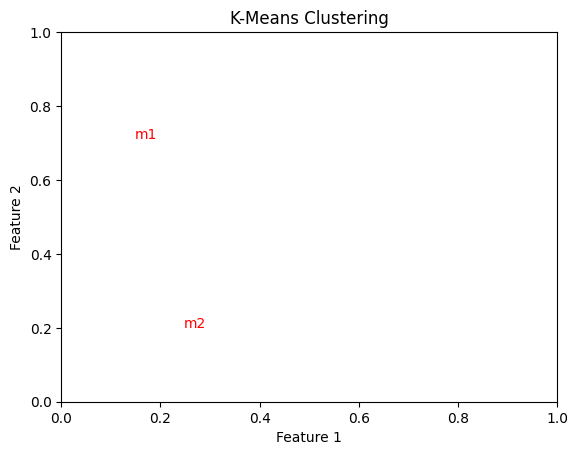
**Disadvantages:**

1. The choice of initial centroids can impact the final clustering results.
2. The algorithm requires specifying the number of clusters beforehand.
3. K-means assumes that clusters are spherical and of similar size.
4. Outliers can significantly affect the cluster centroids and the overall clustering outcome.
5. The algorithm's convergence to a local minimum is not guaranteed to be the global minimum.

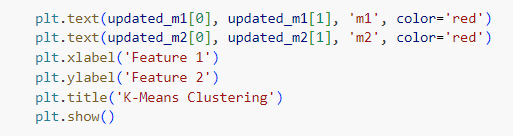
**Applications with example:**

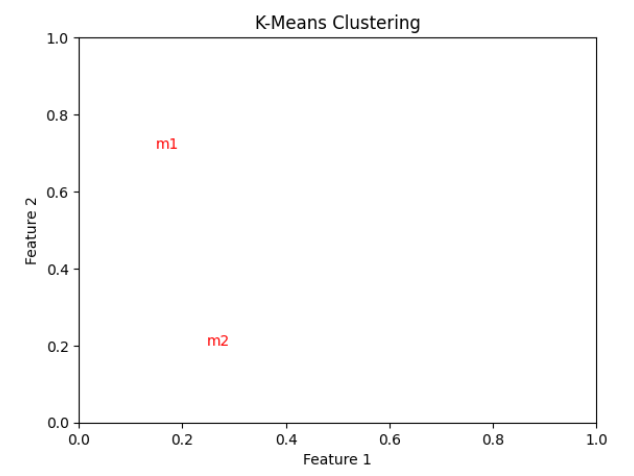
1. Customer Segmentation: In marketing, K-means clustering can be used to segment customers based on their purchasing behavior. For example, a retail company can cluster customers into groups such as high-value customers, frequent buyers, and occasional shoppers.
2. Anomaly Detection: In cybersecurity,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

**Diagram:**



**Screen shots of Output:**

****

****

**Conclusion:**

In conclusion, this assignment demonstrates the effectiveness of K-means clustering in partitioning data into distinct clusters based on similarity. We have explored its simplicity, efficiency, and versatility, showcasing its applicability across various domains such as customer segmentation, anomaly detection, and document clustering. However, the algorithm's performance is influenced by factors like initial centroid selection and the determination of the optimal number of clusters.