**ASSIGNMENT HELP**

**MANUAL**

**ASSIGNMENT-4**



SUBMITTED

TO

VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

FOR THE SKILL AND COMPETENCY EVALUATION OF

**DATA SCIENCE & MACHINE LEARNING**

IN

**CSE AI DEPARTMENT**

BY

**Divyansh Mohta [22210998]**

**Class: S.Y. BTech Division: A Batch: A2**

**Batch Teacher**

**Dr.ANURADHA YENKIKAR.**

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Contents** | **Page No.** |
| **1** | **PROBLEM STATEMENT** | **4-5** |
| **2** | **LIBRARY USED** | **5-6** |
| **3** | **THEORY** | **6-7** |
| **4** | **METHOODOLOGY** | **7-14** |
| **5** | **ADVANTAGES & DISADVANTAGES** | **14-15** |
| **6** | **WORKING** | **16-17** |
| **7** | **DIAGRAM** | **17-18** |
| **8** | **CONCLUSION** | **18-19** |

**1.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 clustering

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?

2. LIBRARY USED:

* scikit-learn (sklearn): Utilized for machine learning tasks such as splitting the dataset into training and testing sets, scaling features, building a decision tree classifier, and evaluating the model's performance using classification metrics.

3. THEORY:

The theory behind the k-means clustering algorithm is as follows:

1. Initialization: The algorithm starts by randomly initializing K cluster centroids. These centroids represent the initial guesses for the centers of the clusters.

2. Assignment Step: Each data point is assigned to the nearest cluster centroid based on a distance metric, commonly the Euclidean distance. This step forms K clusters.

3. Update Step: After all data points are assigned to clusters, the centroids are updated by computing the mean of all data points assigned to each cluster. The new centroids represent the center of mass of the data points in their respective clusters.

4.Iteration: Steps 2 and 3 are repeated iteratively until convergence. Convergence occurs when the cluster assignments and centroids no longer change significantly between iterations or when a predefined number of iterations is reached.

5.Finalization: Once convergence is achieved, the algorithm outputs the final cluster assignments and centroids. Each data point belongs to the cluster associated with the nearest centroid.

K-means clustering is an unsupervised learning algorithm used for partitioning a dataset into K distinct, non-overlapping clusters. It aims to minimize the within-cluster variance or the sum of squared distances between data points and their respective cluster centroids. K-means clustering is widely used for data exploration, pattern recognition, and data compression in various fields such as image processing, data mining, and customer segmentation.

4. METHODS:

1. Initialization:

- Randomly initialize K cluster centroids. Typically, this involves selecting K data points from the dataset as the initial centroids.

2. Assignment Step:

- Assign each data point to the nearest cluster centroid based on a distance metric, usually the Euclidean distance. The data point is assigned to the cluster with the nearest centroid.

3. Update Step:

- After all data points are assigned to clusters, update the cluster centroids by computing the mean of all data points assigned to each cluster. The new centroids represent the center of mass of the data points in their respective clusters.

4. Iteration:

- Repeat the assignment and update steps iteratively until convergence. Convergence occurs when the cluster assignments and centroids no longer change significantly between iterations or when a predefined number of iterations is reached.

5. Finalization:

- Once convergence is achieved, the algorithm outputs the final cluster assignments and centroids. Each data point belongs to the cluster associated with the nearest centroid.

These methods collectively form the k-means clustering algorithm, which aims to partition the dataset into K distinct clusters by minimizing the within-cluster variance.

5. ADVANTAGES AND DISADVANTAGES:

Here are the advantages and disadvantages of the k-means clustering system:

Advantages

1. Simple and Easy to Implement: K-means clustering is easy to understand and implement, making it suitable for a wide range of users, including those with limited experience in machine learning.

2. Efficient:The algorithm is computationally efficient and can handle large datasets with a relatively small number of clusters.

3. Scalability: K-means clustering scales well with the number of data points, making it applicable to datasets of varying sizes.

4. Versatility: It can be applied to various types of data and is not restricted to any particular data distribution. It works well with numerical data but can also be adapted for categorical data with appropriate preprocessing.

5. Interpretability: The clusters generated by k-means clustering are easy to interpret, making it useful for exploratory data analysis and pattern recognition tasks.

Disadvantages:

1. Dependence on Initial Centroids: K-means clustering is sensitive to the initial selection of centroids, which can lead to different final cluster assignments and centroids.

2. Requires Predefined Number of Clusters: The algorithm requires the user to specify the number of clusters (K) beforehand, which may not always be known or intuitive.

3. Sensitive to Outliers: Outliers can significantly impact the clustering results, as they can influence the computation of cluster centroids and distort the cluster boundaries.

4. Assumes Circular Clusters: K-means assumes that clusters are spherical or circular in shape and have similar sizes and densities, which may not always hold true in real-world datasets with irregularly shaped clusters or varying densities.

5. May Converge to Local Optima: The algorithm may converge to a local optimum rather than the global optimum, resulting in suboptimal clustering solutions.

Overall, while k-means clustering is a powerful and widely-used clustering algorithm, it is important to consider its limitations and potential drawbacks when applying it to real-world datasets.

6. WORKING :

Here's how the k-means clustering algorithm works:

1. Initialization:

- Randomly select K data points from the dataset as the initial cluster centroids.

2. Assignment Step:

- Assign each data point to the nearest cluster centroid based on a distance metric, typically the Euclidean distance.

- Form K clusters by assigning each data point to the cluster associated with the nearest centroid.

3. Update Step:

- After all data points are assigned to clusters, update the cluster centroids by computing the mean of all data points assigned to each cluster.

- The new centroids represent the center of mass of the data points in their respective clusters.

4. Iteration:

- Repeat the assignment and update steps iteratively until convergence.

- Convergence occurs when the cluster assignments and centroids no longer change significantly between iterations or when a predefined number of iterations is reached.

5. Finalization:

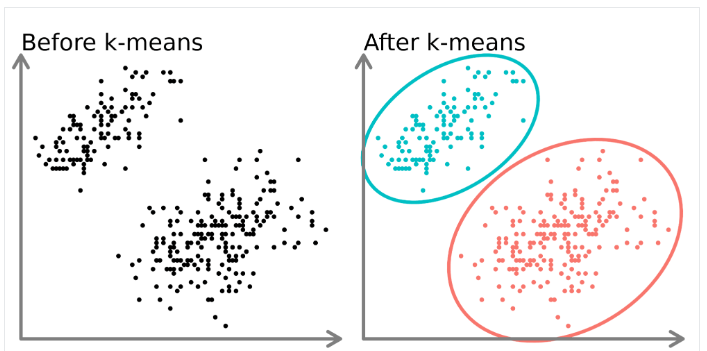
- Once convergence is achieved, the algorithm outputs the final cluster assignments and centroids.

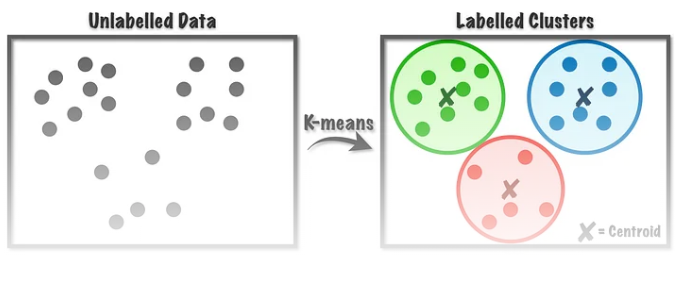
- Each data point belongs to the cluster associated with the nearest centroid.

During each iteration, the assignment step determines the cluster membership of each data point based on the current centroids, while the update step recalculates the centroids based on the current cluster memberships. This iterative process continues until convergence is reached, resulting in stable cluster assignments and centroids.

The goal of k-means clustering is to minimize the within-cluster variance, which is achieved by optimizing the positions of the cluster centroids to better represent the underlying structure of the data.

**7. DIAGRAM :-**





**SOME VISUALIZATION**

**8. CONCLUSION:**

In conclusion, the k-means clustering algorithm provides a straightforward yet effective approach for partitioning a dataset into K distinct clusters. Despite its simplicity, k-means offers scalability, efficiency, and versatility, making it suitable for various applications in data analysis and machine learning. However, it is essential to consider its limitations, such as sensitivity to initial centroids and the need for a predefined number of clusters. Overall, k-means clustering serves as a valuable tool for exploratory data analysis, pattern recognition, and segmentation tasks, offering interpretable and actionable insights into the underlying structure of the data.