Vishwakarma Institute of Information Technology, Pune

**(An Autonomous Institute Maharashtra)**

**A Report**

**on**



**CSE (AI) Department**

**Vishwakarma Institute of Information Technology**

**Academic Year: 2023-24**

**“Assignment 4: K- means Clustering”**

**Submitted by**

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Name** | **Email** |
| **1** | **Omkar Bhaskar Rode** | omkar.22211052@viit.ac.in |

**Under Guidance of**

Dr. Anuradha Yenkikar

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?

Packages / Libraries used: -

• NumPy: Utilized for numerical computations, array manipulation, and handling multi-dimensional arrays. NumPy provides efficient implementations of mathematical functions and operations, making it essential for various machine learning tasks, such as data preprocessing, feature engineering, and model training.

• Matplotlib: Employed for creating static, interactive, and publication-quality visualizations in Python. Matplotlib provides a wide range of plotting functions and customization options, enabling the creation of various types of plots, including line plots, scatter plots, histograms, and bar plots. Matplotlib is highly customizable, allowing users to control every aspect of their plots, from colors and markers to labels and axes.

• scikit-learn (sklearn): A comprehensive machine learning library that provides tools for data preprocessing, model selection, model evaluation, and predictive modeling. The KMeans class from sklearn.cluster is utilized for implementing the K-means clustering algorithm. K-means clustering is an unsupervised learning algorithm used for partitioning a dataset into a specified number of clusters.

Theory: -

The primary objective is to apply the K-means clustering algorithm to a given dataset using Python and the scikit-learn library. K-means clustering is an unsupervised learning algorithm that aims to partition a dataset into a predetermined number of clusters. The algorithm iteratively assigns data points to the nearest cluster centroid and updates the centroids based on the mean of the assigned points. This process continues until convergence, resulting in clusters that minimize the within-cluster sum of squared distances.

By leveraging the K-means clustering algorithm, this task seeks to identify natural groupings or patterns within the dataset, facilitating data exploration and segmentation. Clustering analysis can uncover hidden structures and relationships in the data, enabling insights into the underlying characteristics and similarities among data points. Additionally, visualization techniques such as scatter plots and cluster visualization can aid in interpreting the results and communicating findings effectively.

Overall, this project provides hands-on experience in applying unsupervised learning techniques to real-world datasets, fostering a deeper understanding of clustering algorithms and their applications in data analysis and exploration.

K- means Clustering: -

K-Means Clustering is an [Unsupervised Learning algorithm](https://www.javatpoint.com/unsupervised-machine-learning), which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined 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. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.



Methodology: -

* **Data Preparation:**
* Define the data points to be clustered. In this case, the data points are provided as a numpy array.
* **Initial Centroid Selection:**
* Select initial centroids to initialize the K-means clustering algorithm. In this example, the initial centroids are manually specified as the first and last points in the dataset.
* **K-means Clustering:**
* Utilize the KMeans class from scikit-learn to perform K-means clustering on the data points. Specify the number of clusters (k), initial centroids, and other parameters such as random\_state.
* Fit the KMeans object to the data using the fit() method. This process assigns each data point to the nearest centroid and updates the centroids iteratively until convergence.
* Cluster Assignment and Centroid Update:
* Retrieve the cluster labels assigned to each data point and the updated centroid positions after convergence.
* Determine the cluster to which a specific data point belongs (e.g., P6) and calculate the population of the cluster around the second centroid (m2).
* Visualization:
* Visualize the clusters and updated centroids on a scatter plot. Plot the data points, initial centroids, and updated centroids with distinct colors for clarity.
* Optionally, annotate the centroid positions on the plot for better interpretation.
* Documentation and Reporting:
* Document the findings, including the cluster assignments, centroid positions, and any other relevant observations.
* Provide insights into the clustering results and discuss implications or potential applications of the analysis.

Application: -

* **Customer Segmentation:**
* Businesses can use K-means clustering to segment customers based on their purchasing behavior. By clustering customers into groups with similar buying patterns, companies can tailor marketing strategies and promotions to specific segments, thereby improving customer engagement and increasing sales.
* **Market Basket Analysis:**
* Retailers can apply K-means clustering to analyze transaction data and identify associations between products frequently purchased together. This information can be used to optimize product placement, recommend related items to customers, and design targeted cross-selling campaigns.
* **Image Compression:**
* K-means clustering can be used for image compression by reducing the number of colors in an image. By clustering similar pixel colors together and representing them with the cluster centroid, the image size can be significantly reduced without losing too much visual quality.
* **Anomaly Detection:**
* K-means clustering can help identify anomalies or outliers in datasets by clustering normal data points together and isolating data points that do not belong to any cluster. This approach is commonly used in fraud detection, network security, and monitoring systems to detect unusual behavior or suspicious activities.
* **Document Clustering:**
* Text documents can be clustered based on the similarity of their content using K-means clustering. This technique is useful for organizing large document collections, such as news articles or research papers, into thematic clusters, facilitating document retrieval and topic analysis.
* **Genetic Analysis:**
* In genetics, K-means clustering can be applied to gene expression data to identify patterns and group genes with similar expression profiles. This helps researchers understand gene functions, discover biomarkers for diseases, and classify different types of cells or tissues based on gene expression patterns.

Diagrams: -



Conclusion: -

In conclusion, the application of the K-means clustering algorithm provides valuable insights into the structure and patterns within the dataset. By partitioning the data into distinct clusters, we can identify similarities and relationships among data points, enabling us to make informed decisions and derive actionable insights. Through the iterative process of cluster assignment and centroid update, K-means effectively groups data points based on their proximity to cluster centroids, facilitating data exploration and segmentation. Visualization of the clustering results enhances interpretability and aids in the communication of findings to stakeholders. Overall, the utilization of K-means clustering offers a powerful technique for uncovering hidden structures within data and holds significant potential for a wide range of applications across various industries.