Name: Sumit Avinash Solanke  
Roll No: TYITB121  
Assignment No.: 3  
AIM: Assignment on K-Means Clustering on Iris Dataset  
PREREQUISITE: Python programming

THEORY

K-Means clustering is one of the most popular unsupervised learning algorithms used for clustering problems. The goal of K-Means is to partition a given dataset into *K* distinct, non-overlapping clusters based on feature similarity. It works by minimizing the variance within each cluster and iteratively updating cluster centroids to group similar data points together.

Objective

The objective of this assignment is to implement and analyze the K-Means clustering algorithm on the Iris dataset in order to group the iris flowers into clusters based on their features (sepal length, sepal width, petal length, and petal width). This aims to explore the effectiveness of unsupervised learning in identifying natural groupings within the data and to compare clustering results with actual species labels for validation.

Machine Learning: Unsupervised Learning

Unsupervised learning deals with data that has no predefined labels. The algorithm tries to learn patterns and structure from the data without any supervision.

Two common tasks in unsupervised learning are:  
• Clustering: Grouping data points based on similarity.  
• Dimensionality Reduction: Reducing the number of features while retaining important information (e.g., PCA).

K-Means is widely used in clustering applications where natural groupings or patterns need to be discovered from unlabelled data.

How K-Means Works

K-Means operates through the following steps:

1. Initialization: Select *K* initial cluster centroids randomly.
2. Assignment: Assign each data point to the nearest centroid (based on distance metric like Euclidean distance).
3. Update: Recalculate the centroids as the mean of all data points assigned to each cluster.
4. Repeat: Continue the assign-update process until centroids no longer change or a fixed number of iterations is reached.

Choosing the Optimal K

Determining the best number of clusters (K) is crucial. Common methods include:  
• Elbow Method: Plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters and finding the "elbow" point.  
• Silhouette Score: Measures how similar a data point is to its own cluster compared to other clusters. Higher scores indicate better-defined clusters.

Advantages and Disadvantages of K-Means

Advantages:  
• Easy to implement and computationally efficient  
• Works well with large datasets  
• Produces tighter clusters when the data is well-separated

Disadvantages:  
• Requires pre-specifying the number of clusters (K)  
• Sensitive to the initial placement of centroids  
• Struggles with clusters of different shapes or densities  
• Not suitable for non-spherical clusters

Dataset Description: Iris Dataset

Dataset Name: Iris Dataset  
Source: Introduced by Sir Ronald A. Fisher in 1936, this dataset is widely used for classification and clustering tasks in machine learning.

Overview

The Iris dataset contains 150 records of iris flowers, each with four numerical features and one categorical target class (for validation). The records are equally divided among three iris species:  
• Iris setosa  
• Iris versicolor  
• Iris virginica

Each row represents one flower’s measurements.

Features and Descriptions

| Feature Name | Data Type | Unit | Description |
| --- | --- | --- | --- |
| sepal\_length | Float | Centimeters | Length of the sepal |
| sepal\_width | Float | Centimeters | Width of the sepal |
| petal\_length | Float | Centimeters | Length of the petal |
| petal\_width | Float | Centimeters | Width of the petal |
| species (target) | Categorical | N/A | Actual species (used for evaluation) |

Sample Records

| sepal\_length | sepal\_width | petal\_length | petal\_width | species |
| --- | --- | --- | --- | --- |
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 7.0 | 3.2 | 4.7 | 1.4 | versicolor |
| 6.3 | 3.3 | 6.0 | 2.5 | virginica |

Dataset Characteristics

• Total Instances: 150  
• Classes: 3 (Setosa, Versicolor, Virginica)  
• Balanced: Yes  
• Missing Values: None  
• Type: Multivariate with continuous features  
• Use Cases: Clustering, classification, dimensionality reduction

Conclusion

K-Means clustering is an effective unsupervised learning algorithm for discovering patterns in unlabeled data. Applying it to the Iris dataset demonstrates how natural groupings can be found based on feature similarity. While K-Means is simple and efficient, it also requires careful consideration of the number of clusters and initial centroid positions. Comparing clustering results with actual species labels allows us to evaluate how well the algorithm approximates true classifications.