Aim- Perform Clustering Analysis.

Objectives-

Implement K-Means clustering on Iris.csv dataset. Determine the number of clusters using the elbow method.

Theory-

What is Clustering?

Clustering or Cluster analysis is the method of grouping the entities based on similarities. Defined as an unsupervised learning problem that aims to make training data with a given set of inputs but without any target values. It is the process of finding similar structures in a set of unlabeled data to make it more understandable and manipulative.

It reveals subgroups in the available heterogeneous datasets such that every individual cluster has greater homogeneity than the whole. In simpler words, these clusters are groups of like objects that differ from the objects in other clusters.

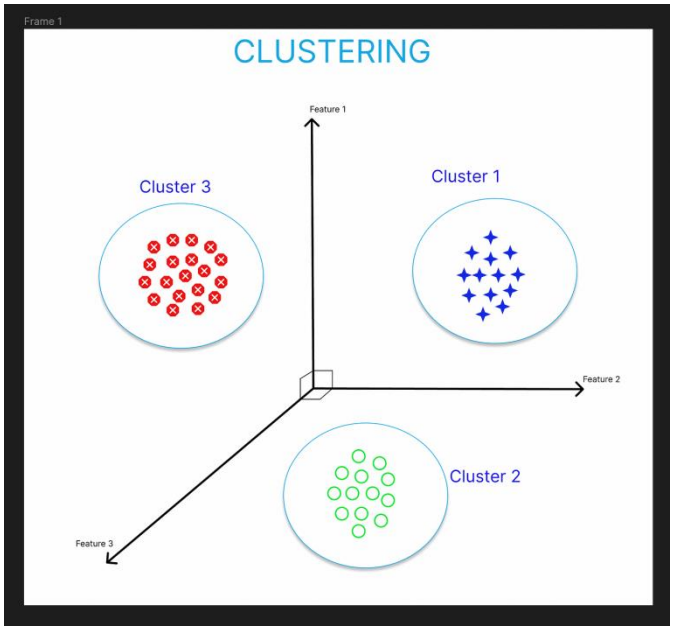
In clustering, the machine learns the attributes and trends by itself without any provided input-output mapping. The clustering algorithms extract patterns and inferences from the type of data objects and then make discrete classes of clustering them suitably.

Types of Clustering Methods

Clustering helps in performing surface-level analyses of the unstructured data. The cluster formation depends upon different parameters like shortest distance, graphs, and density of the data points. Grouping into clusters is conducted by finding the measure of similarity between the objects based on some metric called the similarity measure.

It is easier to find similarity measures in a lesser number of features. Creating similarity measures becomes a complex process as the number of features

increases. Different types of clustering approaches in data mining use different methods to group the data from the datasets. This section describes the clustering approaches.



The various types of clustering are:

1. Connectivity-based Clustering (Hierarchical clustering)
2. Centroids-based Clustering (Partitioning methods)
3. Distribution-based Clustering
4. Density-based Clustering (Model-based methods)
5. Fuzzy Clustering
6. Constraint-based (Supervised Clustering)

Elbow Method for Finding the Optimal Number of Clusters in K-Means-

Clustering is an unsupervised machine-learning technique. It is the process of division of the dataset into groups in which the members in the same group possess similarities in features. The commonly used clustering techniques are K-Means clustering, Hierarchical clustering, Density-based clustering, Model-based clustering, etc. It can even handle large datasets. We can implement the K-Means

clustering machine learning algorithm in the elbow method using the scikit-learn library in Python.

Learning Objectives

- Understand the K-Means algorithm.
- Understand and Implement K-Means Clustering Elbow Method.

What Is the Elbow Method in K-Means Clustering?

The **elbow method is a graphical representation of finding the optimal 'K'** in a K-means clustering. It works by finding WCSS (Within-Cluster Sum of Square) i.e. the sum of the square distance between points in a cluster and the cluster centroid.

Steps involved in K-means clustering:

- Select the number of clusters for the dataset (K)
- Select the K number of centroids randomly from the dataset.
- Now we will use Euclidean distance or Manhattan distance as the metric to calculate the distance of the points from the nearest centroid and assign the points to that nearest cluster centroid, thus creating K clusters.
- Now we find the new centroid of the clusters thus formed.
- Again reassign the whole data point based on this new centroid, then repeat step 4. Continue this for a given number of iterations until the position of the centroid doesn't change, i.e., there is no more convergence.

Finding the optimal number of clusters is an important part of this algorithm. A commonly used method for finding the optimum K value is **Elbow Method.**

K Means Clustering Using the Elbow Method-

- In the Elbow method, we are actually varying the number of clusters (K) from 1 – 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square).
- WCSS is the sum of the squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow.
- As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyze the graph, we can see that the graph will rapidly change at a point and thus creating an elbow shape.
- From this point, the graph moves almost parallel to the X-axis. The K value corresponding to this point is the optimal value of K or an optimal number of clusters.

Conclusion- We implements clustering analysis, to determine the number of clusters using the elbow method.