CLUSTERING

WWW.TANTROSTAV.CO.IN

What is clustering?

- Clustering is a technique in machine learning and data mining used to group similar data points together based on certain criteria.
- In clustering, a set of data points is partitioned into groups, or clusters, such that the data points within each cluster are more similar to each other than to data points in other clusters.
- Clustering is an unsupervised learning technique, which means that it does not rely on labeled data to form the clusters. Instead, it uses similarity measures such as distance or density to group the data points.
- There are several types of clustering algorithms, including k-means clustering, hierarchical clustering, and density-based clustering.

Applications of clustering

- Customer segmentation: Clustering can be used to group customers based on their behavior, preferences, and demographics. This can help businesses to target their marketing campaigns more effectively and provide personalized recommendations.

- Image segmentation: Clustering can be used to separate an image into different regions based on the color, texture, or other visual features. This can be useful for object recognition, image editing, and computer vision applications.

- Document clustering: Clustering can be used to group similar documents based on their content, which can be useful for information retrieval, document classification, and summarization.

K means clustering

- K-means clustering is a popular unsupervised machine learning algorithm used for clustering similar data points together.
- The algorithm works by grouping the data points into a specified number of clusters (k), where each data point is assigned to the cluster with the closest mean (centroid) to it.
- The algorithm iteratively updates the centroids of the clusters until convergence, where the centroids stop changing significantly.

Algorithm working

The k-means algorithm works as follows:

- Choose the number of clusters (k) and randomly initialize k cluster centroids.
- Assign each data point to the nearest centroid (based on Euclidean distance).
- Calculate the new centroid of each cluster as the mean of all the data points assigned to it.
- Repeat steps 2 and 3 until the centroids stop changing significantly, or a maximum number of iterations is reached.

How to choose optimum k-value?

- Choosing an optimum k value is an important task because it ultimately decides how many number of clusters we obtain in our final output.
- So we have many mathematical analysis methods which determine the optimum value of k.
- Some of them are:

1. Elbow method
2. silhoutte score
3. Hierarchical clustering.

- Elbow method:

This method involves plotting the within-cluster sum of squares (WSS) for different values of k and choosing the k value where the decrease in WSS starts to level off, creating an elbow shape in the plot. This is often a good indication of the optimal number of clusters.

- Silhouette score:

The silhouette score measures the quality of clustering, taking into account both the distance between data points within a cluster and the distance between data points in different clusters. The optimal value of k is the one that maximizes the average silhouette score across all clusters

- Hierarchical clustering:

Hierarchical clustering can be used to create a dendrogram that shows how the data points are grouped together at different levels of granularity. The optimal value of k can be chosen by selecting a level of the dendrogram that corresponds to the desired number of clusters.