What is clustering

Clustering is a type of unsupervised machine learning technique used to group similar data points together in a dataset. The goal of clustering is to identify patterns or structures within the data and partition it into subsets or clusters based on the similarity of the data points within each cluster.

It is nothing but a grouping of the data.

How do we find the similarity by finding the distance?

What type of distance is used to find the distance (Euclidean Distance)

Eucedian distance is calculated using Pythagoras’ theorem

To decide how many centroids are required LBOW method is used.

We LBOW by implementing using WCSS

* Inter-Cluster
* Inta Cluster

Evaluation

* Dann Index
* Silhouette coff/Score

K means Clustering

1. Decide n Clusters
2. Init centroids
3. Assign Cluster
4. More Centroids

**K-Means Clustering:**

K-Means is one of the most popular clustering algorithms. It aims to partition the data into K clusters, where K is a user-defined parameter. The algorithm iteratively assigns each data point to the nearest cluster centroid and then recalculates the centroids based on the new assignments. It continues this process until the centroids stabilize.

**Hierarchical Clustering:**

Hierarchical clustering builds a tree-like structure (dendrogram) of nested clusters. It can be agglomerative (bottom-up) or divisive (top-down). In agglomerative hierarchical clustering, each data point starts as its own cluster, and pairs of clusters are merged iteratively based on their similarity until a single cluster remains. In divisive hierarchical clustering, all data points start in one cluster, and the algorithm recursively divides the cluster into smaller clusters.

**DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**

DBSCAN is a popular density-based clustering algorithm used to find clusters of data points in a dataset, particularly when the clusters have irregular shapes and densities. DBSCAN does not require the user to specify the number of clusters beforehand, making it useful when the number of clusters is not known in advance.