SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

Course Name – Data Warehouse & Mining

II Year / IV Semester

Topic – Cluster Analysis

Cluster Analysis

- Grouping similar items
- process of partitioning set of data objects into subset.
- other names of cluster analysis
  - Automatic Classification
  - Data Segmentation
  - outlier detection
  - unsupervised Learning

Types of Data in Cluster Analysis:

- Data Matrix - object by variable structure
- Dissimilarity Matrix - object by object structure

Categories of Clustering Method

- **Hierarchical Clustering** - Builds a hierarchy of clusters, starting with individual data points and merging them iteratively until a single cluster remains, or vice-versa

Types:

Agglomerative (Bottom-up): Starts with each data point as a separate cluster and merges the closest clusters until a single cluster remains.

Divisive (Top-down): Starts with all data points in a single cluster and splits it into smaller clusters until each data point is in its own cluster.

- **Partitioning Clustering** - Divides data into a fixed number of clusters, with each data point belonging to only one cluster.

Types :

K-means: Assigns data points to clusters based on their proximity to cluster centroids (means).

K-medoids: Similar to K-means, but uses medoids (representative data points) instead of means

Categories of Clustering Method

- **Density-based Clustering -** Groups data points based on their density, identifying clusters as dense regions separated by regions of lower density

Types :

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Identifies clusters as dense regions separated by areas of lower density.

OPTICS (Ordering Points To Identify the Clustering Structure): Helps in identifying clusters and hierarchical structures in data.

- **Distribution-based Clustering** - Models clusters using probability distributions, assuming that data points within a cluster follow a certain distribution

Types :

Gaussian Mixture Model (GMM): Assumes that data points are generated from a mixture of Gaussian distributions.

Expectation-Maximization (EM) Clustering: An iterative algorithm used to find the parameters of a GMM

Categories of Clustering Method

Other Clustering Methods:

- **Fuzzy Clustering**: Allows data points to belong to multiple clusters with varying degrees of membership.
- **Constraint-based Clustering:** Incorporates prior knowledge or constraints into the clustering process.
- **Spectral Clustering:** Uses the spectral properties of a data similarity matrix to perform clustering.
- **Grid-based Clustering:** Divides the data space into a grid and performs clustering based on the grid cells.
- **Model-based Clustering:** Assumes a certain model for the data and tries to find the parameters of the model that best fit the data.
- **Affinity Propagation:** A clustering algorithm that uses message passing to identify cluster representatives.
- **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):** An algorithm that efficiently clusters large datasets

THANK YOU