Unit-III (Clustering)

Clustering is an **unsupervised machine learning technique** used to **group similar data points into clusters** based on their characteristics.

The different types of Clustering Algorithms are

- **Partitioning Clustering**
- **Density-Based Clustering**
- **Distribution Model-Based Clustering**
- **Hierarchical Clustering**
- **Fuzzy Clustering**

Partitioning Clustering:

- It is a type of clustering that divides the data into **predefined number of clusters.**
- It is also known as the **centroid-based method**.
- The most commonly used clustering is **K-Means Clustering algorithm.**
- In **K-Means Clustering**, the **centroid** represents the mean position of all points in a cluster.
- **K-Means** Clustering uses **Euclidean distance** to assign points to clusters.
- The **Elbow method** is used to determine the **optimal number of clusters** in K-Means.

Density-Based Clustering:

- The density-based clustering method connects the highly-dense areas into clusters and the arbitrarily shaped distributions are formed as long as the dense region can be connected.
- Clustering algorithm based on Density is **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise)
- **DBSCAN** can identify clusters of arbitrary shapes, making it more effective for **non-spherical clusters.**
- **DBSCAN** can handle **noise and outliers** effectively
- Another example of Density-Based Clustering is **Mean Shift** which does not rely on Distance Metrics

Distribution Model-Based Clustering:

- In the distribution model-based clustering method, the data is divided based on the **probability** of how a dataset belongs to a particular distribution.
- The grouping is done by assuming some distributions commonly **Gaussian Distribution**.
- The example of this type is the Expectation-Maximization Clustering algorithm that uses **Gaussian Mixture Models (GMM).**

Hierarchical Clustering:

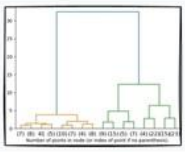
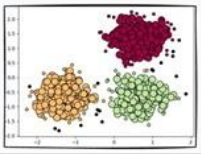
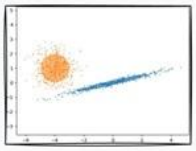
- Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created.
- Hierarchical Clustering is **Connectivity-based Clustering.**

- The dataset is divided into clusters to create a **tree-like structure or nested structure** which is also called a **dendrogram**.
- Hierarchical clustering follows two main approaches:
  Agglomerative (Bottom-Up) – Starts with individual points and **merges** them into clusters.
  Divisive (Top-Down) – Starts with all points in one cluster and **splits** them iteratively.

Fuzzy Clustering

- Fuzzy clustering is a type of soft method in which a data object may belong to more than one group or cluster.
- Each dataset has a set of **membership coefficients**, which depend on the degree of membership to be in a cluster.
- **Fuzzy C-Means (FCM)** is the example of Fuzzy clustering

Types of Clustering Algorithms in Machine Learning

Clustering Algorithm Type		Clustering Methodology	Algorithm(s)
	Centroid-based	Cluster points based on proximity to centroid	KMeans KMeans++ KMedoids
	Connectivity-based	Cluster points based on proximity between clusters	Hierarchical Clustering (Agglomerative and Divisive)
	Density-based	Cluster points based on their density instead of proximity	DBSCAN OPTICS HDBSCAN
	Distribution-based	Cluster points based on their likelihood of belonging to the same distribution.	Gaussian Mixture Models (GMMs)

Applications of Clustering: Clustering is used in following applications.

- Identification of Cancer Cells
- Search Engines
- Customer Segmentation
- GIS Information
- Image Processing & Computer Vision
- Cybersecurity & Anomaly Detection
- Retail & E-Commerce