Parul® University

NAAC A++ ACCREDITED UNIVERSITY

PARUL INSTITUTE OF ENGINEERING & TECHNOLOGY
FACULTY OF ENGINEERING & TECHNOLOGY
PARUL UNIVERSITY

B.Tech, Sem -7th

Subject : Machine Learning

Unit 4 : Unsupervised Learning

Computer Science & Engineering

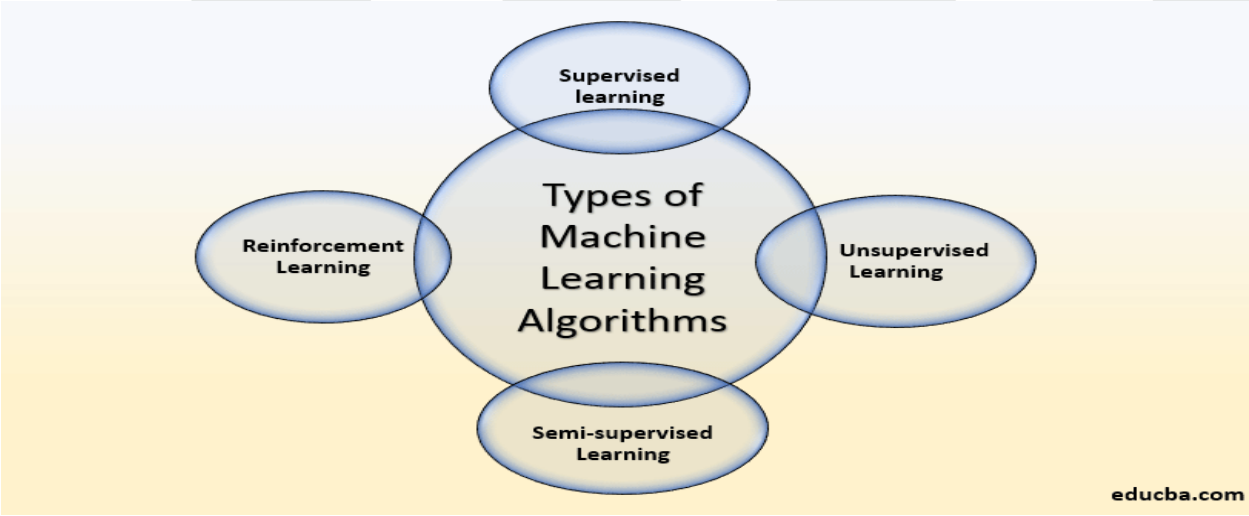
Pragati Mishra(Assistant Prof. PIET-CSE)

Outline

- Clustering basics (Partitioned, Hierarchical and Density based)
- K-Means clustering
- K-Mode clustering
- Self organizing maps
- Expectation maximization
- Principal Component Analysis

What is Machine Learning ?

Machine learning is the study of algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data.



Types of Machine Learning Algorithms

There are three types of machine learning algorithms.

1) Supervised Learning

(i) Regression

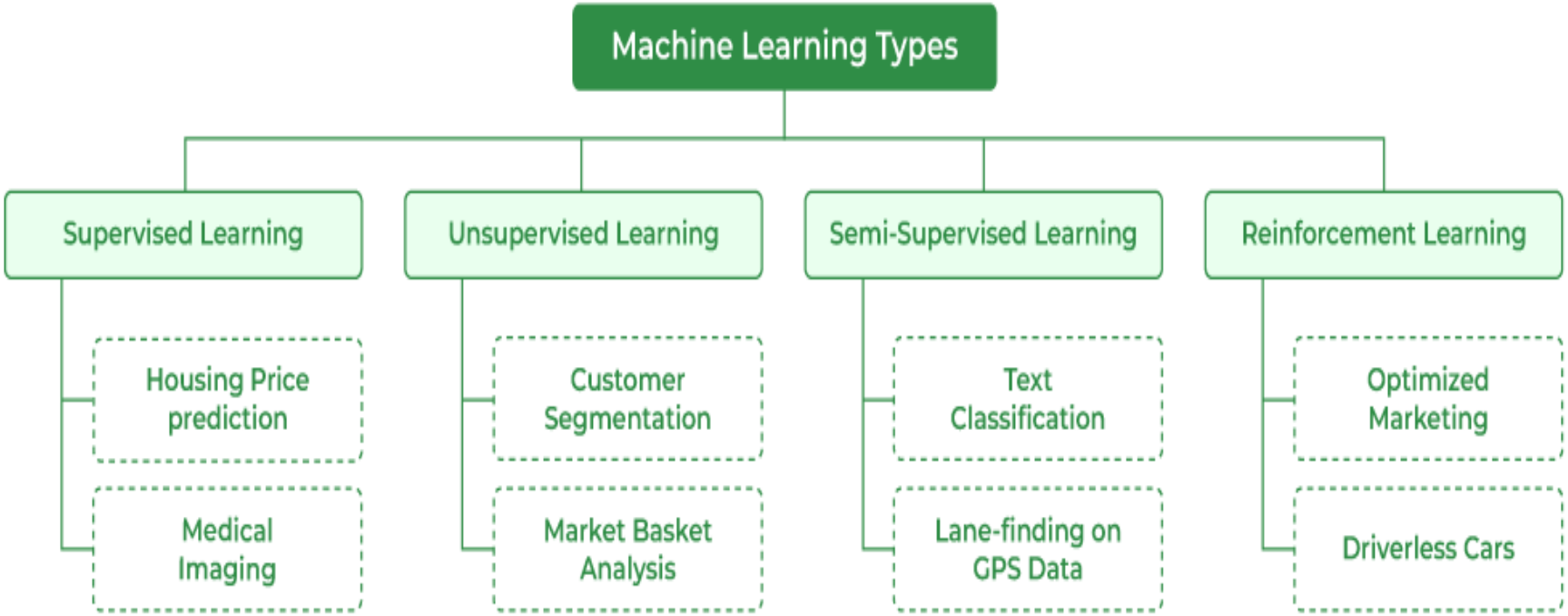
(ii) Classification

2) Unsupervised Learning

(i) Clustering

(ii) Dimensionality Reduction

3) Reinforcement Learning



Parul® University

NAAC A++ ACCREDITED UNIVERSITY

Clustering basics (Partitioned, Hierarchical and Density based)

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as

"A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group.“

- It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.
- It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.
- After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.
- The clustering technique is commonly used for **statistical data analysis.**

Example

Let's understand the clustering technique with the real-world example of Mall: When we visit any shopping mall, we can observe that the things with similar usage are grouped together. Such as the t-shirts are grouped in one section, and trousers are at other sections, similarly, at vegetable sections, apples, bananas, Mangoes, etc., are grouped in separate sections, so that we can easily find out the things. The clustering technique also works in the same way. Other examples of clustering are grouping documents according to the topic.

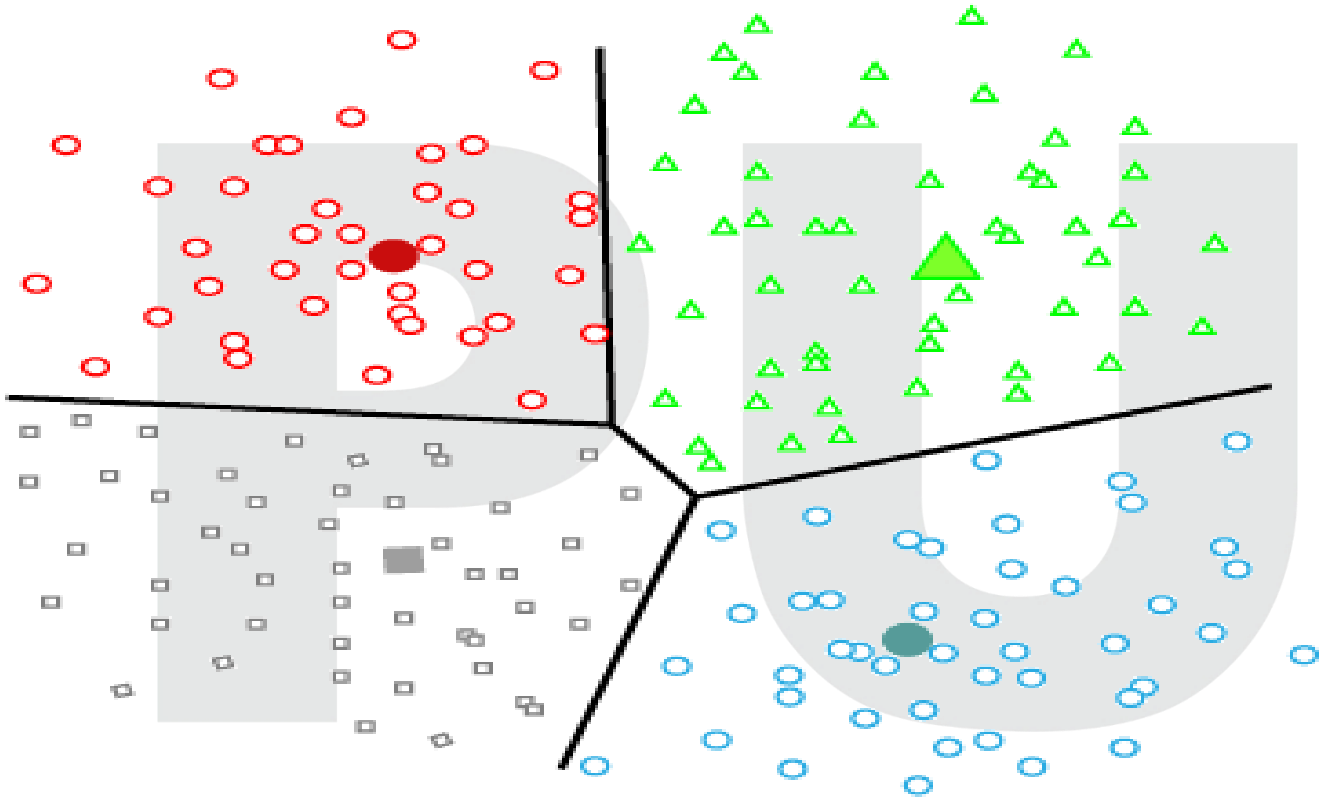
Types of Clustering Methods

- The clustering methods are broadly divided into **Hard clustering** (datapoint belongs to only one group) and **Soft Clustering** (data points can belong to another group also). But there are also other various approaches of Clustering exist. Below are the main clustering methods used in Machine learning:

1. Partitioning Clustering
2. Density-Based Clustering
3. Distribution Model-Based Clustering
4. Hierarchical Clustering
5. Fuzzy Clustering

Partitioning Clustering

- It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the **centroid-based method**. The most common example of partitioning clustering is the **K-Means Clustering algorithm.**
- In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.



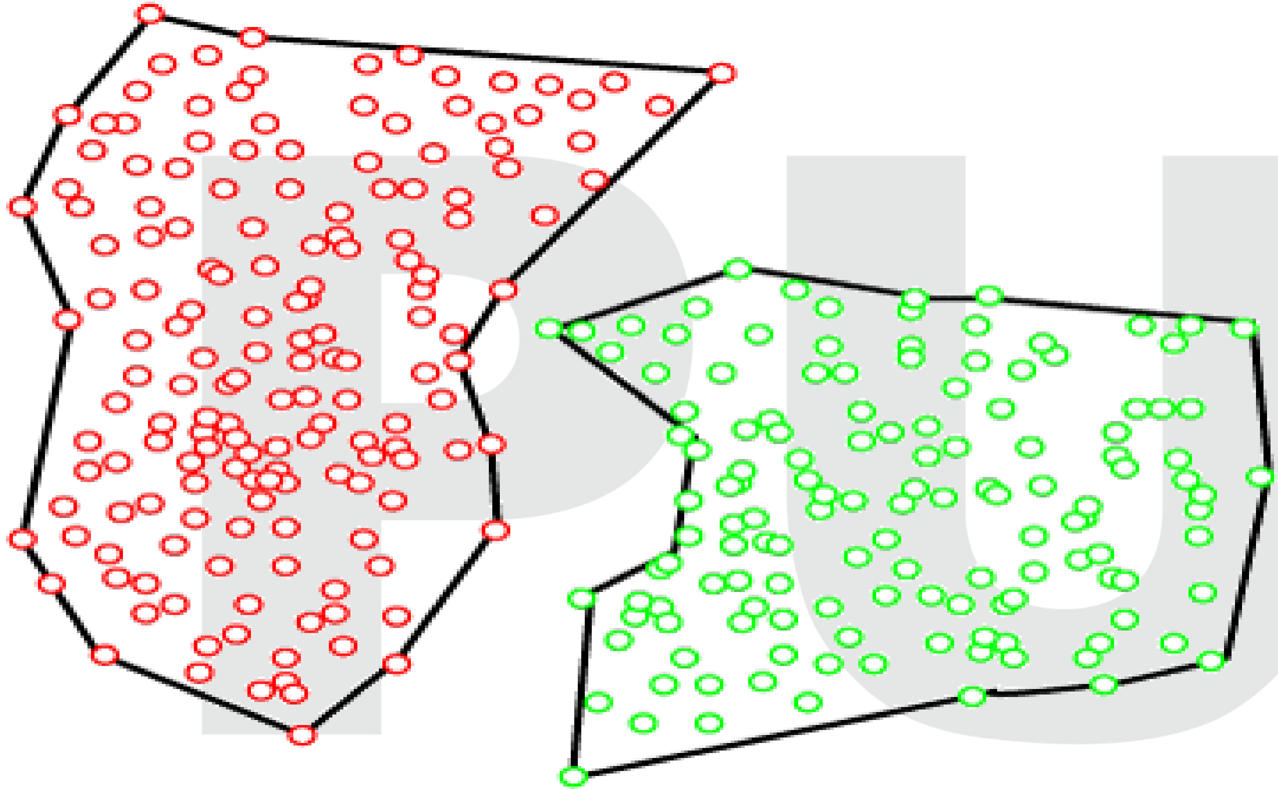
Partitioning Clustering

Parul® University

NAAC A++ ACCREDITED UNIVERSITY

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. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.
- These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.

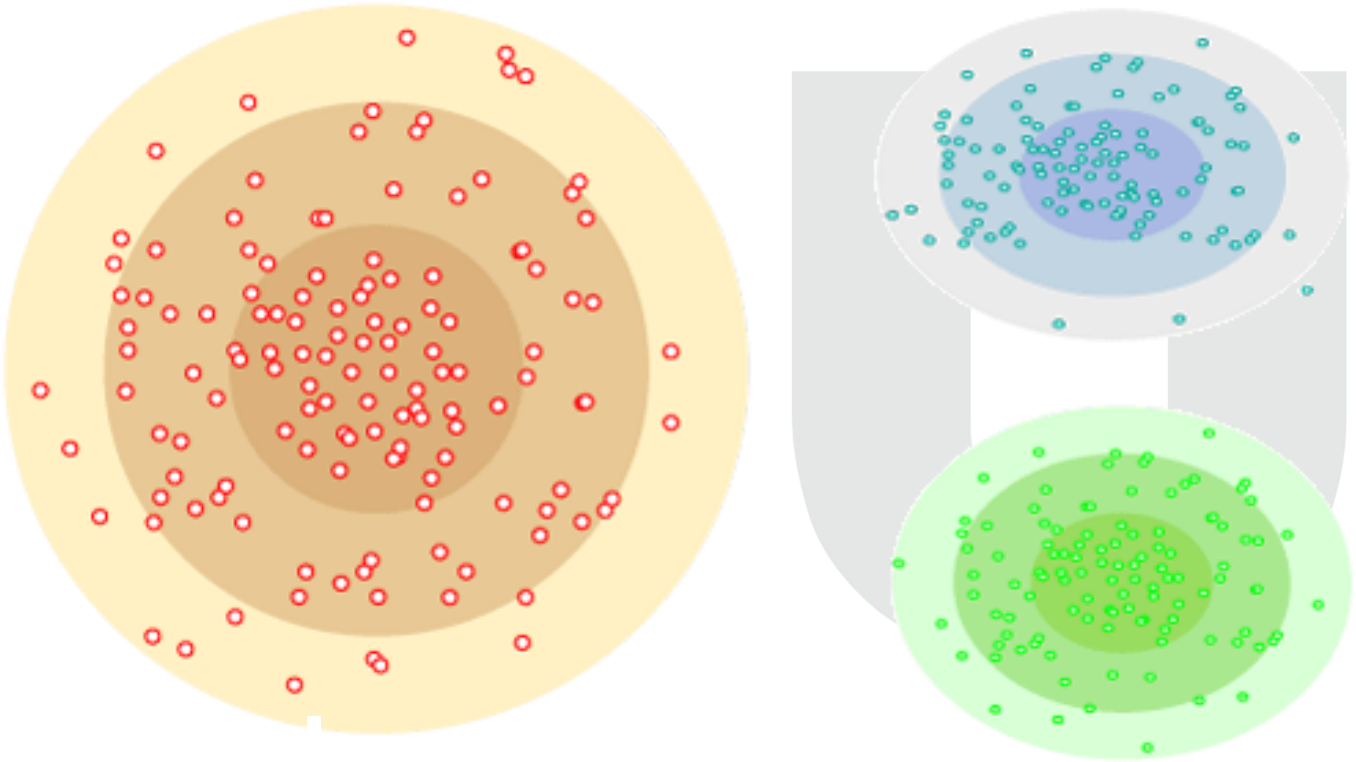


Density-Based Clustering

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**.

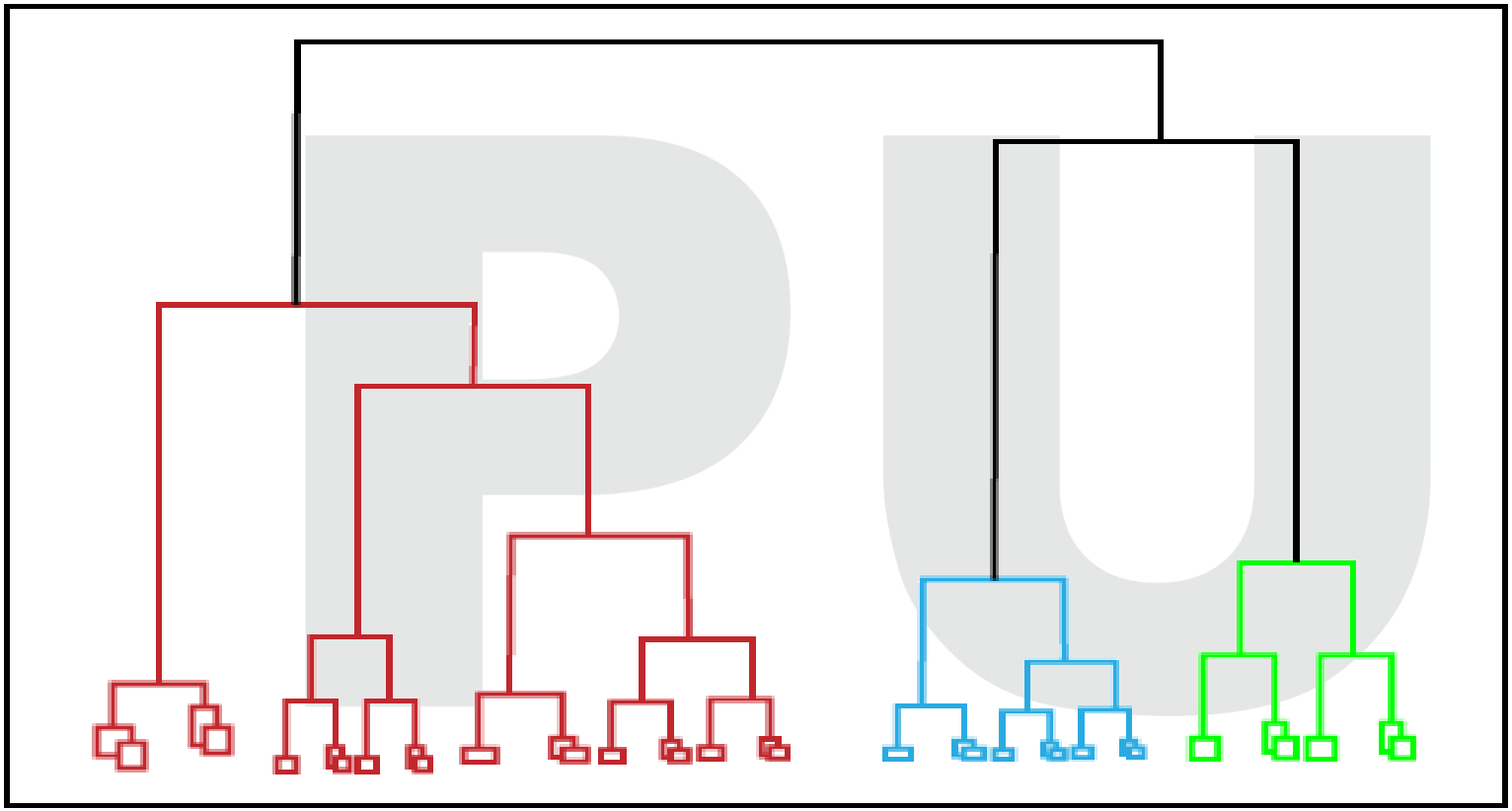
The example of this type is the **Expectation-Maximization Clustering algorithm** that uses Gaussian Mixture Models (GMM).



Distribution Model-Based Clustering

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. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a **dendrogram**. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the **Agglomerative Hierarchical algorithm**.



Hierarchical Clustering

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 algorithm** is the example of this type of clustering; it is sometimes also known as the Fuzzy k-means algorithm.

Applications of Clustering

Below are some commonly known applications of clustering technique in Machine Learning:

•**In Identification of Cancer Cells:** The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.

•**In Search Engines:** Search engines also work on the clustering technique. The search result appears based on the closest object to the search query. It does it by grouping similar data objects in one group that is far from the other dissimilar objects. The accurate result of a query depends on the quality of the clustering algorithm used.

•**Customer Segmentation:** It is used in market research to segment the customers based on their choice and preferences.

•**In Biology:** It is used in the biology stream to classify different species of plants and animals using the image recognition technique.

•**In Land Use:** The clustering technique is used in identifying the area of similar lands use in the GIS database. This can be very useful to find that for what purpose the particular land should be used, that means for

K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.

Parul® University

NAAC A++ ACCREDITED UNIVERSITY

What is K-Means Algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

“It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.”

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

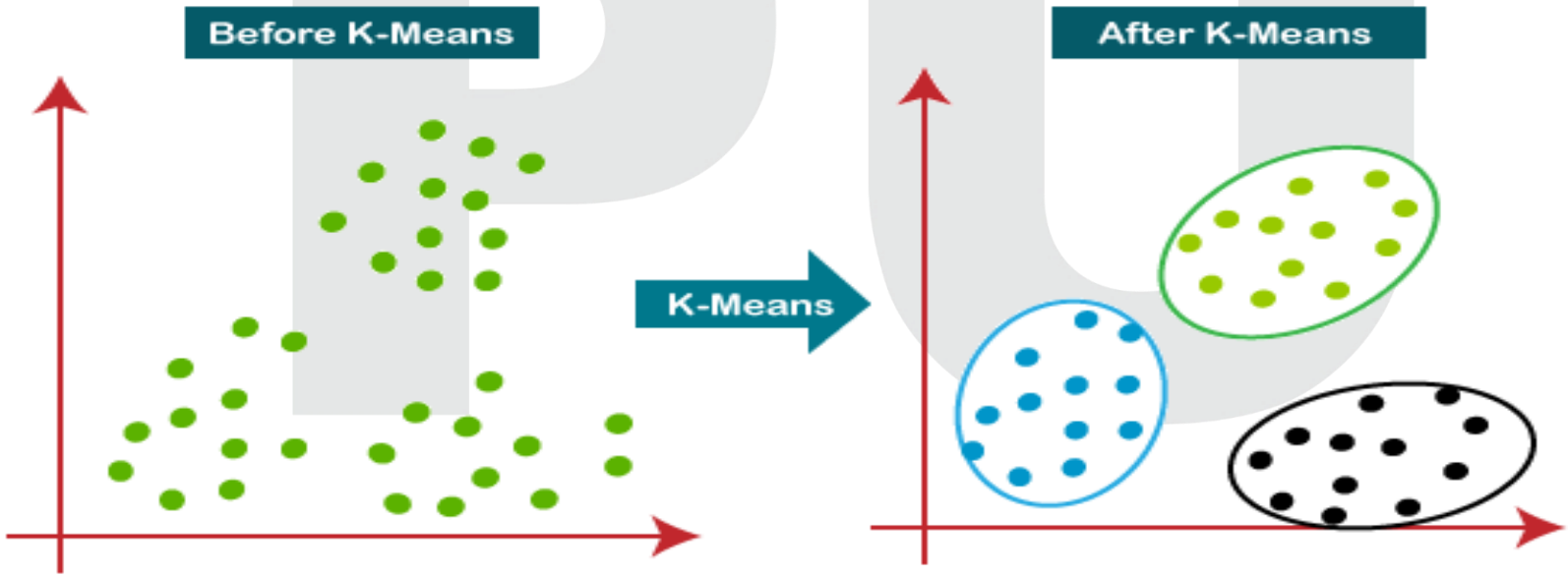
The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



K-Modes Clustering

K-Modes is a clustering algorithm used to partition categorical data into KK clusters. It is an extension of the K-Means algorithm, which is designed for numerical data. K-Modes modifies the clustering process to handle categorical attributes, where it uses modes instead of means and a different dissimilarity measure.

Parul® University
NAAC A++ ACCREDITED UNIVERSITY

Key Concepts

1.Categorical Data: Data that can take on one of a limited, usually fixed, number of possible values, assigning each individual to a particular group or nominal category.

2.Mode: The value that appears most frequently in a dataset.

Steps of the K-Means Algorithm

1. **Initialization**:
    1. Choose the number of clusters KK.
    2. Initialize KK cluster modes. This can be done randomly from the data points or using a heuristic.

2. **Assignment**:
    1. Assign each data point to the nearest cluster mode. This is typically done using a dissimilarity measure suitable for categorical data, such as the Hamming distance (count of

3. Update:

1. Recalculate the modes of each cluster. The new mode for a cluster is the most frequent value for each attribute among the data points in the cluster.

4. Repeat:

1. Repeat the assignment and update steps until convergence (i.e., the modes no longer change significantly) or a maximum number of iterations is reached.

Dissimilarity Measure

For categorical data, the dissimilarity between two data points or between a data point and a cluster mode is often measured by the Hamming distance, which counts the number of mismatches between the attribute values:

$$d(x, y) = \sum_{i=1}^{m} \delta(x_i, y_i)$$

where $\delta(xi,yi)\delta(xi,yi)$ is 0 if $xi=yixi=yi$ and 1 if $xi\neq yixi\neq yi$.

Objective Function

The objective of K-Modes is to minimize the total dissimilarity within clusters:

$$J = \sum_{j=1}^{K} \sum_{x \in C_j} d(x, \mu_j)$$

where μjμj is the mode of cluster CjCj.

Example of K-Modes Algorithm in Action

Let's consider a simple example to illustrate the K-Modes algorithm:

1.Initialization:

1. Suppose we have a dataset with categorical attributes and we choose $K=2$.
2. Randomly initialize 2 cluster modes (e.g., $\mu 1$ and $\mu 2$).

2.First Assignment:

1. Assign each data point to the nearest mode based on the Hamming distance.

3. First Update:

1. Calculate the new modes for each of the 2 clusters based on the most frequent values of the attributes.

4. Repeat:

1. Reassign data points to the nearest mode and update modes until the modes stabilize

Applications

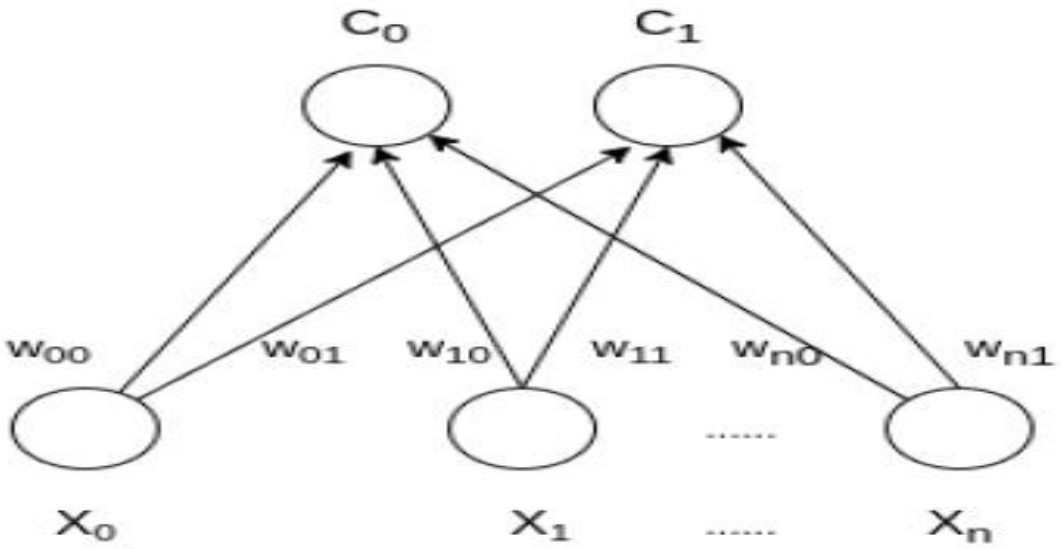
K-Modes is widely used in scenarios involving categorical data:

- **Market Segmentation**: Grouping customers based on purchasing behavior and demographic information.
- **Bioinformatics**: Clustering genetic data with categorical attributes.
- **Social Sciences**: Analyzing survey data where responses are categorical.

Self organizing maps

- Self Organizing Map (or Kohonen Map or SOM) is a type of Artificial Neural Network which is also inspired by biological models of neural systems from the 1970s.
- It follows an unsupervised learning approach and trained its network through a competitive learning algorithm. SOM is used for clustering and mapping (or dimensionality reduction) techniques to map multidimensional data onto lower-dimensional which allows people to reduce complex problems for easy interpretation.
- SOM has two layers, one is the Input layer and the other one is the Output layer.

The architecture of the Self Organizing Map with two clusters and n input features of any sample is given below:



Expectation-Maximization

Expectation-Maximization (EM) is an iterative optimization algorithm used to find maximum likelihood estimates of parameters in probabilistic models, especially when the model involves latent variables or incomplete data. The EM algorithm is widely used in various machine learning applications, including clustering, density estimation, and data imputation.

Key Concepts

1. **Latent Variables**: Variables that are not directly observed but are inferred from the observed data.
2. **Complete Data**: The combination of observed and latent variables.
3. **Incomplete Data**: The observed data alone, without the latent variables.

Steps of the EM Algorithm

The EM algorithm consists of two main steps that are iteratively applied until convergence:

1.Expectation Step (E-Step):

1. Compute the expected value of the log-likelihood function, with respect to the conditional distribution of the latent variables given the observed data and the current parameter estimates.

2. Essentially, this step estimates the missing data given the observed data and current parameter estimates.

2. Maximization Step (M-Step):

1. Maximize the expected log-likelihood found in the E-step with respect to the parameters to obtain updated parameter estimates.

Advantages

- **Handles Missing Data**: EM can efficiently handle datasets with missing or incomplete data.
- **Convergence**: The algorithm is guaranteed to converge to a local maximum of the likelihood function.
- **Flexibility**: EM is applicable to a wide range of models, including those with latent variables.

Disadvantages

- **Local Maxima**: The algorithm may converge to a local maximum rather than the global maximum.
- **Initialization Sensitivity**: The choice of initial parameters can significantly affect the outcome.
- **Computational Cost**: Each iteration can be computationally expensive, especially for large datasets or complex models

Applications

- **Clustering**: EM is used in clustering algorithms like Gaussian Mixture Models (GMMs) where clusters are modeled as mixtures of Gaussian distributions.

- **Density Estimation**: Estimating the probability density function of a dataset.

- **Image Processing**: Segmentation and object recognition tasks.

- **Natural Language Processing**: Topic modeling (e.g., Latent Dirichlet Allocation).

Principal Component Analysis

- Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning.
- It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.
- These new transformed features are called the Principal Components. It is one of the popular tools that is used for exploratory data analysis and predictive modeling.
- It is a technique to draw strong patterns from the given dataset by reducing the variances.

The PCA algorithm is based on some mathematical concepts such as:

- Variance and Covariance
- Eigenvalues and Eigen factors

Some common terms used in PCA algorithm:

•**Dimensionality:** It is the number of features or variables present in the given dataset. More easily, it is the number of columns present in the dataset.

•**Correlation:** It signifies that how strongly two variables are related to each other. Such as if one changes, the other variable also gets changed. The correlation value ranges from -1 to +1. Here, -1 occurs if variables are inversely proportional to each other, and +1 indicates that variables are directly proportional to each other.

- **Orthogonal:** It defines that variables are not correlated to each other, and hence the correlation between the pair of variables is zero.
- **Eigenvectors:** If there is a square matrix M, and a non-zero vector v is given. Then v will be eigenvector if Av is the scalar multiple of v.
- **Covariance Matrix:** A matrix containing the covariance between the pair of variables is called the Covariance Matrix.

Applications

- PCA is mainly used as the dimensionality reduction technique in various AI applications such **as computer vision, image compression, etc.**

- It can also be used for finding hidden patterns if data has high dimensions. Some fields where PCA is used are Finance, data mining, Psychology, etc.

Parul® University

NAAC A++ ACCREDITED UNIVERSITY

Thank You!!!

Parul® University

www.paruluniversity.ac.in