Page No.:
Date:

3rd

# Unit → II

## ☆ Association Rule :-

- Association rule mining finds interesting associations and relationships among large sets of data items.
- This rule shows how frequently a itemset occurs in a transaction.
- A typical example is a Market Based Analysis.
- Market Based Analysis is one of the key techniques used by large relations to show association between items.
- It allows retailers to identify relationships between the items that people buy together frequently.
- We can find rules that will predict the occurrence of an item based on the occurrences of other items in the transactions.

## ✳ Example of Association Rules :-

{Diaper} → {Beer},
{Milk, Bread} → {Eggs, Coke},
{Beer, Bread} → {Milk},
Implication means co-occurrence, not Causality!

Market - Basket transactions

| TID | Items |
|---|---|
| 1. | Bread, Milk |
| 2. | Bread, Diaper, Beer, Eggs |
| 3. | Milk, Diaper, Beer, Coke |
| 4. | Bread, Milk, Diaper, Beer |
| 5. | Bread, Milk, Diaper, Coke |

# ★ Definition: Association Rule

- Association Rule:-

→ An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets.

→ Example:
{Milk, Diaper} → {Beer}

- Rule Evaluation Metrics:-

→ Suppor (s):-
- Fraction of transactions that contain both X and Y.

→ Confidence (c):-
- Measures how often items in Y appear in transactions that contain X.

| TID | Items |
|---|---|
| 1. | Bread, Milk |
| 2. | Bread, Diaper, Beer, Eggs |
| 3. | Milk, Diaper, Beer, Coke |
| 4. | Bread, Milk, Diaper, Beer |
| 5. | Bread, Milk, Diaper, Coke |

Page No.:
Date:

Example:-

{Milk, Diaper} ⇒ Beer

$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$

$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$

# ☆ Mining Association Rules:-

- Two-step approach:-

1. Frequent Itemset Generation:-

- Generate all itemsets whose support ≥ minsup.

2. Rule Generation:-

- Generate high confidence rules from each frequent itemset.
- where each rule is a binary partitioning of a frequent itemset.

- Brute-force approach:-

- Each itemset in the lattice is a candidate frequent itemset.
- Count the support of each candidate by scanning the database.

Page No. :
Date :

Transactions

| TID | Items |
|---|---|
| 1. | Bread, Milk |
| 2. | Bread, Diaper, Beer, Eggs |
| 3. | Milk, Diaper, Beer, Coke |
| 4. | Bread, Milk, Diaper, Beer |
| 5. | Bread, Milk, Diaper Coke |

N ↑ ↓ (rows)

← w →

List of Candidates

M ↑ ↓

- Match each transaction against every Candidate.
- Complexity – O(NMw) => Expensive since $M = 2^d$ !!!

- Rule Generation :-

- Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement.

⇒ If {A, B, C, D} is a frequent itemset, Candidates rules :

- $ABC \rightarrow D$, $ABD \rightarrow C$, $ACD \rightarrow B$, $BCD \rightarrow A$,
- $A \rightarrow BCD$, $B \rightarrow ACD$, $C \rightarrow ABD$, $D \rightarrow ABC$,
- $AB \rightarrow CD$, $AC \rightarrow BD$, $AD \rightarrow BC$, $BC \rightarrow AD$,
- $BD \rightarrow AC$, $CD \rightarrow AB$,

- If $|L| = k$, then there are $2^k \rightarrow 2$ Candidate association rules (ignoring $L \rightarrow \emptyset$ and $\emptyset \rightarrow L$).

☆ Introduction of Holdout Method :-

- Holdout method is the simplest kind of method to evaluate a classifier.
- In this method, the data set is separated into two sets, called the training set and Test set.
- A classifier performs function of assigning data items in a given collection to a target category or class

## Example:-

- E-mail in our inbox being classified into spam and, non-spam.
- Classifier should be evaluated to find out, its accuracy, error rate, and error estimates.
- It can be done using various methods.
- One of most simplistic method in evaluation of classifier is "Holdout method".
- In the holdout method, data set is partitioned, such that maximum data belongs to training set and remaining data belong to test set.

## * Cross-Validation:-

- Cross validation is a standard tool in analytics and is an important feature for helping you develop and fine-tune data mining models.
- You use cross-validation after you have created a mining structure and related mining models

ascertain the validity of the model.

## * Cross Validation has the following application:

- Validating the robustness of a particular mining model.
- Evaluating multiple models from a single statement.
- Building multiple model and then identifying the best model based on statistics.

## * Overview of Cross Validation Process:-

- Cross-validation consists of two phases, training and result generation. These phases include the following steps.

* You select a target mining structure.
* you specify the models you want to test this step is optical; you can test just the mining structure as well.
* you specify the parameters for testing the trained models.
* The predictable attributes, predicated value, and

accuracy threshold.
* The number of folds into which to partition the structure or model data.

# * Bootstrapping :-

- A statistical concept, Bootstrapping is a resampling method used to stimulate samples out of a data set using the replacement technique.
- The process of bootstrapping allows one to infer data about the population, derive standard errors, and ensure that data is tested efficiently.
- In simple terms, the Bootstrapping method, in statistics and Machine learning, is a resampling statistical technique that evaluates statics of a given population by testing a dataset by replacing the sample.
- This technique involves repeatedly sampling a dataset with random replacement.
- A statistical test that falls under the category of resampling methods, this method ensures that the statics evaluated are accurate and unbiased as much as possible.
- The Bootstrapping method uses the samples procedure from a study over and over again in order to use the replacement technique and ensure that the stimulated samples lead to an accurate evaluation.

Page No.:
Date:

# * Bootstrapping Method -

## How does it Work

- Invented by Bradley Efron, the bootstrapping method is known to generates new samples or resamples out of the already existing samples in order to measure the accuracy of a samples statistic.
- Using the replacement technique, the method creates new hypothetical samples that help in the testing of an estimated value.

* Here are 3 quick steps that are involved in the Bootstrapping method :-

1. Randomly choose a sample size.
2. Pick an observation from the training dataset in random order.
3. Combine this observation with the sample chosen earlier.

* The bootstrapping methods involves the bootstrapping samples or the training.

1. Parametric Bootstrap Method :-

- In this method, the distribution parameter must be

known.
- This means that the assumption of the kind of distribution the sample has must be provided beforehand.
- It must be known to the user if the sample has Gaussian Distribution or skewed distribution.
- This type of Bootstrap method is more efficient since it already known the nature of distribution.

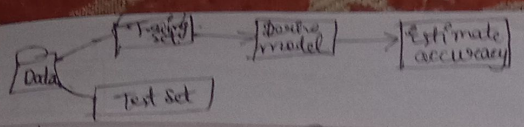
2. Non-Parametric Bootstrap Method:-

- unlike the parametric bootstrap method, this type does not require the parameters of distribution to be known beforehand.
- Therefore this type of bootstrap Method works without assuming the nature of the sample distribution.

## * Random Sampling:-

- Random Sampling is also known as probability Sampling, is a Sampling method that allows for the randomization of sample selection.
- It is essential to keep in mind that samples do not always produce an accurate representation of a population in its entirety, hence, any variations are referred to as sampling errors.

Repeat holdout k-mean and take average accuracy;




Page No. :
Date :

# Types of Random Sampling Methods

1. Simple random Sampling :-

- Simple random Sampling is the randomized selection of a small segement of individuals or members from a whole population.
- It provide each individual or member of a population with an equal and fair probability of being chosen.
- The Simple random Sampling method is one of the most convenient and Simple Sample selection techniques.

2. Systematic Sampling :-

- Systematic Sampling is the selection of specific individuals or members from an entire population.
- The selection often follows a predetermined interval (k). It is less complicated to conduct.

3. Stratified Sampling :-

- It includes the partitioning of a population into subclasses with notable distinctions and variances.

- It allows the researcher to make more reliable and informed conclusions by confirming that each respective subclass has been adequately represented in the selected samples.

4. Cluster Sampling:-

- It is similar to the stratified sampling method includes dividing a population into subclasses.
- Each of the subclass should portray comparable characteristics to the entire selected sample.
- This method entails the random selection of the whole subclass, as opposed to the sampling of members from each subclass.
- This method is ideal for studies that involves widely spread population.

## * Holdout Method:-

- Holdout Method is the simplest sort of method to evaluate a classifier.
- In this method the data set is seperated into two sets, called the training set and test set.
- A classifier performs function of assigning data items in a given collection of a target category

Page No.:
Date:

ore class.

Example → Email in our inbox being classified into Spam and non-Spam.

- Classifier should be evaluated to find out; its accuracy, error rate, and error estimates.
- It can be done using various methods.
- one of the most primitive methods in evaluation of classifier is 'Holdout method'.

## What is the Method for Evaluating the Performance of the Classifier

- There are several methods for estimating the generalization error of a model during training.
- The estimated error supports the learning algorithm to do model choice i.e., to discover a model of the right complexity that is not affected by overfitting.
- Because the model had been constructed.

Page No. :
Date :

- It can be used in the test set to forecast the class labels of earlier unseen data.

- It is often useful to measure the performace of the model on the test set because such a measure provides an unbiased estimate of its generalization error.

* Data

→ It is

→ It

→ It

# Unit → V th

# Cluster Analysis

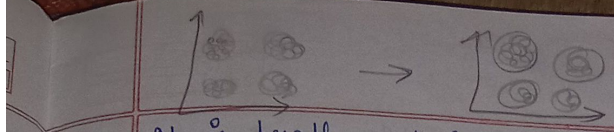
* What is cluster :-

Cluster is a group of objects that belongs to same class. In other words the similar objects are grouped in one cluster and dissimilar are grouped in other cluster.

* What is cluster analysis :-

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) to the objects in other groups.
- The proccess of grouping a set of physical or abstract objects into classes of similar objects is called clustering.

# Application of Cluster Analysis



Page No.:
Date:

- It is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- clustering can also help marketers discover distinct groups in their customer basis. And they can characterize their customer groups based on purchasing patterns.
- In field of biology it can be used to derive plant and animal taxonomies, categorize genes with same functionality.
- clustering also helps in classifying doc. on the web for information discovery.

# Advantages of Cluste

# Analysis

- It is a cheap option as it helps to cut down the Co of preparing the sampling frame or any other administrative factors.
- There is no need of special scales of measurement

- With the help of visual graphics, one can have a clear understanding and comprehension of different cluster.

## Disadvantages of cluster Analysis

- The main point of disadvantages is that the cluster formed are usually not on the basis of any theoretical part. The cluster are rather formed at random.
- Moreover, in a few cases, the process of determining these cluster is very difficult in order to come to a decision.

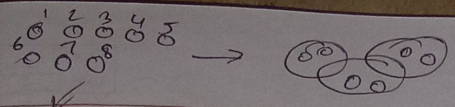
## Types of clustering

It is method to grouping the data items such as each item is only assigned to one cluster.

# Hard clustering :- K-means is one of them.

In hard clustering, each data point either belongs to a cluster completely or not.

# Soft clustering :- Items can exist in multiple clusters, Ex- fuzzy C-means.



Page No. :
Date :

In Soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those cluster is assigned.

# Requirements of clustering

There is they typical requirements of clustering in data mining.

1. **Scalability** → We need highly scalable clustering algorithms to deal with large database.

2. **Ability to deal with different kind of attributes** →

Algorithms should be capable to be applied on any kind of data such as internal based data, categorical, binary data.

3. **Discovery of cluster with attribute shape**:-

The clustering algorithm should be capable of detect cluster of orbitory shape. They should not be bounded to only distance measures that tend to find spherical cluster of small size.

4. **High dimensionality**:-

The cluster algorithm should not only be able handle

low-dimensional data but also the high dimensional space.

5. Ability to deal with noisy data :-

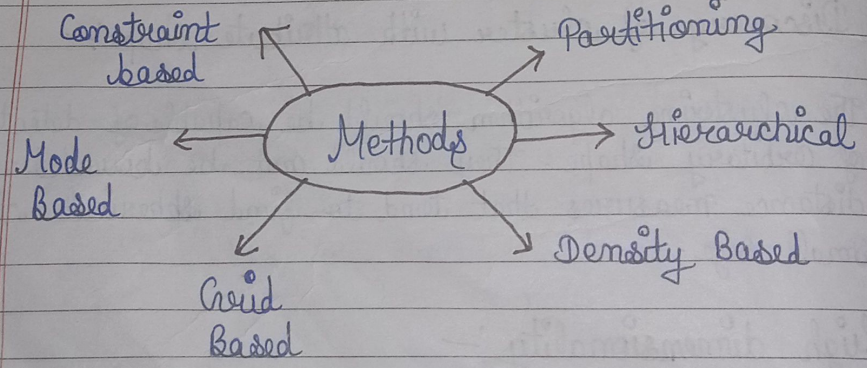
Database contains noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality cluster.

6. Interpretability :-

The cluster results should be interpretable, comprehensible and usable.

# Clustering Methods

The clustering methods can be classified into following categories :-



Page No. :
Date :

## 1. Partitioning Method:-

- Suppose we are given a database of n objects, the partitioning method construct K partition of data.
- Each partition will represents a cluster and $K \leq n$.
- It means that it will classify the data into k groups, which satisfy the following requirements.

- Each group contains at least one object.
- Each object must belong to exactly one group.

### Points to remember:-

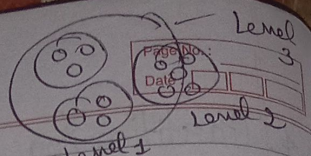
- For a given number of partitions (say k), the partitioning method will create an intial partitioning method will create an intial partitioning.
- Then it uses the iterative relocation data technique to improve the partitioning by moving objects from one group to another.

## 2. Hierarchial Methods:-

- This method creates a hierarchial decomposition of the given set of data objects.
- We can classify hierarchial methods on the basis of how the hierarchial decomposition is formed.

* There are two approaches here:-

1. Agglomerative approach
2. Divisive Approach






3.

## ☆ Agglomerative Approach:-

- This approach is also known as the bottom-up approach.
- In this, we start with each object forming a separate group.
- It keeps on merging the objects or groups that are close to one another.
- It keep on doing so until all of the groups are merged into one or until the termination condition holds.

## ☆ Divisive Approach:-

- This approach is also known as the top-down approach.
- In this, we start with all of the objects in the same cluster.
- In the continuous iteration, a cluster is split-up into smaller clusters.
- It is down until each object in one cluster or the termination condition holds.
- This method is rigid, i.e, once a merging or splitting is done, it can never be undone.

Page No.:
Date:

## 3. Density based Method :-

- This method is based on the notion of density.
- The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e, for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

## 4. Grid-based Method :-

- In this, the object together from a grid. The object space is quantized into finit number of cell that form a grid structure.

## * Advantages :-

- The major advantage of this method is fast.
- It is dependent only on the number of cell that form a grid structure.

## 5. Model-based Methods :-

- In this method, a model is hypothesized for each cluster to find the best file of data for a given model.
- This method locates the cluster by clustering the density function.

It reflects spatial distribution of the data points.

- This method also provides a way to automatically determine the number of cluster based on standard statistics, taking outlier or noise into account.
- It therefore yields robust clustering methods.

# Introduction to k-means Algorithm

- K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid.
- It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means.
- In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance b/w the data points and centroids would be minimim.
- It is to be understood that less variation within the clusters will lead to more similar data points within same cluster.

# Working of k-means Algorithm

* Step 1:-

First, we need to specify the number of clusters, K, need to be generated by this algorithm.

* Step 2:-

Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.

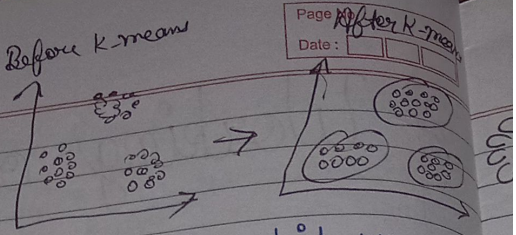
* Step 3:-

Now it will compute the cluster centroids.

* Step 4:-

Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more. first, the sum of squared distance b/w data points and centroids would be computed.

# Example :-




* Let us move to another example in which we are going to apply K-means clustering on sample digits dataset K-means will try to identify similar digits without using the original label information.

- First, we will start by importing the necessary packages :-

```
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()
import numpy as np
from sklearn.cluster import KMeans
```

Next, load the digit dataset from sklearn and make an object of it. we can also find number of rows and columns in this dataset as follows:

- from sklearn.datasets import load_digits
```
digits = load_digits()
digits.data.shape
```

Output :-
(1797, 64)

Page No. :
Date :

# Advantages

- It is very easy to understand and implement.
- If we have large number of variables then, K-means would be faster than Hierarchical clustering.
- On re-computation of centroids, an instance can change the cluster.
- Tighter clusters are formed with K-means as compared to Hierarchical clustering.

# Disadvantages

- It is a bit difficult to predict the number of clusters i.e. the value of K.
- Output is strongly impacted by intial inputs like number of clusters (value of K).
- Order of data will have strong impact on the final output.
- It is very sensitive to rescaling. If we will resc our data by means of normalization or standaridizat

then the output will completely change final output.

# Application of k-Means Clustering Algorithm

* The main goals of cluster analysis are:-

- To get a meaningful intuition from the data we are working with.
- cluster-then-predict where different models will be built for different subgroups.

* To fulfill the above mentioned goals, k-means clustering is performing well enough. It can be used in following applications:-

- Market segmentation
- Document clustering
- Image segmentation
- Image compression
- Customer segmentation
- Analyzing the trend on dynamic data.

# DBSCAN

- DBSCAN stands for Density-based spatial clustering of application with noise.
- It estimates the density by counting the number of points in a fixed-radius neighborhood or $\varepsilon$ and deem that two points are connected only if they lie within each other's neighborhood.

## * Why DBSCAN ?? :-

- Partitioning methods (K-mean, PAM clustering) and hierarchial clustering work for finding spherical shaped clusters or converse cluster.
- In other words, they are suitable only for compact and well-separated clusters.
- Moreover, they are also severely affected by the presence of noise and outliers in the data.

* Real life data may contain irregularities like :-

1. Cluster can be of arbitrary shapes.
2. Data may contain noise.

Page No.:
Date:

☆ DBSCAN algorithm uses two parameters such as ε and Min pts. ε denotes the Eps-neighborhood of a point and Minpts denote the minimum points in an Eps neighborhood.

1. Eps :-

- It defines the neighborhood around a data point i.e if the distance between two point is lower or equal to 'eps' then they are considered as neighbors.
- If eps value is chosen too small then large part of data will be considered as outliers.
- If very large, then the cluster will merge and majority of data point will be in same cluster.
- one way to find the eps value is based on the K-distance graph.

2. Minpts :-

- Minimum number of neighbors within eps radius.
- Larger the dataset, the larger value of Minpts must be chosen.
- As a general rule, the minimum Minpts can be derived from the number of dimensions D in the Dataset as, Minpts >= D+1. The minimum value of Minpts must be chosen at least 3.

Page No.:
Date:

☆ There are a few points to keep in mind :-

☆ Core point :-

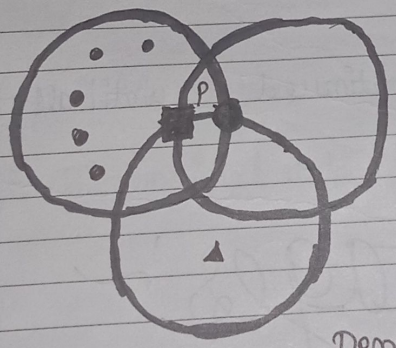
- As the core point is taken that particular point which in Eps - neighborhood, has greater value than a precise number of points that is Minpts.

☆ Border point :-

- As the border point is taken that particular point which in Eps - neighborhood, has less value than a precise number of points that is Minpts.

☆ Noise point :-

- A point that doesnot come under in core or border is said to be a noise point.




■ Core point
● Border point
▲ Noise point

P → neighborhood

Density points

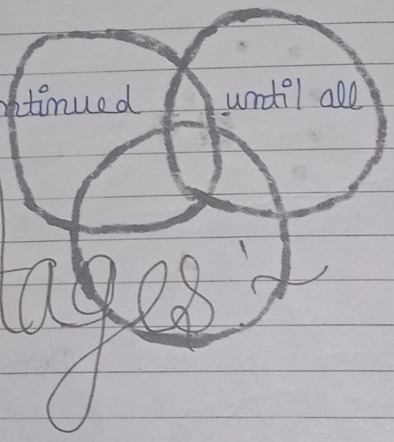
# Algorithm

* Input :-

N object to be cluster and global parameters ε and Minpts.

* Output :- cluster of object.

1. Arbitrary select a point P.
2. Retrieve all point density reachable from P wrt ε and Minpts.
3. If P is a core point a cluster is formed.
4. If P is a border point, then there is no point that is density - reachable and DBSCAN moves to the next point.
5. This process is continued until all the points are processed.



# Advantages :-

→ DBSCAN can find arbitrary shaped cluster using Minpts parameters.

→ The order of the point in the database is insensitive.

→ Handle noise and outliers.

→ Does not require one to specify the number of cluster.

→ It has a notion of noise and is robust to outliers.

## Disadvantages

→ Cannot perform well with large differences in densities.

→ Not suitable when various density involve.

→ High-dimensional data.