# U-5 (BI)

DOMS | Page No.
Date / /

## Impact of ML in BI process.

* Classification problem -
TEL, MLA, HFL, NC, PC.
TMR, P

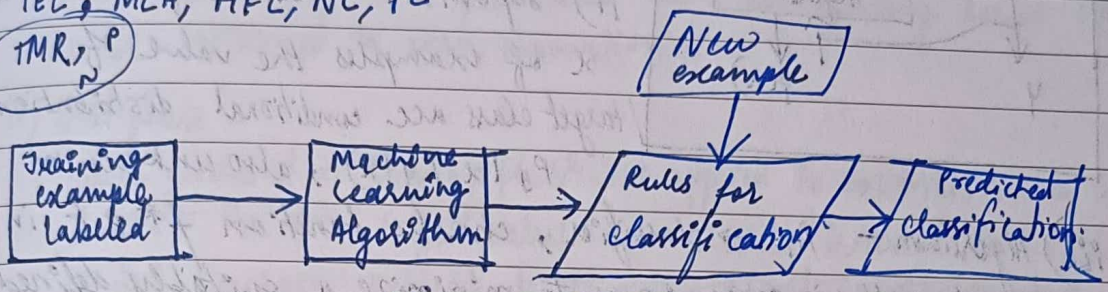



Fig → classification.

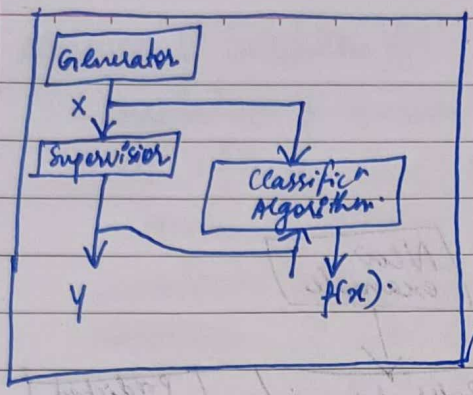
→ Input - Training set of examples with class labels.
→ Aim - To predict categorical class labels for new tuples.
→ Output - classifier is based on training set & class labels.

(1) Classification predicts categorical tables (classes).
Prediction models continous values functions.

(2) Preprocessing data in preparation for classifⁿ & prediction
i) Data cleaning - reduces noise, handles missing values.
ii) Relevance analysis - Removes irrelevant or redundant attributes.
iii) Data Transformation - generalizing data to higher level concepts or normalizing data.

(3) Classification → process of finding a model that describes set of distinguishes data classes or concepts, for purpose of being able to use model to predict the class of objects whose class label is unknown. Derived model is based on analysis of a set of training data.

⇒ Probabilistic structure of learning process for classification.




i) Generator - extracts random vector $x$ of examples acc to an unknown probability distribution.

ii) Supervisor - Returns for each vector $x$ of examples the value of target class acc conditional distribution $P_{y|x}(y|x)$, also unknown.

iii) Algorithm - $A_F \rightarrow$ classifier, chooses function $f^* \in F$ in the hypothesis space so as to minimize a suitably defined loss function.

⇒ Classification Problem.

i) suppose database $D$ is given as $D = \{t_1, t_2, \ldots t_n\}$ & set of desired classes. $C = \{C_1, C_2 \ldots C_n\}$.

the classification problem is to define mapping 'm' in such a way that triple of database $D$ belongs to class of $C$. Actually divides into equivalence classes.

* 2) Classification in 2 step process -

i) Model construction -
- sample object has a predefined class label assigned.
- sample data is training data.
- constructed model based on training data set is represented as classification rules, decision trees or mathematical formulae.

ii) Model usage -
- Fore classifying unknown objects, use construction model.
- compare class label of resultant sample with test sample.
- estimate accuracy

3) Phases of classification Model -

i) Training phase
- classification is applied to subset T for dataset D, called Training set.
- derives classification rules that allow corresponding target class y do be attached to each observation X.

ii) Test phase -
- rules generated in training phase are applied to observation not in dataset D, for target value are already known.
- Test accuracy actual target values ~~are already known~~ class of test set is compared with the predict class.

iii) Prediction Phase -
- Actual use of classification model.
- Assigns target class to not observation to be recorded in literature

* Categories of Classification model - (Heuristic, separation, Probabilistic, Regression).

* Heuristic Model → - make use of classification procedures based on simple & intutive algorithms.
- Example - Nearest Neighbour method.

* Separation Model -
- Divide attribute space $R^n$ into ft disjoint regions $\{S_1, S_2 \ldots S_n\}$
- Separating observation based on target class.
- Example - support vector machine, perceptron methods.

* Probabilistic model -
- Hypothesis is formed regarding functional form of conditional probabilities, $P_{X|Y}(x|y)$ of observations, given the target class.
- Conditional class probability.

DOMS | Page No.
Date / /

## * Regression Model -

- Regression is data mining function that predicts numbers.
- Good choices when all predictor variables are continous valued.

## * Evaluation of Classification Models -

(1) Various methods for estimating a classifier's accuracy are.
A] Holdout Method B] Random sampling C] cross validation.
D] Bootstrap.
All are based on randomly subsampled partitions of data.

(2) Comparision of classifier TO choose the best -
A] confidence intervals.
B] cost benefit analysis & receiver operator curve (BA, ROC).

## * Accuracy & Error Measures -

(1) accuracy of classifier M = acc (M) is percentage of testset tuples that are correctly classified.

$$Acc(M) = \frac{TP + TN}{TP + TN + FP + FN}$$

(2) success - instance class is predicted correctly.
errors - instance class is predicted incorrectly.

(3) Confusion Matrix.

| Actual class | Predicted values: $C_1$ | $C_2$ | total |
|---|---|---|---|
| $C_1$ | TP | FN | P |
| $C_2$ | FP | TN | N |
| | P' | N' | |

TP = class members classified as class mem
FP = non members classified as class members
TN = non class members classified as non class member
FN = class members classified as non class members

P = No. of +ve tuples
N = No. of -ve tuples
P' = No. of tuples labelled +ve
N' = No. of tuples labeled -ve

DOMS | Page No.
Date / /

(3) Sensitivity -
→ True Positive recognition rate.
→ Proportion of Positive tuples correctly identified.

$$\therefore SN = \frac{TP}{P} = \frac{TP}{TP+FN}$$

(4) specificity
→ True negative recognition rate or of negative tuples correctly identify.

$$SP = \frac{TN}{N} = \frac{TN}{FP+TN}$$

Accuracy is function of SN & SP = $\boxed{ACC(M) = SN + SP}$

(5) error rate : % of error made.

$$\text{error rate} = 1 - ACC(M) = 1 - \frac{TP+TN}{TP+FP+FN+TN} = \frac{FP+FN}{TP+FP+FN+TN}$$

(6) Precision - Measure of exactness
No. of positives truly classified as me

$$\boxed{P = \frac{TP}{P'} = \frac{TP}{TP+FP}}$$

(7) Recall
same as Sensitivity

$$\boxed{R = \frac{TP}{P} = \frac{TP}{TP+FN}}$$

(8) F1 score = Harmonic mean of P & R.

$$\boxed{F1 = \frac{2 \times P \times R}{P+R}}$$

(9) $F_\beta$ = weighted measure of P & R assigns $\beta$ times as much weight to recall as precision

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{\beta P + R}$$

$\beta$ is non negative real number.

10) Classifiers can be compared wrt.
i) Robustness
→ Accuracy of rules doesn't change significantly when attaining of test set variable.
→ Ability to handle missing data & outliers.
ii) Scalability.
→ Ability to learn from large datasets.
iii) speed iv) Interpretability:

* Hold out Method :-
(1) Data is split into training dataset & testing dataset.
(2) Training - 2/3, Testing = 1/3.
(3) Train classifier, training dataset is used, & test dataset is used to estimate true error rate.

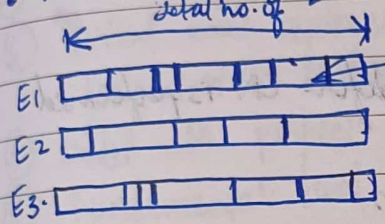
| Training set. | Test set. |
|---|---|

(4) More training leads to better model construction more testing means error estimations are more accurate.
(5) P - samples might not be representation • some classes will have few instances.
S - stratification to ensure both T & V have same no. of instance of same class.
(6) Drawbacks -
i) Requires extra dataset.
ii) Single Train & test experiment - misleading error estimate.

DOMS | Page No.
Date / /

## Random Subsampling -

- Variation of hold out method.
- Holdout method is repeated k times.
- each split randomly selects fixed number examples w/o replacements.
- For each split, we retrain the classifier from stratch.




- overall accuracy is calculated by averaging accuracies from each iteration.

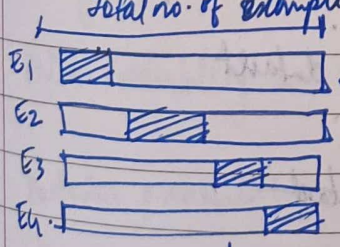
$$E = \frac{1}{k} \sum_{i=1}^{k} E_i$$

## Cross validation (k fold, Leave one out)

1. ensures non overlapping, diff sets.
2. Techniques for evaluating estimating performance by training several ML models on subsets of available of input data & evaluating them on complementary subset of data.
3. used to detect overfitting.

## k cross validation.

Step1 - Data is split into k subsets of equal size usually by RS.
Step2 - each subset is in turn used for both training & testing.




- Advantage is all examples.
- are used for both training & testing.
- ERROR ESTIMATE

$$\hat{E} = \frac{1}{k} \sum_{i=1}^{k} E_i$$

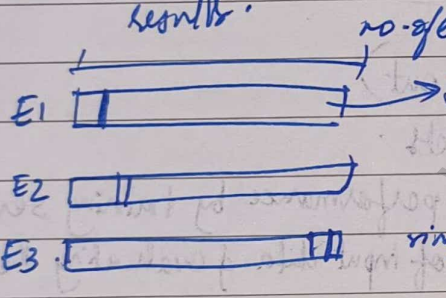
## ⇒ Leave One Out CV -

- If D has N examples, N experiments need to be performed.
- For every experience, training uses N-1 examples remaining for testing
- $E = \frac{1}{N} \sum_{i=1}^{N} E_i$ is avg error rate of test examples, giving true error.

## ⇒ Stratified CV - subsets are stratified before CV is performed

stratified 10 fold CV.

① Gives accurate estimate of evaluation.
② Estimate's variance gets variance gets reduced due to stratified.
③ repeated 10 times & results are all averaged based on previous results.




## ⇒ Bootstrapping -

① CV uses sampling set without replacement ie, once staples or instance is selected, it can't be selected again for training or testing
② bootstrap uses sampling with testing w. replacement to get training set
③ Training set - Dataset of k instances is sampled using replacement k times to form training set of k instances.
④ Test set - separate dataset from original dataset (not part of training set).
⑤ Best error ~~setting~~ estimator for small dataset

DOMS | Page No.
Date / /

## ROC Curve.

1. Receiver Operating Characteristics curve.
2. used to visually compare classification methods.
3. originating roots from signal detection theory.
4. Tradeoff b/w TP rate & FP rate.
5. Accuracy of model can be measured by AUC.
6. vertical axis represents True Positive Rates & horizontal axis represents false positive rates.
7. Model with perfect accuracy will have AUC = 1
8. Fundamental tool for diagonistic test evaluation.

## * Bayesian Methods -

- belongs to family of probabilistic classification model.
- once priori & class conditional probability are known it can explicitly calculate posterior probability.
- Bayesian classifier are statistical classifies.
- Predict class membership probabilities.
- given observation belongs to particular class.
- exhibit high accuracy & speed when it comes to large databases.

Priori probability = Initial probability.

## * Baye's Theorem -

Prior Probability = probability of event based on established knowledge, before any data is collected.

Posterior probability = Revised probability of an event, after taking into consideration new data.

Baye's theorem statement

↳ Let $A_1, A_2 \dots A_n$ be events representing partion of sample S.
Let B be any other defined event on S. If $P(A_i) \neq 0$; $i = 1, 2, n$ & $P(B) \neq 0$ then

$$P(A_i|B) = \frac{P(A_i) \cdot P(B/A_i)}{\sum_{n=1}^{n} P(A_i) \cdot P(B/A_i)} = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}$$

Proof -

$S = A_1 \cup A_2 \cup \dots A_n$ & $B \subset S$.

$\therefore B = B \cap S$

$= B \cap (A_1 \cup A_2 \cup \dots A_n)$.

$= (B \cap A_1) \cup (B \cap A_2) \cup \dots (B \cap A_n)$ (Distributive Law)

$\because (B \cap A_i)$ are mutally disjoint for $i=1$ to $n$;

By addition theorem of probability.

$P(B) = P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n)$.

$= \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)$.

Also we have,

$P(B \cap A_i) = P(B) \cdot P(A_i/B)$.

$P(A_i, B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(A_i) \cdot P(B/A_i)}{\sum_{n=1}^{n} P(A_i) \cdot P(B/A_i)}$

Formula for Baye's Theorem,

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \boxed{\frac{P(A)\ P(B/A)}{\sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)}}$$

{ Red | Domestic | SUV }

$P(R|Y)$ $P(D|Y)$ $P(SUV|Y)$ $P(R|N)$ $P(D|N)$ $P(SUV|N)$.

Yes.

$n = 5$.

$n(R) = 3$.

$P(R|Y) = 3/5$

$P(D|Y) = 1$.

DOMS | Page No.
Date / /

c) Bayesian Belief Networks—

* conditional Independence is defined b/w subsets of variables.
* N/w provides graphical model of casual relⁿ on which learning can be performed.
* Trained N/w can be used for classification.
* Two components.
  i) Directed Cyclic Graph  ii) set of conditional probability tables.

* Directed Acylic Graph—

- each node represents random variable, discrete or continous.
- discrete or continous variables which may correspond to actual attribute or hidden variable
- probabilistic dependence is represented by each graph.
- A → B : A is immediate predecessor or parent of B.
- each variable is conditionaly independat of its non-descendent.

eg— ① Alarm for Burglary & Earthquake.

$$P(AC, VC, A, TE, JB) = P(AC|A) \times P(VC|A) \times P(A|B,E) P(JB) P(JE)$$
$$= 0.90 \times 0.70 \times 0.001 \times 0.99 \times 0.998.$$
$$= 0.00062.$$

② Appl^n of BN —

ML, statistics, computer vision, NLP, speech recognition, error caculation.

## ✱ Logistic Regression

- Regression analysis technique where is outcome variable is binary or dichotomous.
- supervised learning algorithm.
- doesn't involve decision tree more like linear regression.
- can be used with 2 types of target variables.
  i) Categorical TV (binary dichotomous).
  ii) Continous TV (representing probability values as proportion).

✱ Binary logistic Regression model - $P(Y) = \frac{e^z}{1+e^z}$

where $z_0 = \beta_0 + \beta_1 x_1 + \beta_2 \cdot x_2 + \dots \beta_k x_k$

x is independent.

✱ Logistic regression function is written as $= \frac{P(Y=1)}{1-P(Y=1)}$

$$\therefore \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n.$$

- Multiple linear regression allows response variables y to be modelled as linear function of 2+ predictor variables.
- Logistic function is similar to MLR. Such models are called Generalized Linear Mode (GLM).

Here errors do not follow normal distribution & there exists a transformed function of outcome variable but takes linear variable.

DOMS | Page No.
Date / /

## Clustering

1. Unsupervised
2. training sample not provided.
3. Data is not labelled.
4. How can I group their set of items?
5. Unknown no. of classes
6. Used to understand data
7. The function maps the data into one of several clusters which is one of several clusters.
8. which is grouping of data items based on similarities b/w them.
9. No. of clusters is not known before clustering these are identified after completion.

## Classification.

1. Supervised.
2. Training sample provided.
3. Data is labelled.
4. What class does object belong to?
5. Known number of classes.
6. Used to classify further observations
7. Used to classify features observed data into one of several predefined categorical clusters.
8. No. of classes is known before classification as label is predefined. output based on input data.



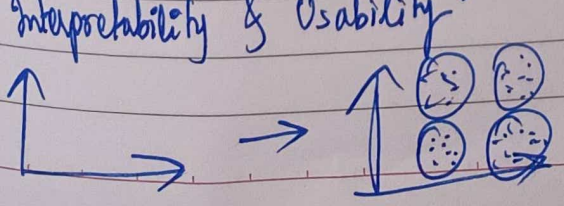
* Clustering -

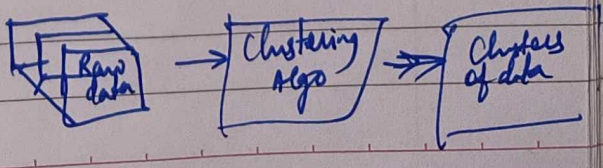
- Clusters are groups such that objects in one group are similar to each other & objects from separate groups are dissimilar
- Clustering analysis is powerful data mining tool. cluster is group of objects that belong to same class.
- Clustering is process of partitioning a set of data into set of meaningful sub-classes.

eg → Hierarchical, Fuzzy, Density based or model based.

- Desirable properties of clustering algorithm -

i) Scalability (wrt time & space).
ii) Ability to deal with different types of data.
iii) Minimal domain knowledge required to determine parameter.
iv) Interpretability & Usability





13/9

DOMS | Page No.
Date / /

- good clustering technique will produce high quality clusters with high intra-class similarities & low inter class similarity
- i) Cluster centroid - point whose parameters values are mean of parameters values of all points in cluster.
  ii) Distance - common metric to see if similarity in population.

(7) Types of data in cluster analysis.

i) Interval Scaled
ii) Binary variables
iii) Nominal, ordinal & ratio variables.
iv) Variable of mixed type.

(8) Types of clustering algorithm -

i) Overlapping
ii) Exclusive
iii) Grid based
iv) Probabilistic
v) Hierarchial
vi) Density Based.

⇒ Appln of Clustering -

1) Marketing → identify distinct group of customer.
2) Land Use → similar land identification for earth database
3) Insurance - moto insurance policy holders with high claim
4) Urban planning - house groups based on type, value & geography
5) Seismology - observed earthquake epicenters.
6) Biology - classify species.
7) Libraries.

DOMS | Page No.
Date / /

→ Typical Requirements / Properties of Clustering in Data Mining.

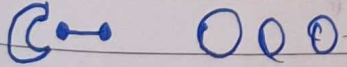
1. Scalability ( wrt time & space)
2. Ability to deal with different types of attributes.
3. Ability to deal with noisy data.
4. Minimal requirement of domain knowledge to determine i/p parameters.
5. Interpretability & usability.
6. constraint based clustering
7. High dimensionality
8. Discovery of clusters with arbitary shape.
9. Incremental clustering & insecurity to order of input records.

→ Types of clusters -

1. Well separated
   - points in nodes are closer than points in seperate ones.

2. Prototype based
   - objects in cluster is closer to prototype center
   - numerical data prototype center.
   - categorical medoid center.

3. Contiguity based
   point increases 2 to it points in same cluster than in others.



4) Density based.
   low density high density.

⇒ Desired features of cluster analysis -

1. Scalability
2. Minimal i/p parameters.
3. only one scan of dataset
4. Ability to stop & resume.
5. Robustness
6. Ability to discover diff cluster shapes.
7. Different data types.
8. Result independant of data i/p order.

⇒ Problems with clustering

1. Do not address all requirements adequantely
2. Complexity due to large dataset + high dimensionality
3. effectiveness depends on distance.
4. If obvious distance measures doesn't as must it which isn't always possible, especially multidimensional space.
5. Result may be interpreted indifferent ways.

# Partition Methods -

1. Partition Method constructs a partition of a database D of n objects into set of k clusters.
2. these are clustering methods used to classify observations into multiple groups based on their similarity.

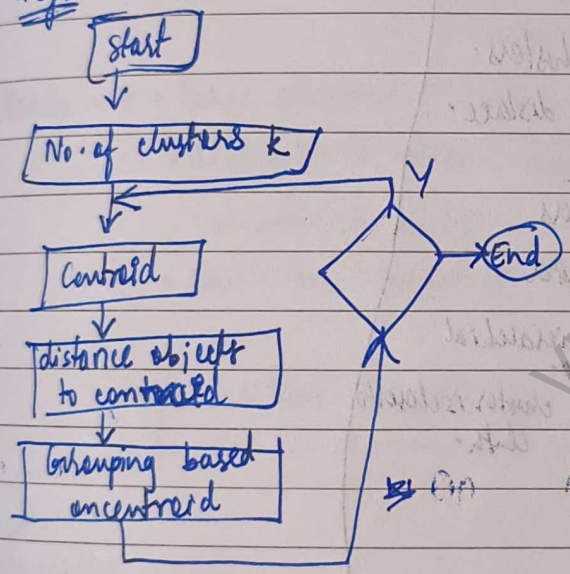
2. Two types
   1. Global optimal method. Exhaustively enumerates all partitions.
   2. Heuristic - k means & k medoids.

DOMS | Page No.
Date / /

* K Means

Attempts to minimize total squared errors.
more efficient O(iKn)
sensitive to outliers
Convex shape required.
efficient for separate clusters
Centroid Based.

Algo -




Strength → 1) efficiency → O(iKn).
2) often terminates local optimum
3) Deterministic annealing to get global solution

Weakness of k means -
1) Applicable only when mean is defined
2) Cannot handle noisy data.
3) Kneeds to be told in advance
4) Not suitable wide non-convex shape.

K - Mediods.

1) Minimizes sum of dissimilarities b/w point labelled to be in classical cluster & points designated as center of cluster.
2) Less efficient
3) not sensitive to outliers
4) convex shape not required.
5) efficient for separate clusters & small datasets.

Representative Object Based

Algo -
- Select k - points as initial medoids
- Assign all points to closest medoid.
- see if any point is better medoid.
- Repeat 2½ x b) until medoids don't change.

DOMS Page No.
Date

# K-Means

↓

- Heuristic Method
- supervised learning algo.
- solves clustering problems
- groups of unlabelled dataset in diff clusters
- k → define no. of predefined clusters needs to be created
  k=2 → 2 clusters.
- each dataset belongs to one group has similar prefect
- min sum of b/w data points & clusters.
- distance calculation → Euclidean distance.

Properties →
- flexway k-clusters
- Atleast one item cluster
- clusters are non-hierarchical & don't overlap.
- every member of cluster is closest to cluth.

Adv →
- efficient in computation.
- easy to implement.

Disadv →
- only when mean is defined
- need to specify k, no. of clusters in advance.
- Trouble with noisy data & outliers.
- not suitable to discover clusters with non complex shapes.

~~Algo →~~

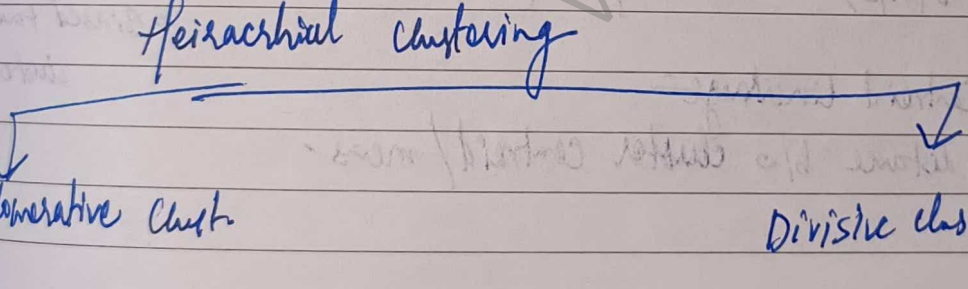
DOMS | Page No.
Date / /

## Hierarchical Methods -

1. Uses distance matrix as clustering criteria.
2. Widely used data analysis tool.
3. Idea is to build binary tree of data that successively merges similar group of points. Visualizing this tree provides useful summary of data

Adv → • simple to implement • easy & results in hierarchy (more info)
• doesn't need prespecify no. of clusters.

Disadv → • large clusters breaks.
• difficult to handle diff sized clusters & shapes.
• sensitive to noise & outliers
• can't be changed or deleted once done.




- bottom-up approach.
- each item → own cluster
- identify small cluster
- iteratively clustered are merged together
- also known as AGNES.

- top down approach
- all item → one cluster
- large cluster
- large clusters are successively divided.
- also known as DIANA.

4 diff methods -

i) Single Lineage -
→ Minimum method, where shortest distance from any object of one cluster to object of another cluster is considered
↳ susceptible to noise (2 outliers).
→ $D(A,B) = \min \{d(i,j)\}$ where $i \in A$, $j \in B$

ii) complete linkage -
→ Maximum method, largest distance is considered.
→ $D(A,B) = \max \{d(i,j)\}$ where $i \in A$, $j \in B$.

iii) Average linkage -
→ distance b/w clusters A & B is average of all distances b/w pairs of $i \in A$ & $j \in B$.
→ Mean distance between elements of cluster.
→ $D(A,B) = \text{Mean} \{d(i,j)\}$.
→ Can handle noise
→ Biased towards global clusters.

iv) Centroid linkage -
→ distance b/w cluster centroid / means.

Evaluation of Clustering Value -

1) Internal validation Methods.

→ Quality of clustering method can be evaluated w/o using external info.

→ Two types -

1) cohesion → metric to evaluate how closely the elements are cluster.
2) separation → Metric to valede level of separation b/w clusters.

→ cohesion & separation do not perform well for density based algorithm.

2) External validation -

→ Associated with supervised learning problem.

→ additional info required.

(A) Silhouette score -

① Measures how elements in one cluster are close to points in neighbouring clusters.

② Principle - Maximum Internal cohesion & Max^m cluster separation.

→ how similar objects is to its own cluster compared to other clustre.

③ Finds optimal value of k, deoring clustering.

④ For each elements silhouette value is calculated.

well clustered elements = 1

poorly clustered elements = -1

DOMS | Page No.
Date / /

## * Association rule -

- unsupervised learning method.
- descriptive method → discover rel$^n$ b/w hidden in large dataset.
- used to mining transactions in databases.
- FREQUENTLY BROUGHT TOGETHER.
- goal → discover interesting relationships among items

- Support → popularity of product of all product transactions.

$$\text{Support}(A) = \frac{\text{No. of transaction in which A appears}}{\text{Total no. of transactions.}}$$

- confidence → likelihood of purashing both products A & B.

$$\text{confidence}(A \to B) = \frac{\text{No. of transaction includes both A \& B}}{\text{No. of transaction include only product A}}$$

- Itemset → set of one or more items.
- Frequent Itemset → an itemset whose support is greater than or equal to minsup threshlod.

## * Market Basket Analysis -

- determine what products customers purchase together.
- name → idea of customers throwing all purchases in shopping (market basket) during grocery shopping.
- Association Analysis & Frequent itemset mining
- creates If-Then scenario rules.

If Item A is purchased then item B is likely to be purchased.

Rule → If {A} then {B}.

- Algorithm → Association Rule & Apriori

Appln → • Retail • Telecommunications - Banks
• Insurance • Medical

DOMS | Page No.
Date / /

* Lift Ratio -
Ratio of confidence to expected confidence.
expected confidence is confidence divided by frequency of tells us how much better the rule is at predicting result.

$$Lift = \frac{(A \cap B)/A}{(B/Total)} = \frac{Confidence}{(B/Total)}$$

* Leverage = Measures diff in $P(x)$ & $P(Y)$ appearing togeme.
Leverage $(x \rightarrow Y)$ = support $(x \wedge Y)$ - support $(x)$ * support $(Y)$

* Appln of MBA
- Retails,
Telecommun.
Banks
Insurance Medical

* Two step process -
i) Find all frequent item sets.
By defn all these items will occur as freq as predetermined support count.
ii) Generate strong association rule from frequent item sets.

$X \rightarrow Y$ means transaction containing items from set $X$ tend to contain items from set $Y$.

* Apriori Algorithm - classic ML algo
- designed to work on databases covering transactions.
- aimed to find subsets which are common to atleast a min. number.
- solves frequent items problem
- used for mining frequent itemsets & deriving association rules from transactions
- major components are support, confidence, lift.
i) calculates support of item size $k=1$.
ii) Prune candidate set by eliminating items w support less than threshold.
iii) Do same for $k \neq 1$ by joining frequent items.

Adv
- easy
- Simple Prune
- easy on large dataset

Limitation
- needs several iterations
- uses uniform minimum support
- difficulties to find rarely occurring events
- some competing approaches focus on partitioning & sampling.

Appln
- education
- Medical Field
- Forestry