Data Mining CSE312A

(Valuable) Search)

→ It is the process of mining knowledge from large amount of data

Raw data ← Big data —Data mining technique→ Useful Data (Extract)

(Social media - Audio, video, text)

Transaction data

why we do data Mining?

1. Companies and organisations get huge amount of data from different sources and platforms.
2. As size of database increasing & it is very difficult to manually search for useful information in it
3. They use data mining technique which include AI & mathematical complex algo. for getting specific & useful data.
4. This specific data helps them for taking data-driven decisions.

• We also get (i) trends, general direction in which data changes over time (+, −, flat)

(ii) pattern → recognizable structure or trend in data that can be useful for making decision or predictions, insights of collected data.

→ Data Mining is also called as 'KDD' (Knowledge Discovery in Database).

Technique:-

Statistics	AI	ML
1. Cluster techniques:-	Diff →	KNN
	AI →	Apriori
	Algo →	K mean
	ANN →	Naive bayes
	Backpropagate	

Ex:- Price Selection in Amazon (0-1000)
Weather forecasting (past data)

* Data Mining is defined as procedure of extracting information from huge sets of data.
Also defined as mining knowledge from data.

Q- What types of data can be mined?

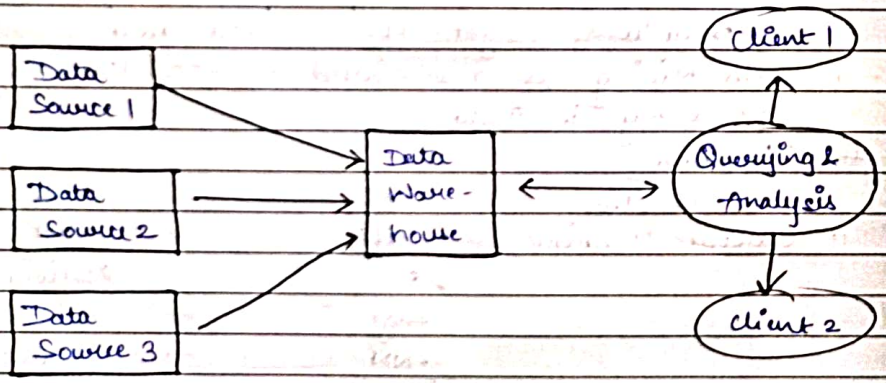
3 types — database
— data ware house
— transactional

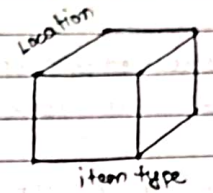
1. Database (RDBMS):- Relational → Tables
→ Set of Tables
has rows & columns
tuples ↳ attributes.

while mining databases, we can search for trends or data patterns (max^m marks, min^m marks)
Eg:- Analysing sales data (any deviation)

2. Data Warehouse data:- Collection of data integrated from different sources with querying (insert, create, delete) and decision making on data

In this, data is stored in multi-dimensional structure (data cube) where each dimension is each attribute





3. Transactional Data:- Each record or attribute is called as Transaction.
↳ Sales, flight booking, user clicks on webpage)

Transaction has transaction ID, list of other items making transaction from transaction database, we can mine frequent patterns.

- Data cleaning :-
↳ To remove noise and inconsistent data ex:- parsing the data
↳ cleaning is performed for detection of syntax error.
↳ Parser decides weather the given string of data is acceptable within data specification.
- Data Integration:-
where multiple data sources are combined.
- Data Selection:-
where data relevant to the analysis task are retrieved from the database.
- Data transformation:-
where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operation for instance.
- Data mining:-
An essential process where intelligent methods are applied in order to extract data pattern.
- Pattern Evaluation:-
To identify the truely interesting patterns representing knowledge base on some interestingness measures.

• Knowledge Representation -
where visualization & knowledge representation technique are used to present the mined knowledge to the user.

ATTRIBUTE TYPES :- Data Mining (NOIR Classification)

→ Nominal :-
- Related to name
- The values are symbols or name of things.
- Represent category, code or state, also called categorial attribute
- Eg:→ hair colour.

→ Ordinal :-
- Represent a meaningful order or ranking.
- Not Quantify (magnitude can't calculate)
- Eg:→ ~~Asst Prof~~ → Associate Prof → Prof

→ Binary :-
- Only 2 states or categories are :- '0' or '1'
- Boolean (True or false)
- 1 (present) / 0 (absent)
- Types:-
  (i) Symmetric :- (equal weightage)
  → Gender (M/F)
  (ii) ~~Assymmetric~~ :- (Unequality)
  → HIV Test (+/-)

→ Numeric :
- Quantitative
- Measurable
- Represented by Integer or Real values.
- Types:-

(i) Interval scaled :-
→ measured on equal sized unit (linear)
→ have order & can be +'ve or -'ve
→ Allow us to compare & Quantify the difference two values
→ Eg:- Temperature

(ii) Ratio scale :-
→ Continuous +'ve
→ measurement on a non-linear scale
→ Add & Sub are allowed.
→ Multiply & Division are not allowed.

NOIR Classification :-
Scale is used to represent usually quantitively, an item's, person's or an event's place in the Spectrum

OR

Scale is defined as the device or an object used to measure or quantify any event or an object.

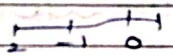
Types of Scales

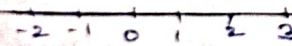
1. Nominal Scale :-
* Nominal Scale is used to categorize data into mutually exclusive categories or groups.
* A nominal scale is the 1st level of measurement scale in which the numbers, serve as "tags" or "labels" to classify or identify the objects.
* The data can be placed into categories but can't be multiplied, divided, added or subtracted from one another
* Ex:- gender, country, player's T-shirt number etc
M=1, F=2

Qualitative (Nominal / ordinal)

* We can't count or order the event (only Identify)
[Labels are assigned cannot be counted
Adhar Number, Phone Number]

2. Ordinal Data :-
* Data is ordered from least to most
* Un-determined numerical distance b/w each category (Ist prof, IInd prof, prof)
(1st place, 2nd place, 3rd place)
* Ordinal scale is used to measure variables in a natural order, such as rating or ranking
Totally satisfied, Satisfied, Neutral, Dissatisfied.

Interval Scaled Data :- (Interval Identify)
- A scale with values
- Same numerical distance b/w each value.
Eg 40°C, 39°C, 38°C
Rating 1, 2, 3
- Has an arbitrary zero point 
0°C, we have temp.
- Interval scale contains properties of nominal & ordered data, but the difference b/w data points can be quantified.
- This type of data shows both the order of the variables & the exact difference b/w the variables.
For eg :- time,

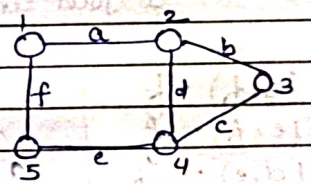


Ratio Scale :- (No Negative Value)
- A scale with zeroes
- Ratio scales of measurement include properties from all four scales of measurement.
- The data is nominal and identity, can be classified in order, contains intervals & can be broken down into exact value.
- Data in the ratio scale can be added, subtracted, divided & multiplied.
- Because of the existence of true zero value, the ratio scale doesn't have negative values
- Example :- Weight, height, distance are all eg. of ratio variables.
- True meaningful zero point (eg :- income, city distance)

Graph Theory

Application :→ Communication, Computation, Bio, chemistry

$G = (V, E)$ V = Vertex (Node) $\{v_1, v_2, v_3 \dots v_n\}$
E = Edges (line) $\{E_1, E_2, E_3 \dots E_n\}$
in which $l_k \in E$ is $l_k = \{v_i, v_j\}$
is called undirected graph



$V = \{1, 2, 3, 4, 5\}$
$E = \{a, b, c, d, e, f\}$

$|V|$ = Order of graph
↳ Total no of vertex in graph
$|E|$ = Size of graph
↳ Total no. of edges in graph.

- Adjacent Vertex — Common edge
- Adjacent edge — that connect to adjacent Vertex.
- Self loop — ∅ उस Vertex से start हो K वहीं पर ही खत्म हो ना।
- Multiedge — //
- Pseudograph — graph in which there is self loop & Multiedge
- Multigraph — No self loop but presence of multiedges
- Simple Graph — No self loop & No multiedges.

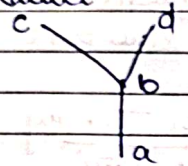
Lattice :- A partially ordered set in which every pair of elements has both a least upper bound and a greatest lower bound is called Lattice.

JSL + MSL

- **Join Semi-Lattice :-(V)**

If in a POSET for every pair of element LEAST UPPER BOUND exists, then the POSET is called Join-Semi Lattice

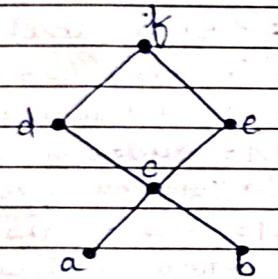
Eg:-



$(a,b) \to b$
$(b,c) \to c$
$(b,d) \to d$
$(c,d) \to X$

This structure is not a Join semi Lattice

Eg:-

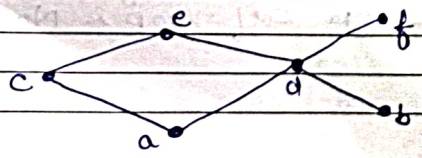


$(d,f) = f$
$(e,f) = f$
$(d,e) = f$
$(c,d) = d$
$(c,e) = e$
$(a,b) = c$
$(a,c) = c$
$(b,c) = c$

$(a,f) = f$
$(a,d) = d$
$(b,c) = e$

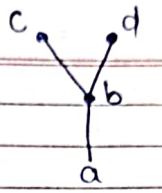
- **Meet Semi-Lattice :(∧) :-**

If in a POSET, for every pair of element GREATEST LOWER BOUND is present, then the POSET is called meet semi Lattice



$(a,b) \to X$
(Not GLB)

eg:-

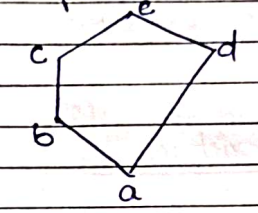


$(a,b) = a$
$(b,d) = b$
$(c,b) = b$
$(c,a) = a$
$(d,a) = a$

$(c,d) = b$

Examples :-

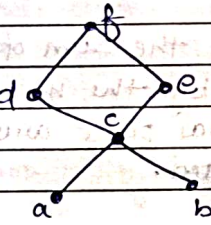
1.



$(a,e) \vee e$
$(a,e) \wedge a$
$(b,d) \vee e$
$(b,d) \wedge a$

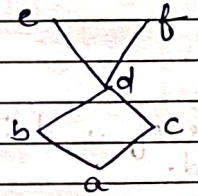
This a Lattice

2.



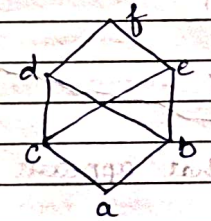
$(a \wedge b) = X$ Not Lattice

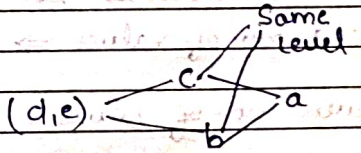
3.



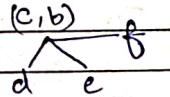
$(e \vee f) \to X$ Not Lattice

4.

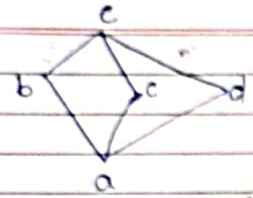




we can't find greatest lower bound because (c,b) are at same level.

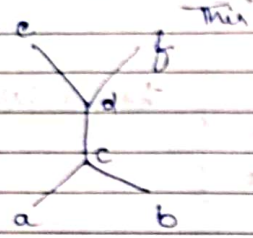


Again both d & e are at same level, so we can't find least upper bound.
So this is not lattice

5. 

$(a,c) \vee c$
$(a,c) \wedge a$
$(a,c) \rightarrow c$ or a
$(b,d) \rightarrow c$ or a

This is lattice

6. 

$(c \vee f) \times$
$(a \wedge b) \times$

Mean :- The mean or average is the sum of all values in a dataset divided by the no. of values.
- It represents the central point around which the data is distributed.

Ex:- • Salary Package in Lakhs
2, 3, 4, 5, 8, 9, 32

Mean $\Rightarrow \frac{2+3+4+5+8+9+32}{7} =$

Median :- It is the middle value of a dataset when it is ordered.

If odd no. of values $\Rightarrow \frac{No.+1}{2}$

If even no. of values $= \frac{N_i + N_j}{2}$

Mode :- The mode is the value that appears most frequently in a dataset

Ex:- 2, 3, 4, 3, 5, 7, 3, 8, 30
Mode = 3

Ex:- 2, 3, 4, 5, 7, 8, 30
Mode = 0

Ex Real time :- E-commerce

Range :- It is a measure of the spread or dispersion of a dataset.
- It is defined as the difference b/w the maximum & minimum values in a set of observation.

Ex:- 2, 3, 4, 5, 8, 9, 32
↑ Min ↑ Max

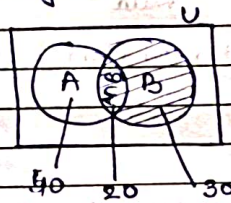
Max - Min = 32 - 2 = (30)

Conditional Probability

$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$ Event A & B.

↳ Probability of event A given that event B is already occur.

Ex:-



$P(B) = \frac{30}{100} = 0.2$

$P(A \cap B) = \frac{20}{100} = 0.2$

$P(A/B) = \frac{0.2}{0.3} = 0.67$

Q:- In a group of 100 Computer buyers, 40 bought CPU, 30 purchased monitor, and 20 purchased CPU and monitors. If a Computer buyer chose at random and bought a CPU, what is the probability they also bought a Monitor?

Ans:- A = bought computer = 40
B = purchased monitor = 30

$P(A) = \frac{40}{100} = 0.4$ $P(B) = \frac{30}{100} = 0.3$

A∩B = Purchased both CPU & monitor = 20

$P(A \cap B) = \frac{20}{100} = 0.2$

We known

$$P(B/A) = \frac{P(A \cap B)}{P(B)} = \frac{0.2}{0.4} = \frac{2}{4}$$

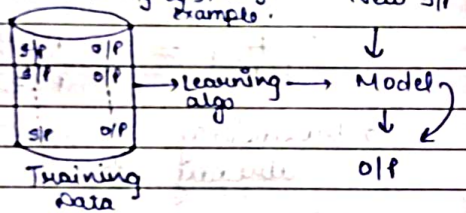
$$= 0.5$$

The probability that a buyer bought a monitor, given that they purchased a CPU is 50%.

SUPERVISED LEARNING: It like teaching a computer how to do something by showing it examples.

- Training data
- Both Input & output
- classification
- Naive Bayes algo. (learning algo)

Ex:- Weather forecasting



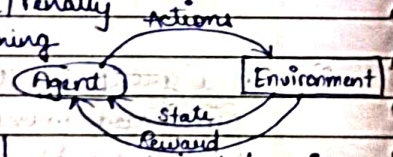
Unsupervised Learning:- (No knowledge)

- only I/P
- clustering

Algo → K-Mean

Reinforcement Learning:-

- Reward / Penalty
- Q-Learning



(Supervised + Unsupervised)

Supervised Learning

- Classification
  - Logistic Regression
  - Naive Bayes
  - Linear Discriminant Analysis (LDA)
  - Decision Tree
  - Random Forest
  - Support Vector Machines (SVM)
- Regression
  - Linear Regression
  - Ridge Regression
  - Lasso Regression

Unsupervised Learning

- PCA
- K-mean clustering
- Hierarchical clustering
- DB. Scan clustering.

Supervised Learning

- → K-Nearest Neighbours (KNN)
- → Gradient Boosting algorithms.
- → Neutral N/Ws.

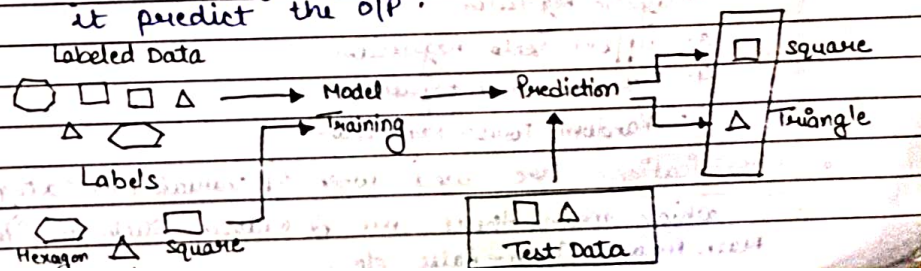
Types of Learning

Supervised learning	Unsupervised Learning	Reinforcement Learning
• Classification	• Dimensionality Reduction	• Gaming
Fraud detection	Text Mining	• Finance Sector
Email Spam Detection	Face Recognition	• Manufacturing
Diagnostics	Big Data Visualization	• Inventory Management
Image classification	Image Recognition	• Robot Navigation
• Regression	• Clustering	
Risk Assessment	Biology	
Score Predictions	City Planning	
	Targetting Marketing	

1. **SUPERVISED LEARNING:-** • Type of machine learning in which machines are trained using well 'labelled' training data, and on basis of that data, machines predict the O/P.
   - The 'Labelled data' means some I/P data is already tagged with the correct O/P.
   - It is a process of providing I/P data as well as correct O/P data to the machine learning model.
   - AIM:- To find a mapping function to map the I/P variable (x) with the O/P variable (y). $Y = F(x)$

Working of Supervised Learning algorithm:-

- Models are trained using labelled dataset, where the model learns about each type of data.
- Once the training process is completed, the model is tested on the basis of test data (a subset of training set), & then it predict the O/P.



- Steps of Supervised learning Algorithm:-

1. First Determine the type of Training dataset
2. Collect/Gather the Labelled Training data
3. Split the Training dataset into training dataset, test dataset, Validation dataset
4. Determine the I/P features of the training dataset, which should have enough knowledge so that the model can accurately predict the O/P.
5. Determine the suitable algorithm for the model, such as Support Vector machine, decision tree etc.
6. Execute the algo. on training dataset
7. Evaluate the accuracy of model by providing the test set.
8. If the model predicts the correct O/P, which means our model is accurate.

Types of Supervised Learning Algo.

- Regression
  - Ex:- what is the temp going to be tomorrow?
- Classification
  - Ex:- will it be cold or hot tomorrow?

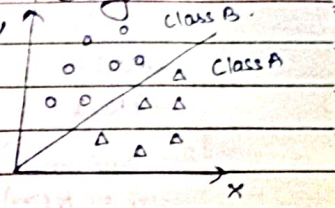
- Regression:- are used if there is a relationship b/w I/P variable and the O/P variable.
  - used for prediction of continuous & real variable or to predict a numerical value based on I/P features.
  - Ex:- Weather Forecasting, Market Trends
    Temp, Age, Salary, Price.

Regression Analysis Algorithm:-
1. Linear Regression
2. Logistic Regression
3. Support Vector Regression
4. Decision Tree Regression
5. Random Forest Regression

- Classification:- are used when O/P variable is categorical, which means there are 2 classes such as Yes-No, Male-Female, True-False etc.

It goals is to assign I/P data to predefined categories or classes.

Ex:- 1. Determining an email is spam or not
2. The outcome can be a binary classification/ logistic regression
3. Something is harmful or not
4. Rainfall tomorrow or not

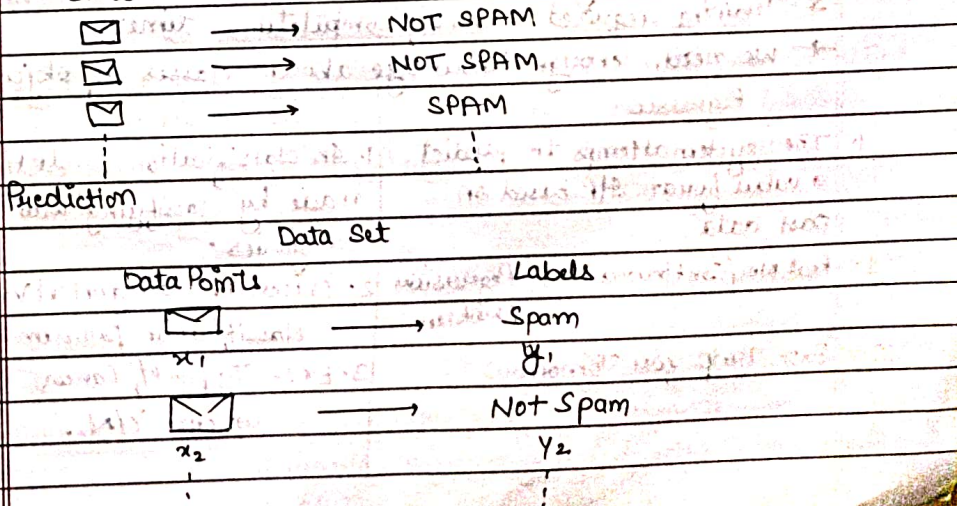


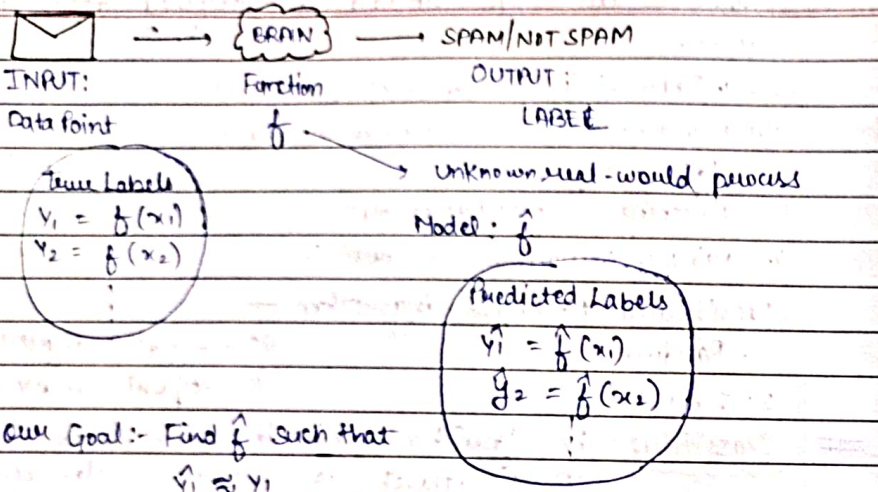
Classification Analysis Algorithm:-
1. Random Forest
2. Decision Trees
3. Logistic Regression
4. Support Vector Machines.

Classifier:- The algorithm which implements the classification on a dataset is known as classifier.

Types of classifications

* Binary Classifier:- If the classification problem has only 2 possible outcomes, then it is called as Binary Classifier (Divided into +ve & -ve Class)
  - Ex:- Yes or No, SPAM or NOT SPAM, CAT or DOG etc.
* Multi-class classifier:- If a classification problem has more than 2 outcomes, then it is called as Multi-class Classifier. Ex:- Classification of types of crops.
  Classification of types of music.





Our Goal:- Find $\hat{f}$ such that

$\hat{Y}_1 \approx Y_1$

$\hat{Y}_2 \approx Y_2$

⋮

• ADVANTAGES:- * with the help of supervised learning, the model can predict the o/p on basis of prior experiences.
* we can have an exact idea about the classes of object
* It helps us to solve various real-world problems such as fraud detection, spam filtering etc.

• DISADVANTAGES:- * are not suitable for handling the complex task
* cannot predict the correct o/p if test data is diff from the training dataset.
* Training required a lots of computations times
* we need, enough knowledge about classes of object.

	Regression	Classification
1.	The system attempts to predict a value for an o/p based on past data	1. In classification, predictions are made by classifying the data into diff. categories.
2.	Real No./Continuous No – Regression problem	2. Discrete / Categorical Variable – classification problem.
3.	Ex:- Temp. for Tomorrow	3. Ex:- Type of Cancer, Cancer Y/N.

	Supervised Learning	Unsupervised Learning.
1.	These are trained using labelled data	1. These are trained using unlabelled data.
2.	In this, I/P data is provided to the model along with the o/p	2. In this, only I/P data is provided to the model.
3.	It predicts the o/p	3. Finds the hidden pattern in data
4.	It needs supervision to train the model	4. It does not require any supervision to train the model.
5.	It produces an accurate result.	5. It gives less accurate result.
6.	It is less complex and less time consuming/fast	6. It is more complex and relatively slower.
7.	The goal is to identify relationship b/w the I/P and o/p variables and then use it to map new unlabeled data	7. The goal of this is to find pattern/meaningful pattern in data based on the relationship b/w data point themselves.
8.	It is used for price predictions, weather forcasting	8. Used for customer personas, recommendation engines.

BOOLEAN FUNCTIONS:- A function $f: A^n \to A$ is called Boolean function if it can be specified by a boolean expression of 'n' variables.

Let $A = \{x_1, x_2, \dots x_n\}$ & 3 operations $+, \cdot, '$ these $\vee, \wedge$

$(\underbrace{x_1 + x_2' \cdot x_3}_{\text{Boolean expression of 3 variables}})' + x_2 \cdot x_1$ OR $x_1 \vee (x_2' \wedge x_3) \vee (x_2 \wedge x_1)$ → Boolean expression in 2 variables.

* Boolean Algebra:- is a branch of algebra in which value of variable are truth values T/F usually

Truth Tables:

x_1	$\bar{x}_1$
0	1
1	0

OR

x_1	x_2	$x_1 \vee x_2$
0	0	0
0	1	1
1	0	1
1	1	1

AND

x_1	x_2	$x_1 \wedge x_2$
0	0	0
0	1	0
1	0	0
1	1	1

Q:- Show that Boolean Functions $f_1 = (x_1 \vee x_2) \vee x_3$ and $f_2 = x_1 \vee (x_2 \vee x_3)$ are equivalent

Variables → x_1, x_2, x_3 → 3 $\quad 2^n = 2^3 = 8$

x_1	x_2	x_3	$x_1 \vee x_2$	$x_2 \vee x_3$	f_1	f_2
0	0	0	0	0	0	0
0	0	1	0	1	1	1
0	1	0	1	1	1	1
0	1	1	1	1	1	1
1	0	0	1	0	1	1
1	0	1	1	1	1	1
1	1	0	1	1	1	1
1	1	1	1	1	1	1

$F_1 \approx f_2$

Types of Algebraic Expressions:-

1. Expressions that contains only one term is called Monomials
   Ex:- $4x^2, 3xy, -7z, 5xy^2, 10y$
2. Expressions that contains or 2 terms is called Binomial
   Ex:- $a+b, 4l+5m, a+4, 5-3y$
3. Expressions that contains 3 terms is called Trinomial.
   Ex:- $a+b+c, 2x+3y-5, x^2y - xy^2 + y^2$
4. In general, an expression containing one or more terms with non-zero co-efficients is called a polynomial.
   Ex:- $a+b+c+d, 3xy, 7xyz-10, 2x+3y+7z.$

Data Mining

ASSOCIATION RULE:- Knowledge Discovery in DB (KDD)

↳ The Association Rule is a learning technique that helps identify the dependencies b/w 2 data items

Items → A → B C
Relation on how they associate

↳ Associations rules are simple if/Then statements, if 'A' then 'B' {A ⇒ B}

antecedent → Consequent
A → B

↳ Two Parameters which is used to finding relationship b/w items. • Support • Confidence.

ASSOCIATON RULE MINING:-

↳ Also called as Market Basket Analysis (MBA) and Affinity Analysis

↳ Set of Items in a transaction is called Market Basket

↳ Mostly used in Retail.

• SUPPORT:- (S) Percentage (%) of transactions (T) that Contains both 'A' & 'B'. measures frequency
("A" ⇒ "B") = P(A∩B) {→ of association

• CONFIDENCE: (C) In a transaction set (T) if 'c' is the % of times 'B' is present in all the transactions containing 'A'.

Conditional Probability:- $C = P(B|A) = \frac{P(A \cup B)}{P(A)}$ {strength of association

→ PARAMETERS:-

• Finding all items that appears frequently in transaction. {min^m Support count}

• finding strong associations among frequent items (confidence)

→ Functions of Association Rules:-

① finding Set of items that has significant impact on business.

② Collecting information from numerous transactions

③ Generating rules from count in transactions

→ Problems in ARM:-

* Levels of frequency of appearance determination
* Finding strong associations among frequent items.

→ Functions of ARM:-

* Finding set of items that has significant impact on business.
* Collating information from numerous transactions
* Generating rules from counts in transactions

→ Strength of ARN

1. Easy Interpretation
2. Easy to start
3. Flexible data formats
4. Simplicity

→ Weakness

1. Exponential Growth in Computations
2. Lumping
3. Rule selection
4. Rare items not applicable for

APRIORI ALGORITHM:-

- Idea is to Generate candidate itemsets of a given size, and then scan dataset to check if their counts are really large.
- The process is Iterative

Steps 1. All singleton itemsets are candidates in the first pass. Any items with less than specified support value is eliminated.
2. Two member candidate itemsets.
3. Then make 3 member candidate itemsets upto 'n' no. of memberss.
4. Frequent itemsets constitutes, set of frequent itemsets.
5. We, generate Association Rules which have confidence values greater than or equal to specified min confidence

The Apriori algorithm is a popular method used for mining frequent itemsets.

Min Support = 2

1. Eg:-

Tid	Items
1	2,3
2	1,3,5
3	1,2,4
4	2,3

Items		Support	
1	→	2	
2	→	3	
3	→	3	
4	→	1	→ eliminated
5	→	1	

Make itemset

	Support	
{1,2}	→ 1	→ eliminated.
{1,3}	→ 1	
{2,3}	→ 2	

{2,3} → 2

Real-life Ex:- Market Basket Analysis
A supermarket wants to understand the purchasing behaviour of its customers to optimize product placement & promotion.

Process → Data collection :- the supermarket collects transaction data, where each transaction is a list of items purchased together

→ Frequent Itemset :- It identifies set of items that frequently appear together in transactions.

Application :- * Product Placement
* Promotion
* Inventory Management

2. Eg:-

Transaction ID	Items Purchased
T_1	Milk, Bread, Butter
T_2	Bread, Butter, Jam
T_3	Milk, Bread, Butter, Jam
T_4	Milk, Bread
T_5	Butter, Jam.

1. Initialize the Candidate 1-itemsets C1

Items		support
Milk	→	3
Bread	→	4
Butter	→	4
Jam	→	3

2. Generate the Frequent 1-itemsets (L1)
Remove items that do not meet minimum Support threshold. Let minimum Support threshold is (3).

Item	Support
Milk	3
Bread	4
Butter	4
Jam	3

3. Generate the Candidate 2 Itemsets (C_2)

Itemset	Support
{Milk, Bread}	3
{Milk, Butter}	2 — elim.
{Milk, Jam}	1 — eli.
{Bread, Butter}	3
{Bread, Jam}	2 — elim.
{Butter, Jam}	3

4. Generate the freq. 2 itemset (L_2)

Itemset	Support
{Milk, Bread}	3
{Bread, Butter}	3
{Butter, Jam}	3

5. Generate the Candidate 3 - itemsets (C_3) from L_2

Itemset	Support
{Milk, Bread, Butter}	2
{Milk, Bread, Jam}	1
{Milk, Butter, Jam}	1
{Bread, Butter, Jam}	2

X.

6. Generate the freq. 3 itemsets (L_3)
Since none of 3-itemsets have a support count of at least 3, L_3 { }.

7. Terminate.

Ex: 3:- Min-Support :- 60%
Min-confidence :- 80%.

Support : $\frac{3}{\text{top}} \times \frac{5}{\text{No. of entities}} = 1$

T	Itemset
T_1	M, O, N, K, E, Y
T_2	O, O, N, K, E, Y
T_3	M, A, K, E
T_4	M, U, C, K, Y
T_5	C, O, O, K, E

(i) Now, find Support count of each Itemset

C_1 =

Itemset	Support
M	3
O	4
N	2 — eliminate
E	4
Y	3
D	1 — eliminate
A	1
U	1
C	2
K	5

Compare min support with each itemset Support count.

L_1 =

Itemset	Supp. Count
M	3
O	4
K	5
E	4
Y	3

(ii) Generate pair to generate C_2.

C_2 →

Itemset	Sup. Count
M, O	1 X
M, K	3
M, E	2 X
M, Y	2 X
O, K	3
O, E	3
O, Y	2 X
K, E	4
K, Y	3
E, Y	2 X

Now again compare C_2 with min-Sup.

Itemset	Sup. Count
M, K	3
O, K	3
O, E	3
K, E	4
K, Y	3

3. Make pair to generate C_3

C_3 =

Itemset	Sup. Count	
M, K, O	1	X
M, K, E	2	X
M, K, Y	2	X
O, K, E	3	
O, K, Y	2	X

Now again compare with min support

L_3 =

Itemset	Supp. Count
O, K, E	3

* Now create association rules with Support & Confidence for O, K, E

$\text{Confidence} = \frac{\text{Support}}{\text{No of time it occurs.}} \times 100$

Association Rule	Support	Confidence	Confidency →
O ∧ K → E	3	3/3 = 1	100
O ∧ E → K	3	3/3 = 1	100
K ∧ E → O	3	3/4 = 0.75	75 X
E → O ∧ K	3	3/4 = 0.75	75 X
K → O ∧ E	3	3/5 = 0.6	60 X
O → K ∧ E	3	3/5 = 0.75	75 X

Compare this with min confidence = 80%.

O ∧ K = E	3	100
O ∧ K = K	3	100

Hence final association rules are

O ∧ K = E
O ∧ E → K
} Market basket Analysis.

Ex 4:-

T_{ID}	Item Bought
T_1	{Bread, Butter, Milk}
T_2	{Bread, Butter}
T_3	{Beer, Cookies, Diapers}
T_4	{Milk, Diapers, Bread, Butter}
T_5	{Beer, Diapers}

For given dataset, apply Apriori algo. to discover stronger association rules among image tags.

Min support = 40%.

Min confidence = 70%.

Sol:- Support = $\frac{2}{5} \times 5 = 2$.

(i) Generate Frequent itemset

C_1 =

Itemset	Sup. Count
Bread	3
Butter	3
Milk	2
Beer	2
Cookies	1 → Eliminate
Diapers	3

L_1

Itemset	Sup. Count
Bread	3
Butter	3
Milk	2
Beer	2
Diapers	3

(ii) Generate pair to generate C_2

C_2

Itemset	Sup. Count	
Bread, Butter	3	
Bread, Diapers	1	X
Bread, Milk	2	
Bread, Beer	0	X
Butter, Diapers	1	X
Butter, Milk	2	
Butter, Beer	0	X
Diapers, Milk	1	X
Diapers, Beer	2	
Milk, Beer	0	X

L_2

Itemset	Sup. Count
Bread, Butter	3
Bread, Milk	2
Butter, Milk	2
Diapers, Beer	2

(iii) Generate pair to Generate C_3

C_3

Itemset	Support	
Bread, Butter, Milk	2	
Bread, Butter, Diapers	1	X
Bread, Butter, Beer	0	X
Bread, Milk, Diapers	1	X

Bread, Milk, Beer →	0	×
Bread, Diapers, Beer →	0	×
Butter, Milk, Diapers →	1	×
Butter, Milk, Beer →	0	×
Butter, Diapers, Beer →	0	×
Milk, Diapers, Beer →	0	×

So L3 =

Bread, Butter, Milk → 2

* Now create association Rule with Support & confidence for Butter, Bread, Milk.

Association Rule	Support	Confidence	confidency.
Bread ∧ Butter → Milk	2	2/3 = 0.66	66 ×
Bread ∧ Milk → Butter	2	2/2 = 1	100
Butter ∧ Milk → Bread	2	2/2 = 1	100
Milk → Bread ∧ Butter	2	2/2 = 1	100
Butter → Bread ∧ Milk	2	2/3 = 0.66	66 ×
Bread → Butter ∧ Milk	2	2/3 = 0.66	66 ×

Compare with min confidence = 70.

Bread ∧ Milk → Butter

Butter ∧ Milk → Bread

Milk → Bread ∧ Butter

Definition:- Imp

- Frequent itemset:- It is an itemset whose Support is greater than some user-specified minimum Support.
- Closed Frequent itemset:- An itemset is closed if none of its immediate Supersets has the same support as that of itemset.
- Maximal Frequent Itemset:- An itemset is maximal Frequent if none of its immediate Supersets is Frequent.
- Downward closure property of frequent pattern:- All Subset of any frequent itemset must also be Frequent.

Ex:- If milk, Bread, Butter is a Frequent itemset, then the following itemsets are frequent:-

(Milk), (Bread), (Butter), (Milk, Bread), (Milk, Butter), (Bread, Butter)

Need of Closed and Maximal itemmset :-

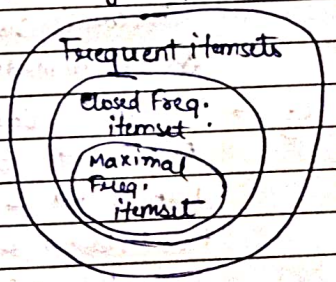
→ These closed Frequent itemset and maximal frequent itemsets are useful when huge amount of data is used in association rule mining.

→ If the length of frequent itemset is 'K', then by downward closure property all of it 2^K subsets are also frequent.

→ When the Computation is very expensive & there is no interest to find additional Subsets. This can be avoided by frequent itemset with max length.

→ One disadvantage with maximal frequent itemsets is that even all its subsets are frequent, we do not know their Suppo for mining rule, support information is very important.

→ ∴ closed freq. itemset is preferred.



Q:- Min Support count = 3

find Frequent, closed, Maximal Itemset.

Tid	List of items
T_1	A, B, C, D
T_2	A, B, C, D
T_3	A, B, C
T_4	B, C, D
T_5	C, D

All items that is A, B, C, D are Frequent because they sup. count is greater than or eq. to minimum support count.

C_1 :→

Item	Count
A →	3
B →	4
C →	5
D →	4

L_2

Item	Count
A, B →	3
A, C →	3
A, D →	2 ×
B, C →	4
B, D →	3
C, D →	4

All 2 items sets are Freq. except AD.

A(3) → AB(3), AC(3), AD(2)

A(Count) is not greater than its immediate Superset.

∴ A is not closed.

In A's immediate Superset, itemset are present with min. Support Count ie. 3.

A is not maximal.

↳ B(4) → AB(3), BC(4), BD(3)

B(Count) is not greater than Im. Superset. B is not closed.

In B's immediate superset, itemset are present with min Support Count i.e 3. ∴ B is not maximal.

↳ C(5) → AC(3), BC(4), CD(2)

C(Count) is greater than immediate Superset. C is closed.

In C's immediate Superset, itemset are present with min. Support Count i.e 3 ∴ C is not maximal.

↳ D(4) → AD(2), BD(3), CD(4)

D(Count) is not greater _____ ∴ D is not closed.

In D's immediate Superset _____ ∴ D is not maximal.

C_3 →

Item	Count
A, B, C	3
A, B, D	2
A, C, D	2
B, C, D	3

A
↳ AB(2) → ABC(3), ABD(2)
AB is not closed.
AB is not maximal.
↳ AC(3) → ABC(3), ACD(2)
AC is not closed.
AC is not maximal.
↳ AD(2) → ABD(2), ADC(2)
AD not frequent.

B
↳ BC(4) → ABC(3), BCD(3)
BC(Count) is greater ∴ BC is closed
BC is not maximal.
↳ BD(3) → ABD(2), BCD(3)
∴ B is not closed
∴ B is not maximal.

C
↳ CD(4) → ACD(2), BCD(3)
CD is closed
CD is not maximal.

C_4

item	Count
A, B, C, D	2

ABCD is not frequent

ABC(3) → ABCD(2)
∴ ABC is closed
In ABC's immediate superset, itemset are not frequent

min. support Count ie. 3,

ABC is maximal.

ABD(2) → Not Freq. ACD(2) → Not Frequent.

↳ BCD(3) → ABCD(2)

BCD is closed ABCD(2) → not freq.

BCD is maximal.

1-itemset

Item	Freq.	Closed	Maximal
A(3)	✓	×	×
B(4)	✓	×	×
C(5)	✓	✓	×
D(4)	✓	×	×

2-itemset

Item	Freq.	closed	Max.
AB(3)	✓	×	×
AC(3)	✓	×	×
AD(2)	×	×	×
BC(4)	✓	✓	×
BD(3)	✓	×	×
CD(4)	✓	✓	×

3-itemset

Item	Freq.	closed	Maximal
ABC(3)	✓	✓	✓
ABD(2)	×	×	×
ACD(2)	×	×	×
BCD(4)	✓	✓	✓

4 itemset

item	Freq.	closed	Max.
ABCD(2)	×	×	×

Q:- Suppose min" support = 2

Items	support
1 →	2
2 →	3
3 →	3

itemset	support	
{1,2}	1	×
{1,3}	1	×
{2,3}	2	✓

Ans {2,3} = 2

Q:- For the following given Transaction Data-set. Generate Rules using Apriori Algorithm.
Consider the Values as. Support - 50% & Confidence = 75%.

Tid	item Purchased
T_1	Bread, cheese, egg, juice
T_2	Bread, cheese, juice
T_3	Bread, milk, Yogurt

T_4 Bread, Juice, Milk

T_5 Cheese, Juice, Milk.

Freq. Item set

Items		frequency	Support
Bread	→	4	4/5 = 80%
Cheese	→	3	3/5 = 60%
Egg	→	1	1/5 = 20% × eliminate
Juice	→	4	4/5 = 80%
Milk	→	3	3/5 = 60%
Yogurt	→	1	1/5 = 20% × eliminate

$$\text{Support (Bread)} = \frac{n_{\text{Bread}}}{m} = \frac{4}{5} = 0.8 = 80\%$$

* Make 2-Items Candidate set & write their Frequency.

Item Pairs		Freq.	Support
Bread, Cheese	→	2	2/5 = 40% ×
Bread, Juice	→	3	3/5 = 60%
Bread, Milk	→	2	2/5 = 40% ×
Cheese, Juice	→	3	3/5 = 60%
Cheese, Milk	→	1	1/5 = 20% ×
Juice, Milk	→	2	2/5 = 40% ×

Support ≥ 50%

Bread, Juice → 3 (60%)

Cheese, Juice → 3 (60%)

1. Bread, Juice

(Bread → Juice) (Juice → Bread)

Formula → confidence $(A \to B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$

• Bread → Juice $= \frac{\text{Support}(\text{Bread} \cup \text{Juice})}{\text{support}(\text{Bread})} = \frac{3.5}{5.4} = \frac{3}{4} = 75\%$

• Juice → Bread $= \frac{\text{Supp.}(J \cup B)}{\text{Supp.}(J)} = \frac{3.5}{5.4} = 75.1\%$

2. (Cheese → Juice) $\frac{3.5}{5.3} = 100\%$ OR Juice → cheese $\frac{3}{5} \cdot \frac{5}{4} = 75\%$

All Rules are Good.

Frequent Itemsets:

• An itemset X is frequent only if support (X) >= min. sup.

$$\text{Support}(X) = \frac{\text{No. of transactions contains } X.}{\text{Total no. of transactions.}}$$

Transaction ID		items
T_1	→	{A, C, D}
T_2	→	{B, C, E}
T_3	→	{A, B, C, E}
T_4	→	{B, E}
T_5	→	{A, B, C, E}

$$\text{Support (A)} = \frac{\text{No. of tra. contain A}}{\text{Total no. of tran.}} = \frac{3}{5}$$

$$\text{Support (D)} = \frac{\text{No. of trans. contain D}}{\text{Total no. of tran}} = \frac{1}{5}$$

let minsup = 2/5

Itemset	Support	F/I
A	3/5	F
B	4/5	F
C	4/5	F
D	1/5	I
E	4/5	F
AB	2/5	F
AC	3/5	F
AD	1/5	I
AE	2/5	F
BC	3/5	F
BD	0/5	I
BE	4/5	F
CD	1/5	I
CE	3/5	F
DE	0/5	I
ABC	2/5	F

Itemset	Support	F/I
ABD	0/5	I
ABE	2/5	F
ACD	1/5	I
ACE	2/5	F
BCD	0/5	I
BCE	3/5	F
BDE	0/5	I
CDE	0/5	I
ABCD	0/5	I
ABCDE	2/5	F
ABDE	0/5	I
ACDE	0/5	I
ADE	0/5	I
BCDE	0/5	I
ABCDE	0/5	I

• let minsup = 2/5
• Total itemsets = 31
• Frequent Itemset = 15
• Infrequent Itemsets = 16.
• $2^5 - 1$

Apriori Principle:-

→ If an Itemset is Infrequent then all its Supersets are Infrequent

→ If an itemset is frequent then all its Subsets are frequent.

- CLOSED frequent itemset:
  An itemset X is closed frequent itemset only if X is frequent & No immediate superset of X has same support as X.
- Maximal frequent itemset:-
  An itemset X is maximal itemset only if X is frequent & No immediate superset of X is freq.

Itemset	Support	F/I	C/NC	M/NM
A	5/5	F	NC	NM
B	4/5	F	NC	NM
C	4/5	F	C	NM
D	1/5	F	NC	NM
E	4/5	F	NC	NM
AB	2/5	F	NC	NM
AC	3/5	F	C	NM
AE	2/5	F	NC	NM
BC	3/5	F	NC	NM
BE	4/5	F	C	NM
CE	3/5	F	NC	NM
ABC	2/5	F	NC	NM
ABE	2/5	F	NC	NM
ACE	2/5	F	NC	NM
BCE	3/5	F	C	NM
ABCE	2/5	F	C	M

F = 15
C = 5
M = 1

∴ M C C C F

Database

dst. min sup = 2/4.

Tid	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

↳ sup (1,3) = 2/4
↳ sup (1,2,3) = 1/4 ↳ sup(2,3,5).
↳ sup(1,3,4) = 1/4
↳ sup(1,3,5) = 1/4
↳ sup(2,3,5) = 2/4
↳ sup(1,2,3,5) = 1/4
↳ sup (2,3,4,5) = 1/4

∴ So {2,3,5} is a Maximal freq. Itemset

Classification & Prediction:-

- Classification technique is a systematic approach to build classification models from an input data-set.
- that is used to identify class-label.
- The i/p dataset is known as training dataset.
- In classification we arrange new data with the help of current or past data.
- There are 2 steps in classification:-
  * Model construction
  * Model usage. → class [Label → fail, Pass]

Students	Marks	Result
A	5	Fail
B	6	Pass
C	2	fail
D	7	Pass
E	9	Pass

⇒ Classification Algorithm (Learn, Analyze, generate model.)

↳ If marks ≤ 5
then Result = Fail → Fail
else Result = Pass → Pass.
Classification model or classifier.

- whenever data is given to this model, based on the rule it will display class label.
- eg:- Decision tree, Neural N/W, Bayesian network etc.

Prediction :- It is a process of identifying the missing or unavailable numerical data for a new obj.

↳ Algorithm which use training data-set to derive a model, that model is predictor. When the new data is given, this model should find a numerical o/p.

↳ Unlike in classification, this model doesnot have the class label, this model predicts a continuous valued functions or ordered value.

Issues regarding classification & Prediction.

* Preparing the data for classification & Prediction
* Performance evaluation measures.

1. Preparing the data for classification & Prediction :-

Initially, data should be pre-processed to improve the accuracy, efficiency, scalability of the classification or prediction process.

→ Pre-Processing techniques :-

1. Data cleaning :- • Removal of noise data
   • Handling of missing values.
2. Relevance Analysis :- • Removing redundant attributes using Correlation Analysis.
   • Removing irrelevant attributes using Attribute Subset selection.
3. Data transformation and reduction :-
   • Normalization :- Scaling of values for a given attributes so that they fall with in small specified Range.
   It can be applied on numeric data only.
   Ex :- min-max, z-score.
   • Generalization :- Generalizating the data to higher conceptual levels. Particularly it can be useful for continuous valued attributes.
   Ex :- attribute "income" can be generalized as : low, medium, high

2. Performance Evaluation measures :-

1. Accuracy :- of a classifier refers to the ability of a given classifier to correctly predict the class label of unseen data.
   • Accuracy of predictor refers to how well a given predictor can guess the value of the predicted attributes for unseen data.
   • Accuracy can be estimated using the Test set that is independent of training set.
2. Speed :- refers to the computational cost involved in generating & using the given classifier or predictor.
3. Robustness :- The ability of classifier or predictor to make correct predictions given noisy data or data with missing values.
4. Scalability refers to ability to construct the classifier or predictor efficiently given large amount of data.
5. Interpretability :- Refers to level of understanding and insight that is provided by classifier or the predictor.

Classification & Prediction issue :-

The major issue is preparing the data for classification & Prediction. Preparing the data involves the following activities :-

1. Data Cleaning :- It involves removing the noise and treatment of missing values. The noise is removed by applying Smoothing techniques and the problem of missing values is solved by replacing a missing values with common occuring value
2. Relevance Analysis :- DB may also have the irrelevant attributes Correlation analysis is used to know whether any 2 given attributes are related.
3. Data Transformation & Reduction :- The data can be transformed by any of the following methods.
   • Normalization :- It involves scaling all values for given attributes in order to make them fall within a small specified range. It is used when in the learning step, the neutral N/ws or the methods involving measurements are used.
   • Generalization :- Generalizing data to higher concept. for this purpose we can use the concept hierarchies.

Instance-Based Learning :- (Also Memory based Learning / Lazy Learning)

↳ Memorising and then apply.

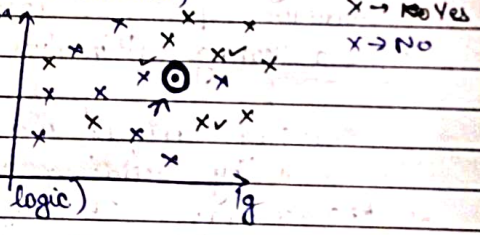
→ instead of performing explicit generalisation, it compares new problems with instance in training, which are stored in memory. eg:- Spam mail.

- Just memorizing but doesn't know what (Rote Learning)
- Not learning with labels.

⇒ when new instance comes in it need to identify its class which delayed the processing time (delay classification in run-time)

IB Approach
- Lazy learners (K-NN)
- Radial Based Function (for weighted or nearest one)
- CBR (Case Based Reasoning (we take previous cases extract knowledge & execute).

eg.	Cgpa	placement
	80	Yes (✓)
	70	No (✗)
	103	7.5



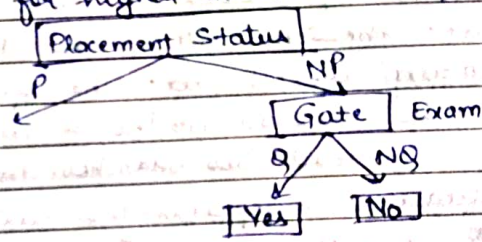
Steps :→ (Its K-NN algorithm logic)
- Place the pt in the graph
- Trying to find similarity with its surrounding (neighbours)
- Try finding distance of that pt with all other neighbours (for instance consider only 3).
- suppose {yes, yes, no}.
- We see 2 pts are for yes & 1 for no.
- So we consider that pt also to be a yes (Placement done).

Note:- Before the new pt the algo. did nothing just Compare already stored data & give similar thing o/p. There is not concept of training & learning.

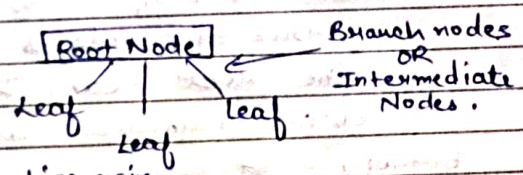
Decision Tree :-

↳ used for both classification & Regression task.
↳ It is a classifier (Tree structure)

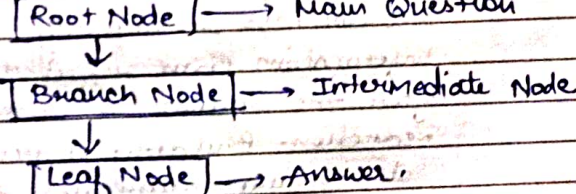
Eg:→ Choose for higher studies or Not?



↳ Tree Structure
- ↳ Decision Node
- ↳ Leaf Node
- ↳ Splitting

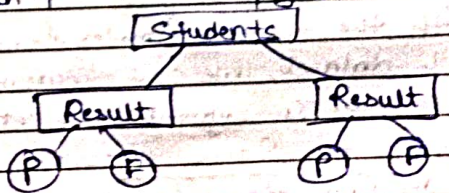


2 factors :→
- Entropy & Information gain
- Pruning



- Entropy :- Measure of how much uncertainity there is in the information
- Info Gain : Measure of how much information the answer to a specific question provides.

⇒ Info Gain ↑ = Entropy ↓



→ Decision Tree is a supervised learning method used in data mining for classification & regression that help us in

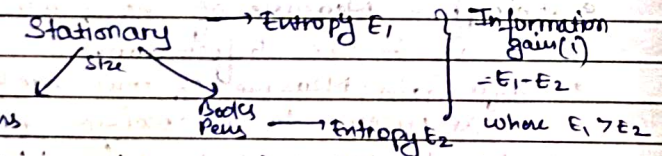
decision-making purposes.

The decision tree creates classification or regression models as a tree structure.

It separates a data set into smaller subsets, & at the same time, the decision tree is steadily developed. The final tree is a tree with the decision nodes & leaf nodes. A decision node has at least two branches. The leaf nodes show a classification or decision. We can't accomplish more split on leaf nodes. The uppermost decision node in a tree that relates to the best predictor called the root node. Decision trees can deal with both categorical and numerical data.

Entropy:- It refers to a common way to measure impurity. In the decision tree, it measures the randomness or impurity in data sets.

Information Gain refers to the decline in entropy after the dataset is split. It is also called Entropy Reduction. Building a decision tree is all about discovering attributes that return the highest data gain.

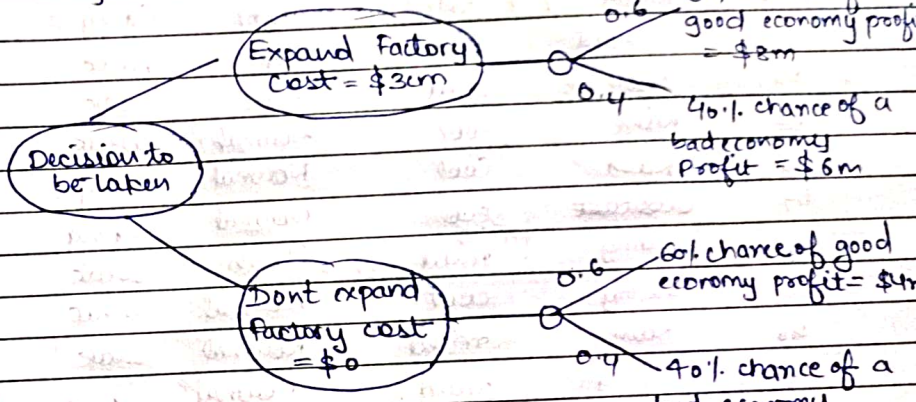


In short, a decision tree is just like a flow chart diagram with the terminal nodes showing decisions. Starting with dataset, we can measure the entropy to find a way to segment the set until the data belongs to the same class.

→ Why are decision trees useful?

It enables us to analyze the possible consequences of a decision throughly.

- It provides us a framework to measure the values of outcomes & the probability of accomplishing them.
- It helps us to make the best decision based on existing data & best speculations.



Expanding factor costs $3 million, the probability of a good economy is 0.6 (60%), which leads to $8 million profit, & the probability of a bad economy is 0.4 (40%), which leads to $6 million profit. Not expanding factor with 0$ cost, the probability of a good economy is 0.6 (60%), which leads to $4 million profit, and the probability of a bad economy is 0.4, which leads to $2 million profit. The management teams need to take a data-driven decision to expand or not based on the given data.

Bayesian classification:-

↳ Bayesian classifiers are statistical classifiers.

↳ They can predict the probabilities of class items (like it gives the probability that a given data item belongs to the class / not)

→ Conditional Probability

→ Bayes theorem

→ Naive Bayes classifier

Naive Bayes classification :- (Probability based)

attributes → Target variable

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D_1	Sunny	Hot	high	Weak	No
D_2	Sunny	Hot	high	Strong	No
D_3	Overcast	Hot	high	Weak	Yes
D_4	Rain	mild	high	weak	Yes
D_5	Rain	cool	Normal	weak	Yes
D_6	Rain	Cool	Normal	Strong	No
D_7	Overcast	cool	Normal	Strong	Yes
D_8	Sunny	mild	High	weak	No
D_9	Sunny	cool	Normal	weak	Yes
D_{10}	Rain	mild	Normal	weak	Yes
D_{11}	Sunny	mild	Normal	Strong	Yes
D_{12}	Overcast	mild	High	Strong	yes
D_{13}	Overcast	hot	Normal	weak	Yes
D_{14}	Rain	mild	High	Strong	No

Given New Instance with Certain condition :—
(outlook = Sunny, temp = cool, Humidity = high, wind = Strong)

we have to check whether the new instance can be classified as yes or No?

Prior Probability

Sol $P(\text{Play Tennis} = yes) = 9/14 = 0.64$

$P(\text{Play Tennis} = No) = 5/14 = 0.36$

now, calculate conditional probability of individual current attributes.

Attribute	Possible outcome	
outlook	Y	N
Sunny	2/9	3/5
Overcast	4/9	0
Rain	3/9	2/5

Temprature	Y	N
	Possible outcome	
hot	2/9	2/5
mild	4/9	2/5
Cool	3/9	1/5

Humidity	Y	N
High	3/9	4/5
Normal	6/9	1/5
Wind	Y	N
weak	6/9	2/5
Strong	3/9	3/5

$$V_{NB}(No) = \frac{V_{NB}(No)}{V_{NB}(yes) + V_{NB}(No)}$$

$$= \frac{0.0206}{0.0259} = 0.795$$

V_{NB} = avg max $P(V_j)$ $\prod_i P(a_i/v_j)$

V_{NB} ↓ $v_j \in \{yes, no\}$ ↓ ↓

Naive Bayes = avg max outcome $P(V_j)$ $P(\text{Sunny}/V_j)$

$v_j \in \{yes, no\}$ $P(\text{Cool}/V_j)$ $P(\text{High}/V_j)$

$P(\text{Strong}/V_j)$

$V_{NB}(yes) = P(yes)\, P(\text{Sunny}/yes)\, P(\text{Cool}/yes)\, P(\text{High}/yes)\, P(\text{Strong}/yes)$

$= 9/14 \times 2/9 \times 3/9 \times 3/9 \times 3/9 = 0.0053$

- $V_{NB}(NO) = P(NO)\,P(\text{sunny}|NO)\,P(\text{cool}|NO)\,P(\text{High}|NO)\,P(\text{Strong}|NO)$

$\Rightarrow 5/14 \times 3/5 \times 1/5 \times 4/5 \times 3/5 = \underset{\text{large}}{.0206}$

- $V_{NB}(\text{yes}) = \dfrac{V_{NB}(\text{yes})}{V_{NB}(\text{yes}) + V_{NB}(NO)}$

$= \dfrac{0.0053}{0.0053 + 0.206} = .205.$

- $V_{NB}(NO) = \dfrac{V_{NB}(NO)}{V_{NB}(\text{yes}) + V_{NB}(NO)} = \dfrac{0.0206}{0.053 + 0.0206}$

$\Rightarrow \underset{\text{large}}{.796}$

* Data Analytics
* Data Engineering
* Data Scientist
* Data Visualisation (software used for data visualisation)
* KDD.

Plan it with No filter

Goal

Checklist
(index)

Question and Answer
Q. _____ ?

a) Me
b) Myself

New Mail
To : My future self
CC : Success
Subject : it's okay to fail but don't give up

Love Future Me

Reminder

Good Vies

Keep Shining

do what you love