DATA MINING

⇒ Data Mining is defined as procedure of extracting info. from huge sets of data.
- also defined as mining knowledge from data.

What types of data can be mined ?

⇒ 3 types → database, datawarehouse, transactional.

1. Database data (RDBMS):

→ set of tables - has rows and columns
↓ ↓
tuples attributes.

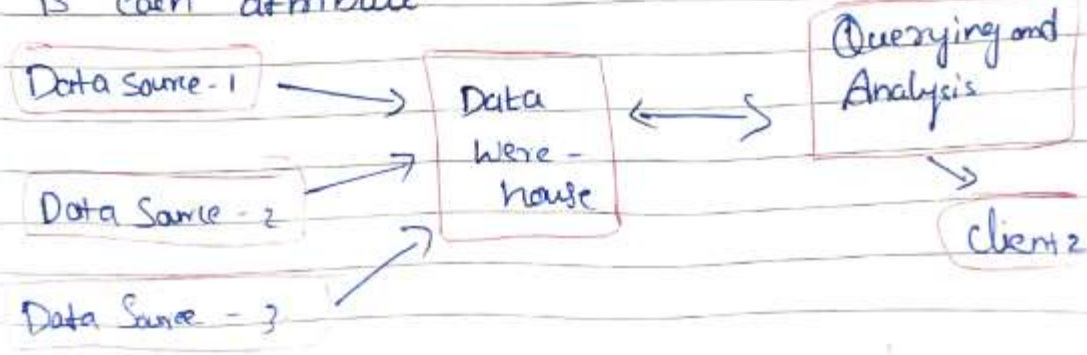
While mining databases, we can search for trends or data patterns.

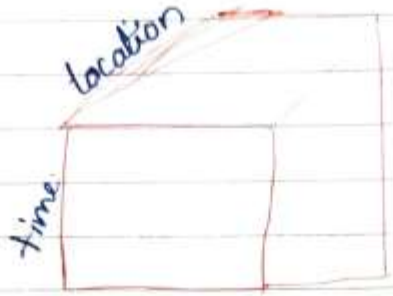
Example: 1. Analysing customer data to predict the credit risks of new customers.
2. Analysing sales data - (any deviations)

2. Datawarehouse data:

collections of data (integrations) integrated from different sources with querying and decision making on data.

In datawarehouse, data is stored in multidimensional structure (data cube) where each dimension is each attribute.





1 unit

item type (pen, pencil)

3 Transactional database :

Each record is called as transaction.
(sales, flight booking, user ↓ clicks on web page

Transaction has transaction ID, list of other items. Making transform transaction db, we can mine frequent patterns.

* other types of data : → cont
→ Maps
→ IC's Sequence data, data streams, Spatial data, Engineering design data, hypertext, Multimedia, Web data etc.

* DATA MINING FUNCTIONALITIES

- (5) functionalities

1 Concept / classif definitions:

data is always associated with class / concepts

description can be done in (2) ways

⇒ Data characterisation:

refers to the Summary of the class

o/p → General overview.

⇒ Data discrimination:

— compares the common features of the class

o/p → barcharts, Curves etc,

2.* mining frequent patterns, associations & correlation

frequent Patterns:

Things which are found most commonly in data

→ frequent itemsets
→ frequent Subsequence
→ frequent substructure

Association Analysis:

It is a way of identifying the relation blw various item.

Example: used to determine sales of items that are frequently purchased together.

Correlation Analysis:

- mathematical technique

- shows how strongly pair of attributes are related together.

Example: Tall people tend to have more weight.

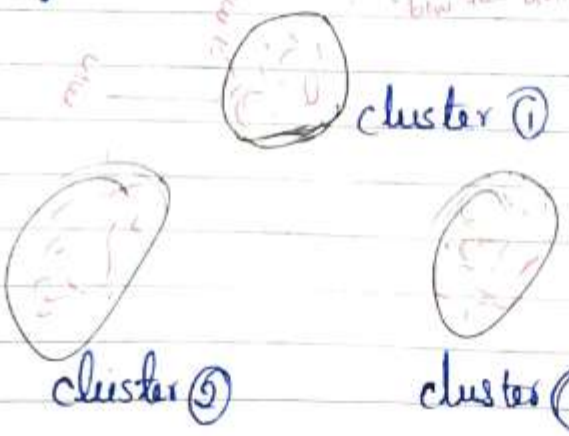
3. classification and regression for predictive Analysis
   ↓
   prediction of data.

classification:
Process of finding a model that distinguishes data items * decision tree is used for classification.

Regression: statistical Methodology that is used for numeric prediction (done based on previous data)

4. **cluster Analysis:**

The data items are clustered based on the principle of maxmising the intraclass Similarity and minimising the interclass Similarity.



Analysis → cluster Analysis.

5. Outlier Analysis: (Anomaly mining)

Among the data items in a db, there may be some items which do not follow the general behaviour of data. Those data items → outliers (noise / exceptions)

* Interestingness of patterns:

In a data mining system, everyday million of data patterns are generated.

Among all these patterns generated, how many are really interesting?

* Actually, a small fraction of patterns generated would be of interest to any given user.

This raises (3) questions:

1. What makes patterns interesting:

A pattern is interesting if it is
- easily (understand) understood by humans.
- valid on new/test data
- potentially useful.

2. Can data mining System generate all of the interesting patterns?
- refers to completeness of a dm System.

In reality it is not possible for a dm system to generate all interesting patterns.

3. Can dm Systems generate only interestingness-pattern?
- refers to optimization of a dm System

→ generating only interesting patterns ⇒ challenging

If only interesting patterns are generated, it becomes easy and efficient for the user (time is saved).

Video 10

* Classification of DATA MINING SYSTEM

Why classification - ?

DM → everywhere and anywhere

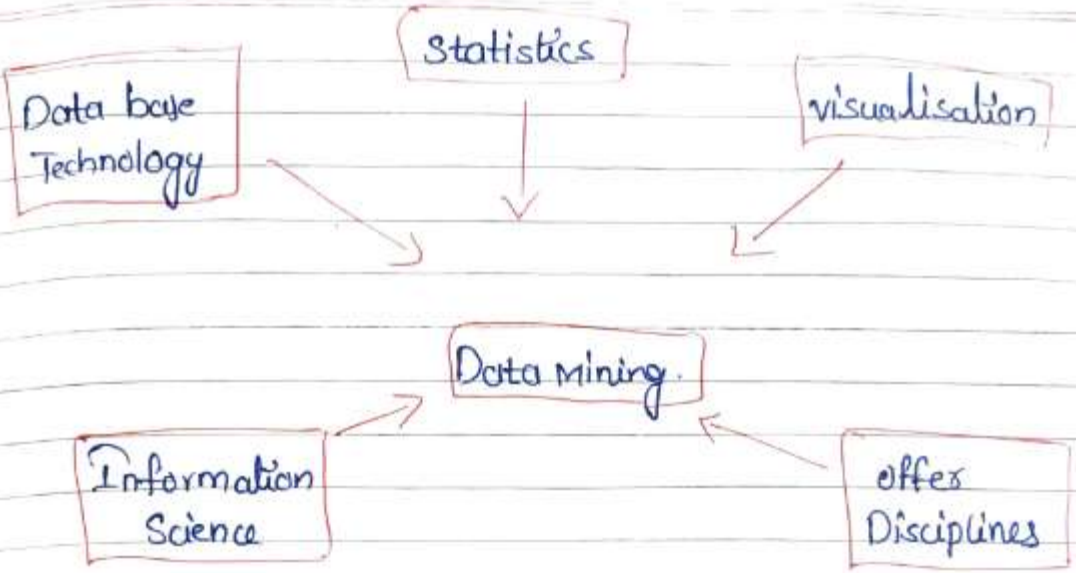
Data mining systems are classified based on several criteria.

1. classification based on mined database:
   based on type of database that is been mined
   - Relational
   - Transactional
   - Object - Relational
   - Data Warehouse

2. classification based on type of knowledge mined.
   - characterization
   - Discrimination.
   - Association and correlation analysis
   - classification.
   - prediction
   - outlier Analysis
   - Evolution Analysis.

3. classification based on kinds of techniques used:
   - ML, statistics, neural network, pattern recognition, data Warehouse oriented techniques etc.

4. classification based on applications adapted
   - finance
   - Telecommunications
   - DNA
   - Stock Market
   - Email etc.



* DATA MINING TASK PRIMITIVES:

A data mining task is represented inform of datamining query. is defined as interms of dm task-primitive

Wi allow the user to interactively communicate with the dm System.

-(5) dm-task primitives:

1. set of task relevant data to be mined.
2. Specifies the kind of knowledge to be mined
3. The background knowledge to be used in discovery process.
4. The interstingness measures and thresholds for pattern Evaluation
5. The Expected representation for visualizing the discovered patterns

* INTEGRATION OF DATAMINING SYSTEM WITH A DATABASE / DATA WAREHOUSE SYSTEM:

Integration → association / combining

If There is no integration - no communication with db.

- We have a total of 4 integration schemes

1. NO Coupling:
DM system will not use any function.

i.e there is no communication with db
for this, it comm. with other storage methods.

2. Loose Coupling

- will use some of the functionalities (only upto some extent)
- better than no coupling.
- Suitable for small data sets.

3. Sometight Coupling:
- linked to the db.
- also, some of the dm primitives are also implented in db.

4. Tight Coupling.
- dm system is completely linked to db
- most efficient among all.

The db system is fully integrated in such a way that it becomes part of the dm system.
- Efficient and optimised implimentation of dm.

* MAJOR ISSUES IN DATA MINING.

1. mining different kinds of knowledge in database.
2. Interactive mining of knowledge of multiple levels of abstraction.
3. Incorporation of background knowledge
4. presentation & visualization of data mining results
5. Handling noisy / incomplete data.
6. Efficiency and scalability of data mining alg.

(V8)

* DATA PREPROCESSING:

The process of transforming raw data into an understandable format.

- ④ major tasks

1. Data cleaning
2. Data integration
3. Data Reduction
4. Data Transformation

* Data cleaning:

process of Removal of incorrect, incomplete, inaccurate data, also replaces Missing data.

1. Handling of missing values:

In place of missing values, we can replace with "NA", with mean values. with median values.

- Sometimes replaced with most probable values.
- missing values can be filled in ② ways

manual — Small

automatic — more efficient - large data set

2 Handling noisy data.

noisy data → inconsistent / error data.

methods to handle (steps)

1. Binning

first, data is sorted. Then sorted data is stored in bins

- ③ methods to handle data in bins
  - Smoothing by bin mean
  - Smoothing by bin median
  - Smoothing by bin boundary

2. Regression:
Numerical prediction of data.

3. Clustering
Similar data items are grouped at one place
dissimilar items - outside the cluster.

* Data integration:

Multiple heterogeneous sources of data are combined into single dataset.

- ② types of data integration

1. Tight coupling:
Data is combined together into a physical location.
2. Loose Coupling.
only an interface is created and data is combined through the i/f and also accessed through i/f.
- Data remains in actual database only.

* Data Reduction:
volume of data is reduced to make analysis easier.

Methods for data Reduction:

1. Dimensionality Reduction
reduces no of i/p variables in the dataset.
bcz, large i/p variable → poor performance

2. Data cube Aggregation
Data is combined to construct a data cube.

(Redundant, noisy Data is removed)

3. Attribute Subset section

Highly Relevant attributes should be used.

Others → discarded (removed)
(∴ Data is Reduced)

4. Numerosity Reduction:

Here, we store only model of data instead of entire data Sample.

* Data Transformation:
Data is transformed into appropriate form Suitable for mining process.

(A) methods

1. Normalisation
Done in order to scale the data values in Specified Range
(-1.0, to 1.0 or from 0 to 1)

2. Attribute selection
New attributes are created using older ones.

3. Discretization
raw values are replaced by interval levels

4. Concept hierarchy Generation
attributes are converted from low level to high level
Ex: City → Country.

UNIT - II

* FREQUENT PATTERNS

the patterns that appear frequently in dataset
↓
(include frequent data items, sequences, substructures)

Example: Milk and bread.

Market Basket Analysis:
process of Analysing customer buying habits by finding the associations b/w the dif. items that a customer will place in their baskets.
- mainly useful for sellers.

Strategies Used:

1. placing them together.
2. placing them at (2) different ends.

- This Analysis will help Sellers to plan their shelf space for increased Sales.
- frequent patterns are represented by assiociation Rule

Ex: Computer and anti-virus.

Support:
identifies how frequently a rule is applied to given dataset.

$$S(P \rightarrow Q) = \frac{(P \cup Q)}{N} \quad (\because N = \text{Total Transactions})$$

$$P(A \cup B)$$

Confidence

defined frequent occurance of items of A in transactions of

$$C(P \rightarrow Q) = p(B/A)$$

* MINING METHODS

- Apriori Algorithm
- Fp Growth Algorithm

Apriori Algorithm

- by R. Agarwal and R. Srikant.
- Shows how objects are associated with each other

objective: To generate an association.

Example:

minimum Support = 50%.
Threshold confidence = 70%.

T/D	items
100	(1) (3) (4)
200	(2) 3 (5)
300	1 2 3
400	2 5

Itemset	Support	miniSupport
1	2	2/4 = 50%.
2	3	3/4 = 75%.
3	3	3/4 = 75%. (X)
4	1	1/4 = 25%.
5	3	3/4 = 75%.

Itemset: (1, 2, 3, 5)

— form pairs

(1,2) (1,3) (1,5) (2,3) (2,5) (3,5)

itemset	Support	minimum Support
(1,2)	1	1/4 = 25%. (x)
(1,3)	2	2/4 = 50%.
(1,5)	1	1/4 = 25%. (x)
(2,3)	2	2/4 = 50%.

itemset = (1,3) (2,3) (2,5) (3,5)

- form triplets

(1,2,3) (1,2,5) (1,3,5) (2,3,5)

itemset	Support	minimum Support.
(1,2,3)	1	1/4 = 25%.
(1,2,5)	1	1/4 = 20%.
(1,3,5)	1	1/4 = 25%.
(2,3,5)	2	2/4 = 50%.

Itemset = (2,3,5)

- now that lets calculate Support and confidence

Confidence = Support (AUB) / Support of A

Using (2,3,5) we can generate association Rules

Rules	Support	confidence.
(2^3) → 5	2	2/2 = 100%.
(3^5) → 2	2	2/2 = 100%.
(2^5) → 3	2	2/3 = 66%. (x)
2 → (3^5)	2	2/3 = 66%. (x)
5 → (2^3)	2	2/3 = 66%. (x)
3 → (2^5)	2	2/3 = 66%. (x)

$(2 \wedge 3) \rightarrow 5$. confidence $= \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$ 2,3,5

$\frac{S((2 \wedge 3) \cup 5)}{S(2 \wedge 3)} = \frac{2}{2} = 100\%$.

$2 \rightarrow (3 \wedge 5) : \frac{S(2 \cup (3 \wedge 5))}{S(2)} = 2/3 = 66\%$.

$\therefore (2 \wedge 3) \rightarrow 5$, $(3 \wedge 5) \rightarrow 2$ are association rules

* Fp GROWTH ALGORITHM

Fp → frequent pattern

- is an efficient and scalable method for mining the complete set of fp using a tree structure for storing information about fp called fp tree.

Example:

minimum Support = 30%.

Trans id	items
1	E, A, D, B
2	D, A, E, C, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, B, E
8	B, C

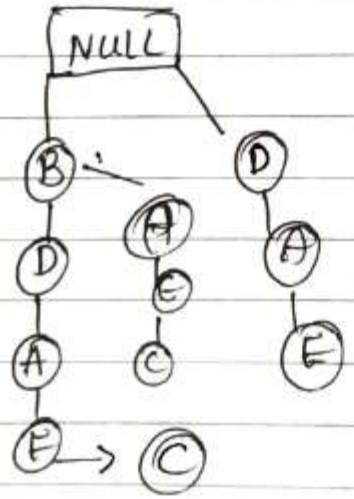
write priorities
more frequency → more priority
same frequency → FCFS

- list out the priorities

itemset	frequency	priority
A	5	→ 3
B	(6)	→ 1
C	3	→ 5
D	(6)	→ 2
E	4	→ 4

- order items according to priority.

Trans ID	Items	ordered Items
1	E A D B	B D A E
2	D A E C B	B D A E C
3	C A B E	B A E C
4	B A D	B D A
5	D	D
6	D B	B D
7	A D E	D A E
8	B C	B C



B - 1, 2, 3
D - 1, 2
A - 1, 2
E - 1, 2
C - 1
A - 1
E - 1
C - 1
D - 1, 2
A - 1

* MINING VARIOUS KINDS OF Association Rules

- ④ types

1. mining Multilevel association Rules
- ~~mining~~ Uniform Support for all levels
  - Using reduced minimum Support at lower levels.
  - using item or group based minimum Support

2. mining Multidimensional association rules from Relational database or data warehouse.

3. mining Multi dimensional association Rules using static discretisation of Quantitave attributes

4. mining quantitative association Rules

* CORRELATION ANALYSIS

used to measure the relationship blw ② variables.

$$r_{A,B} = \frac{\Sigma (A - A')(B - B')}{(n-1)\sigma A \sigma B}$$

$r_{A,B}$ = karle pearson correlation cofficient.

A', B' = mean of A and B

$\sigma A, \sigma B$ = Standard deviation of A and B.

n = no of tuples in db.

$r \rightarrow$ ③ values $(0, -1, +1)$

$r \rightarrow +1 \Rightarrow$ perfect positive correlation

$r \rightarrow 0 \Rightarrow$ NO correlation (no dependence)

$r \rightarrow -1 \Rightarrow$ perfect Negative correlation.

Example:

A	B
20	8
12	34
	4

$$r_{A,B} = \frac{\Sigma (A - A')(B - B')}{(n-1)\sigma A \sigma B}$$

$$A' = \frac{20+12+9}{3} = 13.66 \qquad B' = \frac{8+34+4}{3} = 15.33$$

$$\sigma_A = \sqrt{\frac{\Sigma (A-A')^2}{n-1}}$$

$$= \sqrt{\frac{(20-13.66)^2+(12-13.66)^2+(9-13.66)^2}{2}} = 5.68$$

$$\sigma_B = \sqrt{\frac{\Sigma (B-B')^2}{n-1}}$$

$$= \sqrt{\frac{(8-15.33)^2+(34-16.33)^2+(4-15.33)^2}{2}} = 16.28$$

$$r_{A,B} = \frac{(20-13.66)(8-15.33)+(12-13.66)(34-15.33)+(9-13.66)(4-15.33)}{2 \times 5.68 \times 16.28}$$

$= -1 \ldots$

≈ -1

i.e negative correlation

* CONSTRAINT BASED ASSOCIATION MINING:

Constraint - Condition

- association Rules are generated based on conditions

* Types of constraints:

1. Knowledge Type;

- Specifies the types of knowledge you want to do mining - association, correlation, Regression etc,

2. Data Constraints

- Specifies the type of data on which you want to generate the Rules.
  - only task relevant data

3. dimension level constraints.

Specifies the dimension or level concept hierarchy.

4. Interestingness Constraints

Support, confidence are used to Identify.

5. Rule Constraints

Specifies the form of rules to be mined

↓

(2) ways

1. metarules guided mining
2. constraint pushing

* GRAPH PATTERN MINING

set of tools techniques used to mine frequent Subgroups Subgraphs.

- used to Analyse the properties of real world graphs
- used to Analyse how structure of graph will effect the Rules

2 ways

1. Apriori based approaches
2. pattern growth approaches

Algorithms used:

1. Gspan → all types
2. closed Graph → closed Subgraphs.

Applications

1. in xml Structures
2. anomaly detection
3. Network Analysis
4. control flow Analysis
5. Biological Structures etc.

* SEQUENTIAL PATTERN MINING: (Spm)

Sequence = set of ordered events

Ex: $S = \{e_1, e_2, e_3, e_4, e_5\}$

Spm ⇒ process of finding frequent subsequences from a set of sequences

Sequences are represented by '< >'

Normal transaction data

CID	TID	Transactions
1	100	a, b, c, d
3	111	a, f, d, e
1	122	d, e, f
3	133	b, f, s, a
1	144	b, c, d, e
3	155	a, f, d, c
1	166	a, e, p

Sequential data:

CID	Sequences
1.	<(abcd), (dep), (bcde), (aep)>
3.	<(afde), (bfsa), (afdc)>

min-Sup=2

Challenges in Spm:

- finding all subsequences

Sid	Sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(dfcb)>
40	<eg(af)(bc>

min-Sup=2

<(ab)c> (✓)

<eg> (✗)

Algorithms used:

1. GSP (Generalized sequential patterns)
2. SPADE (vertical format based mining)
3. prefixspan
4. clospm - for closed patterns

Unit - III

* CLASSIFICATION AND PREDICTION

classification

finding a good model that is used to perfect predict the class of objects whose class label is unknown.

- Categorization of new data with the help of current / past data.

Example: Grouping of the patients based on their medical records

Prediction:

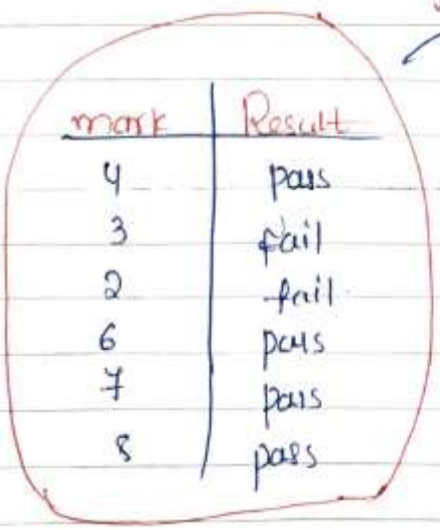
predicting a missing / unknown value based on past / current data.

- o/p is a continous value.

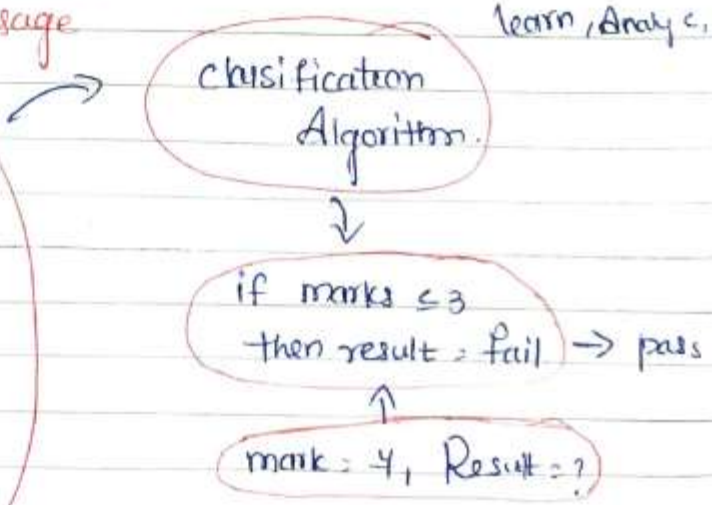
Example: predicting the correct treatment for a person based on their medical condition.

Classification: (2) steps.

1. model construction
2. model usage



Training data.



* DECISION TREE INDUCTION

- Flow chart like tree structure
- Supports in taking decisions

- it defines the rules visually in form of tree.

Types of Nodes.

1. Root Node - main Question
2. Branch Node - Intermediate process
3. Leaf Node — Answer.

* Attribute Selection measures.

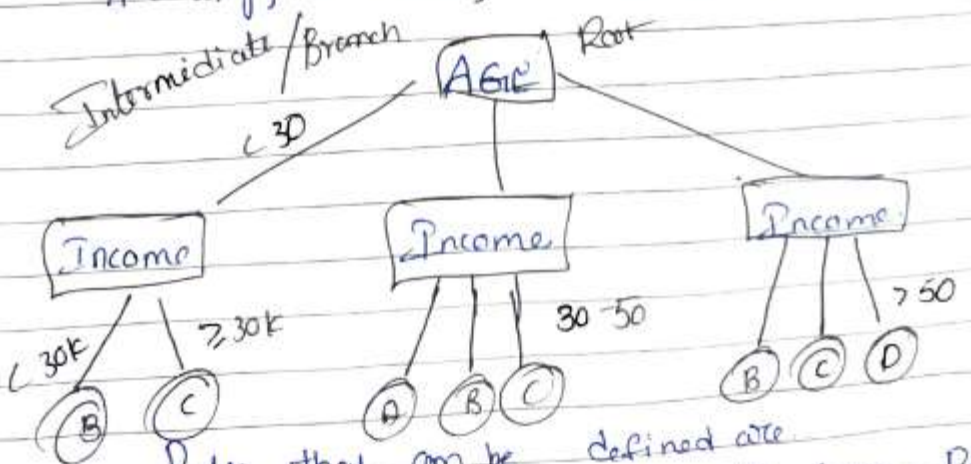
1. Information Gain.
   how much Information does the answer to the specific quetion provide.

2. Entropy:
   measures the amount of uncertainity in the info.

(As IG↑ Entropy↓)

Example: Credit Score Rating

A → Avg; B → Bad; C → Good; D → Excellent.



Rules that can be defined are

If age < 30, income < 30k then credit score = Bad
If age < 30, income ≥ 30k then credit score = Good

* BAYESIAN CLASSIFICATION

Bayesian classifiers are statistical classifiers

- They can predict the probabilities of class items

Like it gives the probability that a given class item belongs to that class/not)

- Bayes Theorem
- naive Bayes (Theorem) classifies

* RULE BASED CLASSIFICATION:

It uses set of IF THEN rules for classification

- 3 important keywords:

If, AND, THEN

If condition THEN conclusion

if part - rule antecedent pre condition

Then part - rule consquent

Example:

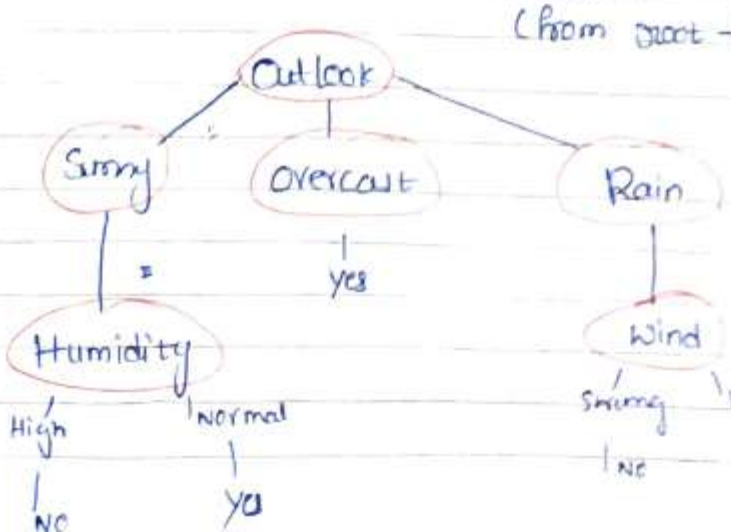
Rule: if age = 'youth AND student = 'yes' AND --
THEN buys computer = 'yes'

- Rule based classification - many ways.

Decision Tree = 1 way.

Extract Rules from decision tree

(From root to leaf nodes)



Rule 1: If outlook = 'sunny' AND humidity = 'high' THEN play = 'no'

Rule 2: If outlook = 'overcast' THEN play = 'yes'

:

many Rules.

* ~~LAZY~~ * LAZY LEARNERS:

learning from neighbours

- simply stores training data and wait until it gets a test tuple

i.e works only when it gets a new example.

- less training time
- more prediction time.

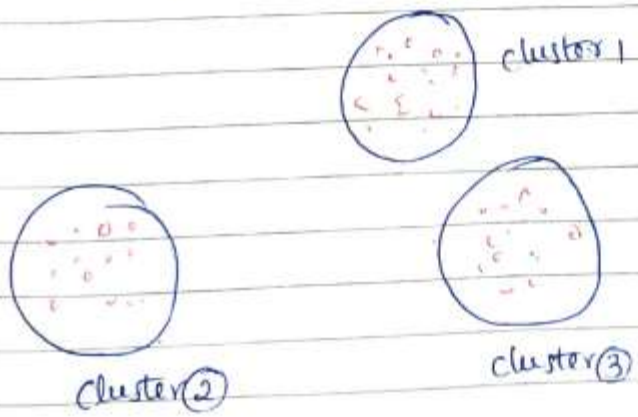
Example: KNN algorithm.

UNIT - IV

* CLUSTER ANALysis:

process of forming similar group of objects together in form of cluster.

- unsupervised ml algorithm.



properties of clusters

1 clustering Scalability - anysize of data.
2. Algorithm usability with multiple - types of data.
3 Dealing with unstructured data.
4. Interoperability.

clustering methods

1. partitioning method
2. Hierarchical method.
3. Density based method.
4. Grid based methods
5. model based methods
6. Constraint based method

} most common.

* TYPES OF DATA IN CLUSTER ANALYSIS

1. / Data Datastructures

1. Data matrix

Here data is represented by table or n by matrix

rows → real world entities (names)

columns → properties of these entities

2. Dissimilarity Matrix:

- represented as nxn matrix
- Identifies dissimilarities b/w (2) objects

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & 33 \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \cdots & 0 \end{bmatrix}$$

Types of data - (4)

1. Interval Scaled Variables.

- Continous Variables

Ex: 10-20, 20-30 etc

Individual data → convert into continous

- do the data Standardisation before that

Standardized data → Calculate mean absolute deviation then divide the data into intervals.

2 Binary Variables:
- has only (2) states <- 0 - nb, 1 - Pr

0 -> variable is absent
1 -> variable is present.

-2 -> Subtypes
1. Symmetric binary
2. Asymmetric binary

3. Categorial Variables

Data that can be divided into categories
-(2) types

1. Normal Variables.

has no particular order to its categories

Ex: Gender
/ \
male female
(can be in any order)

2. ordinal variables
has a particular internal order to its categories

Ex: Temperature
Low / Medium \ High
(should be in an order)

4. mixed variables:
combo of different types of variables

PARTITIONING METHODS

n data items / objects → k partitions
↓
represents a cluster

$k \leq n$

Partitions should satisfy ③ rules

1. each partition → atleast one object
2. Each object should belong to only 1 partition only

Example: k-means algorithm
- data is divided into clusters based on distance and centroid values.

	Height (x)	Weight (Y)
1)	185	72
2)	170	56
3)	168	60
4)	179	68
5)	182	72
6)	188	77
7)	180	71
8)	180	70
9)	183	84
10)	180	88
10)	180	67
11)	177	76

Now, remaining values should be divided based on Euclidian distance (ED)

$$ED = \sqrt{(x_0 - x_c)^2 + (y_0 - y_c)^2}$$

As ③, $k_1 = \sqrt{(168-185)^2 + (60-72)^2} = 20.8$

$$k_2 = \sqrt{(168-170)^2 + (60-56)^2} = 4.48$$

$k_2 < k_1 \therefore$ ③ ∈ cluster 2 (k_2)

Now calculate centroid for K_2

$$\left(\frac{170+168}{2}, \frac{56+60}{2}\right) = (169, 58)$$

$$K_2 = (169, 58)$$

for ④, $K_1 = \sqrt{(179-185)^2 + (68-72)^2} = 6.32$

$$K_2 = \sqrt{(179-169)^2 + (68-58)^2} = 14.14$$

$K_1 < K_2 \therefore$ ④ $\in K_1$

Now centroid for K_1

$$\left(\frac{179+185}{2}, \frac{68+72}{2}\right)$$

$$= (182, 70)$$

$K_1 = \{1, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

$K_2 = \{2, 3\}$

Now the data is divided into ② dif clusters.

* HIERARCHICAL CLUSTERING:

groups the data into free of clusters

↓

dendrogram

↓

has sequence of all merges & splits

2 methods

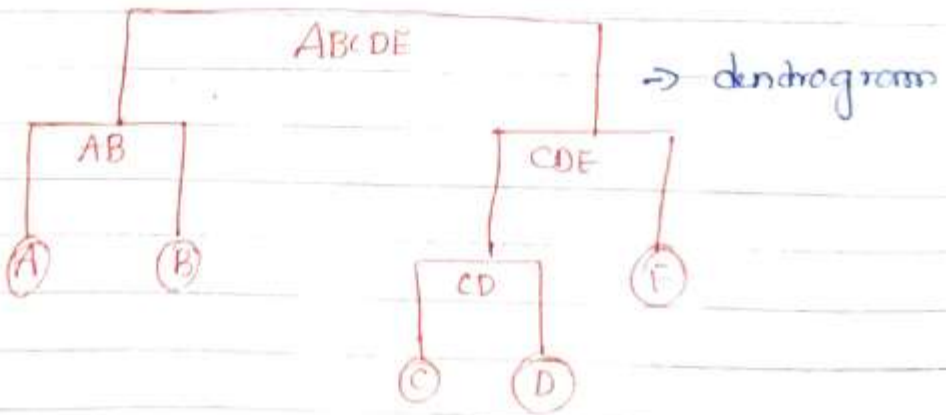
1. Agglomerative method
2. Divisive method

- Agglomerative method:

- bottom up method

Steps Involved

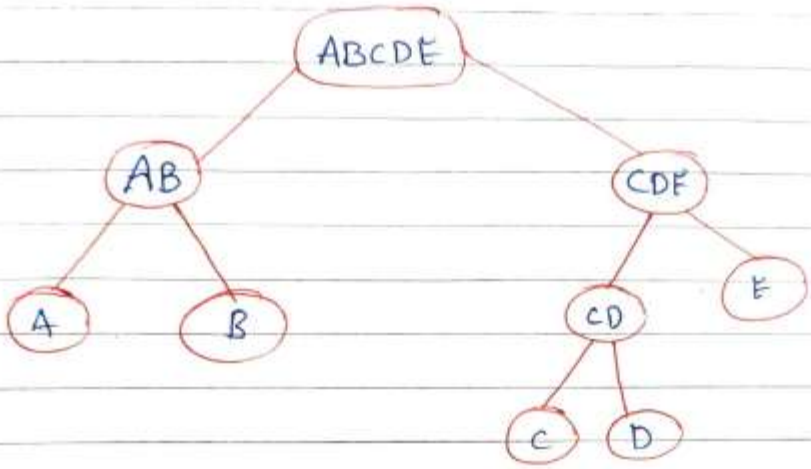
1. calculate the similarity of one cluster with respect to all other clusters
2. consider every data point as individual data.
3. merge the clusters with highest similarity
4. Recalculate similarity for each cluster
5. Repeat step ③ and ④ until single cluster is obtained



→ dendrogram

* 3 modes
  1. Single linkage
  2. Complete linkage
  3. Average linkage

- Divisive method
  Top- bottom method
  - we take all the data items into single cluster and in iteration, we split the data
    - at the end we get N cluster



* DENSITY BASED METHODS:

- data objects are clustered based on density
  Mass/ Volume

Ex: DBSCAN
(density based spatial clustering of application with noise)
- it has (2) inputs (ε and minpts())

ε - radius of circle formed with datapoint as center
minpts() - minimum no of data points inside the circle

- (3) types of data points

1. core point : it should satisfy the condition of minpts
2. boundary point : neighbour of core
3. noise point : not core nor boundary.

p, r → core
q, s → boundary
t → noise

* GRID BASED CLUSTERING:
uses a multi resolution grid data structure
it divides the object into finite no of cells that form a grid like structure
- then density is calue calculated for these cells.
- Sort the cells according to density.
- Identify cluster centers
- update neighbour cells

* Quick processing time

Ex: STING
(Statistical Information Grid clustering Algorithm)
- Spatial data is divided into rectangular is at different levels of resolution (these cells form a tree structure.
cells at higher level contains smaller cells compared to its lower levels.
clustering - done based on parameters
(mean, count, min, max, SD types of dist)
- Calculation of these parameters should start at root and go down till bottom layer.

* OUTLIER ANALYSIS:

outlier → among the data object, one which does not obey general behaviour.

Analysis of outliers — outlier Analysis

* Outlier Detection:
process of detecting outliers and subsequently removing them

methods - (2)

1. Statistical Approach:

based on probability of the data points.
low probability - outlier.
- parametric methods
- Non parametric methods.

2. Proximity Approach:

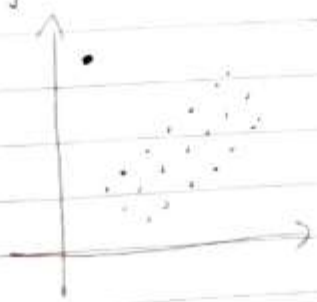
based on location data of points
- Density based approach
- Distance " "
- Grid " "
- Deviation " "

* Types of outlier -

(3) types of outlier

1. Global / point outliers

When a single data object deviations from the rest of data points --> Global / point outliers



3 Contextual / conditional outliers.

Data objects deviates from others because of any specific condition - called Contextual outliers