Unit-3

* Classification - It is the process of finding good model that describe the data class or concept and its purpose to predict the class of object whose label is unknown. (categorical value) - Two step — Model making
↳ model usage.

+ Prediction - Prediction is about predicting a missing/unknown element (continous value) of a dataset.

* Issues regarding classification and prediction

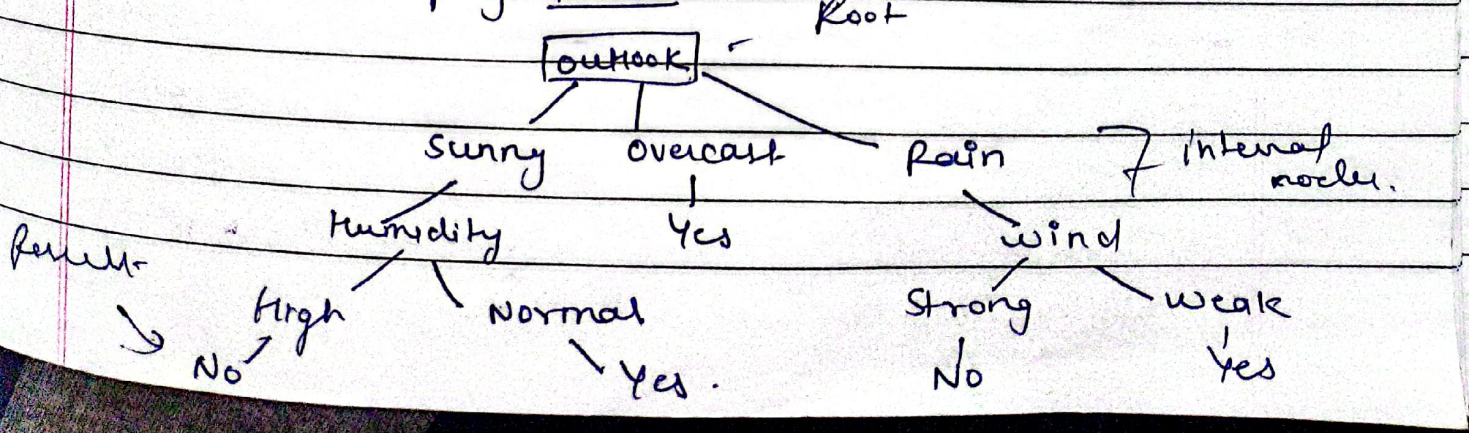
1) Data preparation — Data cleaning
Relevance analysis (feature extraction)
Data transformation

2) Evaluating classification method
- Predictive accuracy
- speed scalability
Robustness (handle noise)
goodness of rule
Interpretability (Easy understand)

* Decision tree classification - supervised machine learning technique used for classifying problem.

- It is a tree structured classifier
- internal node represent feature of dataset
- branch represent rule
- each leaf node represent the outcome.
- Contain root node where decision tree starts.

Decision tree to play tennis



- Bayesian classification - It is collection of classification algorithm based on bayes theorem. It uses bayes theorem to predict occurence of any event. These are statistical classifier with probability understanding.

Bayes's theorem is expressed as -

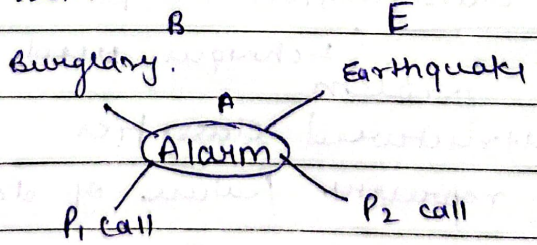
$$P(A/B) = \frac{P(B/A)\ P(A)}{P(B)}$$

P(A/B) conditional probability that of given that B has happened. It true.

P(A) & P(B) are probability of observing A & B independently

Assumptions - No features are dependent- i.e they are independent
- Each feature is give same importance.

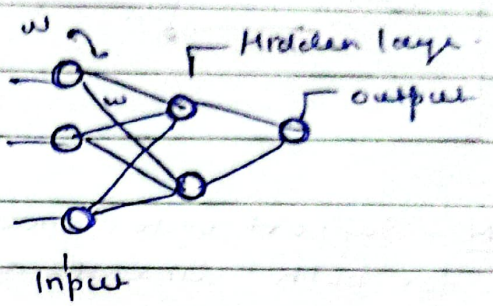
- Bayesian belief networks - It is graphical representation of different probabilistic relationships among random variables in a set.



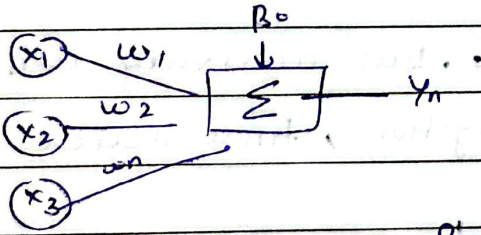
It is probabilistic graphical model which represent a set of variables and their conditional dependencies using a directed acyclic graph.

- Multilayer feed forward Neural Network - It is an interconnected artificial Neural Network with multiple layer that has neuron with weights associated with them and they compute result using activation function. In this flow is from input to output, do not contain loop, feedback, no signal move in backward direction

- Input layer
- Hidden layer
- Output layer.



- Back propagation algorithm - It is an algorithm that back propagates the error from output nodes to the input nodes. Simply called back propagation of error



Steps - Initialise weights → Propagate input forward → Back propagate error → terminating condition.
weights are modified to minimise the error.

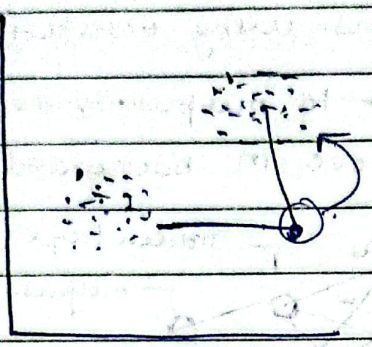
Adv - High tolerance to noisy data, classify untrained patterns

Disadv - long training
- Poor interpretability
- large parameter required

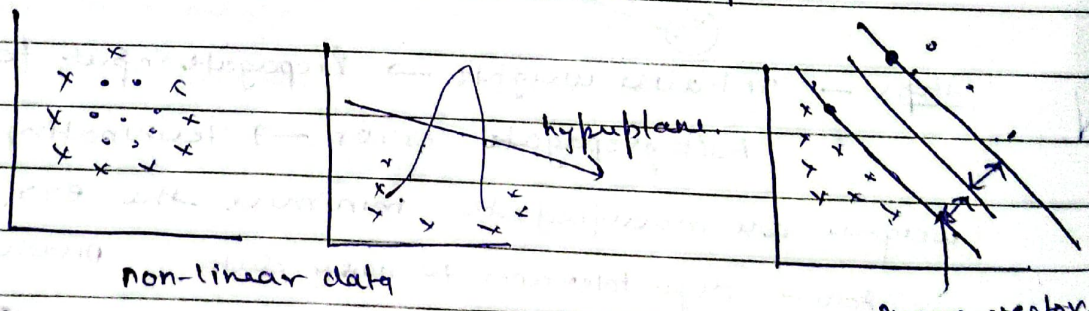
- K-nearest neighbour - knn algorithm assumes the similarity between new data and available data and assign class based on similarity. It is also called a lazy learner algorithm because it does not learn learn from training set immediately instead it stores and at time of classification it perform action.

- It stored data and when it get new data then it classify data into new category similar to new data.

steps - Select k nearest neighbour
- Among them count data point each category
- Assign new point to category which the no. of neighbour is maximum.



- SVM - Support vector machine - classification algorithm for both linear and non linear data
- It uses non-linear mapping to transform data into higher dimension
- with this now data can always be separated by hyper plane.
- training can be slow. but accuracy can be high.
- Applicatn - digit recognition, time series prediction.



- very effective on high dimensional data.

- genetic algorithm - It is based on analogy to biological evolution.
- An initial population is created consisting randomly generated rules.
- Based on survival of fittest a new population is formed
- offspring are generated by crossover and mutation.
- Process continues until a population P evolves when each rule in P satisfies a prespecified threshold.

- Cluster analysis - cluster - a collection of data objects
  - similar data are in one cluster
  - dissimilar to others lie in other clusters.

Cluster analysis - grouping a set of data object into cluster.
It is unsupervised technique

Application - to get insights from data
- As it is a preprocessing step for other algorithm.

general application - Pattern recognition, image processing

Examples - Marketing - discover groups in customer base
Land use - Identification of similar land
Insurance - Identify high claim cost groups
City Planning - Identify group of house based on house type
Earth-quake - observe earth quake epicenters

- good clustering — high quality clusters
  - quality depends on similarity measure
  - ability to find hidden pattern.
  - high inter class similarity.

- Types of data in clustering analysis

- Interval-valued variable - Interval scaled variable are continous measurement of roughly linear scale

- standardise data
  - calculate mean absolute deviation.

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots x_{nf})$$

- calculate standardised measurement (z-score)

$$z = \frac{x_{if} - m_f}{s_f}$$

Similarity and dissimilarity are measured on the basis of distance.

Date..... Page.....

$$d(i,j) = \sqrt[q]{(x_{i1}-x_{j1})^q + (x_{i2}-x_{j2})^q + \cdots + |x_{ip}-x_{jp}|^q}$$

$q=1$ = Manhattan distance $q=2$ Euclidean distance.

Binary variable - Binary variable can take only two values. Example - gender - male or female.

- Symmetric - like male female may or female
- Asymmetric - like disease.
  +ve -ve.

- Nominal Variable A generalization of binary variable it can take more than 2 states eg. red, green, yellow

Method-1 simple matching - $d(i,j) = \frac{p-m}{p}$

p total variables m: m no of matches where i & j have same state.

method-2 - large no. of binary variable.

- Ordinal Variable - It can be discreate or continous.
  - order is important
  - can be treated like interval scaled.
- Ratio scaled variable - A positive measurement on a non-linear scale approximately at exponential scale.

  such as - Ae^{Bt}

methods - - first treat them like scaled variable
- apply logarithmic
- finally treat them as continous ordinal data.

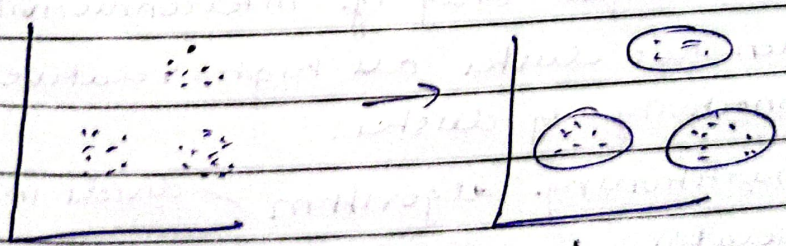
- Variable of mixed type - A database may contain all the six type of variables.

all these combined called mix-type variable.

clustering method.

1) Partitioning algorithms - construct various partitions and evaluate them by some criteria. ($n < p$)
2) Hierarchy algorithm - create a hierarchical decomposition of set of data using some criteria.
3) Density based - based on connectivity and density function
4) Grid Based - based on multilevel granularity structure. In this method grid is formed i.e object space is quantized into finite no. of cell. It is fast
5) Model based - In this all clusters are hypothesized. In order to find the data which is best suited for the model. (Model is hypothesised for each cluster)

* K-means clustering - It is unsupervised machine learning algorithm which group unlabelled data into different cluster. Here K - defines no. of predefined cluster.
- It is a centroid based algorithm where each cluster is associated with a centroid. the main aim of algo is to minimise sum of distance b/w data point



Steps - Select number K.
Select random K point as centroid.
Assign data point to their closest centroid
Calculate new centroid.
Repeat the process.

Date..... Page.....

cure clustering - It is a hierarchical clustering technique that adopt a middle ground b/w centroid based and all point extremes. It start with single point cluster and move to merge until desired clusters are formed

- Used for identifying spherical & non-spherical clusters.
- Useful for discovering groups and identify distribution
- Instead of one point centroid, cure use a set of well defined representative point for efficiently handling cluster & outlier.

Step - Data → Draw Random sample → Partition sample → Partially cluster partition

↑ elimination of outlier.

cluster partial cluster.

label data on disk.

cluster.

- Chameleon clustering - chameleon is a hierarchical clustering algorithm that use dynamic modeling to decide similarity among pair of clusters.
- Two cluster are merged only if interconnectivity and closeness between two cluster are high relative to internal interconnectivity of cluster.
- It use graph partitioning algorithm - divide in sub cluster
- Use an agglomerative hierachical clustering find cluster by combining sub-cluster.

- DBSCAN - Density based spatial clustering of application with noise.

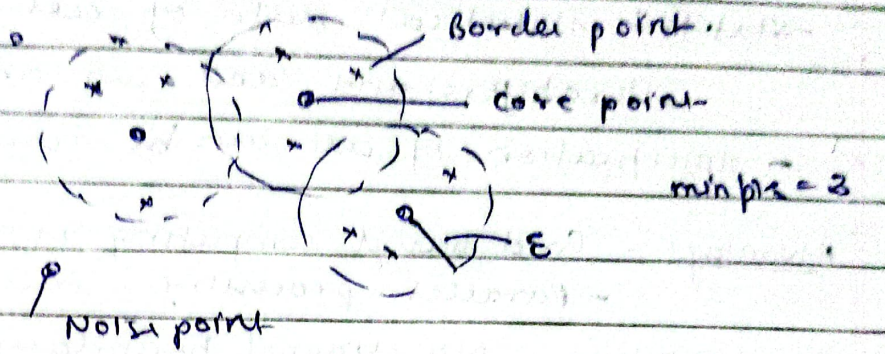
It is a density based clustering method It can

discover clusters of different shapes and size from large amount of data, which contain noise and outliers.

uses two parameters -

minpts - The minimum no. of points in a region

eps - A distance measure that is used to locate point in neighbour to point.



core point - A point is core if it has more than minpoint

Borderpoint - A point which has fewer than minpoint

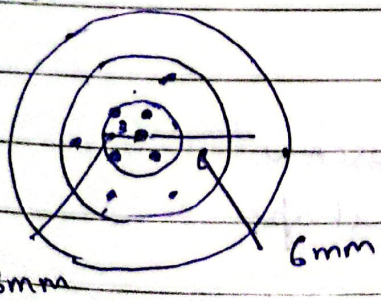
Noise - A point non a core point.

* Optics - ordering points to identify cluster structure.

- draw inspiration from DBSCAN
- Add two more concepts

1) core distance - It is minimum value of radius required to classify a given point as a core point.

If given point is not core then core distance is undefined.



minpts = 5

core distance = 3mm.

2) Reachability distance - It is defined respect to another data point q. Reachability between p & q is max of core distance of p and Euclidean b/w p and q.

Reachability distance is not defined q is not core.

Date..... Page.....

* sting clustering - (Statistical Information grid)
- Spatial area is divided into rectangular cell.
- there are several level of cell correspond to different level of resolution
- each cell at high level is separated to multiple small cell
- ~~Statistist~~ statistical data of cell is computed and stored beforehand and can answer query.
- specification of cell can be count, mean, min, max.

Advantage - Grid based computing is query independent
- Parallel processing, incremental update.

disadvantage - No diagnol boundary detected.

* clique clustering - clique is a density-based and grid based subspace clustering
grid-based - use grid
Density - density based.
- It find out cluster by taking density threshold and no of grid as input parameter.
- large dimension data
- Very scalable
- Uses apriori algorithm.

Advantage - better than kmean, DBSCAN
find cluster of any shape
simple method

disadv - If size of cell is unsuitable estimation will be too much, unable to find cluster.

Date..... Page .....

- Model based clustering is statistical approach to data clustering. The data is considered to have created from a finite combination of component model. model is → propability distribution.

- Statistical approach - It align object to cluster according to weight (probability distribution) New means are computed.

- Neural Network approach - Represent each cluster an an example acting as a prototype of cluster New object are distributed to cluster whose exemplar is most similar according to some distance measure.

- Outlier discovery -

Outlier are set of objects that are dissimilar from remainder of data.

Problem - find top outliers.

Application - credit card fraud detection
telecom fraud.
customer segmentation
medical analysis.

It can be found using distance based approach and deviation based approach.

Date..... Page.....

Unit-V

Data warehouse - data warehouse is collection of data marts representing historical data from different operations in the company. data is stored in structure optimised for querying and data analysis as data warehouse. Table design dimension & organisation should be consistent so that query are consistent.

Data	operational data	Informational data.
Data content	current values	summarised, archived, derived
Organizati	By application	by subject.
Stability	Dynamic	Static until refreshed
Structure	optimised for transaction	optimised for complex query.
Access frequency	High	medium / low.
Response	2-3 s	Several second.
Ex	updateable	non-updateable

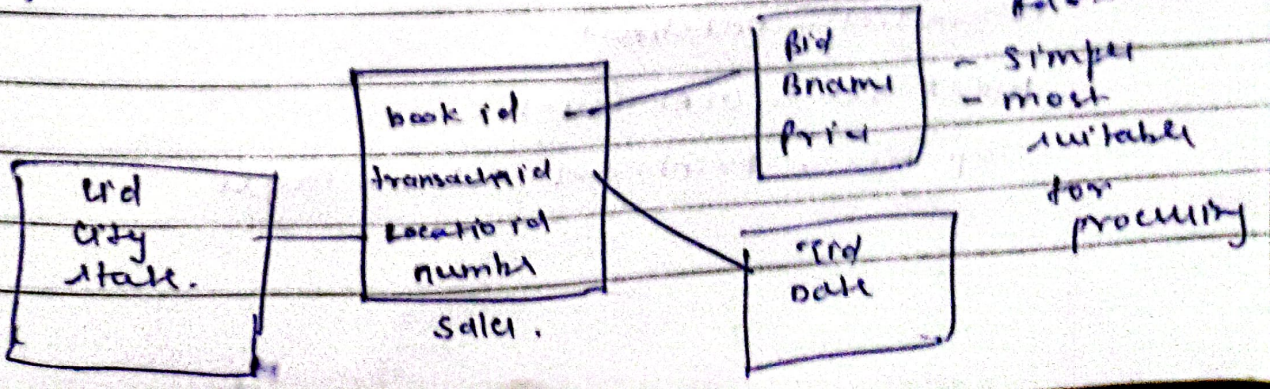
Database system	Data warehouse
• Support operational process	support analysis and performance reporting.
• Capture & maintain data.	Explore data
• Current data	history data.
• Data is update when trasaction occur	Data update on schedule process.
• Application oriented	subject oriented.
• Primitive & highly detailed	summarised and consolidated.
• ER based	Star / Snowflake based.

* multidimensional data model - It is a method which is used for ordering data in database along with good arrangement and assembling of content in database.
They allow women to interogate question comparatively fast.
It represent data in form of data cubes which allow to view data from many dimension & perspective.

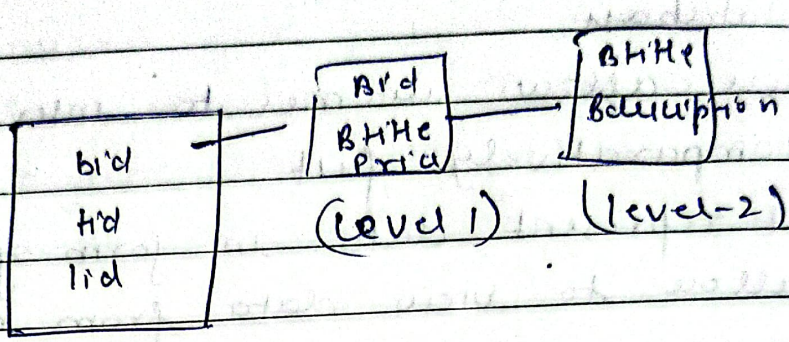
Advantages	disadvantages
- Easy to handle	- Highly complicated
- Easy to maintain	- complicated system
- better peyormance	- dynamic design
- better representation	
- work on complex system	

- Schema for multidimensional database
schema logical representation of entire database
It includes name and discription of all record types.

Star schema - Each dimension in star schema is represented with, only one dimension.
- This dimension table contain the set of attributes.
- Single table for each dimension.



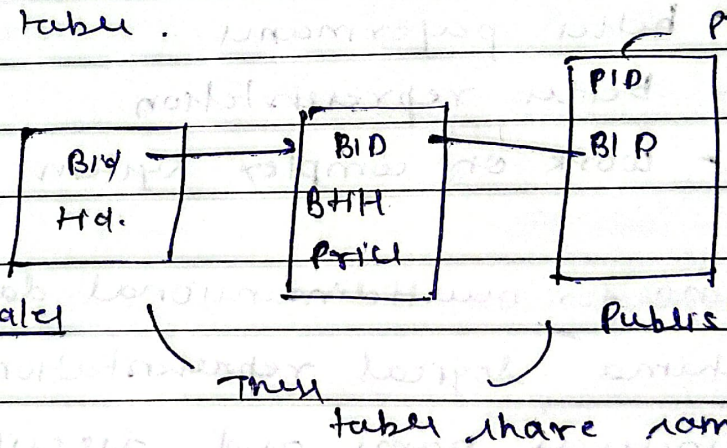
Snowflake schema - Some dimension table in snowflake schema are normalized. Normalization splits up data into additional table.



Adv - less redundancy
Easier to update

Disadv - complex.

Fact constellation schema - also known as galaxy schema. It has multiple tables ~~have~~ shares many dimension table.



Adv

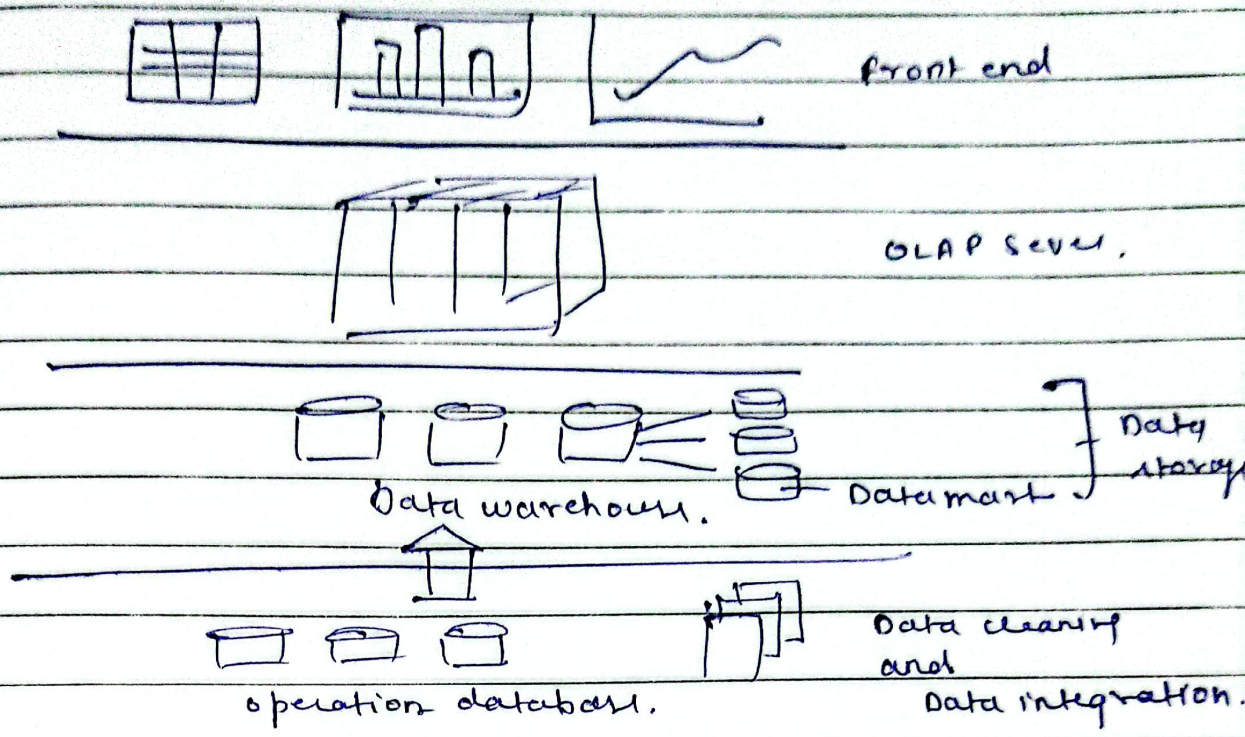
disadv
- complex
- difficult to maintain
- Dimension table are large.

Datawarehouse - 3 tier architecture

Bottom tier - Database

Middle tier - OLAP server

Top tier - Front-end client layer.



architecture

Datamart - A data mart contain subset of coporate data that is of value to specific group of user.

- Independent data mart are sourced from data captured from one or more operational system.
- dependent data mart are sourced directly from enterprise data warehouse.