DMW

Other MCQs

- http://mukhtiar-samoo.blogspot.com/
- http://mukhtiar-samoo.blogspot.com/2015/02/data-warehouse-mcqs-with-answer.html

Unit 1

Data Mining:

- Data mining is the method of processing the large data stored inthe data warehouse and finding the meaningful information or patterns from that large data.

KDD (Knowledge Discovery in Databases)

- KDD is one of the technique of computer science which helps humans in extracting the useful knowledge from data.
- Data maining is one of the step in the KDD process.

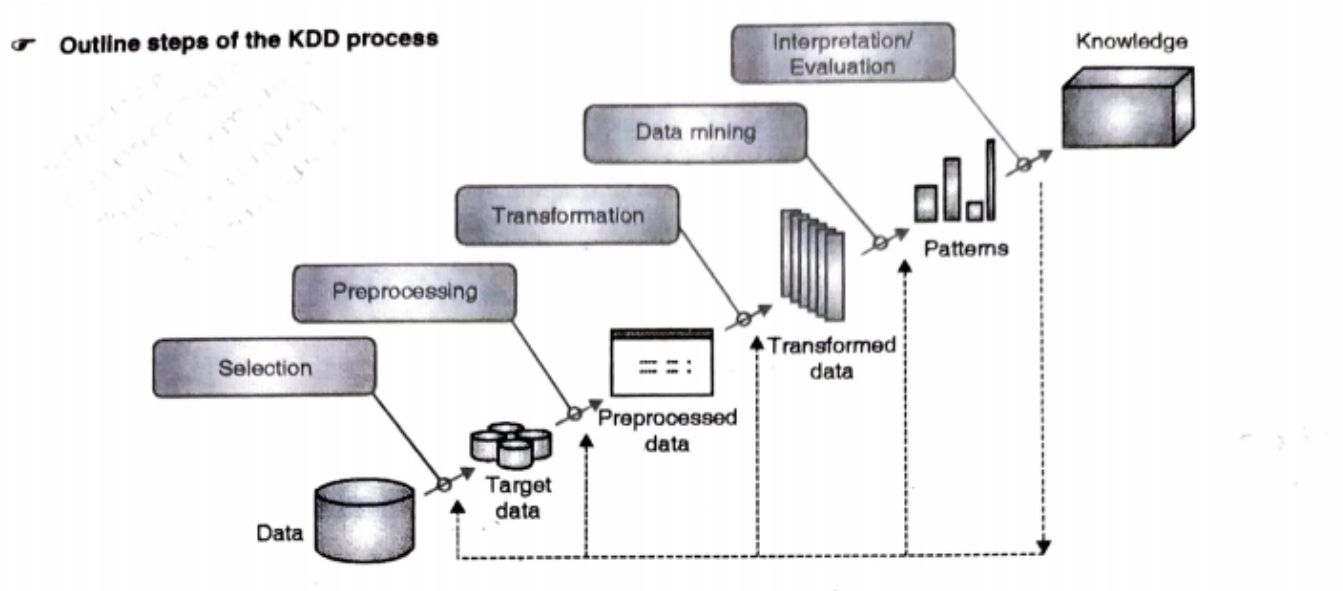


Fig. 1.1.1 : KDD Process

Architecture of a Typical Data Mining System

1.1.4 Architecture of a Typical Data Mining System

Architecture of a typical data mining system may have the following major components as shown in Fig. 1.1.2.

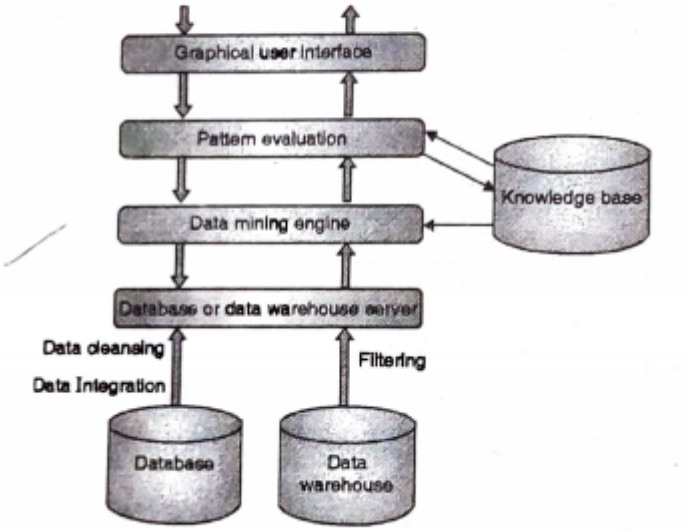


Fig. 1.1.2 : Architecture of typical data mining system

Data Mining Task Primitives:

- Task Relevant Data
- Kinds of knowledge to be mined
  - Associative, clustering, classification, regressstion, etc
- Background knowledge
  - It is the information about the domain to be mined
  - Concept hierarchies are used to discover the knowledge at multiple levele of abstraction
  - Cocept hierarchies types:
    - Schema Hierarchies
      - It is the total or partial order among attributes in the database
      - Ex location hierachy as stree < city < state < country
    - Set grouping
      - organizeed values into sets or groups

        {low, avg, high} € all (salary)

        {1000....5000} € low

        {5001....10000} € avg

        {10001....15000} € high
      -
    - Operation derived

        - It is based on operation specified which may include decoding of encoded strings, etc
        - Ex URL or email address
        - xyz@cs.in gives login name
    - Rule based
        - It occures when whole or portion of concept hierarchy is defined as a set of rules and is evaluted dynamically based on database data and the rule definition.
- Intrestingness measure
    - objective measures of intrestingness
        - Simplicity
        - Support
        - Confidence
        - Novelty: Patterns contributin new information to the give pattern set are called as novel patterns Ex data exception
- Presentation of discovered patterns

Data, Information and Knowledge
Data: Data are facts which are not processed or analyzed
Information: information is obtained after processing of data.
Knowledge: It is useful to take decisions and actions for business. Information is transaformed into knowledge.

Data object and attributes:
Data object is stored in the database in the form of tables.
Data object characteristics are defined by using attributes.

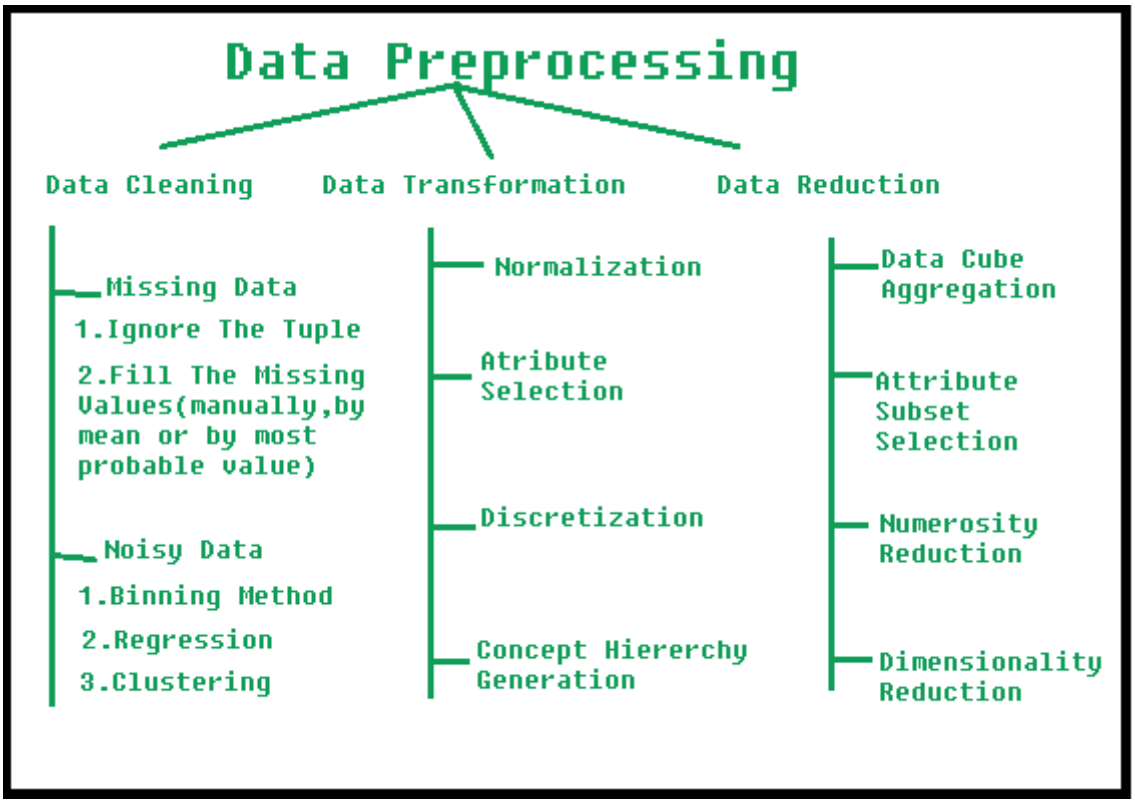
Attribute Types:

- Nominal: Names and categories
- Ordinal: Ranking of categories
- Binary: only two states categories
    - Symmetrical (Equal importance to both states): Ex male and female
    - Assymetrical (UnEqual importance): Ex mediacal test
- Numeric: quantitaive or quatifiable
    - Interval based: measured on linear scale Ex height and weight
    - Ratio based: non-linear scale Ex: temperature. Addition and subtraction can be performed but multiplication and division cannot be performed
- Discrete and Continuous: If an attribute can take any value between two specified values then it is ontinuous else it is discrete. or more simple definition is discrete attributes are generally counted and continuous are measured.

- Discrete Examples: Zip code, dogs, etc
- Continous: height, tempearature, etc.

Data Preprocessing:

- Data processing is the method of converting the data into information through classfication, sorting, translating, etc.
- Data processign can be manual or computer based
- Data processing systems are known as information systems
- When data processing does not involve data manipulation and it just only converts the data type then it may be called as data conversion.



Different forms of data preprocessing:

- Data cleaning
- Data Integration
- Data Transformation and discritization
- Data Reduction

1. Data cleaning

- It is also known as scrubbing
- In this process we remove the errors and inconsistencies form the data and improve the quality of data.

Steps in Data cleansing:

1. Parsing: Individual elements are identified
2. Correcting: We correct those elements

3. Standardizing: Standardize the elements
4. Mathcing: Remove the duplicate elements
5. Consoldating: Merge the similar elements records by identifying the relationships
6. Data cleaning: deal with possible errors
7. Data staging

Handling with missing values:

1. Fill missing values with global constant
2. Fill with column mean
3. Manual entry
4. Use data mining algorithm to find the best predicted value
5. SImply ignore that rows

Noisy data:

- A random error or variance ia a variable is known as noise.
- – **Different types of noise in data :**
  - Unknown encoding : Gender : E
  - Out of range values : Temperature : 1004, Age : 125
  - Inconsistent entries : DoB : 10-Feb-2003; Age : 30
  - Inconsistent formats : DoB : 11-Feb-1984; DoJ : 2/11/2007

To deal with noise data:

- Binning
- Outliers analysis by clustering
- Regression

Binning: In this process the sorted data is placed into bins or buckets.
There are two approaches of binning:

- Equal width (distance) partitioning

- **(a) Equal-width (distance) partitioning**
  - Divides the range into *N* intervals of equal size: uniform grid.

    bin width = (max value – min value) / N
  - **Example :** Consider a set of observed values in the range from 0 to 100.

    The data could be placed into 5 bins as follows :

    width = (100 – 0)/5 = 20

    Bins formed are :

    [0-20], (20-40], (40-60], (60-80], (80-100]
  - The first and the last bin is extended to allow values outside the range : (-infinity-20], (20-40], (40-60], (60-80], (80-infinity)
  - Disadvantages:
    - skewed data cannot be handled
    - outliers may be a problem
- Equal depth (frequency) paritioning or Equal height
  - **Example :** Let us consider sorted data for e.g. Price in INR

    4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
  - Partition into (equal-depth) bins: (N=3)

    Bin 1 : 4, 8, 9, 15

    Bin 2 : 21, 21, 24, 25

    Bin 3 : 26, 28, 29, 34
  - **Smoothing by bin means**

    Replace each value of bin with its mean value.

    Bin 1 : 9, 9, 9, 9

    Bin 2 : 23, 23, 23, 23

    Bin 3 : 29, 29, 29, 29
  - **Smoothing by bin boundaries**

    In this method the minimum and maximum values of the bean boundaries is found and each value is replaced with its nearest value either minimum or maximum.

    Bin 1 : 4, 4, 4, 15

    Bin 2 : 21, 21, 25, 25

    Bin 3 : 26, 26, 26, 34
  - Disadvantages:
    - Handling categorial values may be problem

Clustering: In clustering we can group the data together into cluster and then replace all the values of the cluster with the one value representing the cluster.

2. Data Intergration

- In Data warehouse data is integrated together from mulitple data sources i.e. databases.
- Various problems may encoutered during this like redudent data
- Also as we are merging the data from multiple sources so in some database the atttribute name is differnent and in some other database the attribute name is diff for the same entity.
- Redudancy can be detected by using the correlation analysis.
- X^2 (chi-square) test can be carried out on nominal data (categorial data).
    - Chi square used to test the probability of independece of a distribution of data.
    - – Chi-square test is defined by,

      $$X^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

      Where X^2 = Chi-square

      E = Frequency expected which is the amount of subjects that you would *expect* to find in each category based on known information.

      O = Frequency observed which is the amount of subjects you *actually* found to be in each category in the present data.
    - **Degrees of freedom :** The degrees of freedom(DF) is equal to :

      $$DF = (r-1) * (c-1)$$

      where, r is the number of levels for one categorical variable and c is the number of levels for the other categorical variable.
    - **Expected frequencies :** It is the count which is computed for each level of categorical attribute. The formula for expected frequency is

      $$E_{r,c} = (n_r * n_c) / n$$

      - Where $Er_{,c}$ is the expected frequency count for level r of attribute X and level c of attribute Y,
      - n_r is the sum of sample observations at level r of attribute X,
      - n_c is the sum of sample observations at level c of attribute Y,
      - n is the total size of sample data.
- Correlation coefficient and covariance may be used with numeric data.

3. Data transformation and data discretization

Data transformation:

Processes in data transformation:

- Smoothing: Remove noise from the data
- Aggregation
- Generlization: In this process data us replaced by higher level cocepts using the concept heirarchy.
- Normalization: data is scaled
- Attribute Construction: In this process new attribute is added and used for data mining process.

Data transformation by Normalization:

- Min-Max Normalization
  - This results in a linear alteration of the original data.
  - The values are withing a given range.
  - $$v' = (v - minA)/(maxA - minA) * (new_maxA - new_minA) + new_minA$$
- ZScore
  - This is also called as Zero Mean Normalization.
  - This technique uses the mean and standard deviation to normalize the values.
  - $$v' = (v - meanA) / std_devA$$

    Where, MeanA = sum of the all attribute value of A

    std_devA = Standard deviation of all values of A
- Decimal Scaling
  - In decimal scaling, scaling is performed by moving the decimal point in the value.
  - Movement of decimal point is decided based on the highest value in attribute column.
  - In order to normalize the value we just we have to divide the value with 1(no of zeros equal to length of term)
  - $$v'(i) = v(i)/10^k \text{ for the smallest } k \text{ such that } \max(|v'(i)|) < 1.$$

Ex. 1.6.1 : Consider the following group of data : 200, 300, 400, 600, 1000

(i) Use the min-max normalization to transform value 600 onto the range [0.0, 1.0]

(ii) Use the decimal scaling to transfer value 600. **SPPU - Oct. 16, 4 Marks**

Soln. :

(i) Min = Minimum value of the given data = 200

Max = Maximum value of the given data = 1000

$$V = 600 = \left(\frac{(V - \min)}{(\max - \min)}\right) * (1 - 0) + 0$$

$$= \frac{600 - 200}{1000 - 200} = \left(\frac{400}{800}\right) * 1 = 0.5$$

(ii) Decimal scaling for 600

$$10^K \text{ is } 10^3 = 1000$$

$$\frac{600}{1000} = 0.6$$

Data Discretization:

- In data discretization we divide the conrinuous values into intervals and then replaced the values with the specific interval labels.
- This helps in reducing the size of the data

Data Discretization by Binning:

- This is smoothing technique.
- two approaches: equal width and equal height or depth

Data Discretization by Histogram analysis:
Differnt type hystogram

- Equal width
- Equal depth
- V-optimal: Histogram with least variance is chosen
- MaxDiff

4. Data Reduction

It is not possible to process all the huge data stored into data warehouses because it takes so much time to process that whole data.
Data Reduction is the technique which results in reduced form of the dataset which is small in terms of valume but gives the same result as original dataset.

Different Data Reduction Techniques:

- Data cube aggregaiton
- Dimensionality Reduction
- Numerosity Reduction
- Data Compression

1. Dimensionlity Reduction (Remove the attributes)

- Wavelet transform:
  - Decome signal into differnet freq subbands
  - Ex: Image compression
- PCA (Principle Component Analysis)
  - PCA creates a represetation of the data with orthogonal vectors i.e. eigenvectors of the covariance matrix of the data.
  - This can be derived using SVD (Singular value decomposition)
  - The idea is to reduce the no of eigen vectors used to reconstruct the original data matrix, the amount of storage required is reduced.
- Attrbute Selection
  - In this method we select important attributes from the set of all attributes.
  - Different Methods:
    - Forward Selection: We'll start with empty set and step by step we'll select best attributes and put into the set.
    - Stepwise Backward Elemination: Here we'll start with the full set of atributes and in each step we remove the worst attribute.
    - Combination of forward and backward: At a time select best and remove worst from the remaining attributes.
    - Decision tree induction: Uses decision trees. In decision trees leaf nodes gives classification result, internal nodes represents attribute, and the branch represents attribute value.

2. Numerosity Reduction
   This techniue refers to reduce the volumne of data by choosing smaller forms for data representation.
   Types:

- Histograms
- Clustering
- Sampling
  - Simple Random: Equal prob of selecting any item

    - With replacement: we'll put the item back into populatn after selection so there is a change of repeat of same attribute.
    - Without replacement: remove from the population after selection.
    - Stratified sampling: The data is split into paritions and samples are drawn from each partition randomly.

4. Data Compression

- In data compression we'll reduce the amount of bits required to store the data. Data compression technique can be lossy or lossless.
- In lossy technique there is loss of information whereas in lossless there is no loss of information
- Lossless compression: Huffman encoding, run length encoding
- Lossy: wavelet tranform, PCA

Important Points

- Two types of approaches are used to integrate the heterogenous databases
    - Update driven approach
    - Query driven approach

Unit 2

- Data warehouse was defined by Bill Inmon in 1990
- A warehouse is subject oriented, integrated, time-varient and non-volatile collection of data.

Characterisitcs of Daa warehouse:

- Subject-oriented: In data warwhouse all the operations that we perform is related to partiualr subject. Like finding all the customer who has taken maxium loan amont so this problem is related to subject loan analyis.
- Integrated: Data is integarted from multiple sources.
- Time varient: All data in data warehouse identified with a parituclar time period. Data warehouse system stores all data whcih included historical data. Transactional system only stores current data not historical data.
- Non-volatile: Data cannot be changed or removed but new data is added.

Operational Database systems:

- Operational database systems are not suitabel to provide the strategic information.
- Strategic information extracted form previous operational systems and operational databse systems are not designed for straegic informations.
- Operations systems ex are inventory management, record management, university record system, etc.
- For stretegic information we have to use decision support systems or information systems.

OLTP and OLAP

OLTP

- OLTP is an operational system
- OLTP stands ffor (Online Transaction Processing)
- Application oriented

OLAP

- OLAP stands for (Online Analytical Processing)
- software category
- It support multiple view of data and analysis and run complex quieries on mutiple views of data.
- It is subject oriented and used for analysis
- It is data warehouse

Multidimensional Data Model

- It is logical design technique used for data warehouse.
- in data warehouse data is stored in multidimesional views.
- Dimensional model is the underlying model used for most of the commercial OLAP systems.
- In dimenian model there is at least one tabel with multipart key called as fact tabel and set of other related tabels called dimension tables.
- Mutlidimesnional model consists of two types of tables:
    - Fact table: contains facts or measures
    - Dimensinal tables: contains attribures of dimensions.

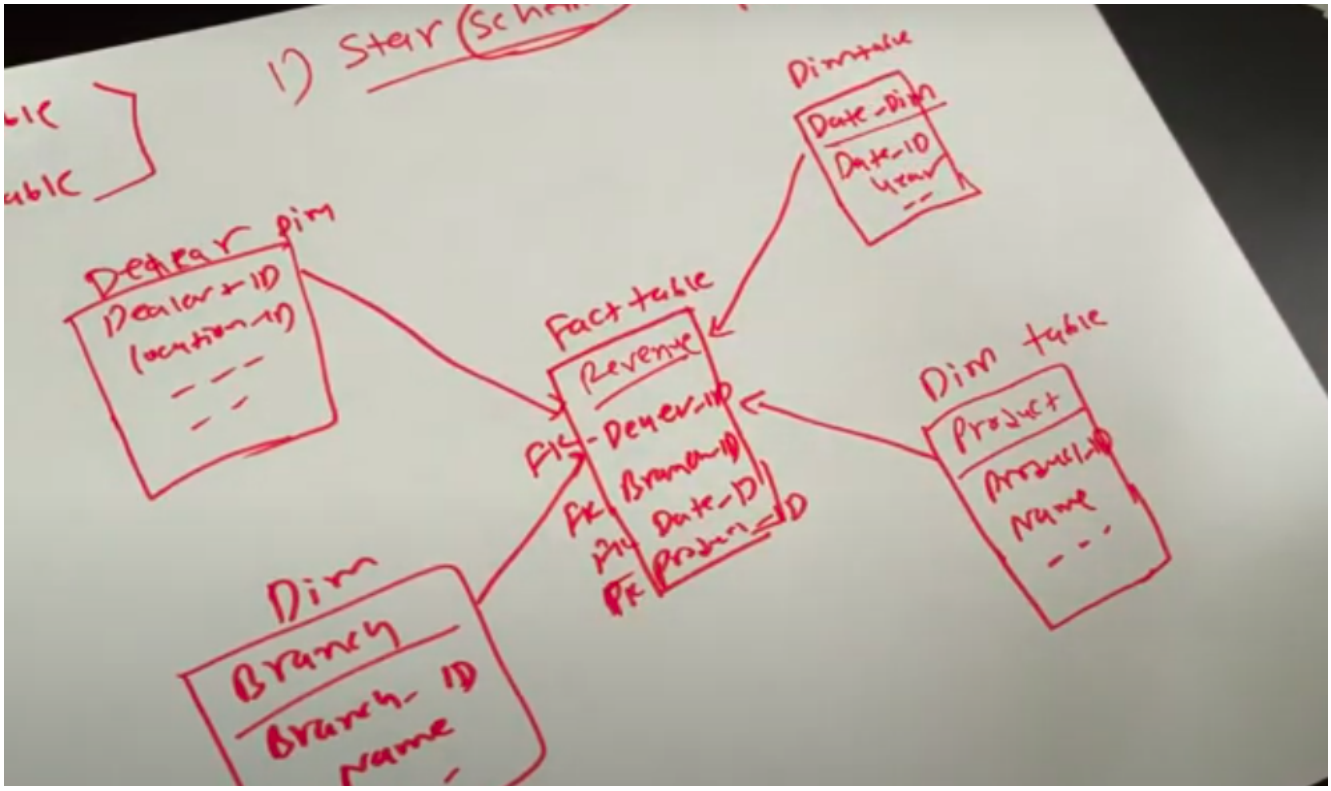
Multidimesnional Data models:

- Data cube
- Star schema
- Snowflake schema

Star Schema

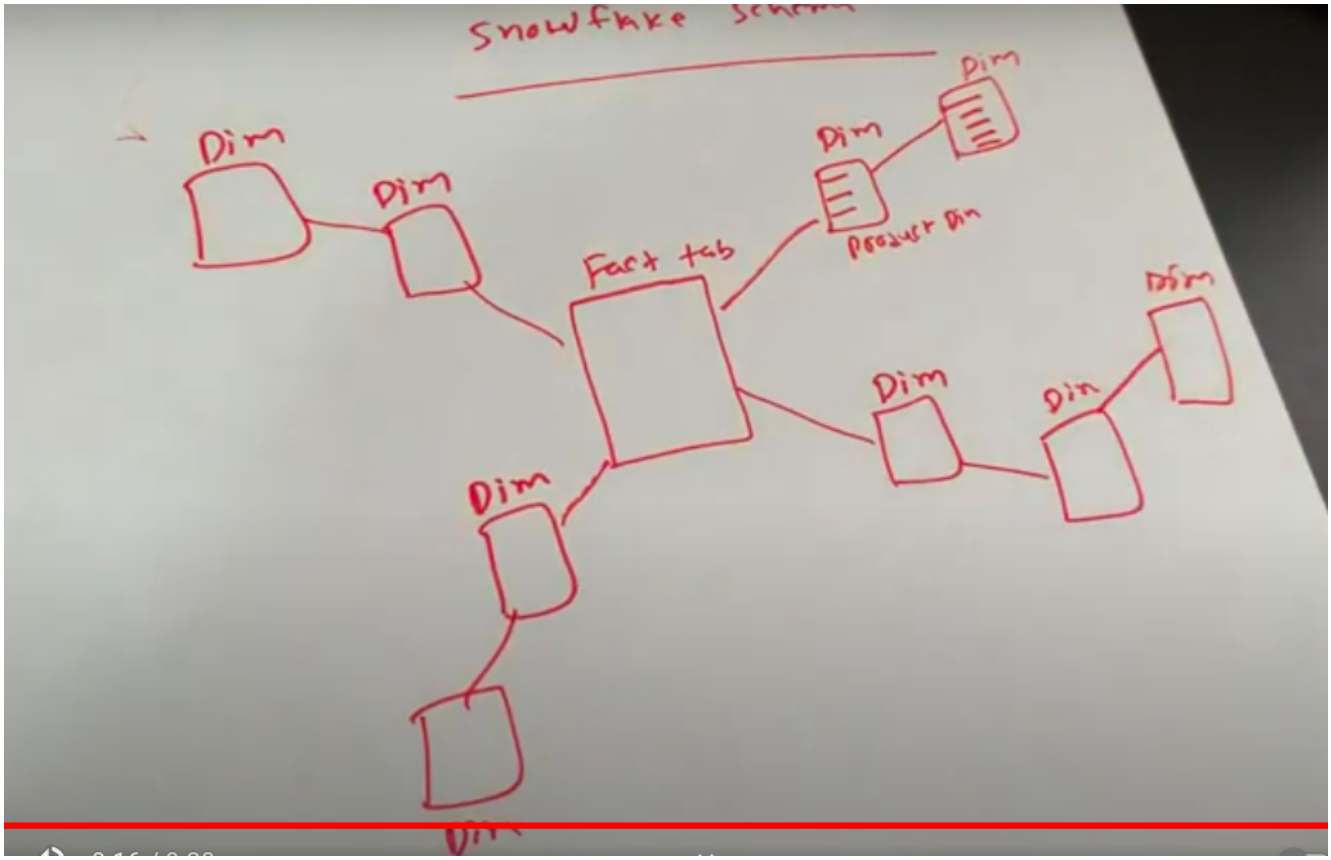
- In star schema the fact table contains the foreign key for all the primary key of the dimension tables.
- The size of the fact table is large as compared to dimension table.
- During dimension modelling fact tabels are normalized whereas dimension tables are not.
- Types of facts in star schema:
    - Fully-aditive:sum up throught all the dimensions in the fact table
    - Semi-additive: sum up throught some of the dimesnions
    - Non-additive: cannot be sum up through any dimesnion
- It is fully denormalized schema
- Reduced joins and faster query operation

- Also called as star join schema



Snowflake Schema

- In snowflake schema we will normalized the dimension tables.
- We divide the information of dimension table into multiple dimension tables.
- A snkwflake schema is used to remove the low cardinality
- It increases the join operations
- In a star schema if all the dimensional tables are normalized then this schema is called as snowflake schema.



- star schemas are usualyy not in BCNF form.
- In snowflake schemas dimesnion table are in 3NF.

Factless fact tables:

- factless fact tables does not contains facts or measures. It only contains the keys.
- Generally used to put the relation between multiple dimensions.
- Used for many to many relationship.
- Types:
  - Event tracking tables
  - coverage tables

Fact constellation schema

- In fact constellation schema we have more than one fact tables and the dimesnsion tables are shared between mutliple fact tables.
- It is also called as families of star.
- More complex design
- Used for the applications that need a high level of sophistication.

☞ **Family of stars**

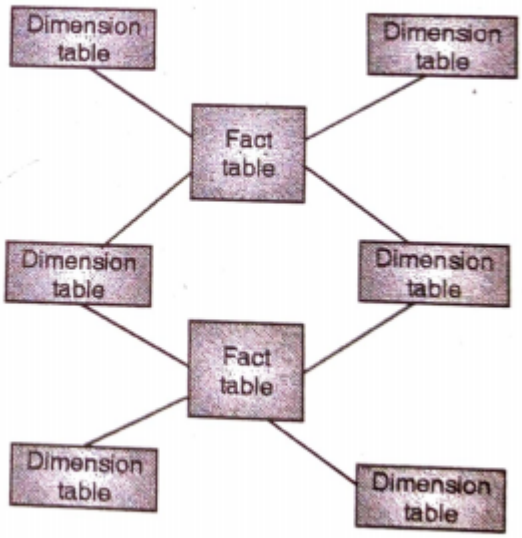
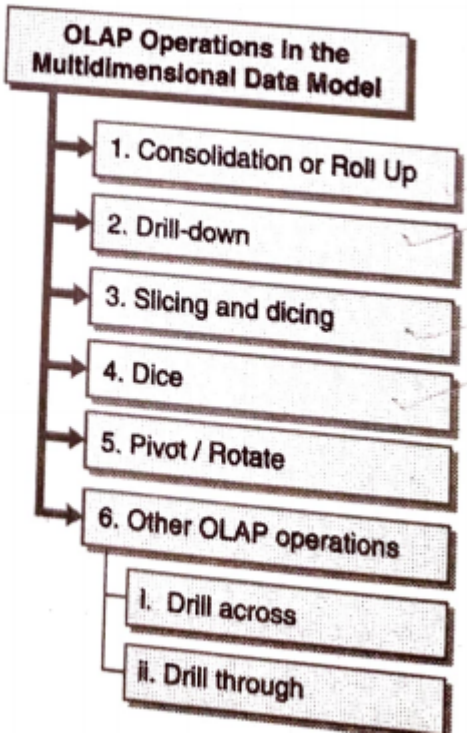


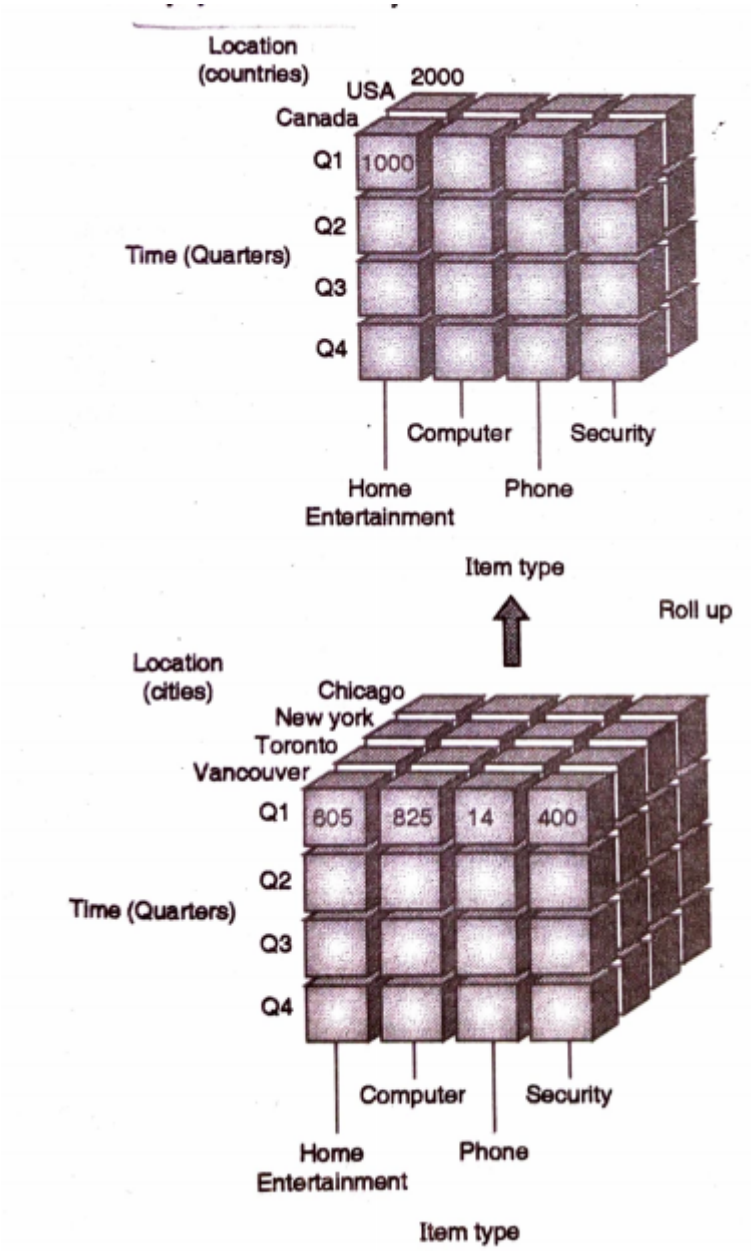
Fig. 2.3.7 : Family of stars

-

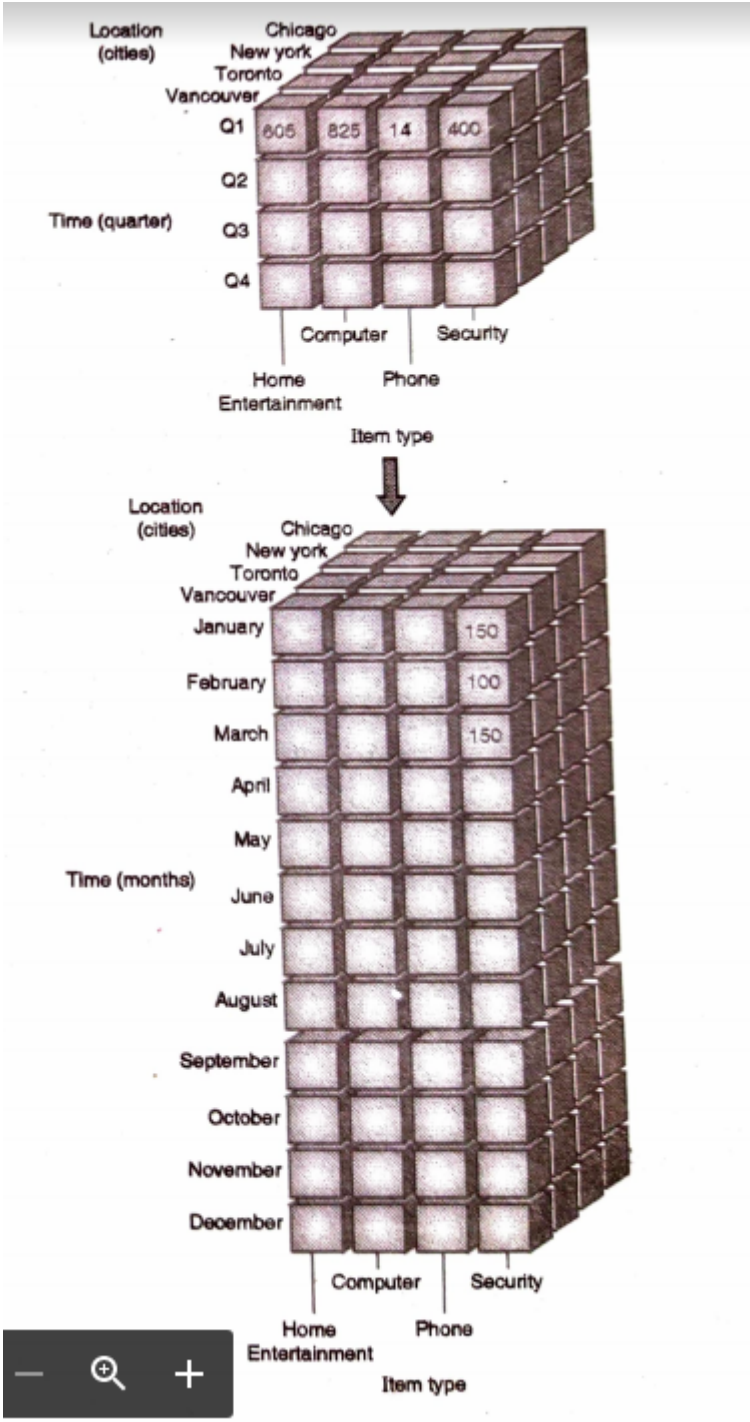
OLAP Operations in the Multidimensional Data Model



Roll up or Drill up: Roll up means generalizign the dimension.



Drill Down: drill down is changing the view of data to greater level of detail.



Slicing or Dicing: It refers to the ability to look at a database from various viewpoints.

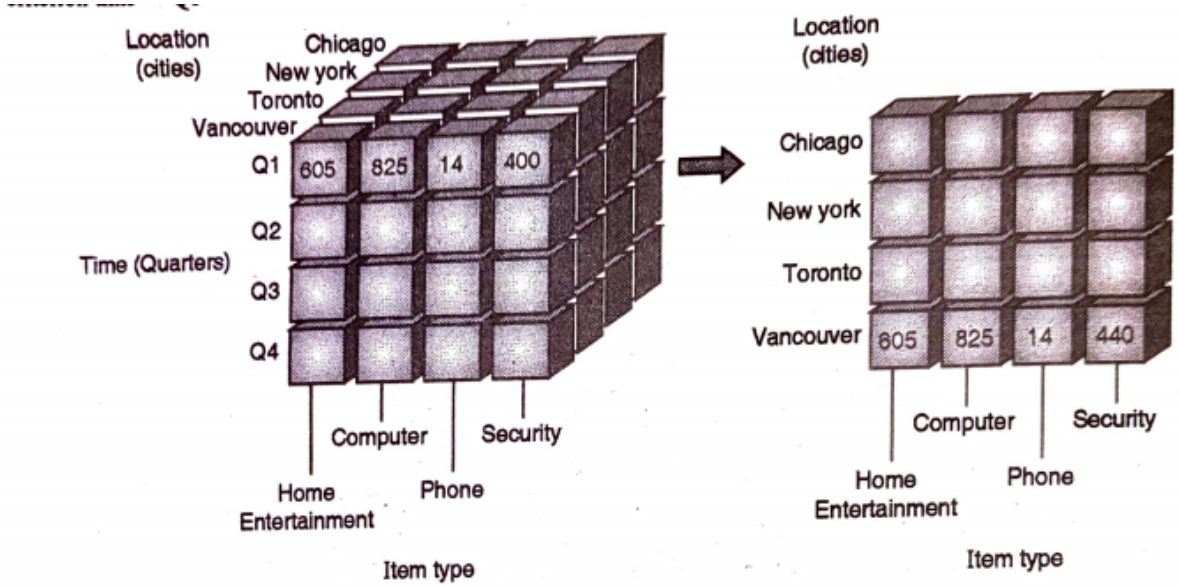


Fig. 2.4.4 : Slice operation

Dice: Dice operation carry out selection with respect to two or more dimensions and return the subcube.

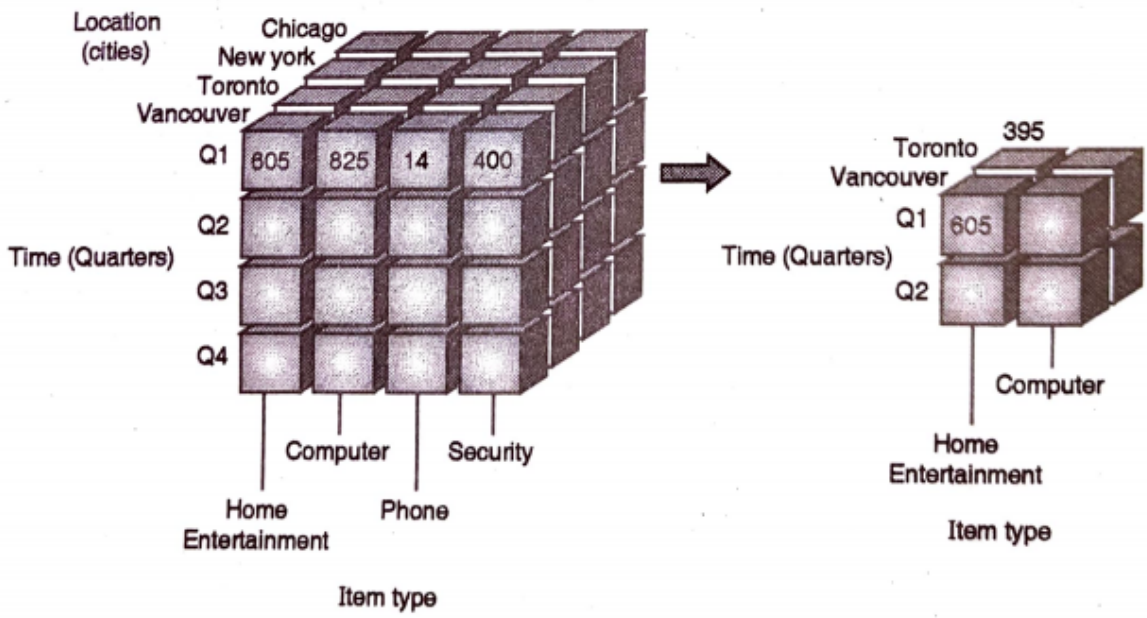


Fig. 2.4.5 : Dice operation

Pivot or Rotate: This technique is used for visulaization of data. By rotating we can view the different representation of the data.

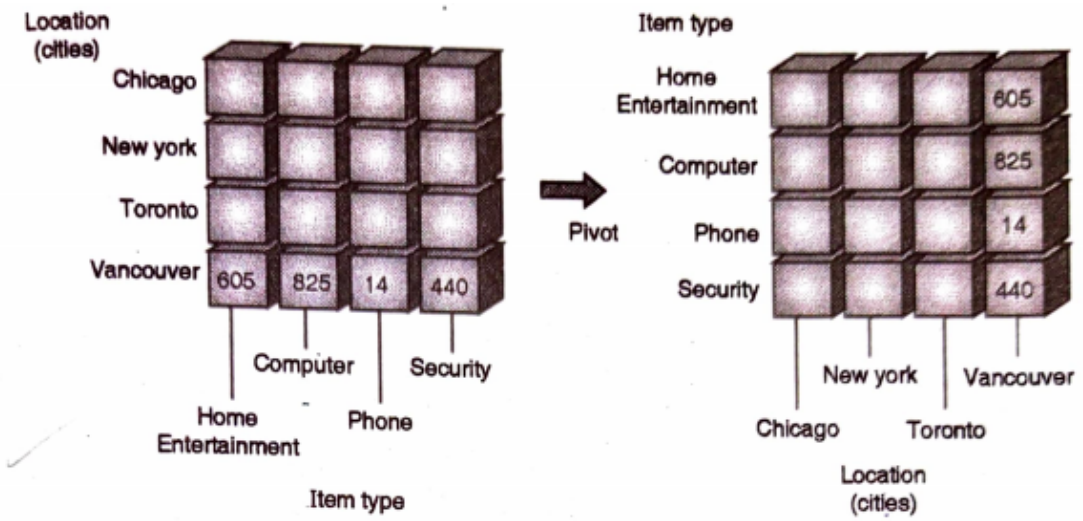


Fig. 2.4.6 : Pivot operation

Drill across: This is used when we have to execute the query that envolves more than one fact table.

Drill through: Drill through bottom level of the data cube. Uses SQL facilities.

Concept Hierarchies

- The amount of data can be reduced by using the concept hierarchies.
- The low level detailed data for ex age may be represented by using the high level data (eg in categories)

Data warehouse architecture

- Three kinds of systems used in data warehouse are:
  - Source systems
  - Data staging area
  - Presentation servers

Source Systems: Source systems are the organization operational systems sources and databases.

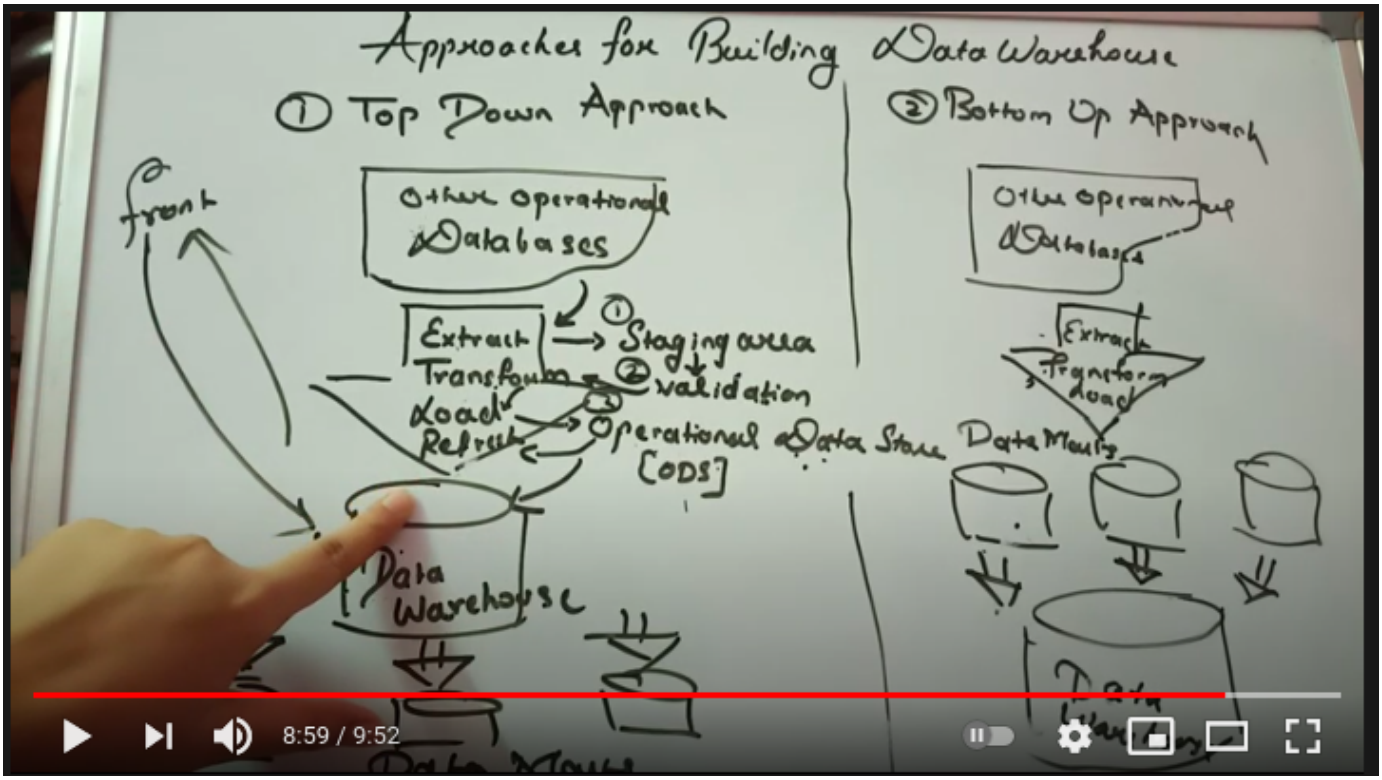
Data staging area: The data extracted from the source systems and stored in aread data staging area where the data is cleaned, transformed, combined. The data staging area does not provide and any query or presentation services.

- The data travels from source systems to presentation server via the data staging area. The entire process is called as ETL (Extract, Tranform and Load) or ETT (Extract Tranform Transfer)
- Oracle ETL tool is called as Oracle warehouse Builder (OWB)
- MS SQL Server ETL tool is called as Data Transmission Services (DTS)

Data warehousing Design Strategies:

- Top Down Approach
- Bottom Up Approach
- Hybrid Approach
- Federated Approach

- Practical Approach



Three-tier data warehousing Architecture

Top tier: consists of frontend tools - client side

Middle tier: OLAP engine implemented using either ROLAP or MOLAP.

Bottom tier: consists of data sources and data storage

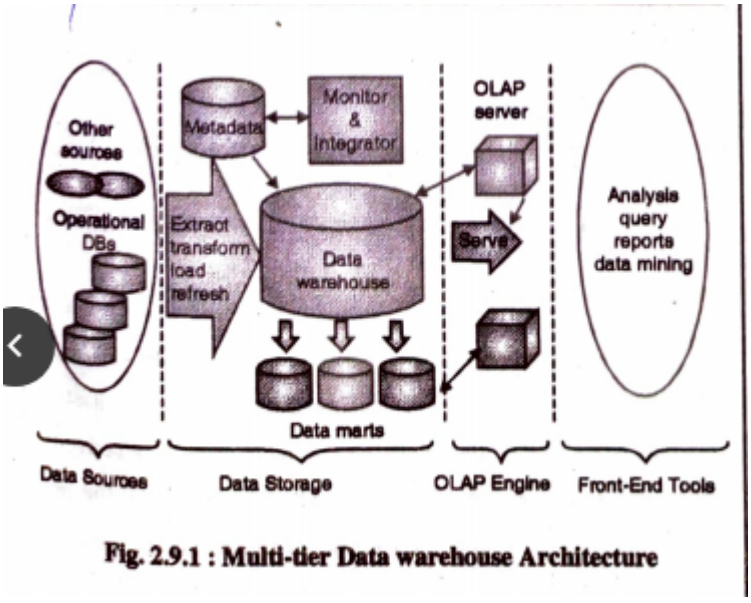


Fig. 2.9.1 : Multi-tier Data warehouse Architecture

Data warehouse Models:

- Enerprice datawarehouse: consist of all informatin together. centralized.
- Data Marts: subject oriented. Contains only particual depratmetn related or subject related data. subset of warehouse. data marts are decentrlized.
- Virtual Datawarehouse: views are created over operational databases systems.

Types of OLAP servers:

- MOLAP (Multidimensional Online Analytical Processing)
- ROLAP (Relational Online Analytical Processing)
- HOLAP (Hybrid Online Analytical Processing)
- DOLAP (Desktop Online Analytical Processing)

MOLAP:

- MOLAP uses multidimension data cube and for transforming relational data into data cubes extra investment is needed because cube technology are often propritetary.
- Fast performance
- Not suitable for large data.

ROLAP:

- ROLAP technology allows to perform OLAP operations directly on the relational data.
- Low performance
- Suitable for large data.

HOLAP:

- combinatio of ROLAP and MOLAP

DOLAP:

- This is the variationof ROLAP.
- Only desktop software is required on the machine to do the processing and analysis.

Unit 3

Measuring Data similarity and Dissimilarity

Similarity and Disimilarity:

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies

Two types of data structure for main memory based clsutering algorithms:

- Data Matrix or object by variable structure
- Disimilaroty Matrix or by object structure

Data Matrix:

- Data matrix stores n object in the realtional form or in matrix form.
- Data matrix is also called as two mode matrix
- $$\begin{bmatrix} x_{11} & \cdots & x_{11} & \cdots & x_{11} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{i1} & \cdots & x_{i1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{n1} & \cdots & x_{n1} \end{bmatrix}$$

Dissimilarity Matrix:

- $$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,1) & \cdots & \cdots & 0 \end{bmatrix}$$
- d(i,j) is close to 0 when the object i and j are similar
- d(i,j) == d(j,i)
- Matrix is symmetric
- One mode matrix as it only represents one dimension i.e. distance
- sim(i,j) = 1 - d(i,j)

Proximity Measures for Nominal Attributes:

- Proximity refers to either similarity or dissimilarity
- – Dissimilarity is given by,

  $$d(i,j) = \frac{p - m}{p}$$

  where, p = Total number of attributes describing the objects

  and m = Number of matches

- Similarity is given by,

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

Table 3.2.1

Id	Types of Property
1	Houses
2	Condos
3	co-ops
4	bungalows

- The Table 3.2.1 represents nominal data for an estate agent classifying different types of property. The dissimilarity matrix for the above example can be calculated as follows :

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

- The value in the above matrix is 0 if the objects are similar and it is a 1 if the objects differ.

Proxity Measures for Binary Attributes:

A contingency Table 3.2.2 for binary data :

Table 3.2.2

		Object n		
Object m		1	0	Sum
	1	A	b	a + b
	0	C	d	c + d
	Sum	a + c	b + d	P

Simple matching coefficient (invariant, if the binary variable is symmetric) as shown in Equation (3.2.1) :

$$d(i,j) = \frac{b+c}{a+b+c+d} \quad \text{...(3.2.1)}$$

Jaccard coefficient (non-invariant if the binary variable is asymmetric) as shown in Equation (3.2.2) :

$$d(i,j) = \frac{b+c}{a+b+c} \quad \text{...(3.2.2)}$$

Table 3.2.4

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jai	M	1	0	1	0	0	0
Raj	F	1	0	1	0	1	0
Jaya	M	1	1	0	0	0	0

- Distance between Jai and Raj (i.e. d(Jai, Raj)) is calculated using Equation (3.2.2) and use contingency Table 3.2.4.

Consider attributes : Fever, cough, Test-1, Test-2, Test-3, Test-4

Consider Jai as object i and Raj as object j

a = Attribute values 1 in Jai and in Raj also = 2

b = Attribute values 1 in Jai but 0 in Raj = 0

c = Attribute values 0 in Jai but 1 in Raj = 1

$$d(i, j) = \frac{b+c}{a+b+c}$$

$$d(\text{Jai, Raj}) = \frac{0+1}{2+0+1} = 0.33$$

Similarly, calculate distance for other combination

$$d(\text{Jai, Jaya}) = \frac{1+1}{1+1+1} = 0.67$$

$$d(\text{Jaya, Raj}) = \frac{1+2}{1+1+2} = 0.75$$

- So, Jai and Raj are most likely to have a similar disease with lowest dissimilarity value.

Proximity Measures for Interval Scaled:

- Here Euclidian distance, manhaten distance and minkowski disatnce is used.

1.	Euclidean L_2	$d_{Eue} = \sqrt{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^2}$
2.	City block L_1	$d_{CB} = \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$
3.	Minkowski L_p	$d_{Mk} = \sqrt[p]{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^p}$

Both Euclidean and Manhattean distance satisfy the following requirements:

- d(i,j) >= 0
- d(i,i) = 0

Proxity Measures for Ordinal Attributes:

- In the ordinal attribute we will firts normalize the value given to the rank and then simply we can use euclidean distance to find the dissmilarity between objects.

  - Let us consider an example :

Emp Id	Income
1	High
2	Low
3	Medium
4	High

  - The three states for the above income variable are low, medium and high, that is $M_f = 3$.
  - Next we can replace these values by ranks 3(low), 2(medium) and 1(High).
  - We can now normalise the ranking by mapping rank 1 to 0.0, rank 2 to 0.5 and rank 3 to 1.0.
  - Next to calculate the distance we can use the Euclidean distance that results in a dissimilarity matrix as :

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

  - From the above matrix it can be seen that objects 1 and 2 are most dissimilar so are the object 2 and 4.

Proxity Measures for Ratio-scaled Attributes:

Theser are the ways to handle the ratio-scaled attributes:

- As interval scale attributes
- As continuous ordinal scale
- Transform th data (ex logarithmic transformation) and then treating the results as interval scaled variables.

Proxity Measure of Mixed types:

- In case of mixed attributes combine all the attributes into a single dissimilarity matrix and computing on a common scale of [0.0,1.0]

Cosine Similarity

- Cosine similarity is used to measure the similarity between two vectors.
- Here data objects are treated as vectors.
- Similairty measured as the angle theta
- Similarity = 1 (Similar) when theata = 0 degree

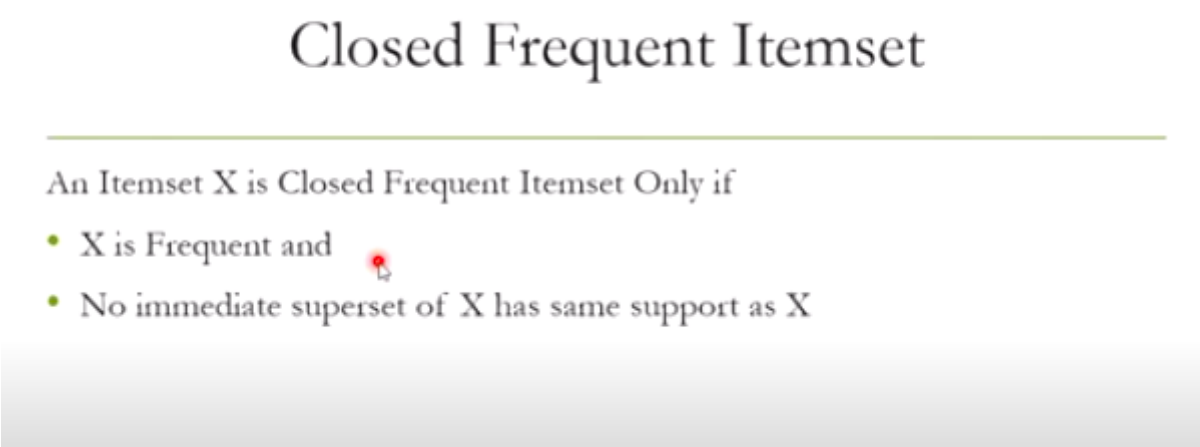
- Similaroty = 0 (Not Similar) when theta = 90 degree

Unit 4

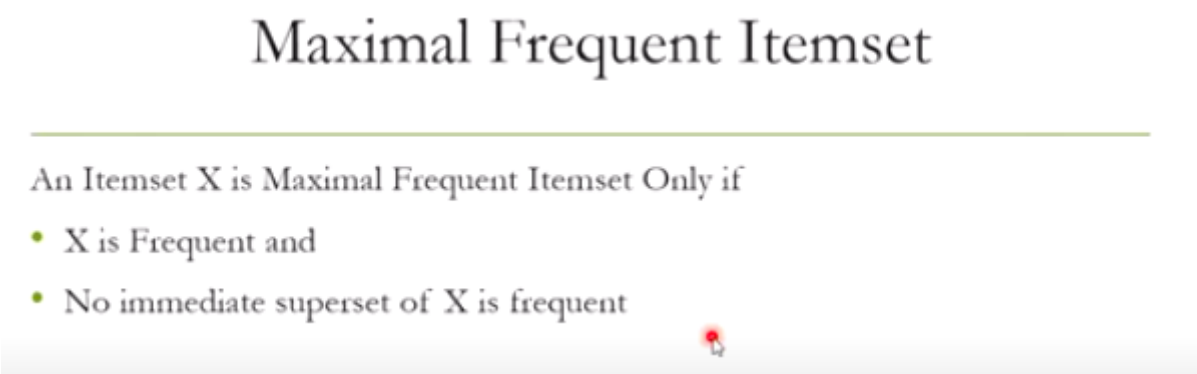
Frequent Itemsets:

- An itemset X is frequent if X's support is no less than a minimum support threshold.
- Procedure to find frequent itemsets:
  - Level wise search
  - Next search for all maximal frequent itemsets

Closed Frequent Itemsets:



Maximal Frequent Itemset:



Different ways to improves the efficiency of Apriory algorithm:

- Paritioning
- Sampling
- Transaction Reduction: remove the column by checking the min support from cube represenataion
- Hashset itemset genetation: uses hashset structure
- dynamic itemset generation
  - Confirmed frequent itemset: rectangle
  - Confirmed infrequent itemset: circle
  - suspected frequent itemset: dashed rectangle

- suspected infrequent itemset: dashed circle

FP Growth Algorithm

- In Apriori algorithm we are also generating itemsets which are not frequent. So this method is not efficient.
- In FP growth we are only generating frequent itemsets.
- In FP growth we first compress the database and generate the FP tree. (Frequent Pattern Tree)
- After FP tree generation we'll divide the database into set of conditional databases.

Algorithm:

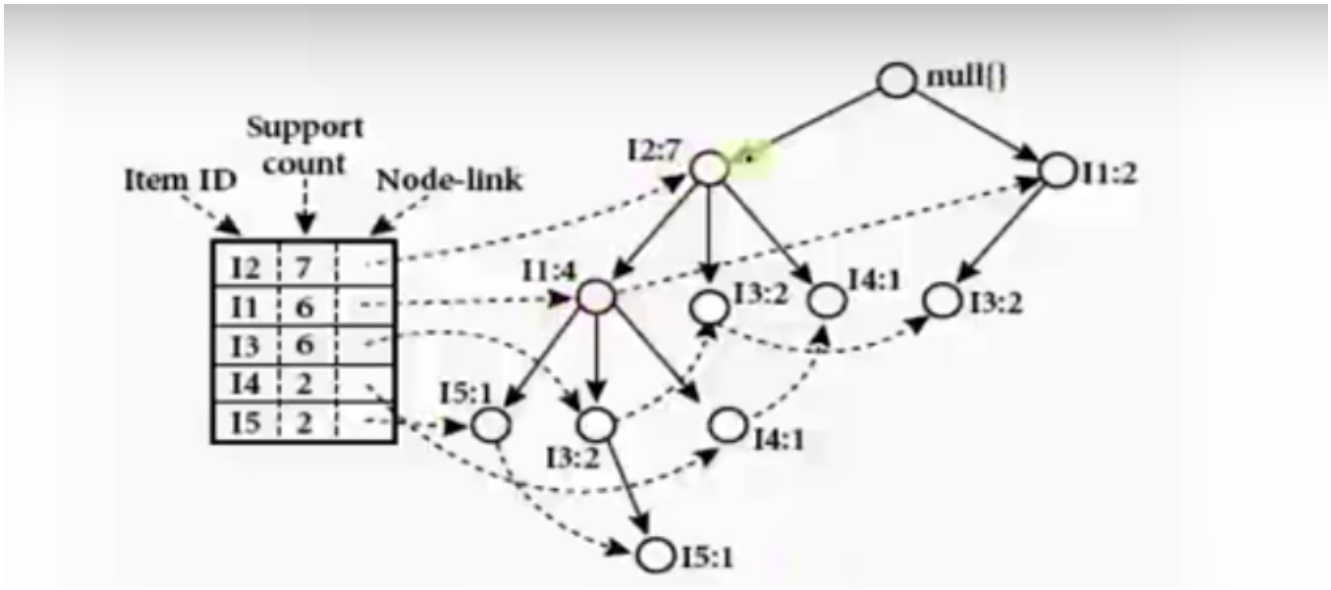
- min sup = 2

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

- Write each item support count in decending order

Items	Sup.count
I2	7
I1	6
I3	6
I4	2
I5	2

- Now start with null node of FP tree and traverse through each transaction in the database and arrange the transaction items according to their descending order of support count and create nodes in that pattern in the FP tree.
- Final FP tree

- 

-

Item	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns Generated
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}	⟨I2: 2, I1: 2⟩	{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}
I4	{{I2, I1: 1}, {I2: 1}}	⟨I2: 2⟩	{I2, I4: 2}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}	⟨I2: 4, I1: 2⟩, ⟨I1: 2⟩	{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}
I1	{{I2: 4}}	⟨I2: 4⟩	{I2, I1: 4}

- In the above diagram we'll start that process with item of least support count. After that we write the path from root node to that node from FP tree in Conditional Pattern Base.
- After that for conditional FP tree we merge the result and remove the item with support count < min support count.
- Also NOTE that in I3 generation while merging we have to consider sets separretly for merging i.e. we have to merge only left sets separtely and right sets separtely.
- Finally we generate frequent itemset conditional FP tree with original item. While doing this process in case of multiple items count we have to take minimum one. For example in the diagram consider I3 item and the last frequent itemset {I2,I1,I3:2} here while merging coditional FP tree with I3 item we have I2:4 and I1:2 so we take miniumum one i.e. 2
- Detailed Video: https://www.youtube.com/watch?v=ToswH_dA7KU&list=PLYT7YDstBQmE50voZ81eLS0hz2gUdZJwp&index=9&ab_channel=CSEGURUS

Mining Differnet Kinds of Association Rules:

- Mining Multilevel Assciation Rules
- Constraint-Based Association Rules
- Metarule Guided Association Rules

Mining Multilevel Association Rules

- Here items are always in the form of hierarchy
- Items which are at left nodes have low support

- An item can be either generalized or specializeed as per the heirarchy
- Two approaches of multilevel association rules:
  - Using uniform min support for all levels
  - Using reduced min support at lower levels
- Search Stretegies:
  - Level by Level independent
    - It is full bread serach method. The parent node is checked depending on whether node is frequent or infrequent other nodes are examined.
  - Level cross filtering by single item
    - The children of only freq nodes are checked
  - Level cross filtering by k-itemset
    - Find the freq k itemset at parent level and then only k itemset at next level is checked.
  - Controlled level cross filtering by single item
    - Here some min support is set for lower level.
    - Children nodes that only satisfy min support are checked.
    - This is also called Level Passage Threshold.

Mining Constraint Based Association Rules:

- Here mining performed based on the user specific constraints .
- Forms of contraints:
  - Data
  - Dimension/Level
  - Intrestingness
  - Rule
  - knowledge type

Mining Metarule Guided Association Rules:

- Here mining is performed based on the constraints specified in the syntactic form.
- $age(C, \text{“30..40”}) \wedge Salary(C, \text{“30K..50K”}) \rightarrow buys(C, \text{“Apple Product”})$

Important Points

- Market baske analysis term propesed by Agrawal
- Data warehouse is defined by Bill Inmon in 1990
- The Fp-tree Growth algorithm was proposed by Hanetal
- Dynamuc Itemset Counting Algorithm was proposed by Bin et al

- KMedioid algorithm - PAM (Partitioning Around Medioids)

Unit 5

Classification

- classification is a two step process:
  - model construction: In this step from the training data model is created. The constructed model based on training data set is represented as classification rules , decision trees or mathematical formulae.
  - model usage: Here we passed the test data to the model and model will assign class labels to this new objets of data. It also calculate the accuracy.
- the data we passed for testing is diff from training data, if we use same data then overfitting will occur.

Evaluating classification methods

- Predictive Accuracy: accuracy score of the unseen data
- Speed and scalability: time to contruct the model and time to use the model
- robustness: handling noise and missing values
- interpretability: understanding and insight provided by the model
- goodness of rules: decision tree size, compactness of classification rules.

Regression

- When we have to predict continuous values then in this case we use the regression.
- Regression is the most popular type in prediction.
- Both classification and regression comes into prediction.
- types of regression:
  - linear regression
  - Multiple regression

Linear regression : $Y = m + nX + u$

Multiple regression : $Y = m + n_1X_1 + n_2X_2 + n_3X_3 + \dots + n_tX_t + u$

Where :

Y = The dependent variable which we are trying to predict

X = The independent variable that we are using to predict variable Y

m = The intercept

n = The slope

u = The regression residual.

- regression uses a group of random variables for prediction and finds a relationship between them. This relatiobship is depicted in the form of straight line (linear regression) that approaximates all the points in the best way.
- In simple linear regression we have only one independent variable and in multiple regression we have more independent variables.
- If the line is straight line then it is a linear model
- If the line is curved lien then it is non linear model.

Other regression models

log linear regression model:

- loglinear models are models that postulate a linear relationship between the indepedent variables and the logarithm of the dependent variable.

$$\log(y) = a_0 + a_1 x_1 + a_2 x_2 \dots + a_N x_N$$

Decision trees

- decision trees are very poplar for classification and prediction.
- rules given by decision trees are easy to understand and can be directly used in SQL to retrieve the records from database.
- Decision tree algorithms
  - ID3 (Iterative Dichotomiser 3)
  - C4.5 (Successor of ID3)
  - CART (Classification And Regression Tree)
  - ASSISTANT
  - CHAID (CHI-squared Automatic Interaction Detector)
- decisoin tree rules are simple if then rules

Decision tree algorithm

- create node N
- D is the training data to the algrithm, so if the dataset contains all the same class labels then make the node N as leaf node and assign that class label to it.
- If the dataset contains some missing or empty values then make the node N as leaf node and assign the class based on majority class label.
- Otherwise use the attribute selection crtetion to get the best attribute and assign it to the node N.
- divide the dataset accoring to the values of the selected attributes values and go with the branches and repeat the process until we get the whole tree.
- time complexity is O(N D^2) where D is the no of features

Tree Pruning

- because of outlier and noise decision tree generated may overfit and to avoid this we use the pruning technique.
- two types:
  - Prepruning
    - here pruning started at the beginning i.e. at the time of model construction.
    - Here we set some threshold value and if the tree passes that threshold value then we stop generating tree.
  - Postpruning
    - Builde the full tree and then start pruning and remove the branches.

Rule Based Classification

- Rule based classification is the classification by using the if then rules.
- syntax is: IF condition THEN conclusion
- Ex: X->Y
- here X is conuction of attributes and Y is the class label of the rule
- LHS of rule is rule antecedent and RHS of rule is consequent.

Rule coverage and Accuracy

- Coverage of a rule: Percentage of tuples that satisfies the anteecdent of a tule.

```
Coverage = Number of tuples covered by rule / total no of tuples in D
```

- Accuracy of rule: Percentage of tuples that satisfy both antecedent and consequent of a rule.

```
Number of tuples correctly classified by rule / no of tuples covered by
rule
```

Example

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(Status=Single) → No

Coverage = 4/10 = 0.4 = 40%

Accuracy = 2/4 = 0.5 = 50%

Characteristics of Rule Based Classifier

Mutually Exclusive Rules: Every record of dataset is covered by at most one rule of classifier and the rules are independent of each other.

Exhaustive Rules: Classifier generates rules for every possible combination of attributes values and each record is covered by at least one rule.

- If more than one rule is triggered then it need conflict resolution.
- In conflict resolution, highest priority is given to the rule which has maximum attribute test.

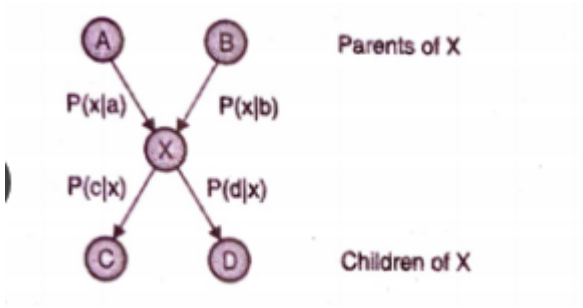
Rule Induction using a sequential covering algorithm

- Using sequential covering algorithm, IF-THEN rules can be exracted without genrating the decision tree.
- The rules can be learned one at a time.
- Sequential algorithms:
  - AQ
  - CN2
  - RIPPER

Bayesian Belief Networks

- In bayesian belief network conditional independence is defined between subsets of variables.
- bayesian belief networks are also called as bayesian networks, belief networks, and probablistic networks.

- A belief network is defined by two components:
  - A directed acyclic graph
  - A set of conditional probability tables



- each node represents a random variable.
- variables may corrspond to actual attributes or to hidden attributes.
- variables may be discrete or continuous valued.
- each arc represents a probablistic dependence.

Training Bayesian Belief Networks

- Training the network if the network topology is known and the variables are observable
  - compute the CPT (Conditional Probability Table) entries.
- Training the network if the network topology is known and some of the variables are hidden
  - Use gradient descent algorithm
  - Algorithm that follow this learning form are called as adaptive probabilistic networks.

Classification using frequent patterns: Associative Classification

- Frequent patterns generated from association can be used for classification is called Associative classification.
- Steps:
  - Find the frequent itemsets
  - Generate the association rules from the frequent itemsets
  - Oragnize the rules to form a rule based classifier
- Associative Classification algorithms:
  - CBA (Classification Based on Associations)
    - simplest associative classification algorithm
    - mine all the association rules satifying the support and confidence.
    - sort the rules by confidence.
    - classify using the rule that satisfies the query and has the highest confidence.
  - CMAR (Classification Based on Multiple Association Rules)
    - Mine rules satisfying confidence and support

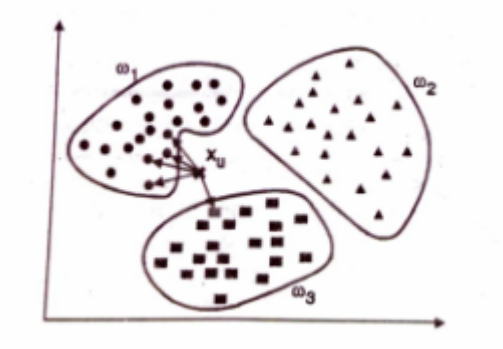
        - sort by confidence
        - find all rules whihh satisfy the given query
        - group them based on their class labels
        - classify the query to the class whose group of CARs has the maximum weight.

Lazy Learners

- classification methods can be classifies as **Eager Learners and Lazy Learners.**
- Eager learnes are those classification techniques in which a given set of training tuples constructs a generalized model and then we use that model for the test data.
- Examples of eager learners are Decision tree, bayesian classificatiom, support vector machine, rule based classification and associative classification ,etc.
- Lazy learners are those classification techniques in which a genetalized model is created only when its see a test tuple. So here learners waits unitl the last minute before doing model construction.
- It stored training data and do very little processing on it but does not create model from it in advance. Also called as instane-based learners.
- Disadvantages of lazy learners:
    - computationally expensive
    - requires efficient storage techniques
- Advantages:
    - support incremental learning
    - well suited to implementation on parallel
- Lazy learners examples: K Nearest Neighbors and case based reasoning.

K-Nearest-Neighbor classifier

- It is used in the field of pattern recorgnition.
- In KNN the for the test tuple we K similar tuples from the training tuples and then based on the resules we give class to the test tuple.
- Here for finding K similar tuples we can use any distance measure like euclidian, manhattan, minskowski, etc.
- KNN are extermely slow when classifying test tuples O(n)
- By simpel presorting and arranging the stored tuples into search tree the no of comparision can be reduced to O(logN)



Case Based Reasoning (CBR)

- case based reasoning means using old experiences to understand and solve new problems.
- In case based reasoning, a reasoner remembers a previous situation similar to the current one and uses that to solve the new problem.
- To solve the current problem: the problem is matched against the cases in the case base and similar cases are retrieved. The retrieved cases are used to suggest a solution which is reused and tested for success. If necessary the solution is then revised. Finally the current problem and the final solution are retained as part of a new case.
- four main processes of CBR: retrieval, reuse, revision and retention.
- Applications of CBR:
  - Diagnosis
  - Help Desk
  - Assessment
  - Decision support
  - Design

Unit 6

Multiclass Classification

- In multiclass classification there are N differernt classes and the test tuple belongs to one of these classes.
- Two approaches in multiclass classification:
  - One-vs-All Classification
    - Here we use the binary classifier and we create N different binary classifiers.
    - classifier i trained using tuples of class i as the positive class and the remaining tuples as the negative class.
    - And then for the test tuple, the set of classifiers vote collectively.
    - If the classifier i predicts the positive classs for X then class I get one vote and if classifier predicts negative class then all the classes execepts positive get one vote.
    - Finally the class with maximum vote is assigne to X.

- All-vs-All Classification
  - This is an alternative approach that learns a classifier for each pair of classes.
  - Given N classes, construct n(n-1)/2 classifiers.
  - For classifiying test tuple X, each classifier votes and the class with maximum votes is assigned to the tuple X.
  - All-vs-All approach is better then One-vs-All approach.

Semi-Supervisted Classification

- In this classification we pass both labelled and unlabelled data to built the classifier.
- two types:
  - Self-training
    - In this method first classifier is created from the labeled data and then the classifier tries to label the unlabled data and try to learn from it. The tuple with most confident label prediction is added to the labeled data.
    - drawback is it may reinforce errors.
  - Co-training
    - In this approach two or more classifiers teach each other.
    - Each learner uses diff and indepdent set of features for each tuple.
    - cp=o-training is less error prone as compared to self training.

Reinforcement Learning

- Based on goal-directed learning
- In machine learning, learner knows which action to take but in reinforcement learning, learner doesn't know the action, it discover which action gives most reward signal.
- Elements of reinforcemnet learning:
  - A policy
  - A reward function
  - A value function
  - A model of the environment (optional)

Multi-Perspective Decision Making and Multi-Perspective Learning

- Multi-Perspective learning is needed for multi-perspective decision making.
- Multi-perspective learning refers to learning from knowledge and information collected from different perspectives.

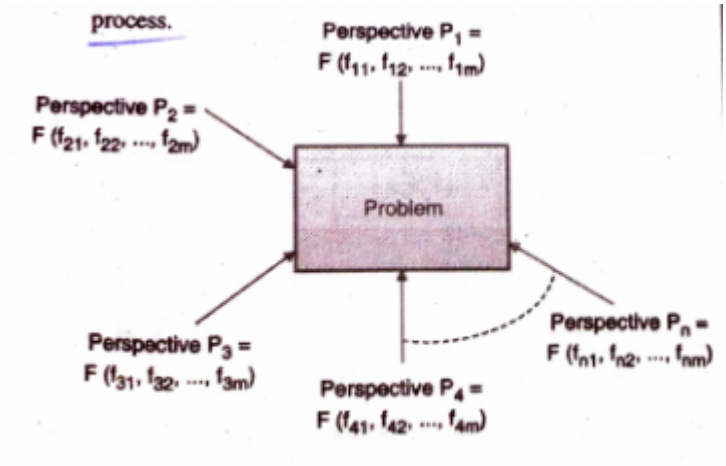


Fig. 6.5.1 : Multi-perspective Learning

- Each of the perspective is represented as a function of features.
- There may be an overlap among the perspectives.
- Perspective based information can be represented as an influence diagram and as an decision tree.

Model Evaluation and Selection

Various methods for estimating a classifier's accuracy are:

- Holdout method
- Random subsampling
- Cross-validation
- Bootstrap

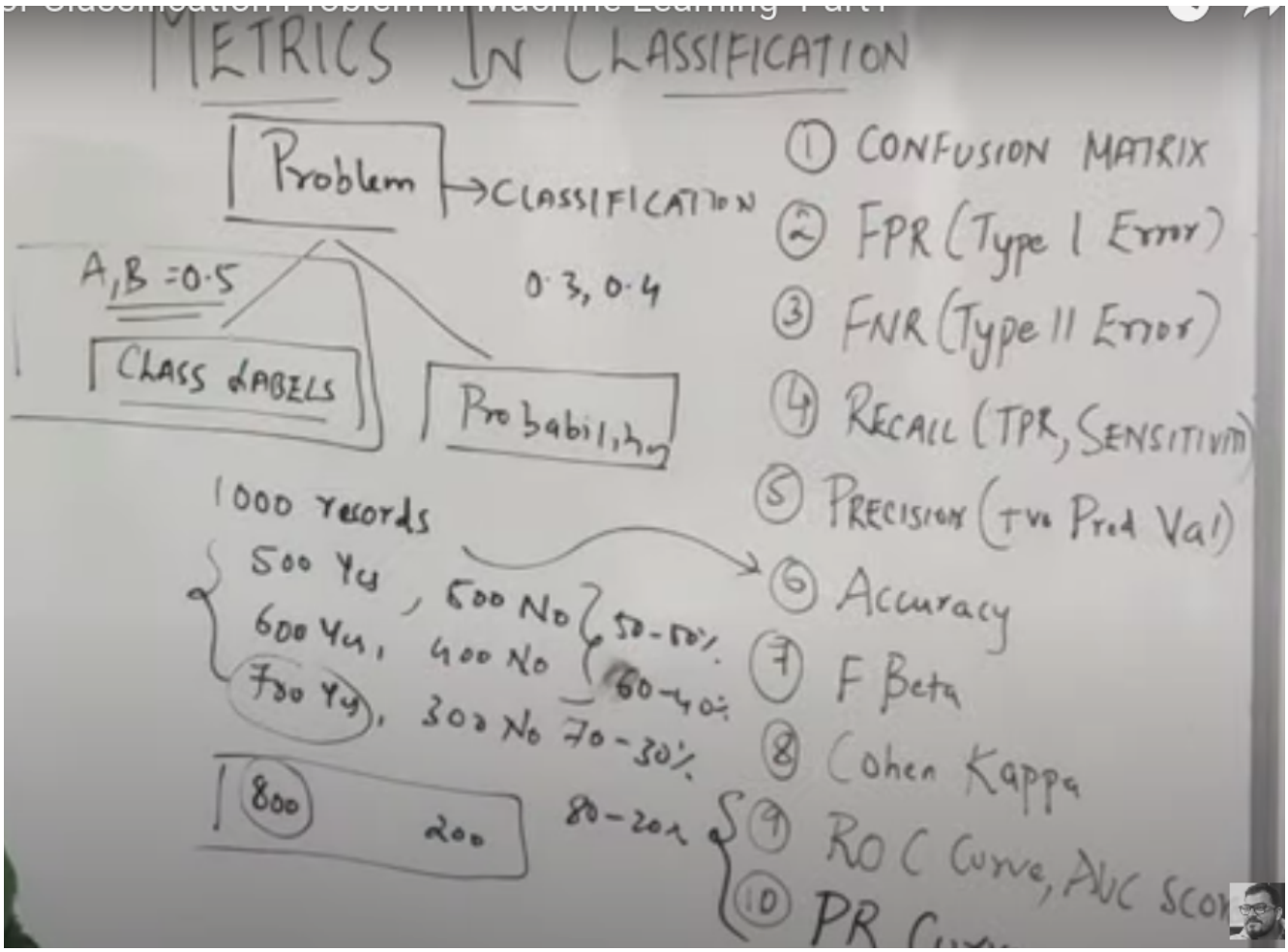
For comparing the classifiers to select the best one then the following methods are used:

- Confidence intervals
- Cost benefit analysis and Receiver Operating Characteristic (ROC) curves.

Diagnostics of classifiers

- For balance dataset we can use auuracy measure.
- If we have imbalance dataset then we should not have to use Accuracy measure.
- In case of imbalance dataset we can use recall, precision, FPR, or FNR.

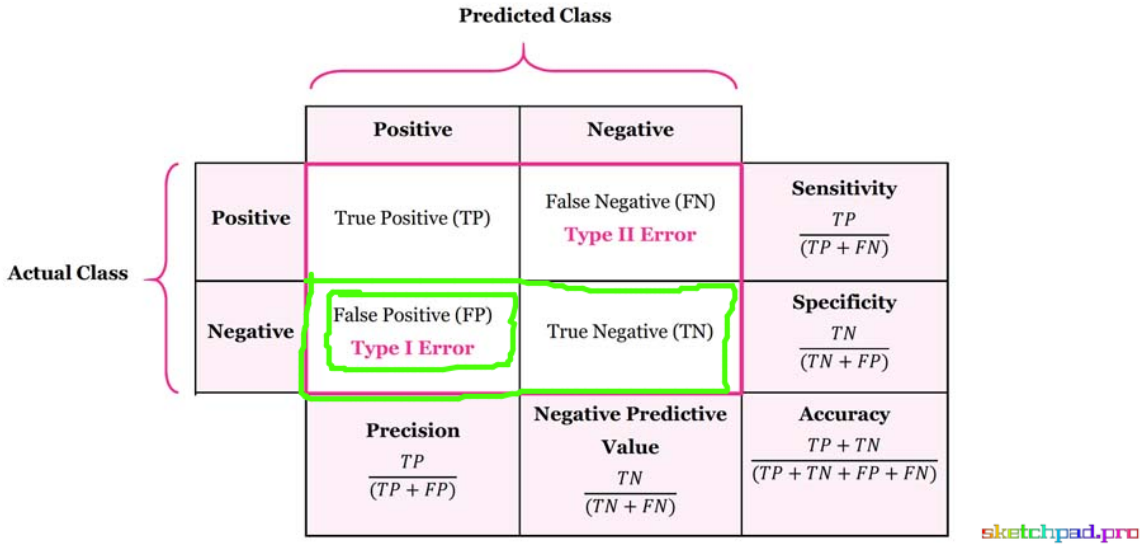
- What is balance dataset? 50-50, 60-40, 70-30 are balanced datasets.



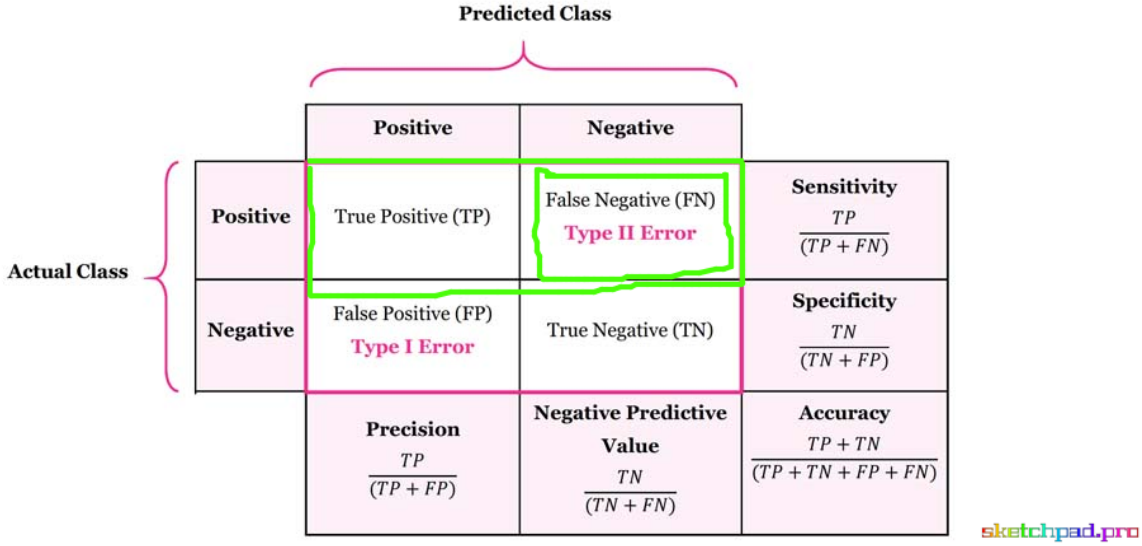
		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

- Accuracy: TP+TN/(TP+TN+FP+FN)

- FPR (False Postive Rate) or TypeI error: FP/(FP+TN)

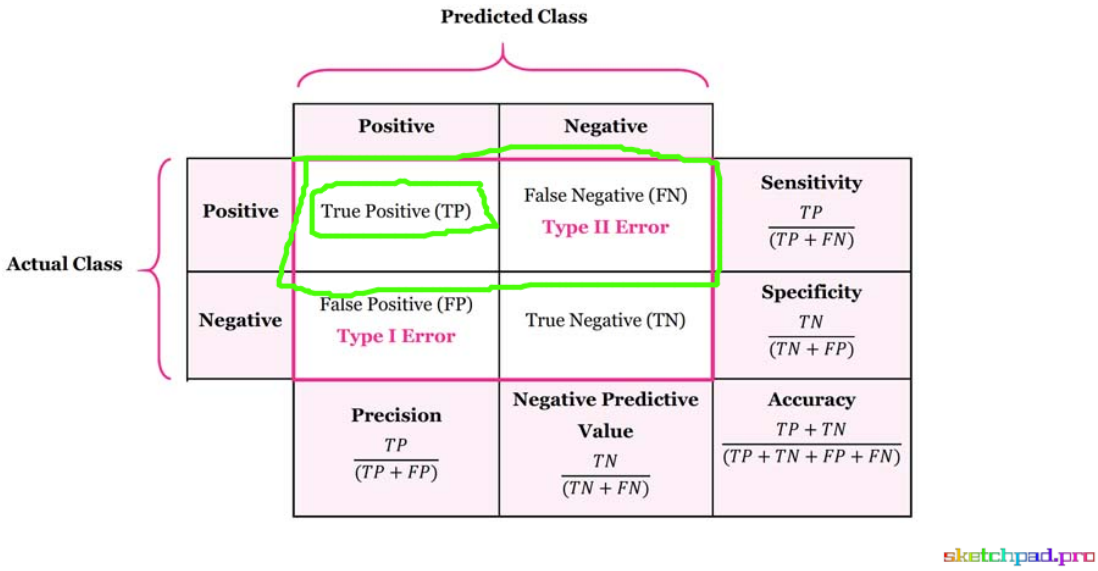


- FNR (False Negetive Rate) or TypeII error: FN/(FN+TP)

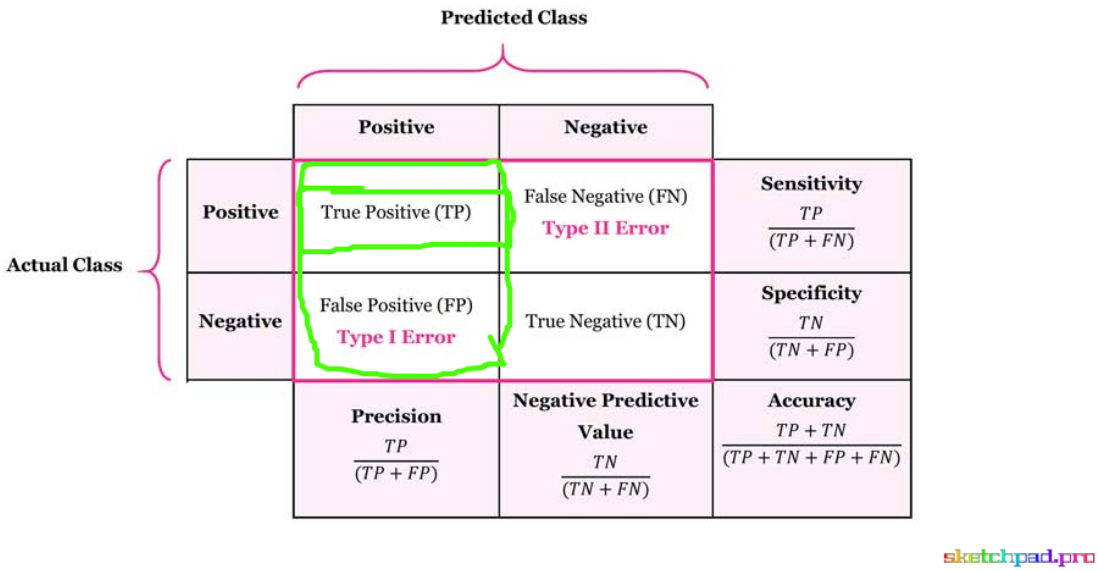


- Error Rate:
  - A percentage of errors made over the whole set of instances used for testing.
  - error rate = 1 - accuracy
  - OR error rate = (FP + FN) / (TP + TN + FP + FN)
- Recall:
  - Out of total positive values how many positives we have predicted correctly
  - It is the measure of completeness
  - TPR
  - Sensitivity

- TP/TP+FN



- Precision:
  - Out of totally actual predicted positive result how many are actually positives
  - It is the measure of exactness
  - Positive prediction value
  - TP/TP+FP

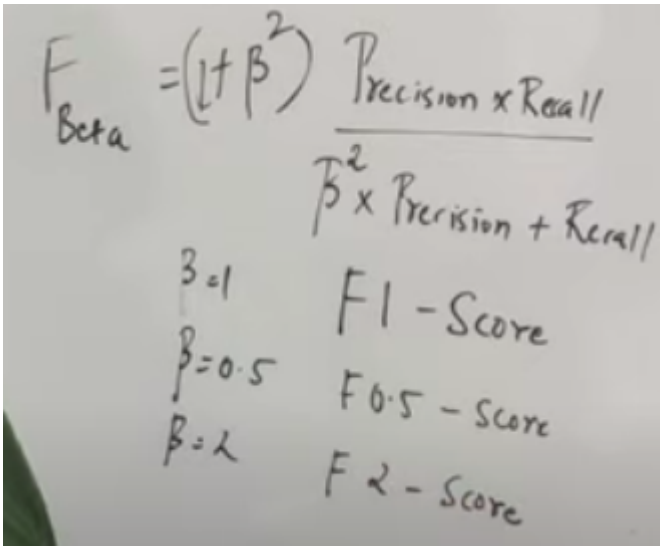


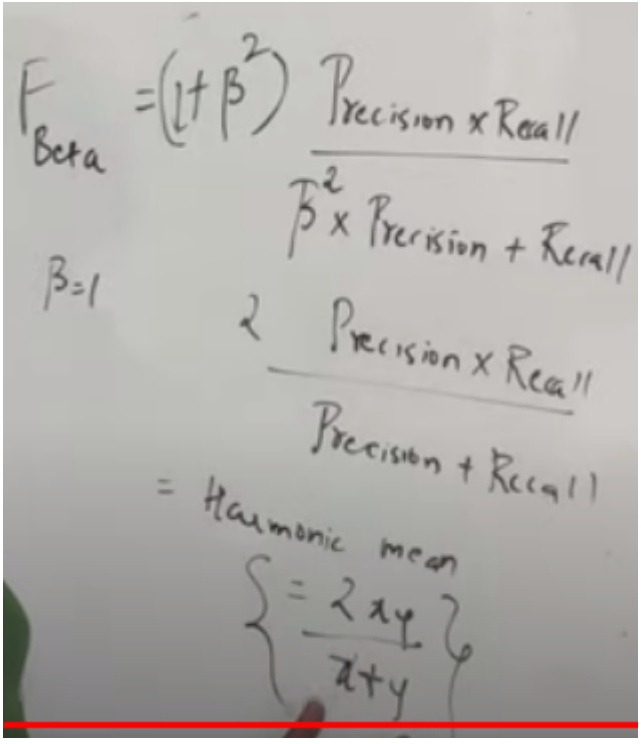
- Specificity:
  - out of total negaive values how many negatives we have predicted correctly.

- TN/TN+FP

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

sketchpad.pro

- F Beta
  - when both precision and recall are important then we use F beta
  - 

  - 

- Select B=1 when FP and FN both are equally important
- Select B=0.5 when FP has more importance then FN
- Select B=2 when FN has more importance then FP

Holdout Method

In holdout method we divide the entire dataset into training and testing data.
Generally 1/3 for testing and 2/3 for training.

Problem: If some class label are present only in test dataset then the model does not know about it because training dataset does not contains that class labels.

Solution: To solve this problem we use stratification, method which ensures that both training and testing data have equal no of samples of same class.

Random subsampling method

- It is a variation of the holdout method
- The holdout method is repeated k times.
- Each split randomly selects a fixed number example withput replacement.
- The overall accuracy is calculated by taking the average of the accuracies obtained from each iteration.

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

Cross-validation method

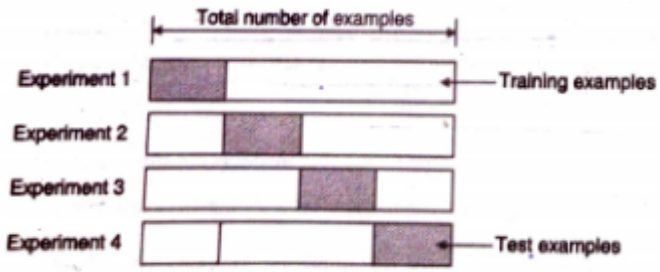
- Avoid overlapping test sets.
- k-fold cross-validation
  - First step: data is split into k subsets of equal size (usually by random sampling)
  - Second step: Each subset in turn is used for testing and thr remainder for training.
  - The advantage is that all the example are used for both training and testing.
  - 

Fig. 6.6.3

- Leave one out cross validation
  - If dataset has N samples then N-1 examples are used for training and remanining example for testing.
  - $$E = \frac{1}{N} \sum_{i=1}^{N} E_i$$
- Stratified ten-fold cross-validation
  - ten-fold cross validation is repeated ten times and finally the results are averaged based on the previous 10 results.
  - The estimate's variance get reduced due to stratification.