10-12-19

Introduction

Data:- collection of raw facts.

↓

processed data

un processed data

→ The data is in the form of textual, video, audio, number, pictorial

→ collection of information or collection of records is known as a file.

→ Database is systematic collection of data. Database support

Database - relations - Table - rows / cols.

Application software - users uses these applications

system software - In built - like compilers, translators, interpreters.

Warehouse:- repository of heterogineous types of data

↓ storing

→ Large amount of data we can stored in it.

warehouse - gb & Tb of data to be stored

In warehouse there is lot of data to be stored and retrive the data also.

↓ mining - extracting

ex: weka tool

bars graphs → audely data

Database

warehouse

cloud

patterns

Different types of data are

db of data
file of data
spread sheet
data mart
↓
data about data

db
spread sheet of data
relational data
↓
mart

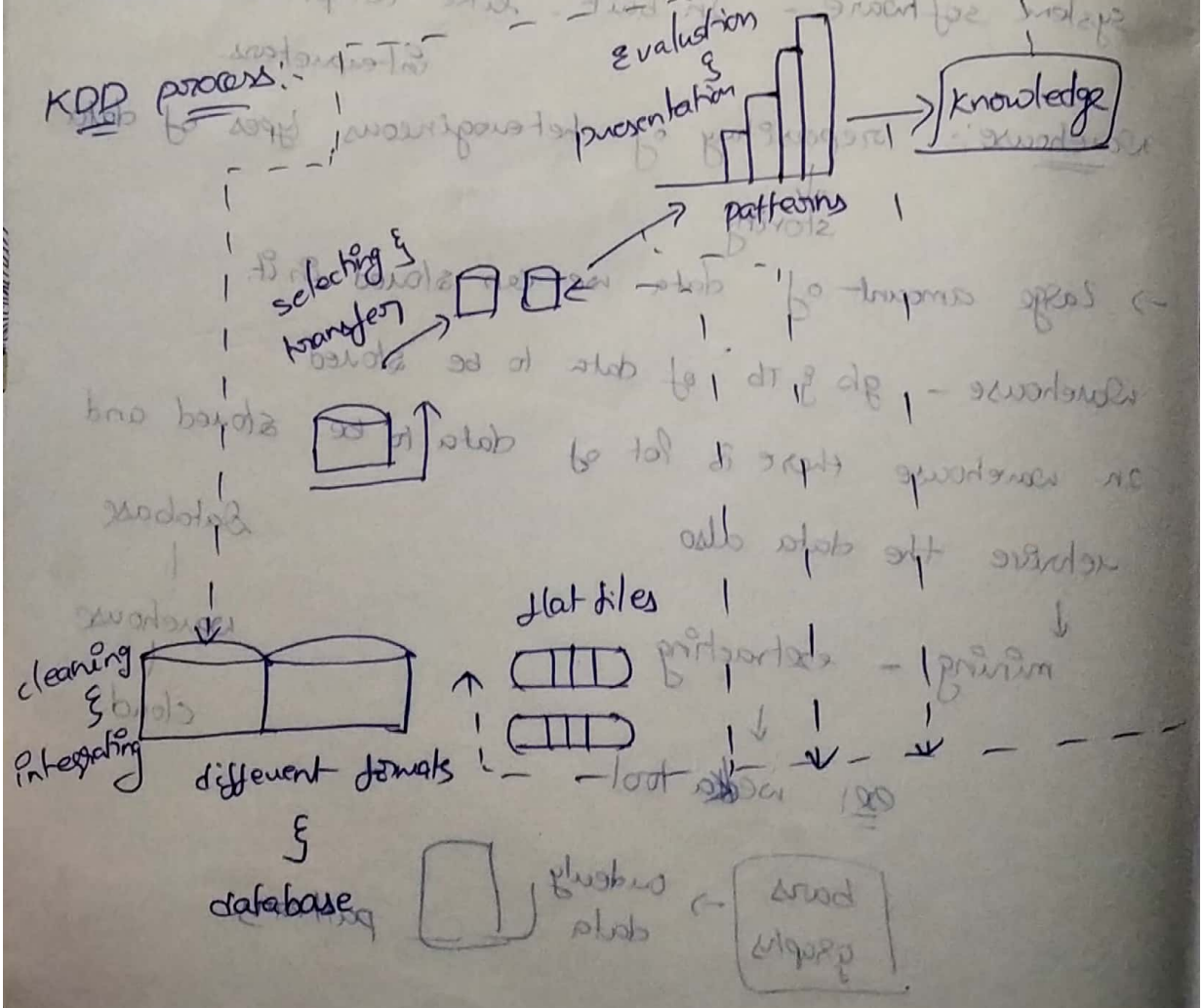
11-12

Unit 1 - Introduction to Datamining

1st topic:- knowledge discovery in database

→ extract the data is known as data mining
→ from the warehouse we can store data
→ warehouse - repository of heterogenous of types of data
file data, flat data, data mart, spread sheet type of data

KDD process:



1. flat

1. Data integration - combining of data it can store different types of data like data base, flat file, spread sheet and data mart. In data base we can store relational form of a data. A flat is also a per line we can take only one record is known as flat file. This flat file doesn't contain any relation to another type of file. table.

→ Data mart is nothing but sub-section of data warehouse

ex:- college → Data warehouse

2. Data cleaning - It remove the noisy, inconsistency, duplication. To remove these data using normalization, for the cleaning purpose.

S.No	name	course
1	A	c, c++
2	B	Java, python

1 A c
1 A c++
2 B Java
2 B Python

Like above table we remove the noisy & duplicate data with the use of 1st normal form.

3. Data selecting & transformation -

→ 90% the cleaned data should be selected for the transformation.

* For selecting the cleaned data we are using some techniques they are called pre-processing techniques.

The following are come under the pre-processed techniques they are 1. feature selection

2. Dymentionality reduction
3. Normalisation
4. Data subsetting

→ Transformation is nothing but convert the one type of format of data to another form

Ex:- dataset - collection of data

Data mining:- post processing classification cluster

Data mining is the process of mine the raw data i.e., post processing can be applied here i.e., on the transformed data we can apply the post processing techniques they are filtering patterns, visualization, pattern interpretation. Than finally from the visualization or from the pattern or from the evaluation the knowledge is an happen or the useful information is provided.

—o—

2nd topic:- motivating challenges in Data mining
↳ strength

The challenges are

1. scalability
2. High dimensionality Reduction
3. Heterogeneous complex data
4. Data ownership & distributed
5. Key challenges

1) scalability - scalable is nothing but big in size we can extract huge amount of data at datawarehouse we can't extract the massive data with help of previous algorithms.

scalability is nothing but one of the challenges faced by Data mining.

1. Originally the datawarehouse can accomidate the data in terms of GB & MB i.e., is in scalable manner (Big in size)
2. But the massive of data sets of data extraction

Is a challege faced by existing datamining algorithms like classification algorithm, clustering algorithm, Aggrigation algorithm

3. For these we need to develop scalable datamining algorithms to mine massive data sets

2. High dimensionality :-

dimensionality - Data, Dataset, attributes
                                  ↳ dimension

The types of data - qualitative, quantitative
                    ↓ character   ↓ number

→ Nominal & ordinary

→ qualitative - categorized data - classification

Nominal - categorized

Ordinal - un "

Numbers - Discrete, continous

Discrete - measuring the finite values

continous - measuring the infinite values.

a) Data set have the attributes
b) when the dataset contain the attribute it is called as dimension
c) At present we have the data sets with minimum number of attributes for data processing is possible but it is not supports the data set with 100's of 1000's of attributes for this it is the challenge faced by datamining due to this need to develop data mining algorithms to handle high dimensionality.
d) It uses low dimensionality i.e., it supports the data is in the form of continuous & categorical

3. Hetrogenous & complex data:-

* Traditional data analysis methods deal with data sets it contains same types of attributes.
* But the complex data set contain diff types of data i.e it supports image data, video data, text video.
* Need to develop Mining methods to handle complex data sets.

4. Data ownership & distribution:-

* Here sometimes the data needed for an analysis is not stored in one location (or) owned by one organisation instead the data is Geo-graphically distributed among resources belonging to multiple entities.

5. key challenges:-

* Scalable is one of the key challenge
* High dimensional is another " "
* Ownership & distribution is another "
* Hetrogenous & complex data " "

— o —

3rd topic The Origins of data mining

* The Origin of data mining is possible through the diff domains
* from the Statictics domain it can be happend from AI (Artificial Intelligence) it was happend from machine learning it was happend, from pattern recognition it way happend.

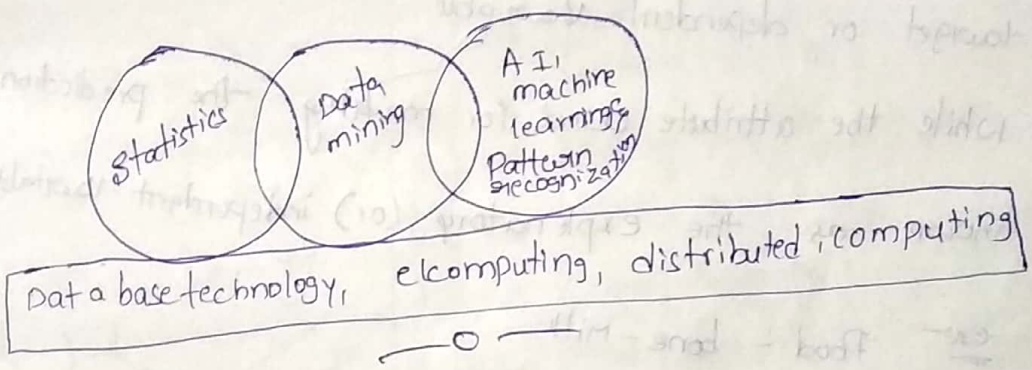
* Sampling, estimation & hypothesis testing form statistics from all of these the data mining coming to picture.

* Search algorthims, modeling techniques, learning theories from AI, (ML) & pattern recognition.
Machine language

* Data base systems are used for storage effective, index & query procesing effective, index & query processing effective storage.

* Paspalled computing is used for massive (Bulk amount) of some data sets address

* According distributed Computing gathers the info from various locations



4th topic Datamining Tasks:

Task meaning work

* They are two types of Datamining task

1. Predictive
2. Despritive

Predictive:- We can estimate the data /variable data may

Processed /mining has two types.

1. Preprocess
2. Post Process

Preprocessing:- We can remove the large, noisy data using Normalization is called Preprocessing.

Post processing:- We can select, transform the data we can appear some algorithms is called post processing.

* we can predict the value of particular attribute based on the value of other attributes
* An attribute can have the not a single value have the multiple values.
* The attribute to be predicted is commonly known as the target or dependent variable.
* While the attribute used for making the prediction or known as the explenatory (or) independent variables

ex:- Food - bone - Milk
Sound - bark - Mew

1.b

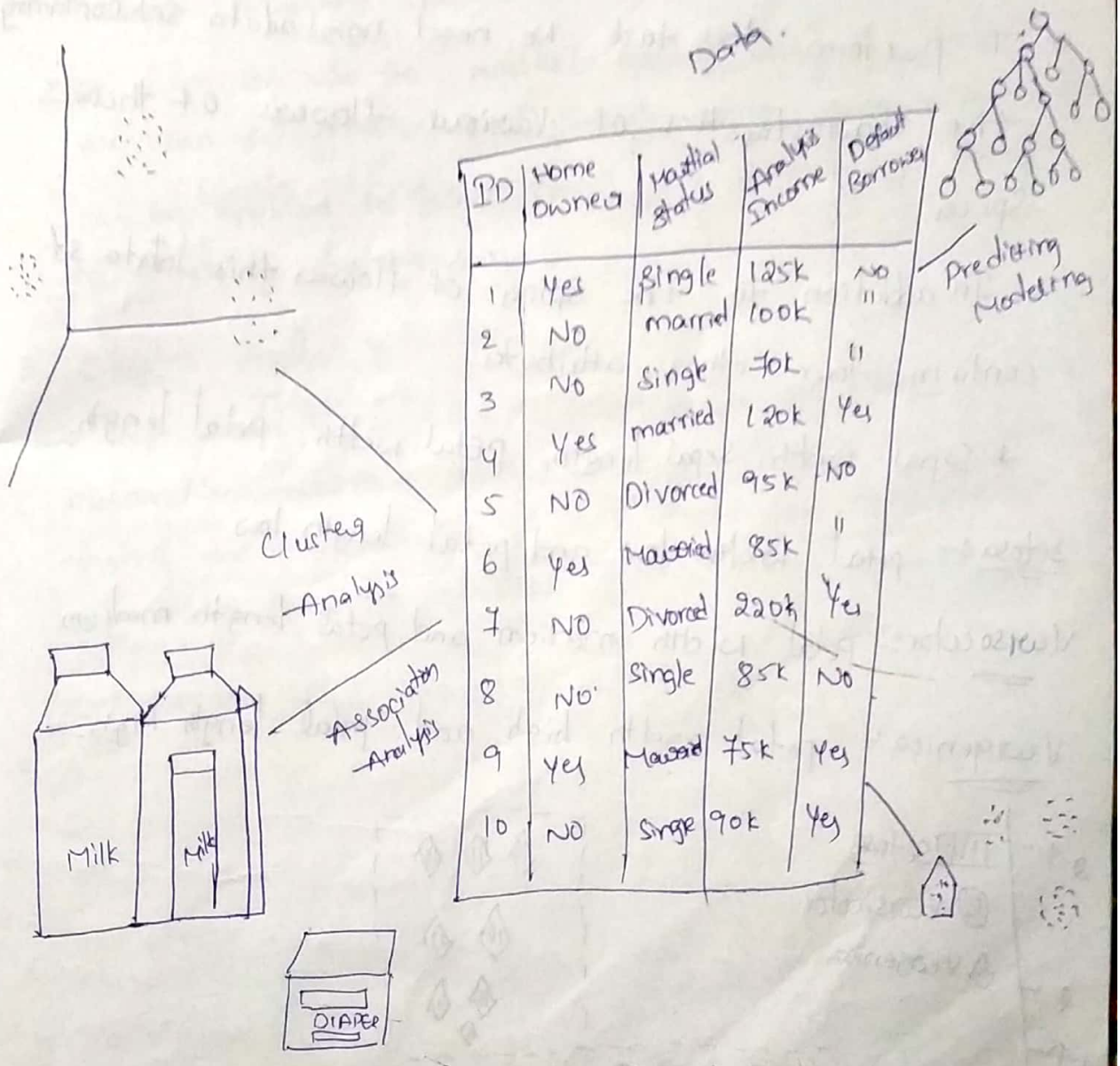
Descriptive:- It is nothing but desgived the pattern

* Pattern is nothing but some standard structure of the data.
* It is used to sumarize the underline relation in data.
* Wheather the data is in the pattern of pre-perceived The data is in the predective model. It can

arises the what type of data it can belongs

* we have furtherly four core data mining tasks they are

1. Predective modeling
2. Association Analysis
3. Clustering Analysis
4. Anamoly detection



Four of the core datamining Tasks

1. Predective Modeling:-

* Consider an "Iris" flower.
* Classifying the Iris flower whether it belongs to one of the following three Iris Spices.

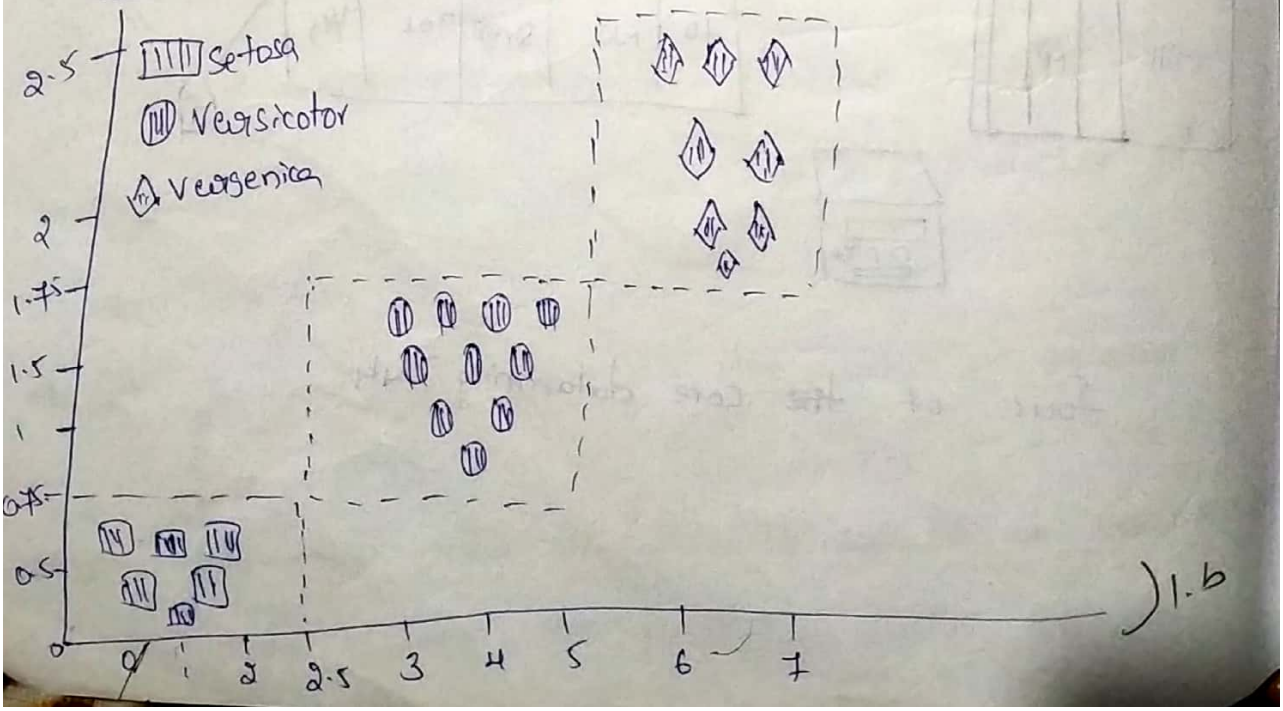
  Setosa, Versicolor and Verginica

* To perform this task we need need a data set containing the characteristics of various flowers of these 3 Spices.
* In addition to the Spices of flower this data set contain four other attributes
  * Sepal width, Sepal length, petal width, petal length.

Setosa:- petal width low and petal length low

Versocolor:- petal width medium and petal length medium

Verginica:- petal width high and petal length high.



2. Association Analysis :-

→ Association Analysis is used to discover patterns (order) that describes strongly associated features in the data

→ The discovered patterns are typically represented in the form of rules

→ The goal of association analysis is to extract the most interesting patterns in an efficient manner

Ex:- we can use the market basket analysis. Here we can concentrate the grocery store association analysis can be applied to find items are frequently bought together by customers.

3. Cluster Analysis :-

→ This analysis is used to find groups of closely related observations so that observation that belongs to the same cluster are more similar to each other then observations that belong to other clusters.

Ex:- we use the document clustering.

Market basket data

Transaction id	Items
1	{Bread, Butter, Diapers, milk}
2	{coffee, sugar, cookies, salmon}
3	{Bread, Butter, coffee, Diapers, milk, eggs}
4	{Bread, Butter, salmon, chicken}
5	{Eggs, Bread, Butter}
6	{salmon, Diapers, milk}
7	{coffee, sugar, chicken, eggs}
8	{Bread, diapers, milk, salt}
9.	{Tea, Eggs, cookies, Diapers, milk}
10.	{Bread, Tea, sugar, Eggs}

collection of news articles

Articles	words
1.	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2
2.	machinery: 2, Labour: 3, market: 4, industry: 2, work: 3, country: 1
3.	Job: 5, Inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3
4.	domestic: 3, forecast: 2, gain: 1, market: 2, sales: 3, rule: 2
5.	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2
6.	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3

4. Anamoly detection:-

→ Anamoly means abnormal that is not in a proper way. we can credit the anamolies of the data.

→ Anamoly detection is one of the core data mine tasks. The task of this detection is identifying observations whoose characteristics are significantly different from the remaining data such observations are known as anomonys (or) outlayers

Ex:- credit card fraud Detection.

A credit card company records the all the transactions made by every credit card holder, along with pesonal information such as credit limits, age, annual income, and address.

→ Anamony detection technique used can be applied to build a profile of legitimate (true) transactions for the users than any new transactions arise it can be checked with the profile of true transactions.

→ The characteristics of the enabled transaction are very different with the profile it can be treated as outlier.

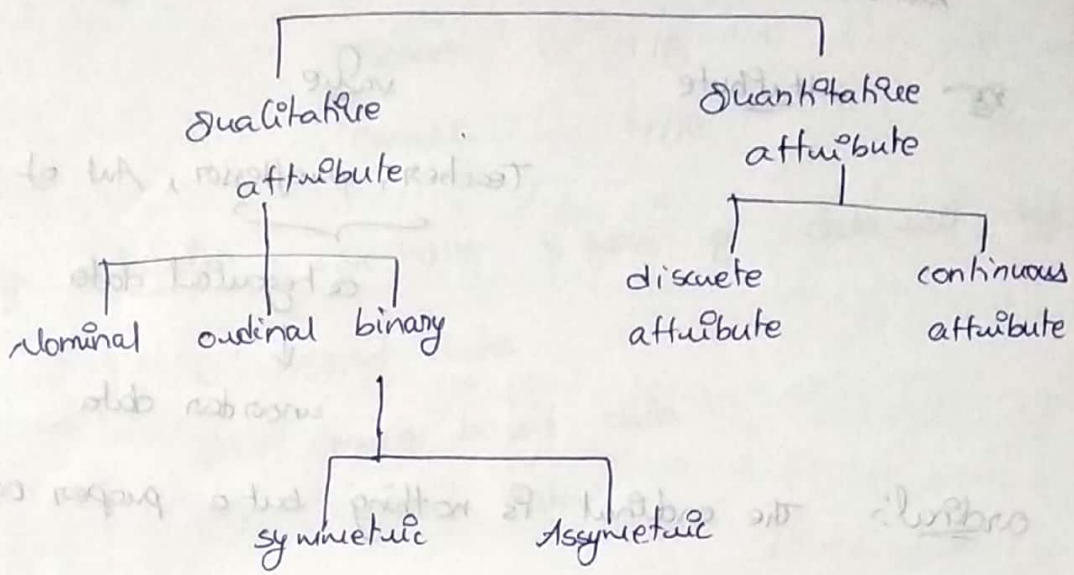
5th topic:- Types of data

1. Data sets differ in a number of ways the attributes used to describe the data objects can be different types
2. We have two types of attributes they are quantitative and qualitative attributes according to the data
3. A data set is nothing but collection of data objects

Different attribute types

Attribute Type		Description	Examples	operations
Categorical (qualitative)	Nominal	The values of a nominal attribute are just different names (ie, nominal values provide only enough information to distinguish one object from another ($=, \neq$)	Zip codes, Employee id numbers, eye color, gender	mode, entropy, contingency, correlation, χ^2 test
	Ordinal	The values of an ordinal attribute provide enough information to order objects ($<, >$)	Hardness of minerals, {good, better, best} grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric (quantitative)	Interval	For interval attribute, the difference btw values are meaningful, ie, unit of measurement exists ($+, -$)	calendar dates, temperature in celsius or fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
	Ratio	For ratio variables both differences and ratios are meaningful ($*, /$)	temperature in kelvin, monetary, quantities, count, age, mass, length, electrical current	geometric mean, harmonic mean, percent, variation

→ Types of attributes :- The attributes are classified into two types



Qualitative attribute :- Qualitative measures something based up on the state.

Quantitative attribute :- The Quantitative measures with the special characters of the data.

Discrete attribute :- The attribute contains the finite value is called discrete attribute

continous attribute :- The attribute contains the infinite value is called continuous attribute.

Ratio :- It is nothing but the two values (a) two attributes b/w the ratio can be contained for example

attribute	values
height	4.5, 5.0
weight	50, 60.

Ration b/w two values ie., 4.5, 5.0 the mean, median and standard deviation are called b/w those ratios.

Nominal:- The nominal is the unordered. Nominal attribute can take the values not in proper order.

Nominal is also called categorical

Ex:- Attribute value

Teacher, professor, Asst of

categorical data

↓

unorder data

ordinal:- The ordinal is nothing but a proper order (or) proper sequence.

Ex:- Attribute values

SSD 1,2,3 - -

Rank 1,2,3 - -

Binary:- The binary attribute contains two values. These are true and false. The both values come under have same equal importance.

These two values are

(i) symmetric

(ii) Asymmetric

symmetric:- The symmetric values have both equal importance is called Asymmetric

Ex:- Attribute value

gender female/male

Asymmetric:- The both values not have equal importance is called Asymmetric.

Eg:-

attribute	values
Report	T/F
cancel	Y/N

→ Types of datasets:- we have 3 types of data sets they are

(i) Record data
(ii) Graph based data
(iii) ordard data

characteristics:- Generally to measured the data. The data belongs to which type otherwise the data belongs to which group whether it is record, grouped data (or) graph based grouped data (or) ordered, grouped data. The dataset can have certain characteristics.

The characteristics are

(i) Dimensionality
(ii) sparesity
(iii) Resolution

(i) Dimensionality:- The dimensionality of a data is the no. of attributes that the object is the dataset posesses.

(ii) sparsity:- some data set have the asymetric feature (asymetric feature nothing but it can have the two values its commanded the binary the attribute)

* most attributes of an object have values of non-zero because the non-zero values need to be stored and manipulated.

* the data mining algorithms work well only for sparse data.

(iii) Resolution: The data possibilities are obtained frequently at different level of resolution.

Ex: (i) Earth seems very an event at resolution of few meters but is a relatively smooth at resolution of lots of kelometers.

(i) The patterns in the data also depend on the level of resolution.

(a) Record data: (a) Record data

Tid	Refund	marital status	Taxable income	Defaulted Borrower
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	20K	No
7	yes	Divorced	220K	No
8	No	single	85K	yes
9	No	married	75K	No

Tid	Items
1	Bread, soda, milk
2	Beer, Bread
3	Beer, soda, Diaper milk
4	Beer, bread, Diaper, milk
5	soda, Diaper, milk

(b) Transaction data

projection of x load	projection of y load	Distance load	thickness
10.23	5.27	15.22	1.2
12.65	6.25	16.22	1.1
13.54	7.23	17.34	1.2
14.27	8.43	18.45	0.9

(c) Data matrix

	Team	coach	play ball	score	game	win	lost	time out	season
Doc-1	3	0	5	0	2	6	0	2	0	2
Doc-2	0	7	0	2	1	0	0	3	0	0
Doc-3	0	1	0	0	1	2	2	0	3	0

(d) document term matrix

Most of the data mining work assumes that the dataset is a collection of records (data objects). Each of which consists of fixed set of data fields (or) attribute.

The most basic form of record data shown in fig (a). There is no explicit relationship among records (or) data fields. But every record has the same set of attributes.

Record data usually stored either in flat files (or) in relational DB's.

Transaction data (or) market-basket data:-

Transaction data is a special type of record data. where each record / transaction involves set of items. The set of products purchased by a customer with the grocery store during one shopping trip constitute one transaction, this type of data is called market-basket data.

Transaction data is collection of set of items, that is viewed as set of records whose fields are asymetric attributes.

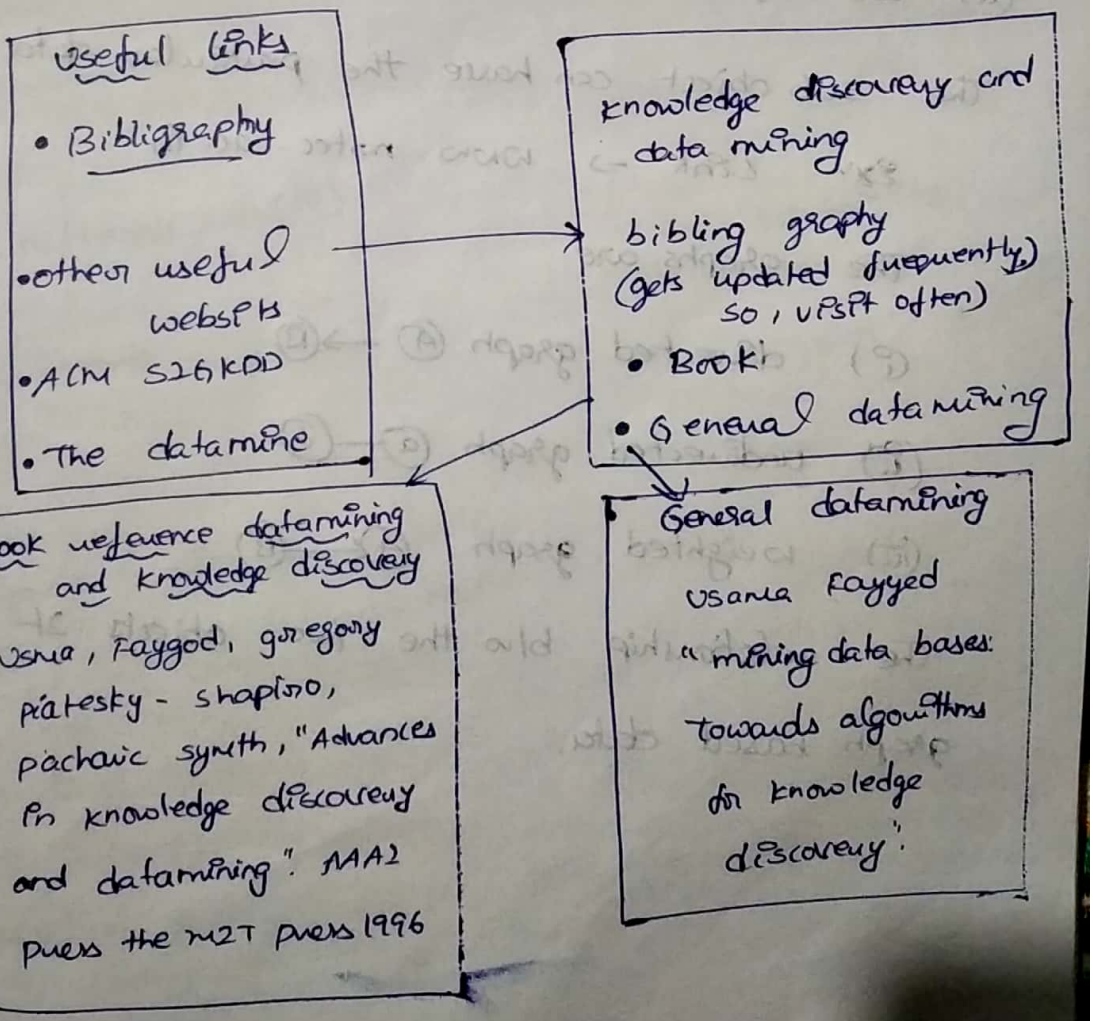
Data matrix:-

The data objects in a collection of data all have the same fixed set of numeric attribute the data objects can be thought of as points in multi-dimensional space.

A set of such data objects can be interpreted as an mxn matrix where there are m rows and n columns. This matrix is called data matrix or pattern matrix.

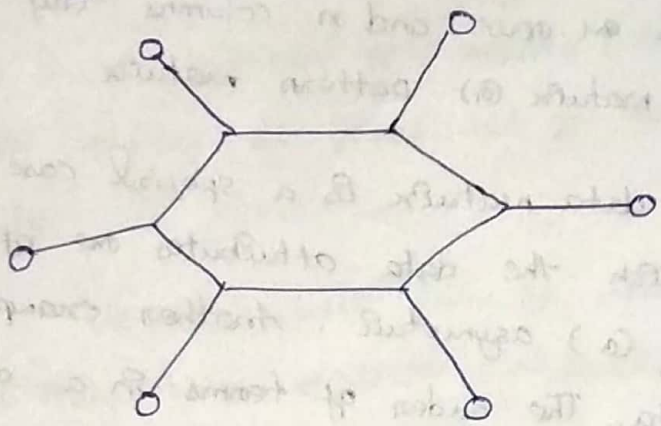
A sparse data matrix is a special case of data matrix. In which the data attributes are of the same type and (a) asymetric. Another example for this is document data. The order of terms in a document is ignored, then the document represented as a term vector where each term is a component is the no. of times of the corresponding terms occures in the document. This representation of a collection of a document is called document-term matrix.

Graph based data:-



Linking web pages

Benzene molecule:-



Graph based data:-

A graph is nothing but a collection of vertices & edges

Here we have the 2 classifications in the graph based data.

(i) The objects can have the relationships

(ii) we object the sub object in one object

The sub object can have the particular data

Ex: Link → www.notec.ac.in

The graphs are

(i) directed graph (A) → (B)

(ii) undirected graph (a) — (b)

(iii) weighted graph (A) ←— (B)

The relationship b/w the two objects. It is called graph based data.

23-1-20

Unit - III

Datawarehouse & OLAP technologies

OLTP - online transaction processing

OLAP - online analytical process

DW — subject oriented — supply, customer, product, sales — view model

heterogenous

Time variant — nonvolatile

seperation

OLAP & OLTD — users, Datacontent, Database, patterns

ER-model - OLTD - database designs

OLTP

ER-model

low level - users -

SQL - patterns - simple, short

tuple - records

OLAP

star schema, snow flake

top level - users

knowledge workers,

managers,

view - model the data of patterns

complex query - start net

tuple -

Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it. Data is read only and periodically refreshed

transaction process, recovery and concurrency control mechanisms

throughput - at a time more of works

it does not require

Differences b/w operational database systems & Data ware houses.

→ Comparision b/w OLTP & OLAP systems

Feature	OLTP	OLAP
characteristic	operational processing	Informational processing
orientation	Transaction	Analysis
User	clerk, DBA, database professional	knowledge worker, ex managers, executive analyst
Function	day-to-day operations	long-term information requirements. decision support
DB design	ER based, application oriented	star / snowflake, subject-oriented
Data	current, guranteed up-to-date	historical; accuracy maintained over time
summerization	primitive; highly detailed	summarized,
View	detailed, flat relational	Summerized, Multi-dimensional
unit of work	short, simple transaction	Complex Query.
Access	read/write	mostly read
Focus	data in	information out
Operations	Index/hash on primary key	lots of scan.
No. of records accessed	tens	Millions.
No. of Users	thousands	Hundreds.
Metric	transaction throughput	query throughput, response time.
priority	High performance, high availability	High flexibility end user autonomy

Multidimensional data model:-

Data ware houses and OLAP tools are based on a multidimensional data model.

→ This multidimensional data model is used to view the data in the form of data cube.

→ Data cube is nothing but allow the data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

→ Dimension is nothing but a entity. Entity is any real world objective. Here we take the entities with respective to an organization want to keep a record.

→ Each dimension may have a table associated with it called a dimension table. The dimension table further describes the dimension.

Ex: A table for item may contain the attributes item name, brand.

→ A multidimensional data model is typically organized around a central theame like sales.

→ This theame is represented by a fact table

→ Facts are numerical measures ie, the quantities by which we want to analyse relationships b/w dimensions.

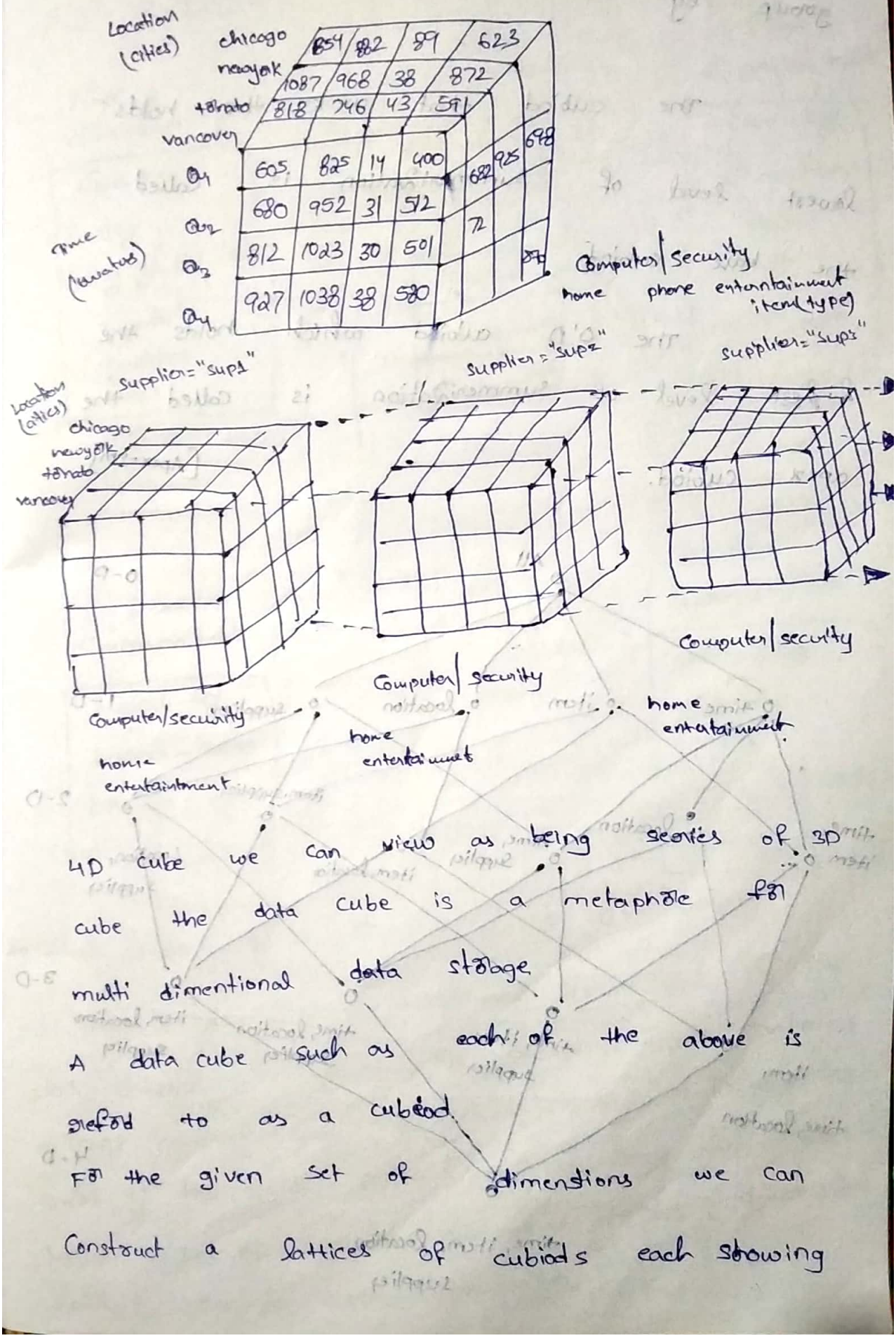
→ The fact table contains the names of the facts or measures as well as keeps to each of the related dimension table

A 2D view of sales data for all electronics according to the dimentions time and item. where the sales are from branches located in the city of Vancover the measure displayed in dollers - sold (in thousands)

Location = "vencover"				
	item type			
time (quarter)	home entertainment	Computer	phone	security
Q_1	605	825	14	400
Q_2	680	952	31	512
Q_3	812	1023	30	501
Q_4	927	1038	38	580

A 3D view of sales data for all electsonics, according to the dimention time, item and location. The measure displayed in dollars - sold (in thousands)

Location = "chicago" item				Location = "newyork" item				Location = "toronto" item				Location = "Vancover" item			
home				home				home				home			
Time	out	Comp	phone	sec	out	comp	phone	sec	out	Comp	phone	sec	out	Comp	phone	sec
Q_1	854	882	89	623	1087	968	38	872	818	746	43	891	605	825	14	400
Q_2	943	890	64	698	1180	1024	41	925	894	769	52	682	680	952	31	512
Q_3	1034	934	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q_4	1109	992	63	870	1142	1091	54	982	978	864	59	784	927	1038	38	580



4D cube we can view as being series of 3D cube the data cube is a metaphor for multi dimentional data storage

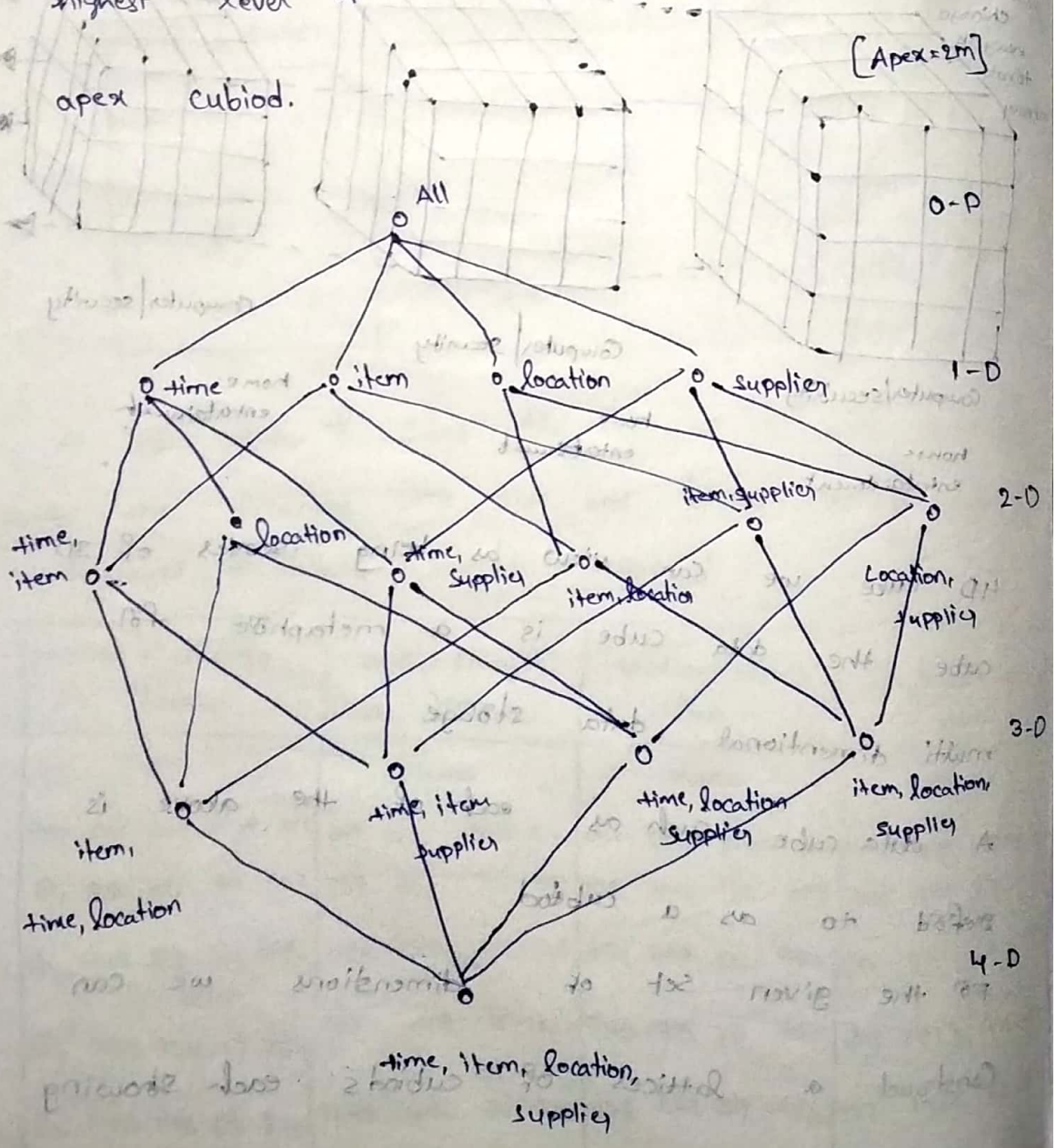
A data cube such as each of the above is refered to as a cubeod.

For the given set of dimentions we can Construct a lattices of cubiods each showing

the data at a different level of group by

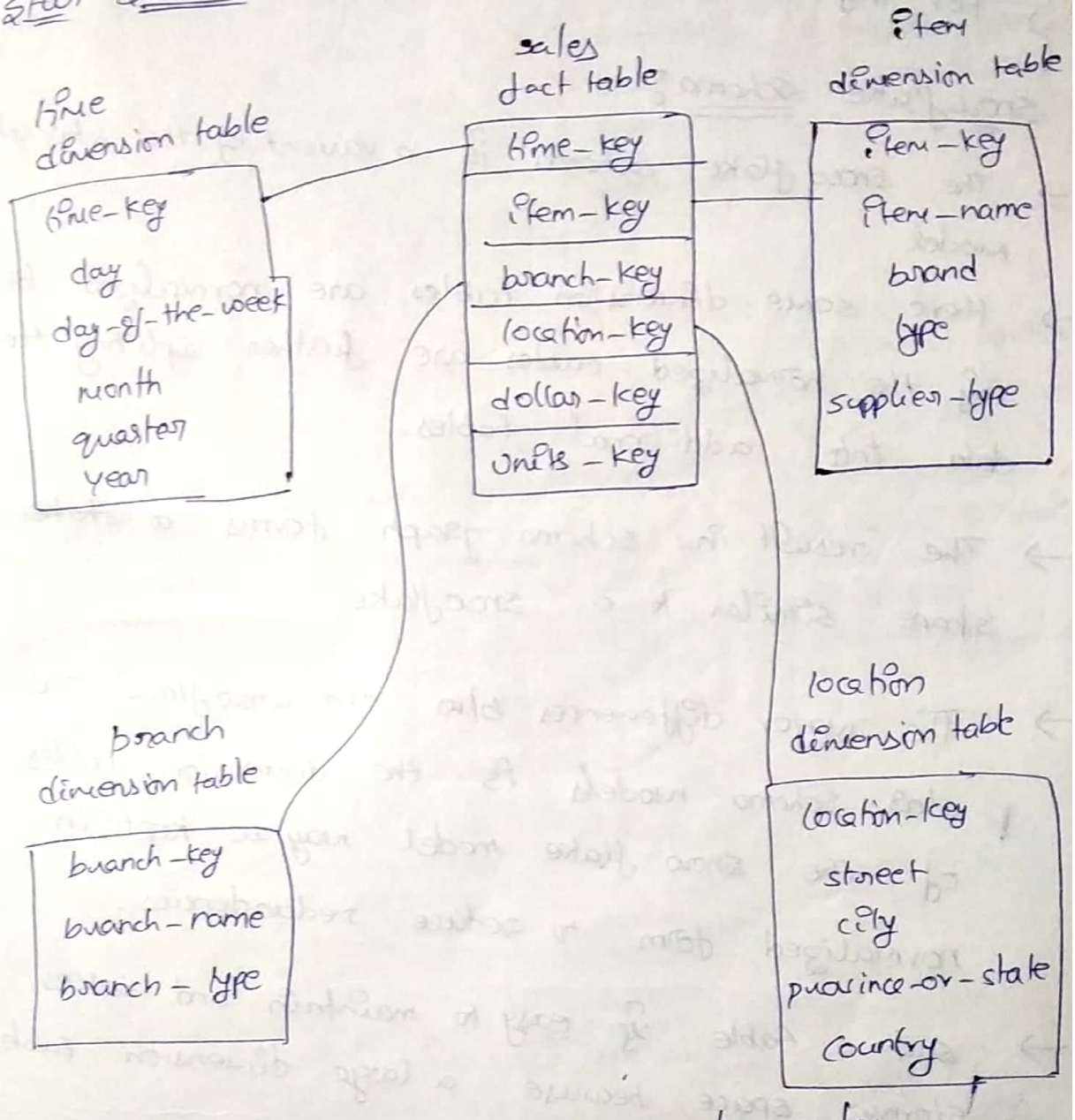
The cubiod that works the holds lowest level of summerization is called the base cubiod.

The 0'D cubiod which holds the highest level of summerization is called the apex cubiod.



→ star, snowflake and fact constellation schemas for multidimensional databases

Star schema :-



Eg. star schema of a datawarehouse for sales.

→ The most common modeling paradigm is the star schema in which the datawarehouse contains a large central table (fact table) (sales) containing the bulk of the data with no redundency and we have a set of smaller attendent tables (dimension tables), one for each dimension.

→ The schema graph resembles (appears) a star burst with the dimension tables displayed in a radial pattern around the central fact table.

→ For this reason they we call it as star schema.

Snowflake schema:-

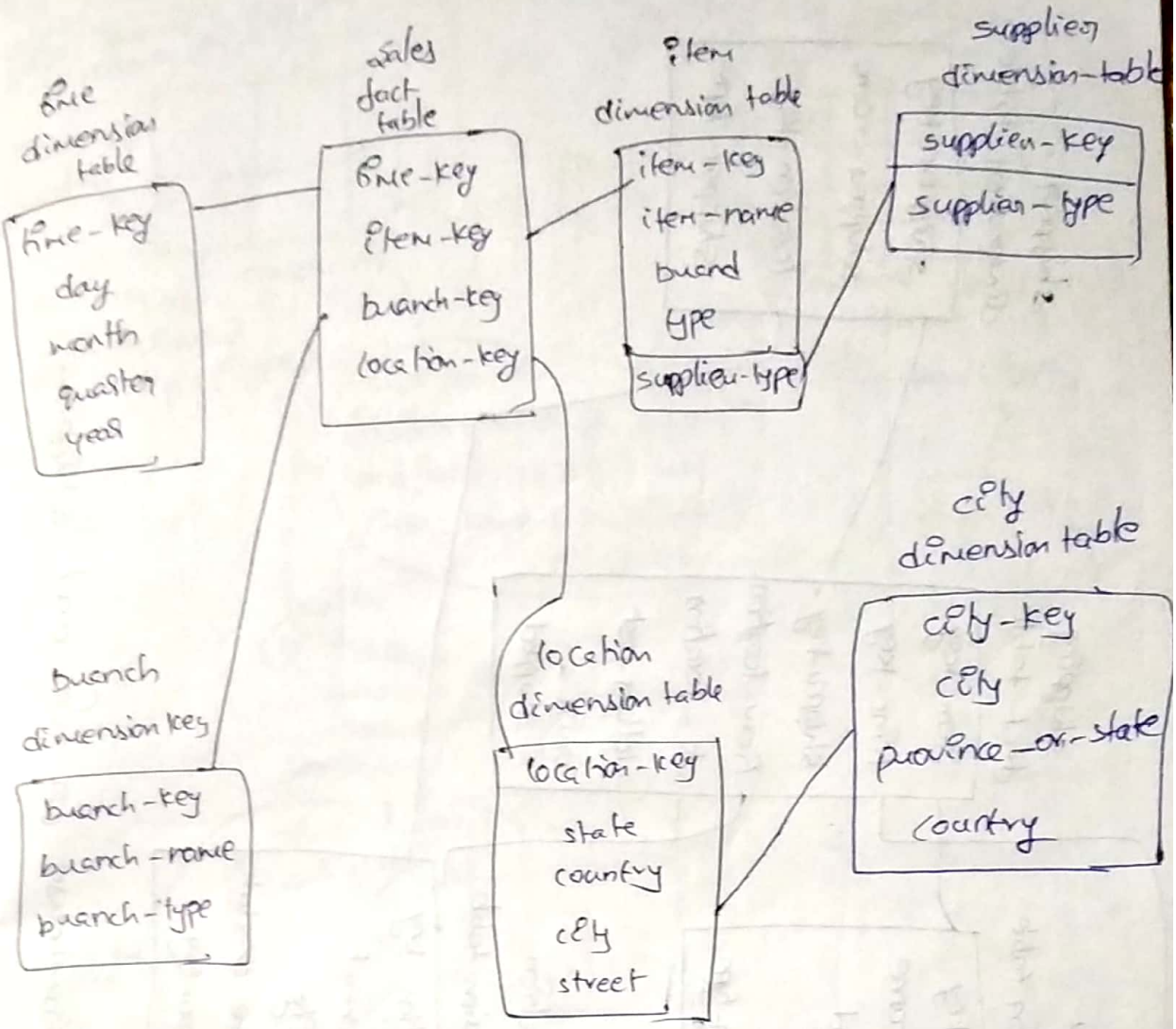
→ The snow flake schema is a varient of the star schema model.

→ Here some dimension tables are normalized for they the normalized tables are further splitting the data into additional tables.

→ The result in schema graph forms a ~~shake~~ shape similar to a snowflake.

→ The major differences b/w the snowflake and star schema models is the dimension tables of the snow flake model may be kept in normalized form to reduce redundencies.

→ such a table is easy to maintain and saves storage space because a large dimension table can be come anomous.

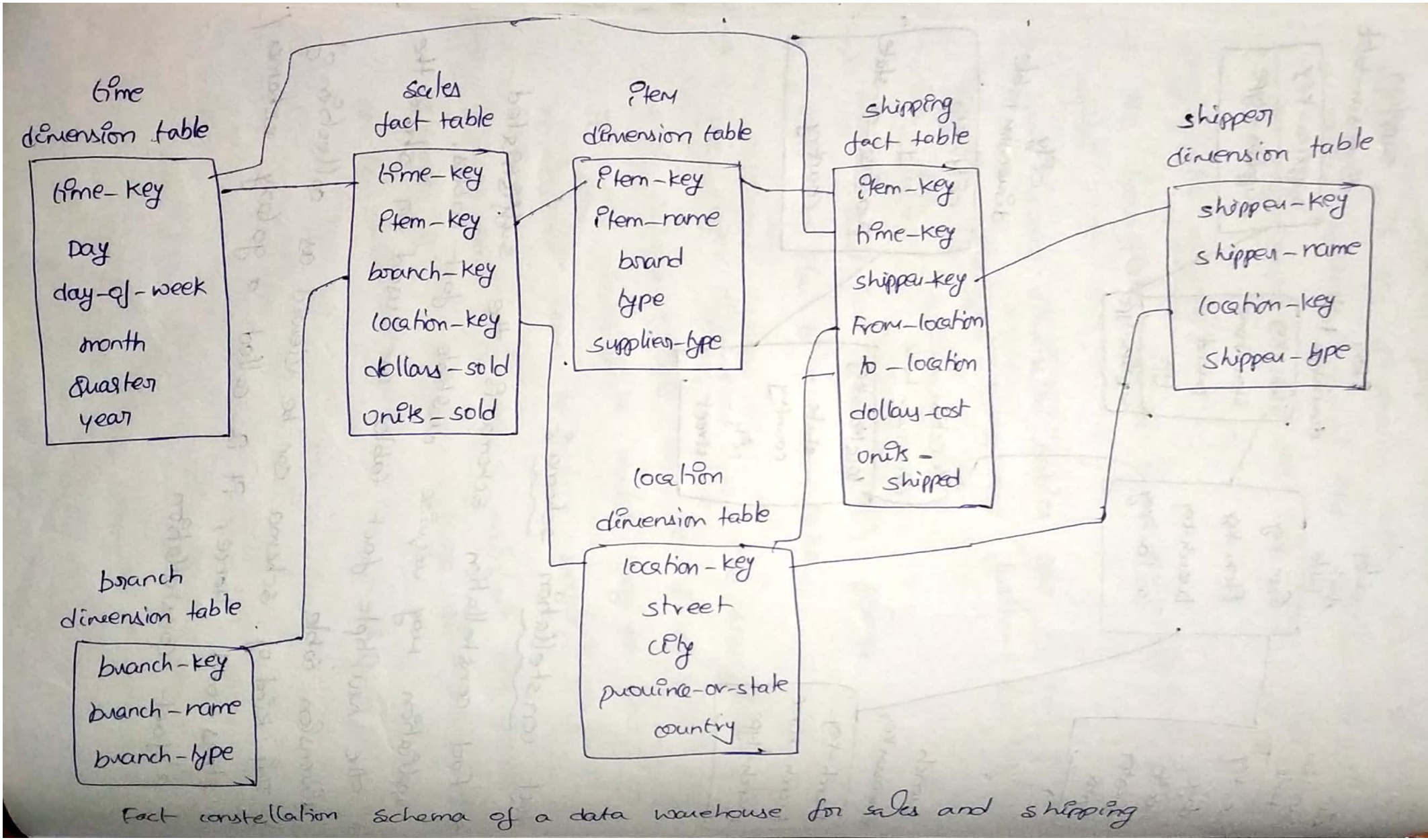


fact constellation schema :-

→ Fact constellation schema is the suffisticated application may require multiple fact tables.

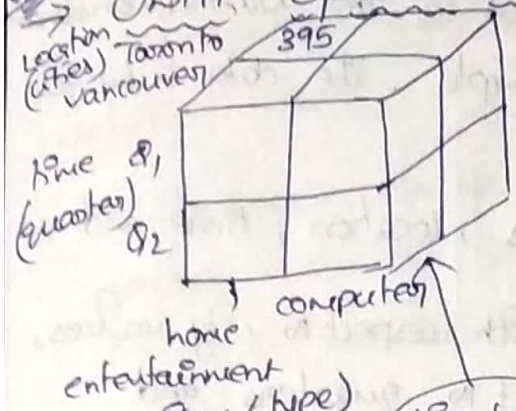
→ the multiple fact tables are used for share the dimension table.

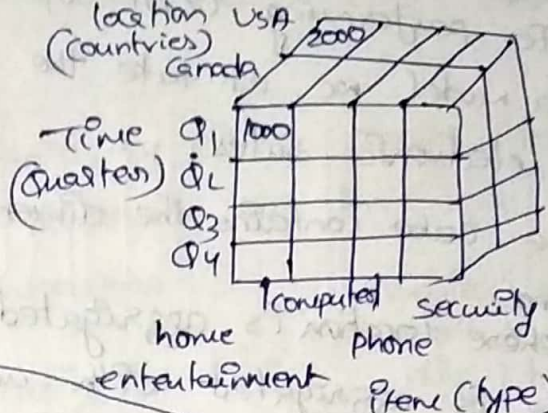
→ this kind of schema can be viewed as collection of stars and Hence, it is called a galaxy schema / a fact constellation



Fact constellation schema of a data warehouse for sales and shipping

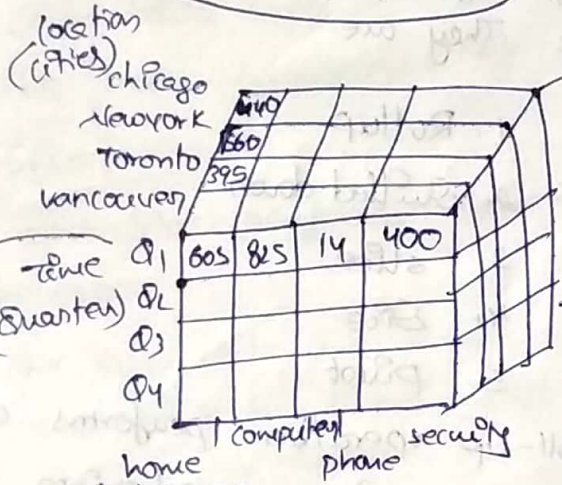
OLAP operations in multidimensional data model:-





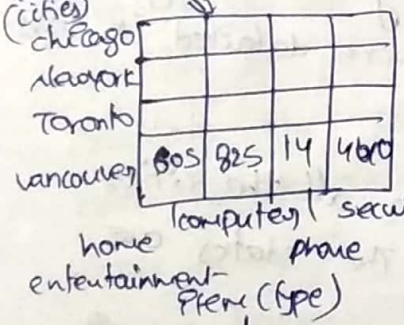
dice for (location = vancouver, Toronto) and (time = Q1 or Q2) and item = home entertainment / computer

Rollup on location from cities to countries

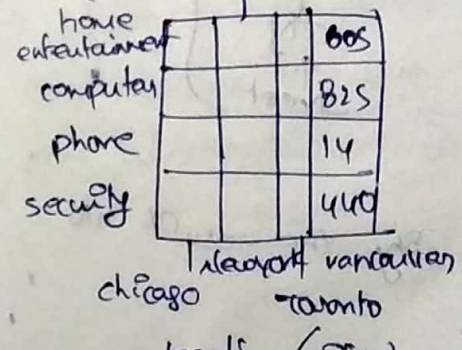


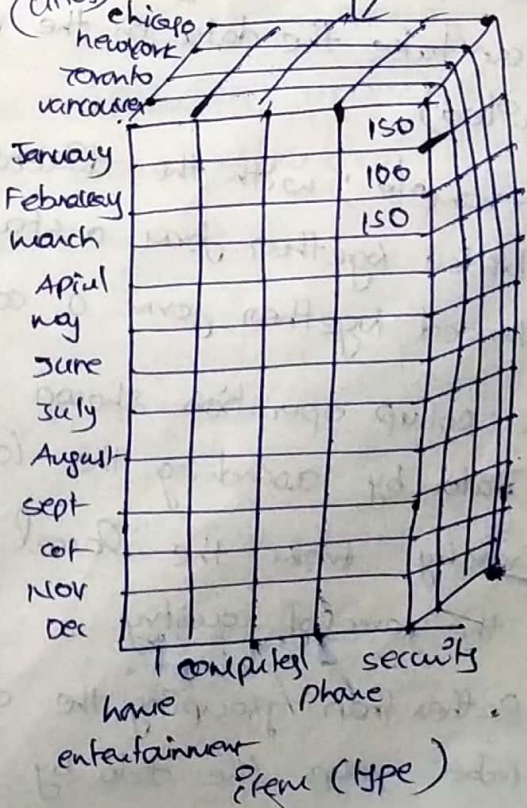
slice for time = Q1

Drilled down on time from quarters to month



pivot





Rollup:

→ For performing OLAP operations in the multidimensional data model, we can take the example. The data cube for all electronic sales.

→ The cube contains the dimensions, location, time and item.

→ where location is aggregated with respect to city values, time is aggregated with respect to quarters and item is aggregated w.r.t to item types. we can take the all electronic sales data cube treated as the central cube

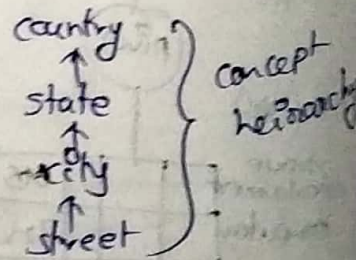
→ We have five types of OLAP operations performed on the central data cube They are

1. Rollup
2. Drilled down
3. slice
4. Dice
5. Pivot

1. Rollup:- The Roll-up operation performs aggregation on data cube either by using a mechanism climbing up a concept hierarchy (concept hierarchy is nothing but we can take the data in the form of more detailed to less detailed)

For example; with the dimension location all the cities combinded together form a state then, the states are combinded together form a country.

→ The Rollup operation shown aggregates the data by ascending the location hierarchy from the level of city to the level of country.



→ Rather than grouping the data by city the result into cube groups the data by country.

2. Drilled down:- Drilled down is the reverse of roll-up

→ It uses the mechanism (i.e. detailed data - more detailed data)

→ Drilled down can be replaced by either stepping-down concept hierarchy for a dimension or we can introduce additional dimensions

→ Drilled down uses descending the time hierarchy from the level of quarter to the more detailed level of month.

Quarter / month / day } stepping down hierarchy

→ The Resulting data cube details the total sales per month rather than summarized by quarter.

3. Slice & dice:-

→ The slice operation performs selection on one dimension of the given cube.

Example the sales data are selected from the central cube for one dimension "time" using the criterion time = "Q_1" as for the slice.

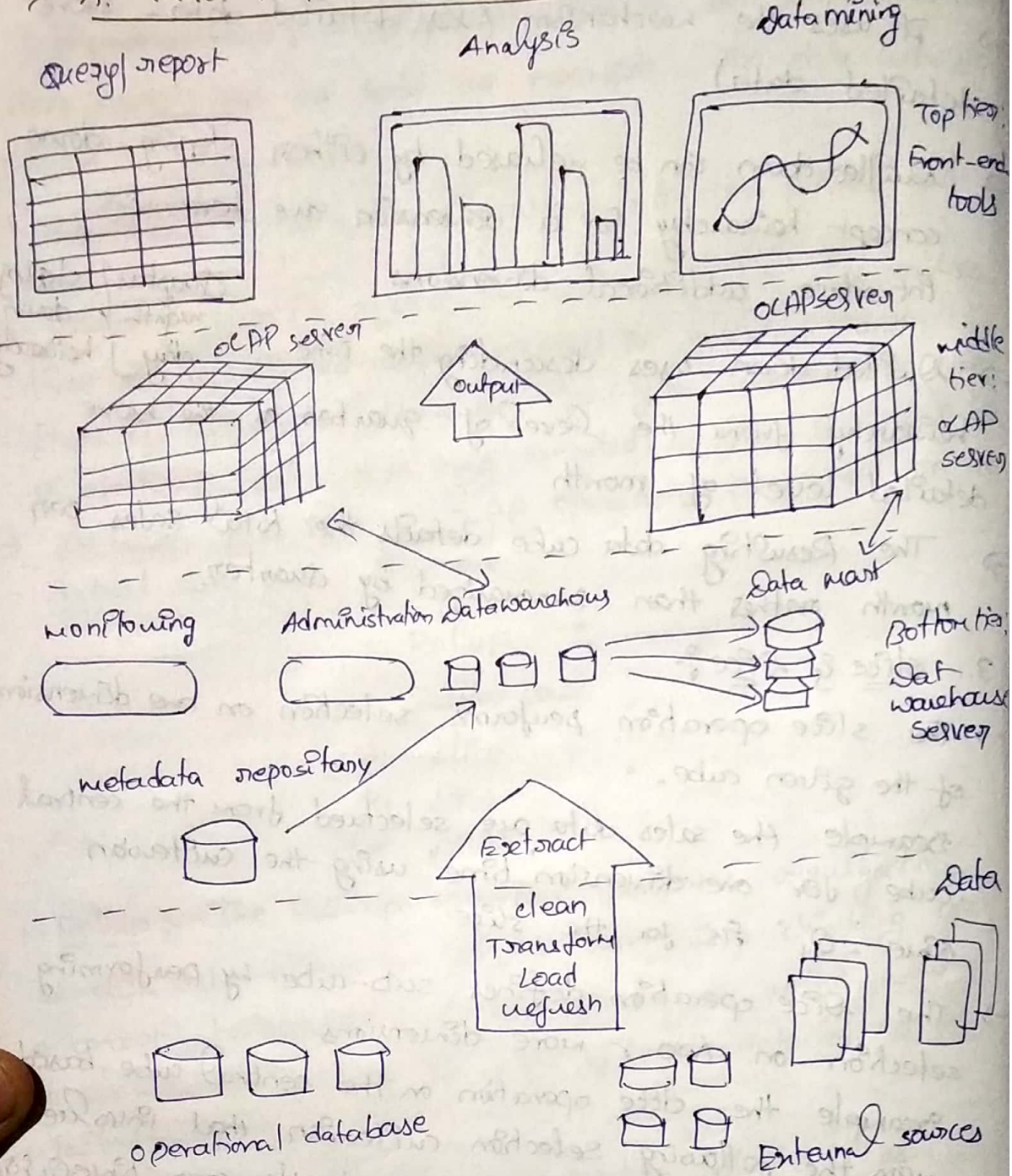
→ The dice operation defines sub-cube by performing selection on two or more dimensions

Example the dice operation on the central cube based upon the following selection criterion that involve three dimension Location = Toronto or Vancouver time = Q_1 or Q_2 item = home entertainment / computer.

4. pivot (rotate):-

→ pivot is also called a rotate is a visualization operation that rotates the data access in view in order to provide an alternative representation of the data.

→ A. three-tier datawarehousing architecture:-



→ steps for design and construction of data warehouse.

a) To design an effective datawarehouse 1st we need to understand business needs, and then construct the data warehouse belongs that need

b) we have 4 different views regarding the design of datawarehouse.

1. Top-down view :-

→ selection of the relevant information necessary for the datawarehouse

→ This information matches current and up-coming business need

2. Data-source view :-

→ The information being captured stored and managed is also we documented at various levels of detailed and accuracy from individual data source tables to integrated data source table.

3. Data-warehouse view :-

→ It includes fact and dimension tables the information is stored in the form these fact and dimension tables

4. Business-Query View :-

→ the data in the data warehouse from the view point of the end user it can be obtained

→ To perform all these views i.e, we design and construct data warehouse 2st of all choose a business process to model (waterfall model) choose the grain of the business process. The grain is nothing but the data to be represented in the fact table for this process.

→ choose the dimensions. The dimensions are applied to each fact table record.

Three-tier data warehousing architecture :-

Bottom-tier :-

→ It is the warehouse database server the data from the data source can take prior it can put into warehouse It can be extracted or cleaned using application program

interface known as gateway.

→ A gateway is supported by DBMS and allows client programs to generate sql code to be executed at a server.

Example. ODBC

middle-tier:-

→ It is an OLAP server is typically implemented using a ROLAP (Relational OLAP model) is an extended relational dbms that maps the operations on multidimensional data to standard relational operations.

→ It's also implemented with MOLAP (multidimensional OLAP). This model is a special purpose server that directly implements multidimensional data and operations

top-tier:-

→ Top tier is nothing but a client here we use front end tools ie., which contains query, reports and performing the analysis of data with the help of determining tools used.

→ Eg.

→ Types of OLAP servers :-

we have 4 types of OLAP servers they are

1. ROLAP
2. MOLAP
3. HOLAP
4. specialized SQL server

1. ROLAP :- (Relational online analytical processing)

→ It stands for ROLAP server. These servers are the intermediate servers.

→ why we call it as intermediate server ? It stands in b/w relational backend server and client front-end tools.

→ They use a relational or extended relational dbms to store and manage the warehouse data.

→ ROLAP servers include optimization for each dbms back-end, Implementation of aggregation and navigation logic and also provide additional tools and services.

→ The ROLAP technology have greater flexibility than MOLAP technology.

2. MOLAP :- (multidimensional OLAP)

→ It stands for multidimensional OLAP server.

→ These servers support multidimensional views of data through array based multidimensional storage in engine.

→ They map multidimensional views directly to data cube array's structure

→ with multidimensional data stores the storage utilization may be low if the data set is sparse.

→ many MOLAP servers adapt a tool level storage representation to handle spars and dense dataset.

3. HOLAP :- (Hybrid OLAP)

→ It stands hybrid OLAP servers.

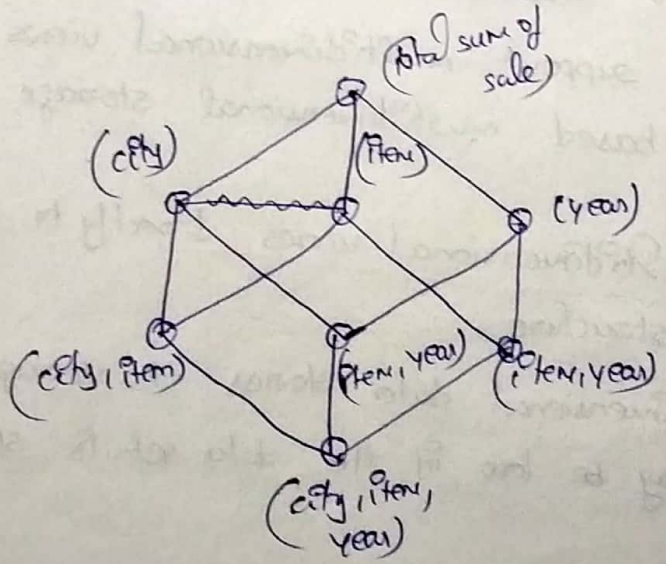
→ The hybrid OLAP approach combines the both the ROLAP & MOLAP technologies, benifiting the greater flexibility of ROLAP and the faster computations of MOLAP.

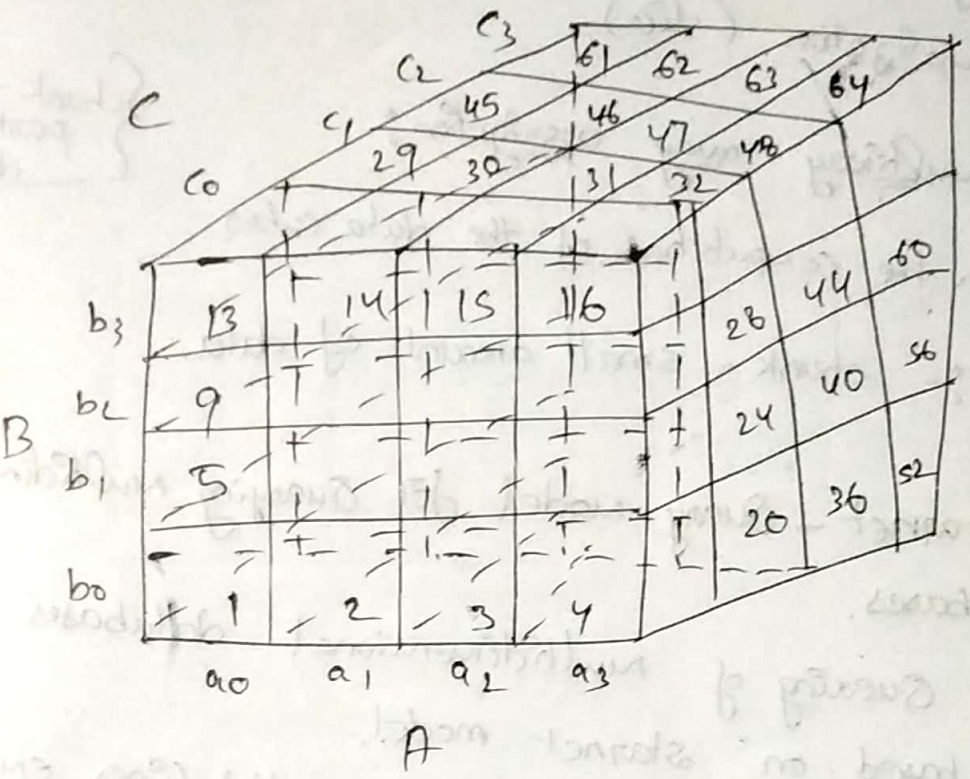
4. specialized SQL servers :-

→ For the growing demand of OLAP processing in relational databases.

→ some relational and dat warehousing firms (Ex:- Red brick of informix) implements specialized SQL servers that provide advanced query language and query processing support for SQL service over star & snaoflake schemas in a read only environment.

Data warehouse implementations :- (a) operations | view of Data warehouse and OLAP technology for data mining





⇔ starnet - Query model for Quering multidimensional databases ⇔

→ The data warehouse implementation can be happned with the help of efficient computations of data cubes

→ The example for this is the above lattice from this we compute the some of sales grouping by item & city - 2D, compute the some of sales grouping by only item 1-D, compute the some of sales grouping by city, item, year - 3-D.

→ partial materialization - From this selected computations of cuboids.

→ compute all of the cuboids is called as full materialization (dice)

→ compute any of the non-base cuboid we call as no materialization (apex)

→ selectively compute proper subset of the whole set of possible cuboids is called as partial materialization (slice).

→ multiway array aggrigation :- {chunk - part of data

→ In the computation of the data cubes

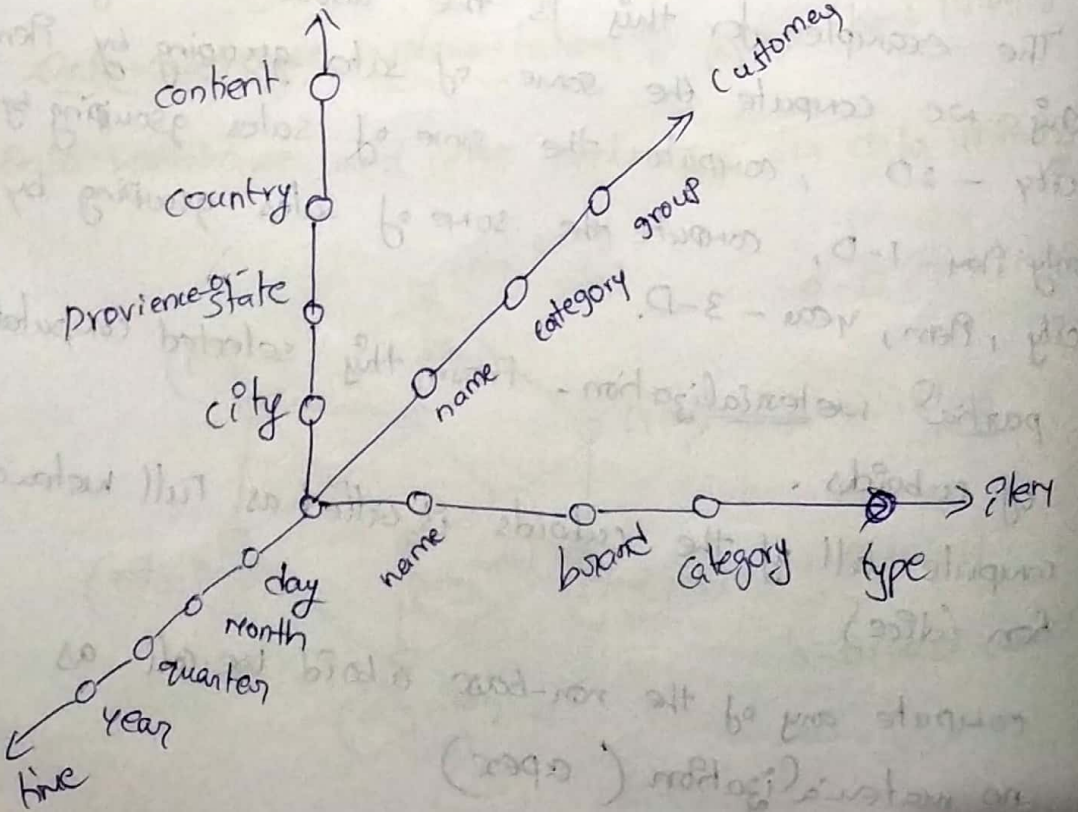
Note :- chunk - small amount of data.

→ starnet - Query model for Querying multidimensional databases.

→ The Quering of multidimensional databases can be based on starnet model.

→ A starnet model consists of radial lines emanating from a central point where each line represents a concept heirarchy for a dimension. each abstraction level in the heirarchy is called a footdwient

Footprint



Unit - IV

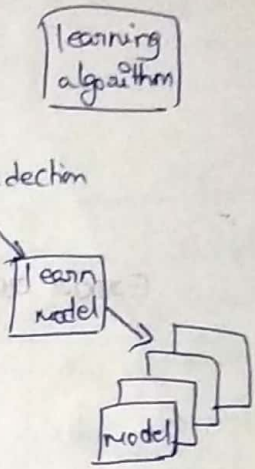
1.C

classification

Def: classification is nothing but process of mapping an attributeset 'x' to the 'y' its classification following nature

General approach

TID	attribute 1	attribute 2	attribute 3	class
1	Yes	large	125K	No
2	No	medium	100K	No
3	No	small	70K	No
4	Yes	medium	120K	No
5	No	large	95K	Yes
6	No	medium	60K	No
7	Yes	large	220K	No
8	No	small	85K	Yes
9	No	medium	75K	No
10	No	small	90K	Yes
11	No	small	55K	?
12	Yes	medium	80K	?
13	Yes	large	110K	?
14	No	small	95K	?
15	No	large	67K	?



confusion binary

		predicted class	
		class : 1	class : 0
Actual class	class : 1 yes	f11	f10
	class : 0 No	f01	f00

evaluation of the performances of the classification model is based on the counts of test records in

correctly predictly the model thy count or tableted in table is known is confusion matrise

confusion matrise for a binary classification problem

$$\text{Accuracy} = \frac{\text{no. of correct production}}{\text{Total no. of production}}$$

$$= \frac{f(11) + f(00)}{f(10)\,f(01)\,f(10)\,f(11)}$$

Descriptive - induction training
predictive - ref
deduction

$$\text{Error rate} = \frac{\text{no. of wrong production}}{\text{Total no. of production}}$$

$$\Rightarrow \frac{f(01) + f(10)}{f(01)\,f(01)\,f(10)\,f(11)}$$

Attribute set (X) $\Longrightarrow$ [classifier model] $\Longrightarrow$ class label (Y)

classification def: classification is a task of learning target function f that mars each attribute set X to one of the predefined class labeles Y.

→ The target function is also know informally is a classification model.

→ The input data for classification collection record.

→ Each record also known in instances is characterise by a tuple (x, y) where x is attribute set Y

specilized attribute disignted as the class label also known as target attribute.

→ descriptive is nothing but training (a) induction.

training - having class labels

induction - "

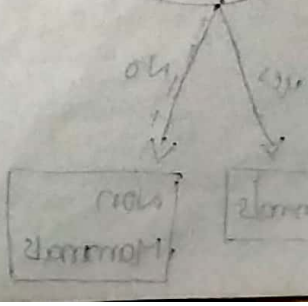
→ predictive is nothing but test (a) deduction

deduction/test - having no class labels, no attributes)/c

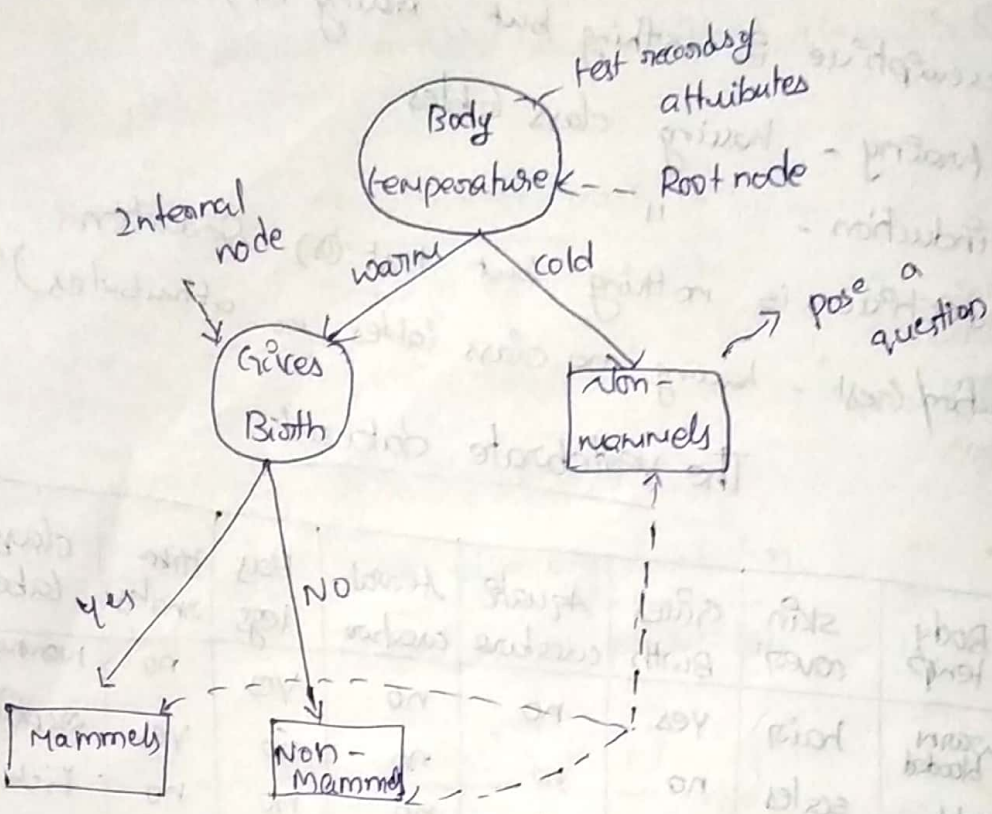
The vertebrate data set.

C 2.9

Name	Body temp	skin cover	Gives Birth	Aquatic creature	Aerial creature	Has legs	Hiber nates	class label
Human	warm blooded	hair	yes	no	no	yes	no	mammel
python	cold	scales	no	no	no	no	yes	reptile
salmon	cold	"	no	yes	no	no	no	Fish
whale	warm	hair	yes	yes	no	no	no	mammel
frog	cold	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold	scales	no	no	no	yes	no	reptile
bat	warm	hair	yes	no	yes	yes	yes	mammel
pigeon	warm	feathers	no	no	yes	yes	no	bird
cat	warm	Fur	yes	no	no	yes	no	mammel
leopard shark	cold	scales	yes	yes	no	no	no	Fish
turtle	cold	scales	no	semi	no	yes	no	reptile
penguin	warm	feathers	no	semi	no	yes	no	bird
porcupine	warm	quills	yes	no	no	yes	yes	mammel
eel	cold	scales	no	yes	no	no	no	Fish
salamander	cold	none	no	semi	no	yes	yes	amphibian



→ Best fit relationship b/w the attribute set and class label of the P/p data.

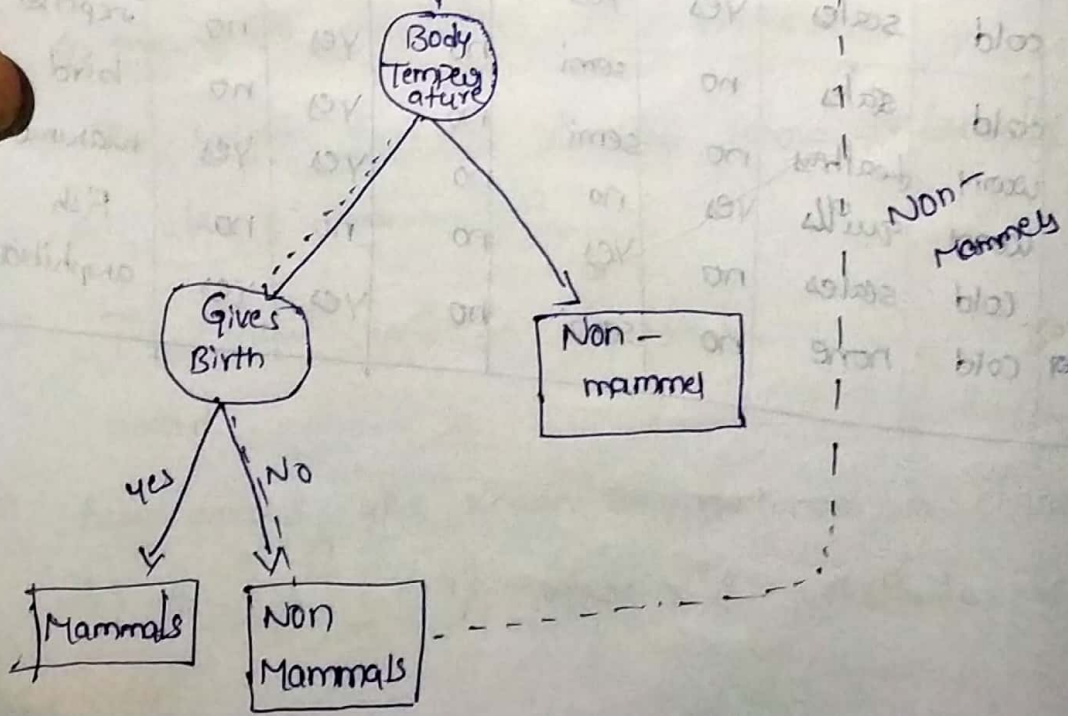


A decision tree for the Mammel classification problem.

How to build a decission tree

unabeled data

Name	Body Temperature	Gives Birth	..	Class
Flamingo	Warm	NO	..	?



classifying an unlabeled vertebrate. The dashed lines represent the outcomes of applying various attributes test conditions on the unabled vertebrate. The vertebrate is eventually assigned to the Non-mammel class.

2.b

2.b

Hunt's Algorithm:- How to build a decision tree with preparly we have the big dataset of construct more decision tree while some of the trees are more accurate than others. Finding the optimal tree is computationaly infusible. For this views the efficient algorithms of the exponential size of search space developed to include a reasonable accurate.

This algorithms usually employed greedy states that grows decision tree by making a series of localy optimum decisions about which attribute to use for partition the data. once such algorithm is hunt's algorithm. which is the basis of many existing decision three induction algorithm.

inducing 2i-3

Hunt's algorithm:-

In hunts algorithm decision tree is grown in recursive by partition by partioning the training records in to successfully purer subset

→ Let D_t be the set of training records that are

associated with node 't' is

$y = \{y_1, y_2 \ldots y_c\}$ be the class labels

step1:- If all the records in D_t belong to the same class 'y_t' then 't' is a leaf node labelled as 'y_t'

step2:- If 'Dt' contains records that belong to more than one class an attribute test condition is selected to partitioning the record into small subsets a child node is created for each outcome of the tests condition and record dt are distributed in the children based outcomes

step3:- The algorithm is recursively applied to each child node.) 2.b

3.a

→ (methods for expressing attribute test condition

Decision Induction algorithm must provide different methods for expressing an attribute test condition and its corresponding outcomes for different attribute types.

we have different attribute types. They are

1. Binary Attribute type
2. Nominal Attribute type
3. Ordinal Attribute type

4. Continuous attribute type

→ 1. Binary Attribute type :-

The test condition for a binary attribute type generates to potential outcomes.

2. Nominal Attribute type :-

The Nominal attribute can have many values its test condition can be expressed in two ways. — 1. multiway split
2. Binary split

3. Ordinal Attribute type :-

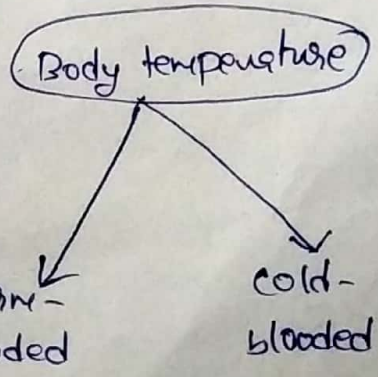
Ordinal Attributes can also produce binary or multiway splits.

Ordinal Attributes values can be grouped as along as the grouping does not violate the order propesty of the attribute values.

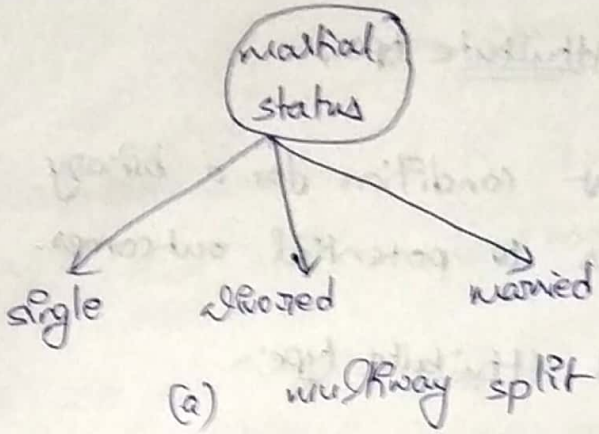
4. Continuous Attribute type :-

For a continuous attribute the test condition can be expressed as a comparision test (<, >) with binary outcomes or multiway outcomes.

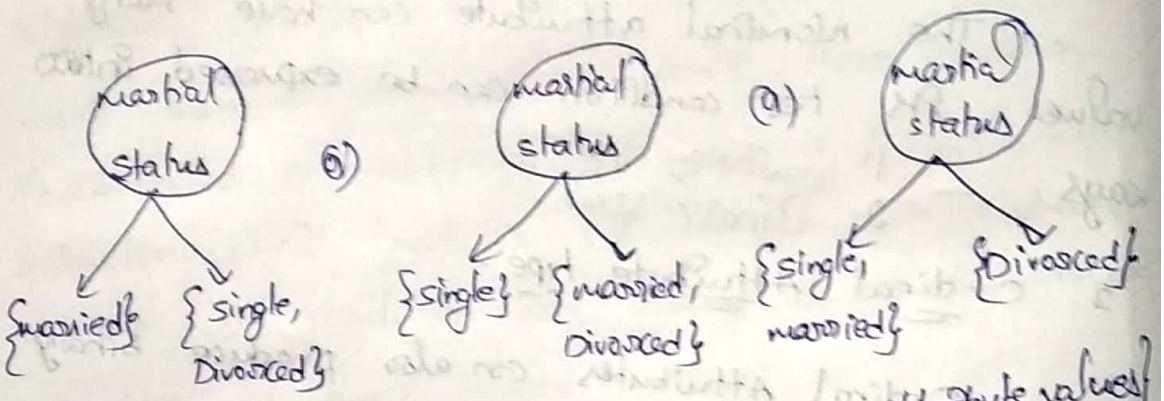
Test condition for binary attributes



Test conditions for nominal attributes

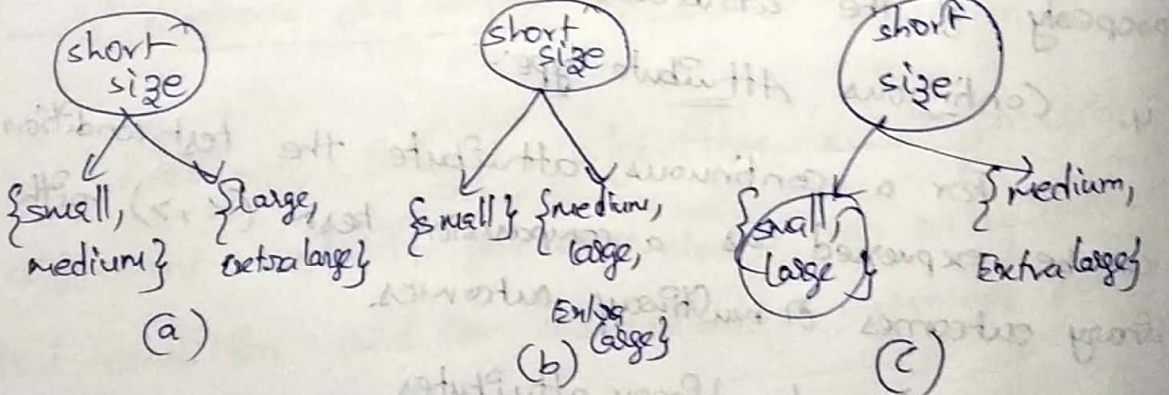


(a) multiway split

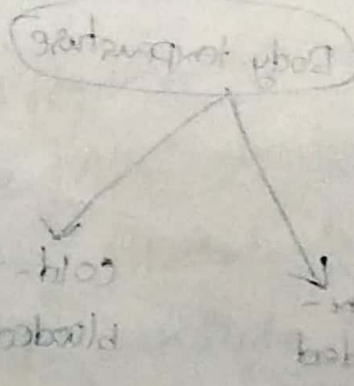


(b) Binary split {by grouping attribute values}

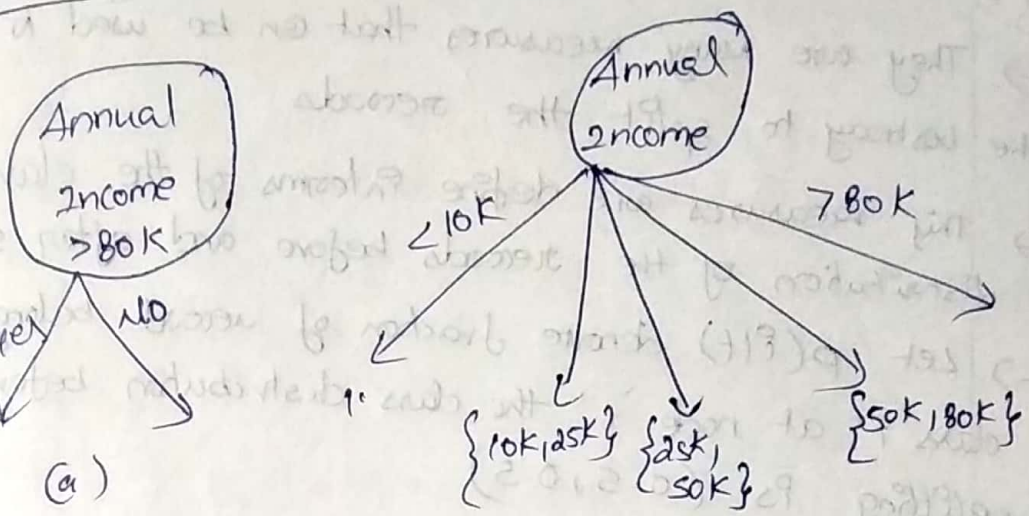
Different ways of grouping Ordinal attribute values



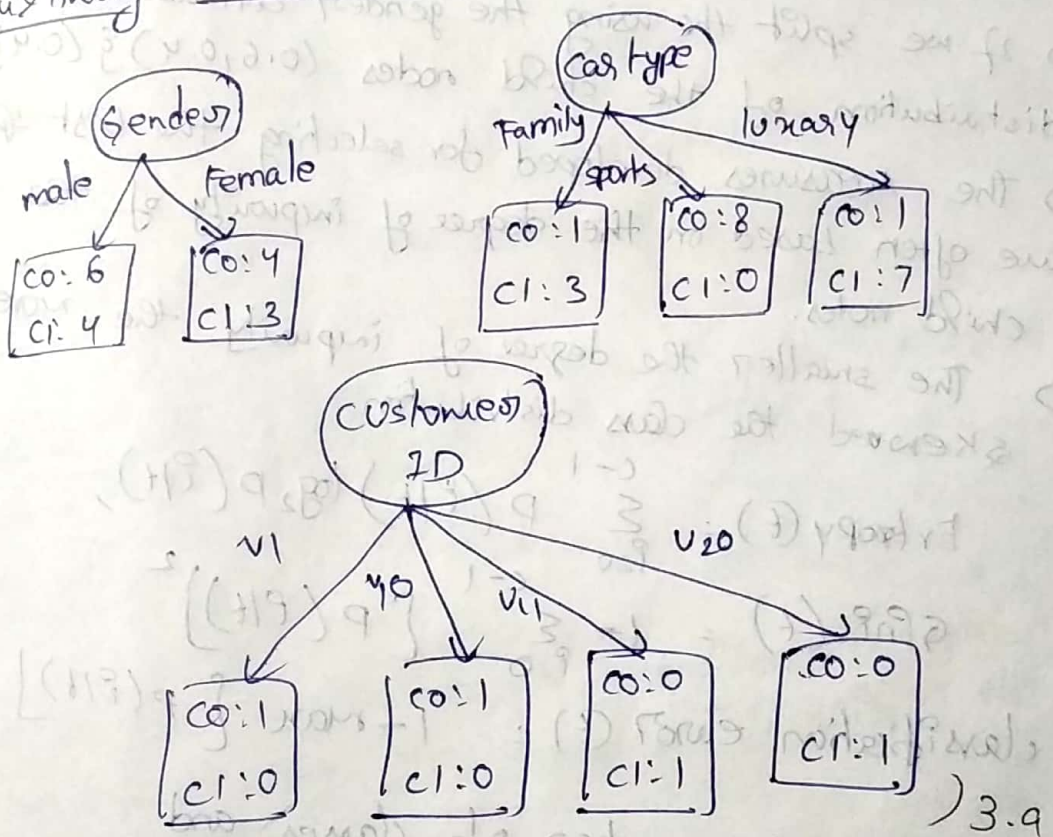
(a) (b) (c)



Test condition for continuous attributes



multiway verses binary splits



) 3.9

3) Measures for selecting the best split

→ They are many measures that can be used to determine the best way to split the records

→ This measures are define in terms of the class distribution of the records before and after splitting

→ Let $p(i|t)$ denote fraction of record belongs to class 'i' at node 't' the class distribution before splitting is (0.5, 0.5)

They are an equal number of records from each class.

→ If we split the using the gender attribute the class distribution of the child nodes (0.6, 0.4) & (0.4 & 0.6)

→ The measures developed for selecting the best split are often based on the degree of impurity of the child notes.

→ The smaller the degree of impurity the more skewed the class distribution.

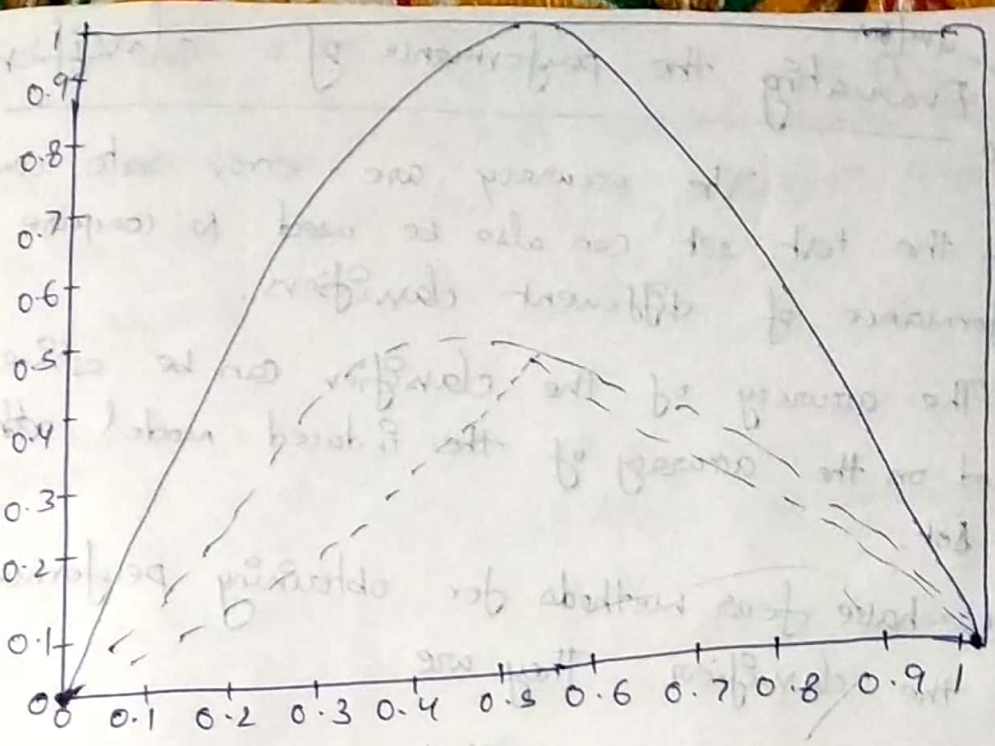
$$\text{Entropy}(t) = \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t),$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

$$\text{classification error}(t) = 1 - \max [p(i|t)]$$

where c is number of classes and

$0 \log_2 0 = 0$ in entropy calculations.



Node N_1	count
class = 0	0
class = 1	6

$Gini = 1 - (0/6)^2 - (6/6)^2 = 0$

$Entropy = -(0/6)\log_2(0/6) - (6/6)\log_2(6/6) = 0$

$Error = 1 - \max[0/6, 6/6] = 0$

Node N_2	count
class = 0	1
class = 1	5

$Gini = 1 - (1/6)^2 - (5/6)^2 =$

To determine how well a test condition performs we need to compare the degree of impurity of the parent node. (3.6)

→ Evaluating the performance of a classifier

The accuracy are error rate computed from the test set can also be used to compare the performance of different classifiers.

→ The accuracy of the classifier can be estimated based on the accuracy of the induced model with the test set.

→ We have four methods for obtaining performance of the classifier. They are

1. Holdout method
2. Random subsampling method
3. cross-validation
4. Bootstrap

1. Holdout method:-

→ In the Holdout method the original data with labled examples is partitioned into two disjoint sets called the training & test sets. A classification model is than induced from the training set and its performance is evaluated on the test set.

Limitations

→ The Holdout method has several limitations

1. Whether the training labled examples are fewer and some of the further records are with holded for testing.

2. Random subsampling method

→ The Holdout method can be repeated several times to improve the estimation of a classifier performance

→ This approach is known as R.S.M

Limitations

Has no control over the no. of times each record is used for testing and training

3. Corss - validation

→ An alternative to random subsampling method is the cross-validation.

→ In this method each record is used the same number of times for training and exeactly one's for testing

→ In this method we partioned the data into two equal sized subsets. Once they just we choose one of the subset for training and the other for testing.

→ We then swap the roles of subsets then the previous training set becomes the test set and the test set becomes training set. This approach is called a two fold cross validation.

4. ~~Bootstrap~~ Limitation

Here we have the chances of duplication of the records are possible.

4. Bootstrap:

→ In this method remove the duplicate records in the training and test sets.

→ In the bootstrap approach the training records are sampled without replacement i.e. A record already

chosen for training is put back into the original pool of the records, so that it is equally likely to be redrawn. u.a

u.b

→ Bayes' theorem :- Bayesian classification.

Dataset

car.no	color	Type	origin	stolen
1	Red	sports	DOM	Yes
2	Red	"	"	No
3	Red	"	"	Yes
4	Yellow	"	IMP	No
5	"	"	"	Yes
6	"	SUV	"	No
7	"	"	"	Yes
8	"	"	"	No
9	Red	"	"	No
10	"	"	"	Yes

Clearly we have the decision tree with the decision tree for the labled records of a particular entry contained in the dataset is easily we can make a decision.

→ Basically for every classification model we have the two classes only either yes / no, true / false.

→ Here we can't make a decision for the possibility of unlabled entries of a record.

→ To overcome this we can go and use bayesian classification.

→ With the bayesian classification we can make a decision for the unlabled entries also with the help of probabilities.

Probability

probability is nothing but getting the chance of outcome of possible-ness.

→ we have two types of probability. They are

1. normal probability
2. Conditional probability

1. normal probability - we have two boxes, named A, B. what is probability for selecting boxe A,

$P(A) = \frac{1}{2}$.

probability for selecting boxe B is $P(B) = \frac{1}{2}$

ex-2 we have - 3 boxes named as A, B, C. Probability of getting boxe A = $P(A) = \frac{1}{3}$.

probability of getting boxe B $P(B) = \frac{1}{3}$

probability of getting boxe C $P(C) = \frac{1}{3}$

2. Conditional probability - It is nothing but, Based upon the condition we can take the probability.

$P(X|A) = R|A$

we have boxe A it can have 5 Red balls, 6 black balls, 4 white balls. Here we can select the Red balls from the boxe A. $= \frac{5}{15}$

Bayes formula:

$$P(A|X) = \frac{P(X|A)\,P(A)}{P(X|A)P(A) + P(X|B)\,P(B)}$$

Normal probability:-

P(Yes) : 5	color, Type, origin
P(NO) : 5	

conditional probability

color : only Red & yellow

P(Red/Yes) = 3/5	P(Red/NO) = 2/5
P(yellow/Yes) = 2/5	P(yellow/NO) = 3/5

Type : suv & sports

P(suv/yes) = 3/5	P(suv/no) = 2/5
P(sports/yes) = 2/5	P(sports/no) = 3/5

origin : Domestic, imported

P(DOM/yes) = 2/5	P(DOM/no) = 1/3
P(IMP/yes) = 3/5	P(IMP/no) = 4/5

Now with the data set we can make a decision we can make a decision regarding the first entry. with the color red and type sports & origin is Domestic. we can have the decision with the class label stolen our answer is yes.

→ with this is also possible to make the decision

regarding the possible values of the sampling can't have the class label.

Ex: The sample

$P(X)$: <Red, suv, domestic>

$P(yes) = P(Red|yes) \cdot P(suv|yes)\, P(dom|yes) \cdot P(yes)$

$P(X|yes) \cdot P(yes) = P(Red|yes) \cdot P(suv|yes) \cdot P(dom|yes) \cdot P(yes)$

$= \frac{3}{5} \times \frac{2}{5} \times \frac{2}{5} = \frac{12}{125} = 0.096$

$P(X|NO) \cdot P(NO) = P(Red|NO) \cdot P(suv|NO) \cdot P(dom|NO) \cdot P(NO)$

$= \frac{2}{5} \times \frac{3}{5} \times \frac{1}{5} = \frac{6}{125} = 0.048$

The class label is stolen the belonged answer for the sample is yes.

4.5

Unit - V

Association Analysis : Basic Concepts &

Association rule mining finds intesting Algorithms association co-relation relationship.

→ find association rule from large amount of data.

→ Example of association rule mining is masket basket analysis.

→ It is a process of Analyses customer buying habits by find in Association b/w the different items that customer plays in their shopping basket.

→ To perform the association b/w the different items we have two measures they are

1- support
2- confidence

for example support is 2% means all the transactions under analysis show that purchased milk product & bread product togethes.

Confidence - 60%. means 60% of the customers who purchased milk also bought the bread.

support $A \Rightarrow B = P(A \cup B)$

confidence $A \Rightarrow B = P(B|A)$

→ Rules that satisfy both minimum support thrush hold and min confidence thrush hold are called strong rules.

→ Support and confidence vasiables occur b/w 0% and 100%.

→ A set of items is refered to as an item set. An item set that contains 'K' items is a k item set.

{milk, bread} – 2 item set

→ The occurence frequency of an item set is the no. of transactions that contain the item set. This is known as frequency, support count, count of item set.

If an itemset satisfies minimum support count then its a frequent item.

→ Association rule mining:-

Association rule mining is a two process that is it can have the two steps

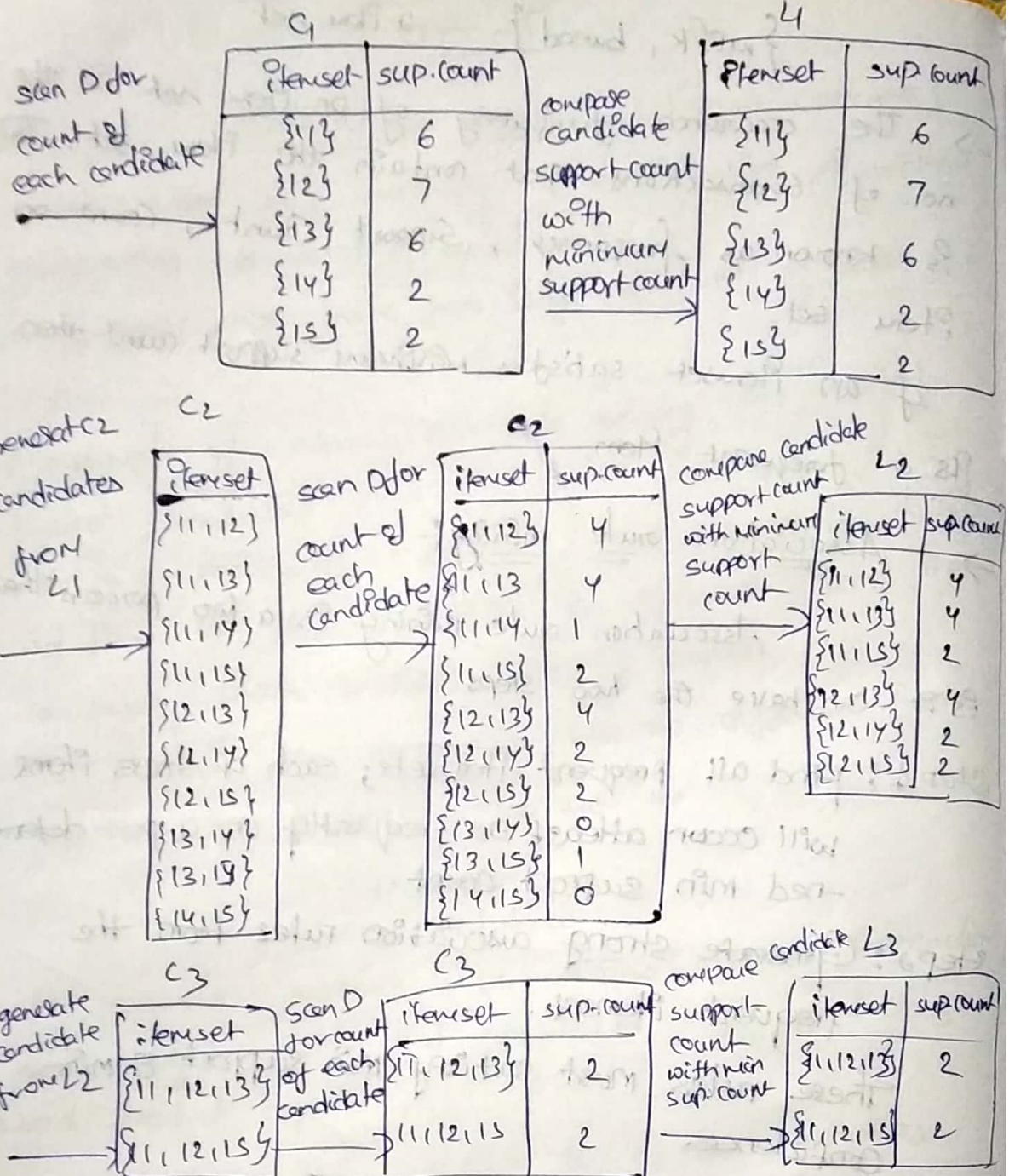
step 1: find all frequent itemsets; each of these items will occur atleast as frequently as a pre-determined min support count.

step 2: Generate strong association rules from the frequent itemset.

These rules must satisfy min support & min confidence.

5

TID	List of item-IDs.
T100	I1, I2, I5
T300	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T400	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3
T900	



Generate of candidate itemset and frequent itemset where min. support count is 2.)5.

The Apriori algorithm give uses two properties for find out frequent Item sets. We have the join and prune

Join:- Is used to find out the $L_1 \cdot - - L_K$ The set of frequent one item set, two item set with join operator themselves the $L_1, L_2 - - L_K$ are join

prune:- Prune using the apriori property all non empty subsets of a frequent item sets must also be frequent

process of Apriori algorithm:-

For performing the Apriori algorithm we have to follow two steps. They are the join step and prune step, with the help of the join step we prepare a set of candidate 'k' item sets, is generated by joining L_{K-1} with itself (∵ For example we can find the L_K $K = 1, 2 - -$)

In the prune step C_K is the super set of L_K $(C_{(1,2\cdots)})$ $(K = 1, 2 - -)$ i.e., its members may or may not frequent but all of the frequent K item sets are included in C_K.

Process:

step-1 - In the 1st iteration of the algorithm each item is member of the set of candidate one-item set C_1.
The algorithm simply scans all of the transactions in order to count the number of occurences of each item.

step2:- Suppose that the minimum transaction support count required two (minimum sup = $\frac{2}{9}$ = 22%.)

The set of frequent one-item sets L_1 can then be determine it consists of the candidate one item sets satisfying minimum support.

step-3 - To discover the set of frequent two item sets L_2 the algorithm uses $L_1 \bowtie L_1$ to generate a candidate set of two item sets C_2. C_2 consists of two item sets of $2^{|L_1|}$.

step-4 - The transaction in D (data set) are scanned and the support count of each candidate item set in C_2 in accumulated. (combinations)

step-5 From the accumulated we have the set of frequent two item sets L_2 is then determine consisting of those candidate two item sets in C_2 having minimum support.

step-6 After that the generation of the set of candidate 3 item sets C_3 1st we can take the C_3 then C_3 is calculated by performing $(L_2 \bowtie L_2)$ joining with itself. Finally the $C_3 = \{I1, I2, I3\}, \{I1, I2, I5\}\ \{I1, I3, I5\}$

$\{I2, I3, I5\}, \{I2, I3, I4\}$

step-7 Based on the apriori property that all subsets of frequent item sets must also be frequent. This can be possible with the help of a apriori prune property.

step-8 The transactions D are scanned in order to determine L_3 consisting of those candidate 3 itemsets having minimum support.

step-8 The algorithm uses $L_3 \bowtie L_3$ to generate Candidate set of 4 item sets C_4. The join result in C_4 is $\{I1, I2, I3, I5\}$ (4) $\{I1, I2, I3, I5\}$. This

Itemset 3 pruned. Since its subsets $\{I_2, I_3, I_5\}$ is not frequent. For this $C_4 = \phi$.

And the algorithm terminates having found all of the frequent item sets.

~~Apriori Al~~

TID	List of items-IDS
T_1	I_2, I_3, I_4
T_2	I_1, I_3, I_5
T_3	I_2, I_4
T_4	I_1, I_2, I_3
T_5	I_5
T_6	I_3, I_4, I_5
T_7	I_1, I_2, I_3
T_8	I_2, I_4, I_5
T_9	I_2, I_4, I_5

C_1

item	sup-count
I_1	3
I_2	6
I_3	6
I_4	5
I_5	5

L_1

item	sup.count
I_1	3
I_2	6
I_3	5
I_4	5
I_5	5

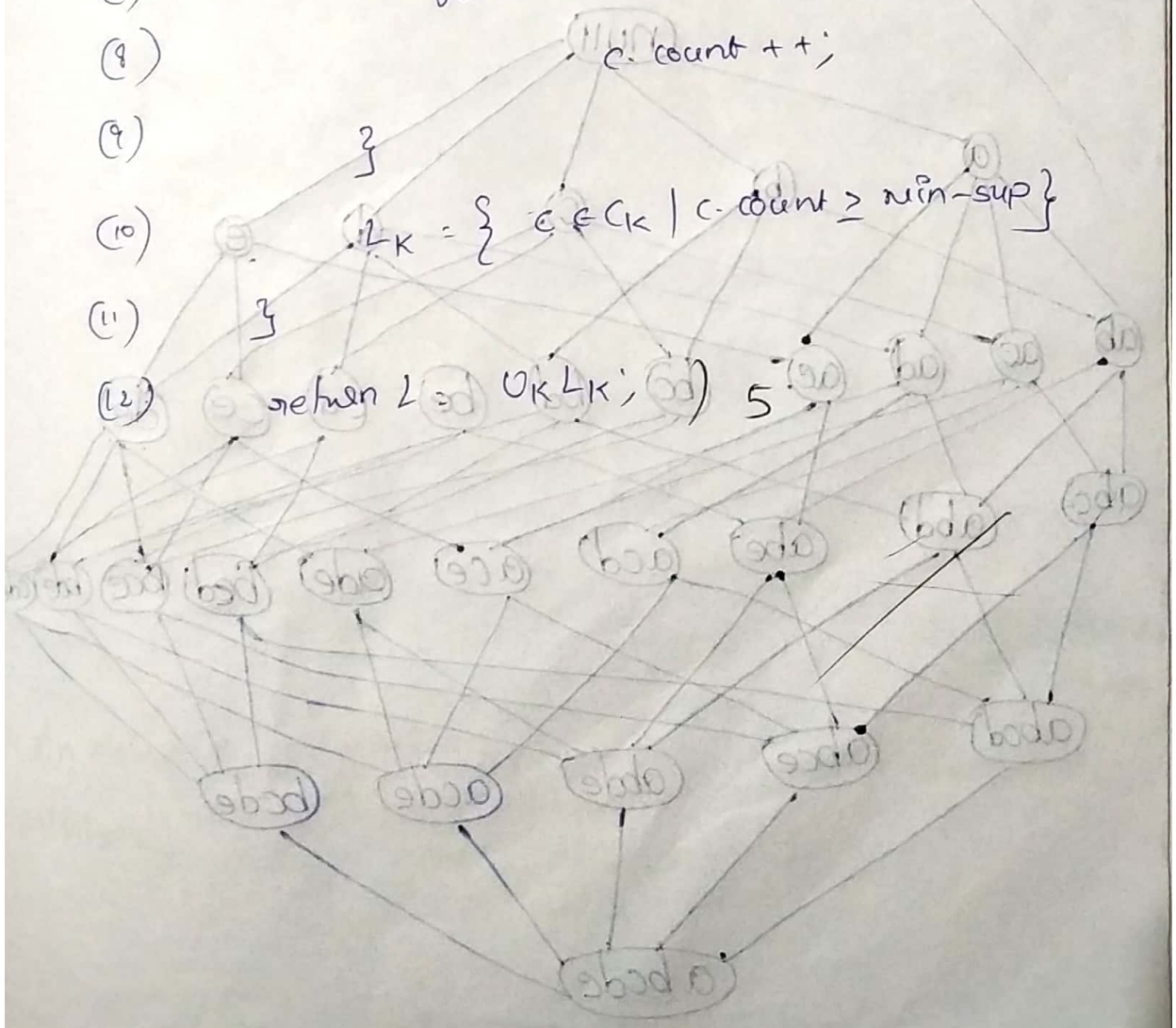
Apriori Algorithm:

Input: Database, D, of transactions; minimum support threshold, min_sup

output: L, frequent itemsets in D

method

(1) L_1 = find_frequent_1-itemset(D);
(2) for ($k=2$; $L_{k-1} \neq \phi$; $k++$){
(3) C_k = apriori_gen(L_{k-1}, min_sup);
(4) for each transaction $t \in D$ { // scan D for counts
(5) C_t = subset(C_k, t); // get the subsets of
(6) t that are candidates
(7) for each candidate $c \in C_t$
(8) c.count++;
(9) }
(10) $L_k = \{ c \in C_k \mid c.count \geq min\_sup \}$
(11) }
(12) return $L = \cup_k L_k$;



→ Compact representation of frequent Itemset

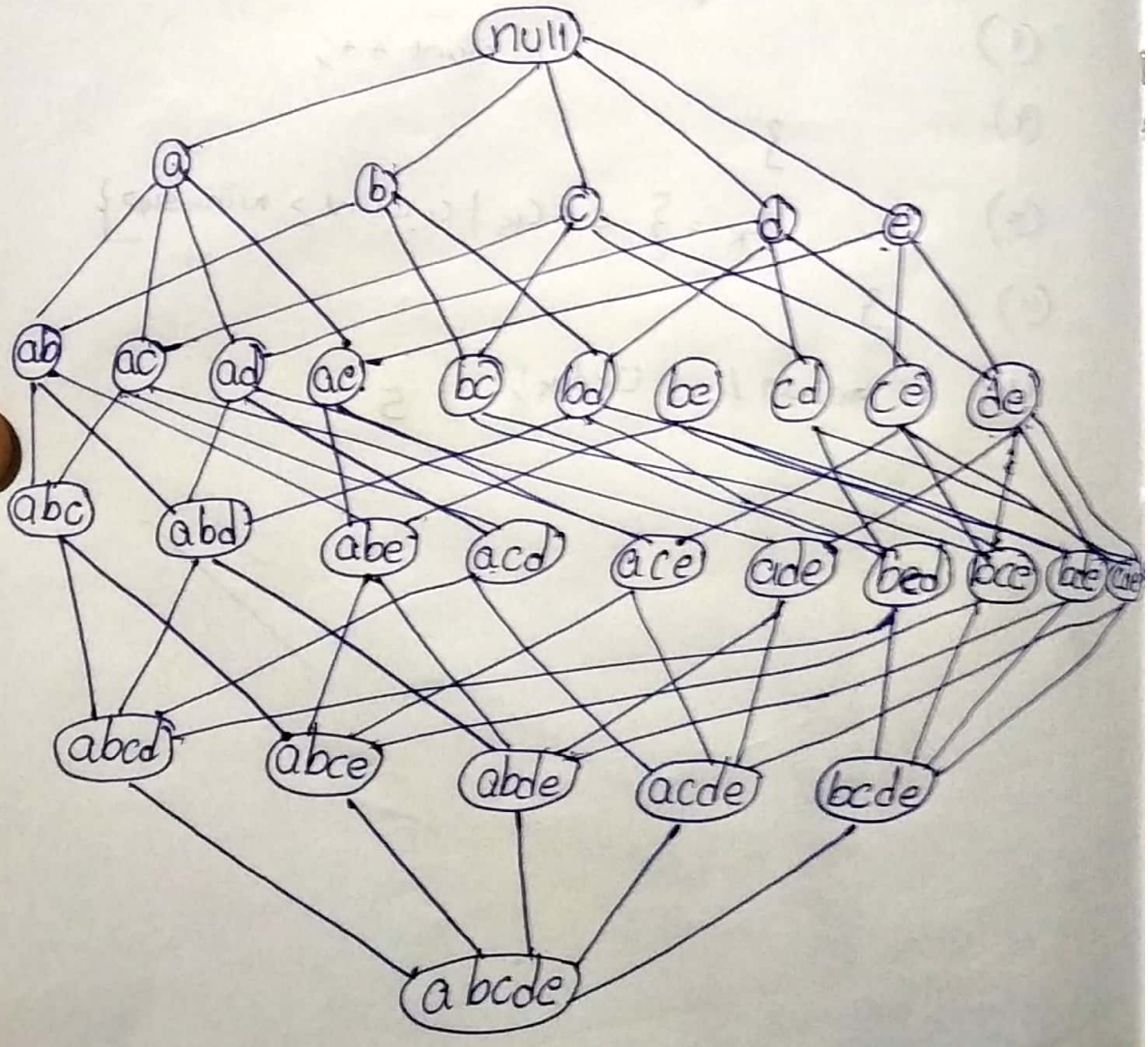
The no. of frequent Itemset produce from a Transaction dataset can be very large it is useful identify a small representative set of item sets from which all other frequent Item sets can be derived.

we have two methods for representating frequency item sets in term of compactness. They are

1. Maximum frequent Item sets
2. closed frequent Item sets

Maximal frequent Item set:-

It is defined as frequent Item sets for which none of the its intermediate super set are frequent



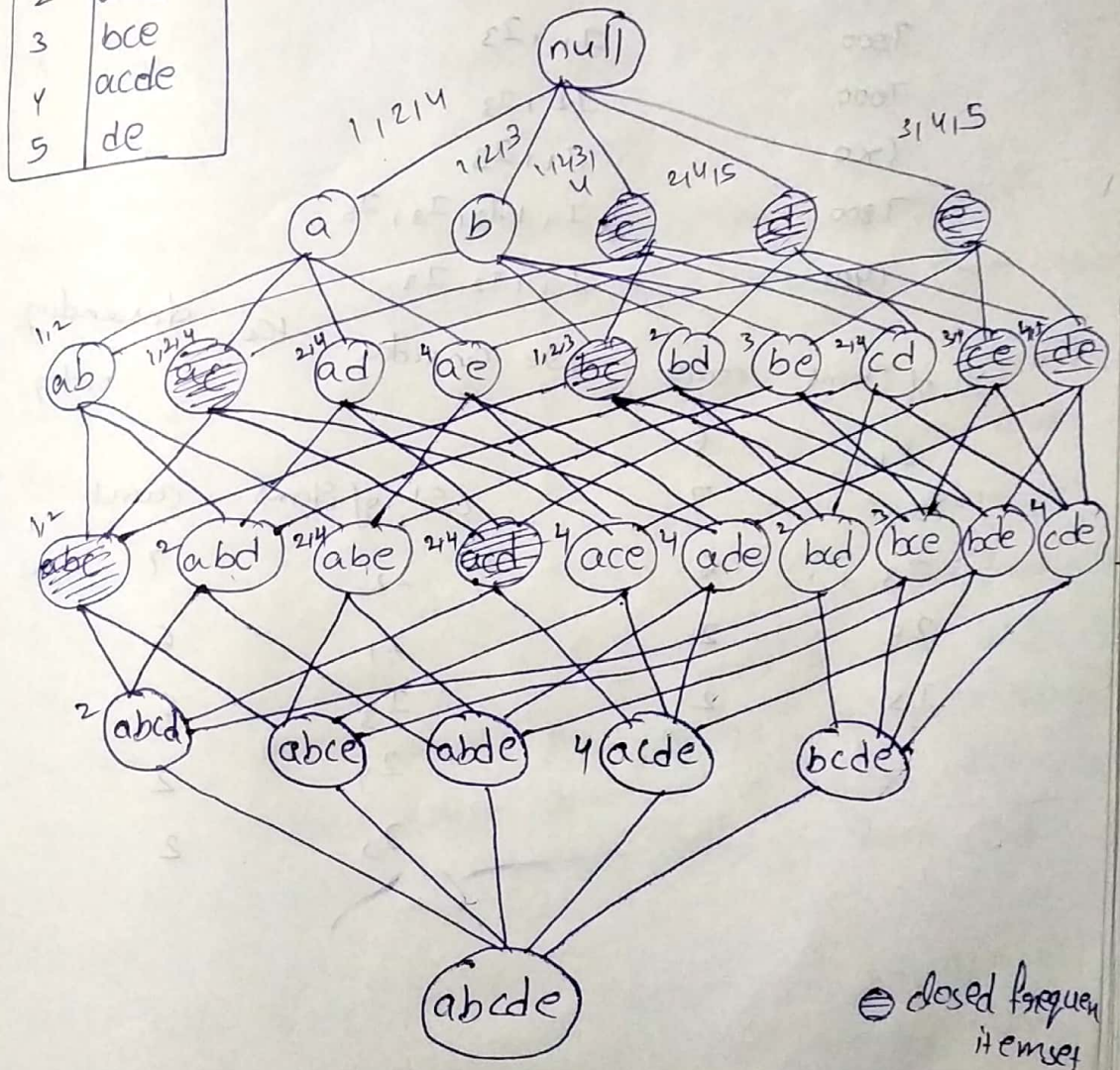
closed frequent item set

closed frequent item set provide minimal representation of item set without losing their support information.

Def An itemset X is closed if none of its immidiate superset has exactly the same support count as X

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de

Minsup = 40%.



● closed frequent itemset

An example of the closed frequent itemsets (with minimum support count equal to 40%.)

→ FP tree :- (Frequent pattern tree)

Apriori Algorithm example

Tid	List of item-ids
T_{100}	I_1, I_2, I_3
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

We consider the descending order

List of items	count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

List of items	count
I_2	7
I_1	6
I_3	6
I_4	2
I_5	2

Here we use FP growth algorithm for finding fp tree representation for the frequent-item sets

→ originally in the Apriori we form the frequent item sets with the help of join and prune steps

→ But with the Apriori we have the problem for finding frequent itemsets for more of the transactions can be happened to be complese for this instead of the apriori. that entire dataset data using a compact data structure called a FP tree and extracts frequent itemsets directly from this structure

→ 1. now for finding the pattern assume of apriori we can take the list of items and we can take the support (minimum support threshold value = 0.2 = 20%.) then I can form a table with the list of items with respect to minimum support of threshold value.

List of Items	sup-count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

2. Now this is a FP tree for this we can find out the pattern.

For finding the pattern we can use descending order (High to low) for the minimum support count happend or Equal or greaterthan happend item sets.

Now we have the pattern is

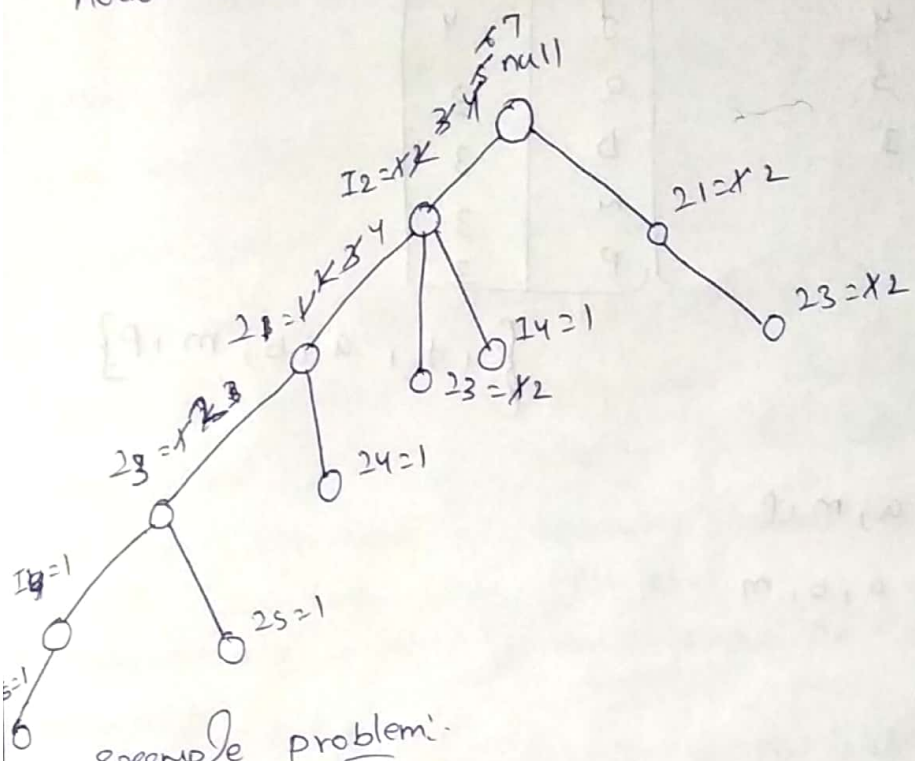
List of item	count
I_2	7
I_1	6
I_3	6
I_4	2
I_5	2

Then the pattern is $\{I_2, I_1, I_3, I_4, I_5\}$

3. Based upon the formed pattern we reordered the items of belonged 9 transactions because we have in our problem the dataset can have 9 transactions

Tid	List of items (ordered)	List of items re-ordered
T_{100}	I_1, I_2, I_3	I_2, I_1, I_3
T_{200}	I_2, I_4	I_2, I_4
T_{300}	I_2, I_3	I_2, I_3
T_{400}	I_1, I_2, I_4	I_2, I_1, I_4
T_{500}	I_1, I_3	I_1, I_3
T_{600}	I_2, I_3	I_2, I_3
T_{700}	I_1, I_3	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5	I_2, I_1, I_3, I_5
T_{900}	I_1, I_2, I_3	I_2, I_1, I_3

4. Now we have the final step from the ordered items we can construct the FP tree with the principle of incoming and outgoing through nodes and links.



→ Example problem:

Tid	Itemsets
1	a, c, d, g, m, p
2	a, b, c, f, l, m, o
3	b, f, h, o
4	b, k, c, p
5	a, f, c, l, p, m, n

min-sup = 3

Tid	Item	Itemsets	count
		a	3 ✓
		b	3 ✓
		c	4 ✓
		d	1
		f	4 ✓
		g	1
		k	1
		l	2
		m	3 ✓
		n	2
		o	2
		p	3 ✓

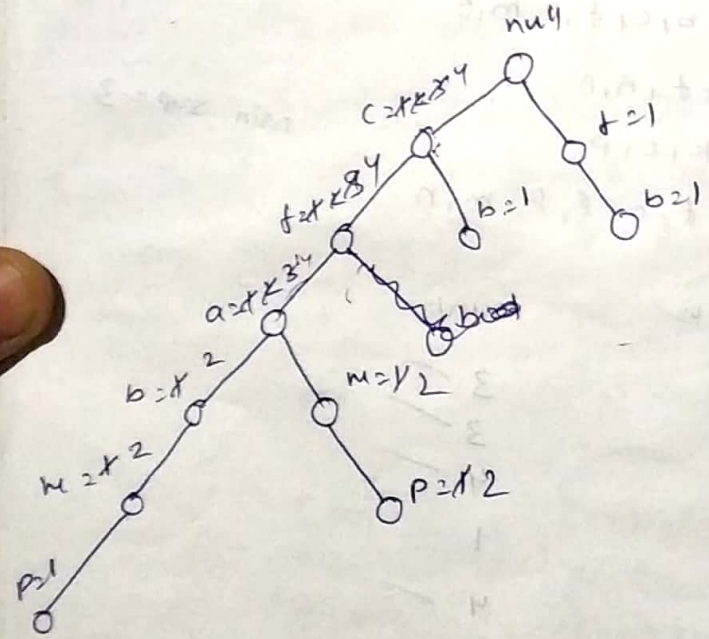
item	count
a	3
b	3
c	4
f	4
m	3
p	3

descending order

c	4
f	4
a	3
b	3
m	3
p	3

{c, f, a, b, m, p}

Tid	list of item
1	c, f, a, m, p
2.	c, f, a, b, m
3.	b, f, b
4	c, b
5	c, f, a, m, p



Unit - VI

K-Mean Alg
DB scan
Agglamarative

Cluster Analysis

cluster is nothing but we form the group with the similar properties of the items are the data are object.

→ Cluster means forming the groups.
→ It was one of the datamining technique.
→ Here we perform unsupervised learning, we use unsupervised learning actually in the classification. Here we have two types of learning supervised. supervised means we know the classlabel.
→ unsupervised means we don't know the class label (reduction)
→ Here we can take the data set. on that dataset we can apply the cluster algorithm. Then we obtain some of the clusters of the data can be happen.

Advantages
→ Not effected by addition of new objects
→ No need to worry about signs of the data
→ computation process is very simple. Because here we are using the formulars are mean, ~~Eqludian~~ Eculdian distance.

Applications
→ Medical - Establish taxanamy of diseasies
→ WWW - world wide web
→ Sismology - trace out the earthquakes of the data
→ Business marketing - find the targeted customers

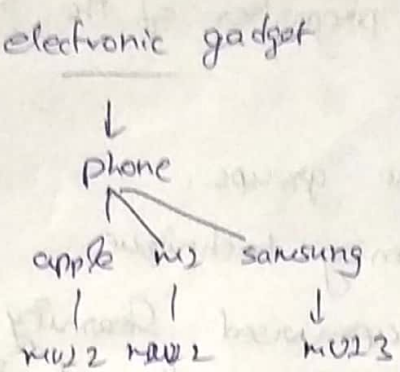
we have the two types of clusters these are the central and heirarical.

→ In the central we form the cluster with the main objects not to form the heirarical manner.

Ex:- Amazon

Bags | Electronic | clothes

→ It's also the cluster it will be happend with the intrinality of the properties of the objects



K-mean algorithm :-

K-mean algorithm is the prototype based or the partional clustering technique that attempts to find a users specified number of clusters (K) which are represented by their centroids / mean values

Procedure -

Step 1 - Take mean values

step 2 - Find nearest number of mean and put in the cluster

step 3 - Repeat the above process ie:- 1, 2 steps until we get the same mean.

K = { 2, 3, 4, 10, 11, 12, 20, 25, 30 }

In this we have to take two mean values randomly

$m_1 = 4$ $m_2 = 12$

$K_1 = \{2, 3, 4, 10\}$ $K_2 = \{11, 12, 20, 25, 30\}$

$m_1 = 19$

$m_1 = 4$

$K_1 = \{2, 3, 4\}$

$m_1 = \frac{9}{3} = 3$

$K_2 = \{2, 3, 4, 10\}$

$m_1 = \frac{19}{4} = 4.75$

$= 5$

$K_3 = \{2, 3, 4, 10, 11, 12\}$

$m_1 = 7$

$K_4 = \{2, 3, 4, 10, 11, 12\}$

$m_1 = 7$

$m_2 = 12$

$K_1 = \{10, 11, 12, 20, 25, 30\}$

$m_2 = \frac{10+11+12+20+25+30}{6}$

$= \frac{18}{6} =$

$K_2 = \{11, 12, 20, 25, 30\}$

$m_2 = \frac{11+12+20+25+30}{5}$

$= 19.6 = 20$

$K_3 = \{20, 25, 30\}$

$m_2 = 25$

$K_4 = \{20, 25, 30\}$

$m_2 = 25$

Then we stop the process, because got the same mean values finally $K_1 = 7$, $K_2 = 25$.

Basic K-mean algorithm:-

1. Select k points as initial centroids
2. repeat
3. From k clusters by assigning each point to its closest centroid
4. Recompute the centroid of each cluster
5. Until centroids do not change.

19-3-20

→ DB Scan :- Density based spatial clustering of a application with noise points

→ DB scan algorithm is based on highly the density.

→ Density based cluser in locates regions of high density that are seperated from one another by regions of low density.

→ DB scan is a simple and effective density based clustering algorithm.

→ In the density based clustering algorithm we can use the center based approach ie., density is estimated for a particular point in the dataset by counting the number of points with in a specified radius thy we call it as Epsilon of that point.

→ The number of points with in a radius of EPS is minimum points are three of that data point itself.

In the DB scan algorithm we have 3 types of points are involved they are

1. In the interior of a dense region (a corepoint)
2. on the edge of dense region (border point)
3. In a sparsely occupied region (a noise or background point)

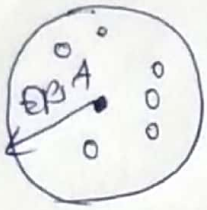
Corepoint - These points are in the interior of a densely based cluster.

A point is the corepoint if the number of points with in a given neighbourhood around the points as determined by the distance function EPS - EPS

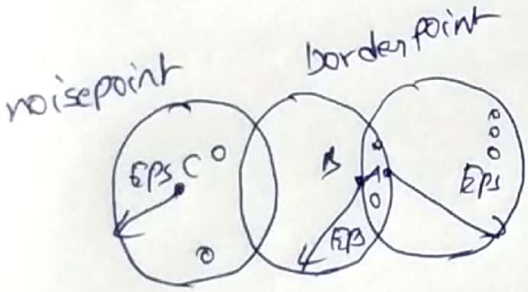
exceeds contain threshold minimum points.

Border point - A borderpoint is not a corepoint. But forms with in the neighbourhood of a corepoint.

Noise point - A noise point is any point ie., neither a core point nor a bordarpoint.



center-based density



DBScan algorithm

1. Label all points as core, border or noise points
2. Eliminate noise points
3. put an edge between all core points that are within Eps of eachother.
4. make each group of connected core points into a seperate cluster.
5. Assign each border point to one of the cluster of its associated core points

1-1-20

Unit-2

Data preprocessing

Dimensionality Reduction: Time/cost parameters

→ Attribute is nothing but a dimension.

Dimensionality reduction means we can reduce the attributes of the dataset.

From old attributes of the data we can follow the new attributes of the data. It is called dimensionality reduction.

we can take the dataset the dataset can have the attributes like sno, sname, sadd, smarks. These attributes are called old attributes. The old attributes are reduced the data by using following the data is in the new attributes.

The old attributes from the subset of the attributes the subset of data is called new attributes

we can measured the dimension of a dataset with an attribute we can measured. the same roll played by the dimension & attribute.

we can create the dataset by the new attributes with the combination of old attributes.