6/2/23 Monday

Module-1

Data Mining

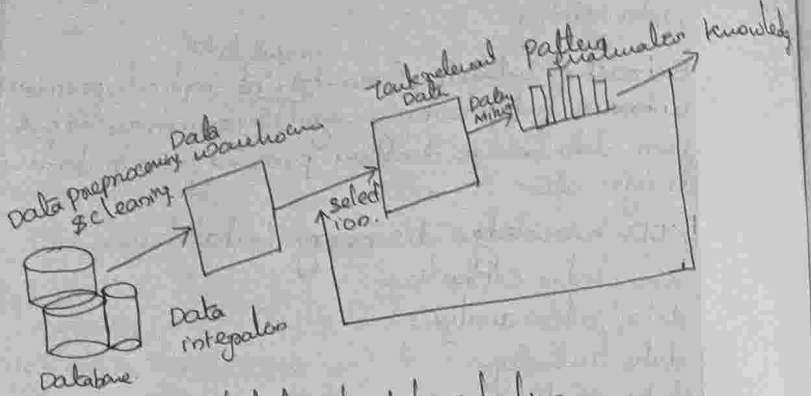
Extract of interesting (non-trivial, implicit, permanently unknown and potentially useful) informations or patterns from data in large database. processed form of data is known as information.

KDD - knowledge discovery in database.
knowledge extraction.
data/pattern analysis.
data archeology.
data dredging.
information harvesting.
business intelligence.
Non data mining.
(Deductive) query processing.
KDD is also known as datamining.

1. Steps in KDD

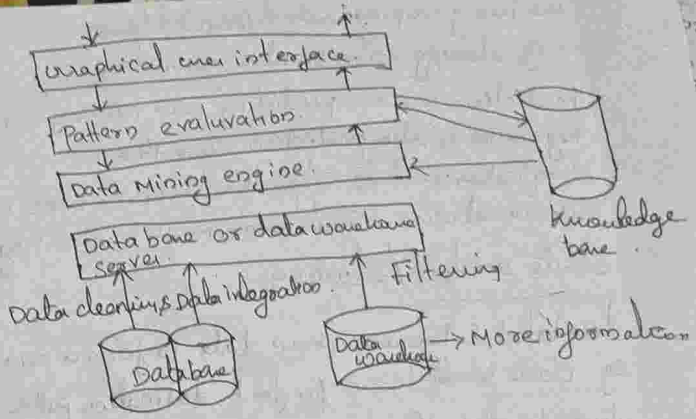
1, Data preprocessing and cleaning.
2, Data selection and Transformation.
3, Data Mining
4, Pattern evaluvation.

Data mining: the core of knowledge recovery process.



creating a target data set : data selection.
Data cleaning and preprocessing : removing uncessantial data.
2) find useful features dimensionality.
choosing functions of data mining
summarization, classification, regression, association, clustering
Choosing the mining algorithm.
Data mining : search for patterns.
Pattern evaluation & knowledge
visualization, transformation, removing redundancy patterns.
use of discovered knowledge.

Architecture a Typical Data mining System



6 Major components
1. Database, data warehouse, WWW, or other information repository
2. Database or data warehouse server.
3. knowledge base.
4. Data Mining engine
5. pattern Evaluation module
6. user interface.

7/2/22
Tuesday

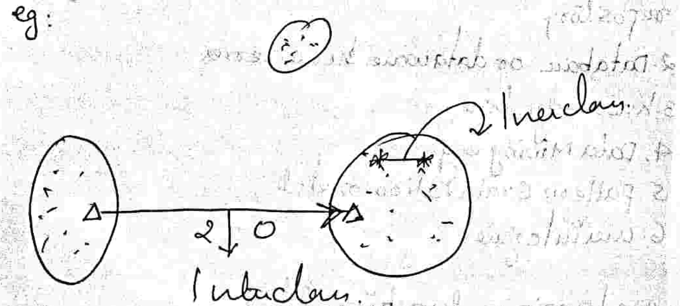
Data mining functionalities
- Concept description: characterization and discrimination
  - Generalize summarize and constant data.
  characteristic eg: dry vs wet regions.
- Association (correlation and causality)
  Multidimensional vs single dimensional association.
- classification and prediction.
  Finding models that distinguish class or concept for future
  Prediction

we are grouping into different clases.
eg: clarify cars based on gas mileag.
Prediction: we are predicting new values from previous values.
Regression is also prediction algorithm.
classification exclude algorithm, decision tree, classification rule, neural n/w. Type - simple, Multi linear, Polynomial - curve logic

Cluster analysis/clustering

class label is known: Group data to form a new classes.
eg: cluster based house to find destination pattern.
cluster based on the principle maximize ~~b/w the class~~ the intraclass value ↑ similarity and minimizing the interclass similarity value ↓.

eg:



Similiarity distance measure;
Inter class similiarity - b/w the class.
Intra " " - within the class.

Outlier analysis (noise)

Outlier: A data object that doesnot comply with general behavior of the data.
Outlier Analysis: It can be considered as noise or exception but it is useful in fraud detection and even analysis

Trend and evaluation analysis

Trend and deviation: regression analysis similiarity based analysis
Sequential pattern mining, periodicity analysis
other pattern: directed or statistical analysis

+ Major issues in data mining.
- Mining methodology and user interaction
  Mining different kinds of knowledge in database.
- Performance and scalability.
- Issues relating to the diversity of the datatypes.
- Issues related to application and social impacts.
- Human interaction.
- overfitting
- changing data.

9/2/23 Thursday Data warehousing

A datawarehouse is subject-oriented, integrated, time-variant, non-volative collection of data in support of management's decision making process.
Decision support database.
Datawarehousing: The process of conducting and using data warehouse.
Time variant: entirely different from database.
non-volatile: operational update of data doesnot occur in the data ware house environment.
Initial loading of data & access of data

Database V/s Datawarehouse

Feature	Database	Datawarehouse.
Purpose	It is designed to record data	It is designed to analyze
Processing method.	The database uses the online transaction processing (OLTP)	The datawarehouse uses OLAP.
usage	Helps to perform fundamental operation for your business.	Allows you to analyze your business.
Tables & joins	complex as they are normalized	simple in because they are denormalized.
orientation	Application-oriented collection of data	A subject-oriented collection of data.
Designing tools	ER modelling techniques are used.	Data modelling techniques are used.
Data type	Data stored in the database is up to date.	current and historical Data is stored in datawarehouse. May not be up to date.
query type	simple hierarchical queries are used.	complex queries are used
Data summary.	Detailed data stored in the database.	It stores highly summarized data

OLTP - Online Transaction processing.

	OLTP	OLAP
ures	clerk, IT professional	knowledge workes.
function	day-to-day operation.	decision support
DB design.	application oriented.	subject oriented.
Data	current up to date detailed, flat relational isolated	historical, summarized multidimensional, integrated consolidated
usage	repetative	ad-hoc
access	read/write/index/hash on primary key.	lots of scans.
unit of work	short, simple transaction	complex query.
#records of accnd	ten	millions.
#user	1000s	100s.
DB size	100MB-GB	100GB-TB.
metric	transaction throughput	query, throughput response

13/2/23 Monday

Multidimensional data model

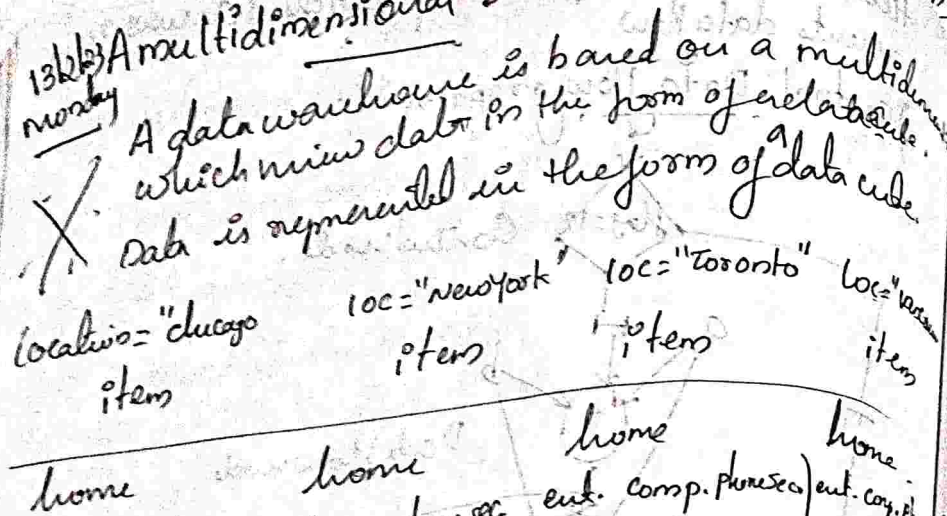
A datawarehouse is based on a multidimensional which view data in the form of a datacube. Data is represented in the form of data.

location="Chicago"	loc="Newyork"	loc="Toronto"	loc="Vancouver"
item	item	item	item

	home ent.	comp.	phone sec	home ent.	comp.	phone sec	home ent.	comp.	phone sec	home ent.	comp.	phone sec
time												
Q1												
Q2												
Q3												
Q4												

13.6.3 A multidimensional Data model

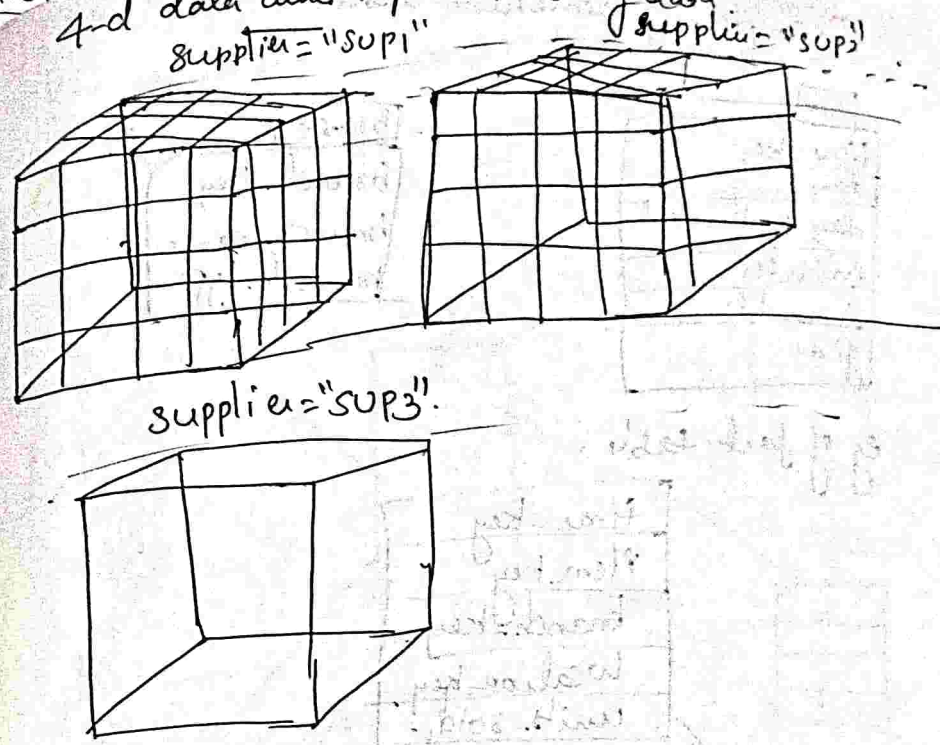
A data warehouse is based on a multidimensional model which views data in the form of a data cube. Data is represented in the form of data cube.



A Comp & Phone

home entertainment

4-d data cube representation of data



Data cube

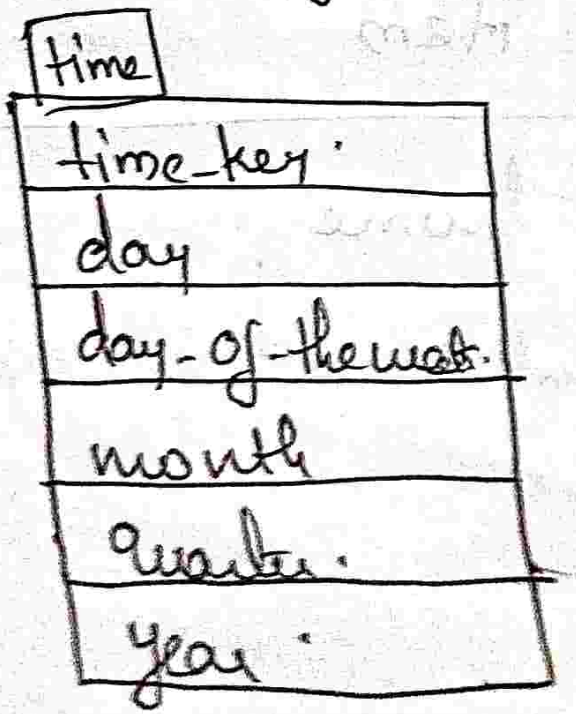
Data cube allows data to be modelled and viewed in multiple dimensions. It can be defined by dimensions and facts.

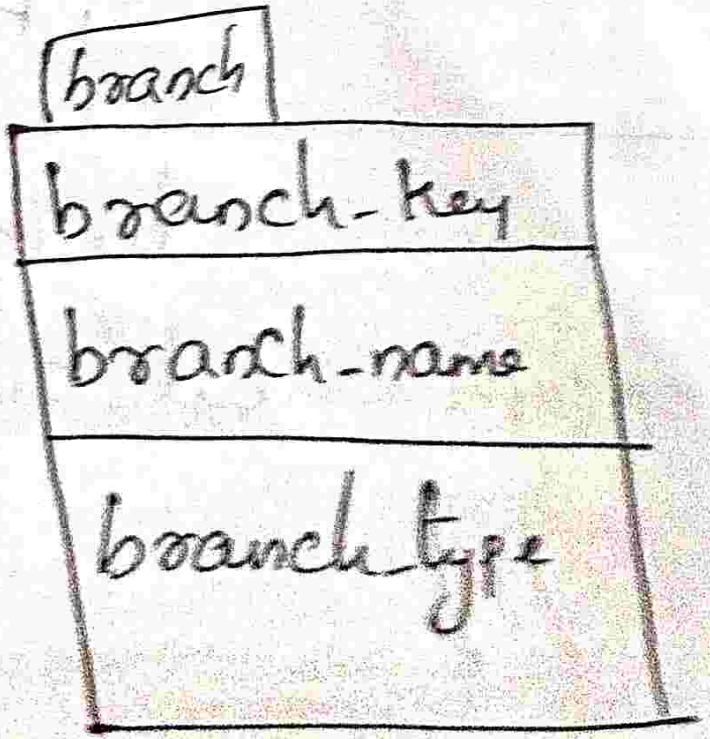
Dimension

Dimensions are the perspectives or entities with which a table, store entity.

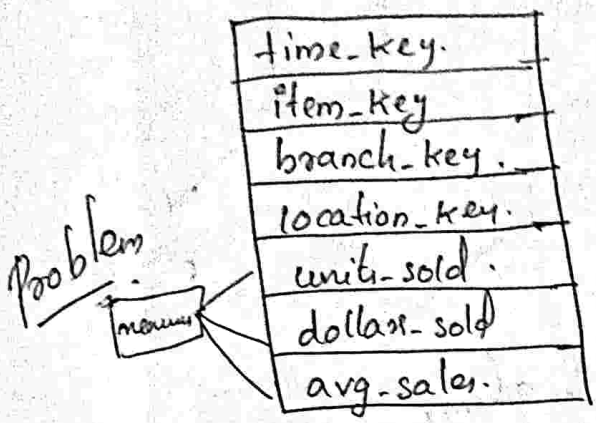
fact - numerical tables, relationship b/w dimension tables.

Example of dimension table.





eg of fact table

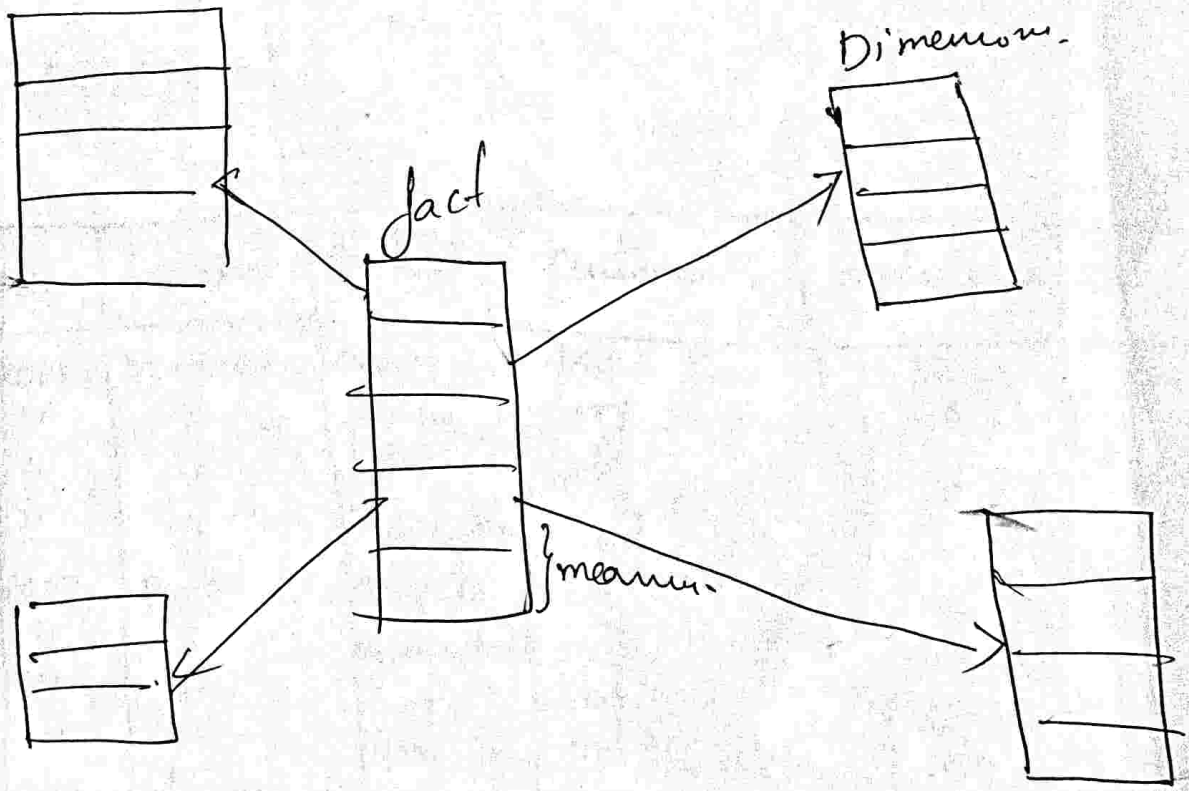


* Conceptual Modelling of Data warehouse / Data warehouse schema.

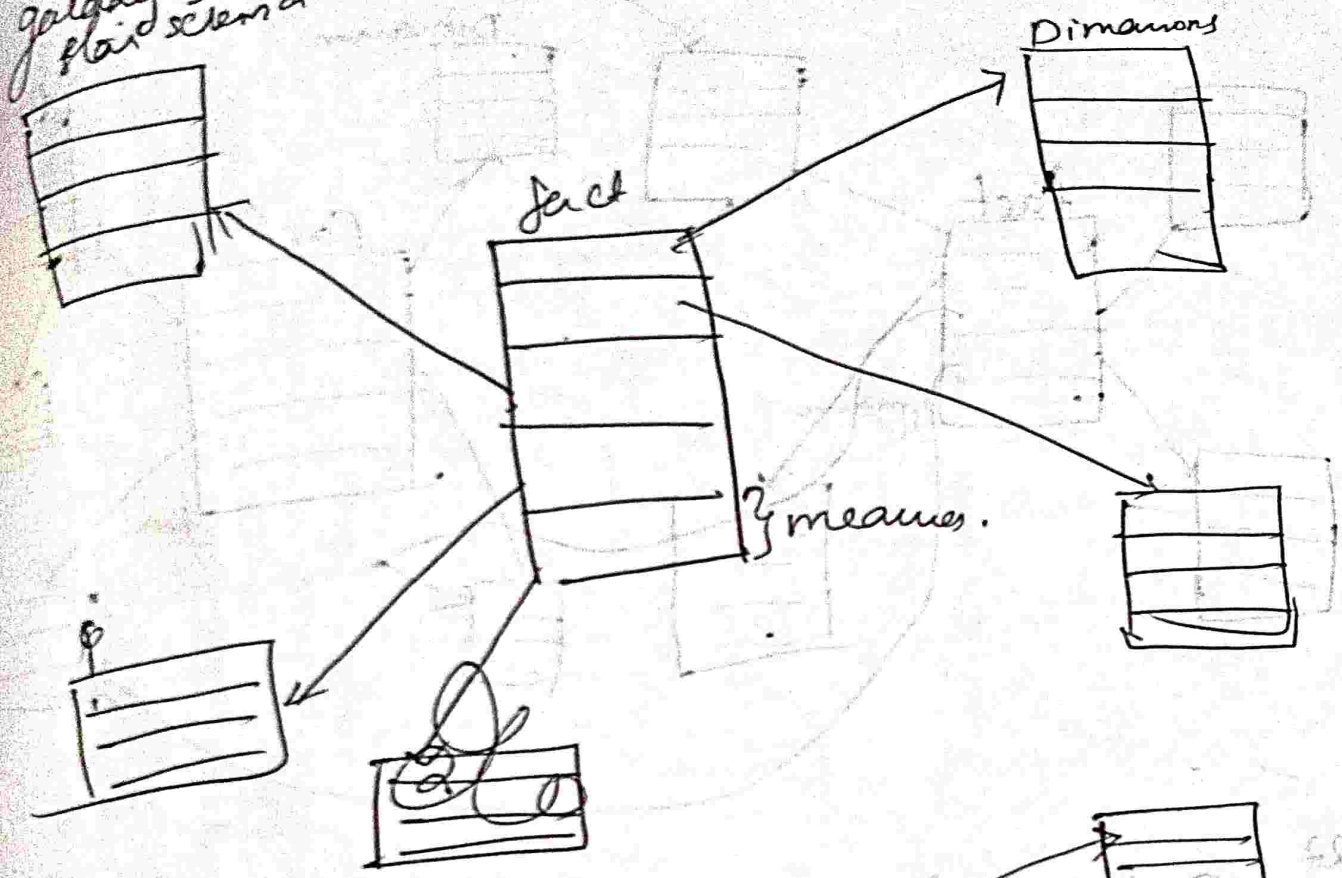
star schema - fact table connected in the middle to a set of dimension table.

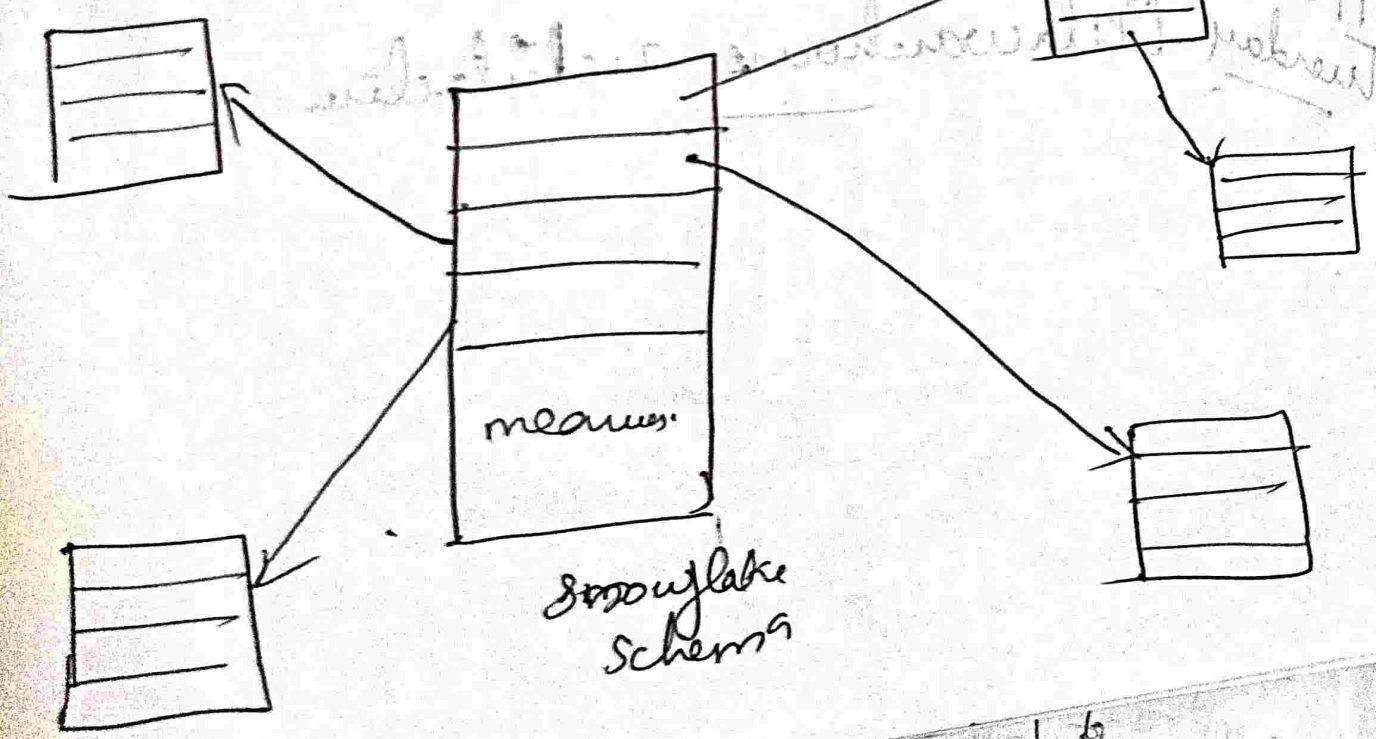
Snowflake schema: A refinement of star schema to normalized into a set of smaller dimension tables, forming a shape similar to.

Galaxy schema: Multiple fact tables share dimension tables, viewed as a collection of stars therefore called galaxy schema.

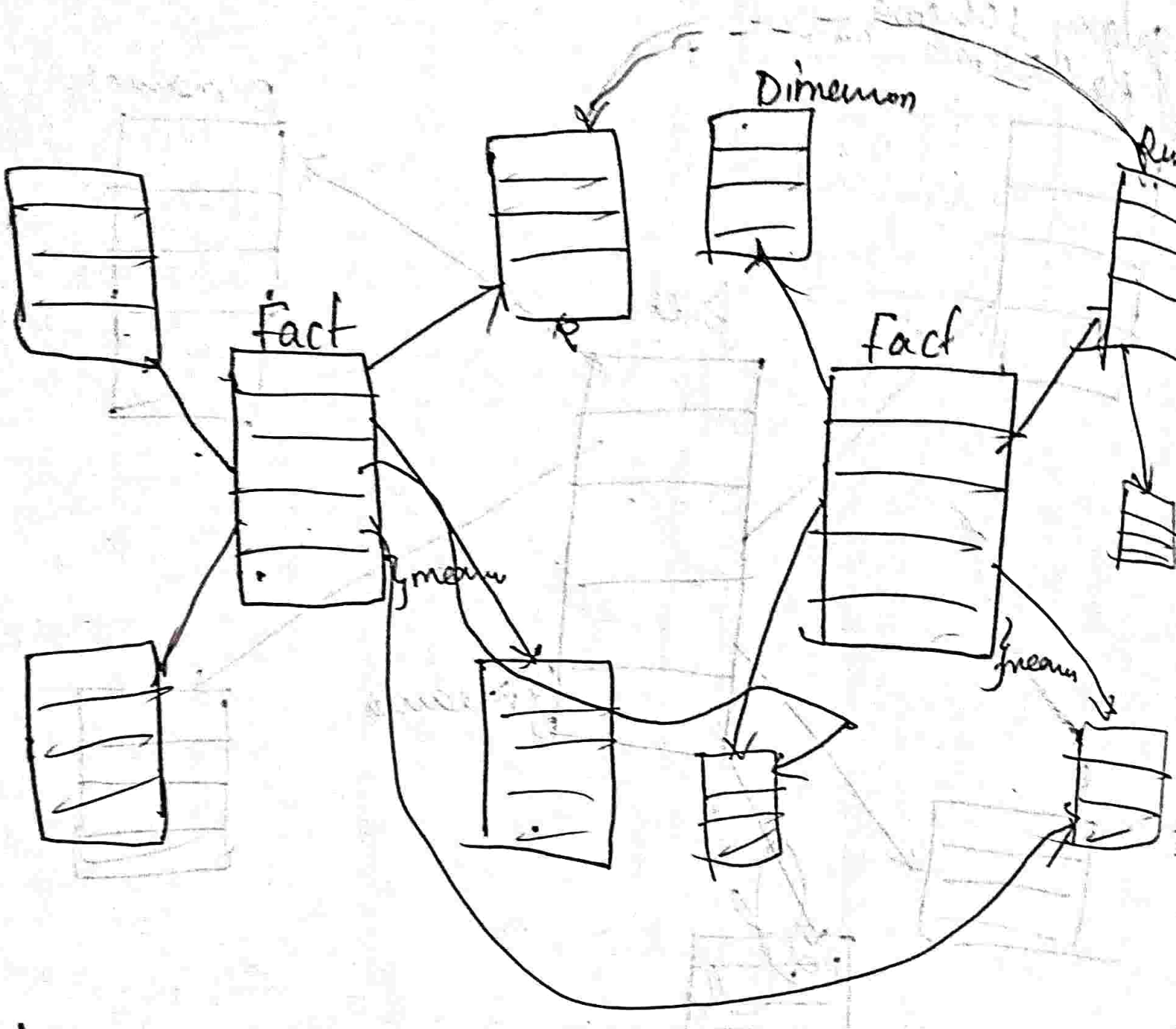


Multiple fact tables share dimension tables, viewed as a collection of stars therefore called galaxy schema, fact constellation schema.





Galaxy schema.

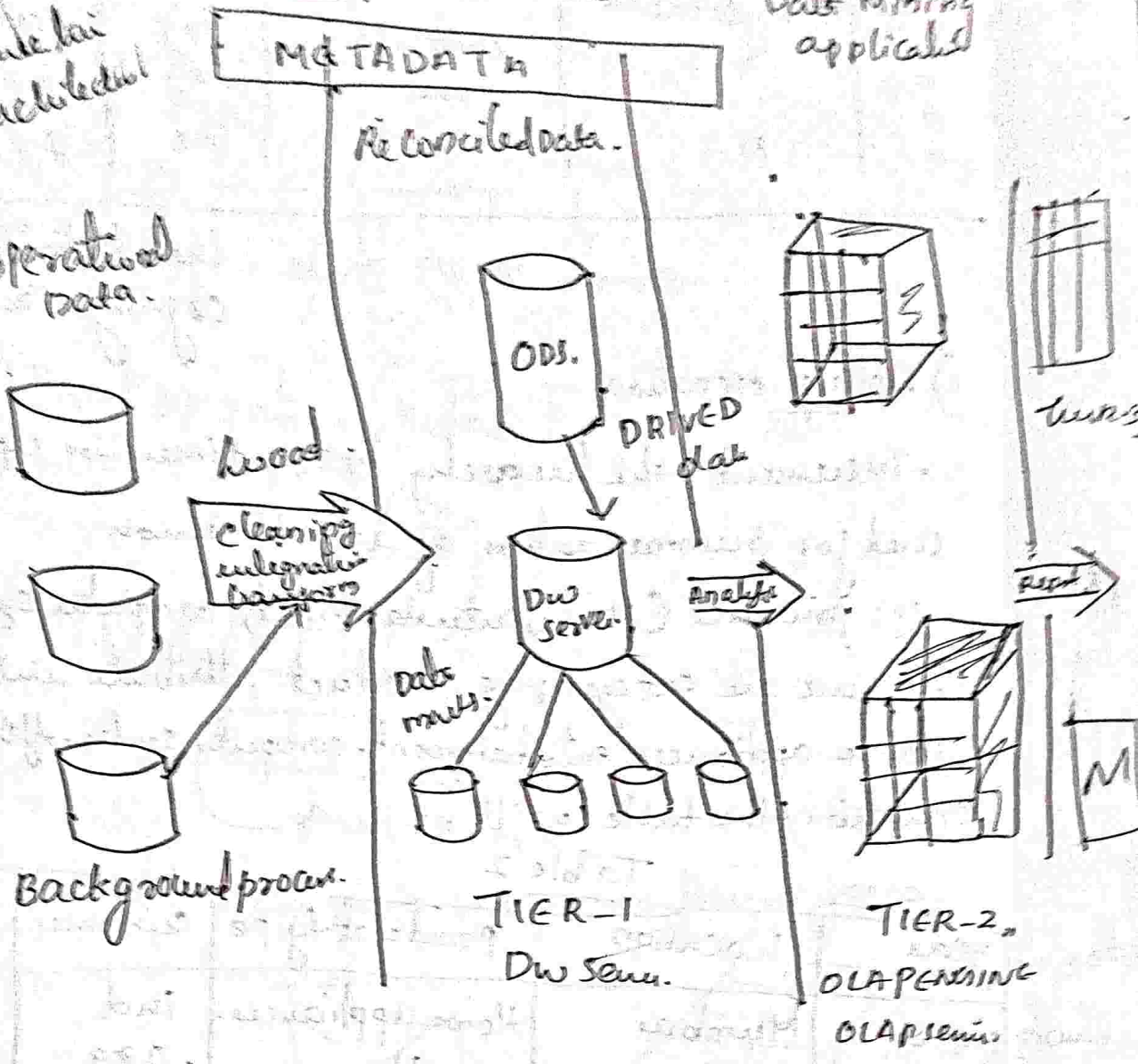


14/2/22
Tuesday Datawarehouse architecture

14/2/23 Tuesday

Data ware house architecture

Architecture



15/2/23 Wednesday

OLAP operations

Table 1

year	location	Product	Customer	Total sale
2008	Mumbai	Oven	Ind	32
"	"	Juicer	org	34
"	"	TV	org	12
"	"	Music system	Ind	11
2009	Pune	Computer	org	43
"	"	Music system	Ind	21
"	"	TV	Ind	14
"	"	play_st	Ind	10
2010	chennai	oven	org	9
"	"	Music sy	Ind	8
"	"	computer	org	13
"	"	TV	Ind	24

2011	Hyderabad	Music-System	Ind	22
"	"	computa	org	15
"	"	ones	org	3
"	"	TV	Ind	7

Ind - Individual
org - organization

1) Rollup operation

- Increment the hierarchy by one dimension up. It is also used for summarization of data dimension

eg: product (Home, entertainment, computer system)

Here, we are categorizing product attribute into 3 categories Home appliances, entertainment, computer systems. After rollup operation the table will be.

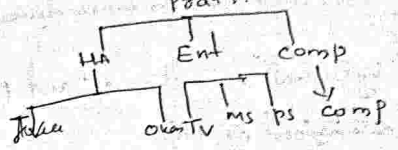
Table 2

year	location	product-type	customer	total sales
2008	Mumbai	Home appliances	Ind	32
"	"	"	org	34
"	"	entertainment	org	12
"	"	Music-system	Ind	4
2009	Pune	computer-system	org	43
"	"	entertainment	Ind	45
2010	chennai	Home appliance	org	09
"	"	entertainment	Ind	26
"	"	computer-system	org	13
2011	Hyderabad	entertainment	Ind	29
"	"	computer-systems	org	45
"	"	HA	org	03

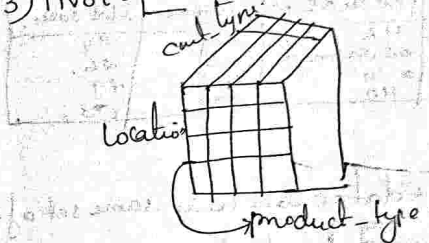
2) Roll down / Drill down

- Reverse operation of roll up.

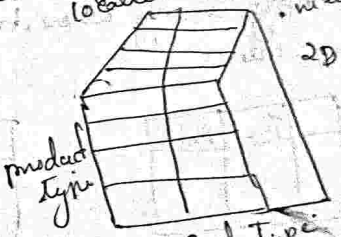
- used to document the dimension hierarchy by one down.



3) Pivot operation (Rotate).



It is used to view data cubes in different form.



- view 3d to 2D planes.

4) Slice: It is used to extract a slice of the original cube corresponding to a single value of a given dimension. (similar to select operation RDBMS)

for eg: Slice Table 2 to on location = "Hyderabad".

year	location	product type	Custom	Total sales
2011	Hyderabad	entertainment	Ind	29
"	"	computer-system	org	45
"	"	HA	org	03

16/2/23 Thursday

5) Dice operation

It is used for selecting two or more dimension and one or more value.

eg: location ("hydrabad" or "chennai" and product type (entertainment or HA).

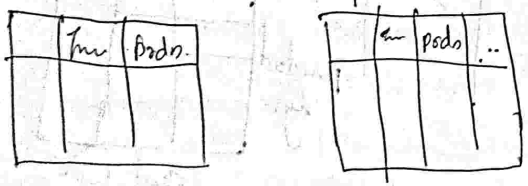
After applying dice operation from Table 2.

year	location	product type	customer	total sales
2010	chennai	HA	org	09
"	"	entertainment	ind	26.
2011	Hydrabad	"	ind	29
"	"	HA	org	03.

6) Drill Across

- Analyse the cells of 2 data cubes using same set of dimensions
- used to move data from one data cube to another.
- Applyied only when data cubes are having same/equal dimensions

eg: Sales.



7) Range query

- Avg
- Sum
- count
- min
- max

It applies a given aggregation operation over a set of selected cells.

For eg: find the total no: of sales for a period 2008-2010.

Then the range query will return the value by aggregating the values total sales from 2008-2010. It can be applied on dimension with numerical values.

Range sum query [sum].

year	location	product	customer	Avg-sales	total avail
2008	Mumbai	Oven	ind	12	03.
"	"	fevicol	org	14	04.
"	"	TV	org	12	01.

Module - 2

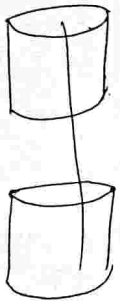
20/2/23
7th session

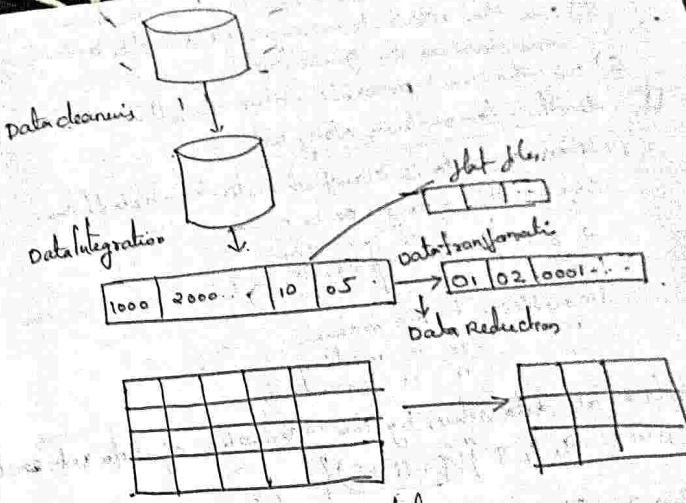
Data preprocessing

3 types of data
Incomplete
Noisy
Inconsistent

why data preprocessing
- Real world data
  - Incomplete - malfunctioning of data.
  - Noisy - unwanted data error.
  - Inconsistent
    - Naming convension : same data and different columns
    - Normalization

Major steps in data preprocessing
(i) Data cleaning
(ii) Data integration
(iii) Data transformation
(iv) Data reduction





Data cleaning : Remove noisy data.
filling missing.

Data transformation : To convert into common format.
(normalization).

Data reduction : To reduce it size after data transformation.

23/2/23
Thursday

Data cleaning : filling the missing value, smoothing noisy data
identifying or remove outliers, resolving inconsistencies

I Handling missing values

1) Ignore the tuple.
Method is not very effective

2) Filling in the missing values manually.
This approach is time consuming, may not be feasible.

3) use a global constant to fill in the missing value.

4) Use the attribute mean to fill in the missing value
5) Use the attribute mean for all samples belonging to the same class as the given tuple.
6) Use the most probable value to fill in the missing value

II Handling / smoothing Noisy Data :

(i) Binning : Data is classified into buckets / bins.

Consider the data for price

Bin 1 | Bin 2 | Bin 3
4, 8, 15 | 21, 21, 24 | 25, 28, 34

Partition

Smoothing by bin mean.
" " " median.
" " " boundaries

Replace all data values by the mean value of data set each bin.

Bin 1 : 9, 9, 9 $\left\{\frac{(4+15+8)}{3}\right\} = 9$.

Bin 2 : 21, 21, 24 = 22.
22, 22, 22

Bin 3 : 25, 28, 34.
29, 29, 29.

Smoothing by median.

Bin 1 : 8
Bin 2 : 21
Bin 3 : 28.

Bin 1 : 8, 8, 8.
" 2 : 21, 21, 21.
" 3 : 28, 28, 28.

Smoothing by bin boundaries.
4, 4, 15
21, 21, 24
25, 25, 34.

Apply smoothing techniques by min mean, median, boundaries on the following data

13, 7, 24, 50, 3, 9, 11, 2, 34, 26, 14, 23, 1, 8, 30

bin size : 5.

Sort

1, 2, 3, 7, 8 | 9, 11, 13, 14, 23 | 24, 26, 30, 34, 50

Bin 1 : 1, 2, 3, 7, 8
" 2 : 9, 11, 13, 14, 23.
" 3 : 24, 26, 30, 34, 50.

Smoothing by bin means.
Bin 1 : 4.2, 4.2, 4.2.
" 2 : 14.6, 14.6, 14.6
" 3 : 32.8, 32.8, 32.8

Smoothing by bin median.
Bin 1 : 3, 3, 3, 3, 3.
" 2 : 13, 13, 13, 13, 13
Bin 3 : 30, 30, 30, 30, 30.

Smoothing by bin boundaries
Bin 1 : 1, 1, 1, 8, 8
Bin 2 : 9, 9, 9, 23, 23
Bin 3 : 24, 24, 24, 50, 50.

Regression (prediction)
It is also used for noise recovery.
Types of regression
Simple linear.
Multiple
Polynomial
Logistic

25/2/23 Saturday

Data cleansing as a process.
Step 1 : Discrepancy detection
2 : Data transformation.

Discrepancy
Human error in data entry

Avoid discrepancy

1) using metadata.

Commercial tools.

Data scrubbing tools.

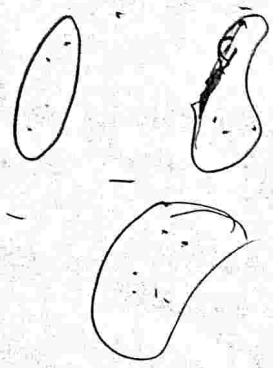
Data auditing tools.

Data transformation

Data Migration tool

ETL (Extraction / Transformation / loading).

clustering



II Data Integration

The data integration is the process of combines data from multiple sources into a coherent data store.

Issues in data Integration

1. Entity identification problem
2. Data Redundancy problem.
3. Detection and resolution of data value conflicts.

1. Entity Identification problem

Eg: Data analyst or computer cannot ensure that customer-id in one database and cust-number in another refer to the same attribute. ie metadata help to avoid errors in schema entity identification problem.

Data Redundancy (repetetion of data)

using normalization to avoid redundancy. Derived attribute are also examples of redundancies.

Solution to data Redundancy

Is correlation Analysis

For numerical attributes we can evaluate the correlation b/w 2 attributes A & B. known as pearson's product moment coefficient (named after its inventor, karl pearson)

$$r_{A,B} = \frac{\sum_{i=1}^{N} (a_i - \bar{A})(b_i - \bar{B})}{N \sigma_A \sigma_B}$$

$$= \frac{\sum_{i=1}^{N} (a_i b_i) - N\bar{A}\bar{B}}{N \sigma_A \sigma_B}$$

N - No: of tuples.

a_i & b_i - respective values of A & B in tuple i.

$\bar{A}$ & $\bar{B}$ - respective mean values of A & B.

σ_A & σ_B - respective standard deviation.

A, B - attributes.

$r_{A,B}$ - correlation coefficient.

$-1 \leq r_{A,B} \leq 1$.

if $r_{A,B} > 0$, A & B are +vely correlated. which means the value of A increase as the values of B increase.

Highly Correlated

if $r_{A,B} = 0$, A & B are independent & there is no correlation b/w them.

If $r_{A,B} < 0$

A & B are negatively correlated, where the values of one increase, the values of the other attribute decrease.

If $r_{A,B} \neq 0$ A & B dependent to each other.

Correlation analysis for categorial attributes

A correlation relationship blw 2 attributes A, B can be discovered by a x^2 (chi-square) test. Suppose A has c distinct values, namely $a_1, a_2, \ldots c$. B has r distinct values, namely $b_1, b_2, \ldots b_r$

$$x^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

The data tuple described by A & B can be shown as a contigency table, with the c values of A

O_{ij} – observed frequency (actual count) of the joint event (A_i, B_j)

e_{ij} – expected frequency of (A_i, B_j)

The sum is computed over all of the

$$e_{ij} = \text{count}(A = a_i) \times \text{count}(B = b_j)$$

Problem of categorial analysis

1) Suppose that a group of 1,500 people was surveyed. The gender of each ~~person~~ person was noted. Each person was polled as to whether their preferred type of reading material was fiction or non fiction. Thus, we have 2 attributes, gender and preferred reading. The observed frequency (or count) of each possible joint events is summarized in the contigency, where the numbers in parantheses are the expected frequencies.

		Male	Female	
b_1	Fiction	250 $(1,1)$	200 $(2,1)$	450
b_2	Nonfiction	50 $(1,2)$	1000 $(2,2)$	1050
		300	1200	1500

Hypothesis

Gender and preferred reading are independent each other.

B_{ij} = Total cost

$$e_{ij} = \left(\frac{a_{ij} \times b_{ij}}{N}\right)$$

$$e_{11} = \frac{a_1 \times b_1}{N} = \frac{300 \times 450}{1500} = 90$$

$$e_{21} = \frac{1200 \times 450}{1500} = 360$$

$$e_{12} = \frac{a_1 \times b_2}{N} = \frac{300 \times 1050}{1500} = 210$$

$$e_{22} = \frac{a_2 \times b_2}{N} = \frac{1200 \times 1050}{1500} = 840$$

		a_1 Male	a_2 Female	Total
b_1	Fiction	250 $(1,1)$ (90)	200 $(2,1)$ (360)	450
b_2	Nonfiction	50 $(1,2)$ (210)	1000 $(2,2)$ (840)	1050

$$\chi^2 = \sum_{i=1} \sum_{j} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 71.11 + 121.90 + 30.48$$

$$= 507.93$$

degrees of freedom $= (r-1)(c-1)$

$= (2-1)(2-1) = 1$.

If the critical value from the chi-square distribution table is checked against the chi-square. If calculate chi-square value < critical value, hypothesis can be accepted.

otherwise hypothesis rejected.

Critical value corresponding to $1d[(r-1)(c-1)]$ with maximum significant levels of 0.001 is 10.828.

$507.93 > 10.828$ hypothesis can be rejected. ∴ A & B are strongly correlated ie, preferred reading and gender are correlated

Detection and resolution of data value conflicts.

To avoid this attribute functional dependencies and referential constraints in the source system match those in the target system.

28/01/23 Tuesday

Data transformation

The data are transformed into forms appropriate for mining.

... for remove noise from the data. such techniques ... include binning, regression, clustering

28/2/23 Tuesday

Data transformation

- ... are applied to the
the data are transformed into forms appropriate for mining.

Smoothing: which works to remove noise from the data. Such techniques include binning, regression, clustering

Aggregation: where summary or aggregation are applied to the data.

eg: the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data @ multiple granularities

Generalization of the data, where low-level data are replaced by higher level concepts through the use of concept hierarchies.

eg: Categorical attributes, like street, can be generalized to higher level concepts like city or country.

Attribute construction (feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

(eg: we may wish to add the attribute area based on the attributes height & width.)

Normalization

where the attribute data are scaled so as to fall within a small specified range.

Methods for data normalization

Min max normalization.

Z-score "

Min-max normalization.

Performs a linear transformation on the original data.

Let min_A and max_A are the minimum and max value of an attribute A.

Maps a value, v, of A to v' in the range $[new\_min_A, new\_max_A]$.

$$v' = \frac{v - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

techniques include ... normalization ...

It will encounter an "out of bounds" error if a future i/p case for normalization falls outside of the original data range for A.

Q: Suppose that the minimum and max value for the attribute income are \$12,000 & \$98,000 respectively. We would like to map income to the range [0.0, 1.0]. By min-max normalization a value of \$73,600 for income is transformed to

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

disadvantage

$min_A = 12000$

$max_A = 98000$

$new\_min_A = 0.0$

$new\_max_A = 1.0$

Given $v = 73600$

$$v' = \left[\frac{73600 - 12000}{98000 - 12000}(1.0 - 0.0)\right] + 0.0$$

$$= \frac{61600}{86000} \times 1 + 0 = 0.716$$

Z-score or Z-mean normalization.

disadvantage: For eg: consider the data 99200.

$$v' = \frac{99200 - 12000}{98000 - 12000}$$

Z-score or Z-mean normalization

Values for an attribute, A, are normalized based on the mean and standard deviation of A.

Maps a value v, of A to v' by computing

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

where $\bar{A}$ and σ_A are the mean & sd. respectively of attribute A

adv: useful when the actual min & max of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

Suppose that the mean & sd. of the values for the attribute income are \$54,000 & \$16,000 respectively with z-score normalization a value of 73,600 for income is transformed.

$v = 73600$

$\bar{A} = 54000 \quad \sigma_A = 16,000$

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

$$= \frac{73600 - 54000}{16000}$$

$$= \frac{19600}{16000} = 1.225$$

Normalization by decimal scaling

Normalizes by moving the decimal point of values of attribute A.

The no: of decimal points moved depends on the max absolute value of A

A value of v_i of A is normalized to v' by computing.

$$v' = \frac{v}{10^j},$$

where j is the smallest integer such that $\text{Max}(|v'|) < 1$.

Q: Suppose that the recorded values of A range from -986 to 917. The max absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., $j = 3$). So that -986 normalizes to -0.986 and 917 normalizes to 0.917.

$v' =$ the following

12/13/23 Use the 2 method to normalize data whether 200, 300, 400, 600, 1000

(A) Min-max normalization by setting min=0 max=1,

(B) Z-score normalization.

A) $$v' = \frac{v - min_A}{max_A - min_A} [new\_max_A - new\_min_A] + new\_min_A$$

$new\_max_A = 1$

$new\_min_A = 0$.

$v = 200$

$min_A = 200$, $max_A = 1000$.

$$v' = \frac{200 - 200}{1000 - 200} [1 - 0] + 0 = 0$$

$v = 300$

$$v' = \frac{300 - 200}{1000 - 200} [1 - 0] + 0 = \frac{-100}{800} = +0.125$$

appropriate formula.

$$v' = \frac{v - 200}{1000 - 200}(1-0) + 0$$

$$= \frac{400 - 200}{800} = \frac{200}{800} = +0.25$$

$$v' = \frac{v - 200}{1000 - 200}(1-0) + 0$$

$$= \frac{600 - 200}{800} = \frac{400}{800} = +0.5$$

$$v' = \frac{1000 - 200}{1000 - 200} = \frac{200 - 1000}{800} = \frac{-800}{800} = -1$$

normalized data = 0, 0.125, 0.25, 0.5, 1

B) Z-score normalization.

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

$$\bar{A} = \frac{200 + 300 + 400 + 600 + 1000}{5}$$

$$= \frac{2500}{5} = 500$$

σ^2 variance $= \frac{\Sigma (x - \bar{x})^2}{N}$

$$SD = \sqrt{\frac{\Sigma (x - \bar{x})^2}{N}}$$

$$\sigma^2 \text{ variance} = \frac{1}{5}\left[(200 - 500)^2 + (300 - 500)^2 + (400 - 500)^2 + \right.$$

$$\sigma^2 \text{ variance} = \frac{1}{5}\left[(200 - 500)^2 + [300 - 500]^2 + (400 - 500)^2 + (600 - 500)^2 + [1000 - 500]^2\right]$$

$$= \frac{1}{5}\left[90000 + 40000 + 10000 + 10000 + 250000\right]$$

$$= 80000$$

$$SD = \sqrt{80000} = 282.84$$

$$v' = \frac{200 - 500}{282.84} = -1.06$$

$$v' = \frac{300 - 500}{282.84} = -0.7$$

$$v' = \frac{400 - 500}{282.84} = -0.35$$

$$v' = \frac{600 - 500}{282.84} = 0.35$$

$$v' = \frac{1000 - 500}{282.84} = 1.76$$

Data Reduction

Data reduction is the process of retaining relevant data and eliminating irrelevant attributes from the data set.

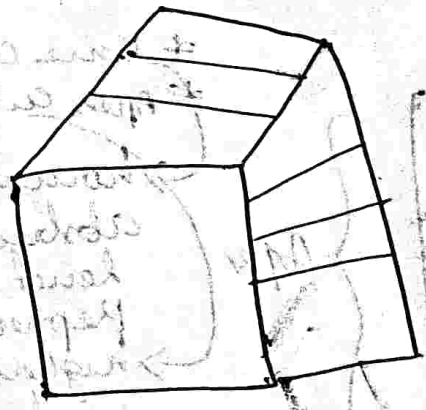
Techniques used for data Reduction

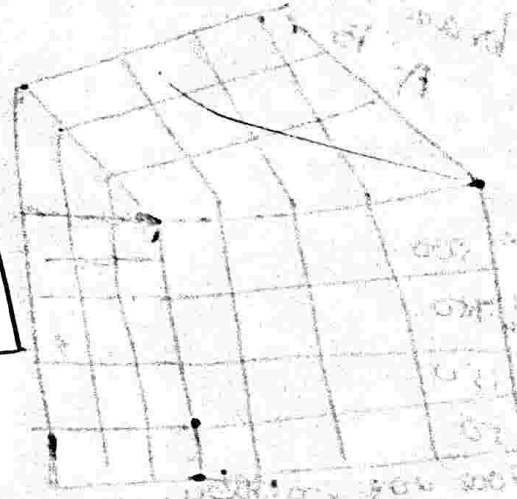
- Data cube Aggregation
- Attribute subset selection
- Dimensionality reduction
- Numerosity reduction
- Discretization & Concept hierarchy generation

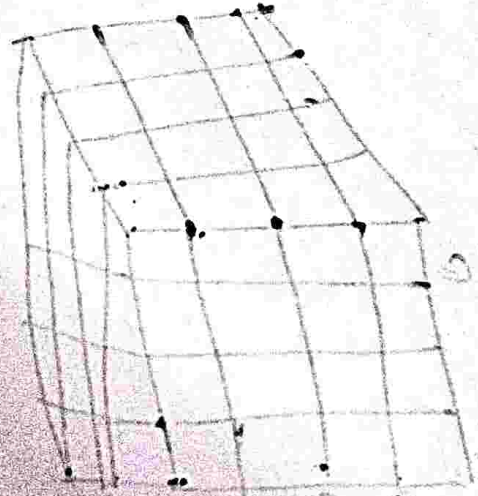
(1) Data cube Aggregation

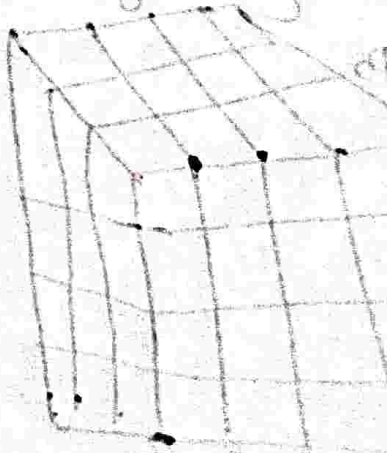
year 2002	
Quarter	Sales

year 2002	
Quarter	Sales
Q_1	
Q_2	
Q_3	
Q_4	





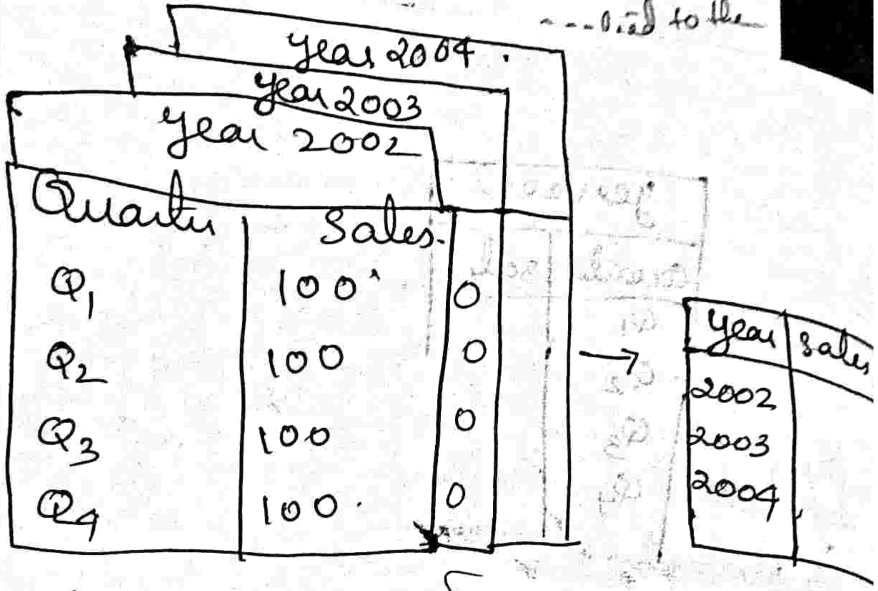


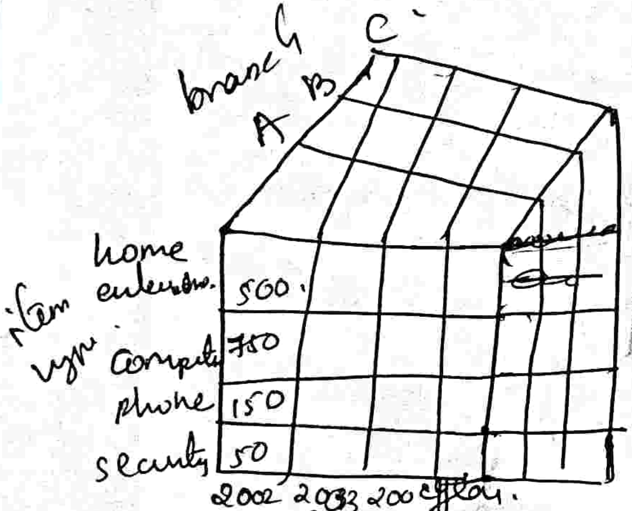


Smoothing ... from the data. Such
techniques ... ...ing, clustering

Aggregation ... ...lied to the
data:
eg: the d...
monthl...
construc...

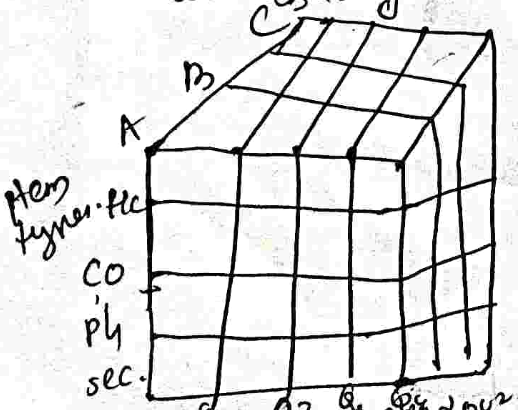
General...

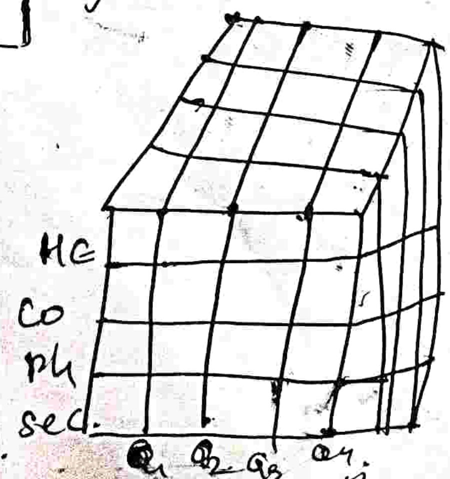




Apex

* Base cuboid,
* Apex cuboid,
→ lowest level abstraction level.
Represented.
→ highest abstract level.





techniques include binning, regression, clustering

Aggregation ... are applied to the data.

eg: the ...

mont...

contr...

Gene...

6/3/23 : Dimensionality Reduction → lossless / lossy (Approx i/p does original)

Monday

1) DWT (Discrete wavelet transform)

- used in signal processing
- original data Replaced by wavelet coefficient.

Lossless : If the original data can be reconstructed from the reduced o/p - then it is termed as lossless reduction

Lossy : If the we can only reconstructed only the approximate i/p.

2) PCA (principal component Analysis)

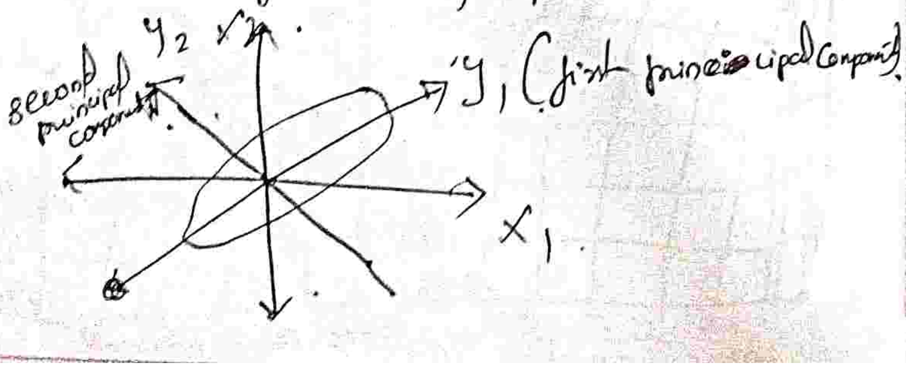
If we have n dimensional data set we have to calculate k (n dimensional) principal components where k<n.

They are orthogonal vector $\perp^r$ to original data.

Principal components are unit vectors calculated from original data i/p.

$$\hat{z} = \frac{\vec{z}}{|z|}$$

It can be diagramatically represented as.



Numerosity Red^n

To reduce the no. of counts.

(i) parametric methods

- Regression.
- Log linear models → probability concept

strictly based on parameters.

uses some parametric models.

2) Non parametric methods.

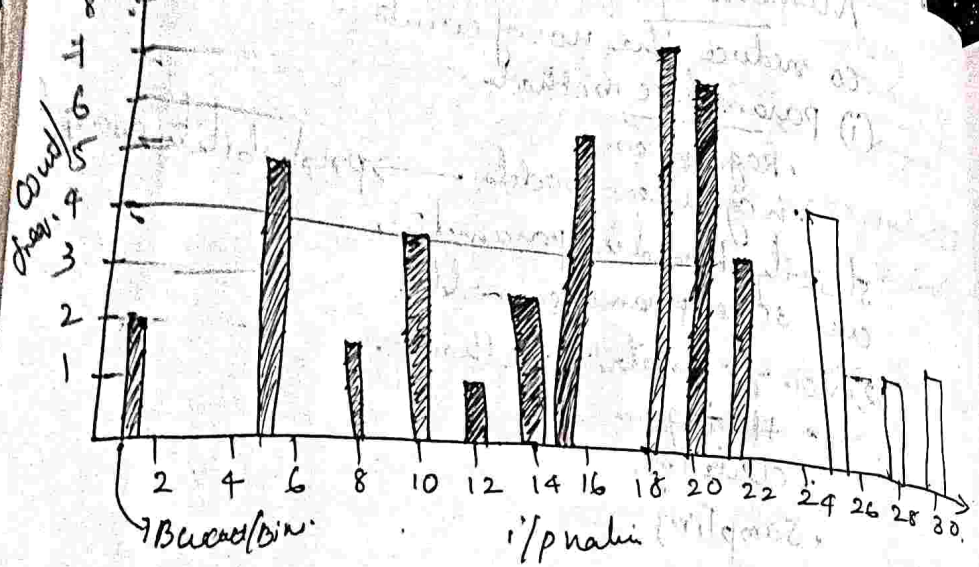
- Histogram
- clustering.
- sampling.

• Histogram

convert the data to a histogram

sorted data: 1,1,5,5,5,5,5,8,8,10,10,10,10,12,14,14,14, 15,15,15,15,15,15,18,18,18,18,18,18,18,18,20, 20,20,20,20,20,20,21,21,21,21,25,25,25,25,25, 28,28,30,30,30.

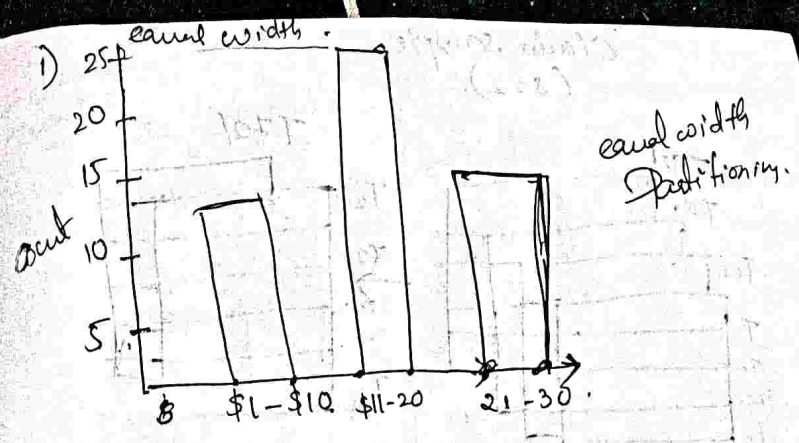
unsorted data first will sorted.



Procen of partitioning Histograms:

- (1) Equal width.
  - ↳ width of each bucket is uniform.
- 2) Equal frequency.
  - frequency of each bucket is a constant
  - Each bucket contains same no. of elements
- 3) V-optimal → Bucket with least variance
- 4) Max Diff - Diff b/w each pair of adj values.

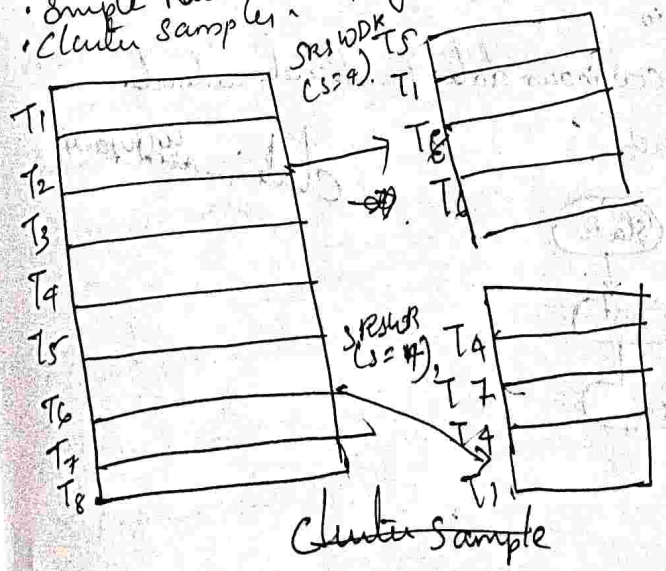
1)



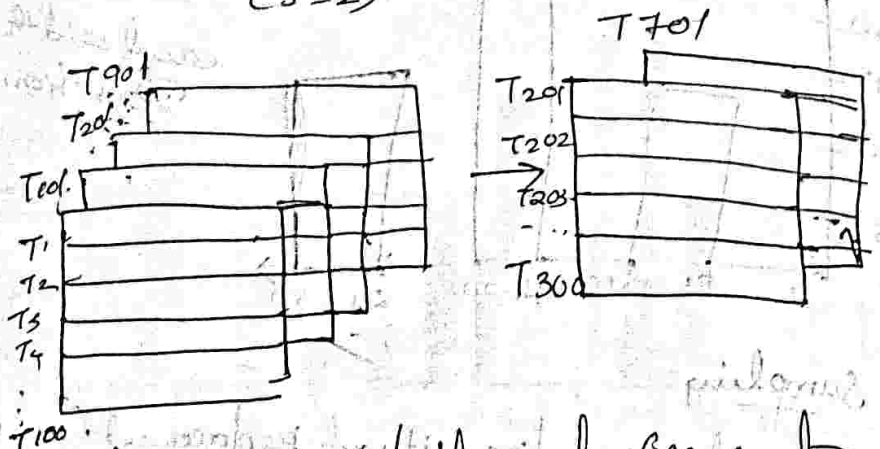
7/3/23 Tuesday

Sampling

- Simple Random sampling without Replacement (SRSWOR)
- Simple Random sampling with Replacing (SRSWR)
- Cluster sample.



Cluster sampling
(cs=2).



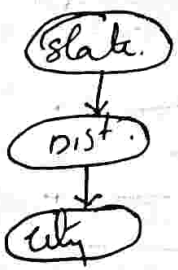
Discretization & concept Hierarchy generation.

Discretization

Converting continous datas to discrete intervals.

concept hierarchy.

Clusters - Unsupervised

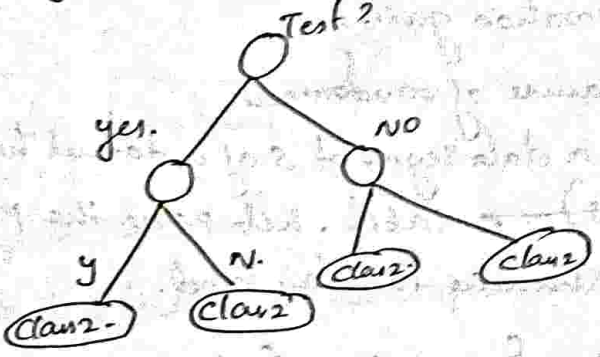


Module - 3.

16/23 Tuesday Decision tree

clustering: unsupervised

Supervised learning method.
It is a hierarchy representation of data.
Simplest type of classification method.



Nodes: Test on attribute.
Branches: outcome for the test.
leaves - final class.
Examples of decision tree algorithm.
ID3, C4.5, CART

Two types of decision tree:
1) classification → o/p are class.
2) Regression tree (predicting new value) - leaf nodes are number.

consider the following data set:

1)

Name	Features				class label.
	give birth	aquatic animal	aerial animal	has legs	
human	yes	no	~~yes~~ no	yes	mammal
Python	no	no	no	no	reptile.
Salmon	no	yes	no	no	fish
frog	yes no	semi	~~yes~~ no	yes	amphibian
bat	no	no	yes	yes	bird
Pigeon	yes	no	yes	yes	bird
cat	no	no	~~yes~~ no	yes	.
shark	yes	yes	no	no	mammal
turtle	no	semi	~~yes~~ no	yes	fish
salamander	no	semi	yes no	yes	amphibian.

8/3/23 Wednesday

Feature selection measures

Two of the popular selection measures are information gain & Gini index

CART → Gini index.

ID3 → information gain.

Entropy - measure of randomness.

Consider a data segment S of a dataset having c no: of class labels. Let p_i be the proportion of examples in S having the i^{th} class label.

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

problem(1) Total no: of samples $|S| = 10$

No: of

Probability of mamal = $\frac{2}{10}$

" " amph = $\frac{3}{10}$

" " bird = $\frac{2}{10}$

" " fish = $\frac{2}{10}$

" " reptile = $\frac{1}{10}$

$$\text{Entropy}(S) = \sum_{i=1}^{5} -p_i \log_2(p_i)$$

$$= -p_{mam}\log_2(p_{mam}) - p_{amp}\log_2(p_{amp}) - p_{bird}\log_2(p_{bird}) - p_{fish}\log_2(p_{fish}) - p_{reptile}\log_2(p_{rep})$$

$$= \frac{-2}{10}\log_2\left(\frac{2}{10}\right) - \frac{3}{10}\log_2\left(\frac{3}{10}\right) - \frac{2}{10}\log_2\left(\frac{2}{10}\right) - \frac{2}{10}\log_2\left(\frac{2}{10}\right) - \frac{1}{10}\log_2\left(\frac{1}{10}\right)$$

$$= 0.46 + 0.52 + 0.46 + 0.46 + 0.33$$

$$= 2.232$$

$$\text{Information Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{S} \times \text{Entropy}(S_v)$$

A - attributes / features

V - values of A

S - complete data set

S_v - subtable contain value v.

Let A = "gives birth"

Values of A = {yes, no}.

$$\text{Information Gain}(S, \text{"gives birth"}) = \text{Entropy}(S) - \frac{|S_y|}{|S|}\text{Entropy}(S_y) - \frac{|S_{no}|}{|S|}\text{Entropy}(S_{no})$$

$$= 2.232 - \text{Entropy}(S, \text{"gives birth=yes"})$$

$$\text{Entropy}(S, \text{"gives birth=yes"}) = -p_{mam}\log_2(p_{mam}) - p_{bird}\log_2(p_{bird}) - p_{fish}\log_2(p_{fish})$$

$$= \frac{-2}{4}\log_2\left(\frac{2}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$$

$$= 1.5$$

$$\text{Entropy}(S, \text{"gives birth=no"}) = -p_{mam}\log_2(p_{mam}) - p_{bird}\log_2(p_{bird}) - p_{fish}\log_2(p_{fish}) - p_{rep}\log_2(p_{rep}) - p_{amp}\log_2(p_{amp})$$

$= -\frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)$

$\frac{3}{6}\log_2\left(\frac{3}{6}\right) = 1.792$

Information gain $= 2.24 - \left(\frac{4}{10}\right) \times 1.5 - \left(\frac{6}{10}\right) \times 1.792$

$= 0.5712$

A = aquatic
Values of A = {Y, N, semi}
Information Gain (S, aquatic animal) = Entropy (S)

$- \frac{|S_Y|}{S}$ Entropy $(S_Y) - \frac{|S_N|}{S}$

Entropy $(S_{N'}) - \frac{|S_S|}{S}$ Entropy (S_S)

Entropy (S, "aquatic animal = yes") $= -P_{fish}\log_2(P_{fish})$

$= 1\log_2(1)$

$= 0$ (pure class)

Entropy (S, "aquatic animal = no") $= -P_{mam}\log_2(P_{mam}) - P_{rep}\log_2$

$P(rep) - P_{bird}\log_2(P_{bird})$

$= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) -$

$\frac{2}{5}\log_2\left(\frac{2}{5}\right)$.

$= 1.5$

Entropy (S, "aquatic animal = semi") $= 0$

Information gain $= 2.24 - 0 - \frac{5}{10} \times 1.5 -$

$0.$

$= 1.49$

A = aerial
values of A = {Y, N}.

Entropy (S, "aerial animal = yes") $= 0$

Entropy (S, "aerial animal = no") $= -\frac{2}{8}\log_2\left(\frac{2}{8}\right) - \frac{1}{8}\log_2$

$-\frac{2}{8}\log_2\left(\frac{2}{8}\right) - \frac{3}{8}\log_2$

$= 1.905$

Information gain (S, aerial animal) = Entropy (S)

$- \frac{|S_Y|}{|S|}$ Entropy $(S_Y) -$

$|S_N$

$= 2.24 - \frac{2}{10} \times 0 - \frac{8}{10} \times 1.905.$

$= 0.716$

A = has legs.
Value of A = {Y, N}
Entropy (S, "has legs = yes") $= -\frac{2}{7}\log_2\left(\frac{2}{7}\right) - \frac{3}{7}\log_2$

$-\frac{2}{7}\log_2\left(\frac{2}{7}\right) -$

1.5

Entropy(S, "has legs = no") = $-P_{no}\log_2(P_{no}) - P_{yes}\log_2(P_{yes})$.

$= \frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= 0.918$

Information gain = Entropy(S) $- \frac{|S_v|}{|S|}$ Entropy

$= 2.24 - \frac{7}{10} \times 1.5 - \frac{3}{10} \times 0.918$

$= 0.8856$

Aquatic animal is the root node because information gain of aquatic animal is high.

9/3/23 Thursday

2. Find the root node of Decision tree using ID3 algorithm for the following data

Age	Competetion	Type	class (profit)
old	yes	software	Down.
old	No	11 2	Down.
old	NO	Hardware	Down.
mid	yes	software	Down.
mid	yes	Hardware	Down.
mid	NO	Hardware	Up.
mid	NO	software	Up.
new	yes	s/w &	up.
new	No	H/w	Up
new	NO	s/w &	Up

Total Entropy (S) = $\sum -P_i \log_2 P_i$

$= -P_{down}\log_2(P_{down}) - P_{up}\log_2(P_{up})$

$= \frac{-5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{2}\log_2\left(\frac{5}{10}\right)$

$= 1$

A = age.

values of A = {old, mid, new}.

Entropy(S, "age = old") = $\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$

$= 1$

Entropy(S, "age = mid") = $\frac{-2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$

$= 1.$

Entropy(S, "age = new") = 0

Information gain = $1 - 0 - \frac{4}{10} \times 1 - \frac{3}{10} \times 0$

$= 0.6$

A = competetion.

values of A = {yes, no}.

Entropy(S, "competetion = yes") = $\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$.

$= 0.8112$

Entropy $(S, \text{"comp = no"}) = -\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right)$

$= 0.918$

(S, comp)

Information gain $= 1 - \frac{4}{10} \times 0.8112 - \frac{6}{10} \times 0.918$

$= 0.124$

Let A = Software Type.

Values of A = {S/w, H/w}.

Entropy $(S, \text{"Type = "S/w"}) = \frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$

$= 1$

Entropy $(S, \text{"Type = H/w"}) = \frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$

$= 1$

(S, Type)

Info gain $= 1 - \frac{6}{10} \times 1 - \frac{4}{10} \times 1$

$= 1 - 0.6 - 0.4 = 0$

Age is the root node.

Instance no.	class label	x_1	x_2
1	1	T	T
2	1	T	T
3	0	T	F
4	1	F	F
5	0	F	T
6	0	F	T

Entropy $(S) = \sum P_i \log_2(P_i)$

$= P_1 \log_2(P_1) - P_0 \log_2(P_0)$.

$= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$

$= 1$.

$A = x_1$.

Values of A = {T, F}.

Entropy $(S, \text{"}x_1 = T\text{"}) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$

$= 0.918$

Entropy $(S, \text{"}x_1 = F\text{"}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= 0.918$

Information gain $(S, x_1) = 1 - \frac{3}{6} \times 0.918 - \frac{3}{6} \times 0.918$

$= 0.082$

$A = x_2$

Values of A = {T, F}.

$= \frac{-2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$

$= 1$

$= \frac{-1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\times\log_2\left(\frac{1}{2}\right)$.

$= 1$

Information(S, ~~x(2)~~) $= 1 - 1\times\frac{2}{6}\cdot 1 - 1\times\frac{2}{6}\cdot 1/2$

$= 0$

$= 1 - \frac{4}{6}\times 1 - \frac{2}{6}\times 1.$

$= 1 - 1 = 0$

13/3/23 Monday

13/3/23 monday

Gini Index

It's a degree or probability of ~~random~~ a particular class being wrongly classified.

Gini index is used in CART algorithm.

Consider a dataset S having class labels $C_1, C_2, \ldots C_c$

Let P_i is the probability of examples having class label C_i

Then Gini index is calculated by

$$Gini(S) = 1 - \sum_{i=1}^{c} (P_i)^2$$

$P_{amp} = \frac{3}{10}$

$P_{rep} = \frac{1}{10}$

$P_{bird} = \frac{2}{10}$

$P_{fish} = \frac{2}{10}$

$P_{mama} = \frac{2}{10}$.

$$Gini(S) = 1 - \left[\left(\frac{3}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right]$$

$= 0.78$

Gini split index

Let S be a set of examples, A - attribute/feature.

S_v - subtable with $A = v$

Values(A) - set of all possible values of A.

Gini split index of an attribute A is defined by,

$Gini_{split}(S, A) =$

$$Gini_{split}(S,A) = \sum_{v \in value(A)} \frac{|S_v|}{|S|} \times Gini(S_v)$$

$Gini_{split}(S, \text{gives birth})$.

A = gives birth
Values of A = {"yes", "no"}

$$Gini_{split}(S, \text{gives birth}) = \frac{|S_{yes}|}{|S|} \times Gini(S_{gives\ birth=yes}) + \frac{|S_{no}|}{|S|} \times Gini(S_{gives\ birth=no})$$

$Gini(S_{gives\ birth=yes}) = 1 - \sum P_i^2$. Total no.

$P(mam) = \frac{2}{4}$

$P(bird) = \frac{1}{4}$.

$P(fish) = \frac{1}{4}$.

$$= 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right]$$

$= 0.625$

$$Gini(S_{gives\ birth=no}) = 1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^2 + \left(\frac{3}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right]$$

$= 0.667$

$$Gini_{split}(S, \text{gives birth}) = \frac{4}{10} \times 0.625 + \frac{6}{10} \times 0.667$$

$= 0.6502$

A = aquatic animal
values of A = {"yes", "no", "semi"}

$Gini(S, \text{aquatic animal} = yes) = 0$

$Gini(S, \text{aquatic} = no) = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right]$.

$= 0.64$

$Gini(S, \text{aquatic} = semi) = 0$

$$Gini_{split}(S, \text{aquatic animal}) = \frac{|S_{yes}|}{|S|} \times Gini(S, aa=yes) + \frac{|S_{no}|}{|S|} \times Gini(S, aa=no) + \frac{|S_{semi}|}{|S|} \times Gini(S, aa=semi)$$

$= 0.64 \times \frac{5}{10} = 0.32$

A = aerial animal
values of A = {"yes", "no"}

$Gini(S, \text{aerial} = yes) = 0$

$Gini(S, \text{aerial} = no) = 1 - \left[\left(\frac{2}{8}\right)^2 + \left(\frac{1}{8}\right)^2 + \left(\frac{2}{8}\right)^2 + \left(\frac{3}{8}\right)^2\right]$

$= 0.71875$

$$Gini_{split}(S, \text{aerial animal}) = \frac{|S_{yes}|}{|S|} \times Gini(S, aerial=yes) + \frac{|S_{no}|}{|S|} \times Gini(S, aerial=no)$$

$0 + \frac{8}{10} \times 0.71875$

$= 0.575$

A = has legs

Values of A = {"yes", "no"}

$Gini(S, has\ legs = yes) = 1 - \left[\left(\frac{2}{7}\right)^2 + \left(\frac{3}{7}\right)^2 + \left(\frac{2}{7}\right)^2\right]$

$= 0.653$

$Gini(S, has\ legs = no) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right]$

$= 0.44$

$Gini_{split}(S, has\ legs) = \frac{|S_{yes}|}{|S|} \times Gini(S, has\ legs = yes) + \frac{|S_{no}|}{|S|} \times Gini(S, has\ legs = no)$

$= \frac{7}{10} \times 0.653 + \frac{3}{10} \times 0.44$

$= 0.5891$

split attribute

Root node (split node) is selected as the attribute with lowest Gini split index. Here, age/is attribute is the lowest Gini split index.

$Gini_{split}$ index is used in CART algorithm.

$P_1 = \frac{3}{6}$

$P_2 = \frac{3}{6}$

$Gini(S) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$

$= 0.5$

$A = x_1$

values of $A = \{T, F\}$

$Gini(S, x_1 = T) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right]$

$= 0.44$

$Gini(S, x_1 = F) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right]$

$= 0.44$

$Gini(S, x_1) = \frac{|S_T|}{|S|} \times Gini(S, x_1 = T) + \frac{|S_F|}{|S|} \times Gini(S, x_2 = F)$

$= \frac{3}{6} \times 0.44 + \frac{3}{6} \times 0.44$

$= 0.44$

$A = x_2$

values of $A = \{T, F\}$

$Gain\ Gini(S, x_2 = T) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$

$= 0.5$

$Gini(S, x_2 = F) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$

$= 0.5$

$Gini(S, X_2) = \frac{4}{6} \times 0.5 + \frac{2}{6} \times 0.5$

$= 0.5$

X_1 is the root node.

14/3/23 Tuesday

Meaning

Performance of a classifier Algorithm

- confusion matrix

predicted class	Actual class: yes	Actual class: NO
yes	TP	FP
NO	FN	TN

TP → True positive
FP → False positive
FN → False negative
TN → True negative

2 class classification Problem

Confusion matrix is a model that is used to describe the performance of a classification model. It is a table that categorises predictions according to whether they match the actual values.

Precision
Recall
accuracy.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

1) Suppose a computer program for recognizing dogs in photograph identifies 8 dogs in a picture containing 12 dogs and some cats. Of the 8 dogs identified, 5 actually are dogs, while the rest are cats. Compute the precision and recall for the computer program.

predicted	Actual: dog	Actual: cat	
dog	5	3	8
cat	7		
	Sum = 12		

TP = 5
FP = 3
FN = 7

$$\text{precision} = \frac{TP}{TP + FP} = \frac{5}{5+3} = \frac{5}{8} = 0.625$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{5}{7+5} = \frac{5}{12} = 0.416$$

2) Let there be 10 balls (6 white & 4 red balls) in a box and it is required to pick 2 red balls from them. Suppose we pick 4 balls as red of which 2 are actually red balls what are the values of precision & recall in picking red ball.

predicted	Actual: red	Actual: white	
red	2	2	4
white	0	4	4
	2	8	10

TP = 2
FP = 2
FN = 0

$$\text{precision} = \frac{2}{2+2} = \frac{2}{4} = \frac{2}{2+2} = \frac{2}{4}$$

$$\text{Recall} = \frac{2}{2+0} = \frac{2}{2} = \frac{2}{2+0} = \frac{2}{2}$$

3) A database contains 80 records on a particular topic of which 55 are relevant to certain investigations. A search was conducted on that topic and 50 records are retrieved 40 were relevant. Construct the confusion matrix for the search and calculate the precision & recall and finally the accuracy whole classification problem

		Actual: relevant	Actual: not relevant	
predicted	relevant	40	10	50
	not relevant	15	15	
		55		

TP = 40
FP = 10
TN = 15
FN = 15

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{40}{40+10} = \frac{40}{50}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{40}{40+15} = \frac{40}{55}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} = \frac{40+10}{40+15+10+15} = \frac{50}{80}$$

15/3/23 Wednesday

The ID3 Algorithm

- Developed by Ross Quinlan
- Iterative Dichotomiser 3
- Assumptions
  - The algorithm uses information gain to select the most useful attribute for classification.
- Assume that there are only two class labels namely "+" and "−". The example with class labels "+" are called positive examples and others negative examples.

Notations

The following notations are used in the algorithm.

S	The set of examples.
C	The set of class labels.
F	The set of features.
A	An arbitary feature (attribute).
values(A)	The set of values of the feature A.
v	An arbitary value of A.
S_v	The set of examples with A = v.

Algorithm ID3 (S, F, C)

1. Create a root node for the tree.
2. if (all examples in S are positive) then
3. return single node tree Root with label "+"
4. end if.
5. if (all examples are negative) then
6. return single node tree Root with label "−".
7. end if.
8. if (number of feature is 0) then
9. return single nodes root tree.
10. else
11. let A be the feature in F with the highest info gain.
12. Assign A to the root node in decision tree

13. for all (values v of A) do.
14. Add a new tree branch below Root node corresponding to v.
15. if (S, is empty) then.
16. Below this branch add a leaf node with label equal to the most common class label in the set S.
17. else
18. Below this branch add the subtree formed by applying the same algorithm ID3 with the values.
ID3 (Sv, C, A-{A}).
19. end if
20. end for
21. end if.

4) Consider a two class classification problem of predicting whether a photograph contains a man or a woman. Suppose we have a test dataset of 10 records with expected outcomes and a set of predictions from our classification algorithm.

	Expected	predicted.
1	man	woman.
2	man	man
3	woman	woman
4	man	man.
5	woman	man
6	woman	woman
7	woman.	woman
8	man.	man
9	man	woman
10	woman.	woman.

a) compute the confusion matrix for the data.
b) compute the accuracy, precision recall.

predicted \ Actual	Man	woman	
Man	3	1	4
woman	2	4	6
	5	5.	

$$\text{precision} = \frac{TP}{TP+FP} = \frac{3}{3+1} = 3/4$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{3+2} = 3/5$$

$$\text{Accuracy} = \frac{3+4}{3+2+1+4} = 7/10$$

5) Suppose 10000 patients tested for flu out of them, 9000 are actually healthy and 1000 are actually sick. For the sick people a test was positive for 620 and negative for 380. For the healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the accuracy, precision and recall for the data.

predicted \ Actual	Sick	Healthy
Sick	620	180.
healthy	380	8820
	1000	9000

Precision $= \frac{620}{620+180} = \frac{620}{800} = \frac{31}{40}$.

Recall $= \frac{620}{620+380} = \frac{620}{1000} = \frac{31}{51}$.

Accuracy $= \frac{620+8820}{620+180+380+8820} = \frac{9440}{10000} = \frac{944}{1000}$

(related to Thursday) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

a) what is the mean of the data? what is the median?

(b) Use smoothing by bin means to smooth the data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

c) How might you determine outliers in the data?

d) what other methods are there for data smoothing?

e) Use min-max algorithm normalization to transform value 35 for age onto the range [0.0, 1.0]

f) comment on which method of normalization you would prefer to use for the given data, giving reasons as to why?

f) Use Z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.

g) Use normalization by decimal scaling to transform the value 35 for

a) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

mean = 29.96

b) median = 25.

b) ~~smoothing by bin means of b~~ =

Bin 1: 13, 15, 16

Bin 2: 16, 19, 20

Bin 3: 20, 21, 22

" 4: 22, 25, 25.

" 5: 25, 25, 30.

" 6: 33, 33, 35

" 7: 35, 35, 35

" 8: 36, 40, 45

" 9: 46, 52, 70.

Smoothing by bin means

Bin 1: 14.6, 14.6, 14.6. replaced by mean value

Bin 2: 18.3, 18.3, 18.3.

Bin 3: 21, 21, 21

" 4: 24, 24, 24.

" 5: 26.6, 26.6, 26.6

" 6: 33.6, 33.6, 33.6

" 7: 35, 35, 35

" 8: 40.3, 40.3, 40.3

" 9: 56, 56, 56

c) Eliminating outliers (values fall outside the data)

d) Smoothing by bin median

" " " boundaries

(e) new_max = 1.0
new_min = 0.0
V = 35.
minA = 13
maxA = 70.

$$v' = \frac{v - minA}{max_A - minA} (new\_max_A - new\_min_A) + new\_min_A$$

$$= \frac{35 - 13}{70 - 13} (1.0 - 0) + 0.$$

$$= \underline{0.385}.$$

(f) $$v' = \frac{v - \bar{A}}{\sigma A}.$$

complete data 2 to 3 digit
divided by 10^3
max digit in decimal scale dividing value based
complete data divide always

$\bar{A} = 29.96.$

$$v' = \frac{35 - 29.96}{12.94}$$

$$= \underline{0.389}.$$

(g) $$v' = \frac{v}{10^3} = \frac{35}{100} = \underline{0.35}.$$

h) Z-Score normalization because c

2) Suppose a hospital tested the age and body fat data for 12 randomly selected adults with the following result.

age	23	23	27	27	39	41	47	49	50	52	54	54	[illegible]
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	[illegible]

	58	60	61	56	57	58	58	60	61
%fat	32.9	41.2	35.7	33.4	30.2	34.1	32.9	41.2	35.9

a) Calculate the mean, median and SD of age and fat
(b) Normalize the two variables based on z-score normalization
(c) Calculate the correlation coefficient (Pearson's product moment coefficient) Are these two variables positively or negatively correlated

a) mean of age = $\underline{46.4}$

Median of age = $\frac{50+52}{2} = \underline{51}$

7.8, 9.5, 17.8, 25.9, 26.5, 27.

decision tree.

Mean of fat = $\underline{26.9}$