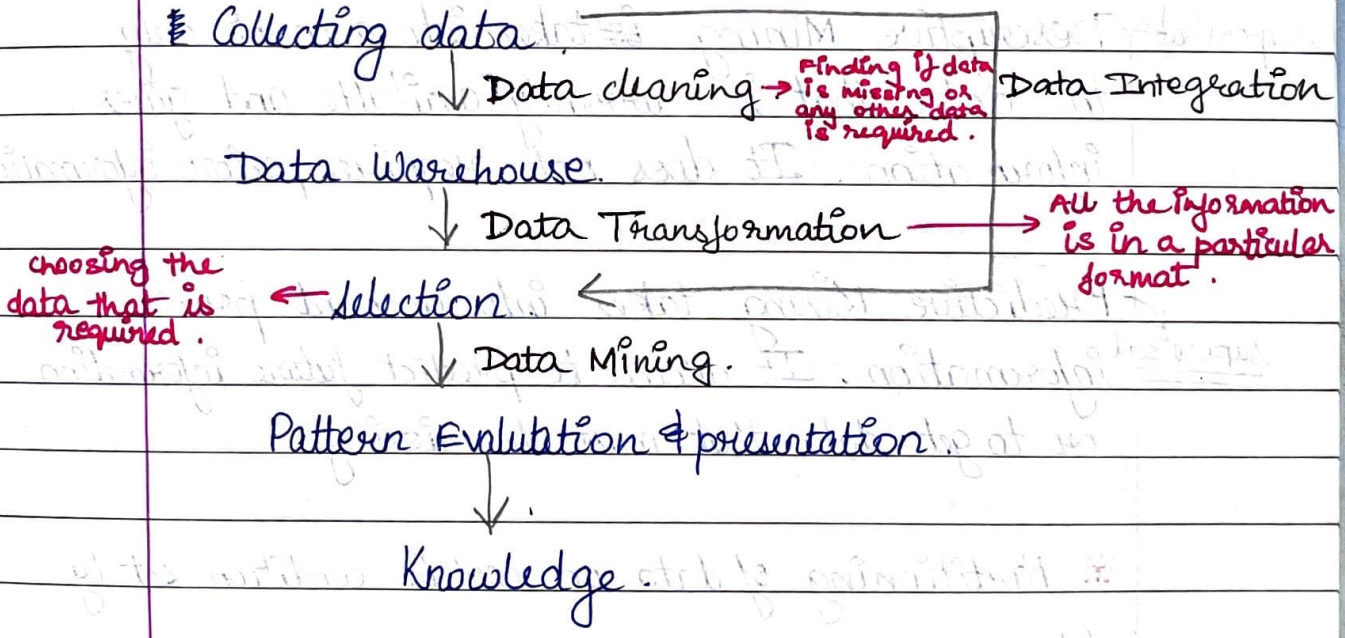
INTRODUCTION TO DATA MINING

Data mining is the extraction of data for processing data or finding data, ~~ie~~ that can be used.
It is knowledge discovery from data.

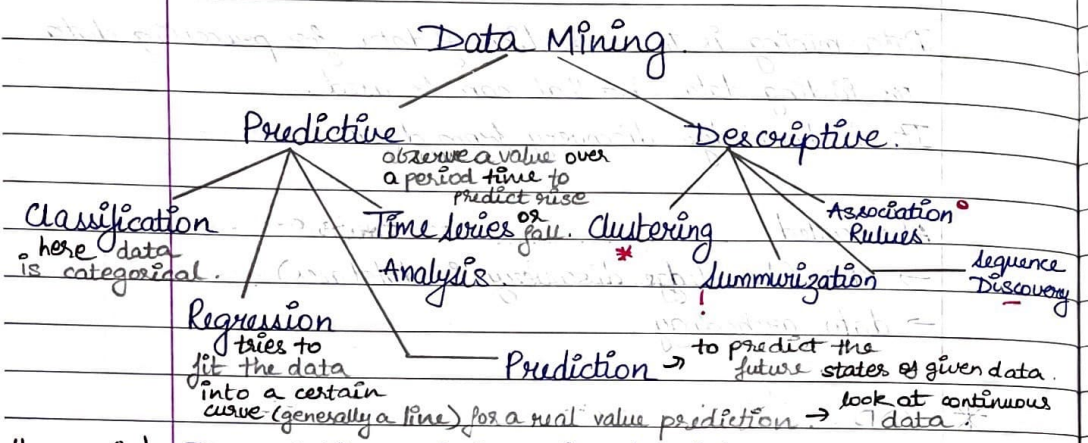
Also called
→ KDD (knowledge discovery in databases).
→ data archeology.

* KDD. (5M)



Data mining is required to find information that can be related to each other. while query processing can be done to only find the information from already available data.

* DATA MINING MODELS AND TASKS



Unsupervised ← Descriptive Mining ~~is~~ takes into account only the data that is currently available and gives information. It does not require prior information.

↙ Predictive Mining takes into account prior information. It allows to predict future information or to get data that must be missing.

Supervised

* Partitioning of data on certain condition set by us.

! describing the data or summarizing or generalisation of data.

° relationship between different data.

this is in same time ← eg. Market-basket analysis used in supermarkets to increase the sale of items.

– to determine a pattern between different data.

~~to make~~ ↘ this is done over a period of time (eg. a month).

* TYPES OF DATA.

1. DATASET

Consists of datapoints or data object.

Can be mapped in any dimension.

Usually mapped in ~~any~~ two dimension.

Each datapoint have attributes or dimensions.

am is generally label class ~~variable~~

di = <ai1, a2i, a3i, ... ami>

↑ FEATURE / ATTRIBUTE VECTOR

Observed values for a given attribute are known as observations.

A set of attributes used to describe a given object is called attribute vector / feature vector.

* DATA OBJECTS & ATTRIBUTE TYPE

1. Nominal / Categorical attributes

Related to names; Values are symbols or names.

Cannot find mean, median but mode can be found easily.

eg. Hair styles, color

Nominal values may be numeric.

They are not ~~quantative~~. quantitative.

2. Binary Attributes.

Can only have 2 states (0/1)

0 – value absent 1 – ~~0~~ value present

Two types.

→ Symmetric ⇒ both have same weight, are equally probable.

→ Asymmetric binary ⇒ both states are not equally valuable. Here one state is more favourable than the other. eg. In medical test, more favourable one is the than -ve

3. Ordinal attributes.
In the form of ranks.
eg Grades ⇒ A+, A, A-, B ....
They have a definite order.

4. Numeric attributes.
Continuous data ; Measurable Quantity.
Two types.
⇒ Interval scaled.
Can be compared, find the difference
Here '0' means absence of ~~the~~ value for attribute.
⇒ Ratio scaled
Cannot be compared, values are ordered.
Mean, median & mode can be found.
Here '0' has a meaning on the scale.
eg. Temperature.

DISCRETE → finite number of values.
eg. medical-test. ↳ points.

CONTINUOUS → ~~numb~~ numeric values.
eg. age.,

COUNTABLE INFINITY → values can grow infinite but still countable. ↳ Needs a function to be plotted.
eg. zipcode.

* MEAN, MEDIAN, MODE

Mean, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Weighted mean $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{w_i}$.

⇒ Median $= L_1 + \left[\frac{N/2 - (\Sigma freq)_e}{freq_{median}}\right] width$

L_1 → lower boundary of the median interval
N → no. of values.
$freq_e$ → sum of frequencies of intervals lower than the median interval.
$freq_{median}$ → frequency of median interval.
width → width of median interval.

Mode → values that occurs more frequently.
Types.
→ Unimodal → Bimodal
→ Trimodal → Multimodal.

Midrange → Average of largest and smallest values in the set.

Grouped mean $= \frac{\Sigma(f_i \times x_{im})}{n}$.

n → total number of observations.
f_i → frequency of i^{th} observation
x_{im} → midpoint of i^{th} x_m.

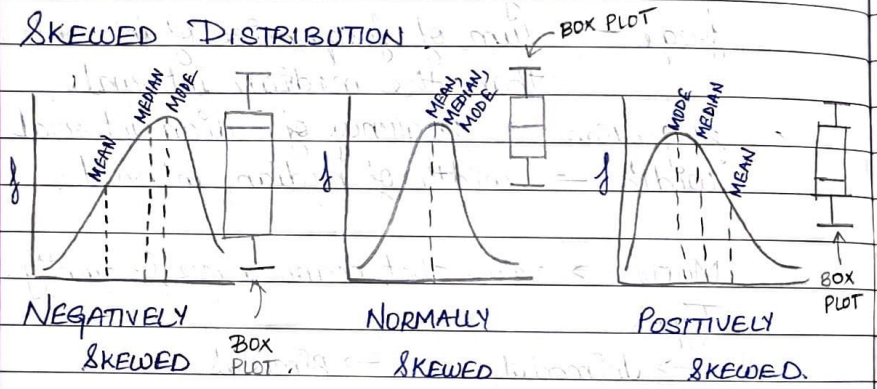
Grouped median :
We first find where N/2 observation lies in which range. & use the above median formula.

Page No.: Date: YOUVA

$$\text{Grouped mode} = L + \left[\frac{f_m - f_{m-1}}{(f_m - f_{m+1}) + (f_m - f_{m-1})}\right] \times \text{width}$$

L → lower limit of the group with the mode (group with highest frequency)
f_m → Frequency of group with mode
f_{m-1} → Frequency of group ~~with the~~ before f_m.
f_{m+1} → Frequency of group after f_m.
width → width of the groups.

✱ SKEWED DISTRIBUTION.



NEGATIVELY SKEWED | NORMALLY SKEWED | POSITIVELY SKEWED.

BOX PLOT.

MEASURING THE DISPERSION DATA.

Range → difference between largest max & smallest min values.

Quartiles → points taken at regular intervals in a distribution.

4-points : Q_1 (25%), Q_2 (50%), Q_3 (75%)

Interquartile range → $IQR = Q_3 - Q_1$.
↳ Used to find if they are any outliers in the distribution.

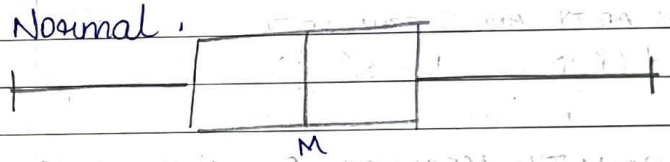
Five no. summary → min, Q_1, M, Q_3, max.

Page No.: Date: YOUVA

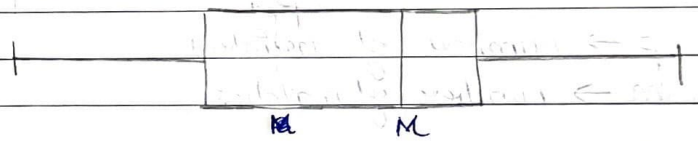
Box plot

Outliers → beyond min & max.
Minimum $Q_1 - 1.5 * IQR$.
Maximum $Q_3 + 1.5 * IQR$.

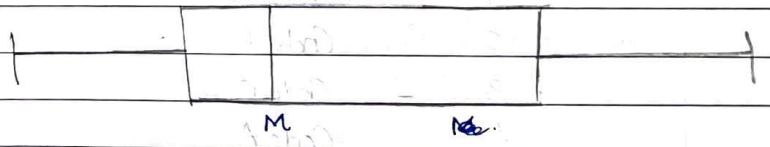
Normal.



Negatively skewed.



Positively skewed.



Varience

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

Standard Deviation

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Z-score

$$Z_{ij} = \frac{x_{ij} - m_j}{s_j}$$

Types of plot:

1. Quantile Plot.
2. Q-Q Plot.
3. Histograms.
4. Scatter plot
5. Loess curve.

* SIMILARITY AND DISSIMILARITY.

$$d(i,j) = 1 - s(i,j)$$

DISTANCE MATRIX
DISSIMILARITY.

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

PROXIMITY MEASURE FOR NOMINAL / CATEGORICAL DATA

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

p → number of variables.
m → number of matches.

eg.

Obs Id.	Test 1
1	Code A
2	Code B
3	Code C
4	Code A.

$p = 1$

$m = \begin{cases} 0, & \text{if the do not match.} \\ 1, & \text{if they match.} \end{cases}$

$$d(1,2) = \frac{1-0}{1} = 1.$$

PROXIMITY MEASURE FOR BINARY VARIABLES.

Contigency table.

		Object j = 1	Object j = 0	sum
Object i	1	a	b	a+b
	0	c	d	c+d
	sum	a+c	b+d	p.

a → no. of attributes i & j = 1
b → no. of attributes. i = 1 & j = 0.
c → no. of attributes. i = 0 & j = 1
d → no. of attributes. i = 0 & j = 0.

Distance measure for symmetric binary variables

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

Distance measure for assymmetric binary variables

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

Jaccard coefficient (similarity measure for assymmetric binary variables)

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

DISSIMILARITY OF NUMERIC DATA.

EUCLIDIAN DISTANCE

$$d(i,j) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2}$$

MINKOWSKI DISTANCE

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

MANHATTAN FORMULA:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

PROXIMITY MEASURE FOR ORDINAL DATA:

→ Has to be normalized using z-score.

$$z = \frac{r_{ij} - 1}{M_j - 1}$$

(r_{ij}: rank of x_{ij}; M_j: Total no. of ranks.)

For ratio-scaled, we find its log values and then apply Manhattan / Euclidian formula.

For numeric data, we can find the minimum & maximum and then use the difference between them as denominator.

For mixed type ←

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

* DATA PREPROCESSING.

Data collected should not be
→ incomplete → noisy → inconsistent

Data collected should be preprocessed for the following reasons.
→ accuracy.
→ completeness.
→ consistency.
→ timeliness.
→ believability.
→ interpretability.
→ accessibility.
→ value added → eg: giving meta data with the data for source information.

Major tasks in Data Preprocessing.

1. DATA CLEANING.
→ Identifying incomplete data.
→ resolve inconsistencies.
→ smooth noisy data.

2. DATA INTEGRATION.
Integrate into multiple databases.

3. DATA TRANSFORMATION.
Normalization & reduction.

4. DATA REDUCTION.

5. DATA DISCREATION.

How to handle Missing Data?
→ Ignore the tuple: usually done when class label is missing. This may result in loss of imp. data.
→ Fill it automatically:
For this we can find the most probable value – regression, inference based such as Bayesian formula or decision tree.
→ Filling manualy can be tedious & infeasible.
→ Using a constant value for replacing missing data may result in the formation of a new class.

How to handle Noisy data?
Binning →
data preprocessing method used to minimize the

effects of small observations errors.
Smooth the data by looking at neighbouring data.

Regression →
smooth by fitting the data into regression functions.

Clustering →
detect and remove outliers. (clusters use centroids)

Combined computer & human inspection.

* BINING METHOD FOR SMOOTHING DATA.

Sorted data → 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into equal-frequency (equi-depth) bins.
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

* DATA CLEANING.

→ Data discrepancy detection.
- make use of meta data (eg. domain, range, etc.)
- check field overloading.
- check unique rule, consecutive rule & null rule.
- use commercial tools.
  - → Data scrubbing: use simple domain knowledge to detect errors and make corrections.
  - → Data auditing: by analysing data to discover rules & relationship to detect violations.

→ Data migration & integration.

* DATA INTEGRATION.

→ Combine data from multiple sources into a coherent store.
→ First step. Entity Identification problem.
- To identify the different attributes and their relationship.
- Mapping of data
- To associate similar data.

→ Second step. Redundancy & Correlation Analysis.
- Redudancy can be found using correlation analysis.

CORRELATION ANALYSIS.

χ^2 - TEST. (CATEGORICAL DATA)

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Larger the value of χ^2, more likely the variables are related

PEARSON'S PRODUCT ~~OF~~ MOVEMENT COEFFICIENT (NUMERICAL DATA).

$$r_{A,B} = \frac{\sum (A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A \sigma_B}$$

$$= \frac{\sum (AB) - n(\bar{AB})}{(n-1)\sigma_A \sigma_B}$$

$r_{A,B} > 0$ → positive correlation

$r_{A,B} < 0$ → negative correlation

$r_{A,B} = 0$ → no relation

$$r_{A,B} = \frac{COV(A,B)}{\sigma_A \sigma_B}$$

$$COV(A,B) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}$$

* DATA REDUCTION.

HISTOGRAMS.

→ equal-frequencies.

→ equal-width.

Clustering.

Sampling

Cluster or stratified sampling

CLASSIFICATION AND PREDICTION

What is classification?

→ Data is generally discrete.

→ Data ~~is also~~ can also be categorical

→ It is supervised learning.

→ Data is classified based on previous historical data.

* SUPERVISED LEARNING.

Classification.

The trai

* CLASSIFICATION – A TWO-STEP PROCESS.

1. MODEL CONSTRUCTION.

→ describing a set of predetermined classes.

→ model is created on a training set.

2. MODEL USAGE

→ for classifying future or unknown object.

→ model is chosen based on its accuracy.

* ISSUES.

Data preparation.

→ Data cleaning.

→ Relevance analysis.

→ Data Transformation.

Evaluating Classification Methods.

→ Accuracy
  Classifier accuracy - predicting class label
  Predictor accuracy - guessing value of predicted attr.

→ Speed

→ Robustness.

→ Scalability.

→ Interpretability.

* BAYESIAN CLASSIFICATION.

Based on Baye's Theorem.

→ Prior knowledge is required for probability as with new ~~entries~~ entries probability may increment or decrement.

X → data sample ; class label is unknown.

H → Hypothesis that X belongs to class C.

P(H|X) → probability that the hypothesis holds given the observed data sample X.

P(H) → prior probability, initial probability.

P(X) → probability the sample data is observed

P(X|H) → posteriori probability.

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}.$$

Informaly, it can be written as

posteriori = likelihood x $\frac{\text{prior}}{\text{evidence}}$

Classification is done on the basis of maximum posteriori.

EXAMPLE ① - For categorical / discrete data.

P(Ci) - Prior probability of ~~class~~ each class.

$P(\text{buys\_comp} = \text{"yes"}) = \frac{9}{14} = 0.643$.

$P(\text{buys\_comp} = \text{"no"}) = \frac{5}{14} = 0.357$.

Compute likelihood of P(X|Ci) of each attribute.

$P(\text{age} = \text{"<=30"} \mid \text{buy\_comp} = \text{"yes"}) = \frac{2}{9}$

$P(\text{age} = \text{"<=30"} \mid \text{buy\_comp} = \text{"no"}) = \frac{3}{5}$.

$P(X|C_i) = P(X|buys\_comp = "yes")$

$= 0.044.$

$P(C_i|X) = P(X|C_i) * P(C_i)$

$P(buys\_comp = "yes"|X) =$

$= 0.028.$

$P(buys\_comp = "no"|X) =$

$= 0.007$

$P(buys\_comp = yes|X) > P(buys\_comp = no|X)$

$\therefore$ X belongs to class (buys_comp = "yes").

If any instance of a class is missing, the probability will be zero, does it is directly classified into other class.

EXAMPLE ② – For continuous attributes.

$P(class = "short") = \frac{4}{15}$

$P(class = "medium") = \frac{8}{15}$

$P(class = "tall") = \frac{3}{15}.$

$X = (gender = M, height = 195)$

$P(gender = "M" | class = "short") =$

$P(gender = "M" | class = "medium") =$

$P(gender = "M" | class = "tall") =$

↖ Categorical / Binary

$P(x_k|C_i) = g(x_k, \mu_{ci}, \sigma_{ci})$

$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

continuous.

Shortest "short" — tallest "short"

$P(height = 1.95 | class = "short") = Range(1.6m - 1.7m)$

$\mu = 1.65 \quad \sigma = 0.05$

$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi} \times 0.05} e^{-\frac{(1.95-1.65)^2}{2 \times 0.05^2}}$

$= 1.2154 \times 10^{-7}$

P(height = 1.95 | class = "medium") = Range (1.75-1.95)

$\mu = 1.85$ & $\sigma = 0.1$

$g(x, \mu, \sigma) =$

P(height = 1.95 | class = "tall") = Range (2m - 2.2m)

$\mu = 2.1 \quad \sigma = 0.1$

$g(x, \mu, \sigma) =$

$= 1.29.$

Posteriori Probability of C conditioned on X.

Example ③

X = <brown, tall, average, no>

Prior probability

$P(\text{class} = \text{"Sunburnt"}) = \frac{6}{10}$

$P(\text{class} = \text{"None"}) = \frac{4}{10}.$

$P(\text{brown} \mid \text{sb}) = \frac{2}{6}$

$P(\text{brown} \mid \text{none}) = \frac{3}{4}$

$P(\text{tall} \mid \text{sb}) = \frac{1}{6}.$

$P(\text{tall} \mid \text{none}) = \frac{1}{4}.$

$P(\text{avg} \mid \text{sb}) = \frac{2}{6}.$

$P(\text{avg} \mid \text{none}) = \frac{2}{4}.$

$P(\text{no} \mid \text{sb}) = \frac{5}{6}.$

$P(\text{no} \mid \text{none}) = \frac{3}{4}.$

Posteriori Probability of class on X.

$P(X \mid \text{sb}) = \frac{6}{10} \times \frac{1}{6} \times \frac{2}{6} \times \frac{5}{6} \times \frac{6}{10}.$

$= 9.34 \times 10^{-3}.$

$P(X \mid \text{none}) = \frac{3}{4} \times \frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} \times \frac{4}{10}.$

$= 0.028.$

So X belongs to class = "none"

Between 2 data, no conditional dependay is present. If 2 data are conditionally dependent then Naive Bayes cannot be used.

LAPLACE CORRECTION

Avoiding the 0-Probability Problem.
We add a tuple ~~to~~ ~~e~~ for each type of attribute.

For example.
- Income (low) = 0
- Income (med) = 990
- Income (high) = 10.

We add 1 to each case.
- Prob (Income (low)) = 1/1003.
- Prob (Income (med)) = 991/1003
- Prob (Income (high)) = 11/1003.

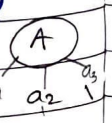
The probability will be very small, but posterior probability won't decrease.

✳ DECISION TREE INDUCTION

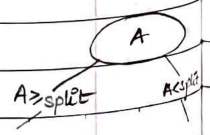
ID 3 — INTERACTIVE DICHOTOMIZER 3.

Internal Nodes — Attributes
Leaf Node — Classes.
Branches — Values of Attribute.

For categorical,
→ Branches are based on the value of attributes.



For continuous,
→ A split point is found



~~Algorithm~~

Attribute selection: Information Gain ID3.
Select attribute with the highest information gain.

p_i → probability that an arbitrary tuple in D belongs to class C.

$$\frac{|C_{i,D}|}{|D|}.$$

Expected Information needed to classify a tuple in D.

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Information needed (after using A to split D into v partitions) to classify D.

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j).$$

Information gained by branching on attribute A.

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D).$$

Example ①.
Class → buys_comp = "yes"
buys_comp = "no".

$$\text{Info}(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= 0.940 \text{ bits.}$$

age	pi	ni	I(pi, ni)
<=30	2	3	0.971
31..40	4	0	0
>40	3	2	0.971

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2)$$

$\frac{5}{14} I(2,3)$ → means "age <=30" has 5 out of 14 sample, with 2 yes's & 3 no's.

$$I(2,3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$
$$= 0.971$$

$$Info_{age}(D) = \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971$$
$$= 0.694 \text{ bits}$$

$$Gain(age) = Info(D) - Info_{age}(D)$$
$$= 0.940 - 0.694$$
$$= 0.246$$

$$I_{income}(D) = \frac{4}{14} I(2,2) + \frac{6}{14} I(4,2)$$
$$+ \frac{4}{14} I(\ ,$$

* the metadata in the data warehouse should be same for every users viewing it. (VIVA Q's → Evolution)

INTRODUCTION TO DATA WAREHOUSE.

* DATA WAREHOUSE

→ single*, complete & consistent store of data obtained from a variety of different sources made available to end users in ~~a~~ what they can understand & use in a business context.

* EVOLUTION.

1960's → Batch reports.
was hard to find & analyse information.

1970's → Terminal based Decision systems (DSS) and Executive information systems (EIS).

1980's → Desktop data access

1990's → Data warehousing.

Terabytes $- 10^{12}$ eg. Walmart data.
Petabytes $- 10^{15}$ Geographic Information Sys.
Exabytes $- 10^{18}$.
Zettabytes $- 10^{21}$
Zottabytes $- 10^{24}$

Data Warehouse is the collection of historic data
Database is the collection of data that is used for day to day transactions
Data warehouses are not replacing databases

OLTP → Online Transaction Process
OLAP → Online Analytic Processing

What is (Banks)
Types of account present
Number of transactions.
Withdrawals & Deposits done.

Data granularity → data should be in summarized format.

DEFINITION *

A data warehouse is a:
→ subject-oriented. eg. Sales, Product, Customer.
→ integrated. → data inconsistency is removed.
→ time-varying → historical data, snapshots of data, allows analysis of past, relates info to present.
→ non-volatile → data cannot be changed.

collection of data that is used primarily in organizational decision making.

* INFORMATION PACKAGE.

What were the sales of Godrej refigerators of white colour in the year 2010 January in Juhu area? → Query.

A.

DIMENSIONS

Product	Time	Region.
Colour(white)	Year(2010)	Juhu
	Month(Jan)	

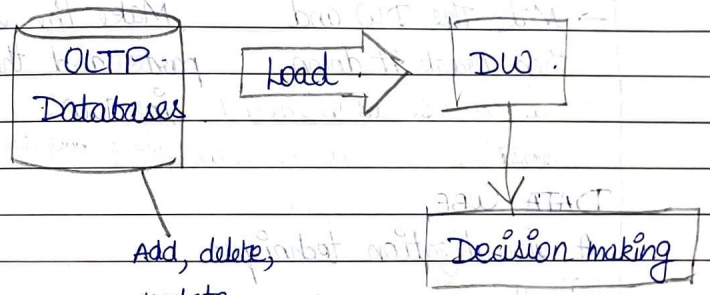
Since data collected is historical, 'time' dimension must always be included.

ER — IP.
Entity → Dimensions
Relationships → Facts.

Pg 120 → DW by Ponniah — IP examples.

(Databases) (Data Warehouses).

OPERATIONAL (OLTP)	INFORMATIONAL. (OLAP)
→ Current values	Archived, derived, summarized
→ Optimized for transaction.	Optimized for complex queries
→ High access frequency.	Medium to low
→ Read, update, delete	Read.
→ Predictable, repetitive usage.	Adhoc, random heuristic usage.
→ Response time in sub-seconds	several seconds to minutes.
→ Large number of users	Relatively small



3 data granularity levels in the bank.
Daily Monthly Quaterly.

For a data cube with dimensions more than 3, are called hypercube.

Identify the subjects for Airline

Page No.: Date: YOUVA

Airline.

Airline Name
Airline sales.
No. of passengers
Cost.
Departure Time
Arrival Time.
Type of seat.
Domestic/Int.
Departure
Destination.
No. of stops in between
No. of seats left.
Flight time.
No. of kg of baggage
Services.
Time seat was booked.

Top down Approach
→ Make the DW and then break it down. (usually should be done)

Bottom Up Approach
Make the small parts and then integrate in DW.
(data marts → dw).

DATA CUBE
A visualization technique.

⇒ IP for Airlines.

TIME	FLIGHT	CUSTOMERS	BOOKING	ROUTE
Year	No. of seats booked	Name	Company	Departure
Month	Type of flight	Gender	Booking Date	Arrival
Day of the Month	Type of classes.	Age		No. of stops
Day of the Week.	Cost	No. of luggage		Time taken
Hour	Services.			
Minute.	Total no of seats.			
Season				
Quarterly				

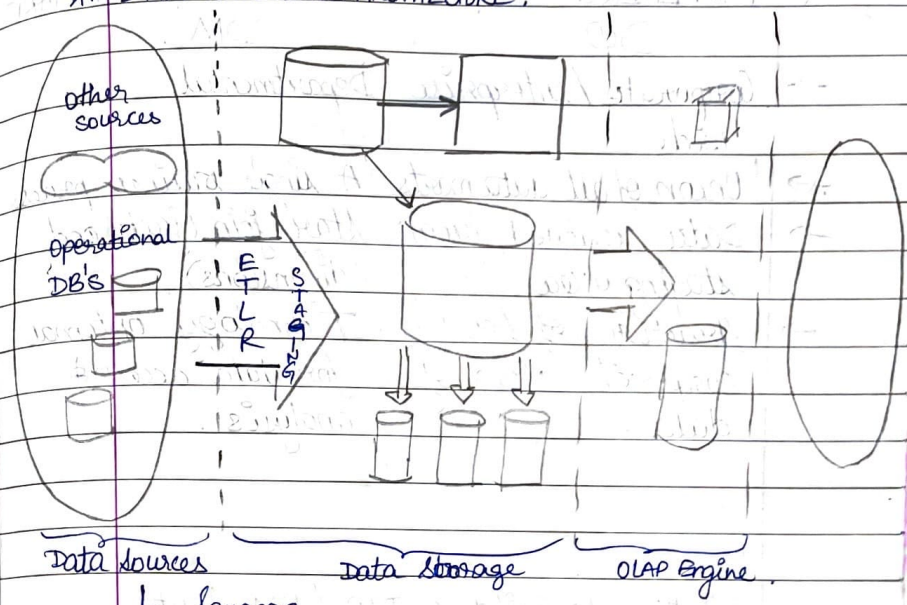
FACTS: No. of seats vacant, most used route, Time flight is booked, Time when most seats are booked, Class in which seats are booked more, if catering was chosen

In paper why we have chosen the given dimension, subject of analysis ~~flight~~.

Meta data – data above data.

Page No.: Date: YOUVA

* DATA WAREHOUSE ARCHITECTURE:



1. SOURCES

2. STAGING.
→ ETL Process.
Extract Transform Load Refresh.

Meta data – varies from business to business. It stores the rules to handle data present correctly.

Data Mart – Every individual business processes have a mini data warehouse in large organizations. These are called data marts. These data refer to one large data warehouse.

3. OLAP.

4. Analysis.

All the ~~data~~ algo present ~~are~~ for AI, DS are applied in stage 3 & 4

IMP

* DIFFERENCE ~~BE~~ / DATA WAREHOUSE & DATA MART

	DW	DM.
→	Corporate / Enterprise wide	Departmental
→	Union of all data marts	A single business process.
→	Data recieved from staging area.	Star join (facts and dimensions).
→	Structure of for corporate view of data.	Technology optimal for data access & analysis.

Whether to Build a DW / DM first.

→ Top down Approach
First data warehouse.
- Truly a ~~cop~~ corporate effort.
-

→ Bottom up Approach.

* POSSIBLE ETL FAILURES.

→ A missing source file.
→ A system failure.
→ Inadequate metadata.
→ Poor mapping information.
→ Inadequate storage planning
→ A source structural change.
→ No contigency plan.
→ Inadequate data validation.

* DESIGN DECISIONS.

→ Choosing the process.
- select subjects from IP.

→ Choosing the grain.
- level of detail is important

→ Identifying and conforming the dimensions.
→ Choosing the facts.
- selecting metrics, units of measurement.

→ Choosing the duration of database.
- how much historical data is required.

* CRITERIA FOR COMBINING TABLES INTO DIMENSION TABLE

→ Model should provide best data access.
→ Whole model should be query centric.
→ Optimized for query and analysis.
→ Every dimension can interact equally with fact table.
→ Should allow ~~table~~ drill up and drill down, along dimensional hierarchy.

* STAR SCHEMA.

A star schema represents the relationship between the different dimensions ~~and it~~. The dimensions directly interact with the fact table.