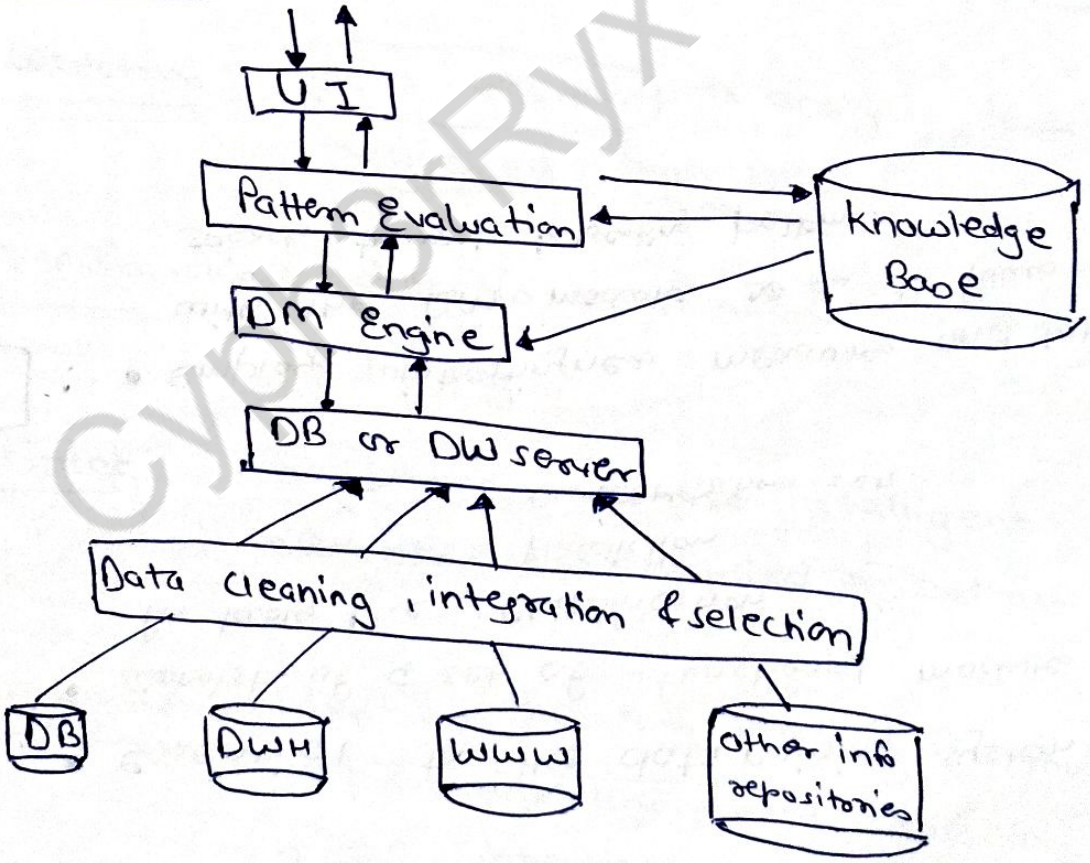
Annexure No :

DATA MINING.

Data mining:

- extraction of knowledge from large ammou of data.
- reffered as knowledge Discovery from data.

Architecture of DM system:



Knowledge Base :

- Domain that is used to search or evaluate the interestingness of patterns.
- This contains user's previously used data and beliefs to assess a pattern's interestingness.

DM engine :

- essential to the data mining system
- consists of a set of functional modules for tasks like classification
  prediction
  characterization etc.

Pattern evolution :

- employs interestingness measures and interacts with the DM modules so as to focus the search toward interesting patterns.

* KDD (Knowledge Discovery from Data)

→ Pre processing operations.

↳ required to make pure data in DWH before using it in DM processes.

1) Data cleaning: Remove noisy & inconsistent data.

2) Data integration: multiple sources are combined.

3) Data selection: data relevant to analysis task are retrieved from DB.

4) Data transformation: data is transformed into forms for mining.

5) Data mining: An process where intelligent methods are applied to extract data

6) Pattern Evaluation: identify the truly interesting patterns representing knowledge based on some interestingness measures.

7) knowledge Patterns: visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Issues in DM:

(1) Performance:

→ as data multiply these factor become critical
→ the algos of DM to extract huge ammount of data
→ the runtime & execution time of DM algo must be short and predictable.

(2) Parallel, distributed & incremental mining algo:

→ Algo's partitions data into pieces

- each piece is processed in parallel by searching for patterns.
- The pattern will eventually merge together.

(3) Diverse DB:

(i) Handling of complex type data:

→ one DM system cant mine all kind of data

(ii) Mining dynamic & global data repos:

→ multiple sources of data are connected by internet forming a huge network of data.

(4) Mining methodology:

(i) mining in multidimensional space
(ii) mining various and new kind of knowledge
(iii) handling noisy data
(iv) Pattern evaluation

APPLICATION of DM:

- market analysis
- fraud detection & management
- Customer retention
- quality analysis
- customer feedback analysis
- Risk analysis & management
- Target marketting
- Web Analysis
- Text mining from documents, emails, news letters.

Disadvantages of DM:

PRIVACY

- Collection of user data
- data leaks / breach
- misuse of data collected
- Buisness politics.

Classification of DM:

(i) Database Technology

(ii) Statistics

(iii) ML model

(iv) Information Science

(v) Visualizations

(vi) other disciplines

Sources of data:

1. WWW
2. Organization
3. Survey & services
4. Hard papers
5. IOT devices.

DM architecture:

(1) No coupling:

- dont use any service from DB or DWH
- data extracted from source
  - Processed → DM algos
- Poor architecture for DM

(2) Loose coupling:

- Uses services from DB or DWH.
- memory based DM.
- data extraction → Processing → DM algos → stores in system.

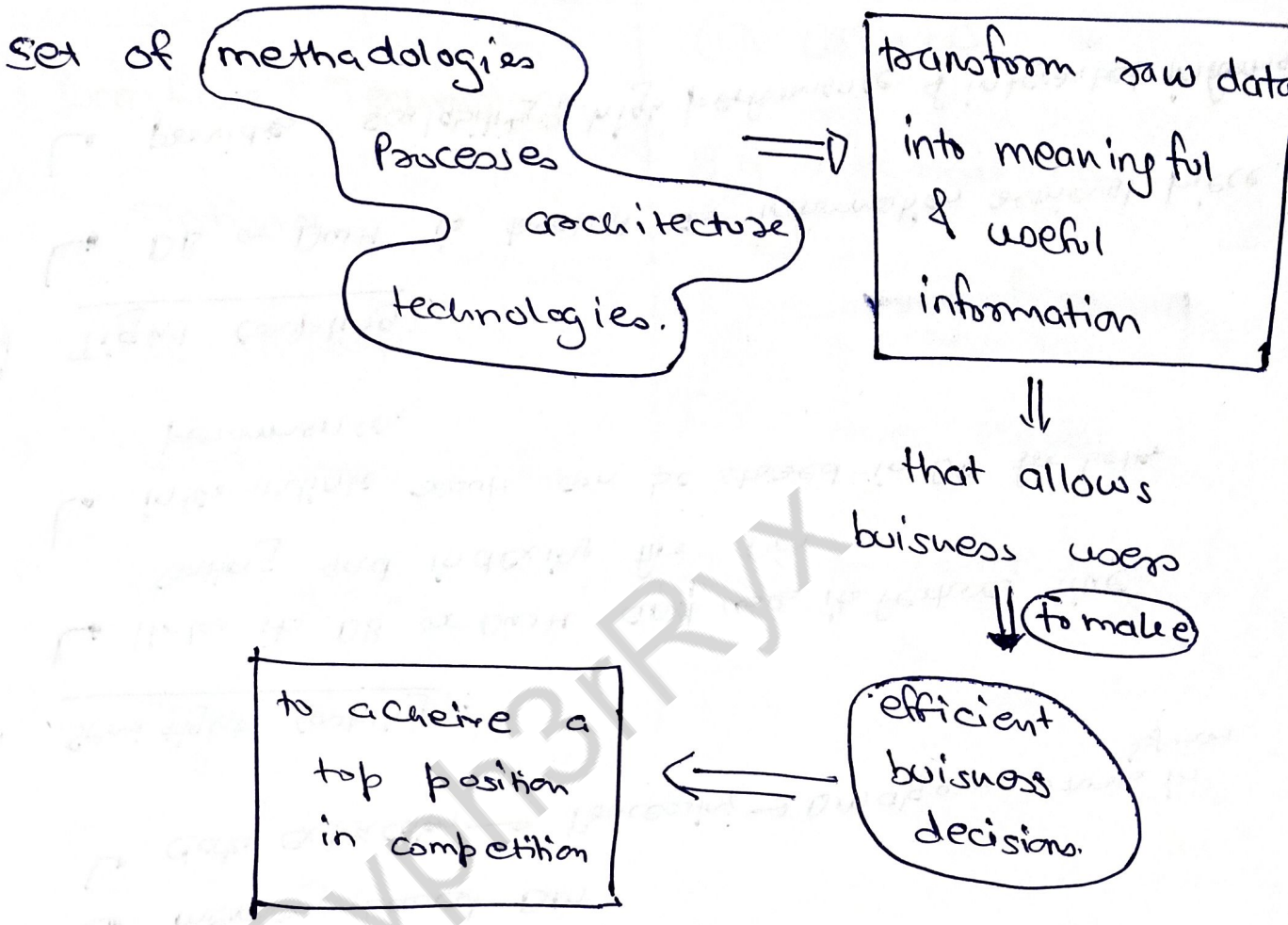
(3) Semi tight coupling:

- links its DB or DWH and uses its features like sorting and indexing for DM
- intermediate result can be stored in DB for better performance.

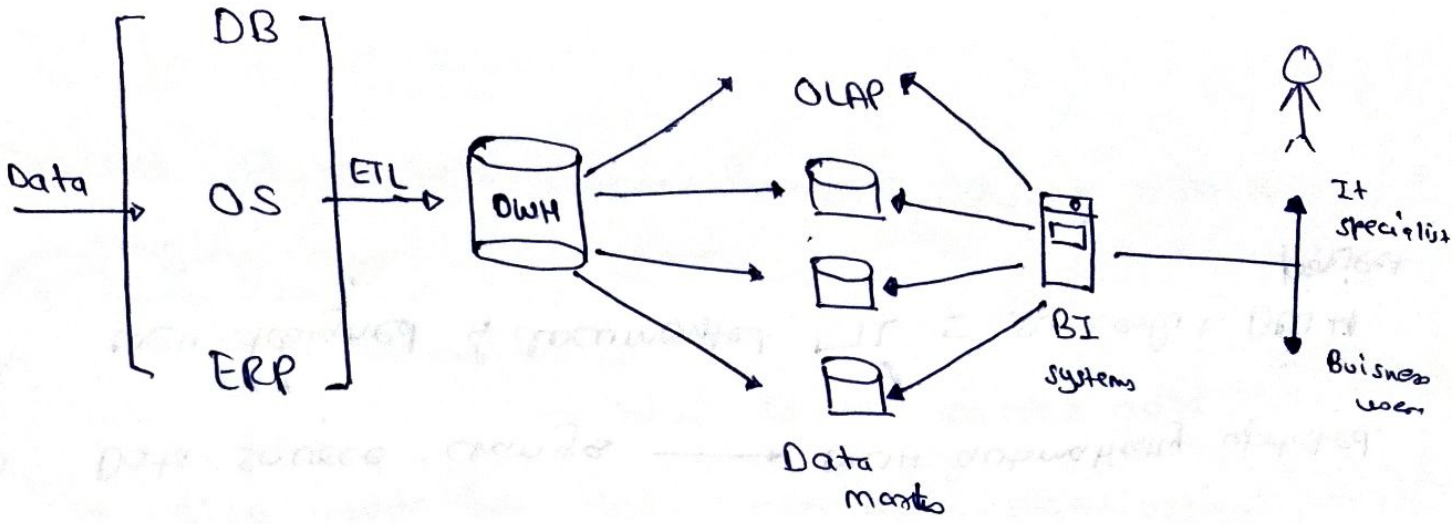
(4) Tight coupling:

- DB or DWH is treated as information retrieval piece.
- provides scalability, high performance & integrated informati

Buisness Intelligence:



Annexure No :



- Data from OS, DB & ERP were extracted & transformed into a more consumable form.
- Data from a DWH is then loaded to OLAP cubes
- Data marts are also filled with data from DWH.
- OLAP cubes facilitate the analysis of data
- With the BI system, IT specialist can setup the system for BI user so that they can analyze & access the data easily from BI system.

ETL Process: Extract – Transform – Load

Process of data integration.

Takes up large volume of raw data ~~~~ Converts it for analysis

Data to DWH ← LOADS

Why do we need ETL?

1. Analysing data
2. Transactional DB can't answer complex business questions
3. Provides method of moving the data from various sources into a DWH.
4. Data source changes → DWH automatically updated.
5. well designed & documented ETL = succesful DWH project

Process ① : Extraction :

Extracted data sets comes from a source into a staging area

↳ acts as buffer

↳ b/w DWH & source data.

• also used for data cleaning & organization

↳ because data might corrupt as it comes from various sources.

Challenge: How ETL handles structured & unstructured data?

↳ Create a custom solution to assist in filtering the unstructured data.

Process ② : Transformation :

• Data from multiple sources is normalized & converted into single system format.

Methods : Cleaning, filtering, sorting, splitting, joining

Process ③ : Loading

• finally transformed data is loaded into DWH from staging area.

• Depending upon buisness needs a data can be loaded in batches or all at once.

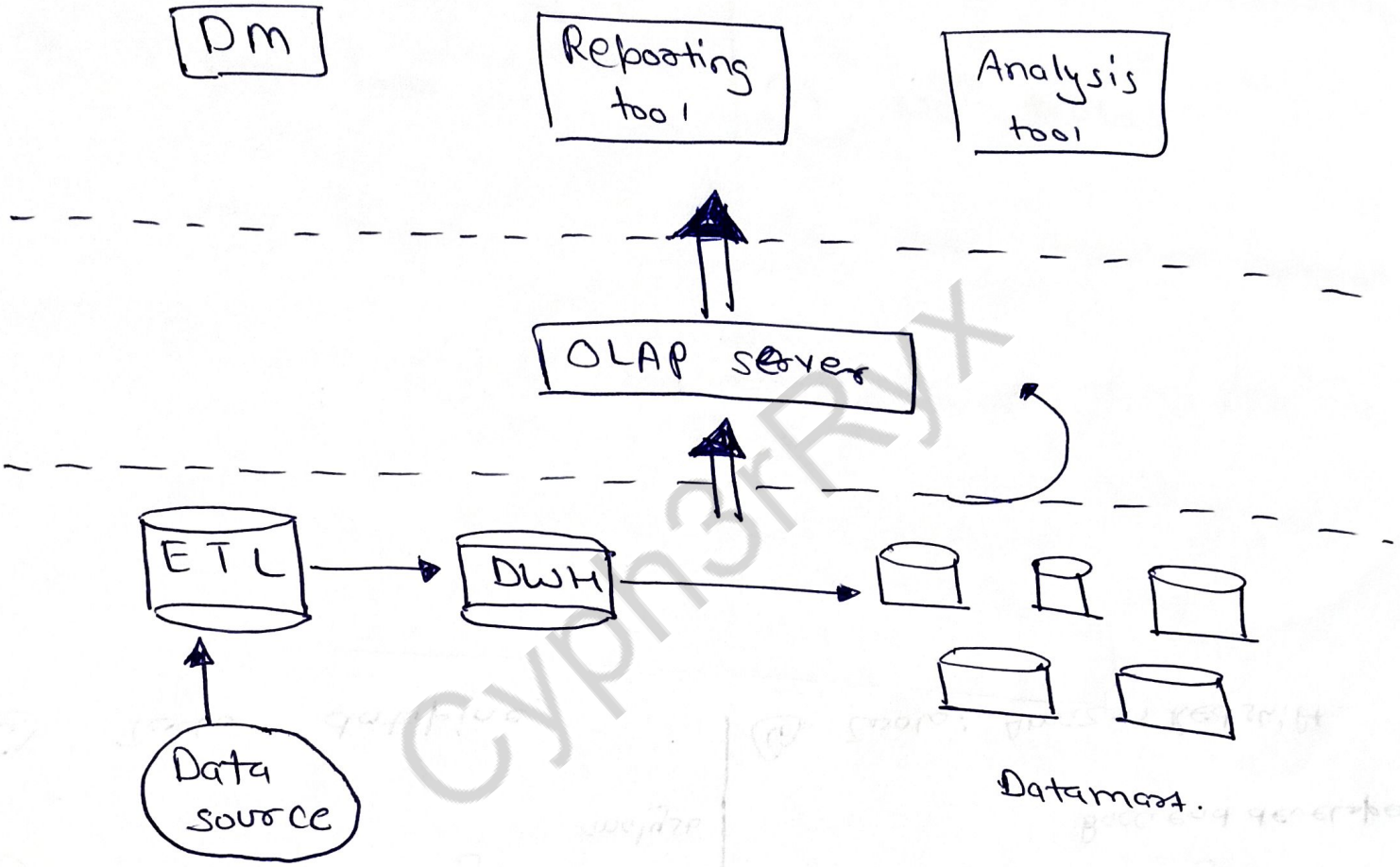
DWH	DB
① Any type of data.	① Structured data only
② Can expand capacity.	② Can't expand capacity
③ DM tools	③ No tools
④ Any format	④ Only tabular format
⑤ All kind of data	⑤ Number & text.
⑥ Designed for analyzing data.	⑥ Designed for recording the data
⑦ Uses OLAP	⑦ Uses OLTP
⑧ Tables are denormalized	⑧ Tables are normalized.
⑨ Subject oriented	⑨ Application oriented
⑩ Techniques: Data Modelling	⑩ Techniques: ER model

Annexure No :

BI	DWH
(1) Generating Buisness insights.	(1) storing data from several sources.
(2) O/P is data visualization, reporting & dashboards.	(2) O/P is unified data for upstream BI apps
(3) Audience: Manager & data analysts	(3) Audience: Data engineers & Backend developers.
(4) Tools: datapine	(4) Tools: Amazon Redshift.

DWH Architecture:

3 tier architecture



Bottom tier consists of DWH, Datamarts, ETL, source data.

Middle tier consists of OLAP server

Top tier consists of Data mining, data analyzer & reporting tool

OLAP server in middle is implemented using

① Relational OLAP model (ROLAP):

DBMS that maps fxn on multidimensional data to standard relational operations.

② Multidimensional OLAP model (MOLAP):

directly implements multidimensional information & operations.

* OLAP

Online Analytical Processing

↳ a category of software tools which provides analysis of data for buisness decisions

Primary goal : Data Analysis not processing.

eg. Amazon analyse customer's buying list to give them a personalised homepage to retain them more.

* OLTP : **Online Transaction Processing**

↳ administer day to day transaction of an organization.

Primary goal : Data Processing not analysis

eg. Sending a text message

or

Adding a book to shopping cart.

OLAP	OLTP
(1) Analysis oriented	(1) Transaction oriented
(2) Informational processing.	(2) Operational processing
(3) Accuracy is maintained of old data.	(3) Only current data is maintained
(4) Focus on information out.	(4) Focus on data in
(5) Uses DWH	(5) Uses DBMS
(6) Database Query management system	(6) Database modifying system
(7) Subject oriented	(7) Application oriented
(8) Volume is stored in TB, PB	(8) Volume stored in MB, GB
(9) Processing time = long	(9) Processing time = fast
(10) Only read operation	(10) Both read & write operation
(11) Improves efficiency of buisness analysts.	(11) Improves efficiency of user's productivity.

OLAP server operations:

① Roll up: • summarize data
• by climbing up heirarchy or by dimension reduction.
(lower – higher)

② Drill down: • reverse of roll up
• from higher level summary to lower level summary
(higher – lower)

③ Slice & dice: Project & select.

→ Slice is performed for the dimension from a given cube and provides a new subcubes.

→ Dice selects 2 or more dimensions from a given cube and provides a new subcube.

④ Pivot: • aka "rotation".
• rotates the axes in view in order to provide an alternative presentation of data.

* Types of OLAP servers :

① **ROLAP** : Relational OLAP

↳ placed betn relational back-end server & client front-end tools

↳ Store and manage DWH, uses extended relational DBMS.

↳ it includes optimization of each DBMS backend & tools.

② **MOLAP**: Multidimensional OLAP

↳ array based multidimensional storage engines for multi.dim. view of data.

↳ storage utilization may be low because of MD stores.

③ **HOLAP**: Hybrid OLAP

↳ combo of ROLAP & MOLAP

↳ offer [high scalability of ROLAP] & [faster computation of MOLAP]

④ **Specialized SQL servers** :

↳ provides advance query language & query processing support for SQL queries over star and snowflake schemas

1. **Mean** = $\frac{\text{(total summation)}}{n}$

2. **mode** = number with the highest frequency.

3. **median** : Middle score

4. **Variance** : $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x - \bar{x})^2$

S.D : $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x - \bar{x})^2}$

eg. 44, 50, 38, 96, 42, 47, 40, 39, 46, 50

x	$(x - \bar{x})$	$(x - \bar{x})^2$
44	-5.2	27.04
50	0.8	0.64
38	11.2	125.44
96	46.8	2190.24
42	-7.2	51.84
47	-2.2	4.84
40	-9.2	84.64
39	-10.2	104.04
46	-3.2	10.24
50	0.8	0.64

$\bar{x} = 49.2$

$\sum (x - \bar{x})^2 = 2600.4$

$S^2 = \frac{2600.4}{9} = 288.7$

$S = \sqrt{288.7} = \sqrt{289} = 17$

Mean = 49.2

SD = 17

$\therefore 49.2 - 17 = 32.2$

$49.2 + 17 = 66.2$

Peps will spend in this range only. between 32.2 - 66.2 $

* Data Preprocessing:

1. **Data Transformation:** Transforms the recieved data in a form that it can be mined.
2. **Data Cleaning:** clean out noisy data & fill in missing values
3. **Data Reduction:** Obtain a reduced representation of the dataset that is much smaller in volume, yet maintains the integrity of original data.
4. **Data Integration:** Combines data from multiple sources into DWH.

DATA CLEANING:

Real world data is noisy, incomplete & inconsistent.
↳ fills data, make data consistent & remove noisy data.

1. **Missing values:**
   - ↳ ignore the tuple
   - ↳ fill in missing values manually
   - ↳ fill in the most probable value
   - ↳ use another attribute which data is with you

Noisy data:

↳ Bining : Smoothes a sorted data value by consulting its neighbourhood values around it.

↳ Regression: A technique that confirms a data values to a fxn.

- ↳ linear : fit 2 attribute in best line so that one attribute can predict other.
- ↳ multiple: fits more than 2 attributes & data is fit in multidimensional surface.

↳ Outlier analysis:

Outlier detected by clustering.

Similar values are organized in groups.

* Data Transformation:

data is transformed in a form such that it can be mined

(i) **Smoothing:**

↳ removes noise from the data
(bining, regression)

(ii) **Aggregation:**

↳ summary or aggregate values are applied
(constructing a data cube)

(iii) **Generalization:**

↳ low level concepts are replaced with high level concepts.
eg. street → city / country

(iv) **Normalization:**

↳ Attribute values are normalized by scaling their values so that they fall in specified range

eg. 2
40
500
1
3
900

All the values are different and nowhere near to each other so if we normalize the data in such a way that its range becomes (0-1).

Method: 1 Min.-max. Normalization

$$v' = \frac{v - \min(x)}{\max(x) - \min(x)}$$

v = original attribute value

min(x) = min. value of all the elements = 1

max(x) = max. value of all the elemnt. = 900

for eg. we need v' of 2:

$$v' = \frac{2-1}{900-1} = \frac{1}{899} = 0.001112$$

In this way we can normalize the data for all elements.

Method - (II) Z- score normalization :

(Zero mean normalization)

$$Z = \frac{V - \bar{x}}{SD\ x}$$

$$SD = \sqrt{\text{variance}}$$

$$= \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
2	-239	57,121
40	-201	40,401
500	259	67,081
1	-240	57,600
3	-238	56,644
900	659	434,281

$$SD = \sqrt{\frac{713128}{6}}$$

$$= \sqrt{118,854.67}$$

$$SD = 344.75$$

$\bar{x} = 241$

$\Sigma (x - \bar{x}) = 713,128$

$$Z_1 = \frac{2 - 241}{344.75} = \frac{-239}{344.75} = -0.693$$

$$Z_2 = \frac{40 - 241}{344.75} = \frac{-201}{344.75} = -0.583$$

$$Z_3 = \frac{500 - 241}{344.75} = \frac{259}{344.75} = 0.7512$$

$$Z_4 = \frac{1 - 241}{344.75} = \frac{-240}{344.75} = -0.696$$

$$Z_5 = \frac{-238}{344.75} = -0.69 \qquad Z_6 = \frac{659}{344.75} = 1.911$$

* Data reduction:

↳ to get a reduced representation of data set that is much smaller in volume.

(1) Data cube aggregation:

Aggregation opⁿ are applied to the data in the construction of a data cube.

(2) Attribute subset selection:

irrelevant attributes are removed & relevant are selected.

(3) Discretization:

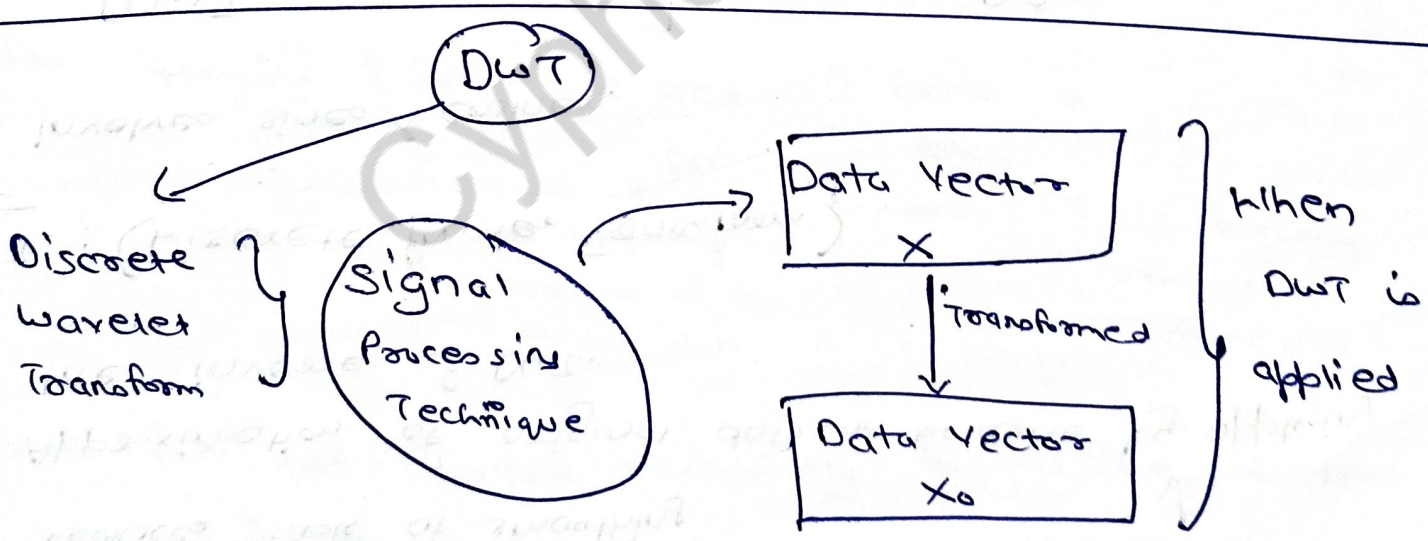
raw data values are replaced by ranges or higher conceptual levels.

DMBI

- Data cleaning
- Data Transformation
- Data Reduction
- Data Integration

Data Cleaning

- fill missing data
- smooth & noisy data
- resolve data inconsistencies



Both X & Xo must have same length

Advantages:

- resultant data is very sparse.
- computation of data is very fast.
- able to remove noise from the data.
- reduces taste of smoothing
- Approximation of original data can be done by applying the inverse of DWT.

DFT: (Discrete fourier Transform)

- involves sines & cosines.

DWT	DFT
• DWT is independent	• DFT is based on the results of DWT
• DWT gains lossy compression.	• DFT dont get lossy compression
• Provides accurate wavelet coefficients.	• dont provide accurate wavelet coefficients.
• Occupies less space	• Occupies more space

Annexure No :

Heirarchical Pyramid Algorithm:

↳ halves data at each iteration.

↳ the data is reduced to half so speed of computation increases.

method:

① input vector : L (L = integer & power of 2)
↳ necessary

② apply 2 fxns for each transform
↳ first performs data smoothing
↳ second finds weighted difference

③ After applying 2 fxns we get 2 data set
- first → low frequency version of original data
- second → high frequency version of original data

④ Both fxns are applied recursively to get the resultant vector of length 2

⑤ Then wavelet coefficients are assigned to the transformation data vectors finally.

DATA MINING Applications:

1. Spatial mining
2. Multimedia mining
3. temporal mining
4. Web mining
5. Text mining.

Spatial mining

- for spatial model DM
- uses geographical analysis
- analysts uses this analysis for creating B.I
- requires specific techniques and resources to get data into relevant & useful format.
- task: search for spatial patterns.

Multimedia mining:

- data is in audio, video, images, graphics
- mining of image, text & wwws.
- methods are: (i) similarity search
  (ii) multidimensional analysis
  (iii) classification & prediction
  (iv) association mining

Temporal mining:

- Time ordered data
  - eg: sales, weblogs, calls, etc.
- Require repeated measurements.
- mining methods require modifications to handle temporal relationship
- Time ordered data contribute to prediction, given the history of events.
- Time ordered data often link certain events to specific pattern

Web mining:

- Automatic discovery & extract information from web documents.
- info. can be extracted from server logs & web activity
- Web mining : web activity
- Categories are:
  - web content mining
  - web structure mining
  - web usage mining.

Text mining:

- retrieve high quality info. from text.
- available data is in unstructured format.
- useful when it is necessary to extract information from large datasets.
- text mining :
  - info. retrieval
  - natural language processing

Clustering:

Unsupervised learning:
- considered as (self learning process)
- discovering patterns from data w/o any labels.

eg. student with all study material but no faculty to guide.

Cluster Analysis: form groups with same similarity.

eg. grouping of students studying similar subjects.

- **Partitioning methods:** division of n items into set of k cluster such that (sum of squared distance is minimized.)

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - c_i)^2 \; ; \; c_i = \text{centroid of cluster.}$$

2 methods: (i) k mean
(ii) k medoids.

Annexure No :

K-mean: ($k=2$ i.e 2 clusters)

S-1 Take mean value

S-2 Find nearest no. of mean and put it in cluster.

S-3 Repeat one and two until we get same mean.

eg. $k = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

$k = 2$

$mean_1 = 4$ $mean_2 = 12$ [Random mean no.]

Now take element (1) - m_1 & (1) - m_2 whichever is closer to m_1 or m_2 put it in.

eg. $2 - 4 = (-2)$ $2 - 12 = (-10)$

-2 is closer than 4 so put it in k_1

(i) $k_1 = \{2, 3, 4\}$; $k_2 = \{10, 11, 12, 20, 25, 30\}$

New $m_1 = (2+3+4)/3 = (3)$ New $m_2 = 108/6 = (18)$

(ii) $k_1 = \{2, 3, 4, 10\}$; $k_2 = \{11, 12, 20, 25, 30\}$

Do it again mean.

Enrollment No :

Page No :

$m_1 = 4.75$ $\quad$ $m_2 = 19.6$

$\approx (5)$ $\quad$ $\approx (20)$

(iii) $u_1 = \{2, 3, 4, 10, 11, 12\}$ $\quad$ $u_2 = \{20, 25, 30\}$

$m_1 = 7$; $\quad$ $m_2 = 25$

(iv) $u_1 = \{2, 3, 4, 10, 11, 12\}$ $\quad$ $u_2 = \{20, 25, 30\}$

$m_1 = 7$ $\quad$ $m_2 = 25$

We got same mean both time so we need to stop

$\therefore$ <u>Final answer</u> : $u_1 = \{2, 3, 4, 10, 11, 12\}$

$u_2 = \{20, 25, 30\}$

Agglomerative methods:

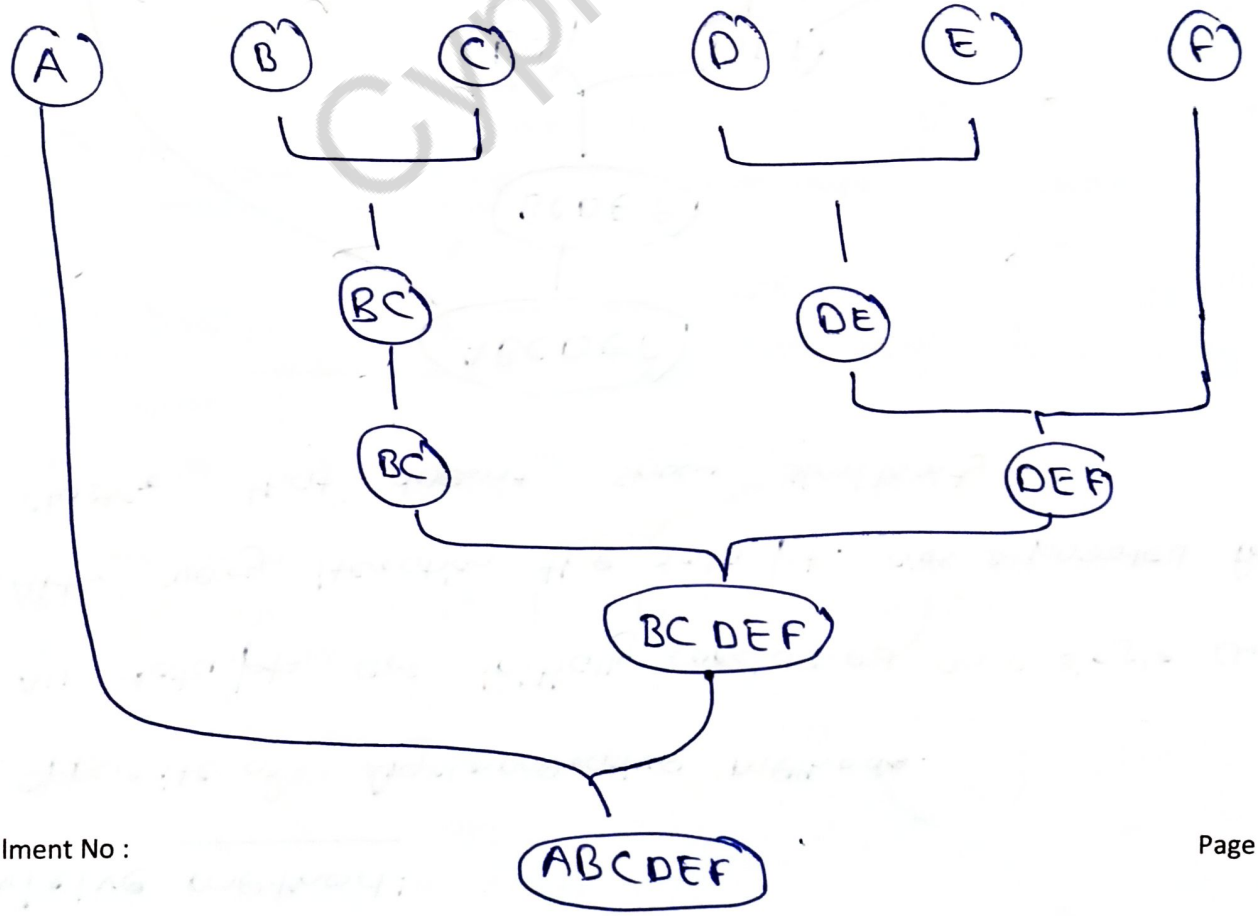
S-1 Calculating similarity among clusters.

S-2 Every datapoint is taken as an individual cluster.

S-3 Merging clusters with higher proximity among each other.

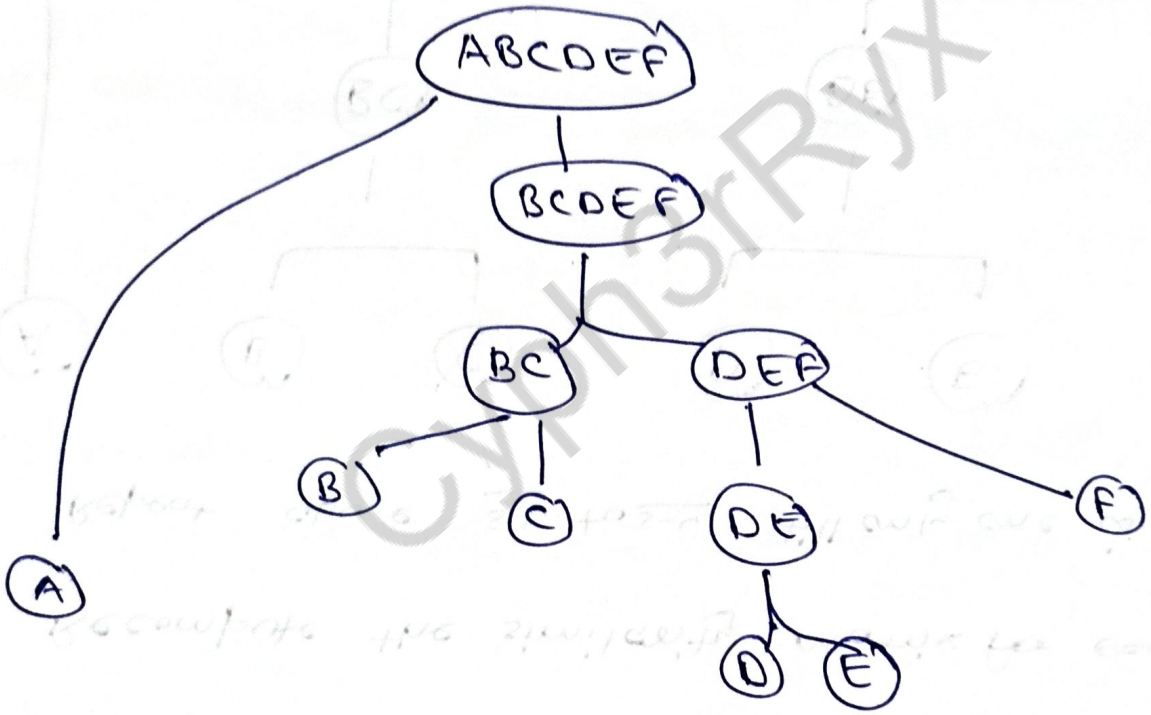
S-4 Recompute the similarity matrix for each cluster.

S-5 Repeat above S-1 to S-4 till only one is left.



Divisive method:

- Opposite of Agglomerative methods.
- All data pts are initially considered as a single cluster.
- After every iteration the data pt. are separated from the cluster that doesn't show similarity



Annexure No :

Density Based Scan (DBscan) : Spatial Clustering of Application with Noise.

↳ Data objects are clustered based on density (M/V)

↳ 2 inputs => ε & min. pts.()

(ε)-radius of circle formed with data objects as center.

(min. pts()) is the minimum no. of data points inside the circle.

3 Types of data points :

① Core point : it should satisfy the condn of min. pts()

② Boundary point : Neighbour of core

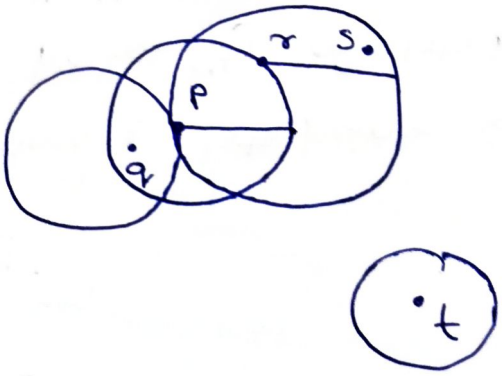
③ Noise point : not core nor boundary

eg. min. pts = 3

p, r → core

q, s → boundary

t → noise.



Evaluation of Clustering:

(1) Clustering Tendency:

- non uniformity among data points is vital
- measuring the probability of data points generated by uniform data distribution.

Hypothesis (H)

→ null hypothesis : Non uniform data distribution

→ Alternate hypothesis : Random data generation

H > 0.5 → data contains cluster
∴ rejects null hypothesis.

H closer to 0 → no clustering tendency.

(2) Number of Cluster :

Depends on → Distribution shape
→ scale in data set
→ Clustering resolution

2 approach

- Domain knowledge
  - initial knowledge on forming no. of clusters.
- Data driven approach
  - Empirical
  - elbow

(3) Quality of Clustering :

Characteristic of Cluster : minimum intra cluster distance

maximum inter cluster distance

intra	inter.
btw 2 data pts.	btw 2 data cluster.

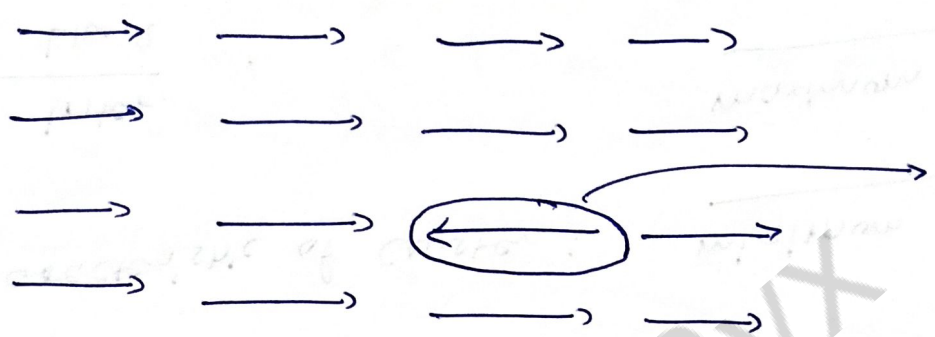
Extrinsic measures : True labels required

Intrinsic measures : not required

Outlier Analysis:

Outlier is the data object which doesn't obey the general behavior.

eg.



Outlier detection: Process of identifying the outliers and then removing it.

2 Approach

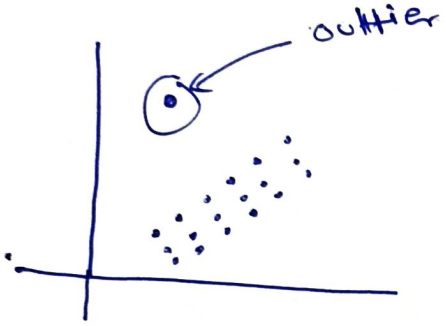
Statistical
↳ Probability based
- Parametric
- Non Parametric

Proximity
↳ Location based
- Density based
- Distance based
- Grid based
- Deviation based

Types of outliers:

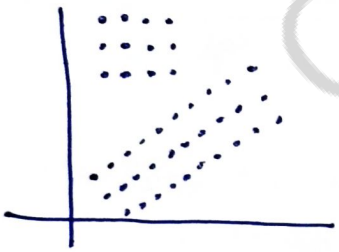
(1) Global / Point outliers:

When a single data object deviates from rest of the data set



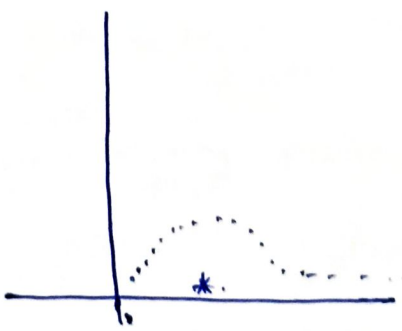
(2) Collective outlier:

When a group of data points deviates from the rest of the data set.



(3) Conditional outlier:

deviation due to specific conditions



Interestingness of pattern:

from millions of DM patterns which of them are really interesting?

3 Questions.:

1. what makes the pattern interesting?
2. Can DM system generates all of the interesting patterns?
3. Can DM system generates only interesting pattern

Ans.(1) A pattern is interesting if it is

- easily understood by humans
- valid on new / test data
- potentially useful.

Ans.(2) Refers to completeness of DM system
As it is not possible to generate all interesting patterns

Ans.(3) Refers to optimization of DM system.

Difficult to generate only interesting patterns.

Classification of DM system:

(1) Based on mined DB:

(i) Relational (ii) Transactional (iii) Object relational

(iv) Data warehouse

(2) Based on knowledge mined:

(i) Characterisation

(ii) Discrimination

(iii) Association & correlation analysis

(iv) Classification

(v) Prediction

(vi) Outlier Analysis

(vii) Evolution Analysis

(3) Based on techniques used:

ML, Statistics, Nueral Networks, etc.

(4) Based on application adapted:

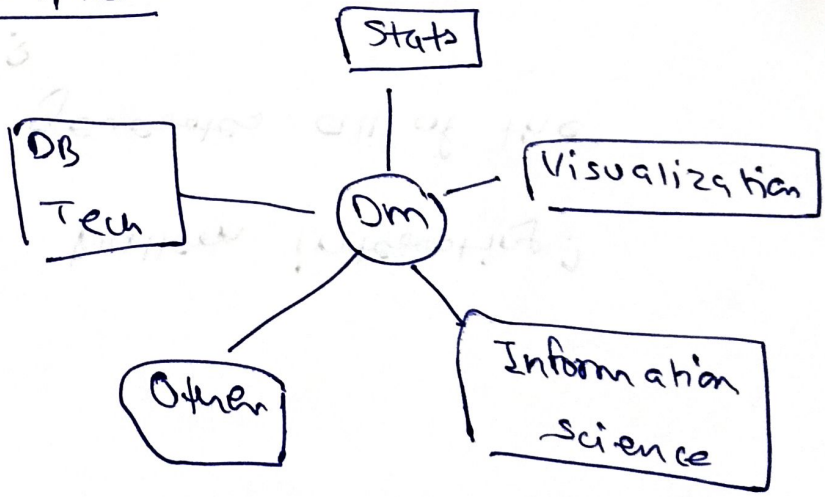
(i) finance

(ii) DNA

(iii) Email

(iv) Stock market

(v) Telecommunication.



Frequent Pattern :

The pattern that appears frequently in dataset

eg. Milk & Bread

Market Basket Analysis :

Process of analysing customer buying habits, by finding the associations between the different items that a customer will place in their basket.

Strategies : (1) Placing them together

(2) Placing them at 2 different ends

- This analysis will help sellers to plan their shelf space for increased sales.

C, Frequent patterns are represented by association rules.

ex. Computer & anti virus.

Computer ⇒ Antivirus software

support ⇒ 2%

Confidence ⇒ 60%.

Support: identifies how frequently a rule is applied to given dataset.

$$S(P \rightarrow Q) = \frac{\sigma(P \cup Q)}{N}$$

N = Total Transactions.

Confidence: defines frequent occurence of items of Q in transactions.

$$C(P \rightarrow Q) = P(B/A) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

B.Tech: ___ Year ___ Semester: ___

Annexure No :

APRIORI Algorithm: By R. Agrawal & R-Srikant (1997)

↳ shows how objects are associated with each other.

Objective: To generate an association.

Eg.

Min. support = 50 %.

Threshold confidence = ~~70~~ 50 %.

ID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Now check how much time each item is occuring

i.e support.

1 occurs (2) time

2 occurs. (3) time

3 4 (3) 4

4 4 (1) 4

5 " (3) "

Item	Support	min. support
1	2	2/4 = 50%.
2	3	3/4 = 75%.
3	3	3/4 = 75%.
4	1	1/4 = 25%. (X)
5	3	3/4 = 75%.

=> Support / ID (total)

Cancel out the item which has support less than suggested

i.e item 4 => 25%.

Remaining itemset = (1,2,3,5)

from pairs. => (1,2), (1,3), (1,5), (2,3), (2,5), (3,5)

Items	support	min. support	
(1,2)	1	1/4 = 25%.	(X)
(1,3)	2	2/4 = 50%.	
(1,5)	1	1/4 = 25%.	(X)
(2,3)	2	2/4 = 50%.	
(2,5)	3	3/4 = 75%.	
(3,5)	2	2/4 = 50%.	

Check how many times (1,2) occurs together and same goes for all items.

Remaining itemset = (1,3), (2,5), (2,3), (3,5)

Now from triplets

i.e (1,3,5) (1,2,5) (1,2,3) (2,3,5)

Check the similar items betⁿ dataset and only form triplets with those.

item set	support	min. support
(1,2,3)	1	1/4 = 25 %. (X)
(1,2,5)	1	1/4 = 25 %. (X)
(1,3,5)	1	1/4 = 25 %. (X)
(2,3,5)	2	2/4 = 50 %.

Only one triplet remains (2,3,5)

Now let calculate support & confidence.

$$\text{Confidence} = \frac{\text{support } (A \cup B)}{\text{support } (A)}$$

Using (2,3,5) we need to find/genesete association rules.

Rules	support	Confidence
$(2 \wedge 3) \rightarrow 5$	2	2/2 => 100 %.
$(3 \wedge 5) \rightarrow 2$	2	2/2 => 100 %.
$(2 \wedge 5) \rightarrow 3$	2	2/3 => 66 %. (X)
$2 \rightarrow (3 \wedge 5)$	2	2/3 => 66 %. (X)
$5 \rightarrow (2 \wedge 3)$	2	2/3 => 66 %. (X)
$3 \rightarrow (2 \wedge 5)$	2	2/3 => 66 %. (X)

$(2 \wedge 3) \rightarrow 5$ $\Rightarrow$ confidence $= \frac{S((2 \wedge 3) \cup 5)}{S(2 \wedge 3)} = \frac{2}{2}$

(A) (B)

$S(A \cup B)$

(by previous tables.)

$S(2,3)$

$= 100\%.$

Do it for all rules.

Remove all the rules which has confidence less than threshold confidence.

i.e @ 70%.

FP Growth Algorithm:

Frequency Pattern Growth is an efficient & scalable method for mining the complete set of FP using a tree structure for storing information about FP called FP tree.

Eg. Min. support = 30 %.

ID	Items.
1	E, A, D, B
2	D, A, E, C, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

⟹

ID	Frequency	Priority
A	5	3
B	6	1
C	3	5
D	6	2
E	4	4

How to write priority?

→ more frequency = more priority

→ same frequency = FCFS

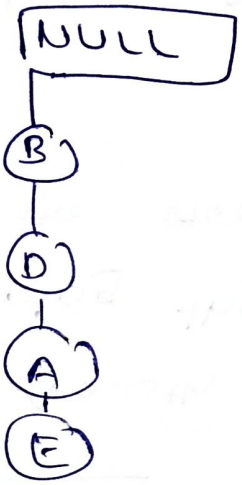
Now rewrite ID based on priority

(B, D, A, E, C)

order items according to their priority (B D A E C)

ID	Item	ordered items.
1	E A D B	B D A E
2	D A E C B	B D A E C
3	C A B E	B A E C
4	B A D	B D A
5	D	D
6	D B	B D
7	A D E	D A E
8	B C	B C

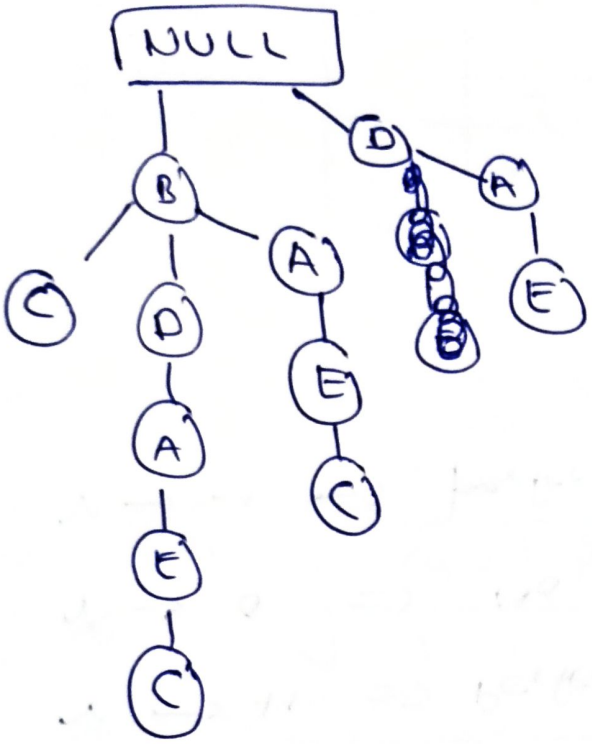
Now lets make a tree with root node NULL



for ordered item - (1)

B D A E

∴ B - 1
D - 1
A - 1
E - 1



B — 1, 2, 3, 4, 5
D — 1, 2, 3, 4
A — 1, 2, 3
E — 1, 2
C — 1
A — 1
E — 1
C — 1
D — 1, 2
A — 1
E — 1
C — 1

Final Answer with FP Tree.

⊂

↳ only 2 passes over data set
↳ compresses dataset
↳ no candidate generation
↳ FP tree may not fit in memory
↳ FP tree is expensive

} Characteristic of FP Growth.

* Correlation Analysis:

Used to measure the relationship b/w 2 variables

$$r_{A,B} = \frac{\Sigma (A - \bar{A})(B - \bar{B})}{(n-1)\,\sigma A\,\sigma B}$$

$r_{A,B}$ = Karle Pearson Correlation coefficient

$\bar{A}, \bar{B}$ = mean of A & B

$\sigma A, \sigma B$ = standard deviation of A & B

n = no. of tuple in DB.

$r \rightarrow$ 3 values $(0, -1, +1)$

$r \rightarrow +1 \Rightarrow$ Perfect positive correlation

$r \rightarrow 0 \Rightarrow$ No correlation

$r \rightarrow -1 \Rightarrow$ perfect negative correlation

Eg.

A	B
20	8
12	34
9	4

$$\bar{A} = \frac{20+12+9}{3}$$

$$= \boxed{13.66}$$

$$\bar{B} = \frac{8+34+4}{3}$$

$$= \boxed{15.33}$$

$$\sigma A = \sqrt{\frac{\Sigma (A-\bar{A})^2}{n-1}}$$

$$\therefore \sigma A = \sqrt{\frac{(20-13.66)^2 + (12-13.66)^2 + (9-13.66)^2}{3-1}}$$

$$\therefore \boxed{\sigma A = 5.68}$$

$$\sigma B = \sqrt{\frac{\Sigma (B-\bar{B})^2}{n-1}} = \sqrt{\frac{(8-15.33)^2 + (34-15.33)^2 + (4-15.33)^2}{3-1}}$$

$$= \boxed{16.28}$$

Now in main formula

$$r_{AB} = \frac{\Sigma (A-\bar{A})(B-\bar{B})}{(n-1)\sigma A \sigma B}$$

$$= \frac{\left[(20-13.66)(8-15.33) + (12-13.66)(34-15.33) + (9-13.66)(4-15.33)\right]}{(3-1)(5.68)(16.28)}$$

negative correlation

$$= \boxed{-1}$$

Classification & Prediction:

finding a good model to predict the class of object whose class label are unknown

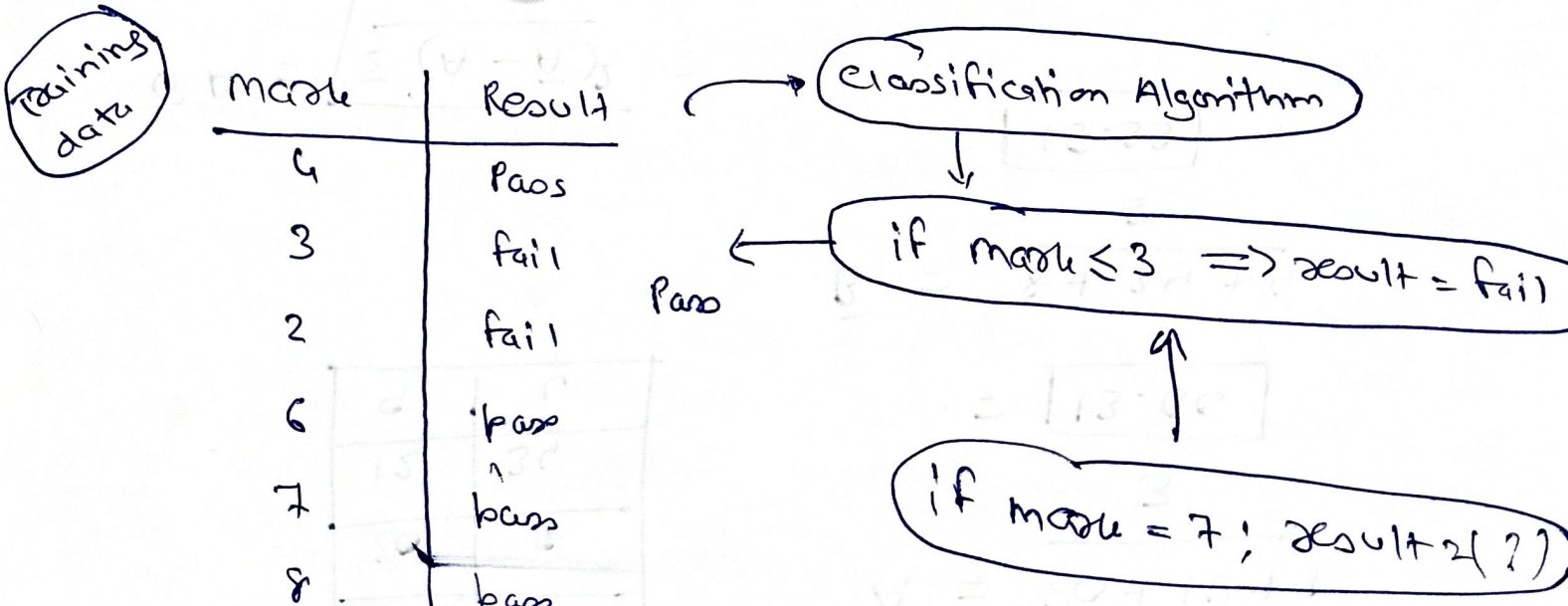
eg:
- grouping of patients based on the medical records.
- categorization of new data with the help of past data.

Predicting a missing / unknown value based on past / current data

eg: predicting the correct treatment for a person based on their medical condition.

Classification has 2 stages:

1. model construction
2. model usage.



Learn, Analyze. Generate classification

* Decision Tree : (ID3 Algorithm)

(1) In the given dataset choose a target attribute.

(2) Calculate I.G (info. gain) of target attribute

$$I.G = \frac{-P}{P+N} \log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N} \log_2 \frac{N}{P+N}$$

(3) for remaining attributes, find entropy

Entropy = I.G x Probability.

$$E(A) = \sum \frac{P_i + N_i}{P+N} \; I(P_i N_i)$$

(4) Calculate Gain = I.G - E(A)

eg.

Age	Completion	Type	Profit.
old	Yes	S/w	Down
old	No	S/w	Down
old	No	h/w	Down
mid	Yes	S/w	Down
mid	Yes	h/w	Down
mid	No	h/w	up
mid	No	S/w	up
new	Yes	S/w	up
new	No	h/w	up
new	No	S/w	up

S-1 Target Attribute => **Profit**

Annexure No :

S-2

$$I.G = \frac{-P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right)$$

P => Count (down) => 5

N => Count (up) => 5

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

$$= \left(\tfrac{1}{2}\log_2(2^{-1}) + \tfrac{1}{2}\log_2(2^{-1})\right)$$

$$= \left[\tfrac{1}{2}\times(-1)\log_2(2) + \tfrac{1}{2}\times(-1)\log_2(2)\right]$$

$$= \left(\tfrac{1}{2}\times -1 + \tfrac{1}{2}\times -1\right) = -\tfrac{1}{2} + \left(-\tfrac{1}{2}\right) = -\tfrac{2}{2}$$

$$= -1$$

$$\boxed{I.G = -1}$$

S-3 Entropy Calculation for remaining attributes:

$$E(A) = \sum \frac{P_i + N_i}{P+N} \; I(P_i N_i) \quad \left[\text{IG} \times \text{Probability}\right]$$

(Age) : Prepare table for each attribute

Rows: values of attribute

Columns: values of target attribute.

	down	up
old	3	0
mid	2	2
new	0	3

Entropy = IG x Probability

$$I_G(\text{old}) = -\left[\tfrac{3}{3}\log\left(\tfrac{3}{3}\right) + \tfrac{0}{3}\log\left(\tfrac{0}{3}\right)\right] = 0$$

Prob. = 3/10

$$E(\text{old}) = 0 \times 3/10 = \boxed{0}$$

$$I.G(mid) = -\left[\frac{2}{4}\log\left(\frac{2}{4}\right) + \frac{2}{4}\log\left(\frac{2}{4}\right)\right] = 1$$

$$E(mid) = I \times \frac{4}{10} = 0.4$$

$$I.G(new) = -\left[\frac{0}{3}\log\left(\frac{0}{3}\right) + \frac{3}{3}\log\left(\frac{3}{3}\right)\right] = 0$$

$$E(new) = 0 \times \frac{3}{10} = 0$$

Entropy (Age) = $0 + 0.4 + 0 = 0.4$

Step 4) Gain = I.G − E(A) = $1 - 0.4 = 0.6$

Same goes for Completion and Type

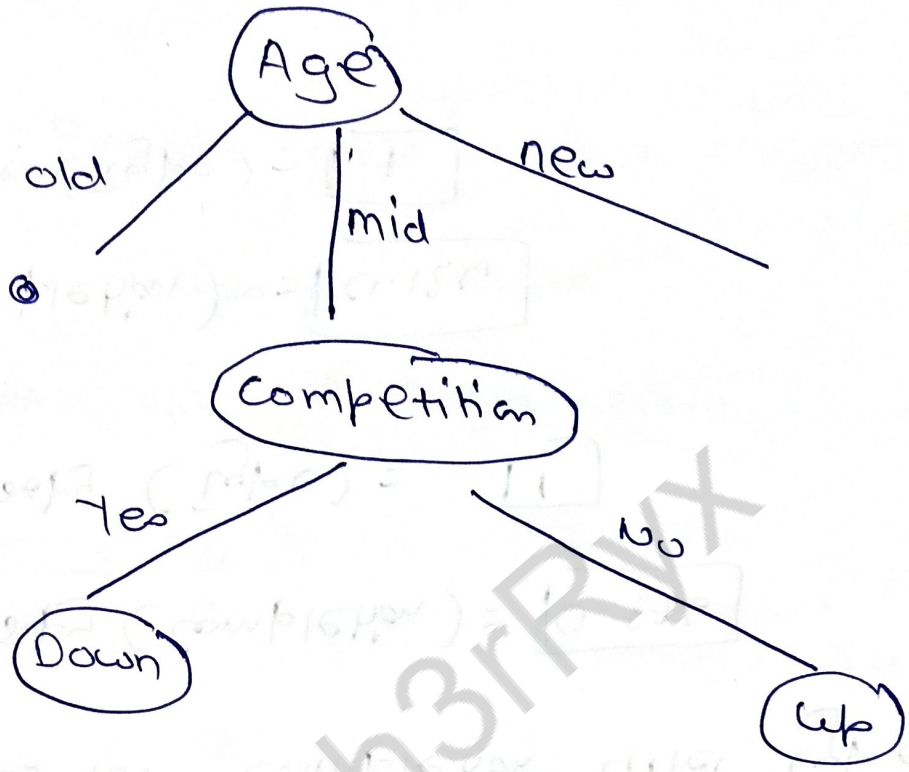
∴ Entropy (Completion) = 0.876

∴ Entropy (Type) = 1

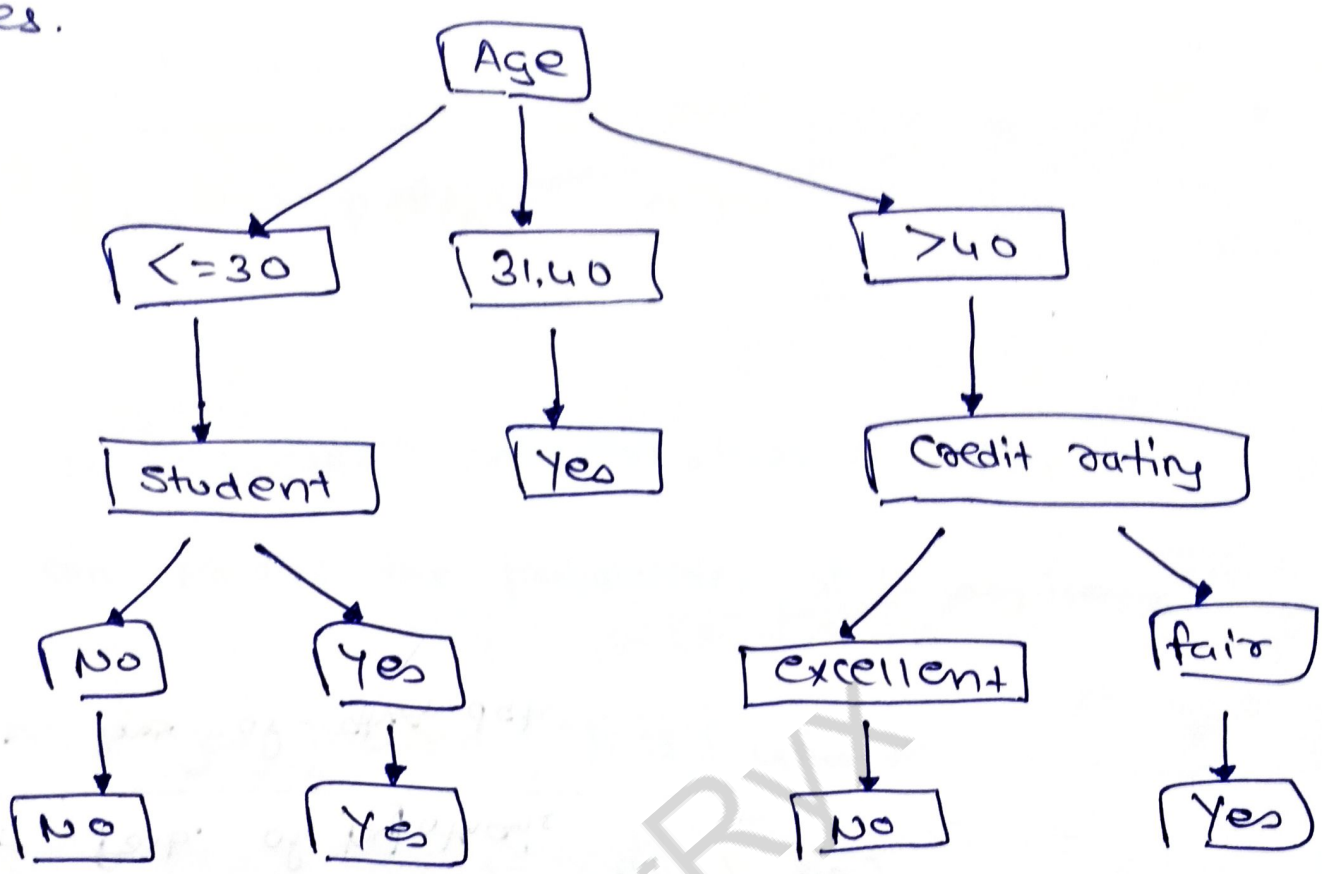
Gain(Completion) = 0.124

Gain (Type) = 1

Highest Gain → root node



eg.



Rules:

- IF age = "<=30" AND Student = "no" THEN buys computer = "no"
- If age = "<=30" AND student = "yes" THEN buys computer = "yes"
- If Age = "31..40" THEN buys computer = "yes"
- If age = ">40" AND Credit rating = "excellent" THEN buys computer = "no"
- If age = ">40" AND Credit rating = "fair" THEN buys computer = "yes".

IF "Cond^n" AND "Cond^n 2" THEN "Conclusion"

Baye's Theorem:

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

A = hypothesis

B = given data

$P(A/B)$ = finding prob. of hypothesis when prob. of training examples are given.

$P(B/A)$ = finding prob. of given data with prob. of hypothesis that is true

$P(A)$ = Prob. of hypothesis

$P(B)$ = Prob of given data

Bayesian Classification:

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

- Bayesian classifiers are statistical classifiers
- They can predict the probabilities of class items.

Naive Bayes Theorem / Classification:

same formula => $P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$

eg

fruit = { Yellow, Sweet, Long }

Fruit	Yellow	Sweet	Long	Total
Orange	350	450	0	650
Banana	400	300	350	400
others	50	100	50	150
Total	800	850	400	1200

We need to select a fruit that is yellow, sweet & long from given data.

$$\therefore P(\text{Yellow}/\text{orange}) = \frac{P(\text{orange}/\text{yellow}) \cdot P(\text{Yellow})}{P(\text{orange})}$$

$$= \frac{(350/500) \cdot (500/1200)}{(650/1200)} = \boxed{0.53}$$

$$\therefore P(\text{Sweet}/\text{orange}) = \frac{P(O/S) \cdot P(S)}{P(O)} = \boxed{0.69}$$

$$\therefore P(\text{Long}/\text{orange}) = \boxed{0}$$

Now $P(\text{orange}/\text{fruit}) = (0.53) \times (0.69) \times (0) = \boxed{0}$

$P(\text{banana}/\text{fruit}) = (1) \times (0.75) \times (0.87) = \boxed{0.65}$

$P(\text{others}/\text{fruit}) = (0.33) \times (0.66) \times (0.33) = \boxed{0.072}$

We will choose the fruit which has highest probability

i.e Banana

Annexure No :

KNN Algorithm: lazy learning

k - nearest neighbour algorithm

Eg. given data query => x => (maths = 6 ; CS = 8)

& k = 3

Classification = Pass / fail

Maths	CS	Result
4	3	F
6	7	P
7	8	P
5	5	F
8	8	P

We need to find that given query x has 2 condn for maths & CS i.e GRT zero. So if the student is pass or fail based on that so let's see.

$$\text{Euclidean Distance } (d) = \sqrt{\left|(x_{O_1} - x_{A_1})^2\right| + \left|(x_{O_2} - x_{A_2})^2\right|}$$

x_O = observed value

x_A = actual value.

$d_1 = \sqrt{(6-4)^2 + (8-3)^2} = \sqrt{29} = 5.38$

$d_2 = \sqrt{(6-6)^2 + (8-7)^2} = \sqrt{1} = 1$

$d_3 = \sqrt{(6-7)^2 + (8-8)^2} = \sqrt{10} = 1$

$d_4 = \sqrt{(6-5)^2 + (8-5)^2} = \sqrt{10} = 3.16$

$d_5 = \sqrt{(6-8)^2 + (8-8)^2} = \sqrt{4} = 2$

Now as $k=3$; we need to pick 3 neighbours.

Choose 3 values and all of them should be as near as they can and as minimum as they can

i.e (1), (1), (2)

d_2 d_3 d_5

⇓ ⇓ ⇓

Pass Pass Pass

3P / 0F

⇓

Given query is Pass

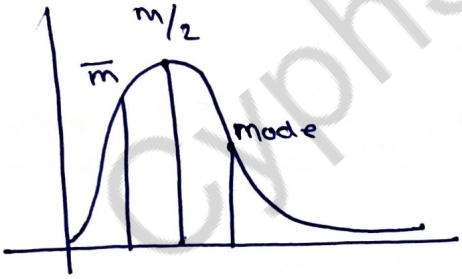
Data Skewness:

when most of the values are skewed to the left or right side from the median then it is called skewed

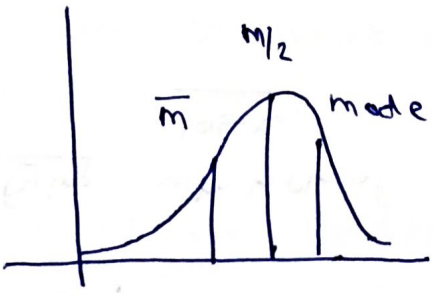
3 types:

(1) Symmetric : mean, median & mode at same pt.

(2) Positively skewed : values are to the left from median.



(3) Negatively skewed : values are to the right from median



Classification	Clustering
(1) Supervised	(1) Unsupervised
(2) Classifying with help of labels.	(2) grouping with help of similarity.
(3) More complex.	(3) less complex.
(4) Naive Bayes, etc.	(4) k means, etc.

Credits:

Cyph3rRyx

For More Handwritten Study Notes, Click Here!