5/8/24 UNIT - I

Date: / /

Objective : To analyse the student's Performance

Target audience : (age) <5, 11>

Step-1 : Data Collection (Raw fact)

Surveys, Questionaires / forms → Primary data
↳ Personal

GPA, Attendance, sports, Course Interest - Secondary data

Wearable devices - (Smart devices / watches)

1) Health - Step count

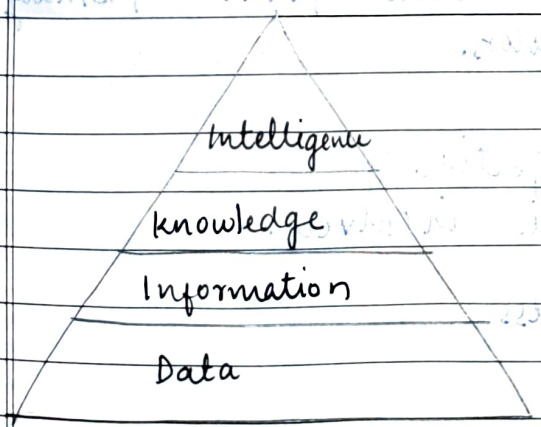
Physical / Mental - sleeping patterns

Latent needs / latent data

step-2 : Information - Processessed data.
(statistics)
Graphs, Tables, charts.

Info

6/8/24 step-3 : Knowledge



pyramid of Data

Summary of statistics

Descriptive — outcome is known

Predictive Analytics — Compute / Analytics

Correlation analysis — study hours & marks among two/more entities with measuring parameters

Step-4 : Intelligence - decision making
Agents / Assitant

Chatbots, GPTS, Smart devices, Robots.

Maths - Personalised / tutoring sessions.

Continous evaluation and monitoring.

Main drawback - training AI. (issues of Subjectivity / Explainability based on context).

subjective - Cannot be defined in a single way on a scale. [^nC_r permutations.]

objective - True / False stmts, Fixed, Binary format Rainbow colors, Rule based approach.

Intelligence is subjective if
{: Human perception is involved.
(thought process }.

IQ is objective measure.

biasid problm

System (Agent) | Human Interaction } Cognitive Science

Study of mind + perception + AI

★ Challenges:

1. Data privacy & Ethical concerns are challenges while collecting data

2. Achieving High Responce rate is another challuge (Biasness)

3. Issues w information
   → Null values / missing data
   → duplicates (names without roll nos) / data inconsistency

S.No	Name	%	Company
1	K. Vinay		Google
2	K. Vinay	80	Google

missing data

if no S.No Name becomes duplicate

→ data Quality / Unwanted data.
[Address in Report card]

4. Challenges of knowledge:
   → Summarization (content usage)
   → Outdated
   [Trend – seasonal changes – pre, during & post
   Pattern – Buying eraser after buying pencil]

5. Challenges in Intelligence

→ Biasness
→ Subjective.

classification :

1. Binary classification
2. Multi classification.
   Review of a movie Feedback.
   o/p: High / Moderate / low Ratings.

Prediction: Based on classified o/p we predict.
Predicting Placement score from CGPA

7/8/24 Data mining : Extracting (mining) knowledge from large amout of data

synonyms : Data mining, Knowledge mining, Data archeology, Knowledge Representation, Data pattern analysis, Data dredging, Knowledge extraction.
KDD - knowledge discovery in databases

Todo Why is it called Data Mining ? & not knowledge mining

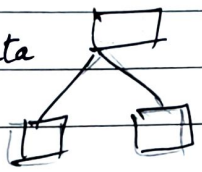
→ Movie Recomendation System
data → Genres, Watchtime, Language, Period
Actors preferences, Age group, favourite movies
Education / Agriculture

* Evolution of Database Technology.

2. Objective : Evolution of Health monitoring system.

small clinics
1) Patient attributes (Pid, Pname)
Issue with Storing large volumes of data.

2) Tree format
Storage issue if data is too vast.



① csv files / Excel flat files.

② Hierarchical / Network database

3) Individual tables for [patients] [doctors] [appointments]

③ - RDBMS.

4) Online appointments

④ Web database
Web mining.

5) Images (unstructured)

⑤ NoSQL / MongoDB. (storage)
↳ MySQL (after structuring)

6) Multispeciality hospitals (large volumes of data)

⑥ Amazon, Google for processing big data

7) Dashboards → visulization - reports/profile
(Tableu) - Assistance

⑦ NLP
Robotic

8) Advanced - Data mining, ML, DL, AI (based on data type)

⑧ (advanced).

types of Data

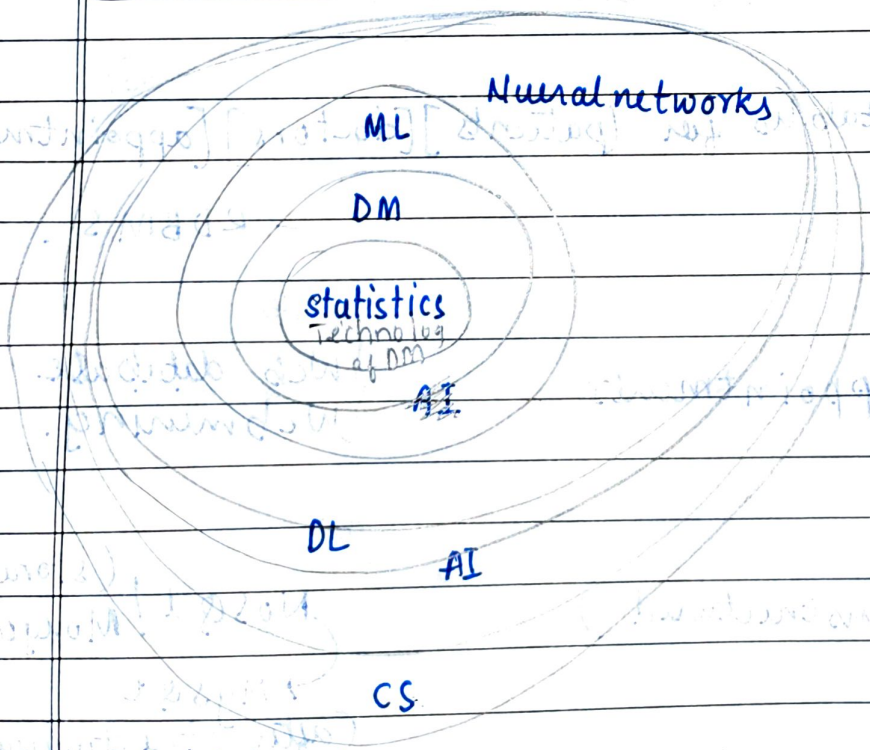
Image, text, audio, video, sign
Numerical

Data
↓
RDBMS
↓
Data processing
↓

Data cleaning
overfitting & underfitting

[nural network]
↓
Back propagations.
↓
system crashed
takes days to solve

ML	DL	AI	
classification prediction clustering Recomendation	(Extension of ML)	Automation	Cognitive science



ELA - CBP, Paper
CA1 - Wit Report
CA2 - Descriptive report

12/8/24 SM Knowledge Discovery Data processes.

1. Data cleaning : Its a procedure to remove the outliers replacing the null values, handling the issues of (to remove outliers, null values, missing data, noisy data) inconsistency, identifying the outliers basing on context of data

Data Preprocessing

2. Data integration : In this step various formats & sources of data are combined/integrated
clinical data - investigations are done (try to perform action)
~~text + numeric~~ + (many)
behavioural data -

These both are merged together.
[text + numeric, image + text, audio + signals, video + signals, ...]

3. Data selection :
collecting data that is required/relevent to analysis is retrieved from database.

4. Data transformation : [image → (conversion) numeric]
to convert one form of data into another.
→ ~~to~~ [summary of statistical ~~data~~ is generated
& we convert different forms of data to similar datatype.]

5. Data mining : Extracting meaningful inkigts from ~~to~~ large

Post processing

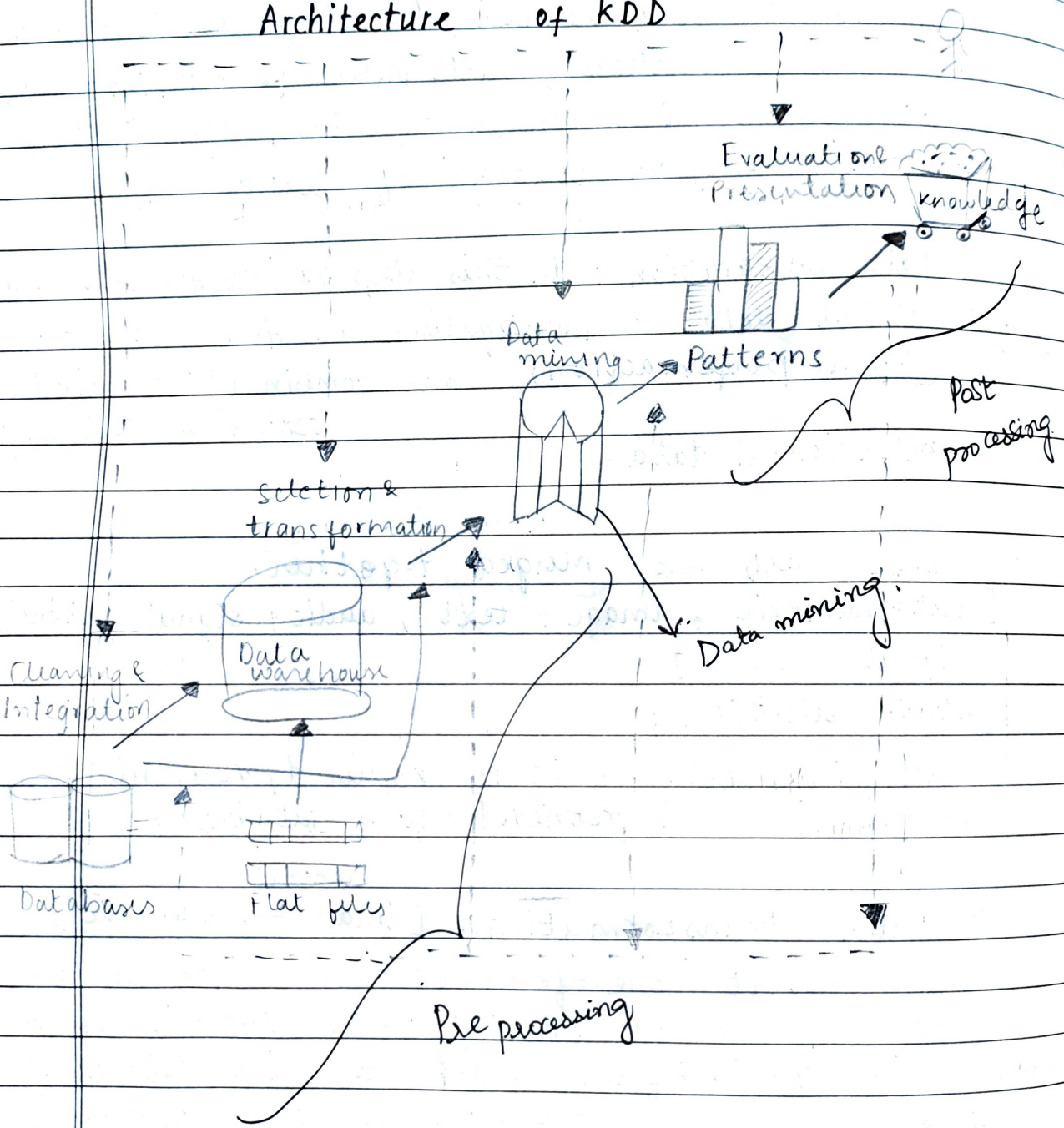
6. Pattern evaluation : Interesting measures.
support & confidence.

Post processing

7. knowledge presentation : Visualization.

→ tableu

Architecture of KDD



(not always structured), Query processing
Database always relies on transactional data
(Present data)

Datawarehouse (Historic data) } Both storage, retrieval / mgmt

Knowledge base always in a structured formats.

Applied by Intelligent agents [Robots]

→ Used for decision making & Business Intelligence

~~Diff types of databases can be~~
1) Numerical data, categorical, text, image, audio data, time series
2) RDBMS, NoSQL, Data warehouses, Distributed DB.

what kinds of data can be trained?
Illustrate various databases wrt datamining.

13-8-24 Classification of Attributes

Attribute / Entity / Object

Emp S.no	Eid	Ename	....

1. Categorical data: [Qualitative data]
a) Nominal: the values of the nominal attribute are different names i.e., they provide the enough info to distinguish one object with

idea

papergrid
Date: / /

other object: $< = \neq >$

zipcodes empidnos eyecolor gender are examples of nominal.

b) Ordinal value: The values of ordinal attributes provide info to order the objects $(<, >)$, Ranking.
{Good, better, best}, {High, Medium, low}
Basing on context we try to segregate it from higher to lower.

~~Interr~~
Numeric data : [Quantitative]

a) Interval data: For interval attributes the diff b/n values are meaningful where a unit of measurement exists. $(+, -)$
Ex: Thermometer (Fixed value) → Low fever / high f
Calendar dates

b) Ratio data: For Ratio variables both differences ratio are meaningful $(\times, \div)$
{Electriccurrent, Temp in Kelvin, Count values, Age, Mass, Length} Physical values.

Discrete and Continuous Variables / data / attributes:
↳ no floating / decimal exists
{1, 2, 3, 4, ... ∞}

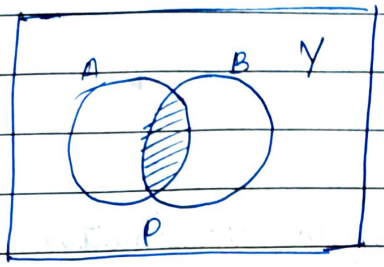
continous - floating / decimal exits
{1, 1.1, 1.2, 1.3, ..., 2 ..., ∞}

14/8/24

Sampling

Represents entire population

↳ <a-z, A-Z>, <1, 2, ∞>, <colleges>



A ∩ B - sampling / ep

Covid -19:

Problems → spreadful, No diagonis, Symptoms are changing. Vaiations

Based on Age, Gendr, localities, Genetic (samples)

diabeties BP.

the made a sample space

Central Limit theorem

S ∈ P. if the soln works for subset 's' then the soln works for whole population 'p'

Precautions

Date: / /

Descriptive statistics

1. Outcome is known
2. Past data
3. Summary / characteristics

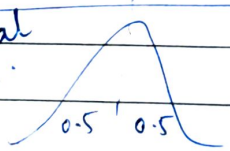
(Precaution - Covid-19)

→ It is a branch of stats that involves organization summary of data

→ Usually, mean, median & mode are used to describe the avg values of data & are reffered to be central tendency in nature

→ We can measure shape & size of probability distribution.

Normal dist.



mean
= median
= mode

Data is uniform

Inferential statistics.

→ Experimentation.

(After vaccination the reactions)

→ It deals with generalisation about a population based on sample data & it draws the conclusion from sample space s'

→ Applications are Regression analysis, Hypothesis tests (Assumptions) (Rel^n b/n independent & dep var's)

Example of Covid -19 :

* In case of covid-19 descriptive stats deals with the following
1) provides info of symptoms for covid
2) Dist of cases accross various regions
3) Frequency of complaints given by individuals
4) Usage of preventive measures like wearing gloves, masks, vaccinated...
5) It doesn't make any assumptions.

* Inferential stats deals w the following:

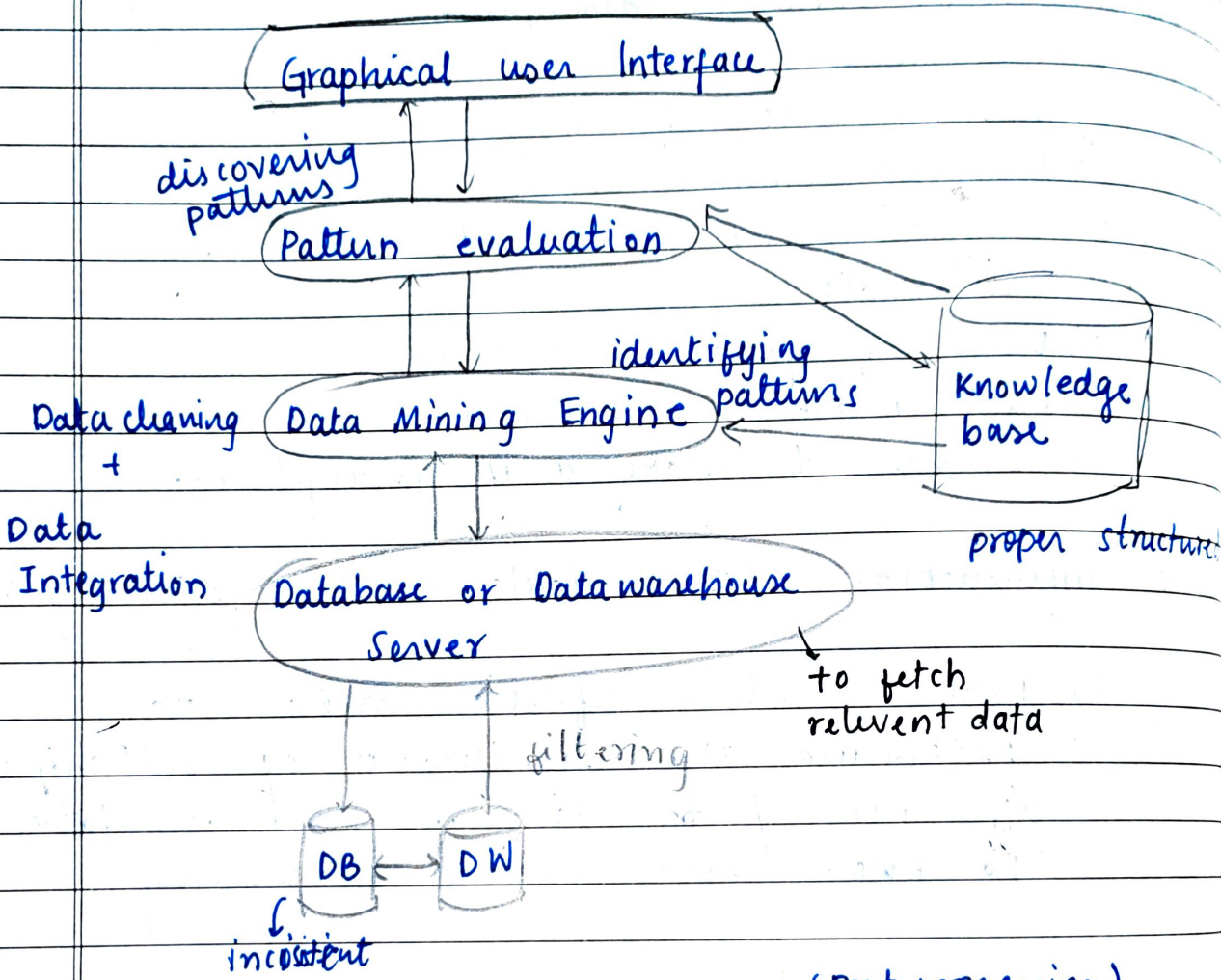
1) Deal w the infection rate, to
2) Determine the significance of rel^n ships, differences / trends observed in data (to know how vaccine is working)

17/8/24

papergrid
Date: / /

Important Question

* Typical Data Mining Architecture



(Post processing)

→ Focuses on pattern evaluation whereas KDD focuses on knowledge discovery. (pre processing)

Although we have knowledge base, we use DB as it is dynamic.

interesting patterns from where we can derive something.

from all the patterns generated from data mining Engine interesting patterns are identified.

Data Mining Engine:

→ Provides the functionality of classification, clustering, prediction, outlier analysis, association analysis, evolution analysis.

Knowledge base engine:

→ Discovering the hidden patterns which are at the measure of interestingness.
→ It includes concept hierarchy
→ It may also include the users belief which relies on multiple hypothesis.

Pattern evaluation module:

Thresholds are set to filter out the discovered patterns

GUI:

Its a communication channel b/n user and client, A query is given, searching takes place through exploratory datamining & then reports are generated based on visualizations.

representing theoritical in numerical form.

19/8/24 ✱ Classification of Mining Techniques : papergrid Date: / /

Frequent patterns finding ⇒ Data mining

Case 1 : Customer id, product purchase, sales - investment → maggi profit/loss %

discount, trends,

primary factor - customer visiting regularly.

2) Text mining : c

case - 2 : Online shopping.

customer reviews, ratings, opinions, feedbacks

↓

+ve/-ve (text) feedback - +ve/-ve/neutral.

sentiment - functionalities.

text - opinions.

↓

text / opinion mining

online - Navigating patterns (Amazon recomendation sys).

↓

web mining

1. Data Mining :
Ex : Customer id, purchase date, product id, Quantity purchased.

2. When you try to apply any sentiment on text / any review. Its termed as opinion mining

opinion mining is subset of text mining

3. Web mining:

case - 3: Social media / Group communities.
. Marketing, Advertising.

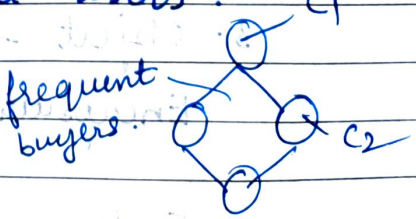
update latest info - (Linked in)
large investment ⇒ investment pattrns.
↓
Network / Graphs mining
bfs, dfs
dynamic.
nodes, links, edges, weights.
(searching becomes easy)

Ex - Amazon reviews Clique stream analysis, topic modelling

We focus mostly on navigating pattrns hence termed to be web mining.

4. Graph mining:

Social network relnsh online media, Group communities, influencial users. for targeted marketing.



5. Trajectory mining:

Any movement patterns analysis of traffic flow, GPS
Main objective of trajjectory mining is to optimize the traffic flow.

case - 4: Moving objects / traffic analysis

Trajectory mining.

6. Association mining :

1. Data mining / On which kind of data can be mined ?

A- 1) We can mine the data on the general

1. Relational Database :

Rows, Colls, Entities.

2. Transactional Database :

wrt. Individual customers

wrt. Product similarity.

} Association Analysis.

Transaction Id	Products

3. Object - oriented database :

Encapsulation & Reuse

Each entity is considered as object

4. Object-Relational Database :

→ handles complex queries.
→ Uses concept of class hierarchy.

5. Spatial database: GPS, Map database,
Medical images. (Raster format)
satellite images
(To find out park near to house)

6. Temporal database:
Focuses on timestamp ordering.

7. Timeseries database:
Focuses on periodic trends.

8. Text database : WWW, ~~huge~~ (unstructured) documents

9. Multimedia database:
audio + video + text + speech combining.
Any Content deliver.

10. Hetro geneous & Legacy (Historic data) database
It combines systems such as Relational, Object oriented, N/w based, hierarchical approaches, file systems, spread sheets & multimedia.

Date: 20/8/24

Data Characterization & Discrimination.

1. Data can be associated with class / concept.
   Ex: In a retailers shop set of items considering
   General / whole → <computer, printer> → Class
   Specific → <bigspenders, budget spenders> → Concept

Data characterization

2. It also describes individual class / concept in precised & summarised manner.
   such type of summarised names / description are called concept hierarchies.
   It also explain features of data.
   Summarizing the data under study is called target class

Examples of Data characterization

- Customer purchasing a product, sales, avg sales

Data Discrimination:

It compares target class w another class from the set of items.

Computer
- s/w - Functionalities
- H/w - cpu, mouse

HP computers	Dell
90% Buyers frequent	10% Buyers In frequent

☆ Data cleaning:

1. Missing values 2. Noisy data

1. Missing values:

1. Ignore the tuples.
if there are 10 diff columns out of which 10th col represent % & 9 cols are missing

Rollno	name	mark	m2	m3	m4	....	per	
-	-	-	-	-	-	-	90%	(ignore it)

2. Fill the missing values manually:
If its a small dataset, the missing values can be filled manually ex: Avg of marks.
3. Use a global constant such as 'unknown' or '-∞'
If all the values are missing from that column, the it takes "unknown" as a pattern
4. Use the attribute mean to fill the missing values:

Ex: Avg income of all electronics customers is $28,000 we use this value to replace the missing, Height of missing individual.

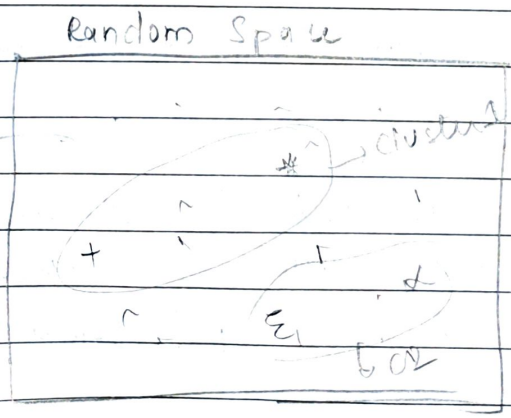
Date: / /

5. Use the attribute mean for all the samples belonging to same class as given tupple.
Ex: If classifying cust acc to credit_risk, replace missing value w avg income value for customrs in the same credit_risk category as that of given tuple.

6. Use the most probable value to fill the missing values.
Regression model, decision tree induction - methods

2. Noisy Data:

Noise is a random error / variance in a measured variable.
Given a numeric attribute such as price, how can we smooth the data to remove the noise, this could be achieved through clustering mechanism.



distance based approach

If we have + & * we try grouping +'s on 1 side & *'s on another

But if we also have Σ & α then + & * form 1 group & Σ, α as one.

3 Regression :

Regression model best fits the relation b/n independant & dep. var's & is formed from the base point,

$y = mx + c$.

4. Combine Human & Computer introspection:

Σ = {alphabet} if '0' & '7' are input they are

Data Transformation & Integration.

→ Smoothing of Data - removing noisy data
We have 3 diff approaches
1. binning
2. Clustering
3. Regression.

→ Data aggregation : sumarize data from (avg. marks, fail %)
different forms & levels (Granularity).
& generalizing (only on sample data)

→ Data Generalization : (On Entire data / population)

→ Data Normalization :
Generally we represent data from (0-1) or
z-score [scaling]
↳ Performing analysis is easy

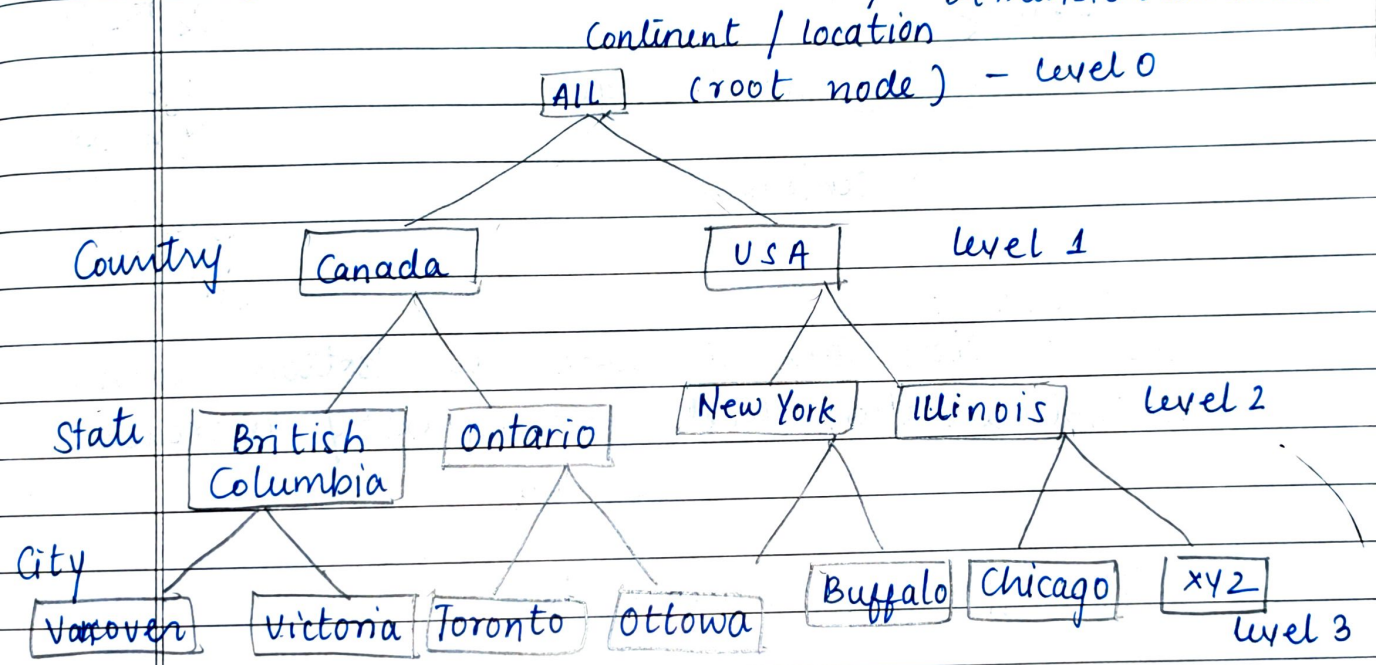
Date: 23/8/24

Expected problems.

5M ✱ **Concept Hierarchies**

Concept hierarchy defines sequence of mappings from a set of low level concepts to higher level in turn resulting in general concepts.

→ subsets are low level whereas supersets are high level



Concept Hierarchy for attribute dimension

Dimension attribute is described by attribute no., street, city, ... country

These attributes are related to total order forming a concept hierarchy such as street < city < state < country (A.O).

↓

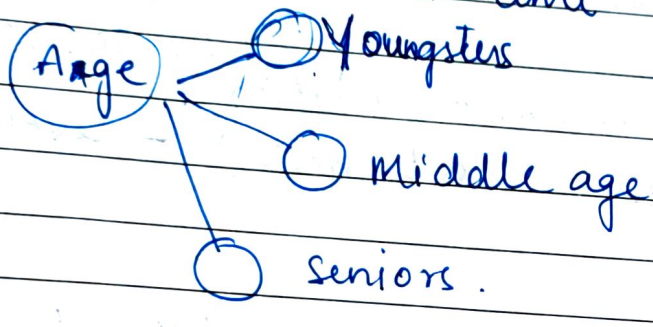
Higher granularity / multiple "

Concept Hierarchy is the total / partial order for Time dimension

day < week < month < year < decade

& this can also be classified as seconds.

Advantage is we are generalising the data. searching traversal time is reduced.



Graph (directional)

drawbacks :

1. Construction time is more at bottom levels.
2. Implementation level - complexity increases

4 diff types of Concept h

1. Schema hierarchy
2. Set grouping hierarchy
3. Operation based hierarchy
4. Rule based hierarchy

Preprocessing & Discovering patterns can be automat but evaluating patterns can't be automated (human intervention is required) Knowledge Engineer

Data Engineer - Analysis Reports/projects visualization

Raw → actual

papergrid

Date: / /

10M

Data mining primitives : (5M) ← tasks

Steps for Data mining tasks / query:

1. Task relevant data (5M)
2. what kind of knowledge to mined? (5M)
3. what type of background knowledge is required?
4. what kind of Interesting measures we require for patterns discovery in Data mining.
5. How would you represent the data (mined)?

Myth: Data mining systems can autonomously pick out all the valuable knowledge that is embedded in a given large database with human intervention.

1. Task relevant data.

Ex.) A group of boys segregated rank wise
- → Physical appearance /
- → Marks
- → Skills.

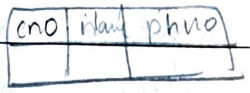
The selection depends on End user biased condn → colour, religion

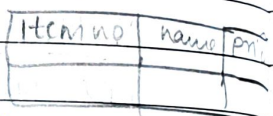
Specificity is → Placement requires skills
→ NCC / Army requires physical

Ex-2) If a data mining task is to study the associations b/n the items that are frequently purchased at all electronics by the customers in Canada, what could be specified in task relevant data.

* dirty data: even after preprocessing, if no meaniful patterns are discovered.

1. → Task relevant data can be specified the name of the database / data warehouse all electronics_db. which includes tuples, tables, containing info such as purchase rate, customer details & items sold.





2. We probably define the conditions for ~~task~~ relevent data.
Ex- Retrieve data i.e. pretraining to purchase made in china for the current year.

3. Relevent attributes / dimensions.
Ex: name & price from item table
income & age from customer table

2. What kind of knowledge to be mined?

→ Electronic store

Customer purchasing computer is likely to buy a printer ← associated (Rel^n).

printer is dependent on computer

user defined { probability - 60%. likely to happen
Confidence - 2%. already happened (not fixed)

→ based on Association Analysis.

3. What kind of Background knowledge is required

By using Concept Hierarchy, they represent low level data & high level data
→ Generalization is used. used in Backend.

1) Task Relavent data

→ Database / Datawarehouse, Conditions for data selection.

2) Knowledge to be mined
→ Characterization & discrimination.
→ Classification & prediction.
→ Clustring & association.

3) Background knowledge to be mined

1. Concept hierarchies
2. User belifs about relationships in da

4) Pattern interesting measures

1. Utility(2), Certainty(1), simplicity(4), novelty(3), support & confidence

Date: / /

⭐⭐⭐ Presentation & Visualisation of discovered patterns

1. Rule based.

age(X, 'young') and income(x, High) => xyz.

Define rule w certain conditions.

generalized version. — space complexities

II Table based.

Age	income	class	count
		A	
		B	
		C	

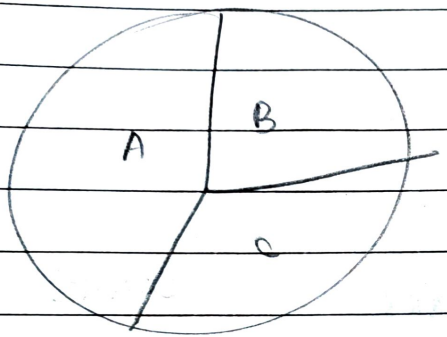
data characterization.

III Cross tab — specific concept

age	income		class		
	high	low	A	B	C
~~Young~~ Old					
~~count~~					

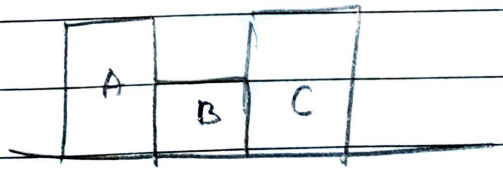
used for data discrimination

(IV) Pie-chart



To highlight some features.

(V) Bar chart



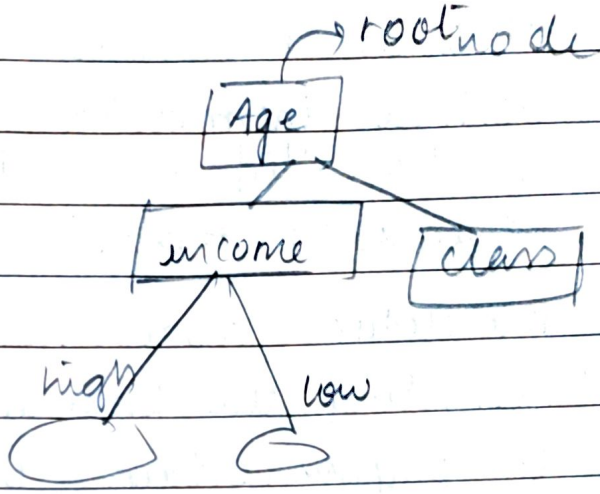
Date: / /

VI Decision tree

tree data structure

ML model

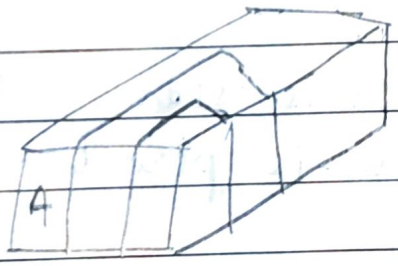
Works very well for large data



VII Data cubes - Data warehouses in xyz axes

Dimension space

each class represent 1 dimension



27/8/24 Knowledge Representation methods. Date: / /

→ o/p is decision making.

What is knowledge ?

Derived information (from unstructured data) is also called knowledge.

Advantage: Analysis

Knowledge representation (KR) is process of extracting the knowledge from derived info, to perform certain tasks for analysing, basing on performance speed evaluating its measures & generating the summary inform of reports, visualisation, graphs, ...

Truth / Logic / Myths:

↓ whatever has mathematical approach

Sun is star - truth → Earth is spherical

Sun is planet → false → Earth is flat

→ myth

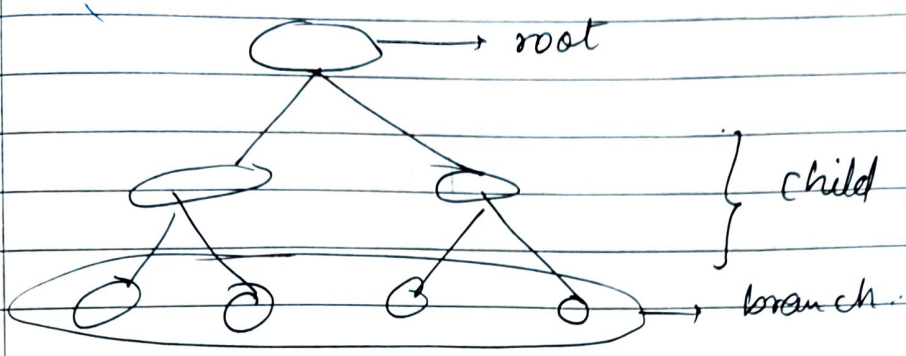
Acc. to Mythology, Sun is God (Planet) → truth (SunGod Ra)

Acc to Flat theory, Earth is flat - truth.

Various forms:

1. Decision Tree: i) Used for classification & Regression problem.
   derived from BST.
   searching becomes easy as well as cleaning
2) It represents knowledge as a tree like structure where each node represents a decision to the attribute & each branch represents

outcome of the decision.



Example: A decision tree for classifying whether a customer will buy a product basing on the attributes like age, income / previous purchase history.

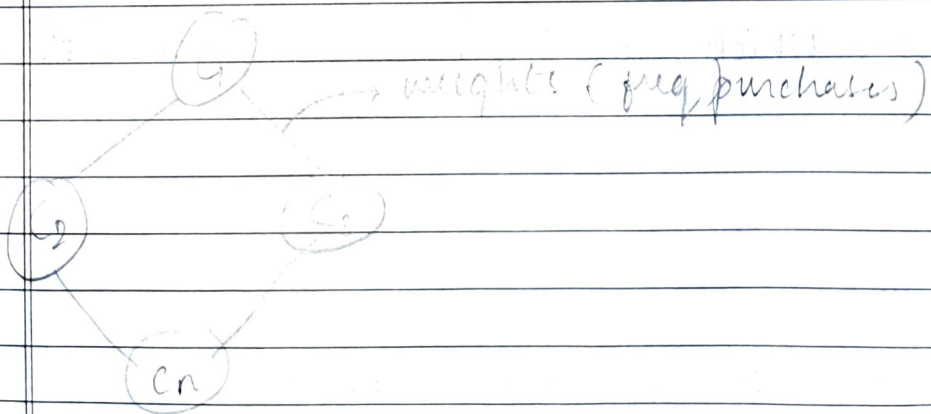
2. Association Rules:

3. Graphs & Networks:

↓

easy to represent large amt of data.

(Social N/w)



Graph represents data as nodes & edges
The main applications of graphs are

Recomendation systems.
Social Network analysis

4. Rule based Systems : Chat GPT
If - then.
Traffic.

→ Maintaining Consistency.

Disadv : Hard Threshold. (no feasibility).
(Objective exams)

Example: In a medical diagnosis sys, a rule might be 'If the patient has high fever & a sore throat then patient might hav severe cough'

5. Ontologies and Semantics:

object = which describes behaviour of entity.

Protege editor

Manual efforts are more

A is a good boy
Entity Rel^n object

3/9/24 UNIT - 2

Market Basket Analysis.

In late 1980's, the Although, the products are present in the shop, the consumer didn't buy them because they are unorganised.

Market Basket.

{milk, bread} ⟶ {butter}. : Rule

Association / coherence Analysis.

1. Given a set of transactions, find the rules that will predict the occurence of an item based on occurence of other items in the transaction. This concept is refered as Market basket analysis / transactions.

Tid	Iname
1	bread, milk
2	bread, diaper, chocolate, eggs
3	milk, diaper, chocolate, coke
4	bread, milk, diaper, chocolate
5	bread, milk, coke, diaper.

bread - item {bread} - itemset

K-items set : $K_d = \{K_1, K_2, K_3, K_4, \ldots\}$.

2. Support count (σ): Frequency of occurence of an item set.

Ex: $\sigma_{milk, bread, diaper} = 2$.

Date: / /

3. Support (s): It is fraction of transaction that contains as item set. (s) = 2/5

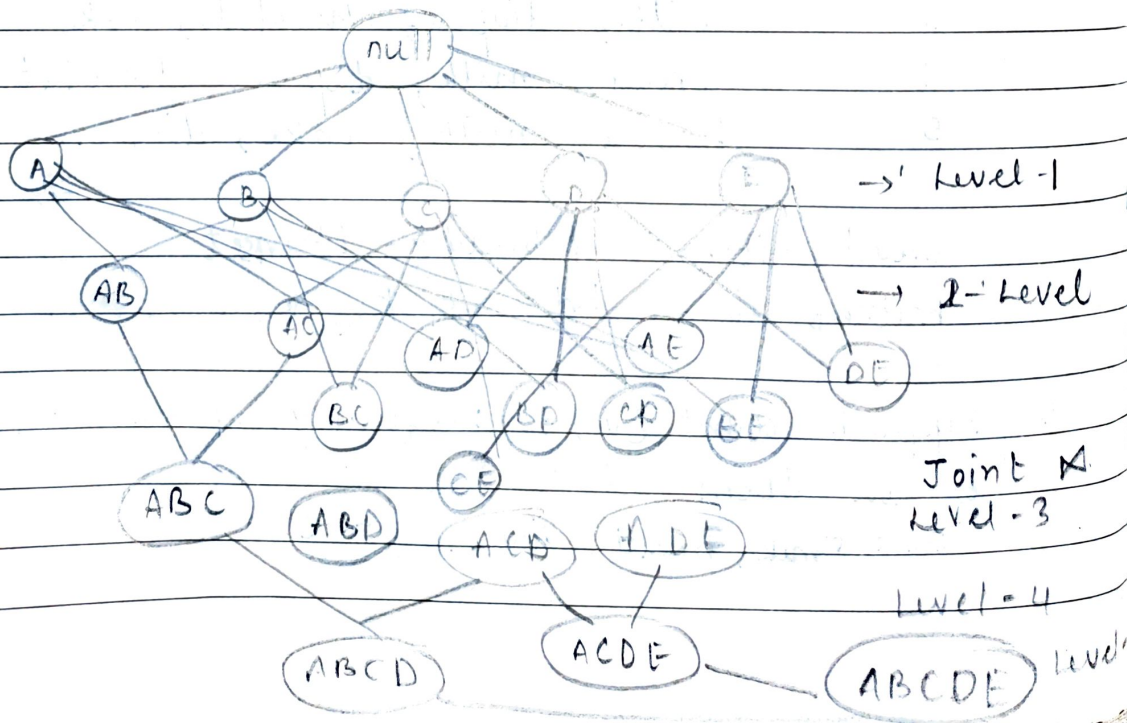
4. Frequent item set:
An item set whose support count is >= minimum support is called a frequent item set

minsupport = threshold value = .

5. Confidence: (c)
It measures how often items in Y appear in transaction that contain X

→ If milk is frequently brought & Rice is frequently brought seperately, it need not be that {milk, Rice} is a frequent item set.

★ Brute Force method / Approach:
If we have d items, possible no.of itemsets are 2^d. (wo repetation).

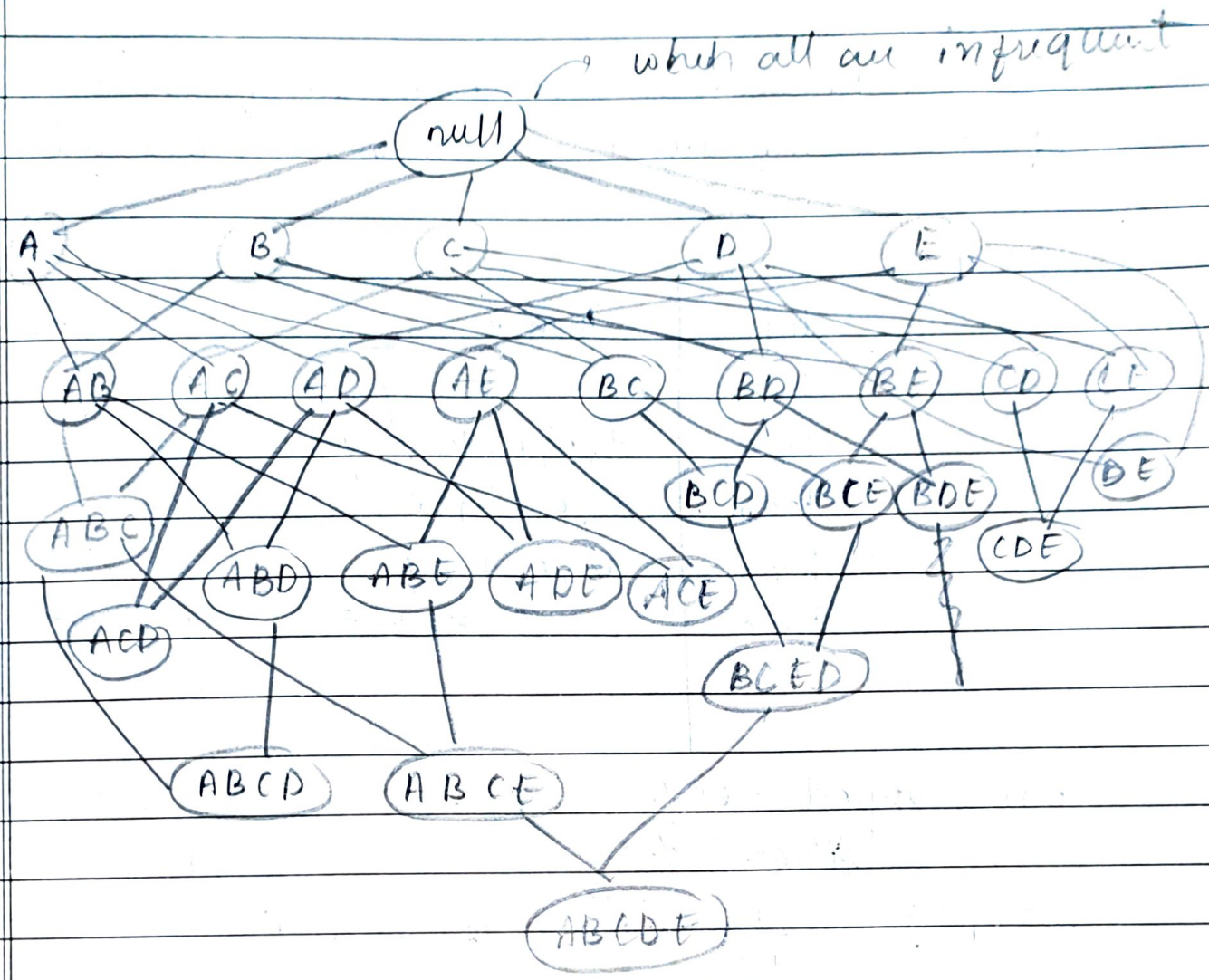


papergrid

Date: / /

Draw back:

→ Computationally prohibited / expensive



4/9/24

drawback:

No track of count

If A is infrequent then all the sets consiting A are inconsistent.

→ We need not draw a tree for lower freq.

Continous semantic

Pruning is difficult.

✱ Apriori Algorithm

Apriori principle:
A is infrequent AB will never be frequent
If AB is infrequent A may be infrequent.

Example:

Tid	Iname
T_{100}	m, o, n, k, e, y
T_{200}	d, o, n, k, e, y
T_{300}	m, a, k, e
T_{400}	m, c, k, y
T_{500}	c, o, o, k, i, e

<m> 3 ✓
<o> 4 ✓
<n> 2
<k> 5 ✓
<e> 4 ✓
<y> 3 ✓
<d> 1
<a> 1
<c> 1
<i> 1

min support = 60%.
min confidence = 80%.

support = $\frac{60}{100} \times 5$ (no of transactions)

= 3.

take sets having >= support value.

L_2

<m, o> 1
<m, k> 3
<m, e> 2
~~<m, y>~~
<o, k> 2
<o, e> 3
~~<k, y>~~
<k, e> 4.

L_1

m	3
o	4
k	5
e	4

L3

$<m, k, o>$	1	
$<m, k, e>$	2	$\Rightarrow$ $<o, k, e>$
$<o, k, e>$	3	

No limit for association rules

Interesting patterns : $<o, k, e>$

Association rule	Support	Confidence	confidence %	
$o \wedge k \Rightarrow e$	3	$3/o \wedge k = 3/3$	1 = 100%	} Interesting patterns.
$o \wedge e \Rightarrow k$	3	$3/o \wedge e = 3/3$	1 = 100%	
$e \wedge k \Rightarrow o$	3	$3/o = 3/4$	= 75%	} having < than 80% confidence
$e \Rightarrow o \wedge k$	3	$3/e = 3/4$	= 75%	
$k \Rightarrow o \wedge e$	3	$3/k = 3/5$	= 0.6	
$o \Rightarrow e \wedge k$	3	$3/o = 3/4$	= 75%	

$n! = 3! = 6$

n = support count

9/9/24

Date: / /

★ FP Growth Algorithm:

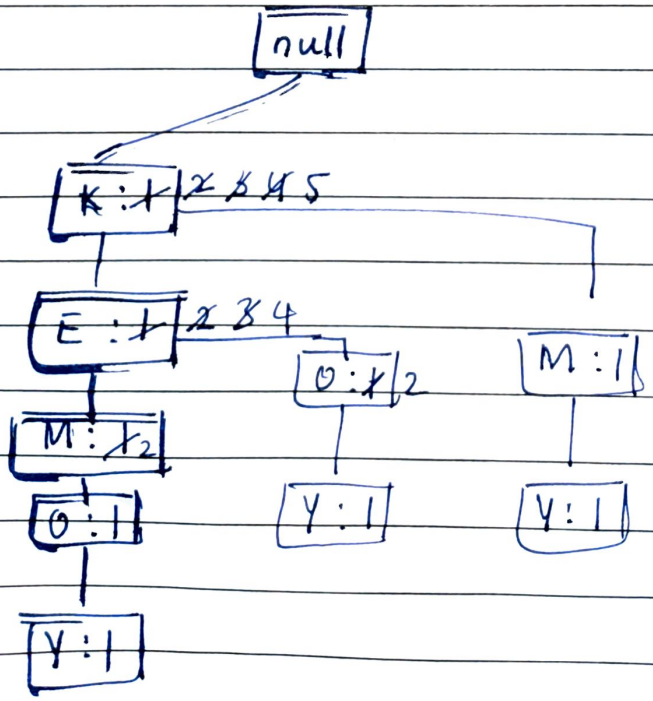
Mining the complete set of frequent patterns using a tree structure for storing info abt FP.

min support ≥ 3.

Tid	Iname	ordered list-item
T_1	e, n, m, e, k, y	{k, e, m, o, y}
T_2	k, y, e, n, d, o	{k, E, O, Y}
T_3	k, m, a, e	{K, E, M}
T_4	m, y, c, k, u	{K, M, Y}
T_5	c, i, e, k, o	{K, E, O}

K=5, E-4, M-3, O-3; Y-3.

FP-Growth tree construction



items	Conditional pattern base	Conditional fptree
Y	{KEMO:1, KEO:1, KM:1}	{K:3}
O	{KEM:1, KE:2}	{KE:3}
M	{KE:2, K:1}	{K:3}
E	{K:4}	{K:4}
K	{∅}	{∅}

Items	FP-Growth
Y	{K, Y:3}
O	{K, O:3, E, O:3}
M	{K, M:3}
E	{K, E:4}
K	{∅}

Drawbacks of FP Growth :

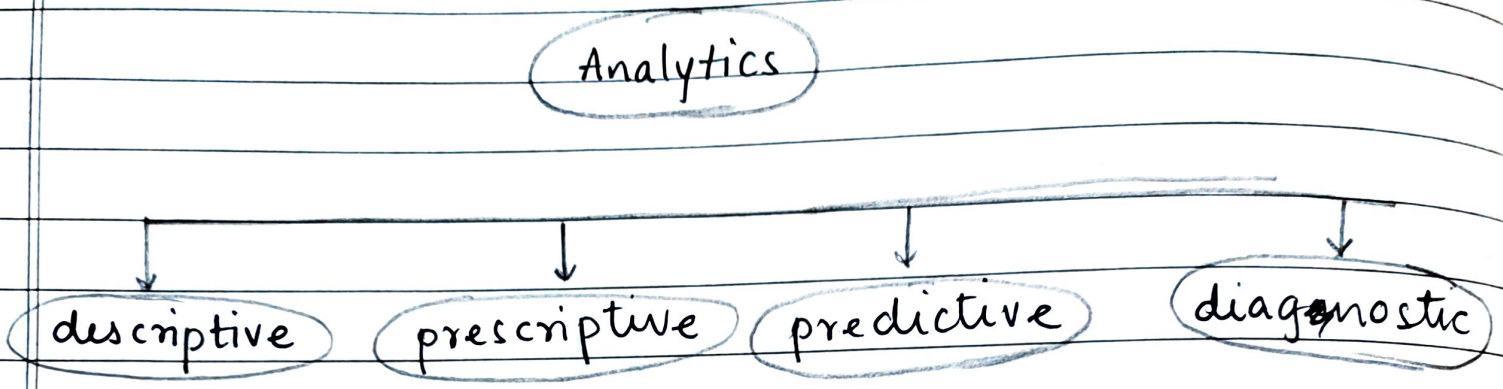
→ If its a large database tree construction becomes complex

→ After every step we have to see association rules.

11/9/24

papergrid
Date: / /

Unit - 3 Data modelling



Forecasting techniques

Heuristic approach

Moving average

Date: / /

10/9/24

1. For each of the following provide an example of an association rule from the market basket domain that satisfies the following condition also describe whether such rules are subjectively interesting.

1) A rule that has high support & high confidence
2) A rule that has high support & low confidence.
3) low support & low confidence
4) A rule that has low support & high confidence

2.

2. Consider the following dataset (8m)

Customer_id	T_id	items bought
1	0001	{a,d,e}
1	0024	{a,b,c,e}
2	0012	{a,b,d,e}
2	0031	{a,c,d,e}
3	0015	{b,c,e}
3	0022	{b,d,e}
4	0029	{c,d}
4	0040	{a,i,e}
5	0030	{a,d,e}
5	0038	{a,b,e}

a. Compute the support for itemsets {e}, {b,d} and {b,d,e} by treating each t-id as a market basket.

b. Use the results of Part(a) to compute the for an association rules {b,d} → {e} & {e} → {b,d} Is confidence a symmetric measure?

c. Repeat part (a) by treating each customer_id as Market basket. Each item should be treated as binary variable. if an item appears in atleast one transaction bought by customer & 0 otherwise

d. Use results in part c to compute the confidence for the association rule

$\{b,d\} \rightarrow \{e\}$ & $\{e\} \rightarrow \{b,d\}$.