Date : 21.08.2023

No :

ST-402 - Data mining

Introduction

Data science involvement becaue of,

- Hardware cost reduced.
- Processing power increased
- Having good internet facilities.
- Networking power increased.

Data minig Applications

Data mining can be classified into 6 groups.

- prediction and Description.
- Customer Relationship Marketing (CRM)

Data

3 v's

- Volume (storage)
- Variety (types, images, videos etc)
- Velocity (speed)

1) Predict and Description

techniques: Selecting some or all the objects in the db to predict other variables of interest.

2) Customer Relationship Marketing (CRM)

Techniques: Cluster analysis to identify customers suitable for cross selling other products.

Date : ..... No : .....

* google → use CRM.
  - provide links
  - Suggestion gives. → through cookies.
* Supermarkets → regular buying. / most buying. } analys pattern
* Fraud Detection (Debit, credit card using)
  - Behaviours analysis (daily transaction detection)

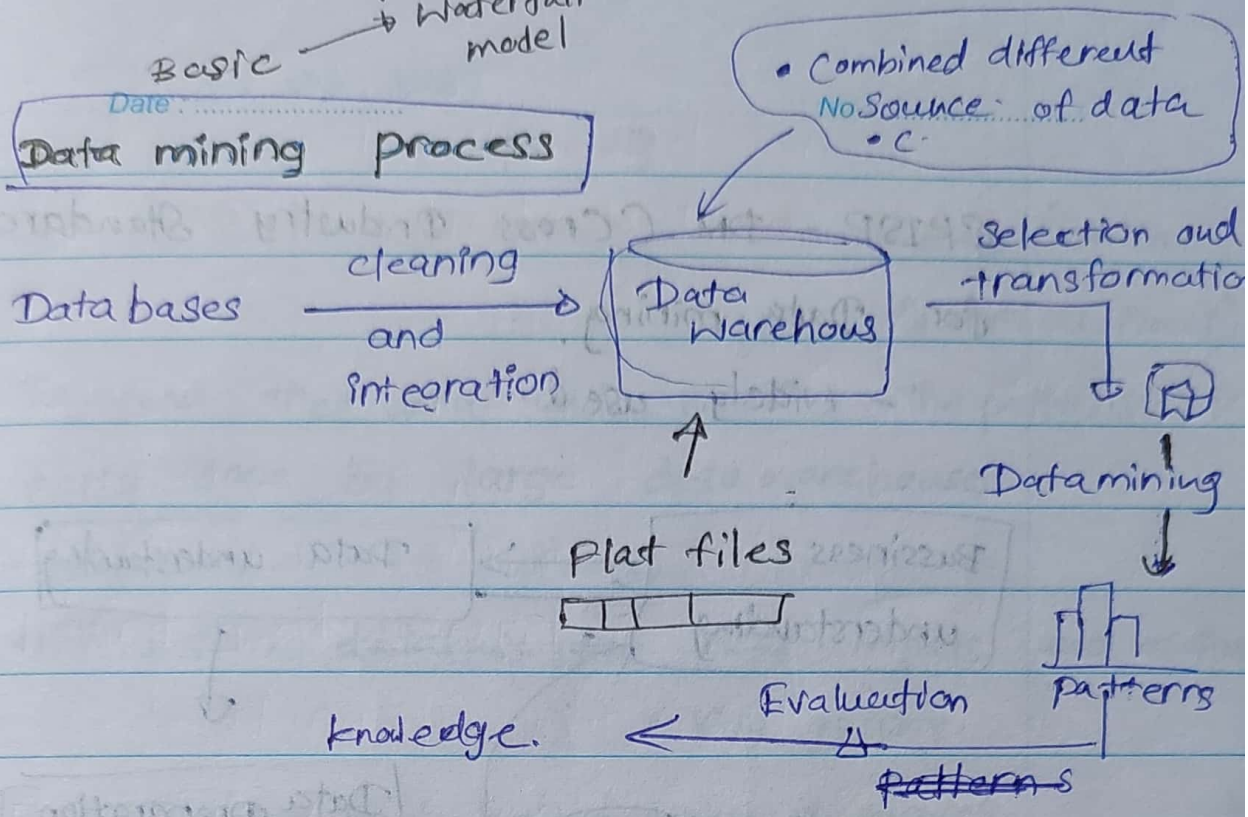
3) Custome profiling
4) - loyality cards.
5) Customer segmentation:
   - good quality items.
   ~~- loyality cards~~.

6) Web site Design and promotion.
   - google Coockies.

Basic → Waterfall model

Date:

Data mining process



preprocessing → Cleaning

missing values: ~~remov~~ i

~~normalizat~~

DB normalization → increased accassability to data & remove duplicates (Main for minimize dupliation)

• to extraction infor quickly.

Data mart → subset of data ware house

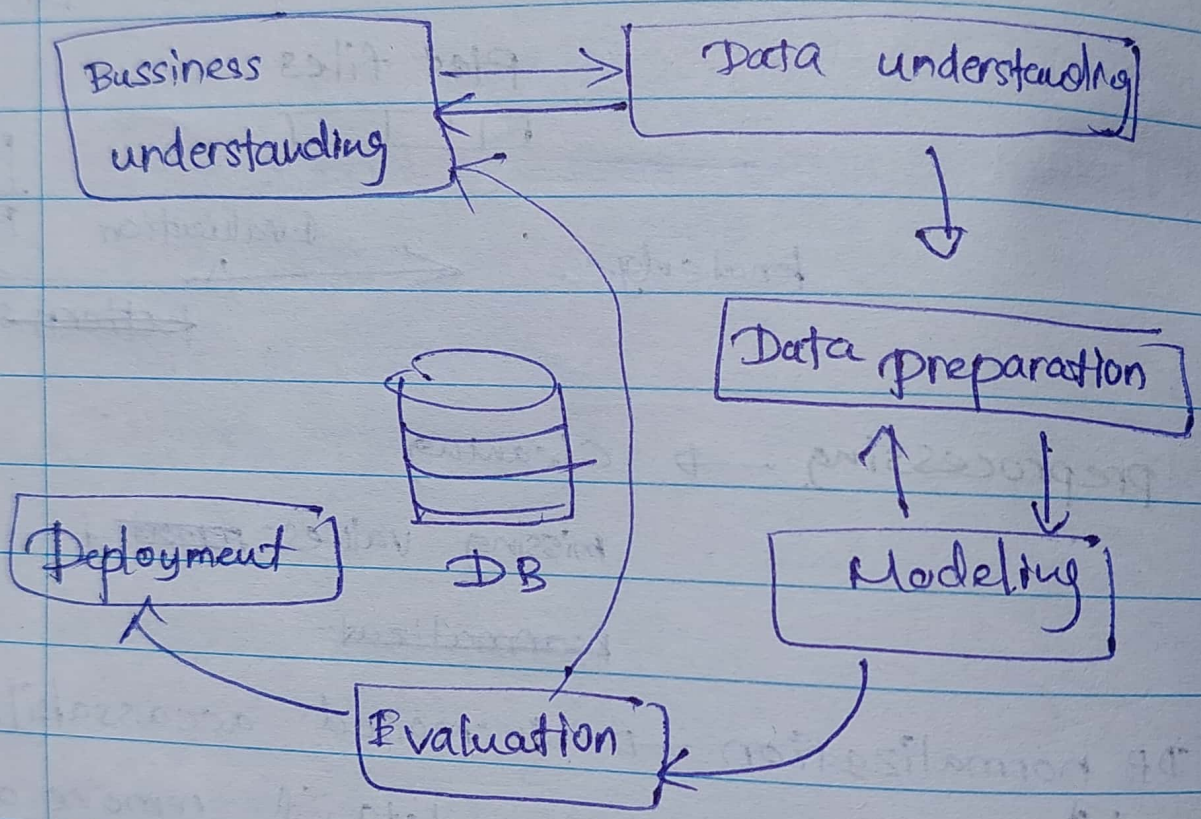
analysing → • identify different hidden patterns

• There are some information that are more impacts and more valuble → knowledge.

Date : ..... No : .....

CRISP - DM (Cross Industry Standard Process for Data mining.

- widely used.



Date : 28.08.2023

No :

Lecture 03

Data Warehouse and OLAP (online Analytical Process)

- To read the data and finding the patterns ⇒ early done by large data warehouse

- transaction database ⇒ need inputs and outputs very quickly.

- data warehous → having many duplicates.

- ~~Various dimension~~

Data warehouse makes it possible to examine.

- Trends and patterns
- Various dimensions

OLAP → process for modeling and viewing data in multiple dimensions.

evolutionary → dB's connected together.

Date : ..... No : .....

Decision Support Systems (DSS)

- Enterprise-wide → Concern about all the aspects.

4 specific aspects of datawarehouse

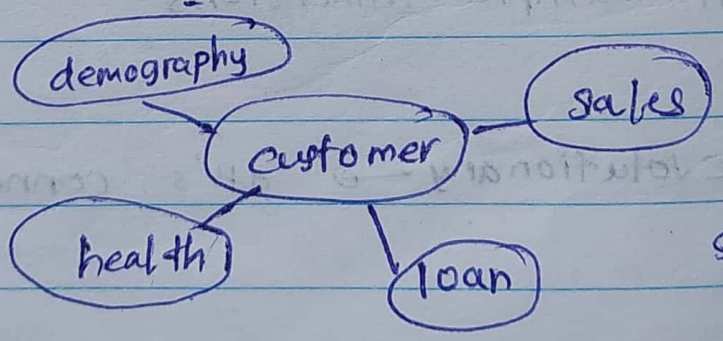
- ~~Specific Subject spe~~
- Subject Oriented
- integrated.
- Time variant
- Non volatile.

Subject Oriented

- consider about different decisions.

eg: focus around customer
focus around suppliers.

eg:



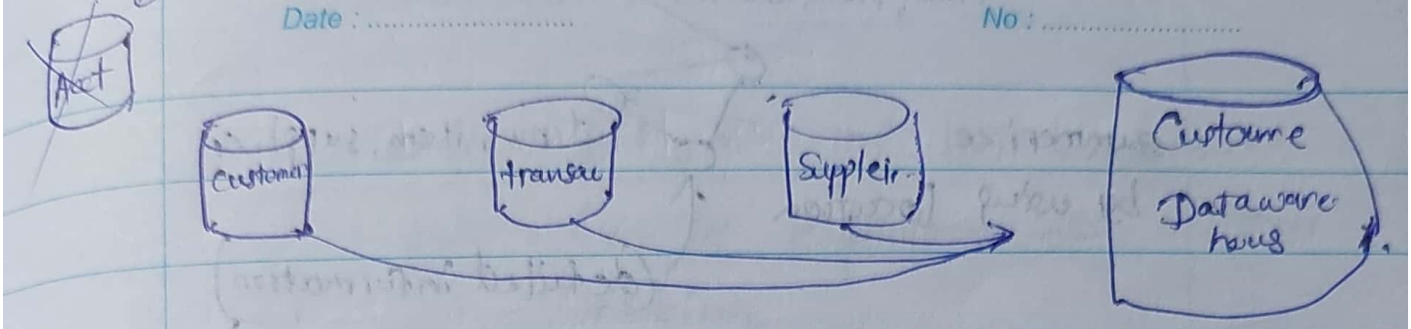
subject → Customer.

Integrated

different dB into a single dB.

may Not relavant

Date : ..... No : .....



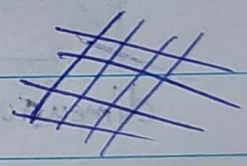
Time - Varient

- timestamp.
- historical data.

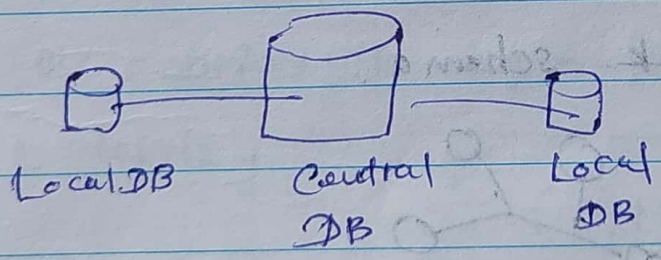
Non volatile

- required two operations.
  - initial loading of data.
  - access of data.

Data Warehouse



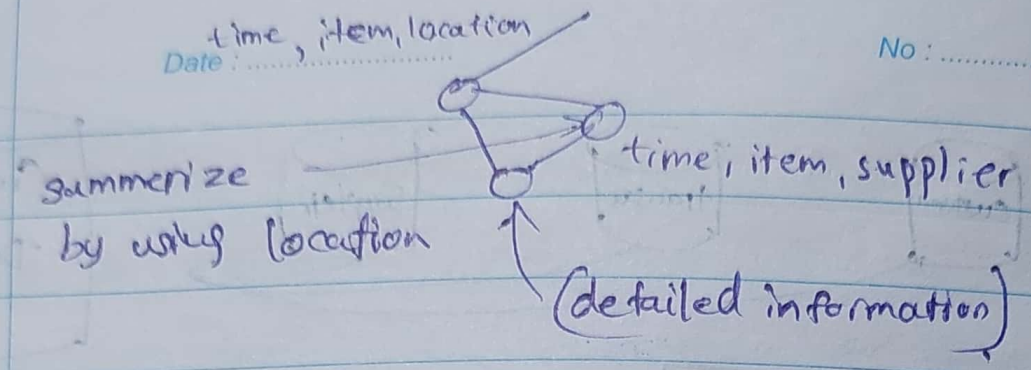
lattice structure



unified schemas ⟹
- relational DB
- Exel sheet.
-

stored in single dB
(integrated/ Combine)

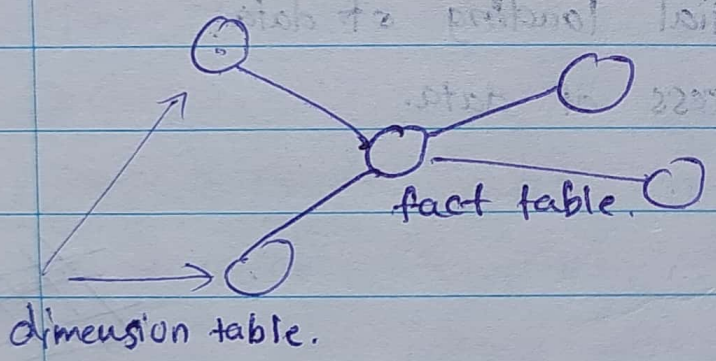
Date : ..... No : .....



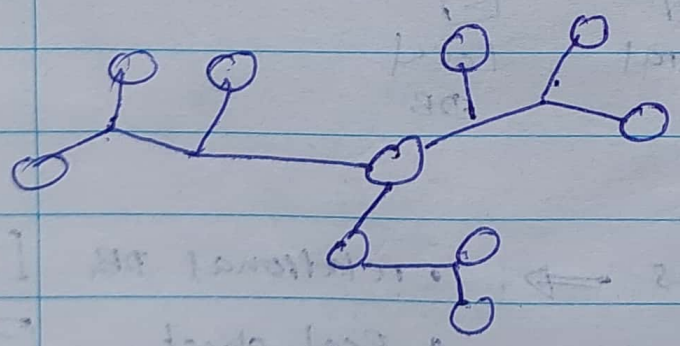
- diferent level of summerization done using a lattice of cuboids.

Multidimensional Data Model.

- star schema.



- snowflak schema.



Date : 2023 . 09.04 No :

Data Marts : • Sub parts of data warehouse
• Use for department wise decision making.

Efficient Data Cube Computation

Data Cube : Summarize the data with different variables.

- If one variable contain L_i levels (for example in time variable, year, monthly, daily, hourly - 4 labels)
- For this it have $L_i + 1$ lables. (∵ We get average record of each variable)

cuboids in an n - dimensional cube with L levels

$$\boxed{T = \prod_{i=1}^{n} (L_i + 1)}$$

add 1 to include the virtual top level of all

Ex: 3 variables each have 2 levels.

$$\# \text{ cuboids} = \prod_{i=1}^{3} (2+1) = 3^3 = 27$$

Date : ..... No : .....

Materialization of the data cube

- Materialize every (cuboid), non or same

full materialization — no mate. — partial

Sales by mont.

Tran #	Item #	date-time	price	units	total

sort using the field.

We can optimize this process if we know we have to get monthly sales or weekly sales, after collecting monthly or weekly data, we can calculate sales and store it in data warehouse for further uses. So we don't have to use the entire table for decisions. This called as "pre-summerize data".

Materialization

Group By.

cube operation

Need compute the following Group-By:

(date, product, customer) ← lowest level

(date, product), (date, customer), (product, customer)

(date) (Product) (Customer)

()

Date : ..... No : .....

Data warehouse usage.

- Information processing
- Analytical processing
- Data mining

Size of cuboid is depend on cardinallity (different variables)

Static data : Data does not change (name)

Dynamic Data : Change over time (age.

Data integrity : All the data are recorded.

2023/09/11

Data Cleaning

- Data in real world is Dirty.
  - lots of potentially incorrect data
  - Instrument faulty
  - Human or computer error
  - Transmission error.

- Incomplete.
  - Lacking attribute values.
  - Lacking certain attributes of interest
  - Containing only aggregate data

eg: Occupation = " " (Missing Data)

Date : ..... No : .....

- Noisy : Containing noise, errors, or outliers
  eg: Salary = "-10" (an error)

- Inconsistent : Containing discrepancies in codes or names.
  eg :- Age = "42", Birthday = "03/07/2010"
  - Was rating "1, 2, 3" now rating A, B, C
    historical data
  - discrepancy between duplicate records.

- International (e.g. disguised missing data)
  - Jan 1 as everyone's birthday.

Incomplete Data

- Data is not always available.
  eg: Many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing Data may due to
  - equipment malfunction.
  - inconsistent with other recorded data and thus deleted.
  - data not entered due to misunderstanding
  - certain data may not be considered impor

Date : ..... No : .....

- not register history or changes of the data.

* How to handle missing data

- common thing we do is ignoring missing data.

Sol^n 1 : manually find. (experties done)

Sol^n 2 : Automatic replacement

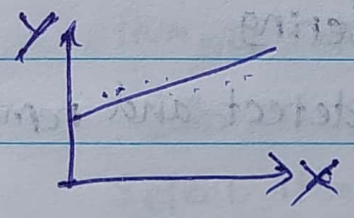
- plot distribution then random generate value for missing value.

Sol^n 3 : • Missing values consider as DON'T CARES.

• Assuming not affected. (Replace by a possible feature value in the domain)

Sol^n 4 : • Can predict the missing values using a model.

X	Y
⋮	⋮



if Y has missing values just take X.

• X and Y are highly correlated

Can use one variable to predict other one.

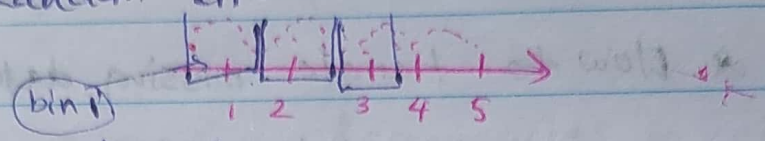
But in data mining

No need to do that. just avoid. Take other variable to that one.

Date : ..... No : .....

Noisy Data

Noise : Random errors or variance.

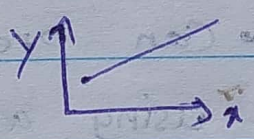


* How handle noisy Data

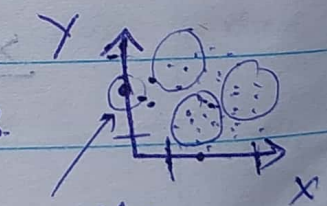
- $\boxed{\text{Binning}}$ ⟹ $\boxed{\text{Reduce random error}}$
  - first sort data and partition into bins. (equal frequencies)
  - Then one can smooth by mean, smooth by median, smooth by bin boundaries (intervals)

    eg: (1.1, 1.23, 0.98, 1.04) ⟹ 1 (Smoothing)
    Replace by

- $\boxed{\text{Regression:}}$
  - Smooth by fitting the data into regression functions.



- Clustering
  - detect and remove outliers.



- Combined computer and human inseption.

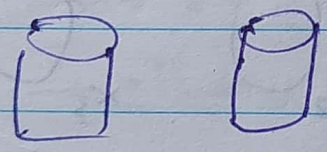
* for x variable that data is a outlier. for y, it is not a outlier. Therefore use clustering

Date : ..... No : .....

Data Cleaning as a process.

- Data migration and integration.
  - Data migration tools:
    - eg: When combining data that has different currencies, this tool use. (Get into same currency)
  - ETL (Extraction / Transformation / Loading) tools: done through GUI
    - eg: Customer centered data house, extracted required data.
- Integration of the two process.
  - Iterative and interactive.

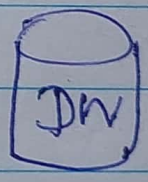
Iterative ⇒



Select item#, unit price from SL_sales into DW

join

Select item#, unitprice from US_Sales into DW



Interactive ⇒ first SL_sales then US_sales.

Date : ..... No : .....

* Data integration problems
  - Entity identification problem
  - ~~Schema integration~~
  - Data value conflicts.
    - attributes values from different sources are different
    - data representation
      eg: British unit and local unit.

Handling redundancy in Data integration
- redundant data occur when integration of multiple dBs.

	Name	Gender	Employment	Salary
R_1	—			
R_2	—	x_1	x_2	x_3

x_1, x_2, x_3 Compared similarity.

- Need to find these two records are same or not. (Similarity measure)

x y dependent $\Rightarrow$ Redundant Data

Date : ..... No : .....

Correlation Analysis (Nominal Data)

if redundant $\Rightarrow$ remove one variable.

else $\Rightarrow$ keep.

- Correlation doesnot imply causality.

  ex: • x and y dependant (x depend on y) and also when x is increasing y also increasing. But we cannot say that x is the reason for increase y.

  (There is a relationship but not reason (cause) for other).

chi-squred test.

$$x^2 = \sum \frac{(\text{observed} - \text{Expected})^2}{\text{Expected}}$$

$x^2 \longrightarrow$ getting large reject null hypothesis.

x^2 test $>$ x^2 table

rejected.

Date : ..... No : .....

Correlation Coefficient

~~$r = \sum a$~~

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

$= [-1, 1]$

if $r_{A,B} > 0$, A and B are positively correlated,

if $r = 0 \Rightarrow$ there is no linear relationship between variable but there are some relationship, so we can't say there is no correlation.

Covariance

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

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

Date : ..... No : .....

Identifying misclassification.

- make sure categorical variables are, all valid and consistent.
- Frequency distribution table are used for cross checking validity of categorical data.

(list down distinct values and their frequencies)

ex. Gend
F
F
M
F
W ← not valid.

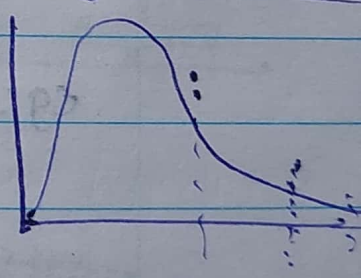
F — 10
M — 20 Likewise.

Transformation of Raw Data

- Normalization
- Data smoothing.

Sol^n

Rs 1200000
Rs 380 mill.
huge variation



Issues in Comparison.

* high variability

(i) Normalization

i) Decimal Scaling.

- move decimal point.

No : .....

Date : .....

eg: Rs 12 00 000.00

Rs 380,000,000.00 [0, 1]

$x/10^9$

0.0012

0.38

$$\boxed{V'(i) = V(i)/10^k}$$

all values should place between [0,1].

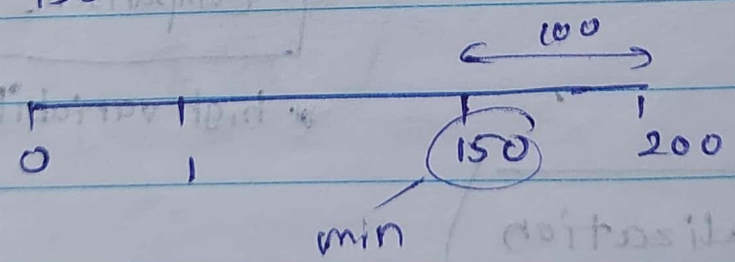
for Salary.

smallest k s.t $(|V'(i)| < 1)$

2) Min-Max normalization

$$v'(i) = (v(i) - \min(v(i)))/\max(v(i)) - \min(v(i))$$

eg: Suppose the data range for feature v is 150 to 250.



$v'(i) = (250 - 150)/100 = 1$

$(150 - 150)/150 - 150 = 0$

Date : ..... No : .....

9) Standard deviation Normalization.

$$v'(i) = (v(i) - mean(v)) / sd(v))$$

$$= \frac{(x_i - \bar{x})}{sd(v)}$$

(2) Data Smoothing

- rounding values. (closest integer rounding)

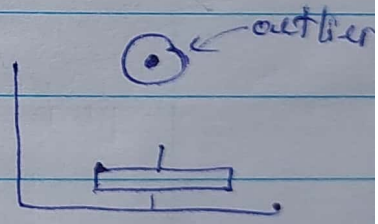
eg: $F = \{\underbrace{0.93, 1.01, 1.001}_{1}, \underbrace{3.02, 2.99}_{3}, \underbrace{5.01, 4.99}_{5}\}$

$F_{smooth} = \{1, 1, 1, 3, 3, 5, 5\}$

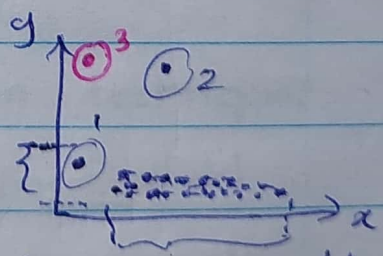
Outlier analysis.

Graphical Methods.

- Histogram.
- Box plots →



- Scatter plots. →



(3) → outlier for both x, y.

- according x variable (1) is a outlier.
- according y variable (2) is a outlier.

* helps to identify misclassifications.

Date : .....

Frequency table

- tabulated summary

USA	1
US	150
France	1
Europe	400.

misclassifiat.

- Numerical Methods.
  - using Z-score standadization.

    But, Mean and variance are sensitive to presence of outliers.

- More robust outlier detection is interquartile range.

Date : ..... No : .....

Data Pre-Processing (II)

Data Reduction Strategies

- Dimensionally reduction
  - → Wavelet transform
  - → Principal component analysis (PCA)
  - → Feature subset selection, feature creation.
- Numerosity reduction.
  - → Regression and log-linear models
  - → Histograms, clustering, sampling
  - → Data Cube aggregation.
- Data compression.

Dimensionally reduction

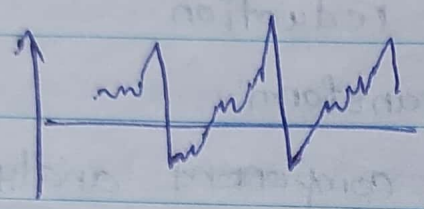
- Curse of dimensionality
- Dimensionality reduction.
- Dimensionality reduction techniques.
  - → Wavelet transforms.
  - → PCA
  - → Supervised and non linear techniques (e.g: feature selection)

PCA → linear combination of all variables ;
(it is not a original information).

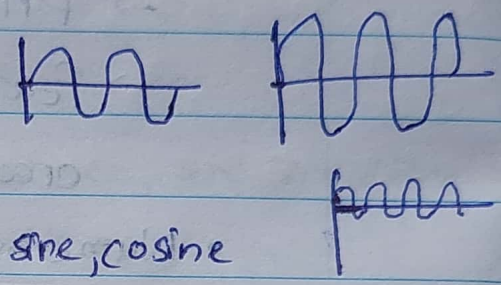
Date : ..... No : .....

Mapping data to a new space.

1) Fourier transform



decompose. → different frequency Bands.



sine, cosine

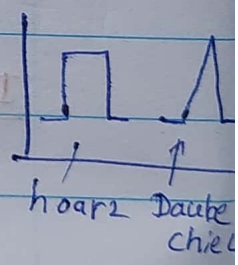
Small piece of signal

2) Wavelet transform.

Consider a window in given data and analyse those windows and move the window

→ Discrete wavelet transform (DWT)

- for linear signal processing, multi-resolution analysis.



Method

- Length, L must be in 2^n (padding with 0's when necessary)
- Each transform has 2 functions : Smoothing, difference

Date : ..... No : .....

ex: $S = [2, 2, 0, 2, 3, 5, 4, 4]$ can be transformed to $\hat{S} = [2\frac{3}{4}, -1\frac{1}{4}, \frac{1}{2}, 0, 0, -1, -1, 0]$

PCA → In order in higher contribution to lower contribution.

- We can change the shape of wave.

2 important parameters in wavelet: S - Scale, L - Location

Why wavelet transforms:

- Use hat-shape filters (band pass filters)
- Effective removal of outliers.
- Multi-resolution.
- Efficiency: Complexity $O(N)$
- Only applicable to low dimensional data.

reduce attribute

Pricipal component analysis (PCA)

$n \rightarrow P$ mappin n variables to P variables

$n >>> P$

- Normalize input data
- Compute k orthonormal (unit) vectors.

Multico → decrease accuracy of model

Date : ..... No : .....

- Each input data is a linear combination of k-principal components.
- PCA sorted in order or decreasing "significance" or strength.
- Size of data can be reduced by elemenatin weakest PCA.

Attribute Subset Selection

- Redundant attributes

  if y depend on x, y is redundant variable.

  ex: price↑ tax↑ ↓ ↓ can use one variat instead of both for analysis.

- Irrelevant attributes

  ex: stu. ID irrelavant for predicting GPA. (can remove it)

Heuristic search in attribute selection

- Best combined attribute selection and eleminatio
- Optimal branch and bound

  Best single attribute

  - use to eleminate and backtracking.
- if there is significant (high) different keep it, o/w elemin

Attribute Creation (Feature Generation)

- Three general methodologies
  - → Attribute extraction. (Extracting new attributes)
    - need Domain specific knowledge
  - → Mapping data to new space. (data reducti

    New Domain

    $P_i = a_1 x_1 + a_2 x_2$

    ex: Fourier transformation, wavelet transformation, manifold approach.

Date : ..... No : .....

→ Attribute construction

A = L × H (Area = length × height) ← Combining features

• Data discretization 0 – 10,000,000

low | Mid | high

18/09/2023

(Amount of data reduce)

Reduce volume of Data

Data Reduction 2 : Numerosity Reduction

1) Parametric method (eg: regression.)

$y = a_1x_1 + a_2x_2 + a_3x_3 \ldots + a_nx_n$

ex:
- Linear
- Multiple
- log linear

* assume regression model. (assume data fits some model).
- Estimate model parameters.
- store only the parameters
- discard the data

2) Non parametric method.

* Do not assume model.
* Major families: histograms, clustering, Sampling

Histogram analysis

partition rules

- Equal width – Equal bucket range.

0 – 10 $0 \le x < 10$

10 – 20 $10 \le x < 20$

Date : ..... No : .....

- Equal Frequencies (or equal depth)

sampling

key principle : should chouse representative sample.

* when have homogenous data → random sampling okay.

o/w it will not give representative sampling

* Therefore need adaptive sampling methods.

ex: stratified sampling.

cluster	stratified.
• don't have preidentified clusters	• clear sampling frame (predefined list of item) (differences are preidentified)

Data cube aggregation

- Summerize data

Date : ..... No : .....

Data Reduction : 3 : Data compression

* String compression (compress data into small files)
* image compression.
  - loss less compression : Recreated images are exactly same as original.
    ex : png
  - lossy compression : having minor changes.
    ex : jpg (slight loss of data), but it is not notable.

lossless → sms → • msg sent as it is (going as raw msg)

~~lossy~~ → whatsapp → • End to end encrypted msg.
(msg is change how its appeared)
* Middle person can't read original msg.

* Dimensionaly reduction → find dependent data.
* Numerosity reduction →

Discretization

* Supervised – more accurate
* Un supervised – faster.

Date : ..... No : .....

Bining

- Equi - width Bining.

Bin boundary in frequency.

ex: 0, 4, $\boxed{12, 16}$, 16, $\boxed{18, 24}$, 26, 28

Bin - Width - 10 min = 0 max = 28

Discrete data.

[0, 10) ⟹ 0, 4

[10, 20) ⟹ 12, 16, 16, 18

[20, 30) ⟶ 24, 26, 28

[0 - 9] ⟶ 0, 4

[10 - 19] ⟶ 12, 16, 16, 18

[19 - 29] ⟶ 24, 26, 28

- Equi - Frequency Binning

ex: Assume 3 elements in each bin.

min = 0 max = 28

Bin number	elements	Boundary
1	0, 4, 12	0 - 14 $\left(\frac{12+16}{2}\right)$
2	16, 16, 18	14 - 21
3	24, 26, 28	21 - 28

Date : ..... No : .....

Entropy Based Discretization

Entropy : Amount of information contains in the given dataset.

* Then we want to put them different classes.

ex: $A_1, A_2, \ldots A_n$ E $E \rightarrow E_A$

↑ use one of this attribute to partition data set.

$E \rightarrow E_A$ (Information gain).

* then select highest information gain.

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

for the target

ex:

Educatial level.

* O/L
* A/L
* Degree.

} if we want classify customers into clusters.

O-Ring Failure	
Y	N
7	17
$7/24$	$17/24$

$= 24$ (observation)

$E(\text{failure}) = E(7, 17) = E(0.29, 0.71)$

$= -0.29 \times \log_2(0.29) - 0.71 \log_2(0.71)$

$=$

$E(3,0) = E(\frac{3}{3}, \frac{0}{3}) = E(1,0)$

$E(4,17) = E(\frac{4}{21}, \frac{17}{21})$

Date ..... No: .....

Entropy for the target given a bin:

$$E(S, A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

O-ring failure

Temperature.	Y	N	
<=60	3	0	= 3
>60	4	17	= 21

one category 0, the entropy 0

$E(\text{failure}, \text{Temp}) = P(<=60) \times E(3,0) + P(>60) \times E(4,17)$

$= \frac{3}{24} \times 0 + \frac{21}{24} \times 0.7$

$E(\frac{3}{24}, \frac{0}{24})$

$= 0.615$

information gain $= E(S) - E(S,A)$

$I.g(\text{failure}, \text{Temp}) = 0.256.$

~~* The information gain for all 3 bins shows that the~~

~~the~~ another bin

Temperature.	Y	N.
<=30		
30-60		
>60		

Calculate information gain.

* Finaly select highest ~~bin~~ gain value.

chisquare → datasets are independent or not.

chisquare > critical value ⟹ reject null hypothesis

Date : .....

chiMerge:

- Assume each ~~an~~ ~~ind~~ an every individual observation as independent.

take adjecent features and calculate chisquare value.

χ^2

$\chi^2 <$ and < thresh. if this smallest can combine that individuals.

* Take adjecent features and calculate chisquare value.
* Then take small chisquare.
* If that chi-square value is less than the threshold (significance level), Merge these intervals. (Not rejecting null hypothesis)

a_1, a_2 $H_0 : a_1 = a_2$

* if no merge increase the threshold

$$E_{11} = \frac{R_1 \times C}{N}$$

	class 1	class 2
[Interval 1]	E_{11}	E_{12}
[2]	E_{21}	E_{22}

Sample | bandwidth F

0.5

1.5 $\frac{1}{3}$

7

0 → replace 0.1 (because get infinite)

then [0.75, 0.95) → can take as single class.

Date: No :

Lecture 7 & 8

Associatio Rule Minning.

When we considering supermarket senario, if the customer has # of same item in basket is summary of the association rule mining.

- Association rule mining
- Frequent item set mining
- Market basket analysis
- Link anglysis

} diff^t name of the assoeiation rule mining.

eg:- Crime scenario → phone outgoing & incoming call logs.
→ Call link analysis.

- Facebook.

Association Rule.

* A customer purchase the item A, he might get item B..

$\boxed{\text{Association Rule : } x \Rightarrow y}$ ← x implies y.

Association Rule, there are parameters need to do this.

- Support
- Confidence.

Date : ..... No : .....

$P(x \cup y)$

Support

(1) Frequency of the support $(x \cup y)$

(2) prob. that transaction contains x & y (both points 1,2 are same).

$n \rightarrow (x, y)$

N = # of total transaction.

$$\boxed{\text{support}(x,y) = n/N = P(x,y)}$$

$n \rightarrow$ support count

support frequency

support $\Rightarrow$ How many x & y are together.

$$\boxed{\text{support} = P(x \cup y)}$$

confidence

(strength of the Rule)

- Conditional prob. that a transcaction having x also contains y.

$x \Rightarrow y$

condition results

$$\boxed{\text{confidence} = P(y/x)}$$

$(A, B, C) \Rightarrow$ Combinations of rules.

$A \Rightarrow B$, $A \Rightarrow C$, $B \Rightarrow C$, $B \Rightarrow A$, $C \Rightarrow A$, $C \Rightarrow B$

- The valid rule must be satisfy the confidence and support parameters for given threshold value (Those values must be higher than the threshold old value)

Date : ..... No : .....

A, B, C, D ⟶ items how many possible rules

$A \Rightarrow B$, $A \Rightarrow C$, $A \Rightarrow D$, $B \Rightarrow C$, $B \Rightarrow D$, $C \Rightarrow D$

$B \Rightarrow A$, $C \Rightarrow A$.

K=4	ABCD.						
K=3	ABC	ABD	ACD	BCD			
K=2	AB	AC	AD	BC	BD	CD	
K=1	A	B	C	D			

$B \Rightarrow C, D$
$C \Rightarrow B, D$
$D \Rightarrow B, C$
$B, C \Rightarrow D$
$D, B \Rightarrow C$
$C, D \Rightarrow B$

$C \Rightarrow D$
$D \Rightarrow C$

(Here we can't create rules because only one variable)

Formula

item(k)	Rules	
1	0	$=(2^1-2)=0$
2	2 (only for one combination)	$=2^2-2=2$
3	6	$=2^3-2=6$
4	14	$=2^4-2=14$

↓

(2^K-2)

m = # of items

Date : ..... No : .....

$$\boxed{\text{No of rules} = \sum_{k=2}^{m} \binom{m}{k} (2^k - 2)} = \sum_{k=2}^{m} {}^mC_k (2^k - 2)$$

- **Item set** : Set of items
- **k- itemset** : An item set that contains k items

  eg: $\{A, B\} \longrightarrow$ 2- item set.
- **Frequency, support count, Count** : # of transaction, that contains the item set.
- **Frequent itemsets** : Itemsets that occurs frequently (more than minimum support.)

  eg: $\{A, B\}$ items,

  $\underbrace{\text{frequent item} > \text{minimum support}}_{\text{called frequent itemset}}$

Support and confidence - Criteria of Rules

Ex:

Transaction ID	Item Brought
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E, F

minimum support = 50%

minimum confidence = 50%.

Date : ..... No : .....

(1) $A \Rightarrow C$?

transaction which has $A, C \Rightarrow 2$.

Support $(A, C) = P(A \cup C) = 2/4 = 0.5 \times 100\%$
$= 50\%$.

total transactions

C given A

Confidence $(A, C) = P(C/A) = \frac{P(C \cup A)}{P(A)}$

$= \frac{n(A \cup C)}{n(A)}$

$= 2/3 = 0.666 \times 100\%$

$= 66.6\%$.

$A \Rightarrow C$ is a valid rule, Because it satisfy the minimum support and minimum confidence

(2) $C \Rightarrow A$?

transaction $(A, C) = 2$.

Support $(A, C) = 2/4 = 0.5 \times 100 = 50\%$.

A given C

confidence $(C \Rightarrow A) = P(A/C) = \frac{P(A \cup C)}{P(C)}$

$= \frac{2}{2} = 1$

Date : ..... No : .....

$C \Rightarrow A$ is also a valid rule.

Because support is 50% and confidence 100%, which is satisfy the minimum support and confidence.

Frequent Item sets

Transactio ID	Item Brought
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E F

min sup = 50%

min conf = 50%

support

$3/4 \leftarrow \{A\} = 111$

$2/4 \leftarrow \{B\} = 11$

$2/4 \leftarrow \{C\} = 11$

$1/4 \leftarrow \{D\} = 1$

$1/4 \leftarrow \{E\} = 1$

$1/4 \leftarrow \{F\} = 1$

$1/4 \leftarrow \{A,B\} = 1$

$3/4 \leftrightarrow \{A,C\} = 11$

$1/4 \leftarrow \{A,D\} = 1$

$1/4 \leftarrow \{B,C\} = 1$

$1/4 \leftarrow \{B,E\} = 1$

$1/4 \leftarrow \{B,F\} = 1$

$1/4 \leftarrow \{E F\} = 1$

$\{A,B,C\}$

$\{A, B D\}$

$\{B E F\}$

freq itemset	support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

> min sup (50%)

By considering all other transaction above are the frequent item set.

Because freq. item > min support

A ⇒ C

Support = Support({A,C}) = 50%

Confidence = support({A,C}) / support({A})

= 66.6%

A priori principle

Any subset of frequent item set are also frequent (Every item sets that contains a non-frequent item set are non-frequent).

eg: {A,C} is a freq. item set.

75% — {A} ⊂ {A,C}

50% — {C} ⊂ {A,C}

* if A is non frequent item, combination which include A are non-frequent.

Date

$Sup(\{A\}) \geq Sup(\{A,B\}) \geq Sup(\{A,B,C\})$

If item set add more than one item it's support is less than or equal to previous support.

(Because of that we can say reverse version of priori principle is true)

Outline of the Association Rule mining Algorithm

(1) Find frequent itemsets in lattice.

(2) Generate Rules from frequent item sets.

Algorithm for finding frequent itemsets

Pesudo code

C_k - Candidate itemset of size k.

$\{A\}, \{B\}, \{C\}, \{D\}, \{E\}$

$\{A\} = 3/4 \Rightarrow 75\%$

$\{B\} = 2/4 \Rightarrow 50\%$

$\{C\} = 2/4 \Rightarrow 50\%$

$\{D\} = 1/4 \Rightarrow 25\%$

$\{E\} = 2/4 \Rightarrow 50\%$

Candidate item set $\{A\}$ $\{B\}$ $\{C\}$ $\{E\}$

min sup count = 2 min sup = 50%

2/4 ← 0.5 transact min conf = 50%

Example

TID	Item
T001	1 3 4
T002	2 3 5
T003	1 2 3 5
T004	2 5

scanD →

C_1 candidate set

1 item set

Item set	sup
{1} II	2
{2} III	3
{3} III	3
{4} I	~~1~~ X
{5} III	3

2,3 > 2 frequent item.

↓

L_1 frequent set

Item set	sup
{1}	2
{2}	3
{3}	3
{5}	3

Creat 2-items candidate set.

first creat C_2

{1,2}
{1,3}
{1,5}
{2,3}
{2,5}
{3,5}

C_2 candidate set

Itemset	sup.
{1,2}	1 X
{1,3}	2
{1,5}	1 X
{2,3}	2
{2,5}	3
{3,5}	2

↓

L_2 frequent set

Item set	sup
{1,3}	2
{2,3}	2
{2,5}	3
{3,5}	2

create 2-item candidate set →

Item set
{2,3,5}

Date : ..... No : .....

{1, 3} , {2, 3} ⟶ × not common

{1, 3} , {2, 5} ⟶ × not common

{1, 3} , {3, 5} ⟶ × not common

{2, 3} , {2, 5} ⟶ (common) we can combine them.

{2, 3, 5}

{2, 5} , {3, 5} ⟶ × not common.

scan D ⟶

Item set	sup
{2, 3, 5}	2

⟶ L3 3-item frequen set

Item set	Sup
{2, 3, 5}	2

(k-1) frequent item set ⟶ k candidate set

Self joinig step

{2, 3}
{2, 5} ⟶ {2, 3, 5}

Prune step:

Here we consider the reverse of prior principle.

Date : ..... No : .....

$\{2, 3, 5\} \longrightarrow \{2,3\}, \{2,5\}, \{3,5\}$

3 item candidate set. 2 item frequent set.

Item set	sup
{1, 3}	2
{2, 3}	2
{2, 5}	3
{3, 5}	2

$\{1, 2, 4\} \longrightarrow \{1,2\}, \{1,4\}, \{2,4\}$

(non frequent) $\longrightarrow \{1,2,4\}$ also non frequent

* Generating Candidate Itemsets (C_k)

Transactions

$$\left.\begin{array}{l} P = \{A_1, A_2, \ldots\ldots, A_n, P\} \\ q = \{A_1, A_2, \ldots\ldots A_n, q\} \end{array}\right\} \underbrace{(A_1, A_2, \ldots A_n)}_{P, q},$$

$\{2,3\}$, $\{2,5\}$ $\longrightarrow \{2,3,5\}$

Common parts, last one of each set.

Date : ..... No : .....

common. c > d. d > e.

- $L_3 = \{abc, abd, acd, ace, bcd\}$

abcd acde. frequent item

- Self joining $L_3 * L_3$

- possible candidate set.

Item set
abcd
acde

$k-1 = L_3$

freq freq freq freq

abcd $\longrightarrow$ {abc}, {abd}, {bcd}, {acd}

frequent ite.

acde $\longrightarrow$ {acd}, {ace} {cde} {ade}

freq freq non freq non freq

$C_4 = \{abcd\} \longrightarrow$ frequent

Date : ..... No : .....

Association Rule

$A \Rightarrow B$

Antecedent — consequent

Generating Rules from frequent Item set.

Eg: $I = \{A, B, C\}$

$\underbrace{A, B}_{S}$

L.H.S R.H.S

$A, B \Rightarrow C$

$S \Rightarrow I - S$

$Sup(S) \geq Sup(I)$

Confidence value of the rule

$S \Rightarrow I - S.$

$(A \Rightarrow B) \longrightarrow \dfrac{sup(A \cup B)}{Sup(A)}$

$(S \Rightarrow I - S) \longrightarrow \dfrac{sup(S \cup (I - S))}{sup(S)}$

$\Longrightarrow \dfrac{Sup(I)}{sup(S)}$

Date : ..... No : .....

$\{A, B, C, E\} \Rightarrow \{D\} \longrightarrow \dfrac{\sup\{A, B, C, D, E\}}{\sup\{A, B, C, E\}}$

$\{B, C, E\} \Rightarrow \{A, D\} \longrightarrow \dfrac{\sup\{A, B, C, D, E\}}{\sup\{B, C, E\}}$

We can use those to find the give rule is valid or not valid (strong or not)

- S C I
- Confidence value of the rule.

If $\{A, B, C, E\} \Rightarrow \{D\}$ is conf is low.

$$\text{Conf} = \frac{\sup(I)}{\sup(A, B, C, E)} \leftarrow \text{sup} \downarrow$$

$A, B, C \Rightarrow \{D, E\}$ conf is also low.

$$\text{Conf} = \frac{\sup(I)}{\sup(A, B, C)} \leftarrow \text{sup} \uparrow$$

Date : ..... No : .....

Improving the Efficiency of Apriori

- Here we didn't want to ~~do the~~ find the candidate set.
- We used the hash-based techniques.
- Hash techiques assign the bin for the each set.

eg: Histogram.

(we get moderate bines)

* disadvantage of Apriori principle
  - We have to find candidate item set. (intermediate set)

Advantage of Hash

- Reduce the # of scans.
- Remove the large candidate that cause high input/output cost

remove from transaction table also.

	items
T001	1 3 ~~4~~
T002	2 3 5

{1,3} {1,4} {3,4}

{2,3} {2,5}, {3,5}

Date :

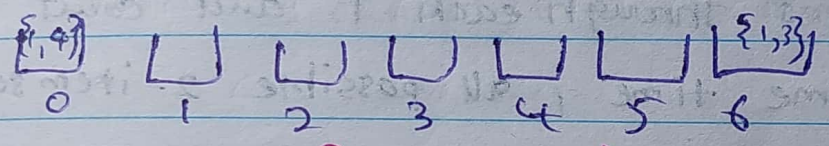
Hash based Techniques.

- reduce the # of k item sets.
- Define hash fun^n for each k-itemsets.

ex: hash function $h(x,y) = ((order_x * 10) + order_y) \% 7$.

{1,3}

$h(1,3) = [(1 \times 10) + 3] \% 7 = 13 \% 7 = 6$



(2-itemset)

1,3,4 ⟹ {1,3}, {1,4}, {3,4}

1,2,3,5 ⟹ {1,2} {1,3}, {1,5}, {2,3}, {2,5} {3,5}

(1-item set)

itemset	sup
{1}	2
{2}	3
{3}	3
~~{4}~~	1
{5}	3

remove

min sup count = 2.

if (min sup count = 2)

* Three item hash function.

hash function $h(x, y, z) = ((order_x * 100) + (order_y * 10) + order_z) \% Z$.

{1,2,3,5} ⟹ {1,2,3}, {1,2,5}, {1,3,5}, {2,3,5}

- if have non frequent ~~trans~~ item can remove whole transaction as well

Date : ..... No : .....

~~ABCD~~ $\boxed{4 \rightarrow \text{transactions.}}$

1 → create frequency table

2 → hash(x, y)

3 → hash(x, y, z)

Direct hashing and Pruning (DHP.

- 3 parts.

<u>part 1</u>

- Goes through each T and count 1-itemset
- Same time, all possible 2-item sets are hashed to hash table.
- Bucket count increase by 1 when new item added.
- if bucket count ≥ min sup count ⟹ Bit vector = 1

  else

  Bit vector = 0.

<u>part 2</u>

ex: C_2 ← Candidate items.

{1,4} — I {2,5} — III

{3,5} — II {1,3} — II

{2,3} — II {3,4} — I

Date : .....

No : .....

Phase 1: C_k is generated using hash table.

pase 2 : • Count support for each itemset in C_k

• Also, hash table for next step is generated.

• Pruning

(Any transaction that doesn't contain any frequent k-item sets cannot contain any frequent (k+1) item sets.

2 item set hash table

using Transaction table.

C_2			L_2
{1,4}	I	1 ✗	{3,5}
{3,5}	II	2	{2,3}
{2,3}	II	2	{2,5}
{2,5}	III	3	{1,3}
{1,3}	II	2	
{3,4}	I	1 ✗	

using 3 itemset hash table.

C_3			L_2
{2,3,5}	II	2	{2,3,5}
{1,2,3}	I	1 ✗	

Date : ..... No : .....

part 3

- Continous until there are no more candidate itemsets.
- Database is now significantly smaller because of pruning.

2023/10/02

Frequent Pattern Growth (FP-Growth) Algorithm

- Generates frequents pattern, without generating candidate key generation.

- Trying to minimize scan.

Step 1: Build FP tree.

Step 2: Extracts frequent item.

Example

transaction table.

TID	items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

Date : ..... No : .....

Frequent table.

min sup = 2.

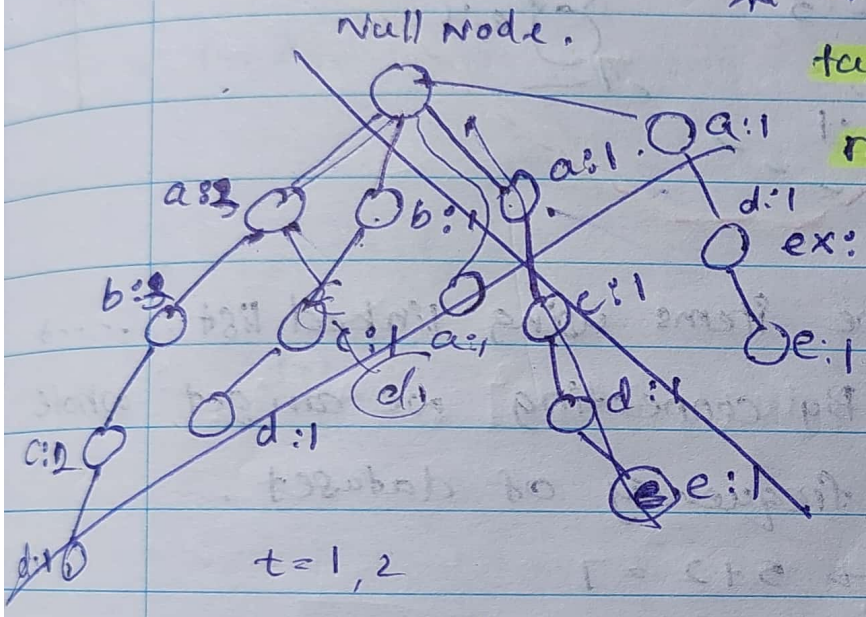
a	~~IIII~~ III	8
b	~~IIII~~ II	7
c	~~IIII~~ I	6
d	~~IIII~~	5
e	III	3

decending order of frequencies. (write it as decending order)

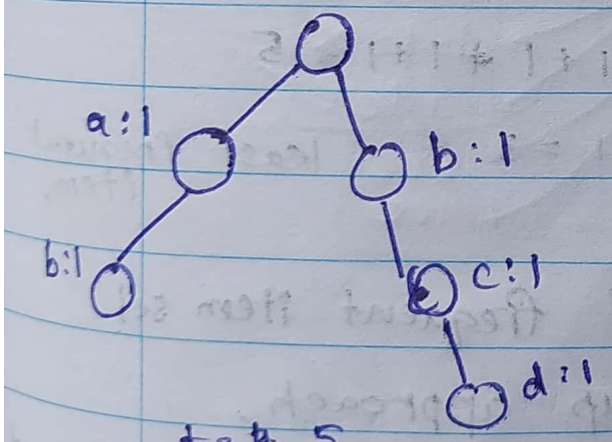
then again transaction table items should rearrange in decending order.

ex: if a → 8, b → 10 transactio table. 1 {b, a}

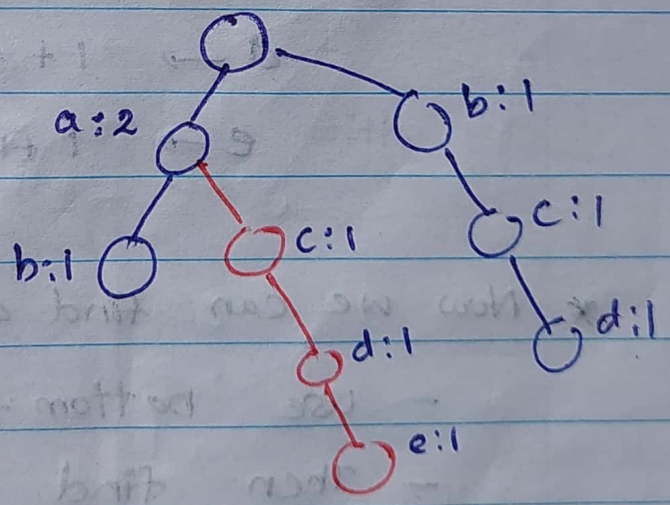
(Here example no need)



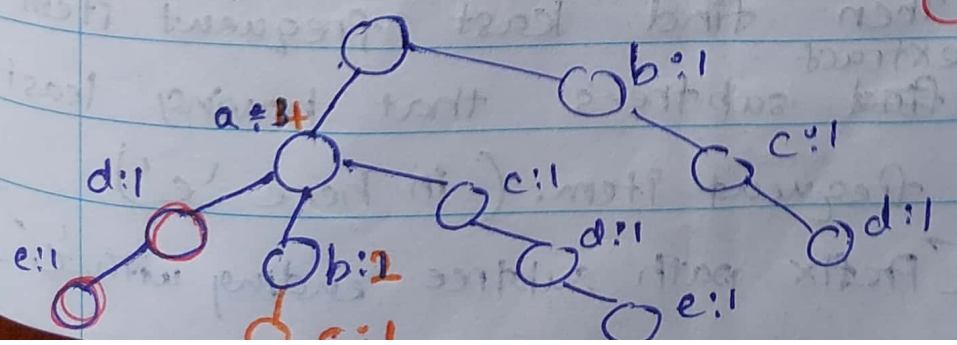
t = 1, 2



t = 3



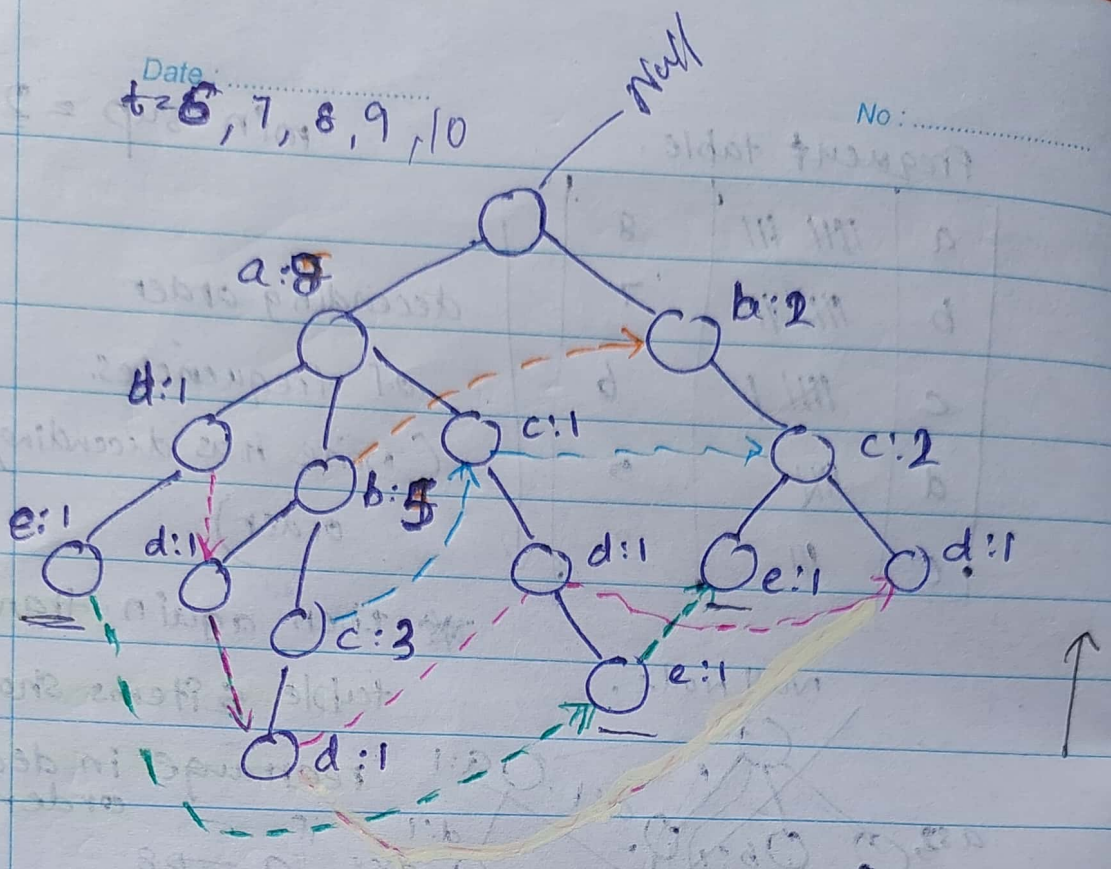
t = 4, 5



Date:

t=6, 7, 8, 9, 10

No:



- Connect some items using linked list: ---->

purpose: By connecting we can get whole frequency of dataset.

eg: b → 5+2 = 7

c → 3+1+2 = 6

d → 1+1+1+1+1 = 5

e → 1+1+1 = 3 ← least frequent item.

* Now we can find all frequent item set

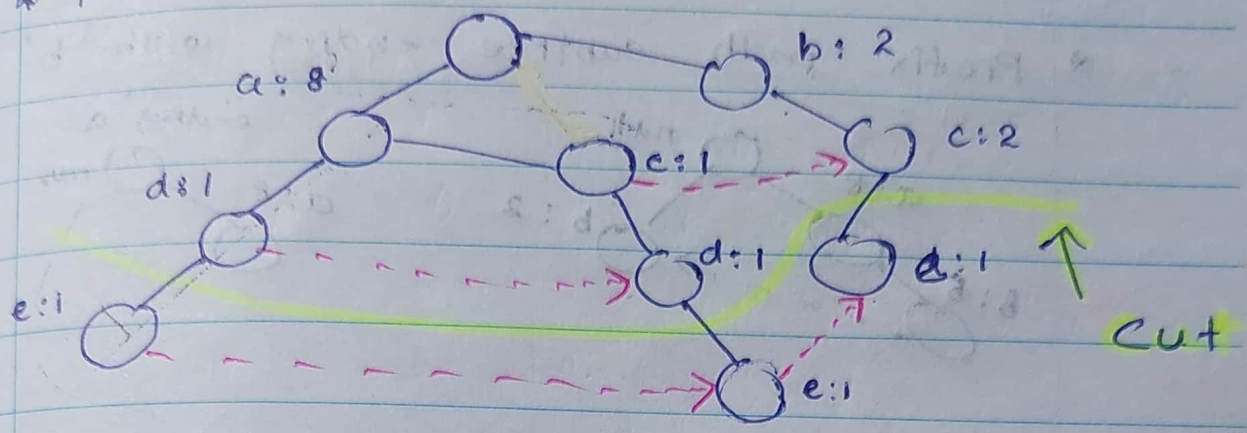
- use bottom-up approach.
- Then find least frequent item and extract ~~find~~ subtrees that having least frequent item. (in here 'e')

[Prefix path subtree ending with 'e']

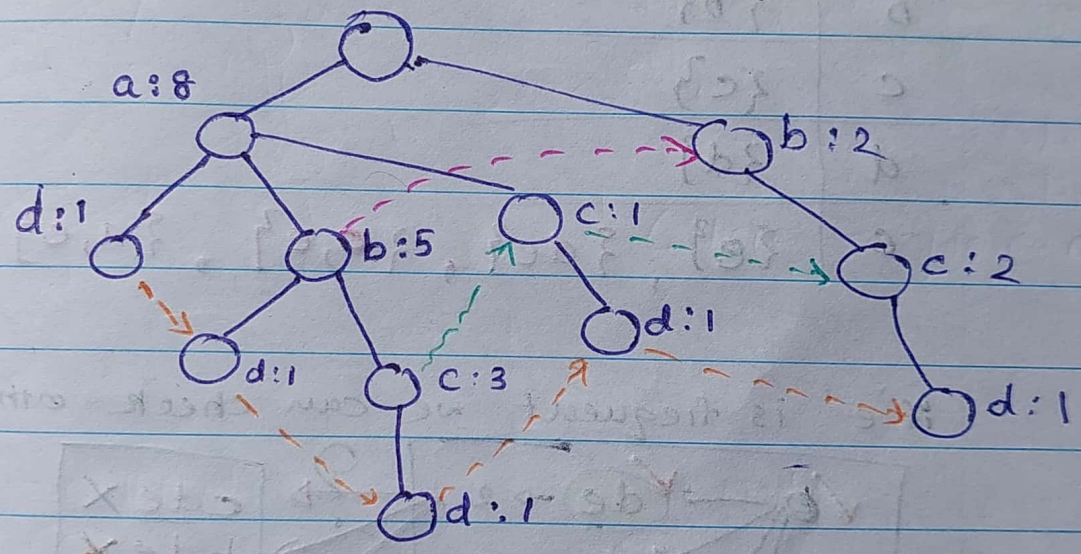
Start with least

Date ..... No .....

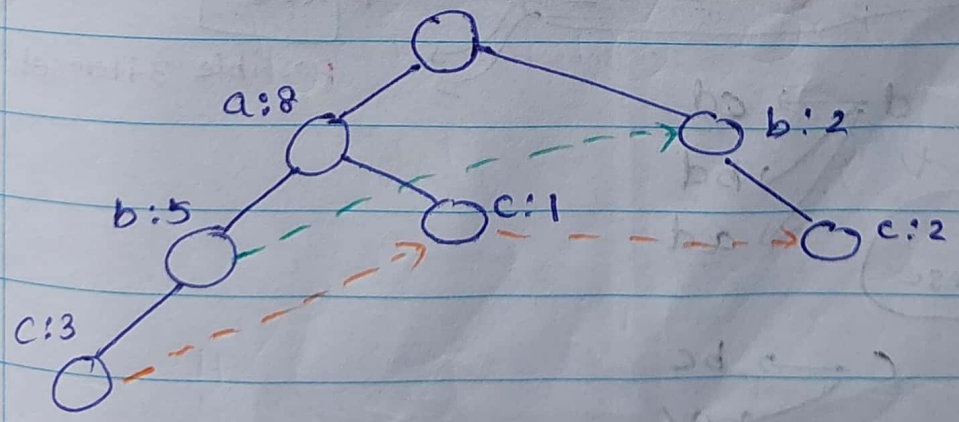
* Prefix path subtree ending with 'e'



* Prefix path subtree ending with 'd'

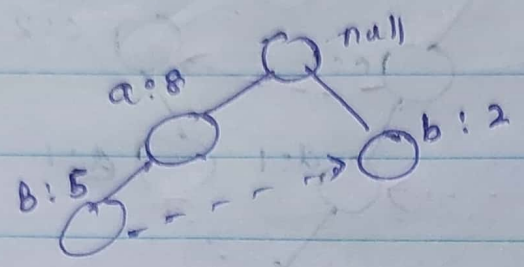


* Prefix path subtree ending with 'c'

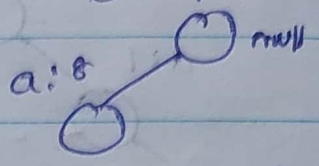


Date : ..... No : .....

* Prefix path subtree ending with 'b'



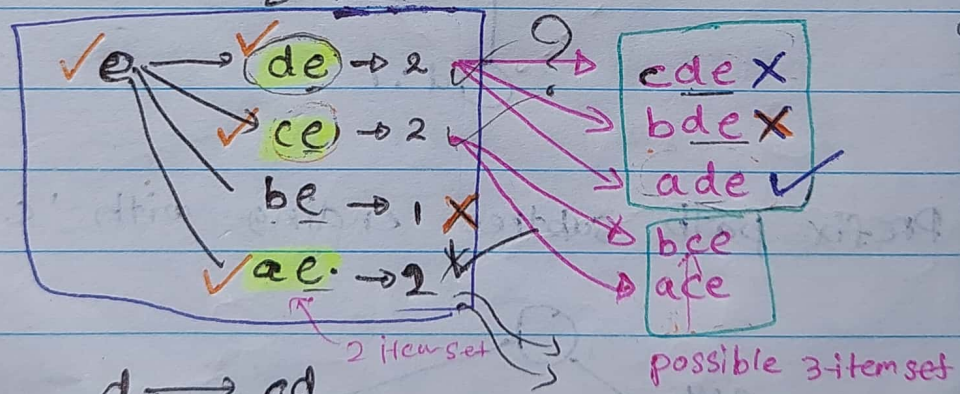
ending 'a'



frequent item

a	{a}
b	{b}
c	{c}
d	{d}
e	{e}, {de}, {a,e}, {a,e}, {a,d,e}

if 'e' is frequent we can check other possibilit
Combinat



use prefix path subtree to find these.

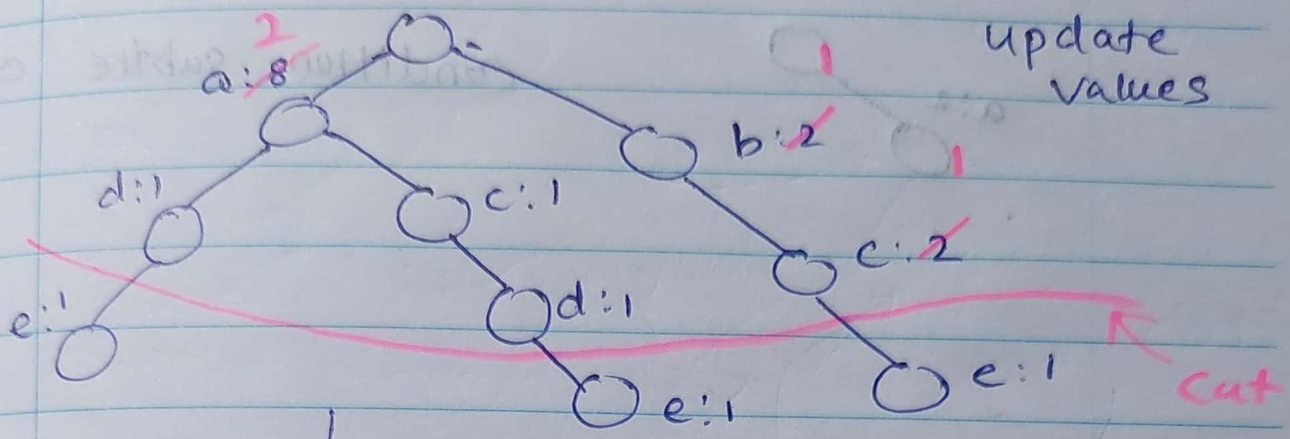
d → cd
→ bd
→ ad

c → bc
→ ac

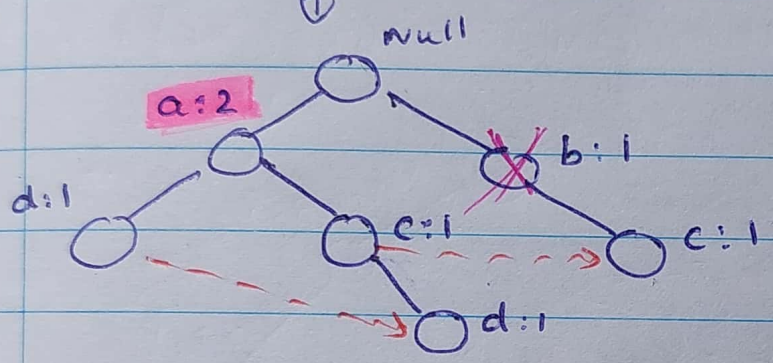
b → ab

Inside prefix subtre 'e'

Date : ..... No : .....



update values

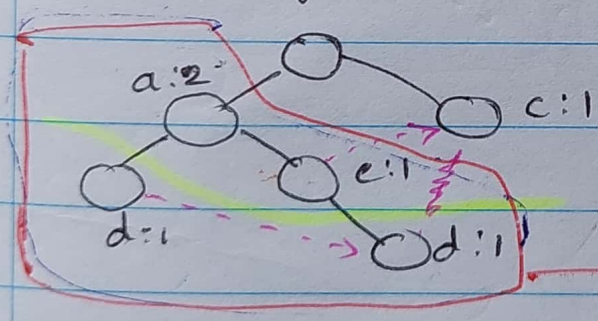


conditional subtree on 'e'

- we remove 'e' then found that d frequency is 2.

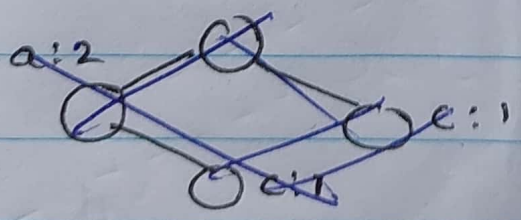
de → 2 , ce → 2 , be → 1 , ae → 2 that mean 'de' freq is 2.

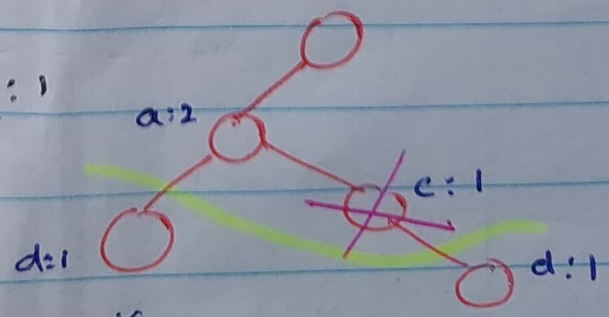
non freq. item remove



select that subtree which is ending 'de' only.

* Prefix subtree ending 'de'



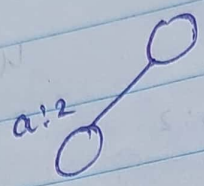


cde → 1 ✗ bde → 0 ✗ ade → 2 ✓

Date :

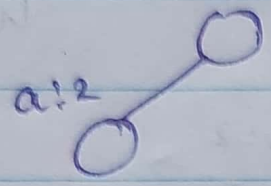
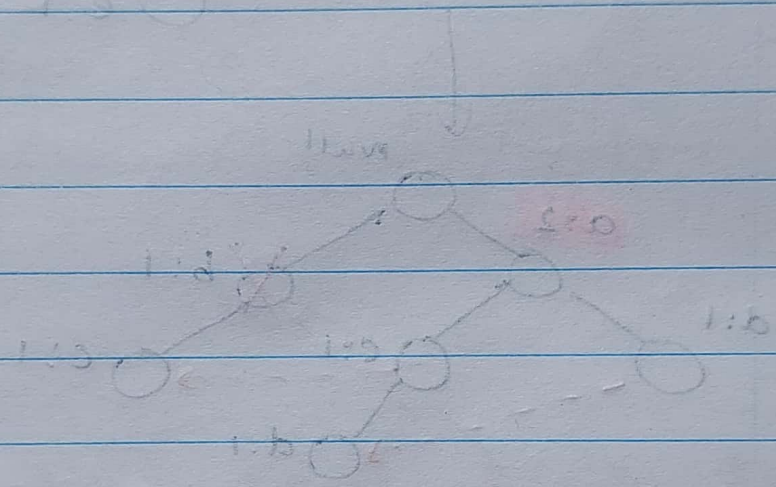
No :

condition subtre on



use
prefi
path
to

Date : ..... No : .....



condition subtre on 'de'