DATA MINING & ANALYSIS:

★ Course Content :

Unit I - • Data Preprocessing (08 hrs)
• Frequent Pattern Mining (08 hrs)

Unit II - • Classification techniques (08 hrs)
• Cluster Analysis (08 hrs)

Unit III - • Advanced Mining Techniques. (08 hrs)

CHAPTER-1

DATA PREPROCESSING

★ KDD Process -

Data → • Data Integration → Data warehouse → Data Selection
• Data Cleaning

• Data Warehouse:
It is a repository of information collected from multiple sources, stored under a unified schema.

• Steps in KDD process — Data Mining (+1) Points
Reduction (+1)

* APPLICATIONS OF DATA MINING:

a). Retail Industry:- Analyze Buying Patterns of Customers.

b). Banking Sector :- Fraud detection, Customer Retention

c). Healthcare Industry :- Effective treatments, Best practices, Fraud detection

d). Social & Information Networks :- New program/ new products / processes or services.

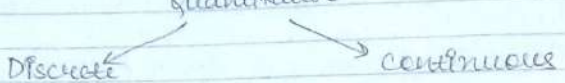
Φ Data Transformation → Model Building (Data Mining) → Model Evaluation

Knowledge Presentation

★ DATA TYPES:

Types of Data : a). Quantitative (Numerical)
b). Qualitative (Categorical)

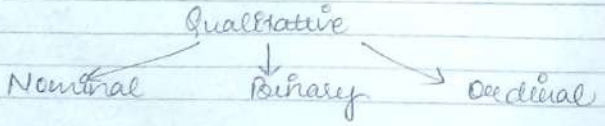
① Quantitative :-



★ Discrete
- takes integer values.
  E.g: no. of people
- They represent the count.
- Represented using Bar charts & Histograms.

★ Continuous
- takes real nos. (can take ∞ no. of values)
  E.g: ht. of person, distance covered.
- Represented using Histograms & Scatter plots.

② Qualitative :-



★ Nominal : words or letters
- variables are labeled using symbols.
  E.g: name of a person, city, color of eyes, etc.
- Represented using Pie charts usually.

★ Ordinal :
- A well-defined order among categories.
  E.g: size → small, medium, large
  grades
  income
- Represented using bar graphs/bar charts.

★ Binary :
- takes only 2 values.
  E.g: pass/fail, male/female, +ve/-ve.
- Represented using Pie charts.

Quiz Select categorical data types—
- Color of a shirt.
- Finish order in a race.

Ex - Gender (Categorical, binary)
Age (Categorical, ordinal, when young, senior...)
Income/Month (Numerical, continuous)
Item purchased (Categorical, nominal)
Quantity (Numerical, Discrete).

* DATA PREPROCESSING : INTRODUCTION

- Why Data Preprocessing is required?
  -Raw (dirty) data is not adequate for analyzing.

  Reasons : • Quality
  • Dimensionality } Too many attributes

- Preprocessing of data assess the quality of the data. There are certain measures to assess quality of data.

  Measures :
  - Completeness, Consistency
  - Timeliness, Accuracy
  - Believability
  - Interpretability

1). Completeness :
-Not recorded, Missing Values

2). Consistency :
~~-Discrepancy~~ Discrepancy in code or names.

3). Accuracy :
-degree to which recorded/noted value of data is correct or wrong.

4). Timeliness : Indicates timely availability (updated, most recent data).

5). Interpretability : It indicates the ease with user can understand the data. E.g values, meanig, info. about that etc.

6). Believability : Indicates the trustworthiness of the data. How trustable the data are correct?

* Preprocessing Techniques :-

Major Tasks in Data preprocessing :—
- Data cleaning
- Data Integration
- Data Reduction
- Data Transformation

1). Data cleaning : fills in the missing values & smoothes the noisy data.

2). Data Integration : It integrates the data by dissolving the issues during integration.

3). Data Reduction : It reduces the dimensionality of the data.

4). Data Transformation : It transforms the set of values of attributes to new set of values of attributes whenever required.

* DATA CLEANSING:

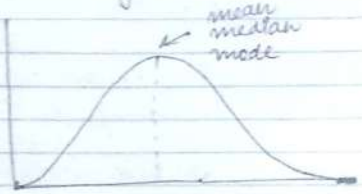
- Incomplete Data (Missing Values)
- Noisy Data

* Techniques to handle missing values:

1). **Ignore the tuple**: deleting those tuples whose values are missing.

2). Use a global constant: Fill in the missing values with either $-\infty$ or ~~+∞~~ or unknown.

3). Use central tendency: the missing values are filled with (mean), when the attributes are normaly distributed this way.



Symmetric data

If attribute values are either +vely skewed or -vely skewed, then we replace the missing values with (median)

- If the missing attribute is a categorical data, then that value is filled with the most frequent occurring value in the column. i.e. (mode)

> mean: normally distributed
> median: skewed
> mode: categorical

4). Fill in the missing values manually.

5). Use most probable value:
- Regression
- Decision Tree

Quiz: Techniques to handle missing values:
- Regression.
- Central Tendency of the attribute.

* NOISY DATA

Noise is a random error or variance in a measured variable.

E.g: Salary → −10 5K.

Techniques to handle Noisy Data:
1). Binning
2). Regression
3). Outlier Analysis

1). Binning:

a) Bin Means Method
b) Bin Median Method
c) Bin Boundary Method

E.g: Data for price (in dollars): 7, 3, 19, 14, 23, 31, 24, 38, 33.

a).

Step(1)- Sort the data in ascending order.

3, 7, 14, 19, 23, 24, 31, 33, 38.

Step(2)- Distribute them equally into bins.

Bin 1: [3, 7, 14] Bin 2: [19, 23, 24] Bin 3: [31, 33, 38]

mean of 1st bin : 8

(3): Replace all the values of the bin with its mean.

∴ Bin 1: [8, 8, 8] Bin 2: [22, 22, 22] Bin 3: [34, 34, 34]

This is Bin Means Method. →

b). Bin Median Method:

In this method, all the data in the bin with the median.

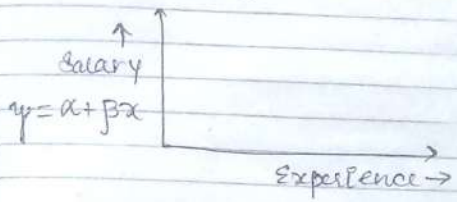
∴ Bin 1: [7, 7, 7]
Bin 2: [23, 23, 23]
Bin 3: [33, 33, 33]

c). Bin Boundary Method:

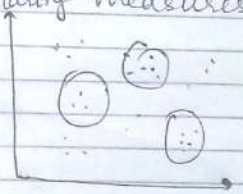
The intermediate value is replaced by its closest boundary.

∴ Bin 1: [3, 3, 14] Bin 2: [19, 24, 24] Bin 3: [31, 31, 38]

2). Regression: Data can be smoothed by fitting the data to a regression function.



3). Outlier Analysis: Uses clustering techniques, where in we group the data based on the similarity measures. The objects outside the clusters are called noise & are eliminated.



2). Non-Parametric Methods:

a). **Histograms:** It reduces the data by storing the average (sum) for each bin.

Bin is a range of values or a single value & the occurrence will be shown.

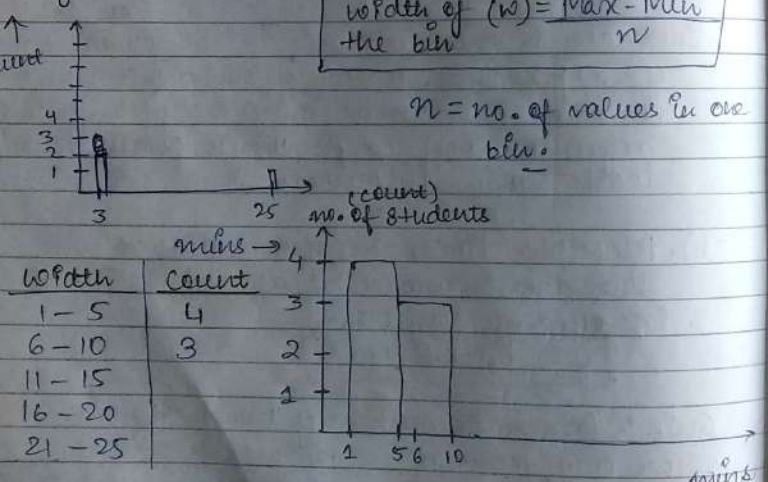
E.g: The given data is the time (in mins) given by students for watching the video.

3, 4, 7, 11, 18, 3, 19, 22, 23, 25, 2, 10, 21, 14, 8, 11, 17, 23, 23, 14.

Singleton Bucket.

$$\text{Width of the bin } (w) = \frac{\text{Max} - \text{Min}}{n}$$

n = no. of values in one bin.



Width	Count
1 – 5	4
6 – 10	3
11 – 15	
16 – 20	
21 – 25	

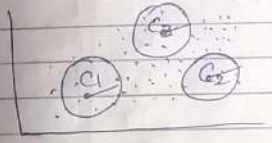
b). **Clustering:** unsupervised method (data mining technique).

- Store cluster representation (E.g. centroid & diameter) only.

The no. of centroids to keep, depends on the no. of clusters. If we want 2 clusters, we will be having 2 centroids, etc.

Let $x_1 = (3,4)$, $x_2 = (5,6)$, $x_3 = (10,12)$ . . . . x_n

find centroids c_1, c_2, c_3 if we want 3 clusters



Each circle will be having diameter.

So, Instead of storing the entire data we can store centroids & diameters only.

c_1 d_1
c_2 d_2
c_3 d_3

Quiz: Clustering in Data Reduction:-

- Partitions data into clusters
- Store only cluster representations (Centroid & Diameter).

c). Sampling:

- Simple Random Sampling (SRS)
  Samples are taken randomly i.e. every tuppe has equal probability of getting selected.
  - SRS without replacement (SRSWOR)
  - SRS with replacement (SRSWR)
  - Stratified Sampling

* SRSWOR:

Let we want two samples of size 3.

S_1	S_2
T_2	T_1
T_4	T_7
T_6	T_9

Here, tuppes are not repeated.

T_1	L
T_2	M
T_3	L
T_4	L
T_5	M
T_6	L
T_7	H
T_8	H
T_9	H
T_{10}	H

* SRSWR:

S_1	S_2
T_2	T_2
T_4	T_7
T_6	T_9

Tuples can be repeated.

T_2, T_4, T_6 are randomly chosen for sample S_1. However, they are not taken out from the tuppes set i.e. they can be again selected for sample S_2.

* Stratified Sampling:

Let us suppose 'Income' attribute has nominal values i.e. Low (L), Medium (M) & High (H).

1st strata: T_1 L, T_3 L, T_4 L, T_6 L

2nd strata: T_2 M, T_5 M

3rd strata: T_7 H, T_8 H, T_9 H, T_{10} H

Now, let us suppose that we need to select 50% from each strata.

From

∴ 1st strata → T_3, T_4
2nd strata → T_5
3rd strata → T_8, T_9

} Stratified Sample (S.S)

- Now, we see that let it be SRSWOR, SRSWR or S.S. the population size is reduced based on the sample size.
- It reduces the data by selecting the data based on the sample size.

Quiz: A selected object is not removed from popul
What type of sampling it is?
SRS with replacement. (SRSWR)

★ DATA TRANSFORMATION:-

- It is a pre processing technique in which attribute values are mapped to a new set of values i.e. replaced by a new value.
- It is required because it leads the data mining process into a more efficient one and also the patterns obtained will be more easy to understand.

Data Transformation Techniques:

- Smoothing
- Normalization
- Concept Hierarchy Generation
- Aggregation
- Discretization

a). Smoothing:

Bin Means Method:

$\frac{3+7+2}{3}$ [3, 7, 2]

$= 4$ [4, 4, 4].

b). Aggregation: summarize the attribute values into one.

	Sales			
J	100	$Q_1 = 600$	Half-yearly	Yearly.
F	200			
M	300			
A	400	$Q_2 =$		
M				
:		$Q_3 =$	Half yearly	
D		$Q_4 =$		

c). Normalization: gives / assigns equal weights to all the attributes.

Normalization Methods:
- Min-Max Normalization
- Z-Score normalization
- Normalization by decimal scaling

★ Min-Max Normalization:-

maps min value to '0' & maximum value to '1' & other values lying b/w will be mapped to any value b/w 0 to 1.

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

(1) (2)

Eg: Income (Before Normalization):-

12,222 → $v' = \frac{(12,222 - 12,222)}{(85000 - 12,222)} \frac{(1-0)}{} = 0$

85,000 → $v' = \frac{(85000 - 12222)}{(85000 - 12222)} \frac{(1-0)}{} = 1$

15,000 → $v' = \frac{(15,000 - 12,222)}{(85000 - 12,222)} \frac{(1-0)}{} = 0.038$

Quiz: Which Normalization is useful when min & max values of att are unknown?
→ Z-score Normalization Technique.

* **Z-score Normalization**: This method is used when the min & the max value of the attribute is not known.

$$v' = \frac{v - \mu_A}{\sigma_A}$$

• μ_A = mean of att. A
• σ_A = std. dev. of A

E.g

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

$$= 1.225$$

(v)

• If value of attribute = mean of att. then v' (value of att) will become 0.

• If v > mean of att. then v' = +ve.

• If v < mean of att. then v' = -ve.

* **Decimal Scaling Method**:
• moves the decimal points of the att. value
• depends on the maximum absolute value of A.

$$v' = v / 10^j$$

j = no. of digits in v.

E.g

v = -986	v = 917
$v' = \frac{-986}{10^3} = -0.986$	$v' = \frac{917}{10^3} = 0.917$

d). **Discretization**: The data values are devided into intervals & labelling them.

E.g: age : 20 80

20..35 (young) | 36...55 (middle) | 56...80 (senior)

• This can be done by equal frequency or equal width method.

• Thus, interval labels replace actual data values.

e). **Concept Hierarchy Generation**: Replaces low level concepts by higher level concepts.

country
↑
state
↑
city
↑
street

Here, street can be replaced by the name of the country only

CHAPTER 2
FREQUENT PATTERN MINING

* Video 1: Basic Concepts - Frequent Itemsets, Support & Confidence

* Frequent Pattern Mining (Association rule mining): It is a process of finding frequent patterns i.e. the relevant hidden pattern from a database E.g transactional database, relational database or any other repositories working together.

* Frequent Patterns: They are the item sets, sub sequences or substructures that occur together.

- Frequent item sets means, the item sets that are purchased together by the customer.
- Sub sequence → DNA Analysis
- Substructures — Protein Structure Analysis
- They represent important properties of the data.

* Frequent Pattern Mining
  - Finding frequent patterns
  - Deriving Association rules

Support & Confidence:

P.T.O

SUPPORT & CONFIDENCE

Tid	Items Purchased	occurence
1	Milk, Bread, Coffee	• Milk : 3 • Egg : 1
2	Milk, Coffee	• Bread : 2
3	Milk, Butter	• Coffee : 2
4	Bread, Egg	• Butter : 1

- We say item is frequent, when the occurrence of that item ≥ threshold. This threshold is known as Support.

E.g: Min_sup = 2 (ut)
∴ frequent items are milk, bread, coffee.
Item set:
milk, bread : 1
milk, coffee : 2 ✓ (frequent item set).

∴ Association rule can be set up for frequent item set i.e. for milk, coffee
milk ⇒ coffee

Now, support (milk, coffee)
support (milk ∪ coffee) = 2 = 50% = 1/4
$ How many milk & coffee are occuring together in the transaction ?.

$$\text{Confidence} = \frac{S(\text{milk} \cup \text{coffee})}{S(\text{milk})} = \frac{2}{3} = 66.6\%$$

- Support = 50% shows that the 50% of items milk & coffee are purchased together.
- Confidence = 66.66% indicates that 66% customer who purchased milk, also purchased coffee.

P.T.O

∴ Association Rule : milk ⇒ Coffee (s=2, c=66.6%)

Quiz: What do you mean by support (A)?

Support(A) = No. of transactions containing A / Total no. of transactions

* Video 2: Apriori Algorithm :-

* APRIORI ALGORITHM (Frequent Item Set Mining method)

* Apriori Property :-

1) If an itemset is frequent, then all of its subsets must also be frequent.
E.g AB, If AB is frequent then {A}, {B}, {AB} all are frequent.

2) If an itemset is infrequent, then all of its supersets must be infrequent too.
E.g If {A} or {B} is infrequent, then their superset i.e {AB} will also be infrequent.

Apriori Algorithm :

It uses a 'levelwise search approach' in which 'k' items sets are used to find out 'k+1' frequent item sets.
→ Join step
→ Prune step

Levelwise Search :- 2 steps

Join step: C_k is generated by joining L_{k-1} with itself.

Prune step: Any (k-1)-item set that is not frequent cannot be a subset of a frequent k-item set.

E.g :

Tid	Items Purchased
1	Milk, Bread, coffee
2	Butter, Bread, Egg
3	Milk, Butter, Bread, Egg
4	Butter, Egg

Min-sup = 2

Step 1 : Generate C_1 & frequent itemset L_1

C_1

Itemset	Support count
M	2
Br	3
C	1
Bu	3
E	3

⇒ Prune step → (≤2) ∴ Ignored

L_1

Itemset	Sup count
M	2
Br	3
Bu	3
E	3

by joining L_1 with L_1

Step 2 : Generate C_2 & 2 frequent itemset L_2 ∧

C_2 Itemset	Freq. count
milk, bread	2
milk, Butter	1 → pruned
milk, Egg	1 → pruned
bread, butter	2
bread, egg	2
butter, egg	3

L_2 Itemset	Sup-Count
milk, bread	2
bread, butter	2
bread, Egg	2
butter, Egg	3

Step 3 : Generate C_3 & 3 freq Itemset L_3 by joining L_2 with L_2.

C_3 =

Itemset	Sup-count
M, Br, Bu	Infrequent (∵ (M, Br) (M, Bu) ~~(Br, Bu)~~ is Infrequent
~~M, Br, Egg~~	Infrequent.
Br, Bu, Egg	

↳ First Apply 2nd Apriori Property.

∴ C_3

Item_set	sup-count
Br, Bu, Egg	2 ✓

→ L_3

Item_set	Sup-Count
Br, Bu, Egg	2

Step 4 : Generate C_4 by joining L_3 with L_3.

$C_4 = \phi$ (∵ in L_3 no 4 data sets)

∴ Frequent Item Set = {Br, Bu, Eg} with Sup-count = 2

Quiz : Let the size 2-frequent item sets obtained from Apriori Algorithm be {A, B}, {A, C}, {A, D}, {B, C}, {B, E} and {C, E}. Select all the potential size-3 item sets from the following :

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

WEEK-3

* Video 1: Closed & Maximal Frequent Itemset

* **Downward Closure Property:** Any subset of a frequent itemset must also be frequent

{A B C} ⟶ {A} {B} {C} {A B} {B C} {C A}

{A, B, C}

A, B, C
N=3 ⟶ 7.

- For N distinct items, frequent itemset will be $2^N - 1$.

* Closed & Maximal Frequent Itemsets:

MAXIMAL FREQUENT ITEMSET:

T_id	Items Purchased
1	Milk, Bread, Coffee
2	Butter, Bread, Egg
3	Milk, Butter, Bread, Egg
4	Butter, Egg.

Using Apriori Algorithm, we get: Min. Support = 2

Step 1: 1 Frequent Itemset

Itemset	Sup. count
{M}	2
{Br}	3
{Bu}	3
{Eg}	3

Step 2: 2 Freq Itemset

Itemset	Sup. count
{M, Br}	2
{Br, Bu}	2
{Br, Eg}	2
{Bu, Eg}	3

Step 3: 3 Freq Itemset

Itemset	Sup. count
{Br, Bu, Eg}	2
{M, Br, Eg}	1
{M, Bu, Eg}	1

• An Itemset is called as maximal frequent Itemset if, its supersets are infrequent.
Eg: {Br, Bu, Eg} ← maximal freq. Itemset

Drawback
Maximal freq Itemsets do not have any information about the support count of the subsets.
Because of this it may give some redundant information.

* CLOSED ITEM SET: A frequent Itemset is also known as closed frequent Item set if its superset doesn't have same support count as of it, or greater sup count.

superset
{M}: 2 → {M, Br}: 2
∴ not a closed freq. Itemset.

{Br}: 3 (closed) → {M, Br}: 2 → Maximal, ∵ superset is infrequent
∴ not closed.
{Br, Bu}: 2
{Br, Eg}: 2

{Eg}: 3 → {Br, Eg}: 2 ∴ Not closed because of {Bu, Eg}: 3
{Bu, Eg}: 3

NOTE:
A maximal frequent Itemset can be a closed one, but closed cannot be a maximal one.

• Thus, there exists a relationship bw frequent, closed frequent & maximal freq Itemset.

∴ [maximal, C closed C freq]
(subset)

* Video 2: Generating Association Rules from frequent k-Itemsets

Steps:
1). For each frequent Itemset l, generate all non empty subsets of l. ∴ For {Br, Bu, Eg}:-
∴ {Br, Bu, Eg} {Br, Bu} {Bu, Eg} {Br, Eg}
{Br} {Bu} {Eg}

2). For every non-empty subset s of l, output the rule "s → (l-s)" if: sc(l) / sc(s) ≥ min-confidence
(sc = support count)

• Rule 1: {Br, Bu} ⇒ Eg // The customer who purchase Br, Bu also purchase Eg.
sup-count (Br, Bu, Eg) / sup-count (Br, Bu) = 2/2 × 100 = 100%

Let us specify min-confidence = 70%.
∴ This rule is a Strong Rule.

Rule 2: {Bu, Eg} ⇒ {Bu}

$\frac{\text{Sup. count}(Bo, Bu, Eg)}{\text{Sup. count}(Bu, Eg)} = \frac{2}{3} \times 100 = 66.66\%$

∴ Not Strong / Not Interesting rule.

* Video 3: FP Growth Algorithm :-

↳ Frequent Item Set Mining Method

* Drawbacks of Apriori :-

- Generates large candidate itemsets if itemset in the database is large.
- Needs multiple scans of database to know the sup-count.
- So, the problem with Apriori is space & time. These drawbacks can be overcome by FP GROWTH ALGORITHM.

FP GROWTH ALGORITHM :-

(Divide & Conquer Strategy)

→ Adv: generates only Freq Itemsets & not the Infrequent Itemsets.

- It compresses the database into a frequent pattern tree.
- Divides the compressed database into a set of conditional databases.
- Each conditional database is associated with one frequent item.

E.g.: min-sup = 3

T-id	Items Purchased
1	I1, I2, I3
2	I2, I3, I4
3	I4, I5
4	I1, I2, I4
5	I1, I2, I3, I5
6	I1, I2, I3, I4

Algorithm :-

- First, scans the database to found sup-count of each item.
- Then, list them in descending order by eliminating the items whose sup-count is less than min_sup-count.

∴ {I2 : 5, I1 : 4, I3 : 4, I4 : 4}

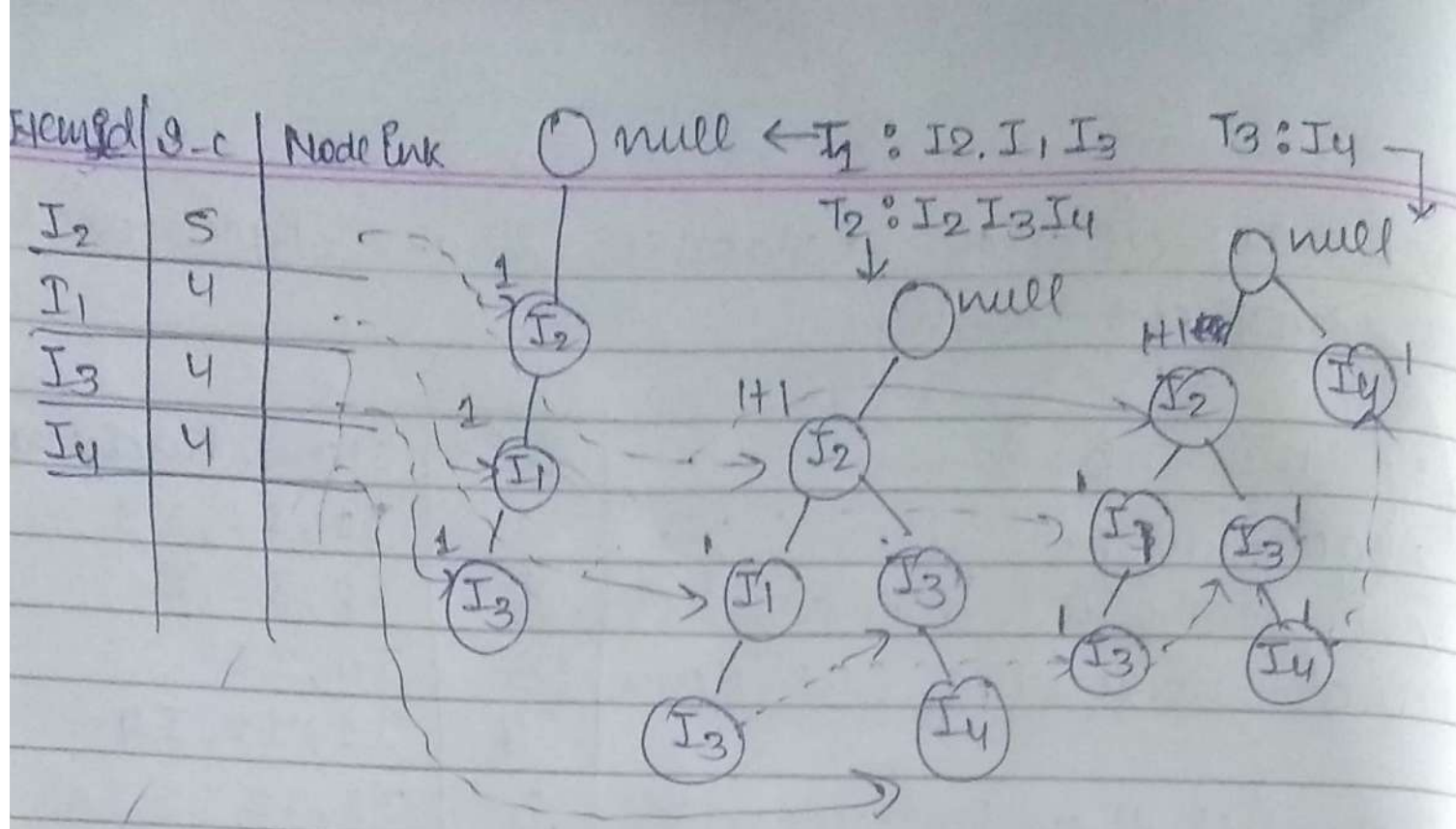
- Now, algorithm will reorganise the table :-

Tid	Items Purchased	Items_Purchased (sorted)
1	I1, I2, I3	I2, I1, I3
2	I2, I3, I4	I2, I3, I4
3	I4, I5	I4
4	I1, I2, I4	I2, I1, I4
5	I1, I2, I3, I5	I2, I1, I3
6	I1, I2, I3, I4	I2, I1, I3, I4

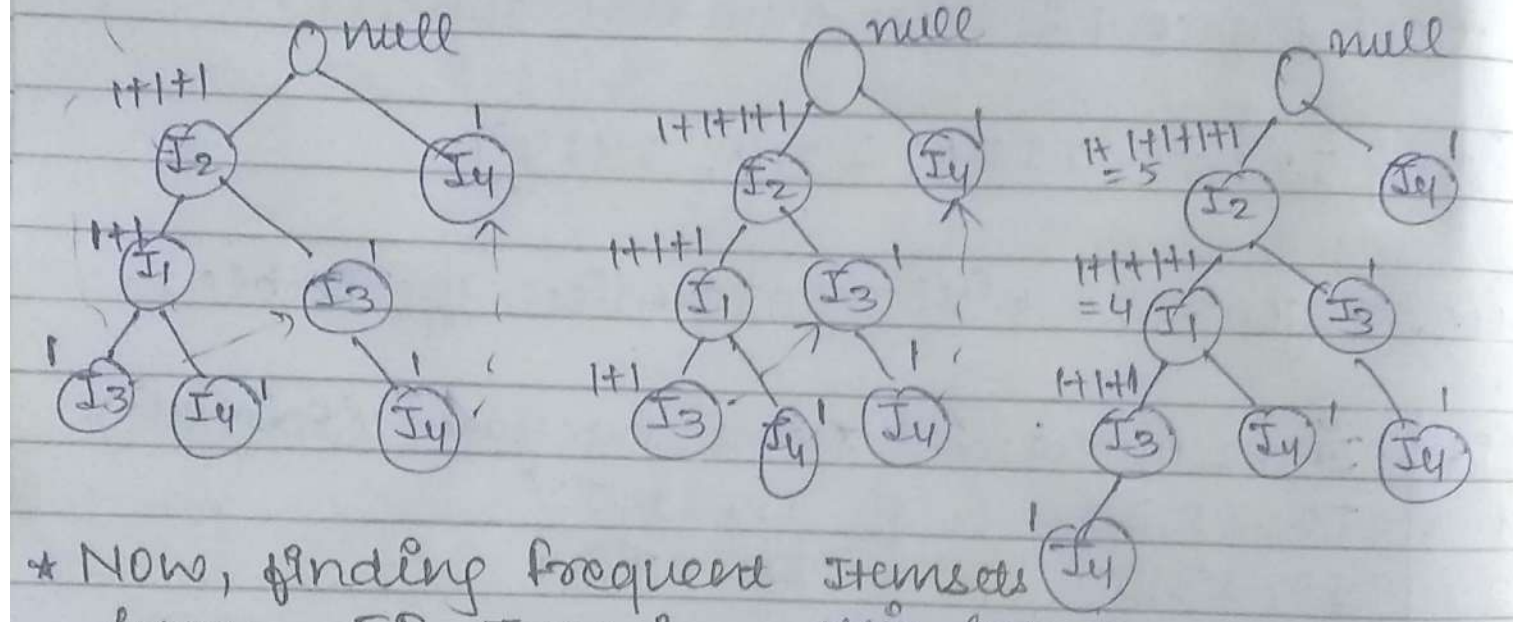
- Header Table :

Itemid	Sup. count	Node-link
I2	5	
I1	4	
I3	4	
I4	4	

Item Id	S-C	Node Link
I_2	5	
I_1	4	
I_3	4	
I_4	4	



$T_4 : I_2 I_1 I_4$ $T_5 : I_2 I_1 I_3$ $T_6 : I_2 I_1 I_3 I_4$



* Now, finding frequent Itemsets from FP Tree :- min. sup = 3

Item	Conditional Pattern Base	Conditional FP Tree	Freq Pattern
I_4	$I_2 I_1 I_3 : 1$, $I_2 I_1 : 1$, $I_2 I_3 : 1$, ~~$I_2 : 1$~~ (Root node)	$I_2 : 3$, $I_1 : 2$, $I_3 : 2$ ($\because$ ~~<~~ 3)	$I_2 I_4 : 3$
I_3	$I_2 I_1 : 3$, $I_2 : 1$	$I_2 : 4$, $I_1 : 3$	$I_2 I_3$; $I_1 I_3 : 3$; $I_1 I_2 I_3 : 3$
I_1	$\{I_2 : 4\}$	$\{I_2 : 4\}$	$\{I_2 I_1 : 4$

28/9/20

Krishna

Data Mining

Unit-1

Data Pre-Processing – Chapter-1

KDD Process

Data

May be structured or not

→ Data Integration & Data cleaning

Needed bcs data might be available in diff sources

Needed bcs "null" values or "duplicate" values may be present in data.

→ Data stored in well defined Data ware house schema

Data selection & Data transformation

Relavent attributes must be selected out of 'n' no. of attributes (Data selection)

attribute values are transformed to new set of values. (Data transformation)

[True-1 false-0]

↓

Model building (Data mining)

Diff class, clustering & pattern designing

↓

Model evaluation using diff metrics

* If evalution results are not upto mark then these steps must be repeated.

Applications:- Retail sector, Banking sector, healthcare Industry & social and information networks etc.

Data types

- Quantitative (Numerical) count, amount
  - * Quantitative
    - Discrete (finite value of integers)
      - Visualize:- Bar charts / histograms
      - Ex:- No. of movies etc
    - continuous (not limited to integers)
      - Ex:- Distance, height etc
      - V:- Scatter / histograms
- Qualitative (Features) symbols, words / letters
  - Nominal
  - Ordinal
  - Binary

*Nominal:- Labeled using symbols / words / letters
Ex:- Name of place, colour of skin etc
V:- Pie charts (percentage of male & female)

*Ordinal:- Order among categories
Ex:- small, medium large etc.
V:- Bar graph.
Good (or) Avg, worst
Grades:- S, A, B, C, D

Binary:- Takes only 2 values

↳ Pass, fail etc.

V:- Pie charts.

Data Preprocessing

Why:- Raw (dirty) data is not adequate for analysis.

Reasons:- Quality (Not good quality.

Dimensionality (Too many attributes)

What:- Access the quality of data.

Measures

1) Completeness, consistency. → Not recorded, Missing values (% of ↑)
2) Timeliness, Accuracy (Or not) → Disturbances happened during data integration in codes / names. (Two tables merging etc)
3) Timeliness, Believability
4) Interpretability.

* Accuracy:- Is recorded value correct / not.
* Timeliness:- Timely availability of data, mean must available / updated data.
* Interpretability:- How easily data is understand
* Believability:- The degree of which user can accept data.

Major tasks in Data Processing

↳ Dirty to quality data

* Data cleaning (fills missing values & duplicacy removed)
* " Integration (Resolves issues during integration)
* " Reduction [dimension reduction]
* " Transformation [attributes values changes to new set of values]

* Data cleaning

* Deals with incomplete data

↳ values are missing

EId	Name	Gender
1	—	—
2	A	—
3	B	F
4	C	M
5	D	M

Methods :- 1. Ignore the tuple.

Ex:- tuple 1 is deleted / ignored.

2. Use a global constant

Ex:- Filling missing values with $-\infty$ or unknown

3. Use central tendency :-

Mean :- Normally distributed data

Median :- Skewed

Mode :- If data is catagorical

like replace Gender of tuple 2 with 'M' ← mode.

4. Fill in the missing values manually.

5. Use most probable value:

* regression
* Decision tree.

Noisy Data

* It is a random error or ~~co~~variance in a measured variable.

tId	Exper	salary
1	10	1000
2	15	-20
3	200	4000
4	15	8000

noise values

* Methods :- 1. Binning ← means, medium, boundary.
2. Regression
3. Outlier Analysis.

Binning

* First sort data
* Partition into equal pins (equal freq) & (equal depth)
* Then apply means/my bar

Ex:- 7 3 9 19 14 23 31 24 38 33

Sort:- 3, 7, 14, 19, 23, 24, 31, 33, 38

Bin 1:- [3 7 14] B2:- [19 23 24] B3 9/3 = 3 in each bin

B1 B2 B3

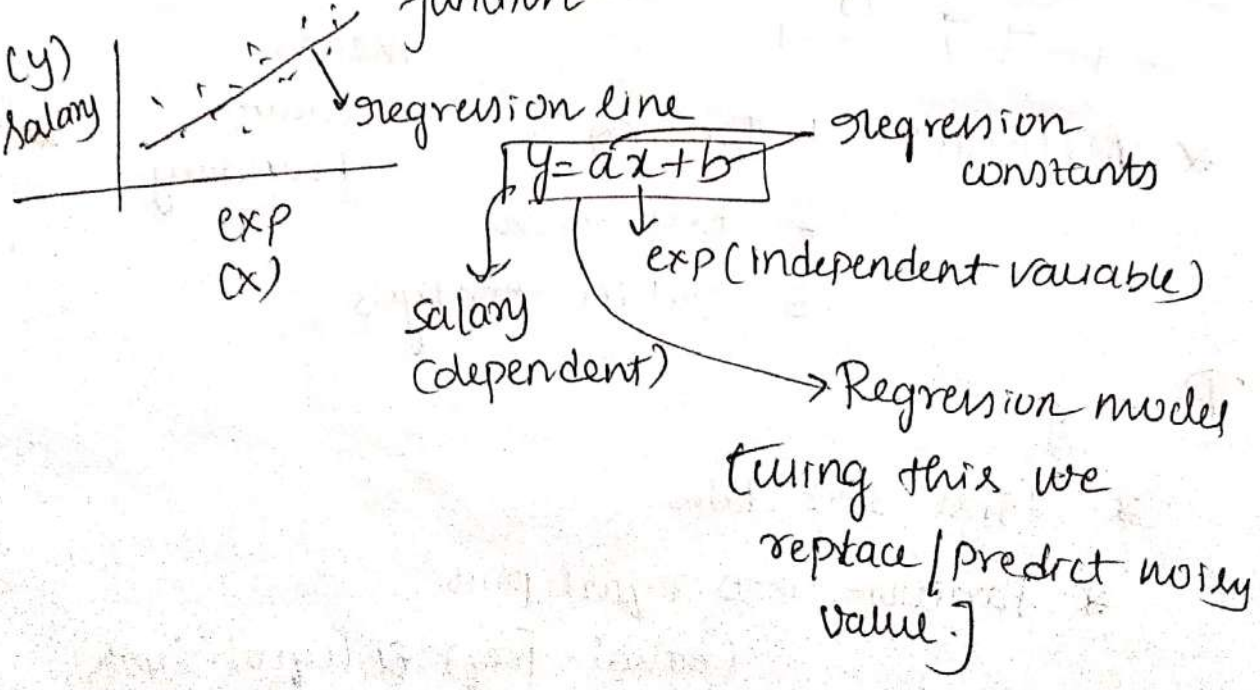
=> Mean:- [8, 8, 8] [22 22 22] [34, 34 34]

=> Median:- [7 7 7] [23 23 23] [33 33 33]

=> Bounday [3 3 14] [19 24 24] [31 31 38]

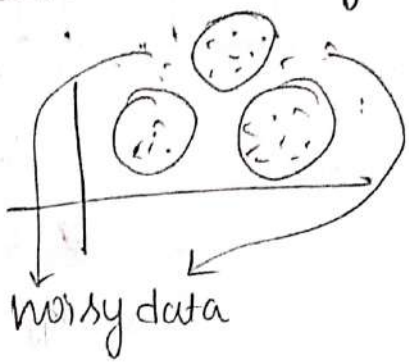
closest bounday value for intermediate values

* Regression: Data can be smoothed by fitting the data to a regression function



Outlier Analysis

* We group data points based on similarity it measures.
* Here noisy data is identified & eliminated



Data integration

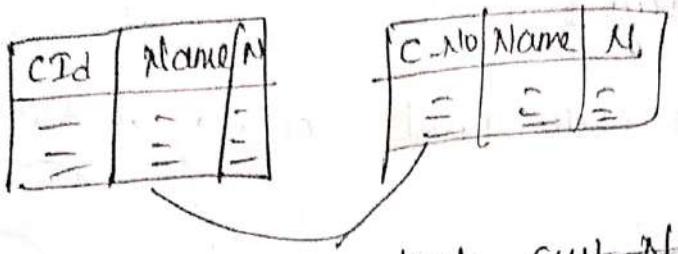
Why:- When data required is available in different sources.

What:- Preprocessing technique, which integrates the data from diff sources into one to provide singe view of data. & It also resolves the problems arrived during integration.

Problems | issues
1) Schema integration.
[Entity identification Prob]
2) Data value conflict Detection & resolution (DVCD&R)
3) Redundancy.

1) EI Prob
→ Objects that respresents to the same-real world entity are represented in diff way in diff tables.

For ex:-



Same but ~~Cust_NO & CIS~~

* C_Id & C_NO are represented as diff here which are same actually.

Sol:- Use metadata.

⇓

which has Name, datatype, meaning range of values allowed & ~~Null~~ rules to handle (0, null & blank values) for attribute.

* So here metadata tells C_Id & C_NO are same, hence ~~EI~~ prob is resolved.

(ii) DVCD & R

↳ data is same but some attributes are represented in different units like Rs & $ for money.

Sol:- Use metadata

*Redundancy

→ Derivable attribute

→ same attribute with diff names.

* Ex:- Age, DOB
↓ in 1st table
↓ in 2nd table

* This occurs bcs there is relationship b/w attributes
* So we have to identify R.S b/w them

Solution:-

(1) correlation / covariance Analysis → for numerical
(2) Chi square test → for catagorical.

Correlation

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

(a_i = value of att in ith tuple; $\bar{A}$, $\bar{B}$ = means)

-0.95 (strong -ve)
+0.2 (week +ve)

if r { <0 -vely; >0 +vely; =0 No relation (independent)

$$-1 < r < +1$$

* It gives "direction & strength"

Covariance

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

* It gives only "direction"

-0.95 (-ve)
+0.2 (+ve)

if r { <0 -vely; >0 +vely; =0 Bnd

* Chi-square test (catagorical data)

Null hypo:- A & B are independent (default)

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

Example

	Male	Female	Sum (row)
Coffee	200 (90)	250 (360)	450
Tea	100 (210)	950 (840)	1050
Sum (col)	300	1200	1500

$$e_{11} = \frac{300 \times 450}{1500} = 90 \text{ ex}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(ob - ex)^2}{ex}$$

$$= \frac{(200-90)^2}{90} + \frac{(250-360)^2}{360} + \frac{(100-210)^2}{210} +$$

$$\chi^2_{cal} = \frac{(950-840)^2}{840} = 240.07$$

DOF = (r-1)(c-1) = (2-1)(2-1) = 1

Sol = 0.001 → $\chi^2_{ob} = 10.828$

compare → $\chi^2_{cal} > \chi^2_{obs}$ → Reject Null

∴ Related

Data Reduction (DR)

→ Huge data analysis diff & inefficient hence DR is needed (bcs of redundancy, noise / null values)

→ It is used to reduce volume of data, so pattern extraction will be quicker and efficient.

→ Results will be same as the pattern extraction done without DR.

Methods 1) Dimensionality Reduction
↓ removes redundant or irrelavent att
2) Numerosity

1) we used attribute subset-selection method (DimeR) (ASS)

2) Numerosity :- Alternate Representation of data
- Parametric method
- Non " "

1) ASS - Heuristic attribute selection method

1. Stepwise forward selection
2. " backward elimination
3. Best combined attribute selection & elimination. (BCASE)

} methods

Stepwise FS

Attributes :- {A1, A2, A3, A4, A5, A6, A7, A8}

Evaluation measures :- {3, 8, 5, 12, 14, 6, 9, 2}
→ highest S.L (most imp)
↓ least
↓ imp of A1 (S.L of att)

Selection threshold :- '5'
→ How many att we want

Procedure

1) { }
2) {A4}
3) {A4 A7}
4) {A4 A7 A2}
5) {A4 A7 A2 A6}
6) {A4 A7 A2 A6 A3}
n=5
5 imp att

Step wise B-E
↓
same eliminates from least

Produce

1) {A1, A2, A3, A4, A5, A6, A7, A8}
2) {" " " A7}
3) {A2, A3 A4 A5 A6 A7}
4) {A1 A3 A4 A6 A7}
n=5
10 terminates
×

BCASE

* In this both forward selection & backward elimination occurs simultaneously.

Procedure

1) { A4 added, A6 eliminated }

2) { A4, A7 added, A1 " }

3) ---- similar way.

* In this way, these 3 methods reduces no. of att based on threshold.

Numerosity - R (NR)

1) Parametric { linear-R, mult. linear-R

2) Non " { Histograms, clustering, sampling.

1) Parametric [Parametric value regression coeffients]

i) linear-R

exp	salary

salary (y) | exp (x)

$y = \alpha x + \beta$ (RC) (drawn using least squares method)

* only RC (α, β) are retained

(ii) Multiple L.R

* In this $\alpha_1, \alpha_2, \alpha_3 \ldots, \beta$ are retained

where $y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \ldots + \beta$

This reduces volume by storing parametric values of data model.

Non Parametric methods

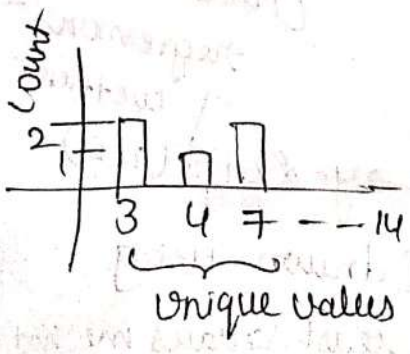
Histograms :- By storing avg (sum) for each bin

Bin - single values or range of values

* Here Bin & occurances of values will be shown

Eg 3, 4, 7 -- 3 - 7 - 14

Singleton bucket

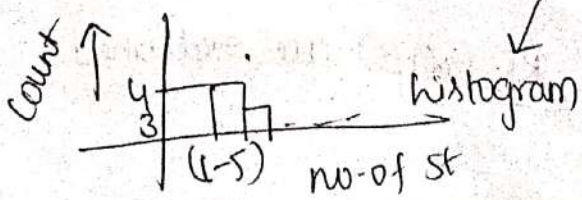


Equal width method

$w = \frac{max - min}{n}$ → no of bins

$w = 5$

width	count
1-5	4
6-10	3
- -	⋮
20-25	10

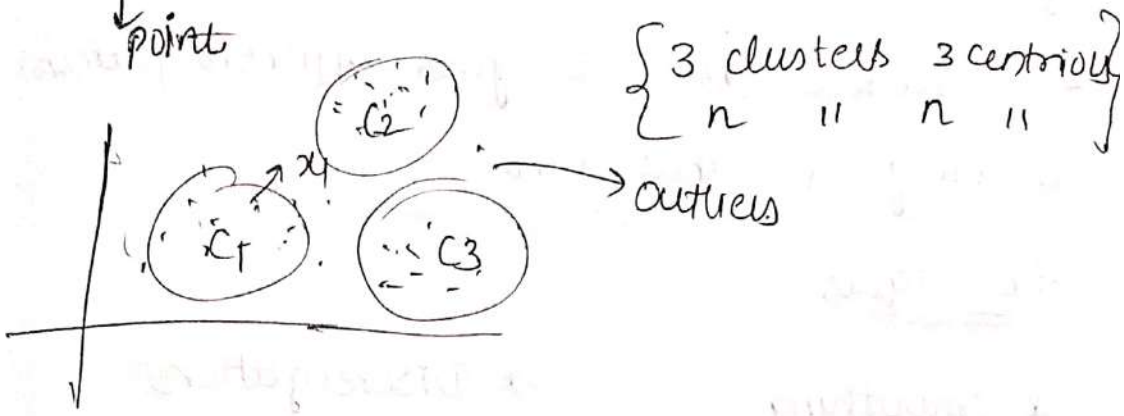


Clustering :- Store cluster parameters

Eg ⇒ centroid & diameter only

$x_1 = (3,4)$, $x_2(5,6)$ --- x_n ( )

↓ point



* Here we are storing (centroid & diameter) values instead of all data points.

Sampling ⟶ Reduces data by sample size.

* Simple random sampling (SRS)
* SRS without replacement (SRSWOR)
* " with " (SRSWR)
* Stratified sampling (SS)

SRSWOR

2 samples with size 3

SRSWR → same

S_1	S_2
T_2	T_1
T_4	T_7
T_6	T_9

S_1	S_2
T_2	T_2
T_4	T_7
T_6	T_4

Income

T_1	L
T_2	M
T_3	L
T_4	L
T_5	M
T_6	L
T_7	H
T_8	H
T_9	H
T_{10}	H

SS

$S_1 = T_1$

L { T_3, T_4, T_6

$S_2 = T_2$

M { T_5

50% = T_2

50% stata T_3 T_6

Data Transformations

→ Attribute values are mapped to new set of values.

→ Needed bcs it gives efficient patterns & easy to understand.

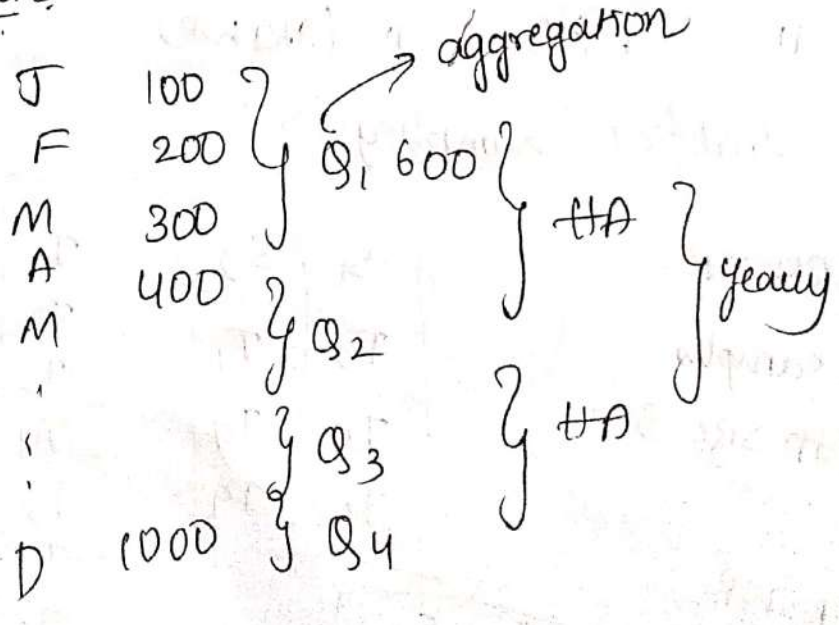
techniques

* Smoothing
* Aggregation
* Normalization
* Discretization
* Concept hierarchy
* Generation

Smoothing :- Using Binning (removes spikes)

[3 7 2] $\xrightarrow[\text{Mean-B}]{}$ [4 4 4]

Aggregation :-



Normalization

* Gives equal weights to all attr

Methods

* 1. Min-Max-N
  2. Z-Score-N
  3. Normalization by decimal scaling.

1) Min-Max-N

→ Maps values to '1' 'or' '0' or any values b/w 0 & 1

→ min = 0 max = 1 rest in b/w 0 & 1

$$V' = \frac{(V - min_A)\,(new\ max_A - new\ min_A)}{(max_A - min_A)}$$

(new max_A = 1, new min_A = 0)

Income

12222 (min) = 0

$$\frac{(12222 - 12222) \times (1-0)}{(85000 - 12222)} = 0$$

18000 (b/w value) → $\frac{(18000 - 12222)(1-0)}{85000 - 12222} = 0.038$

85000 (max) = 1

$$\frac{(85000 - 12222) \times (1-0)}{(85000 - 12222)} = 1$$

2) Z-Score (used when min & max values are not known)

$$V' = \frac{V - \mu_A}{\sigma_A}$$

$$V' = \frac{73600 - 54000\ (mean)}{16000\ (S.D)} = 1.225$$

if $v = \mu$ then $v' = 0$

if $v > \mu$ then v = +ve

if $v < \mu$ then v = -ve.

Normalization by decimal scaling

→ moving the decimal point of values of att A

→ Depends on the max absolute value of A

$$v' = v / 10^j$$

Ex -986 ------ 917

→ $+\frac{917}{10^3} = 0.917$

$= v' \Rightarrow \frac{-986}{10^3} = 0.986$

j → no. of digits

Data Discretization

→ Data values are divided into intervals

age : 20 30 -- 40 --- 80

young middle senior

{ equal freq
equal width → It labels data & replaces data values of actual data

Concept Hierachy Generation

→ lower level concepts are mapped to higher level.

→ Replaces lower-l.c to H.l.c.

Ex:- Country
|
State
↑
City
|
Street (replaced by country)

* Causation:- Causality mean two att A & B have a cause & effect relationship b/w them

→ A & B appear at same time, one after other or wont appear together etc.

* Correlation:- strength & direction of relation -ship b/w A & B

LP

3) 200, 400, 800, 1000, 2000, 2200,

Min-Max Normalization

min = 200 max = 2200

New min = 0 new max = 1

$$V_i' = \frac{(V_i - \min)(N.\max - N.\min)}{(\max - \min)}$$

$$V_1' = \frac{(200-200)(1-0)}{(2200-200)} = 0$$

$$V_2' = \frac{(400-200)(1-0)}{2000} = 0.1$$

$$V_3' = \frac{(800-200)(1)}{2000} = 0.3$$

$$V_4' = \frac{(1000-200)(1)}{2000} = 0.4$$

$$V_5' = \frac{(2000-200)(1)}{2000} = 0.9$$

$$V_6' = \frac{(2200-200)(1)}{2000} = 1$$

∴ 200, 400, 800, 1000, 2000, 2200
↓ ↓ ↓ ↓ ↓ ↓
0 0.1 0.3 0.4 0.9 1 (new values)

Values obtained after normalization

4) 200, 400, 800, 1000, 2000, 2200

n=2 (no of bins)

(i) Equal width partitioning

$$\Rightarrow W = \frac{\max - \min}{n} = \frac{2200-200}{2} = \frac{1000}{x_1}$$

Dividing data into bins:-

For bin 1:

$\Rightarrow$ 1000+200 = 1200 [include data from min to 1200 in bin 1]
$(x_1 + \min)$ x_2 x_2

$\therefore$ Bin1 : [200, 400, 800, 1000]

For bin 2:

$\Rightarrow$ 1200 + 1000 = 2200 [from 1200 to 2200]
$[x_2 + x_4]$

$\therefore$ Bin2 : [2000, 2200]

$\therefore$ New bins are [200, 400, 800, 1000] [2000, 2200]
Bin-1 Bin-2

(ii) Equal Frequency Binning

→ All bins will have equal freq (count of n observations)

n=2 $\therefore$ Total count / n $\Rightarrow$ 6/2 = 3 in each bin

$\therefore$ Bins are [200, 400, 800] [1000, 2000, 2200]
Bin-1 Bin-2

2) Data discretization techniques involves many, some of them are:

(i) histogram analysis
(ii) Binning
(iii) Correlation analysis
(iv) Clustering "
(v) Decision tree "
(vi) Equal width partioning
(vii) Equal depth "

{ explain them (any 2)

[and explain data trans -formation]

Ex:- Ages → 10, 11, 13, 14, 18, 20, 25, 40, 48, 60, 65, 80

input data [Before discretization]

1st We will discretize the data using labels

(i) young (ii) Middle (iii) Old age

young age → [10 11 13 14 18 20]

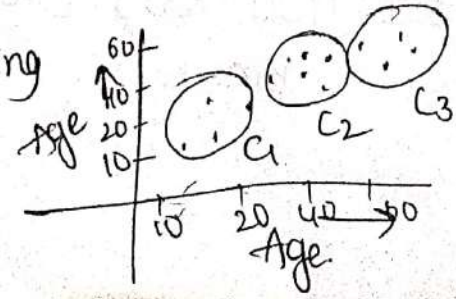
Middle age → [25, 40, 48]

Old age → [60, 65, 80]

Here we did just manually,

Using (vi) & (vii) etc also we can do for given data.

2nd clustering



13/10/20

Krishna

Frequent Pattern Mining - Chapter-2

(FPM)

FPM:- It is process of finding frequent patterns i.e relavent hidden patterns from database.

Frequent Patterns:- Itemsets / sub sequences or substructures that appear together in DB.

They represent very imp patterns / properties of data.

Methods to find

→ Finding frequent patterns

→ Deriving Association rules.

Support & Confidence

Tid	Items Purchased
1.	Milk, Bread, coffee
2.	Milk, Coffee
3	Milk, Butter
4	Bread, Egg

items.

Frequent ↑

Milk = 3 ✓

Bread = 2 ✓

Coffee = 2 ✓

Butter = 1 ✗

egg = 1 ✗

If support_min = 2

Threshold ⇑

Support:- We say that item is frequent only when the occurance of item in Transaction DB is more or equal to threshold of support

Two items frequent itemset → which occur frequently in data (comparing with threshold)

Milk, Bread - 1 ✗
Milk, Coffee - 2 ✓
Bread, coffee - 1 ✗

∴ Milk, coffee frequent itemset of 2

Association rule

Milk ⟹ coffee

Support of this rule = (Milk ∪ coffee)

⇒ S(M ∪ C) = 2

Relative support ⇒ $\frac{S(M \cup C)}{\text{Total No. of transaction}} = \frac{2}{4} = 50\%$

Confidence = $\frac{S(M \cup C)}{S(M)} = \frac{2}{3} = 66.6\%$

* Milk ⟹ Coffee (S = 2 (50%), C = 66.6%

Support:- It indicates that 50% of the transactions in Milk & coffee are purchased together.

Confidence:- It indicates that 66% of users who purchased milk has purchased coffee.

* These 2 measures are used to find interestingness of rule.

Apriori Algorithm

* Method to find frequent itemsets even we have FP growth Algorithm.

Apriori Property.

* If an itemset is frequent, then all of its subsets must be frequent.

. {A B} ⟶ Frequent then {A} {B} {AB} are frequent

* If an IS is infrequent, then all of its supersets are infrequent too

{A} ⟶ In frequent then {AB} {AC} are infrequent.

Algorithm

* Level wise search.

Join step :- C_k is generated by joining L_{k-1} with itself.

Prune step :- Any (k-1) item set that is not frequent cannot be a subset of a frequent k-itemset.

Ex:-

T-id	Items Purchased
1	Milk, Bread, coffee
2	Butter, Bread, Egg
3	Milk, Bread, Butter, Egg
4	Butter, Egg

Min support = 2

Step-1 Generate C_1

C_1

Itemset	support count	
M	2	
Br	3	
C	1	X
Bu	3	
E	3	

We prune step ⟹

L_1

Iset	S-C
M	2
Br	3
Bu	3
E	3

Step-2 C_2 (Joint L_1 with L_1)

Iu	Sc	
M Br	2	
M Bu	1	X
M E	1	X
Br Bu	2	
Br Eg	2	
Bu Eg	3	

prune C_2 ⟹

L_2

Iut	Sc
M Br	2
Br Bu	2
Br E	2
Bu E	3

Step-3 (Join L_2 with L_2)

C_3 (candidate table)

L_3 (3 freq item set)

Iset	S.C
M Br Bu	1 ✗ → (M Bu) not there
M Br Eg	1 ✗ (M E) not there
Br Bu Eg	2

⟹

Iset	S C
Br Bu Eg	2

↑

2 → Use apriori algorithm here if any of these subsets are infrequent then that particular set is infrequent.

Closed & Maximal frequent Item set

* When we generate FIS with the threshold set by user to support, there will problem is if min-support value is low, bcs we will get many no-of subsets.

Downward closure Property

* Any subset of a frequent itemset must also be frequent.

{ABC} - {A}{B}{C}
{AB}{BC}{CA}
$2^3-1=7$

For N distinct items,
FIS will be 2^N-1

* if N is large. is very large like 100 $2^{100}-1$ will be very huge number, hence to overcome this problem we have to as closed & Maximal frequent set.

↓

Downward closure property.

Maximal FIS :- An FIS is called as maximal FIS if its supersets are infrequent.

Consider the last ex

* This MFIS cant have info about support count its subsets

↓

TO overcome this we go with closed FIS

Step3

Iset	S-C
{Br Bu t}	2

→ Any superset of it is infrequent

↓

Maximal FIS
BCS algo has terminated here.

Closed FIS :- An FIS is called as closed FIS if its superset dont have same support count of it.

consider the last ex

Step-1

l_1 l_2

Itset	Sup-count
{M}	2 (Not closed)
{Br}	3 ← closed
{Bu}	3
{E}	3

Itset	S-C
{M Br}	2 (maximal) & ↓ closed bcs superset of it
{Br Bu}	2
{Br Eg}	2
{Bu Eg}	3

same value ✓ ✓ ✓ ✓ ✗

∴ Eg is not closed

{M Br Eg} (=1) infrequent

* If any one of superset is having same or greater value of support count, then it can't be closed set.

{M}, {E} are not closed, {Br} is closed → bcs its superset's S-C is not equal & less than its S-C.

* Maximal FIS can be closed but closed is not maximal.

Relation b/w them

* closed is FIS
* Maximal is closed & FIS

Maximal ⊂ closed ⊂ FIS

So, When we have many subsets with given min-support using apriori algorithm, we can go with maximal FIS or closed FIS

↓

preferred as it will contain info about its subsets.

Generating Association Rules (AR) from frequent k-itemsets

Ex

T-id	Items Purchased
1.	Milk, Br, Cof
2.	Bu, Br, Eg
3.	M, Bu, Br, Eg
4	Bu, Eg

$s = c = 2$

Using apriori Algo, we found FIS as (Bu, Br, Eg)

Rules of generating AR

* For each FIS I, generate all nonempty subsets of I

{Br, Bu, Eg} ← {Br, Bu} {Bu Bg} {Br Eg}

{Br, Bu, Eg}

* For every non empty subset of S of I, O/P the rule "S ⟶ {I-S}" if ⇒ $\frac{Sc(I)}{Sc(S)} \geq$ min-cof

S → Subset
{I-S} → remaining subsets.
then we can generate rule.

Strong Rules

| Min-cof = 70% |

R1: {Br, Bu} ⇒ Eg condition to check: $\frac{Sc(Br, Bu, Eg)}{S(Br, Bu)}$

$\Rightarrow \frac{2}{2} \times 100 = 100\%$. ✓ ≥ 70%

∴ Strong Rule.

R2: {Bu, Eg} ⇒ {Bu} $\Rightarrow \frac{2}{S(Bu, Eg)} = \frac{2 \times 100}{3} = 66.6\%$. ✗ ≤ 70%

∴ Not strong rule.

Other Method of finding FIs

FP growth {Frequent Pattern growth} Algorithm

Drawbacks of Apriori

* Generates large candidate Is if Is in the database is large.
* Needs multiple scans of database.

To overcome this drawback, we go with FP growth algorithm.

FP GM

* Divide & conquer strategy
* It compresses the database into a frequent pattern tree.
* It divides the compressed database into a set of conditional databases.
* Each conditional database is associated with one frequent item.

Example Min-Sup=3

T-id	Items purchased
1	I_1, I_2, I_3
2	I_2, I_3, I_4
3	I_4, I_5
4	I_1, I_2, I_4
5	I_1, I_2, I_3, I_5
6	I_1, I_2, I_3, I_4

Step1 $I_5 : 2 < 3$ ✗

$I_2:5, I_1:4, I_3:4, I_4:4$ (in des Asc order)

* We have reorganise them acc to desc order of their S.C.

step-2

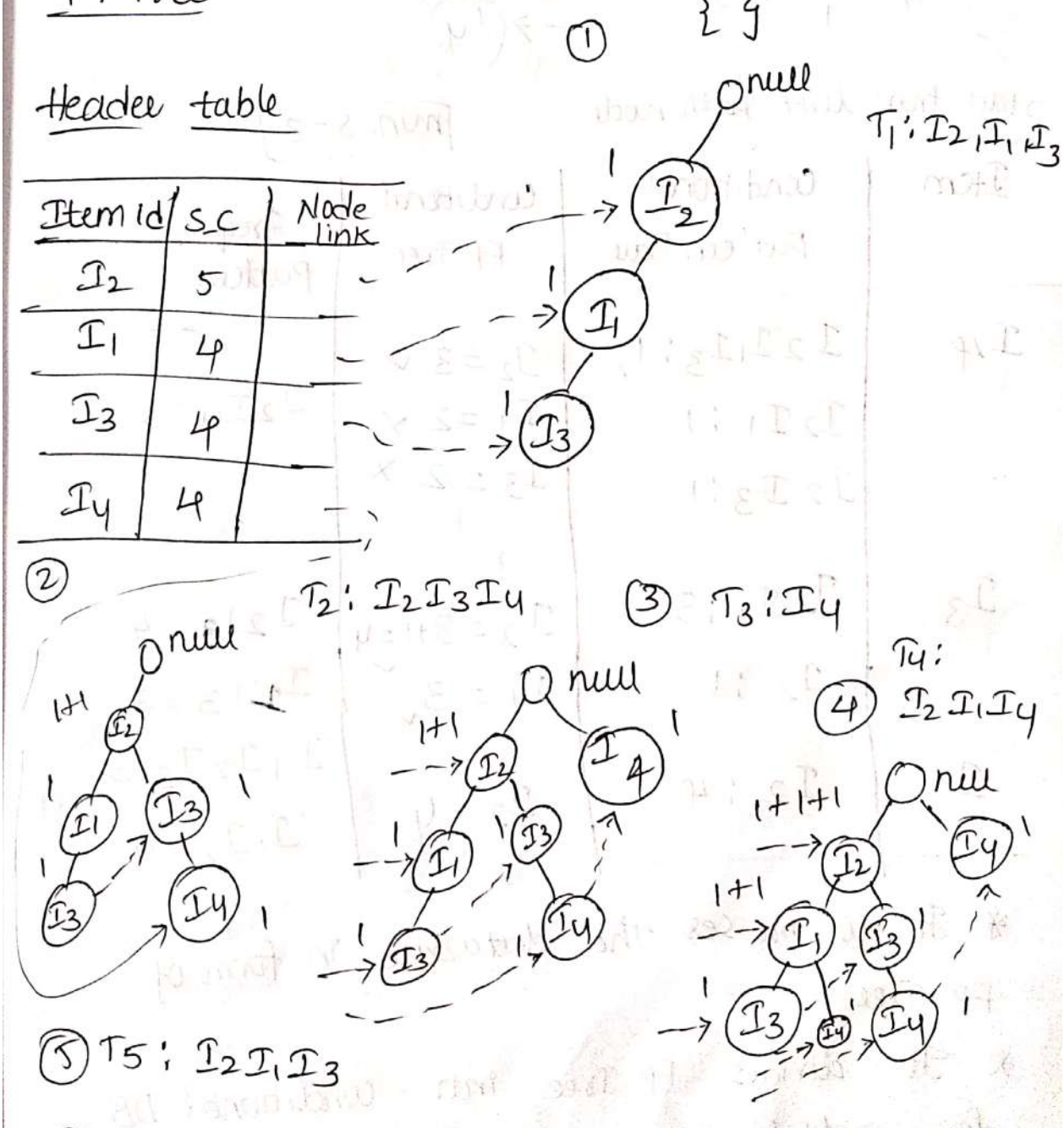
Tid	I.P
1	I_2 I_1 I_3
2	I_2 I_3 I_4
3	I_4
4	I_2 I_1 I_4
5	I_2 I_1 I_3
6	I_2 I_1 I_3 I_4

step-3

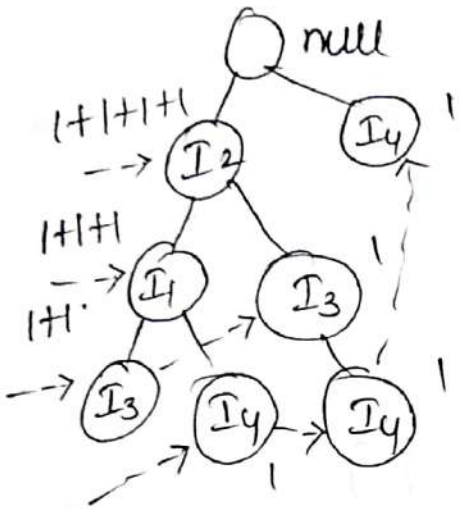
FP Tree

Header table

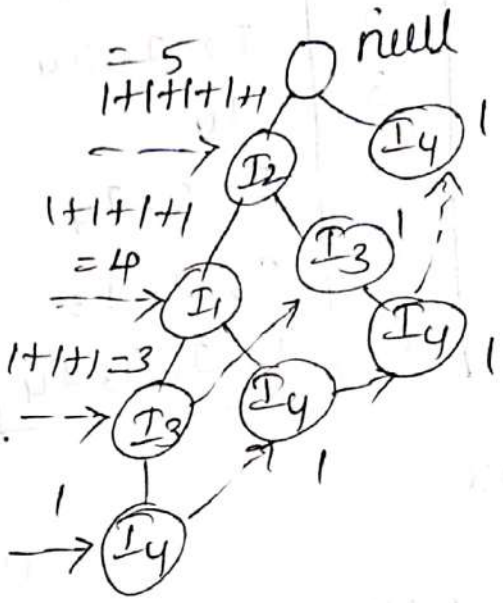
Item id	S.C	Node link
I_2	5	
I_1	4	
I_3	4	
I_4	4	



⑤ I_5: $I_2\ I_1\ I_3$



⑥ I_6: $I_2\ I_1\ I_3\ I_4$



Start from leaf ~~path~~ node

min_s = 3

Item	Conditional Pattern Base	Conditional FP Tree	Freq Patterns
I_4	$I_2\ I_1\ I_3$: 1, $I_2\ I_1$: 1 $I_2\ I_3$: 1	$I_2 = 3$ ✓ $I_1 = 2$ × $I_3 = 2$ ×	$I_2 I_4 = 3$
I_3	$I_2\ I_1$: 3 I_2 : 1	$I_2 = 3+1 = 4$ ✓ $I_1 = 3$ ✓	$I_2 I_3 = 4$ $I_1 I_3 = 3$
I_1	I_2 : 4	$I_2 = 4$	$I_1 I_2 I_3 = 3$ $I_1 I_2 = 4$

* It compresses the database in form of FP Tree
* It divides FP Tree into conditioned DB from which we extract frequent patterns
* It generates only frequent Itemsets.

Limitations of Support & Confidence Framework

Misleading Association Rules

$\{AB\}$ min-sup = 20% value

min-conf = 60%.

$A \Rightarrow B\ [S\ \ C]$

Antecedent — consequent — S: 30% — C: 70%

Strong rule in reality

* In some cases this rule might not be Strong rule though it satisfies min-su & min-co

So, lets solve that problem.

Example

	Horlicks	Not Horlicks	Sum (row)
Coffee	350	800	1150
Not coffee	90	10	100
Sum (col)	440	810	1250

R1:

$H \Rightarrow C\ [S = 28\%, C = 79\%]$ ✓ strong rule.

$S = \frac{S(H \cup C)}{\text{Total. T}}$ $\quad C = \frac{S(H \cup C)}{S(H)}$

$= \frac{350 \times 100}{1250} = 28\%$ $\quad = \frac{350 \times 100}{440} = 79\%$.

R2: $H \Rightarrow {}^{(not)}\overline{C}\ [S = 7.2\%, C = 20.45]$ X not strong rule.

$S = \frac{S(H \cup \overline{C})}{T.T} = 7.2\%$, $\quad C = \frac{S(H \cup \overline{C})}{S(H)} = \frac{90}{440} \times 100 = 20.45\%$.

∴ R_3 :- $\neg H \Rightarrow C$ [S = 64%, C = 98.7%]

✓ strong rule

* Here R_1 & R_3 both are strong rules according to min-s & m-con, But which is at is the challange

$$H \longrightarrow C \quad (or) \quad \neg H \longrightarrow C$$

Here the problem occured due to unf

*To overcome this, we go with '2' rules LIFT & Chi-square test

Lift : $Lift(H,C) = \frac{C(H \rightarrow C)}{S(C)}$

R_1

$$\neq \frac{S(H \cup C)}{S(H)} = \frac{S(H \cup C)}{S(C)} = \frac{S(H \cup C)}{S(H)S(C)}$$

$$= 0.86$$

$$Lift(\neg H, C) = \frac{S(\neg H \cup C)}{S(\neg H)S(C)}$$

R_3

$$= \frac{90/1250}{\frac{440}{1250} \times \frac{100}{1250}} = 2.55$$

* Lift { 1 - Independent
> 1 - correlation is there
< 1 - not correleted

∴ R_1 is not strong (0.86 < 1)

R_3 is strong (2.55 > 1)

Chi-Square

$$\chi^2_{cal} = \sum \frac{(Obs - exp)^2}{Obs}$$

$\chi^2_{cal} = 0$ (Independent)

$\chi^2_{cal} > 0$ (+ve or -vely correlated)

* First calculated expected values for all obs then calculate dof then χ^2_{cal}

if $\chi^2_{cal} > \chi^2_{obs}$ (from table)

→ reject H_0 (null hypothesis) which states they are independent

if Obs = 350
cal exp = 440
then $\chi^2 < 0$ mean 350-440 is -ve
∴ -velycorrelated.

if Obs = 90
exp = 20
90-20 = +ve ∴ +velycorrelated.

* Both Lift & chi-square are correlated measures.
* They overcome the problems of support & confidence.

15) Yes = 7 (P)
No = 3 (n)

$\log_2 x = \frac{\ln x}{\ln 2}$

$$I_D = \frac{-7}{10}\log_2\left(\frac{7}{10}\right) - \frac{3}{10}\log_2\left(\frac{3}{10}\right)$$

$\log_2\left(\frac{7}{10}\right) = \frac{\ln(0.7)}{\ln(2)}$

$$= -0.7 \times (-0.5145) - 0.3\,(-1.7369)$$

$$= 0.8813 \text{ //}$$

Outlook :-

sunny = 5 < 5-Y, 0-N

rainy = 5 < 2Y, 3N

$$\sum \frac{p_i+n_i}{p+n} = \frac{p_i+n_i}{p+n} I(p_i n_i) + \frac{n_i+p_i}{p+n} I(p_i, n_i)$$

$$I_{Outlook} = \frac{(5+0)}{10} I(5,0) + \frac{(2+3)}{10}(2,3)$$

(5+3) ↓ 0 (7+3)

$$= \frac{5}{10}\left[-\frac{5}{5}\log_2\frac{5}{5} - 0\right] + \frac{5}{10}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]$$

($\frac{5}{5}\log_2\frac{5}{5} = 0$)

$$= 0.2\,[-0.4 \times -1.3219 - 0.6 \times -0.7369]$$

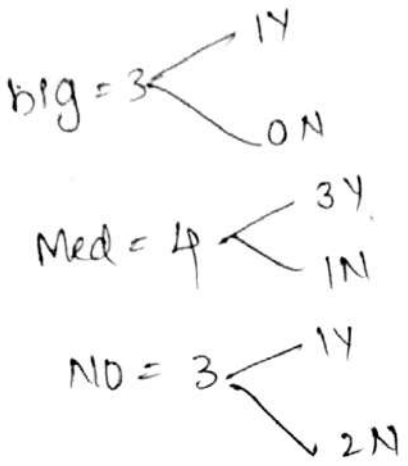
$$= 0.4855$$

$$Gain_{outlook} = I_D - I_{outlook} = 0.8813 - 0.4855 = 0.3958$$

$$G_{outlook} = 0.3958$$

Cam

Company



$$I_{company} = \frac{3}{10} I(3,0) + \frac{4}{10} I(3,1) + \frac{3}{10} I(1,2)$$

$$= \frac{3(0)}{10} + 0.4\left[\frac{-3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right.$$

$$\left. + \frac{3}{10}\left[\frac{-1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)\right]\right.$$

$$= 0.4\left[-0.75x - 0.41503 - 0.25(-2)\right] +$$

$$0.3\left[-0.33x - 1.5849 - 0.66x - 0.5849\right]$$

$$= 0.3245 + 0.2727$$

$$= 0.5972 = 0.6$$

$$\text{Gain}_{\text{company}} = 0.2813$$

Sailboat

Small = 5 < 4Y, 1N

Big = 5 < 3Y, 2N

$$I_{Sailboat} = \frac{5}{10} I(4,1) + \frac{5}{10} I(3,2)$$

$$= \frac{5}{10}\left[\frac{-4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right] +$$

$$\frac{5}{10}\left[\frac{-3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left[\frac{3}{5}\right]\right]$$

$$= 0.5\,[-0.8 \times -0.3219 - 0.2 \times -2.3219] +$$

$$0.5\,[0.4421]$$

$$= 0.8465$$

$$I_{Gain_{Sail}} = 0.0348$$

So, the outlook has higher gain value ∴ it becomes node

Outlook

Sunny

Company	Sailboat	Class
big	Small	Yes
med	Small	"
med	big	"
no	Small	"
big	big	"

done

Rainy

company	Sailboat	Class
~~no~~	small	~~no~~
med	small	Yes
big	big	Yes
no	big	~~no~~
med	big	no

consider this table now

Yes = 2 (p) Ptn = 5

No = 3 (n) $ID = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{2}{5}\right) = 0.9709$

company

$I_{comp} = \frac{+2}{5}I(0,2) + \frac{1}{5}I(1,0) + \frac{2}{5}I(1,1)$

= 0.4

no = 2 — 0 Y, 2 N

big = 1 — 1 Y

med = 2 — 1 Y, 1 N

Gain = 0.5709

Sailboat

Small = 2 < 1Y, 1N

big = 3 < 1Y, 2N

$I_{sailboat} = \frac{2}{5} I(1,1) + \frac{3}{5} I(1,2)$

$= \frac{2}{5}(1) + \frac{3}{5}(0.27253)$

$= 0.4 + 0.1636$

$= 0.9510.$

$G_{sailbatt} = 0.0199$