DATA MINING ANALYTICS

UNIT-1

* OLAP
* OLTP
* DM Architectures
  - ↳ Techniques
  - ↳ Task / Functionalities
  - ↳ Applications.
* Data preprocessing (X)
  - ↳ chi-square
  - ↳ Z-score
  - ↳ Binning
  - ↳ problems
* AOI (Attribute Oriented Analysis)
* Dissimilarity matrix, Distance matrix

UNIT-2

* Association Rule Mining
* Apriori (X)
* FP-growth (X)
* classification (write procedure by our own)
* Decision tree
* Naive Bayes
* KNN.
* Linear models.

UNIT-3 Theory (X)

* Test of hypothesis - linear regression
* Logistic regression - derivation, (Diff)
* Forward & backward
* Multi logistic regression
* GLM (Based on make rs no. of distribution)
* Non-linear regression (WISW)
* Additive, semi-parametric, Non-parametric degress (Notes)

UNIT-4

* Problem -> covariance & correlation (Given time series data)
* Intro of time series data
* Estimation of ARMA
  - ↳ yule - walker (X)
  - ↳ Max- likelihood
  - ↳ least square

Date :

Page No. :
Expt. No. :

Problems on UNIT-1, 2, 3, 4.

UNIT-1.

* Normalisation
  - ↳ Min-Max
  - ↳ z-score
  - ↳ decimal scaling
* Binning method - Data cleaning
* Dissimilarity for all data.
* chi-square test
* Wavelet transform - Data preprocessing

UNIT-2

* Apriori
* FP-growth
* Decision tree
* NB
* KNN
* Linear regression
* Classification metrics → performance evaluation.

UNIT-3

* Testing of hypothesis
* Linear regression
* Logistic regression

UNIT-4

* covariance.
* correlation.
* Prescriptive Analysis : Math optimization Revisions risk Analysis

Teacher's Signature:

Notes.

UNIT-4.

① Apriori algorithm.

sup = 25%, min - conf = 60%.

Tid	Items
T1	{A, B, C, D}
T2	A, C, D, F
T3	C, D, E, G, A
T4	A, D, F, B
T5	B, C, G
T6	D, F, G
T7	A, B, G
T8	C, D, F, G

sup = 25%

$\frac{25}{100} \times 8$

sup. count = 2.

	itemset	sup count
C1 ⇒	{A}	5
	{B}	4
	{C}	5
	{D}	6
	{E}	1 ×
	{F}	4
	{G}	5

	itemset	supcount
L1 ⇒	A	5
	B	4
	C	5
	D	6
	F	4
	G	5

Generate C2 candidates from L1

C2 ⇒ A, B → 3
A, C → 3
A, D → 4
A, F → 2
A, G → 2
B, C → 2
B, D → 2 ✓
B, F → 0 ×
B, G → 2
C, D → 4
C, F → 2
C, G → 3
D, F → 4
D, G → 3
F, G → 2

	itemset	supcount
L2 ⇒	A, B → 3	C, G → 3.
	A, C → 3	D, F → 4
	A, D → 4	D, G → 3
	A, F → 2	F, G → 2
	A, G → 2	
	B, C → 2	
	B, D ~~B, F~~ → 2	
	B, G → 2 ~~B, F~~ → 4	
	C, D	
	C, F P → 2	

Date :

Generate C3 candidates from L2

C3 ⇒ itemset	sup count
ABC	1 ×
ABD	2
ABF	1 ×
ABG	1 ×
ACD	3
ACF	1 ×
ACG	1 ×
ADF	2
ADG	1 ×
AFG	0 ×
BCD	1 ×
BCF	0 ×
BCG	1 ×
BDF	0 ×
BDG	0 ×
BFG	0 ×
CDF	2
CDG	2
~~CD~~ CFG	1 ×
DFG	2

L3 ⇒ itemset	sup count
ABD	2
ACD	3
ADF	2
CDF	2
CDG	2
DFG	2

Generate C4 candidates from L4

C4 ⇒ ABDC → 1 ×

ABDF → 1 ×

ACDF ~~CDFG~~ → 1 ×

ACDG ~~CD~~ → 1 ×

~~ADFG~~ ~~→ 1~~

ADFG → 0 ×

CDFG → 1 ×

~~CDGF~~

Teacher's Signature:

② FP-growth

support = 50% confidence = 60%

Tranaction	List of items
T_1	I_1, I_2, I_3
T_2	I_2, I_3, I_4
T_3	I_4, I_5
T_4	I_1, I_2, I_4
T_5	I_1, I_2, I_3, I_5
T_6	I_1, I_2, I_3, I_4

$$\text{support} = \frac{50}{100} \times 6$$

$$= \frac{5}{10} \times 6$$

$$= \frac{1}{2} \times 6 \; 3.$$

Item	Frequency
I_1	4
I_2	5
I_3	4
I_4	3
I_5	2 ×

Frequent Patterns are;

~~L = {K:~~

$L = \{I_2 : 5, I_1 = 4, I_3 : 4, I_4 : 3\}$

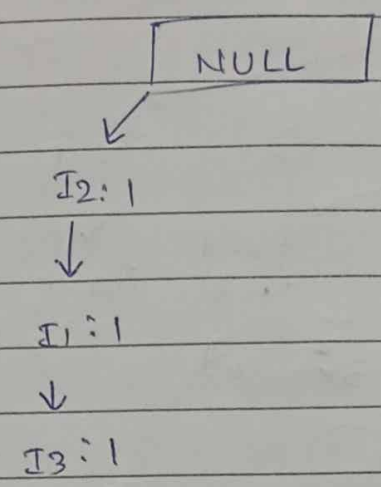
Tranaction ID	Items	Ordered-Item set
T1	I_1, I_2, I_3	I_2, I_1, I_3
T_2	I_2, I_3, I_4	I_2, I_3, I_4
T_3	I_4, I_5	I_4
T_4	I_1, I_2, I_4	I_2, I_1, I_4
T_5	I_1, I_2, I_3, I_5	I_2, I_1, I_3
T_6	I_1, I_2, I_3, I_4	I_2, I_1, I_3, I_4

SRI VIDHYA B

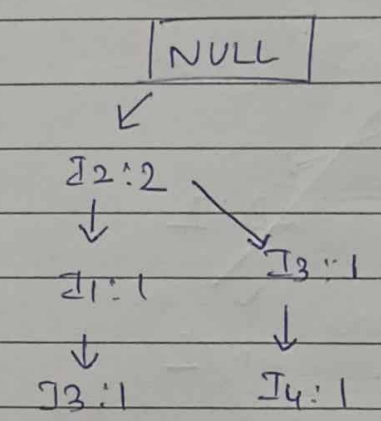
Date :

Page No. :
Expt. No. :

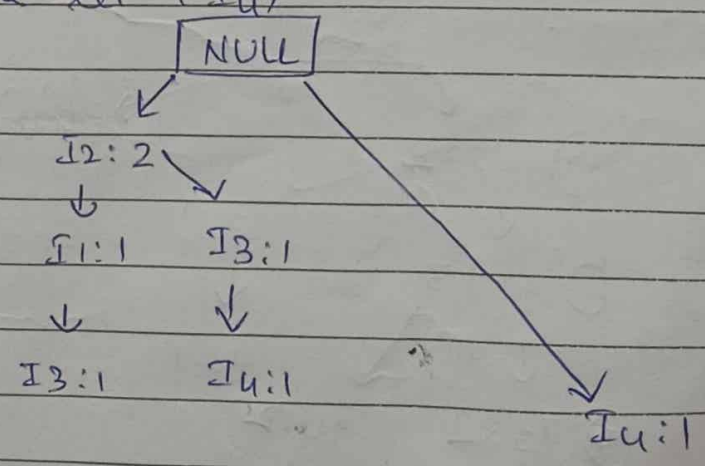
Inserting the set (I_2, I_1, I_3).



Inserting the set (I_2, I_3, I_4)

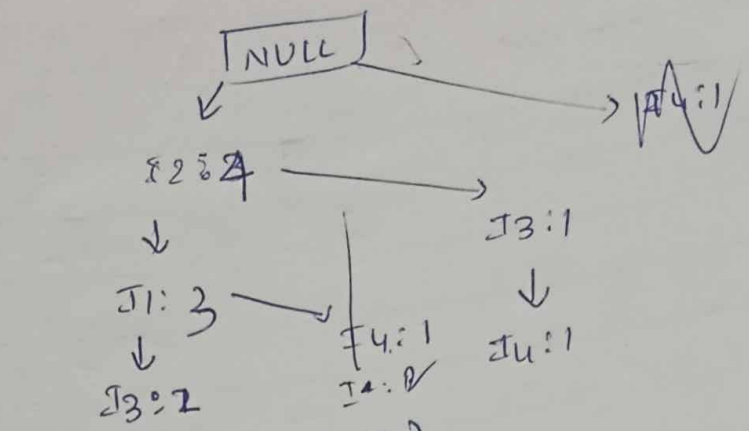


Inserting the set (I_4)

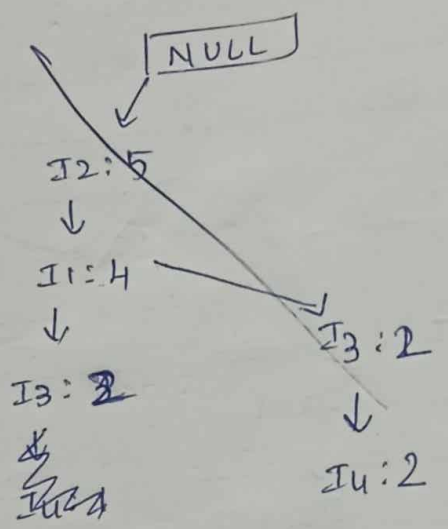


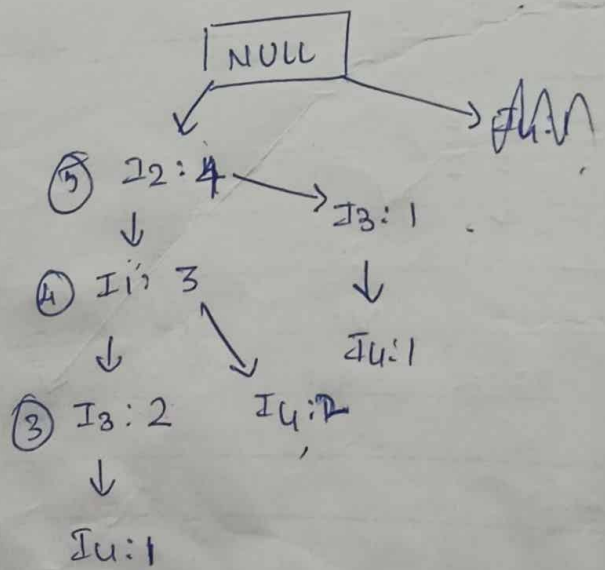
Teacher's Signature:

Inserting the set (I2, I1, I4)



Inserting the set (I2, I1, I3, I4)





∴ {I2:5, I1:4, I3:4, I4:3}

SRI VIDHYA B

Date :

Page No. :
Expt. No. :

compute conditional pattern base,

Items	conditional Pattern Base
I_4	$\{(I_2, I_1, I_3 : 1), (I_2, I_1 : 1), (I_2, I_3 : 1)\}$
I_3	$\{(I_2, I_1 : 3), (I_2 : 1)\}$
I_1	$\{(I_2 : 4)\}$
I_2	$\{(I_2 : 5)\}$

Items	CPB	conditional Frequent Pattern Tree
I_4		$\{I_2, I_4 : 3\}$
I_3		$\{I_2 : 4\}$
I_1		$\{I_2 : 4\}$
I_2		x

(3) KNN.

↳ Non-parameteric algo.

↳ Both classification & regression.

↳ lazy learner algo.

k -> no. of neighbours.

x	y	class
7	7	B
7	4	B
3	4	G
1	4	G

New instance (3, 7) predict the class.

Manhattan = $|(x_2 - x_1) + (y_2 - y_1)|$

Euclidean = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Teacher's Signature:

x_1 y_1
3, 7

6 -> 4
20.

B $|(3-7)+(7-7)| = 4$ ②

B $|(3-7)+(7-4)| = 7$ ④

G $|(3-3)+(7-4)| = 3$ ①

G $|(3-7)+(7-4)| = 5$ ③

$\because k = 3$

$k=1 \rightarrow G$, $k=2 \rightarrow B$, $k=3 \rightarrow G$

$\because k=3 \rightarrow$ good

$(3,7) =$ good.

④ Naive Bayes.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

class: c_1 : playtennis = 'yes'

c_2 : playtennis = 'no'

Data to be classified;

- x = (outlook = sunny, Temp = mild, Humidity = normal, windy = strong)

$P(c_1)$: P(playtennis = "yes") = 9/14 = 0.643

P(playtennis = "no") = 5/14 = 0.357

conditional probability:

outlook	Y	N
sunny	2/9	3/5
overcast	4/9	0
Rainy	3/9	2/5

Temp	Y	N
Hot	2/9	2/5
Mild	4/9	2/5
Cold	3/9	1/5

Date :

Page No. :
Expt. No. :

Humidity	Y	N	windy	Y	N
High	3/9	4/5	weak	6/9	2/5
Normal	6/9	1/5	strong	3/9	3/5

P(outlook = "sunny" | playtennis = "yes") = 2/9 = 0.22
" " " "No") = 3/5 = 0.6
P " " | PT = "yes") = 4/9 = 0.44.
P(Temp = "mild" | PT = "No") = 2/5 = 0.4
" " " "yes") = 6/9 = 0.66.
P(Humidity = "Normal" | PT = "No") = 1/5 = 0.2
" " " "
P(windy = "strong" | PT = "yes") = 3/9 = 0.33.
" " " "No") = 3/5 = 0.6

P(x|ci) = P(x|PT = "yes") = 0.22 × 0.44 × 0.66 × 0.33
= 0.0210

P(x|PT = "No") = 0.6 × 0.4 × 0.2 × 0.6
= 0.0288

P(x|PT = "yes") P(playtennis = "yes") = 0.0210 × 0.643
= 0.0135 ✓

P(x|PT = "No") P(PT = "No") = 0.0288 × 0.357
= 0.010

x = (outlook = sunny, Temp = mild, Humidity = normal,
windy = strong) = YES.

Teacher's Signature:

5. Decision Tree:-

Total = 14

yes = 9

NO = 5

$$Info(D) = -\sum P_i \log_2 (P_i)$$

$$Gain(A) = Info(D) - Info_A(D)$$

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

$$= (-0.6429 \times -0.6374) - 0.3571 \times -1.4854$$

$$= 0.4098 + 0.5304$$

$$\boxed{Info(D) = 0.940}$$

Attributes:-

a) Outlook;

	Y	N
S	2	3
O	4	0
R	3	2

0.5281

0.4422

$$Info_{outlook}(D) = \frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) +$$

$$\frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4} - 0\right) + \frac{5}{14}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$$

$$= \frac{5}{14} \times (0.9710) + \frac{4}{14}(+0) + \frac{5}{14}(0.9710)$$

$$= 0.3468 + 0.3468$$

$$\boxed{Info_{outlook}(D) = 0.6936}$$

b) Temperature;

	Y	N
H	2	2
M	4	2
C	3	1

c) Humidity;

	Y	N
H	3	4
N	6	1

0.33 0.3510 0.3861

Date : 0.52 0.3113 0.66 0.95

Page No. :
Expt. No. :

$$Info_{Temp}(D) = \frac{4}{14}\left(\frac{-2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) +$$

$$\frac{6}{14}\left(\frac{-4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) +$$

$$\frac{4}{14}\left(\frac{-3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)$$

$$= \frac{4}{14}(1) + \frac{6}{14}(0.90) + \frac{4}{14}(0.8110)$$

$$\boxed{Info_{Temp}(D) = 0.911.} \Rightarrow 0.28 + 0.38 + 0.23$$

c) $$Info_{Hum}(D) = \frac{7}{14}\left(\frac{-3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right) +$$

$$\frac{7}{14}\left(\frac{-6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right)$$

$$= \frac{7}{14}(0.5134 + 0.4602) + \frac{7}{14}(0.1890 + 0.3930)$$

$$= \frac{7}{14}(0.9736) + \frac{7}{14}(0.5820)$$

$$\boxed{Info_{Hum}(D) = 0.778}$$

d) Windy;

	Y	N
W	6	2
S	3	3

$$Info_{Windy}(D) = \frac{8}{14}\left(\frac{-6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right) +$$

$$\frac{6}{14}\left(\frac{-3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right)$$

Teacher's Signature:

$$= \frac{8}{14}(0.3237 + 0.5) + \frac{6}{14}(0.5 + 0.5)$$

$$= 0.4707 + 0.4286$$

$$\boxed{Info_{wind}(D) = 0.899}$$

$Gain(A) = Info(D) - Info_A(D)$.

$Gain(outlook) = 0.940 - 0.6936$

$= 0.246$ ✓

$Gain(Temp) = 0.940 - 0.911$

$= 0.029$

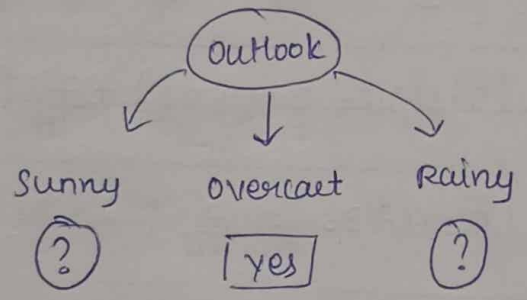
$Gain(Hum) = 0.940 - 0.788$

$= 0.152$

$Gain(Windy) = 0.940 - 0.892$

$= 0.048$

Attribute	Gain
outlook	0.246 ✓
Temp	0.029
Hum	0.152
windy	0.048



Sunny, Rainy table:-

outlook	Temp	Hum	windy	PT
	Hot	High	weak	No
Sunny	Hot	High	strong	No
	Mild	High	weak	No
	Cold	Normal	weak	yes
	Mild	Normal	strong	yes

Total = 5

yes = 2

No = 3

Date :

Page No. :
Expt. No. :

$$Info(sunny) = \frac{-2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.5288 + 0.4422$$

$$Info(sunny) = 0.9710.$$

Attribute;

a) Temp

	Y	N
H	0	2
M	1	1
C	1	0

$$Info_{temp}(D) = \frac{2}{5}\left(0 - \frac{2}{2}\log_2\frac{2}{2}\right) +$$

$$\frac{2}{5}\left(\frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) +$$

$$\frac{1}{5}\left(\frac{-1}{1}\log_2\frac{1}{1} - 0\right)$$

$$= \frac{2}{5}(0) + \frac{2}{5}(+0.5+0.5) + \frac{1}{5}(0)$$

$$\boxed{Info_{temp}(D) = 0.4}$$

b) Humidity;

	Y	N
H	0	3
N	2	0

$$Info_{hum}(D) = \frac{3}{5}\left(0 - \frac{3}{3}\log_2\frac{3}{3}\right) + \frac{2}{5}\left(\frac{-2}{2}\log_2\frac{2}{2}\right)$$

$$= \frac{3}{5}(0) + \frac{2}{5}(0)$$

$$\boxed{Info_{hum}(D) = 0}$$

Teacher's Signature:

c) windy;

	Y	N
W	1	2
S	1	1

-0.33

$0.5230+$

$$\text{Info windy}(D) = \frac{3}{5}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) +$$

$$\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$$= \frac{3}{5}(0.47 + 0.3510) + \frac{2}{5}(1 \quad)$$

$$= \cancel{0.6320 + 0.4} \quad \cancel{0.870 +}$$

$$\cancel{= 1.0320} \quad 0.4+$$

$$= (-0.33 \times -1.58) + (-0.66 \times -0.5850)$$

$$= 0.5214 \times \frac{3}{5} + 0.3861 \times \frac{2}{5}$$

$$= 0.3128 + 0.1544$$

$$= 0.4672 + 0.4$$

$$\boxed{\text{Info windy}(D) = 0.951.}$$

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

$$\text{Gain}(\text{Temp}) = 0.9710 - 0.4$$

$$= 0.571$$

$$\text{Gain}(\text{Hum}) = 0.9710 - 0$$

$$= 0.9710 \checkmark$$

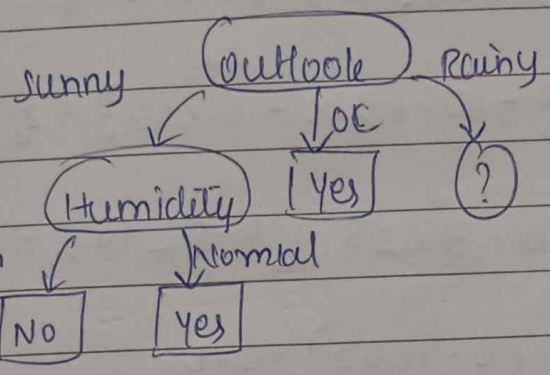
$$\text{Gain}(\text{windy}) = 0.9710 - 0.951$$

$$= 0.02$$

Date :

Page No. :
Expt. No. :

Attribute	Gain
Temp	0.571
Hum	0.9710 ✓
windy	0.02



Rainy table;

Outlook	Temp	Hum	windy	PT
	Mild	High	weak	yes
	Cool	Nor	weak	yes
Rainy	Cool	Nor	Strong	No
	Mild	Nor	weak	yes
	Mild	High	Strong	NO

Total = 5, yes = 3, NO = 2.

$$Info(Rainy) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 0.4422 + 0.5288$$

$$\boxed{Info(Rainy) = 0.9710}$$

Attribute;

Temp	Y	N
M	2	1
C	1	1

Teacher's Signature:

$$\text{Info Temp}(D) = +\frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) +$$

$$\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$$= \frac{3}{5}(0.38 + 0.47) + \frac{2}{5}(0.5 + 0.5)$$

$$= \cancel{0.6980} + 0.51 + 0.4$$

$$= \frac{3}{5}(-0.66 \times -0.58) + (-0.33 \times -1.58)$$

$$= \frac{3}{5}(0.3828 + 0.5214) + 0.4$$

$$= \frac{3}{5}(0.9042) + 0.4$$

$$= 0.9425$$

Hum	Y	N
H	1	1
N	2	1

$$\text{Info Hum}(D) = \frac{2}{5}\left[\cancel{\log_2} -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] +$$

$$\frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right).$$

$$= \frac{2}{5}(-0.5 \times -1) + (0.5) + \frac{3}{5}(-0.66 \times -0.5850) +$$

$$(-0.33 \times -1.58)$$

$$= \frac{2}{5}(1) + \frac{3}{5}(0.3828 + 0.5214)$$

$$= 0.4 + 0.6(0.9)$$

$$\boxed{\text{Info Hum}(D) = 0.94}$$

windy;

W	Y	N
S	3	0
	0	2

Date :

Page No. :
Expt. No. :

$$Info_{windy}(D) = \frac{3}{5}\left(-\frac{3}{3}\log_2\frac{3}{3} - 0\right) + \frac{2}{5}\left(0 - \frac{2}{2}\log_2\frac{2}{2}\right)$$

$$\boxed{Info_{windy}(D) = 0}$$

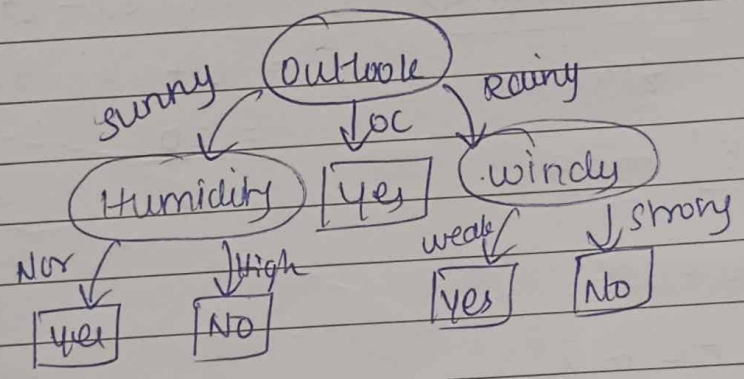
$Gain(D) = Info(D) - Info_D(D)$

$Gain(Temp) = 0.9710 - 0.9425$
$= 0.0285$

$Gain(Hum) = 0.9710 - 0.94$
$= 0.0310$

$Gain(windy) = 0.9710 - 0$
$= 0.9710$ ✓

Attribute	Gain
Temp	0.0285
Hum	0.0310
windy	0.9710 ✓



6. Linear regression :

Least square R^2 method.

a) Least Square method :

$$\boxed{y = b_0 + b_1 x}$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$ → slope (m)

Teacher's Signature:

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	14	-1	-6	1	36
3	24	1	4	1	16
2	18	0	-2	0	4
1	17	-1	-3	1	9
3	27	1	7	1	49.
				4	

$$\bar{x} = \frac{1+3+2+1+3}{5} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{14+24+18+17+27}{5} = \frac{100}{5} = 20.$$

$$y = b_0 + b_1 x$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{(-1 \times -6) + (+1 \times 4) + 0 + 3 + 7}{4}$$

$$= 6 + 4 + 0 + 3 + 7$$

$$= \frac{20}{4}$$

$\boxed{b_1 = 5}$ $m = 5$

$$y = b_0 + 5x$$

$$b_0 = \bar{y} - m\bar{x}$$

$$b_0 = 20 - 5 \times 2$$

$$b_0 = 20 - 10$$

$$\boxed{b_0 = 10}$$

$$y = b_0 + b_1 x$$

$$\boxed{y = 10 + 5x}$$

Date :

Page No. :
Expt. No. :

b) R^2 method

$$R^2 = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}$$

$$\hat{y} = mx + c.$$

x	y	$x_i-\bar{x}$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})^2$	$\hat{y}$	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})^2$
1	14	-1	-6	36	1	15	-5	25
3	24	-1	4	16	1	25	5	25
2	18	0	-2	4	0	20	0	0
1	17	-1	-3	9	1	15	-5	25
3	27	1	7	49	1	25	5	25
2	20							100

$\hat{y} = mx + c$

$m = 5, \; c = 10$

$1 \rightarrow \hat{y} = 5 \times 1 + 10$
$= 15$

$3 \rightarrow \hat{y} = 5 \times 3 + 10$
$= 25$

$$R^2 = \frac{100}{114}$$

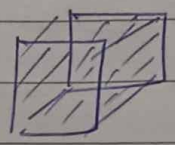
$R^2 = 0.877$

$R^2 = 87.77\%$

7. OLAP:- (Online analytical Processing)
* cube (3D-array)
* multidimensional model / data.

operations of OLAP.

a) Rollup:



Teacher's Signature:

* From small data's to larger data's

Eg: In TN & Kerala info. of there are product like citywise By "Rollup" it will give as statewise

b) Drilldown:

* opp. to rollup, from years to months.

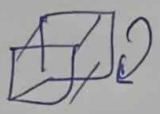
c) Slice:

→ Taking one slice of the portion

d) Dice:

* one size of dice

e) Rotate:

* Rotating 

8. Data mining architectures:

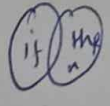
↳ Techniques:

(i) classification → (catogorical data) it involves training a model on labelled data and using it to predict the class labels of new, unseen data.

(ii) Regression → (continous data) Predict numeric (or) continous value based on the relationship btw input variables and target.

(iii) Clustering → (Grouping together) used to group similar data instances together based on their characteristics

(iv) Association Rule → (if then rule) (market-basket analysis)



If then rule.

Date :

Page No. :
Expt. No. :

two measures of association rule;

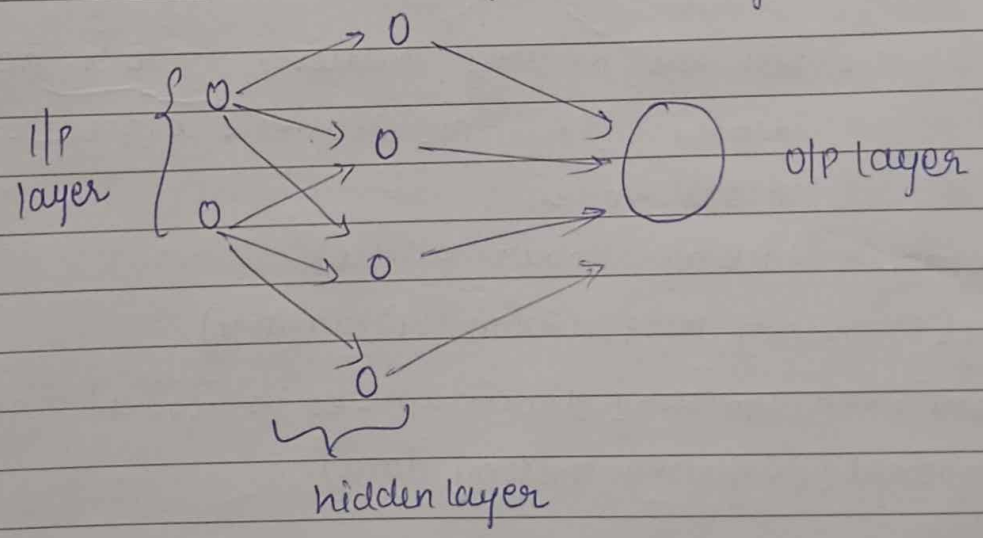
support → Both together is happening

confidence → If a student buy's a lap with the antivirus

(v) Anamoly detection / outlier analysis → Aims to identify rare (or) unusual data instances that deviates significantly from the expected pattern.

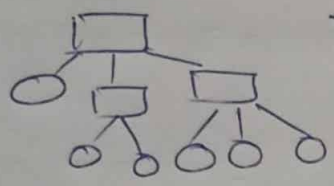
(vi) Time series analysis → (over the period / time related data) Focuses on analyzing and predicting data points collected over time. It involves forecasting, trend analysis etc...

(vii) Neural NW → Interconnected nodes and layers that can learn from data to recognise patterns, performs the classification, regression.



Teacher's Signature:

(viii) Decision tree -> Based on tree line structure we are taking the decisions.



(ix) Ensemble methods -> (combine multiple models) to improve the prediction accuracy and generalisation

(x) Text mining -> (Applied in social media's) using any kind of data we are coming with new kind of data

↳ Applications of DM:

* Intrusion detection, Business Transaction, market-basket analysis, Education, Research, Healthcare & Insurance, Transportation, Finance & Banking sector, scientific analysis

Q. Data preprocessing :-

↳ Data cleaning, Data Integration, Data reduction, Data transformations.

(a) Data cleaning -> to remove noise and correct inconsistencies in data.

-> Fill the missing values, smooth noisy data, identify (or) remove outliers & resolve the inconsistencies.

- Incomplete (lacking attribute values)
- noisy (containing noise, errors (or) outliers)
- Inconsistent (containing discrepancies in codes/names)
- Intentional (disguised missing data)

Date :

Page No. :
Expt. No. :

(i) Incomplete (missing data):
* Data is not available always
* Data not entered due to misunderstandings
* changes of the data

How to handle missing data?

Ans: Ignore tuple > Fill the missing value manually > Fill it automatically with a global constant, attribute mean, the most probable value.

(ii) Noisy (data).
* Random error.

How to handle noisy data?

Ans: Binning method > Regression (linear, multilinear) > clustering > combined human and computer.

=>> Binning method;
↳ smooth a sorted data value by seeing its values around it.
↳ sorted values are distributed into bins
↳ smoothing by bins mean, median, boundaries

Eg: 4, 8, 15, 21, 21, 24, 25, 28, 34.

① Partition into equal bins

Bin 1 : 4 8 15
Bin 2 : 21 21 24
Bin 3 : 25 28 34

② Smoothing bin means

Bin 1 : 9 9 9
Bin 2 : 22 22 22
Bin 3 : 29 29 29

③ Smoothing by bin boundaries;

Bin 1 : 4 4 15
Bin 2 : 21 21 24
Bin 3 : 25 25 34

Teacher's Signature:

=>> Regression (linear, multiple linear)

↓ One attribute can be used to predict

↓ more than 2 attributes.

=>> clustering -> outliers may be detected by using this.

~ b) Data Integration -> Integration of multiple databases, files

-> Merging the data from multiple datastores

- Entity Identification problem
- Redundency and correlation analysis.

=>> correlation analysis (chi-square test)

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

2x2 contingency table

	M	F	Total
fiction	250(90)	200(360)	450
non-fiction	50(210)	1000(840)	1050
Total	300	1200	1500

$$e_{11} = \frac{S \times P}{1500} = \frac{300 \times 450}{1500}$$

$$e_{11} = 90$$

$$e_{12} = \frac{S \times P}{1500} = \frac{1200 \times 450}{1500} = 360$$

$$e_{21} = \frac{300 \times 1050}{1500} = 210$$

$$e_{22} = \frac{1200 \times 1050}{1500} = 840$$

$$\frac{(o-e)^2}{e} = \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$

$$= 284.4 + 71.1 + 121.9 + 30.47$$

$$= 507.93$$

Given value > correlated

independent ignore.

Date :

Page No. :
Expt. No. :

=>> covariance

$$cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

(2,5) (3,8) (5,10) (4,11) (6,14)

$E(A) = 2+3+5+4+6/5 = 4$

$E(B) = 5+8+10+11+14/5 = 9.6$

$\Rightarrow [2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14]/5 - 4\times9.6$

$\Rightarrow 4$

(+)ve covariance → $cov_{A,B} > 0$ A,B both larger than expected value

(-)ve " " → " " < 0 ~~lesser~~ A is > than its expected value, B is smaller.

Independence → $cov_{A,B} = 0$.

c) Data reduction → to reduce the representation of data.

Data reduction strategies:

(i) Dimensionality reduction
- * wavelet transforms
- * PCA
- * Attribute subset selection (By methods of attributes)

(ii) Numerosity reduction
- * parameteric methods
- * Non-parameteric methods

(iii) Data compression
- * Lossy
- * Looseless

Teacher's Signature:

=>> Wavelet transforms

56 40 8 24 48 48 40 16

1st operation: Data smoothing (Average)
2nd " ": WD (weighted difference)

56 40 8 24 48 48 40 16

48 16 48 28 8 -8 0 12

Average WD

(56-48), (8-16), (48-48), (40-28)

32 38 16 10 8 -8 0 12

35 -3 16 10 8 -8 0 12

48 16 48 28 8 -8 0 12

56 40 8 24 48 48 40 16

=>> parameteric.

4 diff - sampling;

1. Simple random sampling
2. Sampling without replacement
3. " " with " "
4. Stratified sampling.

d) Data transformation -> data's are scaled to fall within small range 0.0 to 1.0.

Methods: smoothing (Remove noise from data)
Attribute/feature construction (new attributes added)
Aggregation (combining the data)
Normalization { min-max, z-score, normalization by decimal scaling.
Discretization.
↓
changing numeric value into nominal value

Date :

Page No. :
Expt. No. :

=> Normalization.

a) min - max

$$v' = \frac{v - min_v}{Max - Min} (New\ Max - New\ Min) + new\ Min.$$

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Min - Max normalization value 35 of age [0.0, 1.0]

$$Formula = \frac{value - min}{Max - min}$$

$$\Rightarrow \frac{35 - 13}{70 - 13} (1 - 0) + 0.$$

$$\simeq 0.386$$

b) Z-score normalization.

$$value \leftarrow \frac{v - \mu}{\sigma} \qquad \mu \rightarrow mean \quad , \sigma = 12.94$$

$$\sigma^2 = \frac{1}{N} \Sigma (x_i - \bar{x})^2 \qquad \sigma = \sqrt{\frac{1}{N} \Sigma (x_i - \bar{x})^2}$$

$$\Rightarrow \frac{35 - 26.56}{12.94} \simeq 0.652.$$

c) Decimal scaling: (max value digit num $j = 2$).

Here, 70 so $j = 2$.

$$\text{normalized value } (v') = \frac{v}{10^j} = \frac{35}{10^2} = \frac{35}{100}$$

$$\Rightarrow 0.35$$

Teacher's Signature:

10. AOI :-

* proposed in 1989
* Data cube based on materialized view of data
* AOI based on query oriented, generalization based online data analysis technique.
* First collect task relevant data and perform generalization by attribute removal (or) attribute generalization.
* Aggregation is performed by merging identical tuples and accumulating their counts.

11. Logistic regression derivation :- (sigmoid function)

$$\text{odds}(\theta) = \frac{\text{prob. of event happening}}{\text{prob. of event not happening}}$$

$$\text{odds}(\theta) = \frac{p}{1-p}$$

odds values ranges from $0 \to \infty$

$$Y = \beta_0 + \beta_1 x$$

Taking log on both sides,

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

Exponenting both sides,

$$e^{\ln \frac{p(x)}{1-p(x)}} = e^{\beta_0 + \beta_1 x}$$

Date :

Page No. :

Expt. No. :

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

Let $y = e^{\beta_0+\beta_1 x}$

Then $\frac{p(x)}{1-p(x)} = y$

$P(x) = y[1-P(x)]$

$P(x) = y - y(P(x))$

$P(x) + y(P(x)) = y$

$y = P(x) + y[P(x)]$

$y = P(x)[1+y]$

$P(x)[1+y] = y$

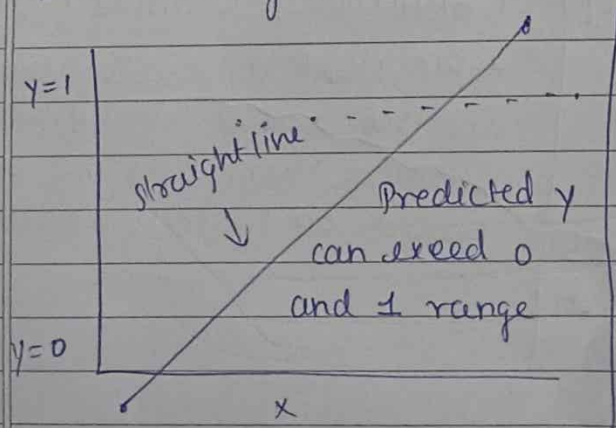
$$P(x) = \frac{y}{1+y}$$

$$P(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

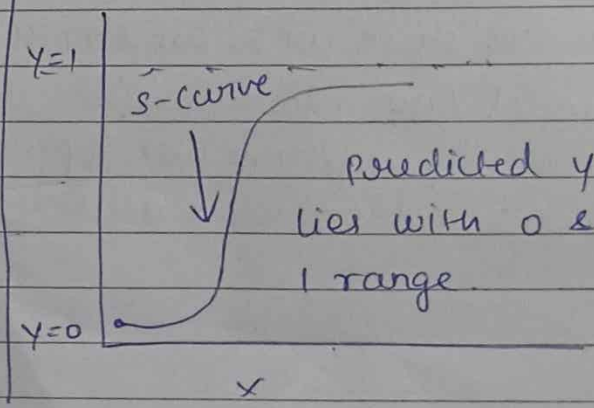
$\div e^{\beta_0+\beta_1 x}$

$$\Rightarrow \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Linear regression



Logistic regression



Teacher's Signature:

Linear regression	Logistic regression
* calculating least square R^2 coefficient method	* Event happening (or) not
* continous variable	* catagorical variable
* predict numeric value	* predicts prob. score
* Regression problem	* classification problem

12. Multiple logistic regression (MLR):

* used to predict binary outcome (yes/No, 0/1) based on multiple predictor variables.
* the regression equation is;

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where, $p \rightarrow$ prob happening

$\beta_0 \rightarrow$ Intercept

$\beta_1 \cdots \beta_k \rightarrow$ coefficients for predictor variables $X_1, X_2 \ldots$

Selection methods for MLR:

Forward selection method

* We start with an empty model (no predictors).
* Add predictors one by one based on some criteria

Backward elimination method

* We start with all predictors in the model and gradually remove them one by one based on certain criteria.

→ Steps in FSM:

a) Start with no predictors

b) Add predictors one by one.

↳ For each step, calculate the contribution of each predictor not yet in the model.

↳ Use metrics like likelihood ratio test etc...

c) Include the most significant predictor.

↳ Add the predictor with the highest significance.

d) Repeat untill NO significant predictors remain.

Advantages-

* when there are many predictors when want to find most ones.
* Avoid overfitting

Disadvantages

* can miss interactions

↓

Starting with large model.

→ Steps in BEM:- → small no. of predictors

a) Start with all predictors.

b) Remove predictors one by one.

↳ Identify the least significant predictor

↳ Remove the least significant predictor from the model.

c) Re-fit the model.

↳ After removing a predictor, refit the model and assess the impact.

d) Repeat until only significant predictors Remain.

↳ continue removing predictors untill all remaining predictors are significant.

Forward selection	Backward elimination
* Starts with no predictor	* Starts with all predictors
* Add predictor one by one	* Removes predictor one by one
* when there are many predictors	* no. of predictors are managable
* Low risk	* High risk
* Lower as small	* Higher it starts with the full model

13. Non-linear regression:

* It is a statistical technique used to model complex relationship btw variables that can't be adequately described by a straight line (or) linear equation

A non-linear regression,

$$Y = f(X, \beta) + \varepsilon$$

where; $f(X, \beta) \rightarrow$ Regression fncn.

$X \rightarrow$ Vector independent.

$\beta \rightarrow$ vector of parameters.

$\varepsilon \rightarrow$ Error term.

Semi-parameteric models:

Date :

Page No. :

Expt. No. :

Teacher's Signature:

14. Yule-walker Estimation :-

$T \rightarrow$ covariance matrix

$\lambda \rightarrow$ covariance vector

$\gamma_0 \rightarrow$ variance.

In order to find correlation, 1st have to find covariance.

consider AR eqn. of $p=2$

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

$(\times) \rightarrow y_{t-k}$ on both sides;

$$y_t \times y_{t-k} = \phi_1 y_{t-k} y_{t-1} + \phi_2 y_{t-k} y_{t-2} + \varepsilon_t y_{t-k}$$

Exceptation both sides;

$$E(y_t y_{t-k}) = \phi_1 E(y_{t-k} y_{t-1}) + \phi_2 E(y_{t-2} y_{t-k}) + E(\varepsilon_t y_{t-k})$$

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2}$$

$\div$ by γ_0 (variance)

$$\frac{\gamma_k}{\gamma_0} = \phi_1 \frac{\gamma_{k-1}}{\gamma_0} + \phi_2 \frac{\gamma_{k-2}}{\gamma_0}$$

$$\boxed{\Rightarrow \rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2}}$$

$\rho \rightarrow$ correlation

Let $k=1 \Rightarrow \rho_1 = \phi_1 \rho_0 + \phi_2 \rho_{-1}$ $\quad (\rho_{-1}) = (\rho_1)$

$$= \phi_1 \rho_0 + \phi_2 \rho_1$$

$$\Rightarrow \rho_1 = \phi_1 + \phi_2 \rho_1 \rightarrow (1)$$

$$\rho_1 - \phi_2 \rho_1 = \phi_1$$

$$\rho_1 (1 - \phi_2) = \phi_1 \Rightarrow \rho_1 = \frac{\phi_1}{1-\phi_2}.$$

$k=2 \Rightarrow \rho_2 = \phi_1 \rho_0 + \phi_2 \rho_{-1}$

$$\cancel{= \phi_1 \rho_0 + \phi_2 \rho_1}$$

Date :

Page No. :
Expt. No. :

$$P_2 = \phi_1 P_1 + \phi_2$$

$$P_2 = \phi_1 \left(\frac{\phi_1}{1-\phi_2} \right) + \phi_2$$

$$\Rightarrow \frac{\phi_1^2 - \phi_2^2 + \phi_2}{1 - \phi_2}$$

15. Least square estimation :

Goal:

16. GLM :-

* when error distribution is not normal we use GLM.

MLR ;

$$Y = X\beta + \epsilon$$

$$E(Y) = X\beta \; ; \; E(\epsilon) = 0$$

here $\boxed{(n(Y)) = X\beta}$

* In GLM, the response variable distribution must be a member of the exponential family.

The exponential Family of distribution :

* A random variable u that belongs to the exponential family with a single parameter θ, pdf (prob. density fnctn)

$$f(u, \theta) = s(u)\, t(\theta)\, e^{a(u)\, b(\theta)}$$

where, s, t, a, b are all known functions

$$f(u, \theta) = \exp\{a(u)\, b(\theta) + d(u) + c(\theta)\}$$

where, $d(u) = \ln(s(u))$ $\quad c(\theta) = \ln(t(\theta))$

where, $a(u) = u$, the distribution is said to be in canonical form and $b(\theta)$ is called natural parameter.

Parameters other than the parameter of interest θ are called nuisance parameters.

a) Normal distribution $N(\mu, \sigma^2)$:-

pdf of normal distribution,

$$f(u, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-1/2\left(\frac{u-\mu}{\sigma^2}\right)}$$

Date $E(u) = \lambda\beta$ $V(u) = 0$

Page No. :
Expt. No. :

$$= \exp\left\{ u \frac{\mu}{\sigma^2} + \left[\frac{-\mu^2}{2\sigma^2} - \frac{1}{2} \ln 2\pi\sigma^2 \right] - \frac{u^2}{2\sigma^2} \right\}$$

$a(u) = u$, $b(\theta) = \frac{\mu}{\sigma^2} = \mu$, $c(\theta) = \frac{-\mu^2}{2\sigma^2} - \frac{1}{2} \ln 2\pi\sigma^2$

$d(u) = \frac{-u^2}{2\sigma^2}$

b) Binomial distribution $Bin(n, p)$:

p - parameter of interest
n - nuisance parameter

$$f(u, p) = \binom{n}{u} p^u (1-p)^{n-u}$$

$$= \binom{n}{u} \left(\frac{p}{1-p}\right)^u (1-p)^n$$

$$= \exp\left\{ u \ln\left(\frac{p}{1-p}\right) + n \ln(1-p) + \ln\binom{n}{u} \right\}$$

$a(u) = u$, $b(\theta) = \ln\left(\frac{p}{1-p}\right)$, $c(\theta) = n \ln(1-p)$, $d(u) = \ln\binom{n}{u}$

c) Poission distribution $P(\lambda)$:

$$f(u, \lambda) = \frac{e^{-\lambda} \lambda^u}{u!}, \quad u = 0, 1, \ldots$$

$$= \exp\left\{ u \ln \lambda - \lambda - \ln u! \right\}$$

$a(u) = u$, $b(\theta) = \ln \lambda$, $c(\theta) = -\lambda$, $d(u) = -\ln u!$

Teacher's Signature:

d) exponential distribution:

$$f(u, \theta) = \theta e^{-\theta u}$$

$$= \exp\{-u\theta + \ln \theta\}$$

$a(u) = u$, $b(\theta) = \theta$.

e) -ve binomial distribution:

$$f(u, \theta) = \binom{r+u-1}{r-1} \theta^{r} (1-\theta)^{u}$$

$$= \exp\left\{u \ln(1-\theta) + r \ln \theta + \ln \binom{r+u-1}{r-1}\right\}$$

$a(u) = u$, $b(\theta) = \ln(1-\theta)$, $c(\theta) = r \ln \theta$, $d(u) = \ln \binom{r+u-1}{r-1}$

17. Non-linear regression:

NLR can be expressed as

$$Y = f(x, \beta) + \varepsilon$$

$f(x, \beta)$ - Reg function.

$x \rightarrow$ vector ind. var to predict dep var.

$\beta \rightarrow$ vector of parameter

$\varepsilon \rightarrow$ Error term.

a) polynomial regression:

* Powerful non-linear technique
* capture more complex relationship b/w ind & dep var.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n$$

$x, x^2, x^3 \rightarrow$ ind. var, $\beta_0, \beta_1, \beta_3 \rightarrow$ parameter

$n \rightarrow$ degree of poly.

b) Exponential regression :-

* Models data following an exp. growth

$$y = \alpha e^{(\beta x)}$$

$\alpha, \beta \rightarrow$ parameters.

c) Logistic regression :-

-> Binary class
-> Prob estimation
-> Non-linear relationship

18. Semi-parameteric regression model :-

* These are statical models that combine both parameteric and non-parameteric components.
* More flexibility.

Addictive Regression model :-
* Specific type of semiparameteric model where the relationship b/w the response variable and predictors modeled as sum of functions, each depending on a single predictor variable.
* Additive model allows non-linear relationship b/w each predictor and the response.

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \epsilon$$

$\beta_0 \rightarrow$ intercept

$f_1, f_2, f_3 \rightarrow f_p$ are smooth, possibly non-linear functions of the predictors $x_1, x_2 \ldots x_p$

* These are typically typically fitted using techniques such as smoothing splines.

Non-parameteric regression methods:

* It doesn't assume a predefined functional form for relationship btw the response variable and predictors.
* ~~Useful when there is little prior knowledge about the form little~~
* when the relationship is too complex to be captured by parameteric models.

common non-parameteric methods,

a) Kernel regression
b) Smoothing splines
c) Local regression

19. Dissimilarity matrix:

Object	Attribute 1 Nominal	Attribute 2 ordinal	Attribute 3 Numerical
1	A1.	Per → 1	23
2	A2.	Good → 2	45
3	A1.	Poor → 3	78
4	A3.	Perfect → 1	31

P - 1
G - 2
Por - 3

Date :

Page No. :
Expt. No. :

Nominal

$$\begin{array}{c|cccc} & A_1 & A_2 & A_1 & A_3 \\ \hline A_1 & 0 & & & \\ A_2 & 1 & 0 & & \\ A_1 & 0 & 1 & 0 & \\ A_3 & 1 & 1 & 1 & 0 \end{array} \rightarrow (1)$$

not same → 1
same → 0

ordinal:-

$$Z_{ij} = \frac{r_{ij} - 1}{M_{ij} - 1}$$

r_{ij} → rank
M_{ij} → max rank.

$$Z_{ij}\ (\text{Perfect}) = \frac{1-1}{3-1} = 0$$

$$(\text{Good}) = \frac{2-1}{3-1} = 0.5$$

$$(\text{Poor}) = \frac{3-1}{3-1} = 1$$

$$\begin{array}{c|cccc} & (0)\ Pe & (0.5)\ Go & (1)\ Po & (0)\ Pe \\ \hline (0)\ Pe & 0 & & & \\ (0.5)\ G & 0.5 & 0 & & \\ (1)\ Poo & 1 & 0.5 & 0 & \\ (0)\ Pe & 0 & 0.5 & 1 & 0 \end{array} \rightarrow (2)$$

Teacher's Signature:

Numerical.

$$\begin{array}{c|cccc} & 23 & 45 & 78 & 31 \\ \hline 23 & 0 & & & \\ 45 & 22 & 0 & & \\ 78 & 55 & 33 & 0 & \\ 31 & 8 & 14 & 47 & 0 \end{array}$$

÷ by 55

$$\begin{array}{c|cccc} & 23 & 45 & 78 & 31 \\ \hline 23 & 0 & & & \\ 45 & 0.4 & 0 & & \\ 78 & 1 & 0.6 & 0 & \\ 31 & 0.15 & 0.25 & 0.85 & 0 \end{array} \rightarrow (3)$$

Dis-matrix.

$$\frac{(1)+(2)+(3)}{3}.$$

Binary :-

$$d_{ij} = \frac{r+s}{q+r+s}$$

$q \rightarrow 1,1$

$r \rightarrow 1,0$

$s \rightarrow 0,1$

A, B

B, C

C, A.

Date :

Page No. :
Expt. No. :

CIA - 1

1. A) Buy/computer Middleage senior value = 3.84

yes	120 (0.36)	3020	3140	0.92
No	4600 (1.07)	13784	18384	1.63
	4720	16804	21524	

$$x^2 = \frac{(o-e)^2}{e}$$

$$\frac{(120-0.36)^2}{0.36} + \frac{(4600-1.07)^2}{1.07} + \frac{(3020-0.92)^2}{0.92} + \frac{(13784-1.63)^2}{1.63}$$

$$\Rightarrow \frac{(119.64)^2}{0.36} + \frac{(4598)^2}{1.07} + \frac{(3019.08)^2}{0.92} + \frac{(13782.37)^2}{1.63}$$

$$\Rightarrow 39760.36 + 19758508.41 + 9907439.181 + 116536026.3$$

$\Rightarrow$

1. A)

Buys	MA	sen	
yes	120 (688.5)	3020 (2451.4)	3140
NO	4600 (4031.4)	13784 (14352.5)	18384
	4720	16804	21524

$$\frac{(120-688.5)^2}{688.5} + \frac{(3020-2451.4)^2}{2451.4} + \frac{(4600-4031.4)^2}{4031.4} + \frac{(13784-14352.5)^2}{14352.5}$$

$$\Rightarrow \frac{(-568.5)^2}{688.5} + \frac{(568.6)^2}{2451.4} + \frac{(568.6)^2}{4031.4} + \frac{(-568.5)^2}{14352.5}$$

Teacher's Signature:

$\Rightarrow 469.41 + 131.88 + 80.19 + 22.5$

$\Rightarrow 704$

B) Handling the missing data

↳ Ignore tuple > Fill the missing values manually >

~~use~~ Fill it automatically with a global constant, by attribute mean, the most probable value.

2. A) Apriori

sup-count = 2

sup = 25.1. conf = 60.1.

tid	Items
A	
B	
C	
D	
E	
F	
G	

(L=) tid	N
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

	Element	sup. count
L1 ⇒	I_1	6
	I_2	7
	I_3	6
	I_4	2
	I_5	2

Date :

Page No. :
Expt. No. :

Generate C2 from L1;

C2 =>		
	I1, I2	4
	I1, I3	4
	I1, I4	1 ×
	I1, I5	2
	I2, I3	4
	I2, I4	2
	I2, I5	2
	I3, I4	0 ×
	I3, I5	1 ×
	I4, I5	0 ×

L2 =>		
	I1, I2	4
	I1, I3	4
	I1, I5	2
	I2, I3	4
	I2, I4	2
	I2, I5	2

C3 =>		
	I1, I2, I3	2
	I1, I2, I5	2
	I1, I2, I4	1 ×
	I1, I3, I5	1 ×
	~~I1, I3, I2~~	×
	I2, I3, I4	0 ×
	I2, I3, I5	1 ×

L3 => I1, I2, I3 -> 2
I1, I2, I5 -> 2

I1, I2, I5 -> support => $\frac{2}{9}$ = 0.2

confidence => I1, I2, I5 -

I1 -> I2, I5 = 8/L = 2/6 =
I2 -> I1, I5 = 2/7 =
I5 -> I1, I2 = 4/2 =
I1, I2 -> I5 = 2/4 =
I2, I5 -> I1 = 6/2 =

Teacher's Signature:

I1, I5 -> I2 = 7/2 =

3. FP-growth

T_1	I_1, I_2, I_3
T_2	I_2, I_3, I_4
T_3	I_4, I_5
T_4	I_1, I_2, I_4
T_5	I_1, I_2, I_3, I_5
T_6	I_1, I_2, I_3, I_4

sup = 50%

$\frac{50}{100} \times 6 = 3$

Item	frequency	
I_1	4	4
I_2	5	5
I_3	~~4~~	4
I_4	~~3~~	4
I_5	2 ×	2

Frequent patterns are;

$L = \{I_2:5, I_1:4, \cancel{I_4:4}, I_3:3\}$

$I_4:4, I_3:3$

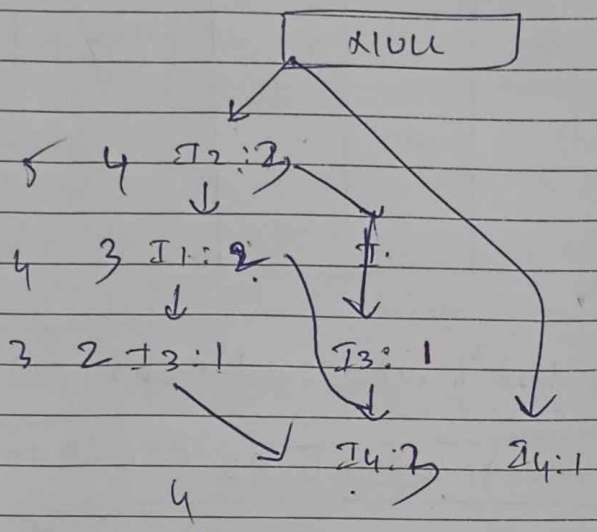
Tid	items	ordered Item set
T_1	I_1, I_2, I_3	I_2, I_1, I_3
T_2	2, 3, 4	I_2, I_3, I_4
T_3	4, 5	I_4
T_4	1, 2, 3, 5	I_2, I_1, I_4
T_5	1, 2, 3, 4	I_2, I_1, I_3, I_4
T_6	1, 2, 3, 4	I_2, I_1, I_3, I_4

$L = \{I_2:5, I_1:4, I_4:4, I_3:3\}$.

Date :

Page No. :
Expt. No. :

Conditional patterns tree;



compute conditional pattern base;

Items	Conditional Pattern Base
I_4	$\{(I_2, I_3 : 1), (I_2, I_1 : 1), (I_2, I_1, I_3 : 1)\}$
I_3	$\{(I_2, I_1 : 3), (I_2 : 1)\}$
I_1	$\{(I_2 : 4)\}$
I_2	×

Items	CPB	CFP Tree
I_4	✓	$\{I_2 : 3\}$
I_3	✓	$\{I_2 : 4\}$
I_1	✓	$\{I_2 : 4\}$
I_2	×	×

Teacher's Signature:

4. A)

X	O_1	O_2	O_3	O_4
O_1	0	79.9		
O_2	79.9 ~~80~~	0		
O_3			0	
O_4				0

Euclidean $= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

Man $= |(x_2-x_1)+(y_2-y_1)|$

$$(O_1, O_2) = \sqrt{(45-12)^2+(12-23)^2+(45-12)^2+(87-23)^2}$$

$$= \sqrt{1089+121+1089+4096} = \sqrt{6395} = 79.9$$

$$(O_1, O_3) = \sqrt{(12-12)^2+(45-23)^2+(21-12)^2+(45-23)^2}$$

$$= \sqrt{\quad} =$$

$$(O_1, O_4) = \sqrt{(24-12)^2+(54-23)^2+(28-12)^2+(37-23)^2}$$

$$= \sqrt{\quad} =$$

$$(O_2, O_3) = \sqrt{(45-12)^2+(45-12)^2+(21-45)^2+(45-87)^2}$$

$$= \sqrt{\quad} =$$

$$(O_2, O_4) = \sqrt{(24-45)^2+(54-12)^2+(28-45)^2+(37-87)^2}$$

$$= \sqrt{\quad} =$$

$$(O_3, O_4) = \sqrt{(24-12)^2+(54-45)^2+(28-21)^2+(37-45)^2}$$

Date :

Page No. :
Expt. No. :

$= \sqrt{} = 18.38.$

B) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

a) Min - Max. (0.0 - 1.0) $v = 35$

$$v' = \frac{v - Min}{Max - Min} (\text{new max} - \text{new min}) + \text{new min}$$

$$v' = \frac{35 - 13}{70 - 13} (1 - 0) + 0$$

$$v' = \frac{22}{57} (1) + 0$$

$$v' = 0.386.$$

b) z-score.

$$v' = \frac{v - \mu}{\sigma} \qquad \sigma = 12.94. \quad v = 35$$

$$v' = \frac{35 - 29.96}{12.94} \qquad \frac{5.04}{12.94}$$

$$v' = 0.3895.$$

c) decimal scaling

$$v' = \frac{v}{10^j} = \frac{35}{10^2} = 0.35$$

Teacher's Signature:

CIA-2

1. {Age <= 30, Income = Medium, student = yes, credit = fairy.

Total = 14, yes = 9, No = 5

Age	Y	N
<=30	2/9	3/5
31-40	4/9	0
>40	3/9	2/5

Income	Y	N
High	2/9	2/5
Medium	4/9	2/5
low	3/9	1/5

$yes = \frac{9}{14} = 0.6429$

$No = \frac{5}{14} = 0.3571$

CR	Y	N
F	6/9	2/5
E	3/9	3/5

Student	Y	N
Y	6/9	1/5
N	3/9	4/5

P(Age <= 30 | Buy (c = "yes") = 2/9 = 0.22

P(Age <= 30 | By c = "No") = 3/5 = 0.6

P(Income = med | Buy c = "yes") = 4/9 = 0.44

P(Income = med | Buy c = "No") = 2/5 = 0.4

P(student = yes | Buy c = "yes") = 6/9 = 0.66

P(student = yes | Buy c = "No") = 1/5 = 0.2

P(credit = fais | Buy c = "yes") = 6/9 = 0.66

P(credit = fair | Buy c = "No") = 2/5 = 0.4.

P(x | Buy com = "yes") = 0.22 x 0.44 x 0.66 x 0.66

= 0.0422

P(x | Buy com = "No") = 0.6 x 0.4 x 0.2 x 0.4

= 0.0192

Date :

Page No. :
Expt. No. :

$P(yes) = 0.0422 \times 0.6429 = 0.0271$ ✓
$P(NO) = 0.0192 \times 0.3571 = 0.0069$

2. $K = 3$

Patient	Age	Glucose	Y/N
1	45	85	NO
2	50	90	NO
3	65	150	yes
4	70	160	yes
5	60	130	?

$x_1 \quad y_1$
$(60, 130)$

Manhattan $= |(x_2 - x_1) + (y_2 - y_1)|$.

1 ->	$(45-60) + (85-130)$	$= 60$ -> 4	N
2 ->	$(50-60) + (90-130)$	$= 50$ -> 3	N
3 ->	$(65-60) + (150-130)$	$= 25$ -> 1	Y
4 ->	$(70-60) + (160-130)$	$= 40$ -> 2	Y

$K = 1 \rightarrow Y$
$K = 2 \rightarrow Y$
$K = 3 \rightarrow N$.

$(60, 130) = yes$.

$\hat{y} = 5 \times 1 + 10$

3.

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\hat{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$\times 2$
1	14	-1	-6	1	36	15	6	36
3	24	+1	4	1	16	25	4	16
2	18	0	-2	1	4	20	-2	4
1	17	-1	-3	0	9	15	0	0
3	27	1	7	1	49	25	7	49.
2	20							

Teacher's Signature:

$y = b_0 + b_1 x$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}$$

$\Rightarrow (-1 \times -6) + (1 \times 4) + (\overset{0}{\cancel{-1}}) + (3) + 7$

$\Rightarrow 6 + 4 + 0 + 3 + 7$

$b_1 = \frac{20}{4}$ $\boxed{b_1 = 5}$ $m = 5$

$y = b_0 + b_1 x$

$b_0 = \bar{y} - b_1 \bar{x}$

$b_0 = 20 - 5 \times 2$

$b_0 = 20 - 10$ 5

$\boxed{b_0 = 10}$ 10

$\boxed{y = 10 + 5x}$

$y = b_0 + b_1 x$

$45 = 10 + 5 \times x$

$10 + 5x = 45$

$5x = 35$

$\hat{y} = mx + c$. $5 = \frac{35}{x}$

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{(y - \bar{y})^2}$$

$\hat{y} = mx + c$

(ii) $y = mx + c$

$y = 5 \times 45 + 10$

$\boxed{y = 235}$

45
5
225

$b_1 = -1.4457$

$c = 32.628$

$\hat{y} =$

Date :

Page No. :

Expt. No. :

Sigmoid fncn :

$$\text{Odds } (\theta) = \frac{p}{1-p}$$

$y = \beta_0 + \beta_1 x$ Odd $0 \to \infty$ -

log on both sides

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

$$e^{\ln} \frac{p(x)}{1-p(x)} = e^{\beta_0 + \beta_1 x}$$

$$\frac{p(x)}{1-p(x)} = e^{\beta_0 + \beta_1 x}, \quad y = e^{\beta_0 + \beta_1 x}$$

$$\frac{p(x)}{1-p(x)} = y$$

$$p(x) = y \times (1 - p(x))$$

$$p(x) = y - y[p(x)]$$

$$y = p(x) + y[p(x)]$$

$$y = p(x)[1+y]$$

$$\frac{y}{1+y} = p(x)$$

$$p(x) = \frac{y}{1+y}$$

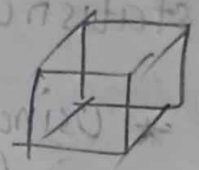
$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$\Rightarrow \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Teacher's Signature:

=> OLAP (Online Analytical Processing):

↳ cube (3D - Array)

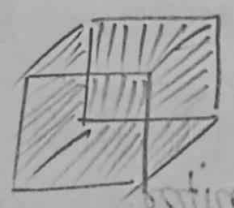


↳ Multidimensional data/model.

Rollup, Drilldown, Slice, Dice, Rotate/pivot.

Operations of OLAP.

(i) Rollup:-



Eg: In TN & Kerala info. of their are products are sold et..., like city wise. By "Rollup" it will give as statewise

* From ~~to~~ small data's to larger data's.

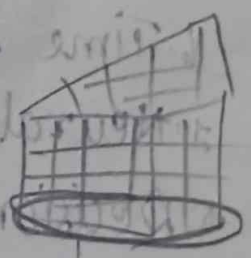
(ii) Drilldown:-

* opp. to rollup.

* From year to month means "drilldown".

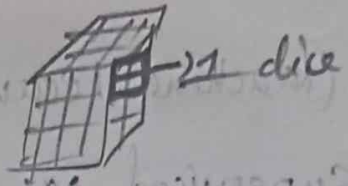
(iii) Slice:

* Taking one slice of portion.



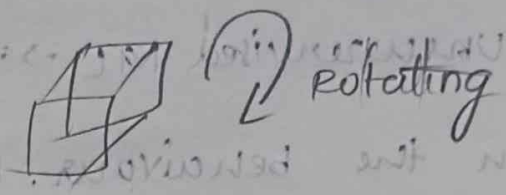
slice

(iv) Dice :



* one size of dice

(v) Rotate :



* Rotating.

=> statistics :

* Using the 54 data (or) collected data's of the previous one to predict the future.

Stages of DM process :

1. Understand the problem & data.
2. Prepare the data.
3. Build the model.
4. Evaluate the result.
5. Implement change & monitor.

DM techniques :

10. Text mining.
1. Classification
2. Regression
3. Clustering
4. Association Rule
5. Anomoly Detection
6. Time series analysis.
7. Neural NW's
8. Decision Tree
9. Ensemble methods.