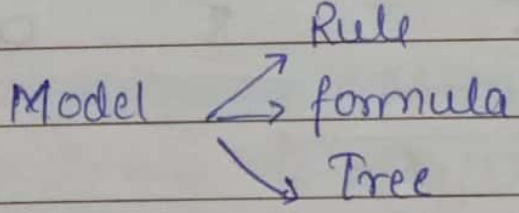
Unit-02

Basic concepts- Classification

Classification Process:- It is a data mining technique / functionality that classifies or categorizes a given test data / given test tuple into a particular category (class)

It has 2 steps:-

1) Model Building step (Training the model)
2) Testing the model or evaluating the performance of the model that is built



Classification accuracy is the measure of accuracy of the classification of a concept that is given by a certain theory.

Scanned with CamScanner

Naive Bayes Classifier

→ It is based on Bayes Theorem

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

- $P(A|B)$ → posterior probability
- $P(B|A)$ → predictor probability given the class
- $P(A)$ → prior probability of the class
- $P(B)$ → prior predictor probability
- A → class label
- B → predictor variables

Data:	Cold	Temp	Body_Pain	Class (F/M)
	Y	H	S	F
	N	H	M	M
	N	M	L	M
	Y	L	L	M

Y-yes N-no

M-moderate L-low

S-severe H-high

F-flu M-malaria.

Test: (yes, Moderate, Moderate)

It is called as naive because

Scanned with CamScanner

i) it assumes all attributes are independent
→ i.e changing value of any one of the attr doesn't change other
ii) it assumes equal priority for all attributes

$$P(F|B) = \frac{P(C|F) * P(T|F) * P(BP|F) \times P(F)}{P(C) \times P(T) \times P(BP)}$$

$$P(M|B) = \frac{P(C|M) \times P(T|M) \times P(BP|M) \times P(M)}{P(C) \times P(T) \times P(BP)}$$

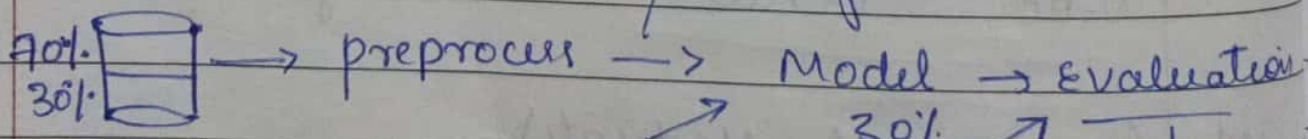
Tuples = 9
Flu = 5
M = 4

Cold: (5: yes)(4: No)

	F	M
yes	4	1
No	1	3

Both for classification & regression

K-nn

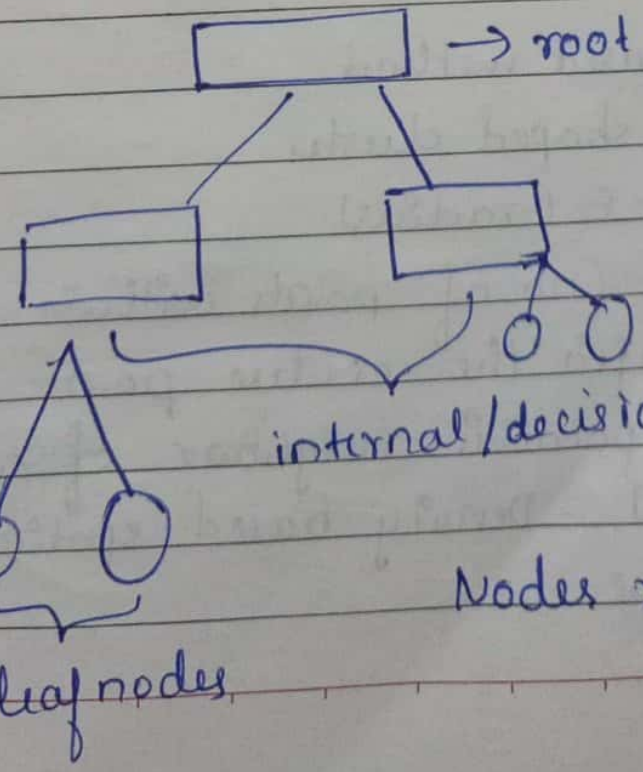


↳ Eages learner
Lazy learner – No training only test.

Scanned with CamScanner

Decision tree Induction

→ Model: Decision tree, mathematical formula, Rule
→ When domain knowledge is not required
→ Construction of decision tree doesn't require domain knowledge
→ We can work on high dimensional data
→ Building of tree is simple & fast
→ Visual representation of tree is easy to understand
→ decision rules infered from the tree " " "



Scanned with CamScanner

* Clustering (Image processing, pattern recognition
→ Unsupervised Business Intelligence, Websearch

Partioning method
→ Grouping of objects takes place iteratively the cluster objects will be changed based on iterative relocation technique
→ Useful for small to med size datasets
→ Data transformation is necessary in pre-processing
Eg: K-means

Hierarchial Method
1) Bottom up 2) Top down.
→ Once grouped can't be undone in BU
Ex: Agglomerative

Density based method
→ Random shaped clusters
→ Epsilon - ε (radius)
Minpts - (no of points within circle of radius
If satisfies the centre point is core point
Border point is neighbor of core point
DBSCAN - Density based spatial clustering app
with noise

Scanned with CamScanner

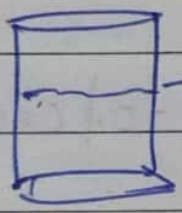
K-NN

→ Lazy learner

→ Data is stored in memory & used only during testing so it's called memory-based classifier

→ It is uses train data when test instance comes into picture. so it's called instance based. classifieder

Eager learners :-



→ preprocessing → Quality data → Model training →

Raw data

Evaluation

↓

accuracy

* Similarity measure / similarity metric

Distance → Manhattan

→ Euclidean

K = no. of neighbors

Scanned with CamScanner

DATE : / /

$d(p,q) = \sqrt{(x_1-x_2)^2+(y_2-y_1)^2}$ Euclid

$d(p,q) = |x_1-x_2| + (y_1-y_2|$ Manhattan

$k = \sqrt{\text{No. of data points}}$

Cons.

It is very slow when no. of records & dimension is more

Whole data in memory (expensive)

Scaling of attributes is needed

Pros

Simple & easy to understand

No assumptions is made and is called non-parametric

App's.

Image Recognition

Handwriting "

Speech "

List of sandioning loan

Scanned with CamScanner

Decision Tree:

Top down recursive divide & conquer - construction of tree.

Information Gain
Gain Ratio
Gini Index.

Decision tree induction.

Attribute selection measure:-

* It minimizes the info or expects no of tests that need to classify a tuple to particular class.

Info gain = Info before partition − Info after partition

(no. of bits used to transfer 'x' from sample)

Entropy
(measure of uncertainty / impurity)

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

$$0 \leq H(x) \leq 1$$

↓ varied distn

↳ impure (large no of attr)

Scanned with CamScanner

DATE: / /

$$H(x) = -\sum_{i=1}^{n} P_i \log_2 P_i$$

N = 14

yes = 9

No = 5.

$$\text{Info}(D) = H(x) = +\left(\frac{9}{14} \log_2 \frac{9}{4} + \frac{5}{14} \log_2 \frac{5}{14} \right) = 0.940$$

$$\text{After partition} = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Info}(D_j)$$

Info gain = Info before partiation / Entropy / H(x) / Info(D) − Info after partition.

$$= -\sum_{i=i}^{n} P_i \log_2 P_i - \left(\sum_{j=1}^{v} \frac{|D_j|}{D} \text{Info}|D_j| \right)$$

For age

Scanned with CamScanner

|D|- Total no of tuple
Dj - No of tuple with attribute

PAGE NO.:
DATE: / /

RID	Age	Income	student	CreditRating	Class
1	Y	H	No	Fair	Yes
2	M-A	M	Yes	Excellent	No
3	S	L			

$$\text{Gain(age)} = H(x) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Info}(D_j)$$

$= 0.940$

Age — Y N — $\text{Info}(2,3) = \frac{5}{14}\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right)$

Youth = 5 — (2, 3)

MA = 4 — (4, 0) — $= \frac{4}{14} - \left(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right)$

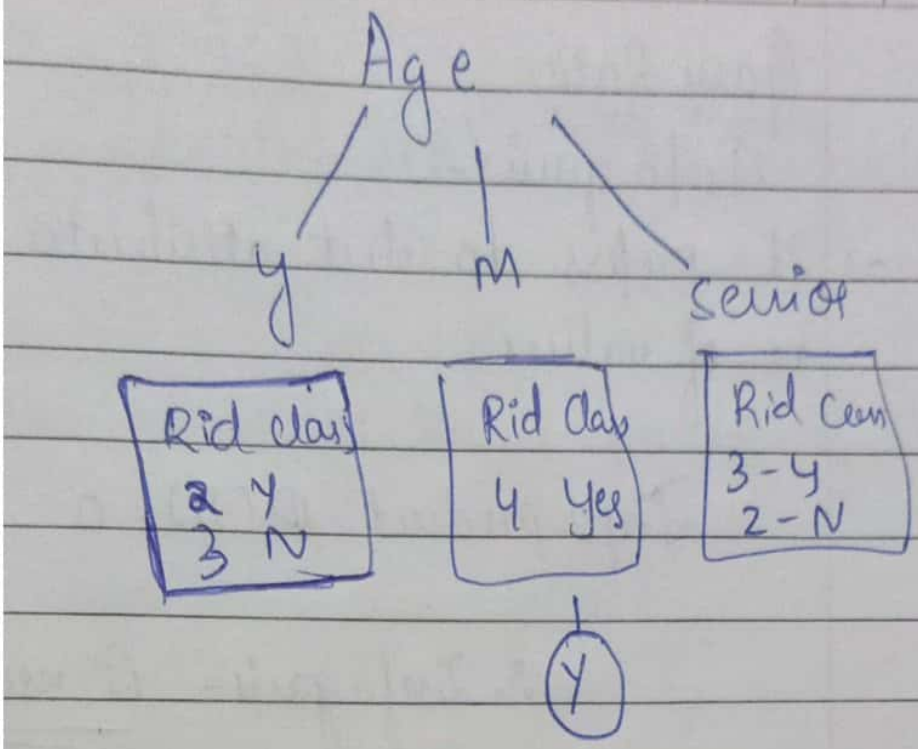
S = 5 — (3, 2)

$= \{0.940 - 0.694\}$ — $= \frac{5}{14} - \left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right)$

$= \{0.246 \text{ bits}\}$

Gain (age) = 0.246 bits (1) - root
Gain (Income) = 0.029 bits
Gain (student) = 0.151 bits (2)
Gain (credit rating) = 0.048 bits

Scanned with CamScanner



Rules.

→ Traverse from root to leaf

→ If ( ) then

sum

{

leaf nodes no = no. of rules.

Accuracy = $\frac{n_{correct}}{n_{covers}}$

$Coverage_{(R_i)} = \frac{n_{covers}}{|D|}$ ← for every rule

↳ Total tuple

↳ Tuples satisfying rule

Scanned with CamScanner

Gain Ratio

Info gain :-

→ It prefers to select attributes having larger no. of values.

Info product - ID(D) = 0

∴ Info gain = is max

but useless

Gain Ratio:

It uses extension of info gain called gain ratio (C4.5 successor of ID3)

$$\text{Split Info}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} \times \log_2 \frac{|D_j|}{D}$$

↳ applies normalization to info gain using a "

Scanned with CamScanner

DATE : / /

GainRatio = $\frac{\text{Gain}(A)}{\text{SplitInfo}_A(D)}$ — o (unstable)

Gini Index: Measures the impurity of D

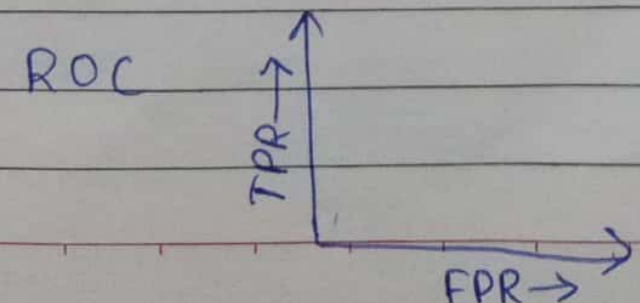
$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{D} Gini(D_2)$$

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

ROC curve

Actual \ Predicted	Y	N
Y	TP	FN
N	FP	TN



Scanned with CamScanner

Sensitivity

$$TPR = \frac{TP}{TP + FN}$$

Specificity

$$TNR = \frac{TN}{FP + TN}$$

$$FPR = \frac{FP}{FP + TN}$$

1 - specificity = FPR

Scanned with CamScanner

UPMP 3) $P(A=1 \mid C=c1)$

$$= \frac{P(c1/A=1)}{P(c1)} = \frac{4/10}{6/10}$$

$$P(A=0 \mid C=c1) = \frac{2/10}{6/10}$$

$$P(B=1 \mid C=c1) = \frac{4/10}{6/10} \qquad P(B=0 \mid C=c1) = \frac{2/10}{6/10}$$

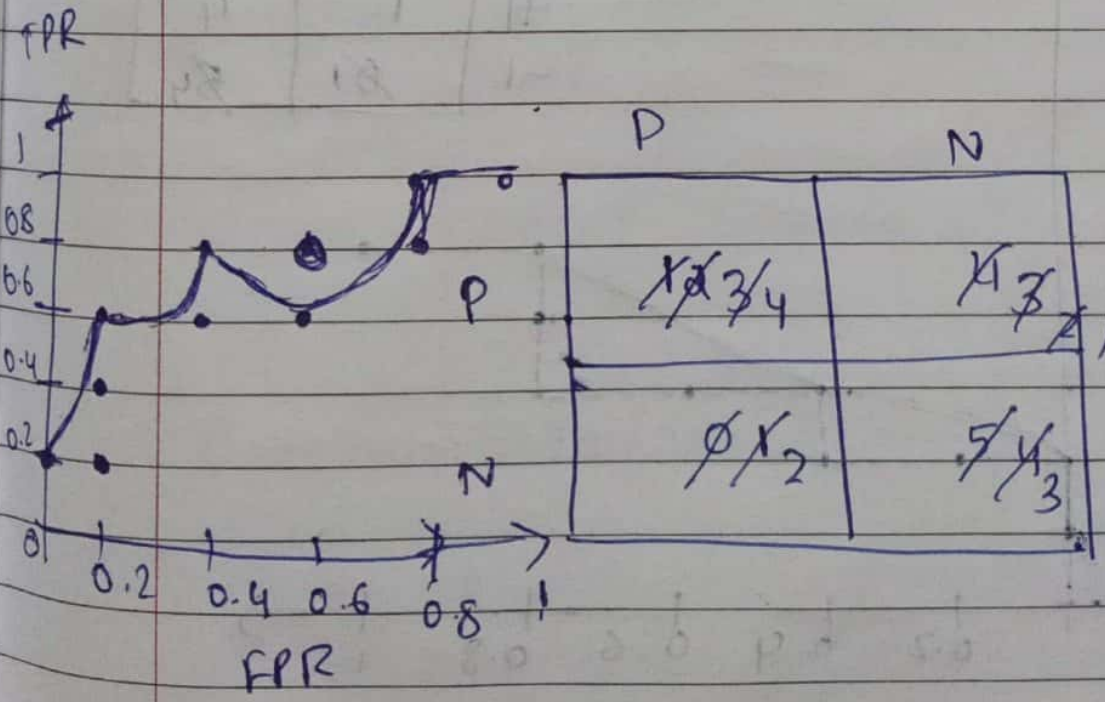
$$P(A=1 \mid C=c2) = \frac{1/10}{4/10} \qquad P(A=0 \mid C=c2) = \frac{3/10}{4/10}$$

$$P(B=1 \mid C=c2) = \frac{0/10}{4/10} \qquad P(B=0 \mid C=c2) = \frac{4/10}{4/10}$$

Scanned with CamScanner

1P3.

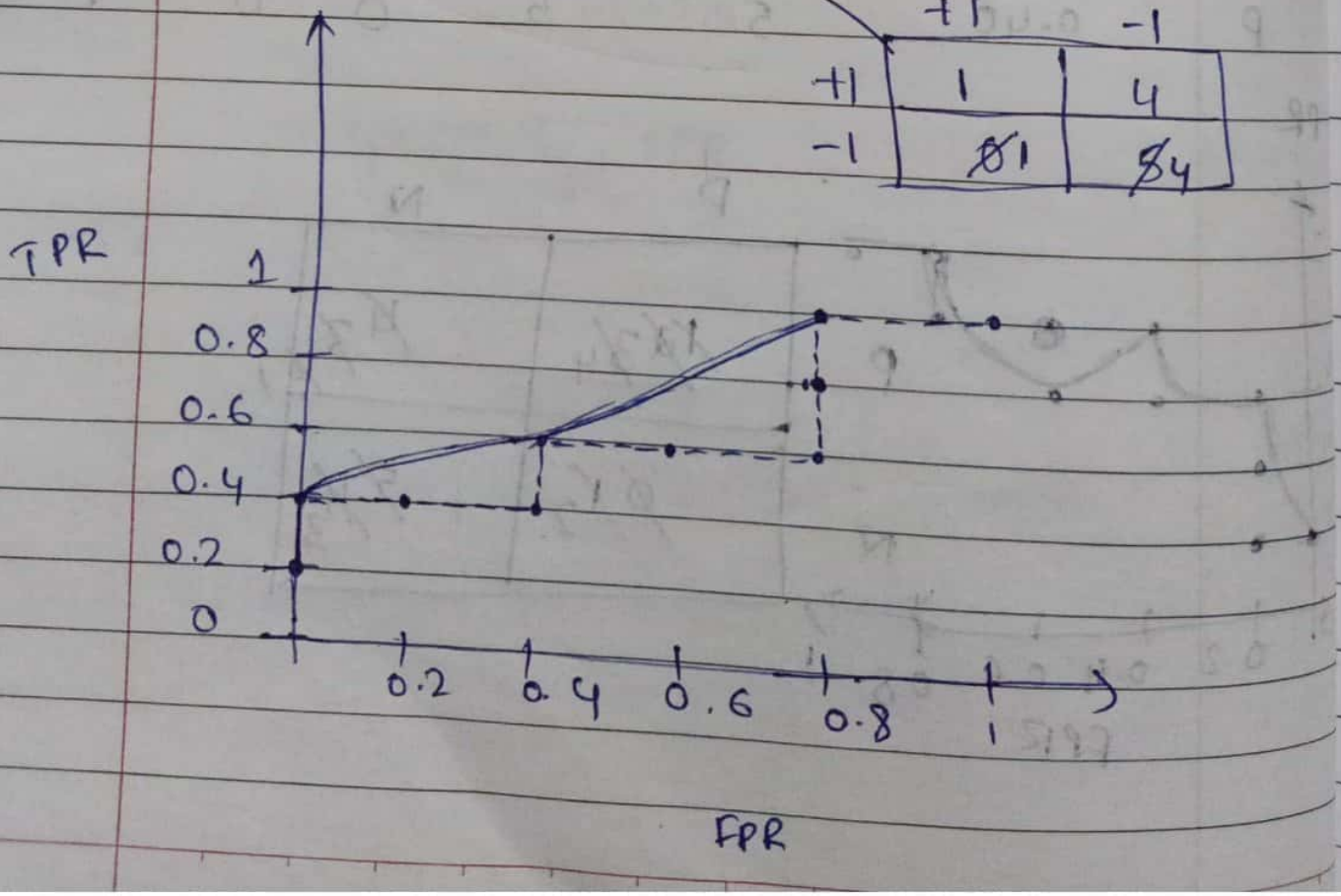
class	prob	TP	FP	TN	FN	TPR	FPR
P	0.95	1	0	5	4	0.2	0
N	0.85	1	1	4	4	0.2	0.2
P	0.78	2	1	4	3	0.4	0.2
P	0.66	3	1	4	2	0.6	0.2
N	0.60	3	2	3	2	0.6	0.4
P	0.55	4	2	3	1	0.8	0.4
N	0.53	4	3	2	1	0.8	0.6
N	0.52	4	4	1	1	0.8	0.8
N	0.51	4	5	0	1	0.8	1
P	0.40	5	5	0	0	1	1



Scanned with CamScanner

P	class	TP	FP	TN	FN	TPR	FPR
0.95	+1	1	0	5	4	0.2	0
0.94	+1	2	0	5	3	0.4	0
0.87	-1	2	1	4	3	0.4	0.2
0.86	-1	2	2	3	3	0.4	0.4
0.86	+1	3	2	3	2	0.6	0.4
0.86	-1	3	3	2	2	0.6	0.6
0.76	-1	3	4	1	2	0.6	0.8
0.53	+1	4	4	1	1	0.8	0.8
0.44	-1	4	5	0	1	0.8	1
0.25	+1	5	5	0	0	1	1

	+1	-1
+1	1	4
-1	~~1~~	~~4~~



Scanned with CamScanner

3b

TP = 100

FN = 40

TN = 300

FP = 60

Precision = $\frac{TP}{\text{predicted yes}}$

$= \frac{TP}{FP + TP}$

Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$

TPR / Sensitivity / Recall $= \frac{TP}{TP + FN}$

Specificity $= \frac{TN}{TN + FP}$

Precision $= \frac{TP}{TP + FP}$

FPR = 1 - specificity $= \frac{FP}{TN + FP}$

f-measure = $2a / 2a + b + c$

a - TP

b - FN

c - FP

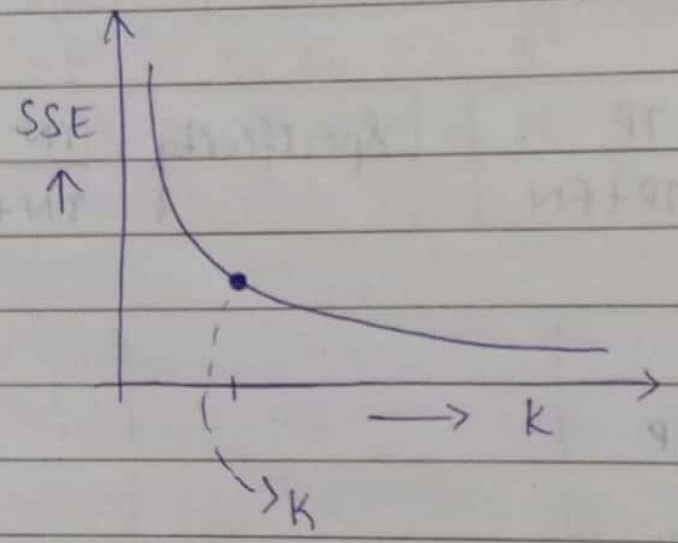
d - TN

Scanned with CamScanner

K-means

→ Initial guess as to cluster centroids
→ Defined distance metric
→ No. of clusters
→ Rescaling is required as distance is calculated

k value can be obtained from plotting K v/s SSE (sum of squared errors)



Scanned with CamScanner

DATE: / /

K-means: Assignment

Manhattan

1. K=3

$C_1 = \{x_1, x_2, x_3\}$ $C_2 = \{x_4, x_5, x_6\}$ $C_3 = \{x_7, x_8\}$

$C_1 = \left\{\left(\frac{2+2+8}{3}\right), \left(\frac{10+5+4}{3}\right)\right\} = C_1 = (4, 6.33)$

$C_2 = \left(\frac{5+7+6}{3}, \frac{8+5+4}{3}\right) = (6, 5.67)$

$C_3 = \left(\frac{1+4}{2}, \frac{2+9}{2}\right) = (2.50, 5.50)$

Iteration 2

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
C_1	5.67	3.33	6.33	2.67	4.33	7.33	7.33	2.67
C_2	8.33	4.67	3.67	3.33	1.67	1.67	8.67	5.3
C_3	5	1	7	5	5	5	5	5
Cluster	C_3	C_3	C_2	C_1	C_2	C_2	C_3	C_1

Calculation

$C_1, x_1 = |4-2| + |6.33-10| = 5.67$

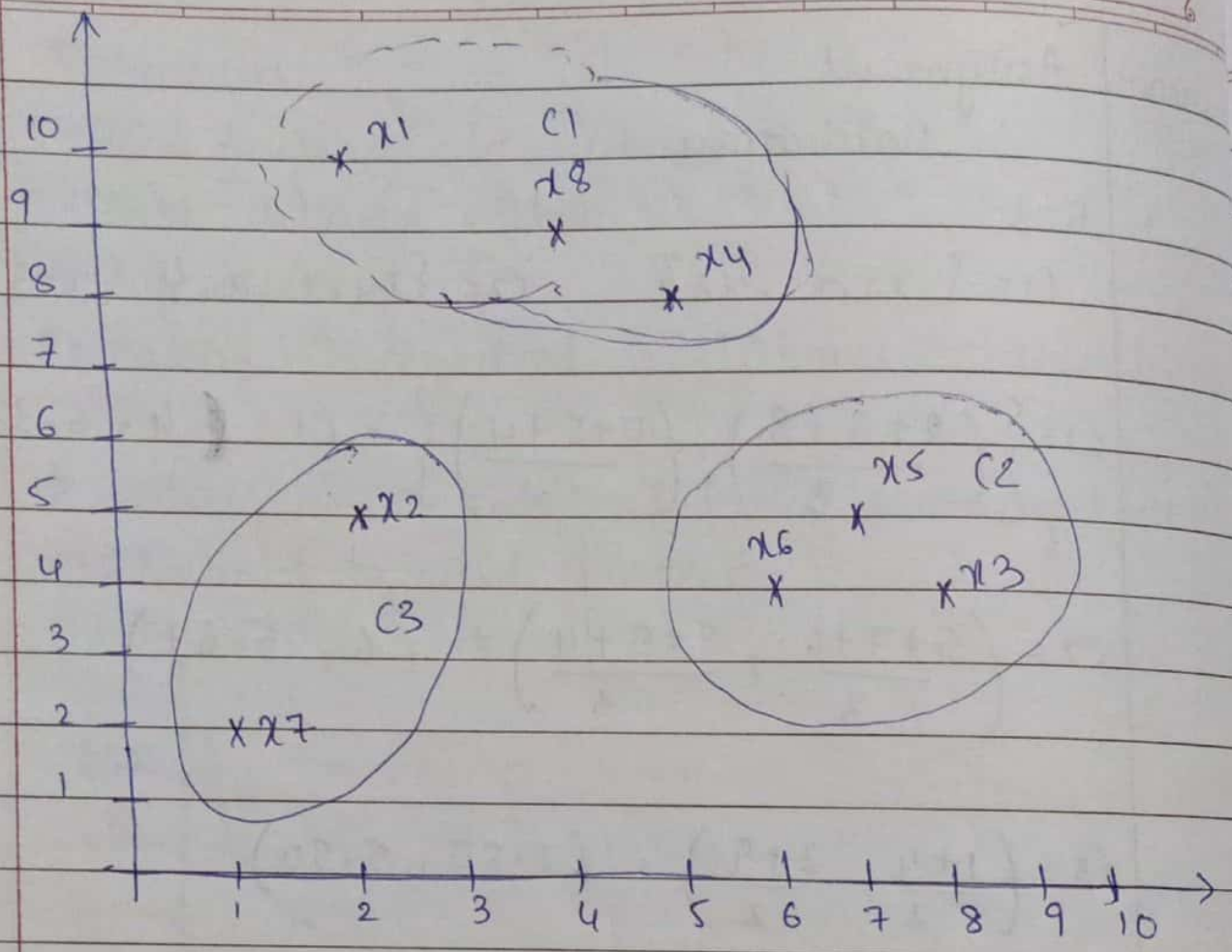
$C_1, x_2 = |4-2| + |6.33-5| = 3.33$

$C_1, x_3 = |4-8| + |6.33-4| = 6.33$

$C_1, x_4 = |4-5| + |6.33-8| = 2.67$

$C_1, x_5 = |4-7| + |6.33-5| = 4.33$

Scanned with CamScanner



$C_1, x_6 = |4-6| + |6.33-4| = 4.33$

$C_1, x_7 = |4-1| + |6.33-2| = 7.33$

$C_1, x_8 = |4-4| + |6.33-9| = 2.67$

$C_2, x_1 = |6-2| + |5.67-10| = 8.33$

$C_2, x_2 = |6-2| + |5.67-5| = 4.67$

$C_2, x_3 = |6-8| + |5.67-4| = 3.67$

$C_2, x_4 = |6-5| + |5.67-8| = 3.33$

$C_2, x_5 = |6-7| + |5.67-3| = 1.67$

$C_2, x_6 = |6-6| + |5.67-4| = 1.67$

Scanned with CamScanner

$C_2, x_7 = |6 - 1| + |5.67 - 2| = 8.67$

$C_2, x_8 = |6 - 4| + |5.67 - 9| = 5.33$

$C_3, x_1 = |2.5 - 2| + |5.5 - 10| = 5$

$C_3, x_2 = |2.5 - 2| + |5.5 - 5| = 1$

$C_3, x_3 = |2.5 - 8| + |5.5 - 4| = 7$

$C_3, x_4 = |2.5 - 5| + |5.5 - 8| = 5$

$C_3, x_5 = |2.5 - 7| + |5.5 - 5| = 5$

$C_3, x_6 = |2.5 - 6| + |5.5 - 4| = 5$

$C_3, x_7 = |2.5 - 1| + |5.5 - 2| = 5$

$C_3, x_8 = |2.5 - 4| + |5.5 - 9| = 5$

Iteration 2

$C_1 = \{x_4, x_8\}$

$C_2 = \{x_3, x_5, x_6\}$

$C_3 = \{x_1, x_2, x_7\}$

$$C_1 = \left(\frac{5+4}{2}, \frac{8+9}{2}\right) = (4.50, 8.5)$$

$$C_2 = \left(\frac{8+7+6}{3}, \frac{4+5+4}{3}\right) = (7, 4.33)$$

$$C_3 = \left(\frac{2+2+1}{3}, \frac{10+5+2}{3}\right) = (1.67, 5.67)$$

Scanned with CamScanner

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
C1	4	6	8	1	6	6	10	1
C2	10.67	5.67	1.33	5.67	0.67	1.33	8.33	7.67
C3	4.66	1	8	5.66	6	6	4.34	5.34
Clusters	C1	C3	C2	C1	C2	C2	C3	C1

$C_1, x_1 = |4.5 - 2| + |8.5 - 10| = 4$

$C_1, x_2 = |4.5 - 2| + |8.5 - 5| = 6$

$C_1, x_3 = |4.5 - 8| + |8.5 - 4| = 8$

$C_1, x_4 = |4.5 - 5| + |8.5 - 8| = 1$

$C_1, x_5 = |4.5 - 7| + |8.5 - 5| = 6$

$C_1, x_6 = |4.5 - 6| + |8.5 - 4| = 6$

$C_1, x_7 = |4.5 - 1| + |8.5 - 2| = 10$

$C_1, x_8 = |4.5 - 4| + |8.5 - 9| = 1$

$C_2, x_1 = |7 - 2| + |4.33 - 10| = 10.67$

$C_2, x_2 = |7 - 2| + |4.33 - 5| = 5.67$

$C_2, x_3 = |7 - 8| + |4.33 - 4| = 1.33$

$C_2, x_4 = |7 - 5| + |4.33 - 8| = 5.67$

$C_2, x_5 = |7 - 7| + |4.33 - 5| = 0.67$

$C_2, x_6 = |7 - 6| + |4.33 - 4| = 1.33$

$C_2, x_7 = |7 - 1| + |4.33 - 2| = 8.33$

$C_2, x_8 = |7 - 4| + |4.33 - 9| = 7.67$

Scanned with CamScanner

PAGE NO.:
DATE: / /

0.33 4.33

$C3,x_1 = |4.67 - 2| + |5.67 - 10| = 4.66$

$C3,x_2 = |4.67 - 2| + |5.67 - 5| = 1$

$C3,x_3 = |4.67 - 8| + |5.67 - 4| = 8$

$C3,x_4 = |1.67 - 5| + |5.67 - 8| = 5.66$ (6.33, 3.33)

$C3,x_5 = |1.67 - 7| + |5.67 - 5| = 6$

$C3,x_6 = |1.67 - 6| + |5.67 - 4| = 6.$ (4.33)

$C3,x_7 = |1.67 - 1| + |5.67 - 2| = 4.34$

$C3,x_8 = |1.67 - 11| + |5.67 - 9| = 5.66$ (2.33, 3.33)

Iteration 3 $C1 = (x_1, x_4, x_8)$ $C2 = (x_3, x_5, x_6)$ $C3 = (x_2, x_7)$

$C1 = (2+5+4/3, 10+8+9/3)$

$C2 = (8+7+6/3, 4+5+4/3)$

$C3 = (2+1/2, 5+2/2)$

$C1 = (3.67, 9)$

$C2 = (7, 4.33)$

$C3 = (1.5, 3.5)$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
C1	2.67	5.67	9.33	2.33	7.33	7.33	9.67	0.33
C2	10.67	5.67	1.33	5.67	0.67	1.33	8.33	7.67
C3	7	2	7	8	7	5	2	8
Cluster	C1	C3	C2	C1	C2	C2	C3	C1

Scanned with CamScanner

$C_1,x_1 = 3.67 - 2 + 9 - 10 = 2.67$

$C_1,x_2 = 11 - 2 + 9 - 5 = 5.67$

$C_1,x_3 = 11 - 8 + 9 - 4 = 9.33$

$C_1,x_4 = 3.67 - 5 + 9 - 8 = 2.33$

$C_1,x_5 = 3.67 - 7 + 9 - 5 = 7.33$

$C_1,x_6 = 3.67 - 6 + 9 - 4 = 7.33$

$C_1,x_7 = 3.67 - 1 + 9 - 2 = 9.67$

$C_1,x_8 = 3.67 - 4 + 9 - 9 = 0.33$

$C_2,x_1 = 7 - 2 + 4.33 - 10 = 10.67$

$C_2,x_2 = 7 - 2 + 4.33 - 5 = 5.67$

$C_2,x_3 = 7 - 8 + 4.33 - 4 = 1.33$

$C_2,x_4 = 7 - 5 + 4.33 - 8 = 5.67$

$C_2,x_5 = 7 - 7 + 4.33 - 5 = 0.67$

$C_2,x_6 = 7 - 6 + 4.33 - 4 = 1.33$

$C_2,x_7 = 7 - 1 + 4.33 - 2 = 8.33$

$C_2,x_8 = 7 - 4 + 4.33 - 9 = 7.67$

$C_3,x_1 = 1.5 - 2 + 3.5 - 10 = 7$

$C_3,x_2 = 1.5 - 2 + 3.5 - 5 = 2$

$C_3,x_3 = 1.5 - 8 + 3.5 - 4 = 7$

$C_3,x_4 = 1.5 - 5 + 3.5 - 8 = 8$

$C_3,x_5 = 1.5 - 7 + 3.5 - 5 = 7$

$C_3,x_6 = 1.5 - 6 + 3.5 - 4 = 5$

$C_3,x_7 = 1.5 - 1 + 3.5 - 2 = 2$

$C_3,x_8 = 1.5 - 4 + 3.5 - 9 = 8$

Scanned with CamScanner

Let no. of clusters = $\sqrt{\frac{n}{2}}$

Each cluster has = $\sqrt{2n}$ points

Scanned with CamScanner

Krishna

Ch-P (HP)

	x	y
⑦ P_1	1	2
P_2	1.5	1.5
P_3	2	1.5
P_4	5	1.5
P_5	4.5	2
P_6	2.5	7
P_7	3	6.5

d = Manhatan

$d(P_1P_2) = |1-1.5| + |2-1.5| = 1$

$d(P_1P_3) = 1 + 0.5 = 1.5$

$d(P_1P_4) = 4 + 0.5 = 4.5$

$d(P_1P_5) = 3.5 + 0 = 3.5$

$d(P_1P_6) = 1.5 + 5 = 6.5$

$d(P_1P_7) = 2 + 4.5 = 6.5$

$d(P_2P_3) = 0.5$

$d(P_2P_4) = 3.5$

$d(P_2P_5) = 3 + 0.5 = 3.5$

$d(P_2P_6) = 1 + 5.5 = 6.5$

$d(P_2P_7) = 1.5 + 5 = 6.5$

$d(P_3P_4) = 3 + 0 = 3$

$d(P_3P_5) = 2.5 + 0.5 = 3$

$d(P_3P_6) = 0.5 + 5.5 = 6$

$d(P_3P_7) = 1 + 5 = 6$

$d(P_4P_5) = 0.5 + 0.5 = 1$

$d(P_4P_6) = 2.5 + 5.5 = 8$

$d(P_4P_7) = 2 + 5 = 7$

$d(P_5P_6) = 2 + 5 = 7$

$d(P_5P_7) = 1.5 + 4.5 = 6$

$d(P_6P_7) = 0.5 + 0.5 = 1$

Scanned with CamScanner

Single linkage

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
P_1	0						
P_2	1	0					
P_3	1.5	0.5	0				
P_4	4.5	3.5	3	0			
P_5	3.5	3.5	3	1	0		
P_6	6.5	6.5	6	8	7	0	
P_7	6.5	6.5	6	7	6	1	0

P_3P_2 → New cluster

	P_1P_2	P_2P_3	P_4	P_5	P_6	P_7
P_1P_2	0					
P_2P_3	1	0				
P_4	4.5	3	0			
P_5	3.5	3	1	0		
P_6	6.5	6	8	7	0	
P_7	6.5	6	7	6	1	0

$P_1P_2P_3$ → New cluster

	$P_1P_2P_3$	P_4	P_5	P_6	P_7
$P_1P_2P_3$	0				
P_4	3	0			
P_5	3	1	0		
P_6	6	8	7	0	
P_7	6	7	6	1	0

P_4P_5 → New

Scanned with CamScanner

	$P_1P_2P_3$	P_4P_5	P_6	P_7
$P_1P_2P_3$	0			
P_4P_5	3	0		
P_6	6	7	0	
P_7	6	6	(1)	0

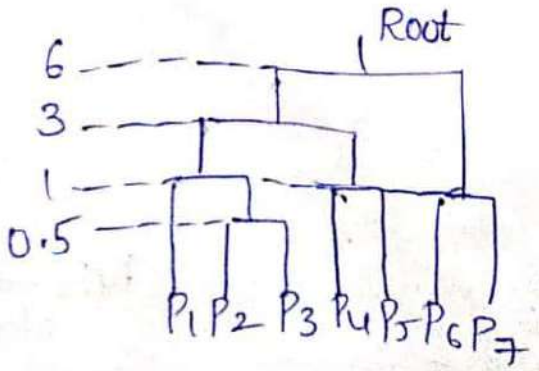
$P_6P_7 \rightarrow$ NIC

	$P_1P_2P_3$	P_4P_5	P_6P_7
$P_1P_2P_3$	0		
P_4P_5	(3)	0	
P_6P_7	6	6	0

$P_1P_2P_3P_4P_5 \rightarrow$ N-C

	$P_1P_2P_3P_4P_5$	P_6P_7
$P_1P_2P_3P_4P_5$	0	
P_6P_7	(6)	0

$P_1P_2P_3P_4P_5P_6P_7 \rightarrow$ One cluster



Scanned with CamScanner

Complete linkage (max dist)

	P_1	P_2P_3	P_4	P_5	P_6	P_7
P_1	0					
P_2P_3	1.5	0				
P_4	4.5	3.5	0			
P_5	3.5	3.5	(1)	0		
P_6	6.5	6.5	8	7	0	
P_7	6.5	6.5	7	6	1	0

$P_4P_5 \rightarrow N.C$

	P_1	P_2P_3	P_4P_5	P_6	P_7
P_1	0				
P_2P_3	1.5	0			
P_4P_5	4.5	3.5	0		
P_6	6.5	6.5	8	0	
P_7	6.5	6.5	7	(1)	0

$P_6P_7 \rightarrow NC$

	P_1	P_2P_3	P_4P_5	P_6P_7
P_1	0			
P_2P_3	(1.5)	0		
P_4P_5	4.5	3.5	0	
P_6P_7	6.5	6.5	8	0

$P_1P_2P_3 \rightarrow N.C$

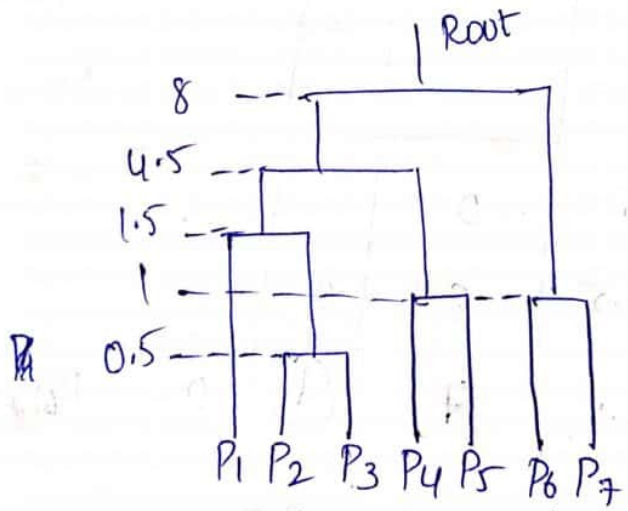
Scanned with CamScanner

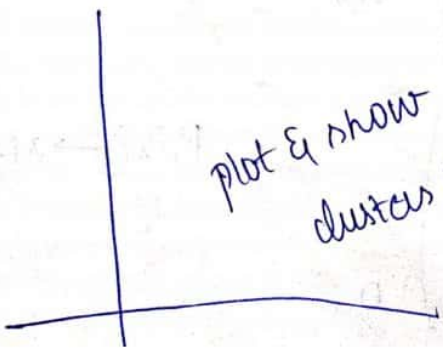
	$P_1P_2P_3$	P_4P_5	P_6P_7
$P_1P_2P_3$	0		
P_4P_5	(4.5)	0	
P_6P_7	6.5	8	0

$P_1P_2P_3P_4P_5 \rightarrow N.C$

	$P_1P_2P_3P_4P_5$	P_6P_7
$P_1P_2P_3P_4P_5$	0	
P_6P_7	(8)	0

$P_1P_2P_3P_4P_5P_6P_7 \rightarrow N.C$





Scanned with CamScanner

Avg linkage

	P_1	P_2P_3	P_4	P_5	P_6	P_7
P_1	0					
P_2P_3	$=1.25$ $\frac{1.5+1}{2}$	0				
P_4	4.5	3.25	0			
P_5	3.5	3.25	(1)	0		
P_6	6.5	6.25	8.	7	0	
P_7	6.5	6.25	7	6	1	0

$P_4P_5 \rightarrow NC$

	P_1	P_2P_3	P_4P_5	P_6	P_7
P_1	0				
P_2P_3	1.25	0			
P_4P_5	4	3.25	0		
P_6	6.5	6.25	7.5	0	
P_7	6.5	6.25	6.5	(1)	0

$P_6P_7 \rightarrow NC$

Scanned with CamScanner

	P_1	P_2P_3	P_4P_5	P_6P_7
P_1	0			
P_2P_3	1.25	0		
P_4P_5	4	3.25	0	
P_6P_7	6.5	6.25	7	0

$P_1P_2P_3 \rightarrow N.C$

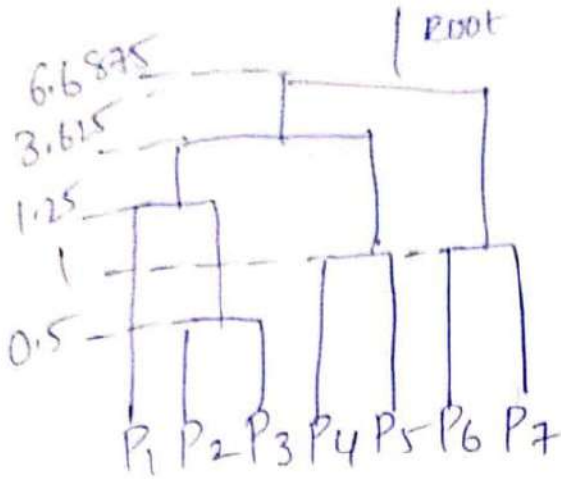
	$P_1P_2P_3$	P_4P_5	P_6P_7
$P_1P_2P_3$	0		
P_4P_5	3.625	0	
P_6P_7	6.375	7	0

$P_1P_2P_3P_4P_5 \longrightarrow N.C$

	$P_1P_2P_3P_4P_5$	P_6P_7
$P_1P_2P_3P_4P_5$	0	
P_6P_7	6.6875	0

$P_1P_2P_3P_4P_5P_6P_7$
$\downarrow N.C$

Scanned with CamScanner



↳ dendogram

(5) with paramaters as Minpts = 6, ε = 0.1 > Noise = 5%

clusters = 4

" " " Minpts = 8, ε = 0.1 (const) > Noise will increase

No. of clusters will ~~decrease increase~~ decrease.

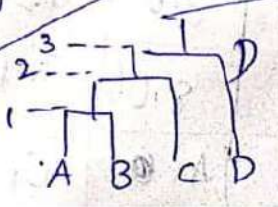
(3) single

AB → NC

	A	B	C	D
A	0			
B	(1)	0		
C	4	2	0	
D	5	6	3	0

	A B	C	D
AB	0		
C	(2)	0	
D	5	3	0

	ABC	D
ABL	0	
D	(3)	0



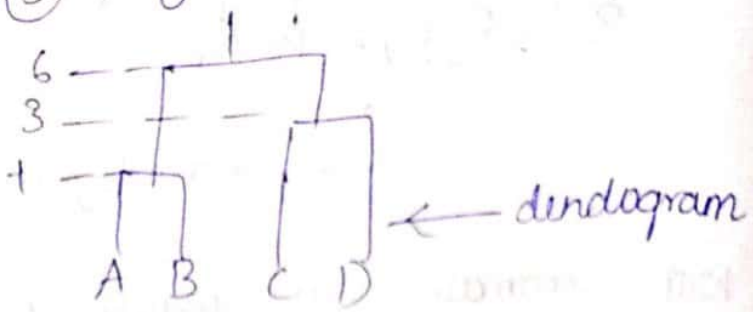
Scanned with CamScanner

Complete

	AB	C	D
AB	0		
C	4	0	
D	6	(3)	0

⇒

	AB	CD
AB	0	
CD	(6)	0



(2) $A_1(2,10)$
$A_2(2,5)$
$A_3(8,4)$
$B_1(5,8)$
$B_2(7,5)$
$B_3(6,4)$
$C_1(1,2)$
$C_2(4,9)$

Distance M:- Euclidian

New clusters { $A_1(2,10)$, $B_1(5,8)$, $C_1(1,2)$ } are initial clusters centres of each clus

K-Means ?

Round-1

	A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2
A_1	0	5	8.48	3.6	7.07	7.21	8.06	2
B_1	3.6	4.24	5	0	3.60	4.123	7.2	1
C_1	8.06	3.16	7.28	7.21	6.7	5.39	0	7
	A_1	C_1	B_1	B_1	B_1	B_1	C_1	C_1

Scanned with CamScanner

Lets calculate distances to update distance Matrix

$d(A_1, A_1) = 0$

$d(A_1, A_2) = \sqrt{0^2 + 5^2} = 5$

$d(A_1, A_3) = \sqrt{6^2 + 6^2} = 8.48$

$d(A_1, B_1) = \sqrt{3^2 + 2^2} = 3.60$

$d(A_1, B_2) = \sqrt{5^2 + 5^2} = 7.07$

$d(A_1, B_3) = \sqrt{4^2 + 6^2} = 7.21$

$d(A_1, C_1) = \sqrt{1^2 + 8^2} = 8.06$

$d(A_1, C_2) = \sqrt{2^2 + 1^2} = 2.23$

~~$d(A_1, C_3) = \sqrt{\quad}$~~

$d(B_1, A_1) = 3.6$

$d(B_1, A_2) = \sqrt{3^2 + 3^2} = 4.24$, $d(B_1, A_3) = \sqrt{3^2 + 4^2} = 5$

$d(B_1, B_1) = 0$

$d(B_1, B_2) = \sqrt{2^2 + 3^2} = 3.60$

$d(B_1, B_3) = \sqrt{1^2 + 4^2} = 4.123$

$d(B_1, B_1, C_1) = \sqrt{4^2 + 6^2} = 7.21$

$d(B_1, C_2) = \sqrt{1^2 + 1^2} = 1.41$

$d(C_1, A_1) = 8.06$

$d(C_1, A_2) = \sqrt{1^2 + 3^2} = 3.16$

$d(C_1, A_3) = \sqrt{7^2 + 2^2} = 7.28$

$d(C_1, B_1) = 7.21$

$d(C_1, B_2) = \sqrt{6^2 + 3^2} = 6.7$

$d(C_1, C_1) = 0$, $d(C_1, B_3) = \sqrt{5^2 + 2^2} = 5.39$

$d(C_1, C_2) = \sqrt{3^2 + 7^2} = 7.61$

Scanned with CamScanner

(a) After first round = New clusters are

A A_1 → $A1_1$

B B_1 → $A_3, B_1, B_2 B_3, C_2$

C C_1 → A_2, C_1

$A = (2, 10)$

$B = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right)$

$= (6, 6)$

$C = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

Round-2

	A_1	B_2	A_3	B_1	B_2	B_3	C_1	C_2
A_1	0							
B A_2								
C A_3								

Scanned with CamScanner

$d(A_1, B_1)$ -

$d(A_1, A_3)$ -

$d(A_1 B_1)$ -

$d(A_1 B_2)$ -

$d(A_1 B_3)$ -

$d(A_1, C_1)$ -

$d(A_1, C_2)$ -

$d(B_1, A_1)$ -

$d(B_1 A_2)$ -

$d(B_1, A_3)$ -

$d(B_1, B_1)$ -

$d(B_1, B_2)$ -

$d(B_1 B_3)$ -

$d(B_1 C_1) =$

$d(B_1, C_2) =$

$d(C_1 A_1)$ -

$d(C_1 A_2)$ -

$d(C_1, A_3)$ -

$d(C_{s1}, B_1)$ -

$d(C_1, B_2)$ -

$d(C_1 B_3)$ -

$d(C_1 C_1)$ -

$d(C_1 C_2)$ -

calculate distance & update distance matrix if last & this centroids DP are same stop

Scanned with CamScanner

Chapter-3 (LP)

(4)

A	B	class label
0	1	C1
0	0	C2
1	1	C1
0	1	C1
1	0	C1
0	0	C2
1	1	C1
0	0	C2
1	0	C1
1	0	C2

$C_1 = 6 \quad C_2 = 4$

Total = 10

A:

	A=0	A=1
C1	2	4
C2	3	1
	5	5

B:

	B=0	B=1
C1	2	4
C2	4	0
	6	4

(a) Gain using Gini Index

The overall gini index before splitting is

$G_{orig} = 1 - ((0.6)^2 + (0.4)^2)$

$= 0.48$

The gain in the Gini index after splitting on A is

A: $G_{A=0} = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 \Rightarrow 1 - 0.16 - 0.36 = 0.48$

$G_{A=1} = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 \Rightarrow 1 - 0.64 - 0.04 = 0.32$

B: $G_{B=0} = 1 - (2/6)^2 - (4/6)^2 = 1 - 0.11 - 0.44 = 0.44$

$G_{B=1} = 1 - (4/4)^2 - 0 \Rightarrow 0$

Scanned with CamScanner

corresponding gain$_A$ is $G_{orig} - \left(\frac{5}{10}\right) G_{A=0} - \frac{5}{10} G_{A=1}$

$\Rightarrow 0.48 - 0.5(0.48) - 0.5(0.32)$

$\Rightarrow 0.08$ //

$Gain_B = 0.48 - (6/10)(0.44) - (4/10)(0)$

$= 0.48 - (0.6)(0.44)$

$= 0.216$ //

As gain for B is more we chose B attribute to split the node.

(b) Entropy before split is E_{orig}

$E_{orig} = -0.6 \log_2(0.6) - 0.4 \log_2(0.4)$

$= -0.6(-0.736) - 0.4(-1.32)$

$= 0.4416 + 0.528$

$= 0.9696$

Info gain after splitting A is

$E_{A=0} = -\frac{2}{5} \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \log_2\left(\frac{3}{5}\right)$

$\Rightarrow -0.4(-1.32) - (0.6)(-0.736)$

$\Rightarrow 0.9696$

$E_{A=1} = -\frac{4}{5} \log_2\left(\frac{4}{5}\right) - \left(\frac{1}{5}\right) \log_2\left(\frac{1}{5}\right)$

$= -(0.8)(-0.32) - (0.2)(-2.32) = 0.1512$

$= 0.72$

$\Delta_A = 0.9696 - 5/10(0.9696) - (5/10)(0.72)$

$\Rightarrow 0.1248$

Scanned with CamScanner

For B

$G_{B=0} \Rightarrow -\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right)$

$\Rightarrow (-0.33)(-1.59) - (0.6)(-0.58)$

$\Rightarrow 0.8727$

$G_{B=1}$

$\Rightarrow \frac{-4}{4}\log_2(4/4)$

$\Rightarrow 0$ //

$\Delta_B = 0.9696 - \left(\frac{6}{10}\right)(0.8727) - \left(\frac{4}{10}\right)(0)$

$\Rightarrow 0.44598$

∴ B is choosen as first attribute to split as B has more Δ (gain) value than A

Scanned with CamScanner

(3)

class	Prob	TP	FP	TN	FN	TPR	FPR
P	0.95	1	0	5	4	0.2	0
N	0.85	1	1	4	4	0.2	0.2
P	0.78	2	1	4	3	0.4	0.2
P	0.66	3	1	4	2	0.6	0.2
N	0.60	3	2	3	2	0.6	0.4
P	0.55	4	2	3	1	0.8	0.4
N	0.53	4	3	2	1	0.8	0.6
N	0.52	4	4	1	1	0.8	0.8
N	0.51	4	5	0	1	0.8	1
P	0.40	5	5	0	0	1	1

P=5 N=5

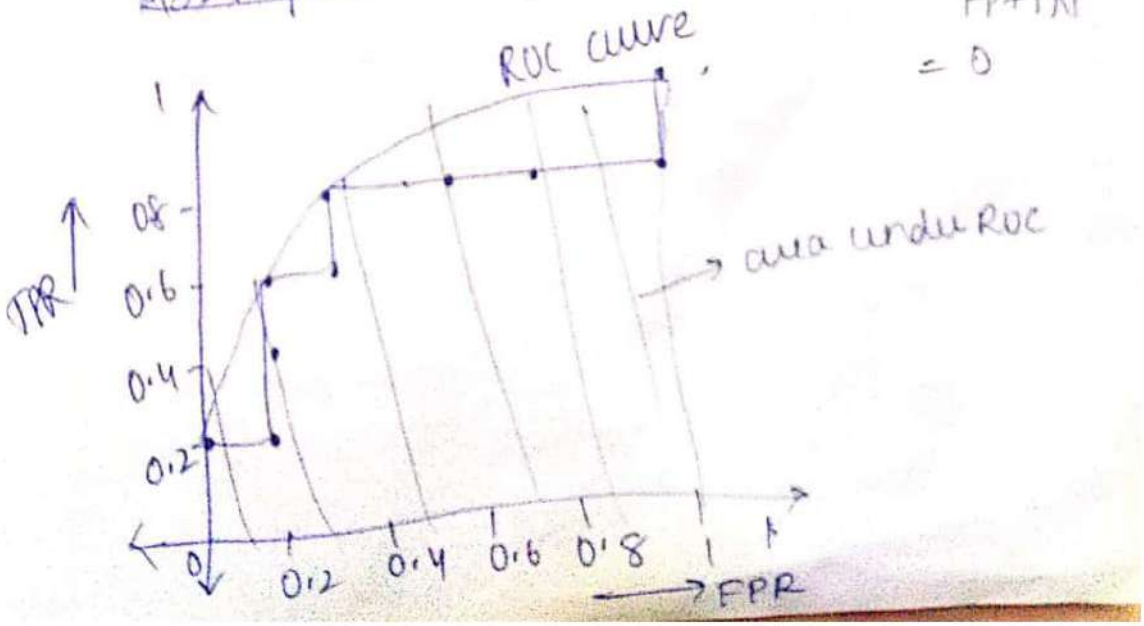
	P'	N'
P	TP 5 4 3 2 1	FN 0 1 2 3 4
N	FP 0 1 2 3 4	TN 0 1 2 3 4 5

$$TPR = \frac{TP}{TP+FN}$$

$$= 1/5 = 0.2$$

$$FPR = \frac{FP}{FP+TN}$$

$$= 0$$



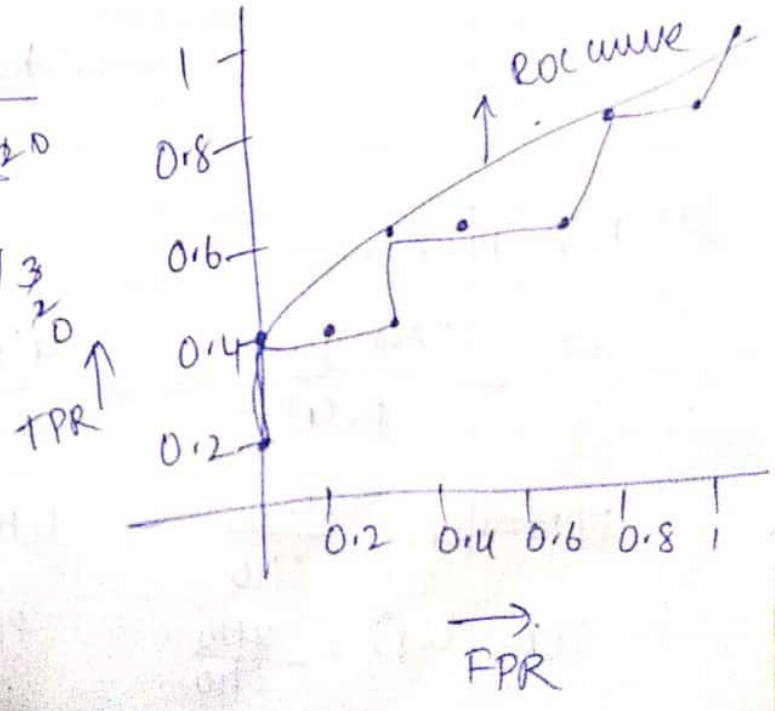
Scanned with CamScanner

(5)

Class.	Prob	TP	FP	TN	FN	TPR	FPR
(P) +1	0.95	1	0	5	4	0.2	0
(P) +1	0.94	2	0	5	3	0.4	0
(N) -1	0.87	2	1	4	3	0.4	0.2
(N) -1	0.86	2	2	3	3	0.4	0.4
(P) +1	0.86	3	2	3	2	0.6	0.4
(N) -1	0.86	3	3	2	2	0.6	0.6
(N) -1	0.76	3	4	1	2	0.6	0.8
(P) +1	0.53	4	4	1	1	0.8	0.8
(N) -1	0.44	4	5	0	1	0.8	1
(P) +1	0.25	5	5	0	0	1	1

+ve (+1)

	P'	N'
P	TP 4 3 2 1	FN 4 3 2 0
N	FP 2 1 0 3 4 5	TN 5 4 3 2 0



Scanned with CamScanner

Krishna

Model Paper

3b)

	P'	N'	
P	100 (TP)	FN 40	= 140
N	60 (FP)	TN 300	= 360
	160	340	

$N = TP + FN + FP + TN$

$\Rightarrow 500$

a) $\text{Accuracy} = \frac{TP + TN}{N} = \frac{400}{500} = 0.8 \times 100 = 80\%$

b) $\text{Precision} = \frac{TP}{P'} = \frac{100}{160} = 0.625 = 62.5\%$

c) $\text{Recall} = \frac{TP}{P} = \frac{100}{140} = 0.714 = 71.42\%$

d) $\text{F1 score} = \frac{2 \times \text{Pre} \times \text{Recall}}{\text{Pre} + \text{Recall}}$

$= \frac{2 \times 0.625 \times 0.714}{0.625 * 0.714} = \frac{0.8925}{1.339} = 66.65\%$

3a) $P(A=1 | C=C_1)$

$\Rightarrow \frac{P(A=1 \text{ given } C_1)}{P(C_1)} = \frac{4/10}{6/10} = 0.66$

$P(B=1|C_1) = \frac{4/10}{6/10}$

$P(B=0|C_1) = \frac{2/10}{6/10}$

$P(A=1|C_2) = \frac{4/10}{4/10}$

$P(A=0|C_2) = \frac{3/10}{4/10}$

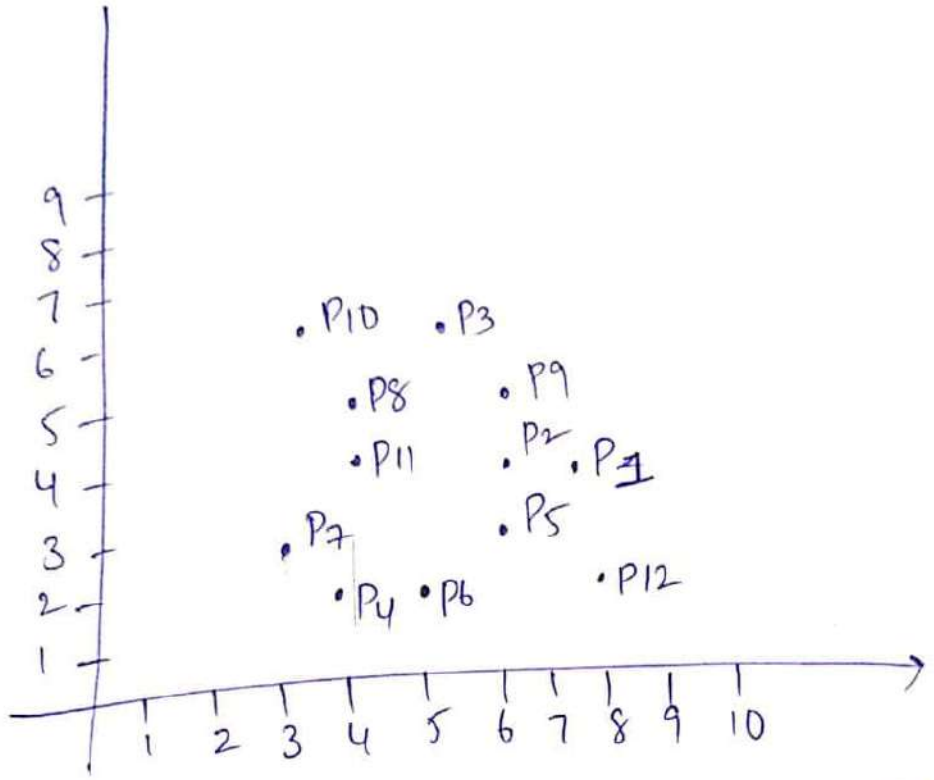
$P(B=1|C_2) = 0$

$P(B=0|C_2) = \frac{4/10}{4/10}$

Scanned with CamScanner

1b) R = 1.9 (ε)

Minpts = 4



$P_4(4,2)$ $P_6(5,2)$ $P_{12}(8,2)$ $P_7(3,3)$ $P_5(6,3)$ $P_1(7,4)$

$P_2(8,4)$ $P_{11}(4,4)$ $P_8(4,5)$ $P_9(6,5)$ $P_3(5,6)$

$P_{10}(3,6)$

Find distances then fill distance matrix And check for $\epsilon = 1.9$ & if mp = 4 then core else if min one pt mean border else noise

$d(P_1P_2)$

$d(P_1P_3)$

$d(P_1P_4)$

$d(P_1P_5)$

$d(P_1P_6)$

$d(P_1P_7)$

$d(P_1P_8)$

$d(P_1P_9)$ ---. done :c

Scanned with CamScanner