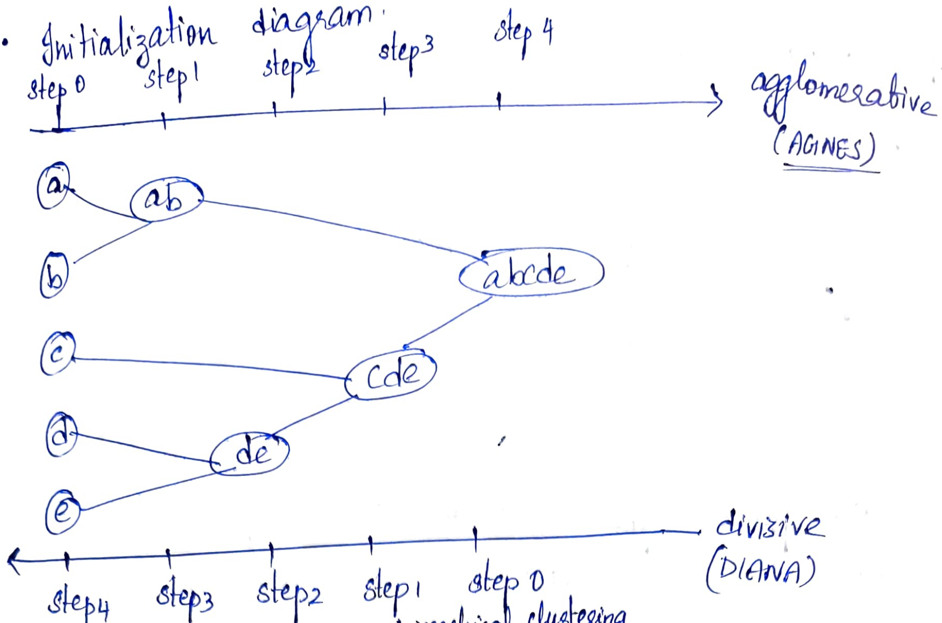
• Initialization diagram:



hierachical clustering

→ Agglomerative clustering	Divisive clustering
• bottom up approach	• top - down approach
• starts by placing each obj. in its own cluster and then merges these atomic clusters, until all the objects are in a single cluster or until certain termination conditions are satisfied.	• start at the top with all documents in one cluster. • The cluster is split using a flat clustering algorithm. • this procedure is applied recursively until each document is in its own singleton cluster.

14/3/25

⇒ Decision tree:

It is an supervised learning technique that gives the input data to be built upon some ML algorithms and predicting the output, based upon the decisions. eg: Yes/No male or female, class A/B/C.

eg:

Tid	Refund (categ)	Marital status (categ)	Taxable income (conti class)	class cheat
1	Yes	Single	125 K	No
2	NO	Married	100 K	No
3	NO	Single	70 K	No
4	Yes	Married	120 K	No
5	No	Divorced	95K	Yes
6	NO	Married	60 K	No
7	Yes	Divorced	220 K	No
8	No	Single	85K	Yes
9	NO	Married	75K	No
10	No	single	90 K	Yes

Splitting Attributes

Refund

No

- splitting attribute based upon the root element. • After finding the class, again split the attributes based on subnodes.
- After splitting the process based on subnodes, after finding the class category, going to take good decisions.

eg:

Name	Rank	Years	Tenured
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant prof	7	yes

→ start from root node -

Refund	marital status	Taxable income	cheat
No	Married	80K	?

Refund

NO

Marrit

⇒ classification:

- It is a supervised learning technique that predicts the value based on categorial data. It also uses label the data to predict the output.
- It classifies data based on testing data.

eg: Naives Bayes, ~~Bayesion~~ Bayesian network.

training data

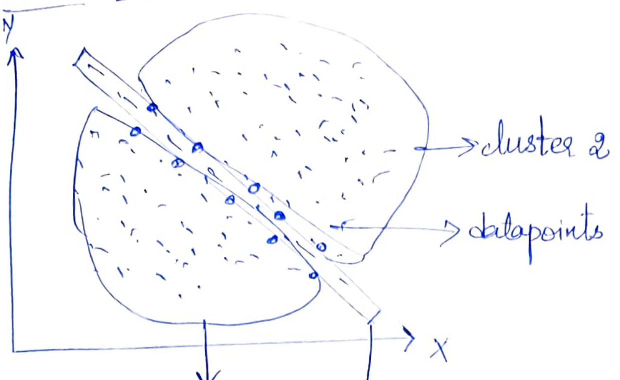
classification algorithm

classifier (model)

If rank = 'professor' OR years > 6 THEN Tenured = 'yes'

18/3/25

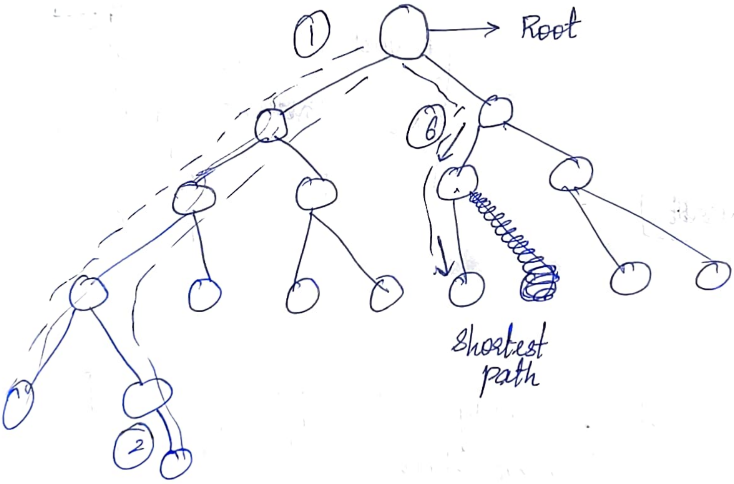
Support vector machine :



The dataset is divided into 2 parts through a separate single straight line

In the hyperlane the datasets of different groups ~~lik~~ lie on the margin to divide the partitions ~~or~~ segments of the whole dataset.

Random Forest :



largest path → 2

shortest path → 6 → optimized path

(choosing the shortest path - optimize solutions.)

0/3/2025 ⊗

⇒ Decision tree Algorithm Problem:

Also known as ID3 Algorithm - Iterative Dichotomiser Algorithm.

Q:

1	Size	color	shape	Class
2	Small	Yellow	Round	A
3	Big	Yellow	Round	A
4	Big	Red	Round	A
5	Small	Red	Round	A
6	Small	Black	Round	B
7	Big	Black	cube	B
8	Big	Yellow	cube	B
9	Big	Black	round	B
10	Small	Yellow	cube	B

find no. of classes

Given:

M = 2

no. of category or class

class A = 4

class B = 5

Total no. of classes = 9

entrophy - measure of how much uncertainity in the info

Info gain - how much info is the answer to a specific question provided.

Information Gain Entrophy.

Formula for entrophy:-

$$\boxed{Info(D) = -\sum_{i=1}^{M} P_i \log_2 P_i} \rightarrow \text{category}$$

Decision

class A class B

$$P_i = \frac{\text{no. of category}}{\text{Training data}}$$

class A = 4/9

class " = 5/9

$$Info(D) = \left[-\sum_{i=1}^{1} \frac{4}{9} \log \frac{4}{9}\right] + \left[\sum_{i=1}^{2} \frac{5}{9} \log \frac{5}{9}\right]$$

$$= \left[(-0.44)(-1.18)\right] + \left[(-0.55) \cdot (-0.86)\right]$$

$$= 0.519 + 0.473 = \underline{0.992}$$

(Info (D) = 0.992)

Given data:

Round = 6
cube = 3

Total no of training data = 9

Info (shape) =

$M = 2$

no. of Round (A) = 4
Round (B) = 2

$$\boxed{Info(D) = -\sum_{i=1}^{M} P_i \log P_i}$$

total round class A round class B

$$Info(shape) = \underbrace{\frac{6}{9}}_{\text{total training data}}\left[\left(-\frac{4}{6}\log_2\frac{4}{6}\right) + \left(-\frac{2}{6}\log_2\frac{2}{6}\right)\right]$$

total cube

$$+ \frac{3}{9}\left[\left(-\frac{0}{3}\log_2\frac{0}{3}\right) + \left(-\frac{3}{3}\log_2\frac{3}{3}\right)\right]$$

total training data cube

$$= \frac{2}{3}\left[\left(-\frac{2}{3}\log_2\frac{2}{3}\right) + \left(-\frac{1}{3}\log_2\frac{1}{3}\right)\right]$$

$$+ \frac{1}{3}\left[0 + \left(-1 \times \log_2(1)\right)\right]$$

$$= \frac{2}{3}\left[\left((-0.66)(-0.59)\right) + \left((-0.33)(-1.59)\right)\right] + 0.33\left[-1 \times 0\right]$$

$$= 0.66\left[0.39 + 0.52\right] = \underline{0.6006}$$

Info (color) = ?

Total no. of color = 3

~~Yellow (A) = 2~~

no. of class = 3

Red = 2
Black = 3
Yellow = 4

Total training data = 9

$$\boxed{info(D) = -\sum_{i=1}^{M} P_i \log P_i}$$

class A (yellow) class B (yellow)

$$Info(color) = \frac{4}{9}\left[\left(-\frac{2}{4}\log_2\frac{2}{4}\right) + \left(-\frac{2}{4}\log_2\frac{2}{4}\right)\right]$$

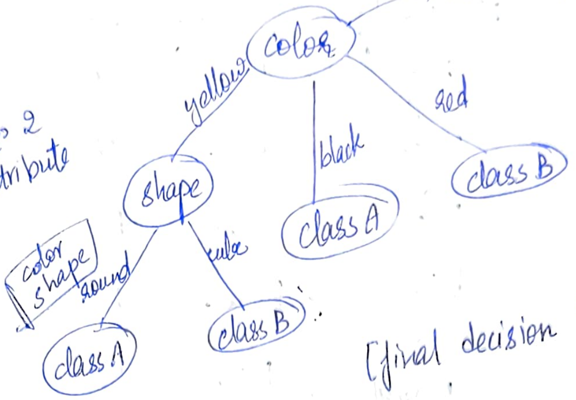
$$+ \frac{3}{9}\left[\left(\frac{-0}{3}\log_2\frac{0}{3}\right) + \left(\frac{3}{3}\log_2\frac{-3}{3}\right)\right] + \frac{2}{9}\left[\left(-\frac{2}{2}\log_2\frac{2}{2}\right) + \left(-\frac{0}{2}\log_2\frac{0}{2}\right)\right]$$

21/3/25

Entropy highest value.

root → highest value in entropy

Take 2 attribute



[final decision tree]

⇒ Bayesion classification:

⇒ Decision tree algorithm:

Example 2:

Owns home	Married	Gender	Employee	class
Y	Y	M	Y	B
N	N	F	Y	A
Y	Y	F	Y	C
Y	N	M	N	B
N	Y	F	Y	C
N	N	F	Y	A
N	N	M	Y	A
Y	N	M	N	B
N	Y	F	Y	A
Y	Y	F	Y	C

no. of classes (M) = 3 (A, B, C)

22/3/25

⇒ Bayesion classification:

outlook	Temperature	Humidity	Windly	class
sunny	Hot	high	False	N
sunny	Hot	high	True	N
overcast	Hot	high	False	P
rain	Mild	high	False	P
rain	Cool	Normal	False	P
rain	cool	Normal	True	N
overcast	cool	normal	True	P
sunny	mild	high	False	N
sunny	cool	normal	False	P
Rain	Mild	Normal	False	P
sunny	Mild	Normal	True	P
over cast	~~hot~~ mild	high	True	P
overcast	~~mild~~ hot	Normal	False	P
rain	mild	high	True	N

Rule:

step 1 $P(x)$ = outlook = Rain, temp = hot, Humidity = high, windy = false

No. of class = 2

Total no. of training data = 14

$P(N) = \frac{5}{14} = 0.357$

$P(P) = \frac{9}{14} = 0.642$

$P(\text{outlook} = \text{rain} / N) = \frac{2}{5} = 0.4$ → Total no. of N

$P(\text{outlook} = \text{rain} / P) = \frac{3}{9} = 0.33$ → Total no. of P.

$p(\text{temp} = \text{hot}/N) = \frac{2}{5} = 0.4$

$p(\text{temp} = \text{hot}/P) = \frac{2}{9} = 0.22$

$p(\text{humidity} = \text{high}/N) = \frac{4}{5} = 0.8$

$p(\text{humidity} = \text{high}/P) = \frac{3}{9} = 0.33$

$p(\text{windy} = \text{false}/N) = \frac{2}{5} = 0.4$

$p(\text{windy} = \text{false}/P) = \frac{6}{9} = \underline{0.66}$

step 2:

Formula ⇒ $P(H/x) = \frac{P(x/H) \cdot P(H)}{P(x)} = P(x_1 C_1) \times P(X_2 C_2) \cdots P(x_n C_n)$

$P(x/N) = \cancel{P(N/x)}\ p(\text{outlook} = \text{rain}/N) \times p(\text{temp} = \text{hot}/N) \times p(\text{humid} = \text{high}/N) \times p(\text{windy} = \text{false}/N)$

$= 0.4 \times 0.4 \times 0.8 \times 0.4 = 0.0512$

$P(x/P) = p(\text{outlook} = \text{rain}/P) \times p(\text{temp} = \text{hot}/P) \times p(\text{humid} = \text{high}/P) \times p(\text{windy} = \text{false}/N)$

$= 0.3 \times 0.2 \times 0.3 \times 0.6 = \underline{0.0108}$

~~st~~ last step:

$p(x/N) = 0.0512$

$p(x/P) = 0.0108$

To compute ⇒ propability of $(x/N) \cdot P(N)$

$= 0.0512 \times 0.357$

$= \underline{0.018}$

To compute $p(x/P) \cdot p(P)$

$= 0.0108 \times 0.642$

$= \underline{0.00679}$

∴ 3 attributes of the rule match with class N.

1/4/25

⇒ Univariate Linear model problem (or) Linear regression Problem

(1) Q Mid Term / Final Exam

Mid Term X	Final Exam Y
72	84
50	63
81	71
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

(1) Decision tree
↓
ID3 Algorithm

(2) Bayesian classification
↓
Naive Bayes

(3) linear regression

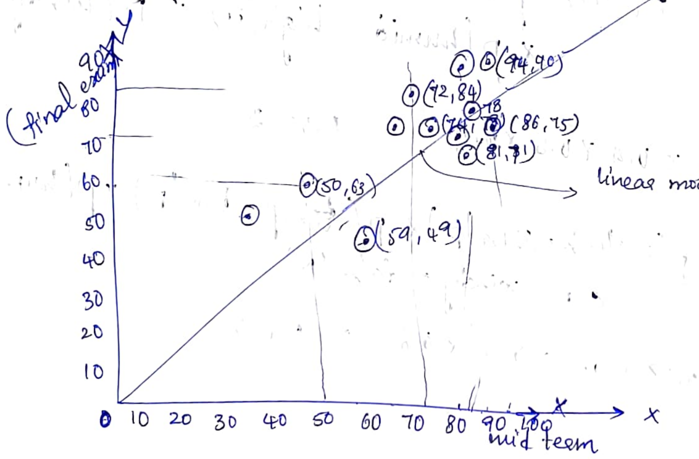
(4) k-means clustering

a) plot the data X & Y

b) use method of least² to find equation for the prediction of student final exam based on the student mid term grades in the grade score.

c) predict final exam grade of student who received 87 marks on the mid term exam.

(a)



X & Y are in linear relationship

(b) $\bar{x}$ = Mean of mid term = $\frac{\text{total}}{12}$ = 72.16

$\bar{y}$ = mean of final exam = $\frac{\text{total}}{12}$ = 73.5

(least square formula)

$$\beta = \frac{\sum_{i=1}^{S} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{S} (x_i - \bar{x})^2}$$

$$\beta_1 = \frac{(72-72.16)(84-73.5)}{(72-72.16)^2} = -65.625$$

(value for 1st row)

$$\beta_2 = \frac{(50-72.16)(63-73.5)}{(50-72.16)^2} = \frac{(-22.16)(-10.5)}{(22.16)^2} = 0.473$$

$$\beta_3 = \frac{(81-72.16)(71-73.5)}{(81-72.16)^2}$$

$$\beta_4 = \frac{(74-72.16)(78-73.5)}{(74-72.16)}$$

$\boxed{\beta = 0.569}$ → final answer.

(c) $\alpha = \bar{y} - \beta\bar{x}$ (predict 87 marks in midterm)

$= 73.5 - (0.569)(72.16)$

$\alpha = 32.44$

$y = \alpha + \beta x$

$x \to 87$

$\beta = 0.569$

$\alpha = 32.44$

$y = 32.44 + (0.569)(87)$

$\boxed{y = 81.943}$

Q ②

x: 10, 30, 15, 40

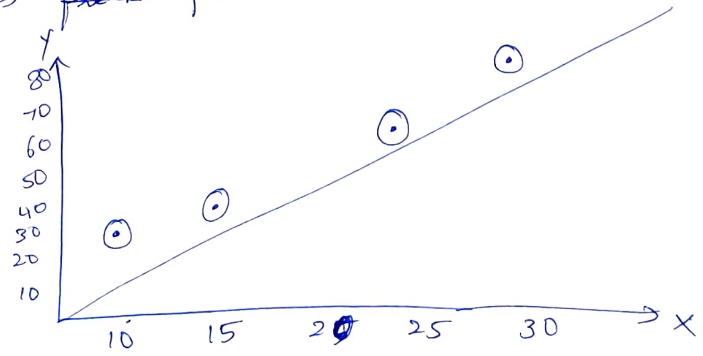
X	Y
10	30
15	40
25	60
30	80

(a) same as Q 1

(b) same as Q 2

(c) ~~predt~~ predict 22

(a)



(b) $\bar{x} = 20$

$\bar{y} = 52.5$

⇒ K-means clustering problem

Q

⇒ Datapoints (2,3), (6,5), (1,1)
(3,3), (8,8)

Initial centroids 1 : (2,3)
" 2 : (6,5)

step ①:

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

step ②:

calculate distances from each datapoint to centroid.

1) Distance from $(2,3)$ [x_1, y_1] to centroid 1 : $(2,3)$ [x_2, y_2]

$$d_1 = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$
$$= \sqrt{(2-2)^2 + (3-3)^2} = \underline{\underline{0}}$$

2) distance from (3,3) to centroid 1 : (2,3)

$$d_2 = \sqrt{(3-2)^2 + (3-3)^2} = \sqrt{1^2+0} = \underline{\underline{1}}$$

3) (6,5) to centroid 1 : (2,3)

$$d_3 = \sqrt{(6-2)^2 + (5-3)^2} = \sqrt{(4)^2 + 2^2}$$
$$= \sqrt{16+4} = \sqrt{20} = 4.47$$

4) (8,8) to c1 : (2,3)

$$d_4 = \sqrt{(-6)^2 + (-5)^2} = \sqrt{36+25} = \sqrt{61} = 7.81$$

5) (1,1) to c1 : (2,3)

$$d_5 = \sqrt{(1-2)^2 + (1-3)^2} = \sqrt{(-1)^2 + (-2)^2}$$
$$= \sqrt{1+4} = \sqrt{5} = 2.236$$

Distance to c2 ;

Distance from (2,3) to centroid 2 : (6,5)

$$d_1 = \sqrt{4^2 + 2^2} = \sqrt{16+4} = \sqrt{20}$$

$$d_2 = \sqrt{3^2 + 2^2} = \sqrt{13}$$

$d_3 = 0$

$d_4 = 3.606$

$d_5 = 6.403$

~~at~~
d1 = 0 at datapoint (2,3) closest to ~~c~~ centroid 1
d2 = 1 at point (3,3) is closest to centroid 1.
d3 = 0 at point (6,5) is closest to " 2.
d4 = 3.606 at " (8,8) " " " " 2
d5 = 2.236 at " (1,1) " " " (1)

step ③:
Assign points to cluster:
no. of clusters = no. of ~~t~~ centroid.

cluster 1 = (2,3), (3,3) (1,1)
cluster 2 = (6,5), (8,8)

step ④: Recalculate centroids ~~stepwise~~ (iteration 1)

Add $x_1 + x_2 + x_3$ / Total no. of datapts

cluster 1: (2,3), (3,3), (1,1)

new centroid 1 = $\left(\frac{x_1 + x_2 + x_3}{3}, \frac{y_1 + y_2 + y_3}{3}\right)$

$= \left(\frac{2+3+1}{3}, \frac{3+3+1}{3}\right)$

$= \left(\frac{6}{3}, \frac{7}{3}\right) = (2, 2.33)$

$2\overline{)13}$ = 6.5, 12, 1

(6,5), (8,8)

new centroid 2 = $\left(\frac{6+8}{2}, \frac{5+8}{2}\right)$

$= \left(\frac{14}{2}, \frac{13}{2}\right) = (7, 6.5)$

C1, (2, 2.3) ↓ C2 (7, 6.5)

	C1	point	C2
d1	$D = \sqrt{(2-2)^2 + (2.3-3)^2} = \cancel{0.67}$	(2,3)	$d1 = \sqrt{(7-2)^2 + (6.5-3)^2} = 6.10$
d2	$= \sqrt{(2-3)^2 + (2.3-3)^2} = 1.20$ ~~0.67~~	(3,3)	$d2 = \sqrt{(7-3)^2 + (6.5-3)^2} = 5.315$
d3	$= \sqrt{(2-6)^2 + (2.3-5)^2} = 4.80$	(6,5)	$d3 = \sqrt{(7-6)^2 + (6.5-5)^2} = 1.80$
d4	$= \sqrt{(2-8)^2 + (2.3-8)^2} = 8.24$	(8,8)	$d4 = \sqrt{(7-8)^2 + (6.5-8)^2} = 1.80$
d5	$= \sqrt{(2-1)^2 + (2.3-1)^2} = 1.66$	(1,1)	$d5 = \sqrt{(7-1)^2 + (6.5-1)^2} = 8.139$

new cluster assignment:
cluster 1 = (2,3) (3,3) (1,1)
cluster 2 = (6,5), (8,8)

clusters do not change after the 2nd iteration

final clusters

C1 : (2,3) (3,3), (1,1)

C2 : (6,5) (8,8) ; ~~int~~ iteration 2

(4,7), (5,6), (7,7), (2,3), (6,5)

Centroid C1: (4,7)

centroid C2: (7,7)

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

centroid c1;

$d1$ ~~wit~~ $= \sqrt{0+0} = 0$

$d2 = \sqrt{(-1)^2+(1)^2} = \sqrt{1+1} = \sqrt{2} = 1.414$

$d3 = \sqrt{(-3)^2+(0)^2} = \sqrt{9} = 3$

$d4 = \sqrt{2^2+4^2} = \sqrt{4+16} = \sqrt{20} = 4.472$

$d5 = \cancel{\sqrt{1^2+2^2}} = \cancel{\sqrt{1+4}} = \cancel{\sqrt{5}} = \cancel{2.236}$

$\sqrt{(-2)^2+2^2} = \sqrt{4+4} = \sqrt{8} = 2.828$

centroid C2;

$d1 = \sqrt{3^2+0^2} = \sqrt{9} = 3$

$d2 = \sqrt{2^2+1^2} = \sqrt{4+1} = \sqrt{5} = 2.236$

$d3 = \sqrt{0+0} = 0$

$d4 = \sqrt{5^2+4^2} = \sqrt{25+16} = \sqrt{41} = 6.403$

$d5 = \sqrt{1^2+2^2} = \sqrt{5} = 2.236$

cluster 1 : (4,7), (5,6), (2,3)

cluster 2 : (7,7), (6,5)

new centroid for cluster 1 : $\left(\frac{4+5+2}{3}, \frac{7+6+3}{3}\right)$

$= \left(\frac{11}{3}, \frac{16}{3}\right) = (3.66, 5.33)$

new centroid for cluster 2: $\left(\frac{7+6}{2}, \frac{7+5}{2}\right)$

$= \left(\frac{13}{2}, \frac{12}{2}\right)$

$= (6.5, 6)$

new centroid c1 ;

d1 = $\sqrt{0.1156 + 2.7889}$ = 1.704

d2 = $\sqrt{1.79 + 0.4489}$ = 1.49

d3 = $\sqrt{11.15 + 2.78}$ = 3.73

d4 = $\sqrt{2.75 + 5.42}$ = 2.85

d5 = $\sqrt{5.47 + 0.10}$ = 2.36

new centroid c2 ;

d1 = $\sqrt{6.25 + 1}$ = 2.69

d2 = $\sqrt{2.25 + 0}$ = 5.06

d3 = $\sqrt{0.25 + 1}$ = 1.11

d4 = $\sqrt{20.25 + 9}$ = $\sqrt{29.25}$ = 5.40

d5 = $\sqrt{0.25 + 1}$ = 1.11

cluster 1 : (4,7) (5,6) ~~(7,7)~~ 2,3)

cluster 2 : {(7,7) (6,5)

new centroid c1 => (3.66, 5.33)

new centroid c2 => (6.5, 6)

~~Int~~ Iteration 2 will also give the same.