USN

CMRIT
CMR INSTITUTE OF TECHNOLOGY, BENGALURU.
ACCREDITED WITH A++ GRADE BY NAAC

Internal Assessment Test 3 – March 2024

Sub:	**Artificial Intelligence and Machine Learning**				Sub Code:	**21CS54**	Branch:	**CSE**
Date:	**14/03/2024**	Duration:	90 mins	Max Marks: 50	Sem/Sec:	V / A, B&C		OBE

<u>Answer any FIVE FULL Questions</u>

		MARKS	CO	RBT
1	Write the C4.5 algorithm. How is a C4.5 algorithm better than ID3? What are the attribute selection measures used in the algorithm?	(5+3+2)	CO2	L2
2	Consider a dataset containing information about patients to predict whether a person has "Heart Disease" or not. Apply the CART algorithm to build a decision tree model	10	CO3	L3
3	State Bayes theorem. Write an algorithm of the Naïve Nayes classifier. List out and write the working principle of variants of Naïve Bayes classifiers.	(2+5+3)	CO2	L2

Chest Pain	**Good Blood Circulation**	**Blocked Arteries**	**Heart Disease**
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	No	Yes
No	No	Yes	Yes
No	Yes	No	No
Yes	No	Yes	Yes

USN

CMRIT
CMR INSTITUTE OF TECHNOLOGY, BENGALURU.
ACCREDITED WITH A++ GRADE BY NAAC

Internal Assessment Test 3 – March 2024

Sub:	**Artificial Intelligence and Machine Learning**				Sub Code:	**21CS54**	Branch:	**CSE**
Date:	**14/03/2024**	Duration:	90 mins	Max Marks: 50	Sem/Sec:	V / A, B&C		OBE

<u>Answer any FIVE FULL Questions</u>

		MARKS	CO	RBT
1	Write the C4.5 algorithm. How is a C4.5 algorithm better than ID3? What are the attribute selection measures used in the algorithm?	(5+3+2)	CO2	L2
2	Consider a dataset containing information about patients to predict whether a person has "Heart Disease" or not. Apply the CART algorithm to build a decision tree model	10	CO3	L3
3	State Bayes theorem. Write an algorithm of the Naïve Nayes classifier. List out and write the working principle of variants of Naïve Bayes classifiers.	(2+5+3)	CO2	L2

Chest Pain	**Good Blood Circulation**	**Blocked Arteries**	**Heart Disease**
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	No	Yes
No	No	Yes	Yes
No	Yes	No	No
Yes	No	Yes	Yes

<table>
<tr><td>4</td><td>(a) Draw and explain the structure of an Artificial Neural Network.
(b) Write an algorithm of the perceptron Learning.</td><td>5
5</td><td>CO2</td><td>L2</td></tr>
<tr><td>5</td><td>Consider the network architecture with 4 input and 2 output units. Consider four training samples each vector of length 4.
Training samples:
I1: (1,1,1,0) I2: (0,0,1,1) I3: (1,0,0,1) I4: (0,0,1,0)
Output units: Unit1 and Unit2, learning rate 0.6.
Initial weight matrix:
Unit 1: 0.3 0.8 0.7 0.2
Unit 2: 0.6 0.7 0.4 0.6
Identify an algorithm to learn without supervision. Show the weight updates of the SOFM during the first epoch of the learning algorithm.</td><td>10</td><td>CO4</td><td>L3</td></tr>
<tr><td>6</td><td>Consider the following dataset. Cluster it using the K-Means algorithm with the initial value of objects 2 and 4 as initial seeds.
<table><tr><td>S.No</td><td>X</td><td>Y</td></tr><tr><td>1</td><td>3</td><td>5</td></tr><tr><td>2</td><td>7</td><td>8</td></tr><tr><td>3</td><td>12</td><td>5</td></tr><tr><td>4</td><td>16</td><td>9</td></tr></table></td><td>10</td><td>CO5</td><td>L3</td></tr>
</table>

CI CCI HoD

<table>
<tr><td>4</td><td>(a) Draw and explain the structure of an Artificial Neural Network.
(b) Write an algorithm of the perceptron Learning.</td><td>5
5</td><td>CO2</td><td>L2</td></tr>
<tr><td>5</td><td>Consider the network architecture with 4 input and 2 output units. Consider four training samples each vector of length 4.
Training samples:
I1: (1,1,1,0) I2: (0,0,1,1) I3: (1,0,0,1) I4: (0,0,1,0)
Output units: Unit1 and Unit2, learning rate 0.6.
Initial weight matrix:
Unit 1: 0.3 0.8 0.7 0.2
Unit 2: 0.6 0.7 0.4 0.6
Identify an algorithm to learn without supervision. Show the weight updates of the SOFM during the first epoch of the learning algorithm.</td><td>10</td><td>CO4</td><td>L3</td></tr>
<tr><td>6</td><td>Consider the following dataset. Cluster it using the K-Means algorithm with the initial value of objects 2 and 4 as initial seeds.
<table><tr><td>S.No</td><td>X</td><td>Y</td></tr><tr><td>1</td><td>3</td><td>5</td></tr><tr><td>2</td><td>7</td><td>8</td></tr><tr><td>3</td><td>12</td><td>5</td></tr><tr><td>4</td><td>16</td><td>9</td></tr></table></td><td>10</td><td>CO5</td><td>L3</td></tr>
</table>

CI CCI HoD

IAT 3

21CS54 AIML

Scheme of valuation

(1.) Write the C4.5 algorithm. How is a C4.5 algorithm better than ID3? What are the attribute selection measures used in the algorithm?

C4.5 algorithm: (5)

1.Compute entropy for the whole training dataset based on the target attribute.

$$Entropy\_Info(T) = -\sum_{i=1}^{m} P_i \log_2 P_i$$

2. Compute Entropy, Information gain, Split information and Gain Ratio for each of the attribute in the training dataset.

$$Entropy\_Info(T, A) = \sum_{i=1}^{v} \frac{|A_i|}{|T|} \times Entropy\_Info\ (A_i)$$

$$Information\_Gain(A) = Entropy\_Info(T) - Entropy\_Info(T, A)$$

$$Split\_Info(T, A) = -\sum_{i=1}^{v} \frac{|A_i|}{|T|} \times \log_2 \frac{|A_i|}{|T|}$$

$$Gain\_Ratio(A) = \frac{Info\_Gain(A)}{Split\_Info(T, A)}$$

3. Choose the attribute for which the maximum Gain ratio is the best-split attribute.
4. The best-split attribute is placed as the root node.
5. The root node is branched into subtrees with each subtree as an outcome of the test condition of the root node attribute. Accordingly, the training dataset is also split into subsets.
6. Recursively apply the same operation for the subset of the training set with the remaining attributes until a leaf node is derived or no more training instances are available in the subset.

(3)

Features	*ID3*	*C4.5*
Type of data	Categorical	Continuous and Categorical
Speed	Low	Faster than ID3
Boosting	Not supported	Not supported
Pruning	No	Pre-pruning
Missing Values	Can't deal with	Can't deal with
Formula	Use information entropy and information Gain	Use split info and gain ratio

Metrics: (2)

Entropy, Information gain, Split information and Gain Ratio

(3) State Bayes theorem. Write an algorithm of the Naïve Nayes classifier. List out and write the working principle of variants of Naïve Bayes classifiers.

Bayes theorem: **(2)**

It can be written as:

$$P(\text{Hypothesis } h \mid \text{Evidence } E) = \frac{P(\text{Evidence } E \mid \text{Hypothesis } h)\, P(\text{Hypothesis } h)}{P(\text{Evidence } E)}$$

Algorithm: Naïve Bayes Classifier: **(5)**

1. Compute the prior probability for the target class.
2. Compute the Frequency matrix and likelihood probability for each of the attributes.
3. Use Bayes theorem to calculate the probability of all hypotheses.

It can be written as:

$$P(\text{Hypothesis } h \mid \text{Evidence } E) = \frac{P(\text{Evidence } E \mid \text{Hypothesis } h)\, P(\text{Hypothesis } h)}{P(\text{Evidence } E)}$$

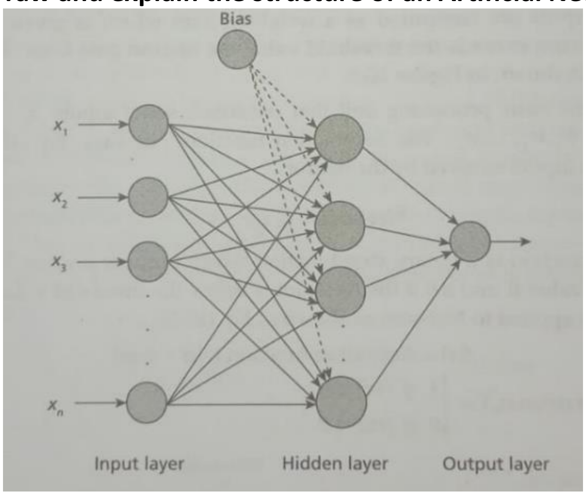
4. Use Maximum A Posteriori (MAP) hypothesis h_{MAP} to classify the test object to the hypothesis with the highest probability.

$$h_{MAP} = \max_{h\in H} P(Hypothesis\,h \mid Evidence\ E)$$
$$= \max_{h\in H} \frac{P(Evidence\ E \mid Hypothesis\ h)P(Hypothesis\ h)}{P(Evidence\ E)}$$
$$= \max_{h\in H} P(Evidence\ E \mid Hypothesis\ h)P(Hypothesis\ h)$$

Variants of Naïve Bayes classifiers: (3)

1) Bernoulli Naïve Bayes classifier
   - Discrete features
   - Boolean variables
2) Multinomial Naïve Bayes classifier
   - Generalization of Bernoulli Naïve Bayes classifier
3) Multi-class Naïve Bayes classifier
   - Multiclassification

(4) (a) Draw and explain the structure of an Artificial Neural Network. (5)



(b) Write an algorithm of perceptron Learning. (5)

...out how a perceptron model works

Algorithm 10.1: Perceptron Algorithm

Set initial weights $w_1, w_2, \ldots, w_n$ and bias θ to a random value in the range [–0.5, 0.5].

For each Epoch,

1. Compute the weighted sum by multiplying the inputs with the weights and add the products.
2. Apply the activation function on the weighted sum:
$$Y = \text{Step}\left((x_1 w_1 + x_2 w_2) - \theta\right)$$
3. If the sum is above the threshold value, output the value as positive else output the value as negative.
4. Calculate the error by subtracting the estimated output $Y_{estimated}$ from the desired output $Y_{desired}$:
$$\text{error } e(t) = Y_{desired} - Y_{estimated}$$
[If error $e(t)$ is positive, increase the perceptron output Y and if it is negative, decrease the perceptron output Y.]
5. Update the weights if there is an error:
$$\Delta w_i = \propto \times e(t) \times x_i$$
$$w_i = w_i + \Delta w_i$$

where, x_i is the input value, $e(t)$ is the error at step t, $\propto$ is the learning rate and Δw_i is the difference in weight that has to be added to w_i.

Q:2

Dataset:

	Chest pain	Good Blood circulation	Blocked Arteries	Heart Disease
①	No	No	No	No
②	Yes	Yes	Yes	Yes
③	Yes	Yes	No	No
④	Yes	No	No	Yes
⑤	No	No	Yes	Yes
⑥	No	Yes	No	No
⑦	Yes	No	Yes	Yes

Solution:

Iteration 1:

Gini-Index for the whole dataset

$$Gini\_Index(T) = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2$$

[formula: $Gini\_Index(T) = 1 - \sum_{i=1}^{m} P_i^2$]

$= 1 - 0.327 - 0.184$

$= 0.489$

First attribute: 'chest pain'

possible values: No, Yes, so subsets: 4

No. of Yes : 4

No. of No : 3

Chest Pain	Heart Disease = Yes	Heart Disease = No
Yes	3	1
No	1	2

2 subsets:

{Yes}, {No}, ~~{Yes, No}~~

Gini-Index (T, chest pain ∈ {Yes})

$= 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$

$= 1 - 0.563 - 0.063$

$= 0.374$

Gini-Index (T, chest pain {No})

$= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2$

$= 1 - 0.11 - 0.44$

$= 0.45$

Gini-Index (T, chest pain {Yes, No})

$= \frac{|S_1|}{T} Gini(S_1) + \frac{|S_2|}{|T|} Gini(S_2)$

$= \frac{4}{7} \times 0.374 + \frac{3}{7} \times 0.45$

$= 0.214 + 0.19$

$= 0.404$

Δ Gini (Chest pain) = Gini(T) − Gini(Chest pain)

$= 0.489 - 0.404$

$= 0.085$

Second attribute: Good Blood Circulation

Yes : 3, NO : 4

Good Blood Circulation	Heart Disease = Yes	Heart Disease = No
Yes	1	2
NO	3	1

2 subsets:

{Yes}, {No}

Gini_index (T, GBC ∈ {Yes})

$= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2$

$= 1 - 0.11 - 0.44 = 0.45$

Gini-Index (T, GBC ∈ {No})

$= 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 1 - 0.563 - 0.063$

$= 0.374$

Gini-Index (Good B.C ∈ {Yes, No})

$= \frac{3}{7} \times 0.45 + \frac{4}{7} \times 0.374$

$= ~~0.193~~\ 0.404$

ΔGini (GBC) = Gini(T) − Gini (GBC)

= 0.085

Third attribute: Blocked Arteries

Yes: 3, No: 4

Blocked Arteries	Heart Disease = Yes	Heart Disease = No
Yes	3	0
No	1	3

Subsets:

{Yes}, {No}

Gini-Index (BA ∈ {Yes})

$= 1 - \left(\frac{3}{3}\right)^2$

$= 0$

Gini-Index (BA ∈ {No})

$= 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$

$= ~~0.3~~\ 0.374$

Gini-Index (BA ∈ {Yes, No})

$= \frac{3}{7} \times 0 + \frac{4}{7} \times 0.374$

$= 0.214$

ΔGini (BA) = Gini(T) − Gini (BA)

$= 0.489 - 0.214$

$= 0.275$

Gini-index & ΔGini for all the attributes

Attribute	Gini-index	ΔGini
chest pain	0.404	0.085
GBC	0.404	0.085
BA	0.214	0.275

Blocked Arteries has the highest ΔGini value.
So 'Blocked Arteries' is the root node.

Blocked Arteries

Yes:

Chest pain	GBC	HD
Yes	Yes	Yes
NO	NO	Yes
Yes	NO	Yes

NO:

Chest pain	GBC	HD
NO	NO	NO
Yes	Yes	NO
Yes	NO	Yes
NO	Yes	NO

Iteration 2: Left subtree

Chest pain	GBC	HD
Yes	Yes	Yes
NO	NO	Yes
Yes	No	Yes

same class

$$\text{Gini\_Index}(T) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2$$

$$= 0$$

Blocked Arteries

Yes: Heart disease = Yes

NO:

CP	GBC	HD
NO	NO	NO
Yes	Yes	NO
Yes	No	Yes
NO	Yes	NO

Iteration 2 :

CP	GBC	HD
NO	NO	NO
Yes	Yes	NO
Yes	NO	Yes
NO	Yes	NO

$$\text{Gini-index}(T) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$$
$$= 0.374$$

First attribute : 'Chest pain'
Yes : 1 , No : 3

Chest pain	Heart disease = Yes	HD = No
Yes	1	1
NO	0	2

Subsets :
{Yes} , {No}

$$\text{Gini-index}(CP \in \{yes\})$$
$$= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2$$
$$= 1 - 0.25 - 0.25$$
$$= 0.5$$

$$\text{Gini-index}(CP \in \{No\})$$
$$= 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2$$
$$= 0$$

$$\text{Gini-index}(\text{Chest pain}) = \frac{2}{4} \times 0.5$$
$$= 0.25$$

$$\Delta \text{Gini} = \text{Gini}(T) - \text{Gini}(\text{chest pain})$$
$$= 0.374 - 0.25$$
$$= 0.124$$

2nd attribute: Good Blood Circulation

Yes: 2, No: 2

GBC	HD = Yes	HD = No
Yes	0	2
No	1	1

~~Gini-index~~

Subset:

{Yes}, {No}

Gini-index (GBC ∈ {yes})

= 0

Gini-index (GBC ∈ {No})

$= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2$

$= 0.5$

~~ΔGini = Gini(T) −~~

Gini_index (Good Blood Circulation)

$= \frac{2}{4} \times 0 + \frac{2}{4} \times 0.5$

$= 0.25$

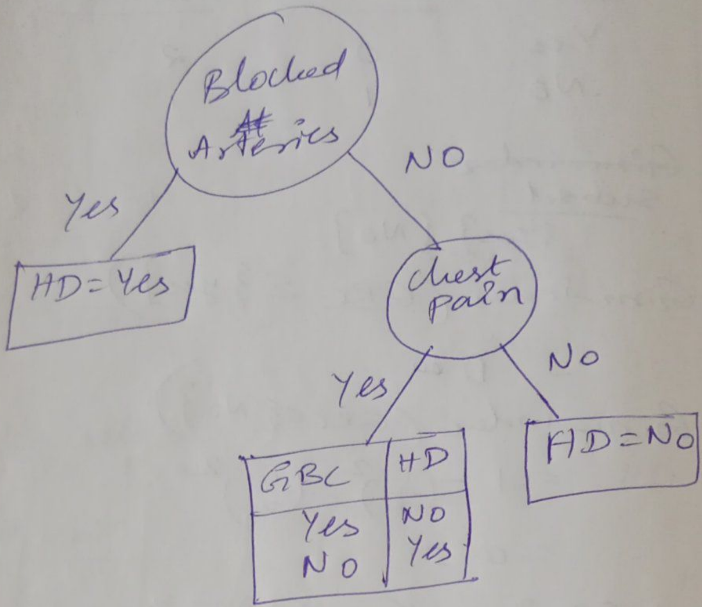
ΔGini(GBC)
= Gini(T) − Gini(GBC)
= 0.374 − 0.25
= 0.124

Gini-index & ΔGini values

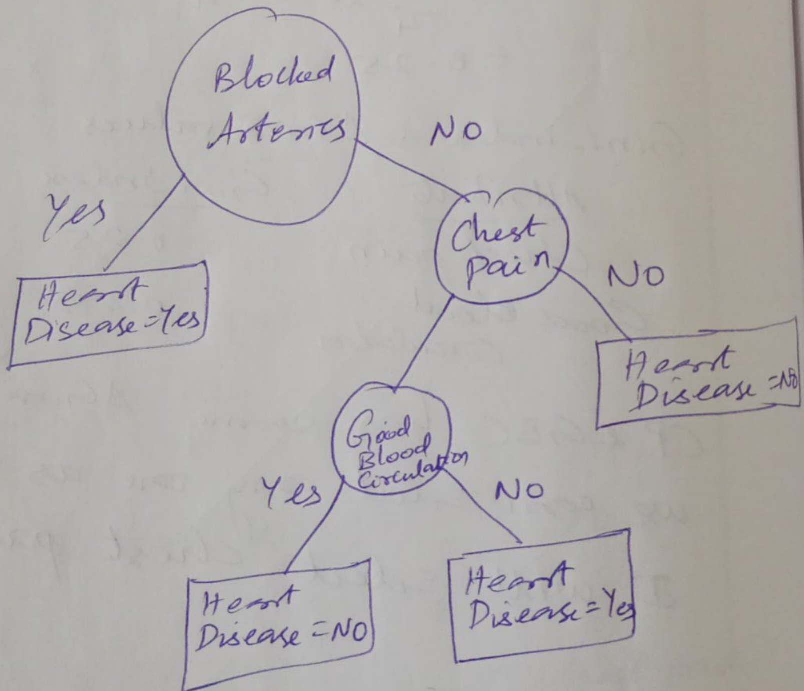
Attribute	Gini_index	ΔGini
Chest pain	0.25	0.124
Good Blood Circulation	0.25	0.124

CP & GBC has same ΔGini value,
we can select any one as root node
I will select 'chest pain' as root node.

The classification decision tree



Final tree:



15/3/2024
Friday

Network Architecture

Q:5
4 inputs units
2 output units (unit1 & unit2)

I/p: 4 training samples (each vector of length 4)

I_1 : (1, 1, 1, 0)

I_2 : (0, 0, 1, 1)

I_3 : (1, 0, 0, 1)

I_4 : (0, 0, 1, 0)

Learning rate = 0.6

Initial weight matrix:

$$\begin{matrix} \text{unit1:} \\ \text{unit2:} \end{matrix} \begin{bmatrix} 0.3 & 0.8 & 0.7 & 0.2 \\ 0.6 & 0.7 & 0.4 & 0.6 \end{bmatrix}$$

Solution:

Iteration 1:

Training sample I_1: (1, 1, 1, 0)

weight matrix

$$\begin{matrix} \text{unit1:} \\ \text{unit2:} \end{matrix} \begin{bmatrix} 0.3 & 0.8 & 0.7 & 0.2 \\ 0.6 & 0.7 & 0.4 & 0.6 \end{bmatrix}$$

Compute Euclidean distance between I_1 and unit 1 weights

$$d^2 = (0.3-1)^2 + (0.8-1)^2 + (0.7-1)^2 + (0.2-0)^2$$

$= 0.49 + 0.04 + 0.09 + 0.04$

$= 0.66$

Compute Euclidean distance between I_1 and unit 2 weights

$$d^2 = (0.6-1)^2 + (0.7-1)^2 + (0.4-1)^2 + (0.6-0)^2$$

$= 0.16 + 0.09 + 0.36 + 0.36$

$= 0.97$

- Distance between I_1 and Unit 1 is shorter, so Unit 1 wins.
- Update the weights of Unit 1

Formula:

$$w(t+) = w(t) + \eta\,(\text{Input} - w(t))$$

unit 1's weight vector

$= [0.3\ \ 0.8\ \ 0.7\ \ 0.2] + 0.6\,([1\ \ 1\ \ 1\ \ 0] - [0.3\ \ 0.8\ \ 0.7\ \ 0.2])$

$= [0.3\ \ 0.8\ \ 0.7\ \ 0.2] + 0.6\,(0.7\ \ 0.2\ \ 0.3\ \ -0.2)$

$= [0.3\ \ 0.8\ \ 0.7\ \ 0.2] + [0.42\ \ 0.12\ \ 0.18\ \ -0.12]$

$= [0.72\ \ 0.92\ \ 0.88\ \ 0.08]$

$$\begin{matrix} \text{unit 1} : \\ \text{unit 2} : \end{matrix} \begin{bmatrix} 0.72 & 0.92 & 0.88 & 0.08 \\ 0.6 & 0.7 & 0.4 & 0.6 \end{bmatrix}$$

Iteration 2: $I_2 : (0, 0, 1, 1)$

Euclidean distance between I_2 & Unit 1

$d^2 = (0.72-0)^2 + (0.92-0)^2 + (0.88-1)^2 + (0.08-1)^2$

$= 0.52 + 0.85 + 0.014 + 0.85$

$= 2.234$

Euclidean distance between I_2 & Unit 2

$d^2 = (0.6-0)^2 + (0.7-0)^2 + (0.4-1)^2 + (0.6-1)^2$

$= 0.36 + 0.49 + 0.36 + 0.16$

$= 1.37$

Unit 2 wins

update weights of unit 2 $= [0.6\ \ 0.7\ \ 0.4\ \ 0.6] + 0.6\,([0\ \ 0\ \ 1\ \ 1] - [0.6\ \ 0.7\ \ 0.4\ \ 0.6])$

$= [0.6\ \ 0.7\ \ 0.4\ \ 0.6] + 0.6\,(-0.6\ \ -0.7\ \ 0.6\ \ 0.4)$

$= [0.6\ \ 0.7\ \ 0.4\ \ 0.6] + [-0.36\ \ -0.42\ \ 0.36\ \ 0.24]$

$= [0.24\ \ 0.28\ \ \sout{0.64}\ 0.76\ \ \sout{0.36}\ 0.84]$

unit 1 : [0.72 0.92 0.88 0.08]
unit 2 : [0.24 0.28 ~~0.04~~ 0.76 ~~0.36~~ 0.84]

Iteration 3:

I_3 : (1, 0, 0, 1)

Euclidean distance between I_3 & unit 1

$d^2 = (0.72-1)^2 + (0.92-0)^2 + (0.88-0)^2 + (0.08-1)^2$
$= 0.08 + 0.85 + 0.77 + 0.85$
$= 2.55$

Euclidean distance between I_3 & Unit 2

$d^2 = (0.24-1)^2 + (0.28-0)^2 + (~~0.04~~\ 0.76 - 0)^2 + (~~0.36~~\ 0.84 - 1)^2$
$= 0.58 + 0.08 + ~~0.002~~\ 0.58 + ~~0.41~~\ 0.026$
$= ~~1.072~~\ 1.27$

Unit 2 wins

update weights of unit 2 } = [0.24 0.28 ~~0.04~~ 0.76 ~~0.36~~ 0.84] + 0.6 ([1, 0, 0, 1] − [0.24 0.28 ~~0.04~~ 0.76 ~~0.36~~ 0.84])

= [0.24 0.28 ~~0.04~~ 0.76 ~~0.36~~ 0.84] + 0.6 (0.76 −0.28 −~~0.04~~ 0.76 ~~0.64~~ 0.16)

= [0.24 0.28 ~~0.04~~ 0.76 ~~0.36~~ 0.84] + [0.46 −0.17 −~~0.024~~ 0.46 ~~0.38~~ 0.1]

= [0.7 0.11 ~~0.02~~ 0.3 ~~0.74~~ 0.94]

unit 1 : [0.72 0.92 0.88 0.08]
unit 2 : [0.7 0.11 ~~0.02~~ 0.3 ~~0.74~~ 0.94]

Iteration 4: I_4 : (0, 0, 1, 0)

Euclidean distance b/w I_4 and Unit 1

$d^2 = (0.72-0)^2 + (0.92-0)^2 + (0.88-1)^2 + (0.08-0)^2$
$= 0.52 + 0.85 + 0.014 + 0.006$
$= 1.39$

Euclidean distance between I_4 & Unit 2

$d^2 = (0.7-0)^2 + (0.11-0)^2 + (~~0.02~~ 0.3 - 1)^2 + (~~0.74~~ 0.94 - 0)^2$

$= 0.49 + 0.012 + ~~0.96~~ 0.49 + ~~0.55~~ 0.88$

$= ~~2.012~~ 1.872$

~~Euclidean distance between I_4 & Un~~

Unit 1 wins

update weights of unit 1 } = [0.72 0.92 0.88 0.08] + 0.6 ([0 0 1 0] − [0.72 0.92 0.88 0.08])

= [0.72 0.92 0.88 0.08] + 0.6 [−0.72 −0.92 0.12 −0.08]

= [0.72 0.92 0.88 0.08] + [−0.43 −0.55 0.07 ~~−0.05~~ −0.05]

= [0.29 ~~1.47~~ 0.37 0.95 0.03]

unit 1 : [0.29 ~~1.47~~ 0.37 0.95 0.03]
unit 2 : [0.7 0.11 ~~0.02~~ 0.3 ~~0.74~~ 0.94]

Best mapping units for each of the inputs taken are

I_1 : (1, 1, 1, 0) → Unit 1
I_2 : (0, 0, 1, 1) → Unit 2
I_3 : (1, 0, 0, 1) → Unit 2
I_4 : (0, 0, 1, 0) → Unit 1

— This is the result after epoch 1.

[illegible]/3/2024
[illegible]ursday
[illegible]:b

Cluster the dataset using the K_Means algorithm with the initial value of objects 2 and 4 as initial seeds

S.No	x	y
1	3	5
2	7	8
3	12	5
4	16	9

Solution:

(7, 8) & (16, 9) are initial seeds.

- These data points are stored as two clusters. So K = 2.
- Centroids: (7, 8) & (16, 9)

Iteration:

- Compare all the data points with the centroid and assign to the nearest cluster.

1st data point: (3, 5)

- calculate the distance between the data point and the centroids of cluster1 and cluster 2.

$$Dist(1, centroid_1) = \sqrt{(3-7)^2 + (5-8)^2}$$
$$= \sqrt{4^2 + 3^2} = \sqrt{16+9}$$
$$= \sqrt{25} = 5$$

$$Dist(1, centroid\ 2) = \sqrt{(3-16)^2 + (5-9)^2}$$
$$= \sqrt{13^2 + 4^2}$$
$$= \sqrt{169+16} = \sqrt{185}$$
$$= 13.60$$

(3, 5) is closer to cluster 1.

Object 3 : (12, 5)

Dist(3, centroid1) $= \sqrt{(12-7)^2 + (5-8)^2}$

$= \sqrt{5^2 + 3^2} = \sqrt{25+9}$

$= \sqrt{34} = 5.831$

Dist(3, centroid2) $= \sqrt{(12-16)^2 + (5-9)^2}$

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

$= \sqrt{32} = 5.66$

Object 3 is closer to cluster 2.

	cluster1	cluster 2
	(7, 8)	(16, 9)
	(3, 5)	(12, 5)
centroid:	$(\bar{x}, \bar{y})$	$(\bar{x}, \bar{y})$
	$\left(\frac{10}{2}, \frac{13}{2}\right)$	$\left(\frac{28}{2}, \frac{14}{2}\right)$
centroid:	(5, 6.5)	(14, 7)

Iteration 2 :

Object 1 : (7, 8) centroid 1 : (5, 6.5)
centroid 2 : (14, 7)

Dist(1, centroid1) $= \sqrt{(7-5)^2 + (8-6.5)^2}$

$= \sqrt{2^2 + 1.5^2}$

$= \sqrt{4 + 2.25}$

$= \sqrt{6.25} = 2.5$

$$Dist(1, centroid2) = \sqrt{(7-14)^2 + (8-7)^2}$$
$$= \sqrt{49+1}$$
$$= 7.07$$

Object 1 belongs to the same cluster 1.

Object 3: (12, 5) centroid 1: (5, 6.5)
centroid 2: (14, 7)

$$Dist(3, centroid1) = \sqrt{(12-5)^2 + (5-6.5)^2}$$
$$= \sqrt{49+2.25} = \sqrt{51.25}$$
$$= 7.16$$

~~object 3 is closer to~~

$$Dist(3, centroid2) = \sqrt{(12-14)^2 + (5-7)^2}$$
$$= \sqrt{2^2+2^2} = \sqrt{8}$$
$$= 2.83$$

Object 3 is closer to cluster 2

cluster 1	cluster 2
(7, 8)	(16, 9)
(3, 5)	(12, 5)
(5, 6.5)	(14, 7)

centroid: (5, 6.5) (14, 7)

- There is no change in the clusters.
- Therefore k-means algorithm terminates with two clusters with the given data points.