USN | | | | | | | | | | |

Internal Assessment Test 2 – May 2025
Scheme & Solution

Sub:	Machine Learning					Sub Code:	BAI602	Branch:	AIML
Date:	**23/05/25**	Duration:	90 min	Max Marks:	50	Sem/Sec:	VI(A&B)		OBE
<u>Answer any FIVE FULL Questions</u>							**MARKS**	**CO**	**RBT**

1	Make use of entropy and information gain to discover the root node for the Decision tree for the following dataset using ID3 algorithm.	10	CO3	L3

S. No.	CGPA	Interactiveness	Practical Knowledge	Communication Skills	Job Offer
1	≥9	Yes	Very good	Good	Yes
2	≥8	No	Good	Moderate	Yes
3	≥9	No	Average	Poor	No
4	<8	No	Average	Good	No
5	≥8	Yes	Good	Moderate	Yes
6	≥9	Yes	Good	Moderate	Yes
7	<8	Yes	Good	Poor	No
8	≥9	No	Very good	Good	Yes
9	≥8	Yes	Good	Good	Yes
10	≥8	Yes	Average	Good	Yes

Solution:

Step 1:

Calculate the Entropy for the target class 'Job Offer'.

Entropy_Info(Target Attribute = Job Offer) = Entropy_Info(7, 3) =

$$= -\left[\frac{7}{10}\log_2\frac{7}{10} + \frac{3}{10}\log_2\frac{3}{10}\right] = -(-0.3599 + -0.5208) = 0.8807$$

Iteration 1:

Step 2:

Calculate the Entropy_Info and Gain(Information_Gain) for each of the attribute in the training dataset.

Table 6.4 shows the number of data instances classified with Job Offer as Yes or No for the attribute CGPA.

Table 6.4: Entropy Information for CGPA

CGPA	Job Offer = Yes	Job Offer = No	Total	Entropy
≥9	3	1	4	
≥8	4	0	4	0
<8	0	2	2	0

Solution:

Step 1:

Calculate the Entropy for the target class 'Job Offer'.

Entropy_Info(Target Attribute = Job Offer) = Entropy_Info(7, 3) =

$$= -\left[\frac{7}{10}\log_2\frac{7}{10} + \frac{3}{10}\log_2\frac{3}{10}\right] = -(-0.3599 + -0.5208) = 0.8807$$

Iteration 1:

Step 2:

Calculate the Entropy_Info and Gain(Information_Gain) for each of the attribute in the training dataset.

Table 6.4 shows the number of data instances classified with Job Offer as Yes or No for the attribute CGPA.

Table 6.4: Entropy Information for CGPA

CGPA	Job Offer = Yes	Job Offer = No	Total	Entropy
≥9	3	1	4	
≥8	4	0	4	0
<8	0	2	2	0

Table 6.6: Entropy Information for Practical Knowledge

Practical Knowledge	Job Offer = Yes	Job Offer = No	Total	Entropy
Very Good	2	0	2	0
Average	1	2	3	
Good	4	1	5	

Entropy_Info(T, Practical Knowledge)

$$= \frac{2}{10}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{3}{10}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right] + \frac{5}{10}\left[-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right]$$

$$= \frac{2}{10}(0) + \frac{3}{10}(0.5280 + 0.3897) + \frac{5}{10}(0.2574 + 0.4641)$$

$$= 0 + 0.2753 + 0.3608$$

$$= 0.6361$$

$$\text{Gain(Practical Knowledge)} = 0.8807 - 0.6361$$
$$= 0.2446$$

Table 6.7 shows the number of data instances classified with Job Offer as Yes or No for the attribute Communication Skills.

Table 6.7: Entropy Information for Communication Skills

Communication Skills	Job Offer = Yes	Job Offer = No	Total
Good	4	1	5
Moderate	3	0	3
Poor	0	2	2

Entropy_Info(T, Communication Skills)

$$= \frac{5}{10}\left[-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right] + \frac{3}{10}\left[-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}\right] + \frac{2}{10}\left[-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right]$$

$$= \frac{5}{10}(0.5280 + 0.3897) + \frac{3}{10}(0) + \frac{2}{10}(0)$$

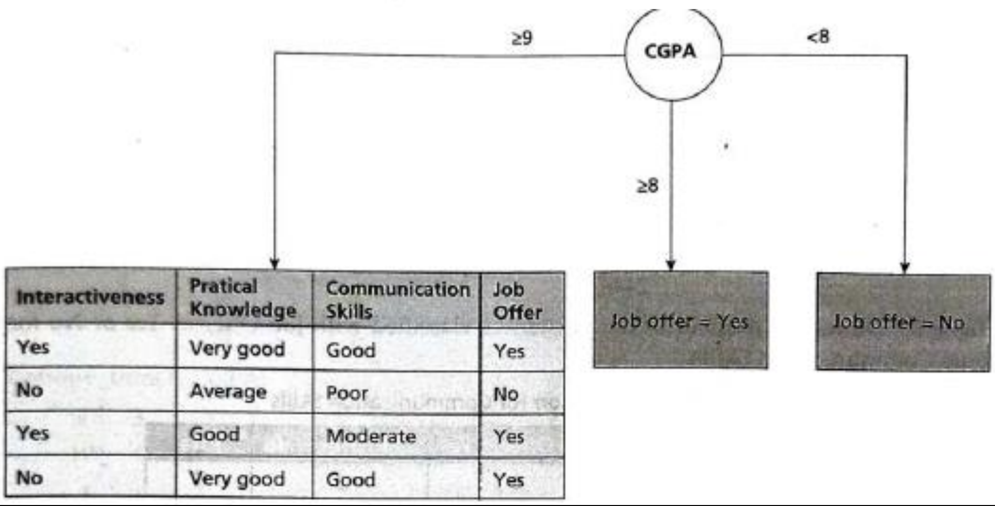
$$= 0.3609$$

$$\text{Gain(Communication Skills)} = 0.8813 - 0.36096$$
$$= 0.5203$$

The Gain calculated for all the attributes is shown in Table 6.8:

Table 6.8: Gain

Attributes	Gain
CGPA	0.5564
Interactiveness	0.0911
Practical Knowledge	0.2246
Communication Skills	0.5203



Interactiveness	Pratical Knowledge	Communication Skills	Job Offer
Yes	Very good	Good	Yes
No	Average	Poor	No
Yes	Good	Moderate	Yes
No	Very good	Good	Yes

2 a	Explain decision tree learning with its structure, advantages, and disadvantages. **Structure explanation- 1M**	5	CO3	L2

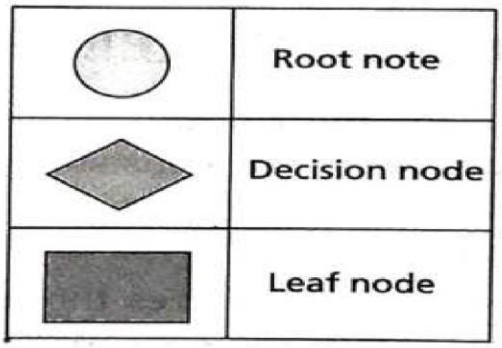


Figure 6.1: Nodes in a Decision Tree

Advantages- 2M

Advantages of Decision Trees

1. Easy to model and interpret
2. Simple to understand
3. The input and output attributes can be discrete or continuous predictor variables.
4. Can model a high degree of nonlinearity in the relationship between the target variables and the predictor variables
5. Quick to train

Disadvantages- 2M

Advantages of Decision Trees

1. Easy to model and interpret
2. Simple to understand
3. The input and output attributes can be discrete or continuous predictor variables.
4. Can model a high degree of nonlinearity in the relationship between the target variables and the predictor variables
5. Quick to train

2 b	Explain pruning in decision tree with an example. Explanation – 5M	5	CO3	L2

Inductive Bias in Decision Trees:

- Inductive bias is necessary for learning algorithms to generalize from training data to unseen data.
- In the ID3 algorithm, the bias favors shorter trees and attributes with high information gain.
- ID3 builds a single decision tree using a hill-climbing search that may not find the global optimum.
- Occam's Razor is used: the simplest tree (shortest) is preferred.

Overfitting in Decision Trees:

- Overfitting occurs when a tree performs well on training data but poorly on test data.
- This happens due to the tree being too complex, capturing noise rather than patterns.
- There is a tradeoff between accuracy and complexity.

Pruning to Prevent Overfitting:

- Pruning improves decision tree generalization.
- **Pre-pruning:** Stops tree growth early.
- **Post-pruning:** Trims the tree after it is fully built.
- Data is split into training (40%), validation, and testing (60%).
- Validation data helps determine where pruning should occur by measuring misclassifications

<table>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>3</td><td>

Define prior probability. Explain Bayes theorem, h_{ML} and h_{MAP} with an example

Prior Probability – 2M
Prior probability is the initial likelihood of an event occurring before any new evidence or observation is taken into account. It reflects what is believed based on existing knowledge, prior to collecting new data.

Bayes Theorem- 3M

P (Hypothesis h | Evidence E) is calculated from the prior probability P (Hypothesis h), the likelihood probability P (Evidence E | Hypothesis h) and the marginal probability P (Evidence E). It can be written as:

$$P(\text{Hypothesis } h \mid \text{Evidence } E) = \frac{P(\text{Evidence } E \mid \text{Hypothesis } h)\, P(\text{Hypothesis } h)}{P(\text{Evidence } E)} \quad (8.1)$$

h_{ML} and h_{MAP} -3M

Maximum A Posteriori (MAP) Hypothesis, h_{MAP}

Given a set of candidate hypotheses, the hypothesis which has the maximum value is considered the *maximum probable hypothesis* or *most probable hypothesis*. This most probable hypothesis is called the Maximum A Posteriori Hypothesis h_{MAP}. Bayes theorem Eq. (8.1) can be used to find the h_{MAP}.

$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(Hypothesis\, h \mid Evidence\, E) \\ &= \max_{h \in H} \frac{P(Evidence\, E \mid Hypothesis\, h) P(Hypothesis\, h)}{P(Evidence\, E)} \\ &= \max_{h \in H} P(Evidence\, E \mid Hypothesis\, h) P(Hypothesis\, h) \quad (8.2) \end{aligned}$$

Maximum Likelihood (ML) Hypothesis, h_{ML}

Given a set of candidate hypotheses, if every hypothesis is equally probable, only $P(E \mid h)$ is used to find the *most probable hypothesis*. The hypothesis that gives the maximum likelihood for $P(E \mid h)$ is called the Maximum Likelihood (ML) Hypothesis, h_{ML}.

$$h_{ML} = \max_{h \in H} P(Evidence\, E \mid Hypothesis\, h) \quad (8.3)$$

Example- 2M

</td><td>10</td><td>CO4</td><td>L2</td></tr>
<tr><td>4a</td><td>

Explain different types of artificial neural network with diagram
Explanation with diagram any 3

Feed Forward Neural Network

- **Structure**: Simple layers where information flows in one direction—from input to output.
- **Features**: May or may not have a hidden layer. No backpropagation.
- **Use**: Suitable for simple classification and image processing tasks.
- **Limitations**: Not suitable for complex learning problems.

Fully Connected Neural Network

- **Structure**: Every neuron in one layer is connected to every neuron in the next layer.
- **Use**: Allows for more complex representations and learning due to full connectivity.
- **Note**: It's a more specific structure within feedforward networks.

Multi-Layer Perceptron (MLP)

- **Structure**: Multiple layers (input, hidden, and output). Fully connected.
- **Features**: Includes forward propagation and backpropagation.
- **Use**: Complex tasks like deep learning, speech recognition, medical

</td><td>5</td><td>CO4</td><td>L2</td></tr>
</table>

diagnosis.

- **Note**: Learning occurs through weight adjustment using errors from predictions.

Feedback Neural Network

- **Structure**: Connections allow signals to flow both forward and backward.
- **Features**: Neurons in later layers can influence earlier layers.
- **Use**: Suitable for dynamic learning tasks. More complex due to feedback loops.

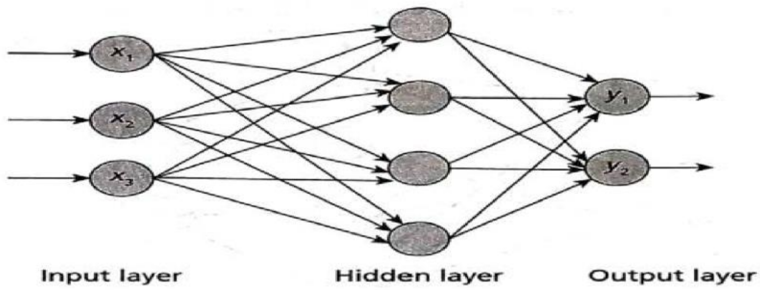


Figure 10.8: Model of a Fully Connected Neural Network

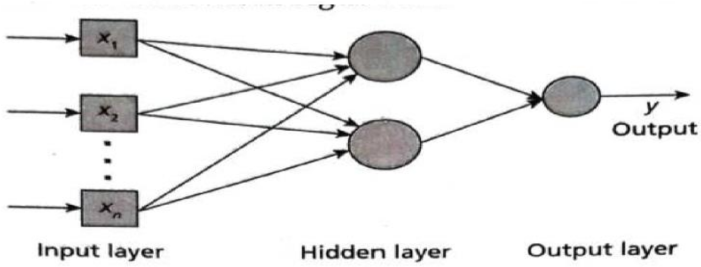


Figure 10.7: Model of a Feed Forward Neural Network

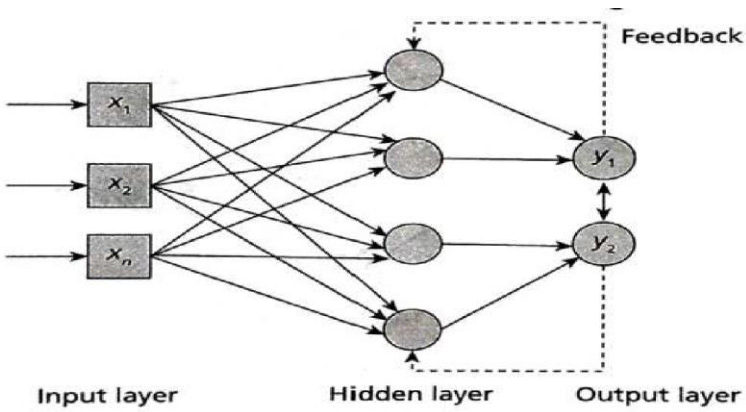


Figure 10.10: Model of a Feedback Neural Network

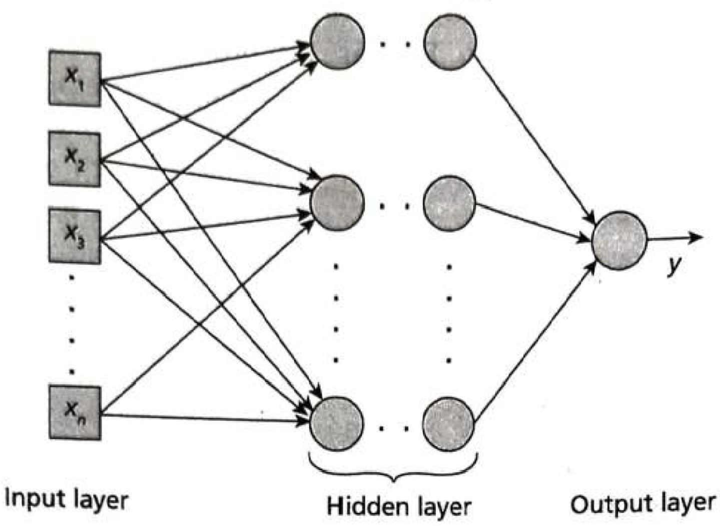


Figure 10.9: Model of a Multi-Layer Perceptron

4b	Define activation function. Explain different types of activation function	5	CO4	L2

	Any 3 Activation Function Below are some of the activation functions used in ANNs: 1. **Identity Function or Linear Function** $$f(x) = x \ \forall x \quad (10.4)$$ The value of $f(x)$ increases linearly or proportionally with the value of x. This function is useful when we do not want to apply any threshold. The output would be just the weighted sum of input values. The output value ranges between $-\infty$ and $+\infty$. 2. **Binary Step Function** $$f(x) = \begin{cases} 1 & if\ f(x) \geq \theta \\ 0 & if\ f(x) < \theta \end{cases} \quad (10.5)$$ The output value is binary, i.e., 0 or 1 based on the threshold value θ. If value of $f(x)$ is greater than or equal to θ, it outputs 1 or else it outputs 0. 3. **Bipolar Step Function** $$f(x) = \begin{cases} 1 & if\ f(x) \geq \theta \\ -1 & if\ f(x) < \theta \end{cases} \quad (10.6)$$ The output value is bipolar, i.e., +1 or −1 based on the threshold value θ. If value of $f(x)$ is greater than or equal to θ, it outputs +1 or else it outputs −1. 4. **Sigmoidal Function or Logistic Function** $$\sigma(x) = \frac{1}{1+e^{-x}} \quad (10.7)$$ It is a widely used non-linear activation function which produces an S-shaped curve and the output values are in the range of 0 and 1. It has a vanishing gradient problem, i.e., no change in the prediction for very low input values and very high input values. 5. **Bipolar Sigmoid Function** $$\sigma(x) = \frac{1-e^{-x}}{1+e^{-x}} \quad (10.8)$$ It outputs values between −1 and +1. 6. **Ramp Functions** $$f(x) = \begin{cases} 1 & if\ x > 1 \\ x & if\ 0 \leq x \leq 1 \\ 0 & if\ x < 0 \end{cases} \quad (10.9)$$ It is a linear function whose upper and lower limits are fixed. 7. **Tanh – Hyperbolic Tangent Function** The Tanh function is a scaled version of the sigmoid function which is also non-linear. It also suffers from the vanishing gradient problem. The output values range between −1 and 1. $$\tan h(x) = \frac{2}{1+e^{-2x}} - 1 \quad (10.10)$$			
5a	Calculate the Euclidean, Manhattan and chebyshev distance (a) (2, 3, 4) and (1, 5, 6) --------2.5M (b) (2, 2, 9) and (7, 8, 9)---------2.5M Solution:	5	CO5	L3

	a. (2 3 4) and (1 5 6) **Solution** **Euclidean distance =** $\sqrt{(2-1)^2+(3-5)^2+(4-6)^2}=\sqrt{9}=3$ Manhattan distance = $\lvert 2-1\rvert+\lvert 3-5)\rvert+\lvert 4-6\rvert=1+2+2=5$ Chebyshev Distance = $\max\{\lvert 2-1\rvert,\lvert 3-5)\rvert,\lvert 4-6\rvert\}=\max\{1,2,2\}=2$ b. (2 2 9) and (7 8 9) Euclidean Distance = $\sqrt{(2-7)^2+(2-8)^2+(9-9)^2}=\sqrt{25+36+09}=\sqrt{61}=7.81$ Manhattan Distance = $\lvert 2-7\rvert+\lvert 2-8)\rvert+\lvert 9-9\rvert=5+6+0=11$ Chebyshev Distance = $\max\{\lvert 2-7\rvert+\lvert 2-8)\rvert+\lvert 9-9\rvert\}=\{5,6,0\}=6$			
5b	For the given pairs of binary vectors, compute the following similarity measures: Cosine Similarity & Simple Matching Coefficient (SMC) (a) (1, 0, 1, 1) and (1, 1, 0, 0)-------2M (b) (1, 0, 0, 0, 1) and (1, 0, 0, 0, 1) and (1, 1, 0, 0, 0)----------3M a. (1 0 1 1) and (1 1 0 0) **Solution** **1 0 1 1** **1 1 0 0** **C = 2, b = 1, d = 1,** **SMC =** $\dfrac{a+d}{a+b+c+d}=\dfrac{1}{4}=0.25$ Cosine Similarity = $\dfrac{(1\times1+0\times1+1\times0+1\times0)}{\sqrt{3}\sqrt{2}}=\dfrac{1}{\sqrt{3}\sqrt{2}}$ = 0.408 **(b) Vectors:** 1. **(1, 0, 0, 0, 1)** and **(1, 0, 0, 0, 1)** 2. **(1, 0, 0, 0, 1)** and **(1, 1, 0, 0, 0)** **Pair 1: (1, 0, 0, 0, 1) and (1, 0, 0, 0, 1)** **Cosine Similarity**: • Vectors are identical ⇒ cosine similarity = 1.0 **SMC**: SMC: • All 5 elements match $\text{SMC}=\frac{5}{5}=\mathbf{1.0}$	5	CO5	L3

Pair 2: **(1, 0, 0, 0, 1)** and **(1, 1, 0, 0, 0)**

Step 1: Cosine Similarity

- Dot product: $\mathbf{1} \cdot \mathbf{1} + 0 \cdot \mathbf{1} + 0 \cdot 0 + 0 \cdot 0 + \mathbf{1} \cdot 0 = \mathbf{1}$
- $|A| = \sqrt{(1^2 + 0 + 0 + 0 + 1^2)} = \sqrt{2}$
- $|B| = \sqrt{(1^2 + 1^2 + 0 + 0 + 0)} = \sqrt{2}$

$$\text{Cosine Similarity} = \frac{1}{\sqrt{2} \cdot \sqrt{2}} = \frac{1}{2} = 0.5$$

- Matches: 3

$$\text{SMC} = \frac{3}{5} = 0.6$$

Q.No	Question	Marks	CO	Level
6	Apply k means clustering algorithm for the given data with initial value of objects 2 and 5 considered as initial seeds. Solution – 10M	10	CO5	L3

Objects	X-Coordinate	Y-Coordinate
1	2	4
2	4	6
3	6	8
4	10	4
5	12	4

Table 13.9: Sample Data

Objects	X-coordinate	Y-coordinate
1	2	4
2	4	6
3	6	8
4	10	4
5	12	4

Solution: As per the problem, choose the objects 2 and 5 with the coordinate values. Hereafter, the objects' id is not important. The samples or data points (4, 6) and (12, 4) are started as two clusters as shown in Table 13.10.

Initially, centroid and data points are same as only one sample is involved.

Table 13.10: Initial Cluster Table

Cluster 1	Cluster 2
(4, 6)	(12, 4)
Centroid 1 (4, 6)	Centroid 2 (12, 4)

Iteration 1: Compare all the data points or samples with the centroid and assign to the nearest sample. Take the sample object 1 (2, 4) from Table 13.9 and compare with the centroid of

the clusters in Table 13.10. The distance is 0. Therefore, it remains in the same cluster. Similarly, consider the remaining samples. For the object 1 (2, 4), the Euclidean distance between it and the centroid is given as:

$$\text{Dist (1, centroid 1)} = \sqrt{(2-4)^2 + (4-6)^2} = \sqrt{8}$$

$$\text{Dist (1, centroid 2)} = \sqrt{(2-12)^2 + (4-4)^2} = \sqrt{100} = 10$$

Object 1 is closer to the centroid of cluster 1 and hence assign it to cluster 1. This is shown in Table 13.11. Object 2 is taken as centroid point.

For the object 3 (6, 8), the Euclidean distance between it and the centroid points is given as:

$$\text{Dist (3, centroid 1)} = \sqrt{(6-4)^2 + (8-6)^2} = \sqrt{8}$$

$$\text{Dist (3, centroid 2)} = \sqrt{(6-12)^2 + (8-4)^2} = \sqrt{52}$$

Object 3 is closer to the centroid of cluster 1 and hence remains in the same cluster 1.

Proceed with the next point object 4(10, 4) and again compare it with the centroids in Table 13.10.

$$\text{Dist (4, centroid 1)} = \sqrt{(10-4)^2 + (4-6)^2} = \sqrt{40}$$

$$\text{Dist (4, centroid 2)} = \sqrt{(10-12)^2 + (4-4)^2} = \sqrt{4} = 2$$

Object 4 is closer to the centroid of cluster 2 and hence assign it to the cluster table. Object 4 is in the same cluster. The final cluster table is shown in Table 13.11.

Obviously, Object 5 is in Cluster 3. Recompute the new centroids of cluster 1 and cluster 2. They are (4, 6) and (11, 4), respectively.

Table 13.11: Cluster Table After Iteration 1

Cluster 1	Cluster 2
(4, 6)	(10, 4)
(2, 4)	(12, 4)
(6, 8)	
Centroid 1 (4, 6)	Centroid 2 (11, 4)

The second iteration is started again with the Table 13.11.

Obviously, the point (4, 6) remains in cluster 1, as the distance of it with itself is 0. The remaining objects can be checked. Take the sample object 1 (2, 4) and compare with the centroid of the clusters in Table 13.12.

$$\text{Dist (1, centroid 1)} = \sqrt{(2-4)^2 + (4-6)^2} = \sqrt{8}$$

$$\text{Dist (1, centroid 2)} = \sqrt{(2-11)^2 + (4-4)^2} = \sqrt{81} = 9$$

Object 1 is closer to centroid of cluster 1 and hence remains in the same cluster. Take the sample object 3 (6, 8) and compare with the centroid values of clusters 1 (4, 6) and cluster 2 (11, 4) of the Table 13.12.

$$\text{Dist (3, centroid 1)} = \sqrt{(6-4)^2 + (8-6)^2} = \sqrt{8}$$

$$\text{Dist (3, centroid 2)} = \sqrt{(6-11)^2 + (8-4)^2} = \sqrt{41}$$

	Object 3 is closer to centroid of cluster 1 and hence remains in the same cluster. Take the sample object 4 (10, 4) and compare with the centroid values of clusters 1 (4, 6) and cluster 2 (11, 4) of the Table 13.12: Dist (4, centroid 1) = $\sqrt{(10-4)^2 + (4-6)^2} = \sqrt{40}$ Dist (3, centroid 2) = $\sqrt{(10-11)^2 + (4-4)^2} = \sqrt{1} = 1$ Object 3 is closer to centroid of cluster 2 and hence remains in the same cluster. Obviously, the sample (12, 4) is closer to its centroid as shown below: Dist (5, centroid 1) = $\sqrt{(12-4)^2 + (4-6)^2} = \sqrt{68}$ Dist (5, centroid 2) = $\sqrt{(12-11)^2 + (4-4)^2} = \sqrt{1} = 1$. Therefore, it remains in the same cluster. Object 5 is taken as centroid point. The final cluster Table 13.12 is given below:			

Table 13.12: Cluster Table After Iteration 2

Cluster 1	Cluster 2
(4, 6)	(10, 4)
(2, 4)	(12, 4)
(6, 8)	
Centroid (4, 6)	Centroid (11, 4)

There is no change in the cluster Table 13.12. It is exactly the same; therefore, the *k*-means algorithm terminates with two clusters with data points as shown in the Table 13.12.

Faculty Signature CCI Signature HOD Signature