CBCS SCHEME

USN

Sixth Semester B.E./B.Tech. Degree Examination, June/July 2025
Machine Learning

Time: 3 hrs. Max. Marks: 100

Note: 1. Answer any FIVE full questions, choosing ONE full question from each module.
2. M : Marks , L: Bloom's level , C: Course outcomes.

		Module – 1	M	L	C
Q.1	a.	State Tom Mitchell's definition of machine learning. List and explain the challenges of machine learning.	7	L1	CO1
	b.	List and explain the visualization aids available for univariate data analysis with example for each.	7	L2	CO1
	c.	For the patients age list {12, 14, 19, 22, 24, 26, 28, 31, 34}. Find the IQR.	6	L3	CO1
		OR			
Q.2	a.	Explain in detail the machine learning process with a neat diagram.	7	L2	CO1
	b.	Explain data preprocessing with measures to solve the problem of missing data.	7	L2	CO1
	c.	Find the 5-point summary of the list {13, 11, 2, 3, 4, 8, 9} and plot the box plot for the same.	6	L3	CO1

Module – 2

Q.3 a. Let the data points be $\begin{pmatrix}2\\6\end{pmatrix}$ and $\begin{pmatrix}1\\7\end{pmatrix}$. Apply Principal Component Analysis (PCA) and find the transformed data. (10, L3, CO1)

b. Apply candidate elimination algorithm on the dataset given in Table Q.3(b) to obtain the complete version space. (10, L3, CO2)

Table Q.3(b)

CGPA	Interactiveness	Practical knowledge	Communication skills	Logical thinking	Job offer
≥ 9	Yes	Excellent	Good	Fast	YES
≥ 9	Yes	Good	Good	Fast	YES
≥ 8	No	Good	Good	Fast	NO
≥ 9	Yes	Good	Good	Slow	YES

OR

Q.4 a. Find Singular Value Decomposition (SVD) of the matrix $A = \begin{pmatrix}1 & 2\\4 & 9\end{pmatrix}$. (10, L3, CO2)

b. Write Find-S algorithm. Apply the algorithm to obtain the hypothesis for the dataset given in the Table Q.4(b). (10, L3, CO2)

Table Q.4(b)

Sky	Air temp	Humidity	Wind	Water	Forecast	Enjoy sport
Sunny	Warm	Normal	Strong	Warm	Same	YES
Sunny	Warm	High	Strong	Warm	Same	YES
Rainy	Cold	High	Strong	Warm	Change	NO
Sunny	Warm	High	Strong	Cool	Change	YES

Module – 3

Q.5 a. Apply K-nearest neighbor algorithm, for the dataset given in Table Q.5(a). Given a test instance (6.1, 40, 5), use the training set to classify the test instance. Choose K = 3. (6, L3, CO3)

Table Q.5(a)

CGPA	Assessment	Project submitted	Result
9.2	85	8	PASS
8	80	7	PASS
8.5	81	8	PASS
6	45	5	FAIL
6.5	50	4	FAIL
5.8	38	5	FAIL

b. Explain types of regression methods and limitations of regression methods. (7, L2, CO3)

c. Explain the structure of a decision tree and write the procedure to construct a decision the using ID3 algorithm. (7, L2, CO3)

OR

Q.6 a. Write the nearest-centroid classifier algorithm. Apply the same to predict the class for the given test instance (6, 5) using the training dataset given in Table Q.6(a). (7, L3, CO3)

X	Y	Class
3	1	A
5	2	A
4	3	A
7	6	B
6	7	B
8	5	B

Table Q.6(a)

b. Distinguish between (6, L2, CO3)
i) Regression and correlation
ii) Regression and causation
iii) Linearity and non-linearity relationships.

c. Explain the advantages and disadvantages of decision tree. Write the general algorithm for decision tree. (7, L2, CO3)

Module – 4

Q.7	a.	Using Naïve bayes classifier classify the new data (Red, SUV, Domestic) using the training dataset given in Table Q.7(a).	10	L3	CO4

Table Q.7(a)

Color	Type	Origin	Stolen
Red	Sports	Domestic	YES
Red	Sports	Domestic	NO
Red	Sports	Domestic	YES
Yellow	Sports	Domestic	NO
Yellow	Sports	Imported	YES
Yellow	SUV	Imported	NO
Yellow	SUV	Imported	YES
Yellow	SUV	Domestic	NO
Red	SUV	Imported	NO
Red	Sports	Imported	YES

	b.	Explain the simple model of an artificial neuron along with the artificial neural network structure.	10	L2	CO4

OR

Q.8	a.	Explain Bayes theorem, Maximum A Posteriori (MAP) hypothesis and Maximum Likelihood (ML) hypothesis in detail.	10	L2	CO4
	b.	Explain different activation functions used in artificial neural network.	10	L2	CO4

Module – 5

Q.9	a.	Consider the following set of data given in Table Q.9(a). Cluster it using K-means algorithm with initial value of objects 2 and 5 with the coordinate values (4, 6) and (12, 4) as initial seeds.	10	L3	CO5

Table Q.9(a)

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

	b.	Explain the various components of reinforcement learning.	10	L2	CO5

OR

Q.10	a.	Find the Manhattan and Chebyshev distance if the coordinates of the objects are (0, 3) and (5, 8).	4	L3	CO5
	b.	Explain the mean shift clustering algorithm.	6	L2	CO5
	c.	List and explain the i) Characteristics of reinforcement learning ii) Challenges of reinforcement learning iii) Applications of reinforcement learning	10	L3	CO5

* * * * *

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

VTUEXAMINATION June-2025/July-2025
SCHEME OF EVALUATION

Sub:	**Machine Learning**				Sub Code:	**BCS602**	Branch:	**ISE**
Exam Date:	**01/07/2025**	Duration:	**3 Hrs**	MaxMarks:	**100**	Sem/Sec:	**VI/ A,B&C**	**OBE**

Answer any **FIVE FULL** Questions	MARKS	CO	RBT
Solutions			

1.a

Tom Mitchell's definition of machine learning states that,

- "A computer program is said to learn from experience E, with respect to task T and some performance measure P, if its performance on T measured by P improves with experience E." The important components of this definition are **experience E, task T, and performance measure P.**

CHALLENGES OF MACHINE LEARNING

Problems that can be dealt with Machine Learning
Problems – Machine learning can deal with the 'well-posed' problems where specifications are complete and available.
2. Huge data – This is a primary requirement of machine learning. Availability of a quality data is a challenge.
A quality data means it should be large and should not have data problems such as missing data or incorrect data.
3. High computation power – With the availability of Big Data, the computational resource requirement has also increased.
Systems with Graphics Processing Unit (GPU) or even Tensor Processing Unit (TPU) are required to execute machine learning algorithms.
4.Complexity of the algorithms – The selection of algorithms, describing the algorithms, application of algorithms to solve machine learning task, and comparison of algorithms have become necessary for machine learning or data scientists now
5. Bias/Variance – Variance is the error of the model. This leads to a problem called bias/ variance tradeoff.

- A model that fits the training data correctly but fails for test data, in general lacks generalization, is called overfitting.
- The reverse problem is called underfitting where the model fails for training data but has good generalization.

MARKS	CO	RBT
7 M	CO1	L1

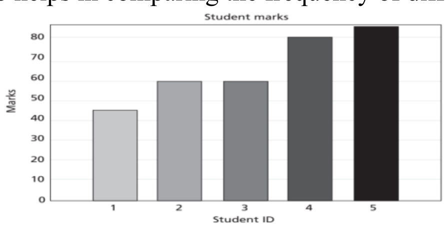
Data Visualization

1b.

- To understand data, graph visualization is must. Data visualization helps to understand data.
- It helps to present information and data to customers. Some of the graphs that are used in univariate data analysis are **bar charts, histograms, frequency polygons and pie charts.**
- The advantages of the graphs are presentation of data, summarization of data, description of data, exploration of data, and to make comparisons of data.

Bar Chart

- A Bar chart (or Bar graph) is used to display the frequency distribution for variables.
- Bar charts are used to illustrate discrete data. The charts can also help to explain the counts of nominal data. It also helps in comparing the frequency of different groups.



MARKS	CO	RBT
7M	CO1	L2

Pie Chart These are equally helpful in illustrating the univariate data. The percentage frequency distribution of students' marks {22, 22, 40, 40, 70, 70, 70, 85, 90, 90} is below in Figure 2.4.

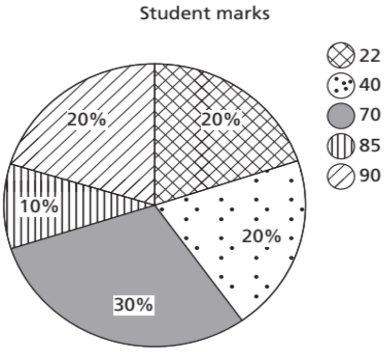


Figure 2.4: Pie Chart

It can be observed that the number of students with 22 marks are 2. The total number of students are 10. So, 2/10 × 100 = 20% space in a pie of 100% is allotted for marks 22 in Figure 2.4.

Histogram It plays an important role in data mining for showing frequency distributions. The histogram for students' marks {45, 60, 60, 80, 85} in the group range of 0–25, 26–50, 51–75, 76–100 is given below in Figure 2.5. One can visually inspect from Figure 2.5 that the number of students in the range 76–100 is 2.

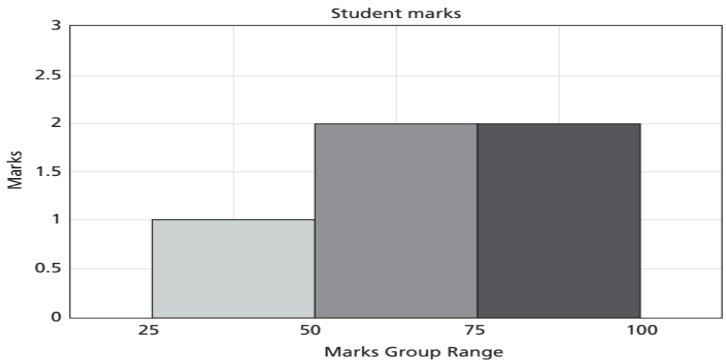
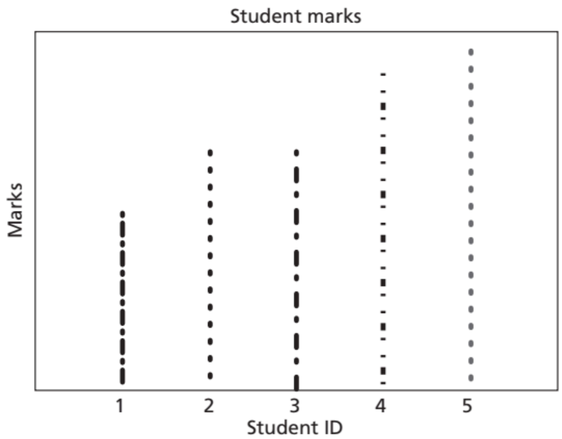


Figure 2.5: Sample Histogram of English Marks

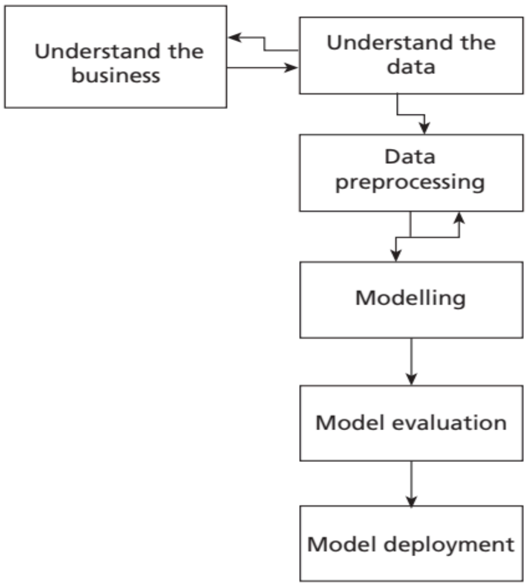
Histogram conveys useful information like nature of data and its mode. Mode indicates the peak of dataset. In other words, histograms can be used as charts to show frequency, skewness present in the data, and shape.

Dot Plots These are similar to bar charts. They are less clustered as compared to bar charts, as they illustrate the bars only with single points. The dot plot of English marks for five students with ID as {1, 2, 3, 4, 5} and marks {45, 60, 60, 80, 85} is given in Figure 2.6. The advantage is that by visual inspection one can find out who got more marks.



1.c	median = 24 $\quad \frac{26+28+31+34}{4} = 29.5$ $Q_{0.25} = 16.5$ $Q_{0.75} = 29.5 \quad \frac{12+14+19+22}{4} = 16.5$ $IQR = Q_{0.75} - Q_{0.25}$ $29.5 - 16.5$ $\boxed{IQR = 13}$	6M	CO1	L3

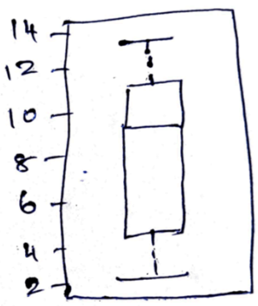
<table>
<tr><td>2.a</td><td>



- 1. Understanding the business – This step involves understanding the objectives and requirements of the business organization. Generally, a single data mining algorithm is enough for giving the solution. This step also involves the formulation of the problem statement for the data mining process.
- 2. Understanding the data – It involves the steps like data collection, study of the characteristics of the data, formulation of hypothesis, and matching of patterns to the selected hypothesis.
- 3. Preparation of data – This step involves producing the final dataset by cleaning the raw data and preparation of data for the data mining process. The missing values may cause problems during both training and testing phases. Missing data forces classifiers to produce inaccurate results. This is a perennial problem for the classification models. Hence, suitable strategies should be adopted to handle the missing data
- 4. Modelling – This step plays a role in the application of data mining algorithm for the data to obtain a model or pattern.
- 5. Evaluate – This step involves the evaluation of the data mining results using statistical analysis and visualization methods. The performance of the classifier is determined by evaluating the accuracy of the classifier. The process of classification is a fuzzy issue. For example, classification of emails requires extensive domain knowledge and requires domain experts. Hence, performance of the classifier is very crucial.
- 6. Deployment – This step involves the deployment of results of the data mining algorithm to improve the existing process or for a new situation.

</td><td>7M</td><td>CO1</td><td>L2</td></tr>
<tr><td>2c</td><td>

Minimum = 2
Maximum = 13
$Q_1 = 3$
$Q_2 = 8$
$Q_3 = 11$
5- point summary is {2, 3, 8, 11, 13)



</td><td>6M</td><td>CO1</td><td>L3</td></tr>
</table>

3a.

Let the data points be $\begin{pmatrix}2\\6\end{pmatrix}$ and $\begin{pmatrix}1\\7\end{pmatrix}$. Apply PCA and find the transformed data.

10M CO1 L3

solution:

The mean vector can be calculated as,

$$\mu = \frac{x_1 + x_2}{2}, \quad \mu = \begin{pmatrix}\frac{2+1}{2}\\ \frac{6+7}{2}\end{pmatrix} = \begin{pmatrix}1.5\\6.5\end{pmatrix}$$

centering the Data:-

The mean must subtracted from the data to get the adjusted data,

$$x_1 = X_1 - \mu = \begin{pmatrix}2-1.5\\6-6.5\end{pmatrix} = \begin{pmatrix}0.5\\-0.5\end{pmatrix}$$

$$x_2 = X_2 - \mu = \begin{pmatrix}1-1.5\\7-6.5\end{pmatrix} = \begin{pmatrix}-0.5\\0.5\end{pmatrix}$$

solving for λ,

case 1 : $0.5 - \lambda = 0.5$

$\lambda = 0.5 - 0.5 = 0$

$\boxed{\lambda = 0}$

case 2 : $0.5 - \lambda = -0.5$

$\lambda = 0.5 + 0.5$

$\boxed{\lambda = 1}$ $\boxed{\text{Eigen values } \lambda_1 = 1, \lambda_2 = 0}$

$$\|v\| = \sqrt{(-1)^2 + (1)^2} = \sqrt{1+1} = \sqrt{2}$$

Normalize the vector, v divide each component by the magnitude,

$$\frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\end{pmatrix} \Rightarrow \begin{pmatrix}\frac{-1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$$

compute AA^T

$$AA^T = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2}+\frac{1}{2} & -\frac{1}{2}+\frac{1}{2} \\ -\frac{1}{2}+\frac{1}{2} & \frac{1}{2}+\frac{1}{2} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$$

The transformation of the matrix using the equation,

$$y = A(x-m)$$

$x \rightarrow$ original datapoint

$m \rightarrow$ mean of the datapoint

$A \rightarrow$ Transformation Matrix

$x-m \rightarrow$ mean-adjusted data.

$$y = A(x-m) = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$ { 0.5 as $\frac{1}{2}$ for convenience }

$$y = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix}$$

transformation

3b	$S_0 = \langle \phi, \phi, \phi, \phi, \phi \rangle$ $S_1 : \langle \geq 9, yes, Ex, Good, fast \rangle$ $S_2 : \langle \geq 9, yes, ?, Good, fast \rangle$ $S_3 : \langle \geq 9, yes, ?, Good, Fast \rangle$ $S_4 : \langle \geq 9, yes, ?, Good, ? \rangle$ $\boxed{\langle \geq 9\ yes\ ?\ ?\ ? \rangle \langle \geq 9, ?, ?, good\ ? \rangle \langle ?\ yes\ ?\ good\ ? \rangle}$ $G_4 : \langle \geq 9, ?\ ?\ ?\ ? \rangle \langle ?\ yes, ?\ ?\ ? \rangle$ $G_3 : \langle \geq 9, ?\ ?\ ?\ ? \rangle \langle ?\ yes, ?\ ?\ ? \rangle \langle ?\ ?\ ?\ ?\ ? \rangle$ $G_2 : \langle ?\ ?\ ?\ ?\ ?\ ? \rangle$ $G_1 : \langle ?\ ?\ ?\ ?\ ? \rangle$ $G_0 : \langle ?\ ?\ ?\ ?\ ? \rangle$	10M	CO2	L3

4a	①. Find SVD of the matrix $A = \begin{pmatrix} 1 & 2 \\ 4 & 9 \end{pmatrix}$ The first step is to compute $AA^T = \begin{pmatrix} 1 & 2 \\ 4 & 9 \end{pmatrix}\begin{pmatrix} 1 & 4 \\ 2 & 9 \end{pmatrix} = \begin{pmatrix} 5 & 22 \\ 22 & 97 \end{pmatrix}$ Eigen values :- subtract λI from AA^T $AA^T - \lambda \Rightarrow \begin{pmatrix} 5-\lambda & 22 \\ 22 & 97-\lambda \end{pmatrix}$ The determinant is $\begin{vmatrix} 5-\lambda & 22 \\ 22 & 97-\lambda \end{vmatrix} = 0$ using the determinant formula for 2x2 matrix. $(5-\lambda)(97-\lambda) - (22\times 22) = 0$ $5\times 97 - 5\lambda - 97\lambda + \lambda^2 - 484 = 0$ $485 - 5\lambda - 97\lambda + \lambda^2 - 484 = 0$ $\lambda^2 - 102\lambda + 1 = 0$ using quadratic formula, $\lambda = \frac{-(-102) \pm \sqrt{(-102)^2 - 4(1)(1)}}{2(1)}$ Compute eigen values, $\lambda_1 = \frac{102 + 101.99}{2} = \frac{203.99}{2} = 101.9902$ $\lambda_2 = \frac{102 - 101.99}{2} = \frac{0.01}{2} = 0.0098$ $\boxed{\lambda_1 = 101.9902}$ $\boxed{\lambda_2 = 0.0098}$ Eigen vector for λ_1 is $v_1 = \begin{bmatrix} 0.2268 \\ 1 \end{bmatrix}$	10M	CO2	L3

Matrix U can be obtained by concatenating the above vectors

$$\boxed{U = [u_1, u_2] = \begin{pmatrix} 0.2212 & -0.9752 \\ -0.9752 & 0.2212 \end{pmatrix}}$$

V is obtained by concatenating the normalized eigen vectors

$$\boxed{V = [v_1 \; v_2] = \begin{pmatrix} 0.4082 & -0.9129 \\ 0.9129 & 0.4082 \end{pmatrix}}$$

The eigen values are

$$\lambda_1 = 101.9902, \quad \lambda_2 = 0.0098$$

Taking square root,

$$\sigma_1 = \sqrt{101.9902} \approx 10.099$$

$$\sigma_2 = \sqrt{0.0098} \approx 0.099$$

The diagonal Matrix S is

$$\boxed{S = \begin{pmatrix} 10.099 & 0 \\ 0 & 0.099 \end{pmatrix}}$$

$\therefore$ The Matrix decomposition $A = USV^T$ is done.

4.b	Algorithm steps — 4m $\boxed{h_3 = \langle \text{sunny, warm, ?, strong, ?, ?} \rangle}$ — 6m	10M	CO2	L3
5a	Given the test instance (6.1, 40, 5) Given K=3, Compute the Euclidean distance b/w test instance and each of the training instance CGPA Ass. Proj Result Euclidean Distance 9.2 85 8 Pass $\sqrt{(9.2-6.1)^2+(85-40)^2+(8-5)^2}$ $= 45.2063$ 8 80 7 Pass 40.095 8.5 81 8 Pass 41.179 6 45 5 Fail 5.001 6.5 50 4 Fail 10.057 5.8 38 5 Fail 2.022 $\boxed{\text{Prediction : Fail}}$	6M	CO3	L3

5b	**Classification of Regression Methods** **Limitation of Regression Methods**	4M 3M	CO3	L2
5c	Decision tree is a concept tree which summarizes the information contained in the training dataset in the form of a tree structure. Once the concept model is built, test data can be easily classified	7M	CO3	L2

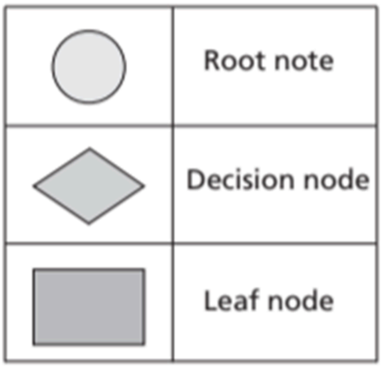


Figure 6.1: Nodes in a Decision Tree

Let P be the probability distribution of data instances from 1 to n as shown in Eq. (6.2).

So, $P = P_1 \ldots\ldots P_n$ (6.2)

Entropy of P is the information measure of this probability distribution given in Eq. (6.3),

$$\text{Entropy\_Info}(P) = \text{Entropy\_Info}(P_1 \ldots P_n)$$
$$= -(P_1 \log_2(P_1) + P_2 \log_2(P_2) + \ldots\ldots + P_n \log_2(P_n)) \quad (6.3)$$

where, P_1 is the probability of data instances classified as class 1 and P_2 is the probability of data instances classified as class 2 and so on.

P_1 = |No of data instances belonging to class 1| / |Total no of data instances in the training dataset|

Entropy_Info(P) can be computed as shown in Eq. (6.4).

Thus,

Entropy_Info(6, 4) is calculated as $-\left[\frac{6}{10}\log_2\frac{6}{10} + \frac{4}{10}\log_2\frac{4}{10}\right]$ (6.4)

Mathematically, entropy is defined in Eq. (6.5) as:

$$\text{Entropy\_Info}(X) = \Sigma_{x \in \text{values}(X)} Pr[X = x] \cdot \log_2 \frac{1}{Pr[X = x]} \quad (6.5)$$

$\Pr[X = x]$ is the probability of a random variable X with a possible outcome x.

Note: $\log_2 \frac{1}{Pr[X = x]} = -\log_2(Pr[X = x])$

2. Split the training dataset into subsets based on the outcomes of the test attribute and each subset in a branch contains the data instances or tuples with the same value for the selected test attribute.
3. Repeat step 1 and step 2 on each subset until we end up in leaf nodes in all the branches of the tree.
4. This splitting process is recursive until the stopping criterion is reached.

Stopping Criteria

The following are some of the common stopping conditions:

1. The data instances are homogenous which means all belong to the same class C_i and hence its entropy is 0.
2. A node with some defined minimum number of data instances becomes a leaf (Number of data instances in a node is between 0.25 and 1.00% of the full training dataset).
3. The maximum tree depth is reached, so further splitting is not done and the node becomes a leaf node.

Expected information or Entropy needed to classify a data instance d' in T is denoted as Entropy_Info(T) given in Eq. (6.8).

$$\text{Entropy\_Info}(T) = -\sum_{i=1}^{m} P_i \log_2 P_i \tag{6.8}$$

Entropy of every attribute denoted as Entropy_Info(T, A) is shown in Eq. (6.9) as:

$$\text{Entropy\_Info}(T, A) = \sum_{i=1}^{v} \frac{|A_i|}{|T|} \times \text{Entropy\_Info}\ (A_i) \tag{6.9}$$

where, the attribute A has got 'v' distinct values $\{a_1, a_2, \ldots a_v\}$, $|A_i|$ is the number of instances for distinct value 'i' in attribute A, and Entropy_Info (A_i) is the entropy for that set of instances.

Information_Gain is a metric that measures how much information is gained by branching on an attribute A. In other words, it measures the reduction in impurity in an arbitrary subset of data.

It is calculated as given in Eq. (6.10):

$$\text{Information\_Gain}(A) = \text{Entropy\_Info}(T) - \text{Entropy\_Info}(T, A) \tag{6.10}$$

It can be noted that as entropy increases, information gain decreases. They are inversely proportional to each other.

6.a	**Solution:** **Step 1:** Compute the mean/centroid of each class. In this example there are two classes called 'A' and 'B'. Centroid of class 'A' = (3 + 5 + 4, 1 + 2 + 3)/3 = (12, 6)/3 = (4, 2) Centroid of class 'B' = (7 + 6 + 8, 6 + 7 + 5)/3 = (21, 18)/3 = (7, 6) Now given a test instance (6, 5), we can predict the class. **Step 2:** Calculate the Euclidean distance between test instance (6, 5) and each of the centroid. Euc_Dist[(6, 5) ; (4, 2)] = $\sqrt{(6-4)^2+(5-2)^2} = \sqrt{13} = 3.6$ Euc_Dist[(6, 5) ; (7, 6)] = $\sqrt{(6-7)^2+(5-6)^2} = \sqrt{2} = 1.414$ The test instance has smaller distance to class B. Hence, the class of this test instance is predicted as 'B'.	7M	CO3	L3
6b.	1. Regression and Correlation Difference 2. Regression and Causation Difference 3. Linearity and Non linearity difference	2M+2M +2M	CO3	L2
6c	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 of Decision Trees •Some of the issues that generally arise with a decision tree learning are that: •1. It is difficult to determine how deeply a decision tree can be grown or when to stop growing it. •2. If training data has errors or missing attribute values, then the decision tree constructed may become unstable or biased. •3. If the training data has continuous valued attributes, handling it is computationally complex and has to be discretized. •4. A complex decision tree may also be over-fitting with the training data. •5. Decision tree learning is not well suited for classifying multiple output classes. •6. Learning an optimal decision tree is also known to be NP-complete.	7M	CO3	L2

6.c

Algorithm 6.1: General Algorithm for Decision Trees

1. Find the best attribute from the training dataset using an attribute selection measure and place it at the root of the tree.

2. Split the training dataset into subsets based on the outcomes of the test attribute and each subset in a branch contains the data instances or tuples with the same value for the selected test attribute.
3. Repeat step 1 and step 2 on each subset until we end up in leaf nodes in all the branches of the tree.
4. This splitting process is recursive until the stopping criterion is reached.

Stopping Criteria

The following are some of the common stopping conditions:

1. The data instances are homogenous which means all belong to the same class C_i and hence its entropy is 0.
2. A node with some defined minimum number of data instances becomes a leaf (Number of data instances in a node is between 0.25 and 1.00% of the full training dataset).
3. The maximum tree depth is reached, so further splitting is not done and the node becomes a leaf node.

7a.

Frequency & likelihood table (color)

	Yes	No
Red	3	2
yellow	2	3

	P(yes)	P(No)
Red	3/5	2/5
Yellow	2/5	3/5

Frequency & likelihood table of Type:

	Yes	No
sports	4	2
SUV	1	3

	P(yes)	P(No)
sports	4/5	2/5
SUV	1/5	3/5

Frequency & Likelihood Table (origin):

	Yes	No
Domestic	2	3
Imported	3	2

	P(yes)	P(No)
Domestic	2/5	3/5
Imported	3/5	2/5

$P(yes/x) = P(Red/yes) * P(SUV/yes) * P(Domestic/yes) * P(yes)$

$= 3/5 \times 1/5 \times 2/5 \times 1 = 0.048$

$P(No/x) = P(Red/No) * P(SUV/No) * P(Domestic/No) * P(No)$

$= 2/5 * 3/5 * 3/5 * 1$

$= 0.144$

$\boxed{\text{Answer = 'NO'}}$

10M CO4 L3

<table>
<tr><td colspan="5">OR</td></tr>
<tr><td>7b</td><td>
Artificial neurons are like biological neurons which are called as nodes. A node or a neuron can receive one or more input information and process it. Artificial neurons or nodes are connected by connection links to one another. Each connection link is associated with a synaptic weight

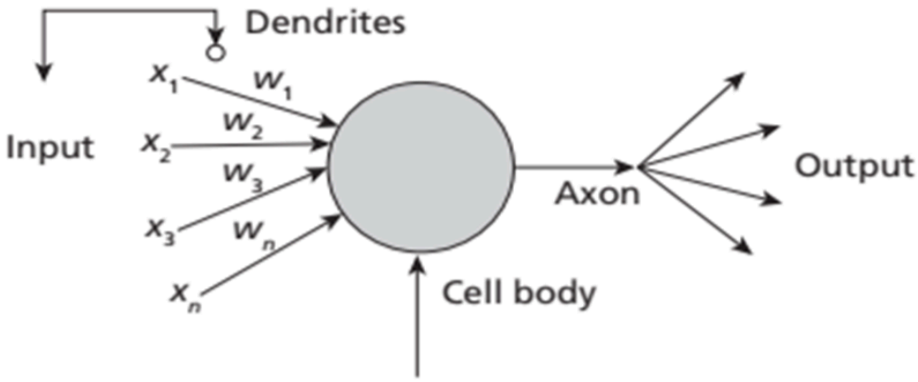


Figure 10.2: An Artificial Neuron

The activation function is a binary step function which outputs a value 1 if the Net-sum is above the threshold value θ, and a 0 if the Net-sum is below the threshold value θ. Therefore, the activation function is applied to Net-sum as shown in Eq. (10.2).

$$f(x) = \text{Activation function } (Net - sum) \qquad (10.2)$$

$$\text{Then, output of a neuron } Y = \begin{cases} 1 & if\ f(x) \geq \theta \\ 0 & if\ f(x) < \theta \end{cases} \qquad (10.3)$$

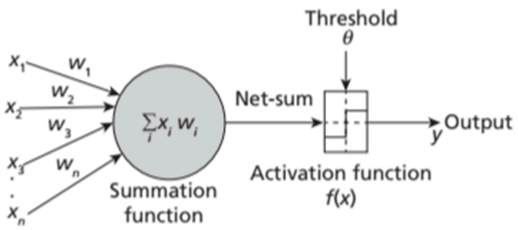


Figure 10.3: McCulloch & Pitts Neuron Mathematical Model
</td><td>10M</td><td>CO4</td><td>L2</td></tr>
</table>

8a

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

Given a set of candidate hypotheses, the hypothesis which has the maximum value is considered as 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}.

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

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

$$= \max_{h \in H} P(Evidence\ E \mid Hypothesis\ h)P(Hypothesis\ h) \quad (8.2)$$

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}.

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

10M CO4 L2

8b

•Activation functions are mathematical functions associated with each neuron in the neural network that map input signals to output signals. (used to introduce non-linearity)

•It decides whether to fire a neuron or not based on the input signals the neuron receives.

•These functions normalize the output value of each neuron either between 0 and 1 or between -1 and +1.

•Typical activation functions can be linear or non-linear.

•Linear functions are useful when the input values can be classified into any one of the two groups and are generally used in binary perceptrons.

•Non-linear functions, on the other hand, are continuous functions that map the input in the range of (0, 1) or (-1, 1), etc.

•These functions are useful in learning high-dimensional data or complex data such as audio, video and images.

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.

10M CO4 L2

4. Sigmoidal Function or Logistic Function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{10.7}$$

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

$$\sigma(x) = \frac{1 - e^{-x}}{1 + e^{-x}} \tag{10.8}$$

It outputs values between –1 and +1.

6. Ramp Functions

$$f(x) = \begin{cases} 1 & if\ x > 1 \\ x & if\ 0 \le x \le 1 \\ 0 & if\ x < 0 \end{cases} \tag{10.9}$$

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.

$$\tan h(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{10.10}$$

9.a	**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.	10M	CO5	L3

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:

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

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:

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

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.

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

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.

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

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.

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

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.

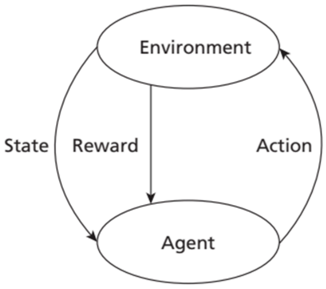
<table>
<tr><td>9b</td><td>•The components of reinforcement learning are shown in Figure 14.4. These are environment, agent, actions and rewards.


Figure 14.4: Basic Components of RL

• There are two types of problems in reinforcement learning – Learning and Planning.

•In learning problems, the environment is unknown and the agent learns by trial and error.

The agent interacts with the environment to improve policy. Planning is another problem where the environment is known and the agent computes with the model and improves policy.

Environment is the world where all actions take place. It is the framework, where the input, output and reward are specified. The environment describes the state or state variables or simply as state. Initially, the environment is in a state called initial state. For example, in a car system, the maps, game rules and obstructions in the road are described in the environment. An agent is an autonomous body that looks at the environment and takes an action. It can be any human or another computer program such as a robot or chatbot</td><td>10M</td><td>CO5</td><td>L2</td></tr>
<tr><td colspan="5"></td></tr>
</table>

10.a	**Solution:** The Euclidean distance using Eq. (13.1) is given as follows: $$\text{Distance}(x_i, x_j) = \sqrt{(0-5)^2 + (3-8)^2} = \sqrt{50} = 7.07$$ The Manhattan distance using Eq. (13.2) is given as follows: $$\text{Distance}(x_i, x_j) = \left|(0-5) + (3-8)\right| = 10$$ The Chebyshev distance using Eq. (13.3) is given as follows: $$\text{Max}\left\{\lvert 0-5\rvert, \lvert 3-8\rvert\right\} = \text{Max}\{5, 5\} = 5$$	4M	CO5	L3
10b	Mean-shift is a non-parametric and hierarchical clustering algorithm. This algorithm is also known as **mode seeking algorithm or a sliding window algorithm**. •It has many applications in **image processing and computer vision**. •There is **no need for any prior knowledge** of clusters or shape of the clusters present in the dataset. •The algorithm **slowly moves from its initial position towards the dense regions**. •The algorithm uses a window, which is basically a **weighting function.** **Algorithm 13.2: Mean-Shift Clustering** **Step 1:** Design a window. **Step 2:** Place the window on a set of data points. **Step 3:** Compute the mean for all the points that come under the window. **Step 4:** Move the center of the window to the mean computed in step 3. Thus, the window moves towards the dense regions. The movement to the dense region is controlled by a mean shift vector. The mean shift vector is given as: $$v_s = \frac{1}{K}\sum_{x_i \in s_k}(x_i - x) \qquad (13.13)$$ Here, K is the number of points and S_k is the data points where the distance from data points x_i and centroid of the kernel x is within the radius of the sphere. Then, the centroid is updated as $x = x + v_s$. **Step 5:** Repeat the steps 3–4 for convergence. Once convergence is achieved, no further points can be accommodated.	6M	CO5	L2

10.c	Characteristics of Reinforcement Learning. •1. Sequential decision making – Consider the Figure 14.3. It can be seen the path from start to goal is not done in one step. It is a sequence of decisions that leads to the goal. One wrong move may result in a failure. This is the main characteristic of reinforcement learning. •2. Delayed feedback – Often, rewards are not immediate. One must spend many moves to get final success or failure. Feedback in terms of reward is often delayed. •3. The agent actions are interdependent as any action affects the subsequent actions. For example, one wrong move of an agent may lead to failure. •4. Time related – All actions are associated with time stamps inherently as all actions are ordered as per the timeline inherently. Challenges of Reinforcement Learning •1. Reward design is a big challenge as in many games, as determining the rewards and its value is a challenge. •2. Absence of a model is a challenge – Games like chess have fixed board and rules. But, many games do not have any fixed environment or rules. There is no underlying model as well. So, simulation must be done to gather experience. •3. Partial observability of states – Many states are fully observable. Imagine a scenario in a weather forecasting where the uncertainty or partial observability exists as complete information about the state is simply not available. •4. Time consuming operations – More state spaces and possible actions may complicate the scenarios, resulting in more time consumption. •5. Complexity – Many games like GO are complicated with much larger board configuration and many possibilities of actions. So, labelled data is simply not available. This adds more complexity to the design of reinforcement algorithms. Applications of Reinforcement Learning There are many applications of RL. Some of the application domains where reinforcement learning is used are listed below: • Industrial automation • Resource management applications to allocate resource • Traffic light controller to reduce congestion of traffic • Personalized recommendation systems like news • Bidding for advertisement • Customized applications • Driverless cars • Along with deep learning games like Chess and GO • Deep mind applications like to generate programs and images	10M	CO5	L3

FacultySignature CCISignature HODSignature