5/5/25 fn

RollNo. | | | | | | | | | | |

ANNA UNIVERSITY (UNIVERSITY DEPARTMENTS)

B.E. /B.Tech / B. Arch (Full Time) - END SEMESTER EXAMINATIONS, APRIL/MAY 2025

INFORMATION TECHNOLOGY
Sixth Semester
IT 5036 Machine Learning
(Regulation2019)

Time:3hrs Max.Marks: 100

CO1	Choose and implement classification or regression algorithms for an application using an open-source tool.
CO2	Implement probabilistic discriminative and generative algorithms for an application and analyse the results.
CO3	Use a tool to implement typical clustering algorithms for different types of applications.
CO4	Design and implement an HMM for a sequence model type of application.
CO5	Implement appropriate learning algorithms for any real-time application using an open-source tool.

BL – Bloom's Taxonomy Levels
(L1-Remembering, L2-Understanding, L3-Applying, L4-Analysing, L5-Evaluating, L6-Creating)

PART- A(10x2=20Marks)
(Answer all Questions)

Q.No.	Questions	Marks	CO	BL
1	Identify the two main sources of error in the bias-variance tradeoff.	2	1	L1
2	List the main types of machine learning based on how models learn from data.	2	1	L1
3	Given a dataset with features such as area, number of bedrooms, and age of the house, explain why linear regression might not always give accurate predictions for house prices. Mention any assumptions involved.	2	2	L2
4	Justify how the KNN algorithm classifies a new data point and explain the role of the parameter 'k' in influencing the result.	2	2	L2
5	How are core points, border points, and outlier points determined in DBSCAN using ε (epsilon) and MinPts? What conditions define each type of point during clustering?	2	3	L2
6	Find the Euclidean and Manhattan distances between the points A(2, 3) and B(7, 8); what does each distance indicate about how the points are spaced?	2	3	L2
7	Name the key components of a Bayesian Network.	2	4	L1
8	Differentiate the assumptions about dependencies that differs between Bayesian Networks and Markov Models and how this affects their use in modelling real-world problems.	2	4	L1
9	State how an agent in reinforcement learning learns from its environment, and describe the role of rewards in shaping its behaviour.	2	5	L1

10	Describe how ensemble learning improves prediction performance and explain why combining multiple models can be more effective than using a single model.	2	5	L1

PART- B(5x 13=65Marks)

(Restrict to a maximum of 2 subdivisions)

Q.No.	Questions	Marks	CO	BL
11 (a)	1. Given a dataset with customer purchase records, outline the steps one would take to build a machine-learning model to predict future purchases. Explain how you would apply the machine learning process in this case. 2. A retail company wants to improve customer experience. Apply your knowledge of supervised, unsupervised, and reinforcement learning to suggest a suitable machine learning type for each of the following tasks: predicting customer churn, segmenting customers by behaviour, and optimising website layout based on user interaction.	6 7	1	L3
	OR			
11 (b)	Apply the Find-S algorithm to the following training data to determine the final hypothesis for the concept 'EnjoySport'. Write the algorithm and explain it step by step with respect to the algorithm.	6+7	1	L3

Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
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

b. Use the candidate elimination method and explain the differences between Find-S and the Candidate elimination method. Write the algorithm and explain it step by step with respect to the algorithm.

12 (a)	1.A data analyst is studying how the number of study hours and sleep hours affect students' test scores. Use the following dataset to apply linear regression and find the relationship between the inputs and the output.	6	2	L3

Study Hours	Sleep Hours	Test Score
4	6	70
6	7	80
8	5	85

Apply linear regression to estimate the coefficients for the model:
Test Score = $\beta_0 + \beta_1 \times$ Study Hours + $\beta_2 \times$ Sleep Hours

2.A teacher wants to predict whether a new student will pass or fail based on historical data. The decision is based on two features: the number of hours studied and the number of classes attended. Use the K-Nearest Neighbors (KNN) algorithm with K=3 to classify the new student with 5 study hours and 7 classes attended.

		7	2	L3

Study Hours	Classes Attended	Result
4	6	Fail
6	8	Pass
5	7	Pass
3	5	Fail
7	9	Pass

Using Euclidean distance, determine the class (Pass/Fail) of the new student.

OR

12(b)	A researcher is using a Support Vector Machine (SVM) to classify data points into two classes: A and B. Each data point has two features: X1 and X2. Analyse the dataset below and determine whether the data is linearly separable. Identify the support vectors and use linear algebra to find the equation of the separating hyperplane.	13	2	L4

X1	X2	Class
2	3	A
3	4	A
4	1	B
5	2	B
3	2	A

Plot the points on a 2D graph. Analyse whether a linear hyperplane can separate the classes. Identify the support vectors (points that lie closest to the decision boundary). Using linear algebra, determine the equation of the hyperplane in the form $w_1 \cdot x_1 + w_2 \cdot x_2 + b = 0$.

13 (a)	A researcher is using hierarchical clustering to group data points based on their pairwise distances. Apply single linkage, complete linkage, and average linkage methods to determine how the clustering structure differs depending on the linkage rule used. The pairwise distance matrix between 4 data points (A, B, C, D) is given below:	13	3	L3

0	A	B	C	D
A	0	2	6	10
B	2	0	5	9
C	6	5	0	4
D	10	9	4	0

Apply the single, complete and average linkage method to cluster the points step-by-step. Compare the dendrograms formed by each method and discuss the differences in cluster formation.

OR

13 (b)	A company is trying to segment its customers based on their behaviour using the K-Means clustering algorithm. The dataset	13	3	L3

below contains four customers, each represented by two features: Annual Spending (in $1000s) and Visit Frequency (per month). Apply the K-Means algorithm with K=2 to perform one full iteration of clustering, starting with customers A and D as initial centroids.

Customer	Annual Spending	Visit Frequency
A	40	5
B	45	8
C	60	2
D	65	1

Plot the customers on a 2D graph based on their features. Euclidean distance should be used to assign each customer to the nearest centroid. Discuss whether the clustering appears reasonable based on the features.

14 (a)	A doctor is using a Bayesian Network to diagnose whether a patient has a certain disease based on symptoms and test results. The network consists of three nodes: Disease (D), Test Result (T), and Symptom (S). The dependencies are as follows: Disease affects both the Test Result and the symptom. Test Results and Symptoms are conditionally independent, given the disease. The conditional probabilities are given below: P(Disease): Disease | Probability Yes / No | 0.1 / 0.9 P(Test Result | Disease): Disease | Positive | Negative Yes | 0.8 | 0.2 No | 0.1 | 0.9 P(Symptom | Disease): Disease | Present | Absent Yes | 0.7 | 0.3 No | 0.2 | 0.8 1. Given that the patient has a positive test result and the symptom is present, use Bayes' Theorem to calculate the posterior probability that the patient has the disease. Analyse how the independence assumption in the network affects the calculation. 2. Discuss how modifying the network structure (e.g., adding a dependency between Test and Symptom) would change the inference.	8 + 5	4	L4
	OR			
14 (b)	You are given an HMM with three hidden states: - Rainy (R), Sunny (S), Cloudy (C) And three observations: - Walk (W), Shop (Sh), Clean (Cl) Initial Probabilities P(R) = 0.5, P(S) = 0.3, P(C) = 0.2	13	4	L4

	Transition Probabilities From \ To / Rainy (R) / Sunny (S) Rainy / 0.6 / 0.2 Sunny / 0.3 / 0.5 Cloudy / 0.4 / 0.3 **Emission Probabilities** State / Walk (W) / Shop (Sh) Rainy / 0.1 / 0.4 Sunny / 0.6 / 0.3 Cloudy / 0.3 / 0.4 **Observation Sequence** O = [Walk, Shop, Clean] O'=[Shop, Walk, Clean] Write the forward algorithm for computing the observation sequence probability in HMM. Apply the algorithm step-by-step to compute P(O) and explain the intermediate steps. Justify the final answer.			
15 (a)	In the following problem, an agent moves in a 2x2 grid world. The agent can move in four directions: Up, down, left, and right. The agent receives a reward of +10 for reaching the goal and 0 elsewhere. Each move has a cost of -1. The agent starts in the top-left corner (A1). A1 (Start) | A2 B1 | B2 (Goal) Write a Q-learning algorithm. Analyse one episode of Q-learning where the agent follows the path: A1 → A2 → B2. Use a learning rate $\alpha = 0.5$ and a discount factor $\gamma = 0.9$. Assume initial Q-values are 0. Calculate Q-values for each visited state-action pair based on the reward received. Discuss how the Q-values are updated and how they guide the learning in future episodes.	**13**	5	L3
	OR			
15 (b)	1. Analyse how AdaBoost determines the importance (weight) of each weak learner during the training process. Explain how the algorithm adjusts the weights of training examples and how this affects the learning of subsequent classifiers. Also, discuss the condition under which a weak learner contributes positively to the final model. 2. Analyse how Random Forest reduces overfitting compared to a single decision tree. Explain the role of bootstrap sampling and random feature selection in creating diverse trees. Discuss how the final prediction is made using aggregation techniques (majority voting or averaging) and the impact of increasing the number of trees on model performance.	**8+5**	5	L3

PART- C(1x 15=15Marks)

(Q.No.16 is compulsory)

<table>
<tr><th>Q.No.</th><th>Questions</th><th>Marks</th><th>CO</th><th>BL</th></tr>
<tr><td>16.</td><td>You are given a dataset to classify whether a person will play tennis based on three attributes: Outlook, Temperature, and Humidity. Use the ID3 algorithm to construct a decision tree and evaluate its effectiveness.
Training Data:
<table>
<tr><th>Outlook</th><th>Temperature</th><th>Humidity</th><th>Play Tennis</th></tr>
<tr><td>Sunny</td><td>Hot</td><td>High</td><td>No</td></tr>
<tr><td>Sunny</td><td>Hot</td><td>Normal</td><td>Yes</td></tr>
<tr><td>Overcast</td><td>Hot</td><td>High</td><td>Yes</td></tr>
<tr><td>Rain</td><td>Mild</td><td>High</td><td>Yes</td></tr>
<tr><td>Rain</td><td>Cool</td><td>Normal</td><td>Yes</td></tr>
<tr><td>Sunny</td><td>Mild</td><td>High</td><td>No</td></tr>
</table>
1. Evaluate the effectiveness of the decision tree created using the ID3 algorithm. Discuss its ability to classify new instances correctly. Is the model likely to overfit the training data? Justify your reasoning based on the structure of the tree and the dataset.

2. Create a simplified version of the decision tree using pruning. Justify your pruning decisions and describe how this might impact the model's generalization to unseen data.</td><td>7 + 8</td><td>5</td><td>L4 and L5</td></tr>
</table>