

[This question paper contains 4 printed pages.]

Your Roll No.....

Sr. No. of Question Paper : 13247

L

Unique Paper Code : 32347607

Name of the Paper : Machine Learning

Name of the Course : B.Sc. (H) Computer Science (CBCS)

Semester : VI

Duration : 3 Hours

Maximum Marks : 90

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. **Section A** is compulsory.
3. Attempt **any 4** of questions from **Section B**.
4. Parts of a question must be answered together.
5. Use of Scientific Calculator is allowed.

SECTION A
(Compulsory)

- 1 (a) A model performs well on training data but poorly on unseen data. Identify the issue and suggest any two corrective techniques. (3)
- (b) What is standardization in machine learning? Why is it important for algorithms like k-Nearest Neighbour and Support Vector Machine? (3)
- (c) What is bias-variance trade-off? Why is it impossible to simultaneously minimize both bias and variance? (3)
- (d) Normalize the given data using mean normalization method (3)
X: [8,4,11,23,27,29]
- (e) A single layer perceptron has two inputs $x_1=0.5$ and $x_2=0.8$ with weights $w_1=-0.2$ and $w_2=0.5$. If the bias $b=-0.1$, calculate the weighted sum(z) and find output using a step activation function (Output 1 if $z \geq 0$, else 0) (3)
- (f) Find the derivative of activation function used in logistic regression. (3)
- (g) A logistic regression model trained on a rare disease dataset reports 97% accuracy, yet a medical expert rejects it. Analyse how class imbalance makes accuracy misleading, identify a more appropriate evaluation metric for this setting. (3)
- (h) How does a cost function differ from a loss function? Discuss any two cost functions in machine learning. (3)
- (i) A Linear regression is applied to a binary classification problem. Analyse why this approach is unsuitable, discuss its impact on predicted outputs and decision boundaries, and suggest a more appropriate alternative model. (3)

P.T.O.

- (j) For each of the following three real-world applications (A1, A2, A3) of machine learning, identify the primary type of machine learning involved (Supervised, Unsupervised, or Reinforcement) and provide a one-sentence reason for your choice. (3)
- A1: An AI model is trained on labelled images (e.g., X-rays, MRIs) where doctors have already identified cancerous vs. healthy tissues. The AI model learns to detect cancerous tissues in new images.
 - A2: A self-driving car learns to navigate a highway by receiving a reward for staying in its lane and a penalty for swerving, without being told every correct steering angle.
 - A3: An online news website groups articles into topics (e.g., sports, politics, tech) based only on the words they contain, without any pre-assigned topic labels.
- (k) Distinguish between linear and non-linear decision boundaries. Verify the truth table of AND operation using a single layer perceptron. (5)

SECTION B

- 2 (a) Find the least square regression line for the given dataset using the Ordinary Least Square method. Show the computation at each step. Predict the cost of a house if number of bedrooms are 12. Note that the Property ID column is an identification label (metadata) and should not be used as a numerical feature in your regression calculations. (5)

Property ID	Bedrooms	Cost (lakhs)
P1	4	7
P2	3	5
P3	5	14
P4	9	10
P5	7	18

- (b) Consider the dataset given below having two inputs variables x_1 , x_2 and one output variable update the coefficients θ_0 , θ_1 , θ_2 using gradient descent for the logistic regression. Assume the learning rate 0.1 and initial value of coefficients as $\theta_0=0$, $\theta_1=0.2$, $\theta_2=-1.0$. Updates the weights after first iteration. (5)

x_1	1	3	5	6
x_2	2	4	3	2
y	1	1	0	0

- 3 (a) In a binary classification problem, a classifier correctly predicts 90 instances as positive, incorrectly predicts 20 instances as positive, correctly predicts 90 instances as negative, and incorrectly predicts 10 instances as negative. Construct a confusion matrix on above results. Also calculate accuracy, precision, and F1-Score. (5)
- (b) Compare the behaviour of Ridge and Lasso in different scenarios, and analyse how model coefficients and performance are affected as the regularization parameter α approaches 0 and as it becomes very large. Also examine how each method handles feature selection. (5)

- 4 (a) A health monitoring system uses K-Nearest Neighbors (KNN) to classify individuals as "Fat" or "Non-Fat" based on Weight (kg) and Height (cm). The training dataset is given below. (5)

Patient Id	Weight(kg)	Height(cm)	Category (y)
P1	50	160	Non-Fat
P2	55	170	Non-Fat
P3	60	165	Non-Fat
P4	80	170	Fat
P5	85	175	Fat

A new individual arrives with weight = 70 and Height = 168, Using the Euclidean Distance metric and (K=3), determine the category of the individual.

- (b) Distinguish between hard margin and soft margin in SVM. Derive the margin width and show that it equals $\frac{2}{\|w\|}$. Also analyse the role of C and ξ_i in controlling margin violations. (5)
- 5 (a) What is the assumption of a Naïve Bayes classifier. Using Naïve Bayes classification for the following data, predict whether a Middle age Employee having High income will buy a House or not. (5)

Id	Age	Income	Employee	Buy House
1	Young	High	No	No
2	Young	High	No	No
3	Middle	High	No	No
4	Old	Medium	No	Yes
5	Old	Low	Yes	Yes
6	Old	Low	Yes	No
7	Middle	Low	Yes	Yes
8	Young	Medium	No	No

- (b) What is cross validation. What are its different types? Discuss k-fold validation with the help of a suitable example. (5)
- 6 (a) Perform K-Means clustering on following customer behaviour data using K=2 and Euclidean distance. The initial centroids are $C_1 = (1,6)$ and $C_2 = (6,1)$. Compute the updated centroids after two iterations. (5)

Customer	Item Purchase	Monthly Visits
C1	1	6
C2	2	5
C3	5	2
C4	6	1
C5	8	4

- (b) Given a dataset of five points: $P1(1, 2)$, $P2(2, 1)$, $P3(3, 4)$, $P4(5, 4)$, and $P5(2, 5)$, using the Single Linkage (nearest neighbor) clustering method perform the followings (5)

- Compute the initial Euclidean distance matrix.
- Show the stepwise merging of clusters and the updated distance calculations.
- Draw the final dendrogram, clearly labelling the merge distances on the vertical axis.

- 7 (a) Using ID3 Decision Tree algorithm which node would you select root node based on information gain: (5)

Weather	Temperature	Play Game
Sunny	Hot	No
Sunny	Hot	No
Overcast	Hot	Yes
Rainy	Mild	Yes
Rainy	Cool	Yes
Sunny	Mild	No

- (b) What happens to the training process if the learning rate is extremely high? Explain how it affects loss convergence and model stability with a suitable diagram. (5)