

SL No of QP : 13356

Unique Paper Code: 2344000027

Name of the paper: Machine Learning

Name of the Course: Generic Elective (Computer Science) (NEP-UGCF)

Semester: VII

Duration: 3 Hours

Maximum Marks: 90

Instructions for candidates

1. Write your Roll no. on the top immediately on receipt of question paper.
2. **SECTION A** is compulsory
3. Attempt **any 4** 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) Identify the learning approach (Supervised (S), Unsupervised (U), or Reinforcement Learning (RL)) for each of the following task: 3
 - (i) A robot arm tunes its moves to maximize the number of parts assembled per hour.
 - (ii) A camera app learns to blur backgrounds after seeing portraits with ground-truth masks.
 - (iii) An e-commerce store wants to segment its users on the basis of their browsing pattern.
- (b) What are the main advantages of using PCA in machine learning? 3
- (c) A classification model performs with a high accuracy on training data but generalizes poorly to new instances. Identify the problem. Enumerate two possible solutions to this problem. 3
- (d) Given $p = 0.8$, compute the odds and log-odds. 3
- (e) Two regression models are used to predict house prices. Model A has an R^2 of 0.55, and Model B has an R^2 of 0.89. Which model provides a better fit to the data? 3
- (f) Give one application area of a Simple Perceptron versus a Multi-Layer Perceptron (MLP). 3
- (g) What is K-means clustering? What is the main drawback of this algorithm? 3
- (h) Draw the architecture of a 3-layer Feedforward Neural Network designed for a binary classification problem. Label the input layer (with 4 neurons), one hidden layer (with 5 neurons), and the output layer. Include weights and biases in your diagram. 3
- (i) Explain any one feature selection strategy. 3
- (j) Explain any one validation approach used in machine learning. 3

SECTION-B

2. (a) Consider the following training data for five persons. Create a decision tree model for binary classification of a person as sick or not sick. Show all the steps. 10

Person No	A1	A2	A3	Class
-----------	----	----	----	-------

1	Yes	Yes	Yes	Not Sick
2	Yes	No	Yes	Sick
3	No	No	Yes	Sick
4	No	Yes	Yes	Not Sick
5	No	Yes	No	Sick

- (b) Apply the K-nearest neighbor algorithm on the following dataset with $K=3$ to predict the class label for a new data point X with Feature A=4 and Feature B = 5 on the given dataset. (Use the Euclidean distance metric)

Feature A	Feature B	Class
1	1	Alpha
2	3	Alpha
4	8	Beta
6	6	Beta
5	4	Beta

3. (a) Consider an example where we are interested to find the effect of studying hours per day over grades in examination and predict the grades of a student for given study hours.

Study Hours per day	Grades
6	5
5	3
4	2
7	6
9	11

Using the least square regression method, find the line of best fit. Predict the grades if study hours are 8 per day and compute R^2 statistics for the model.

- (b) Differentiate between the following:
- Hard margin and Soft margin
 - Overfitting and Under Fitting

4. (a) Consider a dataset used to determine if a student will pass an exam based on their study habits and attendance:

Study_Hours	Attendance	Pass_Exam
High	Low	Yes
High	Low	No
Medium	High	Yes
Medium	Low	No
High	High	Yes
High	High	No

Using this dataset, apply the Naive Bayes classifier to predict the class label "Pass_Exam" for a new student with the following attribute values: Study_Hours = Medium and Attendance = Low.

- (b) Cluster the following set of data objects in two clusters by applying one iteration of the k-means algorithm. Treat objects 2 and 5 as initial centroids. Use Euclidean distance as the distance metric. 5

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

5. (a) A system implemented for the classification of 1000 incoming emails. Out of which 900 emails are non-spam (legitimate), and 100 are confirmed spam emails. The system classified 135 emails as spam out of which only 85 are actually spam. Construct a confusion matrix for the scenario described above and compute the following: 10
- True Positive Rate (TPR)
 - False Positive Rate (FPR)
 - Specificity
 - Accuracy
- (b) Consider a dataset consisting of 100 instances. In the context of 10-fold cross-validation, how many training iterations will be conducted? and what will be the size of the training and testing sets for each iteration? 5
6. (a) Calculate the output of a single layer neural network with two input neurons and one output neuron. The input feature vector is $(x_1, x_2) = (0.6, 0.8)$ the weight vectors is $(w_1, w_2) = (0.4, -0.5)$ bias $b_1=0.2$. Use sigmoid function as the activation function. 10
- (b) Suppose you have three class labels in your dataset i.e. cat, dog and horse. Describe the process of constructing a multi-class prediction model using One-vs-All classification. 5
7. (a) Perform the single link hierarchical clustering for the distance matrix given in the below table. Show your result by drawing a dendrogram. 10

	P1	P2	P3	P4	P5
P1	0	4	7	9	1
P2	4	0	3	5	3
P3	7	3	0	2	6
P4	9	5	2	0	8
P5	1	3	6	8	0

- (b) Explain the impact of the following on machine learning model 5
- High Variance, Low Bias
 - Low Variance, High Bias