Sr. No. of Question Paper:9758

Unique Paper Code: 6333490007

Name of the Paper: Machine Learning Fundamentals

Name of the Course: B.Voc (Web Designing) (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 this question paper.
2. **Section A (Question No. 1)** is compulsory.
3. Attempt any ***four*** questions from Section B (Questions 2 to 7).
4. Parts of a question must be answered together.
5. Simple Calculator is allowed.

Section A

Q1. (a) What is the assumption in the Naive Bayes Classifier ? (2)

(b) What is the difference between Linear Regression and Logistic Regression ? (3)

(c) What do you understand by Standardization ? Explain briefly. (4)

(d) What is AUC–ROC Curve? (4)

(e) What is 'training Set' and 'test Set' in a Machine Learning Model ? How much data will you allocate for your Training and Test sets ? (4)

(f) Define FPR and TPR. (4)

(g) Briefly explain Logistic Regression with a diagram. (4)

(h) What is Regularization ? Explain the difference between L1 Regularization (Lasso) and L2 Regularization (Ridge). (5)

Section B

Q2. (a) What is Bias and Variance in a Machine Learning Model ? (4)

(b) What are the different strategies which can be followed for feature subset selection ? (5)

(c) What Are the Different Types of Machine Learning algorithms ? Explain each with examples. (6)

Q3. (a) Consider a dataset with actual values y and predicted values γ from a regression model as given below. Assess the performance of the model using : (4)

a) Mean absolute error
b) Mean squared error
c) Root mean squared error
d) R^2

S.No	Actual value (y)	Predicted value(γ)
1	3.5	2.8
2	4.0	3.6
3	5	5.4
4	7	6.8
5	9	8.7

(b) Given a dataset with three input features (x1,x2,x3) and target attribute y which can take either a value 'yes' or 'no'. A logistic regression model is trained on the given training set to estimate the parameters as $\beta_0 = 0.1$, $\beta_1 = 0.35$, $\beta_2 = 0.11$ and $\beta_3 = 0.2$. compute the probability of y = "yes" given the instance as (2,4,7). (5)

(c) Describe the k-fold cross-validation method a suitable choice. What are the ideal choices for the value of k ? can we say that Leave-one out cross-validation is a special case of k-fold cross-validation ? justify your answer. (6)

Q4 (a) The following dataset contains information about eight students, including their CGPA, assessment score, project points, and result status. Use the k-Nearest Neighbor (k-NN) algorithm with k = 3 to classify the result (pass/fail) for a new test instance having the following attributes: CGPA = 8.0, Assessment = 62, and Project Points = 8. (5)

Roll No	CGPA	Assessment	Project points	Result
1	8	85	8	Pass
2	7	75	6	Pass
3	6	50	5	Fail
4	7.5	65	7	Pass
5	6.5	55	5	Fail
6	8.2	90	9	Pass
7	5.8	45	4	Fail
8	7.2	70	6	pass

(b) What is the use of Normalization in Machine learning ? Explain its technique with an example. (5)

(c) A machine learning based imperfect email spam classifier flags emails as spam or not. Out of 500 emails, 120 emails are actually spam. The model detects 150 emails as spam, out of which only 100 are truly spam. Also, the model is unable to detect 20 spam emails. Draw the confusion matrix and calculate the following performance metrics for the classifier: (5)

(a) Accuracy

(b) Precision

(c) Recall

Q5. (a) Answer the following questions :- (3+4)

1. List down the steps to perform K-Means clustering.

2. Given the following data points, cluster them into two groups (K=2) using the K-means clustering algorithm. Data Points are given as R1(1.0, 1.0) , R2(1.5, 2.0) , R3(3.0, 4.0) , R4(5.0, 7.0) , R5(3.5, 5.0) , R6(4.5, 5.0) , R7(3.5, 4.5). Initial Centroids are given as C1 = (1.0, 1.0) , C2 = (5.0, 7.0).

(b) A dataset contains information about weather conditions and whether people play tennis. Calculate the entropy of the dataset and determine which attribute (Outlook or Wind) should be selected as the root node. (8)

Day	Outlook	Wind	Play Tennis
1	Sunny	Weak	No
2	Sunny	Strong	No
3	Overcast	Weak	Yes
4	Rain	Weak	Yes
5	Rain	Strong	No
6	Overcast	Strong	Yes

Q6. (a) Consider a fictional dataset that describes the weather conditions for playing a game of golf. Given the weather conditions, each tuple classifies the conditions as fit("Yes") or unfit("No") for playing golf. (5)

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Find the label for the class "Play Golf" for the input X=(Outlook = Sunny , Temperature = Hot, Humidity = Normal , Windy = False) using Naïve Bayes Classifier.

(b) Consider the following 2 classes of points in 2-dim: (5)

class +1: (2, 2), (4, 4), (4, 2) , class -1 : (2, 4), (4, 6), (2, 6)

(a) Plot these points on 2-dimension axes.
(b) Try to draw any line that separates the two classes.
(c) Draw the widest possible margin (i.e., the maximal margin hyperplane) between the classes.
(d) Identify the support vectors.

(c) Find the linear regression equation for the given data: (5)

X	Y
3	8
9	6
5	4
3	2

Q7. (a) List the steps required for performing Principal Component Analysis (PCA). (4)

(b) Find the clusters using a single linkage technique for the given six data points. Draw the Dendogram also. Use Min(dist) to find distance between two points. (5)

18, 22 , 25 , 42 , 27 , 43

(c) In a Covid test of 1000 patients, there were 45 positive tests, of which 30 patients had covid and 15 were falsely tested positive. Of the 955 negative tests there were 5 that were incorrect, these patients had covid but were tested negatively. Draw the confusion matrix and calculate the accuracy, precision, recall, sensitivity from the matrix. (6)