

Sr. No. of Question Paper:10391

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) Name any two distance metrics in Machine Learning ? (2)
- (b) What are outliers and how to handle them ? (3)
- (c) Explain how sigmoid function work in Logistic Regression and why it is not a Regression Model even though it name has it ? (4)
- (d) What is Bias-Variance tradeoff ? (4)
- (e) What is Information Gain and Entropy in Decision Tree? (4)
- (f) What are the assumptions of Naive Bayes? (4)
- (g) What do you understand by Standardization ? Explain briefly. (4)
- (h) What is Overfitting ? What are the measures to be taken to remove it. (5)

Section B

Q2.

- (a) What are the different Machine Learning algorithms. Explain each with an example. (6)
- (b) What is the difference between Cross-validation and Hold-out method ? (4)

(c) Explain how changing the tuning parameter (λ) in ridge regression affects the model. (5)

- i. What happens if $\lambda \rightarrow 0$?
- ii. What happens if $\lambda \rightarrow \infty$?

Q3.

(a) What do you understand by Principal Component Analysis (PCA) ? Enumerate the steps involved in performing PCA ? (5)

(b) Consider the following customer purchase dataset for an e-commerce platform : (10)

Customer ID	Age Group	Income Level	Purchase
1	Young	High	Yes
2	Middle	Low	No
3	Senior	High	Yes
4	Young	Low	No
5	Middle	High	No
6	Senior	Low	Yes
7	Young	High	Yes
8	Middle	Low	No

Using the Entropy criterion construct a decision tree.

Q4.

(a) Cluster the following dataset in two clusters by applying one iteration of k-means algorithm. Treat objects 2 and 5 as initial cluster centres. Use Euclidean distance as the distance metric. Determine updated cluster centre coordinates. (5)

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

- (b) What is the difference between feature scaling and feature selection ? Why is feature selection important ? what are the various feature selection strategies. Explain them briefly with examples. (10)

Q5.

- (a) Consider the following 2 classes of points in 2-dim: (8)

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

- Plot these points on 2-dimension axes.
 - Try to draw any line that separates the two classes.
 - Draw the widest possible margin (i.e., the maximal margin hyperplane) between the classes.
 - Identify the support vectors.
- (b) Consider the following 10 training instances. (7)

No	Color	Type	Mileage	Tested
1	Blue	Sports	Avg	Yes
2	Blue	SUV	Avg	No
3	Blue	Sports	High	Yes
4	Pink	Sports	High	No
5	Pink	Sports	Avg	Yes
6	Pink	SUV	Avg	No
7	Blue	SUV	High	Yes
8	Red	SUV	High	No
9	Red	SUV	Avg	Yes
10	Red	Sports	avg	yes

Compute the following probabilities

- $P(\text{Blue}/\text{yes})$
- $P(\text{Sports}/\text{NO})$
- $P(\text{Red} / \text{Yes})$
- $P(\text{Pink} / \text{yes})$

Q6.

- (a) Calculate intercept and slope the following pairs of (x,y) training values : (2,6) , (5,12) , (8,15) , (12,23) , (17,29). Predict the value of y for test data x=10. (8)
- (b) For the given set of points, identify clusters using complete linkage in agglomerative clustering. Use Euclidean distance to calculate the distance between two points. (7)

Points	x-coordinate	y-coordinate
P1	1	1

P2	1.5	1.5
P3	5	5
P4	3	4

Q7.

- (a) Apply the k-nearest neighbor algorithms on the following dataset : (8)

a	b	Class
7	7	False
7	4	False
3	4	True
1	4	True

Predict the class for X (a=3 , b=7). Assume k=3 (use Euclidean distance to compute the distance).

- (b) Assume a total of 1000 patients are tested for influenza; 900 are found to be healthy, while 100 are found to be sick. A test resulted in 60 being positive and 40 being negative for the sick persons. The same test positive for 120 and negative for 780 in healthy adults. Construct a confusion matrix for the data and determine precision and recall. (7)