M/E/TX

Reg. No:

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of the UGC Act, 1956)

# Final Assessment Test – May 2024

Course: **MDI4001 - Machine Learning for Data Science**

Class NBR(s): **2477/2478** Slot: **C1+TC1**

Time: **Three Hours** Max. Marks: **100**

- **KEEPING MOBILE PHONE/ELECTRONIC DEVICES EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

**Answer any TEN Questions**

**(10 X 10 = 100 Marks)**

1. Consider a dataset with 100 samples and 10 features. The dataset is split into training, validation, and testing sets with a 70:15:15 ratio. Further, three different models, including Linear Regression (Model A), Decision Tree (Max depth 10)(Model B), and Decision Tree (Max depth 2) (Model C), are trained on the training data. After training, model evaluation results in the mean squared error (MSE) values of 200, 20, 15 from the training set and 167, 17, and 100 from the validation set, respectively, for linear regression, Decision Tree (Max depth 10), and Decision Tree (Max depth 2). Based on this information, answer the following questions:

   a. Which model is likely underfitting and overfitting? Justify with a logical explanation. **[5]**

   b. If you had to choose one model to deploy, which one would it be and why? **[5]**

2. Consider the following dataset:

| Purchasing behaviour | Gender | Age | Time spent in online shop (hrs/week) |
|---|---|---|---|
| Buy now | female | 22 | 40 |
| Buy now | male | 18 | 65 |
| Buy later | female | 27 | 28 |
| Buy later | male | 27 | 15 |

Train the dataset using logistic regression and predict the purchasing behaviour where gender is female, age is 30 and time spent in online shop is 30 hrs/week.

3. a) In recent years, polynomial regression has gained significant attention in various fields of study and practical applications. Discuss with the help of an example the need for polynomial regression in predictive modeling and data analysis, elucidating its advantages over logistic regression techniques. **[2]**

   b) Examine and elucidate the Batch Gradient Descent (BGD), Stochastic Gradient Descent (SGD), and Mini-batch Gradient Descent (MBGD) optimization algorithms concerning their implementation within neural network training paradigms. Analyze the distinctive merits, drawbacks, and contextual applicability of each algorithm. **[8]**

4. a) What are the two ways to handle multiclass classification problem using SVM? Discuss with example. [6]

b) What is the difference between hard and soft margin of SVM? [4]

5. Consider the following dataset and draw the hyperplane using SVM.

| Class | X | Y |
|---|---|---|
| A | 1 | 4 |
| A | 2 | 5 |
| A | 3 | 5 |
| A | 3 | 4 |
| B | 6 | 1 |
| B | 4 | 0 |
| B | 5 | 2 |
| B | 5 | 1 |

6. Consider the following dataset with binary classification. Train a backpropagation neural network with suitable loss function and sigmoid activation function classifies the samples into two categories based on these features. Consider the learning rate as 0.1. The neural network will have an input layer with two neurons (corresponding to the two features), one hidden layer with two neurons, and an output layer with one neuron. The weight and bias initialization are as follows: Input to Hidden Layer Weights: $W^1 = \begin{bmatrix} 0.2 & -0.5 \\ 0.8 & 0.1 \end{bmatrix}$, Hidden Layer Biases: $b^1 = \begin{bmatrix} -0.3 \\ 0.5 \end{bmatrix}$ ,Hidden to Output Layer Weights: $W^2 = \begin{bmatrix} -0.7 \\ 0.4 \end{bmatrix}$, Output Layer Biases: $b^1 = 0.2$.

| Feature 1 | Feature 2 | Label |
|---|---|---|
| 0.2 | 0.8 | 0 |
| 0.5 | 0.3 | 1 |

a) Calculate the loss after first iteration. [3]

b) Using backpropagation algorithm fine tune the weight and bias value and show that loss has reduced after the backpropagation. [7]

7. Consider the following dataset:

| Age | Weight | Smoker | Risk of heart attack |
|---|---|---|---|
| 18 | Normal | Yes | low |
| 15 | Obese | Yes | Low |
| 28 | Obese | Yes | High |
| 25 | Normal | No | Low |
| 59 | Normal | No | Low |
| 45 | Obese | Yes | High |
| 49 | Normal | Yes | High |
| 67 | Obese | No | High |

Create a decision tree using Gini impurity to predict the risk of heart disease and validate it using any one observation mentioned in the dataset.

8. a) Discuss the differences in the working principle of bagging and boosting with respect to the dataset mentioned in Qs. 7. **[6]**

b) Discuss the working principle of gradient boosting algorithm. **[4]**

9. Discuss the need and working principle of PCA in machine learning by considering an example.

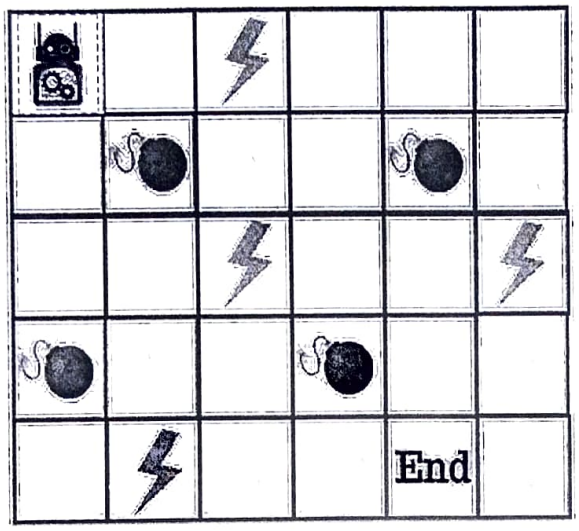
10. Consider the following dataset:

| Point | X | Y |
|---|---|---|
| P1 | 2 | 10 |
| P2 | 2 | 5 |
| P3 | 8 | 4 |
| P4 | 5 | 8 |
| P5 | 7 | 5 |
| P6 | 6 | 4 |
| P7 | 1 | 2 |
| P8 | 4 | 9 |

Using DBSCAN algorithm, find the core points, noise points and clusters by considering Eps = 2 and minimum number of points = 3.

11. Discuss the difference between different hierarchical clustering algorithms with an example.

12. Consider the following diagram and apply the working principle of reinforcement learning with respect to this scenario.




The scoring/reward system is as below:

A. The robot loses 1 point at each step. This is done so that the robot takes the shortest path and reaches the goal as fast as possible.

B. If the robot steps on a mine, the point loss is 100 and the game ends.

C. If the robot gets power ϟ, it gains 1 point

D. If the robot reaches the end goal, the robot gets 100 points.

⇔⇔⇔ M/E/TX⇔⇔⇔