Reg. No															

B.Tech DEGREE EXAMINATION, DECEMBER 2023

Fifth Semester

18CSE363J - MACHINE LEARNING

(For the candidates admitted during the academic year 2020 - 2021 & 2021 - 2022)

Note:

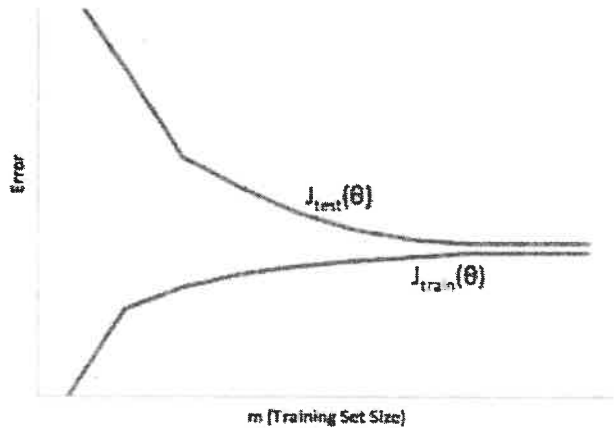
i. **Part - A** should be answered in OMR sheet within first 40 minutes and OMR sheet should be handed over to hall invigilator at the end of 40th minute.

ii. **Part - B** and **Part - C** should be answered in answer booklet.

Time: 3 Hours **Max. Marks: 100**

PART - A (20 × 1 = 20 Marks) — Answer **all** Questions

		Marks	BL	CO
1.	What is the goal of feature selection in machine learning? (A) To transform categorical variables into numerical representations (B) To balance the class distribution in imbalanced datasets. (C) To reduce the dimensionality of the dataset. (D) To normalize the data for better model performance.	1	1	1
2.	Which of the following is a categorical feature? (A) Amount of rainfall in a day. (B) Price of petroleum. (C) Mother tongue of a person. (D) Height of a person.	1	2	1
3.	A feature F1 can take certain values: A, B, C, D, E, F, and represents the grade of students from a college. Which of the following statement is true in the following case? (A) Feature F1 is an example of a nominal variable (B) Feature F1 is an example of an ordinal variable (C) It doesn't belong to any of the above categories (D) Both of these	1	2	1
4.	You train a learning algorithm, and find that it has unacceptably high error on the test set. You plot the learning curve, and obtain the figure below. Is the algorithm suffering from high bias, high variance, or neither?	1	1	1



(A) High bias
(B) High variance
(C) Neither
(D) Low bias

5. Which of the following tasks is NOT a suitable machine learning task(s)? 1 1 1

(A) Finding the shortest path between a pair of nodes in a graph
(B) Predicting if a stock price will rise or fall
(C) Predicting the price of petroleum
(D) Grouping mails as spams or non-spams

6. What should be the loss function used to train the model? 1 1 2

(A) Multi-Class Cross-Entropy Loss
(B) Mean Squared Error
(C) Binary Cross-Entropy Loss
(D) Hinge

7. Naïve Bayes is a popular classification algorithm in machine learning. Which of the following statements is/are true about Naïve Bayes? 1 2 2

(A) Naive Bayes assumes that all features are independent of each other, given the class
(B) It is does not suit for text classification tasks, like spam detection
(C) Naive Bayes can handle missing values in the dataset without any special treatment
(D) It is a complex algorithm that requires a large amount of training data

8. A 4-input neuron has weights 1, 2, 3 and 4. The transfer function is linear with the constant of proportionality being equal to 2. The inputs are 4, 10, 5 and 20 respectively. What will be the output? 1 2 2

(A) 76
(B) 238
(C) 119
(D) 123

9. Which of the following techniques can be used to handle overfitting in decision trees? 1 1 3

(A) Pruning
(B) Increasing the tree depth
(C) Increasing the minimum number of samples required to split a node
(D) Adding more features to the dataset

10. Which of the following is a measure used for selecting the best split in decision trees? 1 1 3

(A) Gini Index
(B) Support Vector Machine
(C) K-Means Clustering
(D) Naive Bayes

11. What is the purpose of the decision tree's root node in machine learning? 1 1 3

(A) It represents the class labels of the training data
(B) It serves as the starting point for tree traversal during prediction
(C) It contains the feature values of the training data
(D) It determines the stopping criterion for tree construction

12. Which of the following statements about linear regression is true? 1 1 3

(A) Linear regression is a supervised learning algorithm used for both regression and classification tasks
(B) Linear regression assumes a linear relationship between the independent and dependent variables
(C) Linear regression is not affected by outliers in the data
(D) Linear regression can handle missing values in the dataset

13. K-Nearest Neighbor is a _____ algorithm 1 1 3

(A) Non-parametric, eager
(B) Parametric, eager
(C) Non-parametric, lazy
(D) Parametric, lazy

14. What is true about K-Mean Clustering? 1 1 4

1. K-means is extremely sensitive to cluster center initializations
2. Bad initialization can lead to Poor convergence speed
3. Bad initialization can lead to bad overall clustering

(A) 1 and 2
(B) 1 and 3
(C) 1,2 and 3
(D) 2 and 3

15. Assume, you want to cluster 7 observations, into 3 clusters using K-Means clustering algorithm. After first iteration the clusters: C1, C2, C3 has the following observations: 1 2 4

C1: {(1,1), (4,4), (7,7)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
What will be the cluster centroids if you want to proceed for second iteration?

(A) C1: (4,4), C2: (2,2), C3: (7,7) (B) C1: (2,2), C2: (0,0), C3: (5,5)
(C) C1: (6,6), C2: (4,4), C3: (9,9) (D) None of these

16. Imagine you are dealing with 20 class classification problem. What is the maximum number of discriminant vectors that can be produced by LDA? 1 1 4

(A) 20 (B) 19
(C) 21 (D) 10

17. Bagging is done to _____ 1 1 5

(A) increase bias (B) decrease bias
(C) increase variance (D) decrease variance

18. Which of the following is a categorical outcome? 1 1 5

(A) RMSE (B) RSquared
(C) Accuracy (D) SVM

19. What technique helps in assessing a machine learning model's performance by dividing the dataset into multiple subsets for training and testing? 1 1 6

(A) Hyper-parameter tuning (B) Feature scaling .
(C) Cross-validation (D) Data augmentation

20. Which library is suitable for implementing neural networks and deep learning in Python? 1 1 6

(A) scikit-learn (B) Keras
(C) Pandas (D) Matplotlib

PART - B (5 × 4 = 20 Marks) **Marks BL CO**

Answer **any 5** Questions

21. Explain the concepts of over fitting and under fitting in machine learning. Outline some strategies to mitigate over fitting and under fitting in machine learning models. 4 2 1

22. How neurons are connected between different layers in a feed forward neural network? 4 2 2

23. Describe a scenario where Multiple Linear Regression would be a suitable choice for modeling a real-world problem. 4 3 3

24. Describe "elbow method," and how it is used to select the optimal number of clusters? 4 4 4

25. What are Latent Variable Models (LVMs), and how do they address the challenges of modeling complex data? 4 3 4

26. Point out the two main categories of ensemble methods, and how do they differ from each other. 4 4 5

27. Explain cross-validation and how it can be implemented in Python. 4 3 6

PART - C (5 × 12 = 60 Marks) **Marks BL CO**

Answer **all** Questions

28. (a) 1. Elucidate the Candidate Elimination Algorithm, its significance and also explain how does it work? [9 Marks] 12 4 1
2. What is the purpose of S and G sets in version space [3 Marks]

(OR)

(b) Explain various learning techniques involved in supervised learning?

29.	(a) Design single layer perception with two iteration. Consider the perceptron having with the initial weights w1=0.5, w2 = 0, learning rate α=0.2 and bias θ =0.4 for AND Boolean function. The activation function is the Step function f(x) which gives the output either 0 or 1. If value of f(x) is greater than or equal to 0, it outputs 1 or else it outputs 0.	12	6	2
	(OR)			
	(b) 1. What is the main assumption of the Naive Bayes Classifier, and why is it called "naïve"? [3 Marks] 2. How does the Naïve Bayes Classifier calculate the probability of a class given input features? [9 Marks]			
30.	(a) Compare and contrast the linear and logistics regression with example.	12	2	3
	(OR)			
	(b) Explain weighted K-nearest Neighbor algorithm.			
31.	(a) Explain Principal Component Analysis (PCA) in detail. Discuss its applications, advantages, and limitations. Provide an example illustrating its use.	12	4	4
	(OR)			
	(b) Explain the following 1. Eigen values and Eigenvectors, and how are they used in Spectral Clustering? [8 Marks] 2. The role of the number of clusters (K) in Spectral Clustering, and how is it determined? [7 Marks]			
32.	(a) Compare the binomial test, approximate normal test, and paired t-test for assessing a classification algorithm's performance.	12	5	5
	(OR)			
	(b) 1. Explain the process of K-Fold Cross-Validation with K=5? [6 Marks] 2. In 5x2 Cross-Validation, how are the data splits typically arranged for training and testing? [6 Marks]			

* * * * *