B.Tech DEGREE EXAMINATION, DECEMBER 2023

Fifth to Seventh Semester

18CSE392T - MACHINE LEARNING - I

(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. Identify the type of learning in which both input and output attributes available. `1  1  1`
   (A) Semi Supervised Learning   (B) Supervised Learning
   (C) Reinforcement Learning   (D) Unsupervised Learning

2. Which of the following statement is TRUE? `1  2  1`
   (A) Outliers should be identified and removed always from a dataset.
   (B) Outliers can never be present in the test set.
   (C) Outliers is a data point that is significantly close to other data points.
   (D) The nature of our business problem determines how outliers are used.

3. What kind of algorithm can be used for "Future currency exchange rates"? `1  2  1`
   (A) Generating Pattern   (B) Recognition Pattern
   (C) Prediction   (D) Anomalies

4. _____ is not an example of classification. `1  2  1`
   (A) Speech Recognition   (B) Character Recognition
   (C) Face Recognition   (D) Game Playing

5. Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target? `1  2  2`
   (A) AUC-ROC   (B) Accuracy
   (C) Logloss   (D) Mean-Squared-Error

6. What's the the hypothesis of logistic regression? `1  2  2`
   (A) to limit the cost function between 0 and 1
   (B) to limit the cost function between -1 and 1
   (C) to limit the cost function between -infinity and +infinity
   (D) to limit the cost function between 0 and +infinity

7. Naïve Bayes algorithm is based on _____ and used for solving classification problems. `1  1  2`
   (A) Candidate elimination algorithm (B) Bayes Theorem
   (C) EM algorithm   (D) RM algorithm

8. The confusion matrix visualizes the ____ of a classifier by comparing the actual and predicted classes. `1  1  2`
   (A) Stability   (B) Connectivity
   (C) Accuracy   (D) Compactability

9. *What's the penalty term for the Ridge regression?* `1  1  3`
   (A) the square of the magnitude of the coefficients
   (B) the square root of the magnitude of the coefficients
   (C) the absolute sum of the coefficients
   (D) the sum of the coefficients

10. PCA works better if there is _____   1   2   3
    1. A linear structure in the data
    2. If the data lies on a curved surface and not on a flat surface
    3. If variables are scaled in the same unit
    (A) 1 and 2   (B) 2 and 3
    (C) 1 and 3   (D) 1, 2 and 3

11. Suppose you are dealing with a 4-class classification problem, and you want to train   1   3   3
    an SVM model on the data. For that, you are using the One-vs-all method.
    How many times do we need to train our SVM model in such a case?
    (A) 1   (B) 2
    (C) 3   (D) 4

12. Which of the following distance measure do we use in case of categorical variables   1   3   3
    in k-NN?
    1. Hamming Distance
    2. Euclidean Distance
    3. Manhattan Distance
    (A) 1   (B) 2
    (C) 3   (D) 1 & 3

13. What is the minimum no. of variables/ features required to perform clustering?   1   2   4
    (A) 0   (B) 1
    (C) 2   (D) 3

14. Which of the following clustering algorithm follows a top to bottom approach?   1   1   4
    (A) K- Means   (B) Divisible
    (C) Agglomerative   (D) Hierarchical

15. Which of the following clustering algorithms suffers from the problem of   1   4   4
    convergence at local optima?
    1. K- Means clustering algorithm
    2. Agglomerative clustering algorithm
    3. Expectation-Maximization clustering algorithm
    4. Diverse clustering algorithm
    (A) 1 only   (B) 2 and 3
    (C) 2 and 4   (D) 1 and 3

16. In which of the following cases will K-Means clustering fail to give good results?   1   4   4
    1. Data points with outliers
    2. Data points with different densities
    3. Data points with round shapes
    4. Data points with non-convex shapes
    (A) 2 and 3   (B) 1 and 2
    (C) 1, 2 and 4   (D) 2 and 4

17. _____ is the measure of uncertainty of a random variable, it characterizes the   1   1   5
    impurity of an arbitrary collection of examples.
    (A) Information Gain   (B) Gini Index
    (C) Entropy   (D) Gain Ratio

18. _____ is a metric to measure how often a randomly chosen element would be   1   1   5
    incorrectly identified.
    (A) Information Gain   (B) Gini Index
    (C) Entropy   (D) Gain Ratio

19. Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?
    1. Both methods can be used for classification task
    2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
    3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
    4. Both methods can be used for regression task

    (A) 1   (B) 1 & 2
    (C) 1 & 4   (D) 2 & 4

         1   3   5

20. Which of the following is not an Machine Learning strategies in ANNs?
    (A) Unsupervised Learning   (B) Reinforcement Learning
    (C) Supreme Learning   (D) Supervised Learning

         1   1   5

PART - B (5 × 4 = 20 Marks)
Answer any 5 Questions

	Marks	BL	CO
21. Compare and contrast Overfitting and Underfitting.	4	2	1
22. Explain the Curse of dimensionality with example.	4	3	1
23. Why K-Fold Cross Validation is required? Explain the concept with suitable diagram.	4	2	2
24. For the below mentioned data, find y using python code. y = mx+c. Write the python code for this. Note: import the required libraries, split the data, train the model and test the model.	4	3	2

Skill Score	Salary
18	76
19	77
20	78
21	78
22	79
23	80

	Marks	BL	CO
25. How does the KNN algorithm make the predictions on the unseen dataset?	4	3	3
26. How the k-medoids algorithm works.	4	1	4
27. What is the role of the Activation functions in Neural Networks? List popular Activation Functions used in Neural Networks.	4	2	5

PART - C (5 × 12 = 60 Marks)
Answer all Questions

	Marks	BL	CO
28. (a) Summarize the types of machine learning with examples.	12	2	1

(OR)

(b) Explain
   (i) Bias, Variance and its tradeoff.
   (ii) Parametric and non-parametric model

29. (a) **Consider the given Dataset ,Apply Naive Baye's Algorithm and Predict that if a fruit has the following properties then which type of the fruit it is**   12   3   2
Fruit = {Yellow , Sweet ,long}

Frequency Table:

Fruit	Yellow	Sweet	Long	Total
Mango	350	450	0	650
Banana	400	300	350	400
Others	50	100	50	150
Total	800	850	400	1200

(OR)

(b) Explain logistic regression with suitable example

30. (a) Write the working principle of different SVM classifiers with suitable example.   12   4   3

(OR)

(b) Using the PCA algorithm, for the given sample

	Example 1	Example 2	Example 3	Example 4
X1	4	8	13	7
X2	11	4	5	14

i) Compute the number of sample, features and covariance matrix of all ordered pairs.
ii) Compute the corresponding Eigenvalues and Eigenvectors.

31. (a) Cluster the following set of data using k-means algorithm with initial value of objects 2 and 5 with the coordinate values (4,6) and (12,4) as initial seeds.   12   3   4

Objects	X coodinates	Y coordinates
1	2	4
2	4	6
3	6	8
4	10	4
5	12	4

(OR)

(b) Explain the following
i) Spectral clustering
ii) Hierarchical clustering
iii) Agglomerative clustering

32. (a) Explain CART (Classification & Regression tree) algorithm with example.   12   4   5

(OR)

(b) With the given dataset of two features (X1 and X2) and a binary target variable (Y) representing whether a customer buys a product (1 for "buys" and 0 for "doesn't buy"). Solve the binary classification using random forest algorithm.

X1	X2	Y
2	3	1
3	3	0
3	2	1
1	2	0
1	3	1

* * * * *