36

Reg. No																

B.Tech DEGREE EXAMINATION, NOVEMBER 2023

Fifth & Sixth Semester

18CSE479T - STATISTICAL 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.	Choose the correct option for Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka), runner age group (junior, senior, master's), and race finishing times in a marathon (A) one-way (B) two-way (C) three-way (D) t-test	1	2	1
2.	The following transformation is called the z-score transformation (A) (X-max(X))/(Max(X)-Min(X) (B) (X-Mean(X))/(Max(X)-Min(X)) (C) (X-Mean(X))/Standard Deviation of X (D) Mean(X)/Max(X)	1	2	1
3.	A _____ is when the output variable is a real value, such as "dollars" or "weight". (A) Classification problem (B) Regression Problem (C) Reinforcement problem (D) Either Classification or Regression	1	1	1
4.	Consider two small-sized samples and population standard deviation is not given. How will you test the significance of the difference of the mean values between them? (A) F-test (B) Chi-square test (C) Z-test (D) t-test	1	2	1
5.	In the general case, imagine that we have d binary features, and we want to count the number of features with value 1. How many leaf nodes would a decision tree need to represent this function? (A) 2^1 leaf nodes (B) 2^d leaf nodes (C) 2^{d-1} leaf nodes (D) 2^d-1 leaf nodes	1	3	2
6.	The measures developed for selecting the best split are often based on the degree of impurity of the child nodes. Which of the following is NOT an impurity measure? (A) Pruning (B) Entropy (C) Gini (D) Classification error	1	1	2
7.	Consider fitting Lasso regression model on a dataset, which has 100 features (x1,x2,...x100). If we rescale one of these features by multiplying with 10 (say the feature is x1), and then refit lasso regression with the same regularization parameter. Choose the correct option given below (A) It is more likely for x1 to be excluded from the model (B) It is more likely for x1 to be included in the model (C) It is more likely for x2 to be included in the model (D) It is more likely for x100 to be included in the model	1	3	2

8. Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with tuning parameter lambda to reduce its complexity. Choose the option below which describes relationship of bias and variance with lambda. | 1 | 2 | 2

(A) in case of very large lambda; bias is low, variance is low
(B) in case of very large lambda; bias is low, variance is high
(C) in case of very large lambda; bias is high, variance is low
(D) in case of very large lambda; bias is high, variance is high

9. What is the purpose of the Laplace estimator in the context of Naïve Bayes classifiers | 1 | 1 | 3

(A) To ensure that probabilities are not negative
(B) To ensure that probabilities sum to one
(C) To ensure non –zero probabilities
(D) To ensure that cdf integrates to 1

10. Suppose 40% of spam messages contain the word "free". 10% of messages are spam.10% of messages contain the word "free". The probability that a message is spam, given that it contains the word "free" is _____ | 1 | 3 | 3

(A) 0.5
(B) 0.1
(C) 0.2
(D) 0.4

11. How do you deal with Euclidean distance for nominal data in the context of KNN classification? | 1 | 2 | 3

(A) Using dummy coding
(B) Ignoring such data
(C) Deleting those observations
(D) Replacing these observations by 0

12. Binning in the context of naive Bayes classifiers refers to | 1 | 2 | 3

(A) Deleting incorrect values
(B) Averaging missing values
(C) Discretizing numeric values
(D) Extrapolating missing values

13. The minimum time complexity for training an SVM is O(n2). According to this fact, what sizes of datasets are not best suited for SVMs? | 1 | 2 | 4

(A) Large datasets
(B) Small datasets
(C) Medium-sized datasets
(D) Size does not matter

14. Assume that all features are used in the dataset and 100% accuracy is achieved on training data set, but ~70% on the validation set of data, Identify the situation from the choices below | 1 | 2 | 4

(A) Under fitting
(B) Nothing; the model is perfect
(C) Over fitting
(D) Model is imperfect

15. Identify the layer that computes the output volume by computing dot product between all filters and image patch | 1 | 2 | 4

(A) Input Layer
(B) Convolution Layer
(C) Pool Layer
(D) Activation Function Layer

16. In which ANN, loops are allowed? | 1 | 1 | 4

(A) FeedForward ANN
(B) ForwardFeed ANN
(C) FeedBack ANN
(D) Recurrent Neural Network

17. How is the optimal number of clusters typically determined in K-means clustering? | 1 | 1 | 5

(A) By employing an elbow plot or silhouette analysis
(B) By using domain knowledge or expert judgment
(C) By selecting the number of clusters that minimizes the within-cluster variance
(D) By using cross-validation

18. Identify the non-zero vector that stays parallel after multiplication with covariance matrix. | 1 | 1 | 5

(A) Eigen value
(B) Linear value
(C) Gaussian value
(D) Eigen vector

19. PCA reduces the dimension by finding a few_____ | 1 | 1 | 5

(A) Hexagonal linear combination (B) Orthogonal linear combination

(C) Pentagonal linear combination (D) Octagonal linear combination

20. Which of the following functions can be used as an activation function in the output layer if we wish to predict the probabilities of n classes (p1, p2, …, pk) such that sum of p over all n equals to 1? | 1 | 2 | 5

(A) Sigmoid (B) ReLU

(C) Tanh (D) Softmax

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

Answer **any 5** Questions

21. Researchers took 20 cars of the same to take part in a study. These cars are randomly doped with one of the four-engine oils and allowed to run freely for 100 kilometres each. At the end of the journey, the performance of each of the cars is noted. Implement the python code for suitable Anova test. | 4 | 3 | 1

22. Find Root Mean Square Error for the following data sample using linear regression model Y1=mx+c where m=0.43 and c=0.72. | 4 | 3 | 2

X	Y
1	1
2	2
3	1.5
4	2.75
5	3.25

23. Suppose 10000 patients get tested for flu; out of them, 9000 are actually healthy and 1000 are actually sick. For the sick people, a test was positive for 620 and negative for 380. For the healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the accuracy, precision and recall for the data. | 4 | 3 | 2

24. Express the curse of dimensionality with 1-D,2-D and 3-D with example | 4 | 2 | 3

25. Consider a point that is correctly classified and distant from the decision boundary. Why would SVMs decision boundary be unaffected by this point, but the one learned by logistic regression be affected? Why does the kernel trick allow to solve SVMs with high dimension feature spaces, without increasing run time? | 4 | 2 | 4

26. 15. Given two data points in a 2-dimensional space: | 4 | 3 | 4

- x1=(2,3)
- x2=(3,4)

(i) Calculate the output for the Linear SVM kernel function.

(ii) Calculate the output for the RBF SVM kernel function $\sigma=2$.

27. Illustrate the use of elbow method with example | 4 | 1 | 5

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

Answer **all** Questions

28. (a) 1. Suppose we have a dataset representing the number of goals scored by a soccer team in 10 matches: {2, 1, 3, 2, 4, 1, 2, 2, 3, 1}. Calculate the statistical mean, median, and mode of this dataset. [4 Marks] — 12 | 3 | 1

2. Illustrate Hypothesis Testing and its steps with an example. [8 Marks] — 8 | 3 | 1

(OR)

(b) A research team is conducting an experiment to compare the effectiveness of three different fertilizers (Fertilizer A, Fertilizer B, and Fertilizer C) in promoting plant growth. They randomly selected five plants and applied each fertilizer to one plant. After a month, they measured the height of each plant (in centimetres), and the results are as follows: — 12 | 3 | 1

Fertilizer A: [32, 34, 30, 33, 31]
Fertilizer B: [35, 37, 34, 38, 36]
Fertilizer C: [31, 32, 30, 33, 29]

At a 5% significance level and degrees of freedom (2, 12), the critical value is approximately 3.89
Is there a significant difference in the mean plant height among the three fertilizers? Perform a one-way ANOVA to find out.

29. (a) 1. Write the logistic regression equation and write the issues solved by logistic regression over linear regression. [4 Marks] — 12 | 3 | 2

2. Discuss about linear regression in detail and find the coefficients value of linear regression model using the following sample data, also draw the learning curve.[8 Marks] — 8 | 3 | 2

X	Y
1	1
2	2
3	1.30
4	3.75
5	2.25

(OR)

(b) Write the importance of ridge and lasso machine learning models and explain with examples along with python codes. — 12 | 3 | 2

30. (a) 1. Express the curse of dimensionality with 1-D,2-D and 3-D with example. [4 Marks]

2. Let's consider a diabetes dataset with two features: BMI and Age. We want to predict whether the person has diabetes (label 1) or not diabetes (label 0) based on these features. Here's a small table of the dataset.Predict the label for a new person **BMI=43.6, Age=40** using KNN with K = 3. [8 Marks]

BMI	Age	Diabetes
33.6	50	1
26.6	30	0
23.4	40	0
35.3	23	1
35.9	67	1
45.5	55	0

(OR)

(b) 1. Explain Bayes Theorem with conditional probability. [4 Marks]

2. The following dataset contains loan information and can be used to try to predict whether a borrower will default (the last column is the classification). Use the naïve Bayes method to determine whether a loan X= (Home Owner = No, Marital Status=Married, Income=High) should be classified as a Defaulted Borrower or not. So, determine which is larger, P(Yes|X) or P(No|X). [8 Marks]

31. (a) 1. Explain the model for support vector machine for the linearly separable case. Write down the equation for the error function and explain it. Explain how it chooses a large margin decision. [8 Marks]

2. Give the importance of kernel functions and briefly explain the concept. [4 Marks]

(OR)

(b) Describe in detail about Forward Propagation technique in neural network with an example.

32. (a) Consider the following 12 instances with their X and Y values. Determine the optimal clusters out of the data k-means working methodology.

Instance	1	2	3	4	5	6	7	8	9	10	11	12
X	7	2	6	3	6	5	3	1	5	7	7	2
Y	8	4	4	2	5	7	3	4	4	7	6	1

(OR)

(b) Consider the below sample data. Find Eigen values and Eigen vectors and reduce the 2D data into one dimension using Principal Component Analysis.

Instance	1	2	3	4	5	6	7	8	9	10
X	0.72	0.18	2.50	0.45	0.04	0.13	0.30	2.65	0.91	0.46
Y	0.13	0.23	2.30	0.16	0.44	0.24	0.03	2.10	0.91	0.32

* * * * *