INDIAN STATISTICAL INSTITUTE

MS QMS

TEST ON MULTIVARIATE DATA ANALYSIS

Date: 30 April 2025 Time: 3 hours Maximum Marks: 50

Answer as many questions as you can. The maximum you can score is 50 marks

1. Suppose we collect data for a group of students in a statistics class with variables x_1 = hours studied, x_2 =undergrad GPA, and y = receive an A grade. We fit a logistic regression and produce estimated coefficient, β_0 = -6, β_1 = 0.05, β_2 = 1.
   a. Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A grade in the class.
   b. How many hours would a student with an undergrad GPA of 3.5 need to study to have a 50% chance of getting an A grade in the class?

[10]

2. Suppose that we wish to predict whether a given stock will issue a dividend this year (*Yes* or *No*) based on x, last year's percent profit. We examine a large number of companies and discover that the mean value of x for companies that issued a dividend was 10, while the mean for those that didn't was 0. In addition, the variance of x for these two sets of companies was $\sigma^2 = 36$. Finally, 80% of companies issued dividends. Assuming that x follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $x = 4$ last year.

[10]

3.
   a. Compare ridge and lasso regression in terms of their regularization techniques. When would you choose one over the other?
   b. Explain ridge regression and describe how it addresses multicollinearity in linear models.
   c. Compare and contrast ridge regression with ordinary least squares regression. What are the key differences
   d. Discuss the impact of the regularization parameter λ in ridge regression. How does it affect the model coefficients?
   e. Explain lasso regression and describe how it performs variable selection.

[10]

4. Fit a regression tree model to the following data to predict the response hardness. Provide detailed explanation of the procedure used especially describe how features are selected, branch point identified, the performance metric optimized at each split?

SL No	Temperature	Time	Hardness	SL No	Temperature	Time	Hardness
1	150	60	58	7	100	90	53
2	150	60	56	8	100	90	54
3	100	60	51	9	100	60	50
4	150	90	60	10	150	60	57
5	150	90	59	11	100	90	55
6	100	60	52	12	150	90	61

[13]

5. Imagine a company wants to understand customer preferences for a new smartphone. They identify three key attributes with different options as shown below:

SL No	Attribute	Option 1	Option 2
1	Battery Life	24 hours	36 hours
2	Camera Quality	12 MP	24 MP
3	Price	Rs. 20,000	Rs. 25,000

Participants are presented with different combinations of these attributes are asked to rank or rate their preferences for each combination. The data collected is given below.

Combination	Battery life	Camera Quality	Price	Preference Score
1	24 hours	12 MP	Rs. 25,000	5
2	24 hours	24 MP	Rs. 20,000	8
3	36 hours	12 MP	Rs. 20,000	7
4	36 hours	24 MP	Rs. 25,000	7

a. Analyze the data and identify the optimum combination of features or attributes?
b. Compute the part worth utility, importance score and interpret the results?
c. Estimate the performance score for the optimum combination of attributes?

[12]