[This question paper contains 4 printed pages.]

**Your Roll No.....**

**Sr. No. of Question Paper : 9545** **K**

Unique Paper Code : 2353200014

Name of the Paper : Mathematical Data Science

Name of the Course : **B.A/B.Sc. (Programme) Mathematics**

Semester : VII

Duration : 3 Hours Maximum Marks : 90

**Instructions for Candidates**

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. Attempt **all** questions by selecting any **two** parts from each.
3. Use of Scientific Calculator is allowed.

1. (a) Differentiate between structured and unstructured data. Give examples of each type.

   (b) What are the common problems faced with data? Explain the issues of dirty and missing data in tabular formats.

   (c) Consider the following dataset collected from a fitness study :

   (i) Type of Exercise (Yoga, Running, Swimming)

   (ii) Gender (Male/Female)

   (iii) Duration of Exercise (minutes)

   (iv) Satisfaction Level (Low, Medium, High)

   (v) Weight (kg)

   Classify each variable according to its data type and explain why.

*P.T.O.*

2. (a) Discuss anomaly detection techniques and their applications in real-world data science problems.

   (b) Define Data quality and Data hygiene. Describe methods used for assessing and improving data quality.

   (c) Explain meta-data and its role in data description. Differentiate between XML and JSON formats with examples.

3. (a) Define the Probably Approximately Correct (PAC) and explain its key parameters. Compare how Chernoff-Hoeffding inequality provide bounds on the deviation between empirical and true error, and relate these bounds to error tolerance and probability failure.

   (b)

   Consider $X_1, X_2, \cdots, X_{20}$ iid random variables with expected value $E[X_i] = 18$ and variance $Var(X_i) = 4$. Assume we also know that each $X_i$ must satisfy $12 \leq X_i \leq 22$.

   i. Use the Chebyshev inequality to find the upper-bound $P[\bar{X} > 19]$.

   ii. Use the Chernoff-Hoeffding inequality to find the upper-bound $P[\bar{X} > 19]$.

   (c)

   Show how Mahalanobis distance generalizes the Euclidean distance. Under what condition do the two coincide? Given two vectors $a = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$;

   $b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$; $and\ M = \begin{bmatrix} \frac{1}{3} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$. Compute the Mahalanobis distance $d_M(a, b)$.

4. (a)

| $x$ | 0.569 | 0.556 | 0.8944 | 0.9532 | 0.5852 |
|---|---|---|---|---|---|
| $y$ | 0.8496 | 0.8134 | 1.7522 | 1.8413 | 0.9571 |

(i) On the above data, run polynomial regression for $p = 3$. Report polynomial models for each. With each of these models, predict the value of y for a new $x$ values of 0.4 and 0.6.

(ii) Use cross-validation to predict an error, with error of a single data point ($x$, y) from a model $M$ as $(M(x) - y)^2$. Describe how you did this, and which data was used for what.

(b)

| X | | | y |
|---|---|---|---|
| 1/3 | 2/3 | 1/3 | 1/4 |
| 4/3 | 1/2 | 2/3 | 1/3 |
| 1/5 | 4/3 | 1/3 | 2/5 |
| 2/5 | 1/4 | 2/5 | 1/8 |
| 1/2 | 1/5 | 1/3 | 2/3 |

Solve for the coefficient

$$\text{Ridge Regression } \alpha_s = (X^T X + s^2 I)^{-1} X^T y$$

with s = 2; that is generate the ridge regression solution. For this, report the error in the estimate $\hat{y}$ of $y$ as $\|y - X\alpha_s\|_2 = \|y - \hat{y}\|_2$.

(c) Consider the function

$$f(x) = (x-3)^2 + (y+1)^2 - 4xy$$

Starting with ($x$, y) = (0,1) run the gradient descent algorithm for each function. Run for two iterations, and report the function value at the end of each step.

(i) First, run with a fixed learning rate of $\gamma = 0.1$ for f.

(ii) Second, run with any variant of gradient descent you want. Try to get the smallest function value after two steps.

5. (a) Find Singular Value Decomposition (SVD) of a matrix $A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$.

(b) Find linear Discriminant projection vector for 2-dimensional data of two classes as follows :

*Class* 1: $X_1$ = { (1,2), (2,3), (3,3), (4,5), (5,5)}

*Class* 2: $X_2$ = {(1,0), (2,1), (3,1), (4,2), (5,3), (6,5) }

(c) Compute the first Principal Component for the following 2-dimensional data :

| | | | | | |
|---|---|---|---|---|---|
| x | 2.5 | 0.5 | 2.2 | 1.9 | 3.1 |
| y | 2.4 | 0.7 | 2.9 | 2.2 | 3 |

6. (a) What are differences between Principal Component Analysis (PCA) and Linear discriminant analysis.

(b) Draw Voronoi diagram with four points A (0,1), B (0,4), C (2,0) and D (1,4).

(c) Use Lloyd's Algorithm for k-means clustering to divide the following data into three cluster with initial cluster centers $C_1$ = (1,10), $C_2$ = (2,4) *and* $C_3$ = (5,9).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| x | 1 | 2 | 2 | 5 | 7 | 5 | 1 | 8 |
| y | 10 | 5 | 4 | 8 | 5 | 9 | 2 | 9 |