Predict –1 or +1 using a kNN (k-nearest neighbor) classifier with $k = 3$ on $x = 3$ and $x = -1$.

(7.5)

[This question paper contains 8 printed pages.]

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

Sr. No. of Question Paper : **1227** **I**

Unique Paper Code : 2353010007

Name of the Paper : DSE-3(i) : Mathematical Data Science

Name of the Course : **B.Sc. (H) Mathematics**

Semester : V, DSE

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 **three** parts from **Q1-Q4**. and two parts from **Q5.-Q6**.

3. Parts of the questions to be attempted together.

4. **All** questions carry equal marks. **All** parts of a question carry equal marks. Marks are indicated.

5. Use of non-programmable Scientific Calculator is allowed.

P.T.O.

1. (a) Differentiate between structured and unstructured data. (5)

   (b) What is preprocessing of data. Explain with the help of an example. (5)

   (c) Consider the data of coffee shops with following attributes: Name of the coffee shop, Revenue in Rs., Pin code of the shop, Average monthly customers. Briefly explain the attributes as quantitative and qualitative. (5)

   (d) Define discrete quantitative and continuous quantitative data. Give an example of each. (5)

2. (a) Write five steps of data science and explain these briefly. (5)

6. (a) Consider a matrix $A \in \mathbb{R}^{8 \times 4}$ with squared singular values $\sigma_1^2 = 10$, $\sigma_2^2 = 5$, $\sigma_3^2 = 2$ and $\sigma_4^2 = 1$.

   (i) Find the rank of matrix A.

   (ii) Find the value of $\|A - A_2\|_F^2$, where A_2 is the best rank-2 approximation of A. (7.5)

   (b) Draw Voronoi diagram with three points A(–6,7), B(–6, –3) and C(2,5). (7.5)

   (c) Consider a set of 1 - dimensional data points

   $(x_1 = 0,\ y_1 = +1)$, $(x_2 = 1, y_1 = -1)$, $(x_3 = 2,\ y_1 = +1)$, $(x_4 = 4, y_1 = +1)$, $(x_5 = 6, y_1 = -1)$, $(x_6 = 7, y_1 = -1)$, $(x_7 = 8, y_1 = +1)$, $(x_8 = 9, y_1 = -1)$.

(b) Use Lloyd's Algorithm for k-means clustering to divide the following data into two clusters with initial cluster centers $C_1 = (2,1)$ and $C_2 = (2,3)$

x	1	2	2	3	4	5
y	1	1	3	2	3	5

(7.5).

(c) Reduce the feature 2-dimensional data of the following into one dimensional using Principal Component Analysis (PCA)

Class → Feature ↓	C_1	C_2	C_3	C_4
X_1	4	8	13	7
X_2	11	4	5	14

(7.5)

(b) Explain any two tabular data formats out of CSV, Excel, Data Frame, XML or similar. (5)

(c) What are anomalies in data science? Explain anomalies with examples. (5)

(d) Give an overview of data source format SQL or HDF5. (5)

3. (a) State Chebyshev Inequality. Consider a probability density function f and a random variable $X \sim f$. If $E[X] = 20$ and $Var[X] = 9$, find the value of $Pr[X \geq 50]$. (5)

(b) Explain the term probably approximately correct with an example. Consider sets $A = \{1,2,4,8\}$ and $B = \{1,2,3\}$. Calculate the Jaccard distance between A and B. (5)

(c) Define L_p distances mathematically and sketch these geometrically for p = 1,2 and ∞. (5)

(d) Consider two vectors a = (1,2, −4,3, −6) and b = (1,2,5, −2,1) in $\mathbb{R}^5$. Find Kullback-Liebler divergence. (5)

4. (a) Find the regression line for the given data

height (in):	66	68	60	70	65	61	74	73	75	67
weight (lbs):	160	170	110	178	155	120	223	215	235	164

(5)

(b) Explain polynomial regression, Consider the input data set with n = 3 points {(2,1), (3,6), (4,5)}. Find polynomial expansion of x generates with p = 5. (5)

(c) Define the Gradient in data sciences and find the value of the Gradient for

$$f(x,y,z) = 3x^2 - 2y^3 - 2xe^z \text{ at } (3,-2,1).$$ (5)

(d) Consider two-variable function $f = (x-5)^2 + (y+2)^2 - 2xy$. Starting with (x, y) = (0,2), using the gradient descent algorithm for the function, perform 3 iterations and report the function value at the end of each step. (5)

5. (a) Find Singular Value Decomposition (SVD) of the matrix $A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$. (7.5)