[This question paper contains 8 printed pages.]

Your Roll No.....

Sr. No. of Question Paper : 4986 **H**

Unique Paper Code : 2342202402

Name of the Paper : Data Mining II

Name of the Course : **B.A. (P) (NEP)**

Semester : IV

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. **Section A** is compulsory.
3. Attempt any **four** questions from **Section B**.
4. Parts of a question must be answered together.
5. Use of scientific calculator is allowed.

Section A

1. (a) How does the number of clusters affect anomaly detection in k-means clustering algorithm? (2)

P.T.O.

(b) In a dataset of monthly sales figures for a retail store, the mean monthly sales are Rs. 50,000 with a standard deviation of Rs. 5,000. In a certain month, the store recorded sales of Rs. 65,000. Calculate the z-score for this month's sales. (2)

(c) Consider a dataset with binary labels. The dataset is trained using Adaboost method. The decision boundary obtained after a single iteration is shown in figure II. (3)

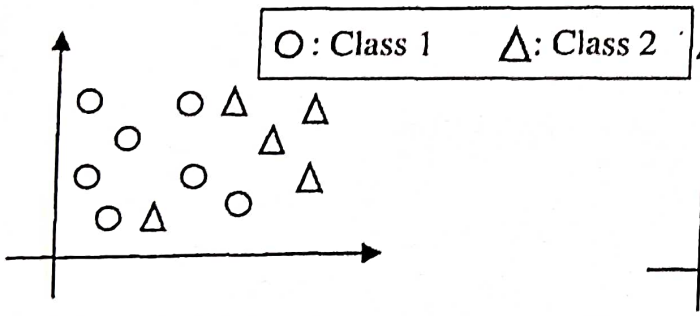

Figure I: Original Dataset

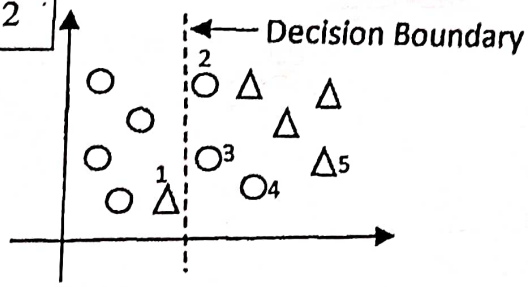

Figure II: Classification after 1st iteration

In figure II, out of the points marked 1, 2..., 5, which points shall have higher weights? Justify your answer.

(d) What is overfitting in the context of classification? Name two methods to prevent it. (3)

(e) Can clustering be used for dimensionality reduction? Justify your answer. (3)

(f) What is downsampling? Perform downsampling on the data given below : (3)

Class 0	9000 samples
Class 1	1000 samples

(g) Discuss any two key issues and their solutions in hierarchical clustering. (4)

(h) How does a proximity matrix differ from data matrix and why is it advantageous for anomaly detection methods? (4)

(i) Consider the following dataset with two documents related to customer's reviews of a product:

Document 1: "The product is excellent. I am very satisfied with its performance."

Document 2: "This product exceeded my expectations. It works flawlessly."

Perform the following tasks :

(i) Calculate the Term Frequency (TF) for the term "product" in each document.

(ii) Compute the Inverse Document Frequency (IDF) for the term "product" in the given document corpus.

(iii) Apply Frequency Damping with a damping parameter k=0.2 to the TF values obtained in part (i) for the term "product" in each document. (6)

Section B

2. (a) Consider the following distance matrix for the five data points P1 to P5. (7)

	P1	P2	P3	P4	P5
P1	0				
P2	0.25	0			
P3	0.23	0.17	0		
P4	0.38	0.21	0.17	0	
P5	0.36	0.15	0.27	0.30	0

Perform hierarchical clustering using centroid linkage on the given distance matrix and show the dendrogram.

(b) Consider the following data set comprising of eight points A1, A2, ..., A8 on a plane. (8)

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

K-means clustering algorithm is applied to find the three clusters by keeping A1(2, 10), A4(5, 8) and A7(1, 2) as initial cluster centers. Find the cluster centres obtained in the next iteration. Use Manhattan distance to find distance/similarity among data points. Show all the intermediate steps.

4986

3. (a) What is the primary objective of time series analysis in data mining? Consider the following time series dataset containing daily temperature fluctuation over a week: (7)

Monday: 25°C

Tuesday: 24°C

Wednesday: 22°C

Thursday: 26°C

Friday: 28°C

Saturday: 29°C

Sunday: 27°C

If the dataset is shifted by two days forward, what will be the temperature on Tuesday of the second week?

(b) What is the significance of the Bagging algorithm in ensemble learning? How does Bagging combine multiple weak learners to create a strong classifier, illustrate with the help of a diagram. Suppose we have a dataset with the following features and labels :

Features (X)	1	2	3	4	5	6	7	8	9	10
Labels (Y)	1	0	1	1	0	0	1	0	1	0

Generate two sub datasets containing four samples each from the original dataset using bootstrap sampling method with replacement and two sub datasets without replacement. (8)

4. (a) Discuss the relationship between entropy and information content and provide mathematical formula to calculate entropy. (9)

Consider the following dataset of 10 instances labelled as either "Positive" or "Negative", calculate entropy on this dataset.

X	Label
1	Positive
2	Negative
3	Positive
4	Positive
5	Negative
6	Positive
7	Negative
8	Negative
9	Positive
10	Positive

(b) Compare the following anomaly detection techniques :– (6)

(i) Model-Based and Model-Free method

(ii) Label and Score method

(iii) Global and Local Perspective method

5. (a) Explain the E-step and M-step in the Expectation-Maximization (EM) algorithm and (9) Probabilistic Latent Semantic Analysis (PLSA). How are the parameters of PLSA updated in each step? Illustrate the generative processes of EM-clustering and PLSA through appropriate diagrams. (9)

(b) Consider the following text in a document D1:

Document D1: "Text mining is a technique used to extract useful information from text documents,"

Perform the following text mining preprocessing steps on the text given and write your answer:

(i) Stop Word Removal

(ii) Stemming

(iii) Removal of punctuation marks (6)

6. (a) What is the principle of the STREAM algorithm? How does it handle data streams effectively? (5)

(b) Consider the following stream of data points arriving with timestamps. The data points are as

follows :

(10)

Time Stamp	Data Point
0	(2,3)
1	(3,4)
2	(5,6)
3	(8,9)
4	(10,12)

Apply the STREAM algorithm with a window size of 2 and determine the micro-clusters formed at the end of the data stream.

7. (a) Describe the steps of a random forest classifier. Discuss the impact of attribute selection at internal nodes on improving the classifier's performance.

(5)

(b) Consider the following dataset of six data points :

A(1, 2), B(2, 3), C(3, 4), D(7, 8), E(8, 9), F(25, 30)

Apply the Density-Based Spatial- Clustering of Applications with Noise (DBSCAN) algorithm on the given data with Epsilon (ε) = 2.5 units, and Minimum Points (minPts)= 2. Identify the clusters, core points, reachable points and noise points.

(10)