VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

REG.NO.:

SLOT: A1+TA1

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

Programme Name & Branch	: BTECH & Common to all branches
Course Codeand Course Name	: BCSE209L and Machine Learning
Faculty Name(s)	: Sanjiban Sekhar Roy
Class Number(s)	: VL2025260102101
Date of Examination	: 5/10/2025, TIME : 2PM -3:30PM
SLOT	: A1+TA1
Exam Duration	: 90 minutes Maximum Marks: 50

General instruction(s): Answer all questions (5X10=50 MARKS)

<table>
<tr><th>Q. No</th><th>Question</th><th>Module</th><th>Marks</th><th>CO</th><th>BL</th></tr>
<tr><td>1.</td><td>You are given the following categorical dataset of 6 customers and their attributes:
<table>
<tr><th>Customer</th><th>Gender</th><th>Marital Status</th><th>Product Preference</th></tr>
<tr><td>C1</td><td>Male</td><td>Single</td><td>Electronics</td></tr>
<tr><td>C2</td><td>Female</td><td>Married</td><td>Clothing</td></tr>
<tr><td>C3</td><td>Male</td><td>Single</td><td>Clothing</td></tr>
<tr><td>C4</td><td>Female</td><td>Single</td><td>Electronics</td></tr>
<tr><td>C5</td><td>Male</td><td>Married</td><td>Clothing</td></tr>
<tr><td>C6</td><td>Female</td><td>Married</td><td>Electronics</td></tr>
</table>
(a) Using K-Modes Clustering with k = 2, perform one iteration of clustering using the following initial modes (cluster centers):
• Cluster 1: ['Male', 'Single', 'Electronics']
• Cluster 2: ['Female', 'Married', 'Clothing']
(i) Compute the dissimilarity (number of mismatches) between each data point and the cluster centers.
(ii) Assign each customer to the nearest cluster based on dissimilarity.
(iii) Recalculate the new modes (centroids) for each cluster after assignment.
(b) Discuss the fundamental assumptions and limitations of K-Modes clustering in handling high-dimensional categorical data. How do these assumptions impact clustering quality, and what potential modifications or hybrid approaches could address these challenges in real-world, large-scale datasets?</td><td>4</td><td>(7+3)</td><td>3</td><td>3</td></tr>
<tr><td>2.</td><td>You are given the following dataset with 4 observations and 2 features:
<table>
<tr><th>Obs</th><th>Feature 1 (X1)</th><th>Feature 2 (X2)</th></tr>
<tr><td>1</td><td>2</td><td>0</td></tr>
<tr><td>2</td><td>0</td><td>1</td></tr>
<tr><td>3</td><td>3</td><td>1</td></tr>
<tr><td>4</td><td>4</td><td>3</td></tr>
</table>
(a) Standardize the data (mean = 0, variance = 1) for each feature.
(b) Calculate the covariance matrix of the standardized data.
(c) Compute the eigenvalues and eigenvectors of the covariance matrix.
(d) Identify the principal components and explain how much variance each component explains.
(e) PCA assumes that the directions of maximum variance correspond to the most important structure in the data. Under what real-world conditions might this assumption fail, and how could this affect model performance in downstream tasks like classification or anomaly detection?</td><td>4</td><td>(1+2+2+2+3)</td><td>3</td><td>3</td></tr>
</table>

REG.NO.:

SLOT: A1+TA1

VIT

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

3.	In an experiment to demonstrate the Expectation-Maximization (EM) Algorithm, two coins A and B are tossed multiple times. The outcomes are as follows: Coin B: 6H,4T; Coin A: 8H,2T; Coin A: 7H,3T; Coin B: 5H,5T; Coin A: 9H,1T a) From the above trials, compute the initial estimates of θ_A and θ_B, where θ denotes the probability of heads for each coin. b) Briefly explain how the EM algorithm updates these probabilities when the coin labels are unknown. c) Using your initial values, perform two iterations of the EM algorithm and report the updated values of d) How can the EM algorithm be extended to estimate the parameters of a Gaussian Mixture Model when the number of clusters is unknown? Write your analysis.	4	(2+2+4+2)	3	3
4.	You are given a dataset of 6 samples with binary class labels: Sample | Feature (X) | Label (Y) 1 | 1 | +1 2 | 2 | +1 3 | 3 | +1 4 | 4 | -1 5 | 5 | -1 6 | 6 | -1 AdaBoost runs for 2 rounds with these weak below classifiers: Weak classifier 1 (h_1): predicts +1 if X < 3.5, else -1 Weak classifier 2 (h_2): predicts +1 if X < 4.5, else -1 (a) Initialize sample weights uniformly. Write the initial weights. (b) For the first classifier (h_1): Calculate weighted error rate ε_1. Compute classifier weight α_1. Update sample weights for the next round (normalize weights). (c) For the second classifier (h_2): Calculate weighted error rate ε_2. Compute classifier weight α_2. Update sample weights after round 2 (normalize weights).	5	(2+4+4)	2	3
5.	Explain how ensemble classifiers address the bias–variance trade-off. Discuss the roles of Bagging, and Boosting in improving model performance and mention any limitations of ensemble methods.	5	(5+5)	3	4