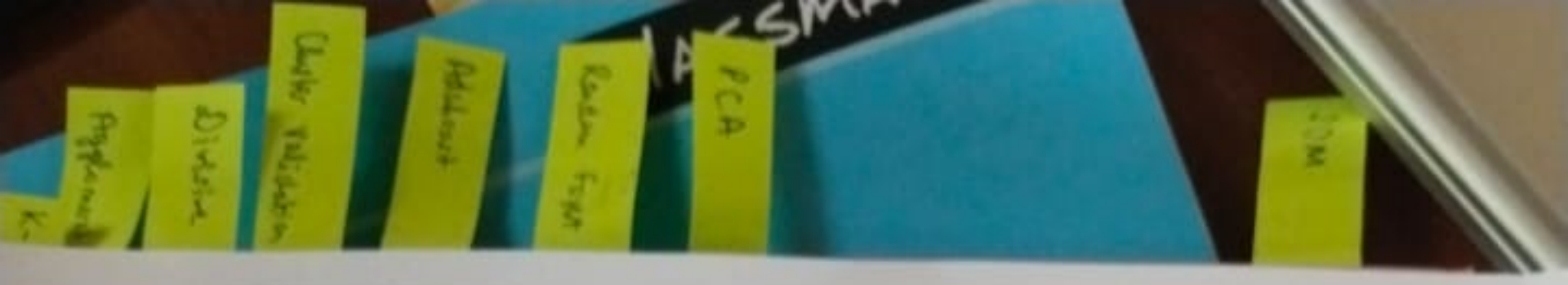


SLOT: B2+TB2

REG.NO.: 21BCE2857

VIT
Vellore Institute of Technology

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2024-2025

Programme & Branch: **B.Tech. CSE** Course Code & Name: **BCSE209L – Machine Learning**
Class Number(s) :**1595, 1598, 1602, 1689, 1697, 1708, 1713, 1719, 1726, 1730, 1737**
Date of Examination : **14/10/2024** Duration : **90 minutes** Maximum Marks: **50**

General instruction(s):

- ***Answer All Questions***
- **Course Outcomes: CO3.** To understand the need for Reinforcement learning in real – time problems. **CO4.** Apply all learning algorithms over appropriate real-time dataset

1. Consider a set of 2-dimensional training data points (x1, x2) belonging to two classes '+1' and '−1', respectively, as shown below. **[2 + 2 + 4 +2 marks][CO3][BL3]**

– **Class '+1'**: (6, 4), (6, 3), (7, 2), (8, 3)

– **Class '-1'**: (2, 3), (3, 2), (4, 4), (4, 2)

(a) Graphically represent the constellation of data points and the optimal separating hyperplane. Write the equation of the optimal separator and mention the width of the margin (figuring it out manually from the diagram/graph you have shown).

(b) List the support vectors and justify your selection of support vectors.

(c) What weight vector and threshold (bias) value are being learnt using hard-margin SVM training algorithm with these eight training points? Show the detailed calculations.

(d) Consider the decision rule: $H(x) = \langle w, x \rangle + b$. Explain how this equation classifies points on either side of a line. Determine the class for the points (3, 4), (7, 4) and (5, 5).

2. Suppose that the points (P1, P2, P3, P4, P5, P6) provided in a 2D plane. **Table 1** is the distance matrix which shows the Euclidean distances between pair of data points. **[10 marks][CO3][BL3]**

	P1	*P2*	*P3*	*P4*	*P5*	*P6*
P1	0					
P2	0.12	0				
P3	0.51	0.25	0			
P4	0.84	0.16	0.14	0		
P5	0.28	0.77	0.70	0.45	0	
P6	0.34	0.61	0.93	0.20	0.67	0

Table 1

Phase →	*1*	*2*	*3*	*4*	*5*
New cluster	{P1, P2}				
Distance to choose the cluster	0.12				

Table 2

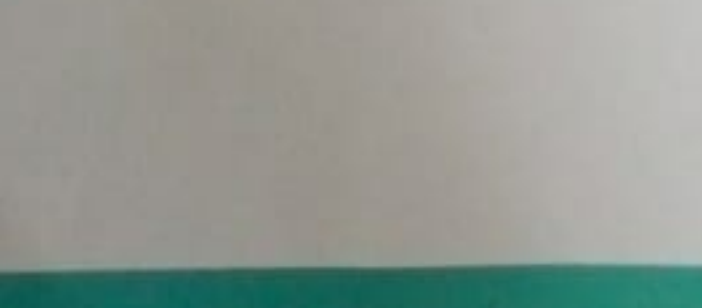
Using **Single and Complete Linkage Hierarchical Agglomerative Clustering** technique, for each phase, find the clusters, show the calculations, and update the distance matrix. Also, fill **Table 2** with the final distances (copy the table for both methods) and draw the final dendrogram. Phase 1 result is shown for you.

3. Given the (X,Y) coordinates of six data points in two-dimensional space: (1,0), (2,0), (3,0), (5,6), (6,6), and (7,6), calculate the first principal component and fill-in the same in a table against each sample. Show your calculations in detail. **[10 marks][CO3][BL3]**

4. Given the following dataset with 10 samples, compute a partitioning using K-modes algorithm with samples 1, 5 and 10 as initial cluster centroids. Do partition for 2 iterations. At the end of each iteration, show the updated distance matrix and the newly constructed centroids. **[10 marks][CO3][BL4]**

Sample No.	*X1*	*X2*	*X3*	*X4*	*X5*
1	A	B	A	B	C
2	A	B	A	B	B
3	C	A	B	B	A
4	A	B	C	A	C
5	C	C	C	B	A
6	A	A	B	A	A
7	A	C	A	C	C
8	C	B	A	B	C
9	A	A	B	C	A
10	B	B	A	A	C

5. **(a)** Varying combinations of bias and variance are shown in the figures A and B below. Identify the combinations and discuss about the same in detail with appropriate examples. Also, write down the formula that connects test MSE to the bias, variance and irreducible error. **[6 marks][CO4][BL4]**



A B

(b) Graphs of different models with their degree of polynomials are given below. Which of them underfits, overfits and just right (best fitting)? Elaborate on your pick on each of these graphs. Note: + represents testing data and circle represents training data. **[4 marks][CO4][BL4]**