SASTRA

School of Computing
First CIA Exam –Aug 2025
Course Code: CSE425
Course Name:MACHINE LEARNING ESSENTIALS
Duration: 90 minutes Max Mark: 50

PART A

Answer ALL the questions **5x10=50**

1.(i) List out the Supervised Learning Algorithms in Machine Learning. Also explain the terms Capacity, Overfitting,Underfitting and Bias in Machine Learning Algorithms. (6)

ii) Outline the key steps involved in Feature Engineering for Machine Learning models. (4)

2.Find all frequent patterns and association rules from the following database by using the FP-Growth algorithm.
Take minimum support = 2.

TID	List of item IDs
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3

3. (i) Explain Information Gain, Gain Ratio, and Gini Index in the context of decision tree algorithms. (5)

(ii) Write the Apriori Algorithm used for mining frequent itemset in Association Rule Learning. (5)

4.(i)List and explain any five Pattern Evaluation Metrics used in Association Rule Mining.For each metric, provide a) Definition, b) Formula and c) Purpose in evaluating association rules. (5)

ii) Explain the method that enhances efficiency during frequent itemset mining by minimizing candidate generation, and support your answer with an example. (5)

5.i)A zoologist is studying two species of frogs: Tree Frog (T) and Ground Frog (G). The characteristics observed are Skin Color, Leg Length, Body Size, and Poisonous (Yes/No). The dataset of observed frogs is given below:

No	Skin Color	Leg Length	Body Size	Poisonous	Species
1	Green	Long	Small	Yes	T
2	Brown	Short	Large	No	T
3	Green	Long	Small	Yes	T
4	Brown	Short	Small	Yes	T
5	Green	Short	Large	No	G
6	Brown	Short	Large	No	G
7	Brown	Short	Small	No	G
8	Brown	Short	Large	Yes	G

Choose an appropriate classification approach to estimate the probability values for the new frog instance (Skin Color = Green, Leg Length = Short, Body Size = Large, Poisonous = No) and classify the frog as Tree Frog (T) or Ground Frog (G) based on the computed probabilities. (7)

ii) Convert the transaction {a, c, k} into the corresponding vertical data format representation using the given table. (3)

TID	Items
10	{a, b, c, d, e, f}
20	{b, c, e, g}
30	{a, c, d, e, f, h}
40	{b, d, f, e, j}
50	{b, c, k, f}

SASTRA
DEEMED TO BE UNIVERSITY
THANJAVUR | KUMBAKONAM | CHENNAI

School of Computing
Second CIA Exam – Sep 2025
Course Code: CSE425
Course Name: MACHINE LEARNING ESSENTIALS
Duration: 90 minutes Max Mark: 50

PART A

Answer Any FIVE the questions **5x10=50**

1. a) Given a set of labeled training data points and a new, unlabeled data point, determine the class of the new point using a K-Nearest Neighbors approach.

Training Data:

- **Point 1:** (2, 3) -> Class A
- **Point 2:** (5, 4) -> Class B
- **Point 3:** (9, 6) -> Class B
- **Point 4:** (4, 7) -> Class A
- **Point 5:** (8, 1) -> Class B and **New Data Point:** (6, 3)

Using the K value as 3 and a distance metric, what is the predicted class for the new data point? (5)

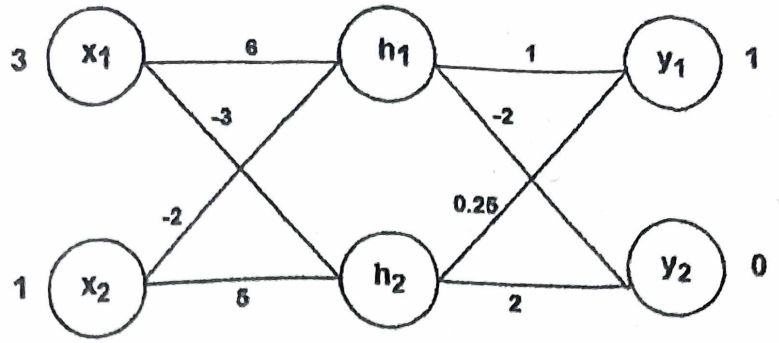
b) Write the step-by-step procedure used in the K-Nearest Neighbors (KNN) algorithm and discuss about key advantages and disadvantages of using the K-Nearest Neighbors (KNN) algorithm for machine learning tasks. (5)

2. A hidden Markov model has two states {S1, S2} with $\pi(S1) = 1$. It can emit the outputs O1 and O2. Find the probability of the output sequence O2, O1, O2 using the forward procedure.

State transition	S1	S2
S1	0.6	0.4
S2	0.7	0.3

Emission Table	O1	O2
S1	0.5	0.5
S2	0.2	0.8

3. Consider the diagram given below with (3,1) as inputs and (1,0) as targets. Perform a forward propagation by assuming sigmoid as activation function and compute the error in the network and write the chain rule for backpropagation.



The output y1=0.73 is somewhat close to 1 and the output y2=0.12 is close to 0.

4. Let X represent the amount of fertilizer used (in kilograms) for a crop field, and Y represent the crop yield (in quintals). Apply linear regression using the least squares method for one iteration using the following training set. Check the performance for the test cases where the fertilizer amount is x = 15 kg and x = 22 kg.

Training Data:

X (Fertilizer in kg)	5	10	12	14	18	20	25
Y (Crop yield in qtl)	12	20	24	28	36	40	50

5. a) You are building a classifier to detect early-stage cancer in patients using medical imaging data. The dataset is imbalanced (5% positive cases, 95% negative cases). The model's performance on the test set is as follows: True Positives: 120, False Positives: 150, True Negatives: 2850, False Negatives: 30. If the goal is to minimize missed cancer cases (False Negatives), which metrics should be used to evaluate the model? Calculate and interpret two relevant metrics based on the given values. (5)

b) Differentiate between Bagging & Boosting in ensemble learning. Write any four key differences and give one real-world application for each. (5)

6. a) Design a perceptron to implement the OR gate using the following parameters: w1=1, w2=1 and b=−0.5 and verify that the perceptron produces the correct output for all input combinations of the OR function. (5)

b) Explain the different regularization techniques used in machine learning. Write their mathematical formulations with proper equations. (5)

SASTRA
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act, 1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA
THANJAVUR | KUMBAKONAM | CHENNAI

School of Computing
Third CIA Exam – Nov 2025
Course Code: CSE425
Course Name: MACHINE LEARNING ESSENTIALS
Duration: 90 minutes Max Mark: 50

PART A

Answer Any FOUR questions **4x10=40**

1 Apply the Apriori Algorithm to find the frequent itemsets and association rules with a minimum support of 40% and minimum confidence of 70%.
Consider the following set of transactions:

Transaction ID	Items Bought
1	{Bread, Butter, Milk}
2	{Bread, Butter}
3	{Beer, Cookies, Diapers}
4	{Milk, Diapers, Bread, Butter}
5	{Beer, Diapers}

2. a) Write the Decision Tree algorithm and explain each step in detail with a suitable example. (8)
b) Discuss the different types of kernel functions used in Support Vector Machines (SVM). (2)

3. Consider a simple weather-based drink choice scenario with the following emission probabilities (probability of choosing a drink given the weather state):

State	cola	iced tea (ice_t)	lemonade (lem)
CP	0.6	0.1	0.3
IP	0.1	0.7	0.2

Assume the state transition probabilities are:

From \ To	CP	IP
CP	0.7	0.3
IP	0.4	0.6

Assume the initial state probabilities are: P(CP) = 0.6, P(IP) = 0.4

You observe the following sequence of drinks over three consecutive days:

observation = [cola, iced tea, lemonade]

1. Write the Viterbi algorithm. (5)
2. Use the Viterbi algorithm to determine the most probable sequence of states that generated the observed drink sequence. (5)

4. A data analyst is studying the spatial distribution of customer locations to identify natural groupings for targeted delivery zones. The analyst collects the following six 2D coordinates representing customer positions: **P1 (2,4), P2 (4,4), P3 (4,6), P4 (5,6), P5 (6,5) and P6 (8,2)** As a data scientist, you are asked to help the analyst group these customers based on proximity using Hierarchical Agglomerative Clustering (HAC) with the Euclidean distance metric and Single-Linkage criterion and construct the dendrogram and explain the sequence of cluster merges.

5. List and explain the different types of points and methods used in the DBSCAN clustering algorithm.

PART B

Answer All the questions **1x10=10**

6. a) A company wants to cluster customer locations based on their coordinates in a 2D plane. They have collected the following 5 customer locations (x, y):

Customer	x	y
C1	2	3
C2	3	5
C3	5	8
C4	8	8
C5	7	5

Using the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm, compute the Clustering Feature (CF) for the cluster containing all 5 customers. (5)

b) Explain how representative points help in clustering in the CURE (Clustering Using REpresentatives) algorithm. Draw a block diagram illustrating the detailed steps of the CURE algorithm. (5)