Total No. of Pages 04 | *Roll No. ......*

FIFTH SEMESTER | **B.Tech[CSE]**

END SEMESTER EXAMINATION ***Nov-Dec, 2023***

CO327 MACHINE LEARNING

Time: 3:00 Hours | *Max. Marks: 40*

Note: Answer *ALL* questions.
Assume suitable missing data, if any.
CO# is course outcome(s) related to the question.

1. Analyze and categorize the following tasks into distinct branches of machine learning, providing reasoning for your classifications:

a. Predicting anomalies in a complex network system using real-time data streams.

b. Identifying emergent patterns in a collection of unstructured textual data without labeled categories.

c. Designing an algorithm to optimize resource allocation in a smart grid system based on fluctuating demands and environmental factors.

d. Creating a model to anticipate traffic congestion patterns in a metropolitan area considering various influencing factors.

e. Developing a system for a humanoid robot to learn and adapt its movements in an unknown environment through continuous interaction and feedback.

f. Extracting meaningful insights from a vast database of satellite images to detect and predict natural disasters.

g. Constructing an algorithm to personalize online shopping recommendations for users based on their browsing history and behavior.

h. Formulating a strategy for an autonomous vehicle to navigate dynamic and unpredictable urban traffic scenarios while ensuring passenger safety.

In your analysis, delineate how each task aligns with either supervised learning, unsupervised learning, reinforcement learning, or a combination thereof. Provide justifications [NOT MORE THAN 2 sentences] for your classifications based on the nature of the learning problem, available data, and the specific characteristics of each machine learning branch. [1x8 = 8] [CO1]

2[a] You are provided with a dataset representing customer profiles and their purchasing decisions (for a sports bike). Your task is to construct a decision tree to predict whether a customer will make a purchase ('Buy'). Find the optimal threshold to make age a binary variable and build the decision tree for dataset in Table I using Gini index. Use median value approach to handle missing values. Additionally, to prevent overfitting, it's necessary to ensure that there are at least two samples at every leaf node following a split. [2+2] [CO1, CO2]

Table I

Customer ID	Age (Years)	Income	Car Owner	Credit Rating	Buy
1	25	Low	No	High	No
2	35	Low	Yes	Low	Yes
3	45	High	No	Medium	Yes
4	20	Medium	Yes	High	Yes
5	40	High	No	Low	No
6	35	Low	Yes	High	Yes
7	50	Medium	No	Medium	No
8	30	High	Yes	Low	Yes
9	22	Low	No	High	No
10	60	High	Yes	Medium	Yes

[b] Consider the dataset in Table I again and built a full decision tree ignoring overfitting condition. Use the decision trees built in 2[a] and 2[b] and find change in accuracy on test dataset in Table II. **[4] [CO2]**

Table II

Customer ID	Age (Years)	Income	Car Owner	Credit Rating	Buy
11	32	High	No	High	Yes
12	28	Low	No	Low	No
13	47	High	Yes	Medium	Yes
14	23	High	Yes	High	Yes
15	55	High	Yes	Low	No
16	37	Low	No	Medium	No
17	41	Low	Yes	High	No

3. Answer *any TWO* of the followings

[a] An epidemiological study was conducted to understand the relationship between lifestyle factors and the incidence of a particular chronic disease. The collected data is presented in Table III. It has three attributes: Age Group = younger (Y) or older (O), Dietary Habit = vegetarian (V) or non-vegetarian (NV), Exercise Frequency = regular (R) or irregular (I), and Smoking Habit = smoker (S) or non-smoker (NS). Predict the likelihood of developing the chronic disease (D = Yes, N = No) for an older individual with a non-vegetarian diet, regular exercise, and who is a smoker using the naïve Bayes classifier. Show each step clearly. **[4] [CO2, CO3]**

Table III

S. No.	1	2	3	4	5	6	7	8	9	10
Age Group	Y	Y	Y	O	O	O	O	O	Y	Y
Dietary Habit	V	V	NV	NV	NV	V	V	NV	V	NV
Exercise Frequency	I	R	I	R	I	R	R	I	I	R
Smoking Habit	S	S	NS	NS	S	NS	S	NS	NS	S
Disease	D	N	D	N	D	N	D	N	N	N

[b] A meteorological research team is analyzing a simplified dataset containing readings from different weather stations. The dataset is given in Table IV with features: Temperature and Humidity. The team decides to use the k-means clustering algorithm tc categorize these samples into two distinct groups (k = 2) based on their similarities in temperature and humidity. The initial centroids are chosen as (10, 80) for Centroid 1 and (20, 60) for Centroid 2. Perform k-means clustering for two iterations and provide the cluster assignments and centroids after each iteration. Using graphical representations, demonstrate how the intra-cluster distances change across iterations. **Hint:** intra-cluster distance in average distance of all points from centroid in a cluster. [4] [CO2]

Table IV

Sample No.	1	2	3	4	5	6
Temperature (°C)	10	11	10	20	21	20
Humidity (%)	80	79	81	60	59	61

[c] A health research team wants to classify patient data for a study on lifestyle diseases. The dataset is given in Table V with features: Body Mass Index (BMI) and Average Daily Steps. Each record has been classified as either "High" or "Low" for lifestyle diseases. Note that there is an outlier preset at record number 7. A new patient's record comes in with BMI = 27, and Average Daily Steps = 2900. Using the *k*-Nearest Neighbors algorithm classify the new patient's risk category based on the provided dataset. Identify the suitable value *k* to mitigate the effect of outlier. [1+1+2] [CO2]

Table V

Patient ID	1	2	3	4	5	6	7	8
BMI	25	30	28	35	24	40	26	45
Average Daily Steps	3000	2500	2800	2000	3200	1500	3100	1000
Risk Category	Low	High	Low	High	Low	High	High	High

4. Answer *any TWO* of the followings

[a] Consider a two-layer neural network used for binary classification. The network has an input layer with 2 neurons, a hidden layer with 2 neurons, and an output layer with 1 neuron. The activation function for the hidden layer is ReLU (Rectified Linear Unit), and for the output layer, it's a sigmoid function. The network is trained using the binary cross-entropy loss function and stochastic gradient descent (SGD) with a learning rate of 0.01. The initial weights and biases are as follows: Weights from input to hidden layer: $W_1 = [[0.5, -0.6], [-0.4, 0.8]]$, Bias for hidden layer: $b_1 = [0.2, -0.2]$, Weights from hidden to output layer: $W_2 = [0.3, -0.5]$, Bias for output layer: $b_2 = 0$. Consider the network is trained with a single training sample ($X = [1.0, 2.0]$, $Y = 0$). Perform the forward pass to calculate activations at hidden layer and output layer, and then compute the loss. [4] [CO2]

[b] Consider the neural network in 4[a] again and perform the backpropagation to update the weights and biases. Calculate the updated weights W_1, W_2, and biases b_1, b_2 after one iteration. Show your calculations for the forward pass, loss calculation, and backpropagation steps. [4] [CO2]

[c] Consider a dataset consisting of three observations, each with two features. The dataset is given in Table VI. Perform Principal Component Analysis (PCA) on this dataset.

[4] [CO2]

Table V

Observation	1	2	3
Feature 1	2.5	0.5	2.2
Feature 2	2.4	0.7	2.9

Hint: Eigen values of the matrix $\begin{bmatrix} 1 & 0.937 \\ 0.937 & 1 \end{bmatrix}$ are 1.937 with Eigenvector [0.707, 0.707] and 0.063 with Eigenvector [-0.707, 0.707].

5[a] A car can operate in three modes: Electric Mode (S1), Hybrid Mode (S2), and Gasoline Mode (S3), each representing a state in the Markov Decision Processes (MDP). The car's system must decide which mode to switch to, aiming to optimize fuel efficiency and battery usage. Actions and their rewards are defined for each state: from S1, switch to S2 (+10) or stay in S1 (0); from S2, switch to S3 (+5) or back to S1 (+2); from S3, end the trip (+20). With a discount factor of 0.5 and initial state values of 0, perform one iteration of value iteration, calculating updated values for each state. [4] [CO1, CO3]

[b] Consider an autonomous robot in a 4x4 warehouse grid, tasked with delivering items to a specific area. Each grid cell represents a different warehouse area. 'G' is the delivery area. The robot starts from area 1. The robot consumes energy (-0.2 reward) for each move and gains a significant energy saving (+5 reward) upon successful delivery to 'G'. The robot's route is: 1 → 2 → 3 → 4 → 8 → 12 → 11 → 15 → G. Using Temporal Difference Learning (TD(0)) with a learning rate (α) of 0.1 and a discount factor (γ) of 1, determine the updated efficiency value of area 1 after one delivery, starting from 0.

[4] [CO2]

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	G

Fig. 1

----Best of Luck----