Q. Code:229119

Reg. No. | | | | | | | | | | | | | |

B.E./B.TECH. DEGREE EXAMINATIONS, MAY 2024

Sixth Semester

CS18604 – MACHINE LEARNING TECHNIQUES

(Computer Science and Engineering)

(Regulation 2018 / 2018A)

TIME:3 HOURS **MAX. MARKS: 100**

COURSE OUTCOMES	STATEMENT	RBT LEVEL
CO 1	Students will be able to understand the basic concepts of machine learning techniques.	**2**
CO 2	Students will be able to apply the concepts of parameter estimation methods.	**2**
CO 3	Students will be able to understand the design of various non parametric methods and dimensionality reduction methods.	**3**
CO 4	Students will be able to differentiate various discriminative learning methods and its applications.	**3**
CO 5	Students will be able to develop the tree models and mixture of experts.	**2**

PART- A (10x2= 20 Marks)

(Answer all Questions)

		CO	RBT LEVEL
1.	Define supervised learning and provide an example of a supervised learning task.	**1**	**1**
2.	Discuss the role of discriminant functions in classification tasks.	**1**	**2**
3.	Describe the trade-off between bias and variance in model selection.	**2**	**1**
4.	Provide examples of real-world applications where Maximum Likelihood Estimation and model selection procedures are used in machine learning.	**2**	**2**
5.	Give advantages of non-parametric density estimation over parametric methods.	**3**	**2**
6.	Name some real-world applications where dimensionality reduction techniques are used?	**3**	**3**
7.	Differentiate logistic regression and linear regression.	**4**	**2**
8.	What is a perceptron, and how does it function in neural networks?	**4**	**3**
9.	Boosting enhance the performance of weak learners? Justify.	**5**	**2**
10.	How are decision trees used for rule extraction?	**5**	**2**

PART- B (5x 14=70 Marks)

		Marks	CO	RBT LEVEL
11. (a)	Consider the given dataset, Apply Naïve Bayes algorithm and predict that if an animal has the following properties: Name=Cow, Size=Medium, Body Color=Black. Can the animal be a pet?	**(14)**	**1**	**2**

S. No	Name	Size	Body Color	Can we pet them?
0	Dog	Medium	Black	Yes
1	Dog	Big	White	No
2	Rat	Small	White	Yes
3	Cow	Big	White	Yes
4	Cow	Small	Brown	No
5	Cow	Big	Black	Yes
6	Rat	Big	Brown	No
7	Dog	Small	Brown	Yes
8	Dog	Medium	Brown	Yes
9	Cow	Medium	White	No
10	Dog	Small	Black	Yes
11	Rat	Medium	Black	No
12	Rat	Small	Brown	No
13	Cow	Big	White	Yes

(OR)

(b)	Apply Aprori algorithm on the grocery store example with support threshold s = 33.34% and confidence threshold c =60%, where H, B, K, C and P are different items purchased by customers.	**(14)**	**1**	**2**

Transaction ID	Items
T1	H, B, K
T2	H, B
T3	H, C, P
T4	P, C
T5	P, K
T6	H, C, P

Show all final frequent itemsets.

Specify the association rules that are generated.

Show final association rules sorted by the confidence.

Represent the transactions as graph.

12. (a) A simple linear regression model is hypothesized to express drain current I (in milliamperes) as a function of ground-to-source voltage V (in volts) for a MOS transistor. The drain current and ground-to-source voltage data were measured as shown in Table below: **(14) 2 2**

Drain Current (mA)	**Gate-to-Source Voltage (volts)**
0.734	1.1
0.886	1.2
1.04	1.3
1.19	1.4
1.35	1.5
1.5	1.6
1.66	1.7
1.81	1.8
1.97	1.9
2.12	2.0

a) Draw a scatter diagram of these data: Does a straight-line relationship seem plausible?

b) Fit a simple linear regression model to these data.

(OR)

(b) Discuss in detail the maximum likelihood estimation and for Gaussian density distribution. **(14) 2 2**

13. (a) Consider the training examples shown in the following table for binary classification. The table shows a training set for a problem of predicting whether a loan applicant will repay his / her loan obligation or defaulting on his / her loan. **(14) 3 3**

S. No	House Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Use kNN algorithm to predict the class label for the test example,

X = (House Owner = No, Marital Status = Married, Income = 120K).

Assume k=3 and the distance is L2 norm (Euclidean distance).

(OR)

(b) Consider the following distance matrix: **(14) 3 3**

	P1	P2	P3	P4	P5	P6	P7	P8
P1	0							
P2	8	0						
P3	6	13	0					
P4	11	2	11	0				
P5	7	20	4	18	0			
P6	10	4	5	5	20	0		
P7	9	11	12	14	13	19	0	
P8	11	13	12	21	17	21	25	0

Use the hierarchical approach to cluster the data points. Use the complete link method to calculate the distances.

In each step

- Show the corresponding distance matrix.
- In the distance matrix, circle the entry whose clusters in the corresponding row and column are to merged.

- Draw the corresponding dendogram for the each clustering.

14. (a) Consider a logistic regression model with two predictors with $\beta_0 = -25.9382$, $\beta_1=0.1109$, $\beta_2=0.9638$, where β_1 and β_2 are for the "Income " and "Lot_Size variables respectively. Using the model with probability cutoff = 0.75, classify the following 6 customers as "Owner" or "Nonowner": if $p \geq 0.75$ then the case as a "Owner". **(14) 4 2**

Customer #	Income	Lot_Size
1	60	18.4
2	64.8	21.6
3	84	17.6
4	59.4	16
5	108	17.6
6	75	19.6

Present the results in a classification matrix.

(OR)

(b) **(i)** Discuss the support vector machine (SVM) algorithm and its use in finding the optimal hyperplane. **(7) 4 2**

(ii) Describe how SVM handles linear and non-linear classification tasks using kernel functions and the kernel trick. **(7) 4 2**

15. (a) You are tasked by Michigan Medicine, a retail company, to build a decision tree model to predict which patients are high-risk for developing heart disease. **(14) 5 2**

Michigan has provided you with a dataset containing the following attributes and records:

S. No	Blood Pressure	Physical Activity	Cholesterol Level	Heart Disease
1	Low	Medium	Medium	NO
2	High	Medium	High	YES
3	Medium	Low	Medium	YES

4	Low	Medium	Low	NO
5	High	Low	Medium	YES
6	High	High	Medium	NO
7	Medium	High	Low	NO
8	Medium	Medium	High	YES
9	Low	High	Low	NO

Build a decision tree model to predict the 'Heart Disease' variable based on the other attributes.

Visualize the decision tree to interpret the rules learned by the model.

(OR)

(b)	**(i)**	List the factors influencing the choice of ensemble learning methods in different scenarios and explain them in detail.	**(7)**	**5**	**2**
	(ii)	Discuss how bagging and boosting address different sources of errors in classification tasks and their impact on model variance and bias.	**(7)**	**5**	**2**

PART- C (1x 10=10 Marks)

(Q.No.16 is compulsory)

			Marks	CO	RBT LEVEL
16.		Consider the five data points P1: (1, 2, 3), P2: (0, 1, 2), P3: (3, 0, 5), P4: (4, 1, 3) and P5: (5, 0, 1). Apply K-means clustering to group these data points into 2 clusters using the Manhattan distance metric and seed centroids C1: (1, 0, 0) and C2: (0, 1, 1). Repeat the iteration of your algorithm until it converges.	**(10)**	**3**	**5**

****