Exam Date & Time: 12-Dec-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

SEVENTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, DEC 2023
DEPARTMENT OF I&CT, MIT, MANIPAL
OPEN ELECTIVE -IV MACHINE LEARNING TOOLS and TECHNOLOGIES-ICT 4304

MACHINE LEARNING TOOLS and TECHNOGIES [ICT4304]

Marks: 50 **Duration: 180 mins.**

A

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) For the data shown in table Q1, apply the K-Means algorithm to compute updated cluster centroids after an iteration. Consider 1st and 5th samples as initial centroids. [Stop updating the clusters when values converge OR Number of iteration is 3 including the first iteration]

A)

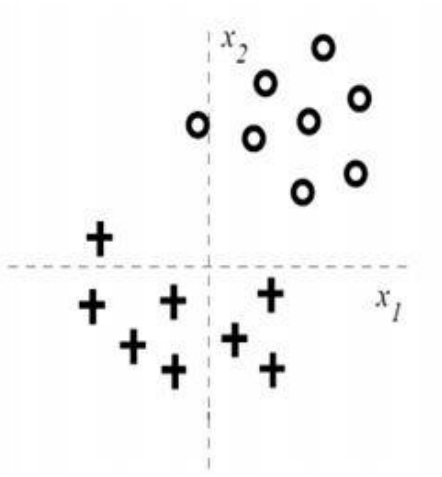
Table Q1

Subject	A	B
1	2.5	0.8
2	3	2.3
3	4	3.5
4	7	5.3
5	5	1.6
6	1.8	2.3
7	2	4.3

(5)

B) What is reinforcement learning? How is learning accomplished in reinforcement learning? Explain the need for reinforcement learning. (3)

C) Consider the data given in Figure Q1C. (2)



FigureQ1C

Apply a logistic regression model for classifying it into two given classes using **logistic regression with L1 regularization.**

$$\sum_{i=1}^{n} \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

Where C is the regularization parameter, and w1 &w2 are the coefficients of x1 and x2. When you increase the value of C from zero to a very large value what will happen to w1 and w2 ? Answer with suitable reasons.

2) **Apply Naïve Bayes to** train the data set given in table Q2A:

Table Q2A

A)

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals

(5)

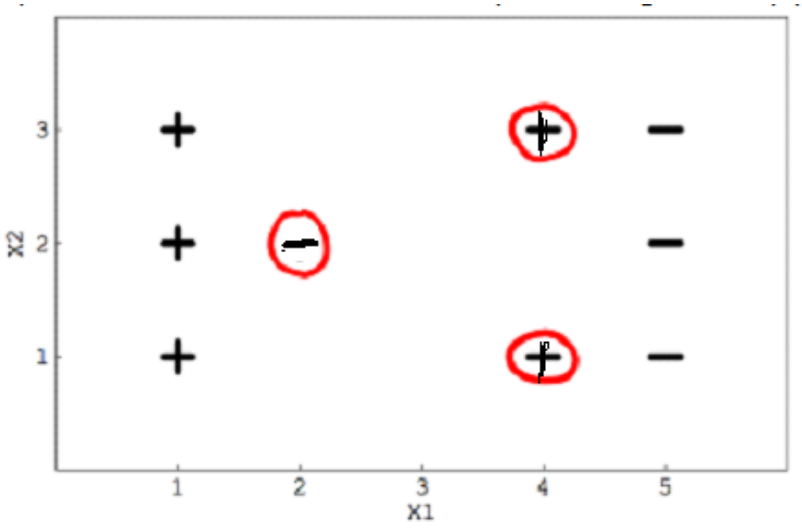
Whether an animal with the following features is a mammal or a non-mammal?

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

B) Assume that you are working with SVM and answer the following Questions:

i. Consider the plot in Figure Q2B. If you remove any one red points from the data shown in the plot, does the decision boundary change? Justify your answer.

ii. In thescenario given in Figure Q2B, we can't have linear hyper-plane between the two classes. So, how does SVM classify these two classes?



(3)

Figure Q2B

C) Consider a simple linear regression model with one independent variable (X). The output variable is 'Y', The equation is: **y= *a* x +*b*,** where **a** is the slope and **b** is the intercept. If we change the (2)

input variable (**x**) by 1 unit, by how much output variable (**y**) will change? Explain with suitable justifications.

3) For the given vector data in the table Q3A, Find the optimal weights to perform the classification using perceptron network. Assume learning rate to be 1 and initial weights to 0.

A) Table Q3A

INPUT				TARGET(t)
X1	X2	X3	X4	
1	1	1	1	1
-1	1	-1	-1	1
1	1	1	-1	-1
1	-1	-1	1	-1

(5)

B) Table Q3B contains the test dataset of 10 records with expected outcomes and a set of predictions from our classification algorithm. Evaluate the model in terms of accuracy, F1 score & Specificity.

Table Q3B

Sl. No.	Expected	Predicted
1	Man	Woman
2	Man,	Man
3	Woman	Woman
4	Man	Man
5	Woman	Man
6	Woman	Woman
7	Woman	Woman
8	Man	Man
9	Man	Woman
10	Woman	Woman

(3)

C) A model is demonstrating high variance across the different training sets. List some specific valid ways which can be used to reduce variance. (2)

4) A research study has been conducted to determine the loss of effect of drug. The data set is given in table Q4A shows the result of the experiment: (5)

A)

Table Q4A

Frequency of prescription	1	2	3	4	5	6
Effect in %	97	84	70	58	53	50

i. Construct the Linear regression model of the effect of drug on frequency of prescription of the drug

ii. According to the Linear model, When will the effect of the drug be 99% ? Also, when does the drug lose its whole activity?

B) Explain the ensemble technique used by random forest? When do you prefer to use random forest over SVM? (3)

C) State whether the following statements are True/False with proper justification:

i. Gradient Descent Methods at all times does not Converge to a Similar Point

ii. The logistic regression classifier performs perfect classification on the data given in figure Q4C. (2)

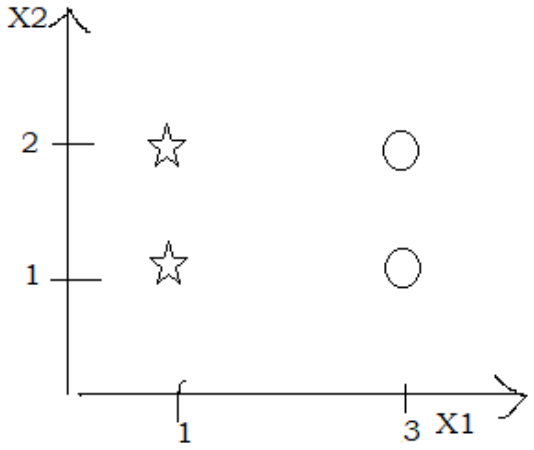


Figure Q4C

5) Cluster the data points given in table Q5A by using hierarchical clustering technique. Illustrate the final hierarchical cluster structure using a dendogram.

A)

Point	A	B
P1	0.07	0.83
P2	0.85	0.14
P3	0.66	0.89
P4	0.49	0.64
P5	0.80	0.46

Table Q5A

(5)

B) The dataset for the credit scoring system is given in table Q5B. A prediction model that is consistent is given in Model Q5B. (3)

Table Q5B

ID	Occupation	Age	Loan-Salary Ratio	Outcome
1	industrial	39	3.40	default
2	industrial	22	4.02	default
3	professional	30	2.7 0	repay
4	professional	27	3.32	default
5	professional	40	2.04	repay
6	professional	50	6.95	default
7	industrial	27	3.00	repay
8	industrial	33	2.60	repay
9	industrial	30	4.5 0	default
10	professional	45	2.78	repay

Model Q5B :

If Age = 50 then

Outcome = Default

Else if Age = 39 then

Outcome = Default

Else if Age =30 and Designation = Senior then

Outcome = Default

Else if Age =27 and Designation = Junior then

Outcome = Default

Else

Outcome = Repay

i. Does the given model generalize correctly to the data instances not contained in the dataset?

ii. Whether the model is overfitted or underfitted? Explain.

C) Mention the uses of decision tree. “Employing 50 smaller decision trees more advantageous compared to using a single large decision tree” Comment on the statement with suitable justification. (2)

-----End-----