Exam Date & Time: 04-Dec-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

FIFTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, NOV/DEC 2023

MACHINE LEARNING [CSE 3171]

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

A

Answer all the questions.

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

1) Evaluate the advantages and limitations of three different forms of machine learning. Provide specific examples of situations where each form of learning excels or faces challenges. (4)

A)

B) Analyze the role of well-posed machine learning problems in the development of intelligent machines. Provide examples to illustrate the importance of problem formulation in machine learning. (3)

C) Compare and contrast the ensemble learning techniques of Bagging and Boosting. Explain the key differences in their approaches, and how they contribute to improving the performance of individual base learners. Provide examples to illustrate their applications. (3)

2) Given the training data listed in Table 4. Apply K nearest neighbour classifier to predict the diabetic patient with the given features BMI, Age. Assume K=3, Test Example BMI=43.6, Age=40, Sugar=? (4)

A)

Table 4

BMI	Age	Sugar
33.6	50	1
26.6	30	O
23.4	40	O
43.1	67	O
35.3	23	1
35.9	67	1
36.7	45	1
25.7	46	O
23.3	29	O
31	56	1

B) Given the training data listed in Table 5. Predict the class of the following new example using Naïve Bayes Clssification: Refund=No, Marital Status=Married, Taxable Income=120k.

Table 5

Tid	Refund	Marital Status	Taxable Income	Evade
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

(3)

C) For the following Figure 1 Bayesian Belief network (3)

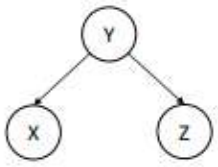

Fig. 1

We know that X and Z are not guaranteed to be independent if the value of Y is unknown. This means that, depending on the probabilities, X and Z can be independent or dependent if the value of Y is unknown. Construct probabilities where X and Z are independent if the value of Y is unknown, and show that they are indeed independent.

3)

Let $\{x_1, \ldots, x_n\}$ be our data set and let $y_i \in \{1,-1\}$ be the class label of x_i So,

A) $y_i \cdot (w^T x_i + b) \geq 1, \forall (x_i, y_i)$. The decision boundary should be as far away from the data of both classes as possible. Show that margin $m = \frac{2}{\|w\|}$. (4)

B) A random sample of eight drivers insured with a company and having similar auto insurance policies was selected. The following Table 8 lists their driving experiences (in years) and monthly auto insurance premiums.

Table 8

Driving Experience (years)	Monthly Auto Insurance Premium
5	$64
2	87
12	50
9	71
15	44
6	56
25	42
16	60

(3)

Find the regression line by choosing appropriate dependent and independent variables. Predict the monthly auto insurance premium for a driver with 10 years of driving experience.

C) Provide examples of real-world applications where specific kernels in SVMs have been particularly effective. Discuss the characteristics of the data that make those kernels suitable. (3)

4) Suppose the visitors to a website need to be grouped using their age as follows: {15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65}. Considering initial clusters as 2, and centroids as c1 =16, c2 = 22, derive the grouping using K-Means algorithm. Furnish all the results in the table format. Show the final listing comprising of iteration number and centroids. (4)

A)

B) Consider applying Expectation Maximization algorithm for training a Gaussian Mixture Model (GMM) to cluster the data. What are the parameters of the Gaussian Mixture Model (GMM)? How are these parameters trained? Write the steps. (3)

C) While evaluating the performance of a classifier, why accuracy may not be a good indicator? Explain with an example. Generate confusion matrix for the following two-class classification problem for predicting whether the patient had cancer and identify all the rows and columns, and arrive at the accuracy of the model.

Cases where the model correctly predicted that the patient had cancer =37

Cases where the model incorrectly predicted that the patient had cancer =3

Case where the model incorrectly predicted that the patient did not have cancer =1

Cases where the model correctly predicted that the patient did not have cancer = 73

(3)

5) Build decision tree for the following training data listed in Table 13 using ID3 algorithm. Show each step of the computation and furnish information gain of attributes in the table form.

A)

Table 13

GPA	Studied	Location	Passes
high	false	hostel	No
high	false	stayOut	No
medium	false	hostel	No
low	true	hostel	Yes
medium	true	stayOut	Yes

(4)

B) Using depth of a decision tree, explain why do underfitting and overfitting occur in decision trees? (3)

C) Why decision tree is considered as a nonlinear classifier? Explain with an example data. In comparison to SVM as a nonlinear classifier, give a scenario where decision tree is expected to have lower performance than SVM. (3)

-----End-----