Question Paper

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

MANIPAL ACADEMY OF HIGHER EDUCATION

SEVENTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, NOV-DEC 2023

MACHINE LEARNING FOR DATA ANALYTICS [ICT 4056]

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

A

Answer all the questions.

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

1) Table Q.1.A shows the expected targets for a small sample test set and a set of predictions made by a model trained for this prediction. (5)

A) i) Construct a confusion matrix for the predictive model.

ii) Calculate the misclassification rate and classification accuracy

Table Q.1.A

ID	Target	Pred.	ID	Target	Pred.
1	spam	ham	11	ham	ham
2	spam	ham	12	spam	ham
3	ham	ham	13	ham	ham
4	spam	spam	14	ham	ham
5	ham	ham	15	ham	ham
6	spam	spam	16	ham	ham
7	ham	ham	17	ham	spam
8	spam	spam	18	spam	spam
9	spam	spam	19	ham	ham
10	spam	spam	20	ham	spam

B) Table Q.1.B. lists a dataset of the scores students achieved on an exam described in terms of whether the student studied for the exam (STUDIED) and the energy level of the lecturer when grading the student's exam (ENERGY). Which of the two descriptive features should we use as the testing criterion at the root node of a decision tree to predict students' scores? (3)

Table Q.1.B.

ID	STUDIED	ENERGY	SCORE
1	yes	tired	65
2	no	alert	20
3	yes	alert	90
4	yes	tired	70

4	yes	tired	70
5	no	tired	40
6	yes	alert	85
7	no	tired	35

C) What is overfitting? How can the use of Regularization eliminate overfitting? (2)

2) Table Q.2.A gives details of symptoms that patients presented and whether they were suffering from meningitis. Using this dataset calculate the following probabilities for the predictive analysis. (5)

A)

i.) P (VOMITING = TRUE)

ii.) P (HEADACHE = FALSE)

iii.) P (HEADACHE = TRUE, VOMITING = FALSE)

iv.) P (VOMITING = FALSE | HEADACHE = TRUE)

v.) P (MENINGITIS | FEVER = TRUE, VOMITING = FALSE)

Table Q.2.A

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	TRUE	TRUE	FALSE	FALSE
2	FALSE	TRUE	FALSE	FALSE
3	TRUE	FALSE	TRUE	FALSE
4	TRUE	FALSE	TRUE	FALSE
5	FALSE	TRUE	FALSE	TRUE
6	TRUE	FALSE	TRUE	FALSE
7	TRUE	FALSE	TRUE	FALSE
8	TRUE	FALSE	TRUE	TRUE
9	FALSE	TRUE	FALSE	FALSE
10	TRUE	FALSE	TRUE	TRUE

B) Instead of explicitly handling problems like noise within the data in an analytical-based table (ABT), some data preparation techniques change the way data is represented just to make it more compatible with certain machine learning algorithms. How the data preparation methods like Binning and Normalization help in this process? Explain. (3)

C) Compare and contrast the use of perceptron with that of linear regression and logistic regression. (2)

3) You have been hired by the Space Agency to build a model that predicts the amount of oxygen that an astronaut consumes when performing five minutes of intense physical work. The descriptive features for the model will be the age of the astronaut and their average heart rate throughout the work. (5)

A)

The regression model is OXYCON = w[0] + w[1] × AGE + w[2] × HEARTRATE

Table Q.3.A shows a historical dataset that has been collected for this task.

Table Q.3.A.

ID	OXYCON	AGE	HEART RATE
1	37.99	41	138
2	47.34	42	153

3	44.38	37	151
4	28.17	46	133
5	27.07	48	126
6	37.85	44	145
7	44.72	43	158
8	36.42	46	143
9	31.21	37	138
10	54.85	38	158
11	39.84	43	143
12	30.83	43	138

i.) Assuming that the current weights in a multivariate linear regression model are w[0] = –59.50, w[1] = –0.15, and w[2] = 0.60, make a prediction for each training instance using this model.

ii.) Calculate the sum of squared errors for the set of predictions generated in part (a).

iii.) Assuming a learning rate of 0.000002, calculate the weights at the next iteration of the gradient descent algorithm.

B) Data analysts believe that in statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. Justify the statement with the concept of the Random Forest algorithm. (3)

C) Why the data augmentation methods are used while handling the non-linear relationships between the feature vctors instead of transforming the data models that were already built? (2)

4) For the following < x,y> pairs shown in Table Q.4.B., simulate the k-means algorithm and Gaussian Mixture Models learning algorithm to identify TWO clusters in the data. Suppose you are given the initial assignment cluster center as {cluster1: #1}, {cluster2: #10} (5)

A) - the first data point is used as the first cluster centre and the 10th as the second cluster centre. If the k-means (k=2) algorithm for ONE iteration is to be simulated. What are the cluster assignments for each of the data points after ONE iteration?

Table Q.4.B.

Data #	x	y
1	1.90	0.97
2	1.76	0.84
3	2.32	1.63
4	2.31	2.09
5	1.14	2.11
6	5.02	3.02
7	5.74	3.84
8	2.25	3.47
9	4.71	3.60
10	3.17	4.96

B) Compare and contrast the following model performance evaluation methods. (3)

i.) k-Fold Cross Validation

ii.) Leave-one-out Cross Validation

iii.) Bootstrapping

C) Use an appropriate data visualization method to represent the data shown in Table Q.4.B. (2)

5) Table Q.5.A. lists a dataset that was used to create a nearest neighbour model that predicts whether it will be a good day to go surfing. (5)

A) Table Q.5.A.

ID	WAVE Size (Ft)	Wave Period (Secs)	Wind Speed (mph)	Good Surf
1	6	15	5	yes
2	1	6	9	no
3	7	10	4	yes
4	7	12	3	yes
5	2	2	10	no
6	10	2	20	no

Assuming that the model uses Euclidean distance to find the nearest neighbour, what prediction will the model return for each of the following query instances?

ID	WAVE Size (Ft)	Wave Period (Secs)	Wind Speed (mph)	Good Surf
Query1	8	15	2	?
Query2	8	2	18	?

B) The basic structure of the multivariable linear regression model allows for only continuous descriptive features, with an example demonstrating how the algorithm can be used in handling categorical descriptive features. (3)

C) Describe the support vector machine algorithm. Use the data from Table Q.4.B to demonstrate the use of SVM. (2)

-----End-----