VIT

Vellore Institute of Technology

Continuous Assessment Test I – May 2023

Programme	:	M.Tech. (Integrated) Computer Science and Engineering with Specialization in Business Analytics	Semester	:	Fall Inter Sem 2022-2023
Course Title	:	Machine Learning	Code Class Nbr	: :	CSE4036 CH2022232500903 CH2022232500902
Faculty	:	Dr.A.Vinothini Dr.A.Kaja Mohideen	Slot	:	D1
Time	:	90 minutes	Max. Marks	:	50

Answer all the Questions

1. (i)Consider the learning problem of the prediction of COVID. Identify and explain in detail the Task, Experience, and Performance measures of the problem. (3 marks).

(ii) Mr. Ram, a bioinformatics analyst collects a huge collection of DNA sequences of various species. He is given the task of classifying the DNA sequence according to their families. The analyst does not have any idea about the dependent variable as the DNA sequence data are not labeled. Identify and explain the type of learning paradigm that can be applied to group the DNA sequence. List any two machine learning algorithms that can be used for implementing the scenario (3 marks).

(iii) Consider the problem of classifying students as pass or fail based on the training samples {CAT_marks, assignment_marks}. c is the target concept and h is the hypothesis. The errors generated by hypothesis h with respect to independent features for the training sample of size 10 are shown in Table 1. Consider the upper bound on an error as 0.05 and the probability of failure in achieving the accuracy as 0.20. Find if h is Probably Approximately Correct or not. Justify your answer (4 marks). 10

Table 1

Sample Number	Error(h)
1	0.032
2	0.064
3	0.023
4	0.042
5	0.040
6	0.021
7	0.061
8	0.034
9	0.050
10	0.072

2. Ram spends one week of summer holidays for fishing in his nearby pond. Table 2 shows the hours spent fishing and the number of fish caught. Ram wishes to analyze the relationship between hours spent fishing and the number of fish caught. Design a regression model by choosing the appropriate dependent and independent variables. Calculate the R-squared value and write your interpretation. 10

Table 2

Hours spend	Number of fishes caught
2	4
3	5
5	7
7	10
9	15

3. The marketing manager of OPPO mobiles wishes to mail regular updates about their new products to customers residing in Chennai. The manager wishes to identify the customers who may buy the new mobile phones. Consider the details of the customers given in Table 3. Customer Type holds the value according to the number of products purchased (Type1>5 products, Type2-1 to 5 products, Type3=1 product). Construct the machine learning model using a decision tree algorithm to predict the customer purchase behavior for new mobile phones. Illustrate the decision tree constructed. Predict the class for the test instance: 15
(Customer_Type="Type3",Income="low",Recently_purchased="no",
Employment="yes")

Table 3

Customer _Type	Income	Recently purchased?	Employment	buy_mobile
Type 1	high	Yes	no	no
Type 1	high	No	no	no
Type 2	high	Yes	no	yes
Type 3	medium	Yes	no	yes
Type 3	low	Yes	yes	yes
Type 3	low	No	yes	no
Type 2	low	No	yes	yes
Type 1	medium	Yes	no	no
Type 1	low	Yes	yes	yes
Type 3	medium	Yes	yes	yes
Type 1	medium	No	yes	yes
Type 2	medium	No	no	yes
Type 2	high	Yes	yes	yes
Type 3	medium	No	no	no

4. (i)A farmer wishes to learn the likeliness of flower blooming based on the height and age of the plant. Table 4 shows the training examples with attributes: height of the plant, age of the plant, and bloom.
Apply an instance-based learning algorithm and create a machine-learning model to predict the flower blooming likeliness for Height=20 and Age =35. (10 marks)

Table 4

Height (cm)	Age(days)	Flowe_bloom
30	10	Yes
40	30	No
50	80	No
10	35	Yes
60	50	No
50	10	Yes
15	50	No

15

(ii) Compute the performance of the flower blooming machine learning model using the measures: Accuracy, Precision, Recall, and F1 Score. Give your inference. (5 marks)

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Continuous Assessment Test II– July 2023

Programme	:	**M.Tech. (Integrated) Computer Science and Engineering with Specialization in Business Analytics**	Semester	:	**Fall Inter Sem 2022-2023**
Course Title	:	**Machine Learning**	Code	:	**CSE4036**
			Class Nbr	:	**CH2022232500903 CH2022232500902**
Faculty	:	**Dr.A.Vinothini Dr.A.Kaja Mohideen**	Slot	:	**D1**
Time	:	**90 minutes**	Max. Marks	:	**50**

Answer all the Questions

1. Mr. Ram is a geologist who is involved in a research project on grouping similar types of rocks. Most rocks can be characterized by their unique physical properties such as hardness and specific gravity as shown in Table 1.

(i) (i) Help Mr. Ram to group the rocks by applying an appropriate partition-based unsupervised machine learning algorithm. The number of clusters to be formed is 2. Elaborate on each step of the algorithm in detail. (12 marks)

(ii)Evaluate the performance of your model with any two suitable performance metrics. (3 marks) [15]

Table 1

Rock ID	Hardness	Specific Gravity
R1	2.1	1.8
R2	3.2	2.3
R3	5.1	3.7
R4	4.0	1.6
R5	1.2	2.2
R6	6.1	3.1

2. The government wishes to build public health centres near villages. The location of the places L1 to L6 is mentioned as two-dimensional data points. L1(2,8), L2(6,3), L3(5,8), L4(9,6), L5(8,3), and L6(3,4) . Assume the maximum radius of the neighborhood is 2.0 and with a minimum of 3 neighbors.

(i) Apply an appropriate density-based clustering technique to form clusters using the given places. Elaborate on each step of the algorithm in detail. (8marks) [15]

(ii)Identify the core locations, border locations, and outliers. Justify your answer. (5 marks).

(iii)Is it possible to apply hierarchical clustering for the given scenario? What are the disadvantages of hierarchical techniques when compared with density-based techniques? (2 marks)

3. As a data scientist, you build a decision tree model for classifying cooking oil into various quality levels (High, Medium, and Low). While constructing the model, you meet the following situations. How do you handle them? [10]

(i)You are excited to receive the training error 0.00. However, the validation error is 45.32. What's wrong with the model? How can you solve this using machine learning

techniques? (4 marks)

(ii)Assume that your model has high bias and low variance. What are the characteristics of your model? (3 marks)

(iii)How will you apply the bagging technique to the existing model? Elaborate your idea with a proper illustration. (3 marks)

4. Consider the problem of predicting tumors in DNA samples. The training data of DNA samples with their features A1, and A2 and class label Y (Yes, No) is shown in Table 2. The class label "Yes" refers to Tumor and "No" refers to Healthy. The data is trained by a weak classifier M. P refers to the predictions of M. Find and apply an appropriate boosting technique to improve the performance of M. Elaborate the algorithm in detail with necessary steps. [10]

Table 2

Sample ID	A1	A2	Y	P
1	88	73	Yes	Yes
2	72	54	Yes	Yes
3	54	55	Yes	No
4	34	45	Yes	No
5	87	92	Yes	No
6	92	67	No	Yes
7	43	87	No	No
8	56	51	No	No
9	44	67	No	No
10	62	23	No	No

⇔⇔⇔