VIT
Vellore Institute of Technology
CHENNAI

Reg. Number: 22BPS1181

Continuous Assessment Test (CAT) – I AUGUST 2025

Programme	:	B.Tech CSE (AI & ML Specialization)	Semester	:	Fall 2025-26
Course Code & Course Title	:	BCSE209L Machine Learning	Class Number	:	CH2025260102134 CH2025260102132 CH2025260102128 CH2025260102130
Faculty	:	Dr. R. Bhargavi, Dr.Rajalakshmi R, Dr. S A Sajidha, Dr. Syed Ibrahim SP	Slot	:	A1 + TA1
Duration	:	1 Hour 30 minutes	Max. Mark		50

General Instructions:

- Write only your registration number on the question paper in the box provided and do not write other information
- Use statistical tables supplied from the exam cell as necessary
- Use graph sheets supplied from the exam cell as necessary
- Only non-programmable calculator without storage is permitted

Answer all questions

Q. No	Sub Sea.	Description	Marks
1		Consider the following dataset where x is the independent feature and y is target variable.	15

ID	x	y
1	10	277
2	3	4
3	6	85
4	15	680
5	5	60
6	12	425
7	9	220
8	13	484

The following predicted linear regression model is learned from the training data

$y_{pred1} = -223.35 + (551.1 * x)$.

But, as a domain expert you have a confidence that instead of regressing x on y, regressing x^2 (only x^2 feature without any other feature(s)) on y would result in a better generalised model.

(a) Train a linear regression with x^2 as the independent parameter and y as the target feature. Call this model as y_{pred2}(8M)

(b) Predict the y value for the test input x = 17 for both the models(i.e y_{pred1} and y_{pred2}) (2M)

(c) Identify which model is better using the performance metric MAE. (5M)

RL

2	A	A medical doctor wants to develop a machine learning system to analyze and predict which patients are more likely to be readmitted to the hospital in the near future (re-admit/No re-admit). Would this problem be considered a supervised, or unsupervised learning task? Explain your choice based on the nature of the data and the objective. (3M)	15
	B	With an appropriate neat diagram, give a detailed roadmap for building this Machine learning problem. Explain the process. (8M)	
	C	How is a classification problem different from regression problem? Given two examples for each with input and output features. (4M)	
3	A	You are a data scientist building a logistic regression model to predict the probability of a customer signing up for a premium subscription service. You have collected a wide range of features, including customer demographics and usage habits. After training your model, you examine the coefficients to understand which features are driving the predictions. You notice a surprising result: * The coefficient for monthly_call_minutes is a large, positive value ($w_1 = +0.65$). * The coefficient for monthly_data_usage_GB is a large, negative value ($w_2 = -0.42$). Your business intuition suggests that both high call minutes and high data usage should be strong positive indicators of a customer who is a heavy user and therefore more likely to sign up for a premium plan. The negative coefficient for data usage seems illogical. You also know that monthly_call_minutes and monthly_data_usage_GB are highly correlated; customers who use a lot of one service also tend to use a lot of the other. The Question: (a) Explain, why the logistic regression model produced these illogical and contradictory coefficients, despite your strong business intuition. (2M) (b) Propose two distinct methods to address this issue and make the coefficients more stable and interpretable. Explain the core mechanism of each method. (4M)	10
	B	You are a data scientist building a multiple linear regression model to predict housing prices. You have a large dataset with 30 potential features (e.g., square footage, number of bedrooms, lot size, age of the house, etc.). After building an initial model with all 30 features, you observe a high R-squared value of 0.95. However, when you	

calculate the Adjusted R-squared, you find it is significantly lower, at 0.75.

1. Explain the fundamental difference between R-squared and Adjusted R-squared, and why the Adjusted R-squared is a more reliable metric for model selection in this scenario. (2M)
2. What is the most likely reason for the large discrepancy between these two metrics? (2M)

4

Assume, you collected the following regular and spam mails to train the classifier, and only three words are informative for this classification, i.e., each email is represented as a 3 dimensional binary vector whose components indicate whether the respective word is contained in the email.

'study'	'free'	'money'	Category
1	0	0	Regular
0	0	1	Regular
1	0	0	Regular
1	1	0	Regular
0	1	0	Spam
0	1	0	Spam
0	1	0	Spam
0	1	0	Spam
0	1	1	Spam
0	1	1	Spam
0	1	1	Spam
0	1	1	Spam

Now, assume that you have a test mail with the features study: 1, free:0, money:1. Predict whether the test mail is spam or Regular.

10

**********All the best **********