Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2023-2024

Mid-Semester Test
(EC-2 Regular)

Course No.	: AIML ZG565
Course Title	: MACHINE LEARNING
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	:

No. of Pages = 3
No. of Questions = 5

Note:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1. a) Tom Mitchell defined ML as, "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E".

Answer the following question with respect to the above definition: [2+1=3]

1) Suppose you provide a learning algorithm with a lot of historical weather data to make a ML model that predicts whether the weather will be Sunny, Windy or Rainy for the given input parameters. In this setting, what is E, T, and P? Name some performance measures, P for the given problem (name any 2).

Answer

Explanation
T := The weather prediction task. Classification task
P := The probability of it correctly predicting a future date's weather/ number of new instances that are correctly classified.
E := The process of the algorithm examining a large amount of historical weather data/ historical weather data

[1.5 marks for correctly identifying all 3 parameters, 0.5 marks each]

P could be Accuracy, Precision, Recall, F score, AUC_ROC, Misclassification rate etc.

[0.25 marks each for correctly identifying classification evaluation metrics]

2) The quantity of rainfall in a day is typically measured in either millimeters (mm) or inches. If you utilize a learning algorithm to forecast tomorrow's rainfall, which learning technique would be suitable? Justify your selection. What would be a suitable choice for P? Name some performance measures P for the given problem (any 2 measures).

Regression **[0.25 marks]**, target variable is continuous **[0.25 marks for justification]**

P could be MSE, R-square **[0.25 marks for each]**

Q1.b) A Google machine learning system beat a human Go player in 2015. This complicated game may contain more playable options than there are universe atoms. The first version won by observing hundreds of thousands of hours of human game play; the second version learnt by getting rewards while playing against itself. How would you describe this transition to different machine learning approaches? **[2]**

Ans - The system went from supervised learning to reinforcement learning. The first version is supervised learning**[1 marks]** ***and second is reinforcement learning*** **[1 marks]**

Q2. You are working in the Computer Science Department of College of Engineering of a reputed national university. Over the past few years, there has been a noticeable decline in student CGPA in the B.Tech program, leading to a loss of students to other universities. In response, your department head has tasked you with conducting an analysis of the data to propose corrective actions to retain or increase student intake. Upon acquiring a small data sample, you notice that are some issues with the dataset. Identify at least 5 potential issues with this dataset and suggest how to resolve these issues (python code is not required). [3]

Name	Age	Date of Birth	Course ID	CGPA
Aishwarya	24	01-Jan-1995	CS104	7.4
Bhargav	23	Dec-01-1996	CS102	7.5
Chandra	25	01-Nov-1994		6.7
Divya	24	Oct-01-1995	CS104	7.9
Bhargav	23	Dec-01-1996	CS102	8.1
Eshan	24	01-Jul-1995	CS103	87.5
Francis	54	01-01-1959	CS105	7.0

Solution:

1. Missing data in Course ID column.
2. Redundant data - Age and DOB columns are redundant. Age can be calculated from DOB.
3. Inconsistent data - Date of Birth format is not consistent across all the records.
4. Duplicate entries - Bhargav has two records with different CGPA.
5. Outlier in Age column – Francis with Age 54 is a student.

[0.25 marks for each identified issues + 0.25 marks for suggesting techniques for each issue to resolve it]

Q3) Answer the following questions with respect to Logistic Regression: (1 + 2 + 3 = 6 marks)

a) Suppose we want to apply regularization techniques to avoid model overfitting such that, we want to penalize the large coefficients to reduce their impact but not settling them to 0. Which regularization technique will be more appropriate for this situation?
b) Are the model evaluation metrics like Precision, Recall, Sensitivity, Specificity affected by changing the decision threshold value for decision boundary? Justify your answer with an example by showing the calculation for any one of the above metrics.
c) In logistic regression, explain how the coefficients obtained from logistic regression relate to the log-odds of the target variable. Provide a mathematical justification for your explanation.

Answers:

a) Ridge Regularization (1 mark)

b) Yes. [0.5 marks] Calculation of any one of precision, recall, sensitivity, specificity is expected as per the formulae: (1.5 marks)

c) The coefficients in logistic regression represent the change in the **log-odds** of the outcome variable **per unit increase** in the predictor variable. (1 mark)

[2 marks for mathematical justification]

$p=\sigma(z) = 1/1+e^{-(\theta^T X)}$

$\text{logit}(p) = \log(p/1-p) = \theta^T X$

Q4)

Given the following confusion matrix for a binary classification problem: (6 marks)

	Predicted Positive	Predicted Negative
Actual Positive	350 (TP)	150 (FN)
Actual Negative	200 (FP)	800 (TN)

a) Calculate the precision, recall, and F1-score for the negative class based on the confusion matrix provided. Show your calculations step-by-step. [1.5]
b) Based on the calculated metrics (precision, recall, and F1-score), discuss which metric (precision, recall, or F1-score) would be most informative in evaluating the model's performance in this scenario with respect to negative class. Justify your stance. [1.5]
c) Give an example of a scenario where optimizing Precision is more important than optimizing Recall. Provide explanation to support your stance. [1.5]
d) Give an example of a scenario where optimizing Recall is more critical than optimizing Precision. Provide explanation to support your stance [1.5]

Answers:

a) Precision for negative class = TN / (TN + FN) (1 mark)
Recall for negative class = TN / (TN + FP) (1 mark)
F1-score for negative class = 2PR/(P+R) (1 mark)
<approximate answers are also acceptable>

b) In this scenario, the **F1-score** would be the most informative metric (0.5 marks) because it provides a balance between precision and recall. It indicates how well the model balances precision and recall. (1 mark)

c) Precision is more critical in scenarios where **minimizing false positives** is crucial. (marks to be awarded for any relevant example) (0.5 marks for example and 1 mark for proper justification)

d) Recall is important where minimizing **false negatives is crucial** (marks to be awarded for any relevant example) (0.5 marks for example and 1 mark for proper justification)

Q5. An educational institution wants to predict student performance based on various factors. They need a decision tree model to classify student performance as 'Excellent', 'Good', or 'Poor' based on specific features. Considering the following data set use the ID3 decision tree algorithm to determine the optimal root node feature among {Exam Preparation Time, Attendance Rate, Participation in Extracurricular

Activities} to predict student performance as 'Excellent', 'Good', or 'Poor'. Show all calculations. [5 marks]

Exam Preparation Time	Attendance Rate	Participation in Extracurricular Activities	Performance
Low	High	Yes	Excellent
Medium	Medium	No	Good
Low	Low	Yes	Poor
High	Medium	Yes	Excellent
Medium	High	No	Good
Low	Medium	Yes	Poor
High	High	Yes	Excellent
Medium	Medium	Yes	Good
Low	Low	No	Poor
High	High	No	Excellent
Medium	High	Yes	Good
Low	Medium	No	Poor
High	High	Yes	Excellent
Medium	Medium	Yes	Good
Low	Low	No	Poor

Answer:

 Calculate Entropy of the Target Variable (Performance):

The target variable has three categories: Excellent, Good, and Poor. Let's calculate the entropy.

- Total instances: 15
- Excellent: 5
- Good: 5
- Poor: 5

p(Excellent)=5/15=0.3333,
p(Good)=5/15=0.3333,
p(Poor)=5/15=0.3333

$H(S)=-(\ 0.3333\log_2(0.3333)+0.3333\log_2(0.3333)+\ 0.3333\log_2(0.3333))$

 Calculate Information Gain for Each Feature:

Feature: Exam Preparation Time

Entropy =6/15*(-(1/6)*LOG(1/6,2)-(5/6)*LOG(5/6,2)) = 0.260

Information Gain:

IG(Exam Preparation Time)=H(S)−H(Exam Preparation Time)=1.58−0.260

Feature: Participation in Extracurricular Activities and Attendance Rate

Similarly, calculate entropy for these 2 features

 Select **Feature with Highest Information Gain**:

The feature "Exam Preparation Time" has the highest information gain, so it will be the root node of the decision tree.

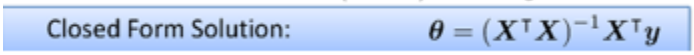
Evaluation:

1.5 marks each for entropy/ Information gain calculation

0.5 marks for correct root selection

*Minimizing entropy is equivalent to maximizing Information Gain. Select the feature for which entropy is the minimum, or IG is the highest.

Q6. a) Suppose you have m = 28 training examples with n = 4 features (excluding the additional all-ones feature for the intercept term, which you should add). The closed form solution is



For the given values of m and n, what are the dimensions of θ, X, and y in this equation? [2]

Answer: X is 28×5 **[1 marks]**, y is 28×1 **[0.5 marks]**, θ is 5×1 **[0.5 marks]**

If any of the dimension coordinates is wrong, 0 marks will be awarded for that parameter.

b) Suppose you have been asked to train a linear regression model using Gradient descent approach with MSE as the cost function. You run gradient descent for 15 iterations with learning rate, α=0.3 and compute J(θ) after each iteration. You find that the value of J(θ) increases over time. Based on this, what conclusion will you derive? Is it necessary to change the learning rate α? If so, should it be increased or decreased? Justify your answer. [2]

1. **Answer:** Learning rate is high, so it should be decreased.. **[0.5]**

   Rather than use the current value of α, it'd be more promising to try a smaller value of α (say α=0.1). **[1.5 marks for justification]**