Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2024-2025

Mid-Semester Test
(EC-2 Make-up)

Course No.	: AIML ZG565/ DSE 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 = 6

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. 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+2=4]**

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
[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]**

Q2. For each of the following statements, does it justify the use of feature scaling in machine learning algorithms? Answer with "Yes" or "No" and provide your reasoning. Marks will be awarded based on the reasoning provided. [6]

- It guarantees that the algorithm converges at global minimum solution
- It prevents the matrix X^TX (used in the normal equation) from being non-invertable (singular/degenerate).
- It is necessary to prevent gradient descent from getting stuck in local optima.
- It speeds up gradient descent by making it require fewer iterations to get to a good solution.

True or False	Statement	Explanation
False	It guarantees that the algorithm converges at global minimum solution	It **does not guarantee convergence to the global minimum**.Many algorithms (e.g., neural networks, logistic regression with non-convex regularization) may have **local minima** or **saddle points**.
False	It prevents the matrix X^TX (used in the normal equation) from being non-invertable (singular/degenerate).	none
False	It is necessary to prevent gradient descent from getting stuck in local optima.	The cost function $J(\theta)$ for linear regression has no local optima.
True	It speeds up gradient descent by making it require fewer iterations to get to a good solution.	Feature scaling speeds up gradient descent by avoiding many extra iterations that are required when one or more features take on much larger values than the rest.

[0.5 marks for True/False, 1 mark for the right explanation]

Q3.You are working as a data scientist for a real estate company to predict apartment prices. You have a dataset of 500 apartments with 15 features (area, bedrooms, bathrooms, floor, age,

distance to metro, etc.). You decide to experiment with different regression models and regularization techniques. [6M]

- Models tested:
  - Model A: Linear regression with 3 features (area, bedrooms, age)
  - Model B: Polynomial regression degree 8 with all 15 features (120 total features after polynomial expansion)
  - Model C: Polynomial regression degree 8 with Ridge regularization ($\lambda = 100$)
  - Model D: Polynomial regression degree 8 with Lasso regularization ($\lambda = 50$)

  Results after training:

Model	Training RMSE	Test RMSE	R^2 (Test)	Features Used
A	$45,000	$47,000	0.66	3
B	$12,000	$68,000	0.42	120
C	$28,000	$36,500	0.76	120
D	$31,000	$34,800	0.79	23

A) Analyse Model A and B in terms of bias and variance. Identify which models suffer from under fitting, overfitting, or are well-balanced. Support your analysis with specific numerical evidence from the results

B) Model B shows training error decreasing continuously while test error increases after 5 gradient descent iterations. Explain the underlying reasons for this divergent behaviour. Explain which regularization techniques can be helpful here? (No lasso or Ridge).

C) Explain the role of R^2 metric in the decision making for above 4 models.

D) For the model c which exhibits moderate bias (not low). How the bias can be further reduced.

E) Justify the number of features used by model C and D

F) If you are asked to use only Model A then how will you modify it to give a better performance.

Solution:

A) Model A :High Bias, Low Variance → UNDERFITTING

- High Bias : Training RMSE ($45,000) and test RMSE ($47,000) are both high and similar
- Small gap between train/validation performance (2,000) indicates low variance

Model B : Low Bias, High Variance → SEVERE OVERFITTING

- High Variance: Large gap between training RMSE ($12,000) and test RMSE ($65,000) = $53,000 gap
- Low bias : Low RMSE ($12,000)

B) The model continues to fit the training data better by learning increasingly complex patterns, including noise which is not present in test data. Model is memorizing the training data and fitting it too well. This is overfitting case as we observe decrease in

test error. Early stopping (stopping the training after 5 iterations) can help to control overfitting.

C) R^2 value explains the goodness of fit. It is a statistical measure that represents the proportion of variance in the dependent variable that is explained by the independent variables in a regression model. Model D is having highest R^2 value. 23 features used by model D are the best predictors.

D) By decreasing the value of hyper parameter λ

E) Model C is Ridge Regression using model B which has 120 features. Ridge regression retains all the feature coefficients but reduces their magnitude. Hence, all 120 features are used for prediction.
Model D is Lasso Regression using model B which has 120 features. Lasso regression zeros some of the feature coefficients. The 23 features are used for prediction as other features' coefficients are zero.

F) Model A suffers from high bias. It uses only 3 features so it is a simple model. We can use all 15 features which might help in reducing the bias i.e. decreasing the RMSE

Q4. You are evaluating three different models for a medical diagnosis system that predicts whether a patient has a rare disease (prevalence = 2%). The confusion matrices for 10,000 test samples are:
Model A: TP=180, FP=200, FN=20, TN=9600
Model B: TP=150, FP=50, FN=50, TN=9750
Model C: TP=190, FP=500, FN=10, TN=9300

a) Calculate Precision, Recall, F1-score, and Specificity for each model. Show all calculations. [3 marks]
b) Which model would you choose for this medical application and why? Consider the cost of false negatives vs false positives. [2 marks]

a) Performance Metrics Calculations [3 marks]
Model A: TP=180, FP=200, FN=20, TN=9600
Precision, Recall, F1, Specificity [1 mark]
Model B: TP=150, FP=50, FN=50, TN=9750
Precision, Recall, F1, Specificity [1 mark]
Model C: TP=190, FP=500, FN=10, TN=9300
Precision, Recall, F1, Specificity [1 mark]

b) Model Selection [2 marks]

Model B is the best choice because in medical diagnosis, false negatives (missing disease) are more costly than false positives. Model B has the best balance with high precision (0.750) and reasonable recall (0.750), minimizing both types of errors. [2 marks]

Q5. You are building a binary classification model using logistic regression to predict whether a house will sell above or below the median market price. Your dataset contains 1,000 examples and 15 features, such as area, number of bedrooms, location score, age of the house, etc. After training the model, you observe the following:

- Training Accuracy = 0.94, Validation Accuracy = 0.70
- Training Log-Loss = 0.15, Validation Log-Loss = 0.55
- Some feature weights (coefficients) are extremely large (e.g., $\beta 1=85$, $\beta 2=-72$)
- Small perturbations in the training data lead to large changes in coefficient values

Analyze this scenario and answer:

a) What issue is your logistic regression model likely suffering from? Justify your answer. [1 mark]

b) Explain the mathematical reasoning behind why this issue arises in logistic regression. [1 mark]

c) Recommend two techniques to mitigate this problem. Provide a brief explanation of how each technique works, including relevant mathematical intuition. [2 marks]

a) **Answer: Overfitting [0.5 marks]**
Justification: [0.5 marks]

- **High training accuracy (0.94)** and **low validation accuracy (0.70)** indicate the model performs very well on training data but generalizes poorly.
- **Low training log-loss (0.15)** vs. **high validation log-loss (0.55)** reinforces this performance gap.
- **Extremely large coefficient values (e.g., $\beta_1 = 85$)** and **sensitivity to small data perturbations** further suggest overfitting, where the model is too closely fitting noise or patterns in the training set.

b) Logistic regression overfits when the optimization process pushes weights β\betaβ to large values to minimize log-loss, especially when the data is (nearly) linearly separable, high-dimensional, or noisy. Without regularization, the model becomes too flexible, fitting noise instead of general patterns. **[1 mark only if correct justification is provided]**

c) Regularization (L1, L2), Early stopping, Increase training data, Reduce Model complexity

[1 mark for each approach only if explanation is provided]

Q6. Answer the following questions: **[5 marks]**

During decision tree induction on 2-class data, a node has {12, 8} objects of class 1 and class 2 respectively. After splitting using an attribute B, two child nodes are created with the following distributions:
Left child node: {9, 1} (class1, class2)
Right child node: {3, 7} (class1, class2)

Tasks:
a) Compute the Entropy of the parent node and the child nodes. (2)
b) Calculate the Information Gain of the split using attribute B. (1)
c) What are the limitations of Information Gain, and how can they be addressed? (2)

a) Entropy of parent node

$$P = \{12, 8\}, \quad \text{Total} = 20, \quad p_1 = \frac{12}{20}, \quad p_2 = \frac{8}{20}$$

$$H(P) = -\left(\frac{12}{20}\log_2\frac{12}{20} + \frac{8}{20}\log_2\frac{8}{20}\right) \approx -\left(0.6\log_2 0.6 + 0.4\log_2 0.4\right)$$

$$= -(0.6 \cdot -0.737 + 0.4 \cdot -1.322) = 0.442 + 0.529 = \boxed{0.971 \text{ bits}}$$

b) Information gain of the attribute B

Child Entropy

Left child: $\{9, 1\}$

$$p_1 = 0.9,\ p_2 = 0.1, \quad H(L) = -(0.9\log_2 0.9 + 0.1\log_2 0.1) = -(0.9 \cdot -0.152 + 0.1 \cdot -3.322)$$

$$= 0.137 + 0.332 = \boxed{0.469 \text{ bits}}$$

Right child: $\{3, 7\}$

$$p_1 = 0.3,\ p_2 = 0.7, \quad H(R) = -(0.3\log_2 0.3 + 0.7\log_2 0.7) = -(0.3 \cdot -1.737 + 0.7 \cdot -0.515)$$

$$= 0.521 + 0.361 = \boxed{0.882 \text{ bits}}$$

Weighted Entropy

$$H_{children} = \frac{10}{20} \cdot 0.469 + \frac{10}{20} \cdot 0.882 = 0.5(0.469 + 0.882) = \boxed{0.6755}$$

Information Gain

$$\text{IG} = H(P) - H_{children} = 0.971 - 0.6755 = \boxed{0.2955}$$

c) Entropy-based information gain tends to **favor attributes** with **many unique values** (e.g., ID numbers), which may **not be meaningful** splits. [1 mark]

This is why **Gain Ratio** (used in C4.5) is introduced — it penalizes high-cardinality attributes. [1 mark]