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

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

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). For the given problem statements, identify the most suitable type of learning (Supervised, Unsupervised, or Reinforcement) for each scenario. Justify your answer. [2]

Problem Statement 1: A self-driving car must navigate through urban environments, adjusting speed, avoiding obstacles, and obeying traffic rules while minimizing fuel consumption and travel time. The car improves its performance over time through interactions with the environment and feedback from its driving outcomes.

Problem Statement 2: An online retail platform wants to predict whether a product review is "positive", "negative", or "neutral" using a large set of previously labeled customer reviews. The goal is to automate sentiment analysis for improving customer feedback handling.

Problem Statement 3: A geneticist wants to identify hidden patterns in DNA sequences by analyzing a dataset of unlabeled genetic codes. The goal is to cluster similar genetic structures together to identify potential links to diseases, without any prior classification.

Problem Statement 4: An insurance company wants to predict the amount of insurance claim a customer might make in a year based on their age, driving history, vehicle type, and location. The dataset contains the historical records of customers and their claim amounts.

Solution

- ***Problem Statement 1:*** *Reinforcement Learning*
- ***Problem Statement 2:*** *Supervised Learning- Classification*
- ***Problem Statement 3:*** *Unsupervised Learning – Clustering*
- ***Problem Statement 4:*** *Supervised Learning – Numeric Prediction/ Regression*

[0.25 marks for correctly identifying the learning approach, 0.25 marks for justification]

Q1. b) What are the key implications of overfitting on model performance? How can techniques like regularization and early stopping be employed to address this issue? Provide a detailed explanation, specifically in the context of logistic regression or linear regression, on how these techniques help improve a model's ability to generalize beyond the training data. [4]

Implications of Overfitting:

- **Loss of generalizability**
- **Unreliable predictions**
- **Increased complexity** ***[2 marks only if at least 2 implications are mentioned]***

Mitigation Techniques:

1. Regularization: Regularization adds a penalty term to the loss function that discourages overly complex models by constraining the magnitude of model parameters.

2. Early Stopping: Early stopping monitors the model's performance on a validation set during training. When the validation loss starts increasing while training loss continues to decrease, the model learning is stopped — preventing it from learning noise in the training data.

[1 mark for each only if explanation is provided wrt linear regression]

Q2. You are interning with the Admissions Analytics Team of a reputed national university. The management is concerned about a fall in the average performance of M.Tech students across various specializations. You are assigned the task of auditing the student academic dataset to identify anomalies or inconsistencies. Upon inspecting a sample of the data, you notice several issues that could affect data reliability.
Identify at least 6 potential issues with this dataset and suggest how to resolve them. (Python code is not required.) **[4]**

Student Name	Age	Email	Specialization	Admission Date	CGPA
Rina	22	rina@example.com	AI	15-Aug-2022	8.2
Dev	23	dev[at]mail.com	Cybersecurity	2022-08-15	7.8
Aarav	22	aarav@example.com	Data Science	15/08/2022	9.1
Shweta	22		Cybersecurity	15-Aug-2022	6.4
Rina	22	rina@example.com	Artificial Intelligence	15-Aug-2022	8.2
Mansi	58	mansi123@gmail.com	Data Science	15-Aug-2022	7.0

Solution:

1) *Inconsistent Date Formats: The "Admission Date" column has varying formats like* `15-Aug-2022`, `2022-08-15`, `15/08/2022`, *and* `August 15, 2022`.

   Resolution: Standardize all dates to a single format, e.g., `YYYY-MM-DD` *using a date parser or format conversion during preprocessing.*

2) *Duplicate Records The record for "Rina" appears twice with identical data.*

Resolution: Remove exact duplicate entries using deduplication based on student name, email, and specialization.

3) *Age Outlier: "Mansi" has an age of 58, which is highly unlikely for a regular M.Tech student.*
4) *Missing Email Address*
5) *Incorrect Email Format: "Dev" has an email recorded as `dev[at]mail.com` instead of `dev@mail.com`. Resolution: Use regex or validation rules to detect and correct improperly formatted email addresses.*
6) *Case of Inconsistent Specialization Naming: Specializations like "AI", "Cybersecurity", "Computer Vision", and "Data Science" should be verified for consistent naming (e.g., AI vs Artificial Intelligence). Resolution: Apply a mapping table to unify specialization categories.*

[Full marks if at least 6 issues have been identified and techniques suggested to resolve them, otherwise deduct marks accordingly]

Q3. You are developing a linear regression model for a high-dimensional dataset that contains hundreds of predictor variables, many of which may be redundant or irrelevant. [2+2 = 4]

1) Describe how L1 (Lasso**)** and L2 (Ridge) regularization alter the objective function of linear regression. How does each method contribute to controlling model complexity and reducing overfitting?
2) Which regularization technique would be more appropriate in this scenario?

In linear regression, the goal is to minimize the MSE between predicted and actual values. Regularization modifies this objective by adding a penalty term that discourages large coefficients:

- ***L1 Regularization (Lasso)** adds the absolute value of coefficients, encouraging sparsity (i.e., driving some coefficients exactly to zero). **[0.5 marks for explanation on how it controls complexity and handle overfitting]***

$$\text{Loss} = \text{MSE} + \lambda \sum_{j=1}^{p} |\beta_j|$$

[0.5 marks if mathematical expression is provided]

- ***L2 Regularization (Ridge)** adds the squared magnitude of coefficients, which shrinks them but typically does not eliminate them.). **[0.5 marks for explanation on how it controls complexity and handle overfitting]***

$$\text{Loss} = \text{MSE} + \lambda \sum_{j=1}^{p} \beta_j^2$$

[0.5 marks if mathematical expression is provided]

Both methods reduce overfitting by penalizing complexity, but they differ in how they affect feature selection and coefficient magnitude.

Given the scenario—a high-dimensional dataset with potentially many irrelevant predictors— ***Lasso*** *is the more appropriate choice. It helps eliminate non-informative features, reduces model complexity, and enhances interpretability without compromising predictive performance.*

*[**0.5 marks for identifying the right approach and 1.5 marks for the justification]***

Q3. You are developing a machine learning model to predict whether a loan application should be approved, using a dataset of 10,000 historical records. Out of these, only 800 applications were approved. [1.5+2+1.5=5 marks]

(a) How would you split the dataset into training, validation, and test sets to ensure fair evaluation and effective model development? Justify your approach.

(b) You train two different models on the dataset:

- Model A achieves 95% accuracy on the test set
- Model B achieves 84% accuracy on the test set

Which model would you consider more reliable for deployment, and why? What additional evaluation metrics would you consider before making a decision?

(c) Why is it important to keep the test set untouched during training phase?

Solution:

Since the dataset contains a small proportion of approved applications (800 out of 10,000), the split should be ***stratified****, meaning the proportion of approved vs. rejected loans is preserved in all subsets. This ensures that each split reflects the original distribution of the target variable and avoids misleading model evaluation.* ***[0.5 marks if stratified is mentioned and 1.5 marks for the correct explanation]***

(b) *Although* ***Model A*** *shows higher accuracy (95%) compared to* ***Model B*** *(84%), accuracy alone can be misleading in this scenario. Given that only 8% of the applicants were approved, a model could predict all applications as "rejected" and still achieve high accuracy.*

Therefore, Model B may actually be better if it is identifying approved loans more effectively.

[0 marks if only Model A or Model B is mentioned without any justification.
0 marks if Model A is chosen and the explanation is simply based on its higher accuracy.
1 mark will be awarded **only if** the reasoning aligns with the correct explanation provided in the answer key (e.g., consideration of class imbalance and the ability to identify approved loans effectively).]

To make an informed decision, we should evaluate additional metrics such as:

- ***Precision*** *and* ***Recall*** *(especially for the minority class)*
- ***F1-score***
- ***Confusion matrix***
- ***ROC-AUC score***

These metrics provide deeper insight into the model's ability to correctly identify approved loans (true positives) without being biased by the majority class.

[1 mark will be awarded if any two metrics are mentioned.]

***(c)** Using the **test set** during this process would lead to **data leakage** and overly optimistic estimates of performance, as the model indirectly "sees" the test data during training. To ensure a fair evaluation of how the model performs on truly unseen data, the test set must be kept untouched until the final model selection is complete.*

[1.5 marks if the correct reasoning is provided]

Q4. Plot the sigmoid function [4 marks]

$$\sigma(wx) = 1/(1 + e^{-wx}),\ x \in R$$

for increasing weights $w \in \{1, 10, 100\}$. A qualitative sketch is enough. Using this plot, explain how changing weights can lead logistic regression to overfit the data.

The sigmoid function is:

$$\sigma(wx) = \frac{1}{1+e^{-wx}}$$

As w increases:

- $w = 1$: Smooth, gradual transition from 0 to 1.
- $w = 10$: Steeper transition.
- $w = 100$: Almost step-like behavior.

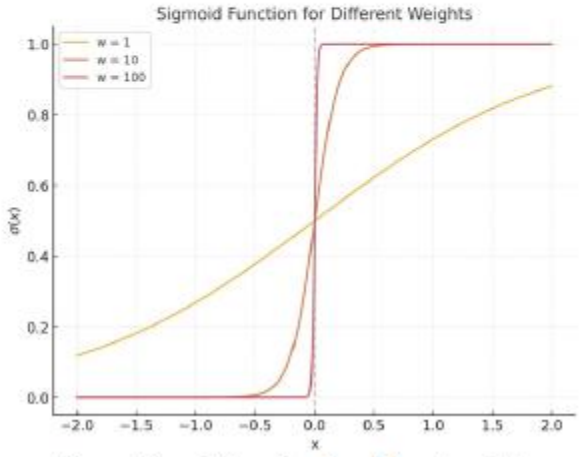

Figure: Sigmoid function for different weights

*Let's examine the function qualitatively for different values of **w**:*

- *When w = 1: The curve is smooth and gradual, transitioning slowly from 0 to 1.*
- *When w = 10: The sigmoid becomes steeper, changing from 0 to 1 more quickly around x=0*
- *When w = 100: The sigmoid becomes very sharp, almost like a step function at x=0.*

How changing weights lead to overfitting:

- *High weights make the sigmoid function very steep, meaning small changes in input x produce large changes in predicted probability.*
- *This causes the model to become overconfident, assigning predictions close to 0 or 1 even when the input is only slightly different from the decision boundary.*
- *Such sharp decision boundaries fit the training data too closely, reducing generalization ability on unseen data — a classic symptom of overfitting.*
- *Loss of Smoothness: The model behaves more like a hard threshold classifier, losing the probabilistic interpretation.*

This shows that large weights reduce the model's ability to generalize, making it prone to overfitting. Hence, increasing weights in logistic regression makes the model more sensitive to input variations and prone to overfitting.

[2 marks for the rough sketch; 2 marks for correctly explaining how changing weights leads to overfitting.]

5) **Answer the following questions:** **[4*1.5 = 6 marks]**

a) You are given a classification problem with 4 classes. What is the minimum number of bits needed to encode the information? Justify your answer mathematically.
b) Suppose all the 4 classes in part (a) have a uniform distribution. Calculate the entropy of this system. Show your complete calculation.
c) If the probability distribution of all the 4 classes is changed from uniform to non-uniform, would the entropy increase or decrease? Explain your reasoning with a suitable example showing numerical calculation.
d) You have two events: Event A with probability 0.2 and Event B with probability 0.8. Which event would need more information to be represented? Justify your answer mathematically.
[Use log base 2 for all your calculations]

Answer Keys

(a) **Minimum bits for encoding (1.5 marks)**

Answer: 2 bits

General Formula: $\lceil \log_2(n) \rceil$ bits for n classes

For 4 classes: $\lceil \log_2(4) \rceil = \lceil 2 \rceil = 2$ bits

(b) **Entropy calculation with uniform distribution (1.5 marks)**

Answer: 2 bits

For uniform distribution with 4 classes, each class probability $= \frac{1}{4} = 0.25$

Using the entropy formula:

$$
\begin{aligned}
H(X) &= -\sum p(x) \times \log_2(p(x)) && (1)\\
&= -4 \times \left(\frac{1}{4} \times \log_2 \left(\frac{1}{4} \right) \right) && (2)\\
&= -\log_2 \left(\frac{1}{4} \right) && (3)\\
&= -\log_2(2^{-2}) && (4)\\
&= -(-2) = 2 \text{ bits} && (5)
\end{aligned}
$$

(c) **Effect of non-uniform distribution on entropy (1.5 marks)**

Answer: Entropy would DECREASE

Reasoning: Uniform distribution maximizes entropy for a given number of classes. Any deviation from uniform results in lower entropy because some outcomes become more predictable.

Example: Non-uniform distribution: $p_1 = 0.7, p_2 = 0.1, p_3 = 0.1, p_4 = 0.1$

$$
\begin{aligned}
H(X) &= -(0.7 \times \log_2(0.7) + 3 \times 0.1 \times \log_2(0.1)) && (6)\\
&= -(0.7 \times (-0.515) + 0.3 \times (-3.322)) && (7)\\
&= -(-0.361 - 0.997) && (8)\\
&= 1.358 \text{ bits} && (9)
\end{aligned}
$$

This is less than 2 bits (uniform case).

(d) **Information content comparison (1.5 marks)**

Answer: Event A (probability 0.2) needs more information

Using the information content formula $I(x) = -\log_2(p(x))$:

Event A ($p = 0.2$)**:**

$$I_A = -\log_2(0.2) = \log_2(5) \approx 2.32 \text{ bits}$$

Event B ($p = 0.8$)**:**

$$I_B = -\log_2(0.9) = \log_2(1.111) \approx 0.15 \text{ bits}$$

Justification: Rarer events (lower probability) carry more information content. Event A is more surprising when it occurs, hence requires more bits to represent. This follows the principle that unexpected events are more informative than expected ones.