Department of Computer Science & Engineering

End Semester Examination

Monsoon Semester, Session: 2025-2026

Subject: Machine Learning (NCSD519 & CSO504)

Model Answer

Time: 2 Hours

Max. Marks: 50

Instructions: Answer the questions as per the instruction within each section.

Sl. No. **Section-I (20 Marks: Attempt All Question)**

1. **(a)** A hospital uses a machine learning system to classify whether a patient has pneumonia using lung X-ray features. A single decision tree performs well on training data but poorly on new patients. The data also contains missing values, noisy pixels, and correlated features. The hospital considers using a random forest to improve performance.

(i)Which components of a random forest help to improve the model generalization on unseen patients.

Sol:- Components of a Random Forest that improve generalization (Any two of the following)

1. Bootstrap Sampling (Bagging)
2. Random Feature Selection
3. Ensemble Averaging

[1x2=2 Marks]

[Full Marks: 2]

(ii) How do these components handle missing values and correlated features effectively?

Bootstrap Sampling (Bagging): Each tree is trained on a random subset of the data sampled with replacement. If a feature value is missing for some samples, it is unlikely to be missing in the same way across all bootstrap samples. The ensemble "fills in the knowledge gap" indirectly since predictions from all trees are aggregated (e.g., majority vote for classification and averaging for regression).

OR

Random feature selection ensures that selected subset of features may not contain missing features. Finally errors from individual trees cancel out and random forest produces stable and robust predictions, improving generalization.

[1.5 Marks]

Random feature selection: At each split, only a random subset of features is considered instead of all features which ensures that dominant corelated features do not appear in every split.

[1.5 Marks]

[Full Marks: 3]

(b) Consider the following 5 training examples:

Customer	**Feature** x	**True Label** y
1	2	+1
2	3	+1
3	5	-1
4	7	-1
5	8	-1

Suppose that you are using AdaBoost with a decision stump (one-level decision tree) as a weak learner and the decision stump predicts based on the following

$$h(x) = +1 \; if \; x < 6 \; else \; -1$$

(i)What is initial weight of each sample?
(ii)Find the weighted classification error.
(iii)Compute the weight (alpha) of weak model.

Ans:-

(i) Initial weight of each sample

In AdaBoost, all samples start with equal weights.

Total samples = 5

Initial weight of each sample:

$w_i = 1 / 5 = 0.2$ [Full Marks: 1]

(ii) Weighted classification error

Prediction using stump:

Customer	True y	h(x)	Result
1	+1	+1	Correct
2	+1	+1	Correct
3	-1	+1	Incorrect
4	-1	-1	Correct
5	-1	-1	Correct

Only Customer 3 is misclassified.

Weighted error = sum of weights of misclassified samples weighted error = 0.2

[Full Marks: 2]

(iii) Weight (alpha) of weak mode (Formula):

$\alpha = 1/2 \ln((1 - \varepsilon) / \varepsilon)$

Substituting $\varepsilon = 0.2$

$\alpha = 1/2 \ln(0.8 / 0.2)$

$\alpha = 1/2 \ln(4)$

$\alpha = 1/2 \times 1.386 = 0.693$

[Full Marks: 2]

2. **(a)**How many binary classifiers are needed in an SVM with N classes when using the One-vs-All (OvA) strategy and the One-vs-One (OvO) strategy? How do the One-vs-All (OvA) and One-vs-One (OvO) strategies in SVM make the final classification decision for a given test sample?

Ans:- Let the number of classes be N. Total number of classifiers required:

One-vs-All (OvA) : N

One-vs-One (OvO) : N *(N − 1) / 2 [1x2=2 Marks]

[Full Marks: 2]

One-vs-All (OvA) :

- Each classifier predicts how likely the test sample belongs to its class.
- All N classifiers evaluate the test sample.
- The class whose classifier gives the **highest decision score** (confidence or margin) is selected as the final class.

Decision rule:

The class with maximum output score is chosen. So:

Final class = argmax(class score) [1.5 Marks]

One-vs-One (OvO):

- Every pairwise classifier votes for one of the two classes it was trained on.
- Each classifier casts one vote.
- The class with the maximum total votes across all classifiers is selected.

Decision rule:
Final class = class with maximum votes. [1.5 Marks]

[Full Marks: 3]

(b) A company wants to build a machine learning model for three different problems:

1. Problem A: Predict whether a transaction is fraud or not fraud based on numerical and categorical features.
2. Problem B: Classify handwritten digits (0–9) from images.
3. Problem C: Predict whether customers will buy a product, with the goal of explaining how each feature influences the probability of purchase.

Which classifier would be most suitable for Problem A, Problem B, and Problem C? Justify your answer.

Ans:-

Problem A: Suitable Choice: Random Forest [1 Marks]
Reason: Fraud detection data is highly imbalanced which is handled by any ensemble method well. [1 Marks]
[Full Marks: 2]

Problem B – Suitable Choice: Convolutional Neural Network (CNN) [1 Marks]
Reason: For images, CNNs generally give the highest accuracy because they learn spatial features. [1 Marks]

[Full Marks: 2]

Problem C –
Suitable Choice: Logistic Regression (Predicting likelihood of purchase with easy Interpretability) [1 Marks]

Section-II (30 Marks: Attempt Any TWO Questions)

3. **(a)** A hospital is developing a machine learning model to predict whether patients have heart disease. After testing the model on 100 patients, the confusion matrix is as follows:

	Predicted Positive	Predicted Negative
Actual Positive	50	10
Actual Negative	30	10

(i) Calculate the F1-score for this model.

(ii)When the primary concern is minimizing missed cases of heart disease rather than avoiding false alarms, how should the F1-score be interpreted in evaluating the model's performance?

(iii)If another model has Precision = 0.9 and Recall = 0.6, which model would you recommend based on the F1-score, and why

Ans:-

Heart Disease Prediction – Model Evaluation

Given Confusion Matrix

TP (True Positive) = 50, FN (False Negative) = 10
FP (False Positive) = 30, TN (True Negative) = 10

(i) Calculation of F1-score : First find Precision and Recall.

Precision = TP / (TP + FP)
= 50 / (50 + 30)
= 50 / 80
= 0.625

Recall = TP / (TP + FN)
= 50 / (50 + 10)
= 50 / 60
= 0.833

F1-score formula:

F1 = 2 × (Precision × Recall) / (Precision + Recall)

F1 = 2 × (0.625 × 0.833) / (0.625 + 0.833)
F1 = 1.041 / 1.458
F1 ≈ 0.71

F1-score = 0.71 [Full Marks: 2]

(ii) Interpretation when minimizing missed heart disease cases

When the primary goal is minimizing missed cases, Recall is more important than precision because high Recall means fewer false negatives (FN)(missing heart disease cases can be life-threatening.) [1 Marks]

Since the F1-score balances Precision and Recall, here it should be interpreted with emphasis on Recall. A high F1-score only is not appropriate to judge the model performance in this case rather priority should be given to Recall even if Precision slightly decreases. So, the model should be judged based on its ability to correctly identify actual heart disease patients. [1 Marks]

[Full Marks: 2]

(iii) Model Comparison

Given second model:

Precision = 0.9

Recall = 0.6

F1-score of second model:

F1 = 2 × (0.9 × 0.6) / (0.9 + 0.6)

= 1.08 / 1.5

= 0.72 [1 Marks]

Comparison:

Recommended model: Original Model [1 Marks]

Reason : Although second model has a slightly higher F1-score (0.72), indicating better overall balance between precision and recall, but recall is more important in this case which is critical for detecting heart disease. Since the original model demonstrates higher recall, it is the preferred choice in this context. [1 Marks]

[Full Marks: 3]

(b) A dataset contains 100 samples for predicting whether a patient has a certain disease (binary classification). A data scientist wants to evaluate the performance of a logistic regression model using cross-validation. He decides to use Leave-One-Out Cross-Validation (LOOCV).

(i)Explain what LOOCV is and how it works in this scenario.

(ii)How many models will be trained during LOOCV for this dataset?

(iii)What are the advantages and disadvantages of using LOOCV compared to, say, 5-Fold Cross-Validation?

Ans:-

(i)LOOCV is a special case of k-Fold Cross-Validation where k = N (number of samples). [1 Marks]

In given scenario works as per following steps.

For each iteration:

o Use N – 1 samples for training.

o Use the 1 remaining sample for testing.

Repeat this N times, so that each sample is used once as a test sample.

[1 Marks]

[Full Marks: 2]

(ii) Number of models trained = 100 [Full Marks: 1]

Each model is trained on 99 samples and tested on 1 sample.

(iii) Any one advantage and disadvantage as per following

Aspect	Advantage	Disadvantage
Accuracy estimate	Uses almost all data for training → low bias	Very high variance if data is noisy
Data usage	Maximizes training data	Computationally expensive for large datasets
Suitability	Small datasets	Not practical for large datasets

[1x2=2 Marks]

[Full Marks: 2]

(c) A language processing company is building models for the following tasks:

1. Named Entity Recognition (NER): Identifying entities in a sentence (e.g., names, locations).
2. Weather forecasting: Predicting temperature over the next week based on historical daily data.
3. Chatbot response generation: Predicting next dialogue response in real-time.

Suggest a suitable RNN variant (Vanilla RNN, Bidirectional RNN, LSTM, GRU) for each task. Justify your choice.

Ans:-

Task	RNN Variant	Justification
NER	Bidirectional RNN (or Bi-LSTM)	Entity recognition depends on context before and after a word; bidirectional RNN captures both directions.
Weather forecasting	LSTM	Time-series prediction with moderate long-term dependencies;
Chatbot response generation	GRU	Requires fast processing; short-term context is often sufficient for immediate response.

[1 x3 = 3 Marks]

[Full Marks: 3]

4. (a)A delivery robot uses epsilon (ε)-greedy Q-learning with ε = 0.2 to choose actions in a warehouse. In state S, it has the following Q-values:

Q(S, Forward) = 5, Q(S, Right) = 8

During one step, the robot unexpectedly chooses the action Forward, even though it has a lower Q-value. After taking this action, it receives a reward r = –1, and the next state S′ has a maximum Q-value of 10. The learning rate α = 0.4 and discount factor γ = 0.9.

(i)Explain why the robot might have selected Forward instead of the higher-valued action Right and identify whether this selection is exploration or exploitation.

(ii) Using the Q-learning update rule, compute the updated Q-value for following: Q(S, Forward).

Ans:-

(a) Given:
ε = 0.2
Q(S, Forward) = 5
Q(S, Right) = 8
Selected action = Forward
Reward r = –1
max Q(S′) = 10
Learning rate α = 0.4
Discount factor γ = 0.9

(i) Reason for selecting Forward due to epsilon-greedy Q-Learning as per following

In ε-greedy Q-learning, the agent:

- Chooses the best action based on maximum Q value (exploitation) with probability (1 − ε)
- Chooses a random action (exploration) with probability ε

Here ε = 0.2 means there is a 20% chance of choosing a random action.

Although Q(S, Right) = 8 is higher than Q(S, Forward) = 5, the robot selected Forward because it was exploring the environment due to the epsilon probability. [1 Marks]

This action selection is Exploration, not exploitation. [1 Marks]

[Full Marks: 2]

(ii) Updated Q-value for Q(S, Forward)

Q-learning update rule:

$Q(S, a) \leftarrow Q(S, a) + \alpha [r + \gamma \max Q(S', a') - Q(S, a)]$

Substituting the given values:

$Q(S, Forward) \leftarrow 5 + 0.4 [-1 + 0.9(10) - 5]$

Step-by-step:

$= 5 + 0.4 [-1 + 9 - 5]$
$= 5 + 0.4 [3]$
$= 5 + 1.2$
$= 6.2$

[Full Marks: 3]

(b) (i)Differentiate between model interpretability and model explainability and discuss how each relates to understanding machine learning predictions. Additionally, explain the trade-off between interpretability, explainability, and model accuracy.

Ans:-

(i) Difference between Model Interpretability and Model Explainability

Model Interpretability:

Model interpretability refers to how easily a human can directly understand the internal working of a model. An interpretable model is inherently transparent, meaning its decision-making logic can be followed without additional tools. [1 Marks]

Model Explainability

Model explainability refers to the ability to provide post-hoc explanations for predictions made by complex or black-box models whose internal structure is not easily understandable in short to explain why a particular prediction was made. [1 Marks]

Trade-off between Interpretability, Explainability, and Model Accuracy

There is often a trade-off between these three:

1. Highly interpretable models: Simple and transparent, Easy to understand, lower accuracy on complex data

2. Highly accurate models: Complex architectures, Hard to interpret, Require explainability techniques

[1 Marks]

[Full Marks: 3]

(ii)A company is designing a convolutional neural network (CNN) for colour images of size 32 × 32 × 3 (Height × Width × Channels). The first convolutional layer has the following parameters:

Number of filters: 8, Filter size: 3 × 3, Stride: 1, Padding: 0 (no padding)

What is output feature map size?

Ans:-

Step 1: Convolutional Layer Output Size
Formula:
Input height = Input width = 32
Filter size = 3
Stride = 1
Padding = 0
Step 2: Depth (Number of Channels)
Depth = Number of filters = 8

Output height (H_out)

$$H_{out} = \frac{H - K + 2P}{S} + 1$$

Output width (W_out)

$$W_{out} = \frac{W - K + 2P}{S} + 1$$

Output channels = number of filters

Hence output feature map size: 30 × 30 × 8 [Full Marks: 2]

(c) A smart city uses an IoT-based environmental monitoring system to record temperature, humidity, and air-quality levels from sensors placed throughout the city. The data is collected every minute and stored in the central system. Below is a description of several unusual observations recorded during one week:

Observation A: One sensor reported a temperature of 85°C at 3:15 PM. All other sensors nearby recorded temperatures between 25°C and 30°C at the same time.

Observation B: A sensor located in a coastal area reports a humidity level of 15% during the early morning hours. However, the same humidity level is commonly observed in that location during noon hours.

Observation C: Over a period of 20 minutes, a cluster of air-quality sensors in the industrial zone recorded a sudden, sharp drop in air-quality index (AQI), indicating heavy pollution. However, each individual reading seems within the normal range when viewed alone.

For each observation (A, B, and C), identify whether it represents a point anomaly, contextual anomaly, or collective anomaly. Justify the reasoning behind your classification.

Ans:-Classification of Anomalies in IoT Environmental Data
Observation A
One sensor reported a temperature of 85°C while nearby sensors recorded 25°C–30°C.
Type: Point Anomaly
Justification:
This single reading is extremely different from the other sensor values at the same time. Since it is an isolated abnormal value compared to normal observations, it represents a point anomaly.

[2 Marks]

Observation B

Humidity level of 15% recorded in the early morning, though such values usually occur at noon.

Type: Contextual Anomaly

Justification:
The value itself (15%) is not abnormal, but it is unusual based on the time context (early morning). Since the anomaly depends on the surrounding condition (time), this is a contextual anomaly.

[1 Marks]

Observation C

A group of sensors shows a sharp AQI drop over 20 minutes, but each individual reading appears normal.
Type: Collective Anomaly
Justification:
Each reading alone does not appear abnormal, but when viewed as a sequence over time, the pattern clearly indicates abnormal behavior. This group behavior forms a collective anomaly. [2 Marks]

[Full Marks : 5]

5. **(a)** (i) Describe how hierarchical clustering allows selecting a specific number of clusters from a dataset, and illustrate your answer with a dendrogram**.**

The following steps are used :

1.Dendrogram Creation : Hierarchical clustering first assumes all points as an individual cluster and merge the clusters using any linkage method to create one cluster (Bottom up approach) [1 Marks]

2.Scan the dendrogram for big vertical distances between merges. Draw a horizontal line through the largest gap.(i.e. Using horizontal cut operation, desired no. of clusters obtained) and diagram illustration.

[2 Marks]

[Full Marks : 3]

(ii)Given Proximity (Distance) Matrix

$$\begin{array}{ccccc} & A & B & C & D \\ A & 0 & 4 & 8 & 7 \\ B & 4 & 0 & 6 & 9 \\ C & 8 & 6 & 0 & 5 \\ D & 7 & 9 & 5 & 0 \end{array}$$

Determine the second merge clusters based on the updated distances after first merge clusters.

Ans:-

Step 1 — Identify the First Merge from distances given in matrix.
Smallest distance = 4, so the first merge is {A, B}. [1 Marks]

Step 2- Update Distances After First Merge (Using linkage method of your choice)
So the updated matrix becomes (single linkage)

$$\begin{array}{cccc} & AB & C & D \\ AB & 0 & 6 & 7 \\ C & 6 & 0 & 5 \\ D & 7 & 5 & 0 \end{array}$$

[2 Marks]

Smallest updated distance = 5

Hence Second merge = {C, D} [1 Marks]

[Full Marks :4]

(b) A company records the monthly spending (in \$100) of 5 customers:

Customer	Spending
A	12
B	15
C	14
D	30
E	28

Use absolute distance to find the medoid of this dataset.

Ans:- Use absolute distance to find the medoid of the dataset.

Absolute distance formula:
Total distance for a point = $\Sigma |x - x_i|$

Step-by-step Calculation:

For A (12):
|12−15| + |12−14| + |12−30| + |12−28|
= 3 + 2 + 18 + 16
= 39

For B (15):
|15−12| + |15−14| + |15−30| + |15−28|
= 3 + 1 + 15 + 13
= 32

For C (14):
$|14-12| + |14-15| + |14-30| + |14-28|$
$= 2 + 1 + 16 + 14$
$= 33$

For D (30):
$|30-12| + |30-15| + |30-14| + |30-28|$
$= 18 + 15 + 16 + 2$
$= 51$

For E (28):
$|28-12| + |28-15| + |28-14| + |28-30|$
$= 16 + 13 + 14 + 2$
$= 45$

Customer	Spending	Total Absolute Distance
A	12	39
B	15	32
C	14	33
D	30	51
E	28	45

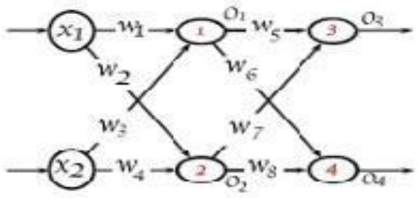
Final Answer

Medoid = Customer B (Spending = 15)

This is because Customer B has the minimum total absolute distance (32) from all other customers.

[Full Marks: 3]

(c) Consider the following Multilayer Neural Network.



Assume that all internal nodes and output nodes use the sigmoid function as activation function. Show, how back propagation algorithm update the values of w_1, w_2, and w_8 for one epoch. Derive an explicit expression for same. Algorithm is given the example (x_1, x_2, y_1, y_2) with y_1 and y_2 being outputs at 3 and 4 respectively (there are no bias terms). Assume that the learning rate of your choice. Let o_1 and o_2 be the output of the hidden units 1 and 2 respectively. Let o_3 and o_4 be the output of the output units 3 and 4 respectively.

Ans:- Updating weight procedure :

```
        For i = 1 to 8
        $Wi = Wi - \alpha (dE/dwi)$
      Where $\alpha$ is the learning rate.
```

$dE/dw8 = (O4 - y2) * O4 * (1 - O4) * O2$ [1 Marks]

$dE/dw2 = \{(O3 - y1) * O3 * (1 - O3) * W7 + (O4 - y2) * O4 * (1 - O4) * W8 \} * O2 * (1 - O2) * x1$

[2 Marks]

$dE/dw1 = \{(O3 - y1) * O3 * (1 - O3) * W5 + (O4 - y2) * O4 * (1 - O4) * W6 \} * O1 * (1 - O1) * x1$

[2 Marks]

[Full Marks : 5]