



SPRING MAKEUP MID SEMESTER EXAMINATION-2025

School of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University
Machine Learning
[CS31002]

Time: 1 1/2 Hours

Full Mark: 20

Answer all the questions.

The figures in the margin indicate full marks.

Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

1. Answer all the questions.

[1 Mark X 5]

a) For the given dataset: $X = [1, 2, 3, 4]$, $Y = [3, 4, 8, 11]$, what is the mean squared error (MSE) if the predicted model is given by $\hat{y} = 2x + 1$?

Step 1: Predict: For each x_i , calculate $\hat{y}_i = 2x_i + 1$:

$$\hat{y}_1 = 2(1) + 1 = 3, \quad \hat{y}_2 = 2(2) + 1 = 5, \quad \hat{y}_3 = 2(3) + 1 = 7, \quad \hat{y}_4 = 2(4) + 1 = 9.$$

Step 2: Difference: Compute the differences $(y_i - \hat{y}_i)$ and square them:

$$(3 - 3)^2 = 0, \quad (4 - 5)^2 = 1, \quad (8 - 7)^2 = 1, \quad (11 - 9)^2 = 4.$$

Step 3: Average: Sum the squared differences and divide by the number of points:

$$\text{MSE} = \frac{0 + 1 + 1 + 4}{4} = 1.5.$$

Model Answer:

$\text{MSE} = 1.5$

Marking: Award 1 mark if all steps (prediction, error computation, and averaging) are clearly stated and the final answer is correct.

b) When would we prefer to use linear regression with gradient descent instead of the least squares method (normal equation), and why?

Gradient descent is preferred when dealing with high-dimensional data or massive datasets, where the computational cost or memory requirement of inverting a matrix (required by the normal equation) is high. It is also useful when using models with added regularisation terms.

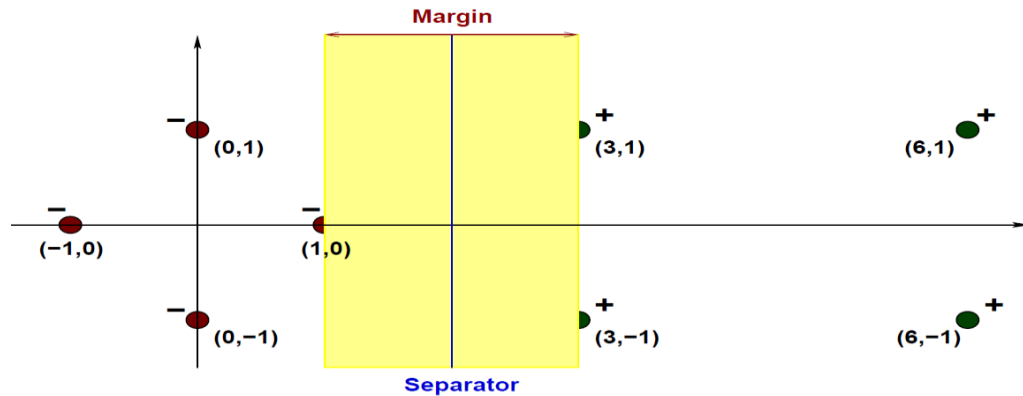
Award 1 mark if at least two valid reasons (e.g., high p , large n , large memory are mentioned.

c) Consider a set of 2-dimensional training data points (x_1, x_2) belonging to two classes '+1' and '-1', respectively, as shown below.

- Class '+1': (3,1) ; (3,-1) ; (6,1) ; (6,-1)
- Class '-1': (1,0) ; (0,1) ; (0,-1) ; (-1,0)

We design a linear hard-margin SVM to classify these linearly separable points. Pictorially (graphically) represent the data points in the 2D plane. Which data points are the support vectors here?

Sol: The support vectors are likely the points closest to the decision boundary that define the margin:



From Class +1: (3,1) and (3,-1)

From Class -1: (1,0)

Award 1 mark for correctly identifying at least two key points.

- d) Consider a feedforward neural network that performs classification task on a p -dimensional input to produce a class label using ' k ' output units. It has ' m ' hidden layers and each of these layers has ' r ' hidden units. What is the total number of trainable parameters (weights and biases) in the network with $p = 10$, $m = 3$, $r = 5$, and $k = 2$?

Sol: A generalized formula for computing the total number of trainable parameters (weights and biases) in a feedforward neural network with:

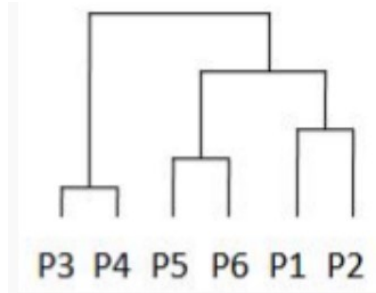
- p : Number of input features
- m : Number of hidden layers
- r : Number of neurons in each hidden layer
- k : Number of output units

$$\begin{aligned}
 \text{Total trainable Parameters} &= (p \cdot r + r) + (m-1)(r \cdot r + r) + (k \cdot r + k) \\
 &= (10 \cdot 5 + 5) + (3-1)(5 \cdot 5 + 5) + (2 \cdot 5 + 2) \\
 &= (50 + 5) + 2(25 + 5) + (10 + 2) \\
 &= 55 + 60 + 12 = 127
 \end{aligned}$$

- e) The pairwise distance between 6 points is given below. Draw dendrogram hierarchy of clusters created by single link clustering algorithm?

	P1	P2	P3	P4	P5	P6
P1	0	3	8	9	5	4
P2	3	0	9	8	10	9
P3	8	9	0	1	6	7
P4	9	8	1	0	7	8
P5	5	10	6	7	0	2
P6	4	9	7	8	2	0

Sol: Apply single-link clustering to determine the order of merges and draw a dendrogram that reflects the clustering structure.



- Merge P3 and P4 at distance 1.
 - Merge P5 and P6 at distance 2.
 - Merge the cluster {P5, P6} with P1 at distance 4.
 - Merge P2 with the cluster from the previous merge at distance 3 (if appropriate after recalculating distances).
 - Finally, merge with the cluster {P3, P4} at the highest distance (e.g., 6).
2. Derive the gradient of the log-likelihood function for logistic regression with respect to the parameters and use it to update the parameters. Assuming that the input data is represented by a matrix X with dimensions $n \times p$ where n is the number of observations and p is the number of features, and the parameters are represented by a vector θ with dimensions $p \times 1$.

[5 Marks]

Sol:

Let:

- $X \in \mathbb{R}^{n \times p}$: Input feature matrix with n samples and p features.
- $\theta \in \mathbb{R}^{p \times 1}$: Parameter vector.
- $y \in \{0, 1\}^n$: Binary class labels for each sample.
- $\hat{y} = \sigma(X\theta)$: Predicted probabilities using the **sigmoid function**:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad \text{where } z = X\theta$$

1. Log-Likelihood Function

For binary classification, the **log-likelihood** $\mathcal{L}(\theta)$ is:

$$\mathcal{L}(\theta) = \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

where $\hat{y}_i = \sigma(X_i\theta)$.

2. Gradient of Log-Likelihood

We compute the gradient with respect to the parameter vector θ :

$$\nabla_{\theta} \mathcal{L}(\theta) = X^T(\mathbf{y} - \hat{\mathbf{y}})$$

Explanation:

- $\hat{\mathbf{y}} = \sigma(X\theta)$ is the vector of predicted probabilities.
- The difference $\mathbf{y} - \hat{\mathbf{y}}$ gives the residual (error).
- Multiplying by X^T propagates this error backward to adjust each parameter.

3. Parameter Update Rule

To **maximize** the log-likelihood, we use **gradient ascent** (or **gradient descent** on the negative log-likelihood):

$$\theta^{(t+1)} = \theta^{(t)} + \alpha X^T(\mathbf{y} - \hat{\mathbf{y}})$$

- α : Learning rate.
- $\theta^{(t)}$: Current parameter vector.
- $\theta^{(t+1)}$: Updated parameter vector.

3. For a binary classification problem, consider the training examples shown in the following table. The features, **A₁** and **A₂**, can take either True or False values and the **Class** label can be either + (positive) or – (negative). Answer the following.

Instance	A ₁	A ₂	Class
1	True	True	+
2	True	True	+
3	True	False	-
4	False	False	+
5	False	True	-
6	False	True	-
7	False	False	-

8	True	False	+
9	False	True	-

- (a) What is the entropy of this collection of training examples with respect to positive (+) class?
- (b) What are the information gains of A_1 and A_2 relative to these training examples?
- (c) Which is the best feature (among A_1 , and A_2) to split according to the information gain?

[5 Marks]

Sol:

(a)

There are 4 positive (+) examples and 5 negative (-) examples. Thus, $\mathbb{P}_{(+)} = \frac{4}{9}$ and $\mathbb{P}_{(-)} = \frac{5}{9}$.
 The entropy w.r.t. the positive (+) class of the training examples is $= -\frac{4}{9} \log_2 \left(\frac{4}{9} \right) = 0.52$.
 The entropy w.r.t. the negative (-) class of the training examples is $= -\frac{5}{9} \log_2 \left(\frac{5}{9} \right) = 0.47$.

The entropy of the training examples is $= -\frac{4}{9} \log_2 \left(\frac{4}{9} \right) - \frac{5}{9} \log_2 \left(\frac{5}{9} \right) = 0.9911$.

(b)

For attribute A_1 , the corresponding counts and probabilities are:

A_1	+	-
True	3	1
False	1	4

The entropy for A_1 is $= \frac{4}{9} \left[-\frac{3}{4} \log_2 \left(\frac{3}{4} \right) - \frac{1}{4} \log_2 \left(\frac{1}{4} \right) \right] + \frac{5}{9} \left[-\frac{1}{5} \log_2 \left(\frac{1}{5} \right) - \frac{4}{5} \log_2 \left(\frac{4}{5} \right) \right] = 0.7616$.
 Therefore, the information gain for A_1 is $(0.9911 - 0.7616) = 0.2294$.

For attribute A_2 , the corresponding counts and probabilities are:

A_2	+	-
True	2	3
False	2	2

The entropy for A_2 is $= \frac{5}{9} \left[-\frac{2}{5} \log_2 \left(\frac{2}{5} \right) - \frac{3}{5} \log_2 \left(\frac{3}{5} \right) \right] + \frac{4}{9} \left[-\frac{2}{4} \log_2 \left(\frac{2}{4} \right) - \frac{2}{4} \log_2 \left(\frac{2}{4} \right) \right] = 0.9839$.
 Therefore, the information gain for A_2 is $(0.9911 - 0.9839) = 0.0072$.

(c)

According to information gain, A_1 produces the best split.

4. Consider the data set shown in the following table.

Instance	A	B	C	Class
1	0	0	1	—
2	1	0	1	+
3	0	1	0	—
4	1	0	0	—
5	1	0	1	+
6	0	0	1	+
7	1	1	0	—
8	0	0	0	—
9	0	1	0	+
10	1	1	1	+

The attributes, A, B and C, can take two values (either 1 or 0) and the Class can be either + or —. Predict the class label for a given test sample, (A = 0, B = 1, C = 1), using the Naive Bayes approach. [5 Marks]

Sol:

The attributes, A, B and C, can take two values (either 1 or 0) and the **Class** can be either + or —. Answer the following.

(a) Estimate the conditional probabilities for the following:

$\mathbb{P}(A = 0 \mid +)$, $\mathbb{P}(B = 1 \mid +)$, $\mathbb{P}(C = 1 \mid +)$, $\mathbb{P}(A = 0 \mid -)$, $\mathbb{P}(B = 1 \mid -)$, $\mathbb{P}(C = 1 \mid -)$.

Solution:

$$\begin{aligned}\mathbb{P}(A = 0 \mid +) &= \frac{2}{5} & \mathbb{P}(A = 0 \mid -) &= \frac{3}{5} \\ \mathbb{P}(B = 1 \mid +) &= \frac{2}{5} & \mathbb{P}(B = 1 \mid -) &= \frac{2}{5} \\ \mathbb{P}(C = 1 \mid +) &= \frac{4}{5} & \mathbb{P}(C = 1 \mid -) &= \frac{1}{5}\end{aligned}$$

$$\begin{aligned}\mathbb{P}(+ \mid A = 0, B = 1, C = 1) &= \mathbb{P}(+) \cdot \mathbb{P}(A = 0, B = 1, C = 1 \mid +) \\ &= \frac{\mathbb{P}(+) \cdot \mathbb{P}(A = 0 \mid +) \cdot \mathbb{P}(B = 1 \mid +) \cdot \mathbb{P}(C = 1 \mid +)}{\mathbb{P}(+) \cdot \mathbb{P}(A = 0 \mid +) \cdot \mathbb{P}(B = 1 \mid +) \cdot \mathbb{P}(C = 1 \mid +) + \mathbb{P}(-) \cdot \mathbb{P}(A = 0 \mid -) \cdot \mathbb{P}(B = 1 \mid -) \cdot \mathbb{P}(C = 1 \mid -)} \\ &= \frac{(\frac{1}{2}) \cdot (\frac{2}{5}) \cdot (\frac{2}{5}) \cdot (\frac{4}{5})}{(\frac{1}{2}) \cdot (\frac{2}{5}) \cdot (\frac{2}{5}) \cdot (\frac{4}{5}) + (\frac{1}{2}) \cdot (\frac{3}{5}) \cdot (\frac{2}{5}) \cdot (\frac{1}{5})} = \frac{8}{11}.\end{aligned}$$

$$\therefore \mathbb{P}(- \mid A = 0, B = 1, C = 1) = 1 - \mathbb{P}(+ \mid A = 0, B = 1, C = 1) = \frac{3}{11}.$$

Since $\mathbb{P}(+ \mid A = 0, B = 1, C = 1) > \mathbb{P}(- \mid A = 0, B = 1, C = 1)$, therefore the predicted class label will be '+'.

*** Best of Luck ***