MADRAS SCHOOL OF ECONOMICS

POSTGRADUATE DEGREE PROGRAMME IN DATA SCIENCE [2025-27]

SEMESTER TWO [JANUARY – MAY, 2026]

SUPPLEMENTARY EXAMINATION, JULY 2026

Course Name: Introduction to Machine Learning, **Course Code:**125

Duration: 2 Hours **Maximum Marks: 60**

Instructions: *Any numerical computation also requires justification using theorems. The programming questions must pass the public and private test cases. More details on each programming question can be found in the online portal.*

Answer any six questions.

1. [10 marks] Explain the bias–variance tradeoff in machine learning. Your answer should clearly define bias and variance, describe how they contribute to prediction error, and illustrate the tradeoff using linear regression (for example, by comparing simple and high-degree polynomial models).

2. [10 marks] Write a note on unsupervised learning algorithms and contrast their use with supervised learning. Explain K-means and PCA with an explicit pseudocode.

3. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and define

$$f(x) = x^T Ax.$$

   (a) [5 marks] Compute the gradient $\nabla f(x)$ and the Hessian $\nabla^2 f(x)$.

   (b) [5 marks] Using the Hessian criterion, show that f is convex on $\mathbb{R}^n$ if and only if A is positive semidefinite.

   Show all steps clearly.

4. In logistic regression, the loss for a single datapoint is:

$$L(w) = -y\log(\sigma(z)) - (1-y)\log(1-\sigma(z)), \quad z = w^T x,$$

   where

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

   You are given:

$$x = \begin{bmatrix}1\\2\end{bmatrix}, \quad y = 1, \quad w = \begin{bmatrix}0\\0\end{bmatrix}.$$

   Compute:

   (a) [5 marks] The gradient $\nabla L(w)$.

   (b) [5 marks] The Hessian matrix $\nabla^2 L(w)$.

   Show all steps clearly.

5. Suppose you are trying to decide whether an email is spam based on the words it contains. You notice that certain words like free and win tend to appear more often in spam emails than in regular emails. You decide to classify a new email based on how frequently these words appear among past emails.

   Here is the mathematical setup: Consider a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where $y_i \in \{0, 1\}$ indicates whether the email is spam, and each $x_i = (x_{i1}, x_{i2})$ consists of two binary features indicating the presence of the words free and win. Each feature takes value 1 if the word is present and 0 otherwise.

   Assume that, given the class y, the two features are independent. Estimate probabilities using relative frequencies from the training data:

$$P(y) = \frac{\#\{i : y_i = y\}}{n}, \quad P(x_j = 1 \mid y) = \frac{\#\{i : x_{ij} = 1 \text{ and } y_i = y\}}{\#\{i : y_i = y\}}.$$

   Example. If in a dataset of 10 emails, 4 are spam and among those 4 spam emails, 3 contain the word free, then:

$$P(y = 1) = \frac{4}{10}, \quad P(x_1 = 1 \mid y = 1) = \frac{3}{4}.$$

   For a new email $x = (x_1, x_2)$, define the score:

$$\text{score}(y) = P(y)\, P(x_1 \mid y)\, P(x_2 \mid y),$$

   and predict the class with the larger score.

   Let the training dataset be:

$$\mathcal{D} = \{(1,1)\to 1,\ (1,0)\to 1,\ (0,1)\to 1,\ (1,1)\to 1,\ (1,0)\to 0,\ (0,1)\to 0,\ (0,0)\to 0,\ (0,0)\to 0\}.$$

   (a) [6 marks] Use the above procedure to classify the email $x = (1, 1)$. Compute all required probabilities from the data, evaluate the scores for both classes, and state the predicted label.

   (b) [4 marks] Suppose that in the training data, the word free appears in exactly the same proportion of emails in both classes. How would this affect the classification rule? Work out the math and explain the effect clearly.

6. In the decision tree algorithm, candidate splits are evaluated based on the reduction in Gini impurity.

   (a) [6 marks] Show that for any split of a node into two child nodes, the impurity after the split cannot exceed the impurity of the parent node. Clearly justify your reasoning.

   (b) [4 marks] Suppose a parent node N contains 10 samples: 4 belonging to Class A and 6 belonging to Class B.

   A candidate split partitions the data into two child nodes:

   - Left child N_L: 5 samples (4 from Class A, 1 from Class B)
   - Right child N_R: 5 samples (remaining samples)

   Compute the decrease in Gini impurity achieved by this split.

7. Consider a two-layer neural network with a single hidden neuron and a single output neuron. For an input $x \in \mathbb{R}$, the network computes:

$$z_1 = w_1 x + b_1, \quad a_1 = \sigma(z_1),$$

$$z_2 = w_2 a_1 + b_2, \quad \hat{y} = \sigma(z_2),$$

   where $\sigma(z) = \frac{1}{1+e^{-z}}$.

   The loss function is:

$$L = \frac{1}{2}(\hat{y} - y)^2.$$

(a) [5 marks] Using the chain rule, compute $\frac{\partial L}{\partial w_2}$ and $\frac{\partial L}{\partial b_2}$. Show all intermediate steps.

(b) [5 marks] Using the chain rule, compute $\frac{\partial L}{\partial w_1}$ and $\frac{\partial L}{\partial b_1}$. You must clearly show how gradients propagate through the hidden layer.

8. In a random forest, each tree is trained on a bootstrap sample obtained by sampling n data points with replacement from a dataset of size n.

   For a fixed data point i, let S_b denote the bootstrap sample used to train tree b.

   (a) [5 marks] Show that the probability that data point i is not included in S_b is

   $$\mathbb{P}(i \notin S_b) = \left(1 - \frac{1}{n}\right)^n.$$

   Hence, show that this probability converges to e^{-1} as $n \to \infty$.

   (b) [5 marks] Explain how this result justifies the use of out-of-bag (OOB) samples to estimate test error in random forests.