MADRAS SCHOOL OF ECONOMICS

POSTGRADUATE DEGREE PROGRAMME IN DATA SCIENCE [2025-27]

SEMESTER TWO [JANUARY – MAY, 2026]

REGULAR END TERM EXAMINATION, MAY 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. (a) [4 marks] Consider the 1-dimensional Ridge objective:

$$L(w) = \frac{1}{2}(w-4)^2 + \lambda w^2, \quad \lambda > 0.$$

Determine whether there exists a finite value of λ for which the optimal solution satisfies $w^* = 0$. Justify your answer by showing all steps of reasoning.

(b) [4 marks] Consider the 1-dimensional Lasso objective:

$$L(w) = \frac{1}{2}(w-4)^2 + \lambda|w|, \quad \lambda > 0.$$

Determine the smallest value of λ for which the optimal solution satisfies $w^* = 0$. You must analyse all relevant cases and show all steps clearly.

(c) [2 marks] Explain briefly how the qualitative behaviour of Ridge regression differs from Lasso.

3. It is often said that a person can be judged by the company they keep. For instance, if your friends are all good people, you are likely to be a good person. This idea can be translated into a simple classification rule: the label of a point is determined by the labels of nearby points.

Here is the mathematical setup: Consider the following classification procedure. Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ in $\mathbb{R}^2$ with $y_i \in \{0, 1\}$, we wish to predict the label of a new point x. The procedure is to compute the Euclidean distance from x to each x_i, select the k closest points, and assign to x the most frequent label among these neighbors.

For example, if $k = 3$ and the three nearest neighbors of a point have labels $\{1, 0, 1\}$, then the predicted label is 1.

Let $k = 3$ and consider the training dataset

$$\mathcal{D} = \{(0,0)\to 0,\ (1,0)\to 0,\ (0,1)\to 1,\ (1,1)\to 1,\ (2,1)\to 1\}.$$

(a) [5 marks] Use the above procedure to classify the point $x = (1, 0.5)$. Show the distances, identify the k nearest neighbors, and state the predicted label.

(b) [5 marks] Predict the behaviour of the above procedure as k varies over the values $\{1, 2, 3, 4, 5\}$. Which is more sensitive to noise and which produces smoother predictions? You dont have to compute but you have to reason and explain.

4. Consider the dataset in $\mathbb{R}^2$

$$\mathcal{D} = \{(\pm 1, 0), (0, \pm 1), (\pm 3, 0), (0, \pm 3)\}.$$

(a) [4 marks] Write the pseudocode of Lloyd's algorithm.

(b) [6 marks] Implement Lloyd's algorithm for K-means with $K = 2$ using the initialization

$$\mu_1^{(0)} = (1, 0), \quad \mu_2^{(0)} = (-1, 0),$$

and report the final centroids and the final value of the K-means objective. Show all intermediate cluster assignments for each iteration.

5. Consider logistic regression with predicted probability

$$\hat{y} = \sigma(w^T x)$$

and a Beta prior

$$P(\hat{y}) \propto \hat{y}^{\alpha-1}(1 - \hat{y})^{\beta-1}.$$

(a) [5 marks] Show that, for a single data point (x, y), the gradient of the negative log-posterior can be written in the form

$$\nabla_w \mathcal{L} = (-\alpha + (\alpha + \beta - 1)\hat{y})\, x.$$

(b) [5 marks] Let

$$x = \begin{bmatrix} 1 & 3 \end{bmatrix}, \quad y = 1, \quad \alpha = 3,\ \beta = 5, \quad w^{(0)} = \begin{bmatrix} 0 & 0 \end{bmatrix}, \quad \eta = 0.05.$$

Compute the updated weight vector $w^{(1)}$ after one gradient descent step on the negative log-posterior.

6. Consider a random forest with B trees, where each tree produces a prediction $T_b(x)$ with

$$\mathrm{Var}(T_b(x)) = \sigma^2, \quad \mathrm{Corr}(T_b(x), T_{b'}(x)) = \rho \text{ for } b \neq b'.$$

The random forest predictor is

$$\mathrm{RF}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x).$$

(a) [4 marks] Show that

$$\mathrm{Var}(\mathrm{RF}(x)) = \frac{\sigma^2}{B} + \frac{B-1}{B}\rho\sigma^2.$$

(b) [3 marks] Let $\mathcal{F}_b, \mathcal{F}_{b'} \subset \{1, \ldots, d\}$ be chosen independently and uniformly at random, each of size m. Prove that

$$\mathbb{E}\big[|\mathcal{F}_b \cap \mathcal{F}_{b'}|\big] = \frac{m^2}{d}.$$

(c) [3 marks] Suppose

$$T_b(x) = \sum_{j \in \mathcal{F}_b} g_j(x_j),$$

where the random variables $g_j(x_j)$ are independent with common variance τ^2. Show that

$$\mathbb{E}\big[\,\mathrm{Cov}(T_b(x), T_{b'}(x))\big] = \tau^2 \frac{m^2}{d},$$

and briefly explain why decreasing m reduces correlation between trees.

7. A single neural layer with two neurons performs a weighted sum followed by a sigmoid activation function.

   For an input x, the neurons compute:

$$a_1 = \sigma(w_1 x + b_1), \quad a_2 = \sigma(w_2 x + b_2),$$

   where

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

   Write a function `forward_pass(x, w, b)` that returns the output vector $[a_1, a_2]$, where $w = [w_1, w_2]$ and $b = [b_1, b_2]$.

8. A neuron computes:

$$\hat{y} = \sigma(w^T x + b),$$

   where

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

   The loss function is:

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

   Write a function `gradient_step(x, y, w, b, lr)` that performs one step of gradient descent:

$$w \leftarrow w - lr \cdot \frac{\partial L}{\partial w}, \quad b \leftarrow b - lr \cdot \frac{\partial L}{\partial b}.$$