Introduction to machine learning
Exercise 3

Fall 2025/26

Submission guidelines, **read and follow carefully**:

- The exercise **must** be submitted in pairs.
- Submit via Moodle.
- The submission should be only a PDF file with your answers to all the questions.
- No need to submit code for Question 1.
- For questions, use the exercise forum, or if they are not of public interest, send them to the course staff email intromlbgu26@gmail.com.
- Grading: Q.1 : 30 points, Q.2: 20 points, Q.3: 10 points, Q.4: 24 points, Q.5: 16 points,

Question 1. **Ridge-Regression and Least Squares**
In this problem, you will implement the Ridge-Regression algorithm and the Linear Least Squares solution and compare them.

Setup: Use the dataset `lsdata.mat` (where each $x \in \mathbb{R}^d$), which is provided on the course web page. You can load the data as follows:

```
import scipy.io as sio
import numpy as np
data = sio.loadmat('lsdata.mat')
X, Y = data['X'], data['Y']
X_test, Y_test = data['Xtest'], data['Ytest']
```

(a) Implement a function that computes the weight vector w for the Least Square Problem, as was shown in class, for $m > d$. Run this for training set sizes $m \in \{100, 110, \ldots, 500\}$. For each m, sample m points from the training set, compute w, and calculate average squared loss on the training set and a separate test set.

  i. Submit a plot showing **average squared loss on the test set** as a function of m.
  ii. Submit a plot showing the **average squared loss on the training set** as a function of m.

(b) Implement the ridge-regression algorithm. Run it using $m = 60$ and $\lambda \in \{0, 0.01, 0.02, 0.05, 0.1, 1, 10, 15\}$. Calculate the solution to the least square problem using the same data.

i. Submit a plot showing **average squared loss on the test set** as a function of λ. On the same plot, include a horizontal line representing the average squared loss obtained by the least square solution.

ii. Describe the results. What is the behavior of the test loss as λ varies from small large? Explain the observed behavior.

iii. Repeat the same experiment for $m = 500$. Submit a plot showing **average squared loss on the test set** as a function of λ. On the same plot, include a horizontal line representing the average squared loss obtained by the least square solution.

iv. What is the difference in the results when using a large value of m (500), compared to a smaller one (60)?

Question 2. Consider a classification problem for input space $\mathcal{X} = \mathbb{R}_+^d$ of d-dimensional vectors of strictly-positive real components, and a label space $\mathcal{Y} = \{-1, 1\}$.

For positive real parameters $a, b > 0$, we define the following function for any two input vectors $x, x' \in \mathbb{R}_+^d$:

$$Q_{a,b}\left(x, x'\right) = a\sqrt{x(2)x'(2)x(3)x'(3)} + \frac{b}{\sqrt{x(2)x'(2)x(3)x'(3)}} + \sum_{i=1}^{d} \sqrt{x(i)x'(i)}.$$

(a) Is $Q_{a,b}$ a kernel function for any $a, b > 0$?

- If yes, formulate a possible feature map function $\psi : \mathcal{X} \to \mathcal{F}$, for a feature space $\mathcal{F}$, that proves that $Q_{a,b}$ is a kernel function for any $a, b > 0$.
- If not, mathematically explain why.

(b) We are given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of input-output examples from $\mathcal{X} \times \mathcal{Y}$. Consider an integer $k > d$, a function $\psi : \mathbb{R}_+^d \to \mathbb{R}^k$, and an optimization problem

$$\min_{w \in \mathbb{R}^k} \frac{1}{m} \sum_{i=1}^{m} y_i \left\langle w, \psi(x_i) \right\rangle + g(w)$$

for a function $g : \mathbb{R}^k \to \mathbb{R}$.

There is at least one $w \in \mathbb{R}^k$ that solves this minimization problem.

Is there a function $g : \mathbb{R}^k \to \mathbb{R}$ for which necessarily exist $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ such that $w = \sum_{i=1}^{m} \alpha_i \psi(x_i)$ solves the minimization problem?

If yes, prove your answer by formulating a possible function g and explain. If no, explain why.

Question 3. For a fixed (constant) parameter $b \in \mathbb{R}$, we define the function

$$f(x) = |x - b|$$

for real input $x \in \mathbb{R}$.

What is the subgradient **set** of f at $x = b$? Mathematically prove your answer.

Note: You should specify the subgradient **set**, i.e., all the possible subgradients of f at the input point $x = b$.

Question 4. Consider a regression problem with input space $\mathcal{X} = \mathbb{R}^d$ and output space $\mathcal{Y} = \mathbb{R}$. The unknown distribution $\mathcal{D}$ is defined over $\mathcal{X} \times \mathcal{Y}$. The given sample $S = \{(x_i, y_i)\}_{i=1}^m$ includes m input-output examples i.i.d. from $\mathcal{D}$. Define $\mathbf{X} = [x_1, \dots, x_m]$ as the $d \times m$ matrix of the input examples from S organized as the matrix rows, and $\mathbf{y} = [y_1, \dots, y_m]^T$ as the m-dimensional column vector of the output examples from S.

The learning of a linear regression predictor is defined here by the following minimization problem for a hyperparameter $\lambda > 0$ and a matrix $A \in \mathbb{R}^{d \times d}$:

$$\widehat{w} \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{m} \left(\langle w, x_i \rangle - y_i \right)^2 + \lambda \|Aw\|_2^2$$

(a) Does this optimization problem (for $\lambda > 0$) have a unique solution?

- If the solution is not (necessarily) unique, formulate a mathematical condition that guarantees a unique solution.
- If the solution is unique, mathematically prove it.

(b) Mathematically formulate the solution $\widehat{w}$ in a closed form.

- If the solution for $\widehat{w}$ is unique, formulate it.
- If there is more than one solution for $\widehat{w}$, formulate the solution for $\widehat{w}$ with the minimal ℓ_2-norm $\|\widehat{w}\|_2$ among all possible solutions.

In the provided formula you can use only $\mathbf{X}, \mathbf{y}, A, \lambda$ that were defined in this question. If needed, you can also use basic mathematical elements and symbols, including the identity matrix.

Provide the mathematical developments that prove the closed-form formula.

Question 5. Consider a dimensionality reduction problem input data that has the probability distribution $\mathcal{D}_x$ over $\mathbb{R}^d$. Consider the dimensionality reduction problem for a sample of m vectors $S = \{x_1, \dots, x_m\} \subset \mathbb{R}^d$ that are i.i.d. drawn from $\mathcal{D}_x$, and $k < d$:

$$\widehat{U}_k = \underset{U \in \mathbb{R}^{d \times k} : U^T U = I_k}{\operatorname{argmin}} \sum_{i=1}^{m} \|x_i - UU^T x_i\|_2^2.$$

(a) The m low-dimensional vectors $z_1, \dots, z_m \in \mathbb{R}^k$ are defined as

$$z_i = \widehat{U}_k^T x_i, \ \forall i \in \{1, \dots, m\}.$$

Define the matrix

$$Z = \sum_{i=1}^{m} z_i z_i^T$$

Mathematically prove or disprove the following claim: The matrix Z cannot have k nonzero eigenvalues.

(b) A group of motivated students would like to find the best value for the hyperparameter k such that the operator $\widehat{U}_k$ that was learned from the sample $S = \{x_1, \dots, x_m\}$ will have a low distortion (representation error) on a new random input drawn from $\mathcal{D}_x$ independently of S.

Is it a good approach to choose the value of k that achieves the minimal distortion (representation error) of $\sum_{i=1}^m \|x_i - U_k U_k^T x_i\|_2^2$? Explain your answer.