Introduction to machine learning
Exercise 3

Fall 2024/25

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 intromlbgu25@gmail.com.
- Grading: Q.1 : 21 points, Q.2: 16 points, Q.3: 20 points, Q.4: 18 points, Q.5: 15 points, Q.6: 10 points

Question 1. Implement the ridge-regression algorithm. **No need to submit your code**. Run the algorithm on the dataset `regdata.mat` provided on the course web page, you may use the following code to load the math file with python:

```
import scipy.io as sio
data = sio.loadmat('regdata.mat')
X = data['X']
Y = data['Y']
Xtest = data['Xtest']
Ytest = data['Ytest']
```

Run the regression using $\lambda \in \{0, 1, 2, \ldots, 30\}$ on the training set X, Y provided in the data file. Try training-set sizes between 10 and 100. For each training set size that you try, find the value of λ that obtains the smallest mean-squared-error (the average squared loss) on the test set provided in the data file.

(a) Submit a plot of the value of λ that minimizes the mean squared error on the test set as a function of the training set size m.

(b) What trend to you expect in the plot based on what we learned in class? Explain.

(c) Did you get this trend in the plot you submitted? If there are any differences, explain why they could occur.

(d) In this data set, the label y of each example x was generated by setting $y = \langle w, x\rangle + \eta$, where w is a fixed vector which is the same for all examples in the data set, and η is a standard Gaussian random variable, $\eta \sim N(0, \sigma)$ for some $\sigma > 0$. η is drawn independently for each example in the data set. What is the Bayes-optimal predictor for this problem with respect to the squared loss? And how about the absolute loss? Prove your claims.

Question 2. We are given a training sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where the input space is $\mathcal{X} = \mathbb{R}^d$ and the output space is $\mathcal{Y} = \mathbb{R}$. Consider the learning of linear regression predictor by the following optimization problem that has two hyperparameters $\lambda_1 \geq 0$ and $\lambda_2 > 0$:

$$\widehat{w} = \operatorname*{argmin}_{w \in \mathbb{R}^d} \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 + \sum_{i=1}^{m} \left(\langle w, x_i\rangle - y_i\right)^2 .$$

Let us define the $d \times m$ matrix of training input examples as $\mathbf{X}_s = [x_1, \ldots, x_m]$, and the vector of training output examples as $\mathbf{y}_s = [y_1, \ldots, y_m]^T$.

Clarification: The solution for this question should not involve quadratic programming.

(a) **In this section consider $\lambda_1 = 0$.**
Formulate a closed-form solution for $\widehat{w}$ that solves the optimization defined in this question. In the formula you may use only $\mathbf{X}_s$, $\mathbf{y}_s$ and λ_2, that were defined in this question, and basic mathematical elements (including the identity matrix).

(b) **In this section consider $\lambda_1 > 0$.**
Assume that the gradient descent algorithm is used for solving the optimization problem in this question with the training sample S and step size η. Formulate the update formula of the $(t+1)^{\text{th}}$ iteration of the gradient descent. Specifically, formulate $w^{(t+1)}$ using η, $w^{(t)}$, λ_1, λ_2, $\mathbf{X}_s$, $\mathbf{y}_s$ and basic mathematical elements (including the identity matrix).

Question 3. Consider a regression problem with input space $\mathcal{X} = \mathbb{R}^d$ and output space $\mathcal{Y} = \mathbb{R}^q$ for $q > 1$. 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 $\mathcal{D}$. Specifically, note that the outputs in this problem are vectors. Define $\mathbf{X} = [x_1, \ldots, x_m]^T$ as the $m \times d$ matrix of the input examples from S organized as the matrix rows (note the transpose), and $\mathbf{Y} = [y_1, \ldots, y_m]^T$ as the $m \times q$ matrix of the output examples from S organized as the matrix rows (note the transpose).

The task is to learn a $d \times q$ matrix $\mathbf{W}$ that can be used as a linear prediction operator for a given $x \in \mathcal{X}$: $\widehat{y} = \mathbf{W}^T x$. The performance is measured by the loss function of $\ell(y, \widehat{y}) = \|y - \widehat{y}\|_2^2$.

The Frobenius norm of the matrix $\mathbf{M} \in \mathbb{R}^{r \times c}$ is $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^r \sum_{j=1}^c m_{r,c}^2}$ where $m_{r,c}$ is the component at the (r, c) coordinate of $\mathbf{M}$. The following derivation rule will be useful for solving this question:

$$\frac{d}{d\mathbf{M}} \|\mathbf{\Theta}\mathbf{M} - \mathbf{\Omega}\|_F^2 = 2\mathbf{\Theta}^T(\mathbf{\Theta}\mathbf{M} - \mathbf{\Omega})$$

where $\mathbf{\Theta} \in \mathbb{R}^{b \times r}$ and $\mathbf{\Omega} \in \mathbb{R}^{b \times c}$.

(a) In this section, the learning is defined as the optimization problem:

$$\widehat{\mathbf{W}} = \operatorname*{argmin}_{\mathbf{W} \in \mathbb{R}^{d \times q}} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_F^2$$

where $\lambda > 0$ is a hyperparameter that determines the regularization strength.

(i) Mathematically formulate the solution for $\widehat{\mathbf{W}}$ in a closed form. Provide the mathematical developments that prove the closed-form formula.

(ii) 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.

(b) In this section, the learning is defined as the optimization problem:

$$\widehat{\mathbf{W}} = \underset{\mathbf{W} \in \mathbb{R}^{d \times q}}{\operatorname{argmin}} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_F^2$$

Does this optimization problem have a unique solution? If the solution is not unique, formulate a mathematical condition that guarantees a unique solution.

Question 4. **Prove or disprove** the following claims.

(a) Consider the dimensionality reduction problem for m vectors $x_1, \ldots, x_m \in \mathbb{R}^d$, and $k < d$:

$$\widehat{U}, \widehat{C} = \underset{U \in \mathbb{R}^{d \times k}, C \in \mathbb{R}^{k \times d}}{\operatorname{argmin}} \sum_{i=1}^{m} \|x_i - UCx_i\|_2^2.$$

Claim: The rank of $\widehat{U}\widehat{C}$ is at most k.

(b) $\mathcal{M} = \{U \in \mathbb{R}^{d \times k}$ such that $U^T U = I_k\}$ where $k < d$, and I_k is the $k \times k$ identity matrix.
Claim: $\mathcal{M}$ is a convex set.

(c) Consider the dimensionality reduction problem for m vectors $x_1, \ldots, x_m \in \mathbb{R}^d$, and $k < d$:

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

Then, the m low-dimensional vectors $z_1, \ldots, z_m \in \mathbb{R}^k$ are defined as

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

Define the matrix

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

Claim: For $k > m \geq 1$, the matrix Z cannot have k nonzero eigenvalues.

Question 5. Consider the algorithm that optimizes the k-means objective that we learned in class. For each of the 3 axioms, answer whether this algorithm satisfies them and prove your claim. Answer according to the following sub-questions:

(a) Scale invariance

(b) Richness

(c) Consistency

Hint: Consider the following example. There are 5 points $x_1, \ldots, x_5$. There are no other points in the space (so the centroids must be one of these points). The distances between them are

$$\rho(x_i, x_j) = 1 \quad \text{for all } i \neq j \text{ such that } i, j \leq 4,$$

$$\rho(x_i, x_5) = 1 + \epsilon \quad \text{for all } i \leq 4,$$

where $\epsilon > 0$ is some small constant. Now consider changing the metric to ρ' such that

$$\rho'(x_1, x_2) = \rho'(x_3, x_4) = \alpha,$$

for some small $\alpha > 0$. All other distances stay the same. Show that for $k = 2$, there exists α, ϵ such that this example disproves consistency.

Question 6. Consider the distribution density $f_\lambda(x)$ with parameter $\lambda > 0$ over the domain $[0, \infty)$, which is defined by

$$f_\lambda(x) := C_\lambda e^{-\lambda x},$$

where C_λ is a constant depending only on λ.

(a) Find a formula for the value of C_λ as a function of λ. Prove your claim.

(b) Let $S = \{x_1, \ldots, x_m\}$ be an i.i.d. sample from the distribution with the density f_λ, where λ is unknown. Derive the Maximum Likelihood Estimator for the value of λ. Prove all your claims.