Exercise 1. (Linear regression)

You are given a data matrix $X \in \mathbb{R}^{N\times(d+1)}$:

$$X := \begin{pmatrix} 1 & x_1^1 & x_2^1 & \dots & x_d^1 \\ 1 & x_1^2 & x_2^2 & \dots & x_d^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^N & x_2^N & \dots & x_d^N \end{pmatrix}.$$

We also define a weight vector $w \in \mathbb{R}^{d+1}$:

$$w := \begin{pmatrix} b \\ w_1 \\ \vdots \\ w_d \end{pmatrix},$$

where b is the bias, and $w_1, \dots, w_d$ are the weights corresponding to the features.

1. Verify that $J(w) = \frac{1}{N}\sum_{i=1}^{N}(w^\top x^i - y^i)^2$ can be equivalently written as $J(w) = \frac{1}{N}(Xw - y)^\top(Xw - y)$.
2. Now consider the regularized loss function $J(w) = \frac{1}{N}\sum_{i=1}^{N}(w^\top x^i - y^i)^2 + \lambda(w_1^2 + \ldots + w_d^2)$. Find the gradient and Hessian of the regularized loss function with respect to the model parameters.

   where $I_{d\times d}$ is the identity matrix of size $d \times d$ and $\mathbf{0_d}$ is zero vector of size $d \times 1$.
3. Now, we consider unregulated loss function denoted by $J(w) = \frac{1}{N}(Xw - y)^\top(Xw - y)$. Assume $X^\top X$ is invertible. Find the optimal weight vector $w^\star$ that minimizes $J(w)$, i.e. $w^\star = \arg\min_w J(w)$.
4. Now we consider k feature functions $\{\phi_i(x)\}_{i=1}^k$, where $\phi_i : \mathbb{R}^d \to \mathbb{R}$. What is the corresponding data matrix?

Exercise 2. (Linear regression and cross-validation)

A server is one of the main energy-consuming components of a data center. It has been found that the variables CPU denoted by $x_1 \in \mathbb{R}$, and the memory load denoted by $x_2 \in \mathbb{R}$, are two of the main contributing factors to the energy consumption of a server. You have made measurements of the CPU x_1^i, memory loads x_2^i, and energy consumption y^i, for $i = 1, 2, \dots, 2000$ instances and aim to use linear regression to come up with a function that predicts a server's energy consumption. You randomly select 400 data samples for testing and the rest for training.

1. You have found that increasing CPU and memory load have a multiplicative effect on energy consumption. Hence, you define a new feature : $\phi(x_1, x_2) = x_1 x_2$. Write the equation for a linear predictor in terms of the features $x_1, x_2, \phi(x_1, x_2)$.
2. Write the regularized mean-square loss function for identifying the parameters of the model; use $\lambda \in \mathbb{R}$ for regularization.
3. Derive the gradient of the loss function with respect to the linear regression parameters.
4. Which is likely to decrease the training error: increasing or decreasing λ and why?

5. Assume we choose the optimal λ using 5-fold cross-validation. Let $\hat{y}^i$ denote the prediction and y^i denote the actual server energy consumption for a given data point. How would you compute the mean validation error over the 5-folds?

6. Based on the result of 5-fold cross validation on the 1600 data points in the training set shown below, for which λ is the test error more likely to be similar to the validation error?

model	mean validation error	variance of error
λ_1	2.35	9.42
λ_2	1.30	4.16
λ_3	1.76	3.50

Exercise 3. (Train vs. test datasets)

You want to predict whether or not a product fails based on historical data on the amount of the different materials used to make the product. You have a set of 1,000 data points, where each data point contains the amount of the 5 different materials ($x^i \in \mathbb{R}^d$) and the information on failure or non-failure of the product, $y^i \in \{0, 1\}$. Here 0 denotes no failure and 1 denotes failure.

1. Determine what type of learning problem you are dealing with: supervised or unsupervised learning? If supervised learning, is it a regression or classification problem?

2. You randomly split the data such that 800 data points are used as the training set and 200 are used as the test set. Why shouldn't we use all the 1,000 available data points to train the model?

3. After training your model, you observe that it performs well on the training data, but poorly on the test data. What is one possible explanation? What could you try to improve the performance on the test data?