INFR10086 Machine Learning (MLG) Semester 2, 2025/6

Exercises 3

Lecturer: Hao Tang

Exercise 1. Given a data set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ whose elements $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, show that the mean-squared error

$$L = \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \tag{1}$$

is convex in $\mathbf{w}$.

Exercise 2. Given a data set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ whose elements $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, show that the optimal solution for the mean-squared error

$$L = \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \tag{2}$$

is

$$\mathbf{w} = (X^\top X)^{-1} X^\top \mathbf{y}, \tag{3}$$

where

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{4}$$

Exercise 3. In this question, we are going to extend the previous result to predicting a vector instead of just a value. Given a data set $\{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$ whose elements $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{y}_i \in \mathbb{R}^q$, show that the optimal solution for the mean-squared error

$$L = \sum_{i=1}^{n} \|\mathbf{y}_i - W\mathbf{x}_i\|^2 \tag{5}$$

is

$$W = Y^\top X (X^\top X)^{-1}, \tag{6}$$

where

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_n^\top \end{bmatrix} \tag{7}$$

Exercise 4. In this question, we are going to train a simple regression model on MNIST to generate an image from a label. MNIST is a labeled data set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^{784}$ is a 28×28 image and $y_i \in \{0, \dots, 9\}$ is a label. Our goal is to train a function $f(y) = W\mathbf{1}_y$, where $W \in \mathbb{R}^{784 \times 10}$ and $\mathbf{1}_y \in \mathbb{R}^{10}$ is a one-hot vector where the y-th coordinate is 1 and is otherwise 0.

1. Write a python script that trains a model $f(y) = W\mathbf{1}_y + \mathbf{b}$ by minimizing the mean-squared error using stochastic gradient descent. You can modify the training script from Tutorial 2.

2. Given the trained f, visualize the rows of W. Each row of W has 784 dimensions. Reshape each row to 28×28 and plot them with `matplotlib.imshow`. Note that $W \in \mathbb{R}^{784 \times 10}$, so you should see 10 images.

3. Write a script that computes the mean of each individual digits, and visualize them with `matplotlib.imshow`. Compare them with those from the rows of W. Do they look nearly identical?

4. Given a data set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, show that the optimal W to the mean-squared error

   $$L = \sum_{i=1}^{n} \|\mathbf{x}_i - (W\mathbf{1}_{y_i} + \mathbf{b})\|^2 \tag{8}$$

   consists of the means of each digit and the optimal $\mathbf{b}$ is the mean of all digits. In other words, the k-th row of the optimal W has the form

   $$\mathbf{w}_k = \frac{\sum_{i=1}^{n} \nVdash_{y_i=k} \mathbf{x}_i}{\sum_{i=1}^{n} \nVdash_{y_i=k}}. \tag{9}$$

   This problem is known as regression to the mean.

5. Since we standardize the images already before training, confirm that the $\mathbf{b}$ after training is nearly a 0 vector by showing that its norm is close to 0.