Homework 3

Jimin Chae
4190.408 001, Artificial Intelligence

November 7, 2025

Problem 2. Linear Regression

(a) Let function L be the loss function for linear regression.

$$L = \frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\bar{\mathbf{x}}_n)^2$$

The gradient of L with respect to $\mathbf{w}$ is:

$$\begin{aligned}
\nabla L &= \nabla\left(\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\bar{\mathbf{x}}_n)^2\right)\\
&= \nabla\left(\frac{1}{2}\sum_{n=1}^{N}(t_n^2 - 2t_n\mathbf{w}^T\bar{\mathbf{x}}_n + (\mathbf{w}^T\bar{\mathbf{x}}_n)^2)\right)\\
&= \nabla\left(\frac{1}{2}\sum_{n=1}^{N}t_n^2 - \sum_{n=1}^{N}t_n\mathbf{w}^T\bar{\mathbf{x}}_n + \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\bar{\mathbf{x}}_n)^2\right)\\
&= \nabla\left(\frac{1}{2}\sum_{n=1}^{N}t_n^2 - \sum_{n=1}^{N}t_n\mathbf{w}^T\bar{\mathbf{x}}_n + \frac{1}{2}\mathbf{w}^T\sum_{n=1}^{N}\bar{\mathbf{x}}_n\bar{\mathbf{x}}_n^T\mathbf{w}\right)\\
&= -\sum_{n=1}^{N}t_n\bar{\mathbf{x}}_n + \sum_{n=1}^{N}\bar{\mathbf{x}}_n\bar{\mathbf{x}}_n^T\mathbf{w}\\
&= 0
\end{aligned}$$

Therefore, the gradient of L with respect to $\mathbf{w}$ is 0 when:

$$\sum_{n=1}^{N}\bar{\mathbf{x}}_n\bar{\mathbf{x}}_n^T\mathbf{w} = \sum_{n=1}^{N}t_n\bar{\mathbf{x}}_n$$

This represents the normal equation for linear regression for $\mathbf{A} = \sum_{n=1}^{N}\bar{\mathbf{x}}_n\bar{\mathbf{x}}_n^T$ and $\mathbf{b} = \sum_{n=1}^{N}t_n\bar{\mathbf{x}}_n$:

$$\mathbf{A}\mathbf{w} = \mathbf{b}$$

(b) By **(a)**, we have:

$$\begin{aligned}
\mathbf{A} &= \sum_{n=1}^{2}\bar{\mathbf{x}}_n\bar{\mathbf{x}}_n^T\\
&= \bar{\mathbf{x}}_1\bar{\mathbf{x}}_1^T + \bar{\mathbf{x}}_2\bar{\mathbf{x}}_2^T\\
&= \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix} + \begin{pmatrix}1 & \epsilon\\\epsilon & \epsilon^2\end{pmatrix}\\
&= \begin{pmatrix}2 & \epsilon\\\epsilon & \epsilon^2\end{pmatrix}
\end{aligned}$$

And, $\mathbf{b}$ is:

$$\begin{aligned}
\mathbf{b} &= \sum_{n=1}^{2}t_n\bar{\mathbf{x}}_n\\
&= t_1\bar{\mathbf{x}}_1 + t_2\bar{\mathbf{x}}_2\\
&= 1\cdot\begin{bmatrix}1\\0\end{bmatrix} + 1\cdot\begin{bmatrix}1\\\epsilon\end{bmatrix}\\
&= \begin{bmatrix}2\\\epsilon\end{bmatrix}
\end{aligned}$$

Therefore, the normal equation for linear regression is:

$$\begin{pmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{pmatrix} \mathbf{w} = \begin{bmatrix} 2 \\ \epsilon \end{bmatrix}$$

Solving this equation, we know that $\mathbf{A}$ is invertible, so we can get:

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{b}$$
$$= \begin{pmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{pmatrix}^{-1} \begin{bmatrix} 2 \\ \epsilon \end{bmatrix}$$
$$= \begin{pmatrix} \frac{1}{2} & -\frac{1}{2\epsilon} \\ -\frac{1}{2\epsilon} & \frac{1}{\epsilon^2} \end{pmatrix} \begin{bmatrix} 2 \\ \epsilon \end{bmatrix}$$
$$= \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Therefore, the solution to the linear regression problem is:

$$\mathbf{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

(c) The normal equation for linear regression is:

$$\begin{pmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{pmatrix} \mathbf{w} = \begin{bmatrix} 2 + \epsilon \\ \epsilon \end{bmatrix}$$

Solving this equation, we know that $\mathbf{A}$ is invertible, so we can get:

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{b}$$
$$= \begin{pmatrix} 2 & \epsilon \\ \epsilon & \epsilon^2 \end{pmatrix}^{-1} \begin{bmatrix} 2 + \epsilon \\ \epsilon \end{bmatrix}$$
$$= \begin{pmatrix} \frac{1}{2} & -\frac{1}{2\epsilon} \\ -\frac{1}{2\epsilon} & \frac{1}{\epsilon^2} \end{pmatrix} \begin{bmatrix} 2 + \epsilon \\ \epsilon \end{bmatrix}$$
$$= \begin{bmatrix} 1 + \epsilon \\ -1 \end{bmatrix}$$

Therefore, the solution to the linear regression problem is:

$$\mathbf{w} = \begin{bmatrix} 1 + \epsilon \\ -1 \end{bmatrix}$$

(d) When $\epsilon = 0.1$, we have:

$$\mathbf{w}_{(b)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$
$$\mathbf{w}_{(c)} = \begin{bmatrix} 1 + 0.1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1.1 \\ -1 \end{bmatrix}$$

Therefore, the difference between the two solutions is:

$$\Delta \mathbf{w} = \mathbf{w}_{(b)} - \mathbf{w}_{(c)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1.1 \\ -1 \end{bmatrix} = \begin{bmatrix} -0.1 \\ 1 \end{bmatrix}$$

Problem 3. Linear Regression with Regularization

(a) The $\mathbf{A}$ matrix is semi-positive definite because:

$$\mathbf{A} = \sum_{n=1}^{N} \bar{\mathbf{x}}_n \bar{\mathbf{x}}_n^T = \sum_{n=1}^{N} (\bar{\mathbf{x}}_n)^2 \geq 0$$

Therefore, each eigenvalue of $\mathbf{A}$ is non-negative for all λ_i ($1 \leq i \leq N$). At that time, the eigenvalues of $(\mathbf{A} + \lambda \mathbf{I})$ are $\lambda_i + \lambda$ for all λ_i ($1 \leq i \leq N$).

$$\text{eig}(\mathbf{A} + \lambda \mathbf{I}) = \lambda_i + \lambda \quad (1 \leq i \leq N)$$

The eigenvalues of inverse matrix is the reciprocal of the eigenvalues of the original matrix. Therefore, the eigenvalues of $(\mathbf{A} + \lambda \mathbf{I})^{-1}$ are $\frac{1}{\lambda_i + \lambda}$ for all λ_i ($1 \leq i \leq N$).

$$\text{eig}((\mathbf{A} + \lambda \mathbf{I})^{-1}) = \frac{1}{\lambda_i + \lambda} \quad (1 \leq i \leq N)$$

Therefore, for the definition of **spectral radius**, we have:

$$\rho(\mathbf{A} + \lambda \mathbf{I})^{-1} = \max_{1 \leq i \leq N} \left| \frac{1}{\lambda_i + \lambda} \right|$$
$$= \frac{1}{\min_{1 \leq i \leq N}(\lambda_i + \lambda)}$$
$$\leq \frac{1}{\lambda}$$

(b) First, $\mathbf{w}$ in **Problem 2-(b)** is:

$$\mathbf{w}_{(b)} = \frac{1}{\epsilon^2 + (2 + \epsilon^2)\lambda + \lambda^2} \begin{bmatrix} \epsilon^2 + \lambda & -\epsilon \\ -\epsilon & 2 + \lambda \end{bmatrix} \begin{bmatrix} 2 \\ \epsilon \end{bmatrix}$$
$$= \frac{1}{\epsilon^2 + (2 + \epsilon^2)\lambda + \lambda^2} \begin{bmatrix} \epsilon^2 + 2\lambda \\ \epsilon\lambda \end{bmatrix}$$
$$\approx \begin{bmatrix} 0.97345 \\ 0.04425 \end{bmatrix}$$

And $\mathbf{w}$ in **Problem 2-(c)** is:

$$\mathbf{w}_{(c)} = \frac{1}{\epsilon^2 + (2 + \epsilon^2)\lambda + \lambda^2} \begin{bmatrix} \epsilon^2 + \lambda & -\epsilon \\ -\epsilon & 2 + \lambda \end{bmatrix} \begin{bmatrix} 2 + \epsilon \\ \epsilon \end{bmatrix}$$
$$\approx \begin{bmatrix} 1.02655 \\ -0.04425 \end{bmatrix}$$

Therefore, the difference between the two solutions is:

$$\Delta\mathbf{w} = \mathbf{w}_{(b)} - \mathbf{w}_{(c)}$$
$$= \begin{bmatrix} 0.97345 \\ 0.04425 \end{bmatrix} - \begin{bmatrix} 1.02655 \\ -0.04425 \end{bmatrix}$$
$$= \begin{bmatrix} -0.0531 \\ 0.0885 \end{bmatrix}$$

Therefore, the difference between the two solutions is:

$$\Delta\mathbf{w} = \mathbf{w}_{(b)} - \mathbf{w}_{(c)} = \begin{bmatrix} -0.0531 \\ 0.0885 \end{bmatrix}$$

(c) Let ϵ is a small noise in training data. When a small noise is added to the training data, the weight vector $\mathbf{w}$ will be changed. And, we know that the difference between the two solutions at **Problem 3-(b)** is smaller then **Probelm 2-(d)**:

$$||\Delta\mathbf{w}_{\textbf{Problem 3-(b)}}|| < ||\Delta\mathbf{w}_{\textbf{Problem 2-(d)}}||$$

Then we can say that the regularization term λ is effective in preventing overfitting.