Linear regression

Machine Learning II (2023-2024)
UMONS

1 Exercise 1

Consider the hat matrix $H = X(X^T X)^{-1} X^T$, where X is an n by $d+1$ matrix, and $X^T X$ is invertible.

(a) Show that H is symmetric.

(b) Show that H is a projection matrix, i.e. $H^2 = H$. So $\hat{y}$ is the projection of y onto some space. What is the space?

(c) Show that $H^k = H$ for any positive integer k.

(d) If I is the identity matrix of size n, show that $(I - H)^k = I - H$ for any positive integer k.

(e) Show that $\text{trace}(H) = d + 1$, where the trace is the sum of diagonal elements. [**Hint:** $\text{trace}(AB) = \text{trace}(BA)$]

Solution

(a) To show H is symmetric, we have to show $H^T = H$.

$$\begin{aligned} H^T &= (X(X^T X)^{-1} X^T)^T \\ &= X(X^T X)^{-T} X^T \\ &= X(X^T X)^{-1} X^T \\ &= H \end{aligned}$$

(b) In the finite-dimensional case, a square matrix P is called a projection matrix if it is equal to its square, i.e., if $P = P^2$.

$$\begin{aligned} H^2 &= (X(X^T X)^{-1} X^T)(X(X^T X)^{-1} X^T) \\ &= X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T \\ &= X(X^T X)^{-1} X^T \\ &= H \end{aligned}$$

So, H is a projection matrix. $\hat{y}$ is the projection of y onto the space spanned by X.

(c) We have to show that $H^k = H$ for $k = 1, 2, 3, \ldots$. We will prove that by using induction.

- For $k = 1$, $H^1 = H$.
- For $k = 2$, $H^2 = H$.
- Consider, it is true for k, $H^k = H$.

- For $k = k + 1$,

$$\begin{aligned} H^{k+1} &= H^k \cdot H \\ &= H \cdot H \\ &= H^2 \\ &= H \end{aligned}$$

(d) If I is the identity matrix of size n, we have to show that $(I - H)^k = I - H$ for $k = 1, 2, 3, \ldots$.

- For $k = 1$, $(I - H)^1 = I - H$.
- For $k = 2$,

$$\begin{aligned} (I - H)^2 &= (I - H)(I - H) \\ &= I - 2H + H^2 \\ &= I - 2H + H \\ &= I - H \end{aligned}$$

- Consider, it is true for k, $(I - H)^k = I - H$.
- For $k + 1$,

$$\begin{aligned} (I - H)^{k+1} &= (I - H)^k \cdot (I - H) \\ &= (I - H) \cdot (I - H) \\ &= (I - H)^2 \\ &= (I - H) \end{aligned}$$

(e) We have to prove $\text{trace}(H) = d + 1$,

$$\begin{aligned} \text{trace}(H) &= \text{trace}(X(X^T X)^{-1} X^T) \\ &= \text{trace}(AB) \qquad [\text{where } A = X(X^T X)^{-1} \text{ and } B = X^T] \\ &= \text{trace}(BA) \qquad [\text{Using Hint}] \\ &= \text{trace}(X^T X (X^T X)^{-1}) \\ &= \text{trace}(I_{d+1}) \qquad [\text{As } X \text{ is } n \times d + 1 \text{ matrix}] \\ &= d + 1 \end{aligned}$$

2 Exercise 2

Consider a noisy target $y = \mathbf{w}^{*T}\mathbf{x} + \epsilon$ for generating the data, where ϵ is a noise term with zero mean and σ^2 variance, independently generated for every example $(\mathbf{x}, y)$. The expected error of the best possible linear fit to this target is thus σ^2.

For the data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, denote the noise in y_i as ϵ_i and let $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]^T$; assume that $X^T X$ is invertible. By following the steps below, show that the expected in-sample error of linear regression with respect to $\mathcal{D}$ is given by

$$\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] = \sigma^2\Big(1 - \frac{d+1}{n}\Big)$$

(a) Show that the in-sample estimate of $\mathbf{y}$ is given by $\hat{\mathbf{y}} = X\mathbf{w}^* + H\boldsymbol{\epsilon}$.

(b) Show that the in-sample error vector $\hat{\mathbf{y}} - \mathbf{y}$ can be expressed by a matrix times $\boldsymbol{\epsilon}$. What is the matrix?

(c) Express $E_{in}(\mathbf{w}_{lin})$ in terms of ϵ using (b), and simplify the expression using Exercise 1(d).

(d) Prove that $\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] = \sigma^2\Big(1 - \frac{d+1}{n}\Big)$ using (c) and their independence of $\epsilon_1, \cdots, \epsilon_n$. [**Hint:** The sum of the diagonal elements of a matrix (the trace) will play a role. See Exercise 1(e)]

For the expected out-of-sample error, we take a special case which is easy to analyze. Consider a test data set $\mathcal{D}_{test} = \{(\mathbf{x}_1, y_1'), \ldots, (\mathbf{x}_n, y_n')\}$, which shares the same input vector $\mathbf{x}_i$ with $\mathcal{D}$ but with different realization of the noise terms. Denote the noise in y_i' as ϵ_i' and let $\boldsymbol{\epsilon}' = [\epsilon_1', \epsilon_2', \ldots, \epsilon_n']^T$. Define $E_{test}(\mathbf{w}_{lin}$ to be the average squared error on $\mathcal{D}_{test}$.

(e) Prove that $\mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[E_{test}(\mathbf{w}_{lin})] = \sigma^2\Big(1 + \frac{d+1}{n}\Big)$.

The special test error E_{test} is a very restricted case of the general out-of-sample error. Some detailed analysis shows that similar results can be obtained for the general case, as shown in Exercise 3.

Solution
We have,

$$\begin{aligned}\mathcal{D} &= \{(\mathbf{x}_i, y_i\}_{i=1}^n \qquad [\text{where } \mathbf{x}_i \in \mathbb{R}^{d+1} \text{ and } y_i \in \mathbb{R}] \\ &= \{X, \mathbf{y}\} \qquad [\text{where } X \in \mathbb{R}^{n\times d+1} \text{ and } \mathbf{y} \in \mathbb{R}^{n\times 1}]\end{aligned}$$

Then the in-sample error can be written as,

$$\begin{aligned}E_{in}(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^{N}(y_i - h(\mathbf{x}_i))^2 \\ &= ||\mathbf{y} - X\mathbf{w}||^2\end{aligned}$$

Now, for linear regression,

$$\mathbf{w}_{lin} = \hat{\mathbf{w}} = (X^T X)^{-1} X^T \mathbf{y}$$

Therefore,

$$\begin{aligned}\hat{\mathbf{y}} = X\mathbf{w}_{lin} &= X\hat{\mathbf{w}} \\ &= X((X^T X)^{-1} X^T \mathbf{y}) \\ &= H\mathbf{y}\end{aligned}$$

(a) The in-sample error estimate is

$$\begin{aligned}\hat{\mathbf{y}} &= H\mathbf{y} \\ &= H(X\mathbf{w}^* + \boldsymbol{\epsilon}) \\ &= HX\mathbf{w}^* + H\boldsymbol{\epsilon} \\ &= (X(X^T X)^{-1} X^T)X\mathbf{w}^* + H\boldsymbol{\epsilon} \\ &= X\mathbf{w}^* + H\boldsymbol{\epsilon}\end{aligned}$$

(b) The in-sample error vector $\hat{y} - y$ can be expressed as below.

$$\begin{aligned}\hat{\mathbf{y}} - \mathbf{y} &= (X\mathbf{w}^* + H\boldsymbol{\epsilon}) - (X\mathbf{w}^* + \boldsymbol{\epsilon}) \\ &= H\boldsymbol{\epsilon} - \boldsymbol{\epsilon} \\ &= (H - I)\boldsymbol{\epsilon}\end{aligned}$$

(c)

$$\begin{aligned}E_{in}(\mathbf{w}_{lin}) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{w}_{lin}^T\mathbf{x}_i)^2 \\ &= \frac{1}{n}||\mathbf{y} - \hat{\mathbf{y}}||^2 \\ &= \frac{1}{n}||(H - I)\boldsymbol{\epsilon}||^2 \\ &= \frac{1}{n}\boldsymbol{\epsilon}^T(H - I)^T(H - I)\boldsymbol{\epsilon} \\ &= \frac{1}{n}\boldsymbol{\epsilon}^T(H^T - I)(H - I)\boldsymbol{\epsilon}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\boldsymbol{\epsilon}^T(H-I)(H-I)\boldsymbol{\epsilon} \\
&= \frac{1}{n}\boldsymbol{\epsilon}^T(H-I)^2\boldsymbol{\epsilon} \\
&= \frac{1}{n}\boldsymbol{\epsilon}^T(I-H)^2\boldsymbol{\epsilon} \\
&= \frac{1}{n}\boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon} \qquad \text{[Using Exercise 1(d)]}
\end{aligned}
$$

(d)

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] &= \mathbb{E}_{\mathcal{D}}[\frac{1}{n}\boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon}] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}[\frac{1}{n}\boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon}] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T H\boldsymbol{\epsilon}] \\
&= \frac{1}{n}(\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] - \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T H\boldsymbol{\epsilon}]) \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] - \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T H\boldsymbol{\epsilon}]) \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] - \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\text{trace}(\boldsymbol{\epsilon}^T H\boldsymbol{\epsilon})] \quad \text{[As } \epsilon \text{ is } n\times 1 \text{ matrix and } H \text{ is } n\times n \text{ matrix]} \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] - \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\text{trace}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T H)] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] - \frac{1}{n}\text{trace}(\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T H]) \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] - \frac{1}{n}\text{trace}(\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]H)
\end{aligned}
$$

Now,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] &= \mathbb{E}_{\boldsymbol{\epsilon}}[\begin{pmatrix}\epsilon_1 & \epsilon_2 & \cdots & \epsilon_n\end{pmatrix}\begin{pmatrix}\epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n\end{pmatrix}] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}[\sum_{i=1}^{n}\epsilon_i^2] = \sum_{i=1}^{n}\mathbb{E}_{\epsilon}[\epsilon_i^2] = n\sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{\epsilon}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] &= \mathbb{E}_{\epsilon}[\begin{pmatrix}\epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n\end{pmatrix}\begin{pmatrix}\epsilon_1 & \epsilon_2 & \cdots & \epsilon_n\end{pmatrix}] \\
&= \mathbb{E}_{\epsilon}[\begin{pmatrix}\epsilon_1^2 & \cdots & \epsilon_1\epsilon_n \\ \vdots & \ddots & \vdots \\ \epsilon_n\epsilon_n & \cdots & \epsilon_n^2\end{pmatrix}] \\
&= \begin{pmatrix}\sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2\end{pmatrix}
\end{aligned}
$$

$$= \sigma^2 I_n$$

Hence,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{in}(\mathbf{w}_{lin})] &= \frac{1}{n}n\sigma^2 - \frac{1}{n}\text{trace}(\sigma^2 I_n H) \\
&= \sigma^2 - \frac{1}{n}\text{trace}(\sigma^2 H) \\
&= \sigma^2 - \frac{\sigma^2}{n}\text{trace}(H) \\
&= \sigma^2 - \frac{\sigma^2}{n}(d+1) \qquad \text{[Using Exercise 1(e)]} \\
&= \sigma^2(1 - -\frac{(d+1)}{n})
\end{aligned}$$

(e)

$$\begin{aligned}
\mathcal{D}_{test} &= \{(\mathbf{x}_i, y_i'\}_{i=1}^n \qquad \text{[where } \mathbf{x}_i \in \mathbb{R}^{d+1} \text{ and } y_i' \in \mathbb{R}] \\
&= \{X, \mathbf{y}'\} \qquad \text{[where } X \in \mathbb{R}^{n \times d+1} \text{ and } \mathbf{y}' \in \mathbb{R}^{n \times 1}]
\end{aligned}$$

So, we have

- For $\mathcal{D}$, $\mathbf{y} = X\mathbf{w}^* + \epsilon$
- For $\mathcal{D}_{test}$, $\mathbf{y}' = X\mathbf{w}^* + \epsilon'$

Now,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D},\mathcal{D}_{test}}[E_{test}(\mathbf{w}_{lin})] &= \frac{1}{n}\mathbb{E}_{\mathcal{D},\mathcal{D}_{test}}[||\mathbf{y}' - \hat{\mathbf{y}}||^2] \\
&= \frac{1}{n}\mathbb{E}_{\mathbf{y},\mathbf{y}'}[||\mathbf{y}' - \hat{\mathbf{y}}||^2] \\
&= \frac{1}{n}\mathbb{E}_{\mathbf{y},\mathbf{y}'}[||X\mathbf{w}^* + \boldsymbol{\epsilon}' - (X\mathbf{w}^* + H\boldsymbol{\epsilon})||^2] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[||\boldsymbol{\epsilon}' - H\boldsymbol{\epsilon}||^2] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[(\boldsymbol{\epsilon}' - H\boldsymbol{\epsilon})^T(\boldsymbol{\epsilon}' - H\boldsymbol{\epsilon})] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[(\boldsymbol{\epsilon}'^T - \boldsymbol{\epsilon}^T H^T)(\boldsymbol{\epsilon}' - H\boldsymbol{\epsilon})] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[(\boldsymbol{\epsilon}'^T\boldsymbol{\epsilon}' - \boldsymbol{\epsilon}^T H^T\boldsymbol{\epsilon}' - \boldsymbol{\epsilon}'^T H\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T H^T H\boldsymbol{\epsilon})] \\
&= \frac{1}{n}(\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[(\boldsymbol{\epsilon}'^T\boldsymbol{\epsilon}'] - \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}^T H^T\boldsymbol{\epsilon}'] - \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}'^T H\boldsymbol{\epsilon}] + \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}^T H^T H\boldsymbol{\epsilon})]) \\
&= \frac{1}{n}(\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[(\boldsymbol{\epsilon}'^T\boldsymbol{\epsilon}'] + \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}^T H\boldsymbol{\epsilon})]) \\
&= \frac{1}{n}(n\sigma^2) + \frac{1}{n}(\sigma^2(d+1)) = \sigma^2\left(1 + \frac{d+1}{n}\right)
\end{aligned}$$

Using the fact that $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}'$ are independent of each other and ϵ_i and ϵ_i' are independent among themselves. Therefore, $\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}^T H^T\boldsymbol{\epsilon}'] = \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}'^T H\boldsymbol{\epsilon}] = 0$ and $H^T H = H$ from Exercise 1(c)

3 Exercise 3

Consider the linear regression problem setup in Exercise 2, where the data comes from a genuine linear relationship with added noise. The noise for the different data points is assumed to be iid with zero mean and variance σ^2. Assume that the 2^{nd} moment matrix $\Sigma = \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T]$ is non-singular. Follow the steps below to show that, with high probability, the out-of-sample error on average is

$$E_{out}(\mathbf{w}_{lin}) = \sigma^2\Big(1 + \frac{d+1}{n} + o(\frac{1}{n})\Big).$$

(a) For a test point $\mathbf{x}$, show that the error $y - g(\mathbf{x})$ is

$$\epsilon - \mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon},$$

where ϵ is the noise realization for the test point and $\boldsymbol{\epsilon}$ is the vector of noise realizations on the data.

(b) Take the expectation with respect to the test point, i.e., $\mathbf{x}$ and ϵ, to obtain an expression for E_{out}. Show that

$$E_{out} = \sigma^2 + \text{trace}(\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1})$$

[**Hint:** $a = \text{trace}(a)$ for any scalar a; $\text{trace}(AB) = \text{trace}(BA)$; expectation and trace commute.]

(c) What is $\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$?

(d) Take the expectation with respect to $\boldsymbol{\epsilon}$ to show that, on average,

$$E_{out} = \sigma^2 + \frac{\sigma^2}{n}\text{trace}(\Sigma(\frac{1}{n}X^TX)^{-1}).$$

Note that $\frac{1}{n}X^TX = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$ is an n sample estimate of Σ. So $\frac{1}{n}X^TX \approx \Sigma$. If $\frac{1}{n}X^TX = \Sigma$, then what is E_{out} on average?

(e) Show that (after taking the expectation over the data noise) with high probability,

$$E_{out} = \sigma^2(1 + \frac{d+1}{n} + o(\frac{1}{n})).$$

[**Hint:** By the law of large numbers $\frac{1}{n}X^TX$ converges in probability to Σ, and so by continuity of the inverse at Σ, $(\frac{1}{n}X^TX)^{-1}$ converges in probability to Σ^{-1}.]

Solution

(a) For a test point $\mathbf{x}_i$,

$$
\begin{aligned}
y_i - g(\mathbf{x}_i) &= \mathbf{x}_i^T\mathbf{w}^* + \epsilon_i - \mathbf{x}_i^T\hat{\mathbf{w}} \\
&= \mathbf{x}_i^T\mathbf{w}^* + \epsilon_i - \mathbf{x}_i^T(X^TX)^{-1}X^Ty \\
&= \mathbf{x}_i^T\mathbf{w}^* + \epsilon_i - \mathbf{x}_i^T(X^TX)^{-1}X^T(X\mathbf{w}^* + \boldsymbol{\epsilon}) \\
&= \mathbf{x}_i^T\mathbf{w}^* + \epsilon_i - \mathbf{x}_i^T(X^TX)^{-1}X^TX\mathbf{w}^* - \mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon} \\
&= \mathbf{x}_i^T\mathbf{w}^* + \epsilon_i - \mathbf{x}_i^T\mathbf{w}^* - \mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon} \\
&= \epsilon_i - \mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}
\end{aligned}
$$

(b) We can compute E_{out} by taking expectation of $(y_i - g(\mathbf{x}_i))^2$ w.r.t. $\mathbf{x}_i$ and ϵ_i.

$$
\begin{aligned}
E_{out} &= \mathbb{E}_{\mathbf{x}_i,\epsilon_i}[(y_i - g(\mathbf{x}_i))^2] \\
&= \mathbb{E}_{\mathbf{x}_i,\epsilon_i}[(\epsilon_i - \mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2] \\
&= \mathbb{E}_{\mathbf{x}_i,\epsilon_i}[\epsilon_i^2 - 2\epsilon_i\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon} + (\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2] \\
&= \mathbb{E}_{\mathbf{x}_i,\epsilon_i}[\epsilon_i^2] - \mathbb{E}_{\mathbf{x}_i,\epsilon_i}[2\epsilon_i\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}] + \mathbb{E}_{\mathbf{x}_i,\epsilon_i}[(\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2] \\
&= \mathbb{E}_{\mathbf{x}_i,\epsilon_i}[\epsilon_i^2] + \mathbb{E}_{\mathbf{x}_i,\epsilon_i}[(\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2] \qquad [\text{As } \mathbb{E}_{\epsilon_i}[\epsilon_i] = 0] \\
&= \sigma^2 + \mathbb{E}_{\mathbf{x}_i}[(\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2] \\
&= \sigma^2 + \mathbb{E}_{\mathbf{x}_i}[\text{trace}((\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2)] \qquad [\text{As } (\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2 \text{ is a scalar}] \\
&= \sigma^2 + \mathbb{E}_{\mathbf{x}_i}[(\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})(\boldsymbol{\epsilon}^TX(X^TX)^{-1}\mathbf{x}_i)] \\
&= \sigma^2 + \mathbb{E}_{\mathbf{x}_i}[\text{trace}(\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1}\mathbf{x}_i)] \\
&= \sigma^2 + \mathbb{E}_{\mathbf{x}_i}[\text{trace}(\mathbf{x}_i\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1})] \\
&= \sigma^2 + \text{trace}(\mathbb{E}_{\mathbf{x}_i}[\mathbf{x}_i\mathbf{x}_i^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1}]) \\
&= \sigma^2 + \text{trace}(\mathbb{E}_{\mathbf{x}_i}[\mathbf{x}_i\mathbf{x}_i^T]\mathbb{E}_{\mathbf{x}_i}[(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1}]) \\
&= \sigma^2 + \text{trace}(\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1})
\end{aligned}
$$

(c)

$$
\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 \times I_n
$$

(d) By taking expectation w.r.t. $\boldsymbol{\epsilon}$, we obtain,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}[E_{out}] &= \mathbb{E}_{\boldsymbol{\epsilon}}[\sigma^2 + \text{trace}(\Sigma(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1})] \\
&= \sigma^2 + \text{trace}(\Sigma(X^TX)^{-1}X^T\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]X(X^TX)^{-1}) \\
&= \sigma^2 + \text{trace}(\Sigma(X^TX)^{-1}X^T\sigma^2 I_nX(X^TX)^{-1}) \\
&= \sigma^2 + \sigma^2\text{trace}(\Sigma(X^TX)^{-1}X^TX(X^TX)^{-1}) \\
&= \sigma^2 + \sigma^2\text{trace}\Big(\Sigma(X^TX)^{-1}\Big) \\
&= \sigma^2 + \sigma^2\frac{n}{n}\text{trace}\Big(\Sigma(X^TX)^{-1}\Big) \\
&= \sigma^2 + \frac{\sigma^2}{n}\text{trace}\Big(\Sigma\Big(\frac{X^TX}{n}\Big)^{-1}\Big) \\
&= \sigma^2 + \frac{\sigma^2}{n}\text{trace}(\Sigma\Sigma^{-1}) \qquad \Big[\Big(\frac{X^TX}{n}\Big) \approx \Sigma\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \sigma^2 + \frac{\sigma^2}{n}\text{trace}(I_{d+1}) \\
&= \sigma^2 + \frac{\sigma^2(d+1)}{n} \\
&= \sigma^2\Big(1 + \frac{(d+1)}{n}\Big)
\end{aligned}
$$

(e)

$$
\begin{aligned}
&\frac{X^TX}{n} \xrightarrow{P} \Sigma \\
&(\frac{X^TX}{n})^{-1} \xrightarrow{P} \Sigma^{-1} \\
&(\frac{X^TX}{n})^{-1} = \Sigma^{-1} + o(1)
\end{aligned}
$$

Now,

$$
\begin{aligned}
E_{out} &= \sigma^2 + \frac{\sigma^2}{n}\text{trace}\Big(\Sigma\Big(\frac{X^TX}{n}\Big)^{-1}\Big) \\
&= \sigma^2 + \frac{\sigma^2}{n}\text{trace}\Big(\Sigma(\Sigma^{-1} + o(1)) \\
&= \sigma^2 + \frac{\sigma^2}{n}[\text{trace}(I_{d+1}) + \text{trace}(\Sigma o(1))] \\
&= \sigma^2 + \frac{\sigma^2}{n}[(d+1) + o(1)] \\
&= \sigma^2(1 + \frac{d+1}{n} + o(\frac{1}{n}))
\end{aligned}
$$

4 Exercise 4

In a regression setting, assume the target function is linear, so $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}^*$, and $\mathbf{y} = Z\mathbf{w}^* + \boldsymbol{\epsilon}$, where the entries in $\boldsymbol{\epsilon}$ are zero mean, iid with variance σ^2. In this problem derive the bias and variance as follows.

(a) Show that the average function is $\bar{g}(\mathbf{x}) = f(\mathbf{x})$, no matter what the size of the data set. What is the bias?

(b) What is the variance? [**Hint:** Exercise 3]

Solution

(a)

$$
\begin{aligned}
y_n &= f(\mathbf{x}) + \epsilon_n = \mathbf{x}^T\mathbf{w}^* + \epsilon \\
\mathbf{y} &= X\mathbf{w}^* + \boldsymbol{\epsilon} \\
g^{\mathcal{D}}(\mathbf{x}) &= \mathbf{x}^T\hat{\mathbf{w}} \\
\hat{\mathbf{w}} &= (X^TX)^{-1}X^T\mathbf{y} \\
\bar{g}(\mathbf{x}) &= \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\mathbf{x})] \\
&= \mathbb{E}_{\mathcal{D}}[\mathbf{x}^T\hat{\mathbf{w}}] \\
&= \mathbb{E}_{\mathcal{D}}[\mathbf{x}^T(X^TX)^{-1}X^T\mathbf{y}] \\
&= \mathbb{E}_{\mathcal{D}}[\mathbf{x}^T(X^TX)^{-1}X^T(X\mathbf{w}^* + \boldsymbol{\epsilon})] \qquad [\text{where } \mathbf{y} = X\mathbf{w}^* + \boldsymbol{\epsilon}] \\
&= \mathbb{E}_{\mathcal{D}}[\mathbf{x}^T\mathbf{w}^* + \mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}[\mathbf{x}^T\mathbf{w}^* + \mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})] \\
&= \mathbf{x}^T\mathbf{w}^* \\
&= f(\mathbf{x}) \\
\text{Bias} &= \mathbb{E}_{\mathbf{x}}[(\mathbb{E}_{\epsilon_i}[y_i] - \bar{g}(\mathbf{x}))^2] \\
&= \mathbb{E}_{\mathbf{x}}[(f(\mathbf{x}) - f(\mathbf{x}))^2] \\
&= 0
\end{aligned}
$$

(b)

$$
\begin{aligned}
\text{Variance} &= \mathbb{E}_{\mathbf{x},\mathcal{D}}[(g^{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\mathbf{x})])^2] \\
&= \mathbb{E}_{\mathbf{x},\mathcal{D}}[(g^{\mathcal{D}}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2] \\
&= \mathbb{E}_{\mathbf{x},\mathcal{D}}[(\mathbf{x}^T\hat{\mathbf{w}} - \mathbf{x}^T\mathbf{w}^*)^2] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}}[(\mathbf{x}^T(X^TX)^{-1}X^T\mathbf{y} - \mathbf{x}^T\mathbf{w}^*)^2] \\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[(\mathbf{x}^T(X^TX)^{-1}X^T(X\mathbf{w}^* + \boldsymbol{\epsilon}) - \mathbf{x}^T\mathbf{w}^*)^2] \\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[(\mathbf{x}^T(X^TX)^{-1}X^TX\mathbf{w}^* + \mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon} - \mathbf{x}^T\mathbf{w}^*)^2] \\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[(\mathbf{x}^T\mathbf{w}^* + \mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon} - \mathbf{x}^T\mathbf{w}^*)^2] \\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[(\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2] \\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[\text{trace}(\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2)] \qquad [\text{As } (\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^2 \text{ is a scalar}] \\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[\text{trace}((\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})(\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})^T)] \\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[(\text{trace}(\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})(\boldsymbol{\epsilon}^TX(X^TX)^{-1}\mathbf{x}))] \\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[\text{trace}((\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon})(\boldsymbol{\epsilon}^TX(X^TX)^{-1}\mathbf{x}))]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[\text{trace}(\mathbf{x}\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1})] \\
&= \text{trace}(\mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}}[\mathbf{x}\mathbf{x}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1}]) \\
&= \text{trace}(\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T\mathbb{E}_{\boldsymbol{\epsilon}}[(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1}]]) \\
&= \text{trace}(\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T\sigma^2(X^TX)^{-1}]) \qquad [\text{where } \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}] = \sigma^2 I] \\
&= \text{trace}(\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T]\sigma^2(X^TX)^{-1}) \\
&= \sigma^2\text{trace}(\Sigma(X^TX)^{-1}) \\
&= \sigma^2\frac{n}{n}\text{trace}(\Sigma(X^TX)^{-1}) \\
&= \frac{\sigma^2}{n}\text{trace}(\Sigma(\frac{X^TX}{n})^{-1}) \\
&= \sigma^2\Big(\frac{d+1}{n} + o(\frac{1}{n})\Big) \qquad [\text{from Exercise 3(e)}]
\end{aligned}
$$

5 Exercise 5

In the text we derived that the linear regression solution weights must satisfy $X^T X\mathbf{w} = X\mathbf{y}$. If $X^T X$ is not invertible, the solution $\mathbf{w}_{lin} = (X^T X)^{-1} X^T \mathbf{y}$ won't work. In this event, there will be many solutions for $\mathbf{w}$ that minimize E_{in}. Here, you will derive one such solution. Let ρ be the rank of X. Assume that the singular value decomposition (SVD) of X is $X = U\Gamma V^T$, where $U \in \mathbb{R}^{n\times\rho}$ satisfies $U^T U = I_\rho$, $V \in \mathbb{R}^{(d+1)\times\rho}$ satisfies $V^T V = I_\rho$, and $\Gamma \in \mathbb{R}^{\rho\times\rho}$ is a positive diagonal matrix.

(a) Show that $\rho < d + 1$.

(b) Show that $\mathbf{w}_{lin} = V\Gamma^{-1}U^T\mathbf{y}$ satisfies $X^T X\mathbf{w}_{lin} = X^T\mathbf{y}$, hence is a solution.

(c) Show that for any other solution that satisfies $X^T X\mathbf{w} = X^T\mathbf{y}$, $\| \mathbf{w}_{lin} \| < \| \mathbf{w} \|$. That is, the solution we have constructed is the minimum norm set of weights that minimize E_{in}.

Solution

(a) We know that, $\text{RANK}(X) = \rho$. Now by the property of rank we can write, $\text{RANK}(X) = \text{RANK}(X^T X)$. $X^T X$ is a $(d+1)\times(d+1)$ matrix and $X^T X$ is not invertible. Therefore,

$$\begin{aligned}
&\text{RANK}(X^T X) < d + 1 \\
&\text{RANK}(X) < d + 1 \\
&\rho < d + 1
\end{aligned}$$

(b) We have $X = U\Gamma V^T$ and $\mathbf{w}_{lin} = V\Gamma^{-1}U^T\mathbf{y}$, then,

$$\begin{aligned}
X^T X\mathbf{w}_{lin} &= V\Gamma U^T U\Gamma V^T V\Gamma^{-1}U^T\mathbf{y} \\
&= V\Gamma^2\Gamma^{-1}U^T\mathbf{y} \\
&= V\Gamma U^T\mathbf{y} \\
&= (U\Gamma V^T)^T\mathbf{y} \\
&= X^T\mathbf{y}
\end{aligned}$$

Hence, $\mathbf{w}_{lin}$ is a possible solution.

(c) Let, $\mathbf{w}$ be any solution and we can write,

$$\mathbf{w} = \mathbf{w}_{lin} + (\mathbf{w} - \mathbf{w}_{lin}) = \mathbf{w}_{lin} + \delta$$

Now,

$$\begin{aligned}
||\mathbf{w}||^2 &= ||\mathbf{w}_{lin} + \delta||^2 \\
&= (\mathbf{w}_{lin} + \delta)^T(\mathbf{w}_{lin} + \delta) \\
&= (\mathbf{w}_{lin}^T + \delta^T)(\mathbf{w}_{lin} + \delta) \\
&= \mathbf{w}_{lin}^T\mathbf{w}_{lin} + \delta^T\mathbf{w}_{lin} + \mathbf{w}_{lin}^T\delta + \delta^T\delta \\
&= ||\mathbf{w}_{lin}||^2 + ||\delta||^2 + \delta^T\mathbf{w}_{lin} + \mathbf{w}_{lin}^T\delta
\end{aligned}$$

Now, $\mathbf{w}$ and $\mathbf{w}_{lin}$ both are possible solutions. Therefore,

$$X^T X(\mathbf{w} - \mathbf{w}_{lin}) = X^T\mathbf{y} - X^T\mathbf{y} = 0$$

$$
\begin{aligned}
&\Rightarrow V\Gamma U^T U \Gamma V^T (\mathbf{w} - \mathbf{w}_{lin}) = 0 \\
&\Rightarrow V\Gamma^2 V^T (\mathbf{w} - \mathbf{w}_{lin}) = 0 \qquad [\text{As } U^T U = I_\rho] \\
&\Rightarrow \Gamma^{-2} V^T V \Gamma^2 V^T (\mathbf{w} - \mathbf{w}_{lin}) = 0 \\
&\Rightarrow V^T (\mathbf{w} - \mathbf{w}_{lin}) = 0 \qquad [\text{As } V^T V = I_\rho]
\end{aligned}
$$

Again,

$$
\begin{aligned}
\mathbf{w}_{lin}^T \delta &= \mathbf{w}_{lin}^T (\mathbf{w} - \mathbf{w}_{lin}) \\
&= (V\Gamma^{-1} U^T \mathbf{y})^T (\mathbf{w} - \mathbf{w}_{lin}) \\
&= \mathbf{y}^T U \Gamma^{-1} V^T (\mathbf{w} - \mathbf{w}_{lin}) \qquad [\text{As } V^T(\mathbf{w} - \mathbf{w}_{lin}) = 0] \\
&= 0
\end{aligned}
$$

Hence,

$$
\begin{aligned}
||\mathbf{w}||^2 &= ||\mathbf{w}_{lin}||^2 + ||\delta||^2 + 0 + 0 \\
&= ||\mathbf{w}_{lin}||^2 + ||\delta||^2 \\
&> ||\mathbf{w}_{lin}||^2
\end{aligned}
$$

So, $\mathbf{w}_{lin}$ is minimum norm set of weights that minimizes E_{in}

Note: This lab is based on Abu-Mostafa et al., 2012.

References

Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H.-T. (2012). *Learning from data.* AMLBook.