Linear regression

Machine Learning II
2021-2022 - UMONS
Souhaib Ben Taieb

1

Do Exercise 3.3 in LFD.

Exercise 3.3

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

(a) Show that H is symmetric.

(b) Show that $\mathrm{H}^K = \mathrm{H}$ for any positive integer K.

(c) If I is the identity matrix of size N, show that $(\mathrm{I}-\mathrm{H})^K = \mathrm{I}-\mathrm{H}$ for any positive integer K.

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

Figure 1: Source: Abu-Mostafa et al. Learning from data. AMLbook.

Solution

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

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

(b) We have to show that $H^K = H$ for $K = 1, 2, 3, \dots$. We will prove that by using induction.

- For $K = 1$, $H^1 = H$.
- For $K = 2$,

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

- Consider, it is true for K, $H^K = H$.

- For $K = K + 1$,

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

(c) 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 $K + 1$,

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

(d) We have to prove $trace(H) = d + 1$,

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

2

Do Exercise 3.4 in LFD.

Exercise 3.4

Consider a noisy target $y = \mathbf{w}^{*\mathrm{T}}\mathbf{x} + \epsilon$ 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_n as ϵ_n and let $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^{\mathrm{T}}$; assume that $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is invertible. By following

(continued on next page)

the steps below, show that the expected in-sample error of linear regression with respect to $\mathcal{D}$ is given by

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{lin}})] = \sigma^2 \left(1 - \frac{d+1}{N}\right).$$

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

(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 $E_{\text{in}}(\mathbf{w}_{\text{lin}})$ in terms of $\boldsymbol{\epsilon}$ using (b), and simplify the expression using Exercise 3.3(c).

(d) Prove that $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{lin}})] = \sigma^2\left(1 - \frac{d+1}{N}\right)$ using (c) and the 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 3.3(d).]*

For the expected out-of-sample error, we take a special case which is easy to analyze. Consider a test data set $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_1, y'_1), \ldots, (\mathbf{x}_N, y'_N)\}$, which shares the same input vectors $\mathbf{x}_n$ with $\mathcal{D}$ but with a different realization of the noise terms. Denote the noise in y'_n as ϵ'_n and let $\boldsymbol{\epsilon}' = [\epsilon'_1, \epsilon'_2, \ldots, \epsilon'_N]^{\mathrm{T}}$. Define $E_{\text{test}}(\mathbf{w}_{\text{lin}})$ to be the average squared error on $\mathcal{D}_{\text{test}}$.

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

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 Problem 3.11.

Figure 2: Source: Abu-Mostafa et al. Learning from data. AMLbook.

Solution

We have,

$$\begin{aligned}\mathcal{D} &= \{(x_n, y_n\}_{n=1}^N \qquad [\text{where } x_n \in \mathbb{R}^{d+1} \text{ and } y_n \in \mathbb{R}] \\ &= \{X, y\} \qquad [\text{where } N \in \mathbb{R}^{N\times d+1} \text{ and } y \in \mathbb{R}^{N\times 1}]\end{aligned}$$

Then the in-sample error can be written as,

$$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}(y_n - h(x_n))^2$$

$$= ||y - Xw||^2$$

Now, for linear regression,

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

Therefore,

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

(a) The in-sample error estimate is

$$\begin{aligned} \hat{y} &= Hy \\ &= H(Xw^* + \epsilon) \\ &= HXw^* + H\epsilon \\ &= (X(X^T X)^- {-1} X^T)Xw^* + H\epsilon \\ &= Xw^* + H\epsilon \end{aligned}$$

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

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

(c)

$$\begin{aligned} E_{in}(w_{lin}) &= \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 \\ &= \frac{1}{N} \sum_{n=1}^{N} (y_n - w_{lin}^T x_n)^2 \\ &= \frac{1}{N} ||y - \hat{y}||^2 \\ &= \frac{1}{N} ||(H - I)\epsilon||^2 \\ &= \frac{1}{N} \epsilon^T (H - I)^T (H - I)\epsilon \\ &= \frac{1}{N} \epsilon^T (H^T - I)(H - I)\epsilon \\ &= \frac{1}{N} \epsilon^T (H - I)(H - I)\epsilon \\ &= \frac{1}{N} \epsilon^T (H - I)^2 \epsilon \\ &= \frac{1}{N} \epsilon^T (I - H)^2 \epsilon \\ &= \frac{1}{N} \epsilon^T (I - H)\epsilon \qquad \text{[Using } \mathbf{3.3\ (c)}\text{]} \end{aligned}$$

(d)

$$\mathbb{E}_{\mathcal{D}}[E_{in}(w_{lin})] = \mathbb{E}_{\mathcal{D}}[\frac{1}{N} \epsilon^T (I - H)\epsilon]$$

$$
\begin{aligned}
&= \mathbb{E}_\epsilon[\frac{1}{N}\epsilon^T(I-H)\epsilon] \\
&= \frac{1}{N}\mathbb{E}_\epsilon[\epsilon^T\epsilon - \epsilon^T H\epsilon] \\
&= \frac{1}{N}(\mathbb{E}_\epsilon[\epsilon^T\epsilon] - \mathbb{E}_\epsilon[\epsilon^T H\epsilon]) \\
&= \frac{1}{N}\mathbb{E}_\epsilon[\epsilon^T\epsilon] - \frac{1}{N}\mathbb{E}_\epsilon[\epsilon^T H\epsilon]) \\
&= \frac{1}{N}\mathbb{E}_\epsilon[\epsilon^T\epsilon] - \frac{1}{N}\mathbb{E}_\epsilon[trace(\epsilon^T H\epsilon)] \quad \text{[As ϵ is $N\times 1$ matrix and H is $N\times N$ matrix]} \\
&= \frac{1}{N}\mathbb{E}_\epsilon[\epsilon^T\epsilon] - \frac{1}{N}\mathbb{E}_\epsilon[trace(\epsilon\epsilon^T H)] \\
&= \frac{1}{N}\mathbb{E}_\epsilon[\epsilon^T\epsilon] - \frac{1}{N}trace(\mathbb{E}_\epsilon[\epsilon\epsilon^T H]) \\
&= \frac{1}{N}\mathbb{E}_\epsilon[\epsilon^T\epsilon] - \frac{1}{N}trace(\mathbb{E}_\epsilon[\epsilon\epsilon^T]H)
\end{aligned}
$$

Now,

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

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

Hence,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{in}(w_{lin})] &= \frac{1}{N}N\sigma^2 - \frac{1}{N}trace(\sigma^2 I_n H) \\
&= \sigma^2 - \frac{1}{N}trace(\sigma^2 H) \\
&= \sigma^2 - \frac{\sigma^2}{N}trace(H) \\
&= \sigma^2 - \frac{\sigma^2}{N}(d+1) \qquad \text{[Using }\mathbf{3.3\ (d)}\text{]} \\
&= \sigma^2(1 - -\frac{(d+1)}{N})
\end{aligned}
$$

(e)

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

So, we have

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

Now,

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

3

Solve Problem 3.11 in LFD.

Problem 3.11 Consider the linear regression problem setup in Exercise 3.4, 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 2nd moment matrix $\Sigma = \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^{\mathsf{T}}]$ is non-singular. Follow the steps below to show that, with high probability, the out-of-sample error on average is

$$E_{\text{out}}(\mathbf{w}_{\text{lin}}) = \sigma^2 \left(1 + \frac{d+1}{N} + o(\tfrac{1}{N})\right).$$

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

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

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

$$E_{\text{out}} = \sigma^2 + \operatorname{trace}\left(\Sigma(\mathrm{X}^{\mathsf{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathsf{T}}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\mathrm{X}^{\mathsf{T}}(\mathrm{X}^{\mathsf{T}}\mathrm{X})^{-1}\right).$$

[Hints: $a = \operatorname{trace}(a)$ for any scalar a; $\operatorname{trace}(\mathrm{AB}) = \operatorname{trace}(\mathrm{BA})$; expecta tion and trace commute.]

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

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

$$E_{\text{out}} = \sigma^2 + \frac{\sigma^2}{N}\operatorname{trace}\left(\Sigma(\tfrac{1}{N}\mathrm{X}^{\mathsf{T}}\mathrm{X})^{-1}\right).$$

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

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

$$E_{\text{out}} = \sigma^2 \left(1 + \frac{d+1}{N} + o(\tfrac{1}{N})\right).$$

[Hint: By the law of large numbers $\frac{1}{N}\mathrm{X}^{\mathsf{T}}\mathrm{X}$ converges in probability to Σ, and so by continuity of the inverse at Σ, $\left(\frac{1}{N}\mathrm{X}^{\mathsf{T}}\mathrm{X}\right)^{-1}$ converges in probability to Σ^{-1}.]

Figure 3: Source: Abu-Mostafa et al. Learning from data. AMLbook.

Solution

(a) For a test point x_i,

$$\begin{aligned} y_i - g(x_i) &= x_i^T w^* + \epsilon_i - x_i^T \hat{w} \\ &= x_i^T w^* + \epsilon_i - x_i^T (X^T X)^{-1} X^T y \\ &= x_i^T w^* + \epsilon_i - x_i^T (X^T X)^{-1} X^T (X w^* + \epsilon) \end{aligned}$$

$$
\begin{aligned}
&= x_i^T w^* + \epsilon_i - x_i^T (X^T X)^{-1} X^T X w^* - x_i^T (X^T X)^{-1} X^T \epsilon \\
&= x_i^T w^* + \epsilon_i - x_i^T w^* - x_i^T (X^T X)^{-1} X^T \epsilon \\
&= \epsilon_i - x_i^T (X^T X)^{-1} X^T \epsilon
\end{aligned}
$$

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

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

(c)

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

(d) By taking expectation w.r.t. ϵ, we obtain,

$$
\begin{aligned}
\mathbb{E}_\epsilon[E_{out}] &= \mathbb{E}_\epsilon[\sigma^2 + trace(\Sigma (X^T X)^{-1} X^T \epsilon\epsilon^T X (X^T X)^{-1})] \\
&= \sigma^2 + trace(\Sigma (X^T X)^{-1} X^T \mathbb{E}_\epsilon[\epsilon\epsilon^T] X (X^T X)^{-1}) \\
&= \sigma^2 + trace(\Sigma (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1}) \\
&= \sigma^2 + \sigma^2 trace(\Sigma (X^T X)^{-1} X^T X (X^T X)^{-1}) \\
&= \sigma^2 + \sigma^2 trace\Big(\Sigma (X^T X)^{-1}\Big) \\
&= \sigma^2 + \sigma^2 \frac{N}{N} trace\Big(\Sigma (X^T X)^{-1}\Big) \\
&= \sigma^2 + \frac{\sigma^2}{N} trace\Big(\Sigma \Big(\frac{X^T X}{N}\Big)^{-1}\Big) \\
&= \sigma^2 + \frac{\sigma^2}{N} trace(I) \qquad [\Big(\frac{X^T X}{N}\Big) \approx \Sigma] \\
&= \sigma^2 + \frac{\sigma^2 (d+1)}{N} \\
&= \sigma^2 \Big(1 + \frac{(d+1)}{N}\Big)
\end{aligned}
$$

(e)

$$
\frac{X^T X}{N} \xrightarrow{P} \Sigma
$$

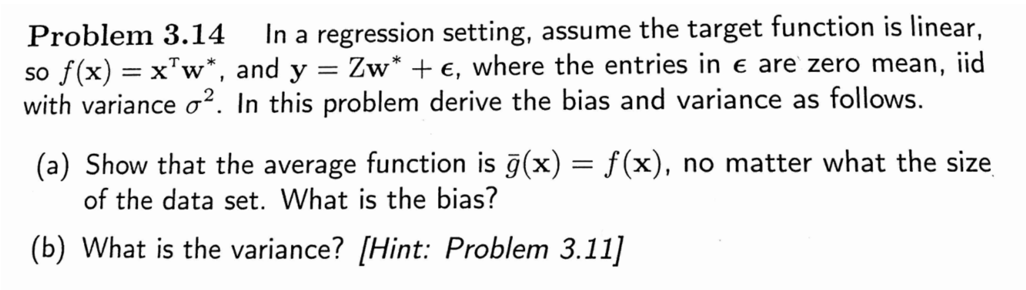
$$
\begin{aligned}
(\frac{X^T X}{N})^{-1} &\xrightarrow{P} \Sigma^{-1} \\
(\frac{X^T X}{N})^{-1} &= \Sigma^{-1} + o(1)
\end{aligned}
$$

Now,

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

4

Solve Problem 3.14 in LFD.



Problem 3.14 In a regression setting, assume the target function is linear, so $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{w}^*$, and $\mathbf{y} = \mathrm{Z}\mathbf{w}^* + \boldsymbol{\epsilon}$, 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: Problem 3.11]*

Figure 4: Source: Abu-Mostafa et al. Learning from data. AMLbook.

Solution

(a)

$$\begin{aligned}
y_n &= f(x) + \epsilon_n = x^T w^* + \epsilon \\
y &= Xw^* + \epsilon \\
g^{\mathcal{D}}(x) &= x^T \hat{w} \\
\hat{w} &= (X^T X)^{-1} X^T y
\end{aligned}$$

$$\begin{aligned}
\bar{g}(x) &= \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)] \\
&= \mathbb{E}_{\mathcal{D}}[x^T \hat{w}] \\
&= \mathbb{E}_{\mathcal{D}}[x^T (X^T X)^{-1} X^T y] \\
&= \mathbb{E}_{\mathcal{D}}[x^T (X^T X)^{-1} X^T (Xw^* + \epsilon)] \qquad [\text{where } y = Xw^* + \epsilon] \\
&= \mathbb{E}_{\mathcal{D}}[x^T w^* + x^T (X^T X)^{-1} X^T \epsilon)] \\
&= \mathbb{E}_{\epsilon}[x^T w^* + x^T (X^T X)^{-1} X^T \epsilon)] \\
&= x^T w^* \\
&= f(x)
\end{aligned}$$

$$\begin{aligned}
\text{Bias} &= \mathbb{E}_x[(\mathbb{E}_{\epsilon_n}[y_n] - \bar{g}(x))^2] \\
&= \mathbb{E}_x[(f(x) - f(x))^2] \\
&= 0
\end{aligned}$$

(b)

$$\begin{aligned}
\text{Variance} &= \mathbb{E}_{x,\mathcal{D}}[(g^{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)])^2] \\
&= \mathbb{E}_{x,\mathcal{D}}[(g^{\mathcal{D}}(x) - \bar{g}(x))^2] \\
&= \mathbb{E}_{x,\mathcal{D}}[(x^T \hat{w} - x^T w^*)^2] \\
&= \mathbb{E}_{x,y}[(x^T (X^T X)^{-1} X^T y - x^T w^*)^2] \\
&= \mathbb{E}_{x,\epsilon}[(x^T (X^T X)^{-1} X^T (Xw^* + \epsilon) - x^T w^*)^2] \\
&= \mathbb{E}_{x,\epsilon}[(x^T (X^T X)^{-1} X^T X w^* + x^T (X^T X)^{-1} X^T \epsilon - x^T w^*)^2] \\
&= \mathbb{E}_{x,\epsilon}[(x^T w^* + x^T (X^T X)^{-1} X^T \epsilon - x^T w^*)^2]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{x,\epsilon}[(x^T(X^TX)^{-1}X^T\epsilon)^2] \\
&= \mathbb{E}_{x,\epsilon}[trace(x^T(X^TX)^{-1}X^T\epsilon)^2)] \qquad \text{[As } (x^T(X^TX)^{-1}X^T\epsilon)^2 \text{ is a scalar]} \\
&= \mathbb{E}_{x,\epsilon}[trace((x^T(X^TX)^{-1}X^T\epsilon)(x^T(X^TX)^{-1}X^T\epsilon)^T)] \\
&= \mathbb{E}_{x,\epsilon}[(trace(x^T(X^TX)^{-1}X^T\epsilon)(\epsilon^TX(X^TX)^{-1}x))] \\
&= \mathbb{E}_{x,\epsilon}[trace((x^T(X^TX)^{-1}X^T\epsilon)(\epsilon^TX(X^TX)^{-1}x))] \\
&= \mathbb{E}_{x,\epsilon}[trace(xx^T(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1})] \\
&= trace(\mathbb{E}_{x,\epsilon}[xx^T(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}]) \\
&= trace(\mathbb{E}_{x}[xx^T\mathbb{E}_{\epsilon}[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}]]) \\
&= trace(\mathbb{E}_{x}[xx^T\sigma^2(X^TX)^{-1}]) \qquad \text{[where } \mathbb{E}_{\epsilon}[\epsilon\epsilon] = \sigma^2 I\text{]} \\
&= trace(\mathbb{E}_{x}[xx^T]\sigma^2(X^TX)^{-1}) \\
&= \sigma^2 trace(\Sigma(X^TX)^{-1}) \\
&= \sigma^2\frac{N}{N}trace(\Sigma(X^TX)^{-1}) \\
&= \frac{\sigma^2}{N}trace(\Sigma(\frac{X^TX}{N})^{-1}) \\
&= \sigma^2\Big(\frac{d+1}{N} + o(\frac{1}{N})\Big) \qquad \text{[from } \textbf{ex 3 (e)}\text{]}
\end{aligned}
$$

5

Solve Problem 3.15 in LFD.

Problem 3.15 In the text we derived that the linear regression solution weights must satisfy $X^T X\mathbf{w} = X^T\mathbf{y}$. If $X^T X$ is not invertible, the solution $\mathbf{w}_{\text{lin}} = (X^T X)^{-1} X^T \mathbf{y}$ 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 $\mathbf{w}_{\text{lin}} = V\Gamma^{-1}U^T\mathbf{y}$ satisfies $X^T X\mathbf{w}_{\text{lin}} = X^T\mathbf{y}$, and hence is a solution.

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

Figure 5: Source: Abu-Mostafa et al. Learning from data. AMLbook.

Solution

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

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

(b) We have $X = U\Gamma V^T$ and $w_{lin} = V\Gamma^{-1}U^Ty$, then,

$$\begin{aligned}
X^TXw_{lin} &= V\Gamma U^TU\Gamma V^TV\Gamma^{-1}U^Ty\\
&= V\Gamma^2\Gamma^{-1}U^Ty\\
&= V\Gamma U^Ty\\
&= (U\Gamma V^T)^Ty\\
&= X^Ty
\end{aligned}$$

Hence, w_{lin} is a possible solution.

(c) Let, w be any solution and we can write,

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

Now,

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

$$
\begin{aligned}
&= w_{lin}^T w_{lin} + \delta^T w_{lin} + w_{lin}^T \delta + \delta^T \delta \\
&= ||w_{lin}||^2 + ||\delta||^2 + \delta^T w_{lin} + w_{lin}^T \delta
\end{aligned}
$$

Now, w and w_{lin} both are possible solutions. Therefore,

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

Again,

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

Hence,

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

So, w_{lin} is minimum norm set of weights that minimizes E_{in}