Linear regression

Big Data Analytics 2020-2021 - UMONS
Souhaib Ben Taieb

1

Do Exercise 3.3 in LFD.

Exercise 3.3

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

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.

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 $\Sigma = \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ 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

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

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

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

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

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

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

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

Note that $\frac{1}{N}\mathrm{X}^{\mathrm{T}}\mathrm{X} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}$ is an N sample estimate of Σ. So $\frac{1}{N}\mathrm{X}^{\mathrm{T}}\mathrm{X} \approx \Sigma$. If $\frac{1}{N}\mathrm{X}^{\mathrm{T}}\mathrm{X} = \Sigma$, 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 $\frac{1}{N}\mathrm{X}^{\mathrm{T}}\mathrm{X}$ converges in probability to Σ, and so by continuity of the inverse at Σ, $\left(\frac{1}{N}\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}$ converges in probability to Σ^{-1}.]

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

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.

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.