Linear regression

Machine Learning II (2022-2023)
UMONS

1 Exercise 1

Consider the hat matrix $H = X(X^TX)^{-1}X^T$, where X is an n by $d+1$ matrix, and X^TX 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:** $\text{trace}(AB) = \text{trace}(BA)$]

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 $\epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]^T$; assume that X^TX 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

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

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

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

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

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

For the expected out-of-sample error, we take a special case which is easy to analyze. Consider a test data set $\mathcal{D}_{test} = \{(\mathbf{x}_1, y'_1), \ldots, (\mathbf{x}_n, y'_n)\}$, 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 $\epsilon' = [\epsilon'_1, \epsilon'_2, \ldots, \epsilon'_n]^T$. Define $E_{test}(\mathbf{w}_{lin}$ to be the average squared error on $\mathcal{D}_{test}$.

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

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 Exercise 3.

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

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

where ϵ is the noise realization for the test point and ϵ 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

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

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

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

(d) Take the expectation with respect to ϵ 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}.]

4 Exercise 4

In a regression setting, assume the target function is linear, so $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}^*$, and $\mathbf{y} = Z\mathbf{w}^* + \epsilon$, where the entries in ϵ 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]

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}.

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.