COL341 Spring 2023
Homework 1
(To be done Individually)

Due Date: 3rd February 2023, Friday, 11:55 PM (No extensions)

Instructions

Type the solutions in LaTeX(you may use Overleaf for ease of use). Submit the `.tex` source and the compiled `pdf` in a single `.zip` file in Moodle. Name the file as `<your-entry-number>.zip`, e.g. `2019CSZ8406.zip`. The homework is to be done individually. Plagiarism and academic dishonesty will be penalized as per the course policy. No deadline extension will be provided.

Question 1 $[1 \times 4 = 4$ marks$]$

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 $H^K = H$ for any positive integer K.

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

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

Question 2 $[1 + 1 + 1 + 2 + 2 = 7$ marks$]$

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 $\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, denote the noise in y_n as ϵ_n and let $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^T$; 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

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

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

(d) Prove Eq. (1) using (c) and the independence of $\epsilon_1, \ldots, \epsilon_N$. [Hint: the sum of the diagonal elements of a matrix (the trace) will play a role.]

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

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

The special test error E_{test} is a very restricted case of the general out of sample error.

Question 3 $[1 + 2 + 2 + 2 + 2 = 9$ marks$]$

Consider the linear regression problem setup in Question2, 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 the second 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

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

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

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

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

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

[Hints: $a=$ trace(a) for any scalar a; trace(AB)= 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,

$$E_{\text{out}} = \sigma^2 + \frac{\sigma^2}{N}\text{trace}\Big(\Sigma(\frac{1}{N}X^TX)^{-1}\Big). \tag{3}$$

Note that $\frac{1}{N}X^TX = \frac{1}{N}\sum_{n=1}^N \mathbf{x}_n\mathbf{x}_n^T$ 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,

$$E_{\text{out}} = \sigma^2\Big(1 + \frac{d+1}{N} + o(\frac{1}{N})\Big)$$

[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}.]