COL341: Homework -1

Simran Mahawar
2020CS10387

February 3, 2023

Question 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.
A matrix is symmetric if it is equal to its transpose. $H^T = H$
T represents the transpose of a Matrix. We will compute the value of matrix H^T.

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

Hence proved

(b) Show that $H^K = H$for any positive integer K.

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

$H^2 = H$ so H is idempotent and $H^K = H$.

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

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

$(I - H)^2 = I - H$. It is also idempotent.
$(I - H)^K = I - H$.

(d) Show that $trace(H) = d + 1$, where the trace is the sum of diagonal elements.

$$\begin{aligned}\text{trace}(H) &= \text{trace}\left(X(X^TX)^{-1}X^T\right) \\ &= \text{trace}\,(AB) \\ &= \text{trace}\,(BA) \\ &= \text{trace}\left(X^TX(X^TX)^{-1}\right) \\ &= \text{trace}\,(I) \\ &= d+1\end{aligned}$$

where the identity matrix I is of size $d + 1$.

Question 2

Consider a noisy target $y = Xw^* + \epsilon$ for generating the data, where ϵ is a noise term with zero mean and σ^2 variance, independently generated for every example (x, y).

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

We have $y = Xw^* + \epsilon$ and $H = X(X^TX)^{-1}X^T$
Using this into the expression for in-sample estimate of y is $\hat{y}$

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

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

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

(c) Express E_in $(w_lin$ in terms of ϵ using (b), and simplify the expression using Question 1(c).

$$
\begin{aligned}
E_{in}(w_{lin}) &= \frac{1}{N}\|Xw_{lin} - y\|^2 \\
&= \frac{1}{N}\|y - \hat{y}\|^2 \\
&= \frac{1}{N}\|(I - H)\epsilon\|^2 \\
&= \frac{1}{N}\epsilon^T(I - H)^T(I - H)\epsilon \\
&= \frac{1}{N}\epsilon^T(I - H)(I - H)\epsilon \\
&= \frac{1}{N}\epsilon^T(I - H)\epsilon
\end{aligned}
$$

(d) Prove Eq. (1) using (c) and the independence of $\epsilon_1, ..., \epsilon_N$.
Using result from part (c) We have to find $E_{\mathcal{D}}\left[E_{in}(w_{lin})\right]$.

$$
\begin{aligned}
E_{\mathcal{D}}\left[E_{in}(w_{lin})\right] &= E_{\mathcal{D}}\left[\frac{1}{N}\epsilon^T(I - H)\epsilon\right] \\
&= \frac{1}{N}\left(E_{\mathcal{D}}[\epsilon^T\epsilon] - E_{\mathcal{D}}[\epsilon^T H\epsilon]\right) \\
&= \frac{1}{N}\left(E_{\mathcal{D}}[\sum_{k=1}^{N}\epsilon_k^2] - E_{\mathcal{D}}[\sum_{i=1}^{N}\sum_{j=1}^{N}\epsilon_i h_{ij}\epsilon_j]\right) \\
&= \frac{1}{N}\left(\sum_{k=1}^{N}E_{\mathcal{D}}\epsilon_k^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}E_{\mathcal{D}}[\epsilon_i h_{ij}\epsilon_j]\right) \\
&= \frac{1}{N}\left(N\sigma^2 - \sum_{i=1}^{N}E_{\mathcal{D}}[\epsilon_i^2 h_{ii}]\right) \\
&= \frac{1}{N}\left(N\sigma^2 - \sum_{i=1}^{N}h_{ii}E_{\mathcal{D}}[\epsilon_i^2]\right) \\
&= \frac{1}{N}\left(N\sigma^2 - \sigma^2\text{trace}(H)\right) \\
&= \sigma^2\left(1 - \frac{\text{trace}(H)}{N}\right) \\
&= \sigma^2\left(1 - \frac{d+1}{N}\right)
\end{aligned}
$$

Here we assumed ϵ_i is independent and H is not random variable. We can pull h_{ii} out of $E_{\mathcal{D}}[\epsilon_i^2 h_{ii}]$.

(e) Prove that $E_{\mathcal{D},\epsilon'}\left[E_{test}(w_{lin})\right] = \sigma^2\left(1 + \frac{d+1}{N}\right)$.

The special test error E_test is a very restricted case of the general out of sample error.
Since X doesn't change, only ϵ changes, we have Here only ϵ is changing X doesn't change. So,

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

ϵ and ϵ' are independent of each other

$$\begin{aligned}
E_{\mathcal{D},\epsilon'}\left[E_{test}(w_{lin})\right] &= E_{\mathcal{D},\epsilon'}\left[\frac{1}{N}\|y' - \hat{y}\|^2\right] \\
&= E_{\mathcal{D},\epsilon'}\left[\frac{1}{N}\|\epsilon' - H\epsilon\|^2\right] \\
&= \frac{1}{N}E_{\mathcal{D},\epsilon'}\left[(\epsilon' - H\epsilon)^T(\epsilon' - H\epsilon)\right] \\
&= \frac{1}{N}E_{\mathcal{D},\epsilon'}\left[(\epsilon'^T - \epsilon^T H^T)(\epsilon' - H\epsilon)\right] \\
&= \frac{1}{N}E_{\mathcal{D},\epsilon'}\left[(\epsilon'^T - \epsilon^T H)(\epsilon' - H\epsilon)\right] \\
&= \frac{1}{N}E_{\mathcal{D},\epsilon'}\left[\epsilon'^T\epsilon' - \epsilon'^T H\epsilon - \epsilon^T H\epsilon' + \epsilon^T H\epsilon\right] \\
&= \frac{1}{N}E_{\mathcal{D},\epsilon'}\left[\epsilon'^T\epsilon' + \epsilon^T H\epsilon\right] \\
&= \frac{1}{N}\left(\sum_{k=1}^{N} E_{\mathcal{D}}\epsilon_k'^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} E_{\mathcal{D}}[\epsilon_i h_{ij}\epsilon_j]\right) \\
&= \sigma^2\left(1 + \frac{d+1}{N}\right)
\end{aligned}$$

Question 3

(a) For a test point x, show that the error $y - g(x)$ is $\epsilon_t - x_t^T(X^TX)^{-1}X^T\epsilon$. Where ϵ is the noise realization for the test point and ϵ is the vector of noise realizations on the data.
Following question 2 and use the fact that $w_{lin} = (X^TX)^{-1}X^Ty$, for a given test point x_t, we have

$$\begin{aligned}
g(x_t) &= x_t^T w_{lin} \\
&= x_t^T(X^TX)^{-1}X^Ty \\
&= x_t^T(X^TX)^{-1}X^T(Xw^* + \epsilon) \\
&= x_t^T(X^TX)^{-1}X^TXw^* + x_t^T(X^TX)^{-1}X^T\epsilon \\
&= x_t^T w^* + x_t^T(X^TX)^{-1}X^T\epsilon
\end{aligned}$$

On the other side, the y at test point x_t is: $y = x_t^T w^* + \epsilon_t$, so we have

$$y - g(x_t) = \epsilon_t - x_t^T(X^TX)^{-1}X^T\epsilon$$

Where ϵ_t 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., x and ϵ, to obtain an expression for Eout. Show that $E_{out} = \sigma^2 + trace(\Sigma(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1})$
Take the expectation w.r.t. to the test point, i.e. x_t and ϵ_t, we have

$$\begin{aligned}
E_{out} &= E[(y - g(x_t))^2] \\
&= E[(\epsilon_t - x_t^T(X^TX)^{-1}X^T\epsilon)^2t] \\
&= E[\epsilon_t^2 - 2\epsilon_t x_t^T(X^TX)^{-1}X^T\epsilon + (x_t^T(X^TX)^{-1}X^T\epsilon)(x_t^T(X^TX)^{-1}X^T\epsilon)^T] \\
&\text{Note the last term is a scalar} \\
&= E[\epsilon_t^2] - 2E[\epsilon_t x_t^T(X^TX)^{-1}X^T\epsilon] + E[x_t^T(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-T}x_t] \\
&= \sigma^2 - 2E[\epsilon_t]E[x_t^T(X^TX)^{-1}X^T\epsilon] + E[trace(x_t^T(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-T}x_t)]
\end{aligned}$$

In the last term we use the fact that trace on a scalar equals to the scalar
We also apply the independence between ϵ_t and x_t.
Also note that X and ϵ are non-random in this expectation

$$\begin{aligned}
&= \sigma^2 + E[trace(x_tx_t^T(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-T})] \\
&= \sigma^2 + trace(E[x_tx_t^T(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-T}]) \\
&= \sigma^2 + trace(E[x_tx_t^T]E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}]) \\
&= \sigma^2 + trace(\Sigma(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1})
\end{aligned}$$

(c) What is $E_\epsilon[\epsilon\epsilon^T]$
$\epsilon\epsilon^T$ is a $N \times N$ matrix, with entries $\epsilon_i\epsilon_j$. So $E_\epsilon[\epsilon\epsilon^T] = \sigma^2 I$ where the expectation of $E[\epsilon_i\epsilon_j] = 0$ when $i \neq j$, otherwise σ^2.

(d) Take the expectation with respect to ϵ to show that, on average,
$E_{out} = \sigma^2 + \frac{\sigma^2}{N} trace(\Sigma(\frac{1}{N}X^TX)^{-1})$

If $\frac{1}{N}X^TX = \Sigma$, then what is E_{out} on average?
Take the expectation w.r.t. ϵ, which is a $N \times 1$ vector. We have

$$\begin{aligned}
E_{out} &= \sigma^2 + E_\epsilon[trace(\Sigma(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1})] \\
&= \sigma^2 + trace(E_\epsilon[\Sigma(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}]) \\
&= \sigma^2 + trace(E_\epsilon[\Sigma(X^TX)^{-1}X^T]E_\epsilon[\epsilon\epsilon^T]E_\epsilon[X(X^TX)^{-1}]) \\
&= \sigma^2 + trace(\Sigma(X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1}) \\
&= \sigma^2 + \sigma^2 trace(\Sigma(X^TX)^{-1}X^TX(X^TX)^{-1}) \\
&= \sigma^2 + \sigma^2 trace(\Sigma(X^TX)^{-1}) \\
&= \sigma^2 + \frac{\sigma^2}{N} trace(\Sigma(\frac{1}{N}X^TX)^{-1})
\end{aligned}$$

Note that $\frac{1}{N}X^TX = \frac{1}{N}\sum_{n=1}^{N} x_n x_n^T$ is an N-sample estimate of Σ. So $\frac{1}{N}X^TX \approx \Sigma$, in such case, we have

$E_{out} = \sigma^2 + \frac{\sigma^2}{N}trace(I) = \sigma^2 + \frac{\sigma^2(d+1)}{N} = \sigma^2(1 + \frac{d+1}{N})$

(e) Show that (after taking the expectation over the data noise) with high probability, $E_{out} = \sigma^2(1 + \frac{d+1}{N} + o(\frac{1}{N}))$

By law of large numbers $\frac{1}{N}X^TX$ converges in probability to Σ, so by continuity of the inverse at Σ, $(\frac{1}{N}X^TX)^{-1}$ converges in probability to Σ^{-1}. $trace(\Sigma(\frac{1}{N}X^TX)^{-1}) = trace(I) + o(1)$, so we have $E_{out} = \sigma^2 + \frac{\sigma^2}{N}(d + 1 + o(1)) = \sigma^2(1 + \frac{d+1}{N} + o(\frac{1}{N}))$