Machine Learning and Statistical Learning Theory
Hw 1 (With Corrections)

Daniel Yu

September 11, 2024

Note. Problem 1

1. Question 1: Find the Expected Value $E_D(\hat{\vec{\theta}})$ of Ridge Regression over $D = (X, \vec{y})$.

   Note. Note we are given that $y = \vec{\theta}^T\vec{x} + \varepsilon$ and $(\hat{\vec{\theta}} = X^TX + \lambda I)^{-1}X^T\vec{y}$. We know $\vec{y} = X\vec{\theta} + \vec{\varepsilon}$ (Recall that in standard notation vectors are column vectors but matrices are formed by row vectors, so we construct $\vec{\theta}^T\vec{x}$ into matrix form using a transpose $X\vec{\theta}$).

   Proof.

   $$\begin{aligned}E_D(\hat{\vec{\theta}}) &= E_D[(X^TX + \lambda I)^{-1}X^T\vec{y}]\\ &= E_D[(X^TX + \lambda I)^{-1}X^T(X\vec{\theta} + \vec{\varepsilon})]\\ &= E_D[(X^TX + \lambda I)^{-1}X^TX\vec{\theta} + (X^TX + \lambda I)^{-1}X^T\vec{\varepsilon}]\\ &= E_D[(X^TX + \lambda I)^{-1}X^TX\vec{\theta}] + E_D[(X^TX + \lambda I)^{-1}X^T\vec{\varepsilon}]\\ &= E_D[(X^TX + \lambda I)^{-1}X^TX\vec{\theta}] + 0\\ &\text{since } (X^TX + \lambda I)^{-1}X^TX\vec{\theta} \text{ is constant w.r.t D}\\ &= (X^TX + \lambda I)^{-1}X^TX\vec{\theta}.\end{aligned}$$

   □

2. Question 2: Is $\hat{\vec{\theta}}$ an unbiased estimator for $\vec{\theta}$?

   Note. No, because the estimator has a bias of $(X^TX+\lambda I)^{-1}X^T\vec{X}$ coming from the λI term which is to be expected as this is the ridge regression hyperparameter to control excessively large coefficients. A true unbiased estimator would have $E_D[\hat{\vec{\theta}}] = \vec{\theta}$

3. Question 3: Find Covariance Matrix $Cov(\vec{\theta})$ of Ridge Regression $\vec{\theta}$ over data $D = (X, \vec{y})$.

   Proof.

   $$\begin{aligned}Cov(\hat{\vec{\theta}}) &= E[(\hat{\vec{\theta}} - E[\vec{\theta}])(\hat{\vec{\theta}} - E[\vec{\theta}])^T]\\ &= E[(\hat{\vec{\theta}} - (X^TX + \lambda I)^{-1}X^TX\vec{\theta})(\hat{\vec{\theta}} - (X^TX + \lambda I)^{-1}X^TX\vec{\theta})^T]\\ &= E[((X^TX + \lambda I)^{-1}X^T\vec{y} - (X^TX + \lambda I)^{-1}X^TX\vec{\theta})(\hat{\vec{\theta}} - (X^TX + \lambda I)^{-1}X^TX\vec{\theta})^T]\\ &= E[((X^TX + \lambda I)^{-1}X^T(\vec{y} - X\vec{\theta}))((X^TX + \lambda I)^{-1}X^T\left(\vec{y} - X\vec{\theta}\right))^T]\\ &= E[((X^TX + \lambda I)^{-1}X^T\vec{\varepsilon})((X^TX + \lambda I)^{-1}X^T\vec{\varepsilon})^T]\\ &= E[\vec{\varepsilon}^2]((X^TX + \lambda I)^{-1}X^T)((X^TX + \lambda I)^{-1}X^T)^T\\ &= \sigma^2((X^TX + \lambda I)^{-1}X^T)((X^TX + \lambda I)^{-1}X^T)^T\\ &\text{since } ((X^TX + \lambda I)^{-1}X^T)^T = X((X^TX + \lambda I)^T)^{-1} = X((X^TX)^T + \lambda I^T)^{-1}\\ &= \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}.\end{aligned}$$

   □

4. Question 4: Decompose the Expected Prediction Error at $\vec{x}^0$:

$$EPE(\vec{x}^0) = E_{D,y^0}[(y^0 - \hat{y^0})^2].$$

as irreducible error, bias, and variance for ridge regression.

Proof.

$$\begin{aligned}
E_{D,y^0}[(y^0 - \hat{y^0})^2] &= E_{y^0} E_D[(y^0 - \hat{y}^0)^2] \\
&= E_{y^0} E_D\{[(y^0 - E_{y^0}[y^0]) + (E_{y^0}[y^0] - E_D[\hat{y}^0]) + (E_D[\hat{y}^0] - \hat{y}^0)]^2\} \\
&= E_{y^0}[y^0 - E_{y^0}[y^0]]^2 + [E_{y^0}[y^0] - E_D[\hat{y}^0]]^2 + E_D[E_D[\hat{y}^0] - \hat{y}^0]^2 \\
&\text{where } y^0 = \vec{\theta}\vec{x^0} + \varepsilon \text{ and } \varepsilon \sim N(\mu, \sigma) \\
&= \sigma^2 + [E_{y^0}[y^0] - E_D[\hat{y}^0]]^2 + E_D[E_D[\hat{y}^0] - \hat{y}^0]^2
\end{aligned}$$

We know the two expressions decompose as follows for ridge regression:

$$\begin{aligned}
[E_{y^0}[y^0] - E_D[\hat{y}^0]]^2 &= [E_{y^0}[\vec{\theta}^T\vec{x^0} + \varepsilon] - E_D[\hat{y}^0]]^2 \\
&\text{since } \hat{y^0} = \vec{x^0}^T\vec{\theta} = \vec{x^0}^T(X^TX - \lambda I)^{-1}X^T\vec{y} = \vec{x^0}^T(X^TX - \lambda I)^{-1}X^T(X\vec{\theta} + \vec{\varepsilon}) \\
&= [\vec{\theta}^T\vec{x^0} - E_D[\vec{x^0}^T(X^TX + \lambda I)^{-1}X^TX\vec{\theta} + \vec{x^0}^T(X^TX + \lambda I)^{-1}X^T\vec{\varepsilon}]]^2 \\
&= [\vec{\theta}^T\vec{x^0} - E_D[\vec{x}^{0^T}(X^TX + \lambda I)^{-1}X^TX\vec{\theta}] + 0]^2 \\
&\text{and } \vec{x}^{0^T}(X^TX + \lambda I)^{-1}X^TX\vec{\theta} \text{ is a constant} \\
&= [\vec{\theta}^T\vec{x^0} - \vec{x}^{0^T}(X^TX + \lambda I)^{-1}X^TX\vec{\theta}]^2
\end{aligned}$$

and,

$$\begin{aligned}
E_D[E_D[\hat{y}^0] - \hat{y}^0]^2 &= E_D[\vec{x}^{0^T}(X^TX + \lambda I)^{-1}X^TX\vec{\theta} - \hat{y}^0]^2 \\
&= E_D[\vec{x^0}^T(X^TX + \lambda I)^{-1}X^T\vec{\varepsilon}]^2 \\
&= E_D[[\vec{x^0}^T(X^TX + \lambda I)^{-1}X^T\vec{\varepsilon}]^2] \\
&= E_D[\vec{\varepsilon}^2][\vec{x^0}^T(X^TX + \lambda I)^{-1}X^T]^2 \\
&\text{since } Var(\varepsilon) = E[(\varepsilon - E[\varepsilon])^2] = E[(\varepsilon - 0)^2] = E[\varepsilon^2] = \sigma^2 \\
&= \sigma^2(\vec{x^0}^T(X^TX + \lambda I)^{-1}X^T)(\vec{x^0}^T(X^TX + \lambda I)^{-1}X^T)^T \\
&= \sigma^2\vec{x^0}^T(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}X^T\vec{x^0}.
\end{aligned}$$

Putting it all together, we get:

$$= \sigma^2 + [\vec{\theta}^T\vec{x^0} - \vec{x}^{0^T}(X^TX + \lambda I)^{-1}X^TX\vec{\theta}]^2 + \sigma^2\vec{x^0}^T(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}X^T\vec{x^0}$$

.

□

Note. Problem 2
With the following assumptions,

$$\hat{h}(\vec{x}) = \vec{B}^TX\vec{x}.$$

Define the cost function as:

$$J(\hat{B}) = \sum_{j=1}^{N} \left(\hat{h}(\vec{x}) - y^i\right)^2 + \lambda \mid X^T \vec{B} \mid^2 .$$

$$= (XX^T\vec{B} - \vec{y})^T(XX^T\vec{B} - \vec{y}) + \vec{B}^T X\lambda X^T\vec{B}.$$

1. Question 1: Find $\hat{B}$ that minimizes cost:

   Note. Recall:

   $$\frac{d}{d\vec{x}}\vec{x}^T A\vec{x} = (A + A^T)\vec{x}.$$

   $$\frac{d}{d\vec{x}}\vec{x}^T\vec{y} = \vec{y}^T.$$

   If A symmetric, then $\frac{d}{d\vec{x}}\vec{x}^T A\vec{x} = 2A\vec{x}$

   Proof.

   $$\begin{aligned}
   \nabla J\left(\vec{B}\right) &= \frac{d}{dB}\left((XX^T\vec{B} - \vec{y})^T(XX^T\vec{B} - \vec{y}) + \vec{B}^T X\lambda X^T\vec{B}\right)\\
   &= \frac{d}{dB}\left(\left(\vec{B}^T XX^T - \vec{y}^T\right)(XX^T\vec{B} - \vec{y}) + \vec{B}^T X\lambda X^T\vec{B}\right)\\
   &= \frac{d}{dB}\left(\vec{B}^T(XX^T)^2\vec{B} - \vec{y}^T XX^T\vec{B} - \vec{B}^T XX^T\vec{y} + \vec{y}^T\vec{y} + \vec{B}^T X\lambda X^T\vec{B}\right)\\
   &= 2(XX^T)^2\vec{B} - 2\vec{y}^T XX^T + 2\vec{B}X\lambda X^T\\
   &= 0.
   \end{aligned}$$

   So,

   $$(XX^T)^2\vec{B} - \vec{y}XX^T + \vec{B}X\lambda X^T = 0.$$

   $$(XX^T)^2\vec{B} + \vec{B}X\lambda X^T = \vec{y}XX^T.$$

   $$\hat{\vec{B}} = \vec{y}XX^T((XX^T)^2 + X\lambda X^T)^{-1}.$$

   □

2. Question 2: Find degrees of freedom of $\hat{h}(\vec{x})$

   Proof. We know:

   $$\begin{aligned}
   \hat{h}(\vec{x}) &= \vec{B}^T X\vec{x}\\
   &= \vec{y}XX^T((XX^T)^2 + X\lambda X^T)^{-1}X\vec{x}.
   \end{aligned}$$

   Substituting into the main formula gives:

   $$\begin{aligned}
   df(\hat{h}(\vec{x})) &= \frac{1}{\sigma^2}Tr(Cov(\hat{\vec{y}}, \vec{y}))\\
   &= \frac{1}{\sigma^2}Tr(Cov(\vec{y}XX^T((XX^T)^2 + X\lambda X^T)^{-1}X\vec{x}, \vec{y}))\\
   &= \frac{1}{\sigma^2}Tr(XX^T((XX^T)^2 + X\lambda X^T)^{-1}X\vec{x}Cov(\vec{y}, \vec{y}))\\
   &= Tr(XX^T((XX^T)^2 + X\lambda X^T)^{-1}X\vec{x})\\
   &= Tr(\vec{x}XX^T X((XX^T)^2 + X\lambda X^T)^{-1}).
   \end{aligned}$$

This makes sense because if $\lambda = 0$ then this reduces to $Tr(X^T XX^T X((XX^T)^2 - 0)^{-1}) = Tr(I_d) = d$ the same as OLS. When $\lambda > 0$, then $(XX^T)^2 + X\lambda X^T)$ increases in magnitude so $((XX^T)^2 + X\lambda X^T)^{-1})$ decreases in magnitude resulting in the final sum of diagonal elements decreasing. $\square$