CMSE 820 HW1

This HW is due on Sep 14th at 11:59 pm.

Question 1: Show that we can decompose the expected prediction error, $\mathbb{E}[(Y_0 - \hat{f}(x_0))^2]$ at an input point $X = x_0$ for a general model $Y = f(X) + \epsilon$ with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$:

$$\mathbb{E}[(Y_0 - \hat{f}(x_0))^2] = \sigma^2 + \text{Var}(\hat{f}(x_0)) + \text{Bias}^2,$$

Question 2: Give the joint probability density, $f_{X,Y}(x, y)$, for X and Y, prove that

$$\hat{f}(X) = \arg\min_f \mathbb{E}\Big[(Y - f(X))^2\Big] = \mathbb{E}(Y|X)$$

Question 3:Consider the usual linear regression setup, with response vector $\mathbf{y} \in \mathbb{R}^n$ and predictor matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$. Let $x_1, \ldots, x_p$ be the rows of $\mathbf{X}$. Suppose that $\hat{\beta} \in \mathbb{R}^p$ is a minimzer of the least squares criterion

$$\|\mathbf{y} - \mathbf{X}^T\beta\|^2.$$

a. Show that if $v \in \mathbb{R}^p$ is a vector such that $\mathbf{X}^T v = 0$, then $\hat{\beta} + c \cdot v$ is also a minimizer of the least squares criterion, for any $c \in \mathbb{R}$.

b. If $x_1, \ldots, x_p \in \mathbb{R}^n$ are linearly independent, then what vectors $v \in \mathbb{R}^p$ satisfy $\mathbf{X}^T v = 0$? We assume $p \leq n$.

c. Suppose that $p > n$. Show that there exists a vector $v \neq 0$ such that $\mathbf{X}^T v = 0$. Argue, based on part (a), that there are infinitely many linear regression estimates. Further argue that there is a variable $i \in \{1, \ldots, p\}$ such that the regression coefficient of variable $\beta_{[i]}$ can have different signs, depending on which estimate we choose. Comment on this.

Question 4: Prove that if we center the rows of $\mathbf{X}$ by replacing $x_{i[j]}$ with $x_{i[j]} - \bar{x}_{[j]}$ where $\bar{x}_{[j]} = \sum_{i=1}^n x_{i[j]}/n$, then the intercept estimate ends up just being $\hat{\beta}_{[1]} = \bar{y} = \sum_{i=1}^n y_{[i]}/n$ for the ridge regression as follows:

$$\hat{\beta} = \arg\min_\beta \|\mathbf{y} - \beta_{[1]}\mathbf{1} - \widetilde{\mathbf{X}}^T\beta_{-[1]}\|^2 + \lambda\|\beta_{-[1]}\|^2,$$

where $\widetilde{\mathbf{X}}$ is a $(p-1)\times n$ matrix removing the first row of $\mathbf{X}$ which are all 1's(we also assume $\mathbf{X}$ is row-centered) and $\beta_{-[1]}$ is a $(p-1)-$dimensional vector after removing the first element of β.

Question 5: Implement the following model (you can use any language)

$$y_i = \beta^*_{[1]} x_{i[1]} + \beta^*_{[2]} x_{i[2]} + \epsilon_i,$$

where $\mathbb{E}(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = 1$, $\text{Cov}(x_i, x_j) = 0$ and $\beta = (-1, 2)^T$. We also assume $x_i \sim N(0, \Sigma_x)$ with

$$\Sigma_x = \text{Cov}(x_i) = \begin{pmatrix} 1 & 0.9999 \\ 0.9999 & 1 \end{pmatrix}.$$

We repeated the following 2000 times:

- Generate $\mathbf{y} = (y_1, \ldots, y_{50})^T$ and $\mathbf{X} = (x_1, \ldots, x_{50})$.
- compute and record $\hat{\beta}^{\text{ols}}$ and $\hat{\beta}^{\text{ridge}}$ (for ridge regression, choose $\lambda = 0.005$).

Then report the followings:

a. The histograms for $\hat{\beta}^{\text{ols}}_{[1]}$ and $\hat{\beta}^{\text{ridge}}_{[1]}$. What conclusion can you make from these histograms?

b. For each replicate of the 2000 repeats, compare $|\beta^*_{[1]} - \hat{\beta}^{\text{ols}}_{[1]}|$ with $|\beta^*_{[1]} - \hat{\beta}^{\text{ridge}}_{[1]}|$. How many times does ridge regression return a better estimate of $\beta^*_{[1]}$?