CMSE 820 HW1

This HW is due on Sep 13th at 11:59 pm.

Question 1: Assume $Y = X^T\beta + \epsilon$, where $X \in \mathbb{R}^P$ is not randomand and $\epsilon \sim N(0,1)$. Given i.i.d. data $\{(x_1, y_1), \dots (x_n, y_n)\}$, we would like to estimate $\beta \in \mathbb{R}^p$ through maximum likelihood framework. Write down the joint log likelihood and compare it with least square method.

Question 2: 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 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 $x_1, \dots, x_p$ 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 $x_1, \dots, x_p \in \mathbb{R}^n$ 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 $i \in \{1, \dots, p\}$ such that the regression coefficient of variable $\beta_{[i]}$ can have different signs, depending on which estimate we choose. Comment on this.

Question 4: Implement the following model (you can use any language)

$$Y = X^T\beta + \epsilon,$$

where $\epsilon \sim N(0,1)$, $X \sim N(0, I_{p\times p})$ and $\beta \in \mathbb{R}^p$ with $\beta_{[1]} = 1$, $\beta_{[2]} = -2$ and the rest of $\beta_{[j]} = 0$. Based on this setting, let's start with $p = 5$ and simulate $\{x_1, \dots, x_{100}\}$ and store it. Then do the followings

(1) Based on the β and $\{x_1, \dots, x_{100}\}$, we first simulate the corresponding Y's and calculate the $\hat{\beta}^{ols}$.

(2) Use the same $\{x_1, \dots, x_{100}\}$, we then simulate another set of $\tilde{Y} = \{\tilde{y}_1, \dots, \tilde{y}_{100}\}$ and calculate the in-sample prediction error using $\hat{\beta}^{ols}$ calculated in (1). This is one realization of PE_{in}.

(3) Repeat (1) - (2) 5000 times and take average of those 5000 calculated PE_{in}. You have an approximate PE_{in}

(4) Repeat the same procedure for $p = 10, 40, 80$. What is the trend for the PE_{in}? Comment your findings.

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 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 $\beta^*_{[1]}$?