Lee Jae Yeol
20308109
Math 4432 Assignment 2

Ch. 5

1. Prove that $\alpha = \dfrac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$ minimizes $Var(\alpha X + (1-\alpha)Y)$.

$Var(X) = \sigma_X^2$ , $Var(Y) = \sigma_Y^2$ , $COV(X,Y) = \sigma_{XY}$

By the property of variance : $Var(aX+b) = a^2 Var(X)$

$$Var(\alpha X + (1-\alpha)Y) = \alpha^2 Var(X) + (1-\alpha)^2 Var(Y) + 2\alpha(1-\alpha) COV(X,Y)$$

$$= \alpha^2 \sigma_X^2 + (1-\alpha)^2 \sigma_Y^2 + 2\alpha(1-\alpha)\sigma_{XY}$$

$$\frac{\partial Var(\alpha X + (1-\alpha)Y)}{\partial \alpha} = 2\alpha\sigma_X^2 - 2\sigma_Y^2 + 2\alpha\sigma_Y^2 + 2\sigma_{XY} - 4\alpha\sigma_{XY}$$

$$= 2\alpha(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}) - 2(\sigma_Y^2 - \sigma_{XY})$$

$$= 0$$

$$\Rightarrow \alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

Perform a second derivative

$$: \frac{\partial^2}{\partial\alpha^2} = 2(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}) = 2 Var(X-Y)$$

which is a non-negative value, stating that $\alpha$ is a minimum value.

$$\therefore \alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$ minimizes $Var(\alpha X + (1-\alpha)Y)$.

2. a) Out of the set of $n$ observations, the probability that the bootstrap observation is the $j$th-observation is $\frac{1}{n}$

Therefore, the probability that it is NOT is $\therefore 1 - \frac{1}{n}$

b) The bootstrap observations may be the same sample since they are drawn randomly. Therefore the probability is actually same as the previous answer. $\therefore 1 - \frac{1}{n}$

c) Since the probability that $j$th observation is not the $j$th bootstrap observation is $1-\frac{1}{n}$ and it is equal for all $n$ bootstrap observations, the probability will be just a product of $1-\frac{1}{n}$. $\therefore (1-\frac{1}{n})^n$

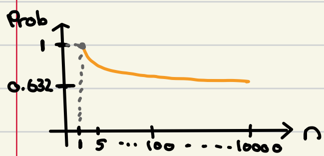
d) The probability that $j$th observation will be selected as bootstrap observation is $1-(1-\frac{1}{n})^n$.

If $n=5$ : $1-(1-\frac{1}{5})^5 = 0.6723$ $\therefore 0.6723$

e) $n=100$ : $1-(1-\frac{1}{100})^{100} = 0.633967$ $\therefore 0.634$

f) $n=10000$ : $1-(1-\frac{1}{10000})^{10000} = 0.632$ $\therefore 0.632$

g) $n=2$ : $1-(1-\frac{1}{2})^2 = 0.75$ , $n=3$ : $1-(1-\frac{1}{3})^3 = 0.7037$




h) As $n \to \infty$, $\lim_{n\to\infty}(1-\frac{1}{n})^n$ is in the form $e^{-1}$.

Therefore, the probability will converge to $1-e^{-1} = 0.632$

5 & 6 are done in .ipynb

Ch. 6

1. Best Subset / Forward Stepwise / Backward Stepwise Selections

a) The Selection with the smallest training RSS is the best subset selection since it takes into account all $2^k$ possible models and select the best one (i.e smallest training RSS) among them.

b) We may assume for the similar reason above that the best subset selection yields the smallest test RSS. However, there is no certain answer since the performance is based on test set.

c) i) True - The k+1 variable model comes from k variable model in forward stepwise selection.

ii) True - the k variable model comes from k+1 variable model in backward stepwise selection.

iii), iv) False - Forward stepwise selection and backward stepwise selection are not related.

v) False - In the best subset selection, different variable sized models are selected independently. They are not related.

2. a) Lasso Model is less flexible relative to least square and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance. ∴ iii)

b) For Ridge model, it is same as Lasso model ∴ iii)

c) For non-linear model, it is more flexible relative to least square and will give improved prediction accuracy when its increase in variance is less than the decrease in bias. ∴ ii)

3. $\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$ subject to $\sum_{j=1}^{p} |\beta_j| \le s$

As we increase $s$ from 0 :

a) The Training RSS will iv) Steadily decrease since increase in the constraint $s$ means the model gets more flexible. Therefore, test RSS will decrease.

b) The test RSS will ii) decrease initially but eventually starts increasing in a U-shape. It forms a U-shape because after some point, the overfitted model will have a higher variance.

c) The variance will iii) Steadily increase because as the model gets more flexible, it means that the model starts to overfit.

d) The squared bias will iv) Steadily decrease with the same reason for the training RSS a).

e) The irreducible error is independent and invariant of the model. So it will v) remain constant.

4. $\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

As we increase $\lambda$ from 0 :

a) The training RSS will iii) steadily increase since the shrinkage penalty increases and the model will become less flexible. Therefore, the training RSS will increase.

b) The test RSS will ii) decrease initially but eventually starts increasing in a U-shape. Initial decrease will happen because variance decreases with respect to the increase in $\lambda$. However after some point, the model will be too underfit and the bias will increase heavily.

c) The variance will iv) steadily decrease because the model becomes less flexible as the shrinkage penalty increases.

d) The square bias will iii) steadily increase for the same reason with the training RSS a).

e) The irreducible error is independent and invariant of the model. So it will v) remain constant.

7. a) $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \quad \sim N(0, \sigma^2)$

$$f(Y \mid X, \beta) = \prod_{i=1}^{n} f(Y \mid X_i, \beta_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left( - \frac{Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{2\sigma^2} \right)$$

$$= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \cdot \exp\left( - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} \left[ Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}) \right]^2 \right)$$

b) From Bayes Rule, we have $\underbrace{P(\theta \mid y)}_{\text{Posterior}} \propto \underbrace{P(y \mid \theta)}_{\text{likelihood}} \cdot \underbrace{P(\theta)}_{\text{Prior}}$

Prior for $\beta$: $\beta_1, \beta_2, \dots \beta_p$ are iid according to a double-exponential distribution with mean 0, $P(\beta) = \frac{1}{2b} \exp\left(-\frac{|\beta|}{b}\right)$

$f(\beta \mid X, Y) \propto f(Y \mid \beta, X) \cdot P(\beta) =$

$$= \frac{1}{2b} \cdot \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \cdot \exp\left( - \frac{|\beta|}{b} - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} \left[ Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}) \right]^2 \right)$$

c) $\underset{\beta}{\arg\max} \log f(Y \mid X, \beta) \cdot P(\beta)$

$$= \underset{\beta}{\arg\max} \left[ \log\left( \frac{1}{2b} \cdot \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \right) - \left( \frac{|\beta|}{b} + \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} \left[ Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}) \right]^2 \right) \right]$$

Since the first log term is irrelevant of $\beta$ and taking $\underset{\beta}{\arg\max}\,(-\gamma\beta) \Rightarrow \underset{\beta}{\arg\min}\,(\gamma\beta)$, we get

$$\underset{\beta}{\arg\min} \quad \frac{|\beta|}{b} + \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} \left[Y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)\right]^2$$

$$= \underset{\beta}{\arg\min} \; \frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} RSS + \frac{2\sigma^2}{b} |\beta| \right)$$

replacing $\frac{2\sigma^2}{b}$ with $\lambda$ we got the lasso estimate.

d) With mean 0 and variance $c$, the prior probability will be

$$p(\beta_i) = \frac{1}{\sqrt{2\pi c}} \cdot \exp\left(-\frac{\beta_i^2}{2c}\right)$$

$$f(\beta \mid Y, X) = f(Y \mid X, \beta) \cdot p(\beta)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \exp\left(-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} \left[Y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)\right]^2\right) \cdot \left(\frac{1}{\sqrt{2\pi c}}\right)^p \cdot \exp\left(-\frac{1}{2c} \sum_{i=1}^{p} \beta_i^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \left(\frac{1}{\sqrt{2\pi c}}\right)^p \cdot \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2} \sum_{j=1}^{n} \left(Y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)\right)^2 + \frac{1}{c} \sum_{i=1}^{p} \beta_i^2\right)\right\}$$

e) $\arg\max \log f(\beta \mid Y, X)$, similar to what we did above in c)

$$\Rightarrow \underset{\beta}{\arg\min} \; \frac{1}{2}\left\{\left(\frac{1}{\sigma^2} \sum_{j=1}^{n} \left(Y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)\right)^2 + \frac{1}{c} \sum_{i=1}^{p} \beta_i^2\right)\right\}$$

$$\Rightarrow \underset{\beta}{\arg\min} \; \frac{1}{2\sigma^2} \left( RSS + \frac{\sigma^2}{c} \cdot \sum_{i=1}^{p} \beta_i^2 \right)$$

replacing $\frac{\sigma^2}{c}$ with $\lambda$ we got the ridge estimate.