HINS Assignment

M. Anusha
1628631

1) a

Best subset selection will have best training RSS, this is because the model will be chosen after considering all the possible models with k parameters for best subset. this is not true for either backward or forward stepwise.

b) its difficult to tell, best selection may have the smallest test RSS because it takes into account more models than the other models However, the other methods might also pick a model with smaller test RSS

c) i) the predictors in the k variable model having (k+1) predictors by augmenting the predictors in the model (True)

ii) the model with k predictors is obtained by removing one predictor from the model (k+1) predictors (True)

iii) there is no link between the models obtained from forward and backward selection (False)

iv) There is no direct link between the models obtained from forward and backward selection (false)

v) the model with (k+1) predictors will be obtained by selecting all possible models with (k+1) predictors so (false)

2)

a) iii) option 3 is correct,
the lasso is more restrictive model as cost function = Los + $\alpha \sum_{i=1}^{n} |a_i|$
that is as selecting best parameters leads to reduce overfitting, As long as it doesnot result to high bias due to its added constraints

b) it is same as lasso so iii) is correct option, because ridge regression having lossfunction = Los + $\alpha \sum_{i=1}^{n} a_i^2$ as it is less flexible and give prediction accuracy when its in bias is less and then its decrease in bias

c) Non-linear methods are more flexible and will give improved prediction accuracy when their increase in variance are less than their decrease in bias.

4)

a) steadily decreases, as we increase s from 0, we are restricting the β_j coefficients less and so the model is becoming more flexible

b) Decreases initially and eventually start increasing in a U shape
As we increase s from 0, we are restricting β_j coefficients less and the model is more flexible

c) steadily increase, As we increase s from 0, we are restricting the β_j coefficients less and so the model is becoming more and more flexible which provokes a steady increase in variance

4) d steadily decreases, as we increase s from 0, we are restricting the β_j coefficients less and so model is more flexible which provokes a steady decrease in bias

e) Remains constant, the error is independant of model but dependent on dataset.

3)

a) steadily increases as we increase λ from 0, we are restricting the β_j coefficients more and more

b decreases intially, and then eventually increasing in u shape becacuse of hyperparameter tuning

c) steadily decrease as we increase λ from 0 it mainly depends on selection of λ value i.e $\lambda=0$ overfitting, to large underfitting so

d) steadily increase, ase we increase λ from 0, as above mentioned due to hyperparameter tuning

e) it remains constant because depends on dataset.

5)

a) for ridge regression to minimize the parameters

$$\Rightarrow \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$= \sum_{j=1}^{p} \left[(y_j - \beta_j)^2 + \lambda \beta_j^2 \right]$$

where $\beta = (\beta_1, \beta_2 \dots \beta_p)$

it can be additively separable

$$V(\beta) = \sum_{j=1}^{p} V(\beta_j)$$

so the derivative with β_j is

$$\frac{d}{d\beta_j}(V\beta) = \frac{d}{d\beta_j} V(\beta_j)$$

thus minimizing with β is equalent to 'p' i.e parameters

$j = 1, 2, 3 \dots p$ so

$$\Rightarrow \frac{d}{d\beta_j} V(\beta_j) = \frac{d}{d\beta_j} \left[(y_j - \beta_j)^2 + \lambda \beta_j^2 \right]$$

$$\Rightarrow \frac{d}{d\beta_j} \left[y_j^2 - 2 y_j \beta_j + (1+\lambda) \beta_j^2 \right]$$

$$\Rightarrow -2y_j + 2(1+\lambda)\beta_j$$

setting to 0 then $2(1+\lambda)\beta_j - 2y_j = 0$

$$\beta_j = \frac{y_j}{1+\lambda}$$

b) For lasso regression,

$$l(\beta) = \sum_{j=1}^{p} \left((y_j - \beta_j)^2 + \lambda |\beta_j| \right)$$

thus for $j = 1, 2 \ldots p$ we must find derivate as similar to Ridge regression i.e

$$\frac{d}{d\beta_j} l(\beta_j) = \frac{d}{d\beta_j} \left[(y_j - \beta_j)^2 + \lambda |\beta_j| \right] = \frac{d}{d\beta_j} \left[y_j^2 + \beta_j^2 - 2 y_j \beta_j + \lambda |\beta_j| \right]$$

Because of the $-\beta_j y_j$ term in the above, we choose β_j to have same sign as y_j to preserve the formation of the problem

1. suppose that $y_j > 0$ then $j = 1, 2 \ldots p$ we must minimize

$$l(\beta_j) = y_j^2 - 2 y_j \beta_j + \beta_j^2 + \lambda \beta_j$$

since $|\beta_j| = \beta_j$ when $\beta_j > 0$ the derivative is

$$\begin{aligned} l'(\beta_j) &= -2 y_j + 2\beta_j + \lambda \\ &= 2 \left[\beta_j - (y_j - \lambda/2) \right] \end{aligned}$$

if $|y_j| \leq \lambda/2$ then $(y_j - \lambda/2) > 0$ so that $l'(\beta) > 0$ giving

$$\beta_j = y_j - \lambda/2 \text{ if } y_j > \lambda/2$$

similarly, $y_j < 0$ we minize

$$l(\beta_j) = y_j^2 - y_j \beta_j + \beta_j^2 - \lambda \beta_j$$

5) Given, $p = 2, 4, 6, 8, 10$

$n = 20$

Given, $AIC = n\ln(RSS/n) + 2(p+1)$ $\quad BIC = n\ln(RSS/n) + \ln(p+1)$

if $p = 2$

$AIC = 20\ln(220/2) + 2(3)$

$= 53.24$

$BIC = 20\ln(220/20) + \ln(3)$

$= 47.70$

if $p = 4$

$AIC = 20\ln(200/20) + 2(5)$

$= 56.05$

$BIC = 20\ln(200/20) + \ln(5)$

$= 47.65$

if $p = 6$

$AIC = 20\ln(190/20) + 2(7)$

$= 69.03$

$BIC = 20\ln(190/20) + \ln(7)$

$= 46.97$

if $p = 8$

$AIC = 20\ln(187/20) + 2(9)$

$= 62.707$

$BIC = 20\ln(187/20) + \ln(9)$

$= 46.98$

if $p = 10$

$AIC = 20\ln(186.8/20) + 2(11)$

$= 59.84$

$BIC = 20\ln\left(\frac{186.8}{20}\right) + \ln(11)$

$= 45.97$

2	4	6	8	10
220	200	190	187	186.8
53.05 47.65	56.05 47.65	59.03 46.97	62.707 46.98	59.84 45.97

AIC = $n\ln(RSS/n) + 2(p+1)$

BIC = $n\ln(RSS/n) + \ln(p+1)$ → Increases if p is more

↓ decreases if p is more

Among AIC and BIC,

if p is large or then AIC is better, if p is smaller then BIC is better

∴ Best to be

AIC	BIC
larger p	smaller p