Name: Jay Joshi

Course: Machine Learning Assignment 2

Student Id: 200440993

Q1

(a) Best subset selection will have smallest training RSS among all the models because the best subset selection model evaluates all the possible models and selects the one with lowest RSS. However, in forward stepwise selection the model with k predictors is the model with smallest RSS among p-k models and not all the possible models are evaluated. Also in backward stepwise selection, model with k predictors is the model with smallest RSS among k-models, hence there are only few possibilities considered in backward -stepwise selection as well.

(b) According to me, since in forward stepwise and backward stepwise are greedy approach, we might not be able to find best possible combinations, there is higher probability that the lowest test RSS will be captured by Best subset selection. However this might not be the case always. Hence it is difficult to choose.

Scanned with Ca

(C) True or False.

i) True. Because since we are augmeting / adding features which are best from previous evaluation, the predictors of k-variable model are subset of predictors of k+1 variable model.

ii True. Since we are removing the least best feature from previous round in backward stepwise selection, k-variable model will be subset of k+1 variable model.

iii) False. No such relation is given by which we can prove this result among forward & backward stepwise seleetion. because the least RSS features identified by both the methods can be completely differented and thus this is not the case.

iV False. As said above, the least RSS features identified by both the methods can be completely different this is not the case.

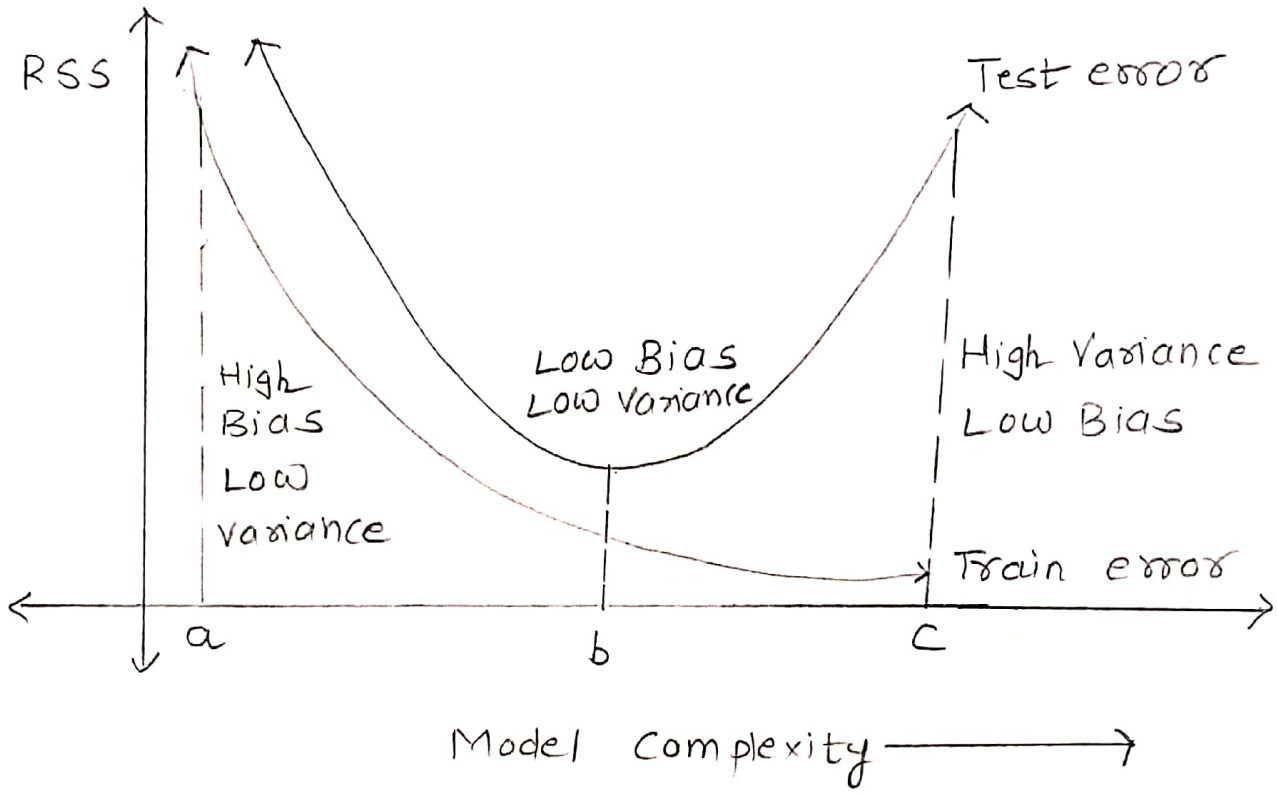
v False. The model identified as best in k-variable subset selection, does not necessarily contains all the features that are in (k+1) variable model identified by best subset selection.

Scanned with Ca

Q2 We are estimating regression coefficients in a Linear regression model by minimizing

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_{ij} x_{ij} \right)^2 \text{ w.r.t to } \sum_{j=1}^{p} |\beta_j| \leq s$$

→ Consider the following graph



As we increase s from 0

(a) The training RSS will

→ (iv) Steadily decrease. Because from the above graph you can see that, since we are minimizing RSS and as the value of s increases from 0, we are adding more complexity in the model subject to model weights, hence from point a to point c training RSS will steadily decrease with increase in model complexity.

Scanned with Ca

(b) The test RSS will

→ (ii) Decrease initially and then eventually start increasing in a U shape. Because, as said earlier, since we are increasing the value of s from 0, we are adding more complexity to the model, so as we will traverse from point a to point c on X-axis, test error, will decrease initially and when the model will start capturing noise at point b with increasing model complexity, test error will start increasing in U-shape.

(c) The variance will,

→ Steadily Increase (iii), because since we are increasing s, there will be more variation among the weights and hence variance will increase as we more from point a to point c. Since we are increasing value of s, which will increase sensitivity and will result in high variation among models and it will increase variance.

(d) The squared bias will

→ (iv) steadily decrease. Since we are increasing the model complexity with increase in s, it will increase model flexibility, and since model is more flexible, the squared bias will decrease steadily.

Scanned with Ca

This phenomenon is shown in graph as we traverse from point a to point c.

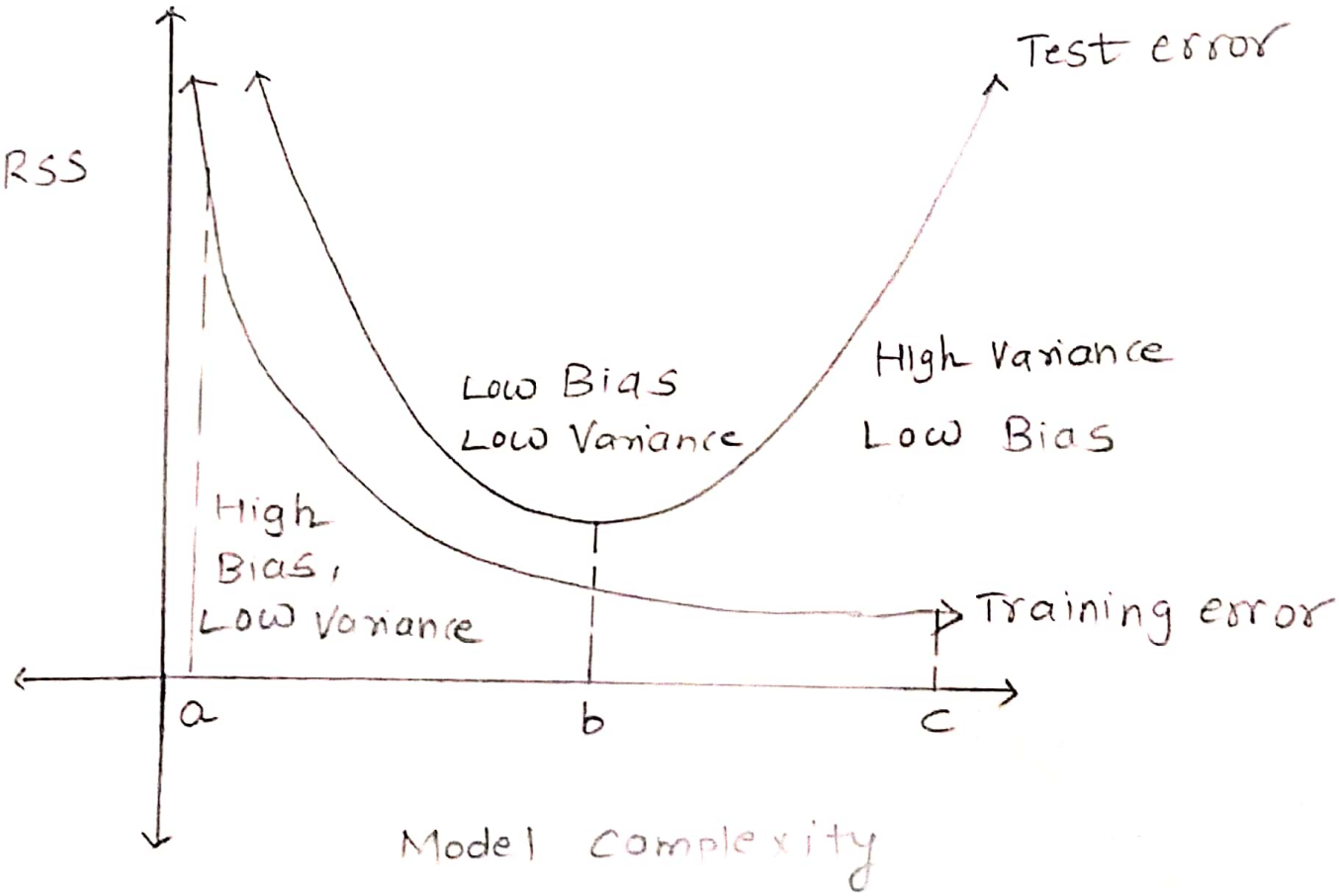
(e) Irreducible error will

→ (r) Remains constant. Since data = Information + noise,

every, data will contain some amount noise, which does not depends upon model complexity and hence it is independent of "s" ~~noise~~. Thus, noise will remain constant.

(3) Suppose we are estimating the regression coefficients by minimizing

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



Scanned with Cam

Consider the following graph as basis for explanation. Since we are using ridge regression to minimize cost function to resolve the issue of model overfitting by changing the values of λ. It's neeessary to note that in model overfitting we are at point C in graph and we need to travel towards point B to resolve issue of model overfitting. Hence term $\lambda \|w\|^2$ helps us to travel towards point B on x-axis. In other words we are adding Bias to make our model more flexible.

(a) The training RSS will

→ (iii) Steadily Increase. Since we are increasing value of λ, we are increasing flexibility of model and we travel from point C to point b. Hence training RSS will steadily increase with increase in flexibility of model.

(b) The test RSS will

→ (ii) Decrease initially and then eventually start increase in a U-shape. Since λ is a regularizer for finding optimal weights, by increasing value of λ weights will decrease, hence less variation among model weights, and hence due to decreasing sensitivity test RSS will decrease initially, but as model flexibility decreases due to large values

Scanned with Ca

of λ, test RSS will start increasing in U-shape

(c) The variance will,

→ (iv) Steadily decrease. Since λ is optimizer for model weights, the model weights will decrease with increase in values of λ, hence variation among model weights will decrease, which will decrease sensitivity and hence variance will decrease steadily.

(d) The squared bias will,

→ (iii) Steadily increase. Since it is already explained that term $\lambda |w|^2$ is adding bias to increase model flexibility and decrease sensitivity. to resolve overfitting issue. Hence, with increase in value of λ squared bias will increase.

(e) The irreducible error will,

→ (v) Remains constant. Since data = Information + noise,

every data will contain some amount of noise, which is independent of model complexity and hence independent of value of λ. Thus, noise will remain constant.

Scanned with Ca

Q4 Coordinate descent algorithm for unnormalized data.

1) Coordinate descent algorithm for RSS when data is not normalized.

$$RSS(w) = \sum_{i=1}^{N} \left(y_i - \sum_{j=0}^{D} w_j h_j(x_i)\right)^2$$

differentiating w.r.t to w_j

$$\frac{\partial RSS(w)}{\partial w_j} = \sum_{i=1}^{N} 2\left(y_i - \sum w_j h_j(x_i)\right)\left[0 - h_j(x_i)\right]$$

$$\therefore \frac{\partial RSS(w)}{\partial w_j} = -2 \sum_{i=1}^{N} h_j(x_i)\left[y_i - \sum_{\substack{k=0 \\ k \neq j}}^{D} w_k h_k(x_i) - w_j h_j(x_i)\right]$$

$$= -2 \sum_{i=1}^{N} h_j(x_i)\left[y_i - \sum_{\substack{k=0 \\ k \neq j}}^{D} w_k h_k(x_i)\right] +$$

$$2 w_j \sum_{i=1}^{N} h_j(x_i)^2$$

Now, $-2 \sum_{i=1}^{N} h_j(x_i)\left[y_i - \sum_{\substack{k=0 \\ k \neq j}}^{D} w_k h_k(x_i)\right] = \rho_j$

$$\therefore \quad -2\rho_j + 2 w_j \sum_{i=1}^{N} h_j(x_i)^2 \qquad \text{——— (1)}$$

Now equating with zero

$$-2\rho_j + 2\hat{w}_j \sum_{i=1}^{N} h_j(x_i)^2 = 0$$

$$\therefore \quad \hat{w}_j = \frac{\rho_j}{\sum_{j=1}^{N} h_j(x_i)^2}$$

Scanned with Cam

Hence, coordinate descent algorithm for RSS with unnormalized data is:

Algorithm:

Initialize w

while not converged:

pick a coordinate j (say, you can pick in round robin fashion)

$$w_j \leftarrow \frac{\rho_j}{\sum_{i=1}^{N} h_j(x_i)^2}$$

2) Coordinate descent algorithm for Ridge regression when data is not normalized

Cost function in Ridge regression is

$$\text{cost}(w) = RSS(w) + \lambda \sum_{j=0}^{D} |w_j|^2$$

As seen earlier from equation ——C1)

$$\frac{\partial RSS(w_j)}{\partial w_j} = -2\rho_j + 2w_j \sum_{i=1}^{N} h_j(x_i)^2$$

Hence

$$\frac{\partial \text{Cost}(w_j)}{\partial w_j} = -2\rho_j + 2w_j \sum_{i=1}^{N} h_j(x_i)^2 + 2\lambda w_j$$

Now, $-2\rho_j + 2\hat{w}_j \sum_{i=1}^{N} h_j(x_i)^2 + 2\lambda \hat{w}_j = 0$

$$\therefore \quad \hat{w}_j = \frac{\rho_j}{\sum_{i=1}^{N} h_j(x_i)^2 + \lambda}$$

Scanned with Cam

Algorithm for ridge regression:

Algorithm:

Initialize w

while not converged:

Pick a coordinate j

$$w_j \leftarrow \frac{\rho_j}{\sum_{i=1}^{N} h_j(x_i)^2 + \lambda}$$

3) Coordinate descent algorithm for Lasso when data is not normalised.

Cost function in Lasso regression is

$$cost(w) = RSS(w) + \lambda \sum_{j=0}^{D} |w_j|$$

As seen earlier from equation —(1)

$$\frac{\partial RSS(w_j)}{\partial w_j} = -2\rho_j + 2w_j \sum_{i=1}^{N} h_j(x_i)^2$$

Now using sub gradients we get

$$\lambda\, \partial|w_j| = \begin{cases} -\lambda & \text{if } w_j < 0 \\ [-\lambda, \lambda] & \text{if } w_j = 0 \\ +\lambda & \text{if } w_j > 0 \end{cases}$$

(sub gradient generalizes gradients to non-differentiable points)

Hence

$$\frac{\partial\, cost(w)}{\partial w_j} = -2\rho_j + 2w_j \sum_{i=1}^{N} h_j(x_i)^2 + \begin{cases} -\lambda & w_j < 0 \\ [-\lambda, \lambda] & w_j = 0 \\ +\lambda & w_j > 0 \end{cases}$$

Scanned with Cam

By equating with zero;

$$-2\rho_j + 2\omega_j \sum_{i=1}^{N} h_j(x_i)^2 - \lambda = 0 \quad \text{if } \omega_j < 0$$

$$[-2\rho_j - \lambda \ , \ -2\rho_j + \lambda \ ] \quad \text{if } \omega_j = 0$$

$$-2\rho_j + 2\omega_j \sum_{i=1}^{N} h_j(x_i)^2 + \lambda = 0 \quad \text{if } \omega_j > 0$$

Now ω_j if $\omega_j < 0$

$$\hat{\omega}_j \sum_{i=1}^{N} h_j(x_i)^2 = \rho_j + \frac{\lambda}{2}$$

$$\therefore \hat{\omega}_j = \frac{\rho_j + \frac{\lambda}{2}}{\sum_{j=1}^{N} h_j(x_i)^2}$$

Since $\omega_j < 0$ & $\sum_{j=1}^{N} h_j(x_i)^2 > 0$

$\rho_j + \frac{\lambda}{2} < 0 \quad \therefore \rho_j < -\frac{\lambda}{2}$

Now ω_j if $\omega_j > 0$

$$\hat{\omega}_j \sum_{i=1}^{N} h_j(x_i)^2 = \rho_j - \frac{\lambda}{2}$$

$$\hat{\omega}_j = \frac{\rho_j - \frac{\lambda}{2}}{\sum_{i=1}^{N} h_j(x_i)^2}$$

Since $\omega_j > 0$ & $\sum_{j=1}^{N} h_j(x_i)^2 > 0$

$\rho_j + \frac{\lambda}{2} > 0 \quad \therefore \rho_j > -\frac{\lambda}{2}$

Algorithm for Lasso regression:

Initialize w

while not converged:

Pick a coordinate j

if $\omega_j < 0$

$$\hat{\omega}_j = \frac{\rho_j + \lambda/2}{\sum_{j=1}^{N} h_j(x_i)^2}$$

Scanned with Cam

if $w_j > 0$

$$\hat{w}_j \leftarrow \frac{\rho_j - \frac{\lambda}{2}}{\sum_{i=1}^{N} h_j(x_i)^2}$$

Scanned with Ca