University of Regina

ENEL/ENSE 865: Applied Machine Learning

Professor: Dr. Abdul Bais

Instructional Designer/TA: Dr. Muhammad Hamza Asad

Submitted by:

Name: Nandish Bakulkumar Bhatt
Student ID: 200441204

Writing Assignment:2

Name:- Nandish . Bakulkumar . Bhatt.

Student ID: 20041204

Q:1

1(a) Best subset selection will have the smallest training RSS when compared with all the models.

→ The main reason behind this is that, the model will be selected after consideration of the all possible models with k parameters for best subset.

→ The model (best subset selection model) evaluates all possible models with k-parameters and selects the one with lowest RSS.

→ In opposite, forward stepwise selection is the model with k-predictors has smallest RSS among p-k models and thus not all the possible models are evaluated. Moreover, in backward stepwise selection, is the model with k-predictors have the smallest RSS among k-models.

1(b)

→ Best subset selection might have smallest test RSS, as it takes into consideration more models than forward stepwise selection and backward stepwise selection.

→ Moreover, these forward and backward stepwise selection are greedy approach, so they are also not able to find the best possible combination.

→ Thus, it is very hard to decide that which has smallest test RSS.

1(c)

(i) True.

- The model with (k+1) predictors can be obtained by adding the predictors in model with k-predictors with one additional predictor.

(ii) True.

- The predictors in k-variable model identified by backward stepwise are a subset of predictors in (k+1)-variable model identified by backward stepwis selection is true, because the method implementing in backward stepwise selection is removing ~~the~~ one predictor from model with (k+1) predictors.

(iii) False.

— Actually, there is no relationship between the stepwise forward and stepwise backward selection.

(iv) False.

— Same as above, as there is no relation between the forward stepwise and backward stepwise selection.

(v) False.

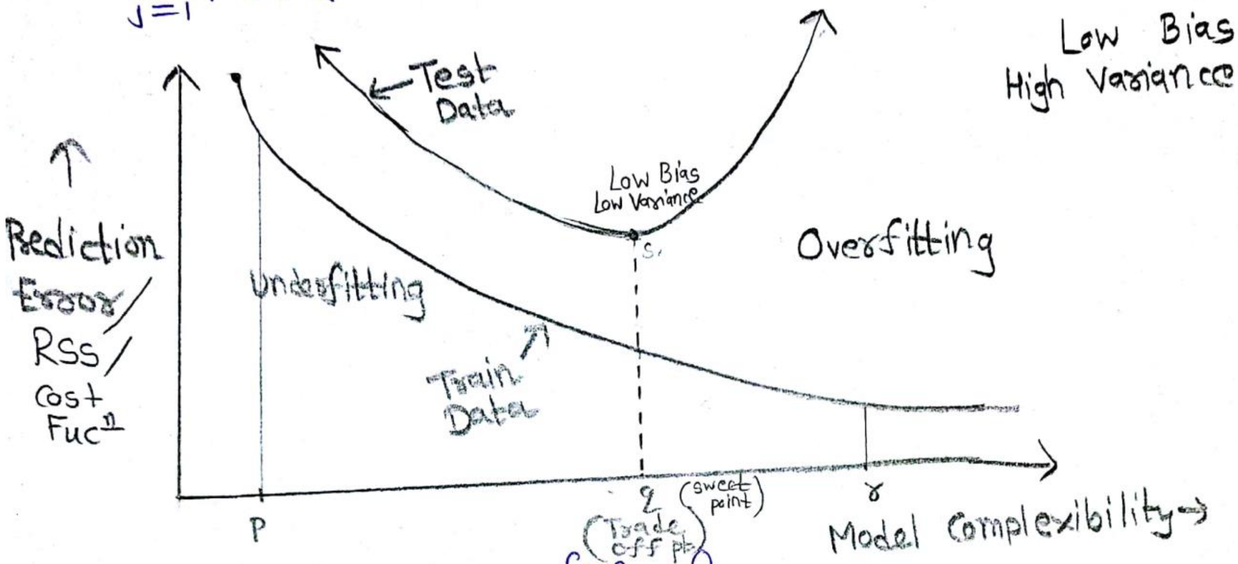
— The model which is assumed to be best in k-variable sub-set selection, does not necessarily contains all the features that are in (k+1)-variable model identified by best subset selection.

Q-2

→ Given that,

Minimizing Equation:
$$\sum_{i=1}^{n}\left[y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right]^2 \text{ subject to}$$

$\sum_{j=1}^{p}|\beta_j| \leq s$ for particular value of s.



→ As we increase s from 0,

2(a) The training RSS will (iv) steadily decrease.

as s increases from 0,

Reason: As shown in graph, as s increases from 0, the model complexity increases i.e., we are adding more new parameters, so training RSS will decrease. Thus, training RSS will steadily decreases.

2(b) The test RSS will (ii) decrease initially and then eventually start increasing in a U-shape.

Reason: As shown in graph, as s increases from 0, initially the model complexity is less, so the test RSS will decrease initially. After Bias-Variance Trade off Point, when s is still increasing, the model complexity is increasing and thus new parameters are adding. But, model will retain its training data and it will not be able to capture test data. So, test RSS then eventually increasing in U-shape

2(c) The Variance will (iii) steadily increase.

Reason: As we increase s from 0, the model complexity increases, the new parameters are added and thus it will create more variation in weights and this is in turn will increase variance.

- By mathematically, as s increases from 0, we are restricting β_j coefficients less and less, and model becomes more and more flexible which in turn increase variance.

2(d) $(Bias)^2$ will (iv) steadily decrease.

Reason: Now, as we increases s from 0, we are restricting β_j coefficients to lesser values and so model will become more flexible, which results steady decrease in bias.

- Also, from graph, as s increases from 0, the model complexity also increases and bias will steadily decrease.

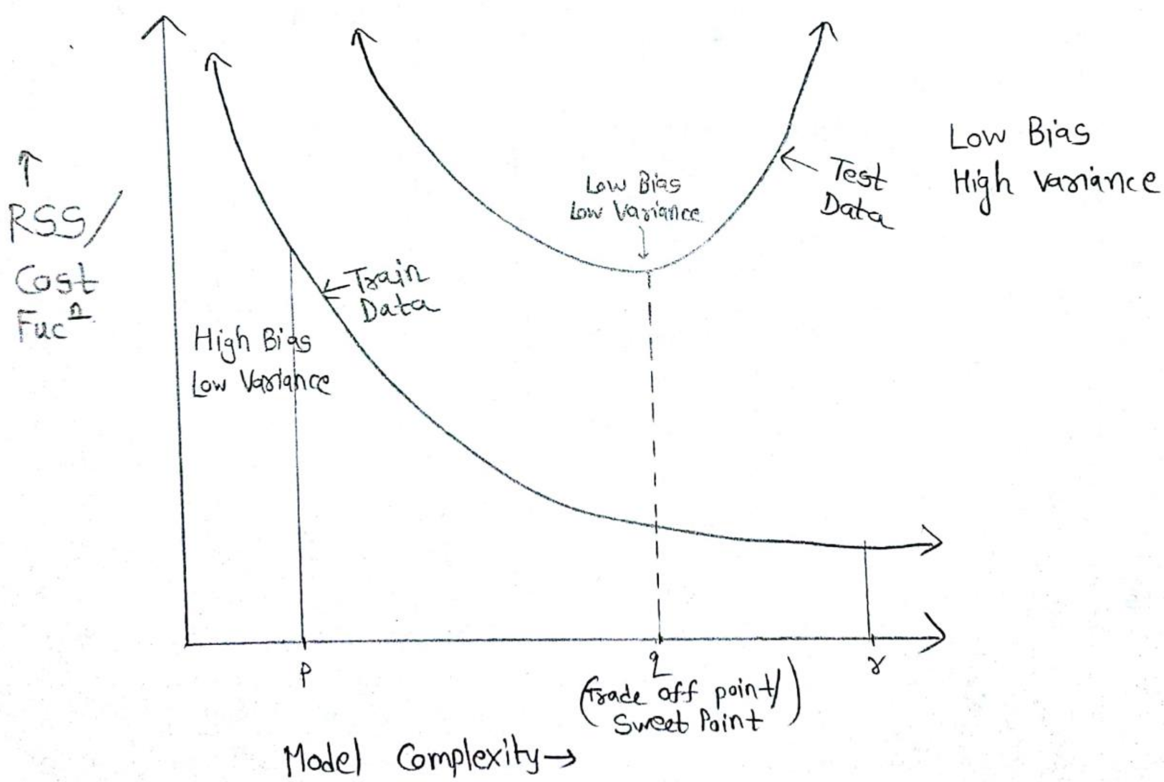
2 (e) Irreducible error will (v) remains constant.

Reason: Now, irreducible error is noise, which is independent of model and any model parameter.
So, it will not dependent on s.
Thus, noise remain constant.

Q:3

Minimizing equation, $$\sum_{i=1}^{n}\left[y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right]^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for particular value of λ.



3(a) As we increase λ from 0, training RSS will (iii) steadily increase.

Reason: since λ increases from 0, then weights vector $||w||^2$ has to decrease, thus the weights are decreasing and so complexity of model decrease, which in turn from graph it is clear that training RSS is steadily increasing. Also, model is becoming less and less flexible.

3(b) Test RSS will (ii) decrease initially and then eventually start increase in a U-shape.

Reason: When λ increases from 0, then weight vector $||w||^2$ has to decrease, and thus the weights are decreasing, so this will decrease the complexity of model. So, graph will travel from point "r" towards "p" and so sensitivity decreases and when graph travels from "r" towards "p" then test RSS will decrease initially and then eventually start increase in a U-shape.

3(c) Variance will (iv) steadily decreases.

Reason: As we increase value of λ from 0, then weight vector $||w||^2$ has to decrease, and thus weights are decreasing, so this ~~d~~ will decrease the complexity of model. Now, complexity decreases, when we travel from "r" towards "p" point in graph, the sensitivity decreases and hence variance will steadily decrease.

3(d) $(Bias)^2$ will (iii) steadily increase.

Reason: As we increases value of λ from 0, then weight vector $\|w\|^2$ has to decrease and thus weights are decreasing, so complexity of model decreases.

- Now, model complexity decreases from "r" to "p" point in graph and thus sensitivity decreases and, bias increases in that region when we travel from "r" to "p". So, $(bias)^2$ will steadily increase in that point.

3(e) Irreducible error will (v) remains constant.

Reason: Since, irreducible error is a noise and it is independent of any parameter of model, and model complexity.

- So, any value of λ won't affect irreducible error.
- Thus, noise will remain constant.

Q:4 Now, we need to find co-ordinate descent algorithm for unnormalized data for least square cost function, Ridge and Lasso.

(1) Co-ordinate Descent Algorithm for Least Square Cost Function using unnormalized data.

→ Now, we know the formula of,

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

Partial Differentiating $RSS(w)$ w.r.t. w_j, so we get—

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

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

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

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

→ Putting eqn (2) into eqn (1),

$$\frac{\partial RSS}{\partial w_j} = -2\rho_j + 2w_j \sum_{j=1}^{N} [h_j(x_i)]^2 \quad \text{——— (3)}$$

→ For finding optimal solution, closed form solution –

$$\frac{\partial RSS}{\partial w_j} = 0 \quad (\because \text{equating with zero for closed form})$$

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

$$\therefore \boxed{\hat{w}_j = \frac{\rho_j}{\sum_{i=1}^{N} [h_j(x_i)]^2}}$$

→ Now, initialize w, while not converged.
Pick a co-ordinate, j, let's say for any- round robin fashion or random pick,

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

(2) Co-ordinate Descent Algorithm for Ridge regression using unnormalized data:-

→ Cost Function in Ridge regression is: $cost = RSS + \lambda \|w\|^2$

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

$$\rightarrow \frac{\partial RSS}{\partial w_j} = -2\rho_j + 2w_j \sum_{i=1}^{N}[h_j(x_i)]^2 \quad \left[\because \text{from eq}^n (3)\right].$$

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

→ For finding optimal solution, for closed form solution, $\frac{\partial RSS(w_j)}{\partial w_j}$ equates to zero.

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

$$\therefore \hat{w}_j \left[\sum_{i=1}^{N}[h_j(x_i)]^2 + \lambda\right] = \rho_j$$

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

→ Now, for gradient descent, initialize w, while not converged.

Pick co-ordinate, j.

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

(3) Co-ordinate descent algorithm for Lasso using unnormalized data

→ For Lasso regression,

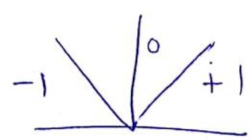
cost Function = RSS + $\lambda \|w\|$.

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

→ Now, from eq^n (3), $\frac{\partial RSS}{\partial w_j} = -2\rho_j + 2w_j \sum_{i=1}^{N} [h_j(x_i)]^2$

→ Now, using sub-gradients, we get—

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



(this are not differentiable, so we use sub-gradient)

→ Thus,

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

→ For finding optimal solution, for closed form solution,

$$\frac{\partial \text{cost}(w_j)}{\partial w_j} = 0.$$

$$\therefore \begin{cases} -2\rho_j + 2w_j \cdot \sum_{i=1}^{N} [h_j(x_i)]^2 - \lambda = 0, & \text{if } w_j < 0 \\ [-2\rho_j - \lambda, \ -2\rho_j + \lambda], & \text{if } w_j = 0 \\ -2\rho_j + 2w_j \sum_{i=1}^{N} [h_j(x_i)]^2 + \lambda = 0, & \text{if } w_j > 0 \end{cases}$$

Case: I $\quad w_j < 0$

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

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

since, $w_j < 0$ and $\sum_{j=1}^{N} [h_j(x_i)]^2 > 0$

$\rho_j + \frac{\lambda}{2} < 0 \quad \therefore \rho_j < \left(-\frac{\lambda}{2}\right)$

Case: II $\quad w_j > 0$

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

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

since, $w_j > 0$ and

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

$\rho_j + \frac{\lambda}{2} > 0$

$\therefore \rho_j > -\frac{\lambda}{2}$

→ Now, for gradient descent, initialize w, while not converged.

Pick a coordinate, j.

→ If $w_j < 0$, then

$$\hat{w}_j \Longleftarrow \frac{\rho_j + \frac{\lambda}{2}}{\sum_{j=1}^{N} [h_j(x_i)]^2}$$

→ If $w_j > 0$, then

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