Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer 1

Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression.

Ridge Regression : In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients.

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

This is equivalent to saying minimizing the cost function in above equation under the condition as below,

$$\text{For some } c>0,\ \textstyle\sum_{j=0}^{p}w_j^2<c$$

So ridge regression puts constraint on the coefficients (w). The penalty term (lambda or alpha) regularizes the coefficients such that if the coefficients take large values the optimization function is penalized. So, ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity. For low value of α (0.01), when the coefficients are less restricted, the magnitudes of the coefficients are almost same as of linear regression. For higher value of α (100), the magnitudes are considerably less compared to linear regression case. This is an example of shrinking coefficient magnitude using Ridge regression.

Lasso Regression: The cost function for Lasso (least absolute shrinkage and selection operator) regression can be written as

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

$$\text{For some } t>0,\ \textstyle\sum_{j=0}^{p}|w_j|<t$$

Just like Ridge regression cost function, for lambda = 0, the equation above reduces to equation for ridge regression. The only difference is instead of taking the square of the coefficients, magnitudes are taken into account. This type of regularization (L1) can lead to zero coefficients i.e. some of the features

are completely neglected for the evaluation of output. So Lasso regression not only helps in reducing over-fitting but it can help us in feature selection.

When alpha value increases the model variance decrease but the bias increases. At optimal value of alpha the model bias and variance becomes balanced. That means the model becomes simple and robust.

If we double the value of alpha for ridge and Lasso regression the model will become simpler but the error term will be increased. The model variance will decrease but bias will increase.

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer 2

During the assignment I have determined the optimal value of Lambda for ridge and lasso regression. The optimal value of lambda for ridge was 100 and for lasso was 500. I choose the lambda for Lasso regression because in Lasso regression the coefficient of most of the variables became 0. So it worked as a feature selection.

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer 3

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer 4

Mostly simpler the model more robust and generalisable it is.

Simpler models are more robust—they are not as sensitive to the specifics of the training data set as theirmore complex counterparts are. Clearly we are learning a 'concept' using a model and not really the training data itself. So ideally the model must be immune to the specifics of the training data

provided and rather somehow pick out the essential characteristics of the phenomenon that is invariant across any training data set for the Model Selection problem. So it is generally better for a model to be not too sensitive to the specifics of the data set on which it has been trained. Complex models tend to change wildly with changes in the training data set. Again using the machine learning jargon simple models have low variance, high bias and complex models have low bias, high variance. Here 'variance' refers to the variance in the model and 'bias' is the deviation from the expected,ideal behaviour. This phenomenon is often referred to as the bias-variance tradeoff.

Simpler models are usually more 'generic' and are more widely applicable (are generalizable).

Simpler models make more errors in the training set — that's the price one pays for greater predictability. Complex models lead to overfitting — they work very well for the training samples, fail miserably when applied to other test samples.