Advanced Regression Assignment subjective questions.

Submitted by - Kunal Arneja

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?

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.

1. The cost function for both ridge and lasso regression are similar. However, ridge regression takes the square of the coefficients and lasso takes the magnitude.
2. Lasso regression can be used for automatic feature selection, as the geometry of its constrained region allows coefficient values to inert to zero.
3. An alpha value of zero in either ridge or lasso model will have results similar to the regression model.
4. The larger the alpha value, the more aggressive the penalisation.

In our notebook the optimal value for alpha is 0.4 for Lasso and 6 for Ridge.

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?

In our notebook the optimal value for alpha is 0.4 for Lasso and 6 for Ridge.

If we use 0.4 for Lasso model, all features had a coefficient 0, So to include only most relevant features we sent the alpha for Lasso to a very small 0.01.

For Ridge, we will apply alpha as 6.

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?

The top 5 features are -

1	MSSubClass	-0.009572
2	LotFrontage	0.012786
3	LotArea	0.018169
4	OverallQual	0.058088
5	OverallCond	0.043713

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?

Bias - Error in data should be kept low.
Variance - Overfit should be controlled while training the model.

To say the model is robust following conditions should be met -

1. R2 scores should be high
2. R2 scores for test should be close to that of train dataset
3. Mean square error should be low.
4. Mean square error fir test should be close to that of train dataset
5. Model should be as simple as possible.