Advanced Regression Assignment - Part -2

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

Optimal Value for Lasso regression as per the findings is 0.01 and for Ridge Regression is 2. When we plot the curve between negative mean absolute error and alpha we see that the value of alpha increases from 0, the error term decreases and the train error shows an increasing trend when value of alpha increases. When alpha=2 in case of ridge minimum test error is observed . In case of lasso when we increase the value of alpha the model tries to penalise more and makes most of the coefficient as zero, hence chose value as 0.01

If we double the value of alpha for lasso regression we are trying to penalise the model and making the coefficients equal to zero and the value of our r2 score will also decreases.
In case of ridge regression if we double the alpha, the model will apply more penalty on the curve and make it more generalised that makes model more simple and will try to fit every data of the data set.

The most important variables after changes implemented for ridge -
1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MsZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

The most important variables for lasso would be-
1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotFrontage

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

Lasso Regression would be better. It uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As lambda increases, lasso shrinks the coefficients towards zero. Lasso also does variable selection. When

lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

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

The top 5 most important predictor variables are-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

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

The model should be simple, robust and generalisable. It can be understood by using the Bias-Variance trade off. Simple models have more bias but less variance and more generalisable. It implies in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data.

Bias: Its error in model, when the model is weak to learn from the data. High Bias means model is unable to learn details in the data. Model performs poor on training and test data.

Variance: Its error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as its unseen for the model.

Its important to have balance in Bias and Variance to avoid overfitting and undercutting of data