Assignment Part-II

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 –

Ridge regression: - When we plot the curve between negative mean absolute error and alpha, we see that as the value of alpha increase from 0 the error term decreases and the train error is showing increasing trend when value of alpha increases.
Here I have taken the optimal value of alpha as 2, when the value of alpha is 2 the test error is minimum so we decided to go with value of alpha equal to 2 for our ridge regression.

Lasso regression: - I have taken very small value of alpha in lasso regression that is 0.01, when we increase the value of alpha the model tries to penalize more and try to make most of the coefficient value zero. Initially it came as 0.4 in negative mean absolute error and alpha.

When we double the value of alpha for our ridge regression then model will apply more penalty on the curve and try to make the model more generalized that is making model simpler and no thinking to fit every data of the data set. From the graph we can see that when alpha is doubled, we get more error for both test and train.
Similarly, when we increase the value of alpha for lasso regression, we try to penalize more our model and more coefficient of the variable will reduced to zero.

Also, when we increase the value of alpha then the value of r2 square decreases.

The most important variable after the changes has been implemented for ridge regression are: -

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 variable after the changes has been implemented for lasso regression are: -

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

Assignment Part-II

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 –

Lambda is a tuning parameter, it applies penalty to the coefficients those having greater values.

As we increase the value of lambda the variance in model decreases and bias remains constant. Ridge regression includes all variables in final model.

Lasso regression also uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. 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.

So, we need to be careful while choosing the tuning parameter i.e. Lambda. In this model I have chosen lambda (alpha) 2 for Ridge regression and lambda 0.01 for Lasso regression.

For Ridge regression when the value of alpha is 2 the test error is minimum so we decided to go with value of alpha equal to 2 for our ridge regression.

For Lasso regression when we increase the value of alpha the model tries to penalize more and try to make most of the coefficient value zero. Hence, I have chosen value of lambda 0.01

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 –

5 most important predictor variables in the original model that will be excluded are: -

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

5 most important predictor variables in the new model after excluding previous one: -

1. BsmtFinSF1
2. Fireplaces
3. LotArea
4. LotArea
5. LotFrontage

Assignment Part-II

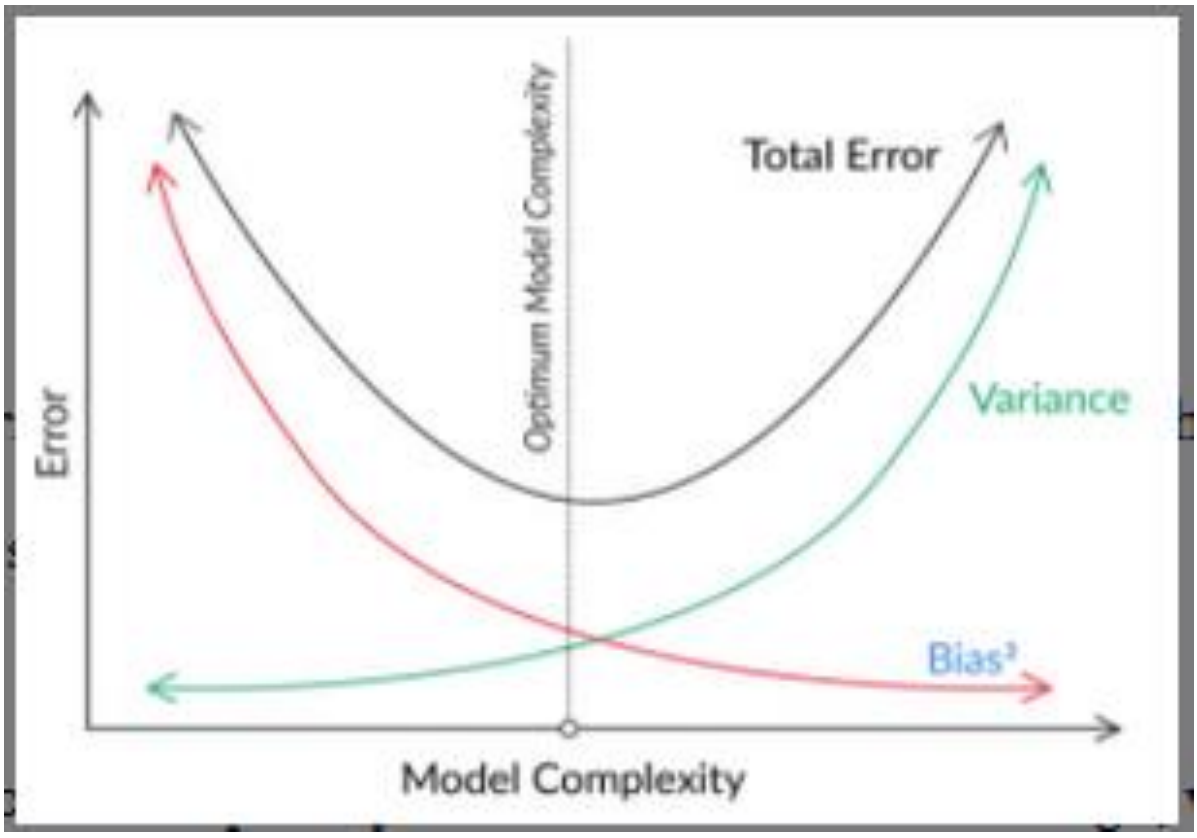
Question 4

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–

Variance: Variance is 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 it was unseen data for the model.

Bias: Bias is 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 testing data.



It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.