Problem Statement - Part II

Question1: What is the optimal value of alpha for ridge and lasso regression? What will be the changes in themodel 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:

In the case of 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 decrease and the train erroris showing increasing trend when value of alpha increases 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 I have decided to keep very small value that is 0.01, when we increase the value of alpha the model try 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 no we will take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set from the graph we can see that when alpha is 10 we get more error for both test and train.
Similarly, when we increase the value of alpha for lasso, we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases.

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

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 as follows: -

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

<u>Question 2: You have determined the optimal value of lambda for ridge and lasso regression during theassignment. Now, which one will you choose to apply and why?</u>

Answer:

It is important to regularize coefficients and improve the prediction accuracy also with thedecrease in variance, and making the model interpretably.

Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression, 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.

<u>Question3:</u>

<u>After building the model, you realized 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 thefive most important predictor variables. Which are the five most important predictor variables now?</u>

Answer:

Those 5 most important predictor variables that will be excluded are :-

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

Question 4:

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

Answer:

Per, Occam's Razor— given two models that show similar 'performance' in the finite training or testdata, we should pick the one that makes fewer on the test data due to following reasons:-

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complexones and hence are easier to train.
- Simpler models are more robust.
    - Complex models tend to change wildly with changes in the training data set
    - Simple models have low variance, high bias and complex models have low bias, highvariance
- Simpler models make more errors in the training set. Complex models lead to overfitting —they work very well for the training samples, fail miserably when applied to other test samples

Therefore to make the model more robust and generalizable, make the model simple but notsimpler which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolutevalues or the squares of the parameters of the model.

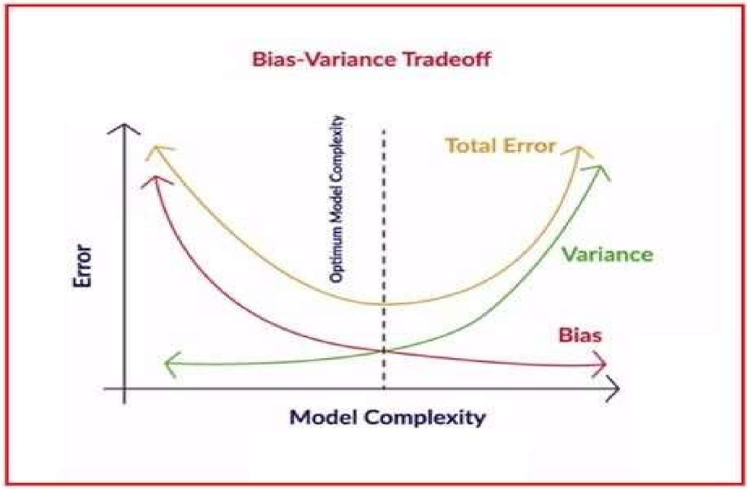
Also, Making a model simple leads to Bias-Variance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is veryunstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given isunlikely to change wildly even if more points are added or removed.

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a verylarge bias as its expected error across all test inputs are very high.

Variance refers to the degree of changes in the model itself with respect to changes in the trainingdata.

Thus, accuracy of the model can be maintained by keeping the balance between Bias and Variance asit minimizes the total error as shown in the below graph


Bias-Variance Tradeoff
Optimum Model Complexity
Total Error
Variance
Error
Bias
Model Complexity