Advanced Regression Assignment part-II

Submission by Ashwini Abhang

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?

Ans:

RIDGE REGRESSION - If you plot the curve between negative mean absolute errors You can see that as the alpha value increases from 0, the error term decreases and the training error decreases. It shows an increasing tendency as the value of alpha increases. If the value of alpha is 2, the test will fail Since it is the smallest, we chose an alpha value equal to 2 for ridge regression.

LASSO REGRESSION - In the Lasso regression, we chose to keep the minimum value of 0.01 when increasing the value. Alpha tries to impose more penalties on the model, making most of the coefficient values zero. Initially it was a negative mean absolute error and an alpha 0.4.

If you double the alpha value of ridge regression, the alpha value will be the same. The 10 model applies more penalties to the curve and tries to make the model more general To simplify the model and not think about adjusting all the data in the dataset. You can see from the figure If Alpha is 10, more bugs will occur in testing and training. Similarly, increasing the value of Lasso's Alpha will try to penalize the model and so on. If you also increase the squared value of r2, the coefficient of the variable will be zero sink.

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. 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?

Ans:

It is important to make the coefficients regular and improve the prediction accuracy. Reduce variance and make the model interpretable

Ridge regression uses an adjustment parameter called a lambda because the penalty is the square of the magnitude of Coefficients identified by cross-validation. Residual sum of squares or squares should be reduced using penalty. Because the penalty is lambda multiplied by the sum of the squares of the coefficients, The higher the value, the more penalized it will be. Increasing the value of lambda reduces the variance of the model, The preload remains constant. Ridge regression, unlike Lasso regression, contains all the variables of the final model.

Since the penalty is an absolute value of size, the lasso regression uses an adjustment parameter called a lambda. Of the coefficients identified by cross-validation. As the lambda value increases, the lasso shrinks The coefficients tend to be zero, making the variable exactly equal to zero. Lasso also performs variable selection. If the lambda value is small, perform a simple linear regression, and as the lambda value increases, A contraction occurs and variables with a value of 0 are ignored 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?

Ans:

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 generalisable? What are the implications of the same for the accuracy of the model and why?

Ans:

The model should be as simple as possible, although its accuracy will decrease but it will be more powerful, general, robust and generalizable. This can also be found using the Bias Variance trade-off. The easiest model is more biased but has less variance and is more generalizable. Its implication for accuracy is that a robust and generalizable model will work on both training and test data, i.e. accuracy doesn't change much for training and test data.

Bias: Bias is an error in the model when the model is weak in learning the data. High bias means that the model cannot learn the details of the data. The model performs poorly on training and test data.

Variance: Variance is an error in the model when the model tries to overlearn from the data. High variance means the model performs exceptionally well on training data as it has been very well trained on this data but performs very poorly on testing data as it was unseen data for the model.

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