

Problem Statement – Part II

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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans: As per my designed experiment model below are my response of above question:

Ridge Regression – We plot the curve between negative mean absolute error and alpha was seen that the value of alpha increase from 0 the error term decrease. As per the plot we are having trend seen that if increasing trend when the value of alpha increase. We have also seen that alpha is 2 the test error is minimum, that is why we decided to go with value of alpha equal to 2 for our Ridge Regression.

Lasso Regression – In Lasso Regression I have taken small value of 0.01 and observed that if we increased the value of alpha the model try to make most of the coefficient value zero. Where initially it came as 0.4 in negative mean absolute error and alpha.

Double the Alpha:

When the alpha value is doubled for our Ridge Regression, we are going to see the model will apply more penalty on the curve and try to make the model more generalized. We also observed 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. Also, the coefficient of the variable will reduce to zero. The r^2 square also got impacted when we increase the value of alpha.

As per my build experiment, below are my important variables after the changes has been implemented for Ridge Regression:

1. GrLivArea
2. SaleCondition_Partial
3. Neighborhood_StoneBr
4. MSZoning_FV
5. MSZoning_RL
6. Neighborhood_Crawfor
7. MSZoning_RM
8. MSZoning_RH
9. SaleCondition_Normal
10. Exterior1st_BrkFace

For Lasso Regression below are the most important variables:

1. GrLivArea
2. OverallQual

Problem Statement – Part II

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?

Ans:

As per my understanding and experiment, it is always important to regularize coefficient and improve the prediction accuracy. It is also required to decrease in variance and make the model more interpretable.

Ridge regression: It uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation.

Residual sum of 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 per my observation if we increase the value of lambda the variance in model is dropped and bias remains constant.

Ridge regression includes all variables in final model in compared to Lasso Regression.

Lasso regression: It 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.

As per my observation when lambda value is small it performs linear regression. It is also observed that if lambda value increases, variables with 0 value are getting neglected by the model and sheen shrinkage in place.

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:

After my experiment model finalization, I have identified those 5 most important predictor variables that will be excluded are given below:

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

Problem Statement – Part II

Assignment Part - II

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?

Ans:

As per my understanding and the experiment performed for building the model, it is observed that a model should be as simple as possible, even though its accuracy decreases but it will be more robust. The simpler the model, the more the bias but less variance and more generalized.

In terms of accuracy, a robust and generalized model will perform equally well on both training and test data.