Advanced Regression

Problem Statement 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:

The optimal value of Alpha for **Lasso Regression** is **0.01**.
The optimal value of Alpha for **Ridge Regression** is **2**.

On Doubling alpha:

In Lasso Regression, doubling the value of alpha penalizes the model, so more coefficients become 0, which increases the r-squared error.

In Ridge Regression, doubling the value of alpha will cause the model to apply a penalty to the curve and allow it to become more generalized and simple, producing more errors in test and train data.

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 to and why?

Answer:

Both of these techniques are regularization techniques that help determine the accuracy of the predicted value for target variables.

In the Lasso Regression, the penalty is the absolute value of the magnitude of coefficients which is identified by cross-validation. As the lambda value increases Lasso shrinks the coefficients towards 0. Hence, Lasso also helps in the feature selections.

In Ridge Regression method, it uses the hyperparameter called lambda as a penalty multiplied by the square of the magnitude of the coefficients which is identified as the cross-validation. The penalty is lambda times the sum of squares of the coefficients.

Hence the coefficients that have greater value get penalized. As we increase the value of lambda the variance of the model increases and bias remains constant. Ridge Regression includes all the features in the final model

Question 3:

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 the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

Now, the most important predictors would be:
`GrLiveArea, OverallQual, OverallCond, TotalBsmtSF, 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:

In accordance with Occum's principle, the model should be simpler. It will be more robust and generalizable for simple models, although accuracy will decrease. Also known as the Bias-Variance Tradeoff.

Bias and Variance

For example, let's consider a model that memorizes all training data.

The model also needs to be drastically changed if the data set is slightly changed. Therefore, the model is unstable and sensitive to changes in training data, which is called high variance. A model's variance is its output on some test data relative to its changes in training data. Here, variance refers to the degree of change in the model itself as a result of changes in training data.

Bias measures how accurate the model will be based on future (test) data. Extremely simple models are likely to fail in predicting complex real-world phenomena. Despite its advantages, simplicity has its disadvantages as well.

Since the expected total error of a model is the sum of both bias and variance errors, we want to reduce both bias and variance as much as possible.