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

When considering ridge regression, we observe that as the value of alpha increases from 0, the error term decreases and the training error exhibits an increasing trend. However, when alpha equals 2, the test error reaches its minimum. Therefore, we select an alpha value of 2 as the optimal choice for ridge regression.

For lasso regression, I have chosen a small value of 0.01. Increasing the value of alpha causes the model to penalize coefficients more aggressively, driving more coefficients towards zero. Initially, the negative mean absolute error is 0.4 when alpha is considered.

If we double the value of alpha for both ridge and lasso regression, the model will apply stronger penalties and strive for greater generalization, aiming to simplify the model and not overfit the data. From the graph, we can observe that when alpha is increased to 10, both the test and train errors increase.

After implementing these changes, the most important predictor variables 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 predictor variables for lasso regression, following the changes, are:

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

Answer

Regularizing coefficients is crucial for improving prediction accuracy, reducing variance, and enhancing model interpretability. In ridge regression, the tuning parameter lambda is utilized to apply a penalty that is proportional to the square of the coefficient's magnitude, determined through cross-validation. By imposing this penalty, the residual sum of squares is minimized. Consequently, coefficients with larger values are penalized more significantly. As lambda increases, the model's variance decreases while the bias remains constant. Unlike lasso regression, ridge regression incorporates all variables into the final model.

On the other hand, lasso regression employs lambda as the penalty, which corresponds to the absolute value of the coefficient's magnitude, identified through cross-validation. As lambda increases, lasso regression shrinks the coefficients towards zero, potentially resulting in some variables being exactly equal to zero. Lasso regression also facilitates variable selection. When lambda is small, lasso regression performs similarly to simple linear regression. However, as lambda increases, shrinkage occurs, leading the model to disregard variables with coefficients reduced to zero.

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

The following five predictor variables, considered to be the most important, will be excluded:

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?

Answer

Striving for simplicity in the model is crucial, even if it leads to a decrease in accuracy, as it enhances robustness and generalizability. This concept can be understood through the Bias-Variance trade-off. A simpler model tends to have higher bias but lower variance, resulting in improved generalizability. In terms of accuracy, a robust and generalizable model performs equally well on both training and test data, displaying minimal changes in accuracy between the two.

Bias refers to the error in the model when it is unable to effectively learn from the data, indicating a weak performance. High bias implies that the model struggles to capture the intricate details within the data, leading to poor performance on both training and testing data.

Variance, on the other hand, refers to the error in the model when it tends to overfit the data by excessively relying on the training data. High variance indicates that the model performs exceptionally well on the training data, which it has extensively learned from, but performs poorly on testing data, as it struggles with previously unseen data.

Maintaining a balance between bias and variance is crucial to avoid both overfitting and underfitting of the data, ensuring optimal model performance.