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:

In the context of ridge regression, our analysis involved plotting the relationship between the negative mean absolute error and various alpha values. We observed that starting from an alpha of 0, as the alpha value increases, there's a noticeable decrease in the error term. Conversely, the training error begins to show an upward trend with increasing alpha values. The minimum test error occurs when the alpha value is set to 2, leading us to select an alpha of 2 for our ridge regression model.

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:

Choosing between ridge and lasso regression hinges on our goals regarding regularization, prediction accuracy, variance reduction, and model interpretability. Both methods apply a tuning parameter, lambda, to penalize the magnitude of coefficients, but they do so in different ways, influencing their applicability to various scenarios.

Ridge regression applies a penalty equal to the square of the coefficient magnitudes. This approach ensures that while coefficients are shrunk towards zero, they are never exactly zero, thus retaining all predictors in the model. The primary advantage here is variance reduction; as lambda increases, model variance decreases, enhancing stability and generalizability at the cost of potentially higher bias. This makes ridge regression particularly suitable for scenarios where each predictor adds some value to the model interpretation or when multicollinearity is present among the predictors.

On the other hand, lasso regression penalizes the absolute value of the coefficient magnitudes, which can shrink some coefficients exactly to zero, thereby performing variable selection. This characteristic of lasso is especially valuable when we suspect that some predictors are irrelevant or when we prefer a more parsimonious model for easier interpretation. As lambda increases, lasso provides the added benefit of feature selection, simplifying the model by excluding non-contributory variables.

In deciding which method to apply, the choice boils down to the specific requirements of our analysis and prediction. If the goal is to maintain all variables for interpretation and deal with multicollinearity, ridge regression is preferable. Conversely, if the aim is to identify a subset of important predictors and potentially simplify the model for interpretation or when dealing with a high-dimensional dataset where many variables are suspected of being irrelevant, lasso regression may be the better choice.

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:

When rebuilding the lasso regression model without the five key predictors—GrLivArea, OverallQual, OverallCond, TotalBsmtSF, and GarageArea—the next set of important variables will be determined by fitting the model to the updated dataset. Generally, the new top predictors could include:

YearBuilt or YearRemodAdd: Reflecting the property's age and updates.
LotArea: Indicating the size of the property.
Neighborhood: Capturing location value.
BsmtQual or GarageQual: Assessing the quality of the basement and garage.
Fireplaces or FullBath: Valued amenities influencing property price.

These variables are chosen based on their potential to significantly impact property valuation, replacing the influence of the initially excluded features.

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:

To ensure a model is robust and generalizable, aim for simplicity to balance the trade-off between bias and variance. A simpler model, while possibly less accurate on training data, tends to be more reliable across both seen and unseen data. High bias results in underfitting, missing key trends in data, whereas high variance leads to overfitting, capturing noise instead of underlying patterns. The goal is to achieve a model that performs consistently well on both training and testing datasets, indicating it has learned the underlying structure of the data without being swayed by noise. This balance is crucial for maintaining accuracy and reliability in predictions on new data.