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 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 in Ridge and Lasso regression depends on the data and the problem being solved. In general, the value of alpha is chosen through cross-validation, which involves selecting a range of possible values for alpha and evaluating the model performance for each value. The value of alpha that results in the best performance on the validation set is chosen as the final value.

If you double the value of alpha for both Ridge and Lasso, the magnitude of the regularization term in the loss function will increase. This will lead to shrinkage of the coefficients towards zero, and potentially result in some coefficients becoming zero. In Ridge regression, all coefficients will shrink, but none will be set to zero. In Lasso regression, some coefficients will be set to zero, effectively removing their contribution to the model.

The most important predictor variables after the change in alpha will depend on the data, but it is likely that the variables with the largest magnitude coefficients will continue to be the most important, as long as they are not set to zero by Lasso. You can check the magnitude of the coefficients after the change in alpha to see which variables are most important.

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

The choice between Ridge and Lasso regression depends on the specific problem and the characteristics of the data.

Ridge regression is a good choice when the goal is to shrink the coefficients towards zero, but not to set any coefficients to exactly zero. This makes it a good choice when there is multicollinearity in the data, meaning that the predictors are highly correlated with each other. Ridge regression helps to address multicollinearity by shrinking the coefficients towards zero, rather than removing any variables entirely.

Lasso regression is a good choice when the goal is to set some coefficients to exactly zero, effectively removing their contribution to the model. This can be useful when there are many

predictors, but only a small number of them are truly relevant. Lasso regression will set the coefficients for the unimportant predictors to zero, making it easier to interpret the model and reducing overfitting.

Ultimately, the choice between Ridge and Lasso regression will depend on the specifics of the data and the problem being solved. It may be necessary to try both and compare their performance through cross-validation to determine the best model for a given data set.

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

If the five most important predictor variables are not available in the incoming data, then their impact on the model cannot be accurately captured. In this case, the five most important predictor variables would change, and the new important predictor variables would depend on the remaining variables in the data set.

To determine the new five most important predictor variables, you could build a new model using the remaining variables and evaluate their importance. This can be done through techniques such as feature selection, which involves ranking the variables based on a criterion such as their coefficient magnitude in a linear regression model, or their importance in a decision tree model.

Once the new five most important predictor variables have been identified, you can use:

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 that a model is robust and generalizable, it is important to consider several key factors:

Data Quality: The quality of the data used to train the model can have a significant impact on its robustness and generalizability. It is important to check for missing values, outliers, and other anomalies that could skew the model's predictions.

Model Complexity: Overfitting can occur when the model is too complex and fits the training data too closely. To prevent overfitting, it is important to balance model complexity and model fit. Cross-validation can be used to evaluate the performance of the model on unseen data and avoid overfitting.

Regularization: Regularization techniques such as Ridge or Lasso can help to prevent overfitting and improve the generalizability of the model by shrinking the coefficients towards zero.

Data Representativeness: The model's performance can be impacted if the training data is not representative of the data it will encounter in production. To ensure that the model is generalizable, it is important to use a diverse and representative sample of the data.

The implications of these factors for the accuracy of the model are significant. If the model is not robust and generalizable, its predictions will be less accurate and unreliable when applied to new, unseen data. On the other hand, if the model is robust and generalizable, it is more likely to produce accurate predictions even when applied to new data, which increases its usefulness in real-world applications.