Machine Learning 2 Assignment

Problem Part 2

Chris Flood

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 can depend on the goals of the analysis as well as the type of data in the problem. We would use grid search or randomized search (cross-validation techniques) as part of the process. Metrics such as R-squared and mean squared error are used to evaluate which value of alpha (tuning parameter) would lead to the most optimal performance from the model. If we double the value of alpha for Ridge and Lasso regression, we see that the model becomes more conservative. Doubling the alpha value will lead to more of the coefficients going to zero which results in a model which is less sparse.

After alpha is doubled, we need to check model performance again. The most important predictors will be those which have the greatest influence over the prediction. These will be the coefficients with the highest values.

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:

Lasso regression is best when there are a high number of predictor variables present and when many of those predictors are not relevant. Lasso will reduce those coefficients to zero. Thereby the most important variables are identified.

Ridge regression is best when dealing with multicollinearity. The variables are reduced to almost zero without being eliminated completely. If there is importance associated with the variables we may need to retain them and not shrink them to completely zero.

I would choose Ridge to reduce multicollinearity otherwise and I would choose Lasso to identify the most important variables.

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:

In this case it would be necessary to retrain the Lasso model as the original variables have changed in terms of the available data. The most important predictor variables will then become the ones identified by Lasso having being retrained on the new data.

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:

1. Use good data that is varied when training and testing the models.
2. Consider good evaluation metrics to evaluate each model.
3. Test the model on data which has not been used in the analysis to determine if the model is robust and does not lead to increasing error.
4. Use cross-validation and test this on different data sets.