Problem Set 9

Trevor Jones

April 11, 2023

1 Questions

1. How many more X variables do you have than in the original housing data? (Q7)

   (a) Using the dim() function, I found that the new housing-train data has the dimensions of 404 by 14, meaning that there are 404 observations of 14 variables. Housing-train-prepped has the dimensions of 404 by 75, meaning that there are 404 observations for 75 variables.

2. In the LASSO model: What is the optimal value of lambda? What is the in-sample RMSE? What is the out-of-sample RMSE (i.e. the RMSE in the test data)?

   (a) The optimal value of lambda here is 0.00139. The in-sample RMSE is 0.170, the out of sample RMSE is 0.0632.

3. In the Ridge model: What is the optimal value of lambda? What is the in-sample RMSE? What is the out-of-sample RMSE (i.e. the RMSE in the test data)?

   (a) The optimal value of lambda here is 0.00569. The in-sample RMSE is 0.175, the out-of-sample RMSE is 0.0712.

4. Would you be able to estimate a simple linear regression model on a data set that had more columns than rows? Using the RMSE values of each of the tuned models in the previous two questions, comment on where your model stands in terms of the bias-variance trade- off.

   (a) You can, but it is certainly not a good idea as it will lead to overfitting – meaning that the model will perform well with training data but poorly when exposed to new data. Both models estimated in this problem set have higher in-sample RMSE than out-of-sample RMSE, suggesting that the models have low variance but could potentially have higher bias.