Problem Set Nine

Devin Williams

April 8, 2025

1 Dimension of Training Data and amount of X Variables: Q7

- What is the dimension of your training data (housing_train)?
  - The dimensions of the training data were 404 to 14.
- How many more X variables do you have than in the original housing data?
  - There are 61 more x variables than the original housing data.

2 Lasso: Q8

- What is the optimal value of Lambda?
  - The optimal Lambda value is 0.00139 based on the code.
- What is the in-sample RMSE?
  - The in-sample RMSE was estimated from the cross-validation results, which is 0.0632 as shown in the "mean" column. This is the average RMSE across the 6-fold cross-validation on the training data.
- What is the out-of-sample RMSE (i.e. the RMSE in the test data)?
  - It shows an RMSE of 0.170 for "Preprocessor1_Model1". This is the RMSE on your test data (out-of-sample).

3 Ridge: Q9

- What is the optimal value of Lambda now?
  - The ridge regression model is 0.0233, as shown in the "penalty" column of the ridge_top_rmse output.

- What is the out-of-sample RMSE?
  - The out-of-sample RMSE for the ridge regression model is 0.173, as shown in the first table.

4 Question 10

- Would you be able to estimate a simple linear regression model on a data set that had more columns than rows?
  - You cannot estimate a simple linear regression model on a dataset that has more predictors (columns) than observations (rows). This creates a under determined system where there are infinite solutions that perfectly fit the training data.
- Where the model stands in terms of the bias-variance trade-off.
  - Both models show significant gaps between in-sample and out-of-sample performance. This indicates some overfitting. Overall, the LASSO model performs better in both in- and out-of-sample fits. The optimal regularization parameters (Lambda = 0.00139 for LASSO and Lambda = 0.0233 for Ridge) show that Ridge requires stronger regularization than LASSO to achieve its best performance.