PS9_Zilles

Andrew Zilles

April 2024

7. What is the dimension of your training data (`housing_train`)? How many more X variables do you have than in the original housing data?

The dimensions in the new dataset `housing_train` has 404 observations (80% of the original) and 14 variables. After applying the transformations, there are now 74 variables in `housing_train_prepped`. I'm wondering if the code provided really wants to interact crim with zn, indus, rm... all together or if that's supposed to be broken out to crim:zn, b:dis, rad:ptratio, etc?

8. What is the optimal value of λ? What is the in-sample RMSE? What is the out-of-sample RMSE (i.e. the RMSE in the test data)?

The optimal value of λ is 5.179 e-05 In-sample RMSE is 0.413 Out-of-sample RMSE is 0.390

9. What is the optimal value of λ now? What is the out-of-sample RMSE (i.e. the RMSE in the test data)?

Optimal value of λ is now 1 e-10 Out-of-sample RMSE is 0.390

10. Would you be able to estimate a simple linear regression model on a data set that had more columns than rows? ...comment on where your model stands in terms of the bias-variance tradeoff.

No, we can't perform an estimate of a simple linear regression model on data that had more columns than rows because OLS requires more observations than variables. Having more variables than observations would make the model overfit. Lasso In-sample vs Out-of-sample RMSE is 0.413, 0.390 ridge In-sample vs Out-of-sample RMSE is 0.413, 0.390 Since these are pretty close then we can say we're balancing bias and variance pretty well.