Econ-5253

Problem Set 9

Mieon Seong

Due by Apr 9th, 2024

Question 7 : Create a new recipe() that takes the log of the housing value, converts chas to a factor, creates 6th degree polynomials of each of the continuous features, and linear interactions of each. What is the dimension of your training data? How many more X variables do you have than in the original housing data?

The dimension of the training data (housingtrain) is 74 variables.

Question 8 : Following the example from the lecture notes, estimate a LASSO model to predict log median house calue, where the penalty parameter λ is tuned by $6-$fold cross validation. What is the optimal value of λ? What is the in-sample RMSE? What is the out-of-sample RMSE?

The optimal value of λ is 0.00356 .

The in-sample RMSE is 0.172.

The out-of-sample RMSE is 0.806.

Question 9 : Repeat the previous question, but now estimate a ridge regression model where again the penalty parameter λ is tuned by $6-$fold CV. What is the optimal value of λ now? What is the out-of-sample RMSE?

The optimal value of λ is 0.0233. The in-sample RMSE is 0.803.

The out-of-sample RMSE is 0.173.

Question 10 : 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.

Estimating a simple linear regression model on a dataset with more columns than rows can be challenging due to the potential for overfitting and multicollinearity.