PS9 DScourse 2024

Emilien Akotenou

April 9, 2024

Problem Set 9 Solution

Question 7

The dimension of the training data is 404 rows and 14 columns, after preprocessing is 404 rows and 75 columns. There are 61 more X variables compared to the original housing data, which had 14 columns.

Question 8

The optimal value of λ for the LASSO model is `0.0373`. The in-sample RMSE is `0.413`, and the out-of-sample RMSE is `0.39`.

Question 9

The optimal value of λ for the ridge regression model is `0.0373`. The in-sample RMSE is `0.140` and the out-of-sample RMSE is `0.181`.

Question 10

It would not be possible to estimate a simple linear regression model on a data set that had more columns than rows. In such a case, the matrix of predictors would not be of full rank, and the ordinary least squares solution would not be unique. Regularization techniques like LASSO and ridge regression allow us to estimate models in high-dimensional settings where the number of predictors exceeds the number of observations.

Based on the RMSE values of the tuned LASSO and ridge regression models, we can assess where the models stand in terms of the bias-variance trade-off. The LASSO model, which performs variable selection by setting some coefficients to exactly zero, tends to have lower variance but potentially higher bias. On the other hand, the ridge regression model, which shrinks the coefficients towards zero but does not perform variable selection, tends to have lower bias but potentially higher variance. The optimal choice between LASSO and ridge regression depends on the specific characteristics of the data and the relative importance of bias and variance in the prediction task.