

Problem Set 9

Ana Gallart

April 11 2023

- Question 7

How many more X variables do you have than in the original housing data?

`dim(housing)` – 506 by 14

`dim(housing train)` – 404 by 14

`dim(housing train x)` – 404 by 74, 60 more x variables

`dim(housing train y)` – 404 by 1

- Question 8

What is the optimal value of λ ? 0.000139 or 0.00910

What is the in-sample (train data) RMSE? 0.170

What is the out-of-sample (test data) RMSE? 0.180

- Question 9

What is the optimal value of λ ? 0.0000000001

What is the in-sample (train data) RMSE? 0.173

What is the out-of-sample (test data) RMSE? 0.173

Not sure why the in-sample and out-of-sample came out the same... probably an error somewhere. I tried to find it but couldn't.

- 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 tradeoff.

We could fit a SLR model on dataset with more columns than rows, but it's overfit the model to have more features than observations. I think the model is in good standing for the bias-variance tradeoff. Since the RMSE is relatively low for the models (less than 0.2) it doesn't have high bias, but since it isn't near zero it doesn't have no bias and thus high variance.