MSDS 422: Practical Machine Learning

Professor Anil D. Chaturvedi

Assignment 3: Evaluating Linear Models

Tiffany Duong

Summary

Supervised machine learning methods such as linear regressions are a great tool when it comes to understand the relationships among the predictive variable and the several explanatory variables. Linear models were deployed to help understand the variables that can help explain the median value of homes (in thousands of dollars) in the Boston metropolitan area during the 1970's.

Research Design

To understand the variables, the data must be explored first. The predictive variable is *mv* (median value) and there were 12 features that can be used to help predict *mv* within the data of 506 observations. From initial data exploration, it was found that the number of rooms, the percentage of land zoned for lots, the weighted distance to employment centers, and whether or not the house is located on the Charles River have positive correlations to the median housing value, while the percentage of the population being of a lower socioeconomic status had a high value of negative correlation.

Technical Overview

Analysis on the survey data was performed using traditional statistical methods and Python, a high-level general program language. Several Python packages were used, including NumPy, pandas, matplotlib, and seaborn, and scikit-learn. Scikit-learn was heavily used when setting up the classification models as well as performing the cross validation to see model performance. The predictive variable we will be looking at is the *mv* variable – the median housing value (in thousands) during the 1970's. There were 12 features from the data set that will be used as explanatory variables for the linear models. The most basic linear model OLS will be used as well as regularized models (ridge regression, lasso regression, and ElasticNet). Regularized

models are a good way to prevent and reduce overfitting of the model. In addition, because *mv* has a substantially skewed distribution, we will be experimenting with using a log transformation for that variable, alongside with the untransformed variable. We will be comparing to see which linear model performs better and to conduct the comparison, we will evaluate the models using k-fold (5 folds in this case) cross validation, using root mean square error (RMSE) as an index of model performance.

Results

Comparing the log-transformed and the untransformed *mv* variable within the models, the RMSE had higher values for the untransformed *mv*. While OLS, ridge regression, lasso regression, and ElasticNet had similar average RMSE results from the 5-fold cross validation, lasso regression had a slightly higher average RMSE value of 0.655 – this could be attributed to lasso regression having the tendency of eliminating the weights of the least important features (explanatory variables), thus resulting in a better fit.

When advising to a real estate brokerage firm on which machine learning method to employ, I would highly suggest employing a lasso regression based on the current housing data available as it had the highest RMSE value of 0.655 compared to the other models. In addition to having a higher RMSE value, lasso regression can also help to identify features that impacts the median housing value the most as well as the least important features.