Data Analytics and Machine Learning | Prof. Lochstoer

Problem Set 5

Question 1:

Simulate 1500 realizations of two uncorrelated standard Normal variables. Call the simulated variables x_1 and x_2 and use these simulated variables as your predictors for y. Simulate 1500 outcomes for y for each of the two models:

a) $y = 1.5x_1 - 2x_2 + \varepsilon$

b) $y = \begin{cases} 1.5x_1 - 2x_2 + \varepsilon, & if\ x_1 < 0 \\ 1.5 \ln x_1 + \varepsilon & if\ x_1 \geq 0 \end{cases}$'

where ε is a Standard Normal uncorrelated with x_1 and x_2. Use the first 1000 observations of x_1, x_2, and y as your training sample and observations 1001-1500 as your test sample. Repeat the simulation exercise above 500 times and plot a histogram of the out-of-sample mean-squared errors for the following methods for each of model a) and b):

(i) OLS regression
(ii) Random Forest with ntree=250 and maxnodes=10
(iii) XGBoost with eta=0.3, gamma=0, and max_depth=6; use 20 rounds and 10 folds for the cross-validation procedure. Make sure that the output of the cross-validation procedure does not appear in your final write-up.

Note that you can use an in-sample cross-validation procedure to determine the optimal values for the decision tree parameters. However, you are not required to do so for this exercise.

Interpret the histograms. Which of the models (i), (ii), and (iii) do best in the out-of-sample exercise for models (a) and (b)? Do the histograms conform to your expectations given the data generating processes in parts (a) and (b)?

Question 2:

Attached to this problem set is a dataset which deals with Boston real estate prices. The dataset was obtained from the UCI Machine Learning Depository: https://archive.ics.uci.edu/ml/index.php.

Our goal in this exercise is to predict house prices in Boston (medv) given 11 explanatory variables (columns 1 through 11). Use the first 400 observations as your training sample and observations 401-506 as your test sample.

(a) Use random forest with ntree=500 and maxnodes=10.
Once you run the random forest, use R's predict() function to obtain predicted values for the test sample. What is the MSE of the prediction? Compare this to the benchmark MSE generated by a model that has as its predicted house value the mean house value in the test sample. As in the class notes, also report the Pseudo-R2 implied by these MSEs.

(b) Repeat the same exercise as above using XGBoost with eta=0.1, gamma=0, max_depth=6. Use 10 folds and 200 rounds for the cross-validation procedure. Make sure that the output of the cross-validation procedure does not appear in your final write-up.

(c) Repeat the exercise in part (a) using elastic net with alpha=0.5. Use a cross-validation procedure to find an optimal lambda. For that exercise, split the training sample into quarters (i.e., the 4-fold cross-validation).
Comment on the performance of the linear model relative to decision trees. In particular, get the MSE for the test sample and compute the Pseudo-R2 relative to the benchmark MSE from a).

(d) Repeat the exercise in part (a) but use log transformations of the following variables: indus, rm, rad, pt, and lstat. Drop the original variables from your model. Comment on the performance of this version of the linear model relative to decision trees in this case.