CHL5212H Assignment 3: Trees and Regularized Regression

The goal of this assignment is to practice fitting regularized regression and regression tree models. Split the data into a training, validation and test set (proportions 1/3, 1/6, 1/2) as in A2.

Variable Selection: Start with a list of all predictors and generate three lists using the training set. Use lasso to identify main effects and retain variable with non-zero coefficient estimates. Next, extend the main effects list by screening for pairwise interactions among the retained main effects. For the third list, fit a full regression tree (CP=0.01) on 80 instances of bootstrapped data and generate a list of variable ranks, averaging over the iterations, retaining variables with importance > 2.

Trees and Regularized Regression: Use the sets of variables (all variables, lasso main effects, lasso main effects with first-order interactions and tree list) and fit an elastic net with alpha = 1 (lasso), 0.5 (mix) and 0 (ridge) for each list, using 10 fold CV. Also fit a full regression tree and estimate a pruned regression tree using the smallest internal CV error for each list.

As in the A2, repeat the process to select the best model using the validation set performance. Retrain the models on the combined training and validation set and get the errors for all models on the test set. Compare the performance of your selected model with the others, as well as the models that were fit in A2. Get the variable ranks for each model in A3 and compare these to the selected model.

Regularization and Robustness: Train a lasso, ridge and forward step-AIC model on the first 50, 100, 200 and 400 observations of the combined training and validation set using the tree variable list obtained using bootstrap. Use 5 fold CV for the regularized models and k=2 for AIC selection. Plot the residual error on the test set for each method vs. sample size.

Interpretation:
1. Was there any advantage to using lasso for variable selection over univariate screening? How are the top variables influenced by the change of method?

2. Was there a lot of discrepancy between the lasso and bootstrapped tree list? What signal structure is likely to produce a highly differentiated list? A very similar list?

3. Based on subsequent model fits, did bootstrapping trees produce in effective list? Why was the bootstrapped list substantially longer than just a full/pruned tree?

4. How did the value of alpha affect the results of the elastic net? Was the performance of the ridge model expected, given the distributions of univariate RE and the results of models fit in A2?

5. Was there a lot of difference between full and pruned trees? What type of signal/covariate structure would result in minimal tree pruning?

6. How did performance compare between AIC, ridge and lasso as sample sizes increased? How robust was each method in terms of performance given sample size?

Evaluation:
Results 40%, interpretation 40%, coding style 20%.

Due: June 9, 2020 at 12:00pm