Math 189 Homework 8

Due March 10th, 2023

Problem

In this problem, we revisit the Boston Housing Price dataset (Boston.csv). The description of the dataset can be found in the lecture slides. The response variable is median house price (medv). Analyze this dataset using tree-based methods through the following steps.

1. Randomly divide the dataset into a training set and a test set with equal or nearly equal sizes.
2. Fit the training set with a single regression tree using all predictors. Use CV to select a subtree size and apply the tree prune to obtain a subtree. Plot, respectively, the large tree and the selected subtree. Use the test set to calculate prediction MSEs for both the large tree and selected subtree.
3. Fit the training set with bagging and random forests. For both methods, use 100 trees. For random forests, set m=4. Again, use the test set to calculate prediction MSEs for bagging and random forests. Compare these results with those obtained for one single regression tree.

For students working with Python:

See the following link for an example of random forests in Python:

https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/