Homework #2

MTH 9899 Baruch College
DATA SCIENCE II: Machine Learning

Due: May 1, 2019 - 18:00

Notes

- Code for this **MUST** be written in **Python 3.x**.
- Do NOT use 3^{rd} Party Packages for the regression functions unless specified.
- **The Due Date is Friday night, not at the beginning of class. Note however that more homework will be assigned the upcoming week so it is best to start early.**

Problem 1 Implement a simple regression tree. We will use point estimates in the leaves and use the CART Variance Reduction measure for a splitting criteria.

$$VR(S) = \text{var}\, S - \sum_{i=0}^{1} \frac{|S_i|}{|S|} \text{var}\, S_i$$

Use the attached code as your starting point

- For simplicity's sake, divide each attribute up into 5 equal sized bins, and test each end point of a bin as a potential split point. Test your algorithm on a 50000 row dataset generated using the attached generate_test_data function. Test against different max_depths and report a graph of depth vs R^2. Now, on the same graph, plot R^2 where you are using a new dataset, generated independently of the one used to train the tree. Does it look different? Why?
- One way to improve this is to build a random forest of these trees. Generate 100 of these trees, each with a different bootstrapped sample of X. At each split point, select $\frac{1}{3}$ of the features randomly as potential split variables. For values of $T \in \{1, 2, 5, 10, 20, 30, 50, 75, 100\}$, select multiple random boostrapped samples of these 100 trees, and for each forest of size T, predict on a test data set and calculate the mean and median R^2 for the forests. Plot these as a function of T and calculate the average pairwise correlation of the predictions from each tree.
- Make the number of features to consider splitting on at each split point a parameter instead of hard-coded to $\frac{1}{3}$. Generate more trees with this set to 1 and calculate the pairwise correltion of these trees. Is it higher or lower than when we used $\frac{1}{3}$?