Homework #2

MTH 9899 Baruch College
DATA SCIENCE II: Machine Learning

Due: May 4, 2018 - 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.

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

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 saame 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 potentially improve this and avoid overfitting would be to use cross-validation when calculating variance reduction. Modify your tree to have the constructor to take a *num_cv_folds* parameter. If this value is > 1, then calculate the variance reduction on a cross-validated dataset instead, ie for each candidate split point you are considering, you should do a CV measurement of the variance reduction. ONLY SPLIT if the CV Variance Reduction is positive. Repeat the experiments from the first part of the question and discuss any differences.