Homework #1

MTH 9899 Baruch College
DATA SCIENCE II: Machine Learning

Due: April 30, 2017 - 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 Sunday night, not at the beginning of class**

Problem 1 In class, we spoke about the equation below - a description of how the variance of the average of identically distributed variables decreases as we add more predictors to an ensemble. Please derive this formula - you can ignore the case of negative correlations. ρ represents the correlation between any two predictors and B represents the number of predictors. (This is Exercise 15.1 from the ESL text).

$$\sigma^2(\bar{\hat{y}}) \;=\; \rho\sigma^2_{\hat{y}_i} + \frac{1-\rho}{B}\sigma^2_{\hat{y}_i}$$

Problem 2 Implement a simple regression tree. We will use point estimates in the leaves and use the CART Variance Reduction measure for a splitting criteria.

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

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.