Yale University
Department of Statistics

STATISTICS 365: MACHINE LEARNING

Issued: 04/14/2015 **Due:** 04/24/2015

Boosting, Model Selection, and Clustering

Notes:
Some coding will be required. When coding please provide printouts of the plots that we request. Please also provide a printout of your code. Use whatever language is most comfortable for you, but for those of you not familiar with Python, R, Matlab, or Octave (the free version of Matlab), it might be worth it to learn. They hae a lot of impressive toolboxes and libraries. Datasets can be found on classesv2.

Problem 1 (Boosting and Overfitting):

- Implement the adaboost algorithm that we discussed in class. Use decision trees of depth 3 as your "weak" learners. You can either implement your own decision tree based on the class or you can use one that exists. However, please implement the adaboost part.
- Given the dataset labeled hw3prob1train.data where the last column are the $y \in \{-1, 1\}$ labels, run the boosting algorithm and plot the test error (data can be found in hw3prob1test.data) as a function of the number of iterations.
- Given the same dataset, now fixed the total number of iterations. Plot the test error versus the depth of the trees in the weak learners.

Problem 2 (Boosting for regression: In class we saw that we can do boosting for classification and that there is a coordinate descent interpretation. Now we will create an algorithm for doing boosted regression using decision trees as the weak learners.

- Write down an algorithm similar to adaboost where at each step we add a new decision tree of a fixed depth at every iteration. Justify your algorithm and explain your steps. (Hint: at each iteration you should be learning a decision tree that tries to approximate the residual error $y_i - f_{t-1}(x_i)$.)
- Given the dataset hw3prob3.data where the last column data are the real $y \in \mathbb{R}$ value, plot the test error as the number of iterations increases using the test dataset hw3prob3test.data.

Problem 3 (Model Selection): In the dataset hw3prob3.data the columns represent real values $y_i = f(x_i) + w_i$. The function takes a simple form as a linear combination of x, x^2, x^3, $cos(\omega x)$ where $\omega \in \{2\pi i/1000\}$ where $i \in [0, 1000]$, and $\exp(-x^2/(2\sigma^2))$ where $\sigma \in \{i/100\}$ for $i \in [0, 1000]$. The function f is only a small linear combination of the above function (at most 10 of them). Find those functions that best approximate f. List the function as well as their coefficients in the linear combination.

Problem 4 (Project Proposal): Write down a brief paragraph stating what topics your projects aims to explore.