Project Proposal
Monday, May 8, 2017
Brennan Kuo
Thomas Redding

Date Preferences

We'd prefer to present on Friday, May 26th.

Overview

Our basic plan is to code the C4.5 algorithm and use this to build trees. These trees will be aggregated with Adaboost - often called "the best out-of-the-box classifier." Because Adaboost requires *weak* learners, we will artificially limit the number of nodes that the C4.5 algorithm creates in each tree. Then, we will run this C4.5-Adaboost hybrid on the a dataset of 300 people's demographics and tea consumption habits/preferences to try and predict whether people prefer drinking Earl Grey, Black, or Green tea. We will use Java. If we have time, we can consider expanding into a softer Adaboost.

C4.5 Algorithm

The C4.5 algorithm works on multivariate data by choosing the best variable to separate the data along (which variable is the best predictor). Then, once the data is bifurcated, each branch is divided in a similar manner, using a variable that has not been used. Once we have included all of the variables that we care about, then we halt, and have our decision tree.

1. http://www.uh.edu/~smiertsc/4397cis/C4.5_Decision_Tree_Algorithm.pdf
2. http://research.ijcaonline.org/volume82/number16/pxc3892444.pdf
3. http://www.ijetae.com/files/Volume3Issue3/IJETAE_0313_57.pdf

Adaboost

At a high level, adaboost does something like this:

1. Initialize a set of weights at 1.
2. Train a weak C4.5 decision tree on the weighted data.
3. Increase the weights of points that the new tree misclassifies; reduce the weights of points the new tree correctly classifies
4. Return to step 2.

To classify a point, the many decision trees perform a weighted vote, where weights are determined by the weighted proportion of points each tree correctly classifies. Here are some references:

1. Wikipedia
2. Widely cited academic paper
3. MIT Powerpoint
4. Czech Technical University Powerpoint