CS 434, Final Project Proposal

Cody Malick & Garrett Smith

May 7, 2016

Our project is going the comparison of different classification algorithms to the problem of spam filtering. Spamfiltering is a fairly common problem that has had many different solutions proposed for it. We would like to go through some of the algorithms that we learned this term, and see how they compare in performance. Performance would be measured by error, speed of classification over increasing data set sizes. We want to compare linear regression, decision tree, and logistic regression.

Our data will be pulled from the online data store from University of California, Irvine. They have a large set of spam from the early 2000s we plan to use. It contains roughly four-thousand emails, of which about fourty percent are spam. This will be a great data set to work with, as we have the actual number of spam emails in the data set already, so we can use that to measure the accuracy and error of the different algorithms we are going to use. Having this data, will also allow us to compare the performance along with the change in performance when change the weights of different features in the data set.

There is a possibility of overfitting with each of these algorithms. Decision tree, for example, has a high probibility for over fitting if we do not set a limit on the depth of the tree. Linear regression is harder to overfit, but it is still a real possibility. To help prevent over fitting, we will test the algorithms using test and training sets, and using cross-validation to ensure that the general model is not overfit.

Performance will be measured by the accuracy, error, and actual runtime. We will be graphing each of these as a function of data set size. We hope to see interesting and different results by comparing these three algorithms.