CS534 Implementation Assignment 3: Bagging and AdaBoost

Amit Bawaskar, Michael Lam
EECS, Oregon State University

May 26, 2013

Abstract

In this assignment we implemented and evaluated bagging and AdaBoost using the decision stump as the base learner.

1 Introduction

Ensemble methods such as bagging and AdaBoost work by taking a base learner and generating a set (ensemble) of hypotheses by varying the training set. The final hypothesis for classification is a majority vote of these hypotheses. The decision stump is a suitable base learner for bagging and especially AdaBoost because it is a weak learner, meaning it classifies slightly better than random.

In this assignment we implemented the decision stump, bagging and AdaBoost. For each ensemble method we evaluated the training and test errors as a function of the ensemble size.

2 Decision Stump

The decision stump is a one level decision tree. In this assignment all features and labels are binary. Therefore the decision stump is a one level binary tree with a binary feature test at the root node connected to two leaf nodes each containing a label.

Learning is done by going through every feature in the training data and computing the maximum information gain (minimum entropy). Once the feature with the maximum information gain is determined, the two leaf nodes are given the label of the majority class from splitting the training data on that feature. Inference is done by simply testing the feature and assigning the label following the appropriate branch.

For AdaBoost, decision stump learning also accepts a distribution of the data as input. Therefore each training example is weighted differently based on the distribution, which affects the information gain computation and leaf node labels in every iteration of AdaBoost.

3 Bagging

Explain bagging and how we implemented

4 AdaBoost

Explain bagging and how we implemented

5 Results

For AdaBoost, figure 1 plots the training errors versus the ensemble size, and figure 2 plots the test errors versus the ensemble size.

6 Discussion

For AdaBoost, the training error appears to decrease as the ensemble size increases but then flattens out at ensemble size 15 and greater for the given training data. The test error decreases and performs the best at around ensemble size 20, but then increases afterward; this trend is typical of test curves in machine learning.

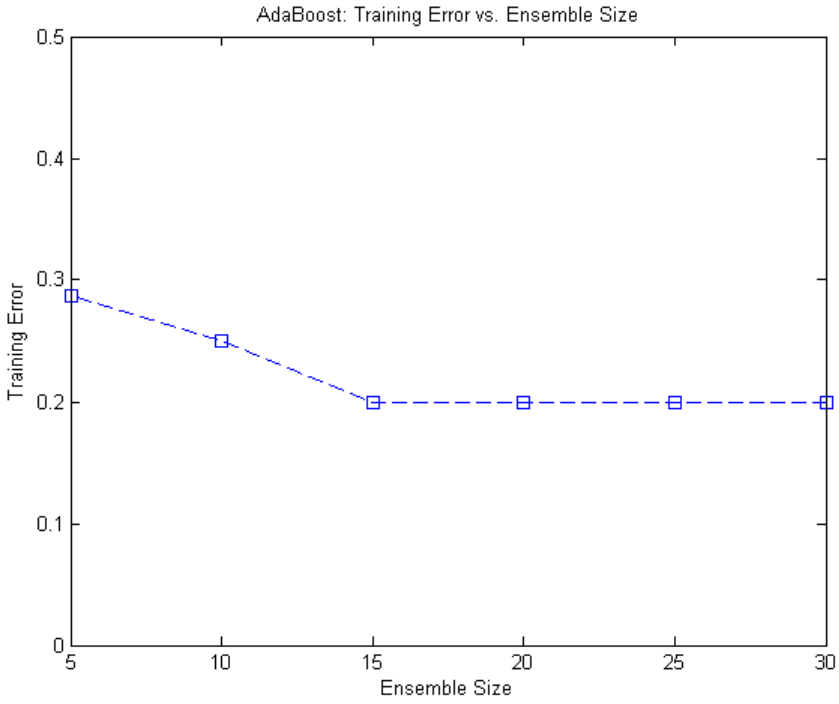

Figure 1: AdaBoost training errors on different ensemble sizes.

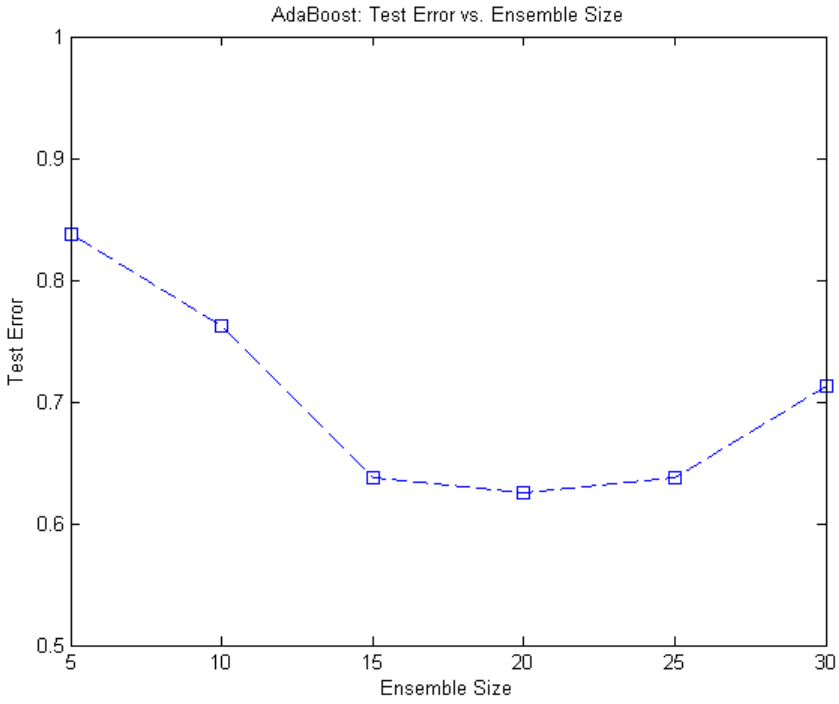

Figure 2: AdaBoost test errors on different ensemble sizes.