CS189 Homework 5 - Decision Trees for Spam Classification
Jeff Tsui, Justin Nguyen, Jeff Nieh

Decision Tree

Features

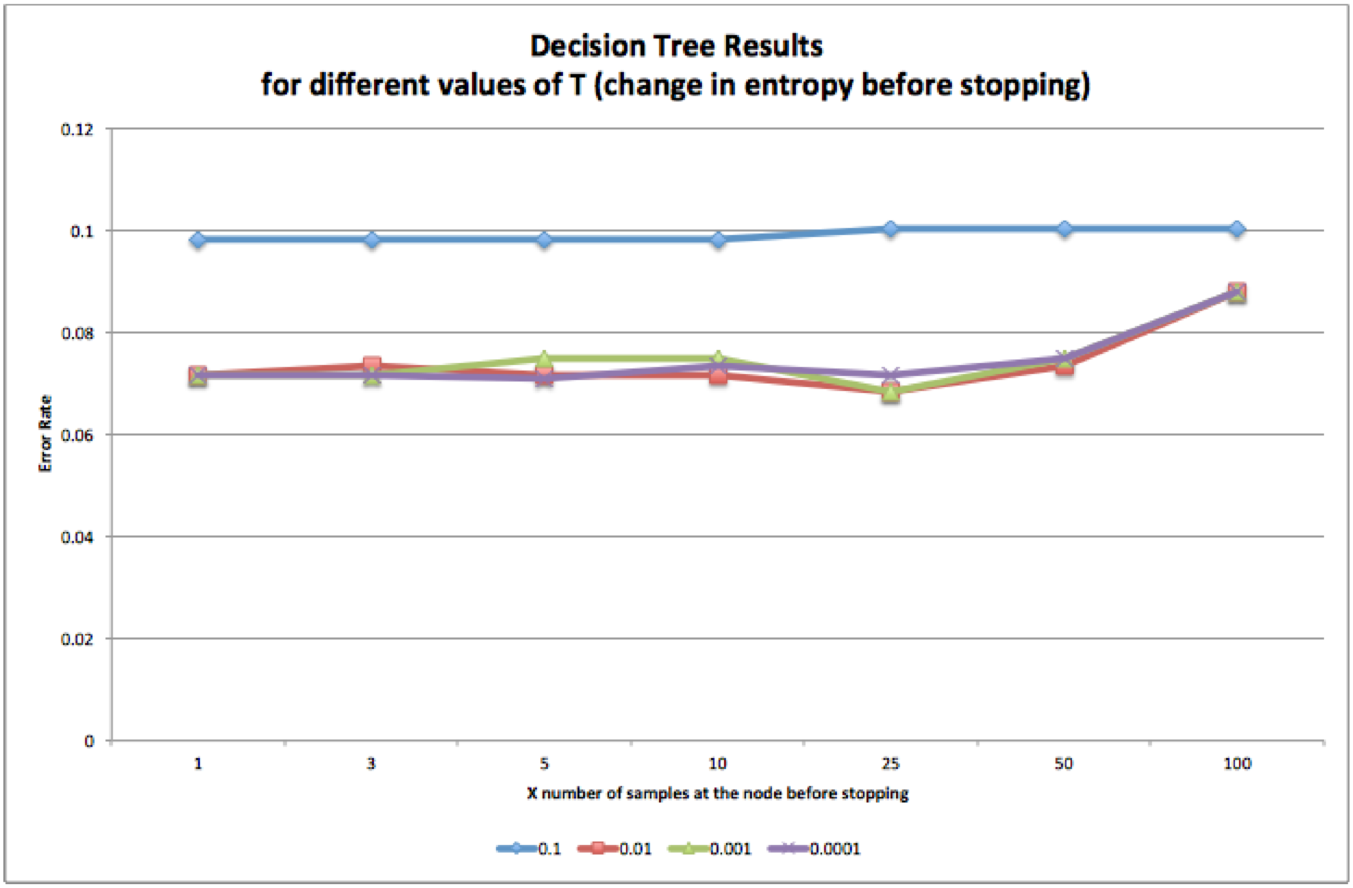
We stop building the tree when we hit fewer than **X** number of samples in the node or when the difference in the node's entropy and the newly calculated entropy is less than **T**.

Results

Lowest error rate = 6.84%
for (X, T) = (25, 0.01) and (25, 0.001)

Decreasing X to below 25 seems to result in overfitting and a slightly higher error rate. Similarly, the best values for T seem to be between 0.01 and 0.001.

X / T (see above)	**0.1**	**0.01**	**0.001**	**0.0001**
1	0.0983	0.0716	0.0716	0.0716
3	0.0983	0.0736	0.0716	0.0716
5	0.0983	0.0716	0.0749	0.0710
10	0.0983	0.0716	0.0749	0.0736
25	0.1003	0.0684	0.0684	0.0716
50	0.1003	0.0736	0.0749	0.0749
100	0.1003	0.0879	0.0879	0.0879



Random Forest

Features

Based on our code for decision trees. We updated our code so that each node stores the probability that it is spam or ham (based on the sample labels at each node), instead of just storing a boolean.

Based on Breiman's algorithm. Includes randomization of a subset of **S** samples for each tree. Each tree also randomly selects a random subset of **F** features from which to test questions. We tested results on varying **N** number of trees.

Same stopping criteria as above based on hyperparameters **X** and **T**.

Results

lowest error rate = 5.79% in the following values of hyperparameters.

The full set of results is in the table at the end of the report.

It's interesting that the lowest error rate was achieved by so many combinations hyperparameters. The row highlighted below may be the best choice since these parameters run very quickly compared to the others. Using only 50 random trees, a random subset of 500 samples in each, and a random subset of 15 out of 57 features achieved just as good results as increasing these values.

error	T	X	num_trees	sample_subset	feat_subset
0.0579	0.001	5	25	2000	45
0.0579	0.001	25	25	2000	45
0.0579	0.01	5	50	500	15
0.0579	0.01	10	50	2000	45
0.0579	0.01	25	50	2000	45
0.0579	0.001	5	50	2000	45
0.0579	0.001	10	50	2000	45
0.0579	0.001	25	50	2000	45
0.0579	0.01	5	100	500	15
0.0579	0.01	10	100	500	15
0.0579	0.01	25	100	500	15
0.0579	0.001	5	100	500	15
0.0579	0.001	5	100	500	30
0.0579	0.001	10	100	2000	30
0.0579	0.001	25	100	2000	30
0.0579	0.01	5	100	2000	45
0.0579	0.01	10	100	2000	45
0.0579	0.01	25	100	2000	45
0.0579	0.001	5	100	2000	45
0.0579	0.001	10	100	2000	45
0.0579	0.001	25	100	2000	45

Adaboost

Features

We used the decision tree from part 1 and removed the stopping criteria T and X for entropy change and number of samples in the node. We added a depth restriction so trees stop growing after depth D.

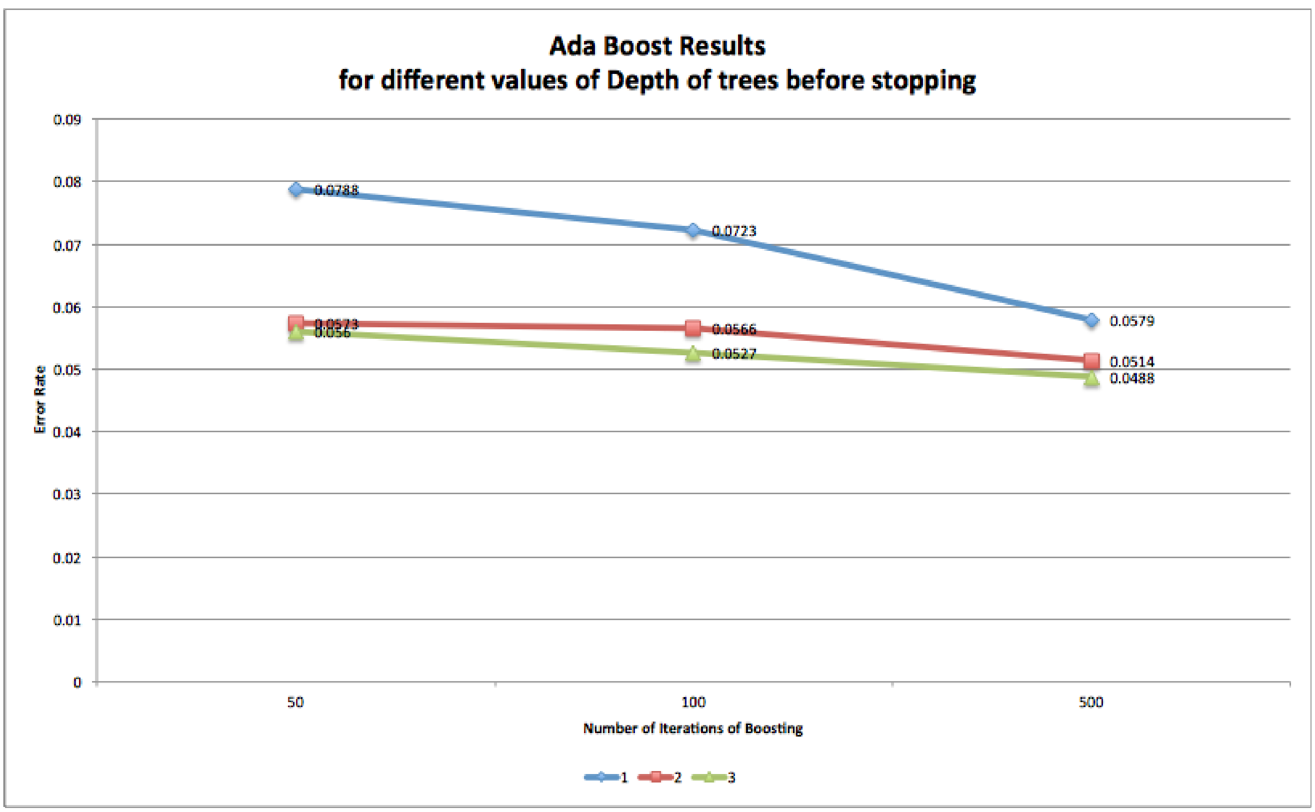
We tuned hyperparameters D and the number of iterations.

Results

lowest error rate 4.88% with depth = 3 and 500 iterations

From these results, we see that increasing the number of iterations always lowers the error rate. The depth of the tree was best for depth 3 (3 child levels), but the main differentiator is number of iterations.

Error	Depth	Iterations
0.0788	1	50
0.0573	2	50
0.056	3	50
0.0723	1	100
0.0566	2	100
0.0527	3	100
0.0579	1	500
0.0514	2	500
0.0488	3	500



Sources

http://www.onlamp.com/lpt/a/6464
http://docs.opencv.org/modules/ml/doc/boosting.html

Random Forest Complete Results

error	T	X	num_trees	sample_subset	feat_subset
0.0658	0.01	5	25	500	15
0.0664	0.01	10	25	500	15
0.0658	0.01	25	25	500	15
0.0664	0.001	5	25	500	15
0.0664	0.001	10	25	500	15
0.0671	0.001	25	25	500	15
0.0677	0.01	5	25	500	30
0.0677	0.01	10	25	500	30
0.0677	0.01	25	25	500	30
0.069	0.001	5	25	500	30
0.0684	0.001	10	25	500	30
0.0697	0.001	25	25	500	30
0.0684	0.01	5	25	500	45
0.069	0.01	10	25	500	45
0.0677	0.01	25	25	500	45
0.0684	0.001	5	25	500	45
0.0677	0.001	10	25	500	45
0.069	0.001	25	25	500	45
0.0677	0.01	5	25	1000	15
0.0671	0.01	10	25	1000	15
0.0658	0.01	25	25	1000	15
0.0671	0.001	5	25	1000	15
0.0664	0.001	10	25	1000	15
0.0671	0.001	25	25	1000	15
0.0671	0.01	5	25	1000	30
0.0671	0.01	10	25	1000	30
0.0671	0.01	25	25	1000	30
0.0664	0.001	5	25	1000	30
0.0658	0.001	10	25	1000	30

0.0664	0.001	25	25	1000	30
0.0651	0.01	5	25	1000	45
0.0645	0.01	10	25	1000	45
0.0638	0.01	25	25	1000	45
0.0632	0.001	5	25	1000	45
0.0632	0.001	10	25	1000	45
0.0638	0.001	25	25	1000	45
0.0632	0.01	5	25	2000	15
0.0618	0.01	10	25	2000	15
0.0638	0.01	25	25	2000	15
0.0625	0.001	5	25	2000	15
0.0618	0.001	10	25	2000	15
0.0632	0.001	25	25	2000	15
0.0625	0.01	5	25	2000	30
0.0612	0.01	10	25	2000	30
0.0612	0.01	25	25	2000	30
0.0612	0.001	5	25	2000	30
0.0612	0.001	10	25	2000	30
0.0605	0.001	25	25	2000	30
0.0605	0.01	5	25	2000	45
0.0605	0.01	10	25	2000	45
0.0599	0.01	25	25	2000	45
0.0579	0.001	5	25	2000	45
0.0592	0.001	10	25	2000	45
0.0579	0.001	25	25	2000	45
0.0579	0.01	5	50	500	15
0.0592	0.01	10	50	500	15
0.0592	0.01	25	50	500	15
0.0599	0.001	5	50	500	15
0.0592	0.001	10	50	500	15
0.0599	0.001	25	50	500	15
0.0599	0.01	5	50	500	30
0.0605	0.01	10	50	500	30

0.0599	0.01	25	50	500	30
0.0599	0.001	5	50	500	30
0.0605	0.001	10	50	500	30
0.0618	0.001	25	50	500	30
0.0618	0.01	5	50	500	45
0.0618	0.01	10	50	500	45
0.0618	0.01	25	50	500	45
0.0618	0.001	5	50	500	45
0.0625	0.001	10	50	500	45
0.0625	0.001	25	50	500	45
0.0632	0.01	5	50	1000	15
0.0618	0.01	10	50	1000	15
0.0625	0.01	25	50	1000	15
0.0625	0.001	5	50	1000	15
0.0632	0.001	10	50	1000	15
0.0632	0.001	25	50	1000	15
0.0632	0.01	5	50	1000	30
0.0632	0.01	10	50	1000	30
0.0632	0.01	25	50	1000	30
0.0625	0.001	5	50	1000	30
0.0625	0.001	10	50	1000	30
0.0632	0.001	25	50	1000	30
0.0618	0.01	5	50	1000	45
0.0618	0.01	10	50	1000	45
0.0612	0.01	25	50	1000	45
0.0612	0.001	5	50	1000	45
0.0612	0.001	10	50	1000	45
0.0612	0.001	25	50	1000	45
0.0618	0.01	5	50	2000	15
0.0612	0.01	10	50	2000	15
0.0612	0.01	25	50	2000	15
0.0605	0.001	5	50	2000	15
0.0599	0.001	10	50	2000	15

0.0599	0.001	25	50	2000	15
0.0586	0.01	5	50	2000	30
0.0586	0.01	10	50	2000	30
0.0586	0.01	25	50	2000	30
0.0586	0.001	5	50	2000	30
0.0586	0.001	10	50	2000	30
0.0586	0.001	25	50	2000	30
0.0586	0.01	5	50	2000	45
0.0579	0.01	10	50	2000	45
0.0579	0.01	25	50	2000	45
0.0579	0.001	5	50	2000	45
0.0579	0.001	10	50	2000	45
0.0579	0.001	25	50	2000	45
0.0579	0.01	5	100	500	15
0.0579	0.01	10	100	500	15
0.0579	0.01	25	100	500	15
0.0579	0.001	5	100	500	15
0.0586	0.001	10	100	500	15
0.0586	0.001	25	100	500	15
0.0586	0.01	5	100	500	30
0.0586	0.01	10	100	500	30
0.0586	0.01	25	100	500	30
0.0579	0.001	5	100	500	30
0.0586	0.001	10	100	500	30
0.0586	0.001	25	100	500	30
0.0592	0.01	5	100	500	45
0.0592	0.01	10	100	500	45
0.0605	0.01	25	100	500	45
0.0612	0.001	5	100	500	45
0.0612	0.001	10	100	500	45
0.0618	0.001	25	100	500	45
0.0618	0.01	5	100	1000	15
0.0618	0.01	10	100	1000	15

0.0605	0.01	25	100	1000	15
0.0605	0.001	5	100	1000	15
0.0605	0.001	10	100	1000	15
0.0612	0.001	25	100	1000	15
0.0605	0.01	5	100	1000	30
0.0605	0.01	10	100	1000	30
0.0605	0.01	25	100	1000	30
0.0605	0.001	5	100	1000	30
0.0612	0.001	10	100	1000	30
0.0605	0.001	25	100	1000	30
0.0612	0.01	5	100	1000	45
0.0612	0.01	10	100	1000	45
0.0612	0.01	25	100	1000	45
0.0599	0.001	5	100	1000	45
0.0599	0.001	10	100	1000	45
0.0599	0.001	25	100	1000	45
0.0599	0.01	5	100	2000	15
0.0592	0.01	10	100	2000	15
0.0592	0.01	25	100	2000	15
0.0592	0.001	5	100	2000	15
0.0592	0.001	10	100	2000	15
0.0586	0.001	25	100	2000	15
0.0592	0.01	5	100	2000	30
0.0592	0.01	10	100	2000	30
0.0592	0.01	25	100	2000	30
0.0586	0.001	5	100	2000	30
0.0579	0.001	10	100	2000	30
0.0579	0.001	25	100	2000	30
0.0579	0.01	5	100	2000	45
0.0579	0.01	10	100	2000	45
0.0579	0.01	25	100	2000	45
0.0579	0.001	5	100	2000	45
0.0579	0.001	10	100	2000	45

0.0579	0.001	25	100	2000	45