HOMEWORK 5
DECISION TREES FOR SPAM CLASSIFICATION

Varun Pemmaraju – cs189-bz
Lauri Takacsi-Nagy – cs189-ag
Thomas Chow – cs189-

Features Implemented

We implemented the following:

- Decision Tree using the ID3 algorithm with an entropy heuristic
- Boosted trees with 100 iterations using Adaboost
- Random forest using 100 trees

Specifically, our decision tree implementation used the ID3 algorithm with no stopping criteria outside of being a completely pure label. It also was enhanced with 4-fold cross-validation.

Regarding boosted trees, we resample with replacement using the weighted distribution by utilizing the `scipy.stats` module. We also are using decision stumps of depth 2 (using a depth-limited implementation of id3) to make up the weak classifiers.

Finally, for random forest we tuned the dataset size and the feature size and again used 4-fold cross-validation. This produces the lowest error rates amongst the various features implemented.

Resulted Obtained

The tables below summarize the estimated accuracies for each implemented feature (obtained via 4-fold cross validation):

Method	Decision Tree	Boosted Tree	Random Forest
Accuracy	90.3%	92.3%	95.1%

References

1. Adaboost Paper
2. http://www.robots.ox.ac.uk/~az/lectures/cv/adaboost_matas.pdf
3. http://en.wikipedia.org/wiki/ID3_algorithm