CS189: Introduction to Machine Learning

Homework 5

Due: 11:59 p.m. April 7, Monday, 2014

Decision Trees for Spam Classification

In this homework, you will implement decision trees (and its extensions) for classifying spam using the spam dataset provided to you as part of Homework 4.

In lectures, you were given a basic introduction to decision trees and how such trees are trained. You were also introduced to random forests and boosting algorithms (AdaBoost). This homework is meant to test your ability to implement decision trees, averaging (random forests) and boosting as explained in class as well as to research different decision tree techniques (stopping criteria, pruning, dealing with missing attributes, splitting criteria other than entropy, heuristics for faster training, complex decisions at nodes, cross-validation etc.) from resources on the internet/books/blogs. The reading attached with this homework might be a useful point to start off! The rules for the homework are as follows:

1. There are three compulsory elements that you have to implement: **Decision trees** (without pruning), **Random forests** and **Boosted trees (AdaBoost)**. All other components will be given extra credit (even if it doesn't improve performance!).
2. You are **NOT** allowed to use any off the shelf decision tree/random forest/boosting implementation for the homework. You can use external libraries for data preprocessing, faster computation etc. as long as you cite them in your report. You can use any programming language you wish to as long as we can run it with minimal effort.
3. You can work in groups of up to **3** (yes 3, not 2) and there should be **ONE** submission per group.
4. As deliverables for this homework, you need to submit your code, a README and a report explaining the features you implemented, the results you obtained, references to all external sources you used (code, articles, papers, books etc.) and anything else you might want to include (analysis of results, performance curves etc.). Submissions are through **bSpace** and **Kaggle**.

Hopefully by the end of this homework everyone will have their own version of decision trees running and can try it on different datasets! Good luck!

Appendix

Below are some of the error rates we obtained during our experiments with decision trees.

Classifier	**Error rate**
1 Classification Tree (no pruning)	0.074
1 Classification Tree (with pruning)	0.068
Boosted trees (500 itearations)	0.061
Random forest (100 trees, no pruning)	0.049
Random forest (100 trees, with pruning)	0.047

All the above experiments were done with no preprocessing of the data and entropy impurity using MATLAB's decision tree implementation (you are NOT supposed to use this!).