CS181 Homework 1

Lucas Freitas and Angela Li

February 16, 2013

1 Decision Trees and ID3

(a)

$$H(\mathrm{D}) = \frac{4}{7}\log_2\frac{7}{4} + \frac{3}{7}\log_2\frac{7}{3} \approx 0.985$$

Calculating mutual information for D and A:

$$H(\mathrm{D}|\mathrm{A} = \text{true}) = 1$$

$$H(\mathrm{D}|\mathrm{A} = \text{false}) = \frac{2}{3}\log_2\frac{3}{2} + \frac{1}{3}\log_2 3 \approx 0.918$$

$$H(\mathrm{D}|\mathrm{A}) = \frac{4}{7}H(\mathrm{D}|\mathrm{A} = \text{true}) + \frac{3}{7}H(\mathrm{D}|\mathrm{A} = \text{false}) \approx 0.965$$

$$I(D;A) = H(D) - H(D|A) \approx 0.02$$

Calculating mutual information for D and B:

$$H(\mathrm{D}|\mathrm{B} = \text{true}) = 1$$

$$H(\mathrm{D}|\mathrm{B} = \text{false}) = \frac{3}{5}\log_2\frac{5}{3} + \frac{2}{5}\log_2\frac{5}{2} \approx 0.971$$

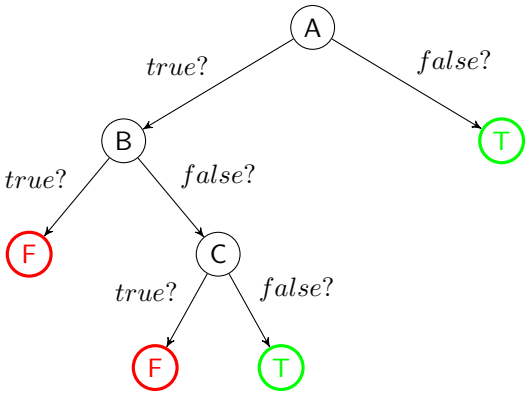
$$H(\mathrm{D}|\mathrm{B}) = \frac{2}{7}H(\mathrm{D}|\mathrm{B} = \text{true}) + \frac{5}{7}H(\mathrm{D}|\mathrm{B} = \text{false}) \approx 0.965$$

$$I(D;B) = H(D) - H(D|B) \approx 0.006$$

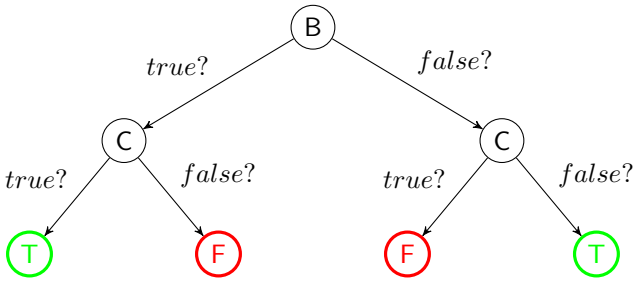
Thus, ID3 would split on A, since that gives a higher information gain. An argument in favor of splitting on A is that it is a more extreme split with more disparate probabilities, meaning it will more resolutely divide the data. An argument in favor of splitting on B is that it will classify a larger subset well.

This shows that the inductive bias of ID3 is to prefer more extreme splits that result in shorter trees.

(b) From looking at the data it is fairly obvious that in figuring out the first attribute to split on, B, C, and D have the same information gain. Since we calculated that A has a higher information gain than B, we'll first split on A. When A is 1, we can split on either B or C, which still have the same information gain at this step; we can recurse in a similar fashion from there, preferring binary splits. When A is 0, we can again split on either B or C, but (assuming we pick B) we run into the base case of no more information gain after traveling down the 0 path. So we can set that leaf to the most common label, 1.



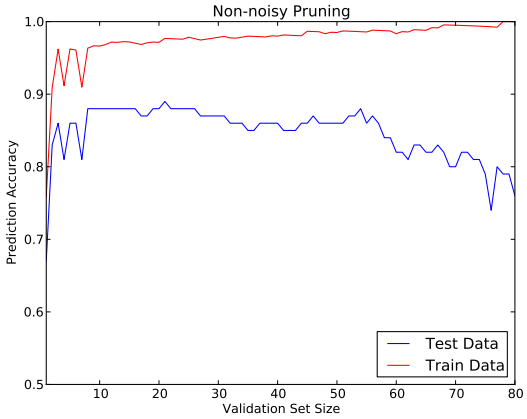
(c) This tree has a training error of 1/7, just like the ID3 tree produced above. It shows us that ID3 certainly does not always generate the shortest trees, and specifically that preferring higher information gain for individual attributes at a time (and not being able to look ahead at combinations of attributes) often does not produce the most concise tree.

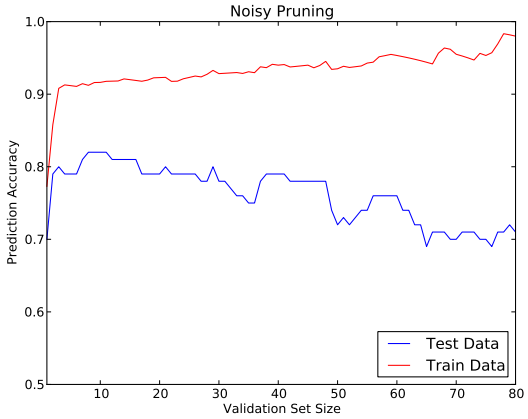


2 ID3 with Pruning

(a) The average cross-validated training and test performance over the ten trials is 0.87 for the non-noisy dataset and 0.78 for the noisy dataset.

(b) i. Graphs outputted using `plot_pruning()`:





ii.
- For non-noisy data: the performance of the train data initially oscillates, but then increases almost monotonically, reaching 100% performance when the validation set size reaches 78. The test data performance oscillates a lot, reaching a maximum at validation set size 21, with performance 89%, which is slightly higher ($\approx$ 2.3% increase) than the one obtained without pruning. The performance then decreases, reaching 76% performance for a validation set of size 80.
- For noisy data: the performance of the train data initially oscillates, but then also increases almost monotonically, reaching $\approx$ 98.3% performance when the validation set size reaches 78. The test data performance oscillates a lot, reaching a maximum at validation set size 8, with performance 82%, which is slightly higher ($\approx$ 5.1% increase) than the one obtained without pruning. The performance then decreases, reaching 71% performance for a validation set of size 80.

iii. For some values of the validation set, yes. As mentioned in the previous item, for the non-noisy data we reach an 89% performance at validation set size 21, which is a $\approx$ 2.3% increase in the performance of ID3, while for noisy data the maximum occurs at validation set size 8 with 82% performance, which is a $\approx$ 5.1% increase) in the performance of ID3. If the validation set is too large, however, the ID3 performance decreases ($\approx$ 12.6% performance decrease for non-noisy and 8.97% for noisy).

iv. There are potentially problems with overfitting with such a small data set; however, since pruning is supposed to reduce the problem of overfitting but didn't improve performance by a significant amount for our data set, we can probably infer that overfitting was not a large problem for the given data.

3 Boosting

(a) For $c = 1$, W_c is $\frac{1}{2}$ because there is one instance with weight $\frac{1}{2}$. For $c = 0$, there are $N - 1$ instances each with weight $\frac{\frac{1}{2}}{(N-1)}$, for a total W_c of $\frac{1}{2}$. The weighted entropy is then $\frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2 = 1$.

i. A. Table showing cross-validated test performances on both datasets for 10 and 30 rounds of boosting, and depths of 1 and 2:

(non-noisy)

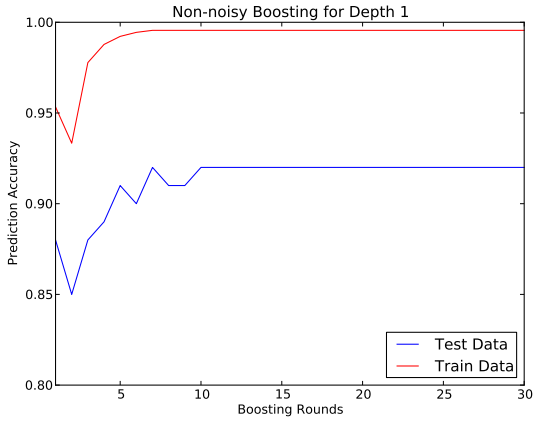
	10	30
1	0.92	0.92
2	0.89	0.89

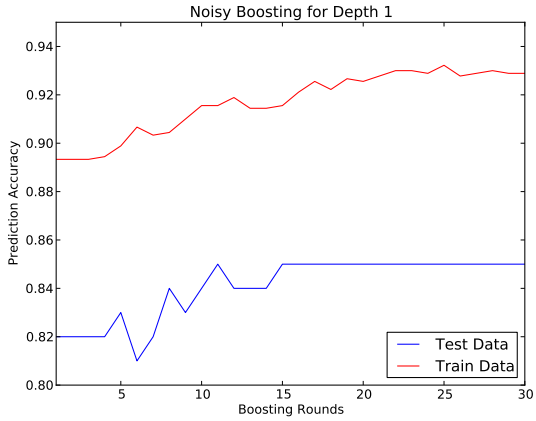
(noisy)

	10	30
1	0.84	0.85
2	0.81	0.83

It appears that increasing the maximum depth of the weak learner actually decreases the overall performance of boosting. This is likely because the learners of depth 2 are too strong, making them overfit the data on their own.

B. Graph of cross-validated test performances of the decision trees learned by boosting over different numbers of rounds (with weak learner depth 1):

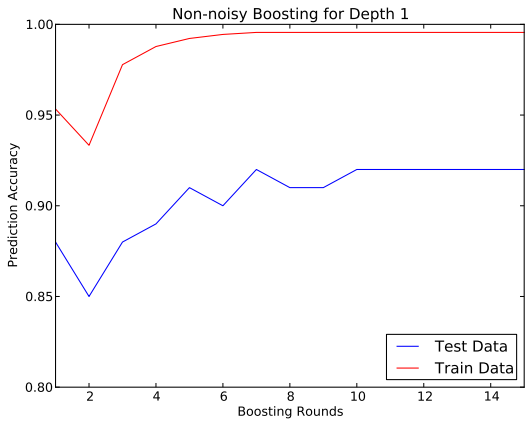


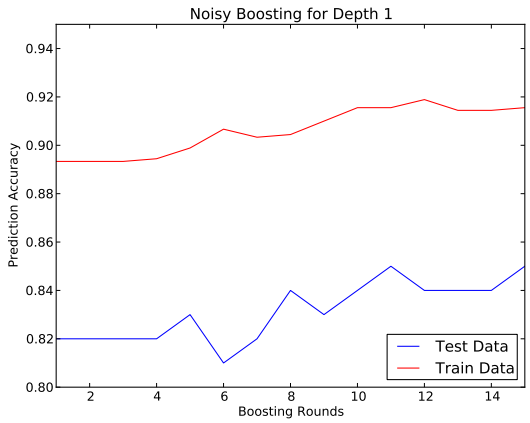


The trend is that as the number of boosting rounds increases, performance also increases up to some asymptotic bound. This seems to fit with what we discussed in class; namely, that a host of weak learners is likely to have very good accuracy (especially when compared with a small number of weak learners).

C. The cross-validated test performance of boosting is better than the performance of simple ID3 without pruning and ID3 with pruning. With boosting, the test performance reaches 0.92 accuracy, while using only ID3 it reached 0.87 (without pruning) and 0.89 (with pruning).

D. Graphs of cross-validated training and test performance for boosting with weak learners of depth 1 over a number of rounds in [1, 15]:



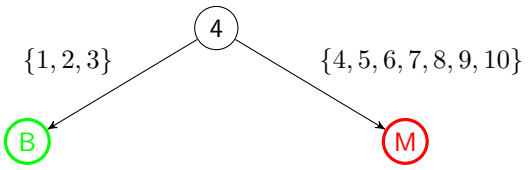


Of course both the training and test performance increase as the number of boosting rounds increase. Furthermore, peaks and valleys in the training performance seem to correspond with peaks and valleys in the testing performance (though they are more pronounced on the test set).

Aside from that, however, we can also see that test performance actually continues to increase even after the training data has reached what appears to be its asymptotic upper bound. This implies that there is more than simple correlation between the training and test performances.

4 Tree Analysis

We were most interested in finding trees that were not only efficient but also extremely simple, so we ran ID3 using increasing maximum depths (starting with 1) in order to find a simple but well-performing tree. We discovered one such tree with depth 1 that is actually capable of outperforming ID3:



Attribute 4 corresponds to "Uniformity of cell shape", and thus it appears that simply looking at uniformity of cell shape is an extremely, extremely good indicator of whether a tumor is benign or malignant. This makes a good argument for the case of making the shortest and simplest trees possible.

Philosophical tangent: This is highly interesting from a theoretical point of view, but in practice, a $\approx 93\%$ success rate of diagnosing tumors isn't nearly good enough when the consequence of a misdiagnosis is potentially a human life. If we wanted a good decision tree for use in early detection of cancer, we might prefer one that maybe has a lower accuracy rate overall but a has a very high sensitivity (that is, a higher probability of identifying malignant tumors). Currently, of course, neither or ID3 or AdaBoost algorithms allow for this. But AdaBoost could be modified to weight data such that malignant examples wrongly classified as benign are given extremely high importance for future hypotheses, while benign examples wrongly classified as malignant are given less importance; similarly hypotheses that wrongly classify malignant examples as benign would be given less importance.