

Joy Ming and Alisa Nguyen (9 February 2013)

1. Decision Trees and ID3

(a) ID3 will chose to split on \boxed{A} because it has a higher information gain.

- Splitting on A will have an information gain of $Gain(X_k, A) = H(A) - Remainder(X_k, A) = 0.02$, where $H(A) = \frac{3}{7} \log_2 \frac{7}{3} + \frac{4}{7} \log_2 \frac{7}{4} = 0.985$, and $Remainder(X_k, A) = \frac{4}{7}(\frac{2}{4} \log_2 2 + \frac{2}{4} \log_2 2 + \frac{3}{7}(\frac{2}{3} \log_2 \frac{3}{2} + \frac{1}{3} \log_2 3)) = 0.965$
- Splitting on B will have an information gain of $Gain(X_k, B) = H(B) - Remainder(X_k, B) = 0.005$, where $H(B) = \frac{2}{7} \log_2 \frac{7}{2} + \frac{5}{7} \log_2 \frac{7}{5} = 0.985$, and $Remainder(X_k, B) = \frac{2}{7}(\frac{1}{2} \log_2 2 + \frac{1}{2} \log_2 2 + \frac{5}{7}(\frac{3}{5} \log_2 \frac{5}{3} + \frac{2}{5} \log_2 \frac{5}{2})) = 0.98$

This example shows that ID3 has an inductive bias of strongly preferring extreme partitions and larger subsets. In this case, looking at the results of the

(b) In this example, a tree that could be formed would split first on A, then B, then C, as shown below.

(c) By eyeballing the data

2. ID3 with Pruning

(a) The average cross-validated training performance was:

- Non-noisy: Training $\boxed{1.0}$ and test $\boxed{0.87}$.
- Noisy: Training $\boxed{0.98}$ and test $\boxed{0.78}$.

(b) After the pruning function:

- Graph
- The cross-validated performance of the validation set pruning improves at first, as the validation increases from 1, peaks at a point around size 40 to 60, and then worsens in performance as the validation set size becomes too large and overfitting becomes an issue.
- The validation set pruning improves the cross-validated performance of ID3 on these data for all the data points leading up to the peak when comparing against validation-size. After the peak, pruning gives us similar and sometimes slightly worse results than the cross-validated performance of ID3. On the nosy data, the average cross-validated test performance with pruning on the non-noisy dataset is 0.8599 and without pruning 0.855.
- Overfitting is an issue for these data, as evidenced by the dropoff after a peak when the validation set size gets too large. ELABORATE.

3. Boosting

(a) The weighted entropy of the set can be calculated:

$$W = 0.5 + \frac{0.5}{N-1}(N-1) = 1 \quad H = 0.5 \log_2 \frac{1}{0.5} + 0.5 \log_2 \frac{1}{0.5} = \boxed{1}$$

i. Analyze the effectiveness of boosting:

A. Effect of maximum depth on cross validated boosting in noisy and not noisy data

Noisy?	Max depth	$R = 10$	$R = 30$
N	1	0.82	0.84
NN	1	0.89	0.91
N	2	0.81	0.79
NN	2	0.87	0.87

For each given set of data that is the same

noisyness and number of rounds, it seems the greater the maximum depth the less accurate the created learner is. This is in part because the bigger the tree, the less of a "weak" learner it is. In this example, the greater the depth the more splits the tree will go through and thus will create a higher probability of overfitting. NEEDS MORE BUZZ WORDS

- B. Effect of number of boosting rounds on cross-validated performance of decision trees (graph)
This is expected based on our theoretical discussion of boosting in class because
- C. Comparing cross-validated test performance of boosting with ID3 with and without pruning.
- D. Comparing cross-validated training and test performance for boosting with weak learners of depth 1 over a number of rounds in $[1, 15]$.
(graph)
The relationship between training and test performance shows that

4. Tree Analysis