CS181 Assignment 1

Joy Ming and Alisa Nguyen (9 February 2013)

1. Decision Trees and ID3

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

   - Splitting on A will have an information gain of $Gain(X_k, A) = H(A) - Remainder(X_k, A) =$??, where $H(A) = \frac{3}{7}\log_2\frac{7}{3} + \frac{4}{7}\log_2\frac{7}{4} =$??, and $Remainder(X_k, A) = \frac{4}{7}(\frac{2}{4}\log_2 2 + \frac{2}{4}\log_2 2 + \frac{3}{7}(\frac{2}{3}\log_2\frac{3}{2} + \frac{1}{3}\log_2 3) = 0.96$
   - Splitting on B will have an information gain of $Gain(X_k, B) = H(B) - Remainder(X_k, B) =$??, where $H(B) = \frac{2}{7}\log_2\frac{7}{2} + \frac{5}{7}\log_2\frac{7}{5} =$??, 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 and 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:

   i. Graph

   ii. The

3. Boosting

4. Tree Analysis