CS 181
Problem Set 1

1. (a)

   (b)

   (c)

2. (a) The average ten-fold cross-validation score for the non-noisy data is 0.87 while the score for noise data is 0.78.

   (b)

3. (a) We are defining our information gain criterion to be the mutual information between an attribute and the label. However, we will modify all the calculated specific conditional entropies according to the weights of the examples. So instead of:

   $$H(X|y) = \Sigma_x p(x|y) log_2 \frac{1}{p(x|y}$$

   We will be replacing the probability distribution p with a weighted distribution based on the weights of the data. This way, data that is "more important" (the ones that tend to be predicted incorrectly) will be weighted more when the algorithm is trying to decide what attribute to split on. Attributes that split the more important data correctly will be favored over those that do not. For the example given where the label of the first example is $y_1 = 1$, etc, the weighted entropy of the set is

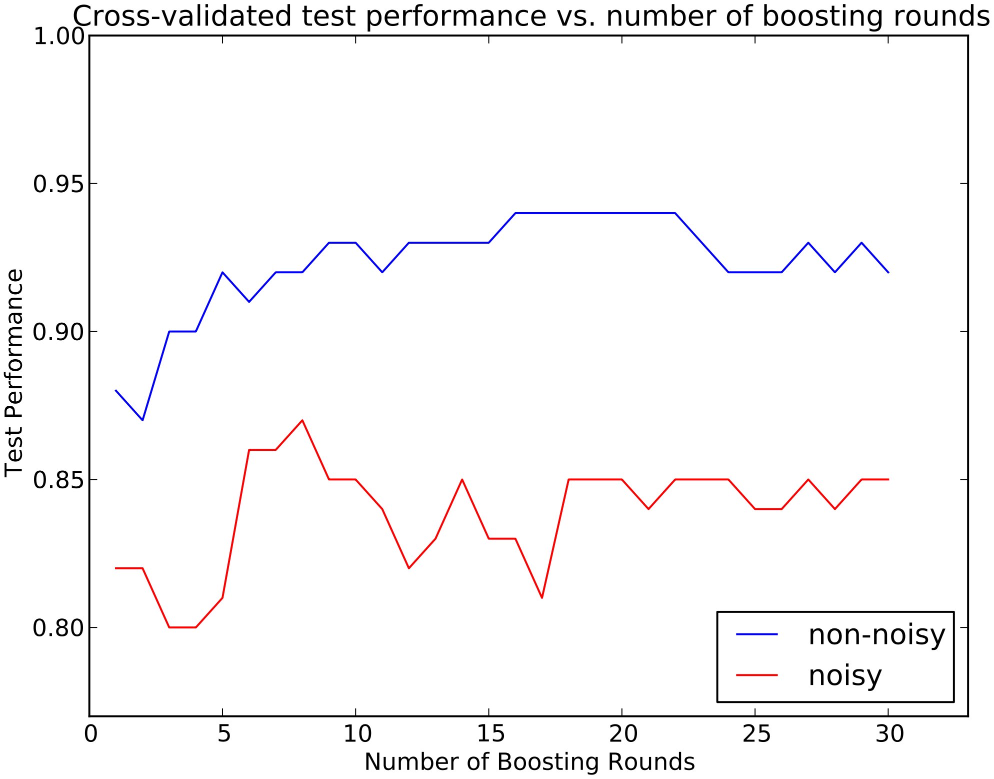
   $$0.5 * log_2 2 + 0.5 * log_2 2 = 1.$$

   (b) i. Table of how the depth of the tree affects boosting

	10 rounds	20 rounds
maxDepth = 1	0.93	0.92
maxDepth = 2	0.86	0.86

   A larger maximum depth actually hurts the cross-validated test performance because when the individual trees are more complex, they are more susceptible to overfitting. This introduces trees with splits that are based more on noise than signal.

   ii. Boosting performs consistently worse on noisy data compared to non-noisy data no matter how many rounds (from 1 to 30) are performed. This is because the algorithm gives highers weights to data points that are predicted incorrectly. Therefore, if there are inconsistencies within the data, the algorithm will focus on the inconsistent data more than the rest of the data.


Cross-validated test performance vs. number of boosting rounds
Test Performance
Number of Boosting Rounds
1.00
0.95
0.90
0.85
0.80
0
5
10
15
20
25
30
non-noisy
noisy

iii. jh

iv. kj