Programming Assignment 1 Write-up
Josh Veltri (jkv13)
Justin Wang (jsw104)

(a) What is the CV accuracy of the classifier on each dataset when the depth is set to 1?

Dataset	**CV Accuracy w/ Depth = 1**
voting	0.9837
volcanoes	0.7222
spam	0.6639

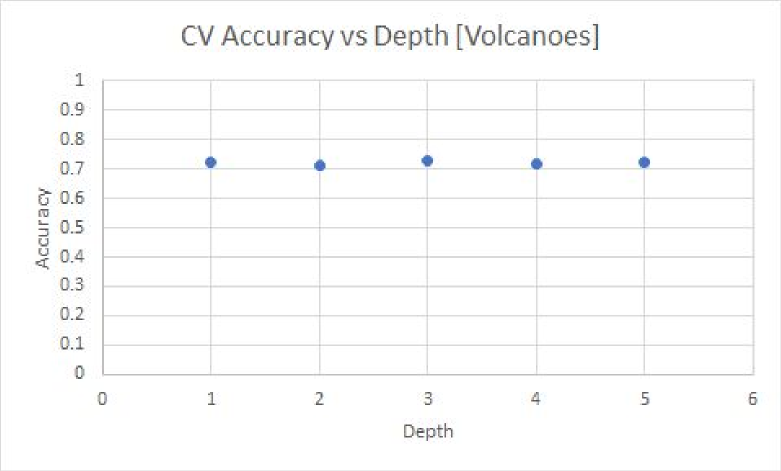
(b) For spam and voting, look at first test picked by your tree. Do you think this looks like a sensible test to perform for these problems? Explain.

voting: The first test selected is for the attribute *Repealing-the-Job-Killing-Health-Care-Law-Act*. This seems like a sensible test -- healthcare is one of the most contentious topics in politics today and it is well known that support/opposition for certain approaches to healthcare divides very cleanly by political party. Thus it is not surprising that we can sort voters almost perfectly by using this attribute alone.

spam: The first test selected is for the attribute *OS*. This at least feels like a sensible test -- the market of mail servers used for legitimate non-spam email is likely dominated by just a few major players, so any mail sent by a mail server running a different OS could have a high risk of being spam.

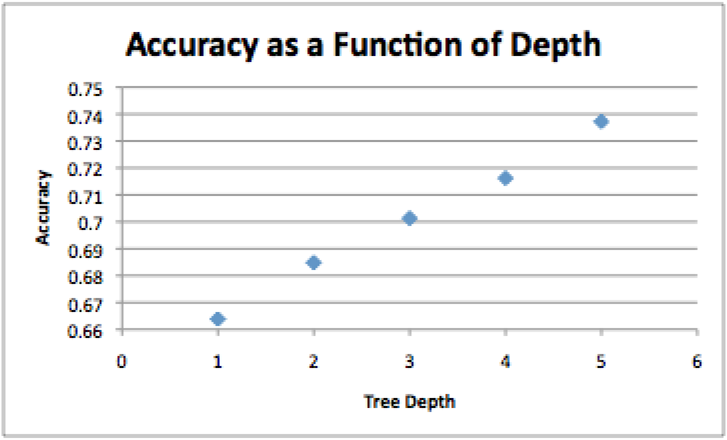
(c) For volcanoes and spam, plot the CV accuracy as the depth of the tree is increased. On the x-axis, choose depth values to test so there are at least five evenly spaced points. Does the accuracy improve smoothly as the depth of the tree increases? Can you explain the pattern of the graph?

volcanoes:



For volcanoes, we see that the CV accuracy remains essentially constant for depths from 1 to 5. This looks like the start of overfitting -- we expect that if we built trees with even larger depths, the CV accuracy would start to trend downward. (These accuracy values correspond with trees grown using the information gain split criteria.)

spam:



For spam, we see that the CV accuracy increases as the depth of the tree increases. Intuitively, the more we grow the tree the less entropy at the leaf nodes, which means the more accurate the tree will become as a predictor. We expect that if we built trees with even larger depths, the graph of accuracy as a function of depth would eventually begin to flatten and eventually decrease (which would represent overfitting). (These accuracy values correspond with trees grown using the gain ratio split criteria.)

(d) Pick 3 different depth values. How do the CV accuracies change for gain and gain ratio for the different problems for these values?

voting

Depth	CV Accuracy with Information Gain	CV Accuracy with Gain Ratio
1	0.9873	0.9873
3	0.9860	0.9860
5	0.9860	0.9860

volcanoes

Depth	CV Accuracy with Information Gain	CV Accuracy with Gain Ratio
1	0.7222	0.7222
3	0.7100	0.7177
5	0.7109	0.7191

spam

Depth	CV Accuracy with Information Gain	CV Accuracy with Gain Ratio
1	0.6639	0.6639
3	0.7394	0.7013
5	0.7559	0.7372

For both voting and volcanoes, the information gain and the gain ratio accuracy values are essentially equal at depths of 1, 3, and 5. Information gain ratio primarily protects against nominal values unique to each example that return a large information gain but provide little generalizable information in practice. Neither the voting nor the volcanoes data set exhibit contain attributes that would suffer from this problem -- the voting attributes are all 3-valued nominal while the volcanoes attributes are nearly all continuous -- so we would expect similar split decisions and therefore are not surprised to see little difference between the two accuracy measures.

For spam, the accuracy using the information gain and the gain ratio are equal at depth 1, but for depths 3 and 5, the gain ratio accuracy begins to fall behind the information gain accuracy.

(e) Compare the CV accuracies and the accuracy on the full sample for depths 1 and 2. Are they comparable?

voting

Depth	CV Accuracy	Full-Sample Accuracy
1	0.9873	0.9886
2	0.9883	0.9932

volcanoes

Depth	CV Accuracy	Full-Sample Accuracy
1	0.7222	0.7297
2	0.7051	0.8669

spam

Depth	CV Accuracy	Full-Sample Accuracy
1	0.6639	0.6639
2	0.7320	0.7323

From the tables of results above, we see that the CV accuracy and the full-sample accuracy are similar for a depth of one for all of, voting, volcanoes, and spam. However, for a depth of 2, the voting and spam accuracies remain similar, while the CV accuracy significantly trails the full-sample accuracy for volcanoes. This would seem to indicate overfitting on the volcano data.

Write down any further insights or observations you made while implementing and running the algorithms, such as time and memory requirements, the complexity of the code, etc. Especially interesting insights may be awarded extra points.

We had to re-think our approach multiple times due to time and memory requirements. First, we reduced the number of nodes from having a node for each possible split value of a continuous attribute to just one node per continuous attribute to save memory. We also had to re-do our calculations for the best split threshold for a continuous attribute multiple times in order to avoid iterating through the example set as much as possible. These two improvements allowed us to run our algorithms in a reasonable amount of time and without destroying our RAM. If we were to have more time, we would have further cut down our memory usage by storing one global example set and simply passing around indices to the nodes to do calculations instead of passing around copies of the example set. We saw this affect the Spam data set the most because of the large number of examples in this test set.