EECS 440: Machine Learning Fall 2015 Write-up for Project 1, assigned 9/8. Max points: 20.

Yingcheng Sun ID: yxs489

1. Answer the following questions:

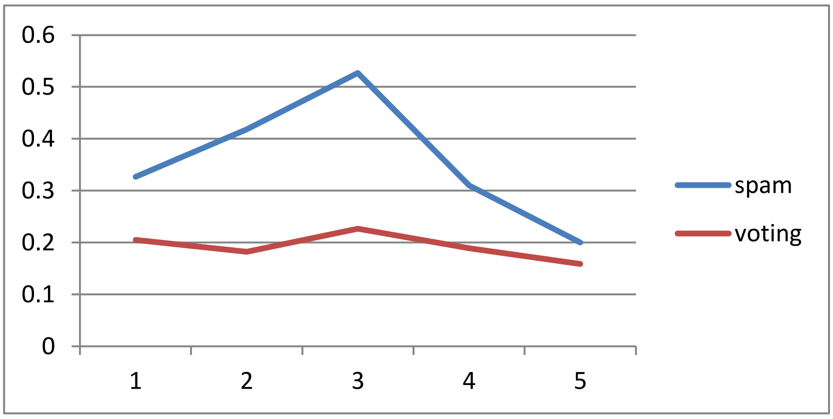
   (a) What is the accuracy of the classifier on each dataset when the depth is set to 1? (i.e. the tree has just 1 test)

   Answer: Voting: 20.7 % Spamming: 21.3 % Volcanoes : 20.4 %

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

   No. Since we use gain ratio, the decision tree does not be sensible to the test data.

   (c) For volcanoes and spam, plot the 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?



   The plot goes down after depth three, maybe it caused overfitting.

2. 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. You can also test your code with classification problems from your research.

   Answer: The method I used to partition the tree is based on the theory that the only split values we need to check to determine a split with max IG(X) lie between points with different labels, so it will decrease the running time and space consuming.

```
for i,(ex1,ex2) in enumerate(zip(dataset[:-1],dataset[1:])):
            part_data = {}
            if ex1[-1] == ex2[-1]: #not a threshold
                continue
```