(364-1-1441) Foundations of Artificial Intelligence

Problem Set 2

Due: 11/1/2019

1 Building a Spam Filter

In this exercise you will use a decision tree to build a spam filter, that allows you to decide if an email is spam (and should be ignored by the user) or not. In order to do this, you shall use the data file `spambase.data`, which you can download from the course Moodle. The description of the data file is found in the `spambase.names` file you can also download from the course Moodle.

Broadly, the first variable indicates if the email was spam or not, and the rest show how many times (out of all the words) did particular words appear in the email. Since these variables are continuous, you will need to set thresholds yourself – which means each of your trees might break different than someone else's (which is fine!).

When you build your decision trees, you are expected to use entropy to calculate the more meaningful variables, and to use the χ^2 test to prune vertices.

Your Python code will include the following functions. You can assume the file `spambase.data` is found in the same directory as your code.

`buildTree(<float> k)` `k`$\in [0,1]$. You need to build a decision tree, using k ratio of the data (so if $k = 0.6$, you arbitrarily choose 60% of the data), and validate it on the remainder. The outcome is printing out the decision tree, and reporting the error.

`treeError(<int> k)` You need to report the quality of the decision tree by building k-cross validation, and reporting the error.

`isThisSpam(<array>)` You receive an an input from the user, of the email they received. You can assume it is in the same order as the data file (but without the bit saying if it's spam or not, of course). You need to return 1 if it is spam and 0 if it is not.