Danial Ghofrani CS486 Assignment 3 July 2015

1-

I chose to implement a solution in C++. I wrote my own parser. I did not pick C++ because I thought it was easy, I chose it because of the superior performance of a compiled solution (compared to python that is almost often interpreted and yields poor performance) I also wanted to demonstrate my object oriented design and C++ prowess (I had used JavaScript in assignment 1 and 2).

The source code is available in the file Classification.cpp please make sure to run the executable in the same folder as the .txt files that are provided.

Please note that my program uses “depth” as the maximum possible depth of the tree starting from the maximum “depth” at the root and ending with depth=0 for the leaves. This is a minor technicality not relevant to the performance of the decision tree.

The following is the graph of accuracy achieved by the program:

Overfitting definitely happens especially after depth=4 in the tree. Therefore depth 4 is the best depth for classification.

Here is the tree:

writes

wrote

G

A

G

A

archive

G

A

A

use

graphics

bible

that

god

image

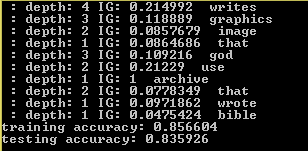
G

G

that

A

A



Red edges indicate that document contains the word, blue edged indicate that the document does not contain the word.

The words chosen make sense especially ones such as graphics and god that are such excellent discriminators at level 2. Notice that the leaf node at level 2 of the tree indicates the documents that contain the word writes and also contain “graphics” turns out all of them are graphics related documents. At next level, documents that do not contain the word “graphics” but contain the word “image” are all graphics documents also.

On the left subtree we can see that documents that contain “god” but don’t contain “use” are all atheism related.

Notice that premature leaf nodes indicate that the leaf is 100% made of either graphics or atheism documents, whereas the leafs at the very bottom are not guaranteed to be pure, rather they use the mode.