# MACHINE LEARNING

## ASSIGNMENT 2: DECISION TREE

SARTHAK CHAKRABORTY (16CS30044)

## Part 2 (20 newsgroup Dataset)

Decision Tree was implemented on the preprocessed 20 newsgroup dataset to classify documents into two groups. Maximum depth is taken as a parameter by the tree. If a branch of the decision tree exceeds the specified depth, then it is not grown further and the label becomes the one having a maximum frequency in that node. This is called **pre-pruning.** The decision tree was built using scikit-learn as well.

A graph showing the accuracy of the classifier with respect to the maximum depth is given below:

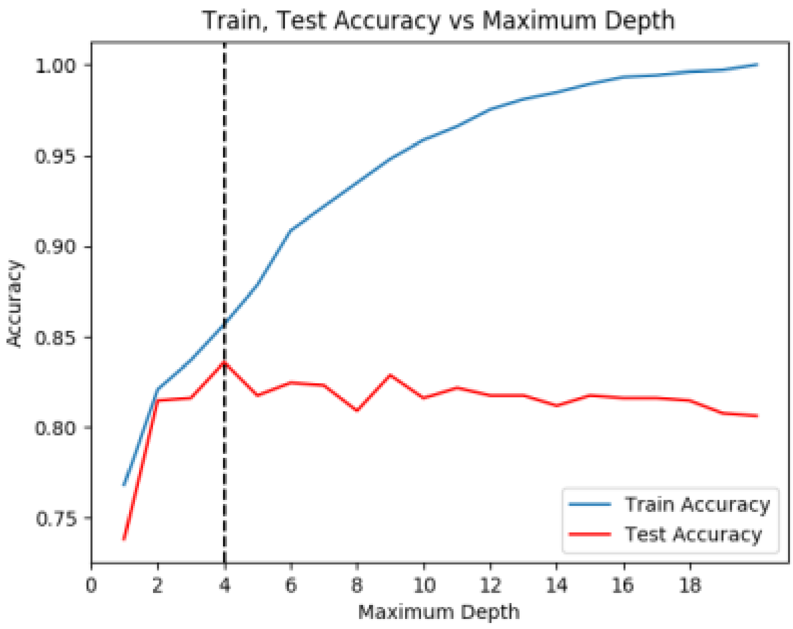


Fig. 1: Accuracy vs Maximum depth for the decision trees implemented by my algorithm.

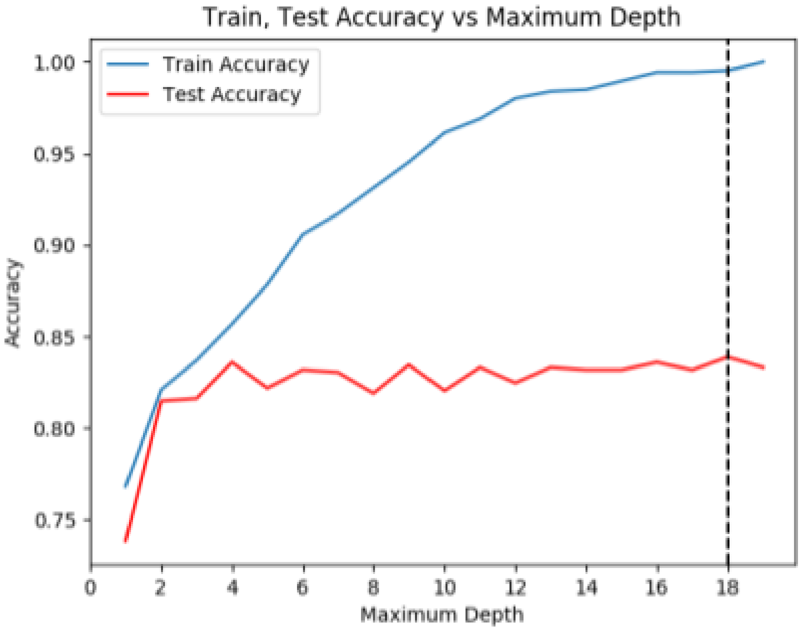
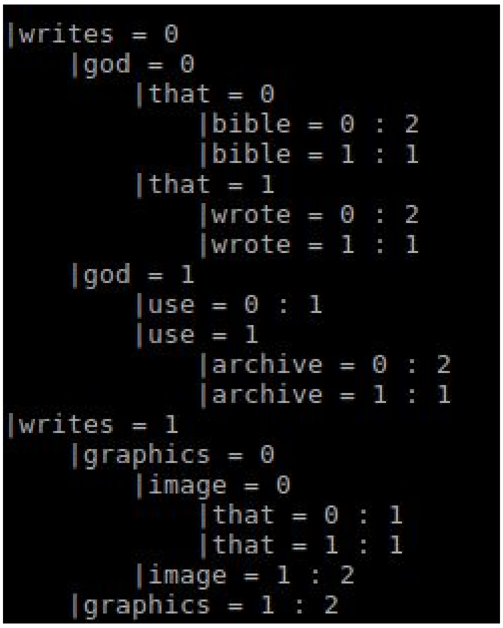


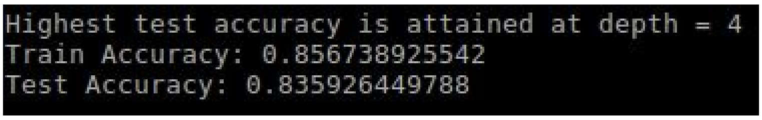
Fig. 2: Accuracy vs Maximum Depth for the decision trees implemented in scikit-learn

As can be clearly seen in Fig. 1, the maximum test accuracy occurs for the decision tree at a depth of 4 (black dotted line). The tree structure is as follows:



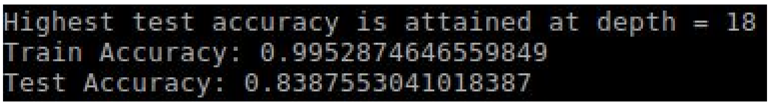
```
|writes = 0
    |god = 0
        |that = 0
            |bible = 0 : 2
            |bible = 1 : 1
        |that = 1
            |wrote = 0 : 2
            |wrote = 1 : 1
    |god = 1
        |use = 0 : 1
        |use = 1
            |archive = 0 : 2
            |archive = 1 : 1
|writes = 1
    |graphics = 0
        |image = 0
            |that = 0 : 1
            |that = 1 : 1
        |image = 1 : 2
    |graphics = 1 : 2
```

Train and test accuracy for the above tree is:



```
Highest test accuracy is attained at depth = 4
Train Accuracy: 0.856738925542
Test Accuracy: 0.835926449788
```

However, for the decision tree that was built by scikit-learn, test accuracy was highest when the tree attained a depth of 18. Train and test accuracy for the tree of depth=18 is:



```
Highest test accuracy is attained at depth = 18
Train Accuracy: 0.9952874646559849
Test Accuracy: 0.8387553041018387
```

## Does overfitting occur?

As can be clearly seen in Fig. 1, **test accuracy is highest at the depth of 4**. Hence overfitting occurs from depth 5 onwards. However, this does not give conclusive evidence of overfitting as the test accuracy keeps on fluctuating and reaches local maxima at a depth of 8. However, test accuracy of the trees having a maximum depth of 9 onwards decreases and after a depth of around 15 is reached, there can be seen a **monotonous decrease of test accuracy**. The decision tree grows to its full depth at max_depth=20, where train accuracy becomes 1.0. Thus, we can conclusively say that **overfitting starts occurring slightly after a depth of 9, and after a depth of 15, overfitting is evident.**

For the scikit-learn implementation, we see that the test accuracy is highest at a depth of 18. Thus overfitting occurs from a depth of 19 onwards.

## A brief discussion on word features selected by the decision tree that achieved the highest testing accuracy.

From the tree structure shown above, we see that the decision tree has used the word 'writes' as the first splitting attribute and then there is a use of 'god', 'bible', 'graphics', 'image' etc. The two classes that we had to classify were **alt.atheism** denoted by class 1 and **comp.graphics** denoted by class 2. Hence, the words that can act as a distinguishing feature among the two classes can indeed be 'god', 'bible', 'image' and 'graphics' as has been used by the decision tree.

A document having the words 'god' and 'bible' are more likely to be in class 1(speaks about atheism), whilst a document having the words 'graphics' and 'image' are more likely to belong to the 2nd class (speaks about something related to graphics). Hence, as rightly predicted by the algorithm, if the document has a word **'god'** or if there is no word 'god' present, then if it has the word **'bible'**, the document belongs to **class 1,** if none present, then belongs to class 2, that is may not be related to atheism. However on the other hand, if there is the words **'graphics'** or **'image'**, the document likely belongs to **class 2**, else class 1.

Not all the word features selected by the algorithm makes sense, for example, the words ‘that’, ‘use’, ‘wrote’, ‘archive’ does not give any conclusive evidence about the class of the document. Hence, the algorithm choosing these words as a classifying feature between atheism news and graphics news does not make sense. However, according to the data that reached the corresponding nodes, using these features resulted in having the least GINI INDEX or the maximum INFORMATION GAIN. Thus these features were selected. Hence, it all depends on the quality of training data that was used and the data that reached the particular nodes.

---