Machine Learning Assignment 2 - Part II

Sayan Sinha
16CS10048

INTRODUCTION

A decision tree is a decision support tool that uses a tree-like model of conditional statements and their possible consequences, such that the conditions match certain criteria of the concerned data. It is one way to display an algorithm that only contains conditional control statements.

MODEL

A dataset is already provided which is divided into train and test sets. The dataset are presented as comma separated values is saved in the root directory.

A decision tree classifier is trained using Information Gain scores. Information gain is given by:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

where entropy is defined as the minimum number of bits required to represent a set of data, i.e.

$$Entropy(t) = -\sum_{j} p(j \mid t) \log_2 p(j \mid t)$$

Higher the information gain, better classification results are obtained by choosing that feature as the next node of the tree.

Scikit-learn is an open source machine learning library for the Python programming language, which was born out of a GSoC project in 2017. It supports regression, clustering and classification based machine learning algorithms.

The model works with a recursive approach, in which the tree creator is a called by the possible child nodes again and again.

The dataset is read as a Python dataframe and is converted to Numpy arrays as and when required.

The given figures highlight the fact that the model behaves almost exactly the same way as Scikit learn does, except for the speed factor. The accuracy values measured were almost the same.

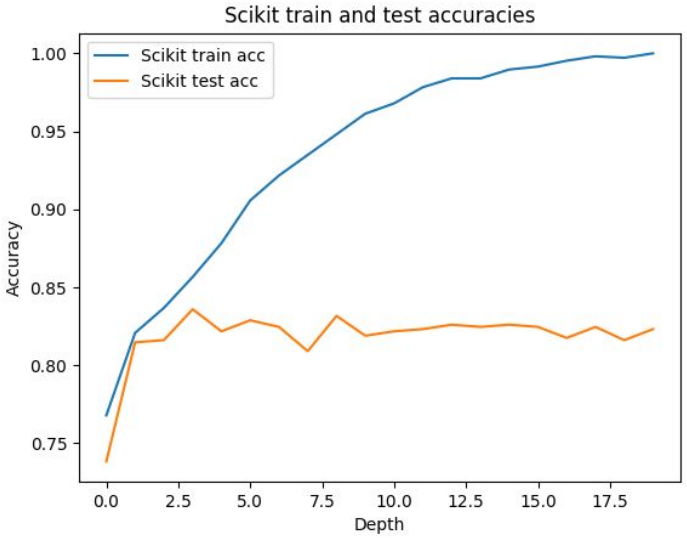

Fig 1

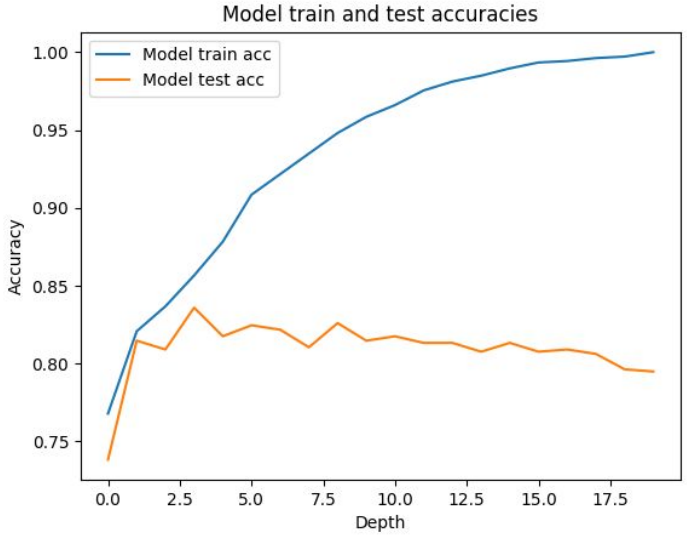

Fig 2

Depth	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**10**	**20**
Train acc	0.77	0.82	0.84	0.86	0.88	0.91	0.92	0.96	1.00
Test acc	0.74	0.81	0.81	0.84	0.84	0.82	0.82	0.81	0.79

Table 1: Accuracies of the model

The highest test accuracy was observed at depth 5 and hence, overfitting occurs beyond depth 5. We see that at depth 20, it completely fits the training set, but the accuracy on the test set decreases by a significant amount.

The tree is:

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

Most of the word features selected here make sense in a way that they are related to God and religion. But we also find some aspects of computing such as “program”, “image” and “graphics” giving the same category as that of sentences containing “atheist”. This might mean, that people more interested in science and technology have less faith in religion. Thus, there is a fair amount of sense in the entire selection of words.

CONCLUSION

We conclude that Decision Tree classifier is indeed a suitable method for making predictions on data of the given nature. The entire program (all tree generation and scikit exploration upto depth 20) takes around 124s to run whereas running just the Scikit part of the program happens in around 15s. Therefore, the current model imposes huge overheads.

REFERENCES

- Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics 21.3 (1991): 660-674.
- Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.