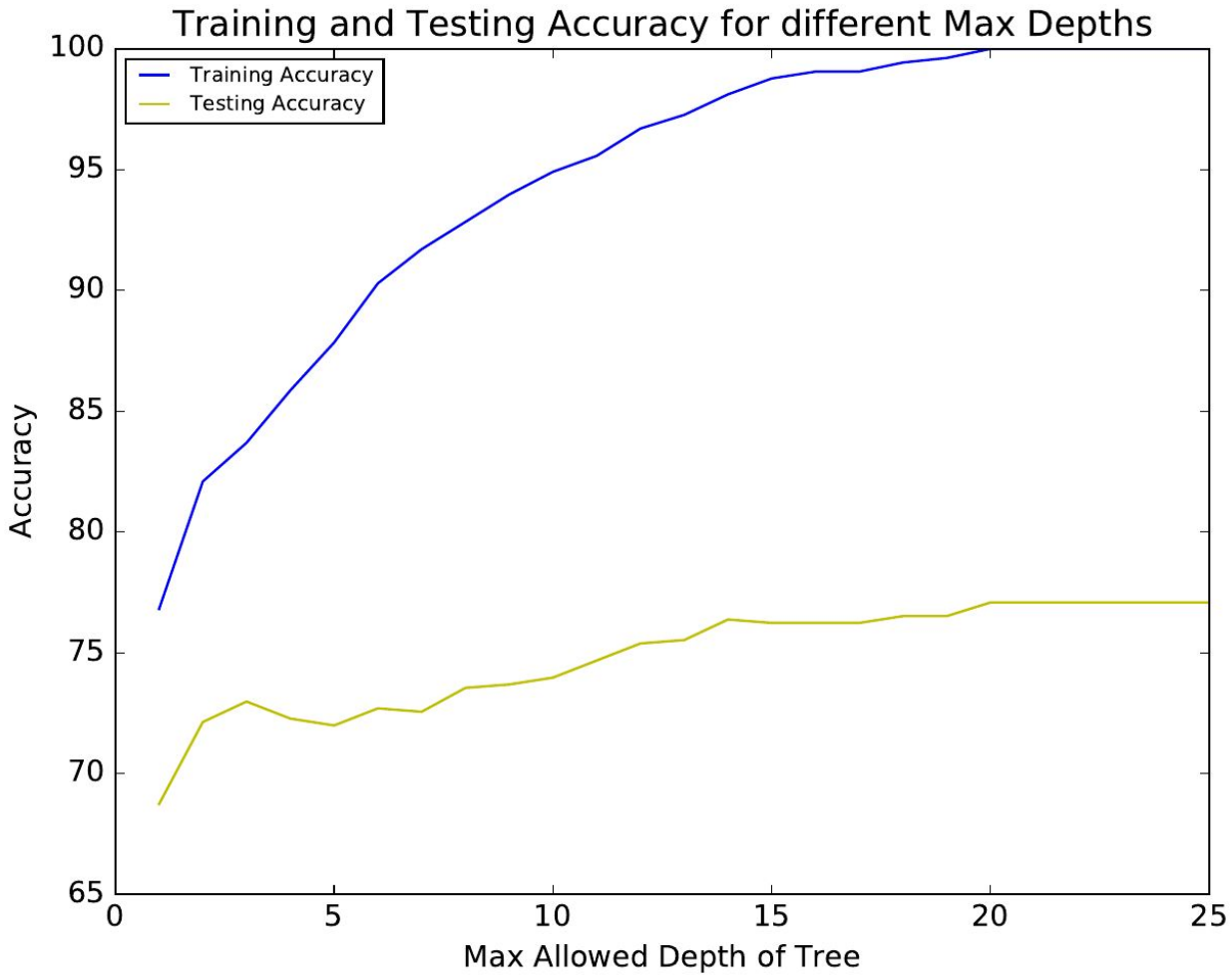
Assignment Report

Machine Learning (CS60050) Assignment 2

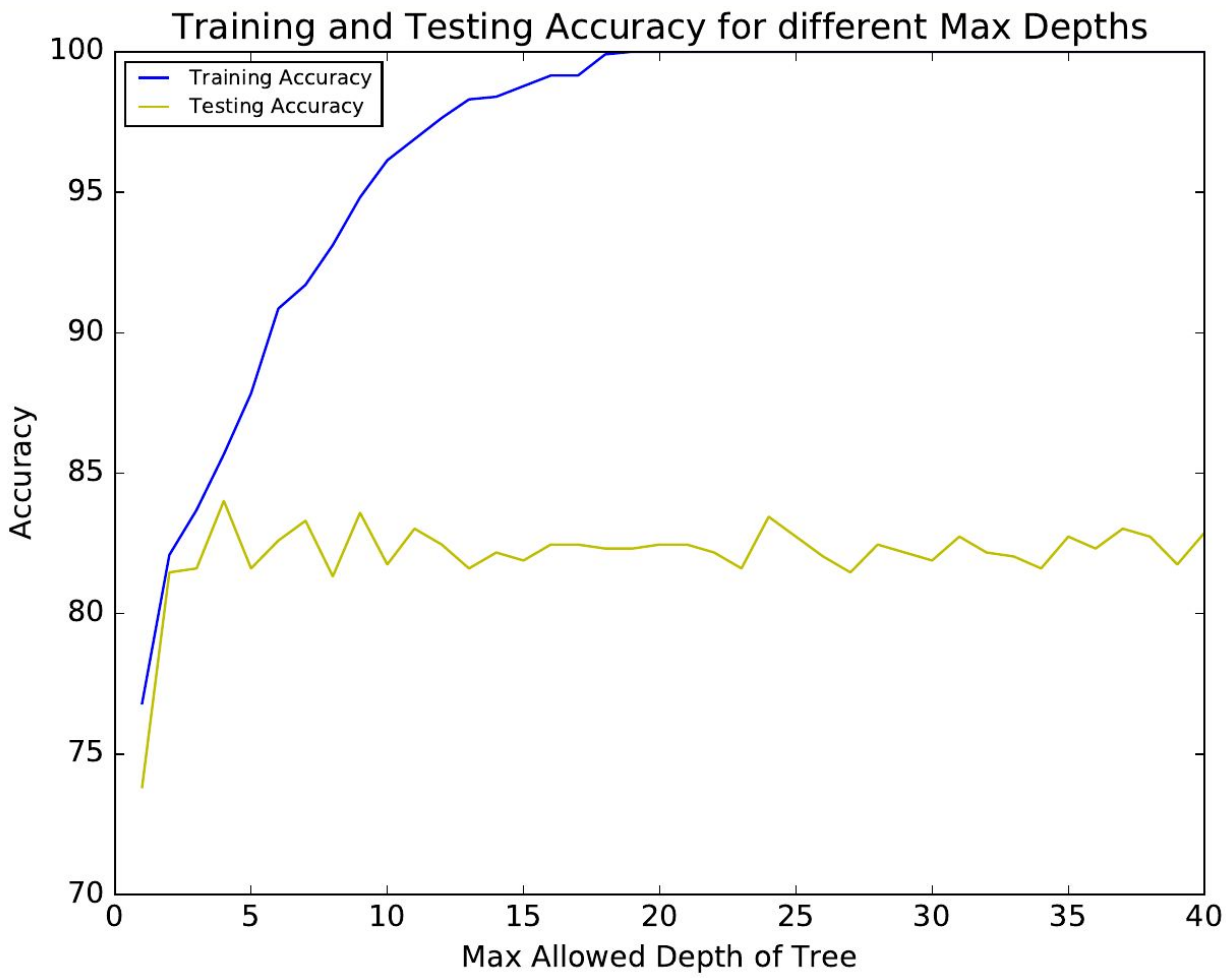
Arijit Kar

15-03-2019
16CS30005

(1.a) Plot of accuracy vs max. allowed depth (My Model):



(1.b) Plot of accuracy vs max. allowed depth (scikit-learn):



(2) Overfitting:

Yes, Overfitting occurs in the decision tree as can from the fact that training accuracy increases to almost 100% while there is little to no increase in the testing accuracy.

Overfitting occurs around the depth of 20 where the model complete fits the training data and the training accuracy becomes 100%.

(3) Selection of word features:

Some of the word features in the Decision Tree with the highest testing accuracy are as follows:

(a) graphics
(b) image
(c) comp
(d) files
(e) format

It can easily seen that these word features were selected by the model as they relate to the target class ‘comp.graphics’ and there were samples in the training set that had these features and belonged to the class ‘comp.graphics’.

Similarly for the target class ‘alt.atheism’ the selection of the following word features which are makes sense since the presence of these words in an article can mean that the article belongs to this class:

(a) god
(b) keith
(c) christian
(d) atheist
(e) bible

But there are also certain random words that were selected by the model as features which do not exactly make sense. They are the following:

(a) that
(b) you
(c) have
(d) more
(e) be

The selection of these word does not make sense.