Assignment 3 COL-774

Abhishek Agarwal - 2014MCS2114

April 13, 2016

1 Question 1

1.1 Part a

By looking at the graph for part a, it is observed that the training set accuracy increases sharply while the test accuracy does not increase that sharply. Also the validation set accuracy is almost always less than the train set accuracy. Thus the classifier is over-fitted a bit here.

Train Set Accuracy Final : 89.36
Validation Set Accuracy Final : 82.4
Test Set Accuracy Final : 75.74

1.2 Part b

By looking at the graph for part b, it is observed that validation set accuracy remains higher than the train set accuracy almost over all the graph and becomes at par with it in the end. Also, not much difference is observed between test and train accuracy (test accuracy decreasing a bit in the end). Thus the classifier is less over-fitted as compared to in part a, which is expected on pruning.

Train Set Accuracy Final : 88.75
Validation Set Accuracy Final : 87.2
Test Set Accuracy Final : 75.74

1.3 Part c

By looking at the graph for part c, it is observed that the train set accuracy is mostly always vastly greater than the test set and validation set accuracy. It is due to the fact that the number of nodes have increased and the classifier learns a bit better as compared to part a. Also, the final validation set and test set accuracy are less as compared to part a. Thus we can say the classifier over-fits here.

Age is repeated 5 times in a branch and is maximum over all such repeating attributes. All these attributes are printed/logged in the code for partC method.

Train Set Accuracy Final : 99.19
Validation Set Accuracy Final : 72.8
Test Set Accuracy Final : 75.74

1.4 Part d

The configuration for which the DecisionTreeClassifier works well in all trial and error settings :

criterion entropy : gini could also have been chosen but changing the other parameters along with gini never gave better results than this setting
maxdepth 6 : decreasing the depth less than 6 tended to under-fit and anything else tended to over-fit in all the trials
minsamplesleaf 3: less than 3 over-fits here and more than 3 under-fits in this setting. A very large value will under-fit
splitter best: other splitter(random) was under-fiiting too much
minsamplessplit 2: if only one sample comes to a node there is no need to split, and more than 2 tended to under-fit. A very large value will under-fit

Ac-curacies obtained for these settings:
Training Accuracy: 90.361
Validation Accuracy: 79.2
Test accuracy: 80.59

2 Question 2

2.1 Part a

By performing 5-fold cross validation(training on each possible combination of 4 splits and the testing on the remaining one), the average test accuracy is : 94.88/100.

2.2 Part b

The accuracy that you would obtain by randomly guessing one of the newsgroups as the target class for each of the articles: 12.46/100
The above result makes sense as randomly guessing out of k=8 target labels will result in approximately 1/8 test accuracy. Our original algorithm as in part 'a' is better by: 82.42/100

2.3 Part c

By cross-posting, we assume an article belonging to category A has also been explicitly posted to category B. This is basically adding erroneous training data to the classifier. This can create problems for the classifier as this will increase confusion between the involved categories.
But, since only 4 percent articles have been cross-posted, so naive bayes would

not actually get affected a lot.

2.4 Part d

The plot has been attached with the submission
The train accuracy is almost always constant and gets plateaued after sufficient examples.
The test accuracy, as expected, increases with the number of training examples.

2.5 Part e

Excel is attached for the confusion matrix according to part a.
Highest diagonal entry: rec.sport.hockey (972)
Highest entry amongst the non-diagonal entries in the confusion matrix: talk.politics.misc is most time wrongly reported as talk.politics.guns (69)
rec.sport.hockey must have some very unique words which doesn't co-occur in other categories, which is why it is the least confused
talk.politics.guns and talk.politics.misc must have many co-occurring words together. Also misc is a broad category.