CSL603-Machine Learning Lab1

Vivek Kumar Verma
2016csb1064

Experiment 1.

- 1000 samples/instances each for Train and Test are randomly selected from labeledBow.feat from respective files.
- 500 are positive instance and other 500 are negative instance.
- After selecting random instance I have saved them in MyTrainData.txt and MyTestData.txt file respecting each containing 1000 instance
- Then I selected features based on expectation value positive polarity >2.2 and negative polarity < -1.2
- Then I selected 5000 randomly from these features with 2500 positive polarity and 2500 negative polarity.
- These features are finally saved in MyVocab.txt using their index of actual vocabulary provided.
- To run experiment1 I have created file named generate.py . This file is executed before running any other file.

Experiment 2

- I used ID3 algorithm to train decision tree.

Original Tree without early stopping

Training Accuracy	92.5%
Test Accuracy	70.1%
Nodes Count	895

Feature Index in Vocabulary	Frequency
3485	3
3533	3
344	4
868	4
427	4
439	6
734	6

(There are many other Features which you will see in the output)

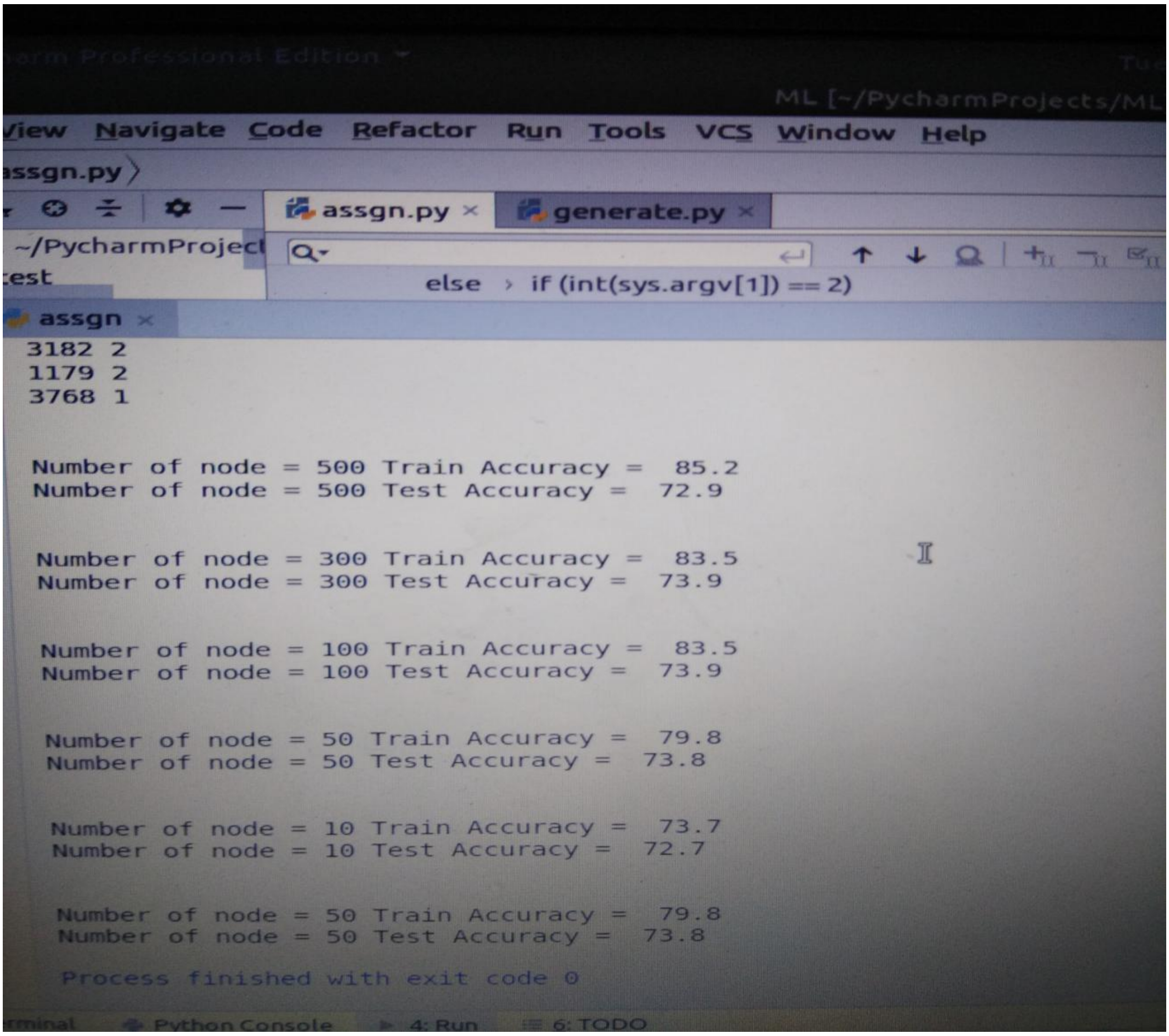
Statistics of early stopping.

I stopped the tree on basis of number of leaf nodes.

Node Restrict Count	Train Accuracy%	Test Accuracy%
500	85.2	72.9
300	83.5	73.9
100	81.5	73.9
50	79.8	73.8

10	73.7	72.7

- It was observed that Training accuracy was decreasing on restricting the node in tree.
- Test Accuracy increased. Then on further decreasing the node it gradually fall.
- Conclusion: This is observed because of reduction in over fitting. On further decreasing the node accuracy decreases

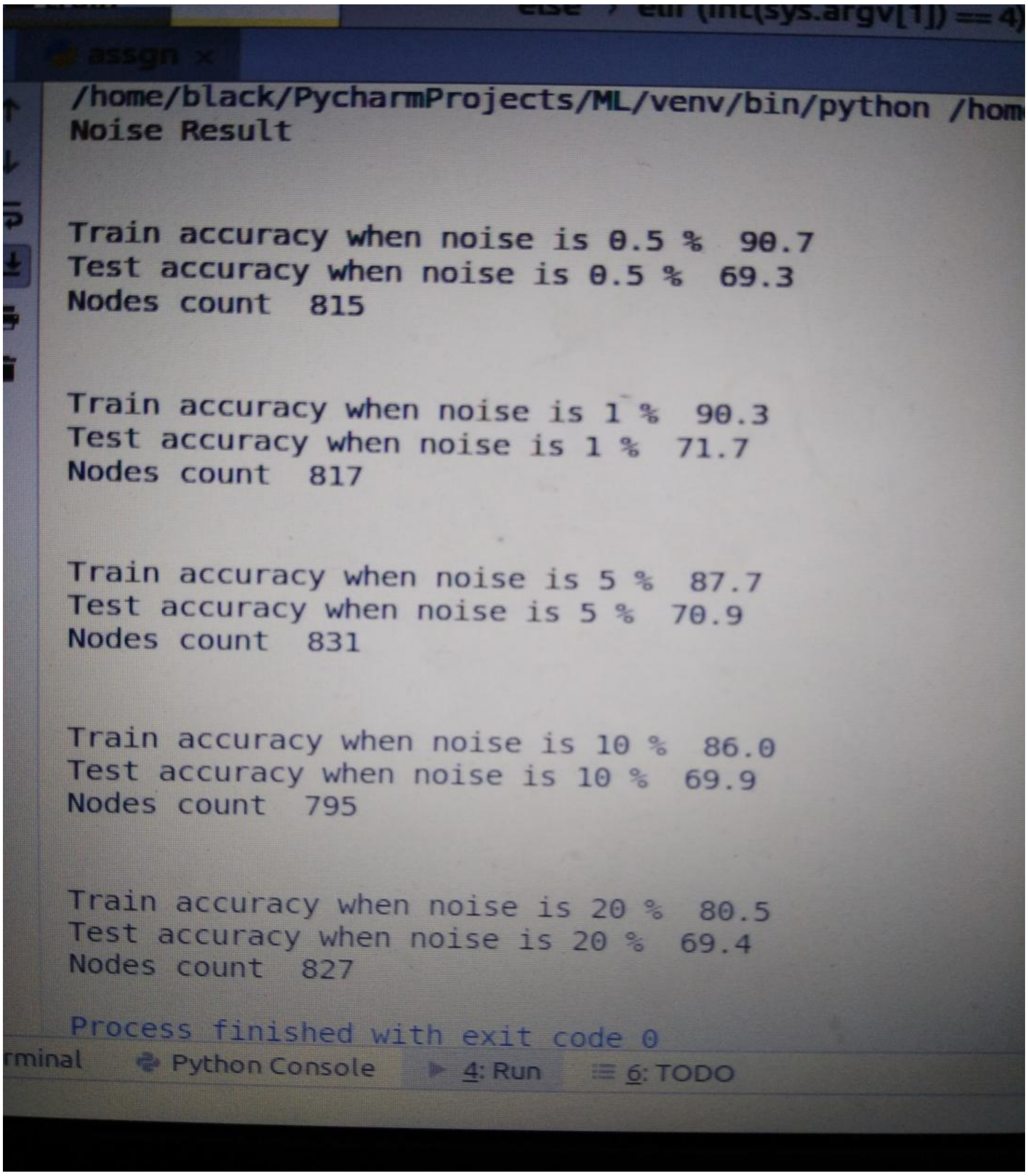

arm Professional Edition
ML [~/PycharmProjects/ML
View Navigate Code Refactor Run Tools VCS Window Help
ssgn.py
assgn.py ×
generate.py ×
~/PycharmProject
est
else › if (int(sys.argv[1]) == 2)
assgn ×
Terminal Python Console 4: Run 6: TODO

```
3182  2
1179  2
3768  1


Number of node = 500 Train Accuracy =  85.2
Number of node = 500 Test Accuracy =  72.9


Number of node = 300 Train Accuracy =  83.5
Number of node = 300 Test Accuracy =  73.9


Number of node = 100 Train Accuracy =  83.5
Number of node = 100 Test Accuracy =  73.9


Number of node = 50 Train Accuracy =  79.8
Number of node = 50 Test Accuracy =  73.8


Number of node = 10 Train Accuracy =  73.7
Number of node = 10 Test Accuracy =  72.7


Number of node = 50 Train Accuracy =  79.8
Number of node = 50 Test Accuracy =  73.8

Process finished with exit code 0
```

Experiment 3

Effect of Noise on accuracy of Decision Tree

Noise Percentage	Train Accuracy%	Test Accuracy%	Node
0.5%	90.7	69.3	815
1%	90.3	71.7	817
5%	87.7	70.9	831
10%	86.0	69.9	795
20%	80.5	69.4	827



```
/home/black/PycharmProjects/ML/venv/bin/python /hom
Noise Result


Train accuracy when noise is 0.5 %  90.7
Test accuracy when noise is 0.5 %  69.3
Nodes count  815


Train accuracy when noise is 1 %  90.3
Test accuracy when noise is 1 %  71.7
Nodes count  817


Train accuracy when noise is 5 %  87.7
Test accuracy when noise is 5 %  70.9
Nodes count  831


Train accuracy when noise is 10 %  86.0
Test accuracy when noise is 10 %  69.9
Nodes count  795


Train accuracy when noise is 20 %  80.5
Test accuracy when noise is 20 %  69.4
Nodes count  827

Process finished with exit code 0
```

Observation:

- It was observed that on increasing the noise Training accuracy decreased rapidly and it reached 80% in case of 20% noise.
- Test accuracy slightly increased but not to large extent. Only few ups and downs were shown.
- Number of nodes increased as noise increased i.e height of tree increase.

Conclusion:

- Training accuracy decreased because of large disturbance in data it s clearly shown in image above.
- Test accuracy showed no general trend
- Number of nodes increased as noise increased.

Experiment 4

ID3 with post pruning

Accuracy without pruning on test 69.5% number of nodes 815

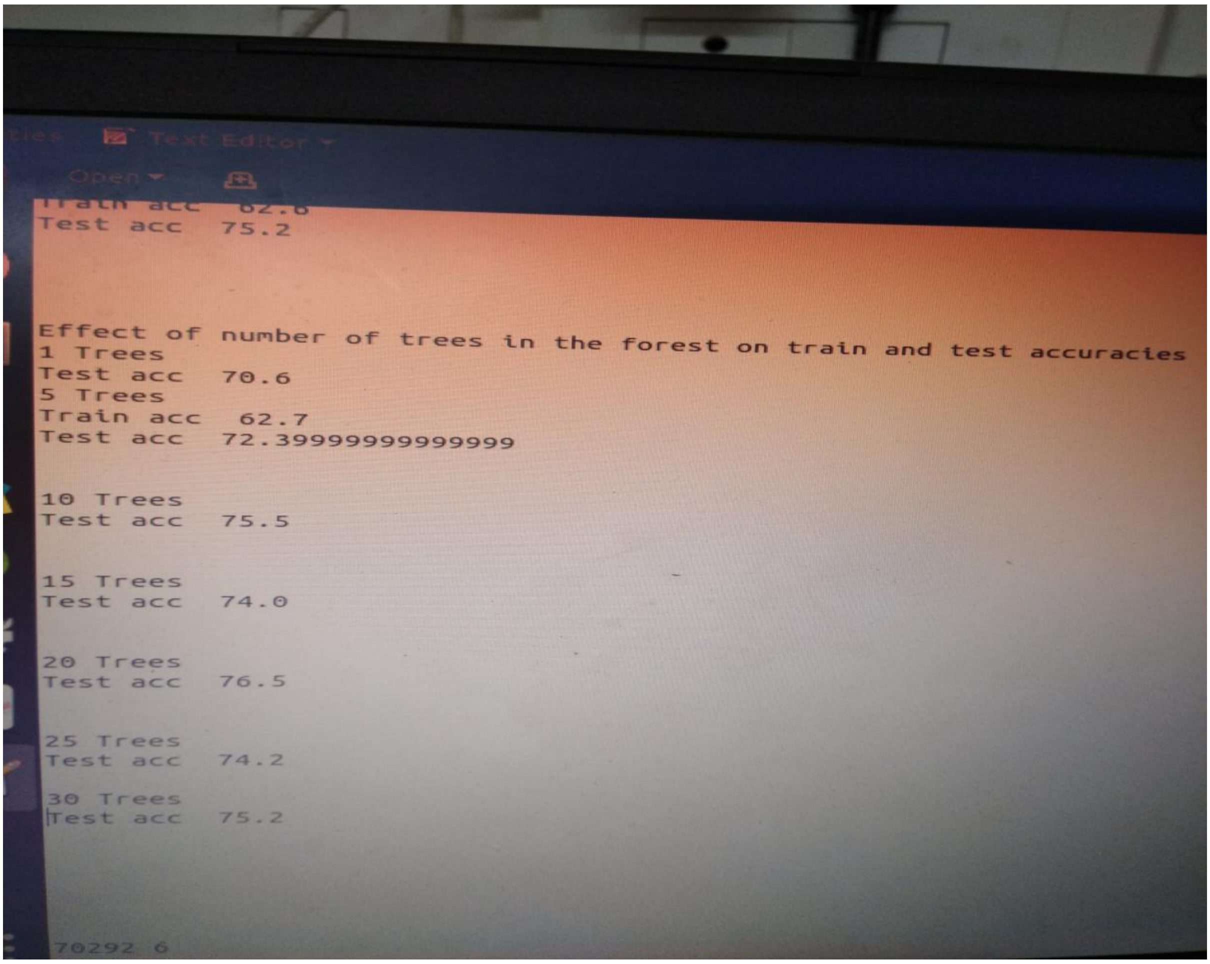
Accuracy when pruning on test 72.01% number of nodes 786

Experiment 5

Random Forest Using Feature Bagging

No of Trees	Accuracy on train
1	70.6
5	72.33333
10	75.5
15	74.0
20	76.5
30	75.2

Thus we can infer from this that accuracy increases with increase in number of trees and then become stable.



```
Train acc  62.6
Test acc   75.2


Effect of number of trees in the forest on train and test accuracies
1 Trees
Test acc   70.6
5 Trees
Train acc  62.7
Test acc   72.39999999999999

10 Trees
Test acc   75.5

15 Trees
Test acc   74.0

20 Trees
Test acc   76.5

25 Trees
Test acc   74.2

30 Trees
Test acc   75.2



70292 6
```