Lab-1 Report Chirag Khurana (2016CSB1037)

# Sentiment classification of movie reviews using decision trees and forests

## 1. Training Set & Attributes

A training set containing a random sample of 1000 (500 +ve & 500 -ve) observations has taken from Large Movie Review Dataset.

A vocabulary (attributes) set of top 5000 (2500+2500) words according to sentiment value has taken.

## 2. Statistics of the learned tree

Using the above training set, the ID3 algorithm is implemented to create a decision tree.

Attributes that are most frequently used as splitting function:

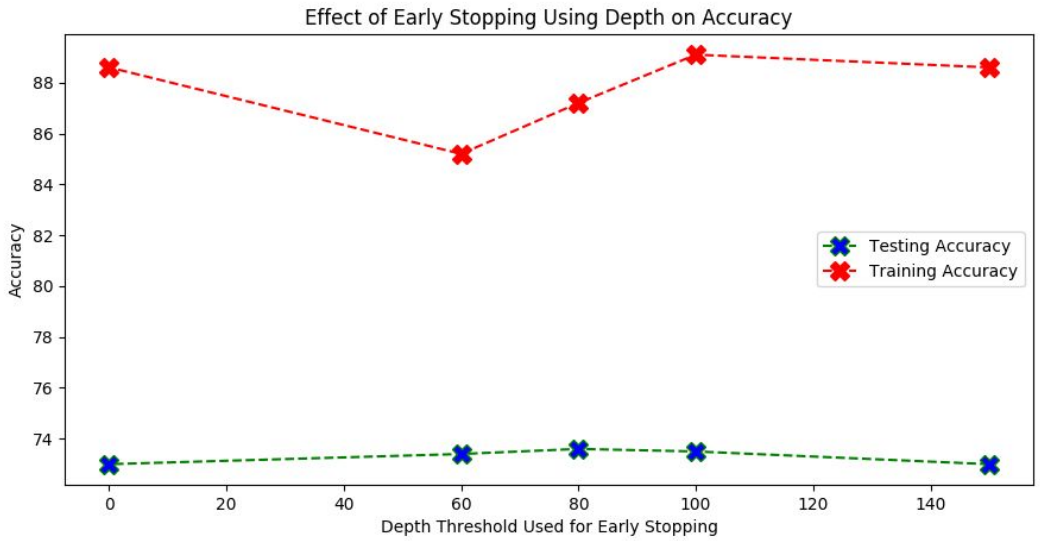
| Index | Word | Count |
|---|---|---|
| 1270 | laughable | 9 |
| 422 | worse | 7 |
| 1897 | unfunny | 6 |
| 1456 | zero | 6 |
| 760 | avoid | 5 |
| 344 | boring | 5 |
| 368 | awful | 5 |
| 503 | horrible | 5 |
| 1197 | garbage | 4 |
| 734 | dull | 4 |
| 1285 | excuse | 4 |
| 813 | lame | 4 |
| 906 | mess | 4 |
| 363 | stupid | 4 |
| 1014 | wasted | 4 |



* May change as Forest is random.

Effect of Early Stopping based on depth of the tree and minimum Info Gain = 0.001:

| Early Stopping Depth | Nodes | Terminal Nodes | Training Accuracy | Testing Accuracy |
|---|---|---|---|---|
| - | 849 | 425 | 88.6 | 73.0 |
| 150 | 637 | 319 | 88.6 | 73.0 |
| 100 | 579 | 290 | 89.1 | 73.5 |
| **80** | **535** | **268** | **87.2** | **73.6** |
| 60 | 495 | 248 | 85.2 | 73.4 |




Observation: Overfitting can be avoided by early stopping.

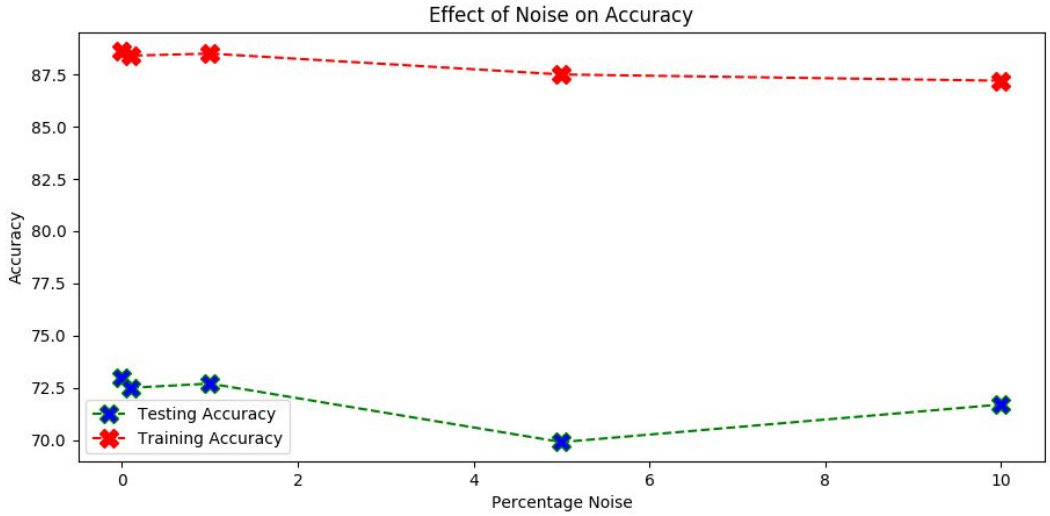


* May change as Forest is random.

# 3. Noisy Dataset

Noise added in the dataset by switching the labels of the instances.

| Noise % | 0 | 0.5 | 1 | 5 | 10 |
|---|---|---|---|---|---|
| Nodes | 849 | 845 | 841 | 859 | 869 |
| Terminal Nodes | 425 | 423 | 421 | 430 | 435 |
| Training Accuracy | 88.6 | 88.4 | 88.5 | 87.5 | 87.2 |
| Testing Accuracy | 73.0 | 72.5 | 72.7 | 69.9 | 71.7 |




Observation: Little-bit noise not affects the accuracy much but large noise will lead to affect the accuracy significantly.



* May change as Forest is random.

# 4. Post-Pruning

A pruned tree has been produced by computing the prediction accuracy on the test set.

| | Normal Tree | Pruned Tree |
|---|---|---|
| Training Accuracy | 88.6 | **82.4** |
| Testing Accuracy | 73.0 | **78.8** |
| Nodes | 849 | **197** |

Observation: Overfitting can be avoided by pruning. For my dataset, pruning came out best method than others and also pruning don't takes much time like random forest.

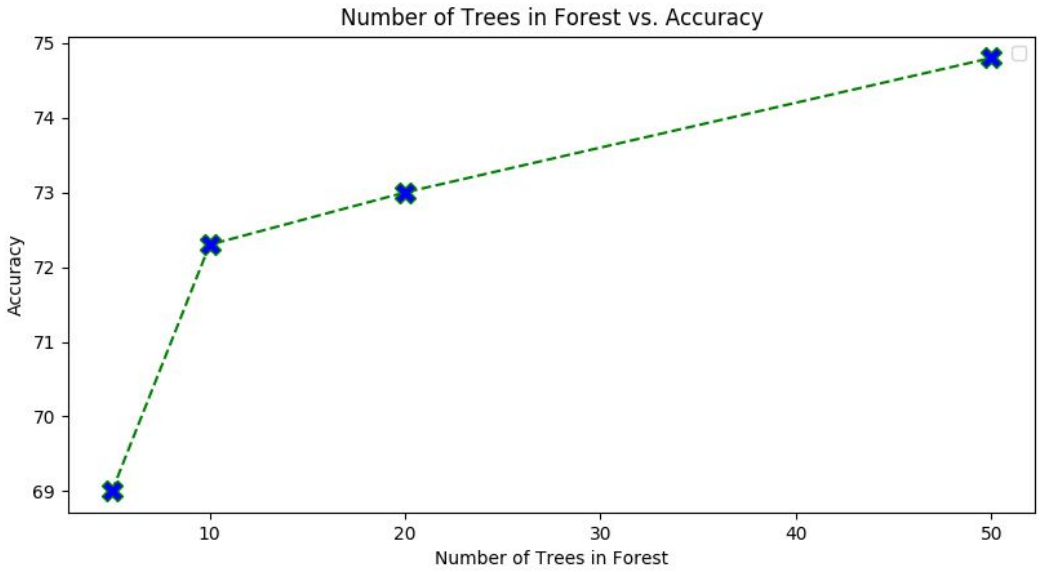


* May change as Forest is random.

# 5. Random Forest (Feature-Bagging)

The effect of number of trees in the forest on the prediction accuracy of the test data set:

| No. of Trees | Testing Accuracy* |
|---|---|
| 5 | 69.0 |
| 10 | 72.3 |
| 20 | 73.0 |
| **50** | **74.8** |




Observation: Accuracy increases as number of trees increases in random forest.



* May change as Forest is random.