CSL603

Decision Tree

2016CSB1050 - Assignment 1

Sizes of Data Sets Used:

1) Words Dictionary: Based on the sentiment value of 89000 English words , I have used top 3000 negative sentiment and top 3000 positive sentiment.
2) Number of instances in training/test set: A random subset of 1000 examples has been used to train/test the Decision tree

Quantisation:

1) The rating value <=4 has been treated as -1 and >=7 has been treated as +1
2) The frequency of occurrence of a particular word >0 has been treated as 1 (rest as 0)

Note: The reason to why train accuracy is not 100% is because of the creation of a subset of dictionary words. Since we are choosing 6000 out of 89000, there exist some reviews (≈ 300 out of 1000) whose none of the words fall into the chosen 6000 and hence they possess an empty feature vector. So no matter what their labels are they all go to the same leaf node.

Statistics of the Decision Tree:
Without any optimisations: The decision tree was built on different combinations of data and following results were observed. The values for similar input but random instances almost remains same with a variation of ±2% in accuracy.

Average Train Accuracy for the model = 92%
Average Test Accuracy for the model = 73%

Most often used Words:

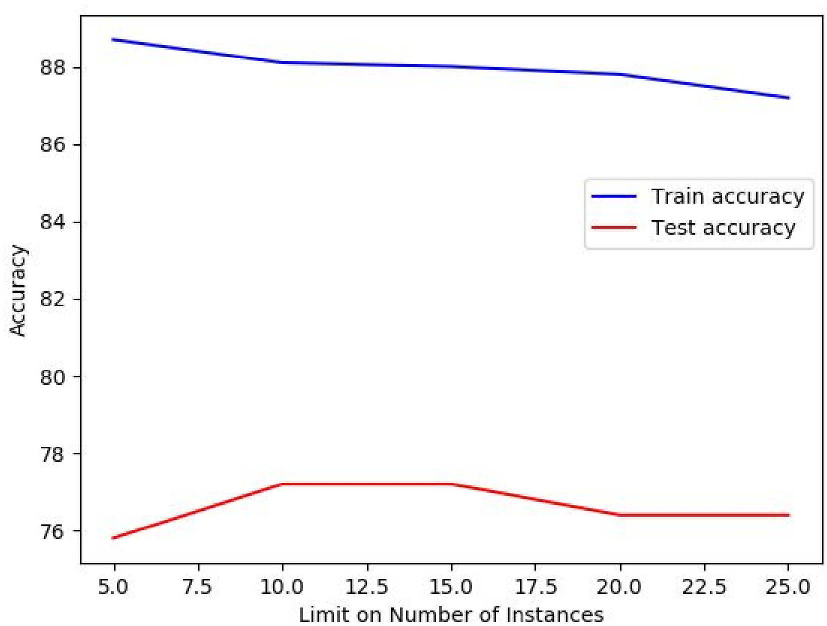
1. Boring (12)
2. Annoying (8)
3. Supposed (7)
4. Horrible (7)
5. Dumb (6)
6. Terrible (6)
7. Joke (5)
8. Worse (5)
9. Bother (5)
10. Avoid (5)
11. Cheap (5)
12. Porn (4)
13. Awful (4)
14. Stupid (4)
15. Excuse (4)
16. Bore (4)
17. Pointless (4)
18. Crap (4)
19. Poorly (4)
20. Wasted (3)

Experiment #2 : Early Stopping

While building the tree if the height of the current node reaches the allowed level, then instead of splitting the node, it is turned into a leaf and assigned a label. The model tends to be highly biased towards the training data and a number of leaves occur with just one instance indicating overfitting. I have tried three different ways of early stopping , restriction on the depth of the tree , number of instances to split and minimum threshold for information gain to cause the split.

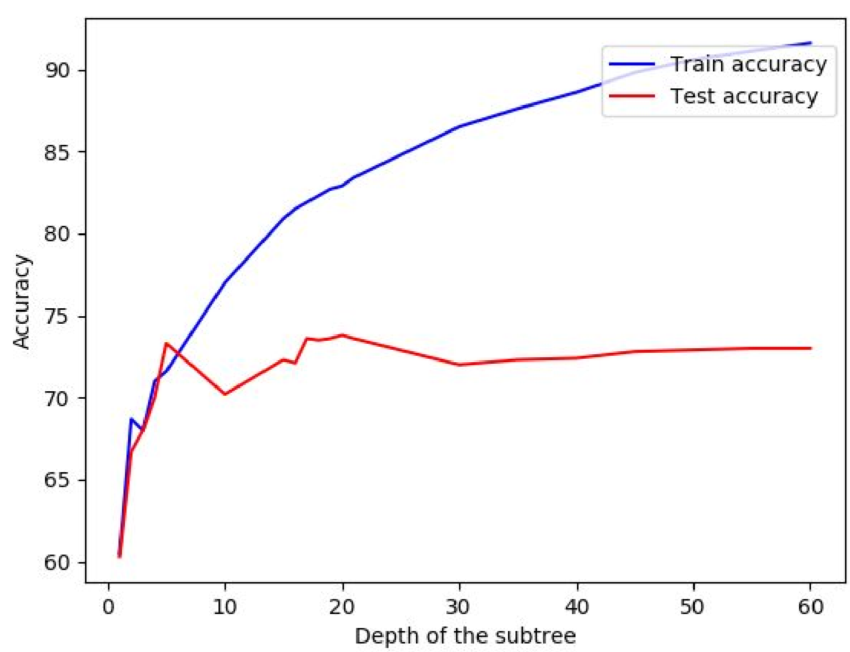
The results obtained using early stopping on number of instances have improved the accuracy considerably.

S. No.	Limit on Number of instances	Number of terminal nodes	Train Accuracy	Test Accuracy
1	5	247	88.7	75.8
2	10	237	88.1	77.2
3	15	231	88.0	77.2
4	20	223	87.8	76.4
5	25	205	87.2	76.4



Limiting the depth of the tree is showing a very little effect on the accuracy. Depth limitation of around 18-20 increases the accuracy by 0.2-0.4%

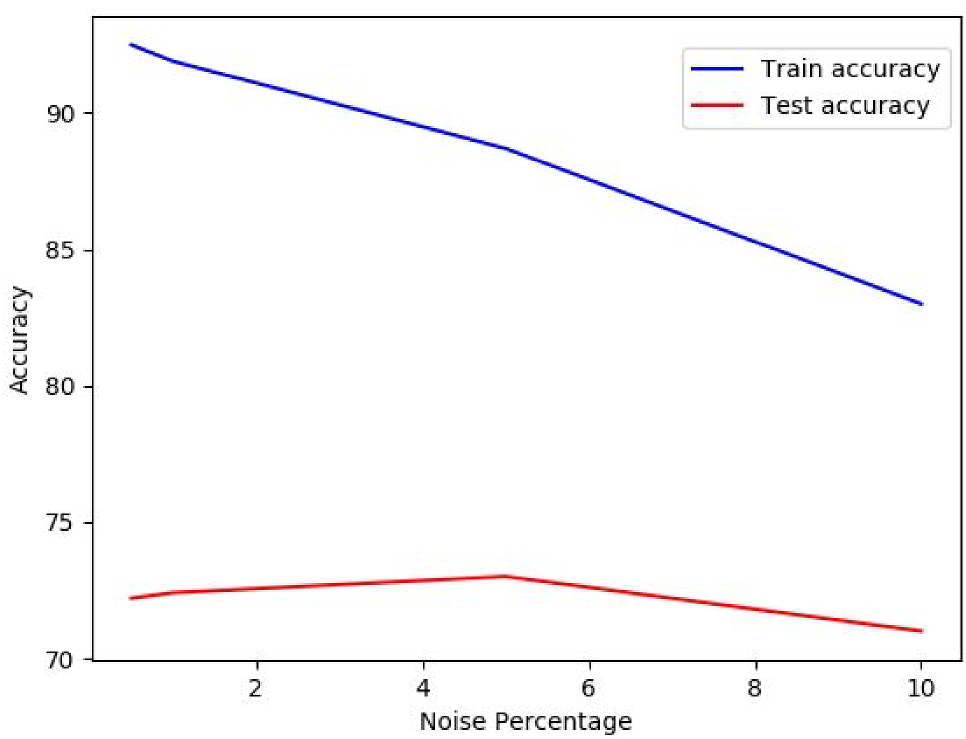
S. No.	Allowed depth of the decision tree	Number of terminal nodes	Train Accuracy	Test Accuracy
1	5	20	71.6	73.3
2	10	61	77.0	70.2
3	15	111	80.9	72.3
4	16	121	81.5	72.1
5	17	131	81.9	73.6
6	18	137	82.3	73.5
7	19	143	82.7	73.6
8	20	147	82.9	73.8



Experiment #3: Add random noise

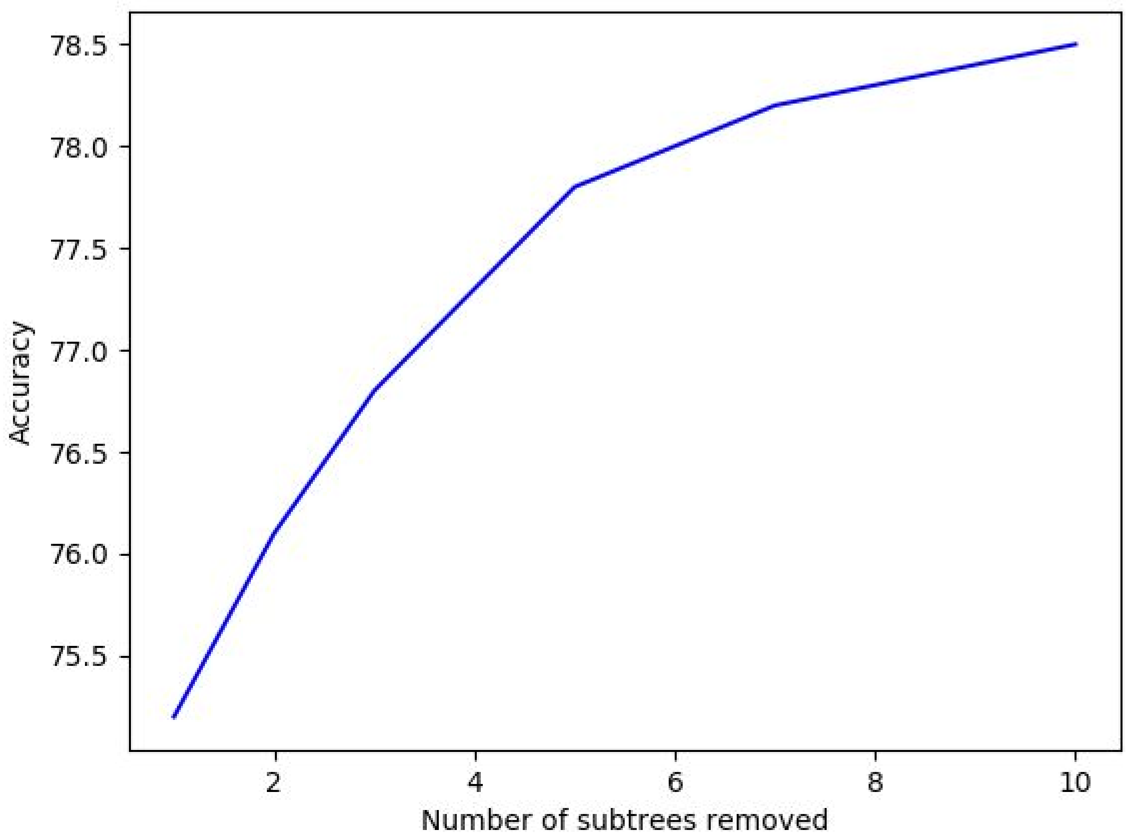
On changing the labels of random instances of train data the accuracy of the model decreases. Following are the statistics for addition of different percentages of noise and their effect on the accuracy. There is no definite trend in the accuracies observed.

S. No.	Noise percentage	Number of terminal nodes	Train Accuracy	Test Accuracy
1	0.5%	374	92.5	72.2
2	1%	379	91.9	72.4
3	5%	381	88.7	73.0
4	10%	396	83.0	71.0



Experiment #4: Pruning

Since the tree is prone to overfitting, removing the subtree such that accuracy increases proves highly effective for the decision tree. The following graph shows the variation of test accuracy with the number of subtrees being removed. Pruning considerably improved the results returning the smallest version of most the accurate tree with an accuracy of 79%.



Experiment #5: Decision Forest

Following are the statistics for 2 runs of the decision forest. Since the selection of attributes is random the results may slightly vary. Theoretically we use √D attributes for each tree of decision forest but in our case √6000 = 78 , Selecting 78 attributes randomly from 6000 words which are in turn selected from 89000 words, I observed a large number of cases ≈90% whose feature vectors are all zero i.e. none of the words present in the instance are there in the dictionary. Therefore I have used D/2 attributes for creating the decision forest and achieved considerably good results.

The result of decision forest is varied with no particular trends observed.

S. No.	Number of trees in the forest	Test Accuracy
1	5	73.4
2	10	75.1
3	15	73.1
4	20	72.2

S. No.	Number of trees in the forest	Test Accuracy
1	5	71.7
2	10	71.3
3	15	72.6
4	20	71.1