CSL 603 – Machine Learning

Lab 1

Decision Trees and Forests

Jatin Goyal

2015csb1014

1 . Introduction

The aim of this lab is to predict the sentiment of a movie review by implementing the **ID3 Decision Tree Algorithm** . To implement this algorithm we used the Large Movie Review Dataset from the Standford University. For this , we made decision tree using some movie reviews as the training set of the algorithm whose sentiments were already known to us. After making the decision tree , we used the tree to predict the sentiments of the movie reviews which were not known to us. After training the tree we checked the accuracy of tree using test dataset whose sentiments were known to us. Then we came across some of the techniques for increasing the accuracy on the test dataset.

2. Pre – Processing

First of all to implement the algorithm we created a file having 1000 movie reviews as our training dataset . We chose the these 1000 reviews randomly , in which 500 reviews have positive sentiments and 500 have negative sentiments . Following these reviews , in the same file we stored 1000 reviews as our validation set having equal number of positive and negative sentiments and following this, we stored 1000 reviews as our test dataset having equal number of positive and negative sentiments. This file is used for all the further experiments.

3. Experiment – 1

The experiment is basically based on finding the accuracy of the formed decision tree on the training set , validation set and test set. After this , we observed that what will be the change in accuracy , number of nodes etc. on changing the different stopping criteria on the decision tree. In this, we did pre-pruning on the decision tree and then observe the results.

The formed decision tree (sampled on training set) shows the various results as follows.

1. Decision Tree Prediction accuracy on training set - 100%
2. Decision Tree Prediction accuracy on validation set - 59.2%
3. Decision Tree Prediction accuracy on test set - 60.4%
4. Number of Nodes in the Decision Tree - 1889
5. Number of Leaf Nodes in the Decision Tree - 945
6. Height of the Decision Tree - 51

Observation

All the various results are shown above. As it is mentioned above that the decision tree was giving 100% accuracy on training set, it is because the tree is sampled on the training set so each review will go to its correct leaf node when traversed from the root of the tree. However on the validation set and the test set the accuracy is around 60% . It is because on the different (new) reviews other than training set , all reviews would not go to their correct leaf node and hence the tree would not show the 100% accuracy. As Hypothesis class comes out to be different than that of the Concept class.

Now let's see the effect of varying the tree height (max height of tree) on the different characteristics of the decision tree.

3 . (a) Effect of varying the Decision Tree Height

In the below table effect of varying tree height (max height of tree) on the accuracy on the test set , the number of nodes of the decision tree formed and the number of leaf nodes of the decision tree formed.

Observation

S.No.	Height of Tree	Accuracy (%)	Number of Nodes	Number of Leaf Nodes
1	51	60.4	1889	945
2	46	60.6	1861	931
3	41	60.9	1761	881
4	36	61.2	1597	799
5	31	61.6	1339	670
6	26	62.2	1077	539
7	21	62.1	797	399
8	16	63.0	519	260
9	11	63.4	297	149
10	05	64.2	79	40
11	01	62.9	3	2

Explanation

According to the table shown above , the statistics of the various characteristics changes with varying the height of the decision tree. On decreasing the height of the tree the accuracy of tree on test set increases. It is because of the fact that the internal nodes now becomes the leaf nodes of the tree . The reviews which passes from a particular node gets split on the nodes below that node. But after pre – pruning those particular nodes gets terminate and returns positive if majority of the reviews passing through that node is positive else returns negative. Since majority is the result so , when a review from test set is tested than it is more likely to give right output because of the majority. The number of nodes will decrease on decreasing the height of the tree as well as the number of leaf nodes will also decrease on decreasing the height of the tree.

4 . Experiment – 2

This experiment is basically based on quantifying the complexity of the learned decision tree. In this experiment , we added noise to the training set and then the tree was sampled on this new training set. We tested the tree on adding different proportion of noise in the training set . We added 0.5 % , 1% , 5% ,10% noise and observed the characteristics of the learned decision tree. Accuracy on the test set, Height of the tree etc.. To add the noise we switched the reviews with positive label to negative label and vice-versa .

Observation

S. No.	Added Noise(%)	Accuracy(%)	Height of Tree	Number of Nodes	Number of Leaf nodes
1	0	60.4	51	1889	945
2	0.5	60.3	65	1249	625
3	1	59.6	83	1329	665
4	5	56.4	88	1107	554
5	10	55.2	66	1441	721

Explanation

According to the table shown above , the statistics of the various characteristics changes with adding noise to the training set. On increasing the proportion of noise in the training set the complexity of the tree increases. The complexity of a binary tree is mainly based on its height. Because trees are not height balanced trees, so their height can goes upto the number of nodes in the tree. The time complexity for predicting the sentiment of a movie review is O(h) where h is the height of the tree . But on increasing the proportion of noise beyond a limit will decrease the tree height. It is because of the fact that on increasing noise beyond a limit the learned tree will be start giving more and more false results and the height of tree will start decreasing. So, if we add 100% noise to the training set, the height of the tree will be equal to that of adding 0% noise. The tree will then predict a positive review as negative and negative as positive.

5. Experiment – 3

This experiment is basically based on finding the accuracy(on test set) of the tree by using post – pruning strategy. In this first of all we find all the characteristics of the decision tree without implementing the post – pruning . Then to implement post-pruning, what we did basically is to remove all the nodes of a tree one by one and check whether on removing that particular node leads to increase in the accuracy of the tree or not. If removing that node increases the accuracy of the tree than we keep that tree , else we again add the removed node to its original position in the tree. In this way , we keep on traversing the tree (all nodes one by one) and eventually we will get the tree with the best accuracy among that.

Observation

Status	Accuracy(%)	Height of Tree	Number of nodes	Number of leaf nodes
Decision Tree **Before Post - Pruning**	60.4	51	1889	945
Decision Tree **After Post - Pruning**	65.2	50	981	491

Explanation

In the above table, the changes in the various characteristics of the decision tree by implementing the post – pruning strategy is shown. It is clear from the observation that on doing post – pruning on decision tree the accuracy of the tree increases. This is because when we sampled our data at training set, there may be some extra splitting which will indeed increase the accuracy on the training set but in actual some splitting may lead to give false results. So when we tested our data on the test set the accuracy decreases by drastic amount as compared to training set. But because of post pruning the extra splitting will not be there and the tree will be having only the needed splitting.

6 . Experiment – 4

This experiment is basically based on finding the accuracy(on test set) of the tree by using feature bagging in the tree. In this experiment first of all we find the accuracy of the decision tree without implementing feature bagging. Then to implement this strategy what we did basically is to make decision forest instead of decision tree. For making decision forest we made forest having trees around 50 – 110 . After making (let’s say) x number of trees, we check all the 1000 reviews on all the trees and considered result as the majority of this. Here for each tree we took 100 attributes and sampled the training set on these 100 attributes.

Observation

S.No.	Number of Trees in Forest	Number of Attributes per Tree	Accuracy(%) On test set	Accuracy(%) on training set
1	50	100	65.6	99.3
2	64	100	66.1	99.3
3	70	100	67.2	99.3
4	75	100	68.5	99.3
5	80	100	67.1	99.2
6	90	100	66.4	99.4
7	100	100	66.6	99.6
8	110	100	65.1	99.5

Explanation

In the above table , the change in the accuracy on the test dataset by using feature bagging is shown. It is clear from the observation table that , the accuracy on the test dataset increases on implementing feature bagging strategy. The reason behind increase of the accuracy is that there are many attributes which appears to be highly predictive in the decision tree (before using feature bagging) when sampled on the training set, but fails miserably when tested on the test set(outside training set). For this reason, feature bagging is used in which if any attribute acts highly predictive will only be limited to some trees and its effect will get diminished on increasing the number of trees. As we can see that on increasing the number of trees the accuracy on the test set increases. But increasing the trees beyond a limit will again starts decreasing the accuracy on the test set.