Performance Of Decision TreesOn Student Performance Data

Shubham Barudwale
School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: barudwaleshubham.dinesh2015@vit.ac.in

Abstract-

Decision tree is the simple algorithm which uses supervised learning to classify the dataset into parts. It is a linear classifier which classifies the given data points into groups by finding the importance of the attributes which affects results majorly.

The Decision tree is a simple if else tree which takes each data point and classify it based on the value of the data. At end it reaches to a leaf node based on which the class is assigned to the data point.

This algorithm is effective on the both linealy and non linearly seperable data. The system performs better than perceptron algorithm and gives about 94% accuracy.

1. Introduction

Machine learning is the process in which the algorithm will take the data input and will try to find out the decision boundary to classify the given input into appropriate classes.

In Decision tree we take the attributes which affect the result in major scence. i.e. weight for attribute is more in classification of the data. Then the decision tree divides all the data into two parts. One which follows the positive and one which does not follow the condition. Each time the clssified data is again classified until the calssification reaches the specified depth. After that the proper target values are assigned to the classified data.

The training data is also classified from the training tree. By tweaking with various adjustments we can get various results

2. Methadology

Decision trees are non parametric supervised learning method used for classification and regression. Decision tree creates the model to predict the decision by creating simple decision rules.

Decision tree takes attributes of the data and try to figure out the importance of each data field according to the target function. It classifies (or splits) the whole data based on such decision making trees. We can give the depth of the tree upto which the algorithm will consider various attributes of the data and will classify them upto n depth. More the depth means more will be the accuracy based on various attributes.

Some of advantages of the decision trees are, they are simple to understand and to interrupt. They can be visualized. Time required to train and prediction reduces exponentially. Able to handle any type of data. Validation is easy using statistics.

But in some cases where data is not simple overfitting can happen by complex trees. Small variation in data can give completely different tree. Could crete biased trees if some classes are dominant.

3. Dataset – Student Performance Dataset

Student performance dataset is the dataset from which we are trying to gain some information from 650 student's academic performance data and various attributes which are affecting the performance. Attributes include sex, age, study time, free time, family information etc.

We will be using the decision tree method to learn and predict the performance. Out of 650 I have taken 455 data entries for lerning and rest for prediction i.e testing the performance of the algorithm.(70:30). The labels or target functions are mentioned in the 'new' field which is added manually. bad(0-7), medium(8-14), good(15-20) .

4. Algorithm

Given training vectors $x_i \in R^n$, i=1,…, l and a label vector $y \in R^l$, a decision tree recursively partitions the space such that the samples with the same labels are grouped together.

Let the data at node m be represented by Q. For each candidate split $\Theta=(j,t_m)$ consisting

of a feature j and threshold t_m, partition the data into $Q_{left}(\Theta)$ and $Q_{right}(\Theta)$ subsets

$$Q_{left}(\Theta) = (x,y) \mid x_j <= t_m$$
$$Q_{right}(\Theta) = Q \setminus Q_{left}(\Theta)$$

The impurity at m is computed using an impurity function H(), the choice of which depends on the task being solved (classification or regression)

$$G(Q,\Theta)=(n_{left}/N_m)H(Q_{left}(\Theta))+ (n_{right}/N_m)H(Q_{right}(\Theta))$$

Select the parameters that minimises the impurity

$$\Theta^* = argmin_\Theta G(Q,\Theta)$$

Recurse of subsets $Q_{left}(\Theta^*)$ and $Q_{right}(\Theta^*)$ until the maximum allowable depth is reached, $N_m < min_{samples}$ or $N_m = 1$

5. Experiments

Initially we will take the three classes that is poor(0-7), medium(8-14) and good(15-20). Initially I set the testing data as 30% and depth of the tree as 3. After running the Decision tree algorithm I got the In sample accuracy as 96.7% and out of sample as 94.4%

After running several variations of the test data and depth of the tree we get following observations:

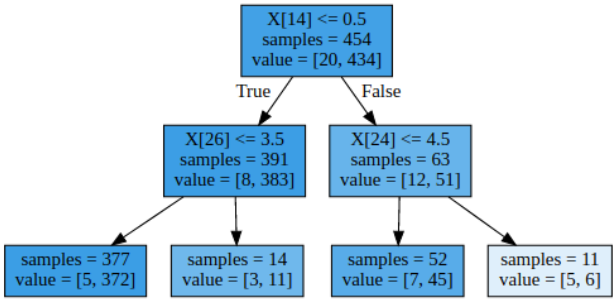
Depth ►	**3**	**5**	**7**	**10**
Testdata ▼				
0.1	0.962 0.954	0.983 0.954	0.991 0.908	0.997 0.908
0.3	**0.967** **0.944**	0.982 0.918	0.993 0.908	1.0 0.918
0.5	0.975 0.923	0.985 0.902	0.994 0.886	1.0 0.902
0.9	1.0 0.899	1.0 0.899	1.0 0.899	1.0 0.899

As we can see from above table that as we decrease the test size of training data the insample and out of sample acuracy increases but we can't test it on more dataset. If we increase the test size insample accuracy tends to one but out of sample accuracy decreases due to lack of training variations.

On other hand as we increases the depth of the tree out of sample accuracy decreases and insample accuracy increases which is of no use. This happens because of memorizing the data and complex trees.

So we get optimum solution at training data as 30% and depth 3. So mediacore depth and training data is best for learning the data.

The Decision tree was performed on the data using sklearn Decision tree which gives us the tree using predefined Decision tree algorithm in the sklearn.



6. Conclusion

We used Decision tree to seperate the Student performance data and classify them according to target function into classes (here taken poor and good only). We created the decision trees using various modifications on the number of testing and training dataset and depth. We got the maximum accuracy and optimum performance on the 70% traiing data and depth of decision tree as 3. We got 94.4% out of sample accuracy.

We also have seen the interrelation of various factors on the target functions. And effect of particular aatribute on the target function.

Here we can conclude that the accuracy of the Decision tree is higher (94.4%) than simple Perceptron algorithm which was about 89%. Therefore the Decision tree algorithm is better for the data which can be classified using superwised learning. Which gives better results for non linearly seperable data.

7. Referances

1)http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

2) https://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf