Decision Tree Classifier for Breast Cancer Wisconsin dataset

Shashank Pathak(15BCE1287)*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: shashank.pathak2015@vit.ac.in*

[be innovative but precise with the title]

***Abstract*—In machine learning, the decision tree is an algorithm for supervised learning for classification (functions that can decide whether an input, represented by a vector of numbers, belongs to some specific class or not).The algorithm allows for learning, in that it processes elements in the training set one at a time.We study the performance of the decision tree for the classification of the Breast Cancer Wisconsin Dataset.The system developed performs with 98.00% insample accuracy and 95.90% accuracy for the test sample in a two class classification**

1. Introduction

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too.

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data(training data). Decision Tree Learning is one of the most simple and widely used practical method for inductive inference over supervised data. A decision tree represents a procedure for classifying data based on attributes or features. It is also an efficient way of processing data ,for this very reason it has wide application in data mining.The construction of a decision tree does not require any domain knowledge or parameter setting, and therefor ebest for exploratory knowledge discovery.This way of representation and analysis of data is quite intutive and easy to assimilate by humans

2. Methodology

The working model of deciion tree is quite easy to implement and can be very effective in most of the classification problems.In decision trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with records attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

We continue comparing our records attribute values with other internal nodes of the tree until we reach a leaf node with predicted class value. As we know how the modeled decision tree can be used to predict the target class or the value.

3. Database - Breast Cancer Wisconsin Dataset(Original)

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used,to distinguish malignant (cancerous) from benign (non-cancerous) samples.This dataset consist of 699 instances and 11 attributes that would help the classiy the data into the two classes.The attributes are described in the fig.1

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

Figure 1. Breast Cancer Wisconsin Dataset Attributes

4. Algorithm

By using information gain as a criterion, we try to estimate the information contained by each attribute. We are going to use some points deducted from information theory. To measure the randomness or uncertainty of a random variable X is defined by Entropy.

$$H(X) = \mathbb{E}_X[I(x)] = -\sum_{x \in \mathbb{X}} p(x) \log p(x).$$

For a binary classification problem with only two classes, positive and negative class.

If all examples are positive or all are negative then entropy will be zero i.e, low. If half of the records are of positive class and half are of negative class then entropy is one i.e, high.

By calculating entropy measure of each attribute we can calculate their information gain. Information Gain calculates the expected reduction in entropy due to sorting on the attribute. Information gain can be calculated.

5. Experiments

First we used Decision tree classifier from scikit learn library to classify the dataset on a reduced set of attributes namely:

'ClumpThickness', 'UCellSize', 'UCellShape','MargAdhes', 'EpiCellSize', 'BareNuc', 'Chromatin', 'Nucleoli' and 'Mitoses'. After this I plotted the decision tree using graphviz and got a visual aspect of a decision tree. In this we can accurately predict the class of any new datapoint based on the selected attributes.

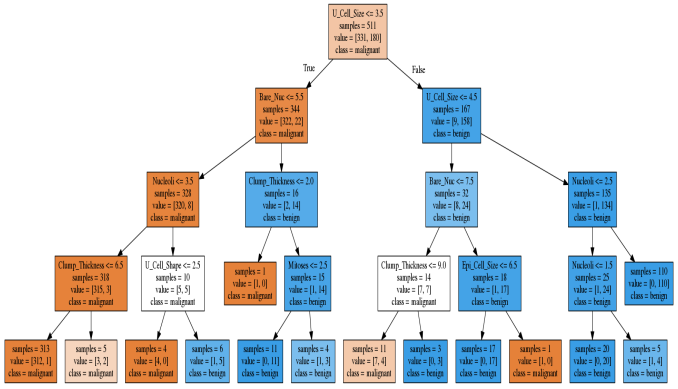

Figure 2.

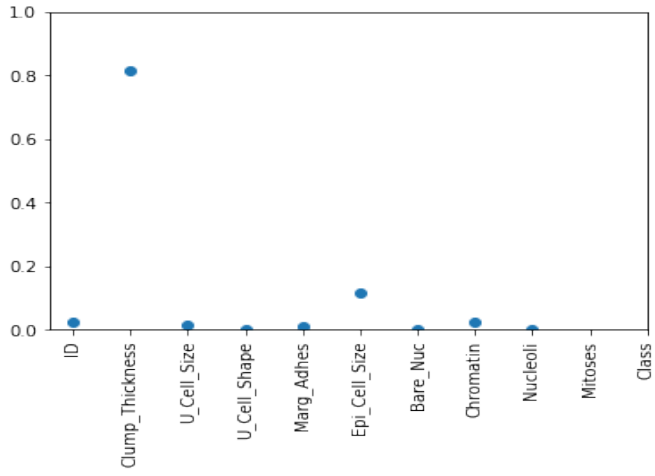

Figure 3. Train error vs Number of Epochs.

After the classification we can see that almost all the points are correctly classified .So we can say that the Decision Tree worked fine to classify the dataset by maximizing the information gain.

6. Conclusion

After the classification we can see that the classification was apt although there were datapoints which belong to wrong category.This indeed suggest that the Decision Tree model is a simple model which helps to classifify the data into two class Benign and Malignant but it is also quiet effective as a classification model to classify the Breast Cancer Dataset. The system developed performs with 98.00% insample accuracy and 95.90% accuracy for the test sample in a two class classification