National Taipei University of Technology

Introduction to Computer Science
Fall 2019-2020

Decision Tree Algorithm

Name: Li-Wei Yeh

StudentID: 108012047

Date: December 30th, 2019

Contents

Summary

Data Science is a field using scientific methods to typically analyze, make decisions or predict data. It consists of two different kinds of algorithms, supervised and unsupervised algorithms. Supervised algorithms rely on labeled data, whereas unsupervised algorithms rely on the model to discover its own information.

The Decision Tree Algorithm is an algorithm that uses a tree-like structure, making decisions based on questions. The questions depend on the kind of Decision Tree Algorithm. The ID3 algorithm uses entropy values for the information gain, whereas the CART algorithm uses Gini values for the information gain. Both however, have a similar approach and thus, a similar outcome. However, due to the Entropy using a log function, it's computationally more intensive. Therefore, generally the Gini value is preferred over the Entropy.

Data science

Data Science is a field that uses scientific methods to extract information from data. It is related to Statistics, Data Mining and Big Data. Data Science uses mathematical concepts to understand and analyze data, like; statistics, machine learning, computer science etc.

Data Science is primarily used to make decisions and predictions, making use of predictive analytics, prescriptive analytics, and machine learning.

Algorithms

An algorithm is a process or a set of rules to be followed in calculations or other problem-solving operations.

Typically, there are two kinds of algorithms, supervised and unsupervised algorithms

Supervised Algorithms

Supervised algorithms are algorithms that rely on labeled data, meaning that the outcome of the data is already known and thus, has the right "answer". Supervised algorithms rely on labeled training data to predict outcome of unforeseen data.

Unsupervised Algorithms

Unsupervised algorithms are algorithms that don't have labeled data. You need to make the model discover its own information. It is generally less accurate than supervised algorithms due to the trouble of having unlabeled data.

The Algorithms

There are numerous different algorithms to use for Data Science. Some of the more popular ones are: Linear Regression, K-means Clustering, Logistic Regression, Support Vector Machines, Feed Forward Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks and Decision Trees. And even those algorithms have different ways to implement them but come from the same idea.

This report will mainly focus on the Decision Trees algorithm.

Decision Trees

The Decision Tree algorithm is one of the predictive models used for statistics, data mining and machine learning. A decision tree can be used to visually represent decisions in decision making, thus making it easier for humans to understand the process it takes, to reach such conclusions.

Decision Tree learning is the construction of a decision tree from class-labeled training tuples. It has a flow-chart-like structure, where each node represents the outcome of the prediction. The trees are essentially a series of questions designed to assign a certain classification.

There are many different kinds of Decision Tree algorithms. Some working better than others. The more notable ones are:

- ID3
- C4.5 (successor of ID3)
- CART (Classification and Regression Trees)
- Chi-square automatic interaction detection (CHAID)

The most popular ones are ID3 and CART. They were both created at roughly the same time independently from each other, whilst still following a similar approach for learning a decision tree from training datasets.

ID3

To create the tree, we first need to have a feature to choose from. The feature we will choose, will depend on the feature which has the most information gain. We first need to define a measure commonly used in information theory, also known as entropy. Entropy is a measure of uncertainty in a dataset.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

The steps we have to take:

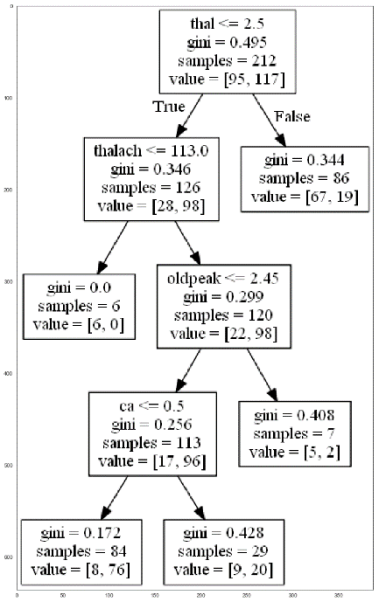
1. Compute the entropy for every feature
2. Calculate the information gain
3. Pick the highest gain
4. Repeat

Doing this will create a tree, resulting in a prediction of certain data.

CART

With cart, we basically take the same steps. Instead of the measure being entropy, this time the unit of measure will be Gini values. The Gini value will decide which feature will be chosen first (comparable to the most information gain). A Gini value gives an indication of how good a split is by how mixed the classes are.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$



The steps we have to take:

1. Compute the Gini index for every feature
2. Calculate Gini gain
3. Pick the best Gini gain attribute
4. Repeat

The result will be a decision tree, which is reliant on the previous choices in the flowchart-like graph.

Due to Entropy having a logarithmic function, which is computationally intensive, people generally prefer to use Gini values, thus the CART algorithm as opposed to ID3

Conclusion

The Decision Tree algorithm has different kinds of algorithms. Two of them are the ID3 algorithm and the CART algorithm. The ID3 algorithm is computationally more intensive, due to it using a log function. Therefore, generally the CART algorithm is used instead of the ID3 algorithm.