WHITE PAPER: DECISION TREES

by Michael J. Cavaretta, PhD

1. What is a Decision Tree?

A decision tree (a.k.a. classification tree) is a machine learning technique used to determine the relationship between a single dependent variable and a set of independent variables. In most cases the dependent variable is something we wish to predict, such as sales dollars, or whether a loan should be approved. The independent variables represent factors that influence the prediction, such as whether an item is promoted and its price discount, or if a loan applicant has filed for bankruptcy and their total assets.

Similar to other machine learning technologies, decision trees require a database of examples. This database is used to find relationships between the dependent and independent variables. This process is composed of six steps:

1. Group all of the values for the dependent variable. This is the top bucket.
2. Make the top bucket the current bucket.
3. Test each independent variable to find the greatest effect on the dependent variable.
4. Split the current bucket using the values of the independent variable found in the above step.
5. Set the current bucket to the first untested bucket generated from the above step.
6. Go to step 3.

Each independent variable is tested for its effect on the dependent variable. The independent variable with the greatest statistically significant effect is chosen for the first "split". That is, the dependent variable is broken into buckets containing a specific value or set of values for the independent variable and the corresponding value of the dependent variable. For example, if the dependent variable is loan approval and the first split is bankruptcy then the buckets would be bankruptcy = 'y' and bankruptcy = 'n'. These buckets contain examples of where the loan was approved and bankruptcy = 'y' for the first bucket and bankruptcy = 'n' for the second bucket. To find the next level of splits, each of the above buckets is tested against the other independent variables. If another independent variable has a statistically significant effect on the dependent variable the bucket is split. This process continues until either no statistically significant independent variables can be found or a pre-set stopping condition is met. Successively splitting buckets produces a graph that resembles a reversed tree, with the root at the top and branches spreading out towards the bottom. This is why the technology is called Decision Trees.

The stopping condition can be implemented in a number of different ways but they are generally based on the number of examples in a bucket. When this value becomes too low the bucket is not split any further. Two examples of stopping conditions include the number of examples in a bucket, and the percentage of total examples in a bucket.

One of the problems with early decision trees was their inability to handle continuous variables. This forced the user to group the continuous values into buckets prior to using the decision tree. Most decision trees in use today automatically group continuous values into ranges.

2. Where have Decision Trees been applied?

Decision tree technology has a long history in statistical analysis and was used extensively in the medicine, psychology, and biology fields for both analysis and prediction. Lately there is an increased interest in decision trees because of data mining. Decision trees fit in very well with the data mining process and compliment other data mining techniques such as neural networks.

3. When are Decision Trees used?

Decision trees can be used as both an analytical tool and as a prediction tool. Decision trees can aid in analysis tasks primarily because they can automatically determine which independent variable has the greatest effect on the dependent variable. Another benefit of using decision trees is that they can graphically show the relationship between the dependent variable and the independent variables. This viewing of the relationships allows a domain expert to perform a “realism” check. If applicable, source code representing the decision tree can be generated. This allows the relationships to be applied to unknown data just like any other model.

There are three conditions that should be met when choosing a decision tree.

- There is only one dependent variable. Decision trees cannot produce a model containing more than one dependent variable. Since each split subsets the data based on the splits that occurred before it, using more than one dependent variable is not possible.
- The relationships between a single independent variable and the dependent variable are generally stronger than between combinations of independent variables and the dependent variable. Decision trees test all of the independent variables for each split but only one individual is used. If the primary relationships are between combinations of independent variables the decision tree is useless.

- It is more important to view the relationships between the independent variables and the dependent variable than build a model with the highest accuracy. Many data mining technologies (e.g., neural networks) can discover relationships. Few of these technologies can present the relationships in an understandable manner. (However, a word of caution should be noted. With large decision trees containing dozens of independent variables, viewing the relationships and understanding the relationships may be very different tasks.) With this ability to display relationships is a corresponding decrease in accuracy. By definition, only one independent variable can be used for each split. This limits the accuracy of the Decision Tree when a number of independent variables are equally good.

4. How do Decision Trees compare with other Artificial Intelligence techniques?

Decision trees belong to group of artificial intelligence techniques called pattern recognizers. These techniques analyze a database of examples and once the pattern has been learned, can make predictions on unseen data. Other examples of pattern recognition technology include multivariate linear regression and neural networks.

While both linear regression and neural networks can easily handle continuous variables, decision trees must have all of the independent variables grouped into ranges. This allows the decision tree to graphically show the relationships between the independent and dependent variables using splits. However, the consequence of grouping continuous values is some sacrifice in accuracy. This can be thought of as the difference between a smooth continuous line going upward and a series of steps leading upward. At the “front” of the steps the decision tree will predict a value higher than the continuous line and “back” of the steps lower than the continuous line.

Decision trees are strictly hierarchical. That is, each prediction of the dependent variable is based on the values of the independent variables all the way from the bottom bucket to the top. This can seriously affect the ability of the decision tree to make predictions on unseen data. If the independent variable used for the first split is not present in the unseen data, the decision tree cannot make a prediction.

5. How would I develop a Decision Tree application?

Developing a decision tree application is very similar to developing a neural network application. The steps include project definition, knowledge engineering, database creation, model training and model validation. In fact, a combination of decision trees and neural networks can be very effective for data mining and pattern recognition problems. At an early stage in the pattern recognition process a decision tree can be used for the initial pattern recognition and finding the important independent variables. Later in the process, a neural network is used for the final model when the final set of factors is complete and accuracy is more important.

Copyright© 2025 Churchill Systems Inc. All rights reserved.