



"turning data into dollars"

Tom's Ten Data Tips – July 2008

Decision Trees

Decision trees are some of the most flexible, intuitive and powerful data analytic tools for exploring complex data structures. Because decision trees can be used for both prediction as well as insight, any data miner can gain from applying them in diverse projects.

The way decision trees work is that after a target variable has been chosen, the algorithm runs through all the columns in the dataset to find the "optimal" splitting point. What "optimal" means is driven by the loss function the tree is trying to minimize. The objective of splitting is to find subsets of records (called "leaves" of the tree) that differ the most with regards to the target variable. This process is then repeated for each subset of records at each leaf, hence the term "recursive partitioning" algorithms that is sometimes used. Afterwards the tree may require "pruning", the practice of eliminating splits that don't generalize to the test data. This is done to avoid overfitting on the training data.

A decision tree may not be the *only* tool you use, but you certainly need to be familiar with them. the *process* of tree building (manual – see tip# 4) is a rich source of insight in the data, and may point to new variables that might successfully be derived.

1. Inductive And Deductive Trees Have A Different Purpose

When data miners refer to decision trees, they actually mean *inductive* trees. Such an inductive decision tree is grown *from* the data.

However, using deductive trees is equally valid and can be just as useful, albeit for completely different reasons. Deductive trees are used to explicate formal decision processes, for instance. Or they may also be used for causal analysis, to link numerous contributing factors in a way they are assumed to exert their influence.

Inductive trees describe patterns that are discovered in the data, deductive trees are used to imbue reality with assumed structure (and are not associated with data mining).

2. The Importance Of Decision Tree Algorithms Is Overrated

The sequel to this is that the features/functionality of decision tree algorithms are of paramount importance! Decision tree algorithms determine what split is considered to be “optimal”, which is determined by the loss function one is minimizing. In the machine learning literature researchers continue to develop new tree growing and pruning algorithms. For practical purposes this is of limited relevance. There are no “bad” algorithms, all do a fair job.

What *is* relevant is that some algorithms can deal with a continuous target variable (e.g. CART), while most deal with a categorical target only. The target variable is mostly dichotomous, with CHAID being one of the few exceptions. For explanatory reasons (see also tip# 5), however, trees with multiple output categories are of limited use because they are too hard to interpret. Some tree algorithms produce binary splits only, others can (and mostly will) generate multi-way splits (see also tip# 6). Sometimes the tool can work directly with continuous variables, and sometimes these first need to be categorized. And last, but certainly not least, the way algorithms are implemented in tools has a huge impact on the user friendliness of interfaces. It should allow the miner to steer the tool, rather than the other way around (Pyle, 2003). All these features are far more important than the specifics of how the loss function is calculated, in particular for “manual” tree building (see also tip# 4).

3. Prediction Scores Need To Be Accompanied By Insight

Predictive modeling is not just about predicting, it is also and *always* about insight, too – and decision trees are very well suited for this. There are 3 reasons why it is good practice to accompany every prediction score with *at least* some rudimentary insight into the characteristics of that group (segment, tree leaf):

- This provides a sanity check: the variables that make up the explanation can guard against the occasional “blooper” when erroneous variables have wound up in the data set, unbeknownst to the miner...
- Providing not only a prediction but also insight helps foster acceptance of the prediction models. Nobody likes to rely on a “black box”.
- Insight may drive not just efficiency gains but improvement of effectiveness by reinvention of existing business practices. Insight helps spur innovations in marketing practices.

4. Manual Tree Building Allows For Maximum Input Of Domain Knowledge

There are basically two ways in which to create decision trees from the data: automatically or manually. In automatic mode, the algorithm in conjunction with the default settings determines which tree will evolve. When you apply manual (also called “interactive”) tree building, the data miner gets to input his domain knowledge, and he will use this to override (on occasion) the choices that are suggested by the algorithm. This approach is also sometimes called “model engineering”.

The advantage of manual tree building is that the process of model building itself, by trying several alternative variables per splitting point, is an extremely rich source of insight in the data. For this reason, we are convinced this is the superior way to build models using decision trees. Manual tree building makes pruning afterwards unnecessary, because the miner runs through all the options and keeps checking whether the chosen splits hold up in the training- as well as the test-data. So pruning is an integral part of the tree growing when you build models in manual (=interactive) mode.

The author has written two papers (together with his esteemed colleague Bas van den Berg) on interactive tree building that can be acquired through the following links:

5. Choose Decision Trees For Insight, *Not* Accuracy

Any data mining algorithm has data sets on which it will excel, and on which it will perform appallingly poor. And it is relatively easy to synthesize a dataset on which one algorithm does well, and another one very poor. The famous x-or problem is incredibly easy to model for a decision tree, yet a neural network will perform appallingly poor. *On average*, though, neural networks and regression models tend to produce models with higher predictive accuracy than decision trees.

The most compelling reason to use decision trees for building predictive models is that quite apart from a prediction score, decision trees are extremely powerful for generating insight in the data (see also tip# 7).

6. Multi-way Splits Are Easier To Read

Some tree algorithms split the record set arriving at a certain node in two branches (e.g.: CART), the majority of algorithms create multi-

way splits (e.g.: CHAID, ID3, C4.5, C5.0, QUEST, etc.). It is possible to *recreate* or copy a multi-way split by a series of dichotomous splits. A dichotomous tree building algorithm can therefore find all the splits a multi-way algorithm could, and more. This increased “flexibility” is the theoretical basis for assumed superiority of dichotomous splitting algorithms. This is mathematically sound.

However, in practice a lot is gained when the analyst can input domain knowledge in the tree building process. It turns out that one multi-way split is much clearer, and easier to read than a perfectly equivalent series of dichotomous splits. This gives multi-way splitting algorithms an advantage for manual tree building (see also tip# 4).

7. Decision Trees Are Useful For Discovering Interactions

One of the particularly useful purposes of decision trees is for discovering (important) interaction variables. When a mining dataset holds 100 variables, there are 9900 (99×100) first order interaction variables possible, 970200 (98×9900) second order interaction variables, etc. And often there's more than 100 columns. Including all these variables in a regression equation wouldn't just be unwieldy, it is mathematical nonsense. That's because you would wind up with more columns than data points to fit the regression equation.

By its very nature a decision tree is made up of interaction variables. You can use a decision tree to run through the most important interactions as they follow from the (preferably interactive) tree building process (see tip# 4). This serves as an efficient and tractable means of discovering interaction variables for use in a regression model or neural network. Regression models or neural networks may be preferred for the final model because they tend to generate predictions with higher accuracy than decision trees (see also tip #5).

8. Decision Trees Are Flexible And Versatile

As a modeling tool, decision trees can be made effective in an incredibly wide variety of circumstances. In contrast to almost all other data mining algorithms, decision trees can easily work with missing values in a data set “as is”. Whereas regression models and neural networks, for example, require special purpose data preparation efforts to replace missings, decision trees can deal with missing values as a separate and unique class of values without the need for replacement.

Decision trees can be used for problems that are focused on *either* insight *or* prediction. They have proven remarkably resilient against the curse of dimensionality. Even on data sets with very many columns decision trees tend to converge very quickly on a decent model. And the computational efficiency makes this feasible, too. The problem only becomes harder as a linear function of the number of columns whereas many other algorithms would simply break down when the number of columns in a data set becomes too large.

9. You Never Find *The* Tree, Merely *A* Tree

Because of their high transparency, decision trees are very easy to explain, even to colleagues with very limited experience in data modeling. Experience shows that only a brief introduction is needed to learn to interpret a decision tree model.

One caution though: what is not apparent (and often not so intuitive) from the display of a tree model is that there can be many trees that all do a more or less equivalent job of predicting the target variable. And these trees can look very, very different. This occurs especially in data sets with a lot of multi-collinearity. So when there is high correlation with variables in the tree, this will not show, nor be apparent from the model. But this should certainly be taken into account when interpreting the relation(s) between input variables and the target variable! By simply choosing a different variable as the first splitter, a completely differently looking tree might emerge that nonetheless has almost identical predictive properties.

10. Decision Trees Are A Means, Not An End

Data miners sometimes fall in love with their favorite methods (algorithms). As superior as decision trees appear in many respects, it is tremendously important to stay flexible with respect to the goals one is pursuing and that data mining is used for. Decision trees can be very useful for insight, but other tools like profiling, heatmaps, or visualization can sometimes shed a completely new light on the same data set. A good data miner is eclectic and will alternate use of techniques in response to the particular task at hand, and idiosyncrasies of any given data set.