Introduction

Trees and Random Forests

Trees

Decision trees may be familiar to many with a background in social science as a convenient way to represent data and assist in decision making. Morgan and Sonquist (1963) derived a way for constructing trees motivated by the specific feature space of data collected from interviews and surveys. Unlike, agricultural data which involves mostly numeric variables, the data collected from interviews is mostly categorical. On top of this, the datasets Morgan and Sonquist dealt with had few participants, and a lot of data collected on each one. To add to their difficulties, there was reason to believe that there were lurking errors in the data that would be hard identify and quantify. Lastly, many of the predictors were correlated. Morgan and Sonquist doubted that the additive assumptions of many models would be appropriate for this data. They noted that while many statistical methods would have difficulty accurately parsing this data, a clever researcher with quite a lot of time could create a suitable model simply by grouping values in the feature space and predicting that the response corresponding to these values would be an average of the observed responses given the grouped conditions. Their formalization of this procedure in terms of "decision rules" laid the ground work for future research on decision trees.

Later researchers proposed new methods for creating trees that improved upon the Morgan and Sonquist model. Leo Breiman et al (1984) proposed an algorithm called CART, *classification and regression trees*, that would allow trees to be fit on various types of data. An alternative to this method is conditional inference trees. Torsten Hothorn, Kurt Hornik, Achim Zeileis argue in their 2006 paper *Unbiased Recursive Partitioning: A Conditional Inference Framework*, CART has a selection bias toward variables with either missing values or a great number of possible splits. This bias can effect the interpretability of all tree models fit using this method. As an alternative to CART and other algorithms, Hothorn et al propose a new method, conditional inference trees.

There is a limit to the predictive capabilities of a single tree as they suffer from high variance. To alleviate this, aggregate methods called forests are often used instead. They function by enlisting the help of many trees, and then by aggregating the responses over all of them. The two most common types of forests are bagged and random forests.

What We Mean When We Talk About Inference

Inferential vs Descriptive Statistics

A note should be made of the difference between inferential and descriptive statistics. This paper's aim is to describe a process of making inferential claims using random forests, not descriptive ones. Descriptive statistics describe the data at hand without making any reference to a larger data generating system that they come from. It follows that inferential statistics then make claims about the data generating system given the data.

Permutations and Populations

As stated in the introduction of the *Chronical of Permutations Statistical Methods* by KJ Berry et al, 2014, there are two models of statistical inference. One is the population model, where we assume that the data was randomly sampled from one (or more) populations. Under this model, we assume that the data generated follows some known distribution. "Under the population model, the level of statistical significance that results from applying a statistical test to the results of an experiment or a survey corresponds to the frequency with which the null hypothesis would be rejected in repeated random samplings from the same specified population(s)", (Berry et al, 2014).

The permutation family of methods, on the other hand, only assumes that the observed result was caused by experimental variability. The test statistics is first calculated for the observed data, then the data is permuted a number of times. The statistic is calculated after each permutation to dervive a distribution of possible values. Then the original test statistic is tested against this distribution. If it is exceptionally rare, then there is evidence that our observation was not simply experimental variability.

A Step Back

A random forest R_f is the set of functions $T_1, ..., T_N$ where each T_j is a piece-wise function from the sample space Ω to the response space Φ. In general, Ω is defined by an n x p matrix where each column is a random variable and Φ is defined by an n x 1 vector Y.

Each tree in a random forest, $T_j \in R_f$, is generated on a subset of both Ω and Φ called the training set. This training set is a bootstrapped sample of the original dataset and is noted as B^t. It is then tested on a disjoint subset of Ω called the test set, $\bar{B}^t$, where $\bar{B}^t = \Omega \backslash B^t$. The image of T_j in Φ is called the predictions of T_j

As outlined in the 1984 textbook, *Classification and Regression Trees*, Brieman, Friedman, Olshen, and Stone described their method for creating, pruning, and testing regression trees. There are essentially three steps: one, decide on a variable to split over, two, partition that variable space in two distinct partitions, and three, set our initial predictions for each partition to be mean value of the response according to the observed responses corresponding to the values in the partitions. Recursively, this process is repeated for each new partition until some stopping condition is reached.This is a top down, greedy algorithm that functions by creating as large a tree as possible and then is pruned down to prevent over fitting.

Random Forests are generated by fitting a large number of trees, each on a boosted sample of the data. The crutial difference, however, between the trees in CART and the trees in a random forest, is that at each node in a random forest, only a subset of the predictors are considered as candidates for possible splits. This decorrelates each tree from its neighbors, and decreases bias of the whole model while slightly increasing variance of each tree.

Inference on Random Forests

The Problem

Random forests create models with great predictive-, but poor inferential capabilities. After Morgan and Sonquist's initial development of decision trees, trees quickly moved to the domain of machine learning and away from statistics. Researchers focused on bettering predictions and improving run times and less on the statistics behind them. Inferential statistics with random forests is usually treated as a variable selection problem, and generally falls behind the predictions in importance. This has limited the applications of random forests in certain fields, as to many the question of "why" the data is the way it is, is just, if not more, important as the predictions. There are several means of performing descriptive statistics with random forests that could be interpreted incorrectly as attempting to answer this but without a statistically backed method for performing variable importance, the use of random forest is limited to prediction-only settings.

Proposed solutions to this problem

Breiman and Cutler proposed a method of permuted variable importance in their paper (cite) to answer this problem. Their method compares the variable importance for each variable in a tree-wise manner. For each tree, the permuted variable importance of the variable X_j is:

$$VI^t(x_j) = \frac{\sum_{i \in |\bar{B}^t|}(y - \hat{y})^2}{|\bar{B}^t|} - \frac{\sum_{i \in |\bar{B}_p^t|}(y - \hat{y_p})^2}{|\bar{B_p}^t|}$$

Where $\bar{B}^t$ is the out of bag sample for tree t, $|B|$ is the number of observations in that sample, $\bar{B}_p^t$ is with X_j permuted, $\hat{y}$ is the predicted outcome, and $*\hat{y}^t$ is the predicted outcomes after variable X_j has been permuted. This value is averaged over all the trees. It's important to note that if the variable X_j is not split on in the tree t, the tree-wise variable importance will be 0.

Creating a permutation-based method is certainly an attractive solution to our problem. One, it allows us to estimate the distribution of variable importance and generate a Z score under the null hypothesis that $PV = 0$.

$$VI_\alpha(x_j) = \frac{\sum_1^n treePV^t(x_j)}{\frac{\hat{\sigma}}{\sqrt{ntree}}}$$

Strobl et al from the University of Munich criticize this method in their 2008 technical report, *Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance*. One, this method has the downside of increasing power with increasing numbers of trees in the forest. This is a more or less arbitrary parameter which we would hope would not affect our importance estimates. Secondly, the null hypothesis under Breiman and Cutler's strategy is that the variable importance V for any variable X_j is not equal to zero given Y, the response. Because random forests are most often used in situations with multicolinearity that would make other methods like the linear model difficult, Strobl argues that any variable importance measure worth its salt should not be mislead by correlation within the predictors.

The researchers at the University of Munich published a fully fleshed response to the Breiman and Cutler method in 2008, titled *Conditional Variable Importance for Random Forests* that address these issues. Strobl et al propose restructuring the Breiman and Cutler algorithm to account for conditional dependence among the predictors.The null hypothesis is that $VI_\beta(X_j) = 0$ given the predictor Y *and all other predictors* $X_1, ..X_n$.This accounts for interactions between X_j and the other predictors.

This paper aims to provide a response to this method. The partitions are made from the random forest corresponding to the formula of $Y\ X_1, ..., X_n$ instead of a model of $X_j\ X_1, ..., X_n$. This ignores the common situation where the predictors are correlated enough, they act as stand ins for each other, so that if one variable is heavily influential in a certain tree at predicting Y, the other variable will be forgotten all together.