

## Frequently Asked Questions and Answers about TreeNet

TreeNet is a revolutionary advance in data mining technology developed by Jerome Friedman, one of the world's outstanding data mining researchers. TreeNet offers exceptional accuracy, blazing speed, and a high degree of fault tolerance for dirty and incomplete data. It can handle both classification and regression problems and has been proven to be remarkably effective in both traditional numeric data mining and text mining. This document introduces TreeNet and answers some basic questions about the methodology.

### How does TreeNet fit into the scheme of data mining tools?

TreeNet is designed for very high accuracy predictive modeling. Because TreeNet attempts to achieve this goal even if very complex models are required, models may be relatively difficult to understand in detail. However, the graphs produced by TreeNet software display the impact of any relevant predictor or pair of predictors on the target, thus revealing the underlying data structure.

We see TreeNet as a tool to be used after the data have been explored with tools such as CART and MARS. CART and MARS produce output that can clearly reveal data errors and inconsistencies, quickly leading to a detailed understanding of the data and potential problems. Once data quality has been assured and basic understanding of the key drivers in the data has been achieved, reanalyzing the data with TreeNet is worthwhile. In most cases, TreeNet will confirm the primary findings reported by CART or MARS while substantially increasing the predictive accuracy of the models.

### How does TreeNet work and what does a TreeNet model look like?

A TreeNet model normally consists of from several dozen to several hundred small trees, each typically no larger than two to eight terminal nodes. The model is similar in spirit to a long series expansion (such as a Fourier or Taylor's series) - a sum of factors that becomes progressively more accurate as the expansion continues. The expansion can be written as:

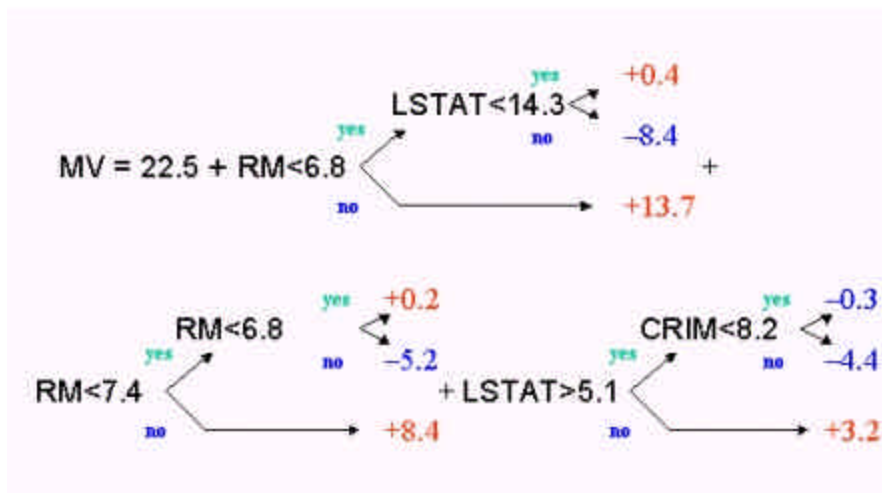
$$F(X) = F_0 + \beta_1 T_1(X) + \beta_2 T_2(X) + \dots + \beta_M T_M(X)$$

where each  $T_i$  is a small tree. An example of the first few terms of a model to predict home values based on the 1970 Census Boston Housing data set is:



The model tells us that we start with the mean home value (in 1970) of \$22,533 and adjust that estimate upwards by \$13,541 for larger homes, and adjust it upwards again by \$2,607 for neighborhoods with good socioeconomic status indicators. In practice the adjustments are usually much smaller than shown in this regression example and hundreds of adjustments may be needed. The final model is thus a collection of weighted and summed trees. For binary classification problems, a yes or no response is determined by whether the sign of the predicted outcome is positive or negative. For multi-class problems a score is developed separately for each class via class specific expansions, and the scores are converted into a set of probabilities of class membership.

The example above uses the smallest possible 2-node tree in each stage. More complicated models tracking complex interactions are possible with 3 or more nodes at each stage.



## What are the advantages of TreeNet?

TreeNet advantages include:

- Automatic selection from thousands of candidate predictors
  - No prior variable selection or data reduction is required
- Ability to handle data without preprocessing
  - Data does not need to be rescaled, transformed, or modified in any way
- Resistance to outliers in predictors or the target variable
- Automatic handling of missing values
- General robustness to dirty and partially inaccurate data
- High Speed
- Trees are grown quickly; small trees grown extraordinarily quickly
- TreeNet is able to focus on the data that is not easily predictable as model evolves
  - Thus, as additional trees are grown less and less data needs to be processed
  - In many cases, TreeNet is able to train effectively on 20% of the data
- Resistance to Over-Training
  - When working with large data bases even models with 2,000 trees show little evidence of over-training
  - Most models show maximum accuracy well before 1,000 trees are grown

TreeNet's robustness extends to data contaminated with erroneous target labels. For example, in medicine there is some risk that patients labeled as healthy are in fact ill and vice versa. This type of data error can be very challenging for conventional data mining methods and will be catastrophic for conventional boosting. In contrast, TreeNet is generally immune to such errors as it dynamically rejects training data points too much at variance with the existing model.

In addition, TreeNet adds the advantage of a degree of accuracy usually not attainable by a single model or by ensembles such as bagging or conventional boosting. Independent real world tests in text mining, fraud detection, and credit worthiness have shown TreeNet to be dramatically more accurate on test data than other competing methods.

Of course no one method can be best for all problems in all contexts. Typically, if TreeNet is not well suited for a problem it will yield accuracies on par with that achievable with a single CART tree.

## What does TreeNet output look like?

The TreeNet model is a complex structure not easily understood by studying its individual components. However, TreeNet produces a number of clear reports and graphs that reveal the core message and predictive content of the model. These include:

- Variable importance ranking
- Graphs of the typical relationship between the target and any one predictor
  - All other variable effects are taken into account to arrive at a typical relationship
  - Technically, we graph  $E(Y|X_i)$  for a single predictor  $X_i$ , integrating out all other relevant predictors
- 3-D graphs of the target against any pair or predictors
- The first few trees of the model may also be displayed as a set of text rules.

## How are prediction and scoring handled?

Optionally, TreeNet 1.0 will score any database and output predictions in the file format or database required. The data management system allows access to any of 85 file formats. Input and output file formats can be different. Alternatively, the TreeNet model can be exported as a SAS® language subroutine.

## What are the advantages of TreeNet over a neural net?

TreeNet is not sensitive to data errors and needs no time-consuming data preparation, preprocessing or imputation of missing values. TreeNet is resistant to over-training and is over 100 times faster than a neural net. Finally, TreeNet is not troubled by hundreds or thousands of predictors.

## Can a neural net do anything a TreeNet cannot?

Yes. Version 1.0 of TreeNet cannot accept more than one target variable at a time. To model a collection of targets a separate TreeNet model must be developed for each target independently. Also, neural nets can simultaneously estimate a function and its derivatives whereas TreeNet is not designed to estimate the target function derivatives.

## What is the technology underlying TreeNet and how does it differ from boosting?

TreeNet uses gradient boosting to achieve the benefit of boosting (accuracy) without the drawback of a tendency to be misled by bad data. In boosting, each tree grown would normally be a fully articulated stand-alone model, with each boosted tree combined with its mates via a weighted voting scheme. In contrast, each TreeNet component is a small tree, often no larger than two terminal nodes; trees are summed together with very small weights on each component.

## What is the TreeNet track record?

TreeNet was developed in 1997 by Stanford University's Jerome Friedman, one of the authors of CART®, the author of MARS™, and the inventor of Projection Pursuit and HotSpotDetector™. The TreeNet technology has been tested in a broad range of industrial and research settings and has demonstrated considerable benefits. In tests in which TreeNet was pitted against expert modeling teams using a variety of standard data mining tools, TreeNet was able to deliver results within a few hours comparable to or better than those that requiring months of hands-on development by expert data mining teams.

## **How does TreeNet fit into the Salford Systems data mining solution?**

The Salford Systems data mining solution rests on three pillars: CART® for clear, interpretable classification via decision trees and rules; MARS™ for clear, interpretable predictive models via regression and logistic regression, and TreeNet for situations in which the analyst may desire to sacrifice interpretability in favor of accuracy. Even in circumstances where interpretability and transparency are mandatory and a model must be expressible in the form of rules, TreeNet can serve a useful function by benchmarking the maximum achievable accuracy against which interpretable models can be compared.

## **What are the hardware and resource requirements of TreeNet?**

TreeNet 1.0 requires that both training and test data reside in RAM. Thus, if large databases are being analyzed, TreeNet will be most effective when running on large-capacity servers. We recommend a minimum of 512 MB RAM and on Windows machines, Windows 2000 or XP or later versions of the OS are preferred platforms for performance. TreeNet is available for Windows 98/NT/2000 and UNIX (IBM AIX, Compaq Alpha, SGI, HP, and Sun) platforms and will run with as little as 64 MB RAM. A Linux version is planned.

## **Where can I learn more about TreeNet?**

Contact Salford Systems at 619.543.8880 or e-mail [support@salford-systems.com](mailto:support@salford-systems.com). We maintain a collection of white papers and academic studies on various data mining topics on the web site and offer tutorials on TreeNet, CART, and MARS in major cities worldwide. Internet meetings to demonstrate and discuss any of our products can be arranged.