Francis Analytics

Actuarial Data Mining Services

Data Mining 101

By Louise Francis

November 3, 2001

Data Mining Definition

- Finding Patterns in data
  - "the computer assisted process of information analysis"
- Usually thought of as being applied to large databases
  - Many records
  - Many variables

Data Mining Definition

- Includes classical approaches such as regression and logistic regression
- More typically thought of as involving techniques developed in AI such as neural networks and Genetic Algorithms as well as recently developed statistical techniques like decision trees (CHAID, CART), regression splines and clustering

Data Mining Definition

- Most of what is regarded as data mining:
  - Does not require assumption of normality or other distributional assumptions
  - Can fit nonlinear relationships
  - Can accommodate other messiness in data such as interactions
  - Tends to be computationally intensive

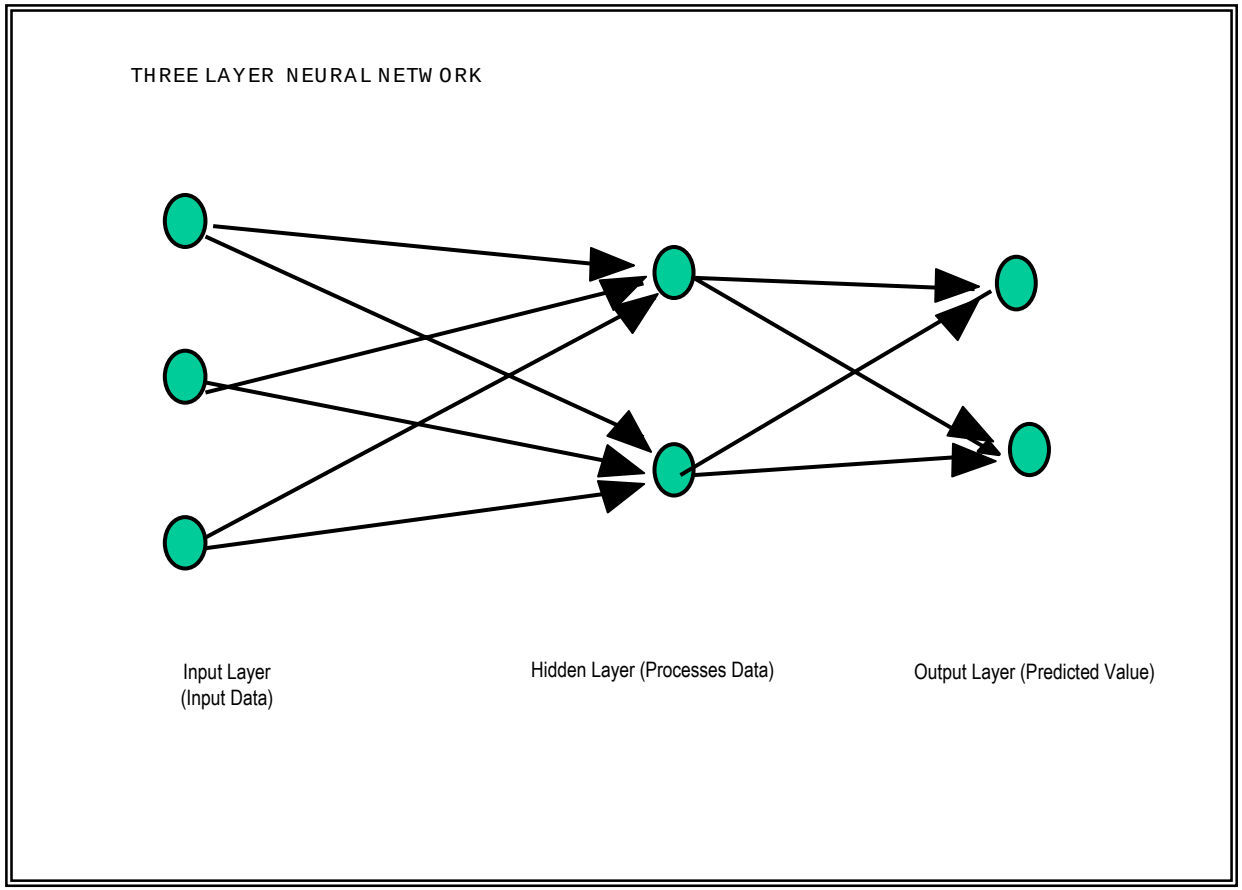
Data Mining Classification

- Supervised Learning
  - Have a dependent variable
- Unsupervised Learning
  - No dependent variable

Data Mining Techniques

- Neural Networks
- MARS
- Decision Trees
- Clustering
- Others
  - Association rules
  - Genetic Algorithms
  - Fuzzy Logic

Neural Networks



Neural Networks

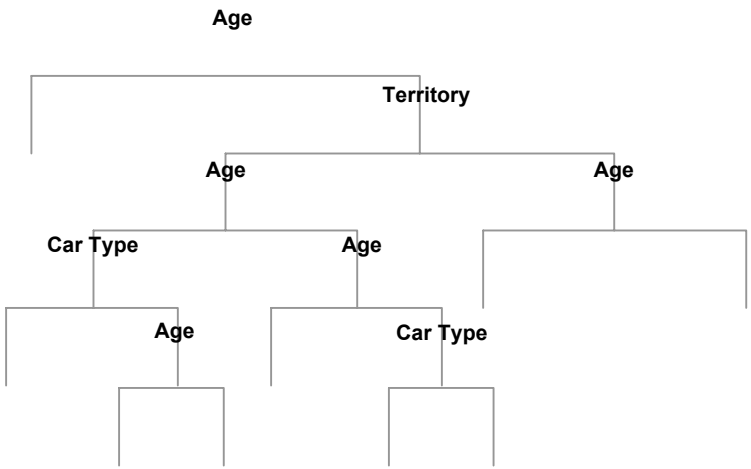
- Developed in Artificial Intelligence discipline
- Based on how neurons in brain work
- From a statistics perspective, a complicated nonparametric regression
- A universal function approximator: theoretically can approximate any nonlinear function

Decision Trees

- Sequentially splits data into categories which have similar values for dependent variables
- Uses a statistic such as chi square statistic or R^2 to do split
- Result can be presented in form of a tree or as rules

Decision Trees

SAMPLE TREE MODEL OF AUTOMOBILE CLAIM FREQUENCY



MARS

- Multivariate Adaptive Regression Splines
- Results look like a regression
- Uses splines to fit non linear functions
- Uses search technique to find significant interactions
- Technique creates new variables or basis functions to capture changes in regression slope and interactions

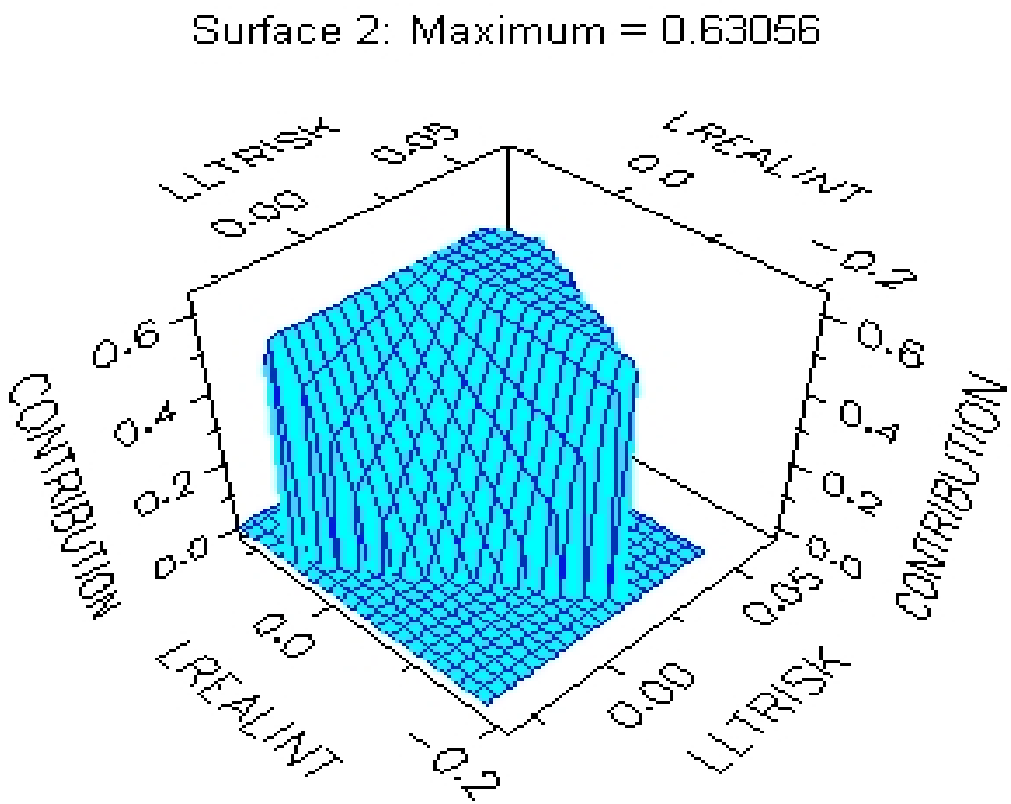
MARS Model Example

Relationship Between Annual Stock Return and Other Financial Variables

```
 BF1 = max(0, REALINT - 0.048);
BF2 = max(0, 0.048 - REALINT );
BF3 = max(0, DIVIDEN - 0.028);
BF5 = max(0, 0.030 - LLTBONDS ) * BF3;
BF7 = max(0, DIVIDEN - 0.054);
BF14 = max(0, 0.016 - LTRISK ) * BF2;
BF17 = max(0, TBILL + .199999E-03) * BF7;
  Y = 0.141 - 8.914 * BF1 - 782.090 * BF5 - 240.926 *     BF14
+ 1529.334 * BF17;
```

LREALINT = real interest rate, DIVIDEN = dividend rate, LTRISK= long term – short term - interest, TBILL = Tbill rate

MARS Model Fitted Function for Real Interest and Long Term Risk



Clustering

- Partitions data into similar records
- Based on similar values on independent variables
- There is no dependent variable

Recommended reading

- Berry, Michael J. A., and Linoff, Gordon, *Data Mining Techniques*, John Wiley and Sons, 1997
- Dhar, Vasant and Stein, Roger, *Seven methods for Transforming Corporate Data Into Business Intelligence*, Princeton Hall, 1997
- Witten, Ian and Frank, Eibe, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufman Publishers, 2000
- Derrig, Richard, "Patterns, Fighting Fraud With Data", *Contingencies*, pp. 40–49.
- Lawrence, Jeannette, *Introduction to Neural Networks: Design, Theory and Applications*, California Scientific Software, 1994
- Smith, Murry, *Neural Networks for Statistical Modeling*, International Thompson Computer Press, 1996