options are provided in an efficient graphical user interface in which you can “brush” the current tree, i.e., select a specific node to grow a branch, delete a branch, etc. As in all modules for predictive data mining, the decision rules contained in the final tree built for regression or classification prediction can optionally be saved in a variety of ways for deployment in data mining projects, including C/C++, STATISTICA Visual Basic, or Predictive Model Markup Language (PMML). Hence, final trees computed via this module can quickly and efficiently be turned into solutions for predicting or classifying new observations.

Multivariate Adaptive Regression Splines (MARSplines)

We’ll use the STATISTICA Data Miner software tool to describe the MARSplines algorithm, but the ideas described can be applied to whatever software package you use.

Note: Many of the paragraphs in this section are adapted from the STATISTICA online help, StatSoft, Inc. (2008). STATISTICA (data analysis software system), version 8.0. www.statsoft.com.

The STATISTICA Multivariate Adaptive Regression Splines (MARSplines) module is a generalization of techniques (called MARS) popularized by Friedman (1991) for solving regression- and classification-type problems, with the goal to predict the value of a set of dependent or outcome variables from a set of independent or predictor variables. MARSplines can handle both categorical and continuous variables (whether response or predictors). With categorical responses, MARSplines will treat the problem as a classification problem; with continuous dependent variables, as a regression problem. MARSplines will automatically determine that for you.

MARSplines is a nonparametric procedure that makes no assumption about the underlying functional relationship between the dependent and independent variables. Instead, it constructs the model from a set of coefficients and basis functions that are entirely “driven” from the data. In a sense, the method follows decision trees in being based on the “divide and conquer” strategy, which partitions the input space into regions, each with its own regression or classification equation. This makes MARSplines particularly suitable for problems with higher input dimensions (i.e., with more than two variables), where the curse of dimensionality would likely create problems for other techniques.

The MARSplines technique has become particularly popular in data mining because it does not assume or impose any particular type or class of relationship (e.g., linear, logistic, etc.) between the predictor variables and the dependent (outcome) variable of interest. Instead, useful models (i.e., models that yield accurate predictions) can be derived even in situations in which the relationships between the predictors and the dependent variables are non-monotone and difficult to approximate with parametric models. For more information about this technique and how it compares to other methods for nonlinear regression (or regression trees), see Hastie et al. (2001).

In linear regression, the response variable is hypothesized to depend linearly on the predictor variables. It’s a parametric method, which assumes that the nature of the relationships (but not the specific parameters) between the dependent and independent variables is known a priori (e.g., is linear). By contrast, nonparametric methods do not make any such assumption as to how the dependent variables are related to the predictors. Instead, it allows the model function to be “driven” directly from data.