How Understandable Do the Classes Have to Be?

One of the strengths of a decision tree model is that it produces results that are easy to understand in terms of the predictor variables and target variables. Results from a neural net model may be more predictive, but the understandability of the results in terms of the predictor variables is less clear. Often, there is a trade-off between accuracy and understandability of the results. This trade-off may be related to the choice of modeling algorithm. Some algorithms do better for some data sets than others. The STATISTICA Data Miner Recipe Interface uses several modeling techniques in the form of a recipe that provides a basis for choosing the right trade-off combination of accuracy and understandability of the model results.

ASSUMPTIONS OF CLASSIFICATION PROCEDURES

Classification in general requires that you accept a number of assumptions. The fidelity of your classes and their predictive ability will depend on how close your data set fits these assumptions. In Chapter 4, we stressed the importance of describing your data set in terms of the nature of its variables, their possible interactions with the target variable and with each other, and their underlying distributional pattern. In classification, you should try to satisfy these assumptions as much as possible.

Numerical Variables Operate Best

Categorical variables can be used, but they should be decomposed into dummy variables, if possible (cf. Chapter 4 for an introduction to dummy variables).

No Missing Values

By default, most data mining algorithms (including those for classification) will eliminate cases with missing values in predictor variables. Imputation of missing values is one way to fix this problem (see Chapter 4). Another way that some classification algorithms (e.g., C&RT) may fill missing values is to use surrogate variables. A surrogate variable has a similar splitting behavior to the variable with the missing value, and its value in this case can be used to replace a missing value in predictor variable.

Variables Are Linear and Independent in Their Effects on the Target Variable

When we say that variables are linear and independent in their effects on the target variable, we mean that there is a straight-line (linear) change in the target variable as each variable is varied over its range and that the effect of one variable is not related to (is independent of) effects of any other variable. There is not much you can do about