to supervised classification methods. Unsupervised classification methods will be discussed in Chapter 17, in relation to the detection and modeling of fraud.

There are two general kinds of supervised classification problems in data mining: (1) binary classification—only one target variable and (2) multiple classification—more than one target variable. An example of analyses with only one target variable are models to identify high-probability responders to direct mail campaigns. An example of analyses with multiple target variables is a diagnostic model that may have several possible outcomes (influenza, strep throat, etc.).

**INITIAL OPERATIONS IN CLASSIFICATION**

Before classification can begin, there are some initial tasks you must perform:

1. Determine what kind of classification problem you have. This means that you have to determine how many target classes you have and define them, at least in general terms.
2. Define the boundaries of each class in terms of the input variables.
3. Construct a set of decision rules from class boundaries to define each class.
4. Determine the prior-probability of each class, based on the frequency of occurrence of a class in the entire data set.
5. If appropriate, you should determine the cost of making the wrong choice in assigning cases to a given class. This task is extremely important in some classification situations (e.g., medical diagnosis).

**MAJOR ISSUES WITH CLASSIFICATION**

There are a number of issues that you must face before proceeding with the classification project. It is important to consider each of these issues, and either resolve the issues before modeling or set some expectations surrounding them.

**What Is the Nature of the Data Set to Be Classified?**

The purpose of the classification should be specified, and it should be related to the expected interpretation of the results. For example, a classification of breast cancer propensity should be accompanied with costs of a wrong diagnosis. You probably would not worry much about misclassifying response rate in a mailing campaign, but failing to diagnose a breast tumor could be fatal.

**How Accurate Does the Classification Have to Be?**

If you are in a time-crunch, a model with 80% sensitivity to prediction accuracy built in 2 days may be good enough to serve the model’s purpose. In business, time is money!