Naïve Bayesian classifiers assume that the predictor variables are independent in their effects on the classification. This assumption is rather naïve in the face of reality, hence the origin of the name. But in spite of this rather too-strong assumption, it performs rather well with many data sets. This classifier can accept any number of either continuous or categorical variables. In fact, the Naïve Bayesian classifier technique is particularly suited when the number of variables (the dimensionality of the inputs) is high. Although the assumption that the predictor variables are independent is not always accurate, it does simplify the classification task dramatically, since it allows the class conditional densities to be calculated separately for each variable; i.e., it reduces a multidimensional task to a number of one-dimensional ones. Furthermore, the assumption does not seem to greatly affect the posterior probabilities, especially in regions near decision boundaries, thus leaving the classification task unaffected. In effect, Naïve Bayesian classifiers reduce a high-dimensional density estimation task to a one-dimensional kernel density estimation. This kernel function can be modeled in several different ways including normal, lognormal, gamma, and Poisson density functions.

**WHAT IS THE BEST ALGORITHM FOR CLASSIFICATION?**

If you have the time to use all the algorithms discussed in the preceding sections to classify your data sets, you will find that the best algorithm to use to classify one of your data sets may not work well for other data sets. In other words, different algorithms work best for different data sets. Using a diversity of algorithms is best. A good example is provided in the results of a study of performance of 10 data mining algorithms by Kalousis et al. (2004). They compared algorithm performance on 80 data sets available from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Each data set was characterized by a number of criteria:

- Error correlation between two classifiers using the data set (EC)

\[
EC = P(i)[\text{Covariance}(X_1, X_2)], \text{ summed for target}(i) = 1 \text{ and target}(i) = 0, \\
EC = \sum_{i=0} P(i)[\text{cov}(error_1, error_2)] \\
\text{where } P(i) = \text{the prior probability of each target class (proportion } = 1 \text{ and } 0), \text{ summed for cases where target } = 1 \text{ and target } = 0.
\]

All data sets were clustered on their matrices of error correlation and grouped into two classes: relatively low EC and relatively high EC.

- Log of the total number of cases/number of attributes
- Sum of the logs of the total number of cases/number of cases where target = 0 and total number of cases/number of cases where target = 1
- Log of the total number of cases/number of cases where target = 1

The data sets with high EC and low EC were further characterized by other variables, as shown in Table 11.3.