results of the shuffled-Y models (mean and standard deviation) and note, for the real model’s results, how many standard deviations above the shuffled mean it is. This is a very good measure of quality.

We recommend learning to use such resampling techniques (cross-validation, leave-one-out, bootstrap, jackknife) as one of the most important topics in data mining and statistics. One useful resource to explore is www.statistics.com.

RE-CAP OF THE MOST POPULAR ALGORITHMS

Some aspects of the algorithms summarized next were discussed in previous chapters (7, 8, 11, and 12). We briefly highlight the distinct ways they work to set the stage for how to enhance them. With each algorithm, we note if it is a consensus method (summarizing and tossing the data) or contributory (keeping and using the data), and also whether it can select variables. It is very rare for a contributory method to be variable-selecting, but most consensus methods can be.

Linear Methods (Consensus Method, Stepwise Is Variable-Selecting)

Methods traditionally used in statistical analysis often contribute significantly to a data mining effort, at the very least providing a baseline against which to compare more modern techniques. Linear regression (LR) predicts a response (dependent) variable by a weighted sum of predictor (independent) variables. The estimation surface is a plane in the space of candidate variables, though those variables can be nonlinear functions of the original ones. The plane is the one minimizing the squared error over the “training” cases. This leads to elegant mathematics and fast computation, but is rarely the exact scoring function most appropriate for estimation penalization. Linear discriminant analysis (LDA) predicts a categorical response variable by creating a discriminating plane separating the groups of the response variable. A quadratic extension allows for nonlinear boundaries but requires estimating covariance matrices for each class.

Decision Trees (Consensus Method, Variable-Selecting)

Decision trees (DTs) are the most popular inductive method in current use. DTs are often built in two stages. When growing, the algorithm finds at each node (subset of data) the best feature discriminating between the classes and then splits the data into two new nodes based on that feature. This is applied recursively to the resulting data subsets until a class assignment can be made for each leaf. The second stage of pruning works by clipping off (reabsorbing) the least useful branches of the tree to best balance the accuracy on the training data with the complexity of the model. (A simpler model is generally more robust, i.e., more accurate on new data.) The final tree partitions the feature space in a number of labeled regions—typically in the form of axis-parallel hyper-rectangles.

Independently developed in multiple fields (including statistics, computer science, artificial intelligence, and psychology), DTs are heavily used in the newer disciplines of machine learning and data mining. Early psychology researchers believed that such trees were a