Data Mining Terminology and Notation

Algorithm: A specific procedure used to implement a particular data mining technique such as classification tree, discriminant analysis, and the like.

Attribute/Feature/Independent Variable/Input Variable/Predictor: A variable, usually denoted by X, used as an input into a predictive model, also called a field from a database perspective.

Case/Observation/Record: The unit of analysis on which the measurements are taken (a customer, a transaction, etc.), also called instance, sample, example, pattern, or row. In spreadsheets, each row typically represents a record. Each column is a variable. Note that the term "sample" here is different from its usual meaning in statistics, where it refers to a collection of observations.

Categorical Variable/Factor Variable: A variable that takes on one of several fixed values, for example, a flight could be on-time, delayed, or canceled.

Confidence: A performance measure in association rules of the type "IF A and B are purchased, THEN C is also purchased". Confidence is the conditional probability that C will be purchased IF A and B are purchased. Confidence also has a broader meaning in statistics (confidence interval), concerning the degree of error in an estimate that results from selecting one sample as opposed to another.

Estimation/Prediction: The prediction of the numerical value of a continuous output variable.

Holdout Data/Holdout Set: A sample of data not used in fitting a model, but instead used to assess the performance of that model. This course uses the terms validation set and test set instead of holdout set.

Model: An algorithm as applied to a dataset, complete with its settings (many of the algorithms have parameters that the user can adjust).

Dependent Variable/Outcome Variable/Output Variable/Response/Target: A variable, usually denoted by Y, which is the variable being predicted in supervised learning.

Conditional Probability: The conditional probability of event A occurring given that event B has occurred. Read as "the probability that A will occur given that B has occurred".

Profile: A set of measurements on an observation (e.g., the height, weight, and age of a person).

Sample: In the statistical community, "sample" means a collection of observations. In the machine learning community, "sample" means a single observation.

Score: A predicted value or class. Scoring new data means using a model developed with training data to predict output values in new data.

Success Class: The class of interest in a binary outcome (e.g., purchasers in the outcome purchase/no purchase).

Supervised Learning: The process of providing an algorithm (logistic regression, regression tree, etc.) with records in which an output variable of interest is known and the algorithm "learns" how to predict this value with new records where the output is unknown.

Test Data/Test Set: The portion of the data used only at the end of the model building and selection process to assess how well the final model might perform on new data.

Training Data/Training Set: The portion of the data used to fit a model.

Unsupervised Learning: An analysis in which one attempts to learn patterns in the data other than predicting an output value of interest.

Validation Data/Validation Set: The portion of the data used to assess how well the model fits, to adjust models, and to select the best model from among those that have been tried.

Variable: Any measurement on the records, including both the input (X) and the output (Y) variables.