



Glossary of data mining terms:

Accuracy

Accuracy is an important factor in assessing the success of data mining. When applied to data, accuracy refers to the rate of correct values in the data. When applied to models, accuracy refers to the degree of fit between the model and the data. This measures how error-free the model's predictions are. Since accuracy does not include cost information, it is possible for a less accurate model to be more cost-effective. Also see precision.

Activation function

A function used by a node in a neural net to transform input data from any domain of values into a finite range of values. The original idea was to approximate the way neurons fired, and the activation function took on the value 0 until the input became large and the value jumped to 1. The discontinuity of this 0-or-1 function caused mathematical problems, and sigmoid-shaped functions (e.g., the logistic function) are now used.

Analytical model

A structure and process for analyzing a dataset. For example, a decision tree is a model for the classification of a dataset.

Anomalous data

Data that result from errors (for example, data entry keying errors) or that represent unusual events. Anomalous data should be examined carefully because it may carry important information.

Antecedent

When an association between two variables is defined, the first item (or left-hand side) is called the antecedent. For example, in the relationship "When a prospector buys a pick, he buys a shovel 14% of the time," "buys a pick" is the antecedent.

Artificial neural networks

Nonlinear predictive models that learn through training and resemble biological neural networks in structure.

Associations

An association algorithm creates rules that describe how often events have occurred together. For example, "When prospectors buy picks, they also buy shovels 14% of the time." Such relationships are typically expressed with a confidence interval.

Back-propagation

A training method used to calculate the weights in a neural net from the data.

Bias

In a neural network, bias refers to the constant terms in the model. (Note that bias has a different meaning to most data analysts.) Also see precision.

Binning

A data preparation activity that converts continuous data to discrete data by replacing a value from a continuous range with a bin identifier, where each bin represents a range of values. For example, age could be converted to bins such as 20 or under, 21-40, 41-65 and over 65.

Bootstrapping

Training data sets are created by re-sampling with replacement from the original training set, so data records may occur more than once. In other words, this method treats a sample as if it were the entire population. Usually, final estimates are obtained by taking the average of the estimates from each of the bootstrap test sets.

CART

Classification and Regression Trees. A decision tree technique used for classification of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. Segments a dataset by creating 2-way splits. Requires less data preparation than CHAID.



Categorical data

Categorical data fits into a small number of discrete categories (as opposed to continuous). Categorical data is either non-ordered (nominal) such as gender or city, or ordered (ordinal) such as high, medium, or low temperatures.

CHAID

Chi Square Automatic Interaction Detection. A decision tree technique used for classification of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. Segments a dataset by using chi square tests to create multi-way splits. Preceded, and requires more data preparation than, CART.

chi-squared

A statistic that assesses how well a model fits the data. In data mining, it is most commonly used to find homogeneous subsets for fitting categorical trees as in CHAID.

Classification

The process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to specific variable(s) you are trying to predict. For example, a typical classification problem is to divide a database of companies into groups that are as homogeneous as possible with respect to a creditworthiness variable with values "Good" and "Bad" .

Classification tree

A decision tree that places categorical variables into classes.

Cleaning (cleansing)

Refers to a step in preparing data for a data mining activity. Obvious data errors are detected and corrected (e.g., improbable dates) and missing data is replaced.

Clustering

The process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to all available variables.

Confidence

Confidence of rule "B given A" is a measure of how much more likely it is that B occurs when A has occurred. It is expressed as a percentage; with 100% meaning B always occurs if A has occurred. Statisticians refer to this as the conditional probability of B given A. When used with association rules, the term confidence is observational rather than predictive. (Statisticians also use this term in an unrelated way. There are ways to estimate an interval and the probability that the interval contains the true value of a parameter is called the interval confidence. So a 95% confidence interval for the mean has a probability of .95 of covering the true value of the mean.)

Confusion matrix

A confusion matrix shows the counts of the actual versus predicted class values. It shows not only how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong.

Consequent

When an association between two variables is defined, the second item (or right-hand side) is called the consequent. For example, in the relationship "When a prospector buys a pick, he buys a shovel 14% of the time," "buys a shovel" is the consequent.

Continuous

Continuous data can have any value in an interval of real numbers. That is, the value does not have to be an integer. Continuous is the opposite of discrete or categorical.

Cross validation

A method of estimating the accuracy of a classification or regression model. The data set is divided into several parts, with each part in turn used to test a model fitted to the remaining parts.

Data mining

Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud



detection, evaluation of retail promotions, and credit risk analysis.

data navigation

The process of viewing different dimensions, slices, and levels of detail of a multidimensional database. See OLAP.

Data visualization

The visual interpretation of complex relationships in multidimensional data.

Data warehouse

A system for storing and delivering massive quantities of data.

Decision tree

A tree-shaped structure that represents a set of decisions. These decisions generate rules for the classification of a dataset. See CART and CHAID.

Degree of fit

A measure of how closely the model fits the training data. A common measure is r-square.

Deployment

After the model is trained and validated, it is used to analyze new data and make predictions. This use of the model is called deployment.

Dimension

In a flat or relational database, each field in a record represents a dimension. In a multidimensional database, a dimension is a set of similar entities; for example, a multidimensional sales database might include the dimensions Product, Time, and City.

Discriminant analysis

A statistical method based on maximum likelihood for determining boundaries that separate the data into categories.

Entropy

A way to measure variability other than the variance statistic. Some decision trees split the data into groups based on minimum entropy.

Exploratory data analysis

The use of graphical and descriptive statistical techniques to learn about the structure of a dataset.

Feed-forward

A neural net in which the signals only flow in one direction, from the inputs to the outputs.

Fuzzy logic

Fuzzy logic is applied to fuzzy sets where membership in a fuzzy set is a probability, not necessarily 0 or 1. Non-fuzzy logic manipulates outcomes that are either true or false. Fuzzy logic needs to be able to manipulate degrees of "maybe" in addition to true and false.

Genetic algorithms

A computer-based method of generating and testing combinations of possible input parameters to find the optimal output. It uses processes based on natural evolution concepts such as genetic combination, mutation and natural selection.

Hidden nodes

The nodes in the hidden layers in a neural net. Unlike input and output nodes, the number of hidden nodes is not predetermined. The accuracy of the resulting model is affected by the number of hidden nodes. Since the number of hidden nodes directly affects the number of parameters in the model, a neural net needs a sufficient number of hidden nodes to enable it to properly model the underlying behavior. On the other hand, a net with too many hidden nodes will overfit the data. Some neural net products include algorithms that search over a number of alternative neural nets by varying the number of hidden nodes, in the end choosing the model that gets the best results without over fitting.

**k-nearest neighbor**

A classification method that classifies a point by calculating the distances between the point and points in the training data set. Then it assigns the point to the class that is most common among its k-nearest neighbors (where k is an integer).

Kohonen feature map

A type of neural network that uses unsupervised learning to find patterns in data. In data mining it is employed for cluster analysis.

Layer

Nodes in a neural net are usually grouped into layers, with each layer described as input, output or hidden. There are as many input nodes as there are input (independent) variables and as many output nodes as there are output (dependent) variables. Typically, there are one or two hidden layers.

Leaf

A node not further split - the terminal grouping - in a classification or decision tree.

Logistic regression

A generalization of linear regression. It is used for predicting a binary variable (with values such as yes/no or 0/1). An example of its use is modeling the odds that a borrower will default on a loan based on the borrower's income, debt and age.

MARS

Multivariate Adaptive Regression Splines. MARS is a generalization of a decision tree.

Maximum likelihood

Another training or estimation method. The maximum likelihood estimate of a parameter is the value of a parameter that maximizes the probability that the data came from the population defined by the parameter.

MPP

Massively parallel processing, a computer configuration that is able to use hundreds or thousands of CPUs simultaneously. In MPP each node may be a single CPU or a collection of SMP CPUs. An MPP collection of SMP nodes is sometimes called an SMP cluster. Each node has its own copy of the operating system, memory, and disk storage, and there is a data or process exchange mechanism so that each computer can work on a different part of a problem. Software must be written specifically to take advantage of this architecture.

Noise

The difference between a model and its predictions. Sometimes data is referred to as noisy when it contains errors such as many missing or incorrect values or when there are extraneous columns.

OLAP

On-Line Analytical Processing tools give the user the capability to perform multidimensional analysis of the data.

Overlay

Data not collected by the organization, such as data from a proprietary database, that is combined with the organization's own data.

Precision

The precision of an estimate of a parameter in a model is a measure of how variable the estimate would be over other similar data sets. A very precise estimate would be one that did not vary much over different data sets. Precision does not measure accuracy. Accuracy is a measure of how close the estimate is to the real value of the parameter. Accuracy is measured by the average distance over different data sets of the estimate from the real value. Estimates can be accurate but not precise, or precise but not accurate. A precise but inaccurate estimate is usually biased, with the bias equal to the average distance from the real value of the parameter.

r-squared

A number between 0 and 1 that measures how well a model fits its training data. One is a perfect fit; however, zero implies the model has no predictive ability. It is computed as the covariance between the predicted and observed values divided by the standard deviations of the predicted and observed values.

**RAID**

Redundant Array of Inexpensive Disks. A technology for the efficient parallel storage of data for high-performance computer systems.

Retrospective data analysis

Data analysis that provides insights into trends, behaviors, or events that have already occurred.

Sequence discovery

The same as association, except that the time sequence of events is also considered. For example, "Twenty percent of the people who buy a VCR buy a camcorder within four months."

Significance

A probability measure of how strongly the data support a certain result (usually of a statistical test). If the significance of a result is said to be .05, it means that there is only a .05 probability that the result could have happened by chance alone. Very low significance (less than .05) is usually taken as evidence that the data mining model should be accepted since events with very low probability seldom occur. So if the estimate of a parameter in a model showed a significance of .01 that would be evidence that the parameter must be in the model.