Machine Learning, 30 (2/3), 271–274 (1998)

Glossary of Terms

RON KOHAVI AND FOSTER PROVOST ronnyk@cs.stanford.edu, foster@basit.com

Editors: Ron Kohavi and Foster Provost

To help readers understand common terms in machine learning, statistics, and data mining, we provide a glossary of common terms. The definitions are not designed to be completely general, but instead are aimed at the most common case.

Accuracy (error rate) The rate of correct (incorrect) predictions made by the model over a data set (cf. coverage). Accuracy is usually estimated by using an independent test set that was not used at any time during the learning process. More complex accuracy estimation techniques, such as cross-validation and the bootstrap, are commonly used, especially with data sets containing a small number of instances.

Association learning Techniques that find conjunctive implication rules of the form "X and $Y \rightarrow A$ and B" (associations) that satisfy given criteria. The conventional association algorithms are sound and complete methods for finding all associations that satisfy criteria for minimum support (at least a specified fraction of the instances must satisfy both sides of the rule) and minimum confidence (at least a specified fraction of instances satisfying the left hand side, or antecedent, must satisfy the right hand side, or consequent).

Attribute (field, variable, feature) A quantity describing an instance. An attribute has a domain defined by the attribute type, which denotes the values that can be taken by an attribute. The following domain types are common:

Categorical A finite number of discrete values. The type *nominal* denotes that there is no ordering between the values, such as last names and colors. The type *ordinal* denotes that there is an ordering, such as in an attribute taking on the values low, medium, or high.

Continuous (quantitative) Commonly, subset of real numbers, where there is a measurable difference between the possible values. Integers are usually treated as continuous in practical problems.

A *feature* is the specification of an attribute and its value. For example, color is an attribute. "Color is blue" is a feature of an example. Many transformations to the attribute set leave the feature set unchanged (for example, regrouping attribute values or transforming multi-valued attributes to binary attributes). Some authors use *feature* as a synonym for *attribute* (e.g., in feature-subset selection).

Classifier A mapping from unlabeled instances to (discrete) classes. Classifiers have a form (e.g., decision tree) plus an interpretation procedure (including how to handle unknowns, etc.). Some classifiers also provide probability estimates (scores), which can be thresholded to yield a discrete class decision thereby taking into account a utility function.

Confusion matrix A matrix showing the predicted and actual classifications. A confusion matrix is of size $\ell \times \ell$, where ℓ is the number of different label values. The following confusion matrix is for $\ell = 2$:

$\downarrow$ actual \ predicted$\rightarrow$	negative	positive
negative	a	b
positive	c	d

The following terms are defined for a 2×2 confusion matrix:

Accuracy $(a+d)/(a+b+c+d)$.

True positive rate (Recall, Sensitivity) $d/(c+d)$.

True negative rate (Specificity) $a/(a+b)$.

Precision $d/(b+d)$.

False positive rate $b/(a+b)$.

False negative rate $c/(c+d)$.

Coverage The proportion of a data set for which a classifier makes a prediction. If a classifier does not classify all the instances, it may be important to know its performance on the set of cases for which it is "confident" enough to make a prediction.

Cost (utility/loss/payoff) A measurement of the cost to the performance task (and/or benefit) of making a prediction $\hat{y}$ when the actual label is y. The use of accuracy to evaluate a model assumes uniform costs of errors and uniform benefits of correct classifications.

Cross-validation A method for estimating the accuracy (or error) of an inducer by dividing the data into k mutually exclusive subsets (the "folds") of approximately equal size. The inducer is trained and tested k times. Each time it is trained on the data set minus a fold and tested on that fold. The accuracy estimate is the average accuracy for the k folds.

Data cleaning/cleansing The process of improving the quality of the data by modifying its form or content, for example by removing or correcting data values that are incorrect. This step usually precedes the machine learning step, although the knowledge discovery process may indicate that further cleaning is desired and may suggest ways to improve the quality of the data. For example, learning that the pattern Wife $\rightarrow$ Female from the census sample at UCI has a few exceptions may indicate a quality problem.

Data mining The term data mining is somewhat overloaded. It sometimes refers to the whole process of knowledge discovery and sometimes to the specific machine learning phase.

Data set A schema and a set of instances matching the schema. Generally, no ordering on instances is assumed. Most machine learning work uses a single fixed-format table.

Dimension An attribute or several attributes that together describe a property. For example, a geographical dimension might consist of three attributes: country, state, city. A time dimension might include 5 attributes: year, month, day, hour, minute.

Error rate See Accuracy.

Example See Instance.

Feature See Attribute.

Feature vector (record, tuple) A list of features describing an instance.

Field See Attribute.

i.i.d. sample A set of independent and identically distributed instances.

Inducer / induction algorithm An algorithm that takes as input specific instances and produces a model that generalizes beyond these instances.

Instance (example, case, record) A single object of the world from which a model will be learned, or on which a model will be used (e.g., for prediction). In most machine learning work, instances are described by feature vectors; some work uses more complex representations (e.g., containing relations between instances or between parts of instances).

Knowledge discovery The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. This is the definition used in "Advances in Knowledge Discovery and Data Mining," 1996, by Fayyad, Piatetsky-Shapiro, and Smyth.

Loss See Cost.

Machine learning In Knowledge Discovery, *machine learning* is most commonly used to mean the application of induction algorithms, which is one step in the knowledge discovery process. This is similar to the definition of empirical learning or inductive learning in Readings in Machine Learning by Shavlik and Dietterich. Note that in their definition, training examples are "externally supplied," whereas here they are assumed to be supplied by a previous stage of the knowledge discovery process. *Machine Learning* is the field of scientific study that concentrates on induction algorithms and on other algorithms that can be said to "learn."

Missing value The value for an attribute is not known or does not exist. There are several possible reasons for a value to be missing, such as: it was not measured; there was an instrument malfunction; the attribute does not apply, or the attribute's value cannot be known. Some algorithms have problems dealing with missing values.

Model A structure and corresponding interpretation that summarizes or partially summarizes a set of data, for description or prediction. Most inductive algorithms generate models that can then be used as classifiers, as regressors, as patterns for human consumption, and/or as input to subsequent stages of the KDD process.

Model deployment The use of a learned model. *Model deployment* usually denotes applying the model to real data.

OLAP (MOLAP, ROLAP) On-Line Analytical Processing. Usually synonymous with MOLAP (multi-dimensional OLAP). OLAP engines facilitate the exploration of data along several (predetermined) dimensions. OLAP commonly uses intermediate data structures to store pre-calculated results on multidimensional data, allowing fast computations. ROLAP (relational OLAP) refers to performing OLAP using relational databases.

Record see Feature vector.

Regressor A mapping from unlabeled instances to a value within a predefined metric space (e.g., a continuous range).

Resubstitution accuracy (error/loss) The accuracy (error/loss) made by the model on the *training* data.

Schema A description of a data set's attributes and their properties.

Sensitivity True positive rate (see Confusion matrix).

Specificity True negative rate (see Confusion matrix).

Supervised learning Techniques used to learn the relationship between independent attributes and a designated dependent attribute (the label). Most induction algorithms fall into the supervised learning category.

Tuple See Feature vector.

Unsupervised learning Learning techniques that group instances without a prespecified dependent attribute. Clustering algorithms are usually unsupervised.

Utility See Cost.