Glossary

Note: This glossary is an extension to one compiled by Ron Kohavi and Foster Provost (1998), used with kind permission of Springer Science and Business Media.

a priori

A priori is a term borrowed from philosophy meaning "prior to experience." In data science, an *a priori* belief is one that is brought to the problem as background knowledge, as opposed to a belief that is formed after examining data. For example, you might say, "There is no *a priori* reason to believe that this relationship is linear." After examining data you might decide that two variables have a linear relationship (and so linear regression should work fairly well), but there was no reason to believe, from prior knowledge, that they should be so related. The opposite of *a priori* is *a posteriori*.

Accuracy (error rate)

The rate of correct (incorrect) predictions made by the model over a dataset (cf. coverage). Accuracy is usually estimated using an independent (holdout) dataset 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 datasets containing a small number of instances.

Association mining

Techniques that find conjunctive implication rules of the form "X and Y → A and B" (associations) that satisfy given criteria.

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 (symbolic)**: 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.

We do not differentiate in this book, but often the distinction is made that 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 multivalued attributes to binary attributes). In this book we follow the practice of many authors and practitioners, and use feature as a synonym for *attribute*.

Class (label)
: One of a small, mutually exclusive set of labels used as possible values for the target variable in a classification problem. Labeled data has one class label assigned to each example. For example, in a dollar bill classification problem the classes could be *legitimate* and *counterfeit*. In a stock assessment task the classes might be *will gain substantially*, *will lose substantially*, and *will maintain its value*.

Classifier
: A mapping from unlabeled instances to (discrete) classes. Classifiers have a form (e.g., classification tree) plus an interpretation procedure (including how to handle unknown values, etc.). Most classifiers also can provide probability estimates (or other likelihood scores), which can be thresholded to yield a discrete class decision thereby taking into account a cost/benefit or utility function.

Confusion matrix
: A matrix showing the predicted and actual classifications. A confusion matrix is of size $l \times l$, where l is the number of different label values. A variety of classifier evaluation metrics are defined based on the contents of the confusion matrix, including *accuracy, true positive rate, false positive rate, true negative rate, false negative rate, precision, recall, sensitivity, specificity, positive predictive value,* and *negative predictive value*.

Coverage
: The proportion of a dataset 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 dataset minus one of the folds and tested on that fold. The accuracy estimate is the average accuracy for the k folds or the accuracy on the combined ("pooled") testing 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 modeling step, although a pass through the data mining process may indicate that further cleaning is desired and may suggest ways to improve the quality of the data.

Data mining
: The term data mining is somewhat overloaded. It sometimes refers to the whole data mining process and sometimes to the specific application of modeling techniques to data in order to build models or find other patterns/regularities.

Dataset
: A schema and a set of instances matching the schema. Generally, no ordering on instances is assumed. Most data mining work uses a single fixed-format table or collection of feature vectors.

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 (error rate).

Example
: See Instance (example, case, record).

Feature
: See Attribute (field, variable, feature).

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.

Induction
: Induction is the process of creating a general model (such as a classification tree or an equation) from a set of data. Induction may be contrasted with deduction: deduction starts with a general rule or model and one or more facts, and creates other specific facts from them. Induction goes in the other direction: induction takes a collection of facts and creates a general rule or model. In the context of this book, model induction is synonymous with *learning* or *mining* a model, and the rules or models are generally statistical in nature.

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 data science work, instances are described by feature vectors; some work uses more complex representations (*e.g.*, containing relations between instances or between parts of instances).

KDD
: originally was an abbreviation for Knowledge Discovery from Databases. It is now used to cover broadly the discovery of knowledge from data, and often is used synonymously with data mining.

Knowledge discovery
: The nontrivial 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," by Fayyad, Piatetsky-Shapiro, & Smyth (1996).

Loss
: See Cost (utility/loss/payoff).

Machine learning
: In data science, machine learning is most commonly used to mean the application of induction algorithms to data. The term is often used synonymously with the modeling stage of the data mining 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 situation where 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 data mining process.

Model deployment
: The use of a learned model to solve a real-world problem. Deployment often is used specifically to contrast with the "use" of a model in the Evaluation stage of the data mining process. In the latter, deployment

usually is simulated on data where the true answer is known.

OLAP (MOLAP, ROLAP)
: Online Analytical Processing. Usually synonymous with MOLAP (multidimensional OLAP). OLAP engines facilitate the exploration of data along several (predetermined) dimensions. OLAP commonly uses intermediate data structures to store precalculated results on multidimensional data, allowing fast computations. ROLAP (relational OLAP) refers to performing OLAP using relational databases.

Record
: See Feature vector (record, tuple).

Schema
: A description of a dataset'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 (record, tuple).

Unsupervised learning
: Learning techniques that group instances without a pre-specified target attribute. Clustering algorithms are usually unsupervised.

Utility
: See Cost (utility/loss/payoff).