Glossary

Bias

Model bias expresses the error that the model makes (how different the prediction is from the training data). A high bias means that the model is too simple and failed to capture the relationship between the features and labels. High bias is a sign that the model is underfitting. This happens, for example, when you make the wrong modeling assumptions, such as training a model on data for which it is not suited.

Decision trees

A popular supervised learning algorithm that relies on recursively splitting the data into partitions. You can keep track of these partitions in a tree structure. During inference, an unlabeled example traverses the tree until it falls into a leaf. Each leaf is associated with one of the data partitions, and you assign the unlabeled example the most common label within that partition (or the average label in the case of regression).

Distance function

A special type of function used to determine nearness in k-nearest neighbors; typically defined between two points.

Entropy

A useful formula that measures dispersion or uncertainty of a discrete random variable; also used in decision trees.

Euclidean distance

The most commonly used distance function; it represents the straightest distance between two points.

Generalization

A model's ability to adapt to new, previously unseen data.

High-dimensional data

A data set with too many features. In such cases, it becomes difficult to train a model that can find the relationship between features and a label.

Hyperparameters

The "knobs" that you tweak during successive runs of training a model; they help guide the learning process. They are parameters in the model that are not learned but set prior to learning. Hyperparameters often trade off complexity vs. simplicity of models.

Information gain

A formula that measures the difference in the average entropy of a variable after segmenting the data into multiple partitions; also used in decision trees.

Instance-based learning

Another type of supervised learning model in which training examples are stored in memory. Those examples are utilized on demand to make predictions for a new, previously unseen example.

K-nearest neighbors

A commonly used supervised learning algorithm that makes the assumption that similar points of data share similar labels. The algorithm predicts the label of a test point through a majority vote among its k-nearest neighbors within the training set.

Loss

A measure of how far a model's predictions are from its label; to phrase it more pessimistically, a measure of how bad the model is. To determine this value, some models use a loss function.

Loss functions

Specialized mathematical functions that represent how well our models predict the labels; i.e., the "error."

Model calibration

Setting unique parameters achieved by using the measurements from the model's predictions. The parameters are used to provide a good description of the system's behavior.

Neighbor count (k)

The number of nearest neighbors to use in prediction.

Normalization

The methodology used to ensure features are on the same scale.

Overfitting

A model failure mode that occurs when a model is too complex. It learns the training data so closely that it does not generalize well to new data. An overfit model has low training error but poor generalization.

Regression

The process of predicting continuous numerical values of some quantity of interest for previously unseen input data based upon prior training of a model (a "regressor") using labeled data examples. One of three categories for predicted labels, y, used when y is a real value; for example, the price of a house. (The other two categories are binary classification and multi-class classification.)

Regression model

A type of model that outputs continuous (typically, floating-point) values. Compare with classification models which output discrete values such as "day lily" or "tiger lily."

Scikit-learn

Software for Python that has a wide range of algorithmic options, covering regression, classification, and unsupervised learning. It also provides rich libraries for data preparation, model selection, and evaluation.

Supervised learning

A class of machine learning problems in which labeled data are available, enabling an algorithm to learn how to associate data values with data labels so that predictive models for classification or regression on unseen data are possible.

Training data

A subset of data used in a supervised learning problem to fit or "train" a predictive model, which can then be used to make predictions about unseen data (e.g., in a testing set).

Underfitting

A model failure mode that occurs when the model is too simple. It is unable to learn important nuances in the training data to properly make predictions. An underfit model has high training error and poor generalization.

Variance

Model variance expresses how consistent the predictions of a model are if it is trained on different sections of the training data set. High variance is a sign that the model is overfitting to the particular data set on which it is trained.