Glossary

Bayes optimal classifier

The optimal classifier if the data distribution P(Y|X) was known. The Bayes optimal classifier predicts the most likely label, given a feature vector x. In practice it cannot be used if P(Y|X) is not known; however, it can be useful as a lower bound on the error or if P(Y|X) is approximated.

Binary classification

A type of classification that categorizes data instances into one of two groups. There are only two possible label values (e.g. +1,-1, or 0/1).

Data augmentation

Data augmentation is a way to artificially increase the size of your training data by augmenting data instances through label-preserving heuristics. For example, natural images can often be flipped horizontally or rotated slightly without changing the class membership.

Distance function

A distance function measures the dissimilarity between two input vectors according to some pre-specified metric. Examples are the Euclidean and Manhattan distance. A distance function is essential for the k-Nearest Neighbor classifier to retrieve the most similar training inputs for a given test point.

Features

The relevant characteristics or attributes that we believe may be predictive of a data instance's class membership. For example, the features we might collect to identify fraudulent bank transactions might include dollar amount, type of transaction, country of origin, frequency, etc.

High-dimensional data

Data is typically represented as vectors, where any single dimension contains a feature of a data point. The more attributes you collect, the more high-dimensional that vector gets. When we have data with very high-dimensional feature vectors — high-dimensional data — the k-Nearest Neighbor algorithm's performance may deteriorate due to the curse of dimensionality.

Hypothesis (Function)

The function or hypothesis is commonly denoted as "h" and represents the program that we learn from our training data. We apply this function to new data in order to make predictions during test time.

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 amongst its k-Nearest Neighbors within the training set.

Labels

The label is what you want to infer about a data point. Your training data has labels so that you can train your function to predict the label of test points.

Label preserving

A transformation is label preserving if it does not change the class membership of a given input.

Loss function

A loss function gives the computer a clear objective, measuring how many mistakes the selected function makes. A lower loss is always better, and a loss of zero is perfect.

Machine learning

The science of how to make computers learn from experience.

Multiclass classification

A type of classification that sorts an element of data into more than two groups of labels (e.g., class1="red", class2="blue", class3="yellow").

NumPy

A Python library specialized for linear algebra operations. It is really fast and convenient.

Regression

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 multiclass classification.)

Supervised learning

A type of machine learning in which a specific label y is predicted based on a specific data set of labeled features.