Unit 1 Glossary

Array

Array is a collection of numbers of a given type, such as float or int, in one or more dimensions.

Axis

A particular dimension (or direction) in an array or a DataFrame.

Broadcasting

Broadcasting is a NumPy feature that enables mathematical operations to be applied to arrays of different sizes and dimensions.

Cell

A unit of structure in a Jupyter Notebook that can contain multiple lines of code to be run as a unit.

Classification

One of two classes of methods in supervised learning, where the label is a categorical value. The two types of classification are binary classification and multi-class classification.

Cross Industry Standard Process for Data Mining (CRISP-DM)

A popular diagram used to represent the process used for building machine learning models.

DataFrame

DataFrame refers to a data table or spreadsheet with row and column headers, where each column contains data of a particular type but which can be of different types in different columns.

Data matrix

A data matrix is a structured table consisting of rows and columns.

Ethical risk

The likelihood for large scale and automated decision systems to cause unintended harms.

Example

An example is an instance of data. It can also be called a data point.

Features

Features are input variables. These variables are predictive data elements of a machine learning problem. They are the data contained in the columns of a data matrix. One feature value is contained in one column.

Generalization

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

Jupyter Notebook

A web-based interpreter which ties together code, analyses, documentation, and graphics.

Label

In supervised learning, the "answer" or "result" portion of an example. Each example in a labeled data set consists of one or more features and a label. For instance, in a housing data set, the features might include the number of bedrooms, the number of bathrooms, and the age of the house, while the label might be the house's price. In a spam detection data set, the features might include the subject line, the sender, and the email message itself, while the label would probably be either "spam" or "not spam."

Labeled example

A labeled example contains features and a label.

Labels

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.

Library

A library is a set of related software items (e.g., functions, objects) that can be called from within a program but which are defined externally to that program, typically to provide a defined set of operations that are useful to a variety of different applications.

Machine Learning (ML)

A broad class of methods and algorithms for building predictive models from data without prescribing the specific form of relationships between inputs and outputs. Machine learning is considered a subfield in the larger field of artificial intelligence but also straddles the world of data science, which is an amalgamation of human insight and automated inference.

Machine Learning Model

A computer program that has been trained to recognize patterns in data to make predictions on future data.

Notebook

A computational environment that generally combines code, documentation, results, and graphics. Jupyter Notebooks are a widely used platform that supports work with Python as well as several other programming languages.

Package

Within the Python ecosystem, a collection of related software items that are bundled and distributed together to provide specific functionality within a Python program. A package might simply be a library (and sometimes the terms are synonymous), or it might contain additional tools beyond a library that support working with Python.

Regression

One of two classes of methods in supervised learning, where the label is any real valued number.

Recommendation systems

Machine learning systems designed to recommend items to you on various websites and apps.

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 is possible.

System risk

System risk is the likelihood for complex and dynamic systems to have failure points.

Training

Training is either creating or learning the model.

Unlabeled example

An unlabeled example contains only features and no label.

Unsupervised learning

A class of machine learning problems in which labeled data are not available, whereby algorithms work to identify various types of patterns in data.

Vectorization

Vectorization is a NumPy feature that performs operations on entire arrays that would normally be performed through the use of loops.