

# Introduction to Machine Learning in Python

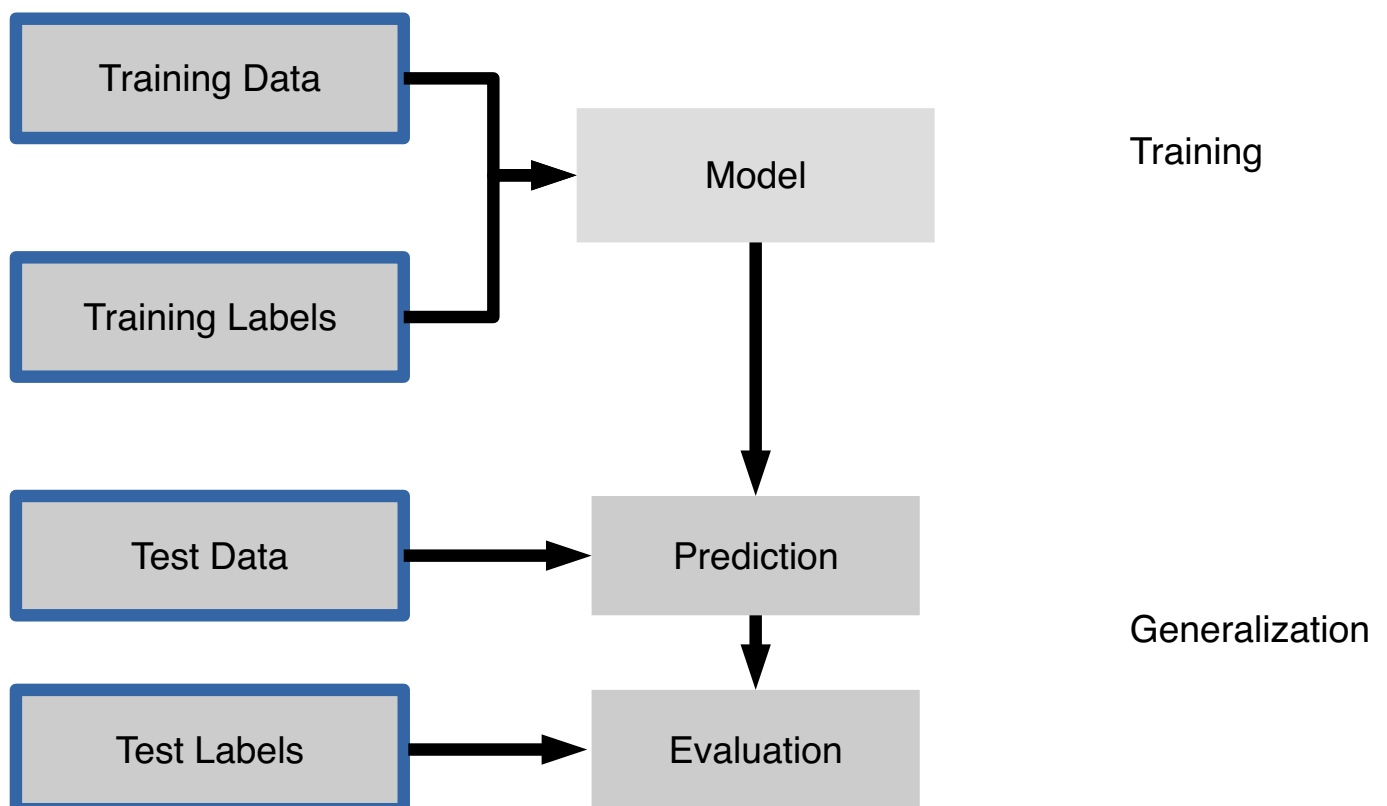
## What is Machine Learning?

Machine learning is the process of extracting knowledge from data automatically, usually with the goal of making predictions on new, unseen data. A classical example is a spam filter, for which the user keeps labeling incoming mails as either spam or not spam. A machine learning algorithm then "learns" a predictive model from data that distinguishes spam from normal emails, a model which can predict for new emails whether they are spam or not.

Central to machine learning is the concept of **automating decision making** from data **without the user specifying explicit rules** how this decision should be made.




For the case of emails, the user doesn't provide a list of words or characteristics that make an email spam. Instead, the user provides examples of spam and non-spam emails that are labeled as such.

The second central concept is **generalization**. The goal of a machine learning model is to predict on new, previously unseen data. In a real-world application, we are not interested in marking an already labeled email as spam or not. Instead, we want to make the user's life easier by automatically classifying new incoming mail.



The data is presented to the algorithm usually as a two-dimensional array (or matrix) of numbers. Each data point (also known as a *sample* or *training instance*) that we want to either learn from or make a decision on is represented as a list of numbers, a so-called feature vector, and its containing features represent the properties of this point.

Later, we will work with a popular dataset called *Iris* -- among many other datasets. Iris, a classic benchmark dataset in the field of machine learning, contains the measurements of 150 iris flowers from 3 different species: Iris-Setosa, Iris-Versicolor, and Iris-Virginica.

Species	Image
Iris Setosa	
Iris Versicolor	
Iris Virginica	

We represent each flower sample as one row in our data array, and the columns (features) represent the flower measurements in centimeters. For instance, we can represent this Iris dataset, consisting of 150 samples and 4 features, a 2-dimensional array or matrix  $\mathbb{R}^{150 \times 4}$  in the following format:

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \dots & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \dots & x_4^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & \dots & x_4^{(150)} \end{bmatrix}.$$

(The superscript denotes the  $i$ th row, and the subscript denotes the  $j$ th feature, respectively.)

There are two kinds of machine learning we will talk about today: **supervised learning** and **unsupervised learning**.

## Supervised Learning: Classification and regression

In **Supervised Learning**, we have a dataset consisting of both input features and a desired output, such as in the spam / no-spam example. The task is to construct a model (or program) which is able to predict the desired output of an unseen object given the set of features.

Some more complicated examples are:

- Given a multicolor image of an object through a telescope, determine whether that object is a star, a quasar, or a galaxy.
- Given a photograph of a person, identify the person in the photo.
- Given a list of movies a person has watched and their personal rating of the movie, recommend a list of movies they would like.
- Given a person's age, education and position, infer their salary

What these tasks have in common is that there is one or more unknown quantities associated with the object which needs to be determined from other observed quantities.

Supervised learning is further broken down into two categories, **classification** and **regression**:

- **In classification, the label is discrete**, such as "spam" or "no spam". In other words, it provides a clear-cut distinction between categories. Furthermore, it is important to note that class labels are nominal, not ordinal variables. Nominal and ordinal variables are both subcategories of categorical variable. Ordinal variables imply an order, for example, T-shirt sizes "XL > L > M > S". On the contrary, nominal variables don't imply an order, for example, we (usually) can't assume "orange > blue > green".
- **In regression, the label is continuous**, that is a float output. For example, in astronomy, the task of determining whether an object is a star, a galaxy, or a quasar is a classification problem: the label is from three distinct categories. On the other hand, we might wish to estimate the age of an object based on such observations: this would be a regression problem, because the label (age) is a continuous quantity.

In supervised learning, there is always a distinction between a **training set** for which the desired outcome is given, and a **test set** for which the desired outcome needs to be inferred. The learning model fits the predictive model to the training set, and we use the test set to evaluate its generalization performance.

## Unsupervised Learning

In **Unsupervised Learning** there is no desired output associated with the data. Instead, we are interested in extracting some form of knowledge or model from the given data. In a sense, you can think of unsupervised learning as a means of discovering labels from the data itself. Unsupervised learning is often harder to understand and to evaluate.

Unsupervised learning comprises tasks such as *dimensionality reduction*, *clustering*, and *density estimation*. For example, in the iris data discussed above, we can use unsupervised methods to determine combinations of the measurements which best display the structure of the data. As we'll see below, such a projection of the data can be used to visualize the four-dimensional dataset in two dimensions. Some more involved unsupervised learning problems are:

- Given detailed observations of distant galaxies, determine which features or combinations of features summarize best the information.
- Given a mixture of two sound sources (for example, a person talking over some music), separate the two (this is called the [blind source separation](http://en.wikipedia.org/wiki/Blind_signal_separation) problem).
- Given a video, isolate a moving object and categorize in relation to other moving objects which have been seen.
- Given a large collection of news articles, find recurring topics inside these articles.
- Given a collection of images, cluster similar images together (for example to group them when visualizing a collection)

Sometimes the two may even be combined: e.g. unsupervised learning can be used to find useful features in heterogeneous data, and then these features can be used within a supervised framework.

### (simplified) Machine learning taxonomy

