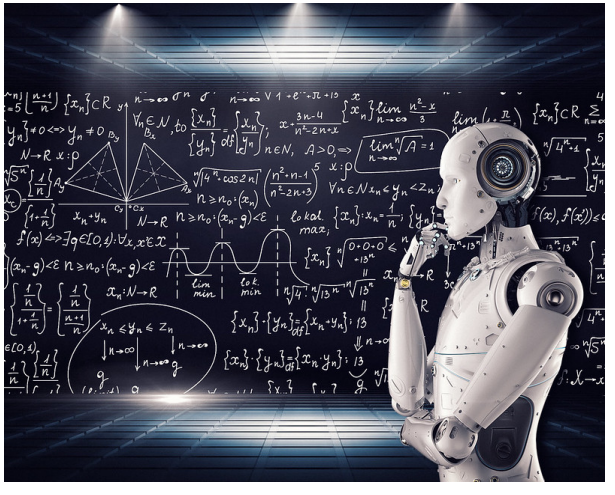
Artificial Intelligence

Supervised Learning



source

Lluís Talavera, 2022

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

ETSEIB

Classification

Given: a dataset of examples divided into classes,

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
6.1	2.9	4.7	1.4	Iris-versicolor
5.6	2.9	3.6	1.3	Iris-versicolor
7.6	3.0	6.6	2.1	Iris-virginica

150 examples (50 per class), 4 columns*

learn how to assign a class label to new, unseen examples.

sepal length	sepal width	petal length	petal width	class
5.7	3.8	1.7	0.3	???

We will learn (build) a **model** to make the predictions.

* *Source* : *Iris* problem UCI repository (Frank & Asunción, 2010)

Terminology

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
6.1	2.9	4.7	1.4	Iris-versicolor
5.6	2.9	3.6	1.3	Iris-versicolor
7.6	3.0	6.6	2.1	Iris-virginica

- The class or target is the column to predict, usually referred to as vector y. It is a categorical value.
- The rest of the columns are called attributes, features, predictors and form a matrix usually referred to as X. The `sklearn` library only accepts numerical values.
- Each row is called an example, an instance o an observation.

What is a model?

It is a term used in different disciplines, with different meanings.

In ML, can be viewed as an *abstraction* or *summary* of the data that can be used to

- make predictions
- discover patterns in data

Many different forms: an equation, a probability distribution, a data structure, a set of rules...

For example, a linear regression equation:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Regression

Given: a dataset of examples that include a numeric target column

density	pH	sulphates	alcohol	quality
0.998	3.16	0.58	9.8	6
0.9948	3.51	0.43	11.4	4
0.9973	3.35	0.86	12.8	8
0.9994	3.16	0.63	8.4	3
0.99514	3.44	0.68	10.55	7

1599 examples & 12 columns (11 attributes + 1 target)*

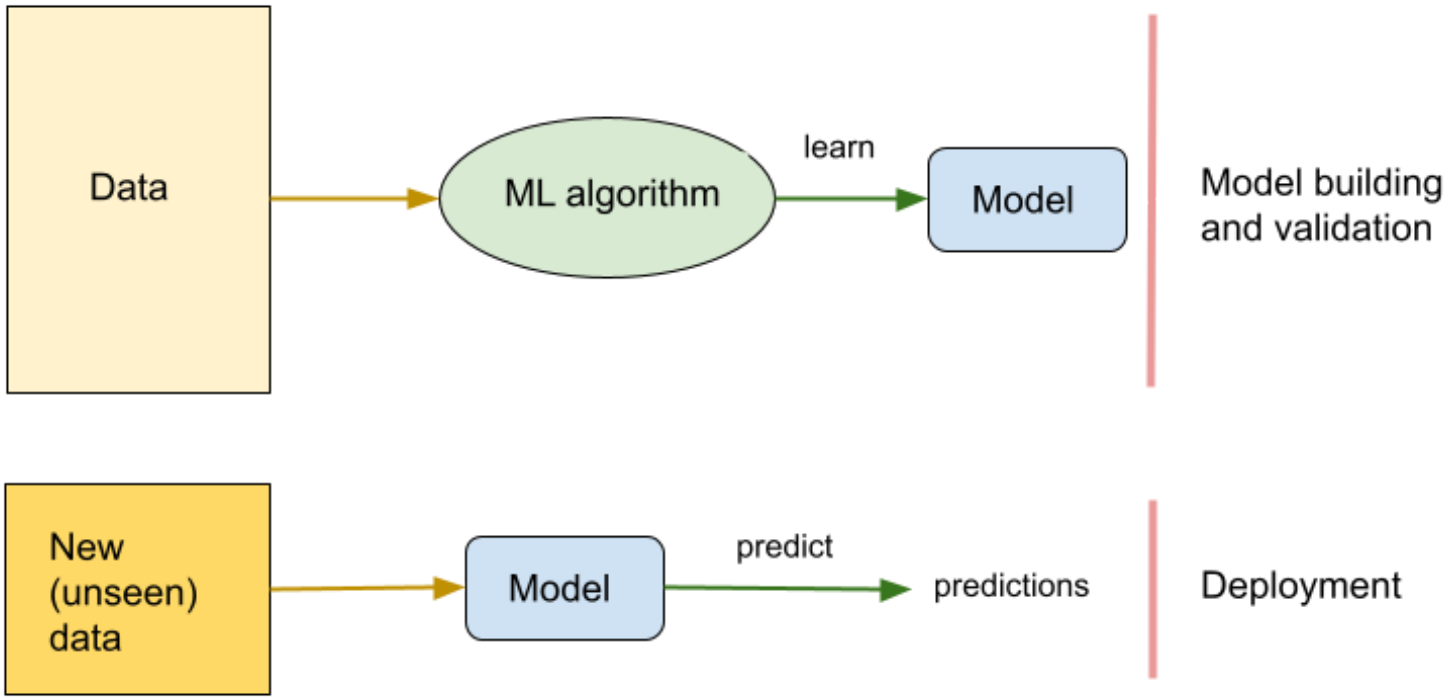
learn how to predict the numeric value of the target for new, unseen examples.

density	pH	sulphates	alcohol	quality
0.9978	3.51	0.56	9.4	???

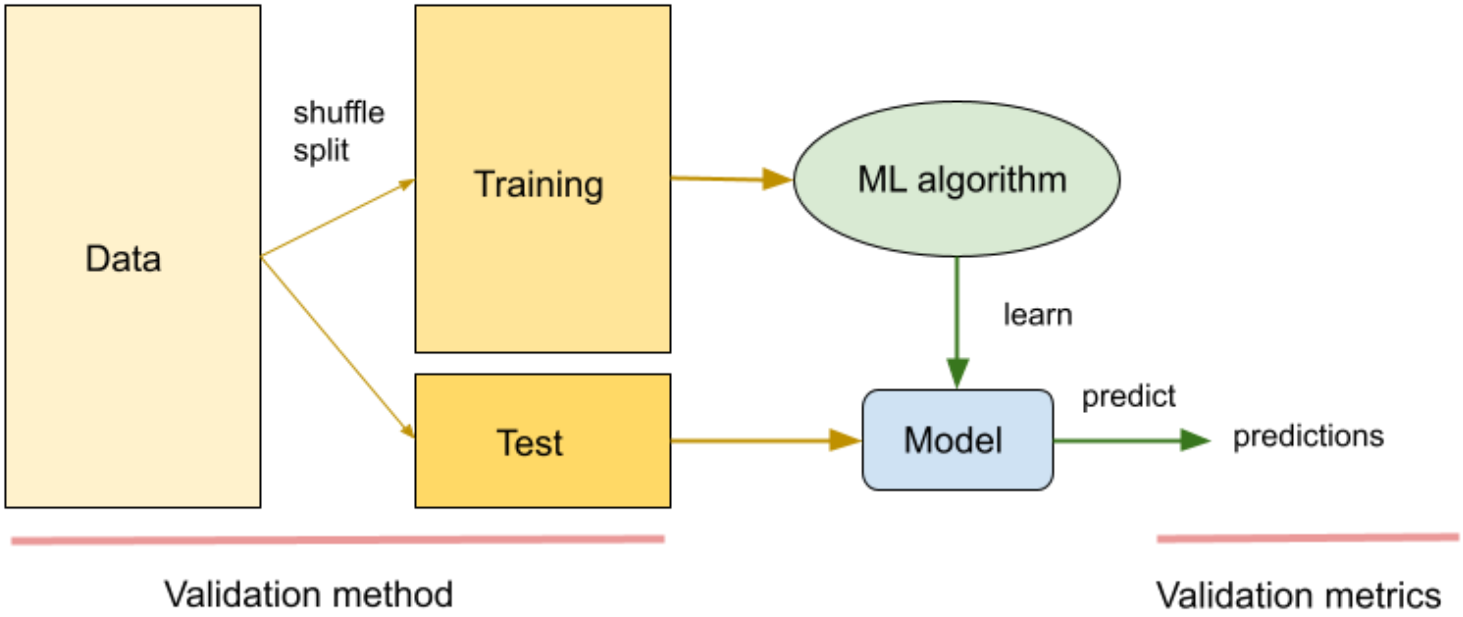
The diference between classification and regression is that classification predicts categorical values while regression predicts numerical quantities.

* *Source* : *wine quality* problem from UCI repository (Frank & Asunción, 2010)

Production architecture



Validation



Later, we will study validation in more detail. For now, we will use this method (**holdout**) and **accuracy** as the metric.

$accuracy = \frac{\text{number of correct predictions}}{\text{total number of predictions}}$