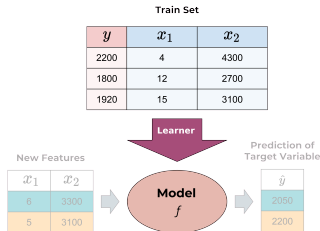


Introduction to Machine Learning

ML-Basics: Learner



Learning goals

- Understand that a supervised learner fits models automatically from training data

SUPERVISED LEARNING EXAMPLE

Imagine we want to investigate how working conditions affect productivity of employees.

- It is a **regression** task since the target *productivity* is continuous.
- We collect data about worked minutes per week (*productivity*), how many people work in the same office as the employee in question, and the employee's salary.

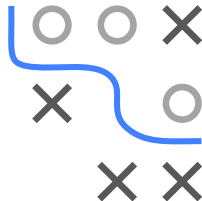
| Features x | | Target y |
|------------------------------------|--------------------------|---------------------------------------|
| People in Office (Feature 1) x_1 | Salary (Feature 2) x_2 | Worked Minutes Week (Target Variable) |
| 4 | 4300 € | 2220 |
| 12 | 2700 € | 1800 |
| 5 | 3100 € | 1920 |

Diagram illustrating the data structure for a regression problem. The data is organized into a table with 3 rows (labeled $n = 3$) and 3 columns. The first two columns represent features x (labeled $p = 2$), and the third column represents the target variable y .

- Feature 1 (x_1):** People in Office. Values: 4, 12, 5.
- Feature 2 (x_2):** Salary (in €). Values: 4300 €, 2700 €, 3100 €.
- Target Variable (y):** Worked Minutes Week. Values: 2220, 1800, 1920.

The diagram also shows the vector notation for the features and target variable:

- $x^{(2)}$ (Feature 1 vector)
- $x_2^{(1)}$ (Feature 2 vector)
- $y^{(3)}$ (Target variable vector)



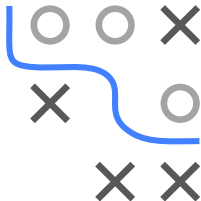
SUPERVISED LEARNING EXAMPLE / 2

How could we construct a model from these data?

We could investigate the data manually and come up with a simple, hand-crafted rule such as:

- The baseline productivity of an employee with salary 3000 and 7 people in the office is 1850 minutes
- A decrease of 1 person in the office increases productivity by 30
- An increase of the salary by 100 increases productivity by 10

=> Obviously, this is neither feasible nor leads to a good model



IDEA OF SUPERVISED LEARNING

Goal: Automatically identify the fundamental functional relation in the data that maps an object's features to the target.

- **Supervised** learning means we make use of *labeled* data for which we observed the outcome.
- We use the labeled data to learn a model f .
- Ultimately, we use our model to compute predictions for **new** data whose target values are unknown.



LEARNER DEFINITION

- The algorithm for finding our f is called **learner**. It is also called **learning algorithm** or **inducer**.
- We prescribe a certain hypothesis space, the learner is our means of picking the best element from that space for our data set.
- Formally, it maps training data $\mathcal{D} \in \mathbb{D}$ (plus a vector of **hyperparameter** control settings $\lambda \in \Lambda$) to a model:

$$\mathcal{I} : \mathbb{D} \times \Lambda \rightarrow \mathcal{H}$$

