Introduction to Machine Learning

Introduction: Supervised Learning

compstat-lmu.github.io/lecture_i2ml

IDEA OF SUPERVISED LEARNING

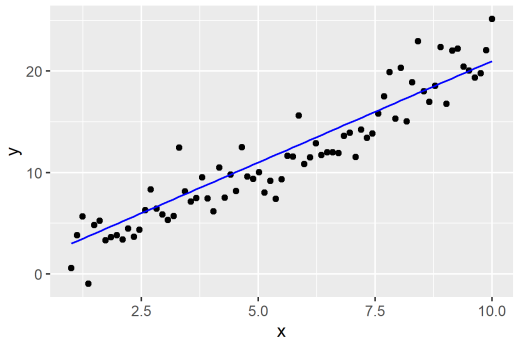
- **Goal:** Identify the fundamental relations in the data that map the features to the target.
  $\rightarrow$ This allows us to make **predictions** for new data.
- Ideally, we would have full knowledge about the data-generating process and thus be able to specify this mapping function precisely.
- However, since this is basically impossible, we must try to **learn** the mapping function: for objects exhibiting certain patterns or properties, certain outcomes are much more likely.
  $\rightarrow$ We call such an assumed mapping a **model** f.

SUPERVISED LEARNING

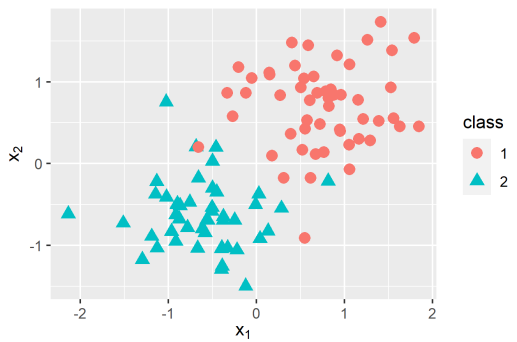
- Humans are pretty good at learning, at least for some domains. In machine learning, we rely on computers, which is why the model itself, as well as all feature and target values, need to be computable.
- In **supervised** learning we make use of **labeled** data, i.e., observations for which we already know the target outcome.
- We try to construct f automatically from an example set of such labeled objects.
- Knowing the “truth” allows us to test how well we have grasped the nature of the underlying mapping by comparing our predictions to the actually observed values.

TASKS IN SUPERVISED LEARNING

- In general, supervised learning comes in two flavors we call **tasks**:
  - **Regression**: Given features **x**, predict corresponding output from $\mathcal{Y} \in \mathbb{R}^m$.

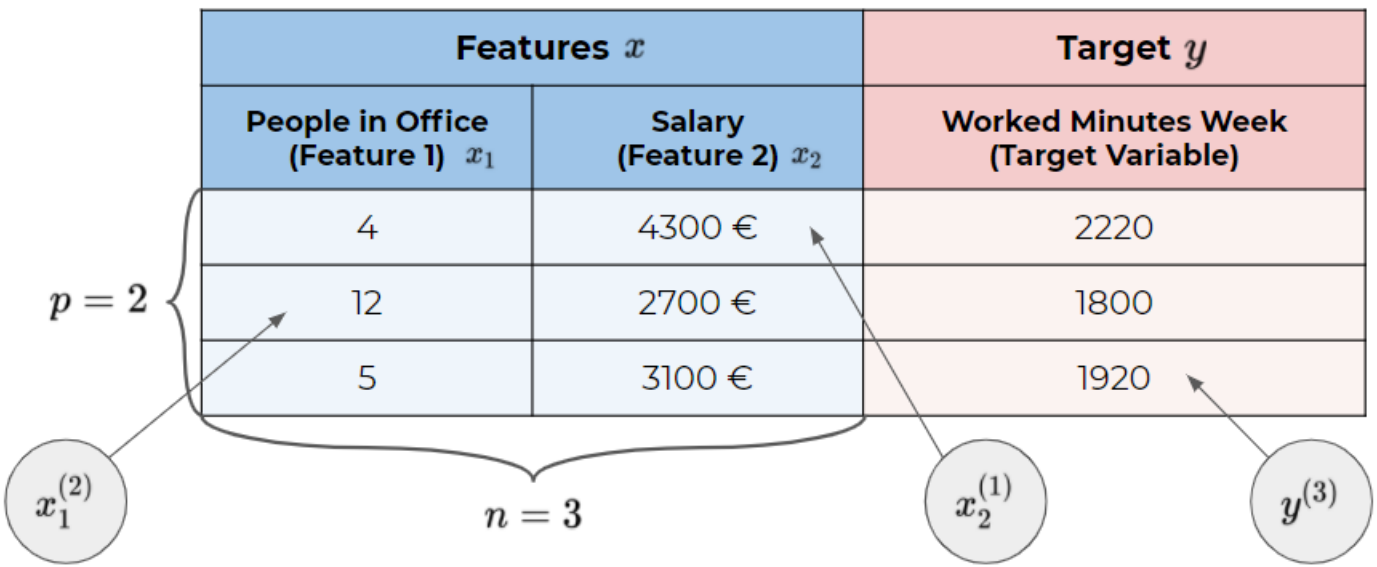


  - **Classification**: Assign an observation with features **x** to one class of a finite set of classes $\mathcal{Y} = \{C_1, ..., C_g\}, g \geq 2$ (details later).



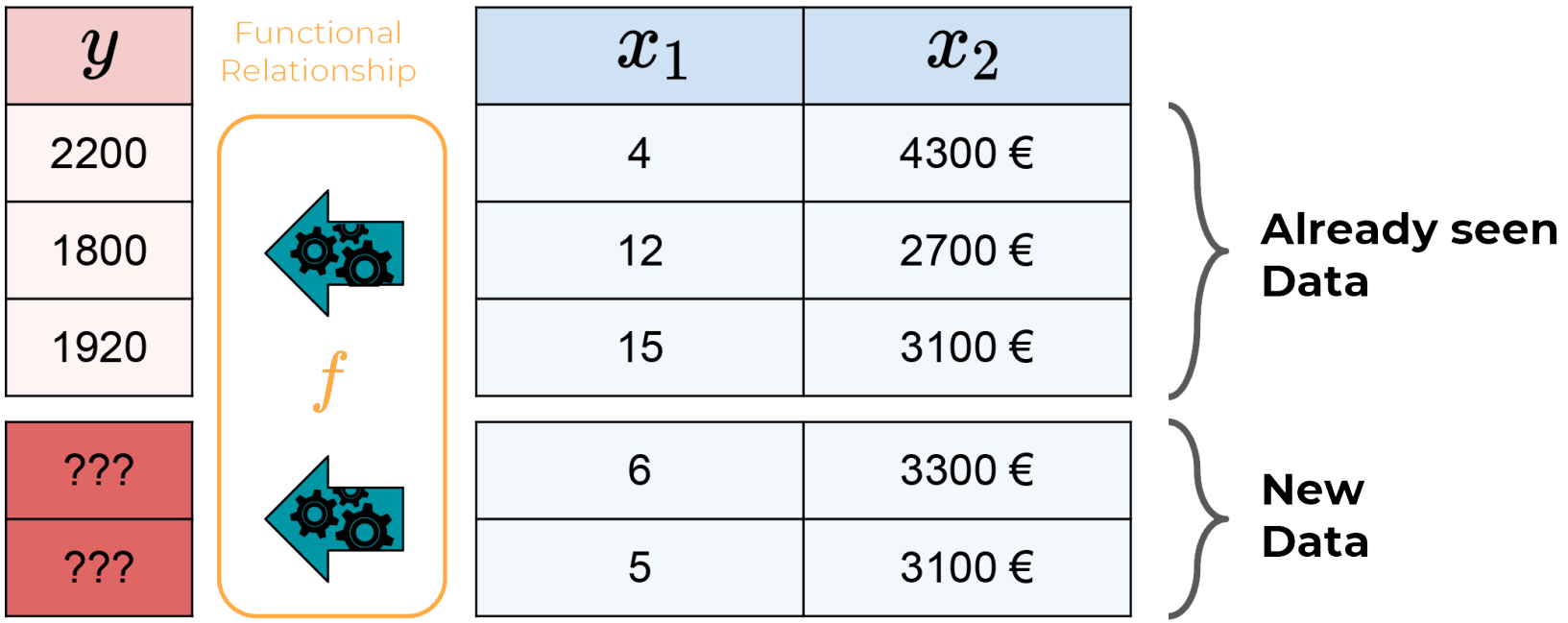
REGRESSION TASKS: EXAMPLE

Imagine you want to investigate how salary and workplace conditions (*features*) affect productivity of employees (*target*) – a standard **regression** task. Therefore, you collect data about their worked minutes per week (productivity), how many people work in the same office as the employees in question, and the employees' salary.



REGRESSION TASKS: EXAMPLE

- For our observed data we know which outcome is produced.
- For new employees can only observe the features but not the target.



MORE REGRESSION TASKS

1. **Predict house prices**
   - **Aim**: Predict the price for a house in a certain area
   - **Features**: e. g.
     - square footage
     - number of bedrooms
     - swimming pool yes/no
2. **Predict the length-of-stay in a hospital at the time of admission**
   - **Aim**: Predict the number of days a single patient has to stay in hospital
   - **Features**: e. g.
     - diagnosis category (heart disease, injury,...)
     - admission type (urgent, emergency, newborn,...)
     - age
     - gender

CLASSIFICATION TASKS: EXAMPLE

- Imagine you work for an insurance company which **classifies** its life insurance customers according to five risk categories, depending on which insurance premiums are charged.
- You might use features such as
  - job type (white collar, carpenter, stuntman, ...)
  - age
  - smoking behavior

  to perform this classification.

PARAMETERS, STATISTICS AND SUPERVISED ML

- Supervised ML additionally assumes that f is of a certain "form" or comes from a certain **class of functions**.
  This is necessary to make the problem of automatically finding a "good" model feasible at all.
- The specific behavior of a mapping from this class can then be described by **parameters** that define its shape.
- Statistics, too, studies how to learn such functions (or, rather: their parameters) from example data and how to perform inference on them and interpret the results.
- For historical reasons though, statistics is mostly focused on fairly simple classes of mappings, like (generalized) linear models.
- Supervised ML also includes more complex kinds of mappings that can typically deal with more complicated and high-dimensional inputs.

SUMMARY

Supervised machine learning is concerned with learning a function that predicts a certain **target** from an object's **features** from a set of examples for which both the features and the target are known.
The function to be learned is restricted to come from a certain class of functions and its precise shape is defined in terms of a set of **parameters**.