Introduction to Machine Learning

Introduction: Data

compstat-lmu.github.io/lecture_i2ml

DATA IN MACHINE LEARNING

- The data we deal with in machine learning usually consists of observations on different aspects of objects:
  - **Target** variable(s): the attribute(s) of interest
  - **Features**: measurable properties that provide a concise description of the object
  - Both features and target variables may be of different data types (categorical, numeric, ...).
- We assume some kind of relationship between the features and the target, in a sense that the value of the target variable can be explained by a combination of the features.

DATA IN MACHINE LEARNING

- Imagine, for instance, you want to investigate how salary and workplace conditions (*features*) affect productivity of employees (*target*). 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.

		Features x			
Worked Minutes Week (Target Variable)	y	**People in Office (Feature 1)**	x_1	**Salary (Feature 2)**	x_2
2220	$y^{(1)}$	4	$x_1^{(1)}$	4300 €	$x_2^{(1)}$
1800	$y^{(2)}$	12	$x_1^{(2)}$	2700 €	$x_2^{(2)}$
1920	$y^{(3)}$	5	$x_1^{(3)}$	3100 €	$x_2^{(3)}$

$n = 3$

$p = 2$

- In practical applications we frequently encounter high-dimensional data, i.e., data with many features and/or observations.

NOTATION FOR DATA

In formal notation, the data sets we are given are of the following form:

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right) \right\} \subset (\mathcal{X} \times \mathcal{Y})^n .$$

We call

- $\mathcal{X}$ the input space with $p = \dim(\mathcal{X})$ (for now: $\mathcal{X} \subset \mathbb{R}^p$),
- $\mathcal{Y}$ the output / target space,
- the tuple $\left(\mathbf{x}^{(i)}, y^{(i)}\right) \in \mathcal{X} \times \mathcal{Y}$ the *i*-th observation,
- $\mathbf{x}_j = \left(x_j^{(1)}, \ldots, x_j^{(n)}\right)^T$ the j-th feature vector.

DATA-GENERATING PROCESS

- We assume the observed data $\mathcal{D}$ to be generated by a process that can be characterized by some probability distribution

$$\mathbb{P}_{xy},$$

  defined on $\mathcal{X} \times \mathcal{Y}$.
- Depending on the context, we denote the random variables following this distribution by $\mathbf{x}$ and y.
- Usually we assume the data to be drawn i.i.d. from the joint probability density function (pdf) / probability mass function (pmf) $p(\mathbf{x}, y)$.

DATA-GENERATING PROCESS

Remarks:

- With a slight abuse of notation we write random variables, e.g., $\mathbf{x}$ and y, in lowercase, as normal variables or function arguments. The context will make clear what is meant.
- Often, distributions are characterized by a parameter vector $\boldsymbol{\theta} \in \Theta$. We then write $p(\mathbf{x}, y \mid \boldsymbol{\theta})$.
- This lecture mostly takes a frequentist perspective. Distribution parameters $\boldsymbol{\theta}$ appear behind the | for improved legibility, not to imply that we condition on them in a probabilistic Bayesian sense. So, strictly speaking, $p(\mathbf{x}|\boldsymbol{\theta})$ should usually be understood to mean $p_{\boldsymbol{\theta}}(\mathbf{x})$ or $p(\mathbf{x}, \boldsymbol{\theta})$ or $p(\mathbf{x}; \boldsymbol{\theta})$.