

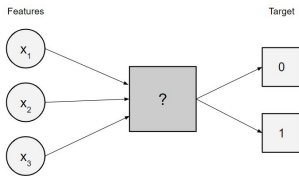
Introduction to Machine Learning

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.



Features x				Target y
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4.3	3.0	1.1	0.1	setosa
5.0	3.3	1.4	0.2	setosa
7.7	3.8	6.7	2.2	virginica
4.4	3.2	1.3	0.2	setosa
5.9	3.0	5.1	1.8	virginica

DATA IN MACHINE LEARNING

- For instance, it is reasonable to assume a relationship between certain features of a job-seeker, such as their field of expertise, academic qualifications and previous job experiences, and their salary.

Your skills impact your salary

Find Skills

Skill Name

Add

Related Skills

Value

⊕ Data science	+ 12%
⊕ Machine learning	+ 9%
⊕ SAS/MACROS	+ 7%
⊕ Clinical trials	+ 7%
⊕ Modeling	+ 6%
⊕ Business ...	+ 6%
⊕ Statistical models	+ 3%
⊕ Biostatistics	+ 3%
⊕ Marketing analytics	+ 3%
⊕ Pharmaceuticals	+ 3%

Statistician Salary Prediction

New York, NY

0 Years of Experience

Skills included in this prediction

R Data analysis SAS Statistics SQL

Does this salary look accurate? [Help us improve it!](#)

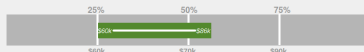
Your Salary Prediction ⓘ

\$60,500 - \$86,000

See how you compare to all other Statistician salaries nationwide

Statistician Range

Your Prediction ⓘ



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

DATA LABELS

- We distinguish two basic forms our data may come in:
 - For **labeled** data we have already observed the target values (*labels*).
 - For **unlabeled** data these remain unknown.
- It is easy to see how labeled data are vastly more informative.
- In practice, however, we will much more frequently encounter the unlabeled sort.
- Figure from slide 1 with some target values, some "?"

NOTATION FOR DATA

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

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

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 $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y}$ the i -th observation,
- $\mathbf{x}_j = \left(x_j^{(1)}, \dots, x_j^{(n)} \right)^T$ 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 .
- It is important to understand that the true distribution is essentially **unknown** to us.

DATA-GENERATING PROCESS

- Usually we assume the data to be drawn *i.i.d.* from the joint probability density function (pdf) / probability mass function (pmf) $p(\mathbf{x}, y)$.
 - i.i.d. stands for independent and identically distributed.
 - We presuppose that all samples are drawn from the same distribution and are mutually independent – the i -th realization does not depend on the previous $i - 1$ ones.
 - It is a strong yet crucial assumption that is precondition to many theoretical implications (e.g., the Central Limit Theorem).
- **FIGURE**

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 $\theta \in \Theta$. We then write $p(\mathbf{x}, y \mid \theta)$.
- This lecture mostly takes a frequentist perspective. Distribution parameters θ appear behind the \mid for improved legibility, not to imply that we condition on them in a probabilistic Bayesian sense. So, strictly speaking, $p(\mathbf{x} \mid \theta)$ should usually be understood to mean $p_\theta(\mathbf{x})$ or $p(\mathbf{x}, \theta)$ or $p(\mathbf{x}; \theta)$.