Introduction to Machine Learning

Introduction: Tasks & Data

compstat-lmu.github.io/lecture_i2ml

SUPERVISED TASKS AND DATA

Supervised Learning comes in two flavours:

- **Regression**: Given features x, predict corresponding output from $\mathcal{Y} \in \mathbb{R}^m$.
- **Classification**: Assigning 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 TASK - INCOME PREDICTION

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%
⊕ Pharmaceutics	+ 3%

Statistician Salary Prediction
New York , NY
0 Years of Experience

Skills included in this prediction

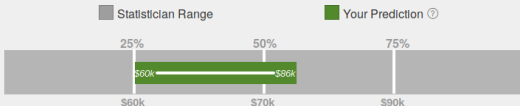R ✕   Data analysis ✕   SAS ✕   Statistics ✕   SQL ✕

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 ⓘ

25%	50%	75%

$60k — $86k

| $60k | $70k | $90k |

MORE REGRESSION TASKS

❶ 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

❷ 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

DATA

Imagine you want to investigate how salary and workplace conditions affect productivity of employees. 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.

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

$$p = 2$$

TARGET AND FEATURES RELATIONSHIP

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

y
2200
1800
1920

x_1	x_2
4	4300 €
12	2700 €
15	3100 €

Already seen Data

???
???

6	3300 €
5	3100 €

New Data

$\Longrightarrow$ The goal is to predict the target variable for **unseen new data** by using a **model** trained on the already seen **training data**.

NOTATION FOR DATA

		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 supervised machine learning, we are given a dataset

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

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 (e.g. $\mathcal{Y} = \mathbb{R}$ for regression or $\mathcal{Y} = \{C_1, ..., C_g\}$, $g \geq 2$, for classification),
- the tuple $\left(\mathbf{x}^{(i)}, y^{(i)} \right) \in \mathcal{X} \times \mathcal{Y}$ the *i*-th observation,
- $\mathbf{x}_j = \left(x_j^{(1)}, ..., x_j^{(n)} \right)$ the j-th feature vector

DATA GENERATING PROCESS

- We assume that a probability distribution

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

  is defined on $\mathcal{X} \times \mathcal{Y}$ that characterizes the process that generates the observed data $\mathcal{D}$.

- Depending on the context, we denote the random variables following this distribution by **x** and y.

- Usually, we assume that the data is 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., $\boldsymbol{x}$ 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 $p(\mathbf{x}, y \mid \boldsymbol{\theta})$.

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