

## Notation

### Darstellung von Daten

- Wir unterscheiden grundsätzlich zwischen:
  - Eigenschaften* (engl. *Features*)  $x$
  - Zielgrößen* (engl. *Targets*)  $y$ .
- Wir bezeichnen einen *Datenpunkt* mit  $\underline{x}^{(i)}$  und beschreiben diesen als Zeilenvektor bestehend aus  $j=1\dots N$  Eigenschaften:  $\underline{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \dots, x_N^{(i)}]$ .
- Ein Set aus  $i=1\dots M$  Datenpunkten bezeichnen wir als *Datensatz*:  $\{\underline{x}^{(1)}, \underline{x}^{(2)}, \dots, \underline{x}^{(M)}\}$ .
- Wir können somit einen Datensatz in einer *Datenmatrix*  $\underline{X}$  darstellen:

$$\underline{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_N^{(1)} \\ x_1^{(2)} & x_2^{(2)} & & x_N^{(2)} \\ \vdots & & \ddots & \\ x_1^{(M)} & \dots & & x_N^{(M)} \end{bmatrix}$$

- Jeder Datenpunkt kann eine (oder mehrere) Zielgrößen  $y^{(i)}$  ( $\underline{y}^{(i)}$ ) besitzen.

### Grundlegende (statistische) Eigenschaften eines Datensatzes

- (Arithmetischer) Mittelwert  $\bar{x}_j$  der Eigenschaft mit Index  $j$ :  $\bar{x}_j = \frac{1}{M} \sum_{i=1}^M x_j^{(i)}$
- Median  $\tilde{x}_j$  der Eigenschaft mit Index  $j$ :
  - Sortiere die Datenpunkte nach Wert von  $x_j$
  - Geordneter Datensatz:  $\{\underline{x}^{(1)}, \underline{x}^{(2)}, \dots, \underline{x}^{(M)}\}$
  - Der Median ist dann:  $\tilde{x}_j = \begin{cases} x_j^m & \text{für ungerades } M = 2m+1 \\ \frac{1}{2}(x_j^m + x_j^{m+1}) & \text{für gerades } M = 2m \end{cases}$
  - Beispiele:
    - $\{1, 2, 3, 4, 5\} \rightarrow \tilde{x} = 3$
    - $\{1, 2, 3, 4, 5, 6\} \rightarrow \tilde{x} = 3.5$
    - $\{0.99, 1.01, 1.01, 1.02, 2\} \rightarrow \tilde{x} = 1.01, \bar{x} = 1.21$
- Stichproben-Varianz  $s_j^2$  der Eigenschaft mit Index  $j$ :  $s_j^2 = \frac{1}{M-1} \sum_{i=1}^M (x_j^{(i)} - \bar{x}_j)^2$
- Stichproben-Standardabweichung  $s_j$  der Eigenschaft mit Index  $j$ :  $s_j = \sqrt{s_j^2}$

## Maschinelles Lernen in der Verfahrenstechnik

- *Stichproben-Kovarianzmatrix*  $\underline{\underline{S}}$ :

$$\underline{\underline{S}} = \begin{pmatrix} s_1^2 & s_{12}^2 & \dots & s_{1N}^2 \\ s_{21}^2 & s_2^2 & & s_{2N}^2 \\ \vdots & & \ddots & \\ s_{N1}^2 & s_{N2}^2 & \dots & s_N^2 \end{pmatrix}$$

Hierbei ist  $s_j^2$  die Stichproben-Varianz von  $j$  (s.o.) und  $s_{jj'}^2$  die *Stichproben-Kovarianz*

von  $j$  und  $j'$ :  $s_{jj'}^2 = \frac{1}{M-1} \sum_{i=1}^M (x_j^{(i)} - \bar{x}_j)(x_{j'}^{(i)} - \bar{x}_{j'})$ .

Stichproben-Kovarianzmatrizen sind symmetrisch!

- *Stichproben-Korrelationsmatrix*  $\underline{\underline{P}}$ :

$$\underline{\underline{P}} = \begin{pmatrix} 1 & P_{12} & \dots & P_{1N} \\ P_{21} & 1 & & \vdots \\ \vdots & & \ddots & \\ P_{N1} & P_{N2} & \dots & 1 \end{pmatrix}$$

mit *Stichproben-Korrelationskoeffizient* zwischen  $j$  und  $j'$ :  $P_{jj'} = \frac{s_{jj'}^2}{\sqrt{s_j^2 s_{j'}^2}}$