# MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2017

Prof. E.G. Schukat-Talamazzini

Stand: 25. August 2017



# Teil II

# Datenaufbereitung



## Grundbegriffe des Data Mining

Datensätze mit expliziter oder impliziter Objektcharakterisierung

### Datensatz

**Menge** oder **Folge** von **Objekten** („*Instanzen*“) des Aufgabenbereichs $\boldsymbol{\Omega}$ mit ihren Eigenschaften und/oder Beziehungen

### Attribut

**Objekteigenschaft** $\hat{=}$ Element eines **Wertebereichs** $\mathcal{X}$ („*Skala*“)

### Beziehung

**Relation** $\mathcal{R} \subset \boldsymbol{\Omega} \times \boldsymbol{\Omega}$ zwischen Objekten oder ...
**Abstand/Ähnlichkeit** $d : \boldsymbol{\Omega} \times \boldsymbol{\Omega} \to \mathbb{R}$ zwischen Objekten

| | $o_1$ | $o_2$ | $o_3$ | $o_4$ |
|---|---|---|---|---|
| $o_1$ | | | $\infty$ | |
| $o_2$ | | | | $\infty$ |
| $o_3$ | $\infty$ | | | |
| $o_4$ | | $\infty$ | | |

| | $o_1$ | $o_2$ | $o_3$ | $o_4$ |
|---|---|---|---|---|
| $o_1$ | 0 | 3 | 8 | 15 |
| $o_2$ | 3 | 0 | 5 | 12 |
| $o_3$ | 8 | 5 | 0 | 7 |
| $o_4$ | 15 | 12 | 7 | 0 |

| | $\mathcal{X}_1$ | $\mathcal{X}_2$ | $\mathcal{X}_3$ | $\mathcal{X}_4$ |
|---|---|---|---|---|
| $o_1$ | + | *low* | 1.2 | +4 |
| $o_2$ | − | *hi* | 0.5 | −3 |
| $o_3$ | + | *hi* | 2.3 | −3 |
| $o_4$ | + | *med* | 2.1 | −7 |



## Datenmatrix $\hat{=}$ Objekte $\times$ Attribute

Objekteigenschaften erster Ordnung

### Definition

Sind $\mathcal{X}_1, \ldots, \mathcal{X}_N$ die Attribute eines Objekts, so heißen die Elemente

$$\boldsymbol{x} \;=\; (x_1, \ldots, x_N)^\top \;\in\; \mathcal{X}_1 \times \ldots \times \mathcal{X}_N =: \boldsymbol{\mathcal{X}}$$

**Datenvektoren** des Objekts.

Die Menge $\boldsymbol{\mathcal{X}}$ heißt **Wertebereich** des Objekts.

Eine (Multi-)Menge $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\} \subset \boldsymbol{\mathcal{X}}$ oder eine Folge $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) \in \boldsymbol{\mathcal{X}}^T$ bezeichnen wir als **Datenmatrix** oder ggf. als **Meßreihe**.

### Variable

{ Typ, Name, Wert }

### Datum

{ Attribut, Objekt, Eintrag }

### Schreibweise

Datenvektoren $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t, \ldots$
Meßwerte $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,N})^\top$

Reelle Datenmatrix $\begin{pmatrix} x_{1,1} & \cdots & x_{1,N} \\ \vdots & \ddots & \vdots \\ x_{T,1} & \cdots & x_{T,N} \end{pmatrix}$ $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)^\top \in \mathbb{R}^{T \times N}$

Werteskalen

Relationen und Distanzen

Skalenkonversion

Detektion von Ausreißern

Imputation von Fehlanzeigen

Zusammenfassung



# Attribute und ihre Skalentypen

Was bedeuten die Spalteneinträge einer Datenmatrix ?

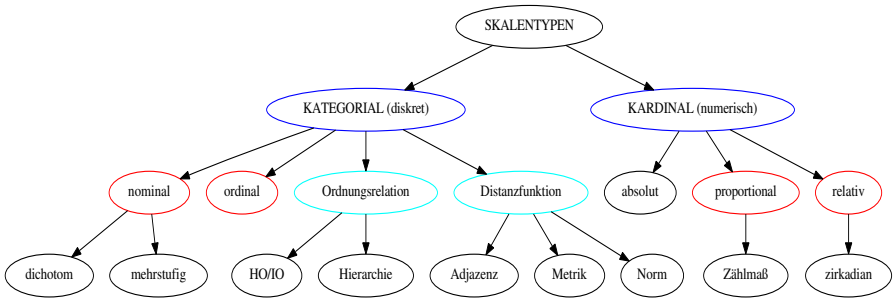
## Beispiel

| | vorbestraft | Partei | Abinote | Geburt | Spenden |
|---|---|---|---|---|---|
| Angela | F | CDU | gut | 1954 | $345 \cdot 10^3$ |
| Guido | F | FDP | ausreichend | 1961 | $137 \cdot 10^3$ |
| Roland | T | CDU | gut | 1958 | $3.6 \cdot 10^6$ |
| Gregor | F | PDS | sehr gut | 1948 | NA |
| Linus | F | Pirat | NA | 1969 | 0 |
| Bill | F | Rep | mangelhaft | 1955 | $-4.2 \cdot 10^9$ |
| Roman | T | ... | ... | 1933 | ... |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | $\{T, F\}$ | $\{\pi_1, \ldots, \pi_9\}$ | $\{\nu_1, \ldots, \nu_5\}$ | $\mathbb{Z} \subset \mathbb{R}$ | $\mathbb{R}$ |



# Skalentypen

Objektattribut $\hat{=}$ (Wertebereich, Operatorenmenge)




## Diskrete Skala

Endlicher Wertebereich

$\mathcal{X} = \{\xi_1, \xi_2, \ldots, \xi_K\}$

## Numerische Skala

Kontinuierlicher Wertebereich

$\mathcal{X} \subseteq \mathbb{R}$



# Typische Wertebereiche

## Nominalskala

- Dichotomien
  $\{0,1\}$, $\{T,F\}$, $\{+,-\}$, $\{m,f\}$ ...
- Zeichensätze $\{C, G, A, T\}$
- Farben („*red*“, green, blue)
- Gruppen & Prädikate

## Ordinalskala

- Notenskala
  „*sehr gut*“, „*gut*“, „*befriedigend*“, ...
- Unscharfe Prädikate
  „*kalt*“, „*kühl*“, „*lau*“, „*warm*“, „*heiß*“
- Eingefrorene Quantitäten
  „*2-türig*“, „*4-türig*“, „*5-türig*“

## Intervallskala (relativ)

- Temperaturen 20°C, 451°F
- Zeitangaben
  1066, 2001/09/11, 469 v.Chr., ...

## Verhältnisskala (absolut/proport.)

- absol. Temperatur 273°K
- Dauer 45 min, $13.7 \cdot 10^9$ Jahre
- Mengenangaben
  $C$= 2.98, 17 cm, 8 $\mu$g, ...

# Binäre Skalenoperationen

Vergleichsoperationen $\mathcal{X} \times \mathcal{X} \to \{T, F\}$ · Rechenoperationen $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$

## Nominalskala

$=$ Gleichheitstest

Alle Attributwerte $\xi_\ell$ sind gleichberechtigt.

## Ordinalskala

$\prec$ Vergleichbarkeit

Abschnittsbildung nach Totalordnung:
$\{\xi \mid \xi \preceq \xi_\ell\}$

## Intervallskala (relativ)

$-$ Differenzbildung

Unterschiede sind durch $x_1 - x_2$ quantifizierbar.

## Verhältnisskala (absolut/proport.)

$\div$ Quotientenbildung

Wohldefiniert:
„*Nullpunkt*“, „*doppelt*“, „*Drittel*“



# Durchschnittswerte

Wie berechnet man/frau einen für $(x_1, \ldots, x_T) \in \mathcal{X}^T$ „(proto)typischen“ Wert ?

## Nominalskala

**Modus** — der häufigste Wert:
$\mu^{\text{mod}} = \xi_{\ell^*}$ mit

$$\ell^* \;=\; \operatorname*{argmax}_{\ell} N_\ell$$

mit den absoluten Häufigkeiten
$N_\ell = \sum_{t=1}^{T} \delta_{x_t, \xi_\ell}$

## Ordinalskala

**Median** — der mittlere Wert:
$\mu^{\text{med}} = \xi_{\ell^*}$ mit

$$\sum_{k=1}^{\ell^*-1} N_k \;\le\; \frac{T}{2} \;\le\; \sum_{k=1}^{\ell^*} N_k$$

falls das Inventar $\mathcal{X}$ geordnet ist:
$\xi_1 < \xi_2 < \ldots < \xi_\ell < \xi_{\ell+1} < \ldots < \xi_L$

## Intervallskala (relativ)

**Arithmetisches Mittel**

$$\mu^{\text{mean}} \;=\; \frac{1}{T} \cdot \sum_{t=1}^{T} x_t \;=\; \frac{1}{T} \cdot \sum_{\ell=1}^{L} N_\ell \cdot \xi_\ell$$

## Verhältnisskala (absolut/proport.)

**Geometrisches Mittel**

$$\mu^{\text{geo}} \;=\; \sqrt[T]{\prod_{t=1}^{T} x_t} \;=\; \sqrt[T]{\prod_{\ell=1}^{L} \xi_\ell^{N_\ell}}$$



# Durchschnittswerte

Verallgemeinerung auf Metriken und normierte Vektorräume

## Beispiel

Für die Wertemenge $\{1, 1, 1, 2, 2, 5, 9\}$ gilt:
$\mu^{\text{mod}} = 1, \quad \mu^{\text{med}} = 2, \quad \mu^{\text{mean}} = 3, \quad \mu^{\text{geo}} = 2.0998$

## Definition

In einem metrischen Raum $(\mathcal{X}, d)$ heißt der Wert

$$\mu^{\text{zen}} \;=\; \operatorname*{argmin}_{z \in \mathcal{X}} \left( \sum_{t=1}^{T} d(z, x_t) \right)$$

das **Zentroid** der (Multi-)Menge $\{x_1, \ldots, x_T\}$.

## Lemma

*(1) Es ist $\mu^{mean}(\cdot)$ das Zentroid zur euklidischen Metrik $d(y, z) = (y - z)^2$.*
*(2) Es ist $\mu^{med}(\cdot)$ das Zentroid zur Betragsmetrik $d(y, z) = |y - z|$.*
*(3) Es ist $\mu^{mod}(\cdot)$ das Zentroid zur diskreten Metrik $d(y, z) = 1 - \delta_{y,z}$.*



# Durchschnittswerte

Verallgemeinerung von (endlichen) Wertemengen auf diskrete Verteilungen

## Definition

Es sei $\mathbb{X}$ eine diskrete Zufallsvariable über dem Wertebereich $\mathcal{X} \subset \mathbb{R}$ mit der Wahrscheinlichkeitsfunktion $\mathrm{P}(\cdot)$. Dann heißt

$$\mu(\mathbb{X}) \;=\; \mathcal{E}[\mathbb{X}] \;\stackrel{\text{def}}{=}\; \sum_{x \in \mathcal{X}} x \cdot \mathrm{P}(\mathbb{X} = x)$$

der **Erwartungswert** von $\mathbb{X}$, es heißt

$$\mu^{\text{med}}(\mathbb{X}) \;=\; \xi \quad \text{mit} \quad \sum_{x < \xi} \mathrm{P}(\mathbb{X} = x) \;\le\; \frac{1}{2} \;\le\; \sum_{x \le \xi} \mathrm{P}(\mathbb{X} = x)$$

der **Median** von $\mathbb{X}$, und es heißt

$$\mu^{\text{mod}}(\mathbb{X}) \;\stackrel{\text{def}}{=}\; \operatorname*{argmax}_{x \in \mathcal{X}} \mathrm{P}(\mathbb{X} = x)$$

der **Modus** von $\mathbb{X}$.

# Durchschnittswerte

Verallgemeinerung von (endlichen) Wertemengen auf stetige Verteilungen

## Definition

Für eine kontinuierliche Zufallsvariable über dem Wertebereich $\mathcal{X} = \mathbb{R}$ mit der Wahrscheinlichkeitsdichtefunktion $f_{\mathbb{X}}(\cdot)$ gilt entsprechend:

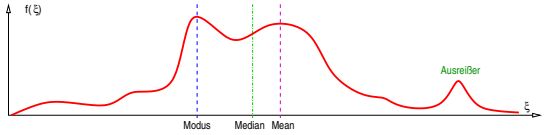
$$\mu(\mathbb{X}) \overset{\text{def}}{=} \mathcal{E}[\mathbb{X}] = \int_{\mathbb{R}} x \cdot f_{\mathbb{X}}(x)\, dx$$

$$\mu^{\text{med}}(\mathbb{X}) \overset{\text{def}}{=} \xi \quad \text{mit} \quad \int_{-\infty}^{\xi} f_{\mathbb{X}}(x)\, dx = \frac{1}{2}$$

$$\mu^{\text{mod}}(\mathbb{X}) \overset{\text{def}}{=} \operatorname*{argmax}_{x \in \mathbb{R}} f_{\mathbb{X}}(x)$$

Bemerkung

Die Mediandefinition erfordert eine stetige und streng monotone Wahrscheinlichkeitsverteilungsfunktion.






Werteskalen
Relationen und Distanzen
Skalenkonversion
Detektion von Ausreißern
Imputation von Fehlanzeigen
Zusammenfassung



# Relationen auf diskreten Attributen

Spezialfall: Objekte besitzen genau ein Attribut $\mathcal{X}$

## Adjazenz

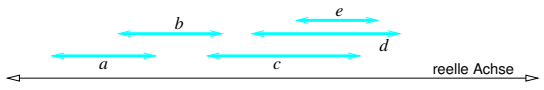
Die Matrix $\boldsymbol{A} \in \{0,1\}^{L \times L}$ repräsentiert eine **(Objekt)nachbarschaft**.

- räumliche Nähe, Verwandtschaft, Interaktion ...
- *„Elter-von"*, Einflußnahme, ...

## Präferenz

Die Relation $\mathcal{R} \subset \mathcal{X} \times \mathcal{X}$ repräsentiert eine (nicht notwendig totale) **Ordnung**.

- Halbordnung, Verband, Boolesche Algebra
- Turnier, (echte) Intervallordnung




Bemerkung

Zyklus: $c \preceq b \prec d \preceq c$

¬transitiv: $c \preceq b \preceq a$



# Abstände und Ähnlichkeiten

Diskrete metrische Attribute

## Definition

Eine Abstandsfunktion $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ heißt **Metrik** auf $\mathcal{X}$, wenn $d(\cdot,\cdot)$ für alle $x, y, z \in \mathcal{X}$ die drei Eigenschaften

1. $d(x,y) = 0 \Leftrightarrow x = y$ — Definitheit
2. $d(x,y) = d(y,x)$ — Symmetrie
3. $d(x,z) \leq d(x,y) + d(y,z)$ — Dreiecksungleichung

besitzt.

Bemerkungen

1. Jede Vektorraumnorm $\|\cdot\| : \mathcal{X} \to \mathbb{R}$ definiert eine Metrik $d(x,y) = \|x - y\|$.
2. Jedes innere Produkt $\langle\cdot,\cdot\rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ definiert eine VR-Norm $\|x\| = \sqrt{\langle x, x\rangle}$.
3. Distanzen transformieren in Ähnlichkeiten $s(x,y) = \exp(-{}^{d(x,y)}\!/_{2\sigma^2})$.
4. Ähnlichkeiten transformieren in Distanzen $d(x,y) = -2\sigma^2 \cdot \log(s(x,y))$.

# Spezialfall Zeichenketten

Attribute mit einem diskreten Wertebereich $\mathcal{X} \subset \mathcal{A}^\star$

## Elementare Operationen auf Zeichenketten

- Ersetzung eines Zeichens durch ein anderes — substitution
- Löschung eines Zeichens — deletion
- Einfügung eines Zeichens — insertion


## Definition

Ist $\mathcal{A}$ ein endliches Alphabet und sind $v$, $w$ zwei Zeichenfolgen aus $\mathcal{A}^\star$, so bezeichnet der **Levenshtein-Abstand** $d^{\text{lev}}(v, w)$ die minimale Anzahl von Elementaroperationen, mit denen $v$ in $w$ überführt werden kann.



# Spezialfall Zeichenketten

Zeichenkettenattribute sind metrisch und erlauben die Durchschnittbildung

## Lemma

*Der Levenshtein-Abstand $d^{lev} : \mathcal{A}^\star \times \mathcal{A}^\star \to \mathbb{R}$ über dem Alphabet $\mathcal{A}$ ist eine definite, symmetrische Distanzfunktion und erfüllt die Dreiecksungleichung — $(\mathcal{A}^\star, d^{lev})$ ist folglich ein **metrischer Raum**.*

## Definition

Sei $(\mathcal{X}, d)$ ein metrischer Raum und $(w_1, \ldots, w_T) \in \mathcal{X}^T$ eine Auswahl (Multimenge) von Elementen. Der Wert

$$\mu^{\text{mid}} \;=\; \underset{z \in \{w_1, \ldots, w_T\}}{\operatorname{argmin}} \left( \sum_{t=1}^{T} d(z, w_t) \right)$$

heißt das **Medoid** der Menge bezüglich der Metrik $d(\cdot, \cdot)$.

### Bemerkung

Das Medoid einer Wortmenge $w_1, \ldots, w_T$ mit maximaler Wortlänge $N_{\max}$ läßt sich mit Aufwand $\mathrm{O}(T^2 N_{\max}^2)$ berechnen.



Werteskalen

Relationen und Distanzen

Skalenkonversion

Detektion von Ausreißern

Imputation von Fehlanzeigen

Zusammenfassung



# Konversion der Attributskalen — wozu ?

## Datensatz mit Attributen unterschiedlichen Skalentyps

Traditionelle Modellierungsverfahren erfordern einheitliche Skalen:

- **Numerische Skalen** $\boldsymbol{\mathcal{X}} = \underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{N\text{-mal}} = \mathbb{R}^N$
  Multivariate Normalverteilung

$$f(\boldsymbol{x}) \;=\; \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{S}) \;=\; |2\pi \boldsymbol{S}|^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \cdot (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}$$

- **Diskrete Skalen** $\boldsymbol{\mathcal{X}} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N$ mit $\mathcal{X}_n = \{1, \ldots, \ell_n\}$
  $N$-dimensionale Wahrscheinlichkeitstabelle

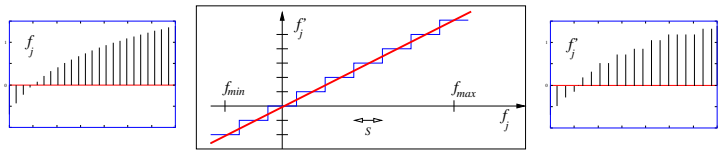
$$\mathrm{P}(\boldsymbol{x}) \;=\; p_{x_1, \ldots, x_N} \qquad \text{mit dem Tensor} \quad \boldsymbol{p} \in [0,1]^{\ell_1 \times \ell_2 \times \ldots \times \ell_N}$$

## Option auf robusteres Datenmodell

Sind die Attributwerte wirklich normalverteilt ?
Kann ich mir eine Tabelle mit $\prod_n \ell_n$ Einträgen leisten ?

# Diskretisierung numerischer Attribute (unüberwacht)

(kardinal ➪ ordinal)




## Äquidistante Intervalle

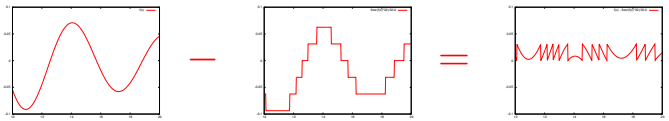
Mißachtet Datenverteilung $\rightsquigarrow$ unglm. Zellenbesetzung & Übersteuern

## Äquifrequente Intervalle

Histogrammegalisierung $\rightsquigarrow$ konstante Zellenbesetzung $T/L$

## Nichtlinearer Skalarquantisierer

Minimiert den Störabstand (SNR): *mittlerer quadratischer Quantisierungsfehler*



Faustregel
$L = \sqrt{T}$



# Nominalisierung ordinaler Attribute

(ordinal ➪ nominal)

## Problem

*Die Quantisierung numerischer Skalen liefert konstruktionsbedingt Werte einer **ordinalen** Skala.*

*Die immanente Reihenfolgeinformation wird aber von den einschlägigen Datenmodellen (W-Tabellen, lineare Modelle, Entscheidungsbäume) nicht genutzt.*

## Ordinale Entflechtung

Die ordinale Skala mit (sortiertem) Wertebereich $\mathcal{X} = \{\xi_1, \ldots, \xi_\ell\}$ wird auf einen Komplex **binärer** Attribute $\mathcal{X}_i = \{0,1\}$, $i = 1, \ldots, \ell-1$, abgebildet:

$$\phi(\xi_j) = (\underbrace{0,\ldots,0}_{(j-1)\text{-mal}}, \underbrace{1,\ldots,1}_{(\ell-j)\text{-mal}}) \in \{0,1\}^{\ell-1}$$

Für jede $\mathcal{X}$-Stufe $\xi - j$ gilt also:
$\phi_i(\xi_j) = 1 \Leftrightarrow i \geq j$.

Fall $\ell = 3$

| $\mathcal{X}$ | $\mathcal{X}_1$ | $\mathcal{X}_2$ |
|---|---|---|
| $\xi_1$ | 1 | 1 |
| $\xi_2$ | 0 | 1 |
| $\xi_3$ | 0 | 0 |

Fall $\ell = 5$

| $\mathcal{X}$ | $\mathcal{X}_1$ | $\mathcal{X}_2$ | $\mathcal{X}_3$ | $\mathcal{X}_4$ |
|---|---|---|---|---|
| $\xi_1$ | 1 | 1 | 1 | 1 |
| $\xi_2$ | 0 | 1 | 1 | 1 |
| $\xi_3$ | 0 | 0 | 1 | 1 |
| $\xi_4$ | 0 | 0 | 0 | 1 |
| $\xi_5$ | 0 | 0 | 0 | 0 |



# Kardinalisierung nominaler Attribute

(nominal ➪ numerisch)

## Problem

*Zahlreiche Methoden (k-nächste-Nachbarn, Bayesregel, Trennfunktionen) der Klassifikation und Vorhersage benötigen Objektabstände oder numerische, besser noch gaußverteilte Objektattribute.*

## Nominale Entflechtung

Die nominale Skala mit Wertebereich $\mathcal{X} = \{\xi_1, \ldots, \xi_\ell\}$ wird auf einen Komplex **reellwertiger** Attribute $\mathcal{X}_i = \{0,1\}$, $i = 1, \ldots, \ell$, abgebildet:

$$\phi(\xi_j) = (\underbrace{0,\ldots,0}_{(j-1)\text{-mal}}, 1, \underbrace{0,\ldots,1}_{(\ell-j)\text{-mal}}) \in \mathbb{R}^\ell$$

Fall $\ell = 5$

| $\mathcal{X}$ | $\mathbb{R}_1$ | $\mathbb{R}_2$ | $\mathbb{R}_3$ | $\mathbb{R}_4$ | $\mathbb{R}_5$ |
|---|---|---|---|---|---|
| $\xi_1$ | 1 | 0 | 0 | 0 | 0 |
| $\xi_2$ | 0 | 1 | 0 | 0 | 0 |
| $\xi_3$ | 0 | 0 | 1 | 0 | 0 |
| $\xi_4$ | 0 | 0 | 0 | 1 | 0 |
| $\xi_5$ | 0 | 0 | 0 | 0 | 1 |

Für diese Darstellung gilt die **Äquidistanzeigenschaft**

$$d(\phi(\xi_i), \phi(\xi_j)) = \|\phi(\xi_i) - \phi(\xi_j)\| = \begin{cases} 0 & \xi_i = \xi_j \\ \sqrt{2} & \xi_i \neq \xi_j \end{cases}$$



# Kontrastmatrizen

Auch im $\mathbb{R}^{\ell-1}$ ist genug Platz für $\xi_1, \ldots, \xi_\ell$

## Ursprung & Einheiten

Einer-gegen-alle: `treatment`

| | | | | |
|---|---|---|---|---|
| $\xi_1$ | 0 | 0 | 0 | 0 |
| $\xi_2$ | 1 | 0 | 0 | 0 |
| $\xi_3$ | 0 | 1 | 0 | 0 |
| $\xi_4$ | 0 | 0 | 1 | 0 |
| $\xi_5$ | 0 | 0 | 0 | 1 |

Distanzen 0, $\sqrt{2}$, aber auch 1

## Gestaffelt

Gegen-Anfangspartie: `helmert`

| | | | | |
|---|---|---|---|---|
| $\xi_1$ | −1 | −1 | −1 | −1 |
| $\xi_2$ | 1 | −1 | −1 | −1 |
| $\xi_3$ | 0 | 2 | −1 | −1 |
| $\xi_4$ | 0 | 0 | 3 | −1 |
| $\xi_5$ | 0 | 0 | 0 | 4 |

Distanzen 0 und viele andere ...

## Spaltenmittelwertfrei

Einer-gegen-alle: `sum`

| | | | | |
|---|---|---|---|---|
| $\xi_1$ | −1 | −1 | −1 | −1 |
| $\xi_2$ | 1 | 0 | 0 | 0 |
| $\xi_3$ | 0 | 1 | 0 | 0 |
| $\xi_4$ | 0 | 0 | 1 | 0 |
| $\xi_5$ | 0 | 0 | 0 | 1 |

Distanzen 0, $\sqrt{2}$, aber auch $\sqrt{\ell+2}$

## Äquidistant

Orthonormalpolynome: `poly`

| | | | | |
|---|---|---|---|---|
| $\xi_1$ | $p_1(r_1)$ | $p_1(r_2)$ | $p_1(r_3)$ | $p_1(r_4)$ |
| $\xi_2$ | $p_2(r_1)$ | $p_2(r_2)$ | $p_2(r_3)$ | $p_2(r_4)$ |
| $\xi_3$ | $p_3(r_1)$ | $p_3(r_2)$ | $p_3(r_3)$ | $p_3(r_4)$ |
| $\xi_4$ | $p_4(r_1)$ | $p_4(r_2)$ | $p_4(r_3)$ | $p_4(r_4)$ |
| $\xi_5$ | $p_5(r_1)$ | $p_5(r_2)$ | $p_5(r_3)$ | $p_5(r_4)$ |

$\|\boldsymbol{u} - \boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 - 2\langle \boldsymbol{u}, \boldsymbol{v}\rangle + \|\boldsymbol{v}\|^2 = 2$

# Redundante Kardinalisierung

Fehlererkennende und fehlerkorrigierende Codes

## Indexcodierung

Holzhammermethode: $\phi(\xi_j) = j$ $\qquad \phi : \{\xi_1, \ldots, \xi_\ell\} \to \mathbb{R}^1$

## Dualcodierung

(Hamming)abstände „fehleranfällig“ $\qquad \phi : \{\xi_1, \ldots, \xi_\ell\} \to \mathbb{R}^{\lceil \log_2 \ell \rceil}$

## Vollständige Korrekturcodes

Erkennt und kompensiert Fehler in einer Komponente (interessant ab $\ell = 4$)

$\phi : \{\xi_1, \ldots, \xi_\ell\} \to \mathbb{R}^L$, $L = 2^{\ell-1} - 1$

| | |
|---|---|
| $\xi_1$ | 1 1 1 1 1 1 1 |
| $\xi_2$ | 0 0 0 0 1 1 1 |
| $\xi_3$ | 0 0 1 1 0 0 1 |
| $\xi_4$ | 0 1 0 1 0 1 0 |

| | |
|---|---|
| $\xi_1$ | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 |
| $\xi_2$ | 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 |
| $\xi_3$ | 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 |
| $\xi_4$ | 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 |
| $\xi_5$ | 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 |

Beinhaltet alle $\{0,1\}^\ell$-Spalten außer Komplementen und den uninformativen Attributen $\mathbf{0}$, $\mathbf{1}$.



# Konversion von Distanzfunktionen

Nachbarschaft — Metrik — normierter Vektorraum

## Metrik ➪ symmetrische Nachbarschaft

Global operierende Schwellwertoperation ($0 < \delta_{\max} \in \mathbb{R}$)

$$\xi_i \propto \xi_j \quad \Leftrightarrow \quad d(\xi_i, \xi_j) \leq \delta_{\max}$$

## Metrik ➪ nichtsymmetrische Nachbarschaft

Lokale Umgebungsdefinition ($k$ nächste Nachbarn, $k \in \mathbb{N}$)

$$\xi_i \propto \xi_j \quad \Leftrightarrow \quad \xi_j \in \mathcal{U}_{\mathcal{X}}^{(k)}(\xi_i)$$

## Adjazenz ➪ Metrik

Geodätische Abstände (minimale Pfadlängen im Adjazenzgraphen)

## Metrik ➪ (euklidischer) Vektorraum

Nicht jede metrische Distanz $\boldsymbol{D} \in \mathbb{R}^{L \times L}$ ist im $\mathbb{R}^{L-1}$ repräsentierbar.



# Floyd-Warshall-Algorithmus

Schnelle Berechnung geodätischer Distanzen mittels dynamischer Programmierung

(Algorithmus)

1. INITIALISIERUNG

$$\text{Setze} \quad D_{ij} = \begin{cases} 0 & i = j \\ 1 & \xi_i,\ \xi_j \text{ adjazent} \\ \infty & \text{sonst} \end{cases}$$

2. REKURSION
Für alle $k, i, j \in \{1, \ldots, L\}$:

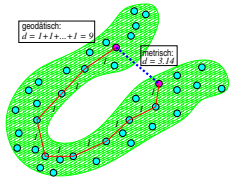
$$D_{ij} \leftarrow \min\{D_{ij}, D_{ik} + D_{kj}\}$$

3. TERMINIERUNG
Die Matrix $\boldsymbol{D}$ enthält alle minimalen Wegelängen zwischen Elementen $\xi_i$, $\xi_j$.

(Algorithmus)

## Wirkungsweise

Der FWA erzwingt in $O(L^3)$ Schritten die Gültigkeit der Dreiecksungleichung.




Bemerkung

Der Algorithmus ist auch anwendbar für gewichtete und nichtsymmetrische Adjazenzen.



# Kardinalisierung von Präferenzrelationen

Schwache Ordnungsrelation $(\mathcal{X}, \prec)$ ➪ ein, zwei, mehrere relative Attribute

## Intervallordnung

Repräsentation durch $\mathcal{X}_1 \times \mathcal{X}_2 = \mathbb{R}^2$ mit

$$a \prec b \quad \Leftrightarrow \quad a_2 < b_1$$

## Inklusionsfreie Intervallordnung

Repräsentation durch $\mathcal{X}_1 = \mathbb{R}^1$ mit $\delta \in \mathbb{R}_+$ und

$$a \prec b \quad \Leftrightarrow \quad a_1 + \delta < b_1$$

## Endliche Halbordnung

Repräsentation durch $\mathcal{X}_1 \times \ldots \times \mathcal{X}_L = \mathbb{R}^L$ mit

$$a \prec b \quad \Leftrightarrow \quad \forall \ell = 1, \ldots, L: \ a_\ell < b_\ell$$

# Standardisierung numerischer Skalen

Vereinheitlichung von Wertebereichen u/o Dynamikeigenschaften

## Min-Max-Normierung

$$f : \begin{cases} \mathbb{R} & \to & [0,1] \\ x & \mapsto & \dfrac{x - x_{\min}}{x_{\max} - x_{\min}} \end{cases}, \qquad f^{-1}(x) = (x_{\max} - x_{\min}) \cdot x + x_{\min}$$

## Statistische Normierung

$$f : \begin{cases} \mathbb{R} & \to & [\mu - C\sigma, \mu + C\sigma] \\ x & \mapsto & \dfrac{x - \mu}{\sigma} \end{cases}, \qquad f^{-1}(x) = \sigma \cdot x + \mu$$

## Reziproke Transformation

$$f : \begin{cases} \mathbb{R} \setminus \{0\} & \to & \mathbb{R} \setminus \{0\} \\ x & \mapsto & 1/x \end{cases}, \qquad f^{-1}(x) = 1/x$$



# Standardisierung numerischer Skalen

Vereinheitlichung von Wertebereichen u/o Dynamikeigenschaften

## Wurzel-Transformation

$$f : \begin{cases} (C, \infty) & \to & \mathbb{R}^+ \\ x & \mapsto & \sqrt[B]{x - C} \end{cases}, \qquad f^{-1}(x) = x^B + C$$

## Logarithmus-Transformation

$$f : \begin{cases} (C, \infty) & \to & \mathbb{R} \\ x & \mapsto & \log_B(x - C) \end{cases}, \qquad f^{-1}(x) = B^x + C$$

## Fisher-Transformation

$$f : \begin{cases} (-1, +1) & \to & \mathbb{R} \\ x & \mapsto & \dfrac{1}{2} \log_e \dfrac{1+x}{1-x} \end{cases}, \qquad f^{-1}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$



Werteskalen

Relationen und Distanzen

Skalenkonversion

Detektion von Ausreißern

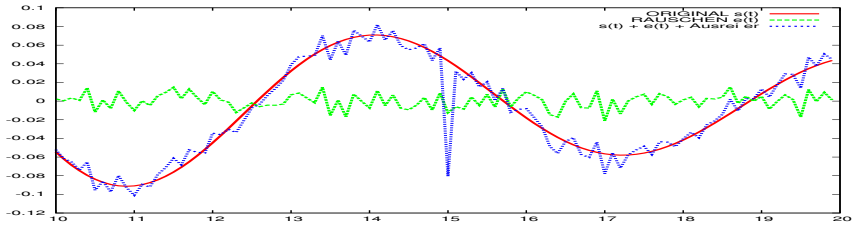
Imputation von Fehlanzeigen

Zusammenfassung



# Meßfehler & Erhebungsfehler

Die „Rohdaten“ sind oft fehlerbehaftet, verrauscht, verzerrt




## Zufällige Fehler

- Meßungenauigkeit
- Übertragungsstrecke
- Modell *additives Rauschen*: $y_n = x_n + e_n,\ e_n \sim \mathcal{N}(0, \sigma^2)$

- **Ausreißer**

## Systematische Fehler

- Kalibrierung
- Skalierung
- Trend, Drift, Saisoneffekt

- **Ausreißer**

# Ausreißerdetektion

Was ist ein Ausreißer und wie erkenne ich ihn ?

## Vertikale Detektion

Ein Wert $x_{tj}$ fällt aus dem Rahmen seines **Attributs** $\mathcal{X}_j$.

Kategoriale Attribute bieten *keine Handhabe* ! $\mathcal{X}_j = \{m, f\}$

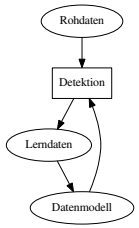
## Horizontale Detektion

Ein Wert $x_{tj}$ fällt aus dem Rahmen seines **Objekts** $\boldsymbol{o}_t$.

Werden Objekte durch Ausreißer erst *interessant* ? $\boldsymbol{o}_t = (\text{„kath.“}, \text{„verh.“})$

## Teufelskreis

vertikale **Detektion** ⇦ **Attributmodell**
horizontale **Detektion** ⇦ **Objektmodell**






# Hypothesentests für Ausreißer

Ausreißer ≙ extrem unwahrscheinliche Attributwerte

## Satz (Tschebyscheff)

*Ist $\mathbb{X}$ eine kontinuierliche Zufallsvariable mit dem Erwartungswert $\mu$ und der Varianz $\sigma^2$, so gilt für jede Konstante $C > 0$ die Ungleichung:*

$$\mathrm{P}(|\frac{\mathbb{X} - \mu}{\sigma}| \geq C) \;\leq\; \frac{1}{C^2}$$

## Beispiel

Zweiseitige Streuungswahrscheinlichkeiten m/o NV-Annahme:

| | $\sigma$ | $2\sigma$ | $3\sigma$ | $4\sigma$ | $5\sigma$ |
|---|---|---|---|---|---|
| Tschebyscheff | $\leq 1$ | $\leq 0.25$ | $\leq 0.11$ | $\leq 0.063$ | $\leq 0.040$ |
| $\mathcal{N}(\mu, \sigma)$ | $= 0.323$ | $= 0.065$ | $= 0.003$ | $= 0.001$ | $\leq 0.0001$ |

⇨ ein „zahnloser“ Test ohne Kenntnis der Dichtefunktion !



# Hypothesentests für Ausreißer

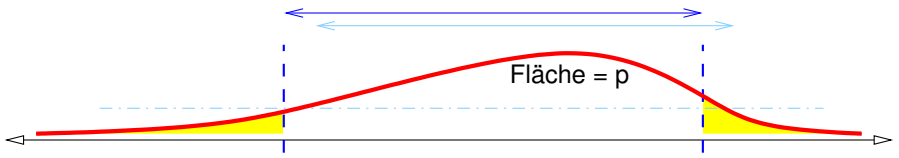
... bei bekannter unimodaler Verteilungsdichtefunktion

## Definition (Quantilmethode)

Ein Wert $x_q \in \mathbb{R}$ heißt $q$-**Quantil** der Dichtefunktion $f_{\mathbb{X}}(\cdot)$ genau dann, wenn gilt:

$$F_{\mathbb{X}}(x_q) \;=\; \mathrm{P}(\mathbb{X} \leq x_q) \;=\; q$$

Ein Wert $x \in \mathbb{R}$ heißt **Ausreißer** der Verteilung zum **Niveau** $p \in [0, 1]$, wenn er außerhalb des Akzeptanzintervalls $[x_{1/2-p/2}, x_{1/2+p/2}]$ liegt.




Bemerkungen

1. Für **symmetrische** Dichtefunktionen gilt für jedes $q \in [0, 1]$ die Identität $f_{\mathbb{X}}(x_q) = f_{\mathbb{X}}(x_{1-q})$. $[\mu - C\sigma, \mu + C\sigma]$
2. Für **multimodale** Dichtefunktionen ergibt das definierte Akzeptanzintervall keinen Sinn.



# Hypothesentests für Ausreißer

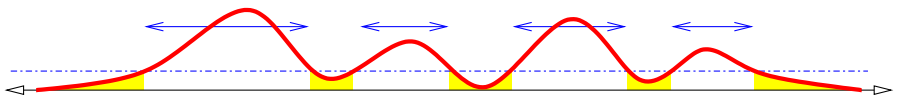
... bei bekannter multimodaler Verteilungsdichtefunktion

## Definition (Bayesträgermethode)

Die Wertemenge $\mathcal{B}_c = \{x \mid f_{\mathbb{X}}(x) \geq c\}$ heißt **Bayesträger** der Verteilung $f_{\mathbb{X}}(\cdot)$ zum **Niveau** $p \in [0, 1]$, wenn gilt:

$$\int_{\mathcal{B}_c} f_{\mathbb{X}}(\xi)\, d\xi \;=\; p$$

Jeder Wert $x \in \mathbb{R}$ mit $f_{\mathbb{X}}(x) < c$ heißt **Ausreißer** der Verteilung zum **Niveau** $p$.



Bemerkungen

1. Für **symmetrisch-unimodale** Dichtefunktionen stimmen Bayesträger und Akzeptanzintervall überein.
2. Nicht verwechseln mit **Bayesintervall**, dem kürzesten Intervall mit Fläche $p$.

# Faustregeln zur Ausreißerdetektion

Treffer als Fehlanzeige (NA=„*not available*“) markieren

## Unimodal

- **Normalverteilung**

$$|x - \mu| \; > \; C \cdot \sigma$$

- **Gleichverteilung**

$$|x - \mu| \; > \; p\text{-Niveau}$$

- **Empirischer Trimm**

$$x \; \notin \; [x_{1/2 - p/2}, x_{1/2 + p/2}]$$

## Multimodal

- **Tschebyscheff**

$$\frac{|x - \mu|}{\sigma} \; > \; \sqrt{\frac{1}{1 - p}}$$

- **Gauß-Mischung**

$$(\forall \ell) \; |x - \mu_\ell| > C \cdot \sigma_\ell$$

- **Lonesome Cowperson**

$$|x - k\text{-NN}(x)| \; > \; d_{\max}$$



# Teufelskreis Parameterschätzung

Ausreißer verändern die genutzten Verteilungsparameter

## Modellrechnung für die $C\sigma$-Regel

- **Datensatz**
  Eine $\mathcal{N}(\mu, \sigma)$-verteilte Probe der Größe $T$
  zuzüglich $M^+$ Ausreißer der Gestalt $a^+ = \mu + c\sigma$
  zuzüglich $M^-$ Ausreißer der Gestalt $a^- = \mu - c\sigma$
  $(T' = T + M^+ + M^-)$
- **Geschätzter Erwartungswert** $\hat{\mu} \; = \; \mu + \frac{M^+ - M^-}{T'} \cdot c\sigma$
  (Im Fall $M^+ = M^-$ gilt einfach $\hat{\mu} = \mu$.)
- **Geschätzte Varianz** $\hat{\sigma} \; = \; \sigma\sqrt{1 + c^2 r - r}$
  Gilt mit $M := {}^{M^+}\!/_2 = {}^{M^-}\!/_2$ wegen
  $\frac{1}{T'} \sum_{x \in \omega'} x^2 \; = \; \mu^2 + \sigma^2 + \frac{M}{T+M}(c^2 - 1)\sigma^2$
  und der Abkürzung $r := {}^{M}\!/_{(T+M)}$.

Für eine nicht verschwindende Anzahl ($r \gg 0$) von markanten Ausreißern ($c \gg 1$) dominiert $c^2 r$ den Wurzelausdruck und die $C\sigma$-Regel ist wegen $\hat{\sigma} \propto c$ entschärft!



Werteskalen

Relationen und Distanzen

Skalenkonversion

Detektion von Ausreißern

Imputation von Fehlanzeigen

Zusammenfassung



# Fehlanzeigen (a.k.a. „*not available*“)

Nicht zugängliche Attributwerte in der Datenmatrix

## Fehlanzeige als Unfall

Sensorkomponente hat versagt
Erhebungsprotokoll unvollständig
Markierte Ausreißer

## Fehlanzeige als Regelfall

Verzicht aus Kostengründen
Nichthomogenes Warehousing
Dünnbesetzung anwendungsbedingt
z.B. Bewertungssysteme für Musik, Bücher, Restaurants, Webseiten, Bordellbetriebe, ...

## Fehlanzeigenbehandlung

- **Objekt löschen**
  Können wir uns das leisten?
- **Eintrag markieren**
  und auf spezielle Weise weiterverarbeiten.
- **Imputieren**
  Leerstelle mit geeignetem Wert auffüllen.

# Imputationstechniken

Welcherart Zusatzinformation wird zur Wertergänzung genutzt ?

| $\mathbb{M}_1$ | $\mathbb{M}_2$ | ... | $\mathbb{M}_n$ | ... | $\mathbb{M}_N$ |
|---|---|---|---|---|---|
| $\mathbb{X}_1$ | $\mathbb{X}_2$ | ... | $\mathbb{X}_n$ | ... | $\mathbb{X}_N$ |

*„Missing (Completely) At Random“*

## Kontextfrei (MCAR)

Attributstatistik (& Ausreißer)

- Ersetzen durch Datenmittel $\hat{\mu}$
- durch $x_{\min}$ bzw. $x_{\max}$
- durch nächsten Nachbarn

$$x_n^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{\xi \in \bar{\omega}} d(x_n, \xi)$$

## Interpolation (MAR)

Zeile/Spalte $\hat{=}$ Meßreihe

- Linearer Ausgleich, z.B.

$$x_n^* \stackrel{\text{def}}{=} (x_{n-1} + x_{n+1}) \,/\, 2$$

- Polynome, Splines (nl.)
- Glättungsfilter

## Regression (MAR)

Probabilistisches Datenmodell

$$x_n^* = \mathcal{E}[\mathbb{X}_n \mid \ldots, x_{n-1}, x_{n+1}, \ldots]$$

## Matrixapproximation (MAR)

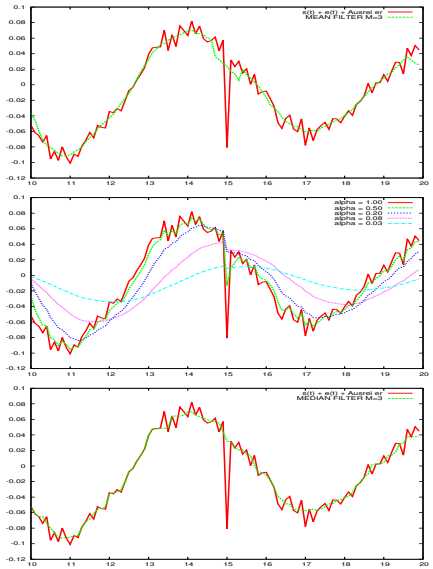
Lückenhafte (num.) Datenmatrix

$$\boldsymbol{X} \stackrel{\neg\text{NA}}{\approx} \boldsymbol{V}^\top \boldsymbol{D}\boldsymbol{U}$$



# Glättungsfilter für Meßreihenfehler

Imputation $\hat{=}$ kontextfrei Ersetzen + Filtern



## Gleitender Mittelwert

der Ordnung $q = 2p + 1$, $p \in \mathbb{N}$:

$$\hat{x}_n = \frac{1}{q} \cdot \sum_{\ell=n-p}^{n+p} x_\ell$$

⇩ Ausreißer, ⇧ Phasentreue

## Exponentialfilter

mit Abklingparameter $\alpha \in [0, 1]$:

$$\hat{x}_n = \hat{x}_{n-1} + \alpha \cdot (x_n - \hat{x}_{n-1})$$

⇔ Ausreißer,
⇩ Phasentreue/Nivellierung

## Gleitendes Medianfilter

der Ordnung $q = 2p + 1$, $p \in \mathbb{N}$:

$$\hat{x}_n = \mu^{\text{med}}(x_{n-p}, \ldots, x_n, \ldots, x_{n+p})$$



# Regression für nominale Datensätze

Beispielszenarium: drei Attribute $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ mit 2, 3 bzw. 4 Wertestufen

⟨Algorithmus⟩

1. ABSOLUTE HÄUFIGKEITEN
   Erstelle Tabelle $\boldsymbol{f} \in \mathbb{N}^{2 \cdot 3 \cdot 4}$ mit den 24 Auftretenszahlen $f_{ijk}$ der Ereignisse $(x_1, x_2, x_3) = (\xi_i, \eta_j, \zeta_k)$.
2. EREIGNISWAHRSCHEINLICHKEITEN
   Erstelle Tabelle der 24 ML-Schätzwerte $\hat{p}_{ijk} = f_{ijk}/T$.
3. BEDINGTE ATTRIBUTWAHRSCHEINLICHKEITEN

$$q_{i|jk}^{(1|23)} = \frac{\hat{p}_{ijk}}{\sum_\ell p_{\ell jk}}, \quad q_{j|ik}^{(2|13)} = \frac{\hat{p}_{ijk}}{\sum_\ell p_{i\ell k}}, \quad q_{k|ij}^{(3|12)} = \frac{\hat{p}_{ijk}}{\sum_\ell p_{ij\ell}}$$

4. IMPUTATION DES BEDINGTEN MODUS

$$\mu_{jk}^{(23)} = \operatorname*{argmax}_i q_{i|jk}^{(1|23)}, \quad \mu_{ik}^{(13)} = \operatorname*{argmax}_j q_{j|ik}^{(2|13)}, \quad \mu_{ij}^{(12)} = \operatorname*{argmax}_k q_{k|ij}^{(3|12)}$$

⟨zumhoglA⟩



Werteskalen
Relationen und Distanzen
Skalenkonversion
Detektion von Ausreißern
Imputation von Fehlanzeigen
Zusammenfassung

# Zusammenfassung (2)

1. Ein **Datensatz** besteht aus **Objekten**, die explizit durch eine Reihe von **Attributwerten** oder implizit durch Beziehungen wie **Abstand**, **Adjazenz** oder **Präferenz** charakterisiert sind.
2. Attribute besitzen eine **diskrete Skala** (**nominal** oder **ordinal**) oder eine **numerische Skala** (**relativ** oder **proportional**).
3. Die Skalen unterscheiden sich hinsichtlich ihres **Wertebereichs**, ihrer **Verknüpfungsoperationen** und ihrer **Durchschnittswertbildung**.
4. Auf **Zeichenketten** ist mit dem Levenshteinabstand eine **Metrik** und mit dem **Medoid** ein Durchschnitt definiert.
5. Skalen lassen sich nötigenfalls mittels **Quantisierung** (numerisch$\rightsquigarrow$ordinal), **Entflechtung** (ordinal$\rightsquigarrow$nominal) bzw. **Kontrastmatrizen** (nominal$\rightsquigarrow$numerisch) konvertieren.
6. Aus **Adjazenzen** leiten sich **geodätische Distanzen** her, aus **Präferenzen** ein oder zwei **Ordinalskalen**.
7. **Ausreißer** werden durch einen der attributbezogenen **Hypothesentests** detektiert.
8. Als **Ersatzwerte** für Ausreißer und andere **Fehlanzeigen** dienen Mittel- und Extremwerte; wenn möglich, imputieren wir durch **Interpolation** oder **Regression**.