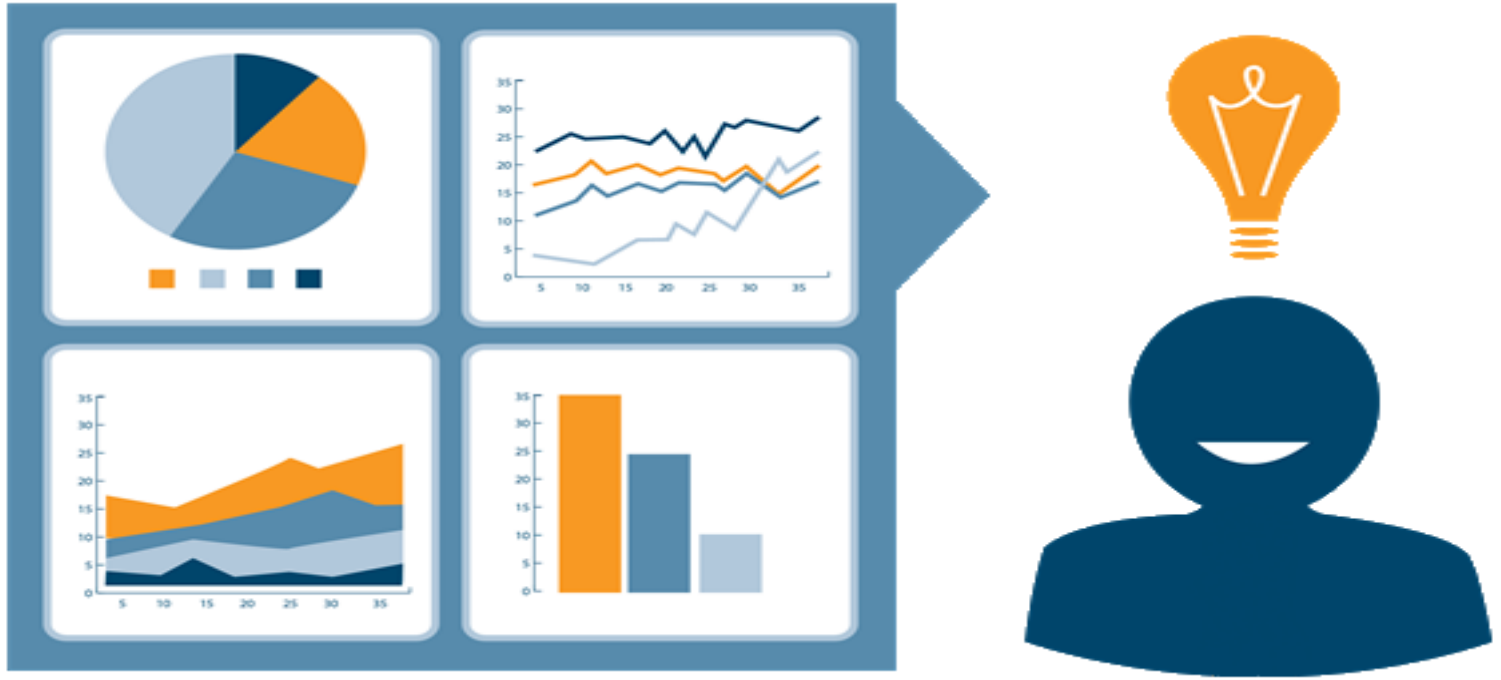
Introduction to Data Science
Prof. Dr. Bernhard Drabant
Duale Hochschule Baden-Württemberg Mannheim
Fakultät Wirtschaft

Methoden und Verfahren

Konzepte, Algorithmen



Methoden und Verfahren – Konzepte und Algorithmen

Ziel dieses Kapitels

- Details einiger ausgewählter Methoden und Verfahren der Wissensgewinnung
  - Ideen
  - Konzepte
  - Algorithmen

Exkurs: Abbildungen mehrerer Variablen

Definition 1 (Abbildung).

X und Y seien zwei Mengen. Unter einer Abbildung (oder Funktion) f von X nach Y versteht man eine Zuordnungsvorschrift, die jedem Element $x \in X$ in eindeutiger Weise genau ein Element $y \in Y$ zuordnet. Das Element y wird mit $f(x)$ bezeichnet: $y = f(x)$. Das Element $f(x)$ heißt Bild von x unter der Abbildung f.

X heißt Definitionsmenge, Y heißt Werte- oder Zielmenge.

Notationen:

$$f : \begin{cases} X \to Y \\ x \mapsto f(x) \end{cases}$$

oder in Kurzform $f : X \to Y$ oder $X \xrightarrow{f} Y$.

Exkurs: Abbildungen mehrerer Variablen

Bemerkung:

In dem Ausdruck $f(x)$ heißt x das Argument der Abbildung f. Um zu betonen, dass in $f(x)$ das Argument x beliebig in M gewählt werden darf, nennt man x Variable von f.

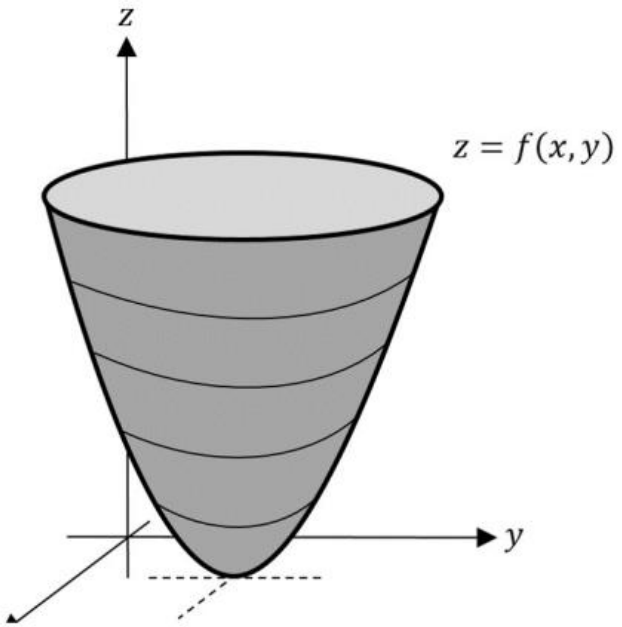
Exkurs: Abbildungen mehrerer Variablen

Beispiel

Sei die Abbildung $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ definiert durch

$$f(x, y) = (x - 1)^2 + (y - 2)^2$$

Diese Funktion lässt sich graphisch wie folgt darstellen:



Exkurs: Abbildungen mehrerer Variablen

Rechnen in $\mathbb{R}^n$

1. $\mathbb{R}^n$ ist ein Vektorraum. Die Elemente $\mathbf{x}$ in $\mathbb{R}^n$ sind Vektoren $\mathbf{x} = (x_1, \ldots, x_n)$ mit n Komponenten $x_j \in \mathbb{R}$, $j \in \{1, \ldots, n\}$.
2. In $\mathbb{R}^n$ können Vektoren $\mathbf{x}$ und $\mathbf{y}$ addiert oder mit einem Skalar $\lambda \in \mathbb{R}$ multipliziert werden:

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_n + y_n)$$
$$\lambda \cdot \mathbf{x} = (\lambda \cdot x_1, \ldots, \lambda \cdot x_n)$$

3. Jeder Vektor $\mathbf{x} \in \mathbb{R}^n$ hat eine euklidische Länge $\|\mathbf{x}\| = \sqrt{x_1^2 + \ldots + x_n^2}$
4. Zwei Elemente $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ haben den Abstand $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$

Exkurs: Abbildungen mehrerer Variablen

Rechnen in $\mathbb{R}^n$

Bemerkung

- $\mathbb{R}^n = \underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{n \text{ fach}}$ ist das n-fache cartesische Produkt von $\mathbb{R}$.
- Sei $f : \mathbb{R}^n \to \mathbb{R}^m$ eine Abbildung. Wie für Abbildungen in $\mathbb{R}$ können somit Stetigkeit und Differenzierbarkeit der Funktion f definiert werden.

Exkurs: Abbildungen mehrerer Variablen

Ableitung

Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine differenzierbare Abbildung. Deren Ableitung $\frac{\partial f}{\partial \mathbf{x}}$ ist ein n-dimensionaler Vektor, der komponentenweise definiert ist:

$$(\frac{\partial f}{\partial \mathbf{x}})_i = \frac{\partial f}{\partial x_i}$$

wobei $\frac{\partial f}{\partial x_i}$ die gewöhnliche Ableitung nach der Variablen x_i ist und die anderen Variablen wie Konstanten behandelt werden.

Exkurs: Abbildungen mehrerer Variablen

Beispiel

Sei die Abbildung $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ gegeben durch

$$f(x, y) = (x - 1)^2 + (y - 2)^2$$

Diese Abbildung ist differenzierbar, und ihre Ableitung ist gegeben durch:

$$(\frac{\partial f}{\partial \mathbf{x}}) = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$$

mit

$$\frac{\partial f}{\partial x} = 2\,(x - 1)$$

$$\frac{\partial f}{\partial y} = 2\,(y - 2)$$

Exkurs: Abbildungen mehrerer Variablen

Bemerkung:

1. Statt $\frac{\partial f}{\partial \mathbf{x}}$ wird gelegentlich die Notation $\partial_{\mathbf{x}} f$ oder ∇f (sprich: "nabla f") verwendet.
2. Statt $\frac{\partial f}{\partial x_i}$ wird gelegentlich die Notation $\partial_{x_i} f$ oder $\partial_i f$ verwendet.

Exkurs: Abbildungen mehrerer Variablen

Lokale Extrema von Abbildungen mehrerer Variablen

Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine differenzierbare Abbildung. Eine notwendige Bedingung, dass f an der Stelle $\mathbf{x}_{(0)}$ einen Extremwert besitzt, ist gegeben durch

$$\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}_{(0)}) = \mathbf{0}$$

Der Punkt $\mathbf{x}_{(0)}$ heißt stationärer Punkt der Abbildung f.

Exkurs: Abbildungen mehrerer Variablen

Lokale Extrema von Abbildungen mehrerer Variablen

Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine (zweimal) differenzierbare Abbildung. Die Matrix

$$H_f(\mathbf{x}) = \left(\partial_i \partial_j f\right)_{i,j}$$

heißt die Hesse-Matrix der Abbildung f.

Definition 2.

Der k. (führende) Hauptminor der Matrix $H_f(\mathbf{x})$ ist die Determinante der linken oberen $(k \times k)$-Untermatrix von $H_f(\mathbf{x})$. Bezeichnung: $M_k(f)$.

Exkurs: Abbildungen mehrerer Variablen

Lokale Extrema von Abbildungen mehrerer Variablen

Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine (zweimal) differenzierbare Abbildung und $H_f(\mathbf{x})$ die Hesse-Matrix von f.
Sei außerdem $\mathbf{x}_{(0)}$ ein stationärer Punkt der Abbildung f.

Dann gilt eines der folgenden ausschließlichen Kriterien:

(1) Alle Hauptminoren $M_k(f) > 0 \Rightarrow \mathbf{x}_{(0)}$ ist ein lokales Minimum von f.

(2) Für alle Hauptminoren $M_k(f)$ gilt $(-1)^k M_k(f) > 0 \Rightarrow \mathbf{x}_{(0)}$ ist ein lokales Maximum von f.

(3) $\det(H_f)(\mathbf{x}_{(0)}) \neq 0$ aber weder (1) noch (2) sind erfüllt $\Rightarrow \mathbf{x}_{(0)}$ ist ein Sattelpunkt von f.

(4) Andernfalls ist keine Aussage möglich.

Exkurs: Abbildungen mehrerer Variablen

Übung

Sei die Abbildung $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ gegeben durch

$$f(x, y) = (x - 1)^2 + (y - 2)^2$$

Bestimmen Sie die lokale Extrema der Abbildung f.

Beachten Sie: $\mathbf{x}^2 = x^2 + y^2$

Exkurs: Abbildungen mehrerer Variablen

Übung

Sei die Abbildung $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ gegeben durch

$$f(x, y) = \mathbf{x}^2 \cdot (\mathbf{x}^2 - 1) + \frac{1}{2}$$

Diskutieren Sie die Abbildung f und bestimmen Sie deren lokale Extrema.

Beachten Sie: $\mathbf{x}^2 = x^2 + y^2$

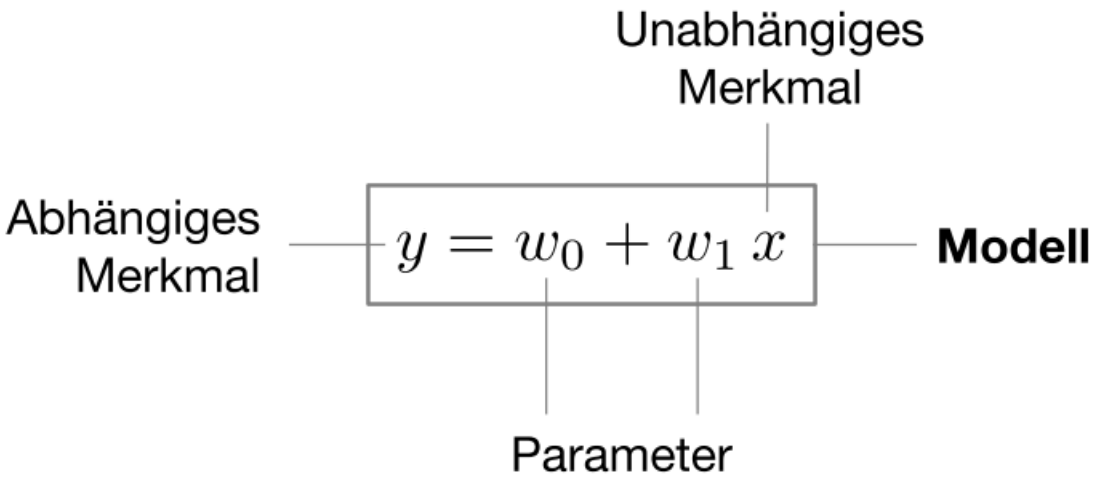
Methode: Lineare Regression

- Einfache lineare Regression betrachtet **Datenpunkte**

$$(x_1, y_1), \ldots, (x_n, y_n)$$

und nimmt an, dass das metrische Merkmal y **linear** vom metrischen Merkmal x **abhängt**

- Das **angenommene Modell** hat somit die Form



Lineare Regression

Beispiel: Fahrzeuge mit x = Leistung, y = (inverser) Verbrauch

- Annahme
  - Inverser Verbrauch (= km pro Liter) y ist linear abhängig von der Leistung x eines Fahrzeugs

- Vorgehen:
  1. Annotierte Trainingsdaten $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ mit n Datensätzen bereitstellen
     - x_j : Leistung des Fahrzeugs j
     - y_j : inverser Verbrauch des Fahrzeugs j
  2. Teste, ob in den Trainingsdaten eine lineare Beziehung angenähert vorliegt
  3. Bestimme das optimale lineare Modell – genauer die Parameter $\mathbf{w_0}$ und $\mathbf{w_1}$:

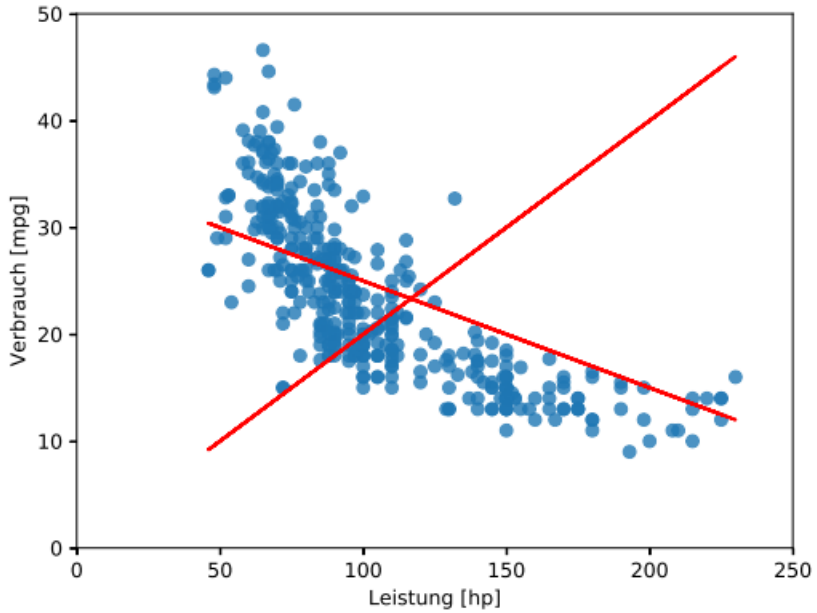
$$\mathbf{y = w_0 + w_1 x}$$

Lineare Regression – (1) Trainingsdaten beschaffen

	mpg	cylinders	cubicinches	hp	weightlbs	time-to-60	year	brand
0	14.0	8	350	165	4209	12	1972	US
1	31.9	4	89	71	1925	14	1980	Europe
2	17.0	8	302	140	3449	11	1971	US
3	15.0	8	400	150	3761	10	1971	US
4	30.5	4	98	63	2051	17	1978	US
5	23.0	8	350	125	3900	17	1980	US
6	13.0	8	351	158	4363	13	1974	US
7	14.0	8	440	215	4312	9	1971	US
8	25.4	5	183	77	3530	20	1980	Europe
9	37.7	4	89	62	2050	17	1982	Japan
10	34.0	4	108	70	2245	17	1983	Japan
11	34.3	4	97	78	2188	16	1981	Europe
...	...	...	...	...	...	...	...	...
250	32.1	4	98	70	2120	16	1981	US
251	24.0	4	121	110	2660	14	1974	Europe
252	36.4	5	121	67	2950	20	1981	Europe
253	13.0	8	350	145	3988	13	1974	US
254	23.5	6	173	110	2725	13	1982	US
255	24.0	4	113	95	2372	15	1971	Japan
256	17.0	8	305	130	3840	15	1980	US
257	36.1	4	91	60	1800	16	1979	Japan
258	22.0	6	232	112	2835	15	1983	US
259	18.0	6	232	100	3288	16	1972	US
260	22.0	6	250	105	3353	15	1977	US

Lineare Regression – (2) Linearitätstest

- **Verschiedene Werte** der Parameter w_0 und w_1 entsprechen **verschiedenen Geraden**



$w_0 = 0$ $w_1 = 0.2$

$w_0 = 35$ $w_1 = -0.1$

Teste, ob lineare Beziehung angenähert vorliegt

- Wir benötigen ein **Gütekriterium**, um zu bestimmen, **welche Gerade die beste** ist

Lineare Regression – (2) Linearitätstest

Für den Linearitätstest benötige wir folgende Größen (Mittelwert, Varianz, Standardabweichung):

- Wir definieren den **Mittelwert** unserer **Merkmale** als

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Die **Varianz** unserer Merkmale ist definiert als

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Die Werte σ_x und σ_y heißen **Standardabweichung** der Merkmale x und y

Lineare Regression – (2) Linearitätstest

- **Kovarianz** $\mathrm{cov}_{x,y}$ misst inwiefern die beiden Merkmale x und y **zusammenhängen**, d.h. sich in die gleiche Richtung bzw. entgegengesetzte Richtungen ändern

$$\mathrm{cov}_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Große Kovarianz deutet auf einen Zusammenhang hin
  - ein **positiver Wert** zeigt an, dass sich die beiden Merkmale in die **gleiche Richtung** ändern
  - ein **negativer Wert** zeigt an, dass sich die beiden Merkmale in **entgegengesetzte Richtungen** ändern

Lineare Regression – (2) Linearitätstest

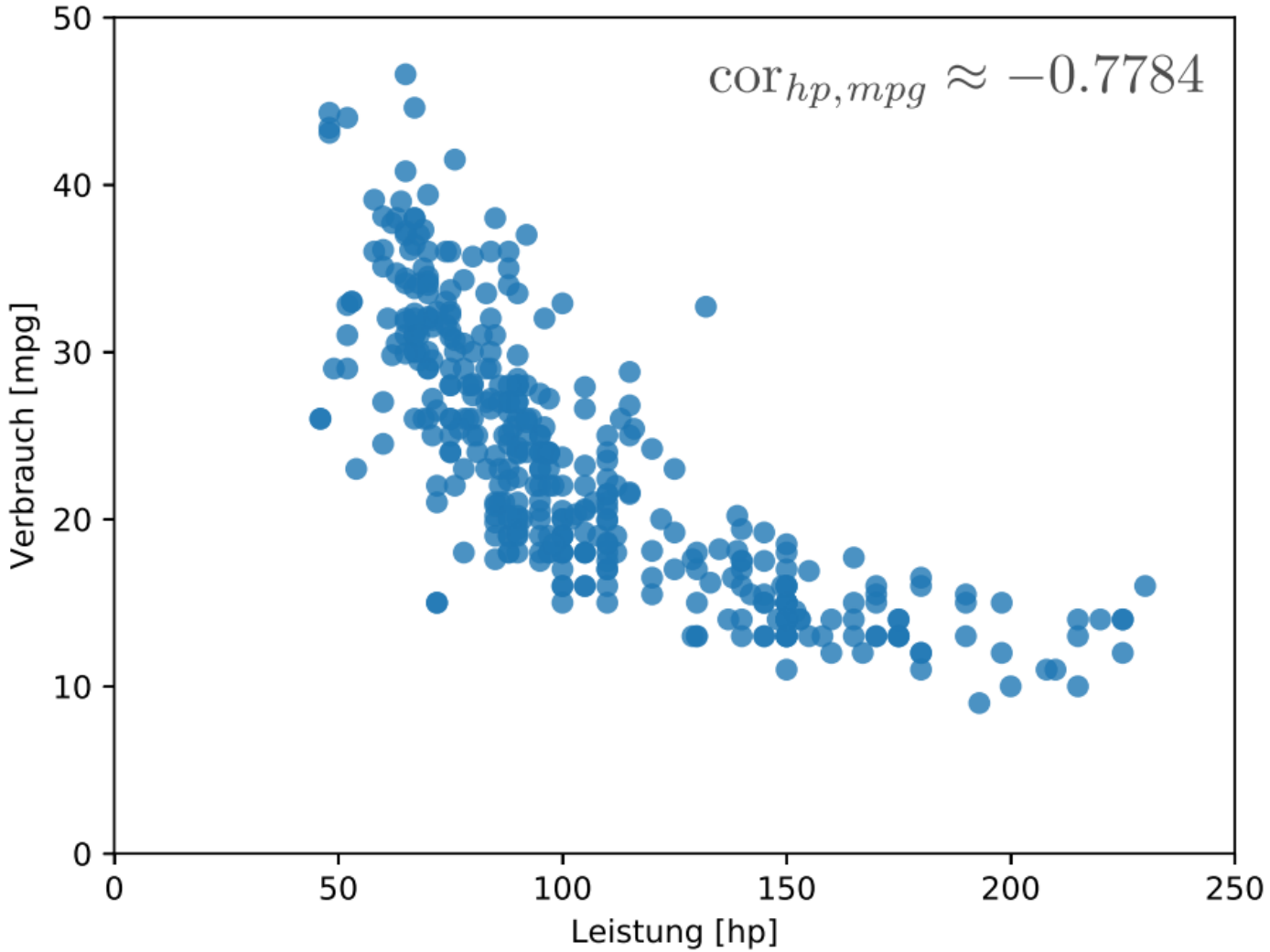
- Pearsons **Korrelationskoeffizient** misst inwiefern ein linearer Zusammenhang zwischen zwei Merkmalen x und y besteht

$$\mathrm{cor}_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\mathrm{cov}_{x,y}}{\sigma_x \sigma_y}$$

- Pearsons Korrelationskoeffizient nimmt Werte in [-1,+1] an
  - Wert -1 zeigt **negative lineare Korrelation** an
  - Wert 0 zeigt **keine lineare Korrelation** an
  - Wert 1 zeigt **positive lineare Korrelation** an

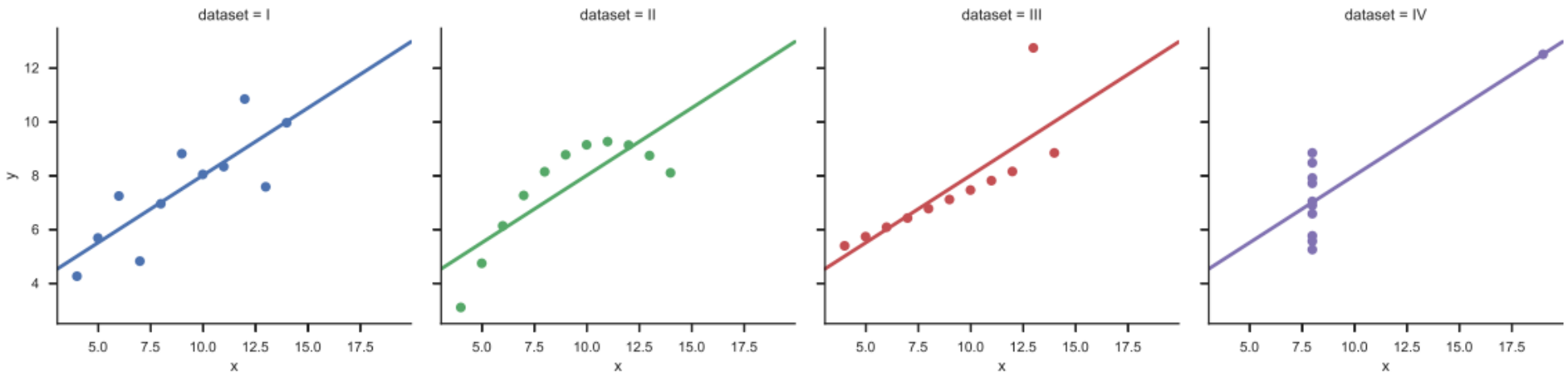
Lineare Regression – (2) Linearitätstest

- Korrelationskoeffizient nach Pearson



Lineare Regression – (2) Linearitätstest

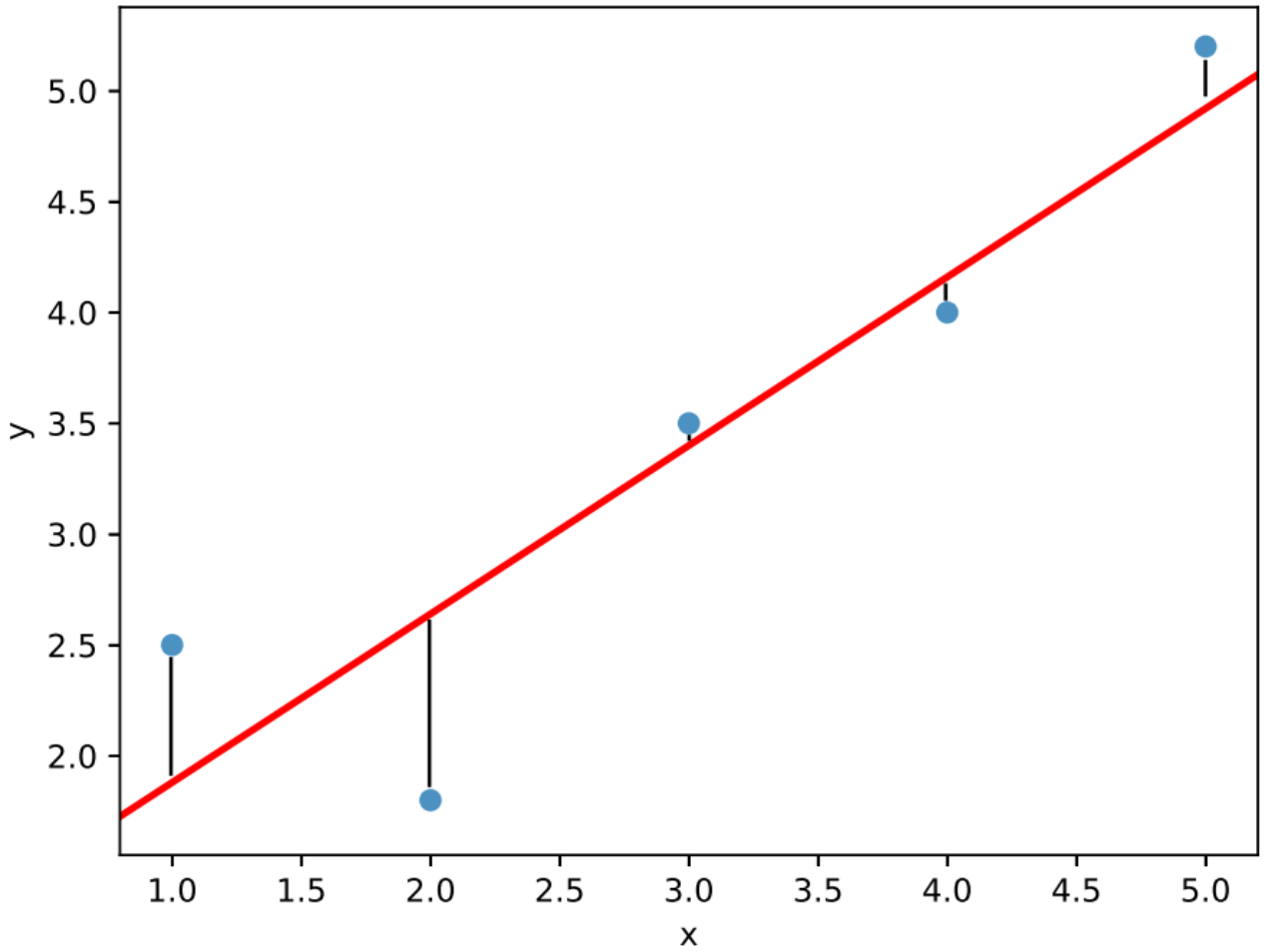
ABER: Linearitätstest ist keine absolute Garantie für das Vorliegen einer linearen Beziehung!



- Alle **vier Datensätze** haben den **gleichen Mittelwert**, die **gleiche Varianz**, den **gleichen Korrelationskoeffizienten** sowie die gleiche **optimale Regressionsgerade**

Lineare Regression – (3) Bestimmung der Optimalen Parameter

Lege Gerade so, dass Abstände zu Trainingsdatenpunkten minimiert werden:



Lineare Regression – (3) Bestimmung der Optimalen Parameter

Verwende dafür

- Mittlere Quadratische Abweichung (MQA)

$$\mathbf{MQA} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y_i} - \hat{\varphi}(\mathbf{X_i}))^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y_i} - w_0 - w_1\mathbf{X_i})^2$$

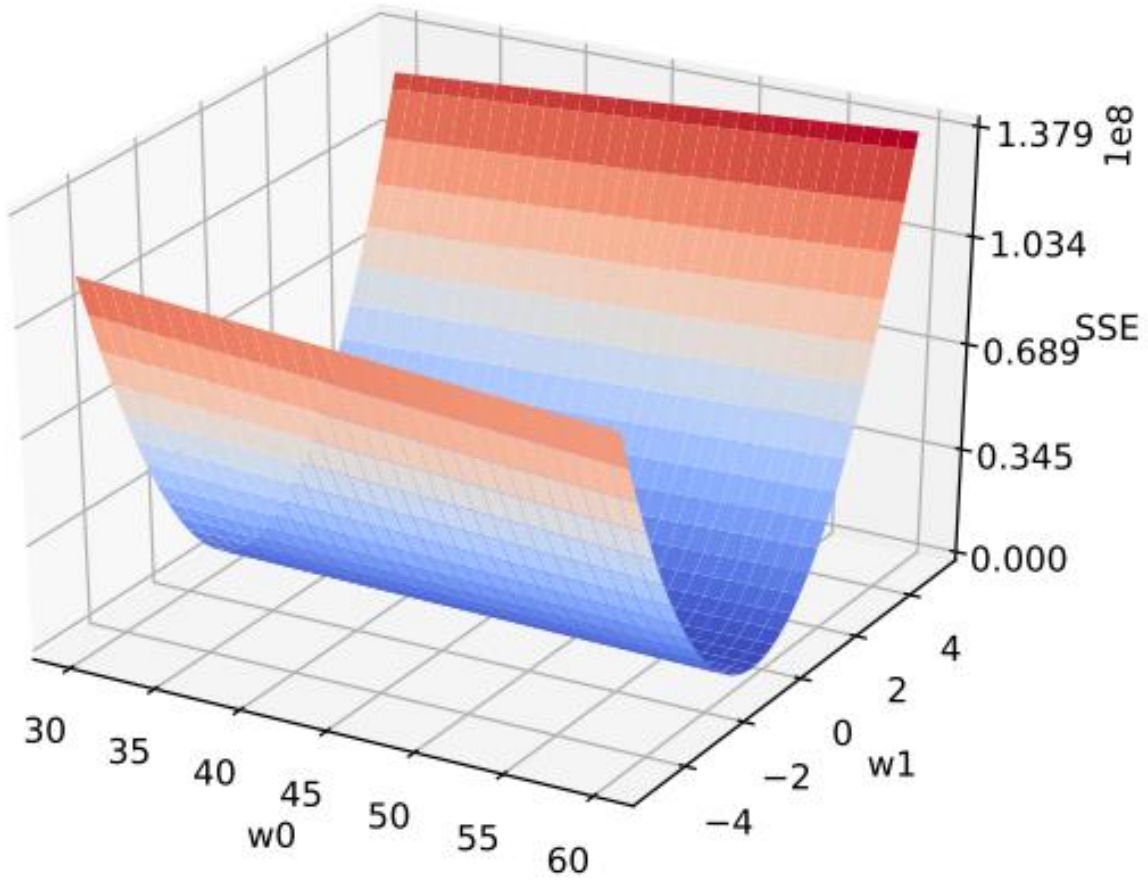
Lineare Regression – (3) Bestimmung der Optimalen Parameter

- **Ziel:**
  - Bestimme Parameter w_0 und w_1 in MQA(w_0 , w_1) so, dass Straffunktion minimal wird.
  - Optimierungsproblem der Abbildung MQA(w_0 , w_1) in den Parametern w_0 und w_1 :

$$\underset{\mathbf{w}_0,\mathbf{w}_1}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (\mathbf{Y_i} - \mathbf{w}_0 - \mathbf{w}_1\mathbf{X_i})^2 \right)$$

Lineare Regression – (3) Bestimmung der Optimalen Parameter

MQA für das Beispiel (Autodaten)



Lineare Regression – (3) Bestimmung der Optimalen Parameter

Optimale Parameterwerte w_0 und w_1 lassen sich im Fall der linearen Regression analytisch bestimmen:

- Berechne die (partiellen) Ableitungen von MQA nach w_0 und w_1:

$$\frac{\partial(\mathrm{MQA})}{\partial w_0} = -\frac{2}{n}\sum_{i=1}^{n}(\mathbf{Y_i} - w_0 - w_1\mathbf{X_i})$$

$$\frac{\partial(\mathrm{MQA})}{\partial w_1} = -\frac{2}{n}\sum_{i=1}^{n}(\mathbf{Y_i} - w_0 - w_1\mathbf{X_i})\,\mathbf{X_i}$$

- Ermittle gemeinsame Nullstelle(n) durch Lösen des Gleichungssystems

$$\frac{\partial(\mathrm{MQA})}{\partial w_0} = 0 \qquad \frac{\partial(\mathrm{MQA})}{\partial w_1} = 0$$

Lineare Regression – (3) Bestimmung der Optimalen Parameter

- Die beiden Gleichungen $\frac{\partial(\mathrm{MQA})}{\partial \mathbf{w}_0} = 0$ und $\frac{\partial(\mathrm{MQA})}{\partial \mathbf{w}_1} = 0$ lassen sich geschlossen lösen.

- Die stationäre Lösung $\mathbf{w_{(0)}} = (w^*_0 , w^*_1)$ ist:

$$w_0^* = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - w_1^* \sum_{i=1}^{n} x_i\right)$$

$$w_1^* = \frac{n\sum_{i=1}^{n}(x_i \cdot y_i) - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

(Übungsaufgabe: Herleitung dieser Lösung)

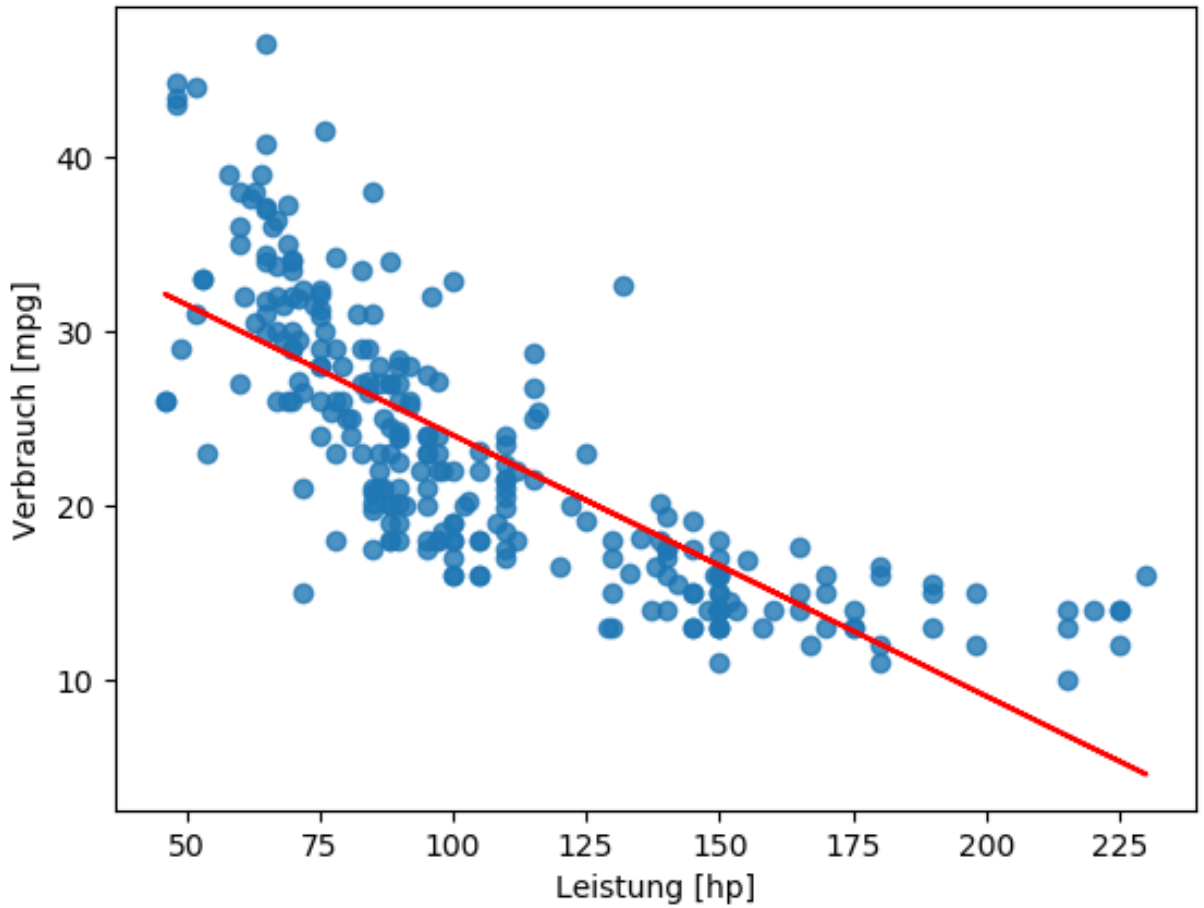
Lineare Regression – (3) Bestimmung der Optimalen Parameter – Python / scikit-learn

```
import numpy as np
import pandas as pd
import os
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import linear_model

# Konfiguration
FILE_DATA = os.path.join(...)

# Autodaten einlesen
cars = pd.read_csv(FILE_DATA)

# Verbrauchs- und Leistungswerte
mpg = cars.iloc[:,0].values
hpAR = cars.iloc[:,[3]].values
```

```
# Einfache lineare Regression
reg = linear_model.LinearRegression()
reg.fit(hpAR,mpg)

# Plot erstellen
g = sns.regplot(x=hpAR, y=mpg, fit_reg=False)
plt.plot(hpAR, reg.predict(hpAR), color='red')
plt.xlabel('Leistung [hp]')
plt.ylabel('Verbrauch [mpg]')
plt.show()
```

Lineare Regression – (3) Bestimmung der Optimalen Parameter – Python / scikit-learn

- Optimale Parameter für die Beispieldaten:
  - $w_0^* = 39{,}9359$ $w_1^* = -0{,}1578$



Lineare Regression – (3) Bestimmung der Optimalen Parameter

Zusammenfassung

- Verwendung der mittleren quadratischen Abweichung MQA zwischen eigentlichem Wert und Vorhersagefunktionswert (im Beispiel: lineares Modell)
- Ziel: Bestimme Parameter des Modells so, dass MQA minimal wird
  - Optimale Parameter sind stationäre Punkte dieses Minimierungsproblems
- Quadratische Abweichung: Parameter lassen sich analytisch bestimmen
  - das ist in der Regel nicht möglich für Beträge oder höhere Potenzen
- Lösung (lineares Modell): Regressionsgerade / Regressionsebene / Regressionshyperebene

Assoziations- und Musteranalyse

Assoziation

- Identifikation von Abhängigkeiten zwischen Objekten oder Attributen
- unüberwachtes Lernen / unsupervised learning
- Methoden:
  - **Frequent Pattern Mining**
  - Korrelationsanalysen
- Beispiele:
  - Warenkorbanalysen
  - Cross-Selling-Angebote

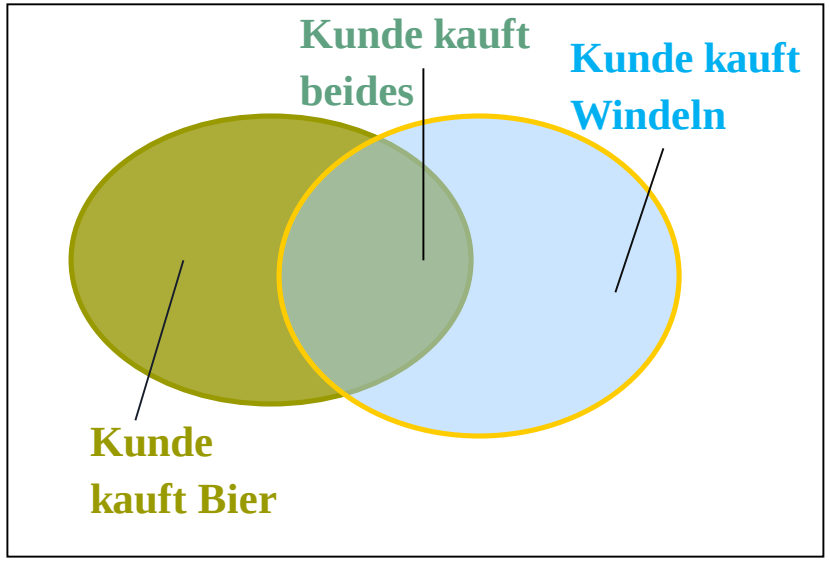
Assoziations- und Musteranalyse

- **Frequent Pattern Mining**
  - Mustererkennung in Produktmengen → “Muster in Warenkörben”
- Motivation: Finde inhärente Regelmäßigkeiten in Daten
  - Welche Produkte werden oft zusammen gekauft – Bier und Windeln?
  - Welche Produkte werden mit einem Computer gekauft?
- Anwendungen
  - Warenkorbanalysen
  - Cross-Selling
  - Entwurf von Gruppen von Katalogen und Verkaufskampagnen

Grundlegendes Konzept: Häufige Artikel-Menge

Datenbank D

Tid	Warenkorb / Transaktion
10	Bier, Nüsse, Windeln
20	Bier, Kaffee, Windeln
30	Bier, Eier, Windeln
40	Bier, Eier, Milch, Nüsse
50	Eier, Kaffee, Milch, Nüsse



- **Artikel-Menge**:
  - $X = \{x_1, \ldots, x_k\}$. Menge von Artikeln aus vorgegebener Grundmenge (Sortiment) G
- **Transaktion:** Satz von Artikeln (“Warenkorb”)
- **Datenbank D:** Liste von Transaktionen (mit ID)
- **support von X:** Relative Häufigkeit des Vorkommens der Artikel-Menge X in allen Transaktionen der Datenbank

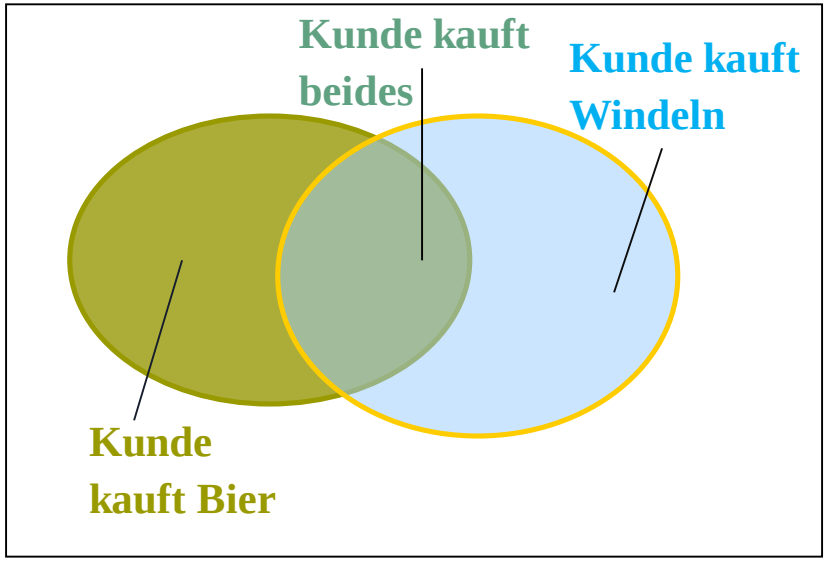
$$\text{support}_D(X) = \frac{|\{\, t \text{ Transaktion} \in D \mid X \sqsubseteq t \,\}|}{|D|}$$

- X heißt **häufige Artikel-Menge**, falls der support von X größer oder gleich einer bestimmten, festgelegten **Support-Schwelle** S ist:

$$\text{support}_D(X) \geq S$$

Grundlegendes Konzept: Häufige Artikel-Menge

Tid	Warenkorb / Transaktion
10	Bier, Nüsse, Windeln
20	Bier, Kaffee, Windeln
30	Bier, Eier, Windeln
40	Bier, Eier, Milch, Nüsse
50	Eier, Kaffee, Milch, Nüsse



Beispiel:

G = {Bier, Eier, Kaffee, Milch, Nüsse, Windeln, Orangen}

X = {Bier, Windeln}

S = 0,5 (festgelegt)

- Somit: $support_D(X) = 3/5 = 0,6$
- Also: Artikel-Menge X ist häufig, weil

$$support_D(X) \geq S$$

Assoziationsregeln

- **Assoziationsregeln**
  - durch Korrelationen zwischen gemeinsam auftretenden Artikel-Mengen X und Y festgelegt

- Für Assoziationsregeln sind (mindestens) folgende Korrelationsmaße relevant:
  - **Support**: Gewicht des gemeinsamen Auftretens
  - **Konfidenz**: Relatives Gewicht des gemeinsamen Auftretens

Assoziationsregeln

- Seien X und Y zwei Artikel-Mengen:
  - $X = \{x_1, \ldots, x_k\}$, $Y = \{y_1, \ldots, y_n\}$
  - $X \cup Y = \{x_1, \ldots, x_k, y_1, \ldots, y_n\}$ (Vereinigungsmenge von X und Y)
- Korrelationsmaße für Assoziationsregel zwischen X und Y

  $support_D (X \cup Y)$ = Gewicht des gemeinsamen Auftretens von X und Y

  $confidence_D (X,Y)$ = Relatives Gewicht des gemeinsamen Auftretens

  $= support_D (X \cup Y) / support_D (X)$

  (Anteil Transaktionen mit X, die auch Y enthalten)

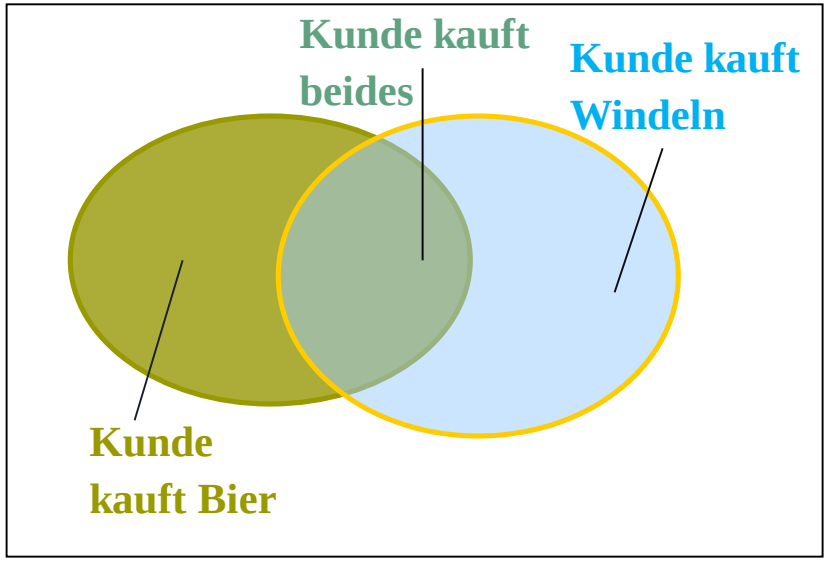
- Dann gelte die **Assoziationsregel X → Y** (“Aus X folgt Y”), wenn
  - $support_D (X \cup Y)$ ≥ **S (festgelegte Support-Schwelle)**
  - $confidence_D (X,Y)$ ≥ **C (festgelegte Konfidenzschwelle)**

In der Sprache von Warenkörben:
X → Y heißt: “Wer X kauft, kauft (häufig) auch Y”

Assoziationsregeln – Beispiel 1

X → Y
$support_D (X \cup Y) \geq$ **S**
$confidence_D (X,Y) = support_D (X \cup Y) / support_D (X) \geq$ **C**

Tid	Warenkorb / Transaktion
10	Bier, Nüsse, Windeln
20	Bier, Kaffee, Windeln
30	Bier, Eier, Windeln
40	Bier, Eier, Milch, Nüsse
50	Eier, Kaffee, Milch, Nüsse

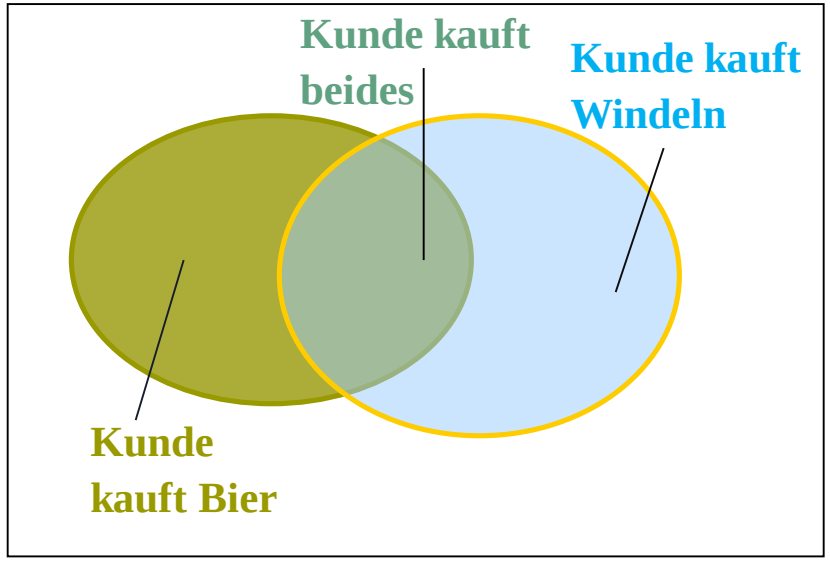


- X = {Bier}, Y = {Windeln}
- Support-Schwelle: S = 0,5
- Konfidenzschwelle: C = 0,8
- Dann:
  - $X \cup Y$ = {Bier, Windeln}
  - $support_D (X \cup Y) = 0,6$
  - $confidence_D (X,Y) = 3/4 = 0,75$
- Also gilt **nicht** die Assoziation X → Y

Assoziationsregeln – Beispiel 2

X → Y
$support_D (X \cup Y) \geq$ **S**
$confidence_D (X,Y) = support_D (X \cup Y) / support_D (X) \geq$ **C**

Tid	Warenkorb / Transaktion
10	Bier, Nüsse, Windeln
20	Bier, Kaffee, Windeln
30	Bier, Eier, Windeln
40	Bier, Eier, Milch, Nüsse
50	Eier, Kaffee, Milch, Nüsse



- X = {Windeln}, Y = {Bier}
- Support-Schwelle: S = 0,5
- Konfidenzschwelle: C = 0,8
- Dann:
  - $X \cup Y$ = {Bier, Windeln}
  - $support_D (X \cup Y) = 0,6$
  - $confidence_D (X,Y) = 3/3 = 1$
- Also gilt die Assoziation X → Y

Assoziationsregeln – Beispiel 3

X → Y
$\text{support}_D(X \cup Y) \geq \mathbf{S}$
$\text{confidence}_D(X,Y) = \text{support}_D(X \cup Y) / \text{support}_D(X) \geq \mathbf{C}$

- Datenbank D mit 10.000 Transaktionen
- Artikel-Grundmenge (Sortiment) G = {Computer, Monitor, Maus, …}
- Support-Schwelle S = 0,4
- Konfidenzschwelle C = 0,6

- X = {Computer} $\text{support}_D(X) = 6.000/10.000 = 0,6$
- Y = {Monitor} $\text{support}_D(Y) = 7.500/10.000 = 0,75$
- $X \cup Y$ = {Computer, Monitor} $\text{support}_D(X \cup Y) = 4.000/10.000 = 0,4$

- Dann:
  - $\text{support}_D(X \cup Y) = 0,4$
  - $\text{confidence}_D(X,Y) = 2/3 = 0,66$
  - $\text{confidence}_D(Y,X) = 0,4/0,75 = 8/15 < 0,6$
  - Also gilt die Assoziation **X → Y** aber **nicht** die Assoziation **Y → X**

Assoziationsregeln – Beispiel 3

X → Y
$\text{support}_D (X \cup Y) \geq \mathbf{S}$
$\text{confidence}_D (X,Y) = \text{support}_D (X \cup Y) / \text{support}_D (X) \geq \mathbf{C}$

Zusammenfassung (drittes Beispiel)

- Wahrscheinlichkeit (Monitor)-Kauf = $\text{support}_D (Y) = 0{,}75$
- Wahrscheinlichkeit (Monitor & Computer)-Kauf = $\text{support}_D (X \cup Y) = 0{,}4$
- Wahrscheinlichkeit Monitorkauf bei Computerkauf = $\text{confidence}_D (X,Y) = 0{,}66$
- Da sowohl Support-Schwelle als auch Konfidenzschwelle überschritten, gilt also
  - X → Y und damit Aussage der Maschine: “Aus Computerkauf folgt Monitorkauf”

Aber:

- Wahrscheinlichkeit Monitorkauf größer als Wahrscheinlichkeit Monitorkauf bei Computerkauf

Aussage der Maschine irreführend! Führt zu falschen Geschäftsentscheidungen.

Konsequenz: Support und Konfidenz oft nicht ausreichend

- (!) Erweiterungen der Assoziationsregel “X → Y” notwendig

Erweiterte Assoziationsregeln

Erweiterte Assoziationsregel

X ⇒ Y [support, confidence, $corr_1$, … , $corr_n$] (“Aus X folgt **stark** Y”)

wenn

- $support_D(X \cup Y)$ ≥ S
- $confidence_D(X,Y)$ ≥ C
- $corr_1(X,Y)$ ≥ K_1 (weiteres Korrelationsmaß mit Schwelle K_1)

  …
- $corr_n(X,Y)$ ≥ K_n (weiteres Korrelationsmaß mit Schwelle K_n)

- Beispiel für solch ein weiteres Korrelationsmaß: **lift**

$$\text{lift}_D(X,Y) = \frac{\text{support}_D(X \cup Y)}{\text{support}_D(X) \times \text{supportD}(Y)} = \frac{\text{confidence}_D(X,Y)}{\text{support}_D(Y)}$$

- positive Korrelation: $lift_D(X,Y) > 1$
- neutrale Korrelation: $lift_D(X,Y) \approx 1$
- negative Korrelation: $lift_D(X,Y) < 1$

Erweiterte Assoziationsregeln – Lift

Für unsere Zwecke relevant:

Erweiterte Assoziationsregel mit Lift

X ⇒ Y [support, confidence, lift] ("Aus X folgt **stark** Y")

wenn

- $\text{support}_D(X \cup Y) \geq S$
- $\text{confidence}_D(X,Y) \geq C$
- $\text{lift}_D(X,Y) > 1$

mit vorgegebenen Schwellenwerten S (Support) und C (Confidence)

Assoziationsregeln – Beispiel 3

Übungsaufgabe: Ermitteln Sie, ob für folgendes Beispiel eine starke Assoziation **X ⇒ Y** oder **Y ⇒ X** vorliegt. (Beachten Sie: Es gilt die Assoziation **X → Y** aber **nicht** die Assoziation **Y → X**)

- Datenbank D mit 10.000 Transaktionen
- Artikel-Grundmenge (Sortiment) G = {Computer, Monitor, Maus, …}
- Support-Schwelle S = 0,4
- Konfidenzschwelle C = 0,6

- X = {Computer} $support_D(X) = 6.000/10.000 = 0,6$
- Y = {Monitor} $support_D(Y) = 7.500/10.000 = 0,75$
- $X \cup Y$ = {Computer, Monitor} $support_D(X \cup Y) = 4.000/10.000 = 0,4$

Frequent-Itemset-Bestimmung

- Die Assoziationen **X → Y** und **X ⇒ Y** verwenden Korrelationsmaße, die auf der Ermittlung des Supports von Artikelmengen und der Bestimmung von deren Häufigkeit (**frequent item sets**) beruhen

- **Ziel:** Bestimmung der häufigen Artikel-Mengen (**frequent item sets**) in einer Transaktionsdatenbank

Frequent-Itemset-Bestimmung – Apriori-Eigenschaft

Für Frequent-Itemset-Bestimmung verwende die **Apriori-Eigenschaft**:

Jede Untermenge einer häufigen Artikel-Menge ist häufig

Veranschaulichung der Apriori-Eigenschaft:

- Wenn X = {Bier, Nüsse, Windeln} häufige Artikel-Menge, dann auch {Bier, Nüsse}
- Denn jede Transaktion, die {Bier, Nüsse, Windeln} enthält, enthält auch {Bier, Nüsse}

Frequent-Itemset-Bestimmung – Apriori-Ausschlusseigenschaft

Aus Apriori-Eigenschaft folgt im Umkehrschluss die ***Apriori-Ausschlusseigenschaft***:

Ist eine Untermenge X einer Artikel-Menge Y nicht häufig, dann ist auch die Artikel-Menge Y selbst nicht häufig.

Begründung dieser Aussage durch Widerspruch:

- Gegeben/Voraussetzung: $X \subseteq Y$ und X nicht häufig
- Annahme: Y häufig
- Aus Annahme folgt dann: X häufig wegen Apriori-Eigenschaft.
  - Aber nach Voraussetzung ist X nicht häufig!
  - Widerspruch!
- Schlussfolgerung: Annahme ist falsch und daher Y nicht häufig

Frequent-Itemset-Bestimmung – Apriori-Ausschneideprinzip

Apriori-Ausschlusseigenschaft führt zu Apriori-Ausschneideprinzip (Pruning Principle):

Wenn eine Artikel-Menge X nicht häufig ist, dann ist jede Obermenge von X nicht häufig und kann bei der Suche nach häufigen Artikelmengen verworfen werden.

Frequent-Itemset-Bestimmung – Apriori-Algorithmus

Verfahren zur Bestimmung häufiger Artikelmengen (Apriori-Algorithmus)

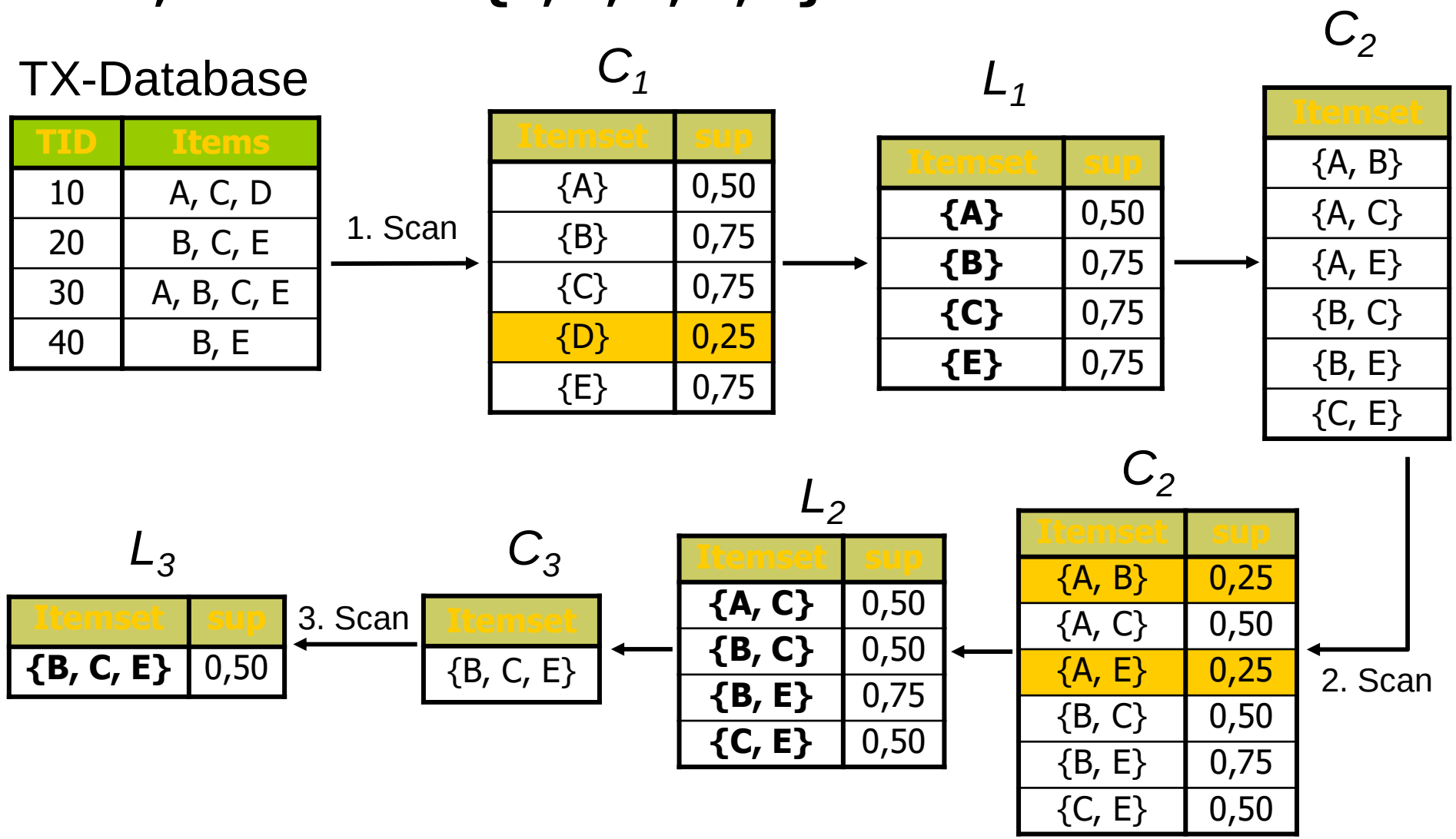
- beruht auf Apriori-Ausschneideprinzip (Pruning Principle)
- Notation: X ist **k-Artikel-Menge**, wenn $X = \{x_1, \ldots, x_k\}$ k Artikel enthält

Apriori-Algorithmus

1. Scanne die Transaktionsdatenbank, um alle häufigen 1-Artikel-Mengen zu bestimmen. Nehme die häufigen 1-Artikel-Mengen in der Menge L_1 auf.
2. Setze nun k = 1 und $L_k = L_1$
3. Erzeuge aus L_k alle (k+1)-Artikel-Mengen, die nur häufige Artikel-Mengen aus $L_k, L_{k-1}, \ldots, L_1$ enthalten und nehme diese (k+1)-Artikel-Mengen in C_{k+1} auf. (Apriori-Ausschneideprinzip!)
4. Test jeden Kandidaten in C_{k+1} auf seine Häufigkeit.
5. Nehme die so ermittelten häufigen (k+1)-Artikel-Mengen in der Menge L_{k+1} auf.
6. Beende den Prozess, falls L_{k+1} die leere Menge ist.
7. Ansonsten setze k := k + 1 und fahre mit Schritt (3) fort.

Apriori-Algorithmus – Beispiel

S = 0,5 **G = {A, B, C, D, E}**



Apriori-Algorithmus (Pseudo-Code)

```
C : Candidate itemsets
L : Frequent itemsets

input (database, threshold)
L = {frequent items};
FIS = L
 for (k = 1 ; L != ∅ ; k++) do begin
    C = (k+1)-candidates generated from L and database
    if (C == ∅) break
    for each transaction t in database do
     increment count of all candidates in C that are contained in t
    L = (k+1)-candidates in C with support ≥ threshold
    FIS = FIS ∪ L
 end
return FIS
```

Frequent-Itemset-Bestimmung – Algorithmen

- Bekannte Vertreter von Frequent-Itemset-Algorithmen:
  - Apriori-Algorithmus
  - FPGrowth: Nutzt das Frequent-Pattern-Growth-Prinzip
  - ECLAT: Frequent Pattern Mining mit vertikalem Datenformat
  - …

Cluster-Analyse

Gruppen-/Cluster-Bildung

- Identifizierung von Gruppen oder Clustern gleichartiger Objekte
- auf Basis von Ähnlichkeitsmerkmalen
  - Objekte in Cluster mit möglichst ähnlichen Merkmalen einteilen
  - Objekte unterschiedlicher Cluster mit möglichst verschiedenen Merkmalen
- Eigenschaften und Merkmale für Clusterbildung i. a. nicht vorgegeben
- unüberwacht, Cluster-Bildung nur durch Datenanalyse ohne Prediktor-Variablen
- Methoden:
  - **Cluster-Analysen wie k-Means, etc.**
  - neuronale Netze

Anwendungsfelder für Cluster-Analyse

- Warenkorbanalysen: Muster von Warenkörben entdecken → Warenkorbgruppen
- Text Mining: Dokumenten-Clustering
- Ausreißerermittlung: Datenobjekte, die “weit weg” von jedem Cluster liegen
- Marketing:
  - Bestimmung unterschiedlicher Kundengruppen oder -segmente für das segmentspezifische Marketing
- Stadtplanung:
  - Identifikation von Häusergruppen nach Haustyp, Wert, Lage
- Klima: Atmosphärische und ozeanische Mustererkennung

Anwendungsfelder für Cluster-Analyse

- Datenreduktion
  - Vorbereitung für Klassifikationsanalyse und Regression
- Prognosen basierend auf Gruppen/Clusters
  - Bestimme characteristisches Muster (Klasse) pro Cluster

Cluster-Analyse

- Cluster-Analyse
  - Finde homogene Teilmengen (Cluster) von Objekten aus heterogener Gesamtheit von Objekten
    - Objekte innerhalb eines Cluster sind homogen
    - Objekte aus verschiedenen Clustern sind heterogen
  - Häufige Methode: Identifiziere Objekt-Merkmale, anhand derer sich Cluster bilden lassen.
    - Ähnlichkeit von Datenobjekten bezüglich dieser Merkmale

Einige Typen von Cluster-Methoden

- Partitionierung
  - Konstruktion von disjunkten Objektmengen
- Frequent-Pattern-basiert
  - Analyse von Frequent Patterns
  - “Warenkorbanalysen” – Muster von Warenkörben
- Verbindungsbasiert
  - Bestimmung der Verbindungen/Assoziationen zwischen Objekten
  - stark zusammenhängende Objekte gehören zusammen → Netzwerkanalyse
- …

Qualitätskriterien für Cluster-Methoden

- Erzeugung hochwertiger Cluster
  - hohe Intra-Cluster-Ähnlichkeit: Kohäsion
  - geringe/keine Inter-Custer-Ähnlichkeit
- Qualität der Methoden hängt ab von
  - Ähnlichkeitsmaß
  - Realisierung/Implementierung
  - Fähigkeit, viele oder alle versteckte Muster/Cluster zu entdecken
- Qualitätsfunktion zur Messung der Güte einer Cluster-Methode oder eines Clustering
  - Was bedeutet “ähnlich genug” oder “gut genug”?
  - Oft subjektive Kriterien

Vorbereitung einer Cluster-Analyse

- Merkmal-Auswahl
  - entsprechend der Aufgabe oder der Vorgaben
  - Minimierung von redundanter Information
- Festlegung eines Ähnlichkeitsmaßes
  - auf den Merkmalen
- Cluster-Kriterium
  - basierend auf Ähnlichkeitsmaß
- Geeignete Wahl eines Cluster-Algorithmen
- Validierung und Interpretation der Ergebnisse

Ähnlichkeitsmaße

- Distanzbasiert
  - in Form von Abstandsfunktionen/Metriken $d(J, K)$ zwischen zwei Datenobjekten J und K
  - z. B. Euklidische Metrik, Straßennetzwerk, …
  - i. a. Verwendung im partitionsbasierten Clustering
- Link-basiert
  - Dichte von Datenpunkten
  - Anzahl von Verbindungen/Links
  - i. a. Verwendung im verbindungsbasierten Clustering

Distanzbasierte Partitionierung

- Voraussetzung: Datenobjekte haben numerische Merkmalsvektoren
- Finde Partitionierung der Datenobjekte in ***k*** (a priori unbekannte) Clusters
  - so dass Summe S der (quadrierten) Abstände zu Cluster-Schwerpunkt c_i des Clusters C_i minimal ist (für alle i $\in$ {1,..,k})

$$S = \sum_{i=1}^{k} \sum_{P \in C_i} (d(P, c_i))^2$$

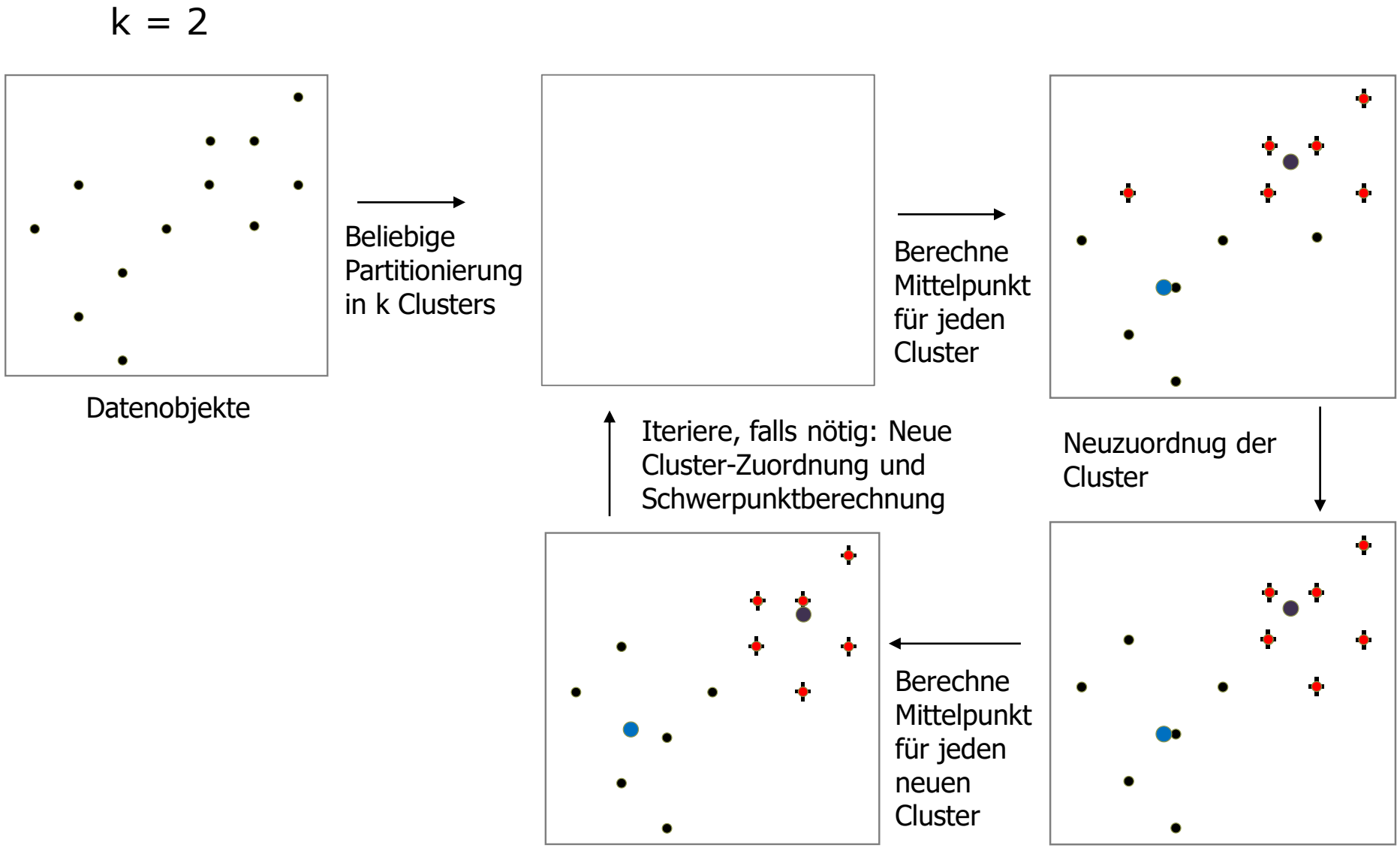
- Wichtiger Vertreter: **k-Means-Algorithmus**

k-Means-Algorithmus zur distanzbasierten Partitionierung

Vorgegeben: $k \in \mathbb{N}$ und Menge von Datenobjekten (Datenbank)

1. Partitioniere die Objekte in *k* nichtleere Clusters C_i , $i \in \{1, .. k\}$
2. Berechne die Schwerpunkte c_i jedes Clusters C_i aus den Datenobjekten dieses Clusters
3. Bilde neue Clusters C'_i durch Zuordnung jedes Datenobjektes zu dem ihm nähesten Punkt c_i
4. Fahre mit Schritt 2 fort, bis sich die Cluster nicht mehr verändern

Beispiel für Clustering durch k-Means-Algorithmus



Bewertung des k-Means-Algorithmus

- Stärke
  - Effizient mit Laufzeit $\mathbf{0}(k * m * n * t)$
    - k = # Clusters, m = # Datenobjekte, n = # Dimensionen, t = # Iterationen
    - Normalerweise k, t << m
- Schwächen
  - Nur in n-dmensionalen numerischen bzw. metrischen Räumen anwendbar
  - Cluster-Anzahl k muss vorab festgelegt werden
    - Erweiterter Algorithmus: Automatische Bestimmung eines “guten” k
  - Empfindlich gegen Ausreißer:
    - Datenobjekt mit sehr großen Werten kann Clusterverteilung erheblich verzerren
  - Algorithmus terminiert oft in einem lokalen Optimum
  - Terminierung nicht immer gewährleistet!

Fragen?

Weiterhin viel Erfolg im Studium!