UNIVERSITÄT TRIER

Business Analytics II
Methodische Ansätze

Dr. Holger Steinmetz
Lehrstuhl für Unternehmensführung
Universität Trier

Kategorien von Ansätzen (Wiederholung)

UNIVERSITÄT TRIER

- **Teil der Informationsstrategie: Definition der Ziele** (→ Was ist die relevante / zu liefernde Information?):
  - **Deskriptiv:** Beschreibung von
    - Zielvariablen (z.B. KPIs)
    - Zeitlichen Verläufen, einfachen Zusammenhängen (z.B. Geschlechtsunterschiede)
    - Identifikation von Mustern (z.B. Clusteranalyse)
  - **Kausal** (Explanation):
    - Frage nach den **Ursachen** einer Target-Variable (z.B. "warum laufen die Kunden weg?")
    - Frage nach der **Wirkung** einer Maßnahme (z.B. "wie wirkt sich eine Preiserhöhung aus")

    → Problem der **kausalen Identifikation**
  - **Vorhersage** (Prediction): Was weiß ich über Y (das Target), wenn ich Wissen über X habe?
    - Querschnittlich: Machine learning-Ansätze (ML)
    - Längsschnittlich: Forecasting (u.U. auch mit in Kombination mit ML)

UNIVERSITÄT TRIER

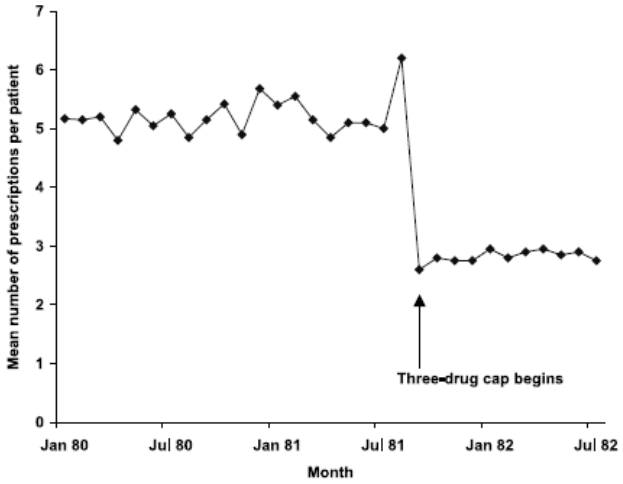
Analysen mit kausalem Fokus

Hauptziel kausaler Analysen: Kausale Identifikation

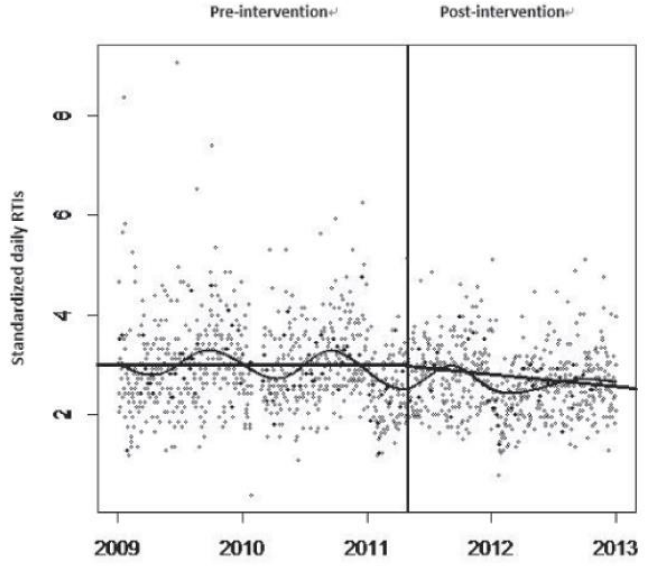
UNIVERSITÄT TRIER

Lösungen bei Interventionen:

- **Experimentell:** Randomized controlled trial (z.B. AB testing bei Pricing Models)
- **Quasi-Experimentell:**
  - **Gruppenvergleich** mit u.U. selbst-selektierten / nicht identischen Gruppen (→ Kontrollvariablen)
  - **Längsschnittdaten: Vorher-Nachher-Vergleich** oder **Interrupted times series**



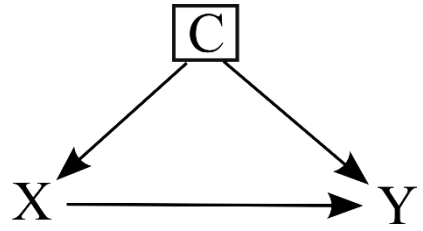
Einführung einer Begrenzung der Medikamente pro Patient



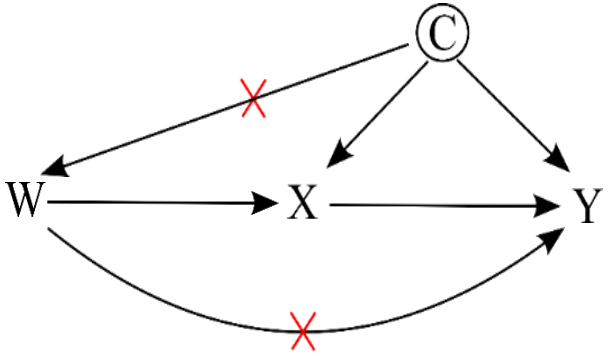
Einführung eines Gesetzes gegen Alkohol am Steuer in einer chinesischen Provinz

Hauptziel kausaler Analysen: Kausale Identifikation

Lösungen bei Vorliegen von nicht-experimentellen Daten (observational data):



- **Kontrollvariablen:** Statistische Kontrolle von "Confoundern" (die Scheinkorrelationen verursachen)



- **Instrumentalvariablen:** Identifikation von Variablen W, die
  a) mit X hoch korrelieren
  b) keinen direkten Effekt auf Y haben und
  c) unkorreliert mit Confoundern C (die nicht bekannt oder als Daten verfügbar sind)

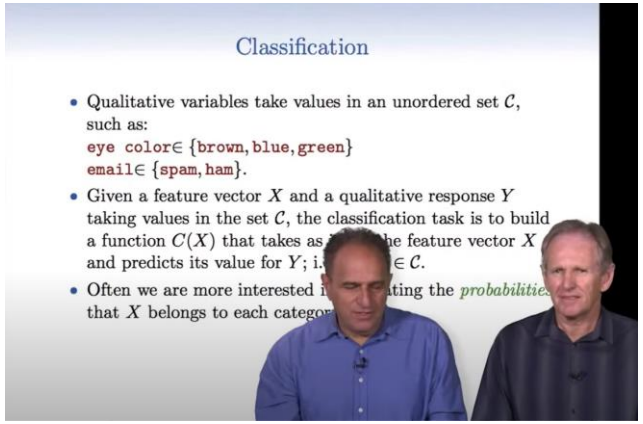
  → Two-stage-least squares regression

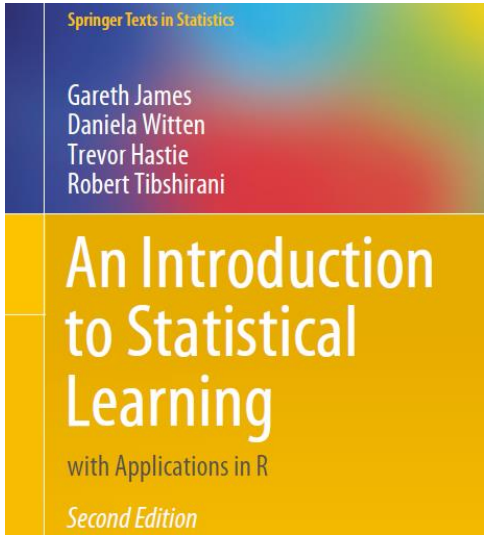
- **Längsschnittliche Modelle:** z.B. VAR-Models
  → Adressiert reverse causation-Möglichkeit, aber nicht Confounding

UNIVERSITÄT TRIER

Analysen mit prädiktivem Fokus: Machine learning

James et al.

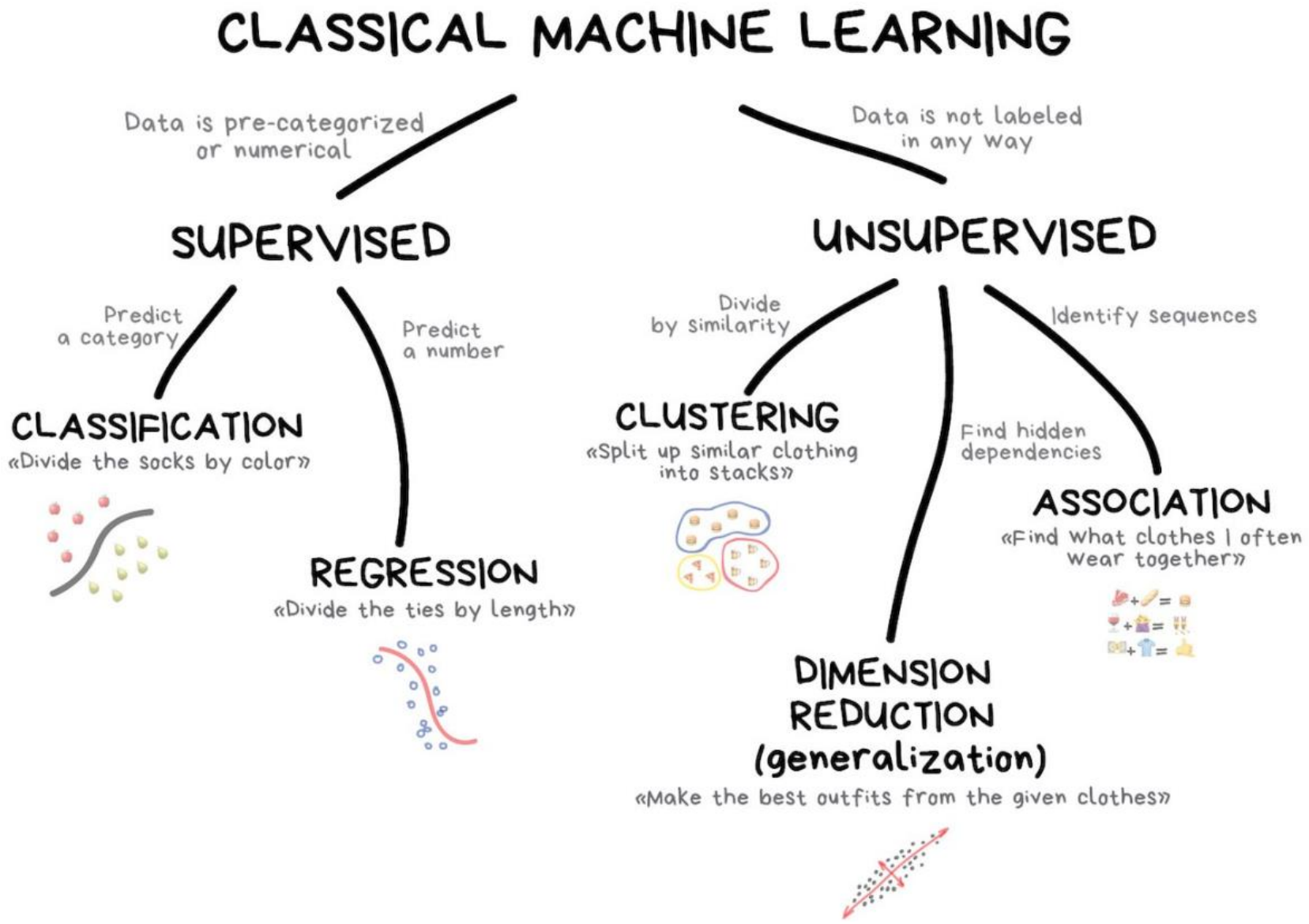




- Kapitel 2 (Statistical Learning): Empfehlenswert (Video: https://shorturl.at/rsxDW)
- Nur bei Bedarf/Interesse
  - Kapitel 3: Refresher lineare Regression (https://shorturl.at/dmxGL
  - Kapitel 4: Refresher logistische Regression: https://t.ly/LBsA (nur bis 4.4 relevant)

    → Aber: Die Szenarien/Daten in Kap. 3 und 4 verwende ich im Tidymodels-Skript
- Kap. 5 (Resampling Methods): Empfehlenswert (https://rb.gy/h4vcz)

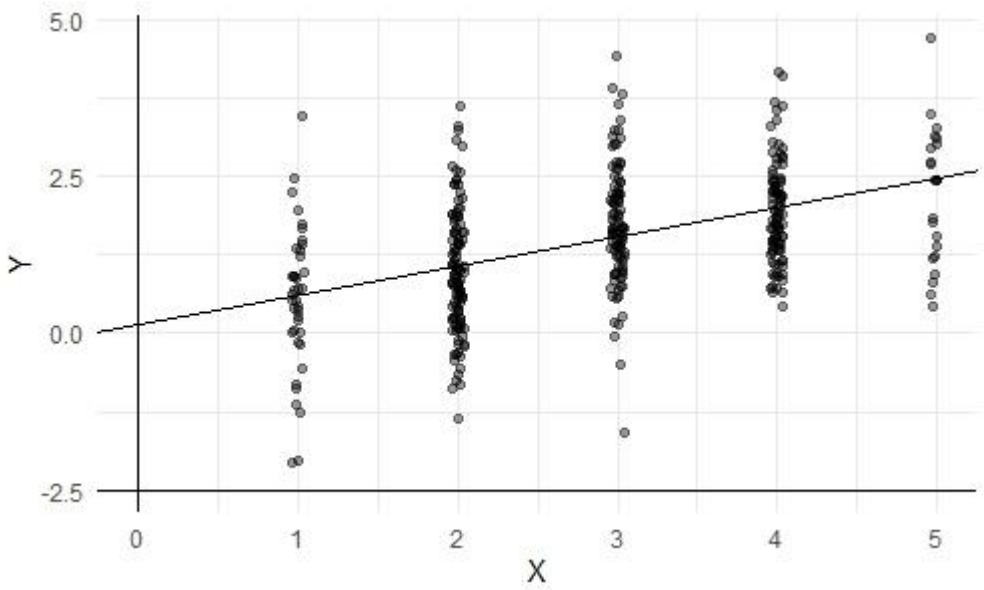
Arten des ML

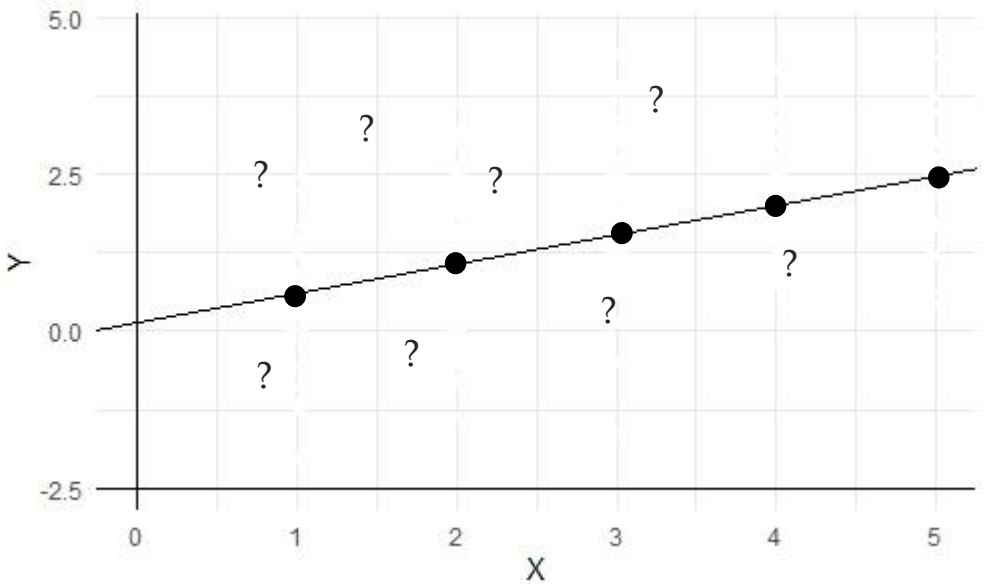

CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
Find hidden dependencies
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

UNIVERSITÄT TRIER

Supervised machine learning: Regression

Das Regressionsmodell als einfachstes Modell





- Szenario: "Wenn ich Informationen über X habe: Sagt mir das etwas über Y"
- **Modell:** Beziehung wird beschrieben durch einen lineare Funktion

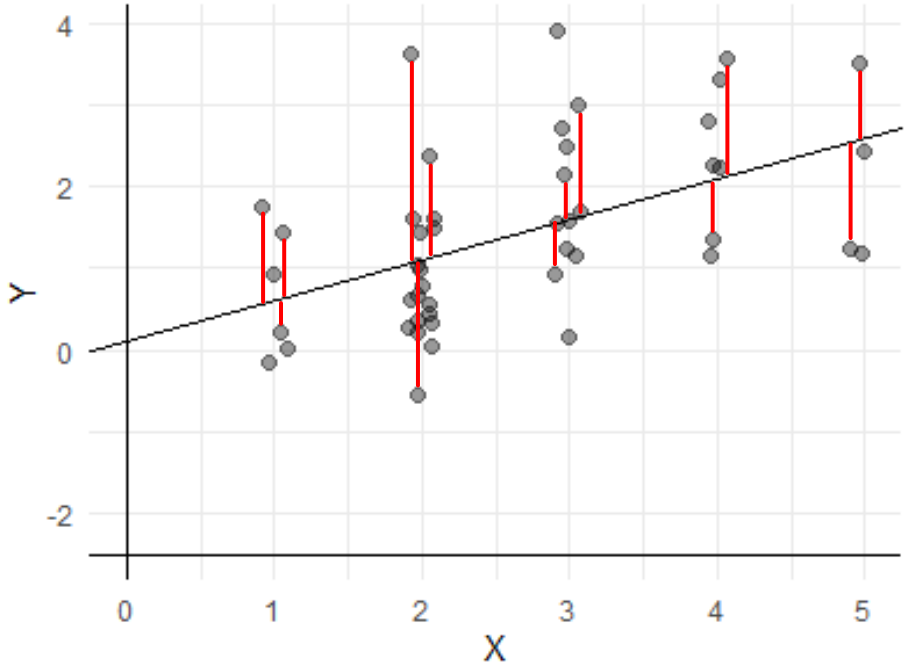
$$E(Y|X) = \beta 0 + \beta 1 X$$

- Schätzung erfolgt mit ordinary least squares (oder kurz least squares, **LS**)
- **Ziel:** *Out-of-sample prediction* (Vermutungen über Y, wenn man keine Daten hat)

→ **Predicted values**

Prediction Errors (Vorhersagefehler)

- **Residuen:** Individuelle Abweichungen von der Regressionsgerade



- Verursacht von
  - Weiteren/nicht-einbezogenen Einflussfaktoren
  - Zufallsfehler
- Konsequenzen:
  - Faktisch (Fehlentscheidung, Kosten etc.)
  - Ethische Relevanz

  → Zentrale Frage: Was ist die Alternative?
- Vorhersagefehler lassen sich reduzieren durch
  - Weitere Prädiktoren (die Informationen liefern)
  - Besseres Modell (z.B. nicht-linear, Interaktionen)

  → Aber: Gefahr des Overfittings

Model Performance

UNIVERSITÄT TRIER

- Wie gut ist das Modell als Entscheidungsgrundlage (wie ausgeprägt sind prediction errors)?

- **Die wichtigsten:**
  - **R-Squared (R^2)**: Anteil der Varianz der abhängigen Variablen, der durch das Modell erklärt wird
  - **Mean Squared Error (MSE):** Mittelwert der quadrierten Residuen
  - **Root Mean Squared Error (RMSE):** Wurzel des MSE. Hebt die Quadrierung beim MSE auf → Interpretation in der originalen Metrik
  - **Mean Absolute Error (MAE):** Durchschnittliche der **absoluten** Residuen → Robuster gegenüber Ausreißern (Outliern).
  - **Mean Absolute Percentage Error (MAPE)**: Durchschnittlicher prozentualer Betrag der Residuen

$$MAPE = \frac{1}{N} \times \sum \frac{|e_i|}{y_i}$$

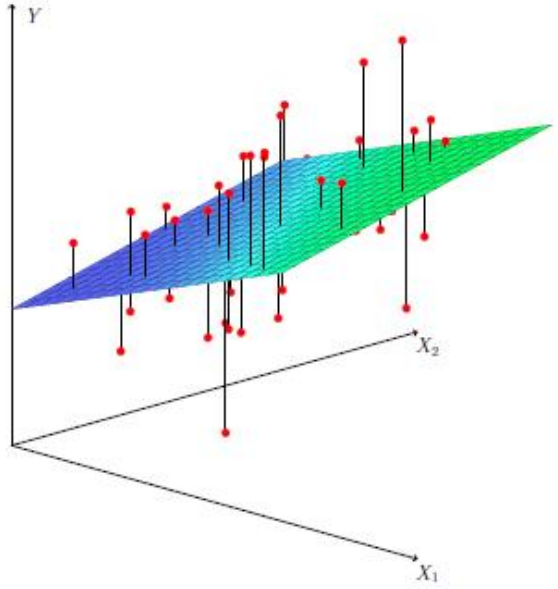
Multiple Regression

Multiple Regression: Berücksichtigung von mehr als einem Prädiktor:

$$E(Y \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p$$

- Anstatt eine **Linie** (Steigung) in ein Streudiagramm einzusetzen
  → **Regressionsfläche (regression surface)** in einem p-dimensionalen "feature space".

Beispiel mit 2 Prädiktoren:



Ziele:

- Y mit hoher Genauigkeit **vorhersagen** (alle Informationen verwenden)
- Interesse an der **spezifischen Rolle** verschiedener Prädiktoren
- (Bei kausaler Perspektive: Verringerung des **Confounding Bias**)

Quelle: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2011). An introduction to statistical learning - with applications in R (Vol. 112). New York: Springer. https://www.statlearning.com/

UNIVERSITÄT TRIER

Supervised machine learning: Klassifikation

Klassifikation

UNIVERSITÄT TRIER

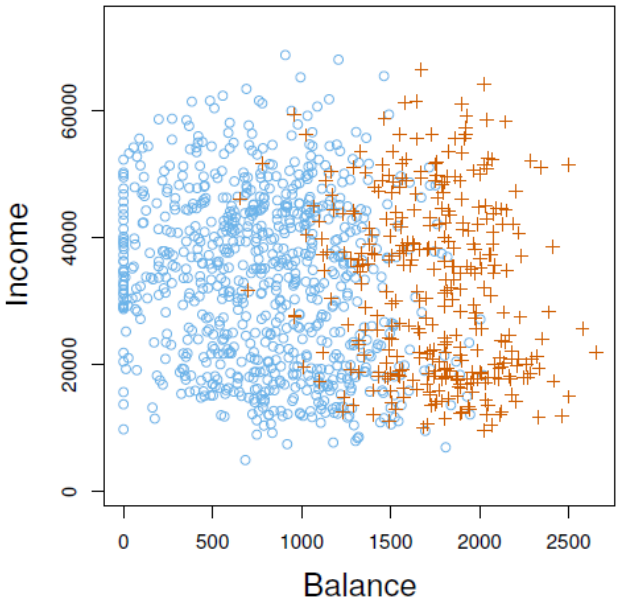
- Häufigster Fall
- Kategoriales Target, z.B.
  - **Customer churn:** Ja (Abo gekündigt) oder nein (noch Kunde)
  - **Customer value:** Wichtig vs. "unwichtig"
  - **Fraud detection (Kreditkartenmissbrauch):** Ja vs. nein
- **Anzahl der Kategorien**
  - **Binär** (0/1)
  - **Multi-class** (z.B. Wahl unter 5 Pricing-Modellen)

→ Gesucht ist ein "Classifier", der auf Basis von Prädiktoren (Features) einen Fall klassifiziert

→ Basis ist die geschätzte **Klassenwahrscheinlichkeit** (Optimal: hohes p für eine Klasse, niedriges für alle anderen

Klassifikation: Beispiel

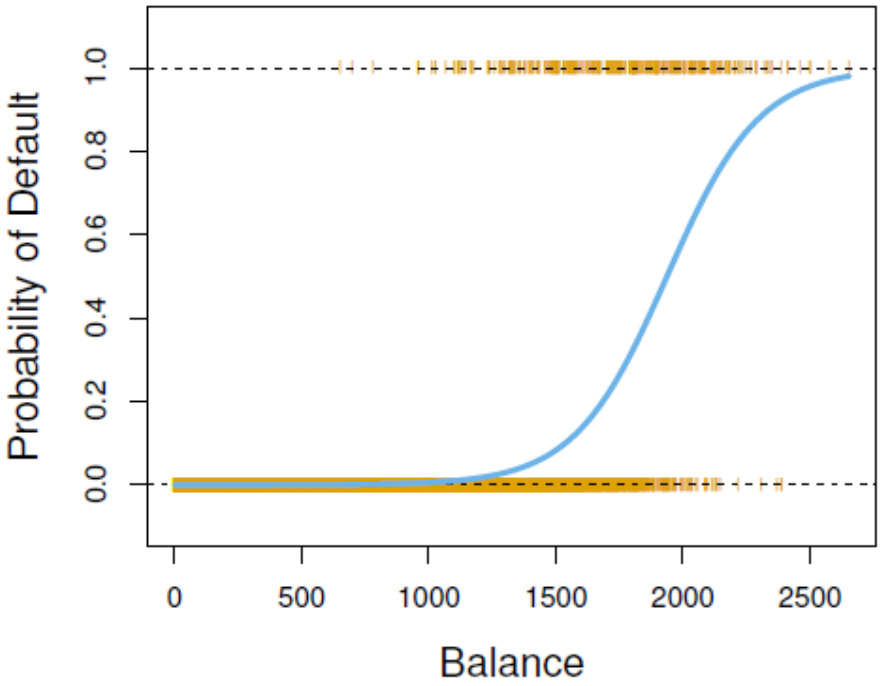
- Beispiel aus James et al. (Kapitel 4)



- Typische Darstellung im Bereich ML
  - X und Y –Achse sind Prädiktoren
  - Das Target ist stattdessen als Symbol oder Farbverlauf gekennzeichnet
- **Target:** Default (Überziehen der Kreditkarte):
  - ja (rot)
  - nein (blau)
- **Features:**
  - Einkommen
  - Monatliches Kreditkartensaldo ("Balance")

Klassifikation: Beispiel

- Standard Workhorse: Logistische Regression



- Modelliert die Wahrscheinlichkeit für Y über eine logistische Funktion

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- Eine Umformung führt zu einer linearen Funktion, deren predicted values die **logarithmierten odds / logits** sind

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- Auf Basis des/der jeweiligen X-Variablen bekommt man p
- Klassifikation in Y=1, wenn der p-Wert einen **manuell gewählten Threshold** überschreitet, z.B. p > .5)

Klassifikation: Performance

UNIVERSITÄT TRIER

- Zentrale Frage: Wie akkurat ist das Modell in der positiven und negativen Klassifikation ?
- **Confusion matrix**

	Klassifikation	
Wahrheit	Y = 0	Y = 1
Y = 0	True Negatives	False Positives
Y = 1	False negatives	True positives

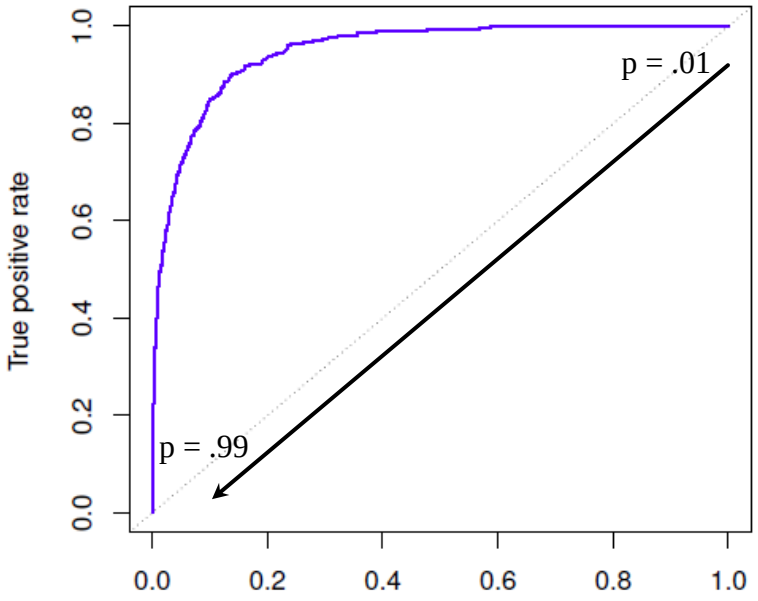
	Klassifikation	
Wahrheit	"Gesund"	"Krebs"
Gesund	True Negatives	False Positives
Krebs	False negatives	True positives

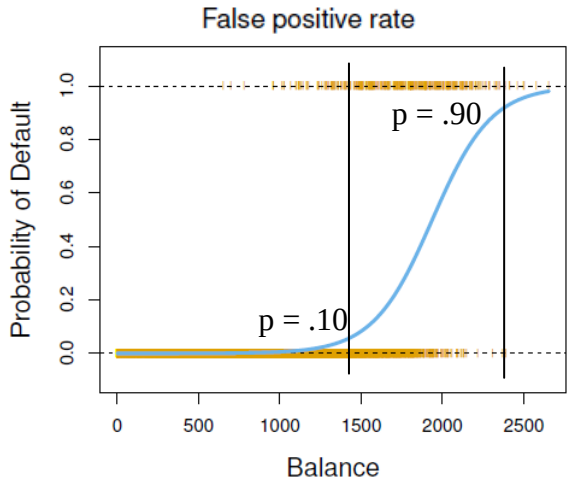
- Intuitiv: Akkuratheit (Accuracy) als primäre Metrik: Anzahl der korrekten Klassifikationen an allen.
- Wichtige Konzepte:
  - **Sensitivity:** Wie häufig erkenne ich Y=1 (=krank), wenn die Person krank ist (untere Zeile)?
  - **Spezifität:** Wie häufig kann ich die Krankheit ausschließen, wenn die Person gesund ist (obere Zeile)?

→ Beide stehen in einem Spannungsverhältnis (eine zu maximieren ist leicht)

Klassifikation: Performance

- Die ROC-Kurve zeigt das Spannungsverhältnis:
  Wie wirkt sich ein laxerer vs. strengerer Threshold auf die True positives v.s false positives aus?



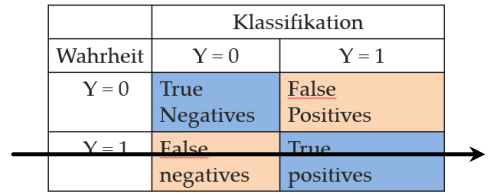


- Graue Diagonale: Raten
- Alle Modelle im oberen Drittel sind besser als Raten.
- "**Area under the curve" (AUC)** ist ein wichtiges performance-Maß
- Die Kurve zeigt, wie ein bestimmter Klassifikations-Erfolg (TP) mit dem Anteil der Fehlalarme (FP) einhergeht.
- Durch die **Wahl des Thresholds** bestimmt man das gewünschte Verhältnis (will man eher eine Krankheit übersehen oder gesunde Patienten beunruhigen?)

Klassifikation: Performance

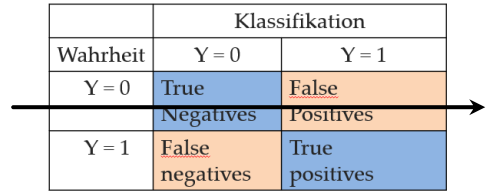
- Auf Basis der Confusion matrix lassen sich andere wichtige Metriken berechnen
  - **Sensitivity / Recall:** Korrektes Erkennen der positiven Fälle (TP + FN)

$$Sensititivity = \frac{TP}{TP + FN}$$



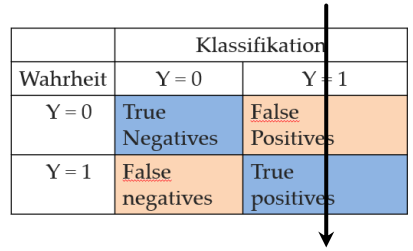
  - **Spezifität:** Korrektes Erkennen der negativen Fälle (TN + FP)

$$Specificity = \frac{TN}{TN + FP}$$



  - **Precision:** Anteil der TP an den als positiv klassifizierten Fällen (korrekte und falsche)

$$Precision = \frac{TP}{TP + FP}$$



  - **F1 score:** Der F1-Score betrifft das optimale Verhältnis beider und ist im Idealfall 1.0

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$