Wirtschaftsinformatik – Informationsmanagement und Data Science

HOCHSCHULE HEILBRONN

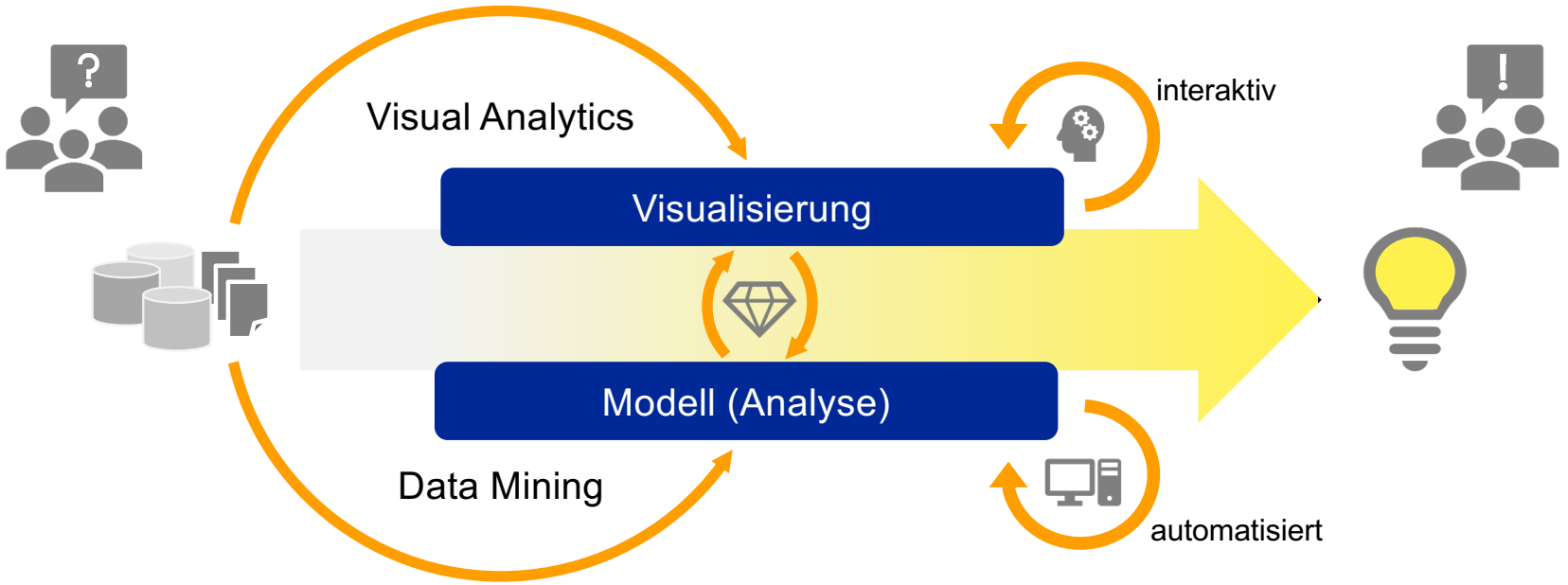
› Data Science Grundlagen
Modeling: Aus Daten lernen

Prof. Dr. Carsten Lanquillon | Fakultät Wirtschaft und Verkehr | Wirtschaftsinformatik

Erkenntnisse aus Daten gewinnen

Von der Fragestellung

zur Erkenntnis

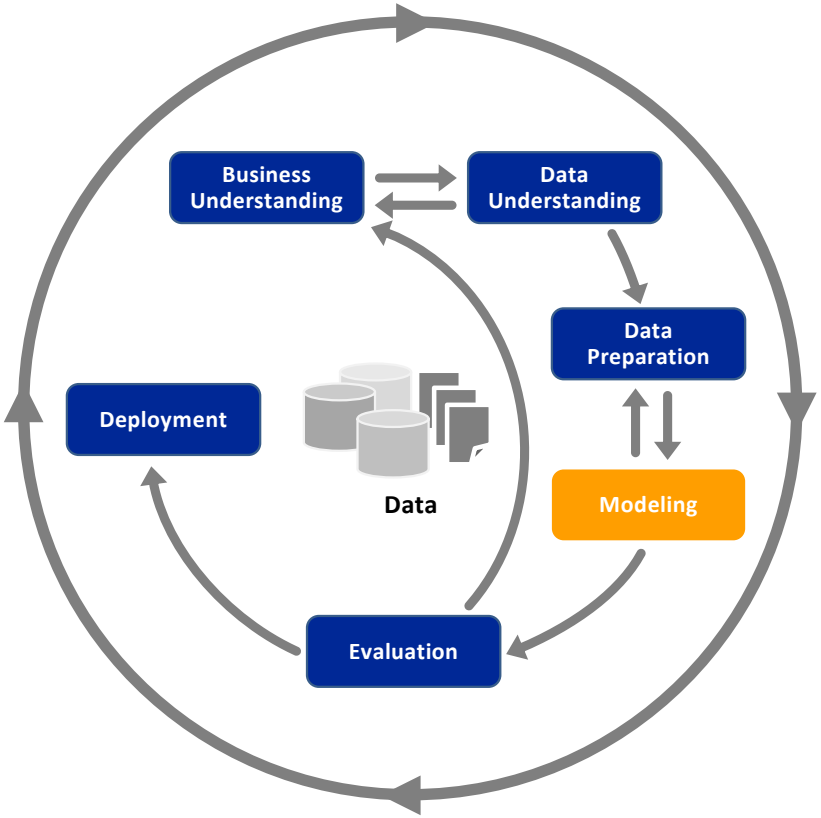


CRISP-DM Phase 4: Modeling

Ziel: Modellerstellung durch maschinelle Lernverfahren

Aufgaben

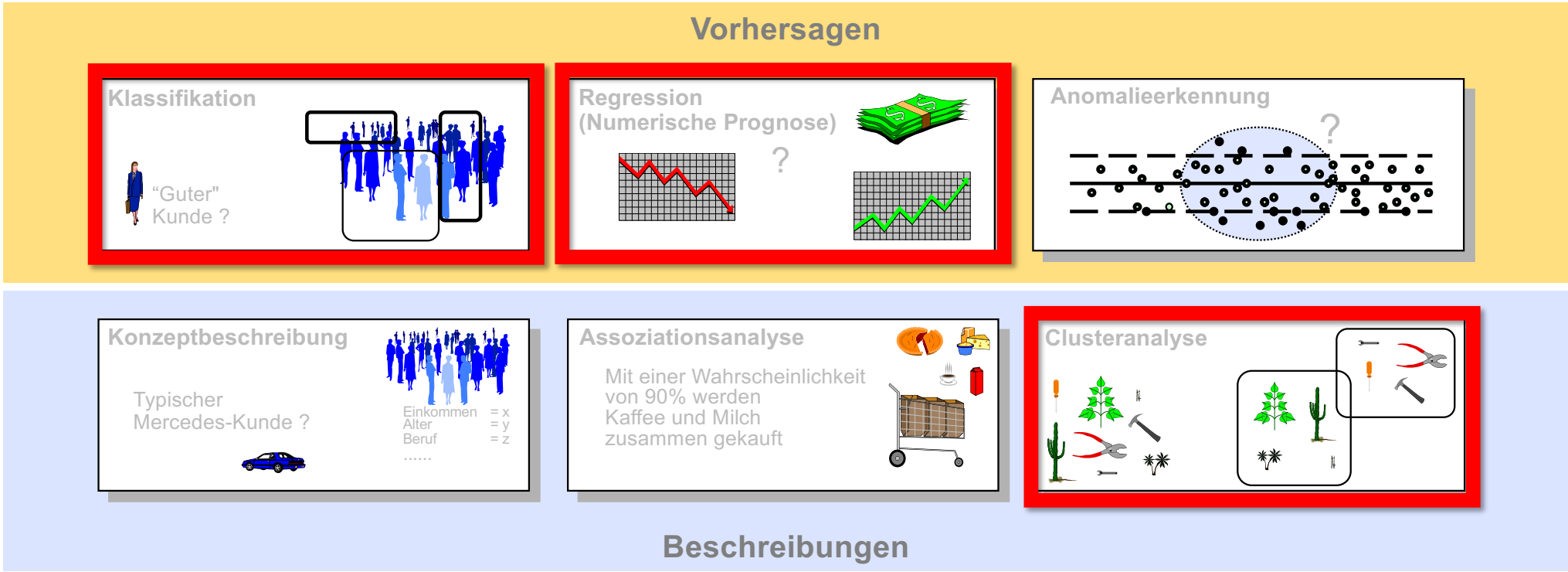
- Auswahl geeigneter Lernverfahren (Model Selection)
- Nicht auf einen Lösungsansatz (Lernverfahren) festlegen („*no free lunch*“)
- Festlegen von Modellparametern
- Aufbau verschiedener Modelle (Lernen aus Daten)
- Auswahl eines Modells für die Aufgabenstellung



Modeling – Aus Daten lernen

- Wiederholung der Data-Mining-Aufgaben
- Muster und Modelle: Was können wir aus Daten lernen?
- Was ist Lernen? – Eine Machine-Learning-Perspektive
- Überwachtes vs. unüberwachtes Lernen
- Gütemaße für Prognoseaufgaben
- Induktiver Bias
- Overfitting
- Modellkomplexität und der Bias-Varianz-Trade-off
- Abschätzung der zukünftigen Güte eines Modells

Klassische Data-Mining-Aufgaben

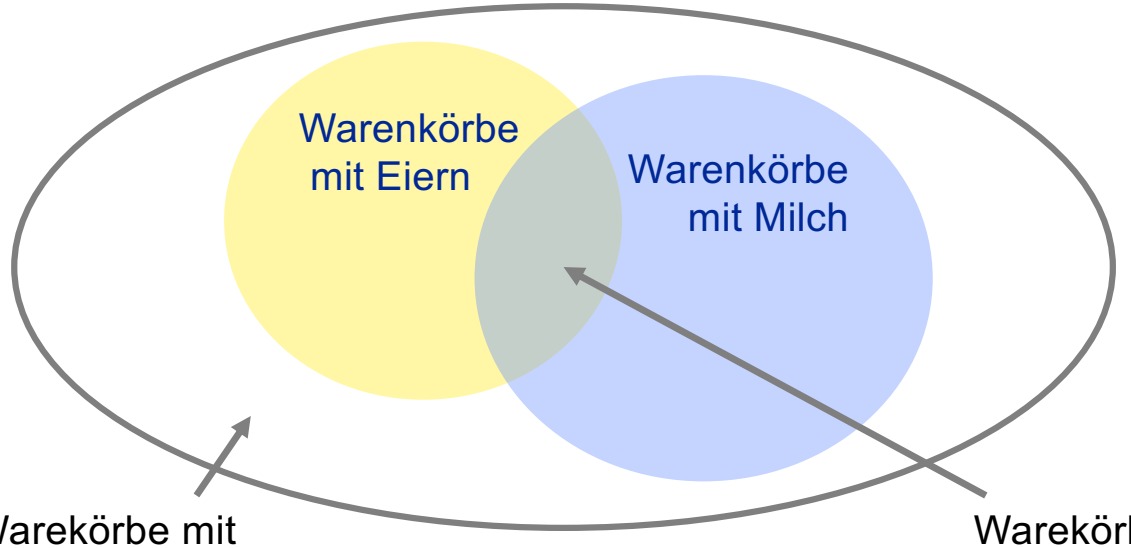


Beispiele für Data-Mining-Aufgaben

> Anomalieerkennung	Ist eine Kreditkartentransaktion legitim oder Betrug?
> Assoziationsanalyse	Welche Waren werden häufig zusammen gekauft?
> Clusteranalyse	Gibt es Kundengruppen mit ähnlichem Verhalten?
> Klassifikation	Ist ein neuer Kunde kreditwürdig oder nicht?
> Konzeptbeschreibung	Was kennzeichnet einen Kunden der kündigt?
> Regression	Wie viel Umsatz machen meine Kunden?

Beispiel: Assoziationsregeln

- Wir betrachten Scanner-Daten aus einem Supermarkt
- Angekommen Milch und Eier sind für uns interessant



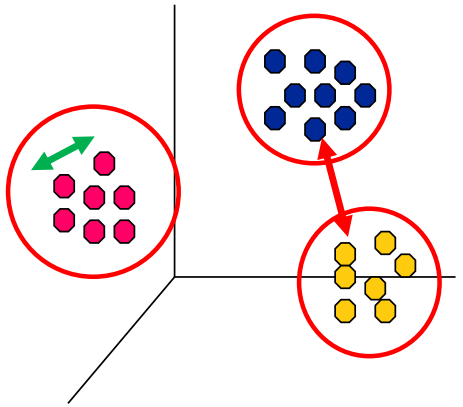
Regel: Milch → Eier
Confidence: 50%
Support: 20%

Beispiel: Clusteranalyse

Erzeuge Gruppen (Cluster, Klassen, Segmente)
von Objekten mit folgenden Eigenschaften:

- **“Within-Cluster Homogeneity”**
  Objekte in einer Gruppe sind ähnlich zueinander

- **“Between-Cluster Heterogeneity”**
  Objects in verschiedenen Gruppen sind unähnlich



Wan sind Objekte ähnlich?

→ **Benötigt Ähnlichkeits- oder Distanzmaße!**

Beispiel: Konzeptbeschreibung

- **Konzeptlernen (concept learning)**
  Eine boolesche Funktion aus Trainingsdaten ableiten

- **Konzeptbeschreibung (concept description)**
  Eine verständliche Beschreibung (typische Kombinationen von Merkmalsausprägungen) für das Konzept (die **positiven** Fälle) aus Trainingsdaten ableiten

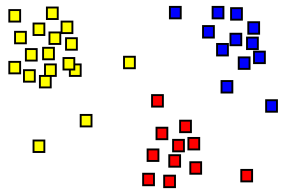
Beispiele:

- Was zeichnet einen guten Kunden aus?
- Was sind typische Eigenschaften einer Kunden der kündigt?
- Was kennzeichnet Fahrzeuge die ein bestimmtes Qualitätsproblem haben?

Klassifikation vs. Clustering

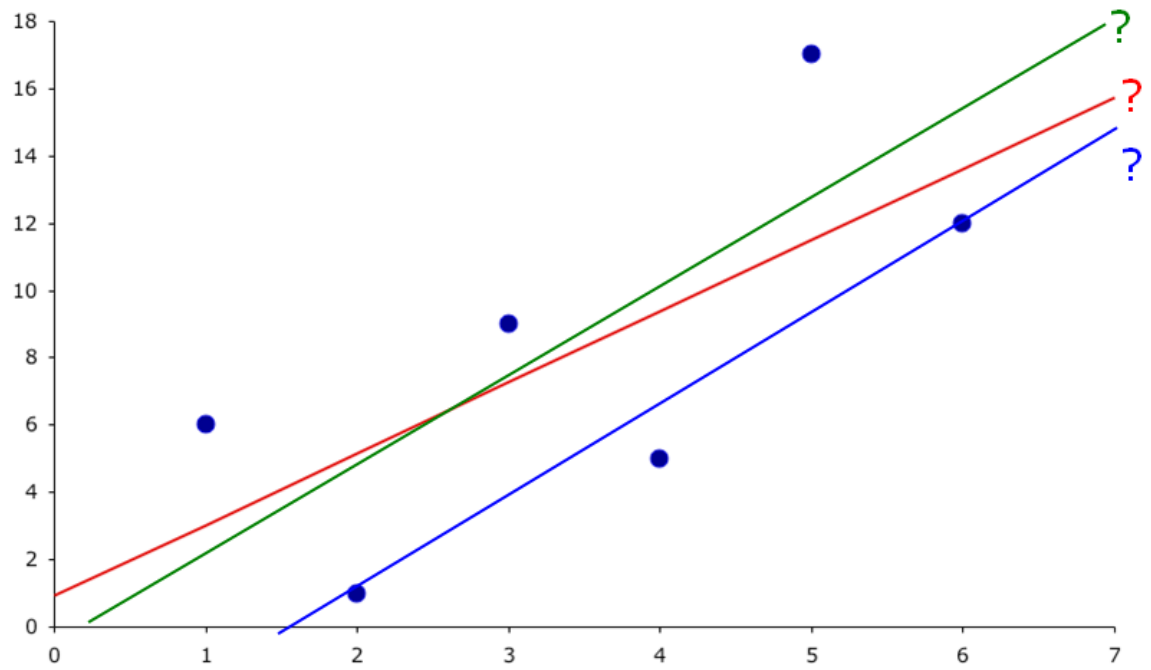
- **Klassifikation** → Zielgröße (Klassenmerkmal) ist bekannt
  - Klassen und Zuordnung von Trainingsdaten zu Klassen (Label) sind gegeben
  - **Ziel**: Lernen einer Abbildung, die diese Zuordnung für neue Objekte durchführen kann

- **Clustering** → keine Zielgröße (Klassenmerkmal) gegeben
  - **Ziel**: Struktur in den Daten entdecken



Beispiele: Einfache lineare Regression

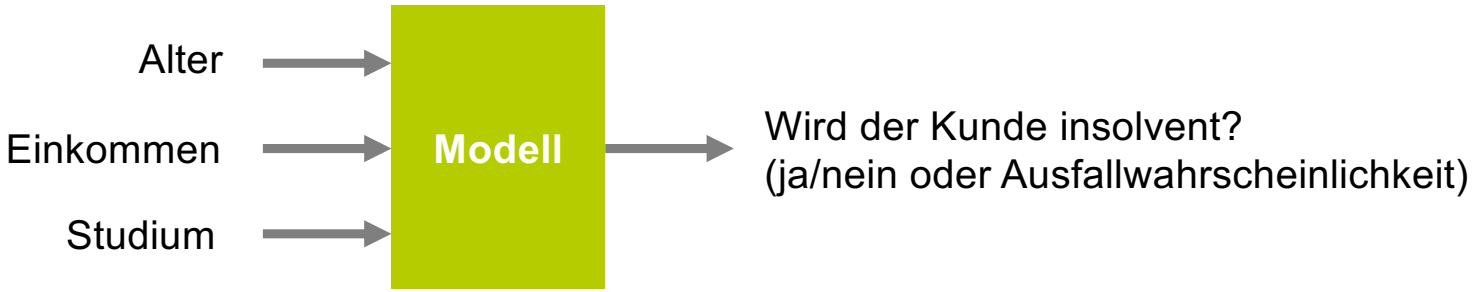
Welche Gerade passt am besten?



Predictive Modeling: Klassifikation und Regression

Mit von beobachten Merkmalen soll der Wert einer Zielgröße vorhergesagt werden

- **Klassifikation**: Zielgröße ist qualitativ
- **Regression**: Zielgröße ist quantitativ
- **Ranking**: Zielgröße ist ordinal skaliert



Default-Klassifikationsregel

Die Default-Klassifikationsregel ist nach der zufälligen Klassifikation die einfachste Zuordnungsregel, die eingesetzt werden kann, wenn über ein Objekt nichts bekannt ist:

- Bestimme die häufigste Klasse in den Trainingsdaten
- Die Wahrscheinlichkeit dieser Klasse sei p_{max}
- Die erwartete Fehlerrate der Default-Regel: $1 - p_{max}$
- Diese Feherrate ist ein erster Benchmark für komplexere Klassifikatoren
- **Ein aus Daten erlernte Klassifikator soll besser sein als die Default-Regel!**

Was ist Lernen?

Wörterbücher definieren “lernen” als

- Wissen oder Kenntnisse aneignen durch Studieren, Erfahrung oder Unterricht
- sich durch Information oder Beobachtung einer Sache bewusst werden
- sich im Gedächtnis einprägen
- informiert werden oder erkunden
- Anweisungen erhalten

→ Die Definitionen sind nicht wirklich adäquat, wenn es um Computer (Maschinen) geht!

Arbeitsdefinition:

> „Dinge lernen, wenn sie ihr Verhalten derart anpassen, dass sie zukünftig besser sind.“
>
> (Witten et al. (2011), p. 7)

Machine-Learning-Definitionen

Samuel (1959):

> *Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed.*

Mitchell (1997):

> *Machine learning is the study of computer algorithms that improve automatically through experience.*
>
> *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

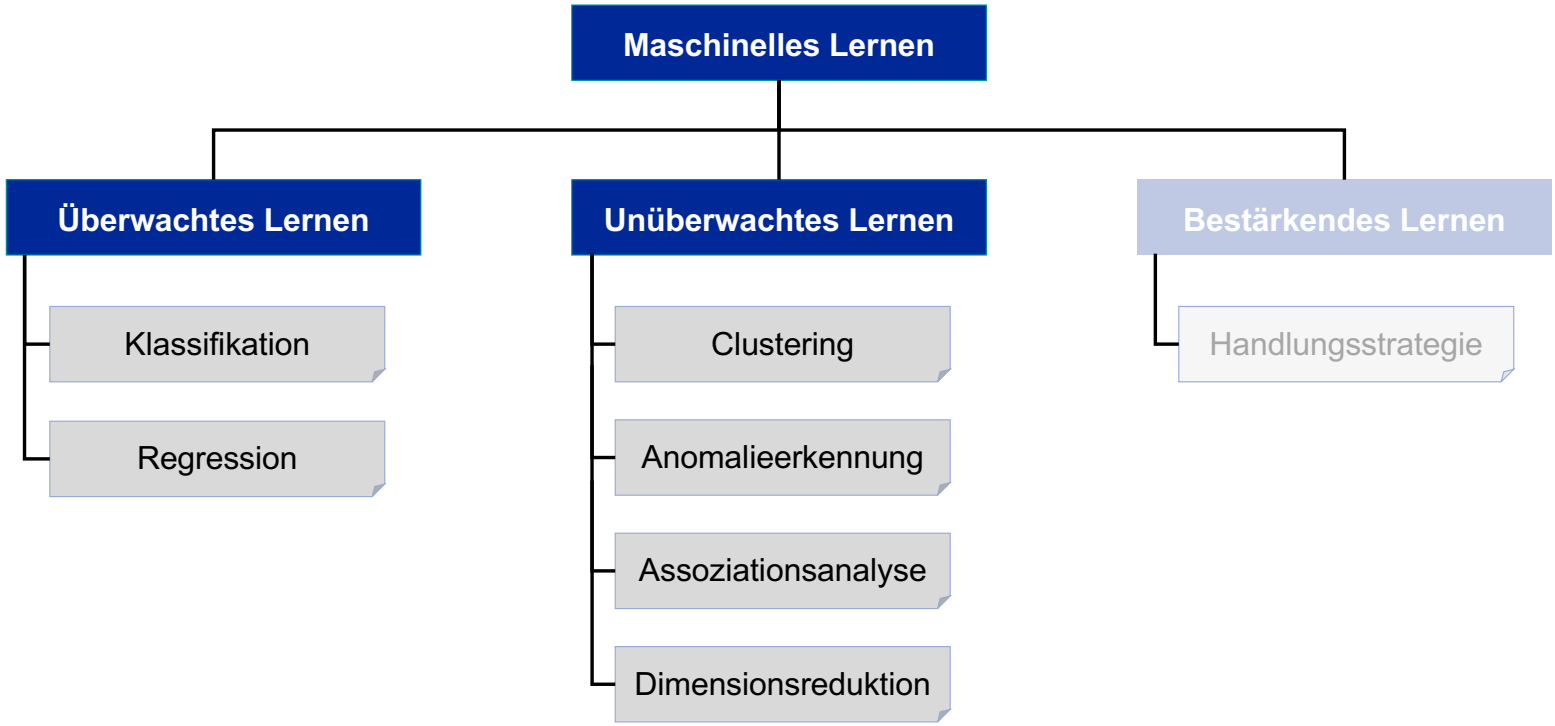
Lernformen beim Machine Learning

- Überwachtes Lernen (Supervised Learning)
  - **Generalisieren basierend auf Beispielen**

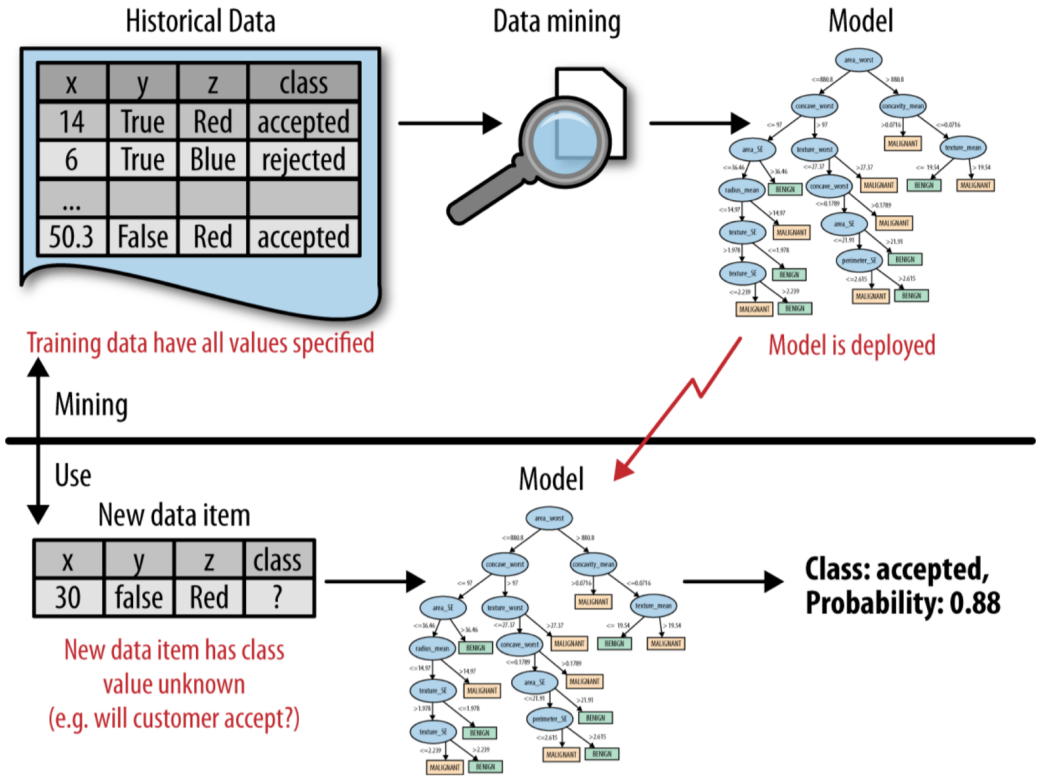
- Unüberwachtes Lernen (Unsupervised Learning)
  - **Strukturentdeckend ohne Vorgaben/Feedback**

- Bestärkendes Lernen (Reinforcement Learning)
  - **Lernen durch Versuch und Irrtum**

Lernformen und Data-Mining-Aufgaben



Modellerstellung vs. Modellanwendung



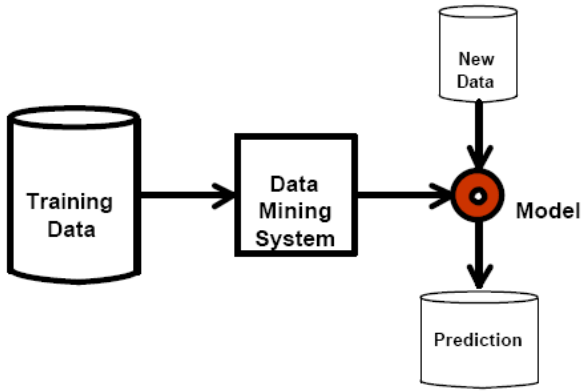
Quelle: Provost & Fawcett (2013): Data Science for Business)

Überwachtes Lernen von Vorhersagemodellen

Basierend auf einer Trainingsmenge wollen wir

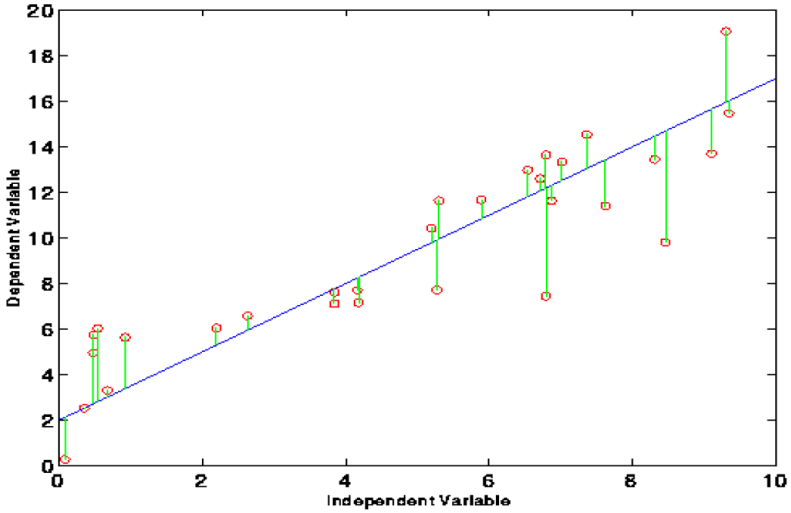
- die unbekannte Zielgröße Y für ungesehene Objekte basierend auf beobachteten Merkmalswerten X möglichst genau vorhersagen
- bei einem Klassifikationsproblem (auch) die Wahrscheinlichkeit (Konfidenz) zum Vorhersagewert ausgeben
- verstehen welche Eingaben die Ausgaben wie beeinflussen
- die Vorhersagequalität beurteilen

past-expenses	age	bonus	gender	accept
low	elder	high	female	no
low	elder	high	male	no
average	elder	high	female	yes
high	average	high	female	yes
high	young	normal	female	yes
high	young	normal	male	no
average	young	normal	male	yes
low	average	high	female	no
low	young	normal	female	yes



Regression: Messung der Anpassungsgüte

Sehr häufig wird der empirische Fehler als Summe der quadrierten Abweichungen der vorhergesagten Werte von den tatsächlichen (beobachteten) Werten verwendet (sum of squared errors, SSE)



$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

tatsächlicher Wert vorhergesagter Wert

Klassifikation: Messung der Klassifikationsleistung

> Die Konfusionsmatrix fasst die Klassifikationsentscheidungen zusammen

> Für binäres Klassifikationsproblem mit den Klassen P und N ergibt sich:

	klassifiziert als P	klassifiziert als N
wahre Klasse P	a	b
wahre Klasse N	c	d

> Gütemaße, die beide Klasse gleichmäßig berücksichtigen:

– (Empirische) Fehlerrate err = (b + c) / (a + b + c +d)

– Erkennungsrate (accuracy) acc = (a + d) / (a + b + c +d)

→ Zusammenhang err = 1 - acc

Spam oder Ham? – Alternative Gütemaße

- In einigen Anwendungen ist die Fehlerrate nicht sinnvoll
  - Bei sehr schiefer Klassenverteilung und/oder unterschiedlichem Interesse an den Klassen
  - E.g. Spam-Filter, Betrugserkennung, medizinische Behandlung

- Alternative Gütemaße fokussieren oft nur auf eine Klasse

True positive rate (recall, sensitivity)	tp = a / (a + b) = r
False negative rate (type II error)	fn = b / (a + b)
True negative rate (specificity)	tn = d / (c + d)
False positive rate (type I error)	fp = c / (c + d)
Precision	p = a / (a + c)
F-Measure	f = 2·p·r / (p + r)

Klasse P

Klasse N

Die Inductive Learning Hypothesis

Aus Daten Lernen = Anpassung eines Modells

Jedes Modell, das die Trainingsdaten hinreichend genau annähert, wird auch für neue Daten die Zielgröße hinreichend genau vorhersagen.

Ziel: Generalisieren anstatt auswendig lernen

Die Lerntheorie beantwortet Fragen wie

- Wann ist die Trainingsmenge groß genug?
- Was ist die bestmögliche Performanz?

Das Problem der Induktion und der induktive Bias

Das Generalisieren über die bekannten Trainingsdaten hinaus ist nie logisch gerechtfertigt, da es stets mehrere Hypothesen (Modelle) gibt, die die Trainingsdaten erklären können.

Wie kann „das richtige“ Modell ausgewählt werden?

Wir betrachten das Lernen als Suche im Hypothesenraum (Menge aller möglichen Modelle)

Der induktive Bias begrenzt und steuert die Suche

- Representation (Language) Bias
- Search (Preference) Bias
- Overfitting Avoidance Bias

→ **Kein Generalisieren ohne induktiven Bias!**

Representation Bias

- Der **Representation Bias** oder **Language Bias** definiert den Hypothesenraum, d.h. die Menge der Hypothesen, d.h. die Menge der Hypothesen die in Frage kommen
- Durch den Lernalgorithmus wird meistens eine Modellart vorgegeben, z.B. Entscheidungsbäume oder Regeln
- Fachliches und analytisches Hintergrundwissen unterstützt die Auswahl einer geeigneten Modellklasse

Search Bias

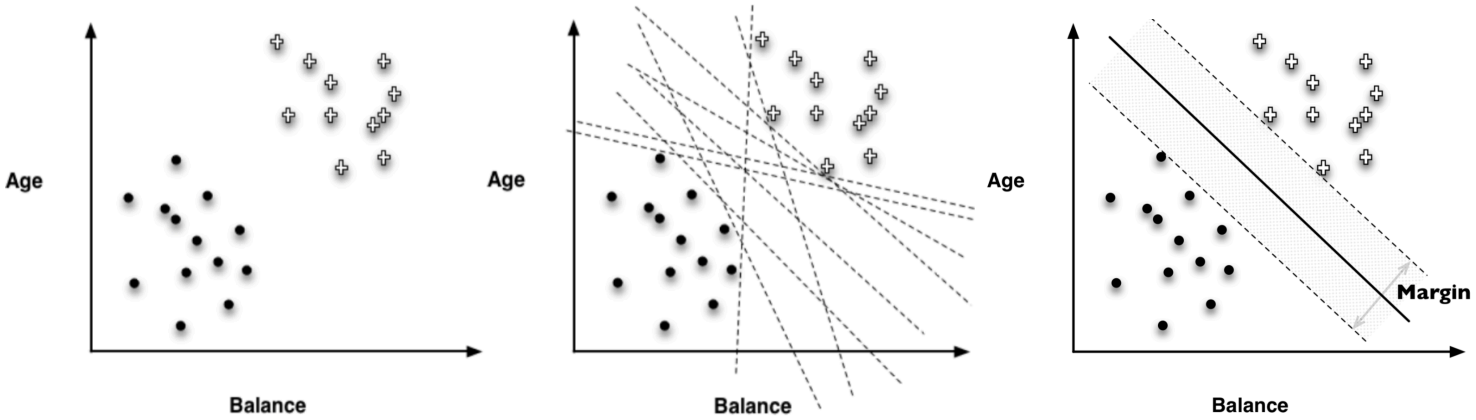
Der **Search Bias** oder **Preference Bias** definiert eine Präferenz für bestimmte Hypothesen (Auswahlkriterium, Evaluierungskriterium, Zielfunktion) oder legt fest, wie Hypothesen generiert werden (z.B. greedy search)

Beispiele

- **Occam‘s razor**
  Wähle das einfachste Modell, das die Daten hinreichend gut erklärt

- **Minimum description length (MDL) Prinzip**
  Minimiere die Summe der Länge (Bits) aus Codierung des Modells und Codierung von Ausnahmen die das Modell nicht erklären kann (Fehler)

Beispiel: Language und Search Bias



Binäre Klassifikation

- Language Bias: Beschränkung auf lineare Modelle
- Search Bias: Bevorzuge Gerade, die den Rand (margin) maximieren

Overfitting und Fähigkeit zum Generalisieren

„If you torture the data long enough, it will confess.“ (Ronald Coase)

- Wenn man lange und intensiv genug sucht, wird man Zusammenhänge in Daten finden – aber diese lassen sich nicht unbedingt auf neue Daten übertragen.
- Overfitting (Überanpassung) ist die Neigung von Lernverfahren Modelle zu stark an die Trainingsdaten anzupassen (z.B. Auswendiglernen).
- Dies geht auf Kosten der Fähigkeit zu Generalisieren, d.h. Aussagen für bislang nicht gesehene Daten zu treffen.

→ **Konsequenz: Geringe Leistungsgüte bei neuen Daten!**

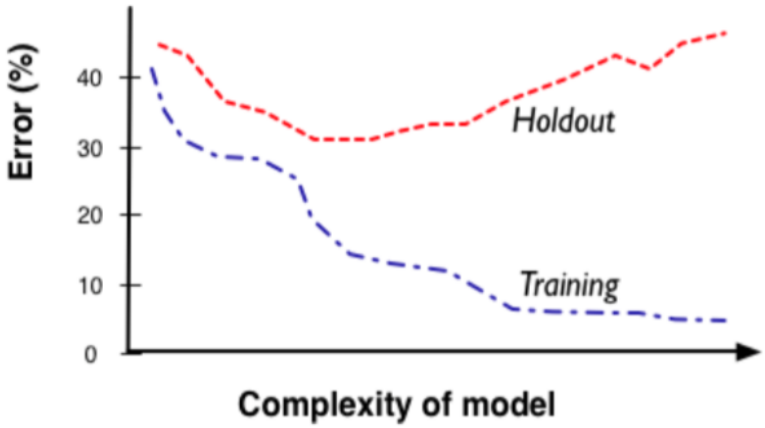
- Zu einem gewissen Grad neigen alle Lernverfahren zum Overfitting!

Wie erkennt man Overfitting (1)

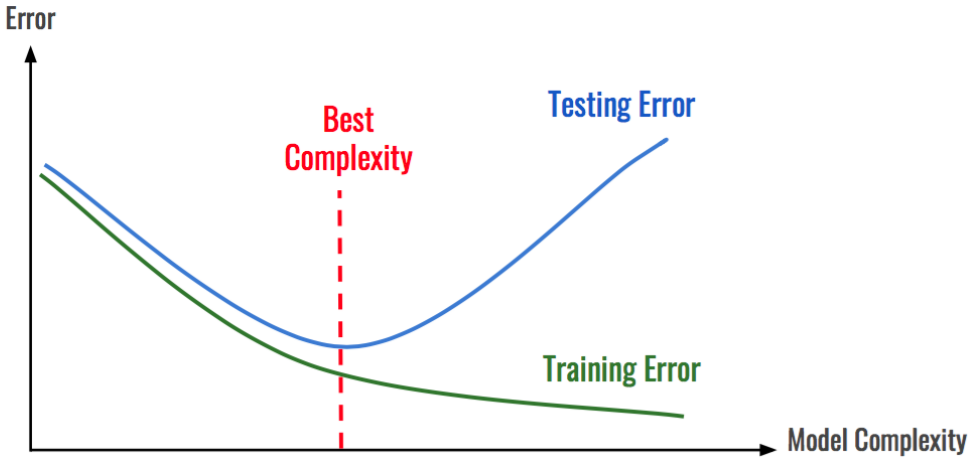
- Die Art und Weise Overfitting zu verhindern hängt stark vom Lernverfahren ab
- Eine gute Strategie ist Overfitting zu erkennen und die Modellkomplexität als Hauptursache dafür zu kontrollieren → **Overfitting Avoidance Bias**
- Kernidee für die Erkennung:

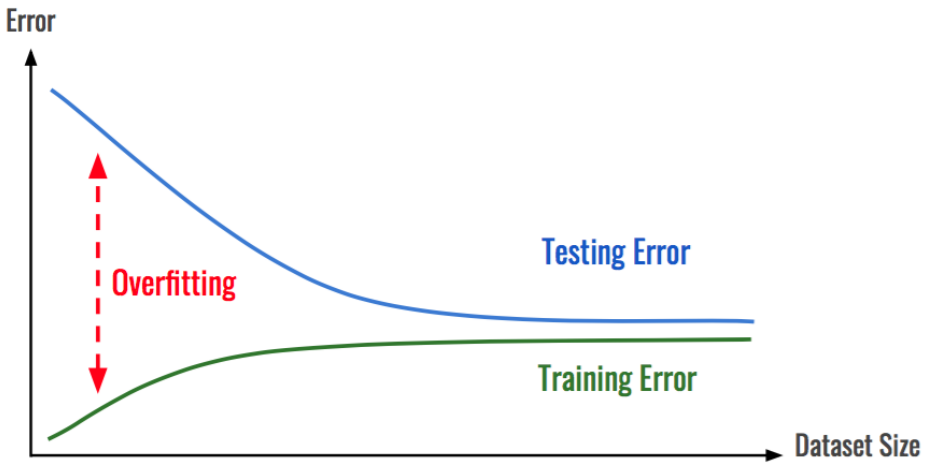
  Lerne und validiere auf unterschiedlichen Daten!

  - **Trainingsdaten** zum Lernen (in-sample)
  - **Hold-out-Daten** zum Validieren (out-of-sample)



Wie erkennt man Overfitting (2)





Quelle: https://hackernoon.com/memorizing-is-not-learning-6-tricks-to-prevent-overfitting-in-machine-learning-820b091dc42

Overfitting Avoidance Bias

Ziel: Modellkomplexität kontrollieren, um Overfitting zu vermeiden

- Regularisierung
  Bestrafung von Modellkomplexität im Rahmen der Modellbewertung
  (z.B. Strafterm bei der Zielfunktion)

- Forward Pruning (Pre-Pruning)
  Modellkomplexität ab einem vorgegebenen Grad unterbinden
  (Modellverfeinerung stoppen)

- Backward Pruning (Post-Pruning)
  Modelle nach dem Lernen wieder vereinfachen solange die Güte nicht signifikant sinkt

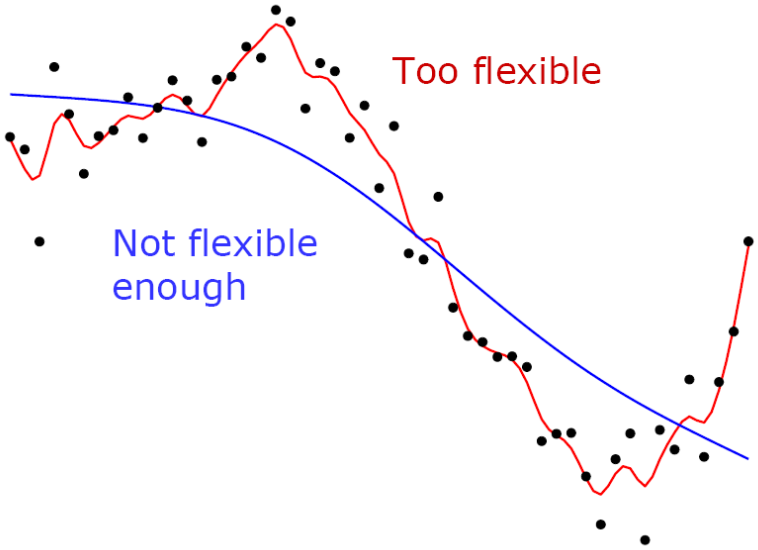
Das Bias-Varianz-Dilemma

Welche Modellkomplexität ist erforderlich?

Modelle mit zu wenig Komplexität (zu wenig Parameter) haben einen zu großen Bias – sie sind nicht flexibel genug.

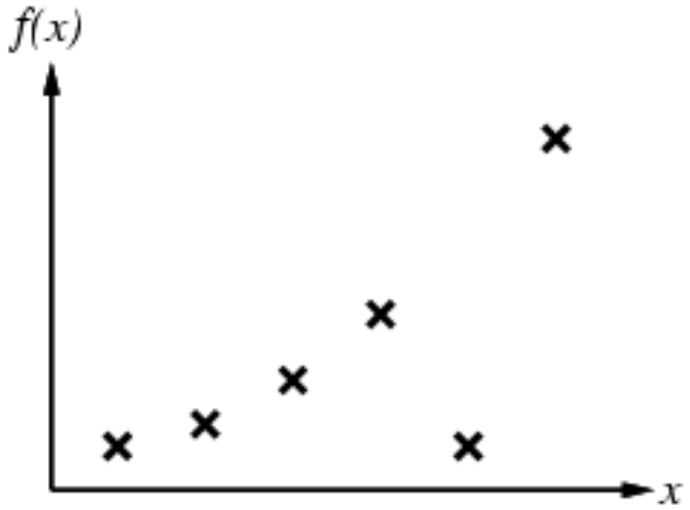
Modell mit zu viel Komplexität (zu viele Parameter) haben eine sehr große Varianz – sie sind zu flexibel.

→ Bestimmung des besten Modells erfordert Festlegung der richtigen Modellkomplexität (Anzahl Parameter)



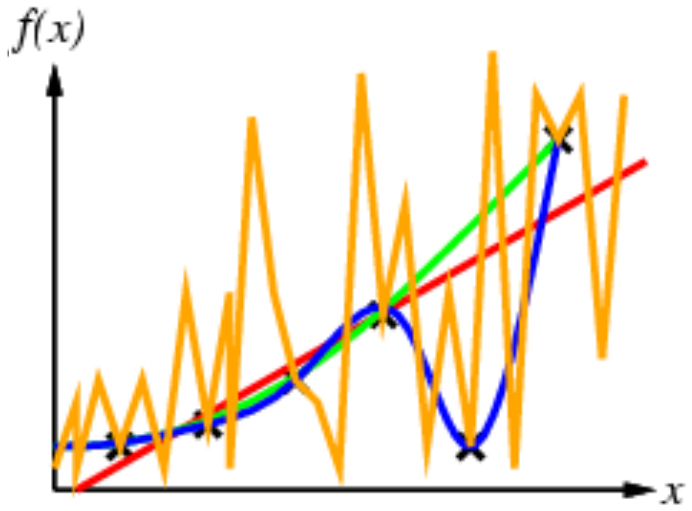
Beispiel: Curve-Fitting (1)

Welches Polynom passt am besten?



Beispiel: Curve-Fitting (2)

Welches Polynom passt am besten?



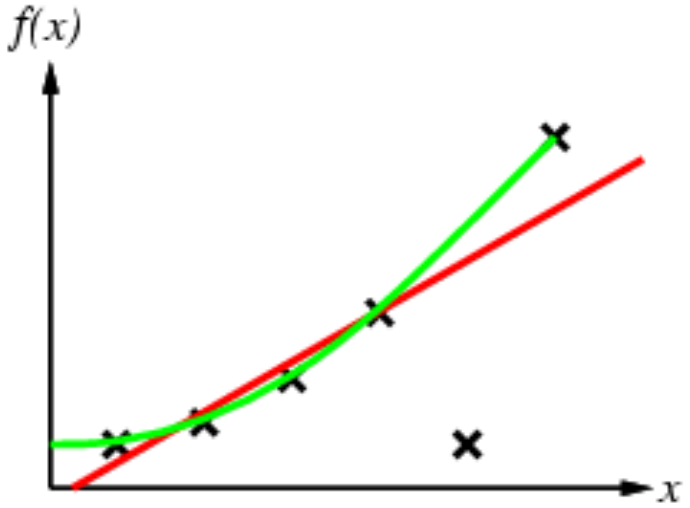
Beispiel: Curve-Fitting (3)

Welches Polynom passt am besten?

→ Ockham's Razor:

Wähle das einfachste Modell, das die Daten hinreichend gut beschreibt!

Hier: Die Parabel



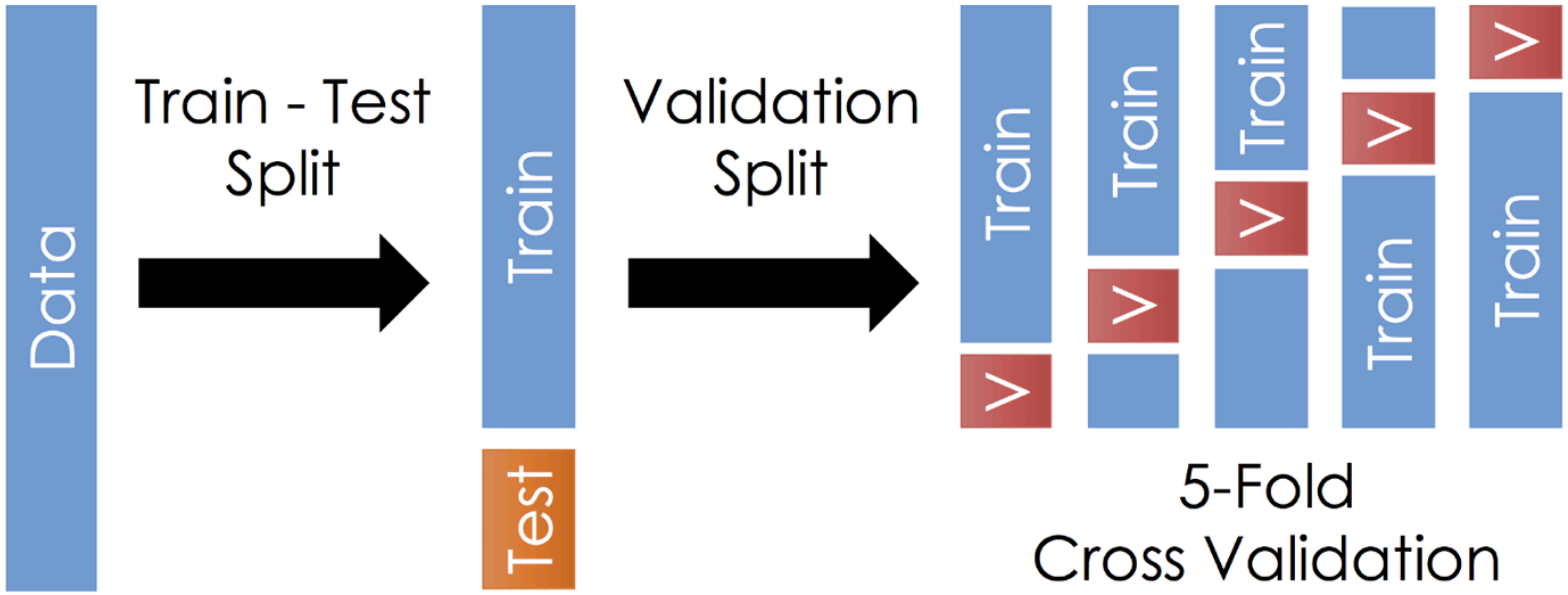
Schätzung der zukünftigen Leistungsgüte

- Überwachte Lernverfahren optimieren das Modell hinsichtlich der Leistungsgüte auf den Trainingsdaten
- Bei hinreichender Modellkomplexität und wenig Rauschen in den Daten ist eine perfekte Güte realistisch

  → Als Schätzung der zukünftigen Leistungsfähigkeit ungeeignet (zu optimistisch)

- Relevant ist die Schätzung der Leistungsfähigkeit auf neuen Daten
  - Trainingsdaten zum Lernen
  - Validierungsdaten zum Auswählen → ggf. Kreuzvalidierung
  - Testdaten für abschließende Bewertung → bereits am Anfang zur Seite legen!

Datenfluss: Aufteilung der Daten



K-fache Kreuzvalidierung

Ziel: Schätzung der zukünftigen Leistungsfähigkeit, ohne zu viele Daten dafür zu opfern

- Teile Daten in k disjunkte möglichst gleich große Teilmengen auf (hier: k=10)

- Lege jeweils eine Menge zum Testen (Validieren) zur Seite und Trainiere auf dem Rest

- Wiederhole Vorgang mit jeder Teilmenge (k-mal)

- Schätzung der Leistungsgüte ergibt sich als Mittelwert der Einzelwerte

Zusammenfassung

- Aus Daten lernen heißt ein Modell an Daten anpassen zum Beschreiben, Erklären oder für Vorhersagen
- Unterscheide zwischen überwachten und unüberwachten Verfahren
- Unüberwachtes Lernen eher als Zwischenschritt mit explorativem Charakter
- Klassifikation und Regression sind die am häufigsten durchgeführten Lernaufgaben
- Overfitting erkennen und vermeiden ist elementar
- Schätze Leistungsfähigkeit von Modellen stets mit ungesehen Daten ab