Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Datenmanagement & -analyse

Übung 9 – Maschinelles Lernern: Von der Erklärung zur Prädiktion

Dr. Nikolai Stein
Lehrstuhl für WI & BA
Julius-Maximilians-Universität Würzburg

Sommersemester 2021

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Recap: Lineare Regression

- Ziel: Erklärung der Varianz in der Zielvariablen durch Linearkombination unabhängiger Variablen
  - Zugehöriges Gütemaß: R^2
  - Interpretation der Ergebnisse durch Koeffizienten und deren Signifikanz

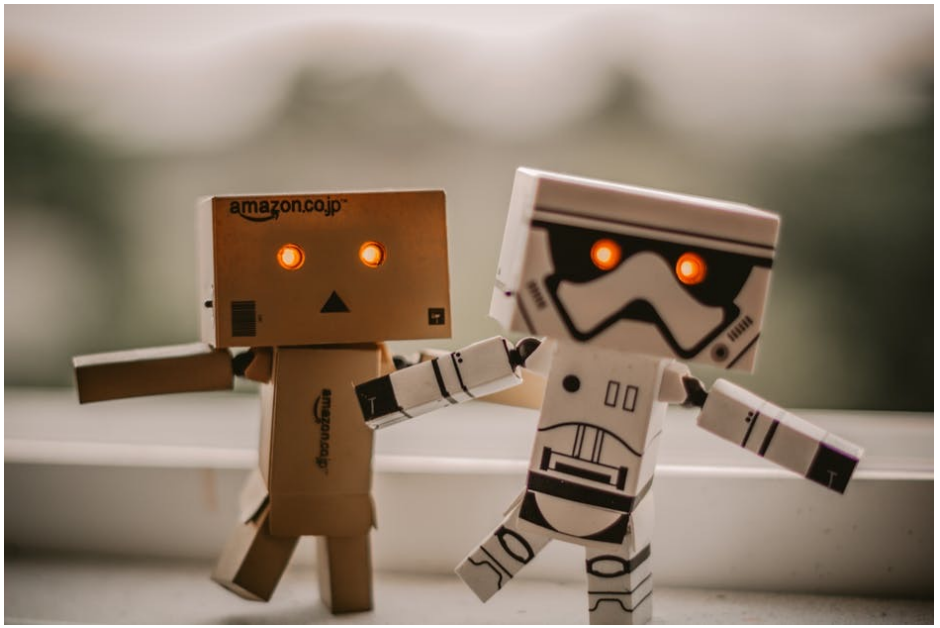
- Probleme:
  - R^2 ist monoton wachsend in der Anzahl unabhängiger Variablen
    - Robustheit des geschätzten Modells nicht gegeben
    - Generalisiert es auf unbekannte Datensätze?
  - Lineare Struktur für komplexe Zusammenhänge zu limitiert

- Möglichkeiten mit diesem Problem umzugehen?
  - Ökonometrie: andere Modellspezifikationen (robuste Standardfehler), Variablentransformation,
  - **Maschinelles Lernen / prädiktive Analyse:**
    - Lernen als zugrunde liegendes Paradigma (Verallgemeinerung)
    - Vielzahl unterschiedlicher Verfahren stehen zur Verfügung
    - Management des Bias-Varianz-Tradeoffs

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

What is Machine Learning?

- Tom Mitchell: “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.”



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

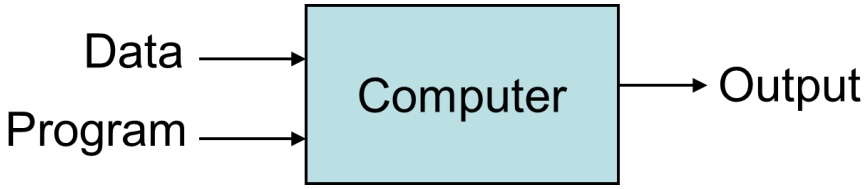
Why let machines “learn”?

- There is no need to “learn” how to calculate the payroll

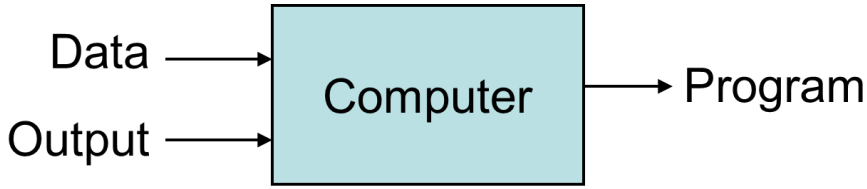
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (traffic routing)
  - Solution needs to be adapted to particular cases (user biometrics)

Ein Paradigmenwechsel

Klassische Programmierung



Maschinelles Lernen



- „Automatisierung automatisieren"
- Computer dazu bringen, sich selbst zu programmieren
- Lassen Sie stattdessen die Daten die Arbeit machen!

Verschiedene Formen des maschinellen Lernen

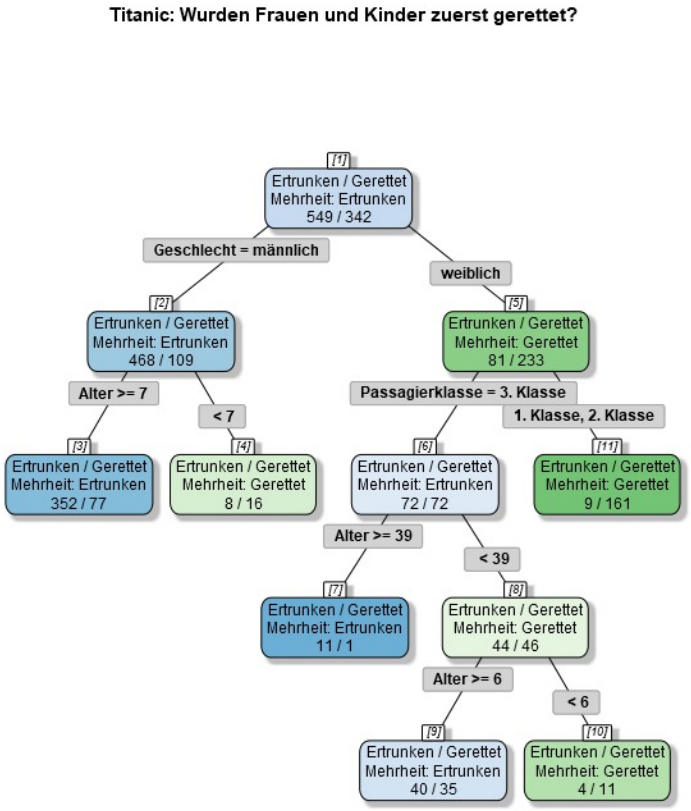
- Überwachtes Lernen
  - Trainingsdaten enthalten gewünschte Ausgaben
- Unüberwachtes Lernen
  - Trainingsdaten enthalten nicht die gewünschten Ausgaben
- Reinforcement Learning
  - Lernen aus einer Folge von Aktionen und Belohnungen

Heute

- Auf Basis von Trainingsdaten (X, Y) eine Funktion $X \mapsto Y = f(x)$ bestimmen
- Die Funktion soll zuverlässig Y-Werte für neue Datenpunkte X bestimmen
  - Diskrete Funktion $f(x)$: Klassifikation
  - Kontinuierliche Funktion $f(x)$: Regression

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Vorteile von Entscheidungsbäumen



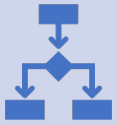
- Relativ schnell im Vergleich zu anderen Klassifikationsmodellen
- Erzielt oft gute Genauigkeit im Vergleich zu anderen Modellen
- Einfach und leicht zu verstehen
- Kann in einfache und leicht verständliche Klassifikationsregeln umgewandelt werden

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Wie werden Entscheidungsbäume erzeugt?

Splittingregel bestimmen (Attribut und Schwellwert)

Datensatz mithilfe der Splittingregel in disjunkte Teildatensätze aufteilen

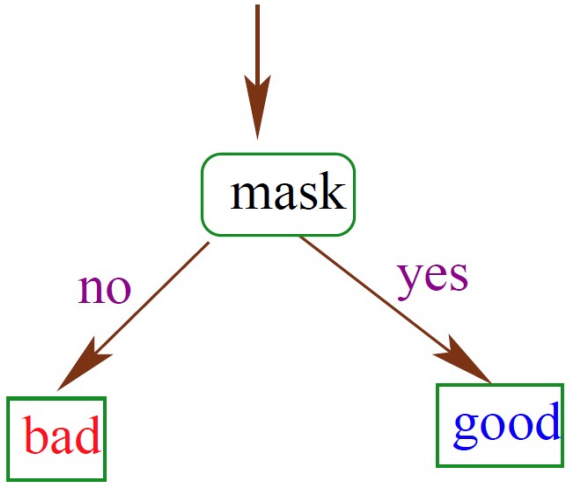


So lange wiederholen bis Teildaten ausreichend rein sind

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

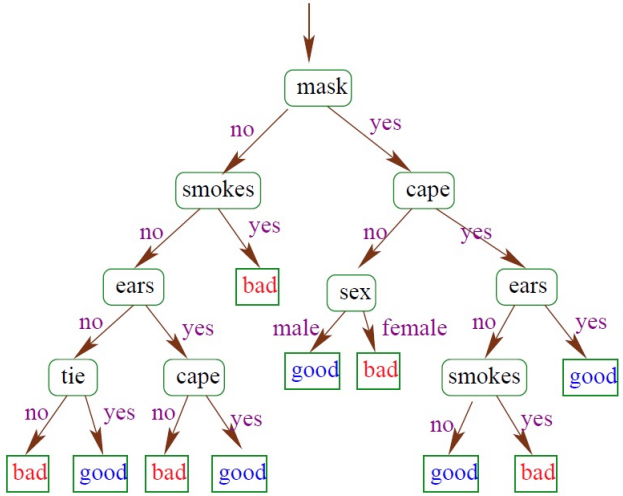
Baumgröße

Unteranpassung



- Zu einfach
- Bereits die Trainingsdaten können nicht gut klassifiziert werden

Überanpassung

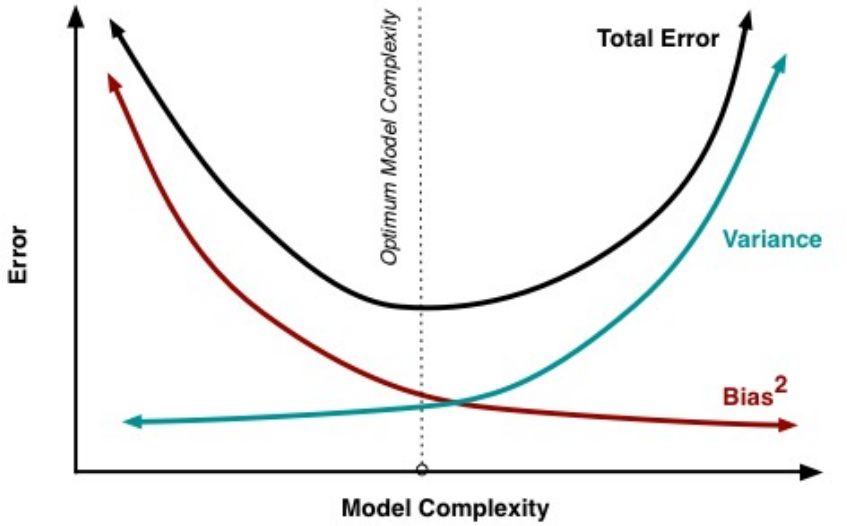


- Sehr kompliziert
- Trainingsdaten warden perfekt gelernt
- Verallgemeinerbar??

Verzerrung-Varianz-Tradeoff

Unteranpassung
Underfitting

Überanpassung
Overfitting



Strategien zur Vermeidung von Überanpassung

- Testdatenmanagement
- Regularisierung
- Ensemblemethoden