Maschinelles Lernen und Datenanalyse

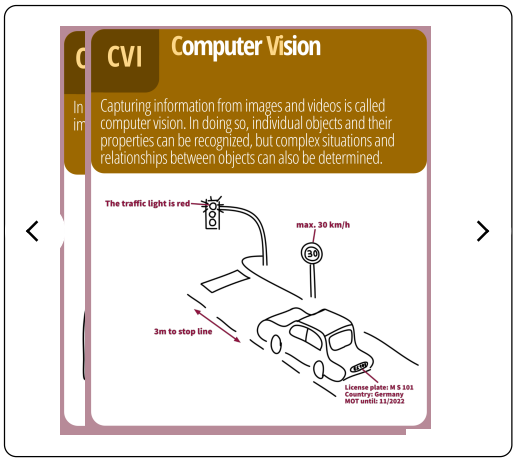
In der Mess- und Prüftechnik

PD Stefan Bosse

Universität Bremen - FB Mathematik und Informatik

Überblick

Anwendungsklassen von Maschinellen Lernen



Motivation

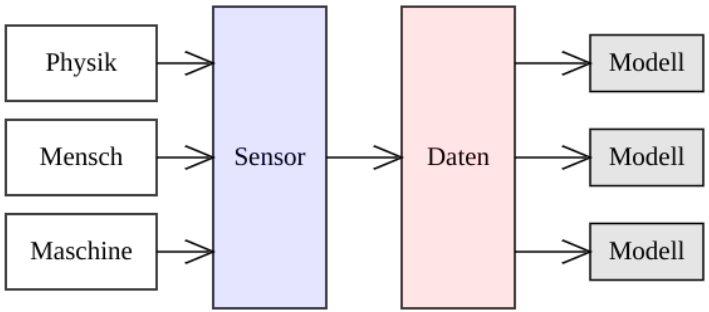
Dieser Kurs mit interaktiven Übungen soll:

- Einen **anwendungsorientierten Einstieg** in die Datenanalyse und Interpretation mit Verfahren des **Maschinellen Lernens** bieten;
- Einen **Überblick** über gängige und weniger gängige **Verfahren** geben;
- **Interaktive Tutorials und Übungen mit zielgruppenorientierten Fallbeispielen** sollen Verfahren begreifbar und erfahrbar machen!

Inhalte

- Die Ontologie des Kurses besteht aus den **Bausteinklassen**:
  - **Modelle** (Datenstrukturen)
  - **Verfahren** (Algorithmen: Training, Test, Inferenz)
  - Überwachtes Training
  - Nichtüberwachtes Training
- Weiterhin aus den Anwendungs- und **Datenklassen**:
  - Sensorische und experimentelle Daten (Mess- und Prüftechnik)
  - Erhebungs- und Umfragedaten (Soziologie) ⇒ Der Mensch als Sensor!
  - Metrische und Kategorische Variablen

Die Grenzen der Datenklassen sind fließend! Material, Maschine und Mensch als Sensoren!



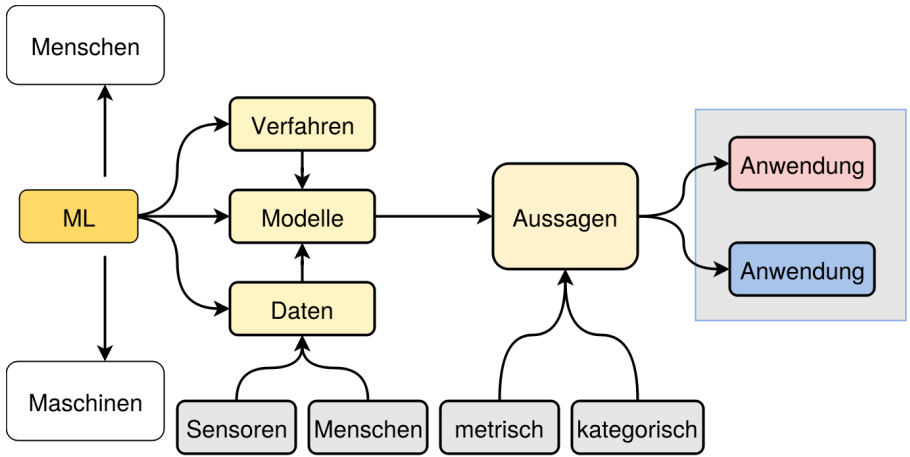


Abb. 1. Gemeinsame Verfahren und Modelle → Unterschiedliche Daten, Aussagen, Anwendungen

Organisation der Veranstaltung

1. **Vorlesungen mit integrierten Übungen**
   - Vermittlung der Grundlagen
   - Unmittelbare Übung und Anwendung der Grundlagen mit einfachen Übungen
2. **Asynchrone Videos und Tutorials**
   - Auch offline seh- und hörbar
3. **Gemeinsame Treffen mit** Videokonferenz (Zoom, falls erforderlich)
4. **Interaktive Tutorials** und Übungen mit *NoteBook* und *WorkBook* (NoteBook-2) im Web Browser!
   - Offline ausführbar (evtl. werden Daten von einem Server geladen)

Organisation der Veranstaltung

5. Texte und Folien
   - Vorlesungsskript (am Anfang: für jedes Modul/jede Einheit) als Ebook
   - Das Vorlesungsskript gibt die Folieninhalte 1:1 wieder (nur anderes Layout und kompaktiert)
   - Alle Folien im HTML Format (auch offline lesbar)
   - Begleitende Literatur (Bücher im PDF)

Services

1. Web Service: Informationen, Dokumente, Folien, Videos: https://edu-9.de/Lehre/ml3k
2. Dokuwiki: **News**, Informationen und Links, **Chats**, **Videostreams**: https://ag-0.de/dokuwiki
   - Registrierung und Login erforderlich
   - Interaktiv!
3. Videos: https://edu-9.de/Lehre/ml3k

Prüfungsleistungen

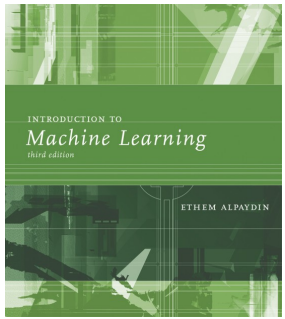
1. Eine mündliche Abschlussprüfung (20 Minuten); **oder alternativ** 2.
2. Eine schriftliche Seminararbeit (Experimentelle Arbeit oder Literaturrecherche)
   - 15-20 Seiten PDF
   - Grundstruktur: Erarbeitung des wissenschaftlichen Standes, Diskussion und Bewertung, Beschreibung und Dokumentation der experimentellen Arbeit, Diskussion von Ergebnissen (bei experimenteller Arbeit), Zusammenfassung
3. Bearbeitung und Abgabe der digitalen Übungen (JSON Dateien)
   - Punktesystem: 0/1/2 für Aufgaben und gesamten Übungszettel
   - Es muss jeder Übungszettel eingereicht werden und wenigstens einen Punkt erhalten.

Literatur

- Zur Vertiefung!

S. Richter, Statistisches und maschinelles Lernen. Springer Spektrum, 2019.

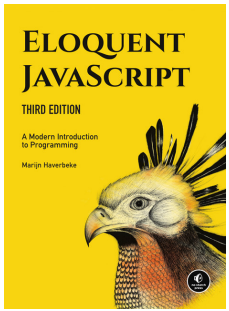
E. Alpaydın, Introduction to Machine Learning. MIT Press, 2010.



Programmierung

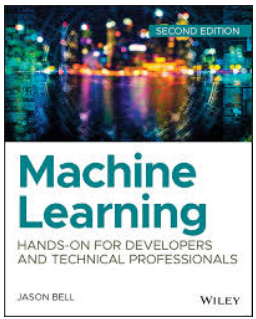
Axel Rauschmayer, JavaScript For Impatient Programmers.

M. Haverbeke, Eloquent JavaScript. 2018.

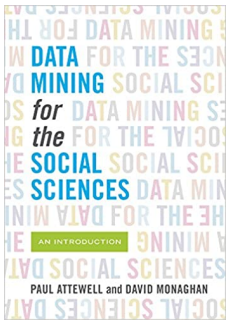


Domainspezifische Literatur

J. Bell, Machine Learning - Hands-On for Developers and Technical Professionals. John Wiley & Sons, Ltd, 2015.



P. Attewell and D. B. Monaghan, Data mining for the social sciences : an introduction. University of California Press, 2015.



Software

NoteBook

- Interaktive vorwiegend praktische Übungen werden rein digital im Web Browser mit den *NoteBooks* durchgeführt
- Ein digitale Übung (oder Turorial) besteht aus:
  - Textabschnitten
  - Informationsblöcken
  - Aufgaben (mit Lösungen)
  - Editoren für Programmcode
  - Ausführungsterminals für Programmcode
  - uvm.

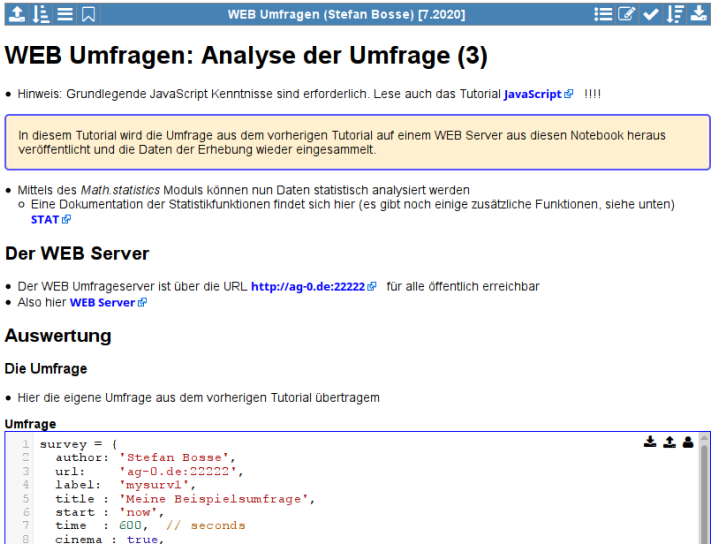


Abb. 2. Ein NoteBook im WEB Browser

NoteBook Konzept

- Top-down Bearbeitungsfluss
- Statische Struktur mit dynamischen Inhalten
- Alle dynamischen Inhalte können in einer JSON Datei gespeichert und wieder geladen werden
- Es können Notizzettel überall im NoteBook angeheftet werden (werden auch gespeichert)
- Musterlösungen (dynamische Inhalte) können eingebettet und mit einem Schlüssel freigeschaltet werden

WorkBook

- Dynamische Struktur mit dynamischen Inhalten
- Ein WorkBook besteht aus
  - Textabschnitten (MarkDown)
  - Codesnippets mit Editoren und Ausgabekonsolen
  - Speziellen Snippets wie editierbare Tabellen oder allg. Formulare
- Programmierung in JavaScript, aber menügesteuerte und geführte Auswahl von Ausführungsblöcken mit einer kursspezigischen Bibliothek
- Alle dynamischen Inhalte und Daten können im JSON Format gespeichert und wieder geladen werden

Die NoteBook Konzepte (Editierbare Aufgaben und Einreichungs-/Hilfefunktion) sind jetzt auch hier integriert

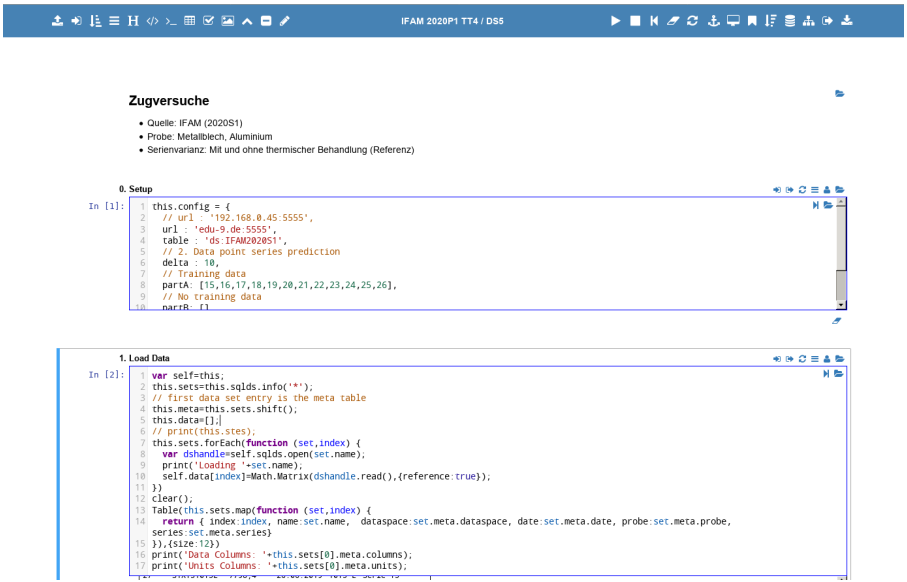


Abb. 3. Ein WorkBook Beispiel

Machinelles Lernen

Schlüsselwörter und Begriffe

Welche Begriffe werden häufig bei ML genannt:

Anwendungsgebiete

Welche Anwendungsgebiete gibt es:

Fragestellungen

Welche Fragestellungen (zu lösende Probleme) gibt es:

Inhalte

1. **Eingabe** x: Daten (Attribute) und Eigenschaften (Analyse)
2. **Sensoren**: Erfassung von Daten, $S(welt)$: $welt \rightarrow x$
3. **Ausgabe** y: Numerische und kategorische Werte
4. **Metriken und Taxonomie**: Grundlagen des Maschinellen Lernens
5. **Algorithmen und Modelle**: $f(x)$: $x \rightarrow y$
6. **Training, Lernen, Prädiktion, Test** $M(<x,y>)$: $<x,y> \rightarrow f$
7. **Anwendungen**

Geschichte

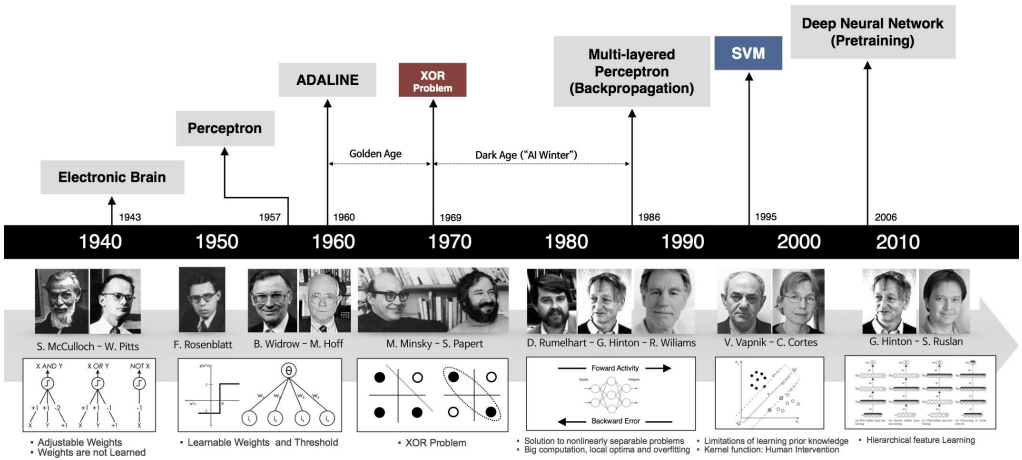


Abb. 4. Die Geschichte fokussiert auf Neuronale Netze. Es gibt mehr.

[www.pinterest.com]

Modelle

1. Entscheidungsbäume (gerichtete Graphen)
2. Funktionen (z.B. Polynome)
3. Funktionsgraphen (z.B. künstliche neuronale Netzwerke)

Algorithmen

1. Entscheidungsbäume: C4.5, ID3, IDT, Regressionsbäume
   - Teilungsverfahren (mit Entropie, Informationsgewinn, usw.)
2. Lineare und nichtlineare Regression, Support Vector Machines (SVM)
   - Least Square Fit (iterativ, mehrschrittig)
   - Lineare Algebra (numerisch, einschrittig)
3. Datenanalyse!: Hauptkomponentenanalyse, statistische Methoden
4. Bayesien Netzwerke mit statistischem Methoden (probabalistische Verfahren)
5. Rückwärtspropagation von Fehlergradienten (vor allem KNN) aus Vorwärtsberechnung
6. Überwachte und nichtüberwachte Trainingsverfahren
7. Zustandsbasierte Funktionen (LSTM) für Datenserien

Datenanalyse und Eigenschaftsselektion

Wir unterscheiden folgende Klassen von Eigenschaften in der Datenanalyse und Prädiktion (Merkmale, Features):

1. Eigenschaften der Eingabedaten, vor allem dominante Eigenschaften abgeleitet aus den Eingabedaten x mit starker y Korrelation
    - Beispiel: Charakteristische Signalfrequenz einer Betriebsschwingung die auf einen Schaden hindeutet
2. Zieleigenschaften, also Werte der Zielvariable y
    - Numerische Eigenschaften (kontinuierlich oder diskret), z.B. Materialdichte, Schadensposition, Bruchdehnung
    - Kategorische Eigenschaften, Z.B. Farbe, Tierart, Schadensklasse, Entscheidungen

Datenanalyse und Eigenschaftsselektion

Häufig sind die rohen sensorischen Daten(variablen) zu hochdimensional und noch abhängig voneinander (schwache Korrelation mit y)

Datenanalyse und Eigenschaftsselektion

Häufig sind die rohen sensorischen Daten(variablen) zu hochdimensional und noch abhängig voneinander (schwache Korrelation mit y)

Reduktion auf wesentliche Merkmale kann ML Qualität deutlich verbessern!

Datenanalyse und Eigenschaftsselektion

Häufig sind die rohen sensorischen Daten(variablen) zu hochdimensional und noch abhängig voneinander (schwache Korrelation mit y)

Reduktion auf wesentliche Merkmale kann ML Qualität deutlich verbessern!

Häufig besitzen einzelne Sensorvariablen keine oder nur geringe Aussagekraft (geringe Entscheidbarkeitsqualität) $\rightarrow$ geringe bis keine Korrelation mit y oder sogar Antikorrelation (Störung)

Datenverarbeitung

- Die Daten die als Grundlage für die Induktion (Lernen) und die Deduktion (Applikation/Inferenz der Zielvariablen) müssen i.A. vorverarbeitet werden → **Merkmalsselektion**

Datenverarbeitung

- Die Daten die als Grundlage für die Induktion (Lernen) und die Deduktion (Applikation/Inferenz der Zielvariablen) müssen i.A. vorverarbeitet werden → **Merkmalsselektion**

[6]

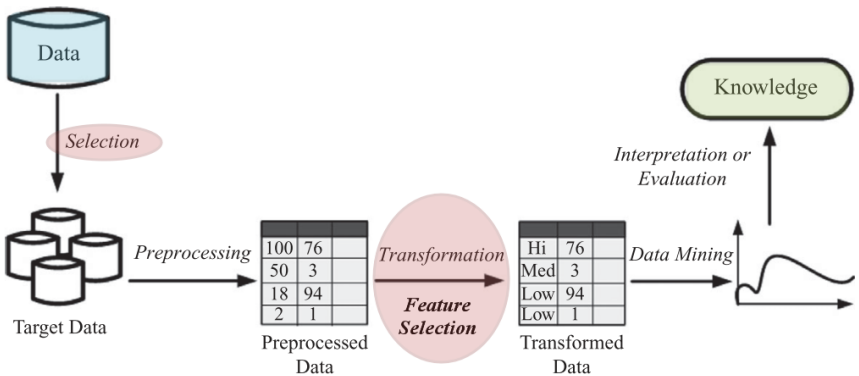


Abb. 5. Maschinelles Lernen ist ein Werkzeug der Datenanalyse und des Data Minings

Modellbildung

[100]

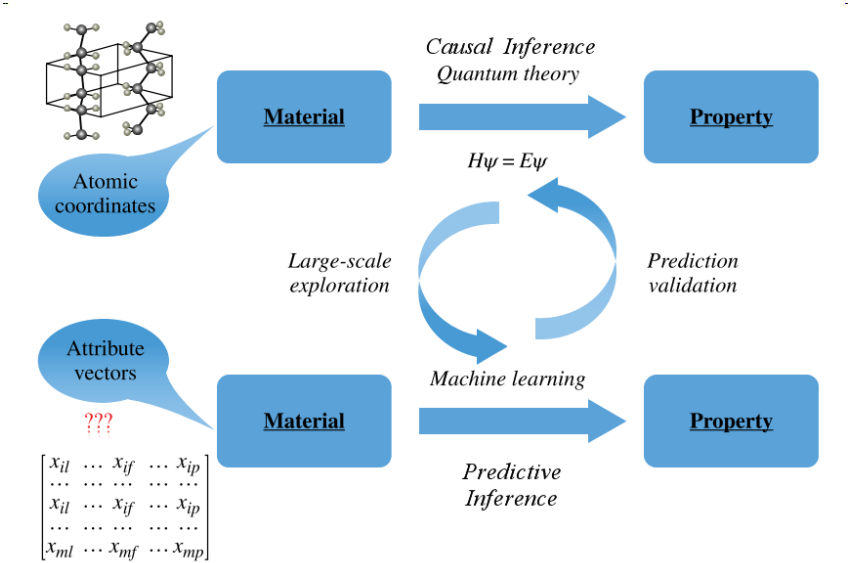


Abb. 6. Kausale vs. Prädiktive Modellbildung und Physikalische Modelle versa algorithmisch bestimmte Modelle (Hypothesen)

Induktion und Deduktion

[6]

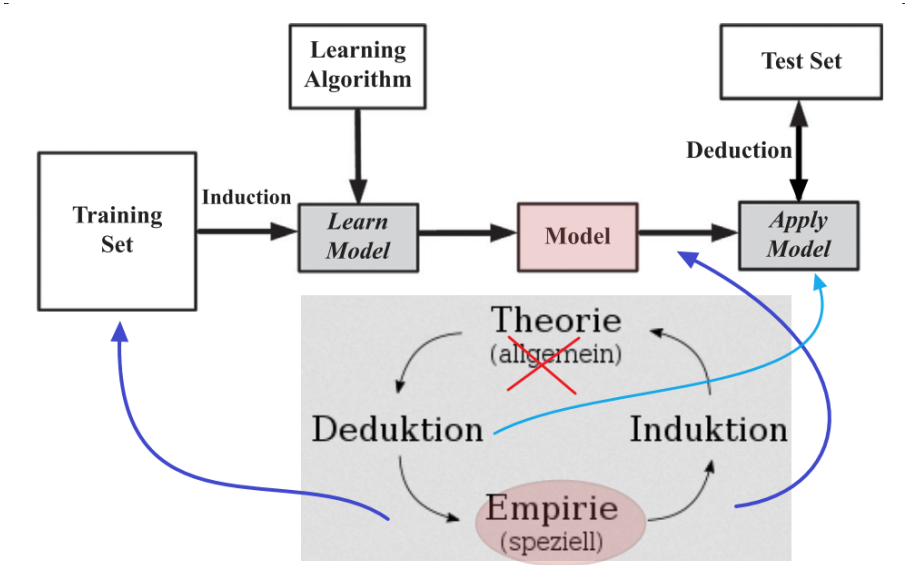


Abb. 7. Ablauf Überwachtes Lernen mit Trainings- (Induktion) und Applikationsphasen (Deduktion). Aber: Meistens keine Verallgemeinerung!