Classification and Clustering

Prof. Dr. Yvonne Gorniak

31. März 2025

Organisatorisches

Unser Ziel

Die Teilnehmer besitzen ein gut fundiertes Wissen über Methoden und Verfahren des Data Mining. Dazu gehören Methoden zur Klassifikation und zum Clustering von mehrdimensionalem Datenmaterial. Sie können Techniken im Bereich des Machine Learning einschätzen und erarbeiten, klassische und moderne Methoden wissenschaftlich fundiert darstellen und in neuartigen Praxisbeispielen einsetzen.

Prüfungsleistung: Referat/Projekt

Wird in einem separaten Dokument bereitgestellt.

Termine - Montags (E,4)

31.03.

07.04.

28.04.

12.05.

02.06.

16.06.

23.06.

30.06.

07.07.

01.09. (Referate, Projektvorstellung)

08.09. (Referate, Projektvorstellung

Literatur

- **Praxiseinstieg Machine Learning mit Scikit-Learn, Keras und TensorFlow**, Géron, O'Reilly, 2019
- **Data Mining for Business Analytics - Concepts, Techniques and Applications in Python**, Shmueli et al., Wiley 2020, `dataminingbook.com`
- **Multivariate Analysemethoden**, Backhaus et al., Springer Gabler, 16. Auflage, 2018, `https://multivariate.de/`

Arbeitsumgebung

Wir nutzen in dieser Vorlesung Jupyter Notebooks zusammen mit Python 3.x.

Mögliche Arbeitsumgebung sind z.B.

- Google Colab `https://colab.research.google.com/`
- Jupyter Notebook oder Jupyter Lab `https://jupyter.org/`

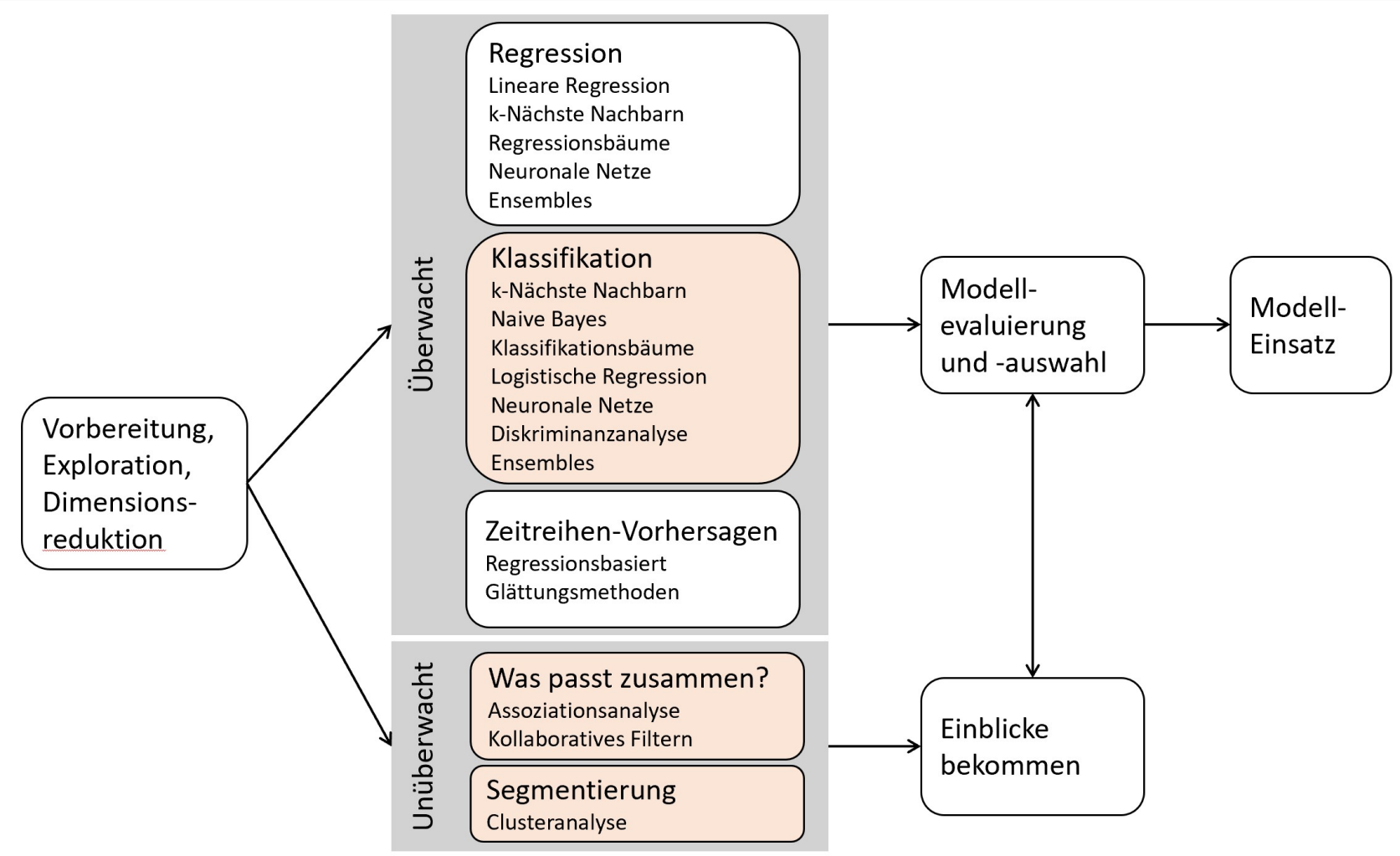
- **A Whirlwind Tour of Python**, Jake VanderPlas, O'Reilly, https://jakevdp.github.io/WhirlwindTourOfPython/index.html
- **Python Data Science Handbook**, Jake VanderPlas, O'Reilly, https://jakevdp.github.io/PythonDataScienceHandbook/index.html

Python Bibliotheken: Data Science

- **NumPy**: Numerische Analyse (Arrays, Matrizen)
- **Pandas**: Komplexe Datananalyse (Dataframes, Zeitreihenanalyse)
- Statsmodels: Statistische Datenanalyse
- SciPy: Lineare Optimierung
- **Scikit-learn**: Machine Learning
- Mlpy: Machine Learning
- Keras: Deep Learning
- TensorFlow: Deep Learning (Google)
- PyTorch: Deep Learning (Facebook)
- NLTK: Text Mining
- **Matplotlib**: Visualisierung
- Bokeh: Interaktive Datenvisualisierung
- **plotnine**: Visualisierung mit der „Grammar of Graphics“ (basiert auf ggplot2)

Einordnung

Einordnung der Inhalte in den Data Mining Prozess



Regression

Regression meint die Vorhersage eines numerischen Wertes.

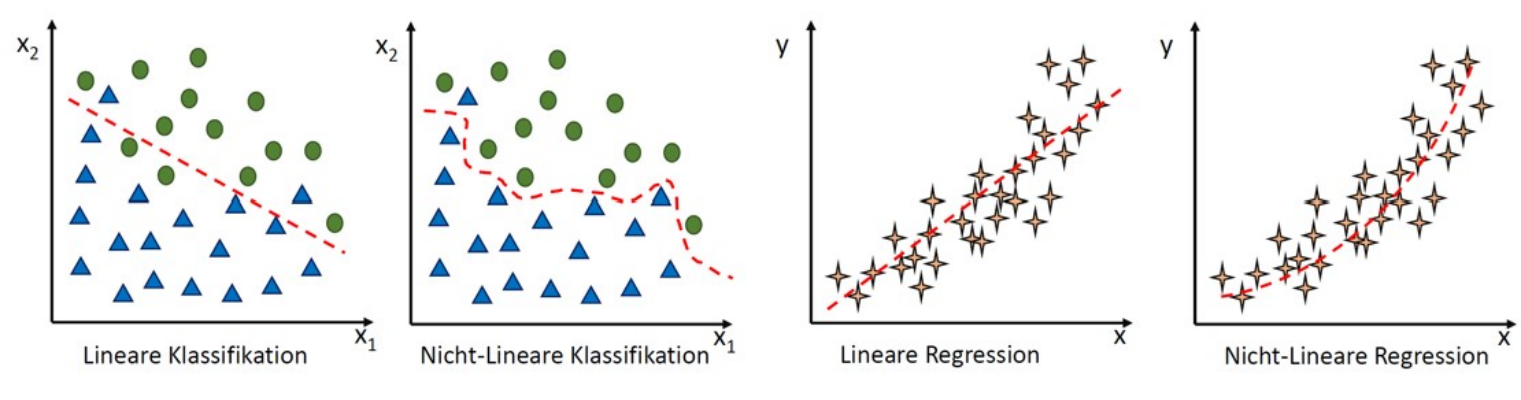
- Vorhersage der Umsatzentwicklung eines Unternehmens basierend auf Performance-Metriken
- Vorhersage von Kriminalitätsraten auf Basis von Nachbarschaft, Lage, Infrastruktur, etc.
- Vorhersage von Einkommen auf Basis von Alter und Bildung

Klassifikation

Klassifikation meint die Vorhersage eines kategorialen Wertes (einer Klasse).

- Bilderkennung (Produktbilder, Tumore, ...)
- Nachrichtenartikel klassifizieren
- Beleidigende Kommentare automatisch markieren
- Spam erkennen
- Kreditnehmer klassifizieren
- Kundenabwanderung erkennen
- u.v.m.

Klassifikation vs. Regression



Viele Regressionsalgorithmen lassen sich auch zur Klassifikation einsetzen und umgekehrt.

Clustering

Ziel des Clustering ist, ähnliche Instanzen in Clustern zu gruppieren. Clusterverfahren werden in verschiedenen Bereichen eingesetzt.

- Datenanalyse
- Kundensegmentierung
- Empfehlungssysteme
- Suchmaschinen
- Bildsegmentierung
- Dimensionsreduktion
- Anomalieerkennung
- u.v.m.

Data Mining

Data Mining (= Daten schürfen) versucht Muster in großen Datenmengen zu ermitteln. Muster sind z.B.:

- Zusammenhänge: Wie hängt der Absatz vom Produktpreis und der Bonität der Kunden ab?
- Ähnlichkeitsstrukturen: Welche Gruppen von Kunden mit gemeinsamen Eigenschaften lassen sich bilden?
- Abhängigkeiten: Was kaufen die Menschen, die Bier kaufen, zusätzlich ein?

Data Mining legt den Fokus auf die Entdeckung neuen Wissens, während Maschinelles Lernen den Fokus auf Vorhersagen legt. Beide Felder verwenden aber oft dieselben Methoden, mit unterschiedlicher Zielsetzung.

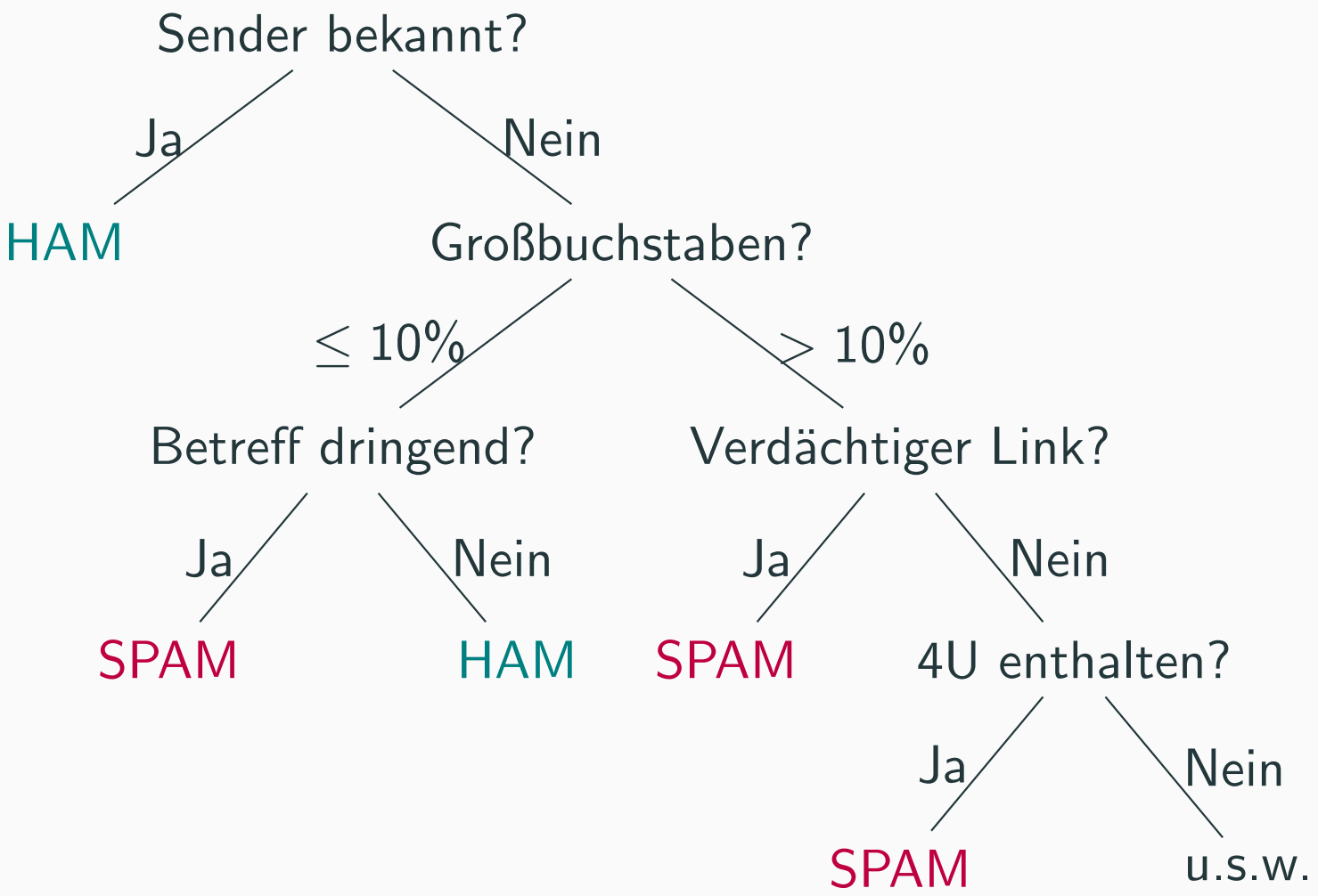
Machine Learning

Machine Learning (= Maschinelles Lernen) beschreibt das automatisierte Erlernen von Wissen auf Basis von Erfahrungswissen. Dieses erlernte Wissen wird durch ein Modell repräsentiert, welches zum Treffen von Vorhersagen verwendet wird.

Machine Learning Ingenieure...

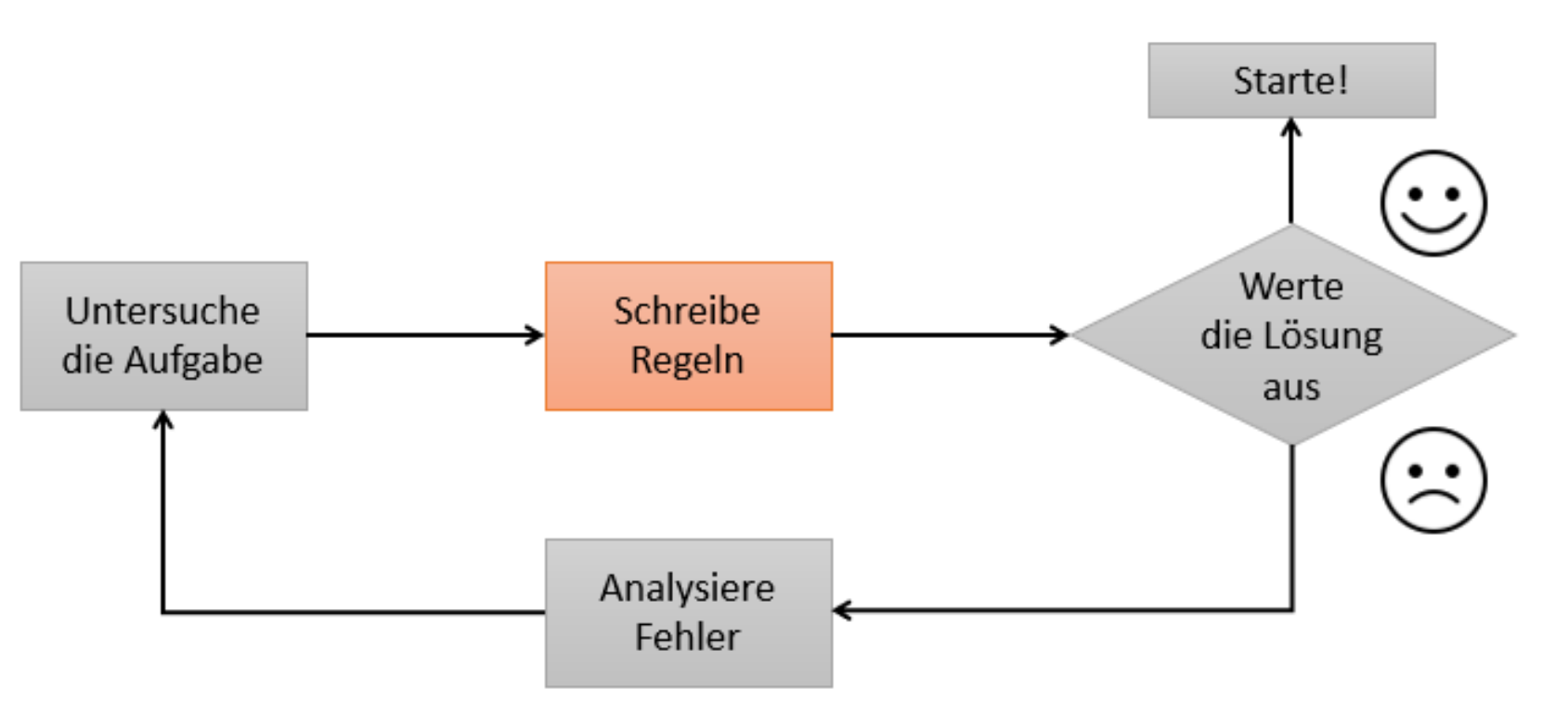
- stellen Machine Learning Pipelines in produktiven Umgebungen bereit
- entwickeln neue oder verbessern bestehende Modelle
- analysieren dazu große und komplexe Datenmengen

Beispiel: Ein Spam Filter



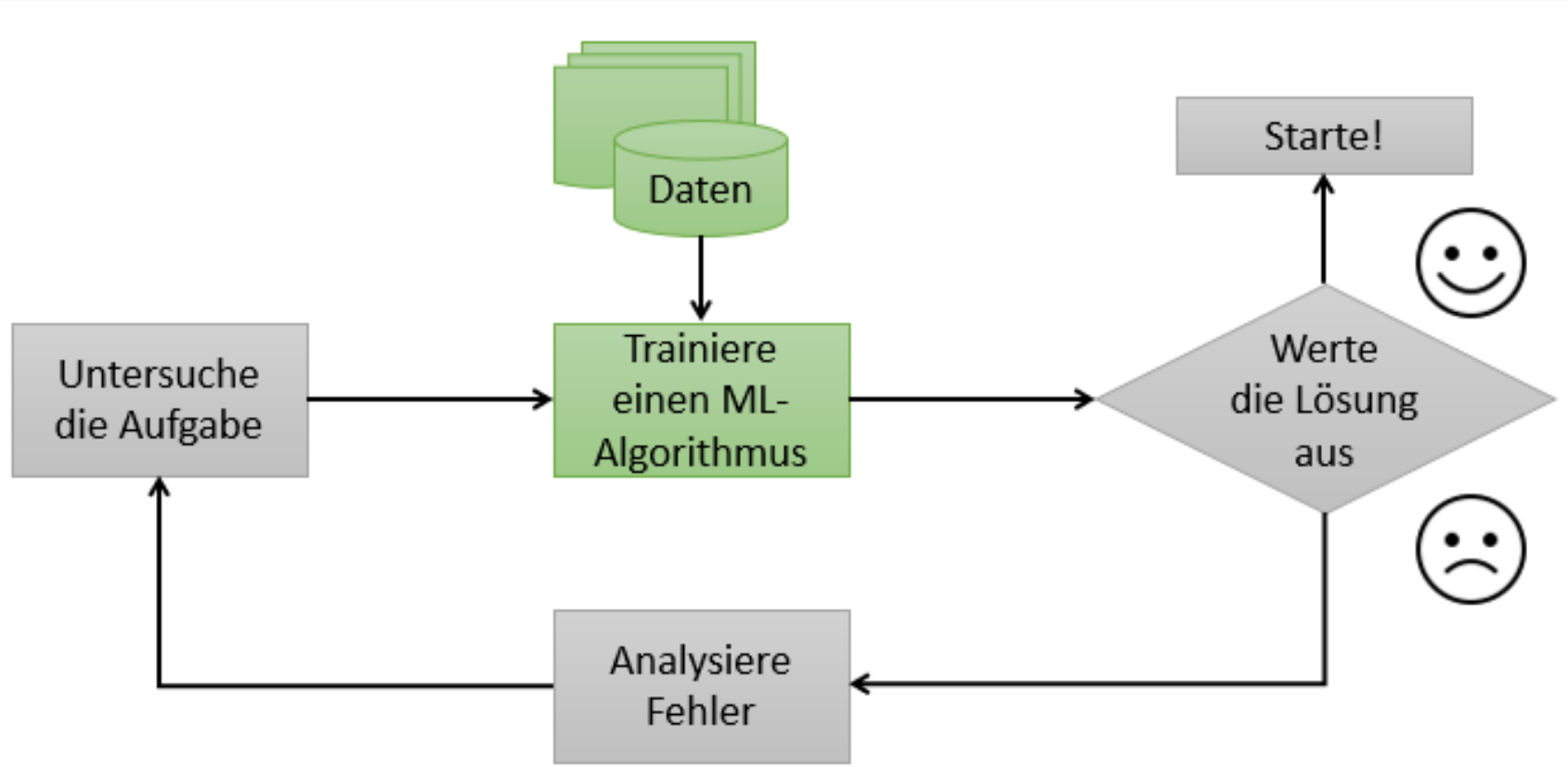
Wie entsteht dieser Baum?

Klassischer Ansatz



Im klassischen Ansatz werden Regeln aufgestellt und diese in einem Programm zur Spam-Erkennung eingebaut.

Ansatz des Maschinellen Lernens



Im ML-Ansatz werden die Regeln nicht *aufgestellt*, sondern selbst durch den Computer gelernt.

Maschinelles Lernen

Künstliche Intelligenz kann auf programmierten Abläufen basieren oder durch maschinelles Lernen erzeugt werden.

Definition
Maschinelles Lernen

Maschinelles Lernen ist das Fachgebiet, das Computern die Fähigkeit zu Lernen verleiht, ohne explizit programmiert zu werden (Arthur Samuel, 1959)

Einsatzgebiete des Maschinellen Lernens

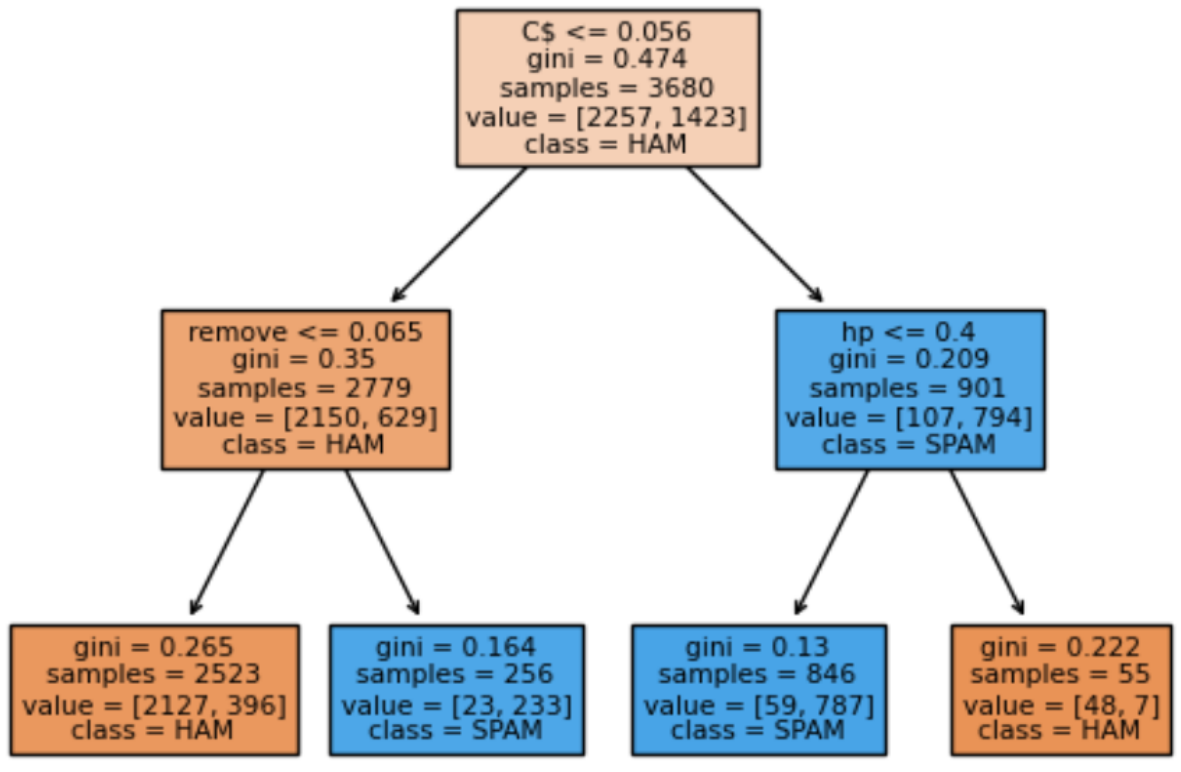
- Aufgaben, die lange Listen von Regeln erfordern
- Aufgaben, die so komplex sind, dass es mit herkömmlichen Methoden keine guten Lösungen gibt
- Systeme in Umgebungen, die sich laufend verändern (ML kann sich gut neuen Daten anpassen)
- Erkenntnisse über komplexe Aufgabenstellungen und große Datenmengen gewinnen

Modelle

Im Maschinellen Lernen müssen Datenzusammenhänge modelliert werden. In diesem Zusammenhang wird das Wort **Modell** mehrdeutig verwendet.

- *Art des Modells*, bspw. Lineare Regression, Entscheidungsbaum, Neuronales Netz
- *Modellarchitektur*, in der bspw. auch die Ein- und Ausgabe der Daten beschrieben ist
- *Trainiertes Modell*, für das bereits die besten Modellparameter gefunden wurden und welches nun für Vorhersagen genutzt werden kann.

Beispiel eines Entscheidungsbaumes zur SPAM Erkennung



Notebook: 0_SPAM_Beispiel

Überwachtes Lernen

- Regression und Klassifikation nutzen überwachte Verfahren. Es müssen Daten vorliegen, für die der Zielwert bekannt ist (gelabelte Daten).
- Anhand dieser Trainingsdaten lernt das Modell die Zusammenhänge zwischen den Prädiktoren und dem Zielwert,das Modell wird trainiert.
- Auf Basis weiterer gelabelter Daten (Validierungsdaten) wird die Performance verschiedene Modelle miteinander verglichen.
- Zuletzt wird für das gewählte Modell anhand weiterer gelabelter Daten (Testdaten) überprüft, ob das Modell eine gute generelle Vorhersage leisten wird. Später mehr hierzu.

Modell: Ein Algorithmus, wie er auf ein Datenset angewendet wird, zusammen mit seinen Voreinstellungen (Hyperparameter).

Unüberwachtes Lernen

- Beim unüberwachten Lernen sind die Daten nicht gelabelt. Das System versucht, ohne Anleitung zu lernen.
- Clustering und Assoziationsanalyse verwenden unüberwachte Lernalgorithmen.

Es gibt auch noch das **halbüberwachte** Lernen für den (häufigen) Fall, dass man viele ungelabelte und wenige gelabelte Instanzen hat (Anwendungbeispiel: Google Photos).

Außerdem das **Reinforcement Learning**, in dem das Lernsystem Belohnungen oder Strafen für seine Aktionen bekommt und mit der Zeit seine Policy anpasst, um möglichst viele Belohnungen zu erhalten (Anwendungsbeispiel: Laufenlernen von Robotern, AlphaGo).

Weitere Einteilungen von ML-Systemen

Batch vs. Online

- **Batch-Learning**: Das System kann nicht inkrementell lernen, sondern muss mit sämtlichen verfügbaren Daten (offline) trainiert werden.
- **Online-Learning**: Das System wird nach und nach trainiert, mit einzelnen Datensätzen (*Mini Batches*) nacheinander.

Instanzbasiert vs. Modellbasiert

- **Instanzbasiertes Lernen**: Das System lernt die Beispiele auswendig und verallgemeinert dann mit Hilfe eines Ähnlichkeitsmaßes auf neue Fälle, wobei es sie mit den gelernten Beispielen vergleicht.
- **Modellbasiertes Lernen**: Das System entwickelt ein Modell aus den Beispielen und verwendet dieses Modell dann für Vorhersagen.

Herausforderungen beim Machine Learning

- Unzureichende Menge an Trainingsdaten
- Nicht repräsentative Trainingsdaten (z.B. durch Stichprobenverzerrung)
- Minderwertige Daten
- Irrelevante Merkmale
- Overfitting der Trainingsdaten
- Underfitting der Trainingsdaten

Testen und Validieren

- Zur Bewertung, ob ein Modell gut verallgemeinert, wird ein *Testdatensatz* beiseite gelegt und das Modell nur mit dem verbleibenden *Trainingsdatensatz* trainiert. Wenn der *Trainingsfehler* gering ist, der *Verallgemeinerungsfehler* jedoch groß, overfittet das Modell die Trainingsdaten.
- Für Zwischenevaluierungen im Auswahlprozess (z.B. Einstellung von Hyperparametern, Auswahl zwischen unterschiedlichen Modellen) darf der Testdatensatz nicht verwendet werden. Für diesen Zweck wird noch ein Teil des Trainingsdatensatzes als *Validierungsdatensatz* beiseite genommen (*Hold-Out-Validierung*).
- Es ist üblich, Validierungen mehrfach mit verschiedenen Validierungsdaten durchzuführen und diese im Mittel zu bewerten. Dieses Verfahren wird *Kreuzvalidierung* genannt.

Aufwärmen: Pandas / Datassets

Pandas

Notebook: 0_Pandas_Uebungen

Lösungen bzw. Lösungsansätze:

Notebook: 0_Pandas_Uebungen_Loesungen

Schon fertig?! `https://calmcode.io/challenge/`

Aufwärmen: Scikit-Learn / ML

Einführung in Scikit-Learn:

https://calmcode.io/scikit-learn/introduction.html

Notebook: 0_CalmCode_scikit-learn

Ein Machine Learning Projekt von A bis Z

Notebook: 0_Ein_Machine_Learning_Projekt