Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Datenmanagement & -analyse

Maschinelles Lernen: Von der Erklärung zur Prädiktion

Prof. Dr. Christoph M. Flath
Lehrstuhl für WI & BA
Julius-Maximilians-Universität Würzburg

Sommersemester 2021

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

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

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

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

Agenda

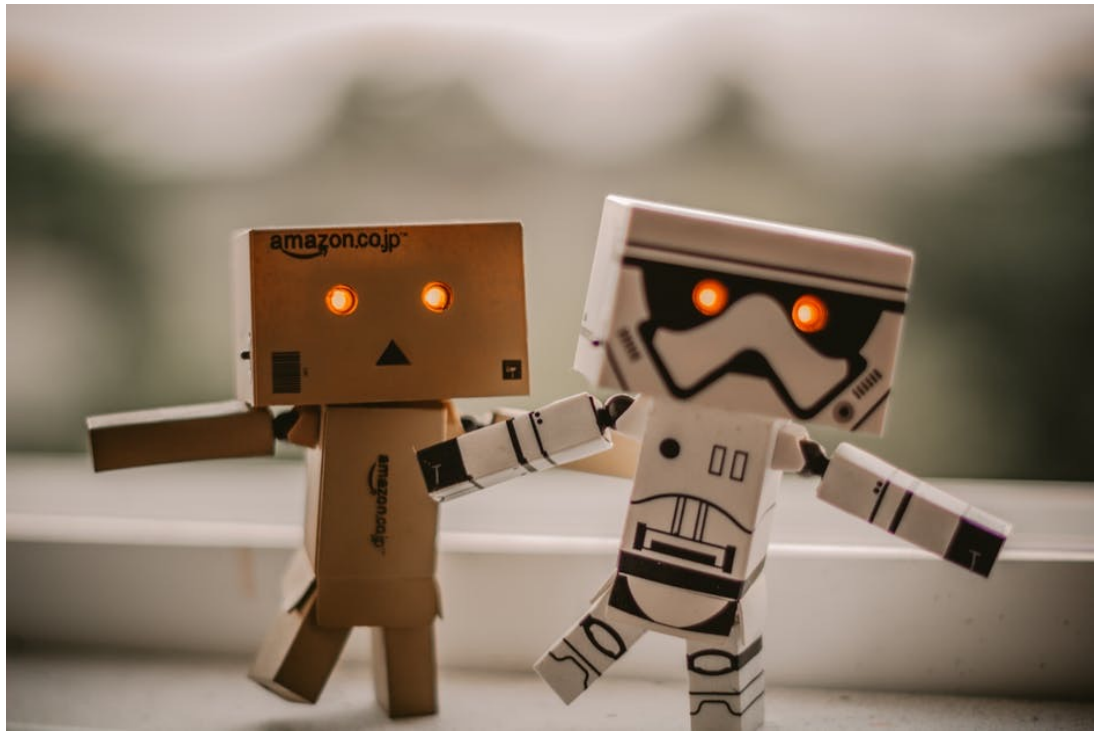
1. **Grundlagen des Lernens**
2. **Entscheidungsbäume**
3. **Bias-Varianz Tradeoff**

What is Learning?

- Herbert Simon: “Learning is any process by which a system improves performance from experience.”

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.”



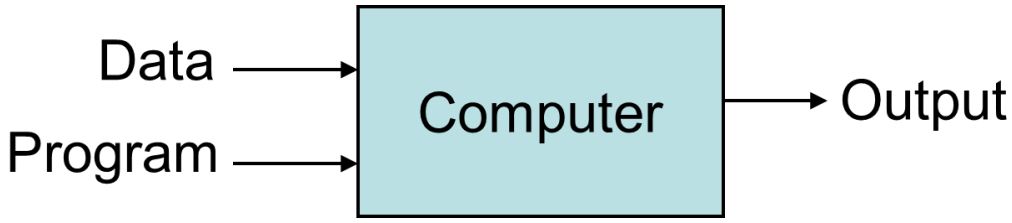
Why let machines “learn”?

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

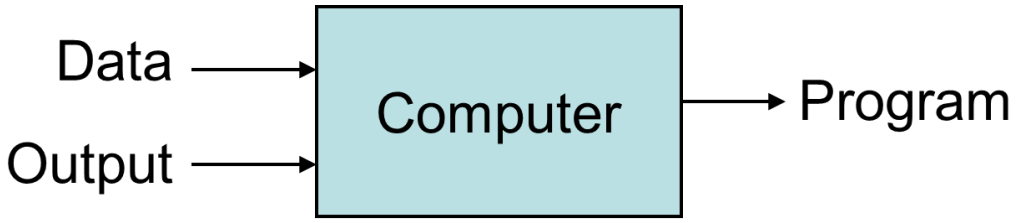
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (traffic routing)
  - Solution needs to be adapted to particular cases (user biometrics)

Ein Paradigmenwechsel

Klassische Programmierung



Maschinelles Lernen



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

Es ist wie bei der Gartenarbeit

- Setzling = Algorithmen
- Dünger = Daten
- (Gärtner = Programmierer)

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

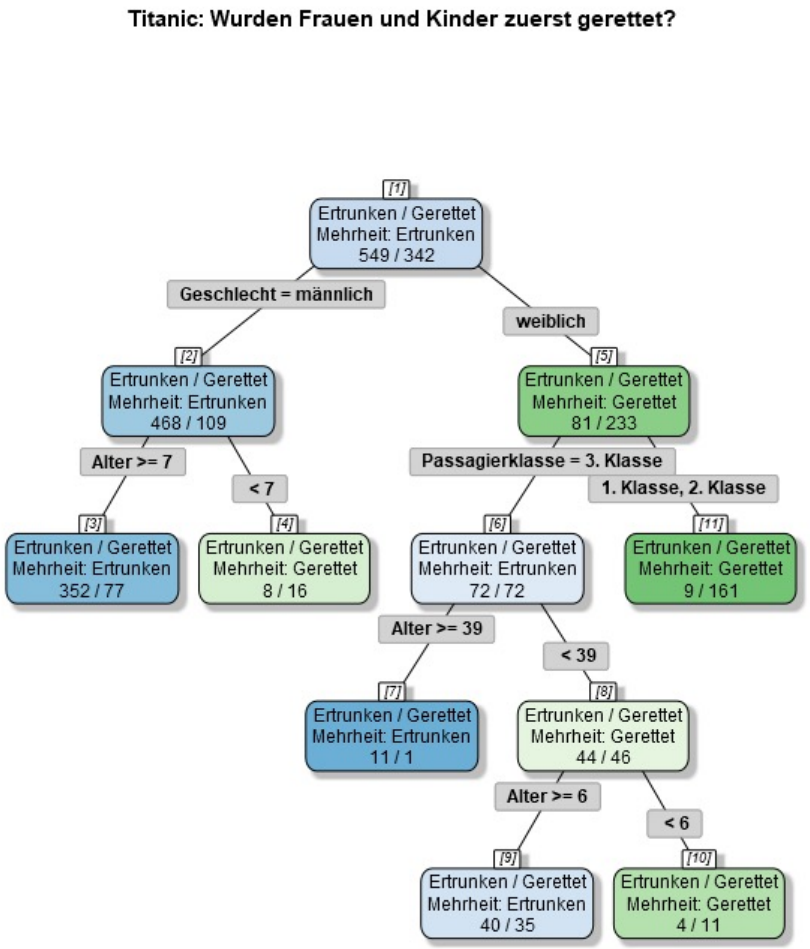
- Auf Basis von Trainingsdaten (X, Y) eine Funktion $X \mapsto Y = f(x)$ bestimmen
- Die Funktion soll zuverlässig Y-Werte für neue Datenpunkte X bestimmen
  - Diskrete Funktion $f(x)$: Klassifikation
  - Kontinuierliche Funktion $f(x)$: Regression

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

1. Grundlagen des Lernens
2. Entscheidungsbäume
3. Bias-Varianz Tradeoff

Vorteile von Entscheidungsbäumen

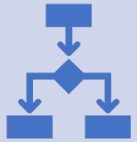


- Relativ schnell im Vergleich zu anderen Klassifikationsmodellen
- Erzielt oft gute Genauigkeit im Vergleich zu anderen Modellen
- Einfach und leicht zu verstehen
- Kann in einfache und leicht verständliche Klassifikationsregeln umgewandelt werden

Wie werden Entscheidungsbäume erzeugt?

Splittingregel bestimmen (Attribut und Schwellwert)

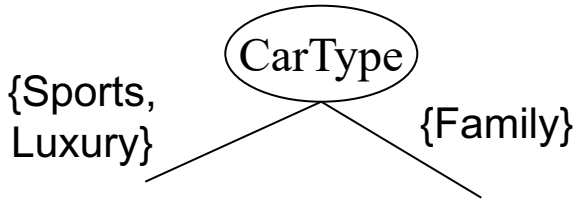
Datensatz mithilfe der Splittingregel in disjunkte Teildatensätze aufteilen

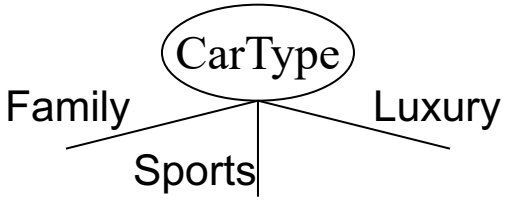


So lange wiederholen bis Teildaten ausreichend rein sind

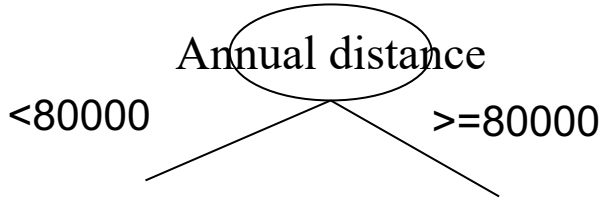
Mögliche Splittinregeln

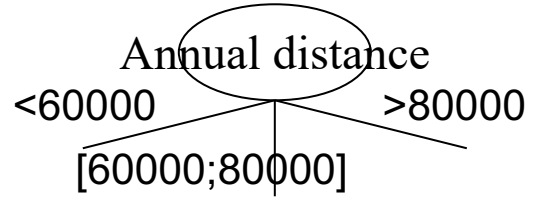
- Abhängig vom Attributtyp
  - Nominal
  - Ordinal
  - Kontinuierlich





- Abhängig von der Anzahl Unterdatensätze
  - 2-Wege Split
  - Mehr-Wege split

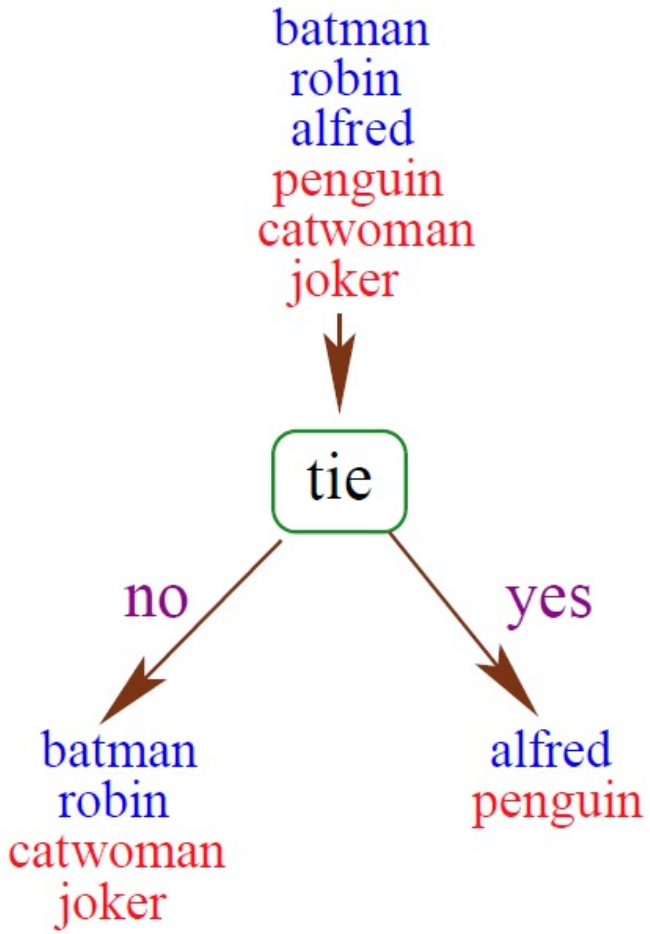


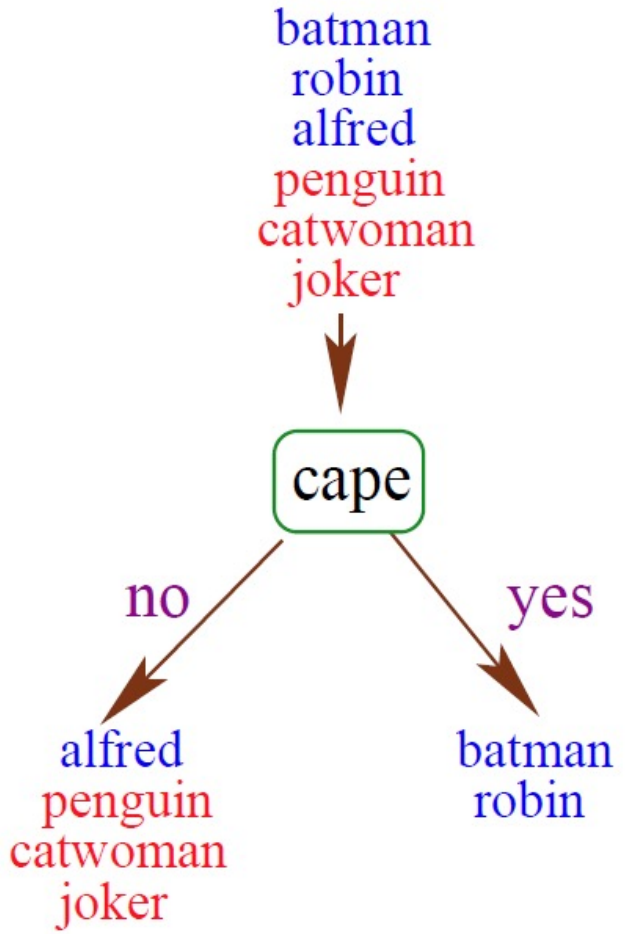


Ein stilisiertes Beispiel: Bestimmen Sie die Gesinnung auf Basis der äußeren Erscheinung

	sex	mask	cape	tie	ears	smokes	class
			training data				
batman	male	yes	yes	no	yes	no	Good
robin	male	yes	yes	no	no	no	Good
alfred	male	no	no	yes	no	no	Good
penguin	male	no	no	yes	no	yes	Bad
catwoman	female	yes	no	no	yes	no	Bad
joker	male	no	no	no	no	no	Bad
			test data				
batgirl	female	yes	yes	no	yes	no	??
riddler	male	yes	no	no	no	no	??

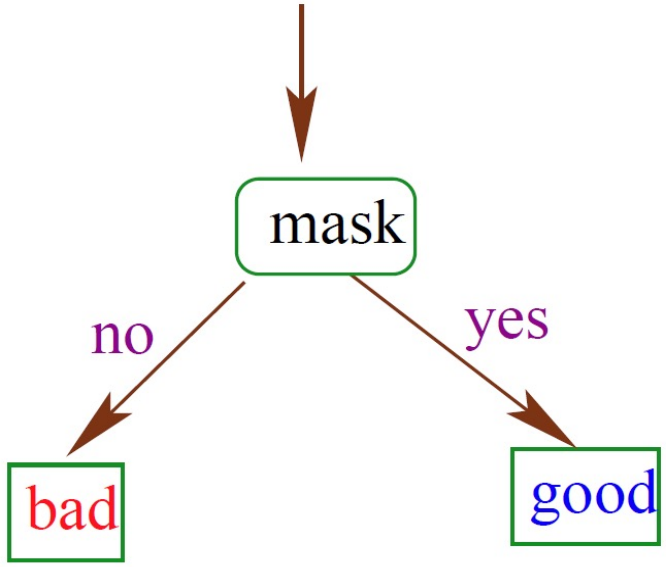
Gute Splits, schlechtes Splits





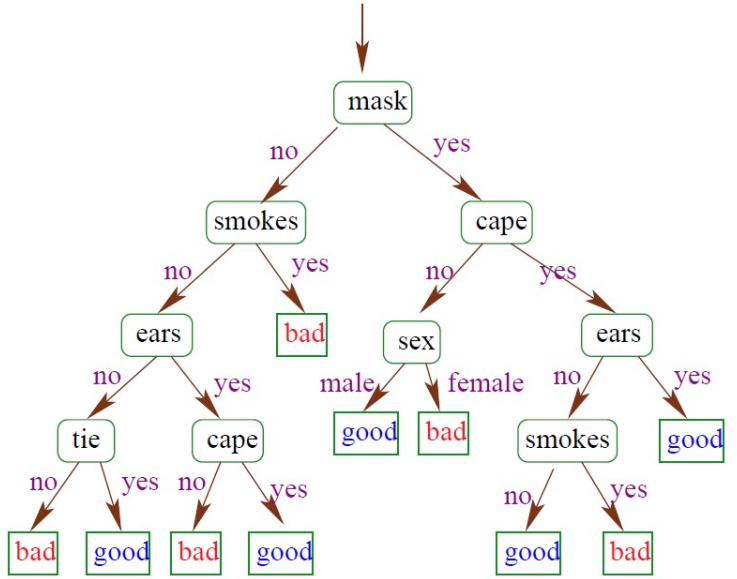
Baumgröße

Unteranpassung



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

Überanpassung



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

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

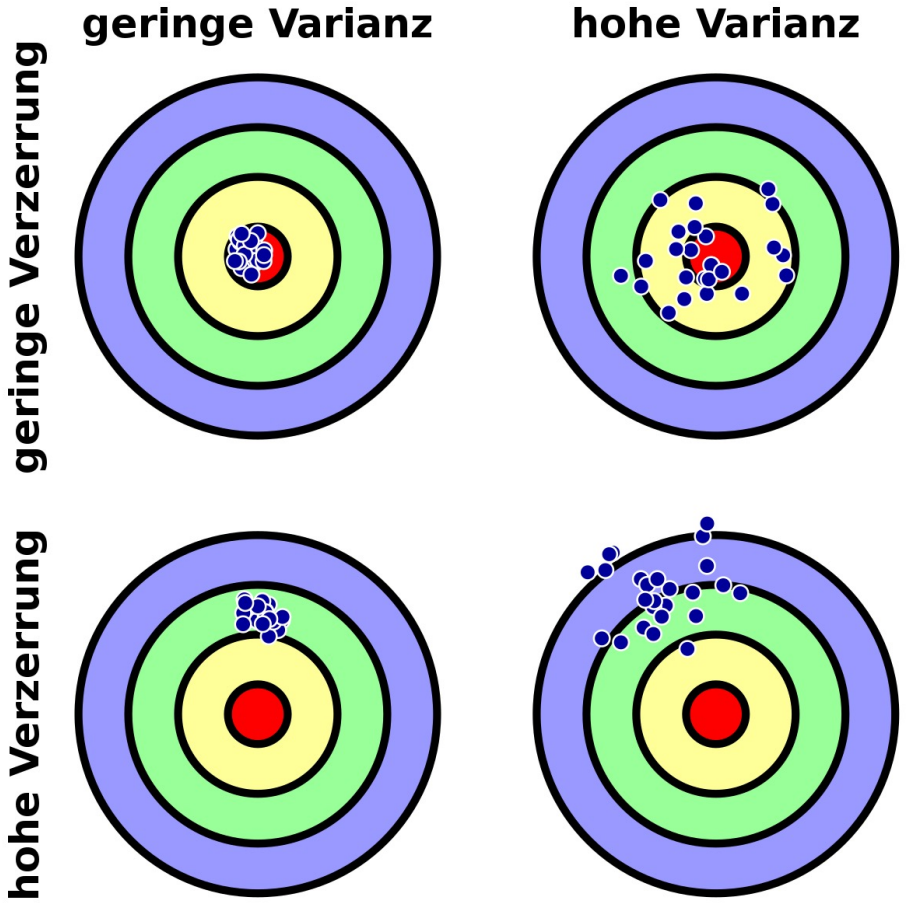
Herausforderung Lernen

- Lernen ist ein unvollständige definiertes Problem: Die Daten sind nicht ausreichend, um eine global gültige Lösung zu bestimmen
- Überwachtes Lernen versucht, Modelle zu identifizieren, die in der Lage sind, **genaue Vorhersagen für ungesehene Daten zu treffen**, die ähnliche Eigenschaften haben wie der Datensatz, der für das Training des Modells verwendet wurde
- Dies wird als Generalisierung bezeichnet und sollte die Modellauswahl leiten
  - Fundamentaler Unterschied zur Ökonometrie!
- Um ein Modell verallgemeinern zu können, muss die Komplexität gegen den beobachteten Fehler abgewogen werden:
- Komplexere Modelle haben einen geringeren beobachteten Fehler bei Trainingsdaten, können aber einen höheren wahren Fehler (bei unbekannten Daten) haben
  - Diese Beobachtung fußt auf dem Bias-Varianz-Trade-off

Verzerrung-Varianz-Dilemma

- Es gibt allgemein zwei Fehlerquellen für Vorhersagen eines statistischen Modells:
  - Verzerrung (Bias): systematische Fehler in den Vorhersagen über alle Instanzen – das Modell liegt daneben
  - Varianz: Fehler durch Schwankung in der Vorhersagegüte über die verschiedenen Instanzen – das Modell ist manchmal besser, manchmal schlechter

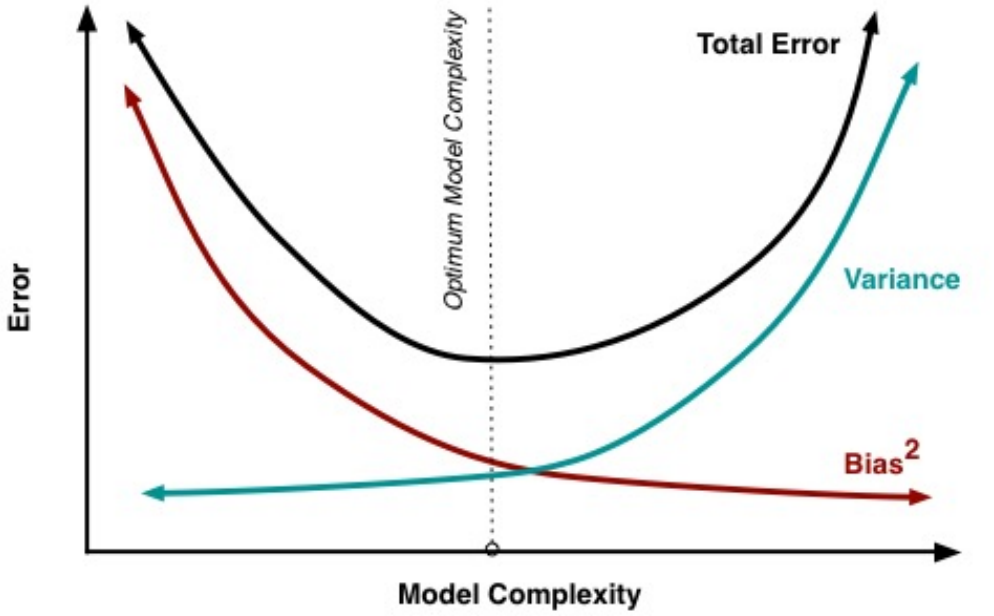
- Gemeinhin laufen die beiden Ziele entgegen:
  - Komplexere Modelle machen geringere systematische Fehler aber Fehler streuen stärker über die Instanzen
  - Einfache Modelle machen große systematische die homogen über die Instanzen sind



Verzerrung-Varianz-Tradeoff

Unteranpassung
Underfitting

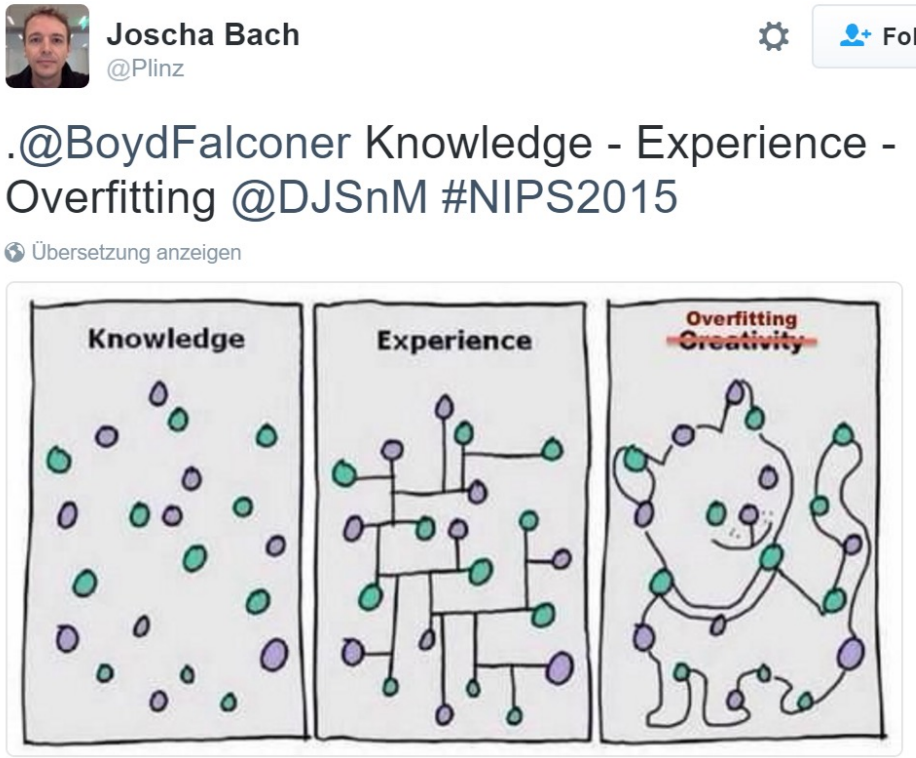
Überanpassung
Overfitting

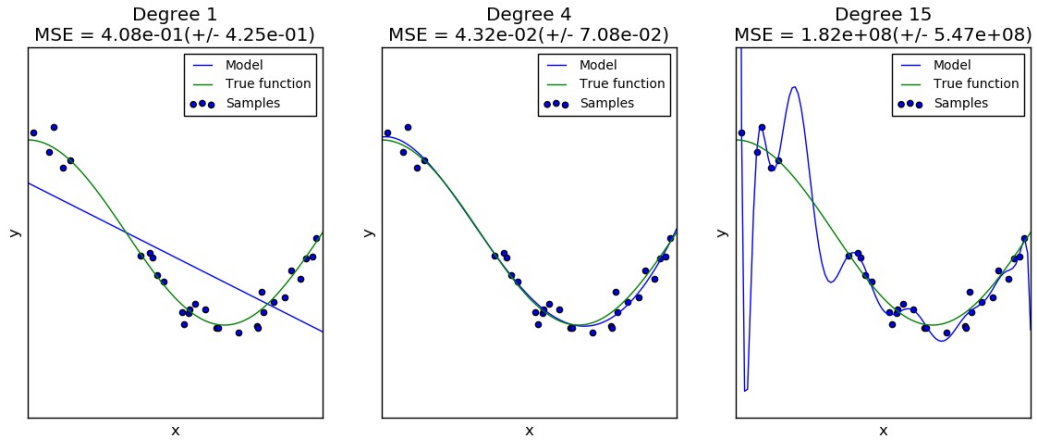


Strategien zur Vermeidung von Überanpassung

- Testdatenmanagement
- Regularisierung
- Ensemblemethoden

Das grundlegende Problem





Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

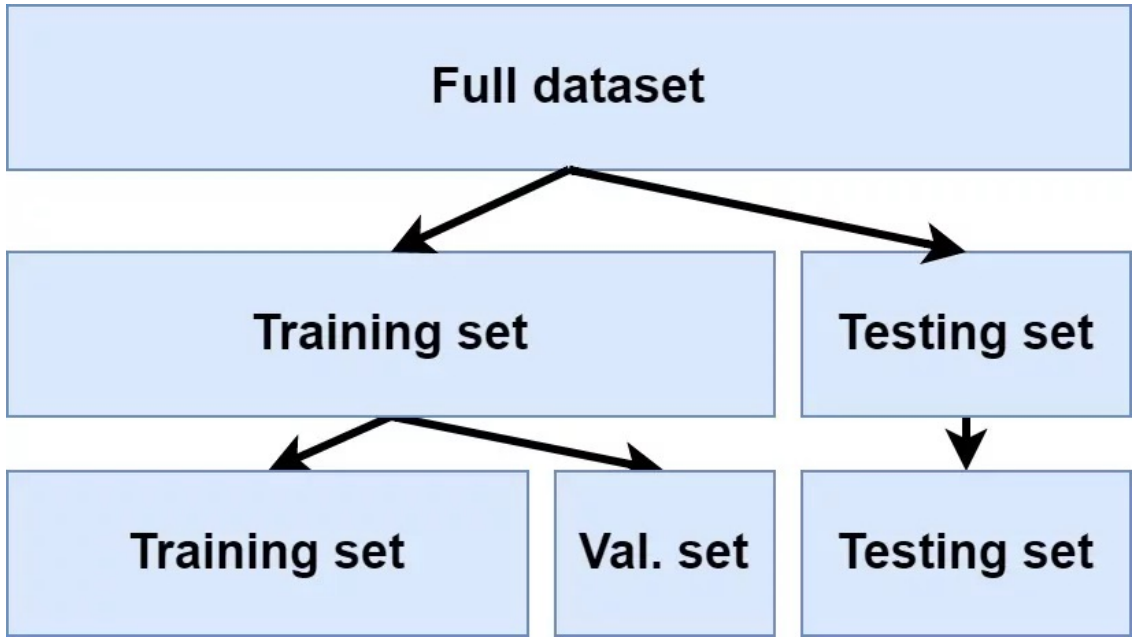
Agenda

Nichtnutzung von Daten erlaubt eine neue Form der Evaluation

- Anstatt die kompletten Daten zu erklären zu wollen können wir auf Teildaten ein Modell lernen und dann auf den restlichen Daten objektiv dessen Güte bewerten: Train-Test Split

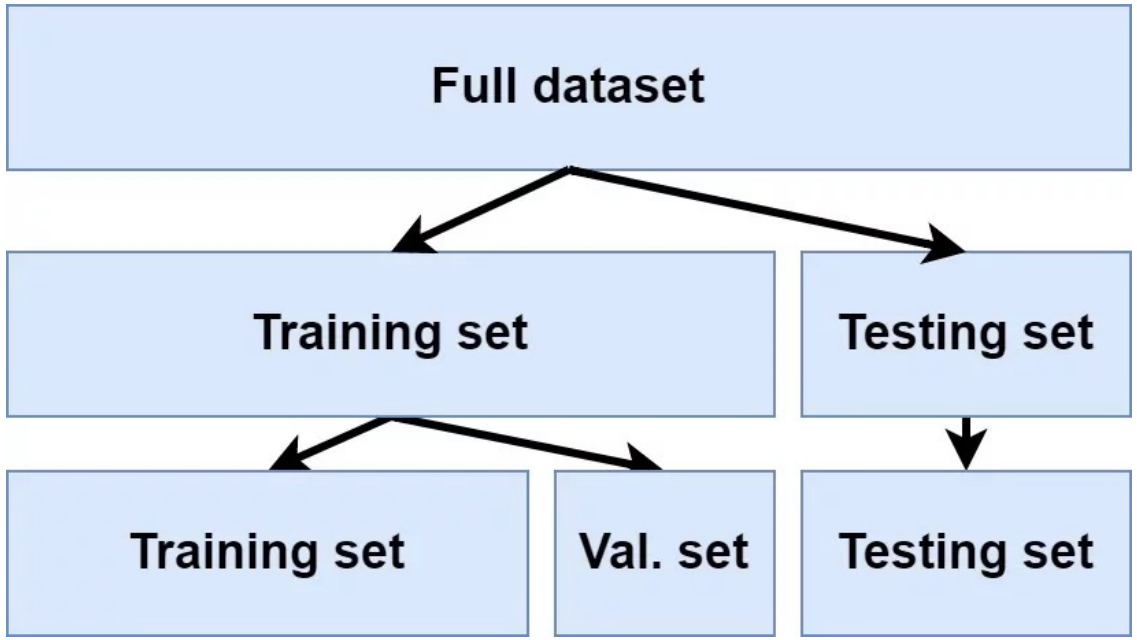
- Die Trainingsdaten werden wiederum in Training und Validation geteilt um eine Bewertung von Konfigurationsvarianten (andere Algorithmen, andere Parameter) zu ermöglichen

- Das Test Set dient nur der Bewertung der Generalisierungsfähigkeiten – es wird nicht für das Training benutzt
  - Insbesondere dürfen auf Basis der Testing Ergebnisse KEINE Anpassungen an den Algorihmen erfolgen



Lohnt sich das Vorgehen?

- Was passiert mit Verzerrung?
- Was passiert mit Varianz?
- Verschwenden wir Daten?



Verzerrung und Varianz

- Verzerrung

- Varianz

Verzerrung und Varianz

- Verzerrung geht hoch:
  - Verfügbare Daten werden nicht komplett für Modelltraining genutzt
  - Hieraus folgt direkt eine geringere Vorhersagegüte des Modelles auf den verfügbaren Daten
- Varianz

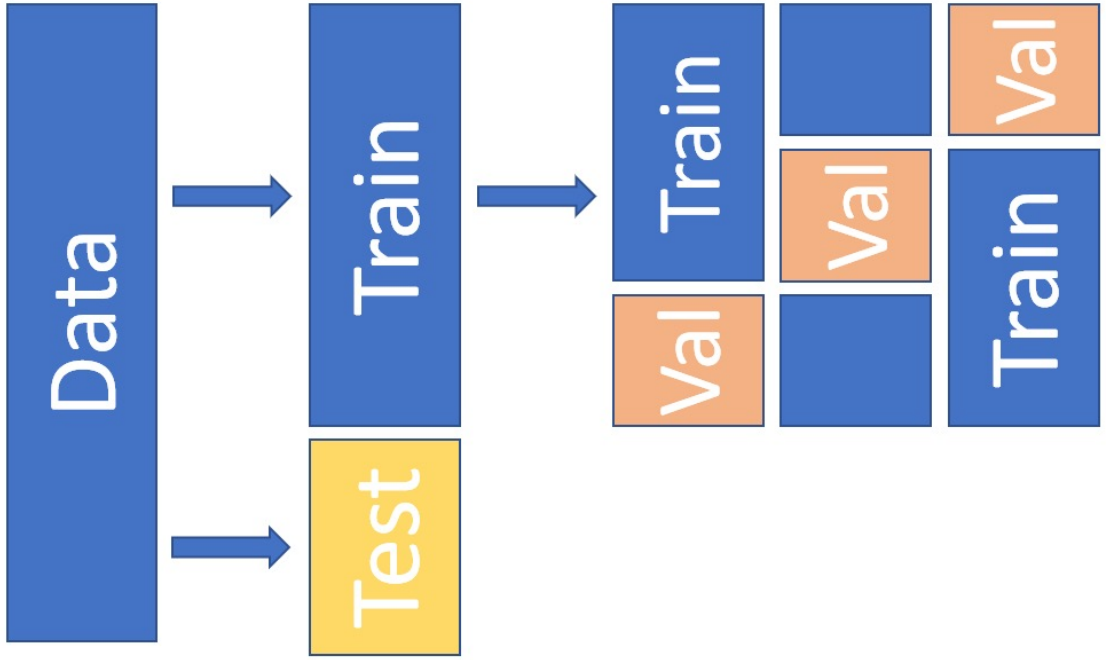
Verzerrung und Varianz

- Verzerrung geht hoch:
  - Verfügbare Daten werden nicht komplett für Modelltraining genutzt
  - Hieraus folgt direkt eine geringere Vorhersagegüte des Modelles auf den verfügbaren Daten
- Varianz geht runter:
  - Vorhersagegüte wird auf Basis der Ergebnisse aus den Testdaten bewertet
  - Nur Muster und Strukturen die generalisierbar sollen genutzt werden

Verschwenden wir Daten?

- Gewissermaßen ja – insbesondere werden die Validationdaten im skizzierten Protokoll nicht im Training berücksichtigt, nur in der Bewertung

- Alternative: Kreuzvalidierung
  - Training wird mit verschiedenen Splits wiederholt und Ergebnis gemittelt



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

Regularisierung

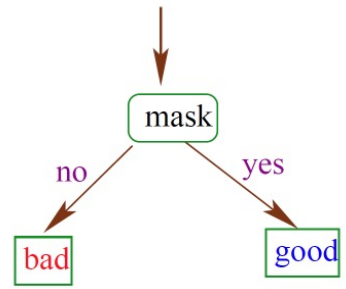
- Regularisierung ist ein Konzept, mit dem verhindert werden soll, dass ML-Algorithmen das trainierte Modell einem Datensatz überanpassen
- Die Regularisierung erreicht dies indem in die Kostenfunktion ein Bestrafungsterm eingeführt wird, der komplexen Hypothesen eine höhere Strafe zuweist

Baumgröße als zentraler Regularisierungsparameter bei Entscheidungsbäume

- Baumgröße kann durch Tiefe oder Anzahl Blätter gemessen werden

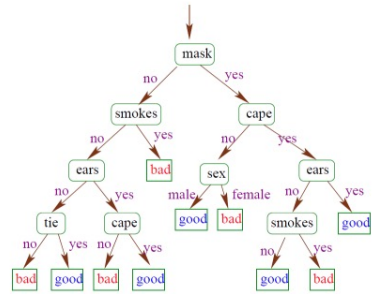
- Unterschiedliches Vorgehen beim Regularisieren möglich:
  - Pre-Pruning: Baumwachstum wird frühzeitig beendet (early stopping) wenn weitere Schritte „zu wenig bringen“
  - Post-Pruning: Baum wird voll ausgewachsen und danach „zurückgestutzt“ auf Basis von Effizienzbewertungen

Unteranpassung



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

Überanpassung



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

Regularisierung in der linearen Regression

Regularisierung in der linearen Regression

- Anzahl Variablen

- Größe der Koeffizienten

Regularisierung in der linearen Regression

- Anzahl Variablen
  - Feature Selection
  - Adjusted R^2
  - Bootstrapping über Teilmengen der Features

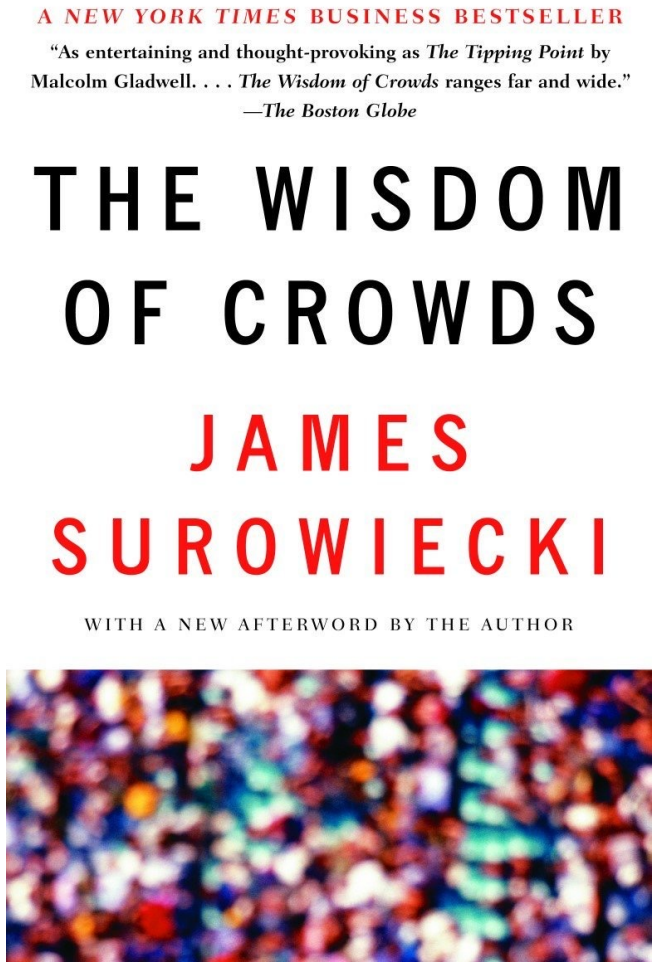
- Koeffizientengröße
  - Begrenzen der Summe der Koeffizienten begrenzt die Anpassungsfähigkeit des Modelles
  - Typische Ansätze Summe der Quadrate (Ridge-Regression) oder Summe der Beträge (Lasso-Regression)

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

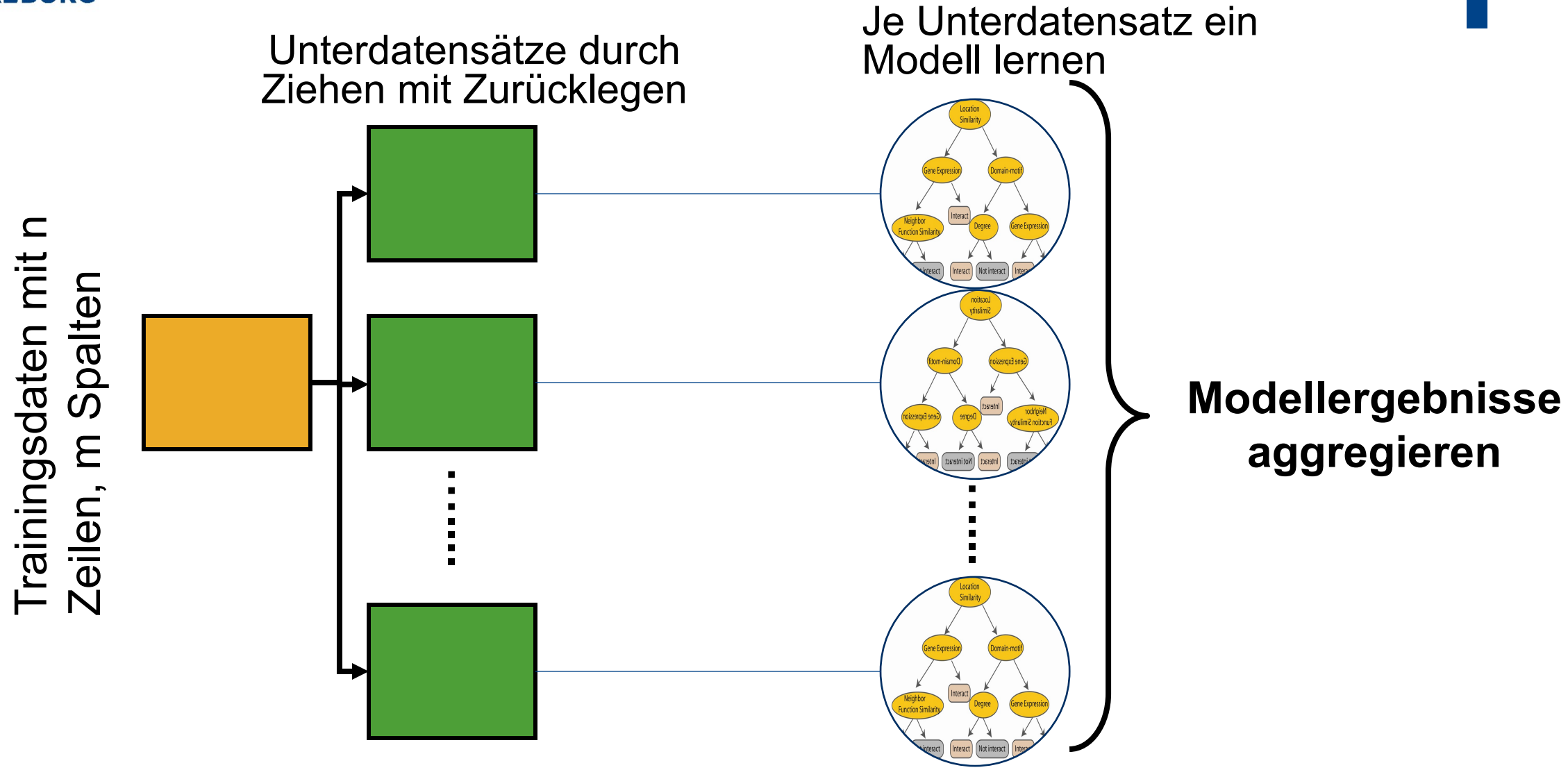
Agenda

Lern-Ensembles

- Idee
  - Lernen Sie mehrere alternative Modelle mit unterschiedlichen Trainingsdaten oder unterschiedlichen Lernalgorithmen.
  - Kombinieren Sie Modellvorhersagen, z. B. mit Hilfe einer (gewichteter Abstimmung)
- Solche Ensembles sind oft wertvoll
  - Beim Kombinieren mehrerer unabhängiger und unterschiedlicher Entscheidungen, von denen jede mindestens genauer ist als zufälliges Raten
  - Zufällige Fehler heben sich gegenseitig auf, richtige Entscheidungen werden verstärkt
  - Menschliche Ensembles sind sehr effektiv (Jelly Bean Jar, WWM Publikums-Joker)



Bagging (Bootstrap Aggregation)



Boosting

- Wurde von Theoretikern des computergestützten Lernens entwickelt, um Leistungsverbesserungen bei passenden Trainingsdaten für einen schwachen Lerner zu garantieren (Schapire, 1990)
- Schwacher Lerner: erzeugt eine Vorhersage mit einer Trainingsgenauigkeit größer als 0,5 (etwas besser als Raten)
- Verfahren zum Aufbau von Ensembles, die empirisch die Verzerrung reduzieren

- Verfahren
- Ausgangspunkt: initialer schwacher Lerner wird $\mathcal{L}_0\ (x)$ angepasst, um y vorherzusagen
- Iterative Erweiterung: Trainiere schwachen Lerner $\mathcal{L}_i\ (x)$ so, dass $\mathcal{L}_{i-1}\ (x)$+$\nu\mathcal{L}_i\ (x)$ y besser vorhersagen kann.
- Trainieren auf $y - \mathcal{L}_{i-1}\ (x)$

- Wichtiges Element: Lernrate ν<1, die als Regularisierungsparameter dient

Bagging vs Boosting

Bagging

- Paralleles Ensemble: jedes Modell wird unabhängig erstellt
- Ziel ist die Verringerung der Varianz, nicht des Bias
- Geeignet für Modelle mit hoher Varianz und geringer Verzerrung (komplexe Modelle)
- Ein Beispiel für eine baumbasierte Methode sind Random Forests, die individuelle Entscheidungsbäume kombinieren

Boosting

- Sequentielles Ensemble: Neue Modelle werde so ergänzt, dass sie die Leistung vorheriger verbessert
- Ziel ist es primär den Bias zu verringern
- Geeignet für Modelle mit niedriger Varianz und hoher Verzerrung