Einführung in Web- und Data-Science

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

Tanya Braun (Übungen)

Teilnehmerkreis und Voraussetzungen

Studiengänge

- Bachelor **Informatik**
  - Pflicht in der kanonischen Vertiefung Web and Data Science

Voraussetzungen

- Keine

Organisatorisches: Übungen

- **Vorlesung**: Donnerstags, ab dem 17. Oktober 2019
- **Übungen**: Mittwochs 13-14 Uhr, ab 23.10.
  Anmeldung über Moodle nach dieser Veranstaltung
- **Übungsaufgaben** stehen jeweils nach der Vorlesung ca. ab 18 Uhr über Moodle bereit (erstes Übungsblatt erscheint am 17.10.2019)
- Aufgaben sollen in einer **2-er Gruppe** bearbeitet werden (also bitte Name(n) und Übungsgruppennummer vermerken)
- **Abgabe der Lösungen** erfolgt bis Montag 12 Uhr in der IFIS-Teeküche (jeweils in der zweiten Woche nach der Ausgabe)
- In den Übungen am Mittwoch wird der Übungszettel besprochen, dessen Lösungen bis zum jeweils vorigen Montag abgegeben werden, und auch Fragen zum jeweils neuen Übungszettel geklärt

Organisatorisches: Prüfung

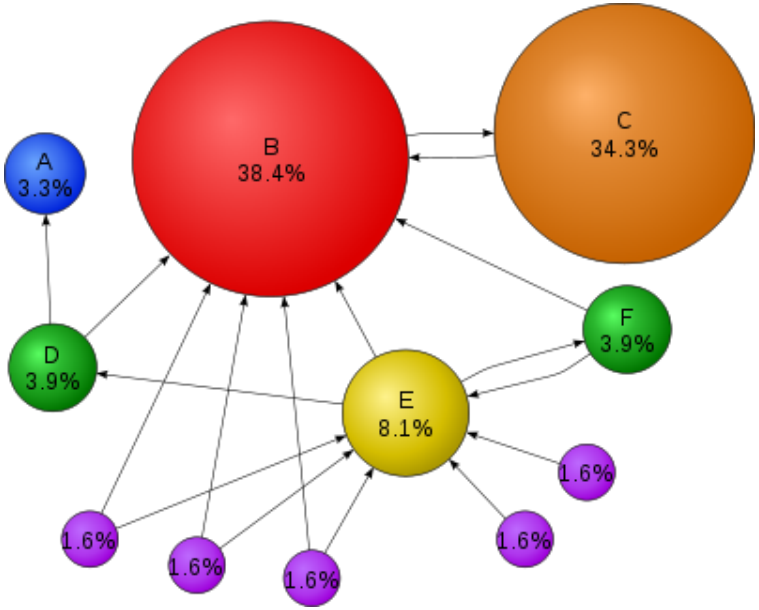
- Die **Eintragung in den Kurs** und in eine Übungsgruppe ist **Voraussetzung**, um an dem Modul Einführung in Web und Data Science teilnehmen zu können
- Am Ende des Semesters findet eine **Klausur** statt
  - Es geht nur um bestehen
- **Voraussetzung** zur Teilnahme an der Klausur sind mindestens **50% der gesamtmöglichen Punkte aller Übungszettel (240 Punkte insgesamt)**

Einführung in Web und Data Science

Begriffsbestimmung

Web und Data Science

- Web Science
  - Analyse von Strukturen im Web (Mensch und Computer)
  - Formalisierung durch große Graphstrukturen und entsprechende Entscheidungsprobleme über Graphen
  - Beispiel: Pagerank (Bewertung von Webseiten)



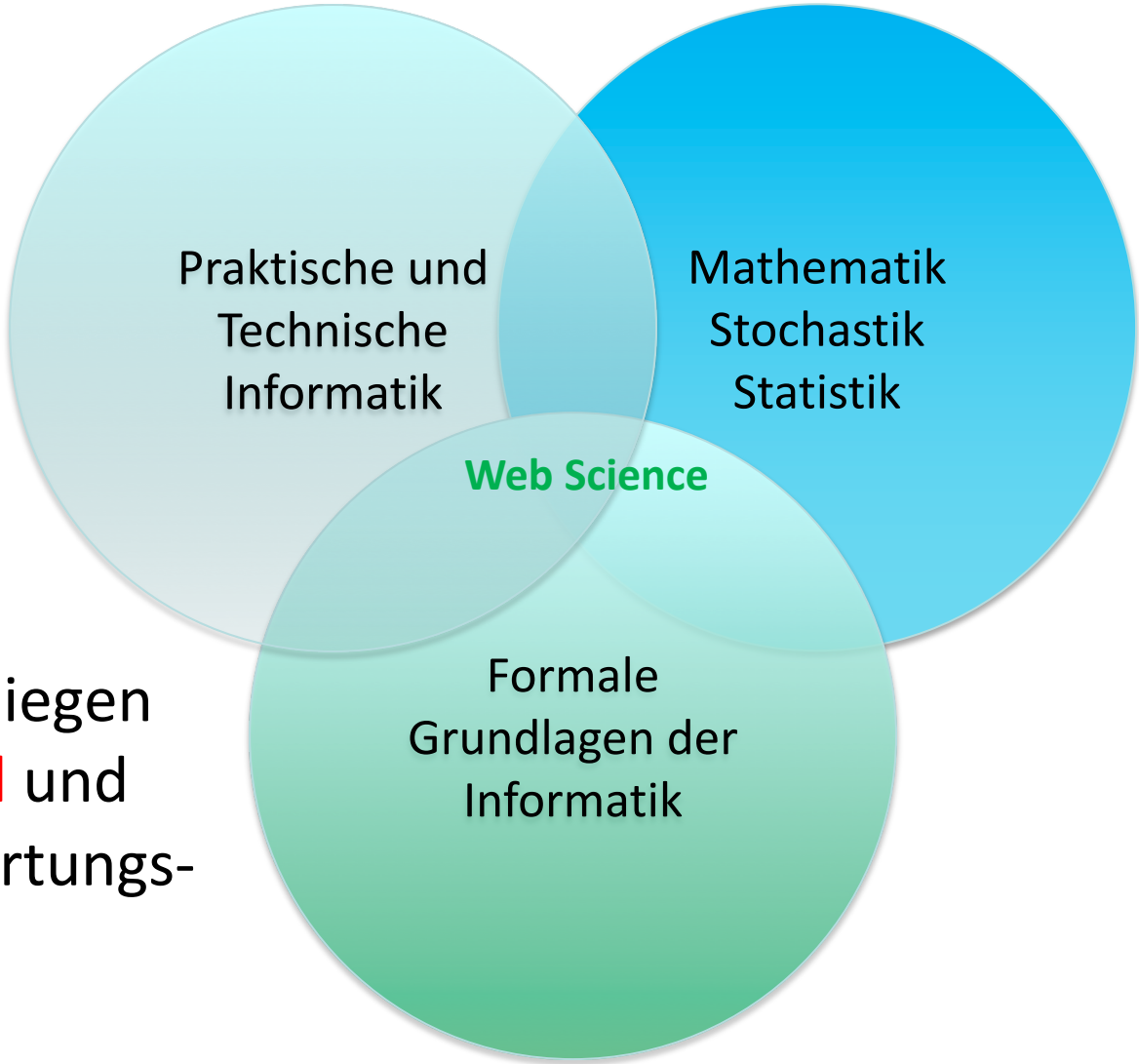
Zufallssurfer-Modell:

Größe der Kreise in etwa proportional der relativen Häufigkeit, mit der sich ein Surfer auf einer Seite befindet

[Wikipedia]

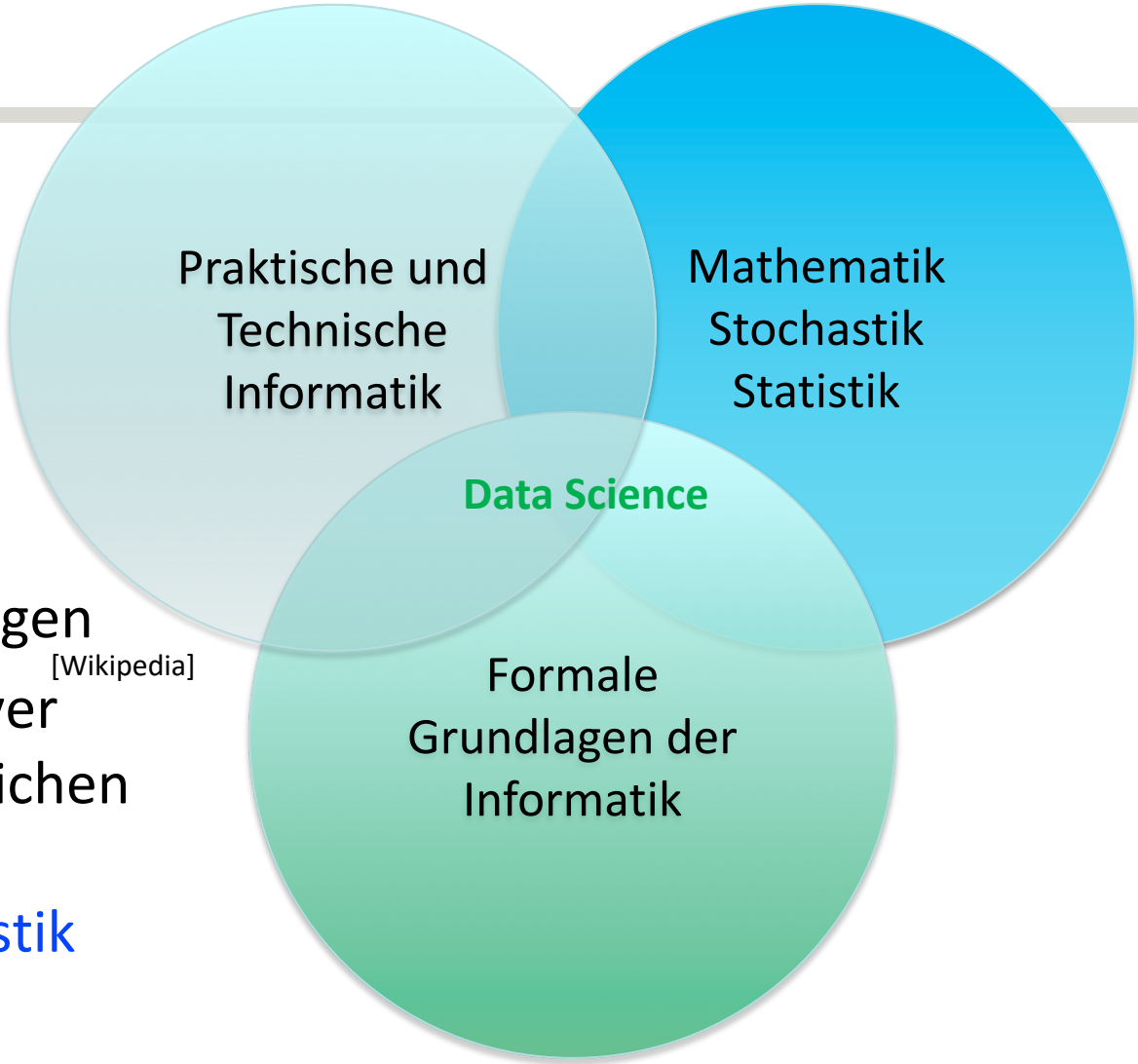
Herausforderungen für die Informatik

- Graphstrukturen extrem groß
- Verfahren zur Lösung von Entscheidungs-problemen extrem aufwendig
- Graphdaten unterliegen ständigem Wandel und so auch die Auswertungs-ergebnisse



Data Science

- Extraktion von Wissen aus Daten (u.a. Graphdaten)
- Begriff schon vor 50 Jahren für Informatik vorgeschlagen [Wikipedia]
- Entwicklung innovativer Konzepte in den Bereichen Logik, Datenbanken und Stochastik / Statistik (Datenanalyse und Wissensentdeckung)
- Verwendung von LADS und Analysis



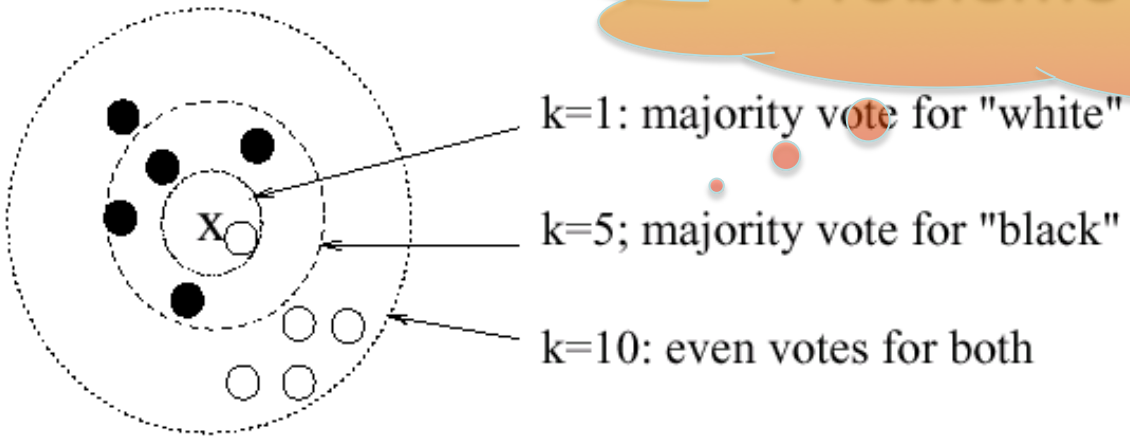
Herausforderungen für die Informatik

- Große Datenbestände
  - Speicher und Zugriffstechnologie
- Starker Zuwachs an Daten, hohe Dynamik
  - Hohe Datenraten und Echtzeitanforderungen
- Heterogene Datenbestände
  - Verteiltes Datenmanagement
  - Datenintegration
- Robuste Modelle der menschlichen Interpretationsvorgänge (Kognition) notwendig für Auswertung und Präsentation von Daten
  - Verarbeitung von Text-, Sprach- und Videodaten
- …

Speicheranforderungen für Anwendungen

- Annahme: Gegeben viele Datenpunkte
  - Beispielmerkmale: (x, y, Farbe), Farbe ∈ { white, black }
- Anfrage: Datenpunkt ohne Wert für bestimmtes Merkmal
  - Beispiel: Merkmal Farbe ohne Wert
- Anfragebeantwortung (Klassifikation des Anfragepunkts): Mehrheitsvotum der k-nächsten Nachbarn (kNN-Verfahren)

Probleme erkennbar?



Fix, E., Hodges, J.L. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951

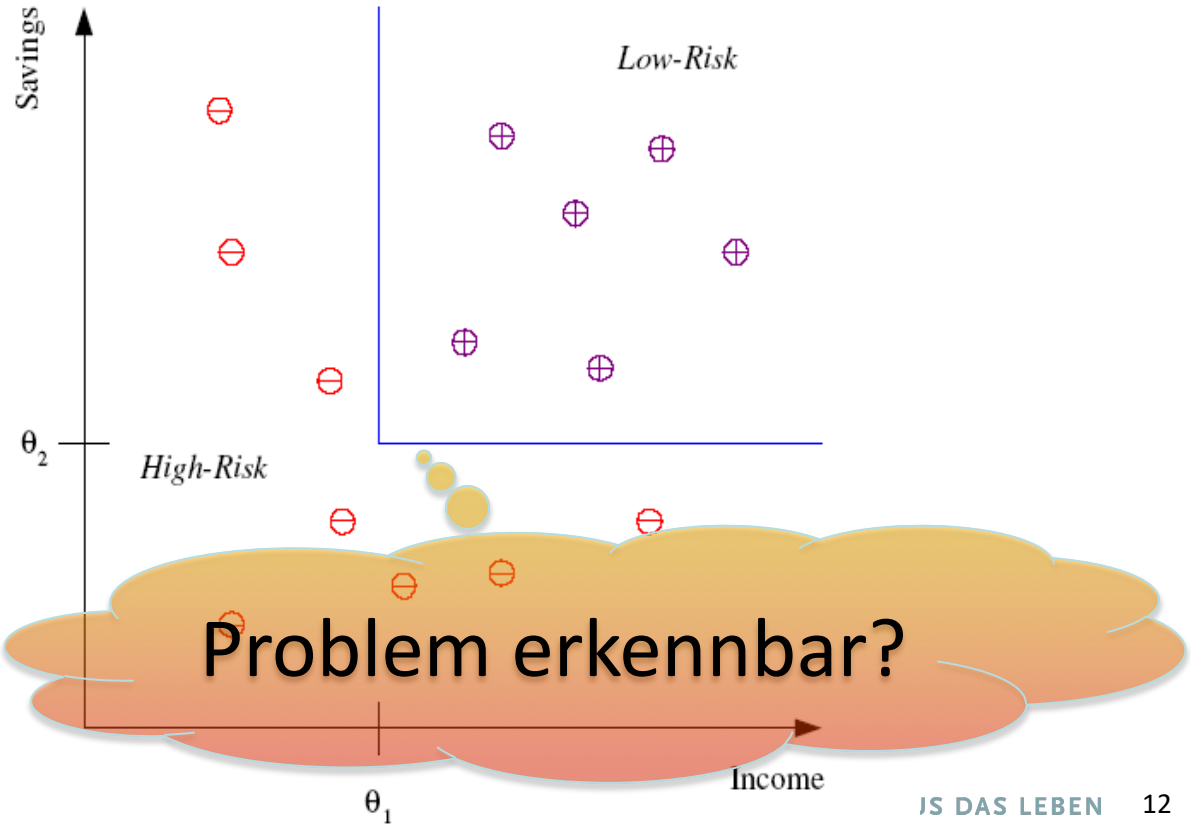
Probleme mit kNN

- Klassifikationsergebnis stark von k abhängig
- Hoher Speicherbedarf
- Effizienter Zugriff auf "Nachbarn" erfordert weitere Maßnahmen (noch mehr Speicherbedarf)

- Klassifikation basierend auf den Daten

Generalisierung von Daten möglich?

- Repräsentation der Daten durch Parameter
  - Wenn (Einkommen > θ_1 ∧ Ersparnisse > θ_2), dann kreditwürdig (⊕), sonst nicht (⊖)
- Nur 2 Parameter nötig: (θ_1, θ_2)
- Modell fordert geringen Speicher



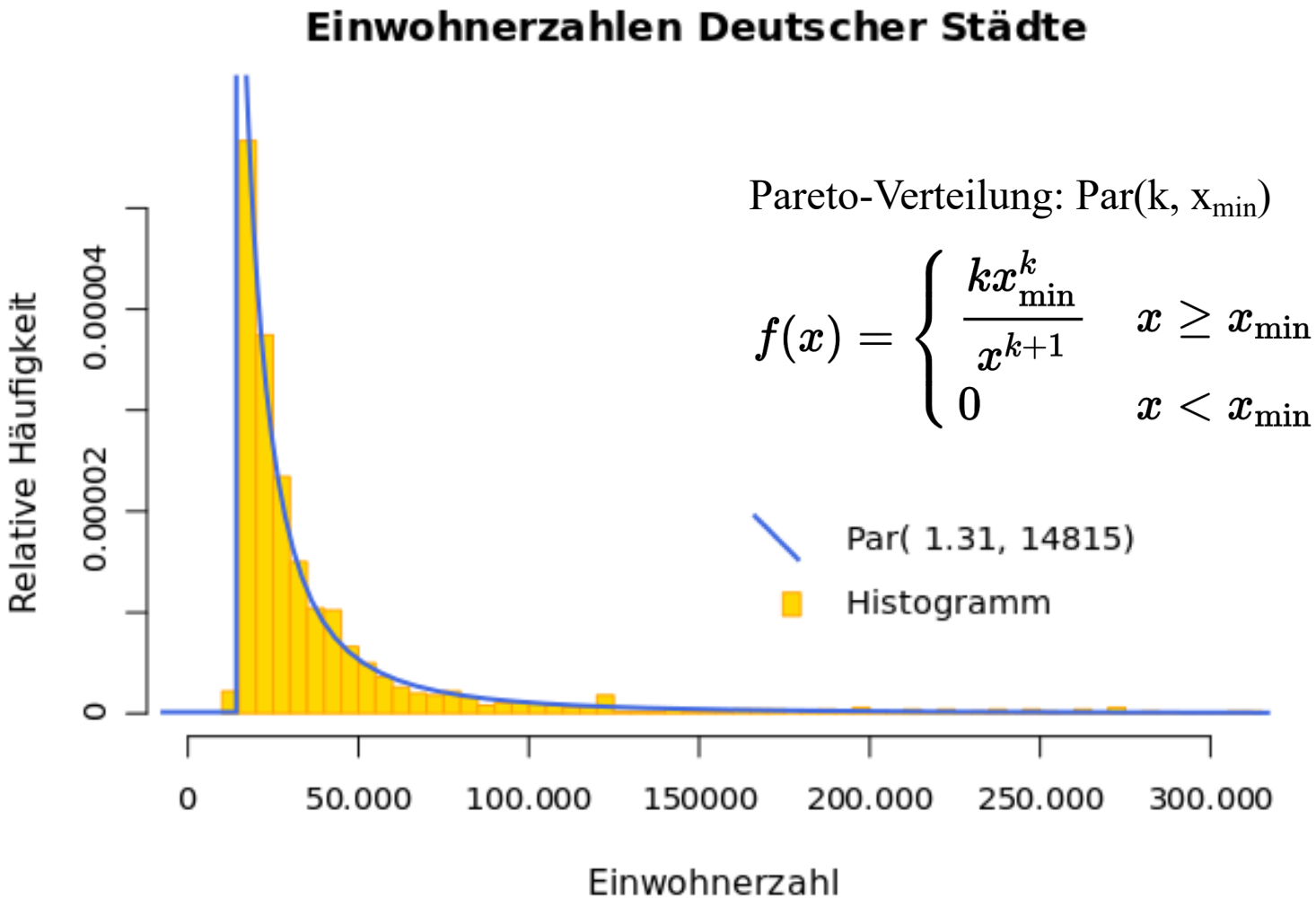
Beispiel

- Anzahl von Städten mit bestimmten Einwohnerzahlen schätzen

- Daten: Liste von Einwohnerzahlen (einige kommen mehrfach vor)

- Explizites Modell: Zählerfeld aufbauen
  - Zählerfeld ggf. sehr groß
- Implizites Modell: Potenzgesetz $y=ax^b$
  - a und b bestimmen
  - Aufwendiges Optimierungsproblem lösen

Begriff der "Verteilung"



Pareto-Verteilung: Par(k, x_{min})

$$f(x) = \begin{cases} \dfrac{k x_{\min}^{k}}{x^{k+1}} & x \geq x_{\min} \\ 0 & x < x_{\min} \end{cases}$$

[Wikipedia]

Über den Kurs

- Repräsentationssprachen
  - Modellierungsansatz
  - Entscheidungs- und Berechnungsprobleme
  - Verfahren zu deren Lösung
- Gewinnung von Modellen aus Daten als Optimierungsproblem (Algorithm. Datenanalyse, Data Mining, Maschinelles Lernen)
  - Überwachtes Lernen,
  - Unüberwachtes Lernen
    - Deep Learning, Transduktives Lernen, Reinforcement-Lernen
  - Transfer-Lernen
- Verwendung von Modellen in Anwendungen
  - Prädiktion, Forensik, Verstehen der Realität (Science-Aspekt!)
  - Autonome Aktivitäten/ Agenten

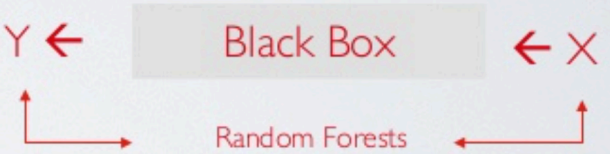
Data Models vs. Algorithmic Models

Data Modeling

vs.

Algorithmic Modeling

$Y \leftarrow F(X, \text{random noise, parameters})$



We understand the world

How well 'my data model' works
Statisticians, Data Analysts, Data Miners
Linear Regression
Logistic Regression
Known Distributions
Confidence Intervals
Predictor Variables & Goodness of Fit

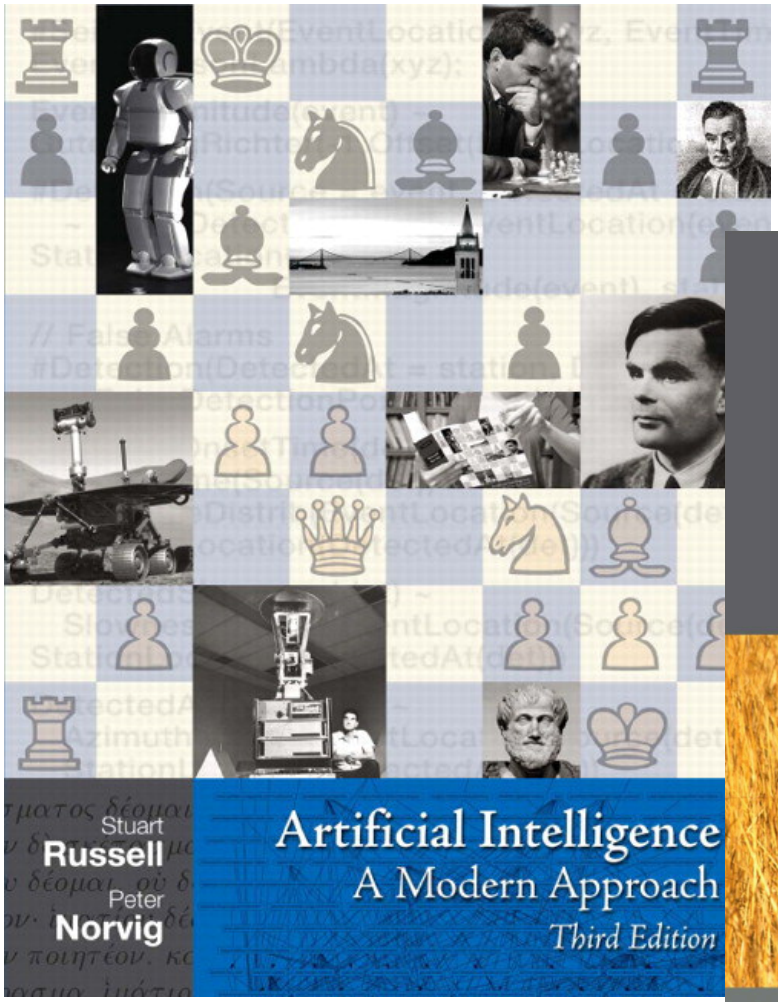
We don't understand the world

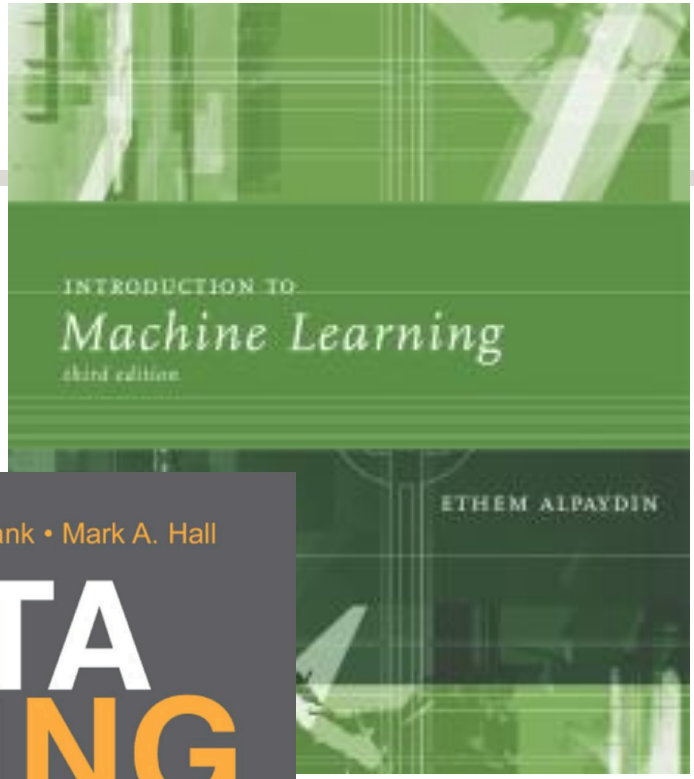
The world produces data in a black-box
Data Scientists
Machine Learning, AI & Neural Nets
Random Forests, SVM, GBT
Unknown Multivariate Distributions
Iterative
Predictive Accuracy

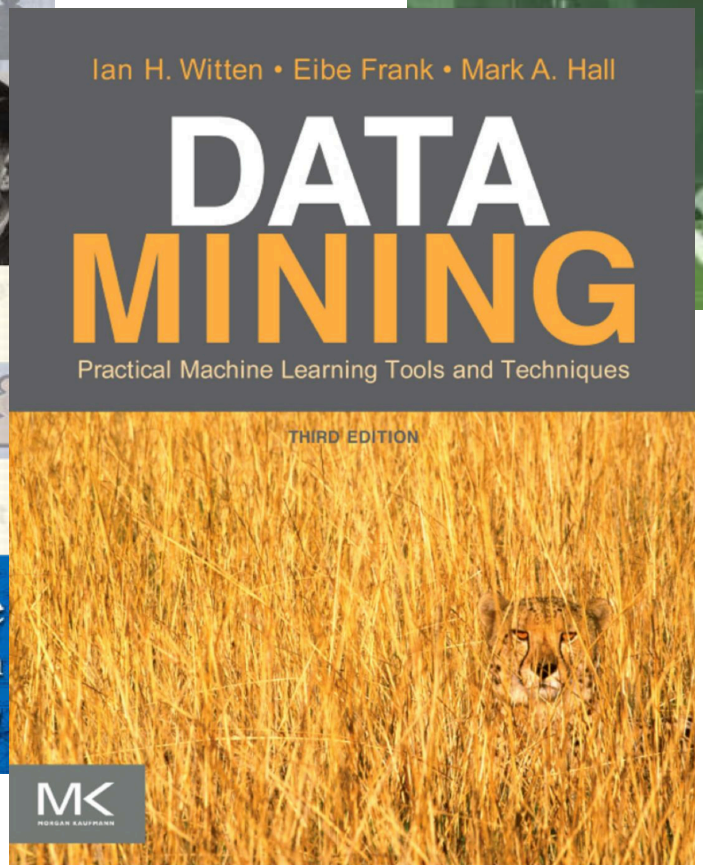
"Statistical Modeling: The Two Cultures" Leo Breiman, 2001

Data Science London

Literatur

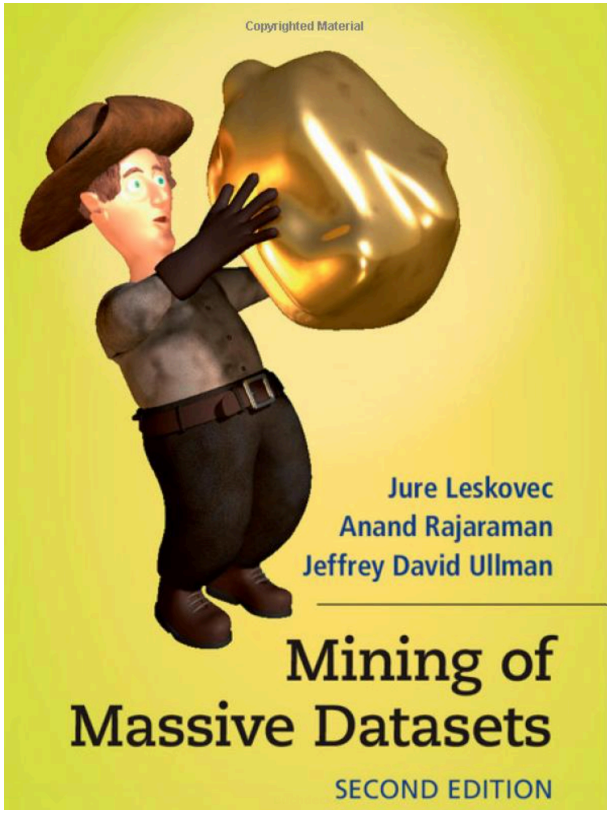

Stuart
Russell
Peter
Norvig
Artificial Intelligence
A Modern Approach
Third Edition


INTRODUCTION TO
Machine Learning
third edition
ETHEM ALPAYDIN


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

Literatur

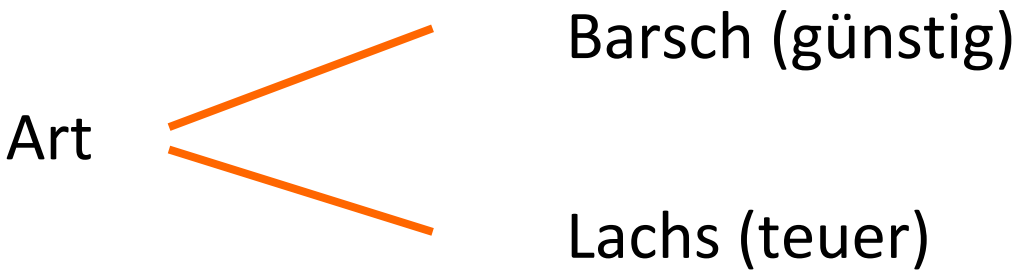
- Stuart Russell, Peter Norvig, **Artificial Intelligence – A Modern Approach**, Pearson, 2009 (oder 2003er Ed.)
- Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: **Practical Machine Learning Tools and Techniques**, Morgan Kaufmann, 2011
- Ethem Alpaydin, **Introduction to Machine Learning**, 3rd Ed., MIT Press, 2014
- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, **Mining of Massive Datasets**, 2nd Ed., Cambridge University Press, 2014



- Viele zusätzliche Bücher, Präsentationen, und Videos im Web

Ein erweitertes Beispiel

“Sortierung von Fischen auf einem Förderband nach Arten durch Bildverarbeitung”



Problemanalyse

Verwende Kamera und nehme Bilder auf,
um Merkmale zu bestimmen:

- Länge
- Helligkeit
- Breite
- Anzahl und Form der Flossen
- Position des Mundes usw.

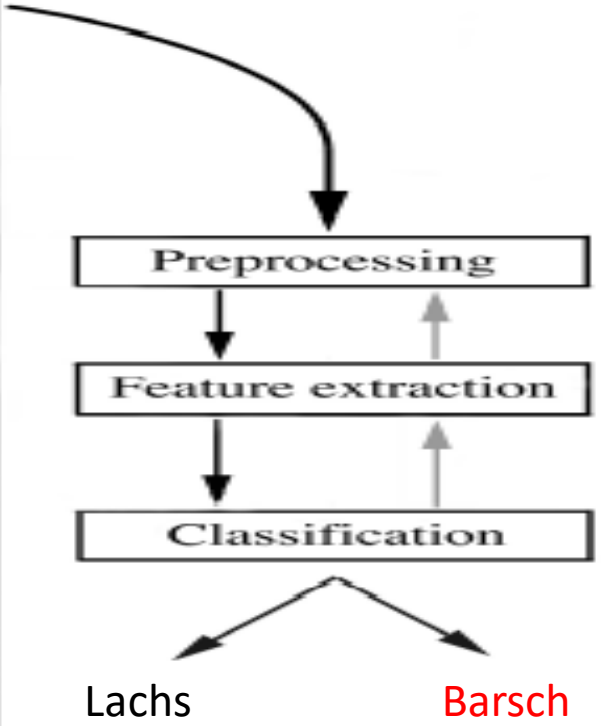
Menge aller möglichen Merkmale

Ziel: Wähle die relevanten aus

Vorverarbeitung

- Verwende Segmentierungsoperator, um Fisch und Hintergrund zu unterscheiden
- Fischregion wird verwendet, um Merkmalswerte zu extrahieren (geeignete Merkmale vorher festgelegt)
- Merkmalswerte weiterverarbeitet durch Klassifikator (Barsch oder Lachs)

Klassifikation



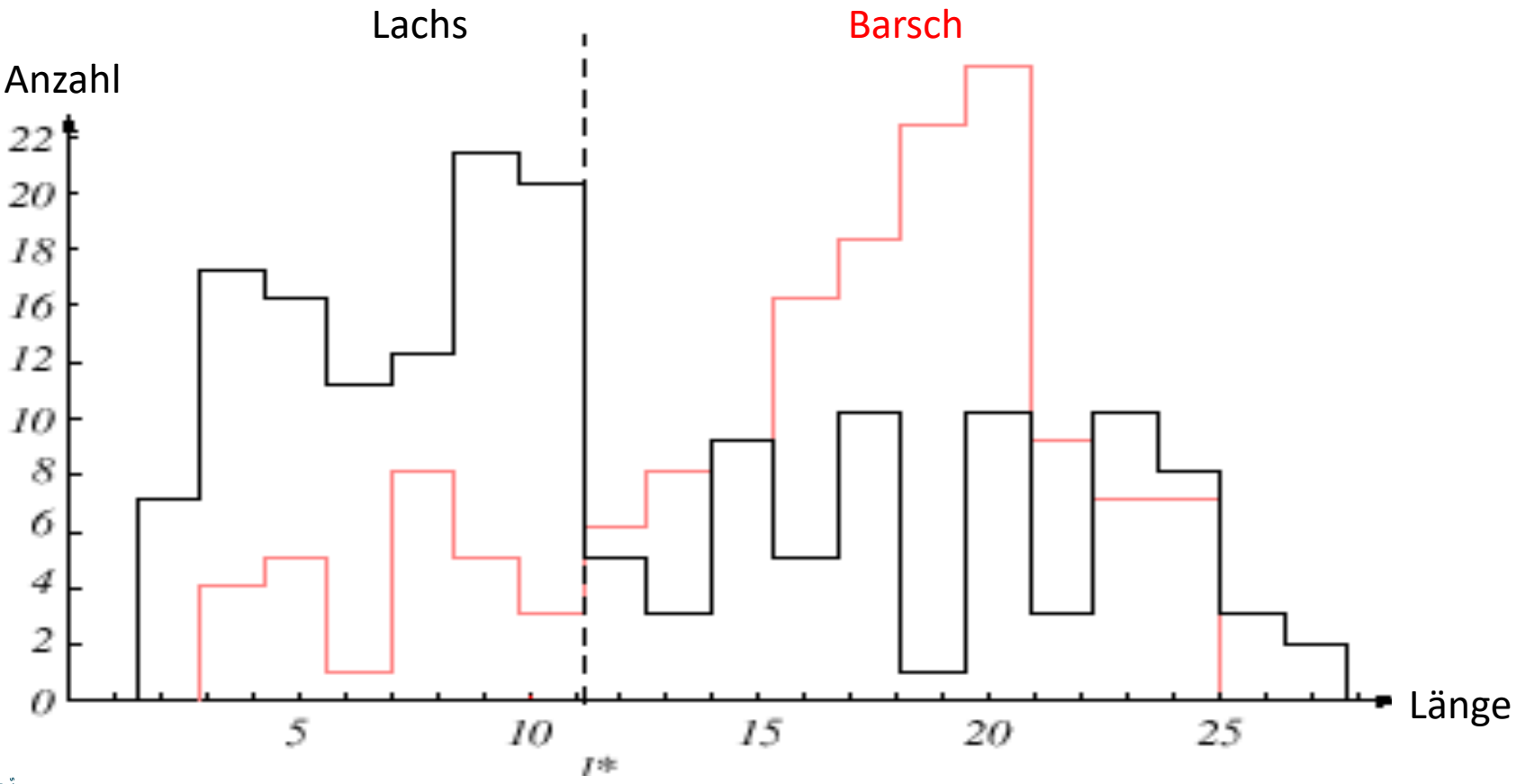
Bestimmung geeigneter Merkmale

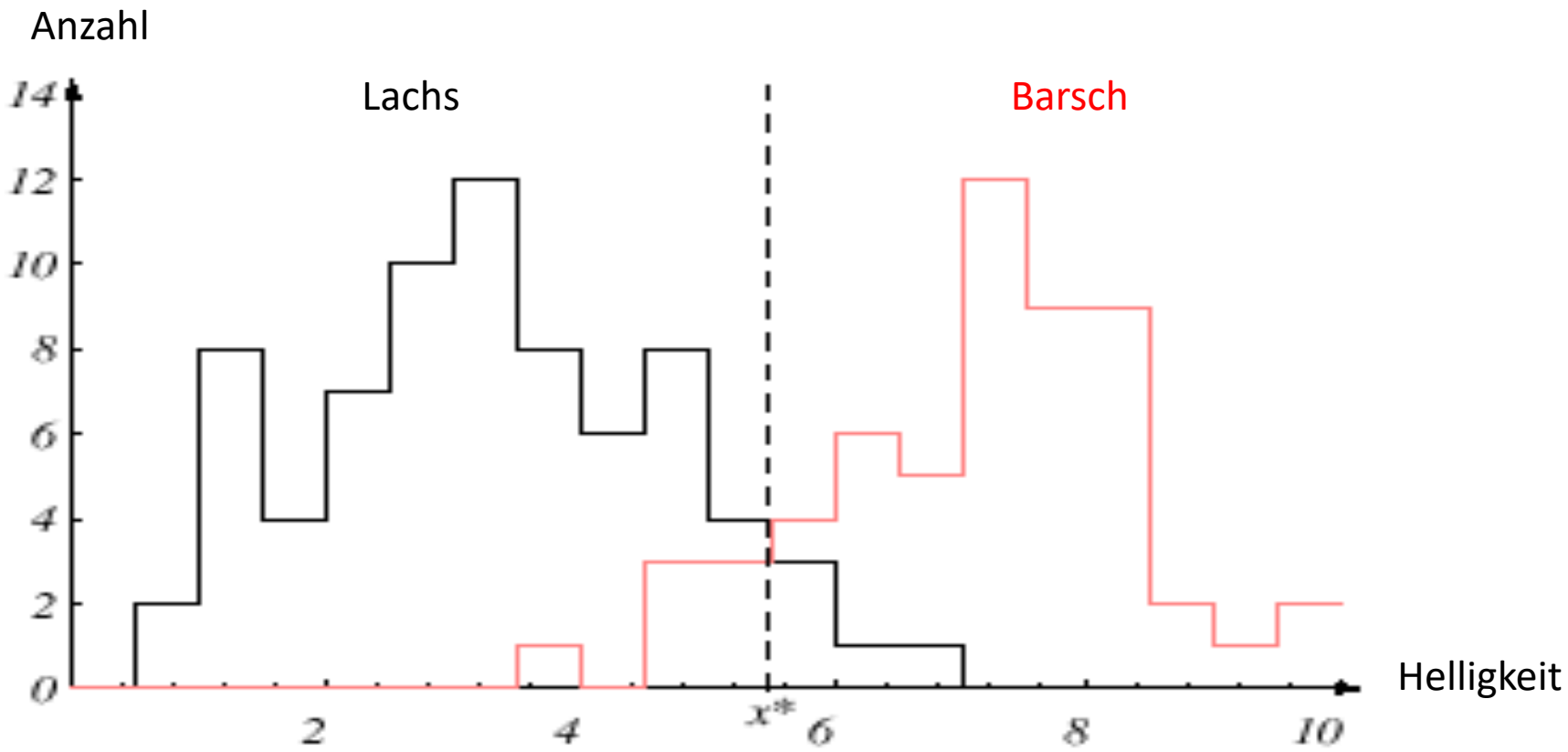
- Wir benötigen einen Experten, um die Merkmale festzulegen, mit denen man Barsche und Lachse richtig klassifizieren kann
- Wie wäre es mit Länge als Merkmal zur Unterscheidung?

Länge allein ist kein gutes Merkmal!

→ Hohe Kosten bei Fehlentscheidung

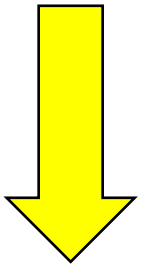
Wie wäre es mit Helligkeit?




Anzahl
Lachs
Barsch
14
12
10
8
6
4
2
0
2
4
x*
6
8
10
Helligkeit

Schwellwert-Entscheidungsgrenze und induzierte Kosten

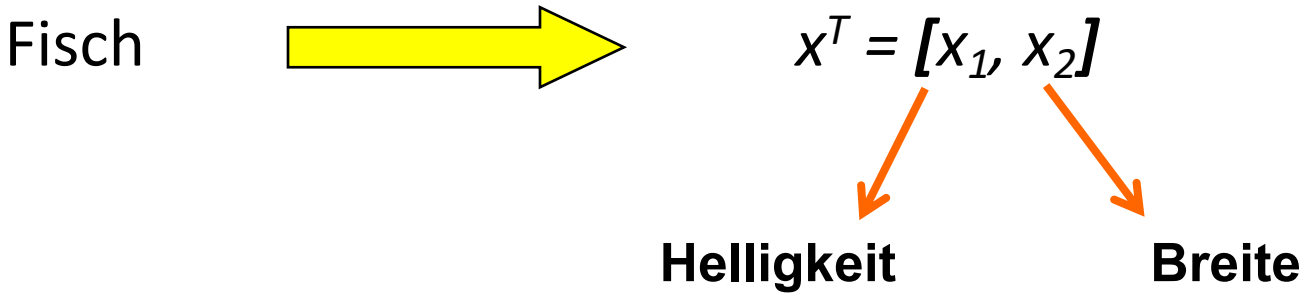
– Schwellwert-Entscheidungsgrenze in Richtung mittlerer Helligkeitswerte minimiert die Kosten (der Fehlklassifikation)

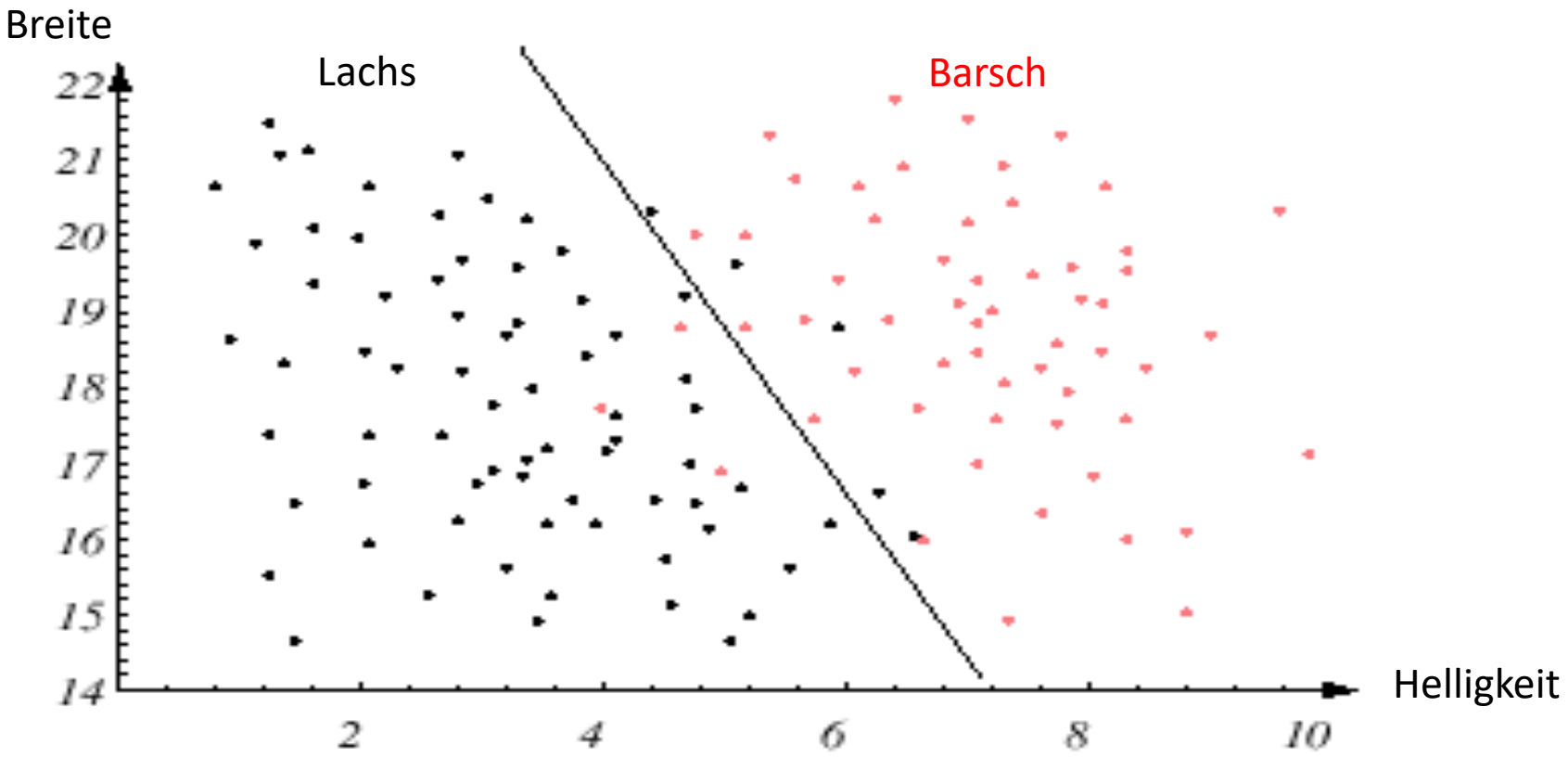


Untersuchung in der sog. Entscheidungstheorie

Ziel ist die automatische Bestimmung von Berechnungsfunktionen für geeignete Merkmale

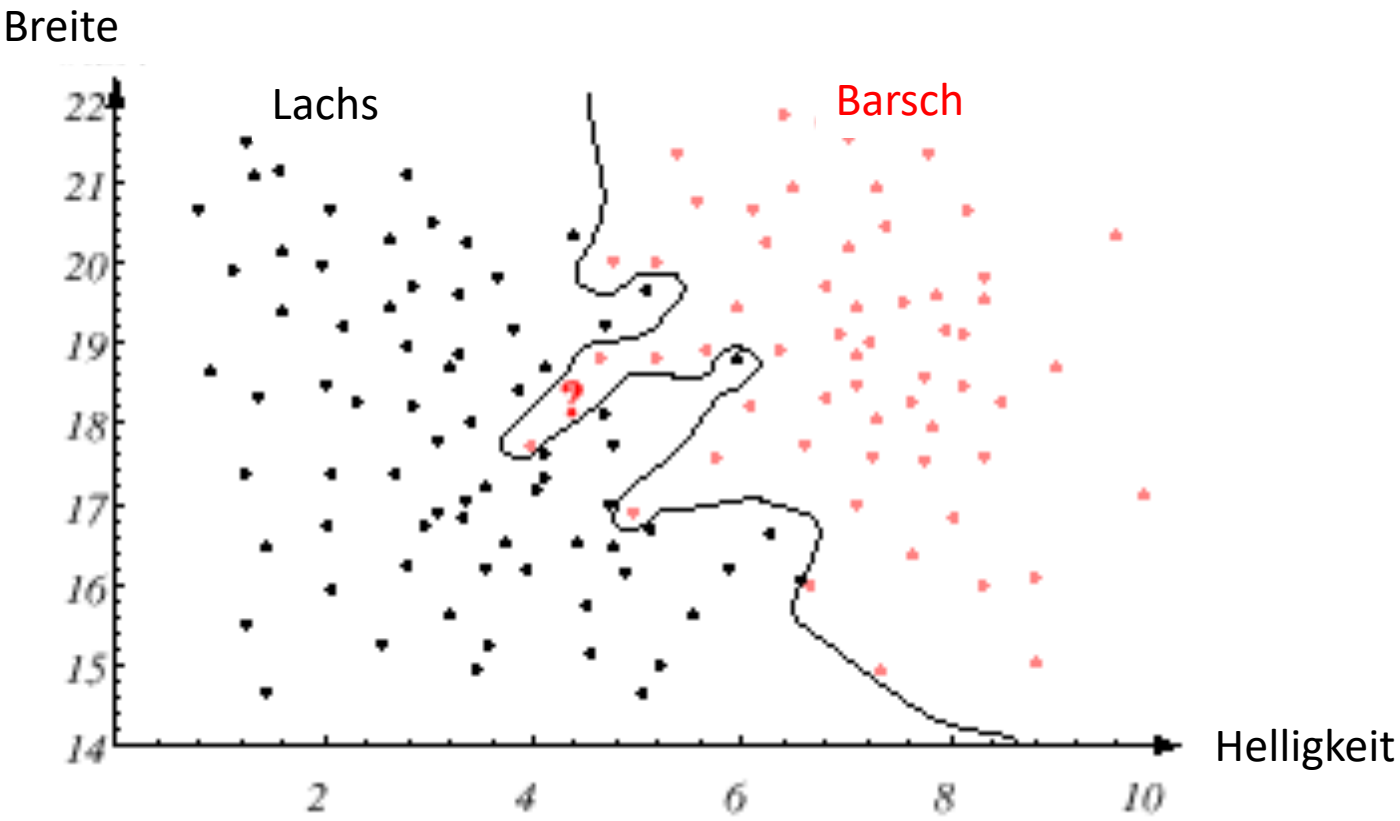
Passe Helligkeit an
und verwende zusätzlich die Breite des Fisches



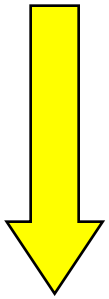

Breite
Lachs
Barsch
22
21
20
19
18
17
16
15
14
2
4
6
8
10
Helligkeit

- Weitere Merkmale, die nicht direkt zu Helligkeit und Breite in Beziehung stehen, könnten hinzukommen
  - Vorsicht aber vor Reduktion durch "verrauschte Merkmale"

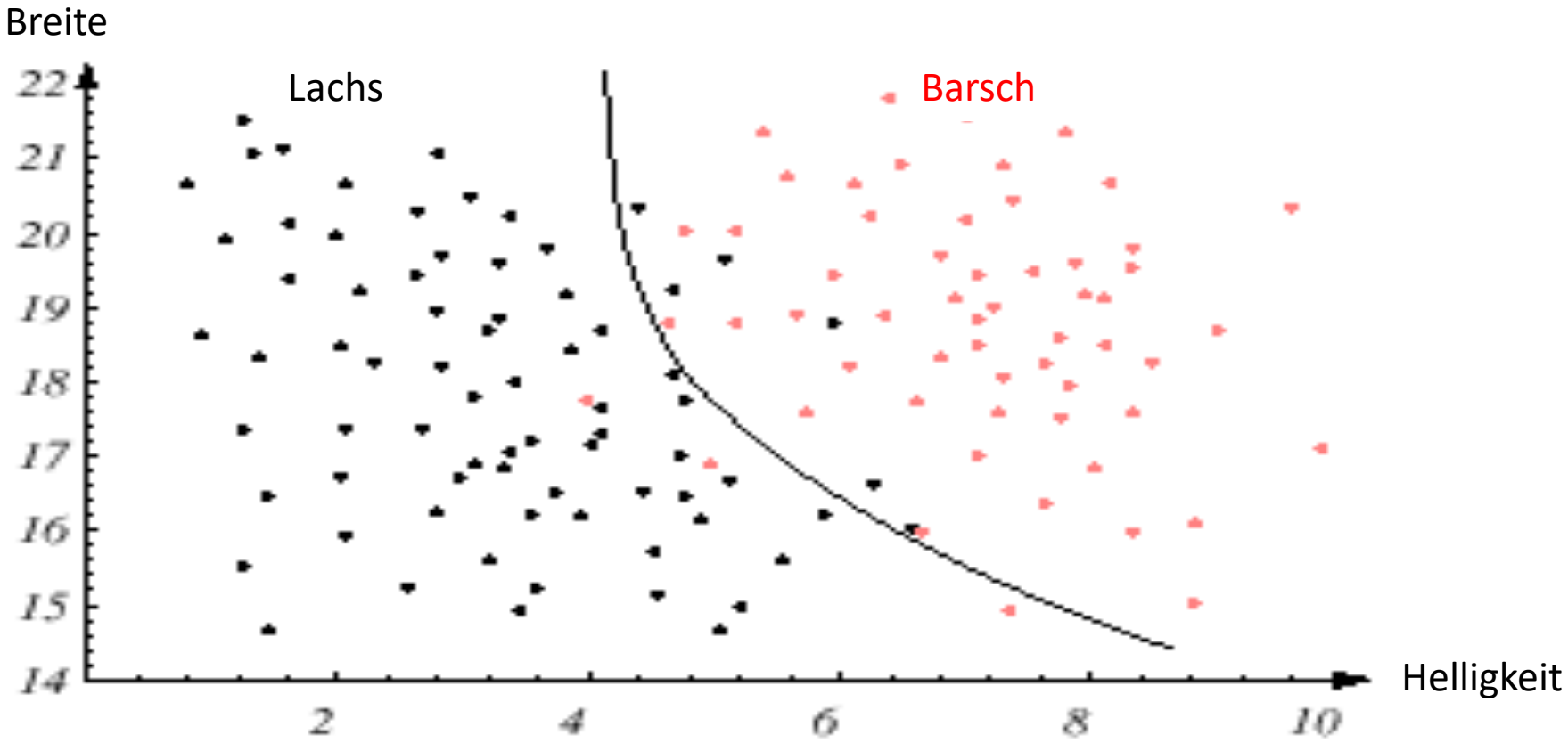
- Wünschenswerterweise ergibt
  die beste Entscheidungsgrenze eine optimale Performanz
  (im Sinne einer Verlustminimierung durch Falschklassifikation)


Breite
Lachs
Barsch
?
22
21
20
19
18
17
16
15
14
2
4
6
8
10
Helligkeit

Vorfreude über Klassifikationsleistung auf Testdaten kann verfrüht sein. Wichtig ist die Leistung auf neuen Daten!



Generalisierungsfähigkeit zählt!


Breite
Lachs
Barsch
22
21
20
19
18
17
16
15
14
2
4
6
8
10
Helligkeit

Klassifikatorentwicklung

- Relevante Merkmale automatisch bestimmbar?
  - Deep Learning, Ensembles
- Dynamische Angepassung des Klassifikators möglich?
  - Ohne spezielle Groundtruth Daten: Transduktives Lernen
  - Mit Belohnungssystem: Reinforcement-Lernen
- Übertragung eines Klassifikators auf neue Anwendung?
  - Transfer-Lernen

Bewertung eines Klassifikators

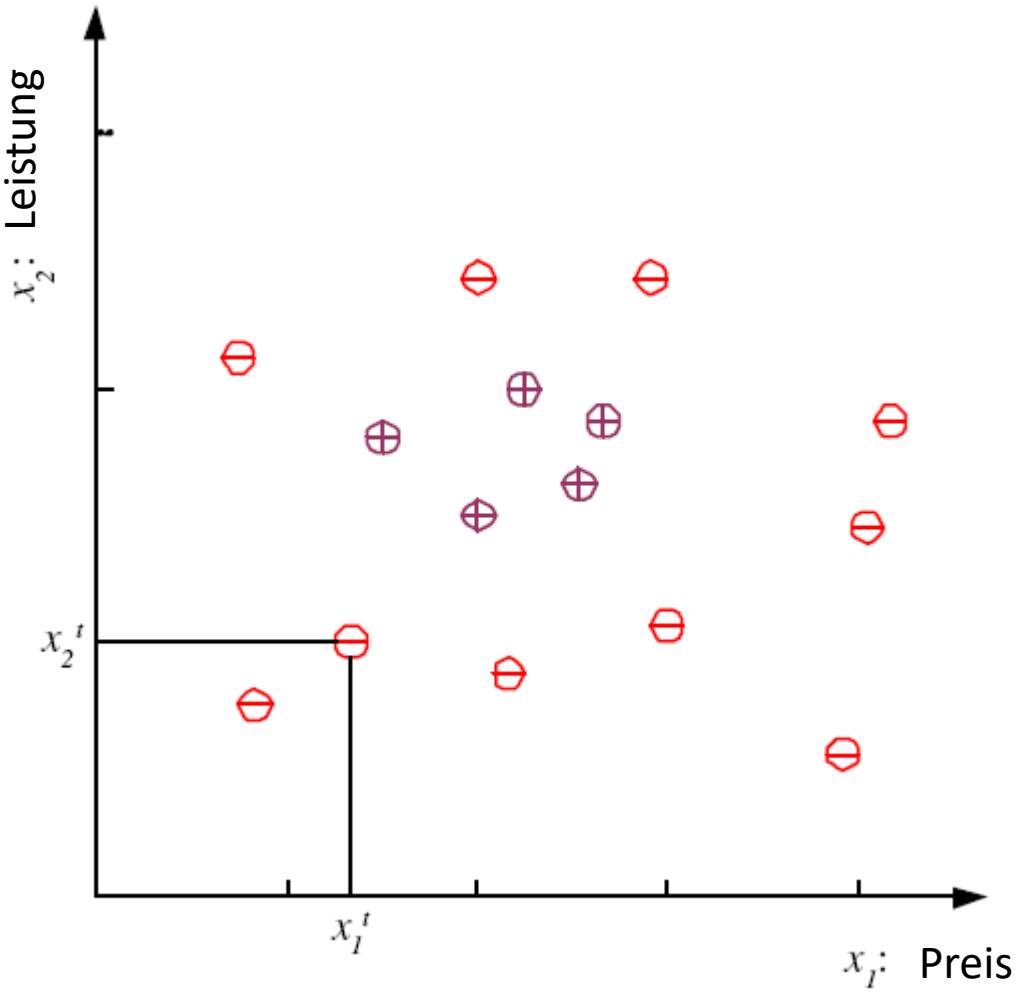
- Klasse C eines “Familienautos”
  - Vorhersage: Ist Auto x ein Familienauto?
  - Wissensextraktion:
    Was erwarten Menschen von einem Familienauto?
- Ausgabe:

  Positive (+) und negative (–) Beispiele
- Repräsentation der Eingabe:

  x_1: Preis, x_2 : Leistung

Frage: Wie gut funktioniert ein bestimmter Klassifikator?

Trainingsmenge $\mathcal{X}$

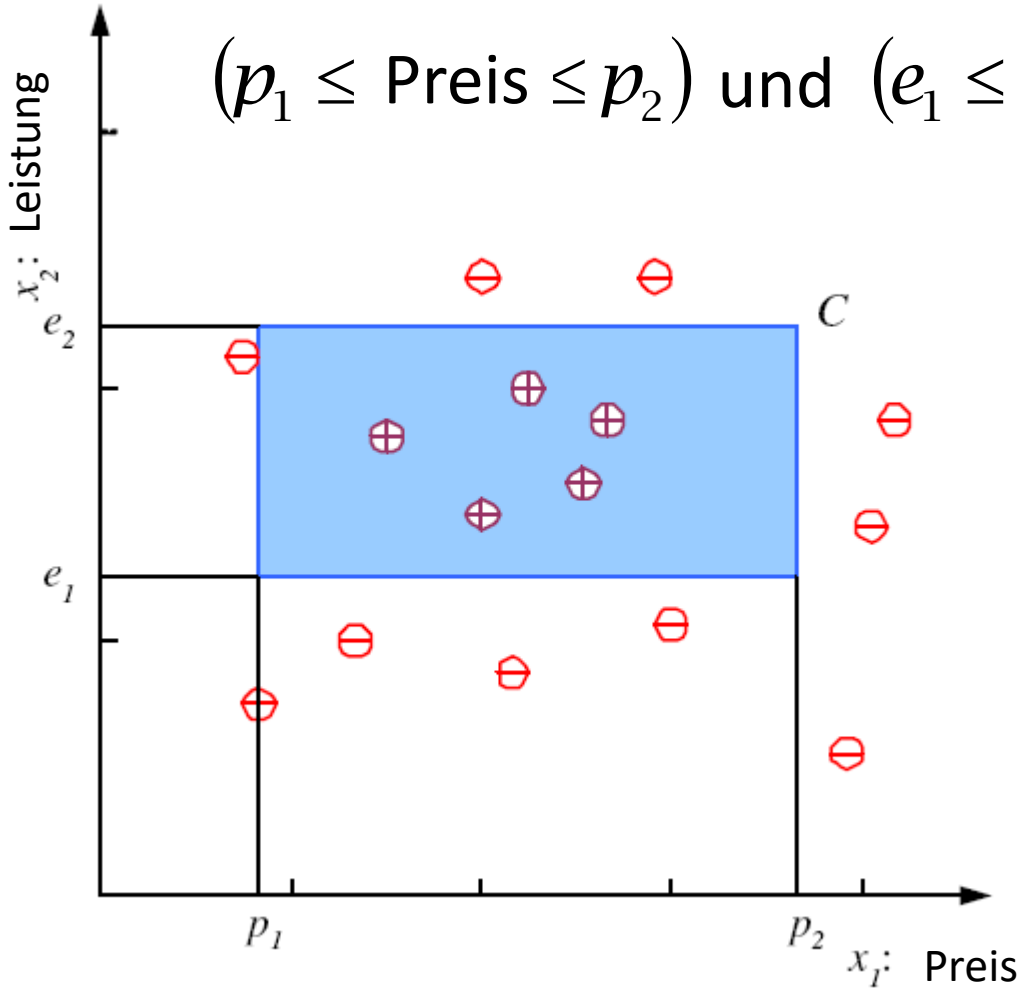


$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_{t=1}^N$$

$$r = \begin{cases} 1 \text{ if } \boldsymbol{x} & \text{ist positiv} \\ 0 \text{ if } \boldsymbol{x} & \text{ist negativ} \end{cases}$$

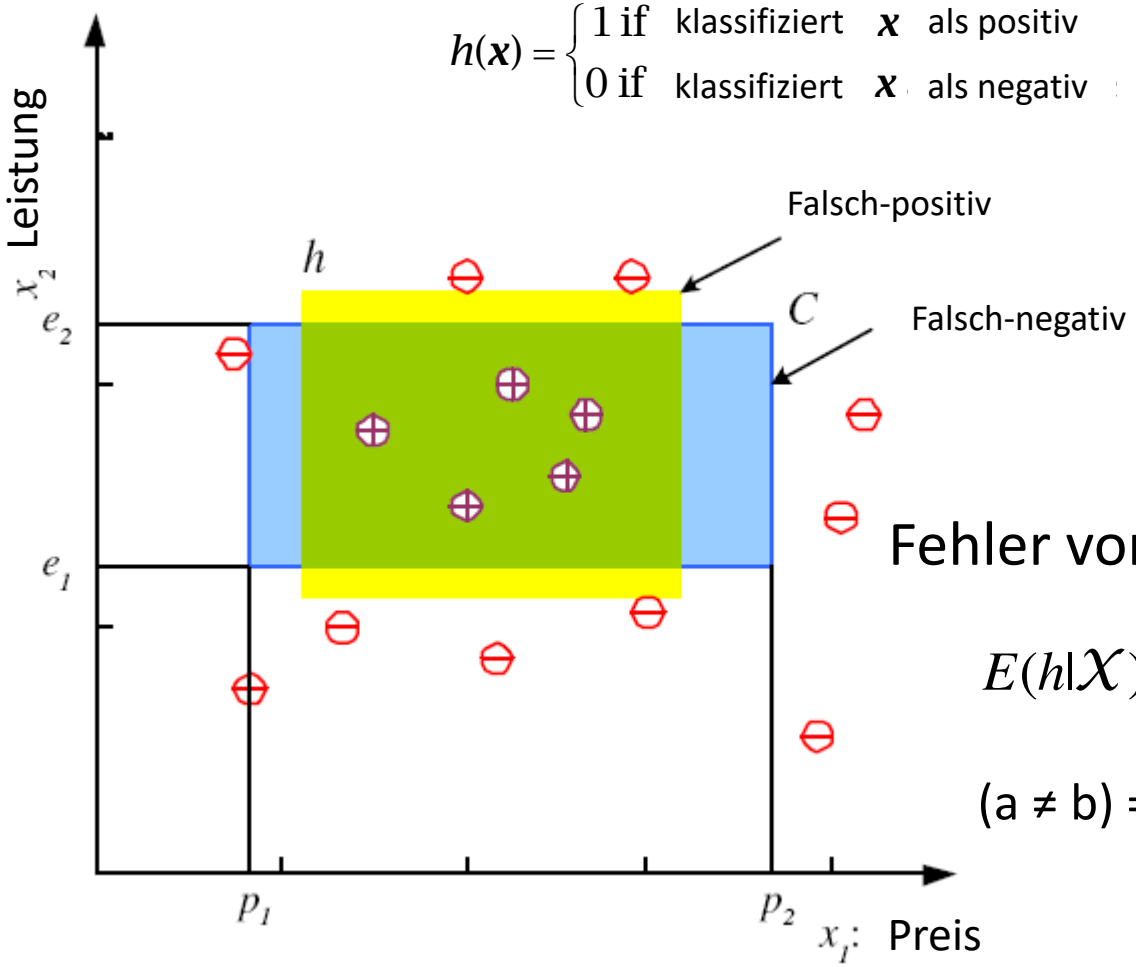
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Richtige Klasse C



Hypothesenklasse $\mathcal{H}$ (z.B. $h \in \mathcal{H}$ in gelb)

$$h(\boldsymbol{x}) = \begin{cases} 1 \text{ if } & \text{klassifiziert } \boldsymbol{x} \text{ als positiv} \\ 0 \text{ if } & \text{klassifiziert } \boldsymbol{x} \text{ als negativ} \end{cases}$$

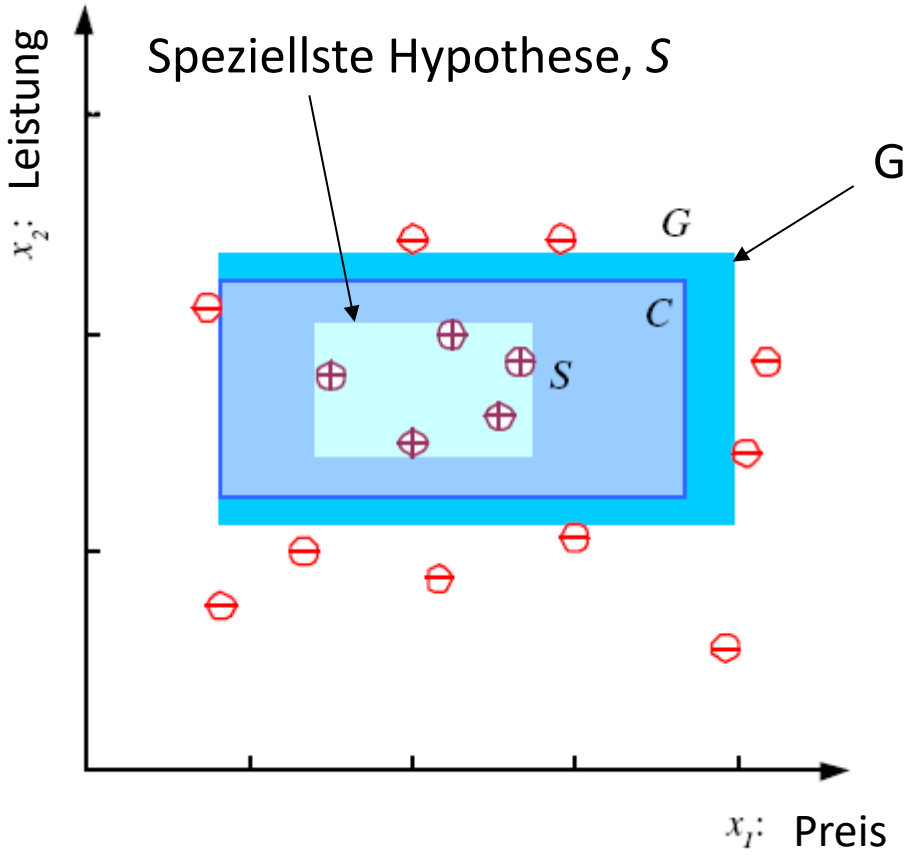


Fehler von h bzgl. $\mathcal{H}$

$$E(h|\mathcal{X}) = (1/N)\sum_{t=1}^{N}\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

(a ≠ b) = 1 if ≠, 0 sonst

S, G, and der Versionsraum (Version Space)



$h \in H$, zwischen S and G ist konsistent und definiert den Versionsraum

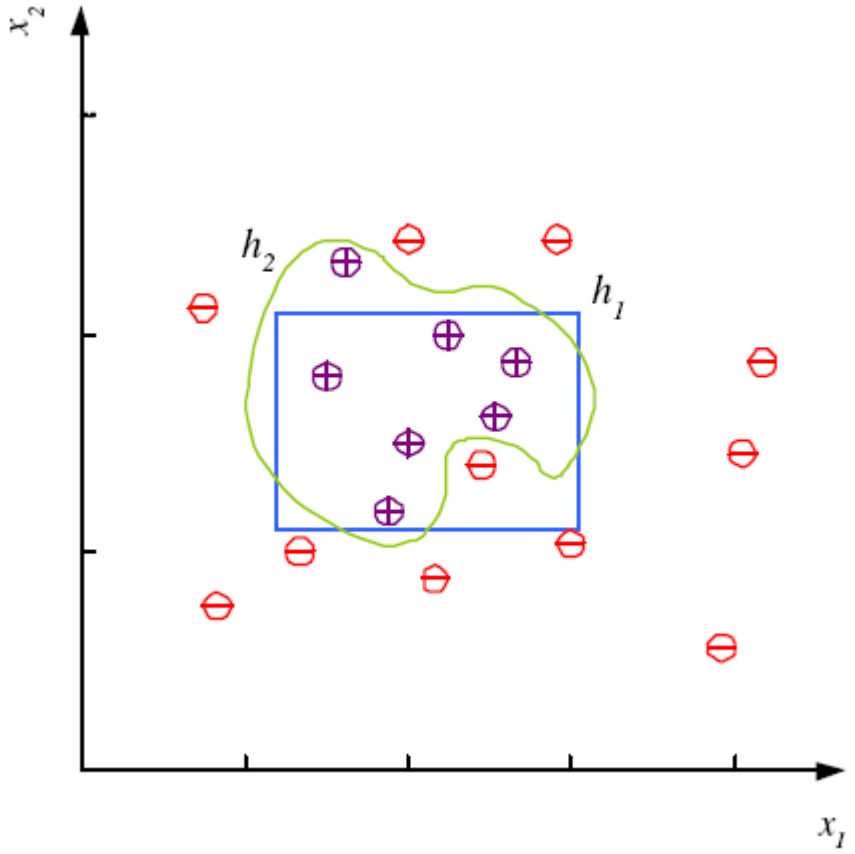
Mitchell, Tom M., Generalization as search. *Artificial Intelligence*. **18** (2): 203–226, **1982**

Rauschen und Modellkomplexität

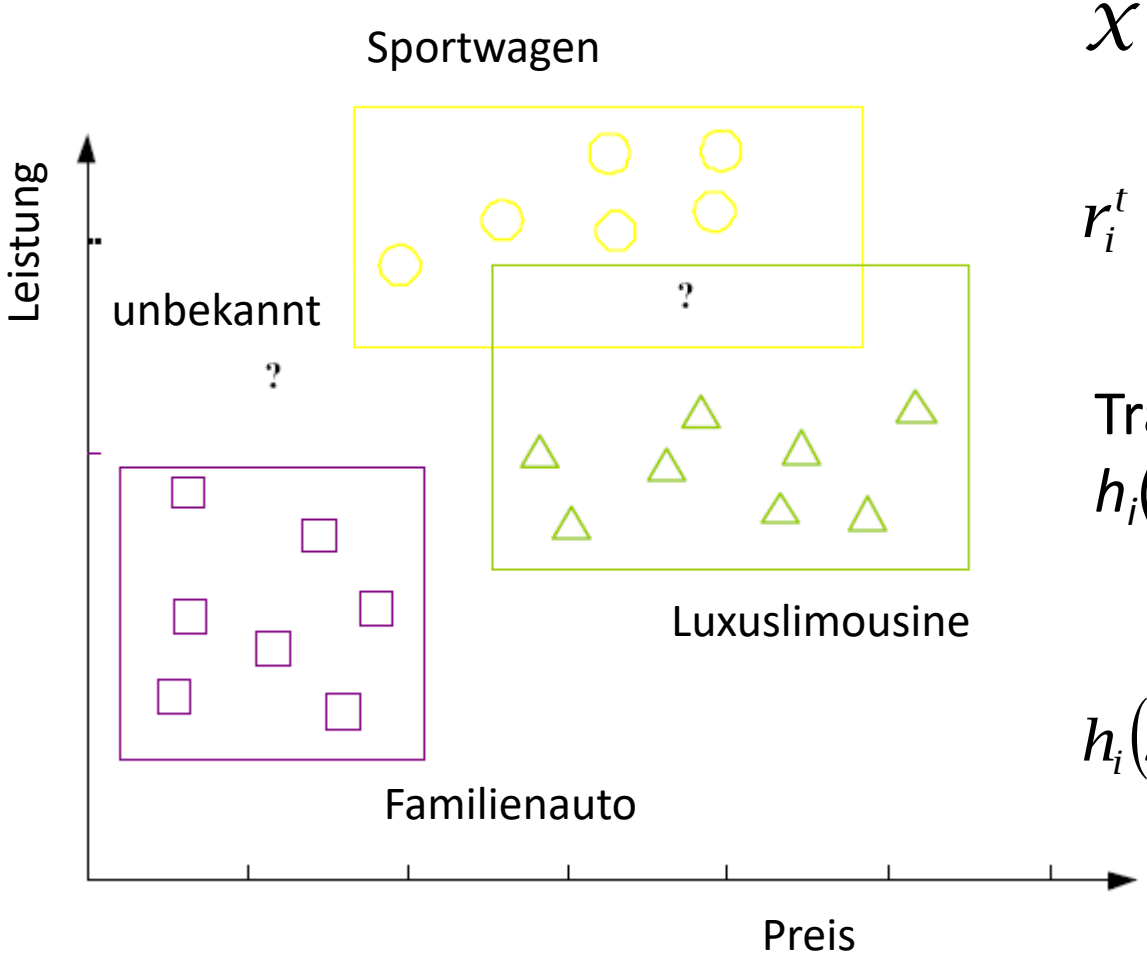
Verwende einfaches Modell:

- Einfacher zu verwenden
  (weniger Berechnungsschnitte)
- Leichter zu trainieren
  (weniger Daten zu speichern)
- Leichter zu erklären
  (besser interpretierbar)
- Bessere Generalisierung
  (Occam's Razor)

Modellkomplexität:
"Größe" der Beschreibung



Clustering: Verschiedene Klassen, C_i i=1,...,K



$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_{t=1}^N$$

$$r_i^t = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

Trainiere Hypothesen $h_i(\boldsymbol{x})$, i =1,...,K:

$$h_i\left(\boldsymbol{x}^t\right) = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

Ausgleichsprobleme: Verschiedene Modellklassen

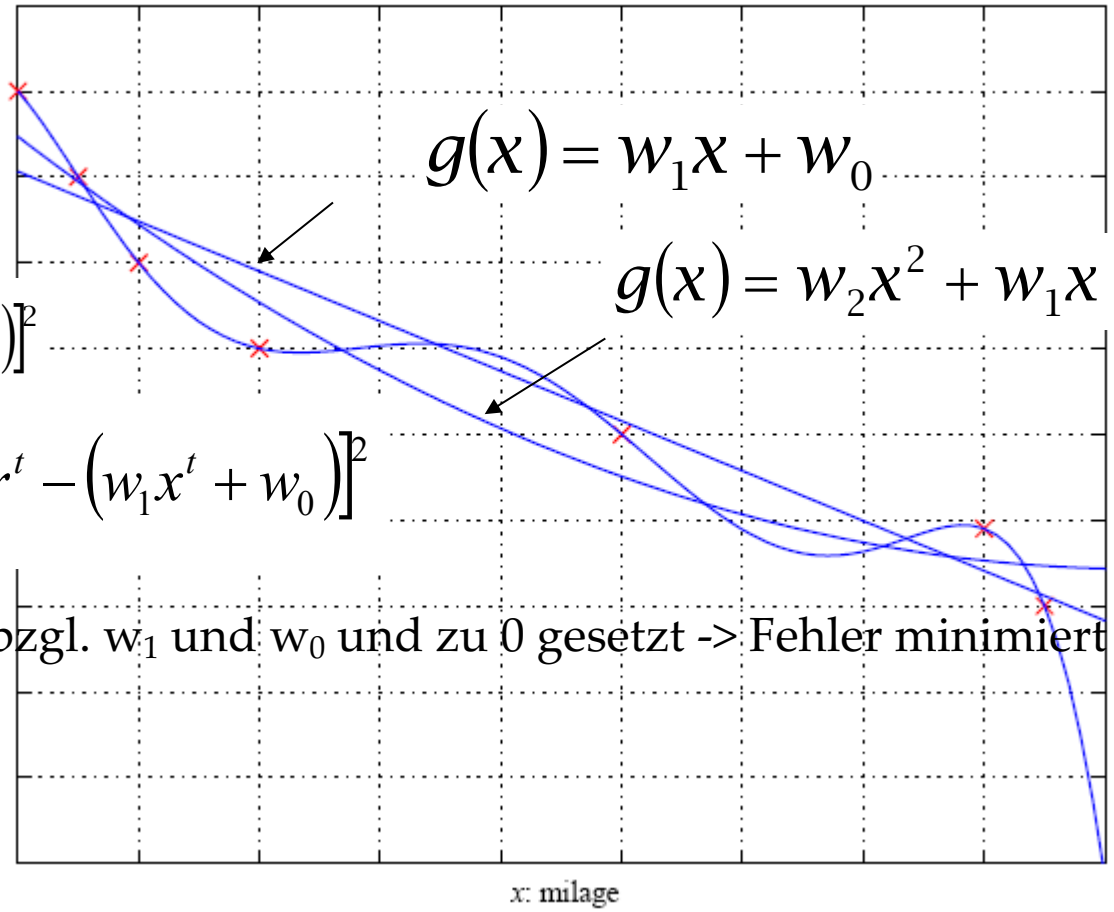
$$\mathcal{X} = \left\{ x^t, r^t \right\}_{t=1}^{N}$$

$$r^t \in \Re$$

$$r^t = f\left(x^t\right)$$

$$E(g \mid \mathcal{X}) = \frac{1}{N} \sum_{t=1}^{N} \left[r^t - g\left(x^t\right) \right]^2$$

$$E\left(w_1, w_0 \mid \mathcal{X}\right) = \frac{1}{N} \sum_{t=1}^{N} \left[r^t - \left(w_1 x^t + w_0\right) \right]^2$$



Partielle Ableitungen von E bzgl. w_1 und w_0 und zu 0 gesetzt -> Fehler minimiert

$$w_1 = \frac{\sum_t x^t r^t - \overline{xr} N}{\sum_t (x^t)^2 - N \overline{x}^2}$$

$$w_0 = \overline{r} - w_1 \overline{x}$$

Überwachtes Lernen

- Beispiel: Ausgleichsrechnung (Regression)[1]
  - Gegeben Datenpunkte, bestimme Parameter, so dass für gegebene x-Werte, die y-Werte geschätzt werden können
  - Optimierungsproblem Minimierung eines Normmaßes

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

- Beispiel: Klassifikation
  - Gegebene Datenpunkte jeweils mit Klassifikationswert, bestimme Klassifikationswert für Datenpunkte ohne diesen (binärer oder mehrwertiger Klassifikator)
  - Kann als Spezialfall eines Ausgleichs gesehen werden

[1] Regression: Rückgang oder Rückschluss von Funktionswerten auf Funktion

Modellauswahl & Generalisierung

- Lernen kann als Optimierungsproblem angesehen werden:
  - Berechne Modell, so dass Fehlerfunktion minimiert
  - Parameter für Repräsentation berechnen → "Parametrisches Lernen"
- Lernen ist i.A. ein schlecht gestelltes Problem
  - In der Regel sind die Daten nicht geeignet, um eine eindeutige Lösung des Optimierungsproblems zu finden
  - Vorannahmen treffen: Annahmen bzgl. H (inductive bias)
  - Regularisierung: Repräsentation für alle (optimalen) Lösungen
- Generalisierung: Wie gut arbeitet Modell auf neuen Daten?
  - Generalisierungsfehler
- Überanpassung (Overfitting): H komplexer als C oder f
- Unteranpassung (Underfitting): H weniger komplex als C oder f

H = Hypothesenklasse, C = Classifier, f = Regressionsfunktion

Drei-Wege-Austauschbeziehung

(Dietterich, 2003):

1. Komplexität von $\mathcal{H}$, $c\,(\mathcal{H})$,
2. Trainingsmengengröße *N*,
3. Generalisierungsfehler, *E*, auf neuen Daten

- Wenn $N\uparrow$, $E\downarrow$
- Wenn $c\,(\mathcal{H})\uparrow$, gilt zuerst $E\downarrow$ und dann $E\uparrow$

Kreuzvalidierung

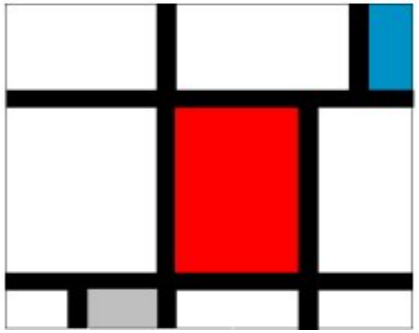
- Um den Generalisierungsfehler abzuschätzen, brauchen wir Daten, mit denen nicht trainiert wurde
- Aufteilung der Daten:
  - Trainingsmenge (50%)
  - Testmenge (z.B. für Publikation) (50%)
- Neuabtastung, wenn wenige Daten vorhanden

Betrachtungsebenen überwachtes Lernen

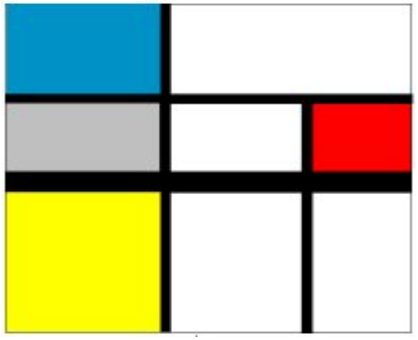
1. Modell: $g(\boldsymbol{x} \mid \theta)$

2. Fehlerfunktion (Verlustfunktion): $E(\theta \mid \mathcal{X}) = \sum_t L(r^t, g(\boldsymbol{x}^t \mid \theta))$

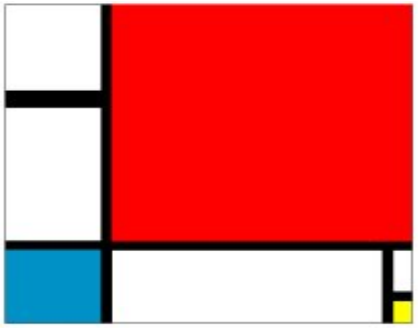
3. Optimierungs verfahren: $\theta^* = \arg\min_\theta E(\theta \mid \mathcal{X})$



1

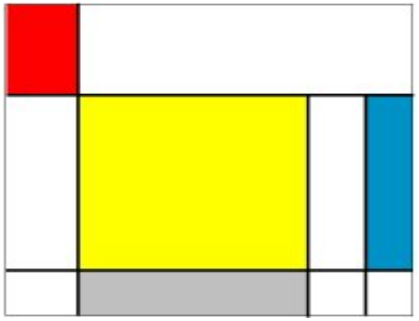


2



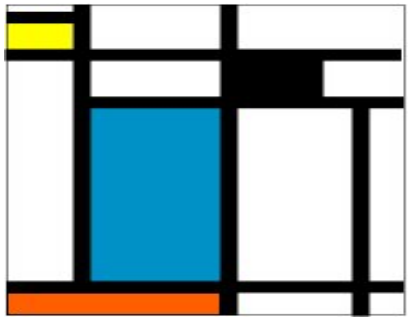
3

4

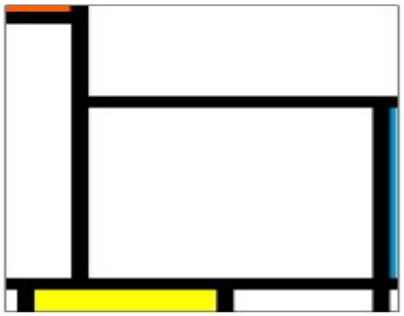


5

6



7



8 ?

Daten in Tabellarischer Form

Number	Lines	Line types	Rectangles	Colours	Mondrian?
1	6	1	10	4	*No*
2	4	2	8	5	*No*
3	5	2	7	4	Yes
4	5	1	8	4	Yes
5	5	1	10	5	*No*
6	6	1	8	6	Yes
7	7	1	14	5	*No*

Anfrage

Number	Lines	Line types	Rectangles	Colours	Mondrian?
8	7	2	9	4	

Analyse von Daten

- Betrachtung einer Spalte x mit n Werten
- Bestimmung des Mittelwerts: $\overline{x} = 1/n \cdot \sum_{i=1}^{n} x_i$
- Große und kleine Werte können sich aufheben
- Mittlere Abweichung vom Mittelwert betrachten (Varianz)
- Bestimmung der Varianz*: $var = 1/n \cdot \sum_{i=1}^{n} (\overline{x}-x_i)^2$
- Meist betrachtet wird die sog. Standardabweichung: $\sigma = \sqrt{var}$

*: Wird später verfeinert.

Halte Daten in normalisierter Form vor

Eine Möglichkeit zur Normalisierung:

$$x_t' \equiv \frac{x_t - \bar{x}_t}{\sigma_t}$$

Gemittelte Abweichung vom Mittel

Normalisierte Trainingsdaten

Number	Lines	Line types	Rectangles	Colours	Mondrian?
1	0.632	-0.632	0.327	-1.021	*No*
2	-1.581	1.581	-0.588	0.408	*No*
3	-0.474	1.581	-1.046	-1.021	Yes
4	-0.474	-0.632	-0.588	-1.021	Yes
5	-0.474	-0.632	0.327	0.408	*No*
6	0.632	-0.632	-0.588	1.837	Yes
7	1.739	-0.632	2.157	0.408	*No*

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{t=1}^{T} [x_{it} - x_{jt}]^2}$$

Number	Lines	Line types	Rectangles	Colours	Mondrian?
8	1.739	1.581	-0.131	-1.021	

Normalisierte Trainingsdaten

Number	Lines	Line types	Rectangles	Colours	Mondrian?
1	0.632	-0.632	0.327	-1.021	*No*
2	-1.581	1.581	-0.588	0.408	*No*
3	-0.474	1.581	-1.046	-1.021	Yes
4	-0.474	-0.632	-0.588	-1.021	Yes
5	-0.474	-0.632	0.327	0.408	*No*
6	0.632	-0.632	-0.588	1.837	Yes
7	1.739	-0.632	2.157	0.408	*No*

$$\sqrt{(0 + 4{,}89 + 5{,}23 + 2{,}04)} = 3{,}489$$

Number	Lines	Line types	Rectangles	Colours	Mondrian?
8	1.739	1.581	-0.131	-1.021	

Distanz der Testinstanz von den Trainingsdaten

Beispiel	Distanz zum Test	Mondrian?
1	2.517	*No*
2	3.644	*No*
3	2.395	Yes
4	3.164	Yes
5	3.472	*No*
6	3.808	Yes
7	3.490	*No*

Klassifikation

1-NN	Yes
3-NN	Yes
5-NN	*No*
7-NN	*No*

Was verwenden wir bei reellwertiger Zielfuntion als Ausgabe?

- Mittel der k-nächsten Nachbarn

Variante von kNN: Distanzgewichtetes kNN

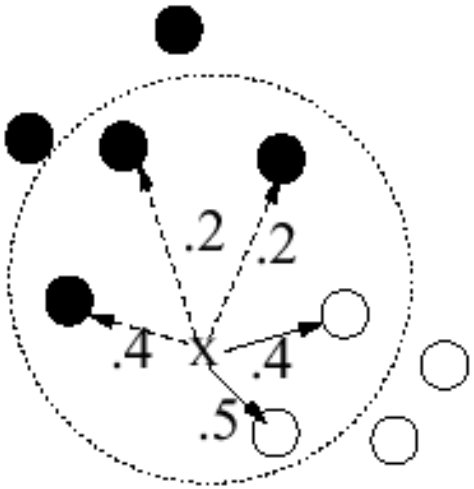
- Nähere Nachbarn haben mehr Einfluss

$$f(\mathbf{x}_q) := \frac{\sum_{i=1}^{k} w_i f(\mathbf{x}_i)}{\sum_{i=1}^{k} w_i} \quad \text{where } w_i = \frac{1}{d(\mathbf{x}_q, \mathbf{x}_i)^2}$$

Dudani, S.A. The distance-weighted k-nearest-neighbor rule. IEEE Trans. Syst. Man Cybern., SMC-6:325–327, **1976**

Variante von kNN: Distanzgewichtetes kNN

k-NN using a weighted-sum voting scheme



kNN (k = 5)

Assign "white" to x because the weighted sum of "whites" is larger then the sum of "blacks".

Each neighbor is given a weight according to its nearness.

Dann könnten wir statt nur k gleich alle Trainingsinstanzen (= Beispiele) nehmen

kNN: Zusammenfassung

- Sehr einfacher Ansatz, nicht-parametrisch
  - Klassifikation (ggf. mit Schwellwert)
  - Regression (Interpolation)
- Verhält sich auch noch gutartig, wenn Daten nicht einfach separiert werden können
- Rang 7 der 10 wichtigsten Data-Mining-Verfahren