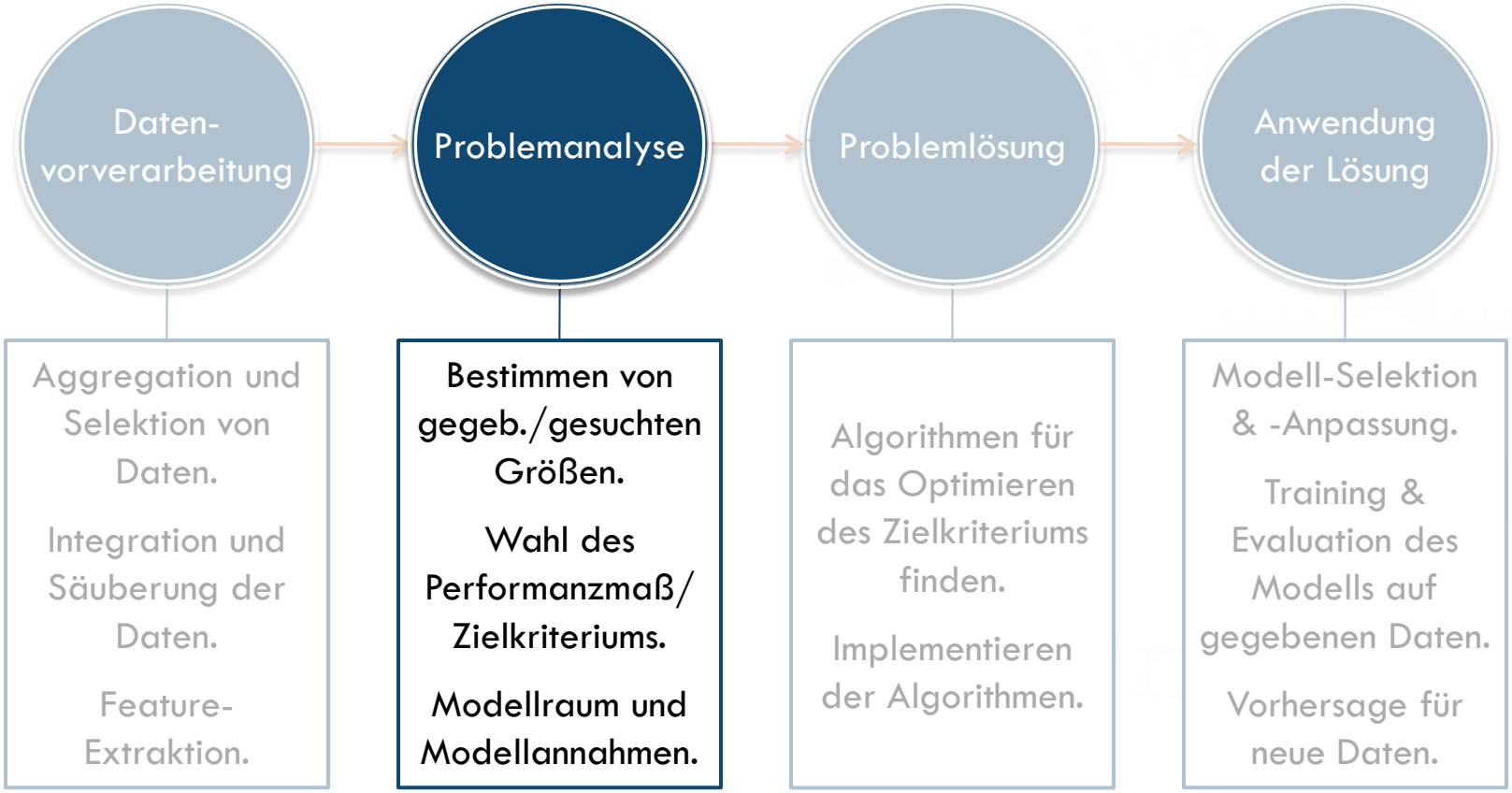
INTELLIGENTE DATENANALYSE IN MATLAB

Problemanalyse & Grundlagen der Lerntheorie

Michael Brückner/Tobias Scheffer

Überblick

- Schritte der Datenanalyse:



Überblick

- Problemanalyse
  - Problemklasse.
  - Daten-Eigenschaften.
- Grundlagen der Lerntheorie.
  - Grundbegriffe.
  - Optimale Entscheidung.

Problemklassen

Überwachtes Lernen von Attributbelegungen

- Gegeben: Trainingsdaten mit bekannten Zielattributen (gelabelte Daten).
- Eingabe: Instanz (Objekt, Beispiel, Datenpunkt, Merkmalsvektor) = Vektor mit Attribut-Belegungen.
- Ausgabe: Belegung des/der Zielattribut(e).
  - Klassifikation: Nominaler Wertebereich des Zielattributs (z.B. {grün, rot, blau}, {Spam, Nicht-Spam}).
  - Ordinale Regression: Ordinaler Wertebereich des Zielattributs (z.B. {klein, mittel, groß}).
  - Regression: Numerischer Wertebereich des Zielattributs (z.B. Temperatur).

Problemklassen

Überwachtes Lernen von Attributbelegungen

- Beispiel *binäre Klassifikation*:

Tag	Bewölkung	Temperatur	Luftfeuchtigkeit	Wind	Tennis spielen?
1	sonnig	warm	hoch	wenig	nein
2	sonnig	warm	hoch	stark	nein
3	bedeckt	warm	hoch	wenig	ja
4	Regen	mild	hoch	wenig	ja
5	Regen	kühl	normal	wenig	ja
6	Regen	kühl	normal	stark	nein
7	bedeckt	kühl	normal	stark	ja
8	sonnig	mild	hoch	wenig	nein
9	sonnig	kühl	normal	wenig	ja
10	Regen	mild	normal	wenig	ja
11	sonnig	mild	normal	stark	?
12	bedeckt	mild	hoch	stark	?
13	bedeckt	warm	normal	wenig	?
14	Regen	mild	hoch	stark	?

Trainingsdaten

Testdaten

Zielgröße

Problemklassen

Unüberwachtes Lernen von Attributbelegungen

- Gegeben: Trainingsdaten mit unbekannten Zielattributen (ungelabelte Daten).
- Eingabe: Instanzen.
- Ausgabe: Belegung der Zielattribute.
  - Clustern: Nominaler Wertebereich des Zielattributs (z.B. Jahreszeiten, {Cluster1, Cluster2, …}).
  - Dichteschätzung: Numerischer Wertebereich des Zielattributs.
  - Visualisierung: Bspw. 3 numerische Zielattribute (Koordinaten).

Problemklassen

Unüberwachtes Lernen von Attributbelegungen

- Beispiel *Clustern* (Tabellendarstellung):

Monat	Bewölkung	Temperatur	Luftfeuchtigkeit	Wind	Jahreszeit
Juli	sonnig	warm	hoch	wenig	?
September	sonnig	warm	hoch	stark	?
August	bedeckt	warm	hoch	wenig	?
April	Regen	mild	hoch	wenig	?
Oktober	Regen	kühl	normal	wenig	?
Dezember	Regen	kühl	normal	stark	?
Januar	bedeckt	kühl	normal	stark	?
Juli	sonnig	mild	hoch	wenig	?
Februar	sonnig	kühl	normal	wenig	?
März	Regen	mild	normal	wenig	?
November	sonnig	mild	normal	stark	?
August	bedeckt	mild	hoch	stark	?
Juni	bedeckt	warm	normal	wenig	?
April	Regen	mild	hoch	stark	?

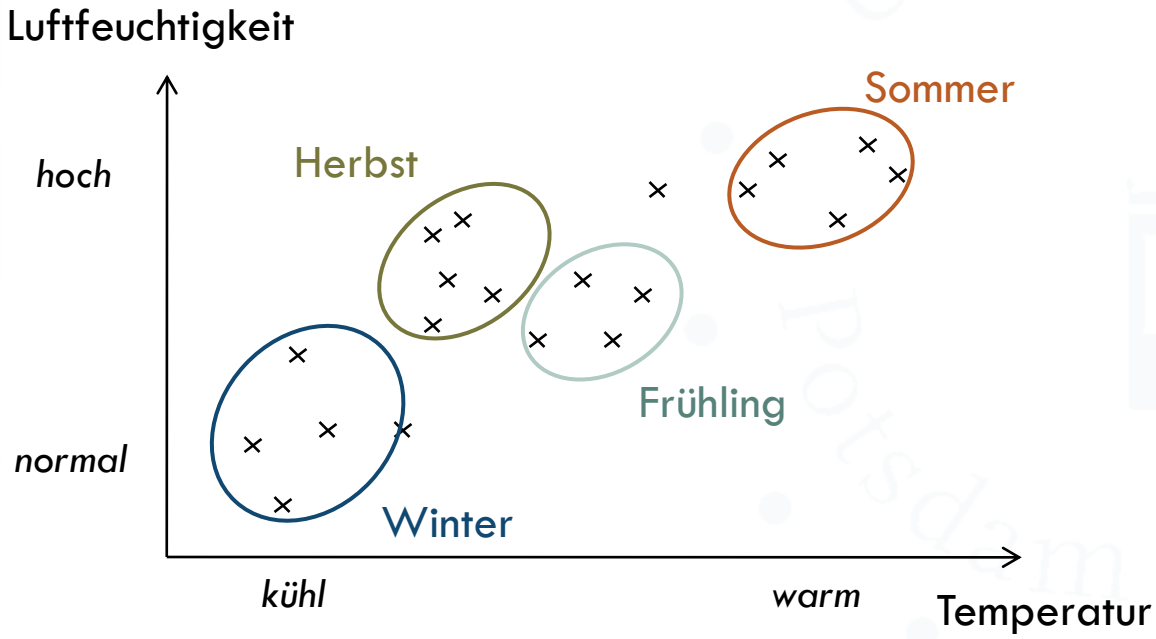
Trainingsdaten

Zielgröße

Problemklassen

Unüberwachtes Lernen von Attributbelegungen

- Beispiel *Clustern* (Diagramm bzgl. der Attribute Luftfeuchtigkeit und Temperatur):



Problemklassen

Suchen von häufigen/seltenen Mustern

- Eingabe: Trainingsdaten mit Relationen zw. den Instanzen.
- Ausgabe: Regeln (allgemeinere Abhängigkeiten).
  - *Frequent Item Sets Detection*: Gegeben sind Tupel von Objekten; Ziel: häufige Teilmengen finden.
  - *Frequent Sequences Detection*: Gegeben sind Sequenzen von Objekten; Ziel: häufige Teilsequenzen finden.
  - *Anomaly/Novelty Detection*: Gegeben sind Sequenzen oder Punktmengen; Ziel: Auffälligkeiten entdecken.

Problemklassen

Halbüberwachtes Lernen von Attributbelegungen

- Gegeben: Trainingsdaten mit teilweise bekannten Zielattributen (gelabelte und ungelabelte Daten).
- Eingabe: Instanz.
- Ausgabe: Belegung des Zielattributs.
  - Überwachtes Clustern: Wertebereich des Zielattributs bekannt (Ziel: Zuordnung der ungelabelten Beispiele).
  - Active Learning: Ungelabelte Daten durch Experten gelabelt (Ziel: möglichst wenig Beispiele labeln zu müssen).
  - Self-Learning, Co-Learning: Ungelabelte Daten durch Modell gelabelt; fließen direkt/indirekt in Trainingsprozess ein.

Problemklassen

Überwachtes Lernen von globalen Ordnungsrelationen

- Gegeben: Trainingsdaten mit einer bekannten Sortierung bzw. paarweisen Ordnung.
- Eingabe: Menge von Instanzen.
- Ausgabe: Eine Sortierung der Instanzen.
  - Ordinale Regression.
  - Rank-Learning.
    - Pairwise Ranking: Ordnung für die Eingabe von zwei Instanzen.
    - Label Ranking: Scoring/Sortierung von mehr als zwei Instanzen.

Problemklassen

Überwachtes Lernen von globalen Ordnungsrelationen

- Beispiel *Rank-Learning*:

Reifenmodell	Preis in EUR	Bremsweg (trocken)	Bremsweg (nass)	Geräusch	Platzierung
Fulda Carat Progresso	44 bis 70	2,0	1,8	3,0	1
Continental PC 2	59 bis 77	1,3	2,0	3,5	2
Bridgestone Turanza	51 bis 75	1,3	2,2	2,9	3
Uniroyal Rain Expert	52 bis 75	1,9	1,9	3,7	4
Semperit Comfort Life	46 bis 66	2,2	2,3	3,3	5
Firestone TZ300 a	48 bis 65	1,8	2,3	3,3	6
Dunlop SP Sport	51 bis 77	1,4	2,8	3,1	7
Vredestein Hi-Trac	37 bis 68	4,1	2,1	4,2	8
Hankook Optimo	43 bis 66	2,1	3,2	3,2	9
Yokohama C.Drive	48 bis 69	1,6	3,2	3,1	10
Goodyear DuraGrip	46 bis 72	2,0	2,8	3,0	?
Michelin Energy Saver	61 bis 86	2,1	2,5	3,0	?
Pneumant PN550	39 bis 65	2,4	3,2	3,6	?
Kumho Solus KH17	42 bis 61	1,8	2,2	3,0	?

Trainingsdaten

Testdaten

Zielgröße

Problemklassen

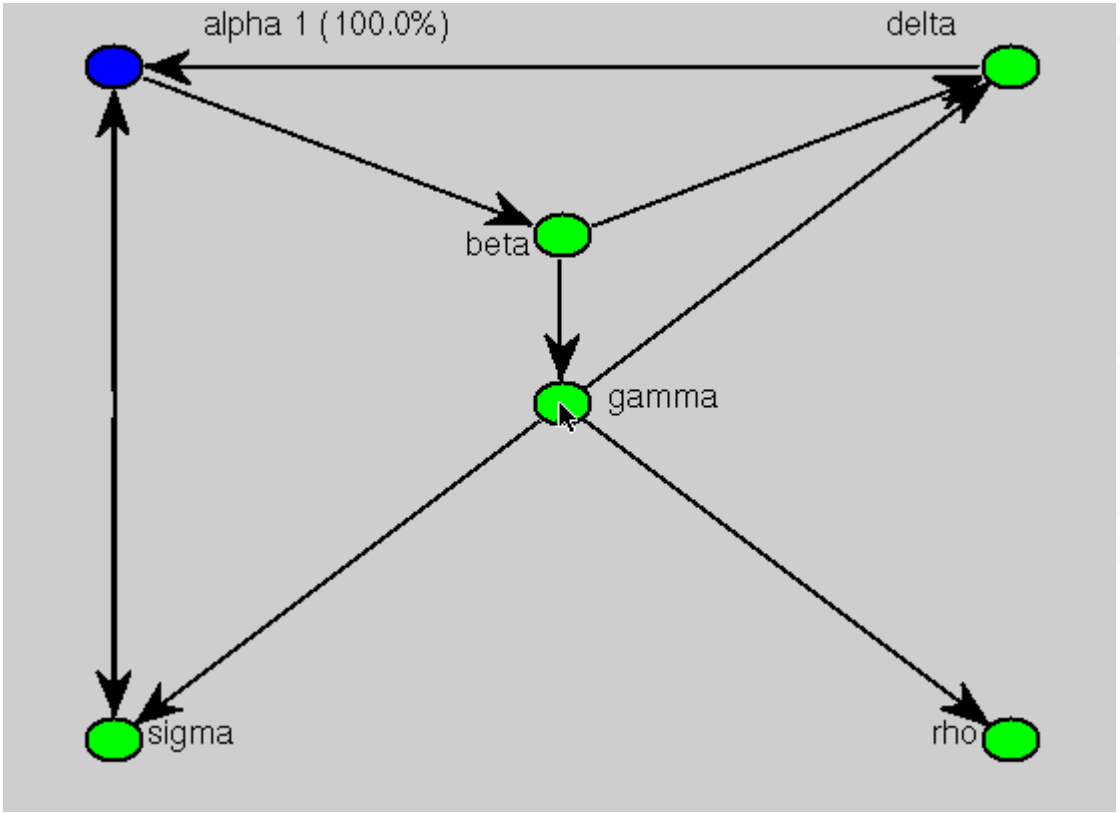
Unüberwachtes Lernen von globalen Ordnungsrelationen

- Gegeben: Trainingsdaten mit einer unbekannten Sortierung bzw. paarweisen Ordnung.
- Eingabe: Menge von Instanzen.
- Ausgabe: Eine Sortierung der Instanzen nach Relevanz.
  - Authority-Ranking.
    - Sortierung unbekannt aber Modell für Relevanz (Authority) bekannt (z.B. Random-Walk-Modell in Graphen).
    - Sortierung/Relevanz nur indirekt bekannt (z.B. durch Anklicken eines Links).

Problemklassen

Un-/Halbüberwachtes Lernen von globalen Ordnungsrelationen

- Beispiel *Authority-Ranking* mit Random-Walk-Modell:



Problemklassen

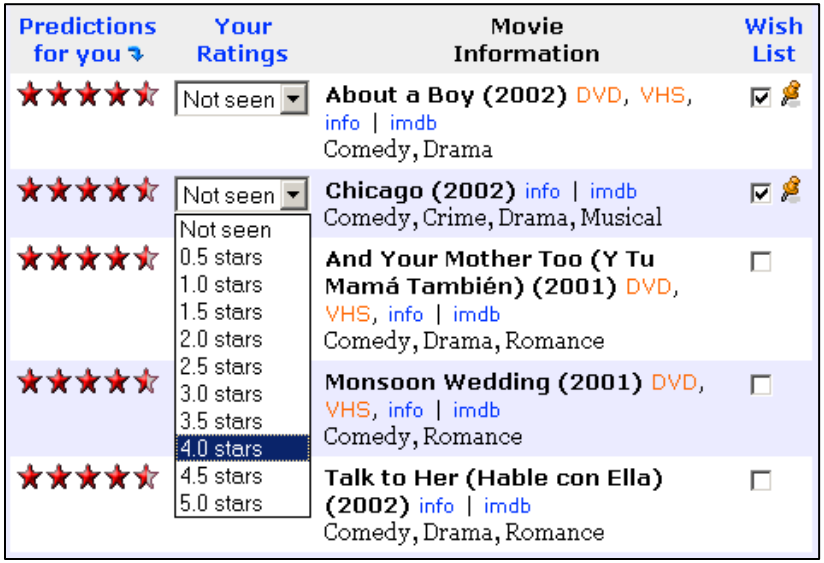
Lernen von lokalen Ordnungsrelationen

- Gegeben: Trainingsdaten mit mehreren (teilweise) bekannten nutzerabhängigen Sortierungen.
- Eingabe: Menge von Instanzen und Nutzern.
- Ausgabe: Sortierung der Instanzen pro Nutzer.
  - Collaborative Filtering: Nur Präferenzen teilweise bekannt.
  - Content-based Filtering: Eigenschaften von Instanzen und Nutzern bekannt; Präferenzen teilweise bekannt.

Problemklassen

Lernen von lokalen Ordnungsrelationen

- Beispiel *Collaborative Filtering*:



Film	Nutzer 1	Nutzer 2	Nutzer 3
About a Boy	5		4
Chicago	4	2	?
And Your M…			5
Monsoon Wedding	4	1	?
Talk to Her			4
Titanic			?
The Bourne Identity	2	2	?
SAW	1	5	?
Se7en		4	1
Earth			5
Stuart Little		1	?

Trainingsdaten

Zielgröße

Problemklassen

Weitere Lernprobleme bzw. Spezialfälle

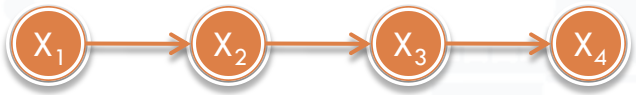
- Adversarial Learning: Lernen in *Static Games*.
- Reinforcement Learning: Lernen in *Repeated Games*.
- Daten-Modellierung & -Visualisierung.
- …

Daten-Eigenschaften

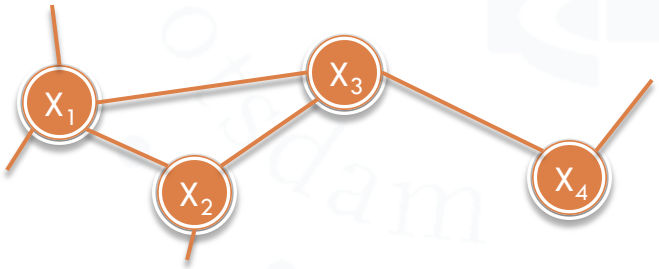
Abhängigkeit zwischen den Datenpunkten

- **Unabhängige Daten:**
  Datenpunkte sind unabhängig von einander (z.B. Kundendaten).

- **Sequenzen:**
  Folge von abhängigen Datenpunkten (z.B. Wetterdaten).



- **Verlinkte Daten:**
  Abhängigkeiten höherer Ordnung (z.B. Webseiten).



Daten-Eigenschaften

Verfügbarkeit

- Batch Learning: Trainingsdaten (und evtl. Testdaten) zum Analysezeitpunkt verfügbar.
  - Beispiel: Kaufverhalten auf Kundendaten vorhersagen.

- Online Learning: Trainingsdaten (und evtl. Testdaten) zum Analysezeitpunkt sequenziell verfügbar.
  - Beispiel: Spam-Filter im Email Client (Nachtrainieren nach jeder neuen, vom Nutzer gelabelten Email).

Daten-Eigenschaften

Umfang und Qualität

- Umfang:
  - Scale: Viele/wenige verschiedene Attribute/Instanzen.
  - Sparsity: Viele/wenige verschiedene Attributbelegungen.
- Qualität:
  - Fehlerhafte Attribute/Zielattribut(e).
  - Fehlende Attributbelegung.
  - Nicht-repräsentative Beispiele.
  - Ungleichmäßige Verteilung der Zielattributbelegungen.
  - …

Grundlagen der Lerntheorie

Grundbegriffe

- Modell (Hypothese): Funktion welche, abhängig von der Problemklasse, den Eingabeattributen/-Instanzen eine Belegung der Zielgröße zuordnet.
- Lernverfahren: Algorithmus welcher für gegebene Daten ein Modell auswählt.
- Modellraum (Hypothesenraum): Menge der Modelle, die vom Lernverfahren berücksichtigt werden.

Grundlagen der Lerntheorie

Grundbegriffe

- Versionsraum: Menge der Modelle, die mit den Trainingsdaten konsistent sind.
  - Versionsraum wird kleiner je mehr Daten vorhanden sind, ist i.d.R. jedoch unendlich groß!
  - Versionsraum leer: Trainingsmenge widersprüchlich oder zu eingeschränkter Modellraum.
  - Alle Elemente des Versionsraums erklären die Daten gleichermaßen gut.
  - Annahme: Genau ein Element des Versionsraums hat die Daten erzeugt.

Grundlagen der Lerntheorie

Grundbegriffe

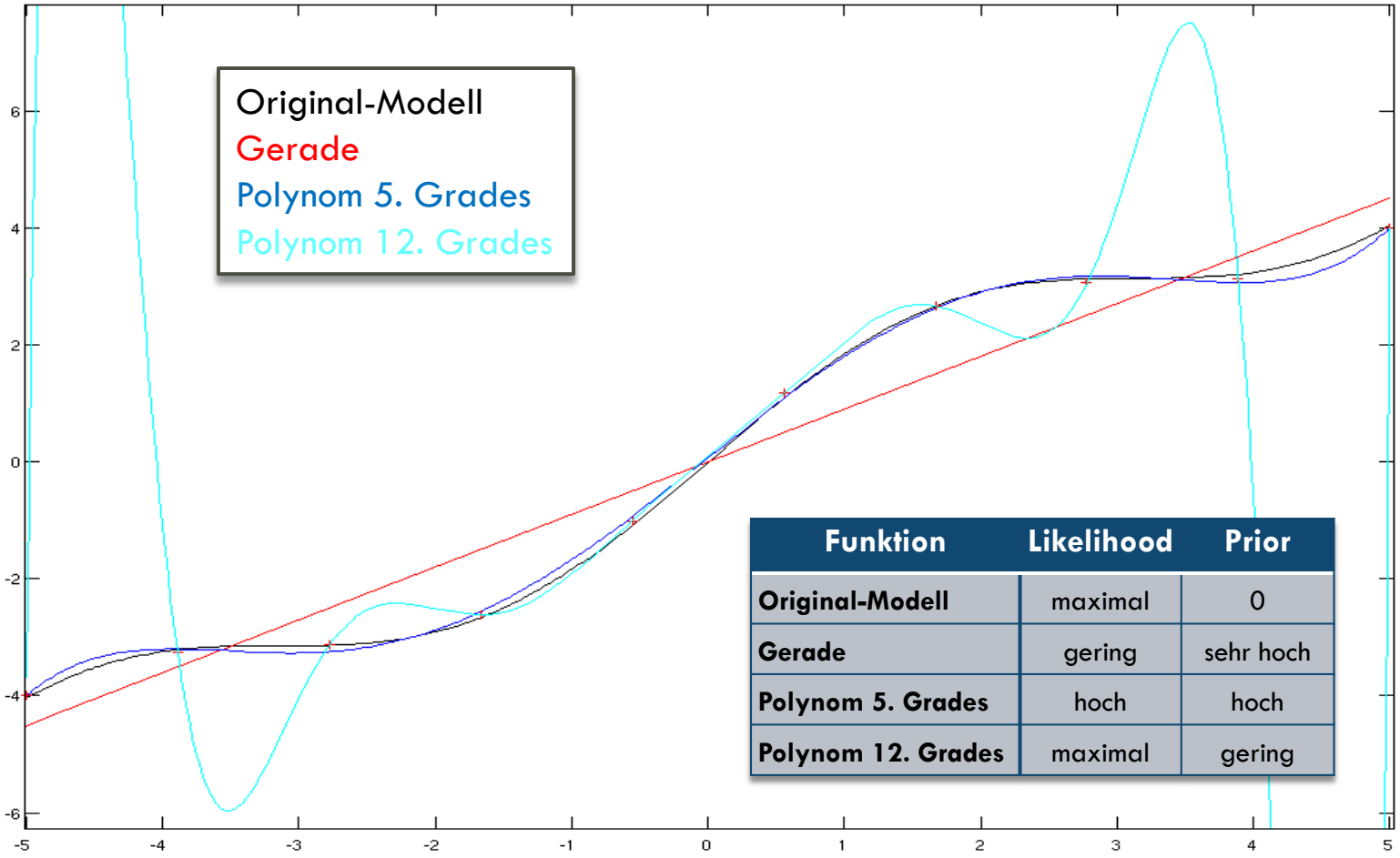
- A-Posteriori-Wahrscheinlichkeit (*Posterior*) eines Modells: Wie wahrscheinlich ist das Modell h nachdem wir die Daten D gesehen haben.
- A-Priori-Wahrscheinlichkeit (*Prior*) eines Modells: Wie wahrscheinlich ist das Modell h bevor wir die Daten D gesehen haben.
- *Likelihood* der Daten: Wie wahrscheinlich wurden die Daten D durch das Modell h erzeugt.

Posterior Likelihood Prior

$$p(h \mid D) = \frac{p(D \mid h)\, p(h)}{p(D)}$$

Grundlagen der Lerntheorie

Grundbegriffe



Funktion	Likelihood	Prior
Original-Modell	maximal	0
Gerade	gering	sehr hoch
Polynom 5. Grades	hoch	hoch
Polynom 12. Grades	maximal	gering

Grundlagen der Lerntheorie

Grundbegriffe

- Verlustfunktion l: Negativer Logarithmus der Daten-Likelihood.
  - Likelihood im Intervall [0,1] $\Rightarrow$ Verlustfunktion im Intervall [0,∞].
  - Je höher die Likelihood, desto geringer der Verlust.
  - Je geringer die Likelihood, desto größer der Verlust.

$$l(D,h) = -\log p(D \mid h)$$

- Regularisierer Ω: Negativer Logarithmus der A-Priori-Wahrscheinlichkeit eines Modells.
  - Analoge Eigenschaften wie Verlustfunktion.
  - Kodiert „Vorwissen“ über die Wahl eines Modells.

$$\Omega(h) = -\log p(h)$$

Grundlagen der Lerntheorie

Optimale Entscheidung

- Wähle das Modell, welches am wahrscheinlichsten die Daten erzeugt hat = Maximum Likelihood (ML).
  - Problem: Likelihood ist gleichgroß für alle Modelle im Versionsraum!

$$\arg\max_h p(D \mid h) = \arg\min_h l(D, h)$$

- Wähle das Modell, welches am wahrscheinlichsten ist, gegeben die Daten = Maximum A-Posteriori (MAP).

$$\arg\max_h p(h \mid D) = \arg\max_h p(D \mid h)\, p(h) = \arg\min_h l(D, h) + \Omega(h)$$

- Mittle über alle Modelle gewichtet mit ihrer A-Posteriori-Wahrscheinlichkeit = Bayes'sches Lernen.

Grundlagen der Lerntheorie

Optimale Entscheidung

- Generatives Lernen:
  - Schätze Likelihood-Funktion explizit aus den Trainingsdaten.
  - Berechne aus geschätzter Likelihood und Prior das wahrscheinlichste Modell.
  - Wähle Entscheidungsfunktion: Vorhersage für Zielgrößen gemäß des Modells.
- Diskriminatives Lernen:
  - Schätze Posterior aus den Trainingsdaten und wähle das wahrscheinlichste Modell.
  - Wähle Entscheidungsfunktion: Vorhersage für Zielgrößen gemäß des Modells.
- Direktes Lernen:
  - Finde Funktion welche direkt aus den Trainingsdaten eine Vorhersage für Zielgrößen trifft (ohne Wahrscheinlichkeiten explizit zu modellieren).

Grundlagen der Lerntheorie

Finden der optimalen Entscheidung

- Problemklasse, Modellraum, Lernstrategie usw. bestimmen.
- Zielkriterium (Optimierungsaufgabe) formulieren.
- Optimierungsaufgabe (OA) lösen:
  - Analytisch: Geschlossene Lösung existiert, z.B. Nullstellen einer Parabel.
  - Numerisch: Keine (bzw. zu aufwendige) geschlossene Lösung, z.B. Nullstellen durch Newton-Verfahren finden.
  - Analytische Approximation: OA durch analytisch lösbare Aufgabe approximieren, z.B. Optimierung einer konvexen oberen (bzw. konkaven unteren) Schranke der Zielfunktion.
  - Numerische Approximation: OA durch numerisch lösbare Aufgabe approximieren, z.B. komplexe OA in mehrere einfache OAs zerlegen.
  - Greedy-Suche.

Zusammenfassung

- Zahlreiche unterschiedliche & unterschiedlich schwere Datenanalyseprobleme, z.B.
  - Klassifikation/Regression,
  - Clustern,
  - Ranking usw.
- Verschiedene Lernstrategien & Lösungsansätze.
- Erkennen des Datenanalyseproblems und Wahl einer geeigneten Lernstrategie entscheidend!