

MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2017

Prof. E.G. Schukat-Talamazzini

Stand: 23. Oktober 2017

Teil I

Methoden und Aufgabenstellungen

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Was ist Lernen ?

Antworten dreier Urväter des maschinellen Lernens

Lernen nach Herbert Simon

„Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task (or tasks drawn from the same population) more efficiently and more effectively the next time.“

(Automatic Performance Improvement)

Trifft Simons Definition unser intuitives Verständnis?

... zu weit?

Schärfen eines Messers
schnellere CPU

... zu eng?

Zwangsarbeiter täuscht Leistung vor

Passant \rightsquigarrow Oper \rightsquigarrow Auskunft

Lernen nach Dana Scott

Prozeß des Aufbaus abrufbarer Repräsentationen von vergangenen Interaktionen mit der Umwelt

Lernen nach Ryszard Michalski

Konstruieren oder Verändern der Repräsentationen von Erfahrungen

Leistungsbegriff?!

Wozu maschinelles Lernen ?

Lernen ist der Schlüssel zur Intelligenz — bei Mensch und Maschine

Knowledge Acquisition Bottleneck

Experten sind oft unfähig, ihr Wissen zu formalisieren.

Wissenserwerb und -einpfege

... sind teuer, langsam und unsicher.

Problemstruktur ist zu komplex

Sprache, Schrift, Szenen, DNA, ...

Maschine findet überlegene Lösungen

Greifende/balancierende Roboter ...

SYNERGIE von Mensch & Maschine

- 👉 Lernfähigkeit des Menschen
- 👉 Kopierfähigkeit des Rechners
- 👉 Lerngeschwindigkeit des Rechners

Ziele des Lernens

Lösung

genauer

Aufgabenbereich

breiter

Arbeitsweise

ökonomischer

Wissensstruktur

einfacher



Alan Turing

den Computer **erziehen**!

Was wird gelernt ?

Kognitionspsychologie des menschlichen (früh/kindlichen) Lernens

Begriffe

{ Struktur
Erwerb
Nutzung }

Aggregation (Extension von Begriffen)

- Gruppieren von Objekten in Kategorien
- Sinnvolle Begriffe \rightsquigarrow Vorhersage von Objektverhalten

Charakterisierung (Intension von Begriffen)

- Gemeinsame Eigenschaften aller Instanzen eines Begriffs
- Welche Merkmale? \leftrightarrow kultureller/sprachlicher Kontext

Klassifikation

- Zuordnen eines Objekts zu „seiner“ Kategorie
- Einordnen in eine Hierarchie von Unter- und Oberbegriffen

Induktives Lernen

Verallgemeinerndes Lernen aus (endlich vielen) Beispielen

$$\gamma_A \hat{=} A(x) \wedge A(y) \wedge A(z)$$

$$\gamma_B \hat{=} B(x) \wedge B(y) \wedge B(z)$$

$$\gamma_V \hat{=} \forall x (A(x) \Rightarrow B(x))$$

Deduktion

allgemein \rightarrow speziell

(formallogisch korrekte Schlußweise)

$$\gamma_V, \gamma_A \vdash \gamma_B$$

Induktion

speziell \rightarrow allgemein

(formallogisch unbeweisbarer, oft lebensnotwendiger Schluß)

$$\gamma_A, \gamma_B \vdash \gamma_V$$

Abduktion

Folgerung \rightarrow hinreichende Voraussetzung

(formallogisch unbeweisbarer, oft unhaltbarer Schluß)

$$\gamma_V, \gamma_B \vdash \gamma_A$$

Induktives Lernen

Philosophisches Reizthema eines Jahrtausends

Francis Bacon (1561–1626)

Relevanz positiver *und* negativer Lernbeispiele

John Stuart Mill (1806–1873)

Vier Methoden für den praktischen Induktionsschluß

Bertrand Russell (1872–1970)

Induktionsschluß ist Grundlage jeglicher Vorhersage, nicht beweisbar und essentiell probabilistischer Natur

Ludwig Wittgenstein (1889–1951) Tractatus Logico-Philosophicus

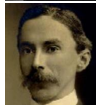
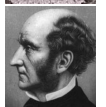
„Suche das einfachste Gesetz, das mit den Fakten harmoniert“

William von Ockham (1285–1347)

Occam's Razor: „*Pluralitas non est ponenda sine necessitate*“

Jorma Rissanen (*1932) 'minimum description length'-Prinzip

MDL \rightsquigarrow minimale Summe codierender & korrigierender Bits



Paradigmen maschinellen Lernens

Der „Lehrer“ befiehlt / demonstriert / präsentiert / fehlt

Lernen aus Instruktionen

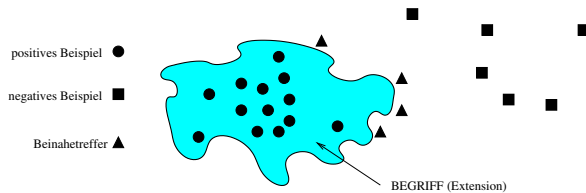
Natürlichsprachliche Systeme · Automatisches Programmieren

Lernen durch Analogiebildung

Wissentransfer auf neue, aber strukturell verwandte Aufgabenstellung

Lernen aus Beispielen (induktiv)

Beispiele, Gegenbeispiele und Beinahetreffer eines Begriffs



Lernen aus Beobachtung (explorativ)

Strukturieren von Objektmengen: $\left\{ \begin{matrix} \text{passiv} \\ \text{aktiv} \end{matrix} \right\} \triangleq$

$\left\{ \begin{matrix} \text{Datenquelle} = \text{Prozeßbeobachtung} \\ \text{Interaktion Lernprogramm-Umwelt} \end{matrix} \right\}$

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Konzeptuelles Lernen

Lernen eines Begriffs — wo kommen die benötigten Lernbeispiele (\pm) her ?

Assistiertes Lernen

Handverlesene Auswahl von \oplus/\ominus -Beispielen

➔ Optimaler Lernerfolg durch kompetenten Reiseführer

Lernen mit Orakel

Lernprogramm wählt *interessante* neue Beispiele

Orakelbefragung liefert \oplus/\ominus -Information

➔ Entdeckungsreise zu den Grenzfällen

Überwachtes Lernen

Beispiele wie vom natürlichen Erzeugungsprozeß produziert

Lehrer vergibt (die korrekten) \oplus/\ominus -Etiketten

➔ Zufälliges Abrastern des Objektraums

Verstärkungslernen ('reinforcement learning')

Lernbeispiele liegen *unetikettiert* vor

Lehrer erteilt summarische Leistungsnote („Lob und Tadel“)

➔ Strategie zwischen Exploration & Exploitation

Beispiele induktiver Lernaufgaben

Aufgabenbereich · Leistungskriterium · Erfahrungsquelle

QUBIC ($4 \times 4 \times 4$ Tic Tac Toe)

AB — alle QUBIC-Partien gegen Bobby Fisher

LK — Prozentsatz aller *gewonnenen* Partien

EQ — die Möglichkeit, 3 Wochen gegen Fisher zu trainieren

Postanschriftenleser

AB — Erkenne Zielorte handgeschriebener Anschriften

LK — Prozentsatz korrekt sortierter Briefsendungen

EQ — 10^5 handadressierte Briefe mit bekanntem Zielort

Steuerung eines (auto-)mobilen Roboters

AB — selbständiges Manövrieren im öffentlichen Fernverkehr

LK — *Geschwindigkeit* / $(1 + \text{Karambolagen})^{1.000.000}$

EQ — 20 Minuten Bewegungsbilder mit Steuerkommandos

Natürlichsprachlicher Datenbankzugang

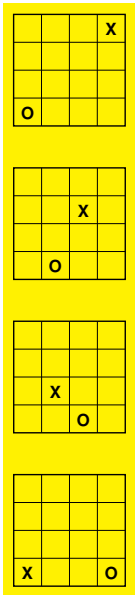
AB — autom. Beantwortung natürlichsprachlicher Datenbankfragen

LK — Prozentsatz korrekter Antworten

EQ — Texte natürlichsprachlicher Benutzeranfragen nebst SQL-Kodierung

Beispiel QUBIC

Dreidimensionales Tic tac toe · Kubus mit $4^3 = 64$ Feldern



Zielfunktion $eval^* : \mathcal{B} \mapsto [-100, +100]$

$$eval^*(\mathbf{b}) = \begin{cases} +100 & \text{wenn 4 X in einer Reihe} \\ -100 & \text{wenn 4 O in einer Reihe} \\ 0 & \text{wenn Remisstellung erreicht} \\ \mathcal{E}[\cdot] & \text{Erwartungswert der Endstellung} \\ & \text{bei optimaler Strategie} \end{cases}$$

Lösungsmodell (lineare Näherung für $eval^*$)

$$eval(\mathbf{b}) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_{10} x_{10} =: \mathbf{w}^T \mathbf{x}$$

mit den Prädiktorvariablen $x_i = x_i(\mathbf{b})$:

- $x_1(x_2) = \#$ offener Reihen mit einem X (O)
- $x_3(x_4) = \#$ offener Reihen mit zwei X (O)
- $x_5(x_6) = \#$ offener Reihen mit drei X (O)
- $x_7(x_8) = \#$ Schnittpunkte von X-Reihen (O-Reihen)
- $x_9(x_{10}) = \#$ Schnittpunkte s.o.; ≥ 2 X (O) je Reihe

Das Münchhausen-Prinzip

Was tun, wenn das Lösungsverfahren die Lösung selbst als Eingabe benötigt ?

Problem

Woher bekommen wir die benötigten Werte

$$eval^*(\mathbf{b}_t) = ?$$

Lösung

Vorwärtssuche mit der der Näherungsfunktion $eval(\cdot)$

$$\begin{aligned} eval^*(\mathbf{b}) &= \max\{eval^*(\mathbf{b}') \mid \mathbf{b}' \text{ Nachfolger von } \mathbf{b}\} \\ &\approx \max\{eval_w(\mathbf{b}') \mid \mathbf{b}' \text{ Nachfolger von } \mathbf{b}\} \end{aligned}$$

- Je besser die Näherung $eval(\cdot)$, desto genauer ist obige Approximation
- Wird dieses „bootstrapping“-Verfahren konvergieren?
- Welche Nachfolger von \mathbf{b} sollten betrachtet werden?
- Kann $eval^*(\cdot)$ überhaupt durch lineare Funktion angenähert werden?

Lernen der Stellungsbewertungsfunktion

Die Kenntnis von $eval^*(\cdot)$ ermöglicht eine optimale Zugauswahl

Benötigte Lernstichprobe

Partiestellungen $\mathbf{b}_1, \dots, \mathbf{b}_T$ mit bekannten Werten $y_t = eval^*(\mathbf{b}_t)$

Minimierung des Modellfehlers

Parameteroptimierung nach LSE-Prinzip („least squared error“)

$$\varepsilon = \sum_{t=1}^T \underbrace{(eval^*(\mathbf{b}_t) - eval(\mathbf{b}_t))^2}_{\varepsilon_t}$$

Iterative Lösung durch Gradientenabstieg

- 1 Initialisiere die Gewichte $w_0, w_1, w_2, \dots, w_{10}$
- 2 Führe je Lernbeispiel \mathbf{b}_t einen Verbesserungsschritt durch:

$$\mathbf{w}' = \mathbf{w} + \frac{2\beta \cdot (eval^*(\mathbf{x}_t) - \mathbf{w}^T \mathbf{x}_t)}{\|\mathbf{x}_t\|^2}$$

Dabei bezeichnet β die **Lernrate** des Verfahrens.

Beispiel: Konzeptuelles Lernen

Unter welchen Witterungsbedingungen empfiehlt sich ein Segelturn ?

GEGEBEN

- Objekte/Instanzen $\hat{=}$ mögliche Kalendertage
- Attribute/Prädikate $\hat{=}$ $\{sky, air, humidity, \dots\}$
- Zielfunktion $\hat{=}$ $gosailing : \mathcal{X} \mapsto \{T, F\}$

Lerndaten

Objekte mit allen Attributwerten & der Begriffzugehörigkeit:

#	sky	air	humidity	wind	water	forecast	gosailing
1	sunny	warm	normal	strong	warm	same	T
2	sunny	warm	high	strong	warm	same	T
3	rainy	cold	high	strong	warm	change	F
4	sunny	warm	high	strong	cold	change	T

Beispiel: Konzeptuelles Lernen

Induktion als Versuch der Datenbeschreibung mit unzureichenden Mitteln

GESUCHT

Passende Hypothese $h \in \mathcal{H}$ aus geeignetem Repräsentationenraum.

- Hypothesenraum $\mathcal{H} \hat{=}$ Konjunktionen von Attribut-Wert-Paaren
(z.B. $sky = sunny \wedge water = cool$)
- Lerndaten $\hat{=}$ positive und negative Beispiele
- Optimale Vorhersage der Urteile $gosailing(.)$ durch h

Postulat des induktiven Lernens

Wenn Hypothese h approximiert Zielfunktion auf (großer) Lernstichprobe

Dann Hypothese h approximiert Zielfunktion auf bislang unbeobachteten Beispielen

Repräsentationsformalismen

für **Datenobjekte** · zugrundeliegende **Begriffe** · gelernte **Hypothesen**

Parametersätze Diskriminanten, Neuronetze, Verteilungsfamilien

Formale Sprachen reguläre Ausdrücke, endliche Automaten, CFG

Produktionsregeln IF-THEN-Regeln, Assoziationen

Logik Aussagen-/prädikatenlogische Formeln, Klauselmengen

Graphen Semantische Netze, Drahtmodelle, Bayes/Markovnetze

Relationen Totale-, partielle- und Intervallordnungen

Frames Attribut-Wert-Paare, Dämonen, Defaults

Prozeduralformen Programme, Operatoren

Hierarchien Taxonomien, Partitionen, Entscheidungsbäume

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Intensionale Repräsentationen

Endliche(!) formalsprachliche Beschreibung unendlicher(!) Gesamtheiten

Logische Formeln

$elefant(x) \Leftrightarrow grau(x) \wedge groß(x) \wedge hat(x, Rüssel) \wedge ist(x, nachtragend) \wedge \neg frißt(x, Rollmops)$

Programme, Algorithmen

```
proc prim (nat n) bool:
  for i from 2 to sqrt(n) do
    if mod(n,i) = 0 then return false fi
  od
  return true
```

Grammatiken

$S \rightarrow NP VP$
 $NP \rightarrow N \mid Det N$
 $VP \rightarrow V \mid VP NP$
 $N \rightarrow John \mid Mary$
 $V \rightarrow loves$

Räumliche Strukturen

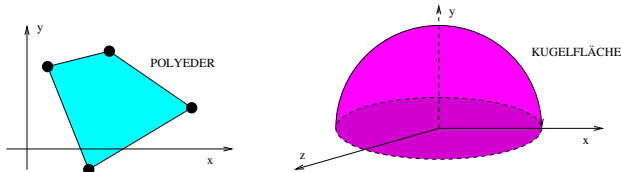
Kontinuum geometrischer Punkte als Lösung einer parametrisierten Gleichung

Polyeder

Drahtmodelle im \mathbb{R}^n :

$$(x_{(1)}, \dots, x_{(m)}), \quad x_{(i)} \in \mathbb{R}^n$$

z.B. ein Viereck $((x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4))$, $x_i, y_i \in \mathbb{R}$, in der Ebene



Punkte auf einer Hyperfläche

z.B. auf einer \mathbb{R}^3 -Sphäre mit Radius r :

$$\mathbf{x} = (r \cos \theta, r \sin \theta, r \cos \omega), \quad \theta, \omega \in [0, 2\pi]$$

Bäume

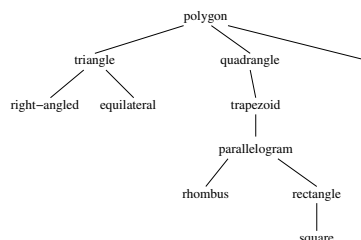
Zyklusfreie zusammenhängende ungerichtete Graphen bzw. ...

Definition

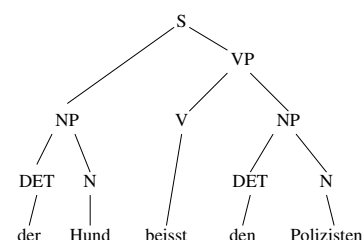
Der gerichtete Graph $\mathcal{G} = (U, L)$ heißt **Baum**, falls gilt:

1. \mathcal{G} ist einfach zusammenhängend.
2. Ex. genau ein **Wurzelknoten** $u_0 \in U$ ohne Vorgängerknoten.
3. Alle $u \in U \setminus \{u_0\}$ besitzen *genau einen* Vorgängerknoten.

Knoten *ohne* Nachfolgerknoten heißen **Blattknoten**.



Taxonomie geometrischer Objekte



Grammatischer PS-Ableitungsbaum

Graphen

Ungerichtet · Gerichtet · Markiert · Gewichtet

Ungerichteter Graph $\mathcal{G} = (U, L)$

$U \triangleq$ Knotenmenge

$L \triangleq$ Kantenmenge, $L \subseteq \{\{u, v\} \mid u, v \in U\}$

Gerichteter Graph $\mathcal{G} = (U, L)$

$U \triangleq$ Knotenmenge

$L \triangleq$ Kantenmenge,

$L \subseteq \{(u, v) \mid u, v \in U\} = U \times U$

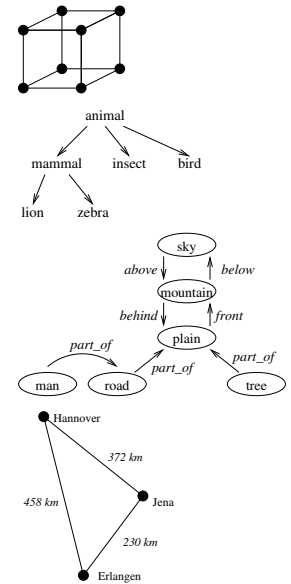
Markierter Graph $\mathcal{G} = (U, L, \ell)$

$A \triangleq$ Symbolvorrat, Alphabet der Markierungen

$\ell \triangleq$ Kantenmarkierungsfunktion, $\ell : L \mapsto A$

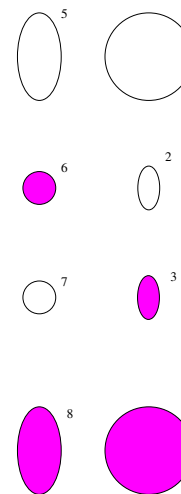
Gewichteter Graph $\mathcal{G} = (U, L, w)$

$w \triangleq$ Kantengewichtungsfunktion, $w : L \mapsto \mathbb{R}$



Listen

Geordnete Folge von (1) Listen oder (2) Symbolen aus Alphabet \mathcal{A}



Objektrepräsentationen

object 1:	((shape circle)	(size large)	(color white))
object 2:	((shape ellipse)	(size small)	(color white))
object 3:	((shape ellipse)	(size small)	(color pink))
object 4:	((shape circle)	(size large)	(color pink))
object 5:	((shape ellipse)	(size large)	(color white))
object 6:	((shape circle)	(size small)	(color pink))
object 7:	((shape circle)	(size small)	(color white))
object 8:	((shape ellipse)	(size large)	(color pink))

Verschachtelte Darstellungen

((object1	((shape circle)	(size large)	(color white))
	(object2	((shape ellipse)	(size small)	(color white))
	(object3	((shape ellipse)	(size small)	(color pink))
	(object4	(... ..)
)

Spezialfälle

Bäume \triangleq Listen ohne Nachfolgerordnung

Zeichenketten \triangleq flache Listen

„Sein oder Nichtsein ...“ oder „GACCTTATAGCT...“

Logische Repräsentationen

Aussagenlogik · Prädikatenlogik · Modal- und Zeitlogik

Hornklausel

(Disjunktive) Klausel mit *höchstens einem* positiven Literal

$$\neg P_1 \vee \dots \vee \neg P_m \vee Q \quad \text{oder} \quad \neg P_1 \vee \dots \vee \neg P_m$$

Schreibweise: «Kopf» \leftarrow «Rumpf»

$$\begin{array}{ll} Q \leftarrow P_1, P_2, \dots, P_m & (\text{allg.}) \\ \leftarrow P_1, P_2, \dots, P_m & (\text{Zielklausel}) \\ Q \leftarrow & (\text{Faktenklausel}) \\ \leftarrow & (\text{leere Klausel}) \end{array}$$

Beispiel

```
female(angela)
male(franz)
mutual_love(franz, angela)
can_marry(x1, x2) ← mutual_love(x1, x2), female(x1), male(x2)
```

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Prozedurale Repräsentationen

Imperative Formen · „if/then“-Regeln · Produktionsregeln

Beispiel

Imperative Darstellung einer Objektbeschreibung der Robotik:

„die kleine rote Schachtel steht auf der großen schwarzen Schachtel“

```
make_on (x,y) {
    cleartop (x);
    cleartop (y);
    puton (x,y);
}
puton (x,y) {
    STORE <on (x,y)>;
}
cleartop (x) {
    for all y DELETE <on (y,x)>;
}
```

Was ist Data Mining ?

... und warum wird seit Beginn des Jahrtausends so viel darüber geredet ?

*„Data Mining is the exploration and analysis,
by automatic or semi-automatic means,
of large quantities of data
in order to discover meaningful patterns and rules.“*

Woher kommt der aktuelle Boom ?

- Massenproduktion von Daten
- Präsentation in *data warehouses*
- Rechnerleistung verfügbar
- Kommerzielle Datamining-Software erhältlich
- Starker Konkurrenzdruck

KDD — Knowledge Discovery in Databases

„We are drowning in information, but we are starving for knowledge.“ (John Naisbett 1996)

Was sind Daten?

- einzelne Objekte
- individuelle Merkmale
- riesige Fallzahlen
- verwirrende Vielfalt
- preiswert zu beschaffen

⊖ **Voraussagen**

Tycho Brahe (1546–1601)

Massendatensammlung zu den Umlaufbahnen der Himmelskörper unseres Planetensystems
geozentrische Koordinaten

Was ist Wissen?

- Klassen von Objekten
- globale Muster
- allgemeine Gesetze
- einfache Prinzipien
- schwer zu bekommen

⊕ **Voraussagen**

Johannes Kepler (1571–1630)

1. Umlaufbahnen sind elliptisch
2. Laufzeit \propto Sektorfläche
3. Umlaufperiode² \propto Großradius³

Was ist das Analyseziel ?

Abstrakter Datensatz $\hat{=}$ Relation (Objekte \times Attribute)

Gruppierung

Partitionierung der Datenobjekte in Häufungsgebiete

Klassifikation

Zuordnung von Datenobjekten zu Kategorien

Dependenzstruktur

Aufdecken der Abhängigkeiten zwischen den Objektattributen

Prädiktion

Vorhersage (noch) nicht verfügbarer Objektattribute

Selektion und Assoziation

Erkennung von Auffälligkeiten & Regelmäßigkeiten

Typische Datenquellen

Industrielle Prozeßdaten

Analyse der Altpapieraufbereitung bei Kübler+Niethammer
8 Deinkingzellen à 54 Sensoren à 9000 Meßwerte/Tag ➡ 3.888.000 Mw/T

Umsatzdatenbanken

Warenkorbanalyse für die Scannerkassen bei WalMart
20 Millionen Transaktionen/Tag ➡ Datenbank 24 Terabytes

Molekularbiologie

Human Genome Database Project
Entschlüsselung des genetischen Codes des Menschen
60 000–80 000 Gene ➡ 3 Milliarden DNA-Basen

Visuelle Daten

NASA Earth Observing System sammelt
Oberflächenbilder tieffliegender Satelliten ➡ 50 Gigabytes/Stunde

Textinformationen

Ca. 10 Milliarden HTML-Seiten im World Wide Web
Suchmaschinen, Indexierer, Extrahierer, Emailfilter

Anwendungsbedarf nach Industriezweigen

Großhandel · Finanzen · Telekommunikation · Verkehr · Gesundheit

Fälschungssicherheit

Mobilfunk — 'cloning' der Geräteerkennung
Kreditkartenmißbrauch — physikalisch/elektronisch
Rechnermißbrauch — Angriff, Einbruch

Kreditwesen

Kreditwürdigkeit, Zahlungsfähigkeit
Risikokapital, Unternehmenssolvenz
Anlageberatung

Kundenbetreuung

Kundenbindung (Beispiel: 5% Reduktion der Fluktuation ➡ 200% Gewinn)
Direktmarketing (Handel, Bank, Versicherung)
Warenkorbanalyse im Einzelhandel

Beispiel Prozeßautomatisierung

Industrielle Herstellung von ICE-Türen aus Verbundwerkstoffen

Fertigungszelle

Prozeßkettenmodell $\hat{=}$ Workflow mit aktiven & passiven Komponenten:

- **Meßwerte** erfassen + auswerten \Rightarrow Sensoren
- **Stellgrößen** berechnen + anlegen \Rightarrow Aktoren

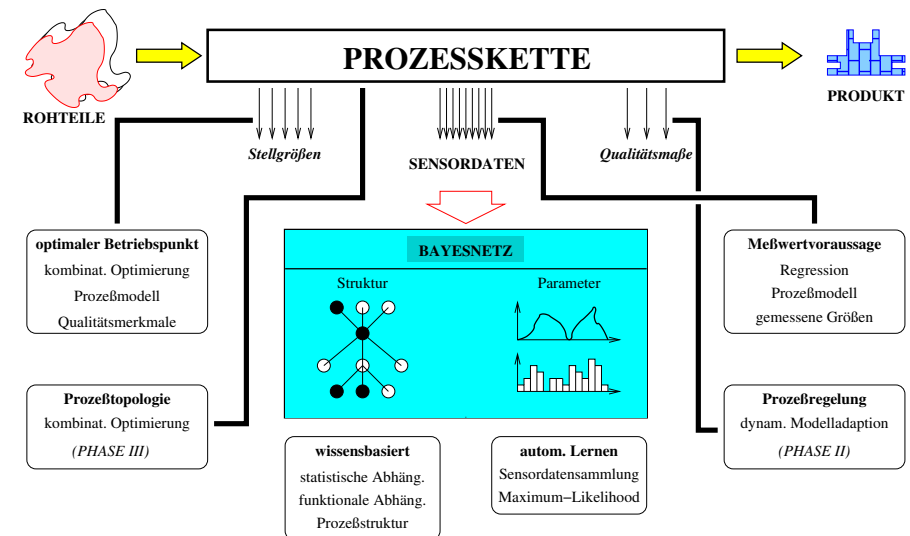
Produktionsoptimierung

Statt Erfahrung, Daumenregel und Intuition ...

- Prozeßvisualisierung
- Entscheidungsunterstützung
- Automatische (adaptive) Regelung
- Optimale Strukturierung der Prozeßkette

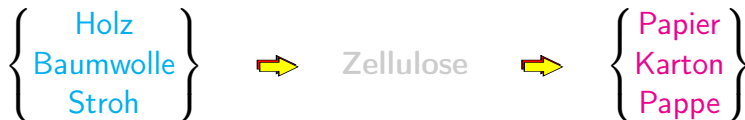
Beispiel Prozeßautomatisierung

Stochastischer Abhängigkeitsgraph zur Vorhersage optimaler Stellgrößen



Beispiel Prozeßautomatisierung

Automatisierung in der Papierindustrie



Industrielle Arbeitsschritte

1. Kocher chemischer Aufschluß, Bleichung
2. Flotationszelle lösen, vorsortieren, entfärben
3. Refiner Fasern mahlen
4. Pulper Wasser zusetzen (Suspension)
5. Trockner Bandsieb, Pressung (Tambouren)
6. Cutter zuschneiden, aufstapeln

Prozeßdatenerhebung

Automatisierung in der Papierindustrie

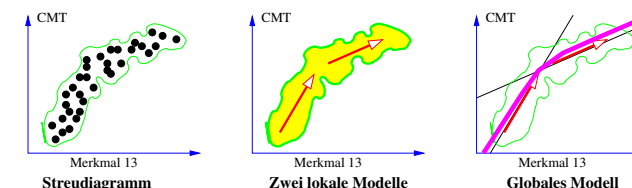
Zielgröße Papierqualität

Concora Medium Test

$$\text{CMT} \stackrel{\text{def}}{=} \text{„Gewicht“} / \text{„Festigkeit“}$$

26 Stellgrößen und Meßwerte

Druck, Temperatur, Menge, Gewicht, Qualität von Rohstoffen und Zwischenprodukten



Elliptotype-Cluster mit $x_{27} = 1.56 \cdot x_{13} + 0.32$ und $x_{27} = 0.60 \cdot x_{13} + 0.48$

Ablauf des Datamining-Prozesses

Automatisierung in der Papierindustrie

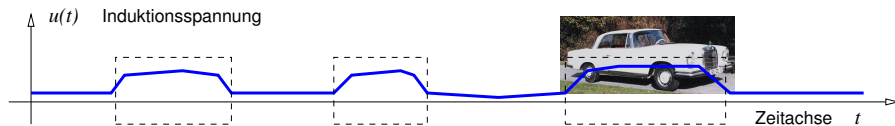
(Algorithmus)

- 0 LAUFZEITBEREINIGUNG
Transformation **physikalischer** Zeit t an Prozeßstation P_i via $\tau = t + \Delta t_i$
Meßwertvektoren $\tilde{\mathbf{x}}_t \in \mathbb{R}^{27} \rightsquigarrow$ Fälle $\mathbf{x}_\tau \in \mathbb{R}^{27}$ mit **synchronisierter** Referenzzeit
- 1 DATENSATZBEREINIGUNG
Ungültige Einträge markieren · Ausreißer nach 4σ -Regel markieren
Fälle mit markierten Werten tilgen
- 2 NORMIERUNG
Jedes der 27 Merkmale wird auf $\mathcal{N}(0, 1)$ normiert.
- 3 DEPENDENZANALYSE
Untersuche Abhängigkeiten der Form (x_i, x_{27}) und (x_i, x_j, x_{27}) .
- 4 REGRESSIONSANSATZ
Linear oder stückweise linear · zwei Elliptotype-Cluster
- 5 REGELERZEUGUNG
Überlagerung lokaler Modelle · Zugehörigkeitsfunktion \rightsquigarrow Regelprämisse

(zumFolienA)

Beispiel Verkehrsplanung und -lenkung

Dienstgüteanalyse der Verkehrszustände auf Autobahnstrecken



Meßverfahren

- **Meßwertreihe** $u(t)$ Induktionsspannung
Impulsfunktion der Induktionsschleife auf der Fahrbahn
- **Verkehrsstärke** q Fahrzeuge/Stunde
Zählung der Anzahl q von Impulsen (in $[1/h]$)
- **Streckenbelegung** β Zeitanteil
Summe der Impulsbreiten $\beta = \frac{1}{u_{\max} \cdot \Delta T} \int_T^{T+\Delta T} u(t) dt$
- **Verkehrsdichte** ρ Fahrzeuge/Kilometer
 $\rho \approx \rho_{\max} \cdot \beta$ und gleichzeitig auch $q \approx \bar{v} \cdot \rho$, aber ρ_{\max} und \bar{v} unbekannt

Vernetzte Systeme

Datenanalyse in granularen Transportsystemen

Aufgabenstellungen

- **Monitoring** · Erfassung des aktuellen Zustandes
- **Modellierung** · Gesetzmäßigkeiten in Transportströmen
- **Prognose** · Vorhersage der Netzbelastung
- **Routing** · Bestimmung optimaler Wege
- **Optimierung** · Verbesserung des Netzzustandes/Netzflusses

Anwendungsgebiete

- Güter- und Personenverkehr
- Telekommunikation
- Energieversorgung
- Rohstoffzufuhr im Fertigungsprozeß

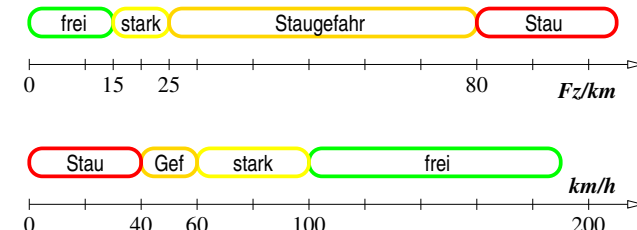
Beispiel Verkehrsplanung und -lenkung

Verkehrsflussmodell und Dienstgütestufen

Mathematisches Verkehrsflussmodell

Den Idealfall einer funktionalen Abhängigkeit $q(\rho) = v(\rho) \cdot \rho$ liefert:

$$v(\rho) = v_0 \cdot \rho \cdot \left(1 - (\rho/\rho_{\max})^{\ell-1}\right)^{\frac{1}{1-m}}$$

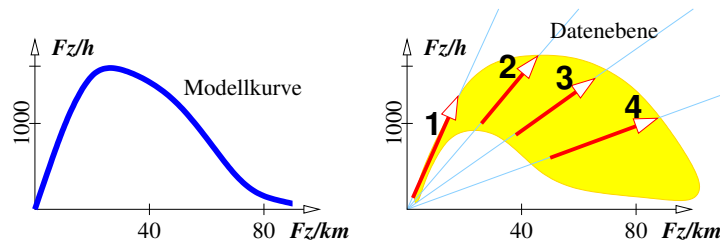


Dienstgütestufen („levels of service“)

- 1 freier Verkehr · 2 starker Verkehr · 3 Staugefahr · 4 Stau

Beispiel Verkehrsplanung und -lenkung

Modellierung und Interpretation der Meßdatensätze



Tagesgangkurven

Viertelstündige Verkehrsstärkemessung

Mediangelättung · Datensätze für Wochenkerntage

Clustering in drei prototypische Gruppen:

1 Urlaubstag · 2 Durchschnittstag · 3 Großveranstaltungstag

96 Werte/Tag
 $M = 5$; Mo, Di, Mi, Do

Struktur der (ρ, q) -Datenebene

Konzentrische Geradenstücke $\hat{=}$ Verkehrssituationen gleicher Geschwindigkeit

4 Dienstgütern \rightarrow konzentrische Längscluster

Beispiel Marketing

Aktive Orientierung an Kundenwünschen \leadsto Wettbewerbsvorteil

Relationale Datenbank eines Versandhauses

Kundentabelle KuNr, PLZ, GJ (Geburtsjahr), ...

Umsatztablelle BestNr, KuNr, Betrag, ...

Datamining-Schritte

Clusteranalyse der Verbundtabelle

$$(PLZ, GJ, Umsatz) \in \mathbb{R}^3$$

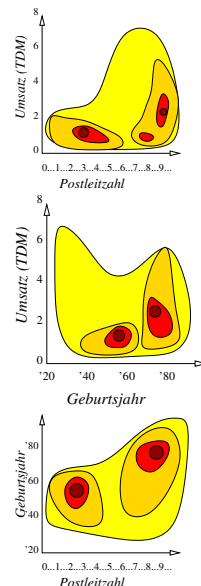
Gewichteter euklidischer Abstand

$$g = (10^{-5}, 10^{-2}, 10^{-4})$$

$$\mu^{(1)} = \begin{pmatrix} 27\,374 \\ 1954.16 \\ 1122.44 \end{pmatrix}, \quad \mu^{(2)} = \begin{pmatrix} 86\,356 \\ 1969.35 \\ 1618.99 \end{pmatrix}$$

Risiken und Nebenwirkungen

„Alter“ \rightarrow „Geburtsdatum“ \rightarrow „1.1.1970“



Beispiel Marketing

Welche Datamining-Methoden für welche Fragestellung ?

Segmentierung

Welche Idealtypen von Kunden besitzt die Firma?

Klassifikation

Ist die konkrete Person ein potentieller Neukunde?

Konzeptualisierung

Welche Attribute charakterisieren ein Kundensegment?

Prädiktion

Welcher Umsatz ist im Folgejahr zu erwarten?

Deviation

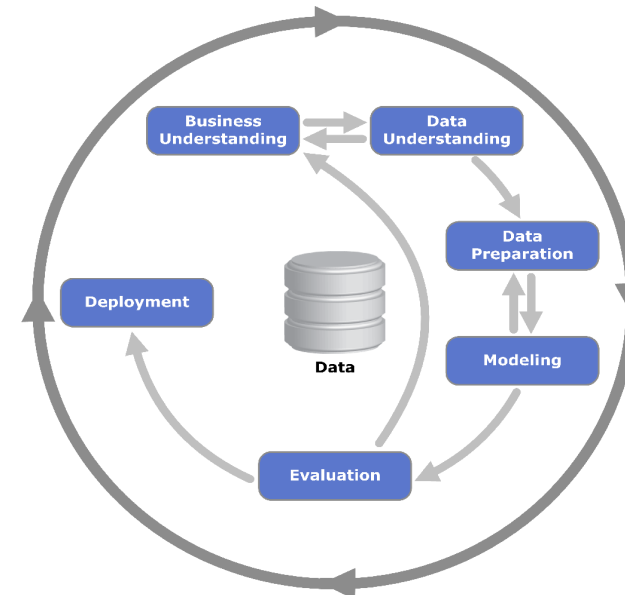
Wo und warum ist Kundenverhalten verändert?

Dependenz

Wie beeinflusst eine Marketingaktion das Kundenverhalten?

Cross-Industry Standard Process for Datamining

CRISP-DM (NCR & Daimler & SPSS/IBM)



SEMMA (SAS)

Sample
 Explore
 Modify
 Model
 Assess

WEKA et al.

data acquisition
 data preprocessing
 data modeling
 data evaluation

Datamining-Projekte

Arbeitsphasen & Grundbausteine eines Datamining-Prozesses

Materialbeschaffung (I)

Planung
Datensammlung
Merkmalberechnung
Datenauswahl



Vorverarbeitung (II)

Normierung
Säuberung
Filterung
Ergänzung
Korrektur



Strukturanalyse (III)

Korrelation
Regression
Modellierung
Klassifikation
Gruppierung



Auswertung (IV)

Visualisierung
Interpretation
Dokumentation



Kommerzielle Softwaresysteme

Anwendungsspezifische Werkzeuge — integrierte Speziallösung

Fälschungsschutz

HNC Falcon/Eagle, Neuraltech Nestor/Minotaur, Nestor

Kreditkontrolle

FairIsaacs, Sigma Analytics, Neuraltech Decider

Kundenbindung

SLP InfoWare, Neuraltech Churn Manager

Kundenprofil

HNC ProfitMax, Neuraltech Gold, RightPoint, AppliedMetrix

(Kommerzielle) Softwaresysteme

Allroundpakete — nicht anwendungsspezifisch, viele Werkzeuge

Paket (Anbieter)	Implementierte Methoden
Clementine (SPSS & IBM)	EB Reg MLP Ru1 kNN SOM Clus
Enterprise Miner (SAS)	EB Reg MLP Ru1 Seq Clus
Darwin (Thinking Machines)	EB MLP kNN
WEKA (OSS/FSW)	EB Reg MLP Ru1 SOM Clus
'R'-Projekt (OSS/FSW)	... das alles und noch viel mehr ...

EB Statistische Entscheidungsbäume (CART)

Reg Regressionsmodelle für Vorhersage & Kategorisierung

MLP Mehrschichtenperzeptron

Ru1 Assoziations- und Fuzzyregelsysteme

kNN k-nächster-Nachbar Klassifikation

SOM Selbstorganisierende Merkmalkarten

Clus (Hierarchische) Gruppierungsverfahren

Seq Statistische Zeitreihenanalyse

Kommerzielle Softwaresysteme

Methodenspezifische Werkzeuge — die Welt sieht aus wie ein Nagel ...

(Tiefe) Neuronale Netze

PittNet, NN/XNN, SNNS; TensorFlow, Caffe, Torch

Nächster-Nachbar-Klassifikator

SGI MLC++, Condor PEBLS

Abhängigkeitsanalyse

SGI MineSet, XPertRule Miner

Graphische Modelle

LEDa, LINK, ViCLAS, Precision Crimelink

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Zusammenfassung (1)

1. **Maschinelles Lernen** verknüpft empirische *Beobachtungen*, menschliches *Vorwissen* und überlegene *Rechnerleistung* zu einer neuen Qualität intelligenter Informationsverarbeitung.
2. **Induktives Lernen**, die Verallgemeinerung auf Basis von Einzelfällen, ist eine unverzichtbare, gleichwohl unbeweisbare Schlußtechnik.
3. Die **Lernbeispiele** zu einem **Begriff** und ihre **Etikettierung** werden vom **Lehrer** und/oder dem **Lernprogramm** vorgegeben.
4. Die Frage nach einer (geeigneten) **Repräsentation** stellt sich bei den präsentierten **Datenobjekten**, den zugrundeliegenden **Begriffen** („*Konzepten*“) und den zu lernenden **Hypothesen**.
5. Die Objektrepräsentation umfasst **numerische**, **symbolische**, **prozedurale**, **relationale** und **metrisch-topologische** Darstellungen.
6. Zur Lösung der Lernaufgabe wird ein **Erfolgskriterium** optimiert.
7. **Datamining** ist die (oft interaktive) Anwendung von ML-, Statistik- und Visualisierungsmethoden auf **große Datenbestände**.
8. Das Anliegen ist das Aufdecken von **Gruppenstrukturen** und **Abhängigkeiten**, das Ermitteln von **Kategoriezugehörigkeiten** sowie Vorhersage und Abgleich zukünftiger oder unzugänglicher **Attributwerte**.
9. Datamining ist ein **zyklischer Prozess** der Schritte **Akquisition**, **Bereinigung**, **Modellierung** und **Evaluierung**.