

tu technische universität dortmund LS 8 Informatik Computergestützte Statistik Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Vorlesung Wissensentdeckung

Einführung

Katharina Morik, Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

7.4.2015

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik Computergestützte Statistik Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Gliederung

- 1 Anwendungen Wissensentdeckung
- 2 Aufgaben der Modellbildung
- 3 Themen, Übungen

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik Computergestützte Statistik Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Bekannte Anwendungen

- Google ordnet die Suchergebnisse nach der Anzahl der auf sie verweisenden Hyperlinks an.
- Amazon empfiehlt einem Kunden, der A gekauft hat, das Produkt B, weil viele Kunden, die A kauften, auch B kauften.
- Der Markt wird beobachtet: wie äußern sich Verbraucher im WWW über ein Produkt? (Sentiment Analysis)
- Versicherungen bewerten ihre Produkte nach den Schadensfällen.
- Verkaufszahlen werden vorhergesagt (Lagerhaltung).
- Daten physikalischer Vorgänge werden analysiert, z.B. Terabytes von Messungen der Astrophysik.
- Verteilte Sensormessungen werden ausgewertet, z.B. zur Verbesserung der Navigationssysteme.

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik Computergestützte Statistik Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Das neue Paradigma: Sehr viele Daten!

20 Petabyte Daten werden bei Google täglich bearbeitet (2011).

- 1 Megabyte (MB) = 1024 Kilobyte = $1024 \cdot 1024$ Byte = 1.048.576 Byte
- 1 Gigabyte (GB) = 10^9 Bytes
- 1 Terabyte (TB) = 10^{12} Bytes
- 1 Petabyte (PB) = 10^{15} = 1.125.899.906.842.624 Bytes
- 1 Exabyte (EB) = 10^{18} Bytes

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik Computergestützte Statistik Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Wikipedia

Wikipedia bietet (30.März 2012)

- 9.239.223 Artikel auf Englisch, 2.323.504 auf Deutsch
- 270 Sprachen insgesamt.
- 9.953.252 mal pro Stunde wird ein Wikipedia Artikel auf Englisch angeschaut, 1.374.452 in der Stunde einer auf Deutsch.
- Aktuelle Statistik:
<http://stats.wikimedia.org/DE/Sitemap.htm>

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik Computergestützte Statistik Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Soziale Netzwerke

- FaceBook verbindet (1.1.2012) 157.412.000 Nutzer in den USA, 22.124.000 in Deutschland.
- Gezählt werden Nutzer, die sich innerhalb der letzten 30 Tage mindestens einmal am entsprechenden Ort eingeloggt haben. Unternehmensaccounts zählen nicht.
- Berlin: 1,32 Millionen
Hamburg: 0,72 Millionen
München: 0,85 Millionen
Frankfurt am Main: 0,62 Millionen
Köln: 0,60 Millionen
Dortmund: 0,31 Millionen
- <http://allfacebook.de/news/facebook-nutzerzahlen-2012-in-deutschland-und-weltweit>

Katharina Morik, Claus Weihs DMV

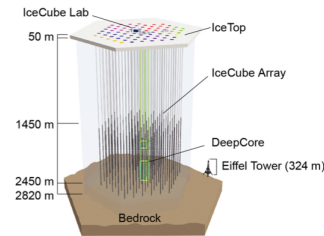
Die Masse macht's!

Googles Philosophie ist: wir wissen nicht, warum diese Seite besser als eine andere ist. Aber wenn viele Menschen das meinen, ist es so.

- Statistik eingehender Verweise auf eine Seite zeigt die Bedeutung, die die verweisenden Menschen der Seite beimessen.
- Peter Norvig, bekannt durch sein Einführungsbuch in die Künstliche Intelligenz (1995, mit Bertrand Russell), seit 2001 bei Google, seit 2006 Google's research director, sagte in einer Rede auf der O'Reilly Emerging Technology Conference 2008: "All models are wrong, and increasingly you can succeed without them." (wired 2008)
- Data Mining soll Massen an Daten indexieren, sortieren, strukturieren, klassifizieren, darin Muster finden, interessante Unterräume bestimmen.

Astroteilchenphysik – IceCube

Die Masse macht's – sehr viele Messungen sind nötig, um ein Neutrino zu fangen!



Neutrinos sind elektrisch neutrale Elementarteilchen mit sehr kleiner Masse. Da sie fast ungeschwächt durch alles (z.B. die Erde) hindurchgehen, können sie Aufschluss über Vorgänge im All geben, Geburt von Sternen, Supernovae... Sie können mit Tscherenkow-Licht-Detektoren im Eis erfasst werden.

Trennung von Signal und Rauschen

Signal Atmosphärisch ~ 14180 Ereignisse in 33,28 Tagen bei IC-59

Hintergrund Falsch rekonstruierte Muons ~ $9,699 \cdot 10^6$ Ereignisse in 33,28 Tagen bei IC-59

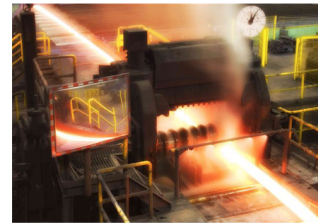
Verhältnis $1,46 \cdot 10^{-3}$

Methode Monte Carlo Simulation ergibt klassifizierte Beobachtungen. Nur relevante Merkmale der Beobachtungen (477) verwenden. RandomForest-Lerner trainieren auf der Simulation, anwenden auf reale Messreihen.

T. Ruhe, K. Morik, W. Rhode

Stahlindustrie

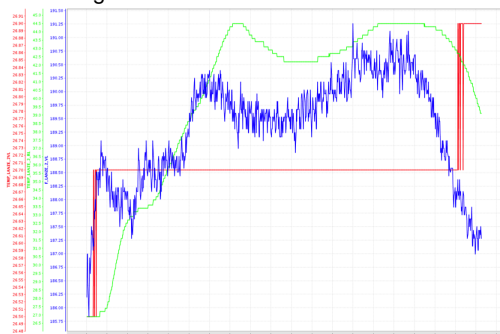
Ressourcenschonung: Kürzerer Brennprozess, niedrigere Temperatur (<1700 Celsius), weniger Metallgabe!



- Online Prognose im Betrieb zur verbesserten Steuerung
 - Temperatur,
 - Fe-Gehalt der Schlacke,
 - Phosphor,
 - Kohlenstoff
- Patent mit Siemens angemeldet.

Prognose anhand von Messreihen

Merkmale aus Zeitreihen extrahieren und dann für das Lernen einer Prognose nutzen.



Navigation

Floating Car Data



- Mobiltelefone mit GPS Empfängern zur Erzeugung von FCD
- Kartengenerierung durch Datenfusion über alle Teilnehmer
- Karte bereits nach wenigen Eingangsdaten übereinstimmend mit Basiskarte

technische universität dortmund

LS 8 Informatik
 Computergestützte Statistik
 Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Sprachtechnologie

Maschinelles Lernen bzw. Statistik macht den Erfolg:

- Suchmaschinen: 100% der verbreiteten Suchmaschinen sind probabilistisch. Die Retrieval Funktion beinhaltet Gewichte, die antrainiert werden müssen.
- Spracherkennung: 100% der verbreiteten Systeme sind probabilistisch – wahrscheinlichste Laufolgen.
- Question answering: Das IBM Watson System, das im Februar 2011 in drei Runden den Quiz Jeopardy gegen zwei Champions gewann, basiert auf maschinellem Lernen und anderen KI-Techniken.
- Part of speech tagging: Die meisten Systeme sind statistisch. Der Brill tagger ist hybrid: er lernt eine Menge deterministischer Regeln aus Daten.
- Parsing: die Mehrheit ist probabilistisch und wird auf eine Sprache hin trainiert.

Katharina Morik, Claus Weihs DMV

technische universität dortmund

LS 8 Informatik
 Computergestützte Statistik
 Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Interesse an Anwendungen

- Werbung soll besser auf die Interessierten zugeschnitten sein und nur an diese gesandt werden.
- Business Reporting soll automatisiert werden. On-line Analytical Processing beantwortet nur einfache Fragen. Zusätzlich sollen Vorhersagen getroffen werden.
- Wissenschaftliche Daten sind so umfangreich, dass Menschen sie nicht mehr analysieren können, um Gesetzmäßigkeiten zu entdecken.
- Geräte sollen besser gesteuert werden, um Ressourcen zu schonen. Industrie 4.0

Katharina Morik, Claus Weihs DMV

technische universität dortmund

LS 8 Informatik
 Computergestützte Statistik
 Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Berufsaussichten

Zur Zeit sind Hochschulabsolventen, die die Datenanalyse beherrschen, in vielen Branchen sehr nachgefragt! Beispiele für Firmen und Institutionen, die von Datenanalyse leben:

- Recommind, Rheinbach (Köln), mehr als 2000 Mitarbeiter weltweit
- PrudSys, Chemnitz, Veranstalter des jährlichen Data Mining Cups
- RapidMiner, Dortmund und Boston, mehr als 50 Mitarbeiter, wachsend
- Fraunhofer, St. Augustin, Intelligente Analyse- und Informationssysteme, mehr als 200 Wissenschaftler

Katharina Morik, Claus Weihs DMV

technische universität dortmund

LS 8 Informatik
 Computergestützte Statistik
 Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Rexer Analytics Umfrage zu Datenanalyse

- Fragebogen mit 40 Fragen
- 710 Antworten aus 58 Ländern
- Benutzer von
 - IBM SPSS Modeler,
 - Statistica und
 - RapidMiner
 sind am zufriedensten mit ihrer Software.
- Die am häufigsten benutzten Algorithmen sind
 - Regression,
 - Entscheidungsbäume,
 - Clustering.

Katharina Morik, Claus Weihs DMV

technische universität dortmund

LS 8 Informatik
 Computergestützte Statistik
 Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Datenanalyse – generische Aufgabe

Population: Eine Menge von Objekten, um die es geht.

Merkmale: Eine Menge von Variablen (quantitativ oder qualitativ) beschreibt die Objekte.

Ausgabe: Ein quantitativer Wert (Messwert) oder ein qualitativer gehört zu jeder Beobachtung (Zielvariable).

Ein **Lernverfahren** findet eine Funktion, die Objekten einen Ausgabewert zuordnet. Oft **minimiert** die Funktion einen **Fehler**.

Modell: Das Lernergebnis (die gelernte Funktion) wird auch als *Modell* bezeichnet.

Katharina Morik, Claus Weihs DMV

technische universität dortmund

LS 8 Informatik
 Computergestützte Statistik
 Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Notation

- Der Raum möglicher Beobachtungen wird als p -dimensionale Zufallsvariable X geschrieben.
- Jede Dimension der Beobachtungen wird als X_i notiert (Merkmal).
- Die einzelnen Beobachtungen werden als $\vec{x}_1, \dots, \vec{x}_N$ notiert.
- Die Zufallsvariable Y ist die Ausgabe (Zielvariable).
- N Beobachtungen von Vektoren mit p Komponenten ergeben also eine $N \times p$ -Matrix.

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Lernaufgabe Clustering

Gegeben

- eine Menge $\mathcal{T} = \{\vec{x}_1, \dots, \vec{x}_N\} \subset X$ von Beobachtungen,
- eine Anzahl K zu findender Gruppen C_1, \dots, C_K ,
- eine Abstandsfunktion $d(\vec{x}, \vec{x}')$ und
- eine Qualitätsfunktion.

Finde

- Gruppen C_1, \dots, C_K , so dass
- alle $\vec{x} \in X$ einer Gruppe zugeordnet sind und
- die Qualitätsfunktion optimiert wird: Der Abstand zwischen Beobachtungen der selben Gruppe soll minimal sein; der Abstand zwischen den Gruppen soll maximal sein.

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Lernaufgabe Klassifikation

Gegeben

- Klassen Y , oft $y \in \{+1, -1\}$,
- eine Menge $\mathcal{T} = \{(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)\} \subset X \times Y$ von Beispielen,
- eine Qualitätsfunktion.

Finde

- eine Funktion $f : X \rightarrow Y$, die die Qualitätsfunktion optimiert.

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Lernaufgabe Regression

Gegeben

- Zielwerte Y mit Werten $y \in \mathcal{R}$,
- eine Menge $\mathcal{T} = \{(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)\} \subset X \times Y$ von Beispielen,
- eine Qualitätsfunktion.

Finde

- eine Funktion $f : X \rightarrow Y$, die die Qualitätsfunktion optimiert.

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Funktionsapproximation

Wir schätzen die wahre, den Beispielen unterliegende Funktion. Gegeben

- eine Menge von Beispielen $\mathcal{T} = \{(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)\} \subset X \times Y$,
- eine Klasse zulässiger Funktionen f_θ (Hypothesensprache),
- eine Qualitätsfunktion,
- eine feste, unbekannte Wahrscheinlichkeitsverteilung $P(X)$.

Finde

- eine Funktion $f_\theta : X \rightarrow Y$, die die Qualitätsfunktion optimiert.

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Themen

- Finden häufiger Mengen
- statistische Grundlagen
- lineare Modelle
- Stützvektormethode (SVM)
- Klassifikation
- Entscheidungsbäume
- Versuchsplanung, Stichproben
- Dimensionsreduktion, Merkmalsselektion
- Clustering
- Zeitreihen

Katharina Morik, Claus Weihs DMV

tu technische universität dortmund LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

Anwendungen Modellbildung – Aufgaben Vorlesungsablauf

Literatur

Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Springer series in statistics. Springer, New York, USA, 2001.

Gerald Teschl and Susanne Teschl.
Mathematik für Informatiker.
Springer, 2006.

Katharina Morik, Claus Weihs DMV