Vorlesung Wissensentdeckung

Einführung

Katharina Morik, Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

3.4.2010

Gliederung

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.

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

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

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

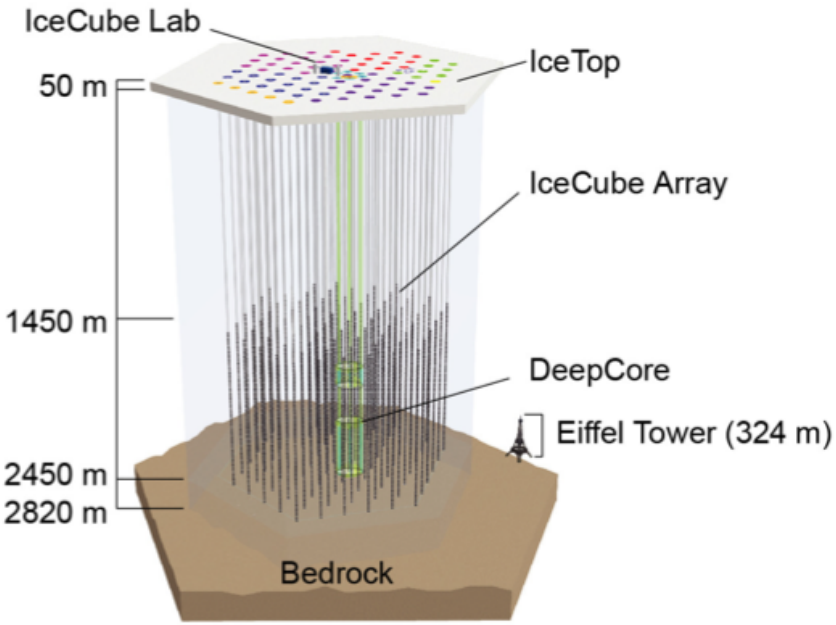
Die Masse macht's!

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 unabgeschwächt 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.

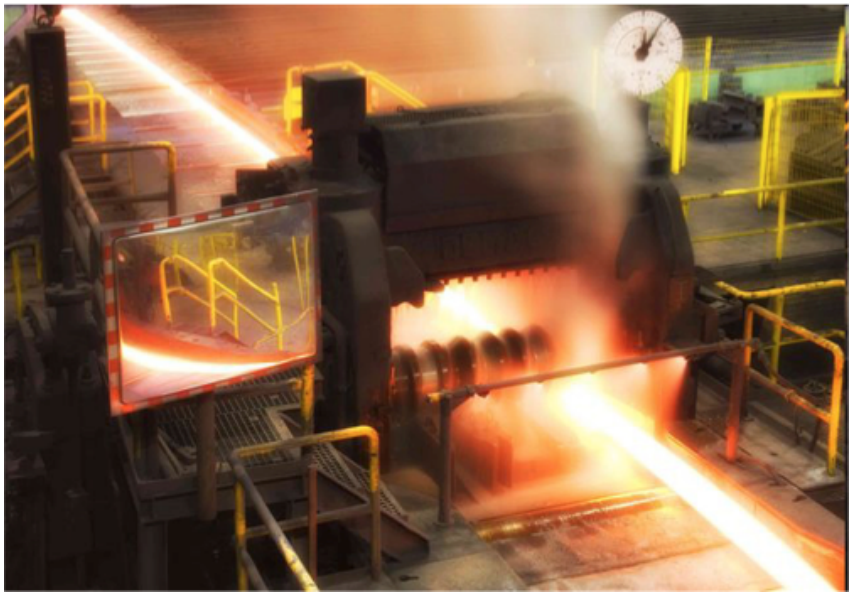
Trennung von Signal und Rauschen

Signal	Atmosphärisch ~ 14180 Ereignisse in 33,28 Tagen bei IC-59
Hintergrund	Falsch rekonstruierte Muons $\sim 9,699 \cdot 10^6$ 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 “Data mining ice cubes” Astronomical Data Analysis Software and Systems, Paris, 2011

Stahlindustrie

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

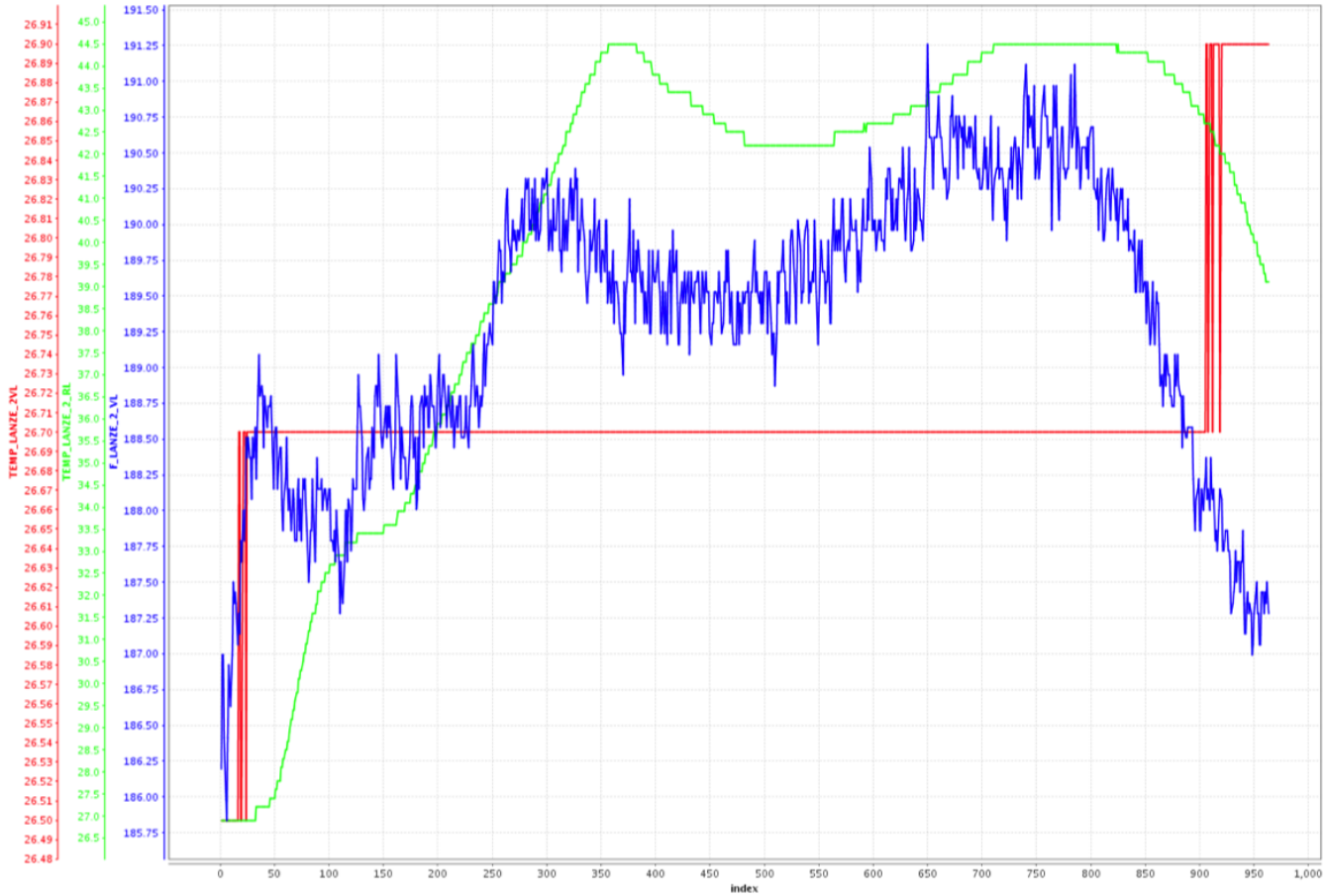


- Online Prognose im Betrieb zur verbesserten Steuerung
  - Temperatur,
  - Fe-Gehalt der Schlacke,
  - Phosphor,
  - Kohlenstoff

  Patent mit Siemag 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

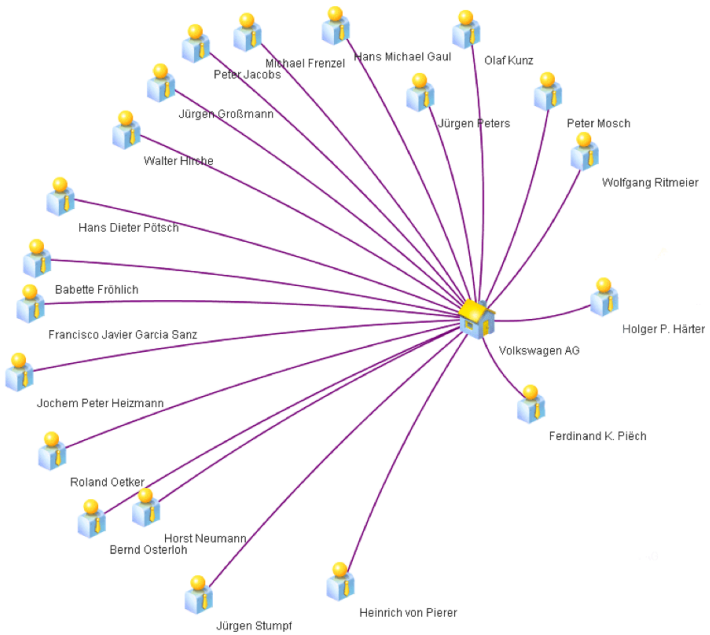
Sprachtechnologie

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 Lautfolgen.
- 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 traniert.

Relationsextraktion aus Web-Seiten

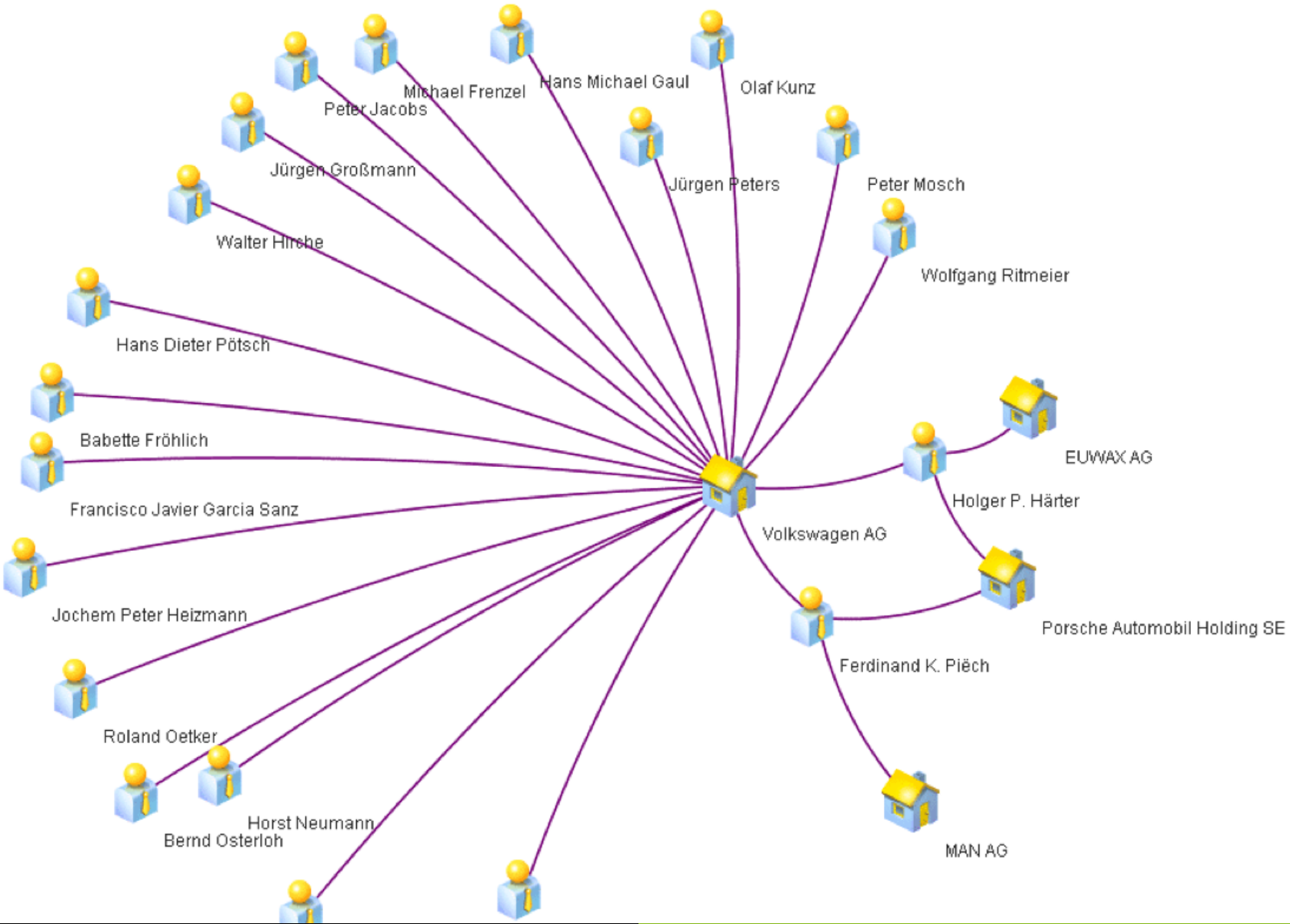
Web of Economy: welche Firmen werden vielleicht fusionieren?



- HTML-Seiten crawled,
- TagSoup - SAX Parser für sauberes HTML,
- XPath für sauberen Text,
- OpenNLP für Aufteilung in Sätze,
- Stanford Parser auf das Deutsche trainiert,
- RapidMiner Information Extraction Plugin

Diplomarbeit am LS 8, Martin Had 2009.

Vorhersage


Michael Frenzel
Hans Michael Gaul
Olaf Kunz
Peter Jacobs
Jürgen Großmann
Jürgen Peters
Peter Mosch
Walter Hirche
Wolfgang Ritmeier
Hans Dieter Pötsch
Babette Fröhlich
EUWAX AG
Holger P. Härter
Francisco Javier Garcia Sanz
Volkswagen AG
Jochem Peter Heizmann
Porsche Automobil Holding SE
Ferdinand K. Piëch
Roland Oetker
Horst Neumann
Bernd Osterloh
MAN AG

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.
- Das Internet soll nicht nur gesamte Dokumente liefern, sondern Fragen beantworten.
- Multimedia-Daten sollen personalisiert strukturiert und gezielter zugreifbar sein.

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 200 Mitarbeiter weltweit
- PrudSys, Chemnitz, Veranstalter des jährlichen Data Mining Cups
- Rapid-I, Dortmund, mehr als 20 Mitarbeiter, wachsend
- Fraunhofer, St. Augustin, Intelligente Analyse- und Informationssysteme, mehr als 200 Wissenschaftler

Und der LS8 sucht auch ;-)

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.

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.

Notation

ExampleSet

Meta Data View | Data View | Plot View

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

row no.	Play	Outlook	Temperat...	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

- 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 $\vec{x_1}, ..., \vec{x_N}$ notiert.
- Die Zufallsvariable Y ist die Ausgabe (Zielvariable).
- N Beobachtungen von Vektoren mit p Komponenten ergeben also eine $N \times p$-Matrix.

Lernaufgabe Clustering

Gegeben

- eine Menge $\mathcal{T} = \{\vec{x_1}, ..., \vec{x_N}\} \subset X$ von Beobachtungen,
- eine Anzahl K zu findender Gruppen $C_1, ..., C_K$,
- eine Abstandsfunktion $d(\vec{x}, \vec{x'})$ und
- eine Qualitätsfunktion.

Finde

- Gruppen $C_1, ..., C_K$, 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.

Lernaufgabe Klassifikation

Gegeben

- Klassen Y, oft $y \in \{+1, -1\}$,
- eine Menge $\mathcal{T} = \{(\vec{x_1}, y_1), ..., (\vec{x_N}, y_N)\} \subset X \times Y$ von Beispielen,
- eine Qualitätsfunktion.

Finde

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

Lernaufgabe Regression

Gegeben

- Zielwerte Y mit Werten $y \in \mathcal{R}$,
- eine Menge $\mathcal{T} = \{(\vec{x_1}, y_1), ..., (\vec{x_N}, y_N)\} \subset X \times Y$ von Beispielen,
- eine Qualitätsfunktion.

Finde

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

Funktionsapproximation

Wir schätzen die wahre, den Beispielen unterliegende Funktion. Gegeben

- eine Menge von Beispielen $\mathcal{T} = \{(\vec{x_1}, y_1), ..., (\vec{x_N}, y_N)\} \subset X \times Y$,
- eine Klasse zulässiger Funktionen f_θ (Hypothesensprache),
- eine Qualitätsfunktion,
- eine feste, unbekannte Wahrscheinlichkeitsverteilung $P(X)$.

Finde

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

Problem

- Wir haben nur eine endliche Menge von Beispielen. Alle Funktionen, deren Werte durch die Beispiele verlaufen, haben einen kleinen Fehler.
- Wir wollen aber für alle Beobachtungen das richtige y voraussagen. Dann sind nicht mehr alle Funktionen, die auf die Beispiele gepasst haben, gut.
- Wir kennen nicht die wahre Verteilung der Beispiele.
- Wie beurteilen wir da die Qualität unseres Lernergebnisses?

Lern- und Testmenge

Wir teilen die Daten, die wir haben, auf:

Lernmenge: Einen Teil der Daten übergeben wir unserem Lernalgorithmus. Daraus lernt er seine Funktion $f(x) = \hat{y}$.

Testmenge: Bei den restlichen Daten vergleichen wir $\hat{y}$ mit y.

Aufteilung in Lern- und Testmenge

- Vielleicht haben wir zufällig aus lauter Ausnahmen gelernt und testen dann an den normalen Fällen. Um das zu vermeiden, verändern wir die Aufteilung mehrfach.

  leave-one-out: Der Algorithmus lernt aus $N-1$ Beispielen und testet auf dem ausgelassenen. Dies wird N mal gemacht, die Fehler addiert.

- Aus Zeitgründen wollen wir den Algorithmus nicht zu oft anwenden.

  Kreuzvalidierung: Die Lernmenge wird zufällig in n Mengen aufgeteilt. Der Algorithmus lernt aus $n-1$ Mengen und testet auf der ausgelassenen Menge. Dies wird n mal gemacht.

Kreuzvalidierung

- Man teile alle verfügbaren Beispiele in n Mengen auf. z.B. $n = 10$.
- Für i=1 bis i=n:
  - Wähle die i-te Menge als Testmenge,
  - die restlichen $n - 1$ Mengen als Lernmenge.
  - Messe die Qualität auf der Testmenge.
- Bilde das Mittel der gemessenen Qualität über allen n Lernläufen. Das Ergebnis gibt die Qualität des Lernergebnisses an.

Was wissen Sie jetzt?

- Sie haben Anwendungsbeispiele gesehen.
- Als Aufgaben der Modellbildung haben Sie Clustering, Klassifikation, Regression gesehen.
- Sie wissen, was die Kreuzvalidierung ist.

Was wissen Sie noch nicht?

- Es gibt viele verschiedene Modellklassen. Damit werden die Lernaufgaben spezialisiert.
- Es gibt unterschiedliche Qualitätsfunktionen. Damit werden die Lernaufgaben als Optimierungsaufgaben definiert.
- Es gibt auch noch mehr Aufgaben: Finden häufiger Mengen!
- Der Gesamtablauf des Data Mining hat eine feste Struktur, die mehr enthält als nur den Lernschritt.
- Das Ziehen von Stichproben und wie diese verwendet werden können, lernen Sie kennen.
- Die Dimensionsreduktion ist ein wichtiger Vorverarbeitungsschritt.

Themen

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

Übungen

Julia Schiffner und Klaus Friedrichs betreuen die Übungen und stehen auch für Fragen zur Verfügung.
Wir verwenden das System RapidMiner und können damit

- (fast) alle Vorverarbeitungsschritte und
- Verfahren und
- Validierungen der Ergebnisse durchführen.

Außerdem verwenden wir R, das Funktionen anbietet für

- (fast) alle Vorverarbeitungsschritte und
- Verfahren und
- Validierungsmethoden.

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.