Vorlesung Wissensentdeckung in Datenbanken

Einführung

Kristian Kersting, (Katharina Morik), Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

8. April 2014

Take-Away Message

Daten sind das neue Öl

Diese Vorlesung

Wie extrahiere ich Wissen aus Daten?

Eine der spannendsten Fragen des Informationszeitalters

Notwendigerweise unvollständig

Statistik+Data Mining

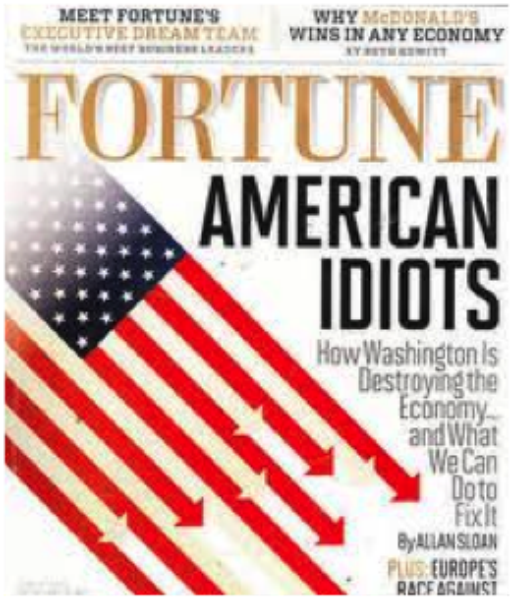
Nur zusammen können wir es schaffen!

Take-Away Message

Aber eigentlich ist die Vorlesung über Euch!

Data scientist: The hot new gig in tech

Datenanalyse Skills sind in Forschung und Industrie gefragt



[Fortune, 5. Sept. 2011]

Gliederung

Bekannte Anwendungen

- Google ordnet die Suchergebnisse nach der Anzahl der auf die Seiten 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 und Benutzerverhalten werden vorhergesagt (Lagerhaltung, Entwicklung, Marketing).
- Daten physikalischer Vorgänge werden analysiert, z.B. Terabytes von Messungen der Astrophysik.
- Verteilte Sensormessungen werden ausgewertet, z.B. zur Verbesserung der Navigationssysteme.

Bekannte Anwendungen

... und viele, viele, viele mehr

Big Data, 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

0.5 PB: DNA aller US-Amerikaner zusammen
1.0 PB: vollständige Erinnerungen von 800 Menschen (R. Kurzweil)
2.0 PB: alle US-Amerikanischen Wissenschaftsbibliotheken

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 ein Artikel auf Deutsch.
- Aktuelle Statistik: http://stats.wikimedia.org/DE/Sitemap.htm

Soziale Netzwerke

- FaceBook verbindet (15.6.2013)
  1.060.627.980 Nutzer weltweit,
  168.000.000 Nutzer in den USA (Rang 1),
  26.000.000 in Deutschland (Rang 10).
- Gezählt werden Nutzer, die sich innerhalb der letzten 30 Tage mindestens einmal am entsprechenden Ort eingeloggt haben. Unternehmensaccounts zählen nicht.
- Auf der Ebene der Städte in Deutschland (1.1.2012)
  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

Big Data: 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 der Seite.
- Peter Norvig, bekannt durch sein Einführungsbuch in die Künstliche Intelligenz (1995, mit Stuart 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)
- "A trillion-word [...] corpus could serve as the basis of [...] certain tasks - if only we knew how to extract the model from the data.
  [Alon Y. Halevy, Peter Norvig, Fernando Pereira: The Unreasonable Effectiveness of Data. IEEE Intelligent Systems 24(2): 8-12 (2009)]

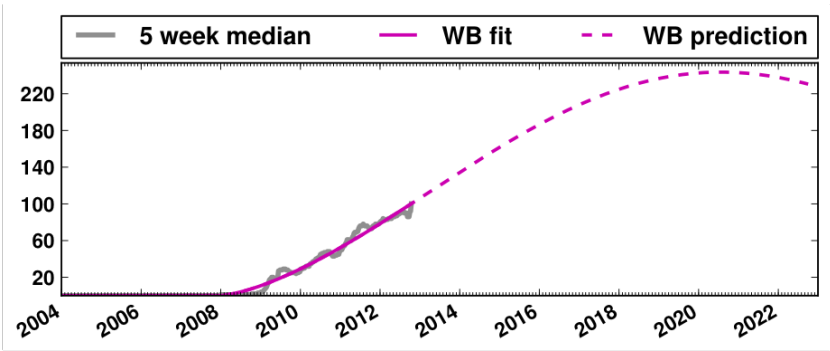
Big Knowledge: Statistik + Data Mining

Data Mining+Statistik soll Massen an Daten indexieren, sortieren, strukturieren, klassifizieren, darin Muster finden, interessante Unterräume und Modelle bestimmen.

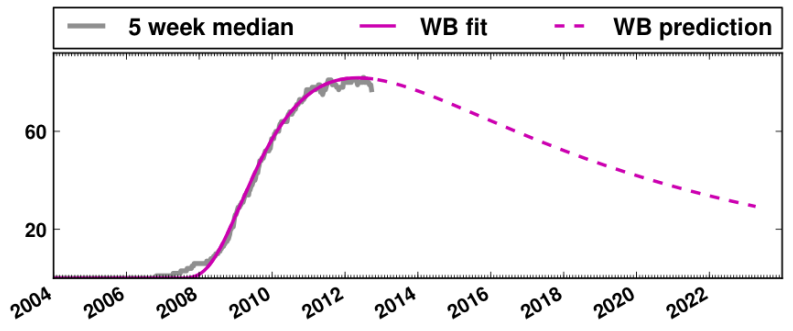
Neue Einsichten in soziale Mechanismen und Massenverhalten

Die Masse macht's – Suchstatistiken von Millionen von Menschen erlauben es physikalisch-plausibel das Interesse an sozialen Medien zu modellieren und vorherzusagen!

Twitter



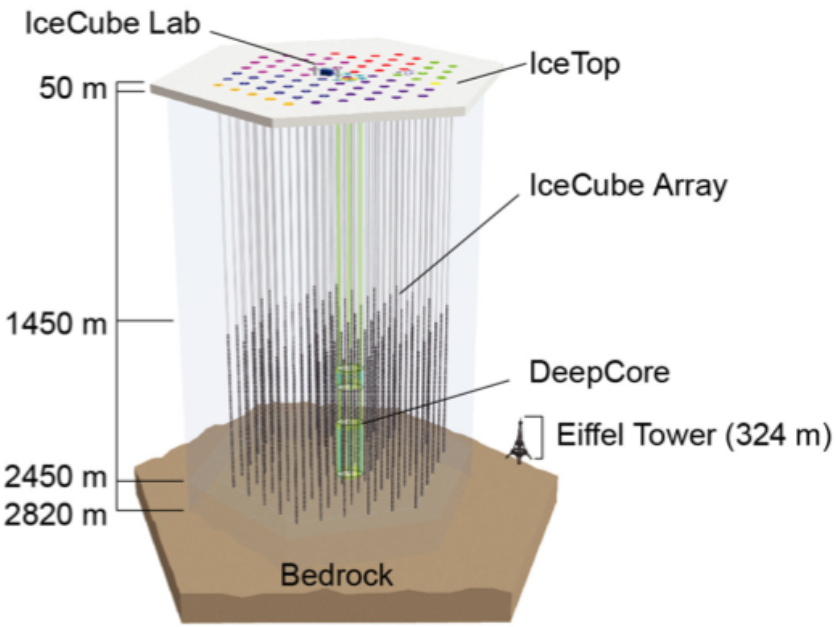
Facebook



[C. Bauckhage, K. Kersting, B. Rastegarpanah: Collective Attention to Social Media Evolves According to Diffusion Models. WWW 2014]

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.

Astroteilchenphysik – IceCube

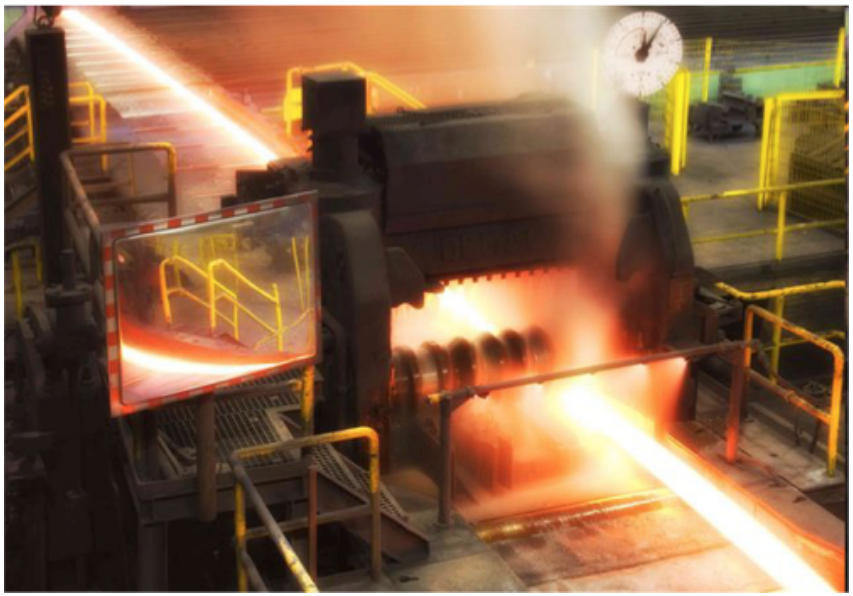
Statistik und Data Mining helfen bei der Trennung von Signal und Rauschen

Signal	Atmosphärisch $\sim$ 14180 Ereignisse in 33,28 Tagen bei IC-59 (IceCube 59 string configuration)
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

Statistik und Data Mining verbessern die Stahlproduktion

Ressourcenschonung:
Kürzerer Brennprozess, niedrigere Temperatur, weniger Metallgabe!

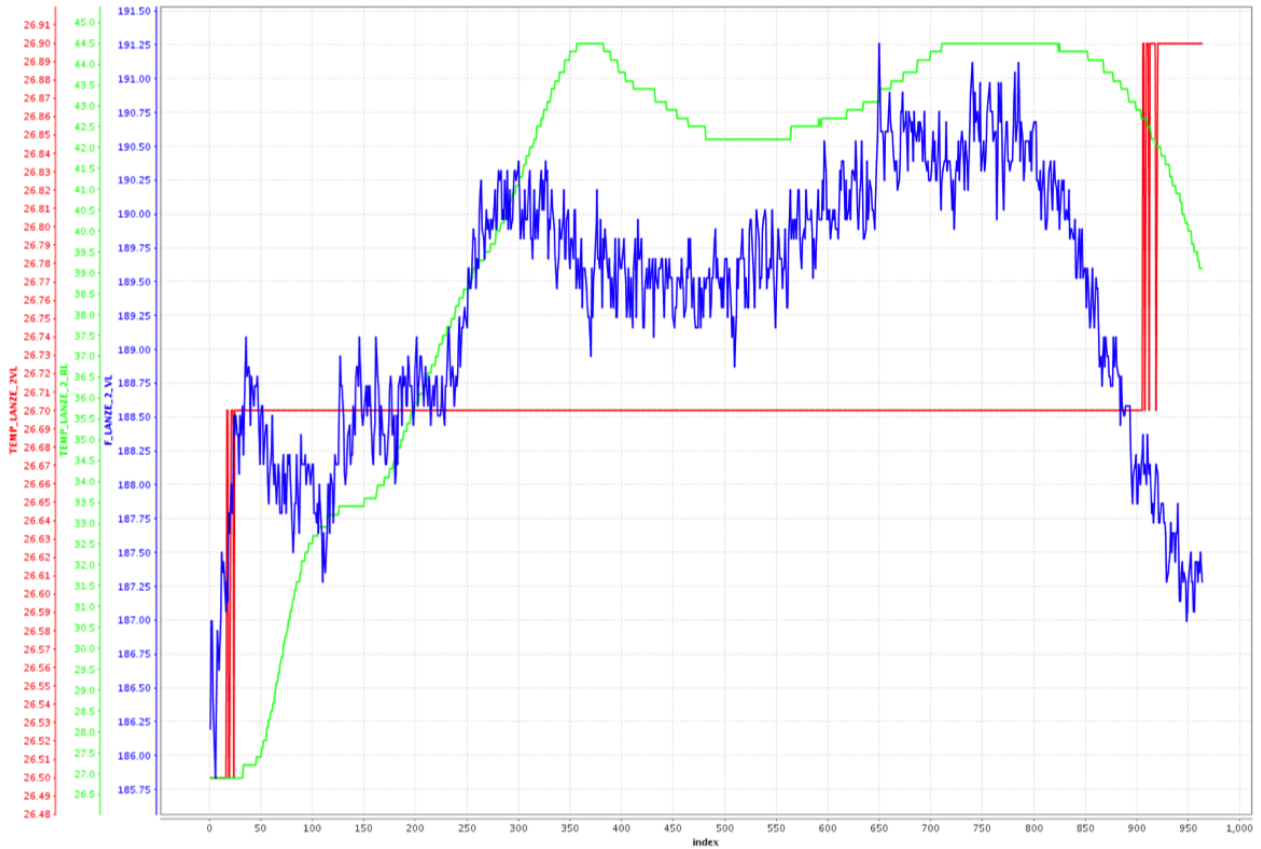


- Online Prognose im Betrieb zur verbesserten Steuerung.
  - Temperatur,
  - Fe-Gehalt der Schlacke,
  - Phosphor,
  - Kohlenstoff

  Patent mit Siemag angemeldet.
- Verteiltes Lernen statistischer Modelle

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 und Statistik zusammen machen 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 Statistik, maschinellem Lernen und 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.

Warum dieses 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
- Amazon, Berlin, Machine Learning HQ mit mehr als 50 Mitarbeiter
- Burda, München, aktiv in Social Media Mining
- Zalando, Berlin, sucht Data Scientists
- Banken, Versicherungen, Hedgefonds, und viele, viele mehr

Und wir suchen auch immer, ;-)

Populärste Software zur Datenanalyse

KDD Nuggets Umfrage 2010

1. RapidMiner
2. R

Rexer Analytics Umfrage 2010 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
  1. Regression,
  2. Entscheidungsbäume zur Klassifikation,
  3. Clustering.

Die klassischen Aufgaben der Datenanalyse

Datenanalyse – generische Aufgabe

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

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

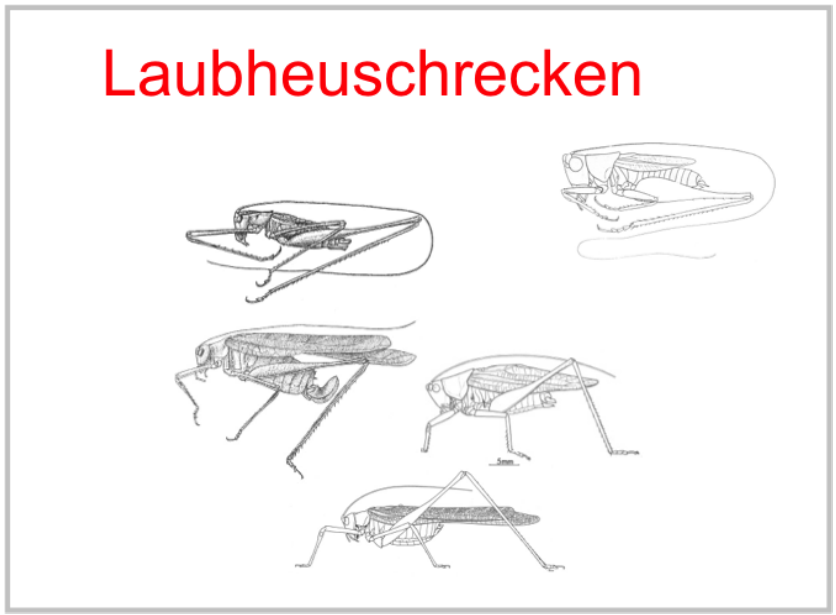
Ausgabe: Ein quantitativer Wert (Messwert) oder ein qualitativer gehört zu jeder Beobachtung (Zielvariable).

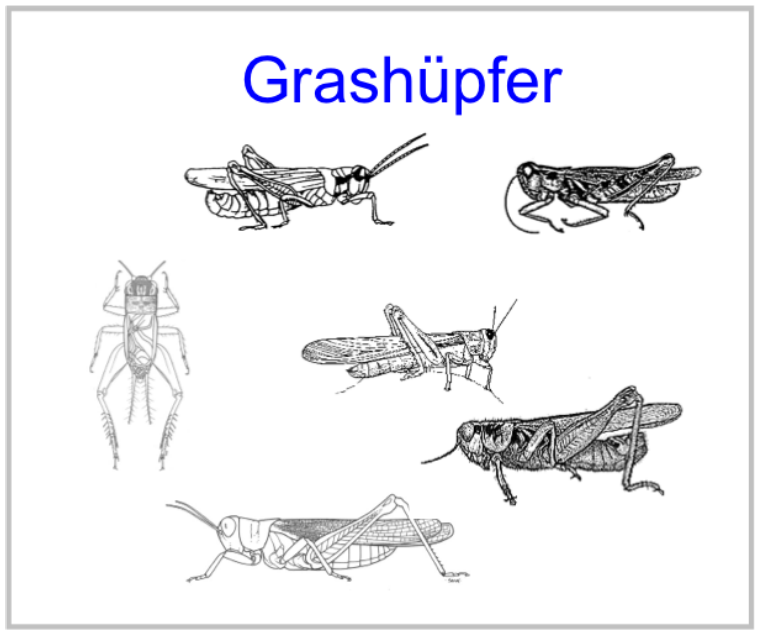
Ein Lernverfahren findet eine Funktion, die Objekten einen Ausgabewert zuordnet. Oft minimiert die Funktion einen Fehler.

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

Ein Beispiel

Wir haben 10 Datenpunkte insgesamt. Jeweils fünf von ihnen beschreiben Laubheuschrecken bzw. Grashüpfer.

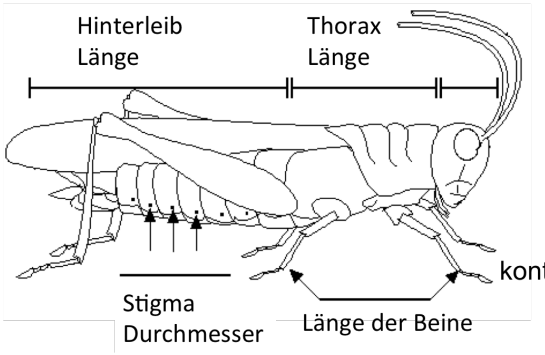




Der erste Schritt: Notation

Farbe {grün, braun, grau, andere}

Binär
Hat Flügel?

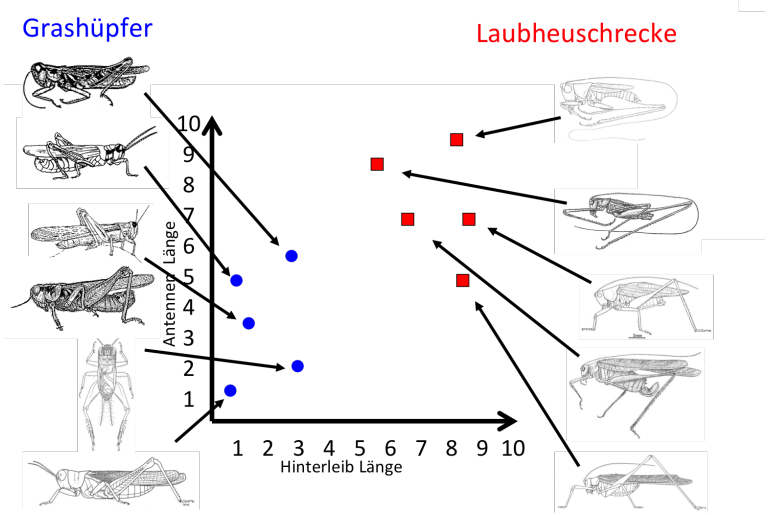


kein Merkmal; nur zur Organization zu benutzen

ID	Hinterleib Länge	Antennen Länge	Klasse
1	2.7	5.5	Grasshüpfer
2	8.0	9.1	Laubheuschrecke
3	0.9	4.7	Grasshüpfer
4	1.1	3.1	Grasshüpfer
5	5.4	8.5	Laubheuschrecke
6	2.9	1.9	Grasshüpfer
7	6.1	6.6	Laubheuschrecke
8	0.5	1.0	Grasshüpfer
9	8.3	6.6	Laubheuschrecke
10	8.1	4.7	Laubheuschrecke

- 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 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.

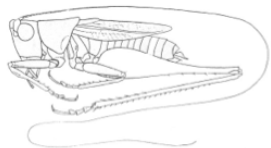
Finde

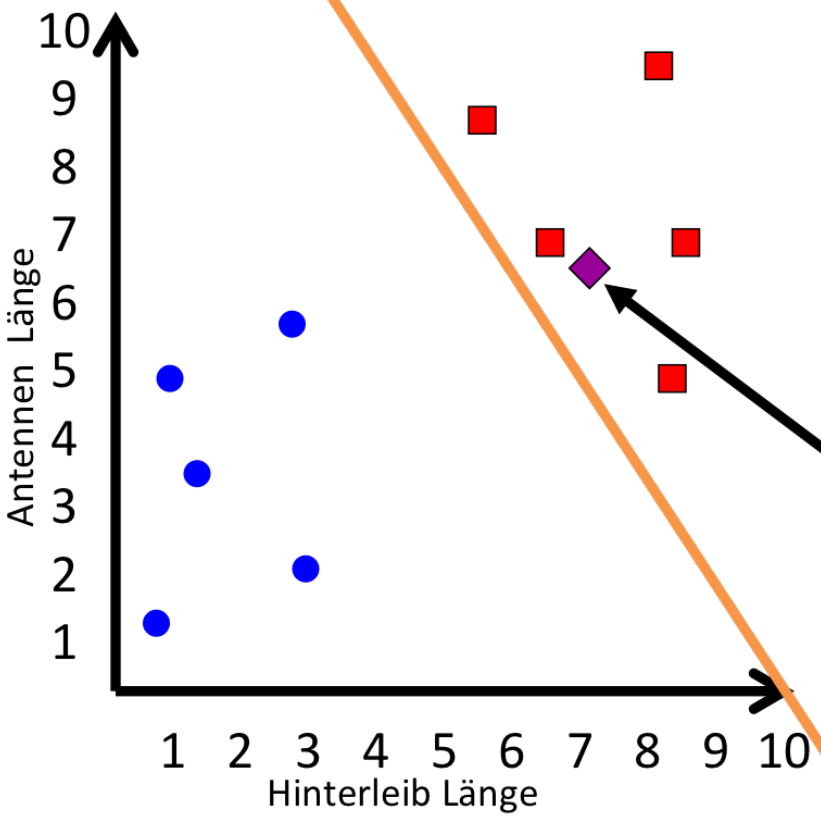
- eine Funktion $f : X \to Y$, die die Qualitätsfunktion optimiert.

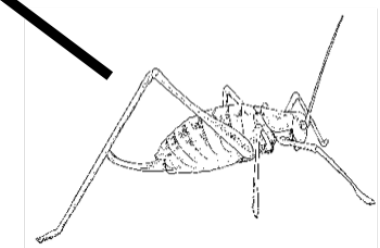
Lernaufgabe Klassifikation

Grashüpfer

Laubheuschrecke







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.

Also keine Klasseninformation!

ID	Hinterleib Länge	Antennen Länge	Klasse
1	2.7	5.5	Grasshüpfer
2	8.0	9.1	Laubheuschrecke
3	0.9	4.7	Grasshüpfer
4	1.1	3.1	Grasshüpfer
5	5.4	8.5	Laubheuschrecke
6	2.9	1.9	Grasshüpfer
7	6.1	6.6	Laubheuschrecke
8	0.5	1.0	Grasshüpfer
9	8.3	6.6	Laubheuschrecke
10	8.1	4.7	Laubheuschrecke

Aber Achtung: Es ist nicht einfach, ein gutes Abstandsmaß zu finden

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.

Es werden also kontinuierliche Werte vorhergesagt

ID	Hinterleib Länge (Klasse)	Antennen Länge	Art
1	**2.7**	5.5	Grasshüpfer
2	**8.0**	9.1	Laubheuschrecke
3	**0.9**	4.7	Grasshüpfer
4	**1.1**	3.1	Grasshüpfer
5	**5.4**	8.5	Laubheuschrecke
6	**2.9**	1.9	Grasshüpfer
7	**6.1**	6.6	Laubheuschrecke
8	**0.5**	1.0	Grasshüpfer
9	**8.3**	6.6	Laubheuschrecke
10	**8.1**	4.7	Laubheuschrecke

Oder ganz allgemein: 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.

Themen

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

Übungen

Daniel Horn, Leonard Judt und Alejandro Molina 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.

Wir sehen uns...

In der ersten Übung wird R vorgestellt.

Übungen finden an 2 von 3 möglichen Terminen statt:

Montags 10-12 Uhr in Raum CDI 120
Dienstag 14-16 Uhr im Raum CDI 120
Mittwochs 14-16 Uhr in Raum E02, OH14

Gruppeneinteilung JETZT im Anschluss!

Zusammenfassung

Daten sind das neue Öl

Data Mining und Statistik raffinieren das Öl zu Wissen

Aber die Vorlesung ist nicht nur über Daten und ihr Analyse. Sie ist über Euch! Datenanalyse Skills sind in Forschung und Industrie gefragt

Viel Spass!
Danke für Ihre Aufmerksamkeit!

Beispiele zur Literatur für den Einstieg

Hastie, T., Tibshirani, R., Friedman, J. (2001). The Elements of Statistical Learning. Springer.

Hand, D., Mannila, H., Smyth, P. (2001). Principles of Data Mining. MIT Press.

Witten, I.H., Frank, E. (2001): Data Mining - Praktische Werkzeuge und Techniken für das maschinelle Lernen.

Mitchell, T. (1997): Machine Learning, McGraw Hill, 1997.

Teschl, G., Teschl, S. (2006): Mathematik für Informatiker, Springer.