

Vorlesung Wissensentdeckung

Einführung

Katharina Morik, Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

5.4.2011

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. Terrabytes von Messungen der Astrophysik.
- Verteilte Sensormessungen werden ausgewertet, z.B. zur Verbesserung der Navigationssysteme.

Gliederung

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

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, indem aus den log-Dateien gelernt wird.
- Das Internet soll nicht nur gesamte Dokumente liefern, sondern Fragen beantworten.
- Multimedia-Daten sollen personalisiert strukturiert und gezielter zugreifbar sein.

Das neue Paradigma: Sehr viele Daten!

- 20 Petabyte Daten werden bei Google täglich bearbeitet (2011).
 - 1 Megabyte (MB) = 1024 Kilobyte = 1024 · 1024 Byte = 1.048.576 Byte
 - 1 Gigabyte (GB) = 10⁹ Bytes
 - 1 Terabyte (TB) = 10¹² Bytes
 - 1 Petabyte (PB) = 10¹⁵ = 1.125.899.906.842.624 Bytes
 - 1 Exabyte (EB) = 10¹⁸ Bytes
- Wikipedia bietet
 - 3.543.418 Artikel auf Englisch, 1.185.247 Deutsch
 - 276 Sprachen
 - Aktuelle Statistik:
<http://stats.wikimedia.org/DE/Sitemap.htm>
- FaceBook verbindet 642.5 Millionen Menschen weltweit (Ende Februar 2011).
- Ein SmartPhone verarbeitet etwa 1000 Systemaufrufe pro Sekunde.

Die Masse macht's!

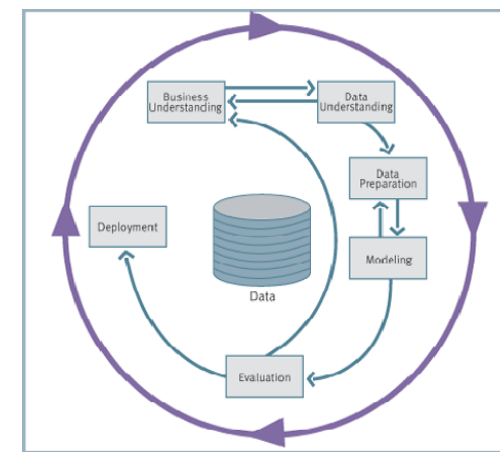
Googles Philosophie ist: wir wissen nicht, warum diese Seite besser als eine andere ist. Aber wenn viele Menschen das meinen, ist es so.

- Statistik eingehender Verweise auf eine Seite zeigt die Bedeutung, die die verweisenden Menschen der Seite beimessen.
- Speaking at the O'Reilly Emerging Technology Conference this past March, Peter Norvig, Google's research director, offered an update to George Box's maxim: "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.

CRISP-DM: CRoss Industry Standard Process for Data Mining (<http://www.crisp-dm.org>)

- Zusammenarbeit von NCR, SPSS und DaimlerChrysler
- NCR: Mehrwert für Data Warehouse Kunden
- SPSS: Konzept für Data Mining Produkt 'Clementine'
- DaimlerChrysler: Praktische Erfahrung
- KEINE theoretische, akademische Entwicklung,
- SONDERN Entwicklung aus praktischer Erfahrung an realen Problemen.

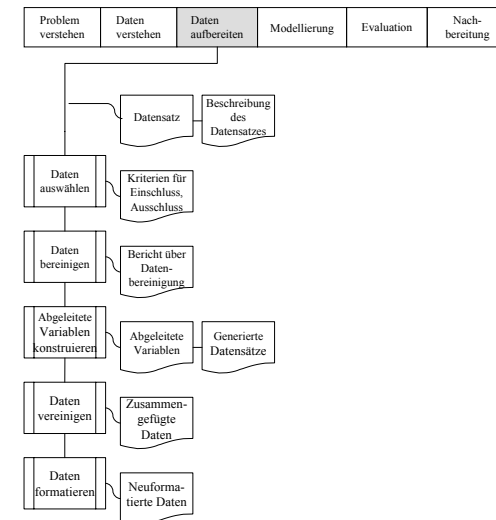
Übersicht



CRISP: Schritte

- **Problem verstehen:** Analyseziele, Situationsbewertung, Datenanalyseziele, Projektplan
- **Daten verstehen:** Sammeln, beschreiben, untersuchen, Qualität von Rohdaten
- **Daten aufbereiten:** Ein- und Ausschluss, Bereinigung, Transformation von Variablen
- **Modellierung:** Methoden- und Testdesignwahl, Schätzung, Modellqualität
- **Evaluierung:** Modell akzeptieren, Prozess überprüfen, nächste Schritte
- **Nachbereitung:** Anwendungs- und Wartungsplan, Präsentation, Bericht

Vorverarbeitung



CRISP: Beispiel: Verteilungsstudien: Problemanalyse

- **Problem:** Charakterisierung der Verteilung von Medikamentenklassen im Körper
- **Studientyp:** Verteilung von ¹⁴C-markierten Substanzen in Ratten 5-6 Minuten nach intravenöser Injektion.
- **Rohdaten:** 26 Experimente, 20 Substanzen, 3-4 Ratten / Experiment, 85 Ratten, 24 Organe, 6 physico-chemische Eigenschaften: 3 Säure-Konstanten (pka), Lipophilität (log p), Molekulargewicht (WE), Wasserlöslichkeit (SO)
- **Datenanalyseziel:** Finden von typischen Verteilungsmustern für Medikamentenklassen auf der Basis von physico-chemischen Eigenschaften
- **Projektplan:** Deskription, Ersetzen fehlender Werte, Klassifikationsregeln

CRISP: Verteilungsstudien: Datenaufbereitung

- **Univariate Analyse:** Viele fehlende Werte, insbesondere wegen nicht definierten Werten bei pka
- **Bivariate Analyse:** Niveau von ¹⁴C-Konzentrationen variiert mit Substanz-Dosen
- **Transformation 1:** Normalisierung der ¹⁴C-Werte mit Blut := 1
- **Transformation 2:** Bilden von physico-chemischen Klassen (z.B. Säuregehaltsklassen mit wohldefinierten pka-Werten)
- **Transformation 3:** Bilden von Therapieklassen für die Medikamente
- **Variablenselektion:** Weglassen der Verdauungsorgane → 20 Organe
- **Beobachtungsselektion:** 78 vollständige Beobachtungen

CRISP: Verteilungsstudien: Klassifikation

- **Therapieklassen:** Klassifikationsregel bei Linearer Diskriminanzanalyse mit 6 Organen
- **Therapieklassen:** Neuroleptica / Antidepressiva, Betablocker / Ca-Antagonisten, alle anderen Klassen zusammen sehr gut getrennt mit 96% Richtigkeit
- **Säuregehaltsklassen:** LDA mit 95% Richtigkeit
- **Säuregehaltsklassen:** RDA (Regularisierte DA) nach Box-Cox-Transformation mit 99% Richtigkeit
- **Säuregehaltsklassen:** Prognosefähigkeit genauso gut (Kreuzvalidierung)

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.

CRISP: Verteilungsstudien: Säuregehaltsklassen

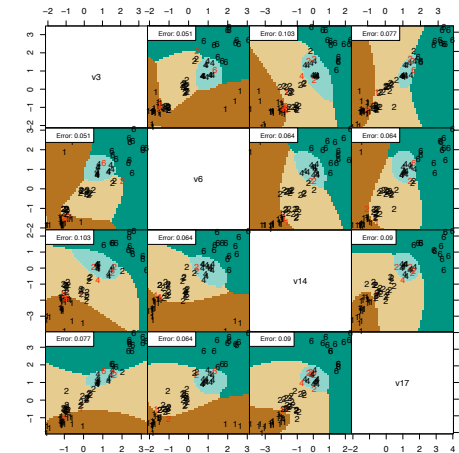


Abbildung: Fehlerraten bei 2-Organ-Kombinationen

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

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

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.

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.

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.

Funktionsapproximation

Wir schätzen die wahre, den Beispielen unterliegende Funktion. Gegeben

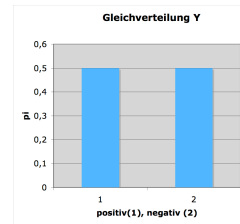
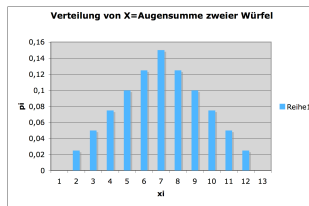
- eine Menge von Beispielen $\mathcal{T} = \{(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)\} \subset X \times Y$,
- eine Klasse zulässiger Funktionen f_θ (Hypothesensprache),
- eine Qualitätsfunktion,
- eine feste, unbekannte Wahrscheinlichkeitsverteilung $P(X)$.

Finde

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

Zur Erinnerung: Verteilung

Eine Zufallsvariable X heißt *diskret*, wenn sie nur endlich oder abzählbar unendlich viele Werte x_1, \dots, x_m annehmen kann. Zu jedem Wert gehört ein Ereignis, das mit der Wahrscheinlichkeit $P(X = x_i)$ eintreten kann. Die Realisationen x_i gemeinsam mit den zugehörigen Wahrscheinlichkeiten heißen **(Wahrscheinlichkeits-)Verteilung** von X .



Verteilungsfunktion

Sei X eine diskrete oder stetige Zufallsvariable. Die Funktion $D(x) = P(X \leq x), x \in \mathcal{R}$ heißt **Verteilungsfunktion** von X .

Bei diskreten Zufallsvariablen gilt: $D(x) = \sum_{i: x_i \leq x} p_i$

Eine Zufallsvariable heißt **stetige Zufallsvariable**, wenn ihre Verteilungsfunktion stetig ist.

Dichtefunktion

Die Ableitung $D'(x)$ wird **Dichtefunktion** genannt. Umgekehrt erhält man die Verteilungsfunktion durch Integration der Dichtefunktion: $D(x) = \int_{-\infty}^x h(t)dt$
 Funktionen, die eine Dichte haben, sind absolut stetig.
 Die Gesamtfläche unter dem Graphen von h ist gleich 1.

Wenn wir die Verteilung kennen, können wir eine gute Prognose machen!

- Wenn wir wissen, dass $p_i = 0,01$ ist, dann ist es nicht so schlimm, wenn wir uns bei x_i irren – wir irren uns dann selten.
- Wenn wir wissen, dass $P(Y = +1) = 0,99$ ist, dann sagen wir immer +1 voraus und sind in 99% der Fälle richtig. Wir haben nur ein Risiko von 1%, uns zu irren.

Qualitätsfunktion – Fehlerfunktion

Fehlerrisiko:

$$R(Y, f(X)) = \sum_{i=1}^N Q(y_i, \vec{x}_i) p(\vec{x}_i) \quad (1)$$

wobei $p(\vec{x}_i)$ die Wahrscheinlichkeit ist, dass das Beispiel \vec{x}_i aus X gezogen wird.

Mittlerer Quadratischer Fehler:

$$MSE(Y, f(X)) = \frac{1}{N} \sum_{i=1}^N (y_i - f(\vec{x}_i))^2 \quad (2)$$

Mittlerer 0-1-Verlust: $Q(Y, f(X)) = \frac{1}{N} \sum_{i=1}^N Q(y_i, f(\vec{x}_i))$, wobei

$$Q(y_i, f(\vec{x}_i)) = \begin{cases} 0, & \text{falls } f(\vec{x}_i) = y \\ 1, & \text{falls } f(\vec{x}_i) \neq y \end{cases}$$



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 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 Gesamtprozess 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.

Was wissen Sie jetzt?

- Sie haben das CRISP kennengelernt, das den gesamten Ablauf der Wissensentdeckung beschreibt.
- Als Aufgaben der Modellbildung haben Sie **Clustering**, **Klassifikation**, **Regression** gesehen.
- Sie wissen, was die **Kreuzvalidierung** ist.

Themen

- Datenbanken analysieren (OLAP)
- Finden häufiger Mengen
- statistische Grundbegriffe
- Versuchsplanung, Stichproben
- Klassifikation
- Entscheidungsbäume
- lineare Modelle
- Stützvektormethode (SVM) und strukturelle Risikominimierung
- stetige Modelle
- Merkmalsselektion
- Zeitreihen
- Clustering

Übungen

Julia Schiffner und Christian Bockermann 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.

In RapidMiner eingebettet ist R, das Funktionen anbietet für

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

Wofür bekommen Sie einen Schein?

- Kommen Sie in jede Vorlesung – dann können Sie auch das Tempo bestimmen und Fragen stellen!
- Gehen Sie in die Übungsgruppe! Sie dürfen nur max. 2 mal unentschuldigt fehlen.
- Lösen Sie jede Übungsaufgabe:
 - Werden 50% der Punkte erreicht,
 - höchstens 3 Blätter nicht abgegeben und
 - mindestens eine Aufgabe in der Übung vorgerechnet
 bekommt man einen Schein.
- Nutzen Sie die Vorlesung/Übung zur Vorbereitung auf eine Fachprüfung! Dies hier ist DIE Prüfungsvorbereitung!

Literatur

Trevor Hastie, Robert Tibshirani, Jerome Friedman.
The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

Springer series in statistics. Springer, New York, USA, 2001.

Jiawei Han, Micheline Kamber.
Data Mining: Concepts and Techniques.

Morgan Kaufmann, Elsevier, San Francisco, USA, 2006.

Gerald Teschl Susanne Teschl.

Mathematik für Informatiker.

Springer, 2006.