Vorlesung Wissensentdeckung

Einführung

Katharina Morik, Uwe Ligges

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

13.4.2010

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

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.

Rexer Analytics Umfrage

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

Funktionsapproximation

Wir schätzen die wahre, den Beispielen unterliegende Funktion. Gegeben

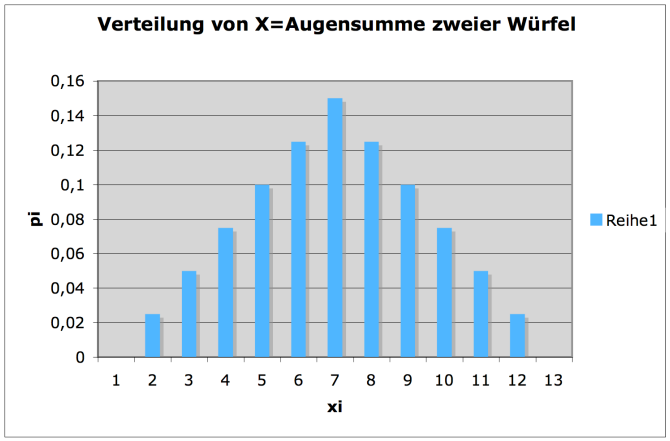
- eine Menge von Beispielen $\mathcal{T} = \{(\vec{x_1}, y_1), ..., (\vec{x_N}, y_N)\} \subset X \times Y$,
- eine Klasse zulässiger Funktionen f_θ (Hypothesensprache),
- eine Qualitätsfunktion,
- eine feste, unbekannte Wahrscheinlichkeitsverteilung $P(X)$.

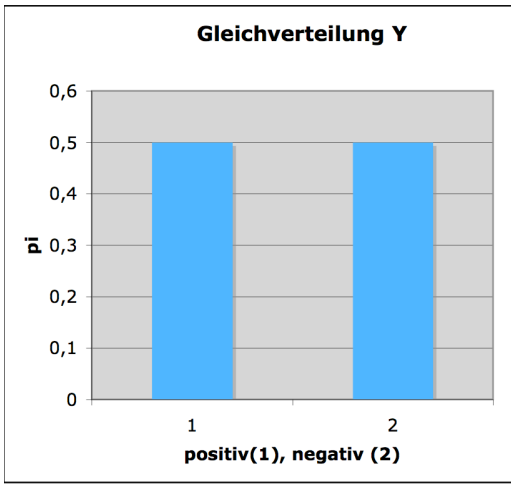
Finde

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

Zur Erinnerung: Verteilung

Eine Zufallsvariable X heißt *diskret*, wenn sie nur endlich oder abzählbar unendlich viele Werte $x_1, ..., x_m$ 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: $Q(Y, f(X)) = \frac{1}{N} \sum_{i=1}^{N} Q(\vec{x_i}, f)$, wobei

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

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 jetzt?

- 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

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

Übungen

Julia Schiffner und Felix Jungermann 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.

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, 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.