Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

30.10.2008

# 1. Übung „Knowledge Discovery“

Wintersemester 2008/2009

## Vorbemerkungen

Vorlesungsfolien und Übungsblätter können Sie im Internet unter der Adresse `http://www.kde.cs.uni-kassel.de/lehre/ws2008-09/kdd` oder mittels folgender RSS-Feeds einsehen:

- Übungen: `http://www.kde.cs.uni-kassel.de/lehre/ws2008-09/kdd/uebung/rss`
- Folien: `http://www.kde.cs.uni-kassel.de/lehre/ws2008-09/kdd/folien/rss`

Bei Fragen wenden Sie sich bitte an Robert Jäschke (jaeschke@cs.uni-kassel.de).

## 1 Allgemeines

1. Was ist KDD und was ist insbesondere das Ziel davon?
2. Was ist der Unterschied zwischen Data Mining und Knowledge Discovery?
3. Geben Sie Beispiele für Bereiche an, in denen KDD angewendet wird.
4. Welche Geschäftsziele werden typischerweise durch KDD unterstützt? Diskutieren sie diese anhand der von Ihnen in 3. genannten Anwendungsbereiche.
5. Geben Sie vier typische Verfahren/Methoden an, die im Rahmen von KDD Anwendung finden und beschreiben Sie diese kurz.

## 2 Überwachte vs. Unüberwachte Verfahren

1. Was sind die wesentlichen Unterschiede zwischen einem überwachten und einem unüberwachten Verfahren?
2. Was für Konsequenzen hat die Verwendung eines überwachten bzw. unüberwachten Verfahrens für die zur Verfügung zu stellenden Daten?
3. Nennen sie jeweils zwei Anwendungen für ein überwachtes und ein unüberwachtes Verfahren.

## 3 CRISP-DM Methodologie

1. Nennen Sie die sechs Phasen der CRISP-DM Methodologie.
2. Was sind die wichtigsten Schritte in der Vorverarbeitung?
3. Was für Probleme ergeben sich dabei typischerweise?
4. Wie hängt diese Phase konzeptuell mit den anderen Phasen zusammen?

## 4 Vorverarbeitung

Ein Versandhändler möchte seinen Kundenbestand analysieren, um den aktivsten Kunden besondere Angebote zu machen. Dazu hat er Ihnen folgende Stichprobe seiner Daten bereitgestellt:

Kunden

| Id | Name | E-Mail-Adresse | Strasse | Ort | PLZ |
|---|---|---|---|---|---|
| 1 | Carla D. Eiffel | | Forsthausweg 2 | Duisburg | 47057 |
| 2 | F. Ganter | ganter@gxm.de | Geschwister-Scholl-Platz 1 | München | 80539 |
| 3 | Jan Klein | jan_klein@gmail.com | Kaiserswerther Str. 16 | Berlin | 14195 |
| 4 | Anton BlÃcher | bluecher@gmx.de | Rosengarten 10 | Halle/Saale | 6132 |
| 6 | Irving, Hans | hans.irving@web.de | Christian-Albrechts-Platz 4 | Kiel | 24118 |
| 7 | Ludwig Mann | lm@lumann.com | Kaiserswerther Strasse 16 | Berlin | 14195 |

Kaufdaten Online-Shop

| Id | Kunden-Id | Datum | Artikel-Id | Preis | Anzahl |
|---|---|---|---|---|---|
| 1 | 1 | 1.1.1970 | 1 | 12,99 | 2 |
| 2 | 1 | 1.1.1970 | 5 | 5,49 | 1 |
| 3 | 2 | 12.6.2006 | 3 | 15,00 | 1 |
| 4 | 5 | 20.6.2007 | 2 | 2,00 | 4 |
| 5 | 3 | 21.6.2006 | 5 | 5,99 | 1 |
| 6 | 1 | 1.1.1970 | 1 | 12,99 | 255 |

Kaufdaten telefonische Bestellung

| Kunden-Nr | Datum | Artikel | Preis | Menge |
|---|---|---|---|---|
| 3 | 3.6.06 | 2 | 2 | 2 |
| 3 | 10.6.06 | 1 | 12,99 | 1 |
| 4 | 4.6.06 | 2 | 2,00 | 1 |
| 1 | 3.6.06 | 1 | 12,99 | 5 |
| 7 | 9.6.06 | 5 | 5,99 | 1 |

1. Wenden Sie (soweit möglich) die in Aufgabe 3, Teil 2 genannten Schritte der Datenvorverarbeitung auf den folgenden Datensatz an. Machen Sie sich insbesondere klar, welches konkrete Vorgehen zu welchem Schritt gehört.

   Wie hängen die vorzunehmenden Schritte vom Ziel der Datenanalyse ab?

2. Diskutieren Sie in der Gruppe die auftretenden Probleme und wie Sie diese lösen könnten.

## 5 Datenbanken

1. Definieren Sie informell ein Datenbanksystem und beschreiben Sie den prinzipiellen Aufbau.

2. Beschreiben Sie den prinzipiellen Unterschied für den Zugriff auf Daten einer Datenbank beim Data Mining gegenüber einer klassischen Datenbankanwendung.

3. Geben Sie die Eigenschaften eines B-Baumes wieder und begründen Sie, warum ein B-Baum balanciert sein muss.

## 6 Einführung Statistik

1. Gegeben sei folgende Tabelle:

| Kredithöhe in Euro | 300 schlechte Kunden (in Prozent) | 700 gute Kunden (in Prozent) |
|---|---|---|
| $0 < \cdots \leq 500$ | 1.00 | 2.14 |
| $500 < \cdots \leq 1000$ | 11.33 | 9.14 |
| $1000 < \cdots \leq 1500$ | 17.00 | 19.86 |
| $1500 < \cdots \leq 2500$ | 19.67 | 24.57 |
| $2500 < \cdots \leq 5000$ | 25.00 | 28.57 |
| $5000 < \cdots \leq 7500$ | 11.33 | 9.71 |
| $7500 < \cdots \leq 10000$ | 6.67 | 3.71 |
| $10000 < \cdots \leq 15000$ | 7.00 | 2.00 |
| $15000 < \cdots \leq 20000$ | 1.00 | 0.29 |
| Frühere Kredite | | |
| gut | 82.33 | 94.85 |
| schlecht | 17.66 | 5.15 |

   Die Tabelle gibt die Merkmale von 1000 Kunden einer Bank wieder, die ihren Kredit mit/ohne Probleme zurückgezahlt haben. Die Bank möchte von Ihnen ein Modell zur Vorhersage, ob neue Kunden Probleme bei der Rückzahlung ihres Kredites machen werden oder nicht. Stellen Sie dazu die Informationen der Tabelle in geeigneter Form dar, um sich einen ersten Eindruck von den Daten zu machen. Interpretieren Sie das Ergebnis und vergleichen Sie vor allen Dingen die Verteilungen der Merkmale. Berechnen Sie das arithmetische Mittel und den Median. Interpretieren Sie die Ergebnisse und setzen Sie die Ergebnisse in Bezug zu den Verteilungen der dargestellten Merkmale.

2. Ein Experiment bestehe aus dem Werfen eines Würfels und einer Münze. Geben Sie einen geeigneten Ergebnisraum $\Omega$ an. Zeigt die Münze Wappen, so wird die doppelte Augenzahl des Würfels notiert, bei Zahl nur die einfache. Wie groß ist die Wahrscheinlichkeit, dass eine gerade Zahl notiert wird?