Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

13.11.2008

3. Übung „Knowledge Discovery“

Wintersemester 2008/2009

1 Statistik

1. Zeigen Sie, dass für $A, B \subseteq \Omega$ mit $P(B) > 0$ die bedingte Wahrscheinlichkeit

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

ein Wahrscheinlichkeitsmaß auf Ω ist.

(A1) wegen $P(A \cap B) \geq 0$ und $P(B) > 0$ gilt $P(A|B) \geq 0$

(A2)

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

(denn $P(B) > 0$)

(A3) Seien $A, C \subseteq \Omega$ mit $A \cap C = \emptyset$. Dann gilt wegen der Regel von de Morgan und weil $P(A)$ ein Wahrscheinlichkeitsmaß ist:

$$\begin{aligned} P(A \cup C|B) &= \frac{P((A \cup C) \cap B)}{P(B)} && (1)\\ &= \frac{P((A \cap B) \cup (C \cap B))}{P(B)} && (2)\\ &= \frac{P(A \cap B) + P(C \cap B)}{P(B)} && (3)\\ &= P(A|B) + P(C|B) && (4) \end{aligned}$$

2. Zur Diagnose einer bestimmten Erkrankung wird eine Reihenuntersuchung durchgeführt. Aus langjährigen Studien ist bekannt, dass 1.5 % der Bevölkerung diese Krankheit haben. Das Diagnoseverfahren erkennt mit 98 %-iger Sicherheit eine erkrankte Person als erkrankt und mit 99 %-iger Sicherheit eine gesunde Person als gesund.

   Schätzen Sie zunächst, wie groß die Wahrscheinlichkeit ist, dass eine zufällig ausgewählte Person, bei der die Krankheit diagnostiziert wird, tatsächlich krank ist.

   Berechnen Sie nun mit der Bayesschen Formel, wie groß die Wahrscheinlichkeit tatsächlich ist. Vergleichen Sie mit Ihrem geschätzen Wert.

- Ereignise: {krank, gesund}
- $P(\text{krank}) = 0,015$, $P(\text{gesund}) = 0,985$,
- Sei nun *diagkrank* das Ereignis, dass einer Person diagnostiziert wurde, dass sie krank ist (entsprechend *diaggesund*).
- $P(\text{diagkrank}|\text{krank}) = 0,98$
- $P(\text{diaggesund}|\text{gesund}) = 0,99$
- $P(\text{diagkrank}|\text{gesund}) = 1 - P(\text{diaggesund}|\text{gesund}) = 0,01$
- gesucht: $P(\text{krank}|\text{diagkrank})$
- Anwendung der Bayesschen Formel:

$$P(\text{krank}|\text{diagkrank}) = \frac{P(\text{krank})P(\text{diagkrank}|\text{krank})}{P(\text{krank})P(\text{diagkrank}|\text{krank}) + P(\text{gesund})P(\text{diagkrank}|\text{gesund})} \tag{5}$$

$$= \frac{0,015 \cdot 0,98}{0,015 \cdot 0,98 + 0,985 \cdot 0,01} \tag{6}$$

$$= \frac{0,01470}{0,02455} \tag{7}$$

$$\approx 0,5988 \tag{8}$$

3. Zur Überprüfung einer Warenlieferung aus einer großen Fertigungsmenge, bei der im Mittel 10 % der Stücke defekt sind, wurden folgende Vorschriften verwendet:

   Die Sendung wird abgelehnt, falls in einer Stichprobe vom Umfang

   a) 15 mehr als ein fehlerhaftes Stück auftritt,

   b) 30 mehr als zwei fehlerhafte Stücke auftreten.

   Bei welcher Methode werden mehr Sendungen abgelehnt?

- Ereignisse: $A =$ Stück ist in Ordnung, $\bar{A} =$ Stück ist Ausschuß; damit:

$$P(A) = 0,9 =: q, P(\bar{A}) = 0,1 =: p$$

- Da aus einer „großen Menge“ gezogen wird, kann das Ziehen als Ziehen mit Zurücklegen angesehen werden. Folglich wird die Binomialverteilung verwendet.
- Wahrscheinlichkeit, dass k von n Teilen Ausschuß sind:

$$P(X = k) = \binom{n}{k} p^k q^{n-k} = \binom{n}{k} 0,1^k \cdot 0,9^{n-k}$$

Fall 3a)

$$P(X > 1) = \sum_{j=2}^{15} \binom{15}{j} 0,1^j \cdot 0,9^{15-j} \approx 0,451 \tag{9}$$

Fall 3b)

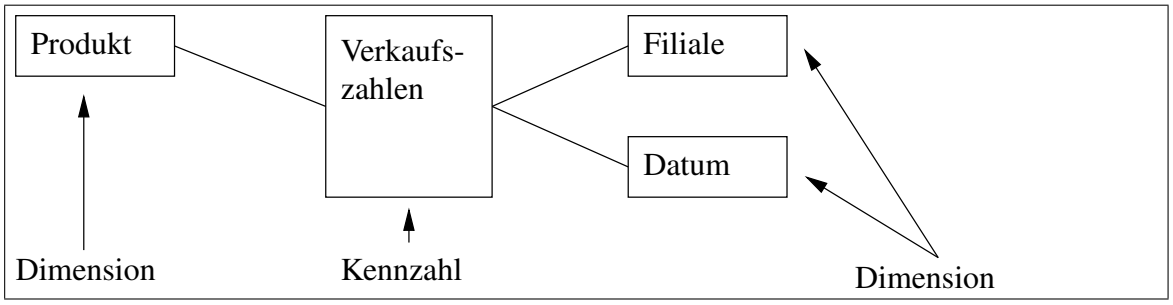
$$P(X > 2) = \sum_{j=3}^{30} \binom{30}{j} 0,1^j \cdot 0,9^{30-j} \approx 0,589 \tag{10}$$

- Aus den Ergebnissen folgt, dass die Wahrscheinlichkeit, bei 30 Stichproben mehr als zwei an Ausschuß zu erwischen größer ist, als bei 15 Proben mehr als eine zu finden. Dadurch werden beim zweiten Verfahren mehr Sendungen abgelehnt.

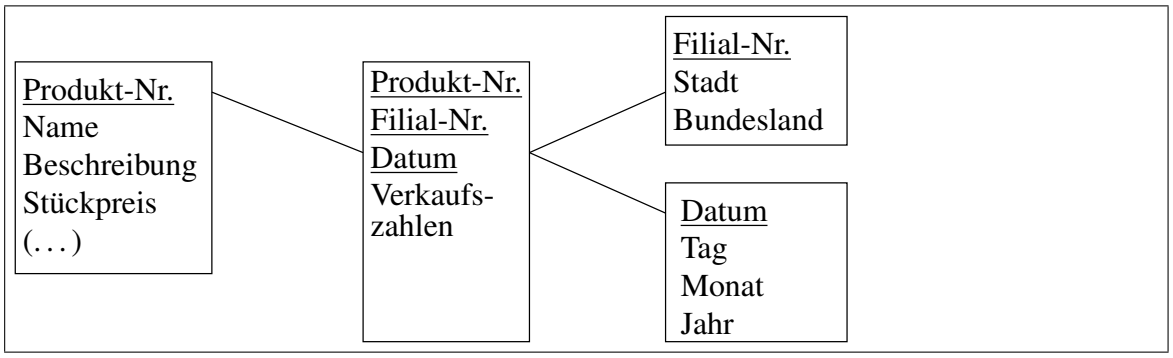
2 Stern-Schema

Die Supermarktkette IDLA möchte ihre Lagerkosten optimieren. Dazu hat sie Daten darüber gesammelt, welche Produkte in welchen Filialen und in welcher Menge über einen Zeitraum von zwei Jahren verkauft worden sind. Die Einheit der Zeitmessung sind Tage. Zur Analyse dieser Daten möchte IDLA ein OLAP System einsetzen.

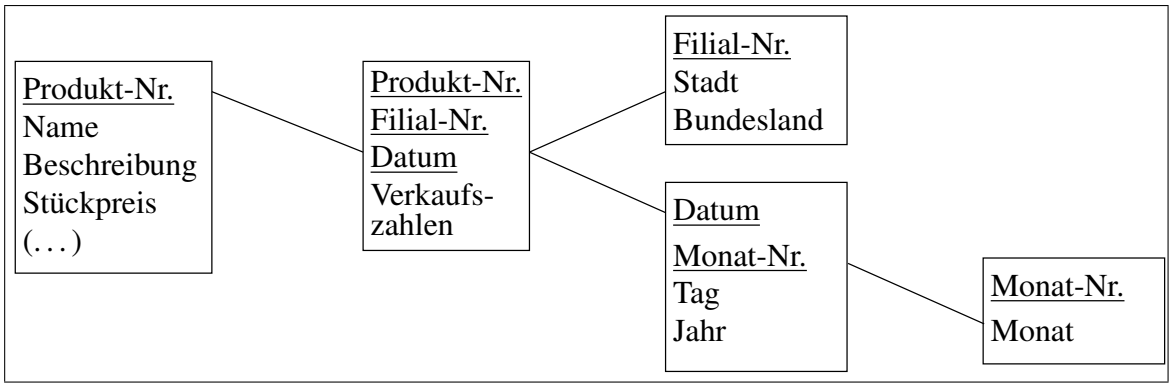
1. Entwerfen sie ein Stern-Schema für die Analyse dieser Daten. Geben sie zuerst die Kennzahl und die Dimensionen an!



2. Skizzieren Sie sowohl die Kennzahlentabelle als auch die Dimensionstabellen.



3. Erweitern Sie das obige Sternschema zu einem Schneeflockenschema wenn zusätzlich die Tagesdaten zu Monatsdaten aggregiert werden sollen.



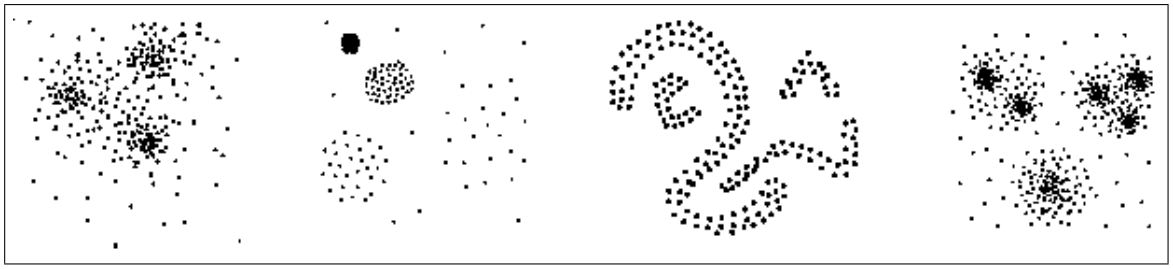
4. Wie würden Sie die Daten visualisieren, damit der Logistikexperte der Firma IDLA möglichst effizient die Logistikplanung für den nächsten Zeitraum vornehmen kann?

3 Allgemeines zum Clustern

1. Beschreiben Sie kurz was man unter Clustern versteht.

   - Identifikation einer endlichen Menge von Kategorien, Klassen oder Gruppen (*Cluster*) in den Daten
   - Objekte im *gleichen* Cluster sollen möglichst ähnlich sein
   - Objekte aus *verschiedenen* Clustern sollen möglichst unähnlich zueinander sein

2. Geben Sie verschiedene Clusterformen an.



3. Diskutieren Sie mögliche Probleme, die beim Entdecken der verschiedenen Cluster durch unterschiedliche Verfahren auftreten können.

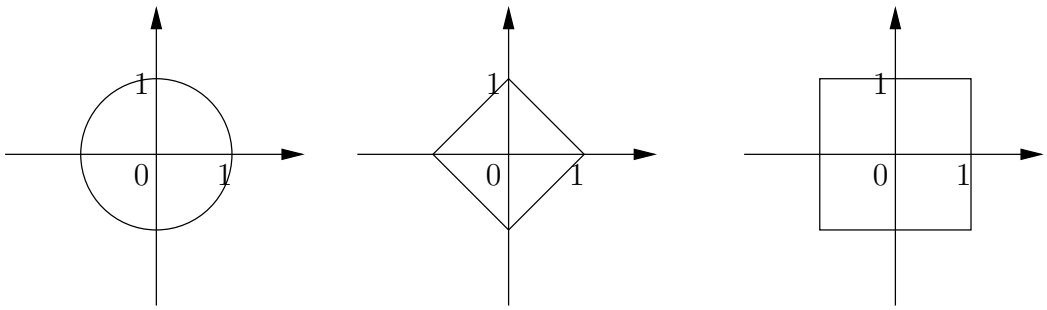
   - Nicht jedes Verfahren kann die gleichen Formen und die gleichen Varianten an Clustern entdecken. So kann KMeans nur konvexe Cluster entdecken.
   - Dichtebasierte Verfahren sind in der Lage, Cluster unterschiedlichster Formen zu entdecken.

4. Geben Sie eine typische Distanz- und eine typische Ähnlichkeitsfunktion an und diskutieren Sie die Beziehung zwischen beiden Funktionen (im allgemeinen).

   - Euklidische Distanz
   - Cosinus-Ähnlichkeit

   Während Distanzfunktionen in einem gegebenen Raum die Entfernung zwischen zwei Objekten widerspiegeln und 0 sind, wenn zwei Objekte an der gleichen Stelle im Raum sind, geben Ähnlichkeitsfunktionen an, wie ähnlich sich zwei Objekte sind und sind am größten wenn die Ähnlichkeit der beiden Objekte am größten ist, also die Distanz am kleinsten.

5. Veranschaulichen Sie sich einige Distanzfunktionen, indem sie für jede Funktion in der reellen Zahlenebene ($\mathbb{R}^2$) alle Punkte mit der Distanz 1 zum Ursprung $(0, 0)$ einzeichnen.



Euklidische Distanz	Manhattan-Distanz	Maximums-Metrik
$d(x,0) = \sqrt{x_1^2 + x_2^2} = 1$	$d(x,0) = \lvert x_1\rvert + \lvert x_2\rvert = 1$	$d(x,0) = \max\{\lvert x_1\rvert, \lvert x_2\rvert\} = 1$