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.

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.

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?

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.

2. Geben Sie verschiedene Clusterformen an.

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

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

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.