9. Übungsblatt

Aufgabe 32 Statistische Signifikanz

Wir betrachten eine Transaktionsdatenbank mit 1000 Transaktionen über 100 Items. Nehmen Sie an, daß jedes Item mit Wahrscheinlichkeit 0.05 in einer Transaktion auftritt, so daß jedes Item im Durchschnitt in 50 Transaktionen enthalten ist. Weiter seien die Transaktionen unabhängig und die Items mögen unabhängig von einander in ihnen auftreten, d.h., die Wahrscheinlichkeit des gemeinsamen Auftretens einer Menge von Items kann berechnet werden als Produkt der Wahrscheinlichkeiten des einzelnen Auftretens der Items.

a) Mit welcher Wahrscheinlichkeit treten die Items mit den Nummern 7, 16, 42 und 78 in *genau* vier Transaktionen zusammen auf?

b) Mit welcher Wahrscheinlichkeit treten die Items mit den Nummern 7, 16, 42 und 78 in *mindestens* vier Transaktionen zusammen auf?

c) Wenn man die Wahrscheinlichkeit berechnen möchte, mit der irgendeine Menge mit vier Items (*nicht* eine bestimmte Menge wie in a) und b), sondern eine beliebige Menge mit vier Items) in mindestens vier Transaktionen zusammen auftritt, wie könnte man ansetzen? Wo liegen die wesentlichen Probleme? Warum ist eine direkte Erweiterung des Ergebnisses aus b) eine Überschätzung?

Aufgabe 33 Surrogatdaten und Mustersignaturen

Um die statistische Signifikanz von gefundenen Itemmengen zu bestimmen, kann man einen auf Surrogat(transaktions)daten und Mustersignaturen basierenden Ansatz verwenden

a) Was ist ein Surrogatdatensatz? Wie könnte man einen Surrogatdatensatz aus einer gegebenen Transaktionsdatenbank erzeugen? Was repräsentieren Surrogatdatensätze?

b) Welche Eigenschaften des Originaldatensatzs sollten beim Erzeugen von Surrogatdatensätzen erhalten werden?

c) Wie viele Surrogatdatensätze sollte man erzeugen? Ist einer genug? Sind zehn genug? Sind 1000 genug? Gibt es eine allgemein richtige Anzahl?

d) Was ist eine Mustersignatur? Warum ist es besser, Mustersignaturen zu betrachten statt konkreter Muster?

e) Wie werden gefundene Itemmengen mit Hilfe von Surrogatdatensätzen gefiltert? Liefert dieses Filtern nur relevante Muster?

Aufgabe 34 Assoziationsregeln

a) Was sind Assoziationsregeln? Mit welchen Maßen werden sie bewertet?
Was ist der Support einer Assoziationsregel (zwei Versionen)?
Was ist die Konfidenz einer Assoziationsregel?

b) Wie werden Assoziationsregeln induziert? Welche Schritte werden benötigt? Wie verhält sich die Induktion von Assoziationsregeln zum Finden häufiger Itemmengen?

c) Mit welchem minimalen Support müssen häufige Itemmengen in der Assoziationsregelinduktion gefunden werden? Ist es der minimale Support der Assoziationsregeln? Oder hängt es von der Wahl der Definition des Regelsupports ab? Wenn ja, wie?

d) Welche Beziehung besteht zwischen den Konfidenzen verschiedener Assoziationsregeln, die aus der gleichen Itemmenge gebildet werden? Wie kann diese Beziehung bei der Erzeugung von Assoziationsregeln ausgenutzt werden?

Aufgabe 35 Assoziationsregeln

a) Sind die beiden Definitionen des Regelsupports äquivalent? D.h., gibt es eine Transaformation, die man auf den verwendeten minimalen Support anwenden kann, so daß die gleichen Regeln gefunden werden? Begründen Sie Ihre Antwort!

a) Mit welchen anderen Maßen, außer dem minimalen Support und der minimalen Konfidenz, kann man Assoziationsregeln bewerten? Aus welchen Größen werden diese Maße gewöhnlich berechnet?

b) Was ist der *lift (value)* einer Assoziationsregel und wie ist er definiert?
Was ist der Informationsgewinn einer Assoziationsregel?
Was ist das χ^2-Maß einer Assoziationsregel?

c) Sind Assoziationsregeln mit mehr als einem Item in Konsequens nützlich?
Welche zusätzliche Information enthalten sie?