

Data Mining und Maschinelles Lernen



TECHNISCHE
UNIVERSITÄT
DARMSTADT

13. Übungsblatt

Aufgabe 1 Apriori

Gegeben seien folgende Beobachtungen vom Kaufverhalten von Kunden:

beer	chips	dip	pizza	wine
1	1	1	0	0
1	1	1	1	0
0	1	1	0	1
1	1	0	1	0
0	1	1	1	1
0	0	0	1	1
1	1	1	1	0
1	0	0	0	1
1	0	0	1	0
1	1	1	1	1

- Wenden Sie den Apriori Algorithmus (Foliensatz “Clustering und Lernen von Assoziationsregeln”, Folie “APRIORI Step 1: FreqSet: Find all Frequent Itemsets”) auf den obigen Datensatz an, um zunächst alle Frequent Itemsets zu bestimmen. Benutzen Sie hierfür einen minimalen Support von 3 Beispielen.
- Erstellen Sie für alle drei- und vierelementigen Itemsets die Assoziationsregeln mit einer minimalen Konfidenz von 0.8, wie im zweiten Teil des Algorithmus beschrieben (Folie “APRIORI Step 2: Generate Association Rules”). Um sich Arbeit zu sparen, können Sie die Antimonotonie der Konfidenz ausnutzen.

Aufgabe 2 Assoziationsregel-Maße

Ein Online-Buchgeschäft möchte eine Datenbank mit 10,000 Kunden analysieren, die jeweils eines oder mehrere von 500 verschiedenen Büchern gekauft haben. Zur Entdeckung von Assoziationsregeln wird der Algorithmus Apriori mit einem Minimum Support von 3% und einer minimalen Konfidenz von 75% verwendet.

- Es wird festgestellt, daß die beiden häufigsten Verkäufe “Harry Potter und der Stein der Weisen” (HP1) und “Harry Potter und die Kammer des Schreckens” (HP2) sind. HP1 wurde von 6,000 Kunden und HP2 von 8,000 Kunden gekauft. 4,000 Kunden kauften beide Bücher.

Welche der beiden Assoziationsregeln findet sich im Output des Assoziationsregel-Lerners?

- HP1 → HP2
- HP2 → HP1
- beide
- keine von beiden

Geben Sie Support und Konfidenz für beide Regeln an.

- Wenn man annimmt, daß alle Kunden, die beide Bücher gekauft haben, zuerst HP1 und später HP2 gekauft haben: Wie interpretieren Sie den Einfluß des Kaufs von HP1 auf den Kauf von HP2?
- Die längste Assoziationsregel, die gefunden wurde, wurde aus einem Itemset der Größe 20 konstruiert. Geben Sie eine möglichst große untere Schranke für die Anzahl der gefundenen Frequent Itemsets an.