Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

20.11.2008

4. Übung „Knowledge Discovery“

Wintersemester 2008/2009

1 Clusterverfahren

Ein Kaufhaus, das seine Kunden in fünf Gruppen klassifiziert hat, möchte eine Werbekampagne durchführen. Da es zu aufwendig wäre, für jede der fünf Gruppen ein spezifisches Werbekonzept zu konzipieren, sollen sie in zwei Hauptgruppen eingeteilt werden. Dazu hat man für die Repräsentanten der einzelnen Gruppe $\{1, 2, 3, 4, 5\}$ die folgenden Abstände d ermittelt:

$D(x, y)$	1	2	3	4	5
1	0	2	2	17	16
2	2	0	4	9	10
3	2	4	0	13	10
4	17	9	13	0	1
5	16	10	10	1	0

1. Entwerfen Sie ein Verfahren, welches ausgehend von einer Anfangsklassifikation K^0 durch den Austausch von Elementen die Klassifikation iterativ bezüglich eines Güteindex optimiert (Austauschverfahren).

2. Ausgehend von der Anfangsklassifikation $K^0 = \{\{1, 2\}, \{3, 4, 5\}\}$ soll mit Hilfe des Austauschverfahrens die bestmögliche Klassifikation K mit dem Güteindex

$$b(K) = \sum_{C \in K} \left(\frac{1}{|C|} \sum_{x,y \in C} d(x, y) \right)$$

   erstellt werden.

3. Welche anderen Verfahren hätte man zur Lösung der Aufgabe auch verwenden können? (Führen Sie ein Verfahren durch und vergleichen Sie die Ergebnisse. Zusatzaufgabe.)

4. Wie kann das Kaufhaus die Ergebnisse zur Aufstellung der Marketingstrategien verwenden?

2 k-Means Clustering

1. Gegeben folgender Datensatz:

x	1	6	8	3	2	2	6	6	7	7	8	8
y	5	2	1	5	4	6	1	8	3	6	3	7

Ermitteln Sie mit Hilfe von k-Means eine Clusterung mit $k = 3$. Verwenden Sie als Centroide die ersten drei Datentupel und verfolgen Sie die Wanderung der Centroide.

2. Betrachten Sie folgenden zweidimensionalen klassifizierten Datensatz zunächst ohne die Information über die Klasse für jedes Tupel.

x	3	3	4	4	5	6	7	7	8	9	1	2	2	3	4	5	5	6	7	7
y	1	2	2	3	3	4	4	6	5	7	3	4	5	6	6	7	8	8	8	9
Klasse	a	a	a	a	a	a	a	a	a	a	b	b	b	b	b	b	b	b	b	b

Welches Problem ergibt sich bei der Anwendung des k-Means-Algorithmus mit $k = 2$ (d. h. zwei Clustern) auf diesem Datensatz?
Hinweis: Überlegen Sie sich, wie das gewünschte Ergebnis aussieht. Was liefert der k-Means-Algorithmus stattdessen? (Sie brauchen das exakte Ergebnis des Algorithmus nicht auszurechnen, eine qualitative Beschreibung reicht.)

3 Erwartungswertmaximierung

1. Geben Sie das prinzipielle Vorgehen des EM-Algorithmus wieder.

2. Geben Sie die Formel zur Berechnung der $P(C_i \mid x)$ und der Modellparameter für $k = 2$ Cluster an. Verwenden Sie dazu folgende Gleichungen:

$$W_i = \frac{1}{n} \sum_{x \in D} P(C_i \mid x) \tag{1}$$

$$\mu_{C_i} = \frac{\sum_{x \in D} xP(C_i \mid x)}{\sum_{x \in D} P(C_i \mid x)} \tag{2}$$

$$\sigma_{C_i} = \frac{\sum_{x \in D} P(C_i \mid x)(x - \mu_{C_i})^2}{\sum_{x \in D} P(C_i \mid x)} \tag{3}$$

3. Zeichnen Sie ein Histogramm für die untenstehenden Daten.

4. Berechnen Sie mittels EM für die untenstehenden Daten eine Clusterung für $k = 2$ Cluster ausgehend von $\mu_1 = 0,12$ und $\mu_2 = 5,28$, $\sigma_1 = 1$ und $\sigma_2 = 1$ sowie gleichwahrscheinlicher prior Wahrscheinlichkeit für die Zugehörigkeit der Objekte zu den Clustern. Führen Sie nur den ersten Schritt aus.

-0,39 0,12 0,94 1,67 1,76 2,44 3,72 4,28 4,92 5,53 0,06 0,48 1,01 1,68 1,8 3,25 4,12 4,6 5,28 6,22

5. Skizzieren Sie in Ihrem Diagramm die Kurven der Funktionen $P(x \mid C_1)$ und $P(x \mid C_2)$.