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 (gegebenen) Güteindex optimiert (Austauschverfahren).

   z. B. gegenseitiges Vertauschen eines Elementes

   - $K^0 = \{\{1, 2\}, \{3, 4, 5\}\}$
   - vertausche Element aus C_1 mit Element aus C_2, so dass der Güteindex sich verbessert

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.

$$b(K^0) = \frac{1}{2}2 + \frac{1}{3}(13 + 10 + 1) = 1 + 8 = 9$$

Element	neue Clusterung	$b()$	Vergleich
1	$\{\{2\},\{1,3,4,5\}\}$	17.25	> 9
2	$\{\{1\},\{2,3,4,5\}\}$	12.25	> 9
3	$\{\{1,2,3\},\{4,5\}\}$	3.17	< 9
4	$\{\{1,2,4\},\{3,5\}\}$	14.33	> 9
5	$\{\{1,2,5\},\{3,4\}\}$	15.83	> 9

Die Clusterung $K^1 = \{\{1,2,3\},\{4,5\}\}$ ist nach diesem Verfahren optimal, es ist keine weitere Verbesserung durch Austausch möglich.

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

z. B. Verwendung des hierarchischen *single-linkage-Verfahrens* mit Verschiedenheitsfunktion $\nu(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y)$

Iterationschritt 1:

$$
\begin{aligned}
K^0 &= \{\{1\},\{2\},\{3\},\{4\},\{5\}\} = \{K_1^0, \ldots, K_5^0\} \\
\nu(C_{t_1}^0, C_{t_2}^0) &= d_{t_1 t_2} \\
\Rightarrow \min_{C_{t_1}^0, C_{t_2}^0 \in K^0} &= \nu(C_{t_1}^0, C_{t_2}^0) = d_{45} = 1 \\
\Rightarrow K^1 &= \{\{1\},\{2\},\{3\},\{4,5\}\}
\end{aligned}
$$

Iterationschritt 2:

$$
\begin{aligned}
K^1 &= \{\{1\},\{2\},\{3\},\{4,5\}\} = \{K_1^1, \ldots, K_4^1\} \\
(\nu(C_{t_1}^1, C_{t_2}^1))_{t_1=1\ldots4, t_2=1\ldots4} &= \begin{pmatrix} 0 & & & \\ 2 & 0 & & \\ 2 & 4 & 0 & \\ 16 & 9 & 10 & 0 \end{pmatrix} \\
\min_{C_{t_1}^1, C_{t_2}^1 \in K^1} &= \nu(C_{t_1}^1, C_{t_2}^1) = \nu(C_1^1, C_2^1) = \nu(C_1^1, C_3^1) = 2 \\
\Rightarrow K^2 &= \{\{1,2\},\{3\},\{4,5\}\}
\end{aligned}
$$

Iterationschritt 3:

$$
\begin{aligned}
K^2 &= \{\{1,2\},\{3\},\{4,5\}\} = \{K_1^2, \ldots, K_3^2\} \\
(\nu(C_{t_1}^2, C_{t_2}^2))_{t_1=1,2,3, t_2=1,2,3} &= \begin{pmatrix} 0 & & \\ 2 & 0 & \\ 9 & 10 & 0 \end{pmatrix} \\
\min_{C_{t_1}^2, C_{t_2}^2 \in K^2} &= \nu(C_{t_1}^2, C_{t_2}^2) = \nu(C_1^2, C_2^2) = 2 \\
\Rightarrow K^3 &= \{\{1,2,3\},\{4,5\}\}
\end{aligned}
$$

Es ergibt sich das gleiche Ergebnis wie im ersten Teil der Aufgabe!

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

Durch Entwurf zweier Werbekonzepte: eines für Kunden der Gruppen 1,2,3 und eines für Kunden der Gruppen 4,5. Eine gruppenspezifische Marketingstrategie lässt sich mit den vorliegenden Daten nicht aufstellen, da über die Kaufeigenschaften und Charakteristiken der Gruppen keine Informationen vorliegen.

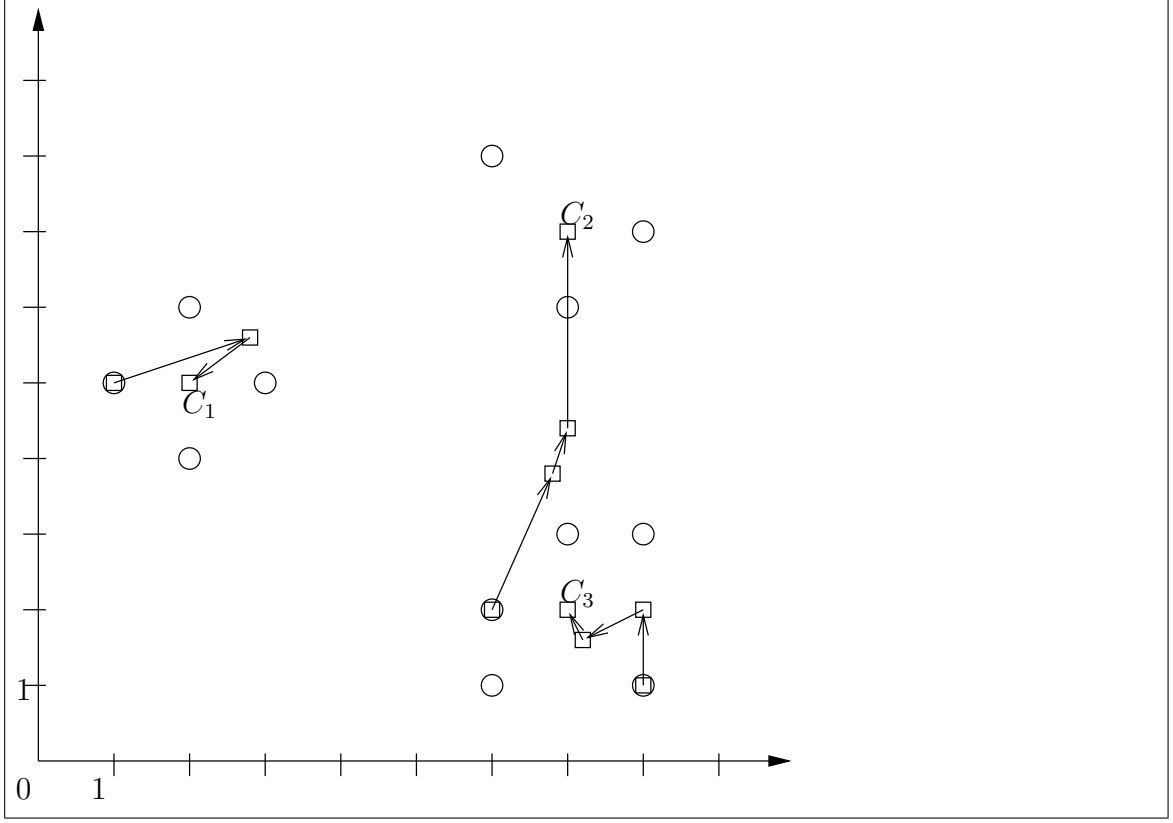
2 k-Means Clustering

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.

It.	C_1	$K(C_1)$	C_2	$K(C_2)$	C_3	$K(C_3)$
1	$(1,5)$	$\{(1,5),(3,5),(2,4),(2,6),(6,8)\}$	$(6,2)$	$\{(6,2),(6,1),(7,3),(7,6),(8,7)\}$	$(8,1)$	$\{(8,1),(8,3)\}$
2	$(\frac{14}{5},\frac{28}{5})$	$\{(1,5),(3,5),(2,4),(2,6),(6,8)\}$	$(\frac{34}{5},\frac{19}{5})$	$\{(6,2),(7,3),(7,6),(8,7)\}$	$(8,2)$	$\{(6,1),(8,1),(8,3)\}$
3	$(\frac{14}{5},\frac{28}{5})$	$\{(1,5),(3,5),(2,4),(2,6)\}$	$(\frac{14}{2},\frac{9}{2})$	$\{(6,8),(7,6),(8,7)\}$	$(\frac{22}{3},\frac{5}{3})$	$\{(6,1),(6,2),(7,3),(8,1),(8,3)\}$
4	$(2,5)$	$\{(1,5),(3,5),(2,4),(2,6)\}$	$(7,7)$	$\{(6,8),(7,6),(8,7)\}$	$(7,2)$	$\{(6,1),(6,2),(7,3),(8,1),(8,3)\}$

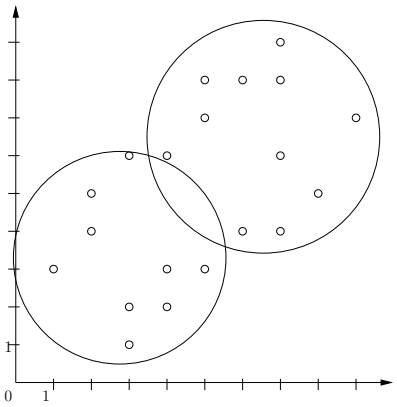


2. Betrachten Sie folgenden zweidimensionalen klassifizierten Datensatz zunächst ohne die Information über die Klasse für jedes Tupel.

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.)



Der Algorithmus liefert eine Clusterung wie im Diagramm gezeigt. Diese entspricht nicht der gewünschten, die gleich der Klassifizierung sein sollte. Die längliche Form der gewünschten Clusterung stellt für k-Means eine instabile Lösung dar, da Punkte des anderen Clusters näher am Centroid sind als Punkte des eigenen Clusters. Kleine Änderungen am Centroid bewirken daher schon ein Verrutschen der Centroide zur obigen Clusterung.

3 Erwartungswertmaximierung

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

   Der EM-Algorithmus arbeitet in einen zweistufigen iterativen Verfahren. Im ersten Schritt werden für jedes Cluster eine Wahrscheinlichkeitsverteilung und damit für jeden Punkt die Wahrscheinlichkeit berechnet, daß er in einem bestimmten Cluster liegt. Im zweiten Schritt werden die Modellparameter neu berechnet.

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:

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

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

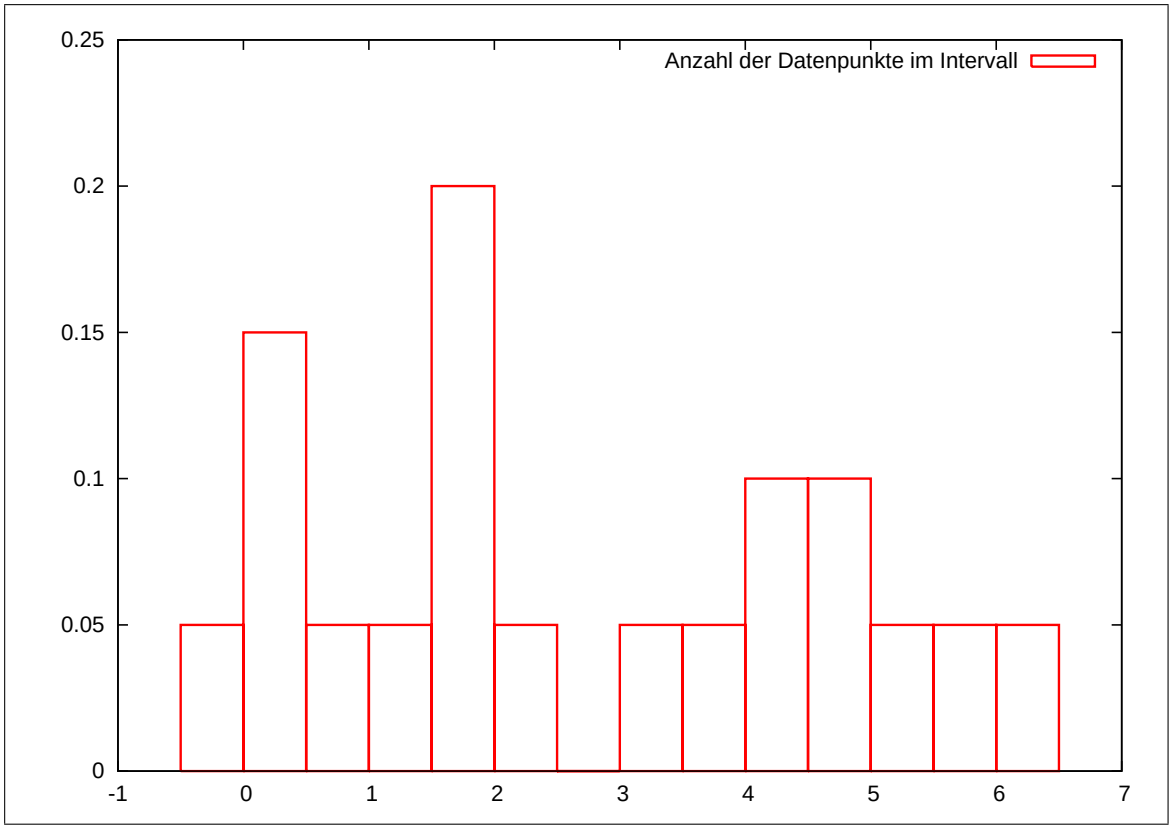
$$\sigma_{C_i} = \frac{\sum_{x \in D} P(C_i \mid x)(x - \mu_{C_i})^2}{\sum_{x \in D} P(C_i \mid x)} \tag{3}$$

$$P(x \mid C_i) = \frac{1}{\sigma_{C_i}\sqrt{2\pi}} e^{-\frac{(x-\mu_{C_i})^2}{2\sigma_{C_i}^2}} \tag{4}$$

$$P(x) = W_1 P(x \mid C_2) + W_2 P(x \mid C_2) \tag{5}$$

$$P(C_i \mid x) = W_i \frac{P(x \mid C_i)}{P(x)} \tag{6}$$

3. Zeichnen Sie ein Histogramm für die untenstehenden Daten.

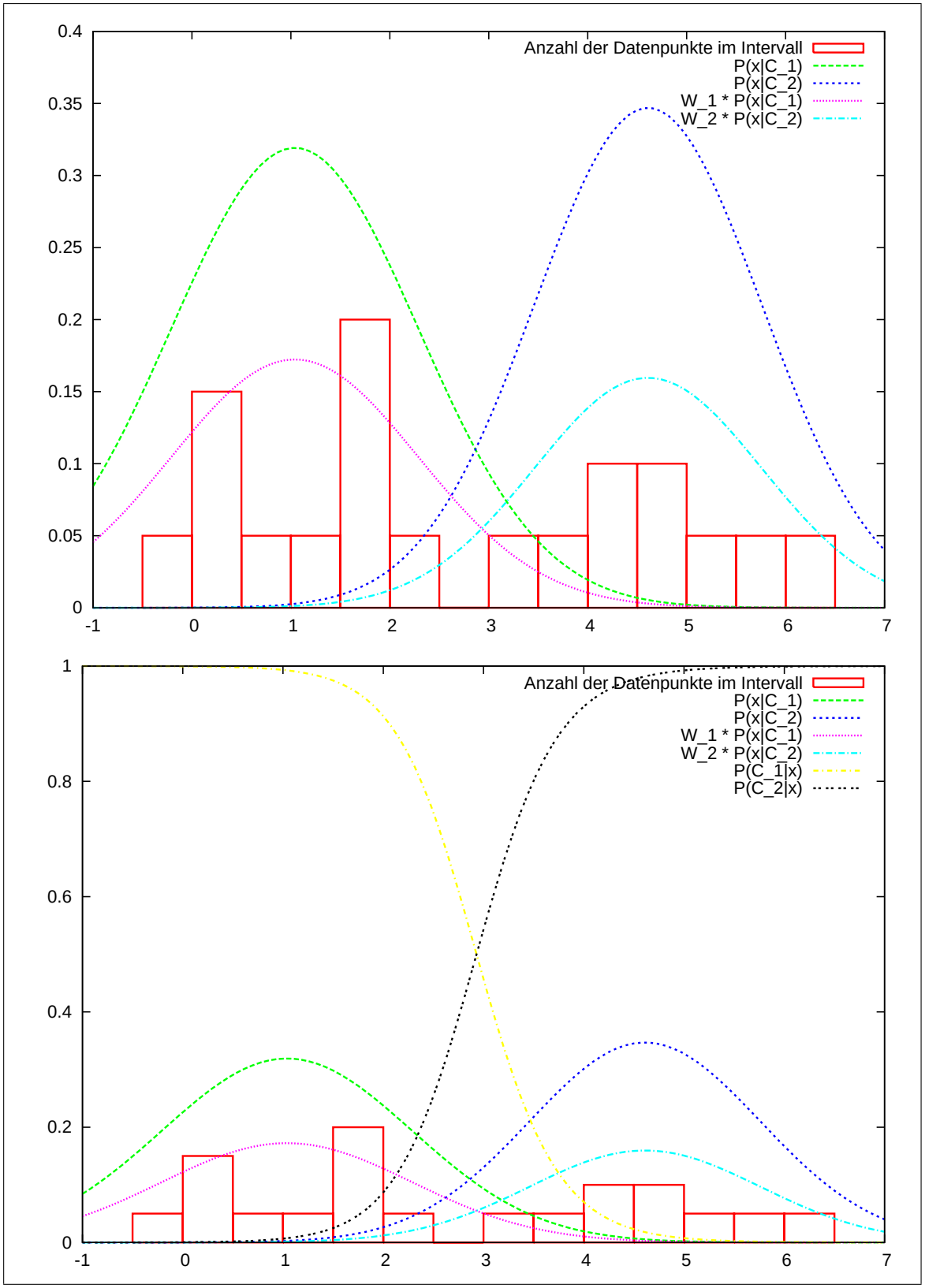


4. Berechnen Sie mittels EM für die untenstehenden Daten eine Clusterung für $k = 2$ Cluster ausgehend von $\mu_1 = 0,12$ und $\mu_2 = 5,28$, $\sigma_1 = 1$ und $\sigma_2 = 1$ sowie gleichwahrscheinlicher prior Wahrscheinlichkeit für die Zugehörigkeit der Objekte zu den Clustern. Führen Sie nur den ersten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

μ_1	μ_2	σ_1	σ_2	W_1	W_2
0,12	5,28	1	1	0,5	0,5
1,04	4,61	1,25	1,15	0,54	0,46

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


Anzahl der Datenpunkte im Intervall
P(x|C_1)
P(x|C_2)
W_1 * P(x|C_1)
W_2 * P(x|C_2)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-1
0
1
2
3
4
5
6
7
Anzahl der Datenpunkte im Intervall
P(x|C_1)
P(x|C_2)
W_1 * P(x|C_1)
W_2 * P(x|C_2)
P(C_1|x)
P(C_2|x)
1
0.8
0.6
0.4
0.2
0
-1
0
1
2
3
4
5
6
7