Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

18.5.2006

4. Übung „Knowledge Discovery“

Sommersemester 2006

1 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

2 DBSCAN

Das dichte-basierte Clustermodell von DBSCAN definiert Kernpunkt und Randpunkte eines Clusters. Randpunkte sind Punkte, die zu einem Cluster gehören, weil sie dichte-erreichbar von Kernpunkten sind, aber selbst keine Kernpunkte sind. Wie der Name schon sagt, sind das Punkte, die am Rand eines Clusters liegen.

1. Kann es Randpunkte geben, die gleichzeitig zu verschiedenen Clustern gehören?
2. Wie geht DBSCAN mit solchen Randpunkten um, d. h. welchem Cluster werden diese Punkte zugeordnet?

   Nennen Sie eine bessere Lösung zur Zuordnung dieser Randpunkte. Begründen Sie Ihre Entscheidung.

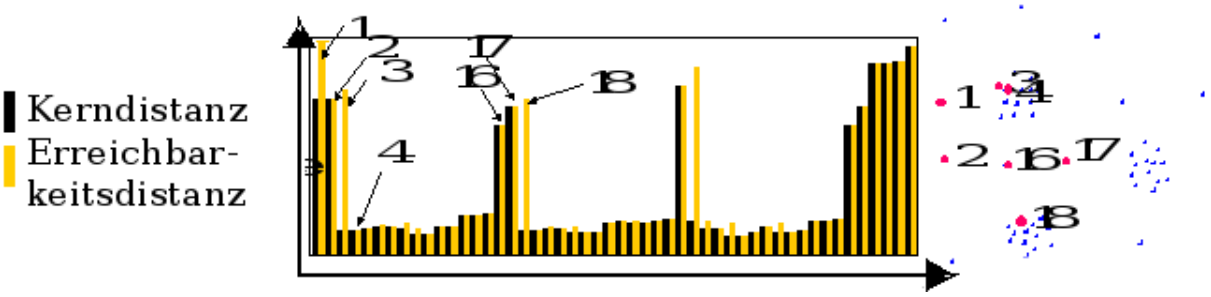
3. Es sei eine Menge von Punkten gegeben. Es sollen Cluster mittels DBSCAN mit den Parametern $minPts = 4$ und $\varepsilon = 10$ gebildet werden. Die folgende Tabelle gibt die Abstände zwischen den Punkten an:

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P1	0	54,1	5,0	8,2	13,0	62,5	69,1	59,4	65,3	49,7	55,9
P2	54,1	0	49,1	46,4	43,9	10,4	15,0	12,4	13,0	7,3	11,7
P3	5,0	49,1	0	4,1	9,5	57,5	64,1	54,5	60,4	44,7	50,9
P4	8,2	46,4	4,1	0	5,4	55,2	61,4	52,5	57,3	42,4	49,0
P5	13,0	43,9	9,5	5,4	0	53,2	58,8	51,0	54,1	40,6	47,7
P6	62,5	10,4	57,5	55,2	53,2	0	9,1	6,3	14,1	12,8	9,0
P7	69,1	15,0	64,1	61,4	58,8	9,1	0	15,3	9,1	20,2	18,0
P8	59,4	12,4	54,5	52,5	51,0	6,3	15,3	0	20,0	10,8	3,6
P9	65,3	13,0	60,4	57,3	54,1	14,1	9,1	20,0	0	20,1	21,5
P10	49,7	7,3	44,7	42,4	40,6	12,8	20,2	10,8	20,1	0	8,1
P11	55,9	11,7	50,9	49,0	47,7	9,0	18,0	3,6	21,5	8,1	0

Bilden Sie die Cluster. Numerieren Sie diese Cluster und tragen Sie in eine Tabelle für jeden Punkt die Punkt-Art (Kernobjekt, Randobjekt, Rauschen) sowie die Nummer des Clusters, dem der Punkt angehört, ein.

In welchem Wertebereich müsste ε liegen (bei $MinPts = 4$), damit alle Punkte als „Rauschen“ identifiziert werden? Begründen Sie ihre Antwort!

3 OPTICS



Es sei OPTICS auf eine Datenbank mit den Parametern ε und $MinPts$ angewendet worden. Geben Sie ein Verfahren an, mit dem man aus dem Resultat des OPTICS-Laufes (Clusterordnung, Erreichbarkeitsdiagramm und Kerndistanzdiagramm) das DBSCAN-Clustering für ein gegebenes $\varepsilon' \leq \varepsilon$ extrahieren kann. Benutzen Sie möglichst intuitiven Pseudocode.
Kann aus dem OPTICS-Ergebnis eine eindeutige Clusterzugehörigkeit abgeleitet werden, die DBSCAN bzgl. des gegebenen $\varepsilon' \leq \varepsilon$ erzeugen würde? Mit anderen Worten, stimmt das Ergebnis ihres Verfahrens exakt mit dem Ergebnis eines DBSCAN-Laufes bzgl. ε' überein? Begründen Sie Ihre Antwort!