

## Knowledge Discovery

### Übungsblatt 4

Sommersemester 2005

#### Aufgabe 1: Clustering EM

- |  |       |
|--|-------|
|  | -0,39 |
|  | 0,12  |
|  | 0,94  |
| a) Geben Sie das prinzipielle Vorgehen des EM-Algorithmus wieder.  | 1,67  |
| b) Geben Sie die Formel zur Berechnung der $P(C_i x)$ und der Modellparameter für $k = 2$ Cluster an. Verwenden Sie dazu folgende Gleichungen:   | 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   |
| c) Zeichnen Sie ein Histogramm für die nebenstehenden Daten.   | 5,28  |
| d) Berechnen Sie mittels EM für die nebenstehenden Daten eine Clustering für $k = 2$ Cluster ausgehend von $\mu_1 = 0,12$ und $\mu_2 = 5,28$ , $\text{std}_1 = 1$ und $\text{std}_2 = 1$ sowie gleichwahrscheinlicher prior Wahrscheinlichkeit für die Zugehörigkeit der Objekte zu den Clustern. Führen Sie nur den ersten Schritt aus. | 6,22  |

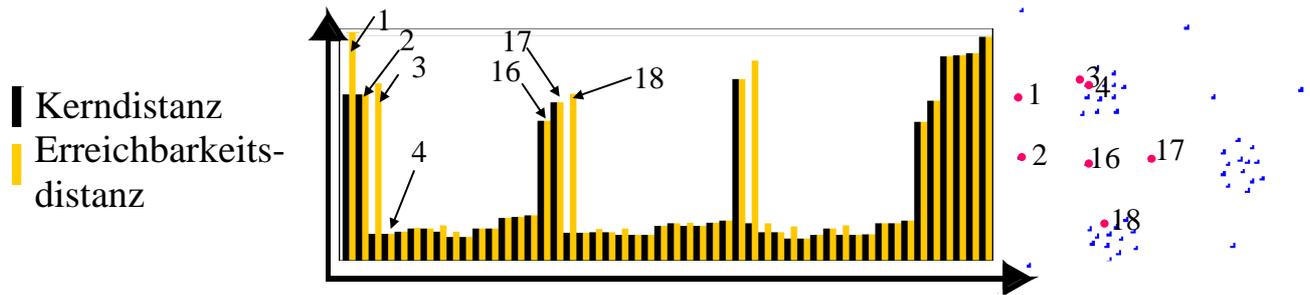
#### Aufgabe 2: DBScan

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.

- Kann es Randpunkte geben, die gleichzeitig zu verschiedenen Clustern gehören?
- 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

### Aufgabe 3: Optics



Es sei OPTICS auf eine Datenbank mit den Parametern  $\epsilon$  und  $\text{MinPts}$  angewendet worden.

Geben Sie ein Verfahren an, wie man aus dem Resultat des OPTICS Laufes (Clusterordnung, Erreichbarkeitsdiagramm und Kerndistanzdiagramm) das DBSCAN-Clustering für ein gegebenes  $\epsilon' \leq \epsilon$  extrahieren kann! Benutzen Sie möglichst intuitiven Pseudocode.

Kann aus dem OPTICS-Ergebnis eine eindeutige Clusterzugehörigkeit abgeleitet werden, die DBSCAN bzgl. des gegebenen  $\epsilon' \leq \epsilon$  erzeugen würde? Mit anderen Worten, stimmt das Ergebnis ihres Verfahrens exakt mit dem Ergebnis eines DBSCAN-Laufes bzgl.  $\epsilon'$  überein? Begründen Sie Ihre Antwort!