Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Peer Kröger, Dr. Matthias Schubert
Markus Bundschus

Knowledge Discovery in Databases

WS 2007/08

Übungsblatt 8

Aufgabe 8-1 EM-Algorithmus

Übungsaufgabe

Gegeben sei eine Datenmenge D mit 100 Punkten, die drei Gausscluster A, B und C und den Punkt p enthält.

Der Cluster A ist repräsentiert durch den Mittelwert aller seiner Punkte (2,2) und die Kovarianzmatrix $\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$.

30 Prozent aller Punkte gehören zu diesem Cluster.

Der Cluster B ist repräsentiert durch den Mittelwert aller seiner Punkte (5,3) und die Kovarianzmatrix $\begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$.

20 Prozent aller Punkte gehören zu diesem Cluster.

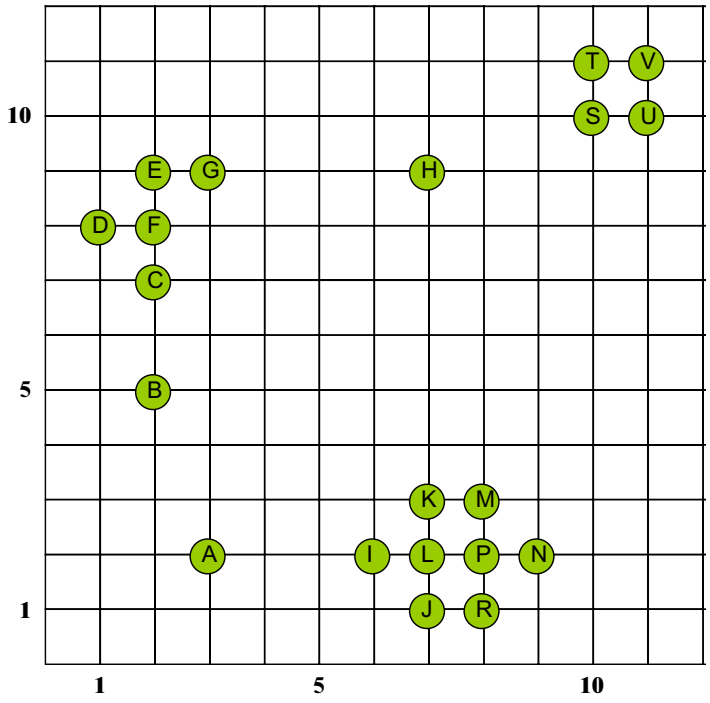
Der Cluster C ist repräsentiert durch den Mittelwert aller seiner Punkte (1,4) und die Kovarianzmatrix $\begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$.

50 Prozent aller Punkte gehören zu diesem Cluster.

Der Punkt p ist durch die Koordinaten (2.5,3.0) gegeben. Geben Sie die drei Wahrscheinlichkeiten an, mit der p zum Cluster A, B bzw. C gehört.

Aufgabe 8-2 Single-Link

Gegeben sei der folgende Datensatz:



Als Distanzfunktion zwischen den Punkten dient Ihnen jeweils wieder die Manhattan-Distanz (L_1-Norm):

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

Berechnen Sie zwei Dendrogramme für diesen Datensatz. Als Distanzfunktion zwischen Mengen von Objekten verwenden Sie

(a) den Single-Link Ansatz,

(b) den Average-Link Ansatz.

Tipp: Innere Knoten müssen nicht binär sein, d.h. sie können mehr als zwei Söhne haben.