Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Matthias Schubert
Erich Schubert

Knowledge Discovery in Databases

WS 2009/10

Übungsblatt 8

Aufgabe 8-1 EM-Algorithmus
Hausaufgabe

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

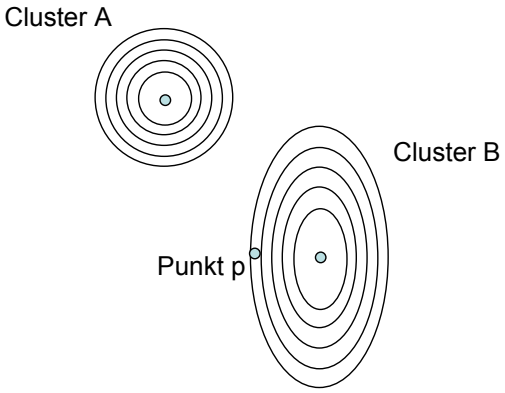
Der Cluster A ist repräsentiert durch den Mittelwert aller seiner Punkte (1,1) und die Kovarianzmatrix $\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$.

60 Prozent aller Punkte gehören zu diesem Cluster.

Der Cluster B ist repräsentiert durch den Mittelwert aller seiner Punkte (3,3) und die Kovarianzmatrix $\begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix}$.

40 Prozent aller Punkte gehören zu diesem Cluster.

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



Achtung: Die Skizze ist nur zu Veranschaulichungszwecken gedacht und nicht maßstabsgetreu!

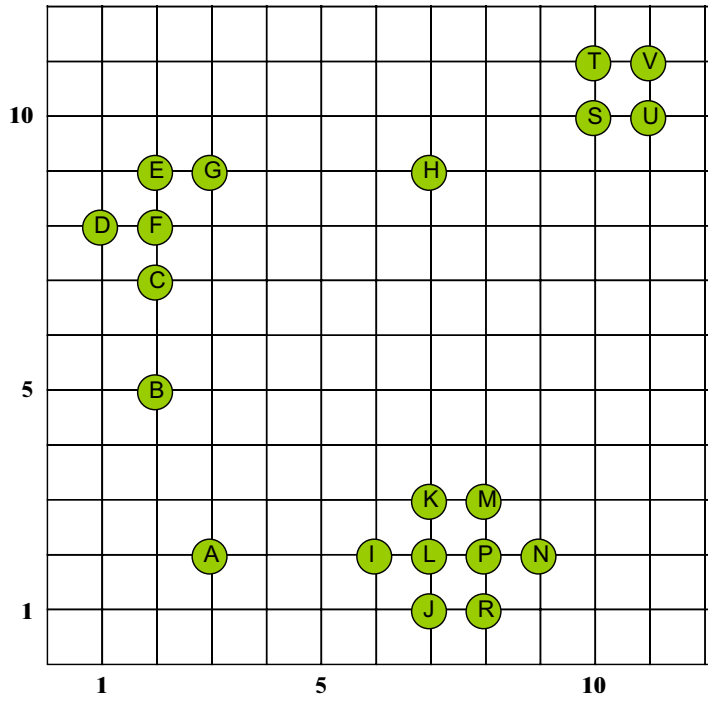
Aufgabe 8-2 DBSCAN: RANDPUNKTE

Das dichte-basiertes 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 eigentlich zu verschiedenen Clustern gehören und wie würde DBSCAN mit solchen Randpunkten umgehen, d.h. welchem Cluster würden diese Punkte zugeordnet werden? Nennen Sie eine vernünftigere Lösung zur Zuordnung dieser Randpunkte. Begründen Sie Ihre Entscheidung.

Aufgabe 8-3 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.