Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Peer Kröger
Johannes Assfalg, Karsten Borgwardt

Knowledge Discovery in Databases

WS 2005/06

Übungsblatt 3

Abgabe aller mit Hausaufgabe *markierten Aufgaben bis Donnerstag, 17.11.2005, 8:30 Uhr, vor der Vorlesung beim Dozenten oder im Übungsbriefkasten*

Aufgabe 3-1 DBSCAN: RANDPUNKTE

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.

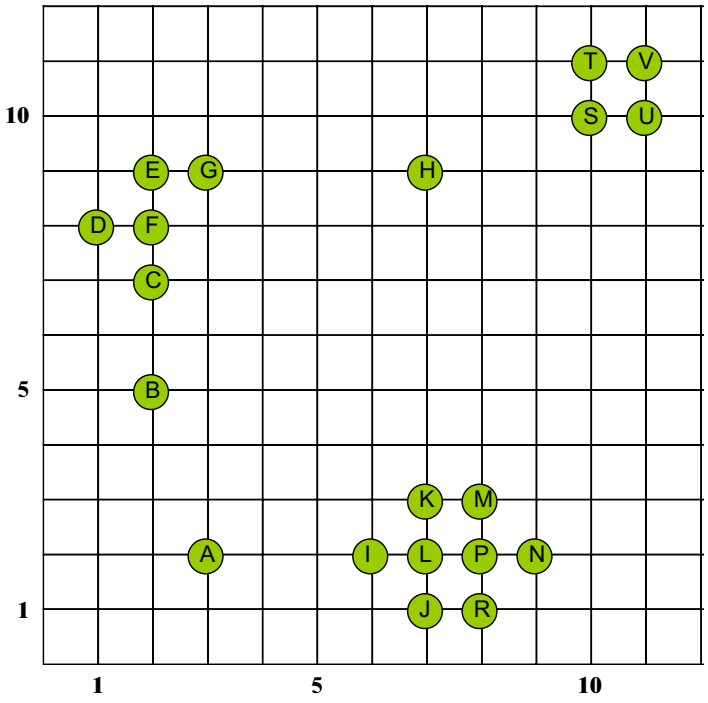
Kann es Randpunkte geben, die eigentlich zu verschiedenen Clustern gehören? Wie würde DBSCAN mit solchen Randpunkten umgehen, d.h. welchem Cluster würden diese Punkte zugeordnet werden? Beschreiben Sie eine vernünftigere Lösung zur Zuordnung dieser Randpunkte. Begründen Sie Ihre Entscheidung.

Aufgabe 3-2 DBSCAN: LEMMA

Beweisen Sie das folgende Lemma: p sei ein Punkt in Datenraum D und $|RQ(p, \varepsilon)| \geq MinPts$. Dann ist die Menge $O = \{o | o \in$ D und o ist dichte-erreichbar von p bzgl. ε und $MinPts\}$ ein Cluster bzgl. ε und $MinPts$.

Aufgabe 3-3 Single-Link
Hausaufgabe

Gegeben sei der folgende Datensatz:



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

$$\mathrm{L}_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

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.