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

Knowledge Discovery in Databases

WS 2005/06

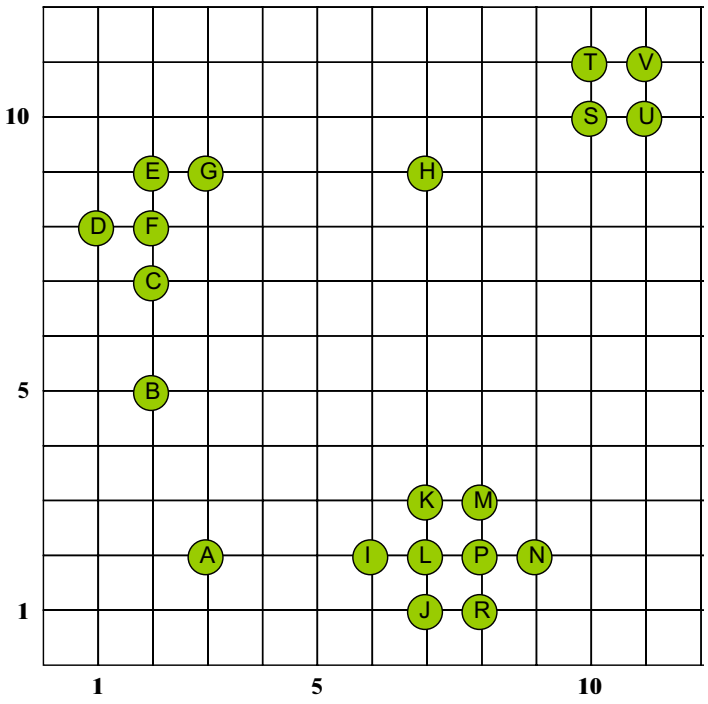
Übungsblatt 4

Abgabe aller mit Hausaufgabe *markierten Aufgaben bis Donnerstag, 24.11.2005, 8:30 Uhr, vor der Vorlesung beim Dozenten oder im Übungsbriefkasten*

Hinweis: In der Woche vom 28.11. bis 2.12. entfallen die Vorlesung und die Übungen. Dieses Übungsblatt wird in den Übungen ab dem 5.12. besprochen.

Aufgabe 4-1 OPTICS
Hausaufgabe

Gegeben sei der folgende 2-dimensionale Datensatz:



Verwenden Sie als Distanzfunktion zwischen den Punkten wieder die Manhattan-Distanz (L_1-Norm):

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

Erzeugen Sie mit OPTICS (Pseudocode am Ende des Übungsblattes) jeweils ein Erreichbarkeitsdiagramm für die folgenden Parameter:

(a) $\epsilon = 5$ und $MinPts = 2$

(b) $\epsilon = 5$ und $MinPts = 4$

(c) $\epsilon = 2$ und $MinPts = 4$

(d) $\epsilon = \infty$ und $MinPts = 4$

(e) Diskutieren Sie, welche Auswirkungen die Parameter $MintPts$ und ϵ haben.

Hinweis: Es reicht, wenn Sie den entstandenen OPTICS-Plot abgeben, Zwischenschritte müssen Sie nicht unbedingt abgeben.

Aufgabe 4-2 Outlier Detection

Gegeben der Datensatz und die Distanzfunktion aus Aufgabe 1. Berechnen Sie für die Punkte H und L den LOF-Wert für $MinPts = 3$.

Aufgabe 4-3 k-Modes

Diskutieren Sie die Vorteile bzw. Nachteile des Clusteringverfahrens k-Modes gegenüber k-Medoid. (Laufzeit, Generierung der Modes, ...)

Pseudocode OPTICS

```
seedlist = ∅    // implemented as a heap
for i = 0 to n-1 do
    if(seedlist = ∅ )  then seedlist = {(random_not_handled_point, ∞)}
    (x, x.reach) = get_and_remove_point_with_min_reach(seedlist)
    x.pos = i
    x.handled = TRUE
    neighbors = rangeQuery(x, ε)
    x.core = nnDist(x, neighbors, MinPts)
    if(x.core < ∞)
        for each y ∈ neighbors with not(y.handled)
            if( y ∉ seedlist)  seedlist = seedlist ∪ {(y, reach-dist(y,x))}
            else
                curr_reach = lookup(seedlist, y)
                update(y, min(curr_reach, reach-dist(y,x)))
        endfor
endfor
```