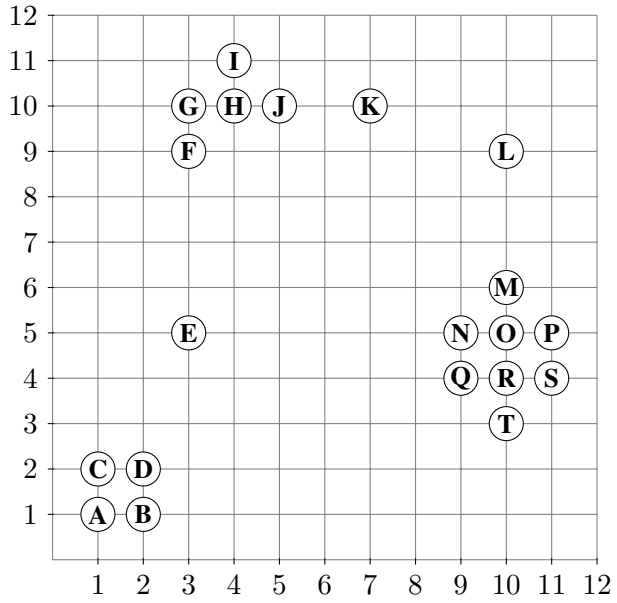
Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Anna Beer, Florian Richter

Knowledge Discovery in Databases
WS 2017/18

Übungsblatt 8: Clusteranalyse – Single-Link und OPTICS

Besprechung: 11. und 12.01.2018

Aufgabe 8-1 OPTICS



Als Distanzfunktion verwenden Sie die Manhattan-Distanz $L_1(a, b) := |a_1 - b_1| + |a_2 - b_2|$.

Konstruieren Sie ein Erreichbarkeitsdiagramm mit dem Algorithmus OPTICS (siehe beiliegenden Pseudocode) für folgende Parameter:

- $\varepsilon = 5$ and *minPts* $= 2$
- $\varepsilon = 5$ and *minPts* $= 4$
- $\varepsilon = 2$ and *minPts* $= 4$
- $\varepsilon = \infty$ and *minPts* $= 4$

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
```

Aufgabe 8-2 Zusammenhang DBSCAN/OPTICS und Single-Link

Was ist der Zusammenhang von DBSCAN bei *minPts* $= 2$ zu single-linkage Clustering?

Warum läuft DBSCAN in $\mathcal{O}(n^2)$ Zeitkomplexität (mit Index typischerweise sogar $\mathcal{O}(n \log n)$), während hierarchische Clusteranalyse auf Distanzmatrizen als $\mathcal{O}(n^3)$ angegeben wird, und SLINK in $\mathcal{O}(n^2)$ läuft?

Warum ist das kein Widerspruch?