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

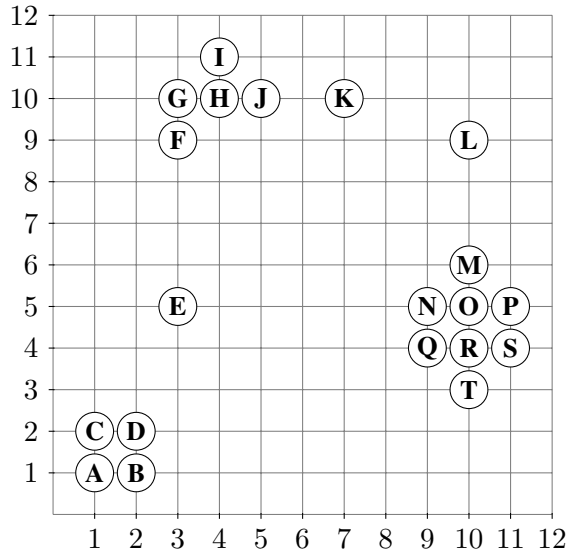
Knowledge Discovery in Databases
WS 2017/18

Übungsblatt 7: Clusteranalyse – DBSCAN und Hierarchical Clustering

Besprechung: 14. und 15.12.2017

Aufgabe 7-1 DBSCAN

Gegeben sei folgender Datensatz:



Als Distanzfunktion verwenden Sie die Manhattan-Distanz:

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

Führen Sie den Algorithmus DBSCAN auf dem Datensatz durch, und notieren Sie, welche Punkte Kernpunkte, Randpunkte und Noise sind.

Verwenden Sie folgende Parameterisierungen:

- Radius $\varepsilon = 1.1$ and *minPts* $= 2$
- Radius $\varepsilon = 1.1$ and *minPts* $= 3$
- Radius $\varepsilon = 1.1$ and *minPts* $= 4$
- Radius $\varepsilon = 2.1$ and *minPts* $= 4$
- Radius $\varepsilon = 4.1$ and *minPts* $= 5$
- Radius $\varepsilon = 4.1$ and *minPts* $= 4$

Sie können ihre Berechnungen / Implementierung mit ELKI verifizieren.

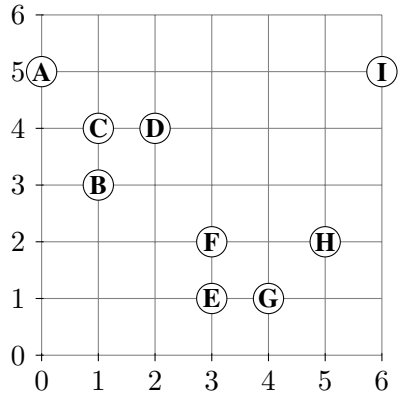
Aufgabe 7-2 Eigenschaften von DBSCAN

Diskutieren Sie folgende Fragen / Aussagen zu DBSCAN:

- Bei *minPts* $= 2$, was passiert mit Randpunkten?
- Das Ergebnis von DBSCAN ist determiniert auf Kern- und Noise-Punkten, aber nicht Randpunkten!
- Ein Cluster in DBSCAN kann weniger als *minPts* Punkte enthalten
- Hat der Datensatz n Objekte, so stellt DBSCAN stets genau n Nachbarschaftsanfragen.
- Auf gleichverteilten Daten wird DBSCAN in der Regel fast alles in einen Cluster clustern, oder alles als Noise klassifizieren. k-means hingegen wird in der Regel die Gleichverteilung in k etwa gleich große Partitionen aufteilen.

Aufgabe 7-3 Hierarchical Clustering

Gegeben sei der folgende Datensatz:



Als Distanzfunktion zwischen den Punkten verwenden Sie 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.