Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Arthur Zimek
Dr. Erich Schubert

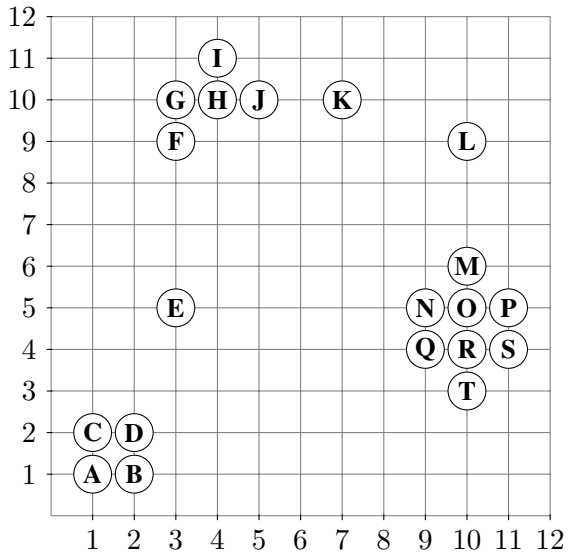
Knowledge Discovery in Databases

SS 2014

Übungsblatt 4: Clusteranalyse – DBSCAN

Aufgabe 4-1 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.

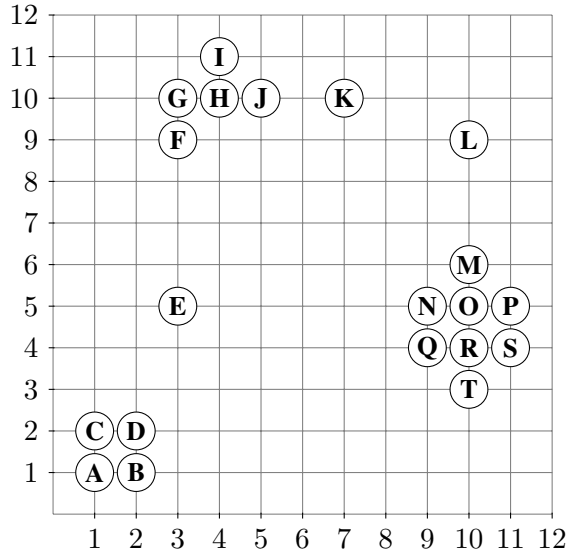
Aufgabe 4-2 Eigenschaften von DBSCAN

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 4-3 Shared Nearest Neighbors

Gegeben sei folgender Datensatz:

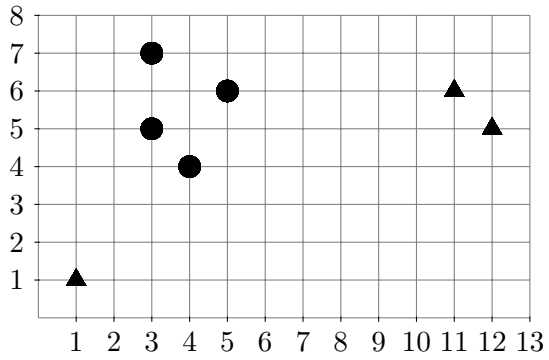


Berechnen Sie die paarweisen Shared-Nearest-Neighbor-Ähnlichkeiten SNN_5 der Objekte M, O, R und T.
Verwenden Sie die Manhattan-Distanz L_1, und die Nachbarschaftsgröße 5.
Der Anfragepunkt sei dabei Bestandteil seiner nächsten Nachbarn.

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

Aufgabe 4-4 Clustering durch Varianzminimierung

Gegeben sei folgender Datensatz mit 7 Punkten (Featurevektoren in $\mathbb{R}^2$).



Im folgenden sollen vollständige Partitionierungen des Datensatzes in $k = 3$ Cluster berechnet werden. Dabei wird jedes Objekt x demjenigen Cluster zugewiesen, bei dem die Summe der Quadrate der Abweichungen vom Clusterzentrum c minimal ist:

$$SSQ(x, c) = \sum_{i=1}^{d} |x_i - c_i|^2$$

Die initialen Cluster*zentren* seien durch die drei mit einem Dreieck markierten Objekte gegeben.

Führen Sie k-Means mit Lloyds Algorithmus durch. Welches Problem ergibt sich?