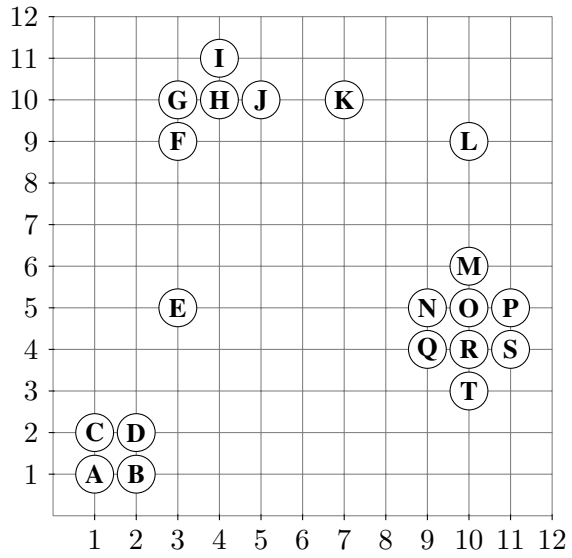
Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Arthur Zimek
Dr. Tobias Emrich

Knowledge Discovery in Databases
SS 2015

Übungsblatt 6: Clusteranalyse – Shared Nearest Neighbours und Single-Link

Aufgabe 6-1 Shared Nearest Neighbors

Gegeben sei folgender Datensatz:

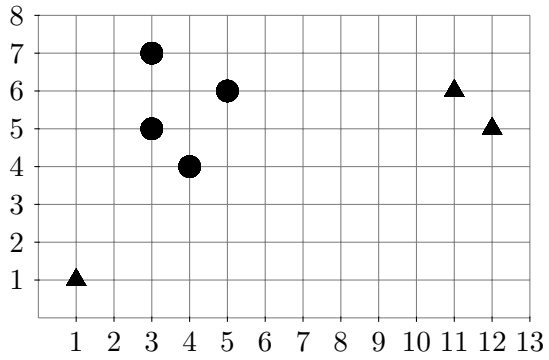


Berechnen Sie die paarweisen Shared-Nearest-Neighbor-Ähnlichkeiten SNN_5 der Objekte M, O, R und T.
Verwenden Sie die Manhattan-Distanz L_1, und die Nachbarschaftsgröße 5.
Der Anfragepunkt sei dabei Bestandteil seiner nächsten Nachbarn.

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

Aufgabe 6-2 Clustering durch Varianzminimierung

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

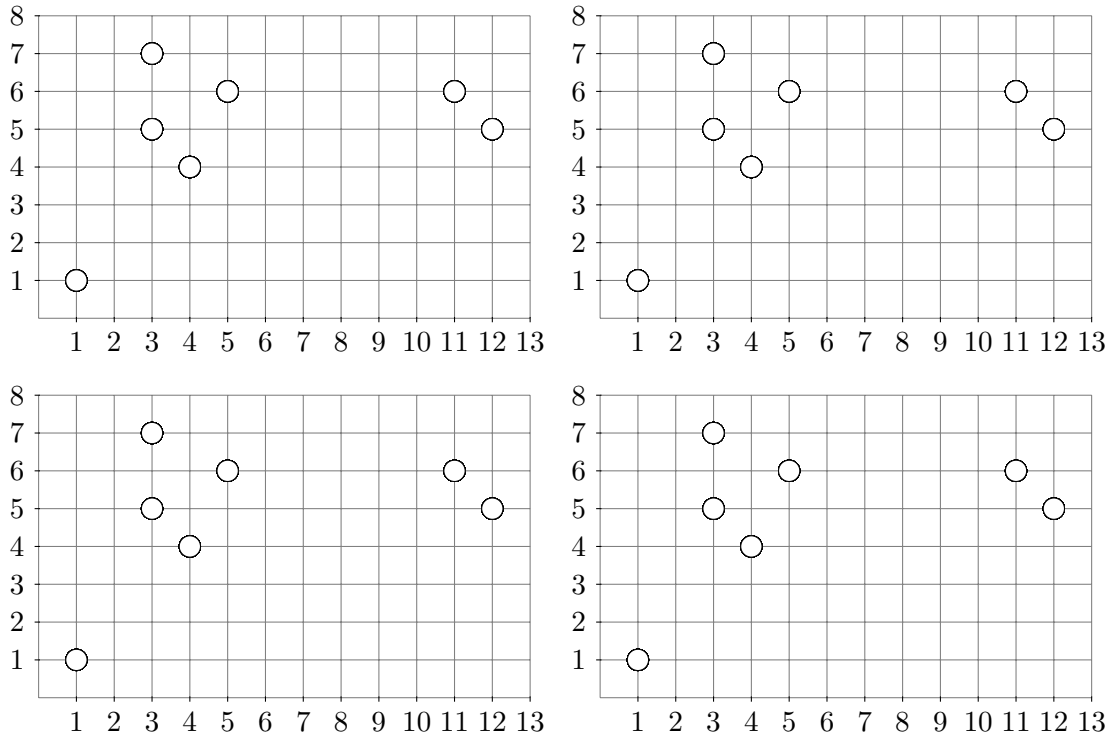


Im folgenden sollen vollständige Partitionierungen des Datensatzes in $k = 3$ Cluster berechnet werden. Dabei wird jedes Objekt x demjenigen Cluster zugewiesen, bei dem die Summe der Quadrate der Abweichungen vom Clusterzentrum c minimal ist:

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

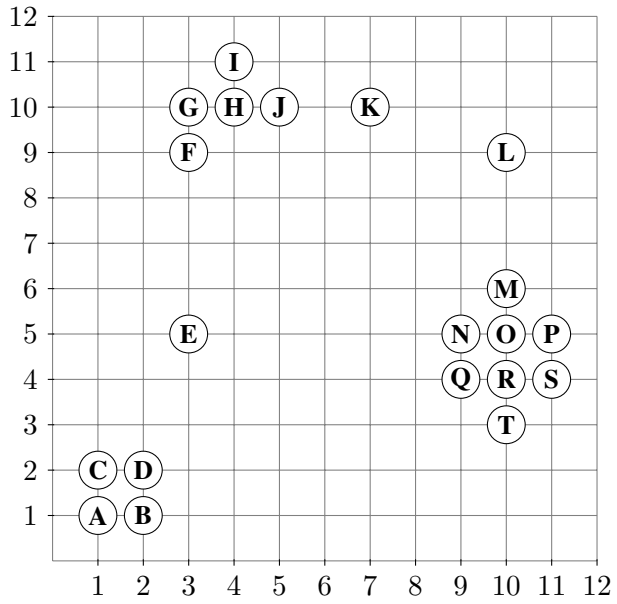
Die initialen Cluster*zentren* seien durch die drei mit einem Dreieck markierten Objekte gegeben.

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



Aufgabe 6-3 Single-Link 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.