Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Matthias Schubert

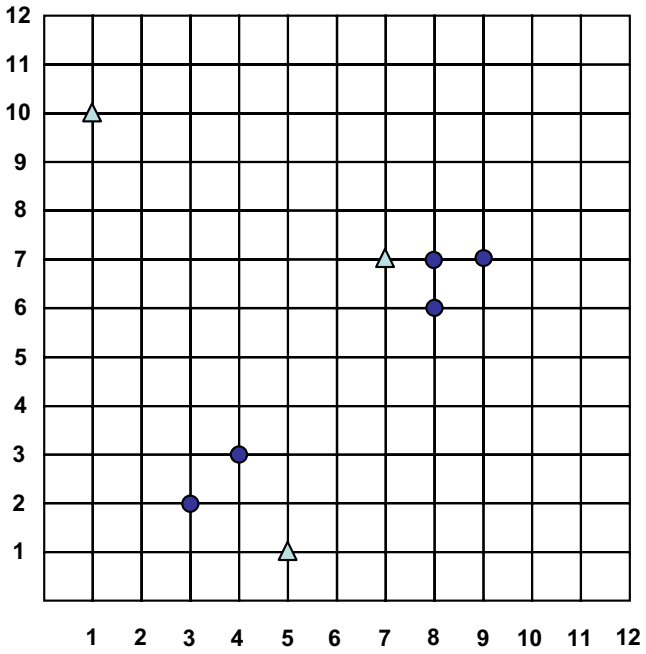
Knowledge Discovery in Databases

WS 2008/09

Übungsblatt 7

Aufgabe 7-1 Clusterung mit Varianzminimierung

Gegeben sei folgender Datensatz mit 8 Punkten (2-dimensionalen Featurevektoren). Stören Sie sich zunächst nicht, dass einige Datenpunkte als Kreise und andere als Dreiecke dargestellt sind, sondern lesen Sie schnell weiter! ;-)



Im folgenden sollen vollständige Partitionierungen des Datensatzes in $k = 2$ Cluster berechnet werden. Als Distanzfunktion zwischen den Punkten soll dabei die Manhattan-Distanz (L_1-Norm) verwendet werden, die für zwei Punkte x, y wie folgt definiert ist:

$$L_1(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

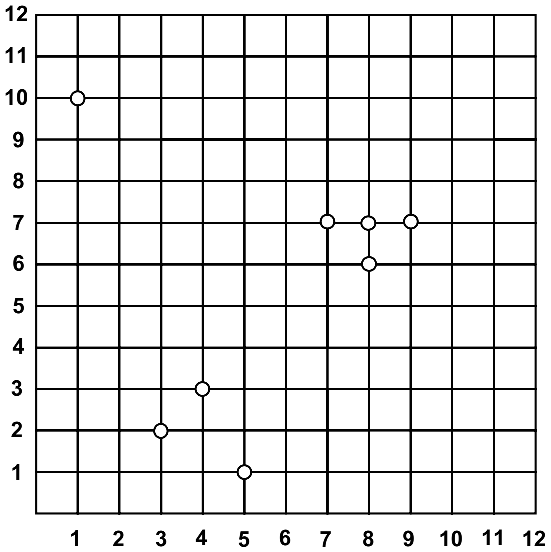
(a) Erzeugen Sie eine Partitionierung in $k = 2$ Cluster mit dem einfachen Verfahren "Clustering durch Varianz Minimierung" (Skript: Folie 185). Die initiale Partitionierung der Daten ist durch die Dreiecke und Punkte gegeben (die Dreiecke bilden einen initialen Cluster, genauso die Punkte). Beschreiben Sie jede Aktion des Algorithmus. Zeichnen Sie nach jedem Schritt die Zentroiden ein und markieren Sie die Punkte anhand ihrer Clusterzugehörigkeit. Denken Sie daran, bei der Zuordnung zu den Zentroiden die L_1-Norm zu verwenden.
Tipp: Hierzu können Sie die Vorlage auf der letzten Seite benutzen, die Sie am besten mehrmals kopieren.

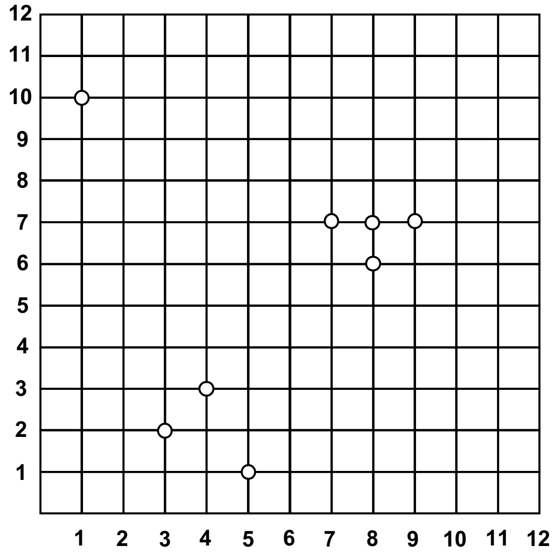
(b) Erzeugen Sie eine Partitionierung in $k = 2$ Cluster mit dem k-means Verfahren (Skript: Folie 187). Die initiale Partitionierung der Daten ist auch hier durch die Dreiecke und Punkte gegeben (die Dreiecke bilden einen initialen Cluster, genauso die Punkte). Beschreiben Sie jede Aktion des Algorithmus. Zeichnen Sie nach jedem Schritt die Zentroiden ein und markieren Sie die Punkte anhand ihrer Clusterzugehörigkeit. Denken Sie daran, bei der Zuordnung zu den Zentroiden die L_1-Norm zu verwenden. Die Reihenfolge der Zuodnung bleibt Ihnen überlassen.
Tipp: Auch hierzu können Sie die Vorlage auf der letzten Seite benutzen.

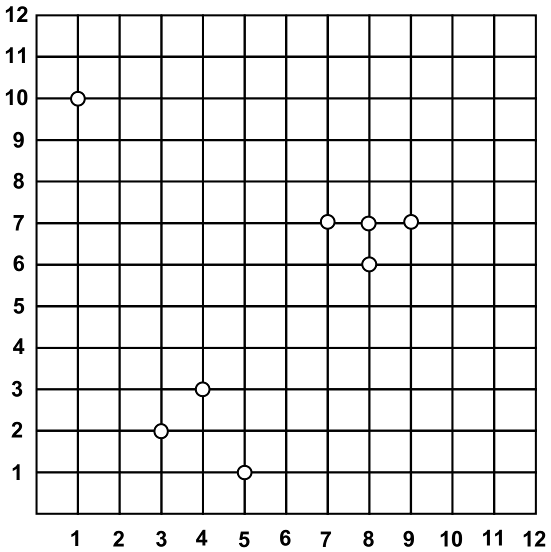
(c) Begründen Sie kurz, warum k-means reihenfolgeabhängig ist.

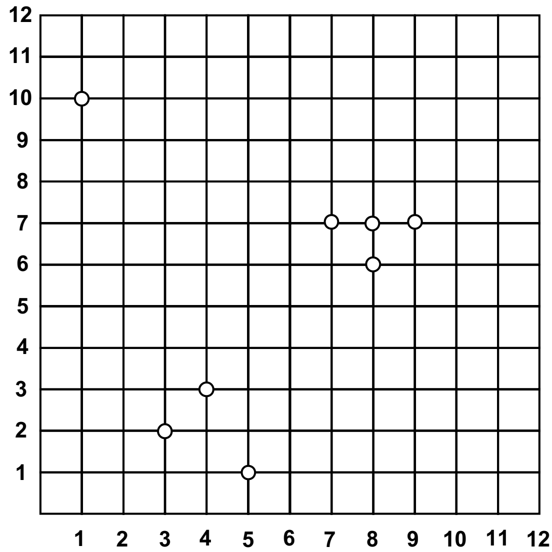
Aufgabe 7-2 PAM

Zeigen Sie, dass der Algorithmus PAM konvergiert.


12
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12


12
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12


12
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12


12
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12