Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Peer Kröger
Johannes Aßfalg, Karsten Borgwardt

Knowledge Discovery in Databases

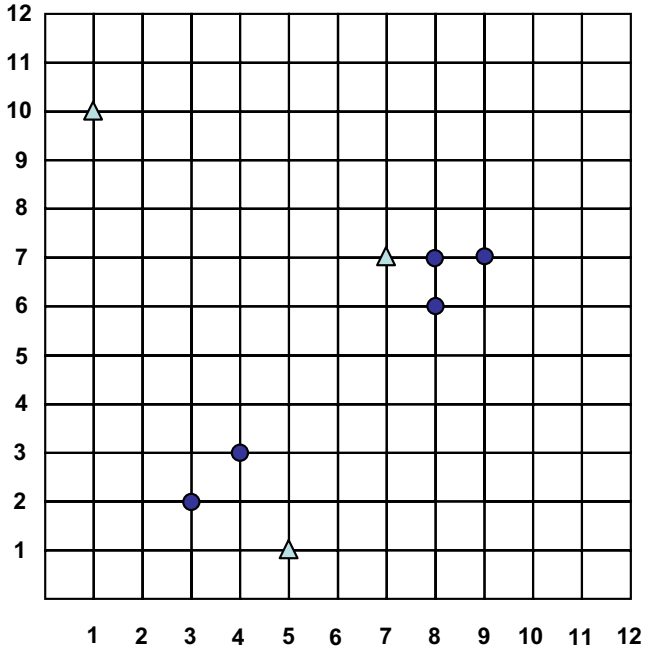
WS 2005/06

Übungsblatt 2

Abgabe aller mit Hausaufgabe *markierten Aufgaben bis Donnerstag, 10.11.2005, 8:30 Uhr, vor der Vorlesung beim Dozenten oder im Übungsbriefkasten*

Aufgabe 2-1 k-Zentroid Verfahren
Hausaufgabe

Gegeben sei folgender Datensatz mit 8 Punkten (2-dimensionalen Featurevektoren):



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 Varianzminimierung". 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. 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 2-2 EM-Algorithmus
Hausaufgabe

Gegeben sei eine Datenmenge D mit 100 Punkten, die zwei Gausscluster A und B und den Punkt p enthält.

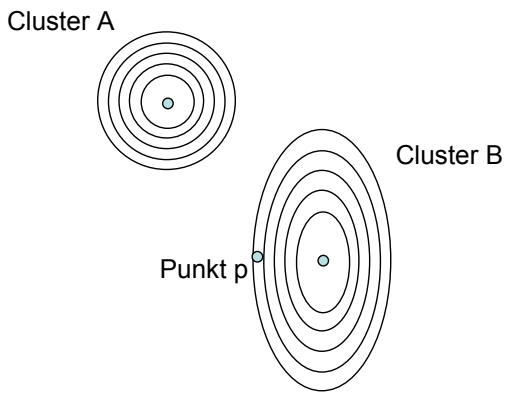
Der Cluster A ist repräsentiert durch den Mittelwert aller seiner Punkte (1,1) und die Kovarianzmatrix $\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$.

60 Prozent aller Punkte gehören zu diesem Cluster.

Der Cluster B ist repräsentiert durch den Mittelwert aller seiner Punkte (3,3) und die Kovarianzmatrix $\begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix}$.

40 Prozent aller Punkte gehören zu diesem Cluster.

Der Punkt p ist durch die Koordinaten (2.5,2.5) gegeben. Geben Sie die beiden Wahrscheinlichkeiten an, mit der p zum Cluster A bzw. B gehört.



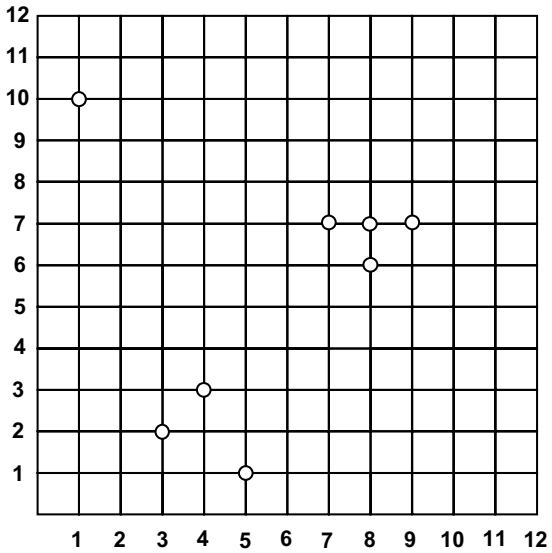
Achtung: Die Skizze ist nur zu Veranschaulichungszwecken gedacht und nicht maßstabsgetreu!

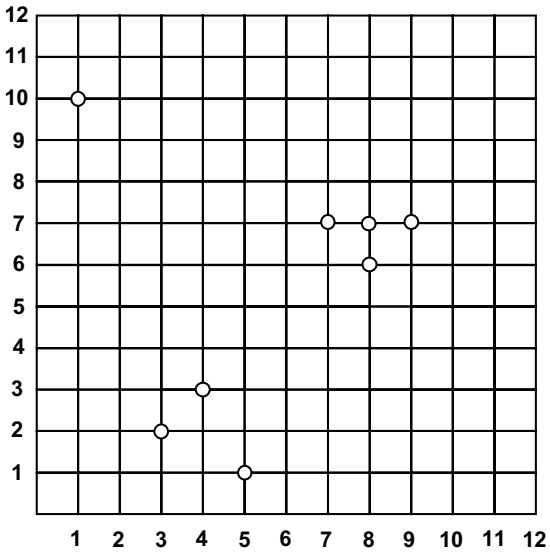
Aufgabe 2-3 PAM

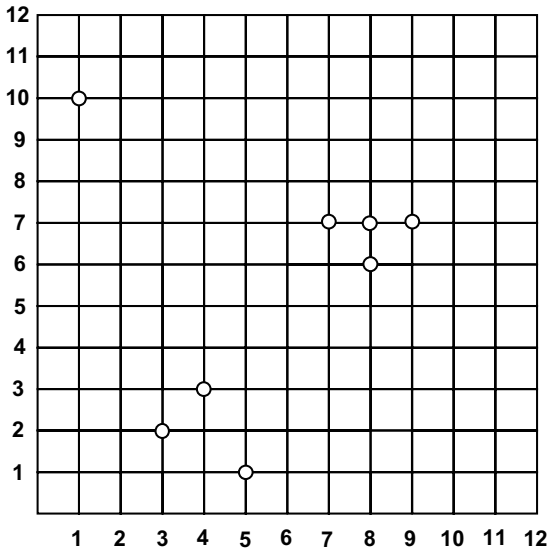
Zeigen Sie, dass der Algorithmus PAM konvergiert.

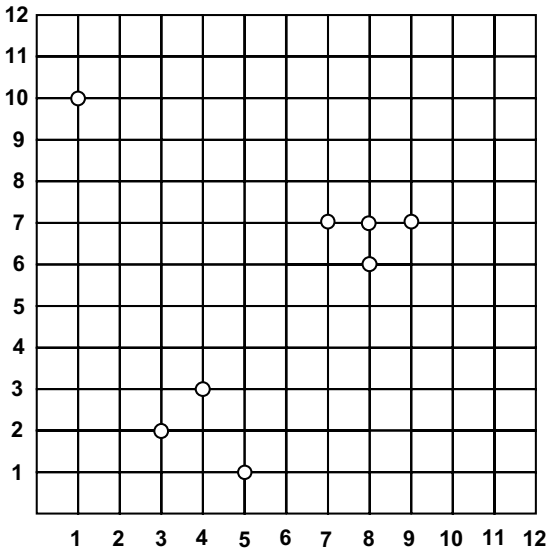
Aufgabe 2-4 Invarianzen räumlicher Ähnlichkeitsmodelle

Diskutieren Sie, welche Invarianzen gegenüber Translation, Rotation, Skalierung und Reflektion das Graphmodell für Protein 3D-Strukturen [Borgwardt et al. 2005] besitzt bzw. nicht besitzt.


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