Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dipl.-Inform. Klaus Friedrichs,
Dipl.-Stat. Julia Schiffner,
Dr. Issam Ben Khediri

Dortmund, 26.06.12

Abgabe: bis Mi, 04.07., 12.00 Uhr an
friedrichs@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2012

Blatt 12

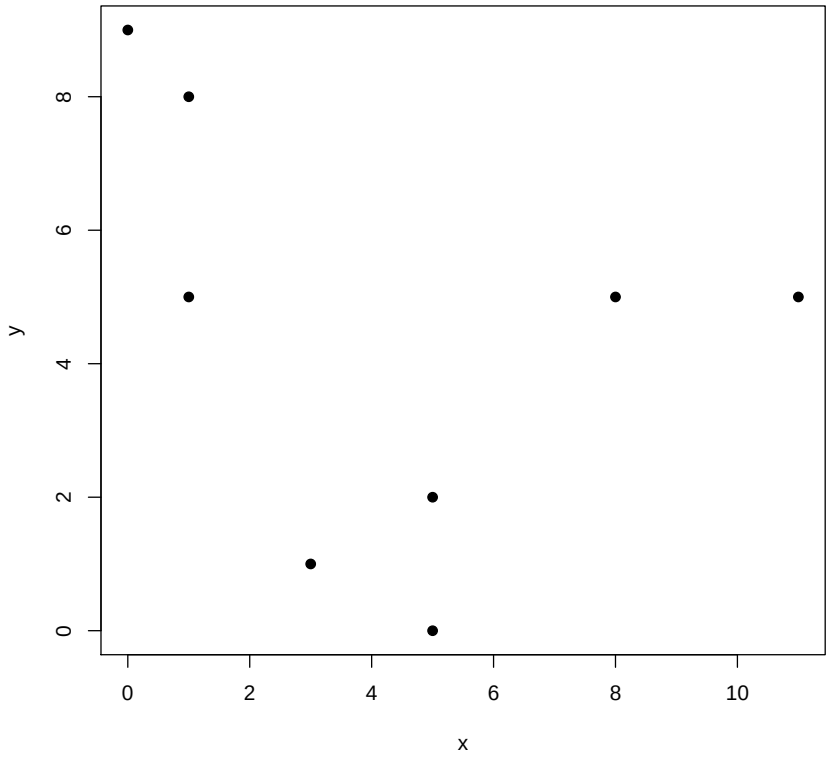
Aufgabe 12.1 (5 Punkte)
Gegeben seien folgende Datenpunkte im euklidischen Raum (siehe auch das Diagramm auf Seite 2):

Punkt	x	y
A	0	9
B	1	8
C	1	5
D	3	1
E	5	2
F	5	0
G	8	5
H	11	5

a) Führen Sie den K-Means-Algorithmus mit k = 3 per Hand aus. Normalerweise werden die Startpunkte für die Mittelpunkte der Cluster (auch Zentroiden genannt) zufällig gewählt. Hier sollen Sie jedoch die Punkte A, B und C als Startzentroiden benutzen. Falls im laufenden Algorithmus ein konkretes Beispiel äquidistant zu zwei Clusterzentroiden ist, so wählen Sie denjenigen, der näher am Nullpunkt liegt.

b) Benutzen Sie nun die Punkte E, F und G als Startzentroiden und führen Sie den Algorithmus ein weiteres Mal durch.

c) Berechnen Sie für Ihre Resultate aus a) und b) jeweils die in der Vorlesung auf Folie 15 vorgestellten Gütekriterien *Innerer Abstand* $W(C)$ und *Zwischenunähnlichkeit* $B(C)$. Interpretieren Sie kurz die Ergebnisse!



Aufgabe 12.2 (5 Punkte)
Bei Cluster-Verfahren, deren Clusteranzahl k vom Benutzer vorgegeben werden muss, ist die automatische Bestimmung dieses k kritisch.

a) Zur Auswahl welcher Cluster tendiert allgemein eine Optimierung, die auf der Formel $W(C)$ beruht? Was ist somit ein 'optimales' Clustering beruhend auf dieser Formel?

b) Benutzen Sie nun RapidMiner, um die Iris-Daten mit verschiedenen k-Werten zu clustern. Benutzen Sie den Operator *Loop Parameter*, um Clusterings für alle ganzzahligen k zwischen 2 und 150 zu erstellen. Benutzen Sie zudem *k-Means*, *Data to Similarity*, *Log* und *Cluster Density Performance* innerhalb der Parameter-Schleife, um die Cluster zu bewerten. *Cluster Density Performance* liefert vergleichbare Ergebnisse wie $W(C)$. Lassen Sie sich die Performanz-Werte für die verschiedenen Parameter-Werte k anzeigen und geben Sie diesen Plot zusammen mit der Experiment-Datei ab!

c) Analysieren Sie den erzeugten Plot und suchen Sie den 'Knick in der Kurve'. Gibt es diesen Knick? Was sind u.U. andere Merkmale, die hier ein gutes Clustering auszeichnen. Ihr Wissen über die Beschaffenheit des Iris-Datensatzes ist hier hilfreich.