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

Knowledge Discovery in Databases

SS 2011

Übungsblatt 6: Kernel, Partitionierendes Clustering

Aufgabe 6-1 *Kernel-Funktionen*

Wie in der Vorlesung erklärt, zeichnet sich eine Kernel-Funktion ("Kernel") durch positive (Semi-)Definitheit aus. Eine Matrix A ist positiv definit, falls ihre Eigenwerte nichtnegativ sind, oder alternativ formuliert, falls für all $x \in \mathbb{R}^d$ gilt: $x^\top \cdot A \cdot x \geq$

Zeigen Sie, dass folgende Funktionen Kernels sind, falls x und $\hat{x}$ Vektoren im $\mathbb{R}^d$ sind:

(a) $k_1(x, \hat{x}) = 1$

(b) $k_2(x, \hat{x}) = 3 \cdot x^\top \cdot \hat{x}$

(c) $k_3(x, \hat{x}) = 3 \cdot x^\top \cdot \hat{x} + 5$

Aufgabe 6-2 *Lineare Regression*

Das Gehalt einer Person hängt von den Jahren ab, in denen die Person ihren Beruf ausgeübt hat. Um diesen Zusammenhang genauer zu untersuchen, kann man ein lineares Regressionsmodel lernen. Als Trainingsmenge stehen uns die Jahre an Berufserfahrung und die Gehälter folgender Personen zur Verfügung.

Erfahrung in Jahren	Gehalt in (1000$)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

(a) Berechnen Sie eine Regressionsgerade, die dazu dienen soll, das voraussichtliche Gehalt auf Basis der Berufserfahrung abzuschätzen. Bestimmen Sie hierzu die Gerade, die den quadratischen Fehler minimiert.

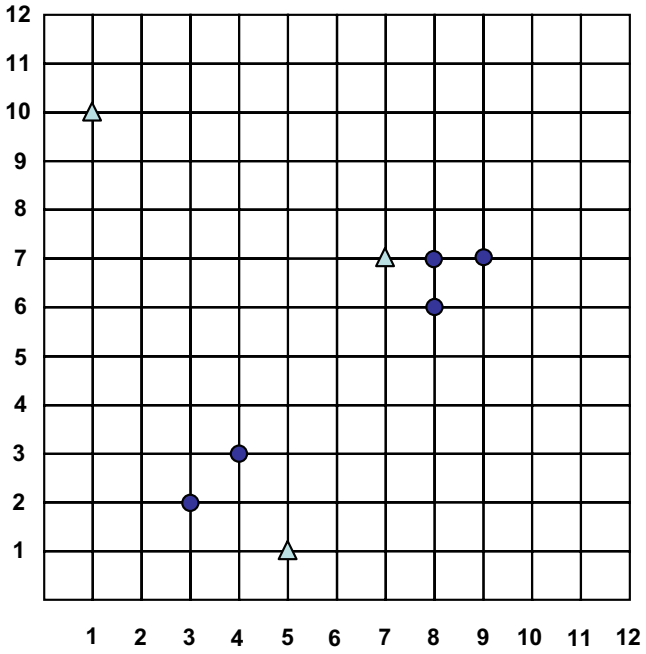
(b) Bestimmen Sie den quadratischen Fehler der berechneten Gerade, um abzuschätzen, wie gut die Regressiongerade den Zusammenhang erklärt.

(c) Berechnen Sie mit Hilfe Ihrer Regressionsgerade das vorraussichtliche Gehalt für Personen mit den folgenden Jahren an Berufserfahrung:

Person A: 20
Person B: 8
Person C: 11

Aufgabe 6-3 *Clusterung mit Varianzminimierung*

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



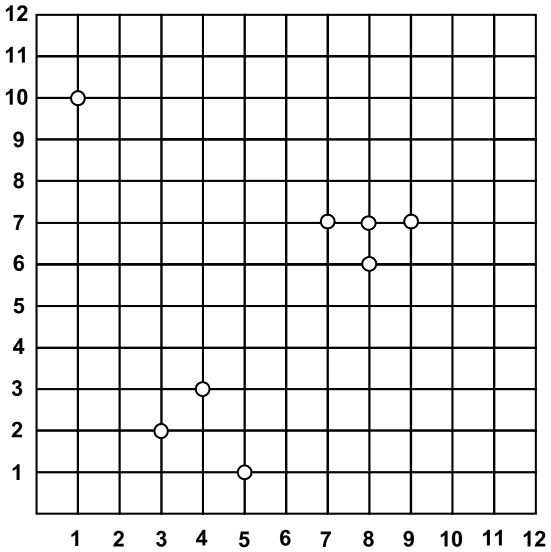
Im folgenden sollen vollständige Partitionierungen des Datensatzes in $k = 2$ Cluster berechnet werden. Als Distanzfunktion zwischen den Punkten soll dabei die Manhattan-Distanz (L_1-Norm) verwendet werden, die für zwei Punkte x, y wie folgt definiert ist:

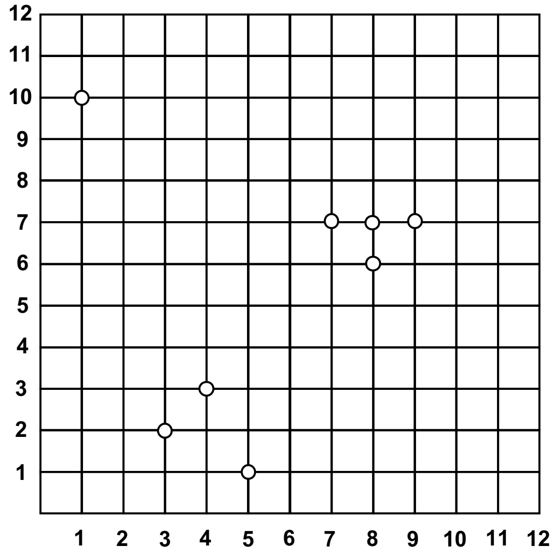
$$L_1(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

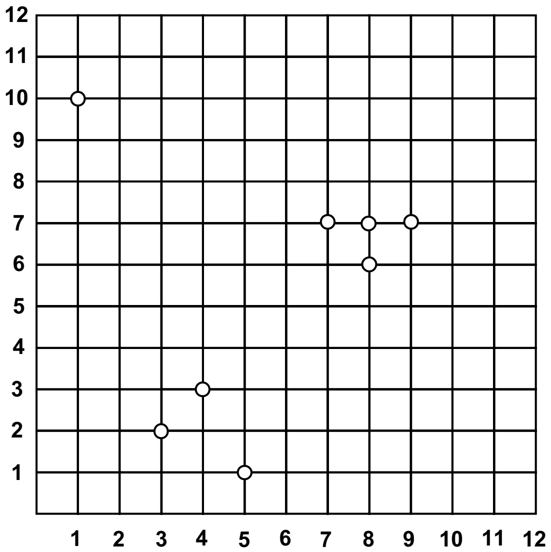
(a) Erzeugen Sie eine Partitionierung in $k = 2$ Cluster mit dem einfachen Verfahren "Clustering durch Varianz Minimierung". Die initiale Partitionierung der Daten ist durch die Dreiecke und Punkte gegeben (die Dreiecke bilden einen initialen Cluster, genauso die Punkte). Beschreiben Sie jede Aktion 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.

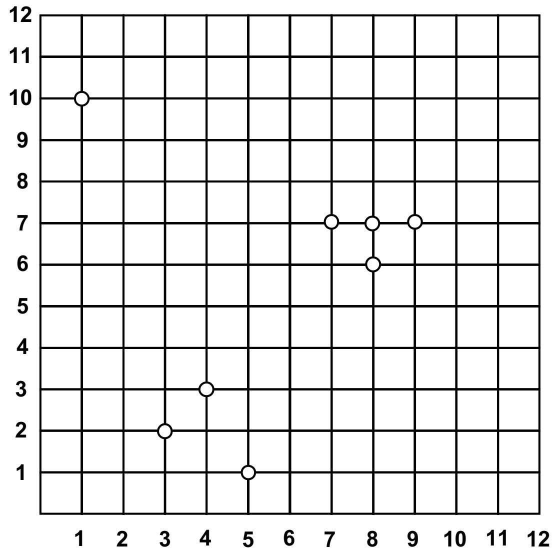
(b) Erzeugen Sie eine Partitionierung in $k = 2$ Cluster mit dem k-means Verfahren (Skript: Folie 187). Die initiale Partitionierung der Daten ist auch hier durch die Dreiecke und Punkte gegeben (die Dreiecke bilden einen initialen Cluster, genauso die Punkte). Beschreiben Sie jede Aktion des Algorithmus. 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.


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