Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Arthur Zimek
Erich Schubert

Knowledge Discovery in Databases

SS 2013

Übungsblatt 2: Clusteranalyse

Aufgabe 2-1 Induzierte Metrik

Gegeben sei eine Pseudo-Metrik d auf der Menge A: $d : A \times A \to \mathbb{R}_0^+$.

Sei $\sim$ die Äquivalenzrelation mit $x \sim y \Leftrightarrow d(x, y) = 0$.

Sei $A^\sim$ die zugehörige Menge der Äquivalenzklassen von A bzgl. $\sim$.

- Welche Eigenschaften hat die Distanzfunktion $d^\sim : A^\sim \times A^\sim \to \mathbb{R}_0^+$ mit $d^\sim(x^\sim, y^\sim) := d(x, y)$?
- Gegeben eine Datenbank wie unten skizziert, welche Eigenschaften hat die folgende Distanzfunktion:

$$\text{euclid}_{xy}((r_1, x_1, y_1), (r_2, x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Datensatz ID	x	y
1	0	1
2	1	1
3	0	1

Datensatz ID	x	y
4	1	1
5	2	2
6	3	3

Erklären Sie, welche Datensätze von der Distanzfunktion als äquivalent behandelt werden, und diskutieren Sie, ob es in einem Datenbank- und Data-Mining-Zusammenhang sinnvoll ist, Pseudo-Metriken zu verwenden.

Aufgabe 2-2 Multivariate Dichte und Mahalanobis-Distanz

Die Dichte der multivariaten Normalverteilung (mit Kovarianzmatrix Σ und Mittelwert μ) wird mit der folgenden Formel berechnet:

$$prob(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

Finden und diskutieren Sie den Zusammenhang dieser Formel zu der Formel der Mahalanobis-Distanz mit Matrix Σ von p zu μ.

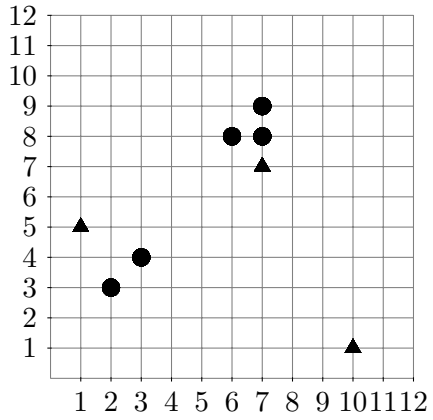
$$d_{\textit{Mahalanobis}}(x, y, \Sigma) := \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

Hinweis: betrachten Sie auch die multivariate Standardnormalverteilung, mit der Dichtefunktion

$$prob(p) := \frac{1}{\sqrt{(2\pi)^d}} \cdot e^{-\frac{1}{2}\langle p,p\rangle} = \frac{1}{\sqrt{(2\pi)^d}} \cdot e^{-\frac{1}{2}\|p\|^2}$$

Aufgabe 2-3 Clustering durch Varianzminimierung

Gegeben sei folgender Datensatz mit 8 Punkten (Featurevektoren in $\mathbb{R}^2$).



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" (nach Lloyd, siehe Skript). 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.

(b) Erzeugen Sie eine Partitionierung in $k = 2$ Cluster mit dem k-means Verfahren (nach MacQueen, siehe Skript). 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.

(d) Optional: probieren Sie auch k-medoids.