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

Knowledge Discovery in Databases

SS 2014

Übungsblatt 2: Skalen, Distanzen und Metriken

Aufgabe 2-1 Distanzmaße

Distanzmaße können wir nach ihren Eigenschaften in folgende Kategorien einteilen:

$d : S \times S \to \mathbb{R}_0^+$ $x, y, z \in S$:	reflexiv reflexive $x = y \Rightarrow d(x, y) = 0$	symmetrisch symmetric $d(x, y) = d(y, x)$	strikt strict $d(x, y) = 0 \Rightarrow x = y$	Dreiecksungleichung Triangle inequality $d(x, z) \leq d(x, y) + d(y, z)$
Unähnlichkeitsfunktion Dissimilarity function	$\times$			
(Symmetrische) Prämetrik (Symmetric) Pre-metric	$\times$	$\times$		
Semimetrik, Ultrametrik Semi-metric, Ultra-metric	$\times$	$\times$	$\times$	
Pseudometrik Pseudo-metric	$\times$	$\times$		$\times$
Metrik Metric	$\times$	$\times$	$\times$	$\times$

D.h., wenn ein Distanzmaß $d : S \times S \to \mathbb{R}_0^+$ für alle $x, y, z \in S$: reflexiv, symmetrisch und strikt ist sowie die Dreiecks-Ungleichung erfüllt, ist es eine Metrik. Wie Sie sehen, muß eine Distanzfunktion nicht *strikt* reflexiv sein. Machen Sie sich den Unterschied zwischen Reflexivität und Striktheit klar!

Anmerkung: Die Namen Distanzfunktion, Semi- und Pseudo-Metrik werden in der Literatur nicht einheitlich definiert.

Entscheiden Sie für die folgenden Funktionen $d(\mathbb{R}^n, \mathbb{R}^n)$ jeweils, ob es sich um ein Distanzmaß handelt, und wenn ja, in welche Kategorie es fällt.

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

(b) $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$

(c) $d(x, y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2}$

(d) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{falls} \quad x_i = y_i \\ 0 & \text{falls} \quad x_i \neq y_i \end{cases}$

(e) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{falls} \quad x_i \neq y_i \\ 0 & \text{falls} \quad x_i = y_i \end{cases}$

Aufgabe 2-2 Skalen-Niveaus von Merkmalen

Entscheiden Sie für jedes Merkmal des folgenden Datensatzes, ob es sich um ordinale, nominale oder metrische Merkmale handelt.

Obs.	Geschlecht	Grösse (cm)	Gewicht (kg)	Haarfarbe	Blutgruppe	Brille	Rauchen	Wohnlage
67	Frau	175	60	dunkelbl./braun	A	nein	gelegentlich	ruhig
68	Frau	176	52	hellblond	AB	ja	gelegentlich	ruhig
69	Frau	176	63	schwarz	A	ja	selten	sehr ruhig
70	Frau	179	65	dunkelbl./braun	0	ja	nie	ruhig
71	Frau	180	62	dunkelbl./braun	B	ja	nie	ruhig
72	Frau	180	70	dunkelbl./braun	A	ja	nie	ruhig
73	Frau	185	72	dunkelbl./braun	B	nein	nie	sehr ruhig
74	Frau	195	62	rot	0	ja	sehr viel	sehr ruhig
75	Frau	203	62	rot	AB	ja	sehr viel	sehr lärmig
76	Mann	165	53	dunkelbl./braun	A	nein	selten	ruhig
77	Mann	169	63	dunkelbl./braun	B	ja	selten	ruhig
78	Mann	169	72	dunkelbl./braun	A	nein	nie	ruhig
79	Mann	170	61	dunkelbl./braun	A	nein	nie	sehr ruhig
80	Mann	171	71	dunkelbl./braun	A	nein	viel	lärmig
81	Mann	173	61	schwarz	A	ja	nie	sehr ruhig
82	Mann	173	63	rot	A	nein	selten	lärmig
83	Mann	173	67	dunkelbl./braun	B	ja	nie	ruhig
84	Mann	175	68	dunkelbl./braun	.	nein	nie	ruhig
85	Mann	175	71	dunkelbl./braun	AB	nein	viel	ruhig
86	Mann	176	60	dunkelbl./braun	A	nein	selten	ruhig
87	Mann	177	64	dunkelbl./braun	AB	nein	nie	sehr lärmig

Aufgabe 2-3 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.