Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Matthias Schubert

Knowledge Discovery in Databases

WS 2009/10

Übungsblatt 1

Aufgabe 1-1 Metrische Distanzfunktionen

Im Data Mining spielen insbesondere metrische Distanzfunktionen eine große Rolle. Eine Distanzfunktion $dist : \mathbb{R}^d \rightarrow \mathbb{R}$ für d-dimensionale Feature-Vektoren ist eine Metrik, wenn folgende Bedingungen für alle $o_1, o_2, o_3 \in \mathbb{R}^d$ erfüllt sind:

(1) $o_1 \neq o_2 : dist(o_1, o_2) > 0$
$\forall o \in \mathbb{R}^d : dist(o, o) = 0$

(2) $dist(o_1, o_2) = dist(o_2, o_1)$,

(3) $dist(o_1, o_3) \leq dist(o_1, o_2) + dist(o_2, o_3)$.

Seien $x = (x_1, \ldots, x_d)$ und $y = (y_1, \ldots, y_d)$, mit $d \in \mathbb{N}$ und $x, y \in \mathbb{R}^d$. Zeigen oder widerlegen Sie, dass die folgenden Distanzfunktionen Metriken sind:

(a) $dist_1(x, y) = \sum_{i=1}^{d}(x_i - y_i)$

(b) $dist_2(x, y) = \sum_{i=1}^{d} \begin{cases} 1 & \text{falls } x_i = y_i \\ 0 & \text{sonst} \end{cases}$

(c) $dist_3(x, y) = \sum_{i=1}^{d} \begin{cases} 1 & \text{falls } x_i \neq y_i \\ 0 & \text{sonst} \end{cases}$

(d) $dist_4(x, y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$

Aufgabe 1-2 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 1-3 Data Mining Aufgaben

Welche Aufgaben für das Data Mining (Clustering, Outlier Detection, Klassifikation, etc.) verbergen sich hinter den folgenden Anwendungen? Ist die Aufgabe überwacht (supervised) oder nicht überwacht (unsupervised)?

(a) **Texterkennung/OCR**:
Beim Passieren der Brennerautobahn existiert seid einigen Jahren die Möglichkeit per E-Maut zu zahlen. Dabei wird bei Zahlungseingang das Nummernschild des Autos registriert. Beim passieren der Mautstation fahren das Auto dann durch einen gesonderte Schranke die nur aufgeht, wenn das Nummernschild des Fahrzeug als registriert erkannt wurde. Die Erkennung erfolgt dabei voll automatisch per digitaler Kamera.

(b) **Computer Added Diagnosis:**
Patienten, die an Blutkrebs leiden, können in zwei Kategorien (ALL und AML) eingeteilt werden. Da sich die Therapien dieser beiden Arten teilweise sehr stark unterscheiden und sogar manchmal die Therapie für AML sehr schädlich für ALL-Patienten sein kann (und umgekehrt), versucht man neue Patienten anhand von speziellen Daten (sog. Gen-Expressionsdaten) zu unterscheiden. Dazu werden die Daten der neuen Patienten mit den Daten der Patienten, deren Blutkrebstyp bereits bekannt ist, verglichen.

(c) **Cheat Detection**
Der Betreiber eines Multiplayer-Online-Spiels will sein System gegen verschiedene Verstöße der Benutzerrichtlinien abdecken. Dazu gehört die Verwendung von Bot-Programmen, das Manipulieren von Zeitstempeln im Kommunikation Protokoll und die Vorhersage verwendeter Zufallszahlen. Zur Erkennung von verdächtigem Verhalten wird Data Mining auf den erhältlichen Benutzerdaten verwendet.

(d) **Mensch und Maschine**
Moderne WWW-Suchmaschinen beantworten Benutzeranfragen, die aus nur einem oder wenigen Suchtermen bestehen. In der Regel liefert eine Anfrage dabei eine sehr große Ergebnismenge, die mit Hilfe eines Ranking Algorithmus nach Relevanz sortiert wird. Durch diese Sortierung kann der User dann

selber entscheiden, wieivele Links er besuchen will. Die Problematik hierbei ist zum einem den Inhalt einer Ergebnisseite richtig zu erkennen. Zum anderen besteht die Notwendigkeit, dass wirklich hilfreiche Seiten höher gerankt werden als weniger hilfreiche Seiten, auch wenn beide inhaltlich zum Suchbegriff passen. Wortmehrdeutigkeiten stellen dabei ebenfalls ein großes Problem dar. Zum Beispiel, kann sich die Suche nach dem Begriff ”Golfäuf das Auto, den Sport oder den geographischen Begriff beziehen. Data Mining Techniken werden hier eingesetzt um das Ranking zu optimieren und mögliche Ergebnisse Mengen nach dem jeweiligen Begriffskontext zu gruppieren.

(e) **Recommendation Systems**
Ein Online-Kaufhaus möchte für registrierte Kunden Artikel bestimmen, die dem Kunden beim Einloggen unaufgefordert angeboten werden. Dabei kann man auf die bereits gekauften Artikel des Kunden zurückgreifen, um so die Interessengebiete des Kunden besser vorhersagen zu können. Zum Beispiel, bietet es sich an jemanden der das Buch ”Herr der Ringe”gekauft auch die DVDs der Verfilmung anzubieten. Eine weitere ähnliche Aufgabe ist die Bestimmung von geeigneten Kombiangeboten zu einem bereits ausgwählten Artikel.