Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Peer Kröger
Johannes Aßfalg, Karsten Borgwardt

Knowledge Discovery in Databases

WS 2005/06

Übungsblatt 1

Abgabe aller mit Hausaufgabe *markierten Aufgaben bis Donnerstag, 03.11.2005, 8:30 Uhr, vor der Vorlesung beim Dozenten oder im Übungsbriefkasten*

Aufgabe 1-1 Metrische Distanzfunktionen
Hausaufgabe

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) $dist(o_1, o_2) \geq 0$,

(2) $dist(o_1, o_2) = 0 \Leftrightarrow o_1 = o_2$,

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

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

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

Hinweis: Auch für Funktionen, für die Sie bereits zu wissen glauben, dass es sich um Metriken handelt, müssen Sie dies sauber beweisen. Sätze wie “Diese Distanzfunktion ist eine Metrik, weil wir das schon aus der Vorlesung wissen.” wird den Korrektoren kein “OK” für die Bewertung entlocken können.

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)? Diskutieren Sie, welche Erkenntnisse das Data Mining für jede Aufgabe zusätzlich bringen könnte.

(a) **Gläserner Kunde**:
Eine Bank möchte einmal mehr ihren Gewinn erhöhen. Die Buchhaltung hat festgestellt, dass Kunden, die über eine gewisse Kontoart verfügen (sog. Gold-Kunden), den höchsten Beitrag zum Gewinn leisten. Als erste Aktion, möchte die Bank nun feststellen, welche Merkmale Gold-Kunden typischerweise auszeichnen.

(b) **Nicht nur für Bioinformatiker:**
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) **Big brother is watching you:**
Ein Rechenzentrum speichert das Benutzerverhalten seiner User mittels Web-Logs. Da in letzter Zeit öfter Missbrauch mit den freien Internetkonten getrieben wurde, möchte der Leiter des Rechenzentrums die entsprechenden User identifizieren.

(d) **Mensch und Maschine:**
Ein englischsprachiger Eigenheimbewohner hat Ärger mit einem seiner Fenster: es ist undicht. Was ist zu tun? Fortschrittlich denkend setzt er sich an seinen PC und füttert die Suchmaschine Google mit dem (nicht wirklich selektiven) Wort “windows”. Überraschenderweise findet Google ein paar Seiten zuviel über das Thema “windows”, so dass der frustrierte User sich nun nicht wirklich schlauer fühlt. Sekundenschnell rechnet er aus, dass er, wenn er alle Trefferseiten browsed, erstens alt und zweitens arm

wird. Zufällig fällt ihm ein, dass er einmal eine Vorlesung namens KDD gehört hat, und eine Methode, die darin vorgestellt wurde erscheint ihm geeignet, die Suche etwas einzugrenzen. Welchen Geistesblitz könnte er gehabt haben?

(e) **Kundenmanipulation, Gläserner Kunde II**:
Eine Supermarktkette sammelt Daten zu jedem Einkauf, d.h. die gesammelten Daten enthalten Informationen darüber, welche Produkte zusammen von einem Kunden gekauft werden. Um die Angebotspalette zu optimieren, möchte die Geschäftsleitung gerne wissen, welche Waren bevorzugt zusammen gekauft werden.