Universität Kassel
Fachbereich Mathematik/Informatik
Fachgebiet Wissensverarbeitung
Hertie-Stiftungslehrstuhl
Wilhelmshöher Allee 73
34121 Kassel

Email: hotho@cs.uni-kassel.de
Tel.: ++49 561 804-6252

Dr. Andreas Hotho
Prof. Gerd Stumme

24.05.04

Knowledge Discovery

Übungsblatt 2

Sommersemester 2004

Aufgabe 1: Preprocessing und Datenbeschaffenheit

a) Welche zwei Typen des Preprocessing gibt es? Diskutieren Sie in diesem Zusammenhang den Begriff Datenverständnis?

b) In der Statistik unterscheidet man die Attributausprägungen einer vorgegebenen Menge von Daten mittels Skalen mit unterschiedlichem Skalenniveau. Nennen Sie die drei wichtigsten Skalenniveaus und beschreiben Sie sie kurz. Was bedingt ein Skalenniveau bei der Untersuchung von Daten?

c) Welches Ziel wird in Bezug auf die spätere Anwendung von Algorithmen mit dem Preprocessing der Daten im Data Mining verfolgt. Nennen Sie in diesem Zusammenhang zwei Beispiele, in denen Algorithmen bestimmte Preprocessingschritte erzwingen.

Aufgabe 2: Entscheidungsbäume

a) Welche Form sollte ein Entscheidungsbaum haben? Möglichst breit oder möglichst tief? Warum?

b) Ein Krankenhaus möchte die Diagnosefähigkeit seiner Ärzte unterstützen. Dazu wurden Daten über gesunde und kranke Patienten gesammelt. Die Krankenhausleitung hat erfahren, dass man mit einem Entscheidungsbaumverfahren anhand vorhandener Beispieldaten ein Modell generieren kann, welches die Entscheidung eines Arztes simuliert. Berechnen Sie mittels der folgenden Daten einen Entscheidungsbaum und zeichnen Sie diesen auf.

Patient Nr.	Heart Rate	Blood Pressure	Klasse
1	irregular	Normal	Ill
2	regular	Normal	Healthy
3	irregular	Abnormal	Ill
4	irregular	Normal	Ill
5	regular	Normal	Healthy
6	regular	Abnormal	Ill
7	regular	Normal	Healthy
8	regular	Normal	Healthy

Nutzen Sie zum Erstellen des Entscheidungsbaumes das *gain-ratio*(*x*) Kriterium:

Ohne Taschenrechner nähern Sie bitte den Logarithmus mittels folgender Formel an: $log_2(x) = 1 - \frac{1}{x}$