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

04.07.05

Knowledge Discovery

Übungsblatt 8

Sommersemester 2005

Aufgabe 1: kNN-Verfahren

a) Geben Sie die prinzipiellen Schritte eines kNN-Verfahrens an und nennen Sie mindestens je ein Abstandsmaß für numerische und kategorische Werte.
b) Diskutieren Sie die Vor- und Nachteile des Verfahrens.
c) Berechnen Sie für k = 4 den Abstand zum Beispiel (sunny,cool, high, true) für den Datensatz aus dem letzten Übungsblatt.

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, welches wie folgt definiert ist:

$$gain\ ratio(T,A) = \frac{\text{Informationsgewinn}(T,A)}{\text{split info}(T,A)} \text{ und } split\ info(T,A) = -\sum_{i=1}^{m} \frac{|T_i|}{|T|} * \log_2\left(\frac{|T_i|}{|T|}\right) \text{ bits}$$

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

c) Definieren Sie den Begriff Overfitting. Schlagen Sie eine Strategie zur Vermeidung vor.

d) Beschreiben Sie das prinzipielle Vorgehen, um das Entscheidungsbaumlernen zu parallelisieren.

Aufgabe 3: SVM

Gegeben sei folgender zweidimensionaler Trainingsdatensatz:

$S = \{(x_i, y_i)\} = \{(2,0;-1), (0,2;-1), (2,2;1), (3,2;1)\}$

a) Bestimmen Sie die d den Abstand der optimalen Hyperebene zum gegeben Trainingsdatasatz.
b) Die Entscheidungsfunktion eines linearen Klassifizieres in erster Form sei:
   $f(x) = sgn(w*x+b)$
   Berechnen Sie die Gewichte w und den Schwellwert b.
c) Welche Trainingsbeispiele sind die Supportvektoren?
d) Klassifizieren Sie das Beispiel: (1,4).