Semestralklausur zur Vorlesung

Maschinelles Lernen und Data Mining

Prof. J. Fürnkranz / Dr. G. Grieser
Technische Universität Darmstadt — Wintersemester 2003/04
Termin: 13. 2. 2004

Name: **Vorname:** **Matrikelnummer:**

Fachrichtung:

Punkte: (1) .... (2) .... (3) .... (4) .... (5) .... **Summe:**

Aufgabe 1

Hinweis: Bei falsch angekreuzten Antworten werden Punkte abgezogen. Die minimale Punktzahl für Aufgabe 1 ist allerdings 0 Punkte; Sie können also keine Punkte verlieren.

1–a Welche der folgenden Algorithmen können direkt mit Multiklassenproblemen umgehen?

☐ ID3

☐ Naive Bayes

☐ SVM

Wie können Sie Multiklassenprobleme mit Algorithmen lösen, die nur 2-Klassenprobleme behandeln können?

1–b Welche der folgenden Algorithmen lassen sich boosten?

☐ ID3

☐ k-means Clustering

☐ Apriori

☐ Feedforward-Netz

1–c Was sind die Hypothesen beim Bagging von ID3 (syntaktische Repräsentation der Hypothesen)?

1–d Angenommen, Sie haben einen Datensatz von 10 000 Beispielen mit jeweils 10 000 Attributen. Welchen der folgenden Algorithmen würden sie keinesfalls einsetzen:

☐ ID3

☐ k-NN

☐ Feedforward-Netz

Begründung:

1–e Erklären Sie kurz die Funktion des Hidden Layers bei Feedforward-Netzen.

1–f Was sind die Grundideen bei Supportvektormaschinen?

Aufgabe 2

2–a Gegeben seien 2 numerische Attribute. Als Hypothesenraum benutzen wir die Menge aller Decision Stumps (Entscheidungsbäume mit einem binären inneren Knoten) über diesen Attributen.

Wie groß ist die VC-Dimension dieses Hypothesenraumes?

Was können Sie über die PAC-Lernbarkeit dieses Hypothesenraumes sagen?

2–b Wie groß ist die VC-Dimension, falls wir als Hypothesen beliebige Entscheidungsbäume zulassen? Was bedeutet das für die PAC-Lernbarkeit?

Aufgabe 3

Gegeben sei folgende Hierarchie von Begriffen:

Beobachtet werden Objekte, die durch Begriffspaare charakterisiert werden, die man an der untersten Ebene dieser Taxonomien finden kann (also z.B. “blaues Dreieck”).

3–a Definieren Sie unter Verwendung der gegebenen Begriffshierarchien eine Vorschrift zur minimalen Generalisierung.

3–b Gegeben seien folgende S und G-Sets:

G: dunkle Vielecke, beliebige Quadrate
S: blaue Quadrate

Skizzieren Sie den Version Space, der durch diese Mengen definiert wird.

3–c Wie würden Sie mit Hilfe des oben gegebenen Version Spaces die folgenden Beispiele klassifizieren (mit Begründung):

Objekt	**Klasse**
blaues Quadrat	?
blauer Kreis	?
blaues Dreieck	?

3–d Gegeben seien wiederum die S- und G-sets aus Punkt 3–b. Wie verändern sich die Sets nach Eintreffen des Beispiels

gelbes Dreieck +

Wie interpretieren Sie dieses Ergebnis?

Aufgabe 4

Der Gini index berechnet sich für eine Menge von Beispielen S wie folgt:

$$g(S) = 1 - \sum_{i=1}^{c} p(i|S)^2$$

wobei c die Anzahl der Klassen in den Beispielen ist, und $p(i|S)$ die Wahrscheinlichkeit (relative Häufigkeit) von Beispielen der Klasse i in S ist.

4–a Diskutieren Sie die Eigenschaften dieses Maßes. Wo nimmt es ein Maximum an, wo nimmt es ein Minimum an? Welcher der beiden Fälle ist für einen Klassifikationsalgorithmus interessanter? Warum?

4–b Definieren Sie analog zu Information Gain das Maß "Gini Gain" als Splitting Criterion für Decision Trees, wobei der Gini index die Rolle der Entropie im Information Gain übernehmen soll.

4–c Berechnen Sie mithilfe des Gini Index einen Entscheidungsbaum für folgende Daten.

A1	A2	A3	Class
0	1	1	–
0	0	1	+
1	1	0	+
0	1	0	–
0	0	0	+
1	0	1	+

[Für die halbe Punktanzahl können Sie auch Information gain zur Konstruktion des Baumes verwenden.]

Aufgabe 5

Gegeben sei folgende Häufigkeitsmatrix:

	kauft CDs	
kauft DVDs	ja	nein
ja	200	100
nein	400	200

5–a Berechnen Sie Support, Confidence, und Lift für die Assoziationsregel DVD → CD

5–b Ist diese Assoziationsregel interessant? Begründung?

5–c Skizzieren Sie, wie Sie einen Assoziationsregel-Lerner mit einem Covering/Separate-and-Conquer Verfahren kombinieren können, um ein Klassifikationsproblem durch Lernen einer Regel-Menge zu lösen.

5–d Die Klassifikationsregeln, die von obigem Algorithmus gelernt werden, erfüllen alle die Kriterien, daß sie mit einem (vom Benutzer spezifizierten) minimalem Support und einer minimalen Konfidenz auftreten. Skizzieren Sie in einem PN-Raum, in welcher Region die gefundenen Regeln liegen müssen?