Semestralklausur zur Vorlesung

Maschinelles Lernen: Symbolische Ansätze

Prof. J. Fürnkranz / Dr. G. Grieser

Technische Universität Darmstadt — Wintersemester 2005/06

Termin: 23. 2. 2006

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

Fachrichtung:

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

Aufgabe 1 (15 Punkte)

1–a Sie haben Q-learning verwendet, um eine Funktion zu lernen, die Ihnen aus einem Labyrinth hinaushilft. Sie befinden sich im Zentrum und können nach Norden, Westen, Süden oder Osten gehen. Wie setzen Sie die gelernte Funktion ein, um die richtige Richtung auszuwählen?

1–b Welchen Algorithmus würden Sie wählen, wenn Ihre Anwendung verlangt, daß die Anzahl der *false negatives* möglichst gering gehalten werden soll?

☐ Find-S

☐ Find-G

☐ Candidate Elimination

Begründung?

1–c Erklären Sie kurz wozu der `FilteredClassifier` in Weka nützlich ist.

1–d Gegeben sei eine Datenbank über 100.000 Informatik-Fachbücher. Jedes Buch ist durch eine Menge von Stichwörtern beschrieben, insgesamt werden in der Datenbank 1000 solcher Stichwörter benutzt.

Ihre Aufgabe sei nun, ein Programm zu schreiben, das aus der Datenbank alle Bücher zum Thema *Künstliche Intelligenz* heraussucht. Als Ausgangsbasis haben Sie eine Menge von 1000 Büchern, von denen Sie wissen, ob sie zum Thema *Künstliche Intelligenz* gehören oder nicht.

Wie würden Sie herangehen?

1–e Nehmen Sie an, Sie haben eine Beispielmenge mit 10 Attributen, von denen 9 Attribute exakte Kopien voneinander sind.

Welchen Effekt erwarten Sie, wenn Sie Naive Bayes auf diese Daten anwenden?

Aufgabe 2 (20 Punkte)

Gegeben sei ein Datensatz mit drei Attributen:

Haarfarbe:	*blond*, *braun*, *schwarz*
Größe:	*klein*, *groß*
Augenfarbe:	*grün*, *blau*

Der Hypothesenraum besteht aus Disjunktionen (Oder-Verknüpfungen) von maximal einem Wert pro Attribut, einer speziellsten Theorie *false*, die keine Beispiele abdeckt, und einer allgemeinsten Theorie *true*, die alle Beispiele abdeckt.

Zum Beispiel deckt die Hypothese *blond* $\vee$ *blau* alle Personen ab, die entweder blond oder blauäugig sind (in der Datenmenge aus Aufgabe b sind das z.B. die Beispiele 1, 3, 4).

Beachte: Hypothesen wie *blond* $\vee$ *braun*, die mehrere Werte desselben Attributs verwenden, sind nicht im Hypothesenraum.

2–a Geben Sie in dieser Hypothesensprache alle minimalen Generalisierungen und Spezialisierungen der Hypothese *blond* $\vee$ *blau* an.

2–b Folgende Beispiele treffen in dieser Reihenfolge ein:

1	*braun*	*groß*	*blau*	+
2	*braun*	*klein*	*grün*	−
3	*schwarz*	*klein*	*blau*	−
4	*blond*	*klein*	*grün*	+

Das erste Beispiel kodiert also eine Person, die braune Haare und blaue Augen hat und groß ist.

Führen Sie auf diesen Beispielen den Candidate-Elimination Algorithmus zur Berechnung des Version Spaces durch und geben Sie nach jedem Schritt das S-Set und das G-Set an.

2–c Nehmen Sie an, die Beispiele aus Aufgabe b kämen in umgekehrter Reihenfolge. Was wäre dann das Resultat des Candidate-Elimination Algorithmus? Begründung?

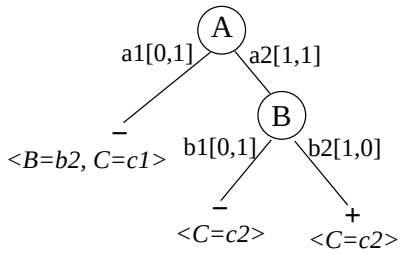
Aufgabe 3 (25 Punkte)

Diese Aufgabe bezieht sich auf ID5R.

Statt des original benutzten Maßes *Gain* benutzen wir hier eine vereinfachte Variante:

> Wähle denjenigen Test t, der folgenden Ausdruck maximiert:
> $$\sum_{v \in Werte(t)} (|S_v^+| - |S_v^-|)^2$$
> wobei S_v^+ die Menge aller Beispiele ist, für die der Test t den Wert v hat und deren Klassifikation positiv (+) sind. (S_v^- analog).
>
> Das Maß summiert also die quadrierte Differenz der Anzahl der positiven und negativen Beispiele in einer Menge.
>
> Falls mehrere Tests den gleichen Wert liefern, nimm denjenigen, der im Alphabet zuerst kommt, d.h. A vor B vor C):

Nehmen wir an, ID5R habe den folgenden Baum erzeugt:



3–a Geben Sie alle Beispiele an (d.h. lesen Sie diese aus dem Baum ab), die zur Erzeugung des Baumes geführt haben.

3–b Nun kommt folgendes Beispiel hinzu:
$$\langle A = a2, B = b1, C = c1; - \rangle$$
Wie lautet die neue Hypothese von ID5R? Erläutern Sie dabei Ihre Vorgehensweise, geben Sie die einzelnen Schritte an. Je detaillierter Sie Ihren Lösungsweg beschreiben, desto mehr erleichtern Sie uns die Vergabe von Punkten.

3–c Eine Möglichkeit, aus unvollständigen Beispielen zu lernen ist, Wahrscheinlichkeiten für jeden möglichen Attributwert zuzuweisen und Anteile der Beispiele in die Teilbäume zu propagieren (siehe Folie 34 im Satz *Entscheidungsbäume*).

Diskutieren Sie diese Idee für ID5R. (Z.B.: Welches grundlegende Problem taucht hier auf? Wie kann man damit umgehen?)

Aufgabe 4 (22 Punkte)

Gegeben sei ein Datensatz mit 300 Beispielen, davon 2/3 positiv und 1/3 negativ.

4–a Ist die Steigung der Isometrien für Accuracy im Coverage Space für dieses Problem

□ > 1

□ $= 1$

□ < 1?

4–b Ist die Steigung der Isometrien für Accuracy im ROC Space für dieses Problem

□ > 1

□ $= 1$

□ < 1?

4–c Sie verwenden einen Entscheidungsbaum, um die Wahrscheinlichkeit für die positive Klasse zu schätzen. Sie evaluieren drei verschiedene Thresholds t (alle Beispiele mit einer geschätzten Wahrscheinlichkeit $> t$ werden als positiv, alle anderen als negativ klassifiziert) und messen folgende absolute Anzahlen von false positives und false negatives:

t	fn	fp
0.7	40	20
0.5	30	60
0.3	10	80

Geben Sie für jeden Threshold an, für welchen Bereich des Kostenverhältnisses $\frac{c(+|-)}{c(-|+)}$ der Threshold optimal ist.

4–d Wie hoch ist die maximale Genauigkeit (Accuracy), die Sie im Szenario von Punkt c bei einer false positive rate von maximal 30% erreichen können? Wie gehen Sie dabei vor?

4–e Sie erfahren, daß in Ihrer Anwendung ein false positive 2 Cents kostet, und ein false negative 5 Cents kostet. Mit welchem Threshold können Sie die Kosten minimieren? Wie hoch sind die entstandenen minimalen Kosten für diese 300 Beispiele?

4–f Sie bekommen die Möglichkeit, zusätzlich zu den vorhandenen 300 Beispielen noch 400 selbst auszuwählen. Wie würden Sie die Auswahl treffen, damit ein Lerner, der Kosten nicht berücksichtigen kann, unter den in Punkt e angegebenen Kosten möglichst effektiv wird?

Aufgabe 5 (18 Punkte)

Sie wissen, daß die Assoziationsregel

`beatles`, `stones` → `dylan`, `cohen`

in einem Datensatz mit 1000 Einträgen über Musikpräferenzen einen Support von 0.4 und eine Konfidenz von 0.8 hat.

5–a Beantworten Sie folgende Fragen unter Angabe eines möglichst kleinen Intervalls der möglichen Werte (kann auch $[-\infty,+\infty]$ oder auch nur ein einziger Wert sein):

- Wie groß ist der Lift dieser Regel?
- Wie viele Personen mögen `beatles` und `stones`?
- Wie viele Personen mögen `stones` und `dylan`?
- Wieviel Prozent der Personen, die `beatles`, `dylan`, und `stones` mögen, mögen auch `cohen`?
- Wie viele Personen, die `stones` mögen, mögen auch `beatles`, `dylan`, und `cohen`?

5–b Sie wissen, daß Apriori die oben angegebenen Assoziationsregel gefunden hat.

Zusätzlich erfahren Sie noch, daß folgende Itemsets und keine ihrer Obermengen frequent waren:

{ `beatles`, `dylan`, `young` }

{ `beatles`, `young`, `stones` }

{ `young`, `dylan`, `stones` }

Geben Sie die positive und die negative Border an.