Semestralklausur zur Vorlesung

Maschinelles Lernen und Data Mining

Prof. J. Fürnkranz / Dr. G. Grieser
Technische Universität Darmstadt — Wintersemester 2004/05
Termin: 14. 2. 2005

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

Fachrichtung:

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

Aufgabe 1 (18 Punkte)

Beantworten Sie die folgenden Fragen nach Möglichkeit kurz und prägnant.

1–a Skizzieren Sie kurz, welche Probleme Sie sich bei einer praktischen Anwendung eines Assoziationsregel-Lerners auf Personen-bezogene Daten erwarten.

1–b Mit welchen Parametern kann man die Baumgröße in Weka's J48 beeinflussen. Beschreiben Sie kurz was eine Veränderung ihrer Default-Werte bewirkt (ohne auf Details wie z.B. Formeln einzugehen).

1–c Worauf müssen Sie achten, wenn Sie eine überwachte Feature Subset Selection Methode wie Relief verwenden und Cross-Validierung zur Genauigkeitsabschätzung einsetzen wollen?

1–d Sie haben ein Lern-Problem mit 100 Klassen. Nennen Sie zwei Methoden, die es Ihnen erlauben, eine Support Vector Maschine auf dieses Problem anzuwenden.

Welches der beiden Verfahren würden Sie wählen, wenn Sie beliebig viel Zeit zum Trainieren der Klassifizierer haben, aber die Antwortzeit bei der Klassifikation neuer Beispiele möglichst gering sein sollte? Warum?

1–e Eine Firma tritt mit dem Problem an Sie heran, eine Steuerung für eine adaptive Ampelanlage zu entwerfen, die in der Lage ist, die Länge der Rot- und Grünphasen an das jeweilige Verkehrsaufkommen anzupassen, sodaß die erwartete Wartezeit für jedes Fahrzeug minimiert wird. Eine realistische Software-Simulation des Verkehrsaufkommens steht zur Verfügung.

Welchen Lern-Algorithmus würden Sie zur Lösung dieses Problems am ehesten in Betracht ziehen?

☐ Ripper

☐ Q-learning

☐ k-means Clustering

Begründung?

1–f Was charakterisiert Ihrer Erfahrung nach die Lernalgorithmen, die in Weka unter `weka.classifiers.meta` zusammengefaßt sind?

Aufgabe 2 (25 Punkte)

Gegeben seien folgende Daten, die angeben, ob eine Person einen Kredit von einer Bank erhalten hat oder nicht.

age	education	married	income	credit
old	secondary	no	medium	yes
young	college	yes	low	yes
old	secondary	yes	medium	yes
young	college	no	high	yes
young	secondary	no	high	yes
old	secondary	no	high	yes
old	college	no	high	yes
old	college	yes	low	yes
young	primary	yes	medium	no
old	primary	yes	low	no
young	secondary	yes	medium	no
young	college	no	medium	no

2–a Listen Sie alle Wahrscheinlichkeiten auf, die Sie aus den Daten schätzen müssen, um einen vollständigen Naive Bayes Klassifizierer zur Vorhersage der Kreditwürdigkeit einer Person zu erhalten.

2–b Berechnen Sie aus den Daten Schätzwerte für diejenigen Wahrscheinlichkeiten, die der Naive Bayes Klassifizierer benötigt, um das Beispiel

age	education	married	income	credit
young	college	no	medium	?

zu klassifizieren.

Welche Klasse wird dann für dieses Beispiel vorhergesagt?

Hinweis: Die Wahrscheinlichkeiten sollen einfach durch die relative Häufigkeit des Auftretens des fraglichen Ereignisses geschätzt werden, und können selbstverständlich als Bruchzahlen angegeben werden. Die Wahrscheinlichkeit, daß eine Person ein mittleres Einkommen hat, würde also mit $Pr(\text{income} = \text{medium}) = 5/12$ geschätzt.

2–c Für das folgende Beispiel ist es (nach obiger Methode) ist es nicht möglich, eine begründete Klassifikation vorzunehmen.

age	education	married	income	credit
old	primary	yes	high	?

Auf welches Problem ist das zurückzuführen? Wie kann man dieses Problem in den Griff bekommen?

2–d Als zusätzliche Informationen erhalten sie für jede Person das Einkommen der letzten neun Jahre. Zufälligerweise ist dieses für alle Personen in der Datenbank identisch mit dem momentanen Einkommen (Spalte income), sodaß sich nun 10 identische Spalten im Datensatz finden.

- Welches Problem ergibt sich dadurch für den Naive Bayes Klassifizierer?
- Was ist der zu erwartende Effekt?
- Wie würde nun das Beispiel aus 2-b klassifiziert werden (unter der Annahme, daß es nun ebenfalls 10 identische income-Werte enthält)?
- Nennen Sie einen Lernalgorithmus, für den diese Situation kein Problem darstellt. Begründung?

Aufgabe 3 (25 Punkte)

Gegeben sei folgender Auszug aus einer Datenbank, bestehend aus einem numerischen Attribut *att* und der Klassen-Variable *class*.

att	*class*
42	-
19	+
51	+
30	-
18	+
22	+

att	*class*
38	-
50	+
21	+
48	+
41	-
29	-

3–a Was wäre für einen Entscheidungsbaum-Lerner die ideale Diskretisierung des Attributs *att*? Glauben Sie, daß die Entropy-Split Methode diese ideale Diskretisierung finden würde? Begründung?

3–b Diskretisieren Sie das Attribut in drei Intervalle unter Verwendung von equal-width.

3–c Diskretisieren Sie das Attribut in drei Intervalle unter Verwendung von equal-frequency.

3–d Clustering Algorithmen lassen sich selbstverständlich auch auf einzelne Attribute anwenden. Wenn man jedem Cluster einen symbolischen Wert zuordnet, läßt sich das Ergebnis als eine Diskretisierungsvorschrift interpretieren.

Verwenden sie Bottom-Up Agglomerative Clustering, um eine Cluster-Hierarchie über dem Attribut *att* zu definieren (*class* wird nicht betrachtet).

- Stellen Sie die Cluster-Hierarchie graphisch als Baumstruktur dar.
- Geben Sie die aus dem Clustering resultierende Diskretisierung in drei Intervalle an.
- Welche Anzahl von Clustern würde eine optimale Diskretisierung produzieren? Was läßt sich daraus über die Verteilung der positiven bzw. der negativen Beispiele folgern?

Hinweis: Die Distanz zwischen zwei Attributwerten a_i und a_j sei als Absolutbetrag ihrer Differenz definiert ($d(a_i, a_j) = |a_i - a_j|$). Für die Distanz zwischen zwei Clustern verwenden Sie die Single-Link Distanz. Das Attribute *class* wird nicht in Betracht gezogen.

3–e Zu welchem der beiden in der Vorlesung besprochenen Verfahren Chi-Merge und Entropy-Split ist das im vorigen Punkt besprochene Verfahren ähnlicher? Wie unterscheidet es sich von beiden Verfahren? Begründen Sie Ihre Antworten.

Aufgabe 4 (14 Punkte)

Ein on-line Buchgeschäft möchte eine Datenbank mit 10,000 Kunden analysieren, die jeweils eines oder mehrere von 500 verschiedenen Büchern gekauft haben. Zur Entdeckung von Assoziationsregeln wird der Algorithmus Apriori mit einem Minimum Support von 3% und einer minimalen Konfidenz von 75% verwendet.

4–a Es wird festgestellt, daß die beiden häufigsten Verkäufe "Harry Potter und der Stein der Weisen" (HP1) und "Harry Potter und die Kammer des Schreckens" (HP2) sind. HP1 wurde von 6,000 Kunden und HP2 von 8,000 Kunden gekauft. 4,000 Kunden kauften beide Bücher.

Welche der beiden Assoziationsregeln findet sich im Output des Assoziationsregel-Lerners?

□ HP1 → HP2

□ HP2 → HP1

□ beide

□ keine von beiden

Geben Sie Support und Konfidenz für beide Regeln an.

4–b Wenn man annimmt, daß alle Kunden, die beide Bücher gekauft haben, zuerst HP1 und später HP2 gekauft haben: Wie interpretieren Sie den Einfluß des Kaufs von HP1 auf den Kauf von HP2?

4–c Die längste Assoziationsregel, die gefunden wurde, wurde aus einem Itemset der Größe 20 konstruiert. Geben Sie eine möglichst große untere Schranke für die Anzahl der gefundenen Frequent Itemsets an.

Aufgabe 5 (18 Punkte)

Gegeben seien 2 numerische Attribute. Als Hypothesenraum benutzen wir die Menge aller Entscheidungsbäume mit maximal zwei inneren Knoten über diesen Attributen, Tests sind jeweils Vergleiche des Werts eines Attributs mit einer beliebigen Konstanten.

5–a Ist der Hypothesenraum endlich oder unendlich?

5–b Wie groß ist die VC-Dimension dieses Hypothesenraumes? Begründen Sie Ihre Antwort.

5–c Wie groß ist die VC-Dimension, wenn beliebig komplexe Formeln (d.h. verschachtelte Konjunktionen und Disjunktionen beliebiger Tests) in den Knoten stehen können? Warum?

5–d Was bedeutet eine unendliche VC-Dimension? Welche Gefahr besteht bei Hypothesenräumen mit unendlicher VC-Dimension?