Semestralklausur zur Vorlesung

Web Mining

Prof. J. Fürnkranz
Technische Universität Darmstadt — Sommersemester 2012
Termin: 17. 7. 2012

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

Fachrichtung:

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

Wichtig!

- **Aufgaben:** Diese Klausur enthält auf den folgenden Seiten 5 Aufgaben zu insgesamt 100 Punkten. Jede Aufgabe steht auf einem eigenen Blatt. Kontrollieren Sie *sofort*, ob Sie alle sechs Blätter erhalten haben!
- **Zeiteinteilung:** Die Zeit ist knapp bemessen. Wir empfehlen Ihnen, daß Sie sich zuerst einen kurzen Überblick über die Aufgabenstellungen verschaffen, und dann mit den Aufgaben beginnen, die Ihnen am meisten liegen.
- **Papier:** Verwenden Sie nur Papier, das Sie von uns ausgeteilt bekommen. Sie können Ihre Lösungen beliebig auf die sechs Blätter verteilen, solange klar ersichtlich ist, welche Lösung zu welcher Aufgabe gehört. Sollten sich allerdings mehrere Lösungen zu derselben Aufgabe finden, suchen wir uns eine aus. Insbesondere können Sie auch auf den Rückseiten schreiben!
  Brauchen Sie zusätzlich Papier (auch Schmierpaper), bitte melden.
- **Hilfsmittel:** Als Hilfsmittel ist darf ein handbeschriebenes DIN-A4-Blatt benutzt werden. Ausländische Studenten dürfen darüber hinaus gedruckte Wörterbücher verwenden. Elektronische Wörterbücher sind nicht erlaubt.
- **Fragen:** Sollten Sie Teile der Aufgabenstellung nicht verstehen, bitte fragen Sie!
- **Abschreiben:** Sollte es sich (wie in den letzten Jahren leider immer wieder) herausstellen, daß Ihre Lösung und die eines Kommilitonen über das zu erwartende Maß hinaus übereinstimmen, werden beide Arbeiten negativ beurteilt (ganz egal wer von wem und egal in welchem Umfang abgeschrieben hat).
- **Ausweis:** Legen Sie Ihren *Studentenausweis* sichtbar auf Ihren Platz.
- **Aufräumen:** Sonst darf außer Schreibgerät, Essbarem, von uns ausgeteiltem Papier und eventuell Wörterbüchern nichts auf Ihrem Platz liegen. Taschen bitte unter den Tisch! Wer bei diesen Temperaturen einen Mantel mithat, wird gebeten, ihn anzubehalten.

Gutes Gelingen!

Aufgabe 1 23 Punkte (3/4/4/4/4/4)

Sie möchten eine mobile App für Smartphones programmieren, die zum aktuellen Standort passende sogenannte Points of Interest (POI) angibt. Ihr Smartphone-System erlaubt es Ihnen, die aktuellen GPS-Koordinaten sowie die aktuelle Adresse in Form von Straßenname, Hausnummer, Postleitzahl und Stadt zu erfahren. Ein anderer Dienst erlaubt es Ihnen, zu einer angegeben Adresse die GPS-Koordinaten zu erhalten, so daß Sie diese z.B. in einer Karte darstellen können.

Die folgenden Lösungsvorschläge sind meist nur sehr kurz wiedergegeben ohne weitere Erklärungen und sollen hauptsächlich angeben, welcher behandelter Aspekt für einen Lösungstext ausgereicht hätte. Die einfache Erwähnung des Aspekts würde im Allgemeinen nicht für die volle Punktzahl ausreichen.

1–a In einem ersten Schritt möchten Sie, daß die App beliebige möglicherweise interessanten Punkte in der Umgebung findet. Dafür steht Ihnen eine einfache Web-Suchmaschine zur Verfügung, die Ihnen zu einer Query eine (positionsunabhängige) Liste von Seiten zurückliefert. Wie könnten Sie diese nützen, um interessante Punkte in einer Stadt-Umgebung zu finden?

Z.B.: nach Strasse, PLZ und Stadt suchen, die Hausnummer und ein Titel jeweils aus den Snippets der Suchmaschine extrahieren oder die Seite selbst crawlen, Verwendung Dienst Adresse→GPS und Anzeige.

1–b Auf der Liste der POI erscheinen Buchläden gemischt mit Cafés, Restaurants, Museen etc. Wie könnten Sie die Ergebnisse automatisiert nach Themen gruppieren?

Clustering, möglicherweise auch Klassifizierung

1–c Jedes Mal, wenn ein Benutzer die App verwendet und einen POI aus der angezeigten Liste auswählt, wird dies registriert. Wie könnten Sie diese lokalen Information nutzen, um die Reihenfolge der Resultate von zukünftigen Queries zu verbessern bzw. um die Ergebnisliste zu filtern?

Relevance Feedback

1–d Sie implementieren eine Funktion, mit der die Informationen aus Aufgabe c) für alle Benutzer der App an einen zentralen Server übermittelt werden. Mit welchem Verfahren könnten Sie auch noch diese Information nützen, um die Reihung der Resultate zu verbessern, und warum?

Collaborative Filtering

1–e Sie möchten zusätzlich zur Gruppierung der POIs diesen auch Kategorien zuweisen. Sie möchten dazu die bereits vorhandene Kategorisierung von Lesezeichen-Portalen wie del.icio.us, digg, Mister Wong oder dmoz.org verwenden. Es gibt jedoch immer wieder Seiten zu ihren POIs, die nicht kategorisiert sind. Hierfür wollen Sie einen Klassifizierer lernen. Welches Verfahren kennen Sie aus der Vorlesung, das sich dabei den invertierten Index Ihrer Suchmaschine zu Nutze machen kann?
k-NN

1–f Einige Web-Seiten enthalten nicht genügend Informationen, um sie verläßlich Kategorien zuzuordnen. Allerdings läßt sich immer eine ausreichende Menge kategorisierter Seiten finden, die mit diesen Seiten verlinkt sind. Wie könnten Sie diese Information zusätzlich für die Klassifizierung nutzen?
Hyperlink-Ensembles

Aufgabe 2 19 Punkte (4/5/7/3)

Auf einem Trainingssatz an Dokumenten mit drei Klassen A, B, C wurde ein *compress* Klassifizierer gelernt, d.h. die Dokumente jeder Klasse wurden jeweils zu einem Archiv komprimiert. Die Archive für die Klassen haben folgende Größen:

Archivname	Klasse	Archivgröße [bits]
c_A	A	536
c_B	B	446
c_C	C	362

Jedes Archiv besitzt ein Kompressions-Wörterbuch mit einer Übersetzungtabelle von Buchstaben zu Bitfolgen. Die folgende Tabelle gibt an, wieviel Bits die Archivierung von Buchstaben unter Verwendung dieser Wörterbücher benötigt:

Archiv	Bitlänge 4	Bitlänge 6	Bitlänge 8	Bitlänge 10
c_A	a e o t	h i n r s	c d f g l m p u w	b j k q v x y z
c_B	e i n r	a d h s t	b c f g l m o u w	j k p q v x y z
c_C	a e s t	i l n r u	c d f g m o p q v	b h j k w x y z

Leerzeichen benötigen immer 2 Bit.

Beispiel: die Kompression des Buchstabens d würde durch c_A 8 bit, durch c_B 6 bit und durch c_C 8 bit benötigen.

2–a Sie wollen mit dem compress Algorithmus das Dokument

"web mining"

klassifizieren. Welche Klasse wird für dieses Dokument vorhergesagt und warum?

Archiv	*w*	*e*	*b*		*m*	*i*	*n*	*i*	*n*	*g*	$\sum$
c_A	*8*	*4*	*10*	*2*	*8*	*6*	*6*	*6*	*6*	*8*	*64*
c_B	*8*	*4*	*8*	*2*	*8*	*4*	*4*	*4*	*4*	*8*	*54*
c_C	*10*	*4*	*10*	*2*	*8*	*6*	*6*	*6*	*6*	*8*	*66*

Es wird B vorhergesagt, da das zur geringsten Erhöhung der Archivgröße führen würde.

2–b Klassifizieren Sie folgende Dokumente, die jeweils nur aus einem Buchstaben bestehen, und tragen Sie das Ergebnis in folgende Konfusionsmatrix ein. Die reale Klassen-Zugehörigkeit der Dokumente ist im Spaltennamen der oberen Tabelle angegeben. Tragen Sie hierbei jedes Dokument (repräsentiert durch seinen Buchstaben) in das entsprechende Feld der Matrix ein.

Berechnen Sie anschließend die Genauigkeit (Accuracy), mit der diese Dokumente klassifiziert werden.

A	B	C
b, i, o, v	d, q , r	l, s, u

	vorhergesagt		
real	A	B	C
A			
B			
C			

	vorhergesagt		
real	*A*	*B*	*C*
A	*o*	*b, i*	*v*
B		*d, r*	*q*
C			*l, s, u*

Accuracy: $\frac{6}{10} = 0.6$

2–c Berechnen Sie für jede Klasse Recall und Precision und micro- und macro-averaged Recall und Precision.

Hinweis: Sie müssen dazu Konfusionsmatrizen für jede einzelne Klasse aufstellen.

	vorhergesagt	
real	*A*	*-A*
A	*o*	*b,i,v*
-A		*d,r,q,l,s,u*

	vorhergesagt	
real	*B*	*-B*
B	*d, r*	*q,*
-B	*b, i*	*o,v,l,s,u*

	vorhergesagt	
real	*C*	*-C*
C	*l, s, u*	
-C	*v, q*	*b,o,i,d,r*

rec_A=1/4, rec_B=2/3, rec_C=3/3, macro=23/36, micro=6/10
$prec_A$=1/1, $prec_B$=2/4, $prec_C$=3/5, macro=7/10, micro=6/10

2–d Für einige Dokumente (z.B. 'a', 'e', 'p', 'g', 'm', etc.) gibt es keine eindeutige Klassenzuordnung. In welcher Art von Lernproblem könnten Sie mehrdeutige Zuordnungen lernen?

Multi-Label Classification

Aufgabe 3 17 Punkte (3/4/4/6)

Betrachten Sie folgendes Dokument:

D_1:

Name: Tim Berners-Lee, Telefonnummer: 0615116 1234 (Darmstadt)
Name: Claude Shannon, Telefonnummer: 0615116 4321 (Darmstadt)
Name: Donald Knuth, Telefonnummer: 0615116 1221 (Darmstadt)

Ziel ist es, mit einem Wrapper die Datentupel Namen und Telefonnummern (z.B. "Claude Shannon" und "0615116 4321") zu extrahieren.

3–a Geben Sie für folgende Wrapper-Algorithmen an, ob es möglich ist, aus dem Dokument einen funktionierenden Wrapper zu lernen. Geben Sie ggf. jeweils den Grund an, wieso ein bestimmter Algorithmus keinen passenden Wrapper lernen kann.

- SoftMealy
- LR-Familie
- Whisk

Alle können.

3–b Geben Sie beispielhaft einen funktionierenden LR-Wrapper für dieses Problem an.

z.B. L_1 = "Name: ", R_1 = ", Tel", L_2 = "efonnummer: ", R_2 = " (Darmstadt)"

3–c Würde Ihr unter b) gelernter Wrapper auf folgendem Dokument funktionieren?

D_t:

Telefonnumer und Name:
Name: Nicholas Kushmerick, Telefonnummer: 0615116 2112 (Darmstadt)

Warum (nicht)?

Falls Ihr Wrapper nicht funktioniert, können Sie eine Wrapper-Klasse aus der LR-Familie nennen, mit der Sie auch dieses einzelne Dokument extrahieren könnten?

Nein, wegen early matches. HLRT könnte D_t aber lösen.

3–d Nehmen Sie an, daß zusätzlich zu D_1 noch folgende Dokumente zum Trainieren verwendet werden sollen.

D_2: Personenname: Marvin Lee Minsky, Mobilnummer: 0170 123456 (Darmstadt)

D_3: Name: M. L. Minsky, Tel. +49615116 1221 (Darmstadt)

D_4: Telefonnummer: 0615116 1221 (Darmstadt), Name: Marvin L. Minsky

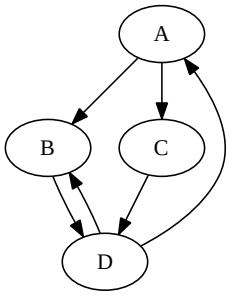
Welche der unter a) genannten Wrapper-Algorithmen sind nach Hinzufügen von D_2, nach D_3 und nach D_4 fähig, einen Wrapper auf dem Trainingsbeispiel zu lernen. Begründen Sie kurz.

- nach D_2:
- nach D_3:
- nach D_4:

D_2: noch alle, D_3: LR nicht mehr, D_4: nur SoftMealy, da Name-Telnr vertauscht

Aufgabe 4 17 Punkte (6/4/4/3)

Gegeben sei folgender Web-Graph, der vier Seiten A, B, C und D, sowie die Verbindungen zwischen den Seiten anzeigt.



4–a Nehmen Sie an, die Seiten haben folgende Hub-Scores $h(p)$ und Authority-Scores $a(p)$:

p	$h(p)$	$a(p)$
A	0.4	0.2
B	0.2	0.4
C	0.1	0.2
D	0.3	0.2

Führen Sie einen Schritt des HITS-Algorithmus durch, d.h. berechnen Sie die Hubs- und Authority-Scores der nächsten Iteration.

p	$h(p)$	$a(p)$
A	*0.6=a(B)+a(C)=0.4+0.2*	*0.3=h(D)*
B	*0.2=a(D)*	*0.7=h(A)+h(D)=0.4+0.3*
C	*0.2=a(D)*	*0.4=h(A)*
D	*0.6=a(A)+a(B)=0.2+0.4*	*0.3=h(B)+h(C)=0.2+0.1*
Normalisierung	*1.6*	*1.7*

4–b Führen Sie die erste Iteration des Page-Rank-Algorithmus aus. Starten Sie mit einem Page-Rank von 1/4 pro Seite und berechnen Sie die nächsten Werte. Sie müssen diese Werte dann *nicht* normalisieren. Der Damping-Faktor d sollte gleich 1/2 sein.

Page Rank: $pr(p) = (1-d) \cdot \frac{1}{N} + d \cdot \sum_{(q,p)\in E} \frac{pr(q)}{o(q)}$

mit $d = \frac{1}{2}$, $pr(q) = \frac{1}{4}$, $N = 4$, $o(p)$*=out degree von* p *und* (q,p)*=Kante von* q *nach* p *(in-Kante)*

p	$pr(p)$
A	$\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2}(\frac{1}{4}/2) = \frac{3}{16}$
B	$\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2}(\frac{1}{4}/2 + \frac{1}{4}/2) = \frac{1}{4}$
C	$\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2}(\frac{1}{4}/2) = \frac{3}{16}$
D	$\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2}(\frac{1}{4} + \frac{1}{4}) = \frac{3}{8}$

4–c Nehmen Sie an, der Random Surfer befindet sich in Knoten A. Wie groß ist die Wahrscheinlichkeit, daß er in Knoten B landet, wenn der Damping Faktor $d = 0$, $d = 1$, oder $d = 1/2$ ist?

$d = 0 : \frac{1}{4}$ die Apriori-Wahrscheinlichkeit, die Wkt. mit einem Random-Jump auf B zu kommen
$d = 1 : \frac{1}{2}$ es gibt zwei ausgehende Klicks, auf die ein Random-Surfer gleichverteilt klicken wird
$d = \frac{1}{2} : \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{8}$ beide Fälle sind gleichverteilt, demnach teilt sich die Wkt. zwischen dem ersten und zweiten Fall

4–d Was gibt der Page-Rank einer Seite an (nur eine Antwort):

☐ die Wahrscheinlichkeit, daß diese Seite von einer Suchmaschine indiziert worden ist

☐ die Wahrscheinlichkeit, daß diese Seite für die gestellte Query relevant ist

☐ die Wahrscheinlichkeit, mit der ein zufälliger Surfer auf dieser Seite landet

☐ die Wahrscheinlichkeit, daß ein zufälliger Surfer durch Klicken von Links auf diese Seite stösst

☒ *die Wahrscheinlichkeit, mit der ein zufälliger Surfer auf dieser Seite landet*

Aufgabe 5 24 Punkte (3/4/5/4/4/4)

5–a Wie können Sie bei einem Memory-Based Collaborative Filtering System berücksichtigen, daß unterschiedliche Benutzer unterschiedlichen Bewertungsskalen verwenden?

Man berechnet die mittlere Abweichung vom Mittelwert jedes Benutzers, diese wird dann zum Mittelwert des aktiven Benutzers addiert.

5–b Geben Sie den Gamma-Code lt. Vorlesung für die Zahl 63 an.

$63 = 2^5 + 31 = 00000\ 1\ 11111$

5–c Der Breakeven-Punkt und der 11-pt-Average sind beides Maße, die versuchen, die Qualität einer Recall- und Precision-Kurve in einem Maß zu fassen. Erklären Sie kurz die Idee hinter diesen beiden Maßen.

Hinweis: Sie müssen *nicht* erklären, wie die Maße berechnet werden, dafür gibt es auch keine Punkte.

Der Break-Even Point nimmt den Diagonalabstand vom idealen Punkt $(1, 1)$ *als Maß dafür, wie nahe die Kurve an diesem Punkt ist.*
Der 11-pt-Average versucht die Fläche unter der Kurve zu approximieren.

5–d k-means Clustering verwendet die Idee, einen Cluster durch einen Prototyp-Vektor zu repräsentieren. Bei welchem in der Vorlesung besprochenen überwachten (supervised) Lernverfahren wird eine ähnliche Idee verwendet?

Der Rocchio Classifier repräsentiert jede Klasse durch einen prototypischen Gewichtsvektor.

5–e Mit welcher Datenrepräsentation können Sie Anfragen der Art "Kunden, die dieses Produkt gekauft haben, kauften auch . . . " effizient beantworten?

Mit einer Item-x-Item Matrix, in der in jedem Eintrag steht, wie oft die jeweilige Kombination gekauft wurde.

5–f Das Token `A` kommt mit einer Häufigkeit von 0.4 in einer Dokumentensammlung vor, das Token `B` mit einer Häufigkeit von 0.2. Nehmen Sie an, dass das Bigramm `A B` mit einer Wahrscheinlichkeit von 0.08 auftritt. Warum würde dieses Bigramm mit einem statistischen Test (z.B. Likelihood Ratio Test) gefiltert werden?

Hinweis: Zur Beantwortung dieser Frage müssen Sie keinen statistischen Test durchführen, Sie erhalten dafür auch keine Punkte.

Der Test testet, ob das Auftreten von B unabhängig vom vorherigen Auftreten von B ist. Das ist hier der Fall, da $0.08 = 0.2 \times 0.4$.