Semestralklausur zur Vorlesung

Web Mining

Prof. J. Fürnkranz
Technische Universität Darmstadt — Sommersemester 2005
Termin: 21. 7. 2005

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

Fachrichtung:

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

Aufgabe 1 19 Punkte (9/2/4/4)

Eine Suchmaschine retourniert für eine Query folgende Liste von 30 Dokumenten:

```
                  1                   2                   3
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
+ + - + + - - + - + - + - - + - + - - - - - - + - - - - - -
```

wobei + für eine relevantes Dokument steht, und - für ein irrelevantes Dokument steht, und die Sortierung des Query-Resultats von links nach rechts erfolgt.

Weiters wissen Sie, daß es insgesamt 1000 Dokumente gibt, wovon 12 relevant und 988 irrelevant sind.

1–a Berechnen Sie für dieses Suchresultat

1. Recall
2. Precision
3. Precision bei einem Recall von 1/2

1–b Berechnen Sie die Genauigkeit (Accuracy) für dieses Suchresultat.

1–c Wie können Sie ohne Probleme 98.8% Genauigkeit für diese Query auf dieser Dokumenten-Sammlung erreichen?

1–d Nennen Sie zwei Maße, die versuchen, Eigenschaften von Recall/Precision-Kurven in einer Maßzahl zusammenzufassen.

Anmerkung: Sie müssen bei den Rechenaufgaben keine Kommazahlen angeben, auch Bruchzahlen sind selbstverständlich zuläßig.

Aufgabe 2 24 Punkte (4/4/4/4/4/4)

Sie erhalten die Aufgabe, einen Web-Katalog zum Thema "Das Schachspiel im 19. Jahrhundert" zu erstellen. Dazu findet sich genügend Material am Web, das Sie "nur" in eine sinnvolle Gliederung in der Form einer Themen-Hierarchie bringen müssen. Sie selbst verstehen von diesem Thema aber gar nichts.

2–a Zuerst beschließen Sie, Material zu sammeln. Sie stellen eine entsprechende Query an Ihre Lieblings-Suchmaschine, stellen jedoch fest, daß die von Ihnen gewählten Such-Terme zu unspezifisch sind. Welche Methode haben wir in der Vorlesung kennen gelernt, die da eventuell Abhilfe schaffen könnte?

2–b Nachdem Sie genügend Material in der Form von Web-Seiten gesammelt haben, wollen Sie diese in eine hierarchische Struktur organisieren. Welche in der Vorlesung besprochenen Methoden würde sich hier anbieten?

2–c Sie wollen nun geeignete kurze Beschreibungen für die Knoten in der hierarchischen Struktur finden. Wie würden Sie vorgehen?

2–d Nachdem Sie nun eine fixe Begriffs-Hierarchie vorgegeben haben, wollen Sie weitere Seiten in diese Struktur einordnen. Wie würden Sie diesen Prozeß zu automatisieren versuchen?

2–e Sie stellen fest, daß die Seiten des gewählten Themenbereichs eine starke Vernetzungsstruktur haben, d.h. daß die Vorgänger und Nachfolger Seiten zu jeder Seite zumeist ebenfalls einem Knoten der Hierarchie zugeordnet werden können. Welches Verfahren haben wir kennen gelernt, das sich dafür anbieten würde?

2–f In Ihren Recherchen sind Sie auf eine Web-Site gestossen, die auf zahlreichen Seiten Tabellen von allen Schach-Turnieren des 19. Jahrhunderts enthält. Sie wollen diese Information verwenden, um einen Namens-Katalog aller aktiven Schachspieler dieser Ära zu erstellen. Welche der in der Vorlesung kennen gelernten Methoden kann Sie bei dieser Aufgabe unterstützen?

Geben Sie kurze (nicht mehr als 1-2 Sätze) Begründungen für Ihre Antworten.

Aufgabe 3 20 Punkte (2/6/3/4/5)

Feature Engineering

3–a Erklären Sie den wesentlichen Unterschied zwischen Supervised und Unsupervised Feature Subset Selection.

3–b Nennen sie ein Beispiel für eine Feature Engineering-Methode, die tendentiell den Recall erhöht, und eine Beispiel für eine Methode, die tendentiell den Recall senkt.

3–c Sie setzen einen Naive Bayes Klassifizierer ein und versuchen ihn mit Hilfe von Feature Subset Selection zu verbesseren. Kann das zu einer Verbesserung der Klassifikationsleistung führen oder nur zu einer Steigerung der Effizienz des Lernens, da man aus weniger Features lernen muß? Begründung?

3–d Warum hat Feature Subset Selection mit einem χ^2-Test ein Problem mit redundanten Features? Welche Methode hat das nicht?

3–e Ein Kollege clustert Dokumente, indem er zuerst Latent Semantic Indexing anwendet, und dann in den 10 signifikantesten Dimensionen des resultierenden Feature-Raums k-means Clustering durchführt. Ist das sinnvoll? Begründen Sie Ihre Antwort.

Aufgabe 4 16 Punkte (5/3/4/2/2)

Lernen

4–a Sie haben ein Klassifikationsproblem mit 10 Klassen, 1000 Dokumenten, die durch 30,000 Features kodiert werden. Wie viele (bedingte oder unbedingte) Wahrscheinlichkeitswerte müssen Sie schätzen, um einen Naive Bayes Klassifizierer auf diesen Daten definieren zu können? Geben Sie eine kurze Begründung Ihrer Antwort.

4–b Sie wenden Ihren Klassifizierer auf neue Texte an. In einem der Texte tritt ein vorher noch nicht dagewesenes Wort auf, d.h. die Auftrittswahrscheinlichkeit für dieses Wort wurde auf den Trainings-Daten mit 0 geschätzt. Welch ein Problem entsteht dadurch? Wie kann man es lösen?

4–c Erklären Sie kurz den Unterschied zwischen multinomialen und binomialen Naive Bayes. Verwenden Sie in der Erklärung keine Formeln.

4–d Erklären Sie den Begriff Semi-Supervised Learning.

4–e Wir haben in der Vorlesung drei Verfahren für Semi-Supervised Learning kennen gelernt. Alle drei stellen die gleiche Anforderung an die zugrunde-liegenden Klassifikations-Algorithmen. Welche ist das?

Aufgabe 5 21 Punkte (6/3/2/3/4/3)

Vernetzung

5–a Berechnen Sie den Hub-Score und den Authority-Score für folgenden Graphen:

```
a -> b
b -> c
a -> c
c -> b
```

5–b Warum wird beim HITS-Algorithmus ein sogenanntes Base-Set konstruiert und bei der Berechnung des PageRank nicht?

5–c Erklären Sie anschaulich (d.h. ohne die Formel hinzuschreiben) die Rolle des "Damping-Factors" d in der Berechnung des Page-Ranks.

5–d In einigen in der Vorlesung besprochenen Untersuchungen hat sich herausgestellt, daß das Betrachten eines Meta-Dokumenten, die durch Zusammenführen aller Vorgänger-Seiten von Dokumenten entstehen, zu keiner Verbesserung der Vorhersagegenauigkeit führt. Können Sie dafür eine Begründung geben?

5–e In der Vorlesung haben wir ein iteratives Verfahren kennen gelernt, das die Information über die Klassenzugehörigkeit der Nachbar-Seiten zur Klassifikation heranzieht. Skizzieren Sie kurz dieses Verfahren.

5–f Würden Sie dieses Verfahren als Supervised, Semi-Supervised, oder Unsupervised einordnen?