Max-Planck-Institut Saarbrücken
AG5: Datenbanken und Informationssysteme
Dozenten: Prof. Dr.-Ing. G. Weikum, Dr. Ralf Schenkel
Betreuer: Dipl.-Inform. Sergej Sizov, Dipl.-Inform. Christian Zimmer

Informationssysteme (SS 05)
Übungsblatt 2

Ausgabe: 26.April 2005

Abgabe: 3.Mai 2005 in der Vorlesung

Aufgabe 2.1: Satz von Bayes

Umfangreiche empirische Untersuchungen haben folgende statistische bzw. probabilistische Zusammenhänge ergeben:

- Von 100 Studenten können 60 mit einem Web-Browser umgehen.
- Von 100 Schülern können 80 mit einem Web-Browser umgehen.
- Unter der restlichen Bevölkerung (also weder Student noch Schüler) können von 100 Personen 20 mit einem Web-Browser umgehen.
- Die Bevölkerungsstruktur ist die folgende:
  - 20 % Studenten
  - 30 % Schüler
  - 50 % sonstige

Wie groß ist die Wahrscheinlichkeit, daß eine Person, die mit einem Web-Browser umgeht (und die Sie z.B. im Internet-Cafe treffen)

a) Student ist?

b) Schüler ist?

Aufgabe 2.2: Satz von Bayes

Ein Statistiker untersucht den Zusammenhang zwischen Personen mit Herzproblemen (H), mit der Gewohnheit, lange zu arbeiten (W), und mit regelmäßiger sportlicher Betätigung (S). Jede dieser Eigenschaften einer Person wird als eine binäre Zufallsvariable betrachtet. Die Bevölkerung wird also eingeteilt in Leute mit Herzproblemen ($H = wahr$) oder ohne Herzprobleme ($H = falsch$), in Leute mit der Gewohnheit, lange zu arbeiten ($W = wahr$), oder mit eher regelmäßigen Arbeitszeiten ($W = falsch$), und in Leute, die regelmäßig Sport betreiben ($S = wahr$) oder eine unzureichende körperliche Betätigung haben ($S = falsch$). Der Statistiker findet die folgenden Zusammenhänge:

- 20% aller Personen haben Herzprobleme.

- Unter allen Personen mit Herzproblemen arbeiten 60% lange.
- Unter allen Personen ohne Herzprobleme arbeiten 50% lange.
- Unter allen Personen mit Herzproblemen betreiben 30% regelmäßig Sport.
- Unter allen Personen ohne Herzprobleme treiben 60% regelmäßig Sport.

Zusätzlich nimmt der Statistiker an:

- Unter allen Personen mit Herzproblemen sind die Arbeitsgewohnheiten und die sportlichen Aktivitäten unabhängig voneinander verteilt (d.h. W ist konditional unabhngig von S bei gegebenem H, formal $P[W|S \wedge H] = P[W|H]$ und S ist konditional unabhngig von W bei gegebenem H, formal $P[S|W \wedge H] = P[S|H]$).
- Unter allen Personen sind die Arbeitsgewohnheiten und die sportlichen Aktivitäten unabhängig voneinander verteilt.

Wie hoch ist die Wahrscheinlichkeit, dass jemand Herzprobleme hat, vorausgesetzt

a) er/sie betreibt regelmäßig Sport,
b) er/sie arbeitet lange und betreibt regelmäßig Sport?

Aufgabe 2.3: Rocchio- und kNN-Klassifikation

Wenden Sie die *kNN* Methode und die *Rocchio* Methode zur automatischen Klassifizierung von Dokumenten auf das Testdokument $d_7 = (0\ 0\ 1\ 2\ 0\ 0\ 3\ 0)$ an. Verwenden Sie die die Trainingsdokumente $d_1, \ldots, d_6$ aus der Vorlesung jeweils mit 2 Dokumenten aus den Kategorien *Algebra*, *Calculus*, and *Stochastics*.

Doc.	**Term Vector**								**Category**
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	
d_1	3	2	0	0	0	0	0	1	c_1 (Algebra)
d_2	1	2	3	0	0	0	0	0	c_1 (Algebra)
d_3	0	0	0	3	3	0	0	0	c_2 (Calculus)
d_4	0	0	1	2	2	0	1	0	c_2 (Calculus)
d_5	0	0	0	1	1	2	2	0	c_3 (Stochastics)
d_6	1	0	1	0	0	0	2	2	c_3 (Stochastics)
d_7	0	0	1	2	0	0	3	0	[?]

Zur Vereinfachung können Sie die TF-Vektoren (statt den normalerweise gewichteten TF-IDF-Vektoren) benutzen und Sie können das einfache Skalarprodukt statt dem Kosinusmaß verwenden.

Aufgabe 2.4: Naive-Bayes-Klassifikator

Die Trainingsmenge aus Aufgabe 2.1 mit 6 Dokumenten wird um 3 weitere Dokumente d_7, d_8 und d_9 verschiedener Länge erweitert. Von Hand teilen wir Dokument d_7 der Kategorie c_1 (Algebra) zu; Dokumente d_8 und d_9 der Kategorie c_3 (Stochastics). In d_7 bis d_9 haben wir zusätzlich zu den 8 Termen noch die Terme *Eigenvalue*, *Differential Equation*, *Laplace Transform*, und *Normal Distribution* (alle werden als einzelne Worte betrachtet), deren Häufigkeiten mit $f_{9,k}$ bis $f_{12,k}$ bezeichnet werden. Insgesamt haben wir die folgende Trainingsmenge:

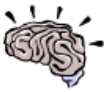
	$\mathbf{f}_1$	$\mathbf{f}_2$	$\mathbf{f}_3$	$\mathbf{f}_4$	$\mathbf{f}_5$	$\mathbf{f}_6$	$\mathbf{f}_7$	$\mathbf{f}_8$	$\mathbf{f}_9$	$\mathbf{f}_{10}$	$\mathbf{f}_{11}$	$\mathbf{f}_{12}$
$\mathbf{d}_1$	3	2	0	0	0	0	0	1	0	0	0	0
$\mathbf{d}_2$	1	2	3	0	0	0	0	0	0	0	0	0
$\mathbf{d}_3$	0	0	0	3	3	0	0	0	0	0	0	0
$\mathbf{d}_4$	0	0	1	2	2	0	1	0	0	0	0	0
$\mathbf{d}_5$	0	0	0	1	1	2	2	0	0	0	0	0
$\mathbf{d}_6$	1	0	1	0	0	0	2	2	0	0	0	0
$\mathbf{d}_7$	0	1	1	0	0	0	0	0	2	0	0	0
$\mathbf{d}_8$	0	0	1	0	0	1	1	0	1	1	1	2
$\mathbf{d}_9$	0	0	0	1	1	2	2	0	0	1	2	1

a) Klassifizieren Sie Dokument d_{10} = (0 0 1 1 1 1 3 1 1 1 3 3) bei dieser Dokumentenmenge mit der Naive Bayes Methode unter der Annahme, dass die Termhäufigkeiten in den Dokumenten multinomial verteilt sind.

b) Wenden Sie die Naive Bayes Methode auf d_{10} an mit binarren (und unabhängigen) Features.

Aufgabe 2.5: Klassifikator mit Laplace-Smoothing

Wenden Sie Laplace-Smoothing auf die Parameterschätzung für das Vorlesungsbeispiel zur multinomialen Naive-Bayes-Klassifikation aus Kapitel 3 an. Berechnen Sie das Klassifikationsergebnis auf der Basis der durch Smoothing verbesserten Parameterschätzung.



Aufgabe 2.6: Naive-Bayes-Klassifikator mit binären Features

Bestimmen Sie die einfachste mögliche Entscheidungsformel für den Spezialfall der binären Klassifikation, z.B die Entscheidung, ob ein neues Dokument einer Kategorie zugeordnet werden sollte oder nicht. Benutzen Sie hierzu die Naive Bayes Methode mit unabhängigen binären Features.

Betrachten Sie den Logarithmus von *P[d gehört zu* C_k *| d hat Feature-Vektor X].*