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

INFORMATIK

Informationssysteme (SS 04)
Beispiellösungen zu Übungsblatt 2

11.Mai 2004

Aufgabe 2.1: Anwendung des Satzes von Bayes

Gegeben:

$$P\,[\text{Web-Browser}|\text{Student}] = 0,6 \tag{1}$$
$$P\,[\text{Web-Browser}|\text{Schüler}] = 0,8 \tag{2}$$
$$P\,[\text{Web-Browser}|\text{Sonst}] = 0,2 \tag{3}$$
$$P\,[\text{Student}] = 0,2 \tag{4}$$
$$P\,[\text{Schüler}] = 0,3 \tag{5}$$
$$P\,[\text{Sonst}] = 0,5 \tag{6}$$

Gesucht: $P\,[\text{Student}|\text{Web-Browser}]$ und $P\,[\text{Schüler}|\text{Web-Browser}]$

$$\begin{aligned} P\,[\text{Web-Browser}] &= \sum_{i\in\{Student,Schüler,Sonst\}} P\,[\text{Web-Browser}|i]\cdot P\,[i] \\ &= 0,6\cdot 0,2 + 0,8\cdot 0,3 + 0,2\cdot 0,5 \\ &= 0,12 + 0,24 + 0,1 \\ &= 0,46 \end{aligned} \tag{7}$$

$$\begin{aligned} \mathbf{P\,[Student|Web\text{-}Browser]} &= \frac{P\,[\text{Web-Browser}|\text{Student}]\cdot P\,[\text{Student}]}{P\,[\text{Web-Browser}]} \\ &= \frac{0,6\cdot 0,2}{0,46} \\ &= \frac{\mathbf{6}}{\mathbf{23}} \\ &\approx \mathbf{0,26} \end{aligned} \tag{8}$$

$$\begin{aligned} \mathbf{P\,[Schüler|Web\text{-}Browser]} &= \frac{P\,[\text{Web-Browser}|\text{Schüler}]\cdot P\,[\text{Schüler}]}{P\,[\text{Web-Browser}]} \\ &= \frac{0,8\cdot 0,3}{0,46} \\ &= \frac{\mathbf{12}}{\mathbf{23}} \\ &\approx \mathbf{0,52} \end{aligned} \tag{9}$$

Aufgabe 2.2: Anwendung des Satzes von Bayes

Gegeben sind die folgende Wahrscheinlichkeiten:

$$
\begin{aligned}
P[H] &= 0.2 && (10)\\
P[W|H] &= 0.6 && (11)\\
P[W|\overline{H}] &= 0.5 && (12)\\
P[S|H] &= 0.3 && (13)\\
P[S|\overline{H}] &= 0.6 && (14)\\
& && (15)
\end{aligned}
$$

Weiter wissen wir, dass W und S unabhängig verteilt sind und W ist bedingt unabhängig von S, wenn H gegeben ist. Um die fehlenden Beziehungen zu finden benötigen wir den Satz von Bayes und den Satz der totalen Wahrscheinlichkeit.

a)

$$
\begin{aligned}
P[H|S] &= \frac{P[S|H] \cdot P[H]}{P[S]} && (16)\\
&= \frac{P[S|H] \cdot P[H]}{P[S|H] \cdot P[H] + P[S|\overline{H}] \cdot P[\overline{H}]} && (17)\\
&= \frac{0.3 \cdot 0.2}{0.3 \cdot 0.2 + 0.6 \cdot 0.8} && (18)\\
&= \frac{0.06}{0.54} && (19)\\
&\approx 0.11 && (20)
\end{aligned}
$$

b)

$$
\begin{aligned}
P[H|W \wedge S] &= \frac{P[W \wedge S|H] \cdot P[H]}{P[W \wedge S]} && (21)\\
&= \frac{P[W \wedge S|H] \cdot P[H]}{P[W] \cdot P[S]} && (22)\\
&= \frac{(P[W|H] \cdot P[S|H]) \cdot P[H]}{\big(P[W|H] \cdot P[H] + P[W|\overline{H}] \cdot P[\overline{H}]\big) \cdot \big(P[S|H] \cdot P[H] + P[S|\overline{H}] \cdot P[\overline{H}]\big)} && (23)\\
&= \frac{0.6 \cdot 0.3 \cdot 0.2}{(0.6 \cdot 0.2 + 0.5 \cdot 0.8)(0.3 \cdot 0.2 + 0.6 \cdot 0.8)} && (24)\\
&= \frac{0.036}{0.52 \cdot 0.54} && (25)\\
&\approx 0.13 && (26)\\
& && (27)
\end{aligned}
$$

Aufgabe 2.3: Rocchio- und kNN-Klassifikator

Gegeben: 6 Trainingsdokumente d_1 bis d_6 und ein Testdokument d_7:

Doc.	**Term Vektor**								**Kategorie**
	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	[?]

Mit den folgenden Features:

f_1: group
f_2: homomorphism
f_3: vector
f_4: integral
f_5: limit
f_6: variance
f_7: probability
f_8: dice

Gesucht: Eine automatische Klassifizierung von d_7 nach der *kNN* Methode und der *Rocchio* Methode unter Benutzung der TF-Gewichte und dem Skalarprodukt als Ähnlichkeitsmaß.

kNN Methode:

$$sim\left(\vec{d}_7, \vec{d}_1\right) = 0 \tag{28}$$
$$sim\left(\vec{d}_7, \vec{d}_2\right) = 3 \tag{29}$$
$$sim\left(\vec{d}_7, \vec{d}_3\right) = 6 \tag{30}$$
$$sim\left(\vec{d}_7, \vec{d}_4\right) = 8 \tag{31}$$
$$sim\left(\vec{d}_7, \vec{d}_5\right) = 8 \tag{32}$$
$$sim\left(\vec{d}_7, \vec{d}_6\right) = 7 \tag{33}$$
$$\tag{34}$$

Wähle $k = 4$. Es folgt:

$$kNN_{k=4}\left(\vec{d}_7\right) = \left\{\vec{d}_3, \vec{d}_4, \vec{d}_5, \vec{d}_6\right\} \tag{35}$$

Nun berechnen wir die Gewichte $f\left(\vec{d}_7, \vec{c}\right)$ für jede der 3 Kategorien c_1, c_2 und c_3 indem die folgende Formel benutz wird:

$$\begin{aligned} f\left(\vec{d}_7, \vec{c}\right) &= \sum_{\vec{v} \in kNN_{k=4}\left(\vec{d}_7\right)} sim\left(\vec{d}_7, \vec{v}\right) \cdot \begin{cases} 1 & \text{falls } \vec{v} \in c \\ 0 & \text{sonst} \end{cases} \\ &= \sum_{\vec{v} \in \left(kNN_{k=4}\left(\vec{d}_7\right) \cap c\right)} sim\left(\vec{d}_7, \vec{v}\right) \end{aligned} \tag{36}$$

Nun gilt:

$$f\left(\vec{d}_7, \vec{c}_1\right) = 0 \tag{37}$$

$$\begin{aligned} f\left(\vec{d}_7, \vec{c}_2\right) &= sim\left(\vec{d}_7, \vec{d}_3\right) + sim\left(\vec{d}_7, \vec{d}_4\right) = 14 \\ f\left(\vec{d}_7, \vec{c}_3\right) &= sim\left(\vec{d}_7, \vec{d}_5\right) + sim\left(\vec{d}_7, \vec{d}_6\right) = 15 \end{aligned} \tag{38}$$

Da $f\left(\vec{d}_7, \vec{c}_3\right) > f\left(\vec{d}_7, \vec{c}_2\right) > f\left(\vec{d}_7, \vec{c}_1\right)$ Dokument d_7 gehört zu Kategorie c_3 (dies gilt $\forall k \in \{3, ..., 6\}$). Für $k = 2$ könnten wir auch d_7 zu Kategorie c_2 oder c_3 zuordnen.

Rocchio Methode: Wähle $\alpha = 16$ und $\beta = 4$. Berechne die Prototyp-Vektoren $\vec{c}_1$, $\vec{c}_2$ und $\vec{c}_3$ der 3 Kategorien c_1, c_2 und c_3. Der Prototyp-Vektor von c_i ist:

$$\vec{c}_i = \frac{\alpha}{|c_i|} \cdot \sum_{\vec{d} \in c_i} \frac{\vec{d}}{\left\|\vec{d}\right\|} - \frac{\beta}{|D \setminus c_i|} \cdot \sum_{\vec{d} \in D \setminus c_i} \frac{\vec{d}}{\left\|\vec{d}\right\|}. \tag{39}$$

Daneben gilt:

$$\left\|\vec{d}_1\right\| = \sqrt{9+4+1} = \sqrt{14} \tag{40}$$
$$\left\|\vec{d}_2\right\| = \sqrt{1+4+9} = \sqrt{14} \tag{41}$$
$$\left\|\vec{d}_3\right\| = \sqrt{9+9} = \sqrt{18} \tag{42}$$
$$\left\|\vec{d}_4\right\| = \sqrt{1+4+4+1} = \sqrt{10} \tag{43}$$
$$\left\|\vec{d}_5\right\| = \sqrt{1+1+4+4} = \sqrt{10} \tag{44}$$
$$\left\|\vec{d}_6\right\| = \sqrt{1+1+4+4} = \sqrt{10} \tag{45}$$

und somit

$$\begin{aligned} \vec{c}_1 &= \frac{16}{2} \cdot \left(\frac{d_1}{\sqrt{14}} + \frac{d_2}{\sqrt{14}}\right) - \frac{4}{4} \cdot \left(\frac{d_3}{\sqrt{18}} + \frac{d_4}{\sqrt{10}} + \frac{d_5}{\sqrt{10}} + \frac{d_6}{\sqrt{10}}\right) \\ &= \frac{8}{\sqrt{14}} \begin{pmatrix} 4 \\ 4 \\ 3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} - \frac{1}{\sqrt{18}} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 3 \\ 0 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ 0 \\ 2 \\ 3 \\ 3 \\ 2 \\ 5 \\ 2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{32}{\sqrt{14}} & & & - & \frac{1}{\sqrt{10}} \\ \frac{32}{\sqrt{14}} & & & & \\ \frac{24}{\sqrt{14}} & & & - & \frac{2}{\sqrt{10}} \\ & - & \frac{3}{\sqrt{18}} & - & \frac{3}{\sqrt{10}} \\ & - & \frac{3}{\sqrt{18}} & - & \frac{3}{\sqrt{10}} \\ & & & - & \frac{2}{\sqrt{10}} \\ & & & - & \frac{5}{\sqrt{10}} \\ \frac{8}{\sqrt{14}} & & & - & \frac{2}{\sqrt{10}} \end{pmatrix} \end{aligned}$$

$$\approx \begin{pmatrix} 8,236 \\ 8,552 \\ 5,782 \\ -1,656 \\ -1,656 \\ -0,632 \\ -1,581 \\ 1,506 \end{pmatrix} \tag{46}$$

$$\begin{aligned}
\vec{c}_2 &= \frac{16}{2} \cdot \left(\frac{d_3}{\sqrt{18}} + \frac{d_4}{\sqrt{10}} \right) - \frac{4}{4} \cdot \left(\frac{d_1}{\sqrt{14}} + \frac{d_2}{\sqrt{14}} + \frac{d_5}{\sqrt{10}} + \frac{d_6}{\sqrt{10}} \right) \\
&= \frac{8}{\sqrt{18}} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 3 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \frac{8}{\sqrt{10}} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \\ 2 \\ 0 \\ 1 \\ 0 \end{pmatrix} - \frac{1}{\sqrt{14}} \begin{pmatrix} 4 \\ 4 \\ 3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} - \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 2 \\ 4 \\ 2 \end{pmatrix} \\
&= \frac{8}{\sqrt{18}} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 3 \\ 0 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ 0 \\ 1 \\ -15 \\ -15 \\ 1 \\ 2 \\ 42 \end{pmatrix} - \frac{1}{\sqrt{14}} \begin{pmatrix} 4 \\ 4 \\ 3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} & - \frac{4}{\sqrt{14}} & - \frac{1}{\sqrt{10}} \\ & - \frac{4}{\sqrt{14}} & \\ & - \frac{3}{\sqrt{14}} & + \frac{7}{\sqrt{10}} \\ \frac{24}{\sqrt{18}} & & + \frac{15}{\sqrt{10}} \\ \frac{24}{\sqrt{18}} & & + \frac{15}{\sqrt{10}} \\ & & - \frac{2}{\sqrt{10}} \\ & & \frac{4}{\sqrt{10}} \\ & - \frac{1}{\sqrt{14}} & - \frac{2}{\sqrt{10}} \end{pmatrix} \\
&\approx \begin{pmatrix} -1,385 \\ -1,069 \\ 1,412 \\ 10,400 \\ 10,400 \\ -0,632 \\ 1,265 \\ -0,900 \end{pmatrix}
\end{aligned} \tag{47}$$

$$\begin{aligned}
\vec{c}_3 &= \frac{16}{2} \cdot \left(\frac{d_5}{\sqrt{10}} + \frac{d_6}{\sqrt{10}} \right) - \frac{4}{4} \cdot \left(\frac{d_1}{\sqrt{14}} + \frac{d_2}{\sqrt{14}} + \frac{d_3}{\sqrt{18}} + \frac{d_4}{\sqrt{10}} \right) \\
&= \frac{8}{\sqrt{10}} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 2 \\ 4 \\ 2 \end{pmatrix} - \frac{1}{\sqrt{10}} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \\ 2 \\ 0 \\ 1 \\ 0 \end{pmatrix} - \frac{1}{\sqrt{14}} \begin{pmatrix} 4 \\ 4 \\ 3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} - \frac{1}{\sqrt{18}} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 3 \\ 0 \\ 0 \\ 0 \end{pmatrix}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{10}} \begin{pmatrix} 8 \\ 0 \\ 7 \\ 6 \\ 6 \\ 16 \\ 31 \\ 16 \end{pmatrix} - \frac{1}{\sqrt{14}} \begin{pmatrix} 4 \\ 4 \\ 3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} - \frac{1}{\sqrt{18}} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 3 \\ 0 \\ 0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} \frac{8}{\sqrt{10}} & - & \frac{4}{\sqrt{14}} & & \\ & - & \frac{4}{\sqrt{14}} & & \\ \frac{7}{\sqrt{10}} & - & \frac{3}{\sqrt{14}} & & \\ \frac{6}{\sqrt{10}} & & & - & \frac{3}{\sqrt{18}} \\ \frac{6}{\sqrt{10}} & & & - & \frac{3}{\sqrt{18}} \\ \frac{16}{\sqrt{10}} & & & & \\ \frac{31}{\sqrt{10}} & & & & \\ \frac{16}{\sqrt{10}} & - & \frac{1}{\sqrt{14}} & & \end{pmatrix} \\
&\approx \begin{pmatrix} 1,461 \\ -1,069 \\ 1,412 \\ 1,190 \\ 1,190 \\ 5,060 \\ 9,803 \\ 4,792 \end{pmatrix} \qquad (48)
\end{aligned}
$$

Wir erhalten die folgenden Ähnlichkeiten von d_7 zu den Prototyp-Vektoren:

$$
\begin{aligned}
sim\left(\vec{d}_7, \vec{c}_1\right) &= 1 \cdot \left(\frac{24}{\sqrt{14}} - \frac{2}{\sqrt{10}}\right) + 2 \cdot \left(\frac{3}{\sqrt{18}} - \frac{3}{\sqrt{10}}\right) + 3 \cdot \left(-\frac{5}{\sqrt{10}}\right) \\
&= \frac{24}{\sqrt{14}} - \frac{2}{\sqrt{10}} - \frac{6}{\sqrt{18}} - \frac{6}{\sqrt{10}} - \frac{15}{\sqrt{10}} \\
&= \frac{24}{\sqrt{14}} - \frac{23}{\sqrt{10}} - \frac{6}{\sqrt{18}} \\
&\approx -2,273 \qquad (49) \\
sim\left(\vec{d}_7, \vec{c}_2\right) &= 1 \cdot \left(\frac{7}{\sqrt{10}} - \frac{3}{\sqrt{14}}\right) + 2 \cdot \left(\frac{24}{\sqrt{18}} + \frac{15}{\sqrt{10}}\right) + 3 \cdot \frac{4}{\sqrt{10}} \\
&= \frac{7}{\sqrt{10}} - \frac{3}{\sqrt{14}} + \frac{48}{\sqrt{18}} + \frac{30}{\sqrt{10}} + \frac{12}{\sqrt{10}} \\
&= \frac{49}{\sqrt{10}} - \frac{3}{\sqrt{14}} + \frac{48}{\sqrt{18}} \\
&\approx 26,007 \qquad (50) \\
sim\left(\vec{d}_7, \vec{c}_3\right) &= 1 \cdot \left(\frac{7}{\sqrt{10}} - \frac{3}{\sqrt{14}}\right) + 2 \cdot \left(\frac{6}{\sqrt{10}} - \frac{3}{\sqrt{18}}\right) + 3 \cdot \frac{31}{\sqrt{10}} \\
&= \frac{7}{\sqrt{10}} - \frac{3}{\sqrt{14}} + \frac{12}{\sqrt{10}} - \frac{6}{\sqrt{18}} + \frac{93}{\sqrt{10}} \\
&= \frac{112}{\sqrt{10}} - \frac{3}{\sqrt{14}} - \frac{6}{\sqrt{18}} \\
&\approx 33,202 \qquad (51)
\end{aligned}
$$

Da $sim\left(\vec{d}_7, \vec{c}_3\right) > sim\left(\vec{d}_7, \vec{c}_2\right) > sim\left(\vec{d}_7, \vec{c}_1\right)$, Dokument d_7 gehört zu c_3 - auch nach der Rocchio Methode. Glücklicherweise bestätigt dies unser erstes Ergebnis.

Aufgabe 2.4: Naive-Bayes-Klassifikator

Gegeben: 9 Trainingsdokumente d_1 bis d_9 und ein Testdokument d_{10}:

Doc.	**Term Vector**												**Category**
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	
d_1	3	2	0	0	0	0	0	1	0	0	0	0	c_1 (Algebra)
d_2	1	2	3	0	0	0	0	0	0	0	0	0	c_1 (Algebra)
d_3	0	0	0	3	3	0	0	0	0	0	0	0	c_2 (Calculus)
d_4	0	0	1	2	2	0	1	0	0	0	0	0	c_2 (Calculus)
d_5	0	0	0	1	1	2	2	0	0	0	0	0	c_3 (Stochastics)
d_6	1	0	1	0	0	0	2	2	0	0	0	0	c_3 (Stochastics)
d_7	0	1	1	0	0	0	0	0	2	0	0	0	c_1 (Algebra)
d_8	0	0	1	0	0	1	1	0	1	1	1	2	c_3 (Stochastics)
d_9	0	0	0	1	1	2	2	0	0	1	2	1	c_3 (Stochastics)
d_{10}	0	0	1	1	1	1	3	1	1	1	3	3	[**?**]

Mit den folgdenen Features:

f_1: group
f_2: homomorphism
f_3: vector
f_4: integral
f_5: limit
f_6: variance
f_7: probability
f_8: dice
f_9: eigenvalue
f_{10}: differential equation
f_{11}: laplace transform
f_{12}: normal distribution

Wir erhalten die folgenden Kategorien und deren Wahrscheinlichkeiten:

$$c_1 = \{d_1, d_2, d_7\} \tag{52}$$
$$c_2 = \{d_3, d_4\} \tag{53}$$
$$c_3 = \{d_5, d_6, d_8, d_9\} \tag{54}$$
$$p_1 = \frac{3}{9} \tag{55}$$
$$p_2 = \frac{2}{9} \tag{56}$$
$$p_3 = \frac{4}{9} \tag{57}$$

Gesucht: Automatische Klassifikation von d_{10}

1. Naives Bayes Methode mit Bag-of-Words Model Berechne alle

$$\begin{aligned} p_{i,k} &= P\left[\text{Term } i \text{ erscheint in } d | d \in c_k\right] \\ &= \frac{\text{Anzahl Vorkommen von Term } i \text{ in Dokumenten der Kategorie } c_k}{\text{Anzahl Terme in Dokumenten der Kategorie } c_k} \\ &= \frac{\sum_{d \in c_k} d_i}{\sum_{d \in c_k} \sum_{t=1}^{12} d_t}: \end{aligned} \tag{58}$$

$p_{i,k}$	$k=1$	$k=2$	$k=3$
$i=1$	$\frac{4}{16}$	0	$\frac{1}{30}$
$i=2$	$\frac{5}{16}$	0	0
$i=3$	$\frac{4}{16}$	$\frac{1}{12}$	$\frac{2}{30}$
$i=4$	0	$\frac{5}{12}$	$\frac{2}{30}$
$i=5$	0	$\frac{5}{12}$	$\frac{2}{30}$
$i=6$	0	0	$\frac{5}{30}$
$i=7$	0	$\frac{1}{12}$	$\frac{7}{30}$
$i=8$	$\frac{1}{16}$	0	$\frac{2}{30}$
$i=9$	$\frac{2}{16}$	0	$\frac{1}{30}$
$i=10$	0	0	$\frac{2}{30}$
$i=11$	0	0	$\frac{3}{30}$
$i=12$	0	0	$\frac{3}{30}$

Berechne $P\left[d_{10}|\, d_{10} \in c_k\right] \cdot P\left[d_{10} \in c_k\right]$ for $k \in \{1,2,3\}$:

$$
\begin{aligned}
P\left[d_{10}|\, d_{10} \in c_k\right] \cdot P\left[d_{10} \in c_k\right] &= \binom{\sum_{i=1}^{12} d_{10,i}}{d_{10,1} \ldots d_{10,12}} \cdot p_{1,k}^{d_{10,1}} \cdot \ldots \cdot p_{12,k}^{d_{10,12}} \cdot p_k \\
&= \binom{16}{0\,0\,1\,1\,1\,1\,3\,1\,1\,1\,3\,3} \cdot p_{1,k}^0 \cdot p_{2,k}^0 \cdot p_{3,k}^1 \cdot p_{4,k}^1 \cdot \\
&\quad p_{5,k}^1 \cdot p_{6,k}^1 \cdot p_{7,k}^3 \cdot p_{8,k}^1 \cdot p_{9,k}^1 \cdot p_{10,k}^1 \cdot p_{11,k}^3 \cdot p_{12,k}^3 \cdot p_k \\
&= \binom{16}{3\,3\,3} \cdot 1 \cdot 1 \cdot p_{3,k} \cdot p_{4,k} \cdot p_{5,k} \cdot p_{6,k} \cdot \\
&\quad p_{7,k}^3 \cdot p_{8,k} \cdot p_{9,k} \cdot p_{10,k} \cdot p_{11,k}^3 \cdot p_{12,k}^3 \cdot p_k \\
&= \frac{16!}{3! \cdot 3! \cdot 3!} \cdot p_{3,k} \cdot p_{4,k} \cdot p_{5,k} \cdot p_{6,k} \cdot \\
&\quad p_{7,k}^3 \cdot p_{8,k} \cdot p_{9,k} \cdot p_{10,k} \cdot p_{11,k}^3 \cdot p_{12,k}^3 \cdot p_k \\
&= \frac{16!}{6^3} \cdot p_{3,k} \cdot p_{4,k} \cdot p_{5,k} \cdot p_{6,k} \cdot \\
&\quad p_{7,k}^3 \cdot p_{8,k} \cdot p_{9,k} \cdot p_{10,k} \cdot p_{11,k}^3 \cdot p_{12,k}^3 \cdot p_k \qquad (59)
\end{aligned}
$$

$$
\begin{aligned}
P\left[d_{10}|\, d_{10} \in c_1\right] \cdot P\left[d_{10} \in c_1\right] &= \frac{16!}{6^3} \cdot \frac{4}{16} \cdot 0 \cdot 0 \cdot 0 \cdot 0^3 \cdot \frac{1}{16} \cdot \frac{2}{16} \cdot 0 \cdot 0^3 \cdot 0^3 \cdot \frac{3}{9} \\
&= 0 \qquad (60)
\end{aligned}
$$

$$
\begin{aligned}
P\left[d_{10}|\, d_{10} \in c_2\right] \cdot P\left[d_{10} \in c_2\right] &= \frac{16!}{6^3} \cdot \frac{1}{12} \cdot \frac{5}{12} \cdot \frac{5}{12} \cdot 0 \cdot \left(\frac{1}{12}\right)^3 \cdot 0 \cdot 0 \cdot 0 \cdot 0^3 \cdot 0^3 \cdot \frac{2}{9} \\
&= 0 \qquad (61)
\end{aligned}
$$

$$
\begin{aligned}
P\left[d_{10}|\, d_{10} \in c_3\right] \cdot P\left[d_{10} \in c_3\right] &= \frac{16!}{6^3} \cdot \frac{2}{30} \cdot \frac{2}{30} \cdot \frac{2}{30} \cdot \frac{5}{30} \cdot \frac{7^3}{30^3} \cdot \frac{2}{30} \cdot \frac{1}{30} \cdot \frac{2}{30} \cdot \frac{3^3}{30^3} \cdot \frac{3^3}{30^3} \cdot \frac{4}{9} \\
&= \frac{2^{22} \cdot 3^{12} \cdot 5^4 \cdot 7^5 \cdot 11 \cdot 13}{2^{19} \cdot 3^{21} \cdot 5^{16}} \\
&= \frac{2^3 \cdot 7^5 \cdot 11 \cdot 13}{3^9 \cdot 5^{12}} \\
&= \frac{19227208}{4805419921875} \\
&\approx 4{,}00 \cdot 10^{-6} \\
&> 0 \qquad (62)
\end{aligned}
$$

Und daher gehört Dokument d_{10} zu Kategorie c_3.

2. Naives Bayes Method mit binären Features Da wir nun nur den Fall betrachten, dass ein Term in einem Dokument vorkommt oder nicht (und nicht die Häufigkeit in einerm Dokument), müssen wir die Wahrscheinlichkeiten $p_{i,k}$ neu berechnen:

Berechne alle:

$$\begin{aligned} p_{i,k} &= P\left[\text{Term } i \text{ erscheint in } d | d \in c_k\right] \\ &= \frac{\text{Anzahl der Dokumente der Kategorie } c_k\text{, die das Feature } X_i \text{ enthalten}}{\text{Anzahl aller Features in Dokumenten der Kategorie } c_k} \\ &= \frac{\sum_{d \in c_k} X_i}{\sum_{d \in c_k} \sum_{t=1}^{12} X_t} \end{aligned} \tag{63}$$

$p_{i,k}$	$k=1$	$k=2$	$k=3$
$i=1$	$\frac{2}{9}$	0	$\frac{1}{22}$
$i=2$	$\frac{3}{9}$	0	0
$i=3$	$\frac{2}{9}$	$\frac{1}{6}$	$\frac{2}{22}$
$i=4$	0	$\frac{2}{6}$	$\frac{2}{22}$
$i=5$	0	$\frac{2}{6}$	$\frac{2}{22}$
$i=6$	0	0	$\frac{3}{22}$
$i=7$	0	$\frac{1}{6}$	$\frac{4}{22}$
$i=8$	$\frac{1}{9}$	0	$\frac{1}{22}$
$i=9$	$\frac{1}{9}$	0	$\frac{1}{22}$
$i=10$	0	0	$\frac{2}{22}$
$i=11$	0	0	$\frac{2}{22}$
$i=12$	0	0	$\frac{2}{22}$

Berechne $\log P\left[c_k | d_{10}\right]$ für $k \in \{1,2,3\}$: Laut Vorlesung gilt, dass $\log P\left[c_k | d_{10}\right]$ ist ungefähr:

$$\log P\left[c_k | d_{10}\right] \approx \sum_{i=1}^{12} X_i \cdot \log \frac{p_{i,k}}{1-p_{i,k}} + \sum_{i=1}^{12} \cdot \log\left(1-p_{i,k}\right) + \log p_k \tag{64}$$

Wir beginnen mit $k=1$:

$$\begin{aligned} \log P\left[c_1 | d_{10}\right] &\approx \left(0+0+1 \cdot log\left(\frac{2/9}{1-2/9}\right) + 1 \cdot \underbrace{log\left(\frac{0}{1-0}\right)}_{-\infty} \ldots \right) \\ &= -\infty \end{aligned} \tag{65}$$

Da die Wahrscheinlichkeit f_4, dass f_4 erscheint in c_1 gerade 0 ist, *darf* dieses Feature nicht in unserem Testdokument vorkommen, wenn es in c_1 klassifiziert werden soll. Ohne Smoothing des Sumanden des Logarithmus ist es $-\infty$, und so ist die Summe selbst $-\infty$ ist ohne weitere Berechnungen. Daher gehört d_{10} nicht zu c_1.

Für $k=2$ haben wir:

$$\begin{aligned} \log P\left[c_2 | d_{10}\right] &\approx \left(0+0+1 \cdot log\left(\frac{1/6}{1-1/6}\right) + 1 \cdot log\left(\frac{2/6}{1-2/6}\right) 1 \cdot log\left(\frac{2/6}{1-2/6}\right) + 1 \cdot \underbrace{log\left(\frac{0}{1-0}\right)}_{-\infty} \ldots \right) \\ &= -\infty \end{aligned} \tag{66}$$

Wieder gilt die Argumentation wie zuvor für c_2 - daher gehört d_{10} nicht zu c_2.

Für $k=3$ gilt:

$$\log P\left[c_3 | d_{10}\right] \approx 0+0+2 \cdot log\left(\frac{1/22}{1-1/22}\right) + 6 \cdot log\left(\frac{2/22}{1-2/22}\right) + 1 \cdot log\left(\frac{3/22}{1-3/22}\right) + 1 \cdot log\left(\frac{4/22}{1-4/22}\right)$$

$$
\begin{aligned}
&+ \quad 3 \cdot log\,(1-1/22) + 6 \cdot log\,(1-2/22) + 1 \cdot log\,(1-3/22) + 1 \cdot log\,(1-4/22) \\
&+ \quad \frac{4}{9} \\
&= \quad 2 \cdot \underbrace{\log 1}_{0} - 2 \cdot \log 21 + 6 \cdot \underbrace{\log 1}_{0} - 6 \cdot \log 20 + 1 \cdot \underbrace{\log 1}_{0} - 1 \cdot \log 19 + 1 \cdot \underbrace{\log 1}_{0} - 1 \cdot \log 18 \\
&+ \quad 3 \cdot \log 21 - 3 \cdot \log 22 + 6 \cdot \log 20 - 6 \cdot \log 22 + 1 \cdot \log 19 - 1 \cdot \log 22 + 1 \cdot \log 18 - 1 \cdot \log 22 \\
&+ \quad \frac{4}{9} \\
&= \quad \log 21 - 11 \cdot \log 22 + \frac{4}{9} \\
&\approx \quad -11,81 \qquad (67)
\end{aligned}
$$

Da $-11,81 > -\infty$ erhalten wir wieder das Resultat, dass d_{10} zu Kategory c_3 gehört.

Aufgabe 2.5: Naive-Bayes-Klassifikator mit Laplace-Smoothing

Auf das Beispiel der Vorlesung soll Lapalace-Smoothing angewendet werden, d.h. für $p_{i,k}$ Werte wird nun die folgende Formel benutzt:

$$p_{i,k} \quad = \quad \frac{f_i + 1}{m + \sum_{d \in c_k} length(d)} \qquad (68)$$

Wir erhalten für die folgenden Kategorien die Werte:

$$
\begin{aligned}
\sum_{d \in c_1} length(d) &= 12 \qquad (69) \\
\sum_{d \in c_2} length(d) &= 12 \qquad (70) \\
\sum_{d \in c_3} length(d) &= 12 \qquad (71) \\
& \qquad (72)
\end{aligned}
$$

Und damit die folgende Tabelle der $p_{i,k}$:

$p_{i,k}$	$k=1$	$k=2$	$k=3$
$i=1$	$\frac{5}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
$i=2$	$\frac{5}{20}$	$\frac{1}{20}$	$\frac{1}{20}$
$i=3$	$\frac{4}{20}$	$\frac{2}{20}$	$\frac{2}{20}$
$i=4$	$\frac{1}{20}$	$\frac{6}{20}$	$\frac{2}{20}$
$i=5$	$\frac{1}{20}$	$\frac{6}{20}$	$\frac{2}{20}$
$i=6$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{3}{20}$
$i=7$	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{5}{20}$
$i=8$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{3}{20}$

Die analoge Berechnung zur Vorlesung ergibt nun

$$
\begin{aligned}
k = 1(Algebra) &: \frac{4}{20^5} \qquad (73) \\
k = 2(Analysis) &: \frac{576}{20^5} \qquad (74) \\
k = 3(Stochastik) &: \frac{50}{20^4} \qquad (75) \\
& \qquad (76)
\end{aligned}
$$

Somit bleibt es dabei, dass d_7 der Klasse Stochastik zugeordnet wird.

Aufgabe 2.6: Naive-Bayes-Klassifikator mit binären Features

Gesucht: Eine Abwandlung der einfahsten möglichen Entscheidungsformel für den Spezialfall der binären Klassifikation der Naive-Bayes-Methode mit unabhängigen binären Features.

Es gilt:

$$P\left[c\left|\vec{X}\right.\right] = 1 - P\left[\neg c\left|\vec{X}\right.\right] \tag{77}$$

$$p_c = 1 - p_{\neg c} \tag{78}$$

Wir setzen $c_1 := c$ und $c_2 := \neg c$ und betrachten den Logarithmus des Quotienten aus $P\left[c_1\left|\vec{X}\right.\right]$ und $P\left[c_2\left|\vec{X}\right.\right]$. Da wir nun binäre unabhängige Features haben gilt:

$$\frac{P\left[c_1\left|\vec{X}\right.\right]}{P\left[c_2\left|\vec{X}\right.\right]} = \prod_i \frac{P\left[c_1\left|X_i\right.\right]}{P\left[c_2\left|X_i\right.\right]} \tag{79}$$

Ob ein Dokument zu Kategorie $c = c_1$ oder zu $\neg c = c_2$ gehört, hängt von der folgenden Bedingung ab:

$$P\left[c_1\left|\vec{X}\right.\right] > P\left[c_2\left|\vec{X}\right.\right] \tag{80}$$

$$\frac{P\left[c_1\left|\vec{X}\right.\right]}{P\left[c_2\left|\vec{X}\right.\right]} > 1 \tag{81}$$

$$\log\frac{P\left[c_1\left|\vec{X}\right.\right]}{P\left[c_2\left|\vec{X}\right.\right]} > 0 \tag{82}$$

$$\tag{83}$$

Die letzte Formel kann umgewandelt werden zu:

$$\begin{aligned}
\log\frac{P\left[c_1\left|\vec{X}\right.\right]}{P\left[c_2\left|\vec{X}\right.\right]} &= \log\frac{\prod_{i=1}^m \left(p_{i,1}^{X_i}\cdot\left(1-p_{i,1}\right)^{1-X_i}\right)\cdot p_1}{\prod_{i=1}^m \left(p_{i,2}^{X_i}\cdot\left(1-p_{i,2}\right)^{1-X_i}\right)\cdot p_2} \\
&= \log\frac{\prod_{i=1}^m \left(\left(\frac{p_{i,1}}{(1-p_{i,1})}\right)^{X_i}\cdot\left(1-p_{i,1}\right)\right)\cdot p_1}{\prod_{i=1}^m \left(\left(\frac{p_{i,2}}{(1-p_{i,2})}\right)^{X_i}\cdot\left(1-p_{i,2}\right)\right)\cdot p_2} \\
&= \log\left(\frac{p_1}{p_2}\cdot\prod_{i=1}^m \frac{\left(\frac{p_{i,1}}{(1-p_{i,1})}\right)^{X_i}\cdot\left(1-p_{i,1}\right)}{\left(\frac{p_{i,2}}{(1-p_{i,2})}\right)^{X_i}\cdot\left(1-p_{i,2}\right)}\right) \\
&= \log\left(\frac{p_1}{p_2}\cdot\prod_{i=1}^m \left(\left(\frac{p_{i,1}\cdot(1-p_{i,2})}{p_{i,2}\cdot(1-p_{i,1})}\right)^{X_i}\cdot\frac{1-p_{i,1}}{1-p_{i,2}}\right)\right) \\
&= \log\frac{p_1}{p_2} + \sum_{i=1}^m \log\left(\left(\frac{p_{i,1}\cdot(1-p_{i,2})}{p_{i,2}\cdot(1-p_{i,1})}\right)^{X_i}\cdot\frac{1-p_{i,1}}{1-p_{i,2}}\right) \\
&= \log\frac{p_1}{p_2} + \sum_{i=1}^m \left(\log\left(\left(\frac{p_{i,1}\cdot(1-p_{i,2})}{p_{i,2}\cdot(1-p_{i,1})}\right)^{X_i}\right) + \log\frac{1-p_{i,1}}{1-p_{i,2}}\right) \\
&= \log\frac{p_1}{p_2} + \sum_{i=1}^m \left(X_i\cdot\log\frac{p_{i,1}\cdot(1-p_{i,2})}{p_{i,2}\cdot(1-p_{i,1})} + \log\frac{1-p_{i,1}}{1-p_{i,2}}\right) \\
&= \log\frac{p_1}{p_2} + \sum_{i=1}^m X_i\cdot\log\frac{p_{i,1}\cdot(1-p_{i,2})}{p_{i,2}\cdot(1-p_{i,1})} + \sum_{i=1}^m \log\frac{1-p_{i,1}}{1-p_{i,2}}
\end{aligned} \tag{84}$$

Nun ist die Frage, ob:

$$-\log\frac{p_1}{p_2}-\sum_{i=1}^{m}\log\frac{1-p_{i,1}}{1-p_{i,2}} \quad < \quad \sum_{i=1}^{m} X_i\cdot\log\frac{p_{i,1}\cdot(1-p_{i,2})}{p_{i,2}\cdot(1-p_{i,1})} \quad \text{or, resp.} \tag{85}$$

$$\log\frac{p_1}{p_2}+\sum_{i=1}^{m}\log\frac{1-p_{i,1}}{1-p_{i,2}} \quad > \quad \sum_{i=1}^{m} X_i\cdot\log\frac{p_{i,2}\cdot(1-p_{i,1})}{p_{i,1}\cdot(1-p_{i,2})} \ . \tag{86}$$

Die rechte Seite der Gleichung ist konstant für eine gegebene Klasse c (mit $c_1 := c$ und $c_2 := \neg c$). Somit haben wir nur den linken Teil für jedes testdokument zu berechnen. Glücklicherweise ist dies nur von den Worten des Dokuments abhängig und nicht von den Worten, die nicht vorkommen (in der Regel eine viel größere Menge).