Zufällig oder signifikant?

Teilnehmer:

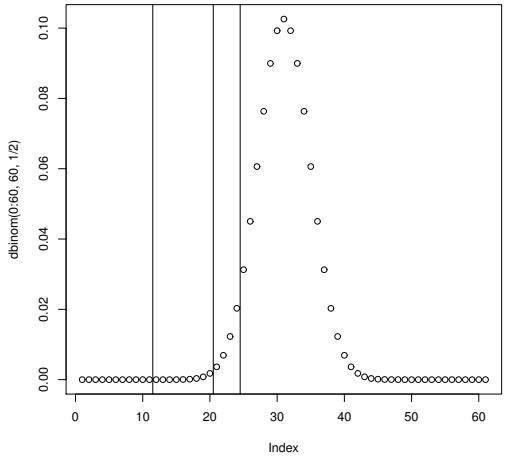
Heinrich-Hertz-Gymnasium	1 Mädchen
Herder-Gymnasium	1 Mädchen, 3 Jungen
Immanuel-Kant-Gymnasium	1 Mädchen, 1 Junge
Käthe-Kollwitz-Gymnasium	1 Junge

mit tatkräftiger Unterstützung durch:

Niklas Sturm	Humboldt-Universität zu Berlin

Gruppenleiter:

Markus Reiß	Humboldt-Universität zu Berlin
Randolf Altmeyer	Humboldt-Universität zu Berlin



1. Einleitung

Zufallsexperimente beschreiben zufällige Vorgänge, die bei beliebig häufiger Wiederholung zu ähnlichen, aber eben nicht unbedingt gleichen Ergebnissen führen. Zum Beispiel erwartet man beim Werfen eines fairen Würfels, dass jede Seite mit gleicher Häufigkeit auftritt. Bei welchen Ergebnissen kann man mit hoher Sicherheit davon ausgehen, dass er nicht fair ist? Dies ist nur eins von vielen Problemen, welche sich mithilfe der Stochastik lösen lassen.
In dieser Woche haben wir Zufallsexperimente mathematisch formalisiert und auf ihre Eigenschaften hin untersucht. Das Ziel war dabei, interessante Strukturen im Zufall zu finden. Für diesen Bericht haben wir aus einer spannenden und umfangreichen Woche eine Auswahl von Resultaten getroffen, die uns besonders begeistert haben. Wir legen unsere Schwerpunkte auf die folgenden Themen:

- das schwache Gesetz der großen Zahlen,
- ein eleganten und kurzen Beweis des Weierstraßsch'en Approximationssatzes,
- optimale einseitige Tests,
- eine Anwendung von statistischen Tests: der SPAM-Filter.

Unser Ziel in diesem Bericht ist mit sauberer Mathematik zu zeigen, dass Stochastik ein interessanter und nützlicher Bereich der Mathematik ist. Im Folgenden gehen wir von Schulwissen in Stochastik aus (Mittelstufe bzw. auch Oberstufe).

2. Das schwache Gesetz der großen Zahlen

2.1. Bernoulli-Ketten und Binomalverteilung

Im Folgenden betrachten wir nur wiederholte Bernoulli-Experimente. Dabei wird ein Zufallsexperiment Bernoulli-Experiment genannt, wenn es nur zwei mögliche Ausgänge gibt (Erfolg, Misserfolg). Die Wahrscheinlichkeit für einen Erfolg sei p. Bei n-maliger Wiederholung des Experiments sprechen wir von einer Bernoulli-Kette. Das Ergebnis dieser Bernoulli-Kette beschreiben wir mit einer Zufallszahl X, die Werte in der Menge $\{0,1\}^n$ annimmt, d.h. jedes mögliche Ergebnis ist ein Vektor der Länge n, dessen Komponenten nur die Werte 0 und 1 annehmen. Die Anzahl der Erfolge in dieser Kette bezeichnen wir mit $S_n(X)$. Bekanntlich ist $S_n(X)$ binomialverteilt zu den Parametern n und p, d.h.

$$\mathbb{P}(S_n(X) = k) = B_{n,p}(k),\ k \in \{0, \dots, n\},$$

wobei

$$B_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Die relative Häufigkeit

$$\frac{1}{n} S_n(X)$$

ist der Quotient aus der Anzahl der Erfolge und der Gesamtanzahl.

2.2. Der Beweis

Der folgende Satz ist von zentraler Bedeutung in der Stochastik und formalisiert die Intuition, dass die relative Häufigkeit gegen die Erfolgswahrscheinlichkeit p konvergiert, wenn das Experiment hinreichend oft wiederholt wird.

Satz 1. *Sei $S_n(X)$ binomialverteilt. Dann gilt für alle $\varepsilon > 0$:*

$$\mathbb{P}\left(|\frac{1}{n}S_n(X) - p| \geq \varepsilon\right) \to 0, \quad \textit{für } n \to \infty.$$

Beweis. Wir betrachten die Abweichungen nach oben und unten getrennt:

$$\mathbb{P}\Big(|\frac{1}{n}S_n(X) - p| \geq \varepsilon\Big) = \mathbb{P}\Big(\frac{1}{n}S_n(X) - p \geq \varepsilon\Big) + \mathbb{P}\Big(\frac{1}{n}S_n(X) - p \leq -\varepsilon\Big).$$

Wir betrachten den linken Summanden für $\alpha \geq 0$ (Beweis für den rechten Summanden folgt analog):

$$\begin{aligned}
\mathbb{P}\Big(\frac{1}{n}S_n(X) - p \geq \varepsilon\Big) &= \mathbb{P}\left(\frac{1}{n}S_n(X) \geq \varepsilon + p\right) \\
&= \sum_{k:\,\frac{k}{n} \geq \varepsilon + p} B_{n,p}(k) \\
&= \sum_{k:\,\frac{k}{n} \geq \varepsilon + p} \left(e^{-\alpha(\varepsilon+p)} \cdot e^{\alpha(\varepsilon+p)}\right) B_{n,p}(k) \\
&\leq e^{-\alpha(\varepsilon+p)} \sum_{k:\,\frac{k}{n} \geq \varepsilon + p}^{n} e^{\frac{\alpha k}{n}} B_{n,p}(k) \\
&\leq e^{-\alpha(\varepsilon+p)} \sum_{k=0}^{n} e^{\frac{\alpha k}{n}} B_{n,p}(k).
\end{aligned}$$

Nach Einsetzen der Binomialwahrscheinlichkeiten und Umstellen ist dies gleich

$$e^{-\alpha(\varepsilon+p)} \sum_{k=0}^{n} \binom{n}{k} \left(e^{\frac{\alpha}{n}} p\right)^k (1-p)^{n-k}.$$

Anwenden des Binomischen Lehrsatzes ergibt:

$$\mathbb{P}\Big(\frac{1}{n}S_n(X) - p \geq \varepsilon\Big) \leq e^{-\alpha(\varepsilon+p)} \left(1 + p(e^{\frac{\alpha}{n}} - 1)\right)^n.$$

Mit $1 + x \leq e^x$ lässt sich weiter abschätzen:

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}S_n(X) - p \geq \varepsilon\right) &\leq e^{-\alpha(\varepsilon+p) + np(e^{\frac{\alpha}{n}} - 1)} \\
&= e^{-\alpha(\varepsilon+p) + \alpha \cdot p \frac{e^{\alpha/n} - e^0}{\alpha/n - 0}}.
\end{aligned}$$

Der Differenzenquotient in der letzten Zeile konvergiert für $n \to \infty$ gegen Eins und daher konvergiert die letzte Zeile gegen

$$e^{-\alpha(\varepsilon+p)+\alpha p} = e^{-\alpha\varepsilon}.$$

Für jedes beliebige $\delta > 0$ können wir $\alpha = -\frac{\ln(\delta)}{\varepsilon}$ wählen, so dass

$$\mathbb{P}\left(\frac{1}{n}S_n(X) - p \geq \varepsilon\right) \leq e^{-\alpha\varepsilon} = \delta.$$

Da δ beliebig ist, folgt die Aussage. □

Intuitiv sagt dieser Satz, dass die Wahrscheinlichkeit mit der die relative Häufigkeit bei steigendem n außerhalb jeder beliebigen ε-Umgebung um die Erfolgswahrscheinlichkeit p liegt, gegen Null konvergiert.

2.3. Approximationssatz von Weierstraß

Mithilfe des schwachen Gesetzes der großen Zahlen können wir einen kurzen und eleganten Beweis für einen wichtigen Satz der Analysis geben.

Satz 2. *Sei f eine stetige Funktion auf $[0,1]$. Dann existiert zu jedem $\varepsilon > 0$ ein Polynom*

$$f_n(p) = \sum_{k=0}^{n} a_k p^k, \quad p \in [0,1],$$

vom Grad n mit

$$|f(p) - f_n(p)| \leq \varepsilon, \quad p \in [0,1].$$

Beweis. Betrachte zu $p \in [0,1]$ das Bernstein-Polynom:

$$f_n(p) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) B_{n,p}(k)$$

und bilde

$$|f(p) - f_n(p)| = \left| \sum_{k=0}^{n} \left(f(p) - f\left(\frac{k}{n}\right)\right) B_{n,p}(k) \right|.$$

Hier wurde $f(p)$ in die Summe gezogen, weil es unabhängig von k ist und weil sich die Binomialwahrscheinlichkeiten $B_{n,p}(k)$ zu Eins aufsummieren, wenn man über k summiert. Nach der Dreiecksungleichung gilt nun für jedes $\delta > 0$

$$\begin{aligned} |f(p) - f_n(p)| \leq & \sum_{k:|\frac{k}{n}-p|\leq\delta} \left| f(p) - f\left(\frac{k}{n}\right)\right| B_{n,p}(k) \\ & + \sum_{k:|\frac{k}{n}-p|>\delta} \left| f(p) - f\left(\frac{k}{n}\right)\right| B_{n,p}(k). \end{aligned}$$

Weil f stetig und auf dem kompakten Intervall $[0,1]$ definiert ist, nimmt f auf $[0,1]$ sein Maximum an. Also ist:

$$\begin{aligned} |f(p) - f_n(p)| \leq & \max_{k:|\frac{k}{n}-p|\leq\delta} \left| f(p) - f\left(\frac{k}{n}\right)\right| \sum_{k:|\frac{k}{n}-p|\leq\delta} B_{n,p}(k) \\ & + 2 \max_{0\leq x\leq 1} |f(x)| \sum_{k:|\frac{k}{n}-p|>\delta} B_{n,p}(k). \end{aligned}$$

Wegen der Stetigkeit von f existiert ein $\delta > 0$, so dass für alle $p, x \in [0,1]$ mit $|p-x| \leq \delta$ gilt $|f(p)-f(x)| < \varepsilon/2$. Für dieses δ folgt also durch Umschreiben des zweiten Summanden:

$$|f(p) - f_n(p)| \leq \varepsilon/2 + 2 \max_{0\leq x\leq 1} |f(x)| \mathbb{P}\left(\left| \frac{1}{n} S_n(X) - p \right| > \delta \right).$$

Nach dem schwachen Gesetz der großen Zahlen findet sich für dieses δ außerdem ein $n_0 \in \mathbb{N}$, so dass für alle $n > n_0$

$$\mathbb{P}\left(\left| \frac{1}{n} S_n(X) - p \right| > \delta \right) \leq \frac{\varepsilon}{2 \cdot 2 \max_{0\leq x\leq 1} |f(x)|}.$$

Insgesamt ergibt dies:

$$|f(p) - f_n(p)| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

□

3. Der beste einseitige Test

3.1. Statistische Tests

Mithilfe eines statistischen Tests kann anhand von Daten entschieden werden, ob eine Hypothese über die Daten beibehalten werden kann oder abgelehnt werden muss. Wir machen die folgenden Annahmen über das Zufallsexperiment, das die Daten erzeugt:

- Es gibt eine Parametermenge $\Theta \subset \mathbb{R}$, die disjunkt zerlegt ist in die Mengen H_0 (Nullhypothese) und H_1 (alternative Hypothese).
- Das Ergebnis des Zufallsexperiments wird durch eine Zufallszahl X beschrieben mit Werten in einer Menge Ω und mit Verteilung $A \mapsto P_\vartheta(X \in A)$ für $A \subset \Omega$, $\vartheta \in \Theta$. Die Verteilung hängt also von ϑ ab.

ϑ ist unbekannt und wir wollen herausfinden, ob $\vartheta \in H_0$ oder $\vartheta \in H_1$. Jede Funktion $\varphi : \Omega \to \{0, 1\}$ wird statistischer Test genannt. Wenn $\varphi(x) = 1$, dann wird die Nullhypothese H_0 abgelehnt (d.h. wir glauben, dass $\vartheta \in H_1$). Wenn $\varphi(X) = 0$, dann wird sie nicht abgelehnt (was auf $\vartheta \in H_0$ hindeutet). Mit jeder Entscheidung kommt auch ein Risiko sich falsch zu entscheiden. Entscheidet der Test die Nullhypothese abzulehnen, obwohl $\vartheta \in H_0$, dann ist dies ein peinlicher Fehler, den wir höchstens mit Wahrscheinlichkeit

$$\mathbb{P}_\vartheta(\varphi(X) = 1) \leq \alpha$$

für ein Signifikanzniveau $\alpha > 0$ machen wollen. Andererseits soll die 'Power' des Tests, also die Wahrscheinlichkeit die Alternative richtig anzunehmen, möglichst groß sein, wenn also $\vartheta \in H_1$.

3.2. Statistischer Test am Beispiel des Würfelwurfs

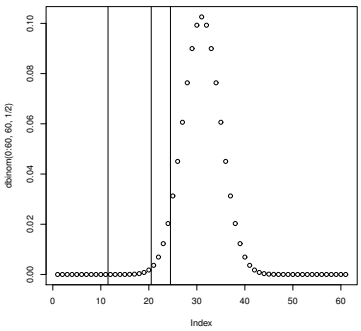
In einem Spaßartikelladen wird ein gezinkter Würfel verkauft. Für den Käufer ist es interessant, ob die Sechs wahrscheinlicher vorkommt als die anderen Zahlen. Wir überprüfen hier die Angabe des Herstellers, dass die Wahrscheinlichkeit eine Sechs zu würfeln größer gleich $\frac{1}{2}$ ist. Der Würfelkäufer würfelt 60 Mal und erhält 31 Sechsen. Ist der Würfel wirklich gezinkt?
Hier liegt eine Bernoulli-Kette vor, da wir uns nur dafür interessieren, ob eine Sechs gewürfelt wird oder nicht. Zur Schätzung der Erfolgswahrscheinlichkeit p wird die Anzahl der Erfolge verwendet. Wir wählen

- $H_0 : \ p \geq 1/2$,
- $H_1 : \ p < 1/2$,
- $n = 60$.

Es müssen noch α und der Test φ gewählt werden. Je nach Wahl wird ein Ablehnbereich definiert, so dass für eine Anzahl von Erfolgen in diesem Bereich die Nullhypothese verworfen wird. Da die Hypothesen einseitig sind, bedeutet dies, dass wir ein $c_\alpha > 0$ suchen, so dass

$$\mathbb{P}_\vartheta(S_n(X) < c_\alpha) \leq \alpha.$$

Dieses c_α findet man mithilfe einer Tabelle mit Binomialwahrscheinlichkeiten, oder liest es aus der folgenden Graphik ab ($\text{Bin}_{60,1/2}$-Verteilung):



Der Test ist dann

$$\varphi(X) = \begin{cases} 1, & S_n(X) < c_\alpha, \\ 0, & S_n(X) \geq c_\alpha. \end{cases}$$

Für verschiedene α ergibt sich

α	lehne H_0 ab
0,1	$S_n(X) < 25$
0,01	$S_n(X) < 21$
10^{-6}	$S_n(X) < 11$

In der Tabelle sieht man den Einfluss von α auf den Ablehnbereich. Je kleiner α ist, desto kleiner wird der Ablehnbereich und desto größer wird die Wahrscheinlichkeit, einen Würfel, der die Angabe des Herstellers erfüllt, falsch als einen Würfel, der die Angabe des Herstellers nicht erfüllt, einzustufen. Anderseits sinkt dadurch auch die Power des Tests, d.h. zu erkennen, ob tatsächlich $p < 1/2$ gilt. Insbesondere können wir für keines dieser α bei 31 Sechsen in 60 Würfen die Nullhypothese ablehnen, dass $p \geq 1/2$.

3.3. Optimale Tests

Wie findet man einen guten Test? Dafür betrachten wir die Likelihood-Funktion

$$\vartheta \mapsto L(\vartheta, x) = \mathbb{P}_\vartheta(X = x),$$

die einen Parameter auf die Wahrscheinlichkeit für $x \in \Omega$ abbildet.

Definition 1. Neyman-Pearson Test
Wenn $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_1$ getestet werden soll, dann heißt jeder Test der Form

$$\varphi(X) = \begin{cases} 1, R(X, \vartheta_1, \vartheta_0) > c \\ 0, R(X, \vartheta_1, \vartheta_0) < c, \end{cases}$$

für $c > 0$ Neyman-Pearson Test, wobei

$$R(x, \vartheta_1, \vartheta_0) = \frac{L(\vartheta_1, x)}{L(\vartheta_0, x)}, \ x \in \Omega.$$

Lemma 1. *(Neyman-Pearson Lemma)*

1. *Jeder Neyman-Pearson Test ist der beste Test für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_1$ zum Niveau α mit $\alpha = \mathbb{P}_{\vartheta_0}(\varphi(X) = 1)$, d.h. φ hat die höchste Power $\mathbb{P}_{\vartheta_1}(\varphi(X) = 1)$ von allen Tests zum Niveau α.*

2. Zu jedem Nivea $\alpha \in [0,1]$ gibt es einen Neyman-Pearson Test.

Beweis. Wir zeigen nur die erste Aussage, und diese nur für endliches Ω. Sei $\varphi^* : \Omega \to [0,1]$ ein beliebiger Test zum Niveau α. Zu zeigen ist, dass $\mathbb{P}_{\vartheta_1}(\varphi(X) = 1) \geq \mathbb{P}_{\vartheta_1}(\varphi^*(X) = 1)$. Es gilt

$$\mathbb{P}_{\vartheta_1}(\varphi(X) = 1) - \mathbb{P}_{\vartheta_1}(\varphi^*(X) = 1) = \sum_{x \in \Omega : \varphi(x) = 1} \mathbb{P}_{\vartheta_1}(X = x) - \sum_{x \in \Omega : \varphi^*(x) = 1} \mathbb{P}_{\vartheta_1}(X = x).$$

Nach Zusammenfassen der beiden Summen sei $x \in \Omega$. Wenn $\varphi(x) < \varphi^*(x)$, dann ist der entsprechende Summand der zusammengefassten Summe gleich $-\mathbb{P}_{\vartheta_1}(X = x)$. Für $\varphi(x) = \varphi^*(x)$ ist er gleich Null, und für $\varphi(x) > \varphi^*(x)$ ist er gleich $\mathbb{P}_{\vartheta_1}(X = x)$. Insgesamt ist also:

$$\mathbb{P}_{\vartheta_1}(\varphi(X) = 1) - \mathbb{P}_{\vartheta_1}(\varphi^*(X) = 1) = \sum_{x \in \Omega : \varphi(x) > \varphi^*(x)} \mathbb{P}_{\vartheta_1}(X = x) - \sum_{x \in \Omega : \varphi(x) < \varphi^*(x)} \mathbb{P}_{\vartheta_1}(X = x).$$

Wenn $\varphi(x) > \varphi^*(x)$, dann muss gelten

$$R(x, \vartheta_1, \vartheta_0) = \frac{L(\vartheta_1, x)}{L(\vartheta_0, x)} > c.$$

Andererseits folgt aus $\varphi(x) < \varphi^*(x)$, dass

$$R(x, \vartheta_1, \vartheta_0) < c.$$

Insgesamt ist also:

$$\begin{aligned}
\mathbb{P}_{\vartheta_1}(\varphi(X) = 1) - \mathbb{P}_{\vartheta_1}(\varphi(X) = 1) = & \sum_{x \in \Omega : \varphi(x) > \varphi^*(x)} R(x, \vartheta_1, \vartheta_0) L(\vartheta_0, x) \\
& - \sum_{x \in \Omega : \varphi(x) < \varphi^*(x)} R(x, \vartheta_1, \vartheta_0) L(\vartheta_0, x) \\
& > c \sum_{x \in \Omega : \varphi(x) > \varphi^*(x)} L(\vartheta_0, x) - \sum_{x \in \Omega : \varphi(x) < \varphi^*(x)} L(\vartheta_0, x) \\
& = \mathbb{P}_{\vartheta_0}(\varphi(X) = 1) - \mathbb{P}_{\vartheta_0}(\varphi(X) = 1).
\end{aligned}$$

Da $\mathbb{P}_{\vartheta_0}(\varphi(X) = 1) = \alpha$, und $\mathbb{P}_{\vartheta_0}(\varphi^*(X) = 1) \leq \alpha$, folgt die Aussage. □

Bemerkung 1. Ein Korollar aus dem Neyman-Person Lemma ist, dass der beste einseitige Test (z.B. wie im Beispiel oben) der beste Test zum Niveau α ist.

4. Spam-Filter

Wir kommen nun zu einem Anwendungsbeispiel der Test-Theorie: Dem Spam-Filter.

4.1. Idee des Spamfilters

Die Aufgabe eines Spamfilters ist es, eingehende E-Mails in Spam und Nicht-Spam zu klassifizieren. Im Folgenden stellen wir einen von uns implementierten Spamfilter vor, basierend auf dem 'naiven Bayes Klassifizierer'.

4.2. Bayes-Klassifizierer

In diesem Kapitel nennen wir den Test φ Klassifizierer. Die Grundlage für seine Definition ist die bedingte

Wahrscheinlichkeit. Im Folgenden sei x eine Realiserung einer vektoriellen Zufallszahl X, deren Komponenten Wörtern aus einem mit Testdaten trainierten Wörterbuch entsprechen und deren Wert angibt, ob ein Wort in einer Email vorkommt oder nicht. Welche Wörter aus der Email berücksichtigt werden, wird wie erwähnt anhand von Trainingsdaten festgelegt. Sei außerdem K eine binäre Zufallszahl, die für eine Spam-Email den Wert Eins und für eine nicht-Spam-Email den Wert Null annimmt. Gesucht ist die bedingte Wahrscheinlichkeit dafür, dass eine gegebene Email zur Klasse K gehört, wobei dafür nach dem Satz von Bayes gilt:

$$\mathbb{P}(K = i|X = x) = \frac{\mathbb{P}(X = x|K = i)}{\mathbb{P}(X = x)} = \frac{\pi_i * \mathbb{P}_i(x)}{\pi_0 * \mathbb{P}_0(x) + \pi_1 * \mathbb{P}_1(x)}.$$

Hierbei ist $\pi_i = \mathbb{P}(K = i)$ die Wahrscheinlichkeit Klasse i zu sehen für $i = 0, 1$.

4.3. Definition des Klassifizierers

Wir modellieren unseren Klassifizierer nach Bayes wie folgt:

$$\varphi(X) = \begin{cases} 1, & \eta(X) \geq \alpha, \\ 0, & \eta(X) < \alpha, \end{cases}$$

für einen Schwellenwert α und mit $\eta(x) = \mathbb{P}(K = 1|X = x)$. Dann ist

$$\varphi(X) = 1 \Longleftrightarrow \frac{P(X = x|K = 1)}{P(X = x|K = 0)} \geq \frac{\alpha}{1 - \alpha} \frac{P(K = 1)}{P(K = 0)}.$$

4.4. Modellierung

Wir betrachten eine unabhängige Verkettung der Gewichtung der einzelnen Wörter, die in einer E-Mail vorhanden sind. Jedes Wort ist dabei Bernoulli-verteilt mit seiner eigenen Erfolgswahrscheinlichkeit in einer Spam-Email beziehungsweise in einer Ham-Email zu erscheinen:

$$\mathbb{P}(X = x|K = 1) = p_{1,1}^{x_1}(1 - p_{1,1})^{1-x_1} * p_{2,1}^{x_2}(1 - p_{2,1})^{1-x_2} \cdots p_{N,1}^{x_N}(1 - p_{N,1})^{1-x_N},$$
$$\mathbb{P}(X = x|K = 0) = p_{1,0}^{x_1}(1 - p_{1,0})^{1-x_1} * p_{2,0}^{x_2}(1 - p_{2,0})^{1-x_2} \cdots p_{N,0}^{x_N}(1 - p_{N,0})^{1-x_N}.$$

Bei dieser Modellierung wird nicht unterschieden, wo in einer Email oder wie oft ein Wort vorkommt. Außerdem werden alle Abhängigkeiten zwischen einzelnen Worten ignoriert. Jede Email ist nur ein 'bag of words'. Da dies sehr naiv ist, heißt dieser Algorithmus auch 'naiver Bayes Klassifizierer'.

4.5. Schätzung $\hat{p}$

Für die Bernoulli-Verkettung benötigen wir die $p_{m,k}$. An dieser Stelle wertet unser Programm Trainings-SMS aus, um diese zu schätzen. Es gilt dabei das Gesetz der großen Zahlen: Je mehr Trainingsdaten verwendet werden, desto mehr entspricht die relative Häufigkeit der tatsächlichen Erfolgswahrscheinlichkeit. Dazu werden die Trainingsdaten in zwei Mengen S für Spam der Länge n_S und H für Ham der Länge n_H aufgeteilt. Die relativen Häufigkeiten sind dann:

$$\hat{p}_{m,1} = \frac{1}{n_S} \sum_{x \in S} x_m, \hat{p}_{m,0} = \frac{1}{n_H} \sum_{x \in H} x_m.$$

Dementsprechend schätzen wir π_1 durch $\hat{pi_1} = n_S$, wobei $n = n_S + n_H$.

4.6. Optimierung durch Summen

Da die Multiplikation von vielen kleinen Zahlen numerisch instabil ist, wird die Multiplikation durch Logarithmieren in eine Summation umgewandelt, d.h. wir betrachten

$$\ln \mathbb{P}(X = x|K = 1) - \ln \mathbb{P}(X = x|K = 0) \geq \ln \frac{\alpha}{1-\alpha} + \ln \mathbb{P}(K = 0) - \ln \mathbb{P}(K = 1),$$

wobei, wenn das Wörterbuch N Worte enthält,

$$\begin{aligned}\mathbb{P}(X = x|K = 1) &= \ln (p_{1,1}^{x_1}(1 - p_{1,1})^{1-x_1}) + \ln (p_{2,1}^{x_2}(1 - p_{2,1})^{1-x_2}) + .. + \ln (p_{N,1}^{x_N}(1 - p_{N,1})^{1-x_N}) \\ &= \sum_{k=1}^{N} x_k \ln \frac{p_{k,1}}{1 - p_{k,1}} + \sum_{k=1}^{N} \ln (1 - p_{k,1}).\end{aligned}$$

4.7. Ergebnisse

Wir haben den Spamfilter in der Programmiersprache R implementiert, mit 4000 SMS (statt Emails) trainiert und mit 1388 neuen SMS getestet.Wir haben die folgenden Ergebnisse erhalten:

	actual		
predicted	0	1	Row Total
ham	1203	32	1235
	0.974	0.026	0.890
	0.997	0.177	
	0.867	0.023	
spam	4	149	153
	0.026	0.974	0.110
	0.003	0.823	
	0.003	0.107	
Column Total	1207	181	1388
	0.870	0.130	

Abbildung 12: Ergebnisse des Spam-Filter

Man sieht, dass der Spamfilter erstaunlich gut funktioniert: Sowohl die Erkennungsrate für Spam, als auch für Ham, liegt etwas bei 97 Prozent.