Karlsruher Institut für Technologie

Fakultät für Mathematik

Institut für Angewandte und Numerische Mathematik 4: Numerische Simulation, Optimierung und Hochleistungsrechnen

Sekretariat
Kollegiengebäude Mathematik (20.30)
Zimmer 3.039

Adresse
Karlsruher Institut für Technologie (KIT)
Institut für Angewandte und Numerische Mathematik
Englerstrasse 2
76131 Karlsruhe

Öffnungszeiten:

Tel.: +49 721 608 - 42062

Fax.: +49 721 608 - 44178

Die Mathematik des Bayes Spamfilters - S. Ritterbusch

Im Artikel *A Plan for Spam* schlug Paul Graham 2002 ein statistische Verfahren zur Klassifizierung von unerwünschten Spam-Mails vor, das inzwischen Bestandteil vieler Spam-Filter ist. Die Erklärungen von P. Graham zur Mathematik hinter den Formeln ist sehr knapp, und das verlinkte Dokument mit einer guten Erklärung bietet viele Varianten des Problems. Hier ist die Mathematik hinter der Technik auf die tatsächlich verwendeten Zusammenhänge reduziert dargestellt.

Annahmen

Wir wollen die Wahrscheinlichkeit dafür berechnen, dass eine empfangene Mail entweder Spam oder eine erwünschte Mail (Ham) ist, unter der Bedingung dass bestimmte Worte in der Mail gefunden wurden. Das vorgeschlagene Verfahren von P. Graham macht **zwei Annahmen**:

(1) Die Wahrscheinlichkeiten, dass eine Mail Spam oder Ham ist, sind gleich 1/2:

$$P(H) = P(S) = \frac{1}{2}$$

Diese Annahme wird auch als "Unvoreingenommenheit des Filters" bezeichnet, tatsächlich kann man aber auch andere Annahmen treffen. Die Erfahrung zeigt, dass man inzwischen mehr Spam als Ham empfängt, und so Spam eigentlich eine höhere Wahrscheinlichkeit zugeordnet werden sollte- doch da es viel schlimmer ist, wenn eine Ham-Mail fälschlicherweise als Spam einsortiert wird, belässt man es bei dieser bewährten Annahme.

(2) Die betrachteten Worte treten in Ham und Spam voneinander stochastisch unabhängig auf:

$$P(\mathrm{wort}_1 \cap \ldots \cap \mathrm{wort}_n | H) = P(\mathrm{wort}_1 | H) \cdot \ldots P(\mathrm{wort}_n | H)$$

$$P(\mathrm{wort}_1 \cap \ldots \cap \mathrm{wort}_n | S) = P(\mathrm{wort}_1 | S) \cdot \ldots P(\mathrm{wort}_n | S)$$

Diese Annahme ist grundlegend für das Verfahren und ist leider für viele Worte unsinnig: Natürlich treten Worte in Sätzen in Zusammenhang zueinander auf (wo "*Baden*" ist, ist "*Württemberg*" nicht weit), aber die Hoffnung ist, dass zumindest für die betrachteten Worte die Annahme eine gute Näherung liefert.

Verwendete Formeln

Es seien $A, B \in \mathcal{A}$ und $(B_j) \subset \mathcal{A}$ Ereignisse in einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dann gelten die folgenden

Zusammenhänge:

(3) Formel von Bayes:

$$P(A|B) = \frac{P(B|A)}{P(B)}\,P(A)$$

(4) Satz von der totalen Wahrscheinlichkeit: Ist $\bigcup_j B_j = \Omega$ und $B_i \cap B_j = \emptyset$ wenn $i \neq j$, so gilt

$$P(A) = \sum_j P(A|B_j)\,P(B_j)$$

Herleitung der Formel

Es seien A und B Ereignismengen für zwei Wörter. Aus einer Datenbank von vorklassifizierten Mails kann man die Wahrscheinlichkeiten dafür, dass diese Wörter in Spam $P(A|S)$ oder Ham $P(A|H)$ Mails auftreten, abschätzen. Genauso kennen wir aus (1) die Wahrscheinlichkeiten $P(S)$ und $P(H)$. Uns interessiert nun die Wahrscheinlichkeit dafür, dass eine Mail Spam ist, unter der Bedingung, dass die Worte A und B in der Mail aufgetreten sind, also die Wahrscheinlichkeit $P(S|A \cap B)$:

$$\begin{aligned}
P(S|A\cap B) &\overset{(3)}{=} \frac{P(A\cap B|S)}{P(A\cap B)}P(S)\\
&\overset{(2)}{=} \frac{P(A|S)\,P(B|S)}{P(A\cap B)}P(S)\\
&\overset{(4)}{=} \frac{P(A|S)\,P(B|S)}{P(S)P(A\cap B|S)+P(H)P(A\cap B|H)}P(S)\\
&\overset{(2)}{=} \frac{P(A|S)\,P(B|S)}{P(S)P(A|S)P(B|S)+P(H)P(A|H)P(B|H)}P(S) \qquad (*)\\
&\overset{(1)}{=} \frac{P(A|S)\,P(B|S)}{P(A|S)P(B|S)+P(A|H)P(B|H)} \qquad (**)
\end{aligned}$$

Am Ende erhalten wir, unter den vorausgesetzten Bedingungen, eine Formel, mit denen wir die Wahrscheinlichkeit, dass wir es mit einer Spam Mail zu tun haben, mit Hilfe von Werten ausrechnen können, die wir schon kennen. Der letzte Schritt ist dabei nur zur Vereinfachung der Formel und der einzige Schritt, wo wir die Annahme (1) verwendet haben- es ist also vollkommen unproblematisch andere Annahmen zu treffen und die Formel $(*)$ zu verwenden. Die entsprechende Formel gilt auch bei mehreren Worten, zur Übersichtlichkeit wurden hier bei der Herleitung nur zwei gefundene Worte berücksichtigt:

$$P(S|A_1\cap\ldots\cap A_n) = \frac{P(A_1|S)\cdot\ldots\cdot P(A_n|S)}{P(A_1|S)\cdot\ldots\cdot P(A_n|S)+P(A_1|H)\cdot\ldots\cdot P(A_n|H)}$$

$$P(H|A_1\cap\ldots\cap A_n) = \frac{P(A_1|H)\cdot\ldots\cdot P(A_n|H)}{P(A_1|S)\cdot\ldots\cdot P(A_n|S)+P(A_1|H)\cdot\ldots\cdot P(A_n|H)}$$

Wir sehen, dass damit auch $P(S|A_1\cap\ldots\cap A_n)+P(H|A_1\cap\ldots\cap A_n)=1$ gilt.

Ein Beispiel

Durch Auszählen in 100 Ham-Mails und 100 Spam-Mails haben wir für die Worte **haben** und **online** folgende Zahlen erhalten (diese Analyse ist die Anlernphase):

	Ham	Spam
haben	30	7
online	3	8

D.h. in 30 von den 100 Ham-Mails kam das Wort "haben" vor (keine Mehrfachzählung). Damit schätzen wir die Wahrscheinlichkeiten unter denen die Wörter in Ham oder Spam vorkommen so ab:

$$\begin{aligned}
P(\text{haben}|H) &= 0{,}30 \qquad P(\text{haben}|S) = 0{,}07\\
P(\text{online}|H) &= 0{,}03 \qquad P(\text{online}|S) = 0{,}08
\end{aligned}$$

Zusammen mit der Information $P(H)=P(S)=\frac{1}{2}$ erhalten wir mit der Formel $(**)$ für eine Mail, in denen die

beiden Worte entdeckt wurden:

$$P(S|\text{haben} \cap \text{online}) = \frac{0,07 \cdot 0,08}{0,07 \cdot 0,08 + 0,30 \cdot 0,03} \approx 0,38 = 38\%$$

Oben wurde schon erwähnt, warum die Annahme (1) so sinnvoll ist- aber nehmen wir zum Beispiel einmal die realistischere Annahme $P(S) = 0,9$ und $P(H) = 0,1$ an, so erhalten wir mit Formel $(*)$ den folgenden Wert:

$$P(S|\text{haben} \cap \text{online}) = \frac{0,9 \cdot 0,07 \cdot 0,08}{0,9 \cdot 0,07 \cdot 0,08 + 0,1 \cdot 0,30 \cdot 0,03} \approx 0,85 = 85\%$$

Damit sieht man, dass das Verfahren dann sehr stark dazu tendiert Mails als Spam einzusortieren, und diese Voreingenommenheit will man durch die Annahme (1) vermeiden und setzt $P(H) = P(S) = 0,5$.

Natürlich sollte man die Entscheidung nicht von 2 Worten alleine abhängig machen, sondern man erstellt in der "Anlernphase" eine Liste der Worte, die am deutlichsten auf Ham oder Spam hinweisen und durchsucht dann die empfangenen Mails nach den zehn signifikantesten Worten in der Mail und führt dann die obige Analyse durch. Sollte eine Mail falsch, oder nur knapp klassifiziert werden, so ist es sinnvoll diese Mail in den Lernkorpus mit aufzunehmen und die Statistiken zu aktualisieren.

Weitere Darstellungen

- A Plan for Spam

Der ursprüngliche Vorschlag von Paul Graham, den er 2002 in *Better Bayesian Filtering* verbesserte.

- T. Arens, F. Hettlich, Ch. Karpfinger, U. Kockelkorn, K. Lichtenegger, H. Stachel: Mathematik. Spektrum Akademischer Verlag, Heidelberg, 2008.

In dieser Darstellung wird zwar ebenfalls die entscheidende Bedingung (2) gefordert, doch am Ende nur ein zu obiger Wahrscheinlichkeit äquivalenter aber quasi einheitenloser Quotient Q von Spam- zu Ham-Warscheinlichkeit berechnet:

$$Q(A_1 \cap \ldots \cap A_n) = \frac{P(S|A_1 \cap \ldots \cap A_n)}{P(H|A_1 \cap \ldots \cap A_n)} = \frac{P(A_1|S) \cdot \ldots \cdot P(A_n|S)}{P(A_1|H) \cdot \ldots \cdot P(A_n|H)}$$

$$P(S|A_1 \cap \ldots \cap A_n) = 1 - \frac{1}{Q(A_1 \cap \ldots \cap A_N) + 1} = \frac{Q(A_1 \cap \ldots \cap A_n)}{Q(A_1 \cap \ldots \cap A_n) + 1}$$

$$P(H|A_1 \cap \ldots \cap A_n) = \frac{1}{Q(A_1 \cap \ldots \cap A_N) + 1}$$

- Wikipedia Artikel zum Bayes-Filter in deutsch und Bayesian spam filtering in English

Der deutsche Artikel ist sehr knapp und gibt einen groben Überblick, dafür ist im deutlich informativeren englischen Artikel auch die Formel zur Berechnung der Wahrscheinlichkeit angegeben und führt auch frühere Quellen wie ifile von 1996 mit Vorschlägen zum Filtern von Spam mit Hilfe von bedingten Wahrscheinlichkeiten und der Bayes-Formel auf.

- POPFile Software

Neben vielen anderen Spamfiltern, die obige Techniken auf die eine oder andere Art einsetzen, erweitert POPFile die Klassifikation darauf, dass es mehr Klassen als nur Spam und Ham zulässt. Man kann die obigen Techniken auch auf mehr als zwei disjunkte Klassen einstufen, und dieses Programmpaket ermöglicht so eine inhaltsbasierte Einteilung der empfangenen Nachrichten.