

Zu Abschnitt 20.7: Bayes-Spamfilter

Allgemeines

Hier zeige ich eine der wichtigsten aktuellen Anwendungen des Satzes von Bayes: den Bayes-Filter zur Selektion von Spam aus der Email-Box. Im Gegensatz zu klassischen, auf "Schwarze" und "Weiße Listen" beruhenden Spamfiltern geht der Bayes Filter allein von der *statistischen Wahrscheinlichkeit* aus, mit der die in einer Email vorkommenden Worte bisher in Spam-Mails bzw. in erwünschten Mails ("Ham") vorkamen. Die einkommende Email wird also *in Worte* zerlegt und diese Wort für Wort analysiert. Ein "Wort" wird dabei sehr weit ausgelegt. Neben Worten im Text, dem Absender, Betreffzeilen etc werden auch Bildverweise, Links sowie Teile des Übertragungsprotokoll-Kauderwelsches analysiert.

Wer dazu näheres wissen will, dem sein auf den sehr guten Artikel in der c't 2003, Heft 17, S. 150 verwiesen.

Die E-Mail als "Zufallsexperiment"

Unser "Zufallsexperiment" ist das Eintreffen einer neuen Email mit den beiden möglichen Ereignissen

- S : "Email ist Spam",
- \bar{S} : "Email ist Ham", d.h. die E-Mail ist erwünscht.

Im Prinzip zerlegt der Bayes-Filter die Email in einzelne Worte $W_1 \dots W_n$, z.B.

$W_1 = \text{"Viagra"}$,
 $W_2 = \text{"Statistik"}$,
 $W_3 = \text{"Republic"}$,

und bestimmt die Spamwahrscheinlichkeit als bedingte Wahrscheinlichkeit

$$P_{\text{spam}} = P(S|W_1, W_2, \dots W_n).$$

Falls diese oberhalb eine Grenzwertes von z.B. 0.9 liegt, wird die Email als Spam klassifiziert, ansonsten nicht.

Frage: Warum wird die Grenz-Wahrscheinlichkeit so hoch gewählt?

Konkretes Beispiel mit einem Wort

In einer neu eintreffenden Email kommt das Wort $W_1 = \text{"Viagra"}$ vor. Zur Bestimmung der statistischen Wahrscheinlichkeiten liegen unserem Spamfilter 300 Emails, darunter 200 Spams zur Analyse vor. In 25% aller Spams kam bisher "Viagra" vor, aber auch in einer Ham-Mail von einem Freund ("Du, bekommst Du in letzter Zeit auch so viele Viagra-Angebote übers Netz?"). Wie groß ist die Spam-Wahrscheinlichkeit der neuen Email, wenn nur dieses eine Wort analysiert wird?

Lösung:

$$P_{\text{spam}} = P(S|W_1) \stackrel{\text{Bayes}}{=} \frac{P(S)P(W_1|S)}{P(W_1)}$$

Die "a-Priori" Wahrscheinlichkeiten $P(S)$ und $P(W_1)$ sowie die bedingte Wahrscheinlichkeit $P(W_1|S)$ bestimmen wir mit der *statistischen Definition* der Wahrscheinlichkeit aus den relativen Häufigkeiten der dem Filter zugänglichen E-Mails der Vergangenheit:

$$P(S) = \frac{200}{300}, \quad P(W_1) = \frac{0.25 * 200 + 1}{300} = \frac{51}{300}, \quad P(W_1|S) = 0.25$$

Damit

$$P_{\text{spam}} = P(S|W_1) = \frac{50}{51}$$

Obwohl "Viagra" nur in einem Viertel aller Spams vorkommt, beträgt dennoch im vorliegenden Mailbox die Bayes-Spamwahrscheinlichkeit einer konkreten Email, die dieses Wort enthält, 98%! Entscheidend ist hier, dass in Ham-Mails dieses Wort eben nur *sehr selten* vorkommt! Im Gegensatz zu den klassischen Filtern "lernt" der Bayes-Filter seine Wahrscheinlichkeiten aus den vergangenen Emails. Jeder Spamfilter ist damit individuell auf seinen "Meister" dressiert! Bei einem Urologen würde z.B. die obige Email sicher als "Ham" durchgehen.



Beispiel mit mehreren Worten

Was ist aber nun mit erwünschten Emails, die "Viagra" o.Ä. enthalten? Werden die auch irrtümlicherweise unter "Spam" abgelegt? Irrtümlich als "schlecht" klassifizierte "gute" Mails, sog. "false positives", stellen immerhin den schlimmsten Fehler dar, den Filter begehen können! Schauen wir uns folgende Email näher an: ¹

From: Arne Kesting <kesting@vwisb7.vkw.tu-dresden.de>

To: treiber <treiber@vwisb7.vkw.tu-dresden.de>

Subject: Übungsaufgaben

Hi Martin, das neue Statistik-Übungsblatt ist fertig. Übrigens: Wirst Du auch so mit Spams bombardiert, die z.B. Viagra oder ein einschlägiges 'Enlargment' anbieten oder - typischerweise von Nigeria aus - eine Million Dollar Gewinn versprechen?

Grüne Worte sind starke Indizien für Ham, rote für Spam. Wie im echten Leben gibt es also "mehrere Meinungen". Wie bestimmt man nun die Gesamt-Spamwahrscheinlichkeit? Zunächst mal enthält diese Email einige "100%ter": Noch nie kamen in einer Spam so spezifische Worte wie "kesting@vwisb7.vkw.tu-dresden.de", "Kesting" oder auch "Übungsblatt" vor. Damit ist diese Mail nach Bayes zu 100% "Ham" (*warum?*)

¹Der Text ist frei erfunden. Ähnlichkeiten mit tatsächlichen Begebenheiten, aktuellen oder vergangenen Emails sowie mit lebenden oder toten Personen sind rein zufällig;-)

Lassen wir nun aus "Sportlichkeit" diese Worte weg und untersuchen nur folgende Worte mit den jeweiligen relativen Häufigkeiten in den bisherigen Spam- und Ham-Mails:

Wort W_i	$P(W_i S)$	$P(W_i \bar{S})$
Viagra	50/200	1/100
Statistik	1/200	25/100

Der Bayes'sche Satz ergibt zunächst:

$$P_{\text{spam}} = P(S|W_1 \cap W_2) = \frac{P(S)P(W_1 \cap W_2|S)}{P(W_1 \cap W_2)}$$

Hier ist $P(W_1 \cap W_2|S)$ die bedingte Wahrscheinlichkeit dafür, dass in einer Spam-Mail die Worte "Viagra" und "Statistik" vorkommen. Macht man die "naive" Annahme, dass die Auftreffwahrscheinlichkeit $P(W_i)$ für ein Wort W_i nicht von anderen Wörtern W_j abhängt, gilt für bedingte Wahrscheinlichkeiten dasselbe Kriterium für Unabhängigkeit wie bei "einfachen" Wahrscheinlichkeiten:

$$P(W_1 \cap W_2|S) = P(W_1|S)P(W_2|S)$$

und damit

$$P_{\text{spam}} = \frac{P(S)P(W_1|S)P(W_2|S)}{P(W_1 \cap W_2)}.$$

Analog gilt für die "Ham"-Wahrscheinlichkeiten

$$P_{\text{ham}} = P(\bar{S}|W_1 \cap W_2) = \frac{P(\bar{S})P(W_1|\bar{S})P(W_2|\bar{S})}{P(W_1 \cap W_2)}.$$

Bildet man den Quotienten, kürzt sich jeweils der Nenner weg und man erhält

$$\frac{P_{\text{spam}}}{P_{\text{ham}}} = \frac{P(S)P(W_1|S)P(W_2|S)}{P(\bar{S})P(W_1|\bar{S})P(W_2|\bar{S})} \approx \frac{N_s}{N_{\bar{s}}} \left(\frac{N_{s,1}N_{\bar{s}}}{N_{\bar{s},1}N_s} \right) \left(\frac{N_{s,2}N_{\bar{s}}}{N_{\bar{s},2}N_s} \right)$$

Hierbei bedeuten

$N_{\bar{s}} = 100$ die Zahl der "guten" E-Mails,

$N_s = 200$ die Zahl der Spams,

$N_{s,1} = 50$ die Zahl der Spams, die "Viagra" enthalten,

$N_{\bar{s},1} = 1$ die Zahl der Nicht-Spams, die "Viagra" enthalten,

usw.

Damit ist $P_{\text{spam}}/P_{\text{ham}} = 1$ und wegen $P_{\text{spam}} + P_{\text{ham}} = 1$ die Spamwahrscheinlichkeit=50%, also kleiner als die Grenz-Wahrscheinlichkeit, so dass die Email als "Ham" angenommen wird.

Mit n Wörtern ist die Verallgemeinerung offensichtlich:

$$\frac{P(S)}{P(\bar{S})} \approx \frac{N_s}{N_{\bar{s}}} \prod_{i=1}^n \left(\frac{N_{s,i}N_{\bar{s}}}{N_{\bar{s},i}N_s} \right)$$

Bemerkungen

- Wegen der zusätzlichen "naiven" Annahme der Wort-Unabhängigkeit heißen die üblichen Bayes-Spamfilter auch "naive Bayes-Filter".
- Hauptvorteil gegenüber den klassischen Filtern ist die individuell erlernte "Whitelist", auf die ein Spammer, im Gegensatz zu "Schwarzen Listen", i.A. keine Reaktionsmöglichkeiten z.B. durch Wortverstümmelungen ("Vi@gra!") hat.