Bayesian Filter\nfighting Spam

Hauptseminar Machine Learning, October 2003
What is Spam?

• It is the act of blindly mass mailing a message that makes it spam, not the actual content.

• However, it seems that the language of spam constitutes a distinctive genre.

• Spam messages are often about topics rarely mentioned in legitimate messages.
Motivation

- Idea: Define rules that trigger some action
- ECA Rules: “Event-Condition-Action” — but who defines those rules?
What we want

- If it’s Spam, throw it away.

- but who decides what is Spam?
Let the machines decide!

- The computer makes the decision
- The user can help in the decision by training the machine in advance. “Offline Learning”
- The user can help in the decision by correction wrong decisions. “Online Learning”
More Classification Tasks

- Speech Recognition
- OCR (Optical Character Recognition)
- Biometric Sensors for authentication
- Quality control in production
- many more…
Abstract Model

- **Object Space**: All emails
- **Sensor Space**: Raw email text
- **Feature Space**: Tokens: word, phrases, sender, etc.
- **Decision Space**: Spam or not Spam

**Steps:**
- **Measurement**
- **Feature Extraction**
- **Decision**
Measurements vs. Features

• Measurements come directly from sensors, like CCD cameras, microphones, etc.
  • usually lots of data
  • contains a lot of unimportant data
• Features are extracted from raw data to reduce complexity
• assume we know nothing about an email, but we know that 20% of the mail we receive are spam.

• Then for a new email, we know with:

\[ P(\omega_1) = 0.2 \quad \omega_1 : \text{Mail is spam} \]
\[ P(\omega_2) = 0.8 \quad \omega_2 : \text{Mail is not spam} \]
a priory Probabilities (2)

- we decide:
  \[ \omega_1 \text{ if } P(\omega_1) > P(\omega_2) \]
  \[ \omega_2 \text{ if } P(\omega_1) \leq P(\omega_2) \]

- but this means that we classify every email not to be spam.
  Obviously, this is not what we want.
Feature Vector

• the Feature Vector contains all our extracted features.

• for example, count the occurrences of words in the email

\[ X = (x_1, x_2, \ldots, x_n) \]

• more on the choice of appropriate features later…
The a posteriori probability is the a conditional probability after a measurement:

\[ P(x_1, x_2, \ldots, x_n | \omega_i) \]

i.e. the probability of the occurrence of certain words in spam (and not spam)

but what we want is:

\[ P(\omega_i | x_1, x_2, \ldots, x_n) \]
Conditional Probabilities

- from highschool we know:

\[ P(A|B) = \frac{P(A \land B)}{P(B)} \]

which leads to

\[ P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \]

- This is called the “Bayes Formula”
Bayes Formula for Spam

\[ P(\text{Spam}|x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n|\text{Spam}) \cdot P(\text{Spam})}{P(x_1, \ldots, x_n)} \]

- A problem remains: How can we calculate \( P(x_1, \ldots, x_n|\text{Spam}) \)
- trivial solution: assume independence of the individual features
Naive Bayes

- Assuming independence, we can compute

\[
P(x_1, \ldots, x_n|\text{Spam}) = \frac{P(x_1, \ldots, x_n \land \text{Spam})}{P(\text{Spam})}
= \frac{P(x_1|x_2 \ldots, x_n \land \text{Spam}) \cdot P(x_2 \ldots, x_n \land \text{Spam})}{P(\text{Spam})}
= \frac{P(x_1|x_2 \ldots, x_n \land \text{Spam}) \cdot P(x_2 \ldots, x_n|\text{Spam}) \cdot P(\text{Spam})}{P(\text{Spam})}
= P(x_1|x_2 \ldots, x_n \land \text{Spam}) \cdot P(x_2 \ldots, x_n|\text{Spam})
= \ldots
= \prod_{i=1}^{n} P(x_i|x_{i+1} \ldots, x_n \land \text{Spam})
= \prod_{i=1}^{n} P(x_i|\text{Spam})
\]
Naive Bayes (2)

• Now we can build our classificator:
  We classify an email as spam, if

\[
\frac{P(Spam|x_1,\ldots,x_n)}{P(Ham|x_1,\ldots,x_n)} \geq \lambda
\]

• The choice of \(\lambda\) depends on the “cost” we imply on missclassification.
Loss Function

• Sometimes the cost of missclassification is different for different classes:
  • mistakenly deleting an important email is much worse than letting a spam mail slip through
  • selling a defect climbing rope is much worse than rejecting a good one in quality assurance.
Loss Function (2)

- Formally, we assign each class a cost by defining a cost function
  \[ \lambda(\alpha_i, \omega_j) \]

- The overall risk is then:
  \[ R(\alpha_i | \vec{x}) = \sum_{i=1}^{n} \lambda(\alpha_i, \omega_j) \cdot P(\omega_i | \vec{x}) \]
Loss Function (3)

- We decide for that class that gives the minimum risk given the observation.
- In the two-category case this is the same as applying a threshold

\[
\frac{P(Spam|x_1,\ldots,x_n)}{P(Ham|x_1,\ldots,x_n)} \geq \lambda
\]
Training Data

- We use our training data to compute the probabilities

\[
\frac{P(\text{Spam}|x_1, \ldots, x_n)}{P(\text{Ham}|x_1, \ldots, x_n)} = \frac{P(\text{Spam}) \cdot \prod_{i=1}^{n} p(x_i|\text{Spam})}{P(\text{Ham}) \cdot \prod_{i=1}^{n} p(x_i|\text{Ham})}
\]

\[
= \frac{N_{\text{Spam}} \cdot \prod_{i=1}^{n} \frac{N_{\text{Spam}, x_i}}{N_{\text{Spam}}}}{N_{\text{Ham}} \cdot \prod_{i=1}^{n} \frac{N_{\text{Ham}, x_i}}{N_{\text{Ham}}}}
\]

\[
= \frac{\prod_{i=1}^{n} N_{\text{Spam}, x_i}}{\prod_{i=1}^{n} N_{\text{Ham}, x_i}}
\]
Precision and Recall

• whenever we have a threshold value, we can write the precision and the recall as function of this parameter

• precision: the percentage of emails classified as spam that are in fact spam

• recall: the percentage of all spam emails that are correctly classified as spam
Precision and Recall

- example for precision/recall curve:
How to select Features

• How to select Features

• words, phrases, meta information: HTML messages, header fields, email address

• removing insignificant features: calculate the mutual information between each feature and the class.
Advanced Rules

Think about how much a new mortgage will save you!

You marked this message as Junk Mail.

- Message-Id: <07aa01c3958d$1e052340$b400a8c0@oemcomputer>
- Mime-Version: 1.0
- Content-Type: text/html; charset="ISO-8859-1"
- X-Priority: 3
- X-Mailer: mailer
- Abc-Tracking: <YmF1ZXJtYUBpbi50dW0uZGU->
- X-Spam-Checker-Version: SpamAssassin 2.60-tuminfo (1.212-2003-09-23-exp) on mailin1.informatik.tu-muenchen.de
- X-Spam-Status: No, hits=6.0 required=6.0 tests=COMPLETELY_FREE,
  HTML_FONTCOLOR_RED,HTML_FONT_BIG,HTML_FONT_INVISIBLE,HTML_MESSAGE,
  MIME_HTML_ONLY,RCVD_IN_BL_SPAMCOP_NET,RCVD_IN_DYNABLOCK autolearn=no version=2.60-tuminfo
- X-Spam-Level: *****
- X-Virus-Scanned: by amavisd-new at informatik.tu-muenchen.de

24118094o97zz859kp0h830qp8819 41057c56ma7q220a5f2510516277j

Take control of your money!

We do the work for you. By submitting your information across to hundreds of lenders, we can get you the best interest rates around.

Interest rates are lower than they have been in over 40 years, but it won't stay that way for long. Our simple form only takes a few moments, the OBLIGATION, and it's 100% FREE. You have nothing to lose, and everything to gain.

Get a free rnor.tgage quote today!
## Advanced Features

<table>
<thead>
<tr>
<th>Header</th>
<th>Description</th>
<th>Rule</th>
<th>Probability, Log Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>header</td>
<td>A domain registration spam subject</td>
<td>DOMAIN_SUBJECT</td>
<td>1.105 0.1650 0.1650</td>
</tr>
<tr>
<td>header</td>
<td>NO MX or A DNS records in the From header</td>
<td>NO_DNS_FOR_FROM</td>
<td>0 1.105 0.1650</td>
</tr>
<tr>
<td>header</td>
<td>Same From and To addresses, but not exactly</td>
<td>FROM_AND_TO_SAME</td>
<td>0.718 1.443 2.097 0.522</td>
</tr>
<tr>
<td>header</td>
<td>Received via a buggy SMTP server (MDaemon 2.7.4SP4R)</td>
<td>MDAEMON_2_7_4</td>
<td>2.900 2.800 2.800 2.700</td>
</tr>
<tr>
<td>header</td>
<td>Received: contains a forged HELO</td>
<td>FORGED_RCVD_HELO</td>
<td>1</td>
</tr>
<tr>
<td>header</td>
<td>Received: contains a numeric HELO</td>
<td>RCVD_NUMERIC_HELO</td>
<td>1.271 0.326 1.526 1.502</td>
</tr>
<tr>
<td>header</td>
<td>Received: contains a name with a fake IP-address</td>
<td>FAKED_IP_IN_RCVD</td>
<td>2.900 2.800 2.800 2.700</td>
</tr>
<tr>
<td>header</td>
<td>Received via SMTPD32 server (SMTPD32-n.n)</td>
<td>SMTPD_IN_RCVD</td>
<td>1</td>
</tr>
<tr>
<td>header</td>
<td>Lots and lots of Cc: headers</td>
<td>LOTS_OF_CC_LINES</td>
<td>2.900 2.800 0 0</td>
</tr>
<tr>
<td>header</td>
<td>Received forged, contains fake AOL relays</td>
<td>FORGED_AOL_RCVD</td>
<td>4.300 4.300 4.100 4.100</td>
</tr>
<tr>
<td>header</td>
<td>Contains forged hostname for a DSL IP in Brazil</td>
<td>FORGED_TELESP_RCVD</td>
<td>4.300 2.800 2.800 2.700</td>
</tr>
<tr>
<td>header</td>
<td>hotmail.com 'Received:' header found</td>
<td>FORGED_HOTMAIL_RCVD</td>
<td>0.470 0.001 0.500 0.500</td>
</tr>
<tr>
<td>header</td>
<td>hotmail.com 'From', address, but no 'Received:'</td>
<td>FORGED_HOTMAIL_RCVD2</td>
<td>0.051 1.884 2.499</td>
</tr>
<tr>
<td>header</td>
<td>Forged eudoramail.com 'Received:' header found</td>
<td>FORGED_EUDORAMAIL_RCVD</td>
<td>2.799 2.796 2.696 2.700</td>
</tr>
<tr>
<td>header</td>
<td>'From' yahoo.com does not match 'Received' headers</td>
<td>FORGED_YAHOO_RCVD</td>
<td>0.375 0.477 1.181 0.901</td>
</tr>
<tr>
<td>header</td>
<td>'From' junocom does not match 'Received' headers</td>
<td>FORGED_JUNO_RCVD</td>
<td>1.538 2.796 2.696 2.058</td>
</tr>
<tr>
<td>header</td>
<td>Forged 'by gw05' 'Received:' header found</td>
<td>FORGED_GW05_RCVD</td>
<td>2.900 2.800 2.800 2.700</td>
</tr>
<tr>
<td>header</td>
<td>Hotmail.com Received 'from mx' header</td>
<td>FORGED_MX_HOTMAIL</td>
<td>2.900 2.800 2.800 2.700</td>
</tr>
<tr>
<td>header</td>
<td>Sent by a known spampaus (qv.es)</td>
<td>RCVD_BY_QVES.COM</td>
<td>2.900 0 0</td>
</tr>
<tr>
<td>header</td>
<td>Character set doesn’t exist</td>
<td>NONEXISTENT_CHARSET</td>
<td>2.900 2.800 2.800 2.700</td>
</tr>
<tr>
<td>header</td>
<td>A language charset used in headers</td>
<td>CHARSET_FARAWAY_HEADER</td>
<td>3.200</td>
</tr>
<tr>
<td>header</td>
<td>'X-Mailer' line contains gibberish</td>
<td>X_MAILER_GIBBERISH</td>
<td>1</td>
</tr>
<tr>
<td>header</td>
<td>Sent with 'X-Priority' set to high</td>
<td>X_PRIORITY_HIGH</td>
<td>1.495 0.516 1.486 1.305</td>
</tr>
<tr>
<td>header</td>
<td>Sent with 'X-MSmaid-Priority' set to high</td>
<td>X_MSMAIL_PRIORITY_HIGH</td>
<td>0.500 0.501 0.501 0.500</td>
</tr>
</tbody>
</table>
Advanced Features

Content analysis details:  (23.6 points, 6.0 required)
3.2 FROM_HAS_MIXED_NUMS From: contains numbers mixed in with letters
0.3 FROM_HAS_MIXED_NUMS From: contains numbers mixed in with letters
1.0 ACCEPT_CREDIT_CARDS BODY: Accept Credit Cards
0.5 CLICK_BELOW_CAPS BODY: Asks you to click below (in capital letters)
0.6 FOR_FREE BODY: No such thing as a free lunch (1)
5.4 BAYES_99 BODY: Bayesian spam probability is 99 to 100%

[score: 1.0000]
0.3 MIME_HTML_ONLY BODY: Message only has text/html MIME parts
0.1 HTML_MESSAGE BODY: HTML included in message
0.5 HTML_LINK_CLICK_CAPS BODY: HTML link text says "CLICK"
0.1 HTML_LINK_CLICK_HERE BODY: HTML link text says "click here"
0.6 MIME_HTML_NO_CHARSET RAW: Message text in HTML without charset
0.5 REMOVE_PAGE URI: URL of page called "remove"
2.6 SUSPICIOUS_RECIP BODY: Similar addresses in recipient list
0.7 RCVD_IN_DSBL RBL: Received via a relay in list.dsbl.org
[<http://dsbl.org/listing?ip=212.214.158.101>]
1.5 RCVD_IN_BL_SPAMCOP_NET RBL: Received via a relay in bl.spamcop.net
[Blocked - see <http://www.spamcop.net/bl.shtml?212.214.158.101>]
2.6 FORGED_MUA_OUTLOOK Forged mail pretending to be from MS Outlook
1.0 FORGED_OUTLOOK_TAGS Outlook can't send HTML in this format
0.0 UPPERCASE_25_50 message body is 25-50% uppercase
1.0 FORGED_OUTLOOK_HTML Outlook can't send HTML message only
1.1 MIME_HTML_ONLY_MULTI Multipart message only has text/html MIME parts
Spammers are fighting back

- HTML tricks:
  - Make mo<foo>ney f<bar>ast
  - used background color to hide words
    - include non-spam words
  - using their own Bayes classifier to test the spam, trying to make it indistinguishable from legitimate email
Simple Example

The demise of my hamster made me cry!

Take control of your money!

We do the work for you. By submitting your information across to hundreds of lenders, we can get you the best interest rates around.

All your efforts to be me have been futile! I rule!

Interest rates are lower than they have been in over 40 years, but it won't stay that way for long. Our simple form only takes a few moments, there is absolutely NO OBLIGATION, and it's 100% FREE. You have nothing to lose, and everything to gain.

Get a free mortgage quote today!

To get off our list, unsubscribe here.
One more Idea...

- Spam email can be roughly divided in two subgroups:
  - pornographic
  - other spam
- what about classifying into three classes instead of two classes?
... that did not work out.

- Tests showed that the combined classifier using porn-spam, other spam and legitimate emails had an overall worse performance

- The reasons for this are:
  - a model with more degrees of freedom must fit many more parameters from the data, and additionally
  - less data for each class is available
Further Reading

• Dr. Andreas Linke
  Spam oder nicht Spam? E-Mail sortieren mit Bayes-Filtern;
c't 17/03, Seite 150

• Mehran Sahami, Susan Dumais, David Heckerman, Eric Horvitz, Microsoft Research

• Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos, George Paliouras, Constantine D. Spyropoulos
  An Evaluation of Naive Bayesian Anti-Spam Filtering (2000)

• Jefferson Provost
  Naive-Bayes vs. Rule-Learning in Classification of Email