Princeton University
COS 217: Introduction to Programming Systems
Spam Filter Knowledge Module: Mathematical Foundations

First, let's define abbreviations for some logical assertions:

f_i: The message contains feature i.
$\sim f_j$: The message does not contain feature j.
ham: The message is ham.
spam: The message is spam.

We wish to compute:

(1)

$P(ham \mid f_1,...,f_m,\sim f_{m+1},...,\sim f_n)$

That is, the probability that the message is ham, given that the message contains some features and does not contain others.

$P(spam \mid f_1,...,f_m,\sim f_{m+1},...,\sim f_n)$

That is, the probability that the message is spam, given that the message contains some features and does not contain others.

That notation is cumbersome. So let's abbreviate "$f_1,...,f_m,\sim f_{m+1},...,\sim f_n$" as "$f_1,...,\sim f_n$". So we wish to compute:

(2)

$P(ham \mid f_1,...,\sim f_n)$

$P(spam \mid f_1,...,\sim f_n)$

But we don't know how to compute those probabilities. So let's apply some mathematics...

Since the AND operator is commutative:

$$P(f_1,...,\sim f_n \text{ AND ham}) = P(\text{ham AND } f_1,...,\sim f_n)$$

By the multiplicative law of probability, $P(x \text{ AND } y) = P(x)\ P(y \mid x)$. And so:

$$P(f_1,...,\sim f_n)\ P(\text{ham} \mid f_1,...,\sim f_n) = P(\text{ham})\ P(f_1,...,\sim f_n \mid \text{ham})$$

Dividing both sides of the equation by $P(f_1,...,\sim f_n)$, we get **Bayes' Rule**:

$$P(\text{ham} \mid f_1,...,\sim f_n) = P(\text{ham})\ P(f_1,...,\sim f_n \mid \text{ham}) / P(f_1,...,\sim f_n)$$

Similarly:

$$P(f_1,...,\sim f_n \text{ AND spam}) = P(\text{spam AND } f_1,...,\sim f_n)$$
$$P(f_1,...,\sim f_n)\ P(\text{spam} \mid f_1,...,\sim f_n) = P(\text{spam})\ P(f_1,...,\sim f_n \mid \text{spam})$$
$$P(\text{spam} \mid f_1,...,\sim f_n) = P(\text{spam})\ P(f_1,...,\sim f_n \mid \text{spam}) / P(f_1,...,\sim f_n)$$

Substituting for the above expressions, we wish to compute:

(3)

$$\frac{P(\text{ham})\ P(f_1,...,\sim f_n \mid \text{ham})}{P(f_1,...,\sim f_n)}$$

$$\frac{P(\text{spam})\ P(f_1,...,\sim f_n \mid \text{spam})}{P(f_1,...,\sim f_n)}$$

We don't know P(ham) or P(spam), and we never will. We'll need to guess them, based upon our perception of the proportion of our e-mail that is, in fact, spam.

$P(f_1,...,\sim f_n \mid \text{ham})$ and $P(f_1,...,\sim f_n \mid \text{spam})$ would be difficult to compute. Doing so would be possible only if we have examples of messages that contain (or do not contain) every combination of features. That would imply that we need a very large number of examples.

So we must make a simplifying assumption. Let's assume that $f_1,...,f_n$ are independent. Then:

$$P(f_1,\ldots,\sim f_n) = P(f_1)\ldots P(\sim f_n)$$

and

$$P(f_1,\ldots,\sim f_n \mid ham) = P(f_1 \mid ham)\ldots P(\sim f_n \mid ham)$$
$$P(f_1,\ldots,\sim f_n \mid spam) = P(f_1 \mid spam)\ldots P(\sim f_n \mid spam)$$

Substituting into expressions (3), we wish to compute:

(4)

$$\frac{P(ham)\ P(f_1 \mid ham)\ldots P(\sim f_n \mid ham)}{P(f_1)\ldots P(\sim f_n)}$$

$$\frac{P(spam)\ P(f_1 \mid spam)\ldots P(\sim f_n \mid spam)}{P(f_1)\ldots P(\sim f_n)}$$

We don't know how to compute $P(f_i)$ or $P(\sim f_i)$ for any i. So we need to transform the denominators.

Certainly, any message is either ham or spam. So:

$$P(ham \mid f_1,\ldots,\sim f_n) + P(spam \mid f_1,\ldots,\sim f_n) = 1$$

After substituting expressions (4) into that equation:

$$\frac{P(ham)\ P(f_1 \mid ham)\ldots P(\sim f_n \mid ham)}{P(f_1)\ldots P(\sim f_n)} + \frac{P(spam)\ P(f_1 \mid spam)\ldots P(\sim f_n \mid spam)}{P(f_1)\ldots P(\sim f_n)} = 1$$

After multiplying both sides of the equation by $P(f_1)\ldots P(\sim f_n)$:

$$P(f_1)\ldots P(\sim f_n) = P(ham)\ P(f_1 \mid ham)\ldots P(\sim f_n \mid ham) + P(spam)\ P(f_1 \mid spam)\ldots P(\sim f_n \mid spam)$$

Substituting into expressions (4), we wish to compute:

$$\text{(5)}\quad \frac{P(ham)\,P(f_1 \mid ham)...P(\sim f_n \mid ham)}{P(ham)\,P(f_1 \mid ham)...P(\sim f_n \mid ham) + P(spam)\,P(f_1 \mid spam)...P(\sim f_n \mid spam)}$$

$$\frac{P(spam)\,P(f_1 \mid spam)...P(\sim f_n \mid spam)}{P(ham)\,P(f_1 \mid ham)...P(\sim f_n \mid ham) + P(spam)\,P(f_1 \mid spam)...P(\sim f_n \mid spam)}$$

As noted previously, the sum of those two expressions is 1. So it would be sufficient to compute only one of them; we easily could derive the other. So, let's compute only the second expression:

$$\text{(6)}\quad \frac{P(spam)\,P(f_1 \mid spam)...P(\sim f_n \mid spam)}{P(ham)\,P(f_1 \mid ham)...P(\sim f_n \mid ham) + P(spam)\,P(f_1 \mid spam)...P(\sim f_n \mid spam)}$$

Note that we can estimate all components of that expression. Specifically, we can estimate:

- P(ham) based upon our perception of the proportion of our e-mail that is, in fact, ham.
- P(spam) based upon our perception of the proportion of our e-mail that is, in fact, spam. Note that P(ham) + P(spam) = 1.
- $P(f_i \mid ham)$ by examining many ham messages, and determining the proportion of them that contain feature f_i.
- $P(f_i \mid spam)$ by examining many spam messages, and determining the proportion of them that contain feature f_i.
- $P(\sim f_i \mid ham)$ by examining many ham messages, and determining the proportion of them that do not contain feature f_i. Or we could compute it as $1 - P(f_i \mid ham)$.
- $P(\sim f_i \mid spam)$ by examining many spam messages, and determining the proportion of them that do not contain feature f_i. Or we could compute it as $1 - P(f_i \mid spam)$.

So, in ***theory***, we can use expression (6) to produce the results that we wish.

However, in ***practice*** the products may become very small (i.e. close to 0), and thus cause may cause floating-point underflow. So, relying upon the equality:

$$x * y = \exp(\log(x) + \log(y))$$

let's compute sums of logarithms instead of products. That is, we wish to compute:

$$\frac{\exp(\log(P(spam)) + \log(P(f_1 \mid spam)) + \ldots + \log(P(\sim f_n \mid spam)))}{\exp(\log(P(ham)) + \log(P(f_1 \mid ham)) + \ldots + \log(P(\sim f_n \mid ham))) + \exp(\log(P(spam)) + \log(P(f_1 \mid spam)) + \ldots + \log(P(\sim f_n \mid spam)))} \tag{7}$$

The logarithms will be negative. So the sums of the logarithms may be large negative numbers (i.e. far from 0). So applying the exp operation to those sums may cause precisely the same floating-point underflow that motivated us to use logarithms in the first place.

Relying upon this equality:

$$\frac{\exp(\log(x))}{\exp(\log(x)) + \exp(\log(y))} = \frac{\exp(\log(x) + k)}{\exp(\log(x) + k) + \exp(\log(y) + k)}$$

let's add some number k to each sum-of-logs before applying the exp operator. So we wish to compute:

$$\frac{\exp(\log(P(spam)) + \log(P(f_1 \mid spam)) + \ldots + \log(P(\sim f_n \mid spam)) + k)}{\exp(\log(P(ham)) + \log(P(\sim f_1 \mid ham)) + \ldots + \log(P(\sim f_n \mid ham)) + k) + \exp(\log(P(spam)) + \log(P(f_1 \mid spam)) + \ldots + \log(P(\sim f_n \mid spam)) + k)} \tag{8}$$

For k, a good choice would be:

$$-(\log(P(spam)) + \log(P(f_1 \mid spam)) + \ldots + \log(P(\sim f_n \mid spam)))$$

that is,

$$-\log(P(spam)) - \log(P(f_1 \mid spam)) - \ldots - \log(P(\sim f_n \mid spam))$$

It's a good choice because it makes the numerator equal exp(0), that is, 1. It also makes the second term of the denominator equal exp(0), that is, 1. So, we wish to compute:

$$\frac{1}{\exp(\ \log(P(ham)) + \log(P(f_1 \mid ham)) + \ldots + \log(P(\sim f_n \mid ham)) - \log(P(spam)) - \log(P(f_1 \mid spam)) - \ldots - \log(P(\sim f_n \mid spam))\ ) + 1} \qquad (9)$$

Note that only one sum-of-logs remains, and it will not evaluate to a large negative number (i.e. far from 0). So applying the exp operation to that sum will not cause floating-point underflow. Thus we have an expression that is correct in theory and in practice.