Combining Evidence using Bayes' Rule

Scott D. Anderson

February 26, 2007

This document explains how to combine evidence using what's called naïve Bayes: the assumption of conditional independence (even though we might know that the data aren't exactly conditionally independent). So, the probability we get won't be *accurate*, but it should at least be a probability and should correlate with the information we want, namely the probability that a message is spam.

I'm basing all this on Russell and Norvig's AI book, section 14.4 (first edition), plus personal communication with David D. Lewis (`http://daviddlewis.com`). I've re-written this document thanks to an email from Ethan Herdrick, who helpfully pointed out that my previous document was not clear.

The context of this problem is spam filters, an honors thesis conducted by Sara "Scout" Sinclair under my supervision. We want to train a Bayesian classifier to classify email. Let's start with an example:

	ham	spam	total
all messages	400	600	1000
with "free"	100	300	400
with "viagra"	10	90	100

The basic application of Bayes' rule allows us to calculate the probability that a message is spam given that it contains any one token.

$$\begin{aligned} P(spam|token) &= P(spam)\frac{P(token|spam)}{P(token)} \\ P(spam|free) &= P(600/1000)\frac{P(300/600)}{P(400/1000)} = 0.6\frac{0.5}{0.4} = 0.75 \\ P(spam|viagra) &= P(600/1000)\frac{P(90/600)}{P(100/1000)} = 0.6\frac{0.15}{0.1} = 0.90 \end{aligned}$$

Our prior probability of spam (given the training data) is 0.6, and if we see a message containing the word "free," we bump that up to 0.75 and if we see "viagra" we bump it up to 0.90.

1 Multiple Evidence

The question is how to combine multiple pieces of evidence. That is, if I see a message with both "free" and "viagra," what is my probability calculation?

Translating Russell and Norvig's example (spam=cavity, toothache=free, catch=viagra), I start with the following equation, which doesn't assume conditional independence. This equation is a straightforward application of Bayes' rule for two pieces of evidence ("free" and "viagra"), and is isomophic to the one in the middle of page 428 of Russell and Norvig:

$$P(spam|free \wedge viagra) = \frac{P(free \wedge viagra|spam)P(spam)}{P(viagra \wedge free)} \tag{1}$$

There are several problems with this equation. The first is the denominator: we are not going to record and train on all subsets of words (let's stipulate that), so the probability of "viagra" co-occuring with "free" is unknown. The same problem is on the numerator, where we would need to know the probability of that pair of terms co-occurring in a spam message.

2 Conditional Independence

One approach is to make the assumption of *conditional independence*. (Russell and Norvig describe this on page 429; thanks to David Lewis for explaining normalization to me.)

Conditional independence means that once you know one piece of information, other features become independent. One classic example is that spelling ability and shoe size are not independent: people with larger feet spell better than people with smaller feet. The missing piece of information is *age*: older kids have larger feet and better spelling. Once you know a child's age, their spelling ability and shoe size are unrelated (independent). When two features are conditionally independent, we can calculate their co-occurrence as a simple multiplication. The general statement is as follows:

$$P(X, Y|Z) = P(X|Z)P(Y|Z) \tag{2}$$

For the spam problem, our assumption is that the occurrence of the words "free" and "viagra" become independent once we know whether the message is spam. (Again, this assumption is probably wrong, but we make it anyhow, because we won't count how many times the words co-occur.)

Now, we make our assumption of conditional independence. Applying equation (2) to the numerator of equation (1), we get:

$$P(viagra \wedge free|spam) = P(viagra|spam)P(free|spam) \tag{3}$$

In words, this means that for spam messages, we expect "viagra" and "free" to be independent, so the probability of their co-occurrence in a spam message is just the product of their conditional probabilities. (You may or may not agree with the assumption, but that's what it means.) Thus equation (1) becomes:

$$P(spam|free \wedge viagra) = \frac{P(viagra|spam)P(free|spam)P(spam)}{P(viagra \wedge free)} \tag{4}$$

We have in our database everything except the second denominator, $P(viagra \wedge free)$, Russell and Norvig explain that, we can eliminate this term with *normalization*, which uses the conditional probabilities and the assumption of conditional independence to calculate this term.

The derivation takes several steps, so be patient. First, we state Bayes' rule for two pieces of evidence, as in equation (1), once for each kind of message:

$$\begin{aligned} P(spam|free \wedge viagra) &= \frac{P(free \wedge viagra|spam)P(spam)}{P(free \wedge viagra)} \\ P(ham|free \wedge viagra) &= \frac{P(free \wedge viagra|ham)P(ham)}{P(free \wedge viagra)} \end{aligned}$$

The two equations sum to 1, since the message is certainly either ham or spam. (This idea can be generalized; see Russell and Norvig.) We can then multiply the whole equation by the common denominator and the left hand side is exactly what we want, namely the unknown denominator in equation (4).

$$\begin{aligned} 1 &= \frac{P(free \wedge viagra|spam)P(spam)}{P(free \wedge viagra)} + \frac{P(free \wedge viagra|ham)P(ham)}{P(free \wedge viagra)} \\ P(free \wedge viagra) &= P(free \wedge viagra|spam)P(spam) + P(free \wedge viagra|ham)P(ham) \end{aligned}$$

This replaces the calculation of the joint probability $P(free \wedge viagra)$, which we don't know, with a calculation involving conditional probabilities. We can proceed by employing, once again, the assumption of conditional independence. Thus:

$$P(free \wedge viagra) = P(free|spam)P(viagra|spam)P(spam) + P(free|ham)P(viagra|ham)P(ham)$$

This, then, is the desired denominator for our probability calculation. Note that the first term is the same as our numerator, the other term is the analogous calculation conditioned on ham rather than spam. The final formula, then, for two pieces of evidence is:

$$P(spam|free \wedge viagra) = \frac{P(spam)P(free|spam)P(viagra|spam)}{P(spam)P(free|spam)P(viagra|spam) + P(ham)P(free|ham)P(viagra|ham)} \quad (5)$$

3 Example

With the example we have, we can compute:

$$\begin{aligned} P(spam|free \wedge viagra) &= \frac{P(spam)P(free|spam)P(viagra|spam)}{P(spam)P(free|spam)P(viagra|spam) + P(ham)P(free|ham)P(viagra|ham)} \\ &= \frac{(600/1000)(300/600)(90/600)}{(600/1000)(300/600)(90/600) + (400/1000)(100/400)(10/400)} \\ &= \frac{.045}{.045 + .0025} \\ &= 0.95 \end{aligned}$$

With "free," we computed the probability that the message was spam to be 75 percent, and with "viagra," we computed a probability of 90 percent, but, with both tokens, the probability that the message is spam goes up to 95 percent.

4 General Combinations

Let's look at how this generalizes to many tokens.

	ham	spam
all messages	H	S
with "free"	h_{free}	s_{free}
with "viagra"	h_{viagra}	s_{viagra}
with $token_i$	h_i	s_i

Also, let $M = H + S$, which is just the total number of training messages.

$$\begin{aligned} P(spam| \wedge_{i=1}^{n} token_i) &= \frac{P(spam)\prod P(token_i|spam)}{P(spam)\prod P(token_i|spam) + P(ham)\prod P(token_i|ham)} \\ &= \frac{(S/M)\prod(s_i/S)}{(S/M)\prod(s_i/S) + (H/M)\prod(h_i/H)} \end{aligned}$$

At this point, we can do some algebraic simplifications (such as eliminating the $1/M$ in the numerator and denominator), but the basic calculation is clear.

5 Bayes vs Graham

I now think we can understand the difference between our calculation, which we believe is correct, and Paul Graham's. His formula looks similar but has some key differences.

Let $p_i = s_i/S$ be the conditional probability that a message contains token i, given that the message is spam and $q_i = h_i/H$ be the conditional probability that a message contains token i given that the message is ham. Our formula is

$$\frac{(S/M)p_1 \ldots p_n}{(S/M)p_1 \ldots p_n + (H/M)q_1 \ldots q_n}$$

While Graham's formula is:

$$\frac{p_1 \ldots p_n}{p_1 \ldots p_n + (1-p_1) \ldots (1-p_n)}$$

We see two differences. First, his formula omits the prior probabilities, or, more precisely, he assumes that the prior probability of spam and ham are equal at 0.5, so they cancel. This is a reasonable, defensible position. The other difference is that his formula supposes that $q_i = 1 - p_i$, which is not true. In our example, for instance, $P(viagra|spam) = 90/600 = 0.15$ and $P(viagra|ham) = 10/400 = 0.025$.