1 Introductory Comments

First, I would like to point out that I got this material from two sources: The first was a page from Paul Graham's website at www.paulgraham.com/ffb.html, and the second was a paper by I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos, titled *An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages*, which appeared in the Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pages 160-167). The Graham paper is interesting, but is written more for those with almost no mathematical background, and it doesn't explain the math behind the algorithm; and, even though Graham's paper gives a link to a page describing the math, *that* linked page also does not do an adequeate job, since it does not place the result proved and used in its proper Bayesian context. Here in these notes I will give a more formal treatment, and will be explicit about the "conditional independence" assumptions that one makes.

2 Bayesian Probability

In this section I will prove a few basic results that we will use. Some of these results are proved in your book, but I will prove them here again anyway, to make these notes self-contained.

First, we have Bayes's Theorem:

Theorem (Bayes's Theorem). Suppose that S is a sample space, and Σ is a σ-algebra on S having probability measure P. Further, suppose that we have a partition of S into (disjoint) events $C_1, C_2, ..., C_k$; that is,

$$S = \bigcup_{i=1}^{k} C_i, \quad \text{and,} \quad \text{for } i \neq j,\ C_i \cap C_j = \emptyset.$$

Then, for any $i = 1, 2, ..., k$, we have

$$P(C_i|A) = \frac{P(A|C_i)P(C_i)}{\sum_{j=1}^{k} P(A|C_j)P(C_j)}.$$

Proof. The proof is really obvious, once you know what everything means. First, we note that A can be partitioned as follows:

$$A \;=\; \bigcup_{j=1}^{k} A \cap C_j,$$

where we notice that the sets $A \cap C_j$ are all disjoint, since the sets C_j are disjoint. Thus, we have

$$P(A) \;=\; \sum_{j=1}^{k} P(A \cap C_j) \;=\; \sum_{j=1}^{k} P(A|C_j)P(C_j). \tag{1}$$

The second equality here just follows from the *definition* of conditional probability

$$P(C|D) \;=\; \frac{P(C \cap D)}{P(D)}.$$

Now, then, via obvious manipulations we get

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

and then replacing the $P(A)$ on the denominator with the right-hand-side in (1) the theorem now follows. ■

The theorem we will actually use is what I will call the "Spam Theorem". The setup is as follows: Suppose we have two classes of objects, spam and Legitimate email. We will use "Spam" to denote the spam messages, and "Legit" to denote the legitimate ones. We also suppose we have a series of k attributes that a Spam or Legit message can have. Our sample space for the probability theory will be as follows: S will be all ordered $(k+1)$-tuples of the form

$$(\text{Spam}, t_1, ..., t_k) \quad \text{or} \quad (\text{Legit}, t_1, ..., t_k),$$

where each t_i is 1 or 0; $t_i = 0$ means the message *does not* have attribute i, and $t_i = 1$ means it *does* have attribute i. In total there are 2^{k+1} elements in S.

These elements of S you can think of as the space of all possible message attributes. For example, if $k = 3$, a typical element of S looks like

(Spam, 1, 0, 1), which would mean that you have a Spam message having attributes 1 and 3, but not attribute 2.

These k attributes you can think of as corresponding to whether or not the message contains a particular word. So, for example, attribute 1 might be "The message contains the word SAVE", attribute 2 might be "The message contains the word MONEY", and attribute 3 might be "The message contains the word NOW". Thus, a message having attribute vector (Spam, 1, 0, 1) would be a spam email containing the words SAVE and NOW, but not the word MONEY.

We assume that we have some probability measure P on $\Sigma = 2^S$. Now, we let "Spam" denote the subset of S consisting of all vectors with "Spam" for the first entry; and we let "Legit" denote the vectors in S that begin with "Legit". We let W_i denote the subset of S consisting of all vectors with $t_i = 1$; that is, you can think of a message as having an attribute vector lying inside W_i if it is a message that contains the ith word somewhere in its text.

We further make the following, crucial conditional independence assumptions:

Conditional Independence Assumptions. Suppose that $W_{i_1}, ..., W_{i_\ell}$ are distinct words chosen from among $W_1, ..., W_k$. Then, we assume that

$$P(W_{i_1} \cap W_{i_2} \cap \cdots \cap W_{i_\ell} \mid Spam) = \prod_{j=1}^{\ell} P(W_{i_j} \mid Spam); \quad \text{and}$$

$$P(W_{i_1} \cap W_{i_2} \cap \cdots \cap W_{i_\ell} \mid Legit) = \prod_{j=1}^{\ell} P(W_{i_j} \mid Legit).$$

Comments: The first assumption here roughly says that for any given piece of spam, the probability that that message contains any given word on our list is "independent" of the probability that it contains any other combination of the other words on our list. This is maybe not a valid assumption, since for instance, if a spam contains the word "SAVE", it very likely contains the word "MONEY"; so, these probabilities are not independent, or "uncorrelated". Nonetheless, we will assume that these conditions hold in order to make our model simple. The second assumption has a similar interpretation.

Now we come to our theorem

Spam Theorem. Suppose that S, Σ, and P as above, and let $p_i = P(Spam|W_i)$. Further, suppose that the conditional independence assumption holds. Then, for any subset $W_{i_1}, ..., W_{i_\ell}$ of $W_1, ..., W_k$ we have that

$$P(Spam \mid W_{i_1} \cap \cdots \cap W_{i_\ell}) \;=\; \frac{p_{i_1} \cdots p_{i_\ell}}{p_{i_1} \cdots p_{i_\ell} + \left(\frac{x}{1-x}\right)^{\ell-1} (1-p_{i_1}) \cdots (1-p_{i_\ell})},$$

where $x = P(Spam)$.

This theorem is telling us that if we are given a document containing the words $W_{i_1}, ..., W_{i_\ell}$, then we can calculate the probability that it is a spam message by plugging in the respective probabilities into the formula. How do we determine the probabilities p_i? That will be described in the next section. For now we will just assume we know what they are.

Let us now see how to prove the theorem.

Proof. We apply Bayes's Theorem with the disjoint events $C_1 = Spam$, and $C_2 = Legit = \overline{C_1}$ (which partition S), and with $A = W_{i_1} \cap \cdots \cap W_{i_\ell}$. With these choices of parameters we get that the numerator in the formula for $P(C_1|A)$ in Bayes's Theorem has the value

$$\begin{aligned} P(A|C_1)P(C_1) &= P(C_1)\prod_{j=1}^{\ell} P(W_{i_j}|C_1) \;=\; P(C_1)\prod_{j=1}^{\ell} P(C_1|W_{i_j})\frac{P(W_{i_j})}{P(C_1)} \\ &= \frac{1}{x^{\ell-1}}\prod_{j=1}^{\ell} p_{i_j} P(W_{i_j}). \end{aligned}$$

To get this expansion we have used our conditional independence assumption, together with the basic fact that $P(F|G) = P(G|F)P(F)/P(G)$. Note here that we also made the substitution $P(C_1) = P(Spam) = x$.

To get the denominator in Bayes's formula for $P(C_1|A)$, we have to find $P(A|C_1)P(C_1) + P(A|C_2)P(C_2)$. We have already found the first term here; so, we just need to find the second term, which is

$$\begin{aligned} P(A|C_2)P(C_2) &= P(C_2)\prod_{j=1}^{\ell} P(W_{i_j}|C_2) \;=\; P(C_2)\prod_{j=1}^{\ell} P(C_2|W_{i_j})\frac{P(W_{i_j})}{P(C_2)} \\ &= \frac{1}{(1-x)^{\ell-1}}\prod_{j=1}^{\ell} P(W_{i_j})(1-p_{i_j}). \end{aligned}$$

So, we have

$$\begin{aligned} P(C_1|A) &= \frac{\frac{1}{x^{\ell-1}}\prod_{j=1}^{\ell} p_{i_j} P(W_{i_j})}{\frac{1}{x^{\ell-1}}\prod_{j=1}^{\ell} p_{i_j} P(W_{i_j}) \;+\; \frac{1}{(1-x)^{\ell-1}}\prod_{j=1}^{\ell}(1-p_{i_j})P(W_{i_j})} \\ &= \frac{p_{i_1}\cdots p_{i_\ell}}{p_{i_1}\cdots p_{i_\ell} + \left(\frac{x}{1-x}\right)^{\ell-1}(1-p_{i_1})\cdots(1-p_{i_\ell})}. \end{aligned}$$

The theorem now follows. ■

3 Applying the Spam Theorem to do the Filtering

The idea is to start with, say 1000 messages, with 500 of them spam, and 500 legitimate. Then, you find the "best" 20 words which appear in a significant percentage of the spam and legit messages. By "best" I mean here that the word is biased towards being in either a spam or legit message. For example, if I were told that I am about to receive a message containing the word "sell", and if I know that I typically get as many spams as legit emails, then there is a better than 50 percent chance that the message is spam. On the other hand, if I was told the message contains the word "the", then that would tell me little about whether the message is spam or legit; that is, I would only know the message is spam with 50 percent chance. So, the word "the" would not go on my list of "best" words, but the word "sell" just might. The correct balance between how often a word appears in emails, and how biased it is to being spam or legit, in determining whether the word is one of the "best words", is left up to the user.

After we have our list of words, we then compute the numbers $p_i = P(Spam|W_i)$ as follows: Given a word W_i, let N be the total number of my 500 spam emails containing W_i, and let M be the total number of my 500 legit emails containing W_i. Then, we let $P(W_i|Spam) = N/500$ and $P(W_i|Legit) = M/500$. The quantity $P(Spam)$ is the parameter x in the spam theorem, which is left up to the user to supply. Finally, by Bayes's Theorem we have

$$p_i = \frac{P(W_i|Spam)P(Spam)}{P(W_i|Spam)P(Spam) + P(W_i|Legit)P(Legit)} = \frac{xN}{xN + M(1-x)}.$$

Now suppose you are given some message. It is not too difficult to have a computer program read through the message to determine which of our 20 best words appear in the message. After this is done, suppose that the words appearing in the message, which are among our 20 best, are $W_{i_1}, ..., W_{i_\ell}$. Then, the computer can calculate the probability that that message is a spam using the spam theorem. If the chance that the message is spam is sufficiently high, then you can have the computer reject the message or to put it into some kind of "holding folder", which you can look through from time to time, just in case a legit message was misclassified as spam.