Conditional Probability - Part 4
by Gerald Fitton

After the flood, Noah said to all the animals “Go forth and multiply” - which most did. Certainly we, (mankind) did! Later he came across a couple of snakes basking in the beautiful sunshine and asked them why they were not multiplying. Their answer was, “We can’t multiply - we’re adders”. So Noah cut up some timber and made a table out of logs. Then he placed the adders on the log table. The rest (as they say) is History.

The moral of this story is, “Even adders can multiply using log tables!”

In this month’s article I shall discuss Bayesian Spam Filters (and the use of logs to multiply) but, as usual, I shall start somewhere else!

One Hypothesis

The favoured method of statistical analysis for Scientists (when trying to decide whether a new theory is worth using) is called Hypothesis Testing. When using this Hypothesis Testing method of analysis, there is only one Hypothesis to consider. We use the statistical data we have collected to decide whether that Hypothesis is probable or improbable. What we hope will happen is that the data we have collected will prove that the Hypothesis we are testing is highly improbable! Then what do we do? How do we believe in anything?

In the next paragraph I shall use the double headed coin experiment as an example.

We choose as our Hypothesis to be tested that the coin is unbiased. After it comes down heads five times in a row then our Hypothesis is looking decidedly shaky. There is only one chance in 32 (about a 3% chance) that we would get five heads in a row if the coin is unbiased. When we get ten heads in a row then there is only 1 chance in 1024 (less than 0.1%) that the coin is unbiased. We would be right to feel there was something fishy!

Two Hypotheses

The Bayesian method of increasing our knowledge is called Bayesian Epistemology. It is the method which we humans use almost instinctively. We can have as many mutually exclusive Hypotheses as we wish. There is only one constraint and it is that when all the probabilities of the different (mutually exclusive) Hypothesis are added together then these probabilities add up to one. You may have to add a ‘catch the rest’ category to allow for all the “I didn’t think of that” possibilities!

The Bayesian Spam Filter is an example which uses only two Hypotheses. These two mutually exclusive Hypotheses are (a) that the email is not-spam and (b) that the email is spam. When we add together the probability that the email is not-spam and that the email is spam then we do get a total probability of 1. The email has to be one or the other.

Using the Bayesian method (Bayesian Epistemology) we analyse the email and then, on the balance of probabilities (note that, unlike a criminal case or the Scientific Method which favours Hypothesis Testing, this is not necessarily a ‘beyond all reasonable doubt’ probability), the computer program decides whether to classify the email as spam - or not.

As human beings, and, dare I suggest it, all animals, we decide what to do almost instinctively. We use Bayesian Epistemology as our preferred method of making those instinctive decisions which have to be made; we do not and often we can not wait for a proof which is "beyond all reasonable doubt". We have to get on with making a decision based on the balance of probabilities. We may make the wrong decision but, if we do, then we shall have more knowledge and, perhaps, we'll make a wiser decision next time.

Hypothesis Testing is a much poorer tool than is the Bayesian method for decision making.

Ham and Spam

I am indebted to Jan-Jaap van der Geer for the word "Ham" to describe an email which is not "Spam". It is much easier to use the abbreviation "Ham" than to write "Not Spam"!

If you receive an email, not knowing whether it is Ham or Spam then you can pass it through a Bayesian Spam Filter (such as Jan-Jaap's SpamStamp) and it will return the probability that the email is Spam. This probability will always be less than certainty. It is only after looking at the email that you will be able to decide whether it is Ham or Spam.

The Ham and Spam Databases

After you have read an email and decided that it is Ham then you tell the program to add every word in that Ham email to the Ham Database. Similarly, after you have read an email and decided that it is Spam then you add every word in that Spam email to the Spam database. You do not let the computer program decide to add words to the Ham and Spam databases until you have looked at the email yourself and declared it to be Ham or Spam.

Some words such as "the" will appear in both Ham and the Spam databases. If the word "the" occurs five times in an email you have declared to be Ham (you have read it and you have declared it to be Ham) then the Ham database increases in size by five words (and not one as you might have thought). The same applies to the Spam database. If the word "the" occurs seven times in an email declared (by you, not by the computer program) to be Spam then "the" is added seven times to the Spam database. The size of the Spam database increases in size by seven words even though every one of those seven words is "the".

Some Definitions

The Ham and Spam databases will contain different numbers of words.

Let W_H be the total number of words in the Ham database.

Let W_S be the total number of words in the Spam database.

Now consider an email which we haven't looked at (and we don't intend to look at - well, not at the moment) but which is being analysed by our Spam Filter computer program. The program 'looks at' the first word in this email (let's call it "word-1") and then it looks at how many times that word, "word-1", appears in each of the two databases.

Let w_{1H} be the number of times the word "word-1" appears in the Ham database.

Let w_{1S} be the number of times the word "word-1" appears in the Spam database.

What can we calculate?

Using the symbols W_H, W_S, w_{1H} and w_{1S} , let's define a couple of probabilities.

w_{1H} / W_H is the proportion of the words in the Ham database which are "word-1".

w_{1S} / W_S is the proportion of the words in the Spam database which are "word-1".

What are these ratios?

The first, w_{1H} / W_H , is the probability that an email which is known to be a Ham email, contains the word, "word-1". This probability is usually pretty small.

The second, w_{1S} / W_S , is the probability that an email which is known to be a Spam email, contains the word, "word-1". This probability is usually pretty small too.

Although both these probabilities are "pretty small", what we are interested in is not their absolute value but the ratio, Ham to Spam (I shall call it r_1 - have you seen r_1 before somewhere?), which theoretically can have a value of anything between zero and infinity!

Bayes' Theorem 'does it backwards'

We don't really want to know anything about the probability that the word "word-1" is contained in the email under consideration - but it's all that we've got. What we want to know is the probability that an email (containing the word "word-1") is a Spam email!

This is where Bayes' Theorem comes into it. What Bayes' Theorem can do is something mathematicians love to do. Bayes' Theorem 'does probability sums backwards'.

Subtraction is Addition done backwards. Division is Multiplication done backwards. Powers are Roots done backwards. Integration is Differentiation done backwards. There are heaps of examples of what are called 'inverse processes' all of which are loved by mathematicians. If an inverse process doesn't exist then mathematicians invent one.

Bayes' Theorem takes us from knowing the probability ratio contained in the number r_1 to the probability that an email containing the word "word-1" is a Spam email!

How does it do this? By using the formula developed by the Rev Thomas Bayes:

$$P = \frac{1}{(1 + r)}$$

When we have calculated the value of r_1 we simply insert this value of r_1 in Bayes' Formula (Theorem or Equation if you prefer) and find P, the probability (based only on the single inclusion of "word-1" in the email) that the email is Spam.

Multiple Events

Let me say right away that $r = r_0*r_1*r_2*r_3*r_4*r_5*...$ follows from Bayes' Theorem. I'm not going to explain how this is proved. Trust me, it can be proved! The Rev T Bayes says so!

The "..." implies that this series of values of r can be extended indefinitely.

Each of these r_n corresponds to one event (one word in the email).

I shall expand on all this by using the example of 'the bus that never came'. You will remember that I waited at the bus stop for a bus and, unknown to me, the buses were not running that day. I could have waited forever - but I didn't. I decided to follow the mathematical (philosophically instinctive) advice of the Rev Thomas - and give up!

In the case of the missing bus, an event occurs every minute that the bus doesn't come! For every minute that goes by we have another value of r_n to add to the chain of events.

For the simple version of the 'missing bus' problem we have two hypotheses for every one of these events. Our first hypothesis is that the buses are not running and our second is that they are running at approximately ten minute intervals. We could have other hypotheses.

Using the first hypothesis we calculate the probability that NO bus will arrive when the buses are NOT running. I'm sure you will agree that if the buses are not running at all then the probability of a bus NOT arriving is 100% (or 1.00 if you prefer).

[Note to Editor: Perhaps a picture of a 'big red bus' might be inserted at this point?]

Then we use the second hypothesis to calculate the probability that NO bus will arrive when the buses are running at approximately ten minute intervals. The probability of a bus arriving in the nth minute is 1 in 10 (10% or 0.1); consequently we calculate that the probability of NO bus arriving is 90% (or 0.9 if you prefer).

The ratio, r_n , is found by dividing the 0.9 of the second hypothesis by the 1.0 of the first hypothesis giving us 0.9. For every minute, the value of r_n is the same; it is 0.9. We can calculate the value of r_n (it is 0.9) for every minute except the 0th minute. We have $r_n = 0.9$ for every value of n except $n = 0$.

The only value we can't calculate is the value of r_0 . Do you remember how to find r_0?

When we use the method called Hypothesis Testing we have only one hypothesis. After collecting enough data we may be able to reject our chosen hypothesis as being improbable. When we use the Bayesian method we can have as many hypotheses as we can think of and, using a suitable spreadsheet, we can test them all at once and then, when we need to make a decision, we can select the one which seems most likely at the time!

Back to $r = r_0*r_1*r_2*r_3*r_4*r_5*...$

We have every value of r_n (except r_0) for the 'missing bus' problem; they are all 0.9. This first value, r_0 , is something which seems to baffle even the expert. It is this first value of r which makes the Bayesian method what it is and distinguishes it from Hypothesis Testing. Also this matter of how to 'calculate' a value for r_0 is why many lecturers shy away from teaching Bayes' Theorem - or teach it very badly.

I shall return to this 'problem' of "What is the value of r_0?" for our email - but not today.

Calculating the values of r_n

Each of the values of r is a ratio of probabilities. In the case of a Bayesian Spam Filter, unlike the problem of the missing bus (where all the values are $r_n = 0.9$) each of the values of r_n are different; there is one value of r_n for each word in the email.

I have said that r_n is the ratio of two probabilities. How can we have two different probabilities for one event? The answer is that we can have two probabilities for the same event when we have two different theories (hypotheses) to consider.

One hypothesis we have to consider is that the email is Ham. The other possibility is that the email is Spam. For each word such as "word-1", "word-2", "word-3", etc, in the email we calculate both probabilities. Then, for each word, we find the ratio of the these two probabilities. The Ham/Spam ratio for the nth word is what I have called r_n.

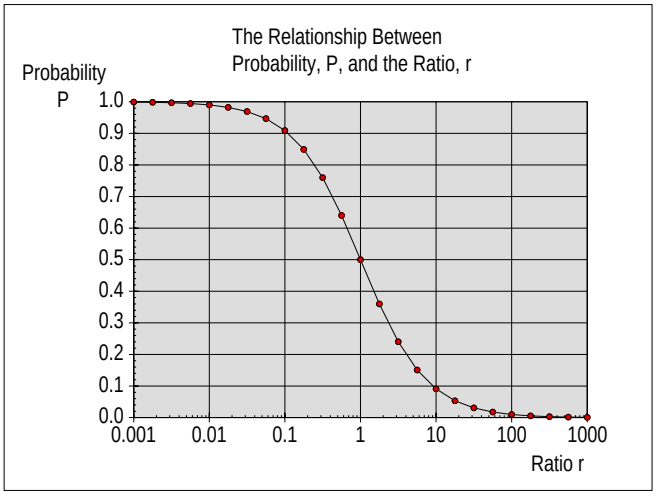
Once again ... Back to $r = r_0*r_1*r_2*r_3*r_4*r_5*...$

We have every value of r_n (except r_0) for the 'missing bus' problem; they are all 0.9. We can calculate a value of r_n for every word in an email with the exception of r_0.

What the Bayesian Spam Filter does is to calculate the value of r_n for every word in the email and enter them into the formula $r = r_0*r_1*r_2*r_3*r_4*r_5*...$ to find r.

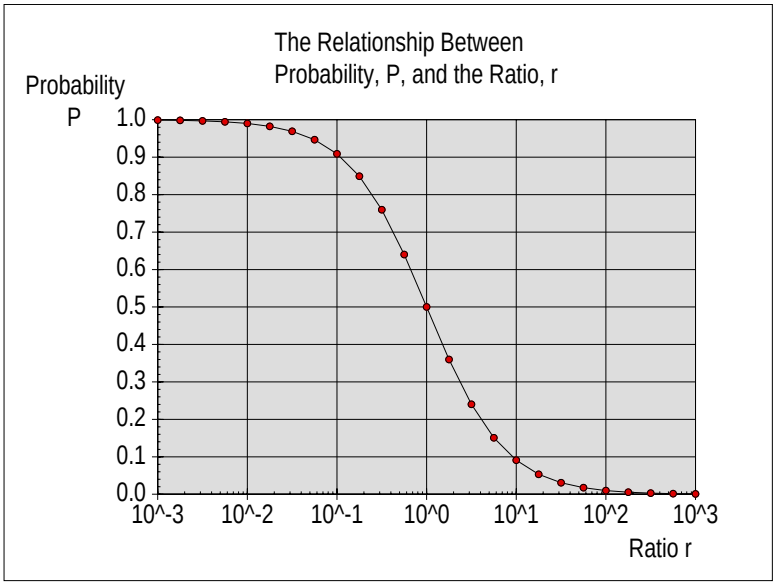
The relationship between P and r

Here is a graph (created using PipeDream) of the relationship between P and r.



You will see that when r is smaller than 0.01 the probability is close to certainty. When r is larger than 100 the probability of the hypothesis being true is negligible.

Here is an alternative version of the same chart:



You will see that, instead of using numbers such as 0.01 or 100 along the x axis I have used the power notation, 10^(-2) and 10^(+2) respectively.

To me (and to many mathematicians) this is a more 'beautiful' graph because there is a certain symmetry about it. We can make it even more 'beautiful' - let's give it a go!

Noah and the 'log tables'

[Note to Editor: Perhaps a picture of the San Maurizio Fresco of Noah's Ark by the Renaissance painter, Bernadino Luini might be an interesting picture to include?]

The advent of reasonably priced hand held electronic calculators in the mid 1980s did away with the need to use log tables and it may be that the very old (mathematical) joke at the begining of this article will have be lost on some of our youngest Archive readers. For our youngest readers I shall try to explain what logarithms are and how they work.

If you have a passion for mathematical symbols you will appreciate my introduction of "s". The definition of s_n is that $s_n = \log(r_n)$; the inverse of this equation is $r_n = 10\text{^}s_n$.

Here is a simple example of using logarithms to carry out a multiplication.

Suppose I want to do the multiplication sum 0.01 * 1000.

First I write the two numbers in a different way. 0.01 = 10^(-2) and 1000 = 10^(+3).

Now I can write the calculation as:

0.01 * 1000 = 10^(-2) * 10^(+3).

Using the law of indices: 10^(-2) * 10^(+3) = 10^(-2+3) = 10^(+1) = 10.

The power to which 10 is raised is called the logarithm of the number.

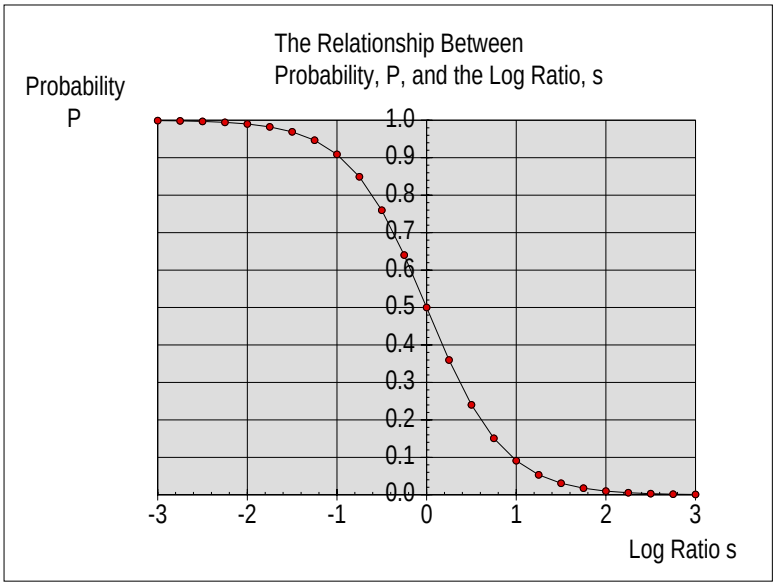
The logarithm of 0.001 is -2 and the logarithm of 1000 is +3. If you know (or can look up in log tables) the logarithm of every number you need to multiply together, then all you have to do is to add those logarithms and then look up 10 to the power of that sum.

The identity I am using is that: $\log(r_n * r_m) \equiv \log(r_n) + \log(r_m)$.

Noah knew this (my guess is by instinct - the Bayesian approach is instinctive) and he provided his adders with some log tables so that they could "go forth and multiply". It would seem that we (mankind) didn't need the log tables in order to over-populate our planet! All we needed was the instruction to multiply - and how! Perhaps we're not taking our custodianship of the Earth seriously enough! All things in moderation (Temperantia)!

The logarithmic version of the graph

Instead of calculating the values of r_n we get the computer to calculate values of $s_n = \log(r_n)$ - and then we plot P, the probability (that the email is a Spam) against s where s is defined as $s = s_0 + s_1 + s_2 + s_3 + s_4 + ...$



In effect what this means is that for every word, "word-n", in the email we calculate the value of $s_n = \log(r_n)$ and then we simply add up all the values of s_n to find s.

I am sure that you can see that if s < (-2) then it is almost certain that the email is Spam and that if s > (+2) then it is almost certain that the email is Ham.

The 'beauty' of using s_n, the logarithm of the value of r_n, is that we can directly compare in our minds the 'pulling power' (towards Ham or Spam) of every word in the email. We are able to say with confidence that a 'ham biased word' with an s value of, say +3, will offset three 'spam biased words' each with an s value of -1. Our brains work so much more intuitively with addition rather than multiplication.

All we have to do is give an “s value” to every word (eg “the” or “viagra” or “Gerald”) in the email and then add together these s values. If the sum is -2, or more negative than -2, then the email is Spam! We can do this without a calculator!

Summary

The theorem developed by the Rev Thomas Bayes allows us to do probability sums backwards.

We start with two probabilities. One of these probabilities is that a word such as “the” or “viagra” or “performance” or “Gerald” will be included in a Ham email. The other is the probability that these same words will be included in a Spam email. Let me emphasise that using Bayesian Epistemology we always have two or more probabilites for the same event. These different probabilities arise because, unlike the Hypothesis Testing method, we decide that we shall consider more than one hypothesis.

We find the ratio of these two probabilities (the Ham probability divided by the Spam probability) and we call it “r”.

If we have lots of events (the number of times we toss the coin, the number of minutes we wait for the bus, the number of words in an email) then we find the s values (the log of the probability ratios, r_n) of these events and then add the s values together. Easy isn’t it?

Issues still to be resolved

There are still a few loose ends to tidy up to make our Bayesian Spam Filter practical.

Here are a couple of them - there are others.

What value of s do we use if a word in the email doesn’t appear in either of the databases?

What value should we choose for r_0?

Let’s leave those issues for another day.

Acknowledgement

I must draw your attention to the excellent Bayesian Spam Filter program created by Jan-Jaap de Geer called SpamStamp. You will find the latest version of SpamStamp at:

http://home.c2i.net/jjvdgeer/riscos/spamstamp.html