Gerald's Column
by Gerald Fitton

The example I shall use for my third variant of Bayes' Theorem is a Bayesian Spam Filter. One example of such a filter finding popularity amongst the RISC OS community is the program SpamStamp by Jan-Jaap van de Greer. In this month's article I shall tackle the 'heavy' stuff, the algebra. The 'how it works' bit will follow next month.

Before I launch into the algebra (and my use of Equasor) there is a question I must address.

The 'Truth Stakes'

You all know that I love questions. It would not be an exaggeration to say that my fondness for questions exceeds my need for answers. Indeed I like to believe that, even if I can't recognise a good answer, I am able to recognise a good question when I see one!

It was Andrew Pinder who first asked me about the relationship between Bayesian Epistemology and Hypothesis Testing; he did so before last month's article was published.

My detailed answer to Andrew's excellent question will have to wait until next month but, in the meantime, for those who are (mathematically) curious, here is my enigmatic reply.

Suffices are a wonderful invention.
A suffix numbered 1 is 'obviously' more important than a suffix numbered 2.
A suffix numbered 0 must be more important than a suffix numbered 1!

Perhaps I should expand a little!

Bayesian Epistemology

Aficionados of Bayes' Theorem give pride of place to Hypothesis 1, H_1. In my flippin' coin example, Hypothesis 1, H_1, is that the coin has two heads. In the 'Truth Stakes', our knowledgeable punter concentrates his attention on H_1. If my two headed coin starts as an outsider at 32 to 1 (a priori probability of 1/33), then at each flippin' hurdle the bookies will reduce the odds until, after successfully flying over the fifth, the odds on H_1 are evens!

Hypothesis 2, H_2, plays a subsidiary role. It starts as the 32 to 1 odds on favourite (win an 'a priori' probability 32/33) to evens as (sequentially) it fails to clear five flippin' hurdles.

Hypothesis Testing

The Bayesian aficionado focuses on H_1 and watches as his outsider gains credibility during the race. The Hypothesis Tester, is more a 'percentage player'. He focuses his attention on H_2, and sees this odds on favourite lose the race in favour of the unknown outsider, H_1.

I have said that it is 'obvious' that a suffix numbered 1 commands much more respect than does a suffix numbered 2. The Hypothesis Tester, who wishes us to focus our attention on the favourite, H_2, knows that this Hypothesis needs a 'make over'.

The biggest single problem for the ‘spin doctor’ is to promote H_2 so that it assumes a more prominent position in the hierarchy of suffices. What can this he do?

First of all he renumbers H_2 as H_0.
The Number 0 is ‘obviously’ of much more importance than is the rival Number 1.
Secondly he gives a fancy sounding name to this Hypothesis, now H_0.
He calls it “The Null Hypothesis”.
Undoubtedly it’s much more important now!

In order to extend his contract (the more he does the more he gets paid) he decides to do something about H_1 (the coin has two heads)?

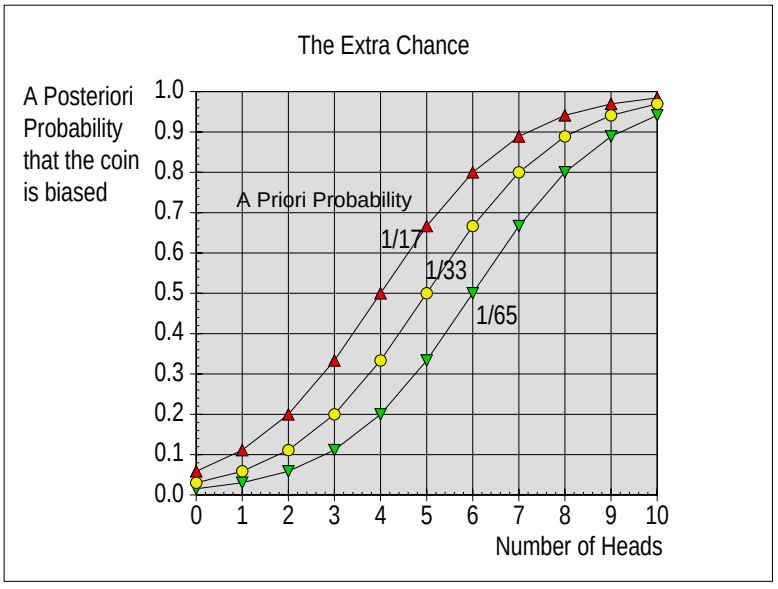
He gives it a name more suited to its subservient stature.
He calls it “The Alternative Hypothesis”.

Indeed, H_1 is now regarded as being so irrelevant that we can (almost) ignore it. We don’t have to worry about guessing its arbitrary ‘a priori’ probability, $P(H_1)$, neither do we have to use Bayes’ Theorem to convert this arbitrarily chosen ‘a priori’ probability to the subjective ‘a posteriori’ ‘Degree of Belief’, $P(H_1|E)$.

The undoubted attraction of Hypothesis Testing is that it takes all the subjectivity out of our search for knowledge. We can all agree that $P(E|H_0) = 1/32$. It is unlikely that we shall ever agree on $P(H_1)$, the ‘a priori’ probability that the coin has two heads.

If $P(E|H_0)$ is small enough then the Hypothesis Tester will reject the Null Hypothesis in favour of the Alternative Hypothesis. No further thought is required. He is blissfully content, even though he knows nothing at all about the hypothesis, Hypothesis 1, to which the Bayesian devotee has given his full attention. There is a certain amount of security in dealing with the objective probabilities of Hypothesis Testing rather than the guesswork and prejudice essential to the application of Bayes’ Theorem.

The Extra Chance



Let's stop there (for this month) and get back to my enigmatic "Extra Chance".

If the Null Hypothesis, H_0, is that the coin is unbiased and the Event E is that the coin comes up heads five times in a row then $P(E|H_0) = 1/32$. This is about 3%.

Null Hypotheses are regarded as suspect if $P(E|H_0)$ is less than 5%. If we reject H_0 because 3% is smaller than 5% (the phrase 'Confidence Limits' comes to mind), then we are left with H_1, a Hypothesis about which we know absolutely nothing!

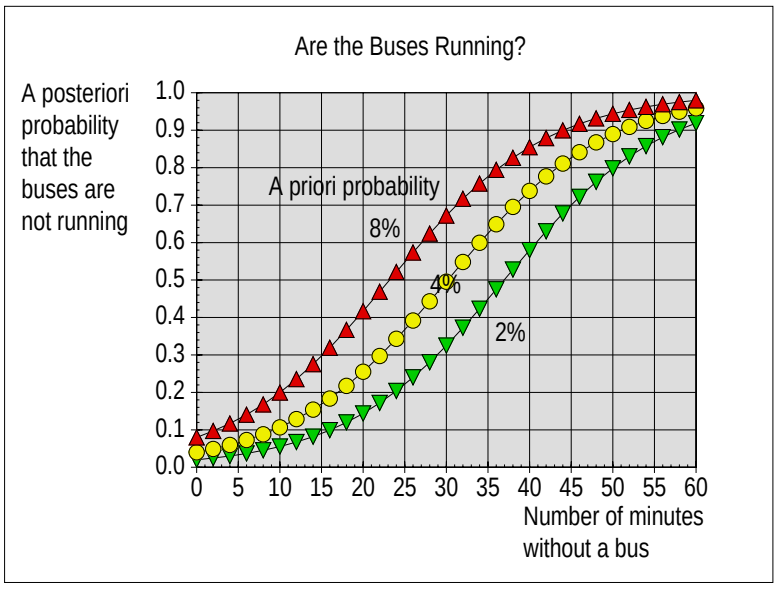
Form a Bayesian viewpoint an 'a priori' probability of $H_1 = 1/33$ increases to an 'a posteriori' of 1/2 after five flippin' heads—so why is $P(E|H_0) = 1/32$ and not 1/33?
Where has the extra chance (1 in 33 instead of 1 in 32) come from?
Or, if you prefer to support Hypothesis Testing viewpoint, then where has it gone?

If you can explain "The Extra Chance" then please write to me ...
Otherwise—you will have to wait until next month!

Chained Events

My third example of Bayes' Theorem addresses the question, "What happens when we experience a sequence of Events all of which are relevant to the same Hypothesis?"

In my first 'Blue eyed Blonde' example there were a couple of Hypotheses but only one Event. In my second 'missing bus' example, I concentrated your attention on the fact that I introduced a third Hypothesis. This may have been sufficiently distracting for you not to have noticed that I introduced the concept of a 'Compound Event' at the same time. The PipeDream spreadsheet illustrates that, as the minutes passed by, my 'Degree of Belief' that a bus would arrive was inexorably eroded by the sequence of related (no bus) Events.



What I did was to I compound a sequence of related Events (the minute by minute non arrival of the bus) into a single Compound Event. I used a PipeDream spreadsheet and a couple of fairly simple Probability Distribution Functions (PDFs) to find the probability of the compound Event rather than deal individually with each (minute by minute) Event.

What I did not do, but will do in this, my third example of Bayes' Theorem, is to regard each Event as a separate Event having its own individual incremental quantitative effect on the 'a posteriori' probability. In this third example each Event, E_n, has its own independent and distinct values for $P(E_n|H_1)$ and $P(E_n|H_2)$.

In general terms we can state the problem as follows:

We have a Hypothesis, H_1, and we are considering our 'Degree of Belief' in it. Perhaps "considering" is the wrong word for what we are doing. We are actively investigating how 'believable' it is. We don't have a single Event, E_1, but a whole sequence of Events such as E_1, E_2, E_3, etc. For each of these Events, E_n, we have values of $P(E_n|H_1)$ and $P(E_n|H_2)$.

How can we combine them together?

SpamStamp

In the case of the Bayesian Spam Filter, SpamStamp, we have a Hypothesis, H_1, "This email is Spam" and fifteen Events. These Events are the most significant 15 words in the email; for each of these words we have values for $P(E_1|H_1)$, $P(E_2|H_1)$, $P(E_3|H_1)$, etc and $P(E_1|H_2)$, $P(E_2|H_2)$, $P(E_3|H_2)$, etc. How can we combine these probabilities together?

There is another more difficult question which I shall postpone until next month. I expect that you will have guessed what it is. "What values shall we use for $P(H_1)$ and $P(H_2)$? These probabilities do not have objective values. What value does SpamStamp use?

Equations

I hope that our Editor, Paul, will be amused rather than upset by my next remark.
Paul 'messes up' my equations.

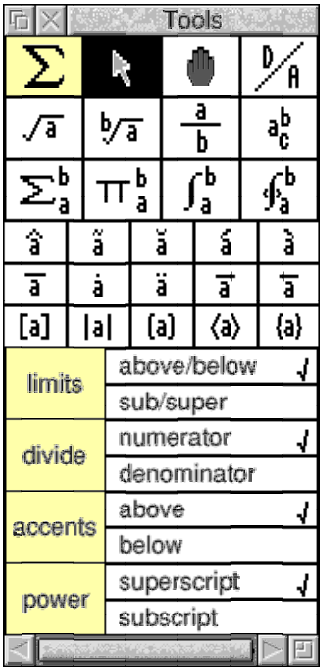
The reason Paul messes them up is because his layout in Archive has a short line length and my equations are longer than his lines. The equations get 'messed up' because Paul has 'Line Wrap' on and this automatically wraps my long equations around at the end of his (far too short) lines. Sometimes this 'line wrap' occurs at a most inappropriate place in an equation. I know that Paul is not the only person who has this problem with Desk Top Publishers and Word Processors because I receive correspondence about it most regularly.

In this section of my article about chained Events I know that the algebra will be much more understandable if equations are written on a single line. What can I do about it? I know. I'll use the utility provided with Impression called Equasor!

Equasor

Of course there are more recent and (some believe) better utilities for writing mathematical equations but my guess is that (almost?) everyone who reads Archive will have a copy of Equasor sitting somewhere on their hard drive. The documentation supplied by Computer Concepts for Equasor (and, for that matter, Impression) is more than comprehensive.

If you haven't used Equasor (or one of the more recent utilities) then I assure you that it is a good way of avoiding the perils of 'Line Wrap'. Furthermore you can introduce Mathematical symbols such as the integral very easily. Just click on the box in the Tools menu. Before using it to write my equations I had a good look at the Equasor manual. I thought I'd better check how to produce subscripts. By comparison with today's 'on disc' manuals, the explanations and examples in the printed manual are wonderfully lucid.



The Equasor Tools Menu

On the Archive monthly disc and elsewhere I have included the following equations in both Equasor and DrawFile format. If you have any problems with the use of Equasor then please write to me. I'll see what I can do to help.

First some Algebra

The two Hypothesis version of Bayes' Formula in the graphic is a DrawFile which I created using Equasor. I hope that you will find this 'easier to read' than the wrapped around versions which have appeared in the last few editions of Archive. Let me know?

$$P(H_1|E_1) = \frac{P(H_1) * P(E_1|H_1)}{[P(H_1) * P(E_1|H_1) + P(H_2) * P(E_1|H_2)]}$$

Bayes' Equation

The 'heavy' algebra you have to do is to divide both the numerator and the denominator of the right hand side of this equation by the numerator. Does the Equasor layout of these equations makes understanding the algebra much easier? I think so.

The result, shown as another DrawFile, saved from Equasor, is:

$$P(H_1|E_1) = \frac{1}{\left[1 + \frac{P(H_2)}{P(H_1)} * \frac{P(E_1|H_2)}{P(E_1|H_1)}\right]}$$

Modified Bayes' Equation

In this second graphic I have doubled the point size of the "1" in the numerator (relative to the numbers in the denominator) so that (to my eye) the size of the numerator and denominator appear more balanced. Perhaps a bit larger would have been even better?

The Effect of E_1

I would like you to look carefully at this modified version of the Bayes' Formula. The important bit is the two ratios in the denominator of the right hand side. I have deliberately constructed these ratios so that both 'a priori' probabilities form one of the ratios and the other depends on the Event, E_1. I have successfully separated the subjective probabilities, H_1 and H_2, from the objective conditional probabilities, $P(E_1|H_1)$ and $P(E_1|H_2)$.

The ratio $P(H_2)/P(H_1)$ has a subjectively chosen value. I shall return to it next month. This month let us concentrate our attention on the ratio $P(E_1|H_2)/P(E_1|H_1)$. If we had some 'metric' to measure it by then we could use this single number to represent quantitatively and objectively the effect of the Event, E_1, on our 'Degree of Belief' in H_1.

This ratio can take values from zero to infinity.
When it is zero then $P(H_1|E)$ becomes unity, certainty (the email is spam).
When it is infinity then $P(H_1|E)$ becomes zero, impossibility (the email is not spam).
Next month we shall look at values between these extremes.

After the Event

By "After the Event" I mean "After the first Event, E_1". The effect of the Event, E_1, is that our 'Degree of Belief' in H_1 has changed from $P(H_1)$ to $P(H_1|E_1)$.

Let us put E_1 behind us and consider the next Event, E_2. Our 'a priori' degree of belief in Hypothesis 1, H1, is no longer $P(H_1)$ but $P(H_1|E_1)$. With this knowledge we can substitute $P(H_1|E_1)$ for $P(H_1)$ in the modified form of Bayes' Theorem. What do we find?

$$P(H_1|E_2) = \frac{1}{\left[1 + \frac{P(H_2|E_1)}{P(H_1|E_1)} * \frac{P(E_2|H_2)}{P(E_2|H_1)}\right]}$$

The Second Event

I won't bore you with the tedious algebra but, substituting $P(H_2|E_1) = 1 - P(H_1|E_1)$ and then simplifying the resulting fraction, it is relatively easy to show that:

$$\frac{P(H_2|E_1)}{P(H_1|E_1)} = \frac{P(H_2)}{P(H_1)} * \frac{P(E_1|H_2)}{P(E_1|H_1)}$$

Chaining Ratios

which in turn leads to:

$$P(H_1|E_2) = \frac{1}{\left[1 + \frac{P(H_2)}{P(H_1)} * \frac{P(E_1|H_2)}{P(E_1|H_1)} * \frac{P(E_2|H_2)}{P(E_2|H_1)}\right]}$$

The Chain of Ratios

The Chain

The time has come for those of you who have struggled with the algebra to reap your much deserved reward. We are so nearly there. Don't give up now!

Have a look at the chain of ratios in the denominator of the right hand side of this equation. You will see that each Event generates its own conditional probability ratio. For the sake of simplicity let me call these ratios r_1, r_2, r_3, r_4, etc where $r_1 = P(E_1|H_2)/P(E_1|H_1)$ etc.

Also, may I ask you to indulge me with a little mathematical licence.
Please may I define r_0 as: $r_0 = P(H_2)/P(H_1)$?
I promise to return to r_0 next month.

If we have five Events which, taken together, we can regard as the compound Event E (without a suffix) then $P(H_1|E) = 1/(1 + r_0*r_1*r_2*r_3*r_4*r_5)$.

Now doesn't that look much simpler?

Of course the chain can be extended indefinitely.

Each value of r can be calculated independently of all the other values of r.
Multiplication of these values of r is commutative.
By this I mean that we can multiply the values of r in any order which we find convenient.

SpamStamp

In the program SpamStamp, a value of r is calculated for every word in an email. These individual values of r are multiplied together and then the result is fed into the denominator of the equation, $P(H_1|E) = 1/(1 + r_0*r_1*r_2*r_3*r_4*r_5)$—with a separate calculable value of r_n for every word in the email. Next month we shall discover why SpamStamp only requires values of r_n for the most significant fifteen words in the email.

Because all the r_n are independent and commutative we can take the words in any order we please, calculate the values of r_n, multiply them together, substitute in the equation and, believe it or not, find $P(H_1|E)$, the 'a posteriori' probability that the email is spam.

Conclusion

This month I have subjected you to a lot of algebra lightened only by my use of Equasor. What I have done is to rewrite Bayes' Formula in a much more usable form so that each Event appears in the modified formula in an independent, isolated and manageable way.

Next month I want to concentrate on the way in which SpamStamp uses this modified version of Bayes' Formula. We shall look at some of the assumptions including the value of r_0, the effect of the notorious 'bias factor', the way in which words missing from the dictionaries are treated by SpamStamp and why it uses only the most significant 15 words.

Communication

Please contact me by email (preferred) or by letter if you have any questions or comments.

Because of spam problems I have set up a Spam Filter which rejects (bounces) emails directed to <gerald@abacusline.demon.co.uk> unless you are on my personal 'white list'. Please email me at <Archive@abacusline.demon.co.uk> if your email relates to articles in the Archive magazine or at <GoldLine@abacusline.demon.co.uk> if your correspondence relates to PipeDream, Fireworkz or GoldLine or, if your correspondence relates to Living with Technology, then please use <LwT@abacusline.demon.co.uk>.