Bayes' Theorem.*

Chilin Shih, Greg Kochanski
greg.kochanski@phon.ox.ac.uk, cls@uiuc.edu

September 15, 2006

1 Introduction

Bayes' Theorem, or the related likelihood ratio, is the key to almost any procedure for extracting information from data.

Bayes' Theorem lets us work backward from measured results to deduce what might have caused them. It will be the basis of most of our later model building and testing.

It UCalgary (2003) is the work of Rev. Thomas Bayes (St.Andrews, 2003), about whom only a modest amount is known, but he has the perhaps unique distinction that two-thirds of his publications were posthumous and the remaining third anonymous.

2 What is a model?

A model is a quantitative description of a part of the universe. It can be a theory, in the sense of Popper (1959) in that it makes falsifiable predictions, but one doesn't really think of a model as being "true" or "false". One difference is that models generally don't pretend to explain why something happens. One usually thinks of a model as being useful within some range of validity. An experiment can disprove a model (if it can be shown not to be valid anywhere), or it can help you know what the model's range of validity is.

One example of a model that is *not* an explanatory theory is Hooke's Law in mechanics. That model states that the force (F) required to stretch a spring is proportional to how far (x) you stretch it: $F = k \cdot x$, where k measures how stiff is the spring. It describes the phenomenon, but does not pretend to describe *why* it happens. For instance, it doesn't mention atoms. Note also that it is

only a partial description: specifically it doesn't tell you how to find k. All it says is that k is a property of a spring, and is presumably invariant.

In linguistics, an example of a model might be a rule-based system for predicting the pronunciation of a word.

3 The Theorem

Bayes' Theorem is pretty simple to write:

$$P(B|A,C) = P(A|B,C) \cdot \frac{P(B|C)}{P(A|C)}, \tag{1}$$

although one often sees it written in an form where condition C is omitted: $P(B|A) = P(A|B) \cdot P(B)/P(A)$.

One derives it by starting with the the symmetry of the "and" operation: $A \cap B = B \cap A$, and then use the algebra for "and":

$$P(A \cap B|C) = P(A|B,C) \cdot P(B|C). \tag{2}$$

As usual, A and B are two events we currently care about, and C represents all the other conditions, some of which may be relevant in that they could change the probabilities we calculate, and others of which are just irrelevant.

The symmetry of "and" allows us to interchange A and B in Equation 2 to yield:

$$P(B \cap A|C) = P(A|B,C) \cdot P(B|C), \tag{3}$$

but we can also write $P(B \cap A|C)$ directly from the algebra for "and" as:

$$P(B \cap A|C) = P(B|A,C) \cdot P(A|C). \tag{4}$$

Since the right sides of Equations 3 and 4 are equal to the same thing ($P(B \cap A|C)$), they must be equal to each other, so one gets

$$P(B|A,C) \cdot P(A|C) = P(A|B,C) \cdot P(B|C), \tag{5}$$

from which one gets Equation 1 (Bayes' Theorem) with a bit of algebraic rearrangement.

The problem that some people have with Bayes' Theorem is not writing it or deriving it but in interpreting what the symbols mean or in supplying values. Let's try applying it and see if we can figure out how it works.

4 Probabilities and models

Q: How do you find a good model of language?
A: Try a billion variations on a theme and see which is best.

If one is out to find a good model, one should think of oneself as an editor, or a judge of a competition. Imagine you have thousands (millions!) of models, all

written out by your friends and devoted admirers: your job is not to create the models, but instead it is to choose the best model from the available selection. Now, in practise, we may not want to go through the trouble of completely building all possible models, and then testing them. There might be too many, or we might not have enough devoted admirers. This selection metaphor always works in principle, and (for small problems) in practise.

We can make it work with many (billions or infinities of) models when there are strong similarities between the models to test: they come from a family. An example might be a family of models for the average height of this class:

- ...
- The average height is 140cm.
- The average height is 141cm.
- The average height is 142cm.
- The average height is 143cm.
- The average height is 144cm.
- ...

From here on down, the word "model" stands for any one of the models in such a family.

Let's rewrite Bayes' Theorem so it applies to data and models. "Datum" stands for any arbitrary measurement. It might be a very complex measurement, even yielding many numbers at once. A measurement can be heads or tails resulting from the flip of a coin, or the weight of a person. It could also be the combination of (Height, Skin Colour, Hair Colour) which might be the information available from the witness of a crime, from which you would like to identify the criminal. It might also be a whole vector of numbers, for instance frequency counts of different letters, which we might use to identify the author of a text or the language in which it is written. It might be the fact that an utterance ("The committee are Smith, Jones and Allyn-Weatherly.") is attested.

5 The Left Side

We will substitute our new names for A and B into Bayes' Theorem and get

$$P(\text{model}|\text{datum}, C) = P(\text{datum}|\text{model}, C) \cdot P(\text{model}|C)/P(\text{datum}|C). \quad (6)$$

Right away, it is obvious how we're going to use it. On the left side is $P(\text{model}|\text{datum}, C)$, which is an extremely useful number: it tells us the probability that each of the models is true, given that we've seen a particular value of the datum under our particular experimental conditions. Recall that we are

thinking about a lot of models, and that the datum could take on two or more values[1], so the notation $P(\text{model}|\text{datum}, C)$ actually stands for a whole array of related numbers: one probability for each combination of model and datum[2].

Now, if our datum is **heads** (as opposed to **tails**), and if we have two competing models, then we might end up with

$$P(\text{Greg's.Model}|\text{heads}, C) = 0.4 \tag{7}$$

(after seeing the data) and

$$P(\text{Chilin's.Model}|\text{heads}, C) = 0.6 \tag{8}$$

(after seeing the data).[3]

Note that if we had observed **tails**, we would be computing $P(\text{Greg's.Model}|\text{tails}, C)$ and $P(\text{Chilin's.Model}|\text{tails}, C)$ instead.

6 The Right Side

The next term, $P(\text{datum}|\text{model}, C)$ translates to "Assuming that a certain model is true, how likely are you to get each possible datum given our experimental conditions?" So, if Greg's model states that the coin is fair, and exactly 50%-50%, then $P(\text{datum}|\text{Greg's.Model}, C)$ stands for two numbers:

- $P(\text{heads}|\text{Greg's.Model}, C)$, which is 0.5, and
- $P(\text{tails}|\text{Greg's.Model}, C)$, which is also 0.5, in that model.

The other competing model is Chilin's model which states that it is a double-headed coin, and implies that the guy flipping the coin is out to cheat you. So, we also have:

- $P(\text{heads}|\text{Chilin's.Model}, C)$, which is 1.0, and
- $P(\text{tails}|\text{Chilin's.Model}, C)$, which is 0.0.

So, the right side is really four numbers! Good! That means we have four equations here, because the left side was also four numbers. We calculate the probability of both models assuming first that the data is heads and then for both models assuming the data is tails.

In practise, once you have the data, you can ignore some of the equations: once you know the coin showed heads, you usually don't care what model you might have gotten if the coin had (instead) turned up tails.

[1] If the datum has only one possible value, it won't be helpful.

[2] That's assuming that we don't chance C. If C could change, there would be even more things to calculate.

[3] If we wanted to be blatant, we could write the same probabilities as $P(\text{model} = \text{Greg's.Model}|\text{datum} = \text{heads}, C)$ and $P(\text{model} = \text{Chilin's.Model}|\text{datum} = \text{heads}, C)$.

7 $P(\mathbf{model}|C)$

The next term is the one that disturbs some people. It is the probability that a particular model is true under condition C. Since it doesn't depend on the data, it must be the probability that the model was true *before* you knew what the data was. This is called the prior probability (or the *à priori* probability). It bothers some people, because you need to put a number in here even if you've never done the experiment before[4].

However, things are not as bad as might seem. We will later show that if you repeat the experiment many times, your initial guess for $P(\text{model}|C)$ becomes less and less important[5]. Also, it's a problem that one cannot avoid, so one might as well be fatalistic and accept it. Using likelihood statistics is simply equivalent to one particular choice of the prior distribution; so dropping the $P(\text{model}|C)$ term is simply equivalent to setting it to a very broad distribution.

Rarely do we operate from a position of entire ignorance. Before we weigh someone, we may not have a detailed prior probability distribution in mind, but we know that adults weigh more than infants, and infants weigh about 4 kg. We also know, from reading the less-serious variety of newspaper ("World's heaviest woman buried in piano crate.") that 500 kg people are quite rare. Consequently, we almost always have at least some hint of a prior probability distribution in mind.

On the bright side, $P(\text{model}|C)$ is where you can legitimately put your hunches, biases, and personal opinions. When you write your paper, you will then explicitly display your hunches for everyone to read and evalaute. Your readers certainly can't expect anything better than that.

Best of all, this is where you can put prior probabilities that you get from actual measurements: for instance, if we had looked at 100,000 coins, and found one with two heads, we could assign $P(\text{Chilin's.Model}|C) = 10^{-5}$.

However, for this example, we'll assume that the guy flipping the coins looks a little shifty. So, we'll assume that while he's probably using a fair coin ($P(\text{Greg's.Model}|C) = 0.9$), there is a noticeable chance, even before we look at the data, that he is cheating ($P(\text{Chilin's.Model}|C) = 0.1$).

8 $P(\mathbf{datum}|C)$

But, now what is $P(\text{datum}|C)$? Is this the *à priori* probability of getting heads? Do we have to guess some more? It is an estimate of the probability of getting heads (before seeing the data), but we don't have to guess any more. We can calculate $P(\text{datum}|C)$ from $P(\text{model}|C)$ and $P(\text{datum}|\text{model}, C)$ using the rules of probability that we already know.

[4] In fact, it bothered some people so much that they created a whole set of likelihood statistics without this term. Any statistic that depends on a likelihood or likelihood ratio is just an application of Bayes' Theorem with all the prior probabilities set to be equal to each other.

[5] As long as it doesn't assign a zero probability in the wrong place.

(At this point, we will drop the condition C, because it has served its purpose of showing that the prior probability is just another conditional probability. Remember it, for later when we apply Bayes' Theorem recursively.)

Specifically, since the different models are assumed to be mutually exclusive, and that the models cover all possibilities, then we can partition the event **data** out across all the different models:

$$\text{datum} = (\text{datum} \cap \text{model}_1) \cup (\text{datum} \cap \text{model}_2) \cup \ldots . \tag{9}$$

In other words, we are treating model_1, model_2, ... as http://kochanski.org/gpk/teaching/0401Oxford/Prob2.pdf. The probabilities of these elementary events are calculated the usual way, via $P(\text{datum} \cap \text{model}) = P(\text{datum}|\text{model}) \cdot P(\text{model})$, and thus $P(\text{datum})$ is just the sum of the probabilities of all the elementary events:

$$P(\text{datum}) = \sum_{\text{all models}} P(\text{datum}|\text{model}) \cdot P(\text{model}). \tag{10}$$

For this particular case,

$$\begin{aligned} P(\text{heads}) = \quad & P(\text{heads}|\text{Greg's.Model}) \cdot P(\text{Greg's.Model}) \\ & + P(\text{heads}|\text{Chilin's.Model}) \cdot P(\text{Chilin's.Model}) \end{aligned} \tag{11}$$

and

$$\begin{aligned} P(\text{tails}) = \quad & P(\text{tails}|\text{Greg's.Model}) \cdot P(\text{Greg's.Model}) \\ & + P(\text{tails}|\text{Chilin's.Model}) \cdot P(\text{Chilin's.Model}). \end{aligned} \tag{12}$$

Note: You get the same answer by considering $P(\text{datum})$ to be just an arbitrary "fudge factor", and then choosing its valus so that $P(\text{Greg's.Model}|\text{datum}) + P(\text{Chilin's.Model}|\text{datum}) = 1$ (the left side sums to one). The left side had better sum to one, because these are probabilities, and they cover all the possible alternatives, and thus (by the definition of probabilities) sum to one.

9 Putting it together

Let's evaluate what we can: We know $P(\text{Greg's.Model})$, $P(\text{Chilin's.Model})$ and $P(\text{tails}|\text{Greg's.Model})$, $P(\text{heads}|\text{Greg's.Model})$, $P(\text{tails}|\text{Chilin's.Model})$, and $P(\text{heads}|\text{Chilin's.Model})$. That means we can calculate the $P(\text{datum})$:

$$P(\text{heads}) = 0.5 \cdot 0.9 + 1.0 \cdot 0.1 = 0.55, \tag{13}$$

and

$$P(\text{tails}) = 0.5 \cdot 0.9 + 0.0 \cdot 0.1 = 0.45. \tag{14}$$

Now, we have all the numbers to run Bayes' Theorem, and we get:

$$P(\text{Greg's.Model}|\text{heads}) = 0.5 \cdot 0.9/0.55 = 0.8181 \tag{15}$$

$$P(\text{Chilin's.Model}|\text{heads}) = 1.0 \cdot 0.1/0.55 = 0.1818 \quad (16)$$

and also two equations we can ignore once we know we have heads:

$$P(\text{Greg's.Model}|\text{tails}) = 0.5 \cdot 0.9/0.45 = 1.0 \quad (17)$$

$$P(\text{Chilin's.Model}|\text{tails}) = 0.0 \cdot 0.1/0.45 = 0.0. \quad (18)$$

Note that (if we had gotten tails), Chilin's Model would have been excluded: it always predicts heads. Therefore, any observation of tails kills that model. Bayes' Theorem agrees: it predicts $P(\text{Chilin's.Model}|\text{tails}) = 0.0$.

However, we *actually* observed heads. That strengthens the case for Chilin's Model, raising it's probability from 0.1 (the prior value) to 0.1818 (after observing heads). Chilin's Model is still not nearly proved though: it takes more than one coin flip to make fraud likely.

9.1 Bayes' Theorem Completely Expanded.

Just for the record, here is an example of fully-expanded Bayes' Theorem expression for two models:

$$\begin{aligned} P(M_1|D,C) &= \frac{P(D|M_1,C)\cdot P(M_1|C)}{P(D|M_1,C)\cdot P(M_1|C)+P(D|M_2,C)\cdot P(M_2,C)} \\ P(M_2|D,C) &= \frac{P(D|M_2,C)\cdot P(M_2|C)}{P(D|M_1,C)\cdot P(M_1|C)+P(D|M_2,C)\cdot P(M_2,C)} \end{aligned} \quad (19)$$

and for three models:

$$\begin{aligned} P(M_1|D,C) &= \frac{P(D|M_1,C)\cdot P(M_1|C)}{P(D|M_1,C)\cdot P(M_1|C)+P(D|M_2,C)\cdot P(M_2,C)+P(D|M_3,C)\cdot P(M_3,C)} \\ P(M_2|D,C) &= \frac{P(D|M_2,C)\cdot P(M_2|C)}{P(D|M_1,C)\cdot P(M_1|C)+P(D|M_2,C)\cdot P(M_2,C)+P(D|M_3,C)\cdot P(M_3,C)} \\ P(M_3|D,C) &= \frac{P(D|M_3,C)\cdot P(M_3|C)}{P(D|M_1,C)\cdot P(M_1|C)+P(D|M_2,C)\cdot P(M_2,C)+P(D|M_3,C)\cdot P(M_3,C)} \end{aligned} \quad (20)$$

As usual, M_1, M_2, M_3, ... are the models, D is the data that you have observed, and C represents all the conditions and knowledge surrounding the experiment.

10 Doing it again.

If we flip the coin again, what can we do? Now, we can just take the probabilities that result from the first coin flip, and use them as priors for the second flip. If we keep getting heads, eventually $P(\text{Chilin's.Model})$ will get close to one, and we will start to suspect that the coin is unfair.

11 Bayes' Theorem and a Naturally Suspicious Mind.

It is really Bayes Theorem (or some in-born mental equivalent) combined with models of likely techniques for cheating that makes us suspicious when a series of coin flips goes HHHHHHHHHHHH....

We aren't suspicious because HHHHHHHHHHH... is improbable. After all, the probability of HHHHHHHHHHH... is the same as HTHTHTHTHTHTHT... or, even, HTHHTTHHHHTHTTTHTTH.... In each case, the probability of getting that precise sequence is just 0.5^N, where N is the length.

We are suspicious because we have models for cheating, like Chilin's.Model, that produce long sequences of heads or tails. We know how people could produce long sequences of heads, we know why people might want to do so, and we know that people sometimes cheat. So, prior to any observation, it's not *too* improbable that HHHHHHHHHHH... will occur.

On the other hand, it's not so easy to think why someone would want to produce HTHHTTHHHHTHTTTHTTH..., and it's harder to actually make that sequence happen. You can't just use a two-headed coin, and (even if you could reliably control the flip of a coin) it would be easy to become confused and accidentally produce HTHHTTHHHHTHTTTHTTH... instead. So, we do not have a mental model of cheating that produces HTHHTTHHHHTHTTTHTTH... particularly often, and therefore we do not suspect cheating when we see that sequence.

12 Web pages

- http://www.anu.edu.au/nceph/surfstat/surfstat-home/
- http://plus.maths.org/issue9/news/banks/
- http://ic.arc.nasa.gov/ic/projects/bayes-group/html/bayes-theorem.html
- http://members.aol.com/johnp71/bayes.html : A little web calculator to do Bayes' Theorem for you.

13 Acknowledgements

Thanks to Peet Morris for comments and corrections.

References

Popper, K. (1959). *The Logic of Scientific Discovery.* Routledge, London, New York.

St.Andrews (2003). *Bayes.* School of Mathematics and Statistics, University of St Andrews Scotland, http://www-gap.dcs.st-and.ac.uk/ history/Mathematicians/Bayes.html. Edited by John O'Connor and Edmund Robertson.

UCalgary (2003). *Bayes Theorem.* University of Calgary, Department of Mathematics and Statistics, Division of Statistics and Actuarial Science, http://balducci.math.ucalgary.ca/.

Who needs condition C?: Bayes theorem can be written as

$$P(M|D,C) = P(D|M,C) \cdot P(M,C)/P(D,C)$$

or as

$$P(M|D) = P(D|M) \cdot P(M)/P(D).$$

(Here, M is a particular model and D is the observation of particular data.) While the second form is equivalent, it doesn't emphasise the need to evaluate $P(M)$ and $P(D)$ under the same conditions.

The conditions for $P(M)$ are fairly clear to everyone, even if they aren't always willing to give a number to it: it is your estimate of the probability of model M, before you see the data. Sometimes that's a well-defined number, sometimes (depending on the statistician) it can be considered either an educated guess or an ill-defined quantity.

On the other hand, it is all too easy to assume that $P(D) = 1$. We always get *some* measurement after all, don't we? (Neglecting, of course, the few cases where the dog eats the printouts and the upstairs plumbing leaks into the computer.) However, there can be two mistakes hidden in this easy answer:

- The event D must be consistent in all three places it appears. Since $P(D|M)$ is the probability of getting a particular datum from a particular model, $P(D)$ must also be the probability of getting one particular datum – a *particular* value from the measurement.

  It may be better to expand $P(D)$ into P(Sentence 4 is grammatical) to make it clear that there are at least two possible values of D. The sentence could be considered either grammatical (D_1) or not (D_0). So, we are not talking about the probability of getting just any data, but instead, certain specific data: one of several (or many) alternatives.

- $P(M)$ is really $P(M|C)$ and $P(D)$ is really $P(D|C)$. In other words, $P(M)$ and $P(D)$ must be obtained under the same conditions, including the same state of knowledge about the universe. So, since we estimate $P(M)$ *before* the measurement of the data, we must also estimate $P(D)$ *before* we measure the data.

  It is trivially obvious that (before the measurement), many different data values are possible, so the probability of getting any particular data value must be less than one.

It may seem doubly hard and silly to estimate $P(D)$ before the measurement, considering that you're already estimating $P(M)$ before you measure any data. However, fortunately, we can compute $P(D)$ from numbers that we already need to have around, specifically the various $P(M)$ numbers and the various $P(D|M)$ numbers.