Hello Bayes Theorem!

George Rebane

Bayesian analysis lets us take existing knowledge or beliefs, combine them with new information, perhaps derived from an experiment or even a conversation, and update our knowledge or beliefs. Bayes works on things which are not known for certain, and that includes about everything in this universe. Everything is known to some level of reliability (or unreliability) that may be measured probabilistically. Probabilities range from zero to one, the former being ascribed to impossible or non-events, and the latter to certain events. (*The Probability Tutoring Book* (1993) by Carol Ash is a good text on elementary probability, and there is always the web where you can google anything.)

This universe may be seen as an awesome ongoing effervescence of events that rise and disappear in time. The events are complex in that they can each have uncounted attributes defining them. For example, {I threw '5' on a die, and drew '9 clubs' from a deck of cards} and {the sun rises, Harry gets newspaper from driveway, the phone rings, …} are two events that each have several contemporary attributes. Another event spread out in time might be {Sam got sick, was diagnosed correctly, received treatment that didn't work, Sam died, …}.

These and myriads of other events are constantly occurring in the dynamic that is our universe. And each of them can be explained or predicted by some set of deterministic or probabilistic rules. Here we look at the working of Bayes Theorem, arguably one of the most powerful rules that underlies our universe that man's intellect has discovered.

Figure 1 shows a sample, perhaps taken over a time interval, of such a myriad of events – the red dots - that occur in a 'universe' (U) of events represented by the rectangle. The indicated universe doesn't have to be The Universe we usually think of, full of galaxies and stuff, but it can be. For most purposes the universe of possible events is a much smaller one like today's population of Chile, or the price histories of all NYSE stocks over the last twenty years. The main thing here is that you can conceive of a more circumscribed universe that still contains all the events from which you want to extract useful information.

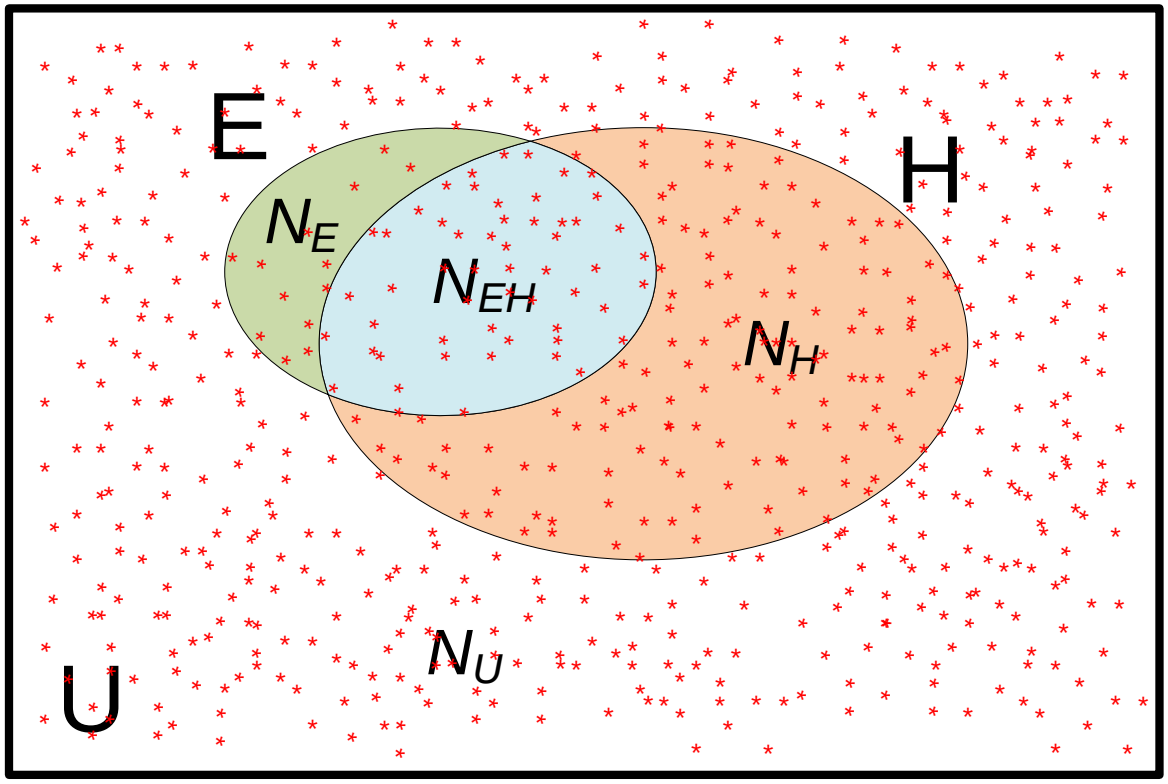

Figure 1

The total number of such events of interest is N_U which includes all the red dots shown. In this universe there occur a subset of events that also have some attribute H that we gather into the big ellipse, and another subset of events that have some attribute E which we have gathered into the smaller ellipse. We notice that these ellipses overlap or intersect (blue area) giving us three distinct regions that comprise the totality of the E and H events. Were we able to count these events, we'd find that N_E (green area) of them are unique to the E events alone, containing no H attribute. And N_H (peach area) of the events are unique to H alone, containing no E attribute.

The events that include both H and E attributes, call them EH events, number N_{EH} and are contained in the intersection of the two ellipses shown. To make sure we don't lose sight of the nature of such a characterization of events, let's make them concrete – the H events are those that include the attribute 'Sally is at home', and the E events include the attribute 'The sound of a radio playing comes from Sally's house'. The EH events are those which have both E and H attributes, namely 'Sally is at home with her radio playing'. And, of course, the universe U here contains all the events whose attributes include Sally's possible whereabouts, and the condition of her radio.

If at random we were to draw out an event from U, what is the probability $P(H)$ that it would be an H event. Intuitively we can see from the figure that $P(H) = (N_H + N_{EH}) / N_U$. It's just the number of such random occurrences in the H ellipse divided by all the relevant possible occurrences in the U rectangle. In the same way we can calculate $P(E) = (N_E + N_{EH}) / N_U$. And extending this process, we can calculate the joint probability $P(E,H)$ that the event includes the joint appearance of both E and H, which from the figure is simply $P(E,H) = N_{EH} / N_U$.

From simple logic we know for certain that any given event, say E, can occur or not occur. And since these events are mutually exclusive – either one or the other, but not both will occur – we know from elementary probability theory that the probabilities of the event occurring, $P(E)$, and not occurring, $P(\neg E)$, must add up to one, the probability of the certain event. Using the symbol ¬ for indicating a logical NOT, we can then write for any event that $P(E) + P(\neg E) = 1$. And, of course, because these occurrences are mutually exclusive, we must have their joint probability $P(E,\neg E) = 0$.

With this under our belts, let's return to the figure and look at how this works out when we put in the indicated numbers for the events.

$$\begin{aligned} P(E) &= \frac{N_E + N_{EH}}{N_U}, \quad P\left(\neg E\right) = \frac{N_U - N_E - N_{EH}}{N_U} \\ P(E) + P\left(\neg E\right) &= \frac{N_E + N_{EH}}{N_U} + \frac{N_U - N_E - N_{EH}}{N_U} = \frac{N_U}{N_U} = 1. \end{aligned} \tag{1}$$

In a similar manner we can see that the H events satisfy

$$P(H) + P\left(\neg H\right) = \frac{N_H + N_{EH}}{N_U} + \frac{N_U - N_H - N_{EH}}{N_U} = \frac{N_U}{N_U} = 1. \tag{2}$$

And using the same line of reasoning, looking at the figure we are able to write the probability of the joint event (E,H) as

$$P\left(E, H\right) = \frac{N_{EH}}{N_U}. \tag{3}$$

The next notion to introduce here is the contingency when one event may or not happen given that another event definitely has happened. In other words we want to find the probability that, say, E happens given that we know H happens. This is a conditional event described by a conditional probability written as $P(E|H)$ – everything to the right of the vertical line is taken to be TRUE, so the uncertain event E is conditioned on H being TRUE.

From the figure, the only region where H happens or is TRUE is in the large ellipse that has a total of $N_H + N_{EH}$ events in it. This means that when we randomly draw out an event now, we are limited to drawing out of only the H ellipse. And in this case there will only be N_{EH} events in there that have the E attribute or for which E will also be TRUE. Proceeding in the usual manner, we can then calculate the desired conditional probability of E given H as

$$P(E \mid H) = \frac{N_{EH}}{N_H + N_{EH}}. \tag{4}$$

The same line of argument can be followed if we desire the probability of H given that E is TRUE. We are now constrained to draw our event sample from only the region defined by the

small ellipse which includes all the events that contain E. Therefore we calculate the conditional probability as

$$P(H \mid E) = \frac{N_{EH}}{N_E + N_{EH}}. \tag{5}$$

Reexamining the joint probability from equation (3) and noting the algebraic form of $P(H|E)$, it occurs to us that we can write $P(E,H)$ as the product of two probabilities that we have already computed, namely

$$P(E,H) = P(H \mid E)P(E) = \left(\frac{N_{EH}}{N_E + N_{EH}}\right)\left(\frac{N_E + N_{EH}}{N_U}\right) = \frac{N_{EH}}{N_U}. \tag{6}$$

Again we can use the same arguments to express this same joint probability of the two events using $P(E|H)$, the other conditional probability we computed in equation (4). The symmetrical argument from the figure confirms this as

$$P(E,H) = P(E \mid H)P(H) = \left(\frac{N_{EH}}{N_H + N_{EH}}\right)\left(\frac{N_H + N_{EH}}{N_U}\right) = \frac{N_{EH}}{N_U}. \tag{7}$$

Since the right hand sides of equations (6) and (7) are equal we can set

$$P(E \mid H)P(H) = P(H \mid E)P(E). \tag{8}$$

Now we solve this simple equation for one of the conditional probabilities, say $P(H|E)$, and we are rewarded with having derived the illustrious and extremely useful formula known as the Bayes Theorem. (Ta Daa!)

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \tag{9}$$

At this point the newly initiated, having gone through the derivation, is wondering what all the fuss is about. Since we put no special restriction on the events in the theorem, we realize that the formula is perfectly general and holds for any two events. The utility of Bayes then comes through the interpretation of its terms. Let's call H an hypothesis and E some kind of evidence from data or experiment. At the beginning we let H be 'Sally is at home' and E be 'the sound of a radio from Sally's house'. Then in (9) we can let $P(H)$ be the prior (also called 'marginal') probability that Sally is at home. This probability is known from some previous knowledge of Sally's wanderings and whereabouts which, of course, includes her being at home. For now, let's say that Sally is home most of the time so that $P(H) = 0.60$.

Since Sally could in fact be just about anywhere – as defined by the events (red dots) in U – we would like to refine our knowledge of Sally's whereabouts by conducting an 'experiment' or incorporating an observation. Namely, we will go to Sally's house to see if there is any evidence that helps us decide one way or the other whether Sally is at home. Arriving, we hear E, the radio playing in Sally's house. How do we incorporate this new evidence E to update our belief

(probability) that Sally is home. Well Bayes in equation (9) tells us exactly how to do it by letting us calculate $P(H|E)$, the probability that Sally is home given that her radio is playing.

From general knowledge or previous data gathering we know that Sally almost always, say 90% of the time, has the radio going when she's home. And that she's also a bit absent minded, and forgets to turn off the radio about one out of five times when she leaves the house. This gives us two needed pieces of information. Remembering how conditional probabilities are defined, we know that $P(E|H) = 0.90$, and $P(E|\neg H) = 1/5 = 0.20$. We are now ready to use Bayes as expressed in equation (9), and plug in what we know.

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} = \frac{0.90 * 0.60}{P(E)}$$

It seems that we have a problem since we don't know $P(E)$, but we do know $P(E|\neg H)$ and this lets us go forward. Actually, from probability theory we know that the marginal probability

$$\begin{aligned} P(E) &= P(E \mid H)P(H) + P(E \mid \neg H)P(\neg H) \\ &= P(E \mid H)P(H) + P(E \mid \neg H)\left[1 - P(H)\right]. \end{aligned} \tag{10}$$

Here it is clear that $P(\neg H)$, the probability that Sally's not home, is just the complement of one minus $P(H)$, the probability that she is at home - see equation (2). So it turns out that we do have all the needed numbers to plug into (9) after we modify it using the expanded formula for $P(E)$ in (10).

$$\begin{aligned} P(H \mid E) &= \frac{P(E \mid H)P(H)}{P(E \mid H)P(H) + P(E \mid \neg H)\left[1 - P(H)\right]} \\ &= \frac{0.90 * 0.50}{0.90 * 0.50 + 0.20 * (1 - 0.60)} = 0.79 \end{aligned} \tag{11}$$

This now tells us that before hearing Sally's radio, our belief (knowledge) was at the 60% level that Sally was home. After incorporating the evidence E that her radio was on, we used Bayes to update our belief to 79% that she is now at home. (As an exercise, recalculate (11) using more and less reliable values for $P(E|H)$ and $P(E|\neg H)$ to characterize the utility of the evidence E as to whether Sally's radio was on or off.)

As an exercise to demonstrate your understanding, you should prove equation (10) by expressing the terms of the equation using Figure 1 and doing the little algebra required. In the meanwhile we can restate Bayes theorem into another useful form by rearranging the right side of (11). We do this by dividing the numerator and denominator by $P(E|\neg H)$. This gives

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E \mid H)P(H) + P(E \mid \neg H)\left[1 - P(H)\right]}$$
$$= \frac{\left\{\frac{P(E \mid H)}{P(E \mid \neg H)}\right\}P(H)}{\left\{\frac{P(E \mid H)}{P(E \mid \neg H)}\right\}P(H) + \left[1 - P(H)\right]} = \frac{L(E \mid H)P(H)}{L(E \mid H)P(H) + \left[1 - P(H)\right]} \quad (12)$$

which is the Bayes Theorem formulated in terms of $L(E|H)$, the likelihood ratio of the 'test' defined in the curly brackets above. The test in our development here was listening for the radio in Sally's house. In the likelihood ratio form, $P(H)$ is sometimes called the 'prior incidence' or 'base rate' of the hypothesis H before any test/experiment is performed to gather evidence E.

The likelihood form of Bayes in (12) emphasizes how new evidence can either reinforce or not the prior belief in an hypothesis. Consider the case when evidence E is such that $P(E|H) = P(E|\neg H)$, the evidence appears with equal probability whether the hypothesis H is true or not. Then $L(E|H) = 1$, and plugging this into (12) gives the expected result that $P(H|E) = P(H)$, the evidence did not affect our belief in the hypothesis. It is clear that the evidence must be such that $L(E|H) > 1$ requiring $P(E|H) > P(E|\neg H)$, for our acceptance of H to increase. If $L(E|H) < 1$, meaning $P(E|H) < P(E|\neg H)$, then our belief in H being true is diminished.

But it's even more straightforward than that. Notice that L in (12) is just the ratio of the seeing the evidence given that the hypothesis is true, to seeing the evidence when the hypothesis is not true. In other words L answers the question 'how many more (or fewer) times does such evidence turn up when the hypothesis is true than when it's false?' We don't even need to know the exact values of the numerator and denominator. We can reason based on our knowledge or experience or even intuition (hunch) about how often or rarely such evidence E turns up to support or degrade our prior belief in H. Please play with the numbers yourself.

In conclusion, we invite the newly initiated into the wonders of Bayes to again look at Figure 1 and consider what Bayes tells us as we move the two ellipses relative to each other – making their intersection larger/smaller, or disappear altogether, and even putting one ellipse inside the other. All this can be done by appropriately modifying the values and locations of the event numbers N_E, N_{EH}, N_H. The discoveries will be illuminating.

