Bayesian Probability Theory:
Inductive (and Deductive) Logic and the Nature of Evidence

By Owen Dix

This discusses: probability theory, logic and evidence (skip forward, if you see fit).

Bayesian probability theory provides a mathematical framework for evaluating inductive logic. Using this, you can show deductive logic is a special case of inductive logic, though it's not hard to grasp this through other arguments. You can also understand what does and does not count as "evidence".

Probability Background

First, some less-than-complete background in probability theory: Bayes theorem is very easily derivable from basic probability theory. The Bayesian interpretation of what a probability is, is very different from other interpretations, for example, the frequentist interpretation. The two do help inform each other, though. You've heard of probability before, sometimes used synonymously (though not technically) with chance, likelihood, etc.

Bayesian interpretation says that that probability that something is true, or something occurring, or something taking on a certain value, is equivalent to the belief that someone has that that thing is true, or that thing occurs, or that thing takes on a certain value.

Notice that two different people can have different belief levels in something. This means it is subjective: it can vary from person to person. If we are careful with how we wield this mathematical tool, we can remove some of that subjectivity, however, to arrive at reasonable values.

For shorthand, I'll use the symbol P(A) to mean "the probability that A is true", where A is some statement or thing that can be true or false. Likewise, to say "the probability that A is false" I'll use the shorthand P(A'), note the apostrophe. Since A is just a variable, we could use it for something else, too. Perhaps we want A to take on a certain value. Maybe A represents the Area taken up by a group of bullet-holes on a target. Then A would have some value, perhaps in squared-centimeters, and we might want to know the probability of A being a certain size. Virtually any question we want to ask, we can assign the probability of getting different answers based off of our educated beliefs. But, for simplicity, let's talk about propositions: statements that are either true or false.

Now, probability is represented as a number from 0 to 1. It can take on any value between and including those numbers. In this case, a 1 means we are completely confident the thing is definitely true (for example), 0 means we are completely confident the thing is NOT true, and 0.5 means we have absolutely no clue. You can also represent these as a percent out of 100%: 0 becomes 0%, 0.5 becomes 50%, and 1 becomes 100% confidence level. You still need to interpret 50% as being completely unsure, and 0% as being sure the thing is false, however. Because of this, and other nice reasons, lets stick with the range from 0 to 1.

It's important to note that all the probability assigned to all the values that A can take on, in our examples A is either true or false, have to add up to 1: P(A) + P(A') = 1. So if P(A) = 0.3, then P(A') is 0.7 because you KNOW it has to be one of them. In terms of percent, you are 100% confident it's either true or false.

Now, no beliefs exist in a vacuum, and since belief is represented as a probability, this means all probabilities that something is true is based on some background information – perhaps it's the set of all your education and experience throughout your life. So, really, we should write P(A|I) where I is the set of all background ideas informing our beliefs, and the vertical bar means "given that" or "conditional on". In classroom practice your background information often comes from the problem statement. These are called conditional probabilities.

Conditional Probability

We can relate conditional probability to some other things and figure out this Bayes's theorem at the heart of this discussion.

Say we want to look at whether two events, A and B, occur at the same time. We would write the probability of this being true as P(AB|I), read as "the probability that both A and B occur given that our background information I is true). We can relate this to the conditional probability:

$$P(AB|I) = P(A|BI)P(B|I) = P(B|AI)P(A|I)$$

The term P(A|BI) is read as "the probability that A is true given that both B and I are true". Note the symmetry of the conditional probability relationship. This comes from the fact that P(AB|I) = P(BA|I), the probability that both A and B are true is equal to the probability that both B and A are true, which makes sense.

Finally, we find Bayes theorem just by dividing by one of the terms in the last two parts of the equality.

$$P(A|BI) = \frac{P(B|AI)P(A|I)}{P(B|I)}$$

This lets us solve for a possibly difficult term to estimate, depending on the problem, P(A|BI), by using some easier terms to estimate, on the right side of the equation. Since A and B were interchangeable in the equation before this, we could easily swap every A for B and B for A if what those two variables represent is easier to calculate that way. Here's how we can relate it to logic.

Logic

Suppose we look at a logical argument, which is a series of statements separable into the premises and the conclusion. In logic, we assume the premises are true for the sake of argument, in order to find what conclusions are true so long as our logic is valid. At least, this is deductive logic. Inductive logic cannot guarantee a conclusion is true if the premises are true; it only argues the likelihood of the conclusion. This is perfect for probability theory and the concept of

conditional probability is ideal for assuming premises to be true; those premises are the variables that you condition your probability on.

Let's take modus ponens:

A → B
A
B

This reads: If A is true then B is true. A is true. Therefore B is true. Note the implied wording "is true" like we did with probabilities involving propositions. This is deductively true: if the premises are true, the conclusion must be true. An example of this could be, "If Billy eats the stinky sandwich, then Billy will get sick. Billy ate the stinky sandwich. Therefore, Billy got sick." The first two statements are the premises and the third is the conclusion, and the conclusion is as guaranteed as the premises are: it is deductive logic. If we want to use probability theory to look at the probability that our conclusion is true given the premises, it would go something like this.

Call the first premise C, so C = A → B. We want to find P(B|AC), the probability that B is true given that A and C are true. Let's use our relation for conditional probability to get at the problem.

$$P(B|AC) = \frac{P(AB|C)}{P(A|C)}$$

For the numerator, we need to evaluate the probability that both A and B are true. From the modus ponens note that if A is true then A and B are true because B becomes true also, given our first premise, C. If A is false then the quantity A and B is false because one of them is false. So the truthiness of the quantity AB is equal to the truthiness of just A by itself. In other words, AB = A, and therefore P(AB|C) = P(A|C). Substituting this into our last equation, the numerator and denominator cancel leaving:

P(B|AC) = 1.

In other words, if our two premises to the modus ponens are true (C and A), then we are guaranteed that B is true: the probability it is true is 1 (or 100%). Probability theory is consistent with deductive logic.

Let's do another, modus tollens:

A → B
B'
A'

This reads: if A is true then B is true. B is false. Therefore A is false. Note the implied "is false" by the apostrophe as mentioned before. This is deductively true: if the premises are true, the

conclusion must be true. An example of this could be, "If Billy eats the stinky sandwich, then Billy will get sick. Billy did not get sick. Therefore, Billy did not eat the stinky sandwich". Like before, we want to use probability theory to look at the probability that our conclusion is true given the premises. Set C = A→ B again. Here, the more useful equation is:

$$P(A|B'C) = \frac{P(AB'|C)}{P(A|C)}$$

In the numerator, we know that B cannot be false and have A simultaneously be true, so P(AB'|C) = 0. This leave us with:

P(A|B'C) = 0.

In other words, if our two premises are true, then there is 0 chance that A is true: A must be false. Here, again, probability theory is consistent with deductive logic.

Let's take another example:

A → B
B
A (Nope)

This fallacy is called affirming the consequent. An example of this could be, "If Billy eats the stinky sandwich, then Billy will get sick. Billy got sick. Therefore, Billy ate the stinky sandwich." We cannot conclude this because something else could have made Billy sick. Finding B to be true doesn't guarantee A is true. Deductive logic fails us here. But probability theory doesn't. We are looking for P(A|BC), where C is the same as in our past examples. Let's look at Bayes theorem to help us.

$$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|C)}$$

The first term in the numerator relates to the result of our modus ponens argument: P(B|AC) = 1. The two premises C and A being true guarantees that B is true. This leaves us with:

$$P(A|BC) = \frac{P(A|C)}{P(B|C)}$$

Note that all probabilities are values between 0 and 1, so

$$P(B|C) \leq 1$$

We can do better, though. We know the content of C, it is equal to the statement If A is true then B is true. Certainly this statement, alone with no others, does not guarantee B to be true. Thus P(B|C) cannot be 1. So,

P(B|C) < 1

Substituting this back into the denominator in the equation above, we find we are dividing by a number less than 1, which means:

P(A|BC) > P(A|C)

The difference between the two terms is that the left one conditions on B being true and the right one doesn't take B into account at all. We should interpret this as saying that learning that B is true makes A more likely. B provides evidence that A is true, but it does not guarantee it! This is inductive logic. In our running example, we could say "If Billy eats the stinky sandwich, then Billy will get sick. Billy got sick. Therefore, learning this means it's more likely (makes us more confident) that Billy ate the stinky sandwich."

Last example:

A → B
A'
B' (Nope)

This fallacy is called denying the antecedent. Finding A to be false doesn't guarantee B is false since something else could be causing B to be true, instead. For example, "If Billy eats the stinky sandwich, then Billy will get sick. Billy did not eat the stinky sandwich. Therefore, Billy did not got sick." We know something else could get Billy sick besides eating a stinky sandwich. Let's use Bayes theorem again and with a little hindsight and foresight, we pick this form.

$$P(B|A'C) = \frac{P(A'|BC)P(B|C)}{P(A'|C)}$$

Our last example concluded that P(A|BC) > P(A|C), but we know that P(A|BC) = 1 – P(A'|BC) because they have to add to 1. It's similar for the P(A|C). Plugging these in:

1 - P(A'|BC) > 1 – P(A'|C), or

P(A'|BC) < P(A'|C).

You can interpret this like we did in the last example: since learning B is true makes it more likely that A is true, then learning B is true makes it less likely that A is false. This means the first term in the numerator is less than the denominator of our Bayes theorem equation. As a ratio, these two terms are less than one. Therefore P(B|A'C) is equal to some number less than one times P(B|C), or:

P(B|A'C) < P(B|C),

Which you can interpret as saying learning that A is false makes you less confident B is true. Again, this is like saying learning that A is false makes you more confident B is false (but

doesn't guarantee it). Look back at our logical structure. What we can really conclude in our running example is, "If Billy eats the stinky sandwich, then Billy will get sick. Billy did not eat the stinky sandwich. Therefore, learning that makes it more likely (makes us more confident) that Billy will not get sick."

One other way of saying it is that the knowledge that Billy did not eat the stinky sandwich is "evidence" that Billy will not get sick. We turn now to the nature of evidence.

Evidence

After you make a good point, have you ever heard someone say "that's not REAL evidence!", as if they have some better understanding at what evidence is at its core. Well maybe they were right. I'm not advocating that there's no such thing as a good, consistent, objective definition of evidence. But evidence can be an ambiguous term. Probability theorists define it one way. I will use a related way that fits better with how the word is more commonly used. After some reflection, you should find it quite up to par. It comes from Bayes theorem.

Evidence doesn't exist in a vacuum, just like our beliefs don't. They have background information that helps us interpret them and they always have to be attached to some, really many different hypotheses. If a hypothesis is just a statement, a claim that could be true or false, then it fits what we were using as variables, A and B, before. This time, we'll call the hypothesis H, for which our particular piece of data D ***might*** be evidence for (it might be evidence against it, instead, or just have no bearing on the hypothesis at all).

We usually, ultimately, want to know if some hypothesis is true. The hypothesis H could stand for the claim that a certain drug treats a certain disease, or that aliens have visited us, or that we are really actually a bad driver, or, in general, anything that can be true or false and has some possible, conceivable data that could inform us about its truthiness. This criteria means the hypothesis has to be testable to be called a hypothesis. Some data needs to potentially exist that could make it more or less likely to be true. Now in reality we usually constrain that data to be empirical in nature because it seems to give more reliable information about things that exist in our universe. However, arguments do play a role, they can influence the truth of a claim, and certainly need to be used to interpret the empirical data. Nevertheless, if we want to know if some hypothesis H is true after collecting some data D, and given our background information I, we need this form of Bayes theorem.

$$P(H|DI) = \frac{P(D|HI)P(H|I)}{P(D|I)}$$

P(H|DI) is read as the probability that our hypothesis is true given the data we get and our all our background information. P(H|I) is known as the prior probability of the hypothesis: how likely the hypothesis is to be true given all our background information. It makes sense that the ratio of P(D|HI) over P(D|I) is called the likelihood because it tells us how to modify our prior probability about the hypothesis to incorporate this new data, D. P(H|DI) is often called the posterior probability, for this reason, because its your confidence in the hypothesis after incorporating new data.

Since the likelihood is ratio of two numbers between 0 and 1, dividing them could make a small number or a large number. The posterior probability is still always bounded by 0 and 1 but multiplying the likelihood by the prior probability can make the posterior probability grow by comparison. It can make us more or less confident in the hypothesis than we were before learning the data. This takes us to a tentative and intuitive definition of evidence:
Evidence for a hypothesis H is any data D that makes us more confident the hypothesis is true.

This Bayes's theorem equation actually lets us do something very clever and very important. We can take the posterior probability for H after some data D (left hand side term) and stick it back into the equation as the prior probability so we can evaluate some new data E. This means we can incorporate as much data as we can to continually update our beliefs as new evidence comes in. This is what a rational person should do: proportion your beliefs to the evidence!

Even still, we can say more about evidence than this with a little more work, though, by looking closer at the likelihood term.

$$\frac{P(D|HI)}{P(D|I)}$$

P(D|HI) is the chance of getting the same data D if we live in a universe where H is really true. This is where most of the probability and statistics work comes in, to inform us on this possibility. P(D|I) is how likely we expect to get the data D regardless of whether the hypothesis is true or false. Recall that P(A|B) is read as "the probability that A is true given B is true". P(D|I) somehow takes into account whether the hypothesis is true AND whether the hypothesis is false. Those familiar with basic probability theory can expand this to see how it does it, but there's a more clear way to get to the nature of what it means to legitimately call something evidence: look at the probability that the hypothesis is <u>false</u> given the data, with Bayes theorem.

$$P(H'|DI) = \frac{P(D|H'I)P(H'|I)}{P(D|I)}$$

Both involve the same denominator but the numerators are different. Let's divide our first Bayes theorem equation by our second to get the "odds form" of Bayes theorem, canceling the common denominator in the process.

$$\frac{P(H|DI)}{P(H'|DI)} = \frac{P(D|HI)}{P(D|H'I)}\frac{P(H|I)}{P(H'|I)}$$

The left hand side is greater than 1 if it's more likely the hypothesis is true than if it's false, given the data (and our background information). It's less than one if its more likely the hypothesis is false, given the data. Note that P(H|DI) + P(H'|DI) = 1 so normally if a probability was greater than 0.5 we'd say its more likely to be true than false but this odds form benefits us in clarity when we get to the right hand side of the equation.

The second ratio on the right hand side is the ratio of our prior probabilities. If there is no information about a hypothesis's truthiness at all, then its equally likely the hypothesis is true or

false and this ratio will just be 1. This ratio tells us how likely the hypothesis is before we start to consider the data D. This takes us to the first ratio on the right hand side.

The first ratio on the right hand side is the ratio of the probability of getting the data D given if the hypothesis is true, versus given the hypothesis is false. If the data D is more likely given the hypothesis is true than if the hypothesis is false, then this number will be greater than 1. And THIS, ladies and gentlemen, is the key to evidence!

- If D is more likely given the hypothesis is true, than if the hypothesis is false, then D counts as evidence for the hypothesis H.
- If D is more likely given the hypothesis is false, than if the hypothesis is true, then D counts as evidence against the hypothesis H.
- If D is equally likely given the hypothesis is true or false, then D does not count as evidence for or against the hypothesis H.
- If D is more is a LOT more likely given the hypothesis is true, than if it is false, then D is strong evidence for the hypothesis H.
- If D is more is only a little more likely given the hypothesis is true, than if it is false, then D is weak evidence for the hypothesis H.

For example, physical evidence for a crime is usually stronger than circumstantial evidence. Circumstantial evidence is considered weaker because a chain of inference makes it only support the hypothesis (person X is the killer) if we assume some chain of events happened. For every assumption we need to make, the probability of the data given the hypothesis goes down.

Anecdotal "evidence" is a lot like this. Since anecdotes, or stories in our everyday experience, aren't like scientific studies, which control for bias and other variables and have lots of trials. Anecdotes are one data point that can easily either be an outlier or be more easily due to bias in interpretation or other variables that a study would control for. All the other possible explanations for the data, the sum of all other hypotheses that could have also led to that data, means the data is not more likely due to the hypothesis being true than to the hypothesis being false (and some other hypothesis, some other explanation for the data, is true). So anecdotal data is weak evidence, at best – being most charitable to the hypothesis – and no evidence, at worst.

Here's the rub. I mentioned at the very beginning that there is some subjectivity in evaluating these probabilities. Different people could assign different numbers. That's why some people discount things as not being evidence that might hold sway to others. Anecdotes are very emotionally persuasive to us as human beings – but they shouldn't be. They should, however, lead us to investigate further. This means we collect new data, think about whether our hypothesis explains the data better than all the other hypotheses out there, and (trying to be as unbiased as we can) update our beliefs in H, whatever it is we want to know.

Knowledge can be defined as a well justified, true, belief. It may be that thinking about evidence and likelihoods can be hard but it can solidify the ground your beliefs rest upon and help you legitimately call some of what you believe "knowledge". This brings me to the last bit of advice, a tweaked quote by philosopher David Hume (1711-1776): A wise person proportions their beliefs to the evidence.

An Example of Evidence

I know that was very abstract. Let's do an example. Let the hypothesis, H, stand for the following:

H = I am a good driver
H' = I am not a good driver

This is what I want to know, the likelihood that H is true (I am a good driver) or H is false (I am not a good driver). Note that H' stands for the statement that H is false, so the hypothesis that I am not a good driver. Let's consider some data D:

D = I got into an accident.

For now, this is all the data we have about my driving. In a minute, we will look at a different piece of data. The odds-form of Bayes theorem says, for convenience:

$$\frac{P(H|DI)}{P(H'|DI)} = \frac{P(D|HI)}{P(D|H'I)}\frac{P(H|I)}{P(H'|I)}$$

The right-most ratio is the prior probability in odds form. If P(H|I)/P(H'|I) = 1, this is only true if P(H|I) = P(H'|I) = 0.5. This means we start from a clean slate about whether or not I am a good driver. This might be the fairest position to take, in this case. We could think about what fraction of all drivers are "good" drivers, for some reasonable definition of good and set P(H|I) equal to that, but what we really care about is how D affects the probability that I am a bad driver.

The likelihood-ratio in odds form is the left ratio on the right side of the equation. If P(D|HI) > P(D|H'I) then the ratio is bigger than one and the odds of me being a good driver increases due to D – you may want to convince yourself that this is true by looking at the equation.

Well, P(D|HI) is the probability of having an accident assuming that I am a good driver. P(D|H'I) is the probability that I got into an accident assuming I am a bad driver. I think we can agree, that even though good drivers get into accidents too, bad drivers are more likely to get into accidents. Bad drivers probability get into as many accidents as good drivers, on average, that are not their fault, plus the ones that they could not avoid because they are bad drivers, plus the ones they caused because they are bad drivers.

All this means that P(D|H'I) is greater than P(D|HI). So the overall ratio is less than one, and the odds of me being a good driver go down after having an accident, any accident. They may not go down much, but they do go down. This means my confidence in the idea that I am a good driver should go down, since we interpret the probability as my belief in the hypothesis being true.

Let's modify D with a new, tweaked example:

D = I got rear ended while stopped at a stop light with other cars in front of me

Now let's look at the likelihood-ratio in odds form. P(D|HI) is the probability of this happening given that I am a good driver. It's pretty clear, I probably could not do anything to avoid the accident. It was not my fault, it was the person who hit me. It's probably roughly as likely for this to happen if I am a bad driver as I am a good driver. Maybe very occasionally the rear-ender will do it out of spite because the person they hit cut them off (a bad driver), so P(D|H'I) might be slightly greater than P(D|HI) but that seems fairly negligible.

In this case, since P(D|HI) ~ P(D|H'I), the likelihood ratio in odds-form is roughly 1 so the left-hand side is unaffected by this new data. The data is not evidence for or against the hypothesis that I am a good driver.

Now, for one a little more complicated. What if we assume that an accident-worthy event occurs as part of our background information, I, and D is the following:

D = I avoided the accident

P(H|DI) stands for the probability that I am a good driver given that an accident-worthy event occurs and I avoided the accident. We are assuming that an accident-worthy event occurs but we are not concerned with how this affects H, just how avoiding it does (D).

Given that there is an accident-worthy event, it's probably more likely that good drivers would avoid the accident (with a reasonable definition of "good", such as one who is skilled, attentive, and law abiding). And bad drivers would be less likely to do so. You may argue, someone could be a bad driver because they aren't law abiding but they're still skilled and attentive so they may be just as likely to avoid the accident. True. But the pool of all "bad drivers" includes those people, but also the inattentive and/or the unskilled drivers. They bring the whole group down.

In this case, P(D|HI) is greater than P(D|H'I); so the likelihood-ratio in odds form is greater than one. And D, now, counts as evidence that I am a good driver.

Bayes-theorem, it turns out, is the only consistent way to define and a belief that some hypothesis is true, and to modify that belief given new evidence (according to the proof in E.T. Jaynes classic book, Probability Theory: The logic of science). If you'll notice, all of these conclusions are actually pretty reasonable, so it passes the gut-feeling test. If you can incorporate it into your thinking, it can serve you well.