PHIL 50 – INTRODUCTION TO LOGIC

MARCELLO DI BELLO – STANFORD UNIVERSITY

PROBABILITY AND LOGIC – WEEK #10

In this note I want to make two points about the relation between probability and logic. The first point is that probability theory needs an underlying logic, typically classical logic. The axioms of probability theory would not make sense without a suitable logic underlying them. Second, it is an open question how to characterize a notion of probabilistic validity that is as systematically and formally defined as the notion of deductive validity. The task of defining probabilistic validity is the subject of what we might call probability logic.

1 PROBABILITY THEORY

We begin with the mathematics of probability. A probability function P is a function from a set of formulas into the real numbers and it follows the Kolmogorov Axioms:

> NORMALITY: $0 \leq P(\varphi) \leq 1$, for any formula φ;
>
> CERTAINTY: $P(\top) = 1$, with $\top$ any logical tautology; and
>
> ADDITIVITY: $P(\varphi \vee \psi) = P(\varphi) + P(\psi)$, provided $\varphi \wedge \psi$ is a contradiction.

In other words, a probability function assigns to every formula a real number between 0 and 1. It assigns the value 1 to tautologies. And to the disjunction of contradictory formulas, it assigns the sum of their probabilities.

Now that we have defined $P(\varphi)$, we can define the *conditional probability* of φ given some other formula ψ, as follows:

> CONDITIONAL PROBABILITY: $P(\varphi|\psi) = \frac{P(\varphi \wedge \psi)}{P(\psi)}$.

The expression $P(\varphi|\psi)$ should be read 'φ conditional on ψ'. For example, consider the statement '*the rain is imminent*' and '*there are dark clouds in the sky*'. It is natural to ask the conditional probability of '*the rain is imminent*' given that '*there are dark clouds in the sky*', which in symbols reads $P($*the rain is imminent*$|$*there are dark clouds in the sky*$)$.

Why do we define conditional probability as $\frac{P(\varphi \wedge \psi)}{P(\psi)}$? The conditional probability of φ given ψ is the probability of having both φ and ψ divided by the probability of having ψ. Here ψ plays the role of the assumption that is taken to be true and against which the probability of φ is determined. We want to know the probability of φ assuming that ψ is true. And that amounts to saying that we want to know the probability of $\varphi \wedge \psi$ assuming that ψ is true, namely $\frac{P(\varphi \wedge \psi)}{P(\psi)}$.

Some lemmas can now be proven, such as:

OVERLAP: $P(\varphi \vee \psi) = [P(\varphi) + P(\psi)] - P(\varphi \wedge \psi)$

EQUIVALENCE: If φ and ψ are logically equivalent, $P(\varphi) = P(\psi)$

LOGICAL CONSEQUENCE: If $\varphi \models \psi$, then $P(\varphi) \leq P(\psi)$

NEGATION: $P(\neg\varphi) = 1 - P(\varphi)$

TOTAL PROBABILITY: $P(\varphi) = P(\varphi|\psi)P(\psi) + P(\varphi|\neg\psi)P(\neg\psi)$

2 THE UNDERLYING LOGIC

It is worth distinguishing probability theory from its underlying logic. Every formulation of a theory of probability, such as the one above, must rest on an underlying logic, typically classical logic. Some theorems fail if the underlying logic is non-classical.

For instance, to prove the negation theorem, one begins with the statement that $P(\varphi \vee \neg\varphi) = 1$, which holds because $\varphi \vee \neg\varphi$ is a classical tautology. It then follows by additivity that $P(\varphi) + P(\neg\varphi) = 1$, whence $P(\neg\varphi) = 1 - P(\varphi)$. The proof gets off the ground because $\varphi \vee \neg\varphi$ is a classical tautology. If the underlying logic were intuitionistic, $\varphi \vee \neg\varphi$ would not be a tautology, and thus the negation rule would not be a theorem.

Similarly, to prove the theorem of total probability, one relies on $P(\varphi) = P(\varphi \wedge \psi) + P(\varphi \wedge \neg\psi)$, and since $P(\varphi \wedge \psi) = P(\varphi|\psi)P(\psi)$ and $P(\varphi \wedge \neg\psi) = P(\varphi|\neg\psi)P(\neg\psi)$, the theorem follows immediately. But the equivalence $P(\varphi) = P(\varphi \wedge \psi) + P(\varphi \wedge \neg\psi)$ holds because, first, φ and $(\varphi \wedge \psi) \vee (\varphi \wedge \neg\psi)$ are (classically) logically equivalent, so they must have the same probability; and second, $\varphi \wedge \psi$ and $\varphi \wedge \neg\psi$ are inconsistent so the probability of their sum must be the same as the probability of their disjunction. And yet, in intuitionistic logic φ can be true without neither $\varphi \wedge \psi$ nor $\varphi \wedge \neg\psi$ being true.

3 THE MEANING OF PROBABILITY

The mathematical treatment of probability is neutral as to what probability values express or what they mean. I shall give here a brief overview of the main interpretations. The most natural interpretation of probability is the *classical* interpretation. A statement of it is given by Laplace (1814) as follows:

> The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability.

What does it mean to say that some cases are equally possible or equally probable? Laplace takes the notion of equi-possibility or equiprobability as primitive, or at best he says that two cases are equally probable when we are 'equally undecided about their existence.' In developing this idea, we can say that two cases are equally probable if the evidence in favour of (or against) each of them is perfectly balanced. (Incidentally, note the ambiguity here between 'we lack evidence for and against proposition A' and 'we have equally strong evidence for and against A.' In both cases, the evidence is perfectly balanced although for very different reasons: lack of evidence in one case and equally strong evidence in the other.) We can also say that two cases are equally probable if they are physically symmetric. Be that as it may, spelling out the notion of equiprobability requires us to invoke a *principle of indifference*, in an epistemological or physical sense.

Moving away from the classical interpretation, some theorists believe that probabilities are *objective* features of the world. In particular, some authors think that probabilities apply to classes of events. They equate probabilities to *relative frequencies* in the case of finite classes of events, or to *limiting relative frequencies* in the case of infinite classes of random events, known as collectives. The frequentist interpretation of probability makes good sense for events that can be repeated in the long run, but it is at odds with probabilities assigned to single-case events.

Other theorists hold that probabilities are not objective, but *subjective* or more generally *epistemic*. Loosely put, the idea is that the probability of a proposition corresponds to an *agent's degrees of belief* in a proposition. In an attempt to spell this out, one can say that degrees of belief in a proposition mirror the strength of one's evidence for that proposition, or more precisely, they mirror what an ideally rational agent takes to be the strength of one's evidence.

We should stop here. We cannot enter into a very difficult debate regarding the foundations and the meaning of probability.

4 BAYES' RULE

Suppose you formulated a hypothesis H and have some evidence E for it, and now you want to know the probability of H given E. Bayes' rule gives you the answer, as follows:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\neg H)P(\neg H)}.$$

Bayes' rule allows us to calculate the probability of H given E from three other items: (i) the probability of H regardless of E; (ii) the probability of the evidence $P(E)$ which, by the rule of total evidence, equals $P(E|H)P(H)+ P(E|\neg H)P(\neg H)$; (iii) the likelihood $P(E|H)$, i.e. the probability of E given H. Probability $P(H)$ measures the probability of H before taking into account the new piece of evidence E. Probability $P(E)$ is a measure of how unusual or surprising it is to obtain E. Finally, likelihood $P(E|H)$ is a measure of how much the hypothesis H "accounts for" or "predicts" E.

(The term 'likelihood' is an unfortunate word choice, but one that is now universally made in the literature. While in common parlance 'probability' and 'likelihood' are interchangeable terms, within the Bayesian framework the term 'likelihood' designates the probability of the evidence given the hypothesis, $P(E|H)$, as opposed to the probability of the hypothesis given the evidence $P(H|H)$.)

5 MEDICAL DIAGNOSIS

Let us now see an example of how to use Bayes' rule. Suppose you get tested for a certain disease and the test is positive. What is the probability that you do in fact have the disease? In order to calculate this probability, Bayes' rule requires information about the reliability of the test as well as information about the base rate of the disease in a reference population. As for the reliability of the test, let's suppose the test has a reliability of 0.8. This means that the test gets it right 80 percent of time. If you have the disease, the test will come out positive 80 percent of time, and if you do not have the disease, the test will come out negative 80 percent of time. This is what it means for the test to be 80 percent reliable. As for the base rate of the disease, let's suppose that the disease in question is had by 15 percent of the reference population.

What we want to know is, how likely are you to have the disease given that you tested positive? Let D abbreviate the statement '*the patient has the disease*'. Let T^+ abbreviate the statement '*the patient tested positive*' and T^- abbreviate the statement '*the patient tested negative*'. What we want to know, then, is the value of $P(D|T^+)$, namely the probability of D conditional on T^+. By Bayes' theorem, we know that

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|\neg D)P(\neg D)}$$

To determine the value of $P(D|T^+)$, we need the values of $P(D)$ and $P(\neg D)$ and the value of $P(T^+|D)$ and $P(T^+|\neg D)$.

Given the suppositions we made earlier, the disease in question is had by 15 percent of the reference population, so $P(D) = 0.15$, so $P(\neg D) = 0.85$. This is the probability that someone in the reference population—picked at random—has the disease regardless of being tested or not. This is a low probability, but not a negligible probability of having the disease. And the question which Bayes' theorem allows us to answer is, given a base rate value of 0.15 for D, what is the probability of D while taking into account the information that that patient tested positive?

(One might object that a patient either has the disease or she does not. It is either true or false, 1 or 0, that the patient has the disease. One might conclude that there is no need for probability in this context. Well, this suggests that probability here is a measure of our information about the world, and in particular it is a measure of how uncertain we are about whether or not a patient has a certain disease. Objectively, the patient has the disease or does not have it. But as far as our epistemic, informational, or subjective standpoint is concerned, we can be more or less sure about the patient's condition.)

Next, to determine the value of $P(D|T^+)$, we also need the probability value of $P(T^+|D)$ and $P(T^+|\neg D)$. The test was assumed to have a reliability of 80 percent. This means that $P(T^+|D^+) = 0.8$, i.e. there is an 80 percent chance that a patient tests positive given that the patient does in fact have the disease. Further, $P(T^-|\neg D) = 0.8$, i.e. there is an 80 percent chance that a patient tests negative given that the patient does not have the disease. The latter implies that $P(T^+|\neg D) = 0.2$ because $P(\varphi) = 1 - P(\neg\varphi)$.

We are now ready to plug these values into the formula for Bayes' theorem, as follows:

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|\neg D)P(\neg D)} = \frac{0.8 * 0.15}{0.29} =\approx 0.41.$$

The interesting result here is that, even if the test is reliable 80 percent of the time, the probability that the patient has in fact the disease is still quite low. The reason is that the probability of D regardless of the tests result is low and equals 0.15. The table below shows that by varying the probability of D we can arrive at different probabilities of D given a positive test holding fixed the test's reliability:

$P(D)$	$P(\neg D)$	$P(T^+|D)$	$P(D|T^+)$
0.15	0.85	0.8	0.41
0.25	0.75	0.8	0.57
0.35	0.65	0.8	0.68
0.45	0.55	0.8	0.76
0.50	0.50	0.8	0.80
0.55	0.45	0.8	0.83
0.65	0.35	0.8	0.88
0.75	0.25	0.8	0.92
0.85	0.15	0.8	0.95

6 PROBABILITY LOGIC

Probability logic offers an account of *probabilistic validity*, whereas deductive logic offers an account of deductive validity. Deductive logic gives us a theory that answers the question:

> Which reasoning patterns bring us from truth premises to a true conclusion?

In the past several weeks we studied the relation of logical consequence $\varphi_1, \ldots, \varphi_k \models \psi$, where $\varphi, \varphi_2, \ldots, \varphi_k$ are the premises and ψ is the conclusion. When the relation $\varphi_1, \ldots, \varphi_k \models \psi$ holds, that means that whenever $\varphi, \varphi_2, \ldots, \varphi_k$ are true, then also ψ is true. So, relation $\models$ can be described as a relation of *truth preservation*.

The task of probability logic is to identify a relation similar to logical consequence but a relation which holds probabilistically. The task of probability logic is to identify a relation that holds between highly probable premises and a highly probable conclusion. Probability logic should give us a theory that answers the question:

> Which reasoning patterns bring us from highly probable premises to a highly probable conclusion?

Instead of truth preservation, probability logic should focus on *high probability preservation*, i.e. in identifying reasoning patterns that bring us from highly probable premises to highly probable conclusions (contrast this with reasoning patterns that brings us from true premises to true conclusions *simpliciter*). This is by no means an easy task, and unlike deductive logic, there is no agreed theory of probabilistic validity.

Let us conclude with a contrast between deductive logic and probability logic. The (deductive) rule of $\wedge I$ tells us that from two true premises, say φ and ψ, the conclusion $\varphi \wedge \psi$ follows. There is no doubt that the reasoning pattern associated with $\wedge I$ is deductively valid. More generally, from true premises $\varphi_1, \varphi_2, \ldots, \varphi_k$, the conjunction $\varphi_1 \wedge \varphi_2 \wedge \cdots \wedge \varphi_k$ follows. Can we say that if the premises $\varphi_1, \varphi_2, \ldots, \varphi_k$ are highly probable, then the conclusion $\varphi_1 \wedge \varphi_2 \wedge \cdots \wedge \varphi_k$ is also highly probable? The answer is no because of what we might call *the aggregation of risk*. Let's see what this means.

Suppose you have a fair lottery with 1,000 tickets, so the probability that a ticket, any ticket, loses is very high, namely $\frac{999}{1{,}000}$. Let L_i be the statement describing the situation in which ticket number i loses. So $P(L_i)$ is very high, for each statement L_i. Now, is the probability of the conjunction $L_1 \wedge L_2 \wedge \cdots \wedge L_{1000}$ still very high? No, it is not. In fact, the probability of the conjunction is zero, because we know that at least one ticket must win, so it cannot possibly be that all tickets lose. What has happened here? We have:

$P(L_1) = \frac{999}{1{,}000}$

$P(L_1 \wedge L_2) = \frac{999}{1{,}000} \times \frac{998}{999}$

$P(L_1 \wedge L_2 \wedge L_3) = \frac{999}{1{,}000} \times \frac{998}{999} \times \frac{997}{998}$

$\ldots$

$P(L_1 \wedge L_2 \wedge L_3 \wedge \cdots \wedge L_{999} \wedge L_{1000}) = \frac{999}{1{,}000} \times \frac{998}{999} \times \frac{997}{999} \times \cdots \times \frac{1}{2} \times 0$

As more and more conjuncts are added, the probability of the conjunction lowers until it becomes zero. Why is that? To see why, let us ask, what is the probability that ticket number 1 and ticket number 2 both lose? This is the probability that the ticket number 1 loses, namely $\frac{999}{1000}$, multiplied by the probability that ticket number 2 loses (taking into account that ticket number one loses), namely $\frac{998}{999}$. And what is the probability that ticket number 1, ticket number 2, and ticket number 3 all lose? This is the probability that ticket number 1 loses, multiplied by the probability that ticket number 2 loses (taking into account that ticket number 1 loses), again multiplied by the probability that ticket number 3 loses (taking into account that both ticket number 1 and ticket number 2 lose). That is, the probability in question is $\frac{999}{1000} \times \frac{998}{999} \times \frac{997}{988}$. More generally, in a fair lottery consisting of 1000 tickets, for a sequence consisting of a k number of tickets, the probability that all such tickets in the sequence lose is $\frac{999}{1000} \times \frac{998}{999} \times \frac{997}{988} \times \cdots \times \frac{1000-k}{1000-k+1}$.

This was just an illustration that it is not easy to develop a notion of probabilistic validity. It is not enough to take deductive validity and turn it into a probabilistic notion. It is not the case that if a reasoning pattern is deductively valid, the same pattern will also be probabilistically valid. There is a divide between deductive validity and probabilistic validity. As of today, we still lack a well-defined theory of probabilistic validity. If you are interested in these topics, please have a look at the book by Ernest Adam, *Probability Logic*, published by Stanford University Press.