Sets, Relations and Probability. Part IA Formal Methods.

Lecture VIII, *Looking Ahead*, 12th March.

Christopher J. Masterman (cm789@cam.ac.uk, christophermasterman.com)

Last week we looked at how we should calculate the non-conditional and conditional probability of repeated trials with independent events—tossing a coin several times—and the non-conditional and conditional probability of repeated trials with dependent events—drawing a cards *without replacement* from a pack. We ended by looking at how conditional probability is highly sensitive to what we conditionalise on.

Throughout these lectures, we've largely focused on *calculating* probabilities; but discussed the *interpretation* of probability very little. At the beginning, we interpreted probabilities as relative frequencies. In Lecture VI, we briefly looked at Bayesian Epistemology and the subjective interpretation of probability. Today, we'll look at that in a bit more detail, as well as some other ways we might *use* probabilities.

1. Subjective Probability Again

1.1. According to a subjective interpretation of probability, the probability you assign to a certain event E measures the degree of confidence you have in E occurring. For example, if I assign the event of it raining today a low probability, this is because I have little confidence that it will rain today. Initially, you might worry that this is not going to work because you can't quantify something like your degree of confidence that an event occurs. Surely, a degree of confidence is not a precise thing—it is just a *feeling* or a *vibe*. If this is right, you should doubt that you can assign it a precise value between 0 and 1.

1.2. Instead, then, we might think of assigning a probability in terms of your assigning a certain *degree of belief*. If we understand degree of confidence as degrees of belief, we could get around the worry above: beliefs states are propositional states—the content of a belief is a proposition—and propositions may be more accommodating to a precise quantitative analysis. How, then, should we measure our degree of belief that a particular event will occur, e.g., that it will rain today? Well, a good starting place would be to focus on *behaviour*. After all, if you have a high degree of confidence that it will rain today, you will behave differently: you will make sure to take an umbrella or wear a rain coat.

1.3. So far, so good; but noting our behaviour and its connection to our degrees of belief that particular events will occur is clearly not enough—we need a way of extracting a precise numerical measure of our degree of belief. One way to do this is to focus on a restricted class of specific behaviours: our willingness to *bet* that particular events will occur. This approach to probability was first outlined by Frank Ramsey, see (Ramsey, 1990[1926]). To understand this approach, we should first talk about betting odds.

- **Example**. Suppose you want to place a bet on a particular horse winning a race, e.g., *Horsey Boy*. The bookies have odds of $2:1$ on Horsey Boy winning. This means that if you place a £10 bet on Horsey Boy winning and Horsey Boy wins, then you will get:
  - Your initial stake of £10 back; and
  - £20 from the bookies. (That is, $2\times$ your initial stake of £10.)

  Of course, if Horsey Boy does not win, you lose your intial stake of £10 and get nothing else.

If the odds are $A:B$ and you bet successfully, then you receive $\left(\frac{A+B}{B}\right)$ pounds for every pound you bet.

- With a £5 bet on $5:1$, if you win you receive $\left(\frac{5+1}{1}\right)$ pounds for every pound. So, $6\times$ £5 =£30.

– With a £10 bet on $1:5$, if you win you receive $\left(\frac{6}{5}\right)$ pounds for every pound. So, $1.2\times$ £10 = £12.

1.4. The idea, then, is that we assign values to your degree of belief that an event E will occur in terms of the *worst* odds you would be prepared to accept on E, at least when the stakes are small. That is, we first work out the worst odds that you would be prepared to accept on E. Then,

(D) The worst odds you would accept on E are $A:B$ iff your degree of belief that E occurs is $\frac{B}{A+B}$

So, for instance, if the worst odds you would accept on E are $2:1$, then your degree of belief that E occurs is $1/3$. If the worst odds on E are $1:2$, your degree of belief is $2/3$. If the worst odds on E are $1:10$, your degree of belief is $10/11$. Notice that as the odds get *worse*, the degree of belief calculated gets higher. This is intuitive: if you think something is likely to happen, then the worst odds on which you bet on it can be pretty bad; but if you think that something is not likely to happen, you'd only be prepared to bet on it if the odds were good. Another way of thinking about this is that if you're certain that an event will occur, you would accept any bet on it; but if you're certain that an event will not occur, you wouldn't accept any bet on it.

2. Dutch Books: Probability as Rationally Compulsed

2.1 In Lecture VI, we talked about the Kolmogorov Axioms and we said that function is a probability function just in case it satisfies those axioms. Then, using those axioms, we calculated various values for these functions and we called those values *probabilities*. Little was done to motivate these axioms. More importantly for us presently, little has been done to motivate the claim that our degrees of belief conform to these axioms.

2.2. There is an interesting series of arguments which show that our degrees of belief, understood as measures of fair betting odds, should satisfy the axioms—and thus the theorems—of probability theory. Why should they? Well, the argument goes, if they do not satisfy the axioms, then our degrees of belief leave us open to accepting odds which guarantee that we will lose money. These are called *Dutch Book Arguments*.

2.3. In more detail, we say that someone is vulnerable or open to a Dutch Book if there some betting odds, arranged in such a way, that guaranteed them to lose money. For instance, suppose that there is a three horse race and the bookies offer Horsey Boy at $10:1$, Speedy at $10:1$, and Lightening at $10:1$ and you have £3 to spend. If you bet £1 on each horse, then regardless of which horse win, you end up with more money, i.e., if Horsey Boy wins, you get £11 from the bookies and lose £2 quid from the failed bets. So, a net gain of £6. The same applies if Lightening or Speedy win. The *bookies* are vulnerable to a Dutch Book.

2.4. If the degrees of belief we assign do not conform to the axioms, and thus the theorems, of probability theory, then we are also vulnerable to Dutch Books. To illustrate, consider the theorem we proved in Lecture V: for any E, $Pr(E) + Pr(\overline{E}) = 1$. Suppose that my degree of belief in E is $3/4$ *and* my degree of belief in $\overline{E}$ is $3/4$. Clearly, my degrees of belief do not conform to the axioms and theorems of probability theory. What's wrong with this? The issue is that it leaves me open to a Dutch Book. If my degree of belief in E is $3/4$, then the worst odds I would accept on E is $1:3$; likewise for $\overline{E}$:

If $Pr(E) = 3/4$, then the worst odds on E acceptable to me is $1:3$
If $Pr(\overline{E}) = 3/4$, then the worst odds on $\overline{E}$ acceptable to me is $1:3$

So, the bookie offers a £3 bet on $\overline{E}$ and a £3 bet on E. From the above, this is acceptable to me. So, I bet £3 on $\overline{E}$ at $1:3$ and £3 on E at $1:3$. But this means I am bound to lose £2, regardless of what happens.

- If $\overline{E}$, then by the first bet, I receive back £4. However, if $\overline{E}$, then I lose £3 on the first bet. Net loss: £2.
- If E, then by the first bet, I receive back £4. However, if E, then I lose £3 on the second bet. Loss: £2.

So, because my degrees of belief do not conform to $Pr(E) + Pr(\overline{E}) = 1$, an intelligent bookies could always make money out of me. This economic irrationality is a consequence of my degrees of belief not conforming to the axioms of probability theory. Therefore, for my degrees of belief to be rational they must conform to the axioms—and thus theorems—of probability theory.

3. Expected Utility

3.1. In Lecture V, I noted that we often reasoned probabilistically and that is crucial therefore to sharpen up our understanding of probability. One kind of argument we might catch ourselves using is the following.

> If I work really hard and do well, then I could get an extremely well-paying job. However, if I work a decent amount, I'm almost guaranteed a decent job. I'd much prefer an extremely well-paying job; but I would be happy with just a good paying job. Unfortunately, the chances that I get an extremely well-paying job, even if I work really hard, are low. So, I shouldn't work very hard.

3.2. The seemingly reasonable thought driving this argument is that when deciding between two options, we should not just consider how great or not great it would be, if the we took one or the other option, but we should also consider how likely each of the options is to succeed. In other words, we should calculate what is called the *expected utility*—how much we would value some outcome multiplied by the likelihood that the outcome occurs. Expected Utility Theory is the view that, in choosing between two or more options, we should pick the one with the greatest expected utility.

3.3. Pascal famously applied this kind of reasoning for the conclusion that you should believe in God. Let's suppose that there's a small, but non-zero chance that God exists, i.e., 0.01. You either believe in God or you do not believe in God. If you don't believe in God, and God doesn't exist, you'll just be quite pleased with yourself. On the other hand, if you believe in God, and God doesn't exist, then you'll be quite annoyed with yourself. But crucially, if you don't believe in God, and God exists, all you are destined for is *eternal damnation*. And if you believe in God, and God does exist, then you'll be guaranteed eternal happiness:

	God exists (0.01)	God doesn't exist (0.99)
You believe	$+100000$ (Eternal Happiness)	-10 (Slightly Annoyed)
You don't believe	-1000000 (Eternal Damnation)	$+10$ (Slightly Please)

The expected utility of believing in God is greater than the expected utility of not believing in God: 10000 -9.9

$$\Big((0.01 \times 100000) + (0.99 \times -10)\Big) > \Big((0.01 \times -1000000) + (0.99 \times +10)\Big) = 9990.1 > -9990.1$$

3.4. This kind of reasoning can very quickly go awry. For instance, consider the case of *Pascal's Mugger*. Suppose someone approaches you in the street and says 'Either give me all the money in your wallet, or I'll denote a nuclear bomb over Manhattan!'. Now, of course, you think it very unlikely that they'll detonate a nuclear bomb if you refuse to give any money. It's unlikely, but it is not impossible. On the other hand, the consequences of you not giving them the money and they *do* detonate a nuclear bomb are catastrophic. Therefore, expected utility theory tells you that you should in fact hand over all of your money!

3.5. This might seem like a ridiculous scenario. However, it is not too far from the kind of argument given by so-called *Longtermists*—a kind of utilitarian who thinks that there is a much great moral imperative to

invest *now* in preventing *future* problems; or, more generally, that we should prioritize having a positive effect on humanity in the *long-term*. For instance, some people think that there's a small non-zero chance that humanity will be wiped completely out by a super-intelligent AI. However, they urge, if we invest *now*, we can prevent this from happening. We can represent this in the following table:

	Super-intelligent AI created (0.01)	Super-intelligent AI not created (0.99)
We invest	-100 (Investing costs)	-110 (Wasted invested costs annoys)
We don't invest	-1000000 (Humanity wiped out)	$+1000$ (We invest elsewhere)

$$\Big((0.01 \times -100) + (0.99 \times -110)\Big) > \Big((-10000000 \times 0.01) + (1000 \times 0.99)\Big) = -109.9 > -100990$$

So, we should invest to prevent, even the improbable, future AI catastrophe.

References

Ramsey, Frank P. (1990). Truth and Probability. In: *Philosophical Papers*. Ed. by Hugh Mellor. Cambridge University Press.

Appendix: Probability Definitions and Theorems

DEFINITION 1. (Outcome Space) *The outcome space V is the set of all possible outcomes.*

DEFINITION 2. (Field of V) *For any outcome space V, the field F_V is the set of all events F_V, i.e., $F_V = \mathcal{P}(V)$.*

- Any F_V is closed under intersection, union, and complement, i.e., for any F_V:
  - If $X, Y \in F_V$, then $X \cap Y \in F_V$ (Intersection)
  - If $X, Y \in F_V$, then $X \cup Y \in F_V$ (Union)
  - If $X \in F_V$, then $(V - X) \in F_V$ (Complement)
- For any event $X \in F_V$, the complement of X is $(V - X)$. Call this $\overline{X}$.

DEFINITION 3. (Kolmogorov Axioms for Probability Functions) *A probability function Pr on a field F_V is any function satisfying the following, for any $X, Y \in F_V$:*

(Axiom 1) $Pr(V) = 1$

(Axiom 2) $Pr(X) \geq 0$

(Axiom 3) *If $X \cap Y = \varnothing$, then $Pr(X \cup Y) = Pr(X) + Pr(Y)$*

DEFINITION 4. (Conditional Probability) *The probability of X given Y, $Pr(X|Y) = \dfrac{Pr(X \cap Y)}{Pr(Y)}$*

DEFINITION 5. (Independence) *For any two events $X, Y \in F_V$, X and Y are independent iff $Pr(X|Y) = Pr(X)$. Any two events $X, Y \in F_V$ are dependent iff they are not independent.*

THEOREM 1. (Probability of Intersection) *For any two events $X, Y \in F_V$, if X and Y are independent, then $Pr(X \cap Y) = Pr(X) \times Pr(Y)$; if X and Y are dependent, $Pr(X \cap Y) = Pr(X|Y) \times Pr(Y)$.*

THEOREM 2. (Bayes' Theorem First Version) *For any two events $X, Y \in F_V$: $Pr(X|Y) = \dfrac{Pr(Y|X) \times Pr(X)}{Pr(Y)}$*

THEOREM 3. (Bayes' Theorem Second Version) *For any two events $X, Y \in F_V$:*

$$Pr(X|Y) = \frac{Pr(Y|X) \times Pr(X)}{(Pr(Y|X) \times Pr(X)) + (Pr(Y|\overline{X}) \times Pr(\overline{X}))}$$