12 July 2005

PROBABILITY LOGIC: Part 2

Axioms for Probability Logic

Based on general considerations, we derived axioms for:

$P(b\,|\,a) =$ measure of the plausibility of proposition b conditional on the information stated in proposition a.

For propositions a, b and c, these are:

P1: $P(b\,|\,a) \geq 0$

P2: $P(b\,|\,a \,\&\, b) = 1$

P3: $P(b\,|\,a) + P(\sim b\,|\,a) = 1$

P4: $P(c \,\&\, b\,|\,a) = P(c\,|\,b \,\&\, a)\; P(b\,|\,a)$

Properties Derived from Axioms P1-P4

There are three properties that we used when deriving the axioms for probability logic that can be derived from Axioms P1-P4 and so they were not included in the list of axioms. We state and prove them in P5:

P5: (a) $P(\sim b\,|\,a \,\&\, b) = 0,$ (b) $P(b\,|\,a) \in [0,1]$, (c) $P(b\,|\,a \,\&\, (b \Rightarrow c)) \leq P(c\,|\,a \,\&\, (b \Rightarrow c))$

Proof:

(a) Follows from P3 with $a \rightarrow (a \,\&\, b)$ & then using P2

(b) Follows from P3 & P1: $P(b\,|\,a) = 1 - P(\sim b\,|\,a) \leq 1$

(c) Let $d = a \,\&\, (b \Rightarrow c)$, then:

$$\begin{aligned} P(b\,|\,d \,\&\, c)\; P(c\,|\,d) &= P(b \,\&\, c\,|\,d) && \text{(from P4)} \\ &= P(b \,\&\, (b \Rightarrow c)\,|\,d) \\ &= P(b\,|\,d \,\&\, (b \Rightarrow c)) P(b \Rightarrow c\,|\,d) && \text{(from P4)} \\ &= P(b\,|\,d) && \text{(from P2)} \end{aligned}$$

where we have used $b \,\&\, c \equiv b \,\&\, (b \Rightarrow c)$. Finally, by P5(b): $P(b\,|\,d \,\&\, c) \leq 1$ and so: $P(b\,|\,d) \leq P(c\,|\,d)$.

P6: (a) $P(c \text{ or } b \,|\, a) = P(c \,|\, a) + P(b \,|\, a) - P(c \,\&\, b \,|\, a)$

(b) $P(c \text{ or } b \,|\, a) = P(c \,|\, a) + P(b \,|\, a)$
if a implies that $(c \,\&\, b)$ is false, i.e. b and c cannot both be true and so are mutually exclusive.

(c) If proposition a implies that only one of the propositions $b_1, b_2, \ldots, b_N$ can be true, i.e. they are mutually exclusive, then:

$$P(b_1 \text{ or } b_2 \text{ or } \ldots \text{ or } b_N \,|\, a) = \sum_{n=1}^{N} P(b_n \,|\, a)$$

If, in addition, proposition a implies that one must be true: $1 = \sum_{n=1}^{N} P(b_n \,|\, a)$

Proof:

(a) From De Morgan's Law: $c \text{ or } b \equiv \sim(\sim c \,\&\, \sim b)$,

so: $P(c \text{ or } b \,|\, a) = P(\sim(\sim c \,\&\, \sim b) \,|\, a)$

$= 1 - P(\sim c \,\&\, \sim b \,|\, a)$ (from P3)

$= 1 - P(\sim c \,|\, \sim b \,\&\, a) P(\sim b \,|\, a)$ (from P4)

$= 1 - [1 - P(c \,|\, \sim b \,\&\, a)]\, P(\sim b \,|\, a)$ (from P3)

$= 1 - P(\sim b \,|\, a) + P(c \,\&\, \sim b \,|\, a)$ (from P4)

$= P(b \,|\, a) + P(\sim b \,|\, c \,\&\, a)\, P(c \,|\, a)$ (from P3, P4)

$= P(b \,|\, a) + [1 - P(b \,|\, c \,\&\, a)] P(c \,|\, a)$ (from P3)

$= P(b \,|\, a) + P(c \,|\, a) - P(b \,\&\, c \,|\, a)$ (from P4)

(b) Follows immediately since $P(c \,\&\, b \,|\, a) = 0$

(c) Exercise: Use Principle of Mathematical Induction.

P7: If proposition a implies that one, and only one, of the propositions $b_1, \ldots, b_N$ is true, then:

(a) $P(c \,|\, a) = \sum_{n=1}^{N} P(c \,\&\, b_n \,|\, a)$ [Marginalization Theorem]

(b) $P(c \,|\, a) = \sum_{n=1}^{N} P(c \,|\, b_n \,\&\, a)\, P(b_n \,|\, a)$ [Total Probability Theorem]

(c) For k=1, …, N:

$$P(b_k \,|\, c \,\&\, a) = \frac{P(c \,|\, b_k \,\&\, a)\, P(b_k \,|\, a)}{\sum_{n=1}^{N} P(c \,|\, b_n \,\&\, a)\, P(b_n \,|\, a)}$$ [Bayes Theorem]

Proof:

(a) Let $b \triangleq b_1 \text{ or } b_2 \text{ or } \ldots \text{ or } b_N$, then since a implies b is true, $a = a \& b$, so from P2:

$$P(b \mid a) = 1 \text{ and } P(c \& b \mid a) = P(c \mid b \& a)\, P(b \mid a) = P(c \mid a) \text{, so}$$
$$P(c \mid a) = P(c \& (b_1 \text{ or } \ldots \text{ or } b_N) \mid a) = P((c \& b_1) \text{ or } \ldots \text{ or } (c \& b_N) \mid a)$$

Since a implies that b_n and b_m cannot both be true if $m \neq n$, it must be that $(c \& b_n) \& (c \& b_m) = c \& (b_n \& b_m)$ is false, i.e. $(c \& b_n)$ and $(c \& b_m)$ are mutually exclusive. From $P6\,(c)$ with $b_n \to c \& b_n : P(c \mid a) = \sum_{n=1}^{N} P(c \& b_n \mid a)$

(b) Directly from (a) using P4.

(c) From (b), the denominator on RHS is $P(c \mid a)$, so we need only show that $P(b_k \mid c \& a)\, P(c \mid a) = P(c \mid b_k \& a)\, P(b_k \mid a)$. This follows from P4 since both are equal to $P(b_k \& c \mid a) = P(c \& b_k \mid a)$.

Derivation of Kolmogorov’s Axioms for Probability Measure of a Set

Recall that the axioms for probability logic are expressed in terms of propositions a, b, and c as:

P1: $P(b\,|\,a) \geq 0$

P2: $P(b\,|\,a \,\&\, b) = 1$

P3: $P(b\,|\,a) + P(\sim b\,|\,a) = 1$

P4: $P(c \,\&\, b\,|\,a) = P(c\,|\,b \,\&\, a)\; P(b\,|\,a)$

These axioms imply the property:

P6: $P(c \text{ or } b\,|\,a) = P(c\,|\,a) + P(b\,|\,a) - P(c \,\&\, b\,|\,a)$

Consider a real-valued quantity whose value x is uncertain but we know that $x \in X$, the *set of possible values* of the quantity, and assume that X is finite. Let π denote the proposition that specifies the *probability model* for the quantity, so π states that $x \in X$ and gives the probability (degree of plausibility) of the quantity having the value x for each $x \in X$.

Let $A \subset X$, then from axiom P1:

K1: $P(x \in A\,|\,\pi) \geq 0$

From axiom P2:

K2: $P(x \in X\,|\,\pi) = 1$ (since π states $x \in X$)

Also, $P(x \in A \cup B\,|\,\pi) = P(x \in A \text{ or } x \in B\,|\,\pi)$

$$= P(x \in A\,|\,\pi) + P(x \in B\,|\,\pi) - P(x \in A \,\&\, x \in B\,|\,\pi)$$

$$= P(x \in A\,|\,\pi) + P(x \in B\,|\,\pi) - P(x \in A \cap B\,|\,\pi)$$

If $A, B \subset X$ are disjoint, i.e. $A \cap B = \phi$ (nullset), then $x \in A$ and $x \in B$ are mutually exclusive, so:

K3: $P(x \in A \cup B\,|\,\pi) = P(x \in A\,|\,\pi) + P(x \in B\,|\,\pi)$

Introduce the shortened notation $\widetilde{P}(A)$ for $P(x \in A\,|\,\pi)$ where $A \subset X$, i.e. leave conditioning on π as implicit, then we can rewrite the above as:

K1’: $\widetilde{P}(A) \geq 0, \quad \forall A \subset X$

K2’: $\widetilde{P}(X) = 1$

K3’: $\widetilde{P}(A \cup B) = \widetilde{P}(A) + \widetilde{P}(B), \; \forall A, B \subset X$ with $A \cap B = \phi$

These are like A. Kolmogorov’s axioms in “Foundations of the Theory of Probability,” 1950. In fact, we could define a real-valued function $\widetilde{P}$, called a *probability measure*, on all the subsets of X by: $\tilde{P}(A) = P(x \in A \,|\, \pi), \forall A \subset X$. We would then have exactly Kolmogorov's axioms K1'-K3' in terms of his probability measure.

In Kolmogorov's presentation, the set function $\tilde{P}(A)$ is not given an operational interpretation. Probability logic gives it an interpretation in terms of the degree of plausibility that the uncertain quantity has a value $x \in A$, conditional on the probability model specified by the proposition π .

We have not yet applied axiom P4, but:

$$\begin{aligned} P(x \in A \cap B \,|\, \pi) &= P(x \in A \,\&\, x \in B \,|\, \pi) \\ &= P(x \in A \,|\, x \in B \,\&\, \pi)\; P(x \in B \,|\, \pi) \end{aligned}$$

Introduce the abbreviated notation again, then:

$$\tilde{P}(A \cap B) = \tilde{P}(A \,|\, B)\tilde{P}(B)$$

Note that in any probability $P(b \,|\, a)$, proposition a cannot be self-contradictory, so we cannot have:

$$\tilde{P}(B) = P(x \in B \,|\, \pi) = 0$$

because this means that $\pi \Rightarrow \sim (x \in B)$ and so the proposition $(x \in B \,\&\, \pi)$ appearing as conditioning information would be self-contradictory.

Thus, $\tilde{P}(B) > 0$ and so:

$$\tilde{P}(A \,|\, B) = \frac{\tilde{P}(A \cap B)}{\tilde{P}(B)}$$

In Kolmogorov's approach, this appears as a definition of conditional probability, but in the probability logic approach, all probabilities are conditional from the outset and so the corresponding result appears as axiom P4.

Note: In Kolmogorov's approach, the uncertain-valued variable x is called a random variable, the set of X of possible values of x is called the sample space and the subsets of X are called events. We have little use for this terminology because probability logic has a much wider scope – it's domain is propositions. Also, we try to avoid the vague words deterministic and random in analyzing systems and natural phenomena, usually preferring instead complete information and incomplete (or partial) information, respectively. When there is missing information, we choose a probability model to give the probability of each possibility in a set of conceived possibilities.

Probability Models for Discrete Variables

Defn If the set X of possible values of a variable x is finite, then the variable is discrete (Terminology is also used if X is countably infinite, but we here we focus on finite X).

Convention: Introduce the shortened notation: $P(A|\pi) = P(x \in A|\pi), \forall\, A \subset X$, i.e. think of A as also representing the proposition "$x \in A$". Also, write:
$P(x|\pi)$ for $P(\{x\}|\pi), \quad \forall\, x \in X$, i.e. when subset $A = \{x\}$.

Here, π is a proposition specifying the probability model for the quantity whose uncertain value x is a discrete variable. Thus, π specifies the set X of possible values of x and some function $f: X \to [0,1]$ such that $P(x|\pi) = f(x), \quad \forall x \in X$.

Let $X \triangleq \{x_1, \ldots, x_N\} = \bigcup_{n=1}^{N} \{x_n\}$ and let $A \subset X$, then by repeated use of K3:

$$\begin{aligned} P(A|\pi) &= P(\bigcup_{x_n \in A} \{x_n\} | \pi) \\ &= \sum_{x_n \in A} P(x_n|\pi) = \sum_{x_n \in A} f(x_n) \end{aligned}$$

because singleton sets $\{x_n\}$ and $\{x_m\}$, $n \neq m$, are disjoint (i.e. propositions "$x = x_n$" and "$x = x_m$" are mutually exclusive). In particular, using K2:

$$1 = P(X|\pi) = \sum_{x_n \in X} P(x_n|\pi) = \sum_{n=1}^{N} f(x_n),$$

so this normalization property must be satisfied by the probability model .

Discrete-Variable Example to Illustrate Concepts

Suppose a gambler offers you a wager that in N rolls of a die, a 6 appears more than some specified n_o times. Before accepting the wager, you want to compute the probability that you will win, so you make a stochastic predictive analysis.

Let x denote the number of 6's in N rolls, then you want to compute:

$$P(\text{You win} \mid \pi) = P(x \le n_o \mid \pi)$$

where π specifies on appropriate probability model. You reason that on a roll, each face is equally plausible to appear (i.e. you see no reason to believe any number on the die to be more likely to show up). Also, you feel that knowing what number appeared on a roll has no influence on what will appear on the next roll, i.e. the information from one roll is irrelevant to predicting what will happen on another roll. Therefore, you choose the binomial model, so π represents the proposition that the set of possible values for x is $X = \{0,1,\ldots,N\}$ and that:

$$P(x \mid \pi) = \binom{N}{x}\left(\tfrac{1}{6}\right)^x \left(\tfrac{5}{6}\right)^{N-x}, \forall x \in X$$

where $\binom{N}{x} = \dfrac{N!}{(N-x)!x!}$ (a combinatorial coefficient).

You win if $x \in A \triangleq \{0,1,\ldots n_o\} \subset X$ and so $P(\text{You win}|\pi)=\text{P}(\text{A}|\pi)=\sum_{\text{x}\in\text{A}} \text{P}(\text{x}|\pi)$.

After computing this probability, you are told that the gambler is dishonest and may have an altered die. He showed you only one face and it was a 6, so you now wonder whether one or more of the other five faces have a 6. You decide to revise your predictive analysis as follows. [To make this scenario plausible, assume you only see the top face on each roll]

Let π_k specify the binomial probability model for the case where exactly *k* faces on the die have a 6, then for *k* = 1,.., 6: $P(x \mid \pi_k) = \binom{N}{x}\ \theta_k{}^x (1-\theta_k)^{N-x}, \forall x \in X$

where $\theta_k = k/6$ is the probability of rolling a six. [Note that for π_6, all faces are 6's, so x=N with certainty, i.e. $P(x \mid \pi_6) = \delta_{xN}$, the Kronecker delta]. Since you are unsure which proposition $\pi_k,\ k = 1,\ldots,6,$ is the appropriate one to assume true, you choose a probability model for these propositions:

$$P(\pi_k \mid M) = g(\theta_k), k = 1,\ldots,6$$

where M is a proposition specifying each π_k and your choice of probability model, $g(\theta_k)$. Now you calculate:

$$P(\text{You win} \mid M) = P(A \mid M) = \sum_{x\in A} P(x \mid M)$$

and use the Total Probability Theorem P7(b) to get:

$$P\left(x \mid M\right)=\sum_{k=1}^{6} P\left(x \mid \pi_k \,\&\, M\right) P\left(\pi_k \mid M\right)$$

because M implies that one, and only one, of $\pi_1,...,\pi_6$ is true. Note that the first factor is the prediction of the k^{th} model and the second factor is the probability of the k^{th} model. Also, $P\left(x \mid \pi_k \,\&\, M\right)=P\left(x \mid \pi_k\right)$ because π_k states the probability of each $x \in X$ and hence M is irrelevant, i.e. $P\left(x \mid \pi_k\right)$ is independent of M. We say that $P\left(x \mid M\right)$is a robust predictive probability because it uses a set of probability models for *x* as specified by M, in contrast to $P\left(x \mid \pi\right)$.

Notes:

1) When we write $P\left(c \mid b \,\&\, a\right)=P\left(c \mid a\right)$, we are stating that, given a, the information stated by b is irrelevant to the probability of c, i.e. probabilistic independence is about information independence and should not be confused with causal independence when the propositions refer to the occurrence of actual events.

2) There are many possible choices for M. You may feel that each π_k is equally plausible, so: $P\left(\pi_k \mid M\right)=\frac{1}{6}, \forall k=1,...,6$. But you may feel that the gambler is unlikely to have a die with 6 on every face because you would get suspicious if every roll gave a 6, so you choose $P\left(\pi_k \mid M\right)$ to be a decreasing function of k, subject to the constraint:

$$\sum_{k=1}^{6} P\left(\pi_k \mid M\right)=1$$

Your calculated probability of winning, and hence your decision whether to wager with the gambler, is conditional on your choice of M and this is inescapable. You cannot get certainty in the choice of π_k in the absence of information about how many faces of the die have a 6.

You could take different choices for M, say $M_j, j=1,...J,$ and let M specify a probability model for them, i.e. $P(M_j \mid M)=h(j), j=1,...,J,$ then:

$$P\left(x \mid M\right)=\sum_{j=1}^{J} P\left(x \mid M_j\right) P\left(M_j \mid M\right) \quad \text{[Total Probability Theorem]}$$

3) π_k specifies a probability model for *x*:

$P\left(x \mid \pi_k\right)= f_k(x), \ \forall x \in X$

and *M* specifies a probability model for the π_k:

$P\left(\pi_k \mid M\right)=g\left(\theta_k\right), \ k=1,...6$.

However, introducing the notation of $f_k\left(x\right)$ and $g\left(\theta_k\right)$ is really unnecessary because we can use $P\left(x\,|\,\pi_k\right)$ and $P\left(\pi_k\,|\,M\right)$ instead in our analysis.

Using Data to Update the Probability of Probability Models

We use the gambling example to examine how we can use data D to learn more about the probability models.

Suppose the gambler says after 2 rolls giving a 3 and a 6 that if you want, you can increase your bet. Therefore, you wish to calculate how information from data $D=\{x=1,N=2\}$ changes your probability of winning, based on your choice of the class of probability models for *x* specified by *M*, as before.

With $A'=\{0,1,...,n_o-1\}$: $P\left(\text{You win}\,|\,D,M\right)=P\left(A'\,|\,D,M\right)=\sum_{x'\in A'}P\left(x'\,|\,D,M\right)$

where the shorthand *D, M* means *D & M* (a common convention) and where x' denotes the number of 6's in (*N*-2) rolls of the die.

From the Total Probability Theorem:

$$P\left(x'\,|\,D,M\right)=\sum_{k=1}^{6}P\left(x'\,|\,\pi_k\right)P\left(\pi_k\,|\,D,M\right)$$

because $P\left(x'\,|\,\pi_k,D,M\right)=P\left(x'\,|\,\pi_k\right)=\binom{N-2}{x'}\theta_k^{x'}\left(1-\theta_k\right)^{N-2-x'}$

(i.e. *D* and *M* are irrelevant because π_k specifies the probability of x'). This has the same form as before except that the original probability $P\left(\pi_k\,|\,M\right)$ is updated to $P\left(\pi_k\,|\,D,M\right)$ because of the new information specified by *D.*

Applying Bayes Theorem P7(c) (with $b_k=\pi_k, c=D$ and $a=M$):

$$P\left(\pi_k\,|\,D,M\right)=c\;P\left(D\,|\,\pi_k,M\right)P\left(\pi_k\,|\,M\right),\forall k=1,...6$$

where $c^{-1}=\sum_{k=1}^{6}P\left(D\,|\,\pi_k,M\right)P\left(\pi_k\,|\,M\right)$ (it ensures that $\sum_{k-1}^{6}P\left(\pi_k\,|\,D,M\right)=1$) and:

$$P\left(D\,|\,\pi_k,M\right)=\binom{2}{1}\theta_k\left(1-\theta_k\right)$$

$$P\left(\pi_k\,|\,M\right)=g\left(\theta_k\right),\forall k=1,...,6$$

Thus, Bayes Theorem gives:

$$P\left(\pi_k\,|\,D,M\right)=2c\theta_k\left(1-\theta_k\right)g\left(\theta_k\right)$$

[Posterior to data] [Influence of data] [Prior to data]

[Note that $P\left(\pi_6\,|\,D,M\right)=0$, as expected, since you now know that not all faces of the die have a 6].

Substituting:

$$P\left(x'\,|\,D,M\right)=\frac{\sum_{k=1}^{5}\binom{N-2}{x'}\theta_k^{x'+1}\left(1-\theta_k\right)^{N-1-x'}g\left(\theta_k\right)}{\sum_{k=1}^{5}\theta_k\left(1-\theta_k\right)g\left(\theta_k\right)}$$

We say that $P\left(x'\,|\,D,M\right)$ is the posterior robust predictive probability whereas $P\left(x\,|\,M\right)$ given earlier is the prior robust predictive probability, meaning after and before, respectively, the information in the data is utilized.

Note on Notation for Specification of Probability Models:

There is another notation that we can use for specifying the probability models that will prove to be useful later.

Introduce parameter θ as the probability that a roll of the die gives a 6, then instead of the notation π_k to specify the probability model for exactly *k* faces of the die having a 6, we could use $\pi(\theta_k)$ with $\theta_k \triangleq \frac{k}{6}$ and so:

$P\left(x \mid \pi(\theta)\right) = \binom{N}{x}\theta^x(1-\theta)^{N-x}$ and $P\left(\pi(\theta) \mid M\right) = g(\theta)$, $\forall\theta = \theta_1, ..., \theta_6$.

Thus, $\pi(\theta)$ is a variable proposition with $\theta \in \{\theta_1, ..., \theta_6\}$. But the notation $\pi(\theta)$ seems unnecessarily cumbersome, so we write instead:

$$P(x \mid \theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x}, \forall x \in X$$

$$P(\theta \mid M) = g(\theta), \ \forall\theta \in \Theta$$

where $\Theta = \{\theta_k : k = 1, ..., 6\}$.

Superficially, it now looks as if we are specifying the probability of the parameter θ, and this is how most people using Bayesian analysis think of it, but strictly speaking θ now represents the variable proposition that specifies the probability model for *x*, i.e. θ has a dual role as representing this proposition and a parameter, both of which specify the probability model in a set of probability models given by the proposition M, i.e.

$\theta =$ "Probability model for $x \in X$ is $\binom{N}{x}\theta^x(1-\theta)^{N-x}$."

$M =$ "Probability model for the set of probability models
$\{\binom{N}{x}\theta^x(1-\theta)^{N-x} : \theta \in \Theta\}$ is $g(\theta)$."

Recall that *x* in $P(x \mid \theta)$ also has a dual role as the variable $x \in X$ and as the variable proposition: $x =$ "number of 6's in N rolls of the die is x."

Why do we use these dual roles? Because it brings consistency between the notation that almost everybody uses in probability theory and the probability logic interpretation that we have of its meaning. However, if we were starting from scratch, we may have come up with the same notation anyway, because of its efficiency.

To summarize, in general we use *M* to represent the proposition specifying the set of probability models for $x \in X$, denoted $\{P(x \mid \theta) : \theta \in \Theta\}$, and the probability model for this set, denoted by $P(\theta \mid M)$. We say that *M* specifies the <u>class</u> of probability models for x.