The $A_p$ Distribution and the Rule of Succession

Bayesian Memory?
Bayesian Memory

• What is your prior belief that a fair coin will come up heads on the first toss? $\rightarrow 0.5$

• What is your prior belief that it will be raining tomorrow? $\rightarrow 0.5$

• What happens if you get new information?
  – Coin: you tossed the coin five times and it comes up head every time $\rightarrow$ I would say it is still $0.5$
  – Rain: it is raining today $\rightarrow$ I would say it it will increase to say $0.7$
Bayesian Memory

• My state of belief is very stable in the case of predicting the outcome of tossing a coin, but very unstable in the case of predicting whether it will be raining tomorrow.
• So a single number is not enough to encode my prior belief! Perhaps, we need two numbers: one representing plausibility, and the other how stable it is in the face of new evidence.
Bayesian Memory

\[ \text{prob}(A|A_p, E) \equiv p \]

\(A_p \equiv \) regardless of anything else you may have been told, the probability of A is \(p\)

If \(X\) says nothing about \(A\) except that it is possible for \(A\) to be true or false, then

\[ \text{prob}(A_p|X) = 1, \text{ for } 0 \leq p \leq 1 \]
Bayesian Memory

\[ \text{prob}(A_p|E, X) = \text{prob}(A_p|X) \frac{\text{prob}(E|A_p, X)}{\text{prob}(E|X)} \]

\[ = \frac{\text{prob}(E|A_p)}{\text{prob}(E|X)} \]
Bayesian Memory

$$\text{prob}(A|E) = \int_{0}^{1} dp \ \text{prob}(A, A_p|E)$$

$$= \int_{0}^{1} dp \ \text{prob}(A|A_p, E) \times \text{prob}(A_p|E)$$

$$= \int_{0}^{1} dpp \times \text{prob}(A_p|E)$$
Suppose we have have some evidence E that consists of two parts, \( E = E_aE_b \), where \( E_a \) is relevant to \( A \) and, given \( E_a \), \( E_b \) is not relevant:

\[
\text{prob}(A|E) = \text{prob}(A|E_a, E_b) = \text{prob}(A|E_a).
\]  

(1)

By Bayes’ theorem, it follows that, given \( E_a \), \( A \) must also be irrelevant to \( E_b \):

\[
\text{prob}(E_b|A, E_a) = \frac{\text{prob}(E_b|E_a) \text{prob}(A|E_a, E_b)}{\text{prob}(A|E_a)} = \text{prob}(E_b|E_a)
\]

(2)

Let’s call this property 'weak irrelevance'. Does this imply that \( E_b \) is also irrelevant for \( A_p \)? No, it does not, because

\[
\text{prob}(A|E) = \int_0^1 p \times \text{prob}(A_p|E)dp
\]

\[
= \int_0^1 p \times \text{prob}(A_p|E_a, E_b)dp
\]

\[
= \int_0^1 p \times \text{prob}(A_p|E_a)dp,
\]

(3)
i.e. only the first moment of \( \text{prob}(A_p|E_a, E_b) \) and \( \text{prob}(A_p|E_a) \) are the same. But suppose that

\[
\text{prob}(A_p|E_a, E_b) = \text{prob}(A_p|E_a);
\]

(4)
call this 'strong irrelevance’ then from Bayes' theorem it follows that \( A_p \) is irrelevant to \( E_b \):

\[
\text{prob}(E_b|A_p, E_a) = \text{prob}(E_b|E_a).
\]

(5)

Now, suppose we gather a new piece of evidence, \( F \). How does this change our state of knowledge about \( A \)?

\[
\text{prob}(A|E, F) = \int_0^1 p \times \text{prob}(A_p|E, F)dp
\]

\[
= \int_0^1 p \times \text{prob}(A_p|E) \frac{\text{prob}(F|A_p, E)}{\text{prob}(F|E)}
\]

(6)

In the likelihood ratio, any part of \( E \) that is irrelevant to \( A_p \) can be struck out:

\[
\frac{\text{prob}(F|A_p, E)}{\text{prob}(F|E)} = \frac{\text{prob}(F|A_p, E_a, E_b)}{\text{prob}(F|E_a, E_b)}
\]

\[
= \frac{\text{prob}(F|A_p, E_a) \left[ \frac{\text{prob}(E_b|F, A_p, E_a)}{\text{prob}(E_b|E_a)} \right]}{\text{prob}(F|E_a)}
\]

\[
= \frac{\text{prob}(F|A_p, E_a)}{\text{prob}(F|E_a)}.
\]

(7)

We arrive at

\[
\text{prob}(A|EF) \sim \int_0^1 p \times \text{prob}(A_p|E) \times \text{prob}(F|A_p)dp
\]

(8)
The preposition E, which represents our prior information, now appears only in the density \( (A_p|E) \). That is, the only property of E which we need to reason out the effect of new information F is the density \( (A_p|E) \). We need to memorize only this density, not how we came to it!!!
Two $A_p$ distributions
Let’s imagine that a ‘random’ experiment is being performed. From the results of the experiment in the past, we want to predict the results in the future. We introduce the propositions:

\[ X \equiv \text{For each trial we admit two prior outcomes: A true, and A false} \]

The prior information:
- the probability assigned to A at the nth trial does not depend on n
- evidence concerning the results of past trials retains its relevance for all times

\[ N_n \equiv \text{A true n times in N trials in the past} \]
\[ M_n \equiv \text{A true n times in M trials in the future} \]

Thus,
\[ \text{prob}(A_p|X) = 1, \quad \text{for } 0 \leq p \leq 1 \] (1)

We are after \( \text{prob}(M_m|N_n) \). We start from the binomial distribution:

\[ \text{prob}(N_n|A_p) = \binom{N}{n} p^n (1-p)^{N-n} \]
\[ \text{prob}(M_m|A_p) = \binom{M}{m} p^m (1-p)^{M-m} \] (2)

What are the prior probabilities for \( N_n \) and \( M_m \) given the information we have?

\[ \text{prob}(N_n|X) = \int_0^1 \text{prob}(N_n, A_p|X) dp \]
\[ = \int_0^1 \text{prob}(N_n|A_p) \times \text{prob}(A_p|X) dp \]
\[ = \binom{N}{n} \int_0^1 p^n (1-p)^{N-n} dp \] (3)

The integral is the complete Beta-function

\[ \int_0^1 x^r (1-x)^s dx = \frac{r! s!}{(r+s+1)!} \] (4)

Thus, we have

\[ \text{prob}(N_n|X) = \begin{cases} \frac{1}{N+1} & 0 \leq n \leq N \\ 0 & \text{otherwise} \end{cases} \] (5)

(analogous for \( \text{prob}(M_n|X) = \frac{1}{M+1} \))

Now, turn around (2) by Bayes’ theorem:

\[ \text{prob}(A_p|N_n) = \frac{\text{prob}(A_p|X) \text{prob}(N_n|A_p)}{\text{prob}(N_n|X)} \]
\[ = (N+1)\text{prob}(N_n|A_p) \] (6)

and so finally the desired probability is:

\[ \text{prob}(M_m|N_n) = \int_0^1 \text{prob}(M_m, A_p|N_n) dp \]
\[ = \int_0^1 \text{prob}(M_m|A_p, N_n) \times \text{prob}(A_p|N_n) dp. \] (7)

Since \( \text{prob}(M_m|A_p, N_n) = \text{prob}(M_m|A_p) \), substituting into (7) yields:

\[ \text{prob}(M_m|N_n) = \frac{(n+m)! (N+M-n-m)!}{n! (N-n)! (M+1)!} \] (8)

For the case \( M = m = 1 \), i.e. the probability for A being in the next trial, we get:

\[ \text{prob}(A|N_n) = \frac{n+1}{N+2} \] (9)

This is Laplace’s Rule of Succession!
Objections to the Rule of Succession

Consider the case $N = n = 0$. Without seeing any data, the probability of success is 0.5.

Thus, the probability that there are exactly 137 elephants on Mars is 0.5. Also, there is a probability of 0.5 that there are 138 elephants on Mars. Therefore it is certain that there are at least 137 elephants on Mars! But the rule also says that there is a probability of 0.5 that there are no elephants on Mars. The rule is logically self-contradictory.
Generalization

\[ \text{prob}(A_i|n_i, N, K) = \frac{n_i + 1}{N + K} \]