Probability Background

CSC 261 -Artificiall Intelligence - Professor Weinman

Summary: We briefly review the principle of indifference, hypergeometric distribution, and rule of succession for making predictions from observations with limited prior knowledge.

We used Simon Pierre Laplace's principle of indifference to assign equal prior probabilities to N indistinguishable, mutually exclusive and exhaustive propositions $\{A_k\}_{k=1}^{N}$when we had no additional information:

$$P(A_k \mid I) = \frac{1}{N}.$$

Laplace derived another general and useful principle called the "Rule of Succession," one of the earliest practical uses of combining Bayesian updates with the principle of indifference for probabilistic learning. Suppose we have seen M examples of a scenario we know has N possible outcomes (as above). If we start from the principle of indifference with $M = 0$, then we should have that the first observation is any of the $k = 1 \ldots N$ outcomes would be$\frac{1}{N}$ by the principle of indifference. Intuitively, as we observe more and more examples of outcome A_k, our belief that the next observation might be A_k should increase.

When $N = 2$, the hypergeometric distribution

$$P(T = t \mid W = w, Y = y, Z = z, I) = h(t; w, y, z)$$

tells us the probability of seeing exactly $T = t$ observations where proposition A_1 holds from a finite population of size $Z = z$ after $W = w$ observations when there are $Y = y$ such possibilities in the population. For example, the probability of seeing 5 cards of the spades suit after ten draws from a standard 52 card deck (in which there are 13 cards of each suit) would be $h(5; 10, 13, 52)$. The form is

$$h(t; w, y, z) = \frac{\binom{y}{t}\binom{z-y}{w-t}}{\binom{z}{w}}$$

so that our query is $h(5; 10, 13, 52) \approx 0.0468$. But what if we don't know enough about the population to say how many examples there are to which A_1 applies? If we know the population size and no other prior information, then the principle of indifference applies to the $z + 1$ possibilities, giving us

$$P(Y = y \mid Z = z, I) = \begin{cases} \frac{1}{z+1} & y \leq z \\ 0 & \text{otherwise.} \end{cases}$$

What if we want to predict E_{w+1}, the result of the $w + 1$th observation? What is our degree of belief? As it turns out, if we marginalize over the possibilities for Y the resulting predictive probability is

$$P(E_{w+1} \mid T = t, W = w, Z = z, I) = \frac{t+1}{w+2},$$

which Laplace showed 1799. Ponder this astounding result for a moment. Our prior information that both possibilities are realizable is akin to adding a "pseudo" observation for each possibility–the plus one in the numerator. Even when we have no observed matches for A_1 $(t = 0)$, we refuse to assign it a zero probability.

However, as we gather more and more actual observations, our belief approaches zero. In fact, our posterior belief is very nearly the relative frequency of the observed values. Our prior information (only "all cases are possible") is quickly dwarfed by the evidence of the observations. One other remarkable fact is worth observing: the result does not depend whatsoever on the size of the population, Z.

This same analysis generalizes to our starting scenario where we have $N \geq 2$ possibilities, each of which is possible on any observation. In that case, the probability of observing an instance of the kth category on the $w+1$th observation, E_{w+1}^k is

$$P\left(E_{w+1}^k \mid \mathbf{T} = \mathbf{t}, W = w, Z = z, I\right) = \frac{t_k + 1}{w + N},$$

where t_k is the number of examples of category k observed among the w; note that $\sum_{k=1}^{N} t_k = w$.

Although Laplace's rule of succession has been the subject of some controversy over the intervening centuries (mostly due to his poor choice of example application), it is a remarkably powerful demonstration of Bayesian reasoning when applied to a problem for which there truly is little prior knowledge.