Notes on the general background to Bayesian probability theory

Jaynes, E. T. (1986), *Bayesian Methods: General Background*, In Maximum Entropy and Bayesian Methods in Applied Statistics, ed. Justice, J. H., Proceedings of the Fourth Maximum Entropy Workshop, 1984, Cambridge University Press.

F. H. Evans
January 2002

Bernoulli (1713) – The Art of Conjecture

Suppose we have a set of equally possible cases (events or outcomes) $\{x_1, x_2..., x_n\}$. This defines the *hypothesis* space, H_0. Now suppose we have some proposition of interest, *A*, defined as being true on some specified subset *H(A)* of *M* points of H_0, false on the others. *M*, the number of ways in which *A* could be true, is called the multiplicity of *A*, and the probability of *A* is defined as the proportion $p(A) = M / N$. The rules of reasoning consist of finding the probabilities *p(A), p(B)* etc of different propositions by counting the number of ways they can be true. Moreover, if we learn that *A* is true, our hypothesis space contracts to *H(A)* and the probability of *B* is changed to the proportion of *H(A)* on which *B* is true.

James Bernoulli also discussed Nature's hypothesis space H_N of possibilities: in many situations, it is impossible to determine the true hypothesis space and assign probabilities. However, we may probe H_N by making repeated observations of some event *A*. Bernoulli proved the first mathematical connection between frequency and probability: **the weak law of large numbers**:

If we make *n* repeated observations and find *A* true *m* times, the observed frequency $f(A) = m / n$ is to be compared with the probability $p(A) = M / N$. In the limit of large *n*, it it becomes practically certain that *f(A)* is close to *p(A)*. Laplace showed later that as *n* tends to infinity, the probability remains more than 0.5 that *f(A)* is in the shrinking interval $p(A) \pm q$, where $q^2 = p(1-p)/n$.

Bayes (1763)

"An essay towards solving a Problem in the Doctrine of Chances" by Rev. Thomas Bayes was published posthumously in 1763.

Laplace

In 1774, Laplace published a work that re-discovered Bayes' principle in greater clarity and generality, as follows:

Denoting various propositions by *A*, *B*, *C* etc, let:

AB	$=_{def}$ "both A and B are true"	the logical product	AND
$A+B$	$=_{def}$ "either A or B is true"	the logical sum	OR
$\bar{A}$	$=_{def}$ "A is false"	the denial	NOT
$A \setminus B$	$=_{def}$ "the probability that A is true given B is true" or "A give B"		

Then the basic rules of probability are:

The Product Rule

$$P(AB \setminus C) = P(A \setminus BC)\,P(B \setminus C) \tag{1}$$

The Sum Rule

$$P(A \setminus B) + P(\bar{A} \setminus B) = 1 \tag{2}$$

From these we derive what is known as **Bayes' theorem** (although Bayes never wrote it):

$$p(A \setminus BC) = \frac{p(A \setminus C)\,p(B \setminus AC)}{p(B \setminus C)} \tag{3}$$

This become useful in the case where A represents some hypothesis whose truth we wish to determine, B represents some new data from some observation and C represents what we knew about A getting the data B. We call $p(A \setminus C)$ the prior probability of A, when we know only C. $p(A \setminus BC)$ is the posterior probability, updated as a result of knowing B.

Note:

1. This assumes that there is some prior information before we obtain the data B.
2. We can apply Bayes' theorem repeatedly as new pieces of information B_1, B_2,… are obtained.

Laplace also published (1812) a two-volume treatise on probability theory. The first volume contains, in his methods for solving finite difference equations, almost all of the mathematics we find today in the theory of digital filters. Yet, because some difficult concepts were ill-explained by Laplace, Bayesian analysis was no well-accepted for more than a century after his work was published.

Jeffreys

Sir Harold Jeffreys re-discovered the work of Laplace in the early 19th century, and explained it much more clearly in the 1930s. He later published "Theory of probability" (Oxford University Press, 1961)

Some notes on assigning priors

If our hypothesis space is large enough to accommodate repetitions, we can calculate the probability of an event by counting the frequency with which it occurs. But note that "a *probability* is an abstract concept, a quantity we assign theoretically, for the prupose of representing a state of knowledge, or that we calculate from previously assigned probabilities using the rules (1)-(3) of probability theory. A *frequency* is, in situations where it makes sense to

speak of repetitions, a factual property of the real world, that we measure or estimate. So instead of committing the error of saying that the probability is the frequency, we ought to calculate the probability p(f)df that the frequency lies in certain intervals df – just a Bernoulli did."

Cox

R. T. Cox (1946) published a paper that showed that any set of rules for inference, in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to the Laplace-Jeffreys rules, that is (1)-(3), or inconsistent.

Shannon

Claude Shannon (1948) used Cox's method. He sought a measure of the "amount of uncertainty" in a probability distribution. The conditions of consistency again took the form of functional equations whose general solution he found. The resulting measure proved to be $\sum p_i \log(p_i)$.

Shannon's work gives us the means to escape from Bernoulli and Laplace's assumption that events in the hypothesis space be equally possible, that is to construct non-uniform prior distributions. We can define a hypothesis space H_0 by enumerating some perceived possibilities $\{x_1, x_2..., x_n\}$; but we do not regard them as equally likely, because we have some additional evidence E. It is not useable as the data B in Bayes' theorem, because E is not an event and does not have a "sampling distribution" $p(E \setminus C)$. But E leads us to impose some constraint on the probabilities $p_i = p(x_i)$ that we assign to the elements of H_0, which forces them to be nonuniform, but does not fully determine them (the number of constraints is less than N).

We interpret Shannon's theorem as indicating that, out of all of the distributions p_i that agree with the constraints, the one that maximizes the Shannon entropy represents the "most honest" description of our state of knowledge, in the following sense: it expresses the enumeration of the possibilities, the evidence E; and assumes nothing beyond that.

References

Cox, R. T. (1946), Probability, frequency, and reasonable expectation, Am. J. Physics 14, 1-13.

Jeffreys, H. (1939), Theory of probability, Oxford University Press.

Molina, E. C. (1963), Two papers by Bayes with commentaries, Hafner Publishing Co., New York.

Shannon, C. E. (1948), A mathematical theory of communication. Bell System Tech. J. 27, 379-423, 623-656. Reprinted in C. E. Shannon and W. Weaver, The mathematical theory of communication, University of Illinois Press, Urbaba, 1949.