CAMBRIDGE

Cambridge University Press
978-0-521-59271-0 - Probability Theory: The Logic of Science
E. T. Jaynes
Excerpt
More information

Part 1

Principles and elementary applications

1
Plausible reasoning

> The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.
>
> *James Clerk Maxwell (1850)*

Suppose some dark night a policeman walks down a street, apparently deserted. Suddenly he hears a burglar alarm, looks across the street, and sees a jewelry store with a broken window. Then a gentleman wearing a mask comes crawling out through the broken window, carrying a bag which turns out to be full of expensive jewelry. The policeman doesn't hesitate at all in deciding that this gentleman is dishonest. But by what reasoning process does he arrive at this conclusion? Let us first take a leisurely look at the general nature of such problems.

1.1 Deductive and plausible reasoning

A moment's thought makes it clear that our policeman's conclusion was not a logical deduction from the evidence; for there may have been a perfectly innocent explanation for everything. It might be, for example, that this gentleman was the owner of the jewelry store and he was coming home from a masquerade party, and didn't have the key with him. However, just as he walked by his store, a passing truck threw a stone through the window, and he was only protecting his own property.

Now, while the policeman's reasoning process was not logical deduction, we will grant that it had a certain degree of validity. The evidence did not make the gentleman's dishonesty *certain*, but it did make it extremely *plausible*. This is an example of a kind of reasoning in which we have all become more or less proficient, necessarily, long before studying mathematical theories. We are hardly able to get through one waking hour without facing some situation (e.g. will it rain or won't it?) where we do not have enough information to permit deductive reasoning; but still we must decide immediately what to do.

In spite of its familiarity, the formation of plausible conclusions is a very subtle process. Although history records discussions of it extending over 24 centuries, probably nobody has ever produced an analysis of the process which anyone else finds completely satisfactory. In this work we will be able to report some useful and encouraging new progress, in which conflicting intuitive judgments are replaced by definite theorems, and *ad hoc* procedures are replaced by rules that are determined uniquely by some very elementary – and nearly inescapable – criteria of rationality.

All discussions of these questions start by giving examples of the contrast between deductive reasoning and plausible reasoning. As is generally credited to the *Organon* of Aristotle (fourth century BC)[1] deductive reasoning (*apodeixis*) can be analyzed ultimately into the repeated application of two strong syllogisms:

$$\begin{array}{c}\text{if } A \text{ is true, then } B \text{ is true}\\ A \text{ is true}\\ \hline \text{therefore, } B \text{ is true,}\end{array} \tag{1.1}$$

and its inverse:

$$\begin{array}{c}\text{if } A \text{ is true, then } B \text{ is true}\\ B \text{ is false}\\ \hline \text{therefore, } A \text{ is false.}\end{array} \tag{1.2}$$

This is the kind of reasoning we would like to use all the time; but, as noted, in almost all the situations confronting us we do not have the right kind of information to allow this kind of reasoning. We fall back on weaker syllogisms (*epagoge*):

$$\begin{array}{c}\text{if } A \text{ is true, then } B \text{ is true}\\ B \text{ is true}\\ \hline \text{therefore, } A \text{ becomes more plausible.}\end{array} \tag{1.3}$$

The evidence does not prove that A is true, but verification of one of its consequences does give us more confidence in A. For example, let

$$\begin{aligned} A &\equiv \text{it will start to rain by 10 AM at the latest;}\\ B &\equiv \text{the sky will become cloudy before 10 AM.}\end{aligned}$$

Observing clouds at 9:45 AM does not give us a logical certainty that the rain will follow; nevertheless our common sense, obeying the weak syllogism, may induce us to change our plans and behave *as if* we believed that it will, if those clouds are sufficiently dark.

This example shows also that the major premise, 'if A then B' expresses B only as a *logical* consequence of A; and not necessarily a causal physical consequence, which could be effective only at a later time. The rain at 10 AM is not the physical cause of the clouds at

[1] Today, several different views are held about the exact nature of Aristotle's contribution. Such issues are irrelevant to our present purpose, but the interested reader may find an extensive discussion of them in Lukasiewicz (1957).

9:45 AM. Nevertheless, the proper logical connection is not in the uncertain causal direction (clouds $\Longrightarrow$ rain), but rather (rain $\Longrightarrow$ clouds), which is certain, although noncausal.

We emphasize at the outset that we are concerned here with *logical* connections, because some discussions and applications of inference have fallen into serious error through failure to see the distinction between logical implication and physical causation. The distinction is analyzed in some depth by Simon and Rescher (1966), who note that all attempts to interpret implication as expressing physical causation founder on the lack of contraposition expressed by the second syllogism (1.2). That is, if we tried to interpret the major premise as 'A is the physical cause of B', then we would hardly be able to accept that 'not-B is the physical cause of not-A'. In Chapter 3 we shall see that attempts to interpret plausible inferences in terms of physical causation fare no better.

Another weak syllogism, still using the same major premise, is

$$\begin{array}{c}\text{If } A \text{ is true, then } B \text{ is true}\\ A \text{ is false}\\ \hline \text{therefore, } B \text{ becomes less plausible.}\end{array} \tag{1.4}$$

In this case, the evidence does not prove that B is false; but one of the possible reasons for its being true has been eliminated, and so we feel less confident about B. The reasoning of a scientist, by which he accepts or rejects his theories, consists almost entirely of syllogisms of the second and third kind.

Now, the reasoning of our policeman was not even of the above types. It is best described by a still weaker syllogism:

$$\begin{array}{c}\text{If } A \text{ is true, then } B \text{ becomes more plausible}\\ B \text{ is true}\\ \hline \text{therefore, } A \text{ becomes more plausible.}\end{array} \tag{1.5}$$

But in spite of the apparent weakness of this argument, when stated abstractly in terms of A and B, we recognize that the policeman's conclusion has a very strong convincing power. There is something which makes us believe that, in this particular case, his argument had almost the power of deductive reasoning.

These examples show that the brain, in doing plausible reasoning, not only decides whether something becomes more plausible or less plausible, but that it evaluates the *degree* of plausibility in some way. The plausibility for rain by 10 AM depends very much on the darkness of those clouds at 9:45. And the brain also makes use of old information as well as the specific new data of the problem; in deciding what to do we try to recall our past experience with clouds and rain, and what the weatherman predicted last night.

To illustrate that the policeman was also making use of the past experience of policemen in general, we have only to change that experience. Suppose that events like these happened several times every night to every policeman – and that in every case the gentleman turned out to be completely innocent. Very soon, policemen would learn to ignore such trivial things.

Thus, in our reasoning we depend very much on *prior information* to help us in evaluating the degree of plausibility in a new problem. This reasoning process goes on unconsciously, almost instantaneously, and we conceal how complicated it really is by calling it *common sense*.

The mathematician George Pólya (1945, 1954) wrote three books about plausible reasoning, pointing out a wealth of interesting examples and showing that there are definite rules by which we do plausible reasoning (although in his work they remain in qualitative form). The above weak syllogisms appear in his third volume. The reader is strongly urged to consult Pólya's exposition, which was the original source of many of the ideas underlying the present work. We show below how Pólya's principles may be made quantitative, with resulting useful applications.

Evidently, the deductive reasoning described above has the property that we can go through long chains of reasoning of the type (1.1) and (1.2) and the conclusions have just as much certainty as the premises. With the other kinds of reasoning, (1.3)–(1.5), the reliability of the conclusion changes as we go through several stages. But in their quantitative form we shall find that in many cases our conclusions can still approach the certainty of deductive reasoning (as the example of the policeman leads us to expect). Pólya showed that even a pure mathematician actually uses these weaker forms of reasoning most of the time. Of course, on publishing a new theorem, the mathematician will try very hard to invent an argument which uses only the first kind; but the reasoning process which led to the theorem in the first place almost always involves one of the weaker forms (based, for example, on following up conjectures suggested by analogies). The same idea is expressed in a remark of S. Banach (quoted by S. Ulam, 1957):

> Good mathematicians see analogies between theorems; great mathematicians see analogies between analogies.

As a first orientation, then, let us note some very suggestive analogies to another field – which is itself based, in the last analysis, on plausible reasoning.

1.2 Analogies with physical theories

In physics, we learn quickly that the world is too complicated for us to analyze it all at once. We can make progress only if we dissect it into little pieces and study them separately. Sometimes, we can invent a mathematical model which reproduces several features of one of these pieces, and whenever this happens we feel that progress has been made. These models are called *physical theories*. As knowledge advances, we are able to invent better and better models, which reproduce more and more features of the real world, more and more accurately. Nobody knows whether there is some natural end to this process, or whether it will go on indefinitely.

In trying to understand common sense, we shall take a similar course. We won't try to understand it all at once, but we shall feel that progress has been made if we are able to construct idealized mathematical models which reproduce a few of its features. We expect that any model we are now able to construct will be replaced by more complete ones in the future, and we do not know whether there is any natural end to this process.

The analogy with physical theories is deeper than a mere analogy of method. Often, the things which are most familiar to us turn out to be the hardest to understand. Phenomena whose very existence is unknown to the vast majority of the human race (such as the difference in ultraviolet spectra of iron and nickel) can be explained in exhaustive mathematical detail – but all of modern science is practically helpless when faced with the complications of such a commonplace fact as growth of a blade of grass. Accordingly, we must not expect too much of our models; we must be prepared to find that some of the most familiar features of mental activity may be ones for which we have the greatest difficulty in constructing any adequate model.

There are many more analogies. In physics we are accustomed to finding that any advance in knowledge leads to consequences of great practical value, but of an unpredictable nature. Röntgen's discovery of X-rays led to important new possibilities of medical diagnosis; Maxwell's discovery of one more term in the equation for curl H led to practically instantaneous communication all over the earth.

Our mathematical models for common sense also exhibit this feature of practical usefulness. Any successful model, even though it may reproduce only a few features of common sense, will prove to be a powerful extension of common sense in some field of application. Within this field, it enables us to solve problems of inference which are so involved in complicated detail that we would never attempt to solve them without its help.

1.3 The thinking computer

Models have practical uses of a quite different type. Many people are fond of saying, 'They will never make a machine to replace the human mind – it does many things which no machine could ever do.' A beautiful answer to this was given by J. von Neumann in a talk on computers given in Princeton in 1948, which the writer was privileged to attend. In reply to the canonical question from the audience ('But of course, a mere machine can't really *think*, can it?'), he said:

> You insist that there is something a machine cannot do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that!

In principle, the only operations which a machine cannot perform for us are those which we cannot describe in detail, or which could not be completed in a finite number of steps. Of course, some will conjure up images of Gödel incompleteness, undecidability, Turing machines which never stop, etc. But to answer all such doubts we need only point to the existence of the human brain, which *does* it. Just as von Neumann indicated, the only real limitations on making ‘machines which think’ are our own limitations in not knowing exactly what ‘thinking’ consists of.

But in our study of common sense we shall be led to some very explicit ideas about the mechanism of thinking. Every time we can construct a mathematical model which reproduces a part of common sense by prescribing a definite set of operations, this shows us how to ‘build a machine’, (i.e. write a computer program) which operates on incomplete information and, by applying quantitative versions of the above weak syllogisms, does plausible reasoning instead of deductive reasoning.

Indeed, the development of such computer software for certain specialized problems of inference is one of the most active and useful current trends in this field. One kind of problem thus dealt with might be: given a mass of data, comprising 10 000 separate observations, determine in the light of these data and whatever prior information is at hand, the relative plausibilities of 100 different possible hypotheses about the causes at work.

Our unaided common sense might be adequate for deciding between two hypotheses whose consequences are very different; but, in dealing with 100 hypotheses which are not very different, we would be helpless without a computer *and* a well-developed mathematical theory that shows us how to program it. That is, what determines, in the policeman’s syllogism (1.5), whether the plausibility for A increases by a large amount, raising it almost to certainty; or only a negligibly small amount, making the data B almost irrelevant? The object of the present work is to develop the mathematical theory which answers such questions, in the greatest depth and generality now possible.

While we expect a mathematical theory to be useful in programming computers, the idea of a thinking computer is also helpful psychologically in developing the mathematical theory. The question of the reasoning process used by actual human brains is charged with emotion and grotesque misunderstandings. It is hardly possible to say anything about this without becoming involved in debates over issues that are not only undecidable in our present state of knowledge, but are irrelevant to our purpose here.

Obviously, the operation of real human brains is so complicated that we can make no pretense of explaining its mysteries; and in any event we are not trying to explain, much less reproduce, all the aberrations and inconsistencies of human brains. That is an interesting and important subject; but it is not the subject we are studying here. Our topic is the *normative principles of logic*, and not the principles of psychology or neurophysiology.

To emphasize this, instead of asking, ‘How can we build a mathematical model of human common sense?’, let us ask, ‘How could we build a machine which would carry out useful plausible reasoning, following clearly defined principles expressing an idealized common sense?’

1.4 Introducing the robot

In order to direct attention to constructive things and away from controversial irrelevancies, we shall invent an imaginary being. Its brain is to be designed *by us*, so that it reasons according to certain definite rules. These rules will be deduced from simple desiderata which, it appears to us, would be desirable in human brains; i.e. we think that a rational person, on discovering that they were violating one of these desiderata, would wish to revise their thinking.

In principle, we are free to adopt any rules we please; that is our way of *defining* which robot we shall study. Comparing its reasoning with yours, if you find no resemblance you are in turn free to reject our robot and design a different one more to your liking. But if you find a very strong resemblance, and decide that you want and trust this robot to help you in your own problems of inference, then that will be an accomplishment of the theory, not a premise.

Our robot is going to reason about propositions. As already indicated above, we shall denote various propositions by italicized capital letters, $\{A, B, C, \text{etc.}\}$, and for the time being we must require that any proposition used must have, to the robot, an unambiguous meaning and must be of the simple, definite logical type that must be either true or false. That is, until otherwise stated, we shall be concerned only with two-valued logic, or Aristotelian logic. We do not require that the truth or falsity of such an 'Aristotelian proposition' be ascertainable by any feasible investigation; indeed, our inability to do this is usually just the reason why we need the robot's help. For example, the writer personally considers both of the following propositions to be true:

$A \equiv$ Beethoven and Berlioz never met.

$B \equiv$ Beethoven's music has a better sustained quality than that of
Berlioz, although Berlioz at his best is the equal of anybody.

Proposition B is not a permissible one for our robot to think about at present, whereas proposition A is, although it is unlikely that its truth or falsity could be definitely established today.[2] After our theory is developed, it will be of interest to see whether the present restriction to Aristotelian propositions such as A can be relaxed, so that the robot might help us also with more vague propositions such as B (see Chapter 18 on the A_p-distribution).[3]

1.5 Boolean algebra

To state these ideas more formally, we introduce some notation of the usual symbolic logic, or Boolean algebra, so called because George Boole (1854) introduced a *notation* similar to the following. Of course, the principles of deductive logic itself were well understood centuries before Boole, and, as we shall see, all the results that follow from Boolean algebra were contained already as special cases in the rules of plausible inference given

[2] Their meeting is a chronological possibility, since their lives overlapped by 24 years; my reason for doubting it is the failure of Berlioz to mention any such meeting in his memoirs – on the other hand, neither does he come out and say definitely that they did *not* meet.

[3] The question of how one is to make a machine in some sense 'cognizant' of the conceptual meaning that a proposition like A has to humans, might seem very difficult, and much of the subject of artificial intelligence is devoted to inventing *ad hoc* devices to deal with this problem. However, we shall find in Chapter 4 that for us the problem is almost nonexistent; our rules for plausible reasoning automatically provide the means to do the mathematical equivalent of this.

by (1812). The symbol

$$AB, \tag{1.6}$$

called the *logical product* or the *conjunction*, denotes the proposition 'both A and B are true'. Obviously, the order in which we state them does not matter; AB and BA say the same thing. The expression

$$A + B, \tag{1.7}$$

called the *logical sum* or *disjunction*, stands for 'at least one of the propositions, A, B is true' and has the same meaning as $B + A$. These symbols are only a shorthand way of writing propositions, and do not stand for numerical values.

Given two propositions A, B, it may happen that one is true if and only if the other is true; we then say that they have the same *truth value*. This may be only a simple tautology (i.e. A and B are verbal statements which obviously say the same thing), or it may be that only after immense mathematical labor is it finally proved that A is the necessary and sufficient condition for B. From the standpoint of logic it does not matter; once it is established, by any means, that A and B have the same truth value, then they are logically equivalent propositions, in the sense that any evidence concerning the truth of one pertains equally well to the truth of the other, and they have the same implications for any further reasoning.

Evidently, then, it must be the most primitive axiom of plausible reasoning that two propositions with the same truth value are equally plausible. This might appear almost too trivial to mention, were it not for the fact that Boole himself (Boole, 1854, p. 286) fell into error on this point, by mistakenly identifying two propositions which were in fact different – and then failing to see any contradiction in their different plausibilities. Three years later, Boole (1857) gave a revised theory which supersedes that in his earlier book; for further comments on this incident, see Keynes (1921, pp. 167–168); Jaynes (1976, pp. 240–242).

In Boolean algebra, the equal sign is used to denote not equal numerical value, but equal truth value: $A = B$, and the 'equations' of Boolean algebra thus consist of assertions that the proposition on the left-hand side has the same truth value as the one on the right-hand side. The symbol '$\equiv$' means, as usual, 'equals by definition'.

In denoting complicated propositions we use parentheses in the same way as in ordinary algebra, i.e. to indicate the order in which propositions are to be combined (at times we shall use them also merely for clarity of expression although they are not strictly necessary). In their absence we observe the rules of algebraic hierarchy, familiar to those who use hand calculators: thus $AB + C$ denotes $(AB) + C$; and not $A(B + C)$.

The *denial* of a proposition is indicated by a bar:

$$\overline{A} \equiv A \text{ is false.} \tag{1.8}$$

The relation between A, $\overline{A}$ is a reciprocal one:

$$A = \overline{A} \text{ is false,} \tag{1.9}$$