THE PROBABILITY THAT TWEETY IS ABLE TO FLY

Giangiacomo Gerla
Dipartimento di Matematica e Fisica, Università di Camerino ITALY.

Abstract. Consider the question of assigning a probabilistic valuation to a statement as *"Tweety (a particular bird) is able to fly"*. We suggest that a natural way to proceed is to rewrite it as *"a (randomly chosen) bird with the same observable properties of Tweety is able to fly"*, and consequently to assume that the probability of *"Tweety is able to fly"* is equal to the percentage of the past observed birds similar to Tweety that are able to fly.

Note: A similar paper was published in the International Journal of Intelligent Systems, 9 (1994).

1. Introduction

F.Bacchus in [1] and J.Y.Halpern in [2] compare statements like:

(i) *"a (randomly chosen) bird will fly "*;

(ii) *" Tweety (a particular bird) flies "*

and emphasize their complete difference with respect to the question of assigning them probabilistic valuations. This induces to consider two different semantics. Indeed, the claim that the probability of (i) is, for example, 0.9, expresses statistical information about the proportion of fliers among the set of birds. Such information is not subjective in nature, it follows, in a sense, from the objective state of the world. To express more formally such an idea, recall the notion of type-1 probability structure proposed by Halpern. Let L be a first order language, denote by F its set of formulas and by F_n the set of formulas whose free variables are among $x_1,...,x_n$. Given a model M of L whose domain we denote by D, a formula $\alpha \in F_n$ and $d_1,...,d_n$ elements of D, as usual we write $M \vDash \alpha\ [d_1,...,d_n]$ to denote that α is true in m provided that $x_1,...,x_n$ are interpreted by the elements $d_1,...,d_n$, respectively. Also, let $\mu : P(D) \to [0,1]$ be a probability defined in the class $P(D)$ of all the subsets of D (in the sequel by *probability* we mean a finitely additive probability). Then, we say that the pair (M,μ) is *a type-1 probability structure* and, given a formula $\alpha \in F_1$, we define *the probability of* α by the formula

$$p(\alpha) = \mu(\{d \in D : M \vDash \alpha\ [d]\}),$$

that is the measure of the extension of the predicate α in D. In our example, we have to consider a type-1 probability structure in which L is a language with a monadic predicate *"Flies"*, D is the set of birds and "Flies" is interpreted in D by the set of flying birds. Then (i) is valued by the number $\mu(\{d \in D : M \vDash Flies(x_1)\ [d]\})$.

Instead, if we say that the probability of (ii) is 0.9, such an interpretation is meaningless. In fact, since (ii) is a closed formula, it is either true or false and we have to interpret 0.9 as a degree of belief about (ii). More generally, it seems impossible to assign probability valuations to closed formulas like (ii) by using only one model as in the case (i). This leads Halpern to propose a different mathematical device, the type-2 probability structures, constructed by using the notion of possible world. Namely, a *type-2 probability structure* is defined by a nonempty set S, whose elements are called *states* or *possible worlds*, together with a probability $\mu : P(S) \to [0,1]$ and a family $(M_s)_{s \in S}$ of models of L with the same domain D. Then, (ii) is evaluate by the number $\mu(\{s \in S : M_s \vDash Flies(Tweety)\})$, that is the probability of a state where Tweety flies.

Now, I agree with this suggestion if Tweety is a bird that sometimes flies and sometimes does not. So, if its behavior has been examined in a set S of past observations, the probability of (ii) is assumed to be equal to the relative frequency in S of the flights of Tweety. But the idea that Halpern and Bacchus had in mind is different. *"Tweety flies"* means *"Tweety is a bird able to fly"*, and a probabilistic valuation of (ii) is a *degree of belief* imposed by my lack of knowledge about the capabilities of Tweety and not by the randomness of its ability to fly. So, they use possible worlds semantics in an epistemic sense: since my knowledge is incomplete, I can imagine worlds where Tweety is able to fly and worlds where it does not. But, in spite of the mathematical elegance of

the notion of type-2 probability structure, the consideration of a set S of possible worlds, each with a different extension of the predicate "Flies", seems to me a rather artificial device. This is a matter of taste, obviously, but, as an example, it is hard for me to connect the type-2 probability structures with the effective processes leading an agent to formulate a degree of belief. As a matter of fact, I am convinced that we derive our degrees of belief from the past experience about a large class of similar cases, in general, and that such an experience is stored in our mind in a statistical (not necessarily conscious) form. In accordance with this opinion, I suggest to restate (ii) as follows:

"a (randomly chosen) bird with the same observable properties of Tweety flies"

and to assume that the claim that the probability of (ii) is 0.9 means that

"90% of all birds satisfying the same properties I am able to know about Tweety fly".

Thus, I may assign a probability to (ii) in the same manner as for (i).

2. Probability and non observable properties.

The notion of "probability valuation" of a formalized language is on the basis of any approach to probability logic. Recall that, if L_0 is a zero order language and F its set of formulas, a *probability valuation* of L_0 is any function $p : F \to [0,1]$ such that

$$\alpha \text{ tautology} \Rightarrow p(\alpha)=1 \quad ; \quad \alpha\wedge\beta \text{ contradiction} \Rightarrow p(\alpha) + p(\beta) = p(\alpha\vee\beta).$$

It is well known that the probability valuations are not truth-functional since the knowledge of the probability of two formulas α and β is not sufficient to determine the probability of $\alpha\wedge\beta$, in general. This is a basic obstacle to manage information probabilistic in nature and induces to introduce, as an intermediate step, non-numerical information Boolean in nature. Recall that, if $\boldsymbol{B}$ is a Boolean algebra, then that a *Boolean valuation* is a map $v : F \to \boldsymbol{B}$ such that

(i) α tautology $\Rightarrow v(\alpha)=1$;

(ii) $v(\alpha\vee\beta) = v(\alpha)\vee v(\beta)$;

(iii) $v(\alpha\wedge\beta) = v(\alpha)\wedge v(\beta)$;

(iv) $v(-\alpha) = -v(\alpha)$.

If $\boldsymbol{B} = \{0,1\}$, the Boolean valuations coincide with the usual classical interpretations of L_0. A Boolean valuation is truth-functional and is determined by its values in the propositional variables. Moreover, it is well known that p is a probability valuation if and only if a Boolean algebra $\boldsymbol{B}$, a Boolean valuation $v : F\to\boldsymbol{B}$ and a probability $\mu : \boldsymbol{B}\to[0,1]$ exist such that $p(\alpha) = \mu(v(\alpha))$ for every $\alpha\in F$. The following definition enables us to define Boolean valuations and related probability valuations in a rather natural way.

Definition 1. A *statistical data-base* in L_0 is a pair (Db,μ) where $Db =(v_c)_{c\in C}$ is a finite family of classical interpretations of L_0 and $\mu : P(C)\to[0,1]$ is a probability such that $\mu(X) \neq 0$ whenever $X \neq \varnothing$.

Elements of C are called *past cases* and we may imagine a statistical data-base as a result of statistical observations. Namely, that

- C is a finite set of past cases that were being examined
- for every $c\in C$, v_c is a description of the properties satisfied by c
- for every $X\subseteq C$, $\mu(X)$ is the relative frequency of X in C.

A statistical data-base determines a Boolean valuation $v : F \to P(C)$ by

$$v(\alpha)=\{c\in C \mid v_c(\alpha)=1\} \qquad \forall\alpha\in F,$$

that is $v(\alpha)$ is the set of past cases satisfying α. Also, a probability valuation $p: F \to[0,1]$, the *priori probability*, is obtained by

$$p(\alpha) = \mu(v(\alpha)) \qquad \forall\alpha\in F,$$

that is $p(\alpha)$ is the percentage of past cases satisfying α.

Definition 2. A *probability framework* in L_0 is a structure (Db, μ,K,m) where:

- (Db,μ) is a statistical data-base in L_0
- K is a set of formulas in L_0 which we call *observable*
- m is an interpretation of L_0, the *actual case*, such that $\exists c \in C \; \forall \alpha \in K \; m(\alpha) = v_c(\alpha)$.

We have to look at K as at the set of (relevant) properties of the actual case an agent is able to recognize. We call *evidence* of the actual case m the class $K(m) = \{\alpha \in K : m(\alpha)=1\}$ of the observable formulas satisfied by m, we say that a case $c \in C$ is *K-similar* (in brief, *similar*) to m if it satisfies any formula in $K(m)$. In other words, a past case c is similar to m if it satisfies the same relevant observable properties of m. We denote by $S(m)$ the set of cases similar to m. The last condition in Definition 2 expresses the fact that we may use a statistical data-base to an actual case m only if past cases similar to m exist. The basic idea is that, to evaluate a non observable property about the actual case, we have to refer to the class of similar cases in the statistical data-base.

Definition 3. The *probability of* α in m is the number $p(\alpha,m)$ defined by:

$$p(\alpha,m) = \mu(v(\alpha)/S(m)) = \frac{\mu(\{c \in S(m) / v_c(\alpha) = 1\})}{\mu(S(m))}.$$

In other words, $p(\alpha,m)$ is the percentage of past cases similar to m satisfying α or, equivalently, the probability of α given the evidence of m. Notice that $p(\alpha,m)$ depends on the evidence of m and not on m, in a sense, so may happen that two agents with different evidences formulate different valuations. The following proposition is obvious.

Proposition 1. The function $p(\_,m) : F \rightarrow [0,1]$ is a probability valuation extending the valuation of the observable properties of m, that is, for every observable formula α, $p(\alpha,m) = 1$ if α is satisfied by m and $p(\alpha;,m) = 0$ otherwise.

As a consequence, if we are able to test directly the validity of α in the actual case, then, since $p(\alpha,m) = m(\alpha)$, there is no reason to search for its probability in the statistical data-base. Otherwise, Definition 3 gives a default rule to evaluate α in m.

3. First order approach

Bacchus and Halpern work in the framework of first order logic. We may give a first order version of the probability frameworks as follows. We call *probability framework in first order form* an object of type (M,μ,K,m) where
- (M, μ) is a 1-type probability structure in a first order language L,
- K is a set of formulas in L we call *observable*,
- $m \subseteq F_1$ is a consistent complete type realized in M, we call *actual case*.

In such a way, the actual case is identified with the set of its properties in the language L. Notice that, by hypothesis, $d \in D$ exists such that $m = \{\alpha \in F_1 : M \vDash \alpha\,[d]\}$. We set $K(m)$ equal to the set of observable formulas satisfied by m, that is $K(m) = m \cap K$. Given a probabilistic framework in first order form, the set of cases similar to m is $S(m) = \{d \in D : \forall\ \alpha \in K(m)\ M \vDash \alpha\,[d]\}$, the Boolean valuation of a formula α is $v(\alpha) = \{d \in D : M \vDash \alpha\,[d]\}$ and the probabilistic valuation by

$$p(\alpha,m) = \mu(v(\alpha)/S(m)) = \frac{\mu(\{d \in S(m) : M \models \alpha[d]\})}{\mu(S(m))}.$$

Such approach contains the zero order approach proposed in the previous section. Indeed, given a probability framework (Db,μ,K,m) in a zero order language L_0, we may build up the first order probability framework (M,μ,K^*, m^*) by assuming that:
- L is the first order language obtained by considering a monadic predicate r_i for every propositional variable p_i in L_0 ;

- (M,μ) is the 1-type probability structure whose domain is the set C of past cases and whose predicate r_i are interpreted by $\{c \in C : v_c(p_i) = 1\}$.
- $K^* = \{\alpha^* : \alpha \in K\}$
- $m^* = \{\alpha^* : m(\alpha) = 1, \alpha \in F\}$

where, for every formula α in L_0, α^* is the formula of L obtained by substituting each p_i occurring in α with $r_i(x_1)$. It is easy to prove that the *probabilistic* valuation $p(\alpha,m)$ in (Db, μ, K, m) coincides with the probabilistic valuation $p(\alpha^*,m^*)$ in (M, μ, K^*, m^*).

In conclusion, a one-model semantics for statements as (ii) seems to be possible, and the 1-type models are adequate both for statements of type (i) and (ii).

4. Coming back to Tweety.

Let L_0 be a zero order language whose propositional variables express properties as

"is able to fly" , *"is a penguin"*, *"is boring"*, *"has little wings"*

and so on. Moreover, consider the probability framework in which C is the set of birds examined in past observations and Tweety is the actual case. Then, if we denote by α the property "is able to fly" and assume that α is not observable, the probability of (ii) is $p(\alpha,Tweety) = \mu(v(\alpha)/S(Tweety))$ i.e. the probability of flying for a bird similar to Tweety. Such valuation depends on the "evidence" about Tweety, i.e. the class of properties of Tweety we are able to know. So, if we know that Tweety is a penguin, then we can conclude that the probability of (ii) is zero. Instead, if we only know that Tweety is a bird, then we cannot distinguish (ii) from (i) and we are forced to assign to (ii) the same probability as (i).

Observe also that, in the semantics proposed by Bacchus and Halpern, Tweety and the other birds play similar roles while in a probability framework the role of actual case is different in nature from the one of the past cases. As a matter of fact, the past cases and therefore the statistical data-base, represent the "general system of knowledge" and therefore the inferential engine of the probabilistic judgements. Instead the actual case represents the source of information we have to insert in this engine. At this regard it is interesting to notice that from $p(\mathrm{m}, \alpha)=1$ we cannot infer that α is true in m. Indeed, from $p(\alpha,m) = 1$ we may infer only that $\mu(\{c \in S(m) : v_c(\alpha)=1\}) = \mu(S(m))$ and therefore that $\mu(\{c \in \mathrm{S}(m) : v_c(\alpha) = 0\} = 0$. This is equivalent to say that every past case similar to m satisfies α and not that α is actually true in m.

5. Medical diagnosis.

I will attempt to apply the concepts given in this note to give a model of the inferential process of a physician in stating a diagnosis for a given patient. Obviously, I refer to an abstract and "simplified" process. Now, two things play a role in this case: the past experience of the physician and of the physician's community (the general theory) and the information about the patient that it is possible to obtain. In terms of the definition we have proposed, the past experience determines a statistical data-base, the information about the patient the evidence. Namely, we can define a probability framework by imagining that:

- in the used medical language L_0 there is a set K of observable formulas for symptoms or results of possible tests, while the possible diagnosis are not observable;
- the past experience was stored in a statistical data-base containing for every (past) clinical case the related symptoms and diagnosis ;
- the patient is the actual case.

Then if α is a diagnosis, the probability $p(m, \alpha)$ is the percentage of the past clinical cases with the same symptoms as m satisfying α.

As a matter of fact, the physician's community cannot be able to record all the past cases, but this is not necessary since it may refer to a smaller data-base of "typical" cases. Indeed, call *similar* two cases satisfying the same formulas in L_0, and consider the statistical data-base obtained as a quotient of the initial data-base modulo this relation. Namely, we call a *typical case* any complete

class of equivalence $c^* = \{x \in C : x \text{ similar to } c\}$ and we refer to the set $C^* = \{c^* : c \in C\}$ of typical cases. Moreover, we define the probability $\mu : P(C^*) \to [0,1]$ by setting $\mu(\{c^*\}) = Card(c^*)/\text{Card}(C)$ (so, the typical cases are not equiprobable). The statistical data-base so obtained is no too large since if in L_0 there are n propositional variables the cardinality of C^* is less or equal to 2^n. Also, it can be continuously updated simply by modifying μ in an obvious manner, as new data becomes available. In this sense, a probability framework, like a neural network, is able to learn from the experience.

6. Some final observations.

We conclude this note by considering two questions. The first is:

"is the notion of probability framework a step toward the construction of "a logic of the probabilistic inferences" ?

The answer depends on the meaning we assign to the word "logic", obviously. For example, the answer is positive provided that we interpret:

- the statistical data-base as a deductive apparatus
- the set $K(m)$ of formulas as a system of axioms
- the function $p(\_,m) : F \to [0,1]$ as the set of (logical) consequences of $K(m)$.

Instead, if we think that the existence of an explicit system of rules to manipulate formulas is on the basis of any logic and if we think that the nature of any logic is linguistic, then the answer is negative. This in spite of the fact that, if we refer at the conscious level, a probabilistic inference looks to be formulate by statements as

"Since it is very probable that α ... " , *"now, in general, if α then β ... "* , *"since α is less probable than β, ... "*

and so on.

But, as matter of fact the information carried on by these statements is only a little part of the actual (non conscious) information possessed and used by the subject. Moreover, the actual inferential mechanism is hidden and the conscious and verbal activity accomplishing a probabilistic inference is only an echo of a (non conscious) primitive independent activity. Then, in a sense, no logic of probabilistic inference is possible.

The second question is the following. Since it seems that the data-base enabling us to build up a probability framework is obtainable in a mechanical way (simply by storing past cases) one would be tempted to conclude that, for example, a diagnostic system can be actually available in a mechanical way. Unfortunately, things ore not so simple as it is well known by scholars of inductive processes. As an example, consider the following paradox. Assume that, as it is usual, among the available information we consider the age of the patient but that, since the physician is a very fussy and precise person, this age is measured by considering also the months. As a consequence is rather likely that in the past cases there is no patient with exactly the age of the actual case so that a too complete information as *"he is seventy years and two months old"* annihilates any possible inferential process. The situation is not better if in the statistical data-base (equivalently, in the memory of the physician) only one past case with such an age exists. Indeed, if for example in this case the patient was killed by a cancer, then we have to conclude that the knowledge of the age of the actual patient is sufficient to conclude that he has a cancer! Obviously, this is completely unsatisfactory and it is due to the unimportance of the exact age, in spite of the relevance of the approximate age. In conclusion, the way how a probability framework may be build up is a very complicate question, but this question does not originate with the flying of Tweety. It belongs to the more general problem of the induction. In spite of that, I think that a large class of inferential processes probabilistic in nature should be simulated by suitable probability frameworks and that such a notion should be useful for the construction of diagnostic systems.

References.

[1] F.Bacchus, *Lp*, a logic for representing and reasoning with statistical knowledge, *Comput. Intell.*, **6** (1990) 209-231.
[2] J.Y.Halpern, An analysis of first-order logic of probability, *Artificial Intelligence*, **46** (1990) 331-350.