PROBABILITIES THAT IMPLY CERTAINTIES

Haim Shvaytser (Schweitzer)
SRI, David Sarnoff Research Center
CN5300
Princeton, NJ 08543-5300
haim@vision.sarnoff.com

Table 1: the four worlds of R and C

world	R	C	probability
w_1	FALSE	FALSE	p_1
w_2	FALSE	TRUE	p_2
w_3	TRUE	FALSE	p_3
w_4	TRUE	TRUE	p_4

Abstract

A method is described for deriving rules of inference from relations between probabilities of sentences in Nilsson's probabilistic logic.

Introduction

One intuitive interpretation of probability is a measure of uncertainty. In many of the application areas for artificial intelligence it is important to be able to reason with uncertain information; this has motivated research in developing methods for probabilistic inference. See, for example, [Nilsson, 1986, Fagin and Halpern, 1988, Pitt, 1989].

A precise model for dealing with probabilities of sentences in predicate calculus was suggested by Nilsson in [Nilsson, 1986]. In Nilsson's probabilistic logic the probability of a sentence is its average truth value in possible worlds. Consider the following example: Let R and C be two sentences; in a specific world a sentence is either TRUE or FALSE. The truth table of all worlds of R and C is given in Table 1. The probabilities of worlds are determined by an arbitrary probability distribution, i.e., four values p_1, p_2, p_3, p_4, such that $p_i \geq 0$ for $i = 1, \dots, 4$, and:

$$p_1 + p_2 + p_3 + p_4 = 1.$$

From Table 1 we see that R is true in the worlds w_3 and w_4, so that its average truth value is $p_3 + p_4$, while C is true in the worlds w_2 and w_4, and its average truth value is $p_2 + p_4$. Therefore, $\mathsf{Prob}(\mathsf{R}) = p_3 + p_4$ and $\mathsf{Prob}(\mathsf{C}) = p_2 + p_4$.

The probability of other formulae involving R and C can also be computed from Table 1. Thus, since $\mathsf{R} \rightarrow \mathsf{C}$ is true in w_1, w_2, w_4, we have:

$$\mathsf{Prob}(\mathsf{R} \rightarrow \mathsf{C}) = p_1 + p_2 + p_4,$$

and from similar arguments:

$$\mathsf{Prob}(\mathsf{C} \rightarrow \mathsf{R}) = p_1 + p_3 + p_4.$$

Now if R stands for the sentence "it rains" and C for the sentence "it is cloudy", the world w_3 is impossible. In this case the value of p_3 in Table 1 is 0, $\mathsf{Prob}(\mathsf{R} \rightarrow \mathsf{C}) = 1$, and $\mathsf{Prob}(\mathsf{C} \rightarrow \mathsf{R}) = p_1 + p_4$.

In the process of reasoning with probabilistic information we are given probabilities of sentences, and either reason about probabilities of other sentences or learn new information about a specific world. Thus, since $\mathsf{Prob}(\mathsf{R} \rightarrow \mathsf{C}) = 1$ we can deduce that it is cloudy in a world w' if we know that it rains in w'. On the other hand, R cannot be deduced from C without the additional information that in a specific world $p_2 = 0$ because if $p_2 \neq 0$ then $\mathsf{Prob}(\mathsf{C} \rightarrow \mathsf{R}) < 1$.

In this paper we describe a method for identifying sentences that are true with probability 1 (i.e., in all possible worlds) from probabilities of sentences that are not necessarily true in all possible worlds. As an example, notice that for any two sentences X, Y:

$$X \rightarrow Y \equiv (X \wedge Y) \vee (\neg X)$$

so that:

$$\mathsf{Prob}(X \rightarrow Y) = \mathsf{Prob}(X \wedge Y) + 1 - \mathsf{Prob}(X).$$

Therefore, $\mathsf{Prob}(X \to Y) = 1$ if and only if $\mathsf{Prob}(X \wedge Y) = \mathsf{Prob}(X)$. Specifically, if it is known that, say, $\mathsf{Prob}(\mathsf{R}) = 0.7$ and $\mathsf{Prob}(\mathsf{R} \wedge \mathsf{C}) = 0.7$ then it must be that $\mathsf{R} \to \mathsf{C}$ in all possible worlds. This is a special case of results that are described in the paper.

Definitions

The following definitions of possible worlds and probabilities of sentences are the same as those in [Nilsson, 1986].

Let $\phi_1, \ldots, \phi_n$ be n sentences in predicate calculus. A *world* is an assignment of truth values to $\phi_1, \ldots, \phi_n$. There are 2^n worlds; some of these worlds are *possible worlds* and the others are *impossible worlds.* A world is impossible if and only if the truth assignment to $\phi_1, \ldots, \phi_n$ is logically inconsistent. For example, if $\phi_2 = \neg\phi_1$ then all worlds with both $\phi_1 = \mathsf{TRUE}$ and $\phi_2 = \mathsf{TRUE}$ are impossible.

We denote by PW the set of possible worlds. An arbitrary probability distribution D is associated with PW such that a world $w \in \mathsf{PW}$ has probability $D(w) \geq 0$, and:

$$\sum_{w \in \mathsf{PW}} D(w) = 1.$$

The truth value of a formula ϕ in the primitive variables $\phi_1, \ldots, \phi_n$ is well defined in all possible worlds. The probability of ϕ is defined as:

$$\mathsf{Prob}(\phi) = \sum_{\substack{w \in \mathsf{PW} \\ \phi \text{ is true in } w}} D(w). \tag{1}$$

Random variables

A *random variable* X_w is a function that has a well defined (real) value in each possible world. With a formula ϕ we associate the random variable $w(\phi)$ that has the value of 1 in worlds where ϕ is true and the value of 0 in worlds where ϕ is false. Equation (1) can now be written as:

$$\mathsf{Prob}(\phi) = \sum_{w \in \mathsf{PW}} w(\phi) \cdot D(w). \tag{2}$$

Definition: The expected value of the random variable X_w is:

$$E(X_w) = \sum_{w \in \mathsf{PW}} X_w \cdot D(w). \tag{3}$$

From Equation (2) we see that for any formula ϕ:

$$\mathsf{Prob}(\phi) = E(w(\phi)). \tag{4}$$

Rules of inference

We consider (deterministic) rules of inference of the following type:

Let X_w be a random variable and ϕ a formula. If:

$$w(\phi) = X_w \quad \textit{in possible worlds}$$

then from $X_w = 1$ infer ϕ and from $X_w = 0$ infer $\neg\phi$.

We investigate only a restricted case of these rules in which X_w can be expressed as a linear combination of the variables $w(\phi_i)$:

If there are coefficients a_{ij} such that:

$$w(\phi_j) = \sum_{i \neq j} a_{ij} w(\phi_i) \quad \textit{in possible worlds}$$

then from $\sum_{i \neq j} a_{ij} w(\phi_i) = 1$ infer ϕ_j and from $\sum_{i \neq j} a_{ij} w(\phi_i) = 0$ infer $\neg\phi_j$.

We call rules of inference of this type *linear rules of inference.*

The main result of this paper is a method for deriving a *complete* set of linear rules of inference. By this we mean a finite set of linear rules of inference RI such that: if there is a set of linear rules of inference that can infer a formula ψ then ψ can also be inferred from RI.

Algebraic structure

Linear rules of inference can be expressed as:

$$w(\phi_j) - \sum_{i \neq j} a_{ij} w(\phi_i) = 0 \quad \textit{in possible worlds.}$$

The left hand side is a linear combination of the random variables $w(\phi_i)$ $i = 1, \ldots, n$, that vanishes in all possible worlds. A complete set of linear rules of this type can be obtained by observing that these rules are all elements of a finite dimensional vector space, and therefore, any basis of this vector space is a complete set of linear rules of inference.

In order to determine a basis to the vector space of linear rules of inference we consider three vector spaces:

- $V = \mathsf{Span}\{w(\phi_1), \cdots, w(\phi_n)\}$.
  An element $v \in V$ is a random variable that can be expressed as $v = \sum_i a_i w(\phi_i)$.
- $W = \{v \in V : v = 0 \text{ in possible worlds}\}$.
  W is the vector space of elements of V that vanish in all possible worlds. Therefore, each element of W can be used as a linear rule of inference.
- $U = V/W$.
  U is the quotient space of V by W. See Chapter 4 in [Herstein, 1975] (or any other basic text on Algebra) for the exact definition. Its elements are subsets of V in the form of $v + W$, where $v \in V$.

There is a natural homomorphism of V onto U with the kernel W. The elements of U are the equivalence classes of V, where two elements $v_1, v_2 \in V$ are equivalent if and only if $v_1 = v_2$ in all possible worlds. We use the notation $v \pmod W$ for the equivalence class (element of U) of v. Thus, if $v_1 = v_2$ in all possible worlds we write $v_1 = v_2 \pmod W$.

The bases of the vector spaces V, W, U are related in a simple way. If $v_1, \cdots, v_t$ is a basis of V, and the equivalence classes of $v_1, \cdots, v_d$ form a basis of U $(d \leq t)$, then there are coefficients b_{ij} for $i = 1, \ldots, t-d$ such that:

$$v_{d+i} = \sum_{j=1}^{d} b_{ij} v_j \pmod W. \tag{5}$$

Furthermore, the $t-d$ random variables w_i, $i = 1, \ldots, t-d$, that are given by:

$$w_i = v_{d+i} - \sum_{j=1}^{d} b_{ij} v_j$$

form a basis of W. (See Chapter 4 in [Herstein, 1975].)

We conclude that a basis for W is a complete set of linear rules of inference which can be found by computing the linear dependencies in the vector space U that are given by Equation (5).

Example: Let R and C be the two sentences from the example that was discussed in the introduction, where $\mathsf{R} = \mathsf{TRUE}$, $\mathsf{C} = \mathsf{FALSE}$ is an impossible world. Let $\phi_1 = \mathsf{R}$, $\phi_2 = \mathsf{C}$, and $\phi_3 = \mathsf{R} \wedge \mathsf{C}$. The corresponding random variables are: $x_1 = w(\phi_1)$, $x_2 = w(\phi_2)$, and $x_3 = x_1 \cdot x_2 = w(\phi_3)$. If we take V as $\mathsf{Span}\{x_1, x_2, x_3\}$ then $\{x_1, x_2, x_3\}$ is a basis of V, and the equivalence classes of x_1 and x_2 form a basis of U. The formula $\mathsf{R} \rightarrow \mathsf{C}$ can be expressed in terms of x_1, x_2, x_3 as $x_3 = x_1$, which is a linear rule of inference, so that:

$$x_3 = x_1 \pmod W,$$

and $x_3 - x_1 \in W$. It can be shown that $x_3 - x_1$ is a basis of W.

Correlations and the correlation matrix

Let E be the expected value operator as defined in Equation (3). The following observations enable easy computation of linear dependencies in U by standard statistical techniques. For any two random variables $x, y \in V$:

- $x = y \pmod W \implies E(x) = E(y)$.
- $x = 0 \pmod W \iff E(x^2) = 0$.

Based on these observations we show that linear dependencies in the vector space U can be computed by applying standard statistical techniques.

The correlation of two random variables x, y is defined in the standard way as $E(xy)$. Let $\{x_1, \ldots, x_t\}$ be t random variables from V. Their correlation matrix is the $t \times t$ matrix $R = (r_{ij})$, where r_{ij} is the correlation value of x_i and x_j. The matrix R depends on the probability distribution D, but the following properties of R hold *for all* probability distributions. (For proofs see Chapter 8 in [Papoulis, 1984].)

a) If the equivalence classes of $x_1, \ldots, x_t$ are linearly independent in U then R is nonsingular.

b) If the equivalence classes of $x_1, \ldots, x_t$ are linearly dependent in U then R is singular.

c) If the equivalence classes of $x_1, \ldots, x_{t-1}$ are linearly independent in U, but the equivalence classes of $x_1, \ldots, x_t$ are linearly dependent in U then

$$x_t = a_1 x_1 + \cdots + a_{t-1} x_{t-1} \pmod W \tag{6}$$

and $a_1, \ldots, a_{t-1}$ can be obtained from the system of linear equations

$$R \cdot \begin{pmatrix} a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_{t-1} \end{pmatrix} = \begin{pmatrix} r_{1,t} \\ \cdot \\ \cdot \\ \cdot \\ r_{t-1,t} \end{pmatrix} \tag{7}$$

where the matrix R is the correlation matrix of $x_1, \ldots, x_{t-1}$.

Table 2: specific world probabilities

world	R	C	probability
w_1	FALSE	FALSE	0.1
w_2	FALSE	TRUE	0.2
w_3	TRUE	FALSE	0
w_4	TRUE	TRUE	0.7

The following algorithm uses the properties of the correlation matrix to generate a basis for W in the form of linear rules of inference. Its input is the correlation values of a set of random variables $S = \{x_1, \ldots, x_n\}$. In the algorithm we denote by $I \subset S$ a set of random variables that are linearly independent modulo W (i.e, their equivalence classes are linearly independent in U), and R is their correlation matrix.

Algorithm: Initially, let $I = \{x_1\}$, so that R is (r_{11}), a matrix of size 1×1.

For each $x_t \in S$:

1- Let R' be the correlation matrix of the random variables in $I \cup \{x_t\}$.

2- If R' is singular solve the system of equations (7) and output the linear rule of inference (6); otherwise, $I \leftarrow I \cup \{x_t\}$, and $R \leftarrow R'$.

Since the algorithm computes the linear dependencies that are given by Equation (6) it generates a basis for W, which is a complete set of linear rules of inference.

Example: Let x_1, x_2, x_3, be the random variables from the example that was given at the end of the previous section, with $x_1 = w(\mathsf{R})$, $x_2 = w(\mathsf{C})$, and $x_3 = w(\mathsf{R} \wedge \mathsf{C})$. If the four worlds of R and C appear with probabilities as given in Table 2 we have:

$$\begin{aligned} r_{11} &= E(x_1 \cdot x_1) = \mathsf{Prob}(\mathsf{R}) = 0.7 \\ r_{12} = r_{21} &= E(x_1 \cdot x_2) = \mathsf{Prob}(\mathsf{R} \wedge \mathsf{C}) = 0.7 \\ r_{13} = r_{31} &= E(x_1 \cdot x_2) = \mathsf{Prob}(\mathsf{R} \wedge \mathsf{C}) = 0.7 \\ r_{22} &= E(x_2 \cdot x_2) = \mathsf{Prob}(\mathsf{C}) = 0.9 \\ r_{23} = r_{32} &= E(x_2 \cdot x_3) = \mathsf{Prob}(\mathsf{R} \wedge \mathsf{C}) = 0.7 \\ r_{33} &= E(x_3 \cdot x_3) = \mathsf{Prob}(\mathsf{R} \wedge \mathsf{C}) = 0.7 \end{aligned}$$

The correlation matrix of $\{x_1, x_2\}$ is:

$$\begin{pmatrix} 0.7 & 0.7 \\ 0.7 & 0.9 \end{pmatrix}.$$

The correlation matrix of $\{x_1, x_2, x_3\}$ is:

$$\begin{pmatrix} 0.7 & 0.7 & 0.7 \\ 0.7 & 0.9 & 0.7 \\ 0.7 & 0.7 & 0.7 \end{pmatrix}.$$

The correlation matrix of $\{x_1, x_2\}$ is non-singular, but the correlation matrix of $\{x_1, x_2, x_3\}$ is singular, and the system of equations (7) gives:

$$\begin{pmatrix} 0.7 & 0.7 \\ 0.7 & 0.9 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix}.$$

The solution is $a_1 = 1$ and $a_2 = 0$, which gives the rule of inference $x_3 = x_1$, i.e.,

$$w(\mathsf{R} \wedge \mathsf{C}) = w(\mathsf{R}) \quad \textit{in possible worlds}$$

which is equivalent to

$$\mathsf{R} \rightarrow \mathsf{C} \quad \textit{in possible worlds.}$$

Notice that this result was obtained from the probabilities of the sentences R, C, and R $\wedge$ C, and *not* from Table 2.

Inference rules as CNF formulae

Our algorithm for deriving rules of inference from probabilities can be used only when linear rules of inference exist. In this section we show how the algorithm can be applied to derive other types of rules of inference.

We consider rules of inference that are variations of modus ponens:

Let X, Y be two sentences such that $X \rightarrow Y$ in all possible worlds. Then in a world where $X = \mathsf{TRUE}$ infer $Y = \mathsf{TRUE}$.

Let $\phi_1, \ldots, \phi_n$ be n sentences. We would like to derive rules of the type:

$$\Psi \rightarrow \phi_i \tag{8}$$

where Ψ is a formula in the sentences ϕ_j, $j \neq i$. Notice that Equation (8) can also be written as:

$$\Psi = \Psi \wedge \phi_i \quad \textit{in possible worlds.}$$

Therefore, using the random variables $w(\phi_i)$ and $w(\Psi \wedge \phi_i)$ we can write Equation (8) in the equivalent form:

$$w(\Psi) = w(\Psi \wedge \phi_i) \quad \textit{in possible worlds.} \tag{9}$$

The reason that the results of previous sections cannot be applied directly to derive rules of the type of (9) is that Equation (9) is not a linear rule of inference for the sentences $\phi_1, \ldots, \phi_n$.

The basic idea of this section is that rules of the type of Equation (9) can be linearized by adding sentences to $\phi_1, \ldots, \phi_n$. For example, notice that if we add the $\binom{n}{2}$ sentences $\phi_i \wedge \phi_j$ for $i \neq j$ to $\phi_1, \ldots, \phi_n$ then all formulae of the type $\phi_\alpha \rightarrow \phi_\beta$ can be expressed as the linear rules:

$$w(\phi_\alpha \wedge \phi_\beta) = w(\phi_\alpha).$$

Clearly, any rule of inference can be regarded as a linear rule for some formulae. However, if too many sentences are added to $\phi_1, \ldots, \phi_n$ then the algorithm of the previous section may become impractical. We investigate the case in which the formulae Ψ are expressed in conjunctive normal form and show that if they have a small size of clauses then the number of formulae that need to be added to $\phi_1, \ldots, \phi_n$ is polynomial in n.

A formula in conjunctive normal form (CNF) of $\phi_1, \ldots, \phi_n$ is a conjunction $p_1 \wedge \cdots \wedge p_r$ of clauses, where each clause p_i is a disjunction $q_1 \vee \cdots \vee q_{j_i}$ of literals. A literal is either a sentence ϕ or the negation $\bar{\phi}$ of a sentence. A k-CNF is a CNF expression with clauses that are disjunctions of at most k literals. For example, $(\phi_1 \vee \bar{\phi}_2) \wedge (\bar{\phi}_1 \vee \bar{\phi}_2 \vee \phi_3)$ is a 3-CNF.

Theorem: Let Θ be the set of sentences that can be obtained from disjunctions of at most $k+1$ sentences from $\phi_1, \ldots, \phi_n$.

$$\Theta = \{\theta : \theta = \phi_{i_1} \wedge \cdots \wedge \phi_{i_j}, j \leq k+1\}.$$

A formula of the type

$$\Psi \rightarrow \phi_i$$

where Ψ is a k-CNF of $\phi_1, \ldots, \phi_n$ can be expressed as a linear rule of inference of sentences from Θ.

Proof: Let $c_1, \ldots, c_m$ be the clauses of Ψ:

$$\Psi = c_1 \wedge \cdots \wedge c_m.$$

This means that

$$\Psi = \mathsf{TRUE} \quad \Leftrightarrow \quad \sum_{\alpha=1}^{m} w(c_\alpha) = m,$$

and $\Psi \rightarrow \phi_i$ if and only if

$$(\sum_{\alpha=1}^{m} w(c_\alpha) - m)(w(\phi_i) - 1) = (w(\phi_i) - 1). \quad (10)$$

Each clause c_α, for $\alpha = 1, \ldots, m$ is a Boolean formula of at most k variables ϕ_i, $i \leq n$, so that $w(c_\alpha)$, can be expressed as a multilinear form of degree at most k of $w(\phi_i)$, $i \leq n$. Therefore, Equation (10) is a multilinear form of degree at most $k+1$ of $w(\phi_i)$, $i \leq n$. Since each monomial of the multilinear form is linear in formulae from Θ the rule in Equation (10) is linear in formulae from Θ. $\square$

Application

The ability to derive crisp information from probabilities is most useful in cases where probabilities can be computed easily. We have shown in [Shvaytser, 1988] how similar ideas enable learning from examples in the sense of Valiant. (The probabilities were obtained from samples of examples that correspond to possible worlds.) However, there seem to be cases in which it is more natural to have information as probabilities and not as examples.

Consider a system of n computers that are connected in a parallel architecture. From time to time the system is required to handle a problem which is distributed among $n/2$ of the computers. Let ϕ_i be the sentence: "Computer i is busy working on the problem". In this case a possible world is a world in which exactly half of the n computers are busy working on the problem.

Let $x_i = w(\phi_i)$. By introducing an additional sentence, ϕ_0, which is always TRUE, and its corresponding random variable $x_0 \equiv 1$, there are linear rules of inference since:

$$x_i = \frac{n}{2} x_0 - \sum_{j=1, j \neq i}^{n} x_j. \quad (11)$$

Now consider the case in which the system malfunctions, and we suspect that there are problems with the distribution of tasks among the computers. This can be verified by checking the condition:

$$\sum_{i=1}^{n} x_i = n/2, \quad (12)$$

but verifying this condition takes time proportional to n when checked by a single computer, and at least time proportional to $\log n$ even with many computers. Therefore, verifying the above condition for each instance of the problem may cause long delays and may not allow a verification in real time.

In this case we are not interested in a probabilistic answer such as that the condition holds "with

Table 3: distribution of instances

x_0	x_1	x_2	x_3	x_4	x_5	x_6	# instances
1	1	1	1	0	0	0	500,000
1	1	0	1	0	1	0	400,000
1	0	1	1	1	0	0	100,000

high probability". We would like to verify that *for all instances* of the problem condition (12) holds.

Since Equation (12) can be expressed as a linear rule of inference it can be inferred from probabilities that can be computed in real time. By assigning a processor to each pair of computers, the number of times in which they are both activated can be computed in a constant time. For the pair i and j this is equivalent to the probability of the formula $\phi_i \wedge \phi_j$ when scaled properly.

As a numerical example, consider the case in which $n = 6$, the number of instances is $1,000,000$, and they are given in Table 3. The correlation matrix of $x_0, \cdots, x_6$, is:

$$\frac{1}{10}\begin{pmatrix} 10 & 9 & 6 & 10 & 1 & 1 & 0 \\ 9 & 9 & 1 & 2 & 0 & 1 & 0 \\ 6 & 1 & 6 & 2 & 1 & 0 & 0 \\ 10 & 2 & 2 & 10 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Applying the algorithm we get three linear rules of inference:

$$\begin{aligned} x_3 &= x_0 \\ x_5 &= 2x_0 - x_1 - x_2 - x_4 \\ x_6 &= 0 \end{aligned}$$

and one can easily verify that they can infer anything that can be inferred from Equation (11). Furthermore, they imply Equation (12).

Conclusions

We have shown that relations between probabilities of sentences can always be used to determine linear rules of inference, whenever such rules exist. This shows that in many cases probabilities can be used to infer crisp (non-probabilistic) knowledge.

References

[Fagin and Halpern, 1988] R. Fagin and J. Y. Halpern. Reasoning about knowledge and probability. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 277–293. Morgan Kaufman, 1988.

[Herstein, 1975] I. N. Herstein. *Topics in Algebra.* John Wiley & Sons, second edition, 1975.

[Nilsson, 1986] N. J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

[Papoulis, 1984] A. Papoulis. *Probability, random Variables, and Stochastic Processes.* McGraw-Hill, second edition, 1984.

[Pitt, 1989] L. Pitt. Probabilistic inductive inference. *Journal of the ACM*, 36(2):383–433, April 1989.

[Shvaytser, 1988] H. Shvaytser. Representing knowledge in learning systems by pseudo boolean functions. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 245–259. Morgan Kaufman, 1988.