Qualitative Reasoning in Bayesian Networks

Paolo Garbolino
Scuola Normale Superiore
56126 Pisa. Italy
e-mail: Garbolino@sabsns.sns.it

Abstract

Some probabilistic inference rules which can be compared with the inference rules of preferential logic are given and it will be shown how they work in graphical models, allowing qualitative plausible reasoning in Bayesian networks.

1 Introduction

The purpose of this paper is to give a small example of the fact that "Probability is not really about numbers; it is about the structure of reasoning" (G.Shafer). B. de Finetti, in his days, was well aware of this fact, when he wrote: "il calcolo delle probabilit insegna a dedurre la maggiore o minore verosimiglianza o probabilit di certe conseguenze dalla maggiore o minore verosimiglianza o probabilit di certe premesse" [de Finetti 1930, p. 259], and forty years later he stressed that "most applications of Bayesian standpoint in everyday life, in scientific guessing, and often also in statistics, do not require any mathematical tool nor numerical evaluations of probabilities; a qualitative adjustment of beliefs to changes in the relevant information is all that may be meaningfully performed" [de Finetti 1974, p. 117]. It has taken more than forty years to fully appreciate the resources of qualitative probability, and we have also to thank the challenge offered to Bayesians by those people who have developed nonmonotonic logic and other theories of plausible reasoning. There have been two milestones in the development of a qualitative theory of probabilistic inference: the axiomatic study of the properties of the relation of conditional independence [Dawid 1979], and its application to causal graphs [Pearl 1988]; and the application of the theory of Markov random fields [Kindermann, Snell 1980] to causal graphs [Lauritzen, Spiegelhalter 1988], [Spiegelhalter, Dawid, Lauritzen, Cowell 1993], [Garbolino 1993].

The principal obstacle to the use of the axioms and the theorems of classical probability theory as rules of inference, having as their premises probabilistic inequalities, and as conclusions other probabilistic inequalities, is perhaps the idea that such a logic would have the complexity of arithmetic inequalities. This was the feeling, for example, of Pearl [Pearl 1988, p. 494]. The turning point in the implementation of probabilistic expert systems, represented by the above mentioned theoretical developments, consisted precisely in allowing a dramatic reduction of the computational complexity of probabilistic knowledge bases. The main idea of this paper is that such a reduction of complexity can also be obtained for a logic of probabilistic inequalities.

The bridge between arithmetic and logic is provided by the following result [Pearl, Geiger, Verma 1990]: if the directed acyclic graph (DAG) associated to a set of conditional independence statements (called the causal input list) satisfies the graphical condition called d-separation, then a Markov blanket is uniquely determined for any given node x of the graph, that is, a set $B(x)$ of nodes that renders x independent of all nodes not in $B(x)$. In other terms, we can say that the nodes in $B(x)$, different from x, "screen off" all the nodes not in $B(x)$ from x, and in this case we say that the Markov condition holds between x and every node not in $B(x)$. The Markov condition has a straightforward translation in the syntax of probability theory. It can be proved [Garbolino 1995a, 1995b] that the Markov condition is sufficient to derive from the axioms and theorems of probability theory a set of inference rules that show a strict parallelism with the inference rules of rational logic [Kraus, Lehmann, Magidor 1990], [Lehmann, Magidor 1992]. Happily, the members of the Markov blanket for a node x in a DAG are exactly the members of the set of neighbors of a node x in a Makov random field, yielding an equivalent probabilistic representation for directed and undirected graphs.

2 A logic of qualitative probability

Our language is a set L of well-formed propositional formulas over a finite set S of propositional variables $A, B, C, \ldots$. L is closed under the classical propositional connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$. A set of models M is associated to L, and validity is defined as usual.

Let assume that a finitely additive probability function does exist for the Boolean algebra B whose elements are the equivalence classes of L, modulo $\leftrightarrow$, i. e., it does exist a function P that assigns to each element A of the Boolean algebra B a real number $P(A)$ such that:

(2.1) $P(A) \geq 0$.

(2.2) $P(A) = 1$ if A is a valid formula.

(2.3) $P(A \vee B) = P(A) + P(B)$, if the formula $A \wedge B$ is unsatisfiable in M.

By abuse of language, equivalence classes of propositional formulas will be denoted by the same symbols $A, B, C, \ldots$, which denote propositional formulas.

The conditional probability of B given A, denoted $P(B|A)$, and the relation of probabilistic independence between A and B are defined in the usual way.

If $P(C|B \wedge A) = P(C|B \wedge \neg A)$, then B is said to screen off C from A, or that C is independent from A given B. Of course, if B screens off C from A, also $\neg B$ screens off C from A:

$$P(C|\neg B \wedge A) = P(C|\neg B \wedge \neg A).$$

If B screens off C from A, then we shall say that the Markov condition holds between A and C [Eells, Sober 1983], [Eells 1991], and:

(2.4) $P(C|A \wedge B) = P(C|B)$.

A conditional assertion of the form "if A is true, then typically B is true", denoted by $A| \sim B$, is interpreted as a conditional probability judgment:

Definition 2.1. *Let P be a finitely additive probability function on the language L. For any two formulas A and B, such that* $0 < P(A) < 1$, $0 < P(B) < 1$, *and* $P(B|A) < 1$, *then $A| \sim B$ iff:*

(2.5) $P(B|A) > P(B|\neg A)$.

From a basic fact of probability theory, (2.5) is equivalent to:

(2.6) $P(B|A) > P(B)$.

Let's make a brief comment about the proposed interpretation of the conditional assertion $A| \sim B$.

Why do we not interpret it probabilistically as $P(B|A) > P(\neg B|A)$? Because this inequality trivially implies $P(B|A) > 0.5$, and this seems too strong a requirement.

The following inference rules are valid for a finitely additive P on L (the proofs are in [Garbolino 1995b], and the rules are named after their likeness with the rules of preferential logic).

Markov Right Weakening.

If $A \rightarrow B, P(A|C) > P(A)$, and $P(B|\neg A \wedge C) = P(B|\neg A)$,

then $P(B|C) > P(B)$.

Markov And.

If $P(B|A) > P(B), P(C|A) > P(C)$, and $P(C|A \wedge B) = P(C|A)$,

then $P(B \wedge C|A) > P(B \wedge C)$.

Markov Or.

$P(C|A) > P(C), P(C|B) > P(C), P(A \wedge B) = P(A)P(B)$, and $P(A|B \wedge C) = P(A|C),$ then $P(C|A \vee B) > P(C).$

Markov S

If $P(C|A) > P(C), P(C|B) > P(C), P(A \wedge B) = P(A)P(B),$

and $P(B|A \wedge C) = P(B|C),$ then $P(B \to C|A) > P(B \to C).$

Markov Cut.

If $P(C|A \wedge B) > P(C), P(B|A) > P(B),$

and $P(C|A \wedge B) = P(C|B),$ then $P(C|A) > P(C).$

Markov Cautious Monotonicity.

If $P(B|A) > P(B), P(C|A) > P(C),$ and $P(C|A \wedge B) = P(C|A),$

then $P(C|A \wedge B) > P(C).$

The following rules have no a counterpart in preferential logic:

Markov Transitivity [Eells 1991].

If $P(B|A) > P(B), P(C|B) > P(C),$

and $P(C|A \wedge B) = P(C|B),$ then $P(C|A) > P(C).$

Markov Left Conjunction

If $P(C|A) > P(C), P(C|B) > P(C), P(B|A) = P(B),$

and $P(B|A \wedge C) = P(B|C),$ then $P(C|A \wedge B) > P(C).$

Contraposition.

If $P(B|A) > P(B),$ then $P(\neg A|\neg B) > P(\neg A).$

Common cause.

If $P(B|A) > P(B), P(C|A) > (C),$ and

$P(C|A \wedge B) = P(C|A),$ then $P(C|B) > P(C).$

As it is well known, Markov Left Conjunction and Contraposition are simply the formalization of two widely accepted principle of the logic of empirical enquiry: independent favorable observations increase the plausibility of an hypothesis, and unfavorable evidence diminishes it. The "common cause principle" [Salmon 1984], [Eells 1991] is called after the fact that, if we interpret the probabilistic inequalities as causal relations or if they are statistical correlations, it explains the "spurious" correlation observable between B and C in terms of the screening off "common cause" A.

3 A Bayesian network

The definition of the d-separation criterion mentioned above is the following [Pearl, Geiger, Verma 1990]:

Definition 3.1. *If X, Y, and Z are three disjoint subsets of nodes in a DAG, then Z is said to d- separate X from Y iff there is no path from a node in X to a node in Y (a path in a directed acyclic graph is a sequence of edges in the underlying undirected graph obtained just by ignoring the direction of the arrows) along which the following two conditions hold:*

(1) every node with converging arrows either is or has a descendant in Z;

(2) very other node is outside Z.

The following theorem [Pearl, Geiger, Verma 1990] ensures that a DAG is a good graphical representation for probabilistic dependencies.

Completeness. *For every DAG there exists a probability distribution such that for every three disjoint sets of nodes X, Y, and Z, X and Y are conditionally independent given Z iff Z d-separates X from Y.*

Let's consider now the following purely qualitative medical model. It has been greatly simplified and it is not to be understood as a realistic representation of causal relationships, but it is a well known standard example in the literature.

"Metastatic cancer (A) is a possible cause of a brain tumor (C) and is also an explanation for increased total serum calcium (B). In turns, either of these could explain a patient falling into a coma (D). A papilledema (E) is also possibly associated with a brain tumor".

The figure shows the directed acyclic graph that contains the dependencies declared in the model.

According to our definition (2.1), the probabilistic hypotheses of the model are:

(3.1) $P(B|A) > P(B)$;

(3.2) $P(C|A) > P(C)$;

(3.3) $P(D|B) > P(D)$;

(3.4) $P(D|C) > P(D)$;

(3.5) $P(D|B \wedge C) > (D)$;

(3.6) $P(E|C) > P(E)$.

The graph in the figure contains also the conditional independence relationships implied by the model, as we can verify by checking definition (3.1).

The set $\{B, C\}$ is d-separated by $\{A\}$, providing further premises for the inference machine:

(3.7) $P(B|A \wedge C) = P(B|A)$;

(3.8) $P(C|A \wedge B) = P(C|A)$.

The set $\{A, D\}$ is d-separated by $\{B, C\}$, and thus:

(3.9) $P(D|A \wedge B \wedge C) = P(D|B \wedge C)$.

The set $\{A, D, E\}$ is d-separated by $\{C\}$:

(3.10) $P(E|A \wedge C) = P(E|C)$;

(3.11) $P(E|D \wedge C) = P(E|C)$.

The set $\{B, C\}$ is not d-separated by $\{D\}$:

(3.12) $P(B|D \wedge C) \neq P(B|D \wedge \neg C)$;

(3.13) $P(C|D \wedge B) \neq P(C|D \wedge \neg B)$.

The set $\{A, D\}$ is not d-separated either by $\{B\}$ or by $\{C\}$:

(3.14) $P(D|B \wedge A) \neq P(D|B \wedge \neg A)$;

(3.15) $P(D|C \wedge A) \neq P(D|C \wedge \neg A)$.

4 Firing the rules

A Bayesian inference machine could now apply all the rules allowed by the set of premises (3.1)-(3.15). By the rule of Markov And and the premises (3.1), (3.2) and (3.8), we get:

(4.1) $P(B \wedge C|A) > P(B \wedge C)$,

By the same premises plus (3.7), and the rule of Markov Cautious Monotonicity, we obtain:

(4.2) $P(C|A \wedge B) > P(C)$;

(4.3) $P(B|A \wedge C) > P(B)$.

By the rule of Markov And, again, and the premises (3.4), (3.6) and (3.11), we obtain:

(4.4) $P(E \wedge D|C) > P(E \wedge D)$

These are not very interesting inferences, indeed, and they have been given just to show that probabilistic rules satisfy elementary inferences that we are intuitively good to do. More interesting, of course, are inferences that go beyond our intuitive skills, as it is the case for the common causes and transitivity.

By the Common Cause principle and from (3.1), (3.2), (3.7) and (3.8), we deduce:

(4.5) $P(C|B) > (C)$;

(4.6) $P(B|C) > (B)$.

By Markov Transitivity, from (3.2), (3.6) and (3.10), we get:

(4.7) $P(E|A) > P(E)$.

And Markov Transitivity, together with the premises (4.1), (3.5), and (3.9), allows us to infer:

(4.8) $P(D|A) > P(D)$.

It is worthwhile to notice that if one of the variables B and C, say C, were negatively relevant for D, that is, $P(\neg D|C) > P(\neg D)$, then we could not derive (4.8). Indeed, $P(\neg D|C) > P(\neg D)$ implies:

(4.9) $P(D|C) < P(D)$,

and we have a "Nixon Diamond" case. Premise (3.4) is false and (3.3) and (3.5) can also be false, one of them or both. As it has been proved by Eells and Sober [Eells, Sober 1983], (3.1), (3.2), (3.3), (3.4), (3.7), (3.8) and (3.9) are sufficient but not necessary conditions for transitivity. It could still hold, but purely qualitative information is not sufficient to establish it, but we have to estimate the numerical probabilities $P(D|B \wedge C)$, $P(D|B \wedge \neg C)$ and $P(D|\neg B \wedge C)$.

5 Qualitative Jeffrey's Kinematics

We have already mentioned the fact that the members of the Markov blanket of a node x in a DAG are the neighbors of the node x in the underlying undirected graph. Let's put it in a more formal way.

Definition 5.1. *The set of neighbors of a node x in the undirected graph underlying a DAG is given by the "parents" of x in the DAG, together with its "children" and all the "parents" of its "children".*

For example, in our figure the set of neighbors of A is $\{B, C\}$, the set of neighbors of B is $\{A, C, D\}$, and the set of neighbors of D is $\{B, C\}$.

As it is well known, there exists a quantitative rule that makes possible local probability computations on undirected graphs [Lauritzen, Spiegelhalter 1988], [Wen 1992]. Starting from the neighborhood system given by definition (5.1), with respect to which the graph is a Markov random field, we form the set of all cliques.

Definition 5.2. *A clique is a set C of nodes such that every pair of them belonging to C are neighbors.*

In our figure, the cliques are $\{A, B, C\}$, $\{B, C, D\}$, and $\{C, E\}$.

Let X be the set of nodes belonging to a clique C, Y a proper subset of X, and let's denote by $P(X)$ and $P(Y)$ the marginal distributions of X and Y. Suppose we come to know a new marginal $Q(Y) \neq P(Y)$. The updating rule for the clique marginal is:

(5.1) $Q(X) = P(X)[Q(Y)/P(Y)]$

The algorithm (5.1) is also known as the Jeffrey's rule [Jeffrey 1983], [Garbolino 1993]. There is a straightforward qualitative version of Jeffrey's rule that works in our probabilistic logic.

Let A and B be the propositions associated to two nodes belonging to the same clique, and suppose that the following condition holds, where P and Q are probability distributions at different times:

(5.2) $Q(B|A) = P(B|A)$.

Then, by the definition of conditional probability, we get at once the formula (5.1):

(5.3) $Q(A \wedge B) = P(A \wedge B)[Q(A)/P(A)]$.

An equivalent formulation follows from the total probabilities theorem:

(5.4) $Q(B) = Q(B|A)Q(A) + Q(B|\neg A)Q(\neg A)$.

From (5.4) and (5.2) it follows:

(5.5) $Q(B) = P(B|A)Q(A) + P(B|\neg A)Q(\neg A)$.

From the formula (5.5) the following list of inference rules can be easily proved [Garbolino 1995b]:

Theorem 5.1.

(1) If $Q(A) > P(A)$, and $P(B|A) > P(B)$, then $Q(B) > P(B)$.

(2) If $Q(A) < P(A)$, and $P(B|A) > P(B)$, then $Q(B) < P(B)$.

(3) If $Q(A) > P(A)$, and $P(B|A) < P(B)$, then $Q(B) < P(B)$.

(4) If $Q(A) < P(A)$, and $P(B|A) < P(B)$, then $Q(B) > P(B)$.

In words: "if A becomes more (less) plausible, and B is a plausible consequence of A, then B becomes more (less) plausible". As an example, let's suppose that:

(5.6) $Q(E) > P(E)$.

By (3.6), and Bayes theorem written under the form:

(5.7) $P(C|E)P(E) = P(E|C)P(C)$,

it follows that:

(5.8) $P(C|E) > P(C)$.

Therefore, from (5.6), (5.8), and Theorem (5.1)-(1), we obtain:

(5.9) $Q(C) > P(C)$.

A cascade of inferences follows by a mechanical application of Bayes theorem and Jeffrey's rule. From (5.9) and (3.2), we get:

(5.10) $Q(A) > P(A)$.

From (5.10) and (4.1):

(5.11) $Q(B \wedge C) > P(B \wedge C)$.

From (5.11) and (3.5), one finally infers:

(5.12) $Q(D) > P(D)$.

The numerical propagation algorithm for Bayesian networks provides a coherent qualitative inference mechanism as well.

"Multiple extension" problems can arise, of course, and they normally will. Take the case mentioned at the end of preceding section: if C is negatively relevant for D then, applying the updating rule to the path $A \rightarrow B \rightarrow D$ will yield $Q(D) > P(D)$, whereas the same rule applied to the paht $A \rightarrow C \rightarrow D$ will yield $Q(D) < P(D)$. But this is not a problem of the inference rule: as it has been said before, we need more information in order to be able to decide which one of the two conclusions holds.

Inference machines are like physical machines, and information is like energy: we cannot have informative outputs if we don't have informative enough inputs.

6 Conclusion: learning new propositions.

Grdenfors and Makinson have recently investigated nonmonotonic logics with an approach similar to that one has inspired our work [Grdenfors, Makinson 1994]. Their key idea is to use a belief ordering of sentences to determine when A nonmonotonically implies B. This ordering is weaker than the probability ordering ("B is at least as probable as A") used in the axiomatizations of probability la Savage

[Savage 1954], and it is incompatible with it. Indeed, Grdenfors and Makinson's "belief valuations" share, as they themselves have pointed out, the basic properties of Shafer's "consonant belief functions" [Shafer 1976] and are closely related to "Possibility measures" [Dubois and Prade 1988]. There is also a formal relationship between belief revisions based on orderings of "epistemic entrenchment" [Grdenfors 1988] and nonmonotonic inference relations based on "belief valuations" [Makinson, Grdenfors 1991]. They identify three kinds of change, in the dynamics of belief states [Grdenfors, Makinson 1994]:

(i) Expansion, when a new sentence is added to a given belief set.

(ii) Revision, when a new sentence which is inconsistent with the given belief set is added.

(iii) Contraction, when some sentence is retracted without adding any new sentence.

If the belief set is a probabilistically coherent set, that is, it can be graphically represented by a Bayesian network, the third case only can be dealt with without re-inizializing all the network, for it corresponds to Bayesian conditionalization:

(6.1) $P(A) > 0 \rightarrow Q(A) = 0$, for some proposition A.

Expansion and revision, on the contrary, both require introducing a new proposition in the set (a new node in the graph), and in this case neither Bayes conditionalization nor Jeffrey conditionalization are possible without previously establishing the probabilistic dependencies among the new node and the old ones. The modularity of Bayesian networks allows to minimize re- inizialization, in the sense that we have to assessing the marginals for those cliques only in which the new node appears, and then the updating of the remaining part of the network can be carried on as usually.

References

Dawid, A. P. [1979] "Conditional independence in statistical theory", J. Roy. Statist. Soc., B 41, 1- 31.

de Finetti, B. [1930] "Fondamenti logici del ragionamento probabilistico", Boll. Un. Mat. Ital., 9, 258-261.

de Finetti, B. [1974] "Bayesianism: its unifying role for both the foundations and applications of statistics", Internat. Statist. Rev., 42, 117-130.

Dubois, D., Prade, H. [1988] Possibility Theory: An Approach to Computerized Processing of Uncertainty, Plenum Press, New York.

Eells, E. [1991] Probabilistic Causality, Cambridge University Press, Cambridge.

Eells, E., Sober, E. [1983] "Probabilistic causality and the question of transitivity", Philosophy of Science, 50, 37-57.

Garbolino, P. [1993] “Bayes machines”, in: Scozzafava, R. (ed.), Probabilistic methods in expert systems, Soc. Ital. Statist., Roma, 105-120.

Garbolino, P. [1995a] “Nonmonotonic logic: Much ado about nothing?”, Communication at 10th International Congress of Logic, Methodology and Philosophy of Science, Florence, August 19-25, 1995.

Garbolino, P. [1995b] “Nonmonotonic Probabilistic Logic”, forthcoming in Annali Univ. Ferrara, Sez. III, Filosofia, Discussion Paper.

Grdenfors, P.[1988] Knowledge in Flux: Modelling the Dynamics of Epistemic States, MIT Press, Cambridge MA.

Grdenfors, P., Makinson, D. [1994] “Nonmonotonic inference based on expectations”, Artificial Intelligence, 65, 197-245.

Jeffrey, R. C. [1983] The Logic of Decision, Chicago University Press, Chicago.

Kindermann, R., Snell, J. L., [1980] Markov Random Fields and their Applications, Am. Math. Soc., Providence.

Kraus, S., Lehmann, D., Magidor, M. [1990] “Nonmonotonic reasoning, preferential models and cumulative logics”, Artificial Intelligence, 44, 167-207.

Makinson, D., Grdenfors, P. [1991] “Relations between the logic of theory change and mommonotonic logic”, in: Fuhrmann, A., Morreau, M. (eds.), The Logic of Theory Change, Lecture Notes in Art. Intell., 465, Springer, Berlin, 185-205.

Lauritzen, S. L., Spiegelhalter, D. J. [1988] “Local computations with probabilities on graphical structures and their application in expert systems”, J. Roy. Statist. Soc., B 50, 157-224

Lehmann, D., Magidor, M. [1992] “What does a conditional knowledge base entail?”, Artificial Intelligence, 55, 1-60.

Pearl, J. [1988] Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Mateo.

Pearl, J., Geiger, D., Verma, T. [1990] “Conditional independence and its representations”, in: Pearl, J., Shafer, G. (eds.), Readings in Uncertain Reasoning, Morgan Kauffman, San Mateo, 55-60.

Salmon, W. C. [1984] Scientific Explanation and the Causal Strutcure of the World, Princeton University Press, Princeton.

Savage, L. J. [1954] The Foundations of Statistics, Wiley, New York.

Shafer, G. [1976] A Mathematical Theory of Evidence, Princeton University Press, Princeton.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., Cowell, R. G. [1993] “Bayesian analysis in expert systems”, Statist. Sc., 8, 219-283.

Wen, W. X. [1992] “Parallel distributed belief networks that learn”, Proc. IJCAI 92, vol. 2, 1210- 1215.