BAYES DECISION METHODS

Judea Pearl

June 1985
CSD-850023

BAYES DECISION METHODS

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California
Los Angeles, CA 90024
(judea@UCLA-locus)
(judea@LOCUS.UCLA.EDU)

Written for Wiley's forthcoming "Encyclopedia of AI"

*This work was supported in part by the National Science Foundation, Grant #DSR 83-13875

(Draft -- written for Wiley's "Encyclopedia of AI")

BAYES DECISION METHODS

by

Judea Pearl

Basic Formulation

Bayes methods provide a formalism for reasoning about partial beliefs under conditions of uncertainty. In this formalism propositions are quantified with numerical parameters signifying the degree of belief accorded them by some body of knowledge, and these parameters are combined and manipulated according to the rules of probability theory. For example, if A stands for the statement "Reagan will seek re-election in 1988", then $P(A|K)$ stands for a person's subjective belief in A given a body of knowledge K which may include that person's assumptions about American politics, specific proclamations made by the White House, an assessment of Reagan's age, personality, and so on. The symbol K, indicating the source of the belief in A, is often suppressed from belief expressions, and we simply write $P(A)$ or $P(\sim A)$. This is justified when K remains constant, since the main purpose of the quantifier P is to *summarize* K without explicating its details. However, when this background information undergoes changes, we specifically need to identify which assumptions account for our beliefs, and an explicit mentioning of K or some of its elements is then required.

In Bayes formalism, belief statements obey the three basic assumptions of probability theory:

1. $0 \leq P(A) \leq 1$ (1)
2. $P(\text{sure proposition}) = 1$ (2)

3. $P(A \text{ or } B) = P(A) + P(B)$ if A and B are incompatible. (3)

Thus, a proposition and its negation must be assigned a total belief of unity,

$$P(\sim A) = 1 - P(A) \tag{4}$$

to account for the fact that one of the two is certain to be true.

The heart of Bayesian techniques lies in the celebrated inversion formula:

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)} \tag{5}$$

stating that the belief we accord a hypothesis H upon obtaining evidence e can be computed by multiplying together our prior belief $P(H)$ and the likelihood $P(e|H)$ that e will materialize assuming that H is true. The denominator $P(e)$ of (5) hardly enters into consideration because it is independent on H and can always be computed by requiring that $P(H|e)$ and $P(\sim H|e)$ sum to unity.

Whereas a formal mathematician will dismiss (5) as a straight-forward identity stemming from the definition of conditional probabilities:

$$P(A|B) = \frac{P(A, B)}{P(B)} \quad , \quad P(B|A) = \frac{P(A, B)}{P(B)} \tag{6}$$

the Bayesian subjectivist regards (5) as a normative rule for updating beliefs in response to evidence. The left-hand side of (5) expresses a quantity $P(H|e)$ that people often find hard to assess, in terms of more readily judged quantities, often available directly from the natural encodings of our experiential knowledge. For example, if a person at the next gambling table declares an outcome "TWELVE" and we wish to know whether he was rolling a pair of dice or turning a roulette wheel, the quantities P *(TWELVE | dice)* and P *(TWELVE | roulette)* are readily known from our model of the

gambling devices (giving 1/36 to the former and 1/51 for the latter). Similarly, we can judge the *prior* probabilities, P *(dice)* and P *(roulette)*, by estimating the number of roulette wheels and dice-rolling tables at the gambling casino. However, issuing a direct judgment of P *(dice* | *TWELVE)* is a much harder mental task, which could not be performed reliably unless one becomes a specialist of such guesses, at the very same casino establishment.

Combining Prospective and Retrospective Supports

The essence of the rule in (5) is conveniently portrayed using the *odds* and *likelihood-ratio* parameters. Dividing (5) by the complementary form for $P(\sim H|e)$, we obtain:

$$\frac{P(H|e)}{P(\sim H|e)} = \frac{P(e|H)}{P(e|\sim H)} \frac{P(H)}{P(\sim H)} \tag{7}$$

Defining the *prior odds* on H to be

$$O(H) = \frac{P(H)}{P(\sim H)} = \frac{P(H)}{1-P(H)} \tag{8}$$

and the *likelihood ratio by*

$$L(e|H) = \frac{P(e|H)}{P(e|\sim H)}, \tag{9}$$

we see that the *posterior odds*

$$O(H|e) = \frac{P(H|e)}{P(\sim H|e)} \tag{10}$$

is given by the product:

$$O(H|e) = L(e|H)O(H) \tag{11}$$

Thus, Bayes rule dictates that the overall strength of belief in a hypothesis H, based on both our previous knowledge K and a given evidence e, should be the *product* of two factors: the prior odds $O(H)$ and the likelihood ratio $L(e|H)$. The former measures the

causal or *prospective* support accorded to H by the background knowledge alone, while the latter represents the *diagnostic* or *retrospective* support given to H by the evidence actually observed.

Strictly speaking, the likelihood ratio $L(e|H)$ may also depend on other propositions in the tacit knowledge base K. However, we shall later see that the power of Bayes techniques comes primarily from the fact that in causal reasoning the relation $P(e|H)$ is fairly local; namely, given that H is true the probability of e can be estimated fairly naturally and it is not dependent on many other propositions in the data base. For example, once we establish that a patient suffers from a given disease, it is fairly natural to estimate the probability that he will develop a certain symptom. This is what physicians learn in medical schools; it is considered a stable characteristic of the disease and, therefore, should be fairly independent of other factors such as epidemic conditions, previous diseases, the tests that help identify the disease, and so on. It is for this reason that the conditional probabilities $P(e|H)$ can meet the modularity requirements of rule-based expert systems in that it can serve to quantify our confidence in rules such as "if H then e", and retain its viability regardless of other rules or facts that may reside in the knowledge base at any given time.

Ex. 1. Imagine being awakened one night to the shrill sound of your burglar alarm system. What would be your degree of belief that a burglary attempt has taken place? For illustrative purposes we make the following judgments: (a) There is a 95% chance that an attempted burglary will trigger the alarm system, $P(Alarm \mid Burglary) = .95$; (b) There is slight (.01) chance that the alarm sound would be triggered by a mechanism other than an attempted burglary; thus, $P(Alarm \mid No\ Burglary) = .01$; (c) Previous crime patterns indicate that there is a one-in-ten-thousands chance that a given house will be burglarized on any given night; i.e., $P(Burglary) = 10^{-4}$.

Putting these assumptions together we obtain, using (5),

$$O\,(Burglary|Alarm) = L\,(Alarm|Burglary)\;O\,(Burglary)$$

$$= \frac{.95}{.01}\,\frac{10^{-4}}{1-10^{-4}} = .0095$$

and so, from

$$P\,(A) = \frac{O(A)}{1 + O(A)} \tag{12}$$

we have:

$$P(\text{Burglary} \mid \text{Alarm}) = \frac{.0095}{1+.0095} = .00941$$

Thus, the retrospective support imparted to the burglary hypothesis by the alarm evidence, has increased its degree of belief from 1 in ten-thousands to 94.1 in ten thousands. Notice that it was not necessary to estimate the absolute values of the probabilities P *(Alarm | Burglary)* and P *(Alarm | No Burglary),* only their ratio enters the calculation and, therefore, a direct estimate of this ratio could have been used instead.

Pooling of Evidence

Assume that the alarm system consists of not one, but a collection of N burglary detection devices, each sensitive to a different physical mechanism (e.g. air turbulences, temperatures variation, pressure, sound, etc.) and each equipped with its own distinctive sound of alarm.

Let H stand for the event that a burglary took place and let e^k stand for the evidence obtained from the k^{th} detector, with e_1^k representing an activated detector and e_0^k representing a silent detector. The reliability (and sensitivity) of each detector is characterized by the probabilities $P(e_1^k|H)$ and $P(e_1^k|\sim H)$ or, more parsimoniously, by their ratio:

$$L(e^k|H) = \frac{P(e_1^k|H)}{P(e_1^k|\sim H)} \tag{13}$$

If some detectors are triggered while others remain deactivated, we have ***conflicting evidence*** on our hands, and the combined belief in the hypothesis H would be computed by equation (11):

$$O(H|e^1,e^2, \cdots e^N) = L(e^1, ..., e^N|H)O(H) \tag{14}$$

Strictly speaking, equation (14) requires the use of an enormous data base, because we need to specify the probabilities of activation for every subset of detectors, conditioned on H and on $\sim H$. Fortunately, reasonable assumptions of independence can drastically cut this storage requirement. Assuming that the state of activation of each detector depends only on whether a burglarly took place, but is thereafter independent of the activation of other detectors, we can write:

$$P(e^1,e^2, \cdots e^N|H) = \prod_{k=1}^{N} P(e^k|H) \tag{15}$$

and

$$P(e^1,e^2, ..., e^N|\sim H) = \prod_{k=1}^{N} P(e^k|\sim H) \tag{16}$$

which lead to

$$O(H|e^1, ..., e^N) = [\prod_{k=1}^{N} L(e^k|H)]O(H) \tag{17}$$

Thus, the individual characteristics of each detector are sufficient for determining the combined impact of any group of detectors.

Multihypothesis Variables

The assumptions of conditional independence in equations (15) and (16) will be justified if the failure of a detector to react to an attempted burglary and the factors which may cause it to fire prematurely both depend solely on mechanisms intrinsic to

the individual detection systems such as insufficient sensitivity, internal noise, and so on. However, if these may be caused by external circumstances affecting a selected group of sensors, such as a power failure or an earthquake, the two hypotheses H = Burglary and $\sim H$ = No Burglary may be too coarse to induce sensors' independence, and additional refinement of the hypotheses space may be necessary. This usually happens when the negation of a proposition entails several possible states of the world, each having its own distinct characteristics. For example, the state of "No Burglary" entails the possibilities of an "ordinary peaceful night", a "night with earthquake", "an attempted entry by the neighbor's dog", each influencing the sensors present in a unique way. Equation (16) may be justified with respect to each one of these conditions, but not with respect to their aggregate "No Burglary". For this reason it is often necessary to refine the hypotheses space beyond that of binary-propositions and group them into multi-valued ***variables,*** where each variable consists of a set of exhaustive and mutually exclusive hypotheses.

Ex. 2. We may choose to assign the variable name $H = \{H_1, H_2, H_3, H_4\}$ to the following set of conditions:

H_1 = no burglary, equipment malfunction $(\sim b, m)$
H_2 = attempted burglary, no malfunction $(b, \sim m)$
H_3 = attempted burglary combined with equipment malfunction (b, m)
H_4 = no Burglary, no malfunction $(\sim b, \sim m)$

Each evidence variable e^k can also be multi-valued (e.g. e_1^k=no sound, e_2^k=low sound, e_3^k=high sound) and, so, the causal link between H and e^k will be quantified by an $m \times n$ matrix where m and n are the number of values, respectively, that H and e^k may take, and the $(i, j)^{th}$ entry of M^k stands for

$$M_{ij}^k = P(e_j^k|H_i) \tag{18}$$

For example, the matrix below could represent the various sensitivities of the k^{th} detector to the four conditions in H:

	e_1^k (no sound)	e_2^k (low sound)	e_3^k (high sound)
H_1	.5	.4	.1
H_2	.06	.5	.44
H_3	.5	.1	.4
H_4	1	0	0

Given a set of evidence readings $e^1, e^2, \cdots e^k, \cdots e^N$, the overall belief in the i^{th} hypothesis is given by (4),

$$P(H_i|e^1, ..., e^N) = \alpha P(e^1, \cdots e^N|H_i)P(H_i) \tag{19}$$

where α is a normalizing constant, and assuming conditional independence with respect to each H_i, we obtain

$$P(H_i|e^1, \cdots, e^N) = \alpha[\prod_{k=1}^{N} P(e^k|H_i)]P(H_i) \tag{20}$$

Thus, we see that the matrices $P(e^k|H_i)$ now play the role of the likelihood ratios in equation (17). If for each detector reading e^k we define the *likelihood vector*

$$\lambda^k = (\lambda_1^k, \lambda_2^k, ... \lambda_m^k) \tag{21}$$

$$\lambda_i^k = P(e^k|H_i) \tag{22}$$

then (20) is computed by a simple vector-product process. First, the individual likelihood vectors are multiplied together, term by term, to form an overall likelihood vector $\Lambda = \lambda^1 \odot \cdots \odot \lambda^N$, where

$$\Lambda_i = \prod_{k=1}^{N} P(e^k|H_i) \tag{23}$$

Then we obtain the overall belief vector $P(H_i|e^1, \cdots e^N)$ by the product

$$P(H_i|e^1, \cdots e^N) = \alpha\Lambda_i P(H_i) \qquad (24)$$

reminiscent of equation (17).

Note that only the *relative magnitude of* the conditional probabilities in (22) need be estimated; their absolute magnitude does not affect the final result because α is to be determined by the requirement $\sum_i P(H_i|e^1, \cdots e^N) = 1$

Ex. 3. Let us assume that our system contains two detectors having identical characteristics, given by the matrix above. Further, let our prior probabilities for the hypothesis in Example 2 be represented by the vector $P(H) = \{.099, .009, .001, .891\}$ and assume that detector-1 was heard to issue a high sound while detector-2 remained silent. From (22) we have

$$\lambda^1 = (.1, .44, .4, 0) \quad , \quad \lambda^2 = (.5, .06, .5, 1)$$

$$\Lambda = \lambda^1 \odot \lambda^2 = (.05, .026, .2, 0)$$

$$P(H_i|e^1, e^2) = \alpha(4.95, .238, .20, 0)10^{-3} = (.919, .0439, .0375, 0)$$

from which we conclude that the chances of attempted burglary (H_2 or H_3) is .0439 + .0375 = 8.14%.

The updating of belief need not wait, of course, until all the evidence is collected, but can be carried out incrementally. For example, if we first observe e^1 = high sound, our belief in H calculates to

$$P(H_i|e^1) = \alpha\,(.0099, .00396, .0004, 0) = (.694, .277, .028, 0)$$

This now serves as a prior belief with respect to the next datum and, after observing e^2 = no sound, it updates to:

$$P(H_i|e^1, e^2) = \alpha'\lambda_i^2 \cdot P(H_i|e^1) = \alpha'(.347, .0166, .014, 0) = (.919, .0439, .0375, 0),$$

as before. Thus, the quiescent state of detector-2 lowers the chances of an attempted burglary from 30.5% to 8.14%.

Uncertain evidence (cascaded inference)

One often hears the claims that Bayes techniques cannot handle uncertain evidence

because the relation $P(A|B)$ requires that the conditioning event B is known with certainty. To see the difficulties that led to this myth let us make a slight modification in the story of the alarm system:

Ex. 4. Mr. Holmes receives a telephone call from his neighbor Dr. Watson, stating that he hears a burglar alarm sound from the direction of Mr. Holmes's house. Preparing to rush home, Mr. Holmes recalls that Dr. Watson is known to be a tasteless practical joker and, therefore, he decides to first call his other neighbor, Mrs. Gibbons, who, in spite of occasional drinking problems, is far more reliable.

Since the evidence variable S= ***"sensor output"*** is now uncertain, we cannot use it as evidence in equation (11) but, rather, apply equation (11) to the actual evidence at hand: W = ***Mr. Watson's testimony***, and write:

$$P(H|W) = L(W|H)O(H) \tag{25}$$

Unfortunately, the task of estimating $L(W|H)$ will not be as easy as that of $L(S|H)$ because the former requires the mental tracing of a two-step process, as shown in Figure 1.

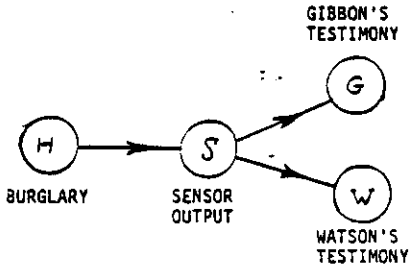

Figure 1

Moreover, even if we could obtain $L(W|H)$ we will not be able to combine it with other possible testimonies, say Mrs. Gibbon's (G), by a simple process of multiplication (23) because those testimonies will no longer be conditionally independent with respect to H. What Mrs. Gibbon is about to say may prove conclusively that Watson's phone call originated from an honest neighborly concern (i.e. S = alarm sound) and, so, it would be wrong to assume $P(W|\text{Burglary}, G) = P(W|\text{Burglary})$ because the r.h.s. also entails the

possibility that Watson makes a prank phone call under the condition of a silent burglary. Given the level of detail used in our story, it is more reasonable to assume that the testimonies W and G are independent on each other once we know whether the alarm sensor was actually triggered. In other words, each testimony depends directly on the alarm system (S), and is only indirectly influenced by the possible occurrence of a burglary (H) or by the other testimony (see Figure 1).

These considerations can be easily incorporated into Bayes formalism; we simply condition equation (19) on the intermediate variable S and obtain

$$\begin{aligned} P(H_i|G, W) &= \alpha P(G, W|H_i)P(H_i) \\ &= \alpha P(H_i)\sum_j P(G, W|H_i, S_j)P(S_j|H_i) \end{aligned} \tag{26}$$

where S_j, j=1, 2 stand for the two possible activation states of the alarm system, namely, S_1 = "alarm triggered", S_2 = "alarm not triggered". Moreover, since G, W and H_i only influence each other via the mediating variable S we can write

$$P(G, W|H_i, S_j) = P(G|S_j)P(W|S_j) \tag{27}$$

and (26) becomes

$$P(H_i|G, W) = \alpha P(H_i)\sum_j P(G|S_j)P(W|S_j)P(S_j|H_i) \tag{28}$$

The computation in (28) can be interpreted as a three state process: first, the local likelihood vectors $P(G|S_j)$ and $P(W|S_j)$ are multiplied together, term by term, to obtain the likelihood vector $\Lambda_j(S) = P(e|S_j)$, where e stands for the total evidence collected, G and W. Second, the vector $P(e|S_i)$ is multiplied by the link matrix $M_{ij} = P(S_j|H_i)$ to form the likelihood vector of the top hypothesis $\Lambda_i(H) = P(e|H_i)$. Finally, using the product rule of (4) (see also (19) or (24)), $\Lambda_i(H)$ is multiplied by the prior $P(H_i)$ to give the

overall belief in H_i.

This process demonstrates the psychological and computational role of the mediating variable S. It permits us to use local chunks of information taken from diverse domains (e.g. $P(H_i), P(G|S_j), P(W|S_j), P(S_j|H_i)$), and fit them together to form a global, cross-domain inference $P(H|e)$ in stages, using simple and local vector operations. It is this role which prompted some philosophers to posit that conditional independence is not an accident of nature for which we must passively wait but rather, a psychological necessity which we actively dictate, e.g., by coining names to new, hypothetical variables, as the need develops. In medical diagnosis, for example, when some symptoms directly influence each other, the medical profession ***invents*** a name for that interaction (e.g. complication, pathological state, etc.) and treats it as a new auxiliary variable which induces conditional independence; knowing the exact state of the auxiliary variable renders the interacting symptoms independent of each other.

Implicit (intangible) Evidence:

Let us imagine the following development in the story of Mr. Holmes:

Ex. 5. When Mr. Holmes called Mrs. Gibbons, he soon realized that she was in a somewhat tipsy mood. Instead of answering his question directly, she went on and on describing her latest operation and how terribly noisy and crime-ridden the neighborhood had become. As he finally hung up, all Mr. Holmes could make out of the conversation was that there probably was an 80% chance that Mrs. Gibbons did hear an alarm sound from her window.

The Holmes-Gibbons conversation is the kind of evidence that is hard to fit into any formalism. If we try to estimate the probability $P(e|\text{Alarm Sound})$ we would get ridiculous numbers because it would entail anticipating, describing, and assigning probabilities to all possible courses Mrs. Gibbons's conversation might have taken under the

circumstances.

These difficulties arise whenever the task of gathering evidence is delegated to autonomous interpreters who, for various reasons, cannot explicate their interpretive process in full detail but, nevertheless, often produce informative conclusions that summarize the evidence observed. In our case, Mr. Holmes's conclusion is that, on the basis of his judgmental interpretation of Gibbons's testimony (alone!), the hypothesis "alarm sound" should be accorded a confidence measure of 80%. Our task is to integrate this probabilistic judgment into the body of hard evidence previously collected.

In Bayes formalism the integration of implicit evidence is straightforward. Although the evidence e cannot be articulated in full detail, we interpret the probabilistic conclusion to convey likelihood-ratio information. In our story, for example, identifying e with G = Gibbons's testimony, Mr. Holmes's summary of attributing 80% credibility to the "Alarm Sound" event, will be interpreted as the statement $P(G|\text{Alarm Sound}) : P(G|\text{No Alarm Sound}) = 4{:}1$. More generally, if the variable upon which the tacit evidence e impinges most directly has several possible states $S_1, S_2, \cdots S_i, \cdots$ we would instruct the interpreter to estimate the relative magnitudes of the terms $P(e|S_i)$ (e.g. by eliciting estimates of the ratios $P(e|S_i) : P(e|S_1)$) and, since the absolute magnitudes do not affect the calculations, we can proceed to update beliefs as if this likelihood vector originated from an ordinary, logically crisp event e. For example, assuming that Mr. Watson's phone call already contributed a likelihood ratio of 9:1 in favor of the hypothesis "alarm sound", the combined weight of Watson's and Gibbons's testimonies would yield a likelihood vector $\Lambda_i(S) = P(W, G|S_i) = (36, 1)$.

We can now integrate this vector into the computation of equation (28) and, using the numbers given in example 1, we get

$$\Lambda_i(H) = \sum_j \Lambda_j(S) P(S_j|H_i) = \begin{pmatrix} .95 & .05 \\ .01 & .99 \end{pmatrix} \begin{pmatrix} 36 \\ 1 \end{pmatrix} = \begin{pmatrix} 34.25 \\ 1.35 \end{pmatrix}$$

$$P(H_i|G, W) = \alpha \Lambda_i(H)\, P(H_i) = \alpha(34.25, 1.35) \odot (10^{-4}, 1-10^{-4}) = (.00253, .99747)$$

Note that it is important to verify that Mr. Holmes 80% summarization is indeed based ***only*** on Mrs. Gibbons's testimony and does not include prejudicial beliefs borrowed from previous evidence (e.g. Watson's testimony or crime-rate information) or else, we stand the danger of counting the same information twice. The likelihood ratio is, indeed, unaffected by such information. Bayesian practitioners claim that people are capable of retracing the origins of their beliefs by answering hypothetical questions such as "What if you didn't receive Watson's call?" or "estimate the ***increment*** increase in belief due to Gibbons's testimony alone".

An effective way of eliciting pure likelihood-ratio estimates, unaffected by previous information, would be to first imagine that prior to obtaining the evidence, we are in the standard state of total ignorance, then to estimate the final degree of belief captured by a proposition as a result of observing the evidence. In our example, if prior to conversing with Mrs. Gibbons Mr. Holmes had a "neutral" belief in S, i.e., $P(alarm) = P(noalarm) = 1/2$, then the post-conversation estimate $P(alarm|G) = 80\%$ would indeed correspond to a likelihood ratio of 4:1 in favor of "alarm."

Predicting Future Events

One of the attractive features of causal models in the Bayes formulation is the ease by which they facilitate the prediction of yet unobserved events such as the possible developments of social episodes, the outcomes of a given test, the prognosis of a given disease and so on. The need to facilitate such predictive tasks may, in fact, be the very reason that human beings have adopted causal schema for encoding experiential knowledge.

Ex. 6. Immediately after his conversation with Mrs. Gibbons, as Mr. Holmes is preparing to leave his office, he recalls that his daughter is due to arrive home any minute and, if confronted by an alarm sound, would probably (.7) phone him for instructions. Now he wonders whether he should not wait a few more minutes in case she calls.

To estimate the likelihood of our new target event: D = "Daughter will Call", we have to add a new causal link to the graph of Fig. 1 and, assuming that hearing an alarm sound is the only reason that may induce the daughter to call, that link should emanate from the variable S, and be quantified by the following $P(D|S)$ matrix:

		D	$\sim D$
		will call	will not call
S	on	.7	.3
	off	0	1

Accordingly, to compute $P(D|\text{all evidence})$ we write

$$P(D|e) = \sum_j P(D|S_j, e)P(S_j|e) = \sum_j P(D|S_j)P(S_j|e)$$

which means that all the lengthy episodes with Mr. Watson and Mrs. Gibbons impart their influence on D only via the belief they induced on S, $P(S_j|e)$.

It is instructive to see now how $P(S_j|e)$ can be obtained from the previous calculation of $P(H_i|e)$. A natural temptation would be to use the updated belief $P(H_i|e)$ and the link matrix $P(S_j|H_i)$ and habitually write the conditioning equation

$$P(S_j|e) = \sum_i P(S_j|H_i)P(S_j|e)$$

also known as Jeffrey's rule of updating [1]. This equation, however, is only valid in a very special set of circumstances. It will be wrong in our example because the changes in the belief of H actually originated from the corresponding changes in S, and so, reflecting these back to S would amount to counting the same evidence twice. Formally, this objection is reflected by the inequality $P(S_j|H_i) \neq P(S_j|H_i, e)$, stating that the evidence obtained affects, not only the belief in H and S, but also the strength of the causal link between H and S. On the surface, this realization may seem detrimental to the usefulness of Bayes methods in handling a large number of facts; having to calculate all links parameters each time a new piece of evidence arrives would be an insurmountable computational burden. Fortunately, however, there is a simple way of updating beliefs which circumvents this difficulty and uses only the original link matrices [2]. The calculation of $P(S_j|e)$, for instance, can be performed as follows. Treating S as an intermediate hypothesis, Equation (4) dictates

$$P(S_j|e) = \alpha P(e|S_j)P(S_j)$$

The term $P(e|S_j)$ is the likelihood vector $\Lambda_j(S)$ which was calculated earlier to (36, 1), while the prior $P(S_j)$ is given by the matrix multiplication

$$P(S_j) = \sum_i P(S_j|H_i)P(H_i) = (10^{-4}, 1-10^{-4}) \begin{pmatrix} .95 & .01 \\ .01 & .99 \end{pmatrix} = (.0101, .9899)$$

Thus, together, we have

$$P(S_j|e) = \alpha(36, 1) \odot (.0101, .9899) = (.2686, .7394)$$

which gives the event S_1 = "Alarm Sound On" a credibility of 26.86%, and predicts that the event D = "Daughter will call" with the probability of

$$P(D|e) = \sum_j P(D|S_i)P(S_j|e) = (.2686, .7314)\begin{pmatrix} .7 \\ 0 \end{pmatrix} = .188$$

Multiple Causes

Tree structures, such as that used in the preceding section, require that only one variable be considered a cause of any other variable. This structure simplifies computations but its representational power is rather limited, because it forces us to group together all causal factors sharing a common consequence into a single node. By contrast, when people associate a given observation with multiple potential causes, they weigh one causal factor against another as independent variables, each pointing to a specialized area of knowledge. As an illustration, consider the following situation:

Ex. 7. As he is pondering this question, Mr. Holmes remembers having read in the instruction manual of his alarm system that the device is sensitive to earthquakes and can be triggered (.2) by one accidentally. He realizes that if an earthquake had occurred, it would surely (.9) be on the news. So, he turns on his radio and waits around for either an announcement or a call from his daughter.

Mr. Holmes perceives two episodes which may be potential causes for the alarm sound, an attempted burglary and an earthquake. Even though burglaries can safely be assumed independent of earthquakes, still the radio announcement reduces the likelihood of a burglary, as it "explains away" the alarm sound. Moreover, the two causal events are perceived as individual variables (see Figure 2); general knowledge about earthquakes rarely intersects knowledge about burglaries.

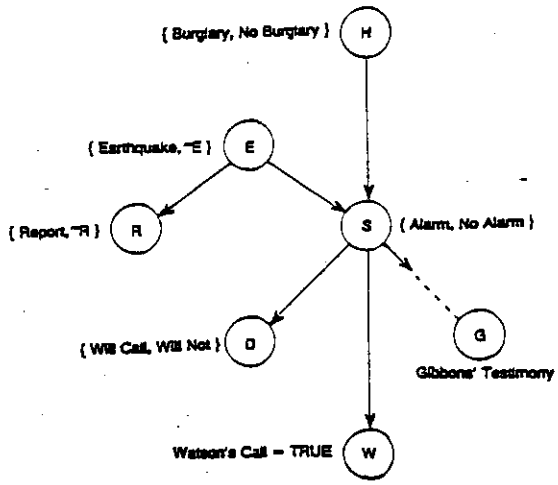

Figure 2

This interaction among multiple causes is a prevailing pattern of human reasoning. When a physician discovers evidence in favor of one disease, it reduces the credibility of other diseases, although the patient may as well be suffering from two or more disorders simultaneously. A suspect who provides an alternative explanation for being present at the scene of the crime appears less likely to be guilty, even though the explanation furnished does not preclude his committing the crime.

To model this "sideways" interaction a matrix M should be assessed giving the distribution of the consequence variable as a function of every possible combination of the causal variables. In our example, we should specify $M = P(S|E,H)$ where E stands for the variable $E = \{earthquake,\ no\ earthquake\}$. Although this matrix is identical in form to the one described in Ex. 2, Eq.(18), where the two causal variables were combined into one compound variable $\{H_1, H_2, H_3, H_4\}$, treating E and H as two separate entities has an advantage in that it allows us to relate each of them to a separate set of evidence without consulting the other. For example, we can quantify the relation between E and R (the radio announcement) by the probabilities $P(R|E)$ without having to consider the irrelevant event of burglary, as would be required by compounding the

pair (E,R) into one variable. Moreover, having received a confirmation of R, we can update the beliefs of E and R in two separate steps, mediated by updating S, closely resembling the process used by people. An updating scheme for networks with multiple-parent nodes is described in [3].

If the number of causal factors k is large, estimating M may be troublesome because, in principle, it requires a table of size 2^{k+1}. In practice, however, people conceptualize causal relationships by creating hierarchies of small clusters of variables and, moreover, the interactions among the factors in each cluster are normally perceived to fall into one of few prestored, prototypical structures, each requiring about k parameters. Common examples of such prototypical structures are: noisy OR gates (i.e., any one of the factors is likely to trigger the effect), noisy AND gates, and various enabling mechanisms (i.e., factors identified as having no influence of their own except enabling other influences to become effective).

Bayesian Networks

In the preceding discussion we have resorted quite often to the use of diagrams such as those in Figures 1 and 2. These diagrams were not used merely for mnemonic or illustrative purposes. They, in fact, convey important conceptual information, far more meaningful than the numerical estimates of the probabilities involved. The formal properties of such diagrams, called ***Bayesian Networks*** [4], will be discussed next.

Bayesian Networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify the existence of direct causal influences between the linked propositions, and the strengths of these influences are quantified by

conditional probabilities (Figure 3). Thus, if the graph contains the variables

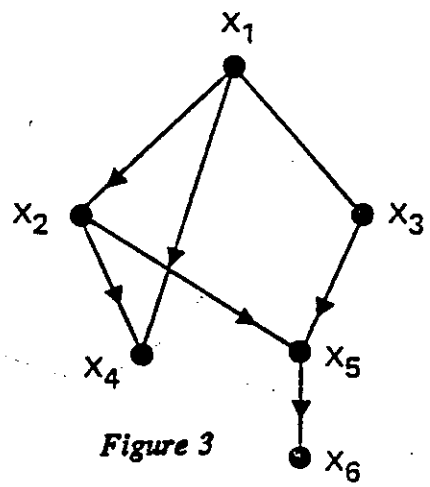

Figure 3

$x_1, \ldots, x_n$, and S_i is the set of parents for variable x_i, then a complete and consistent quantification can be attained by specifying, for each node x_i, an assessment $P'(x_i \mid S_i)$ of $P(x_i \mid S_i)$. The product of all these assessments,

$$P(x_1, \ldots, x_n) = \prod_i P'(x_i|S_i)$$

constitutes a joint-probability model which supports the assessed quantities. That is, if we compute the conditional probabilities $P(x_i \mid S_i)$ dictated by $P(x_1, \ldots, x_n)$, the original assessments are recovered. Thus, for example, the distribution corresponding to the graph of Figure 3 can be written by inspection:

$$P(x_1,x_2,x_3,x_4,x_5,x_6) = P(x_6|x_5)\,P(x_5|x_2,x_3)\,P(x_4|x_1,x_2)\,P(x_3|x_1)\,P(x_2|x_1)\,P(x_1).$$

An important feature of Bayes network is that it provides a clear graphical representation for many independence relationships embedded in the underlying probabilistic model. The criterion for detecting these independencies is based on ***graph separation***: namely, if all paths between x_i and x_j are "blocked" by a a subset S of variables, then x_i is independent of x_j given the values of the variables in S. Thus, each variable x_i is independent of both its siblings and its grandparents, given the values of the variables in

its parent set S_i. For this "blocking" criterion to hold in general, we must provide a special interpretation of separation for nodes that share common children. We say that the pathway along arrows meeting head-to-head at node x_k is normally "blocked", unless x_k or any of its descendants is in S. In Figure 1, for example, x_2 and x_3 are independent given $S_1 = \{x_1\}$ or $S_2 = \{x_1, x_4\}$, because the two paths between x_2 and x_3 are blocked by either one of these sets. However, x_2 and x_3 may not be independent given $S_3 = \{x_1, x_6\}$, because x_6, as a descendant of x_5 , "unblocks" the head-to-head connection at x_5, thus opening a pathway between x_2 and x_3.

Belief Propagation In Bayesian Networks

Once a Bayesian network is constructed, it can be used to represent the generic causal knowledge of a given domain, and can be consulted to reason about the interpretation of specific input data. The interpretation process involves instantiating a set of variables corresponding to the input data and calculating its impact on the probabilities of a set of variables designated as hypotheses. In principle, this process can be executed by an external interpreter who may have access to all parts of the network, may use its own computational facilities, and may schedule its computational steps so as to take full advantage of the network topology with respect to the incoming data. However, the use of such an interpreter seems foreign to the reasoning process normally exhibited by humans. Our limited short-term memory and narrow focus of attention, combined with our inflexibility of shifting rapidly between alternative lines of reasoning seem to suggest that our reasoning process is fairly local, progressing incrementally along prescribed pathways. Moreover, the speed and ease with which we perform some of the low level interpretive functions, such as recognizing scenes, comprehending text, and

even understanding stories, strongly suggest that these processes involve a significant amount of parallelism, and that most of the processing is done at the ***knowledge level itself***, not external to it.

A paradigm for modeling such active knowledge base would be to view a Bayesian network not merely as a passive parsimonious code for storing factual knowledge but also as a computational architecture for reasoning about that knowledge. That means that the links in the network should be treated as the only pathways and activation centers that direct and propel the flow of data in the process of querying and updating beliefs. Accordingly, we can imagine that each node in the network is designated a separate processor which both maintains the parameters of belief for the host variable and manages the communication links to and from the set of neighboring, logically related, variables. The communication lines are assumed to be open at all times, i.e., each processor may at any time interrogate the belief parameters associated with its neighbors and compare them to its own parameters. If the compared quantities satisfy some local constraints, no activity takes place. However, if any of these constraints is violated, the responsible node is activated to revise its violating parameter and set it straight. This, of course, will activate similar revisions at the neighboring nodes and will set up a multidirectional propagation process, until equilibrium is reached.

The fact that evidential reasoning involves both top-down (predictive) and bottom-up (diagnostic) inferences has caused apprehensions that, once we allow the propagation process to run its course unsupervised, pathological cases of instability,

deadlock, and circular reasoning will develop [5]. Indeed, if a stronger belief in a given hypothesis means a greater expectation for the occurrence of its various manifestations and if, in turn, a greater certainty in the occurrence of these manifestations adds further credence to the hypothesis, how can one avoid infinite updating loops when the processors responsible for these propositions begin to communicate with one another?

It can be shown that the Bayesian network formalism is supportive of self-activated, multidirectional propagation of evidence that converges rapidly to a globally-consistent equilibrium [4]. This is made possible by characterizing the belief in each proposition by a ***vector*** of parameters, similar to the likelihood vector of Eq. (20), with each component representing the degree of support that the host proposition obtains from one of its neighbors. Maintaining such a breakdown record of the origins of belief facilitates a clear distinction between belief based on ignorance and those based on firm but conflicting evidence. It is also postulated as the mechanism which permits people to trace back evidence and assumptions for the purpose of either generating explanations, or modifying the model.

As a computational architecture, Bayesian networks exhibit the following characteristics:

1. New information diffuses through the network in a single pass, i.e., equilibrium is reached in time proportional to the diameter of the network.

2. The primitive processors are simple, repetitive, and they require no working memory except that used in matrix multiplication.

3. The local computations and the final belief distribution are entirely independent of the control mechanism that activates the individual operations. They can be activated by either data-driven or goal-driven (e.g., requests for evidence) control strategies, by a clock, or at random.

Thus, this architecture lends itself naturally to hardware implementation, capable of real-time interpretation of rapidly changing data. It also provides a reasonable model of neural nets involved in cognitive tasks such as visual recognition, reading comprehension, and associative retrieval where unsupervised parallelism is an uncontested mechanism.

Rational Decisions and Quality Guarantees

Bayesian methods, unlike many alternative formalisms of uncertainty, provide coherent prescription for choosing actions and meaningful guarantees on the quality of these choices. The prescription is based on the realization that normative knowledge, that is, judgments about values, preferences, and desirability, represents a valuable abstraction of actual human experience and that, like its factual-knowledge counterpart, it can be encoded and manipulated to produce useful recommendations. While judgments about the occurrence of events are quantified by probabilities, the desirability of action-consequences is quantified by utilities (also called payoffs, or values) [6].

Choosing an action amounts to selecting a set of variables in a Bayesian network and fixing their values unambiguously. Such a choice normally alters the probability distribution of another set of variables, judged to be *consequences* of the decision variables. If to each configuration of the consequence set C we assign a utility measure

$u(C)$, representing its degree of desirability, then the overall expected utility associated with action a is given by

$$U(a) = \sum_C u(C)\, P(C|a,e) \qquad (29)$$

where $P(C|a,e)$ is the probability distribution of the consequence set C conditioned upon selecting action a and observing evidence e.

Bayesian methodologies regard the expected utility $U(a)$ as a figure of merit of action a and treat it, therefore, as a prescription for choosing among alternatives. Thus, if we have the option of choosing either action a_1, or a_2, we calculate both $U(a_1)$ and $U(a_2)$ and select that action that yields the highest value. Moreover, since the value of $U(a)$ depends on the evidence e observed up to the time of decision, the outcome of the Maximum-Expected-Utility criterion will be an evidence-dependent ***plan*** (or decision rule) of the form: If e_1 is observed, choose a_1; if e_2 is observed, choose a_2,...

The same criterion can also be used to rate the usefulness of various ***information sources*** and to decide which piece of evidence should be acquired first. The merit of querying variable x can be decided prior to actually observing its value, by the following consideration. If we query x and find the value v_x, the utility of action a will be

$$U(a|v_x) = \sum_c u(C|a,\, x{=}v_x) P(C|a,\, e,\, x{=}v_x)$$

We will be able, at this point, to choose the best action among all pending alternatives and attain the value

$$U(v_x) = \max_a U(a|v_x)$$

However, since we are not sure of the actual outcome of querying x, we must average

$U(v_x)$ over all possible values of v_x, weighed by their appropriate probabilities. Thus, the utility of querying x calculates to

$$U_x = \sum_{v_x} P(x=v_x|e)U(v_x)$$

where e is the evidence available so far.

This criterion can be used to schedule many control functions in knowledge-based systems. For example, we can use it to decide what to ask to user next, what test to perform next, or which rule to invoke next. The expert system PROSPECTOR [7] actually employed a scheduling procedure (called J*) based on similar considerations. If the consequence set is well defined and not too large, this information-rating criterion can also be computed distributedly, concurrently with the propagation of evidence. Each variable x in the network stores an updated value of U_x and, as more evidence arrives, each variable updates its U_x parameter in accordance with those stored at its neighbors. At query time, attention will be focused on the observable node with the highest U_x value.

It is important to mention that the Maximum-Expected-Utility rule was not chosen as a prescription for decisions for sheer mathematical convenience. Rather, it is founded on pervasive patterns of psychological attitudes towards risk, choice, preferences, and likelihoods. These attitudes are captured by what came to be known as the axioms of Utility theory [8]. Unlike the case of repetitive long series of decisions (e.g. gambling), where the expected-value criterion is advocated on the basis of a long run accumulation of payoffs, the expected utility criterion is applicable to single-decision situations. The summation operation in Eq.(29) originates not with additive accumula-

tion of payoffs but, rather, with the additive axiom of probability theory (Eq. 3).

In summary, the justification of decisions made by Bayes methods can be communicated in intuitively meaningful terms and the assumptions leading to these decisions can be traced back with ease and clarity.

References

[1] Jeffrey, R., "The Logic of Decisions," McGraw-Hill, Chapter 11, 1965.

[2] Pearl, J., "Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach," *Proceedings of AAAI Conference on Artificial Intelligence*, Pittsburgh, Pennsylvania, pp. 133-136, 1982.

[3] Kim, J. and Pearl, J., "A Computational Model for Combined Causal and Diagnostic Reasoning in Inference Systems," *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 190-193, 1983.

[4] Pearl, J., "Fusion, Propagation and Structuring in Bayesian Networks," Technical Report CSD 850022, Cognitive Systems Laboratory, UCLA, June 1985.

[5] Lowrance, J., "Dependency-Graph Models of Evidential Support," COINS Technical Report 82-26, University of Massachusetts at Amherst, 1982.

[6] Raiffa, H., "Decision Analysis: Introductory Lectures on Choices under Uncertainty," Addison-Wesley, Reading, Massachusetts, 1968.

[7] Duda, R.O., Hart, P.E., and Nilsson, N.J., "Subjective Bayesian Methods for Rule-Based Inference Systems," *Proceedings of the 1976 National Computer Conference (AFIPS Conference Proceedings)*, 45, pp. 1075-1082, 1976.

[8] von Neumann, J., and Morgenstern, O., "Theory of Games and Economic Behavior," 2nd edition, Princeton University Press, Princeton, New Jersey, 1947.