Massachusetts Institute of Technology

6.034 Artificial Intelligence

Handout 6034 Item # **14**

Notes on Probability Theory and Bayesian Belief Networks

(to be read in addition to the paper by P. Szolovits on Uncertainty in Medical Decision Making)

1 Introduction: Probability Theory and Learning

Probability theory forms a natural bridge in AI from representation to learning. At heart, probability deals with *uncertainty*; in a sense all learning can be viewed as a change in one's uncertainty about the world. Indeed, probability theory is *provably* the only "coherent" calculus to deal with uncertainty, and as such underpins much current work in learning and decision theory — in AI and elsewhere. By "coherent" we mean precisely this: namely, that if one acts in accordance with the calculus of probability theory (as described by the axioms below), then it will be *impossible* for anybody (including nature) to construct a "Dutch book" against you. A Dutch book is some set of apparently fair bets on events such that no matter what the outcome of the events, you will lose money. But the choice remains yours: you are free to lose money — all you have to do is violate the probability calculus.[1]

In the following sections we outline (1) the basic concepts and formulas of probability theory, including the result most directly relevant to learning (or uncertainty reduction) in AI, namely, Bayes' Theorem; and (2) the application of Bayes' Theorem so that one can pass from evidence to hypotheses, testing one hypothesis against another. Bayesian reasoning has proved to be an especially fruitful research area for practical AI uncertainty-based systems, ranging from software that tracks the NASA shuttle launches to diagnostic aids for printer malfunctions, Microsoft's help systems, and Bayesian data searches of keywords from texts on the internet.

[1] For a detailed exposition of this claim, see the chapter by D. Osherson in Osherson and Smith, *An Introduction to Cognitive Science*, MIT Press, Cambridge, MA, v. 3, "Reasoning".

2 Basic Formulas

In probability theory we are concerned with the occurrence of "events" that may be thought of as the outcome of experiments, and our "degree of belief", or probability measure, of those events. The probability of an event A is denoted by:

$$P(A) = \text{probability that the event } A \text{ occurs.}$$

It is sometimes helpful to think of probability in the following way. First we imagine all the possible outcomes of the experiment and call this collection of outcomes U, for the Universe of possible events. A smaller collection of outcomes we are interested in, A, has its probability defined as the "area" of A divided by the "area" of U, with "area" suitably defined — see figure 1. Particular probability models give different definitions for what the "area of A" really means.[2] In any case,

$$P(A) = (\text{area of } A)/(\text{area of } U)$$

Note that we have divided by the "area" of all the possible events U as a kind of scaling factor, so that $P(U) = P(U)/P(U)$=1, as we would like. Continuing to use this figure and the definition of probability, we see that the probability that either one of the two events A *or* B occurs is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Still referring to the set-based figure, $P(A \text{ or } B)$ is simply the area of both A and B minus their intersection, divided by the area of U. (We subtract out the intersection so that we don't double count the area contribution from A and B.)

The probability $P(A \text{ and } B)$ is simply the measure of how often *both* A and B occur, so naturally it is just the area of the intersection of A and B, divided by the area of U. Henceforth we will write it as $P(A, B)$, which is also equivalent to $P(B, A)$.

Example. Roughly, this approach works for discrete as well as continuous spaces: for instance, if we toss a fair coin twice, then there are 4 "points" or events possible in the universe U: HH, HT, TH, and TT, where T=a tail came up; H=a head came up. Thus if A=two heads, then P(A)=1/4 because the "area" of A is 1 point, divided by the universal area of 4 points.

[2]Here of course we really mean to introduce the notion of the *measure* of A in the technical mathematical sense as a way to calculate "area", but we put this technical machinery aside here.

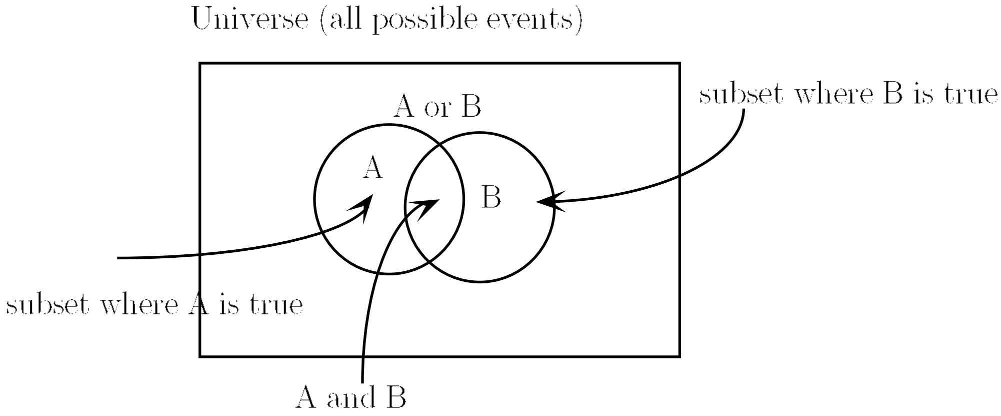

Figure 1: A set-based view of probability.

Definition: Given two probabilities P_1, P_2 defined over the same universe, we call the ratio of the two probabilities P_1/P_2 the *odds* or *likelihood ratio* in favor of P_1 vs. P_2. For example, if the two probabilities are P_1= probability of getting a heads on one toss of a fair coin = 1/2, and P_2= probability of getting a tail on one toss of a fair coin = 1/2, then the odds P_1/P_2= (1/2)/(1/2)=1, i.e., equal odds, as we expect by intuition. We sometimes use odds or likelihood ratios to simplify calculations when the "normalizing" factor introduced by the "area" of the universe need not be taken into account (odds are also sometimes easier for people to grasp intuitively).

2.1 Probability axioms

While informal notions of probability have been known to gamblers since at least the 1500s and far earlier, it was not until this century that Kolmogorov and others demonstrated that only a few basic axioms suffice to account for all one needs to say (though Kolmogorov's particular formulation, especially the last axiom, has been disputed on other grounds).

The probability $P(A)$ of an event A is a number between 0 and 1, inclusively, with 0 representing falsity and 1 truth. One way to list the axioms is as follows.

1. [No probability can be negative]
   $P(A) \geq 0$.
2. [The probability of logical truth is 1]
   If event A is logically true, then $P(A) = 1$.
3. [The probability that one of two logically exclusive events is true is the *sum* of their probabilities]
   If A and B are logically exclusive, then $P(A \text{ or } B) = P(A) + P(B)$.

4. [The probability of a conjunction of events equals the probability of the first, *assuming* the second is true, times the probability of the second]
   If $P(B)$ is not 0, then $P(A, B) = P(A|B) \cdot P(B)$.

From axioms 1–4 we can derive (at least) the following result (introducing some new notation):

$$P(\neg A) = 1 - P(A)$$

Alternative notation: $P(\overline{A})$.

2.2 Conditional probability

The last Kolmogorov axiom introduces the notion and the notation of *conditional probability*, which demands more explanation. Referring again to figure 1, what is the probability that B occurred, **given** our knowledge about A? If A occurred, then the collection of all possible outcomes is no longer U, but must be A — typically a value distinct from U. From the definition of probability, the probability of B *given* A, written $P(B|A)$, must therefore be the ratio of two areas: the area common to both A and B and the area of A. It is as if we had now *restricted* our focus to a new universe of events, namely, just those in A — we have just "shrunk" our view down to look at A and the events in B that overlap with it. In other words, since we limit our universe to just that subset of events where A holds, we now scale (divide) by the size of A in order to make this subset be the size of the entire universe, the value 1. Writing this out, we have:

$$P(B|A) = P(A, B)/P(A)$$

or rewriting,

$$P(A, B) = P(B|A) \cdot P(A)$$

By analogy, since A and B are completely interchangeable, we must also have that:

$$P(A|B) = P(A, B)/P(B)$$

or, rewriting again,

$$P(A, B) = P(A|B) \cdot P(B)$$

It should be stressed that $P(B|A) = x$ does *not* mean that whenever A is true then the probability of B is x. It means that if A is true, and *everything else known is irrelevant for* B, then $P(B) = x$.

Example: Let us look at our 2 coin tosses experiment again. Initially, the probability of getting two heads is 1/4. Now suppose someone else observes the two tosses and tells us, "at least one of the tosses was a Head". Note that this has given us new information — so we might expect it would change our degree of belief in the outcome "two heads turned up". Note also that this is, in effect, *learning* — gaining knowledge (reducing uncertainty) from new information.

Looking at our notion of conditional probability, if we let A= the event that two heads turned up, and B=at least one of the two tosses was a head. We want to calculate the probability of A **given** B, i.e., given that at least one of the two tosses was a head. Now, the discrete "area" in which at least one of the tosses is a head is simply 3: HH, HT, or TH (note that it is smaller than our original space=4). The "area" in which both A (two heads) and B overlap is simply 1: HH (there is just one event where at least one of the tosses is a head *and* both tosses are heads). So by the definition of conditional probability, the answer must just be the ratio of these two "areas", or 1/3. We can calculate the answer as follows, since $P(A, B)$ with respect to the *original* universe =1/4 and $P(B)$ with respect to the original universe = 3/4 (3 out of 4 times we get at least one head):

$$P(A|B) = P(A, B)/P(B) = (1/4)/(3/4) = 1/3$$

Indeed, in this case our degree of belief has increased from 1/4 initially to 1/3 — our uncertainty has been reduced — given that we know something about the event B. (We have a *new* event space or universe we are looking at.)

Example: The universe shown in figure 2 consists of 64 discrete events (shown as squares). A, shaded by lines tilted to the right, has an area of 16 squares, so $P(A)$=16/64= 1/4. B, shaded by lines running down to the left, consists of 10 squares, so $P(B)$=10/64 = 5/32. Their intersection (overlap), cross-hatched in both directions, has an area of 6 squares, so $P(A, B) = 6/64$. From these we can compute the conditional probabilities, either using the formulas above or directly:

$$\begin{aligned} P(A|B) = \quad & P(A, B)/P(B) = (3/32)/(5/32) = 3/5 = \\ & 6/10 \text{ (directly: the ratio of } A\text{'s area overlap to the area of } B) \\ P(B|A) = \quad & P(A, B)/P(A) = (3/32)/(1/4) = 12/32 = \\ & 6/16 \text{ (directly: the ratio of } B\text{'s overlap to the area of } A) \end{aligned}$$

2.3 Independence

Two events are *independent* if knowing whether one event is true or not tells us nothing more about whether the other is true or not — in other words, if knowing the truth or falsity of one

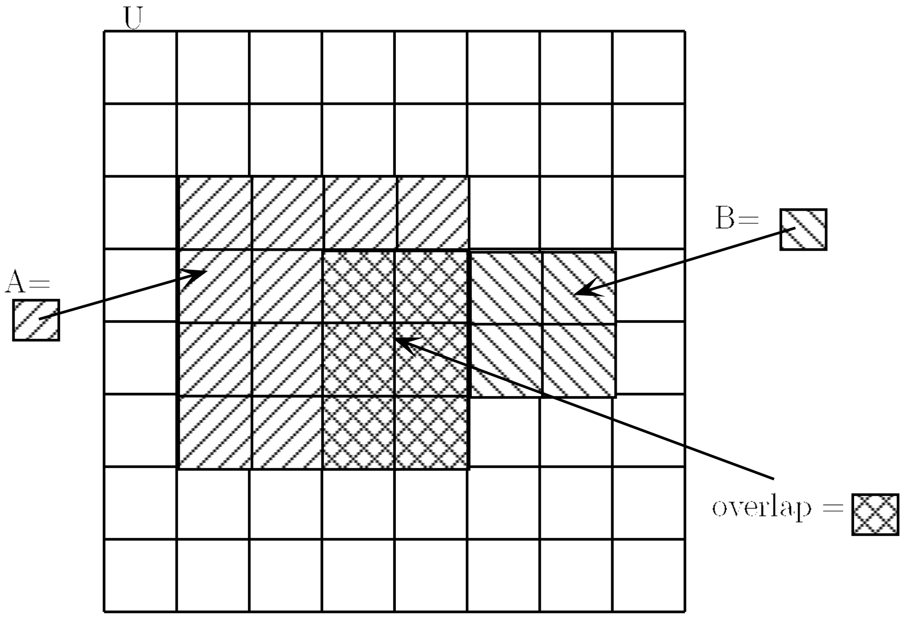

Figure 2: An example of conditional probability. The Universe U consists of 64 "squares".

event does nothing to change our estimate of the probability of the other event. Expressed in terms of conditional probability, if A and B are independent,

$$\begin{aligned} P(A|B) = &\ P(A) \text{ and} \\ P(B|A) = &\ P(B) \end{aligned}$$

Using these formulas in the equation for conditional probability, we see that if two events are independent,

$$P(A, B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$$

That is, we can just multiply two *independent* events together to get their joint probability (the probability of both occurring). That fits with our usual interpretation of a sequence of two coin flips, which are assumed independent: $P(\text{head,head}) = P(\text{head}) \cdot P(\text{head}) = 1/2 \times 1/2 = 1/4$.

Put another way, independence asserts that the proportion of B that is in A (or vice-versa) is the *same* as the proportion of A that is in the entire universe U. Figure 3 illustrates independent and dependent events, using a 16-square universe.

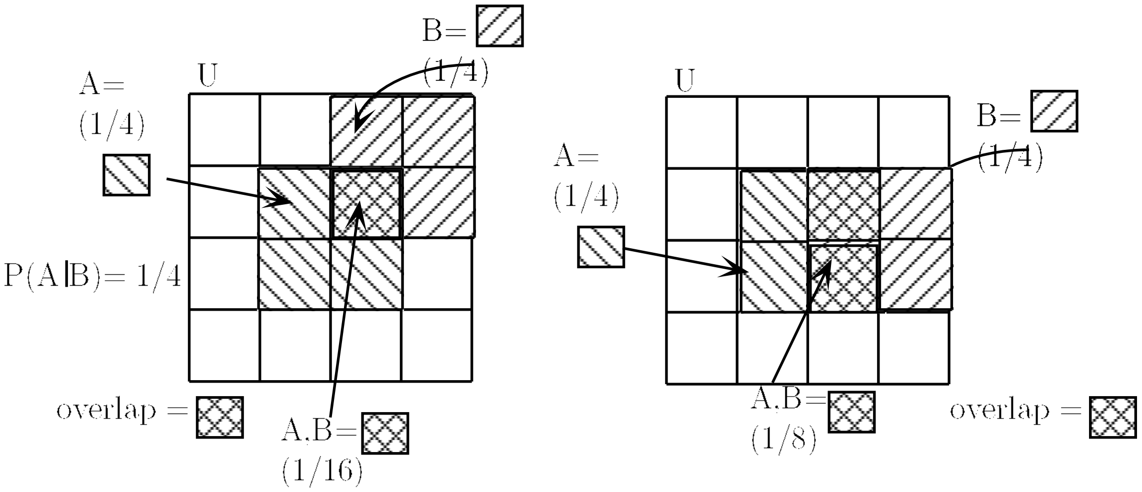

Figure 3: An illustration of probabilistic independence. In the figure on the left, A is independent of B. In contrast, on the right, A is *not* independent of B: once we know that we are working in event space B, then the ratio of A to that new event space is different than before. Thus, on the left, $P(A|B)$=1/4= $P(A)$. On the right, $P(A|B)$=1/2 which is *greater* than $P(A)$=1/4 (A and B are not independent).

2.4 Conditional Independence

In fact, from a certain perspective, probabilities are *always* conditioned by *some* background context, call it C. That is, whenever we write $P(A|B)$ we in fact mean $P(A|B,C)$ — that is, A given *both* B and C. If nothing else, the background is really the universe U — that is, in the case of $P(B|A)$, we really are writing $P(B|A,C)$.

If C tells us nothing more than what B already does, so that knowing C does not change the probability of A, then we have $P(A|B,C) = P(A|B)$. This is called the *conditional independence* of A and C on B. Intuitively, when two variables depend on the same set of variables, but not directly on each other, they are conditionally independent.

In other words, under the conditional independence of A and C, once the state of B is known, no knowledge about C alters the probability of A. (If event B is empty, or null, we simply say that A and C are independent.) For A, C conditionally independent on B we have:

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

Mathematically, the conditional independence of A and C with respect to B may be *defined* as:

$$P(A, C|B) = P(A|B) \cdot P(C|B)$$

that is, we can multiply out the joint probability of conditionally independent events, just like as in the case of "ordinary" independence:

$$P(A, C) = P(A) \cdot P(C)$$

but now we include the conditioning event B. For example, if an infection causes both a fever and diarrhea, but there is no other correlation between these symptoms, then we can write:

$$P(\text{fever}, \text{diarrhea}) = P(\text{fever}|\text{infection}) \cdot P(\text{diarrhea}|\text{infection})$$

Note: Conditional independence (with respect to, say, B) **does not** mean that A and C are in general independent or dependent with respect to each other (*without* the conditioning context B). Rather, conditional independence has to do with whether "information" can be transmitted from A to C across an intervening event B. Suppose that the value of B is fixed or instantiated (we *know* its true value). Then C can no longer contribute any new information to our knowledge about A. We say then that C is conditionally independent of A (with respect to B). It is this notion that is used extensively when we talk about the machinery of Bayes networks.

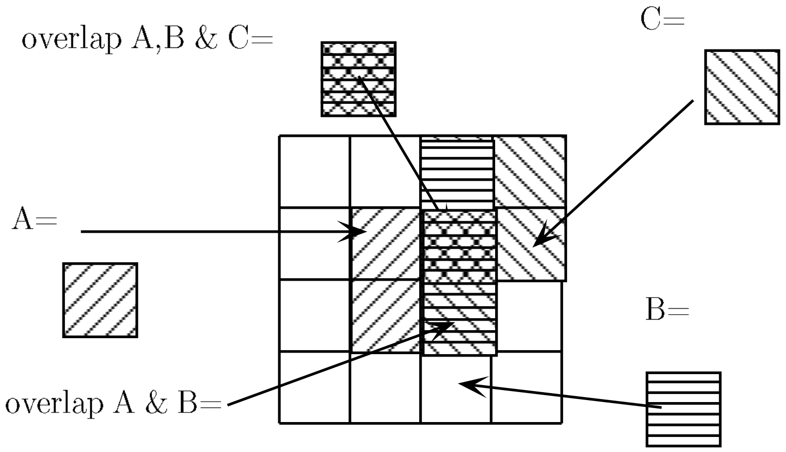

Figure 4: This figure shows that having A and C independent does *not* mean that these events are conditionally independent given B — it depends on how B overlaps with respect to the other events. The right-hand part of the figure shows that *given* B, we know with certainty the probability of A, that is, $P(A|B, C) = 1$, which is *not* equal to the probability of $P(A|B) = 2/3$.

2.5 Bayes' Theorem

The challenge in almost all statistical science is to determine how to use the evidence, E, contained in data. Bayes' Theorem is a very powerful tool for doing just this. It can be shown to be the only "coherent" way to pass from specific evidence to general hypotheses, i.e., to combine evidence (in the same sense of coherence mentioned earlier). That is, it is THE rule for combining evidence and for choosing between competing hypotheses, and thus of great importance in AI (or statistical science generally).

Let us first take an intuitive pass at Bayes' Theorem. Bayes' Theorem says that we can figure out whether one hypothesis is more likely than another (has greater odds, or greater likelihood than another hypothesis) simply by considering two numbers:

- The **prior probabilities** of the hypotheses: what is your degree of belief in the hypotheses *before* you look at any evidence or carry out any experiment at all? (For instance, you might believe that there the prior chance in Boston on any day that it will rain is 0.40, i.e., it will rain 40% of the time.) Call this $H1$.

  For our other hypothesis, $H2$, suppose we our prior probability assessment that it will snow on any given day in Boston is 0.30.

  Then the odds or relative probability of rain vs. snow on any given day in Boston is simply the ratio of these two numbers, or $0.4/0.3 = 4/3$ — 4:3 or a greater chance that it will rain as opposed to snow.

Next, we *scale* these odds by an estimate of *how much* some observed evidence, E, boosts (or decreases) our belief in each hypothesis.

- The **likelihoods**: if the hypotheses *were* true, what is the probability that the evidence as you now see it would have appeared? These are simply the two conditional probabilities $P(E|H1)$ and $P(E|H2)$. (For example, if it's raining, then an observed temperature of 40 degrees F, say, is quite high, $P(E = t > 40|\text{rain}) = 0.9$; while if it's snowing, then it's quite unlikely, but not impossible, that the temperature is greater than 40, $P(E = t > 40|\text{snow}) = 0.1$.) The odds that the first conditional event will happen as opposed to the second is, as usual, the ratio of these two numbers, or,

  $$P(E = t > 40|\text{rain}) = 0.9/P(E = t > 40|\text{snow}) = 0.1 = 0.9/0.1 = 9/1\ or 9 : 1.$$

Our idea is now to scale the original odds by the new odds based on what was observed (temperature greater than 40 degrees) to get a *revised* or **posterior probability** estimate of the relative odds of the two hypotheses. Our scaling factor of 9:1 means that our original ratio of 4/3 ought to be multiplied by 9/1, giving 36/3 or 12/1. That makes sense: if we observe that the temperature's greater than 40 degrees, then that ought to boost our confidence that it's raining as opposed to snowing. Our evidence has increased assessments of the odds in favor of rain nine time.

Since this is an example where new information (the experimental evidence, E) *changes* our degree of certainty or belief in an event, it may also be considered a kind of learning. In Bayes' original paper, published only after his death by a friend, he wanted to figure out how to go from the usual deduction of the probability of a specific result *given* a general hypothesis, $P(E|H1)$, in Bayes' case, the hypothesis that "the sun will rise today, *given* that the sun rises every day" and **turn it around**, showing how to pass from a specific evidential result, "the sun rose today" to the probability of the general case, "the sun will rise every day", i.e., $P(H1|E)$, or the probability of the hypothesis *given* the evidence. This switch from *P(evidence* | *hypothesis* to *P(hypothesis* | *evidence)* is the hallmark of Bayes' result.

We can *derive* Bayes' Theorem from our probability axioms (including the notion of conditional probability) in the following way. By the definition for joint probabilities, we have,

$$P(A, B) = P(A|B)P(B)$$

Substituting for $P(A, B)$ in the following definition for conditional probability,

$$P(B|A) = P(A, B)/P(A) = P(A|B)P(B)/P(A)$$

That is,

$$P(H|E) = P(E|H)P(H)/P(E)$$

Here, $P(H)$ is the **prior** probability of a hypothesis H; and $P(H|E)$ is the **posterior** probability of a hypothesis, *given* that we observed evidence E.

Note that when we are *comparing* two hypotheses, or even trying to just *maximize* the probability $P(H|E)$, then the denominator $P(E)$ is generally fixed (the evidence is fixed), so in such cases we can ignore it.

When there are a number of possible but mutually exclusive outcomes $B_1, B_2, \ldots, B_n$, one of which must occur before A occurs we can also get a corresponding result. The natural generalization is to calculate the probability that B_i occurs given that A occurs. This is just:

$$P(B_i|A) = P(A|B_i)/\Sigma_{j=1}^{n} P(A|B_j)P(B_j)$$

We can imagine all the B_i corresponding to all possible disease states and A is a symptom. Then the total probability of the symptom is just the sum of the joint probabilities of the symptom with each disease.

3 Applications of Bayes' Theorem

Example. Assuming exclusivity, we can use Bayes' formula to do disease diagnosis based on the outcomes of various tests. Assume we are given a cancer test that is 95% reliable, which means that there is 5% chance that the test will be negative for someone with cancer, and that, similarly, there will be a 5% chance that the test will be positive for someone without cancer.

For a prior probability, we assume further that one person in 500 have cancer, i.e., that $P(C) = 1/500$ or 0.002.

If the test is positive, what is the probability that the person has cancer? That is, we want to calculate $P(C|+)$ where $+$ denotes a positive test outcome. We can calculate this by Bayes' Theorem (with one additional twist):

$$P(C|+) = P(+|C) \cdot P(C)/P(+)$$

The "twist" is that we have to calculate $P(+)$ — it is not given to us directly by the data. It is the probability of *all* ways of getting a positive test outcome — which can happen even if we don't have cancer with a chance of 5%. (Having cancer is just one way of getting a positive result, with a successful detection of 95%.)

We know:

$$\begin{aligned}
P(+|C) &= 0.95 \\
P(-|C) &= 0.05 \\
P(+|\neg C) &= 0.05 \\
P(C) &= 0.002 = 1/500
\end{aligned}$$

By the assumption that the ways of getting a positive test outcome are exhausted by the cancer/no cancer possibilities we have,

$$\begin{aligned}
P(+) &= P(+|C) \cdot P(C) + P(+|\neg C) \cdot P(\neg C) \\
&= 0.95 \cdot 0.002 + 0.05 \cdot 0.998 \\
&= 0.0019 + 0.0499 \\
&= 0.0518 \text{ (about } 1/20)
\end{aligned}$$

(Note that 1/20 is about right: out of about 500 people who take the test, 5% or 25 will test positive even if they don't have cancer, and the remaining 1/500 who really do have cancer will add another positive outcome — so roughly 26 people out of 500 will test positive. This yields roughly 0.052 for $P(+)$.)

Concluding: $P(C|+) = 0.95 \cdot 0.002/0.0518 = 0.0019/0.0518 = 0.03668$, or about 4%.

Conclusion: the "accurate" test increases our knowledge of whether we have cancer from .002 to roughly .04. ().2 % to 4%) This is quite startling to most people — who don't reason using probabilities very well, and certainly don't seem to use Bayes' Theorem to combine probabilities (at least on the surface — there is on-going debate as to whether people do). Most people expect a positive test outcome to raise the probability of cancer much higher — most say 95% or, if they are clever, they adjust the probability downwards to something like 50% due to the "rare incidence" factor. But we have just seen that Bayes' Theorem is really the right way to factor in the evidence.

You might try to figure out how accurate the test would have to be in order to, say, raise the probability that one has cancer if the test turns positive from 4% to 50%. (It must be a very, very, accurate test.)

4 Probability Modeling

We now extend our view of probability a bit in order to develop the notion of Bayesian belief networks in the next section.

Assume the world is described by the state values of a number of (discrete) *random variables*. Each random variable X has a *domain* of possible values $\{x_1, x_2, \ldots, x_n\}$, which for our purposes we can just assume to be boolean, that is, ranging over the domain {true, false}.

We use the notation $P(A)$ for the *unconditional* or prior probability that a proposition A is true. A usually refers to the particular value of some random variable. For instance, if the variable is W, for *weather* and the domain is {Cloudy, Sunny, Snow}, we might write $P(W = \text{cloudy})$ to refer to this value. For shorthand, one could write, $P(\text{cloudy})$.

The probability values for each value in the domain of a random variable is a *probability distribution*. Note that for discrete-valued variables, we shall have to specify a *table* of probability entries.

Example. The table for our weather example might look like the following. Note that the sum of entries in a probability distribution must be 1.

Weather	P(Weather=that value)
Cloudy	0.3
Sunny	0.6
Snow	0.1

The table of probabilities for the combination of two random variables X, Y is a *joint probability distribution*. Now we will have to specify 2^{n-1} numbers in the table — 1 for each combination of variable values, except the last, since the sum must be 1 (we have one fewer degrees of freedom).

For a conditional probability table based on Boolean-valued random variables X and Y, we write the following:

Y	$P(X = \text{true}\|Y)$	$P(X = \text{true}\|Y)$
T(rue)	a	$1 - a$
F(alse)	b	$1 - b$

Note that we actually need to specify only one column, since the numbers in one column are dependent on the other. For Boolean variables, the number of independent entries we need goes up exponentially in the number of random variables, as mentioned — it is 2 numbers here, in the case of 2 random variables; it is 4 in the case of 3 random variables, X, Y, Z, and so forth.

To summarize, in order to be more precise about events that take on discrete values (like *true* or *false*, we will write $P(A = a)$ to mean the probability that A happened with the value a and $P(A = \bar{a})$ to mean the probability that A didn't happen (the lowercase letter is now standing in for the actual outcome of the event). We will reserve $P(A)$ to mean a full table of probabilities (including all the possible outcome values). So $P(A = a|B = b)$ is now the probability that A happened given B happened, and $P(A = a|B = \bar{b})$ is the probability that A happened given that B did not happen. To review, $P(A|B)$ may be considered as a *table* that contains information on $P(A = a|B = b)$, $P(A = \bar{a}|B = b)$, $P(A = a|B = \bar{b})$ and $P(A = \bar{a}|B = \bar{b})$. All these equal signs can get confusing, so as a shorthand we will write $P(a|\bar{b})$ to mean $P(A = a|B = \bar{b})$. The clue that we mean a single probability, rather than a table, is the use of lowercase letters.

With this background, we now turn to Bayesian nets (aka Bayes nets)

5 Bayesian Belief Networks

The classic example of the use of Bayesian belief nets is given by Judah Pearl. You have the following situation that can be described by five random, Boolean-valued variables:

- You have a burglar alarm that may be triggered by a burglary or an earthquake. We let B denote the random variable for whether a burglary occurs; it has the values either *true* or *false*. We let DB denote the random variable for whether the earthquake occurs; it has the values either *true* or *false*.
- Your neighbor John will call if he hears the alarm, but also just calls for the fun of it. J is the random variable for whether John calls or not.

- Your neighbor Mary will call when she hears the alarm but she sometimes will miss it, because she listens to MTV (or whatever). M is the random variable for whether Mary calls or not.

The question of interest: what can you conclude about the probability of a burglary *given* knowledge of whether John calls, or Mary calls, or both, whether an earthquake has happened, etc.?

Note first that in the absence of any other structure, with 5 random variables there are $2^5 = 32$ possible states of the world. Thus, we would in general need 31 numbers to completely specify the joint probability table of the events, and compute any probability of interest (like those in the previous paragraph). The Bayes net is a way to reduce this computational load, by systematically taking into account conditional independence and dependence. In this case, we can assume:

- Earthquakes are independent of burglaries
- Phone calls depend only on the alarms

Now, one can then develop the full jointly specified state of the world, for some particular set of values for B, E, A, J, M. Note that we will now use a notation for describing particular values that B, E, A etc. can take on. In particular,

$$P(B = b, E = e, A = a, J = j, M = m)$$

where b, e, a, j, m are the values that range over {true, false}.

Using the rule for conditional probabilities, we can calculate this as follows:

$$\begin{aligned} P(B = b, E = e, A = a, J = j, M = m) = \\ & P(M = m | B = b, E = e, A = a) \times \\ & P(B = b, E = e, J = j, A = a) \end{aligned}$$

But now the second term, with 4 random variables, can be further expanded by the rule for conditional probability:

$$\begin{aligned} P(B = b, E = e, J = j, A = a) = \\ & P(J = j | B = b, E = e, A = a) \times \\ & P(A = a, B = b, E = e) \end{aligned}$$

So we can substitute this back into our original formula:

$$\begin{aligned} P(B = b, E = e, A = a, J = j, M = m) = \\ & P(M = m | B = b, E = e, A = a) \times \\ & P(J = j | B = b, E = e, A = a) \times \\ & P(A = a, B = b, E = e) \end{aligned}$$

Continuing, we can expand the second term of this last factor, with three random variables, and then expanding again until we get down to just one random variable (all by using the rule for conditional probability), with the final result:

$$\begin{aligned} P(B=b,E=e,A=a,J=j,M=m) = \;& P(M=m|B=b,E=e,A=a)\times \\ & P(J=j|B=b,E=e,A=a)\times \\ & P(A=a|B=b,E=e)\times \\ & P(E=e|B=b) \\ & P(B=b) \end{aligned}$$

So, this can be done, but it is computationally intensive — all those 31 numbers. If we look at the conditional independence assumptions, one can show that we need only 10 numbers, because Mary and John calling only depend on the alarm (the first two lines given in the table below); while the third line says that earthquakes are independent of burglaries.

$$P(M=m|B=b,e=E,J=j,A=a) = \quad P(M=m|A=a)$$

Then we can now simplify our previous example so that we need to specify only 10 numbers. In the next section we shall see how to do this in general, without all this fuss:

$$\begin{aligned} P(B=b,E=e,A=a,J=j,M=m) = \;& P(M=m|A=a)\times \\ & P(J=j|A=a)\times \\ & P(A=a|B=b,E=e)\times \\ & P(E=e)\times \\ & P(B=b) \end{aligned}$$

That's fine in this particular case, but we want a more general way to figure this out, without resorting to tedious thought each time — we want a method a computer can use. Such an alternative is a graphical notation called *Bayes Networks*, which makes the assumptions of conditional independence explicit, and thereby localizes the probability table entries and calculations that have to be carried out.

5.1 Bayesian Networks: Definitions and example

Bayesian networks combine a graphical component, which depicts conditional dependence, and a numerical component, which depicts probability distribution tables. They consist of the following:

- A set of *variables* and as set of *directed edges* between variables
- Each variable has a finite set of mutually exclusive states

- The variables together with the directed edges form a *directed acyclic graph* (DAG). (A directed graph is *acyclic* if there is no directed path $A_1 \rightarrow A_2 \rightarrow \ldots A_n$ such that $A_1 = A_n$.) Thus the variables are the nodes in the graph.
- To each variable A with parents $B_1, \ldots, B_n$ there is attached a conditional probability table $P(A|B_1, \ldots, B_n)$

Informally, each directed link from B to A in the graph says that B is a possible cause of A. Importantly, the *absence* of a link (arrow) between any two nodes (variables) signals independence (either ordinary independence or conditional independence) between those two variables.

Example. Figure **??** shows the directed graph for our earthquake example, along with the associated conditional probability tables — one for each variable in the graph, or five tables in all. Note that the tables for B and E are just the *prior* probabilities for these events (since there are no links leading into these nodes).

The conditional probability tables are precisely those required to evaluate the probability of any specified state, that is, any combination of values for every variable. The key point that makes Bayesian network computations simpler is that in general, to find the probability distribution of a state, we only have to compute the *product* of the probabilities of its parents. This *localizes* our probability calculations. You should note that this corresponds exactly to our example in the previous section. For instance, in our earthquake example,

$$\begin{aligned} P(B = b, E = e, A = a, J = j, M = m) = & \\ & P(M = m|A = a) \times P(J = j|A = a) \times \\ & P(A = a|B = b, E = e) \times \\ & P(E = e) \times \\ & P(B = b) \end{aligned}$$

which is precisely the sample multiplication we carried out earlier, but much simpler. In general, the computation is now easier:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \Pi_{i=1}^{n} P(X_i = x_i|\text{Parents}(X_i))$$

because all we have to do is compute the product of table entries, where $\text{Parents}(X_i)$ is just the values of the *parent* variables in the net, rather than *all* possible variables.

For example, suppose we wanted the probability of any *completely specified state*, that is, for any combination of values, *true* or *false*, for every variable. So we have, where $B = b$, for b in the set {T(rue), F(alse)}, etc.:

$$\begin{aligned} P(B = b, A = a, J = j, M = m) = & \\ & P(M = m|A = a) \times P(J = j|A = a) \times \\ & P(A = a|B = b) \times \\ & P(B = b) \end{aligned}$$

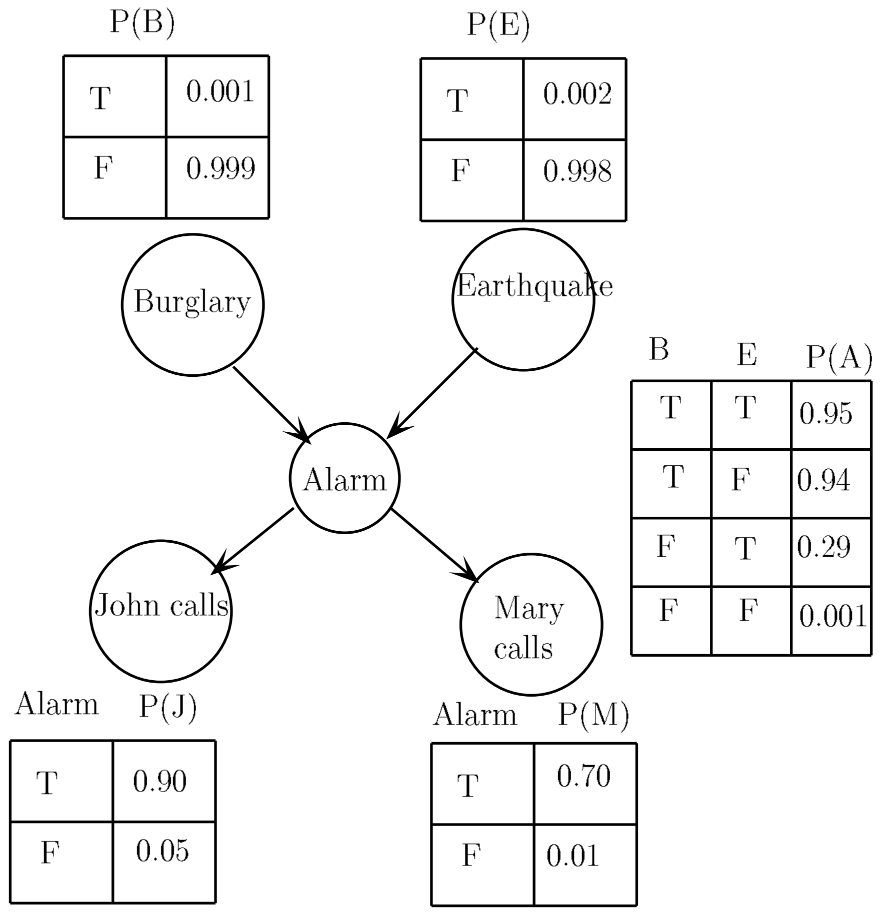

Figure 5: A Bayesian network for the burglary/earthquake example.

What we want to compute for some particular question will have some of these values fixed, and some not. That is, only some of the states will be instantiated or specified. For example, if Mary and John have called, each these values must be *true*, but we don't know whether the alarm has actually gone off or not, nor anything about the earthquake.

In this case, we want to compute the probability that B is true, given J and M are true. Using the conditional probability formula we have:

$$P(B = T|J = T, M = T) = P(B = T, J = T, M = T)/P(J = T, M = T)$$

which is the sum over both alarm states, and is equal to:

$$P(B = T, A = *, J = T, M = T)/P(B = *, A = *, J = T, M = T)$$

where $P(B = *, M = T)$ denotes the sum $P(B = T, M = T) + P(B = F, M = T)$ (all values for B).

Now we want to compute, for the numerator,

$$P(B = t, A = *, J = T, M = T)$$

and this is just the *sum* over all possible values (fully specified states) for the A(larm)={T,F}:

$$P(B = T, A = T, J = T, M = T) + P(B = T, A = F, J = T, M = T)$$

For the first term, where the alarm as gone off, A=T(rue), and using the conditional tables,

$$\begin{aligned} P(B = T, A = T, J = T, M = T) = & \\ & P(M = T|A = T)\times \\ & P(J = T|A = T)\times \\ & P(A = T|B = T)\times \\ & P(B = T) \end{aligned}$$

which is just $0.7 \times 0.9 \times 0.95 \times 0.001 = .000599$

And for the second term, A=f(alse),

$$\begin{aligned} P(B = T, A = T, J = T, M = T) = & \\ & P(M = T|A = F)\times \\ & P(J = T|A = F)\times \\ & P(A = F|B = T)\times \\ & P(B = F) \end{aligned}$$

which is just $0.01 \times 0.05 \times 0.001 \times 0.001 =$ or nearly 0.

We add these to get the numerator.

For the denominator, where both B and A are arbitrary, we have to *double* sum over both the B and A states being true or false. That's 4 possibilities — so you have to break it down into 4 terms, for each combination of B and A values:

$$\Sigma_{E \in T,F} \Sigma_{A \in T,F} P(B = T, E = e, A = a, J = T, M = T)$$