LECTURE NOTES–UNIT V

UNCERTAINTY & LEARNING FROM OBSERVATIONS

ECS302

ARTIFICIAL INTELLIGENCE

Module V Lecture Notes [*Uncertainty & Learning*]

Syllabus

Uncertain Knowledge: *Uncertainty: Acting under uncertainty, basic probability notation, the axioms of Probability, Inference using full joint distributions, independence, Baye's rule and its use, the wumpus world revisited.* ***Learning:*** *Learning from Observations: Forms of learning, Inductive learning, learning decision trees, ensemble learning. Why Learning Works: Computational learning theory.*

1. *Acting Under Uncertainty*

When an agent knows enough facts about its environment, the logical approach enables it to derive plans that are guaranteed to work. But unfortunately, *agents never have access to the whole truth about their environment.* This is called ***uncertainty*****.**

- For example, an agent in the wumpus world consists of sensors that report only local information; most of the world is not immediately observable. A wumpus agent often will find itself unable to discover which of two squares contains a pit. If those squares are *en route* to the gold, then the agent might have to take a chance and enter one of the two squares.
- The real world is far more complex than the wumpus world. For a logical agent, it might be impossible to construct a complete and correct description of how its actions will work.
- Suppose, for example, that the agent wants to drive someone to the airport to catch a flight and is considering a plan, A90, that involves leaving home 90 minutes before the flight departs and driving at a reasonable speed. Even though the airport is only about 15 miles away, the agent will not be able to conclude with certainty that "Plan Ago will get us to the airport in time." Instead, it reaches the weaker conclusion "Plan Ago will get us to the airport in time, as long as my car doesn't break down or run out of gas, and I don't get into an accident, and there are no accidents on the bridge, and the plane doesn't leave early and . . . ." None of these conditions can be deduced, so the plan's success cannot be inferred.

- The information that the agent has cannot guarantee any of these outcomes for A90, but it can provide some degree of belief that they will be achieved.
- Other plans, such as A120, might increase the agent's belief that it will get to the airport on time, but also increase the likelihood of a long wait. "***The right thing to do-the rational decision—therefore depends on both the relative importance of various goals and the likelihood that,*** *and* ***degree to which, they will be achieved".***

1.1Handling uncertain knowledge:

Here we consider the nature of uncertain knowledge; Let us see a simple-diagnosis example to illustrate the concepts involved. "Diagnosis for medicine". So write rules for dental diagnosis using first-order logic.

$$\forall p \;\; Symptom(p, Toothache) \Rightarrow Disease(p, Cavity)$$

The problem is that this rule is wrong. Not all patients with toothaches have cavities; some of them have gum disease, an abscess, or one of several other problems:

$$\begin{aligned}&\forall p \;\; Symptom(p, Toothache) \Rightarrow \\ &\quad Disease(p, Cavity)\, V\, Disease(p, GumDisease) \vee Disease(p, Abscess)\ldots\end{aligned}$$

In order to make the rule true, we have to add an almost unlimited list of possible causes. We could try turning the rule into a causal rule:

$$\forall p \;\; Disease(p, Cavity) \Rightarrow Symptom(p, Toothache)$$

But this rule is not right either; not all cavities cause pain. The only way to fix the rule is to make it logically exhaustive: i.e., the left-hand side with all the qualifications required for a cavity to cause a toothache. Trying to use first-order logic to cope with a domain like medical diagnosis thus fails for three main reasons:

Laziness: It is too much work to list the complete set of antecedents or consequents needed to ensure an exception less rule and too hard to use such rules.

Theoretical ignorance: Medical science has no complete theory for the domain.

Practical ignorance: Even if we know all the rules, we might be uncertain about a particular patient because not all the necessary tests have been or can be run.

- The connection between toothaches and cavities is just not a logical consequence in either direction. This is typical of the medical domain, as well as most other judgmental domains: law, business, design, automobile repair, gardening, and so on.
- The agent's knowledge *DEGREE OF BELIEF* can at best provide only a ***degree of belief*** in the relevant sentences. Our main tool for dealing *PROBABILITY THEORY* with degrees of belief will be ***probability theory*;** it assigns to each sentence a numerical degree of belief between 0 and 1.

"Probability provides a way of summarizing the uncertainty that comes from our laziness and ignorance"

- This belief could be derived from statistical data-80% of the toothache patients seen so far have had cavities. The 80% summarizes those cases in which all the factors needed for a cavity to cause a toothache are present and other cases in which the patient has both toothache and cavity but the two are unconnected.
- The missing 20% summarizes all the other possible causes of toothache that we are too lazy or ignorant to confirm or deny.
- The sentence is false, while assigning a probability of 1 corresponds to an unequivocal belief that the sentence is true. Probabilities between 0 and 1 correspond to intermediate degrees of belief in the truth of the sentence.
- Thus, probability theory makes the same ontological commitment as logic namely; the facts either do or do not hold in the world. Degree of truth, as opposed to degree of belief, is the subject of **fuzzy logic.**

Evidence: In logic, a sentence such as "The patient has a cavity" is true or false depending on the interpretation and the world; it is true just when the fact it refers to is the case. In probability theory, a sentence such as "The probability that the patient has a cavity is 0.8" is about the agent's beliefs, not directly about the world. These beliefs depend on the percepts that the agent has received to date. These percepts constitute the **evidence** on which probability assertions are based.

- Before the evidence is obtained, we talk about **prior** or **unconditional** probability; after the evidence is obtained, we talk about **posterior** or **conditional** probability.

1.2 Uncertainty and Rational decisions:

To make choices among alternatives, an agent must first have ***preferences*** between the different possible ***outcomes*** of the various plans. A particular outcome is a completely specified state, including such factors as whether the agent arrives on time and the length of the wait at the airport. We will be using ***utility theory*** to represent and reason with preferences. Utility theory says that every state has a degree of usefulness, or utility, to an agent and that the agent will prefer states .with higher utility.

The utility of a state is relative to the agent whose preferences the utility function is supposed to represent. For example, the utility of a state in which White has won a game of chess is obviously high for the agent playing White, but low for the agent p1aying Black.

So, Preferences, as expressed by utilities, are combined with probabilities in the general theory of ra.tiona1 decisions called ***decision theory*****:**

Decision theory = probability theory + utility theory

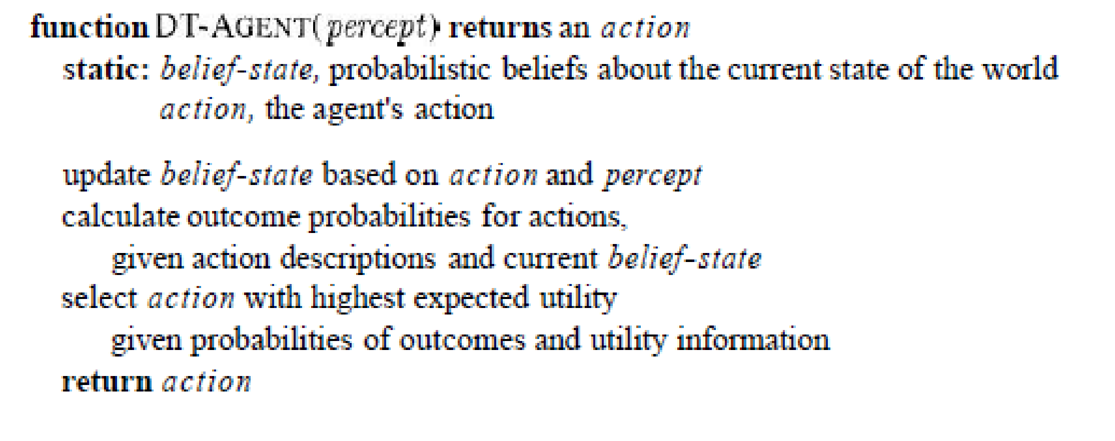
The fundamental idea of decision theory is that *an agent is rational if and only if it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action.* This is called the principle of **Maximum Expected Utility (MEU).**

1.3 Design for a decision-theoretic agent:

Figure 1.3.1 sketches the structure of an agent that uses decision theory to select actions. The agent is identical, at an abstract level, to the logical agent. The primary difference is that the decision-theoretic agent's knowledge of the current state is uncertain; the agent's **belief state** is a representation of the probabilities of all possible actual states of the world.

As time passes, the agent accumulates more evidence and its belief state changes.

Given the belief state, the agent can make probabilistic predictions of action outcomes and hence select the action with highest expected utility.



```
function DT-AGENT(percept) returns an action
   static: belief-state, probabilistic beliefs about the current state of the world
           action, the agent's action

   update belief-state based on action and percept
   calculate outcome probabilities for actions,
       given action descriptions and current belief-state
   select action with highest expected utility
       given probabilities of outcomes and utility information
   return action
```

Figure 1.3.1 A decision-theoretic agent that selects rational actions.

2. Basic Probability Notation

Any notation for describing degrees of belief must be able to deal with two main issues:

1) The nature of the sentences to which degrees of belief are assigned
2) The dependence of the degree of belief on the agent's experience.

Propositions:

Probability theory typically uses a language that is slightly *more expressive than propositional logic.* The basic element of the language is the *random variable*, which can be thought of as referring to a "part" of the world whose "status" is initially unknown.

- For example, *Cavity* might refer to whether my lower left wisdom tooth has a cavity. Random variables play a role similar to that of CSP variables in constraint satisfaction problems and that of proposition symbols in propositional logic. We will always capitalize the names of random variables. For example: P(a) = 1 - P(⌐a)).
- Each random variable has a *domain* of values that it can take on. For example, the domain of *Cavity* might be *(true, fail)*.
- For example, *Cavity* = *true* might represent the proposition that I do in fact have a cavity in my lower left wisdom tooth.

As with CSP variables, random variables are typically divided into *three kinds*, depending on the type of the domain:

- Boolean random variables, such as *Cavity,* have the domain *(true, false).* We will often abbreviate a proposition such as *Cavity* = *true* simply by the lowercase name *cavity.* Similarly, *Cavity* = *false* would be abbreviated by 1 *cavity.*
- Discrete random variables, which include Boolean random variables as a special case, take on values from a *countable* domain. For example, the domain of *Weather* might be *(sunny, rainy, cloudy, snow).* The values in the domain must be mutually exclusive and exhaustive. Where no confusion arises, we: will use, for example, *snow* as an abbreviation for *Weather* = *snow.*
- Continuous random variables take on values from the: real numbers. The domain can be either the entire real line or some subset such as the interval [0, 1]. For example, the proposition X = 4.02 asserts that the random variable .*X* has the exact value 4.02.

Elementary propositions, such as *Cavity* = *true* and *Toothache* =*false,* can be combined to form complex propositions using all the standard logical connectives. For example, *Cavity* = *true A Toothache* =*false* is a proposition to which one may ascribe a degree of belief. As explained in the previous paragraph, this proposition may also be written as *cavity* ∧ *toothache.*

Atomic events:

The notion of an **atomic event** is useful in understanding the foundations of probability theory.

An atomic event is a complete specification of the state of the world about which the agent is uncertain. It can be thought of as an assignment of particular values to all the variables of which the world is composed. For example, if my world consists of only the Boolean variables *Cavity* and *Toothache,* then there are just four distinct atomic events; the proposition

Cavity =*false* ∧ *Toothache* = *true* is one such event.

Atomic events have some important properties:

- They are mutually exclusive-at most one can actually be the case. For example, *cavity* A *toothache* and *cavity* ∧ -*toothache* cannot both be the case.
- The set of all possible atomic events is exhaustive-at least one must be the case. That is, the disjunction of all atomic events is logically equivalent to *true.*
- Any particular atomic event entails the truth or falsehood of every proposition, whether simple or complex. This can be seen by using the standard semantics for logical connectives. For example, the atomic event *cavity* ∧ ⌐ *toothache* entails the truth of *cavity* and the falsehood of *cavity* => *toothache.*
- Any proposition is logically equivalent to the disjunction of all atomic events that entail the truth of the proposition. For example, the proposition *cavity* is equivalent to disjunction of the atomic events *cavity* ∧ *toothache* and *cavity* ∧ ⌐*toothache.*

Prior Probability:

The *unconditional* or *prior probability* associated with a proposition *'a'* is the degree of belief accorded to it in the absence of any other information; it is written as *P (a).* For example, if the prior probability that I have a cavity is 0.1, then

P (*Cavity = true*) = 0.1 or P (*cavity*) = 0.1

It is important to remember that *P (a)* can be used only when there is no other information. As soon as some new information is known, we must reason with the conditional probability of *a* given that new information. Now if we talk about the probabilities of all the possible values of a random variable. In that case, we will use an expression such as **P (***Weather),* which denotes a ***vector*** of values for the probabilities of each individual state of the weather.

So, Instead of writing the four equations

P (Weather = sunny) = 0.7

P (Weather = rain) = 0.2

P (Weather = cloudy) = 0.08

P (Weather = snow) = 0.02.

We may simply write

P (Weather) = (0.7, 0.2, 0.08, 0.02).

This statement defines a *prior probability distribution* for the random variable *Weather*

We will also use expressions such as *P(Weather, Cavity)* to denote the probabilities of all combinations of the values of a set of random variable^ In that case, *P(Weather, Cavity)* can be represented by a 4 x 2 table of probabilities. This is called the *joint probability distribution* of *Weather* and *Cavity.*

- A joint probability distribution that covers this complete set is called the *full joint probability distribution*. For example, if the world consists of just the variables *Cavity, Toothache,* and *Weather,* then the full joint distribution is given by

  P (Cavity, Toothache, Weather)

This joint distribution can be represented as a 2 x 2 x 4 table with 16 entries. So, any probabilistic query can be answered from the full joint distribution. But, *for continuous variables*, it is not possible to write out the entire distribution as a table, because there are infinitely many values. Instead, one usually defines the probability that a random variable takes on some value *x* as a parameterized function of *x*. For example, let the random variable X denote tomorrow's maximum temperature in Berkeley. Then the sentence

$$P(X = x) = U\,[18, 26]\,(x)$$

X is distributed uniformly between 18 and 26 degrees Celsius. Probability distributions for continuous variables are called *probability density functions*. Density functions differ in meaning from discrete distributions. For example, using the temperature distribution given earlier, we find that

$$P\,(X = 20.5) = U\,[18, 26]\,(2\;0.5) == 0\;.125/C.$$

The technical meaning is that the probability that the temperature is in a small region around 20.5 degrees is equal, in the limit, to 0.125 divided by the width of the region in degrees Celsius:

$$\lim_{dx \to 0} P\,(20.5 \leq X \leq 20.5 + dx)/dx = 0.125/C.$$

Conditional probability:

Once the agent has obtained some evidence concerning the previously unknown random variables making up the domain, prior probabilities are no longer applicable. Instead, we use **conditional** or **posterior** probabilities. The notation used is *P (a / b),* where *a* and *b* are any proposition. This is read as "the probability of *a,* given that *all* we know is *b.*"

For example,

P (cavity / toothache) = *0.8*

Indicates that if a patient is observed to have a toothache and no other information is yet available, then the probability of the patient's having a cavity will be 0.8. A prior probability, such as *P (cavity),* can be thought of as a special case of the conditional probability *P (cavity /*), where the probability is conditioned on no evidence. Conditional probabilities can be defined in terms of unconditional probabilities. The defining equation is which holds whenever *P (b)* > 0. This equation can also be written as

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

Which holds whenever *P (b)* > 0. This equation can also be written as

$$\mathbf{P\ (a \wedge b) = P\ (a / b)\ P\ (b)}$$

Which is called the ***product rule*****.** It comes from the fact that, for *a* and *b* to be true, we need *b* to be true, and we also need *a* to be true given *b.* We can also have it the other way;

$$\mathbf{P\ (a \wedge b) = P\ (b / a)\ P\ (a)}$$

We can also use the *P* notation for conditional distributions. *P(X / Y)* gives the values of *P(X* = *xi / Y* = *yj)* for each possible *i, j.* As an example consider applying the product rule to each case where the propositions *a* and *b* assert particular values of *X* and *Y* respectively. We obtain the following equations:

$$\begin{aligned} P(X=x_1 \wedge Y=y_1) &= P(X=x_1|Y=y_1)P(Y=y_1) \\ P(X=x_1\ A\, Y=y_2) &= P(X=x_1|Y=y_2)P(Y=y_2) \end{aligned}$$

We can combine all these into the single equation

$$\mathbf{P(X, Y) = P(X / Y)\ P(Y)}$$

It is wrong, because to view conditional probabilities as if they were logical implications with uncertainty added. For example, the sentence ***P* (*a* / b) = *0.8*** *cannot* be interpreted to mean "whenever *b* holds, conclude that *P (a)* is 0.8." Such an interpretation would be wrong on two counts:

- first, *P(a)* always denotes the prior probability of *a,* not the posterior probability given some evidence;
- Second, the statement ***P* (*a* / b) = *0.8*** is immediately relevant just when *b* is the *only* available evidence. When additional information *c* is available, the degree of belief in *a* is *P (a / b ^ c)*, which might have little relation to *P (a / b).*
- For example, *c* might tell us directly whether *a* is true or false. If we examine a patient who complains of toothache, and discover a cavity, then we have additional evidence *cavity,* and we conclude (trivially) that ***P (cavity / toothache ^ cavity) = 1.0.***

3. *The Axioms Of Probability*

So far, we have defined a syntax for propositions and for prior and conditional probability statements about those propositions. Now we must provide some sort of semantics for probability statements. We begin with the basic axioms that serve to define the probability scale and its endpoints:

1. All probabilities are between 0 and 1. For any proposition a,

$$0 \leq P(a) \leq 1$$

2. Necessarily true (i.e., valid) propositions have probability I, and necessarily false (i.e., unsatisfiable) propositions have probability 0.

$$P(true) = 1 \qquad P(false) = 0 .$$

Next, we need an axiom that connects the probabilities of logically related propositions. The simplest way to do this is to define the probability of a disjunction as follows:

3. The probability of a disjunction is given by

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b) .$$

This rule is easily remembered by noting that the cases where *a* holds, together with the cases where *b* holds, certainly cover all the cases where *a* V *b* holds; but summing the two sets of cases counts their intersection twice, so we need to subtract *Y (a* ∧ b).

These three axioms are often called **Kolmogorov's axioms.**

Using the axioms of probability:

We can derive a variety of useful facts from the basic, axioms. For example, the familiar rule for negation follows by substituting ~*a for* b in axiom 3, giving us:

$$\begin{aligned} P(a \vee \neg a) &= P(a) + P(\neg a) - P(a\, A \neg a) && \text{(by axiom 3 with } b = \neg a) \\ P(true) &= P(a) + P(\neg a) - P(false) && \text{(by logical equivalence)} \\ 1 &= P(a) + P(\neg a) && \text{(by axiom 2)} \\ P(\neg a) &= 1 - P(a) && \text{(by algebra).} \end{aligned}$$

The third line of this derivation is itself a useful fact and can be extended from the Boolean case to the general discrete case. Let the discrete variable D have the domain (dl, . . . , d,). Then it is easy to show that

$$\sum_{i=1}^{n} P(D = d_i) = 1 .$$

The probability of a proposition is equal to the sum of the probabilities of the atomic events in which it holds; that is,

$$P(a) = \sum_{e_i \in \mathbf{e}(a)} P(e_i) .$$

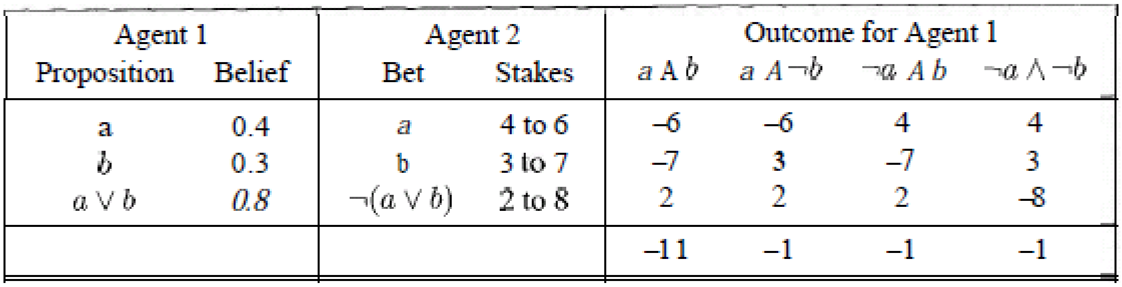
Why the axioms of probability are reasonable:

The axioms of probability can be seen as restricting the set of probabilistic beliefs that an agent can hold. Where a logical agent cannot simultaneously believe A, *B*, and ~ *(A* ∧ *B)* for example. In the logical case, the semantic definition of conjunction means that at least one of the three beliefs just mentioned *must be false in the world,* so it is unreasonable for an agent to believe all three. With probabilities, on the other hand, statements refer not to the world directly, but to the agent's own state of knowledge. Why, then, can an agent not hold the following set of beliefs, which clearly violates axiom 3?

$$P(a) = 0.4 \qquad P(a \wedge b) = 0.0$$
$$P(b) = 0.3 \qquad P(a \vee b) = 0.8$$

Finetti proved something much stronger: I*f Agent* I *expresses a set of degrees of belief that violate the axioms of probability theory then there is a combination of bets by Agent 2 that* guarantees *that Agent I will lose money* every *time.*

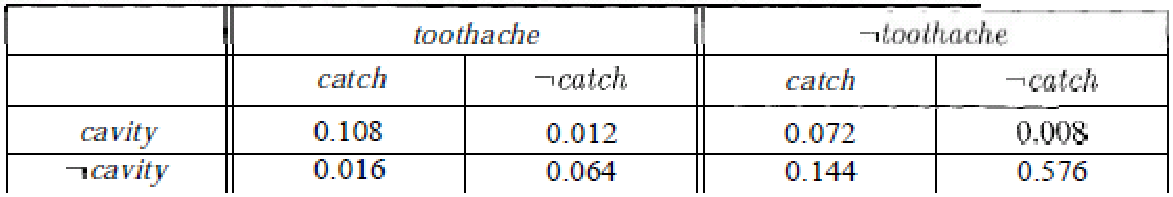
We will not provide the proof of de Finetti's theorem, but we will show an example. Suppose that Agent 1 has the set of degrees of belief from Equation given above. Figure 3.1 shows that if Agent 2 chooses to bet \$4 on a, \$3 on *b*, and \$2 on ~ *(a V b)*, then Agent 1 always loses money, regardless of the outcomes for *a* and *b.*



Agent 1		Agent 2		Outcome for Agent 1			
Proposition	Belief	Bet	Stakes	a A b	a A$\neg b$	$\neg a$ A b	$\neg a \wedge \neg b$
a	0.4	a	4 to 6	–6	–6	4	4
b	0.3	b	3 to 7	–7	3	–7	3
$a \vee b$	*0.8*	$\neg(a \vee b)$	2 to 8	2	2	2	–8
				–11	–1	–1	–1

4. *Inference Using Full Joint Distributions*

Here we will use the full joint distribution as the "knowledge base" from which answers to all questions may be derived. Along the way we will also introduce several useful techniques for manipulating equations involving probabilities. We begin with a very simple example: a domain consisting of just the three Boolean variables *Toothache, Cavity,* and *Catch.* The full joint distribution is a 2 x 2 x 2 table as shown In Figure 4.1.



	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

Figure 4.1 A full joint distribution for the Toothache, Cavity, and Catch world.

Now identify those atomic events in which the proposition is true and add up their probabilities. For example, there are six atomic events in which *cavity V toothache* holds:

$$P(cavity \lor toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

One common task is to extract the distribution over some subset of variables or a single variable. For example, adding the entries in the first row gives the unconditional or marginal probability of *cavity:*

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

This process is called marginalization, or summing out-because the variables other than *Cavity* are summed out. We can write the following general marginalization rule for any sets of variables Y and Z:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{z})$$

That is, a distribution over Y can be obtained by summing out all the other variables from any joint distribution containing Y. A variant of this rule involves conditional probabilities instead of joint probabilities, using the product rule:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}|\mathbf{z})P(\mathbf{z})$$

This rule is called **conditioning.** Marginalization and conditioning will turn out to be useful rules for all kinds of derivations involving probability expressions.

For example, we can compute the probability of a cavity, given evidence of a toothache, as follows:

$$\begin{aligned} P(cavity|toothache) &= \frac{P(cavity\ A\ toothache)}{P(toothache)} \\ &= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6 \end{aligned}$$

Just to check, we can also compute the probability that there is no cavity, given a toothache:

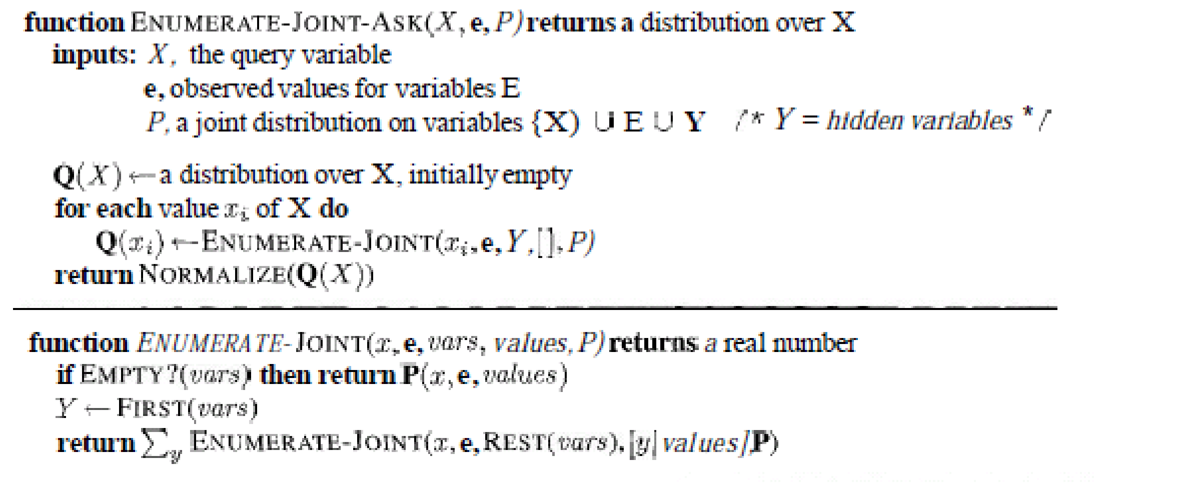
$$P(\neg cavity | toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

In these two calculations the term *1/P (toothache)* remains constant, no matter which value of *Cavity* we calculate. In fact, it can be viewed as a ***normalization*** constant for the distribution *P(Cavity / toothache),* ensuring that it adds up to 1.

We will use *a* to denote such constants. With this notation, we can write the two preceding equations in one:

$$\begin{aligned}\mathbf{P}(Cavity\,(toothache) &= a\mathbf{P}(Cavity, toothache)\\ &= a[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]\\ &= \alpha[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] = \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle\,.\end{aligned}$$

Algorithm for probabilistic inference:



```
function ENUMERATE-JOINT-ASK($X$, e, $P$) returns a distribution over X
   inputs: $X$, the query variable
           e, observed values for variables E
           $P$, a joint distribution on variables {X} ∪ E ∪ Y   /* $Y$ = hidden variables */

   Q($X$) ← a distribution over X, initially empty
   for each value $x_i$ of X do
      Q($x_i$) ← ENUMERATE-JOINT($x_i$, e, $Y$, [], $P$)
   return NORMALIZE(Q($X$))

function ENUMERATE-JOINT($x$, e, $vars$, $values$, $P$) returns a real number
   if EMPTY?($vars$) then return P($x$, e, $values$)
   $Y$ ← FIRST($vars$)
   return $\sum_y$ ENUMERATE-JOINT($x$, e, REST($vars$), [$y$| $values$], P)
```

Figure 4.2 An algorithm for probabilistic inference by enumeration of the entries in a full joint: distribution.

Given the full joint distribution to work with, ENUMERATE-JOINT-ASK is a complete algorithm for answering probabilistic queries for discrete variables. It does not scale well, however: For a domain described by n Boolean variables, it requires an input table of size 0(2 *pow n*) and takes

0(2 pow n) time to process the table. So, the full joint distribution in tabular form is not a practical tool for building reasoning systems.

5. *Independence*

Let us expand the full joint distribution in Figure 13.3 by adding a fourth variable, *Weather* .The full joint distribution then becomes *P(Toothache, Catch, Cavity, Weather),* which has *32* entries (because *Weather* has four values). It contains four "editions" of the table shown in Figure 4.1, one for each kind of weather. Here we may ask what relationship these editions have to each other and to the original three-variable table. For example, how are *P(toothache, catch, cavity, Weather = cloudy)* and *P(toothache, catch, cavity)* related?

To answer this question is to use the product rule: *P(toothache, catch, cavity, Weather = cloudy)*

$$= P\ (Weather = cloudy\ /\ toothache,\ catch,\ cavity)\ P\ (toothache,\ catch,\ cavity).$$

One should not imagine that one's dental problems influence the weather. Therefore, the following assertion seems reasonable:

$$P(\ Weather = cloudy\ /\ toothache,\ catch,\ cavity) = P\ (\ Weather = cloudy)\text{---------}(1)$$

From this, we can deduce;

$$P(toothache,\ catch,\ cavity,\ weather = cloudy)\ \ = P(\ Weather = cloudy)P(toothache,\ catch,\ cavity).$$

Similar equation exists for *every entry* in *P(Toothache, Catch, Cavity, Weather).* In fact, we can write the general equation;

$$P(Toothache,\ Catch,\ Cavity,\ Weather) = P(Toothache,\ Catch,\ Cavity)P(\ Weather)\ .$$

Thus, the 32-element table for four variables can be constructed from one 8-element table and one four-element table. This decomposition is illustrated schematically in Figure 5.1(a). The property we used in writing Equation (1) is called "***independence".***

Independence between propositions *a* and *b* can be written as

$$P\ (\ a\ /\ b\ )= P\ (\ a\ ) \text{ or } P(\ b\ /\ a\ )= P\ (\ b\ ) \text{ or } P\ (\ a\ A \text{ b}) = P\ (\ a\ )\ P\ (\ b\ ).$$

Independence between variables X and Y can be written as follows (again, these are all equivalent):

$$P\ (\ X\ /\ Y\ )= P\ (\ X\ ) \ or\ P\ (\ Y\ /\ X\ )= P(Y) \quad or \quad P\ (\ X\ ,Y\ )= P\ (\ X\ )\ P\ (\ Y\ ).$$

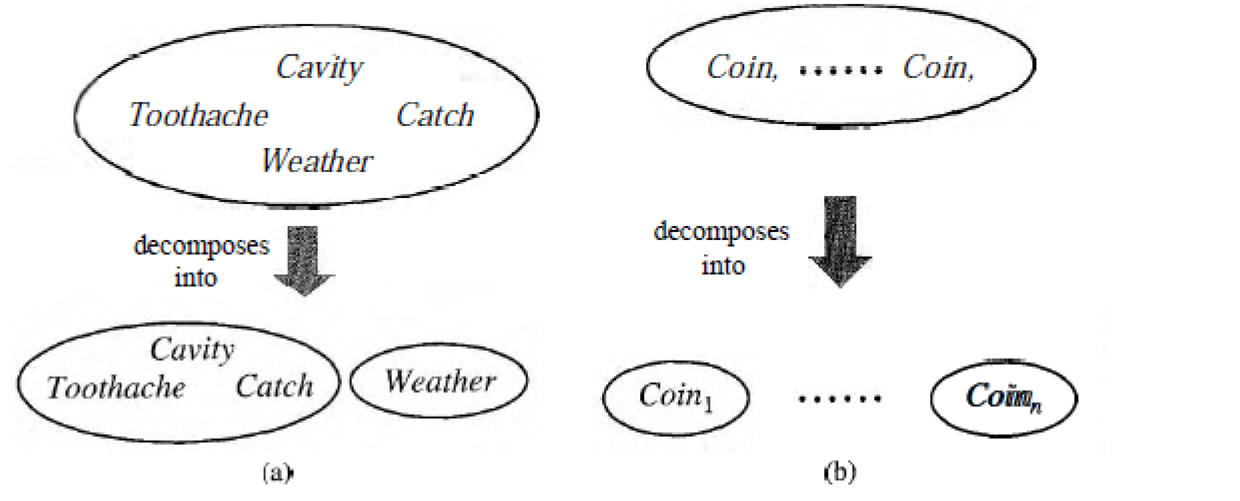


Figure 5.1 Two examples of factoring a large joint distribution into smaller distributions, using absolute independence. (a) Weather and dental problems are independent. (b) Coin flips are independent.

6. Baye's rule and its use

We defined the **product rule** and pointed out that it can be written in two forms because of the commutativity of conjunction:

$$P(a \wedge b) = P(a|b)P(b)$$
$$P(a \wedge b) = P(b|a)P(a)$$

Equating the two right-hand sides and dividing by *P(a),* we get

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

This equation is known as **Bayes' rule** (also Bayes' law or Bayes' theorem) .This is simple equation underlies all modern AI systems for probabilistic inference. The more general case of multi valued variables can be written in the P notation as;

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)}$$

Where again this is to be taken as representing a set of equations, each dealing with specific values of the variables. We will also have occasion to use a more general version conditionalized on some background evidence **e:**

$$\mathbf{P}(Y|X,\mathbf{e}) = \frac{\mathbf{P}(X|Y,\mathbf{e})\mathbf{P}(Y|\mathbf{e})}{\mathbf{P}(X|\mathbf{e})}$$

Applying Bayes' rule: The simple case:

It requires three terms-a conditional probability and two unconditional probabilities-just to compute one conditional probability. Bayes' rule is useful in practice because there are many cases where we do have good probability estimates for these three numbers and need to compute the fourth. In a task such as medical diagnosis, we often have conditional probabilities on causal relationships and want to derive a diagnosis. A doctor knows that the disease meningitis causes the patient to have a stiff neck, say, 50% of the time. The doctor also knows some unconditional facts: the prior probability that a patient has meningitis is 1150,000, and the prior probability that any patient has a stiff neck is 1120. Letting *s* be the proposition that the patient has a stiff neck and *m* be the proposition that the patient has meningitis, we have;

$$\begin{aligned} P(s|m) &= 0.5 \\ P(m) &= 1/50000 \\ P(s) &= 1/20 \\ P(m|s) &= \frac{P(s|m)P(m)}{P(s)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002\,. \end{aligned}$$

That is, we expect only 1 in 5000 patients with a stiff neck to have meningitis. Notice that, even though a stiff neck is quite strongly indicated by meningitis (with probability 0.5), the probability of meningitis in the patient remains small. This is because the prior probability on stiff necks is much higher than that on meningitis.

The same process can be applied when using Bayes' rule. We have

$$\mathbf{P}(M|s) = a\,\langle P(s|m)P(m), P(s|\neg m)P(\neg m)\rangle$$

Thus, in order to use this approach we need to estimate *P(s /~ m)* instead of *P(s)*

The general form of Bayes' rule with normalization is

P(Y / X) = a P(X /Y) P(Y), where *a* is the normalization constant needed to make the entries in *P(Y / X)* sum to 1.

Using Bayes' rule: Combining evidence:

We have seen that Bayes' rule can be useful for answering probabilistic queries conditioned on one piece of evidence-for example, the stiff neck. In particular, we have argued that probabilistic information is often available in the form *P(effect / cause)*. What happens when we have two or more pieces of evidence? For example, what can a dentist conclude if her nasty steel probe catches in the aching tooth of a patient? If we know the full joint distribution, one can read off the answer:

$$\mathbf{P}(Cavity\ (toothacheA\ catch) = a!\langle 0.108, 0.016\rangle \approx \langle 0.871, 0.129\rangle$$

We know, however, that such an approach will not scale up to larger numbers of variables. We can try using Bayes' rule to reformulate the problem:

$$\mathbf{P}(Cavity|toothache\ A\ catch) = \alpha\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity)$$

For this reformulation to work, we need to know the conditional probabilities of the conjunction *toothache* A *catch* for each value of *Cavity.* That might be feasible for just two evidence variables, but again it will not scale up. If there are n possible evidence variables (X rays, diet, oral hygiene, etc.), then there are *2n* possible combination so f observed values for which we would need to know conditional probabilities. We might as well go back to using the full joint distribution.

Rather than taking this route, we need to find some additional assertions about the domain that will enable us to simplify the expressions. The notion of **independence** provides a clue, but needs refining. It would be nice if *Toothache* and *Catch* were independent, but 'they are not: if the probe catches in the tooth, it probably has a cavity and that probably causes a toothache. These variables *are* independent.

Mathematically, this property is written as;

$$\mathbf{P}(toothache\ A\ catch|Cavity) = \mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)$$

This equation expresses the **conditional independence** of *toothache* and *catch* given *Cavity*. We can plug it into above equation to obtain the probability of a cavity:

$$\mathbf{P}(Cavity)toothache\ A\ catch) = a!\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity).$$

The general definition of conditional independence of two variables X and *Y*, given a third variable *Z* is

$$\mathbf{P}(X, Y|Z) = \mathbf{P}(X|Z)\mathbf{P}(Y|Z)$$

In the dentist domain, for example, it seems reasonable to assert conditional independence of the variables *Toothache* and *Catch*, given *Cavity:*

$$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity).$$

Which asserts independence only for specific values of *Toothache* and *Catch?* As with absolute independence in

Equation

$$\mathbf{P}(X|Y, Z) = \mathbf{P}(X|Z) \quad \text{and} \quad \mathbf{P}(Y|X, Z) = \mathbf{P}(Y|Z)$$

It turns out that the same is true for conditional independence assertions. For example, given the assertion in Equation, We can derive decomposition as follows:

$$\begin{aligned}&\mathbf{P}(Toothache, Catch, Cavity)\\ &= \mathbf{P}(Toothache, Catch|Cavity)\mathbf{P}(Cavity) \quad \text{(product rule)}\\ &= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)\end{aligned}$$

In this way, the original large table is decomposed into three smaller tables.

7. *The Wumpus World Rivisited*

Uncertainty arises in the wumpus world because the agent's sensors give only partial, local information about the world. For example, Figure 13.6 shows a situation in which each of the three reachable squares-[1,3], [2,2], and [3,1]-might contain a pit. Pure logical inference can conclude nothing about which square is most likely to be safe, so a logical agent might be forced to choose randomly. We will see that a probabilistic agent can do much better than the logical agent.

Our aim will be to calculate the probability that each of the three squares contains a pit. (For the purposes of this example, we will ignore the wumpus and the gold.) The relevant properties of the wumpus world are that

(1) a pit causes breezes in all neighboring squares, and

(2) each square other than [1,1] contains a pit with probability 0.2.

The first step is to identify the set of random variables we need: As in the propositional logic case, we want one Boolean variable *Pij* for each square, which is true iff square [i, *j]* actually contains a pit. We also have Boolean variables *Bij* that are true iff square, *[i, j]* is breezy; we include these variables only for the observed squares-in this case, [1,1], [1,2], and [2,1]. The next step is to specify the full joint distribution, *P (Pl,1,..., P4,4,B1,1,B1,2,B2,1)* .Applying the product rule, we have

$$\mathbf{P}(P_{1,1},\ldots,P_{4,4},B_{1,1},B_{1,2},B_{2,1}) = \\ \mathbf{P}(B_{1,1},B_{1,2},B_{2,1}\,|\,P_{1,1},\ldots,P_{4,4})\mathbf{P}(P_{1,1},\ldots,P_{4,4})$$

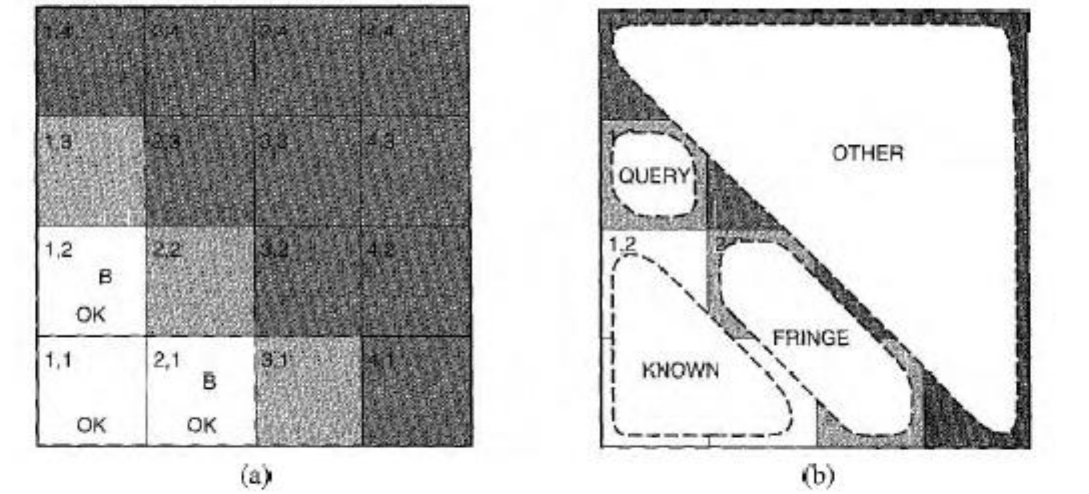


Figure 7.1 (a) After finding a breeze in both [1,2] and [2,1], the agent is stuck-there is no safe place to explore. (b) Division of the squares into Known, Fringe, and Other, for a query about [1, 3].

This decomposition makes it very easy to see what the joint probability values should be. The first term is the conditional probability of a breeze configuration, given a pit configuration; this is 1 if the breezes are adjacent to the pits and 0 otherwise. The second term is the prior probability of a pit configuration. Each square contains a pit with probability 0.2, independently of the other squares; hence,

$$\mathbf{P}(P_{1,1},\ldots,P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j})$$

In the situation in Figure 13.6(a), the evidence consists of the observed breeze (or its absence) in each square that is visited, combined with the fact that each such square contains no pit. We'll abbreviate these facts as

$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$ and $known, = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$.

. We are interested in answering queries such as *P(P1,3 / known,* b) : how likely is it that [1,3] contains a pit, given the observations so far?

To answer this query, we can follow the standard approach suggested by Equation and implemented in the ENUMERATE-JOINT-ASK, namely, summing over entries from the full joint distribution. Let *Unknown* be a composite variable consisting of the *Pi, j* variables for squares other than the *Known* squares and the query square [1,3]. Then, by Equation, we have

$$\mathbf{P}(P_{1,3} | known, \mathrm{b}) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

Let *Fringe* be the variables (other than the query variable) that are adjacent to visited squares, in this case just [2,2] and [3,1]. Also, let *Other* be the variables for the other unknown squares; in this case, there are 10 other squares, as shown in Figure 7.1(b). The key insight is that the observed breezes are *conditionally independent* of the other variables, given the known, fringe, and query variables. The rest is, as they say, a small matter of algebra. To use the insight, we manipulate the query formula into a form in which the breezes are conditioned on all the other variables, and then we simplify using conditional independence:

$$\begin{aligned}
&\mathbf{P}(P_{1,3} | known, b) \\
&= \alpha \sum_{unknown} \mathbf{P}(b | P_{1,3}, known, unknown)\mathbf{P}(P_{1,3}, known, unknown) \\
&\qquad\qquad \text{(by the product rule)} \\
&= a \sum_{frznge} \sum_{other} \mathbf{P}(b | known, P_{1,3}, fringe, other)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b | known, P_{1,3}, fringe)\mathbf{P}(P_{1,3}, known, fringe, other) ,
\end{aligned}$$

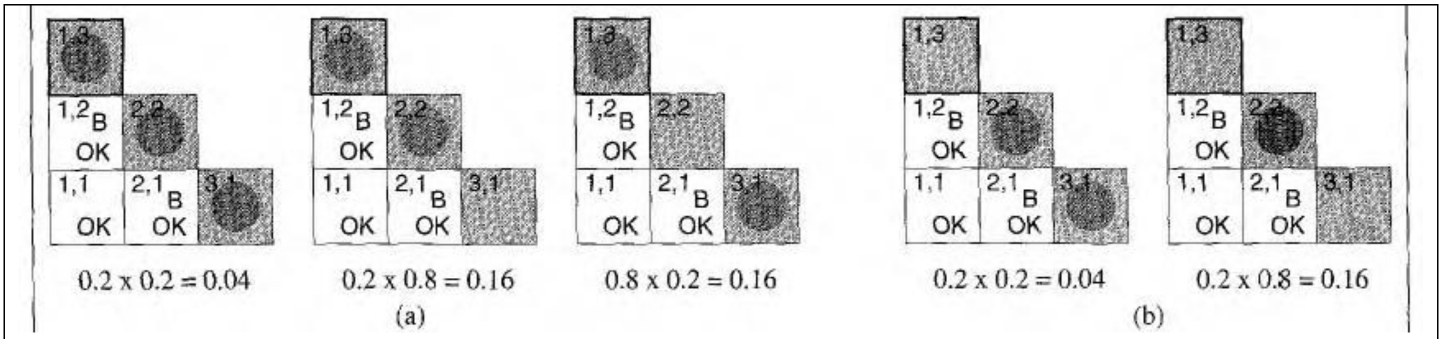


Figure 7.2 Consistent models for the fringe variables P2,2 and P 3,1 showing P(fringe) for each model: (a) three models with = true showing two or three pits, and (b) two models with P1, 3 =false showing one or two pits.

Where the final step uses conditional independence. Now, the first term in this expression does not depend on the other variables, so we can move the summation inwards:

$$\mathbf{P}(P_{1,3}|known, b) \\ = a \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other)$$

By independence, the prior term can be factored, and then the terms can be reordered:

$$\mathbf{P}(P_{1,3}|known, b) \\ = a \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(fringe)P(other) \\ = a\, P(known)\mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \sum_{other} P(other) \\ = a'\, \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe)\,,$$

where the last step folds $P(known)$ into the normalizing constant and uses the fact that $\sum_{other} P(other)$ equals 1.

To get *efticient* solutions, independence and conditional independence relationships can be used to simplify the summations required.