**Unit III Uncertainty CS6659 Artificial Intelligence**

**ACTING UNDER UNCERTAINTY**

1. When an agent knows enough facts about its environment, the logical approach enables it to derive plans that are guaranteed to work.
2. This is a good thing. Unfortunately, *agents almost never have access to the whole truth about their environment.* Agents must, therefore, act under **uncertainty.**
3. If a logical agent cannot conclude that any particular course of action achieves its goal, then it will be unable to act.
4. Conditional planning can overcome uncertainty to some extent, but only if the agent's sensing actions can obtain the required information and only if there are not too many different contingencies.
5. Another possible solution would be to endow the agent with a simple but incorrect theory of the world that does enable it to derive a plan;presumably, such plans will work ***most*** of the time, but problems arise when events contradict the agent's theory.
6. Moreover, handling the tradeoff between the accuracy and usefulness of the agent's theory seems itself to require reasoning about uncertainty.
7. 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.

**Handling uncertain knowledge**

In this section, we look more closely at the nature of uncertain knowledge.

We will use a simple diagnosis example to illustrate the concepts involved.

Diagnosis-whether for medicine, automobile repair, or whatever-is a task that almost always involves uncertainty.

Let us try to write rules for dental diagnosis using first-order logic, so that we can see how the logical approach breaks down.

Consider the following rule:



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:



Unfortunately, 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:



The only way to fix the rule is to make it logically exhaustive: to augment 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:

1. **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.
2. **Theoretical ignorance:** Medical science has no complete theory for the domain.
3. **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, dating, and so on.
* The agent's knowledge can at best provide only a **degree of belief** in the relevant sentences. Our main tool for dealing with degrees of belief will be **probability theory,** which assigns to each sentence a numerical degree of belief between 0 and 1.

1. *Probability provides a way of* ***summarizing*** *the uncertainty that comes from our laziness and ignorance.*
2. But we believe that there is, say, an 80% chance-that is, a probability of 0.8-that the patient has a cavity if he or she has a toothache.
3. This belief could be derived from statistical data-80% of the toothache patients seen so far have had cavities-or from some general rules, or from a combination of evidence sources.
4. The missing 20% summarizes all the other possible causes of toothache that we are too lazy or ignorant to confirm or deny.

* Assigning a probability of 0 to a given sentence corresponds to an unequivocal belief that 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.
* The sentence itself is in *fact* either true or false.
* It is important to note that a degree of belief is different from a degree of truth.
* A probability of 0.8 does not mean "80% true" but rather an 80% degree of belief-that is, a fairly strong expectation.
* Thus, probability theory makes the same ontological commitment as logic namely, that 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.**

1. 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.
2. 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.
3. These beliefs depend on the percepts that the agent has received to date. These percepts constitute the **evidence** on which probability assertions are based.
4. Just as entailment status can change when more sentences are added to the knowledge base, probabilities can change when more evidence is acquired.
5. All probability statements must therefore indicate the evidence with respect to which the probability is being assessed.
6. As the agent receives new percepts, its probability assessments are updated to reflect the new evidence.
7. Before the evidence is obtained, we talk about **prior** or **unconditional** probability; after the evidence is obtained, we talk about **posterior** or **conditional** probability.
8. In most cases, an agent will have some evidence from its percepts and will be interested in computing the posterior probabilities of the outcomes it cares about.

**Uncertainty and rational decisions**

1. The presence of uncertainty radically changes the way an agent makes decisions.
2. A logical agent typically has a goal and executes any plan that is guaranteed to achieve it.
3. An action can be selected or rejected on the basis of whether it achieves the goal, regardless of what other actions might achieve.
4. To make such choices, an agent must first have **preferences** between the different possible **outcomes** of the various plans.
5. 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.
6. We will be using **utility theory** to represent and reason with preferences.
7. The term **utility** is used here in the sense of "the quality of being useful," not in the sense of the electric company or water works.
8. 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.
9. The utility of a state is relative to the agent whose preferences the utility function is supposed to represent.
10. **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 playing Black.
11. Or again, some players (including the authors) might be happy with a draw against the world champion, whereas other players (including the former world champion) might not.
12. A utility function can even account for altruistic behavior, simply by including the welfare of others as one of the factors contributing to the agent's own utility.

* Preferences, as expressed by utilities, are combined with probabilities in the general theory of rationa1 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).

**Design for a decision-theoretic agent**

* The structure of an agent that uses decision theory to select actions shows in figure.
* 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.



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

**The Axioms of Probability**

We have defined 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,



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



1. The probability of a disjunction is given by



* 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 P (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 la for b in axiom 3, giving us:



Let the discrete variable D have the domain (d1.........., dn). Then it is easy to show (Exercise) that



* That is, any probability distribution on a single variable must sum to It is also true that any joint probability distribution on any set of variables must sum to 1: this can be seen simply by creating a single megavariable whose domain is the cross product of the domains of the original variables.
* Recall that any proposition a is equivalent to the disjunction of all the atomic events in which a holds; call this set of events e(a).
* Recall also that atomic events are mutually exclusive, so the probability of any conjunction of atomic events is zero, by axiom 2.
* Hence, from axiom 3, we can derive the following simple relationship: The probability of a proposition is equal to the sum of the probabilities of the atomic events in which it holds; that is,



* This equation provides a simple method for computing the probability of any proposition, given a full joint distribution that specifies the probabilities of all atomic events.

**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.
* 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?

equation (1)

* Here, we give one argument for the axioms of probability, first stated in 193 1 by Bruno de Finetti.
* The key to de Finetti's argument is the connection between degree of belief and actions.
* The idea is that if an agent has some degree of belief in a proposition *a,* then the agent should be able to state odds at which it is indifferent to a bet for or against *a.*
* Think of it as a game between two agents: Agent 1 states "my degree of belief in event *a* is 0.4." Agent 2 is then free to choose whether to bet for or against *a,* at stakes that are consistent with the stated degree of belief.
* That is, Agent 2 could choose to bet that *a* will occur, betting $4 against Agent 1's $6. Or Agent 2 could bet $6 against $4 that A will not occur.
* If an agent's degrees of belief do not accurately reflect the world, then you would expect that it would tend to lose money over the long run to an opposing agent whose beliefs more accurately reflect the state of the world.
* But de Finetti proved something much stronger: *ZfAgent* 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.*
* So if you accept the idea that an agent should be willing to "put its money where its probabilities are," then you should accept that it is irrational to have beliefs that violate the axioms of probability.
* Suppose that Agent 1 has the set of degrees of belief from Equation (1). Figure shows that if Agent 2 chooses to bet $4 on a, $3 on b, and $2 on ⌐(a˅b), then Agent 1 always loses money, regardless of the outcomes for a and b.



* Because Agent 1 has inconsistent beliefs, Agent 2 is able to devise a set of bets that guarantees a loss for Agent 1, no matter what the outcome of *a* and b.

**Inference using full joint distribution:**

* A simple method for **probabilistic inference**-that is, the computation from observed evidence of posterior probabilities for query propositions.
* We will use the full joint distribution as the "knowledge base" from which answers to all questions may be derived.
* We begin with a very simple example: a domain consisting of just the three Boolean variables Toothache, Cavity, and Catch (the dentist's nasty steel probe catches in my tooth).
* The full joint distribution is a 2 x 2 x 2 table as shown in Figure



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

* Notice that the probabilities in the joint distribution sum to 1, as required by the axioms of probability.
* For Example, there are six atomic events in which cavity V toothache holds: 
* One particularly 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:



* 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:



* 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:



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

**Conditional probabilities** can be found by first using Equation P(a|b) = P(a˄b)/ P(b) to obtain an expression in terms of unconditional probabilities and then evaluating the expression from the full joint distribution.



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



* Notice that in these two calculations the term l/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.
* Throughout the chapters dealing with probability, we will use a to denote such constants. With this notation, we can write the two preceding equations in one:



* Normalization will turn out to be a useful shortcut in many probability calculations.
* From the example, we can extract a general inference procedure.
* We will stick to the case in which the query involves a single variable.
* We will need some notation: let X be the query variable (Cavity in the example), let E be the set of evidence variables (just Toothache in the example), let e be the observed values for them, and let Y be the remaining unobserved variables (just Catch in the example).
* The query is P(X|e) and can be evaluated as



* Where the summation is over all possible ys (i.e., all possible combinations of values of the unobserved variables Y).
* Notice that together the variables X, E, and Y constitute the complete set of variables for the domain, so P(X, e, y) is simply a subset of probabilities from the full joint distribution.
* It loops over the values of X and the values of Y to enumerate all possible atomic events with e fixed, add up their probabilities from the joint table, and normalizes the results.

**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.
* For a domain described by n Boolean variables, it requires an input table of size 0(2n) and takes *0(2n)* time to process the table.
* In a realistic problem, there might be hundreds or thousands of random variables to consider, not just three.
* It quickly becomes completely impractical to define the vast numbers of probabilities required-the experience needed in order to estimate each of the table entries separately simply cannot exist.
* For these reasons, the full joint distribution in tabular form is not a practical tool for building reasoning systems.
* Instead, it should be viewed as the theoretical foundation on which more effective approaches may be built.

**BA.YER'S R ULE AND ITS USE**

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



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



* This equation is known as **Bayes' rule** (also Bayes' law or Bayes' theorem).
* This simple equation underlies all modern A1 systems for probabilistic inference.
* The more general case of multivalued variables can be written in the P notation as



Where again this is to be taken as representing a set of equations, each dealing with specific values of the variables.



**Applying Bayes' rule: The simple case**

* On the surface, Bayes' rule does not seem very useful. 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 1/50,000, and the prior probability that any patient has a stiff neck is 1/20.
* Letting *s* be the proposition that the patient has a stiff neck and m be the proposition that the patient has meningitis, we have



* 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.
* A process by which one can avoid assessing the probability of the evidence (here, *P(s))* by instead computing a posterior probability for each value of the query variable (here, *m* and ⌐*m)* and then normalizing the results.
* The same process can be applied when using Bayes' rule. We have
* 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

****

Where *a* is the normalization constant needed to make the entries in *P(Y | X)* sum to 1.

**Using Bayes' rule: Combining evidence**

In particular, we have argued that probabilistic information is often available in the form *P (effect | cause).*

****

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:

****

* For this reformulation to work, we need to know the conditional probabilities of the conjunction *toothache* A *catch* for each value of *Cavity.*
* These variables *are* independent, however, *given the presence or the absence of a cavity.*
* Each is directly caused by the cavity, but neither has a direct effect on the other: toothache depends on the state of the nerves in the tooth, whereas the probe's accuracy depends on the dentist's skill, to which the toothache is irrelevant.
* Mathematically, this property is written as

****

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

****

* Now the information requirements are the same as for inference using each piece of evidence separately: the prior probability *P(Cavity)* for the query variable and the conditional probability of each effect, given its cause.
* The general definition of conditional independence of two variables X and Y , given a third variable Z is



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



* Notice that this assertion is somewhat stronger than Equation (13.13), which asserts independence only for specific values of *Toothache* and *Catch.* As with absolute independence in Equation (13.8), the equivalent forms.

can also be used.

* Absolute independence assertions allow a decomposition of the full joint distribution into much smaller pieces. It turns out that the same is true for conditional independence assertions.
* For example, given the assertion in Equation (13.14), we can derive a decomposition as follows:



* In this way, the original large table is decomposed into three smaller tables.
* The original table has seven independent numbers (23 - 1, because the numbers must sum to 1). The smaller tables contain five independent numbers (2 x (21 - 1) for each conditional probability distribution and *21* - 1 for the prior on *Cavity).*
* This might not seem to be a major triumph, but the point is that, for n symptoms that are all conditionally independent given *Cavity,* the size of the representation grows as *O(n)* instead of *O(2").*
* Thus, *conditional independence assertions can allow probabilistic systems to scale up; moreover; they are much more commonly available than absolute independence assertions.*
* Conceptually, *Cavity* **separates** *Toothache* and *Catch* because it is a direct cause of both of them.
* The decomposition of large probabilistic domains into weakly connected subsets via conditional independence is one of the most important developments in the recent history of AI.
* The dentistry example illustrates a commonly occurring pattern in which a single cause directly influences a number of effects, all of which are conditionally independent, given the cause.
* The full joint distribution can be written as



* Such a probability distribution is called a **naive Bayes** model-"naive" because it is often used (as a simplifying assumption) in cases where the "effect" variables are *not* conditionally independent given the cause variable.
* (The naive Bayes model is sometimes called a **Bayesian classifier,** a somewhat careless usage that has prompted true Bayesians to call it IDIOT BAYES the **idiot Bayes** model.) In practice, naive Bayes systems can work surprisingly well, even when the independence assumption is not true.