Artificial Intelligence

Uncertainty

Christopher Simpkins

Kennesaw State University

Acting Under Uncertainty

Logical agents maintain a belief state and generate contingency plans that account for every possibility. This breaks down for nontrivial problems due to:

- Exhaustivity of belief state, which must contain all possible states, even unlikely states.
- Exhaustivity of contingency plan, which must account for every possible, however unlikely, action outcome.
- Unsatisfiabilty. There may be no guaranteed plan to achieve the goal. If we must act anyway, how do we choose the best plan?
- Qualification problem. The closed-world assumption allows us to simplify logical environment specifications for simple domains, but real-world domains contain far more detail which must be accounted for.

From Logic to Probability Theory

Consider:

$$Toothacke \implies Cavity$$

Not all patients with toothaches have cavities. How about:

$$Toothache \implies Cavity \vee GumProblem \vee Abscess \ldots$$

We'd need large list after the dots. Try causal direction:

$$Cavity \implies Toothache$$

But not all cavities hurt. Logic fails to deal with complex domains like medical diagonosis due to:

- **Exhaustivity** (laziness). Too many antecedents or consequents to list.
- **Theoretical ignorance**. Rare to have complete logical theory of nontrivial domain.
- **Practical ignorance**. Even with a complete domain theory, hard to measure all the necessary inputs (medical tests, etc.)

We need to deal with uncertain knowledge, with **degrees of belief**. For that, we use probability theory.

Probability Theory as Knowledge Representation

- Ontological commitments of logic and probability theory same:
  - World composed of facts that do or do not hold in any particular case.
- Epistemological commitments differ:
  - Logic assigns true, false, or no opinion to each sentence.
  - Probability theory assignesnumerical degree of belief between 0 (for sentences that are certainly false) and 1 (certainly true) to each sentence.

Probability theory solves the qualification problem by summarizing the uncertainty stemming from our laziness and ignorance.

What do probabilities mean?

Probabilities are about our uncertain knowledge.

"80% chance (probability 0.8) patient with a toothache has a cavity."

- Out of all the situations that are indistinguishable from the current situation *as far as our knowledge goes*, the patient will have a cavity in 80% of them.
- Belief could come from statistical data, or domain theory, or combination of sources.

> *There is no uncertainty in the actual world. The patient either has a cavity or does not. Probabilities refer to our knowledge of the world state, not the actual world state.*

If our knowledge changes, e.g., we find out patient has history of gum disease, we make a different statement.

"Given patient has a history of gum disease and a toothache, 40% chance patient has gum disease."

Rational Decisions

How do we choose between different plans that acheive the goal?

- Each plan has a likelihood of success, e.g.:
  - Plan A has a 95% chance of achieving a goal state.
- Each plan may lead to goal states with different outcomes, each of which is a goal state.
  - An **outcome** of an action is a completely specified state, which includes elements that are not part of the goal.

Elements of outcomes can be more or less desirable – more or less *useful* – to a given agent.

- Utility is the quality of being useful.
- That "usefulness" may simply be pleasure, or even altruism.

Utility theory associates a utility value with each possible outcome, which induces a preference ordering over outcomes.

> *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.*

Decision-Theoretic Agents

Decision theory = probability theory + utility theory

function DT-AGENT(*percept*) **returns** an *action*
  persistent: *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*

Each action, a, leads to a probability distribution over outcomes, or "result states":

$$\sum_{s \in S} Pr(s) = 1$$

And each outcome state has a utility. A rational decision-theoretic agent chooses the action that maximize expected utility, that is:

$$\underset{a}{\operatorname{argmax}} \sum_{s \in S} Pr(s)U(s)$$

KENNESAW STATE UNIVERSITY

Probability Models

In logic we have satisfiabilty: $M(\alpha)$ is the set of all possible worlds in which sentence α is true.

In probability theory the set of possible worlds is a **sample space**, Ω, which must be mutually exclusive and exhaustive (one and only one possible world must be the case).

$$0 \leq Pr(\omega) \leq 1 \text{ for every } \omega \text{ and } \sum_{\omega \in \Omega} Pr(\omega) = 1$$

Each ω is a possible world. Probability textbooks typically call these outcomes. For example, each of these pairs is a $\omega \in \Omega$ for the roll of two dice:

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	(4, 4)	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

In probability textbooks, an **experiment** leads to a sample space. Here, the set of possible worlds comes from the task environment.

Propositions and Events

An **event**, which we denote with ϕ here, is a set of possible worlds, some subset of Ω, or set of ω, e.g., the event "doubles" contains the 6 boxed elements ω:

$\boxed{(1,1)}$	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	$\boxed{(2,2)}$	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	$\boxed{(3,3)}$	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	$\boxed{(4,4)}$	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	(4, 5)	$\boxed{(5,5)}$	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	$\boxed{(6,6)}$

A **proposition** is an event expressed in a formal language; specifically, for each proposition, the corresponding set contains just those possible worlds in which the proposition holds.

The probability associated with a proposition is defined to be the sum of the probabilities of the worlds in which it holds:

$$\text{For any proposition } \phi, Pr(\phi) = \sum_{\omega \in \phi} Pr(\omega)$$

Example: $Pr(Doubles) = Pr((1,1)) + \cdots + Pr((6,6)) = \frac{1}{36} + \cdots + \frac{1}{36} = \frac{1}{6}$

KENNESAW STATE UNIVERSITY

Prior and Conditional Probabilities

An **unconditional** or **prior** probability is a degree of belief in a proposition with no other information.

A **conditional** or **posterior** probability is a degree of belief in a proposition given some relevant information.

Definition of conditional probability:

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

From that definition we get the **product rule**:

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

Note that $Pr(a \wedge b)$ can also be written $Pr(a, b)$ or $Pr(ab)$.

Conditional probability is not implication.

The assertion

$$Pr(cavity|toothache) = 0.6$$

does not mean "Whenever toothache is true, conclude that cavity is true with probability 0.6"

It means "Whenever toothache is true *and we have no further information*, conclude that cavity is true with probability 0.6."

If we had the further information that the dentist found no cavities, certainly not the case that cavity is true with probability 0.6; instead we have

$$Pr(cavity|toothache \wedge \neg cavity) = 0$$

Probability Assertions

A **random variable**, which begins with capital letter, is a function mapping from possible worlds Ω to a set of possible values the variable can take.

- $Die_1 : \{1, \ldots, 6\} \rightarrow \{1, \ldots, 6\}$
- $Doubles : \{1, \ldots, 6\} \times \{1, \ldots, 6\} \rightarrow \{true, false\}$
- $Weather : \{sun, rain, cloud, snow\} \rightarrow \{sun, rain, cloud, snow\}$

Individual values are written in lowercase and often abbreviated.

- $Pr(sun)$ stands for $Pr(Weather = sun)$.

A **probability distribution** is an assignment of probabilities to all values of a rendom variable, e.g.:

$$Pr(Weather = sun) = 0.6$$
$$Pr(Weather = rain) = 0.1$$
$$Pr(Weather = cloud) = 0.29$$
$$Pr(Weather = snow) = 0.01$$

KENNESAW STATE UNIVERSITY

Joint Probability Distributions

$Pr(Weather, Cavity)$ denotes the probabilities of all combinations of $Weather$ and $Cavity$. This notation is a compact representation of a **joint probability distribution**. The notation

$$Pr(Weather, Cavity) = Pr(Weather|Cavity)Pr(Cavity)$$

stands for the $4 \times 2 = 8$ equations:

$$\begin{aligned}
Pr(W = sun \wedge C = true) &= Pr(W = sun|C = true)Pr(C = true) \\
Pr(W = rain \wedge C = true) &= Pr(W = rain|C = true)Pr(C = true) \\
Pr(W = cloud \wedge C = true) &= Pr(W = cloud|C = true)Pr(C = true) \\
Pr(W = snow \wedge C = true) &= Pr(W = snow|C = true)Pr(C = true) \\
Pr(W = sun \wedge C = false) &= Pr(W = sun|C = false)Pr(C = false) \\
Pr(W = rain \wedge C = false) &= Pr(W = rain|C = false)Pr(C = false) \\
Pr(W = cloud \wedge C = false) &= Pr(W = cloud|C = false)Pr(C = false) \\
Pr(W = snow \wedge C = false) &= Pr(W = snow|C = false)Pr(C = false)
\end{aligned}$$

KENNESAW STATE UNIVERSITY

Probability Axioms

All of probability theory can be built from **Kolmogorov's axioms**.

Law of normalization:

$$0 \leq Pr(\omega) \leq 1 \text{ for every } \omega \text{ and } \sum_{\omega \in \Omega} Pr(\omega) = 1$$

Probability of a disjunction, **inclusion-exlusion principle**:

$$Pr(a \vee b) = Pr(a) + Pr(b) - Pr(a \wedge b)$$

Why is probability theory a valid basis for rational behavior?

De Finetti's Theorem:

> *If Agent 1 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 1 will lose money every time.*

Example:

Proposition	Agent 1's belief	Agent 2 bets	Agent 1 bets	Agent 1 payoffs for each outcome			
				a, b	$a, \neg b$	$\neg a, b$	$\neg a, \neg b$
a	0.4	\$4 on a	\$6 on $\neg a$	–\$6	–\$6	\$4	\$4
b	0.3	\$3 on b	\$7 on $\neg b$	–\$7	\$3	–\$7	\$3
$a \vee b$	0.8	\$2 on $\neg(a \vee b)$	\$8 on $a \vee b$	\$2	\$2	\$2	–\$8
				–\$11	–\$1	–\$1	–\$1

KENNESAW STATE UNIVERSITY

Inference Using Full Joint Distributions

Knowledge base is full joint distribution of boolean random variables $Toothache, Cavity, Catch$.

	toothache		$\neg$*toothache*	
	catch	$\neg$*catch*	*catch*	$\neg$*catch*
cavity	0.108	0.012	0.072	0.008
$\neg$*cavity*	0.016	0.064	0.144	0.576

- All probabilities above sum to 1.
- $Pr(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28.$

Unconditional or **marginal probability** of cavity:

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

KENNESAW STATE UNIVERSITY

Marginalization

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

General rule: $Pr(Y) = \sum_z Pr(Y, Z = z)$ where Z is set of all possible values of variables other than Y. Example, let $Y = Cavity$

$$
\begin{aligned}
Pr(Cavity) &= Pr(Cavity, toothache, catch) + Pr(Cavity, toothache, \neg catch) \\
&+ Pr(Cavity, \neg toothache, catch) + Pr(Cavity, \neg toothache, \neg catch) \\
&=< 0.108, 0.016 > + < 0.012, 0.064 > + < 0.072, 0.144 > + < 0.008, 0.576 > \\
&=< 0.2, 0.8 >
\end{aligned}
$$

KENNESAW STATE UNIVERSITY

Conditioning

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

Using the product rule in $Pr(Y) = \sum_z Pr(Y, Z = z)$ we obtain the **conditioning** rule: $Pr(Y) = \sum_z Pr(Y|z)Pr(z)$

To get the conditional probabilites, we use the definition:

$$Pr(cavity|toothache) = \frac{Pr(cavity \wedge toothache)}{Pr(toothache)} = \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

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

KENNESAW STATE UNIVERSITY

Normalization

Notice that $Pr(toothache)$ appears in denominator in both equations in the preceding conditional probability calculations. We can vew it as a **normalization constant**, denoted α, that ensures that the conditional probability distribution $Pr(Cavity|toothache)$ sums to 1. Then we can write:

$$\begin{aligned} Pr(Cavity|toothache) &= \alpha Pr(Cavity, toothache) \\ &= \alpha[Pr(Cavity, toothache, catch) + Pr(Cavity, toothache, \neg catch)] \\ &= \alpha[< 0.108, 0.16 > + < 0.012, 0.064 > \\ &= \alpha < 0.12, 0.08 > = < 0.6, 0.4 > \end{aligned}$$

All of the preceding can be summarized in a general inference procedure using full joint probability distributions. Let E be a list of evidence variables, e be a list of observed values for the evidence variables, and Y be the remaining unobserved variables. Then:

$$Pr(X|\boldsymbol{e}) = \alpha Pr(X, \boldsymbol{e}) = \alpha \sum_{y} Pr(X, \boldsymbol{e}, \boldsymbol{y})$$

Great, so we're done! All we need is a full joint parobability distribution and we can answer any query. Unfortunately, not practical.

- For a domain of n boolean variables, we have a table of size $O(2^n)$

So we need different approaches, which we cover next.

Independence

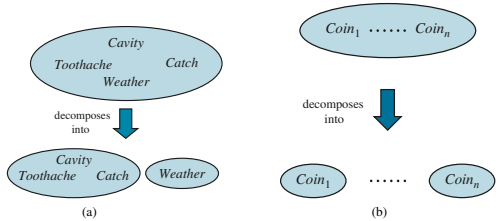
Weather is not affecteed by teeth. So we can assert:

$$Pr(cloud|toothache, catch, cavity) = Pr(cloud)$$

In general, if a and b are independent:

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

Can be a big help. For example, if you have n independent coin flips, then instead of 2^n full joint table, you have a product of n distributions $Pr(C_i)$.



Unfortunately, independence rarely holds in the real world.

Bayes' Rule

Using the symmetry $Pr(X, Y) = Pr(Y, X)$ and the product rule we can derive **Bayes' rule**:

$$\begin{aligned} Pr(X, Y) &= Pr(Y, X) \\ Pr(Y|X)Pr(X) &= Pr(X|Y)Pr(Y) \\ Pr(Y|X) &= \frac{Pr(X|Y)Pr(Y)}{Pr(X)} \end{aligned}$$

The usefulness of Bayes' rule becomes apparent if we consider X as an effect and Y as a cause and we want to determine the cause of some effect (evidence) we observe:

$$Pr(cause|effect) = \frac{Pr(effect|cause)Pr(cause)}{Pr(effect)}$$

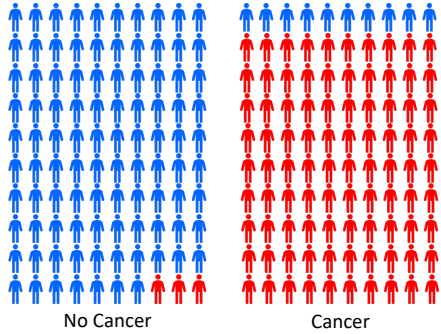
- $Pr(effect|cause)$ quantifies the **causal** direction.
- $Pr(cause|effect)$ quantifies the **diagnostic** direction.

Reasoning from effects to causes is also called **abductive reasoning**. (Is Sherlock Holmes truly employing deduction?)

KENNESAW STATE UNIVERSITY

A Medical Screening Example

A cancer with occurence rate of 1% (.01) has a "90% accurate" test, and:



False positive rate: .03, False negative rate: 0.10

Questions:

- If we screen someone, what is the probability that they test positive?
- If someone tests positive, what is the probability that they have cancer?

The test is an effect. Cancer is the cause.

Analysis of Medical Screening Example

With our probabilistic machinery we can analyze this cancer screening example. First, we model the problem in the language of Bayesian probability theory:

$$p(C = 1) = \frac{1}{100} \quad \text{(Prior probability that someone has cancer)}$$

$$p(C = 0) = \frac{99}{100} \quad \text{(Prior probability that someone has no cancer)}$$

$$p(T = 1|C = 1) = \frac{90}{100} \quad \text{(Conditional probability of positive test given cancer)}$$

$$p(T = 0|C = 1) = \frac{10}{100} \quad \text{(Conditional probability of negative test given cancer)}$$

$$p(T = 1|C = 0) = \frac{3}{100} \quad \text{(Conditional probability of positive test given no cancer)}$$

$$p(T = 0|C = 0) = \frac{97}{100} \quad \text{(Conditional probability of negative test given no cancer)}$$

Now we can answer the two questions we posed at the outset:

- If we screen someone, what is the probability that they test positive?
- If someone tests positive, what is the probability that they have cancer?

KENNESAW STATE UNIVERSITY

Analysis of Medical Screening Example

$$p(C=1) = \frac{1}{100}$$
$$p(C=0) = \frac{99}{100}$$
$$p(T=1|C=1) = \frac{90}{100}$$
$$p(T=0|C=1) = \frac{10}{100}$$
$$p(T=1|C=0) = \frac{3}{100}$$
$$p(T=0|C=0) = \frac{97}{100}$$

If we screen someone, probability that they test positive:

$$\begin{aligned} p(T=1) &= p(T=1|C=0)p(C=0) + p(T=1|C=1)p(C=1) \\ &= \frac{3}{100} \times \frac{99}{100} + \frac{90}{100} \times \frac{1}{100} \\ &= \frac{387}{10,000} \\ &= .0387 \end{aligned}$$

If someone tests positive, probability they have cancer:

$$\begin{aligned} p(C=1|T=1) &= \frac{p(T=1|C=1)p(C=1)}{p(T=1)} \\ &= \frac{90}{100} \times \frac{1}{100} \times \frac{10,000}{387} \\ &= \frac{90}{387} \\ &\approx 0.23 \end{aligned}$$

KENNESAW STATE UNIVERSITY

Bayes' Rule and Combining Evidence

If dentist's probe catches and patient has a toothache, then using the full joint:

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

we can simply read off the answer:

$$P(Cavity|toothache \wedge catch) = \alpha < 0.108, 0.016 > \equiv < 0.871, 0.129 >$$

But we know this approach doesn't scale. With n evidence variables we have $O(2^n)$ possible combinations of observed values.

We could reformulate the problem using Bayes' rule:

$$Pr(Cavity|toothache \wedge catch) = \alpha Pr(toothache \wedge catch|Cavity)PR(Cavity)$$

But, again, we have $O(2^n)$ combinations of observed evidence. We need some additional knowledge.

Conditional Independence

Toothache and Catch are not independent: if the probe catches in the tooth, then it is likely that the tooth has a cavity and that the cavity causes a toothache.

However, Toothache and Catch are independent 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. Mathematically, we write this fact as:

$$Pr(toothache \wedge catch|Cavity) = Pr(toothache|Cavity)Pr(catch|Cavity)$$

In general, the **conditional independence** of two variables X and Y , given a third variable Z, is defined as:

$$Pr(X, Y|Z) = Pr(X|Z)Pr(Y|Z)$$

Given these independence assertions we can say $Pr(X|Y, Z) = Pr(X|Z)$ and $Pr(Y|X, Z) = Pr(Y|Z)$.

Factoring a Joint Distribution using Conditional Independence

Given the conditional independence assertion

$$Pr(Toothache, Catch|Cavity) = Pr(Toothache|Cavity)Pr(Catch|Cavity)$$

We can decompose the full joint for $Toothache, Catch, Cavity$:

$$\begin{aligned} Pr(Toothache, Catch, Cavity) &= Pr(Toothache, Catch|Cavity)Pr(Cavity) \text{ (product rule)} \\ &= Pr(Toothache|Cavity)Pr(Catch|Cavity)Pr(Cavity) \\ &\qquad \text{(cond. ind. assertion above)} \end{aligned}$$

This decomposes the large table smaller tables. In general, this technique turns representation that grows as $O(2^n)$ to one that grows as $O(n)$. Conditional independence assertions:

- allow probabilistic systems to scale, and
- are much more commonly available than absolute independcence assertions.

Conceptually, we say that $Cavity$ **separates** $Toothache$ and $Catch$ because it is a direct cause of both of them.

KENNESAW STATE UNIVERSITY

Naive Bayes Model

If a single cause influences n effects each of which is independent given the cause, then the full joint can be written:

$$Pr(Cause, Effect_1, \ldots, Effect_n) = Pr(Cause) \prod_i Pr(Effect_i|Cause)$$

This is called a **naive Bayes** model – "naive" because it is often used as a simplifying assumption in cases where the "effect" variables are not strictly independent given the cause variable. In practice, naive Bayes systems often work very well, even when the conditional independence assumption is not strictly true.

With some algebraic manipulation using previous results we get, for effects $\boldsymbol{e}$:

$$Pr(Cause|e) = \alpha Pr(Cause) \prod_J Pr(e_j|Cause)$$

This model is useful in text classification, for example, in early spam filters. For the spam filtering problem the causes are Spam and Not-Spam, and the effects are keywords.

KENNESAW STATE UNIVERSITY