Problem of Logic Agents
- Logic-agents almost never have access to the whole truth about their environments.
- A rational agent is one that makes rational decisions in order to maximize its performance measure.
- Logic-agents may have to either risk falsehood or make weak decisions in uncertain situations.
- A rational agent’s decision depends on relative importance of goals, likelihood of achieving them.
- Probability theory provides a quantitative way of encoding likelihood

Foundations of Probability
- Probability Theory makes the same ontological commitments as FOL
- Every sentence $S$ is either true or false.
- The degree of belief, or probability, that $S$ is true is a number $P$ between 0 and 1.
- $P(S) = 1$ iff $S$ is certainly true
- $P(S) = 0$ iff $S$ is certainly false
- $P(S) = 0.4$ iff $S$ is true with a 40% chance
- $P(\text{not } A) = \text{probability that } A \text{ is false}$
- $P(A \text{ and } B) = \text{probability that both } A \text{ and } B \text{ are true}$
- $P(A \text{ or } B) = \text{probability that either } A \text{ or } B \text{ (or both) are true}$

Axioms of Probability
- All probabilities are between 0 and 1
- Valid propositions have probability 1. Unsatisfiable propositions have probability 0. That is,
  - $P(A \text{ v } A) = P(\text{true}) = 1$
  - $P(\text{A } \& \text{A}) = P(\text{false}) = 0$
  - $P(A) = 1 - P(A)$
- The probability of disjunction is defined as follows.
  - $P(A \text{ v } B) = P(A) + P(B) - P(A \& B)$
  - $P(A \& B) = P(A) + P(B) - P(A \text{ v } B)$

Exercise Problem I
Prove that
- $P(A \text{ v } B \text{ v } C) = P(A) + P(B) + P(C) - P(A \& B) - P(A \& C) - P(B \& C) + P(A \& B \& C)$

How to Decide Values of Probability
- $P(\text{the sun comes up tomorrow}) = 0.999$
- Frequentist
  - Probability is inherent in the process
  - Probability is estimated from measurements
  - Probs can be wrong!
A Question
Jane is from Berkeley. She was active in anti-war protests in the 60’s. She lives in a commune.

• Which is more probable?
  1. Jane is a bank teller
  2. Jane is a feminist bank teller

Conditional Probability

• $P(A)$ is the unconditional (or prior) probability
• An agent can use unconditional probability of $A$ to reason about $A$ only in the absence of no further information.
• If some further evidence $B$ becomes available, the agent must use the conditional (or posterior) probability:

$$P(A|B)$$

the probability of $A$ given that the agent already knew that $B$ is true.
• $P(A)$ can be thought as the conditional probability of $A$ with respect to the empty evidence:

$$P(A) = P(A|\emptyset).$$

Random Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>{ 1, 2, ..., 120 }</td>
</tr>
<tr>
<td>Weather</td>
<td>{ sunny, dry, cloudy, raining }</td>
</tr>
<tr>
<td>Size</td>
<td>{ small, medium, large }</td>
</tr>
<tr>
<td>Raining</td>
<td>{ true, false }</td>
</tr>
</tbody>
</table>

• The probability that a random variable $X$ has value $\text{val}$ is written as $P(X=\text{val})$
• $P$: domain $\{0, 1\}$
• Sums to 1 over the domain:
  - $P(\text{Raining} = \text{true}) + P(\text{Raining} = \text{false}) = 0.2$
  - $P(\text{Raining} = \text{false}) = P(\neg \text{ Raining}) = 0.8$

Probability Distribution

• If $X$ is a random variable, we use the bold case $P(X)$ to denote a vector of values for the probabilities of each individual element that $X$ can take.
• Example:
  - $P(\text{Weather} = \text{sunny}) = 0.6$
  - $P(\text{Weather} = \text{rain}) = 0.2$
  - $P(\text{Weather} = \text{cloudy}) = 0.18$
  - $P(\text{Weather} = \text{snow}) = 0.02$
• Then $P(\text{Weather}) = <0.6, 0.2, 0.18, 0.02>$ (the value order of “sunny”, “rain”, “cloudy”, “snow” is assumed).
• $P(\text{Weather})$ is called a probability distribution for the random variable Weather.
• Joint distribution: $P(X_1, X_2, ..., X_n)$
• Probability assignment to all combinations of values of random variables
Joint Distribution Example

<table>
<thead>
<tr>
<th>Toothache</th>
<th>Toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
</tr>
<tr>
<td>-Cavity</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.89</td>
</tr>
</tbody>
</table>

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
  - P(cavity) = 0.1  [add elements of cavity row]
  - P(toothache) = 0.02  [add elements of toothache column]

Joint Distribution Example

<table>
<thead>
<tr>
<th>Toothache</th>
<th>Toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
</tr>
<tr>
<td>-Cavity</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.89</td>
</tr>
</tbody>
</table>

- The sum of the entries in this table has to be 1
- Given this table, one can answer all the probability questions about this domain
  - P(cavity) = 0.1  [add elements of cavity row]
  - P(toothache) = 0.02  [add elements of toothache column]

Joint Probability Distribution (JPD)

- A joint probability distribution \( P(X_1, X_2, ..., X_n) \) provides complete information about the probabilities of its random variables.
- However, JPD’s are often hard to create (again because of incomplete knowledge of the domain).
- Even when available, JPD tables are very expensive, or impossible, to store because of their size.
- A JPD table for \( n \) random variables, each ranging over \( k \) distinct values, has \( k^n \) entries!
- A better approach is to come up with conditional probabilities as needed and compute the others from them.

Bayes’ Rule

- Bayes’ Rule
  - \( P(A \mid B) = P(B \mid A) \cdot P(A) \div P(B) \)
  - What is the probability that a patient has meningitis (M) given that he has a stiff neck (S)?
    - \( P(M \mid S) = P(S \mid M) \cdot P(M) \div P(S) \)
    - \( P(S \mid M) \) is easier to estimate than \( P(M \mid S) \) because it refers to causal knowledge:
      - meningitis typically causes stiff neck.
    - \( P(M) \) can be estimated from past medical cases and the knowledge about how meningitis works.
    - Similarly, \( P(M), P(S) \) can be estimated from statistical information.
Bayes' Rule

- Bayes' Rule: \( P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)} \)
- The Bayes rule is helpful even in absence of (immediate) causal relationships.
- What is the probability that a blonde (B) is Swedish (S)?
  - \( P(S \mid B) = \frac{P(B \mid S) P(S)}{P(B)} \)
- All \( P(B \mid S), P(S), P(B) \) are easily estimated from statistical information.
  - \( P(B \mid S) = (\# \text{ of blonde Swedish})/(\text{Swedish population}) = \frac{9}{10} \)
  - \( P(S) = \text{Swedish population/world population} = \ldots \)
  - \( P(B) = \# \text{ of blondes/world population} = \ldots \)

Conditional Independence

- Conditioning
  - \( P(A) = P(A \mid B) P(B) + P(A \mid \neg B) P(\neg B) \)
  - In terms of exponential explosion, conditional probabilities do not seem any better than JPD's for computing the probability of a fact, given \( n \geq 1 \) pieces of evidence.
  - \( P(\text{Meningitis} \mid \text{StiffNeck} \& \text{Nausea} \& \ldots \& \text{DoubleVision}) \)
- However, certain facts do not always depend on all the evidence.
  - \( P(\text{Meningitis} \mid \text{StiffNeck} \& \text{Astigmatic}) = P(\text{Meningitis} \mid \text{StiffNeck}) \)
- Meningitis and Astigmatic are conditionally independent, given StiffNeck.

Independence

- A and B are independent iff
  - \( P(A \& B) = P(A) \cdot P(B) \)
  - \( P(A \mid B) = P(A) \)
  - \( P(B \mid A) = P(B) \)
- Independence is essential for efficient probabilistic reasoning
  - A and B are conditionally independent given C iff
    - \( P(A \mid B, C) = P(A \mid C) \)
    - \( P(B \mid A, C) = P(B \mid C) \)
    - \( P(A \& B \mid C) = P(A \mid C) \cdot P(B \mid C) \)

Examples of Conditional Independence

- Toothache (T)
- Spot in Xray (X)
- Cavity (C)
  - None of these propositions are independent of one other
  - T and X are conditionally independent given C
Examples of Conditional Independence

• Toothache (T)
• Spot in Xray (X)
• Cavity (C)
• None of these propositions are independent of one another
• T and X are conditionally independent given C
• Battery is dead (B)
• Radio plays (R)
• Starter turns over (S)

Uncertainty

Let action \( A_t = \) leave for airport \( t \) minutes before flight
Will \( A_t \) get me there on time?

Problems:
1. partial observability (road state, other drivers’ plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Hence a purely logical approach either
1. risks falsehood: "\( A_{25} \) will get me there on time", or
2. leads to conclusions that are too weak for decision making:
   "\( A_{25} \) will get me there on time if there’s no accident on the bridge and it doesn’t rain and my tires remain intact etc etc."

\( A_{25} \) might reasonably be said to get me there on time but I’d have to stay overnight in the airport ...

Methods for handling uncertainty

• Default or nonmonotonic logic:
  • Assume my car does not have a flat tire
  • Assume \( A_{25} \) works unless contradicted by evidence

• Issues: What assumptions are reasonable? How to handle contradiction?

• Rules with fudge factors:
  • \( A_{25} \leftarrow \text{no flat tire} \) get there on time
  • Sprinkler \( \rightarrow 0.99 \) WetGrass
  • WetGrass \( \rightarrow 0.7 \) Rain

• Issues: Problems with combination, e.g., Sprinkler causes Rain??

• Probability
  • Model agent’s degree of belief
  • Given the available evidence, \( A_{25} \) will get me there on time with probability 0.04

Inference by enumeration

• Start with the joint probability distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>~toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>work</td>
<td>0.18</td>
<td>0.82</td>
</tr>
<tr>
<td>work</td>
<td>.006</td>
<td>.994</td>
</tr>
</tbody>
</table>

• For any proposition \( \phi \), sum the atomic events where it is true: \( P(\phi) = \sum_{\omega \in \phi} P(\omega) \)
Inference by enumeration

- Start with the joint probability distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>~toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>cavity</td>
<td>0.108</td>
<td>0.012</td>
</tr>
<tr>
<td>catch</td>
<td>~catch</td>
<td>0.016</td>
</tr>
</tbody>
</table>

For any proposition φ, sum the atomic events where it is true:

\[ P(φ) = \sum_{ω: ω \models φ} P(ω) \]

- \( P(\text{toothache} \lor \text{cavity}) = 0.108 + 0.012 + 0.016 + 0.064 + 0.072 + 0.008 = 0.28 \)

Normalization

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>~toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>cavity</td>
<td>0.108</td>
<td>0.012</td>
</tr>
<tr>
<td>catch</td>
<td>~catch</td>
<td>0.016</td>
</tr>
</tbody>
</table>

- Denominator can be viewed as a normalization constant \( α \)

\[ P(\text{Cavity} | \text{toothache}) = \frac{α}{P(\text{Cavity, toothache})} \]

\[ = \frac{α}{P(\text{Cavity, toothache, catch}) + P(\text{Cavity, toothache, ~catch})} \]

\[ = \frac{α}{<0.12, 0.08> + <0.6, 0.4>} \]

\[ = \frac{1}{P(\text{toothache})} \]

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

Conditional independence

- \( P(\text{Toothache, Cavity, Catch}) \) has \( 2^3 - 1 = 7 \) independent entries

- If I have a cavity, the probability that the probe catches in it doesn’t depend on whether I have a toothache:
  1. \( P(\text{catch} | \text{toothache, cavity}) = P(\text{catch} | \text{cavity}) \)
  2. The same independence holds if I haven’t got a cavity:
  3. \( P(\text{catch} | \text{toothache, ~cavity}) = P(\text{catch} | \text{~cavity}) \)

- Catch is conditionally independent of Toothache given Cavity:

\[ P(\text{Catch} | \text{Toothache, Cavity}) = P(\text{Catch} | \text{Cavity}) \]

- Equivalent statements:

\[ P(\text{Toothache} | \text{Catch, Cavity}) = P(\text{Toothache} | \text{Cavity}) \]

\[ P(\text{Toothache, Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity}) P(\text{Catch} | \text{Cavity}) \]

Inference by enumeration

- Start with the joint probability distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>~toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>cavity</td>
<td>0.108</td>
<td>0.012</td>
</tr>
<tr>
<td>catch</td>
<td>~catch</td>
<td>0.016</td>
</tr>
</tbody>
</table>

- Can also compute conditional probabilities:

\[ P(\text{~cavity} | \text{toothache}) = \frac{P(\text{~cavity} \times \text{toothache})}{P(\text{toothache})} \]

\[ = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} \]

\[ = 0.4 \]

Normalization

- Denominator can be viewed as a normalization constant \( α \)

\[ P(\text{Cavity} | \text{toothache}) = \frac{α}{P(\text{Cavity, toothache})} \]

\[ = \frac{α}{P(\text{Cavity, toothache, catch}) + P(\text{Cavity, toothache, ~catch})} \]

\[ = \frac{α}{<0.12, 0.08> + <0.6, 0.4>} \]

\[ = \frac{1}{P(\text{toothache})} \]

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

Conditional independence

- Write out full joint distribution using chain rule:

\[ P(\text{Toothache, Catch, Cavity}) = P(\text{Toothache} | \text{Catch, Cavity}) P(\text{Catch} | \text{Cavity}) P(\text{Cavity}) \]

\[ = β_1 P(\text{Catch} | \text{Cavity}) P(\text{Cavity}) P(\text{Cavity}) \]

\[ = β_2 P(\text{Toothache} | \text{Cavity}) P(\text{Cavity}) P(\text{Cavity}) \]

\[ = β_3 P(\text{Toothache} | \text{Cavity}) P(\text{Catch} | \text{Cavity}) P(\text{Cavity}) \]

I.e., \( 2 + 2 + 1 = 5 \) independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in \( n \) to linear in \( n \).

- Conditional independence is our most basic and robust form of knowledge about uncertain environments.
Bayes' Rule

- Product rule $P(a \land b) = P(a | b) \cdot P(b) = P(b | a) \cdot P(a)$
  - Bayes' rule: $P(a | b) = P(b | a) \cdot P(a) / P(b)$

- or in distribution form
  $P(Y|X) = P(X|Y) \cdot P(Y) / P(X) = \alpha \cdot P(X|Y) \cdot P(Y)$

- Useful for assessing diagnostic probability from causal probability:
  $P(\text{Cause} | \text{Effect}) = P(\text{Effect} | \text{Cause}) \cdot P(\text{Cause}) / P(\text{Effect})$

- E.g., let $M$ be meningitis, $S$ be stiff neck:
  $P(m | s) = P(s | m) \cdot P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008$

- Note: posterior probability of meningitis still very small!

Bayes' Rule and conditional independence

- $P(\text{Cavity} | \text{toothache} \land \text{catch}) = \alpha \cdot P(\text{toothache} \land \text{catch} | \text{Cavity}) \cdot P(\text{Cavity})$

- This is an example of a naive Bayes model:
  $P(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = P(\text{Cause}) \cdot \prod P(\text{Effect}_i | \text{Cause})$

- Total number of parameters is linear in $n$

Summary

- Probability is a rigorous formalism for uncertain knowledge
- Joint probability distribution specifies probability of every atomic event
- Queries can be answered by summing over atomic events
- For nontrivial domains, we must find a way to reduce the joint size
  - Independence and conditional independence provide the tools

Bayesian Networks

- To do probabilistic reasoning, you need to know the joint probability distribution
- But, in a domain with $N$ propositional variables, one needs $2^N$ numbers to specify the joint probability distribution
- We want to exploit independences in the domain
- Two components: structure and numerical parameters

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
  - Syntax:
    - a set of nodes, one per variable
    - a directed, acyclic graph (link = "directly influences")
    - a conditional distribution for each node given its parents: $P(X_i | \text{Parents}(X_i))$
  - In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over $X_i$ for each combination of parent values

Example

- Topology of network encodes conditional independence assertions:
  - Weather is independent of the other variables
  - Toothache and Catch are conditionally independent given Cavity
Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.

- Burglary
- Earthquake
- Alarm
- JohnCalls
- MaryCalls

Compactness

- A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values
- Each row requires one number $p$ for $X_i = \text{true}$
- If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Semantics

The full joint distribution is defined as the product of the local conditional distributions:

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

Example

- Suppose we choose the ordering $M, J, A, B, E$

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \ldots, X_n$
- 2. For $i = 1$ to $n$
  - add $X_i$ to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that
    $$P(X_i | \text{Parents}(X_i)) = P(X_i | X_1, \ldots, X_{i-1})$$

This choice of parents guarantees:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})$$

(by chain rule)

(by construction)
• Suppose we choose the ordering $M, J, A, B, E$

$P(J | M) = P(J)\? 
No 
$P(B | A, J, M) = P(B | A)\? Yes 
$P(E | B, A, J, M) = P(E | A)\? 
$P(E | B, A, J, M) = P(E | A, B)\?

Example contd.

- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct