Intro to Artificial Intelligence
CS 171
Reasoning Under Uncertainty
Chapter 13 and 14.1-14.2
Andrew Gelfand
3/1/2011
Today...

- Representing uncertainty is useful in knowledge bases
  - Probability provides a coherent framework for uncertainty

- Review basic concepts in probability
  - Emphasis on conditional probability and conditional independence

- Full joint distributions are difficult to work with
  - Conditional independence assumptions allow us to model real-world phenomena with much simpler models

- Bayesian networks are a systematic way to build compact, structured distributions

- Reading: Chapter 13; Chapter 14.1-14.2
History of Probability in AI

- Early AI (1950’s and 1960’s)
  - Attempts to solve AI problems using probability met with mixed success

- Logical AI (1970’s, 80’s)
  - Recognized that working with full probability models is intractable
  - Abandoned probabilistic approaches
  - Focused on logic-based representations

- Probabilistic AI (1990’s-present)
  - Judea Pearl invents Bayesian networks in 1988
  - Realization that working w/ approximate probability models is tractable and useful
  - Development of machine learning techniques to learn such models from data
  - Probabilistic techniques now widely used in vision, speech recognition, robotics, language modeling, game-playing, etc.
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.”

($A_{1440}$ might reasonably be said to get me there on time but I'd have to stay overnight in the airport ...
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:**
  - $A_{25} \rightarrow_{0.3} \text{get there on time}$
  - $\text{Sprinkler} \rightarrow_{0.99} \text{WetGrass}$
  - $\text{WetGrass} \rightarrow_{0.7} \text{Rain}$

- **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
Probability

Probabilistic assertions **summarize** effects of

- **laziness**: failure to enumerate exceptions, qualifications, etc.
- **ignorance**: lack of relevant facts, initial conditions, etc.

**Subjective** probability:

- Probabilities relate propositions to agent's own state of knowledge
  
e.g., \( P(A_{25} \mid \text{no reported accidents}) = 0.06 \)

These are **not** assertions about the world

Probabilities of propositions change with new evidence:

- e.g., \( P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15 \)
Making decisions under uncertainty

Suppose I believe the following:

\[
P(A_{25} \text{ gets me there on time } | \ ... ) = 0.04
\]
\[
P(A_{90} \text{ gets me there on time } | \ ... ) = 0.70
\]
\[
P(A_{120} \text{ gets me there on time } | \ ... ) = 0.95
\]
\[
P(A_{1440} \text{ gets me there on time } | \ ... ) = 0.9999
\]

Which action to choose?

Depends on my preferences for missing flight vs. time spent waiting, etc.

- Utility theory is used to represent and infer preferences
- Decision theory = probability theory + utility theory
Syntax

- Basic element: random variable
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.
- **Boolean** random variables
e.g., *Cavity* (do I have a cavity?)
- **Discrete** random variables
e.g., *Dice* is one of \(<1,2,3,4,5,6>\)
- Domain values must be exhaustive and mutually exclusive
- Elementary proposition constructed by assignment of a value to a random variable:
e.g., *Weather = sunny, Cavity = false* (abbreviated as \(\neg \text{cavity}\))
- Complex propositions formed from elementary propositions and standard logical connectives e.g., *Weather = sunny \lor Cavity = false*
**Syntax**

- **Atomic event**: A complete specification of the state of the world about which the agent is uncertain.

- e.g. Imagine flipping two coins
  - The set of all possible worlds is: 
    \[ S = \{ (H,H), (H,T), (T,H), (T,T) \} \]
    
    Meaning there are 4 distinct atomic events in this world.

- Atomic events are mutually exclusive and exhaustive.
Axioms of probability

- Given a set of possible worlds $S$
  - $P(A) \geq 0$ for all atomic events $A$
  - $P(S) = 1$
  - If $A$ and $B$ are mutually exclusive, then:
    \[ P(A \lor B) = P(A) + P(B) \]

- Refer to $P(A)$ as probability of event $A$
  - e.g. if coins are fair $P\{\text{H,H}\} = \frac{1}{4}$
Probability and Logic

- Probability can be viewed as a generalization of propositional logic

- **P(a):**
  - *a* is any sentence in propositional logic
  - Belief of agent in *a* is no longer restricted to *true, false, unknown*
  - **P(a)** can range from 0 to 1
    - **P(a) = 0, and P(a) = 1** are special cases
    - So logic can be viewed as a special case of probability
Basic Probability Theory

- General case for $A$, $B$:
  \[ P(A \lor B) = P(A) + P(B) - P(A \land B) \]

- e.g., imagine I flip two coins
  - Events $\{(H,H),(H,T),(T,H),(T,T)\}$ are all equally likely
  - Consider event $E$ that the 1st coin is heads: $E=\{(H,H),(H,T)\}$
  - And event $F$ that the 2nd coin is heads: $F=\{(H,H),(T,H)\}$
  - $P(E \lor F) = P(E) + P(F) - P(E \land F) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$
Conditional Probability

- The 2 dice problem
  - Suppose I roll two fair dice and 1\textsuperscript{st} dice is a 4
  - What is probability that sum of the two dice is 6?

- 6 possible events, given 1\textsuperscript{st} dice is 4
  - (4,1),(4,2),(4,3),(4,4),(4,5),(4,6)

- Since all events (originally) had same probability, these 6 events should have equal probability too

- Probability is thus 1/6
Conditional Probability

- Let A denote event that sum of dice is 6
- Let B denote event that 1st dice is 4
- Conditional Probability denoted as: \( P(A \mid B) \)
  - Probability of event A given event B

- General formula given by: \( P(A \mid B) = \frac{P(A \land B)}{P(B)} \)
  - Probability of \( A \land B \) relative to probability of B

- What is \( P(\text{sum of dice} = 3 \mid 1^{\text{st}} \text{ dice is 4}) \)?
  - Let C denote event that sum of dice is 3
  - \( P(B) \) is same, but \( P(C \land B) = 0 \)
Random Variables

- Often interested in some function of events, rather than the actual event
  - Care that sum of two dice is 4, not that the event was (1,3), (2,2) or (3,1)

- Random Variable is a real-valued function on space of all possible worlds
  - e.g. let Y = Number of heads in 2 coin flips
    - \( P(Y=0) = P(\{T,T\}) = \frac{1}{4} \)
    - \( P(Y=1) = P(\{H,T\} \lor \{T,H\}) = \frac{1}{2} \)
Prior (Unconditional) Probability

- **Probability distribution** gives values for all possible assignments:

  \[
  \begin{array}{|c|c|c|c|}
  \hline
  & Sunny & Rainy & Cloudy & Snowy \\
  \hline
  P(Weather) & 0.7 & 0.1 & 0.19 & 0.01 \\
  \hline
  \end{array}
  \]

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables:

  \[
  \begin{array}{|c|c|c|c|}
  \hline
  P(\text{Weather,Cavity}) & Sunny & Rainy & Cloudy & Snowy \\
  \hline
  \text{Cavity} & 0.144 & 0.02 & 0.016 & 0.006 \\
  \hline
  \text{¬Cavity} & 0.556 & 0.08 & 0.174 & 0.004 \\
  \hline
  \end{array}
  \]

- \( P(A,B) \) is shorthand for \( P(A \land B) \)

- Joint distributions are normalized: \( \sum_a \sum_b P(A=a, B=b) = 1 \)
Computing Probabilities

Say we are given following joint distribution:

Joint distribution for $k$ binary variables has $2^k$ probabilities!

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>.108</td>
<td>.012</td>
<td>.072</td>
<td>.008</td>
</tr>
<tr>
<td>¬ catch</td>
<td>.016</td>
<td>.064</td>
<td>.144</td>
<td>.576</td>
</tr>
<tr>
<td>cavity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>¬ cavity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Computing Probabilities

Say we are given following joint distribution:

What is \( P(\text{cavity}) \)?

\[
P(\text{cavity}) = P(\text{cavity}, \text{catch}, \text{toothache}) + \\
P(\text{cavity}, \neg \text{catch}, \text{toothache}) + \\
P(\text{cavity}, \text{catch}, \neg \text{toothache}) + \\
P(\text{cavity}, \neg \text{catch}, \neg \text{toothache}) \\
= .108 + .012 + .072 + .008 = .2
\]

Law of Total Probability (aka marginalization)

\[
P(a) = \sum_b P(a, b) \\
= \sum_b P(a \mid b) P(b)
\]
Computing Probabilities

What is $P(\text{cavity}|\text{toothache})$?

$$P(\text{cavity}|\text{toothache}) = \frac{P(\text{cavity}, \text{toothache})}{P(\text{toothache})}$$

$P(\text{cavity}, \text{toothache}) = P(\text{cavity, catch, toothache}) + P(\text{cavity, } \neg \text{catch, toothache})$

$= 0.108 + 0.012 = 0.12$

$P(\text{toothache}) = P(\text{cavity, toothache}) + P(\neg \text{cavity, toothache})$

$= 0.12 + (0.016 + 0.064) = 0.2$

$$P(\text{cavity}|\text{toothache}) = \frac{P(\text{cavity, toothache})}{P(\text{toothache})} = \frac{0.12}{0.2} = 0.6$$

Can get any conditional probability from joint distribution
Computing Probabilities: Normalization

What is $P(\text{Cavity} | \text{Toothache} = \text{toothache})$?

This is a distribution over the 2 states: \{cavity, $\neg$-cavity\}

$$P(\text{Cavity} | \text{Toothache} = \text{toothache}) = \alpha P(\text{Cavity}, \text{Toothache} = \text{toothache})$$

Distributions will be denoted w/ capital letters; Probabilities will be denoted w/ lowercase letters.
Computing Probabilities: The Chain Rule

- We can always write
  \[ P(a, b, c, \ldots z) = P(a \mid b, c, \ldots z) \ P(b, c, \ldots z) \]
  (by definition of joint probability)

- Repeatedly applying this idea, we can write
  \[ P(a, b, c, \ldots z) = P(a \mid b, c, \ldots z) \ P(b \mid c,\ldots z) \ P(c \mid \ldots z) \ldots P(z) \]

- Semantically different factorizations with different orderings
  \[ P(a, b, c, \ldots z) = P(z \mid y, x, \ldots a) \ P(y \mid x,\ldots a) \ P(x \mid \ldots a) \ldots P(a) \]
Independence

- $A$ and $B$ are independent iff
  
  \[ P(A \mid B) = P(A) \]
  
  or equivalently,
  
  \[ P(B \mid A) = P(B) \]
  
  or equivalently,
  
  \[ P(A, B) = P(A) \cdot P(B) \]

- e.g., for $n$ independent biased coins, $O(2^n) \rightarrow O(n)$

- Absolute independence is powerful but rare

- e.g., consider field of dentistry. Many variables, none of which are independent. What should we do?
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 doesn't depend on whether I have a toothache:
  
  \[ (1) P(\text{Catch} \mid \text{Toothache, cavity}) = P(\text{Catch} \mid \text{cavity}) \]

- The same independence holds if I haven't got a cavity:
  
  \[ (2) P(\text{Catch} \mid \text{Toothache, } \neg \text{cavity}) = P(\text{Catch} \mid \neg \text{cavity}) \]

- Catch is conditionally independent of Toothache given Cavity:
  
  $P(\text{Catch} \mid \text{Toothache, Cavity}) = P(\text{Catch} \mid \text{Cavity})$

- Equivalent statements:
  
  $P(\text{Toothache} \mid \text{Catch, Cavity}) = P(\text{Toothache} \mid \text{Cavity})$

  $P(\text{Toothache, Catch} \mid \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) \cdot P(\text{Catch} \mid \text{Cavity})$
Conditional independence...

- Write out full joint distribution using chain rule:
  \[ P(\text{Toothache}, \text{Catch}, \text{Cavity}) \]
  \[ = P(\text{Toothache} \mid \text{Catch, Cavity}) \, P(\text{Catch, Cavity}) \]
  \[ = P(\text{Toothache} \mid \text{Catch, Cavity}) \, P(\text{Catch} \mid \text{Cavity}) \, P(\text{Cavity}) \]
  \[ = P(\text{Toothache} \mid \text{Cavity}) \, P(\text{Catch} \mid \text{Cavity}) \, P(\text{Cavity}) \]

<table>
<thead>
<tr>
<th>( P(\text{Toothache} \mid \text{Cavity}) )</th>
<th>toothache</th>
<th>¬toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = cavity</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Cavity = ¬cavity</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( P(\text{Catch} \mid \text{Cavity}) )</th>
<th>catch</th>
<th>¬catch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = cavity</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>Cavity = ¬cavity</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( P(\text{Cavity}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = cavity</td>
</tr>
<tr>
<td>Cavity = ¬cavity</td>
</tr>
</tbody>
</table>

\[ P(\text{toothache,catch,¬cavity}) = \frac{}{} \]
\[ = 0.4 \cdot 0.5 \cdot 0.45 = 0.09 \]
Conditional independence...

Write out full joint distribution using chain rule:

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

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

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

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

<table>
<thead>
<tr>
<th>( P(\text{Toothache} \mid \text{Cavity}) )</th>
<th>( \text{toothache} )</th>
<th>( \neg\text{toothache} )</th>
<th>( P(\text{Catch} \mid \text{Cavity}) )</th>
<th>( \text{catch} )</th>
<th>( \neg\text{catch} )</th>
<th>( P(\text{Cavity}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Cavity = cavity} )</td>
<td>0.8</td>
<td>0.2</td>
<td>( \text{Cavity = cavity} )</td>
<td>0.7</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>( \text{Cavity = } \neg\text{cavity} )</td>
<td>0.4</td>
<td>0.6</td>
<td>( \text{Cavity = } \neg\text{cavity} )</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

\( P(\text{Cavity}) \)

<table>
<thead>
<tr>
<th>( \text{Cavity = cavity} )</th>
<th>( \text{Cavity = } \neg\text{cavity} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Requires only \( 2 + 2 + 1 = 5 \) parameters!

Use of conditional independence can reduce size of 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.
Conditional Independence vs Independence

- Conditional independence does not imply independence

- Example:
  - A = height
  - B = reading ability
  - C = age

  - \( P(\text{reading ability} \mid \text{age}, \text{height}) = P(\text{reading ability} \mid \text{age}) \)
  - \( P(\text{height} \mid \text{reading ability}, \text{age}) = P(\text{height} \mid \text{age}) \)

- Note:
  - Height and reading ability are dependent (not independent) but are conditionally independent given age
Bayes’ Rule

- Two jug problem
  - Jug 1 contains: 2 white balls & 7 black balls
  - Jug 2 contains: 5 white balls & 6 black balls
  - Flip a fair coin and draw a ball from Jug 1 if heads; Jug 2 if tails

- What is probability that coin was heads, given a white ball was selected?
  - Want to compute \( P(H|W) \)
  - Have \( P(H) = P(T) = \frac{1}{2} \), \( P(W|H) = \frac{2}{9} \) and \( P(W|T) = \frac{5}{11} \)

\[
P(H|W) = \frac{P(H,W)}{P(W)} = \frac{P(W|H)P(H)}{P(W)} = \frac{P(W|H)P(H)}{P(W,H) + P(W,T)}
\]

\[
= \frac{P(W|H)P(H)}{P(W|H)P(H) + P(W|T)P(T)} = \frac{\frac{2}{9} \cdot \frac{1}{2}}{\frac{2}{9} \cdot \frac{1}{2} + \frac{5}{11} \cdot \frac{1}{2}} \approx \frac{22}{67} \approx 0.328
\]
Bayes' Rule...

- Derived from product rule: \( P(a \land b) = P(a \mid b) \ P(b) = P(b \mid a) \ P(a) \)
  \[ \Rightarrow P(a \mid b) = P(b \mid a) \ P(a) / P(b) \]

- or in distribution form
  \[ P(Y \mid X) = \frac{P(X \mid Y) \ P(Y)}{P(X)} = \alpha P(X \mid Y) \ P(Y) = \alpha P(X,Y) \]

- Useful for assessing **diagnostic** probability from **causal** probability:
  - \( P(Cause \mid Effect) = \frac{P(Effect \mid Cause) \ P(Cause)}{P(Effect)} \)
  - e.g., let \( M \) be meningitis, \( S \) be stiff neck:
    \[ P(m \mid s) = P(s \mid m) \ P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008 \]
  - Note: posterior probability of meningitis still very small!
Bayes' Rule...

- \( P(a \mid b, c) = ?? \)
  \[ = P(b, c \mid a) P(a) / P(b,c) \]

- \( P(a, b \mid c, d) = ?? \)
  \[ = P(c, d \mid a, b) P(a, b) / P(c, d) \]

Both are examples of basic pattern \( p(x \mid y) = p(y \mid x)p(x)/p(y) \)
(it helps to group variables together, e.g., \( y = (a,b), \ x = (c, d) \))
Decision Theory – why probabilities are useful

- Consider 2 possible actions that can be recommended by a medical decision-making system:
  - a = operate
  - b = don’t operate

- 2 possible states of the world
  - c = patient has cancer, ¬c = patient doesn’t have cancer

- Agent’s degree of belief in c is P(c), so P(¬c) = 1 - P(c)

- Utility (to agent) associated with various outcomes:
  - Take action a and patient has cancer: utility = $30k
  - Take action a and patient has no cancer: utility = -$50k
  - Take action b and patient has cancer: utility = -$100k
  - Take action b and patient has no cancer: utility = 0.
Maximizing expected utility

- What action should the agent take?
  - Rational agent should maximize expected utility

- Expected cost of actions:
  \[ E[ \text{utility}(a) ] = 30 \, P(c) - 50 \, [1 - P(c)] \]
  \[ E[ \text{utility}(b) ] = -100 \, P(c) \]
  Break even point?
  \[ 30 \, P(c) - 50 + 50 \, P(c) = -100 \, P(c) \]
  \[ 100 \, P(c) + 30 \, P(c) + 50 \, P(c) = 50 \]
  \[ \Rightarrow P(c) = 50/180 \sim 0.28 \]

  If \( P(c) > 0.28 \), the optimal decision is to operate

- Original theory from economics, cognitive science (1950’s)
  - But widely used in modern AI, e.g., in robotics, vision, game-playing

- Can only make optimal decisions if know the probabilities
What does all this have to do with AI?

- Logic-based knowledge representation
  - Set of sentences in KB
  - Agent’s belief in any sentence is: true, false, or unknown

- In real-world problems there is uncertainty
  - $P(\text{snow in New York on January 1})$ is not 0 or 1 or unknown
  - $P(\text{pit in square 2,2 | evidence so far})$
  - Ignoring this uncertainty can lead to brittle systems and inefficient use of information

- Uncertainty is due to:
  - Things we did not measure (which is always the case)
    - E.g., in economic forecasting
  - Imperfect knowledge
    - $P(\text{symptom | disease}) \rightarrow$ we are not 100% sure
  - Noisy measurements
    - $P(\text{speed > 50 | sensor reading > 50})$ is not 1
Agents, Probabilities & Degrees of Belief

- What we were taught in school ("frequentist" view)
  - $P(a)$ represents frequency that event $a$ will happen in repeated trials

- Degree of belief
  - $P(a)$ represents an agent’s degree of belief that event $a$ is true
  - This is a more general view of probability
    - Agent’s probability is based on what information they have
    - E.g., based on data or based on a theory

- Examples:
  - $a =$ "life exists on another planet"
    - What is $P(a)$? We will all assign different probabilities
  - $a =$ "Mitt Romney will be the next US president"
    - What is $P(a)$?

- Probabilities can vary from agent to agent depending on their models of the world and how much data they have
More on Degrees of Belief

- Our interpretation of $P(a \mid e)$ is that it is an agent’s degree of belief in the proposition $a$, given evidence $e$
  - Note that proposition $a$ is true or false in the real-world
  - $P(a \mid e)$ reflects the agent’s uncertainty or ignorance

- The degree of belief interpretation does not mean that we need new or different rules for working with probabilities
  - The same rules (Bayes rule, law of total probability, probabilities sum to 1) still apply – our interpretation is different
Constructing a Propositional Probabilistic Knowledge Base

- Define all variables of interest: A, B, C, ... Z
- Define a joint probability table for P(A, B, C, ... Z)
  - Given this table, we have seen how to compute the answer to a query, P(query | evidence), where query and evidence = any propositional sentence

- 2 major problems:
  - Computation time:
    - P(a|b) requires summing out other variables in the model
    - e.g., O(m^{K-1}) with K variables
  - Model specification
    - Joint table has O(m^K) entries – where do all the numbers come from?
  - These 2 problems effectively halted the use of probability in AI research from the 1960’s up until about 1990
Bayesian Networks
A Whodunit

- You return home from a long day to find that your house guest has been murdered.
  - There are two culprits:
    1) The Butler; and 2) The Cook
  - There are three possible weapons:
    1) A knife; 2) A gun; and 3) A candlestick

- Let’s use probabilistic reasoning to find out whodunit?
Representing the problem

- There are 2 uncertain quantities
  - Culprit = {Butler, Cook}
  - Weapon = {Knife, Pistol, Candlestick}

- What distributions should we use?
  - Butler is an upstanding guy
  - Cook has a checkered past
  - Butler keeps a pistol from his army days
  - Cook has access to many kitchen knives
  - The Butler is much older than the cook
Representing the problem...

What distributions should we use?

- Butler is an upstanding guy
- Cook has a checkered past
- Butler keeps a pistol from his army days
- Cook has access to many kitchen knives
- The Butler is much older than the cook

\[ P(Culprit) \]

<table>
<thead>
<tr>
<th>Butler</th>
<th>Cook</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.7</td>
</tr>
</tbody>
</table>

\[ P(weapon \mid Culprit=Butler) \]

<table>
<thead>
<tr>
<th>Pistol</th>
<th>Knife</th>
<th>Candlestick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

\[ P(weapon \mid Culprit=Cook) \]

<table>
<thead>
<tr>
<th>Pistol</th>
<th>Knife</th>
<th>Candlestick</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Solving the Crime

- If we observe that the murder weapon was a pistol, who is the most likely culprit?

\[
P(\text{culprit} = \text{Butler} | \text{weapon} = \text{Pistol}) = \frac{P(\text{culprit}=\text{Butler}, \text{weapon}=\text{Pistol})}{P(\text{weapon}=\text{Pistol})}
\]

\[
P(\text{Butler}, \text{Pistol}) = P(\text{Pistol} | \text{Butler}) \cdot P(\text{Butler}) = 0.7 \cdot 0.3
\]

\[
P(\text{Pistol}) = P(\text{Pistol} | \text{Butler}) \cdot P(\text{Butler}) + P(\text{Pistol} | \text{Cook}) \cdot P(\text{Cook})
\]

\[
P(\text{Pistol}) = 0.7 \cdot 0.3 + 0.2 \cdot 0.7
\]

\[
P(\text{culprit} = \text{Butler} | \text{weapon} = \text{Pistol}) = \frac{0.21}{0.21 + 0.14} = 0.6
\]

The Butler!
Your 1st Bayesian Network

- Each node represents a random variable
- Arrows indicate cause-effect relationship
- Shaded nodes represent observed variables

Whodunit model in “words”:
- Culprit chooses a weapon;
- You observe the weapon and infer the culprit
Bayesian Networks

- Represent dependence/independence via a directed graph
  - Nodes = random variables
  - Edges = direct dependence
- Structure of the graph ⇔ Conditional independence relations
- Recall the chain rule of repeated conditioning:
  \[
P(X_1, X_2, X_3..., X_N) = P(X_1|X_2, X_3..., X_N)P(X_2|X_3, ..., X_N) \cdots P(X_N)
  \]
  \[
P(X_1, X_2, X_3..., X_N) = \prod_{i=1}^{n} P(X_i|\text{parents}(X_i))
  \]
  The full joint distribution
  The graph-structured approximation
- Requires that graph is acyclic (no directed cycles)
- 2 components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)
Example of a simple Bayesian network

\[
p(A,B,C) = p(C|A,B)p(A|B)p(B) = p(C|A,B)p(A)p(B)
\]

Probability model has simple factored form

Directed edges \(\Rightarrow\) direct dependence

Absence of an edge \(\Rightarrow\) conditional independence

Also known as belief networks, graphical models, causal networks

Other formulations, e.g., undirected graphical models
Examples of 3-way Bayesian Networks

\[ p(A, B, C) = p(A) \cdot p(B) \cdot p(C) \]
Examples of 3-way Bayesian Networks

Conditionally independent effects:
\[ p(A,B,C) = p(B|A)p(C|A)p(A) \]

B and C are conditionally independent
Given A

e.g., A is a disease, and we model
B and C as conditionally independent
symptoms given A

e.g. A is culprit, B is murder weapon
and C is fingerprints on door to the
guest’s room
Examples of 3-way Bayesian Networks

Independent Causes:
\[ p(A,B,C) = p(C|A,B)p(A)p(B) \]

“Explaining away” effect:
Given C, observing A makes B less likely
e.g., earthquake/burglary/alarm example

A and B are (marginally) independent
but become dependent once C is known
Examples of 3-way Bayesian Networks

Markov chain dependence:
\[ p(A,B,C) = p(C|B) \cdot p(B|A) \cdot p(A) \]

e.g. If Prof. Lathrop goes to party, then I might go to party. If I go to party, then my wife might go to party.
Bigger Example

Consider the following 5 binary variables:

- B = a burglary occurs at your house
- E = an earthquake occurs at your house
- A = the alarm goes off
- J = John calls to report the alarm
- M = Mary calls to report the alarm

Sample Query: What is \( P(B|M, J) \)?

Using full joint distribution to answer this question requires

- \( 2^5 - 1 = 31 \) parameters

Can we use prior domain knowledge to come up with a Bayesian network that requires fewer probabilities?
Constructing a Bayesian Network (1)

- Order variables in terms of causality (may be a partial order)
  e.g., \{E, B\} -> \{A\} -> \{J, M\}

- \[ P(J, M, A, E, B) = P(J, M \mid A, E, B) \cdot P(A \mid E, B) \cdot P(E, B) \]
  \[ \approx P(J, M \mid A) \cdot P(A \mid E, B) \cdot P(E) \cdot P(B) \]
  \[ \approx P(J \mid A) \cdot P(M \mid A) \cdot P(A \mid E, B) \cdot P(E) \cdot P(B) \]

- These conditional independence assumptions are reflected in the graph structure of the Bayesian network
The Resulting Bayesian Network

- **Burglary**: $P(B) = 0.001$
- **Earthquake**: $P(E) = 0.002$
- **Alarm**:
  - $P(E|B) = 0.95$
  - $P(E|\neg B) = 0.94$
  - $P(\neg E|B) = 0.29$
  - $P(\neg E|\neg B) = 0.001$
- **JohnCalls**:
  - $P(J|B) = 0.90$
  - $P(J|\neg B) = 0.05$
- **MaryCalls**:
  - $P(M|E) = 0.70$
  - $P(M|\neg E) = 0.01$
Constructing this Bayesian Network (2)

- \[ P(J,M,A,E,B) = P(J|A) \cdot P(M|A) \cdot P(A|E,B) \cdot P(E) \cdot P(B) \]

- There are 3 conditional probability tables to be determined:
  - \( P(J|A), P(M|A), P(A|E,B) \)
  - Requires \(2 + 2 + 4 = 8\) probabilities

- And 2 marginal probabilities \(P(E), P(B)\)

10 parameters in Bayesian Network; 31 parameters in joint distribution

- Where do these probabilities come from?
  - Expert knowledge
  - From data (relative frequency estimates) see Sections 20.1 & 20.2 (optional)
Number of Probabilities in Bayes Nets

- Consider \( n \) binary variables

- Unconstrained joint distribution requires \( O(2^n) \) probabilities

- If we have a Bayesian network, with a maximum of \( k \) parents for any node, then we need \( O(n \ 2^k) \) probabilities

Example

- Full unconstrained joint distribution
  - \( n = 30 \): need \( 10^9 \) probabilities for full joint distribution

- Bayesian network
  - \( n = 30, \ k = 4 \): need 480 probabilities
The Bayesian Network from a different Variable Ordering

\{M\} -> \{J\} -> \{A\} -> \{B\} -> \{E\}

\[
\]
Inference (Reasoning) in Bayes Nets

Consider answering a query in a Bayesian Network

\[ Q = \text{set of query variables} \]
\[ e = \text{evidence (set of instantiated variable-value pairs)} \]

Inference = computation of conditional distribution \( P(Q \mid e) \)

**Examples**

\[ P(\text{burglary} \mid \text{alarm}) \]
\[ P(\text{earthquake} \mid \text{JohnCalls, MaryCalls}) \]

Can we use structure of the Bayesian Network to answer queries efficiently?

Answer = yes

Generally speaking, complexity is inversely proportional to sparsity of graph
Inference by Variable Elimination

- Say that query is $P(B | j, m)$
  - $P(B | j, m) = P(B, j, m) / P(j, m) = \alpha P(B, j, m)$

- Apply evidence to expression for joint distribution
  - $P(j, m, A, E, B) = P(j | A)P(m | A)P(A | E, B)P(E)P(B)$

- Marginalize out A and E

\[
P(B | j, m) = \alpha \sum_a \sum_e p(j | a)p(m | a)p(a | e, B)P(e)P(B) = \alpha P(B) \sum_e P(e) \sum_a p(j | a)p(m | a)p(a | e, B)
\]

Distribution over variable B – i.e. over states \{b, ¬b\}

Sum is over states of variable A – i.e. \{a, ¬a\}
Complexity of Bayes Net Inference

- Assume the network is a polytree
  - Only a single directed path between any 2 nodes

- Complexity scales as $O(n \cdot m^{K+1})$
  - $n =$ number of variables
  - $m =$ arity of variables
  - $K =$ maximum number of parents for any node
  - Compare to $O(m^{n-1})$ for brute-force method

- If network is not a polytree?
  - Can cluster variables to render ‘new’ graph that is a tree
  - Complexity is then $O(n \cdot m^{W+1})$, where $W =$ # variables in largest cluster
Naïve Bayes Model

\[ P(C \mid X_1, \ldots, X_n) = \alpha \prod P(X_i \mid C) \, P(C) \]

Features \( X \) are conditionally independent given the class variable \( C \)

Widely used in machine learning
  e.g., spam email classification: \( X \)'s = counts of words in emails

Probabilities \( P(C) \) and \( P(X_i \mid C) \) can easily be estimated from labeled data
Two key assumptions:
1. hidden state sequence is Markov
2. observation $Y_t$ is Conditionally Independent of all other variables given $S_t$

Widely used in speech recognition, protein sequence models
Since this is a Bayesian network polytree, inference is linear in $n$
Summary

- Bayesian networks represent joint distributions using a graph.
- The graph encodes a set of conditional independence assumptions.
- Answering queries (i.e. inference) in a Bayesian network amounts to efficient computation of appropriate conditional probabilities.
- Probabilistic inference is intractable in the general case:
  - Can be done in linear time for certain classes of Bayesian networks.