CS 687
Jana Kosecka

Uncertainty, Bayesian Networks
Chapter 13, Russell and Norvig
Chapter 14, 14.1-14.3

Outline

• Uncertainty
• Probability
• Syntax and Semantics
• Inference
• Independence and Bayes' Rule
Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - $R$ = Is it raining?
  - $T$ = Is it hot or cold?
  - $D$ = How long will it take to drive to work?
  - $L$ = Where is the ghost?
- We denote random variables with capital letters
- Random variables have domains
  - $R$ in $\{\text{true}, \text{false}\}$ (often write as $\{+r, -r\}$)
  - $T$ in $\{\text{hot}, \text{cold}\}$
  - $D$ in $[0, \infty)$
  - $L$ in possible locations, maybe $\{(0,0), (0,1), \ldots\}$

Probability Distributions

- Associate a probability with each value
  - Temperature:
    \[
    P(T) \begin{array}{c|c} 
    \text{T} & P \\
    \hline 
    \text{hot} & 0.5 \\
    \text{cold} & 0.5 \\
    \end{array} 
    \]
  - Weather:
    \[
    P(W) \begin{array}{c|c} 
    \text{W} & P \\
    \hline 
    \text{sun} & 0.6 \\
    \text{rain} & 0.1 \\
    \text{fog} & 0.3 \\
    \text{meteor} & 0.0 \\
    \end{array} 
    \]
Probability Distributions

- Unobserved random variables have distributions

\[
P(T) \quad P(W)
\]

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hot</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>P(T)</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sun</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rain</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>fog</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>meteor</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

- A distribution is a TABLE of probabilities of values

- A probability (lower case value) is a single number
  \[P(W = \text{rain}) = 0.1\]

- Must have: \(\forall x \ P(X = x) \geq 0\) and \(\sum_x P(X = x) = 1\)

Shorthand notation:

\[P(\text{hot}) = P(T = \text{hot}),\]
\[P(\text{cold}) = P(T = \text{cold}),\]
\[P(\text{rain}) = P(W = \text{rain}).\]

\[\ldots\]

OK if all domain entries are unique

Joint Distributions

- A joint distribution over a set of random variables \(X_1, X_2, \ldots X_n\) specifies a real number for each assignment (or outcome):

\[P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)\]

\[P(x_1, x_2, \ldots x_n)\]

- Must obey:
  \[P(x_1, x_2, \ldots x_n) \geq 0\]
  \[\sum_{x_1, x_2, \ldots x_n} P(x_1, x_2, \ldots x_n) = 1\]

- Size of distribution if \(n\) variables with domain sizes \(d\)?
  - For all but the smallest distributions, impractical to write out!

\[
P(T, W)
\]

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>
Probabilistic Models

• A probabilistic model is a joint distribution over a set of random variables

• Probabilistic models:
  – (Random) variables with domains
  – Assignments are called outcomes
  – Joint distributions: say whether assignments (outcomes) are likely
  – Normalized: sum to 1.0
  – Ideally: only certain variables directly interact

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>T</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>F</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>F</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>T</td>
</tr>
</tbody>
</table>

Events

• An event is a set E of outcomes

\[ P(E) = \sum_{(x_1, \ldots, x_n) \in E} P(x_1 \ldots x_n) \]

• From a joint distribution, we can calculate the probability of any event

  – Probability that it’s hot AND sunny?
  – Probability that it’s hot?
  – Probability that it’s hot OR sunny?

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

• Typically, the events we care about are partial assignments, like \( P(T=\text{hot}) \)
Quiz: Events

- $P(+x, +y)$?
- $P(+x)$?
- $P(-y \text{ OR } +x)$?

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$+x$</td>
<td>$+y$</td>
<td>0.2</td>
</tr>
<tr>
<td>$+x$</td>
<td>$-y$</td>
<td>0.3</td>
</tr>
<tr>
<td>$-x$</td>
<td>$+y$</td>
<td>0.4</td>
</tr>
<tr>
<td>$-x$</td>
<td>$-y$</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables.
- Marginalization (summing out): Combine collapsed rows by adding.

<table>
<thead>
<tr>
<th>$T$</th>
<th>$W$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

\[
P(T) = \sum_t P(t, s) \]

\[
P(W) = \sum_s P(t, s) \]

\[
P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)\]
Quiz: Marginal Distributions

\[ P(X, Y) \]

\[
\begin{array}{c|c|c}
X & Y & P \\
\hline
+x & +y & 0.2 \\
+x & -y & 0.3 \\
-x & +y & 0.4 \\
-x & -y & 0.1 \\
\end{array}
\]

\[ P(x) = \sum_y P(x, y) \]

\[ P(y) = \sum_x P(x, y) \]

\[ P(X) \]

<table>
<thead>
<tr>
<th>X</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+x</td>
<td></td>
</tr>
<tr>
<td>-x</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(Y) \]

<table>
<thead>
<tr>
<th>Y</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>+y</td>
<td></td>
</tr>
<tr>
<td>-y</td>
<td></td>
</tr>
</tbody>
</table>

Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the definition of a conditional probability

\[ P(a|b) = \frac{P(a, b)}{P(b)} \]

\[ P(W = s | T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4 \]

\[ = P(W = s, T = c) + P(W = r, T = c) \]

\[ = 0.2 + 0.3 = 0.5 \]
Quiz: Conditional Probabilities

\[
P(X, Y)
\]

<table>
<thead>
<tr>
<th>(X)</th>
<th>(Y)</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+x</td>
<td>+y</td>
<td>0.2</td>
</tr>
<tr>
<td>+x</td>
<td>-y</td>
<td>0.3</td>
</tr>
<tr>
<td>-x</td>
<td>+y</td>
<td>0.4</td>
</tr>
<tr>
<td>-x</td>
<td>-y</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- \(P(+x \mid +y)\)
- \(P(+x \mid -y)\)
- \(P(-x \mid +y)\)
- \(P(-x \mid -y)\)

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

\[
P(W \mid T = \text{hot})
\]

\[
P(W \mid T = \text{cold})
\]

Joint Distribution

<table>
<thead>
<tr>
<th>T (\times) W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot (\times) sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot (\times) rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold (\times) sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold (\times) rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Normalization Trick

\[ P(T, W) \]

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

\[ P(W \mid T = t) = \frac{P(W \mid T = t)}{P(T = t)} \]

\[ = \frac{P(W = s, T = t) + P(W = r, T = t)}{P(W = s, T = t) + P(W = r, T = t)} \]

\[ = \frac{0.2 + 0.3}{0.4 + 0.6} = 0.4 \]

\[ P(W \mid T = c) \]

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>rain</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Normalization Trick

\[ P(W \mid T = t) = \frac{P(W = s, T = t) + P(W = r, T = t)}{P(T = t)} \]

\[ = \frac{P(W = s, T = t) + P(W = r, T = t)}{0.4 + 0.6} \]

\[ = \frac{0.2 + 0.3}{0.4 + 0.6} = 0.4 \]

\[ P(W \mid T = c) \]

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>rain</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Select the joint probabilities matching the evidence

NORMALIZE the selection (make it sum to one)
Normalization Trick

\[ P(T, W) \]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**SELECT** the joint probabilities matching the evidence

\[ P(c, W) \]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**NORMALIZE** the selection (make it sum to one)

\[ P(W|T = c) \]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>rain</td>
<td>0.6</td>
</tr>
</tbody>
</table>

\[ P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{P(x_1, x_2)}{\sum_{x_1} P(x_1, x_2)} \]

Quiz: Normalization Trick

- \( P(X \mid Y=-y) \) ?

\[ P(X, Y) \]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+x</td>
<td>+y</td>
<td>0.2</td>
</tr>
<tr>
<td>+x</td>
<td>-y</td>
<td>0.3</td>
</tr>
<tr>
<td>-x</td>
<td>+y</td>
<td>0.4</td>
</tr>
<tr>
<td>-x</td>
<td>-y</td>
<td>0.1</td>
</tr>
</tbody>
</table>
To Normalize

- (Dictionary) To bring or restore to a normal condition

- Procedure:
  - Step 1: Compute $Z = \sum$ over all entries
  - Step 2: Divide every entry by $Z$

- Example 1

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Normalize $Z = 0.5$

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>rain</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Example 2

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>20</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>5</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>10</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>15</td>
</tr>
</tbody>
</table>

Normalize $Z = 50$

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g., conditional from joint)

- We generally compute conditional probabilities:
  - $P$($on$ $time$ | no reported accidents) = 0.90
  - These represent the agent’s beliefs given the evidence

- Probabilities change with new evidence:
  - $P$($on$ $time$ | no accidents, 5 a.m.) = 0.95
  - $P$($on$ $time$ | no accidents, 5 a.m., raining) = 0.80
  - Observing new evidence causes beliefs to be updated
Inference by Enumeration

- General case:
  - Evidence variables: $E_1 \ldots E_k = e_1 \ldots e_k$
  - Query* variable: $Q$
  - Hidden variables: $H_1 \ldots H_r$

- We want:
  * Works fine with multiple query variables, too
  $P(Q|e_1 \ldots e_k)$

- Step 1: Select the entries consistent with the evidence

- Step 2: Sum out $H$ to get joint of Query and evidence

- Step 3: Normalize

\[
P(Q|e_1 \ldots e_k) = \frac{1}{Z} \sum_{h_1 \ldots h_r} P(h_1 \ldots h_r, e_1 \ldots e_k)
\]

\[
Z = \sum_{Q} P(Q, e_1 \ldots e_k)
\]

\[
P(Q|e_1 \ldots e_k) = \frac{1}{Z} P(Q, e_1 \ldots e_k)
\]

---

Inference by Enumeration

- $P(W)$?

- $P(W | winter)$?

- $P(W | winter, hot)$?
Inference by Enumeration

- Obvious problems:
  - Worst-case time complexity $O(dn)$
  - Space complexity $O(dn)$ to store the joint distribution

The Product Rule

- Sometimes have conditional distributions but want the joint

$$ P(y)P(x|y) = P(x, y) \quad \iff \quad P(x|y) = \frac{P(x, y)}{P(y)} $$
The Product Rule

\[ P(y) P(x|y) = P(x, y) \]

- Example:

|   | P(W) | P(D|W) | P(D, W) |
|---|------|--------|---------|
| R | sun  | wet    | D       |
|   | 0.8  | sun    | 0.1     |
|   | rain | wet    | 0.7     |
|   | 0.2  | dry    | 0.3     |

The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions:

\[ P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \]

\[ P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1}) \]

- Why is this always true?
Bayes Rule

- Two ways to factor a joint distribution over two variables:
  \[ P(x, y) = P(x|y)P(y) = P(y|x)P(x) \]

- Dividing, we get:
  \[ P(x|y) = \frac{P(y|x)}{P(y)} P(x) \]

- Why is this all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we'll see later (e.g. ASR, MT)

- In the running for most important AI equation!
Inference with Bayes’ Rule

• Example: Diagnostic probability from causal probability:
  \[ P(\text{cause} | \text{effect}) = \frac{P(\text{effect} | \text{cause}) P(\text{cause})}{P(\text{effect})} \]

• Example:
  – M: meningitis, S: stiff neck
  \[ P(+m) = 0.0001 \]
  \[ P(+s | m) = 0.8 \]
  \[ P(+s | m') = 0.01 \]

– Note: posterior probability of meningitis still very small
– Note: you should still get stiff necks checked out! Why?

\[
P(+m | +s) = \frac{P(+s | +m) P(+m)}{P(+s | +m) P(+m) + P(+s | -m) P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999} = 0.007937
\]

Quiz: Bayes’ Rule

• Given:

<table>
<thead>
<tr>
<th>P(W)</th>
<th>0.8</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>sun</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>rain</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

| P(D | W)     | P       | P   |
|---------|---------|-----|
| D, W    | wet, sun | 0.1 |
| D, dry  | wet, rain | 0.7 |
| D, rain | dry, rain | 0.3 |

• What is P(W | dry) ?
Probabilistic Models

- Models describe how (a portion of) the world works

- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - “All models are wrong; but some are useful.” — George E. P. Box

- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)
  - Example: value of information

Independence

- Two variables are independent if:

\[ P(x, y) = P(x)P(y) \]

  - This says that their joint distribution factors into a product of two simpler distributions

  - Another form:

\[ P(x|y) = P(x) \]

  - We write:

\[ X \perp Y \]

- Independence is a simplifying modeling assumption

  - Empirical joint distributions: at best “close” to independent

  - What could we assume for {Weather, Traffic, Cavity, Toothache}?
Example: Independence

\[ P(T) \]
\[
\begin{array}{c|c}
T & P \\
\hline
\text{hot} & 0.5 \\
\text{cold} & 0.5 \\
\end{array}
\]

\[ P_1(T, W) \]
\[
\begin{array}{c|c|c}
T & W & P \\
\hline
\text{hot} & \text{sun} & 0.4 \\
\text{hot} & \text{rain} & 0.1 \\
\text{cold} & \text{sun} & 0.2 \\
\text{cold} & \text{rain} & 0.3 \\
\end{array}
\]

\[ P_2(T, W) \]
\[
\begin{array}{c|c|c}
T & W & P \\
\hline
\text{hot} & \text{sun} & 0.3 \\
\text{hot} & \text{rain} & 0.2 \\
\text{cold} & \text{sun} & 0.3 \\
\text{cold} & \text{rain} & 0.2 \\
\end{array}
\]

\[ P(W) \]
\[
\begin{array}{c|c}
W & P \\
\hline
\text{sun} & 0.6 \\
\text{rain} & 0.4 \\
\end{array}
\]

----

Example: Independence

- N fair, independent coin flips:

\[ P(X_1) \]
\[
\begin{array}{c|c}
H & 0.5 \\
T & 0.5 \\
\end{array}
\]

\[ P(X_2) \]
\[
\begin{array}{c|c}
H & 0.5 \\
T & 0.5 \\
\end{array}
\]

\[ \ldots \]

\[ P(X_n) \]
\[
\begin{array}{c|c}
H & 0.5 \\
T & 0.5 \\
\end{array}
\]

\[ P(X_1, X_2, \ldots X_n) \]
\[
2^n \]
Conditional Independence

- \( P(\text{Toothache, Cavity, Catch}) \)

- If I have a cavity, the probability that the probe catches in it doesn’t depend on whether I have a toothache:
  - \( P(+\text{catch} \mid +\text{toothache}, +\text{cavity}) = P(+\text{catch} \mid +\text{cavity}) \)

- The same independence holds if I don’t have a cavity:
  - \( P(+\text{catch} \mid +\text{toothache}, -\text{cavity}) = P(+\text{catch} \mid -\text{cavity}) \)

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

- Equivalent statements:
  - \( P(\text{Toothache} \mid \text{Catch, Cavity}) = P(\text{Toothache} \mid \text{Cavity}) \)
  - \( P(\text{Toothache, Catch} \mid \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) \)
  - \( P(\text{Catch} \mid \text{Cavity}) \)
  - One can be derived from the other easily

Conditional Independence

- Unconditional (absolute) independence very rare (why?)

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

- \( X \) is conditionally independent of \( Y \) given \( Z \)
  \[ X \perp Y \mid Z \]

  if and only if:
  \[ \forall x, y, z : P(x, y \mid z) = P(x \mid z)P(y \mid z) \]

  or, equivalently, if and only if
  \[ \forall x, y, z : P(x \mid z, y) = P(x \mid z) \]
Conditional Independence

• What about this domain:
  – Traffic
  – Umbrella
  – Raining

Conditional Independence and the Chain Rule

• Chain rule:
  \[ P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots \]

• Trivial decomposition:
  \[ P(\text{Traffic, Rain, Umbrella}) = \]
  \[ \frac{P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain, Traffic})}{P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})} \]
  • With assumption of conditional independence:
    \[ P(\text{Traffic, Rain, Umbrella}) = \]
    \[ \frac{P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})}{P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})} \]
  • Bayes’ nets / graphical models help us express conditional independence assumptions
Bayes’ Nets: Big Picture

Two problems with using full joint distribution tables as our probabilistic models:

- Unless there are only a few variables, the joint is WAY too big to represent explicitly
- Hard to learn (estimate) anything empirically about more than a few variables at a time

Bayes’ nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)

- More properly called graphical models
- We describe how variables locally interact
- Local interactions chain together to give global, indirect interactions
- For about 10 min, we’ll be vague about how these interactions are specified
Example Bayes’ Net: Insurance

Example Bayes’ Net: Car
Graphical Model Notation

- **Nodes**: variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)

- **Arcs**: interactions
  - Similar to CSP constraints
  - Indicate "direct influence" between variables
  - Formally: encode conditional independence (more later)

- For now: imagine that arrows mean direct causation (in general, they don’t!)

---

Example: Coin Flips

- N independent coin flips

\[
\begin{align*}
X_1 & \\
X_2 & \\
\cdots & \\
X_n & 
\end{align*}
\]

- No interactions between variables: absolute independence
Example: Traffic

- Variables:
  - R: It rains
  - T: There is traffic

- Model 1: independence
  - Model 2: rain causes traffic

- Why is an agent using model 2 better?

Example: Traffic II

- Let’s build a causal graphical model!
- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame
  - C: Cavity
Example: Alarm Network

• Variables
  – B: Burglary
  – A: Alarm goes off
  – M: Mary calls
  – J: John calls
  – E: Earthquake!

Bayes’ Net Semantics

• A set of nodes, one per variable X
• A directed, acyclic graph
• A conditional distribution for each node
  – A collection of distributions over X, one for each combination of parents’ values
  \[ P(X|a_1 \ldots a_n) \]
  – CPT: conditional probability table
  – Description of a noisy "causal" process

A Bayes net = Topology (graph) + Local Conditional Probabilities
Probabilities in BNs

- Bayes’ nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:
    
    \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]

- Example:

\[ P(+\text{cavity}, +\text{catch}, -\text{toothache}) \]

Probabilities in BNs

- Why are we guaranteed that setting 
  
  \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]

  results in a proper joint distribution?

- Chain rule (valid for all distributions): 
  
  \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|x_1 \ldots x_{i-1}) \]

- Assume conditional independences: 
  
  \[ P(x_i|x_1, \ldots, x_{i-1}) = P(x_i|\text{parents}(X_i)) \]

  → Consequence: 
  
  \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]

- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies
Only distributions whose variables are absolutely independent can be represented by a Bayes net with no arcs.

Example: Coin Flips

\[ P(h, h, t, h) = \]

Example: Traffic

\[ P(R) \]

\[ P(T|R) \]

\[ P(\pm r, -t) = \]
Example: Alarm Network

Example: Traffic

• Causal direction
Example: Reverse Traffic

• Reverse causality?

\[
\begin{array}{c|cc}
\text{T} & \text{P(T)} & \\
+ & +t & 9/16 \\
- & -t & 7/16 \\
\end{array}
\]

\[
\begin{array}{c|cc}
\text{R} & \text{P(R|T)} & \\
+ & +r & 1/3 \\
- & -r & 2/3 \\
\end{array}
\]

\[
\begin{array}{c|cc}
\text{P(T, R)} & & \\
+ & +r & 3/16 \\
+ & -r & 1/16 \\
- & +r & 6/16 \\
- & -r & 6/16 \\
\end{array}
\]

Causality?

• When Bayes’ nets reflect the true causal patterns:
  – Often simpler (nodes have fewer parents)
  – Often easier to think about
  – Often easier to elicit from experts

• BNs need not actually be causal
  – Sometimes no causal net exists over the domain (especially if variables are missing)
  – E.g. consider the variables Traffic and Drips
  – End up with arrows that reflect correlation, not causation

• What do the arrows really mean?

\[ P(x_i|x_1, \ldots, x_{i-1}) = P(x_i|\text{parents}(X_i)) \]

  – Topology may happen to encode causal structure
  – Topology really encodes conditional independence
Bayes’ Nets

- So far: how a Bayes’ net encodes a joint distribution
- Next: how to answer queries about that distribution
  - Today:
    - First assembled BNs using an intuitive notion of conditional independence as causality
    - Then saw that key property is conditional independence
  - Main goal: answer queries about conditional independence and influence
- After that: how to answer numerical queries (inference)