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
Independence

- Two variables are *independent* if:

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

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

- Another form:

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

- 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?

### $P(T)$

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

### $P_1(T, W)$

<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>

### $P_2(T, W)$

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

### $P(W)$

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

- N fair, independent coin flips:

\[
P(X_1) \quad \quad P(X_2) \quad \quad \ldots \quad \quad P(X_n)
\]

<table>
<thead>
<tr>
<th>H</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\[
2^n \left\{ P(X_1, X_2, \ldots, X_n) \right\}
\]
Conditional Independence
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:
  - $P(\text{Catch} \mid \text{Toothache}, \text{Cavity}) = P(\text{Catch} \mid \text{Cavity})$

- 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

- 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

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

\[ P(x \mid z, y) = \frac{P(x, z, y)}{P(z, y)} = \frac{P(x, y \mid z)P(z)}{P(y \mid z)P(z)} = \frac{P(x \mid z)P(y \mid z)P(z)}{P(y \mid z)P(z)} \]
Conditional Independence

- What about this domain:
  - Traffic
  - Umbrella
  - Raining
Conditional Independence

- What about this domain:
  - Fire
  - Smoke
  - Alarm
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:**
  \[ P(\text{T}rafic, \text{Rain}, \text{Umbrella}) = P(\text{Rain})P(\text{T}rafic|\text{Rain})P(\text{Umbrella}|\text{Rain}, \text{T}rafic) \]

- **With assumption of conditional independence:**
  \[ P(\text{T}rafic, \text{Rain}, \text{Umbrella}) = P(\text{Rain})P(\text{T}rafic|\text{Rain})P(\text{Umbrella}|\text{Rain}) \]

- Bayes’ nets / graphical models help us express conditional independence assumptions
- Each sensor depends only on where the ghost is.

- That means, the two sensors are conditionally independent, given the ghost position.

- T: Top square is red
  B: Bottom square is red
  G: Ghost is in the top

- Givens:
  \( P(\ +g\ ) = 0.5 \)
  \( P(\ -g\ ) = 0.5 \)
  \( P(\ +t\ |\ +g\ ) = 0.8 \)
  \( P(\ +t\ |\ -g\ ) = 0.4 \)
  \( P(\ +b\ |\ +g\ ) = 0.4 \)
  \( P(\ +b\ |\ -g\ ) = 0.8 \)

\[
P(T,B,G) = P(G) P(T|G) P(B|G)
\]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>B</td>
<td>G</td>
<td>P(T,B,G)</td>
<td></td>
</tr>
<tr>
<td>+t</td>
<td>+b</td>
<td>+g</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>+t</td>
<td>+b</td>
<td>-g</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>+t</td>
<td>-b</td>
<td>+g</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>+t</td>
<td>-b</td>
<td>-g</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>+b</td>
<td>+g</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>+b</td>
<td>-g</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>-b</td>
<td>+g</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>-b</td>
<td>-g</td>
<td>0.06</td>
<td></td>
</tr>
</tbody>
</table>
Bayes’Nets: Big Picture
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
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

\[ X_1 \quad X_2 \quad \ldots \quad X_n \]

- 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: Alarm Network

- **Variables**
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!
Example: Humans

- $G$: human’s goal / human’s reward parameters
- $S$: state of the physical world
- $A$: human’s action
Example: Alarm Network

- **Variables**
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!
Example: Traffic II

- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame
  - C: Cavity
Bayes’ Net Semantics
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:

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

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

  \rightarrow 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: Traffic

\[ P(R) \]

\[
\begin{array}{|c|c|}
\hline
+r & 1/4 \\
-r & 3/4 \\
\hline
\end{array}
\]

\[ P(T|R) \]

\[
\begin{array}{|c|c|c|}
\hline
+r & +t & 3/4 \\
-\text{t} & 1/4 \\
-\text{t} & 1/2 \\
\hline
\end{array}
\]

\[
P(+r, -t) = P(+r)P(-t|+r) = \frac{1}{4} \times \frac{1}{4}
\]
Example: Alarm Network

### Alarm Network

- **Burglary** (B)
  - P(B) = 0.001
  - P(B) = 0.999

- **Earthquake** (E)
  - P(E) = 0.002
  - P(E) = 0.998

- **Alarm**

- **John calls**

- **Mary calls**

### Probabilities

#### Conditional Probabilities

| A  | J  | P(J|A) |
|----|----|--------|
| +a | +j | 0.9    |
| +a | -j | 0.1    |
| -a | +j | 0.05   |
| -a | -j | 0.95   |

| A  | M  | P(M|A) |
|----|----|--------|
| +a | +m | 0.7    |
| +a | -m | 0.3    |
| -a | +m | 0.01   |
| -a | -m | 0.99   |

### Joint Probabilities

P(A|B,E)P(E)P(B)P(M|A)P(J|A)
## Example: Traffic

### Causal direction

$$P(R)$$

<table>
<thead>
<tr>
<th></th>
<th>+r</th>
<th>1/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>-r</td>
<td>3/4</td>
<td></td>
</tr>
</tbody>
</table>

$$P(T \mid R)$$

<table>
<thead>
<tr>
<th></th>
<th>+r</th>
<th>+t</th>
<th>3/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td></td>
<td>+t</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td></td>
<td>-t</td>
<td>1/4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-r</th>
<th>+t</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td></td>
<td>+t</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td></td>
<td>-t</td>
<td>1/2</td>
</tr>
</tbody>
</table>

$$P(T, R)$$

<table>
<thead>
<tr>
<th></th>
<th>+r</th>
<th>+t</th>
<th>3/16</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td></td>
<td>+t</td>
<td></td>
</tr>
<tr>
<td>+r</td>
<td></td>
<td>-t</td>
<td>1/16</td>
</tr>
<tr>
<td>-r</td>
<td></td>
<td>+t</td>
<td>6/16</td>
</tr>
<tr>
<td>-r</td>
<td></td>
<td>-t</td>
<td>6/16</td>
</tr>
</tbody>
</table>
Example: Reverse Traffic

- Reverse causality?

\[
\begin{align*}
P(T) & \\
+ t & 9/16 \\
- t & 7/16 \\
\end{align*}
\]

\[
\begin{align*}
P(R | T) & \\
+ t & + r & 1/3 \\
- r & & 2/3 \\
- t & + r & 1/7 \\
+ t & & 6/7 \\
- r & - t & 6/16 \\
- r & - t & 6/16 \\
\end{align*}
\]
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?**
  - Topology may happen to encode causal structure
  - **Topology really encodes conditional independence**
    \[ P(x_i | x_1, \ldots x_{i-1}) = P(x_i | \text{parents}(X_i)) \]