CSE 573: Artificial Intelligence

Hanna Hajishirzi
Bayes Nets

slides adapted from
Dan Klein, Pieter Abbeel ai.berkeley.edu
And Dan Weld, Luke Zettelmoyer
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)
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: $X \independent Y$

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?

<table>
<thead>
<tr>
<th>$P_1(T, W)$</th>
<th>$P(T)$</th>
<th>$P_2(T, W)$</th>
<th>$P(W)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T$</td>
<td>$P$</td>
<td>$T$</td>
</tr>
<tr>
<td>T</td>
<td>hot</td>
<td>0.5</td>
<td>hot</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>0.5</td>
<td>cold</td>
</tr>
<tr>
<td>W</td>
<td>hot</td>
<td>0.4</td>
<td>hot</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>0.2</td>
<td>cold</td>
</tr>
</tbody>
</table>

$P(T)$ is the probability of the temperature, $P_1(T, W)$ and $P_2(T, W)$ are the joint probability distributions of temperature and weather, and $P(W)$ is the probability of the weather.
Example: Independence

- N fair, independent coin flips:

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

\[
\begin{array}{|c|c|}
\hline
H & 0.5 \\
\hline
T & 0.5 \\
\hline
\end{array}
\quad \begin{array}{|c|c|}
\hline
H & 0.5 \\
\hline
T & 0.5 \\
\hline
\end{array}
\quad \ldots 
\quad \begin{array}{|c|c|}
\hline
H & 0.5 \\
\hline
T & 0.5 \\
\hline
\end{array}
\]

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

\[
2^n
\]
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 \textit{conditionally independent} of Toothache given Cavity:
  - $P(\text{Catch} \mid \text{Toothache, Cavity}) = P(\text{Catch} \mid \text{Cavity})$

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

  $$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

  or, equivalently, if and only if

  $$\forall x, y, z : P(x|z, y) = P(x|z)$$
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:
  \[ P(\text{Traffic, Rain, Umbrella}) = P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain, Traffic}) \]

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

- We can represent joint distributions by multiplying these simpler local distributions.
- 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
  - 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: Traffic II

- 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!
Example: Alarm Network

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

![Diagram of Alarm Network]
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

\[ P(X | A_1 \ldots A_n) \]

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:

\[
P(\text{+cavity, +catch, -toothache}) = P(\text{-toothache}|\text{+cavity})P(\text{+catch}|\text{+cavity})P(\text{+cavity})
\]
Bayes’ Net Representation

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over $X$, one for each combination of parents’ values
    \[ P(X|a_1 \ldots a_n) \]
- 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:
    \[ P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]
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: Traffic

\[ P(R) \]

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

\[ P(T|R) \]

\begin{tabular}{|c|c|c|}
\hline
+r & +t & 3/4 \\
\hline
+r & -t & 1/4 \\
\hline
-r & +t & 1/2 \\
\hline
-r & -t & 1/2 \\
\hline
\end{tabular}

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

Burglary

Earthquake

Alarm

John calls

Mary calls

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+b</td>
<td>0.001</td>
</tr>
<tr>
<td>-b</td>
<td>0.999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+e</td>
<td>0.002</td>
</tr>
<tr>
<td>-e</td>
<td>0.998</td>
</tr>
</tbody>
</table>

| 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  |

| B     | E   | A    | P(A|B,E) |
|-------|-----|------|---------|
| +b    | +e  | +a   | 0.95    |
| +b    | +e  | -a   | 0.05    |
| +b    | -e  | +a   | 0.94    |
| +b    | -e  | -a   | 0.06    |
| -b    | +e  | +a   | 0.29    |
| -b    | +e  | -a   | 0.71    |
| -b    | -e  | +a   | 0.001   |
| -b    | -e  | -a   | 0.999   |

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

- Causal direction

\[ P(R) \]
\[
\begin{array}{cc}
+r & 1/4 \\
-r & 3/4 \\
\end{array}
\]

\[ P(T | R) \]
\[
\begin{array}{ccc}
+r & +t & 3/4 \\
+r & -t & 1/4 \\
-r & +t & 1/2 \\
-r & -t & 1/2 \\
\end{array}
\]

\[ P(T, R) \]
\[
\begin{array}{ccc}
+r & +t & 3/16 \\
+r & -t & 1/16 \\
-r & +t & 6/16 \\
-r & -t & 6/16 \\
\end{array}
\]
Example: Reverse Traffic

- Reverse causality?

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

\[
P(R|T) = \begin{array}{|c|c|c|} \hline +t & +r & 1/3 \\
& -r & 2/3 \\
+t & -r & 1/7 \\
- & +r & 6/7 \\
- & -r & 6/7 \\
\hline \end{array}
\]

\[
P(T, R) = \begin{array}{|c|c|c|} \hline +r & +t & 3/16 \\
+ & - & 1/16 \\
- & +t & 6/16 \\
- & - & 6/16 \\
\hline \end{array}
\]
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?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence
    \[ P(x_i|x_1, \ldots, x_{i-1}) = P(x_i|\text{parents}(X_i)) \]
Bayes’ Net Representation

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over $X$, one for each combination of parents’ values
    $$P(X|a_1 \ldots a_n)$$
- 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:
    $$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i))$$