CSEP 573: Artificial Intelligence

Bayesian Networks

Ali Farhadi

Many slides over the course adapted from either Luke Zettlemoyer, Pieter Abbeel, Dan Klein, Stuart Russell or Andrew Moore
Outline

- Probabilistic models (and inference)
  - Bayesian Networks (BNs)
  - Independence in BNs
  - Inference in BNs
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:

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

- 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

- N fair, independent coin flips:

<table>
<thead>
<tr>
<th></th>
<th>$P(X_1)$</th>
<th>$P(X_2)$</th>
<th>$P(X_n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>0.5</td>
<td>h 0.5</td>
<td>h 0.5</td>
</tr>
<tr>
<td>t</td>
<td>0.5</td>
<td>t 0.5</td>
<td>t 0.5</td>
</tr>
</tbody>
</table>

$P(X_1, X_2, \ldots, X_n)$

$2^n$
Conditional Independence

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

- Conditional independence is our most basic and robust form of knowledge about uncertain environments:
  \[
  \forall x, y, z : P(x, y | z) = P(x | z) P(y | z)
  \]
  \[
  \forall x, y, z : P(x | z, y) = P(x | z)
  \]

- What about these domain:
  - Traffic, Umbrella, Raining
  - Toothache, Cavity, Catch
Conditional Independence and the Chain Rule

- Trivial decomposition:

  \[ P(\text{Traffic, Rain, Umbrella}) = P(\text{Rain})P(\text{Traffic|Rain})P(\text{Umbrella|Rain, Traffic}) \]

- With assumption of conditional independence:

  \[ P(\text{Traffic, Rain, Umbrella}) = P(\text{Rain})P(\text{Traffic|Rain})P(\text{Umbrella|Rain}) \]

- Bayes’ nets/ graphical models help us express conditional independence assumptions
Ghostbusters Chain Rule

- 2-position maze, each sensor indicates ghost location

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

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

- Can assume:
  \( P( +g ) = 0.5 \)
  \( P( +t \mid +g ) = 0.8 \)
  \( P( +t \mid -g ) = 0.4 \)
  \( P( +b \mid +g ) = 0.4 \)
  \( P( +b \mid -g ) = 0.8 \)

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

<table>
<thead>
<tr>
<th>T</th>
<th>B</th>
<th>G</th>
<th>( P(T,B,G) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>+b</td>
<td>+g</td>
<td>0.16</td>
</tr>
<tr>
<td>+t</td>
<td>+b</td>
<td>-g</td>
<td>0.16</td>
</tr>
<tr>
<td>+t</td>
<td>-b</td>
<td>+g</td>
<td>0.24</td>
</tr>
<tr>
<td>+t</td>
<td>-b</td>
<td>-g</td>
<td>0.04</td>
</tr>
<tr>
<td>-t</td>
<td>+b</td>
<td>+g</td>
<td>0.04</td>
</tr>
<tr>
<td>-t</td>
<td>+b</td>
<td>-g</td>
<td>0.24</td>
</tr>
<tr>
<td>-t</td>
<td>-b</td>
<td>+g</td>
<td>0.06</td>
</tr>
<tr>
<td>-t</td>
<td>-b</td>
<td>-g</td>
<td>0.06</td>
</tr>
</tbody>
</table>
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
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)
Example: Flip Coins

• N independent flip coins

• No interactions between variables
  – Absolute independence
Example Bayes’ Net: Car
Example Bayes’ Net: Insurance
Example: Traffic

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

Model 1: independence

Model 2: rain is conditioned on traffic
- Why is an agent using model 2 better?

Model 3: traffic is conditioned on rain
- Is this better than model 2?
Example: Traffic II

- Let’s build a 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

- Let’s formalize the semantics of a Bayes’ net
- 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

\[ A \text{ Bayes net} = \text{Topology (graph)} + \text{Local Conditional Probabilities} \]
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:

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

- This lets us reconstruct any entry of the full joint
- Not every BN can represent every joint distribution
  - The topology enforces certain independence assumptions
  - Compare to the exact decomposition according to the chain rule!
Probabilities in BN

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:

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

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

Not every BN can represent every joint distribution

- The topology enforces certain conditional independencies
Bayes Net Probabilities

• Bayes nets compactly represent joint distributions (instead of big joint table)
  – A joint distribution using chain rule

\[ P(x_1...x_n) = \prod_i P(x_i \mid \text{parents}(x_i)) \]

• \{Cavity, Toothache, Catch\}
  P(Cavity, Toothache, ~Catch) ?

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

P(\sim\text{catch} \mid \text{cavity})
Example: Flip Coins

• N independent flip coins

\[
\begin{array}{|c|c|}
\hline
\text{P} & 0.5 \\
\hline
\text{Head} & 0.5 \\
\hline
\text{Tail} & 0.5 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
\text{P} & 0.5 \\
\hline
\text{Head} & 0.5 \\
\hline
\text{Tail} & 0.5 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
\text{P} & 0.5 \\
\hline
\text{Head} & 0.5 \\
\hline
\text{Tail} & 0.5 \\
\hline
\end{array}
\]

• \( P(h, h, t, h) \)?

- No interactions between variables: absolute independence
Example: Traffic

\[ P(R) \]

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

\[ P(T|R) \]

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

\[ P(T, R) \]

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

\[
P(+b, -e, +a, -j, +m) = P(+b)P(-e)P(+a|+b, -e)P(-j|+a)P(+m|+a) = 0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7
\]
Changing Bayes’ Net Structure

- The same joint distribution can be encoded in many different Bayes’ nets

- Analysis question: given some edges, what other edges do you need to add?
  - One answer: fully connect the graph
  - Better answer: don’t make any false conditional independence assumptions
Example: Independence

- For this graph, you can fiddle with $\theta$ (the CPTs) all you want, but you won’t be able to represent any distribution in which the flips are dependent!

\[
\begin{array}{c|c}
X_1 & X_2 \\
\hline
P(X_1) & P(X_2) \\
\hline
h & h \\
0.5 & 0.5 \\
t & t \\
0.5 & 0.5 \\
\end{array}
\]
Example: Coins

- Extra arcs don’t prevent representing independence, just allow non-independence

- Adding unneeded arcs isn’t wrong, it’s just inefficient
Size of a Bayes Net

- How big is a joint distribution over $N$ Boolean variables?
  \[ 2^N \]

- How big is an $N$-node net if nodes have up to $k$ parents?
  \[ O(N \times 2^{k+1}) \]

- Both give you the power to calculate
  \[ P(X_1, X_2, \ldots, X_n) \]

- BNs: Huge space savings!
- Also easier to elicit local CPTs
- Also faster to answer queries (coming)