CIS 471/571 (Winter 2019): Introduction to Artificial Intelligence

Lecture 12: Bayes Nets

Thanh H. Nguyen

Source: http://ai.berkeley.edu/home.html
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
Probability Recap

- Conditional probability
  \[ P(x|y) = \frac{P(x, y)}{P(y)} \]

- Product rule
  \[ P(x, y) = P(x|y)P(y) \]

- Chain rule
  \[ P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots \]
  \[ = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \]
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 of simpler distributions
- Another form:
  \[ \forall x, y : P(x | y) = P(x) \]
  
  - We write: \( X \perp\!\!\!\!\!\!\perp Y \)

- 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?

\[
P(T)
\]

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

\[
P_1(T, W)
\]

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

\[
P_2(T, W)
\]

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

\[
P(W)
\]

\[
\begin{array}{|c|c|}
\hline
W & P \\
\hline
\text{sun} & 0.6 \\
\text{rain} & 0.4 \\
\hline
\end{array}
\]
Example: Independence

- N fair, independent coin flips:

\[
\begin{array}{c|c}
X_1 & P(X_1) \\
\hline
\text{H} & 0.5 \\
\text{T} & 0.5 \\
\end{array}
\quad \begin{array}{c|c}
X_2 & P(X_2) \\
\hline
\text{H} & 0.5 \\
\text{T} & 0.5 \\
\end{array}
\quad \ldots
\quad \begin{array}{c|c}
X_n & P(X_n) \\
\hline
\text{H} & 0.5 \\
\text{T} & 0.5 \\
\end{array}
\]

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

- \( P(\text{Toothache}, \text{Cavity}, \text{Catch}) \)
- 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:
  - \( P(\text{Toothache} \mid \text{Catch}, \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) \)
  - \( P(\text{Toothache}, \text{Catch} \mid \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) \cdot P(\text{Catch} \mid \text{Cavity}) \)
  - 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

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

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

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

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

<table>
<thead>
<tr>
<th></th>
<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></td>
<td>0.16</td>
</tr>
<tr>
<td>+t</td>
<td>+b</td>
<td>-g</td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>+t</td>
<td>-b</td>
<td>+g</td>
<td></td>
<td>0.24</td>
</tr>
<tr>
<td>+t</td>
<td>-b</td>
<td>-g</td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>-t</td>
<td>+b</td>
<td>+g</td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>-t</td>
<td>+b</td>
<td>-g</td>
<td></td>
<td>0.24</td>
</tr>
<tr>
<td>-t</td>
<td>-b</td>
<td>+g</td>
<td></td>
<td>0.06</td>
</tr>
<tr>
<td>-t</td>
<td>-b</td>
<td>-g</td>
<td></td>
<td>0.06</td>
</tr>
</tbody>
</table>
Bayes' Nets: Big Picture

- Encoding Complex Distributions
- In 12 Easy Steps!
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

\[ X_1, X_2, \ldots, X_n \]

- No interactions between variables: absolute independence
Example: Traffic

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

- Model 1: independence

- Why is an agent using model 2 better?

- Model 2: rain causes traffic
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 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 \textit{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:
  \[ P(x_i|x_1, \ldots x_{i-1}) = P(x_i|\text{parents}(X_i)) \]
  \[ \rightarrow \text{Consequence:} \quad P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]

- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies
Only distributions whose variables are absolutely independent can be represented by a Bayes’ net with no arcs.
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>(3/4)</td>
</tr>
</tbody>
</table>

\[ P(T|R) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>-t</th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>(3/4)</td>
<td>(1/4)</td>
</tr>
<tr>
<td>-r</td>
<td>(1/2)</td>
<td>(1/2)</td>
</tr>
</tbody>
</table>

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

<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>

| 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   |

| 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  |
Example: Alarm Network

$P(\, +b, \, -e, \, +a, \, -j, \, +m) = \frac{0.95}{0.99}$
Example: Traffic

- Causal direction

\[ P(R) \]

| \(+r\) | \(1/4\) |
| \(-r\) | \(3/4\) |

\[ P(T | R) \]

| \(+r\) | \(+t\) | \(3/4\) |
| \(-r\) | \(+t\) | \(1/4\) |
| \(+t\) | \(-t\) |
| \(-r\) | \(+t\) | \(6/16\) |
| \(-r\) | \(-t\) | \(6/16\) |

\[ P(T, R) \]

| \(+r\) | \(+t\) | \(3/16\) |
| \(+r\) | \(-t\) | \(1/16\) |
| \(-r\) | \(+t\) | \(6/16\) |
| \(-r\) | \(-t\) | \(6/16\) |
**Example: Reverse Traffic**

- Reverse causality?

\[ P(T) \]

\[ \begin{array}{c|c}
+t & 9/16 \\
-t & 7/16 \\
\end{array} \]

\[ P(R|T) \]

\[ \begin{array}{c|ccc}
+t & +r & 1/3 \\
-r & 2/3 \\
\hline
-t & +r & 1/7 \\
-r & 6/7 \\
\end{array} \]

\[ P(T, R) \]

\[ \begin{array}{c|c}
+r & +t & 3/16 \\
+r & -t & 1/16 \\
-r & +t & 6/16 \\
-r & -t & 6/16 \\
\end{array} \]
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 * 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)
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' 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)