CSEP 573: Artificial Intelligence

Bayesian Networks

Luke Zettlemoyer

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

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

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

- 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!
Example Bayes’ Net: Insurance
Example: Independence

- **N** fair, independent coin flips:

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

<table>
<thead>
<tr>
<th></th>
<th>0.5</th>
<th></th>
<th>0.5</th>
<th></th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>0.5</td>
<td>h</td>
<td>0.5</td>
<td>\ldots</td>
<td>h</td>
</tr>
<tr>
<td>t</td>
<td>0.5</td>
<td>t</td>
<td>0.5</td>
<td></td>
<td>t</td>
</tr>
</tbody>
</table>

\[2^n \quad P(X_1, X_2, \ldots, X_n)\]
Example: Coin Flips

- $N$ independent coin flips

- No interactions between variables: absolute 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?

\[ P(T) \]

\begin{align*}
T & \quad P \\
\text{warm} & \quad 0.5 \\
\text{cold} & \quad 0.5
\end{align*}

\[ P(W) \]

\begin{align*}
W & \quad P \\
\text{sun} & \quad 0.6 \\
\text{rain} & \quad 0.4
\end{align*}

\[ P_1(T, W) \]

\begin{align*}
T & \quad W & \quad P \\
\text{warm} & \quad \text{sun} & \quad 0.4 \\
\text{warm} & \quad \text{rain} & \quad 0.1 \\
\text{cold} & \quad \text{sun} & \quad 0.2 \\
\text{cold} & \quad \text{rain} & \quad 0.3
\end{align*}

\[ P_2(T, W) \]

\begin{align*}
T & \quad W & \quad P \\
\text{warm} & \quad \text{sun} & \quad 0.3 \\
\text{warm} & \quad \text{rain} & \quad 0.2 \\
\text{cold} & \quad \text{sun} & \quad 0.3 \\
\text{cold} & \quad \text{rain} & \quad 0.2
\end{align*}
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:
  - $P(+\text{catch} | +\text{toothache}, +\text{cavity}) = P(+\text{catch} | +\text{cavity})$
- The same independence holds if I don’t have a cavity:
  - $P(+\text{catch} | +\text{toothache}, ¬\text{cavity}) = P(+\text{catch} | ¬\text{cavity})$
- Catch is *conditionally independent* of Toothache given Cavity:
  - $P(\text{Catch} | \text{Toothache}, \text{Cavity}) = P(\text{Catch} | \text{Cavity})$
- Equivalent statements:
  - $P(\text{Toothache} | \text{Catch }, \text{Cavity}) = P(\text{Toothache} | \text{Cavity})$
  - $P(\text{Toothache, Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity}) P(\text{Catch} | \text{Cavity})$
  - 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:

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

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

- What about fire, smoke, alarm?
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:
- \( P( +g ) = 0.5 \)
- \( P( +t \mid +g ) = 0.8 \)
- \( P( +t \mid -g ) = 0.4 \)
- \( P( +b \mid +g ) = 0.4 \)
- \( P( +b \mid -g ) = 0.8 \)

\[
\begin{align*}
P(T,B,G) &= P(G) \, P(T \mid G) \, P(B \mid G) \\
\text{P(T,B,G)} &= 0.16 \\
\text{P(T,B,G)} &= 0.16 \\
\text{P(T,B,G)} &= 0.24 \\
\text{P(T,B,G)} &= 0.04 \\
\text{P(T,B,G)} &= 0.04 \\
\text{P(T,B,G)} &= 0.24 \\
\text{P(T,B,G)} &= 0.06 \\
\text{P(T,B,G)} &= 0.06 
\end{align*}
\]
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: Alarm Network

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

### Burglary (B)
- **P(B)**: 0.001
- **¬b**: 0.999

### Earthquake (E)
- **P(E)**: 0.998
- **+e**: 0.002

### John calls (J)
- **P(J|A)**:
  - +a: 0.9
  - +a: 0.1
  - +a: 0.05
  - +a: 0.95

### Mary calls (M)
- **P(M|A)**:
  - +a: 0.7
  - +a: 0.3
  - +a: 0.01
  - +a: 0.99

### Alarm (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
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
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!

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

$X_1 \perp X_2$

All distributions
Example: Coins

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

- Adding unneeded arcs isn’t wrong, it’s just inefficient
Topology Limits Distributions

- Given some graph topology $G$, only certain joint distributions can be encoded.
- The graph structure guarantees certain (conditional) independences.
- (There might be more independence)
- Adding arcs increases the set of distributions, but has several costs.
- Full conditioning can encode any distribution.
Independence in a BN

- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - If yes, can prove using algebra (tedious in general)
  - If no, can prove with a counter example
  - Example:

```
  X -> Y -> Z
```

- Question: are X and Z necessarily independent?
  - Answer: no. Example: low pressure causes rain, which causes traffic.
  - X can influence Z, Z can influence X (via Y)
  - Addendum: they *could* be independent: how?
This configuration is a “causal chain”

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

- Is X independent of Z given Y?

\[ P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} = P(z|y) \]

Yes!

- Evidence along the chain “blocks” the influence
Another basic configuration: two effects of the same parent
- Are X and Z independent?
- Are X and Z independent given Y?

\[ P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)} = P(z|y) \]

Yes!

Observing the cause blocks influence between effects.
Common Effect

- Last configuration: two causes of one effect (v-structures)
  - Are X and Z independent?
    - Yes: the ballgame and the rain cause traffic, but they are not correlated
    - Still need to prove they must be (try it!)
  - Are X and Z independent given Y?
    - No: seeing traffic puts the rain and the ballgame in competition as explanation?
  - This is backwards from the other cases
    - Observing an effect activates influence between possible causes.

X: Raining
Z: Ballgame
Y: Traffic
Any complex example can be analyzed using these three canonical cases

General question: in a given BN, are two variables independent (given evidence)?

Solution: analyze the graph
Recipe: shade evidence nodes

Attempt 1: if two nodes are connected by an undirected path not blocked by a shaded node, they are conditionally independent

Almost works, but not quite
- Where does it break?
- Answer: the v-structure at T doesn’t count as a link in a path unless “active”
Reachability (D-Separation)

- **Question:** Are X and Y conditionally independent given evidence vars \( \{Z\} \)?
  - Yes, if X and Y “separated” by Z
  - Look for active paths from X to Y
  - No active paths = independence!

- **A path is active if each triple is active:**
  - Causal chain \( A \rightarrow B \rightarrow C \) where B is unobserved (either direction)
  - Common cause \( A \leftarrow B \rightarrow C \) where B is unobserved
  - Common effect (aka v-structure) \( A \rightarrow B \leftarrow C \) where B or one of its descendents is observed

- All it takes to block a path is a single inactive segment
Example: Independent?

\[ R \perp B \]
\[ R \perp B | T \]
\[ R \perp B | T' \]

Yes
Example: Independent?

\[ L \perp T' | T \quad \text{Yes} \]
\[ L \perp B \quad \text{Yes} \]
\[ L \perp B | T \]
\[ L \perp B | T' \]
\[ L \perp B | T, R \quad \text{Yes} \]
Example

- **Variables:**
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I’m sad

- **Questions:**

\[
\begin{align*}
T \perp D \\
T \perp D | R \\
T \perp D | R, S
\end{align*}
\]

Yes
Summary

- Bayes nets compactly encode joint distributions

- Guaranteed independencies of distributions can be deduced from BN graph structure

- D-separation gives precise conditional independence guarantees from graph alone

- A Bayes’ net’s joint distribution may have further (conditional) independence that is not detectable until you inspect its specific distribution