CSE 473: Artificial Intelligence

Bayes’ Nets

Luke Zettlemoyer --- University of Washington

[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley. All CS188 materials are available at http://ai.berkeley.edu.]
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: \( X \perp \!\!\!\!\!\!\!\!\!\perp Y \)

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

$P_1(T, W)$

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

$P(T)$

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>0.3</td>
<td></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></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:

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

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

\[ X \perp Y \mid Z \]

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

\[
\begin{align*}
P(T,B,G) &= P(G) P(T | G) P(B | G) \\
\begin{array}{cccc}
T & B & G & P(T,B,G) \\
+ & t & + & b & + & g & 0.16 \\
+ & t & + & b & - & g & 0.16 \\
+ & t & - & b & + & g & 0.24 \\
+ & t & - & b & - & g & 0.04 \\
- & t & + & b & + & g & 0.04 \\
- & t & + & b & - & g & 0.24 \\
- & t & - & b & + & g & 0.06 \\
- & t & - & b & - & g & 0.06 \\
\end{array}
\end{align*}
\]
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

- battery age
- alternator broken
- fanbelt broken
- battery dead
- no charging
- battery flat
- no oil
- no gas
- fuel line blocked
- starter broken
- lights
- oil light
- gas gauge
- car won’t start
- dipstick
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?
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
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:

\[
P(\text{+cavity, +catch, -toothache})
\]
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
Example: Coin Flips

Only distributions whose variables are absolutely independent can be represented by a Bayes’ net with no arcs.
Example: Traffic

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

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

\[P(+r, -t) = \]
Example: Traffic

- Causal direction

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

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

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

- Reverse causality?

\[ P(T) \]

<table>
<thead>
<tr>
<th></th>
<th>+t</th>
<th>-t</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>9/16</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>7/16</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(R|T) \]

<table>
<thead>
<tr>
<th></th>
<th>+r</th>
<th>-r</th>
</tr>
</thead>
<tbody>
<tr>
<td>+t</td>
<td>1/3</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>2/3</td>
<td></td>
</tr>
<tr>
<td>-t</td>
<td>1/7</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>6/7</td>
<td></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>3/16</td>
<td></td>
</tr>
<tr>
<td>+r</td>
<td>1/16</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>6/16</td>
<td></td>
</tr>
<tr>
<td>-r</td>
<td>6/16</td>
<td></td>
</tr>
</tbody>
</table>
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
    \[ P(x_i|x_1, \ldots x_{i-1}) = P(x_i|\text{parents}(X_i)) \]
Example: Alarm Network

\[
P(\text{+b}, \text{-e}, \text{+a}, \text{-j}, \text{+m}) =
\]

\begin{tabular}{|c|c|}
\hline
B & P(B) \\
\hline
+\text{b} & 0.001 \\
-\text{b} & 0.999 \\
\hline
\end{tabular}

\begin{tabular}{|c|c|}
\hline
E & P(E) \\
\hline
+\text{e} & 0.002 \\
-\text{e} & 0.998 \\
\hline
\end{tabular}

\begin{tabular}{|c|c|c|}
\hline
A & J & P(\text{J}|A) \\
\hline
+\text{a} & +\text{j} & 0.9 \\
+\text{a} & -\text{j} & 0.1 \\
-\text{a} & +\text{j} & 0.05 \\
-\text{a} & -\text{j} & 0.95 \\
\hline
\end{tabular}

\begin{tabular}{|c|c|c|}
\hline
A & M & P(\text{M}|A) \\
\hline
+\text{a} & +\text{m} & 0.7 \\
+\text{a} & -\text{m} & 0.3 \\
-\text{a} & +\text{m} & 0.01 \\
-\text{a} & -\text{m} & 0.99 \\
\hline
\end{tabular}

\begin{tabular}{|c|c|c|c|}
\hline
B & E & A & P(\text{A}|\text{B},\text{E}) \\
\hline
+\text{b} & +\text{e} & +\text{a} & 0.95 \\
+\text{b} & +\text{e} & -\text{a} & 0.05 \\
+\text{b} & -\text{e} & +\text{a} & 0.94 \\
+\text{b} & -\text{e} & -\text{a} & 0.06 \\
-\text{b} & +\text{e} & +\text{a} & 0.29 \\
-\text{b} & +\text{e} & -\text{a} & 0.71 \\
-\text{b} & -\text{e} & +\text{a} & 0.001 \\
-\text{b} & -\text{e} & -\text{a} & 0.999 \\
\hline
\end{tabular}
Example: Alarm Network

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

- Representation
  - Conditional Independences
  - Probabilistic Inference
  - Learning Bayes’ Nets from Data
Conditional Independence

- X and Y are independent if
  \[ \forall x, y \ P(x, y) = P(x)P(y) \rightarrow X \perp Y \]

- X and Y are conditionally independent given Z
  \[ \forall x, y, z \ P(x, y|z) = P(x|z)P(y|z) \rightarrow X \perp Y|Z \]

- (Conditional) independence is a property of a distribution

- Example: \( \text{Alarm} \perp \text{Fire}|\text{Smoke} \)
Bayes Nets: Assumptions

- Assumptions we are required to make to define the Bayes net when given the graph:

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

- Beyond above “chain rule $\rightarrow$ Bayes net” conditional independence assumptions
  - Often additional conditional independences
  - They can be read off the graph

- Important for modeling: understand assumptions made when choosing a Bayes net graph
Example

- Conditional independence assumptions directly from simplifications in chain rule:

- Additional implied conditional independence assumptions?
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:

![Diagram of nodes X, Y, Z with arrows]

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?
D-separation: Outline
D-separation: Outline

- Study independence properties for triples
- Analyze complex cases in terms of member triples
- D-separation: a condition / algorithm for answering such queries
Causal Chains

- This configuration is a “causal chain”

Guaranteed $X$ independent of $Z$? **No!**

- One example set of CPTs for which $X$ is not independent of $Z$ is sufficient to show this independence is not guaranteed.

- Example:
  - Low pressure causes rain causes traffic, high pressure causes no rain causes no traffic

- In numbers:
  \[
P(+y \mid +x) = 1, \quad P(-y \mid -x) = 1, \quad P(+z \mid +y) = 1, \quad P(-z \mid -y) = 1
\]
Causal Chains

- This configuration is a “causal chain”

- Guaranteed X independent of Z given Y?

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

- Evidence along the chain “blocks” the influence

\[ P(x, y, z) = P(x)P(y|x)P(z|y) \]
This configuration is a “common cause”

Guaranteed $X$ independent of $Z$?  **No!**

- One example set of CPTs for which $X$ is not independent of $Z$ is sufficient to show this independence is not guaranteed.

Example:
- Project due causes both forums busy and lab full

In numbers:

$$P(+x | +y) = 1, P(-x | -y) = 1,$$
$$P(+z | +y) = 1, P(-z | -y) = 1$$
Common Cause

- This configuration is a “common cause”

- Guaranteed 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.
Last configuration: two causes of one effect (v-structures)

Are X and Y 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 Y independent given Z?
- 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.
The General Case
General question: in a given BN, are two variables independent (given evidence)?

Solution: analyze the graph

Any complex example can be broken into repetitions of the three canonical cases
Reachability

- Recipe: shade evidence nodes, look for paths in the resulting graph

- 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”
Question: Are X and Y conditionally independent given evidence variables \( \{Z\} \)?
- Yes, if X and Y “d-separated” by Z
- Consider all (undirected) 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
D-Separation

- **Query:** $X_i \nparallel X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$

- Check all (undirected!) paths between $X_i$ and $X_j$
  - If one or more active, then independence not guaranteed
    $X_i \nparallel X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$
  - Otherwise (i.e. if all paths are inactive), then independence is guaranteed
    $X_i \nparallel X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$
Example

\[ R \perp B \]
\[ R \perp B \mid T \]
\[ R \perp B \mid T' \]
Example

$L \perp T'|T$  \hspace{1cm} Yes
$L \perp B$  \hspace{1cm} Yes
$L \perp B|T$  
$L \perp B|T'$
$L \perp B|T, R$  \hspace{1cm} Yes
**Example**

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

- **Questions:**
  
  \[
  T \perp D \\
  T \perp D | R \\
  T \perp D | R, S
  \]

  Yes
Structure Implications

- Given a Bayes net structure, can run d-separation algorithm to build a complete list of conditional independences that are necessarily true of the form

  \[ X_i \perp\!
\perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\} \]

- This list determines the set of probability distributions that can be represented
Computing All Independences

Compute \textbf{ALL THE INDEPENDENCES}!
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.
Bayes Nets Representation 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
Bayes’ Nets

- Representation
- Conditional Independences
  - Probabilistic Inference
    - Enumeration (exact, exponential complexity)
    - Variable elimination (exact, worst-case exponential complexity, often better)
    - Probabilistic inference is NP-complete
    - Sampling (approximate)
- Learning Bayes’ Nets from Data