A Bayes’ net is an efficient encoding of a probabilistic model of a domain.

Questions we can ask:
- Inference: given a fixed BN, what is P(X | e)?
- Representation: given a fixed BN, what kinds of distributions can it encode?
- Modeling: what BN is most appropriate for a given domain?

Example Bayes’ Net

Building the (Entire) Joint
- We can take a Bayes’ net and build any entry from the full joint distribution it encodes
  \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \text{parents}(X_i)) \]
- Typically, there’s no reason to build ALL of it
- We build what we need on the fly
- To emphasize: every BN over a domain implicitly represents some joint distribution over that domain, but is specified by local probabilities

Example: Alarm Network

\[
P(b, e, \neg a, j, m) = \]
Size of a Bayes’ Net

- How big is a joint distribution over N Boolean variables?
- How big is an N-node net if nodes have k parents?

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 turns out to be faster to answer queries (next class)

Bayes’ Nets

- So far: how a Bayes’ net encodes a joint distribution
- Next: how to answer queries about that distribution
  - Key idea: conditional independence
  - Last class: assembled BNs using an intuitive notion of conditional independence as causality
  - Today: formalize these ideas
  - Main goal: answer queries about conditional independence and influence
- After that: how to answer numerical queries (inference)

Conditional Independence

- Reminder: independence
  - \( X \) and \( Y \) are independent if
    \[
    \forall x, y \quad P(x, y) = P(x)P(y) 
    \Rightarrow \quad X \perp Y
    \]
- \( X \) and \( Y \) are conditionally independent given \( Z \)
  \[
  \forall x, y, z \quad P(x, y|z) = P(x|z)P(y|z) \quad \Rightarrow \quad X \perp Y|Z
  \]
- (Conditional) independence is a property of a distribution

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!

\[
\begin{array}{c|c|c}
X_1 & X_2 & P(X_1) \\
\hline
h & 0.5 & 0.3 \\
1 & 0.5 & 0.7 \\
\end{array}
\]

Independence in a BN

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

- Question: are \( X \) and \( Z \) independent?
  - Answer: not necessarily, we’ve seen examples otherwise: low pressure causes rain which causes traffic.
  - \( X \) can influence \( Z \), \( Z \) can influence \( X \) (via \( Y \))
  - Addendum: they could be independent: how?

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

Independence in a BN

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

- Question: are \( X \) and \( Z \) independent?
  - Answer: not necessarily, we’ve seen examples otherwise: low pressure causes rain which causes traffic.
  - \( X \) can influence \( Z \), \( Z \) can influence \( X \) (via \( Y \))
  - Addendum: they could be independent: how?
Causal Chains

- 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?

\[ P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} = P(z|y) \quad \text{Yes!} \]

- Evidence along the chain "blocks" the influence

Common Cause

- Another basic configuration: two effects of the same cause
  - Are X and Z independent?

\[ P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)} = P(z|y) \quad \text{Yes!} \]

- Observing the cause blocks influence between effects.

Common Effect

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

- Are X and Z independent?
  - Yes: remember the ballgame and the rain causing traffic, no correlation?
  - Still need to prove they must be (homework)

- Are X and Z independent given Y?
  - No: remember that seeing traffic put the rain and the ballgame in competition?
- This is backwards from the other cases
  - Observing the effect enables influence between effects.

The General Case

- Any complex example can be analyzed using these three canonical cases

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

- Solution: graph search!

Reachability

- 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 shaded

Reachability (the Bayes’ Ball)

- Correct algorithm:
  - Shade in evidence
  - Start at source node
  - Try to reach target by search

- States: pair of (node X, previous state S)

- Successor function:
  - X unobserved:
    - To any child
    - To any parent if coming from a child
  - X observed:
    - From parent to parent

- If you can’t reach a node, it’s conditionally independent of the start node given evidence
**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

- 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
  - 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 only guaranteed to encode conditional independencies

**Example: Traffic**

- Basic traffic net
- Let’s multiply out the joint

<table>
<thead>
<tr>
<th>R</th>
<th>P(R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>3/4</td>
</tr>
<tr>
<td>¬r</td>
<td>1/4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T</th>
<th>P(T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>1/2</td>
</tr>
<tr>
<td>¬r</td>
<td>1/2</td>
</tr>
</tbody>
</table>

| P(T|R) |
|-------|
| r     | 3/16 |
| ¬r    | 1/16 |

| P(R|T) |
|------|
| r    | 1/4  |
| ¬r   | 3/4  |

**Example: Reverse Traffic**

- Reverse causality?

<table>
<thead>
<tr>
<th>T</th>
<th>P(T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>1/2</td>
</tr>
<tr>
<td>¬r</td>
<td>1/2</td>
</tr>
</tbody>
</table>

| P(T|R) |
|-------|
| r    | 3/16 |
| ¬r   | 1/16 |

| P(R|T) |
|------|
| r    | 1/2  |
| ¬r   | 2/3  |

<table>
<thead>
<tr>
<th>P(R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
</tr>
<tr>
<td>¬r</td>
</tr>
</tbody>
</table>
Example: Coins

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

| X_1 \| X_2 \| P(X_1) | P(X_2) | P(X_1, X_2) |
|-------|-------|----------|----------|-------------|
| h     | 0.5   | h 0.5    | t 0.5    | t 0.5       |
| t     | 0.5   | h 0.5    | t 0.5    | t 0.5       |

Alternate BNs

Summary

- Bayes nets compactly encode joint distributions
- Guaranteed independencies of distributions can be deduced from BN graph structure
- A Bayes’ net may have other independencies that are not detectable until you inspect its specific distribution
- The Bayes’ ball algorithm (aka d-separation) tells us when an observation of one variable can change belief about another variable

Inference

- Inference: calculating some statistic from a joint probability distribution
- Examples:
  - Posterior probability:
    \[ P(Q|E_1 = e_1 \ldots E_k = e_k) \]
  - Most likely explanation:
    \[ \arg\max_q P(Q = q|E_1 = e_1 \ldots) \]

Inference by Enumeration

- P(sun)?

<table>
<thead>
<tr>
<th>S</th>
<th>T</th>
<th>R</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>summer</td>
<td>warm</td>
<td>sun</td>
<td>0.30</td>
</tr>
<tr>
<td>summer</td>
<td>warm</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>summer</td>
<td>cold</td>
<td>sun</td>
<td>0.10</td>
</tr>
<tr>
<td>winter</td>
<td>warm</td>
<td>sun</td>
<td>0.10</td>
</tr>
<tr>
<td>winter</td>
<td>warm</td>
<td>rain</td>
<td>0.05</td>
</tr>
<tr>
<td>winter</td>
<td>cold</td>
<td>sun</td>
<td>0.15</td>
</tr>
<tr>
<td>winter</td>
<td>cold</td>
<td>rain</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Reminder: Alarm Network
Inference by Enumeration

- Given unlimited time, inference in BNs is easy
- Recipe:
  - State the marginal probabilities you need
  - Figure out ALL the atomic probabilities you need
  - Calculate and combine them
- Example:
  \[ P(b|j, m) = \frac{P(b, j, m)}{P(j, m)} \]

Example

- In this simple method, we only need the BN to synthesize the joint entries

\[
P(b, j, m) = \frac{P(b)P(e)P(a|b, e)P(j|a)P(m|a)}{P(b)P(e)P(\bar{a}|b, e)P(j|\bar{a})P(m|\bar{a})} + \frac{P(b)P(\bar{e})P(a|b, \bar{e})P(j|\bar{a})P(m|\bar{a})}{P(b)P(\bar{e})P(\bar{a}|b, \bar{e})P(j|\bar{a})P(m|\bar{a})}
\]

Normalization Trick

\[
P(B|j, m) = \frac{P(B, j, m)}{P(j, m)}
\]

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

\[
P(b, j, m) = \frac{P(b)P(e)P(a|b, e)P(j|a)P(m|a)}{P(b)P(e)P(\bar{a}|b, e)P(j|\bar{a})P(m|\bar{a})} + \frac{P(b)P(\bar{e})P(a|b, \bar{e})P(j|\bar{a})P(m|\bar{a})}{P(b)P(\bar{e})P(\bar{a}|b, \bar{e})P(j|\bar{a})P(m|\bar{a})}
\]

Inference by Enumeration

- General case:
  - Evidence variables: \((E_1, \ldots, E_n) = (e_1, \ldots, e_n)\)
  - Query variables: \(Y_1, \ldots, Y_n\)
  - Hidden variables: \(H_1, \ldots, H_n\)
- We want: \(P(Y_1 \ldots Y_n|e_1 \ldots e_n)\)
- First, select the entries consistent with the evidence
- Second, sum out \(H\):
  \[
P(Y_1 \ldots Y_n, e_1 \ldots e_n) = \sum_{h_1, \ldots, h_n} P(Y_1 \ldots Y_n, h_1 \ldots h_n, e_1 \ldots e_n)
  \]
  \[
  \frac{1}{X_1, X_2, \ldots, X_n}
  \]
- Finally, normalize the remaining entries to conditionalize
- Obvious problems:
  - Worst-case time complexity \(O(d^n)\)
  - Space complexity \(O(d^n)\) to store the joint distribution

Inference by Enumeration?
Nesting Sums

- Atomic inference is extremely slow!
- Slightly clever way to save work:
  - Move the sums as far right as possible
  - Example:
    \[
P(b, j, m) = \sum_{e, a} P(b, e, a, j, m) \\
    = \sum_{e, a} P(b) P(e) P(a|b, e) P(j|a) P(m|a) \\
    = P(b) \sum_r P(e) \sum_a P(a|b, e) P(j|a) P(m|a)
    \]

Variable Elimination: Idea

- Lots of redundant work in the computation tree
- We can save time if we cache all partial results
- This is the basic idea behind variable elimination

Sampling

- Basic idea:
  - Draw N samples from a sampling distribution S
  - Compute an approximate posterior probability
  - Show this converges to the true probability \( P \)
- Outline:
  - Sampling from an empty network
  - Rejection sampling: reject samples disagreeing with evidence
  - Likelihood weighting: use evidence to weight samples

Prior Sampling

- This process generates samples with probability
  \[
  S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(x_i)) = P(x_1 \ldots x_n),
  \]
  ...i.e. the BN’s joint probability
- Let the number of samples of an event be \( N_{PS}(x_1 \ldots x_n) \)
- Then \( \lim_{N \to \infty} P(x_1 \ldots x_n) = \lim_{N \to \infty} N_{PS}(x_1 \ldots x_n)/N = S_{PS}(x_1 \ldots x_n) = P(x_1 \ldots x_n) \)
- I.e., the sampling procedure is consistent

Example

- We’ll get a bunch of samples from the BN:
  - \( c, \neg s, r, w \)
  - \( c, s, r, \neg w \)
  - \( \neg c, s, \neg r, w \)
  - \( \neg c, s, \neg r, \neg w \)
- If we want to know \( P(W) \)
  - We have counts \( <w:4, \neg w:1> \)
  - Normalize to get \( P(W) = <w:0.8, \neg w:0.2> \)
  - This will get closer to the true distribution with more samples
  - Can estimate anything else, too
  - What about \( P(C| \neg r)? \)  \( P(C| \neg r, \neg w)? \)
Rejection Sampling

- Let’s say we want P(C)
  - No point keeping all samples around
  - Just tally counts of C outcomes
- Let’s say we want P(C|s)
  - Same thing: tally C outcomes, but ignore (reject) samples which don’t have S=s
  - This is rejection sampling
  - It is also consistent (correct in the limit)

Likelihood Weighting

- Problem with rejection sampling:
  - If evidence is unlikely, you reject a lot of samples
  - You don’t exploit your evidence as you sample
  - Consider P(B|a)

- Idea: fix evidence variables and sample the rest

- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents

Likelihood Sampling

- Sampling distribution if z sampled and e fixed evidence
  \[ S_{W}^{Y}(z, e) = \prod_{i=1}^{m} P(c_{i}|\text{Parents}(Z_{i})) \]
  - Now, samples have weights
  \[ w(z, e) = \prod_{i=1}^{m} P(c_{i}|\text{Parents}(E_{i})) \]
  - Together, weighted sampling distribution is consistent
  \[ S_{W}^{Y}(z, e)w(z, e) = \prod_{i=1}^{m} P(c_{i}|\text{Parents}(E_{i})) \prod_{i=1}^{m} P(c_{i}|\text{Parents}(E_{i})) \]
  \[ = P(z, e) \]

Likelihood Weighting

- Note that likelihood weighting doesn’t solve all our problems
- Rare evidence is taken into account for downstream variables, but not upstream ones
- A better solution is Markov-chain Monte Carlo (MCMC), more advanced
- We’ll return to sampling for robot localization and tracking in dynamic BNs