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)
  - Aka graphical model
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions

Bayes' Net Semantics

Formally:

- A set of nodes, one per random variable
- Directed edges, forming an acyclic graph
- A CPT for each node
  - CPT = “Conditional Probability Table”
  - Collection of distributions over X, one for each combination of parents’ values
  \[ P(X|a_1 \ldots a_n) \]

A Bayes Net = Topology (graph) + Local Conditional Probabilities

Hidden Markov Models

- An HMM is defined by:
  - Initial distribution: \[ P(X_1) \]
  - Transitions: \[ P(X_i|X_{i-1}) \]
  - Emissions: \[ P(E|X) \]
**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:
    \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]
  - Does this always work? Why?
  - 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: Independent Coin Flips**

- N independent coin flips
  - \( X_1, X_2, \ldots, X_n \)
  - \[ P(X_1) = \begin{cases} h & 0.5 \\ t & 0.5 \end{cases} \]
  - \[ P(X_2) = \begin{cases} h & 0.5 \\ t & 0.5 \end{cases} \]
  - \[ P(X_n) = \begin{cases} h & 0.5 \\ t & 0.5 \end{cases} \]
  - \[ P(X_1, X_2, \ldots, X_n) = 2^{n-1} \]
  - No interactions between variables

**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 fire, smoke, alarm?

**Example: Alarm Network**

- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!
- How big is joint distribution?
  - \[ 2^n - 1 = 31 \] parameters

**Example: Alarm Network**

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+b</td>
<td>0.001</td>
<td>+e</td>
<td>0.002</td>
</tr>
<tr>
<td>¬b</td>
<td>0.999</td>
<td>¬e</td>
<td>0.999</td>
</tr>
</tbody>
</table>

| 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  |

| A   | M   | P(A|B,E) |
|-----|-----|---------|
| +a  | +e  | +α    | 0.95   |
| +a  | +e  | ¬α    | 0.05   |
| +b  | +e  | +α    | 0.94   |
| +b  | +e  | ¬α    | 0.06   |
| +b  | ¬e  | +α    | 0.29   |
| +b  | ¬e  | ¬α    | 0.71   |
| ¬b  | +e  | +α    | 0.001  |
| ¬b  | +e  | ¬α    | 0.999  |
| ¬b  | ¬e  | +α    | 0.001  |
| ¬b  | ¬e  | ¬α    | 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

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 \( (\text{the CPTs}) \) all you want, but you won’t be able to represent any distribution in which the flips are dependent!

  \[
  \begin{array}{c|c}
  \text{X}_1 & \text{X}_2 \\
  \text{P} & 0.5 & 0.5 \\
  \text{h} & 0.5 & h \\
  \text{t} & 0.5 & t \\
  \end{array}
  \]

Example: Coins

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

  \[
  \begin{array}{c|c}
  \text{X}_1 & \text{X}_2 \\
  \text{P} & 0.5 & 0.5 \\
  \text{h} & 0.5 & h \\
  \text{t} & 0.5 & t \\
  \end{array}
  \]

  \[
  \begin{array}{c|c}
  \text{X}_1 & \text{X}_2 | \text{X}_1 \\
  \text{h} & h | 0.5 & t | 0.5 \\
  \text{t} & t | 0.5 & h | 0.5 \\
  \end{array}
  \]

  - 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:

  \[
  \text{X} \rightarrow \text{Y} \rightarrow \text{Z}
  \]

  Question: are \( X \) and \( Z \) independent?
  - Answer: no.
    - Example: low pressure causes rain, which causes traffic.
    - Knowledge about \( X \) may change belief in \( Z \)
    - Knowledge about \( Z \) may change belief in \( X \) (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?

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

Yes!

Evidence along the chain “blocks” the influence

Common Parent

- Another basic configuration: two effects of the same parent

X: Low pressure
Y: Rain
Z: Traffic

- Are X and Z independent?

Yes!

Evidence along the chain “blocks” the influence

Common Parent

- Another basic configuration: two effects of the same parent

Y: Project due
X: Forum busy
Z: Lab full

- Are X and Z independent?

Yes!

Evidence along the chain “blocks” the influence

Common Effect

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

Y: Traffic
X: Raining
Z: Ballgame

- 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 the cause blocks 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: analyze the graph
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?

- Variables:
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I'm sad
- Questions:
  - \( R \perp T \) Yes
  - \( R \perp B \) Yes
  - \( R \perp B | T \) No
  - \( R \perp B | T' \) No
  - \( L \perp B | T \) Yes
  - \( L \perp B | T' \) No
  - \( L \perp B | T, R \) Yes

Given Markov Blanket, X is Independent of All Other Nodes

\[ MB(X) = \text{Par}(X) \cup \text{Childs}(X) \cup \text{Par}(\text{Childs}(X)) \]
Summary

- Bayes nets compactly encode joint distributions (JDS)
- Other graphical models too: factor graphs, CRFs, …
- 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 JD may have further (conditional) independence known only from specific CPTs

Outline

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

Inference in BNs

This graphical independence representation yields efficient inference schemes

- We generally want to compute
  - Marginal probability: \( P(Z) \)
  - \( P(Z|E) \) where \( E \) is (conjunctive) evidence
    - \( Z \): query variable(s),
    - \( E \): evidence variable(s)
    - everything else: hidden variable
  - Computations organized by network topology

\[
P(B | J=true, M=true) = \alpha \sum_{e,a} P(b, j, m, e, a)
\]

\[
P(b|m) = \alpha \sum_e P(b, e) \sum_a P(j|a) P(m|a)
\]

Variable Elimination

\[
P(b,j,m) = \alpha \sum_e P(b, e) \sum_a P(a, e) P(j|a) P(m|a)
\]

Repeated computations \(\rightarrow\) Dynamic Programming
Reducing 3-SAT to Bayes Nets

- **Theorem**: Inference in a multi-connected Bayesian network is NP-hard.

Boolean 3CNF formula $\phi = (w \lor \neg w \lor w) \land (u \lor \neg u \lor y)$

Bayes Net is a generative model
- We can easily generate samples from the distribution represented by the Bayes net
  - Generate one variable at a time in topological order

Approximate Inference in Bayes Nets
Sampling based methods

(Based on slides by Jack Breese and Daphne Koller)

Bayes Net is a generative model

- We can easily generate samples from the distribution represented by the Bayes net
  - Generate one variable at a time in topological order

Stochastic simulation $P(B|C)$

Use the samples to compute probabilities, say $P(c)$ or $P(nc)$
### Rejection Sampling

- Sample from the prior
- Reject if do not match the evidence
- Returns consistent posterior estimates
- Hopelessly expensive if \( P(e) \) is small
- \( P(e) \) drops exponentially with num of evidence vars

### Likelihood Weighting

#### Idea:
- Fix evidence variables
- Sample only non-evidence variables
- Weight each sample by the likelihood of evidence
Likelihood weighting $P(B|C)$

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burglary</td>
<td>0.8</td>
</tr>
<tr>
<td>Earthquake</td>
<td>0.05</td>
</tr>
<tr>
<td>Alarm</td>
<td>0.2</td>
</tr>
<tr>
<td>Cold</td>
<td>0.7</td>
</tr>
<tr>
<td>NewsCast</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Samples:

- $B, E, A, C, N$
- $b, e, a, c, n$

- $P(b) = \frac{\text{weight of samples with } B=b}{\text{total weight of samples}}$

Likelihood Weighting

- Sampling probability: $S(z,e) = \prod_i P(z_i | \text{Parents}(Z_i))$  
  - Neither prior nor posterior
- Weight for a sample $<z,e>$: $w(z,e) = \prod_i P(x_i | \text{Parents}(E_i))$
- Weighted Sampling probability $S(z,e)w(z,e) = \prod_i P(z_i | \text{Parents}(Z_i)) \prod_i P(x_i | \text{Parents}(E_i))$
- $= P(z,e)$  
  - returns consistent estimates
  - performance degrades w/ many evidence vars
  - a few samples get majority of the weight
  - late occurring evidence vars don’t guide sample generation

MCMC with Gibbs Sampling

- Fix the values of observed variables
- Set the values of all non-observed variables randomly
- Perform a random walk through the space of complete variable assignments. On each move:
  1. Pick a variable $X$
  2. Calculate $P(X=\text{true} \ | \ \text{all other variables})$
  3. Set $X$ to true with that probability
- Repeat many times. Frequency with which any variable $Y$ is true = its posterior probability.
- Converges to true posterior when frequencies stop changing significantly
  - stable distribution, mixing

Given Markov Blanket, $X$ is Independent of All Other Nodes

$MB(X) = \text{Par}(X) \cup \text{Childs}(X) \cup \text{Par(Childs}(X))$

Markov Blanket Sampling

- How to calculate $P(X=\text{true} \ | \ \text{all other variables})$?
  - Recall: a variable is independent of all others given it’s Markov Blanket
    - parents
    - children
    - other parents of children
  - So problem becomes calculating $P(X=\text{true} \ | MB(X))$
    - Fortunately, it is easy to solve exactly
      $P(X) = \alpha P(X \ | Parents(X)) \prod_{\text{ForChilds}(X)} P(Y \ | Parents(Y))$

Example

$P(X) = \alpha P(X \ | Parents(X)) \prod_{\text{ForChilds}(X)} P(Y \ | Parents(Y))$

$P(X \ | A, B, C) = \frac{P(X, A, B, C)}{P(A, B, C)}$

$= \frac{P(A)P(X \ | A)P(C)P(B \ | X, C)}{P(A, B, C)}$

$= \frac{P(A)P(C)}{P(A, B, C)} P(X \ | A) P(B \ | X, C)$

$= \alpha P(X \ | A) P(B \ | X, C)$
Example

- Evidence: s, b
- Randomly set: h, g
- Sample H using P(H|s,g,b)
- Suppose result is ~h

Sample G using P(G|s,~h,b)
Example

- Evidence: s, b
- Randomly set: ~h, g
- Sample H using P(H|s,g,b)
  Suppose result is ~h
  Sample G using P(G|s,~h,b)
  =>Suppose result is g
  Sample G using P(G|s,~h,b)

Gibbs MCMC Summary

\[
Pr(X|E) = \frac{\text{number of samples with } X=x}{\text{total number of samples}}
\]

- Advantages:
  - No samples are discarded
  - No problem with samples of low weight
  - Can be implemented very efficiently
  - 10K samples @ second

- Disadvantages:
  - Can get stuck if relationship between vars is deterministic
  - Many variations devised to make MCMC more robust

Other inference methods
- Exact inference
  - Junction tree

- Approximate inference
  - Belief Propagation
  - Variational Methods

Outline
- Probabilistic models
  - Bayesian Networks (BNs)
  - Independence in BNs
  - Efficient Inference in BNs
  - Learning