Probabilistic Models - Outline

- Bayesian Networks (BNs)
- Independence
- Efficient Inference in BNs
  - Variable Elimination
  - Direct Sampling
  - Markov Chain Monte Carlo (MCMC)
- Learning

Bayes’ Nets: Big Picture

- Problems with using full joint distribution:
  - Unless very few variables, the joint is WAY too big
  - Unless very few variables, hard to learn (estimate empirically)

- Bayesian networks: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - A kind of “graphical model”
  - We describe how random variables interact, locally
  - Local interactions chain together to give global distribution

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!
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 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?

Reachability (D-Separation)

- Question: Are X and Y conditionally independent given evidence vars [Z]?
  - Yes, if X and Y "separated" by Z
  - No active paths = independence
  - A path is active if each triple is active:
    - Causal chain A → B → C where B is unobserved (either direction)
    - Common cause A ← B → C where B is unobserved
    - Common effect (aka v-structure) A → B ← C where B or one of its descendents is observed
    - All it takes to block a path is a single inactive segment

Example

- Variables:
  - R: Raining
  - W: Wet
  - P: Plants growing
  - T: Traffic bad
  - D: Roof drips
  - S: I'm sad
- Questions:
  - W ⊥ D

Example

- Variables:
  - R: Raining
  - W: Wet
  - P: Plants growing
  - T: Traffic bad
  - D: Roof drips
  - S: I'm sad
- Questions:
  - W ⊥ D
  - P ⊥ D | R, S

Example

- Variables:
  - R: Raining
  - W: Wet
  - P: Plants growing
  - T: Traffic bad
  - D: Roof drips
  - S: I'm sad
- Questions:
  - W ⊥ D
  - P ⊥ D | R, S

Example

Given Markov Blanket, X is Independent of All Other Nodes

\[ MB(X) = \text{Par}(X) \cup \text{Childs}(X) \cup \text{Par(Childs)(X)} \]
Given Markov Blanket, X is Independent of All Other Nodes

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

Inference in BNs

- The graphical independence representation
- yields efficient inference schemes
- We generally want to compute
  - Marginal probability: \( \Pr(Z) \),
  - \( \Pr(Z|E) \) 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)

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

Variable Elimination

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

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

Use the samples to compute marginal probabilities, say $\Pr(c)$
Stochastic simulation $P(B|C)$

Samples:

$B E A C N$

$b e a c$

$P(b) = 0.03$

$P(B|C) = 0.04$

$P(a) = 0.08$

$P(c) = 0.05$

$P(e) = 0.08$

$P(n) = 0.30$

$P(e) = 0.06$

$P(n) = 0.30$

$P(e) = 0.06$

$P(n) = 0.30$

$P(e) = 0.06$

$P(n) = 0.30$

$P(e) = 0.06$

$P(n) = 0.30$

$P(e) = 0.06$

$P(n) = 0.30$

$P(e) = 0.06$

$P(n) = 0.30$
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 off exponentially with no. of evidence vars
Likelihood Weighting

- Idea:
  - fix evidence variables
  - sample only non-evidence variables
  - weight each sample by the likelihood of evidence
Likelihood weighting $P(B|C)$

- Sampling probability: $S(z,e) =$
  - Neither prior nor posterior
  - Wt for a sample $<z,e>$: $w(z,e) = \prod_i P(c_i | Parents(Z_i))$
  - Weighted Sampling probability $S(z,e)w(z,e)$
    $$ = \prod_i P(z_i | Parents(Z_i)) \prod_i P(c_i | Parents(E_i))$$
    $$ = P(z,e)$$
  - returns consistent estimates
  - performance degrades w/ many evidence vars
    • but a few samples have nearly all the total weight
    • late occurring evidence vars do not guide sample generation

- $P(B|C)$
  - $P(B|C)_{\text{Alarm}}$
  - $P(B|C)_{\text{Newscast}}$

- Samples:
  - $B E A C N$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
  - $b e a c n$
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 \( P(X=\text{true} \mid \text{all other variables}) \)
  3. Set X to true with that probability
- Repeat many times. Frequency with which any variable X is true is its posterior probability.
- Converges to true posterior when frequencies stop changing significantly
  - stable distribution, mixing

Markov Blanket Sampling

- How to calculate \( P(X=\text{true} \mid \text{all other variables}) \)?
- Recall: a variable is independent of all others given its Markov Blanket
  - parents
  - children
  - other parents of children

- So problem becomes calculating \( P(X=\text{true} \mid \text{MB}(X)) \)
- We solve this sub-problem exactly
- Fortunately, it is easy to solve
  
  \[
  P(X) = \alpha P(X \mid \text{Parents}(X)) \prod_{\text{ForChildren}(X)} P(Y \mid \text{Parents}(Y))
  \]

Example

\[
\begin{align*}
P(X) &= \alpha P(X \mid \text{Parents}(X)) \prod_{\text{ForChildren}(X)} P(Y \mid \text{Parents}(Y)) \\
&= P(A \mid B, C) P(B \mid X, C) P(A) P(C) \\
&= \alpha P(X \mid A) P(B \mid X, C)
\end{align*}
\]
Example

- Evidence: s, b
- Randomly set: h, b

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)
**Example**

- **Heart Disease**
  - 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 $P(G|s,~h,b)$

- **Lung Disease**
  - 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**

$$P(X|E) = \frac{\text{number of samples with } X=x}{\text{total number of samples}}$$

- **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