Announcements

• Updated syllabus on web site
• HW5 is available
  (due after Spring Break but don’t wait)
• Pick up midterm exams after class
  – for those who missed Tuesday
  – must stay for regrade opportunity
  – else email me for an individual time next week
  – Regrades, conflicts, ... next week
• Term Projects
  – should be underway (don’t wait)
  – proposals & approval strongly suggested
Dentist Example

BN represents a distribution. Here over 3 Boolean Random Variables:

C – Patient has a cavity

A – Patient reports a toothache

B – Dentist’s probe catches on tooth

Good for recommender systems:

Distribution of features of people, features of movies, peoples’ preferences of movies

Online items for sale, Restaurants, Credit card applications...
Inference with BNs

Five Boolean Random Variables:

B – a burglary is in progress
E – an earthquake is in progress
A – the alarm is sounding
J – John calls
M – Mary calls

What’s missing?
Bayes Net Model

**Structural Part**

- **B** → **A**
  - $\Pr(B) = 0.001$

- **E** → **A**
  - $\Pr(E) = 0.002$

**Parametric Part**

- **A** → **J**
  - $\Pr(J|A)$
    - T: 0.9
    - F: 0.05

- **A** → **M**
  - $\Pr(M|A)$
    - T: 0.7
    - F: 0.01

- **B** → **A**, **E** → **A**
  - $\begin{array}{ccc}
    B & E & \Pr(A|B,E) \\
    T & T & 0.95 \\
    T & F & 0.94 \\
    F & T & 0.29 \\
    F & F & 0.001 \\
  \end{array}$

Often denoted as $\theta$

- $\Pr(A,B,E,J,M; \theta)$
- or $\Pr(A,B,E,J,M | \theta)$
Inference with BNs

Five Boolean Random Variables:

- B – a burglary is in progress
- E – an earthquake is in progress
- A – the alarm is sounding
- J – John calls
- M – Mary calls

Are these numbers reasonable?

How would they be different if they were Joint entries?

We can compute Joint entries.
Inference with BNs

What’s the probability of a burglary?

One in a thousand: 0.001

Are burglaries or earthquakes more likely?

Earthquakes are twice as likely

What’s the distribution of the alarm sounding when there is an earthquake but no burglary?

Pr(A | B=F, E=T) or Pr(A | ¬b, e)

0.29  [also Pr(a | ¬b, e); Pr(¬a | ¬b, e) = 0.71]

Pr(J | a, ¬b, e)?

= Pr(J | a) = 0.9

Pr(b, e, a, j, m)?

.001 * .002 * .95 * .9 * .7 = .000001197
Inference with BNs

Pr(A|J=T)?
Oops, can’t look it up!
Bayes: Pr(A|J) = Pr(J|A) · Pr(A) / Pr(J)
Pr(J=T|A=T) – from the BN; = 0.9
Pr(A=T)?
Marginalize over B and E;
Blend using be  b¬e  ¬be  ¬b ¬e
Pr(b,e) = 0.001 * 0.002; Pr(a|b,e) = 0.95
so from be: 0.001 * 0.002 * 0.95 +
from b¬e: 0.001 * 0.998 * 0.94 +
.999*.002*.29 + .999*.998*.001 (¬be & ¬b ¬e)
= ~ 0.0025
Inference with BNs

\[ \Pr(A \mid J=T)? \]

Oops, can’t look it up!

Bayes: \( \Pr(A \mid J) = \Pr(J \mid A) \cdot \Pr(A) / \Pr(J) \)

\( \Pr(J=T \mid A=T) \) – from the BN; = 0.9

\( \Pr(A=T) = \sim 0.0025 \)

\( \Pr(J=T) \) – marginalize over A; = \( \sim 0.052 \)

Turns out \( \Pr(A=T \mid J=T) \) is quite low \( \sim 0.043 \) Why?

John often false positives: 0.05 and the

Alarm turns out to be unlikely \( \Pr(A=T) \sim 0.0025 \)

Why not marginalize over M?
Pr(A | ¬b)?

Marginalizing over E

blending of = Pr(A | ¬b,e) and Pr(A | ¬b,¬e)

0.29 · 0.002 + 0.001 · 0.998 = 0.001578

Why don’t we marginalize over J? or M?

Can CPTs at J or M influence our opinion on A when B=F?

What if we know that John called?

We can ignore J & M when neither is an evidence or query variable. Observing John called makes J an evidence variable

What are the general rules???
Which are true?

1) J and M are independent
2) J and M are conditionally independent knowing A
3) J and M are conditionally independent not knowing A
4) B and E are independent
5) B and E are conditionally independent knowing A
6) B and E are conditionally independent not knowing A

A: 1 & 4          B: 2 & 5
C: 2 & 6          D: 3 & 5
E: 3 & 6
When are Variables Conditionally Independent from Evidence?

A Variable is conditionally independent of all other nodes in the network given its parents, children, and children’s parents (co-parents). This is its Markov blanket.

A Variable is conditionally independent of its non-descendants given its parents.

(Text Figure 14.4)

A set of Variables X is conditionally independent of a set of Variables Y given a set of Evidence Variables E if all paths connecting an x to a y are “d-separated”
X is conditionally independent of everything else given its Markov blanket
When are Variables Conditionally Independent from Evidence?

A Variable is conditionally independent of all other nodes in the network given its parents, children, and children’s parents (co-parents). This is its Markov blanket.

A Variable is conditionally independent of its non-descendants given its parents.

(Text Figure 14.4)

A set of Variables X is conditionally independent of a set of Variables Y given a set of Evidence Variables E if all paths connecting an x to a y are “d-separated”
X is conditionally independent of its non-descendants given its parents.
When are Variables Conditionally Independent from Evidence?

A Variable is conditionally independent of all other nodes in the network given its parents, children, and children’s parents (co-parents). This is its Markov blanket.

A Variable is conditionally independent of its non-descendants given its parents.

(Text Figure 14.4)

A set of Variables X is conditionally independent of a set of Variables Y given a set of Evidence Variables E if all paths connecting an x to a y are “d-separated”
d-separation
Automated Inference in BNs

- Recover the joint
- Conditionalize on evidence
- Marginalize over unknowns
- Brute Force calculation of the Answer
- May be intractable for large BNs
- Joint grows exponentially in #RV’s while real world BN CPTs often grow no more than polynomially
- Earthquake BN calculations were all much simpler
  - Did not recover the full joint
  - Did not touch irrelevant values (Our cleverness in appreciating cond. Indep.)
- Several exact inference algorithms
  - Make use of same structural conditional independence properties
  - Message-passing algorithms
  - Exact & efficient for polytrees
Polytrees / Non-Polytrees

Polytree: BN with at most one undirected path between any two nodes

Non-polytree: existence of undirected cycles, there are NEVER directed cycles in a BN
Space Complexity of Distribution Data Structures

BNs win big if direct dependencies (fan in) are bounded (limited to k)

Joint complexity:
O\(a^n\): \(2^n - 1\) for n Boolean variables

BN w/ at most k parents / node:
O\(n\): \(\leq n \cdot 2^k\) for Boolean variables
Time Complexity of Inference

• Inference in non-polytrees can be exponential
  – Even approximate inference
  – Often works better than worst-case in practice
  – Belief propagation
    • Several related message-passing algorithms
    • Only guaranteed to work for polytrees
    • Else may not converge; may converge to incorrect answers
  – For non-polytrees
    • Loopy belief propagation
    • Often works efficiently and quite accurately
    • No one understands why or when...
Cluster Forming Algorithms for Non-Polytrees

- Join tree / junction tree algorithm
- Restore polytree property
- Identify clusters of nodes (super nodes)
- Collect all parents
- Exponential in cluster size since cross product of contexts
Non-Polytrees
approximate inference

• Monte Carlo sampling algorithms
  – Generate examples from the conditional distribution
  – Conditionalized on the evidence
  – Estimate the distribution by counting

• Principled use of randomization

• Construct a generator of samples from the desired distribution and count appropriately

• Often this is far easier than trying to construct the conditionalized distribution itself
X is conditionally independent of its non-descendants given its parents.
Monte Carlo Algorithms I
Rejection Sampling

• Start w/ all RV’s unknown
• Choose a *topological* ordering
  (parents before children)
• Get random numbers to assign values
• Descendents are always unknown when assigning a RV
• Generate lots of samples
• Count, but ignore (reject) samples violating evidence
• Can require lots of samples; when?
Rejection Sampling Example

Pr(A | B=False)

Boolean RV’s C: cavity, A: ache, B: probe catches
Rnd() give uniform random floats [0.,1.]
Topological ordering: C, A, B

Assign C: get Rnd(); set C: True if < .5 else False
  suppose Rnd() gives 0.718652 so C=False
Assign A: get Rnd(); set A: True if < .25 else False
  suppose Rnd() gives 0.1582941 so A=True
Assign B: get Rnd(); set B: True if < .25 else False
  suppose Rnd() gives 0.065269 so B=True
Reject this sample
Rejection Sampling Example

Pr(A | B=False)

Boolean RV’s C: cavity, A: ache, B: probe catches
Rnd() give uniform random floats [0.,1.]
Repeat many times:

Assign C: get Rnd(), set C: True if < .5 else False
Assign A: get Rnd(),
    if C=True, set A: True if < .5 else False
    if C=False, set A: True if < .25 else False
Assign B: get Rnd(),
    if C=True, set B: True if < .95 else False
    if C=False, set B: True if < .25 else False
Keep the sample only if B=False

| C  | Pr(A|C) |
|----|---------|
| T  | 0.5     |
| F  | 0.25    |

| C  | Pr(B|C) |
|----|---------|
| T  | 0.95    |
| F  | 0.25    |

Sample |
---|---|---|---|
1  | F  | T  | F  |
2  | T  | T  | F  |
3  | F  | F  | F  |
...

Pr(A | B=False) = (# A=T) / (# Saved Samples)
Monte Carlo algorithms II
Likelihood Weighting

• Generate a sample randomly, assigning values topologically as before
• Except at evidence nodes luckily you get the evidence value
  – how “lucky” was this?
  – compute $\Pr(\text{evidence} \mid \text{parents})$
  – this value and its frequency reflects previous random numbers
• Generate lots of samples
• Reject NOTHING. All samples are consistent w/ evidence!
• But count each fractionally, weighted by its total likelihood
  (product of the “lucky” evidence assignments)
• Some may count for very little
X is conditionally independent of everything else given its Markov blanket
Monte Carlo algorithms III

Gibbs Sampling

• Gibbs: simple version of Markov Chain Monte Carlo (MCMC)
• MCMC is a general, important, and simple algorithm
• Why it works is somewhat involved
• Fix evidence variables to their values
• Initialize others randomly
• Iterate through nodes (must visit all many times)
• Resample each non-evidence variable given its Markov blanket
• Each change generates a (slightly different) example (with values assigned to each random variable)
• After ‘burn in,’ accumulate each k’th sample
• Count