

Bayes IV (Approximate Inference)

Lecture #20
11/11/08

Announcements

- Programming Assignment #2 is done
- Programming Assignment #3 to be distributed shortly
- Reading: Bayes Net Chapters
 - Chapter 13: subjective vs frequentist probabilities, axioms of probability, Bayes' Rule.
 - Chapter 14: Bayes nets, exact inference, approximate inference
- ACM club: Wed. @5:00 USC 110
 - Learn to interact with shadows!

Approaches to inference

- Exact inference
 - Variable elimination
 - Join tree algorithm (*not covered*)
- Approximate inference
 - Simplify the structure of the network to make exact inference efficient (variational methods, loopy belief propagation) (*one slide*)
- Probabilistic methods
 - Stochastic simulation / sampling methods
 - Markov chain Monte Carlo methods

Network simplification

Typical simplifications:

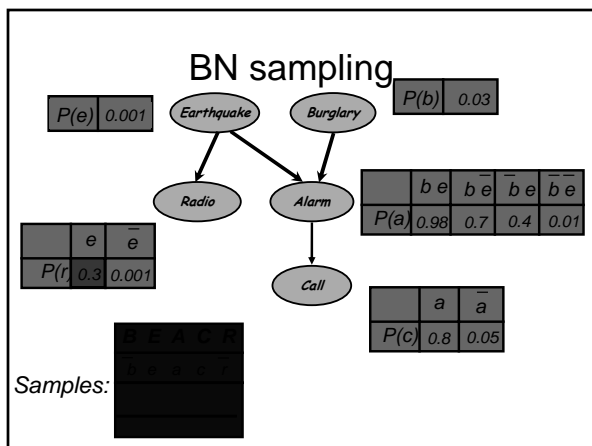
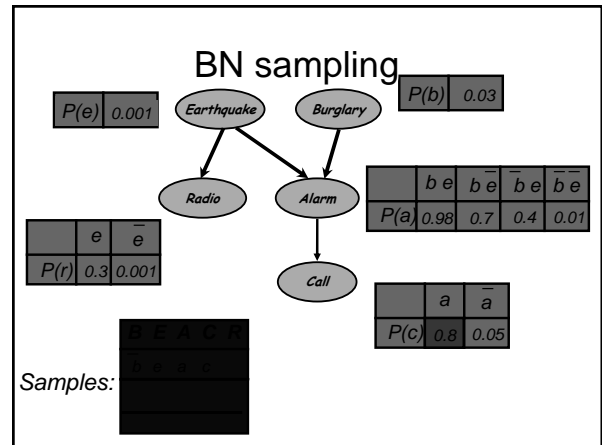
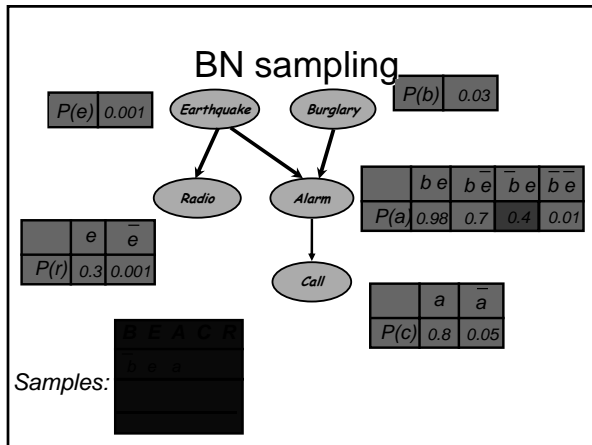
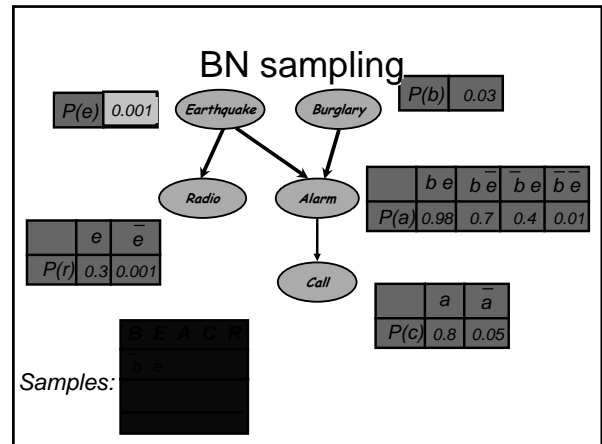
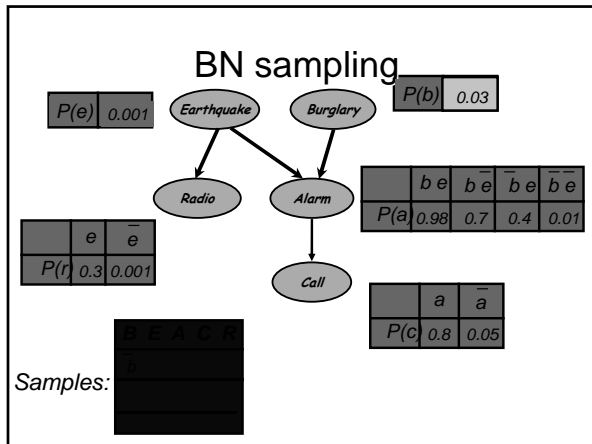
- Remove parts of the network
- Remove edges
- Reduce the number of values (value abstraction)
- Replace a sub-network with a simpler one (model abstraction)
- These simplifications are often w.r.t. to the particular evidence and query

Inference by sampling

- Suppose we can sample instances $\langle X_1, \dots, X_n \rangle$ according to $P(X_1, \dots, X_n)$
- Want to compute $P(e)$
- The probability that a random sample $\langle X_1, \dots, X_n \rangle$ satisfies e is approximately $P(e)$
- We can view each sample as tossing a biased coin with probability $P(e)$ of "Heads"

Sampling a Bayesian Network

- If $P(X_1, \dots, X_n)$ is represented by a Bayesian network, can we efficiently sample from it?
- Idea: sample according to structure of the network
 - Write distribution using the chain rule, and then sample each variable given its parents



BN sampling

- Let X_1, \dots, X_n be order of variables consistent with arc direction
- for $i = 1, \dots, n$ do
 - sample x_i from $P(X_i \mid Pa(X_i))$
 - (Note: since $Pa(X_i) \subseteq \{X_1, \dots, X_{i-1}\}$, we already assigned values to them)
- return x_1, \dots, x_n

BN sampling

- Sampling a complete instance is linear in number of variables
 - Regardless of structure of the network
- However, if $P(e)$ is small, we need many samples to get a decent estimate

Can we sample from $P(X_1, \dots, X_n | e)$?

- If evidence is in roots of network, easily
- If evidence is in leaves of network, we have a problem
 - Our sampling method proceeds according to order of nodes in graph
- Rejection sampling: keep those instantiations that are consistent with the values of the evidence variables
- Estimate $P(X|e)$ by $N(X,e) / N(e)$ where $N(\cdot)$ counts the number of times an event was sampled.

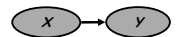
Likelihood Weighting

- Can we ensure that all of our sample satisfy e ?
- One simple solution:
 - When we need to sample a variable that is assigned value by e , use the specified value
- For example: we know $Y = 1$
 - Sample X from $P(X)$
 - Then take $Y = 1$
- Is this a sample from $P(X, Y | Y = 1)$?



Likelihood Weighting

- **Problem:** these samples of X are from $P(X)$
- **Solution:**
 - Penalize samples in which $P(Y=1|X)$ is small
- We now sample as follows:
 - Let $x[i]$ be a sample from $P(X)$
 - Let $w[i]$ be $P(Y = 1 | X = x[i])$



$$P(X = x | Y = 1) \approx \frac{\sum_i w[i] P(X = x | x[i])}{\sum_i w[i]}$$

Markov chain Monte Carlo sampling

- Generates events by making random changes to the state variable.
- The next state is generated by sampling a value for one of the nonevidence variables conditioned on the current values.

Digression: Markov chains

- A Markov chain is a random process (infinite sequence of random variables) $(X(0), X(1), \dots, X(t), \dots)$ that satisfies:

$$P(X(t) | X(0), \dots, X(t-1)) = P(X(t) | X(t-1))$$
- The probability of a particular state at time t depends only on the state at time $t-1$
- If the transition probabilities are fixed for all t , the chain is called *homogeneous* and is characterized by a transition matrix T .



Markov chains

- In order for a Markov chain to be useful for sampling from $P(x)$, we require that for any starting state $x(0)$:

$$\lim_{t \rightarrow \infty} P_t(x) = P(x)$$

- Equivalently, the stationary distribution of the Markov chain must be $P(x)$:

$$P_{t+1}(x') = \sum_{\mathbf{x}} P_t(\mathbf{x})Q(\mathbf{x} \rightarrow \mathbf{x}')$$

the transition probability from x to x'

$$P(x') = \sum_{\mathbf{x}} P(\mathbf{x})Q(\mathbf{x} \rightarrow \mathbf{x}')$$

Using a Markov chain to sample

- If the Markov chain indeed converges to the desired distribution from, we can start in an arbitrary state, use the Markov chain to do a random walk for a while, and output the $x(t)$.
- The resulting state will be sampled from $P(x)$.

Stationary distribution

$$Q = \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{pmatrix}$$

- The stationary distribution of this chain is $(0.33, 0.33, 0.33)$

Markov chains for sampling

- To ensure that the chain converges to a unique stationary distribution the following conditions are sufficient:
 - *Irreducibility*: every state is eventually reachable from any start state; for all x, y there exists a t such that $P_t(y) > 0$ when starting at x
 - *Aperiodicity*: the chain doesn't get caught in cycles.
- The process is *ergodic* if it is both irreducible and aperiodic

Detailed balance

- To ensure that the stationary distribution of the Markov chain is $P(x)$ it is sufficient for P and Q to satisfy the *detailed balance (reversibility)* condition:

$$P(\mathbf{x})Q(\mathbf{x} \rightarrow \mathbf{x}') = P(\mathbf{x}')Q(\mathbf{x}' \rightarrow \mathbf{x})$$

Given that detailed balance holds:

$$\begin{aligned} \sum_{\mathbf{x}} P(\mathbf{x})Q(\mathbf{x} \rightarrow \mathbf{x}') &= \sum_{\mathbf{x}} P(\mathbf{x}')Q(\mathbf{x}' \rightarrow \mathbf{x}) \\ &= P(\mathbf{x}') \sum_{\mathbf{x}} Q(\mathbf{x}' \rightarrow \mathbf{x}) \\ &= P(\mathbf{x}') \end{aligned}$$

Gibbs sampling

- Idea: To transition from one state (variable assignment) to another by:
 - Pick a variable X_j ,
 - Sample its value from the conditional distribution $P(x_j / x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n)$
- In a Bayesian network x_j depends only on a subset of the variables.

Markov Blanket

- Variables are independent of their non-descendants given their parents
- Variables are independent of *everything else in the network* given their *Markov blanket*.
- So, to sample a node, only need to condition on its Markov blanket:

$$P(x_j / x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n) = P(x_j / MB(x_j))$$

The Gibbs sampling algorithm

GIBBS(X, e, bn, N) returns estimate of $P(X|e)$
 $N[x]$ - counts the number of times each value of X was observed
 $x[j]$ - the current state of the network $x[0]$ initialized with random values for the nonevidence variables
for $j = 1$ to N do
 for each nonevidence variable X_i
 sample X_i from $P(X_i|MB(X_i))$
 $N[x] = N[x] + 1$, where x is the value of X in $x[j]$

Convergence of Gibbs sampling

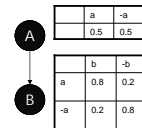
- Gibbs sampling satisfies detailed balance:

$$\begin{aligned} P(\mathbf{x}|e)P(x'_i|\bar{\mathbf{x}}_i, e) &= P(x_i, \bar{\mathbf{x}}_i|e)P(x'_i|\bar{\mathbf{x}}_i, e) \\ &= P(x_i|\bar{\mathbf{x}}_i, e)P(\bar{\mathbf{x}}_i, e)P(x'_i|\bar{\mathbf{x}}_i, e) \\ &= P(x_i|\bar{\mathbf{x}}_i, e)P(x'_i, \bar{\mathbf{x}}_i|e) \\ &= P(\mathbf{x}'|e)P(x_i|\bar{\mathbf{x}}'_i, e) \end{aligned}$$

$\bar{\mathbf{x}}_i$ denotes the variables other than x_i

Gibbs sampling example

- Consider a 2 variable network:



- Initialize randomly
- Sample variables alternately

Practical issues

- How many iterations?
- When to stop?