CSC 444: Logical Methods in AI

Bayesian Network

Shantonu Hossain, Adam Purtee

Outline

- Motivation
- Non-categorical Reasoning
  - Objective Probability
  - Subjective Probability
  - Vague Predicates
- Basic concepts of probability
- Bayesian/Belief Network
  - Syntax
  - Semantics
  - Inference in Bayesian Network
  - Applications
- Summary

Motivation

- Logical Reasoning needs sufficient information of the world to prove any assertion or reach any conclusion
  - Agents never have access to the whole truth about their environment
  - Agent may have incomplete or incorrect understanding of its environment
- Example: Agent under uncertainty
- Goal: Drive someone to the airport to catch a flight
- Plan A90:
  - leave home 90 mins before the flight departure
  - Drive at a reasonable speed
- Fact:
  - Distance to airport is 15 miles

Example: Agent under Uncertainty

- Agent can't decide 'Plan A90 will get us to the airport in time'
- Reaches weaker conclusion – 'Plan A90 will get us to the airport in time as long as
  - My car doesn't break down or out of gas
  - I don't get into any accident
  - There is no road blocking on the way
  - The plane doesn't leave early
  - …

- Need to build an uncertain-reasoning system
  - Capture uncertain knowledge in an efficient way
  - Reach rational decision even when there is not enough information to prove an action
- Expand our interpretation of P-> Q using probabilities
  - introduce number to avoid categorical nature of binary logical values (true/false)
  - ‘All birds fly’ to ‘95% of birds fly’

Non-Categorical Reasoning

- 3 types of modification may be performed to make our standard logic flexible:
- Relax the strength of the quantifier
  - for all x <=> for most x
  - Our use of probabilities is objective, not subject to the interpretation or degrees of confidence
- Relax the applicability of the predicate
  - everyone in our class is absolutely tall<=> everyone in our class is moderately tall
  - Vague predicate, a person can be simultaneously both tall(strongly) and not tall(weakly)

- Relax our degree of belief in the sentence as a whole
  - Everyone in this room has finished the AI project <=> I believe that everyone in this room has finished the AI project, but I am not very sure.
  - We are dealing with uncertain knowledge, reflects individuals personal degree of belief, subjective probability
- All these 3 representation can work together:
  - 'I am pretty sure that most of the persons in the class is fairly tall'
  - connects all 3 approaches

Objective Probability

- A statistical interpretation =>frequency of occurrence of an event
- Requires repeatable experiments
- Doesn’t depend on subject’s interpretation
- Doesn’t depend on degrees of confidence
- Doesn’t need prior knowledge
- Example:
- What is the probability of head of an unbiased coin?
  - Toss coin for 10,000 times
  - Count number of heads = num_head
  - P(head) = num_head/10,000 ≈ 0.5

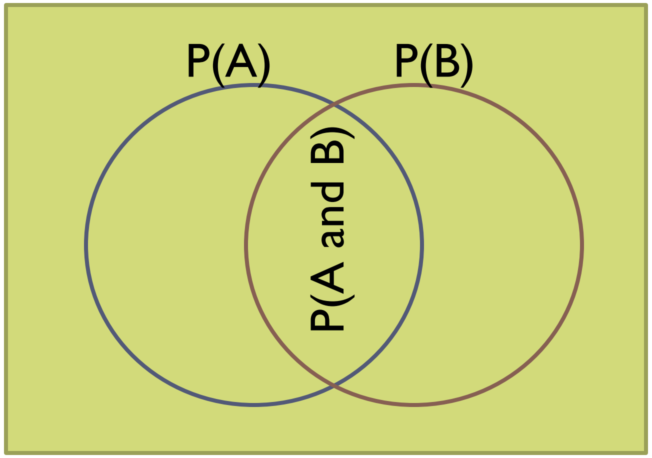
Subjective Probability

- An subjective interpretation => individual's degrees of belief in the occurrence of an event
- Derives from observations of group of things in the world
- Evidence combines to achieve new confidence level in the belief (posterior probability) from the previous level (prior probability)
  - Prior probability + Evidence = Posterior probability
- Example
- P(rain) = 0.2 ⇐ Prior probability
- P(rain| grass is wet) = 0.8
- P(rain| grass is wet ^ rain) = 1.0

Basic Concepts of Probability Theory

The Axioms of Probability

- Probability of an event A , P(A) is a number expressing the chance that A will occur
- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(~A) = 1 - P(A)
- P(A U B) = P(A) + P(B) - P(A and B)



Basic Notions of Probability

- **Unconditional/Prior Probability**
  - The probability that a proposition is true in the absence of any other information

    P(Weather = Sunny) = 0.2
- **Joint Probability**
  - A table which specifies the probability of every combination of values for a set of random variables.
  - P(Sunny, Cavity, Toothache)

sunny	cavity	toothache	probability
0	0	0	
0	0	1	
0	1	0	
0	1	1	
1	0	0	
1	0	1	
1	1	0	
1	1	1	

Basic Notions of Probability

- Conditional Probability
  - P(A|B) – the probability that A occurs given that B occurs

    $P(A|B) = P(A \wedge B) / P(B)$
  - Also written as the product rule: $P(A \wedge B) = P(A|B)*P(B)$
- Independence
  - A and B are said to be independent exactly if

    $P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A \wedge B) = P(A)*P(B)$

    (note: these statements are equivalent.)
- Conditional Independence
  - Two events A and B are conditionally independent given E if

    $P(A \wedge B|E) = P(A|E)*P(B|E)$

Basic Notions of Probability

- Bayes' rule
  - P(B|A) = (P(A|B) * P(B)) / P(A)
  - Usefulness: causation knowledge is more frequent than diagnostic knowledge.
- Bayes' rule with evidence
  - P(B|A ^ E) = (P(A | B ^ E) * P(B | E)) / P(A | E)

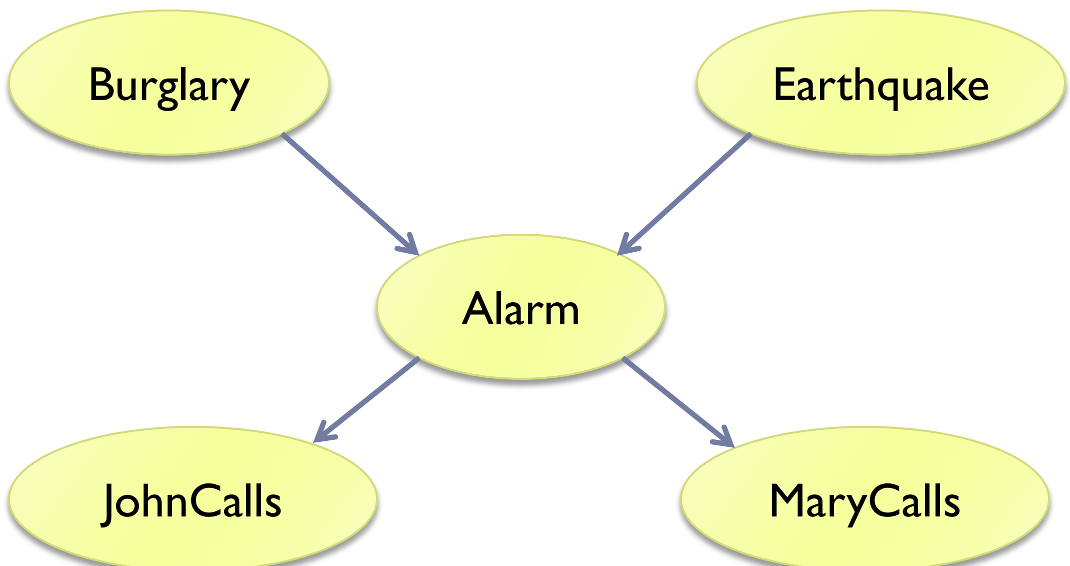
Bayesian Network

Bayesian/Belief Network

- A reasoning system –
  - uses graph theory to reason with uncertainty
  - follows the laws of probability theory
- Definition: A graphical model that represents a set of random variables and their conditional dependencies by Directed Acyclic Graph (DAG)
  - Graphical model = Probability theory + graph theory
- Syntax:
  - One node per random variable
  - A directed link between one node to another if there is any dependency
  - A conditional probability table (CPT) for each node given its parents: $\mathbf{P}(x_i \mid \text{Parents}(X_i))$

Example

- A topology of belief network
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call
- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*



Semantics

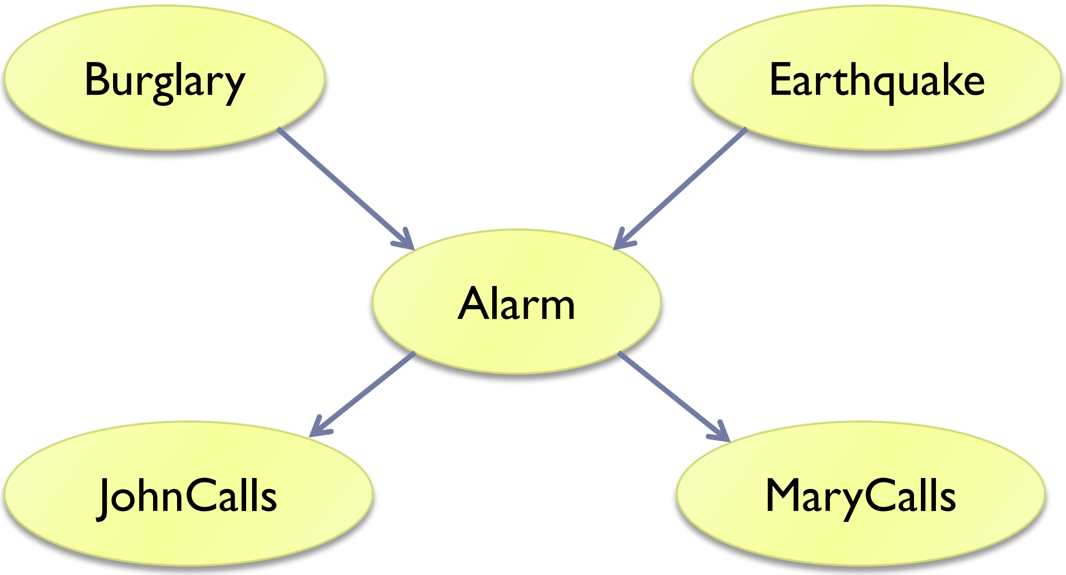
- Full joint distribution is defined as the product of the conditional distribution of each node

  $\boldsymbol{P}(x_1, \ldots, x_n) = \prod_{i=1} \boldsymbol{P}(x_i \mid Parents(X_i))$

- CPT provides decomposed representation of joint distribution
- Explanation

$$\begin{aligned}\boldsymbol{P}(x_1, \ldots, x_n) &= P(x_n \mid x_{n-1}, \ldots x_1) P(x_{n-1}, \ldots x_1) \\ &= P(x_n \mid x_{n-1}, \ldots x_1) P(x_{n-1} \mid x_{n-2} \ldots x_1) \ldots P(x_2 \mid x_1) P(x_1) \\ &= \prod_{i=1} \boldsymbol{P}(x_i \mid x_{i-1}, \ldots x_1) \\ &= \prod_{i=1} \boldsymbol{P}(x_i \mid Parents(X_i))\end{aligned}$$

Example



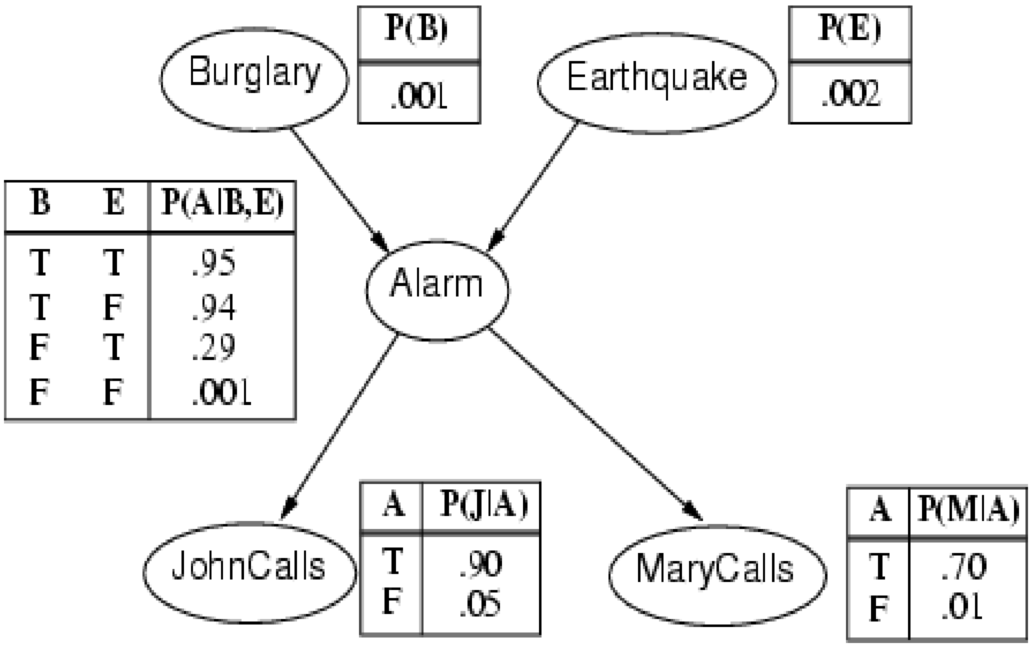
$$\mathbf{P}(x_1, \ldots, x_n) = \prod_{i=1} \mathbf{P}(x_i \mid Parents(X_i))$$

P(JohnCalls ^ MaryCalls ^ Alarm ^ Burglary ^ Earthquake)
= P(JohnCalls|Alarm) x P(MaryCalls|Alarm) x P(Alarm|Burglary^Earthquake)
x P(Burglary) x P(Earthquake)

Construction of Belief Network

- Choose the set of relevant variables X_i that describe the domain
- Choose an ordering of variables $X_1, \ldots, X_n$
- For $i = 1$ to n
  - pick a variable X_i and add a node to the network for it
  - select parents from $X_1, \ldots, X_{i-1}$ such that $\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$
  - define conditional probability table for X_i

Problem 1

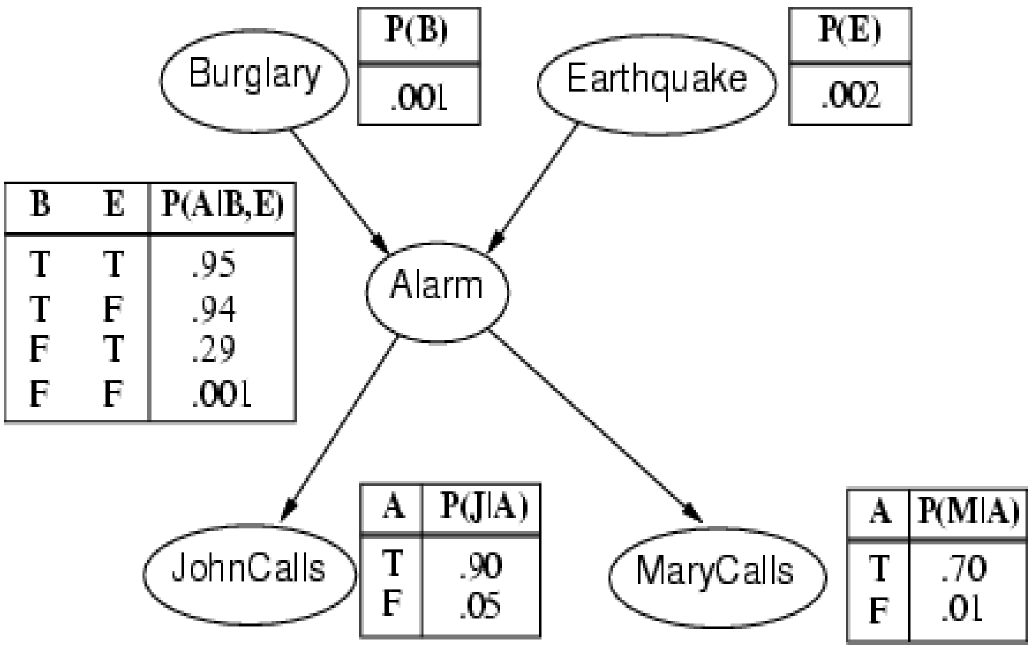


J: JohnCalls
M: MaryCalls
A: Alarm
B: Burglary
E: Earthquake

What is the probability of the event that the alarm has sounded and no burglary but an earthquake has occurred and both Mary and John call?

$$P(J \wedge M \wedge A \wedge \sim B \wedge E) = P(J|A) \times P(M|A) \times P(A|\sim B \wedge E) \times P(\sim B) \times P(E)$$
$$= 0.90 \times 0.70 \times 0.29 \times 0.999 \times 0.002 = 0.00036$$

Problem 2



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

J: JohnCalls
M: MaryCalls
A: Alarm
B: Burglary
E: Earthquake

What is the probability of the event that the alarm has sounded but neither a burglary nor an earthquake has occurred and John call and Mary didn't call?

$$P(J \wedge \sim M \wedge A \wedge \sim B \wedge \sim E) = P(J|A) \times P(\sim M|A) \times P(A|\sim B \wedge \sim E) \times P(\sim B) \times P(\sim E)$$
$$= 0.90 \times 0.30 \times 0.001 \times 0.999 \times 0.998 = 0.00027$$

Compactness of Bayesian Network

Suppose that the maximum number of variables on which any variable directly depends is k. Then a Bayesian network can be specified by n*2^k numbers, as opposed to 2^n for the full joint distribution.
Moreover, the full joint distribution can be computed from the Bayesian network.

AIMA Example: n = 32, k = 5 → 960 vs 4bn

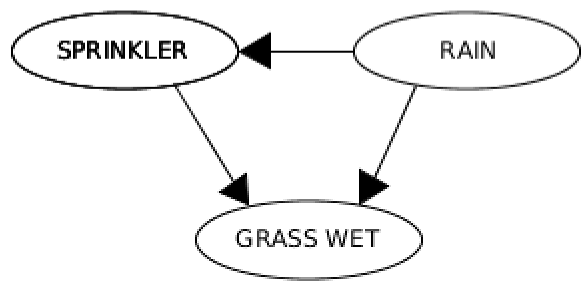
Compactness vs. Accuracy

Compactness and Node Ordering
Nodes for root causes should be added before the nodes they influence.

Exact Inference

Simple, intuitive algorithm:
enumeration of joint distribution and Bayes' rule.

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

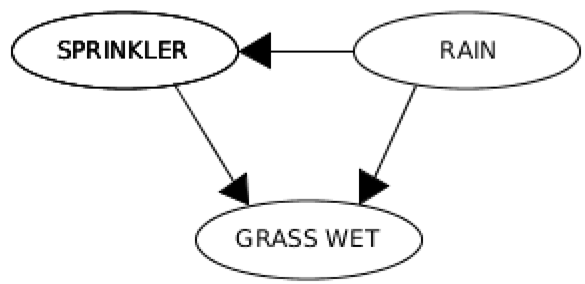
		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

What is P(S|G)?

Exact Inference

Simple, intuitive algorithm:
enumeration of joint distribution and Bayes' rule.

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

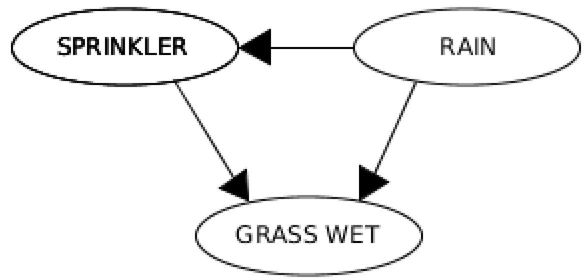
What is P(S|G)?

P(S|G) = P(S^G)/P(G)

P(S^G) = P(S^G^R) + P(S^G^~R)
P(G) = P(S^G) + P(~S^G)
P(~S^G) =P(S^G^R) + P(S^G^~R)

Exact Inference

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

What is P(S|G)?

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

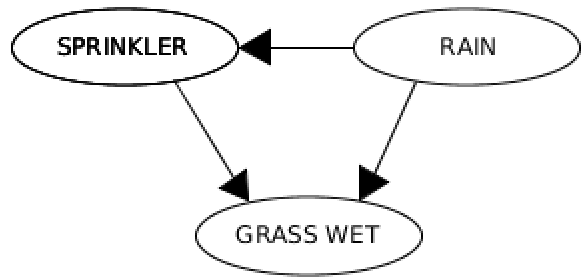
P(S|G) = P(S^G)/P(G)

P(S^G) = P(S^G^R) + P(S^G^~R) = 0.00198+0.288 = .28998
P(G) = P(S^G) + P(~S^G) = .28998 + 0.1584 = 0.44838

P(~S^G) =P(~S^G^R) + P(~S^G^~R) = 0.1584 + 0
P(S^G^R)=P(S|R)P(G|S^R)P(R) = (0.01)(0.99)(0.2) = 0.00198
P(S^G^~R) = P(S|~R)P(G|S^~R)P(~R) = (0.4)(0.9)(0.8) = 0.288
P(~S^G^R) = P(~S|R)P(G|~S^R)P(R) = (0.99)(0.8)(0.2) = 0.1584
P(~S^G^~R) = P(~S|~R)P(G|~S^~R)P(~R) = (0.6)(0.0)(0.8) = 0

Exact Inference

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

What is P(S|G)?

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

P(S|G) = P(S^G)/P(G)
0.6467

P(S^G) = P(S^G^R) + P(S^G^~R) = 0.00198+0.288 = .28998
P(G) = P(S^G) + P(~S^G) = .28998 + 0.1584 = 0.44838

P(~S^G) =P(~S^G^R) + P(~S^G^~R) = 0.1584 + 0
P(S^G^R)=P(S|R)P(G|S^R)P(R) = (0.01)(0.99)(0.2) = 0.00198
P(S^G^~R) = P(S|~R)P(G|S^~R)P(~R) = (0.4)(0.9)(0.8) = 0.288
P(~S^G^R) = P(~S|R)P(G|~S^R)P(R) = (0.99)(0.8)(0.2) = 0.1584
P(~S^G^~R) = P(~S|~R)P(G|~S^~R)P(~R) = (0.6)(0.0)(0.8) = 0

Exact Inference

- Simple, intuitive algorithm:
- enumeration of joint distribution and Bayes' rule.

What is P(S|G)... a lot of work!

Our “algorithm” has time complexity O(n*2^n).

Using dynamic programming, we can get this down to linear time for well-behaved networks (polytrees), but the general case still requires exponential time O(2^n).

The general case is NP-hard (even #P-hard), so exact inference in Bayesian networks is not always feasible

Approximate Inference

Direct Sampling   (Wonky Demo)

Grab a probability for a specific row of Joint distribution

Rejection Sampling (Wonky Demo)

Compute a conditional probability via repeated direct sampling, rejecting the samples in which the evidence does not hold.

Error bounds: stddev(error) ~ 1/sqrt(N)

Problem: rare occurrences

Approximate Inference

Likelyhood Weighting

Compute a conditional probability, but generate only samples consistent with evidence. Weight these samples by their likelyhood, and compute .

To generate a sample:
Let w = 1.
For each variable X_i (i = 1, 2, ...)
If X_i is in the evidence set, set w = w * P(X_i) and X_i = t.
Otherwise Sample variable X_i

After assigning values to each variable, you have a weighted sample. This can be repeated to generate N weighted-samples, where the total weight of the target samples when divided by the total weight of the samples yields the desired conditional probability.

Approximate Inference

Markov Chain Monte Carlo Simulation (MCMC)

Partition the variables into hidden (X) and evidence (E).
Compute a “state” by randomly initializing all variables.
Iteratively sample the hidden variables given, updating the state.
(Keep the evidence variables fixed.)
Maintain an |X| length array N where N[i] is the number of times variable X_i was true.
After desired number of runs, compute the ratio as before.

From AIMA: The sampling process settles into a “dynamic equilibrium” in which the long-run fraction of time spent in each state is exactly proportional to its posterior probability.

Extensions

Arbitrary Discrete Random Variables (We used boolean)

Continuous random variables: discretization & pdfs.

Hybrid models: continuous and discrete variables

Applications / Real-world Examples

Computational Biology

Medicine: Diagnosis

Document Classification

Information Retrieval

Finance

Law

References

- Knowledge Representation and Reasoning, Ronald J. Brachman, Hector J. Levesque, chapter 12
- Artificial Intelligence- A Modern Approach, Stuart Russel, Peter Norvig, chapter 13, 14
- Wikipedia article on Bayesian Network, http://en.wikipedia.org/wiki/Bayesian_network
- A tutorial on Inference in Bayesian Networks, Scott Davies and Andrew Moore, http://www.cs.cmu.edu/afs/cs/Web/People/awm/tutorials/bayesinf.html