Bayes Nets for Representing and Reasoning About Uncertainty

Ronald J. Williams

CSG220

Spring 2007

Contains slides adapted from: (i) an Andrew Moore tutorial with the same title; (ii) Russell & Norvig's *AIMA* site; and (iii) Alpaydin's *Introduction to Machine Learning* site.

What we'll discuss

- Recall the numerous and dramatic benefits of Joint Distributions for describing uncertain worlds
- Reel with terror at the problem with using Joint Distributions
- Discover how Bayes Net methodology allows us to build Joint Distributions in manageable chunks
- Discover there's still a lurking problem…
- …See workarounds for that problem

Why this matters

- One of the most important conceptual advances in the Machine Learning / AI field to have emerged in the last 20 years.
- A clean, clear, manageable language and methodology for expressing what you're certain and uncertain about
- Already, many practical applications in medicine, factories, helpdesks:

  P(this problem | these symptoms)

  anomalousness of this observation

  choosing next diagnostic test | these observations

Why this matters

- One of the most important conceptual advances in the Machine Learning / AI field to have emerged in the last 20 years.
- A clean, clear, manageable language and methodology for expressing what you're certain and uncertain about
- Already, many practical applications in medicine, factories, helpdesks:

  P(this problem | these symptoms)

  anomalousness of this observation

  choosing next diagnostic test | these observations

Active Data Collection

Inference

Anomaly Detection

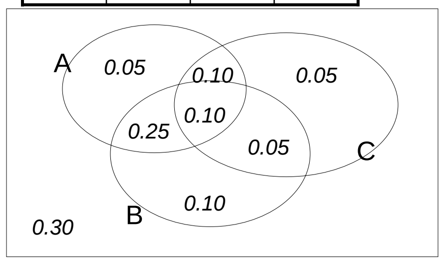
The Joint Distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

Example: Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Joint distributions

- Good news

  Once you have a joint distribution, you can ask important questions about stuff that involves a lot of uncertainty

- Bad news

  Impossible to create for more than about ten attributes because there are so many numbers needed when you build the damn thing.

Using fewer numbers

Suppose there are three events:

- W: the grass is wet in the morning
- S: the sprinkler was on during the night
- R: it rained during the night

The joint for these events contain 8 entries.

If we want to build the joint we'll have to invent those 8 numbers. OR WILL WE??

- We may not have to specify these numbers for all bottom level conjunctive events such as P(~W^S^~R)
- Instead it may sometimes be more convenient for us to specify things like: P(S), P(R), P(W|S^R), etc.

But in general these quantities are not sufficient to determine the entire joint distribution.

Using fewer numbers

Suppose there are three events:

- W: the grass is wet in the morning
- S: the sprinkler was on during the night
- R: it rained during the night

The joint for these events contain 8 entries.

If we want to build the joint we'll have to invent those 8 numbers. OR WILL WE??

- We may not have to specify these numbers for all bottom level conjunctive events such as P(~W^S^~R)
- Instead it may sometimes be more convenient for us to specify things like: P(S), P(R), P(W|S^R), etc.

But in general these quantities are not sufficient to determine the entire joint distribution.

What extra assumptions could help?

Using fewer numbers

Suppose there are three events:

- W: the grass is wet in the morning
- S: the sprinkler was on during the night
- R: it rained during the night

The joint for these events contain 8 entries.

If we want to build the joint we'll have to invent those 8 numbers. OR WILL WE??

- We may not have to specify these numbers for all bottom level conjunctive events such as P(~W^S^~R)
- Instead it may sometimes be more convenient for us to specify things like: P(S), P(R), P(W|S^R), etc.

But in general these quantities are not sufficient to determine the entire joint distribution.

What extra assumptions could help?

Answer: (Conditional) Independence

Independence

"Whether it rained during the night does not depend on and does not influence whether the sprinkler was on during the night."

This can be specified very simply:

$$P(S | R) = P(S)$$

This is a powerful statement!

It requires domain knowledge of a different kind than numerical probabilities. It needs an understanding of causation.

Independence

From $P(S|R) = P(S)$, the rules of probability imply:

$P(\sim S|R) = P(\sim S|\sim R) = P(\sim S)$
$P(R|S) = P(R|\sim S) = P(R)$
$P(S|\sim R) = P(S|R) = P(S)$
$P(\sim R|S) = P(\sim R|\sim S) = P(\sim R)$

Can you prove these?

$P(R \wedge S) = P(R)P(S)$
$P(\sim R \wedge S) = P(\sim R)P(S)$
$P(R \wedge \sim S) = P(R)P(\sim S)$
$P(\sim R \wedge \sim S) = P(\sim R)P(\sim S)$

Bayes Nets: Slide 11

Independence

From these statements, we can derive the full joint for S and R from the two numbers P(S) and P(R):

S	R	Prob
T	T	
T	F	
F	T	
F	F	

What about the joint for W, S, and R?

Bayes Nets: Slide 12

Chain Rule for Probabilities

For any r.v.'s $X_1,...,X_n$

$$P(X_1,\ldots,X_n)=\prod_{i=1}^{n}P(X_i \mid X_{i+1},\ldots,X_n)$$

Easily proved by induction

Applied to our W, S, R case:

P(W=a,S=b,R=c)

= P(W=a | S=b,R=c)P(S=b | R=c)P(R=c)

= P(W=a | S=b,R=c)P(S=b)P(R=c)

for any combination of a,b,c = True/False

Bayes Nets: Slide 13

Independence implies fewer numbers

Thus the full joint for W,S,R requires only the 6 numbers

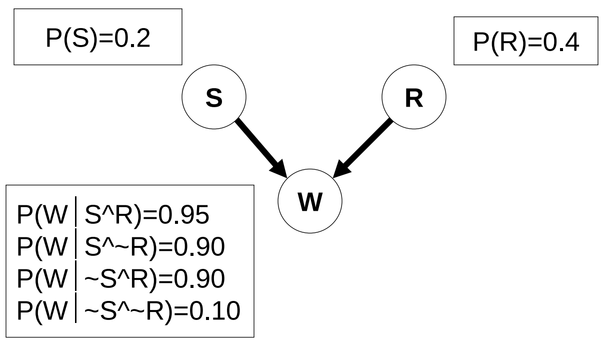
- P(S)
- P(R)
- P(W | S,R)
- P(W | S,~R)
- P(W | ~S,R)
- P(W | ~S,~R)

So we've expressed a full joint pdf using a "mere" 6 numbers instead of 7*

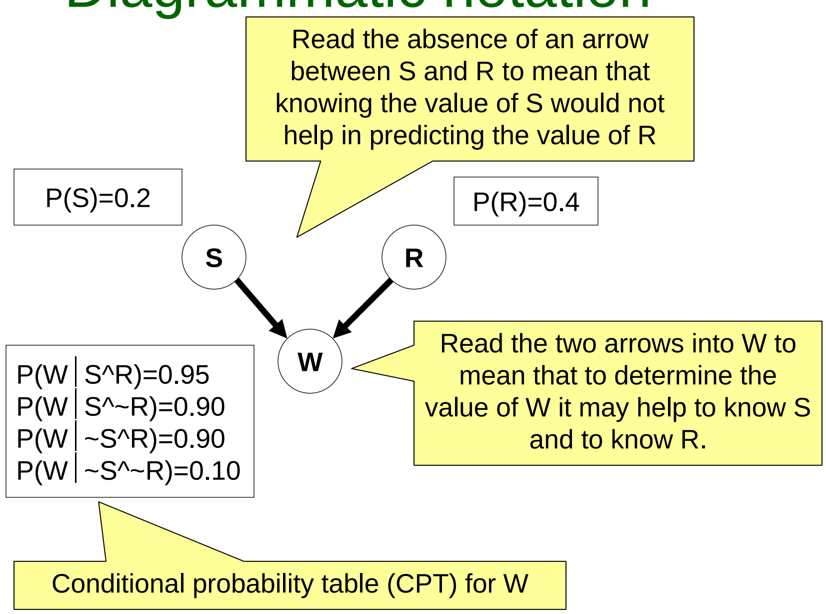
**Savings are larger for larger numbers of variables.*

Bayes Nets: Slide 14

Diagrammatic notation



Diagrammatic notation



The full joint

W	S	R	Prob.
T	T	T	0.076
T	T	F	0.108
T	F	T	0.288
T	F	F	0.048
F	T	T	0.004
F	T	F	0.012
F	F	T	0.032
F	F	F	0.432

E.g., P(~W ^ S^ ~R) = P(~W | S^ ~R) P(S | ~R) P(~R)

= P(~W | S^ ~R) P(S) P(~R)

= (0.10)*(0.20)*(0.60)

= 0.012

Causal Inference

W	S	R	Prob.
T	T	T	0.076
T	T	F	0.108
T	F	T	0.288
T	F	F	0.048
F	T	T	0.004
F	T	F	0.012
F	F	T	0.032
F	F	F	0.432

If the sprinkler was on, what is the probability that the grass is wet?

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

= (0.076 + 0.108) / (0.076 + 0.108 + 0.004 + 0.012)

= 0.92

Diagnostic Inference

W	S	R	Prob.
T	T	T	0.076
T	T	F	0.108
T	F	T	0.288
T	F	F	0.048
F	T	T	0.004
F	T	F	0.012
F	F	T	0.032
F	F	F	0.432

If the grass is wet, what is the probability that the sprinkler was on?

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

= (0.076 + 0.108) / (0.076 + 0.108 + 0.288 + 0.048)

= 0.35

Diagnostic Inference

W	S	R	Prob.
T	T	T	0.076
T	T	F	0.108
T	F	T	0.288
T	F	F	0.048
F	T	T	0.004
F	T	F	0.012
F	F	T	0.032
F	F	F	0.432

If the grass is wet and we know it rained, what is the probability that the sprinkler was on?

P(S|W ^ R) = P(W ^ S ^ R) / P(W ^ R)

= 0.076 / (0.076 + 0.288)

= 0.21

Explaining away

$P(S \mid W \wedge R) = 0.21 < 0.35 = P(S \mid W)$

Knowledge that one possible cause of wet grass (rain) actually occurred makes it less likely that another possible cause (sprinkler on) occurred.

The fact that it rained *explains away* the fact that the grass is wet (making other potential causes less probable than otherwise)

Conditional Independence

Perhaps Sprinkler and Rain are not actually independent. If rain is expected, it makes sense not to run the sprinkler.

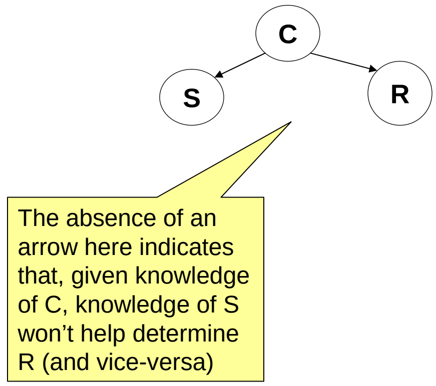
Consider adding the event
C : it was cloudy in the evening

It's reasonable to assume that S and R are conditionally independent, once the value of C is known:

$$P(S \mid C \wedge R) = P(S \mid C) \text{ and}$$
$$P(S \mid \sim C \wedge R) = P(S \mid \sim C)$$

Diagrammatic notation

"S and R are conditionally independent given C"
is represented diagrammatically as follows:



Conditional independence formalized

X and Y are conditionally independent given Z if
for all x,y,z in {T,F}:

$P(X=x \mid Y=y \wedge Z=z) = P(X=x \mid Y=y)$

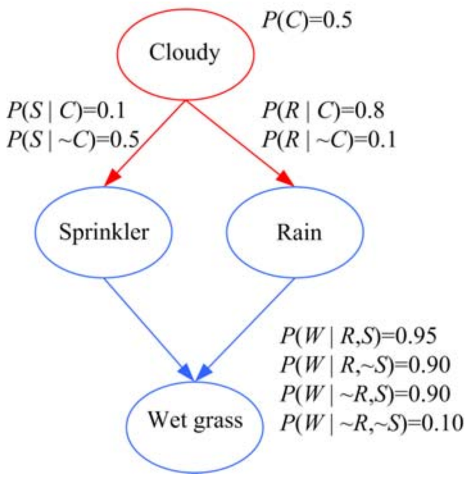
More generally:

Let S_1 and S_2 and S_3 be sets of variables.

Set-of-variables S_1 and set-of-variables S_2 are conditionally independent given S_3 if for all assignments of values to the variables in the sets,

P(S_1's assignments | S_2's assignments & S_3's assignments)=
P(S_1's assignments | S_3's assignments)

Expanded example



The absence of an arrow from Cloudy to Wet grass implies that the value of Cloudy has no direct influence on the value of Wet grass, only an indirect influence through Sprinkler and Rain.

More formally, Wet grass is conditionally independent of Cloudy, given Sprinkler and Rain.

Bayes Network

A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

Syntax:

a set of nodes, one per variable

a directed, acyclic graph (link ≈ "directly influences")

a conditional distribution for each node given its parents:

$\mathbf{P}(X_i|Parents(X_i))$

In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

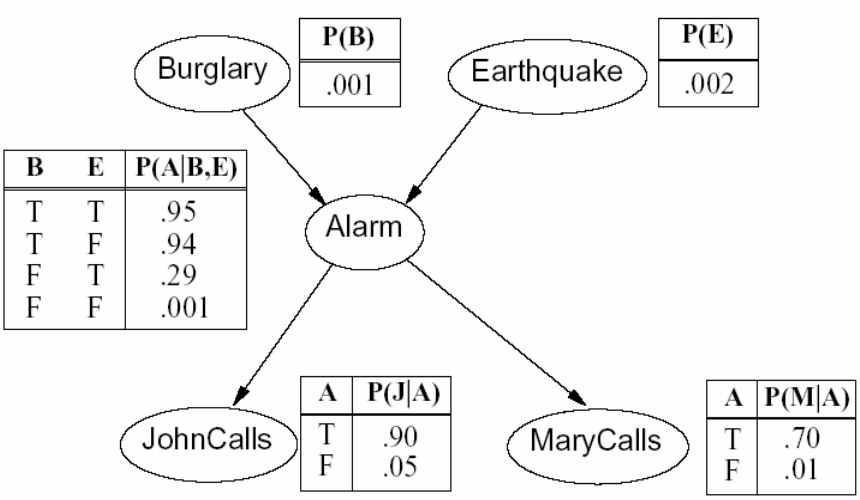
Another Example

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

Variables: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
Network topology reflects "causal" knowledge:

- 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

Corresponding Bayes Net



P(B)
.001

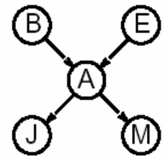
P(E)
.002

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

Compactness



A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values

Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1 - p$)

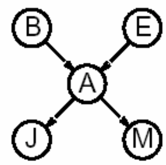
If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers

I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Bayes Nets: Slide 29

Global Semantics



"Global" semantics defines the full joint distribution as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

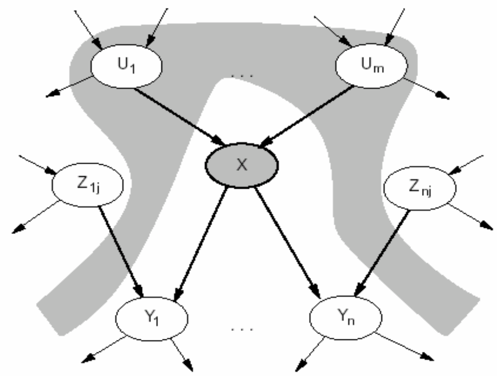
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$

Bayes Nets: Slide 30

Local Semantics

Local semantics: each node is conditionally independent of its nondescendants given its parents

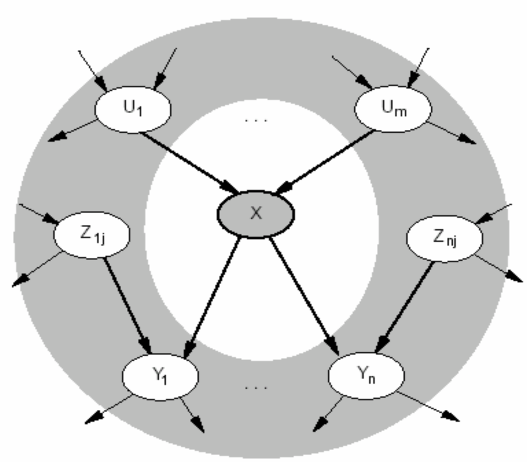


Theorem: Local semantics ⇔ global semantics

Bayes Nets: Slide 31

Markov Blanket

Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents



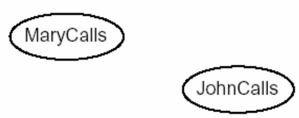
Bayes Nets: Slide 32

Building a Bayes Net

1. Choose a set of relevant variables.
2. Choose an ordering for them
3. Assume they're called X_1 .. X_n (where X_1 is the first in the ordering, X_2 is the second, etc)
4. For $i = 1$ *to* n:
   1. Add the X_i node to the network
   2. Set *Parents*(X_i) to be a minimal subset of $\{X_1 \ldots X_{i-1}\}$ such that we have conditional independence of X_i and all other members of $\{X_1 \ldots X_{i-1}\}$ given *Parents*(X_i)
   3. Define the probability table of $P(X_i = v \mid$ Assignments of *Parents*(X_i)).

Example

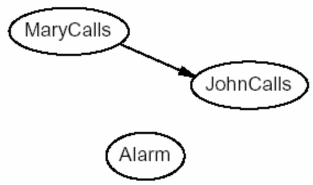
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$?

Example (cont.)

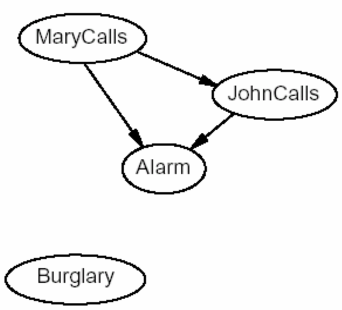
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$?

Example (cont.)

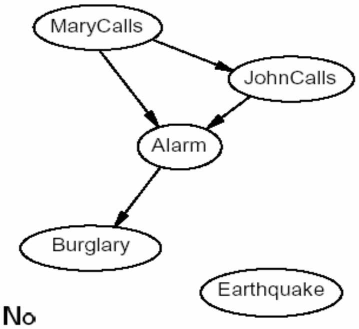
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No
$P(B|A, J, M) = P(B|A)$?
$P(B|A, J, M) = P(B)$?

Example (cont.)

Suppose we choose the ordering M, J, A, B, E

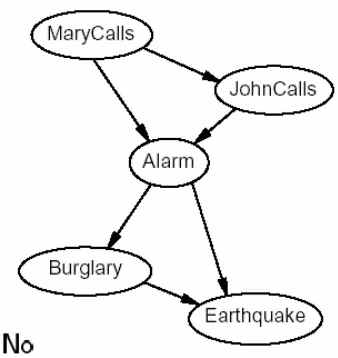


$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No
$P(B|A, J, M) = P(B|A)$? Yes
$P(B|A, J, M) = P(B)$? No
$P(E|B, A, J, M) = P(E|A)$?
$P(E|B, A, J, M) = P(E|A, B)$?

Bayes Nets: Slide 37

Example (cont.)

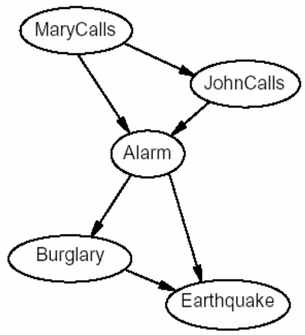
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No
$P(B|A, J, M) = P(B|A)$? Yes
$P(B|A, J, M) = P(B)$? No
$P(E|B, A, J, M) = P(E|A)$? No
$P(E|B, A, J, M) = P(E|A, B)$? Yes

Bayes Nets: Slide 38

Example (cont.)



Deciding conditional independence is hard in noncausal directions

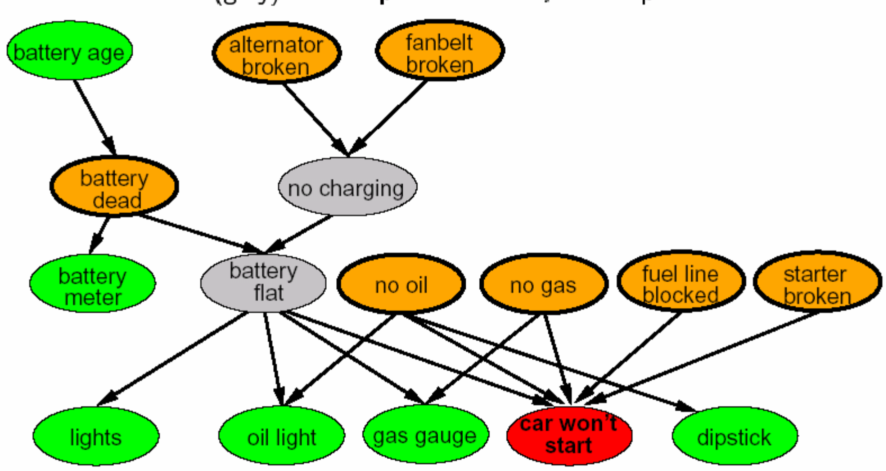
(Causal models and conditional independence seem hardwired for humans!)

Assessing conditional probabilities is hard in noncausal directions

Network is less compact: $1+2+4+2+4=13$ numbers needed

Larger Example

Initial evidence: car won't start
Testable variables (green), "broken, so fix it" variables (orange)
Hidden variables (gray) ensure sparse structure, reduce parameters



Inference in Bayes nets

- Inference tasks
  - e.g., P(Burglary | JohnCalls ^ ~MaryCalls)
- Can determine the entire joint, but that's exponential to represent explicitly
- Instead, can compute desired conditional probabilities more directly from the CPTs without representing the entire joint
  - Enumeration
  - Variable elimination

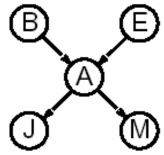
Bayes Nets: Slide 41

Enumeration

Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation

Simple query on the burglary network:

$\mathbf{P}(B|j,m)$
$= \mathbf{P}(B,j,m)/P(j,m)$
$= \alpha\mathbf{P}(B,j,m)$
$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B,e,a,j,m)$



Rewrite full joint entries using product of CPT entries:

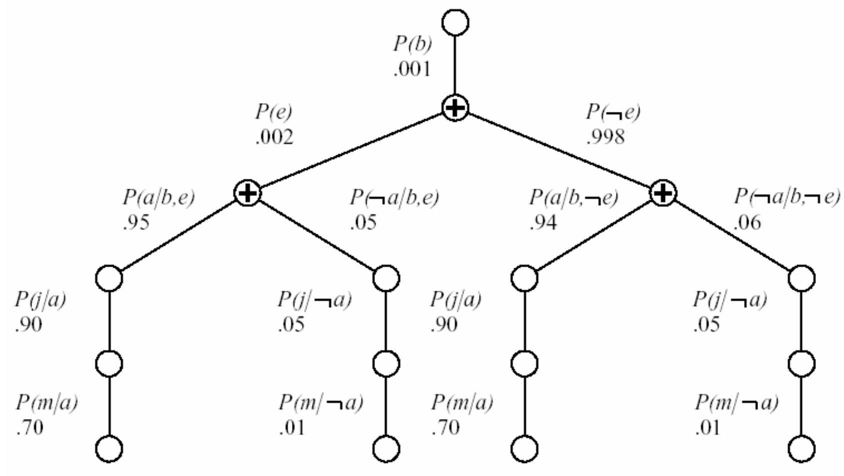
$\mathbf{P}(B|j,m)$
$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)$
$= \alpha\mathbf{P}(B)\Sigma_e P(e)\Sigma_a\mathbf{P}(a|B,e)P(j|a)P(m|a)$

Recursive depth-first enumeration: $O(n)$ space, $O(d^n)$ time

Bayes Nets: Slide 42

Evaluation tree for enumeration

Enumeration is inefficient: repeated computation
e.g., computes $P(j|a)P(m|a)$ for each value of e



Bayes Nets: Slide 43

Variable Elimination

Variable elimination: carry out summations right-to-left, storing intermediate results (factors) to avoid recomputation

$$
\begin{aligned}
\mathbf{P}(B|j,m) \\
&= \alpha \underbrace{\mathbf{P}(B)}_{B} \Sigma_e \underbrace{P(e)}_{E} \Sigma_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&= \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) P(j|a) f_M(a) \\
&= \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) f_J(a) f_M(a) \\
&= \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a f_A(a,b,e) f_J(a) f_M(a) \\
&= \alpha \mathbf{P}(B) \Sigma_e P(e) f_{\bar{A}JM}(b,e) \text{ (sum out } A\text{)} \\
&= \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (sum out } E\text{)} \\
&= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}
$$

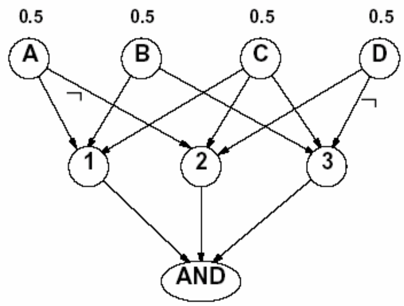
Bayes Nets: Slide 44

Complexity of exact inference

Singly connected networks (or polytrees):
- any two nodes are connected by at most one (undirected) path
- time and space cost of variable elimination are $O(d^k n)$

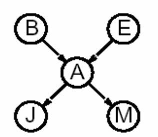
Multiply connected networks:
- can reduce 3SAT to exact inference ⇒ NP-hard
- equivalent to **counting** 3SAT models ⇒ #P-complete

1. A v B v C
2. C v D v ¬A
3. B v C v ¬D

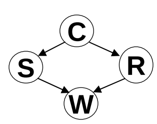


What's a polytree?

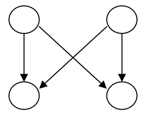
A polytree is a directed acyclic graph in which no two nodes have more than one undirected path between them.



A polytree



Not a polytree

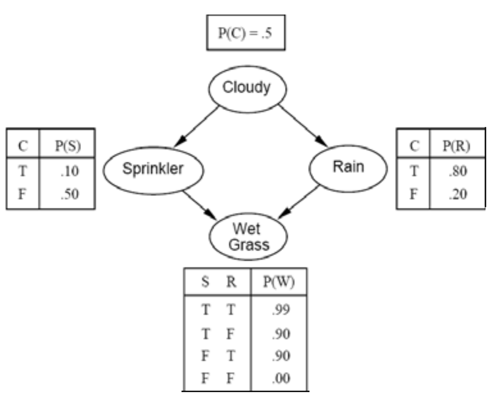


Not a polytree

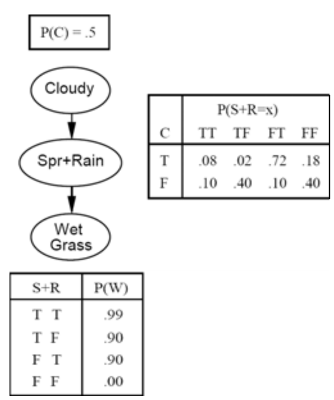
Clustering

- Used to convert a non-polytree Bayes net into a polytree Bayes net by combining certain nodes (and their CPTs)
- Resulting Bayes net allows efficient exact inference
- But the conversion process may be computationally expensive – CPTs of clustered nodes grow exponentially

Clustering example



Original Bayes net



Clustered version

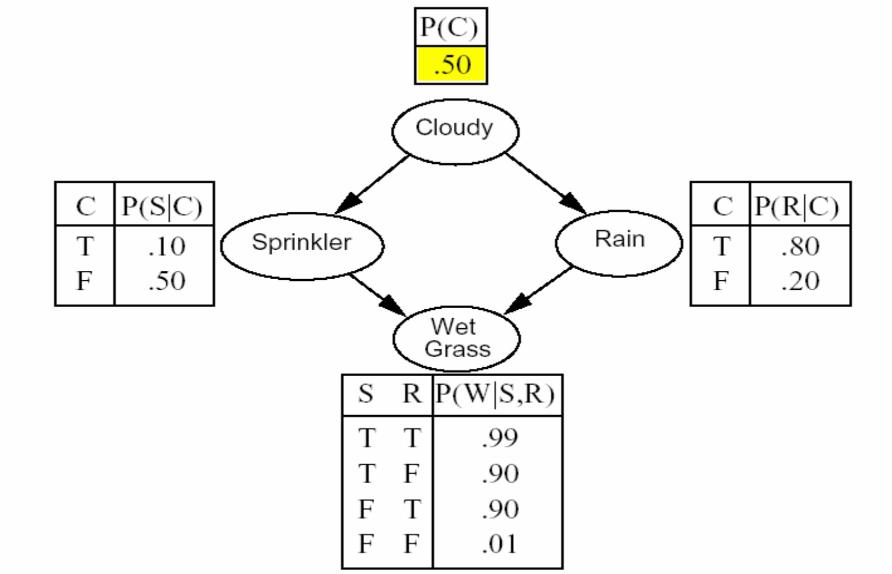
Approximate Inference

- Direct sampling methods
  - Rejection sampling
  - Likelihood weighting
- Markov chain Monte Carlo

Direct Sampling

- Efficiently generates samples from the joint
- Generate samples by going through the net sequentially from nodes having no parents to the leaves
- Use each node's CPT to generate its value randomly from the values of its parents
- Treats Bayes net as a *generative model* for the joint distribution it represents

Example: start from “empty” net



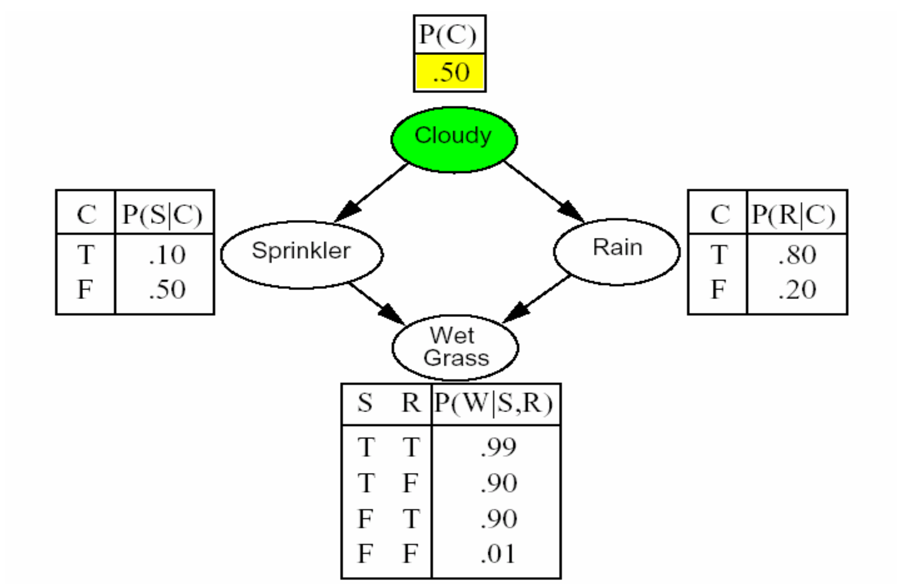
P(C)
.50

C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Example (cont.)



P(C)
.50

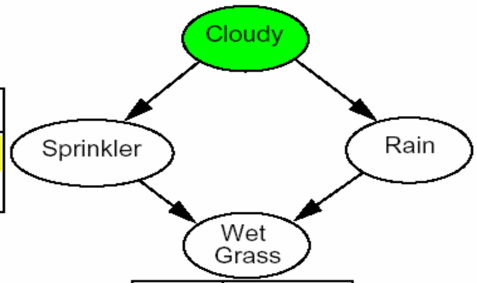
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Example (cont.)

P(C)
.50



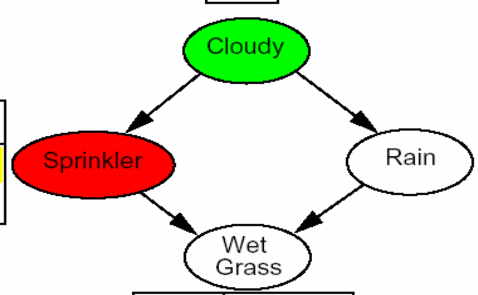
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Example (cont.)

P(C)
.50

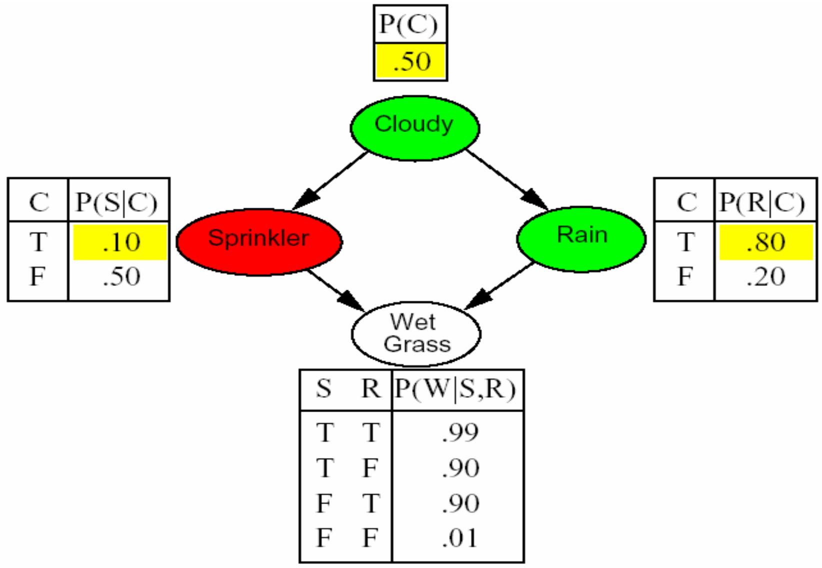


C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Example (cont.)



P(C)
.50

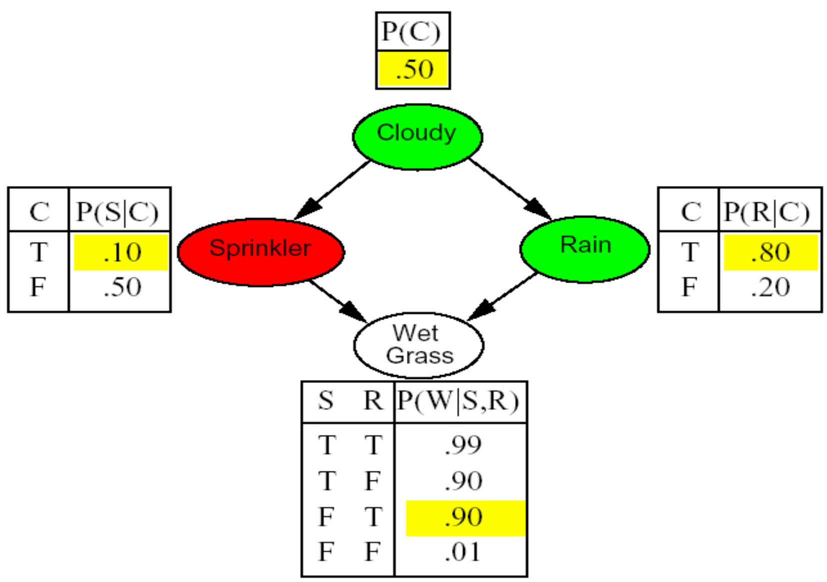
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Bayes Nets: Slide 55

Example (cont.)



P(C)
.50

C	P(S|C)
T	.10
F	.50

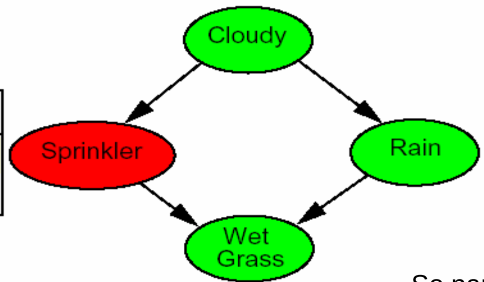
C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Bayes Nets: Slide 56

Example (cont.)

P(C)
.50



C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

So particular sample is
- C = True
- S = False
- R = True
- W = True

Rejection sampling

- Given values of *evidence variables*, discard all samples generated that are inconsistent with these values.

E.g., estimate $\mathbf{P}(Rain|Sprinkler = true)$ using 100 samples
27 samples have $Sprinkler = true$
Of these, 8 have $Rain = true$ and 19 have $Rain = false$.

$$\hat{\mathbf{P}}(Rain|Sprinkler = true) = \text{NORMALIZE}(\langle 8, 19 \rangle) = \langle 0.296, 0.704 \rangle$$

- Problem is that with a lot of evidence variables or unlikely values for them, most of the samples will be discarded.

Likelihood weighting

Imagine we're part way through our simulation

One of our evidence variables X_i is constrained to equal v

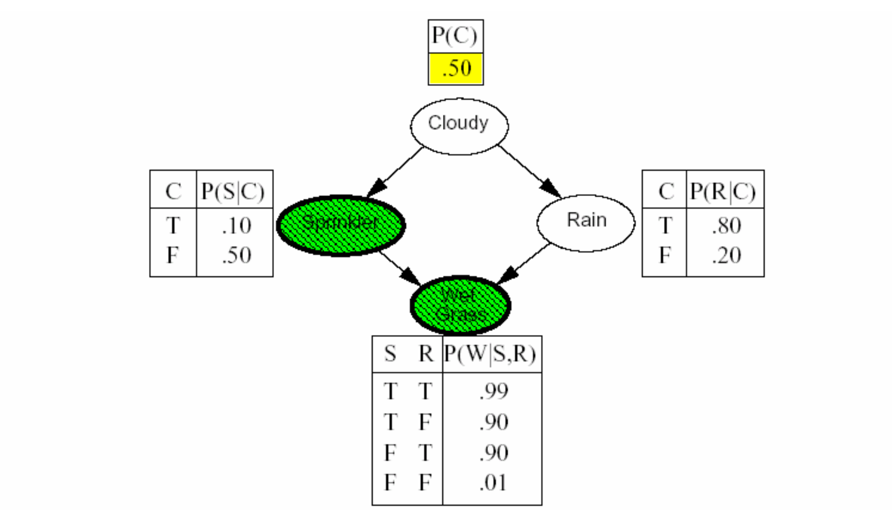
We're just about to generate a value for X_i at random. Given the values assigned to the parents, the CPT for X_i tells us that $P(X_i = v \mid \text{parents}) = p$.

Now we can anticipate what will happen if we randomly generate a value for X_i :

- we'll generate X_i = v fraction *p* of the time and proceed
- we'll generate a different value fraction 1-*p* of the time, and the simulation will be wasted.

Instead, always generate X_i = v, but weight the answer by weight *p* to compensate.

Example



P(C)
.50

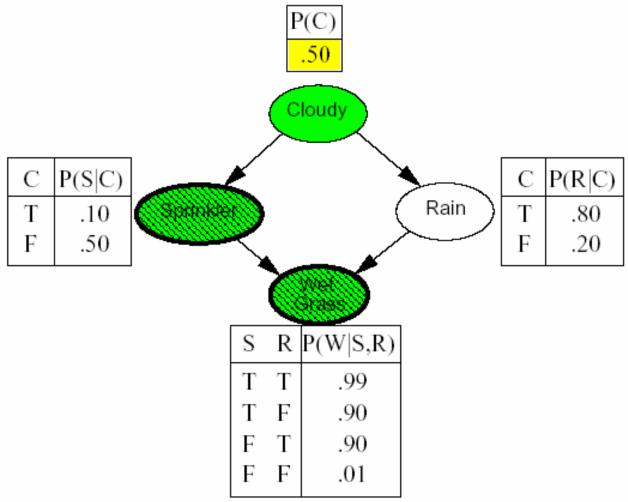
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0$

Example (cont.)



P(C)
.50

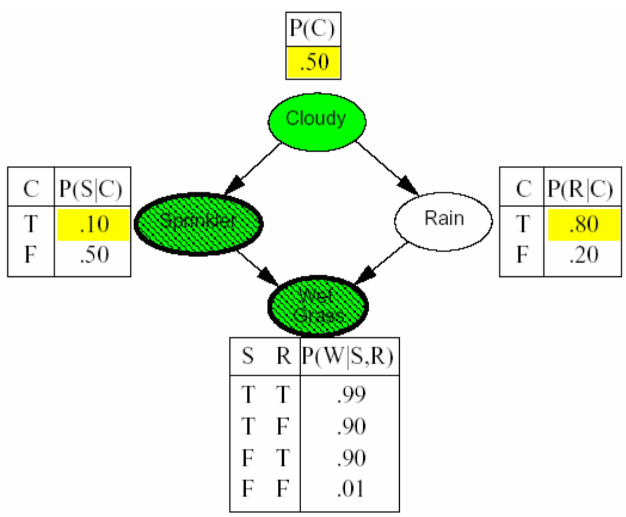
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0$

Example (cont.)



P(C)
.50

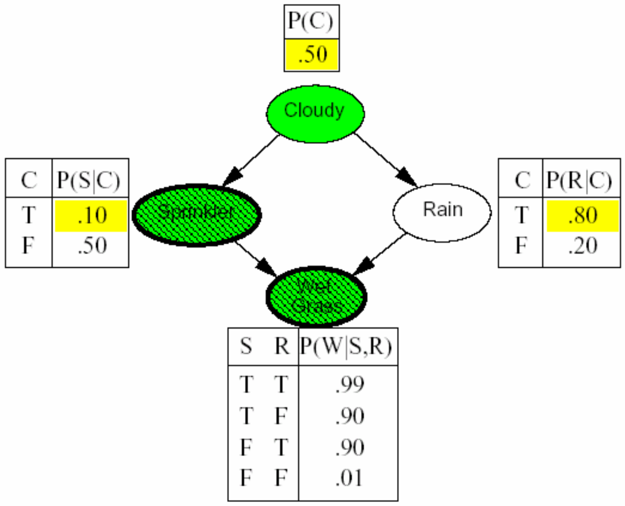
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0$

Example (cont.)



P(C)
.50

C	P(S|C)
T	.10
F	.50

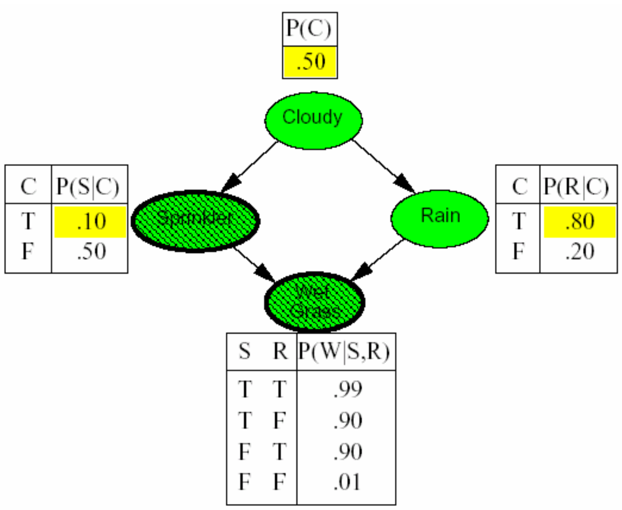
C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0 \times 0.1$

Bayes Nets: Slide 63

Example (cont.)



P(C)
.50

C	P(S|C)
T	.10
F	.50

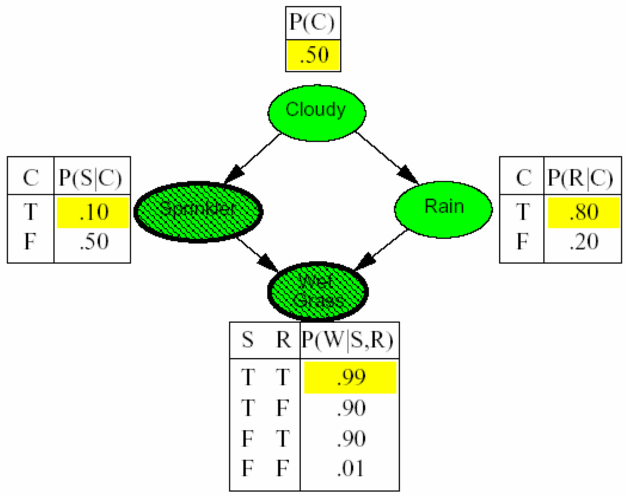
C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0 \times 0.1$

Bayes Nets: Slide 64

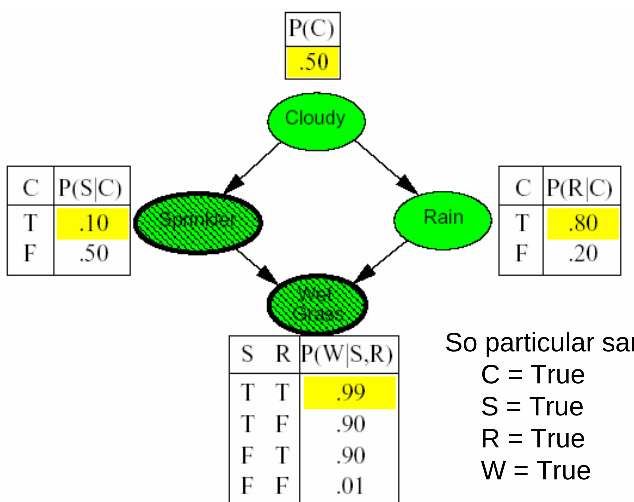
Example (cont.)



$w = 1.0 \times 0.1$

Bayes Nets: Slide 65

Example (cont.)



So particular sample is

- C = True
- S = True
- R = True
- W = True

with weight of 0.099

$w = 1.0 \times 0.1 \times 0.99 = 0.099$

Bayes Nets: Slide 66

Likelihood weighting vs. rejection sampling

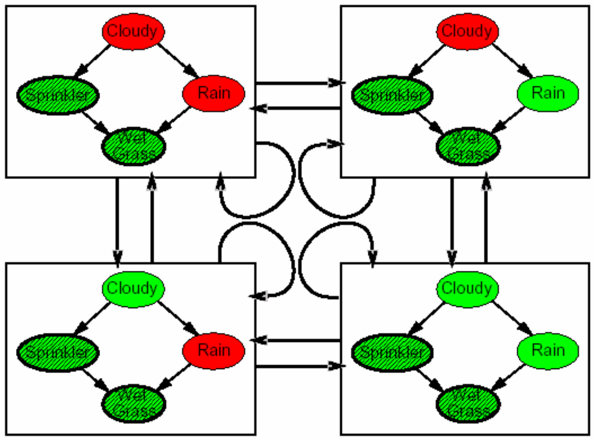
- Both generate consistent estimates of the joint distribution conditioned on the values of the evidence variables
- Likelihood weighting converges faster to the correct probabilities
- But even likelihood weighting degrades with many evidence variables because a few samples will have nearly all the total weight

Markov Chain Monte Carlo (MCMC)

- “State” of network = current assignment to all variables
- Generate next state by sampling one variable given its Markov blanket
- Sample each variable in turn, keeping evidence fixed

Example

With $Sprinkler = true, WetGrass = true$, there are four states:



Wander about for a while, average what you see

Bayes Nets: Slide 69

Example (cont.)

Estimate $\mathbf{P}(Rain|Sprinkler = true, WetGrass = true)$

Sample $Cloudy$ or $Rain$ given its Markov blanket, repeat.
Count number of times $Rain$ is true and false in the samples.

E.g., visit 100 states
31 have $Rain = true$, 69 have $Rain = false$

$\hat{\mathbf{P}}(Rain|Sprinkler = true, WetGrass = true)$
$= \text{NORMALIZE}(\langle 31, 69 \rangle) = \langle 0.31, 0.69 \rangle$

Theorem: chain approaches stationary distribution:
long-run fraction of time spent in each state is exactly proportional to its posterior probability

Bayes Nets: Slide 70

Case Study I

Pathfinder system. (Heckerman 1991, Probabilistic Similarity Networks, MIT Press, Cambridge MA).

- Diagnostic system for lymph-node diseases.
- 60 diseases and 100 symptoms and test-results.
- 14,000 probabilities
- Expert consulted to make net.
  - 8 hours to determine variables.
  - 35 hours for net topology.
  - 40 hours for probability table values.
- Apparently, the experts found it quite easy to invent the causal links and probabilities.
- Pathfinder is now outperforming the world experts in diagnosis. Being extended to several dozen other medical domains.

What you should know

- The meanings and importance of independence and conditional independence
- The definition of a Bayes net
- That a Bayes net is essentially just a potentially concise way of representing a joint distribution
- That there are methods for performing arbitrary inferences in a Bayes net
  - Complete but potentially exponential method
  - More efficient but only approximate stochastic simulation methods