Bayes Nets for representing and reasoning about uncertainty

Andrew W. Moore
Associate Professor
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

 Oct 15th, 2001

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 built Joint Distributions in manageable chunks
- Discover there's still a lurking problem…
- …Start to solve that problem

Why this matters

- In Andrew's opinion, the most important technology in the Machine Learning / AI field to have emerged in the last 10 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

- In Andrew's opinion, the most important technology in the Machine Learning / AI field to have emerged in the last 10 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

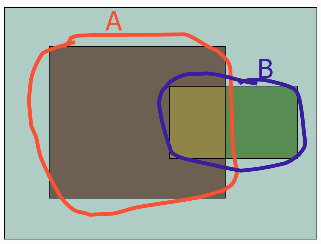
Active Data Collection

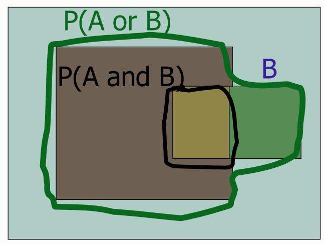
Inference

Anomaly Detection

Remember the Axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)





Simple addition and subtraction

Multivalued Random Variables

- Suppose A can take on more than 2 values
- A is a *random variable with arity k* if it can take on exactly one value out of $\{v_1, v_2, .. v_k\}$
- Thus…

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$
$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

Definition of Conditional Probability

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\, P(B)$$

Bayes Rule

$$P(B|A) = \frac{P(A \wedge B)}{P(A)} = \frac{P(A|B)\, P(B)}{P(A)}$$

This is Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

The Joint Distribution

Example: Boolean variables A, B, C

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).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

Example: Boolean variables A, B, C

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.

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

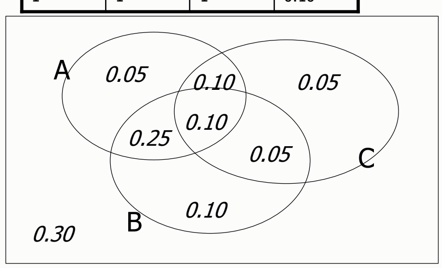
The Joint Distribution

Example: Boolean variables A, B, C

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.

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



Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

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 two events:

- M: Manuela teaches the class (otherwise it's Andrew)
- S: It is sunny

The joint p.d.f. for these events contain four entries.

If we want to build the joint p.d.f. we'll have to invent those four numbers. OR WILL WE??

- We don't have to specify with bottom level conjunctive events such as P(~M^S) IF…
- …instead it may sometimes be more convenient for us to specify things like: P(M), P(S).

But just P(M) and P(S) don't derive the joint distribution. So you can't answer all questions.

Using fewer numbers

Suppose there are two events:

- M: Manuela teaches the class (otherwise it's Andrew)
- S: It is sunny

The joint p.d.f. for these events contain four entries.

If we want to build the joint p.d.f. we'll have to invent those four numbers. OR WILL WE??

- We don't have to specify with bottom level conjunctive events such as P(~M^S) IF…
- …i…d it may sometimes be more convenient for us to … like: P(M), P(S).

But just P(M) … derive the joint distribution. So you can't answer …

What extra assumption can you make?

Independence

"The sunshine levels do not depend on and do not influence who is teaching."

This can be specified very simply:

P(S | M) = P(S)

This is a powerful statement!

It required extra domain knowledge. A different kind of knowledge than numerical probabilities. It needed an understanding of causation.

Independence

From P(S | M) = P(S), the rules of probability imply: (*can you prove these?*)

- P(~S | M) = P(~S)
- P(M | S) = P(M)
- P(M ^ S) = P(M) P(S)
- P(~M ^ S) = P(~M) P(S), (PM^~S) = P(M)P(~S), P(~M^~S) = P(~M)P(~S)

Independence

From P(S | M) = P(S), the rules of probability imply: (*can you prove these?*)

- P(~
- P(M
- P(M

And in general:

P(M=u ^ S=v) = P(M=u) P(S=v)

for each of the four combinations of

u=True/False

v=True/False

- P(~M ^ S) = P(~M) P(S), (PM^~S) = P(M)P(~S),
  P(~M^~S) = P(~M)P(~S)

Independence

We've stated:

P(M) = 0.6
P(S) = 0.3
P(S | M) = P(S)

From these statements, we can derive the full joint pdf.

M	S	Prob
T	T	
T	F	
F	T	
F	F	

And since we now have the joint pdf, we can make any queries we like.

A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

Let's begin with writing down knowledge we're happy about:

$P(S \mid M) = P(S)$, $P(S) = 0.3$, $P(M) = 0.6$

Lateness is not independent of the weather and is not independent of the lecturer.

A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

Let's begin with writing down knowledge we're happy about:

P(S | M) = P(S), P(S) = 0.3, P(M) = 0.6

Lateness is not independent of the weather and is not independent of the lecturer.

We already know the Joint of S and M, so all we need now is

P(L | S=u, M=v)

in the 4 cases of u/v = True/False.

A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

P(S | M) = P(S)
P(S) = 0.3
P(M) = 0.6

P(L | M ^ S) = 0.05
P(L | M ^ ~S) = 0.1
P(L | ~M ^ S) = 0.1
P(L | ~M ^ ~S) = 0.2

Now we can derive a full joint p.d.f. with a "mere" six numbers instead of seven*

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

A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

P(S | M) = P(S)
P(S) = 0.3
P(M) = 0.6

P(L | M ^ S) = 0.05
P(L | M ^ ~S) = 0.1
P(L | ~M ^ S) = 0.1
P(L | ~M ^ ~S) = 0.2

Question: Express

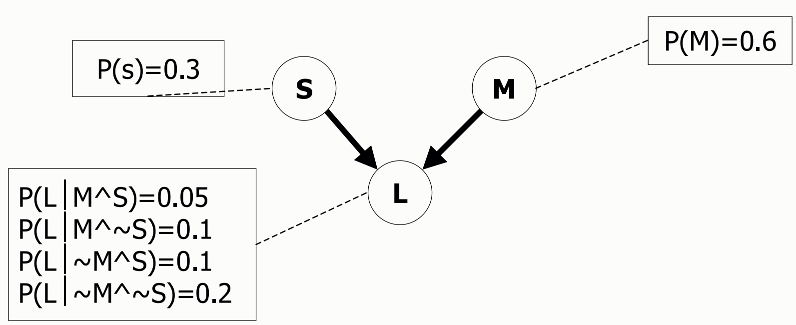
P(L=x ^ M=y ^ S=z)

in terms that only need the above expressions, where *x*,*y* and *z* may each be True or False.

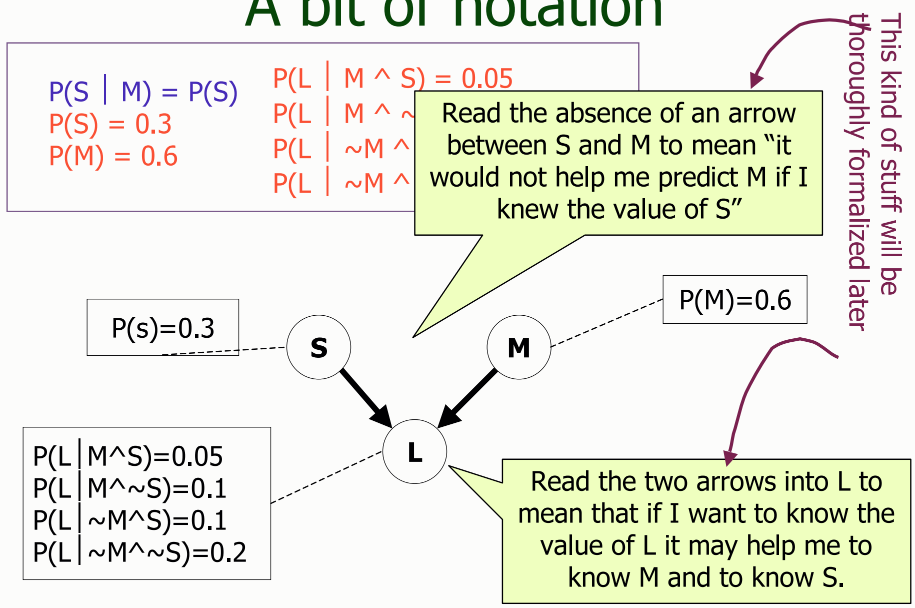
A bit of notation

P(S | M) = P(S)
P(S) = 0.3
P(M) = 0.6

P(L | M ^ S) = 0.05
P(L | M ^ ~S) = 0.1
P(L | ~M ^ S) = 0.1
P(L | ~M ^ ~S) = 0.2



A bit of notation



 Bayes Nets: Slide 31

An even cuter trick

Suppose we have these three events:

- M : Lecture taught by Manuela
- L : Lecturer arrives late
- R : Lecture concerns robots

Suppose:

- Andrew has a higher chance of being late than Manuela.
- Andrew has a higher chance of giving robotics lectures.

What kind of independence can we find?

How about:

- P(L | M) = P(L) ?
- P(R | M) = P(R) ?
- P(L | R) = P(L) ?

 Bayes Nets: Slide 32

Conditional independence

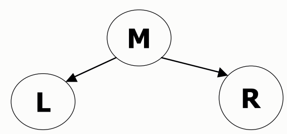
Once you know who the lecturer is, then whether they arrive late doesn't affect whether the lecture concerns robots.

$$P(R \mid M,L) = P(R \mid M) \text{ and}$$
$$P(R \mid \sim M,L) = P(R \mid \sim M)$$

We express this in the following way:

"R and L are conditionally independent given M"

..which is also notated by the following diagram.



Given knowledge of M, knowing anything else in the diagram won't help us with L, etc.

Conditional Independence formalized

R and L are conditionally independent given M if for all x,y,z in {T,F}:

$$P(R=x \mid M=y \wedge L=z) = P(R=x \mid M=y)$$

More generally:

Let S1 and S2 and S3 be sets of variables.

Set-of-variables S1 and set-of-variables S2 are conditionally independent given S3 if for all assignments of values to the variables in the sets,

$P(S_1\text{'s assignments} \mid S_2\text{'s assignments} \;\&\; S_3\text{'s assignments}) =$
$P(S1\text{'s assignments} \mid S3\text{'s assignments})$

Example:

R and L are
for all x,y,z
P(R=

More gener
Let S1 and S2 and S3 be sets of va

> "Shoe-size is conditionally independent of Glove-size given height weight and age"
> means
> forall s,g,h,w,a
> P(ShoeSize=s|Height=h,Weight=w,Age=a)
> =
> P(ShoeSize=s|Height=h,Weight=w,Age=a,GloveSize=g)

Set-of-variables S1 and set-of-variables S2 are conditionally independent given S3 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(S1's assignments | S3's assignments)

Example:

R and L are
for all x,y,z
P(R=

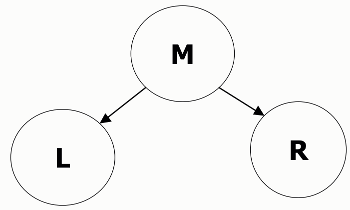
More gener
Let S1 and S2 and S3 be sets of va

> "Shoe-size is conditionally independent of Glove-size given height weight and age"
> does not mean
> forall s,g,h
> P(ShoeSize=s|Height=h)
> =
> P(ShoeSize=s|Height=h, GloveSize=g)

Set-of-variables S1 and set-of-variables S2 are conditionally independent given S3 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(S1's assignments | S3's assignments)

Conditional independence

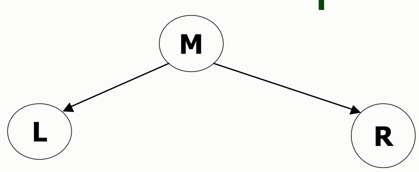


We can write down P(M). And then, since we know L is only directly influenced by M, we can write down the values of $P(L|M)$ and $P(L|\sim M)$ and know we've fully specified L's behavior. Ditto for R.

$P(M) = 0.6$
$P(L \mid M) = 0.085$
$P(L \mid \sim M) = 0.17$
$P(R \mid M) = 0.3$
$P(R \mid \sim M) = 0.6$

'R and L conditionally independent given M'

Conditional independence



$P(M) = 0.6$
$P(L \mid M) = 0.085$
$P(L \mid \sim M) = 0.17$
$P(R \mid M) = 0.3$
$P(R \mid \sim M) = 0.6$

Conditional Independence:

$P(R|M,L) = P(R|M)$,

$P(R|\sim M,L) = P(R|\sim M)$

Again, we can obtain any member of the Joint prob dist that we desire:

$P(L=x \wedge R=y \wedge M=z) =$

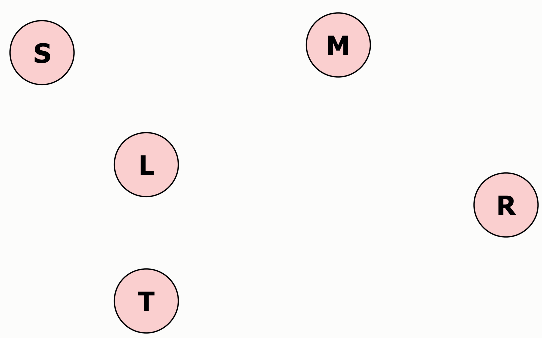
Assume five variables

T: The lecture started by 10:35
L: The lecturer arrives late
R: The lecture concerns robots
M: The lecturer is Manuela
S: It is sunny

- T only directly influenced by L (i.e. T is conditionally independent of R,M,S given L)
- L only directly influenced by M and S (i.e. L is conditionally independent of R given M & S)
- R only directly influenced by M (i.e. R is conditionally independent of L,S, given M)
- M and S are independent

Making a Bayes net

T: The lecture started by 10:35
L: The lecturer arrives late
R: The lecture concerns robots
M: The lecturer is Manuela
S: It is sunny

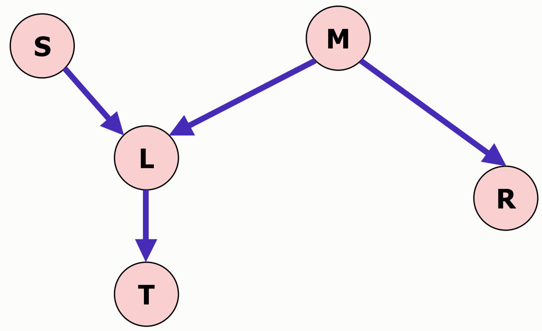


Step One: add variables.

- Just choose the variables you'd like to be included in the net.

Making a Bayes net

T: The lecture started by 10:35
L: The lecturer arrives late
R: The lecture concerns robots
M: The lecturer is Manuela
S: It is sunny

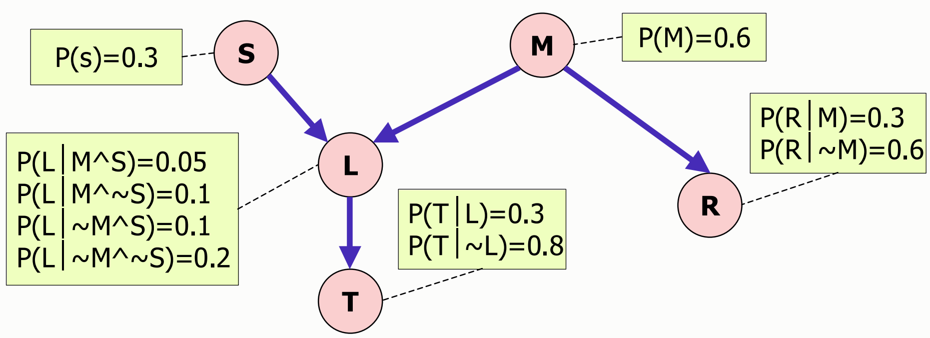


Step Two: add links.

- The link structure must be acyclic.
- If node X is given parents $Q_1,Q_2,..Q_n$ you are promising that any variable that's a non-descendant of X is conditionally independent of X given $\{Q_1,Q_2,..Q_n\}$

Making a Bayes net

T: The lecture started by 10:35
L: The lecturer arrives late
R: The lecture concerns robots
M: The lecturer is Manuela
S: It is sunny

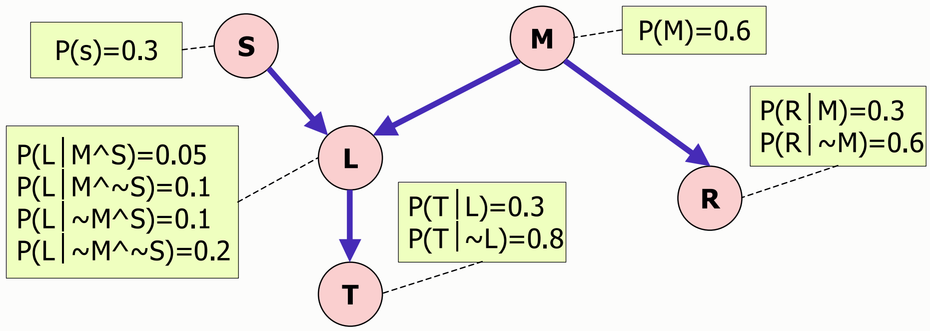


Step Three: add a probability table for each node.

- The table for node X must list P(X|Parent Values) for each possible combination of parent values

Making a Bayes net

T: The lecture started by 10:35
L: The lecturer arrives late
R: The lecture concerns robots
M: The lecturer is Manuela
S: It is sunny



- Two unconnected variables may still be correlated
- Each node is conditionally independent of all non-descendants in the tree, given its parents.
- You can deduce many other conditional independence relations from a Bayes net. See the next lecture.

Bayes Nets Formalized

A Bayes net (also called a belief network) is an augmented directed acyclic graph, represented by the pair V , E where:

- V is a set of vertices.
- E is a set of directed edges joining vertices. No loops of any length are allowed.

Each vertex in V contains the following information:

- The name of a random variable
- A probability distribution table indicating how the probability of this variable’s values depends on all possible combinations of parental values.

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_m (where X_1 is the first in the ordering, X_1 is the second, etc)
4. For $i = 1$ *to* m:
   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 $P(X_i = k \mid \text{Assignments of } Parents(X_i))$.

Example Bayes Net Building

Suppose we're building a nuclear power station.
There are the following random variables:

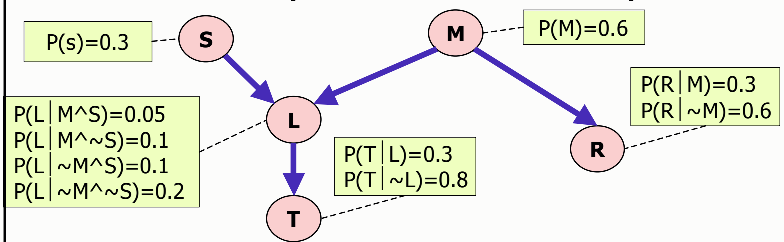
GRL : Gauge Reads Low.
CTL : Core temperature is low.
FG : Gauge is faulty.
FA : Alarm is faulty
AS : Alarm sounds

- If alarm working properly, the alarm is meant to sound if the gauge stops reading a low temp.
- If gauge working properly, the gauge is meant to read the temp of the core.

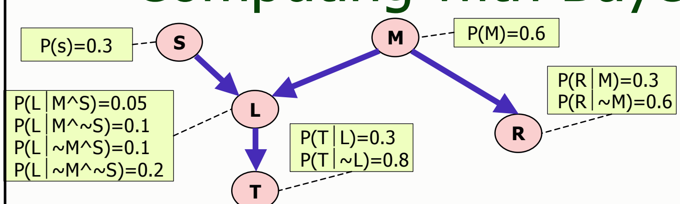
Computing a Joint Entry

How to compute an entry in a joint distribution?

E.G: What is P(S ^ ~M ^ L ~R ^ T)?



Computing with Bayes Net



P(T ^ ~R ^ L ^ ~M ^ S) =
P(T | ~R ^ L ^ ~M ^ S) * P(~R ^ L ^ ~M ^ S) =
P(T | L) * P(~R ^ L ^ ~M ^ S) =
P(T | L) * P(~R | L ^ ~M ^ S) * P(L^~M^S) =
P(T | L) * P(~R | ~M) * P(L^~M^S) =
P(T | L) * P(~R | ~M) * P(L|~M^S)*P(~M^S) =
P(T | L) * P(~R | ~M) * P(L|~M^S)*P(~MS)*P(S) =
P(T | L) * P(~R | ~M) * P(L|~M^S)*P(~M)*P(S).

The general case

$P(X_1{=}x_1 \wedge X_2{=}x_2 \wedge \ldots X_{n-1}{=}x_{n-1} \wedge X_n{=}x_n) =$

$P(X_n{=}x_n \wedge X_{n-1}{=}x_{n-1} \wedge \ldots X_2{=}x_2 \wedge X_1{=}x_1) =$

$P(X_n{=}x_n \mid X_{n-1}{=}x_{n-1} \wedge \ldots X_2{=}x_2 \wedge X_1{=}x_1) * P(X_{n-1}{=}x_{n-1} \wedge \ldots X_2{=}x_2 \wedge X_1{=}x_1) =$

$P(X_n{=}x_n \mid X_{n-1}{=}x_{n-1} \wedge \ldots X_2{=}x_2 \wedge X_1{=}x_1) * P(X_{n-1}{=}x_{n-1} \mid \ldots X_2{=}x_2 \wedge X_1{=}x_1) *$

$P(X_{n-2}{=}x_{n-2} \wedge \ldots X_2{=}x_2 \wedge X_1{=}x_1) =$

:

:

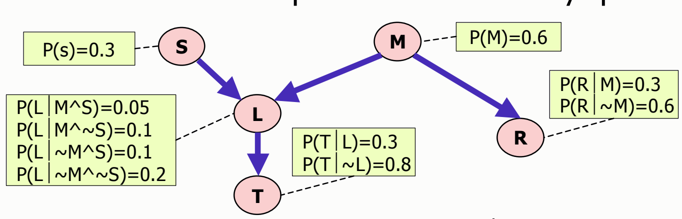
$$= \prod_{i=1}^{n} P\left((X_i = x_i) \mid ((X_{i-1} = x_{i-1}) \wedge \ldots (X_1 = x_1))\right)$$

$$= \prod_{i=1}^{n} P\left((X_i = x_i) \mid \text{Assignments of Parents}(X_i)\right)$$

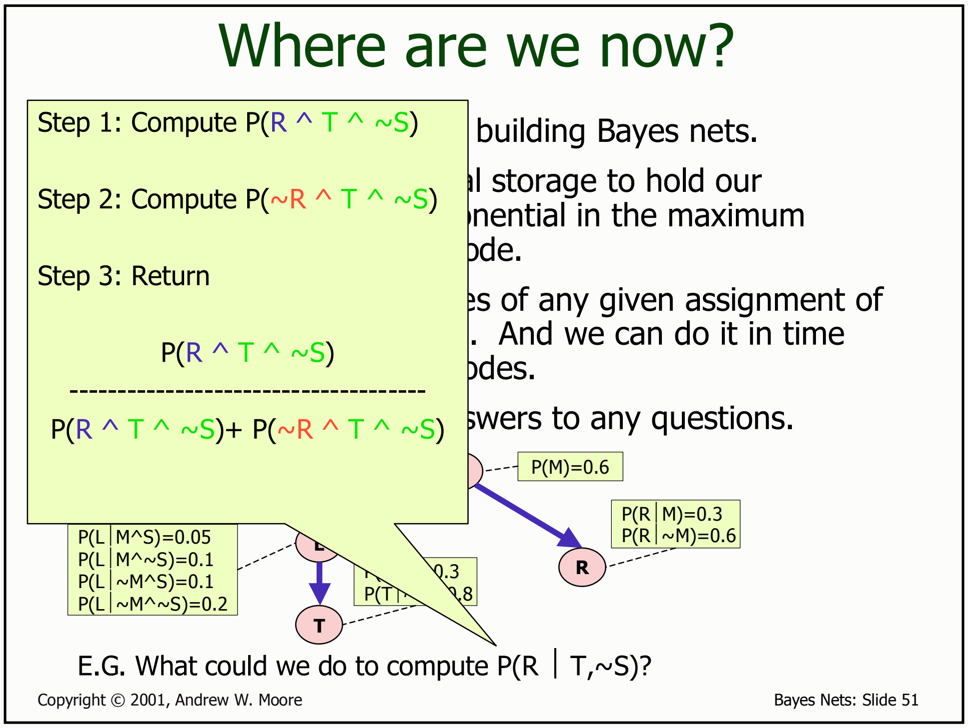
So any entry in joint pdf table can be computed. And so **any conditional probability** can be computed.

Where are we now?

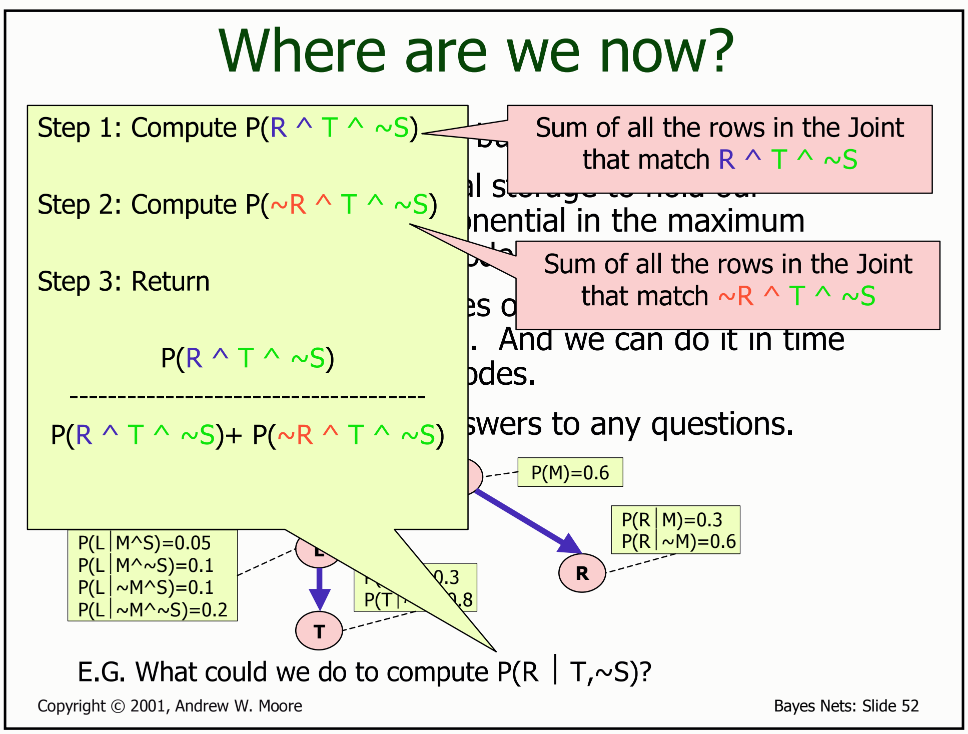
- We have a methodology for building Bayes nets.
- We don't require exponential storage to hold our probability table. Only exponential in the maximum number of parents of any node.
- We can compute probabilities of any given assignment of truth values to the variables. And we can do it in time linear with the number of nodes.
- So we can also compute answers to any questions.



E.G. What could we do to compute P(R | T,~S)?


Where are we now?
Step 1: Compute P(R ^ T ^ ~S)
Step 2: Compute P(~R ^ T ^ ~S)
Step 3: Return
P(R ^ T ^ ~S)

P(R ^ T ^ ~S)+ P(~R ^ T ^ ~S)
building Bayes nets.
l storage to hold our
nential in the maximum
de.
s of any given assignment of
. And we can do it in time
des.
wers to any questions.
P(M)=0.6
P(R|M)=0.3
P(R|~M)=0.6
P(L|M^S)=0.05
P(L|M^~S)=0.1
P(L|~M^S)=0.1
P(L|~M^~S)=0.2
R
T
E.G. What could we do to compute P(R | T,~S)?
Copyright © 2001, Andrew W. Moore
Bayes Nets: Slide 51


Where are we now?
Step 1: Compute P(R ^ T ^ ~S)
Sum of all the rows in the Joint that match R ^ T ^ ~S
Step 2: Compute P(~R ^ T ^ ~S)
Sum of all the rows in the Joint that match ~R ^ T ^ ~S
Step 3: Return
P(R ^ T ^ ~S)

P(R ^ T ^ ~S)+ P(~R ^ T ^ ~S)
nential in the maximum
. And we can do it in time
des.
wers to any questions.
P(M)=0.6
P(R|M)=0.3
P(R|~M)=0.6
P(L|M^S)=0.05
P(L|M^~S)=0.1
P(L|~M^S)=0.1
P(L|~M^~S)=0.2
R
T
E.G. What could we do to compute P(R | T,~S)?
Copyright © 2001, Andrew W. Moore
Bayes Nets: Slide 52

Where are we now?

Step 1: Compute P(R ^ T ^ ~S)

Step 2: Compute P(~R ^ T ^ ~S)

Step 3: Return

P(R ^ T ^ ~S)

P(R ^ T ^ ~S)+ P(~R ^ T ^ ~S)

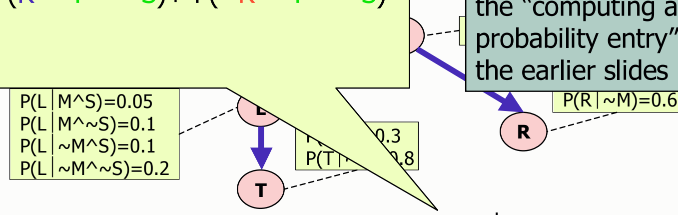
Sum of all the rows in the Joint that match R ^ T ^ ~S

4 joint computes

Sum of all the rows in the Joint that match ~R ^ T ^ ~S

4 joint computes

Each of these obtained by the "computing a joint probability entry" method of the earlier slides



E.G. What could we do to compute P(R | T,~S)?

The good news

We can do inference. We can compute any conditional probability:

P(Some variable | Some other variable values)

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{joint entries matching } E_1 \text{ and } E_2} P(\text{joint entry})}{\sum_{\text{joint entries matching } E_2} P(\text{joint entry})}$$

The good news

We can do inference. We can compute any conditional probability:

P(Some variable | Some other variable values)

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{joint entries matching } E_1 \text{ and } E_2} P(\text{joint entry})}{\sum_{\text{joint entries matching } E_2} P(\text{joint entry})}$$

Suppose you have *m* binary-valued variables in your Bayes Net and expression E_2 mentions *k* variables.

How much work is the above computation?

The sad, bad news

Conditional probabilities by enumerating all matching entries in the joint are expensive:

Exponential in the number of variables.

The sad, bad news

Conditional probabilities by enumerating all matching entries in the joint are expensive:

Exponential in the number of variables.

But perhaps there are faster ways of querying Bayes nets?

- In fact, if I ever ask you to manually do a Bayes Net inference, you'll find there are often many tricks to save you time.
- So we've just got to program our computer to do those tricks too, right?

The sad, bad news

Conditional probabilities by enumerating all matching entries in the joint are expensive:

Exponential in the number of variables.

But perhaps there are faster ways of querying Bayes nets?

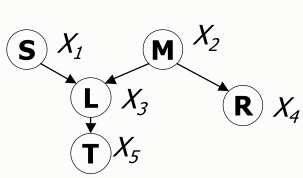
- In fact, if I ever ask you to manually do a Bayes Net inference, you'll find there are often many tricks to save you time.
- So we've just got to program our computer to do those tricks too, right?

Sadder and worse news:

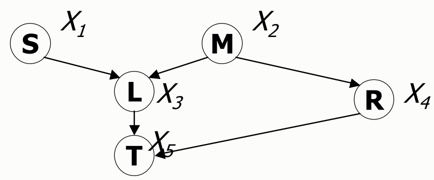
General querying of Bayes nets is NP-complete.

Bayes nets inference algorithms

A poly-tree is a directed acyclic graph in which no two nodes have more than one path between them.



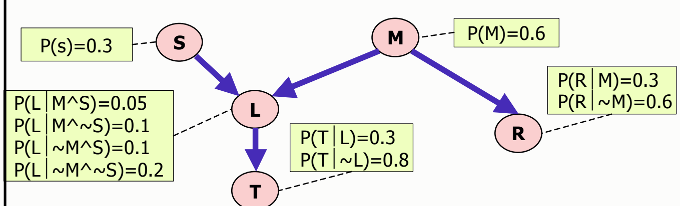
A poly tree



Not a poly tree
(but still a legal Bayes net)

- If net is a poly-tree, there is a linear-time algorithm (see a later Andrew lecture).
- The best general-case algorithms convert a general net to a poly-tree (often at huge expense) and calls the poly-tree algorithm.
- Another popular, practical approach (doesn't assume poly-tree): Stochastic Simulation.

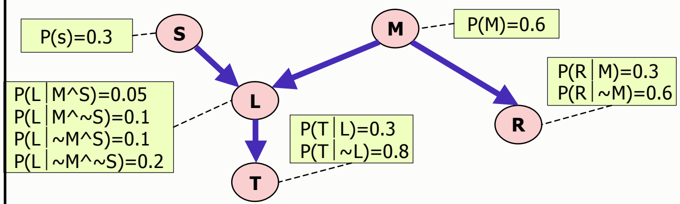
Sampling from the Joint Distribution



It's pretty easy to generate a set of variable-assignments at random with the same probability as the underlying joint distribution.

How?

Sampling from the Joint Distribution



1. Randomly choose S. S = True with prob 0.3
2. Randomly choose M. M = True with prob 0.6
3. Randomly choose L. The probability that L is true depends on the assignments of S and M. E.G. if steps 1 and 2 had produced S=True, M=False, then probability that L is true is 0.1
4. Randomly choose R. Probability depends on M.
5. Randomly choose T. Probability depends on L

A general sampling algorithm

Let's generalize the example on the previous slide to a general Bayes Net.

As in Slides 16-17 , call the variables X_1 .. X_n, where $Parents(X_i)$ must be a subset of $\{X_1 .. X_{i-1}\}$.

For $i=1$ to n:

1. Find parents, if any, of X_i. Assume $n(i)$ parents. Call them $X_{p(i,1)}, X_{p(i,2)}, \ldots X_{p(i,n(i))}$.
2. Recall the values that those parents were randomly given: $x_{p(i,1)}, x_{p(i,2)}, \ldots x_{p(i,n(i))}$.
3. Look up in the lookup-table for: $P(X_i=True \mid X_{p(i,1)}=x_{p(i,1)}, X_{p(i,2)}=x_{p(i,2)} \ldots X_{p(i,n(i))}=x_{p(i,n(i))})$
4. Randomly set $x_i=True$ according to this probability

$x_1, x_2, \ldots x_n$ are now a sample from the joint distribution of $X_1, X_2, \ldots X_n$.

Stochastic Simulation Example

Someone wants to know P(R = True | T = True ^ S = False)

We'll do lots of random samplings and count the number of occurrences of the following:

- N_c : Num. samples in which T=True and S=False.
- N_s : Num. samples in which R=True, T=True and S=False.
- N : Number of random samplings

Now if N is big enough:

N_c / N is a good estimate of *P(T=True and S=False).*

N_s / N is a good estimate of *P(R=True ,T=True , S=False)*.

P(R|T^~S) = P(R^T^~S)/P(T^~S), so N_s / N_c can be a good estimate of P(R|T^~S).

General Stochastic Simulation

Someone wants to know P(E_1 | E_2)

We'll do lots of random samplings and count the number of occurrences of the following:

- N_c : Num. samples in which E_2
- N_s : Num. samples in which E_1 and E_2
- N : Number of random samplings

Now if N is big enough:

N_c / N is a good estimate of $P(E_2)$.

N_s / N is a good estimate of $P(E_1 , E_2)$.

P(E_1 | E_2) = P(E_1^ E_2)/P(E_2), so N_s / N_c can be a good estimate of P(E_1| E_2).

Likelihood weighting

Problem with Stochastic Sampling:

With lots of constraints in E, or unlikely events in E, then most of the simulations will be thrown away, (they'll have no effect on Nc, or Ns).

Imagine we're part way through our simulation.

In E2 we have the constraint Xi = v

We're just about to generate a value for Xi at random. Given the values assigned to the parents, we see that P(Xi = v | parents) = p .

Now we know that with stochastic sampling:

- we'll generate "Xi = v" proportion p of the time, and proceed.
- And we'll generate a different value proportion 1-p of the time, and the simulation will be wasted.

Instead, always generate Xi = v, but weight the answer by weight "p" to compensate.

Likelihood weighting

Set N_c :=0, N_s :=0

1. Generate a random assignment of all variables that matches E_2. This process returns a weight w.
2. Define w to be the probability that this assignment would have been generated instead of an unmatching assignment during its generation in the original algorithm.Fact: w is a product of all likelihood factors involved in the generation.
3. *$N_c := N_c + w$*
4. If our sample matches E_1 then $N_s := N_s$ + w
5. Go to 2

Again, N_s / N_c estimates P(E_1 | E_2)

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.

Questions

- What are the strengths of probabilistic networks compared with propositional logic?
- What are the weaknesses of probabilistic networks compared with propositional logic?
- What are the strengths of probabilistic networks compared with predicate logic?
- What are the weaknesses of probabilistic networks compared with predicate logic?
- (How) could predicate logic and probabilistic networks be combined?

What you should know

- The meanings and importance of independence and conditional independence.
- The definition of a Bayes net.
- Computing probabilities of assignments of variables (i.e. members of the joint p.d.f.) with a Bayes net.
- The slow (exponential) method for computing arbitrary, conditional probabilities.
- The stochastic simulation method and likelihood weighting.