Fundamentals of AI

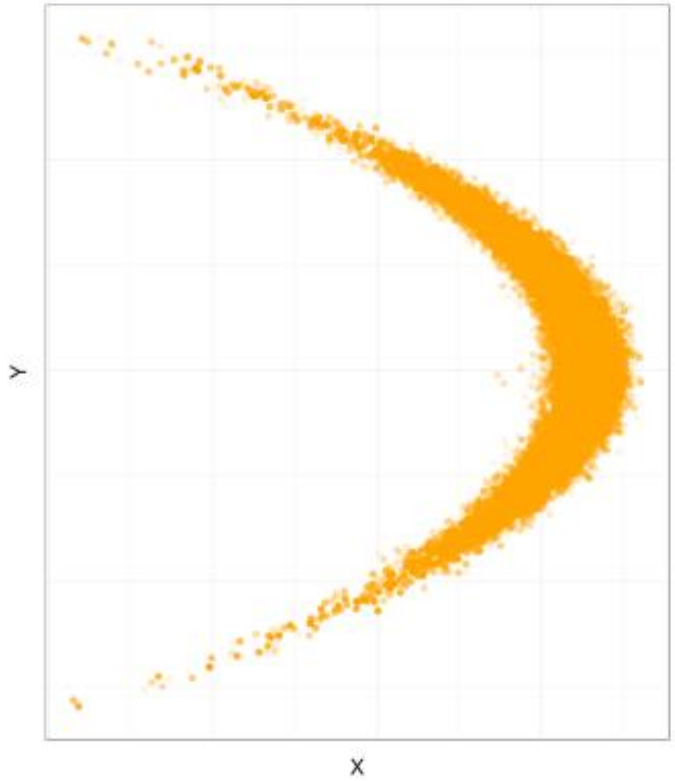
Introduction and the most basic concepts

Conditional independence, Naïve Bayes and Bayesian Networks

Joint Probability Distribution

- Probability of any combination of features to happen

‘Banana-shaped probability distribution’



Probability density function (PDF)

$$f(x,y) = \exp\left(-\frac{x^2}{200} - \frac{1}{2}(y + Bx^2 - 100B)^2\right)$$

Conditional Probability

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

Bayes rule

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

The story of Andrew (Moore) and Manuela

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.

Event ***M***

True

False

Event ***S***

True

False

Independence

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

This can be specified very simply:

$$P(S \mid M) = P(S)$$

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 $P(S \mid M) = P(S)$, the rules of probability imply: (*can you prove these?*)

- $P(\sim S \mid M) = P(\sim S)$
- $P(M \mid S) = P(M)$
- $P(M \wedge S) = P(M)\,P(S)$
- $P(\sim M \wedge S) = P(\sim M)\,P(S)$, $(PM\wedge\sim S) = P(M)P(\sim S)$, $P(\sim M\wedge\sim S) = P(\sim M)P(\sim S)$

Independence

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

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

And in general:

$$P(M=u \wedge S=v) = P(M=u)\,P(S=v)$$

for each of the four combinations of

u=True/False

v=True/False

Independence

We've stated:

$P(M) = 0.6$

$P(S) = 0.3$

$P(S \mid M) = P(S)$

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

M	S	Prob
T	T	0.18
T	F	0.42
F	T	0.12
F	F	0.28

Most probable

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.

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.

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

$P(L \mid S=u, M=v)$

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

Event ***M***

True

False

Event ***S***

True

False

Event ***L***

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.

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

$P(L \mid M \wedge S) = 0.05$
$P(L \mid M \wedge \sim S) = 0.1$
$P(L \mid \sim M \wedge S) = 0.1$
$P(L \mid \sim M \wedge \sim S) = 0.2$

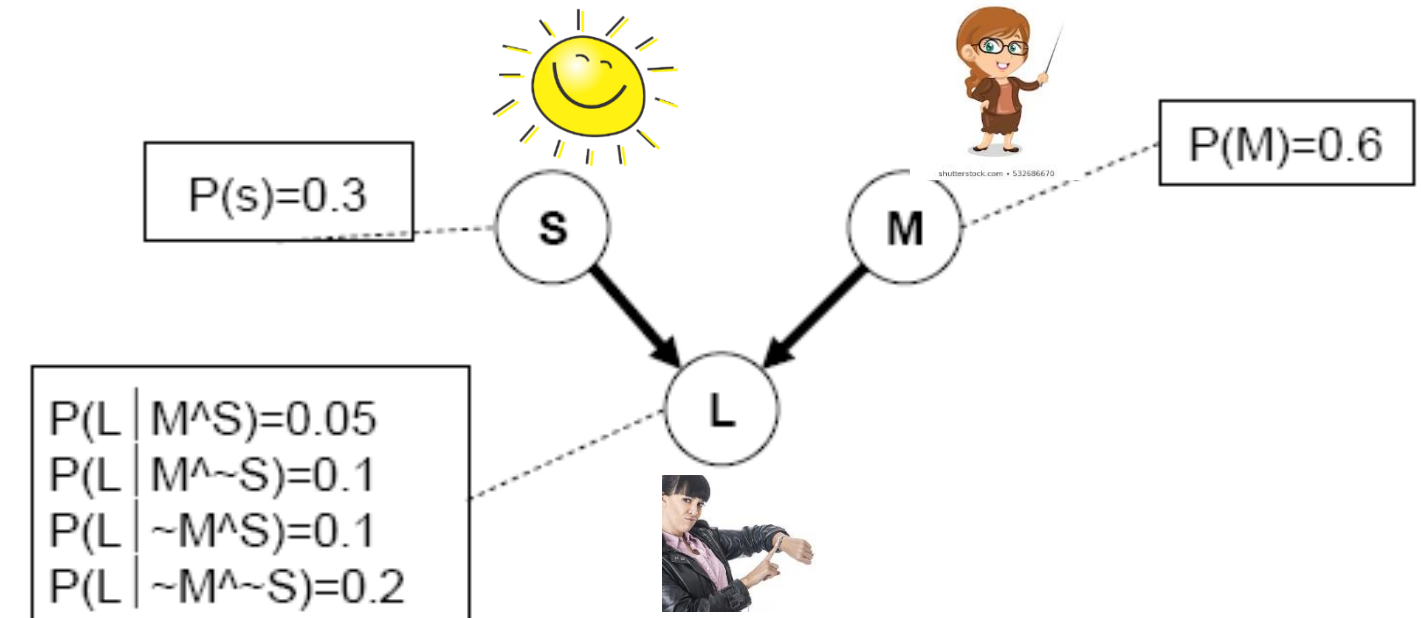
Now we can derive a full joint p.d.f. with a "mere" six numbers instead of seven*

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

A bit of notation

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

$P(L \mid M \wedge S) = 0.05$
$P(L \mid M \wedge \sim S) = 0.1$
$P(L \mid \sim M \wedge S) = 0.1$
$P(L \mid \sim M \wedge \sim S) = 0.2$

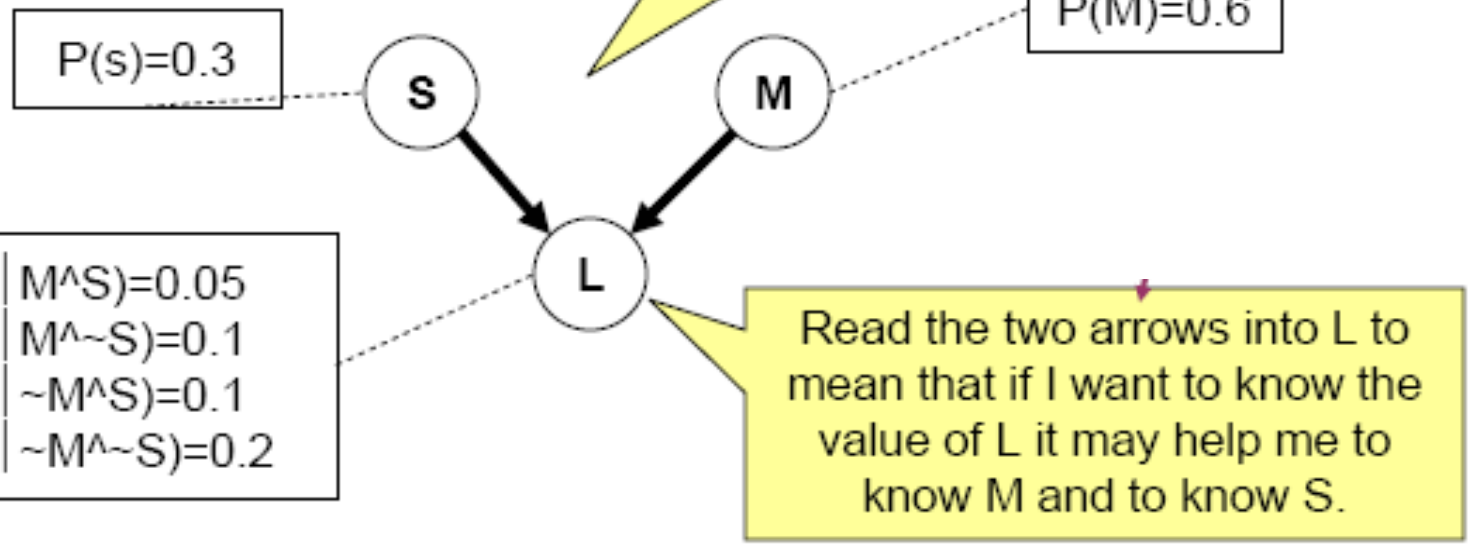


A bit of notation

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

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

Read the absence of an arrow between S and M to mean "it would not help me predict M if I knew the value of S"



Read the two arrows into L to mean that if I want to know the value of L it may help me to know M and to know S.

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:

- $P(L \mid M) = P(L)$?
- $P(R \mid M) = P(R)$?
- $P(L \mid R) = P(L)$?

Event ***M***

True

False

Event ***S***

True

False

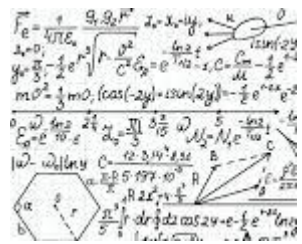
Event ***L***

True

False

Event ***R***

True



False

Conditional independence

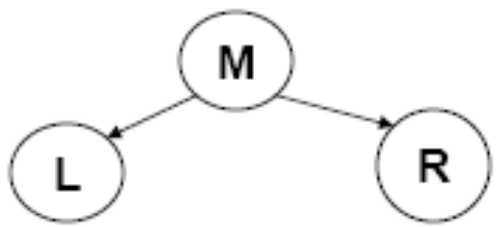
Once you know who the lecturer is, then whether they arrive late doesn't affect whether the lecture concerns robots.

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

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'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"

means

forall s,g,h,w,a

P(ShoeSize=s|Height=h,Weight=w,Age=a)

=

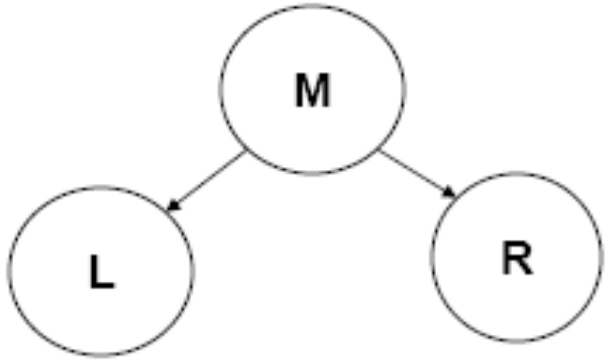
P(ShoeSize=s|Height=h,Weight=w,Age=a,GloveSize=g)

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,

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

$P(S1\text{'s assignments} \mid S3\text{'s assignments})$

Conditional independence

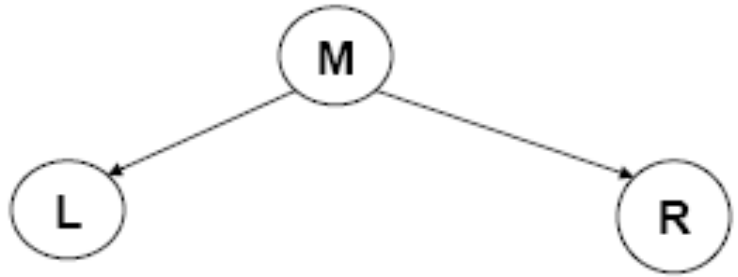


We can write down P(M). And then, since we know L is only directly influenced by M, we can write down the values of P(L | M) and P(L | ~M) and know we've fully specified L's behavior. Ditto for R.

P(M) = 0.6
P(L | M) = 0.085
P(L | ~M) = 0.17
P(R | M) = 0.3
P(R | ~M) = 0.6

'R and L conditionally independent given M'

Conditional independence



P(M) = 0.6

P(L | M) = 0.085

P(L | ~M) = 0.17

P(R | M) = 0.3

P(R | ~M) = 0.6

Conditional Independence:

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

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

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

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

Assume five variables

T: The lecture started by 10:35

L: The lecturer arrives late

R: The lecture concerns robots

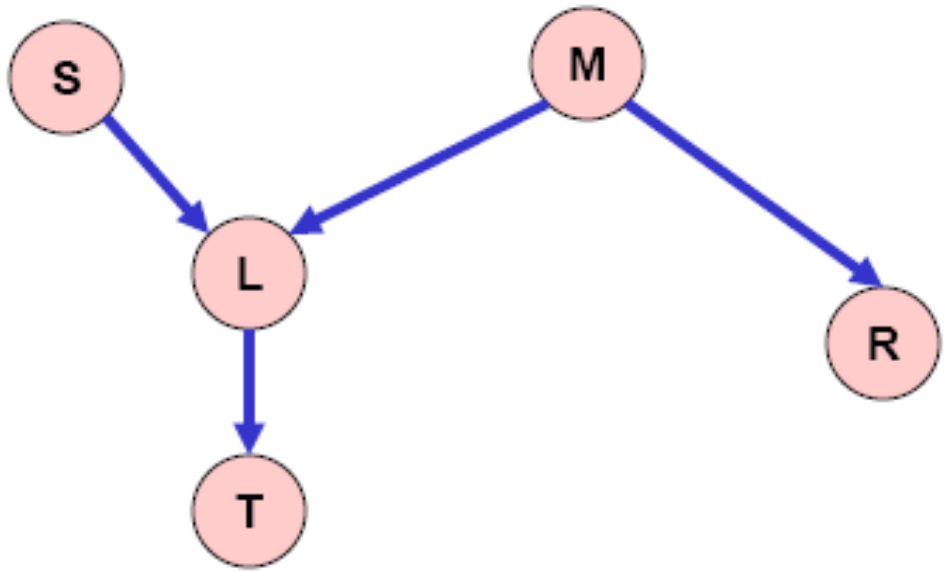
M: The lecturer is Manuela

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

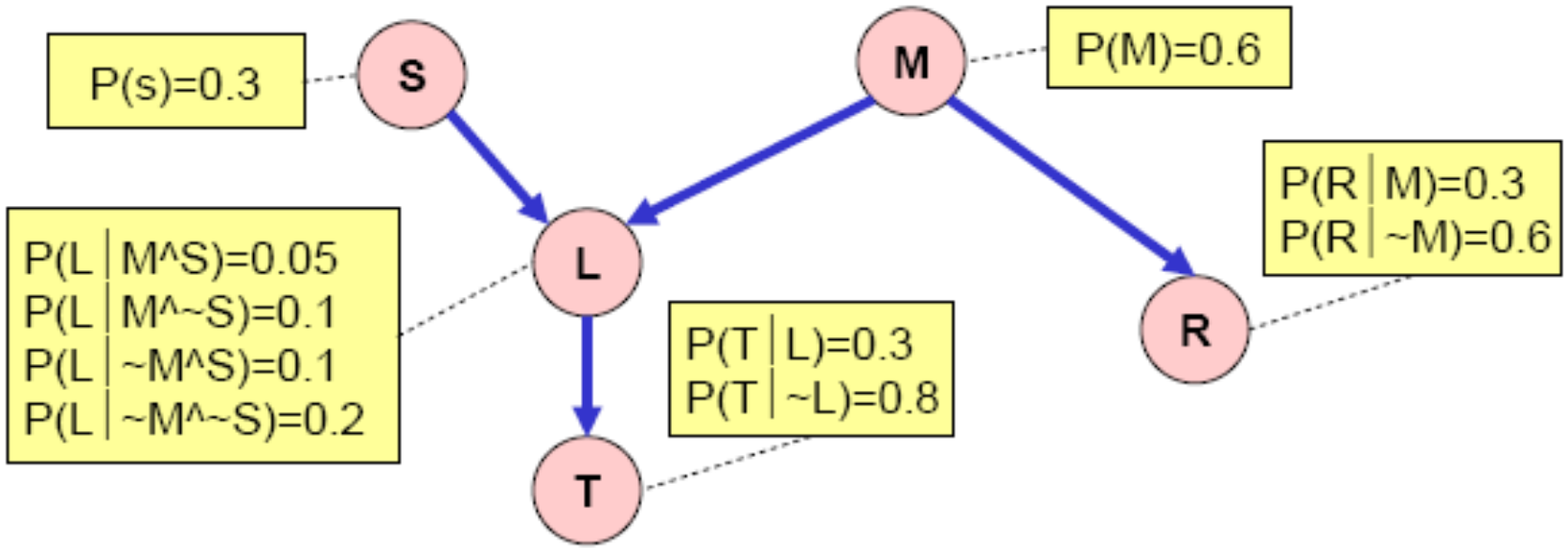


Step Two: add links.

- 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-descendent 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

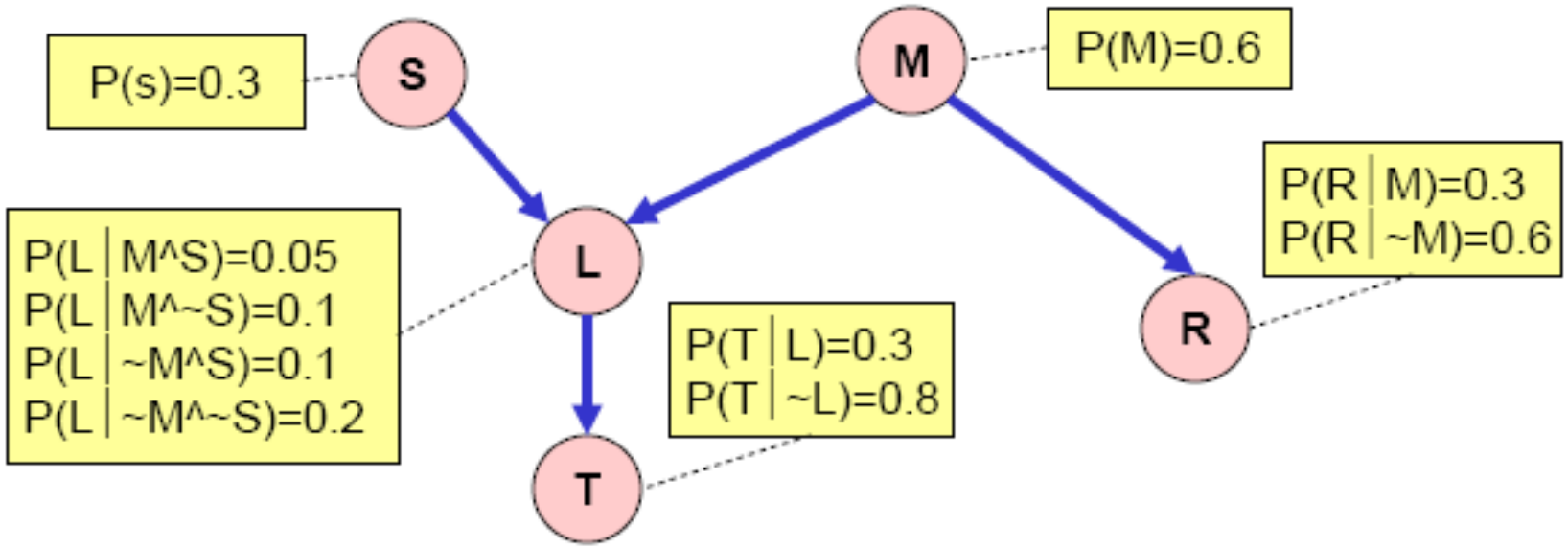


Step Three: add a probability table for each node.

- 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.

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.

Example Bayes Net Building

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

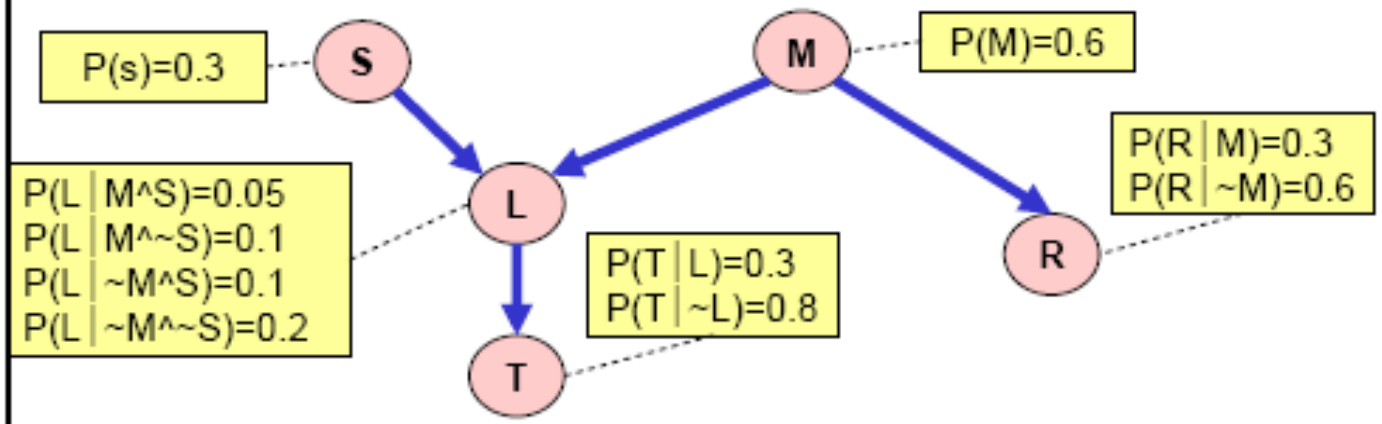
GRL : Gauge Reads Low.
CTL : Core temperature is low.
FG : Gauge is faulty.
FA : Alarm is faulty
AS : Alarm sounds

- 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.

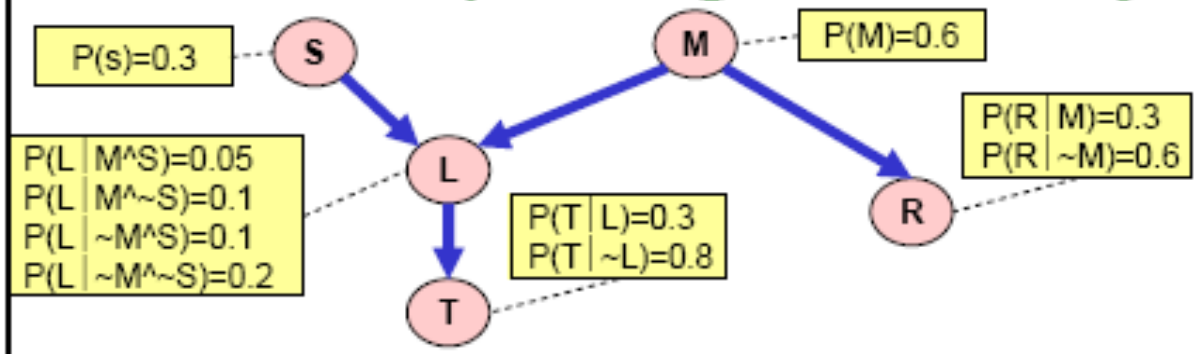
Computing a Joint Entry

How to compute an entry in a joint distribution?

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



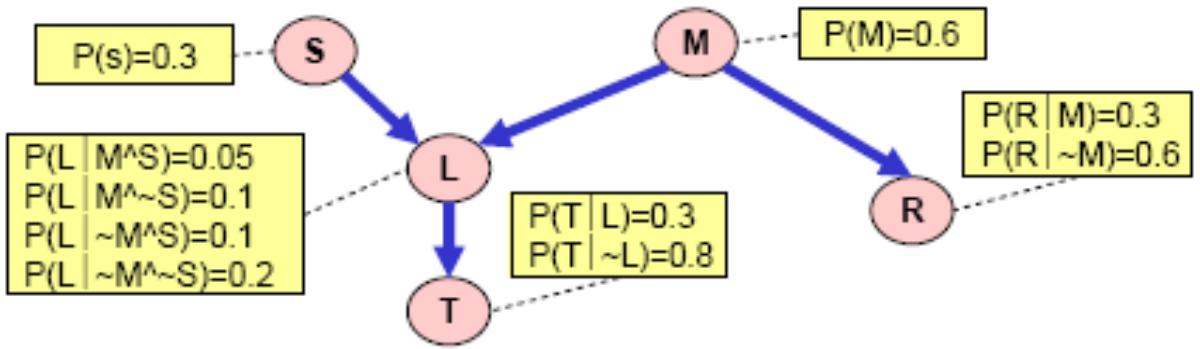
Computing with Bayes Net



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

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?

building Bayes nets.

l storage to hold our probability
ne maximum number of parents

es of any given assignment of
. And we can do it in time
des.

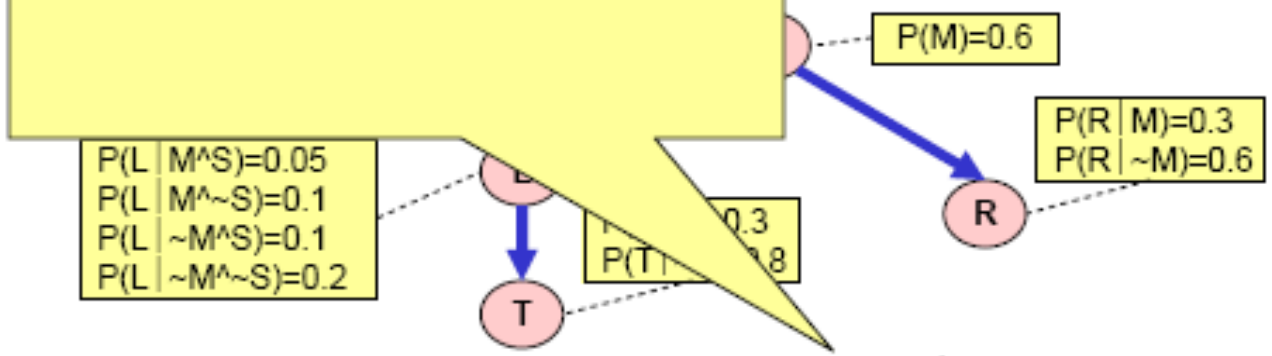
swers to any questions.

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

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

Step 3: Return

$$\frac{P(R \wedge T \wedge \sim S)}{P(R \wedge T \wedge \sim S) + P(\sim R \wedge T \wedge \sim S)}$$



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

$$\frac{P(R \wedge T \wedge \sim S)}{P(R \wedge T \wedge \sim S) + P(\sim R \wedge T \wedge \sim S)}$$

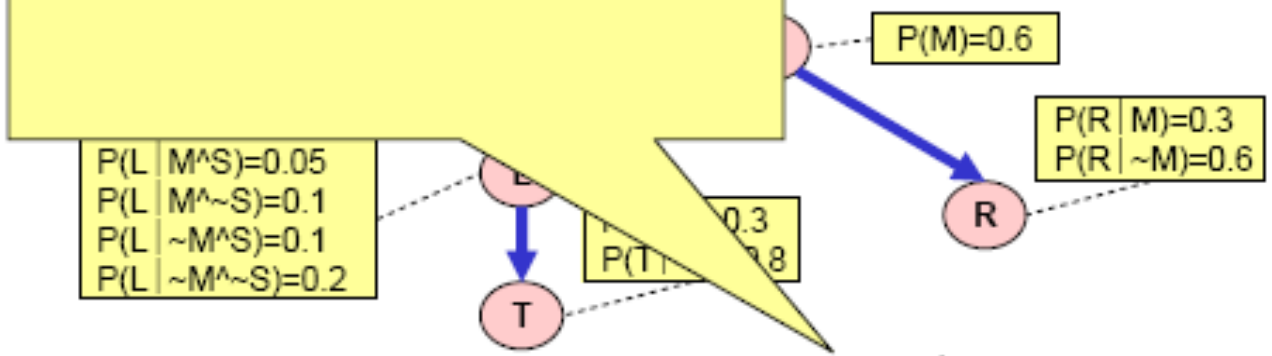
Sum of all the rows in the Joint that match R ^ T ^ ~S

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

...l s... ...he maximum number of parents

...es c... ... And we can do it in time ...des.

...swers to any questions.



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 sad, bad news

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

Exponential in the number of variables.

Example from real-life

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.

Example from real-life

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.

Example of AI not based on learning model parameters!

Example from real-life

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

- Diagnostic system for lymph-node dis…
- 60 diseases and 100 symptoms a…
- 14,000 probabilities
- Expert cons…
  - 8 h…
  - 35 h…
  - 40 hou…
- Apparently, …t the causal links and probabi…
- Pathfinder is … the world experts in diagnosis. Being extended to se… other medical domains.

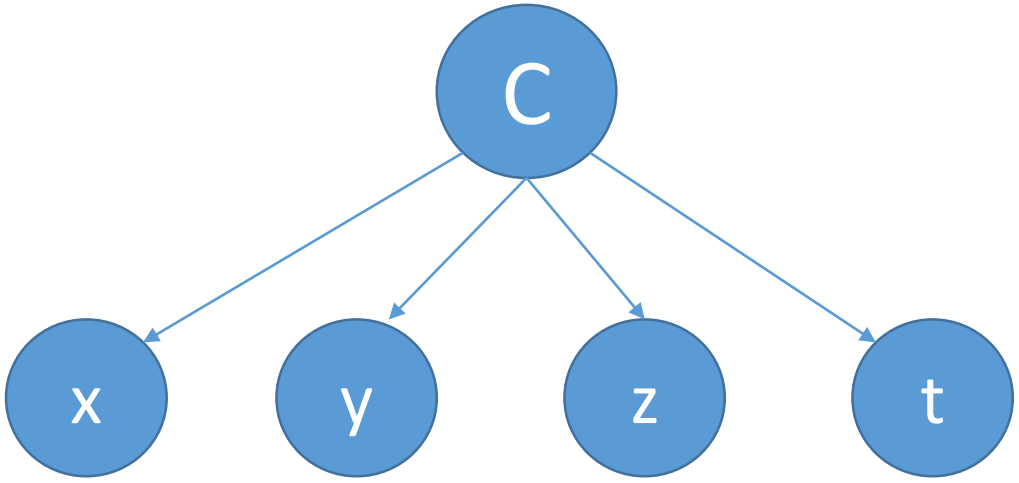
But the parameters can be learnt if we have enough data (Big Data)

Eg, this is the purpose of <u>Belief Propagation</u> approach

Example of AI not based … learning model …rameters!

Now, what is *naïve Bayesian assumption*?

- In simple words, it assumes that all variables (or a set of variables) are all conditionally independent : the Bayesian net is not connected
- Or, we have an unconnected Bayesian net connected to a single node



x,y,z,t are conditionally independent given C

This construction can be used to predict C from x,y,z,t values: this is **Naïve Bayes classifier**

What you should take with you

- Conditional independence of evens given other events
- Bayesian networks: convenient graphical way to represent known causalities and compute joint probability distribution
- Naïve Bayesian assumption is the simplest case: we assume that a set of variables is conditionally independent