Reasoning about uncertainty

Lecture notes for 15-381

22 February 2000

corresponding to Russell & Norvig, Chapters 14 & 15

David Cohn <http:/www.cs.cmu.edu/~cohn>

What you should get out of this section

- The the time we’re done with the next few lectures, you should
  - understand joint, marginal and conditional probabilities
  - be able to derive and manipulate Bayes’ rule
  - be comfortable with conditional independence
  - know what a poly-tree is
  - know what a Bayes net
  - be able to describe three strategies for doing inference with Bayes nets
  - have a vague bit of angst about what it actually takes to fill in entries in a joint or conditional probability table

The limitations of logic

- Logic has a finite number of possible states - true, false, unprovable, etc.
  – if it can be derived from rule base, it’s true, if its converse can be, its false
  – beyond that, you’re on your own
- How to encode
  – whether or not it will rain today?
  – whether the cafeteria is serving eggplant soufflé?
  – how many students will miss class because they tried the eggplant soufflé?
  – whether your lecturer will run out of quarters?
- All rely on information which is either unknown or hidden
  – can’t prove or disprove, so how do we reason about these vital issues?
  – the answer is (drum roll, please)....

...Probability

- Why probability?

  10) universally accepted language for dealing with uncertainty

  9) well-developed semantics and theory

  8) concise mathematical description

  ...

  but the #1 reason for representing uncertainty with probabilities is...

  1) coupled with a utility function, it's the method of choice for being a "rational agent"

- So what on earth does that mean?
  - rationality is a couple of lectures in and of itself,
  - but in brief: given:
    - an additive *utility* function for all states of the world,
    - an estimate of probabilities of various states given various action
  - a "rational agent" takes actions that maximize average utility

Some notation

- *p(a)* - means probability that event *a* is true
  - shorthand for *p(a=true)*
  - frequently just call it "probability of *a*"
  - *a* could be
    - "cafeteria is serving eggplant soufflé"
    - "the next hand dealt will be a full house"
- *p(x=blue)* - means probability that variable *x* takes value *true*
  - values for a variable are assumed to be mutually exclusive
- *p(payoff=0.5)* - probability that today's take is $0.50
  - can also have *continuous* distributions, but we'll concentrate on discrete ones here
- $p(a \wedge b)$ - probability that event *a* and *b* both occur
  - "probability that cafeteria is serving eggplant soufflé *and* Pittsburgh wins the next Superbowl"
- $p(a \vee b)$ - probability either event *a* or *b* occurs (or both)

Great, so we've got probabilities - now what do we do with them?

- We start off with some axioms that let us generate a few useful tools for manipulating probabilities
  – an *axiom* is an assumption that we're willing to make in order to get any reasoning done (should generally be pretty plausible)
    1) $0 \leq p(a) \leq 1$
    2) $p(true) = 1$
    3) $p(false) = 0$
    4) $p(a \lor b) = p(a) + p(b) - p(a \land b)$

- From here, can derive a lot of other relations:
  – $p(\sim a) = 1 - p(a)$
  – $p(a) = p(a \land b) + p(a \land \sim b)$
    - more generally,if b_1, b_2, b_3, b_n are mutually exclusive, and one of b_i must be true
    - $p(a) = p(a \land b_1) + p(a \land b_2) + p(a \land b_3) + ... + p(a \land b_n)$

- *Can you compute XOR?*

So let’s take it on the road

- Practical application: knowing when to panic
  - You have a pair of lucky socks that you frequently wear
    - {L, U}
  - You are enrolled in 15-381. On days when you don’t have class, you sleep in, except for Sundays, when you visit your parents to do laundry and and ask for money
    - {C, S, P}

s	m	t	w	th	f	st
P	S	C	S	C	S	S
U	U	U	L	L	L	L

Some more notation

- We've already met the *joint probability*
  - probability of a specific assignment for each variable
  - event completely describes one state of the world

s	m	t	w	th	f	st
P	S	C	S	C	S	S
U	U	U	L	L	L	L

  - Summarize as a *joint probability table*

	C	N	P
L	1/7	3/7	0
U	1/7	1/7	1/7

  - Given the joint probability table, can compute anything about data
    - What is the probability that you're wearing your lucky socks and you have class today?
    - What's the probability that you get to sleep late?

Some more notation

- Also touched on *marginal probability p(L)*
  - probability of some subset of variables, marginalizing over the values that don’t appear (in this case, what you happen to be doing today besides wearing socks)

	C	N	P
L	1/7	3/7	0
U	1/7	1/7	1/7

  - compute using one of our derived probabilistic devices:

$$p(L) = p(L \wedge C) + p(L \wedge N) + p(L \wedge P)$$
$$= 1/7 + 3/7 + 0 = 4/7$$

  - what other marginals can you compute?

Last bit of notation (for a while)

- Meet the *conditional probability* $p(C|L)$
  - probability of *C* given that we know *L*

	C	N	P
L	1/7	3/7	0
U	1/7	1/7	1/7

- You wake up - oh boy, maybe today's the day I get to go home and wash my lucky socks!
  - What's the probability? $p(P)$
- Then you look down and notice that you're not wearing your lucky socks
  - Does this make you happy? Why?
  - What's $p(P|U)$?

Conditional probabilities

- Conditional probability can be computed from joint and marginal via the chain rule:
  $p(A|B) = p(A \wedge B)/p(B)$
- Also useful to remember it as
  $p(A \wedge B) = p(A|B) * p(B)$

	C	N	P
L	1/7	3/7	0
U	1/7	1/7	1/7

- So now, armed with conditional probabilities,
  - You startle awake at 2:00 and notice that you're in class - what's the probability you're wearing your lucky socks?
- Extra credit: how would you use the chain rule to decompose $p(A \wedge B \wedge C \wedge D)$ into many simpler terms?

Bayes rule - putting it all together

- Funny thing about the chain rule - starting with joint $p(A \wedge B)$, we can run it both ways:

  $p(A \wedge B) = p(B|A) * p(A)$, and
  $p(A \wedge B) = p(A|B) * p(B)$, so...

  $p(A|B) = p(B|A) * p(A) / p(B)$

  – This is Bayes Rule. People have been doing clever bits of inference with it for hundreds of years.
  – It is simple, yet surprisingly powerful
  – Using it properly is a religion for many people

Bayes rule example

- In the utopic future, you're a sysadmin
  - From experience, you know that running Windows2537 on a computer will cause its fusion core reactor to melt down 50% of the time.
  - You know from Slashdot's archives that only one in 50,000 computers these days are running Windows2537
  - You've just read in USA Today's trivia blurb that if you pick a computer at random, there's a 1/20 chance that it will suffer fusion core meltdown today.
- You get a haptic relay call from a customer whose computer has just suffered fusion core meltdown - what's the probability you're going to have to deal with Windows2537?

Ahahaha! We can now compute anything, and no one can stop us!

- True, but this still generally requires filling out the entire joint probability table
  - *Why might this be a problem?*
- Let's say our problem has 4 variables, each of which can take 3 values
  - how many probabilities do we have to know in order to fill in our JPT?
    $3^4 = 81$
  - what if we add another variable with 7 values?
    $3^4 * 7 = 567$
  - how about a realistic sized problem of 20 variables with 3 values apiece?
    we'd have to estimate 3,486,784,401 distinct probabilities!
- The problem: the number of entries in our JPT is exponential in the number of variables
- Ouch.

A digression - computing entries for JPT

- A couple of different ways to do this
  - guess what we think is right
  - estimate from data
- “Bayesian approach” (the religious one) combines these two. For some entry X:
  - describe *prior* belief on the distribution of possible values for X
  - gather data Y (called *evidence* by Bayesians)
  - combine prior with evidence to form *posterior* belief distribution
- Typically, *prior* distribution is spread out
  - evidence focuses the prior into a narrower *posterior* distribution
- Describing one’s knowledge about the state of the world as a distribution is what gives probabilistic reasoning its power

An example (adapted from Andrew Moore)

- You have three identical laundry baskets: B1, B2, and B3
  - in B1 there is 1 pair of clean socks and 3 pairs of dirty socks
  - in B2 there are 2 pairs of clean socks and 2 pairs of dirty socks
  - in B3 there are 4 pairs of clean socks, and no dirty ones
- You wake up. It's dark. You fumble around the room until you find a laundry basket. You reach in, grab a pair of socks, and give them a sniff.
  - *What can you infer about which laundry basket you've found?*
- You throw the pair you've drawn back into the basket and reach in again.
  - *Now how well can you predict whether you'll pick a clean pair?*
- These questions are the foundations of *reasoning with uncertainty* and *machine learning*

Two problems with JPTs

1) Computing entries can require a lot of data

2) There are a lot of entries
   - What can we do about this?

- Let's say we're running late
  - we reach in, grab a pair of socks, and go
  - we run out the door and hope we haven't missed the bus
  - *how many entries do we need for the JPT?*
  - do we really need that many?
  - *What if we could encode the fact that catching the bus does not depend on having clean socks?*
  - We can: $p(B|C) = p(B)$
- This is called *conditional independence*
  - you will learn to love it (or hate it) almost as much as Bayes' rule
- If B and C are conditionally independent, then $p(B \wedge C) = p(B)\, p(C)$

Conditional independence - a more complicated example

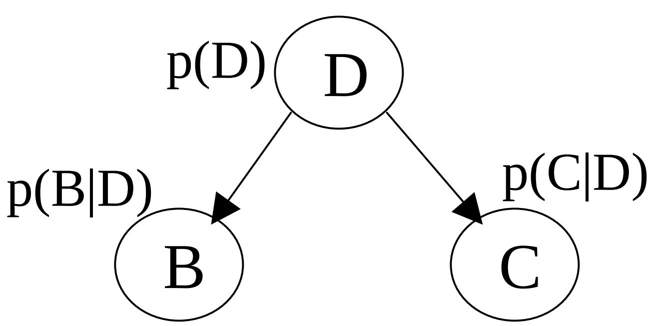
- You sit down and analyze the problem:
  - as the week progresses and you use up clean socks, you have an increasing chance of grabbing a dirty pair
  - as the week progresses, the bus driver becomes increasingly surly and impatient, and is more likely to try to finish early
- So the two *are* related! (How?)
- We now have three variables:
  - B - whether you catch the bus
  - C - whether you have clean socks
  - D - what day of the week it is

  how big is the JPT?

- **Warning - major insight ahead. Sit up straight, pay attention, and prepare to ponder the wonders of the universe...**

Conditional independence example

- Take a deep breath.
- *B* and *C* are clearly correlated because they both depend on *D*. But once you know *D*, *B* and *C* are conditionally independent.
- **So *B* and *C* are conditionally independent given *D***
  - We can use two smaller JPTs: one for *p(C^D)*, and one for *p(B^D)*
  - Or three smaller *conditional* probability tables, which are a little easier to reason with: *p(D)*, *p(C|D)*, and *p(B|D)*
- Denote graphically as:



Conditional independence

- Formally, *B* and *C* are conditionally independent given *D* if and only if

  for all values of *B*, *C* and *D*:

  $p(B|D \wedge C) = p(B|D)$

- This is very, very powerful
  - JPT - many entries must be generated/estimated
    - each data point we get only helps with one entry
  - CPT - many fewer entries need be generated
    - each data point helps with each entry!

- Example: Pathfinder system (1991)
  - diagnostic system for lymph node diseases
  - 60 diseases, 100 symptoms
  - how big is the joint probability table?
    - 1.46×10^{48} entries!
  - by identifying conditional independences, able to get it down to 14,000 probabilities

Bayes nets and graphical models

- What we've just done is started building a Bayes network
  - also called a belief network, probabilistic net, causal net, graphical model, etc.
- A Bayes net is a directed acyclic graph, augmented with
  - a node for each variable in the model
  - a directed edge from each node to all the other nodes it *directly* influences
  - unconditional probabilities on all nodes that have no parents
  - conditional probability tables on all edges
- An example
  - you're building a nuclear reactor in machine shop
  - you have the following random variables
    - GRH - temp gauge reads high
    - CTH - core temp is high
    - AS - alarm sounds
    - GF - gauge is faulty
    - AF - alarm is faulty
  - how would we build a probabilistic model of this?

Building a Bayes net

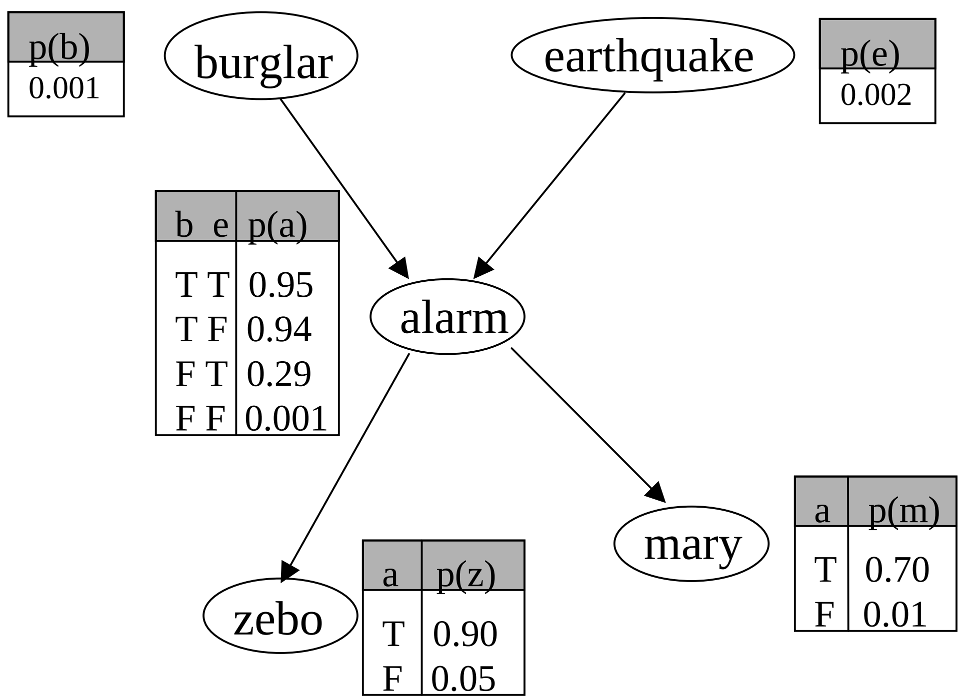
- An algorithm
  1) choose a set of relevant variables
  2) pick an ordering for your variables $X_1 \ldots X_n$
     - hint: make sure it's the *right** ordering of variables
  3) for each X_i in turn:
     - add it as a node in the network
     - set $parents(X_i)$ to be the minimal set of $X_1 .. X_{i-1}$ such that X_i is conditionally independent of $X_1 .. X_{i-1}$ given $parents(X_i)$
     - define conditional probability table for $p(X_i | parents(X_i))$

*There may be many "right" orderings

Bayes net example (from Pearl)

- You live in southern California. You drive a Porsche. You have a burglar alarm. It can be set off by
  - burglars
  - earthquakes
- You have neighbors who will call you at work if they hear an alarm
  - Mary plays in a thrash metal band and has lost most of her hearing
    - If she hears the alarm, she'll call, but she frequently misses it
  - Zebo almost always hears alarms, but also hears voices inside his head
    - If the alarm goes off, he'll definitely hear it, but he frequently hears it even when it's not there
- Let's build the network for a typical day at work

Bayes net example

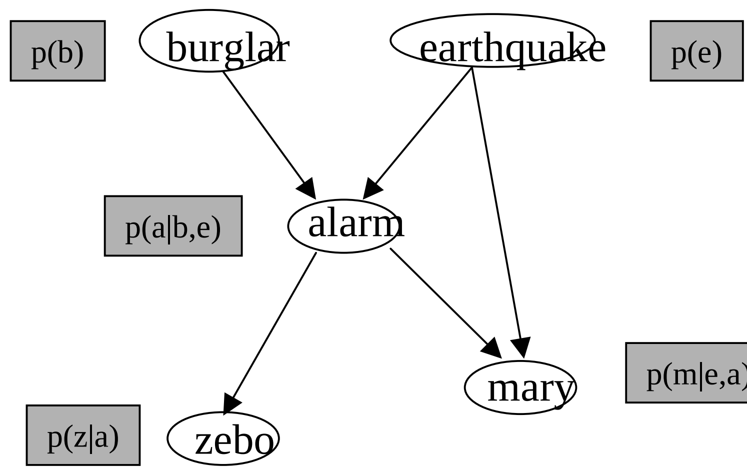


- Okay, let's try to compute with this thing
  - Can we compute an entry the JPT?
  - How about a marginal?
  - How about a conditional: *Zebo calls and tells us he hears an alarm - what's the probability we're being burgled?*

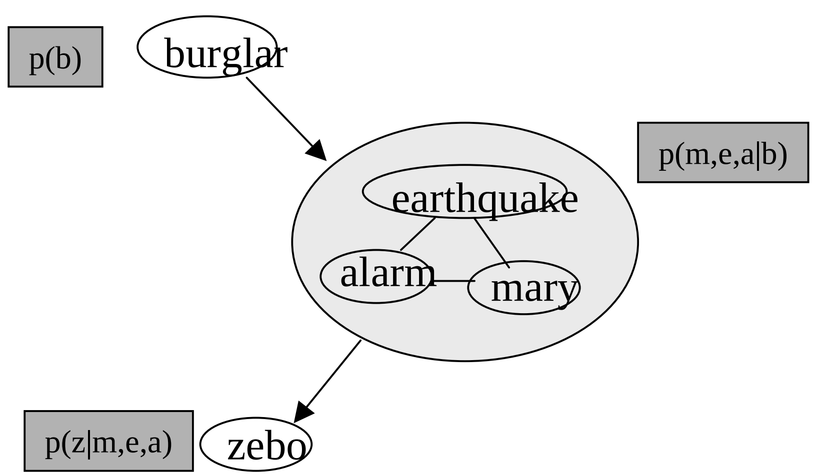
Bayes nets, poly-trees, and inference

- What we've got here is a *poly-tree*
  - a poly-tree is a Bayes network in which no two nodes have more than one path between them
  - inference in poly-trees is linear in the number of variables - hurrah! (see Russell & Norvig)
  - inference in more general Bayes networks is NP-complete - urgh.
- What do we do when our Bayes net isn't a poly-tree?
  - clustering
  - conditioning
  - stochastic simulation

Clustering for Bayes net inference

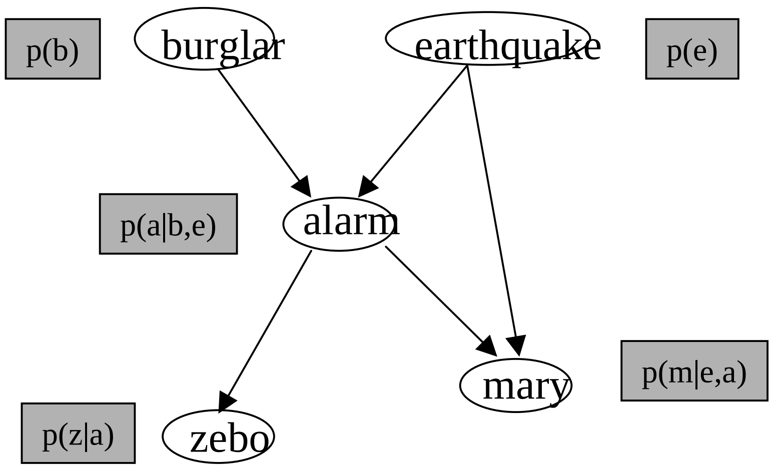


- Given a BN which is not a a poly-tree, we can merge nodes into clusters such that we do have a poly-tree:

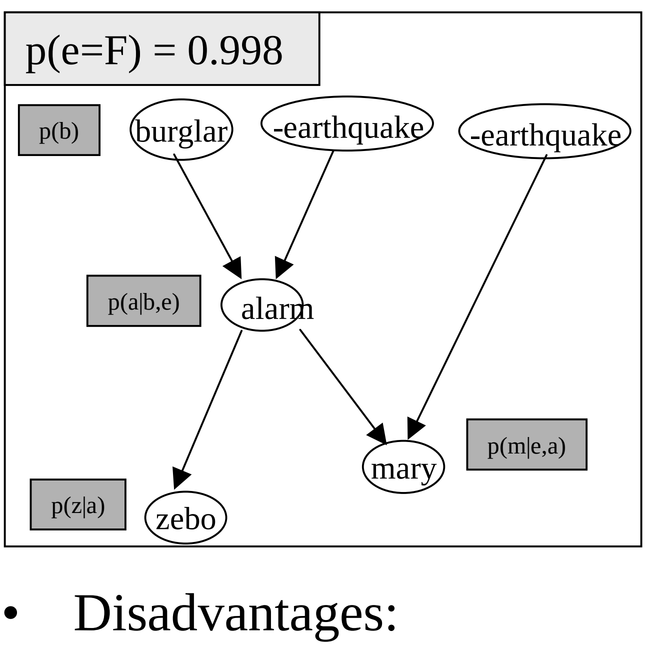


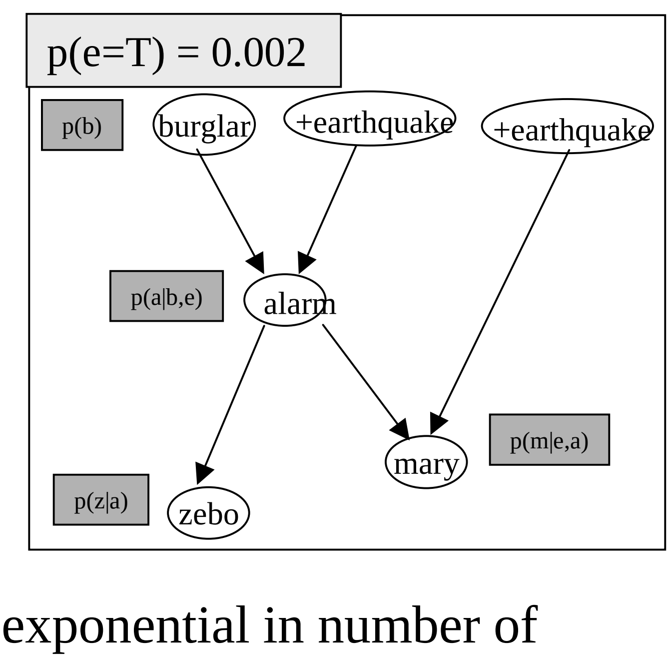
- Disadvantages:
  - needs to be done carefully to minimize blowup of CPTs

Conditioning for Bayes net inference



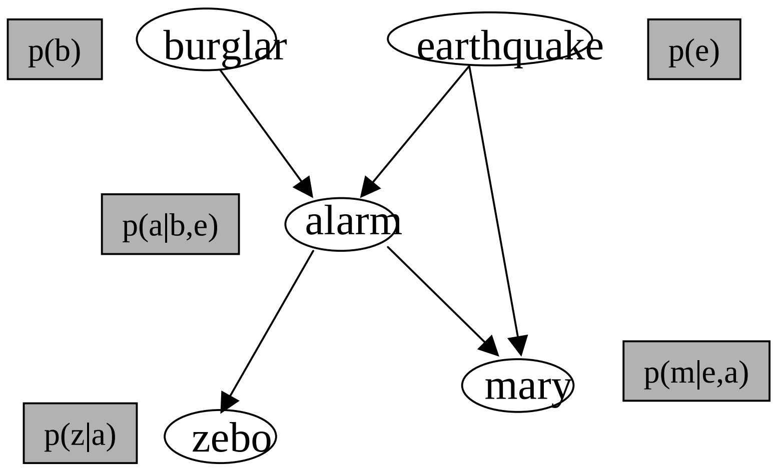
- We could also instantiate the troublesome variable(s) and compute a weighted average of the resulting networks:





- Disadvantages:
  - number of networks is exponential in number of instantiated variables- want minimum cutset

Sampling for Bayes net inference



- We could also generate random samples from the JPT and compute the average
  - Let's say Mary calls and says she heard an alarm *What's the probability that we're being burgled?*
  - How to compute $p(b|m)$?
    - generate a random sample of values from the network
      - set unconditionals *b* and *e* at random according to $p(b)$ and $p(e)$
      - set a at random according to $p(a|b,e)$, etc.
      - continue until we have a complete "sample" - all variables have an assignment
    - repeat a zillion times, until we have gathered many samples
    - look at the fraction remaining for which $b=T$
- Disadvantages:
  - can require lots of samples, especially when computing rare events
  - special techniques exist for improving sampling efficiency