More Probability and Bayes Nets

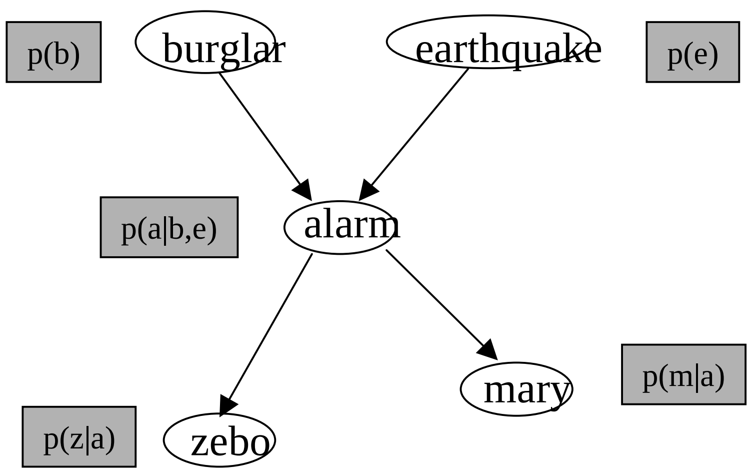
15-381 Lecture Notes

24 Feb 2000

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

Poly-tree inference algorithms

- Joint probability of all variables is easy
- Marginals and conditionals a bit harder
- We've generated a Bayes net
  - Just by looking at the net, can we infer whether *A* is conditionally independent of *B* given *C*?
  - Is it always easy?



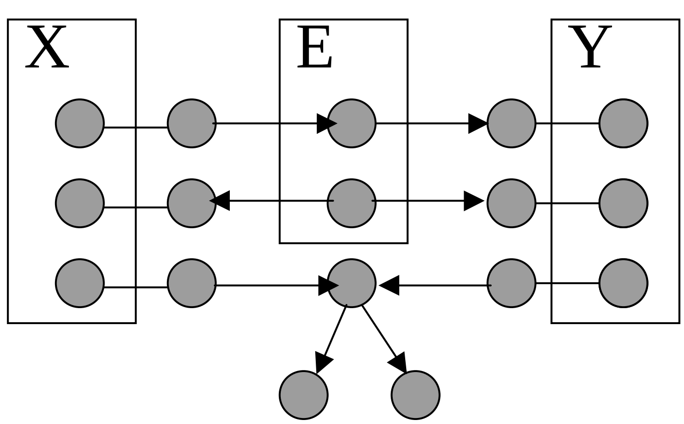
- Zebo calls - what's the probability that we're being burgled? *[p(b|z)]*
- We vaguely remember that there was an earthquake a few minutes ago [p(b|z,e)]
  - *what happened to independence!?!?*
  - this phenomenon is called "explaining away"

D-separation

- We can assert conditional independence of a set of variables *X* and *Y*, given the values of a set of variables E, if *X* and *Y* are **d-separated** by *E*
- *X* and *Y* are d-separated by *E* iff all paths of inference between *X* and *Y* are blocked by *E*

D-separation

- What does it mean for a path of inference to be "blocked"?
  - and what is a "path of inference" anyway?
- **path of inference** - an series of connected edges, regardless of their orientation, that lead from a node in *X* to a node in *Y*
- **blocked** - a path of inference is blocked given *E* if
  - variable *Z* on the path is in *E* and has one arrow going in and one arrow going out
  - *Z* is in *E* and has both arrows leading out
  - neither *Z* nor any of its descendents is in *E*, and both arrows lead in to *Z*



Poly-tree inference

- $p(x|E) = p(x|E_x^+, E_x^-)$ - by independence
- $p(x|E) = p(E_x^-, |x, E_x^+)\, p(x|E_x^+) / p(E_x^-|E_x^+)$
  - by Bayes rule
- $p(x|E) \propto p(E_x^-, |x)\, p(x|E_x^+)$
  - by independence and normalization by $1/ p(E_x^-|E_x^+)$
- Causal support $p(x|E_x^+)$
  - affects x only via parents of x: u_i
  - compute via marginalization
    - $p(x|E_x^+) = \Sigma_u\, p(x|u, E_x^+) p(u|E_x^+)$
    - $p(x|E_x^+) = \Sigma_u\, p(x|u, E_x^+) \Pi_i p(u_i|E_x^+)$
      - since E_x^+ d-separates each of the u_i
    - $p(x|E_x^+) = \Sigma_u\, p(x|u, E_x^+) \Pi_i p(u_i|\, E_{Ui\backslash x})$
      - since $E_{Ui\backslash x}$ d-separates u_i from other evidence in E_x^+

    - now have a recursive equation!

Poly-tree inference

- Evidential support $p(x|E_x^-)$
- Not only take into account *Y*, but *Z*
- *A bit more complicated, but end up with*

  $p(E_x^-|x) = \Pi_i \Sigma_{yi}\, p(E^-_{Yi}|y_i) \Sigma_{zi}\, p(y_i,z_i|x)\ \Pi_j\, p(z_{ij}|E^-_{zij\backslash yi})$
- End up with

  $p(E^-_{Yi}|y_i)$ is a recursive instance of $p(E_x^-|x)$

  $p(y_i|x,z_i)$ is a CPT table entry

  $p(z_{ij}|E^-_{zij\backslash yi})$ is a recursive instance of the $p(x|E)$ calculation
- Causal and evidential support are both incorporated as recursive calls from *x*
  - each recursive call eliminates the node it was called from
  - each node can only be reached from one parent/descendent

Where do probabilities come from?

- How to get the numbers for our CPTs?
  - the *frequentist* approach: life is an empirical experiment - we can with reason with observed frequencies of events
  - *objectivist* approach: the truth is out there - we just have to find it
  - *subjectivist* approach: all we're doing is measuring "belief" (as modified by evidence, when we have it)
- But practically, what do people do? Most common are subjectivist, semi-Bayesian approaches called *maximum likelihood estimation* (MLE) and *maximum a posteriori* (MAP)
- Crossing the murky waters from probability to statistics

Maximum likelihood estimation

- Let's say that the burglar decides whether to burgle you today based on a coin flip
  - Given his coin, find best guess for table entry $p(b)$
  - we'll call the value of this entry J
  - want to combine prior knowledge about coins with evidence from coin flips in probabilistic way to get best guess for J
    - *MLE gives us a guess that is "maximally likely" given the evidence*
  - Bayes rule says $p(J=\gamma|e) = p(e|J=\gamma)\, p(J=\gamma) / p(e)$
    - want to find value of γ that maximizes $p(J=\gamma|e)$
    - MLE assumes a flat prior
      - $p(J=\gamma)$ is equal for all values $0 \leq \gamma \leq 1$, so we can take it out of equation
      - $p(e)$ doesn't depend on γ, so we can take it out of the equation
    - left with $p(J=\gamma|e) \propto p(e|J=\gamma)$
  - So, the γ that is maximally likely is the γ that makes the observed data most likely

Maximum likelihood estimation

- Estimating *p(heads) = J*
  - gather evidence *e* in the form of coin flips
    - flip coins *m* times, get *n* heads
    - *what's obvious best guess for J?*
  - for each possible value of γ, compute

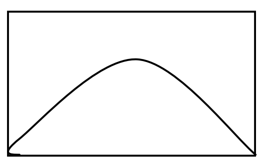
    $p(J=\gamma|e) \propto p(e|J=\gamma)$
- From binomial distribution for coin flips,

  $p(e|J=\gamma) = \gamma^n(1-\gamma)^{m-n}$
  - find γ that maximizes $\gamma^n(1-\gamma)^{m-n}$
    - differentiate and find root
- amazingly, likelihood is maximized for

  $\gamma = n/m$

Maximum a posteriori estimation

- So you flip the coin twice, and get 0 heads
  - what's your estimate of *p(heads)*?
  - need to incorporate some sort of non-uniform *prior* belief about coins
    - example prior: coins can be biased, but can't *always* land on just heads or tails
      - one possibility: $p(J=\gamma) = 0.25 - (\gamma - 0.5)^2$



  - as before, *p(e)* doesn't depend on again γ, so equation reduces to

$$p(J=\gamma|e) \propto p(e|J=\gamma)\, p(J=\gamma)$$

  - find value of γ that maximizes $p(e|J=\gamma)\, p(J=\gamma)$
- From binomial distribution,

$$p(e|J=\gamma)\, p(J=\gamma) = \gamma^n(1-\gamma)^{m-n}\, (0.25 - (\gamma - 0.5)^{2)}$$

  - differentiate and find root
  - prior biases estimates towards 0.5
  - will never estimate J=0 or $J=1$, no matter how many coin flips we make