CISC 6525

Bayesian Networks

Chapter 14

Outline

- Syntax
- Semantics
- Efficient representations
- Inference

Bayesian networks

- 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:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- 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

Independence

- Two random variables A and B are independent iff $P(A|B) = P(A)$

  Or $P(A,B)=P(A|B)P(B)=P(A)P(B)$

- If n boolean variables are independent then their full joint distribution is

  $P(X1,X2,\ldots,Xn) = \Pi_i \; P(X_i)$

  $= P(X1)P(X2)\ldots P(Xn)$

Absolute independence is a strong requirement.

Conditional independence

Recall dentist example: *Toothache*, *Cavity*, *Catch*.

Joint distribution has 2^3-1=7 independent entries

But P(Catch|Toothache,Cavity)= P(Catch|Cavity)

And P(Toothache,Catch|Cavity)=

P(Toothache|Cavity)

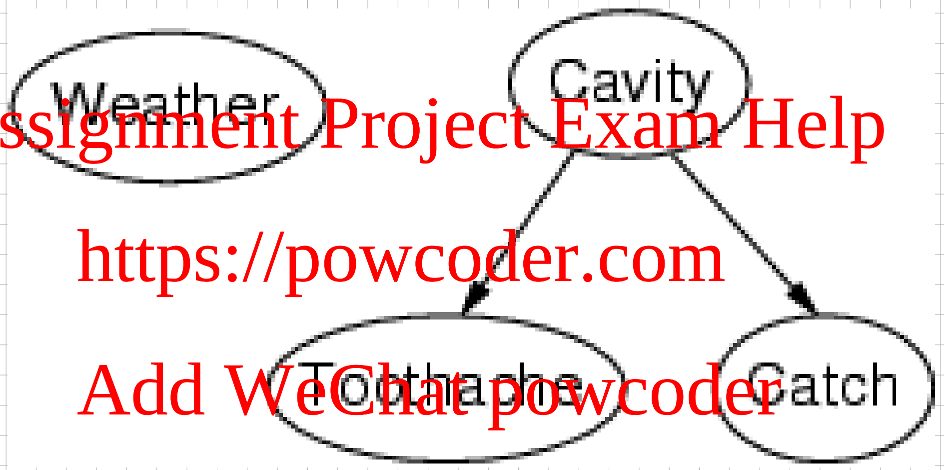
P(Toothache,Catch,Cavity)=

P(Toothache|Cavity)P(Catch|Cavity)P(Cavity)

Full joint distribution only has 5 independent numbers

Example

- Topology of network encodes conditional independence assertions:



- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

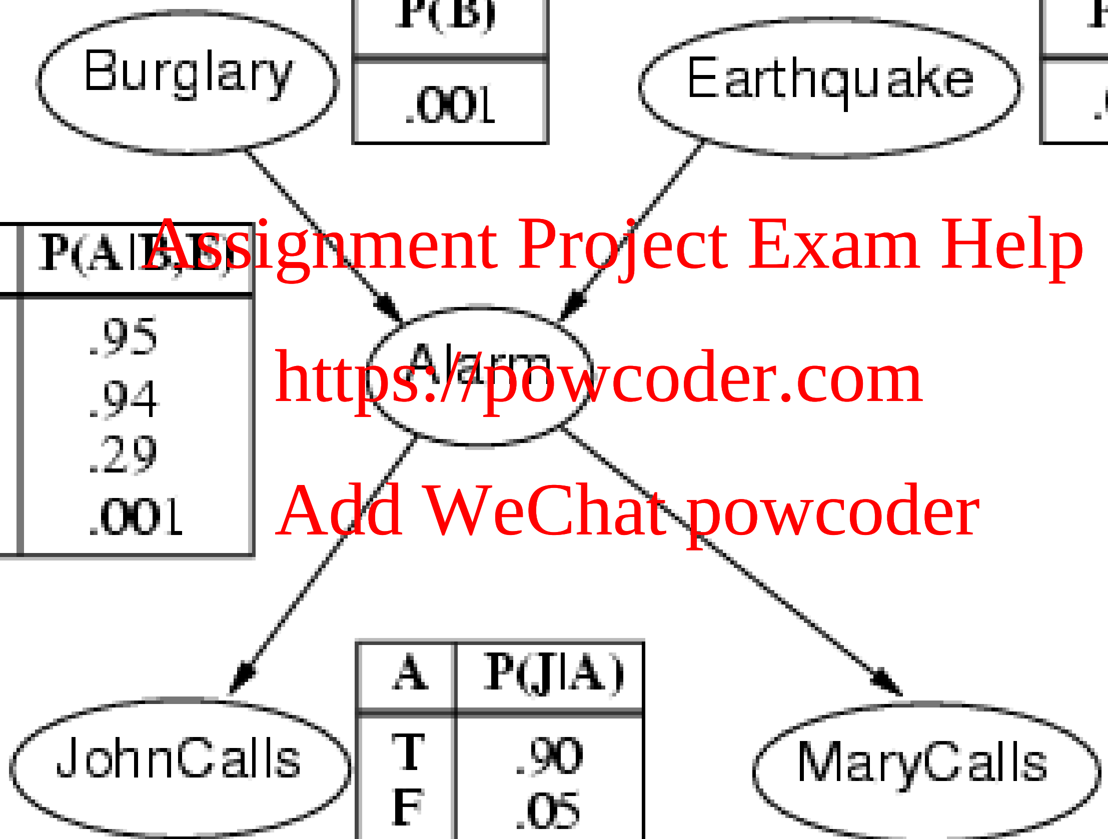
Example

- 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: *Burglary*, *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

Example contd.



P(B)
.001

P(E)
.002

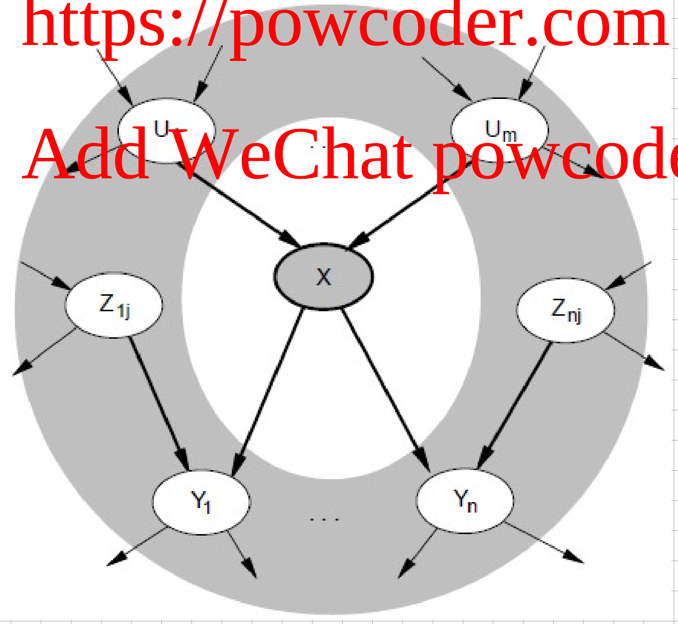
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

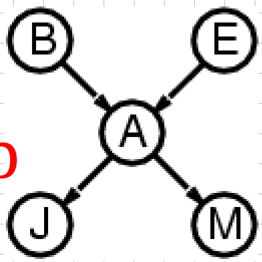
Markov Blanket

Each node is conditionally independent of all others given the *Markov blanket:* <u>parents+ children+children's parents</u>



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 $1-p$)

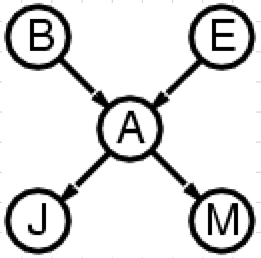
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers

- 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)

Semantics

The full joint distribution is defined as the product of the local conditional distributions:



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

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

$$= \mathbf{P}(j \mid a)\, \mathbf{P}(m \mid a)\, \mathbf{P}(a \mid \neg b, \neg e)\, \mathbf{P}(\neg b)\, \mathbf{P}(\neg e)$$

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \ldots, X_n$
- 2. For $i = 1$ to n
  - add X_i to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that

$$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$$

This choice of parents guarantees:

$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$$

(chain rule)

$$= \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$

(by construction)

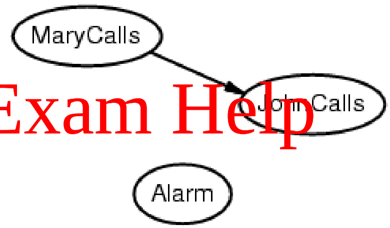
Example

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

$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

Example

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



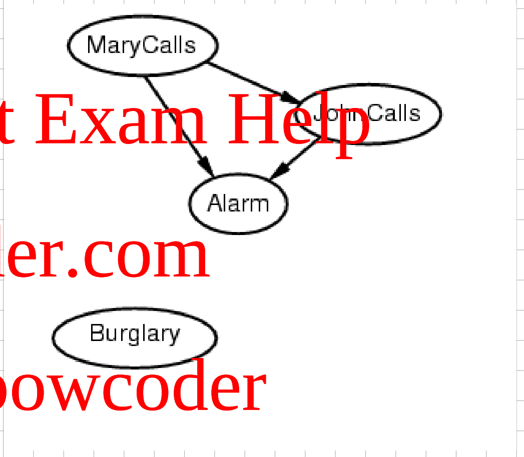
$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

No

$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$?

Example

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



$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

No

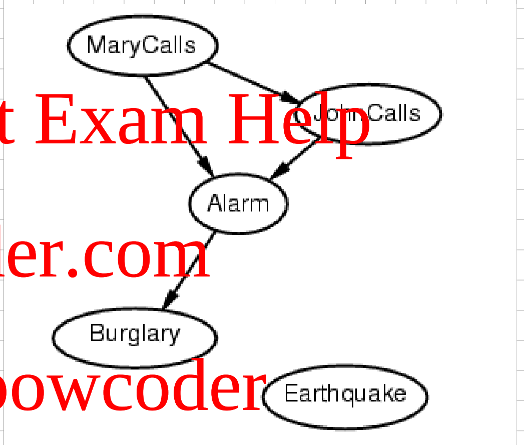
$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$?

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$?

Example

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



$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

No

$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Yes**

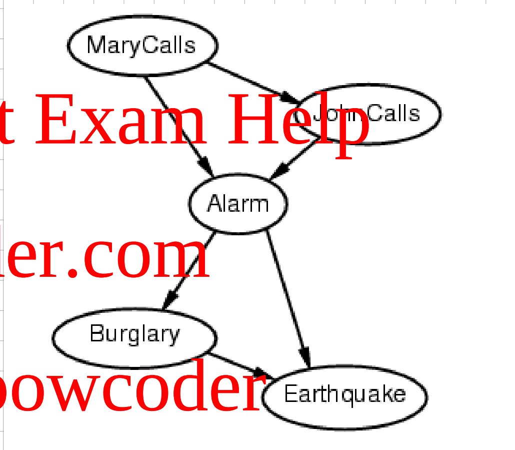
$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**

$\mathbf{P}(E \mid B, A, J, M) = \mathbf{P}(E \mid A)$?

$\mathbf{P}(E \mid B, A, J, M) = \mathbf{P}(E \mid A, B)$?

Example

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



$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

No

$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**

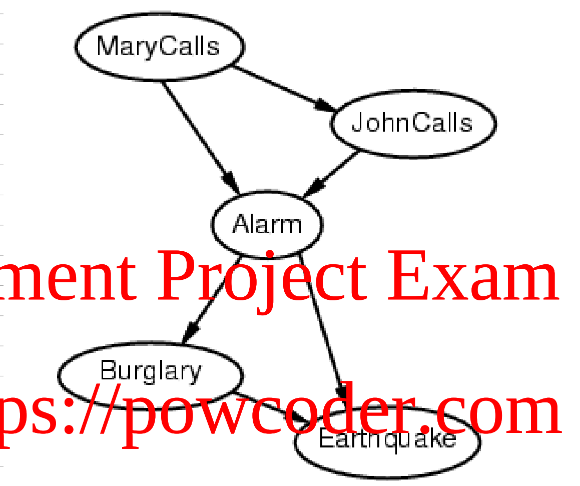
$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Yes**

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**

$\mathbf{P}(E \mid B, A, J, M) = \mathbf{P}(E \mid A)$? **No**

$\mathbf{P}(E \mid B, A, J, M) = \mathbf{P}(E \mid A, B)$? **Yes**

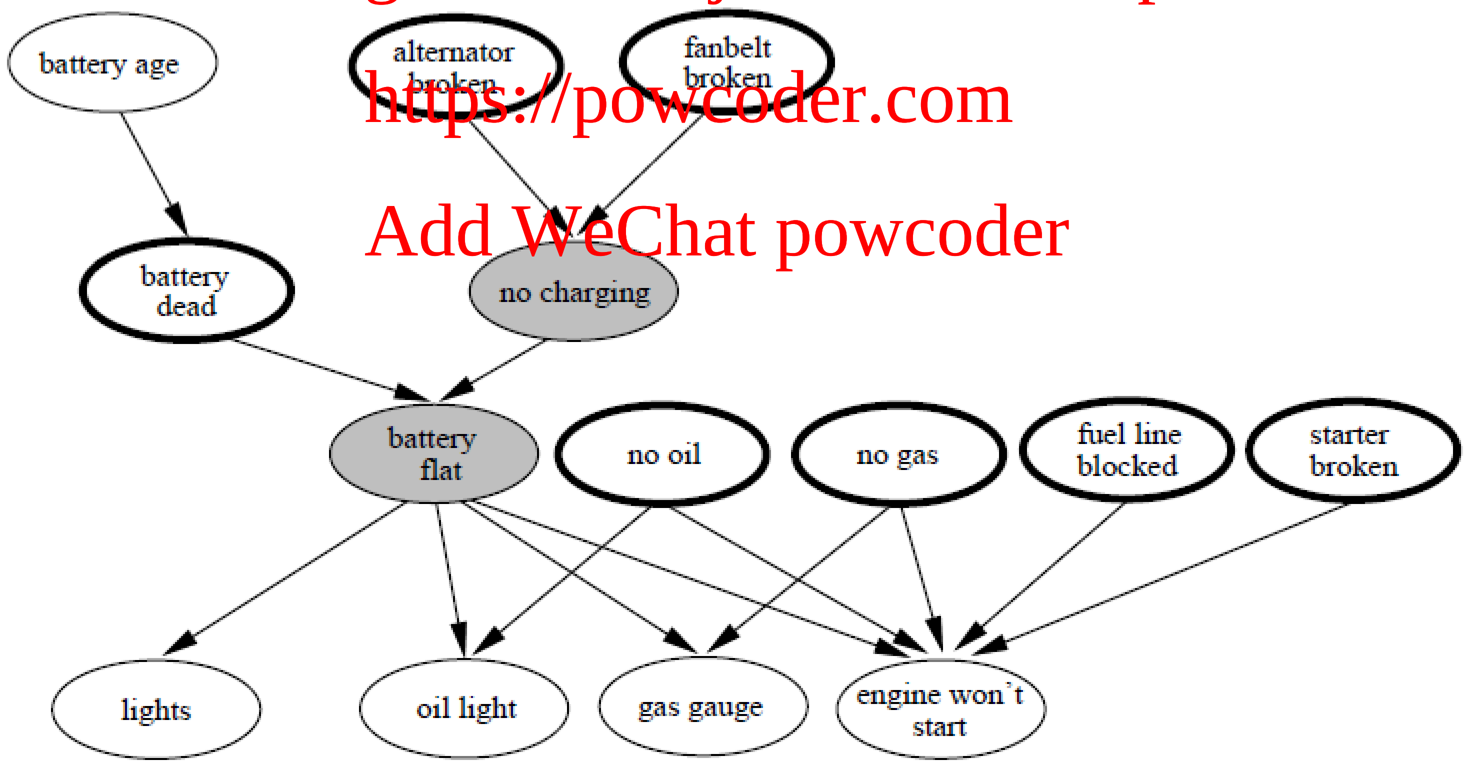
Example contd.



- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed

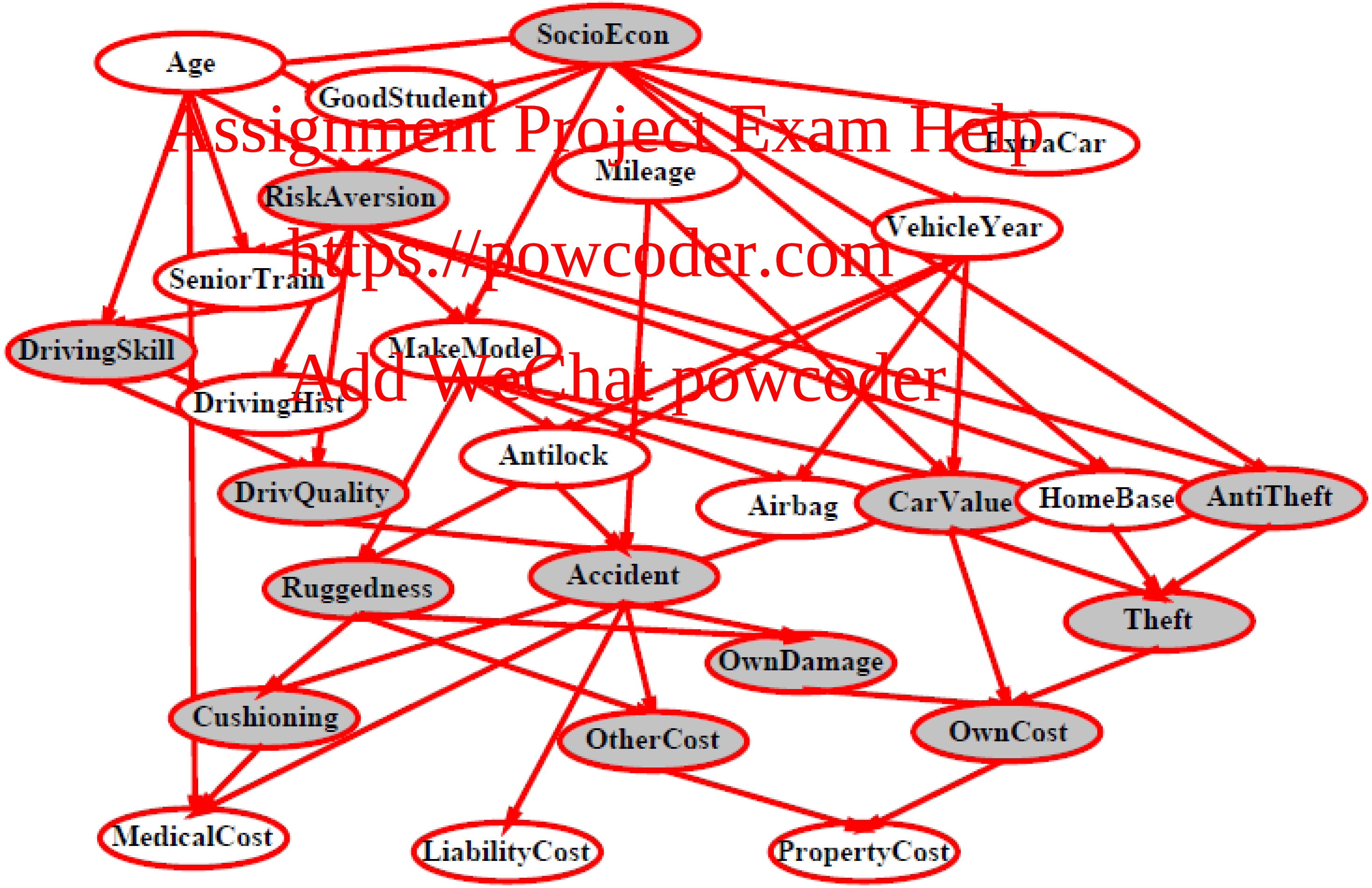
Example 1

Initial evidence: engine won't start
Testable variables (thin ovals), diagnosis variables (thick ovals)
Hidden variables (shaded) ensure sparse structure, reduce parameters



Example 2

Predict claim costs (medical, liability, property)
given data on application form (other unshaded nodes)



Compact Condition Distributions

CPT grows exponentially with num. parents and is infinite with continuous valued nodes.

Solution: Canonical Distributions – distributions defined in terms of a small number of parameters.

Canonical Distributions

- Boolean functions:

  $NorthAmerican \Leftrightarrow Canadian \vee US \vee Mexican$

- Numerical relationships

  $\partial L/\partial t = inflow + precip - outflow - evap$

Canonical Distributions

<u>Noisy-OR</u> distributions model multiple noninteracting causes

1) Parents $U_1 \ldots U_k$ include all causes (can add <u>leak node</u>)

2) Independent failure probability q_i for each cause alone

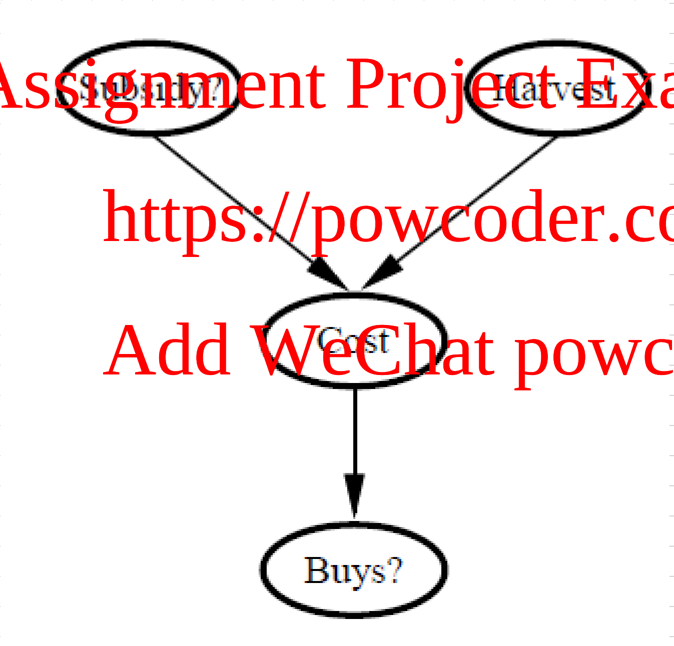
$$\Rightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	0.0	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Number of parameters <u>linear</u> in number of parents

Hybrid Distributions

Discrete (*Subsidy*, *Buys*); Continuous (*Harvest*, *Cost*)



1. Continuous variable, discrete+ continuous parents
2. Discrete variable, continuous parents

Continuous Canonical

Need one conditional density function for child variable given continuous parents, for each possible assignment to discrete parents

Most common is the linear Gaussian model, e.g.,:

$$P(Cost = c \mid Harvest = h, Subsidy? = true)$$

$$= N(a_t h + b_t, \sigma_t)(c)$$

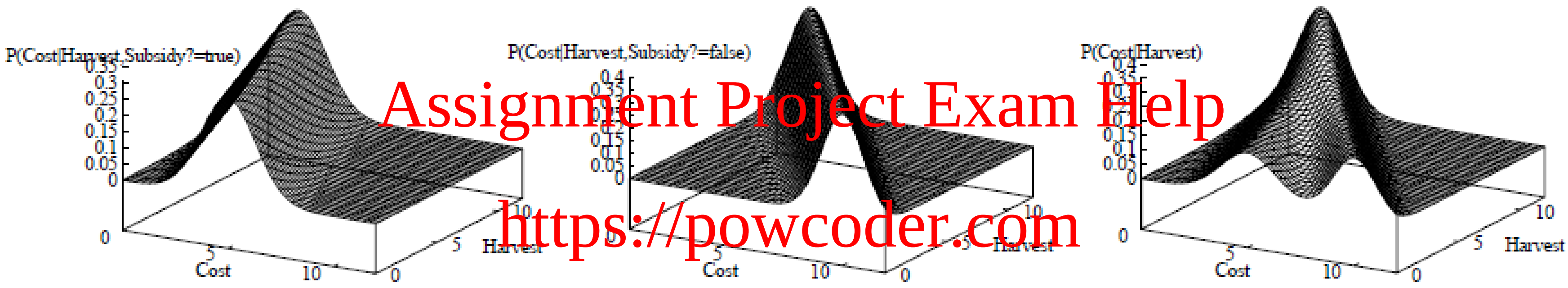
$$= \frac{1}{\sigma_t \sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{c - (a_t h + b_t)}{\sigma_t}\right)^2\right)$$

Mean $Cost$ varies linearly with $Harvest$, variance is fixed

Linear variation is unreasonable over the full range

but works OK if the likely range of $Harvest$ is narrow

Continuous Canonical



All-continuous network with LG distributions

$\Rightarrow$ full joint is a multivariate Gaussian

Discrete+continuous LG network is a <u>conditional Gaussian</u> network i.e., a multivariate Gaussian over all continuous variables for each combination of discrete variable values

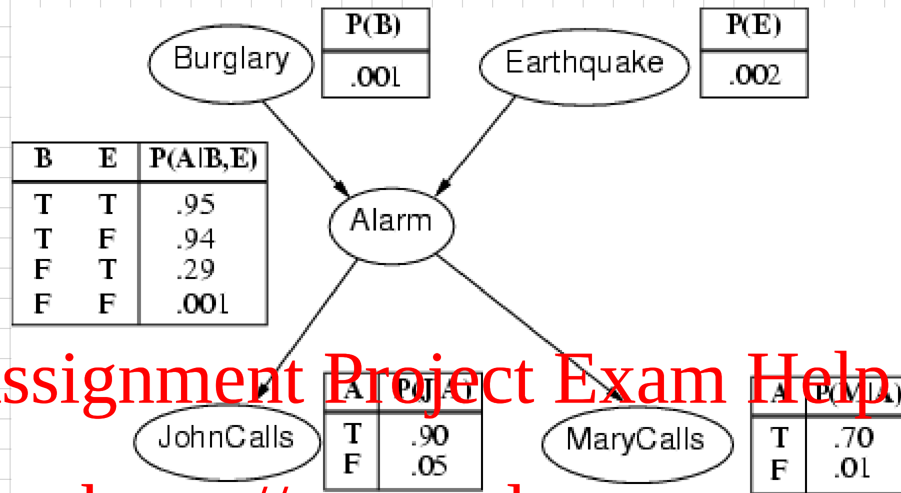
Inference in Bayesian networks

Exact Inference

A query P(X | e) can be answered by computing the sums of products of conditional probabilities over the hidden variables y.

$$P(X \mid e) = \alpha \sum_y P(X, e, y)$$

Exact Inference



$P(B \mid j,m) = \alpha \sum_e \sum_a P(B, j, m, e, a)$

$P(b \mid j,m) = \alpha \sum_e \sum_a P(b)P(e)P(a|b,e)P(j|a)P(m|a)$

Computational complexity is $O(n2^n)$

Singly-connected/polytrees, even $O(n)$

Inexact Inference

- Direct sampling:

  Sample each variable turn to generate event, generate probabilities from the sample proportions

- Rejection sampling

  Reject samples that do not match the evidence

- Likelihood weighting (importance sampling)

  Fix the evidence variables and sample from nonevidence, weighted by the likelihood of the evidence

- Markov Chain Monte Carlo (MCMC, Gibbs Sampling)

  Generate next sample by making a random change to the previous conditioned on the Markov blanket

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct