Belief networks

Chapter 15.1–2

Outline

◊ Conditional independence

◊ Bayesian networks: syntax and semantics

◊ Exact inference

◊ Approximate inference

Independence

Two random variables A B are (absolutely) <u>independent</u> iff

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

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

e.g., A and B are two coin tosses

If n Boolean variables are independent, the full joint is

$$\mathbf{P}(X_1, \ldots, X_n) = \Pi_i \mathbf{P}(X_i)$$

hence can be specified by just n numbers

Absolute independence is a very strong requirement, seldom met

Conditional independence

Consider the dentist problem with three random variables:
$Toothache$, $Cavity$, $Catch$ (steel probe catches in my tooth)

The full joint distribution has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
(1) $P(Catch|Toothache, Cavity) = P(Catch|Cavity)$
i.e., $Catch$ is conditionally independent of $Toothache$ given $Cavity$

The same independence holds if I haven't got a cavity:
(2) $P(Catch|Toothache, \neg Cavity) = P(Catch|\neg Cavity)$

Conditional independence contd.

Equivalent statements to (1)

(1a) $P(Toothache|Catch, Cavity) = P(Toothache|Cavity)$ Why??

(1b) $P(Toothache, Catch|Cavity) = P(Toothache|Cavity)P(Catch|Cavity)$
Why??

Full joint distribution can now be written as

$$\mathbf{P}(Toothache, Catch, Cavity) = \mathbf{P}(Toothache, Catch|Cavity)\mathbf{P}(Cavity)$$
$$= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$$

i.e., $2 + 2 + 1 = 5$ independent numbers (equations 1 and 2 remove 2)

Conditional independence contd.

Equivalent statements to (1)

(1a) $P(Toothache|Catch, Cavity) = P(Toothache|Cavity)$ Why??

$$
\begin{aligned}
&P(Toothache|Catch, Cavity)\\
&\quad = P(Catch|Toothache, Cavity)P(Toothache|Cavity)/P(Catch|Cavity)\\
&\quad = P(Catch|Cavity)P(Toothache|Cavity)/P(Catch|Cavity) \text{ (from 1)}\\
&\quad = P(Toothache|Cavity)
\end{aligned}
$$

(1b) $P(Toothache, Catch|Cavity) = P(Toothache|Cavity)P(Catch|Cavity)$
Why??

$$
\begin{aligned}
&P(Toothache, Catch|Cavity)\\
&\quad = P(Toothache|Catch, Cavity)P(Catch|Cavity) \text{ (product rule)}\\
&\quad = P(Toothache|Cavity)P(Catch|Cavity) \text{ (from 1a)}
\end{aligned}
$$

Belief 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:
$\mathbf{P}(X_i|Parents(X_i))$

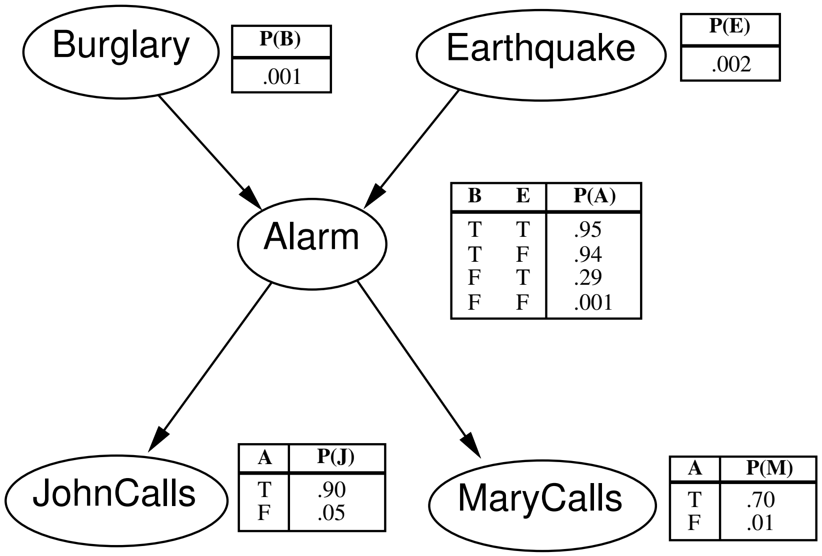
In the simplest case, conditional distribution represented as a <u>conditional probability table</u> (CPT)

AIMA Slides ©Stuart Russell and Peter Norvig, 1998

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: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
Network topology reflects "causal" knowledge:



Note: $\leq k$ parents $\Rightarrow O(d^k n)$ numbers vs. $O(d^n)$

Semantics

"Global" semantics defines the full joint distribution as the product of the local conditional distributions:

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

e.g., $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$ is given by??

$$=$$

Semantics

"Global" semantics defines the full joint distribution as the product of the local conditional distributions:

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

e.g., $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$ is given by??

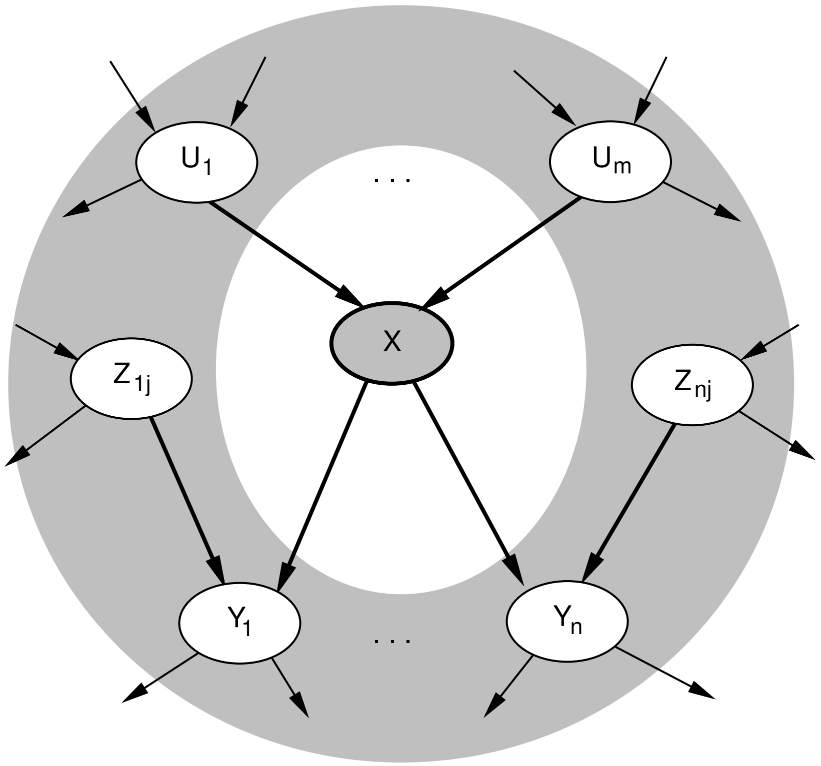
$= P(\neg B)P(\neg E)P(A|\neg B \wedge \neg E)P(J|A)P(M|A)$

"Local" semantics: each node is conditionally independent of its nondescendants given its parents

Theorem: Local semantics $\Leftrightarrow$ global semantics

Markov blanket

Each node is conditionally independent of all others given its
Markov blanket: parents + children + children's parents



Constructing belief networks

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

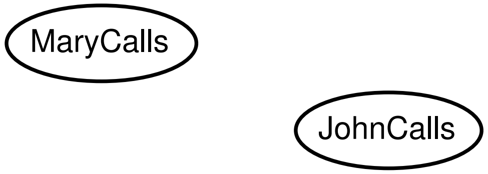
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i|Parents(X_i)) = \mathbf{P}(X_i|X_1,\ \ldots,\ X_{i-1})$

This choice of parents guarantees the global semantics:

$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i|X_1,\ \ldots,\ X_{i-1})$ (chain rule)
$= \prod_{i=1}^{n} \mathbf{P}(X_i|Parents(X_i))$ by construction

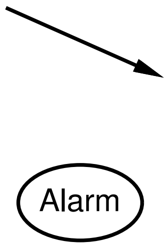
Example

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



$P(J|M) = P(J)$?

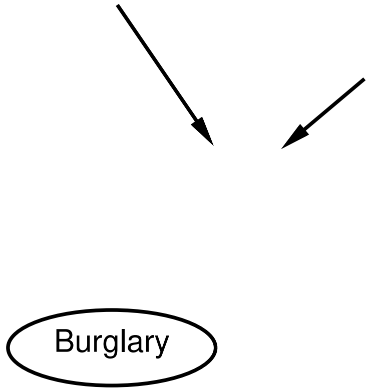
.



. No

$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$?

.



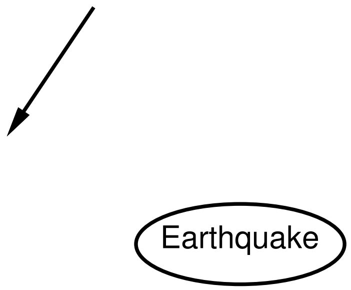
.

. No

$P(B|A,J,M) = P(B|A)$?

$P(B|A,J,M) = P(B)$?

.



.

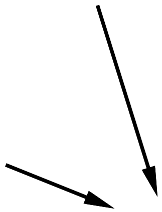
.

. Yes

. No

$P(E|B, A, J, M) = P(E|A)$?

$P(E|B, A, J, M) = P(E|A, B)$?

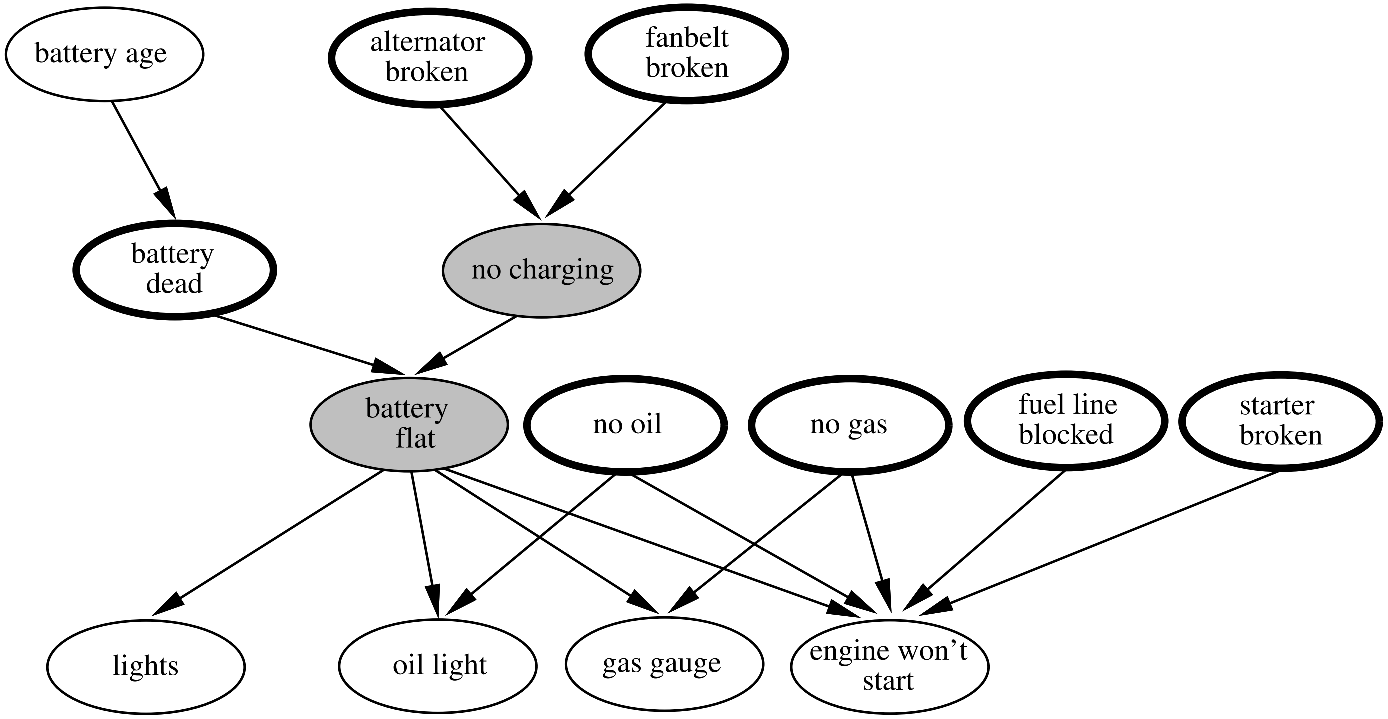


No

Yes

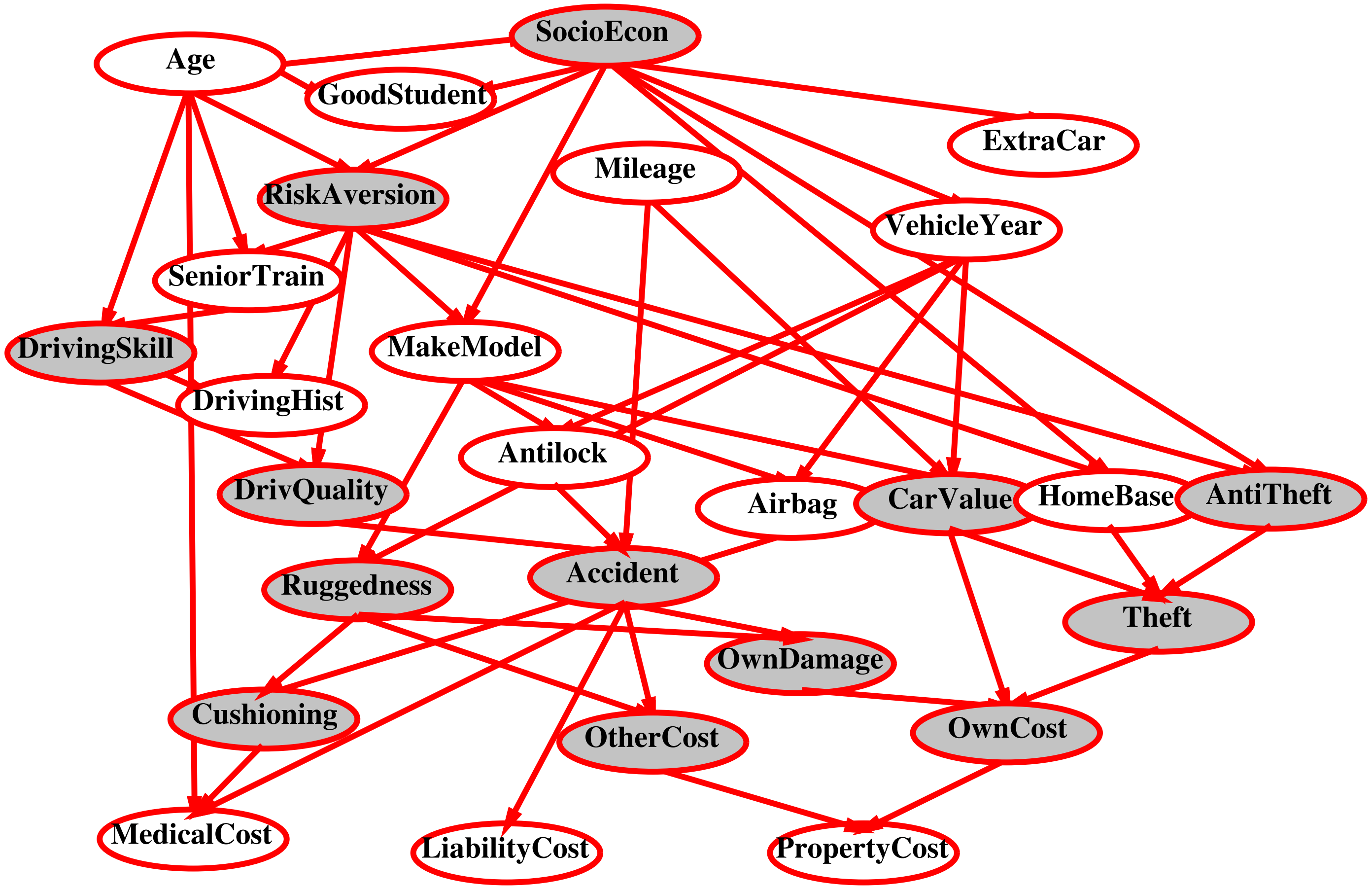
Example: Car diagnosis

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



Example: Car insurance

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



Compact conditional distributions

CPT grows exponentially with no. of parents
CPT becomes infinite with continuous-valued parent or child

Solution: canonical distributions that are defined compactly

Deterministic nodes are the simplest case:
$X = f(Parents(X))$ for some function f

E.g., Boolean functions
$NorthAmerican \;\Leftrightarrow\; Canadian \lor US \lor Mexican$

E.g., numerical relationships among continuous variables

$$\frac{\partial Level}{\partial t} = \text{ inflow + precipation - outflow - evaporation}$$

Compact conditional distributions contd.

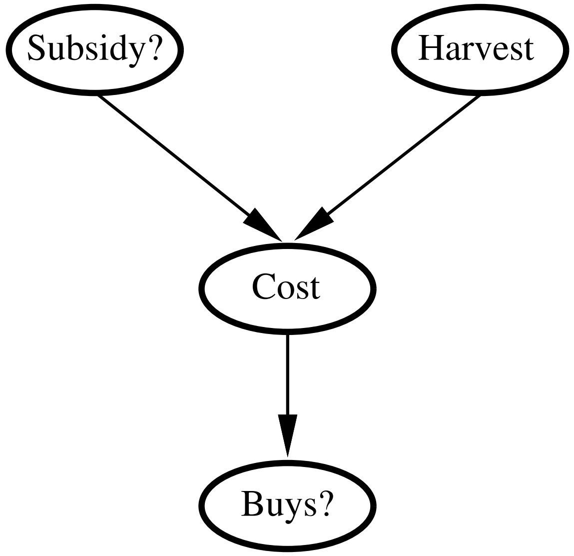
Noisy-OR distributions model multiple noninteracting causes
1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)
2) Independent failure probability q_i for each cause alone
$\Rightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \Pi_{i=1}^{j} q_i$

$Cold$	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 linear in number of parents

Hybrid (discrete+continuous) networks

Discrete ($Subsidy?$ and $Buys?$); continuous ($Harvest$ and $Cost$)



Option 1: discretization—possibly large errors, large CPTs

Option 2: finitely parameterized canonical families

1) Continuous variable, discrete+continuous parents (e.g., $Cost$)
2) Discrete variable, continuous parents (e.g., $Buys?$)

Continuous child variables

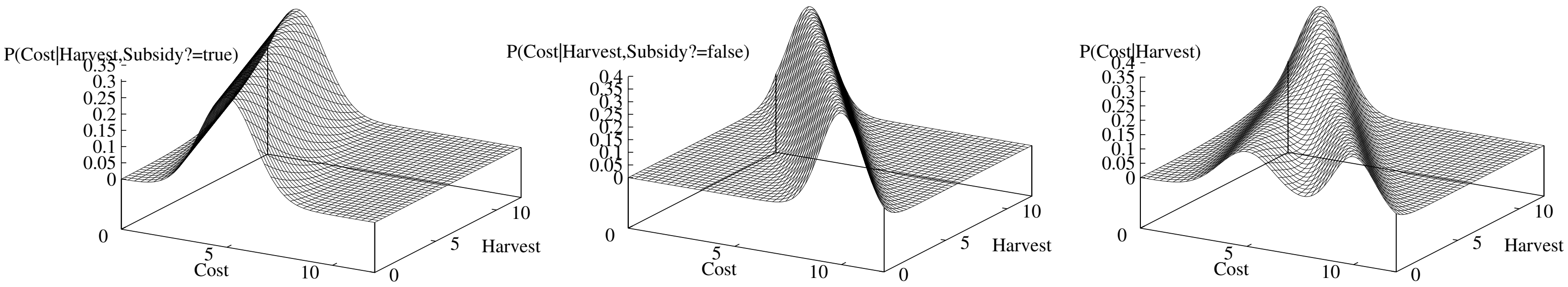
Need one conditional density function for child variable given continuous parents, for each possible assignment to discrete parents

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

$$
\begin{aligned}
&P(Cost = c|Harvest = h, Subsidy? = true)\\
&= N(a_t h + b_t, \sigma_t)(c)\\
&= \frac{1}{\sigma_t\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{c-(a_t h + b_t)}{\sigma_t}\right)^2\right)
\end{aligned}
$$

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 child variables

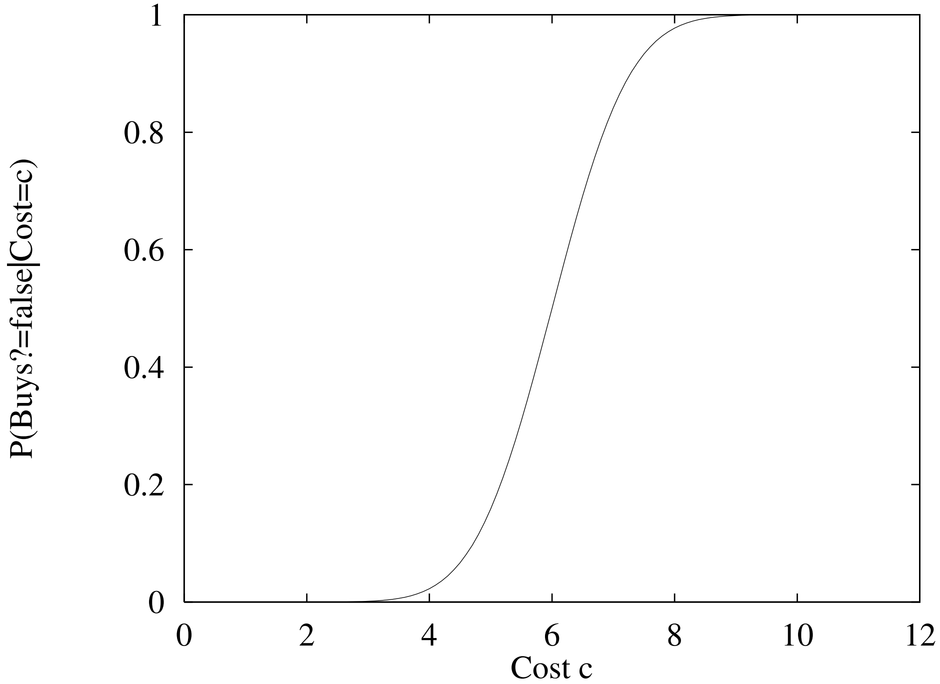


All-continuous network with LG distributions
$\Rightarrow$ full joint is a multivariate Gaussian

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

Discrete variable w/ continuous parents

Probability of $Buys?$ given $Cost$ should be a "soft" threshold:



Probit distribution uses integral of Gaussian:

$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$

$P(Buys? = true \mid Cost = c) = \Phi((-c + \mu)/\sigma)$

Can view as hard threshold whose location is subject to noise

Discrete variable contd.

Sigmoid (or logit) distribution also used in neural networks:

$$P(Buys? = true \mid Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}$$

Sigmoid has similar shape to probit but much longer tails: