USTC

Chapter 14a

Bayesian Networks

王子磊 (Zilei Wang)

Email: zlwang@ustc.edu.cn

http://vim.ustc.edu.cn

提纲

- 贝叶斯网络 (Bayesian networks)
  - 语法
  - 语义
- 参数化分布 (parameterized distributions) *

贝叶斯网络 (Bayesian networks)

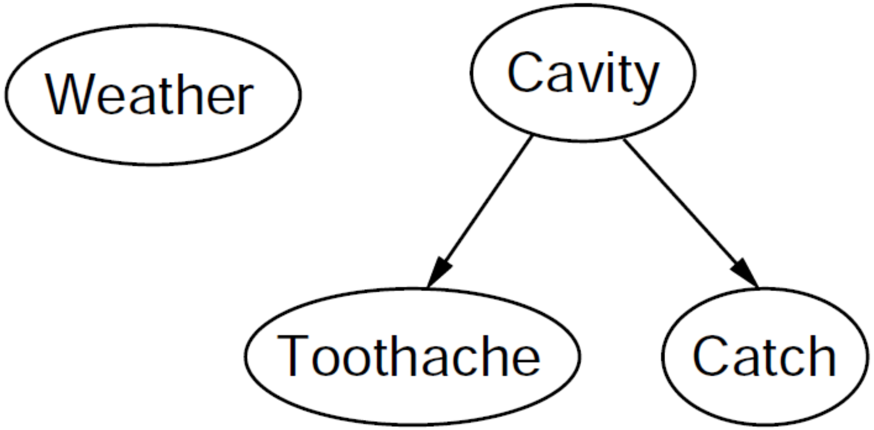
- 条件独立断言的一种简单图 (graphical) 表示，用于完全联合概率分布的紧致表达
- 语法：
  - 一个节点集合，每个节点对应一个随机变量
  - 一个有向无环图 DAG (link ≈ "directly influences")
  - 每个节点上一个条件概率（相对于父节点的）

$$P\,(X_i \mid Parents\,(X_i))$$

- 最简单情况下，条件概率可以表示为 条件概率表 (conditional probability table, CPT)，它给出了 X_i 相对于父节点值组合的概率分布

示例

- 网络拓扑 编码了 条件独立断言：



- *Weather* 与其他变量是独立的
- 给定 *Cavity* 条件下，*Toothache* 和 *Catch* 是条件独立的

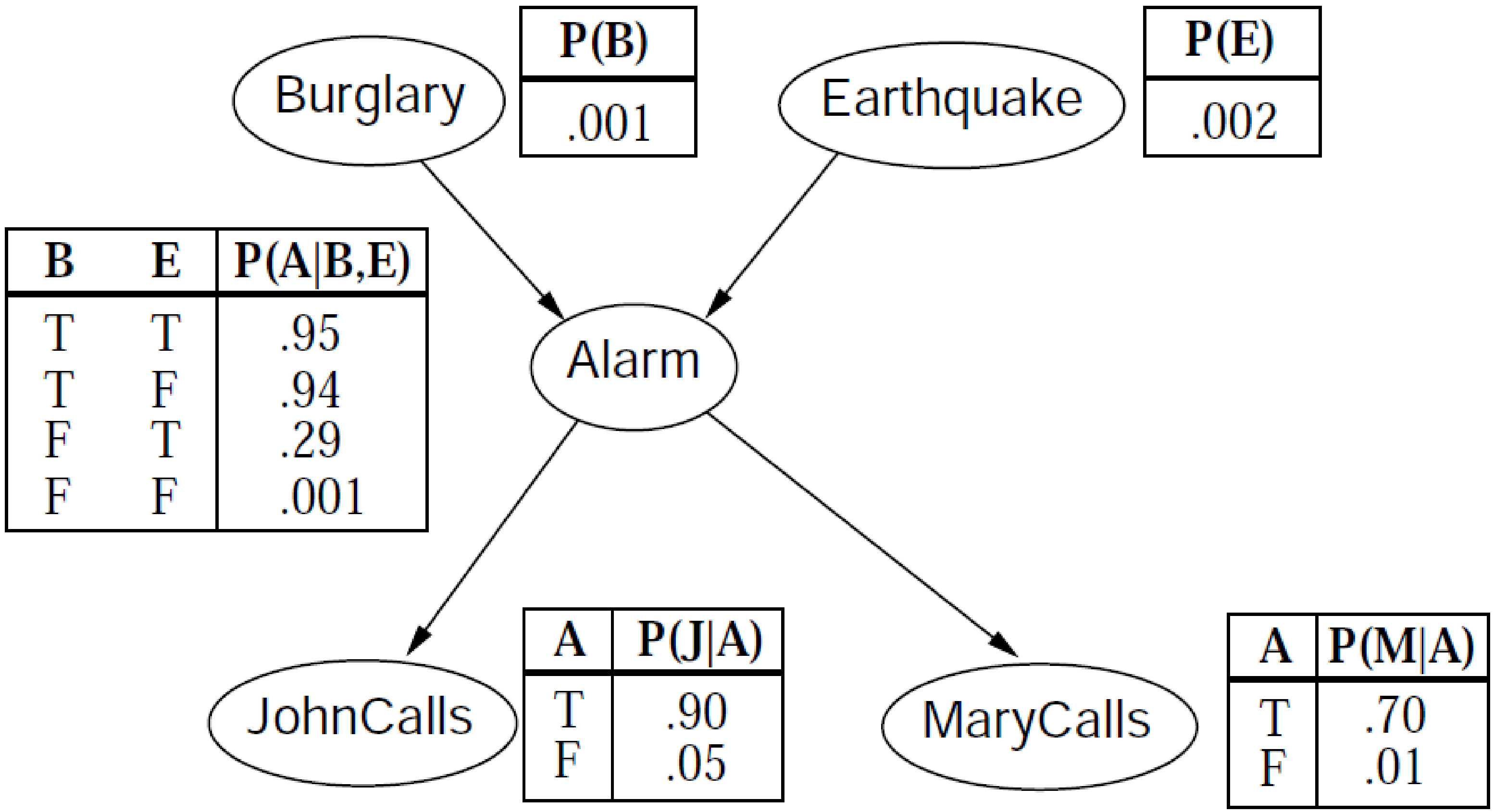
示例

- 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?

- 变量：*Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

- 网络拓扑反应了 因果关系 (causal) 知识
  - 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

示例



P(B)
.001

P(E)
.002

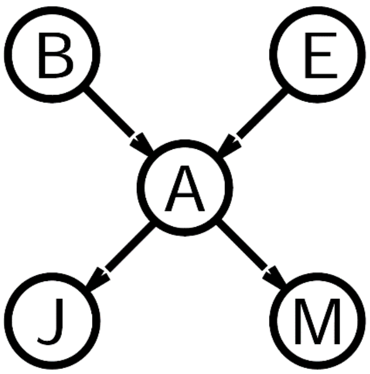
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

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

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

紧致性 (Compactness)

- 一个 Boolean 变量 X_i 相对于 k 个 Boolean 父节点的 CPT 有 2^k 行
  - 每行需要一个数值 p 表示 X_i = *true* 的条件概率
    (X_i = *false* 的条件概率为 $1-p$)

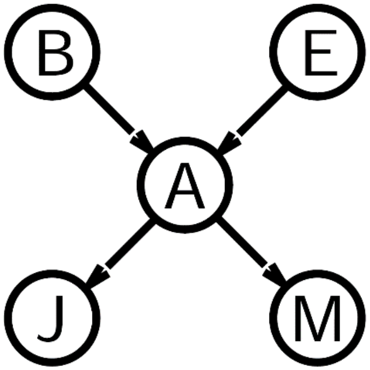


- 如果每个变量的父节点数不超过 k 个，完整的网络需要 $O(n \cdot 2^k)$ 个数值
  - i.e., 随 n 线性增长, vs. 完全联合概率分布的 $O(2^n)$
  - 对 burglary 网络，有 1 + 1 + 4 + 2 + 2 = 10 个数字 (vs. 2^5-1 = 31)

全局语义 (Global Semantics)

完全联合概率分布由局部条件概率分布的乘积进行定义（链式法则）

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$



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

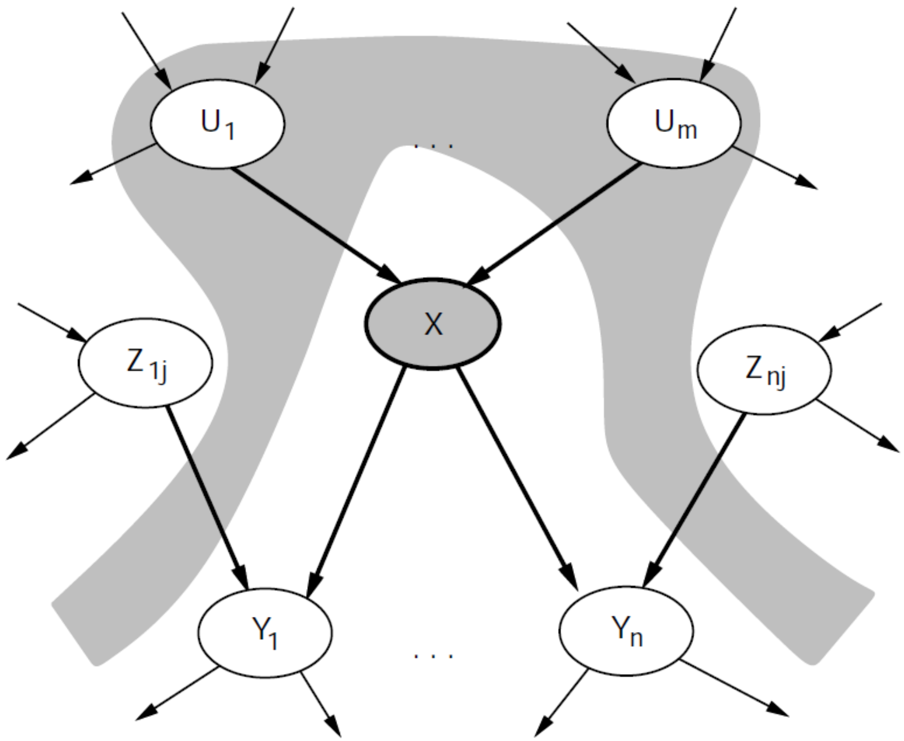
$$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$

$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$$

$$\approx 0.00063$$

局部语义 (Local Semantics)

局部语义：给定父节点，一个节点条件独立于它的其他祖先节点



定理：Local semantics $\Leftrightarrow$ global semantics

局部语义 (Local Semantics)

局部语义：给定父节点，一个节点条件独立于它的其他祖先节点

给定 x_i 的任一其他祖先节点 x_j

$$P\left(x_i, x_j \middle| Parent(x_i)\right) = P(x_i|x_j, Parent(x_i))P(x_j|Parent(x_i)) \quad \text{(联合概率定义)}$$

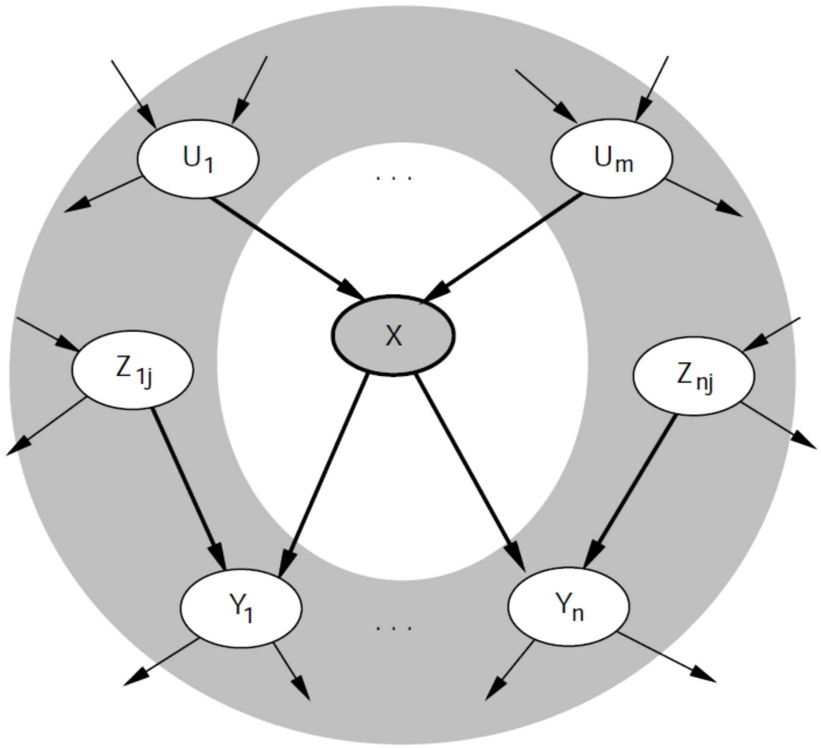
$$= P(x_i|Parent(x_i))P(x_j|Parent(x_i)) \quad \text{(贝叶斯网络)}$$

➔ 条件独立性

马尔科夫覆盖 (Markov blanket)

给定它的 Markov blanket， 该节点条件独立于网络中的所有其他节点

Markov blanket = parents + children + children’s parents



- http://maider.blog.sohu.com/307235658.html

马尔科夫覆盖 (Markov blanket)

给定它的 Markov blanket， 该节点条件独立于网络中的所有其他节点

给定 x_i 的Markov 覆盖 $MB(x_i) = (Parent(x_i), Son(x_i), Parant(Son(x_i)))$，
以及其他所有节点 $\boldsymbol{z}$

$$P(x_i, MB(x_i), \mathbf{z}) = P(x_i|Parent(x_i)) \prod P(MB_j(x_i)|Parent(MB_j(x_i))) \prod P(z_k|Parent(z_k))$$

$$P(x_i|MB(x_i), \mathbf{z}) = P(x_i, MB(x_i), \mathbf{z}) / \int P(x_i, MB(x_i), \mathbf{z}) \mathrm{d}x_i$$ (条件概率定义)

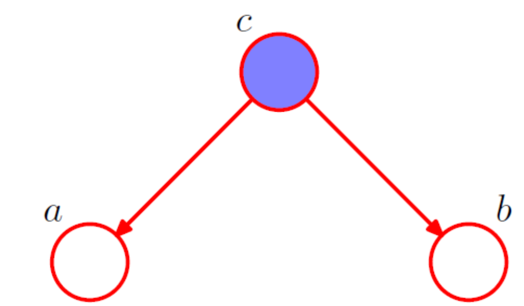
$$= {P(x_i|Parent(x_i)) \prod P(MB_j(x_i)|Parent(MB_j(x_i)))} \Big/ {P(x_i|Parent(x_i)) \prod P(MB_j(x_i)|Parent(MB_j(x_i)))}$$

$$= P(x_i|MB(x_i))$$ (代入消除)

➔ 条件独立性

马尔科夫覆盖 (Markov blanket) *

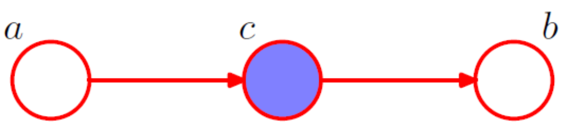
d 分离 (d-separation)：DAG 的三种基本结构



分连结构：

a 和 *b* 关于 *c* 条件独立

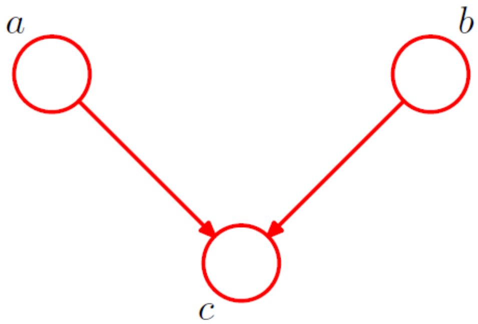
父节点



顺连结构：

a 和 *b* 关于 *c* 条件独立

子节点



汇连结构：

a 和 *b* 边缘独立，若已知结果 *c* 反而不独立

子节点的父节点

构建贝叶斯网络

1. 选择一组排好序的随机变量 $X_1, \ldots, X_n$
2. For $i = 1$ to n
   - 将 X_i 添加到网络中
   - 从 $X_1, \ldots, X_{i-1}$ 中选择它的父节点，使得

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

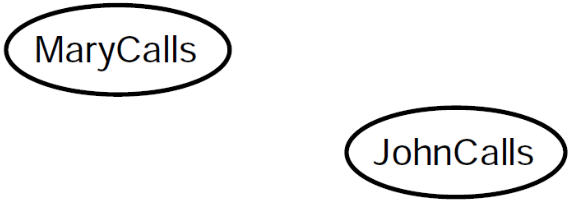
这种父节点选择方法保证了贝叶斯网络的全局语义：

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) \quad \text{(chain rule)}$$

$$= \prod_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i)) \quad \text{(by construction)}$$

构建示例

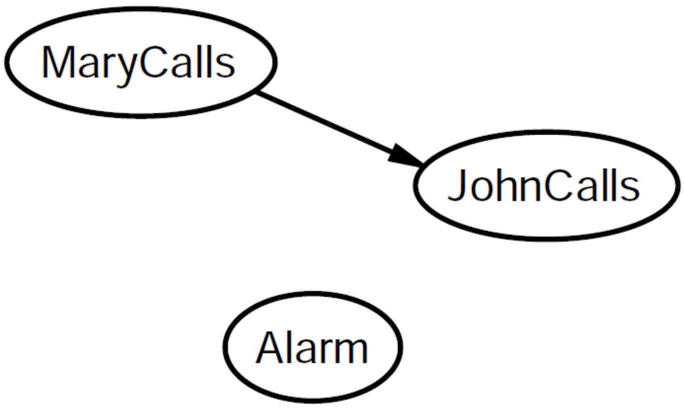
- 假设我们选定了 M, J, A, B, E 的次序



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

构建示例

❖ 假设我们选定了 M, J, A, B, E 的次序

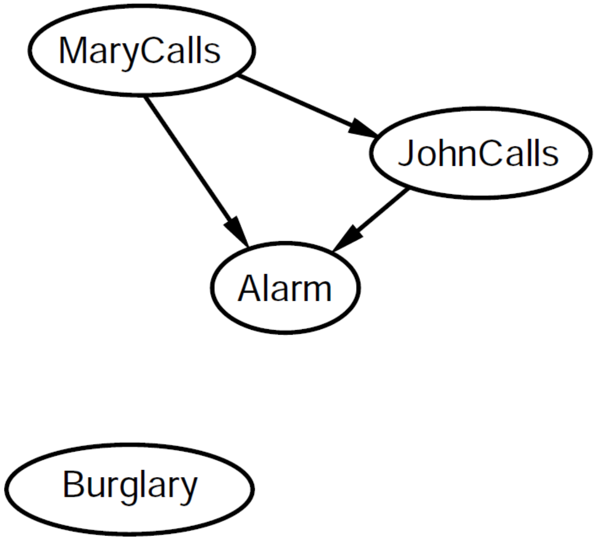


$\boldsymbol{P}(J \mid M) = \boldsymbol{P}(J)$? **No**

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

构建示例

- 假设我们选定了 M, J, A, B, E 的次序



$\boldsymbol{P}(J \mid M) = \boldsymbol{P}(J)$? **No**

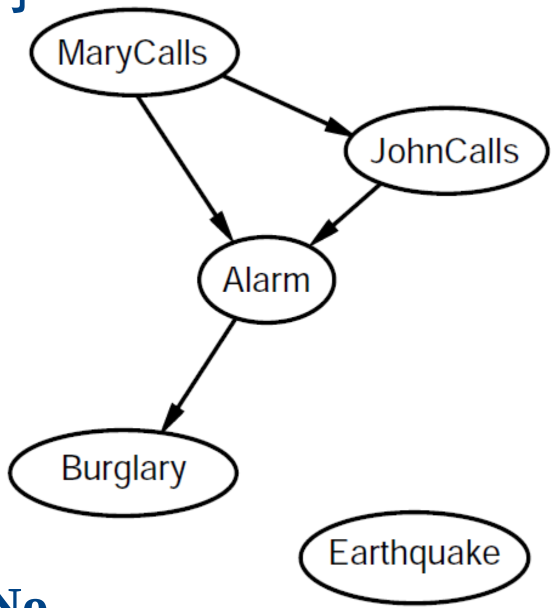
$\boldsymbol{P}(A \mid J, M) = \boldsymbol{P}(A \mid J)$? $\boldsymbol{P}(A \mid J, M) = \boldsymbol{P}(A)$? **No**

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

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

构建示例

❖ 假设我们选定了 M, J, A, B, E 的次序



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

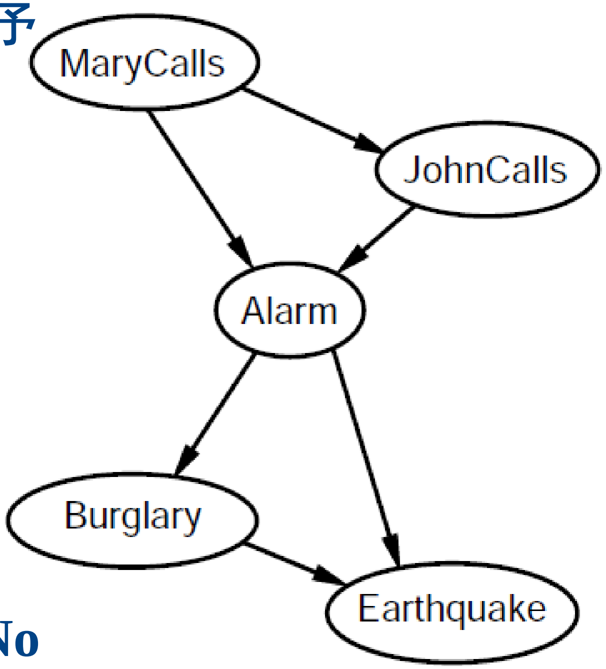
$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**

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

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

构建示例

- 假设我们选定了 M, J, A, B, E 的次序



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

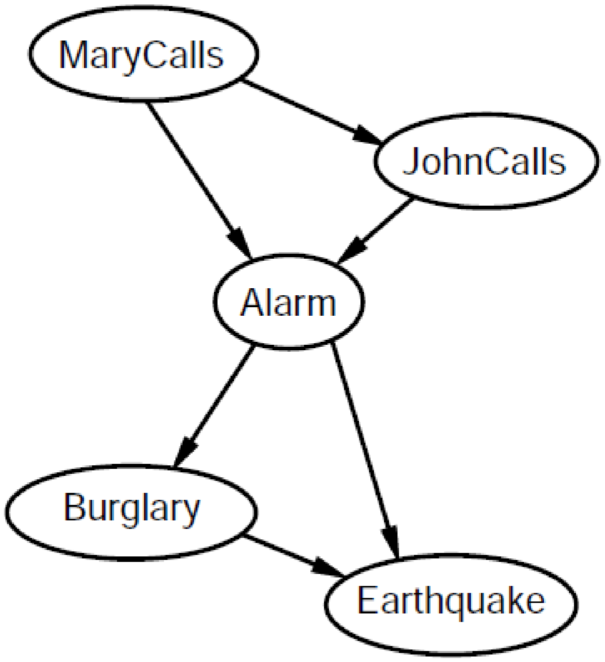
$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**

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

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

构建示例

- 在非因果 (noncausal) 方向上判断条件独立性是困难的
- 对人类而言，因果模型和条件独立性看起来是根深蒂固的 (hardwired)!
- 这个网络是相对不紧致的：需要 1 + 2 + 4 + 2 + 4 = 13 个数值



紧致的条件分布

- CPT 随着父节点数呈指数级增长
- 对连续变量父节点或子节点，CPT 将变成无穷大

- 解决方案：规范分布 (canonical distribution)
  - 确定性节点是最简单的情况：

    $$X = f(Parents(X)) \text{ for some function } f$$

  - E.g., Boolean functions

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

  - E.g., 连续变量之间的数值关系

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

紧致的条件分布

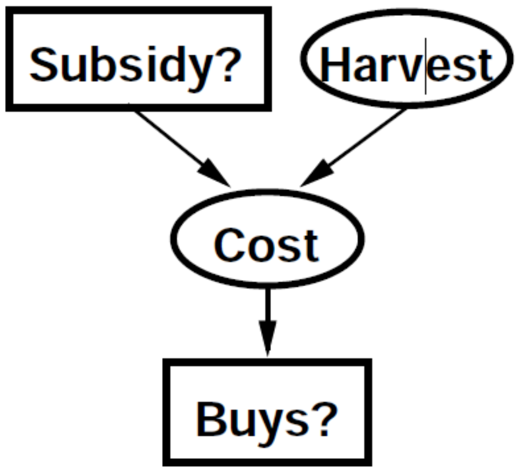
- 不确定情况：噪声或 (noisy-or) 逻辑关系能够刻画多个不交互的子句
  - 父节点 $U_1 \ldots U_k$ 列出了所有可能的原因（可以增加遗漏节点）
  - 每个父节点具有独立的失败概率 q_i

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

$Cold$	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	$0.012 = 0.6 \times 0.2 \times 0.1$

混合网络（离散+连续）*

❖ 离散量 (*subsidy?* 和 *Buys?*) + 连续量 (*Harvest* 和 *Cost*)



❖ 选项1：离散化 —— 可能导致大的误差和庞大的 CPT

❖ 选项2：有限参数的规范化表示

1) 连续变量 (e.g., *Cost*)：离散+连续的父节点

2) 离散变量 (e.g., *Buys?*)：连续的父节点

连续子变量 *

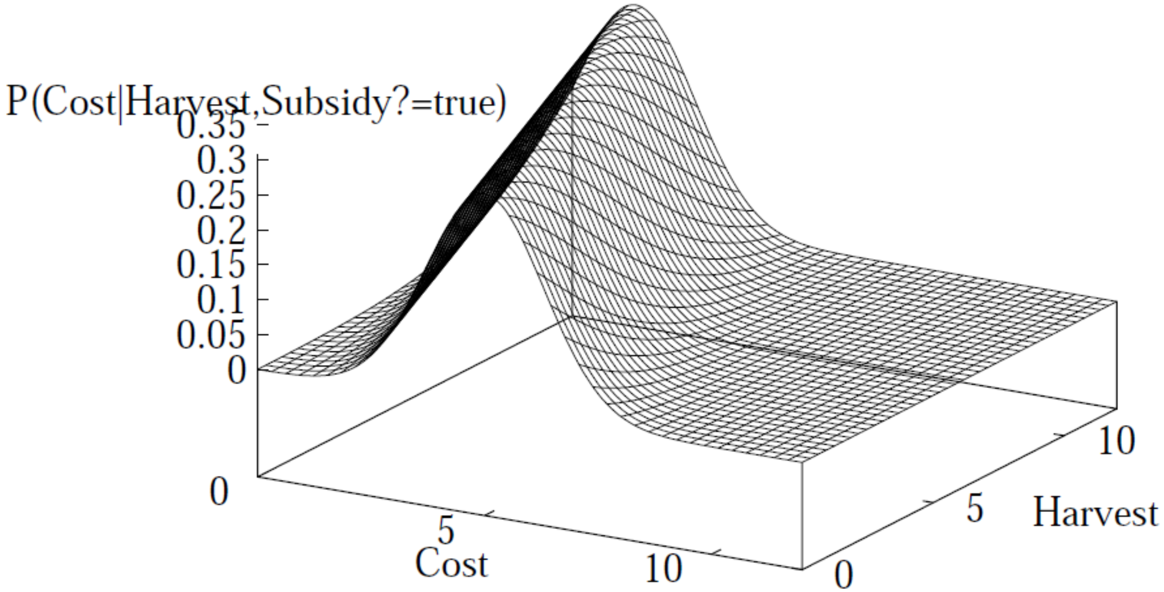
- 需要一个条件概率密度函数
  - 对应单个可能的离散数值，刻画了给定连续父节点下，子节点的概率分布
  - 最常用的是线性高斯 (Linear Gaussian, LG) 模型，e.g.,

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

表示了 *Cost* 随着 *Harvest* 的线性变化，这儿方差是固定的

  - 全范围内的线性关系是不合理的，不过，如果 *Harvest* 的可能范围很小，它是适用的

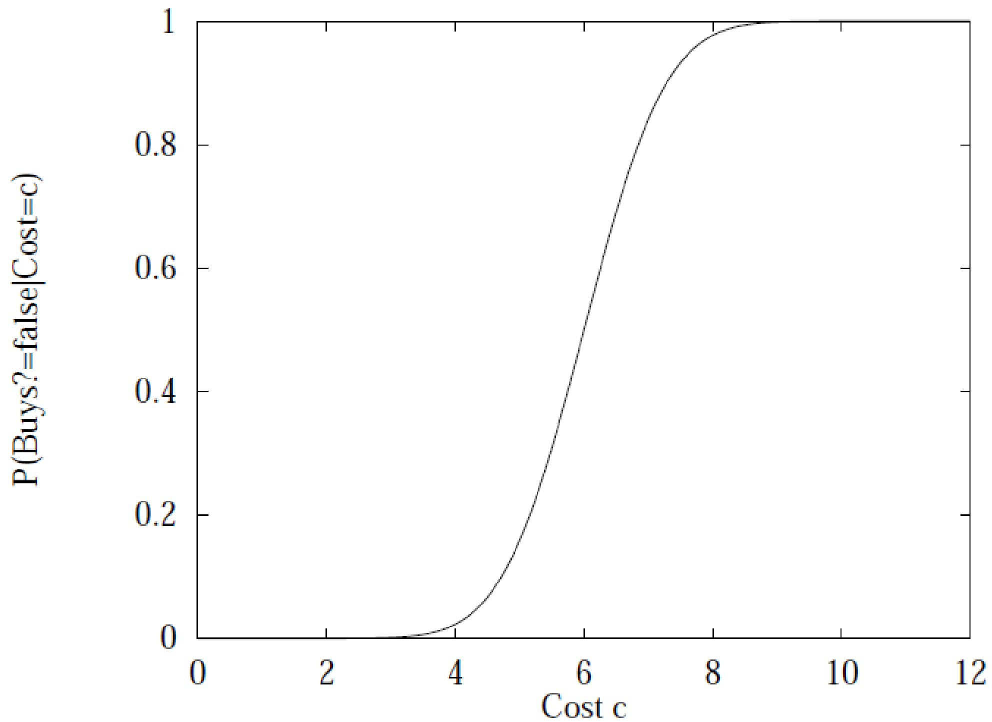
连续子变量 *



- 具有 LG 分布的全连续网络
  - 完全联合分布是一个多变量高斯分布
- 离散 + 连续的 LG 网络是条件高斯网络
  - 针对每个离散量的组合，是其他连续量的多变量高斯分布

离散变量 w/o 连续父节点 *

- 给定 *Cost*，*Buys?* 的概率应该是 "soft" threshold 的

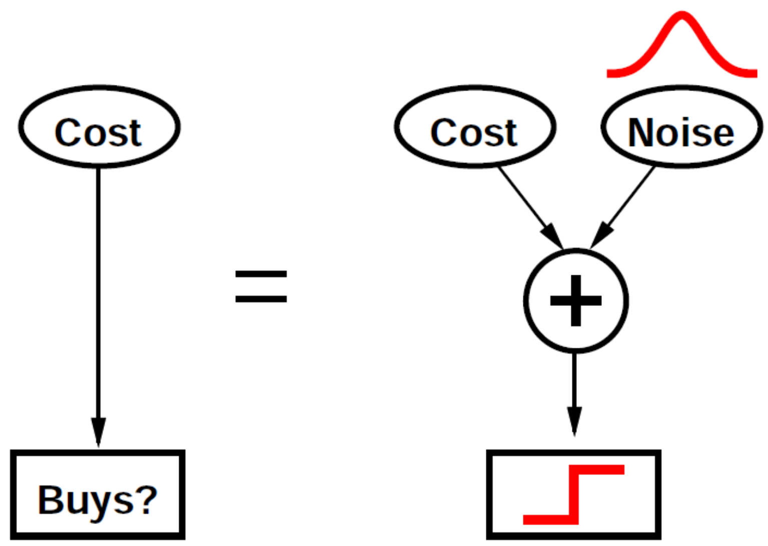


- Probit 分布使用 Gaussian 积分

$$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$$
$$P(Buys? = true \mid Cost = c) = \Phi((-c+\mu)/\sigma)$$

为什么使用 Probit？*

- 它具有合适的概率分布曲线
- 可以看作为位置被噪声干扰的 hard threshold

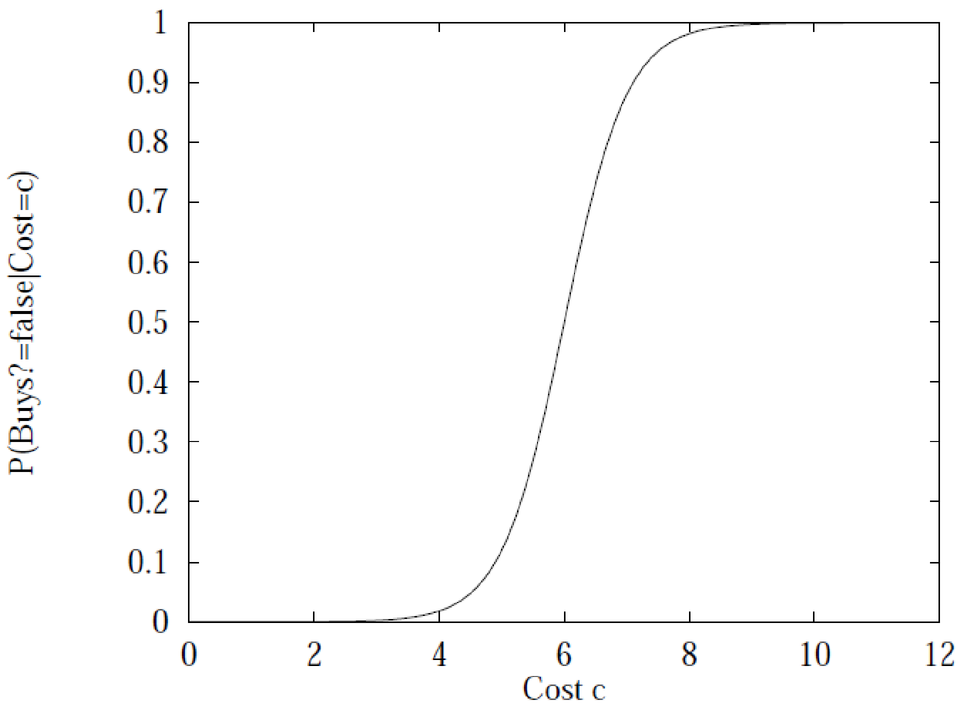


离散变量 *

- Sigmoid (or logit) 分布在神经网络中经常使用

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

- Sigmoid 与 Probit 具有相似的形状，但具有更长的尾



总结

- 贝叶斯网络为条件独立性（因果推理）提供了一种天然表达方法
- 拓扑 + CPTs = 联合概率分布的紧致化表达
- 对领域专家来说，通常是容易构建的

- 规范化分布 (e.g., noisy-or) = CPTs 的紧致化表达
- 连续变量 ➔ 参数化分布 (e.g., linear Gaussian) *

谢谢聆听！