Artificial Intelligence

8.不确定知识与推理b

对应课本13-15章

贝叶斯网络

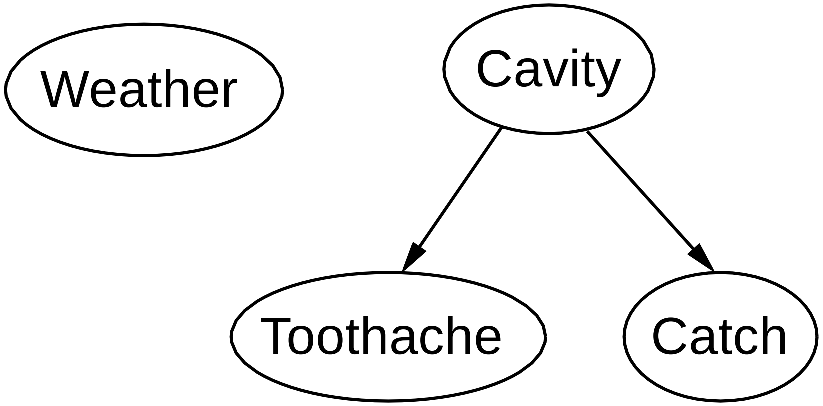
- 语法
- 语义
- 参数化分布

贝叶斯网络

- 贝叶斯网络是由简单的、图形化的符号组成，用于做条件独立的推理，因此其比完整的联合分布表示更为紧凑
- 语法表示：
  - 一组节点，每个变量一个
  - 有向无环图（链接 ≈ “直接影响”）
  - 给定每个节点的父节点的条件分布 $\mathrm{P}(X_i | Parents(X_i))$
- 在最简单的情况下，条件分布表示为条件概率表(CPT)，给出每个父值组合的X_i上的分布

例子

- 网络拓扑编码条件独立性断言：



- *Weather* 相对其它变量是独立的
- *Toothache* 和 *Catch* 和给定的 *Cavity* 是条件独立

家庭报警器

所以其实美国西部的 House，和中国很不一样，这里没有高楼大厦，建筑扁扁平平宽度很大，甚至越有钱的人越住在荒不溜秋的山上，偶尔有野生动物爬进院子，一点都不奇怪。这也带来一个问题：美国人怎么管理住房安全？

就以我这个月住朋友家的经历为例，请看下面指示图，它是装在家门口的一个白色机器：

如果睡得早，通常我会在凌晨五点半去星巴克。出门前，我先要在这个机器中间的“键盘”处，按几个数字键。这是一串由女主人自设的密码，如果不按就出门，随即家里会警报大作，因为我已经动了门窗，它会假设你是不知密码的外来闯入者。

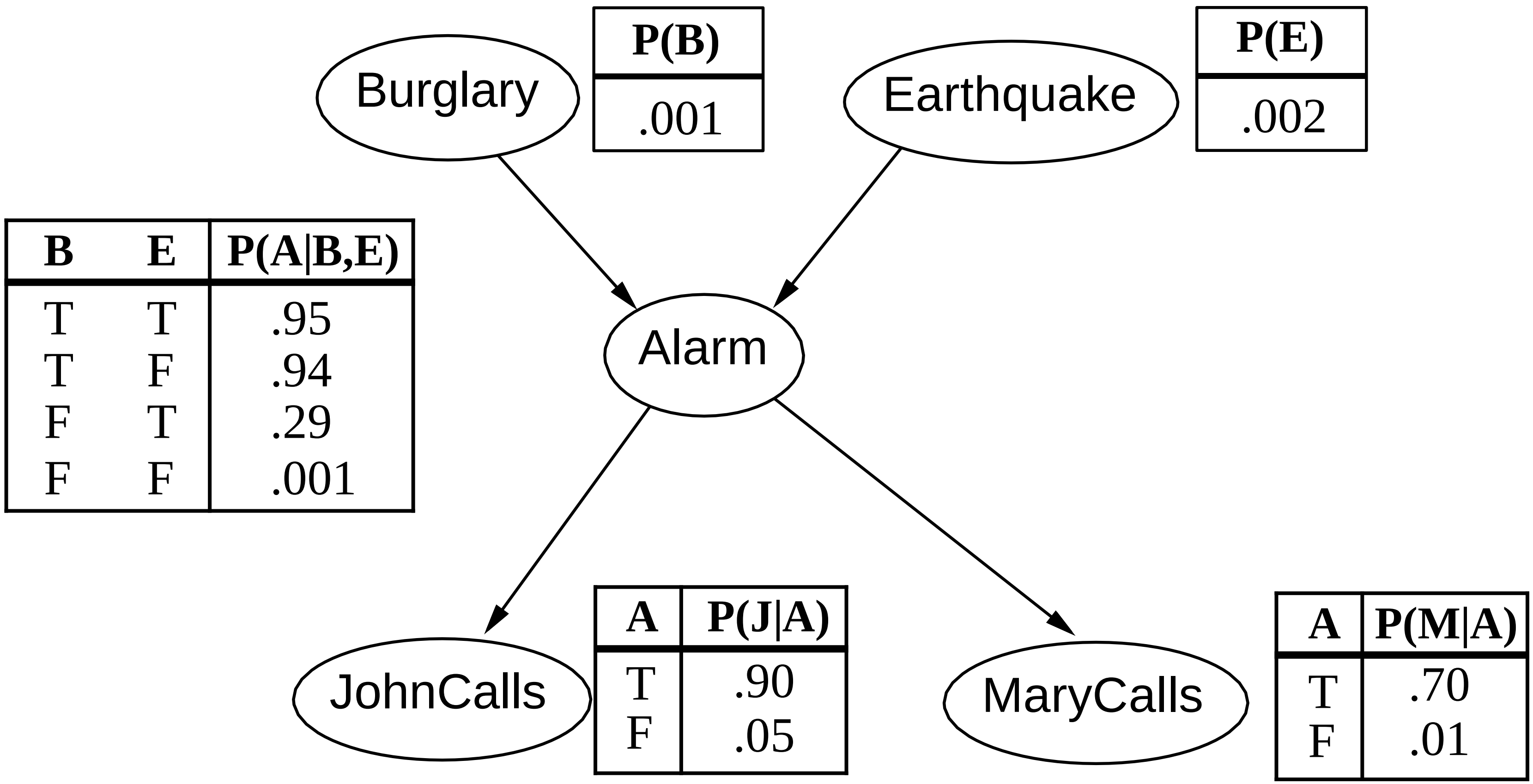
初到朋友家时，我曾在地下室推开落地窗走到大山坡，马上就被楼上朋友知道了。“你去干嘛啦？”“我去看鸵鸟了。”所以任何门窗有动静，我朋友即使坐客厅一动不动，也能知道具体是哪个部位发生异常。

但按完“数字键”还没完，我还要按数字键上面左边的 “Doors+Windows”，这表示：我在解锁、开门后，再把门窗锁上，因为我要让房子重新处于一个被 Home Security System 监控状态。

家庭报警器

- 我在上班，邻居约翰打电话来说我的家里的警报响了，但是邻居玛丽没有打电话。有时它是由小地震引起的。有窃贼吗？
- 变量：*Burglar*，*Earthquake*，*Alarm*，*JohnCalls*，*MaryCalls*
- 我们可以用网络拓扑来定义“因果性”的知识：
  - 夜贼能触发报警器
  - 地震可以触发警报
  - 警报会让玛丽打电话
  - 警报会让约翰打电话

家庭报警器



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

紧凑度

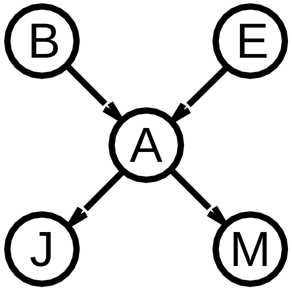
- 对于布尔变量 X_i，它有 k 个布尔型父代，那么将一共有 2^k 个不同的父代值的组合，称这一指标为紧凑度（CPT，Compact）
- 每一组父代值组合都有一个概率值 p 来使得 $X_i = true$
- 同样，对于 $X_i = false$，父代值的组合概率值是 $1 - p$
- 如果每个变量的父代不超过 k 个
- 那么完整的网络需要个 $O(n \cdot 2^k)$ 概率值
- 也就是说如果网络节点规模按照 n 线性增长，则联合概率的复杂度将按照 $O(2^n)$ 增长
- 对于警报网络，一共有个概率值 $1 + 1 + 4 + 2 + 2 = 10$
  - 作为比较 $2^5 - 1 = 31$

全局语义

- 全局语义将完全联合分布定义为局部条件分布的乘积：

$$P(x_1, \dots, xn) = \prod_{i=1}^{n} P(x_i \mid parents(Xi))$$

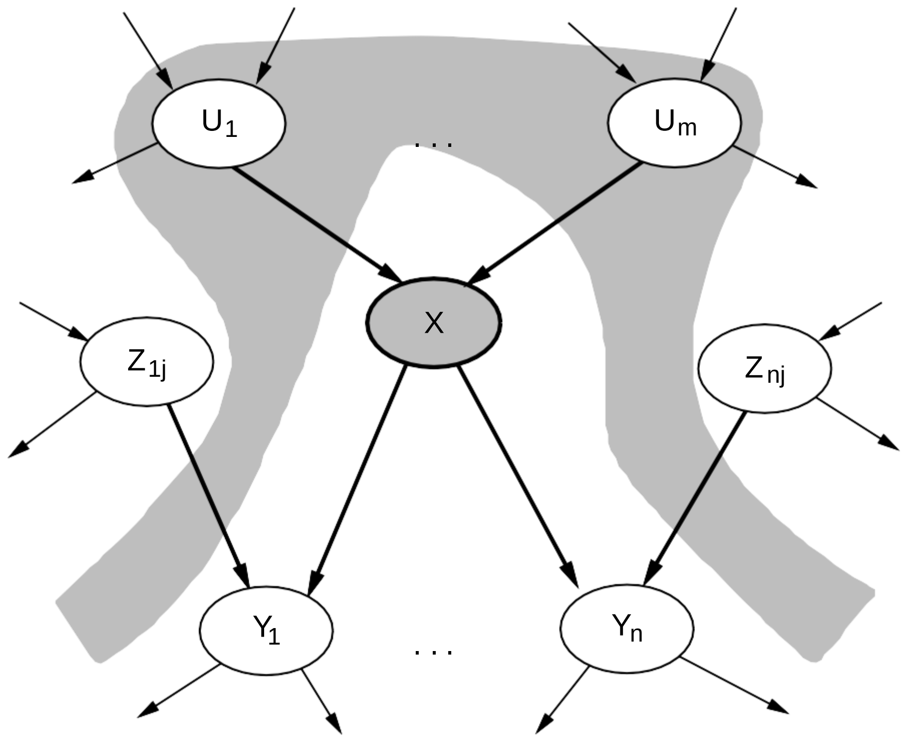
$$\begin{aligned} P(j \land m \land a \land \neg b \land \neg e) \\ &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$



$$P(j \land \neg m \land b) = P(j|a)P(m|a)P(a|b) + P(j|\neg a)P(m|\neg a)\ P(\neg a|b)\ P(b)$$

局部语义

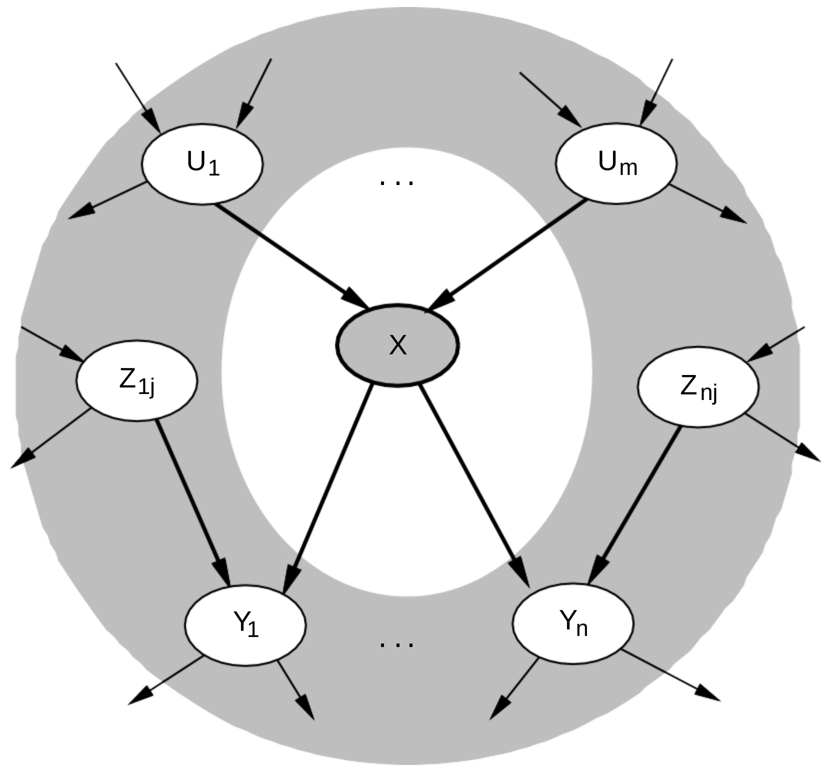
- 局部语义：每个节点都有条件地独立于给定父节点的非子节点



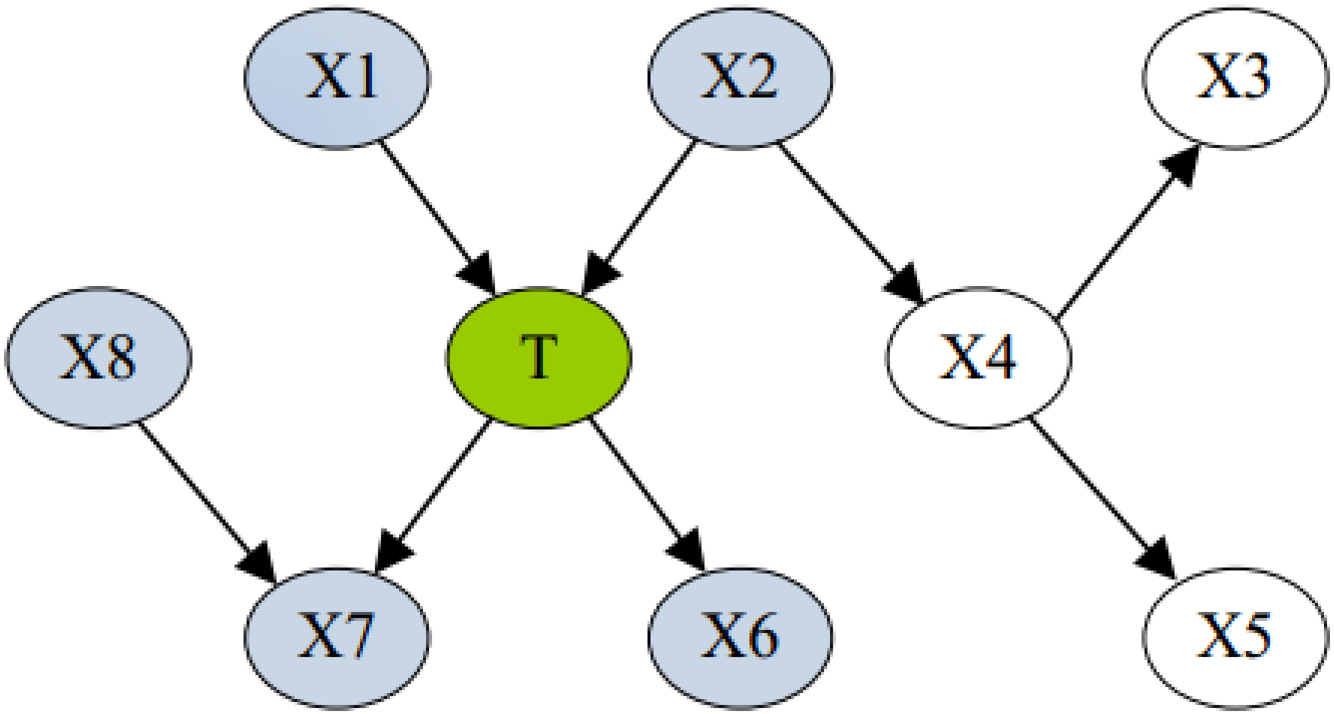
- 定理：局部语义⇔全局语义

马尔科夫毯

- 每个节点都有条件地独立于所有其他节点，给出其马尔科夫毯：父节点+子节点+子节点的父节点



马尔科夫毯示例



马尔科夫毯作用

- 在一个随机变量的全集U中，目标变量T和其非马尔科夫毯变量条件独立
- 这意味着对于T而言，所有非马尔科夫毯变量都是冗余的
- 如果要了解目标变量的分布情况，仅需要了解马尔科夫毯的信息即可，不需要对整个数据集进行了解
- 这项特性主要用于在特征空间中选择特征，利用马尔科夫毯可以去除两种类型的特征：不相关特征和冗余特征
- 马尔科夫毯是特征冗余性分析的一种常用工具

构建贝叶斯网络

- 需要一种方法，使一系列条件独立性的本地可测试断言能够保证所需的全局语义

1. 选择变量 $X_1, \ldots, X_n$ 的一种排序
2. For i=1 to n
3. 将 X_i 加入网络
4. 选择 $X_1, \ldots, X_{i-1}$ 的父代节点，使得

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

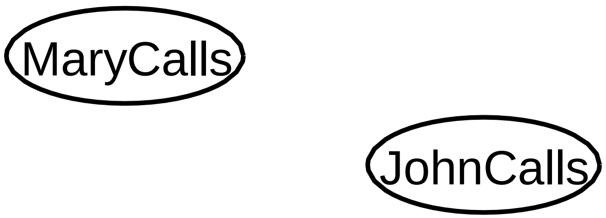
- 这种对父代节点的选择方式保证了全局语义：

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \quad \text{(链式法则)}$$

$$= \prod_{i=1}^{n} P(X_i|Parents(X_i)) \quad \text{(变换后)}$$

构建贝叶斯网络示例

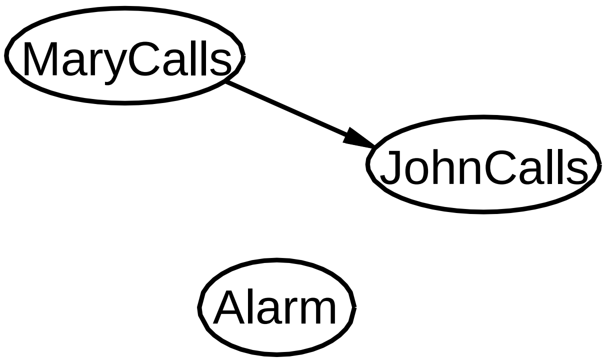
- 假设我们选择顺序按照M , J, A, B, E



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

构建贝叶斯网络示例

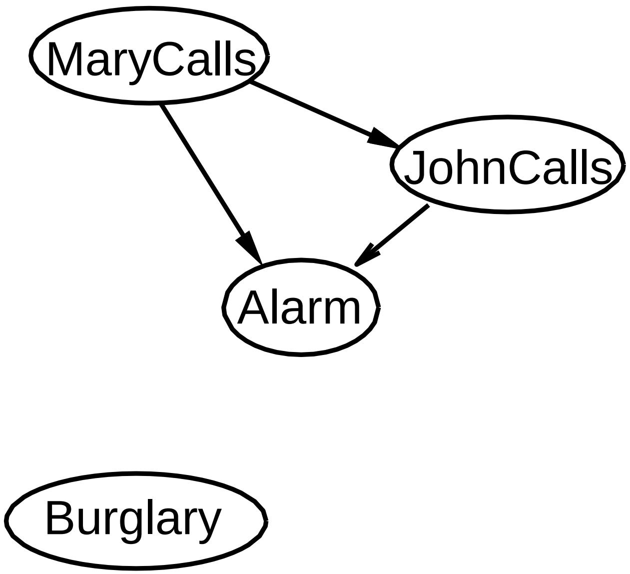
- 假设我们选择顺序按照M , J, A, B, E



- $P(J|M) = P(J)$? No
- $P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$?

构建贝叶斯网络示例

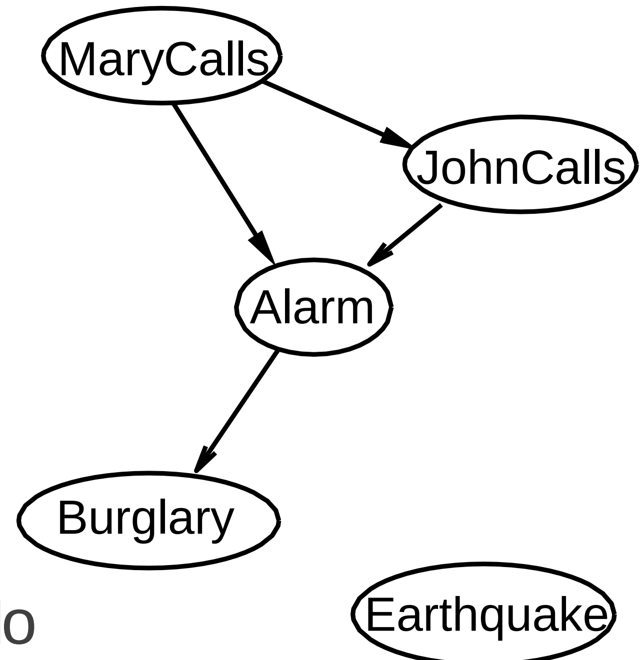
- 假设我们选择顺序按照M , J, A, B, E



- $P(J|M) = P(J)$? No
- $P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No
- $P(B|A,J,M) = P(B|A)$?
- $P(B|A,J,M) = P(B)$?

构建贝叶斯网络示例

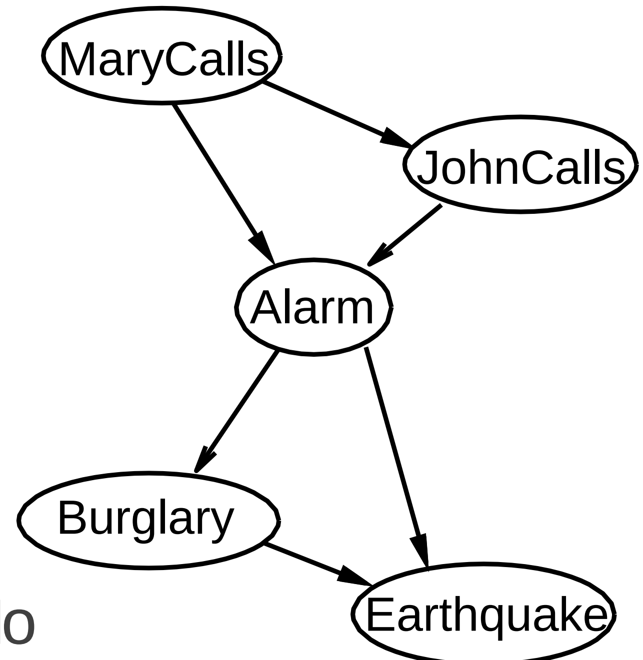
- 假设我们选择顺序按照M , J, A, B, E



- $P(J|M) = P(J)$? No
- $P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No
- $P(B|A,J,M) = P(B|A)$? Yes
- $P(B|A,J,M) = P(B)$? No
- $P(E|B,A,J,M) = P(E|A)$?
- $P(E|B,A,J,M) = P(E|A,B)$?

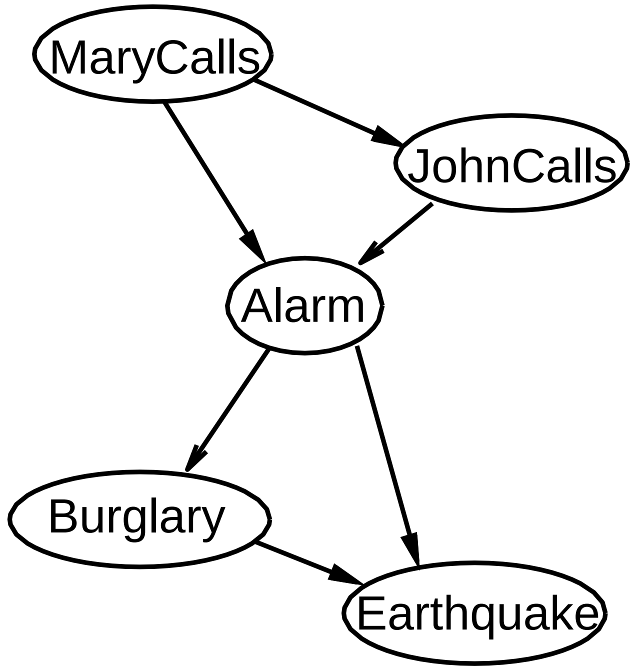
构建贝叶斯网络示例

- 假设我们选择顺序按照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
- $P(B|A,J,M) = P(B|A)$? Yes
- $P(B|A,J,M) = P(B)$? No
- $P(E|B,A,J,M) = P(E|A)$? No
- $P(E|B,A,J,M) = P(E|A,B)$? Yes

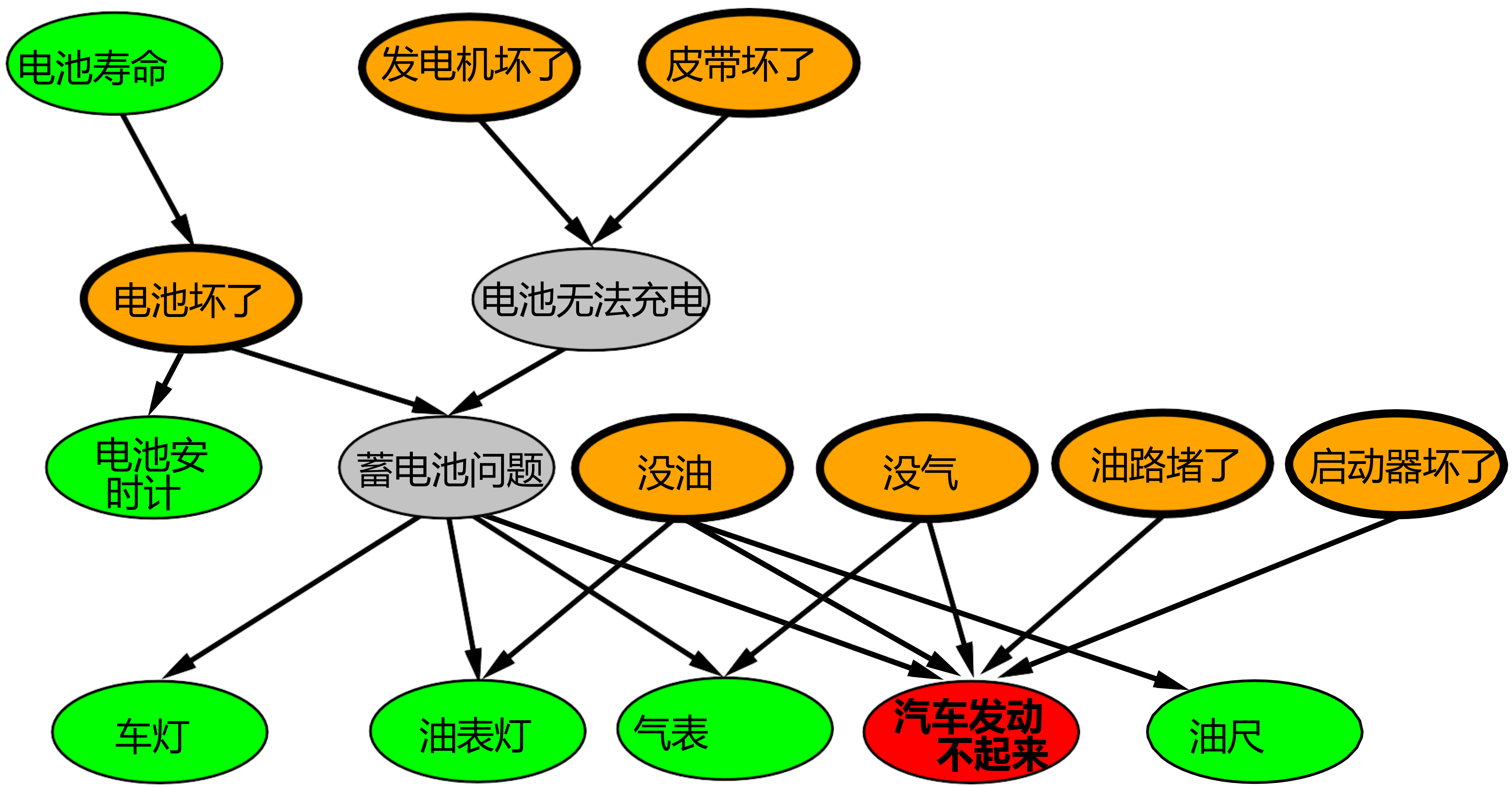
构建贝叶斯网络示例



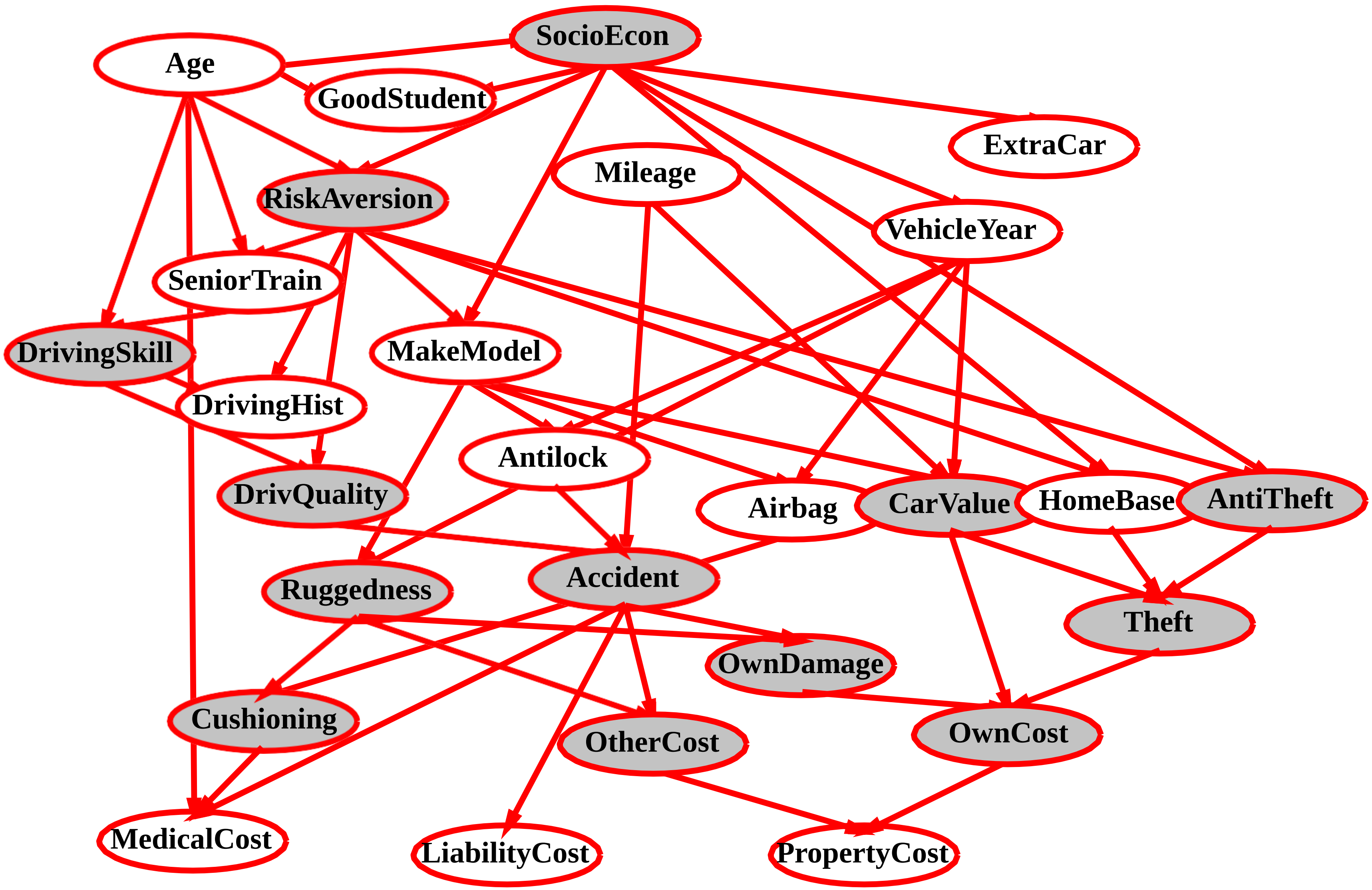
- 在非因果方向上，实现条件独立很困难（人类与生俱来似乎就可以得到因果模型和条件独立性的能力）
- 这也意味着，在非因果方向上评估条件概率是困难的
- 因为网络不那么紧凑，需要 $1+2+4+2+4=13$ 个概率值
    - 作为比较，总共的状态组合是 $2^5-1=31$

实例：汽车诊断

- 汽车发动不起来，如何确定问题
- 可测试的变量、 “坏掉的”变量、隐藏变量



实例：汽车保险


SocioEcon
Age
GoodStudent
ExtraCar
Mileage
RiskAversion
VehicleYear
SeniorTrain
DrivingSkill
MakeModel
DrivingHist
Antilock
DrivQuality
Airbag
CarValue
HomeBase
AntiTheft
Ruggedness
Accident
Theft
OwnDamage
Cushioning
OtherCost
OwnCost
MedicalCost
LiabilityCost
PropertyCost

紧凑条件分布

- 随着父节点的增加，紧凑度CPT程指数增长
- 如果父节点或子节点是连续值的时候，CPT会变为无限大
- 解决方案：将分布规范定义为已知节点上的最简单的情况

对某一个函数 f 来说 $X = f(Parents(X))$

- 例如：对于布尔函数

华中地区⇔湖北省∨湖南省 ∨ 我省

- 例如：对于某地水量（连续值）的和时间的数值关系

$$\frac{\partial Level}{\partial t} = \text{流入} + \text{降水} - \text{流出} - \text{蒸发}$$

紧凑条件分布

- 噪声或分布模型有多个非相互作用的原因
  - 父节点$U_1 \ldots U_k$包括所有的原因
  - 对于每一个原因，都有概率 q_i 导致独立性失败：

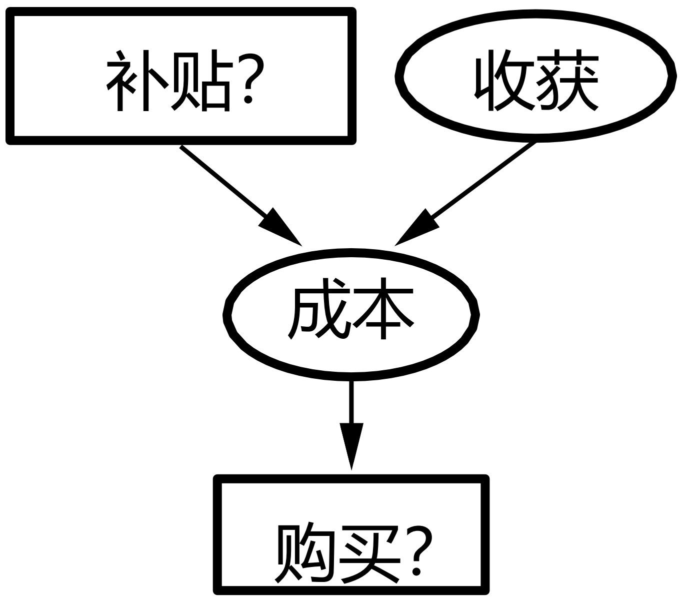
$$\Rightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1} q_i$$

感冒	流感	疟疾	P(发烧)	P(¬发烧)
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 × 0.1
T	F	F	0.4	0.6
T	F	T	0.94	0.06 = 0.6 × 0.1
T	T	F	0.88	0.12 = 0.6 × 0.2
T	T	T	0.988	0.012 = 0.6 × 0.2 × 0.1

- 参数个数与父节点个数成线性关系

混合(离散+连续)网络

- 离散(补贴和购买)、连续的(收获和成本)



- 选择1：离散化——可能有较大的误差，较大的CPT
- 选择2：有限参数化正则族
- 连续变量，离散+连续双亲(例如：成本)
- 离散变量，连续双亲(例如：购买?)

连续子代变量

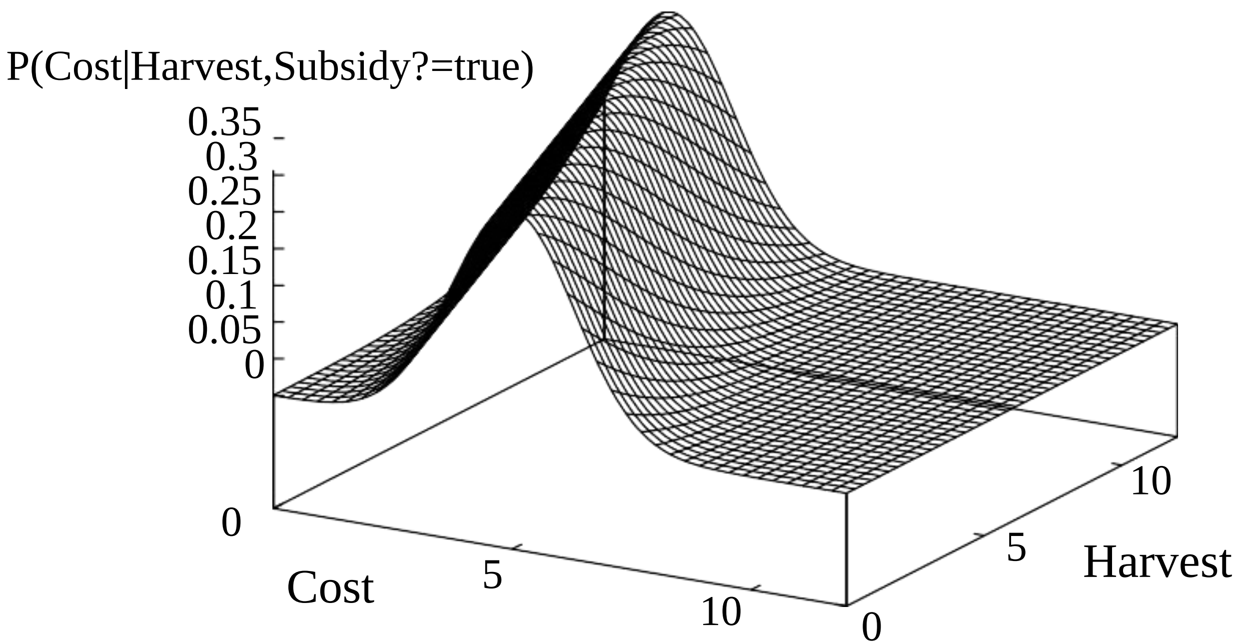
- 对于给定连续双亲的子变量，需要一个条件密度函数，对于每个可能的离散双亲赋值
- 最常见的是线性高斯模型（LG），例如：

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

- 平均$Cost$随$Harvest$线性变化，方差是固定的
- 线性变化在整个定义域范围内是不合理的，但如果可能的自变量范围很窄也可以（为什么？）

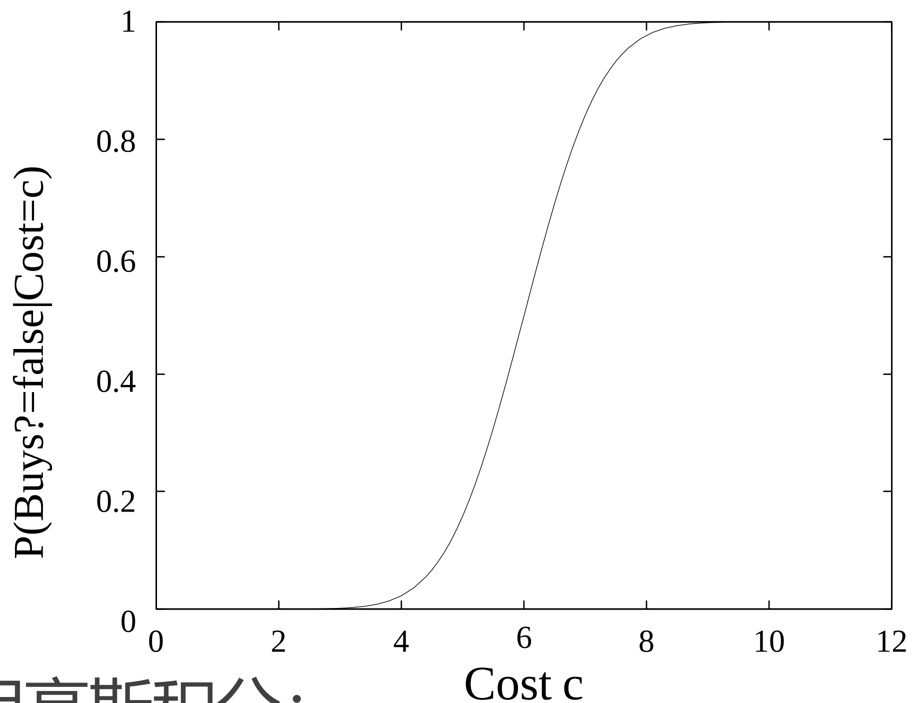
连续子代变量

- 具有线性高斯分布（LG）的全连续网络
- ⇒ 完全联合分布是一个多元高斯分布
- 离散+连续的LG网络是一个条件高斯网络。对于每个离散变量值的组合的所有连续变量满足多元高斯分布（multivariate Gaussian）



P(离散变量w | 连续双亲)

- 给定成本曲线，如果要计算购买的可能性，那么成本的阈值将是一个“软性”的阈值



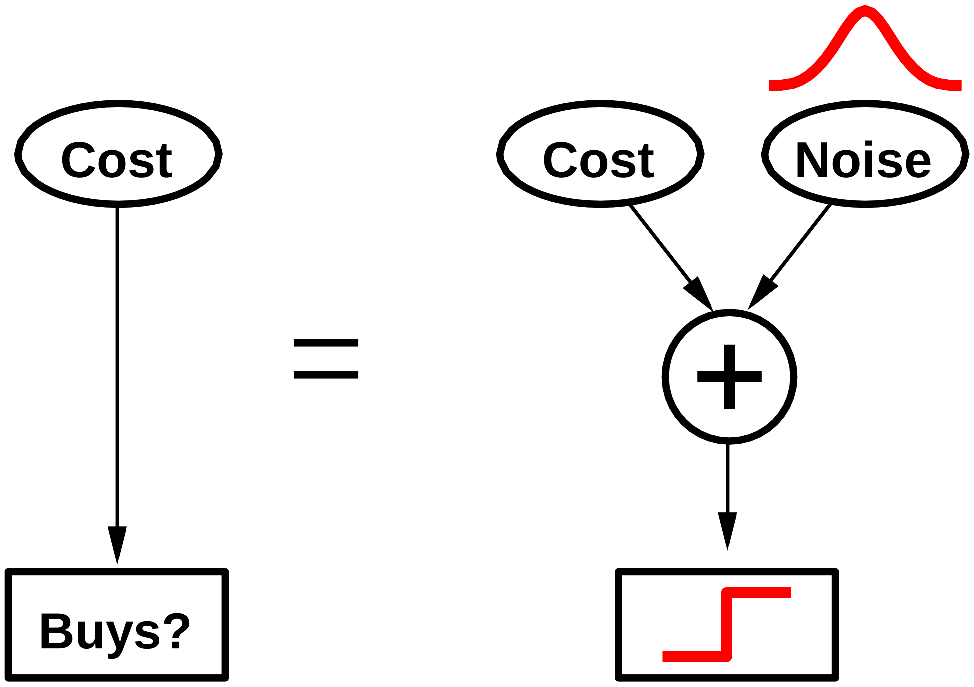
- probit分布采用高斯积分：

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

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

为什么使用probit?

- 1、概率有正确的形状
- 2、概率可以当做“硬”阈值，其位置受噪音的影响

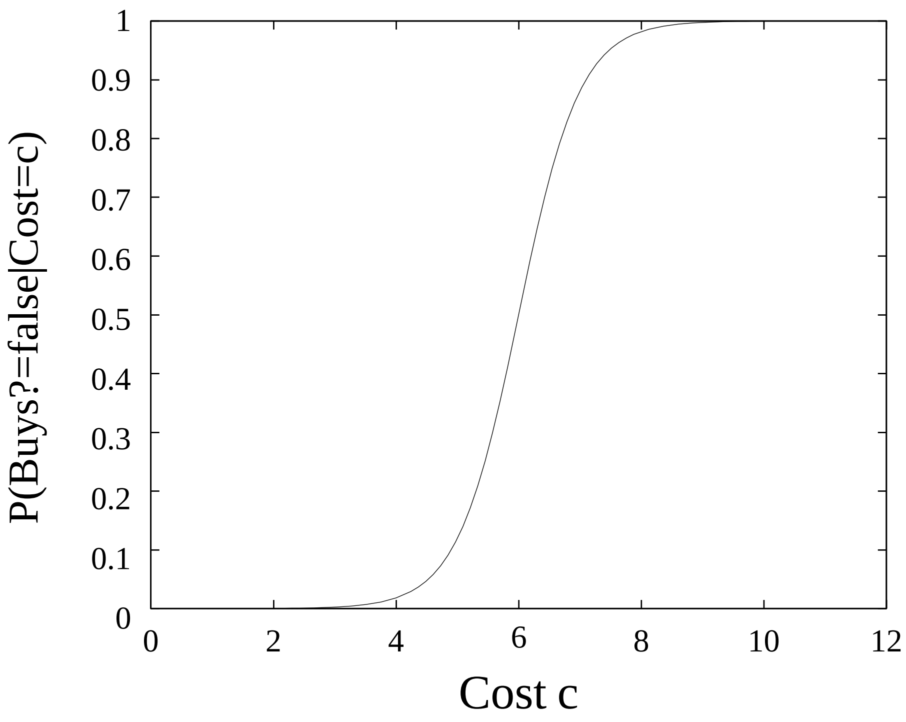


离散变量

- Sigmoid(或logit)分布也用于神经网络：

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

- Sigmoid的形状与probit相似，但尾巴要长得多：



总结

- 贝叶斯网络为（因果性）条件独立提供了一种自然表示
- 拓扑+ CPTs = 联合分布的紧凑表示
- 通常易于 (非) 专家构造概率分布使用
- 正则分布（例如噪声）=CPT的紧凑分布
- 连续变量⇒参数化分布（如线性高斯模型LG）