今天的内容

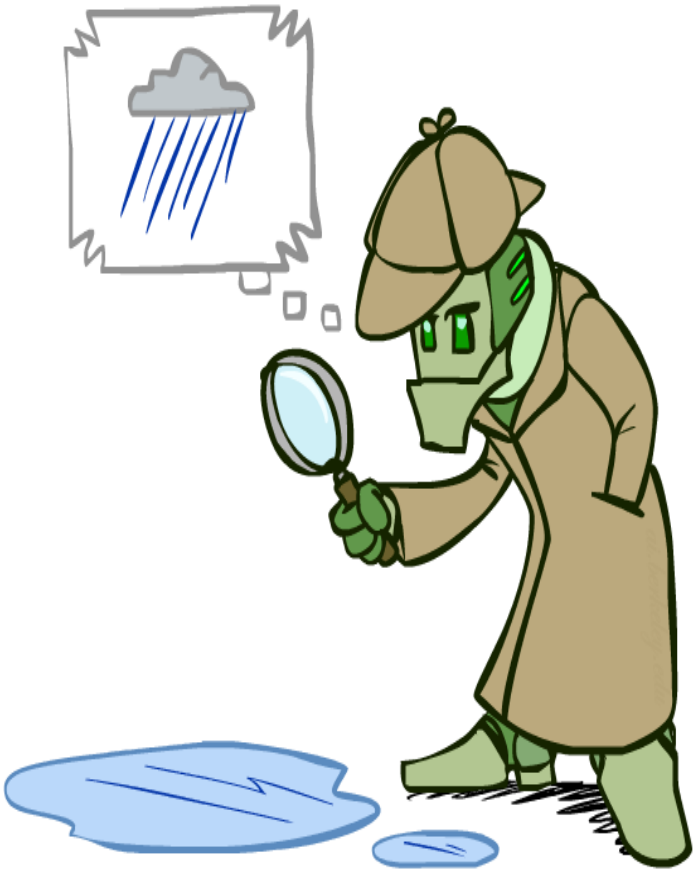
- 概率推理（基于列举法）
- 条件独立性
- 贝叶斯网络：语法和语义

概率推理(Probabilistic Inference)

- 概率推理: 从其他已知概率里计算一个想知道的概率 (例如，从联合概率中计算条件概率)

- 通常我们计算的都是条件概率
  - P(准时到机场 | 没有交通事故发生) = 0.90
  - 这些代表了智能体的*信念(beliefs)*，在给定证据(evidence)下

- 概率会随新的证据而变化:
  - P(准时到达 | 无交通事故, 早上5 点出发) = 0.95
  - 准时到达 | 无交通事故, 早上5 点出发, 下雨) = 0.80
  - 观察到新的证据时，会引发*信念(beliefs)* 的更新



通过列举(Enumeration)来推理

** 多个查询变量也可以*

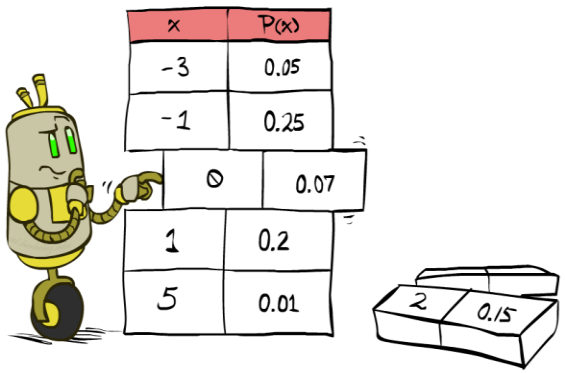
- 通常情况:
  - 证据变量: $E_1 \dots E_k = e_1 \dots e_k$
  - 查询* 变量: Q
  - 隐藏变量: $H_1 \dots H_r$

$\left.\begin{matrix} \\ \\ \end{matrix}\right\} X_1, X_2, \dots X_n$ *所有变量*

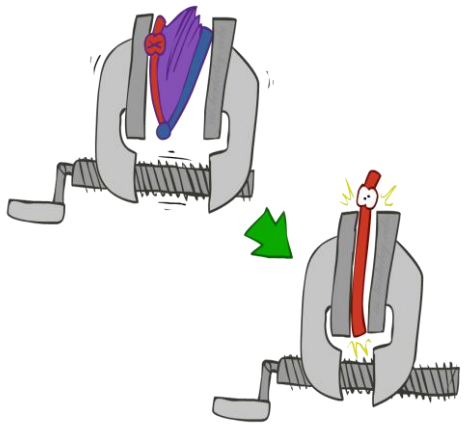
- 我们想要的:

$$P(Q|e_1 \dots e_k)$$

- 第一步: 选择和证据相一致的项



- 第二步: 求和消掉隐藏变量H，以得到查询和证据变量的联合分布



$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} \underbrace{P(Q, h_1 \dots h_r, e_1 \dots e_k)}_{X_1, X_2, \dots X_n}$$

- 第三步: 正规化

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

通过列举来推理

- P(W)?
- P(W | winter)?
- P(W | winter, hot)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

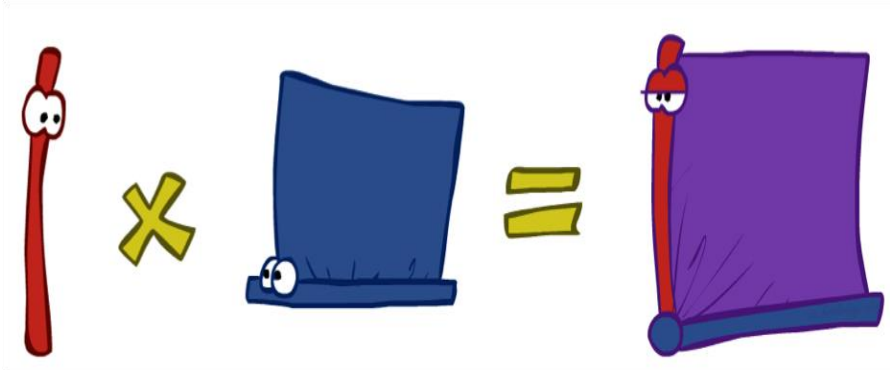
列举推理

- 明显的问题:
  - 最差情况下时间复杂度 $O(d^n)$
  - 空间复杂度 $O(d^n)$，需要存储联合分布

乘法规则(The Product Rule)

- 已有条件分布，想要计算联合分布

$$P(y)P(x|y) = P(x,y) \quad \Leftrightarrow \quad P(x|y) = \frac{P(x,y)}{P(y)}$$



乘法规则

$$P(y)P(x|y) = P(x,y)$$

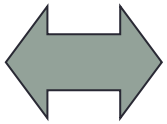
- 举例:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3



$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

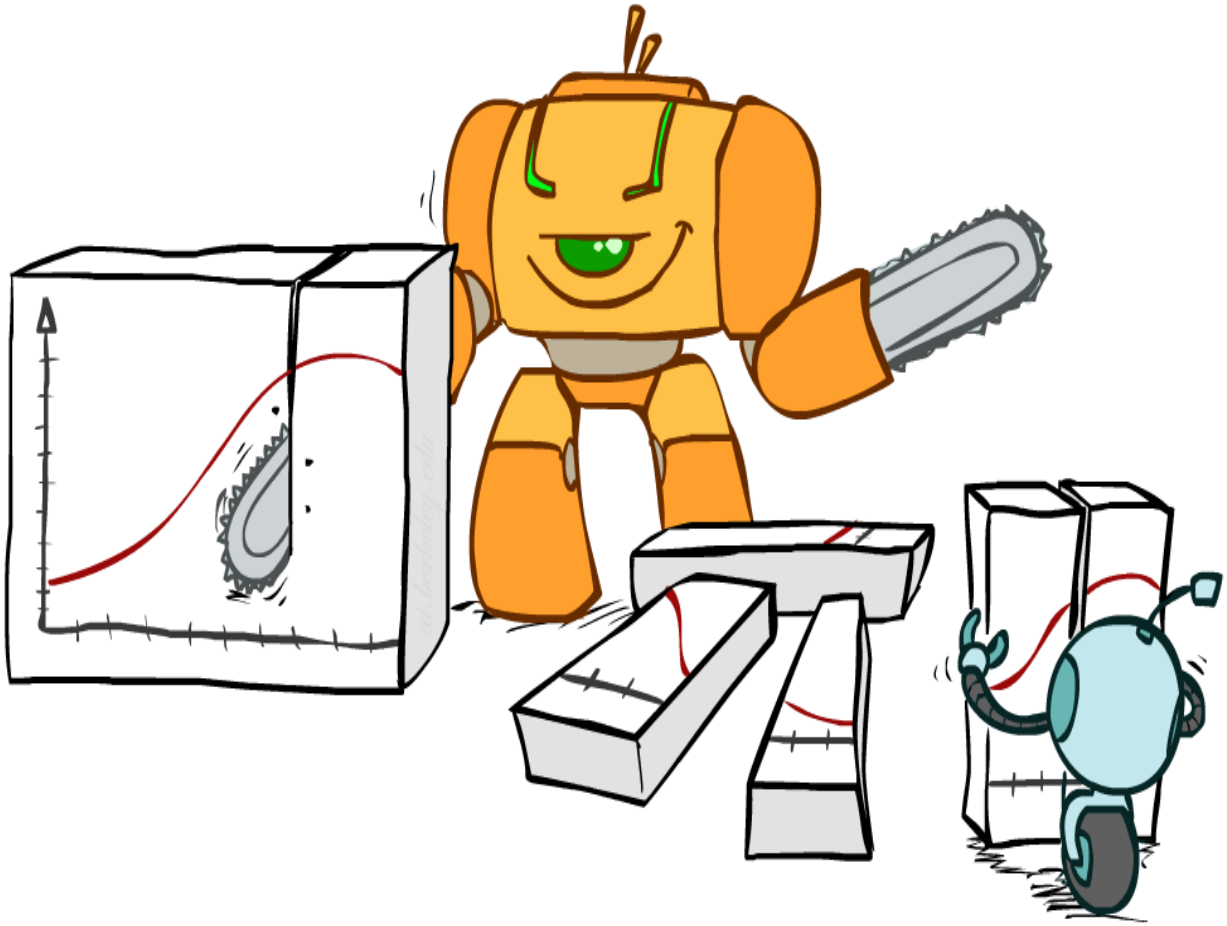
链式法则(The Chain Rule)

- 更普遍化的, 任何联合分布可以写成条件分布的增量相乘

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1})$$

贝叶斯法则



贝叶斯法则(Bayes' Rule)

- 两种方法因式分解一由两个变量组成的联合分布:

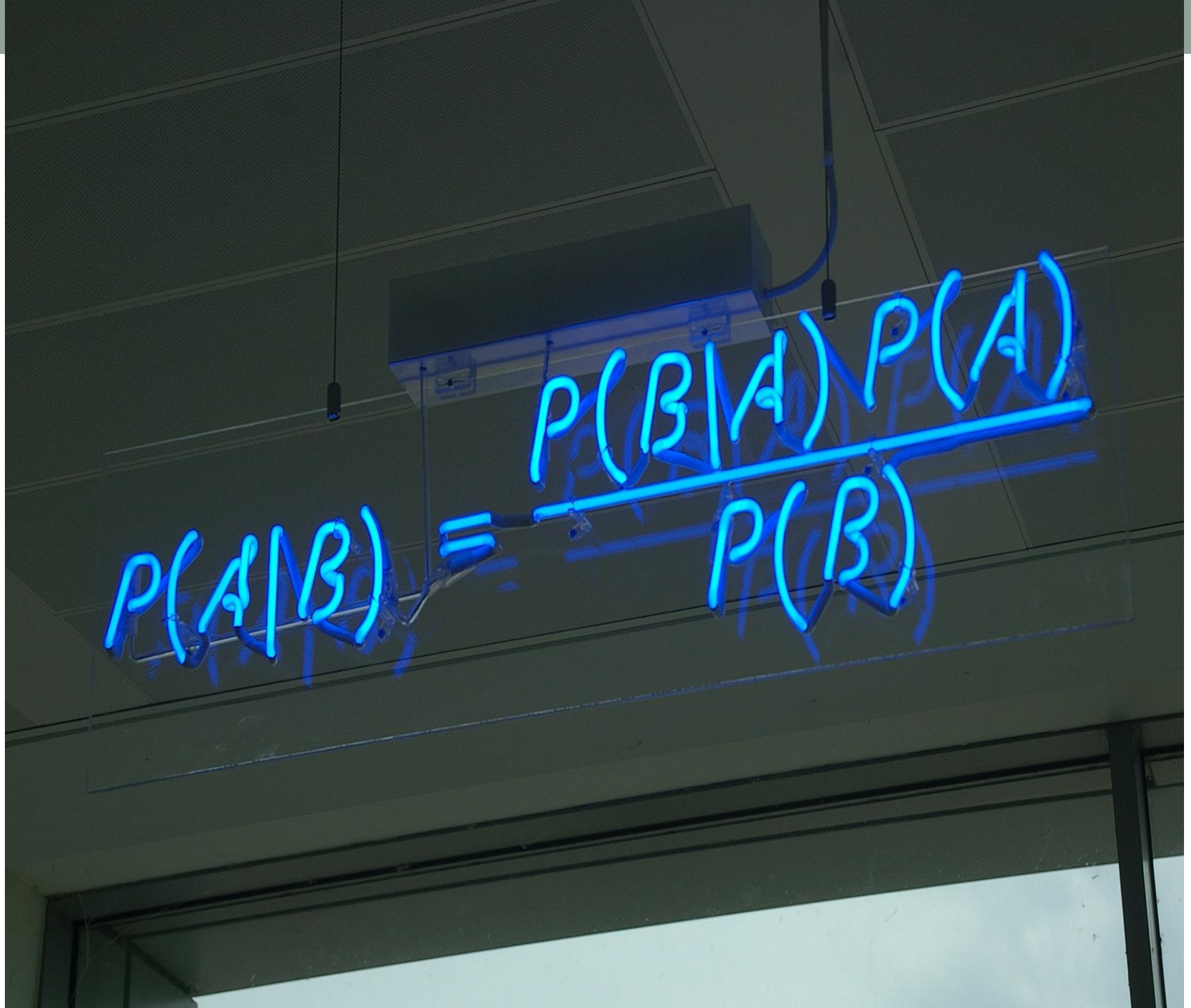
$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- 相除后, 我们得到:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- 为什么这个有用?
  - 让我们计算一个条件概率，从它的相反的形式
  - 通常一个条件概率很难计算，但是相对应的另一个却很简单
  - 许多人工智能系统的基础

- 最重要的人工智能公式之一!


$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$

用贝叶斯法则进行推断

- 举例: 从因果关系概率推断诊断概率:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- 举例:
  - M: 脑脊膜炎, S: 脖子发僵

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\} \text{例子中给定的}$$

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m)+P(+s|-m)P(-m)} = \frac{0.8\times 0.0001}{0.8\times 0.0001+0.01\times 0.999}$$

  - M的后验概率(posterior probability) 仍旧非常小: 0.007944 (将近 80x 大 – 为什么?)
  - 如果患了脖子僵硬仍需去检查! 为什么?

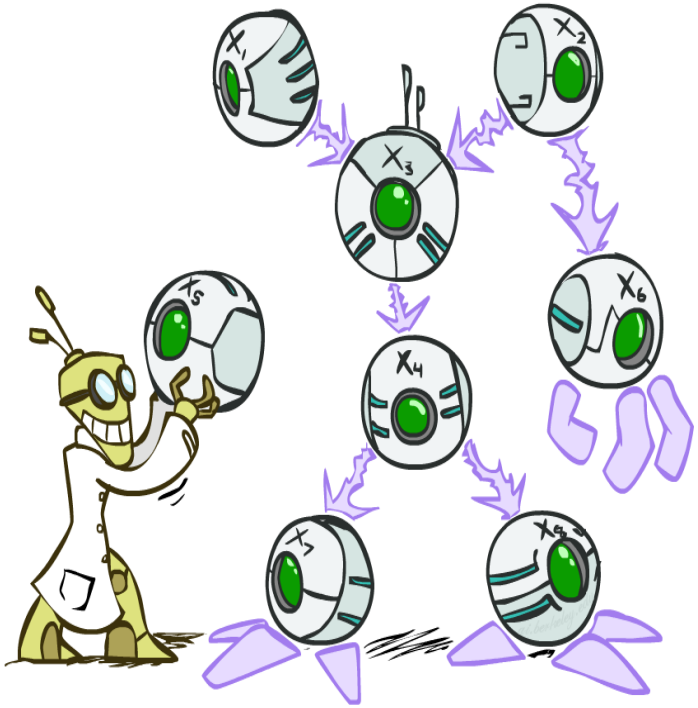
概率基础总结

- 基本法则: $0 \le P(\omega) \le 1$ $\quad \Sigma_{\omega \in \Omega} P(\omega) = 1$
- 事件: Ω 的子集: $P(A) = \Sigma_{\omega \in A} P(\omega)$
- 随机变量 $X(\omega)$ 对于每个 ω 有一个值
  - 分布 $P(X)$ 对每一个 x 值给出一个概率
  - 联合分布 $P(X,Y)$ 对每个 x,y 的组合给出概率
- 求和消除: $P(X=x) = \Sigma_y P(X=x,Y=y)$
- 条件概率分布: $P(X|Y) = P(X,Y)/P(Y)$
- 通过列举法概率推理: $P(Q|e_1,..,e_k) = \alpha\Sigma_{h_1,..,h_m} P(Q,e_1,..,e_k,h_1,..,h_m)$
  - 这里 α (即是 $1/Z$) 是一个正规化因子，使得 $P(Q|\ldots)$ 之和为 1

总结继续

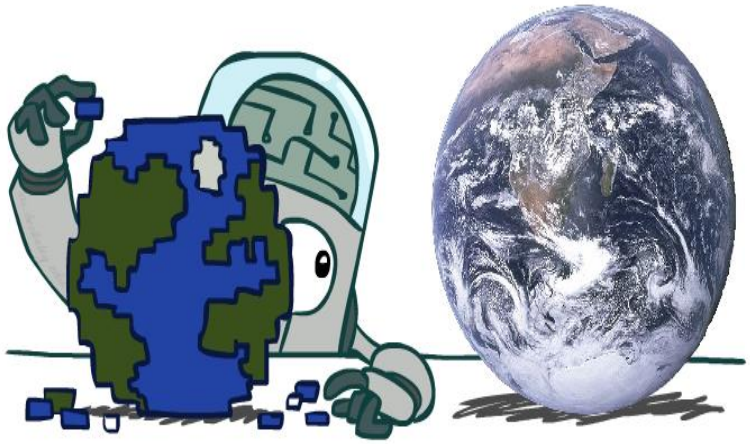
- 乘法规则: $P(X|Y)P(Y) = P(X,Y) = P(Y|X)P(X)$
  - 推广到连锁法则: $P(X_1,..,X_n) = \Pi_i P(X_i \mid X_1,..,X_{i-1})$
- 贝叶斯规则: $P(X|Y) = P(Y|X)P(X) / P(Y)$

贝叶斯网络(Bayes Nets)



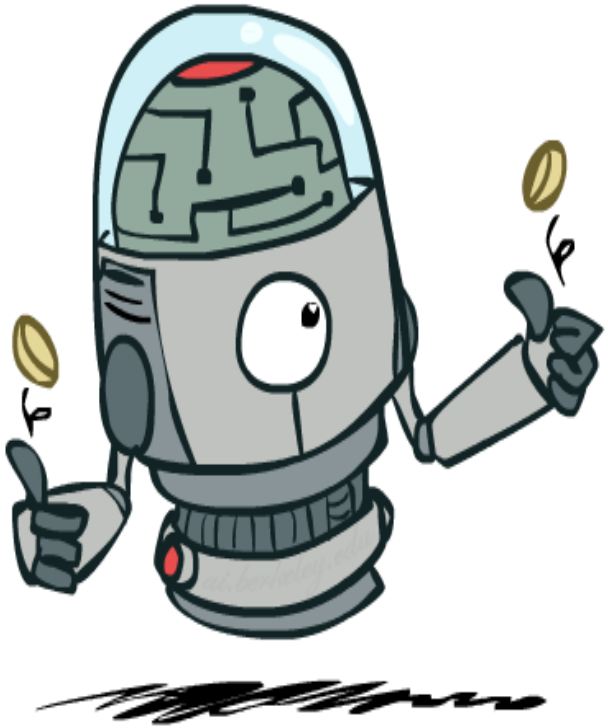
概率模型(Probabilistic Models)

- 模型描述的是世界（或某一部分）是如何工作的



- 模型总是一种简化
  - 可能忽略了某些变量和之间的交互关系
  - “所有模型都是错的; 但某些是有用的.”
    – George E. P. Box
- 概率模型能用来做什么?
  - 我们(或我们的人工智能体) 需要对未知变量进行推理，当给定一些证据后
  - 例如: 解释 (诊断推理)
  - 例如: 预测 (因果推理)
  - 例如: 基于期望利益值的决策
- 如何建立模型，并避免 d^n 的复杂性?

独立性



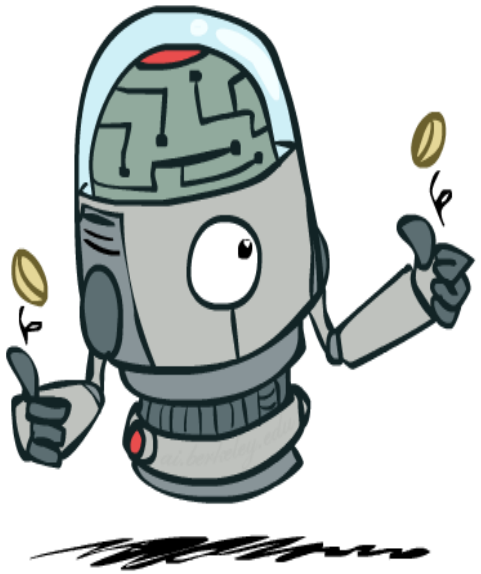
独立性

- 两个变量 X 和 Y 是 ***独立的*** 如果

$$\forall x,y \qquad P(x,y) = P(x)P(y)$$

  - 这说明他们的联合分布 ***因式分解*** 两个简单的分布的之乘积
  - 结合乘法规则： $P(x,y) = P(x|y)P(y)$ 我们可以获得另一种形式:

$$\forall x,y\ P(x|y) = P(x) \quad \text{or} \quad \forall x,y\ P(y|x) = P(y)$$



- 举例: 两个骰子 $Roll_1$ 和 $Roll_2$
  - $P(Roll_1=5, Roll_2=5) = P(Roll_1=5)P(Roll_2=5) = 1/6 \times 1/6 = 1/36$
  - $P(Roll_2=5 \mid Roll_1=5) = P(Roll_2=5)$

举例: 独立性

- n 个公平，独立的硬币翻转:

$P(X_1)$

H	0.5
T	0.5

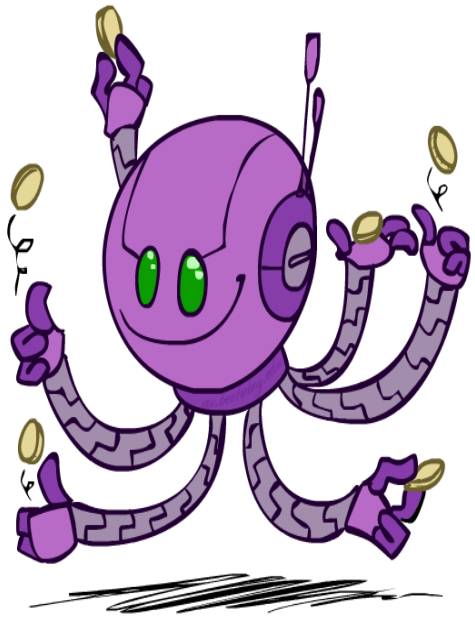
$P(X_2)$

H	0.5
T	0.5

$\ldots$

$P(X_n)$

H	0.5
T	0.5



$P(X_1, X_2, \ldots X_n)$

2^n

真实世界里的（概率事件）独立性

- 独立性是简化建模的假设
  - 有时对于真实世界的变量是合理的
  - 我们可以做什么样的假设，对于这些变量 {天气, 温度, 蛀牙, 牙疼}?
    - 蛀牙和牙疼 *大致* 是独立于天气和温度的
    - 蛀牙和牙疼相互间 *不是* 独立的
    - 天气和温度相互间 *不是* 独立的

天气和温度（的独立性）？

计算边缘分布

验证独立性，与P1对比

$P(T)$

T	P
hot	0.5
cold	0.5

$P_1(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P_2(T,W)$

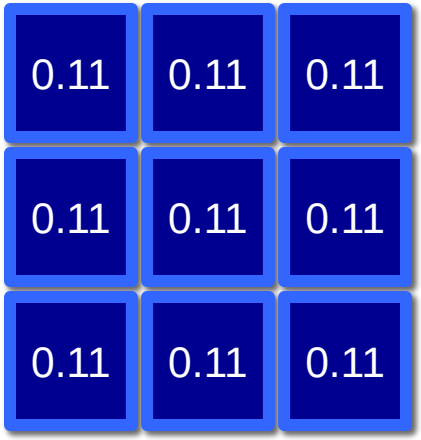
T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

$P(W)$

W	P
sun	0.6
rain	0.4

幽灵破坏者

- 变量和值域:
  - G (幽灵位置) 在 {(1,1),…,(3,3)}
  - $C_{x,y}$ (在方格 x,y 探测到的颜色；颜色越深离幽灵越近)： {red,orange,yellow,green}
- 假设我们有两个概率分布:
  - ***先验概率(Prior distribution)*** 幽灵的位置: $P(G)$
    - 假设是均匀(uniform)分布
  - ***传感器模型(Sensor model)***: $P(C_{x,y} \mid G)$ (只依赖于到 G的距离)
    - 例如 $P(C_{1,1} = \text{yellow} \mid G = (1,1)) = 0.1$



PURPLE YELLOW RED

BLACK RED GREEN

RED YELLOW ORANGE

BLUE PURPLE BLACK

RED GREEN ORANGE

幽灵破环者 计算: 第一步

- 当检测到 $C_{1,1}$ = yellow 时，幽灵可能在哪?
- 换句话说, 概率分布 $P(G \mid C_{1,1} = \text{yellow})$?
- 需要应用Bayes'
  - $P(G \mid C_{1,1} = \text{yellow}) = P(C_{1,1} = \text{yellow} \mid G)\, P(G) / P(C_{1,1} = \text{yellow})$
  - $= \alpha\, P(C_{1,1} = \text{yellow} \mid G)\, P(G)$ (贝叶斯***Bayesian 更新***)

正规化 **Normalize**

可能性 **Likelihood**

先验 **Prior**

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

0.17	0.10	0.10
0.09	0.17	0.10
0.01	0.09	0.17

幽灵破环者 计算: 第二步

- 当看到 $C_{1,1}$ = yellow, $C_{3,1}$ = green ， 幽灵在哪?
- 换句话说， $P(G \mid C_{1,1} = \text{yellow}, C_{3,1} = \text{green})$ 概率是什么 ?
- 我们的模型给出了 $P(C_{x,y} \mid G)$ 需要应用贝叶斯规则 Bayes' rule:
  - $P(G \mid C_{1,1}=y, C_{3,1}=g) = \alpha\, P(C_{1,1}=y, C_{3,1}=g \mid G)\, P(G)$
  - $= \alpha\, P(C_{3,1}=g \mid C_{1,1}=y, G)\, P(C_{1,1}=y \mid G)\, P(G)$
  - 条件后多了一项

应用乘法规则 | G

幽灵破环者 计算: 第二步

- 如何计算 $P(C_{3,1}=g \mid C_{1,1}=y, G)$?
- ***给定幽灵的位置***, 在位置 1,1 观察到的黄色是否影响到在 3,1 为绿色的概率?
- 不影响!
  - (只依赖于到幽灵的距离 distance from ghost)
- 在位置 3,1的颜色是 条件独立（无关的）(***conditionally independent)*** 对于在位置1,1的颜色，给定幽灵的位置
- $P(C_{3,1}=g \mid C_{1,1}=y, G) = P(C_{3,1}=g \mid G)$

幽灵破环者 计算: 第二步

- 观察到 $C_{1,1}$ = yellow, $C_{3,1}$ = green 后，幽灵的位置在哪?
- $P(G \mid C_{1,1} = \text{yellow}, C_{3,1} = \text{green})$?
- 我们的模型给出 $P(C_{x,y} \mid G)$ 只需应用 Bayes' rule:
  - $P(G \mid C_{1,1}=y, C_{3,1}=g) = \alpha\, P(C_{1,1}=y, C_{3,1}=g \mid G)\, P(G)$
  - $= \alpha\, P(C_{3,1}=g \mid C_{1,1}=y, G)\, P(C_{1,1}=y \mid G)\, P(G)$
  - $= \alpha\, P(C_{3,1}=g \mid G)\, P(C_{1,1}=y \mid G)\, P(G)$

依据 $C_{3,1}$ 和 $C_{1,1}$条件独立（无关性） 给定 G

距离	P(green | 距离)
0	0.01
1	0.1
2	0.2
3	0.3
4	0.4

条件独立性（条件无关）Conditional Independence

- 无条件的 (绝对的) 独立性非常稀少 (为什么?)
- 条件独立性是我们对于不确定环境的最基本和强健的知识蕴藏形式
- X 是 条件独立于(conditionally independent) Y, 给定 Z

  当且仅当:

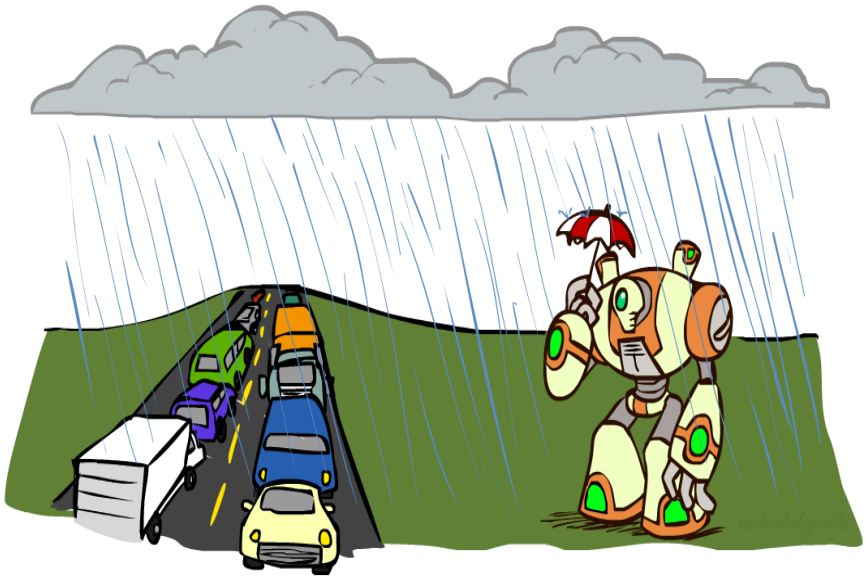
  $\forall x,y,z \quad P(x \mid y,z) = P(x|z)$

  或，等价地, 当且仅当

  $\forall x,y,z \quad P(x,y \mid z) = P(x|z)P(y|z)$

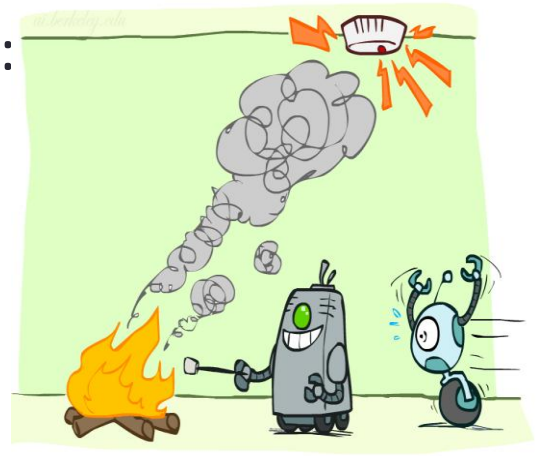
条件独立性（举例）

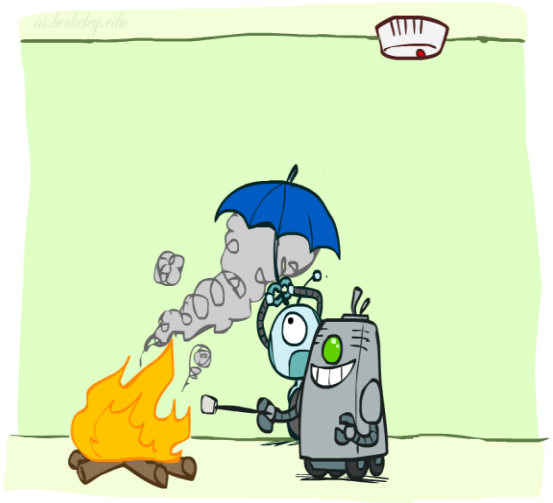
- 关于以下环境的独立性是怎样的:
  - 交通流量
  - 雨伞
  - 下雨



条件独立性

- 关于以下环境的独立性是怎样的:
  - 燃火
  - 冒烟
  - 报警器



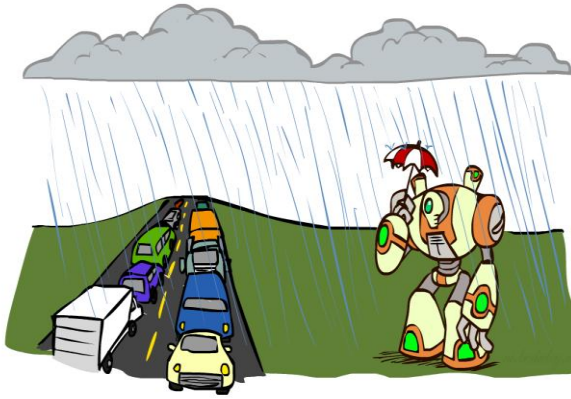


条件独立性与连锁法则(Chain Rule)

- Chain rule: $P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots$

- 简单的分解:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain}, \text{Traffic})$$

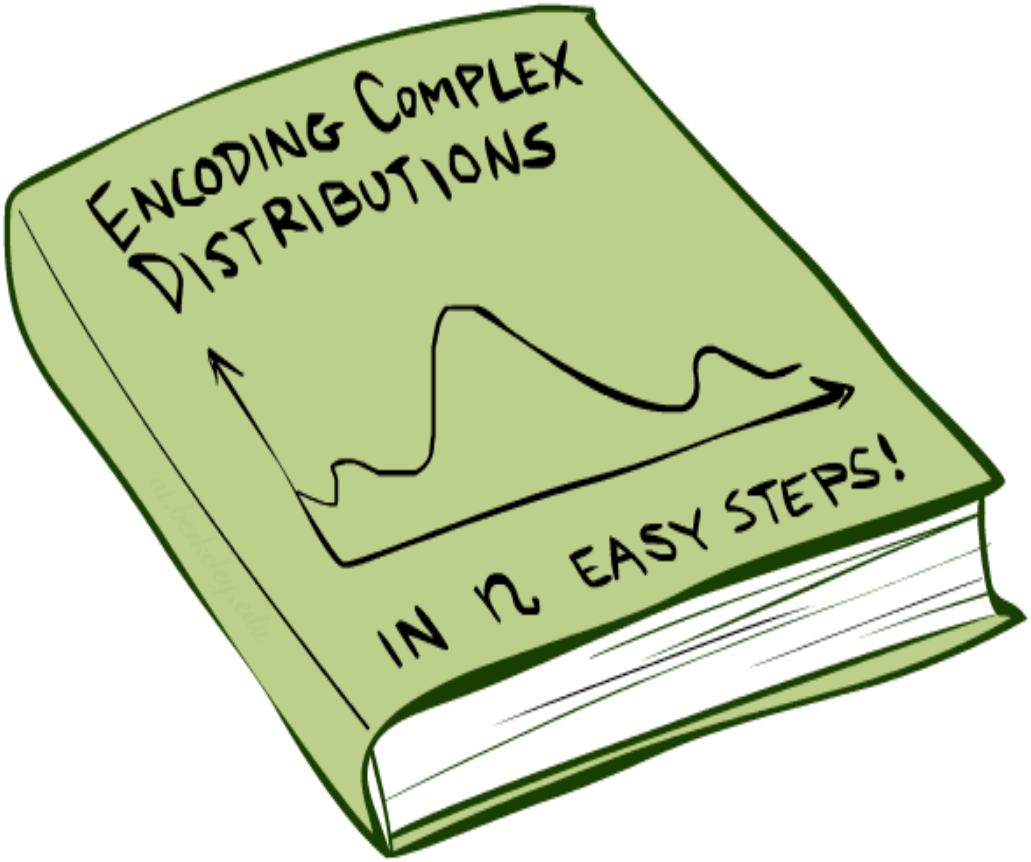


- 利用了条件独立性的假设后:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$

- 贝叶斯网络 / 图形模型 帮助表达条件独立性的假设

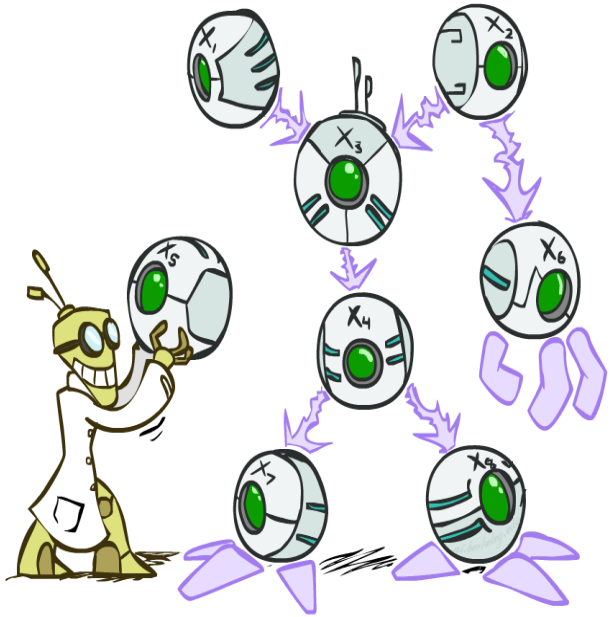
贝叶斯网络(BayesNets): 宏观介绍



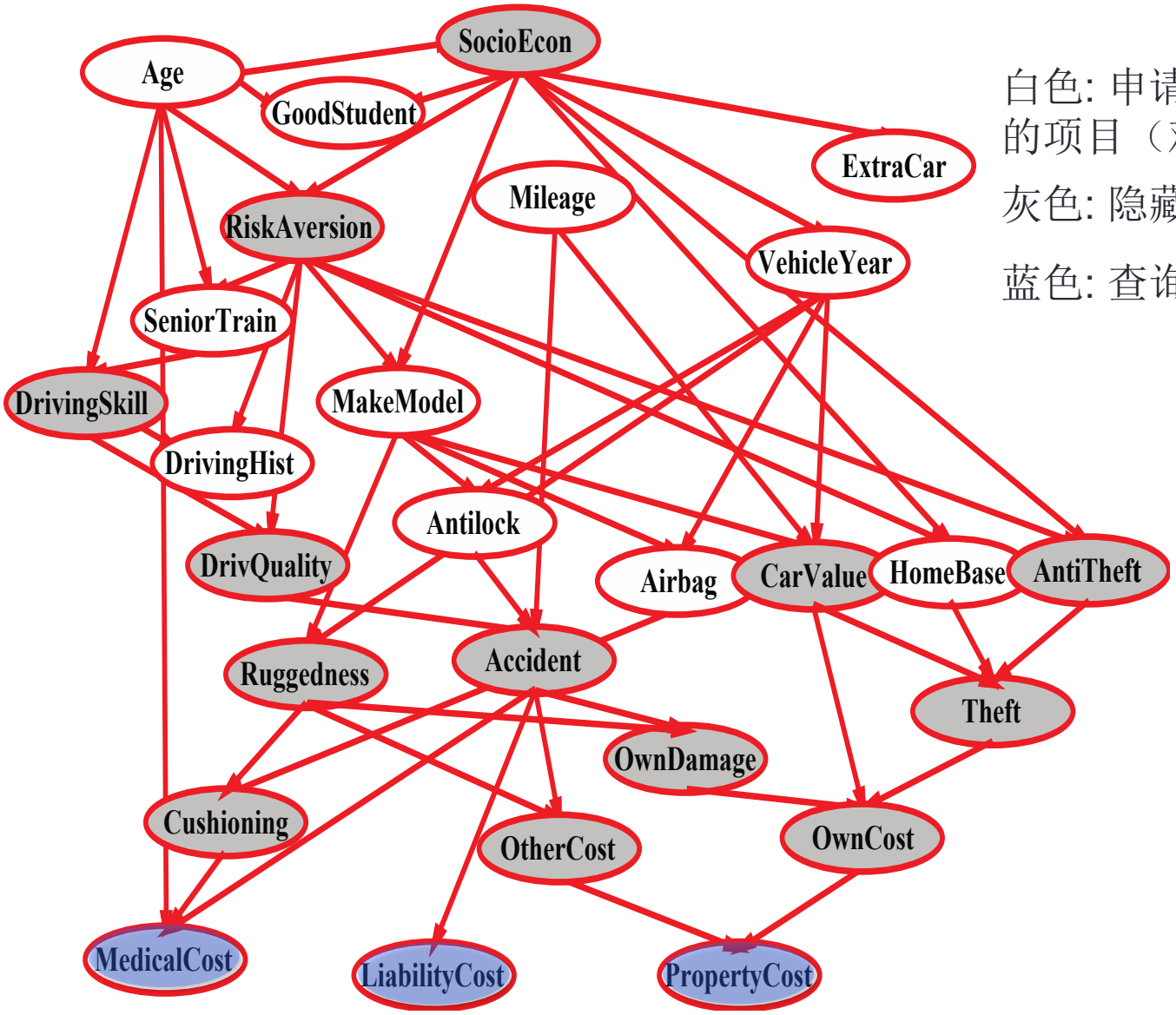
贝叶斯网络: 宏观介绍

- 完全的联合分布表可以回答每一个问题，但是：
  - 表的大小是变量数的指数级
  - 需要大量的例子来学习相应的概率
  - 用列举法 (加和消掉隐藏变量)进行推理太慢

- 贝叶斯网络(Bayesian networks):
  - 表达了一个由变量组成的领域里所有的条件独立性关系
  - 联合分布因式分解为小规模条件概率分布的乘积
  - 分布表达的量级从指数减少为线性
  - 从较少的例子快速学习出模型
  - 快速推理 (在某些重要实例里可以达到线性时间复杂度)
  - "*Microsoft's competitive advantage lies in its expertise in Bayesian networks*"
    -- Bill Gates, quoted in LA Times, 1996



贝叶斯网络举例: 汽车保险

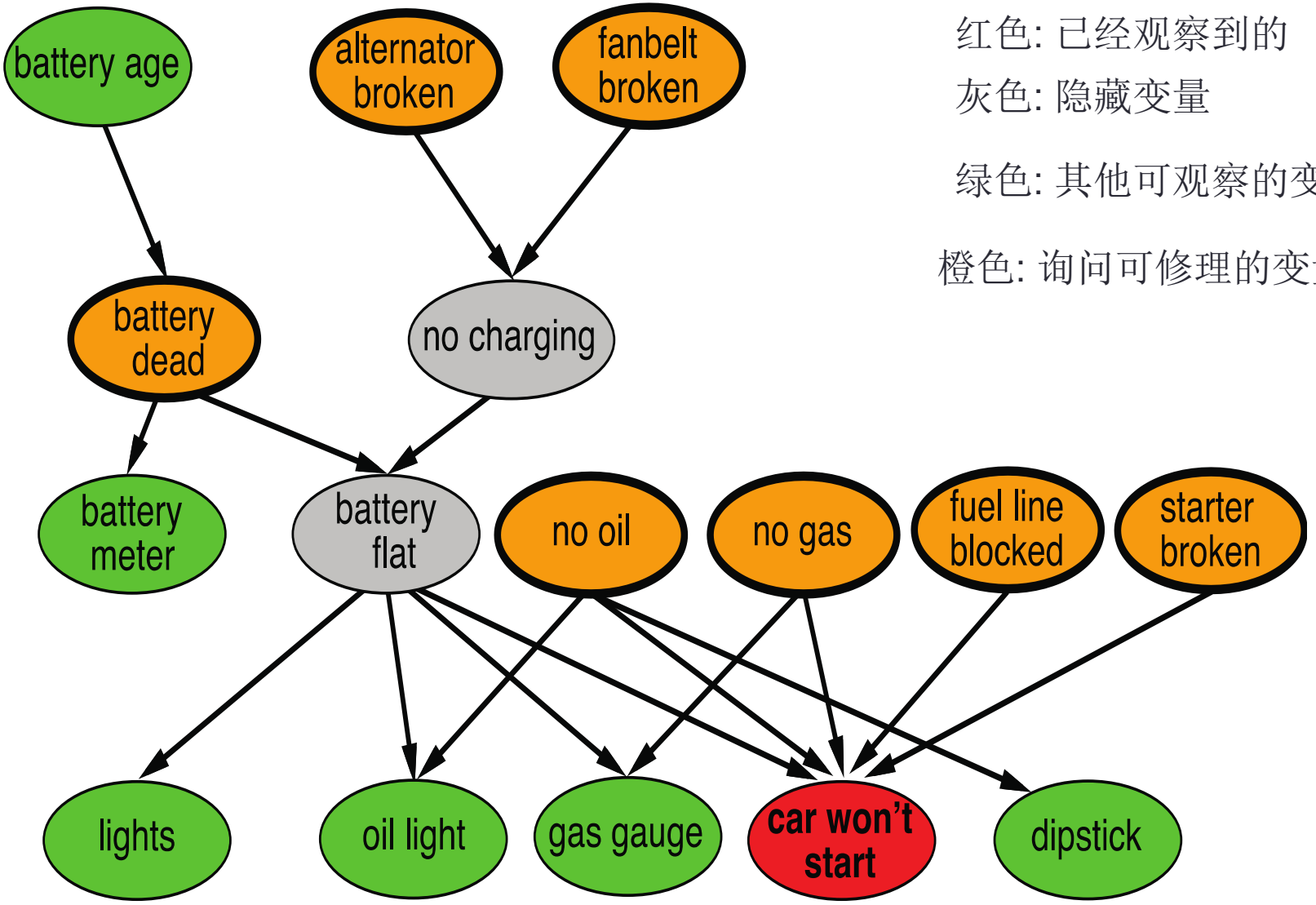


白色: 申请表格里出现的项目（观察到的）

灰色: 隐藏变量

蓝色: 查询变量

贝叶斯网络举例: 简单汽车维修



红色: 已经观察到的

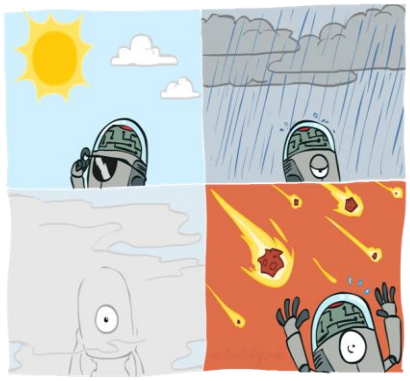
灰色: 隐藏变量

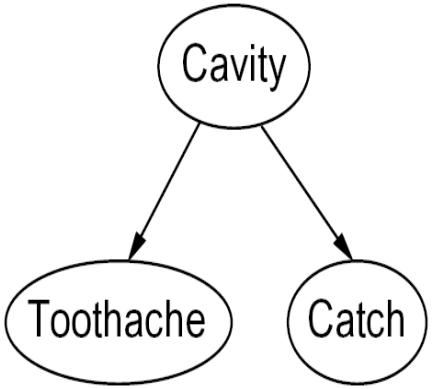
绿色: 其他可观察的变量

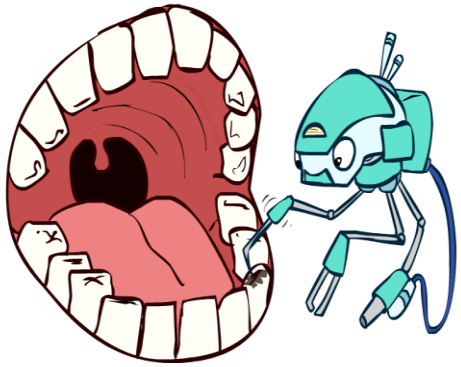
橙色: 询问可修理的变量

图形模型的表达

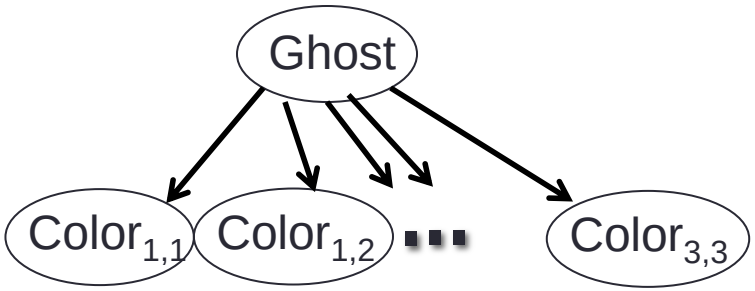
- 节点: 变量 (每个变量有个值域)
- 弧/边: 相互作用
  - 指示 变量间的“直接的影响”
  - 正式的表达: 编码条件独立性
- 现在可以简单地: 箭头意味着直接的因果关系 (通常情况下, 它们并不一定是这样!)





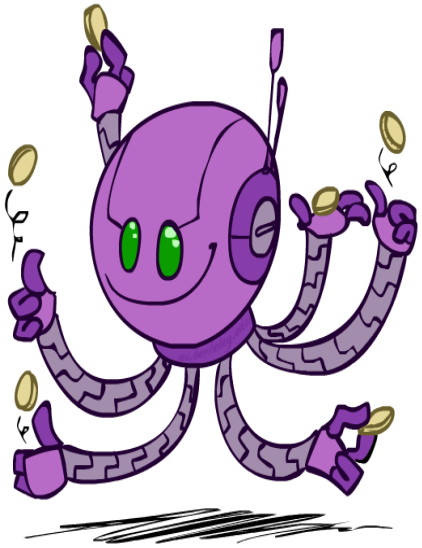


Catch，这里指的是牙医的探测器是否捕获到你的牙是否出问题；有可能会有遗漏。



举例: 硬币上抛翻转

- N 次硬币上抛（也可以理解为N个硬币抛一次）



X_1 X_2 $\cdots$ X_n

- 变量间没有相互作用: 绝对独立

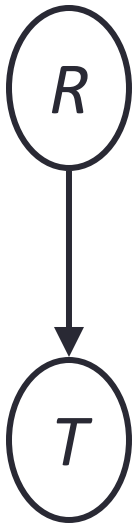
举例: 交通流量预测

- 变量:
  - R: 下雨
  - T: 交通状况

- 模型 1: 独立的

- 为什么用模型 2 更好?

- 模型 2: 下雨导致交通状况变化



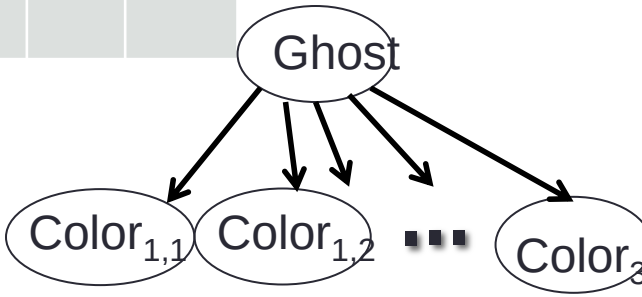
贝叶斯网络语法和语义

贝叶斯网络语法

- 一个节点对应一个变量 X_i
- 一个有向, 无环图
- 一个条件概率分布，对每个节点 给定图中它的 ***父节点***
  - ***CPT***: 条件概率分布表:
    - 每一行是子节点的一个分布，在给定父节点的一个配置以后
  - 一个近似的“因果”过程的描述

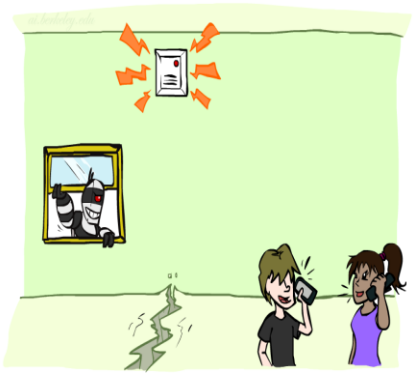
P(Ghost)			
(1,1)	(1,2)	(1,3)	…
0.11	0.11	0.11	…

Ghost	P($Color_{1,1}$ | Ghost)			
	g	y	o	r
(1,1)	0.01	0.1	0.3	0.59
(1,2)	0.1	0.3	0.5	0.1
(1,3)	0.3	0.5	0.19	0.01
…				



贝叶斯网络 = 拓扑结构(图形) + 局部条件概率

举例: 报警器网络

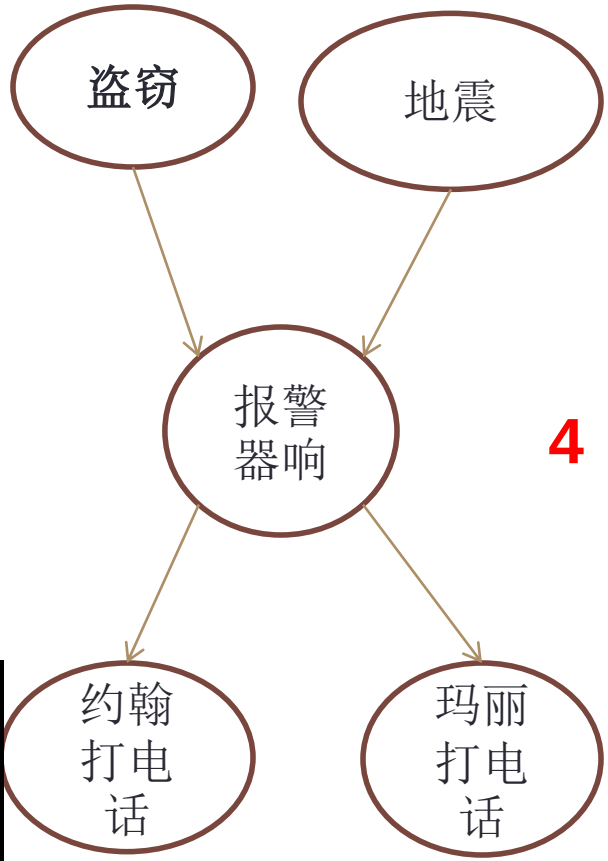


P(B)	
true	false
0.001	0.999

1

P(E)	
true	false
0.002	0.998

1



B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

4

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

2

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

2

条件概率分布表CPT的自由参数的个数总共有:

父变量的值域大小: $d_1,\ldots,d_k$

子变量的值域为 d 表中每一行概率值之和为 1

$(d-1)\ \Pi_i\ d_i$

对于稀疏的贝叶斯网络(BNs)，通用的规模计算公式

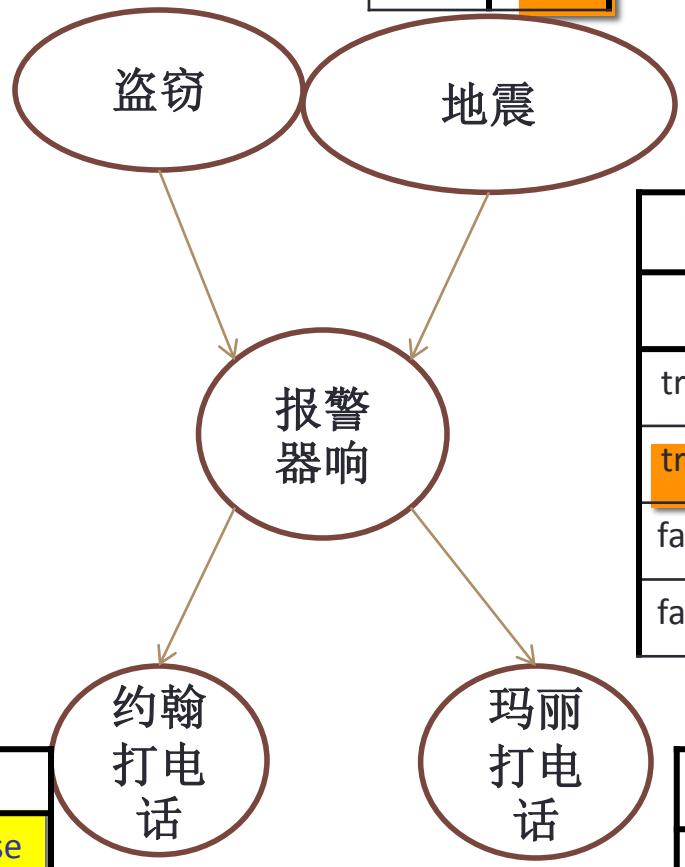
- 假定：
  - n 个变量
  - 最大值域大小是 d
  - 最大父节点数是 k
- 完全的联合分布的规模： $O(d^n)$
- 贝叶斯网络的规模： $O(n \cdot d^{k+1})$
  - n 的线性比例，只要因果结构是局部的

贝叶斯网络的全局语法

- 贝叶斯网络整体表达了（编码）联合分布，作为每一个变量上条件分布的乘积：

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

举例



P(B)	
true	false
0.001	0.999

P(E)	
true	false
0.002	0.998

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

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

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

$=.001 \times .998 \times .94 \times .1 \times .3 = .000028$

贝叶斯网络里的概率

- 为什么我们可以保证以下公式反映的是正确的联合分布

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

- 连锁法 (对所有分布有效): $P(X_1,..,X_n) = \prod_i P(X_i \mid X_1,\ldots,X_{i-1})$

- 假定 条件独立性: $P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid Parents(X_i))$
  - 当加入节点 X_i, 保证了其父节点 “屏蔽” 它与其他祖先节点的联系

→ 结果: $P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$

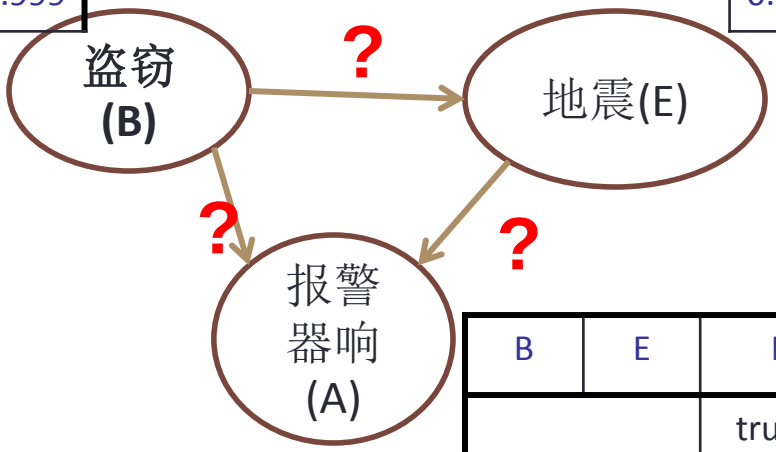
- 所以，网络的拓扑结构暗示着肯定的条件独立性的成立

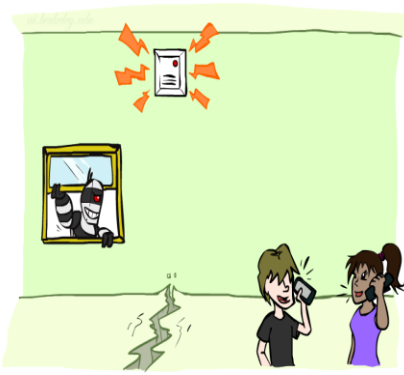
举例: 入室偷盗报警

- 入室盗窃
- 地震
- 报警器

P(B)	
true	false
0.001	0.999

P(E)	
true	false
0.002	0.998

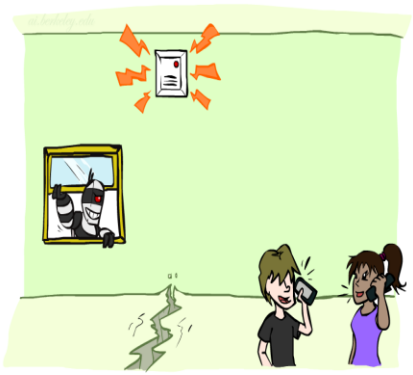


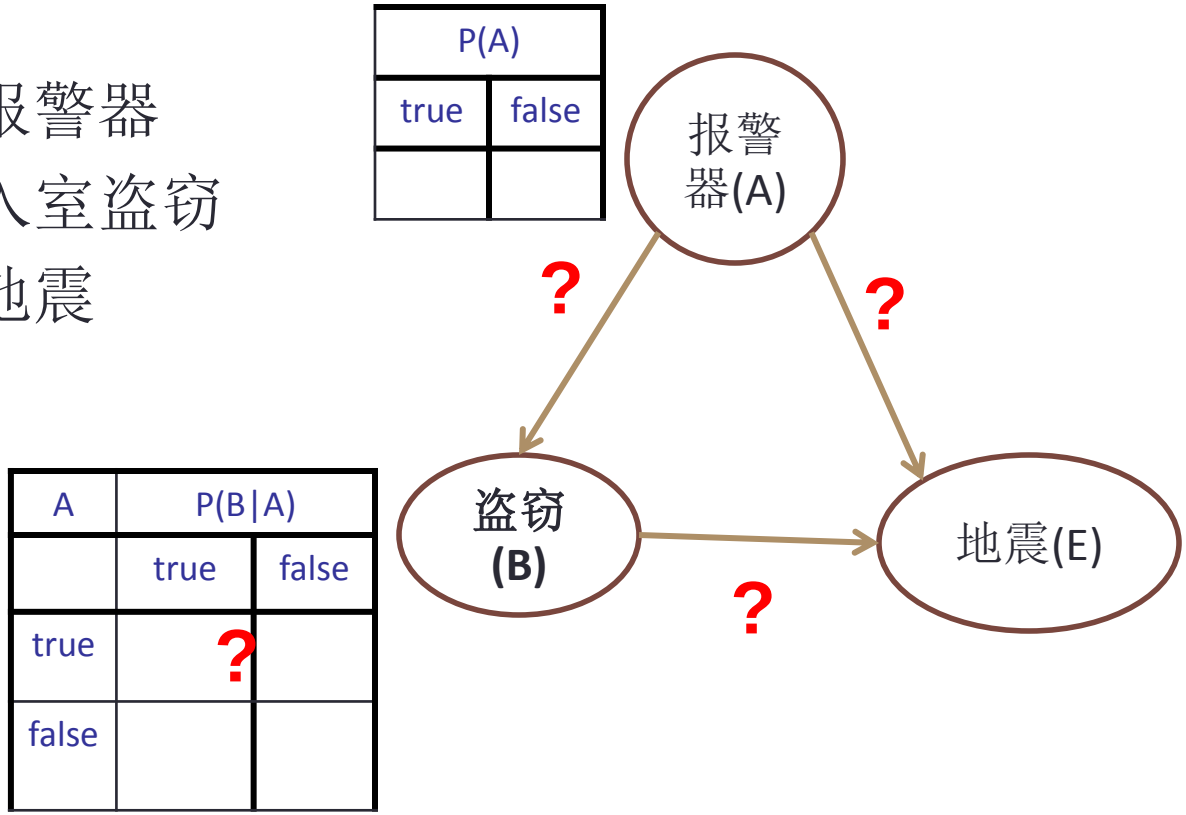


B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

举例: 入室偷盗报警

- 报警器
- 入室盗窃
- 地震





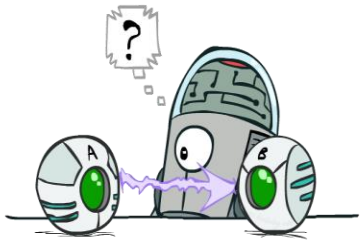
A	P(B|A)	
	true	false
true	?	
false		

A	B	P(E|A,B)	
		true	false
true	true	?	
true	false		
false	true		
false	false		

因果关系(Causality)?

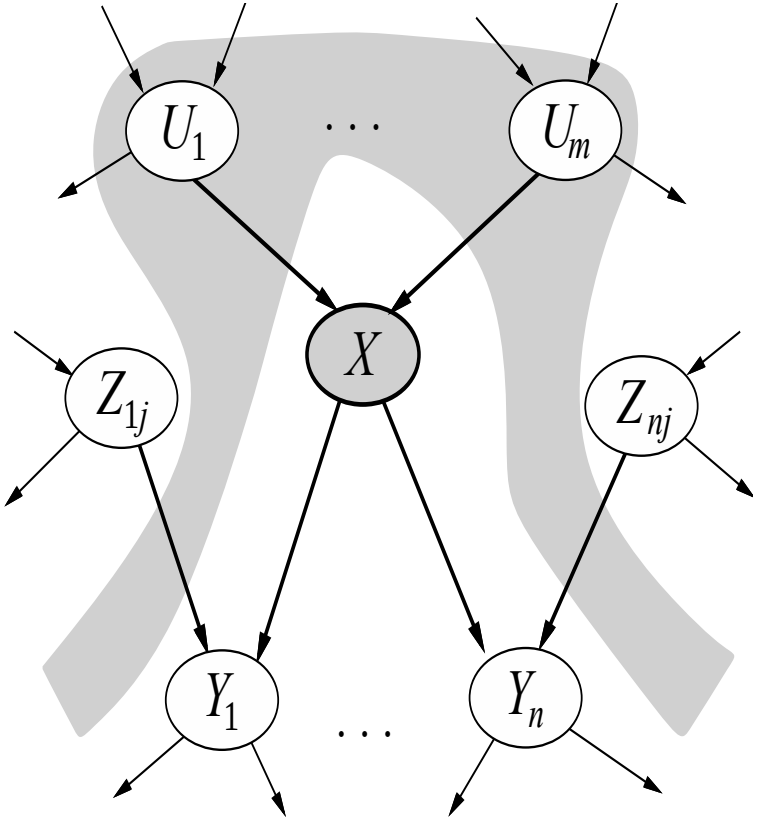
- 当贝叶斯网络反映了真实的因果关系模式时:
  - 通常更简单的网络 (较少的父节点, 较少的参数)
  - 通常更容易评估概率
  - 通常鲁棒性更强， 比如修改盗窃(B)的频率后应该不影响模型里的其他部分!
- BNs 不需要实际上表达因果关系
  - 有时没有因果网络存在于一个领域 (尤其是在一些变量丢失的情况下)
  - 例如，考虑变量 *交通状况* 和 *屋檐滴水*
  - 其结果是箭头关联反映的是相关性(correlation), 而不是因果关系
- 箭头实际表示的是什么?
  - 可能碰巧表达的是因果关系
  - 拓扑结构真正表达（编码）的是条件独立性:

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



条件独立性的语义

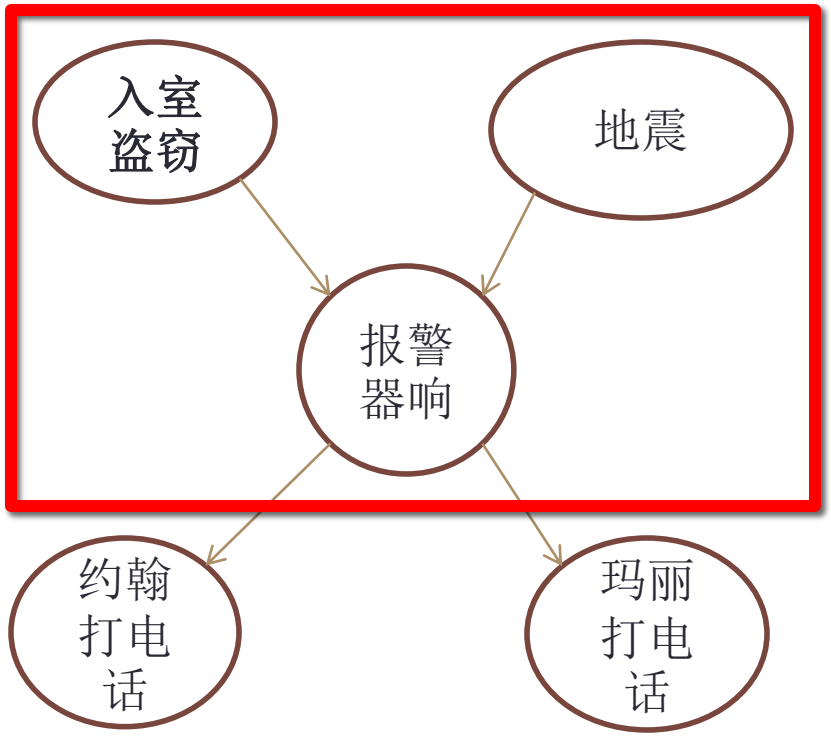
- *每个变量在给定它的父变量节点情况下，则是条件独立于它的非 后代变量*



举例

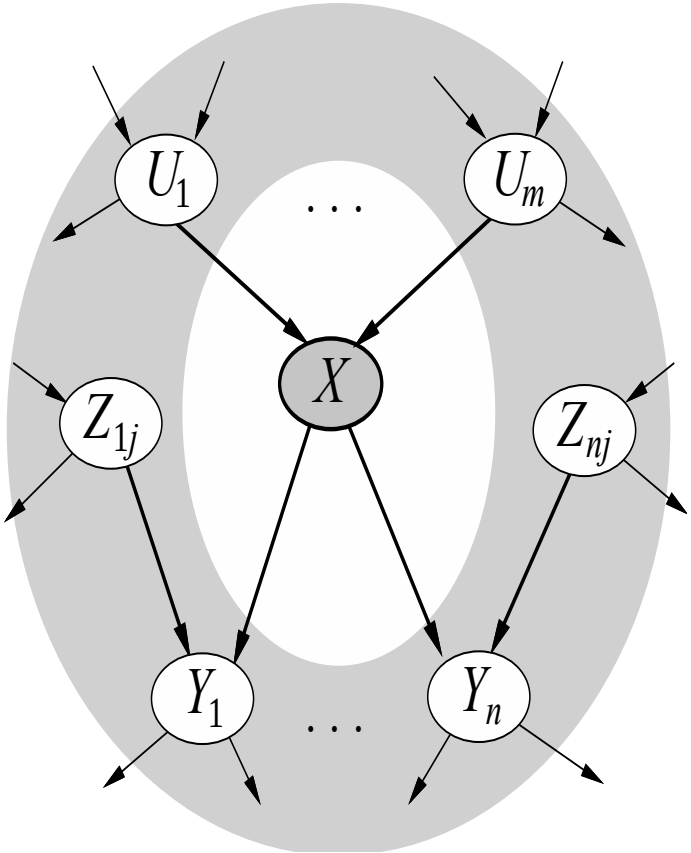
- 给定报警器响，约翰打电话 是否 独立于 入室盗窃的发生？
  - 是的
- 给定报警器响，约翰打电话 是否 独立于 玛丽打电话?
  - 是的
- 盗窃 是否独立于 地震?
  - 是的
- 盗窃 是否独立于 地震 当报警器响后？
  - 不是!
  - 报警器已响，入室盗窃和地震都变得很有可能发生过
  - 但是，如果我们得知一个入室盗窃已经发生，那么报警器响的原因被 ***解释***，则地震发生的概率降低

V-结构



马尔科夫毯(Markov blanket)

- 一个变量的马尔可夫毯包括父节点，子节点, 子节点的其他父节点
- ***每个变量给定它的马尔科夫毯，则是条件独立于所有其他变量***



贝叶斯网络(Bayes Nets)

- 已经介绍：贝叶斯网络如何实现了对联合分布的表达
- 下次：如何回答查询，计算查询变量在给定（观察）证据下的条件概率