贝叶斯网（Bayesian Networks）

吉建民

USTC
jianmin@ustc.edu.cn

2024 年 4 月 14 日

Used Materials

Disclaimer: 本课件采用了部分网络资源

Outline

贝叶斯网概述

- 贝叶斯网（Bayesian Networks）是一种帮助人们将概率统计应用于复杂领域、进行不确定性推理和数值分析的工具。
- 贝叶斯网是一种系统地描述随机变量之间关系的语言。
- 构造贝叶斯网的主要目的是进行概率推理，即计算一些事件发生的概率。
  - 联合概率太复杂（随变量个数指数增长）
  - 贝叶斯网把联合概率分解成一系列简单模块，从而降低难度
- 贝叶斯网是概率论与图论结合的产物，一方面用图论的语言直观揭示问题的结构，另一方面按概率论的原则对问题结构加以利用。
- 许多经典多元概率模型都是贝叶斯网的特例：隐马尔科夫模型、卡尔曼滤波器等。
- 贝叶斯网学习：从数据出发获得贝叶斯网的过程。

Outline

样本空间和事件

- 随机试验：事先不能完全预知其结果的试验
  - 抛掷骰子是一个随机试验
- 样本空间：随机试验的所有可能结果组成的集合，记为 Ω
  - 掷骰子的样本空间 $\Omega = \{1, 2, 3, 4, 5, 6\}$
- 原子事件（样本点）：样本空间中的点，即随机试验的可能结果，记为 ω
- 事件：样本空间的子集，记为 A, B, $\ldots$
  - $A = \{1, 3, 5\}$ 表示“掷出结果为奇数”这一事件
  - Ω 本身为必然事件，$\emptyset$ 为不可能事件
- 若两事件 $A \cap B = \emptyset$，称为互斥事件（不相容事件）
- 若两事件 $A \cap B = \emptyset$ 且 $A \cup B = \Omega$，称为互补事件

概率

- 概率测度：给样本空间中的每一个事件 A 赋予一个数值（概率） $P(A) \in [0, 1]$
- 概率测度（形式化）是一个从样本空间 Ω 的幂集 2^{Ω} 到区间 $[0, 1]$ 的映射 $P: 2^{\Omega} \to [0, 1]$，且满足以下三个 Kolmogorov 公理：

  (1) $P(\Omega) = 1$; （规范性）

  (2) $P(A) \geq 0, \forall A \in 2^{\Omega}$; （非负性）

  (3) $P(A \cup B) = P(A) + P(B), \forall A, B \in 2^{\Omega}, A \cap B = \emptyset$. （有限可加性）

- $P(A)$ 称为事件 A 的概率

随机变量和概率函数

- 随机变量是定义在样本空间 Ω 上的函数，记为 X, Y, Z
- 随机变量的取值随试验结果而定，记为 x, y, z
- 随机变量 X 的所有可能取值的集合称为其值域（状态空间），记为 Ω_X
- 设 X 为一随机变量，x 是它的一个取值，在样本空间 Ω 中，所有使 X 取值为 x 的原子事件组成一个事件，记为 $\Omega_{X=x} = \{w \in \Omega \mid X(\omega) = x\}$，简记为 "$X = x$"
- 事件 "$X = x$" 的概率 $P(X = x) = P(\Omega_{X=x})$ 依赖于 X 的取值 x，让 x 在 Ω_X 上变动，$P(X = x)$ 就称为 Ω_X 的一个取值于 $[0, 1]$ 的函数，称为随机变量 X 的概率质量函数，记为 $P(X)$
- 根据概率测度的定义

$$P(X = x) \geq 0, \forall x \in \Omega_X \text{ 简记为 } P(X) \geq 0$$

$$\Sigma_{x \in \Omega_X} P(X = x) = 1 \text{ 简记为 } \Sigma_X P(X = x) = 1.$$

概率的解释

- 概率的解释主要有 5 种：古典解释、频率解释、主观解释、特性解释、逻辑解释
  - 特性解释：概率被看作一种物理性质或倾向，或者给定一种物理情景以产生某类结果的趋向，或者产生这些结果的长序列中的相对频率。
    - 这种解释不能找到概率和频率之间的联系，特性指派和概率演示之间的关系是不清晰的，也难于实用
  - 逻辑解释：概率是对知识状态的总结，是由从证据到假设的逻辑关系所决定的。一旦相关的知识得到确定，则事件的可能性就已经被客观的确定下来，并应该能够通过逻辑分析得到
    - 古典解释可以看做逻辑解释的一个特例，它从等可能性的前提条件出发来计算概率
    - 同特性解释一样，逻辑解释的缺点在于它没能为概率提供一个可操作的运算方法

古典解释和频率解释

- 古典解释：概率被同等的分配在所有的可能的结果之间，所以一个事件的古典概率是该事件出现的可能场合数与总的场合数的分数
  - 事件 A 的概率为：$P(A) = \dfrac{\text{事件 } A \text{ 包含的样本数}}{\text{样本空间的总样本数}} = \dfrac{m}{n}$
  - 古典概率的一个前提条件是等可能性，在实际应用中，这个前提一般很难满足，因此古典概率的应用范围很有限
  - 例如，古典概率无法解释，给定一个质地不均匀的骰子，掷出 6 的概率为多大
- 频率解释：一个事件的概率是在一个恰当选定的参考类中的那种事件的相对频率
  - 对于一个可在同样条件下重复进行的试验，如果事件 A 在所有 N 次试验中共发生了 M 次，则它的概率可以用其发生的频率来近似：$P(A) \approx \dfrac{M}{N}$
  - 支持大数定律：当 N 趋于无穷大时，频率几乎处处趋于概率
  - 按频率解释，概率只有当试验可以在同等条件下无限次重复时才有意义，然而，实际中人们往往需要研究一些不可重复的事件发生的概率
  - 频率解释无法处理一次性事件，例如总统竞选或体育比赛的结果

主观解释

- 主观解释（贝叶斯解释）：一个事件的概率是人们根据经验对该事件发生可能性所给出的个人置信度（信念度），反映个体的知识状态和主观信念
  - 所谓置信度，就是相信某事件发生或某个命题为真的程度
  - 对置信度的标准的分析是诉求打赌行为
    - 例如，巴西队赢得下届世界杯足球赛冠军的概率是多大？
    - 概率轮评估方法：设想一个质地均匀的概率轮，划定一片黑色的区域，评估“巴西队夺冠的可能性大还是指针旋转后停在黑色区域的可能性大”，据此来调整黑区的大小，最后根据黑区大小计算概率
  - 评估精度问题：概率值为 0.201 或 0.202，有什么根本差别？
    - 在贝叶斯网应用中，精度问题不大，因为：(1) 概率值的微小差别对决策的影响一般不大；(2) 实际中往往会同时考虑多个事件的概率，由于概率必须满足 Kolmogorov 公理，因此不同事件的概率之间存在一定的关系，而这些关系限制了主观概率的任意性；(3) 在数据分析中，当数据量足够大时，主观概率的影响不大

主观解释 (con't)

- 主观概率必须满足 Kolmogorov 公理，否则可以构造一个赌局，使人认为合理而接受，但又必输无疑，因此理性个体的主观概率必须满足 Kolmogorov 公理
- 贝叶斯网早期主要用于专家系统，其结构和参数是通过咨询专家而获得的，所以主要是主观概率；之后，贝叶斯网更多的被用于分析数据（基于数据建立贝叶斯网模型），即参数学习（已知结构，学习参数）和结构学习（同时学结构和参数），当数据足够多时，主观概率对数据分析的影响不大
- 一些人指责概率的主观主义解释，认为它对所有命题、逻辑全知者等等指派精确概率的要求是不合情理的理想化

联合概率分布

- 对多个随机变量 $X_1, \ldots, X_n$，用联合概率分布 $P(X_1, \ldots, X_n)$ 来描述各变量所有可能的状态组合的概率。
- 联合分布是定义在所有变量状态空间的笛卡尔乘积上的函数：
  - $P(X_1, \ldots, X_n) : \otimes_{i=1}^{n} \Omega_{X_i} \to [0, 1]$
  - $\Sigma_{X_1,\ldots,X_n} P(X_1, \ldots, X_n) = 1$
- 联合分布通常表示为一张表，包含 $\Pi_{i=1}^{n} |\Omega_{X_i}|$ 个状态组合及其概率值。例，香港租房市场

	public	private	others
low	0.17	0.01	0.02
medium	0.44	0.03	0.01
upper medium	0.09	0.07	0.01
high	0	0.14	0.01

- 记 $\mathbf{X} = \{X_1, \ldots, X_n\}$，$\mathbf{Y}$ 是 $\mathbf{X}$ 的真子集 $(\mathbf{Y} \subset \mathbf{X})$，$\mathbf{Z} = \mathbf{X} \setminus \mathbf{Y}$。则相对于 $P(\mathbf{X})$，$\mathbf{Y}$ 的边缘分布 $P(\mathbf{Y})$ 定义为 $P(\mathbf{Y}) = \Sigma_{\mathbf{Z}} P(X_1, \ldots, X_n)$，称为边缘化

条件概率分布

- 条件概率：

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- 条件概率分布：

$$P(X = x \mid Y = y) = \frac{P(X = x,\ Y = y)}{P(Y = y)}$$

固定 y，让 x 在 Ω_X 上变动，得到函数 $P(X \mid Y = y)$（在给定 $Y = y$ 时变量 X 的条件概率分布）；
$P(X \mid Y) = \{P(X \mid Y = y) \mid y \in \Omega_Y\}$（给定 Y 时变量 X 的条件概率分布）

$$P(\mathbf{X} \mid \mathbf{Y}) = \frac{P(\mathbf{X}, \mathbf{Y})}{P(\mathbf{Y})}$$

- 链规则：

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2 \mid X_1) \cdots P(X_n \mid X_1, \ldots, X_{n-1})$$

边缘独立与条件独立

- 事件 A 与 B 相互独立：$P(A \cap B) = P(A)P(B)$
  - 当 $P(A) > 0$ 时，$P(B) = P(B \mid A)$.
- 事件 A 与 B 在给定 C 时相互条件独立：

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$$

  - 当 $P(B \cap C) > 0$ 时，$P(A \mid C) = P(A \mid B \cap C)$.
- 两个变量 X 和 Y 相互（边缘）独立，记为 $X \perp Y$:

$$P(X, Y) = P(X)P(Y)$$

  - 若 $P(Y = y) > 0$，则 $P(X) = P(X \mid Y = y)$.
- 三个随机变量 X, Y 和 Z，设 $P(Z = z) > 0, \forall z \in \Omega_Z$，$X$ 和 Y 在给定 Z 时相互条件独立，记为 $X \perp Y \mid Z$:

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

.

  - 若 $P(Y = y, Z = z) > 0$，
    则 $P(X \mid Y = y, Z = z) = P(X \mid Z = z)$.

边缘独立和条件独立 (con't)

考虑 3 个随机变量 X, Y 和 Z，设 $P(Z) > 0$，下列条件相互等价 $(X \perp Y \mid Z)$:

1. $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$;
2. $P(X \mid Y, Z) = P(X \mid Z)$，当 $P(Y, Z) > 0$;
3. $P(X, Y \mid Z) = f(X, Z)\, g(Y, Z)$, f 和 g 均为函数;
4. $P(X \mid Y, Z) = f(X, Z)$, f 为一函数，当 $P(Y, Z) > 0$;
5. $P(X, Y, Z) = P(X \mid Z)P(Y \mid Z)P(Z)$;
6. $P(X, Y, Z) = \frac{P(X,Z)\, P(Y,Z)}{P(Z)}$;
7. $P(X, Y, Z) = f(X, Z)\, g(Y, Z)$, f 和 g 均为函数。

贝叶斯定理

- 在考虑证据 $E = e$ 之前，对事件 $H = h$ 的概率估计 $P(H = h)$ 称为先验概率；而在考虑证据之后，对 $H = h$ 的概率估计 $P(H = h \mid E = e)$ 称为后验概率
- 贝叶斯定理（贝叶斯规则、公式）

$$P(H = h \mid E = e) = \frac{P(H = h)P(E = e \mid H = h)}{P(E = e)}$$

$$P(X \mid E = e) = \frac{P(X)P(E = e \mid X)}{P(E = e)}$$

例：贝叶斯定理

- 有一病人眼睛呈黄色，医生要判断他患乙肝的可能性
  - $D = t$：病人患有乙肝
  - $C = y$：病人眼睛呈黄色
  - 乙肝病人眼黄的概率为 0.8：$P(C = y \mid D = t) = 0.8$
  - 所有病人中，0.5% 的人患有乙肝：$P(D = t) = 0.005$
  - 所有病人中，10% 的人眼黄：$P(C = y) = 0.1$
- 利用贝叶斯公式，有：

$$P(D = t \mid C = y) = \frac{P(D = t)\, P(C = y \mid D = t)}{P(C = y)}$$
$$= \frac{0.005 \times 0.8}{0.1} = 0.04$$

即，病人患乙肝的概率是 0.04，这是先验概率 0.005 的 8 倍

概率论和人工智能

- 不确定性是人工智能系统所面临的一个重要问题
  - 智能系统对外部环境的观测往往是不完备的或是有误差的；
  - 推理涉及对复杂世界一定程度的抽象和简化，因此推理的前提往往因为例外情况而得不到完全满足。在实际应用中，推理前提的例外很多，不能穷举，不确定性提供了一个总结各种例外情况的机制；
  - 多数复杂领域并没有一套完备的理论，从而推理必须在不确定中进行；
  - 有些客观规律本身就是统计的、随机的、非确定性的。
- 早期人工智能研究处理不确定性的方式：
  - 发明新逻辑，用非单调推理的方式来处理推理前提的例外：默认逻辑、非单调模态逻辑、自认知逻辑、限定逻辑等
  - 为推理规则附加一个数字来量化不确定性：附加数字的语义解释有多种，包括确定因子、概率数、模糊集合论中的隶属度等

概率论和人工智能 (con't)

- 早期处理不确定性方式的主要困难：计算复杂性太高，推理结果的正确性不能保证
- 自 80 年代以来，概率论逐渐成为处理人工智能中不确定性的主流方法：
  - 到 70 年代末，越来越多的人认同概率的贝叶斯解释，将概率当做人的主观信念程度来看待，这使得在频率数据缺乏的情况下也可以使用概率；
  - 决策论与人工智能的结合推动了概率论的复兴；
    - 规范性系统：基于决策论（最大期望效用原则）的智能系统
    - 描述性系统：描述和总结专家解决问题的逻辑推理过程或所使用的启发式规则
  - 在 80 年代初，发现利用问题的结构可以把联合概率分布进行分解，从而极大的降低计算复杂度。

Outline

不确定性推理

- 不确定性推理：概率方法、非单调逻辑、模糊逻辑等
- （普通）使用概率方法进行不确定性推理：
  1. 把问题用一组随机变量 $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ 来刻画；
  2. 把关于问题的知识表示为一个联合概率分布 $P(\mathbf{X})$；
  3. 按概率论原则进行推理计算。
- 直接使用联合分布的复杂度极高
  - n 个二元变量的联合概率分布包括 $2^n - 1$ 个独立参数

利用条件独立降低复杂度

$$P(X_1, \ldots, X_n) = P(X_1)P(X_2 \mid X_1) \cdots P(X_n \mid X_1, X_2, \ldots, X_{n-1})$$
$$= \Pi_{i=1}^{n} P(X_i \mid X_1, X_2, \ldots, X_{i-1})$$

对任意 X_i，若存在 $\pi(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$，使得给定 $\pi(X_i)$，X_i 与 $\{X_1, \ldots, X_{i-1}\}$ 中的其他变量条件独立，即

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

则

$$P(X_1, \ldots, X_n) = \Pi_{i=1}^{n} P(X_i \mid \pi(X_i)).$$

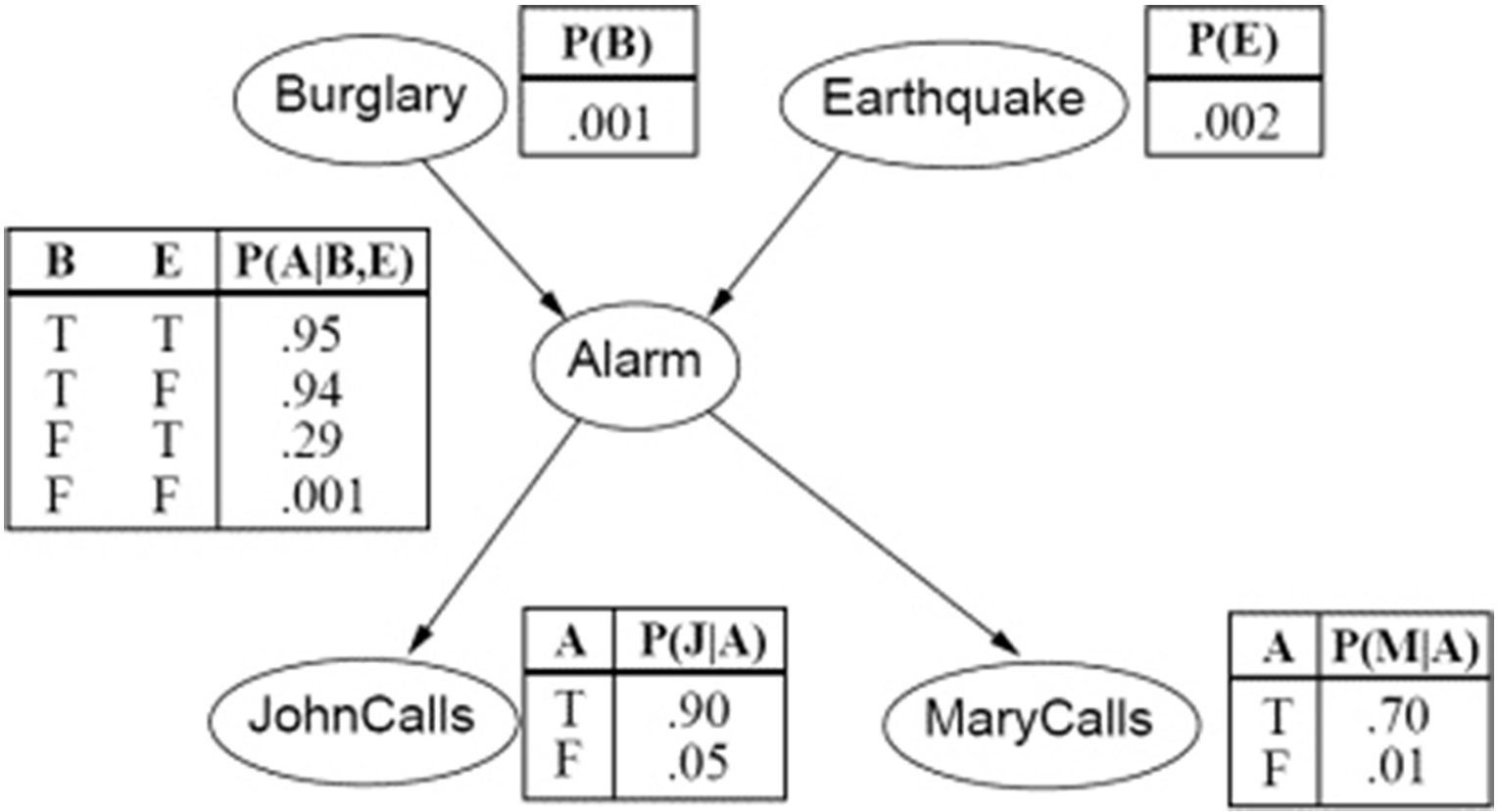
- 假设对任意 X_i, $\pi(X_i)$ 最多包含 m 个变量，上式右端的独立参数最多为 $n2^m$ 个

贝叶斯网的概念

- 根据 X_i 和 $\pi(X_i)$ 构造有向图：
  1. 把每个变量都表示为一个节点；
  2. 对每个节点 X_i，都从 $\pi(X_i)$ 中的每个节点画一条有向边到 X_i
- 贝叶斯网是一个有向无圈图，其中节点代表随机变量，节点间的边代表变量之间的直接依赖关系。每个节点都附有一个概率分布，根节点 X 的是它的边缘分布 $P(X)$，而非根节点 X 所附的是条件概率分布 $P(X \mid \pi(X))$.
- 贝叶斯网的语义：

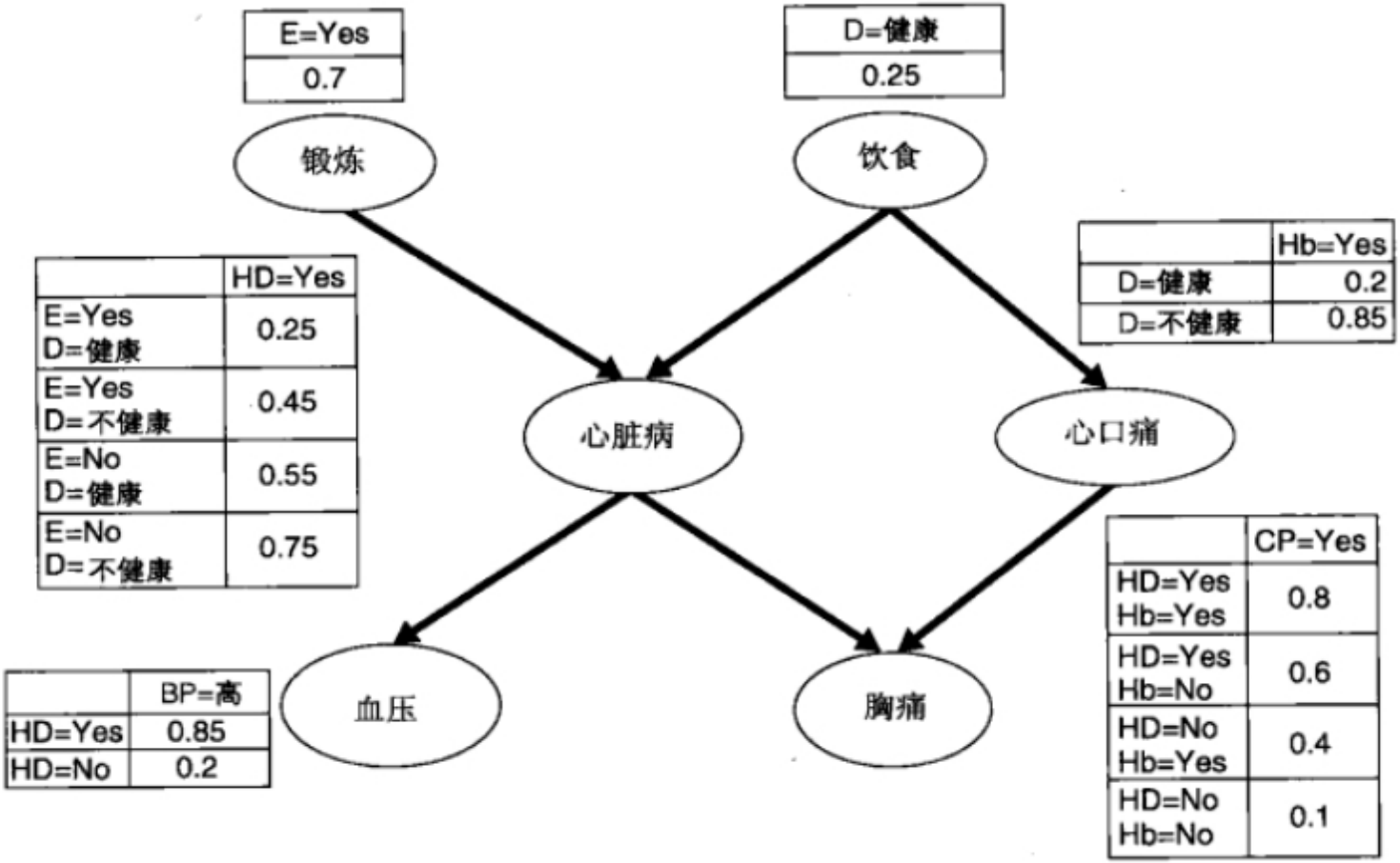
$$P(X_1, \ldots, X_n) = \Pi_{i=1}^{n} P(X_i \mid \pi(X_i))$$

贝叶斯网例子


Burglary
P(B)
.001
Earthquake
P(E)
.002
B E P(A|B,E)
T T .95
T F .94
F T .29
F F .001
Alarm
JohnCalls
A P(J|A)
T .90
F .05
MaryCalls
A P(M|A)
T .70
F .01

贝叶斯网例子

$$P(\text{心脏病}=\text{No}|\text{锻炼}=\text{No},\ \text{饮食}=\text{健康})$$
$$= 1 - P(\text{心脏病}=\text{Yes}|\text{锻炼}=\text{No},\ \text{饮食}=\text{健康})$$
$$= 1 - 0.55 = 0.45$$



贝叶斯网的构造

- 确定网络结构
  1. 选定一组刻画问题的随机变量 $\{X_1, X_2, \ldots, X_n\}$；
  2. 选择一个变量顺序 $\alpha = \langle X_1, X_2, \ldots, X_n \rangle$；
  3. 从一个空图出发，按照顺序 α 逐个将变量加入 $\mathcal{G}$ 中；
  4. 在加入变量 X_i 时，$\mathcal{G}$ 中的变量包括 $X_1, X_2, \ldots, X_{i-1}$：
     - 4.1 利用问题的背景知识，在这些变量中选择一个尽可能小的子集 $\pi(X_i)$，使得假设“给定 $\pi(X_i)$，X_i 与 $\mathcal{G}$ 中的其他变量条件独立”合理；
     - 4.2 从 $\pi(X_i)$ 中的每一个节点添加一条指向 X_i 的有向边。

  - 不同的变量顺序导致不同的网络结构，不同的网络结构表示了联合分布的不同分解，而不同的分解则意味着不同的复杂度
  - 建议用因果关系来决定变量顺序，原因在前，结果在后

- 确定网络参数：通过数据分析或从问题的特性直接得到

贝叶斯网构造举例

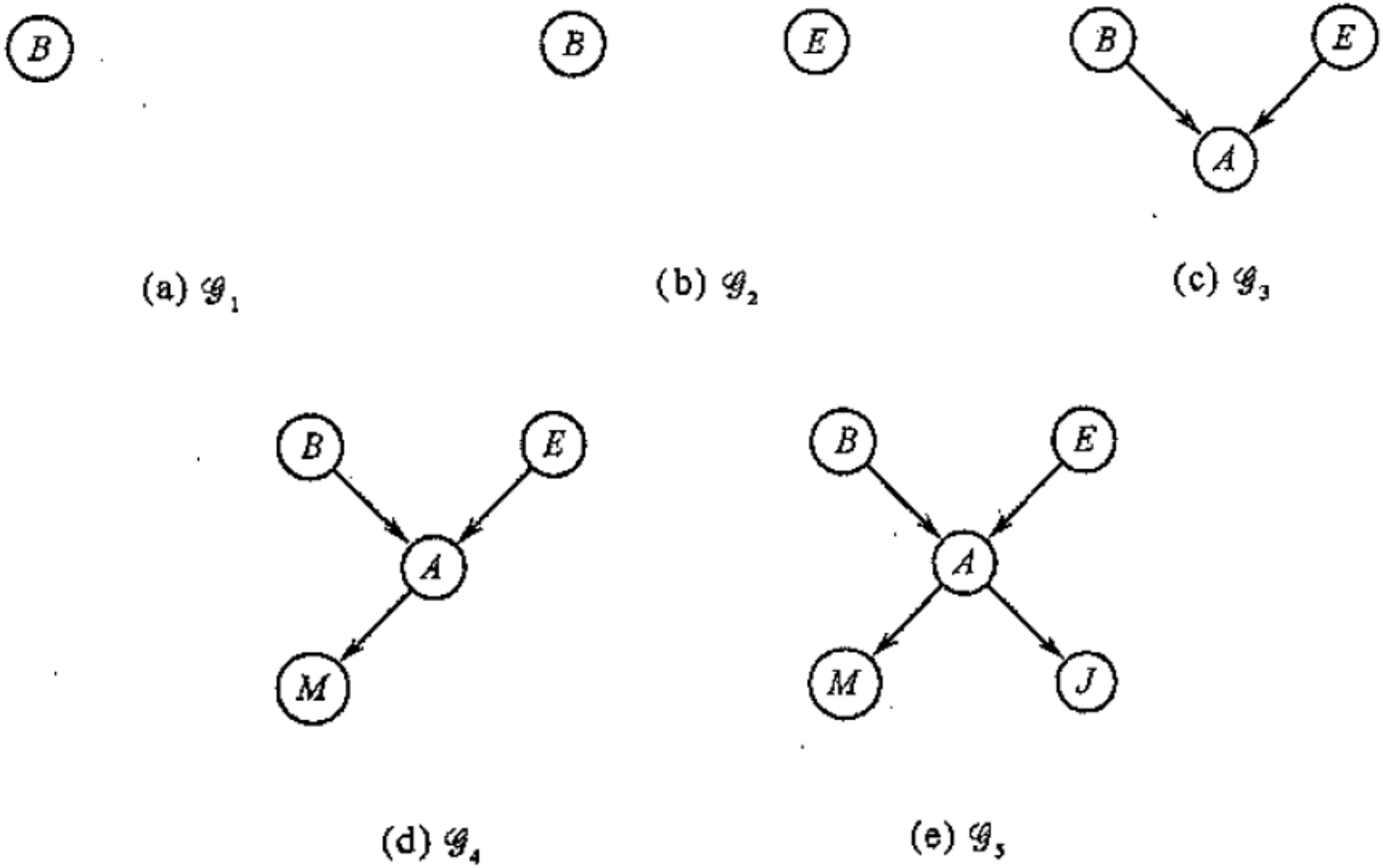


图 2.2　用序 $\alpha_1 = \langle B, E, A, M, J \rangle$ 构造 Alarm 贝叶斯网结构的过程

贝叶斯网构造举例 (con't)

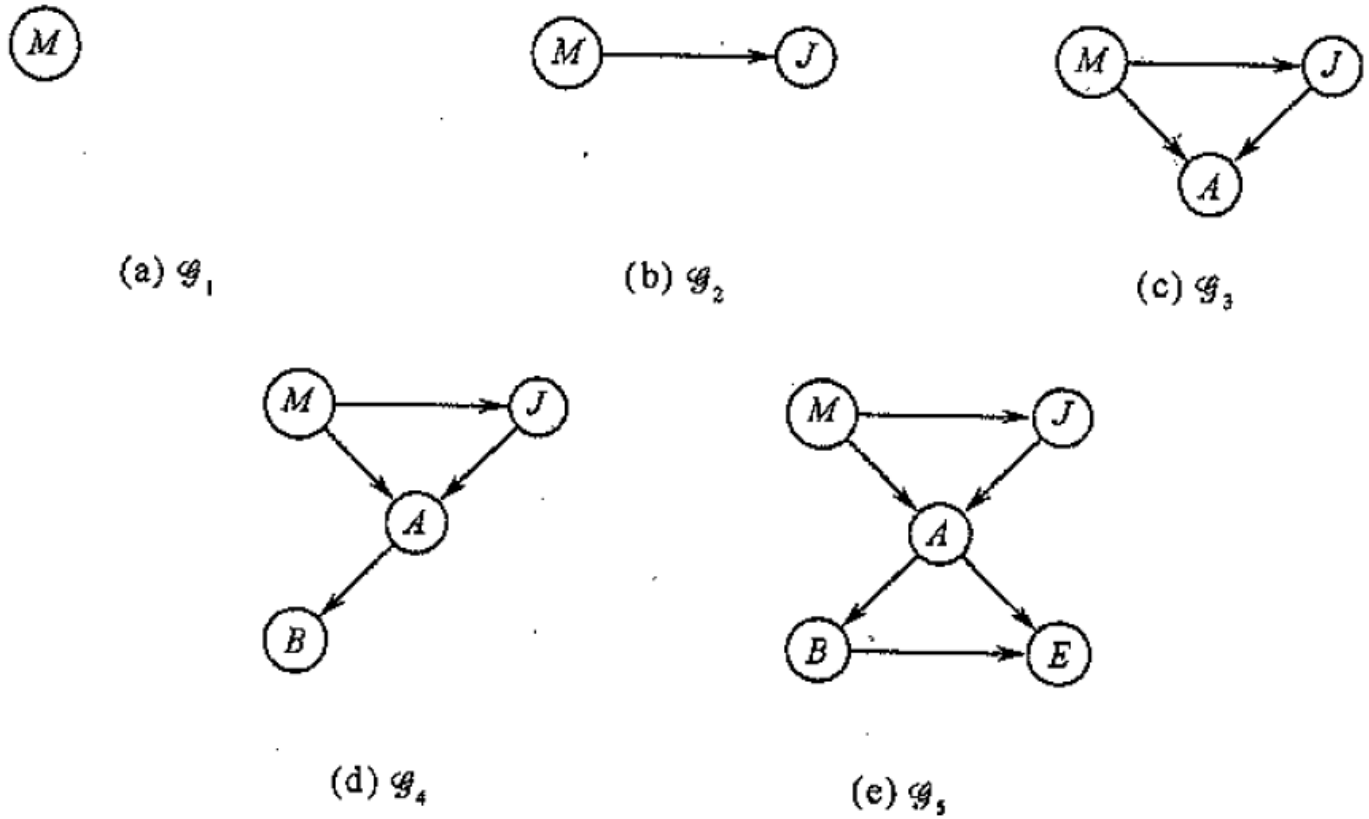


图 2.3　用序 $\alpha_2 = \langle M, J, A, B, E \rangle$ 构造 Alarm 贝叶斯网结构的过程

贝叶斯网构造举例 (con't)

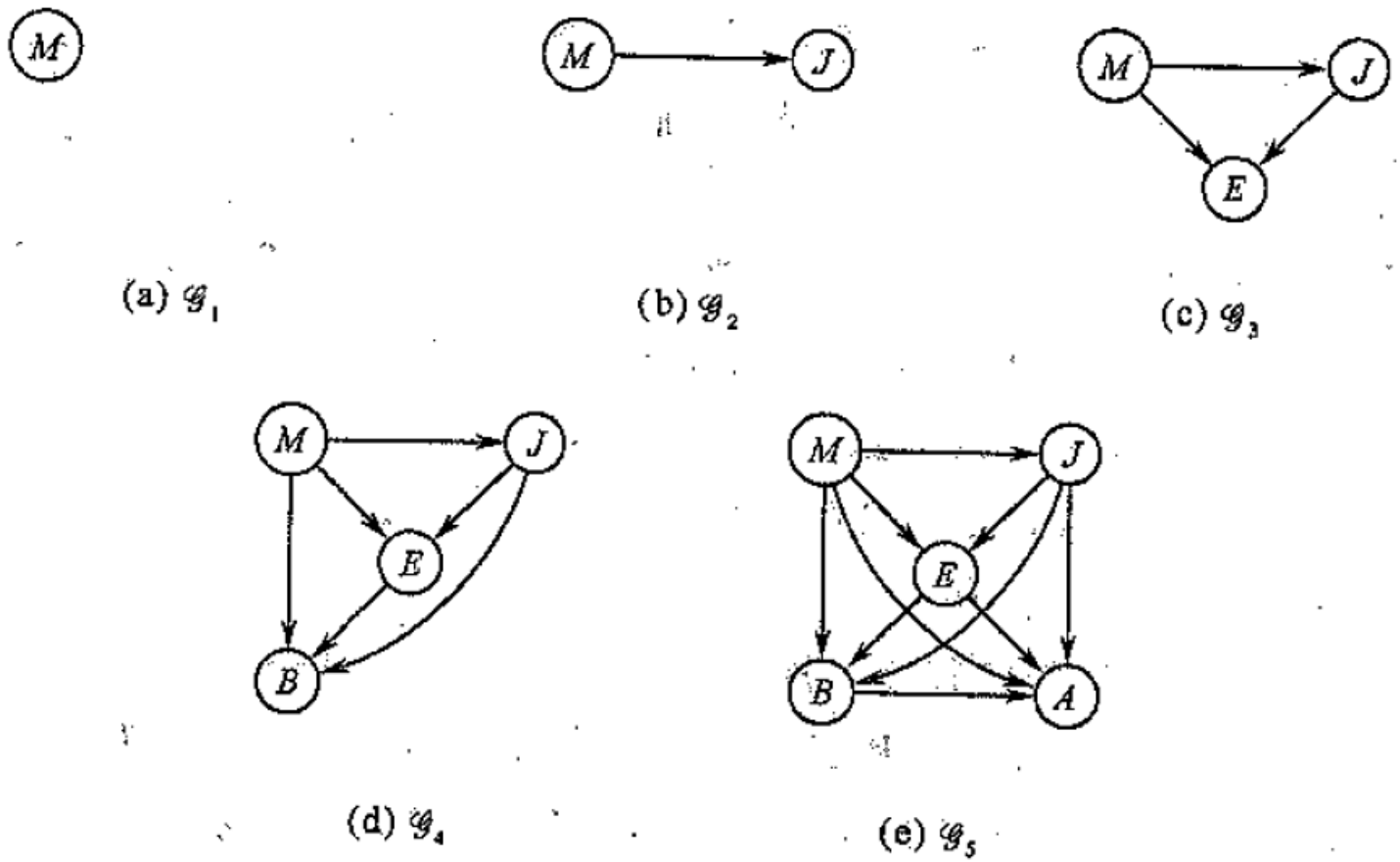


图 2.4 用序 $\alpha_3 = \langle M, J, E, B, A \rangle$ 构造 Alarm 贝叶斯网结构的过程

变量顺序选择

- 不同的变量顺序导致不同的网络结构，不同的网络结构表示了联合分布的不同分解，而不同的分解则意味着不同的复杂度。应该按照什么原则来选择变量顺序？
  - Smith (1989) 认为应以模型的复杂性为标准
  - Howard and Matheson (1984) 认为变量顺序的选取应以条件概率评估的难易程度为标准
  - Pearl (2000) 提出应该用因果关系来决定变量顺序，原因在前，结果在后

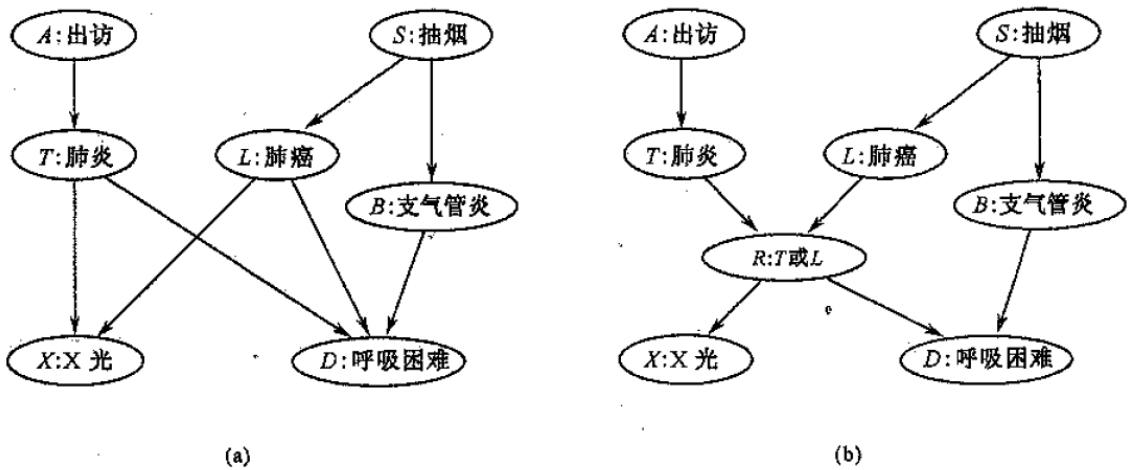


图 2.5　肺病诊断因果网

因果关系与贝叶斯网

- 因果关系还没有一个能被广泛接受的严格定义。对它到底是客观世界本是的属性，还是人的意识为了理解世界而创造出来的主观概念，还没有达成共识
  - 有些因果关系十分明显，例如：支气管炎会导致呼吸困难
  - 有些因果关系不明显，例如：抽烟（S）与肺癌（L）的统计关联也可以通过基因（G）来解释

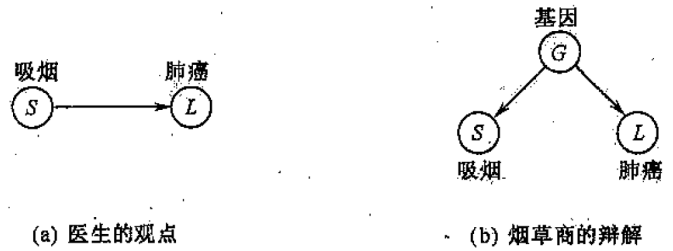


图 2.6　吸烟与肺癌的因果关系

- 实际应用中，采用如下方法：假设一个万能的上帝可以介入改变任何变量的状态，如果知道变量 X 的状态被上帝改变后会影响你对 Y 的信度，而反过来知道 Y 的状态被改变并不影响你对 X 的信度，那么就说 X 是 Y 的原因
- 当利用因果关系建立贝叶斯网时，实际上是在基于因果关系进行条件独立假设，所有的假设可以归纳为因果马尔科夫假设

Outline

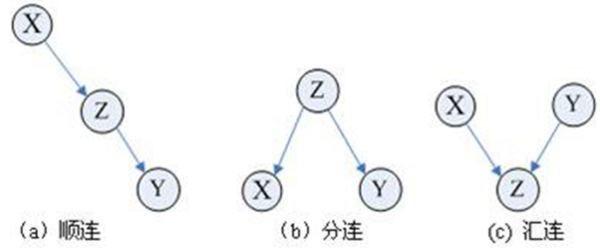
概述

- 贝叶斯网是概率论和图论相结合的产物
- 可以从概率论的角度讨论变量间的依赖与独立，也可以从图论的角度讨论节点间的连通与分隔；两者有深刻的联系
  - 通过图论准则可以判别变量间条件独立关系
  - X 与 Y 不直接相连，通过其他变量才能在两者间传递信息；如果 X 和 Y 之间的所有信息通道都被阻塞，那么信息就无法再它们之间传递

图分隔

图分隔，有向分隔（d-separate，d-分隔）

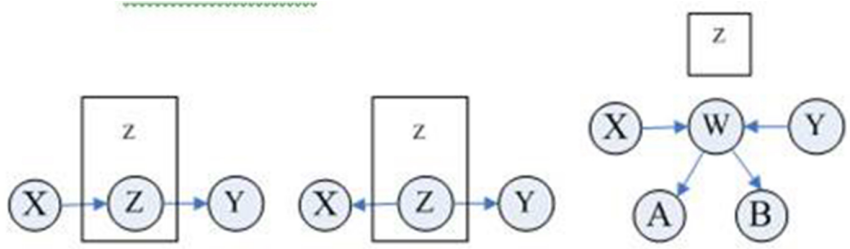
- 变量 X 和 Y 通过第三个变量 Z 间接相连的三种情况
  - 顺连：若 Z 未知，则对 X 的了解会影响对 Z 的信度，进而影响对 Y 的信度；若 Z 已知，则对 X 的了解就不会影响关于 Z 的信度
  - 分连：当 Z 未知，信息可以在 X 和 Y 之间传递，它们关联；当 Z 已知，信息不能在 X 和 Y 之间传递，从而条件独立
  - 汇连：分连代表一因多果，汇连代表多因一果。当 Z 未知，源自 X（或 Y）的信息会从 Z “漏掉”；当 Z 已知，“漏洞”被堵上



- 阻塞（block）：设 $\mathbf{Z}$ 为一节点集合，X 和 Y 是不在 $\mathbf{Z}$ 中的两个节点。考虑 X 和 Y 之间的一条通路 α（无向图中路径）。如果满足下面条件之一，则称 α 被 $\mathbf{Z}$ 所阻塞：
  1. α 上有一个在 $\mathbf{Z}$ 中的顺连节点；
  2. α 上有一个在 $\mathbf{Z}$ 中的分连节点；
  3. α 上有一个汇连节点 W，它和它的后代节点均不在 $\mathbf{Z}$ 中。

图分隔（con't）

- X 和 Y 之间的通路被 $\mathbf{Z}$ 阻塞的三种情况



(a) 顺连节点 $z \in \mathbf{Z}$　(b)分连节点 $z \in \mathbf{Z}$　(c) 汇连节点 W 及其后代均不在 Z 内

- 如果 X 和 Y 之间的所有通路都被 $\mathbf{Z}$ 阻塞，则说 $\mathbf{Z}$ 有向分隔 (directed separate) X 和 Y，简称 d-separate, d-分隔。
- 定理（整体马尔科夫性）：设 X 和 Y 为贝叶斯网 $\mathcal{N}$ 中的两个变量，$\mathbf{Z}$ 为 $\mathcal{N}$ 中一个不包含 X 和 Y 的节点集合。如果 $\mathbf{Z}$ d-分隔 X 和 Y，那么 X 和 Y 在给定 $\mathbf{Z}$ 时条件独立，即

$$X \perp Y \mid \mathbf{Z}$$

- d-分隔是图论的概念，而条件独立是概率论的概念，所以定理揭示了贝叶斯网络图论侧面和概率论侧面之间的关系

图分隔举例

- 设 $\mathbf{Z}=\emptyset$，先考虑节点 A 和 E，它们之间有通路 "$A \to D \to B \to E$"，它未被 $\mathbf{Z}$ 阻塞，因此 A 和 E 不被 $\mathbf{Z}$ d-分隔。再考虑节点 A 和 C，两者之间有 4 条通路，都存在汇连节点，且这些节点及其后代均不在 $\mathbf{Z}$ 中，所以 A 和 C 被 $\mathbf{Z}$ d-分隔
- 设 $\mathbf{Z}=\{B, M\}$，考察 A 和 C，两者之间的 4 条通路中，2 条在节点 B 处有顺连结构，所以被 B 阻塞；另 2 条在 K 处有汇连结构，但 K 有一个后代节点 $M \in \mathbf{Z}$，所以这 2 条通路不被阻塞，故 A 和 C 不被 $\mathbf{Z}$ d-分隔

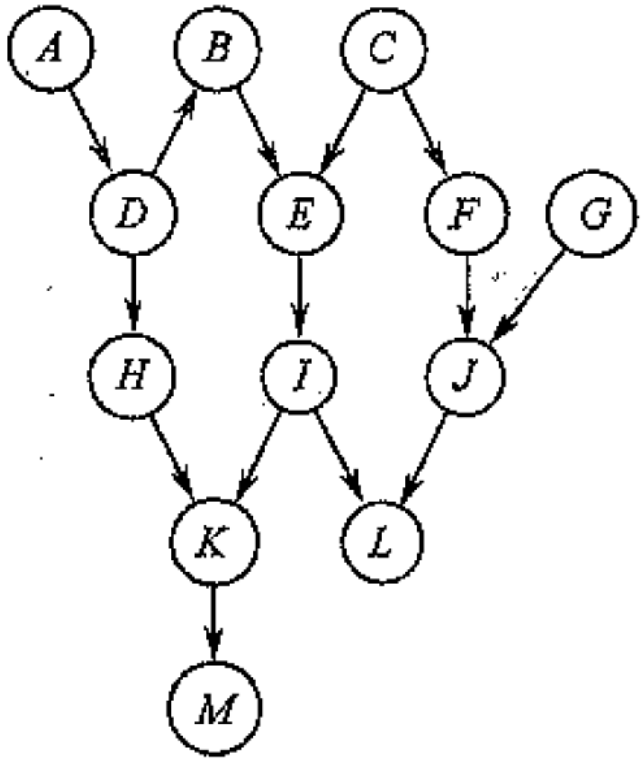


图 3.4　d-分隔示例

马尔科夫边界与端正图

- 马尔科夫边界 (Markov boundary)：在贝叶斯网络中，一个节点 X 的马尔科夫边界包括其父节点、子节点、以及子节点的父节点
- 推论：在一个贝叶斯网络中，给定变量 X 的马尔科夫边界 $mb(X)$，则 X 条件独立于网络中所有其它变量
- 端正图 (Moral graph)：设 G 为一有向无环图，如果将 G 中每个节点的不同父节点结合，即在他们之间加一条边，然后去掉所有边的方向，所得到的无向图称为 G 的端正图

有向分隔和无向分隔

- 设 $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ 为无向图中的两两不相交的节点集合，如果从图中除去 $\mathbf{Z}$ 中的节点后，$\mathbf{X}$ 和 $\mathbf{Y}$ 之间没有通路存在，则称 $\mathbf{Z}$ 无向分隔（undirected separate）$\mathbf{X}$ 和 $\mathbf{Y}$，简称为 $\mathbf{Z}$ u-分隔
- 定理（有向分隔与无向分隔）：设 $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ 是有向无圈图 $\mathcal{G}$ 中三个两两不相交的节点集合。$(\mathcal{G}_{an(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})})^m$ 表示把 $\mathcal{G}$ 限制在 $an(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})$ 上，再将其结果端正化而得到的无向图。那么，集合 $\mathbf{Z}$ 在 $\mathcal{G}$ 中 d-分隔 $\mathbf{X}$ 和 $\mathbf{Y}$ 的充分必要条件是它在 $(\mathcal{G}_{an(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})})^m$ 中 u-分隔 $\mathbf{X}$ 和 $\mathbf{Y}$.
  - $an(\mathcal{X})$ 表示包含 $\mathcal{X}$ 的最小祖先闭集（每个节点的祖先节点都在集合中）
  - 无圈图 $\mathcal{G}$ 在 $\mathbf{Y}$ 上的限制，是从 $\mathcal{G}$ 中除去不属于 $\mathbf{Y}$ 的节点及其相连的边得到的图

Outline

贝叶斯网络推理 (Inference)

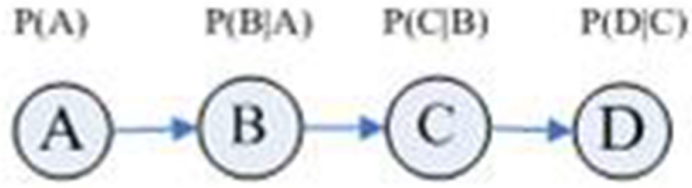
- 贝叶斯网络可以利用变量间的条件独立对联合分布进行分解，降低参数个数
- 推理 (inference) 是通过计算来回答查询的过程
- 贝叶斯网中的推理问题有三大类：
  - 后验概率问题：$P(\mathbf{Q} \mid \mathbf{E} = \mathbf{e})$
    - 从结果到原因的诊断过程，从原因到结果的预测推理，同一结果的不同原因之间的原因关联推理，混合推理
  - 最大后验假设问题 (Maximum A Posteriori hypothesis, MAP)：

$$\mathbf{h}^* = argmax_{\mathbf{h}} P(\mathbf{H} = \mathbf{h} \mid \mathbf{E} = \mathbf{e})$$

    - 给定证据，计算一些变量的后验概率最大的状态组合
  - 最大可能解释问题 (Moist Probable Explanation, MPE)
    - 解释指网络中全部变量的一个与 $\mathbf{E} = \mathbf{e}$ 相一致的状态组合，概率最大的那个解释就是最大可能解释
    - 最大可能解释问题可视为最大后验假设问题的一个特例，即假设变量为网络中所有非证据变量

变量消元算法 (Variable Elimination)

利用概率分解降低推理复杂度



$$P(D) = \sum_{A,B,C} P(A,B,C,D) = \sum_{A,B,C} P(A)P(B \mid A)P(C \mid B)P(D \mid C)$$

$$P(D) = \sum_{C} P(D \mid C) \sum_{B} P(C \mid B) \sum_{A} P(A)P(B \mid A)$$

- 使得运算局部化。消元过程实质上就是一个边缘化的过程
- 不同的消元顺序导致不同的计算复杂度
- 寻找最优消元顺序是 NP-hard 问题，可以用启发式算法：最大势搜索，最小缺边搜索

变量消元法 (con't)

VE($\mathcal{N}$,**E**,**e**,**Q**,ρ)

输入：$\mathcal{N}$—— 一个贝叶斯网；**E** —— 证据变量；

e —— 证据变量的取值；**Q** —— 查询变量；

ρ —— 消元顺序，包含所有不在 **Q**∪**E** 中的变量.

输出：$P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$.

1：$\mathcal{F} \leftarrow \mathcal{N}$中所有概率分布的集合；

2：在$\mathcal{F}$的因子中，将证据变量 **E** 设置为其观测值 **e**；

3：**while** $(\rho \neq \phi)$

4：　　设 Z 为 ρ 中排在最前面的变量，将 Z 从 ρ 中删去；

5：　　$\mathcal{F} \leftarrow \mathrm{Elim}(\mathcal{F}, Z)$；

6：**end while**

7：将 $\mathcal{F}$ 中所有因子相乘，得到一个 **Q** 的函数 $h(\mathbf{Q})$；

8：**return** $h(\mathbf{Q}) \Big/ \sum_{\mathbf{Q}} h(\mathbf{Q})$.

Elim ($\mathcal{F}$, Z)

输入：$\mathcal{F}$—— 一个函数集合；

Z —— 待消元变量.

输出：另一个函数集合.

1：从 $\mathcal{F}$ 中删去所有涉及 Z 的函数，设这些函数是 $\{f_1, \cdots, f_k\}$；

2：$g \leftarrow \prod_{i=1}^{k} f_i$；

3：$h \leftarrow \sum_{Z} g$；

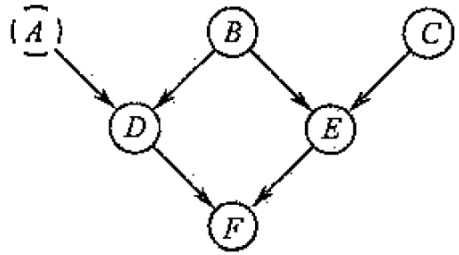
4：将 h 放回 $\mathcal{F}$；

5：**return** $\mathcal{F}$.

图 4.3　变量消元算法

变量消元法举例

在下图所示贝叶斯网中，设证据为 $\{F=0\}$，考虑调用 VE 算法计算 $P(A \mid F=0)$



设变量消元顺序 $\rho=\langle C, E, B, D\rangle$，贝叶斯网给出的联合分布的分解为：

$$\mathcal{F}=\{P(A), P(B), P(C), P(D \mid A, B), P(E \mid B, C), P(F \mid D, E)\};$$

(1) VE 算法首先设置证据 $F=0$，得

$$\mathcal{F}=\{P(A), P(B), \underline{P(C)}, P(D \mid A, B), \underline{P(E \mid B, C)}, P(F=0 \mid D, E)\};$$

(2) 要消去 C，与之关联的函数为 $P(C)$ 和 $P(E \mid B, C)$，消去 C，得

$$\mathcal{F}=\{P(A), P(B), P(D \mid A, B), \underline{P(F=0 \mid D, E)}, \underline{\psi_1(B, E)}\},$$

这里 $\psi_1(B, E)=\Sigma_C P(C) P(E \mid B, C)$;

变量消元法举例 (con't)

(3) 消去 E，得

$$\mathcal{F} = \{P(A), \underline{P(B)}, \underline{P(D \mid A, B)}, \underline{\psi_2(B, D)}\},$$

这里 $\psi_2(B, D) = \Sigma_E P(F = 0 \mid D, E)\psi_1(B, E)$;
(4) 消去 B，得

$$\mathcal{F} = \{P(A), \underline{\psi_3(A, D)}\},$$

这里 $\psi_3(A, D) = \Sigma_B P(B)P(D \mid A, B)\psi_2(B, D)$;
(5) 消去 D，得

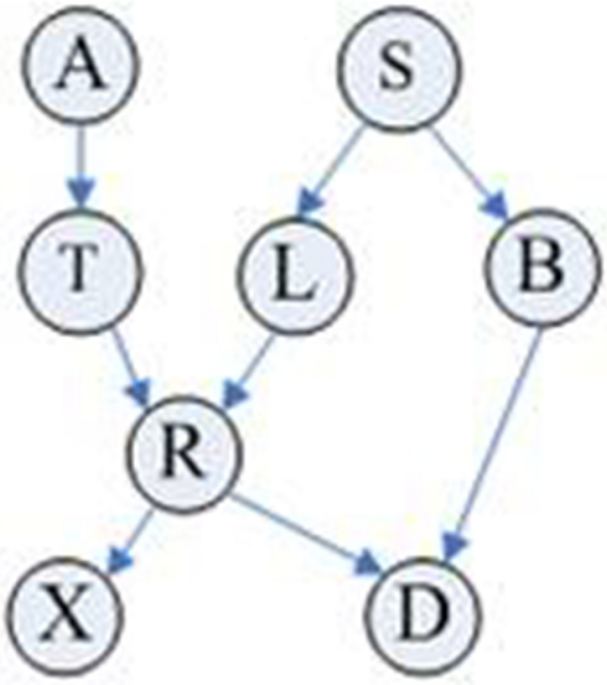
$$\mathcal{F} = \{P(A), \psi_4(A)\},$$

这里 $\psi_4(A, D) = \Sigma_D \psi_3(A, D)$;
(6) 计算 $h(A) = P(A)\psi_4(A)$;
(7) 返回 $P(A \mid F = 0) = \frac{h(A)}{\Sigma_A h(A)}$.

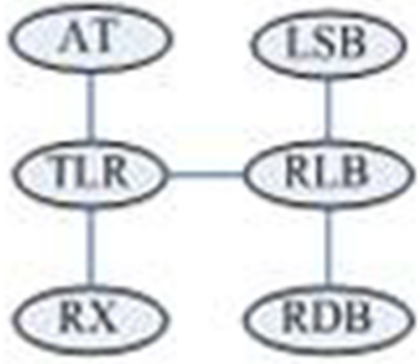
团树传播算法

- 实用中，需要在同一贝叶斯网中进行多次不同的推理，而两次不同的推理之间往往存在一些相同的步骤
- 团树传播算法利用步骤共享来加快推理
- 团树（clique tree）是一种无向树，其中每一个节点代表一个变量集合，称为团 (clique)。团树必须满足变量连通性，即包含同一变量的所有团所导出的子图必须是连通的
  - 如果团树中的两个团 $\mathbf{C}_1$ 和 $\mathbf{C}_2$ 同时包含某变量 X，那么在连接 $\mathbf{C}_1$ 和 $\mathbf{C}_2$ 的通路上的所有团都必须包含 X
- 团树 $\mathcal{T}$ 覆盖（cover）贝叶斯网 $\mathcal{N}$，如果它满足以下两个条件：
  1. 对 $\mathcal{N}$ 中任一变量 X，在 $\mathcal{T}$ 中有一个团 $\mathbf{C}$，使得 $X \in \mathbf{C}$ 且 $\pi(X) \subseteq \mathbf{C}$（$\pi(X)$: X 的父节点集合）；
  2. $\mathcal{T}$ 中所有团的并集刚好是 $\mathcal{N}$ 中所有变量的集合。

团树传播算法（con't）



(a) 贝叶斯网络



(b) 团树

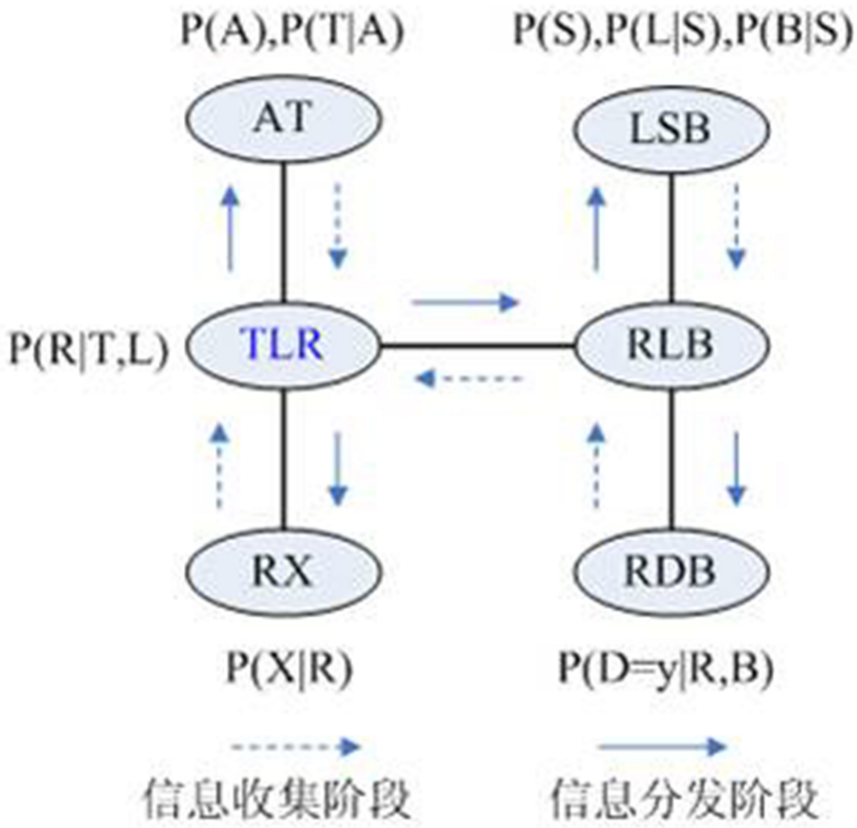
团树传播算法（con't）

- 团树传播算法：用团树组织变量消元的算法。计算共享
- 团树传播算法基本步骤：
  1. 将贝叶斯网络转化为团树
  2. 团树初始化：将贝叶斯网中的概率函数分配到团树的各节点加以储存
  3. 设置证据：改变相应团中所存储的概率函数
  4. 选择一个包含查询变量 Q 的团 $\mathbf{C}_Q$ 作为推理的枢纽节点
  5. 对 $\mathbf{C}_Q$ 的相邻节点逐一调用 `CollectMessage`（从相邻节点获取信息）
  6. 根据所收到的信息及自身的概率函数，得到关于 $\mathbf{C}_Q$ 的函数
  7. 消去 $\mathbf{C}_Q$ 中除 Q 以外的变量，并将结果归一化

团树传播算法（con't）

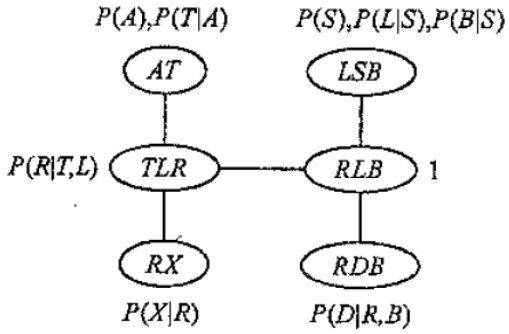
- 团树传播本质上与变量消元法没有区别，它是变量消元的另一种组织形式
- 用团树组织变量消元的优点在于，它能使我们清楚地看到两次不同的推理计算之间哪些步骤是相同的，从而可以进行步骤共享
- 为了在两次不同的推理之间实现计算共享，引入 `SaveMessage` 和 `RetrieveMessage` 来存储和利用中间结果

团树传播算法（con't）

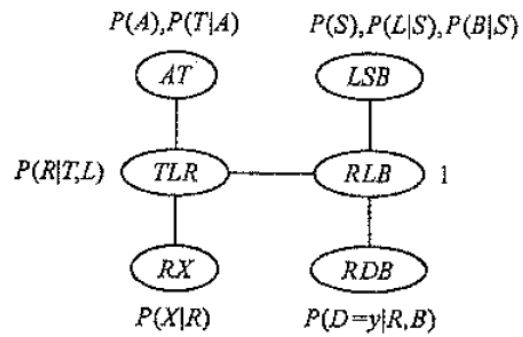


- 变量消元和团树传播算法都是精确推理算法

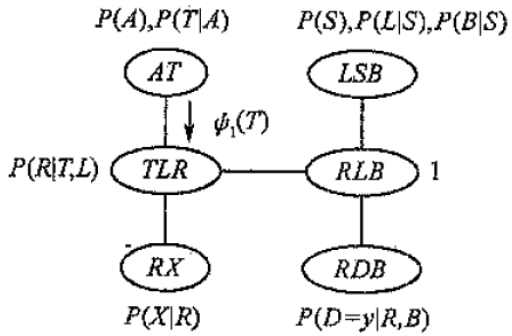
团树传播算法举例



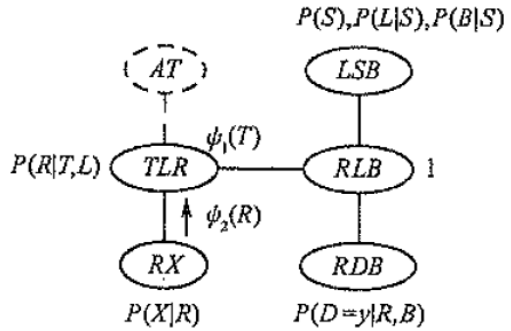
(a) 团树初始化



(b) 设置证据 $D=y$



(c) [AT]传送ψ_1到[TLR]

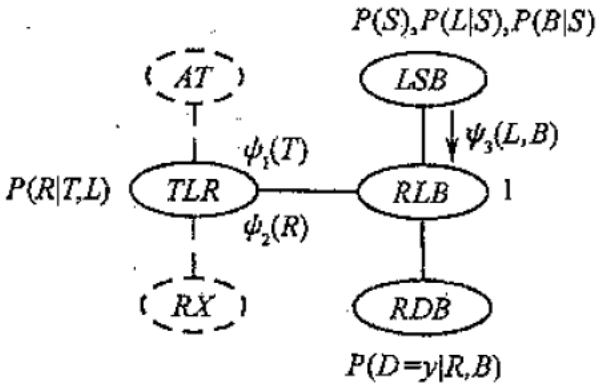


(d) [RX]传送ψ_2到[TLR]

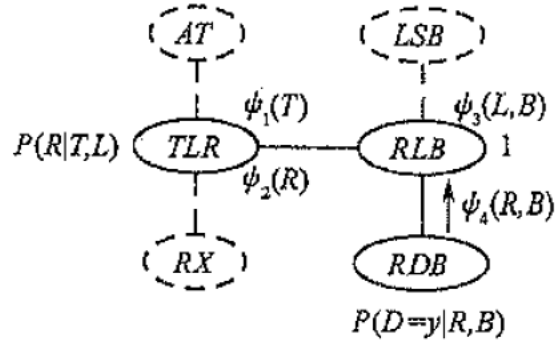
(1) 从 $[AT]$ 到 $[TLR]$ 的信息: $\psi_1(T) = \sum_A P(A)P(T \mid A)$;

(2) 从 $[RX]$ 到 $[TLR]$ 的信息: $\psi_2(R) = \sum_X P(X \mid R)$;

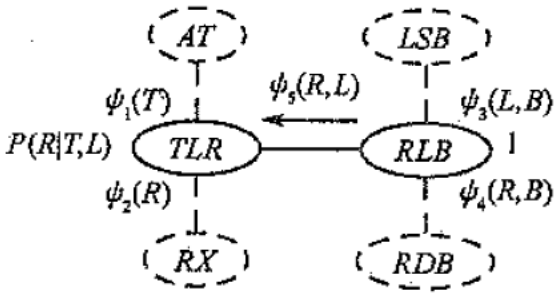
团树传播算法举例 (con't)



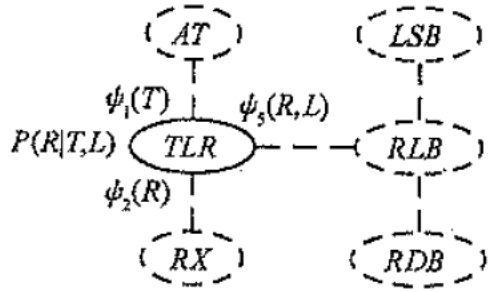
(e) [LSB]传送ψ_3到[RLB]



(f) [RDB]传送ψ_4到[RLB]



(g) [RLB]传送ψ_5到[TLR]



(h) 从[TLR]所收到的信息和储存的函数计算$P(T|D=y)$

团树传播算法举例 (con't)

(3) 从 $[RLB]$ 到 $[TLR]$ 的信息：由于除 $[TLR]$ 以外，$[RLB]$ 还有两个邻居 $[LSB]$ 和 $[RDB]$，所以先要计算它们的信息，分别为：

$$\psi_3(L,B) = \sum_S P(S)P(L \mid S)P(B \mid S)$$

$$\psi_4(R,B) = P(D = y \mid R,B)$$

最后从 $[RLB]$ 到 $[TLR]$ 的信息为：
$\psi_5(R,L) = \sum_B \psi_3(L,B) \cdot \psi_4(R,B)$;
(4) 枢纽节点得到关于 $\mathbf{C}_Q$ 的函数

$$h(T,L,R) = \psi_1(T)\psi_2(R)\psi_5(R,L)P(R \mid T,L)$$

(5) $P(T \mid D = y) = \frac{\sum_{L,R} h(T,L,R)}{\sum_{T,L,R} h(T,L,R)}$

随机抽样算法

- 当网络节点众多并且连接稠密时，它们的计算复杂度高，可以考虑近似推理算法
  - 随机抽样算法
  - 变分法
- 随机抽样算法：是一类应用于数值积分和统计物理中的近似计算方法。基本思想是从某个概率分布随机抽样，生成一组样本，然后从样本出发近似估计要计算的量
- 随机抽样算法分为：
  - 重要性抽样（important sampling）算法
  - 马尔科夫链蒙特卡洛（Markov chain Monte Carlo, MCMC）算法

重要性抽样算法

- 考虑计算积分

$$I = \int_{\Omega_{\mathbf{X}}} f(\mathbf{X})\ d\mathbf{X}$$

- 为了近似计算这一积分，重要性抽样法将上式改写为

$$I = \int_{\Omega_{\mathbf{X}}} \frac{f(\mathbf{X})}{p(\mathbf{X})} p(\mathbf{X})\ d\mathbf{X}$$

  要求，对 $\mathbf{X}$ 的任一取值 $\mathbf{x}$，如果 $f(\mathbf{X} = \mathbf{x}) \neq 0$，那么 $p(\mathbf{X} = \mathbf{x}) \neq 0$

- 重要性抽样法从 $p(\mathbf{X})$ 独立的抽取 m 个样本 $D_1, D_2, \ldots, D_m$，并给这些样本对积分 I 进行估计：

$$\hat{I}_m = \frac{1}{m} \Sigma_{i=1}^{m} \frac{f(D_i)}{p(D_i)}$$

- 可以证明，当样本量 m 趋于无穷时，$\hat{I}_m$ 几乎必然收敛于 I

重要性抽样算法与概率推理

- 设 $\mathbf{W}$ 为一些变量的集合，$\mathbf{Y}$ 是 $\mathbf{W}$ 的一个子集合，$\mathbf{Z} = \mathbf{W} \setminus \mathbf{Y}$，并设 $\mathbf{y}$ 为 $\mathbf{Y}$ 的一个取值。定义：

$$\chi_{\mathbf{Y}=\mathbf{y}}(\mathbf{W}) = \chi_{\mathbf{Y}=\mathbf{y}}(\mathbf{Y}, \mathbf{Z}) = \begin{cases} 1, & \text{若 } \mathbf{Y} = \mathbf{y} \\ 0, & \text{若否} \end{cases}$$

- 条件概率可表示为如下期望形式：

$$\begin{aligned} P(\mathbf{Q} = \mathbf{q} \mid \mathbf{E} = \mathbf{e}) &= \frac{P(\mathbf{Q} = \mathbf{q}, \mathbf{E} = \mathbf{e})}{P(\mathbf{E} = \mathbf{e})} \\ &= \frac{\Sigma_{\mathbf{X}} \chi_{\mathbf{Q}=\mathbf{q}}(\mathbf{X}) \chi_{\mathbf{E}=\mathbf{e}}(\mathbf{X}) P(\mathbf{X})}{\Sigma_{\mathbf{X}} \chi_{\mathbf{E}=\mathbf{e}}(\mathbf{X}) P(\mathbf{X})} \end{aligned}$$

- 重要性分布可以有多种选择，若选择 $P(\mathbf{X})$ 本身作为重要性分布，则称为逻辑抽样 (logic sampling)

MCMC 算法

- 对于重要性抽样算法，不同样本之间相互独立；而对于 MCMC 算法，不同样本之间不是相互独立的
- MCMC 算法—吉布斯抽样 (Gibbs sampling)。它首先随机生成一个与证据 $\mathbf{E} = \mathbf{e}$ 相一致的样本 D_1 作为起始样本。此后，每个样本 D_i 的产生都依赖于前一个样本 D_{i-1}，且 D_i 与 D_{i-1} 最多只有一个非证据变量的取值不同，记改变量为 X
- X 的取值可以从非证据变量中随机选取，也可以按某个固定顺序轮流决定
- 在 D_i 中，X 的值通过随机抽样决定，抽样分布是：

$$P(X \mid D_{i-1} \setminus X) = P(X \mid mb(X)_{D_{i-1}})$$

  $mb(X)_{D_{i-1}}$ 表示 X 的马尔可夫边界的取值等于在 D_{i-1} 中的当前取值
- 当样本数趋于无穷时，马氏链理论保证了算法返回的结果收敛于真正的后验概率。吉布斯抽样的缺点是收敛速度慢，因为马氏链往往需要花很长时间才能真正达到平稳分布

贝叶斯网学习

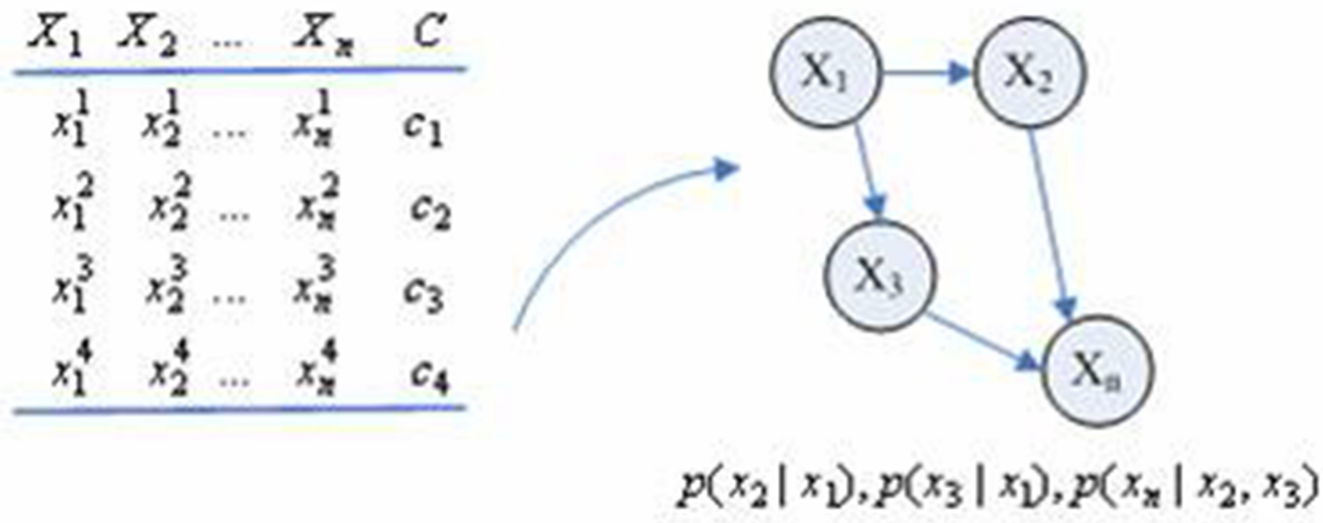


图 5.4 从数据中学习贝叶斯网络

- 结构学习：发现变量之间的图关系
- 参数学习：决定变量之间互相关联的量化关系

贝叶斯网应用

- 医疗诊断
- 故障诊断：工业设备，计算机系统故障诊断，垃圾邮件过滤
- 金融分析
- 模式识别：分类，语义理解
- 军事（目标识别，多目标跟踪，战争身份识别等）
- 生态学：分析人类活动对环境及动物的影响
- 生物信息学（贝叶斯网络在基因连锁分析中应用）
- 编码学
- 机器学习：分类、聚类
- 时序数据和动态模型

小结

- 贝叶斯网的目的是进行概率推理，$P(\mathbf{Q} \mid \mathbf{E} = \mathbf{e})$
- 贝叶斯网是概率论与图论结合的产物，一方面用图论的语言直观揭示问题的结构，另一方面按概率论的原则对问题结构加以利用
- 贝叶斯网把联合概率分解成一系列简单模块，从而降低难度
- 如果 $\mathbf{Z}$ d-分隔 X 和 Y，则 $X \perp Y \mid \mathbf{Z}$
- 贝叶斯网中后验概率问题，可以通过变量消元法或团树传播算法精确求解，也可以通过随机抽样算法近似求解