贝叶斯网（Bayesian Networks）

University of Science and Technology of China

June 4, 2013

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

1. 概率论基础
2. 贝叶斯网基础
3. 图分隔与变量独立
4. 贝叶斯网推理

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

$$P(X = x) \geq 0, \forall x \in \Omega_X \text{ 简记为 } P(X) \geq 0$$
$$\Sigma_{x \in \Omega_X} P(X = x) = 1 \text{ 简记为 } \Sigma_X P(X = x) = 1.$$

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

	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：

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

- 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 $P(X \mid Y = y, Z = z) = P(X \mid Z = z)$.

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）

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

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

Outline

不确定性推理

- 不确定性推理：概率方法、非单调逻辑等
- （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

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

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

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

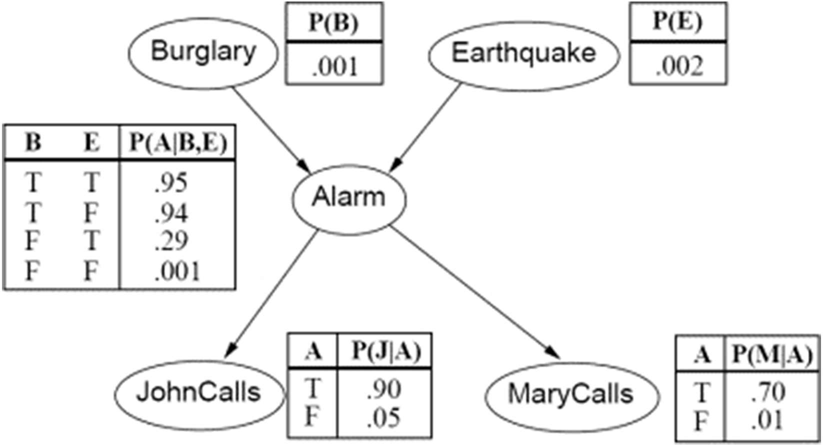
- 假设对任意 $X_i, \pi(X_i)$ 最多包含 m 个变量，上式右端的独立参数最多为 $n2^m$ 个

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：

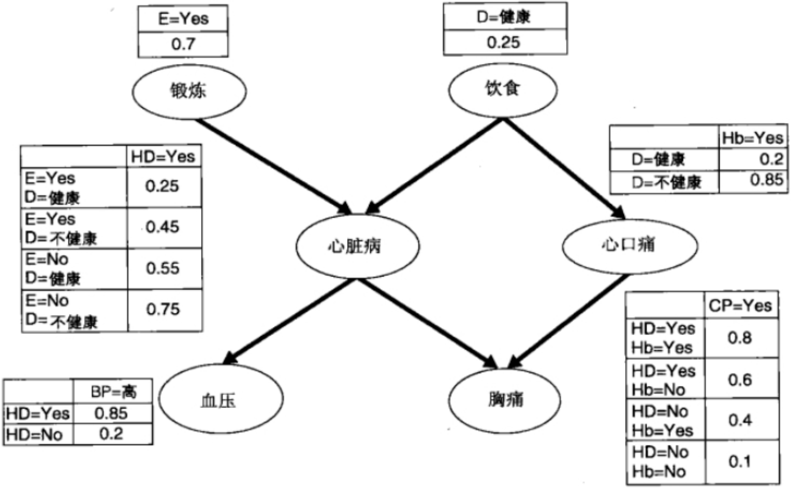
$$P(X_1, \ldots, X_n) = \Pi_{i=1}^{n} P(X_i \mid \pi(X_i))$$

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

$$
\begin{aligned}
&P(\text{心脏病}=\text{No}\mid\text{锻炼}=\text{No},\ \text{饮食}=\text{健康})\\
&=1-P(\text{心脏病}=\text{Yes}\mid\text{锻炼}=\text{No},\ \text{饮食}=\text{健康})\\
&=1-0.55=0.45
\end{aligned}
$$



E=Yes
0.7

D=健康
0.25

	HD=Yes
E=Yes D=健康	0.25
E=Yes D=不健康	0.45
E=No D=健康	0.55
E=No D=不健康	0.75

	Hb=Yes
D=健康	0.2
D=不健康	0.85

	CP=Yes
HD=Yes Hb=Yes	0.8
HD=Yes Hb=No	0.6
HD=No Hb=Yes	0.4
HD=No Hb=No	0.1

	BP=高
HD=Yes	0.85
HD=No	0.2

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 $X_1, X_2, \ldots, X_{i-1}$：
     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 $\pi(X_i)$，是的假设“给定 $\pi(X_i)$，X_i 与 $\mathcal{G}$ 中的其他变量条件独立”合理；
     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

Outline

1. 概率论基础
2. 贝叶斯网基础
3. 图分隔与变量独立
4. 贝叶斯网推理

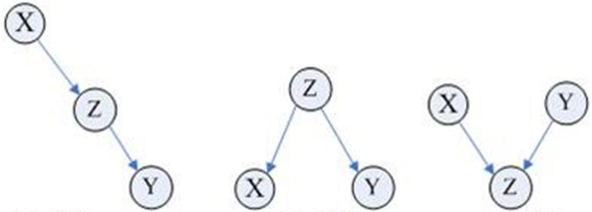
概述

- 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 X 间接相连的三种情况

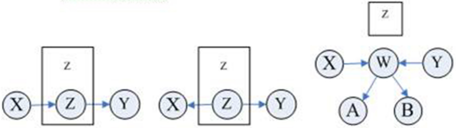


(a) 顺连　(b) 分连　(c) 汇连

- 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 $z\in Z$　(b)分连节点 $z\in Z$　(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

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，所得到的无向图成为 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)

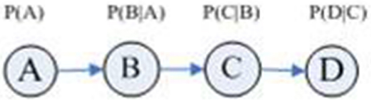
- 贝叶斯网络可以利用变量间的条件独立对联合分布进行分解，降低参数个数
- 推理 (inference) 是通过计算来回答查询的过程
- 贝叶斯网中的推理问题有三大类：
  - 后验概率问题：$P(\mathbf{Q} \mid \mathbf{E} = \mathbf{e})$
  - 最大后验假设问题（Maximum A Posteriori hypothesis, MAP）：

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

  - 最大可能解释问题（Moist Probable Explanation, MPE）

变量消元算法 (Variable Elimination)

利用概率分解降低推理复杂度



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

$$\downarrow$$

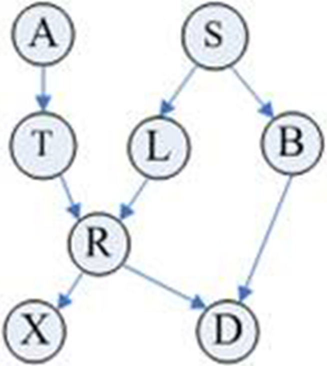
$$P(D) = \sum_{C} P(D \mid C) \sum_{B} P(C \mid B) \sum_{A} P(A)P(B \mid A)$$

- 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

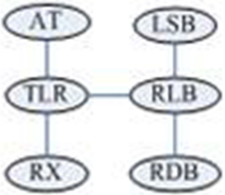
团树传播算法

- 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 $\mathcal{J}$ 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 $\mathcal{J}$ 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. $\mathcal{J}$ 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



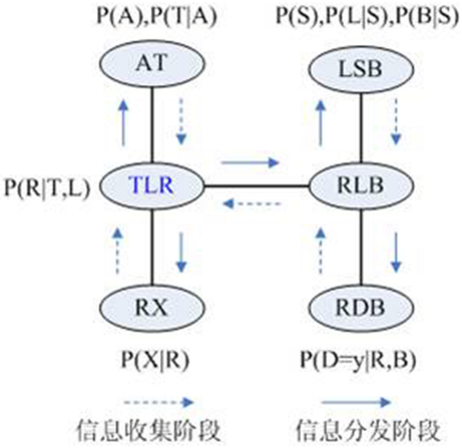
(b) 团树

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



- 变量消元和团树传播算法都是精确推理算法

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

- 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

- 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：

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

- 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

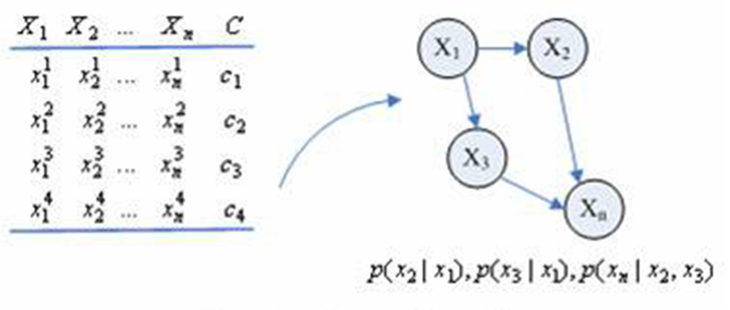


图 5.4 从数据中学习贝叶斯网络

- 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