第5章 不确定性推理与因果推理

南京信息工程大学
计算机学院

应龙

2024年秋季

内容概要

1. 贝叶斯网络

2. 因果推理

Bibliography:
[1] Stuart J. Russel, Peter Norvig, “Artificial Intelligence: A Modern Approach (4th Ed.)”, Pearson, 2020; 中译版“人工智能 现代方法 (第4版)”, 人民邮电出版社, 2022. Ch 13
[2] 周志华, “机器学习”, 清华大学出版社, 2016. Ch 7.5
[3] 吴飞 编著, “人工智能导论：模型与算法”, 高等教育出版社, 2020. Ch 2.4
[4] Judea Pearl, Madelyn Glymour and Nicholas P. “Jewell Causal Inference in Statistics: A Primer” (2016); 中译版“统计因果推断入门”2020, 高等教育出版社.

知识与推理

知识表示

- 谓词逻辑表示
- 产生式表示
- 框架表示
- 语义网络 Knowledge graph
- Large language model

- 确定性推理
- 知识图谱推理
- Fuzzy logic

概率图模型

- Markov logic networks
- Bayesian network 贝叶斯网络
- Topic model Latent Dirichlet allocation

- Chain of Thought

- causal inference 因果推理

马尔科夫逻辑网络

- Domingos 和 Richardson 首次提出了马尔科夫逻辑网络 (Markov logic networks, MLNs)，它是马尔科夫网络与一阶逻辑相结合的一种统计关系学习模型 [Domingos 2004] [Richardson 2006]。
- 一阶逻辑知识库可看作是在一个可能世界的集合上建立一系列硬性规则，即如果一个世界违反了其中的某一条规则，那么这个世界的存在概率即为零。Markov logic networks 的基本思想是让那些硬性规则有所松弛，即当一个世界违反了其中的一条规则，那么这个世界存在的可能性将降低，但并非不可能。一个世界违反的规则越少，那么这个世界存在的可能性就越大。为此，给每个规则都加上了一个特定的权重，它反映了对满足该规则的可能世界的约束力。若一个规则的权重越大，则对于满足和不满足该规则的两个世界而言，它们之间的差异将越大。
- Domingos 和 Richardson 从以下方面论证了马尔科夫逻辑网络作为关系推理学习统一框架的可能性：从概率统计的角度来看，马尔科夫逻辑网络不仅简明地描述了庞大马尔科夫网 (Markov networks, MNs) 所存在的各种关系，还在马尔科夫网络中灵活融入结构化知识；从一阶谓词逻辑的角度来看，马尔科夫逻辑网可在一阶谓词逻辑中添加不确定性 [Richardson 2006]。
- 马尔科夫逻辑网络在自然语言处理、复杂网络、信息抽取、计算机视觉等领域都有重要的应用前景。

内容概要

1. 贝叶斯网络

2. 因果推理

Bibliography:
[1] Stuart J. Russel, Peter Norvig, “Artificial Intelligence: A Modern Approach (4th Ed.)”, Pearson, 2020; 中译版“人工智能 现代方法 (第4版)”, 人民邮电出版社, 2022. Ch 13
[2] 周志华, “机器学习”, 清华大学出版社, 2016. Ch 7.5

贝叶斯网络

贝叶斯网络 (Bayesian Network) 是由美国加州大学的 Judea Pearl (珀尔) 于1985年首先提出的一种模拟人类推理过程中关联关系的不确定性处理模型。它是概率论与图论的结合，其拓扑结构是一个有向无环图 (Directed Acyclic Graph, DAG)。

定义3.17 设$X=\{X_1,X_2,\cdots,X_n\}$是任何随机变量集，其上的贝叶斯网络可定义为$BN=\{B_s,B_p\}$。其中：

① B_s 是贝叶斯网络的结构，即一个定义在X上的有向无环图 (DAG)。并且，其中的每一个节点 X_i 都惟一地对应着X中的一个随机变量，并需要标注定量的概率信息；每条有向边都表示所连接的两个节点之间的条件依赖关系。若存在一条从节点 X_j 到节点 X_i 的有向边，则称 X_j 是 X_i 的父母节点， X_i 是 X_j 的孩子节点。

② B_p 为贝叶斯网络的条件概率集合，$B_p=\{P(X_i|\mathrm{par}(X_i))\}$。其中，$\mathrm{par}(X_i)$ 表示 X_i 的所有父母节点的相应取值，$P(X_i|\mathrm{par}(X_i))$ 是节点 X_i 的一个条件概率分布函数，它描述 X_i 的每个父母节点对 X_i 的影响，即节点 X_i 的条件概率表。

贝叶斯网络中的弧是有方向的，且不能形成回路，因此图有始点和终点。在始点上有一个初始概率，在每条弧所连接的节点上有一个条件概率。

贝叶斯网络

例3.21 假设学生在碰见难题和遇到干扰时会产生焦虑，而焦虑又可导致思维迟缓和情绪波动。请用贝叶斯网络描述这一问题。

解：图3.4是对上述问题的一种贝叶斯网络描述。在该贝叶斯网络中，大写英文字母 A, D, I, C, E 分别表示节点（随机变量）“产生焦虑”、“碰见难题”、“遇到干扰”、“认知迟缓”和“情绪波动”，并将各节点的条件概率表置于相应节点的右侧。

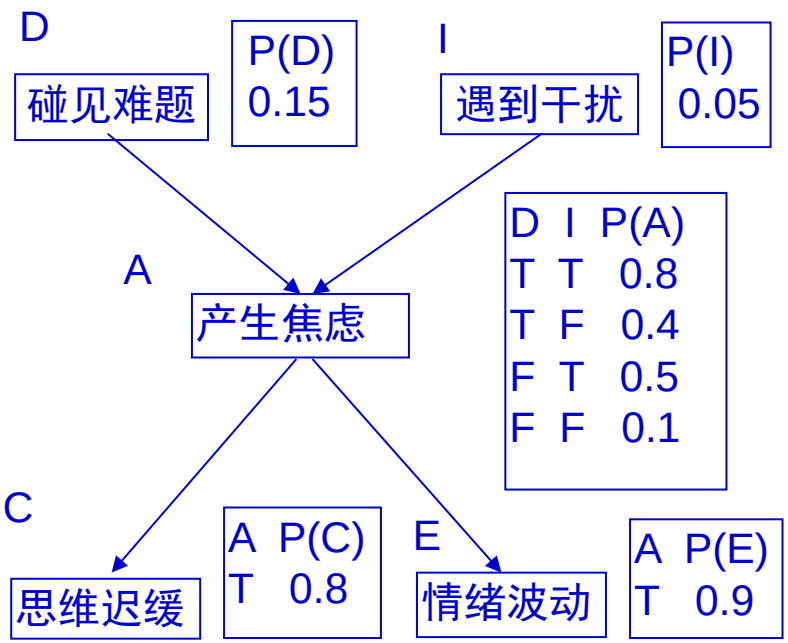


图3.4 关于学习心理的贝叶斯网络

其中的所有随机变量取布尔变量，因此可以分别用小写英文字母 a, d, i, c, e 来表示布尔变量 A, D, I, C, E 取逻辑值为 “True”，用 $\neg a, \neg d, \neg i, \neg c, \neg e$ 来表示布尔变量 A, D, I, C, E 取逻辑值为 “False”。

此外，上述贝叶斯网络中每个节点的概率表就是该节点与其父节点之间的一个局部条件概率分布，由于节点 D 和 I 无父节点，故它们的条件概率表由其先验概率来填充。

贝叶斯网络

- 贝叶斯网络的语义

语义定义了语法如何对应于网络变量的联合分布。

全联合概率分布亦称为全联合概率或联合概率分布，它是概率的一种合取形式，其定义如下

定义3.18 设 $X=\{X_1,X_2,\cdots,X_n\}$ 为任何随机变量集，其全联合概率分布是指当对每个变量取特定值时 $x_i(i=1,2,\cdots,n)$ 时的合取概率，即

$$P(X_1=x_1\wedge X_2=x_2\wedge\cdots\wedge X_n=x_n)$$

其简化表示形式为 $P(x_1,x_2,\cdots,x_n)$。

由全联合概率分布，再重复使用乘法法则

$$P(x_1,x_2,\cdots,x_n)=P(x_n|x_{n-1},x_{n-2},\cdots,x_1)P(x_{n-1},x_{n-2},\cdots,x_1)$$

可以把每个合取概率简化为更小的条件概率和更小的合取式，直至得到如下全联合概率分布表示：

$$\begin{aligned}&P(x_1,x_2,\cdots,x_n)\\&=P(x_n|x_{n-1},x_{n-2},\cdots,x_1)P(x_{n-1}|x_{n-2},x_{n-3},\cdots,x_1)\cdots P(x_2|x_1)P(x_1)\\&=\prod_{i=1}^{n}P(x_i|x_{i-1},x_{i-2},\cdots,x_1)\end{aligned}$$

这个恒等式对任何随机变量都是成立的，该式亦称为链式法则。

贝叶斯网络

根据贝叶斯网络的定义，对孩子节点变量 X_i，其取值 x_i 的条件概率仅依赖于 X_i 的所有父母节点的影响。按照前面的假设，我们用 $\mathrm{par}(X_i)$ 表示 X_i 的所有父母节点的相应取值，$P(x_i|\mathrm{par}(X_i))$ 是节点 X_i 的一个条件概率分布函数，则对随机变量集 X 的所有节点，应有如下联合概率分布：

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\mathrm{par}(X_i))$$

这个公式就是贝叶斯网络的联合概率分布表示。

可见，贝叶斯网络的联合概率分布要比全联合概率分布简单得多。贝叶斯网络能够大大降低计算复杂度的一个重要原因是其具有局部结构化 (locally structured) [也称为 sparse]特征：指每个节点只受到整个节点集中少数别的节点的直接影响，而不受这些节点外的其它节点的直接影响。

贝叶斯网络是一种线性复杂度的方法。即在贝叶斯网络中，一个节点仅受该节点的父节点的直接影响，而不受其它节点的直接影响。例如，在一个包含有n个布尔随机变量的贝叶斯网络中，如果每个随机变量最多只受k个别的随机变量的直接影响，则贝叶斯网络最多可由$2^k \times n$个数据描述。

贝叶斯网络

作为贝叶斯网络简单示例，下面以图3.4所示的贝叶斯网络进行讨论。

例3.22 对例3.21所示的贝叶斯网络，若假设已经产生了焦虑情绪，但实际上并未碰见难题，也未遇到干扰，请计算思维迟缓和情绪波动的概率。

解：令相应变量的取值分别为：

$$a, \neg d, \neg i, c, e$$

其中，无否定符号表示变量取值为True，有否定符号表示变量取值为False，则按贝叶斯网络的联合概率分布表示

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | par(X_i))$$

有：

$$P(c \wedge e \wedge a \wedge \neg d \wedge \neg i)$$
$$= P(c|a)P(e|a)P(a|\neg d \wedge \neg i)P(\neg d)P(\neg i)$$
$$= 0.8 \times 0.9 \times 0.1 \times 0.85 \times 0.95$$
$$= 0.05814$$

即所求的概率为 0.05814。

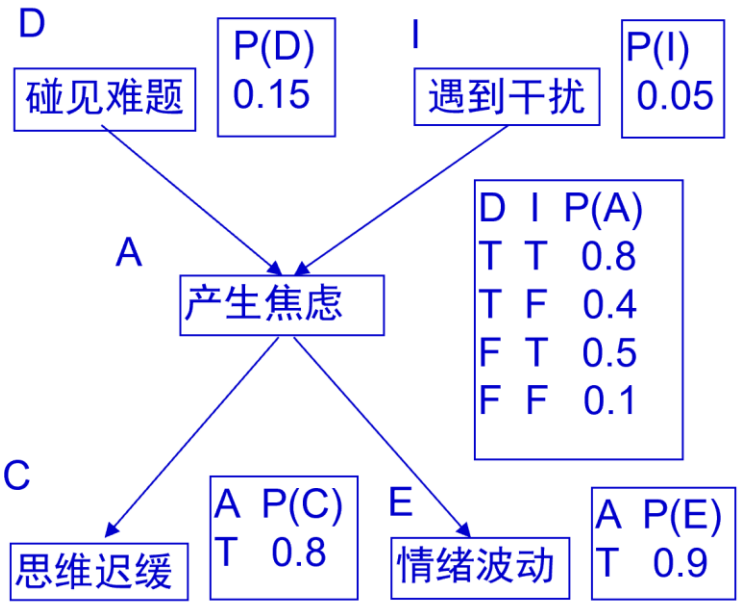


图3.4 关于学习心理的贝叶斯网络

贝叶斯网络

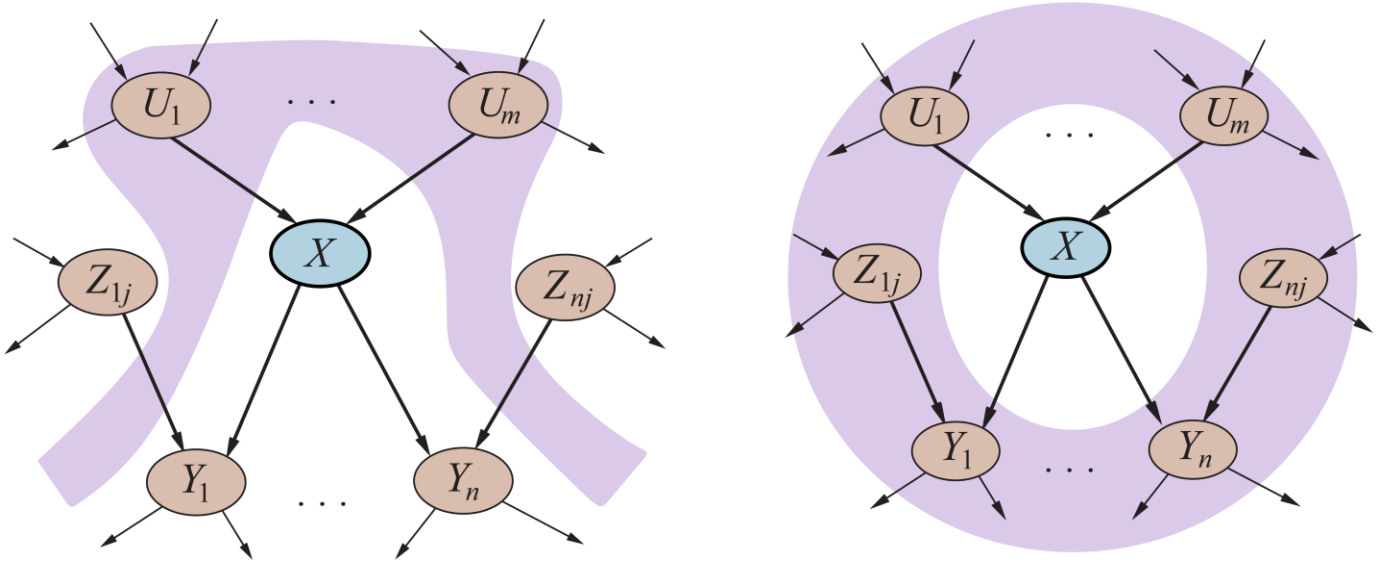
- 贝叶斯网络的条件独立性关系

从贝叶斯网络的局部化特征可以看出，贝叶斯网络能实现简化计算的最根本基础是条件独立性，即一个节点与它的祖先节点之间是条件独立的。下面从网络拓扑结构去定义下面两个等价的条件独立关系的判别准则：

(1) 给定父节点，一个节点与非其后代(descendant)的节点之间是条件独立的。

$$X_v \perp X_{V\backslash \mathrm{de}(v)} | X_{pa(v)}$$

(2) 给定一个节点，该节点与其父节点、子节点和子节点的父节点一起构成了一个马尔科夫毯 (Markov blanket)，则该节点与马尔科夫毯以外的所有节点之间都是条件独立的。

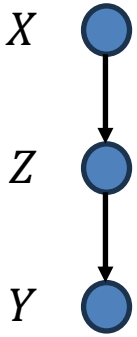


贝叶斯网络

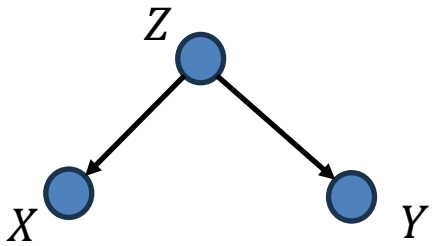
- Bayesian Network 的基本结构：

Junction patterns

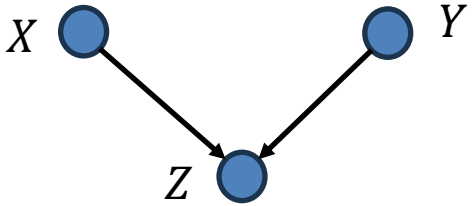
Pattern	Model
Chain	$X \rightarrow Z \rightarrow Y$
Fork	$X \leftarrow Z \rightarrow Y$
Collider	$X \rightarrow Z \leftarrow Y$



Chain



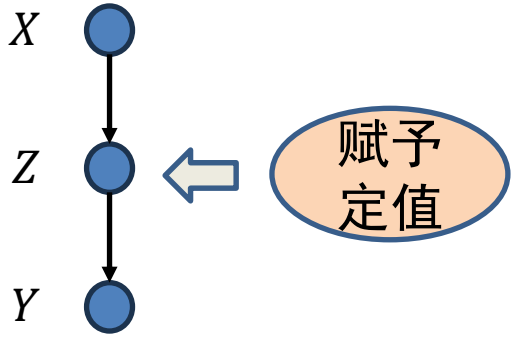
Fork



Collider

贝叶斯网络

- Bayesian Network 的基本结构 链 (chain)：它包含三个节点两条边，其中一条边由第一个节点指向第二个节点，另一条边由第二个节点指向第三个节点。



- 在链 (chain) 结构中，给定Z时，X和Y的联合概率：

$$P(X,Y|Z)=\frac{P(X,Y,Z)}{P(Z)}=\frac{P(X)P(Z|X)P(Y|Z)}{P(Z)}$$
$$=P(X|Z)P(Y|Z)$$

即在链式图$X\to Z\to Y$中，X和Y在给定Z时条件独立。

其中，上式的第一步使用了条件概率的定义，第二步使用了乘积分解规则，最后一步使用了贝叶斯公式$P(X)P(Z|X)=P(Z)P(X|Z)=P(X,Z)$。

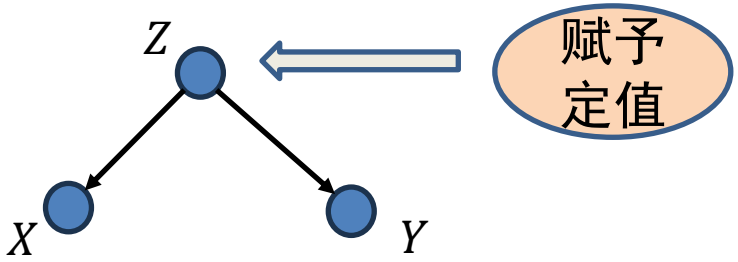
定理2.5 (链中的条件独立性)对于变量X和Y，若X和Y之间只有一条单向的路径，变量Z是截断 (intercept) 该路径的集合中的任一变量，则在给定Z时，X和Y条件独立。

- 不给定Z时，X和Y的联合概率：

$$P(X,Y)=\int P(X)P(Z|X)P(Y|Z)dZ$$
$$P(X)P(Y)=P(X)\int P(Z)P(Y|Z)dZ=P(X)\int P(X)P(Z|X)(Y|Z)dZ=$$
$$P(X)P(X,Y)$$ 因此X和Y相关。

贝叶斯网络

- Bayesian Network 的基本结构 分连 (fork): 它包含三个节点两条边，两条边分别由第一个节点指向第二个节点和第三个节点。



- 在分连 (fork) 结构中，给定Z时，X和Y的联合概率：

$$P(X,Y|Z)=\frac{P(X,Y,Z)}{P(Z)}=\frac{P(Z)P(X|Z)P(Y|Z)}{P(Z)}=P(X|Z)P(Y|Z)$$

即在分连图$X \leftarrow Z \rightarrow Y$中，$X$和$Y$在给定$Z$时条件独立。上式的第一步使用了条件概率的定义，第二步使用了乘积分解规则。

定理 2.6 (分连中的条件独立性)若变量Z是变量X和Y的共同原因，且X到Y只有一条路径，则在给定Z时，X和Y条件独立。

- 不给定Z时，X和Y的联合概率：

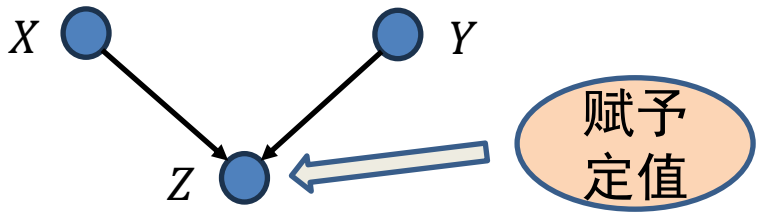
$$P(X,Y)=\int P(X,Y,Z)dZ=\int P(Z)P(X|Z)P(Y|Z)dZ$$

$$P(X)P(Y)=\int P(Z)P(X|Z)dZ\int P(Z_1)P(Y|Z_1)dZ_1=\int P(Z)P(X|Z)P(Y)dZ$$

因此X和Y相关。

贝叶斯网络

- Bayesian Network 的基本结构 汇连 (collider): 它包含三个节点两条边，两条边分别由第一个节点和第二个节点指向第三个节点。



- 不给定Z时，X和Y的联合概率：

$$P(X,Y)=\int P(X)P(Y)P(Z|X,Y)dZ=P(X)P(Y)$$

因此X和Y独立。

- 在汇连结构中，给定Z时，X和Y的联合概率：

$$P(X,Y|Z)=\frac{P(X,Y,Z)}{P(Z)}=\frac{P(X,Y,Z)P(Z)}{P(Z)P(Z)}=\frac{P(X,Y,Z)\sum_X P(X,Z)}{P(Z)P(Z)}$$

$$P(X|Z)P(Y|Z)=\frac{P(X,Z)}{P(Z)}\cdot\frac{P(Y,Z)}{P(Z)}=\frac{P(X,Z)P(Y,Z)}{P(Z)P(Z)}=\frac{P(X,Z)\sum_X P(X,Y,Z)}{P(Z)P(Z)}$$

对任意函数 $P(X,Y,Z)\sum_X P(X,Z)\neq P(X,Z)\sum_X P(X,Y,Z)$

汇连结构 $X\rightarrow Z\leftarrow Y$ 中， X 和 Y 在给定 Z 时条件依赖。

定理 2.7 (汇连中的条件独立性)若变量Z是变量X和Y的汇连节点，且X到Y只有一条路径，则X和Y相互独立，但在给定Z或Z的后代时，X和Y是相关的。

因果推理

- *D*-分离(directional separation, *D*-separation) 和 *D*-链接 (directional connected, *D*-connected) 通常用于判断任意两个节点的相关性和独立性。

给定一个结点限定集合（可以为空集），若两个结点存在某关联路径可被该结点集阻塞 (block)，则称这两个结点在该路径上是有向分离的 (D-separation)；若两个结点间存在某条关联路径没有被该结点集合阻塞 (block)，它们之间的状态称为有向连接 (D-connected)，即这两个节点是相关的。若不存在任何路径联通两个结点，则这两个结点相互独立。

定义 2.18 D-分离：路径p被限定集Z阻塞 (block) 当且仅当：
(1) 路径p含有链结构$A \to B \to C$或分连结构$A \leftarrow B \to C$且中间节点B在Z中，或
(2) 路径p含有汇连结构$A \to B \leftarrow C$且汇连节点B及其后代都不在Z中。
若Z阻塞了节点X和节点Y之间的每一条路径，则称给定Z时，X和Y是D-分离，即给定Z时，X和Y条件独立。

因果推理

例 2.25 分析如下贝叶斯网络（因果图）中节点的关系

不妨考虑X和T的关系：

(1)若限定集为∅时，X和T相互独立。因为X和T之间只有一条路径且这条路径，含有一个汇连结构$X \to Z \leftarrow Y$，且Z及其后代都不在限定集∅中，此时X和T是有向分离，即X和T相互独立；

(2)若限定集为$\{W\}$时，X和T是相关的。因为X和T之间含有一个链式结构$Y \to S \to T$ (S不在限定集中)，一个分连结构$Z \leftarrow Y \to S$ (Y不在限定集$\{W\}$中)，一个汇连结构$X \to Z \leftarrow Y$ (Z的后代W在限定集$\{W\}$中)，根据D-分离的定义，这些结构都无法阻塞X和T之间的路径，因此X和T是有向连接的，即X和T是相关的；

(3)若限定集为$\{W, Y\}$时，X和T条件独立。因为X和T之间只有一条路径，且这条路径含有一个分连结构$Z \leftarrow Y \to S$，且Y在限定集$\{W, Y\}$中，因此Y阻塞了X和T之间的唯一路径，X和T是有向分离，即限定集为$\{W, Y\}$时，X和T条件独立。

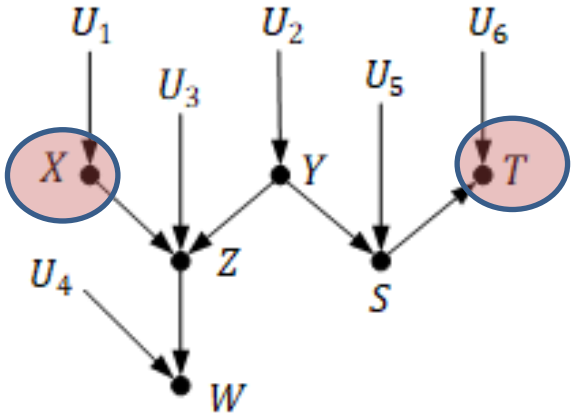


图2.8 包含链、分连和汇连结构的因果图

贝叶斯网络

- 构造贝叶斯网络的方法

依据贝叶斯网络的联合概率分布表示，其构造过程如下：

1. 首先建立不依赖于其它节点的根节点，根节点可以不止一个。
2. 加入受根节点影响的节点，并将这些节点作为根节点的子节点。此时，根节点已成为父节点。
3. 进一步建立依赖于已建节点的子节点。重复这一过程直到叶节点为止。
4. 对每个根节点，给出其先验概率；对每个中间节点和叶节点，给出其条件概率表。

遵循的主要原则：

① 忽略过于微弱的依赖关系

② 随机变量之间的因果关系是最常见、最直观的依赖关系，可以用来指导贝叶斯网络的构建过程

例如，图3.4所示贝叶斯网络的构建过程如下：

① 先建立根节点“碰见难题”和“遇到干扰”；

② 加入受根节点影响节点“产生焦虑”，并将其作为两个根节点的子节点。

③ 进一步加入依赖于已建立节点“产生焦虑”的子节点“思维迟缓”和“情绪波动”。由于这两个新建节点已为叶节点，故节点构建过程终止。

④ 对每个根节点，给出其先验概率；对每个中间节点和叶节点，给出其条件概率表。

贝叶斯网络

- 贝叶斯网络推理

贝叶斯网络推理是指利用贝叶斯网络模型进行计算的过程，其基本任务就是要在给定一组证据变量观察值的情况下，利用贝叶斯网络计算一组查询变量的后验概率分布。

假设，用 X 表示某查询变量，E 表示证据变量集 $\{E_1, E_2, \cdots, E_n\}$，$s$ 表示一个观察到的特定事件，Y 表示一个非证据变量（亦称隐含变量）集 $\{y_1, y_2, \cdots, y_m\}$，则全部变量的集合 $V = \{X\} \cup E \cup Y$，其推理就是要查询后验概率 $P(X|s)$。

例如，在例3.21所示的贝叶斯网络中，若已观察到的一个事件是“思维迟缓”和“情绪波动”，现在要询问的是“遇到干扰”的概率是多少。这是个贝叶斯网络推理问题，其查询变量为 I，观察到的特定事件 $s = \{c, e\}$，即求 $P(I|c, e)$。

贝叶斯网络

- 贝叶斯网络精确推理

贝叶斯网络精确推理的主要方法包括基于变量消元的算法 (Variable Elimination) 、基于团树传播的算法 (Clique tree Propagation) 和 Belief Propagation等。

其中，最基本的方法是基于枚举的算法，使用全联合概率分布去推断查询变量的后验概率：

$$P(X|s) = \alpha P(X,s) = \alpha \sum_{Y} P(X,s,Y)$$

其中，各变量的含义如前所述， X 表示查询变量； s 表示一个观察到的特定事件 Y 表示隐含变量集 $\{y_1, y_2, \cdots, y_m\}$； α 是归一化常数，用于保证相对于 X 所有取值的后验概率总和等于1。

为了对贝叶斯网络进行推理，可利用贝叶斯网络的概率分布公式

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|par(X_i))$$

将上式中的 $P(X,s,Y)$ 改写为条件概率乘积的形式。这样，就可通过先对 Y 的各个枚举值求其条件概率乘积，然后再对各条件概率乘积求总和的方式去计算查询变量的条件概率。

贝叶斯网络

例3.23 仍以例3.21所示的贝叶斯网络为例，假设目前观察到的一个事件 $s = \{c, e\}$，求在该事件的前提下碰见难题的概率 $P(D|c,e)$ 是多少？

解：按照精确推理算法，该询问可表示为：

$$P(D|c,e) = \alpha P(D,c,e) = \alpha \sum_I \sum_A P(D,I,A,c,e)$$

其中，α 是归一化常数，D 取 d 和 $\neg d$，应用贝叶斯网络的概率分布公式：

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|par(x_i))$$

先对 D 的不同取值 d 和 $\neg d$ 分别进行处理。

当 D 取值 d 时，有

$$\begin{aligned}
P(d|c,e) &= \alpha \sum_I \sum_A P(d,I,A,c,e) \\
&= \alpha \sum_I \sum_A P(d)P(I)P(A|d,I)P(c|A)P(e|A) \\
&= \alpha P(d) \sum_I P(I) \sum_A P(A|d,I)P(c|A)P(e|A) \\
&= \alpha P(d)[P(i)\big(P(a|d,i)P(c|a)P(e|a) + P(\neg a|d,i)P(c|\neg a)P(e|\neg a)\big) + \\
&\quad P(\neg i)\big(P(a|d,\neg i)P(c|a)P(e|a) + P(\neg a|d,\neg i)P(c|\neg a)P(e|\neg a)\big)] \\
&= \alpha \times 0.15 \times [0.05 \times (0.8 \times 0.8 \times 0.9 + 0.2 \times 0.2 \times 0.1) + 0.95 \times (0.4 \times 0.8 \\
&\quad \times 0.9 + 0.6 \times 0.2 \times 0.1)] \\
&= \alpha \times 0.15 \times 0.314 = \alpha \times 0.047
\end{aligned}$$

贝叶斯网络

当 D 取值 $\neg d$ 时，有

$$P(\neg d|c,e) = \alpha \sum_I \sum_A P(\neg d, I, A, c, e)$$
$$= \alpha \sum_I \sum_A P(\neg d)P(I)P(A|\neg d, I)P(c|A)P(e|A)$$
$$= \alpha P(d) \sum_I P(I) \sum_A P(A|d, I)P(c|A)P(e|A)$$
$$= \alpha P(\neg d)[P(i)(P(a|\neg d, i)P(c|a)P(e|a) + P(\neg a|\neg d, i)P(c|\neg a)P(e|\neg a)) +$$
$$P(\neg i)(P(a|\neg d, \neg i)P(c|a)P(e|a) + P(\neg a|\neg d, \neg i)P(c|\neg a)P(e|\neg a))]$$
$$= \alpha \times 0.85 \times [0.05 \times (0.5 \times 0.8 \times 0.9 + 0.5 \times 0.2 \times 0.1) + 0.95 \times (0.1 \times 0.8 \times 0.9 + 0.9 \times 0.2 \times 0.1)]$$
$$= \alpha \times 0.85 \times [0.05 \times 0.37 + 0.95 \times 0.09] = \alpha \times 0.85 \times 0.104$$
$$= \alpha \times 0.088$$

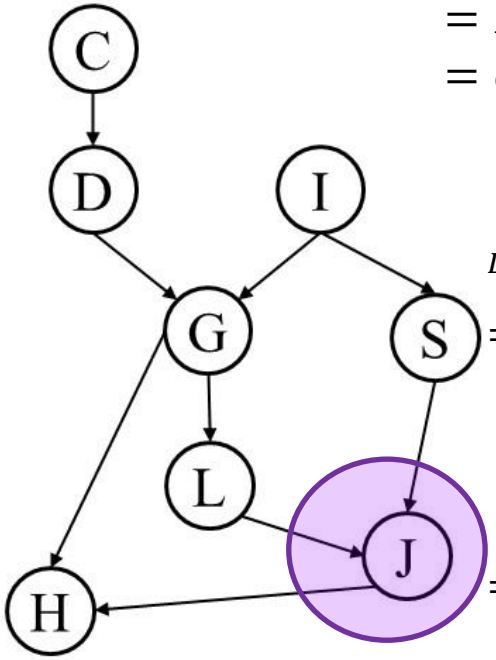
取 $\alpha = 1/(0.047 + 0.088) = 1/0.135$。因此有

$$P(D|c,e) = \alpha(0.047, 0.088) = (0.348, 0.652)$$

即在思维迟缓和情绪波动都发生时，遇到难题的概率是 $P(d|c,e) = 0.348$，不是因为遇到难题的概率是 $P(\neg d|c,e) = 0.652$.

贝叶斯网络

- Variable Elimination



$$P(C,D,I,G,S,L,J,H)$$
$$= P(C)P(D|C)P(I)P(G|I,D)P(S|I)P(L|G)P(J|L,S)P(H|G,J)$$
$$= \phi_C(C)\phi_D(D,C)\phi_I(I)\phi_G(G,I,D)\phi_s(S,I)\phi_L(L,G)\phi_J(J,L,s)\phi_H(H,G,J)$$

$$\sum_{L,S,G,H,I,D,C} \phi_c(C)\phi_D(D,C)\phi_I(I)\phi_G(G,I,D)\phi_S(S,I)\phi_L(L,G)\phi_J(J,L,S)\phi_H(H,G,J)$$

$$= \sum_{L,S,G,H,I,D} \phi_I(I)\phi_G(G,I,D)\phi_S(S,I)\phi_L(L,G)\phi_J(J,L,S)\phi_H(H,G,J) \sum_{C} \phi_c(C)\phi_D(D,C)$$

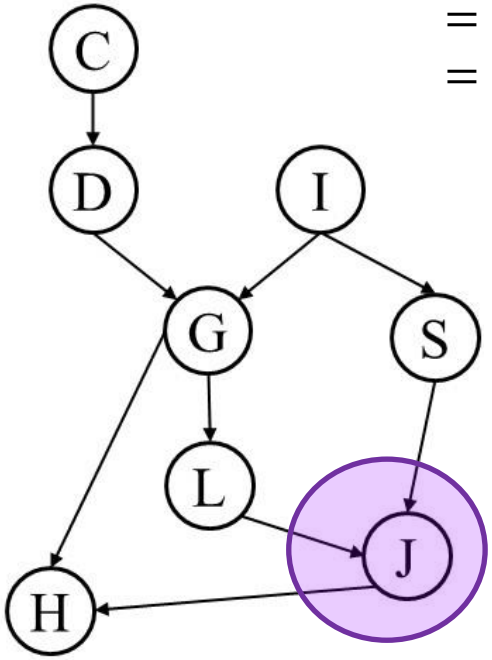
$$= \sum_{L,S,G,H,I,D} \phi_I(I)\phi_G(G,I,D)\phi_S(S,I)\phi_L(L,G)\phi_J(J,L,S)\phi_H(H,G,J)\tau_1(D)$$

$$= \sum_{L,S,G,H,I} \phi_I(I)\phi_S(S,I)\phi_L(L,G)\phi_J(J,L,S)\phi_H(H,G,J) \sum_{D} \phi_G(G,I,D)\tau_1(D)$$

$$= \sum_{L,S,G,H,I} \phi_I(I)\phi_S(S,I)\phi_L(L,G)\phi_J(J,L,S)\phi_H(H,G,J)\tau_2(G,I)$$

贝叶斯网络

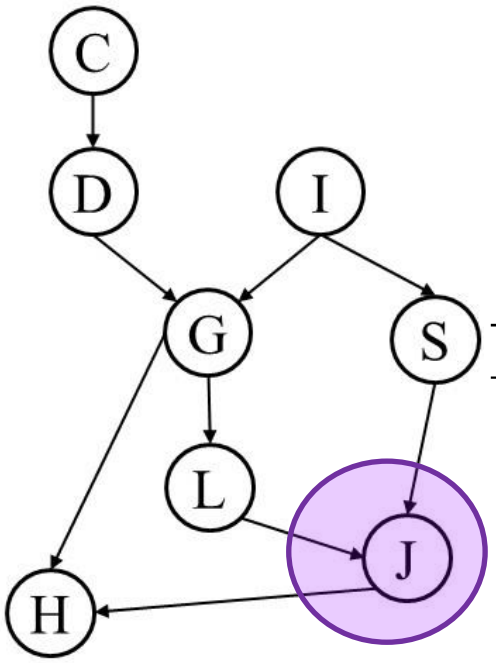
- Variable Elimination



$$P(C,D,I,G,S,L,J,H)$$
$$= P(C)P(D|C)P(I)P(G|I,D)P(S|I)P(L|G)P(J|L,S)P(H|G,J)$$
$$= \phi_C(C)\phi_D(D,C)\phi_I(I)\phi_G(G,I,D)\phi_S(S,I)\phi_L(L,G)\phi_J(J,L,s)\phi_H(H,G,J)$$
$$= \sum_{L,S,G,H} \phi_L(L,G)\phi_J(J,L,S)\phi_H(H,G,J) \sum_I \phi_I(I)\phi_S(S,I)\tau_2(G,I)$$
$$= \sum_{L,S,G,H} \phi_L(L,G)\phi_J(J,L,S)\phi_H(H,G,J)\tau_3(G,S)$$
$$= \sum_{L,S,G} \phi_L(L,G)\phi_J(J,L,S)\tau_3(G,S) \sum_H \phi_H(H,G,J)$$
$$= \sum_{L,S,G} \phi_L(L,G)\phi_J(J,L,S)\tau_3(G,S)\tau_4(G,J)$$
$$= \sum_{L,S} \phi_J(J,L,S) \sum_G \phi_L(L,G)\,\tau_3(G,S)\tau_4(G,J)$$
$$= \sum_{L,S} \phi_J(J,L,S)\tau_5(L,S,J)$$
$$= \sum_L \tau_6(L,J) = \tau_7(J)$$

贝叶斯网络

- Variable Elimination



步骤	消元变量	涉及因子	涉及变量	新因子
1	C	$\phi_C(C), \phi_D(D,C)$	C, D	$\tau_1(D)$
2	D	$\phi_G(G,I,D), \tau_1(D)$	G, I, D	$\tau_2(G,I)$
3	I	$\phi_I(I), \phi_S(S,I), \tau_2(G,I)$	G, S, I	$\tau_3(G,S)$
4	H	$\phi_H(H,G,J)$	H, G, J	$\tau_4(G,J)$
5	G	$\tau_4(G,J), \tau_3(G,S), \phi_L(L,G)$	G, J, L, S	$\tau_5(J,L,S)$
6	S	$\tau_5(J,L,S), \phi_J(J,L,S)$	J, L, S	$\tau_6(J,L)$
7	L	$\tau_6(J,L)$	J, L	$\tau_7(J)$

步骤	消元变量	涉及因子	涉及变量	新因子
1	G	$\phi_G(G,I,D), \phi_L(L,G), \phi_H(H,G,J)$	G, I, D, L, J, H	$\tau_1(I,D,L,J,H)$
2	I	$\phi_I(I), \phi_S(S,I), \tau_1(I,D,L,S,J,H)$	S, I, D, L, J, H	$\tau_2(D,L,S,J,H)$
3	S	$\phi_J(J,L,S), \tau_2(D,L,S,J,H)$	D, L, S, J, H	$\tau_3(D,L,J,H)$
4	L	$\tau_3(D,L,J,H)$	D, L, J, H	$\tau_4(D,J,H)$
5	H	$\tau_4(D,J,H)$	D, J, H	$\tau_5(D,J)$
6	C	$\tau_5(D,J), \phi_C(C), \phi_D(D,C)$	D, J, C	$\tau_6(D,J)$
7	D	$\tau_6(D,J)$	D, J	$\tau_7(J)$

- 变量消元法把全局概率推理计算，转化为局部变量的因子乘积和因子求和运算；变量消元是概率图模型精确推断的中心思想。
- 不同的消元顺序影响算法时间复杂度。寻找最优的消元顺序是一个NP难的问题。可以通过启发式方法选择消元顺序：从具有最小邻节点的节点进行变量消元。
- 主要的缺点是每求取一个节点都要从头算一遍，没有对中间计算过程进行存储，重复计算。

贝叶斯网络

- Belief propagation

Factor Graph

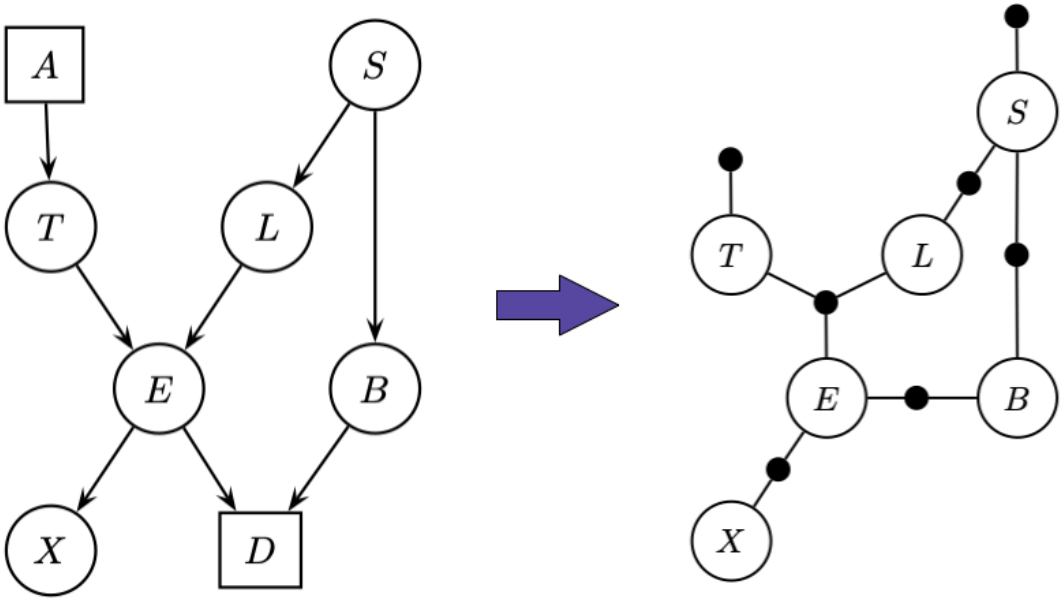
Variables: $\mathcal{X} = \{X_1, \ldots, X_i, \ldots, X_n\}$

Factors: $\Psi = \{\psi_\alpha, \psi_\beta, \psi_\gamma, \ldots\}$, where $\alpha, \beta, \gamma, \ldots \subseteq \{1, \ldots, n\}$

The joint mass function:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{\alpha} \psi_\alpha(\mathbf{x}_\alpha)$$

Where $\mathbf{x}_\alpha$ is the vector of neighboring variable nodes to the factor node .

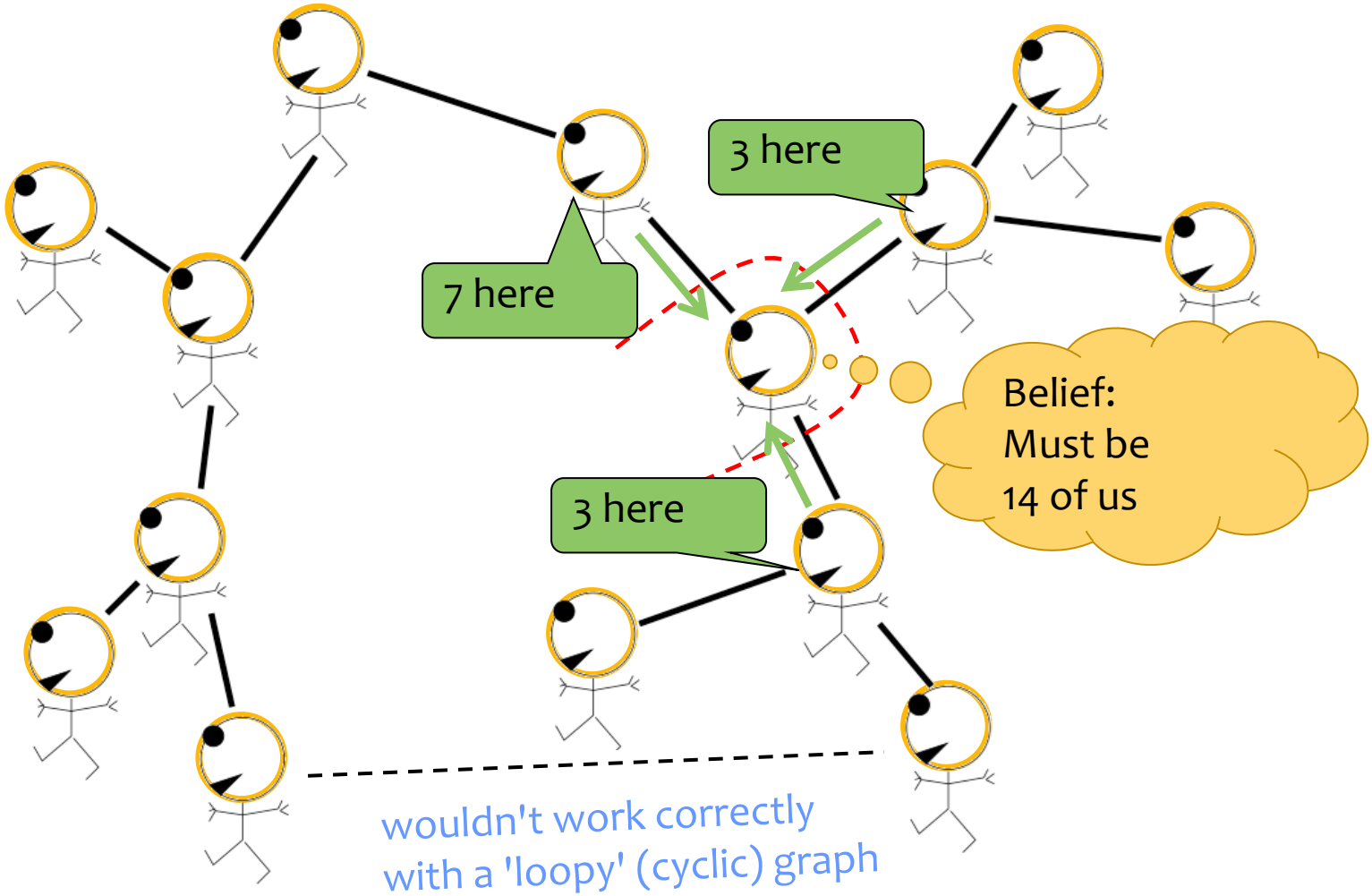


贝叶斯网络

- Belief propagation

Great Ideas in ML:

Each soldier receives reports from all branches of tree

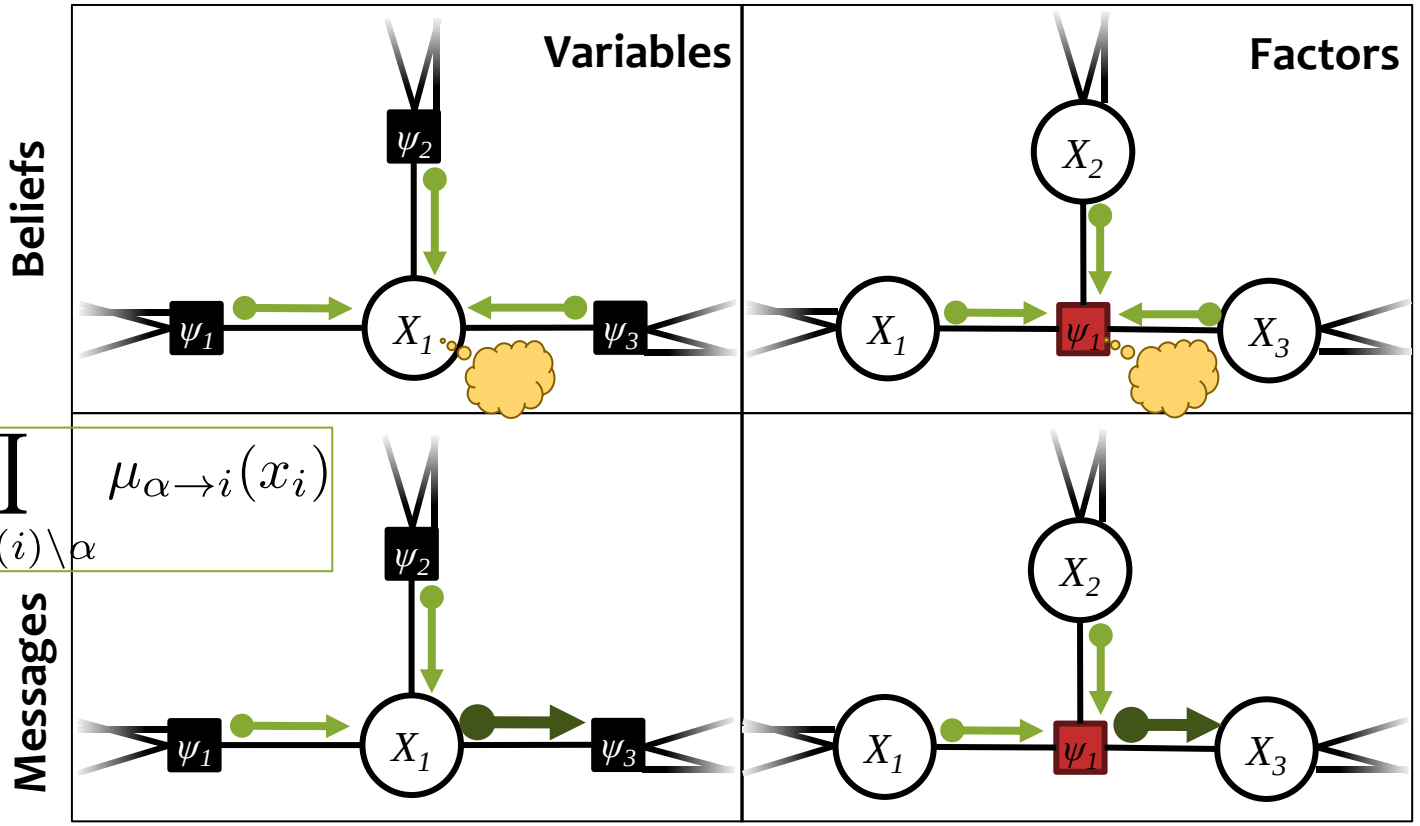


贝叶斯网络

- Belief propagation

$$b_i(x_i) = \prod_{\alpha \in \mathcal{N}(i)} \mu_{\alpha \to i}(x_i)$$

$$b_\alpha(\boldsymbol{x}_\alpha) = \psi_\alpha(\boldsymbol{x}_\alpha) \prod_{i \in \mathcal{N}(\alpha)} \mu_{i \to \alpha}(\boldsymbol{x}_\alpha[i])$$



$$\mu_{i \to \alpha}(x_i) = \prod_{\alpha \in \mathcal{N}(i) \setminus \alpha} \mu_{\alpha \to i}(x_i)$$

$$\mu_{\alpha \to i}(x_i) = \sum_{\boldsymbol{x}_\alpha : \boldsymbol{x}_\alpha[i] = x_i} \psi_\alpha(\boldsymbol{x}_\alpha) \prod_{j \in \mathcal{N}(\alpha) \setminus i} \mu_{j \to \alpha}(\boldsymbol{x}_\alpha[i])$$

贝叶斯网络

- Belief propagation

Input: a factor graph with no cycles

Output: exact marginals for each variable and factor

Algorithm:

1. Initialize the messages to the uniform distribution.

$$\mu_{i\to\alpha}(x_i) = 1 \qquad \mu_{\alpha\to i}(x_i) = 1$$

1. Choose a root node.
2. Send messages from the **leaves** to the **root**.
   Send messages from the **root** to the **leaves**.

$$\mu_{i\to\alpha}(x_i) = \prod_{\alpha\in\mathcal{N}(i)\setminus\alpha} \mu_{\alpha\to i}(x_i) \qquad \mu_{\alpha\to i}(x_i) = \sum_{\boldsymbol{x}_{\boldsymbol{\alpha}}:\boldsymbol{x}_{\boldsymbol{\alpha}}[i]=x_i} \psi_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}}) \prod_{j\in\mathcal{N}(\alpha)\setminus i} \mu_{j\to\alpha}(\boldsymbol{x}_{\boldsymbol{\alpha}}[i])$$

1. Compute the beliefs (unnormalized marginals).

$$b_i(x_i) = \prod_{\alpha\in\mathcal{N}(i)} \mu_{\alpha\to i}(x_i) \qquad b_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}}) = \psi_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}}) \prod_{i\in\mathcal{N}(\alpha)} \mu_{i\to\alpha}(\boldsymbol{x}_{\boldsymbol{\alpha}}[i])$$

2. Normalize beliefs and return the **exact** marginals.

$$p_i(x_i) \propto b_i(x_i) \qquad p_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}}) \propto b_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}})$$

贝叶斯网络

- Belief propagation

Input: a factor graph with cycles

Output: approximate marginals for each variable and factor

Algorithm:

1. Initialize the messages to the uniform distribution.

$$\mu_{i\to\alpha}(x_i) = 1 \qquad \mu_{\alpha\to i}(x_i) = 1$$

1. Send messages until convergence.
   Normalize them when they grow too large.

$$\mu_{i\to\alpha}(x_i) = \prod_{\alpha\in\mathcal{N}(i)\setminus\alpha} \mu_{\alpha\to i}(x_i) \qquad \mu_{\alpha\to i}(x_i) = \sum_{\boldsymbol{x}_{\boldsymbol{\alpha}}:\boldsymbol{x}_{\boldsymbol{\alpha}}[i]=x_i} \psi_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}}) \prod_{j\in\mathcal{N}(\alpha)\setminus i} \mu_{j\to\alpha}(\boldsymbol{x}_{\boldsymbol{\alpha}}[i])$$

1. Compute the beliefs (unnormalized marginals).

$$b_i(x_i) = \prod_{\alpha\in\mathcal{N}(i)} \mu_{\alpha\to i}(x_i) \qquad b_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}}) = \psi_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}}) \prod_{i\in\mathcal{N}(\alpha)} \mu_{i\to\alpha}(\boldsymbol{x}_{\boldsymbol{\alpha}}[i])$$

2. Normalize beliefs and return the **approximate** marginals.

$$p_i(x_i) \propto b_i(x_i) \qquad p_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}}) \propto b_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}})$$

贝叶斯网络

贝叶斯网络推理的一般步骤：①首先确定各相邻节点之间的初始条件概率分布；②然后对各证据节点取值；③接着选择适当推理算法对各节点的条件概率分布进行更新；④最终得到推理结果。(迭代算法)

类型：贝叶斯网络推理的算法可根据对查询变量后验概率计算的精确度，分为精确推理和近似推理两大类。

- 精确推理是一种可以精确地计算查询变量的后验概率的一种推理方法。由于计算复杂度，它的一个重要前提是要求贝叶斯网络具有单连通特性，即任意两个节点之间至多只有一条无向路径连接。

  但现实世界中复杂问题的贝叶斯网络往往不具有单连通性，而是多连通的。例如，在例3.21所示的贝叶斯网络中，若节点“遇到干扰”到节点“思维迟缓”之间存在有向边，则这两个节点之间就有两条无向路径相连。

  多连通贝叶斯网络的复杂度是指数级的。因此，精确推理算法仅适用于规模较小的贝叶斯网络推理。而对复杂的多连通贝叶斯网络，则应该采用近似推理方法。
- 近似推理算法是在不影响推理正确性的前提下，通过适当降低推理精确度来提高推理效率的一类方法。常用的近似推理算法主要有马尔科夫链蒙特卡洛(Markov Chain Monte Carlo, MCMC) 算法等。

贝叶斯网络

- 贝叶斯网络近似推理

马尔科夫链蒙特卡洛 (MCMC) 算法是目前使用较广的一种贝叶斯网络似推理方法。它通过对前一个世界状态作随机改变来生成下一个问题状态，通过对某个隐变量进行随机采样来实现对随机变量的改变。

例3.24 我们知道，学习情绪会影响学习效果。假设有一个知识点，考虑学生在愉快学习状态下对该知识点的识记、理解、运用的情况，得到了如图3.5所示的多连通贝叶斯网络。如果目前观察到一个学生不但记住了该知识，并且还可以运用该知识，询问这位学生是否理解了该知识。

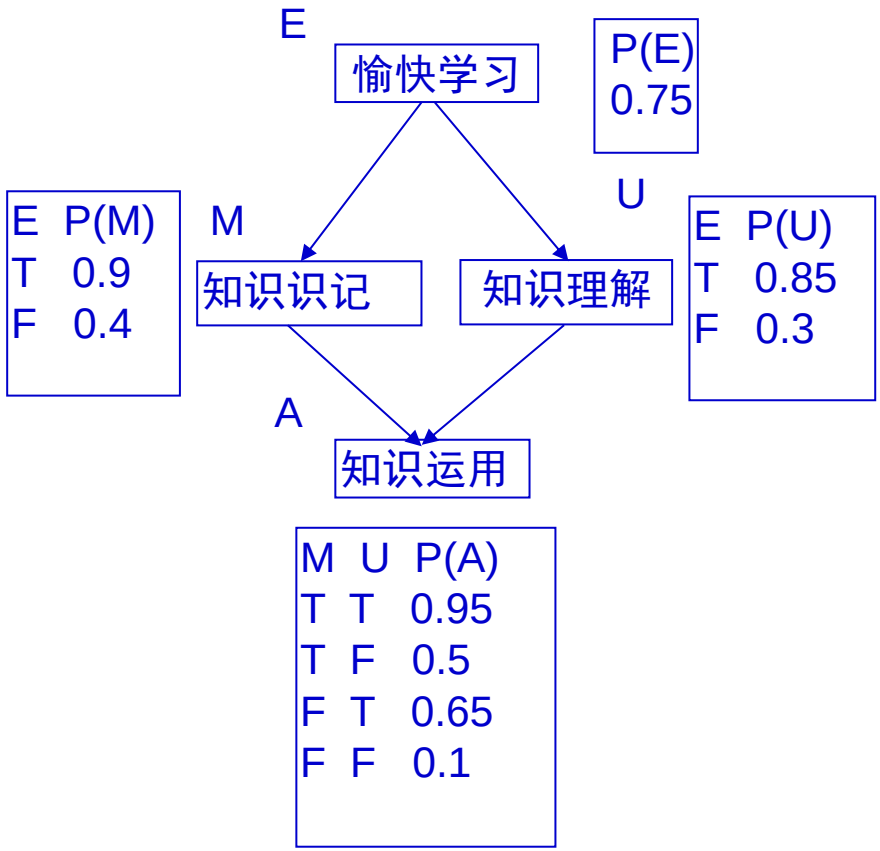


图3.5 关于愉快学习的贝叶斯网络

贝叶斯网络

解：为解决这一问题，令 E, M, U, A 分别表示布尔变量节点“愉快学习”、“知识识记”、“知识理解”和“知识运用”，e, m, u, a 分别表示这些变量取值为“True”，各节点边上的表格为相应节点的条件概率表。

本例的询问句为 $P(U|m, a)$。应用MCMC算法的推理步骤如下：

1. 将“知识识记”节点 M 和“知识运用”节点 A 作为证据变量，并保持它们的观察值不变；
2. 将“愉快学习”节点 E 和“知识理解”节点 U 作为隐变量，并进行随机初始化。假设，取值分别为 e 和 $\neg u$，问题的初始状态为 $\{e, m, \neg u, a\}$；
3. 反复执行如下步骤：
   - 对隐变量 E 进行采样，由于 E 的马尔科夫毯（其父节点、子节点和子节点的父节点）仅包含节点 M 和 U，可以按照变量 M 和 U 的当前值进行采样，若采样得到 $\neg e$，则生成下一状态 $\{\neg e, m, \neg u, a\}$；
   - 对隐变量 U 进行采样，由于 U 的马尔科夫毯包含节点 E, M, A，可以按照变量 E, M, A 的当前值进行采样，若采样得到 u，则生成下一状态 $\{\neg e, m, u, a\}$。

这一反复执行过程中生成的每个状态都作为一个样本，用于估计“愉快学习”的概率的近似值。只要生成的状态足够多，就可得到查询的近似值。

贝叶斯网络

在上述采样过程中，每次采样都需要两步。以对隐变量 E 的采样为例，每次采样步骤如下：

- 第一步，先依据该隐变量的马尔科夫覆盖所包含的变量的当前值，计算该状态转移概率 p ；
- 第二步，确定状态是否需要改变。其基本方法是，生成一个随机数 $r \in [0,1]$ ，将其与第一步得到的转移概率 p 进行比较，若 $r < p$，则 E 取 $\neg e$，转移到下一状态；否则，还处在原状态不变。

例如，对图3.5所给出的问题，在初始状态下，对随机变量 E 进行采样：

第一步可根据 $P(E|m,\neg u)$ 去计算转移到下一状态 $\{\neg e, m, \neg u, a\}$ 的概率。即

$P(e|m,\neg u) = P(e,m,\neg u)/\sum_E P(E,m,\neg u)$

$= P(e)P(m|e)P(\neg u|e)/[P(e)P(m|e)P(\neg u|e) + P(\neg e)P(m|\neg e)P(\neg u|\neg e)]$

$= (0.75 \times 0.9 \times 0.3)/[0.75\times0.9\times0.3+0.25\times0.4\times0.3]$

$= 0.2025/0.2325 = 0.8710$

第二步，假设产生的随机数 $r = 0.46$，有 $0.46 < 0.871$，则 E 取 $\neg e$，转移到下一状态 $\{\neg e, m, \neg u, a\}$。

上述基于转移概率的采样方式亦称为吉布斯 (Gibbs) 采样。

贝叶斯网络

输入: 贝叶斯网 $B = \langle G, \Theta \rangle$;
采样次数 T;
证据变量 $\mathbf{E}$ 及其取值 $\mathbf{e}$;
待查询变量 $\mathbf{Q}$ 及其取值 $\mathbf{q}$.

过程:

$n_q = 0$
$\mathbf{q}^0 =$ 对 $\mathbf{Q}$ 随机赋初值
for $t = 1, 2, \ldots, T$ **do**
  for $Q_i \in \mathbf{Q}$ **do**
    $\mathbf{Z} = \mathbf{E} \cup \mathbf{Q} \setminus \{Q_i\}$;
    $\mathbf{z} = \mathbf{e} \cup \mathbf{q}^{t-1} \setminus \{q_i^{t-1}\}$;
    根据 B 计算分布 $P_B(Q_i \mid \mathbf{Z} = \mathbf{z})$;
    $q_i^t =$ 根据 $P_B(Q_i \mid \mathbf{Z} = \mathbf{z})$ 采样所获 Q_i 取值;
    $\mathbf{q}^t =$ 将 $\mathbf{q}^{t-1}$ 中的 q_i^{t-1} 用 q_i^t 替换
  end for
  if $\mathbf{q}^t = \mathbf{q}$ **then**
    $n_q = n_q + 1$
  end if
end for

输出: $P(\mathbf{Q} = \mathbf{q} \mid \mathbf{E} = \mathbf{e}) \simeq \frac{n_q}{T}$

吉布斯采样是在贝叶斯网所有变量的联合状态空间与证据 $E = e$ 一致的子空间中进行随机游走 (random walk). 每一步仅依赖于前一步的状态，这是一个马尔可夫链 (Markov chain).

在一定条件下，无论从什么初始状态开始，马尔可夫链第t步的状态分布在 $t \to \infty$ 时必收敛于一个平稳分布 (stationary distribution); 对于吉布斯采样来说，这个分布恰好是 $P(\boldsymbol{Q}|\boldsymbol{E} = \boldsymbol{e})$

贝叶斯网络

- 学习

若网络结构已知，即变量间的依赖关系已知，则贝叶斯网的学习过程相对简单，只需通过对训练样本“计数”，估计出每个结点的条件概率即可。但在现实应用中往往并不知晓网络结构。首要任务：根据训练集找出结构最“恰当”的贝叶斯网。

- Let: $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ denote a set of n random variables.
- $\mathcal{D} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\}$ be a dataset of m observations, where each $\mathbf{x}^{(j)} = \{x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}\}$ is a vector of observed values for $\mathcal{X}$.
- Aim to find a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where: $\mathcal{V} = \{X_1, X_2, \ldots, X_n\}$ is the set of vertices, each corresponding to a variable in $\mathcal{X}$. $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of directed edges, where an edge $(X_i \to X_j)$ implies that X_i is a direct parent of X_j. $\mathcal{G}$ encodes conditional dependencies among variables in $\mathcal{X}$ such that the joint probability distribution $P(\mathcal{X})$ factorizes as: $P(\mathcal{X}) = \prod_{i=1}^{n} P(X_i | Pa_{\mathcal{G}}(X_i))$, where $Pa_{\mathcal{G}}(X_i)$ denotes the set of parent variables of X_i in $\mathcal{G}$.

- Constraints: ① DAG Constraint: $\mathcal{G}$ must be a Directed Acyclic Graph (DAG), i.e., it must have no directed cycles. ② Dependency Constraints (for constraint-based methods): The edges in $\mathcal{G}$ should respect the (in)dependencies observed in the data. This is generally achieved through conditional independence tests that dictate edge presence or absence.
- The objective is to identify a DAG $\mathcal{G}$ that best represents the dependencies in the data $\mathcal{D}$. This is typically achieved by maximizing a score function $\mathcal{S}(\mathcal{G}|\mathcal{D})$ that balances model fit and complexity, such as: $\mathcal{G}^* = \text{argmax}_{\mathcal{G}}\, \mathcal{S}(\mathcal{G}|\mathcal{D})$.

贝叶斯网络

- 学习

Junction patterns

Pattern	Model
Chain	$X \rightarrow Y \rightarrow Z$
Fork	$X \leftarrow Y \rightarrow Z$
Collider	$X \rightarrow Y \leftarrow Z$

The first 2 represent the same dependencies (X and Z are independent given Y) and are, therefore, indistinguishable. The collider, can be uniquely identified, since X and Z are marginally independent and all other pairs are dependent.

Thus, while the skeletons (the graphs stripped of arrows) of these three triplets are identical, the directionality of the arrows is partially identifiable. The same distinction applies when X and Z have common parents, except that one must first condition on those parents.

Algorithms have been developed to systematically determine the skeleton of the underlying graph and, then, orient all arrows whose directionality is dictated by the conditional independences observed.

贝叶斯网络

- 学习

基于约束的方法 (Constraint-Based Methods) 主要依赖于条件独立测试来确定变量之间的关系。

➢ PC (Peter-Clark) Algorithm:

- Skeleton Construction: The algorithm begins by assuming a fully connected, undirected graph among the variables. For each pair of variables (X, Y), it tests conditional independence given all possible conditioning sets of neighboring nodes. If X and Y are found to be conditionally independent given any set, the edge between them is removed.
- Edge Orientation: After constructing the skeleton, the algorithm attempts to orient the edges to satisfy certain constraints and avoid cycles. The orientation is often based on v-structures (triplets $X \rightarrow Z \leftarrow Y$) that are identifiable through independence tests. Additional rules, called orientation rules, are used to direct other edges in the graph while preventing cycles.

- Common Conditional Independence Tests
  ① For Discrete Data: The Chi-Square Test; ② For Continuous Data: Partial Correlation Tests; ③ Nonparametric Tests; ④ Mutual Information Tests.

贝叶斯网络

- 学习
  - The Fast Causal Inference (FCI) Algorithm: aims to provide a causal graph known as a Partial Ancestral Graph (PAG), which reveals conditional dependencies while accounting for ambiguity in causal directions due to hidden variables.

    1）**Skeleton Discovery Phase**: determines which variables are (conditionally) dependent and identifies pairs of variables that should remain connected in the final graph；

    2）**Edge Orientation Phase**: uses the skeleton and separation sets to orient the edges, ensuring that the resulting graph represents causal directions accurately even in the presence of hidden variables.

The resulting graph is a **Partial Ancestral Graph (PAG)**, where edges can be oriented in various ways to indicate conditional dependencies and causal directions:

- **Directed Edges** ($X \rightarrow Y$): Represent clear causal relationships from X to Y.
- **Bidirectional Edges** ($X \leftrightarrow Y$): Indicate a possible hidden confounder influencing both X and Y.
- **Undirected Edges** ($X - Y$): Reflect ambiguities in directionality due to latent variables.
- **Partially Directed Edges** ($X \circ\!\!-\!\!\circ Y$): Represent relationships with some level of causal uncertainty.

✓ Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence, 172(16-17), 1873-1896.

贝叶斯网络

- 学习

定义一个评分函数 (score function) 来评估贝叶斯网与训练数据的契合程度。显然，评分函数引入了关于希望获得什么样的贝叶斯网的归纳偏好。

常用评分函数通常基于信息论准则，此类准则将学习问题看作一个数据压缩任务，学习的目标是找到一个能以最短编码长度描述训练数据的模型，此时编码的长度包括了描述模型自身所需的字节长度和使用该模型描述数据所需的字节长度。“最小描述长度”(Minimal Description Length, MDL) 准则

对贝叶斯网学习而言，模型就是一个贝叶斯网络，同时，每个贝叶斯网描述了一个在训练数据上的概率分布。

贝叶斯网络

- 学习

给定训练集 $D=\{x_1,x_2,\cdots,x_m\}$，贝叶斯网 $B=\langle G,\Theta\rangle$, 在 D 上的评价函数可以写为：

$$s(B|D)=f(\theta)|B|-LL(B|D)$$

描述每个参数 θ (每个变量的条件概率表) 所需的字节数

贝叶斯网络参数的个数

$LL(B|D)=\sum_{i=1}^{m}\log P_B(x_i)$ 是贝叶斯网的对数似然。

若贝叶斯网络 $B=\langle G,\Theta\rangle$ 的网络结构G 固定，则评分函数 $s(B|D)$ 的第一项为常数。此时，最小化 $s(B|D)$ 等价于对参数 Θ 的极大似然估计。参数$\theta_{x_i|\pi_i}$能直接在训练数据 D 上通过经验估计获得，即

$$\theta_{x_i|\pi_i}=\hat{P}_D(x_i|\pi_i)$$

为了最小化评分函数 $s(B|D)$，只需对网络结构进行搜索，而候选结构的最优参数可直接在训练集上计算得到.

贝叶斯网络

- 学习

给定训练集 $D=\{x_1,x_2,\cdots,x_m\}$，贝叶斯网 $B=\langle G,\Theta\rangle$, 在 D 上的评价函数可以写为：

$$s(B|D)=f(\theta)|B|-LL(B|D)$$

描述每个参数 θ (每个变量的条件概率表) 所需的字节数

贝叶斯网络参数的个数

$LL(B|D)=\sum_{i=1}^{m}\log P_B(x_i)$ 是贝叶斯网的对数似然。

从所有可能的网络结构空间搜索最优贝叶斯网结构是一个NP Hard 问题，难以快速求解。有两种常用的策略能在有限时间内求得近似解:

- 贪心法，例如从某个网络结构出发，每次调整一条边（增加、删除或调整方向），直到评分函数值不再降低为止;
- 通过给网络结构施加约束来削减搜索空间，例如将网络结构限定为树形结构等。

内容概要

1. 贝叶斯网络

2. 因果推理

代表学者：Judea Pearl, Donald Rubin

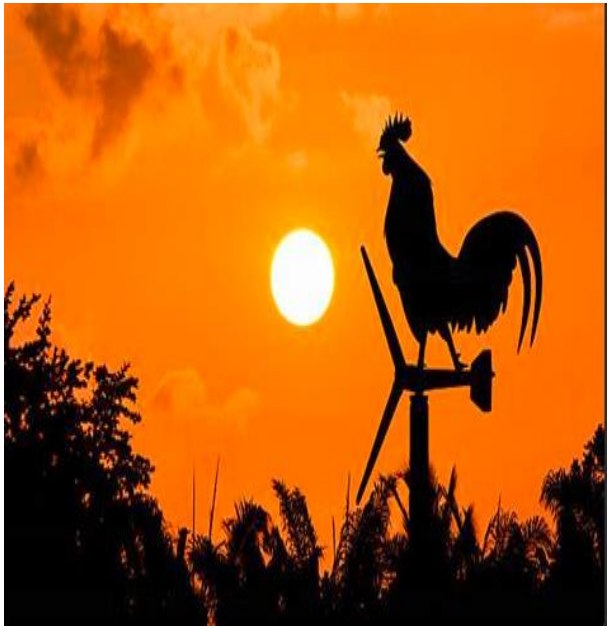
Bibliography:

[3] 吴飞 编著,“人工智能导论：模型与算法”, 高等教育出版社, 2020. Ch 2.4

[1] Stuart J. Russel, Peter Norvig, “Artificial Intelligence: A Modern Approach (4th Ed.)”, Pearson, 2020; 中译版“人工智能 现代方法 (第4版)”, 人民邮电出版社, 2022. Ch 13

[4] Judea Pearl, Madelyn Glymour and Nicholas P. “Jewell Causal Inference in Statistics: A Primer” (2016); 中译版“统计因果推断入门”2020, 高等教育出版社.

因果推理



公鸡打鸣与太阳升起

“力，形之所以奋也”（墨经）

- 计算的可解释：因果是现象加解释，是一种人类文化，即人类在与自然的反复观察、预报、检验中得到的很好考验。
- 一切学习的前提，是我们要假设，这个世界是有规律的

哲学上把现象和现象之间那种“引起和被引起”的关系，叫做因果关系，其中引起某种现象产生的现象叫做原因，被某种现象引起的现象叫做结果。

因果推理 (Causal Inference) 是一种重要的推理手段，是人类智能的重要组成

因果推理

- Simpson's Paradox (辛普森悖论)

	不用药	用药
恢复人数	289	273
总人数	350	350
恢复率(%)	83	78

表2.4.1 某组病人在是否尝试新药以后的恢复情况

	不用药		用药	
	男性	女性	男性	女性
恢复人数	234	55	81	192
总人数	270	80	87	263
恢复率(%)	87	69	93	73

表2.4.2 以性别分组后的某组病人在是否尝试新药以后的恢复情况

- 表2.4.1列出了某组病人在是否尝试新药以后的恢复情况：不用药病人的恢复率高于用药病人的恢复率。
- 然而，当对所有病人按照性别分组后，可得到表2.4.2。当分别比较按照性别分组的两类病人的恢复率时，却发现用药病人的恢复率均高于不用药病人的恢复率，这与表2.4.1的统计结果正好相反。这就是著名的**辛普森悖论 (Simpson's paradox)**，其指出在总体样本上成立的某种关系却在分组样本里恰好相反。

因果推理

- Simpson's Paradox (辛普森悖论)

	不用药	用药
恢复人数	289	273
总人数	350	350
恢复率(%)	83	78

表2.4.1 某组病人在是否尝试新药以后的恢复情况

	不用药		用药	
	男性	女性	男性	女性
恢复人数	234	55	81	192
总人数	270	80	87	263
恢复率(%)	87	69	93	73

表2.4.2 以性别分组后的某组病人在是否尝试新药以后的恢复情况

- 从数学角度而言，上述悖论可写成初等数学不等式

$$\frac{b}{a}<\frac{d}{c},\frac{b'}{a'}<\frac{d'}{c'},\frac{b+b'}{a+a'}>\frac{d+d'}{c+c'}$$

- 辛普森悖论表明，在某些情况下，忽略潜在的"第三个变量"（本例中性别就是用药与否和恢复率之外的第三个变量），可能会改变已有的结论，而我们常常却一无所知。从观测结果中寻找引发结果的原因、考虑数据生成的过程，由果溯因，就是因果推理

因果推理

- Simpson's Paradox (辛普森悖论)

1973年伯克利本科生录取率

	男生		女生	
	申请数	录取率	申请数	录取率
整体	8442	44%	4321	35%

男生录取率(44%)远高于女生(35%)

学院	男生		女生	
	申请数	录取率	申请数	录取率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	**37%**	593	34%
D	417	33%	375	35%
E	191	**28%**	393	24%
F	373	6%	341	7%

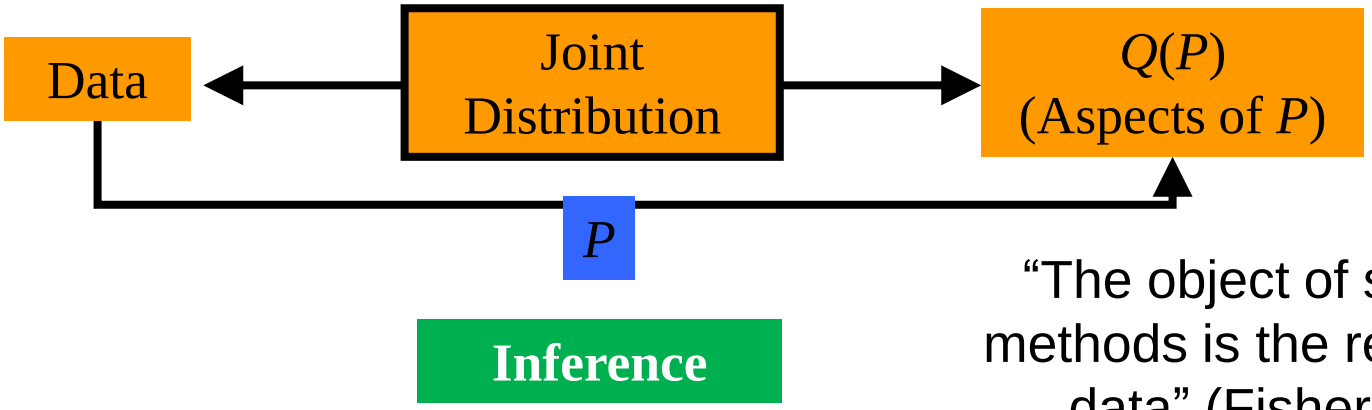
6个最大的院系中，4个院系女生录取率大于男生。如果按照这样的分类，女生实际上比男生的录取率还高一点。

女生更愿意申请那些竞争压力很大的院系（比如英语系），但是男生却更愿意申请那些相对容易进的院系（比如工程学系）。

Peter J. Bickel, Eugene A. Hammel,O'Connell, J. W, Sex bias in graduate admissions: Data from Berkeley, *Science*, 187(4175):398-404,1975

因果推理

传统以统计建模为核心的推理手段



“The object of statistical methods is the reduction of data” (Fisher 1922).

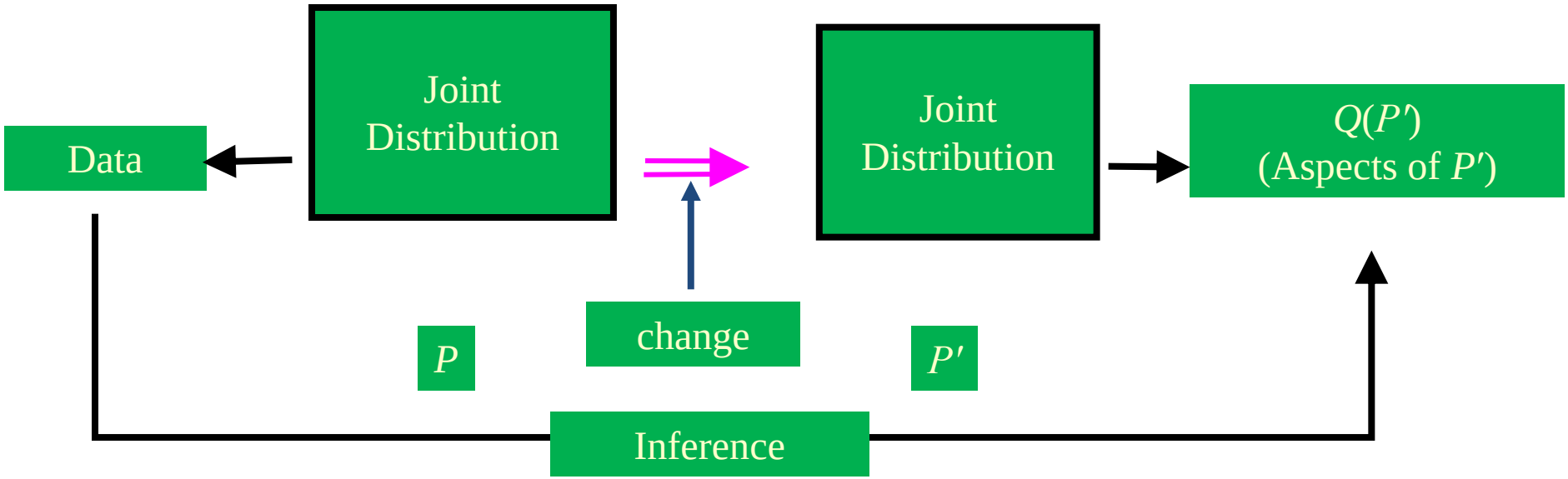
购买了A商品的顾客是否会购买B商品 (对A和B的联合分布建模)

$$Q = P(B|A)$$

Judea Pearl, Causality: models, reasoning, and inference (second edition)，Cambridge University Press, 2009
Judea Pearl, Madelyn Glymour, Nicholas P. Jewell, *Causal inference in statistics: a primer*, John Wiley & Sons, 2016

因果推理

从统计建模推断到因果推理



数据分布从 P 变换到 P'

- 如果商品价格涨价一倍，预测销售量 $P'(sales)$ 的变化
- 如果放弃吸烟，预测癌症 $P'(cancer)$ 的概率

因果推理

- 从关联到因果推理 (Causal Inference)

代表学者：Judea Pearl, Donald Rubin

观测问题	What if we see A (what is?)	$P(y\|A)$
干预问题	What if we do A (how?)	$P(y\|do(A))$ (如果采取A行为，y将如何)
反事实问题 (Counter-factual)	What if we had things differently (why?)	$P(y'\|A)$ (如果A不发生，y'将如何)
Options: with what probability		

关联 (association):
直接可从数据中计算得到的统计相关

干预 (intervention):
无法直接从观测数据就能得到关系，
如“某个商品涨价会产生什么结果”

反事实 (counterfactual):
某个事情已经发生了，则在相同环境中，
这个事情不发生会带来怎样的新结果

因果推理

因果推理 (Causal Inference) 的主要模型

- 结构因果模型 (structural causal model, SCM)：又称 causal model。用严谨的数学符号来表示随机变量之间的因果关系，从而准确地对一个数据集的数据生成过程进行描述。因果图 (causal graph/diagram)：Judea Pearl 1995年提出。一般而言，因果图是有向无环图 (directed acyclic graphs, DAG)。即从图中任意一个节点出发经过任意条边，均无法回到该节点。有向无环图刻画了图中所有节点之间的依赖关系。DAG 可用于描述数据的生成机制。
- 潜结果框架 (potential outcomes framework): 又称 Neyman–Rubin causal model (RCM)。这一模型最早可追溯于Jerzy Neyman在1923年用波兰语所完成论文中提出的“潜在结果” (potential outcome) 的概念。之后，Donald Rubin发展了“潜在结果”这一概念，并将其和缺失数据的理论联系在一起。

因果推理

- 结构因果模型 (structural causal model, SCM)

- 定义 2.15 结构因果模型：结构因果模型由两组变量集合U和V以及一组函数f组成。 $f = \{f_X : W_X \to X | X \in V\}$, 其中 $W_X \subseteq (U \cup V) - \{X\}$ 其它变量
- 定义 2.16 结构因果模型中的原因：如果变量 X 出现在给变量 Y 赋值的函数中，则X是Y的直接原因 (direct cause)。如果X是Y的直接原因或者其它原因（如原因的原因），均称X是Y的原因。
- U中的变量被称为外生变量 (exogenous variables)，即这些变量处于模型之外，不对其阐述和解释； V 中的变量称为内生变量 (endogenous variables)。
- 以图中的节点来说明内生变量和外生变量的关系：每一个内生变量都至少是一个外生变量的后代；而每一个外生变量都不是其他外生或内生变量的后代，它们没有祖先，也就是说，外生变量都是图中的根节点。如果知道了每一个外生变量的值，就可以使用函数f来计算出每一个内生变量的值。

因果推理

- 结构因果模型 (structural causal model, SCM)

例 2.17 在结构因果模型框架下讨论某种治疗方案X对肝脏功能Y是否产生影响的因果关系。在讨论X对Y的因果关系时，可能会假设肝脏功能Y会受到水污染Z的影响，由于水污染Z不会受到治疗方案X和肝脏功能Y的影响，因此，可将X和Y作为内生变量，Z作为外生变量来进行研究。

每个结构因果模型M都与一个因果图G相对应。因果图中的节点是结构因果模型中U和V所包括的变量，节点之间的边表示函数f。在M中，若变量X的函数f_x包含了变量Y（X的取值依赖于Y），则在G中有一条从Y到X的有向边。本书主要讨论因果图为有向无环图的结构因果模型。

因果推理

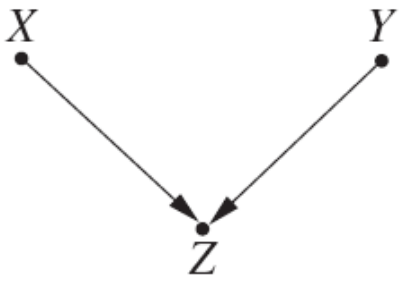
- 因果图 (causal diagram)

定义 2.18 因果图中的原因：在因果图中，若变量Y是另一个变量X的孩子，则X是Y的直接原因；若Y是X的后代，则X是Y的潜在原因。

例 2.18 假设某一商品销量Z、该商品与其他商品的价格差X以及为该商品促销所支出的广告费Y三个变量所形成的结构因果模型如下表示：

$$V = \{X, Y, Z\}, \qquad F = \{f_Z\}$$
$$f_Z: Z = 2X + 5Y$$

由于X和Y都出现在函数f_Z中，因此，X和Y都是Z的直接原因。若X和Y有其他祖先，则这些祖先是Z的潜在原因。



商品销量Z，价格差X和广告费Y的因果图

例 2.19 当例2.18中的商品销量、价格差和广告费的量化关系无法给出，并含有某些未知因素时，其部分指定的结构因果模型如下：

$$U = \{U_1, U_2, U_3\}, \qquad V = \{X, Y, Z\}, \qquad F = \{f_X, f_Y, f_Z\}$$
$$X = f_X(U_1)$$
$$Y = f_Y(U_2)$$
$$Z = f_Z(X, Y, U_3)$$

其中U_1, U_2, U_3代表某些未知的外生变量，有时又被称为“误差项”（error term）或“忽略因素”（omitted factor）。

因果推理

- 因果图 (causal diagram)

对于任意的有向无环图模型，模型中d个变量的联合概率分布由每个节点与其父母节点之间条件概率$P(child|parents)$的乘积给出：

$$P(x_1, x_2, \cdots, x_d) = \prod_{j=1}^{d} P(x_j | x_{pa(j)})$$

其中，$x_{pa(j)}$表示节点x_j的父母节点集合（所有指向x_j的节点）。这里包含了变量之间某种普遍成立的独立性假设。与贝叶斯网络的计算公式相同

对于一个简单的链式图$X \to Y \to Z$，其联合概率分布可直接写成：

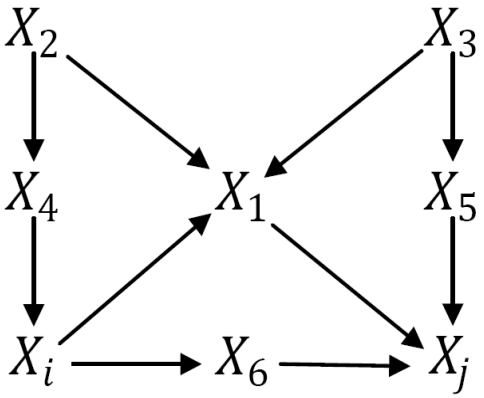
$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y|X = x)P(Z = z|Y = y)$$

因果推理

例 2.20 考虑链式图$X \to Y \to Z$，其中X表示气候好，Y表示水果产量高，Z表示水果价格低，给出P(气候好, 水果产量高, 水果价格低)的联合概率。

解： 使用乘积分解规则，将P(气候好, 水果产量高, 水果价格低)转换为：
P(气候好) P(水果产量高|气候好) P(水果价格低|水果产量高)

根据常识，假设 P (气候好) $= 0.5$ ，当然也可设置为更低或更高的取值。类似地，在气候好的时候水果应该生长得比较好，产量会比较高，因此假设P(水果产量高|气候好) $= 0.8$。同样，由于市场机制，水果产量高的时候，水果价格会比较低，设P(水果价格低|水果产量高) $= 0.9$。因此计算出P(气候好，水果产量高，水果价格低)= P(气候好) P(水果产量高|气候好) P(水果价格低|水果产量高) $= 0.5 \times 0.8 \times 0.9 = 0.36$。



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_i, X_j)$$
$$= P(X_2) \times P(X_3) \times P(X_1|X_2, X_3, X_i) \times P(X_4|X_2)$$
$$\times P(X_5|X_3) \times P(X_6|X_i) \times P(X_i|X_4) \times P(X_j|X_1, X_5, X_6)$$

因果推理

- 因果图的基本结构 链(chain)：包含三个节点两条边，其中一条边由第一个节点指向第二个节点，另一条边由第二个节点指向第三个节点。

例 2.22 考虑如下的结构因果模型：

$$U = \{U_1, U_2, U_3\}, V = \{X, Y, Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X: X = U_1$$

$$f_Z: Z = 3X + 10 + U_2$$

$$f_Y: Y = 5Z + U_3$$

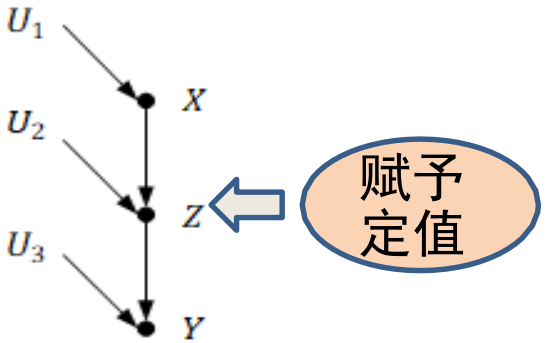


图2.5 链式因果图

链式图$X \to Z \to Y$中，X和Y在给定Z时条件独立也可以这样理解：在给定Z时，若X的取值发生变化，则为了维持Z的取值不变，U_2的取值会发生变化；而由于Y的取值只依赖于Z和 U_3，但Z是给定的，因此Y的取值不会发生变化。也就是说，给定Z时，X的取值不会影响Y的取值，因此给定Z时，Y和X是条件独立的。

定理2.5 (链中的条件独立性)对于变量X和Y，若X和Y之间只有一条单向的路径，变量Z是截断 (intercept) 该路径的集合中的任一变量，则在给定Z时，X和Y条件独立。

不给定Z时，X和Y不独立。

因果推理

- 因果图的基本结构 分连 (fork): 包含三个节点两条边，两条边分别由第一个节点指向第二个节点和第三个节点。

例 2.23 考虑如下的结构因果模型：

$$U = \{U_1, U_2, U_3\}, \qquad V = \{X, Y, Z\},$$
$$F = \{f_X, f_Y, f_Z\}$$
$$f_Z: Z = U_1$$
$$f_X: X = 3Z + 7 + U_2$$
$$f_Y: Y = 6Z + U_3$$

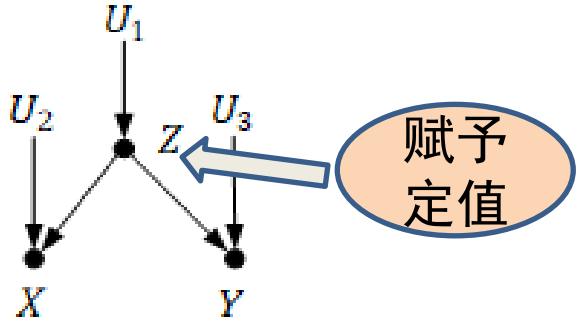


图2.6 分连 (fork) 因果图

在分连图$X \leftarrow Z \rightarrow Y$中， X和Y在给定Z时条件独立可以这样理解：若给定Z，Y和X的取值不会变化，它们的值只会分别随U_3和U_2变化，因此给定Z时，Y和X是条件独立的。

定理 2.6 (分连中的条件独立性)若变量Z是变量X和Y的共同原因，且X到Y只有一条路径，则在给定Z时，X和Y条件独立。
不给定Z时，X和Y不独立。

因果推理

- 因果图的基本结构 汇连(collider): 又称为碰撞。它包含三个节点两条边，两条边分别由第一个节点和第二个节点指向第三个节点。

例 2.24 考虑如下的结构因果模型：

$$U = \{U_1, U_2, U_3\}, \qquad V = \{X, Y, Z\},$$

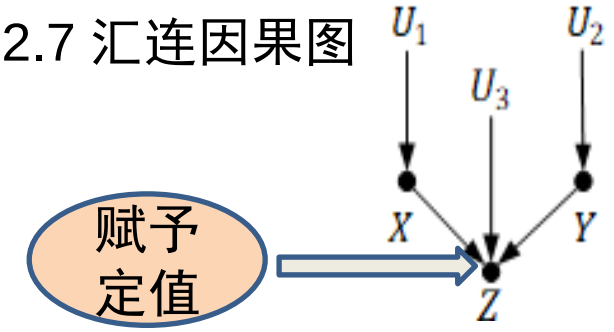
$$F = \{f_X, f_Y, f_Z\}$$

$$f_X: X = U_1$$

$$f_Y: Y = U_2$$

$$f_Z: Z = X + Y + U_3$$

图2.7 汇连因果图



在汇连图$X \to Z \leftarrow Y$中， X和Y在给定Z时条件相关可以这样理解：给定Z，当X的取值发生变化时，为了保证Z的取值不变，Y的取值也一定会发生变化，因而给定Z时，Y和X是相关的。

定理 2.7 (汇连中的条件独立性)若变量Z是变量X和Y的汇连节点，且X到Y只有一条路径，则X和Y相互独立，但在给定Z或Z的后代时，X和Y是相关的。

给定Z的后代，X和Y是相关的，可以这样理解：不妨设Z的某个后代为W，给定W时，则W的祖先也给定，W的祖先给定时，则W的祖先的祖先也给定……因为Z是W的祖先，因此Z也会给定，此时，X和Y是相关的。

不给定Z时， X和Y独立。

因果推理

- 干预的因果效应

干预：干预 (intervention) 指的是固定 (fix) 系统中的变量，然后改变系统，观察其他变量的变化。

为了与X自然取值x时进行区分，在对X进行干预时，引入“do算子”(do-calculus)，记作$do(X = x)$。

因此，$P(Y = y|X = x)$表示的是当$X = x$时，$Y = y$的概率；而$P(Y = y|do(X = x))$表示的是对X进行干预，固定其值为x时，$Y = y$的概率。用统计学的术语来说，$P(Y = y|X = x)$反映的是在取值为x的个体X上，Y的总体分布；而$P(Y = y|do(X = x))$反映的是如果将每一个X取值都固定为x时，Y的总体分布。

以变量为条件是改变了看世界的角度，而干预则改变了世界本身

因果推理

- 干预的因果效应

	不用药	用药
恢复人数	289	273
总人数	350	350
恢复率(%)	83	78

	不用药		用药	
	男性	女性	男性	女性
恢复人数	234	55	81	192
总人数	270	80	87	263
恢复率(%)	87	69	93	73

X表示用药情况(1表示用药，0表示不用药)，Y表示病人的恢复情况(1表示恢复，0表示未恢复)，Z表示性别(1表示男性, 0表示女性)。

为了比较病人用药与否的恢复情况，可以对用药情况分别进行干预，即将病人都分别固定成用药病人和不用药病人。不妨设$do(X=1)$和$do(X=0)$分别表示这两种干预，并估计其中的差别：

$$P(Y=1|do(X=1))-P(Y=1|do(X=0))$$

这被称为“因果效应差” (causal effect difference)或“平均因果效应” (average causal effect, ACE), $P(Y=y|do(X=x))$ 被称为因果效应。

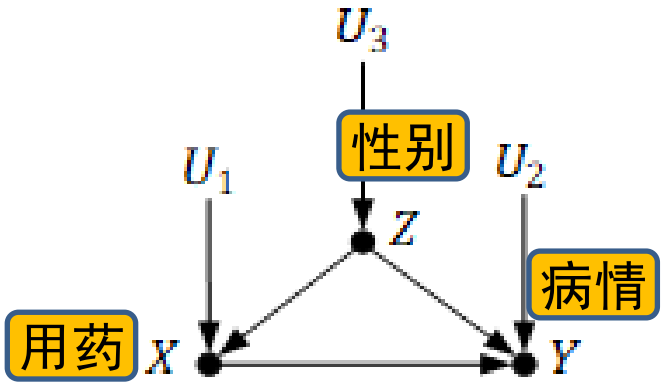


图2.9 辛普森悖论的因果图

因果推理

- 干预的因果效应

对用药情况X进行干预并固定其值为x时，可将所有指向X的边均移除，则因果效应$P(Y=y|do(X=x))$等价于图2.10所示的引入干预的操纵图模型 (manipulated model) 中的条件概率$P_m(Y=y|X=x)$。

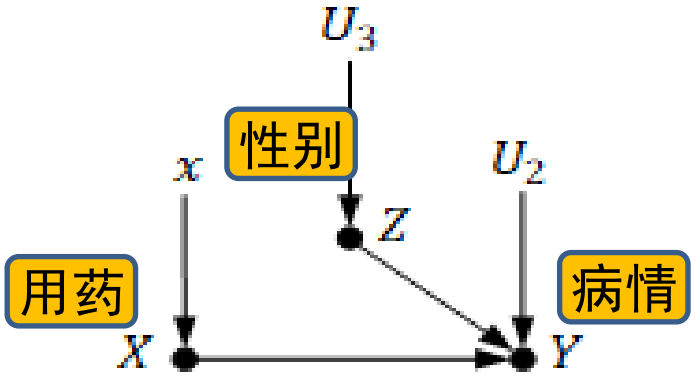


图2.10 引入干预的操纵图模型

计算因果效应的关键在于计算操纵概率 (manipulated probability) P_m。P_m与正常 (无干预，如图2.9) 条件下的概率P有两个相同的重要性质

- 边缘概率$P(Z=z)$不随干预而变化，因为Z的取值不会因为去掉从Z到X的箭头而变化。
- 条件概率$P(Y=y|X=x,Z=z)$不变，因为Y关于X和Z的函数$f_Y=(X,Z,U_2)$并未改变。

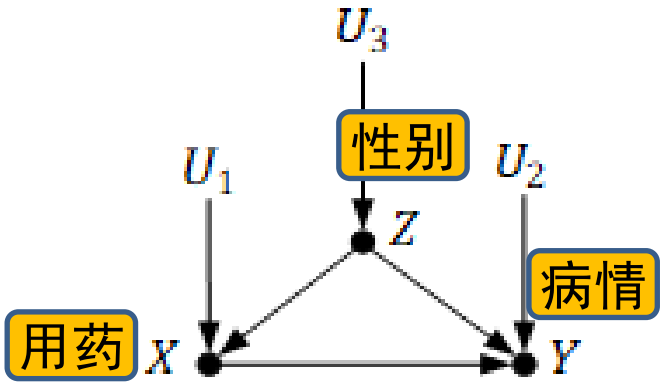


图2.9 辛普森悖论的因果图

因此，有如下等式：

$$P_m(Y=y|X=x,Z=z)=P(Y=y|X=x,Z=z)$$
$$P_m(Z=z)=P(Z=z)$$

因果推理

- 干预的因果效应

在操纵图模型中，X和Z是D-分离，即

$$P_m(Z=z|X=x)=P_m(Z=z)=P(Z=z)。$$

将三个等式组合在一起，则因果效应$P(Y=y|do(X=x))$：

$$P(Y=y|do(X=x))=P_m(Y=y|X=x)$$
$$=\sum_z P_m(Y=y|X=x,Z=z)P_m(Z=z|X=x)$$
$$=\sum_z P_m(Y=y|X=x,Z=z)P_m(Z=z)$$

得到如下正常(无干预)条件下的概率表示的因果效应：

$$P(Y=y|do(X=x))$$
$$=\sum_z P(Y=y|X=x,Z=z)P(Z=z)$$

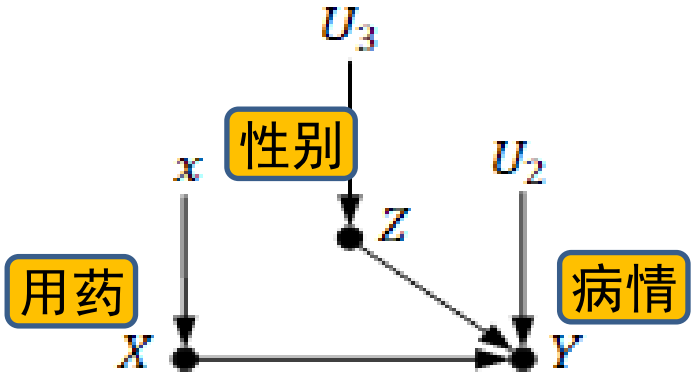


图2.10 引入干预的操纵图模型

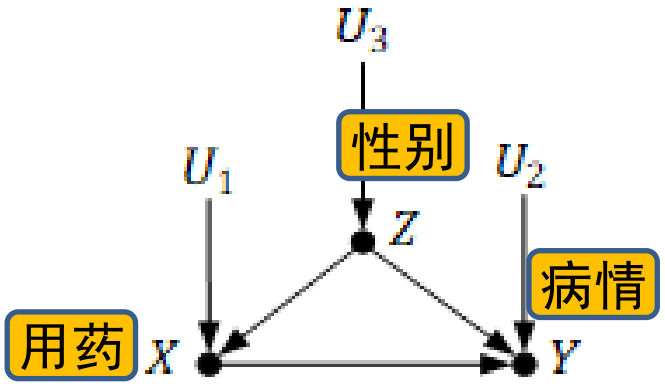


图2.9 辛普森悖论的因果图

这被称为调整公式(adjustment formula)，对于Z的每一个取值z，上式计算X和Y的条件概率并取均值。这个过程称之为“Z调整”(adjusting for Z)或“Z控制”(controlling for Z)。上式右端只包含正常(无干预)条件下的概率P，即可用正常(无干预)条件下的条件概率来计算干预后的条件概率。

因果推理

- 干预的因果效应

其中$X=1$表示用药，$Y=1$表示病人恢复，$Z=1$表示男性，$Z=0$表示女性，则有：$P(Y=1|do(X=1))$

$= P(Y=1|X=1,Z=1)P(Z=1) + P(Y=1|X=1,Z=0)P(Z=0)$

代入表2.10的数据，则将病人干预为用药病人的因果效应为：

$$P(Y=1|do(X=1)) = 0.93 \times \frac{(87+270)}{(350+350)} + 0.73 \times \frac{(263+80)}{(350+350)} = 0.832$$

类似地，将病人干预为不用药病人的因果效应为：

$$P(Y=1|do(X=0)) = 0.87 \times \frac{(87+270)}{(350+350)} + 0.69 \times \frac{(263+80)}{(350+350)} = 0.7818$$

在分别计算完用药病人和不用药病人的干预因果效应后，再计算其因果效应差：

$ACE = P(Y=1|do(X=1)) - P(Y=1|do(X=0)) = 0.832 - 0.7818 = 0.0502$

说明用药病人的恢复率高于不用药病人的恢复率，即该新药能帮助治愈病人。

注意：先前有 $P(Y=1|X=1)=0.78$ 和 $P(Y=1|X=0)=0.83$ （存在辛普森悖论）。可见如果不加干预仅从原始数据归纳所得条件概率无法得到正确结论。

因果推理

- 干预的因果效应

在上面的例子中，通过将Z放入调整公式中，能够利用正常(无干预)条件下的条件概率计算出干预后的因果效应。那么，哪些变量(或变量集合)可以放入调整公式中呢？在进行干预时，由于要将X固定，并将所有指向X的箭头都去掉，也就是说需要让X的父节点失效，因此，应该将X的父节点放入调整公式中。令X的父节点集合为$PA(X)$，则有如下定理：

定理 2.8 (因果效应)给定因果图G，PA表示X的父节点集合，则X对Y的因果效应为：

$$P(Y=y|do(X=x))$$
$$=\sum_{z} P(Y=y|X=x,PA=z)P(PA=z)$$

其中，z是PA的非空子集。

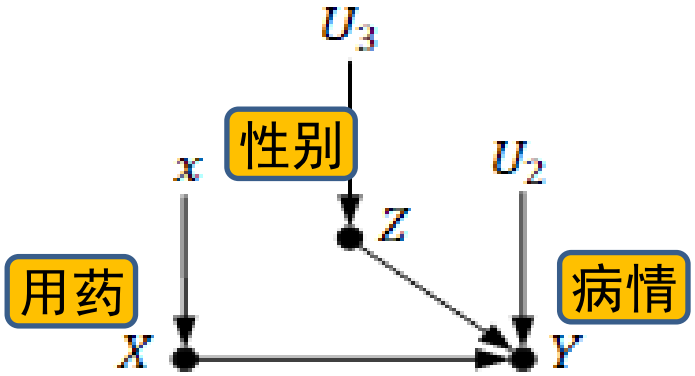


图2.10 引入干预的操纵图模型

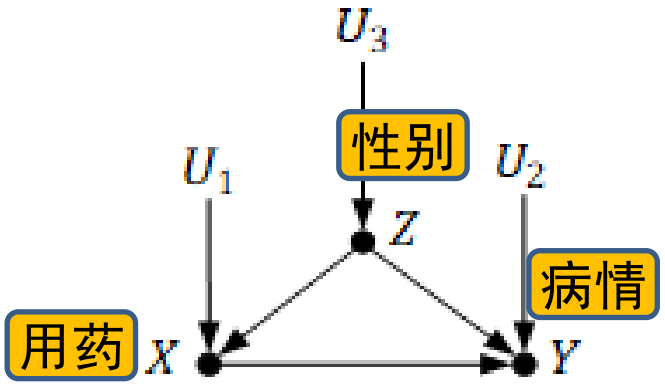


图2.9 辛普森悖论的因果图

因果推理

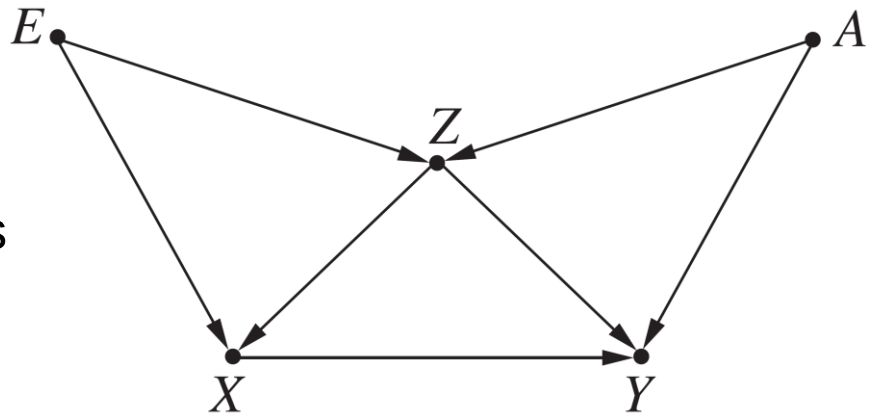
- 后门准则

给定有向无环图中的一对有序变量 (X, Y), 如果变量集合 Z 满足：Z 中没有 X 的后代结点，且 Z 阻断了 X 与 Y 之间的每条含有指向 (contains an arrow into) X 的路径，则称 Z 满足关于 (X, Y) 的后门准则。

如果 Z 满足 (X, Y) 的后门准则，那么 X 对 Y 的因果效应为：

$$P(Y = y|do(X = x)) = \sum_{z} P(Y = y|X = x, Z = z)P(Z = z)$$

注意：$PA(X)$ 总是满足后门准则。



后门准则背后的逻辑：

以这样的结点集合 Z 为条件：

(1) 阻断 X 和 Y 之间所有伪路径 (spurious path)，

(2) 保持所有从 X 到 Y 的有向路径不变，

(3) 不会产生新的伪路径。

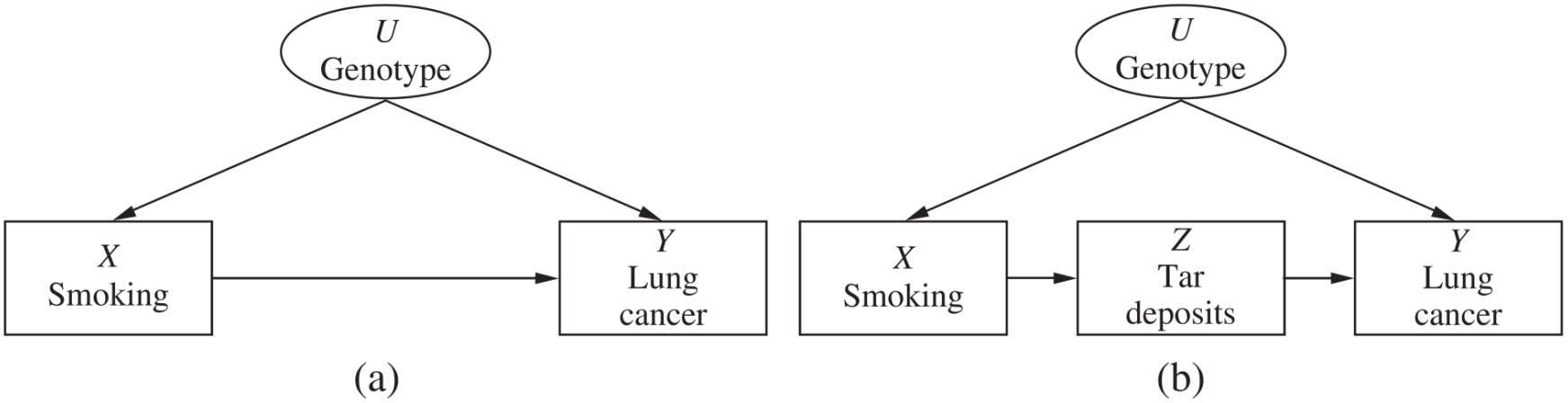
A spurious path in a causal graph introduces a misleading association between variables due to factors like common causes or incorrect conditioning on colliders, making variables appear related when they are not causally connected.

因果推理

- 前门准则

定义 前门变量集合 Z 被称为满足关于有序变量对 (X, Y) 的前门准则，如果：

(1) Z 阻断所有 X 到 Y 的有向路径；
(2) X 到 Z 没有后门路径；
(3) 所有 Z 到 Y 的后门路径都被 X 阻断。

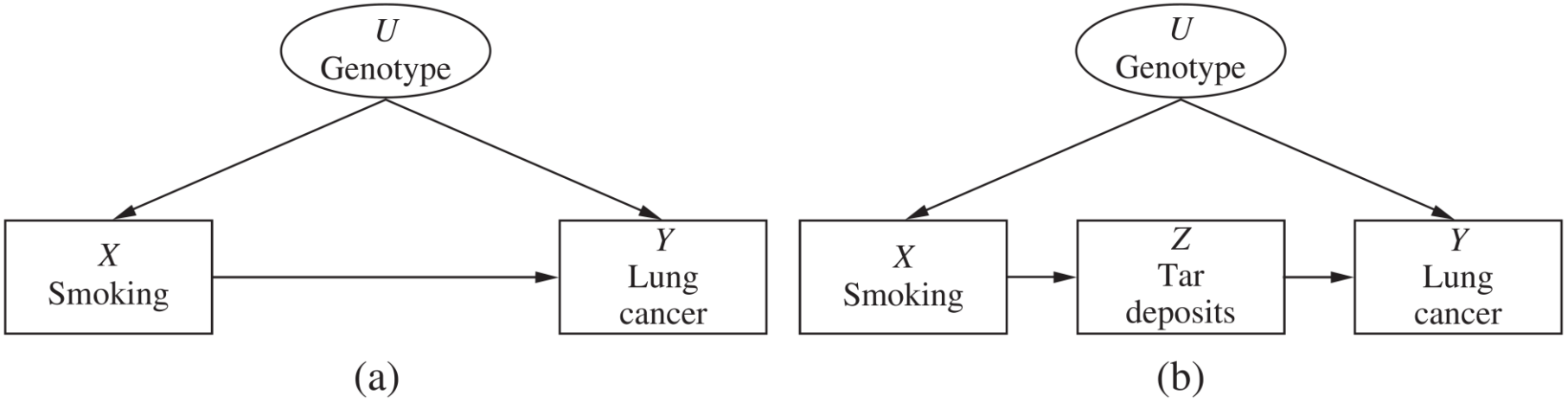


因果推理

- 前门准则

前门校正 如果变量集合 Z 满足变量对(X,Y) 的前门准则，且 $p(x,z)>0$，那么 X 对 Y 的因果效应是可识别的，且由下式计算：

$$P(Y=y|do(X=x))=\sum_{z}P(z|x)\sum_{x'}P(y|x',z)\,P(x')$$



因果推理

- 逆概率加权

如果 Z 满足 (X,Y) 的后门准则，那么 X 对 Y 的因果效应为：

$$P(Y=y|do(X=x))=\sum_{z}P(Y=y|X=x,Z=z)P(Z=z)$$

不必做实际模拟干预即可做出预测 $P(y|do(x))$ 。需要分别查看 Z 的每个值或值的每一种组合，估计每种情况中给定 的条件概率，然后对结果求平均值。随着规模增大，对 Z 的校正会产生计算上和估计上的困难。

$$P(y|do(x))=\sum_{z}\frac{P(Y=y,X=x,Z=z)}{P(X=x|Z=z)}$$

使用有限样本估计 $P(Y=y|do(X=x))$ 的简单方法：如果通过因子 $1/P(X=x|Z=z)$ 对每个可用的样本进行加权，那么将重新加权的样本看作是从 P_m 生产的而不是 P，然后用它们来估计 $P(Y=y|do(x))$。Z 可能的取值很多时效果明显。

因果推理

因果推理 (Causal Inference) 的主要模型

- 潜结果框架 (potential outcomes framework):

To identify causal effects from observational data, certain assumptions are necessary:

- Stable Unit Treatment Value Assumption (SUTVA): SUTVA has two parts:
  - No interference: The treatment of one unit does not affect the outcome of another unit.
  - Consistency: The observed outcome for a unit matches the potential outcome corresponding to the actual treatment received.
- Ignorability (Unconfoundedness): Ignorability assumes that, given a set of observed covariates X, the treatment assignment is independent of the potential outcomes: $\{Y_i^1, Y_i^0\} \perp X_i$

  In observational studies, this assumption is more challenging but essential for unbiased causal inference.
- Overlap (Positivity):

  The overlap assumption requires that each unit has a positive probability of receiving either treatment or control, given covariates X: $0 < P(T = 1|X) < 1$

  This ensures there is sufficient data for comparison between treated and control units across different values of X.

因果推理

因果推理 (Causal Inference) 的主要模型

- 结构因果模型 (structural causal model, SCM)：又称 causal model。用严谨的数学符号来表示随机变量之间的因果关系，从而准确地对一个数据集的数据生成过程进行描述。因果图 (causal graph/diagram)：Judea Pearl 1995 年提出。一般而言，因果图是有向无环图 (directed acyclic graphs, DAG)。即从图中任意一个节点出发经过任意条边，均无法回到该节点。有向无环图刻画了图中所有节点之间的依赖关系。DAG 可用于描述数据的生成机制。
- 潜结果框架 (potential outcomes framework): 又称 Neyman–Rubin causal model (RCM)。这一模型最早可追溯于Jerzy Neyman在1923年用波兰语所完成论文中提出的“潜在结果” (potential outcome) 的概念。之后，Donald Rubin发展了“潜在结果”这一概念，并将其和缺失数据的理论联系在一起。

➢ Pearl (2000) argues that all potential outcomes can be derived from Structural Equation Models (SEMs) thus unifying econometrics and modern causal analysis. Richardson and Robins [2013] propose to use single world intervention graphs. These graphs allow us to set variables to certain values and therefore construct graphical correspondences to counterfactual variables.

因果推理

- 反事实模型 (counterfactual model)
  - 反事实模型 (counterfactual model，也叫potential outcomes) 是大卫•刘易斯 (David Lewis) 等人提出的推断因果关系的标准。
  - 反事实描述的是：假设存在一个虚拟的平行世界，里面的所有因素与现实世界一模一样，两个相同的个体他和“他”，分别在现实世界和平行世界中同时同地做了不同的选择，现在他知道了现实世界中的结果，他想知道平行世界中的那个“他”的选择所带来的结果。然而，平行世界并不存在。幸运的是，反事实将告诉他另一个“他”的选择所带来的结果。

因果推理

- 反事实模型 (counterfactual model)

若用符号$U = u$表示某个叫“张三”的个体的特征，X表示某个称为“身高”的变量，则$X(u)$表示张三的身高。反事实语句“在环境$U = u$下，若$X = x$，则$Y = y$”可表示成$Y_x(u) = y$，其中Y和X是V中的两个变量。考虑如下的因果模型M：

$$X = aU + 1$$
$$Y = bX + U + 2$$

首先计算反事实$Y_x(u)$，即在环境$U = u$时，若$X = x$，则Y应如何取值。将$X = x$代入第一个等式中，则有“修正” (modified) 模型M_x：

$$X = x$$
$$Y = bX + U + 2$$

将$U = u$和$X = x$代入第二个等式中，则有：

$$Y_x(u) = bx + u + 2$$

该结果与预期一致，即“Y的值本来就应该为X的b倍加上u，再加上常数2”。

因果推理

- 反事实模型 (counterfactual model)

若用符号 $U = u$ 表示某个叫“张三”的个体的特征，X 表示某个称为“身高”的变量，则 $X(u)$ 表示张三的身高。考虑如下的因果模型 M：

$$X = aU + 1$$
$$Y = bX + U + 2$$

计算反事实$X_y(u)$，即“在环境 $U = u$ 下，若 $Y = y$，则 X 如何取值”。将 $Y = y$ 代入模型M中，则有修正模型M_y：

$$X = aU + 1$$
$$Y = y$$

再将$U = u$代入M_y的第一个等式，此时有$X_y(u) = au + 1$，即“若$Y = y$，X的取值不变”。X在反事实条件下的不变性表明，“对未来结果的假设并不会改变过去的选择”。

因果推理

- 反事实模型 (counterfactual model)

若用符号$U = u$表示某个叫“张三”的个体的特征，X表示某个称为“身高”的变量，则$X(u)$表示张三的身高。考虑如下的因果模型M：

$$X = aU + 1$$
$$Y = bX + U + 2$$

上述的模型M中包含多种反事实，不妨设U的取值范围为$\{1, 2, 3\}$，$a = 1$，$b = 2$，则在不同x和y的取值情况下，$X(u)$，$Y(u)$，以及反事实$Y_x(u)$和$X_y(u)$的取值情况如表2.11所示。

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	2	7	5	7	9	2	2	2
2	3	10	6	8	10	3	3	3
3	4	13	7	9	11	4	4	4

反事实与前述的干预的区别：干预计算的是概率分布，它的计算结果是一个概率；反事实计算的是在假设$X = x$下，Y的取值，它的计算结果是一个值。从实验者的角度来看，干预描述的是总体的行为；反事实描述的是在环境$U = u$下，某个个体的行为。

因果推理

- 反事实模型 (counterfactual model)

图2.11表示的是兴趣爱好与知识渊博程度的因果图。其中，X表示求知欲望的程度，Z表示看书的数量，Y表示知识的渊博程度。为了表示方便，将三个变量均进行归一化，即它们的均值均为0，方差均为1。

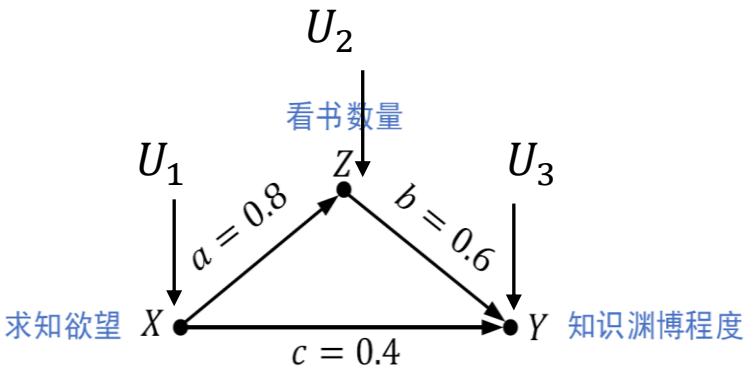


图2.11 求知欲望与知识渊博程度因果图

设图2.11的因果模型为：

$$X = U_1$$
$$Z = aX + U_2$$
$$Y = bZ + cX + U_3$$

其中，$U_i(i = 1, 2, 3)$为外生变量，它们相互独立。a, b, c的取值可从统计数据中进行估计，不妨设$a = 0.8, b = 0.6, c = 0.4$。设有一个叫张三的同学，发现其$X = 0.3, Z = 0.4, Y = 0.5$，那么，如果张三将其看书的数量提高一倍，那么张三的知识渊博程度为多少呢？

因果推理

- 反事实模型 (counterfactual model)

在该因果模型中，变量的值由系数和外生变量决定，且其中的外生变量将作用于所有个体。因此，可将张三的观测值$X = 0.3, Z = 0.4, Y = 0.5$作为证据$E$，反推外生变量的值。即有如下等式：

$$U_1 = 0.3$$
$$U_2 = 0.4 - 0.8 \times 0.3 = 0.16$$
$$U_3 = 0.5 - 0.6 \times 0.4 - 0.4 \times 0.3 = 0.14$$

接下来，用$Z = 0.8$表示张三的看书数量提升一倍，修正后的模型如图2.12所示。则反事实 $Y_{Z=0.8}(U_1 = 0.3, U_2 = 0.16, U_3 = 0.14)$：

$$Y_{Z=0.8}(U_1 = 0.3, U_2 = 0.16, U_3 = 0.14)$$
$$= 0.6 \times 0.8 + 0.4 \times 0.3 + 0.14 = 0.74$$

即，若将张三的看书数量提升一倍，则他的知识渊博程度将由0.5变为0.74。

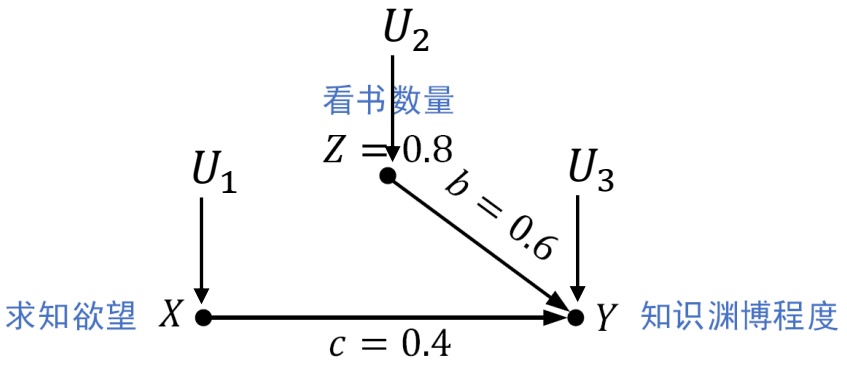


图2.12 将看书数量提升一倍后的反事实模型图

因果推理

- 反事实模型 (counterfactual model)

反事实计算的三个步骤：

- (1) 诱拐 (abduction)：利用现有的证据 E 确定环境 U；
- (2) 动作 (action)：对模型 M 进行修改，移除等式 X 中的变量并将其替换为 $X = x$，得到修正模型 M_x；
- (3) 预测 (prediction)：利用修正模型 M_x 和环境 U 计算反事实 $Y_x(U)$ 的值。

因果推理

- 因果分析的层次化

表2.12 因果模型的层次化示意图

观测问题	What if we see A (what is?)	$P(y\|A)$	关联(association): 直接可从数据中计算得到的统计相关
干预问题	What if we do A (how?)	$P(y\|do(A))$ (如果采取A行为，则y将如何)	干预(intervention): 无法直接从观测数据就能得到关系，如“某个商品涨价会产生什么结果”
反事实问题	What if we had done things differently (why?)	$P(y'\|A)$ (如果A 不发生，则y'将如何)	反事实(counterfactual): 某个事情已经发生了，则在相同环境中，这个事情不发生会带来怎样的新结果

因果推理

- 因果分析的层次化

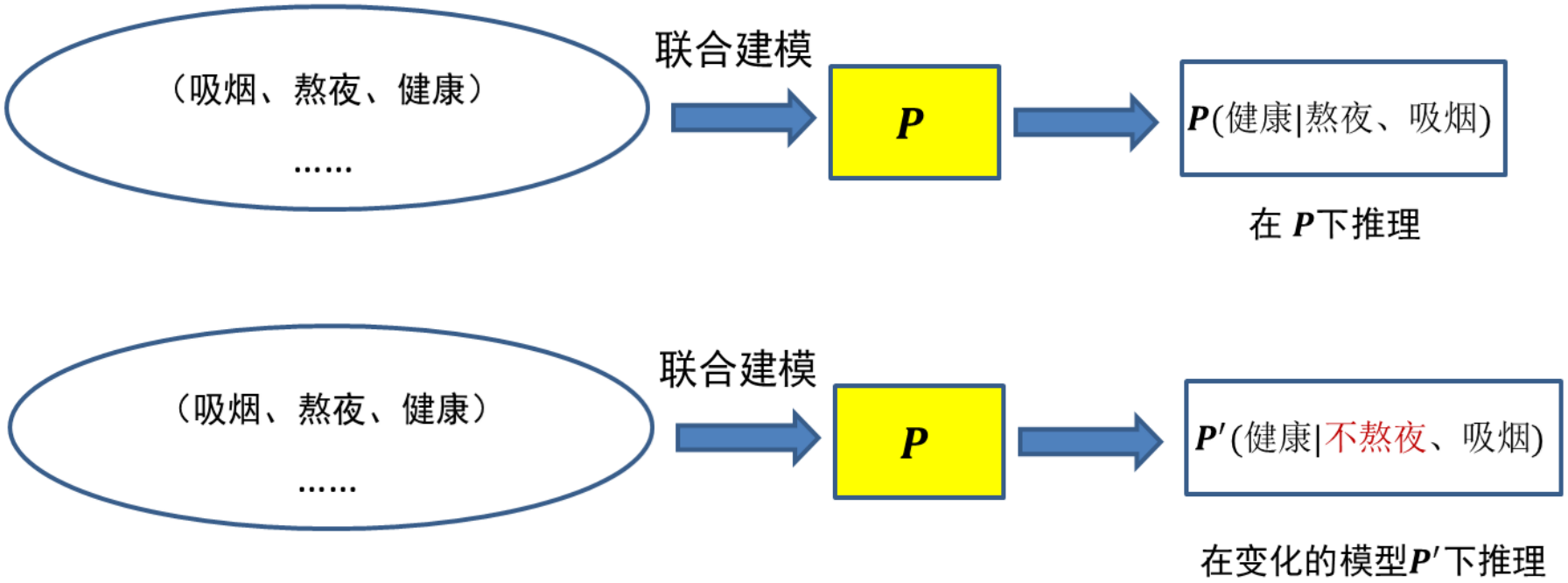


图2.13 统计学习与因果推理区别示意图

- 统计学习：从收集的观测数据中已经训练优化得到一个数据联合发布模型P.
- 反事实推理：因为模型变量出现了变化，因此需要在一个新的模型P'下进行分析。