因果推断

北京大学　龚诚欣
https://wqgcx.github.io/

2022 年 5 月 11 日

目录

目录

1 统计和因果模型

Reichenbach 的共同原因原则: 如果两个随机变量 X 和 Y 统计相关 ($X \not\perp\!\!\!\perp Y$), 则存在第三个变量 Z 对两者都具有因果影响 (Z 可能与 X 或 Y 重合). 此外, 变量 Z 可以屏蔽 X 和 Y, 即在给定 Z 的情况下, X 和 Y 将变得相互独立 ($X \perp\!\!\!\perp Y|Z$).

2 因果推断假设

系统变量的因果生成过程由互不知晓或互不影响的自治模块组成. 在概率情况下, 这意味着每个变量在给定其原因是的条件分布没有通知或影响其他条件分布. 在只有两个变量的情况下, 这就归纳为原因分布和产生效应分布的机制之间的独立性.

3 原因-效果模型

3.1 结构因果模型

$C := N_C, E := f_E(C, N_E), N_E \perp\!\!\!\perp N_C$. C: 原因变量, E: 效果变量. 称 C 是 E 的直接原因, $C \to E$ 是因果图. 模型记为 Structural Causal Model (SCM).

3.2 干预

硬干预: $E := 4$, $\mathrm{do}(E := 4)$. 软干预: $\mathrm{do}(E := g_E(C) + \widetilde{N}_E)$.

3.3 反事实

一种疗法对 99% 的患者有效, 患者痊愈 ($B = 0$), 否则失明 ($B = 1$). 其余 1% 患者无效, 患者失明 ($B = 1$), 否则痊愈 ($B = 0$). 记为 N_B. 医生不知道是哪一类患者, 是否给予治疗 ($T = 1$) 与 N_B 无关, 记为 N_T. 假定 SCM 为 $T := N_T, B := T \cdot N_B + (1 - T) \cdot (1 - N_B)$, 其中 $N_B \sim \mathrm{Bernoulli}(0.01)$. 因果图为 $T \to B$.

如果某给定患者 $T = 1 \Rightarrow B = 1$, 则有 $T = 0 \Rightarrow B = 0$. 这等价于赋值 $N_B = 1$. 上述例子表明, 可以使用反事实的陈述来证伪潜在因果模型.

3.4 结果因果模型的标准表示

考虑 $E := f_E(C, N_E)$. 对于 N_E 的每个取值 n_E, E 是 C 的确定函数. 这也是说, 噪声 N_E 在从 $\mathcal{C}$ 到 $\mathcal{E}$ 的不同函数之间切换. 可以假设 N_E 在从 $\mathcal{C}$ 到 $\mathcal{E}$ 的函数集合中取值, 表示为 $\mathcal{E}^{\mathcal{C}}$. 因此可以重写为 $E = n_E(C)$.

两个标准表示不同的 SCM 可以引入相同的干预概率.

4 学习原因-效果模型

4.1 结构可识别性

图结构的非唯一性: 对于两个实值变量任意联合分布 $P_{X,Y}$, 存在 SCM: $Y = f_Y(X, N_Y), X \perp\!\!\!\perp N_Y$, 其中 f_Y 可测, N_Y 是实值噪声变量. 这说明被动的观察无法推断出观测变量之间的因果方向.

Proof: 考虑 $F_{Y|X} := P(Y \leq y|X = x)$ 并定义 $f_Y(x, n_Y) := F^{-1}_{Y|X}(n_Y), N_Y \sim U(0, 1)$.

线性非高斯模型 (Linear Non-Gaussian Models) 的可识别性: 设 $P_{X,Y}$ 满足线性模型 $Y = \alpha X + N_Y, N_Y \perp\!\!\!\perp X$, X, Y, N_Y 是连续型随机变量, 则存在 $\beta \in \mathbb{R}$ 和一个随机变量 N_X 满足 $X = \beta Y + N_X, N_X \perp\!\!\!\perp Y$ 当且仅当 N_Y 和 X 是高斯分布. 这意味着如果 C 或 N_E 是非高斯分布, 就足以确定因果方向.

Darmois-Skitovic 定理: $X_1, \cdots, X_d$ 是独立的非退化随机变量, 如果存在非零系数 $a_1, \cdots a_d$ 和 $b_1, \cdots, b_d$ 使得线性组合 $l_1 = \sum_{i=1}^d a_i X_i, l_2 = \sum_{i=1}^d b_i X_i$ 独立, 则每个 X_i 服从高斯分布.

加性噪声模型 (Additive Noise Model, ANM): $Y = f_Y(X) + N_Y, N_Y \perp\!\!\!\perp X$.

ANM 的可识别性: 如果 N_Y 和 X 有严格的正密度 p_{N_Y} 和 p_X, 并且 f_Y, p_{N_Y} 和 p_X 是三阶可微的, 则称 ANM 平滑. 假设 $P_{Y|X}$ 允许一个从 X 到 Y 的平滑 ANM, 并且存在一个这样的 $y \in \mathbb{R}$ 满足 $(\log p_{N_Y})''(y - f_Y(x))f'_Y(x) \neq 0$ 对几乎所有的 x 都成立. 然后, 对数密度 $\log p_X$(满足得到的联合分布 $P_{X,Y}$ 允许一个从 Y 到 X 的平滑 ANM) 的集合包含在一个三维仿射空间 (线性流形) 中的.

离散加性噪声模型 (Discrete ANM) 的可识别性: 假设一个分布 $P_{X,Y}$ 允许一个从 X 到 Y 的 ANM $Y = f(X) + N_Y$, 其中 X 和 Y 有有限的支撑集. 当且仅当存在不相交分解 $\cup_{i=0}^l C_i = \text{supp}(X)$, 且满足以下 3 个条件时, 存在 $P_{X,Y}$ 允许一个 Y 到 X 的 ANM: (1) C_i 为彼此移位版本: $\forall i \exists d_i \geq 0 : C_i = C_0 + d_i$, f 为分段常值函数, $f|_{C_i} \equiv c_i$; (2) 对 $C_i s$ 上得到概率分布进行平移和缩放, 平移常数与上面相同: 对于 $x \in C_i$, $P(X = x)$ 满足 $P(X = x|X \in C_i) = P(X = x - d_i|X \in C_0)$; (3) 集合 $c_i + \text{supp}N_Y := \{c_i + h : P(N_Y = h) > 0\}$ 是不相交的.

对于离散 ANM 模 m 也有相似的结果. 注意均匀噪声分布起着特殊的作用: $Y \equiv f(X) + N_Y \mod m$ 加上一个均匀分布在 $\{0, 1, \cdots, m - 1\}$ 上的噪声导致独立的 X 和 Y, 因此也允许 Y 到 X 的 ANM.

后非线性模型 (Post-nonlinear Models): $Y = g_Y(f_Y(X) + N_Y), N_Y \perp\!\!\!\perp X$.

后非线性模型的可识别性: $P_{X,Y}$ 表示一个从 X 到 Y 的后非线性模型, p_X, f_Y, g_Y 三阶可微. 只有当它们相互调整后能满足 (Zhang and Hyvarinen, 2009) 中描述的微分方程时, $P_{X,Y}$ 表示一个从 Y 到 X 的后非线性模型.

信息-几何因果推断 (Information-Geometric Causal Inference, IGCL): $Y = f(X)$ 的微分同胚映射 f 在 $[0, 1]$ 上严格单调且 $f(0) = 0, f(1) = 1$. 此外, P_X 具有连续密度 p_X 满足 $\text{cov}(\log f', p_X) = \int_0^1 \log f'(x)p_X(x)dx - \int_0^1 \log f'(x)dx \int_0^1 p_X(x)dx = 0$.

IGCI 模型的可识别性: 假设分布 $P_{X,Y}$ 满足一个从 X 到 Y 的 IGCI 模型, 则逆函数 f^{-1} 满足 $\text{cov}(\log(f^{-1})', p_Y) \geq 0$ 当且仅当 f 是恒等映射.

Trace 条件: 设 X 和 Y 分别取 $\mathbb{R}^d$ 和 $\mathbb{R}^e$ 中的值, 满足线性模型 $Y = AX + N_X, N_X \perp\!\!\!\perp X$. 如果协方差矩 Σ_{XX} 和 A 满足 $\tau_e(A\Sigma_{XX}A^T) = \tau_d(\Sigma_{XX})\tau_e(AA^T)$, 我们就说 $P_{X,Y}$ 满足 trace 条件. 其中 $\tau_k(B) := \text{tr}(B)/k$ 是归一化迹.

Trace 条件的可识别性: 变量 X, Y 都是 d 维的, 并且 $Y = AX, |A| \neq 0$. 如果从 X 到 Y 满足 trace 条件, 则反向模型为 $X = A^{-1}Y$ 满足 $\tau_d(A^{-1}\Sigma_{YY}A^{-T}) \leq \tau_d(\Sigma_{YY})\tau_d(A^{-1}A^{-T})$, 当且仅当 A 的所有奇异值均具有相同的绝对值时等式成立. 这说明一般情况下反方向的 trace 条件将被违反, 且具有相同的符号.

4.2 结构识别方法

ANM: 残差的独立性: X 上的回归 $Y = \hat{f}_Y(X) + R_Y$, 检验 R_Y 是否独立与 X, 重复交换 X 和 Y 的过程, 如果独立性被一个方向接受而被另一个拒绝, 则推断前者为因果方向.

ANM: 极大似然法: 考虑带有加性高斯误差项的非线性结构因果模型, 比较两个模型的似然分数来区分 $X \to Y$ 和 $Y \to X$. 首先从 Y 对 X 进行非线性回归, 以获得残差 R_Y, 然后比较 $L_{X\to Y} = -\log\widehat{\mathrm{var}}(X) - \log\widehat{\mathrm{var}}(R_Y)$ 和类似版本 $L_{X\leftarrow Y} = -\log\widehat{\mathrm{var}}(R_X) - \log\widehat{\mathrm{var}}(Y)$, 因果关系倾向于分高者. 如果噪声没必要服从高斯分布, 我们必须调整分数函数, 利用误差项的微分熵的估计值取代残差经验方差的对数.

IGCI: 独立性条件 $\mathrm{cov}(\log f', p_X) = 0$ 意味着 $C_{X\to Y} \le C_{Y\to X}$, 其中 $C_{X\to Y} := \int_0^1 \log f'(x)p(x)dx$. 在这里, 使用简单的估计量 $\hat{C}_{X\to Y} = \frac{1}{N-1}\sum_{j=1}^{N-1}\log\frac{|y_{j+1}-y_j|}{|x_{j+1}-x_j|}$, 其中 $x_1 < x_2 < \cdots < x_N$ 是观察到的 x 值, 按升序排列. 如果 $\hat{C}_{X\to Y} < \hat{C}_{Y\to X}$ 可推断出 $X \to Y$.

IGCL: 独立性条件 $\mathrm{cov}(\log f', p_X) = 0$ 意味着 $H(X) \le H(Y)$, 其中 H 表示微分香农熵 $H(X) := -\int_0^1 p(x)\log p(x)dx$.

Trace 方法: 计算跟踪相关率 $r_{X\to Y} := \frac{\tau(A_Y\Sigma_{XX}A_Y^T)}{\tau(A_YA_Y^T)\tau(\Sigma_{XX})}$ 和 $r_{Y\to X}$, 接近于 1 的那个对应因果方向.

5 与机器学习的联系 1

5.1 半监督学习

独立同分布数据点 $(X_1,Y_1),\cdots,(X_n,Y_n) \sim P_{X,Y}$, 未标记数据点 $X_{n+1},\cdots,X_{n+m} \sim P_X$.

机器学习问题: 因果关系 (从因预测果), 反因果关系 (从果预测因).

所有对半监督学习有帮助的案例都是反因果性的、混乱的或者因果结构不清楚的例子. 额外的 x 值只能告诉人们更多关于 P_X 的信息, 而预测需要关于独立对象 $P_{Y|X}$ 的信息.

然而, 即使 P_X 没有告诉人们关于 $P_{Y|X}$ 的任何信息, 了解 P_X 仍然可以帮助人们更好地估计 Y, 因为在学习场景中获得更小的风险值.

5.2 协变量偏移

协变量移位: 即使 P_X 已更改, 仍使用相同的条件 $P_{Y|X}$.

在反因果场景中, P_Y 和 $P_{X|Y}$ 的变化可能是相关的, 这可能是由于 P_X 发生了变化而 $P_{Y|X}$ 保持不变, 换句话说, P_{effect} 和 $P_{\text{cause|effect}}$ 的改变存在相关性, 因为 P_{cause} 和 $P_{\text{effect|cause}}$ 的改变是相互独立的.

6 多变量因果模型

6.1 图的术语

多个随机变量 $X = (X_1,\cdots,X_d)$, 索引集 $V = (1,\cdots,d)$, 联合分布 P_X, 密度 $p(x)$. 一个图 $G = (V,E)$ 由节点 V 和边 $E \subset V^2$ 组成且 $(v,v) \notin V^2$.

如果 $(i,j) \in E$ 但是 $(j,i) \notin E$, 则 i 是 j 的父亲, j 是 i 的孩子. j 的父节点集合记为 PA_j, 孩子集合记为 CH_j. 如果 $(i,j) \in E$ 或者 $(j,i) \in E$, 则称 i,j 相邻. 若 $(i,j) \in E$ 且 $(j,i) \in E$, 则称无向边. 如果一个节点是另外两个节点的孩子, 而且这两个节点本身不相邻, 则称为 v 结构.

G 中的路径是不同顶点 $i_1,\cdots,i_m$ 的序列, 满足 i_k, i_{k+1} 之间存在边. 若 $i_{k-1} \to i_k$ 且 $i_{k+1} \to i_k$, 则称 i_k 为对撞节点. 若都有 $i_k \to i_{k+1}$, 则称有向路径, 称 i_1 是祖先, i_m 是后代. 祖先记为 AN_i, 后代记为 DE_i, 除去自身的非后代记为 ND_i. 没有孩子的节点称为汇聚节点, 没有祖先的节点称为源节点. 排列 $\pi : \{1,2,\cdots,d\} \to \{1,2,\cdots,d\}$, 如果任意 $j \in DE_i$ 都有 $\pi(i) < \pi(j)$, 则称 π 是拓扑排序/因果排序.

若不同时存在 j 到 k 和 k 到 j 的有向路径, 则称部分有向无环图 (PDAG). 如果所有边有向, 则称有向无环图 (DAG).

d 分离: 在有向无环图 G 中, 节点 i_1 和 i_m 之间的任意路径被一个集合 S 阻塞, 即该路径上存在一个节点 i_k, 满足下面任意一条性质: (1) $i_k \in S$ 且 $i_{k-1} \to i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \to i_{k+1}$; (2) $i_k \notin S$ 及其后代都不在 S 中且 $i_{k-1} \to i_k \leftarrow i_{k+1}$. 类似定义集合 A 和 B 被 S d 分离, 记为 $A \perp\!\!\!\perp_G B|S$.

6.2 结构因果模型

SCM $C := (S, P_N)$ 由 d 个赋值的集合 S 组成: $X_j := f_j(PA_j, N_j), j = 1, 2, \cdots, d$, 其中噪声变量独立. 有时也称 PA_j 是 X_j 的直接原因, X_j 是直接效果.

一个 SCM 在变量 X 上定义了一个唯一的分布满足 $X_j = f_j(PA_j, N_j), j = 1, 2, \cdots, d$, 称它为蕴含分布 P_X^C.

6.3 干预

考虑一个 SCM $C := (S, P_N)$ 和它的蕴含分布 P_X^C, 替换一个或几个结构赋值获得一个新的 SCM $\widetilde{C}$, e.g. $X_k := \widetilde{f}(\widetilde{PA_k}, \widetilde{N_k})$, 将新 SCM 的蕴含分布称为干预分布, 记为 $P_X^{\widetilde{C}} =: P_X^{C;\mathrm{do}(X_k := \widetilde{f}(\widetilde{PA_k}, \widetilde{N_k}))}$.

干预通常不同于条件作用, e.g.$X_1 \to Y \to X_2$, $P_Y^{C;\mathrm{do}(X_2 := X)}(y) = P_Y^C(y) \neq P_Y^C(y|X_2 = x)$.

总因果效应: 给定一个 SCM C, X 到 Y 的总因果效应存在, 当且仅当 $X \not\perp\!\!\!\perp Y$ 在 $P_X^{C;\mathrm{do}(X := \widetilde{N_X})}$ 中对于一些随机变量 $\widetilde{N_X}$.

给定一个 SCM C, 下列语句是等价的: (1) 有一个从 X 到 Y 的总因果效应; (2) 存在 x_1, x_2 使得 $P_Y^{C;\mathrm{do}(X:=x_1)} \neq P_Y^{C;\mathrm{do}(X:=x_2)}$; (3) 存在 x_1 使得 $P_Y^{C;\mathrm{do}(X:=x_1)} \neq P_Y^C$; (4) 在 $P_{X,Y}^{C;\mathrm{do}(X:=\widetilde{N_X})}$ 中, $X \not\perp\!\!\!\perp Y$ 对于任何有完全支撑集的分布 $\widetilde{N_X}$.

总因果效应的图准则: 考虑 SCM C, 其相应图为 G. (1) 如果 X 到 Y 不存在有向路径, 那么没有总因果效应; (2) 有时候有一条有向的路径, 但没有总因果效应 (相互干预抵消).

考虑一个肾结石治愈数据集. 在 700 名患者中, 有一半进行了外科手术 (治疗方案 $T = a, 78\%$ 治愈率), 有一半是经皮肾镜取石术 ($T = b, 83\%$ 治愈率). 人们什么都不知道, 让他们选择, 更多人喜欢治疗方案 b. 但若将肾结石分为小石头和大石头, 我们又发现外科手术表现得更好.

	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

6.4 反事实

考虑一个 SCM $C := (S, P_N)$. 给定一些观测 x, 通过替换噪声变量的分布, 定义一个反事实的 SCM: $C|_{X=x} := (S, P_{N|X=x})$. 新的噪声变量集合不需要联合独立, 反事实陈述可以被看作 do 语句.

有时不能从因果图模型和随机试验、观察数据等方面来预测反事实、区分两个 SCM.

反事实陈述不可传递, 即不能简单引入新变量来解释逻辑.

6.5 马尔可夫性、忠实性和因果最小性

马尔可夫性: 给定一个 DAG G 和一个联合分布 P_X, 这个分布被认为满足: (1) 关于 DAG G 的全局马尔可夫性, 如果 $A \perp\!\!\!\perp_G B|C \Rightarrow A \perp\!\!\!\perp B|C$ 对所有不相交的顶点集合 A, B, C; (2) 关于 DAG G 的局部马尔可夫性, 如果给定它的父节点, 每个变量与它的非后代节点独立; (3) 关于 DAG G 的马尔可夫因式分解性, 如果 $p(x) = p(x_1, \cdots, x_d) = \prod_{j=1}^{d} p(x_j|PA_j)$. 乘积因子称为因果马尔可夫核 (Causal Markov Kernel). 如果 P_X 具有密度函数 p, 则上述定义等价.

图的马尔可夫等价性: 用 $M(G)$ 表示关于 G 满足马尔可夫性的分布的集合: $M(G) = \{P : P$满足关于G的全局马尔可夫性$\}$. 如果 $M(G_1) = M(G_2)$, 则两个 DAG G_1 和 G_2 马尔科夫等价. 当且仅当 G_1 和 G_2 有相同的 d 分离, 这意味着马尔可夫性蕴含了相同的条件独立性.

与某一 DAG 马尔科夫等价的所有 DAG 的集合称为马尔科夫等价类 G, 它可以通过完全的 PDAG 表示, 记为 CPDAG(G)=(V, E). 它包含边 $(i, j) \in E$, 当且仅当马尔科夫等价类中有一个成员包含该边.

马尔可夫等价的图标准: 两个 DAG G_1 和 G_2 是马尔可夫等价, 当且仅当它们具有相同的骨架和相同的 v 结构.

马尔可夫毯: 考虑一个 DAG $G = (V, E)$ 和一个目标节点 Y. Y 的马尔科夫毯是最小的集合 M, 满足 $Y \perp\!\!\!\perp_G V \backslash (\{Y\} \cup M)$. 换句话说, 给定 M, 其他变量不提供 Y 的更多信息.

给定一个 DAG G 和目标节点 Y. 则 Y 的马尔可夫毯 M 包含它的父节点、子节点和子节点的父节点: $M = PA_Y \cup CH_Y \cup PA_{CH_Y}$.

Reichenbach 的共同原因原则: 假设任何一对变量 X, Y 在如下意义下可以嵌入到一个更大的系统中: 在包含 X, Y 的随机变量集合 Z 的图 G 上存在一个正确的 SCM. 如果 X 和 Y 相关, 则必满足以下 3 种因果解释之一: (1) 一个从 X 到 Y 的有向路径; (2) 一个从 Y 到 X 的有向路径; (3) 存在一个节点 Z, 有从 Z 到 X 和 Z 到 Y 的有向路径.

选择偏差: 根据 Reichenbach 原则, 从两个相关的随机变量开始, 得到了一个有效的结论. 然而, 在实际应用中这可能是因为隐含了第三个变量 (选择偏差), 这可能导致 X 和 Y 之间的依赖关系, 尽管这三个条件都不成立. 例如 $X \to Z \leftarrow Y \Rightarrow X \perp\!\!\!\perp Y$ but $X \not\perp\!\!\!\perp Y|Z = z$.

SCM 隐含马尔可夫性: 假定 P_X 是由图 G 的 SCM 产生的, 则 P_X 关于图 G 具有马尔可夫性.

因果图模型: 随机变量 $X = (X_1, \cdots, X_d)$ 上的因果图模型包含一个图 G 和一个函数集 $f_j(x_j, x_{PA_j})$, 该函数的积分为 1: $\int f_j(x_j, x_{PA_j}) dx_j = 1$. 这些函数在 X 上产生了一个分布 $P_X : p(x_1, \cdots, x_d) = \prod_{j=1}^{d} f_j(x_j, x_{PA_j})$, 因此扮演了条件分布的角色: $f_j(x_j, x_{PA_j}) = p(x_j|x_{PA_j})$. 因果图模型可产生干预分布: $p^{\mathrm{do}\left(X_k := q(\cdot|x_{\widehat{PA_k}})\right)}(x_1, \cdots, x_d) = \prod_{j \neq k} f_j(x_j, x_{PA_j}) q(\cdot|x_{\widehat{PA_k}})$. 其中, $q(\cdot|x_{\widehat{PA_k}})$ 积分为 1, 且新的父节点不能产生一个循环.

忠实性和因果最小性 (Faithfulness and causal minimality): 考虑分布 P_X 和 DAG G.
(1) P_X 对于 DAG G 具有忠实性, 如果 $A \perp\!\!\!\perp B|C \Rightarrow A \perp\!\!\!\perp_G B|C$ 对于所有不相交节点集合 A, B, C;
(2) 一个分布关于 G 满足因果最小性, 如果它关于图 G 具有马尔可夫性, 但不适用于 G 的任何子图.

忠实性隐含因果最小性: 如果 P_X 关于 G 具有马尔可夫性和忠实性, 那么它具有因果最小性.

因果最小的等价性: 考虑随机矢量 $X = (X_1, \cdots, X_d)$, 假定联合分布关于乘积测度有一个密度. 假定 P_X 关于 G 具有马尔可夫性. 则 P_X 关于 G 满足因果最小性, 当且仅当如果 $\forall X_j, \forall Y \in PA_j$, 有 $X_j \not\perp\!\!\!\perp Y|PA_j \backslash \{Y\}$.

因果最小性可以解释为在描述干预模型时避免冗余的惯例. 没有因果最小性, 观测数据的可识别性是不可能存在的.

6.6 通过协变量调整计算干预分布

给定一个 SCM C, 对于任何由 C 通过干预 (一些)X_k 而不是 X_j 构建的 SCM $\widetilde{C}$, 有 $p^{\widetilde{C}}(x_j|x_{PA_j}) = p^C(x_j|x_{PA_j})$. 这表明在干预下, 因果关系是自治的.

考虑一个带有结构赋值的 SCM $C : X_j := f_j(X_{PA_j}, N_j), j = 1, 2, \cdots, d$ 和密度 p^C. 根据马尔可夫性, 有 $p^C(x_1, \cdots, x_d) = \prod_{j=1}^d p^C(x_j|x_{PA_j})$. 现在考虑执行 $(X_k := \widetilde{N_k})$ 后得到的 SCM $\widetilde{C}$, 其中 $\widetilde{N_k}$ 考虑密度 $\widetilde{p}$. 根据马尔可夫假设, $p^{C;\mathrm{do}(X_k:=\widetilde{N_k})}(x_1, \cdots, x_d) = \prod_{j\neq k} p^{C;\mathrm{do}(X_k:=\widetilde{N_k})}(x_j|x_{PA_j}) \cdot p^{C;\mathrm{do}(X_k:=\widetilde{N_k})(x_k)} = \prod_{j\neq k} p^C(x_j|x_{PA_j})\widetilde{p}(x_k)$.

有效调整集 (Valid adjustment set): 考虑节点集合 V 上的一个 SCM C, 并设 $Y \notin PA_x$ (否则, 有 $p^{C;\mathrm{do}(X:=x)}(y) = p^C(y)$). 称集合 $Z \subset V\backslash\{X, Y\}$ 为有序对 (X, Y) 的一个有效调整集, 如果 $p^{C;\mathrm{do}(X:=x)}(y) = \sum_z p^C(y|x, z)p^C(z)$.

混淆 (Confounding): 考虑节点集合 V 上的一个 SCM C, 有一条从 X 到 Y 的有向路径, $X, Y \in V$. 从 X 到 Y 的因果效应是混淆的, 如果满足 $p^{C;\mathrm{do}(X:=x)}(y) \neq p^C(y|x)$. 否则, 称因果效应是不混淆的.

如果有 $p^{C;\mathrm{do}(X:=x)}(y|x, z) = p^C(y|x, z)$ 和 $p^{C;\mathrm{do}(X:=x)}(z) = p^C(z)$, 则称 Z 是一个有效的调整集.

考虑 SCM 上的变量 V, 且 $X, Y \in V, Y \notin PA_X$. 以下三条结论成立:

(1) "父亲调整": $Z := PA_X$ 是 (X, Y) 的有效调整集;

(2) "后门调整": 任何 $Z \subset V\backslash\{X, Y\}$, 有 (i) Z 不包含 X 的后代; (ii) Z 阻止通过后门进入 $X(X \leftarrow \cdots)$ 的从 X 到 Y 的所有路径; 则 Z 是 (X, Y) 的有效调整集;

(3) "通向必要性": 任何 $Z \subset V\backslash\{X, Y\}$, 有 (i) Z 不包含任何从 X 到 Y 的有向路径上节点的后代; (ii) Z 阻止从 X 到 Y 的所有非有向路径; 则 Z 是 (X, Y) 的有效调整集.

6.7 do-calculus

如果一个干预分布可以通过观测分布和图来计算, 则称它是可识别的.

给定一个图 G 和不相交的子集 X, Y, Z, W, 有:

(1) "观测值的插入/删除": 如果在一个 X 的入边已经被删除的图中有 X, W d 分离 Y, Z, 则成立 $p^{C;\mathrm{do}(X:=x)}(y|z, w) = p^{C;\mathrm{do}(X:=x)}(y|w)$;

(2) "干预/观测的交换": 如果在一个 X 的入边已经被删除且 Z 的出边已经被删除的图中有 X, W d 分离 Y, Z, 则成立 $p^{C;\mathrm{do}(X:=x,Z:=z)}(y|w) = p^{C;\mathrm{do}(X:=x)}(y|z, w)$.

(3) "干预的插入/删除": 如果在一个 $X, Z(W)$ 的入边已经被删除的图中有 X, W d 分离 Y, Z, 则成立 $p^{C;\mathrm{do}(X:=x,Z:=z)}(y|w) = p^{C;\mathrm{do}(X:=x)}(y|w)$. 这里 $Z(W)$ 是在删除 X 的所有入边后, 不是 W 中任何节点祖先的 Z 中节点构成的集合.

6.8 因果模型的等价性和可证伪性

称两个模型 (概率/干预/反事实) 等价, 如果它们蕴含相同的 (观测分布/观测分布和干预分布/观测分布, 干预分布和反事实分布).

假设两个 SCM C_1, C_2 产生严格正的连续条件密度 $p(x_j|x_{PA_j})$, 并满足因果最小性. 进一步假设当某个变量 X_j 被设为具有完全支撑集的变量 $\widetilde{N_j}$ 时, 它们蕴含相同的干预分布: $p_X^{C_1;\mathrm{do}(X_j=\widetilde{N_j})} = p_X^{C_2;\mathrm{do}(X_j:=\widetilde{N_j})}, \forall j, \forall\widetilde{N_j}$. 那么 C_1, C_2 干预等价, 也就是说它们在任何可能的干预上一致.

考虑两个有相同噪声分布 P_N 且仅在第 k 个结构赋值不同的 SCM $C, C^* : f_k(PA_k, n_k) = f_k^*(PA_k^*, n_k), \forall PA_k, \forall n_k \text{ with } p(n_k) > 0, PA_k^* \subset PA_k$. 那么, 这两个 SCM 是反事实等价的.

6.9 潜在结果

$T=1/0$ 表示是否接受治疗, $B=1/0$ 表示是否失明, 用 $B_u(t=1/0)$ 表示. 如果 $B_u(t=1)=0$ 且 $B_u(t=0)=1$, 那么认为治疗对个体 u 有积极影响.

因果推理的根本问题: 对于每个个体可以观察到 $B_u(t=1)$ 或 $B_u(t=0)$, 而不可能同时观察到它们. 未观察到的结果成为反事实.

稳定个体治疗价值假设 (SUTVA): 个体之间不会产生干预, 潜在结果不取决于接受治疗的方式或原因.

平均因果效应: $CE=\frac{1}{n}\sum_{u=1}^{n}(B_u(t=1)-B_u(t=0))$.

在一个完全随机的实验中, $\widehat{CE}=\frac{1}{\#U_0}\sum_{u\in U_0}B_u(t=1)-\frac{1}{\#U_1}\sum_{u\in U_1}B_u(t=0)$ 是无偏估计量.

7 学习多变量因果模型

图结构的非唯一性: 考虑随机变量 $X=(X_1,\cdots,X_d)$ 上的分布 P_X, 它关于 Lebesgue 测度有一个密度, 且关于 G 是马尔可夫的. 则存在一个关于图 G 的 SCM $C=(S,P_N)$, 蕴含分布 P_X.

7.1 结构可识别性

马尔可夫等价类的可识别性: 假定 P_X 关于 G^0 是马尔科夫的、忠实的, 则对于每个图 $G\in \mathrm{CPDAG}(G^0)$, 可以找到一个蕴含 P_X 的 SCM. 并且对于任何 $G\notin \mathrm{CPDAG}(G^0)$, P_X 关于 G 不具有马尔可夫性和忠实性.

因果最小性和 ANM: 考虑加性噪声模型, 并假设函数 f_j 在任何参数上都不是常数. 那么联合分布关于相应的图满足因果最小性.

线性高斯模型的可识别性: 考虑一个图 G_0 的 SCM 及其赋值: $X_j:=\sum_{k\in PA_j}\beta_{jk}X_k+N_j, j=1,\cdots,d$. 所有的 N_j 都是 i.i.d. 的高斯分布, 噪声方差 σ^2 不依赖于 j. 一般地, 设 $\beta_{jk}\neq 0,\forall j\in\{1,2,\cdots,p\},\forall k\in PA_j$. 则图 G_0 可以从联合分布中识别出来.

线性非高斯模型的可识别性: 考虑一个图 G_0 的 SCM 及其赋值: $X_j:=\sum_{k\in PA_j}\beta_{jk}X_k+N_j, j=1,\cdots,d$. 所有的 N_j 都是联合独立的非高斯分布, 且具有严格的正密度. 一般地, 设 $\beta_{jk}\neq 0,\forall j\in\{1,2,\cdots,p\},\forall k\in PA_j$. 则图 G_0 可以从联合分布中识别出来.

非线性高斯加性噪声模型的可识别性: 假定 $P_X=P_{X_1,\cdots,X_d}$ 由如下 SCM 产生: $X_j:=f_j(PA_j)+N_j$. 其中噪声变量服从正态分布: $N_j\sim\mathcal{N}(0,\sigma_j^2)$. 三阶可导函数 f_j 对于每个变元都是非线性的. 则图 G_0 可以从联合分布中识别出来.

7.2 结构识别方法

引理: (1) 在 $\mathrm{DAG}(X,E)$ 中的两个节点, 当且仅当它们不能被任何一个子集 $S\subset V\backslash\{X,Y\}$ d 分离时, 它们是相邻的;

(2) 如果在 $\mathrm{DAG}(X,E)$ 中的两个节点 X,Y 不相邻, 则它们由 PA_X 或 PA_Y d 分离.

代表算法: IC 算法和 SGS 算法: 搜索所有可能的不包含 X,Y 的变量集合 $A\subset V\backslash\{X,Y\}$, 检查给定 A, X,Y 是否 d 分离; PC 算法: 从一个完全连通的无向图开始, 逐步增大条件集合 A 的大小, 从 $\#A=0$ 开始, 在迭代 k 中, 它考虑集合 A, 其大小为 $\#A=k$, 使用以下技巧: 检测 X,Y 是否可以被 d 分离, 只需通过集合 A, 这里 A 是 X 的邻接点或 Y 的临界点的子集.

边的方向: Meek 方向规则.

可满足性方法: 将图关系形式转化为布尔变量, 将独立性语句转化为包含布尔变量和操作 “and” 与 “or” 的公式. 然后, SAT 问题询问是否我们能为每个布尔变量赋值 “true” 还是 “false”, 使整个公式为真. SAT 求解器不仅检查是否属于这种情况, 而且还向人们提供是否整个公式都为真的赋值信息. 尽管布尔 SAT 问题是 NP 完全问题, 但有启发式算法可以解决设计数百万个变量大问题的实例.

条件独立性测试: 统计显著性检验, 基于核的测试.

偏相关的自然非线性拓展: (1) (非线性) 在 Z 上回归 X 并测试残差是否独立于 Y; (2) (非线性) 在 Z 上回归 Y 并测试残差是否独立于 X; (3) 如果上述独立性有一个满足, 可以推断 $X \perp\!\!\!\perp Y|Z$.

最佳评分图: $\hat{G} := \text{argmax}_G S(D, G)$.

惩罚似然分数: $S(D, G) = \log p(D|\hat{\theta}, G) - \frac{\#\text{参数}}{2} \log n$. 贝叶斯评分函数: $S(D, G) = \log p(G|D)$.

贪婪搜索方法: 每一步有一个候选图和一组相邻图, 计算得分并将最佳评分图作为新的候选图. 如果没有相邻图获得更好的分数, 则搜索过程终止.

动态规划: 基于分数可分解性, $\log p(D|\hat{\theta}, G) = \sum_{j=1}^{d} \sum_{i=1}^{n} \log p(X_j^i | X_{PA_j}^i, \hat{\theta}), D = (X^1, \cdots, X^n)$.

整数线性规划 (Integer Linear Programming, ILP): 不仅基于分数可分解性, 也假设评分函数对马尔可夫等价图给出相同的分数. 将图结构表示为向量, 这样评分函数在这个向量表示中变成仿射函数 (最高次数为 1 的多项式函数). 是 NP 问题, 但可以使用 ILP 的现成方法, 比如说限制父节点的数量.

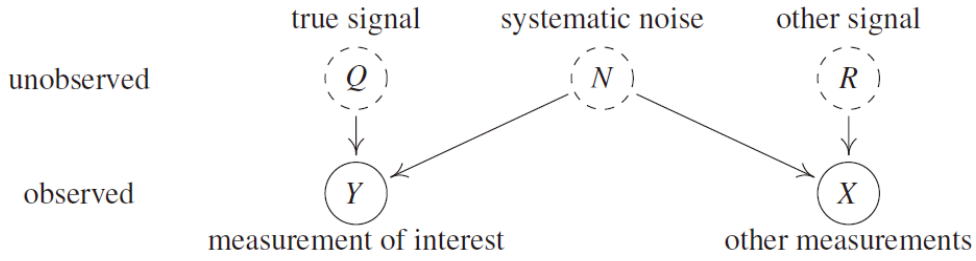
加性噪声模型: $\log p(D|G) = \sum_{j=1}^{d} -\log \widehat{\text{var}}(R_j)$, 其中 $\widehat{\text{var}}(R_j)$ 是残差 R_j 的经验方差.

8 与机器学习的联系 2

8.1 半同胞回归

利用给定的因果结构, 来减少预测任务中系统噪声, 目标是重建未观察到的信号 Q.

可以通过去除所有的能被其他测量 X 解释的信息来去噪信号 Y, 或者说 Y 中所有可以由 X 解释的都必定是由系统噪声 N 引起的, 考虑 $\hat{Q} := Y - \mathbb{E}(Y|X)$ 作为 Q 的估计.



对于满足 $Q \perp\!\!\!\perp X$ 的任意随机变量 Q, X, Y, 成立 $\mathbb{E}(Q - \hat{Q})^2 \leq \mathbb{E}(Q - Y)^2$. 若系统噪声是加性噪声, 即 $Y = Q + f(N)$, 则 $\mathbb{E}(Q - \hat{Q})^2 = \mathbb{E}[\text{var}(f(N)|X)]$.

8.2 因果推断与场景强化学习

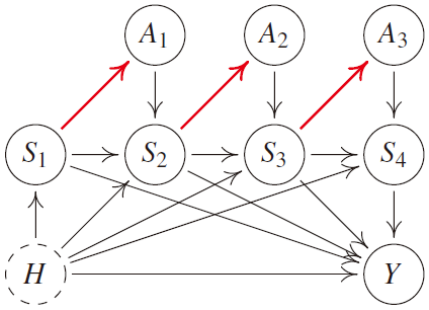
逆概率加权: 从观测分布 P_X^C 中观察到一个样本, 但对干预分布 $P_X^{\widetilde{C}}$ 感兴趣. 这里, 新的 SCM $\widetilde{C}$ 是通过在原来的 C 上对节点 X_k 进行干预来构造的, 即 $\text{do}(X_k := \widetilde{f}(X_{\widetilde{PA_k}}, \widetilde{N_k}))$. 特别是, 可能需要估计新分布 $P_X^{\widetilde{C}}$ 的某一属性: $\widetilde{\mathbb{E}} l(X) := \mathbb{E}_{P_X^{\widetilde{C}}} l(X)$. 如果密度存在, 可以因式分解为 $p(x_1, \cdots, x_d) := p^C(x_1, \cdots, x_d) = \prod_{j=1}^{d} p^C(x_j | x_{PA_j}), \widetilde{p}(x_1, \cdots, x_d) := p^{\widetilde{C}}(x_1, \cdots, x_d) = \prod_{j \neq k} p^C(x_j | x_{PA_j}) \widetilde{p}(x_k | x_{\widetilde{PA_k}})$. 除干预变量项分解一致, 因此 $\xi := \widetilde{\mathbb{E}} l(X) = \int l(x) \widetilde{p}(x) dx = \int l(x) \frac{\widetilde{p}(x)}{p(x)} p(x) dx = \int l(x) \frac{\widetilde{p}(x_k | x_{\widetilde{PA_k}})}{p(x_k | x_{PA_k})} p(x) dx$.

于是, 给定来自分布 p_X^C 的样本 $X^1,\cdots,X^n$, 可以构造估计量 $\hat{\xi}_n := \frac{1}{n}\sum_{i=1}^n l(X^i)\frac{\widetilde{p}(X_k^i|X^i_{PA_k})}{p(X_k^i|X^i_{PA_k})} = \frac{1}{n}\sum_{i=1}^n l(X^i)w_i$. 这里, 权重 w_i 被定义为条件密度的比率. 在 $p_X^{\widetilde{C}}$ 下有很高可能性的数据点获得了较大的权重, 对 $\hat{\xi}_n$ 的估计贡献比小权重的贡献更大.

(1) 假设 $X=(Y,Z)$ 只包含一个目标变量 Y 和因果协变量 Z, 即 $Z\to Y$, 考虑 Z 中的一种干预以及函数 $l(X)=l(Z,Y)=Y$, 则估计式简化为 $\hat{\xi}_n := \frac{1}{n}\sum_{i=1}^n Y^i\frac{\widetilde{p}(Z^i)}{p(Z^i)}$, 称为 Horvitz-Thompson 估计量;

(2) 假设 $X=Z$, 可以利用从 p 中采样的数据估计 $\widetilde{p}$ 下的期望 $\widetilde{\mathbb{E}}(l(Z))$, 则估计式简化为 $\hat{\xi}_n := \frac{1}{n}\sum_{i=1}^n Z^i\frac{\widetilde{p}(Z^i)}{p(Z^i)}$, 称为重要性采样.

场景强化学习: 世界状态 S_t 和行动 A_t, 世界状态根据马尔可夫决策过程变化, 也就是说进入一个新状态的概率 $P(S_{t+1}=s)$ 仅取决于当前状态 S_t 和行动 A_t. 此外将获得一些奖赏 R_{t+1}, 它依赖于 S_t,A_t 和 S_{t+1}. 所有奖赏的总和被称为回报 $Y:=\sum_t R_t$. 回报 Y 依赖于状态, 行动的方式对于试图改进策略 $(a,s)\mapsto\pi(a|s):=P(A_t=a|S_t=s)$ 的智能体是未知的, 也就是说选择行动的条件依赖于世界状态的观察部分. 状态经过有限的动作后被重置.



上图是一个场景强化学习问题. 这个回报 Y 可能依赖于动作 (为清晰而省略边). 它通常被建模为在每个决策之后收到的 (可能加权的) 奖赏总和. 整个系统可以被一个未观察到的变量混淆. 红粗线表示玩家可以影响的条件, 也就是策略.

在一定策略 $(a,s)\mapsto\pi(a|s)$ 下玩 n 个游戏, 每个游戏都是一个场景. 这个函数 π 不依赖于迄今使用的 "移动" 次数, 而仅仅取决于状态值. 只要该策略对任何动作指派正概率, 可以估计不同策略的性能: $\hat{\xi}_n = \frac{1}{n}\sum_{i=1}^n Y^i\frac{\prod_{j=1}^K\widetilde{\pi}(A_j^i|S_j^i)}{\prod_{j=1}^K\pi(A_j^i|S_j^i)}$. 这个估计量往往具有较大方差, 在连续环境中可能不收敛. 可以重新加权或者忽略 5 个最大权重, 以抵消偏差.

9 隐藏变量

9.1 干预充分性

因果充分性: 称变量集 X 因果充分, 如果没有隐藏的共同原因 $C\notin X$, 它是 X 中不止一个变量的原因.

共同原因: 有一条从不包括 X,Y, 分别从 C 到 X,Y 的有向路径, 则称 C 是 X,Y 的共同原因.

干预充分性: 称变量集 X 具有干预充分性, 如果存在一个 X 上的 SCM, 它不能篡改成一个干预模型. 也就是说, 它产生的观察和干预分布与人们在实践中观察到的一致.

干预充分和因果充分: 假设 C 是变量 X 上的 SCM, 不能被篡改为干预模型. (1) 如果一个子集 $O\subset X$ 因果充分, 则干预充分; (2) 反过来通常不成立.

有以下 3 个结论:

(1) 假定存在 X, Y, Z 上的 SCM, 图结构为 $X \to Y \to Z$, 且 $X \perp\!\!\!\perp Z$, 产生正确的干预. 由于 X, Z 上的 SCM 满足 $X \to Z$, X, Z 干预充分;

(2) 假定存在 X, Y, Z 上的 SCM, 图结构为 $X \to Y \to Z$, 产生正确的干预, 附加地 $X \to Z$, 进一步假定 $P^C_{X,Y,Z}$ 关于图是忠实的. 同样地, 由于 X, Z 上的 SCM 满足 $X \to Z$, X, Z 干预充分;

(3) 条件和 (2) 类似, 不同之处为 $P^C_{Z|X=x} = P_Z^{\mathrm{do}(X:=x)} = P^C_Z, \forall x$, 则由于 X, Z 上的 SCM 是空图, X, Z 干预充分. 注意反事实可能没有正确表示.

9.2 Simpson 悖论

在章节 6.3 中我们看到了共同原因 (肾结石大小) 的确实可能导致颠倒的因果断言. 有没有可能存在另一个没有纠正的混杂变量呢? 原则上, 这可能导致一个任意长的反向因果序列. 这表明写下潜在的因果图时要多么小心.

9.3 工具变量

考虑 $Y := \alpha X + \delta H + N_Y$. 称 SCM 中的变量 Z 为 (X, Y) 的工具变量, 如果 (1)Z 独立于 H; (2) Z 不独立于 X; (3) Z 仅通过 X 影响 Y.

考虑 $X := \beta Z + \gamma H + N_X$. 由于 (H, N_X) 和 Z 独立, 所以将 $\gamma H + N_X$ 作为噪声, 则 $Y = \alpha(\beta Z) + (\alpha\gamma + \delta)H + N_Y$. 可以通过在 βZ 上回归 Y 来估计 α. 总的来说, 现在 Z 上回归 X, 然后在预测值 $\hat{\beta} Z$ 上回归 Y. 这种方法称为两阶段最小二乘 (two-stage least squares).