hw2

谭昊童 231300039@smail.nju.edu.cn

2025 年 12 月 2 日

1 (25 points) 逆归结核心规则应用

逆归结是一种从已知事实（证据）和背景知识反向推导出有效假设的逻辑推理方法。在逆归结中，核心操作包括四种规则：吸收（Absorption）、辨识（Identification）、内构（Intra-construction）和互构（Inter-construction）（提示：详细规则可参考机器学习（西瓜书）第 361 页）。现在考虑如下情境，Tom 报名参加了市级举重大赛。与比赛结果相关的部分规则和事实如下：

- R1：若一个人“强壮”且“有技巧”，则他会“获胜”（记为：获胜 $\leftarrow$ 强壮 $\wedge$ 有技巧）；
- R2：若一个人“坚持锻炼”，则他会“强壮”（记为：强壮 $\leftarrow$ 坚持锻炼）；
- R3：若一个人“强壮”且“对手弱”，则他会“获胜”（记为：获胜 $\leftarrow$ 强壮 $\wedge$ 对手弱）；
- R4：一个人不能同时“有技巧”和“运气好”（记为：$\neg$ 有技巧 $\vee \neg$ 运气好）；

已知事实：Tom 最终在举重大赛中获胜了。

1. 假如大赛前 Tom 因意外受伤, 没有坚持锻炼。结合上述背景知识, 你能否判断 Tom 是否会取得胜利? 请详细说明推理过程及依据。(8 分)

2. 有观众猜测, Tom 获胜是因为存在一条未公开的“神秘规则”(形式为获胜 $\leftarrow$ 强壮 $\wedge$ X, X 是某个未提及的条件)。结合背景知识, 解释为什么“X 是运气好”不符合逻辑约束。(8 分)

3. 教练复盘比赛时, 仔细观察 R1 (获胜 $\leftarrow$ 强壮 $\wedge$ 有技巧) 和 R3 (获胜 $\leftarrow$ 强壮 $\wedge$ 对手弱) 后提出: 两条规则中“获胜”的核心前提“强壮”是基础, 但真正决定胜负的是一种未被明确的“神秘制胜条件”(记为 S)。该条件是“有技巧”和“对手弱”的共性本质, 且与“强壮”存在紧密逻辑关联。请明确 S 与“强壮”“有技巧”“对手弱”之间的逻辑关系 (9 分)

Solution. 1. 不能。已知 $\neg$ 坚持锻炼，但是仅由 R2 我们无法推出 Tom 是否强壮，而且缺少关于“有技巧”和“对手弱”的信息，因此无法确定 Tom 是否获胜。

2. 假设 X 是运气好，则神秘规则为获胜 $\leftarrow$ 强壮 $\wedge$ 运气好。已知获胜 $\leftarrow$ 强壮 $\wedge$ 有技巧，由辨识规则可以得到

$$\text{有技巧} \leftarrow \text{运气好} \quad \text{获胜} \leftarrow \text{强壮} \wedge \text{有技巧}$$

但是 R4 可以转化为

$$\neg\text{有技巧} \leftarrow \text{运气好}$$

矛盾，因此 X 不能是运气好。

3. 由内构规则可以得到

$$S \leftarrow \text{有技巧} \quad \text{获胜} \leftarrow \text{强壮} \wedge S \quad S \leftarrow \text{对手弱}$$

关系：S 是“有技巧”和“对手弱”的共性本质，“有技巧”和“对手弱”都能推出 S.S 和强壮构成获胜的充分条件。

□

2 (15 points) FOIL 算法信息增益计算应用题

小明正在开发一套家族关系智能识别系统，核心需求是让系统自动归纳“祖父（Grandfather, x, y）”的判定规则——即通过已知的“父亲”“男性”等基础关系，让系统自主学习“x 是 y 的祖父”的逻辑条件。为实现这一目标，小明收集了相关数据并设计了候选规则，现需使用 FOIL 算法（一阶归纳学习算法）通过计算信息增益筛选最优规则。目标概念为需归纳的谓词 Grandfather(x, y)，语义为“x 是 y 的祖父”；系统已知基础关系谓词包括 Father(a, b)（表示“a 是 b 的父亲”）和 Male(a)（表示“a 是男性”）。

小明标注的训练样本中，正例（即 Grandfather(x, y) = True，x 确实是 y 的祖父）共 3 个，分别为 P_1(Peter,Tom)、P_2(John,Lily)、P_3(Mike, Jack)；反例（即 Grandfather(x, y) = False，x 不是 y 的祖父）共 3 个，分别为 N_1(Peter,Bob)、N_2(John,Jack)、N_3(Mike, Lily)。

一阶规则学习的初始规则为“最一般规则”（无任何前提条件）：Grandfather(x, y) ← True，该规则会覆盖所有训练样本，因此初始正例覆盖数 $p_0 = 3$，初始反例覆盖数 $n_0 = 3$。

父亲关系（Father）：

Father(Peter, David), Father(David, Tom), Father(David, Ann), Father(John, Paul), Father(Paul, Lily), Father(Paul, Bob), Father(Mike, Steve), Father(Steve, Jack), Father(Steve, Lucy)

性别关系（Male）：

Male(Peter), Male(David), Male(John), Male(Paul), Male(Mike), Male(Steve)

小明基于背景谓词设计了 4 条候选规则，具体如下：

1. 规则 A: Grandfather(x, y) ← Father(x, z)
2. 规则 B: Grandfather(x, y) ← Father(z, y)
3. 规则 C: Grandfather(x, y) ← Father(x, z)∧ Father(z, y)
4. 规则 D: Grandfather(x, y) ← Male(x)∧ Father(x, z)∧ Father(z, y)

FOIL 算法通过信息增益衡量候选规则对“区分正例、排除反例”的贡献，信息增益越大，规则性能越优，计算公式为：

$$\text{Gain} = p_1 \times \log_2\left(\frac{p_1}{p_1 + n_1}\right) - p_0 \times \log_2\left(\frac{p_0}{p_0 + n_0}\right)$$

其中：p_0, n_0 为初始规则的正例覆盖数、反例覆盖数（已知 $p_0 = 3, n_0 = 3$）；p_1, n_1 为候选规则的正例覆盖数、反例覆盖数。

1. 结合辅助事实，逐一推导规则 A、B、C、D 的 p_1（覆盖正例数）和 n_1（覆盖反例数）分别计算 4 条规则的 FOIL 信息增益（8 分）

2. 比较 4 条规则的信息增益大小，指出信息增益最大的规则，并结合规则简洁性说明小明应选择的最终最优规则及理由（7 分）

Solution. 1. (A) $p_1 = 3, n_1 = 3, Gain = 0$

命题	A 规则值	B 规则值	C 规则值	D 规则值
Grandfather(Peter,Tom)	True	True	True	True
Grandfather(John,Lily)	True	True	True	True
Grandfather(Mike, Jack)	True	True	True	True
Grandfather(Peter,Bob)	True	True	False	False
Grandfather(John,Jack)	True	True	False	False
Grandfather(Mike, Lily)	True	True	False	False

(B) $p_1 = 3, n_1 = 3, Gain = 0$

(C) $p_1 = 3, n_1 = 0, Gain = 3 \times \log_2(1) - 3 \times \log_2(\frac{3}{6}) = 3$

(D) $p_1 = 3, n_1 = 0, Gain = 3 \times \log_2(1) - 3 \times \log_2(\frac{3}{6}) = 3$

2. 规则 C 和规则 D 信息增益最大，均为 3。规则 C 较 D 来说少一条对性别的约束，更为简洁。虽然在现实生活中 male 是一个 grandfather 的合理约束，但在本题的样本中属于冗余信息。所以应该选择 C 作为最优规则。

□

3 (20 points) 马尔可夫决策过程与 Bellman 方程

给定一个 MDP 四元组 $E = \langle \mathcal{S}, \mathcal{A}, P, R \rangle$, 其中 $\gamma \in [0, 1)$ 为折扣因子, $R^a_{s \to s'}$ 表示在状态 s 执行动作 a 转移到 s' 时的期望奖赏。

1. 请写出最优状态值函数 $V^*(s)$ 和最优状态-动作值函数 $Q^*(s, a)$ 之间的相互转换关系式（即最优 Bellman 方程）。（10 分）

2. 策略改进定理指出，如果我们将策略选择的动作改变为当前 Q 值最大的动作，策略会得到提升。设新策略为 $\pi'(s) = \arg\max_a Q^\pi(s, a)$，请证明对于任意状态 s，有 $V^{\pi'}(s) \geq V^\pi(s)$。（10 分）

Solution. 1. $V^*(s)$ 是最优状态值函数，即

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$$

最优状态-动作值函数 $Q^*(s, a)$ 是执行动作 a 后的最优期望回报，即

$$Q^*(s, a) = \sum_{s' \in \mathcal{S}} P(s'|s, a) \left[R^a_{s \to s'} + \gamma V^*(s') \right]$$

两式结合可得

$$V^*(s) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'|s, a) \left[R^a_{s \to s'} + \gamma V^*(s') \right]$$

2. 根据新策略的定义 $\pi'(s) = \arg\max_a Q^\pi(s, a)$，对于任意状态 s，我们可以得到：

$$Q^\pi(s, \pi'(s)) = \max_a Q^\pi(s, a) \geq Q^\pi(s, \pi(s)) = V^\pi(s) \tag{1}$$

利用 Q 值的定义展开式 1 可得

$$\begin{aligned} V^\pi(s) &\leq Q^\pi(s, \pi'(s)) \\ &= \mathbb{E}_{\pi'} \left[R_{t+1} + \gamma V^\pi(S_{t+1}) \mid S_t = s \right] \end{aligned}$$

这里 Q^π 是基于旧策略 π 的价值函数，但动作是由新策略 $\pi'(s)$ 选择的.

在上面的期望中，出现了一项 $V^\pi(S_{t+1})$。我们可以对这一项再次应用 (式1)，即 $V^\pi(S_{t+1}) \leq Q^\pi(S_{t+1}, \pi'(S_{t+1}))$。代入上式：

$$\begin{aligned}
V^\pi(s) &\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma V^\pi(S_{t+1}) \mid S_t = s\right] \\
&\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma Q^\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s\right] \\
&\text{(再次展开 Q 值)} \\
&= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma\left(R_{t+2} + \gamma V^\pi(S_{t+2})\right) \mid S_t = s\right] \\
&= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 V^\pi(S_{t+2}) \mid S_t = s\right] \\
&\leq \mathbb{E}_{\pi'}\left[\sum_{i=0}^{k-1} \gamma^i R_{t+i+1} + \gamma^k V^\pi(S_{t+k}) \mid S_t = s\right]
\end{aligned}$$

当 $k \to \infty$ 时，由于 $\gamma \in [0, 1)$，衰减项 $\lim_{k\to\infty} \gamma^k V^\pi(S_{t+k}) = 0$（假设奖励是有界的）。剩下的求和项正是新策略 π' 下的状态值函数 $V^{\pi'}(s)$ 的定义：

$$\begin{aligned}
V^\pi(s) &\leq \lim_{k\to\infty} \mathbb{E}_{\pi'}\left[\sum_{i=0}^{k-1} \gamma^i R_{t+i+1} \mid S_t = s\right] \\
&= V^{\pi'}(s)
\end{aligned}$$

所以对于任意状态 s，都有 $V^{\pi'}(s) \geq V^\pi(s)$。

□

4 (20 points) 蒙特卡洛方法与重要性采样

1. 在蒙特卡洛强化学习中，请简述同策略（On-policy）和异策略（Off-policy）的主要区别。(5 分)

2. 在异策略学习中, 我们需要估计目标策略 π 下的期望回报, 但数据是根据行为策略 b 采样的。请写出利用重要性采样 (Importance Sampling) 比率 $\rho_{t:T-1}$ 来修正回报 G_t 的公式, 并解释为什么需要这样做。(10 分)

3. 普通的重要性采样虽然是无偏的, 但存在什么主要问题? (5 分)

Solution. 1.
- 同策略：用于生成采样数据的策略（行为策略）与我们要评估或改进的策略（目标策略）是同一个策略。
- 异策略：用于生成采样数据的策略（行为策略 b）与我们要评估或改进的策略（目标策略 π）是不同的。

2.

$$\rho_{t:T-1} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

$\pi(A_k|S_k)$ 指 π 策略在状态 S_k 下选择动作 A_k 的概率，$b(A_k|S_k)$ 指 b 策略在状态 S_k 下选择动作 A_k 的概率。

利用重要性采样比率修正回报的公式为：

$$\hat{G}_t = \rho_{t:T-1} G_t$$

原因：我们的目标是计算回报 G_t 在目标策略 π 下的期望值 $\mathbb{E}_\pi[G_t]$。然而，我们手头的数据（轨迹）是根据行为策略 b 采样生成的，实际上我们是在计算 $\mathbb{E}_b[G_t]$。

设一条轨迹为 $\tau = \{S_t, A_t, S_{t+1}, A_{t+1}, \ldots, S_T\}$。$P(\tau|\pi)$ 表示在策略 π 下这条轨迹发生的概率。那么期望的定义公式为（这里用求和符号 $\sum$ 代表离散情况下的积分）：

$$\mathbb{E}_\pi[G_t] = \sum_\tau P(\tau|\pi) G_t(\tau)$$

$$\begin{aligned}
\mathbb{E}_\pi[G_t] &= \sum_\tau P(\tau|\pi) G_t(\tau) \\
&= \sum_\tau P(\tau|\pi) \cdot \frac{P(\tau|b)}{P(\tau|b)} \cdot G_t(\tau) \\
&= \sum_\tau P(\tau|b) \cdot \left(\frac{P(\tau|\pi)}{P(\tau|b)} \right) \cdot G_t(\tau) \\
&= \mathbb{E}_b \left[\frac{P(\tau|\pi)}{P(\tau|b)} G_t \right]
\end{aligned}$$

因为

$$\begin{aligned}
P(\tau|\pi) &= \pi(A_t|S_t) \cdot p(S_{t+1}|S_t, A_t) \cdot \pi(A_{t+1}|S_{t+1}) \cdot p(S_{t+2}|S_{t+1}, A_{t+1}) \ldots \\
P(\tau|b) &= b(A_t|S_t) \cdot p(S_{t+1}|S_t, A_t) \cdot b(A_{t+1}|S_{t+1}) \cdot p(S_{t+2}|S_{t+1}, A_{t+1}) \ldots
\end{aligned}$$

所以

$$\frac{P(\tau|\pi)}{P(\tau|b)} = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)} = \rho_{t:T-1}$$

所以乘上系数 $\rho_{t:T-1}$ 即可修正回报。

3. 方差极大，估计不稳定。

原因：重要性采样比率 $\rho_{t:T-1}$ 是多个概率比值的连乘积（$\prod \frac{\pi}{b}$）。当轨迹很长时，只要分母 $b(A_k|S_k)$ 中出现很小的值（即行为策略探索到了目标策略认为大概率的动作，但行为策略自身产生该动作的概率很低），或者分子远大于分母，这样的事情出现几次，连乘的结果就会呈指数级爆炸; 反之则会接近于 0。

□

5 (20 points) 时序差分学习

Sarsa 与 Q-learning 时序差分学习结合了动态规划和蒙特卡洛方法的优点。

1. 请分别写出 Sarsa 算法和 Q-learning 算法的单步更新公式。(8 分)

2. 基于上述公式, 说明为什么 Q-learning 是异策略 (Off-policy) 算法, 而 Sarsa 是同策略 (On-policy) 算法?

3. 在悬崖行走 (Cliff Walking) 等具有惩罚区域的环境中, Q-learning 和 Sarsa 学到的路径通常有何不同? 为什么? (5 分)

Solution. 1. 设 S_t 为当前状态，A_t 为当前动作，R_{t+1} 为获得的奖励，S_{t+1} 为下一状态，α 为学习率，γ 为折扣因子。

Sarsa 更新公式：

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right]$$

Q-learning 更新公式：

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right]$$

2. 在 Sarsa 的公式中，更新 $Q(S_t, A_t)$ 时用到的 TD 目标是 $R_{t+1} + \gamma Q(S_{t+1}, \mathbf{A_{t+1}})$。这里的 A_{t+1} 是智能体在下一时刻实际将会执行的动作。它是根据当前的行为策略（例如 ϵ-greedy）采样出来的。这意味着，算法在评估的策略，正是智能体当前用来行动的策略。学的就是正在做的，两者一致，所以是同策略。

   在 Q-learning 的公式中，更新 $Q(S_t, A_t)$ 时用到的 TD 目标是 $R_{t+1} + \gamma \max_{\mathbf{a}} \mathbf{Q}(\mathbf{S_{t+1}}, \mathbf{a})$。这里它假设下一时刻会执行最优动作（Greedy Action，即 Q 值最大的动作），而不管智能体实际上下一时刻会采取什么动作。这意味着，算法在评估的策略，是基于当前 Q 值的最优策略，而不是智能体当前实际用来行动的策略。学的和正在做的不一样，所以是异策略。

3. Q-learning：通常会学到一条最优路径（Shortest Path），即贴着悬崖边缘走。这条路最近，步数最少。因为它在更新时使用的是 $\max Q$，它默认自己未来会完全按照最优策略行事（不会犯错）。只要不掉下去，边缘的路就是分数最高的。

   Sarsa：通常会学到一条安全路径（Safe Path），即远离悬崖边缘，稍微绕远一点的路。因为它在更新时使用的是实际动作 A_{t+1}。如果走在悬崖边缘，由于 ϵ-greedy 的存在，A_{t+1} 有一定概率（ϵ）是“跳入悬崖”。这个负回报会被计入当前边缘状态的价值中，从而降低边缘状态的 Q 值，最终学会选择一条远离危险的安全路线。

□