SWITCH估计器阈值推导详解

背景知识

问题设定

在情境赌博机(Contextual Bandits)中，我们想评估一个目标策略π的价值：

$$v^\pi = \mathbb{E}_\pi[r] = \mathbb{E}_{x\sim\lambda}\mathbb{E}_{a\sim\pi(\cdot|x)}\mathbb{E}_{r\sim D(\cdot|a,x)}[r]$$

但我们只有从另一个策略μ收集的数据！

重要性权重

为了纠正策略不匹配，我们定义**重要性权重**：

$$\rho(x,a) = \frac{\pi(a|x)}{\mu(a|x)}$$

SWITCH估计器的直觉

为什么需要SWITCH?

1. **IPS估计器**：无偏但在重要性权重大时方差很高
2. **直接方法(DM)**：方差小但可能有偏
3. **DR估计器**：结合两者，但仍受大权重影响

核心思想：根据重要性权重的大小**切换**使用不同的方法！

SWITCH估计器公式

$$\hat{v}_{\text{SWITCH}} = \frac{1}{n}\sum_{i=1}^{n}[r_i\rho_i 1(\rho_i \le \tau)] + \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\hat{r}(x_i,a)\pi(a|x_i)1(\rho(x_i,a) > \tau)$$

其中：

- **第一项**：当ρ≤τ时用IPS（重要性权重小，可信）
- **第二项**：当ρ>τ时用DM（重要性权重大，改用模型）
- **τ**：阈值参数，是我们要优化的！

阈值τ的理论推导

第1步：分解均方误差(MSE)

根据定理2，SWITCH的MSE上界为：

$$\mathrm{MSE} \leq \frac{2}{n}\left\{\mathbb{E}_\mu\left[(\sigma^2+R_{\max}^2)\rho^2 1(\rho\leq\tau)\right]+\mathbb{E}_\pi\left[R_{\max}^2 1(\rho>\tau)\right]\right\}+\mathbb{E}_\pi\left[\epsilon^2 1(\rho>\tau)\right]$$

这里：

- ε(x,a) = r̂(x,a) - E[r|x,a] 是奖励模型的偏差
- σ^2是奖励的方差
- R_max是奖励的上界

第2步：理解各项含义

分解成三部分：

方差部分（来自IPS）：

$$\mathrm{Var}_{\mathrm{IPS}} = \frac{2}{n}\mathbb{E}_\mu\left[(\sigma^2+R_{\max}^2)\rho^2 1(\rho\leq\tau)\right]$$

- 当τ增大时，这项**增大**（包含更多大权重样本）

方差部分（来自DM）：

$$\mathrm{Var}_{\mathrm{DM}} = \frac{2}{n}\mathbb{E}_\pi\left[R_{\max}^2 1(\rho>\tau)\right]$$

- 当τ增大时，这项**减小**（使用DM的样本变少）

偏差部分：

$$\mathrm{Bias}^2 = \mathbb{E}_\pi\left[\epsilon^2 1(\rho>\tau)\right]$$

- 当τ增大时，这项**减小**（使用不准确模型的次数减少）

第3步：偏差-方差权衡

关键观察：**τ控制着偏差和方差的trade-off**！

τ很小 → 更多用DM → 方差小但偏差可能大
τ很大 → 更多用IPS → 无偏但方差可能很大

自动调参方法

方差估计

对于给定的τ，定义：

$$Y_i(\tau) = r_i \rho_i 1(\rho_i \leq \tau) + \sum_{a \in A} \hat{r}(x_i, a) \pi(a|x_i) 1(\rho(x_i, a) > \tau)$$

$$\bar{Y}(\tau) = \frac{1}{n} \sum_{i=1}^{n} Y_i(\tau)$$

方差估计（样本方差）：

$$\widehat{\mathrm{Var}}_\tau = \frac{1}{n^2} \sum_{i=1}^{n} (Y_i(\tau) - \bar{Y}(\tau))^2$$

偏差上界

由于我们不知道真实的ε，使用保守上界：

$$\widehat{\mathrm{Bias}}_\tau^2 = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\pi [R_{\max} 1(\rho > \tau) | x_i] \right)^2$$

最优阈值

通过最小化估计的MSE来选择τ：

$$\hat{\tau} = \arg\min_\tau \left\{ \widehat{\mathrm{Var}}_\tau + \widehat{\mathrm{Bias}}_\tau^2 \right\}$$

详细推导示例

假设简单场景

假设：

- 两个动作：$A = \{a_1, a_2\}$
- 均匀上下文分布
- $\mu(a_1|x) = 0.8$, $\pi(a_1|x) = 0.5$
- $\mu(a_2|x) = 0.2$, $\pi(a_2|x) = 0.5$

计算重要性权重：

- $\rho(x, a_1) = 0.5/0.8 = 0.625$（小）
- $\rho(x, a_2) = 0.5/0.2 = 2.5$（大！）

选择阈值

如果我们设$\tau = 1$：

- 对a_1：$\rho = 0.625 \leq 1$，使用IPS
- 对a_2：$\rho = 2.5 > 1$，使用DM

这样可以避免在ρ=2.5时的高方差！

关键洞察

1. 为什么这个阈值有效?

Minimax最优性：论文证明了适当选择τ后，SWITCH是minimax最优的：

$$\text{MSE(SWITCH)} \leq C \cdot \frac{\mathbb{E}_\mu[\rho^2\sigma^2] + \mathbb{E}_\mu[\rho^2 R_{\max}^2]}{n}$$

这与IPS和DR的理论下界匹配！

2. 实际意义

- **$\tau = 0$**：退化为DM（全用模型）
- **$\tau \to \infty$**：退化为IPS（全用重要性采样）
- **中间值**：智能混合，获得两者优点

3. 保守估计的原因

论文使用R_max作为偏差上界而不是精确估计，因为：

- 保证了不会过度信任不准确的模型
- 更倾向于使用无偏的IPS部分
- 实践中更稳健

数学技巧总结

技巧1：指示函数分解

$$v^\pi = \mathbb{E}_\pi[r] = \mathbb{E}_\pi[r\mathbb{1}(\rho \le \tau)] + \mathbb{E}_\pi[r\mathbb{1}(\rho > \tau)]$$

第一项用IPS无偏估计，第二项用DM估计。

技巧2：方差分解

$$\mathrm{Var}(X + Y) \le 2\mathrm{Var}(X) + 2\mathrm{Var}(Y)$$

用于分别处理SWITCH的两个部分。

技巧3：条件期望

$$\mathrm{Var}(Z) = \mathbb{E}[\mathrm{Var}(Z|W)] + \mathrm{Var}(\mathbb{E}[Z|W])$$

用于计算SWITCH估计器的方差。

实验验证

论文在UCI数据集上验证：

- SWITCH-DR在大多数情况下优于IPS、DM和DR
- 自动调参(Eq. 9)接近oracle最优τ
- 对噪声奖励更稳健

学习建议

理解顺序

1. 先理解IPS为什么方差大（大权重问题）
2. 理解DM为什么有偏（模型不准确）

3. 理解SWITCH如何结合两者优点
4. 最后理解阈值优化

关键公式

记住这三个：

- 重要性权重：$\rho = \pi/\mu$
- SWITCH公式（两部分）
- 阈值优化：$\min\{\mathrm{Var} + \mathrm{Bias}^2\}$

直觉

大权重 = **不可靠** → **改用模型** **小权重** = **可靠** → **使用数据**