作业 4

本文是谷歌翻译和人工校对的结果。如有题意理解偏差，请参考英文版。

1 离线策略评估和因果推断

在课堂上，我们讨论了 Markov 决策过程 (MDP)、从数据中学习 MDP 的方法，以及从该 MDP 计算最佳策略的方法。然而，在使用该策略之前，我们通常希望对其性能进行评估。在某些设置（例如游戏或模拟）中，您可以直接实施该策略并直接衡量其性能，但在许多情况下（例如医疗保健），实施和评估策略非常昂贵且耗时。

因此，我们需要在不实际实施策略的情况下评估策略的方法。这项任务通常称为离线策略评估 (Off Policy Evaluation) 或因果推断 (Causal Inference)。在这个问题中，我们将探索评估离线策略评估的不同方法，并证明这些估计器的一些属性。

我们讨论的大多数方法都适用于一般的 MDP，但为了解决这个问题，我们将考虑具有单个时间步长的 MDP。我们考虑一个由状态 S、动作 A、奖励函数 $R(s,a)$ 组成的宇宙，其中 s 是状态，a 是动作。一个重要因素是我们的数据集中通常只有 a 的一个子集。例如，每个状态 s 可以代表一个患者，每个动作 a 可以代表我们为该患者开出哪种药物，而 $R(s,a)$ 是他们开出该药物后的寿命。

策略由函数 $\pi_i(s,a) = p(a|s,\pi_i)$ 定义。换句话说，$\pi_i(s,a)$ 是在给定特定状态和策略的情况下执行某个动作的条件概率。

我们得到一个由 $(s,a,R(s,a))$ 元组组成的观察数据集。

令 $p(s)$ 表示该数据集中状态 s 值分布的概率密度函数。令我们的观测数据中的 $\pi_0(s,a) = p(a|s)$。π_0 对应于我们的观测数据中存在的基线策略。回到患者示例，$p(s)$ 表示看到特定患者 s 的概率，而 $\pi_0(s,a)$ 表示患者在观测数据中接受药物治疗的概率。

我们还获得了一个目标策略 $\pi_1(s,a)$，它给出了我们希望评估的最佳策略中的条件概率 $p(a|s)$。需要特别注意的是，尽管这是一个分布，但我们希望评估的许多策略都是确定性的，即给定一个特定状态 s_i，对于单个动作，$p(a|s_i) = 1$，对于其他动作，$p(a|s_i) = 0$。

我们的目标是计算与我们的观测数据相同的总体中 $R(s,a)$ 的预期值，但使用 π_1 而不是 π_0 的策略。换句话说，我们试图计算：

$$\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}R(s,a)$$

关于符号和简化假设的重要说明：

我们还没有在课堂上真正讲过多个变量的期望值，例如 $\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}R(s,a)$。为了回答这个问题，你可以做一个简化的假设，即我们的状态和动作是离散分布。这个多变量的期望值只是表明我们取联合对 (s,a) 的期望值，其中 s 来自 $p(s)$，a 来自 $\pi_1(s,a)$。换句话说，你有一个 $p(s,a)$ 项，它是观察该对的概率，我们可以将该概率分解为 $\mathrm{p}(s)p(a|s) = p(s)\pi_1(s,a)$。用数学符号表示，这可以写成：

$$\begin{aligned}\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}R(s,a) &= \sum_{(s,a)} R(s,a)p(s,a)\\ &= \sum_{(s,a)} R(s,a)p(s)p(a|s)\\ &= \sum_{(s,a)} R(s,a)p(s)\pi_1(s,a)\end{aligned}$$

不幸的是，我们无法直接估计这一点，因为我们只有根据策略 π_0 而不是 π_1 创建的样本。或者这个问题，我们将研究使用我们实际可以估计的 π_0 下的期望来近似这个值的公式。

我们将做出一个额外的假设，即每个动作在观察到的策略 $\pi_0(s,a)$ 中都有非零概率。换句话说，对于所有动作 a 和状态 s，$\pi_0(s,a) > 0$。

回归：最简单的估计器是直接使用我们学到的 MDP 参数来估计我们的目标。这通常称为回归估计器。在训练我们的 MDP 时，我们学习一个估计器 $\hat{R}(s,a)$ 来估计 $R(s,a)$。我们现在可以直接估计

$$\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}R(s,a)$$

用

$$\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}\hat{R}(s,a)$$

如果 $\hat{R}(s,a) = R(s,a)$，那么这个估计量显然是正确的。

我们现在将考虑替代方法，并探讨为什么您可能使用一个估算器而不是另一个估算器。

(a) 重要性抽样

一种常用的估计量称为重要性抽样 (Importance Sampling) 估计量。设 $\hat{\pi}_0$ 为真实 π_0 的估计量。重要性抽样估计量使用该 $\hat{\pi}_0$，其形式如下：

$$\mathbb{E}_{s\sim p(s),a\sim\pi_0(s,a)}\frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)}R(s,a)$$

请证明，如果 $\hat{\pi}_0 = \pi_0$，则重要性采样估计量等于：

$$\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}R(s,a)$$

请注意，由于我们有观测数据中项目的 $R(s,a)$ 值，因此该估计量仅要求我们对 π_0 进行建模。

(b) 加权重要性采样

重要性抽样估计量的一种变体称为加权重要性抽样 (Weighted Importance Sampling) 估计量。加权重要性抽样估计量具有以下形式：

$$\frac{\mathbb{E}_{s\sim p(s),a\sim\pi_0(s,a)}\frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)}R(s,a)}{\mathbb{E}_{s\sim p(s),a\sim\pi_0(s,a)}\frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)}}$$

请证明，如果 $\hat{\pi}_0 = \pi_0$，则加权重要性抽样估计量等于

$$\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}R(s,a)$$

(c)

加权重要性抽样估计器的一个问题是，在许多有限样本情况下，它可能会产生偏差。在有限样本中，我们将预期值替换为观测数据集中可见值的总和。请证明加权重要性抽样估计器在这些情况下会产生偏差。

提示：考虑一下你的观察数据集中只有一个数据元素的情况。

助教注：这道题要求你证明，当数据量“有限”，尤其是等于 1 时，加权重要性抽样公式的值在某种程度上不等于 $\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}R(s,a)$。

(d) 双重稳健

最后一种常用的估计量是双重稳健 (Doubly Robust) 估计量。双重稳健估计量的形式如下：

$$\mathbb{E}_{s\sim p(s),a\sim\pi_0(s,a)}((\mathbb{E}_{a\sim\pi_1(s,a)}\hat{R}(s,a))+\frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)}(R(s,a)-\hat{R}(s,a)))$$

双重稳健估计量的一个优点是，当 $\hat{\pi}_0=\pi_0$ 或 $\hat{R}(s,a)=R(s,a)$ 时，它有效。

1 请证明当 $\hat{\pi}_0=\pi_0$ 时，双重稳健估计量等于 $\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}R(s,a)$。

2 请证明当 $\hat{R}(s,a)=R(s,a)$ 时，双重稳健估计量等于 $\mathbb{E}_{s\sim p(s),a\sim\pi_1(s,a)}R(s,a)$。

(e)

现在，我们将考虑几种情况，在这些情况下，您可以选择重要性抽样估计量和回归估计量。请说明在每种情况下，重要性抽样估计量或回归估计量是否可能效果最佳，并解释为什么它会更好。在所有这些情况下，您的状态 s 由患者组成，您的动作 a 代表要给某些患者的药物，您的 $R(s,a)$ 是患者服用药物后的寿命。

1 药物是随机分配给患者的，但药物、患者和寿命之间的相互作用非常复杂。

2 药物以非常复杂的方式分配给患者，但药物、患者和寿命之间的相互作用非常简单。

2 PCA

在课堂上，我们展示了 PCA 可以找到将数据投影到的“方差最大化”方向。在这个问题中，我们发现了 PCA 的另一种解释。

假设我们得到一组点 $\{x^{(1)},\dots,x^{(m)}\}$。假设我们像往常一样对数据进行了预处理，使每个坐标的均值为零，方差为单位。对于给定的单位长度向量 u，让 $f_u(x)$ 成为点 x 在 u 给定方向上的投影。即，如果 $\mathcal{V}=\{\alpha u:\alpha\in\mathbb{R}\}$，则

$$f_u(x)=\arg\min_{v\in\mathcal{V}}||x-v||^2$$

证明最小化投影点和原始点之间的均方误差的单位长度向量 u 对应于数据的第一个主成分。即证明

$$\arg\min_{u:u^Tu=1}\sum_{i=1}^{m}\|x^{(i)}-f_u(x^{(i)})\|_2^2$$

注. 如果要求我们找到一个 k 维子空间来投影数据，以便最小化原始数据和其投影之间的平方和，那么我们应该选择由数据的前 k 个主成分所构成的 k 维子空间。该问题表明，该结果适用于 $k=1$ 的情况。

3 Markov 决策过程

考虑具有有限状态和动作空间的 MDP，以及折扣因子 $\gamma < 1$。令 B 为 Bellman 更新运算符，V 为每个状态的值向量。即，如果 $V' = B(V)$，则

$$V'(s) = R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P_{sa}(s')V(s')$$

(a)

证明，对于任何两个有限值向量 V_1, V_2，有以下事实成立：

$$||B(V_1) - B(V_2)||_\infty \leq \gamma ||V_1 - V_2||_\infty$$

其中

$$||V||_\infty = \max_{s \in S} |V(s)|$$

（这表明 Bellman 更新算子是最大范数中的“γ-收缩”。）

(b)

如果 $B(V) = V$ ，则我们称 V 是 B 的不动点。利用 Bellman 更新算子是最大范数中的 γ 收缩这一事实，证明 B 最多有一个不动点，即 Bellman 方程最多有一个解。您可以假设 B 至少有一个不动点。

注：您在部分 (a) 中证明的结果意味着值迭代几何收敛到最优值函数 V^*。也就是说，经过 k 次迭代后，V 和 V^* 之间的距离最多为 γ^k。

4 单类 SVM

给定一组未标记的示例 $\{x^{(1)}, \ldots, x^{(m)}\}$，单类 SVM 算法尝试找到一个最大程度地将数据与原点分开的方向 w。更准确地说，它解决了（原始）优化问题：

$$\begin{aligned} \min_{w} \quad & \frac{1}{2} w^\top w \\ \text{s.t.} \quad & w^\top x^{(i)} \geq 1 \quad \text{for all} i = 1, \ldots, m \end{aligned}$$

如果 $w^\top x \geq 1$，则新的测试示例 x 标记为 1，否则标记为 0。

(a)

上面给出了单类 SVM 的原始优化问题。写下相应的对偶优化问题。尽可能简化你的答案。特别是，w 不应该出现在你的答案中。

(b)

在训练和测试中，单类 SVM 可以核化 (kernelized) 吗？证明你的答案。

(c) (选做)

给出一个类似 SMO 的算法来优化对偶。即，给出一个算法，该算法在每个优化步骤中都针对最小可能的变量子集进行优化，给出其闭式更新方程。您还应该说明，为什么在每个步骤中一次考虑这么多变量就足够了。

助教注：你应该指出每个步骤优化的最小变量数量，并根据 SMO 算法的原理对其进行证明。