Лабораторная работа 2
Богданов Александр, Б05-003

Запишем информацию о ручках:

Ручка	Средння награда	Количество использований
1	4.6	181
2	4.3	21
3	4.7	384

Общее количество использований: t = 586.

1 Задача

1.1 Задача

Условие: Найдите ε-жадная стратегию. π_ε (положите $\varepsilon = 0.01$).

$$a_t = \begin{cases} \arg\max\limits_{a \in A} Q(a) & \text{с вероятностью } \varepsilon \\ \text{random}(A) & \text{с вероятностью } 1 - \varepsilon \end{cases}$$

В ε случаях выбирается случайная ручка, в $1 - \varepsilon$ выбирается оптимальная ручка, тогда ε-жадная политика: $\pi_\varepsilon = \left[\frac{\varepsilon}{3}; \frac{\varepsilon}{3}; \frac{\varepsilon}{3} + 1 - \varepsilon\right] = \left[\frac{\varepsilon}{3}; \frac{\varepsilon}{3}; 1 - \frac{2\varepsilon}{3}\right] = [0.0033, 0.0033, 0.9933]$.

1.2 Задача

Условие: Найдите UCB стратегию π_{UCB} ($\alpha = 0.5$).

$$a_t = \arg\max_{a \in A} Q(a) + \alpha \sqrt{\frac{\log t}{N_t(a)}}$$

Ручка	Средняя награда (отнормированная)	Добавочная константа	Сумма
1	0.920	0.094	1.014
2	0.860	0.275	1.135
3	0.980	0.064	1.044

То есть UCB политика: $\pi_{\text{UCB}} = [0, 1, 0]$.

1.3 Задача

Условие: Что нужно чтобы применить здесь томсоновское сэмплирование?

1. Ввести априорное распределение.
2. Уметь к априорному распределению строить сопряженное. Для бинарного автомата обычно берут *beta* распределение. Наш автомат не бинарный, но можно провести бинаризацию: 1, 2, 3 -неудача; 4, 5 - успех.

2 Задача

2.1 Задача

Условие: посчитайте logging policy π_0.

Оценим частотной оценкой: $\pi_0 = [0.309, 0.036, 0.655]$.

2.2 Задача

Условие: Оцените стратегию π_1 = [0.3, 0.04, 0.66].

$$\hat{V}(\pi_1, D) = \mathbb{E}_{p(x)\pi_1(a|x)p(r|x,a)}[r] = \mathbb{E}_{\pi_1(a)p(r|a)}[r] = \sum_{a,r} rp(r|a)\pi_1(a) = \sum_a \mathbb{E}[r|a]\pi_1(a)$$

$$\hat{V}(\pi_1, D) = 0.3 * 4.6 + 0.04 * 4.3 + 0.66 * 4.7 = 4.654$$

2.3 Задача

Условие: Оцените стратегию π_2 = [0.3, 0.66, 0.04].

$$\hat{V}(\pi_2, D) = \mathbb{E}_{p(x)\pi_2(a|x)p(r|x,a)}[r] = \mathbb{E}_{\pi_2(a)p(r|a)}[r] = \sum_{a,r} rp(r|a)\pi_2(a) = \sum_a \mathbb{E}[r|a]\pi_2(a)$$

$$\hat{V}(\pi_2, D) = 0.3 * 4.6 + 0.66 * 4.3 + 0.04 * 4.7 = 4.406$$

2.4 Задача

Условие: Оцените стратегию π_ε и π_{UCB}.

$$\hat{V}(\pi, D) = \mathbb{E}_{p(x)\pi(a|x)p(r|x,a)}[r] = \mathbb{E}_{\pi(a)p(r|a)}[r] = \sum_{a,r} rp(r|a)\pi(a) = \sum_a \mathbb{E}[r|a]\pi(a)$$

1. Так как политика π_ε не меняется, то можно оценить как в предыдущих задачах:

$$\hat{V}(\pi, D) = 0.0033 * 4.6 + 0.0033 * 4.3 + 0.9933 * 4.7 = 4.698$$

2. Политика π_{UCB} = [0, 1, 0] после некоторого количества использований поменяется на π_{UCB} = [0, 0, 1], поэтому ее можно оценить так:

$$4.3 \leq \hat{V}(\pi, D) \leq 4.7$$

2.5 Задача

Условие: Проанализируйте результаты. Возможно ли оценить стратегии из 3 предыдущих пунктов с адекватной точностью?

π_1, π_2, π_ε, можно оценить явно посчитав. π_{UCB} можно оценить интервалом – интервал не очень большой, поэтому оценка достаточно точная.

3 Задача

3.1 Задача

Условие: Докажите что оценивание стратегий через IPS несмещенное.

$$\hat{V}_{\text{IPS}}(\pi_{test}, D) = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_{test}(a_i, x_i)}{\pi_0(a_i, x_i)} \cdot r_i,$$

где

$$D = \{(x_i, a_i, r_i)\}_{i=1}^{n} \sim \prod_{i=1}^{n} p(x_i)\pi_0(a_i|x_i)p(r_i|x_i, a_i)$$

$$\begin{aligned}
\mathbb{E}_D\left[\hat{V}_{\text{IPS}}(\pi_{test}, D)\right] &= \frac{1}{n}\sum_{i=1}^{n}\int p(x_i)\pi_0(a_i|x_i)p(r_i|x_i, a_i)\frac{\pi_{test}(a_i, x_i)}{\pi_0(a_i, x_i)}r_i \\
&= \int p(x_i)\pi_{test}(a_i, x_i)p(r_i|x_i, a_i)r_i \\
&= \mathbb{E}_{p(x)\pi_{test}(a|x)p(r|x,a)}[r] = V(\pi_{test})
\end{aligned}$$

3.2 Задача

Условие: При каких необходимых условиях выполняется несмещенность?

Если π_0 может генерировать все действия, которые может сгенерировать π_{test}, иначе будет интеграл не по всему множеству.