№ 1

score:	4,6	4,3	4,2
samples	181	21	384

1) ε-greedy ($\varepsilon = 0{,}01$)

$$a = \begin{cases} \underset{a \in \{1,2,3\}}{\operatorname{argmax}} (S_a), & p = 1-\varepsilon \\ \text{random}(\{1,2,3\}), & p = \varepsilon \end{cases}$$

$$\pi_\varepsilon = \begin{pmatrix} \frac{\varepsilon}{3} \\ \frac{\varepsilon}{3} \\ \frac{\varepsilon}{3} + (1-\varepsilon) \end{pmatrix} = \begin{pmatrix} \frac{1}{300} \\ \frac{1}{300} \\ \frac{1}{300} + \frac{297}{300} \end{pmatrix} = \begin{pmatrix} \frac{1}{300} \\ \frac{1}{300} \\ \frac{298}{300} \end{pmatrix}$$

2) UCB

Оценка Q-функции.

$\hat{Q}(a) = [0{,}92;\ 0{,}86;\ 0{,}84]$

$\alpha = 0.5$

$$u_t = \sqrt{\frac{-\log p}{2N_t(a)}} \Rightarrow \Delta = \alpha \sqrt{\frac{\log t}{N_t(a)}}$$

$t = 181 + 21 + 384 = 586$

$\Delta = [0{,}0938;\ 0{,}275;\ 0{,}06]$

$$\pi_{UCB} = \hat{Q} + \Delta = \begin{pmatrix} 1{,}03 \\ 1{,}13 \\ 1{,}00 \end{pmatrix}$$

Нормализуем вектор π_{UCB} и получаем политику

$$\pi_{UCB} = \begin{pmatrix} 0{,}321 \\ 0{,}360 \\ 0{,}319 \end{pmatrix}$$

3) Томпсоновское семплирование

Для томпсоновского семплирования необходимо

а) Ввести пространство наград и функцию правдоподобия

б) ~~Ввести~~ Ввести априорное распределение на награды.

1) Все пользователи могут выставить оценки от 1 до 5.

по $\{1, \dots, 5\}$

Тогда на пространстве наград введём мультиномиальное распредел

№1

score:	4,6	4,3	4,2
samples	181	21	384

1) ε-greedy ($\varepsilon = 0{,}01$)

$$a = \begin{cases} \underset{a\in\{1,2,3\}}{\operatorname{argmax}} (~~score~~\ S_a), & p = 1-\varepsilon \\ \operatorname{random}(\{1,2,3\}), & p = \varepsilon \end{cases}$$

$$\pi_\varepsilon = \begin{pmatrix} \frac{\varepsilon}{3} \\ \frac{\varepsilon}{3} \\ \frac{\varepsilon}{3} + (1-\varepsilon) \end{pmatrix} = \begin{pmatrix} \frac{1}{300} \\ \frac{1}{300} \\ \frac{1}{300} + \frac{297}{300} \end{pmatrix} = \begin{pmatrix} \frac{1}{300} \\ \frac{1}{300} \\ \frac{298}{300} \end{pmatrix}$$

2) UCB

Оценка Q-функции.

$\hat{Q}(a) = [0{,}92;\ 0{,}86;\ 0{,}94]$

$\alpha = 0.5$

$$u_t = \sqrt{\frac{-\log p}{2N_t(a)}} \Rightarrow \Delta = \alpha\sqrt{\frac{\log t}{N_t(a)}}$$

$t = 181 + 21 + 384 = 586$

$\Delta = [0{,}0938;\ 0{,}275;\ 0{,}05]$

$$\pi_{UCB} = \hat{Q} + \Delta = \begin{pmatrix} 1{,}03 \\ 1{,}13 \\ 1{,}00 \end{pmatrix}$$

Нормализуем вектор π_{UCB} и получим политику

$$\pi_{UCB} = \begin{pmatrix} 0{,}321 \\ 0{,}360 \\ 0{,}319 \end{pmatrix}$$

3) Байесовское оптимизирование

Для Байесовского оптимизирования необходимо

а) Ввести пространство наград с функцией правдоподобия

б) Ввести априорное распределение на награды

1) Все пользователи могут выставить в оценки

от 1 до 5

по $\{1, \dots, 5\}$

Тогда на пространстве наград введём мультиномиальное распр

p_k^i – вероятность оценки i для k-ого продукта.

2) Необходимо ввести функцию распред. $p(p_k^i|\theta)$,

p – плотн. распред. с парамет. θ

Т.к. распред.

Это распределение не должно удовлетворять многим свойствам и желательно быть сопряженным априорным.

Этим свойствам для мультиномиального распредел. обладает распредел. Дирихле

$p(p_k^i|\theta) = Dir_k(p_k^i|\alpha_k)$ – для каждого продукта своё распредел.

$(\alpha_k| = 5$ – совпадает с целевой оценкой.

3) Как только мы определились с приорным наблюдением, что необходимо для апостериорного это уметь пересчитать параметры апостериорного распредел.

$\vec{\alpha}_k' = \vec{\alpha}_k + \sum_{i=1}^{n} \vec{X}_k^i$, где $\vec{X}$ – это оценка выставленная i-ым пользователем по k-ому продукту по результатам испытаний в виде one-hot вектора

Алгоритм.

$p_k^i \sim Dir(\vec{\alpha}_k)$

$\hat{i}_k = \arg\max \sum_{i=1}^{5} p_{k-i}$

$\vec{\alpha}_k = \vec{\alpha}_k + \vec{X}_k$, где $\vec{X}_k$ – оценка награды от пользоват. в виде one-hot вектора

№2

1) logging policy

$\hat{\pi}_0 = \begin{pmatrix} 181 \\ 21 \\ 394 \end{pmatrix}$, $\pi_0 = \frac{\hat{\pi}_0}{\|\hat{\pi}_0\|_1} = \begin{pmatrix} 0{,}3 \\ 0{,}04 \\ 0{,}66 \end{pmatrix}$

2) Evaluate. $\pi_2 = [0.3, 0.04, 0.66]$

$\hat{V}(\pi_0, D) = E_{p(a) \pi_2(a|x) p(r|x,a)}[r]$

Для оценки $\hat{V}(\pi_2, D)$ используем IPS

Для подсчёта IPS и оценки её variance необходимо

подсчитать propensity-score стратег.

~~$w_j z$~~ ~~π_{best}~~: $w(x,a) = \frac{\pi(x,a)}{\pi_0(x,a)}$, где $\pi_0(x,a)$ – logging policy

~~$w(x|a)$~~ $\hat{V}_{IPS} = \frac{1}{n} \sum w(x_i, a_i)$ ~~$\hat{V}_{IPS} = \frac{1}{n} \sum w(x_i, a_i) r_i$~~

$N = 181 + 21 + 394 = 596$

$\hat{V}_{IPS}(\pi) - V(\pi) \le \hat{w}_{max} \sqrt{\frac{1}{2N} \log \frac{2}{\delta}} \cdot r_{max}$ (из неравенства Хёффдинга)

$r_{max} = 1$ (оценки нормализованы)

$\delta = 0{,}05$ – необходимый уровень значимости.

$\hat{V}_{IPS} = \frac{1}{N} \sum w(x_i, a_i) r_i = \sum \pi(x,a) \cdot \bar{r}_i$, т.к. нам уже даны

усреднённые оценки по log-policy

$w_1 = w_2 = w_3 = 1$ (Политика совпадает с log-policy)

$\bar{r}_i = [0{,}92; 0{,}86; 0{,}94]$

$\hat{V}_{IPS} = 0{,}3 \cdot 0{,}92 + 0{,}04 \cdot 0{,}86 + 0{,}66 \cdot 0{,}94 = 0{,}93$

$Var_{0{,}05}(\pi_2) = 0{,}04$

Ответ: $0{,}93 \pm 0{,}04$

3) $\pi_3 = [0{,}3, 0{,}66, 0{,}04]$

$w_1 = 1$ $w_2 = \frac{0{,}66}{0{,}04} = 16{,}5$ $w_3 = \frac{0{,}04}{0{,}66} = 0{,}07$

$Var_\delta(\pi_3) = 16{,}5 \cdot 0{,}04 = 0{,}66$ – слишком большая

ошибка оценки из-за того, что $\pi_0^2 \ll \pi_3^2$

$V_{IPS} = 0,3 \cdot 0,92 + 0,66 \cdot 0,86 + 0,04 \cdot 0,94 = 0,88$

Ответ: $0,88 \pm 0,66$

4.) Возьмём ε-greedy стратегию

$$\pi_3 = \begin{pmatrix} \frac{1}{300} \\ \frac{1}{300} \\ \frac{298}{300} \end{pmatrix}$$

$w_1 = \frac{\frac{1}{300}}{0,3} = 0,01 \quad w_2 = 0,08 \quad w_3 = \frac{\frac{298}{300}}{0,66} = 1,5$

$\cancel{Var} \quad Var_{обс}(\pi_3) = 1,5 \cdot 0,04 = 0,06$

$V_{IPS} = \frac{1}{300} \cdot 0,92 + \frac{1}{300} \cdot 0,86 + \frac{298}{300} \cdot 0,94 = 0,94$

Ответ: $0,94 \pm 0,06$

5.) Стратегии сравнить на уровне значимости невозможно, т.к. ~~интервала~~ доверительные интервалы всех стратегий накладываются друг на друга.

В основном это связано с тем, что стратегии сильно отличаются [illegible] от log policy.

Требуются живые эксперименты, например A/B тестирование.

№3

1) Несмещённость оценки IPS

$$\mathbb{E}_D[\hat V_{IPS}(\pi;D)] = V(\pi) = \mathbb{E}_{p(x)\pi(a|x)\cdot p(r|a,x)}[r]$$

Рассмотрим слагаемое связанное с одним из действий

$$\hat V_{IPS}(x_i,a) = r_i\frac{\mathbb{1}(a_i=a)}{\mu(a_i|x_i)},\ \mu(a_i|x_i) \text{ – logging policy}$$

Предположение: $\mu(a|x)>0$, т.е. ~~политика~~ log policy проверяет все действия и ни одно действие не имеет вероятность 0.

$$\mathbb{E}[\hat V_{IPS}(x_i,a)\,|\,x_i,a] = \cancel{\sum_{a_i\in A}\mu(a'|x_i)\mathbb{E}\hat V_{IPS}}$$

A – пространство действий

$$= \sum_{a_i\in A}\mu(a'|x_i)\,\mathbb{E}[\hat V_{IPS}(x_i,a)\,|\,x_i,a,a_i=a'] =$$

$$= \sum_{a'\in A}\mu(a'|x_i)\cdot V(x_i,a')\cdot\frac{\mathbb{1}(a'=a)}{\mu(a'|x_i)} = /\mu(a'|x_i)\neq 0/ =$$

$$= V(x_i,a)$$

Данные семплируем независимо из $\sim D$

Тогда

$$\hat V_{IPS}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\pi(a|x_i)\,\hat V_{IPS}(x_i,a) = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\mu(a_i|x_i)}r_i$$

$$\mathbb{E}_{p(x)}\mathbb{E}[\hat V_{IPS}(\pi)\,|\,x_i,a\sim\pi(x)] =$$

$$= \mathbb{E}_{p(x)}\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\pi(a|x_i)\,\hat V_{IPS}(x_i,a)\,|\,x_i,a] =$$

$$= \mathbb{E}_{p(x)}\sum_{a\in A}\pi(a|x_i)\,V(x_i,a)] = V(\pi)$$

2) $\mu(a|x)>0$ – нет действий с 0 вероятностью в log policy

$(x_i,r_i,a_i)\sim D$ – независимо и о.р. ~~...~~ (т.е. выборка i.i.d)