ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Домашнее задание #4

Hard deadline по теоретическому заданию: ***23:59 21 апреля***
Hard deadline по заданию на программирование: ***23:59 28 апреля***

В этом задании суммарно можно набрать 140 баллов. Первые 100 набранных баллов идут в зачет этого домашнего задания. Баллы, превышающие 100, идут на повышение результатов прошлых домашних заданий.

1 Теоретическое задание [70 баллов]

Пункт 1 **15 баллов**

Опишите разницу между фреймворком A/B тестирования и фреймворком многоруких бандитов (multi-armed bandits). Для каждого фреймворка приведите пример задачи, для которой он подходит лучше, чем другой.

Пункт 2 **15 баллов**

Опишите фреймворк contextual bandits. В чем ключевая разница с классическими бандитами?

Пункт 3 **20 баллов**

В этой задаче мы работаем с упрощенным фреймворком contextual bandit в котором есть всего два возможных значения контекста ($x \in \{c_1, c_2\}$) и два возможных действия ($a \in \{a_1, a_2\}$). Рассмотрим алгоритм $\mathcal{A}$. В каждом раунде на вход алгоритму подается контекст $x \in \{c_1, c_2\}$ и алгоритм должен выбрать одно из двух действий $a \in \{a_1, a_2\}$. По результатам раунда алгоритм получает награду $r \in [0, 1],\ r \sim \mathcal{P}(x_t, a_t)$.

Предположим, что алгоритм $\mathcal{A}$ это классический алгоритм UCB-1, который игнорирует контекст, то есть выбирает действие a без учета x. Придумайте пример, в котором алгоритм $\mathcal{A}$ работает плохо, то есть получает регрет $\mathcal{O}(T)$.

Вам нужно подобрать последовательность контекстов $\{x_1, x_2, \ldots, x_T\}$ и модель генерации награды $\mathcal{P}(x_t, a_t)$. В этой задаче регрет измеряется относительно лучшей policy $\pi : x \to a$, которая для каждого контекста x выбирает лучшее действие a.

Подсказка: Вы можете предположить, что вам известен код алгоритма $\mathcal{A}$ и в имплементации алгоритма нет элементов случайности.

Пункт 4 **20 баллов**

В условия предыдущей задачи придумайте алгоритм $\mathcal{B}$, который построен на основе алгоритма $\mathcal{A}$ и терпит регрет $o(T)$. Алгоритм $\mathcal{B}$ должен использовать алгоритм $\mathcal{A}$ как функцию и может запускать несколько независимых версий алгоритма $\mathcal{A}$.

2 Задание на программирование [70 баллов]

В этой части вам предстоит реализовать два алгоритма для задачи многоруких бандитов. Подробная постановка дана в Google Colab:

`https://colab.research.google.com/drive/199VOtGsQBTy8BOGbxK_bjkzAkPfX2yJD?usp=sharing`

Для решения скопируйте ноутбук себе и оформите решение в нем. Прикрепите ссылку на ваш ноутбук к решению на EDU и убедитесь, что вы выдали доступ на просмотр вашего решения. Пожалуйста, убедитесь, что весь ноутбук можно запустить после перезапуска ядра.

Удачи!