Семинар 12.
ММП, весна 2013
7 мая

Илья Толстихин
iliya.tolstikhin@gmail.com

Темы семинара:

- Задача о многоруком бандите;
- Hoeffding race, Bernstein race;
- Применение в скользящем контроле.

На этом семинаре мы рассмотрим задачу о *случайном многоруком бандите* (stochastic multi-armed bandit), рассмотренной на лекциях, в более простой постановке. Она формулируется следующим образом. Пусть у нас имеется конечное множество *действий* $A = \{1, \dots, K\}$. Будем считать, что каждому действию $i \in \{1, \dots, K\}$ соответствует фиксированное распределение ν_i на отрезке $[0, 1]$ с математическим ожиданием μ_i. Обозначим *оптимальное действие* $i^* = \arg\max_{i=1,\dots,K} \mu_i$ — действие, распределение которого имеет наибольшее среднее значение. Предполагается, что *применение действия* $i \in \{1, \dots, K\}$ заключается в получении реализации случайной величины, распределенной согласно ν_i. Причем каждое следующее вытягивание происходит независимо от прошлых. Задача будет заключаться в поиске стратегии, то есть схемы последовательного применения действий, которая позволит за как можно меньшее число шагов найти оптимальное действие i^*.

Строго говоря, на каждом шаге $t \in \mathbb{N}$ *игрок* выбирает действие $I_t \in \{1, \dots, K\}$, для которого получает реализацию случайной величины с распределением ν_{I_t}. Обозначим с помощью $T_i(t)$ число шагов, на которых игрок выбирал действие i к началу t-ого шага. С помощью $\bar{X}_{i,T_i(t)}$ обозначим среднее выборочное $T_i(t)$ i.i.d. наблюдений действия i, накопившихся к началу t-го шага. По истечении определенного числа шагов (которое мы опишем позже) игрок должен выбрать единственное действие I, которое ему кажется оптимальным. Задача игрока заключается в том, чтобы среднее значение μ_I распределения ν_I выбранного им действия I оказалось как можно ближе к оптимальному значению μ_{i^*}.

В литературе принято рассматривать самые разные ограничения на число шагов и критерии остановки. Все они обусловлены конкретными приложениями. Мы рассмотрим следующий критерий остановки: процедура должна остановиться, как только она нашла стратегию I, удовлетворяющую условию $\mu_{i^*} - \mu_I \leqslant \varepsilon$ с вероятностью не меньше $1 - \delta$, и при этом число примененных процедурой действий не может превосходить N. Здесь N, δ и ϵ — параметры задачи.

Название задачи происходит от игровых автоматов (слот-машин), часто установленных в казино, которые принято называть «однорукими бандитами» из-за рычагов, прикрепленных к ним. Многоруким бандит становится в тот момент, когда игрок стоит перед K автоматами и выбирает, в каком бы порядке их дергать.

На этом семинаре мы рассмотрим две стратегии решения задачи о случайном многоруком бандите: Hoeffding race и Bernstein race.

1 Hoeffding and Bernstein races

Алгоритм Hoeffding race

Вход: параметр δ, максимальное число действий N;
Выход: выбранная стратегия $I \in \{1, \ldots, K\}$;

1: $t := 1$;
2: $A := \{1, \ldots, K\}$;
3: **пока** $|A| > 1$ и число примененных действий не превосходит N
4: Получаем t-ые реализации каждого действия $i \in A$
5: Удаляем из A все действия со средними выборочными $\bar{X}_{i,t}$, отличающимися более чем на $\sqrt{2\log(2NK/\delta)/t}$ от максимального среднего выборочного $\max_i \bar{X}_{i,t}$:

$$A := A \setminus \left\{ j \in A \colon \bar{X}_{j,t} \leq \max_{i=1,\ldots,K} \bar{X}_{i,t} - \sqrt{\frac{2\log\frac{2NK}{\delta}}{t}} \right\}.$$

6: $t := t + 1$.
7: Возвращаем любой элемент из A.

Задача. *Докажите, что (а) с вероятностью не меньше $1 - \delta$ оптимальное действие i^* не будет исключено в ходе итераций, (б) в случае, если после остановки алгоритма $|A| > 1$ (алгоритм завершился по достижению максимального числа запросов), все оставшиеся по завершению алгоритма действия $i \in A$ с вероятностью не менее $1 - \delta$ удовлетворяют*

$$\mu_{i^*} - \mu_i \leqslant 4\sqrt{\frac{\log\frac{2NK}{\delta}}{2N/K}}.$$

Доказательство: нам достаточно воспользоваться нашим любимым неравенством Хевдинга, утверждающим, что если $X_1, \ldots, X_n$ — i.i.d. выборка случайных величин, принимающих значение в интервале $[0, 1]$ и имеющих математическое ожидание $\mathbb{E}[X]$, то для $\delta > 0$

$$\mathbb{P}\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \right| \leqslant \sqrt{\frac{\log\frac{2}{\delta}}{2n}} \right\} \geqslant 1 - \delta.$$

Наш алгоритм явно пользуется доверительными интервалами на математические ожидания μ_i, описываемыми неравенством Хевдинга. Чтобы оптимальное действие не было исключено в ходе его работы, достаточно, чтобы все доверительные интервалы выполнялись, то есть действительно содержали в себе математическое

ожидание действия. Поскольку наперед мы не знаем, какие действия мы будем удалять, и, соответственно, для каких действий придется использовать доверительные интервалы на тех или иных шагах, мы можем заранее потребовать, чтобы для каждого действия K на каждом из N шагов выполнялось неравенство Хевдинга. Таким образом мы хотим, чтобы одновременно выполнились NK неравенств

$$\mu_i - \sqrt{\frac{\log\frac{2NK}{\delta}}{2k}} \leqslant \bar{X}_{i,k} \leqslant \mu_i + \sqrt{\frac{\log\frac{2NK}{\delta}}{2k}}, \quad k = 1, \ldots, N,\ i = 1, \ldots, K. \tag{1}$$

Если все эти неравенства выполнены, то оптимальное действие мы не исключим. А поскольку неравенство Буля говорит нам, что все эти неравенства выполнены одновременно с вероятностью не меньше $1 - \delta$, то утверждение (а) доказано.

Чтобы доказать утверждение (б) докстаточно заметить, что за N шагов каждое из действий будет испытано уж точно не меньше N/K раз. Поскольку мы только что показали, что с вероятностью не меньше $1 - \delta$ выполнены сразу все неравенства (1), то и в конце алгоритма эти неравенства выполнены для оставшихся действий A. А раз так, то математические ожидания отличаются не более чем удвоенную длину доверительного интервала. Взяв в (1) в качестве k нижнюю оценку N/K, мы завершаем доказательство.

Алгоритм Bernstein race заключается в замене доверительных интервалов неравенства Хевдинга на доверительные интервалы неравенства Бернштейна, которые, как мы знаем, могут оказаться точнее в тех случаях, когда у нас есть хорошие верхние оценки дисперсии случайной величины. Оказывается, в неравенстве Бернштейна дисперсию случайных величин можно заменить ее выборочной оценкой ценой небольшого увеличения длины доверительного интервала.

Алгоритм Bernstein race

Вход: параметр δ, максимальное число действий N;
Выход: выбранная стратегия $I \in \{1, \ldots, K\}$;

1: $t := 1$;
2: $A := \{1, \ldots, K\}$;
3: **пока** $|A| > 1$ и число примененных действий не превосходит N
4: Получаем t-ые реализации каждого действия $i \in A$
5: Удаляем из A все действия со средними выборочными $\bar{X}_{i,t}$, отличающимися более чем на $\sqrt{2\log(2NK/\delta)/t}$ от максимального среднего выборочного $\max_i \bar{X}_{i,t}$:

$$A := A \setminus \left\{ j \in A\colon \bar{X}_{j,t} + \sqrt{\frac{2V_{j,t}\log\frac{2NK}{\delta}}{t}} + 6\frac{\log\frac{2NK}{\delta}}{t} \leqslant \max_{i=1\ldots,K} \left(\bar{X}_{i,t} - \sqrt{\frac{2V_{i,t}\log\frac{NK}{\delta}}{t}} \right) \right\},$$

где $V_{i,t} = \frac{1}{t}\sum_{s=1}^{t}(X_i, s - \bar{X}_{i,t})^2$ — выборочная дисперсия действия i.
6: $t := t + 1$.
7: Возвращаем любой элемент из A.

Список литературы

[1] *Audibert, J.-Y., Bubeck, S., Rémi, M.* Bandit View on Noisy Optimization. Optimization for Machine Learning, editors S. Sra, S. Nowozin and S. J. Wright. MIT Press, 2011.