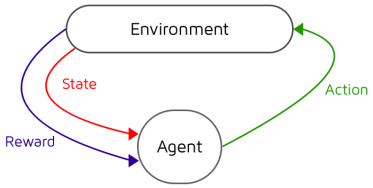
Statistical Learning Theory
Проект (реферат по статье)

Сергей Володин, 374 гр.

1 Обучение с подкреплением

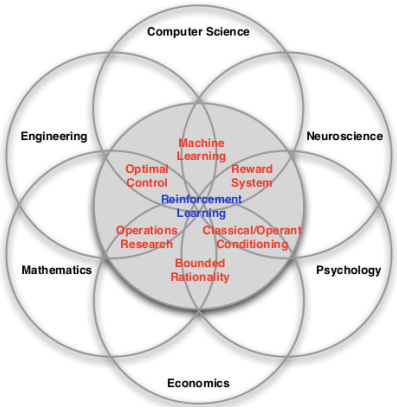
Обучение с подкреплением — область машинного обучения, идея которой основана на поведенческой психологии. Рассматривается *агент*, взаимодействующий со *средой* путем наблюдения ее *состояния*, совершения *действий* и получения от среды *награды*.

Рис. 1: Блок-схема обучения с подкреплением



Обучение с подкреплением находится на стыке многих областей, таких как машинное обучение, оптимальное управление и neuroscience.

Рис. 2: Положение обучения с подкреплением. Лекция Дэвида Сильвера [2]



В последнее время обучение с подкреплением набирает популярность благодаря успехам в решении важных практические задачи. Например, при помощи обучения с подкреплением удается обучать агентов, способных играть в игры Atari [3], Go [4] и backgammon.

2 Введение

Поскольку агент обучается правильным действиям в среде постепенно, сначала он может принимать неоптимальные действия. Таким образом возникает понятие сожаления (regret) — величины, характеризующей то, насколько неоптимально агент выбирает действия. Данный реферат рассказывает о статье «On Lower Bounds for Regret in Reinforcement Learning» (Ian Osband, Benjamin Van Roy) [1]. В статье рассматриваются нижние границы на сожаление. Данная величина характеризует «сложность» самой среды в терминах разницы между наградой данного алгоритма и некоего «наилучшего» алгоритма обучения с подкреплением.

Далее будут даны необходимые определения, а затем будет рассмотрен пример самой простой среды — многорукий бандит. Для нее будет получена нижняя оценка на сожаление.

В оригинальной статье разбирается также случай среды с двумя состояниями, но в данном реферате данный раздел не представлен.

3 Постановка задачи

Определение 3.1. (Марковский процесс принятия решений)
ММПР — это кортеж $(\mathcal{S}, \mathcal{A}, R, P)$, где

1. $\mathcal{S} = \{1, ..., S\}$ — множество состояний среды
2. $\mathcal{A} = \{1, ..., A\}$ — множество действий, доступных агенту
3. $R(s, a)$ — функция награды. Для данных $s \in \mathcal{S}$ и $a \in \mathcal{A}$ случайная величина $R(s, a) \in [0, 1]$ — награда за действие
4. $P(s, a)$ — функция переходов. Для данных $s \in \mathcal{S}$ и $a \in \mathcal{A}$ случайная величина $P(s, a) \in \mathcal{S}$ — новое состояние среды

Определение 3.2. (Взаимодействие агента со средой)
Имеется ММПР $(\mathcal{S}, \mathcal{A}, R, P)$. Вводится время $t \in \mathbb{N}$. Для каждого момента времени 1:

1. Агент получает состояние $s_t \in \mathcal{S}$
2. Агент выбирает действие $a_t \in \mathcal{A}$
3. Агент получает награду $r_t \sim R(s_t, a_t) \in [0, 1]$
4. Среда переходит в новое состояние $s_{t+1} \sim P(s_t, a_t)$

Определение 3.3. (Политика) Имеется ММПР $(\mathcal{S}, \mathcal{A}, R, P)$.
Политика μ — отображение $\mu \colon \mathcal{S} \to \mathcal{A}$. То есть, каждому состоянию $s \in \mathcal{S}$ сопоставляется действие агента $a \in \mathcal{A}$

Определение 3.4. (Средняя награда) Имеется ММПР $M = (\mathcal{S}, \mathcal{A}, R, P)$, а также политика μ.

$$\lambda_\mu^M(s) = \lim_{T \to \infty} \mathbb{E}_{M,\mu} \left[\frac{1}{T} \sum_{t=1}^{T} \overline{r}(s_t, a_t) \middle| s_1 = s \right]$$

где $\overline{r}(s, a) = \mathbb{E}R(s, a)$ — средняя награда в (s, a)
То есть, $\lambda_\mu^M(s)$ — средняя награда за бесконечное время при следовании политике μ в ММПР M, если стартовать из состояния $s \in \mathcal{S}$. В некотором смысле это «ценность» состояния s.
Политика μ^M оптимальна для M, если $\mu^M \in \arg\max_\mu \lambda_\mu^M(s)$ для всех $s \in \mathcal{S}$. То есть, при старте из любого состояния s политика μ максимизирует среднюю награду за бесконечное время.
Величина $\lambda_*^M(s) = \lambda_{\mu^M}^M(s)$ называется оптимальной средней наградой.

Определение 3.5. (История) Имеется ММПР $M = (\mathcal{S}, \mathcal{A}, R, P)$. История к моменту времени t — кортеж

$$\mathcal{H}_t = (s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}, r_{t-1})$$

То есть, это последовательный «журнал» всех состояний, действий и наград, которые произошли во время взаимодействия агента со средой.

Определение 3.6. (Алгоритм обучения с подкреплением) Имеется ММПР $M = (\mathcal{S}, \mathcal{A}, R, P)$. Алгоритм обучения с подкреплением π — это последовательность функций $\pi = \{\pi_t \big| t \in \mathbb{N}\}$, где π_t — функция, сопоставляющая истории $\mathcal{H}_t$ распределение над политиками $\pi_t(\mathcal{H}_t)$
То есть, алгоритм π получает историю $\mathcal{H}_t$ в момент времени t и возвращает распределение над политиками $\pi_t(\mathcal{H}_t)$. Далее выполняется сэмплинг $\mu_t \sim \pi_t(\mathcal{H}_t)$ из этого распределения, и агент возвращает действие $\mu_t(s_t)$.

Каждое распределение π_t должно обладать следующим свойством: если переставить действия и состояния биекциями $s\colon \mathcal{S} \to \mathcal{S}$ и $a\colon \mathcal{A} \to \mathcal{A}$ в истории $\mathcal{H}_t$, получив историю $\mathcal{H}'_t$, то распределение над «исходными» политиками не изменится, а именно:

$$(\pi_t(\mathcal{H}'_t))(\mu') = (\pi_t(\mathcal{H}_t))(a^{-1} \circ \mu' \circ s^{-1})$$

Это свойство означает, что алгоритм действительно получает информацию о среде, а не просто выбирает всегда одно и то же действие.

Данное свойство необходимо для симметрии в Теореме 4.1. В оригинальной статье [1] это свойство на самом деле не выполняется, а в более ранней [5] используется другой приём

Определение 3.7. (Сожаление) Имеется ММПР $M = (\mathcal{S}, \mathcal{A}, R, P)$ и агент π.

Сожаление агента π в момент времени T для начального состояния s в цепи M определяется следующим образом:

$$\mathrm{Regret}(T, \pi, M, s) = \sum_{t=1}^{T} (\lambda^M_{\mu^M}(s) - r_t) = T\lambda^M_*(s) - \sum_{t=1}^{T} r_t$$

То есть, агент взаимодействует со средой, которая изначально находится в состоянии s в течение T шагов времени. В этом процессе им получаются награды $\{r_i\}_{i=1}^T$.

За всё время T агент мог бы действовать оптимально. Тогда можно ожидать, что он бы получил около $T\lambda^M_*(s)$ награды, так как λ^M_* — средняя награда для состояния s. Вместо этого он получил только $\sum\limits_{t=1}^{T} r_t$.

Заметим, что $\mathrm{Regret}(T, \pi, M, s)$ — случайная величина (случайность берётся как из недетерминированности агента, так и из недетерминированности среды)

4 Многорукие бандиты

Определение 4.1. (Многорукий бандит) Многорукий бандит M — ММПР с всего одним состоянием: $S = 1$. Действия $a \in \mathcal{A}$ называются «руками».

Для многорукого бандита оптимальная средняя награда вырождается в максимальное $\overline{r}(a)$:

$$\lambda^M_* = \max_a \overline{r}(a)$$

То есть, один раз выбирается действие, для которого средняя награда $\overline{r}$ максимальна и повторяется каждый раз.

Теорема 4.1. (Нижняя граница на сожаление для многоруких бандитов)

Имеется многорукий бандит M. Тогда для любого алгоритма обучения с подкреплением π существует функция награды R, такая что

$$\mathbb{E}\mathrm{Regret}(T, \pi, M) \geqslant \frac{1}{24}\sqrt{AT}$$

Доказательству этой теоремы посвящен остаток этого раздела.

Рассмотрим следующую среду: пусть M — многорукий бандит с $A \geqslant 2$ и следующей функцией награды:

$$R(a) = \begin{cases} \mathrm{Be}(\delta), & a \neq a^* \\ \mathrm{Be}(\delta + \varepsilon), & a = a^* \end{cases}$$

То есть, все «руки» одинаковые и имеют распределение награды по Бернулли с параметром δ, кроме руки a^*, имеющей распределение $\mathrm{Be}(\delta + \varepsilon)$.

Определяется изменённая награда

$$\tilde{r}_t(a) = \begin{cases} r_t(a), & a \neq a^* \\ \sim \mathrm{Be}(\delta), & a = a^* \end{cases}$$

То есть, награда для действий, отличных от a^* остается без изменений. Для действия a^* производится дополнительный сэмплинг из независимого распределения $\mathrm{Be}(\delta)$. Заметим, что распределение такой награды $\tilde{r}_t(a)$ всегда является $\mathrm{Be}(\delta)$. То есть, она никак не информирует агента о преимуществе a^* перед другими действиями.

Рассматривается изменённая история для $\tilde{a}_t \sim \pi_t(\tilde{\mathcal{H}}_t)$

$$\tilde{\mathcal{H}}_t = (\tilde{a}_1, \tilde{r}_1, ..., \tilde{a}_{t-1}, \tilde{r}_{t-1})$$

В этой истории агент получает измененную награду $\tilde{r}_t$, которая никак не информирует его о том, что рука a^* «лучше», чем остальные.

Определяются величины $n_T(a) = \sum_{t=1}^{T}[a_t = a]$ и $\tilde{n}_T(a) = \sum_{t=1}^{T}[\tilde{a}_t = a]$ — количества выборов действия a в историях $\mathcal{H}_{T+1}$ и $\tilde{\mathcal{H}}_{T+1}$ соответственно.

Лемма 4.1. (Сожаление неинформированного агента)

Рассмотрим построенного многорукого бандита M. Для всех δ, $\varepsilon > 0$ и всех алгоритмов обучения с подкреплением π сожаление

$$R_u = T\lambda_*^M(a) - \mathbb{E}\sum_{t=1}^{T} r(\tilde{a}_t) \geqslant \frac{A-1}{A}T\varepsilon$$

Доказательство. Рассмотрим $T\lambda_*^M = T\max_a \overline{r}(a)$. Тогда

$$R_u = \mathbb{E}_{\text{agent}} \sum_{t=1}^{T} \mathbb{E}_{\text{env}}[r(a^*) - r(\tilde{a}_t)]$$

Эта разница равна ε каждый раз, когда $\tilde{a}_t \neq a^*$ и 0 в противном случае. Перепишем сумму как сумму по действиям и количеству действий:

$$R_u = \mathbb{E}\sum_a \tilde{n}_T(a)(\overline{r}(a^*) - \overline{r}(a)) = \mathbb{E}\sum_{a \neq a^*} \varepsilon\tilde{n}_T(a) = \mathbb{E}\varepsilon(T - \tilde{n}_T(a^*)) = \varepsilon T\frac{A-1}{A}$$

Где последний шаг произведен из соображений симметрии:

$$\mathbb{E}\tilde{n}_T(a^*) = \frac{T}{A}$$

Симметрия возникает, поскольку агент не может отличить a^* от других действий, получая награду $\tilde{r}_t$.

В оригинальной статье [1] данный переход произведен безосновательно. В более ранней [5] используется матожидание по различным средам $a^* \sim u\{1, ..., A\}$, то есть, вычисляется среднее сожаление по всем средам с различными «лучшими» руками a^*. В этом тексте этот переход верен из-за условий на перестановки в Определении 3.6 □

Перейдём к следующему этапу доказательства Теоремы 1. В этой части докажем следующее утверждение: если ε достаточно мал, тогда распределение награды $\tilde{r}_t(a_t)$ близко к $r_t(a_t)$.

Определение 4.2. (Total variation distance)

Пусть $(\Omega, \mathcal{F}, P)$ и $(\Omega, \mathcal{F}, Q)$ — два вероятностных пространства.

Тогда total variation distance между P и Q определяется как:

$$\delta(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

То есть, это максимальный модуль разности мер по всем измеримым подмножествам.

Аналогично total variation distance определяется для случайных величин

Определение 4.3. (Расхождение Кульбака-Лейблера) Пусть P и Q — два распределения. Тогда расхождением Кульбака-Лейблера Q относительно P называется

$$d_{\text{KL}}(P||Q) = \int_X p(x)\log_2\frac{p(x)}{q(x)}dx$$

Теорема 4.2. (Неравенство Пинскера) Пусть P, Q — две случайные величины.

Тогда

$$\delta(P, Q) \leqslant \sqrt{\frac{1}{2}d_{\text{KL}}(P||Q)}$$

Доказательство. См. [7] □

Определим две последовательности наград:

$$\begin{cases} r_t^T = (r_t, ..., r_T) \\ \tilde{r}_t^T = (\tilde{r}_t, ..., \tilde{r}_T) \end{cases}$$

Определим условное распределение последовательностей наград r_t^T и $\tilde{r}_t^T$ при заданной истории в точке $z_t^T \in \mathbb{R}^{T-t+1}$

$$\begin{cases} P(z_t^T | \mathcal{H}_t) = \mathbb{P}(r_t^T = z_t^T \big| \mathcal{H}_t) \\ \tilde{P}(z_t^T \big| \tilde{\mathcal{H}}_t) = \mathbb{P}(\tilde{r}_t^T = z_t^T \big| \tilde{\mathcal{H}}_t) \end{cases}$$

Рассмотрим матожидание расхождения Кульбака-Лейблера $P(z_t^T | \mathcal{H}_t)$ относительно $\tilde{P}(z_t^T | \tilde{\mathcal{H}}_t)$:

$$d_t^T = d_{\mathrm{KL}}(\tilde{P}(z_t^T | \tilde{\mathcal{H}}_t) || P(z_t^T | \mathcal{H}_t) = \mathbb{E} \sum_{z_t^T} \tilde{P}(z_t^T | \tilde{\mathcal{H}}_t) \log_2 \frac{\tilde{P}(z_t^T | \tilde{\mathcal{H}}_t)}{P(z_t^T | \mathcal{H}_t)}$$

Лемма 4.2. (КЛ-расхождение неинформированного распределения)
Рассмотрим построенного многорукого бандита M. Для всех ε, $\delta > 0$ и всех алгоритмов обучения с подкреплением π

$$d_1^T \leqslant \frac{T}{A} \left(\delta \log_2 \frac{\delta}{\delta + \varepsilon} + (1 - \delta) \log_2 \frac{1 - \delta}{1 - \delta - \varepsilon} \right)$$

Доказательство. По цепному правилу для расхождения Кульбака-Лейблера [5], [6]

$$d_t^T = \sum_{t=1}^{T} d_t^t$$

Поскольку награда $\tilde{r}_t$ всегда распределена по $\mathrm{Be}(\delta)$,

$$\tilde{P}(z_t^t) = \begin{cases} \delta, & z_t^t = 1 \\ 1 - \delta, & z_t^t = 0 \end{cases}$$

Для r_t распределение зависит от действия. Для $a \neq a^*$ распределение совпадает с $\tilde{P}$. Но для $a = a^*$

$$P(z_t^t) = \begin{cases} \delta + \varepsilon, & z_t^t = 1 \\ 1 - \delta - \varepsilon, & z_t^t = 0 \end{cases}$$

Эти числа как раз стоят в верхней оценке. Дополнительные вычисления [5] дают:

$$d_1^T \leqslant \sum_{t=1}^{T} P(\tilde{a}_t = a^*) \left(\delta \log_2 \frac{\delta}{\delta + \varepsilon} + (1 - \delta) \log_2 \frac{1 - \delta}{1 - \delta - \varepsilon} \right)$$

Поскольку действия $\tilde{a}_t$ выбираются без информации о различии между «руками», используем тот же прием, использующий симметрию:

$$P(\tilde{a}_t \neq a^*) = \frac{A - 1}{A}$$

В оригинальной статье [1] данный переход произведен безосновательно. В более ранней [5] используется матожидание по различным средам $a^* \sim u\{1, ..., A\}$, то есть, вычисляется среднее сожаление по всем средам с различными «лучшими» руками a^*. В этом тексте этот переход верен из-за условий на перестановки в Определении 3.6 □

Теперь мы покажем, что если распределение P близко к $\tilde{P}$, то получающееся сожаление близко к сожалению неинформированного агента:

Лемма 4.3. (Ограничение на сожаление в терминах КЛ-расхождения) Пусть M — построенный многорукий бандит. Тогда для любого ε, $\delta > 0$ и для любого алгоритма обучения с подкреплением π

$$R = T \max_a \overline{r}(a) - \mathbb{E} \sum_{t=1}^{T} \overline{r}(a_t) \geqslant \varepsilon T \left(1 - \frac{1}{A} - \sqrt{\frac{1}{2} d_{\text{KL}}(\tilde{P}(z_1^T) || P(z_1^T))} \right)$$

Доказательство. По неравенству Пинскера [5]:

$$\mathbb{E} \left[\frac{n_T(a^*)}{T} - \frac{\tilde{n}_T(a^*)}{T} \right] \leqslant \sqrt{\frac{1}{2} d_{\text{KL}}(\tilde{P}(z_1^T || P(z_1^T))}$$

Далее, поскольку $\mathbb{E}\tilde{n}_T(a^*) = \frac{T}{A}$ из симметрии, получаем

$$\mathbb{E} \frac{n_t(a^*)}{T} \leqslant \sqrt{\frac{1}{2} d_{\text{KL}}(\tilde{P} || P)} + \frac{1}{A}$$

Далее рассуждения аналогичны рассуждениям в доказательстве Леммы 4.1:

$$R = \mathbb{E} \sum_{a \neq a^*} n_T(a) \varepsilon = \varepsilon (T - n_T(a^*)) \geqslant \varepsilon T \left(1 - \frac{1}{A} - \sqrt{\frac{1}{2} d_{\text{KL}}(\tilde{P} || P)} \right)$$

□

Далее, оценим величину в Лемме 4.2.

Лемма 4.4. (Ограничение на КЛ-расхождение)
Рассмотрим функцию

$$f(\delta, \varepsilon) = \delta \log_2 \frac{\delta}{\delta + \varepsilon} + (1 - \delta) \log_2 \frac{1 - \delta}{1 - \delta - \varepsilon}$$

При $\delta \in [0, \frac{1}{2}]$ и $\varepsilon \leqslant 1 - 2\delta$ значение $f(\delta, \varepsilon) \leqslant \frac{\varepsilon^2}{\delta \ln 2}$.

Доказательство. См. [8] (используется первая производная) □

Теперь докажем Теорему 1.

Доказательство. (Теорема 1) Рассмотрим

$$R = T \max_a \overline{r}(a) - \mathbb{E} \sum_{t=1}^{T} \overline{r}(a_t)$$

По Лемме 4.3

$$R \geqslant \varepsilon T \left(1 - \frac{1}{A} - \sqrt{\frac{1}{2} d_{\text{KL}}(\tilde{P}(z_1^T) || P(z_1^T))} \right)$$

По Лемме 4.2

$$d_{\text{KL}}(\tilde{P} || P) \leqslant \frac{T}{A} \left(\delta \log_2 \frac{\delta}{\delta + \varepsilon} + (1 - \delta) \log_2 \frac{1 - \delta}{1 - \delta - \varepsilon} \right)$$

Значит, по Лемме 4.4

$$d_{\text{KL}}(\tilde{P} || P) \leqslant \frac{T}{A} \frac{\varepsilon^2}{\delta \ln 2}$$

Выбираем $\varepsilon^2 = \frac{\delta A}{8T}$, подставляем последнюю оценку в оценку R:

$$R \geqslant \varepsilon T (1 - \frac{1}{A} - \frac{1}{\sqrt{16 \ln 2}}) \geqslant c_0 \sqrt{\delta A T}$$

□

Список литературы

[1] Ian Osband, Benjamin Van Roy *On Lower Bounds for Regret in Reinforcement Learning.* arxiv.org: paper in pdf

[2] David Silver *UCL Course on RL.* slides

[3] Volodymyr Mnih et al. *Playing Atari with Deep Reinforcement Learning.* paper in pdf

[4] AlphaGo article on Wikpedia

[5] Sebastien Bubeck, Nicolo Cesa-Bianchi *Regret analysis of stochastic and nonstochastic multi-armed bandit problems* paper in pdf

[6] Nicolo Cese-Bianchi, Gabor Lugosi *Prediction, Learning, and Games* book in pdf

[7] Khudanpur Sanjeev, Yaqiao Li *Information theoretic methods in statistics (scribe)* pdf

[8] Thomas Jaksch et al. *Near-optimal Regret Bounds for Reinforcement Learning* pdf