Обучение с подкреплением (Reinforcement Learning)

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

ШАД Яндекс • 5 ноября 2019

Содержание

Задача о многоруком бандите (multi-armed bandit)

Имеется множество допустимых *действий* (ручек, arm),
с различными распределениями размера *премии* (reward, payoff).

Как быстрее найти самое выгодное действие?
Какие возможны стратегии?

Задача о многоруком бандите (multi-armed bandit)

A — множество возможных *действий*
$p(r|a)$ — неизвестное распределение *премии* $r \in \mathbb{R}$ для $a \in A$
$\pi_t(a)$ — *стратегия* (policy) агента в момент t, распределение на A

Игра агента со средой:

инициализация стратегии $\pi_1(a)$;
для всех $t = 1, \dots, T, \dots$
  агент выбирает действие $a_t \sim \pi_t(a)$;
  среда генерирует премию $r_t \sim p(r|a_t)$;
  агент корректирует стратегию $\pi_{t+1}(a)$;

$$Q_t(a) = \frac{\sum_{i=1}^{t} r_i [a_i = a]}{\sum_{i=1}^{t} [a_i = a]}$$ — средняя премия в t раундах

$$Q^*(a) = \lim_{t\to\infty} Q_t(a) \to \max_{a\in A}$$ — *ценность действия a*

Примеры прикладных задач

- Рекомендация новостных статей пользователям
- Показ рекламы в Интернете
- Управление технологическими процессами
- Управление роботами
- Управление ценами и ассортиментом в сетях продаж
- Игра на бирже
- Маршрутизация в телекоммуникационных сетях
- Маршрутизация в беспроводных сенсорных сетях
- Стратегические игры: шахматы, го, Dota2, StarCraft2, ...

Задача о многоруком бандите впервые рассмотрена в статье
H. Robbins. Some aspects of the sequential design of experiments.
Bulletin of the American Mathematics Society, 58:527–535, 1952.

Жадная стратегия

Множество действий с максимальной текущей оценкой ценности:

$$A_t = \operatorname{Arg}\max_{a \in A} Q_t(a)$$

Жадная стратегия — выбирать любое действие из A_t:

$$\pi_t(a) = \frac{1}{|A_t|}[a \in A_t]$$

Недостаток жадной стратегии — по некоторым действиям a можем так и не набрать статистику для оценки $Q_t(a)$.

Компромисс «изучение–применение» (exploration–exploitation)
ε-*жадная стратегия*:

$$\pi_t(a) = \frac{1-\varepsilon}{|A_t|}[a \in A_t] + \frac{\varepsilon}{|A|}$$

Эвристика: параметр ε уменьшать со временем.

Стратегия softmax (распределение Гиббса)

Мягкий вариант компромисса «изучение–применение»:
чем больше $Q_t(a)$, тем больше вероятность выбора a:

$$\pi_t(a) = \frac{\exp\bigl(\frac{1}{\tau} Q_t(a)\bigr)}{\sum\limits_{b\in A} \exp\bigl(\frac{1}{\tau} Q_t(b)\bigr)}$$

где τ — параметр *температуры*,
при $\tau \to 0$ стратегия стремится к жадной,
при $\tau \to \infty$ — к равномерной, т.е. чисто исследовательской

Эвристика: параметр τ уменьшать со временем.

Какая из стратегий лучше?
— зависит от конкретной задачи,
— решается в эксперименте

Метод UCB (upper confidence bound)

Выбор действия с максимальной верхней оценкой ценности:

$$A_t = \operatorname*{Arg\,max}_{a \in A} \left(Q_t(a) + \delta \sqrt{\frac{2 \ln t}{k_t(a)}} \right),$$

где $k_t(a) = \sum\limits_{i=1}^{t} [a_i = a]$, δ — параметр exr/exp-компромисса.

Интерпретация:
чем меньше $k_t(a)$, тем менее исследована стратегия,
тем выше должна быть вероятность выбрать a;

чем больше δ, тем стратегия более исследовательская.

Эвристика: параметр δ уменьшать со временем.

P. Auer, N. Cesa-Bianchi, P. Fischer. Finite-time analysis of the multiarmed bandit problem, Machine Learning, 2002.

Модельные эксперименты в обучении с подкреплением

«10-рукая испытательная среда»:

Генерируется 2000 задач, в каждой задаче

$|A| = 10,$

$p(r|a) = \mathcal{N}(r; Q^*(a), 1),$

$Q^*(a) \sim \mathcal{N}(0, 1).$

Строятся графики зависимости

— средней премии (average reward),

— доли оптимальных действий (% optimal action),

от числа шагов t, усреднённые по 2000 задачам.

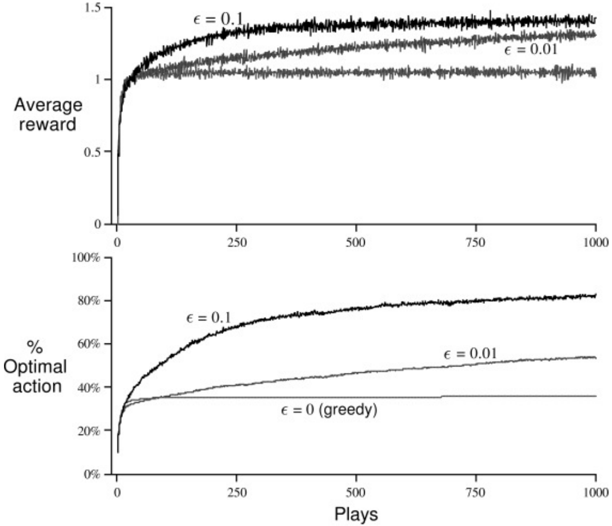
Richard Sutton, Andrew Barto. Reinforcement Learning: An Introduction. The MIT Press. 1998, 2004.

http://webdocs.cs.ualberta.ca/~sutton/book/ebook/the-book.html

Русский перевод:

Р. Саттон, Э. Барто. Обучение с подкреплением. Изд-во «Бином». 2011.

Сравнение жадных и ε-жадных стратегий



Рекуррентная формула для эффективного вычисления средних

Общая формула вычисления Q_t для корректировки стратегии:

$$Q_{t+1}(a) = (1-\alpha_t)Q_t(a) + \alpha_t r_{t+1} = Q_t(a) + \alpha_t\bigl(r_{t+1} - Q_t(a)\bigr)$$

При $\alpha_t = \frac{1}{k_t(a)+1}$ это среднее арифметическое, $k_t(a) = \sum\limits_{i=1}^{t}[a_i = a]$

При $\alpha_t = \text{const}$ это экспоненциальное скользящее среднее

Условие сходимости к среднему:

$$\sum_{t=1}^{\infty} \alpha_t = \infty, \qquad \sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

Среднее арифметическое — для стационарных задач

Экспоненциальное скользящее среднее — для нестационарных (в этом случае сходимости нет, но она и не нужна)

Экспоненциальное скользящее среднее (напоминание)

Задача прогнозирования временного ряда $y_0, \dots, y_t, \dots$:
— простейшая регрессионная модель — константа $y_t = c$,
— наблюдения учитываются с весами, убывающими в прошлое,
— прогноз $\hat{y}_{t+1}$ методом наименьших квадратов:

$$\sum_{i=0}^{t} w_{t-i}(y_i - c)^2 \to \min_c, \quad w_i = \beta^i, \quad \beta \in (0,1)$$

Аналитическое решение — формула Надарая-Ватсона:

$$c \equiv \hat{y}_{t+1} = \frac{\sum_{i=0}^{t} \beta^i y_{t-i}}{\sum_{i=0}^{t} \beta^i}$$

Запишем аналогично $\hat{y}_t$, оценим $\sum_{i=0}^{t} \beta^i \approx \sum_{i=0}^{\infty} \beta^i = \frac{1}{1-\beta}$,

получим $\hat{y}_{t+1} = \hat{y}_t \beta + (1-\beta) y_t$, заменим $\alpha = 1 - \beta$:

$$\hat{y}_{t+1} = (1-\alpha)\hat{y}_t + \alpha y_t = \hat{y}_t + \alpha(y_t - \hat{y}_t)$$

Метод сравнения с подкреплением (reinforcement comparison)

Идея: использовать в softmax не сами значения премий, а их разности со средней (эталонной) премией:

$\bar{r}_{t+1} = \bar{r}_t + \alpha(r_t - \bar{r}_t)$ — *средняя премия* по всем действиям

$p_{t+1}(a_t) = p_t(a_t) + \beta(r_t - \bar{r}_t)$ — *предпочтения* действий

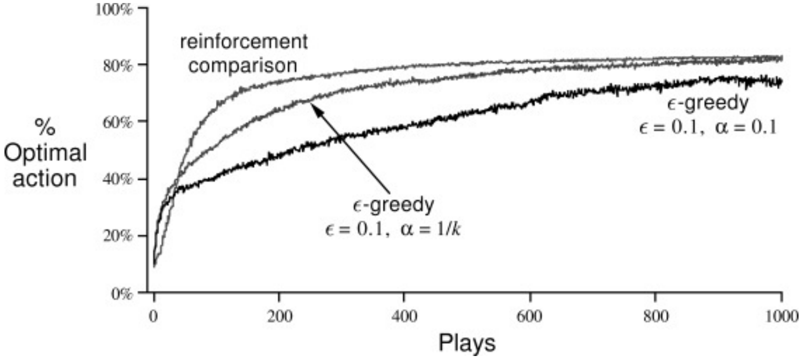
$\pi_{t+1}(a) = \dfrac{\exp\bigl(p_{t+1}(a)\bigr)}{\sum\limits_{b \in A} \exp\bigl(p_{t+1}(b)\bigr)}$ — softmax-стратегия агента

Эвристика: оптимистично завышенное начальное $\bar{r}_0$ стимулирует изучающие действия в начале

Экспериментальный факт: сравнение с подкреплением сходится быстрее ε-жадных стратегий.

Сравнение с подкреплением лучше ε-жадных стратегий

Эксперимент с 10-рукой испытательной средой:



Richard Sutton, Andrew Barto. Reinforcement Learning: An Introduction. The MIT Press. 1998, 2004.
Р. Саттон, Э. Барто. Обучение с подкреплением. Изд-во «Бином». 2011.

Метод преследования (pursuit) жадной стратегии

Вместо собственно *жадной стратегии*

$$\pi_{t+1}(a) = \frac{[a \in A_t]}{|A_t|}$$

предлагается *преследование* (сглаживание) жадной стратегии:

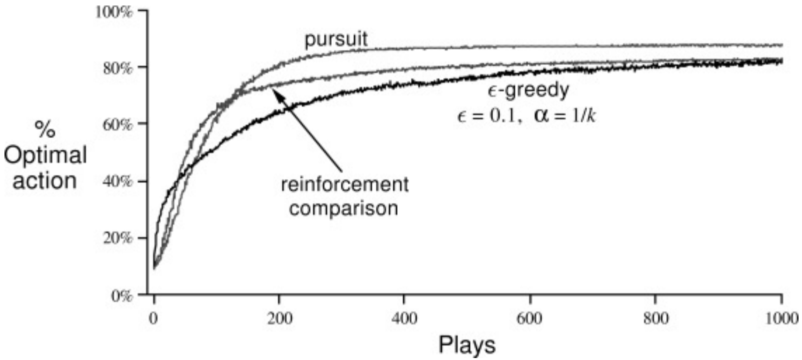
$$\pi_{t+1}(a) = \pi_t(a) + \beta \left(\frac{[a \in A_t]}{|A_t|} - \pi_t(a) \right)$$

Эвристика: начальное $\pi_0(a)$ можно взять равномерным.

Экспериментальный факт: метод преследования, сравнение с подкреплением и ε-жадные стратегии имеют каждый свою область применения.

Стратегия преследования ещё лучше

Эксперимент с 10-рукой испытательной средой:



Richard Sutton, Andrew Barto. Reinforcement Learning: An Introduction. The MIT Press. 1998, 2004.
Р. Саттон, Э. Барто. Обучение с подкреплением. Изд-во «Бином». 2011.

Постановка задачи в случае, когда агент влияет на среду

A — конечное множество возможных *действий* (action)
S — конечное множество состояний среды (state)

Игра агента со средой:

инициализация стратегии $\pi_1(a \,|\, s)$ и состояния среды s_1;
для всех $t = 1, \dots T, \dots$
- агент выбирает действие $a_t \sim \pi_t(a \,|\, s_t)$;
- среда генерирует премию $r_t \sim p(r \,|\, a_t, s_t)$
- и новое состояние $s_{t+1} \sim p(s \,|\, a_t, s_t)$;
- агент корректирует стратегию $\pi_{t+1}(a \,|\, s)$;

Это *марковский процесс принятия решений* (МППР), если

$$\mathsf{P}\bigl(s_{t+1}, r_t \,|\, s_t, a_t, r_{t-1}, s_{t-1}, a_{t-1}, r_{t-2}, \dots, s_1, a_1\bigr) = \\ = \mathsf{P}\bigl(s_{t+1}, r_t \,|\, s_t, a_t\bigr)$$

Понятия выгоды и ценности действия

Суммарная выгода (return):

$$R_t = r_t + r_{t+1} + \cdots + r_{t+k} + \cdots$$

Обобщение — *дисконтированная выгода* (discounted return):

$$R_t = r_t + \gamma r_{t+1} + \cdots + \gamma^k r_{t+k} + \cdots$$

где $\gamma \in [0, 1]$ — коэффициент дисконтирования,
$1 + \gamma + \gamma^2 + \cdots = \frac{1}{1-\gamma}$ — горизонт дальновидности агента.

Функция ценности действия a в состоянии s при стратегии π:

$$Q^\pi(s, a) = \mathsf{E}_\pi(R_t \,|\, s_t = s,\ a_t = a) = \mathsf{E}_\pi\Bigl(\sum_{k=0}^{\infty} \gamma^k r_{t+k} \Bigm| s_t = s,\ a_t = a\Bigr)$$

E_π — мат.ожидание при условии, что агент следует стратегии π

Жадные стратегии максимизации ценности

Рекуррентная формула для ценности действия $Q^\pi(s,a)$:

$$\begin{aligned} Q^\pi(s,a) &= \mathsf{E}_\pi\bigl(\textstyle\sum_{k=0}^\infty \gamma^k r_{t+k} \bigm| s_t = s,\ a_t = a\bigr) = \\ &= \mathsf{E}_\pi\bigl(r_t + \gamma \textstyle\sum_{k=0}^\infty \gamma^k r_{t+k+1} \bigm| s_t = s,\ a_t = a\bigr) = \\ &= \mathsf{E}_\pi\bigl(r_t + \gamma Q^\pi(s_{t+1}, a_{t+1}) \bigm| s_t = s,\ a_t = a\bigr) \end{aligned}$$

Уравнение Беллмана для оптимальной функции ценности Q^*:

$$Q^*(s,a) = \mathsf{E}_\pi\bigl(r_t + \gamma \max_{a' \in A} Q^*(s_{t+1}, a') \bigm| s_t = s,\ a_t = a\bigr)$$

Утв. *Жадная стратегия* π относительно $Q^*(s,a)$ «выбирать то действие, на котором достигается максимум в уравнениях Беллмана», является оптимальной:

$$A_t = \operatorname{Arg}\max_{a \in A} Q^*(s_t, a)$$

Метод Q-обучения

Аппроксимируем оптимальную функцию ценности действия экспоненциальным скользящим средним:

$$Q(s_t, a_t) := Q(s_t, a_t) + \alpha_t \bigl(r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \bigr)$$

инициализация стратегии $\pi_1(a \,|\, s)$ и состояния среды s_1;
для всех $t = 1, \dots T, \dots$
- агент выбирает действие $a_t \sim \pi_t(a \,|\, s_t)$, например, $a_t := \arg\max_a Q(s_t, a)$ — жадная стратегия;
- среда генерирует $r_t \sim p(r \,|\, a_t, s_t)$ и $s_{t+1} \sim p(s \,|\, a_t, s_t)$;
- $Q(s_t, a_t) := Q(s_t, a_t) + \alpha_t \bigl(r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \bigr)$;

Утв. Если α_t уменьшается ($\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$), и все s посещаются бесконечное число раз, то $Q \xrightarrow{\text{пн}} Q^*$, $t \to \infty$

Градиентная оптимизация стратегии (policy gradient)

Обобщение:

- $p(x|\theta) = \pi(a \,|\, s, \theta)$ — параметризованная стратегия агента
- x — описание текущего состояния и действия (s, a)
- $f(x)$ — функция ценности или её оценка

Задача: оптимизировать f по вектору параметров стратегии θ:

$$\mathsf{E}_\pi f(x) \equiv \mathsf{E}_{x \sim p(x|\theta)} f(x) \equiv \mathsf{E}_{x|\theta} f(x) \to \max_\theta$$

Градиентный метод: $\theta^{(t+1)} := \theta^{(t)} + \eta \nabla_\theta \mathsf{E}_{x|\theta^{(t)}} f(x)$;

$$\nabla_\theta \mathsf{E}_{x|\theta} f(x) = \nabla_\theta \sum_x f(x) p(x|\theta) = \sum_x f(x) \nabla_\theta p(x|\theta) =$$

$$= \sum_x f(x) p(x|\theta) \frac{\nabla_\theta p(x|\theta)}{p(x|\theta)} = \mathsf{E}_{x|\theta} \bigl[f(x) \nabla_\theta \ln p(x|\theta) \bigr]$$

Градиентная оптимизация стратегии (policy gradient)

Заменяем $\mathsf{E}_\pi R_t$ эмпирической оценкой и накапливаем вектор градиента g_t с помощью экспоненциальной скользящей средней:

$$g_{t+1} := g_t + \alpha_t \bigl(R_t \nabla_\theta \ln \pi(a_t \,|\, s_t, \theta) - g_t \bigr)$$

Фактически, это SGD для максимизации log-правдоподобия:

$$\sum_t R_t \ln \pi(a_t \,|\, s_t, \theta) \;\to\; \max_\theta$$

Основные отличия от максимизации log-правдоподобия:

- вместо предсказания меток классов y_t — действия a_t
- вместо обучения по бинарным y_t — вещественные R_t

Что можно использовать в качестве R_t:

- функции ценности $\bar{r}_t$, как-либо усредняемые
- оценку преимущества (advantage) $\mathsf{E}_\pi(R_t \,|\, s_t = s) - \bar{r}_t$

Постановка задачи в случае, когда имеется информация о среде

A — множество возможных *действий*

X — пространство контекстов, описаний состояния среды

$x_{ta} \in X$ — состояние среды в раунде t в случае выбора $a \in A$

$p(r \mid a, x)$ — неизвестное распределение *премии* $r \in \mathbb{R}$ для $a \in A$

$\pi_t(a \mid x)$ — *стратегия* агента в момент t, распределение на A

Игра агента со средой (contextual bandit):

инициализация стратегии $\pi_1(a)$;

для всех $t = 1, \dots T, \dots$

- агенту сообщается контекст x_{ta} для всех $a \in A$;
- агент выбирает действие $a_t \sim \pi_t(a \mid x_{ta})$;
- среда генерирует премию $r_t \sim p(r \mid a_t, x_{ta})$;
- агент корректирует стратегию $\pi_{t+1}(a \mid x)$;

Context-free bandit — когда $\pi_t(a \mid x) = \pi_t(a)$, т.е. не зависит от x.

Регрессия с инкрементным обучением и доверительной оценкой

$r(a,x)$ — функция премии за действие a в контексте x,
$\hat{r}(a,x)$ — регрессионная оценка этой функции,
$UCB(a,x)$ — верхняя оценка отклонения $\hat{r} - r$,
δ — параметр (чем больше, тем больше exploration).

Игра агента со средой (contextual bandit):

инициализация стратегии $\pi_1(a)$;
для всех $t = 1, \dots T, \dots$
- агенту сообщается контекст x_{ta} для всех $a \in A$;
- агент выбирает действие
  $$a_t = \arg\max_{a \in A} \Big(\hat{r}(a, x_{ta}) + \delta UCB(a, x_{ta}) \Big);$$
- среда генерирует премию $r_t = r(a_t, x_{ta_t})$;
- регрессия $\hat{r}(a,x)$ дообучается на точке $(a_t, x_{ta_t}; r_t)$;

Пример. Рекомендация новостных статей пользователям

Агент — рекомендательная система для персонализации показов новостных статей (Yahoo! Today).

F1..F4 — позиции для показа заголовков новостей.

A — новостные статьи, действия системы;
$x_{ta} \in X$ — признаковое описание пары (u_t, a);
u_t — пользователь, которому агент даёт рекомендацию;
$r_t \in \{0, 1\}$ — пользователь u_t кликнул на предложенную статью;
$Q_t(a)$ — средняя премия, CTR (click-through rate) статьи.

Цель — повышение среднего CTR и «счастья пользователя».

Lihong Li, Wei Chu, John Langford, Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. WWW-2010.

Линейная модель премий и гребневая регрессия

Пусть $x_{ta} \in X = \mathbb{R}^n$, $w \in \mathbb{R}^n$.
Линейная модель премий для действия $a \in A$:

$$Q^*(a) = \mathsf{E}\bigl[r_t \,|\, x_{ta}\bigr] = \langle x_{ta}, w_a \rangle.$$

Гребневая регрессия: обучение w_a для действия a в момент t:

$$\sum_{i=1}^{t} \bigl[a_i = a\bigr] \bigl(\langle x_{ia}, w_a \rangle - r_i\bigr)^2 + \frac{\tau}{2} \|w_a\|^2 \;\to\; \min_{w_a}.$$

$w_a = \bigl(F_a^{\mathsf{T}} F_a + \tau I_n\bigr)^{-1} F_a^{\mathsf{T}} y_a$ — решение задачи МНК, где

$F_a = \bigl(x_{ia}\bigr)_{i=1:\, a_i=a}^{t}$ — $\ell \times n$-матрица объекты–признаки,

$y_a = \bigl(r_i\bigr)_{i=1:\, a_i=a}^{t}$ — $\ell \times 1$-вектор ответов,

$\ell = k_t(a) = \sum_{i=1}^{t} [a_i = a]$ — объём обучающей выборки.

LinUCB: линейная модель с верхней доверительной оценкой

Доверительный интервал с коэффициентом доверия $1-\alpha$ для линейной модели регрессии:

$$y = \langle x, w \rangle \pm \hat{\sigma} Z_\alpha \sqrt{x^{\mathsf{T}} (F^{\mathsf{T}} F)^{-1} x},$$

$Z_\alpha \equiv t_{\ell-n, 1-\frac{\alpha}{2}}$ — квантиль распределения Стьюдента,
$\hat{\sigma}^2 = \frac{1}{\ell-n} RSS$ — оценка дисперсии отклика y.

Стратегия выбора действия с максимальной верхней оценкой ценности UCB (upper confidence bound):

$$A_t = \operatorname*{Arg\,max}_{a \in A} \Bigl(\langle x_{ta}, w_a \rangle + \delta \hat{\sigma} Z_\alpha \sqrt{x_{ta}^{\mathsf{T}} \bigl(F_a^{\mathsf{T}} F_a + \tau I_n \bigr)^{-1} x_{ta}} \Bigr).$$

Чем больше параметр δ, тем больше исследования.

LinUCB: особенности реализации и обобщения

- *Инкрементный алгоритм* пересчёта w_a и матрицы $\bigl(F_a^\top F_a + \tau I_n\bigr)^{-1}$ при добавлении каждой строки в F_a.
- Гибридная линейная модель $Q^*(a) = \langle \tilde{x}_t, v\rangle + \langle x_{ta}, w_a\rangle$, где $\tilde{x}_t$ — часть контекста, не зависящая от действия a.
- «Сырые признаки»:
  пользователи: 12 соцдем, 200 география, ~1000 категорий,
  статьи: ~100 категорий.
- Используется кластеризация и *понижение размерности*:
  $\dim w_a = 6$, $\dim v = 36$.
- Можно было бы использовать любую другую модель с инкрементным обучением и доверительными оценками.

Lihong Li, Wei Chu, John Langford, Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. WWW-2010.

Оценивание модели по историческим данным

Проблема off-line оценивания стратегии π:
исторические данные накоплены при использовании другой стратегии (logging policy) $\pi_0(a)$, отличной от π

Идея:
для оценивания $Q_t(a)$ отбираются только те события (x_{ta}, a, r_t), для которых стратегии π и π_0 выбирали одинаковое действие:

$$a = \arg\max_a \pi(a, x_{ta}) = \arg\max_a \pi_0(a)$$

(нужны очень большие данные или сходство стратегий)

Утв. Если $\pi_0(a)$ — равномерное распределение, то оценка $Q_t(a)$ по отобранной выборке является несмещённой.

Lihong Li, Wei Chu, John Langford, Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. WWW-2010.

Резюме в конце лекции

- В обучении с подкреплением нет ответов учителя, есть только ответная реакция среды
- В марковских процессах принятия решений накапливается информация о ценности действий в каждом состоянии
- Если состояний слишком много, то вводится непрерывная параметризация и применяется Policy Gradient
- В контекстных бандитах используются модели машинного обучения, удовлетворяющие двум требованиям:
  - существует эффективный инкрементный метод обучения
  - существуют доверительные оценки средней премии $Q^t(a)$
- Компромисс «изучение–применение» при любом обучении с подкреплением подбирается экспериментальным путём
- Объём исследовательских действий приходится уменьшать в случае конечного горизонта игры