Лекция 11. Многорукие бандиты

Александр Юрьевич Авдюшенко

МКН СПбГУ

28 апреля 2022

Факультет
математики
и компьютерных
наук
СПбГУ

Пятиминутка

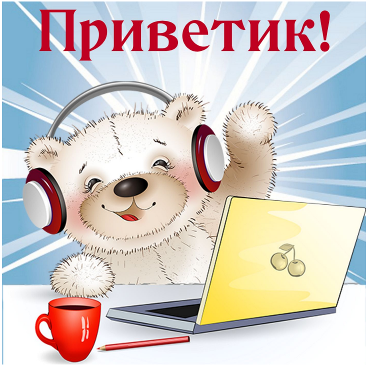
- Назовите три основных подхода к построению моделей ранжирования
- Расшифруйте аббревиатуру RMSE
- Перечислите нетривиальные свойства, влияющие на качество рекомендаций, которые трудно измерять

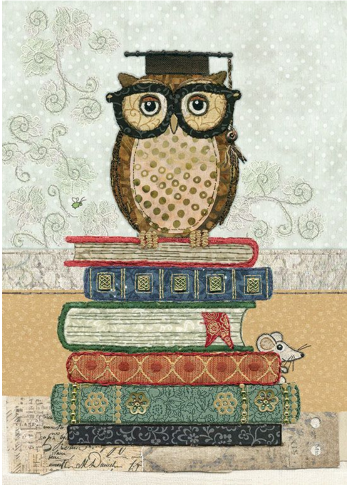
Постановка задачи

- До этого момента мы либо восстанавливали функцию по обучающей выборке (X, Y) (supervised learning), либо искали структуру в наборе объектов X (unsupervised learning)
- Как происходит обучение в реальной жизни? Обычно делаем какое-то действие и получаем результат, постепенно обучаясь

Ещё мотивация

Например, нам нужно выбрать главную страницу сайта магазина открыток, чтобы привлечь пользователя





Какие подходы есть?

- A/B тестирование — потенциально плохие варианты видят многие пользователи
- многорукие бандиты — частный случай обучения с подкреплением

Однорукий бандит Бернулли

Вероятность выиграть $\theta = 0.05$

Многорукие бандиты

Вероятность выиграть $\theta = 0.02$

Вероятность выиграть $\theta = 0.01$(min)

Вероятность выиграть $\theta = 0.05$

Вероятность выиграть $\theta = 0.1$(max)

Мы не знаем истинные вероятности, но хотим придумать стратегию, максимизирующую выигрыш (награду).

Математическая постановка задачи

Даны возможные действия

$$x_1, \dots, x_n$$

На очередной итерации t при каждом совершаемом действии x_i^t мы получаем ответ

$$y_i^t \sim q(y|x_i^t, \Theta),$$

который приносит нам награду **reward**

$$r_i^t = r(y_i^t)$$

Существует оптимальное действие x_{i^*} (иногда $x_{i_t^*}$)

$$\forall i: \;\; E(r_{i_t^*}^t) \geq E(r_i^t)$$

Мера качества

Вопрос

Как оценивать различные стратегии?

Мера качества

Вопрос

Как оценивать различные стратегии?

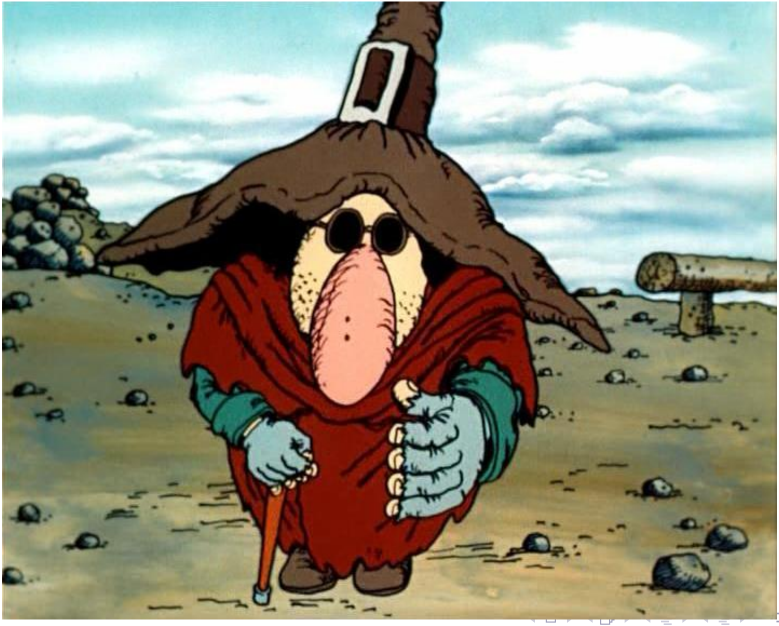
Мерой качества алгоритма многоруких бандитов a обычно является **regret**

$$R(a) = \sum_{t=1}^{T} \Big(E(r^t_{i^*_t}) - E(r^t_{i^a_t}) \Big)$$

В синтетических условиях (когда знаем вероятности) можно рассмотреть

$$E\left(R\right) = \int_{\Theta} R(a) d\Theta$$

ε-жадный подход



Жадный подход

- на основе исторических данных оценить распределение параметров модели $P(\Theta|X)$
- всегда использовать действие, ведущее к наибольшему выигрышу в среднем на основании полученного распределения

$$x_k = \arg\max_{x_k} E_{y|x_k,\hat{\Theta}}\left(r(y)\right),$$

где $\hat{\Theta} = E\Theta$

Формула Байеса

напоминание

$$P(\Theta|Y) = \frac{Q(Y|\Theta)P(\Theta)}{Q(Y)}$$

Формула Байеса

напоминание

$$P(\Theta|Y) = \frac{Q(Y|\Theta)P(\Theta)}{Q(Y)}$$

$$P_t(\Theta|Y_t) = \frac{q(y^t|x^t,\Theta)P_{t-1}(\Theta|Y_{t-1})}{q(y^t|x^t)}$$

Формула Байеса

напоминание

$$P(\Theta|Y) = \frac{Q(Y|\Theta)P(\Theta)}{Q(Y)}$$

$$P_t(\Theta|Y_t) = \frac{q(y^t|x^t,\Theta)P_{t-1}(\Theta|Y_{t-1})}{q(y^t|x^t)}$$

Вопрос

Что получится для бандита Бернулли?

Пример. Применение формулы Байеса для бандита Бернулли

$Q(Y|\Theta)$ представляется в виде $Q(Y|\theta_1), Q(Y|\theta_2), \dots, Q(Y|\theta_k)$, где

$$Q(Y|\theta) = C_{T_i}^{\sum\limits_{t \in |T_i|} r_i^t} \theta^{\sum\limits_{t \in |T_i|} r_i^t} (1-\theta)^{|T_i| - \sum\limits_{t \in |T_i|} r_i^t}$$

$\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$, θ_k — коэффициент для k-го бандита (параметр распределения Бернулли)

$$y_k \sim \mathrm{Bernoulli}(\theta_k)$$

Распределение параметров распределений Бернулли — Бета-распределение:

$$p(\theta_k) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \theta_k^{\alpha_k - 1} (1 - \theta_k)^{\beta_k - 1}$$

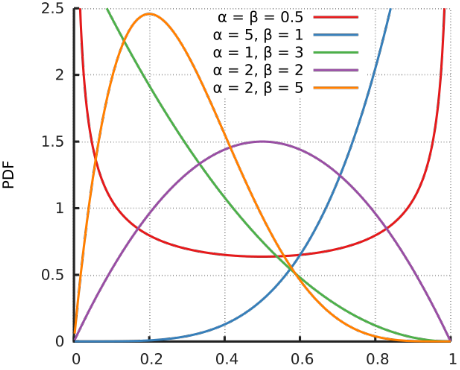
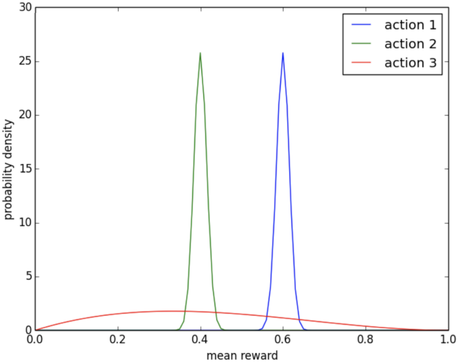


Схема обновления распределения параметров после t-го шага:

$$(\alpha_k, \beta_k) \leftarrow (\alpha_k, \beta_k) + (r_k^t, 1 - r_k^t)$$

Проблемы с жадностью



Апостериорные распределения после того, как:

- бандит 1 был использован 1000 раз и дал выигрыш 600 раз
- бандит 2 был использован 1000 раз и дал выигрыш 400 раз
- бандит 3 был использован 3 раза и дал выигрыш 1 раз

ε-жадный алгоритм

- на основе исторических данных оценить распределение параметров модели Θ, $P(\Theta|X)$
- с вероятностью ε выбирать случайное действие
- с вероятностью $1-\varepsilon$ выбирать действие, ведущее к наибольшему выигрышу в среднем на основании полученного распределения

  $$x_k = \arg\max_{x_k} E_{y|x_k,\hat{\Theta}}[r(y)],$$

  где $\hat{\Theta} = E\Theta$
- повторить

Недостатки

- изучение новых возможностей полностью случайно и никак не зависит от уже известной информации
- непонятно, когда нужно прекращать исследовать и начинать использовать

Сэмплирование Томпсона

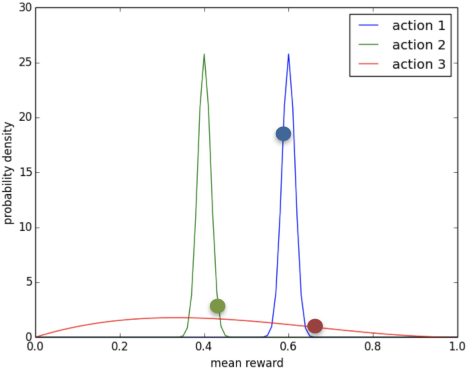
- на основе исторических данных оценить распределение параметров модели Θ, $P(\Theta|X)$
- семплировать один раз $\hat{\Theta}$ из его текущего распределения $P(\Theta|X)$
- выбирать действие, ведущее к наибольшему выигрышу в среднем на основании полученного распределения

$$x_k = \arg\max_{x_k} E_{y|x_k,\hat{\Theta}}[r(y)]$$

- повторить

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband and Zheng Wen «A Tutorial on Thompson Sampling», 2018

Сэмплирование Томпсона. Пример



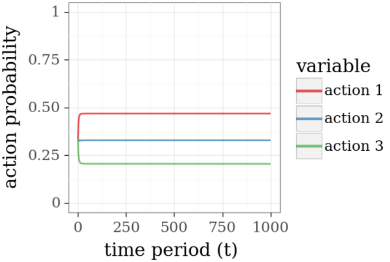
Сэмплировали, например:

- $\theta_1 = 0.59$
- $\theta_2 = 0.45$
- $\theta_3 = 0.67$

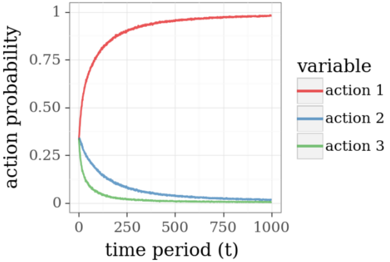
Выбираем action3

Сходимость к оптимальным значениям

$\theta_1 = 0.9, \theta_2 = 0.8, \theta_3 = 0.7$, априорное распределение равномерное. Проводим 1000 независимых испытаний по 1000 шагов t в каждом.



(a) greedy algorithm



(b) Thompson sampling

Figure 3.1: Probability that the greedy algorithm and Thompson sampling selects an action.

В каждой точке — доля испытаний, в которой выбрано соотвествующее действие

Теоретические оценки

Сверху

$$\max_{\theta'} \mathbb{E}[\text{Regret}(T) | \theta = \theta'] = O\left(\sqrt{KT \log(T)}\right)$$

Снизу Показано, что существует распределение такое, что

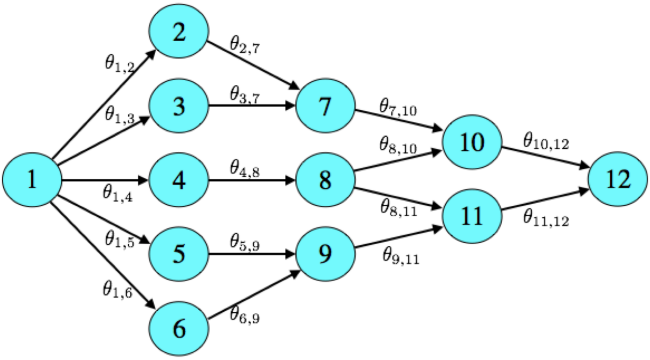
$$\max_{\theta'} \mathbb{E}[\text{Regret}(T) | \theta = \theta'] = \Omega\left(\sqrt{KT}\right)$$

Пример. Путь в графе

$G(V, E)$

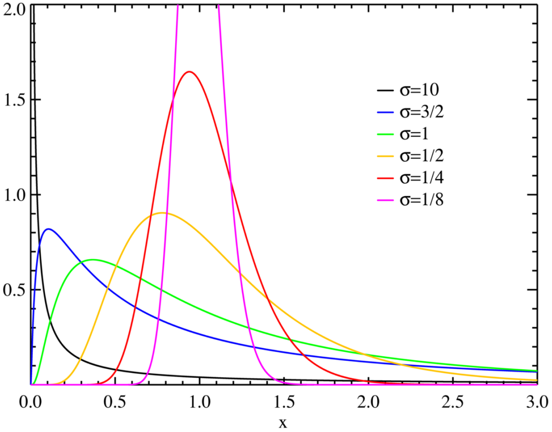
$x_t = e_{1'} \to e_{2'} \cdots \to e_{U'}$

$r_t = -\sum_{e \in x_t} y_e^t$



Пример. Путь в графе

$\theta_e \sim LN(\mu_e, \sigma_e^2)$ (логнормальное распределение)
Математическое ожидание $E[\theta_e] = e^{\mu_e + \sigma_e^2/2}$

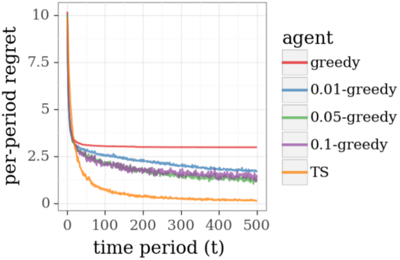


Плотность вероятности, $\mu = 0$

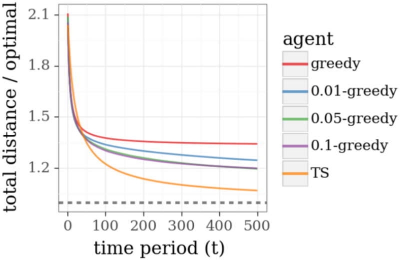
$y_e^t \sim LN(\ln(\theta_e) - s^2/2, s^2)$, $E[y_e^t|\theta_e] = \theta_e$

Обновление распределений параметров

$$(\mu_e, \sigma_e^2) \leftarrow \left(\frac{\frac{1}{\sigma_e^2}\mu_e + \frac{1}{\tilde{\sigma}_e^2}\left(\ln(y_e^t) + \frac{\tilde{\sigma}_e^2}{2}\right)}{\frac{1}{\sigma_e^2} + \frac{1}{\tilde{\sigma}_e^2}}, \frac{1}{\frac{1}{\sigma_e^2} + \frac{1}{\tilde{\sigma}_e^2}} \right)$$



(a) regret



(b) cumulative travel time vs. optimal

Upper Confidence Bound

- на основе исторических данных оценить распределение параметров модели Θ, $P(\Theta|X)$
- выбираем действие

$$x_k = \arg\max_{x_k}(\mu_k^t + c \cdot u_k^t), \ \hat{\Theta} = E\Theta$$

- повторяем

u_k^t характеризует нашу неуверенность в действии x_k и не обязательно выражается через апостериорное распределение
Например:

$$u_k^t = \sqrt{\frac{\log(t)}{|T_k|}}$$

Расширения и эвристики для семплирования Томпсона

Возможные проблемы

- бизнес-ограничения
- контекст
- нестационарность
- многопоточность (например, много пользователей)
- подсчет апостериорного распределения

Контекст

Пусть $y_i^t \sim q(y|x_i^t, z^t, \Theta)$
В таком случае можно:

- $X = \{(x_i, z_j)\}$
- контекстуальные бандиты (LinUCB) — идея метода в том, что награда теперь зависит от контекста z^t, в качестве которого может выступать, например, вектор пользователя

Kuan-Hao Huang and Hsuan-Tien Lin. Linear Upper Confidence Bound Algorithm for Contextual Bandit Problem with Piled Rewards https://www.csie.ntu.edu.tw/ htlin/paper/doc/pakdd16piled.pdf

Нестационарность

Пусть $y_i^t \sim q_t(y|x_i^t, \Theta)$
В таком случае можно:

- аппроксимировать апостериорную вероятность по последней истории
- добавляем неуверенность

$$P_t(\Theta|Y_t) = \frac{q(y^t|x^t, \Theta) P_{t-1}^{1-\gamma}(\Theta|Y_{t-1}) \overline{P}^{\gamma}(\Theta)}{q(y^t|x^t)}$$

к примеру в задаче про бандитов Бернулли

$$(\alpha_k, \beta_k) \leftarrow \begin{cases} ((1-\gamma)\alpha_k + \gamma\overline{\alpha}, (1-\gamma)\beta_k + \gamma\overline{\beta}), & x_t \neq k \\ ((1-\gamma)\alpha_k + \gamma\overline{\alpha} + r_t, (1-\gamma)\beta_k + \gamma\overline{\beta} + 1 - r_t), & x_t = k \end{cases}$$

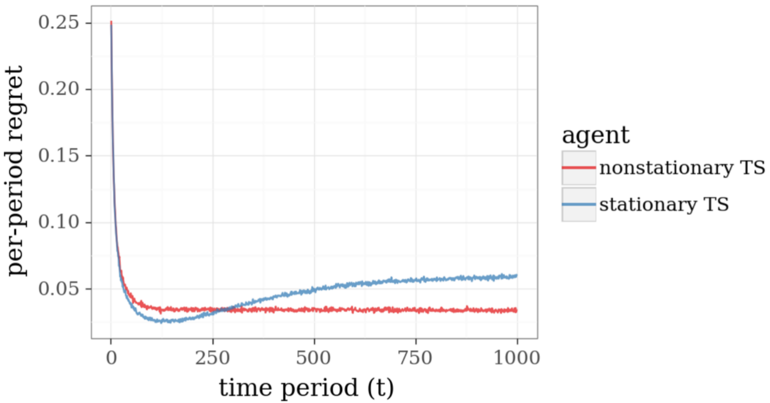
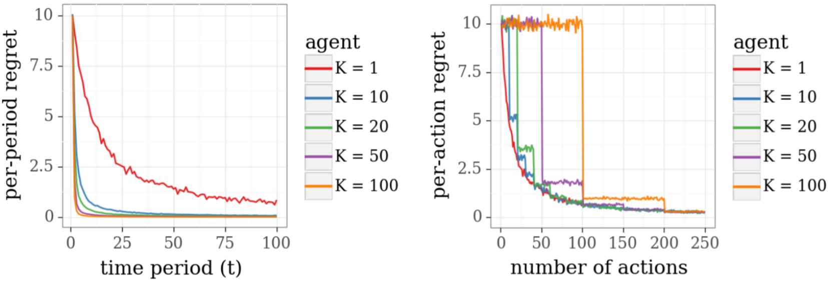


Figure 6.3: Comparison of TS versus nonstationary TS with a nonstationary Bernoulli bandit problem.

Многопоточность



(a) per-action regret over time **(b)** per-action regret over actions

Figure 6.4: Performance of concurrent Thompson sampling.

Снова аппроксимации апостериорного распределения

- сэмплирование Гиббса
- Langevin Monte Carlo
- приближенный вариационный вывод (аппроксимация Лапласа)
- bootstrap

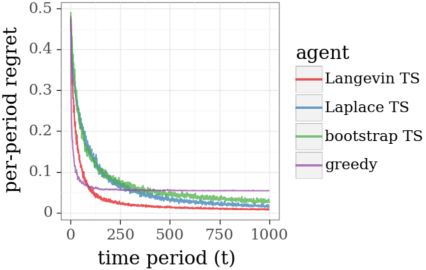


Figure 5.1: Regret experienced by approximation methods applied to the path recommendation problem with binary feedback.

Области применения

- revenue management
- оптимизация сайтов
- интернет-реклама
- рекомендательные системы
- продвижение контента
- обучение нейронных сетей

Резюме

- многорукие бандиты — частный случай обучения с подкреплением
- ε-жадный алгоритм и сэмплирование Томпсона
- UCB — upper confidence bound
- способы аппроксимации апостериорного распределения

Резюме

- многорукие бандиты — частный случай обучения с подкреплением
- ε-жадный алгоритм и сэмплирование Томпсона
- UCB — upper confidence bound
- способы аппроксимации апостериорного распределения

Что ещё можно посмотреть?

- Видео «A short introduction to multi-armed bandits»
- Книга «Introduction to Multi-Armed Bandits», Aleksandrs Slivkins
- Книга «A Tutorial on Thompson Sampling»