MAB ALGORITHM DISCUSSION

FEI WU

NON CONTEXTUAL BANDIT

- 贪心法：Epsilon-greedy strategy， Epsilon-decreasing strategy…
- Probability matching: Thomason sampling/Bayesian Bandits
- Pricing strategies
- Strategies with 伦理限制(多用于医学)

CONTEXTUAL BANDIT

- Upper confidence bound algorithms
- Deep learning method: 神经网络，random forest

Adversarial bandit

Infinite-armed bandit (无限臂)

GREEDY ALGORITHM(贪心法)

- Epsilon-greedy strategy

贪心法每次有ϵ几率进行探索，$1-\epsilon$的几率使用已知的最优解。探索时每条臂有相同的概率被选取。

ϵ的值可以在学习过程变化，比较常见的是epsilon-decreasing 和 Adaptive epsilon-greedy strategy based on value differences (VDBE)。会根据不同的算法来调整ϵ的值以适应不同情况。[1]

贪心法也可以适用于CONTEXTUAL BANDIT， 其方法为对相应的context 计算其对应的ϵ[2]

1. Tokic, Michel (2010), "Adaptive ε-greedy exploration in reinforcement learning based on value differences", KI 2010: Advances in Artificial Intelligence (PDF), Lecture Notes in Computer Science, 6359
2. Bouneffouf, D.; Bouzeghoub, A.; Gançarski, A. L. (2012). "A Contextual-Bandit Algorithm for Mobile Context-Aware Recommender System". Neural Information Processing. Lecture Notes in Computer Science.

PROBABILITY MATCHING

- Thompson sampling

Thompson sampling 算法一般使用beta distribution 来模拟奖赏的的概率分布。

Reward~beta(A, B), 如果是binary case :

R = 1→ A = A+1

R = 0→B = B+1

Thomason sampling 也可以应用为contextual bandit （amazon paper）

- Other probability match 算法[1]

1. Bouneffouf, D.; Bouzeghoub, A.; Gançarski, A. L. (2012). "A Contextual-Bandit Algorithm for Mobile Context-Aware Recommender System". Neural Information Processing. Lecture Notes in Computer Science

PRICING STRATEGIES

- Pricing strategies[1]

也叫 The POKER strategy, 其中POKER 代表Price of Knowledge and Estimated Reward

其基本思路为：

1) 对已经访问的arm，建立相应的price；
2) 对没有访问过的arm，从已经访问过的arms种推测他们的特性
3) 剩余的试验次数对arm选择决策至关重要

POKER strategy 尤其适用于那些arm个数（远）大于可以试验次数（T）的情况

1. Vermorel, Joannes; Mohri, Mehryar (2005), Multi-armed bandit algorithms and empirical evaluation

UPPER CONFIDENCE BOUND ALGORITHMS:INTRO

- Upper confidence bound (UCB)

我们假设每一个arm的reward 都是均值为Ri的一个随机变量。定义Ri的confidence interval是一个区间[Rilow, Rihi]， Ri在这个区间内的概率大于98%。 其核心思路为：

在选取arm时， 预期按照Ri均值最大的来选， 不如按照Ri的 confidence interval的上限来选（upper bound）。

- 一个简单的实现[1]

对每个arm j, 其参数为xj : reward的均值， nj： armj 被选中的次数。 n 是一共进行实验的次数。

每次选择arm时， 选择armj 使 $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$ 最大

Regret 保证这个在范围内

$$\sum_{j=1}^{K} \frac{4 \ln n}{\Delta_j} + \left(1 + \frac{\pi^2}{3}\right) \Delta_j$$

where $\Delta_j = \mu^* - \mu_j$.

1. The algorithm UCB1 [Auer et al.(2002)Auer, Cesa-Bianchi, and Fischer]

UPPER CONFIDENCE BOUND ALGORITHMS

- 各种UCB的衍生算法
- **LinUCB (Upper Confidence Bound) algorithm**: the authors assume a linear dependency between the expected reward of an action and its context and model the representation space using a set of linear predictors
- **UCBogram algorithm**: The nonlinear reward functions are estimated using a piecewise constant estimator called a *regressogram* in Nonparametric regression. Then, UCB is employed on each constant piece. Successive refinements of the partition of the context space are scheduled or chosen adaptively[1]
- **KernelUCB algorithm**: a kernelized non-linear version of linearUCB, with efficient implementation and finite-time analysis.[2]

1. Rigollet, Philippe; Zeevi, Assaf (2010), *Nonparametric Bandits with Covariates*, Conference on Learning Theory
2. *Michal Valko; Nathan Korda; Rémi Munos; Ilias Flaounas; Nello Cristianini (2013), Finite-Time Analysis of Kernelised Contextual Bandits*

DEEP LEARNING ALGORITHM

- NeuralBandit algorithm (神经网络)[1]

特点：不需要假设稳定的环境变量和奖赏分布。

核心：对特定的环境，训练复数的神经网络来模拟奖赏分布

- Bandit Forest algorithm： random forest[2]

1. Rillesiardo, Robin; Féraud, Raphaël; Djallel, Bouneffouf (2014), "A Neural Networks Committee for the Contextual Bandit Problem", Neural Information Processing - 21st International Conference, ICONIP 2014,
2. *Féraud, Raphaël; Allesiardo, Robin; Urvoy, Tanguy; Clérot, Fabrice (2016). "Random Forest for the Contextual Bandit Problem*