Exploitation vs. Exploration

柏爱俊

阿里巴巴集团
一淘及搜索事业部

2014 年 12 月 19 日

主要内容

1. Exploitation vs. Exploration
2. Multi-Armed Bandits
3. Contextual MABs
4. Summary

探索和利用（Exploitation vs. Exploration）困境

- 不确定性环境下，决策者面临的基本挑战
  - Exploitation：选择目前看似最好的行动/方案
  - Exploration：探索尚未尝试充分的行动/方案

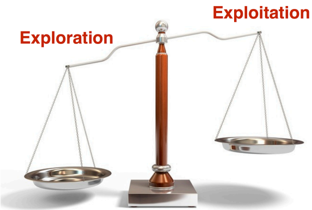


Figure 1：探索和利用平衡

E&E 举例 1

- 到食堂吃饭
  - Goal：最大化期望的用餐满意度
  - Uncertainty：所选食堂的每次用餐满意度
  - Exploitation：选择目前最喜欢的食堂
  - Exploration：尝试一个新的/看似较差的食堂

E&E 举例 2

- 商品推荐
  - Goal：最大化期望的点击率
  - Uncertainty：推荐商品的用户点击行为
  - Exploitation：推荐目前最受欢迎的商品
  - Exploration：推荐一个没有推荐过的/看似较差的商品

多臂赌博机（Multi-Armed Bandits）

Figure 2： 赌博机

MAB 问题

- 形式化
  - 行动空间 $\mathcal{A}$：N 个可选行动/方案（赌博机）
  - 未知收益值分布：$X_a \sim f_a(x \mid \theta_a)$
- 决策流程
  - 选择一个行动 $a_t \in \mathcal{A}$
  - 观察行动结果 $r_t = X_{a_t}$
- 最小化累积剩余值（Cumulative Regret）：

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}(X_{a^*} - r_t)\right] \tag{1}$$

ϵ-贪心策略

- ϵ-贪心策略
  - 以概率 $1-\epsilon$，选择 $a_t = \mathrm{argmax}_{a \in A} \bar{R}(a)$
    * $\bar{R}(a) = \frac{\sum_{1 \leqslant t \leqslant T} X_t \mathbf{1}[a_t = a]}{\sum_{1 \leqslant t \leqslant T} \mathbf{1}[a_t = a]}$ 是 a 的实验平均收益
  - 以概率 ϵ，随机选择一个动作 $a \sim \mathrm{Uniform}(A)$
- 贪心策略：$\epsilon = 0$
- 随机策略：$\epsilon = 1$
- 衰减 ϵ-贪心策略：$\lim_{t \to \infty} \epsilon_t = 0$

贪心策略的 Regret 曲线

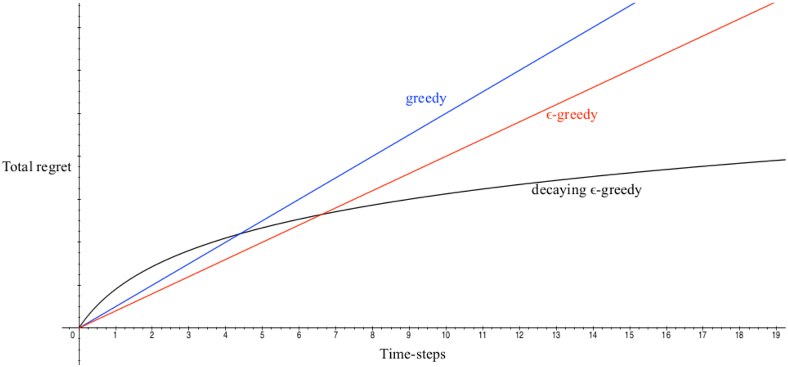


Figure 3：几种贪心策略的 Regret 曲线

UCB 策略

- 令 $Q(a)$ 为行动 a 的真实收益值（未知）
- $Q(a)$ 的置信区间上界（Upper Confidence Boud）：

$$\mathrm{UCB}(a) = \bar{R}(a) + c\sqrt{\frac{\log T}{N(a)}} \tag{2}$$

  - $\bar{R}(a)$ 是行动 a 的实验平均收益
  - $N(a) = \sum_{1 \leqslant t \leqslant T} \mathbf{1}[a_t = a]$ 是选择行动 a 的次数
  - T 是目前为止的所有行动次数
  - c 是 Exploitation-Exploration 平衡因子
- UCB 策略：$a_t = \operatorname{argmax}_{a \in A} \mathrm{UCB}(a)$
- 渐近最优性：Regret 曲线呈对数增长

Thompson 采样策略

- 根据一个动作成为最优动作的后验概率来随机选择该动作
  - 两个动作：a、b
  - $\Pr(a\ \text{最优} \mid \text{历史}) = 0.3$、$\Pr(b\ \text{最优} \mid \text{历史}) = 0.7$
- 形式上：

$$\Pr(a) = \int \mathbb{1}\left[a = \operatorname*{argmax}_{a'} \mathbb{E}\left[X_{a'} \mid \theta_{a'}\right]\right] \prod_{a'} \Pr(\theta_{a'} \mid Z)\, d\theta \tag{3}$$

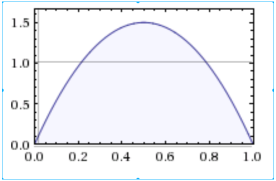
  - Z：行动－收益历史数据
  - θ_a：收益分布 X_a 的未知参数
  - $\Pr(\theta_a \mid Z)$：θ_a 的后验分布

Thompson 采样策略（续）

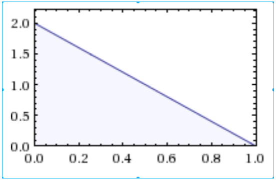
- 可以使用采样方法高效实现
  - 采样一组参数 $\theta_a \sim \Pr(\theta_a \mid Z)$
  - 选择具有最高期望值 $\mathbb{E}[X_a \mid \theta_a]$ 的行动
- 渐近最优性：Regret 曲线呈对数增长
- 实验上效果上比流行的 UCB 算法更好
- 近年来 MAB 问题的研究热点

Thompson 采样举例

- 两个动作：a 和 b
- 收益分布：Bernoulli 分布
- 未知参数：p_a 和 p_b
- 先验分布：$\mathrm{Uniform}(0,1)$
- 动作一收益历史：a, 1, b, 0, a, 0, ?
- 后验分布
  - $p_a \sim \mathrm{Beta}(2,2)$
  - $p_b \sim \mathrm{Beta}(1,2)$
- 采样 p_a 和 p_b
- 比较 $\mathbb{E}[X_a \mid p_a]$ 和 $\mathbb{E}[X_b \mid p_b]$



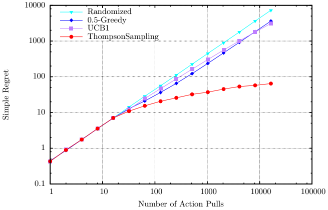
(a) $\mathrm{Beta}(2,2)$.



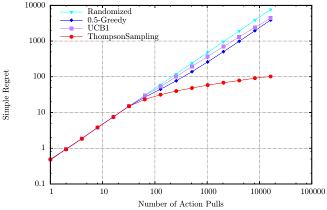
(b) $\mathrm{Beta}(1,2)$.

Figure 4：后验分布

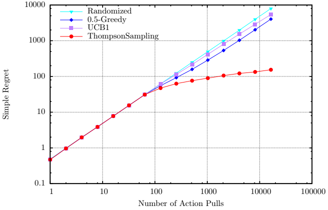
MAB 实验结果



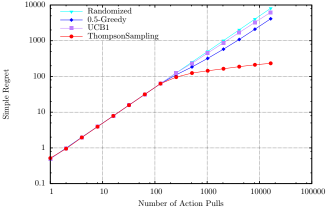
(a) 16 个动作



(b) 32 个动作



(c) 64 个动作



(d) 128 个动作

一个简单推荐系统的 MAB 建模

- 场景：推荐用户可能感兴趣的商品
- 行动：推荐方案（也就是商品的排序）
- 收益：用户的点击行为（1 或 0）
- 目标：提高商品的期望点击率

UCB 解决方案

- 记录每个商品的展示次数 n_i 和点击次数 m_i
- 点击率的置信区间上界

$$\mathrm{CTR}_i' = \frac{m_i}{n_i} + c\sqrt{\frac{\log \sum_j n_j}{n_i}} \tag{4}$$

- 根据 CTR'_i 对商品进行排序

Thompson 采样解决方案

- 记录每个商品的展示次数 n_i 和点击次数 m_i
- 点击率的后验分布

$$f_i(\mathrm{CTR} \mid m_i, n_i) = \mathrm{Beta}(1 + m_i, 1 + n_i - m_i) \tag{5}$$

- 采样商品的点击率

$$\mathrm{CTR}_i'' \sim f_i(\mathrm{CTR} \mid m_i, n_i) \tag{6}$$

- 根据 CTR_i'' 对商品进行排序

简单推荐系统的问题

- 没有考虑推荐的上下文
  - Query: 迈克尔 · 乔丹
    - 篮球？
    - 机器学习？
- 上下文相关的 MAB（Contextual MABs）

上下文相关的 MAB 问题

- 形式化
  - 行动空间 A
  - 上下文空间 S：用户和其 query 的特征
  - 未知收益值分布：$X_{s,a} \sim f_{s,a}(x \mid \theta_{s,a})$
- 决策流程
  - 观察当前上下文 s_t
  - 选择一个行动 $a_t \in A$
  - 观察行动结果 $r_t = X_{s_t,a_t}$
- 最小化累积剩余值（Cumulative Regret）：

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}(X_{s_t,a^*} - r_t)\right] \tag{7}$$

多个 MAB 问题

- 看成 $|S|$ 个独立的 MAB 问题
- 优点：处理比较简单
- 缺点：泛化性能不好
  - 所有新上下文的策略都为空

函数近似（Function Approximation）

- 令 $Q(s,a) = \mathbb{E}[X_{s,a}]$ 为 $X_{s,a}$ 的期望值
- 提取 (s,a) 的特征向量 $\phi(s,a)$
- 线性假设：用 $\phi(s,a)$ 的线性函数估计 $Q(s,a)$

$$Q(s,a) \approx Q_\theta(s,a) = \phi(s,a)^T\theta \tag{8}$$

- 观察行动历史：$\{(s_0,a_0,r_0),(s_1,a_1,r_1),\ldots,(s_T,a_T,r_T)\}$
- 通过线性回归估计参数 θ

$$A_t = \sum_{1\leqslant t\leqslant T} \phi(s_t,a_t)\phi(s_t,a_t)^T \tag{9}$$

$$b_t = \sum_{1\leqslant t\leqslant T} \phi(s_t,a_t)r_t \tag{10}$$

$$\theta_t = A_t^{-1}b_t \tag{11}$$

线性近似：UCB 解决方案

- 参数 θ_t 的协方差：A_t^{-1}
- $Q(s, a)$ 的期望值：$\phi(s, a)^T \theta_t$
- $Q(s, a)$ 的方差：$\phi(s, a)^T A^{-1} \phi(s, a)$
- $Q(s, a)$ 的置信区间上界

$$\mathrm{UCB}(s, a) = \phi(s, a)^T \theta_t + c\sqrt{\phi(s, a)^T A_t^{-1} \phi(s, a)} \tag{12}$$

- UCB 策略：$a_t = \operatorname{argmax}_{a \in A} \mathrm{UCB}(s_t, a)$

线性近似：Thompson 采样解决方案

- 高斯假设：$X_{s,a} \sim \mathcal{N}(\phi(s,a)^T\theta_t, \nu^2)$
- θ 的后验分布：$f_t(\theta \mid \theta_t, A_t, \nu) = \mathcal{N}(\theta_t, \nu^2 A_t^{-1})$
  - 其中，$\nu = R\sqrt{\frac{24}{\epsilon} d \ln(\frac{1}{\delta})}$ 为常数
  - R 满足：$r_t \in [\phi(s,a)^T\theta_t - R, \phi(s,a)^T\theta_t + R]$
  - (ϵ, δ) 保证算法以 $1-\delta$ 的概率找到“精度”为 ϵ 的解
- 采样 $\theta' \sim f_t(\theta \mid \theta_t, A_t, \nu)$
- Thompson 采样策略：$a_t = \mathrm{argmax}_{a \in A}\, \phi(s_t, a)^T\theta'$

总结

- Exploitation 和 Exploration 的平衡
  - 最优化长期收益
- MAB 问题：渐近最优策略
  - 衰减 ϵ-贪心策略
  - UCB 策略
  - Thompson 采样策略
- Contextual MAB 问题：函数近似
  - 线性假设
  - 高斯假设