Lecture 6: 多臂老虎机算法基础与应用

一、随机多臂老虎机

1.一般的机器学习任务

常见：Image recognition, Preference learning for recommendations, Speech recognition, Next token/word prediction(for language models)

以上都是基于模式识别的学习问题(recognition-based learning problems);
- 学习环境通常是给定的：输入图片/语言…输出标签；
- 通常使用监督学习：学习特征到标签的映射。

本讲介绍的是基于决策的学习问题(decision-based learning problems);
- 学习环境通常是动态的；
- 通过决策过程中获得的奖励/代价学习到最优策略(AlphaGo)；
- 强化学习(reinforcement learning)；
- 多臂老虎机：无状态转移的强化学习。

2.多臂老虎机问题：引入

一个赌鬼要玩多臂老虎机，摆在他面前有K个臂(Arms)或动作选择(Actions)，每一轮游戏中，他要选择拉动一个臂并会获得一个随机奖励(reward)(这一随机奖励来源于一个赌场设定好的分布，但赌鬼一开始不知道这一分布)。如果总共玩T轮，他该如何最大化奖励？

$$r_i \sim D_i(\text{for } i = 1, .., k)$$

其它场景：新闻网站(推荐系统)、动态定价、投资组合

问题本质：explore and exploit。例如新闻网站的推荐，可以先尝试几次随机推荐探索用户的喜好，然后利用这几次尝试学习到的结果做出比较准确的推荐。

3.基本概念

根据反馈类型分类：
- 完全反馈(full feedback)：能看见所有臂的奖励，例如投资组合股票的涨跌；
- 部分反馈(partial feedback)：能看见部分臂的奖励，例如动态定价中任何更低(高)价格都被接受(拒绝)；
- 老虎机反馈(bandit feedback)：只能看见选择的臂的奖励，例如新闻网站上该新闻是否被用户点击。

根据奖励类型分类：
- 随机奖励/IID 奖励：玩家或者的奖励随机曲子一个未知概率分布
- 对抗性奖励：奖励可以任意，能由一个"对手"有针对性的选择

4.随机多臂老虎机问题模型

在每一步$t = 1, 2, ..., T$中：

1. 玩家选择一个臂$a_t \in A = \{a_1, ..., a_K\}$;
2. 玩家获得该臂对应的随机奖励$r_t \sim R(a_t)(r_t \in [0, 1])$;
3. 玩家依据过往轮次的奖励情况调整选择策略，实现奖励最大。

说明：
- 奖励分布的均值记为$\mu(a_k) = \mathbb{E}[R(a_k)], k \in [K]$;
- 最优臂a^*的奖励均值$\mu^* = \max_{a \in A} \mu(a)$;
- 奖励均值差异$\Delta(a) = \mu^* - \mu(a)$。

5.遗憾分析

我们需要设计 MAB 算法实现最大化奖励，实际上就是找最优臂。那么，分析 MAB 算法的性能就是在分析算法能否找到最优臂。我们用遗憾(regret)来度量实际选择和最优选择的差异。

定义：

1. 伪遗憾(pseudo-regret)：

$$R(T)=\sum_{t=1}^{T}(\mu^*-\mu(a_t))=\mu^*\cdot T-\sum_{t=1}^{T}\mu(a_t)$$

2. 期望遗憾(expected regret):$\mathbb{E}[R(T)]$。

即伪遗憾就是选择最优臂的期望收益减去实际收益，期望遗憾是伪遗憾的期望(玩家策略可能存在随机性，因此$\mu(a_t)$可能是随机变量)。显然，最大化奖励的目标可以等价为最小化遗憾的目标。

在 MAB 问题中，wine 吧常常关注算法遗憾界(regret bound)。一个好的遗憾界是次线性的(sub-linear)，这意味着算法能逐渐学到最优臂，即

$$\frac{\text{regret bound}}{T}\to 0, T\to\infty$$

6.Hoeffding 不等式

假设$X_1,X_2,...,X_n$是$[0,1]$上的独立随机变量，样本均值为$\overline{X_n}=\frac{1}{n}\sum_{i=1}^{n}X_i,\mu=\mathbb{E}\left[\overline{X_n}\right]$。对于任意$\varepsilon>0$有：

$$\mathbb{P}\Big(|\mu-\overline{X_n}|\geq\varepsilon\Big)\leq 2\exp(-2n\varepsilon^2)$$

该不等式属于集中不等式(concentration inequalities)，直观上来说$\mathbb{P}\Big(|\mu-\overline{X_n}|\leq\text{small}\Big)\geq 1-\text{small}$，即样本均值与实际均值的差距很小的概率是很大的，并且随机变量个数越多(N越大)差距大的概率越小。

与多臂老虎机问题的关联：将$X_1,...,X_n$视为选择一个臂n次得到的n个奖励，这一不等式表明，当n很大时，采样出的奖励均值和臂的真实均值非常接近的。

称$[\mu-\varepsilon,\mu+\varepsilon]$是置信区间(confidence interval)，ε是置信半径(confidence radius)。若令$\varepsilon=\sqrt{\frac{\alpha\log T}{n}}$，则有

$$\mathbb{P}\Big(|\mu-\overline{X_n}|\geq\varepsilon\Big)=\mathbb{P}\left(|\mu-\overline{X_n}|\geq\sqrt{\frac{\alpha\log T}{n}}\right)\leq 2T^{-2\alpha},\forall\alpha>0$$

在接下来的讨论中一般取$\alpha=2$

7.贪心算法

为了找到那个最好的臂，一个朴素的解决思路是，将所有臂都尝试一遍，然后选择表现最好的臂，并在后续的时间步中保持这种选择。贪心算法就是这么一种算法，它采用"完全偏向利用"的策略，其基本策略是“始终选择当前估计奖励最高的臂，并利用历史经验更新对每个臂的奖励估计值。

1: 探索阶段：将每个臂各尝试 N 次
2: 利用阶段：
3: **for** $t>KN$ **do**
4: 选择平均奖励最高的臂 $\hat{a}=\arg\max_a Q_{t(a)}$
5: 观察奖励 r_t, $N_{t+1}(\hat{a})=N_{t(\hat{a})}+1$, $Q_{t+1}(\hat{a})=Q_{t(\hat{a})}+\frac{r_t-Q_{t(\hat{a})}}{N_{t+1}(\hat{a})}$

6: **end for**

注意$N_t(\hat{a})$的含义是第t轮前$\hat{a}$被选过的次数，

$$Q_{t+1}(\hat{a}) = Q_t(\hat{a}) + \frac{r_t - Q_t(\hat{a})}{N_{t+1}(\hat{a})}$$

则是在更新$\hat{a}$的平均奖励。第t轮未被选中的臂对应的N和Q值不变。

8.贪心算法的遗憾分析
定理：贪心算法的遗憾界为$O\left(T^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\right)$。

首先考虑$K=2$的情况，即只有两个臂。遗憾产生当且仅当选择了次优臂$a \neq a^*$。显然，探索阶段的遗憾为

$$R(\text{exploration}) \leq N$$

对于利用阶段，分为两种情况考虑(通常的分析套路)：
1. 事件E：所有臂a均满足$|\mu(a)-Q(a)| \leq \sqrt{\frac{2\log T}{T}}$，即两个臂的采样期望奖励$Q(a)$与真实期望$\mu(a)$间差距都不大的情况；
2. 事件$\overline{E}$：事件E的补集。

则有：

$$\begin{aligned}
&\mathbb{E}[R(\text{exploitation})] \\
&\leq \mathbb{E}[R(\text{exploitation}) \mid E] \times \mathbb{P}(E) + \mathbb{E}\left[R(\text{exploitation}) \mid \overline{E}\right] \times \mathbb{P}\left(\overline{E}\right) \\
&\leq \mathbb{E}[R(\text{exploitation}) \mid E] + T \times O\left(\frac{1}{T^4}\right)
\end{aligned}$$

其中$O\left(\frac{1}{T^4}\right)$来源于

$$\mathbb{P}\left(|\mu(a)-Q(a)| \leq \sqrt{\frac{2\log T}{N}}\right) \geq 1 - 2T^{-4}$$

由此可以看出，如上拆分成两个事件的目标是说明大概率发生的事件遗憾 小，遗憾大的事件发生概率小，综合二者可以证明遗憾是比较小的。

记$\text{rad} = \sqrt{\frac{2\log T}{N}}$，在事件$E$下产生遗憾时，有

$$\mu(a) + \text{rad} \geq Q(a) > Q(a^*) \geq \mu(a^*) - \text{rad}$$

其中首尾两个$\geq$来源于E的定义，中间的$>$来源于此时产生了遗憾，即最好的臂的采样期望比次优的臂小。整理得$\mu(a^*) - \mu(a) \leq 2\ \text{rad}$。那么

$$\begin{aligned}
&\mathbb{E}[R(\text{exploitation})] \\
&\leq \mathbb{E}[R(\text{exploitation}) \mid E] + T \times O\left(\frac{1}{T^4}\right) \\
&\leq (T-2N) \cdot 2\ \text{rad} + O\left(\frac{1}{T^3}\right)
\end{aligned}$$

综合探索和利用的遗憾可得$\mathbb{E}[R(T)] \leq N + 2\ \text{rad}\ T + O\left(\frac{1}{T^3}\right)$。若令$N = T^{\frac{2}{3}}(\log T)^{\frac{1}{3}}$，则有$\mathbb{E}[R(T)] \leq O\left(T^{\frac{2}{3}}(\log T)^{\frac{1}{3}}\right)$。

下面考虑$K>2$的情况，探索阶段的遗憾为

$$R(\text{exploitation}) \leq N(K-1)$$

利用阶段的遗憾为

$$\begin{aligned}
&\mathbb{E}[R(\text{exploitation})] \\
&\leq \mathbb{E}[R(\text{exploitation}) \mid E] \times \mathbb{P}(E) + \mathbb{E}\big[R(\text{exploitation}) \mid \overline{E}\big] \times \mathbb{P}\big(\overline{E}\big) \\
&\leq (T - NK) \cdot 2 \text{ rad} + O\left(\frac{1}{T^3}\right)
\end{aligned}$$

综合探索和利用的遗憾可得$\mathbb{E}[R(T)] \leq NK + 2 \text{ rad } T + O\big(\frac{1}{T^3}\big)$。若令 $N = \big(\frac{T}{K}\big)^{\frac{2}{3}} \cdot O(\log T)^{\frac{1}{3}}$，则有$\mathbb{E}[R(T)] \leq O\Big(T^{\frac{2}{3}}(K \log T)^{\frac{1}{3}}\Big)$

8.ε-贪心算法

回想上面的贪心算法存在两个问题：

- “探索”阶段的尝试带来遗憾；
- “利用”阶段陷于局部最优带来遗憾。

为了解决这两个问题，ε-贪心算法引入了随机性，在“探索”与“利用”之间实现了较好的权衡；

- 其核心思想是：以$1-\varepsilon$的概率选择当前已在最优的臂$a' = \arg\max_a Q_t(a)$，以ε的概率随机选择一个臂；
- 前一步代表对当前知识的“利用”，后一步代表对可能最优的“探索”，从而避免陷入局部最优；
- 通过调整$\varepsilon(0 \leq \varepsilon \leq 1)$的值，可以控制探索和利用之间的平衡；较小的$\varepsilon$值倾向于更多的利用，而较大的$\varepsilon$值倾向于更多的探索；通常而言会将$\varepsilon$设置成一个较小的值。

1: **for** $t = 1, 2, ..., T$ **do**
2: 　　以 ε_t 的概率探索：随机选择一个臂
3: 　　以 $(1-\varepsilon_t)$ 的概率利用：选择 $a_t = \arg\max_a Q_{t(a)}$
4: 　　观察奖励 r_t, $N_{t+1}(\hat{a}) = N_{t(\hat{a})} + 1$, $Q_{t+1}(\hat{a}) = Q_{t(\hat{a})} + \frac{r_t - Q_{t(\hat{a})}}{N_{t+1}(\hat{a})}$
5: **end for**

通过选择合适的ε值，可以证明ε-贪心算法的遗憾上界：

定理：令$\varepsilon_t = t^{-\frac{1}{3}}(K \log t)^{\frac{1}{3}}$，$\varepsilon$-贪心算法的遗憾界为$O\Big(T^{\frac{2}{3}}(K \log T)^{\frac{1}{3}}\Big)$。（定理的证明在 hw1.typ 中）

该算法的优势是简单易实现，并且可以通过调整ε灵活控制探索与利用。

9.上置信界算法

ε-贪心策略存在一个问题：虽然每个动作都有被选择的概率，但是这种选择太过于随机，导致最优臂被访问的概率较低，这并不能有助于智能体很大概率的发现最优选择，上置信界算法（upper confidence bound，UCB）很好地改进了这一点。

UCB 算法是多臂赌博机问题中一种经典的基于置信区间的探索-利用策略。其核心思想是为每个臂的奖励估计构建一个置信区间上界，选择上界最大的臂，从而在探索和利用之间自动平衡。

1: 对于每个候选项 $k = 1, ..., K$, 令 $Q_1(a_k) = 0, N_1(a_k) = 0$
2: **for** $t = 1, ..., T$ **do**
3: 　　**if** $t \leq K$ **then**
4: 　　　　初始化顺序选择每个臂

5: **else**
6: 选择 $a_t = \arg\max_a \left(Q_{t(a)} + \sqrt{\frac{2 \ln t}{N_{t(a)}}} \right)$
7: **end if**
8: 观察奖励 r_t, $N_{t+1}(a_t) = N_{t(a_t)} + 1$, $Q_{t+1}(a_t) = Q_{t(a_t)} + \frac{r_t - Q_{t(a_t)}}{N_{t+1}(a_t)}$
9: **end for**

简而言之，UCB 的算法流程是：首先将每一个候选臂都选择一遍，作为初始化；然后在后续的时间步中，选择奖励均值估计量的上置信界最大的臂，其中均值估计量的上置信界定义为

$$Q_t(a) + \sqrt{\frac{2 \ln t}{N_t(a)}}$$

最后更新被选中的臂的相关参数。其中上置信界的直观理解是：

- 上置信界的前一项$Q_t(a)$代表臂的估计奖励，这个值越大说明对应臂的历史表现越好；
- 后一项$\sqrt{\frac{2 \ln t}{N_t(a)}}$是置信区间的半径，其会随着选择次数的增加而变小，并且该值越大则说明估计的不确定性越大，因此能过鼓励玩家尝试较少被选择的臂，避免陷入次优；
- 因此，选择上置信界最大的臂有利于偏向于选择表现较好或是较少选择的臂，从而算法能够逐渐收敛到最优臂。

总而言之。UCB 算法同时考虑了估计奖励与不确定性，较好的平衡了探索与利用，也可以得到更好的遗憾界：

定理：UCB 算法的遗憾界为$O(\sqrt{KT \log T})$。

我们来补充此定理的证明：我们先来分析在什么情况下一个次优臂i(即$\Delta_i > 0$)会在第t轮被选择。根据 UCB 的规则，如果臂i被选中，那么它的 UCB 值必须大于等于最优臂a^*的 UCB 值：

$$\hat{\mu}_{i,N_i(t-1)} + \sqrt{\frac{2 \log t}{N_i(t-1)}} \geq \hat{\mu}_{a^*,N_{a^*}(t-1)} + \sqrt{\frac{2 \log t}{N_{a^*}(t-1)}}$$

为了使这个不等式成立，以下三种情况中至少有一种必须发生：

1. 最优臂被低估 (Pessimistic estimate for optimal arm): 最优臂的经验平均值远低于其真实平均值。

$$\hat{\mu}_{a^*,N_{a^*}(t-1)} < \mu^* - \sqrt{\frac{2 \log t}{N_{a^*}(t-1)}}$$

2. 次优臂被高估 (Optimistic estimate for sub-optimal arm): 次优臂i的经验平均值远高于其真实平均值。

$$\hat{\mu}_{a^*,N_{a^*}(t-1)} < \mu^* + \sqrt{\frac{2 \log t}{N_i(t-1)}}$$

3. 臂i的拉动次数还不够多 (Arm i is not pulled enough times): 此时置信区间仍然很大，导致其真实均值和最优均值的差距被置信区间覆盖。

$$\mu^* - \mu_i \leq 2\sqrt{\frac{2 \log t}{N_i(t-1)}}$$

假设三种情况都没发生，将他们结合起来得到：

$$\mathrm{UCB}_{a^*}(t) = \hat{\mu}_{a^*,N_{a^*}(t-1)} + \sqrt{\frac{2\log t}{N_{a^*}(t-1)}} \geq \mu^*$$

$$\mathrm{UCB}_i(t) = \hat{\mu}_{i,N_i(t-1)} + \sqrt{\frac{2\log t}{N_i(t-1)}} \leq \mu_i + 2\sqrt{\frac{2\log t}{N_i(t-1)}} < \mu^*$$

这就得出了$\mathrm{UCB}_i(t) < \mathrm{UCB}_{a^*}(t)$，与臂$i$被选中的前提相矛盾。因此，只要臂$i$被选中，上述三种情况必有其一为真。

现在我们来分析这三种情况发生的次数：

- 对于情况 3，我们对其进行代数变换：

$$\Delta_i \leq 2\sqrt{\frac{2\log t}{N_i(t-1)}} \Longrightarrow \Delta_i^2 \leq \frac{8\log t}{N_i(t-1)} \Longrightarrow N_i(t-1) \leq \frac{8\log t}{\Delta_i^2}$$

  由于$t \leq T$，所以臂i因为这种情况被选择的次数不会超过$\frac{8\log T}{\Delta_i^2}$。我们称这个阈值为$l_i = \left\lceil \frac{8\log T}{\Delta_i^2} \right\rceil$。

10.汤普森采样算法

二、对抗性多臂老虎机

三、多臂老虎机的应用