【Bandit Algorithms学习笔记】EXP3算法理论证明

在bandit问题中我们更应该关心随机bandit模型是不是能用而不是模型正不正确。模型正确与否通常用预测结果表示，模型是否好用通常用模型预测结果的准确率表示。

对抗式赌博机

对抗式赌博机不去假设奖励是如何生成的，我们通常叫对抗式赌博机的环境为**对手**（adversary）。**目标**是用来表述一个策略是否能够很好地和最优的动作对抗。

你可以想象一下和朋友在玩这样一个游戏，游戏流程如下：

1. 你告诉你的朋友你要选择的动作，动作有1或2.
2. 你的朋友秘密选择奖励x1 {0,1} 和x2 {0,1}
3. 你使用你的策略去选择动作A1或A2 ，然后得到奖励x_A
4. 遗憾就等于 R = max{x_1, x_2} -x_A

因此我们可以看到，adversarial bandit和stochastic bandit的区别是，前者的奖励是对手随机给出的（你可以想象一个赌场中，赌博机是能被人为操控的），而后者的奖励概率在我们之前讨论过的章节中是确定的（也就是每个bandit以概率 P_i 吐钱或者不吐钱）。

对抗式赌博机问题和随机赌博机问题相似，都有许多不同的推论，在下面的内容中我们将开始讨论。

Exp3算法

对抗式赌博机环境

对于n轮游戏，k臂对抗时赌博机会得到一串n个值的序列。每一轮学习者会选择动作的分布 P_t 属于 P_k-1 ，然后根据 P_t 采样得到动作 A_t 。

策略函数pai是历史序列到动作分布的一个映射。策略函数在环境中的表现可以用遗憾的期望来衡量，具体到下面这个公式：

k臂对抗式赌博机的交互协议如下：

加载失败，请点击重试

可见遗憾的随机性唯一来源于学习者动作的随机性，当然，和环境的交互意味着在t轮时刻的动作选择也许会依赖于动作t轮之前的动作，以及t轮及以前的奖励。

对于所有环境而言，worst-case遗憾为:

对于确定的策略，有结论 $R_n^*(\pi) \geq n(1-1/k)$ ，因此像Explore-then-commit和 UCB算法这种确定的策略是不适用于对抗式的场景。

根据琴生不等式以及最大不等式的凸性质，有：

EXP3算法流程

关于adversarial bandit最常见的算法就是EXP3，EXP3的计算流程如下：

- 根据先前计算得到的 P_{ti} 采样的得到 A_t
- 执行动作得到奖励，根据奖励的观测值 X_{ti} 估计每个动作的奖励估计值
- 用奖励的估计值更新概率 P_{ti}

第一轮，初始化每个动作被执行的概率为1/K。

奖励的估计值

所有对抗式赌博机算法的关键都是一套机制，采用这种机制去**估计**没有玩过的赌博机臂的**奖励**。

P_t 是第t轮动作的条件概率。对于i属于[k]， P_{ti} 的条件概率是（**执行A1，得到奖励X1，执行A2，得到奖励X2，一直到执行A_t-1得到奖励X_t-1的条件下**）执行动作A_i的概率。

我们定义 x_{ti} 的估计值为：

$$\hat{X}_{ti} = \frac{\mathbb{I}\{A_t = i\}}{P_{ti}} X_t \,.$$

X_{ti} 的条件期望满足， $E_{t-1}[X_{ti}] = x_{ti}$ ，意味着 X_{ti} 在t-1轮历史观测的条件下是 x_{ti} 的无偏估计。我们设定 $A_{ti} = I\{A_t = i\}$ 因此有 $X_t A_{ti} = x_{ti} A_{ti}$ ，并且 $E_t[A_{ti}] = P_{ti}$ ，并且因为 P_{ti} 是 $\sigma(A_1, X_1, ..., A_{t-1}, X_{t-1})$ 可测的。因此有：

$$\mathbb{E}_t\left[\hat{X}_{ti}\right] = \mathbb{E}_t\left[\frac{A_{ti}}{P_{ti}} x_{ti}\right] = \frac{x_{ti}}{P_{ti}} \mathbb{E}_t\left[A_{ti}\right] = \frac{x_{ti}}{P_{ti}} P_{ti} = x_{ti}$$

奖励估计值的方差

我们还需要计算奖励估计值的方差，因为如果奖励估计值的方差很小，证明我们的算法就越稳定，相反如果奖励估计值的方差太大，或许这个算法就不值得被使用。因此，我们需要考虑奖励估计值 X_{ti} 的方差 $V_t[X_{ti}]$ 。

根据条件方差的定义式

$$\mathbb{V}_t\left[U\right] \doteq \mathbb{E}_t\left[\left(U-\mathbb{E}_t\left[U\right]\right)^2\right]$$

结合

$$\hat{\boldsymbol{X}}_{ti}^2=\frac{A_{ti}}{P_{ti}^2}\boldsymbol{x}_{ti}^2$$

$$\mathbb{E}_t\left[\hat{\boldsymbol{X}}_{ti}^2\right]=\frac{\boldsymbol{x}_{ti}^2}{P_{ti}}$$

我们容易得到：

$$\mathbb{V}_t\left[\hat{\boldsymbol{X}}_{ti}\right]=\boldsymbol{x}_{ti}^2\frac{1-P_{ti}}{P_{ti}}$$

但是，观察这个式子我们能够看出一个问题：**概率越小的动作，方差越大。**因此，这样设定估计值对概率小的动作非常不友好，为了解决这个问题，设定另外一个估计值：

$$\hat{Y}_{ti}=\frac{\mathbb{I}\left\{A_t=i\right\}}{P_{ti}}Y_t$$

其中， $y_{ti}=1-x_{ti}$，$Y_t=1-X_t$ 。如此一来，根据方差的性质我们能够得到：前一种估计方式和后一种估计方式的方差是相等的，即：

$$\mathbb{V}_t\left[\hat{X}_{ti}\right]=\mathbb{V}_t\left[\hat{Y}_{ti}\right]=y_{ti}^2\frac{1-P_{ti}}{P_{ti}}$$

前者估计值的范围是0到正无穷，后者的范围是负无穷到1，所以后者估计值的方差较小，对“好动作” 的估计值会更加精准。

概率计算

我们定义前t轮奖励估计值的和为：

$$\hat{S}_{ti}\doteq\sum_{s=1}^{t}\hat{X}_{si}$$

对于前t轮奖励和越高的动作，我们自然是希望执行这个动作的概率越高，这样便于指导我们在接下来的几轮中获得更多的奖励。我们需要将奖励和映射为概率，这样便于我们选取动作：

$$P_{ti}\doteq\frac{\exp(\eta\hat{S}_{t-1,i})}{\sum_j\exp(\eta\hat{S}_{t-1,j})}$$

你可以理解 η 为学习率，值越大，越倾向于选择奖励高的动作。

为了便于计算，我们常常采用增量形式的概率计算：

$$P_{t+1,i}=\frac{P_{ti}\exp(\eta\hat{X}_{ti})}{\sum_j P_{tj}\exp(\eta\hat{X}_{tj})}$$

虽然这样的计算方式对于较大的n和K而言会使得数据不置信。

EXP3遗憾计算

上述的前置知识介绍完之后，我们进入EXP3算法的遗憾期望计算。

推论1

对于上述adversarial bandit问题，其遗憾满足下面式子：

$$R_n \leq 2\sqrt{nK\log(K)}\,.$$

证明

根据遗憾的定义式： $Rn = \Sigma_{t=1}^{n} X_{ti} - E[\Sigma_{t=1}^{n} X_t]$ ，因为奖励的估计值是无偏的，因此，$E[S_{ni}] = \Sigma_{t=1}^{n} x_{ti}$ ，并且 $E[X_t] = \Sigma_i P_{ti} x_{ti} = \Sigma_i P_{ti} E[X_{ti}]$ ，因此 $E[\Sigma_{t=1}^{n} X_t] = \Sigma_{t,i} P_{t,i} X_{t,i}$ 因此我们定义 Sn 拔等于 $\Sigma_{t,i} P_{t,i} X_{t,i}$ ，因此就有遗憾改写为：

$$\hat{S}_{ni} - \hat{S}_n$$

为了便于下文推导，我们将上式指数化，得到 $exp(\eta S_{ni})$ ，我们还定义:

$$W_n \doteq \sum_j \exp(\eta(\hat{S}_{nj}))$$

因此有 $W_0 = K, S0i = 0$ ，那么便有以下式子：

$$\exp(\eta \hat{S}_{ni}) \leq \sum_j \exp(\eta(\hat{S}_{nj})) = W_n = W_0 \frac{W_1}{W_0} \cdots \frac{W_n}{W_{n-1}}$$

根据上式规律，我们需要找出Wt / Wt-1，

$$\frac{W_t}{W_{t-1}} = \sum_j \frac{\exp(\eta \hat{S}_{t-1,j})}{W_{t-1}} \exp(\eta \hat{X}_{tj}) = \sum_j P_{tj} \exp(\eta \hat{X}_{tj})\,.$$

观察上式最右边的指数形式，我们根据不等式 $e^x \leq x^2 + x + 1$，$x < 1$ 和不等式 $e^x \geq x + 1$ ，得到：

$$\frac{W_t}{W_{t-1}} \leq 1+\eta\sum_j P_{tj}\hat{X}_{tj}+\eta^2\sum_j P_{tj}\hat{X}_{tj}^2 \leq \exp(\eta\sum_j P_{tj}\hat{X}_{tj}+\eta^2\sum_j P_{tj}\hat{X}_{tj}^2)$$

综上变形得到：

$$\exp(\eta\hat{S}_{ni}) \leq K\exp(\eta\hat{S}_n+\eta^2\sum_{t,j} P_{tj}\hat{X}_{tj}^2)$$

上式取对数后变形得到：

$$\hat{S}_{ni}-\hat{S}_n \leq \frac{\log(K)}{\eta}+\eta\sum_{t,j} P_{tj}\hat{X}_{tj}^2$$

我们可以找到上式最右边一项的上界为：

$$\mathbb{E}_t\left[\sum_j P_{tj}\hat{X}_{tj}^2\right]=\sum_j p_{tj}(1-2y_{tj})+\sum_j y_{tj}^2 \leq K$$

再放缩得到：

$$R_{ni} \leq \frac{\log(K)}{\eta}+\eta nK$$

根据高中所学的对勾函数的性质，我们容易得到遗憾上界为：

$$R_n \leq 2\sqrt{nK\log(K)}\,.$$

证明完毕。

课后思考题

我们之前的证明是采用X估计值，也就是第一种估计值证明遗憾的期望，小伙伴们可以试着用Y的估计值证明一下遗憾的期望，看看能得出什么结论吧～