Bandit Algorithms学习笔记2

7 UCB算法

优势

UCB算法对比ETC算法的优势在于：

1. 它不依赖于对于次优间隙的先验知识
2. 对于两个以上臂的赌博机表现更加优秀

算法原理

UCB公式

$$\mathrm{UCB}_i(t-1,\delta)=\begin{cases}\infty & \text{if } T_i(t-1)=0\\ \hat{\mu}_i(t-1)+\sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} & \text{otherwise}\,.\end{cases} \tag{7.2}$$

有人问deta是怎么得到的：根据这个不等式，用deta表示epsilon得到第二个式子

$$\mathbb{P}\left(\hat{\mu}\leq\mu-\varepsilon\right)\leq\exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

$$\mathbb{P}\left(\mu\geq\hat{\mu}+\sqrt{\frac{2\log(1/\delta)}{n}}\right)\leq\delta\qquad\text{for all }\delta\in(0,1)\,.$$

UCB算法流程

```
Input  k and δ
for t ∈ 1, ..., n do
    Choose action A_t = argmax_i UCB_i(t − 1, δ)
    Observe reward X_t and update upper confidence bounds
end for
```

Algorithm 3: UCB(δ).

在探索阶段阶段，我们需要探索更多的臂的原因有：

1. 臂i的t-1轮的奖励均值很大
2. 臂i在t-1轮内被执行的次数很小，没有被充分探索

在臂i被执行足够多的次数之后，我们希望t-1轮内臂i的历史奖励均值能够趋近于它的理论均值。

当以下式子成立时，我们假设臂1是最优的臂， δ 称为置信界：

$$\hat{\mu}_i(t-1)+\sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}\leq\mu_1\approx\hat{\mu}_1(t-1)+\sqrt{\frac{2\log(1/\delta)}{T_1(t-1)}}\,, \tag{7.3}$$

相关定理及证明

定理7.1

对于k臂1-次高斯赌博机问题，对于任意n轮选择，如果 $\delta=1/n^2$ 则采用上述UCB算法有遗憾上界如下：

$$R_n\leq 3\sum_{i=1}^{k}\Delta_i+\sum_{i:\Delta_i>0}\frac{16\log(n)}{\Delta_i}\,.$$

证明

臂i前s次采样的平均奖励可以定义为 $\mu_{is}=1/s\sum_{u=1}^{s}X_{ui}$

根据遗憾分解引理：有 $R_n=\sum_{i=1}^{k}\triangle_i\mathrm{E}[T_i(n)]$ ，算法要求我们至少对每个动作的至少执行一次，

动作i的UCB值大于最优动作的UCB值时，动作i会被选取，此时一下至少有一项发生，

1. 动作i的UCB值大于最优动作的理论均值

2. 最优动作的UCB小于它的理论均值

我们定义一个好的事件为：

$$G_i = \left\{\mu_1 < \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \cap \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\log\left(\frac{1}{\delta}\right)} < \mu_1\right\},$$

第一项：说明动作1的UCB值尚未收敛到它的理论均值

第二项：动作i的奖励上界小于动作1的奖励的理论均值

G_i 的定义揭露了两件事情：

1. $T_i(n) \leqslant u_i$
2. G_i^c 以较低的概率发生

因此前n轮动作i执行的次数的期望可以分解为(*)式：

$$\mathbb{E}\left[T_i(n)\right] = \mathbb{E}\left[\mathbb{I}\left\{G_i\right\}T_i(n)\right] + \mathbb{E}\left[\mathbb{I}\left\{G_i^c\right\}T_i(n)\right] \leq u_i + \mathbb{P}\left(G_i^c\right)n\,.$$

事件1可以用反证法证明，此处略。

定义完 G_i 之后， G_i^c 定义为它的补集：

$$G_i^c = \left\{\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \cup \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right\}.$$

接下来，我们将上式拆成两项进行分析。

分析第一项：根据UCB的定义，我们能够得到，

$$\left\{\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \subset \left\{\mu_1 \geq \min_{s\in[n]} \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}$$
$$= \bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}$$

μ_1 大于等于n轮内动作1UCB值的最小值的集合就真包含于n轮内 μ_1 大于等于动作1的UCB值的并集。

由上式得到如下如下不等式，最后一项采用置信度 δ 的定义进行放缩：

$$\mathbb{P}\left(\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right) \leq \mathbb{P}\left(\bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}\right)$$
$$\leq \sum_{s=1}^{n} \mathbb{P}\left(\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right) \leq n\delta\,. \quad (7.7)$$

分析第二项：

我们假设 **（**）式**:

$$\Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq c\Delta_i$$

结合次优间隙的定义： $\mu_1 = \mu_i + \triangle_i$ ，得到如下不等式：

$$\mathbb{P}\left(\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right) = \mathbb{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}}\right)$$
$$\leq \mathbb{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i\right) \leq \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$$

前两项我们容易理解，是简单的代入，**最后一项用了霍夫丁引理**进行放缩。

结合一、二两项，我们容易得到 G_i^c 发生的概率为：

$$\mathbb{P}\left(G_i^c\right) \leq n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$$

把它代入(*)式得到**(***)式**：

$$\mathbb{E}\left[T_i(n)\right] \leq u_i + n\left(n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\right)$$

我们再将(**)式变形得到：

$$u_i = \left\lceil \frac{2\log(1/\delta)}{(1-c)^2 \Delta_i^2} \right\rceil$$

最后，我们假设 $\delta = 1/n^2$ 并将上式代入**(***)式**并化简：

$$\mathbb{E}[T_i(n)] \leq u_i + 1 + n^{1-2c^2/(1-c)^2} = \left\lceil \frac{2\log(n^2)}{(1-c)^2 \Delta_i^2} \right\rceil + 1 + n^{1-2c^2/(1-c)^2}$$

向上取整要+1

小于1

不难看出最后一项是小于1的，因为 $2c^2/(1-c)^2$ 大于等于1。如果c趋近于1那么第一项会趋近于无穷大，因此我们选择代入c等于1/2，最终得到前n轮动作i执行的次数的期望的上界为：

$$\mathbb{E}\left[T_i(n)\right] \leq 3 + \frac{16\log(n)}{\Delta_i^2}$$

定理7.2

如果 $\delta = 1/n^2$ ，对于1-次高斯环境采用UCB算法的遗憾上界为：

$$R_n \le 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k}\Delta_i$$

证明

根据遗憾分解引理，遗憾可以定义为：

$$R_n = \sum_{i=1}^{k}\Delta_i \mathbb{E}\left[T_i(n)\right]$$

定理7.1中我们已经证明得到

$$\mathbb{E}\left[T_i(n)\right] \le 3 + \frac{16\log(n)}{\Delta_i^2}$$

代入遗憾的定义式可以得到：

$$\begin{aligned}
R_n &= \sum_{i=1}^{k}\Delta_i\mathbb{E}\left[T_i(n)\right] = \sum_{i:\Delta_i<\Delta}\Delta_i\mathbb{E}\left[T_i(n)\right] + \sum_{i:\Delta_i\ge\Delta}\Delta_i\mathbb{E}\left[T_i(n)\right] \\
&\le n\Delta + \sum_{i:\Delta_i\ge\Delta}\left(3\Delta_i + \frac{16\log(n)}{\Delta_i}\right) \le n\Delta + \frac{16k\log(n)}{\Delta} + 3\sum_i\Delta_i \\
&\le 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k}\Delta_i\,,
\end{aligned}$$

分解为大于delta和小于delta两种情况

小于1到k
次优间隙的和

最后一步用到高中数学中常见的 $ax + b/x \ge 2\sqrt{ab}$ 当且仅当 $ax = b/x$ 时等式去到最大。

证明完毕。