UCB 原理推导

2017 年 6 月 23 日

之前已经学习过 bandit 中的 ucb 策略，但是关于探索和利用的平衡策略公式一直没有深究，这次打算学习一下。证明过程涉及到几个不等式，下面先将不等式的证明说明一下。

马尔可夫不等式指的是一个非负随机变量 X 偏离它的期望的概率有多大。$P(X \geq kE[X]) \leq 1/k$, 即如果随机变量 X 的期望是 10，那么 X>100 的概率不超过 1/10。证明如下

$$E[X] \geq P[X \geq t] * t + P[X \leq t] * 0 = P[X \geq t] * t$$

切比雪夫不等式用来衡量随机变量随机的有多不均匀，定义随机变量的方差 $var[X] = E[(X - E(x))^2]$，简单推导一下，

$$var[X] = E[X^2 - 2XE[x] + E[x]^2] = E[X^2] - 2E[x]E[x] + E[x]^2 = E[x^2] - E[X]^2$$

切比雪夫的定义如下，对于一个随机变量 X，μ 是它的期望，σ 是标准差 ($\sqrt{var[X]}$)，那么对于任意正数 t，都有 $p(|X - \mu| > t\sigma) \leq 1/t^2$
证明如下

$$p(|x - \mu| > t\sigma) \Rightarrow p((x - \mu)^2 > t^2 var[x] = t^2 E[(x - \mu)^2])$$

然后利用马尔可夫不等式，直接得到结论。

切尔诺夫－霍夫丁界。先从期望和方差的性质谈起。$X = X_1 + X_2 + ... + X_n$，$E[X] = \sum_i E[X_i]$，如果 X_i 之间互不相关，则 $var[X] =$

$\sum_i var[X_i]$, 证明如下，

$$E[X^2]-E[X]^2=\sum_i\sum_j E[X_iX_j]-\sum_i\sum_j E[X_i]E[X_j]$$

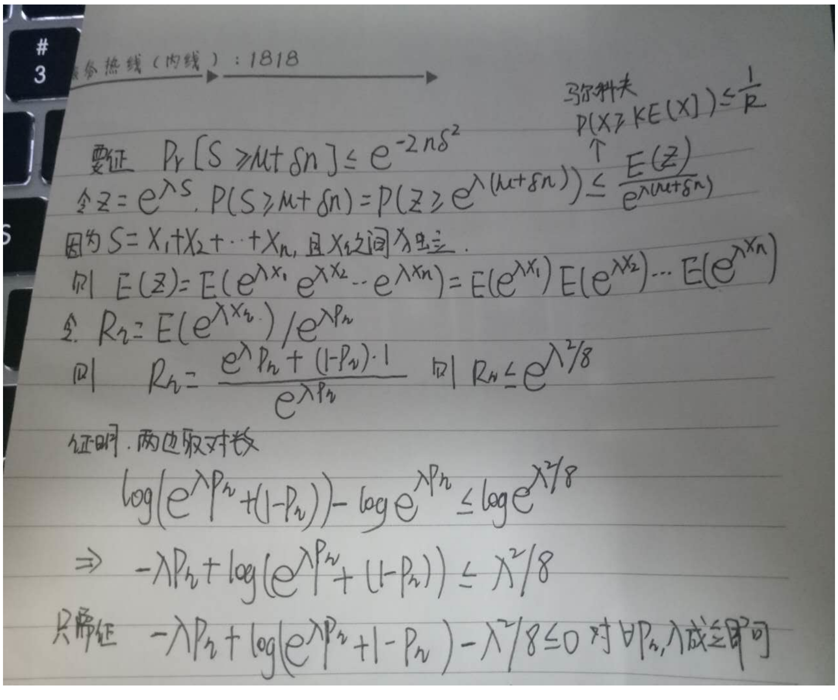
当 X_i, X_j 独立时，$E[X_iX_j]=E[X_i]E[X_j]$, 则消掉 $i\neq j$ 的项，则上化简为 $\sum_i E[X_i^2]-\sum_i E[X_i]^2$, 得证。

掷 n 枚硬币，正面得 1 分，反面− 1 分，计算最后得分随机变量 X。X_i 满足 $p(x_i=1)=p(x_i=-1)=1/2$，那么 $var[x_i]=1, var[X]=n, \sigma(X)=\sqrt{n}, \mu=0$ 利用切比雪夫不等式 $p(|x|>t\sqrt{n})\leq 1/t^2$。

X_i 是 n 个相互独立的随机变量，其中 $P(X_i=1)=p_i$, 令 $S=\sum_i X_i$, 令 $\mu=E[S]$, 则对于所有的 $0<\sigma<1$, 有

$$P(S\geq \mu+\sigma n)\leq e^{-2n\sigma^2}$$

证明过程直接手写了。


备热线（内线）：1818

马尔科夫 $P(X\geq kE(X))\leq \frac{1}{k}$

要证 $Pr[S\geq \mu+\delta n]\leq e^{-2n\delta^2}$

令 $Z=e^{\lambda S}$, $P(S\geq\mu+\delta n)=P(Z\geq e^{\lambda(\mu+\delta n)})\leq \frac{E(Z)}{e^{\lambda(\mu+\delta n)}}$

因为 $S=X_1+X_2+\cdots+X_n$, 且 X_i 之间独立.

则 $E(Z)=E(e^{\lambda X_1}e^{\lambda X_2}\cdots e^{\lambda X_n})=E(e^{\lambda X_1})E(e^{\lambda X_2})\cdots E(e^{\lambda X_n})$

令 $R_i=E(e^{\lambda X_i})/e^{\lambda p_i}$

则 $R_i=\frac{e^{\lambda}p_i+(1-p_i)\cdot 1}{e^{\lambda p_i}}$ 则 $R_i\leq e^{\lambda^2/8}$

证明. 两边取对数

$\log(e^{\lambda}p_i+(1-p_i))-\log e^{\lambda p_i}\leq \log e^{\lambda^2/8}$

$\Rightarrow -\lambda p_i+\log(e^{\lambda}p_i+(1-p_i))\leq \lambda^2/8$

只需证 $-\lambda p_i+\log(e^{\lambda}p_i+1-p_i)-\lambda^2/8\leq 0$ 对 $\forall p_i,\lambda$ 成立即可

图 1: 切尔诺夫－霍夫丁界证明过程 1

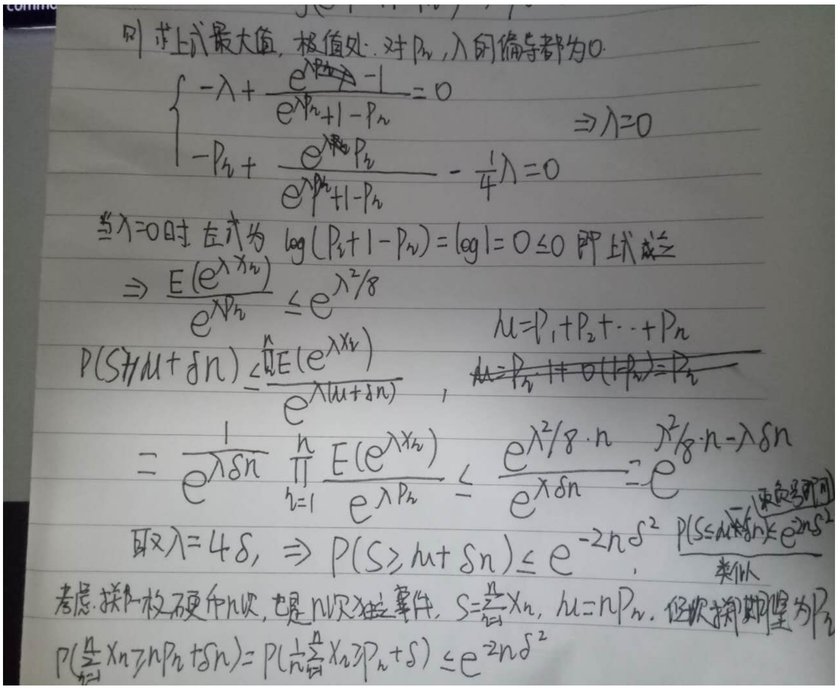


图 2: 切尔诺夫－霍夫丁界证明过程 2

现在回到 ucb 上，ucb(upper confidence bound) 是平衡探索和利用的策略，基本思想是不选择平均收益最高的，而选置信上限最大的，这么做是因为平均收益只是对实际收益期望的估计，当试验次数较少时，平均收益可能与实际期望偏离较大，因此不能把样本均值当作实际期望，而是对期望给出一个置信区间和置信概率。如何计算置信区间，就利用到了上面的切尔诺夫－霍夫丁界。$P(1/n\sum_{i\in 1,n} X_i \le \mu - \sigma) \le e^{-2a^2n}$ 即为 $P(\mu \ge 1/n\sum_{i\in 1,n} X_i + \sigma) \le e^{-2a^2n}$. ucb 选择的置信概率为 $1/t^4$，其中 t 指的是总共实验次数，n 指的是摇这个 arm 的次数，令 $1/t^4 = e^{-2a^2n}$ 得到 $\sigma = \sqrt{2lnt/n}$ 则置信上限如下

$$1/n\sum_{i\in 1,n} X_i + \sqrt{2lnt/n}$$