Concentration Inequalities and Multi-Armed Bandits

Nan Jiang

September 6, 2018

【强化学习理论 59】StatisticalRL 3

张楚珩
清华大学 交叉信息院博士在读

这是UIUC姜楠老师开设的CS598统计强化学习（理论）课程的第二讲。

原文传送门

CS598 Note2
nanjiang.cs.illinois.edu

一、Hoeffding's Inequality

初等的统计学过中心极限定理，它们告诉我们当i.i.d.样本很多的时候，在sample上的统计量就会趋向于真实的统计量。这里就更定量化的告诉我们sample上的统计量以怎样的程度趋向于真实的统计量，即concentration inequalities。Hoeffding's inequality就是其中最常用的一种。

Theorem 1. *Let $X_1, \dots, X_n$ be independent random variables on $\mathbb{R}$ such that X_i is bounded in the interval $[a_i, b_i]$. Let $S_n = \sum_{i=1}^n X_i$. Then for all $t > 0$,*

$$\Pr[S_n - \mathbb{E}[S_n] \geq t] \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}, \tag{1}$$

$$\Pr[S_n - \mathbb{E}[S_n] \leq -t] \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}. \tag{2}$$

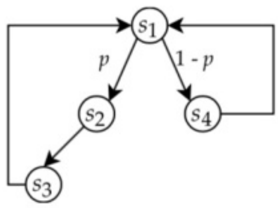
Remarks:

- By union bound, we have $\Pr[|S_n - \mathbb{E}[S_n]| \geq t] \leq 2e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$.
- We often care about the convergence of the empirical mean to the true average, so we can devide S_n by n: $\boxed{\Pr\left[\left|\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\right| \geq t\right] \leq 2e^{-2n^2t^2/\sum_{i=1}^n (b_i - a_i)^2}}$.
- A useful rephrase of the result when all variables share the same support $[a, b]$: with probability at least $1 - \delta$, $\left|\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\right| \leq (b - a)\sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$.
- $X_1, \dots, X_n$ are not necessarily identically distributed; they just have to be independent.
- The number of variables, n, is a constant in the theorem statement. When n is a random variable itself, for Hoeffding's inequality to apply, n cannot depend on the realization of $X_1, \dots, X_n$.

个人认为最好理解的就是红圈里面的不等式，即sample mean是如何趋向于true mean的。

为了说明最后一点，Note中给了一个例子。

Example: Consider the following Markov chain:



Say we start at s_1 and sample a path of length T (T is a constant). Let n be the number of times we visit s_1, and we can use the transitions from s_1 to estimate p.

1. Can we directly apply Hoeffding's inequality here with n as the number of coin tosses? If you want to derive a concentration bound for this problem, look up Azuma's inequality.

2. What if we sample a path until we visit s_1 N times for some constant N? Can we apply Hoeffding's inequality with N as the number of random variables?

对于第二种情况来说， N 固定，然后给定 N 个随机变量 X_i ，每个随机变量要么是3（走左边的环路）要么是2（走右边的环路），直接应用Hoeffding's inequality，就能得到该随机变量的样本上平均值距离真是平均值的界。

对于第一种情况需要用另一种concentration inequality，Azuma's inequality。它是对于鞅的concentration inequality，在这个问题上具体的用法我不太确定。这里简述一下。第一种情况下走的步数 T 是固定的，构建一个随机变量 X_i 表示新增访问 s_1 的次数，随机变量 $X_i' = X_i - \frac{1}{2+p}$ ， $Z_i = \sum_{j=1}^{i} X_i'$ 就是鞅，然后应用Azuma's inequality，注意到随机变量 $\sum_{i=1}^{T} X_i = n$ ，可以得到关于 p 和 n 的

concentration inequality由此来估算 p 。

In probability theory, the **Azuma–Hoeffding inequality** (named after Kazuoki Azuma and Wassily Hoeffding) gives a concentration result for the values of martingales that have bounded differences.

Suppose $\{ X_k : k = 0, 1, 2, 3, \ldots \}$ is a martingale (or super-martingale) and

$$|X_k - X_{k-1}| < c_k,$$

almost surely. Then for all positive integers N and all positive reals t,

$$P(X_N - X_0 \geq t) \leq \exp\left(\frac{-t^2}{2\sum_{k=1}^{N} c_k^2}\right).$$

And symmetrically (when X_k is a sub-martingale):

$$P(X_N - X_0 \leq -t) \leq \exp\left(\frac{-t^2}{2\sum_{k=1}^{N} c_k^2}\right).$$

If X is a martingale, using both inequalities above and applying the union bound allows one to obtain a two-sided bound:

$$P(|X_N - X_0| \geq t) \leq 2\exp\left(\frac{-t^2}{2\sum_{k=1}^{N} c_k^2}\right).$$

Azuma's inequality applied to the Doob martingale gives McDiarmid's inequality which is common in the analysis of randomized algorithms.

二、Multi-Aramed Bandits（MAB）

2.1. 问题描述

A MAB problem is specified by K distributions over $\mathbb{R}$, $\{R_i\}_{i=1}^{K}$. Each R_i has bounded supported $[0,1]$ and mean μ_i. Let $\mu^\star = \max_{i\in[K]} \mu_i$. For round $t = 1, 2, \ldots, T$, the learner

1. Chooses arm $i_t \in [K]$.
2. Receives reward $r_t \sim R_{i_t}$.

A popular objective for MAB is the pseudo-regret, which poses the *exploration-exploitation* challenge:

$$\text{Regret}_T = \sum_{t=1}^{T} (\mu^\star - \mu_{i_t}).$$

Another important objective is the simple regret:

$$\mu^\star - \mu_{\hat{i}},$$

where $\hat{i}$ is the arm that the learner picks after T rounds of interactions. This poses the "pure exploration" challenge, since all it matters is to make a good final guess and the regret incurred within the T rounds does not matter. A related objective is called Best-Arm Identification, which asks whether $\hat{i} \in \arg\max_{i\in[K]} \mu_i$; Best-Arm Identification results often require additional gap conditions.

MAB问题的目标主要有几种

- Pseudo-regret：希望在 T 轮内总的 reget 最小，即目标是一边探索一边利用；
- Simple regret：希望在 T 轮之后找到的那个 arm 的 regret 小，即目标是在 T 轮内尽可能做探索；

- Best-arm identification：希望在 T 轮之后以最大的概率找到最好的 arm，这也是一个鼓励探索的目标。

2.2. 均匀采样

这里我们的目标是最小化 simple regret，方式是在 T 轮中对于每个 arm 都 play 一样多的次数，然后我们推导出一个 simple regret 关于 T 的上界。这也可以看做一个Hoeffding不等式 的应用。

We consider the simplest algorithm that chooses each arm the same number of times, and after T rounds selects the arm with the highest empirical mean. For simplicity let's assume that T/K is an integer. We will prove a high-probability bound on the simple regret. The analysis gives an example of the application of Hoeffding's inequlaity to a learning problem; the algorithm itself is likely to be suboptimal.

For simplicity let's assume that T/K is an integer. After T rounds, each arm is chosen T/K times, and let $\hat{\mu}_i$ be the empirical average reward associated with arm i. By Hoeffding's inequality, we have:

$$\Pr[|\hat{\mu}_i - \mu_i| \geq \epsilon] \leq 2e^{-2T\epsilon^2/K}.$$

Now we want accurate estimation for *all* arms simultaneously. That is, we want to bound the probability of the event that *any* $\hat{\mu}_i$ deviating from μ_i too much. This is where union bound is useful:

$$\begin{aligned} &\Pr\left[\bigcup_{i=1}^{K}\{|\hat{\mu}_i - \mu_i| \geq \epsilon\}\right] && \text{(the event that estimation is ϵ-inaccurate for at least 1 arm)} \\ \leq &\sum_{i=1}^{K} \Pr\left[|\hat{\mu}_i - \mu_i| \geq \epsilon\right] \leq 2Ke^{-2T\epsilon^2/K}. && \text{(union bound, then Hoeffding's inequality)} \end{aligned}$$

注意这里使用了 union bound，讲的是（ K 个事件中任意一个事件发生）的概率小于 K 个事件（每个事件独立发生）的概率的和。

To rephrase this result: with probability at least $1-\delta$, $|\hat{\mu}_i - \mu_i| \leq \sqrt{\frac{K}{2T}\ln\frac{2K}{\delta}}$ holds for all i simultaneously.

Finally, we use the estimation error to bound the decision loss: recall that $\hat{i} = \arg\max_{i\in[K]} \hat{\mu}_i$, and let $i^\star = \arg\max_{i\in[K]} \mu_i$.

$$\begin{aligned} \mu^\star - \mu_{\hat{i}} &= \mu_{i^\star} - \hat{\mu}_{i^\star} + \hat{\mu}_{i^\star} - \mu_{\hat{i}} \\ &\leq \mu_{i^\star} - \hat{\mu}_{i^\star} + \hat{\mu}_{\hat{i}} - \mu_{\hat{i}} \leq 2\sqrt{\frac{K}{2T}\ln\frac{2K}{\delta}}. \end{aligned}$$

We can rephrase this result as a sample complexity statement: in order to guarantee that $\mu^\star - \mu_{\hat{i}} \leq \epsilon$ with probablity at least $1-\delta$, we need $T = O\left(\frac{K}{\epsilon^2}\ln\frac{K}{\delta}\right)$.

2.3. 下界

这里讲的是对于所有的用来解决 MAB 问题的算法而言，性能都不可能超过一个什么样的界。概括说起来就是，如果最优的 arm 的均值比次优的 arm 的均值没有大太多，这样我们就很难在较少的次数内以较高精度来分辨谁是最优的，因此找一个最优的 arm 成功率就不会太高。

The linear dependence of the sample complexity on K makes a lot of sense, as to choose a arm with high reward we have to try each arm at least once. Below we will see how to mathematically formalize this idea and prove a lower bound on the sample complexity of MAB.

Theorem 2. *For any $K \geq 2$, $\epsilon \leq \sqrt{1/8}$, and any MAB algorithm, there exists an MAB instance where $\mu^\star$ is ϵ better than other arms, yet the algorithm identifies the best arm with no more than $2/3$ probability unless $T \geq \frac{K}{72\epsilon^2}$.*

The theorem itself is stated as a best-arm identification lower bound, but it is also a lower bound for simple regret minimization. This is because all arms except the best one is ϵ worse than $\mu^\star$, so missing the optimal arm means a simple regret of at least ϵ.

See the proof in [1] (Theorem 2); the technique is due to [2] and can be also used to prove the lower bound on the regret of MAB.

另外，我导师前年的高等理论计算机课程也讲过这个问题，证明了 upper

confidence bound（UCB） 算法在此问题上的性能。参见ATCS Note4。

编辑于 2019-05-14