强化学习学习笔记

强基数学 *002* 吴天阳

第二章　多臂赌博机

定义 1 (多臂赌博机)**.** k 臂赌博机（*k-armed Bandits*）有一台机器包含 k 种**动作**（***Action***）可进行选择，每个动作对应一个概率分布，在做出动作选择后，会从对应的概率分布中采样得到对应的**收益**（***Reward***），目标是在有限的时间内与赌博机进行交互，并达到最大总收益.

数学表示：设存在 k 个不同的分布 $f_k(\cdot)$，分别对应 k 个动作，总共存在 T 个时刻，在时刻 t 选择的动作记为 A_t，得到的收益（*Reward*）记为 R_t，并且 R_t 来自分布 f_{A_t}，通过在每个时刻与机器进行交互从而最大化 $\sum\limits_{t=1}^{T} R_t$.

我们将每个动作收益的期望值记为 $q_*(a)$，则其满足 $q_*(a) := \mathbb{E}[R_t|A_t = a]$，通过不断地和机器进行交互，从而得到 $q_*(a)$ 的估计. 于是在 t 时刻，我们将 $q_*(a)$ 的估计量记为 $Q_t(a)$，称为 a 对应的**价值（Value）**.

假设我们有 $q_*(a)$，并且 $q_*(a)$ 不会随时间发生变换，那么通过贪心的思想，不难得到，每次选择最大收益对应的动作即可最大化全局收益，但是事实并不如此，其一我们仅有 $q_*(a)$ 的估计量 $Q_t(a)$，所以不能保证估计量的准确性；其二，现实场景中的往往是非平稳的（Nonstationary Problem），也就是指收益的分布会随时间等因素发生变换，而不是保持一个稳定的分布不变. 所以我们的算法不能一味地贪心选择当前最优价值，而是以一定概率探索新的动作，从而得到可能更优的价值. 具体而言，每次动作的选择会分为**探索与利用**两种：

$$\begin{cases}\text{1. 利用（Exploitation）：贪心操作，选择 } \arg\max\limits_a Q_t(a) \text{ 作为当前执行的动作.}\\ \text{2. 探索（Exploration）：以非贪心操作执行动作.}\end{cases}$$

强化学习中很重要的一个问题就是如何去平衡探索与利用两种操作.

2.1　动作-价值方法

动作-价值方法（Action-value Methods）是指用价值来进行动作的选择. 一种自然的估计价值的方法是用收益的均值：

$$Q_t(a) := \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

其中 $\mathbb{1}_{A_i=a} = \begin{cases}1, & A_i = a,\\ 0, & \text{否则.}\end{cases}$，下面引入的 ε-贪心算法（ε-greedy）是一种常用的平衡探索与利用的方法.

算法 1 (ε-贪心)**.** 该算法以 ε 的概率在全部动作集合中随机选择（探索），以 $1-\varepsilon$ 的概率以贪心的方法选择 $\arg\max\limits_a Q_t(a)$（利用）.

2.1.1 均值估计的增量法

增量法（Incremental Implementation）是对均值估计的改进，如果直接通过均值公式计算时间复杂度会不断上升，我们考虑通过递推的方式求解 $Q_t(a)$，下面只考虑对于某个特定的动作为 a，当前时刻之前总共选择了 n 次动作 a，每次选择所获得的收益为 $\{R_1, R_2, \cdots, R_n\}$，则

$$\begin{aligned} Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) = \frac{1}{n}\left(R_n + (n-1)Q_n\right) \\ &= Q_{n-1} + \frac{1}{n}(R_n - Q_n) = Q_{n-1} + \alpha(t)(R_n - Q_n) \end{aligned} \tag{2.1}$$

其中 Q_{n+1}, R_n 分别表示第 n 次选择动作 a 后动作 a 的价值与收益，$\alpha(t) := 1/n$ 称为**步长（StepSize）**，有时为常量，有时可随时间发生变换，例如这里与时间成反比关系.

2.1.2 指数近因估计

在上述均值估计中，使用的是变换的步长，如果我们将其取为 $(0, 1)$ 中的常量 α，则

$$\begin{aligned} Q_{n+1} = Q_n + \alpha(R_n - Q_n) &= \alpha R_n + \alpha(1-\alpha)R_{n-1} + \cdots + \alpha(1-\alpha)^{n-1}R_1 + (1-\alpha)^n Q_1 \\ &= (1-\alpha)^n Q_1 + \sum_{k=1}^{n} \alpha(1-\alpha)^{n-k} R_k \end{aligned} \tag{2.2}$$

由于系数之和满足

$$(1-\alpha)^n + \alpha\sum_{k=1}^{n}(1-\alpha)^{n-k} = (1-\alpha)^n + \alpha\sum_{k=1}^{n}(1-\alpha)^{k} = (1-\alpha)^n + \frac{1-(1-\alpha)^n}{\alpha}\alpha = 1$$

则 (2.2) 式是对 $\{Q_1, R_1, \cdots, R_n\}$ 的一种加权平均，且随时间差的增大，权重以指数形式递减，越靠近当前时刻的权重越大，于是这种方法也称为**指数近因加权平均（exponential recency-weighted average）**.

在随机逼近论中有以下定理，常用于判断估计量是否能依概率收敛到真实值上：

定理 2.1. 设 α_n 为某个动作第 n 步的步长，若 $\{\alpha_n\}$ 满足

$$\sum_{n=1}^{\infty} \alpha_n = \infty \quad 且 \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty$$

即 $\{\alpha_n\} \in \ell^2 \backslash \ell^1$ 时，Q_n 能以概率 1 收敛到真实值 q_*，即 $\forall \varepsilon > 0$，有

$$\lim_{n\to\infty} \mathrm{P}(|Q_n - q_*| < \varepsilon) = 1$$

上述定理中，$\{\alpha_n\} \notin \ell^1$ 说明步长需要足够大，以克服初始条件或随机波动，$\{\alpha_n\} \in \ell^2$ 是保证收敛性.

注意到：当 $\alpha_n = 1/n$ 时，Q_n 收敛，这也是大数定律所保证的；但当 $\alpha \in (0, 1)$ 为常值时，上述定理失效，说明估计永远无法完全收敛，而是随最近得到的收益变换而变换，但在非平稳环境中这种方法的效果比收敛的效果更好.

例 1 (练习 2.5). *设计实验来证实使用均值估计方法取解决非平稳问题的困难，使用一个 10 臂赌博机，其中所有的 $q_*(a)$ 初始时均相等，然后进行随机游走，每一步所有的 $q_*(a)$ 都加上一个服从 $N(0, 0.01^2)$ 的增量，分别使用均值估计方法和指数近因加权估计方法且步长 $\alpha = 0.1$ 进行决策，采用 ε-贪心进行动作选择，且总步数为 $T = 10000$.*

解答. 如图1所示进行了 2000 次不同的多臂赌博机实验的平均结果，从中非常容易得出，指数近因估计在处理多臂赌博机问题上比均值估计要好.

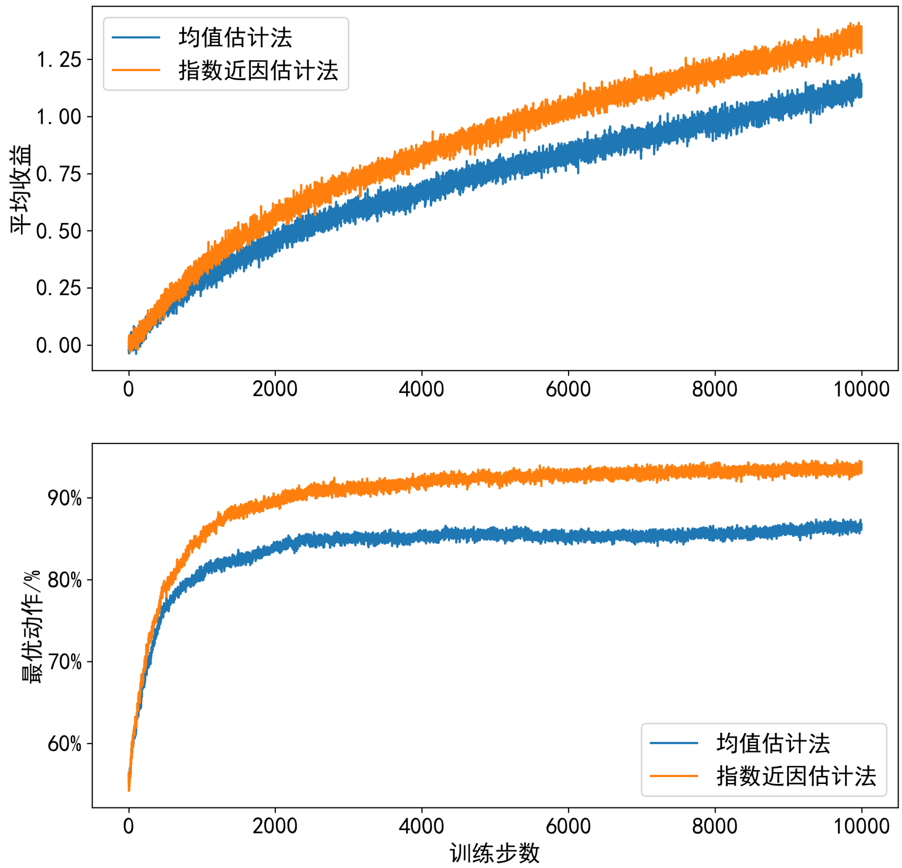

图 1: 非稳定问题中均值法与指数近因法的对比