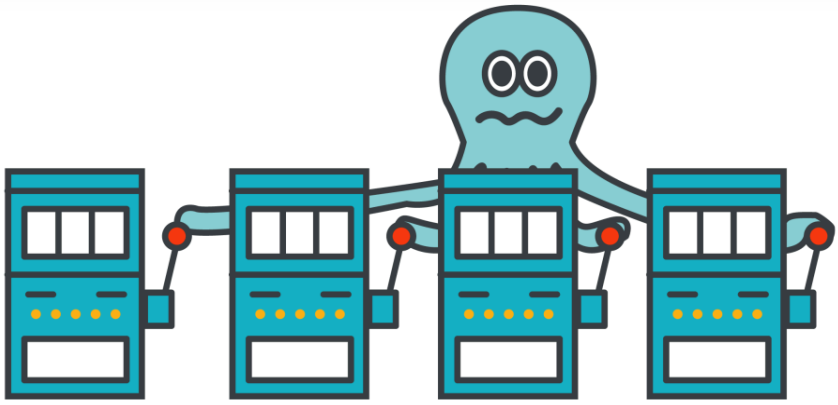
知乎 首发于 RL from scratch



【强化学习】多臂老虎机的上置信界算法（全网最通俗）

TRiddle
阿里巴巴 员工

关注他

143 人赞同了该文章 〉

收起

机
法
算法
定性度量
策略

有一位大佬曾经说过：“不了解多臂老虎机就等于没接触过强化学习。”

虽然我觉得这句话是夸张了，但至少它传达出一个很重要的信息：多臂老虎机是一个特别经典的问题，理解它能够帮助我们学习强化学习。多臂老虎机的解法有很多种，其中上置信界算法✦是比较有意思的。它不仅能够简洁而又漂亮地应用在多臂老虎机中，而且还是蒙特卡洛树✦的重要组成部分。

因此，本文会简单地介绍多臂老虎机，然后详细地讲解上置信界算法。在讲解上置信界算法时，先尽量不使用任何数学公式介绍它的思想。如果你很感兴趣，还可以跟随本文的思路深入了解它的数学细节。或者干脆直接跳过数学部分看结论，应该也不会影响阅读。

多臂老虎机

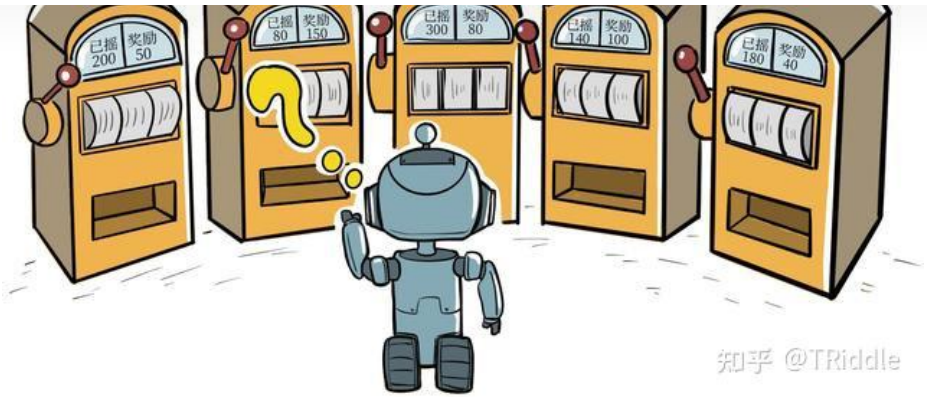
多臂老虎机（multi-armed bandit，MAB）是一个很有趣的问题。很多实际问题（例如寻找最优广告投放策略）都可以建模成这个问题。另外，它跟**强化学习**（reinforcement learning）有非常深的联系。了解多臂老虎机的探索与利用问题，对学习强化学习环境探索有很重要的帮助。

那么先来描述一下这个问题吧。

假设你有5个按钮（或者拉杆，如果你了解老虎机的话），每按一次按钮就有一定的概率获得1块钱。按下每个按钮获得奖励的概率是不同的，麻烦的是，你事先并不知道这些概率分别是多少。好在你一共能按100次按钮，每次除了有可能获得奖励外，还能积累一些经验。不过每按一次按钮，你都会**懊悔**（regret），懊悔的程度为最优按钮（也就是概率最大的那个）的奖励概率减去本次按下的按钮的奖励概率。请问，有什么办法能让你在按完所有按钮后懊悔的总和最小。

如果是第一次遇见这个问题，不妨在此处停留一下，先自己思考一下怎么解决这个问题.....

知乎 首发于 RL from scratch



多臂老虎机（图片来自参考资料1）

简单的算法

好了，现在我们来探讨一下。

显然，一个简单的切入点是：按了几次按钮之后，你能够获得一些经验。比方说，你可以估计每个按钮的奖励概率。假设你按了5次1号按钮，其中有3次获得了奖励，那么你一定会认为，这个按钮的奖励概率会比较接近60%。按照这个思路，你应该：

1. **先将每个按钮各按一次来获得初始预估奖励概率**
2. **接着每次选择预估奖励概率最高的按钮按下（如果有多个就随机选一个）**
3. **然后根据是否获得奖励来更新刚刚按下的按钮的预估概率**
4. **重复前两个步骤直到不能再按下按钮。**

我们不妨就按照这种策略比划一下。假如现在5个按钮的初始预估奖励概率分别是0, 1, 1, 0, 0（每个按钮各按1次，但只在按2号和3号时获得了奖励）。

- 第一次选择按钮时，因为2号和3号按钮的预估奖励概率都是1，所以你随机选择到了2号按钮，你按下它并得到了奖励，这时候2号按钮的预估奖励概率是 $\frac{1+1}{1+1}=1$
- 在第二次选择按钮时，你又随机选择到了预估奖励概率最高的2号按钮，按下按钮的同时并没有得到奖励，这时候2号按钮的预估奖励概率就变成了 $\frac{2+0}{2+1}\approx 0.67$
- 在第三次选择按钮时，你只会选择预估奖励概率最高的3号.....

这个过程中，各个按钮的预估奖励概率如下：

	按钮1	按钮2	按钮3	按钮4	按钮5
初始情况	0	1	1	0	0
按第1次后	0	1	1	0	0
按第2次后	0	0.67	1	0	0
.....	.....	.....	.....	.....	.....

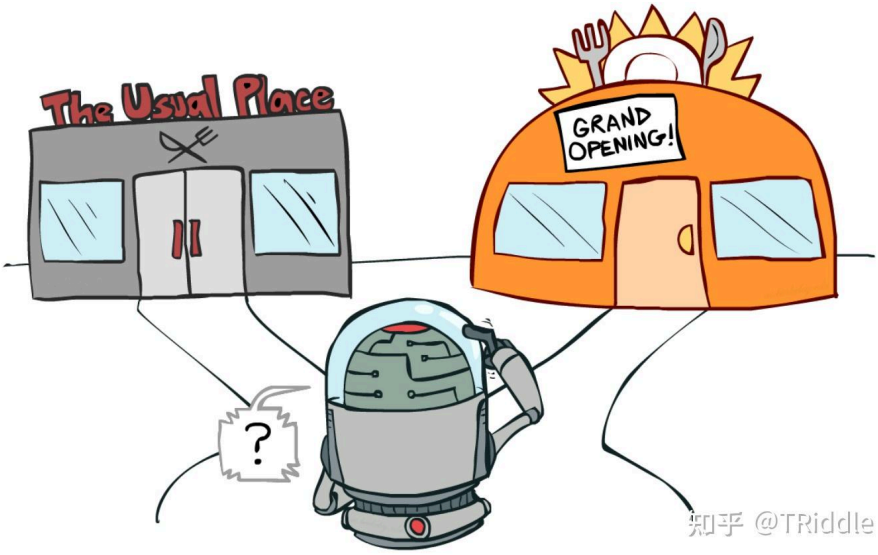
相信通过这个简单的例子，这种简单的算法已经比较清晰了。接下来，很自然地，你会思考一个问题：这个简单的算法会不会有什么问题呢?

虽然这个简单的算法的效果不会太差，但是反复利用现有的知识很可能会让你陷入**局部最优解**。例如，你按了10次2号按钮，有7次都得到了奖励，因此它的预估奖励概率是0.7；而3号按钮只被你按了1次，并且很不巧没有得到奖励，因此它的预估概率是0。这时候能说选择2号按钮一定更好吗？也许多按几次3号按钮，它的预估奖励概率会变成0.9也说不定呢?

	按钮1	按钮2	按钮3	按钮4	按钮5
.....	.....	.....	.....	.....	.....
按第N次后	.....	0.7（10次）	0（1次）	.....	.....
.....	.....	.....	.....	.....	.....

知乎 首发于 RL from scratch

活化的例子是：我今天到底是该去最喜欢的餐厅，还是去新开张的餐厅呢？上置信界算法能够更好地处理这个问题。



探索和利用问题（图片来自链接6）

上置信界算法

上置信界（upper confidence bound，UCB）算法，又称置信区间上界算法。他的思想是：显然，当前成功率高的按钮是有高利用价值的，不确定性高的按钮是有高探索价值的。我们能不能综合这两种价值，给每种按钮一个评分，从而选择评分最高的按钮呢？

让我们想一想。用预估奖励概率（得到奖励的次数 / 按下的次数）来衡量某个按钮的“利用价值”就行了，“探索价值”可以用这个按钮的某种不确定性度量来衡量。而按钮的评分，最简单的方法就是用两个价值的和来表示。也就是说：

评分 = 预估奖励概率 + 不确定性度量

到了这个阶段，找到一种能够衡量按钮的不确定性的方法就至关重要了。下面说一下UCB算法的思路：

假设当前某个按钮的真实奖励概率是p，它被按下了n次，其中有 n_r 次获得了奖励，它的预估奖励概率是 $\tilde{p} = \frac{n_r}{n}$ 。如果能够根据我们按按钮的经验找到一个 δ ，使得 $p \leq \tilde{p} + \delta$ 恒成立，那么 $\tilde{p} + \delta$就是真实概率p的上界，或者说 $\tilde{p} + \delta$ 决定了p的取值范围。因此在 $\tilde{p}$ 确定的情况下， δ 越大p的取值范围也就越大。显然 δ 能够在一定程度上反映按钮的不确定性。

那怎么找到这个 δ 呢？我们需要引入一个著名的数学原理——**霍夫丁不等式**（Hoeffding's inequality）。如果你对这部分不感兴趣的话，可以直接跳过下一节的推导，看找到 δ 之后的部分。

推导不确定性度量

先来看看霍夫丁不等式的内容（暂时不用透彻地理解，后面会详细解释）：

> 【霍夫丁不等式】设 $X_1, \ldots, X_n$ 是取值范围为[0, 1]的n个独立同分布随机变量，用 $\tilde{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 表示它们的样本均值，用p表示它们的分布的均值。那么有 $P\{p - \tilde{p} \leq \delta\} \geq 1 - e^{-2n\delta^2}$ 。

所以，霍夫丁不等式怎么帮助我们找到之前说的δ呢？

在多臂老虎机问题中，对于某个按钮而言，按一次的结果（即得到的钱）可以用服从伯努利分布的随机变量X表示。其中有p的[illegible]0。得到的钱的期望也正是p

1	p	按下按钮后得到了奖励

按n次该按钮的结果分别是取值范围为[0, 1]的独立同分布的随机变量$X_1, \ldots, X_n$。它们的样本均值为$\tilde{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$。

你发现了吗？这其实就是霍夫丁不等式的条件的描述。

霍夫丁不等式本身又是什么意思呢？就是说$p - \tilde{p} \leq \delta$这件事发生的概率一定大于或等于$1 - e^{-2n\delta^2}$。我们对$p - \tilde{p} \leq \delta$做移项，可以得到$p \leq \delta + \tilde{p}$。

相信你一定又发现了，这不正是我们定义不确定信度量δ时用的不等式吗？

前面说过，我们想要让$p \leq \tilde{p} + \delta$恒成立。可惜这是不现实的，但是好在根据霍夫丁不等式，我们可以想办法让$p \leq \tilde{p} + \delta$这件事发生的概率足够大，大到我们可以睁一只眼闭一只眼，近似地认为它一定会发生。也就是说我们要想办法让霍夫丁不等式右侧的$1 - e^{-2n\delta^2}$在小于1的限制下（不清楚的话可以分析一下它的值域）足够大。

这下问题又来了，多大才算足够大？不妨用当前按下按钮的总次数N来定义"足够大"，看看会得到什么样的结论。$\frac{N-1}{N}$看上去就足够大了，而且也满足小于1的限制。我们令$1 - e^{-2n\delta^2} = \frac{N-1}{N}$解这个方程可以得到$\delta = \sqrt{\frac{\ln N}{2n}}$。

结论就是我们令$\delta = \sqrt{\frac{\ln N}{2n}}$，可以几乎使得$p \leq \tilde{p} + \delta$恒成立，因此$\delta$是一个不错的不确定性度量。

其实还有一种方式让霍夫丁不等式右侧的值足够大。因为$1 - e^{-2n\delta^2}$关于δ是单调递增的，所以只要让δ趋向于无穷大，$1 - e^{-2n\delta^2}$也会无限趋近1。只不过当δ落在$[1, \infty)$区间内的时候，根据概率的定义，任何按钮都能满足$p \leq \tilde{p} + \delta$恒成立（两个概率的差值不可能大于1），这样的$\delta$没有携带任何额外的信息，显然不足以用来刻画不确定性度量。反观$\delta = \sqrt{\frac{\ln N}{2n}}$，在绝大多数情况下都是小于1的。

按钮选择策略

现在是时候结束在数学世界中的旅程，回到现实世界中来了。我们刚刚得到了不确定性度量的表达式$\delta = \sqrt{\frac{\ln N}{2n}}$，这很好。但是别忘了，推导过程中有一个量是我们拍脑门定义出来的，虽然有道理，但是理由不够充分。所以我们必须拿出更有说服力的理由来说明为什么这个表达式真的能够度量不确定性。观察这个式子，

- 当n不变时，δ随着N的增大而增大。也就是说，当我们始终不按某个按钮，却多次按下其它按钮时，这个按钮的不确定性（相对别的按钮而言）就升高了
- 当N不变时，δ随着n的增大而减小。也就是说，当我们按某个按钮的比例增大后，这个按钮的不确定性（相对别的按钮而言）就降低了

这两条分析都是合理的所以可以认为我们得到的不确定性表达式是合理的。

至此，我们得到了UCB算法选择按钮的策略：

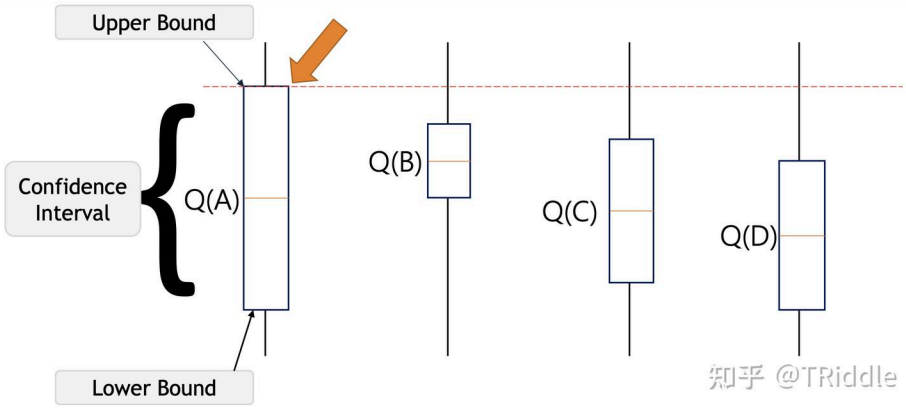
每次选择评分$\frac{n_r}{n} + \sqrt{\frac{\ln N}{2n}}$最高的按钮

为了防止分母爆零，我们可以将分母加上1。另外，为了调节探索与利用的权重，我们可以在根号前乘上一个超参数c，以便通过设置一个较大的c使策略更加注重探索。相当于

在工程上会使用这个评分表达式：$\frac{n_r}{n+1} + c\sqrt{\frac{\ln N}{2n+1}}$

实验证明（有严格的理论证

知乎 首发于 RL from scratch



各个按钮在某一时刻的置信区间（来自参考资料4）

总而言之，如果要用一句话来概括上置信界算法的话，那就是：**用上置信界来给按钮评分的算法**。

参考资料

1. 多臂老虎机
2. 机器学习A-Z～置信区间上界算法 Upper Confidence Bound or UCB
3. 冯伟：Multi-Armed Bandit: UCB (Upper Bound Confidence)
4. Upper Confidence Bound for Multi-Armed Bandits Problem
5. Pikachu5808：霍夫丁不等式（Hoeffding's inequality）
6. The Epsilon-Greedy Algorithm for Reinforcement Learning

编辑于 2023-07-12 12:16 · IP 属地北京

内容所属专栏

RL from scratch 订阅专栏

强化学习 (Reinforcement Learning) 机器学习 算法

理性发言，友善互动

35 条评论 默认 最新

芝士熊
当n不变时， δ\delta\delta 随着N的增大而增大。也就是说，当我们始终不按某个按钮，却多次按下其它按钮时，这个按钮的不确定性（相对别的按钮而言）就降低了当N不变时，δ\delta\delta 随着n的增大而减小。也就是说，当我们按某个按钮的比例增大后，这个按钮的不确定性（相对别的按钮而言）就升高了
——这一段是不是写反了？
2022-12-09 · 浙江 回复 2

TRiddle 作者 ▸ 好多好吃
按下其它按钮的次数减少，N就不变了
2023-11-05 · 河北 回复 喜欢

好多好吃 ▸ TRiddle
这个N按下按钮的总次数，在n增加时，N怎么不变的
2023-11-05 · 上海 回复 喜欢

展开其他 1 条回复 ›

梦里啥都有