强化学习（三）——术语总览

前两篇文章中，我们讨论的是非常简单的问题：

$$奖励信号要么完全独立，要么只与时间步相关$$

也就是说，行为的价值和环境毫无关系。
这显然是过于理想的假设，实际上，大多数时候，行为的价值和环境息息相关。比如：下棋时，不同的布局下，同一个走法可能价值天差地别。

在正式翻开新篇章之前，我将先介绍我们所使用的基本术语，以便后续的写作。

Agent & Environment & Policy & Action & Reward

代理（Agent）类似于一个玩家，通过不停地与环境（Environment）互动完成学习过程。

代理可以从环境中观察，得到状态（s_t（t 表示时间步）），再将 状态 s_t 映射到执行的动作 a_t，这个映射就是 代理的 策略（Policy）。我们用 $\pi_t(s,a)$ 表示代理在状态 s 下执行 动作 a 的概率。

代理执行动作 a_t 后，环境将会改变，并且产生一个奖励信号 r_{t+1} 。代理将会接受奖励信号，同时观察新环境（s_{t+1}），如此往复轮回。

Rewards & Goals & Episodic

奖励（rewards）是强化学习系统的关键，决定着系统的目标（goals）。

比如，迷宫游戏中，我们可以在代理走出迷宫后给予一定的 rewards，代理为了最大化奖励，会尝试走出迷宫，走出迷宫也就成了代理的目标。

rewards 的设定方式数不胜数，在迷宫游戏中，我们除了在代理走出迷宫后给予一个正向的奖励信号，还可以在代理走了一步却没有出迷宫的情况下给予一个负向的惩罚信号，鼓励代理尽可能快地走出迷宫。这里可以看出，**rewards 设定方式不同，任务的目标往往不同。**

值得补充的是，我们往往不会令代理自己生成奖励信号，虽然人类的奖励信号实际上是在人体内部生成的。这可能导致，代理学会直接生成高奖励信号，但是却没能实现我们希望的目标。当然，**很多研究也在尝试把奖励过程交给代理。**

迷宫这类任务，一旦走出迷宫，意味着结束（存在着一个终止状态（s_T）），我们称之为 片段化（Episodic）任务；与之相对应的，存在一种不会结束（不存在终止状态）的任务，我们称之为连续性（Continuing）任务，比如：仿生机器人。

Returns

Returns 是一个我们用来调整代理策略的信号，它衡量了某时刻后所有的 rewards。如果代理执行一个行为后，期望的 Returns 很大，那么我们认为代理执行这个行为是比较正确的选

择，反之，类似。

既然 Returns 衡量某时刻后所有的 rewards，那么最朴素的想法自然是直接求和。

$$R_t = \sum_{k=0}^{T} r_{t+k+1}$$

但是我们会发现一个问题，$T \to \infty$ 时（连续性任务），R_t 也会趋于无穷大（我们认为奖励信号不是无穷小）。

故而实际上，我们会使用下面这个式子计算 Returns

$$R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1}$$

$0 < \gamma < 1$，被称作折扣因子。

我们会发现，上面这个式子不仅可以用在片段化任务中，也可以用在连续性任务中。而且，它还提供了一个很好的特性：

在回报相同的情况下，代理优先考虑更简单的实现方式

我们可以通过调节折扣因子的大小来调整代理的‘**格局**’，折扣因子越接近 1，代理目光就越长远；折扣因子越接近 0，代理目光就越短浅。