马尔可夫决策过程（二）

关键词：
#value_functions
#transition_graph
#Monte_Carlo_Methods
#Parameterized_function

我们用 $\mathcal{P}_{ss'}^{a}$ 表示 $s_t = s, a_t = a$ 时，状态转移到 $s_{t+1} = s'$ 的概率，用 $\mathcal{R}_{ss'}^{a}$ 表示 $s_t = s, a_t = a$ $s_{t+1} = s'$ 时代理获得的期望奖励。

Value Functions

价值函数用来形容某个状态或者某个状态下某动作的好坏。
明显，**价值是与策略有关的**，不同的策略下，同一个状态或动作可能会有不同的价值。

比如：游戏中，激进的游戏策略下，代理往往采取激进的行为，这些行为可能会带来很弱的奖励信号。因此，此时我们倾向于认为产生大量激进行为的不稳定状态的价值偏低。然而，在另一种稳健的策略下，即使面对产生大量激进行为的不稳定状态，代理依然会选择稳健的行为，而这些稳健的行为也许反而反馈较高的奖励信号，此时我们会认为这些状态的价值偏高。

综上，状态价值函数是**状态和策略**的函数，特定状态下特定动作的价值函数是**动作、状态和策略**的函数。我们分别记为 $V^{\pi}(state)$，$Q^{\pi}(state, action)$ 。π 代表策略。

结合我们之前的折扣因子的思想，可以这么量化：

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\}$$

$$Q^{\pi}(s, a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\}$$

即用未来可以获得的期望奖励信号作为价值。这是符合直觉的，未来的奖励信号越是丰厚，代表价值越是高。

估算价值函数

- Monte Carlo Methods
  根据大数定律，我们只需要采样足够多的 $(s_t, a_t, r_{t+1}, s_{t+1})$，（接下来以 V^{π}为例），计算 $V^{\pi}(s)$ 时，将 状态s后 可以产生的所有轨迹找出来计算未来能获得的奖励信号的期望。很显然，这种方法只适用于状态数较少的情况，当状态数增大时，我们需要采样的点也急速增大，大大降低了学习的效率。
- Parameterized Functions
  利用带参数的函数将 **状态** 映射到 **价值**。这里的前提是，**状态有一定规律地影响价值，**

我们可以用较少的参数表达出这种规律，从而很好的将状态映射到价值上，如果状态中不含任何规律，那么我们只能一个状态直接对应一个价值（或者用过量的参数进行过拟合），而不能用一个比较简单的函数去表达。
我们往往可以通过学习算法，将参数函数的参数逐步更新，拟合实际的价值函数。

价值函数的递归性质

我们可以通过数学变换构造出价值函数的一个递归方程：

$$
\begin{aligned}
V^{\pi}(s) &= E_{\pi}\{R_t|s_t=s\} \\
&= E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t=s\right\} \\
&= E_{\pi}\left\{r_{t+1}+\gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_t=s\right\} \\
&= \sum_a \pi(s,a)\sum_{s'}\mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'}+\gamma E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_{t+1}=s'\right\}\right] \\
&= \sum_a \pi(s,a)\sum_{s'}\mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'}+\gamma V^{\pi}(s')\right],
\end{aligned}
$$

这个递归方程的物理意义十分鲜明，即为 $s\to s'$ 的奖励信号以及 折扣的 $V^{\pi}(s')$ 之和。
这就是十分重要的 **贝尔曼方程**。