最优价值函数

#optimal_value_functions
#tabluar_methods
#parameterized_function
#Bellmen_equation

假设我们现在有两个策略 π ， π' 。我们定义，$\pi >= \pi'$ 当且仅当对所有状态 s，$V^{\pi}(s) >= V^{\pi'}(s)$ 。我们记最优策略为 π^* （最优策略就是 大于等于 其他所有策略的策略），记对应的价值函数为 $V^*(s)$ 、$Q^*(s,a)$。

最优价值函数的递归性质

我们将给出最优价值函数的递归性质，这可以很容易的得到。

$$
\begin{aligned}
V^*(s) &= \max_{a \in \mathcal{A}(s)} Q^{\pi^*}(s,a) \\
&= \max_a E_{\pi^*}\left\{R_t \mid s_t = s, a_t = a\right\} \\
&= \max_a E_{\pi^*}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right\} \\
&= \max_a E_{\pi^*}\left\{r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s, a_t = a\right\} \\
&= \max_a E\left\{r_{t+1} + \gamma V^*\left(s_{t+1}\right) \mid s_t = s, a_t = a\right\} \\
&= \max_{a \in \mathcal{A}(s)} \sum_{s'} \mathcal{P}_{ss'}^a \left[\mathcal{R}_{ss'}^a + \gamma V^*\left(s'\right)\right].
\end{aligned}
$$

$$
\begin{aligned}
Q^*(s,a) &= E\left\{r_{t+1} + \gamma \max_{a'} Q^*\left(s_{t+1}, a'\right) \mid s_t = s, a_t = a\right\} \\
&= \sum_{s'} \mathcal{P}_{ss'}^a \left[\mathcal{R}_{ss'}^a + \gamma \max_{a'} Q^*\left(s', a'\right)\right].
\end{aligned}
$$

我们可以发现，对于一个有限的马尔可夫决策过程（不妨假设状态数为 N），则对 N 个状态我们都可以得到一个 递归方程。**总的来说，我们可以得到 N 个非线性方程。**如果我们已知任意的 ${\mathcal{P}_{ss'}}^a$、${\mathcal{R}_{ss'}}^a$。那么这 N 个方程就仅仅只有 N 个未知数，我们可以求得 $V^*(s)$。

一旦得到了 $V^*(s)$，我们就可以找到最优策略。

最优策略：对 状态 s，做一步搜索。状态 s 可以由不同的动作转移到其他的状态，每一个动作都可以获得一定的奖励信号，这个奖励信号加上后续状态的最佳奖励信号就是这个动作的最佳奖励信号，所有动作中，最佳奖励信号最大的就是最佳动作。

类似地，我们可以求 $Q^*(s,a)$。值得一提的是，一旦得到 $Q^*(s,a)$，我们就可以很容易地找到最优策略（不再需要一步搜索）。

优化和近似

在绝大多数我们感兴趣的问题中，直接得到 V^* 或 Q^* 都是不现实的，而且由于实际问题的复杂性，想要获得 准确的 最优价值函数 很难，即使可以完成，也需要大量的计算资源。因此，我们往往找到最优价值函数的近似即可。

- Tabluar Methods：
  我们将最优价值函数储存为 表（Table） 的形式。
- Parameterized Functions：
  我们构建函数，将状态映射到价值，这可以解决 状态过大时，表无比庞大，既难储存、又难维护的缺点。
  无论有多少状态，我们只要将状态映射到价值即可。
  以后，我们将会介绍具体的学习算法。