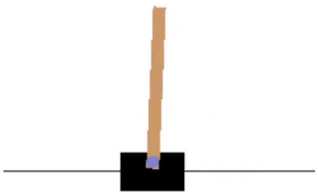
强化学习概述

深度学习如图像识别和语音识别解决的是感知问题，而强化学习相当于大脑，解决的是智能决策问题或者说序贯决策问题，就是随着环境的变化连续不断的作出决策，实现最终的目标。

强化学习最初应用在倒立摆问题上，这里的决策是指应该给台车施加什么方向、多大的力，使倒立摆系统收敛到目标点即保持竖直。



马尔科夫决策过程 MDP

强化学习方法适用于马尔科夫决策过程，所要解决的问题要满足马尔科夫性。即系统的下一个状态 S_{t+1} 仅与当前的状态 S_t 有关，而与之前的状态无关。

1、马尔科夫决策过程

马尔科夫决策过程由(S, A, P, R, γ)描述，其中 S 为有限的状态集；A 为有限的动作集；P 为状态转移概率，它是包含动作的，$P_{ss'}^{a}$ = $P[S_{t+1} = s' | S_t = s, A_t = a]$；R 为回报函数；γ 为折扣因子，用来计算累积回报。

2、策略 π(a|s)

强化学习的目标是给定一个马尔科夫决策过程，寻找最优策略。所谓策略是指状态到动作的映射，通常用 π 表示，它是指给定状态 s 时，动作集上的一个分布：$\pi(a|s)=p[A_t=a|S_t=s]$。这里的最优是指得到的总回报最大。

3、累积回报 G_t

当有策略 π 后，就可以计算累积回报了。时刻 t 之后得到的累积回报定义如下：

$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{n} \gamma^k R_{t+k+1}$ 其中 γ 为折扣因子表示将来奖励的影响程度，当 γ=0 时，只用即时奖励来评判。由于 π 是服从一定概率分布的随机变量，所以 G_t 为随机变量。

4、状态值函数 $V_\pi(s)$与状态行为值函数 $Q_\pi(s,a)$

用状态值函数来评价某一状态 s 的价值，用 $V_\pi(s)$表示。可以用累积回报来定义，但由于累积回报 G_t 是随机变量，所以将从状态 s 出发使用策略 π 所带来的累积奖赏的期望值 $E[G_t]$ 定义为状态值函数：$V_\pi(s) = E_\pi[R_{t+1} + \gamma R_{t+2} + \cdots] = E_\pi[\sum_{k=0}^{n}\gamma^k R_{t+k+1} | S_t=s] = E_\pi[G_t | S_t=s]$

定义状态行为值函数来评价在某一状态 s 下动作 a 的价值，用 $Q_\pi(s,a)$表示。定义为从状态 s 出发执行动作 a 后再使用策略 π 所带来的累积奖赏：$Q_\pi(s,a) = E_\pi[\sum_{k=0}^{n}\gamma^k R_{t+k+1} | S_t=s, A_t=a]$

5、贝尔曼方程

贝尔曼方程就是 $V_\pi(s)$与 $V_\pi(s')$，$Q_\pi(s,a)$与 $Q_\pi(s',a)$，以及 $V_\pi(s)$与 $Q_\pi(s,a)$之间的关系，一切都是建立在定义（累积回报）的概念之上，具体理解时想着他们的定义（累积回报）会容易理解一些。

1、当$P_{ss'}^{a}$=1 时，

当$P_{ss'}^{a}$=1 时，即在策略 π 下当发出一个动作后会到达一个确定的状态 s_i，已知之后每个状态的值函数 V_i，并且有相应的回报 r_i。

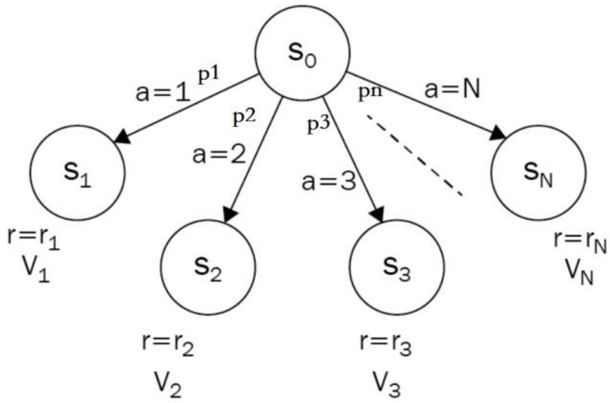
由状态值函数的定义可以得到：

$$
\begin{aligned}
\upsilon_\pi(s) &= E_\pi\left[G_t \mid S_t = s\right] \\
&= E_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \mid S_t = s\right] \\
&= E_\pi\left[R_{t+1} + \gamma\sum_{k=0}^{\infty}\gamma^k R_{t+k+2} \mid S_t = s\right] \\
&= E_\pi\left[R_{t+1} + \gamma\upsilon_\pi(S_{t+1}) \mid S_t = s\right]
\end{aligned}
$$

同样可以得到状态-动作值函数的贝尔曼方程：

$$
q_\pi(s,a) = E_\pi\left[R_{t+1} + \gamma q(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right]
$$

状态值函数与状态-行为值函数的具体推导过程：



有：$Q_\pi(s_0,a_i) = r_i+\gamma V_i(s')$； $V_\pi(s_0) = \Sigma_a p_i \cdot Q_\pi(s_0,a_i) = \Sigma_a \pi(a|s) \cdot Q_\pi(s_0,a_i)$;

则：$V_\pi(s_0) = \Sigma_a \pi(a|s) \cdot (r_i+\gamma V_i(s'))$ $Q_\pi(s_0,a_i) = r_i+\gamma \Sigma_a \pi(a|s) \cdot Q_\pi(s',a_i)$

2、当 $P_{ss'}^{a} \neq 1$时

当 $P_{ss'}^{a} \neq 1$时，发出动作 a 之后，可能转移到三个不同的状态。

有：$Q_\pi(s_0,a_i) = r_i+\gamma \cdot \Sigma_{s'} P_{ss'}^{a} \cdot V_i(s')$； $V_\pi(s_0) = \Sigma_a p_i \cdot Q_\pi(s_0,a_i) = \Sigma_a \pi(a|s) \cdot Q_\pi(s_0,a_i)$;

则：$V_\pi(s_0) = \Sigma_a \pi(a|s) \cdot (r_i+\gamma \cdot \Sigma_{s'} P_{ss'}^{a} \cdot V_i(s'))$