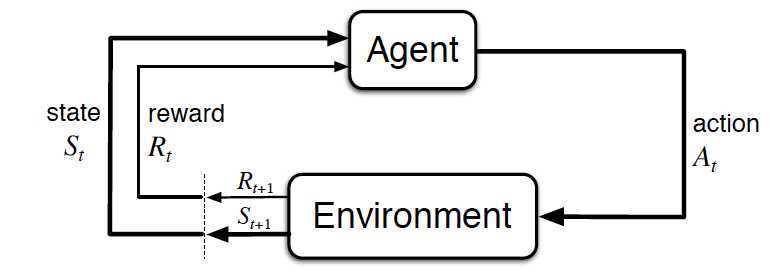
强化学习(Reinforcement Learning, RL)

1. **基本概念**



强化学习中，假如时间是离散的，称为时间步（time step）。

智能体（Agent）：执行强化学习的个体，游戏中为玩家，无人驾驶中为汽车，机器人捡垃圾中为机器人。

状态（State）：（第t个time step环境提供给智能体的状态），游戏中玩家的得分，无人驾驶中汽车是否出车祸了等。State会受到Action的影响，会有一定的概率（转移概率）转换成下一个State。

动作（Action）：（第t个time step中Agent的动作）。Agent根据环境的State和Reward执行的操作。

奖励（Reward）：（第t个time step环境给Agent的奖励）。由Agent获得，可以评价Agent的Action好坏，是否正确。Reward类似一种反馈机制，可以告诉Agent的Action是否正确。Reward有可能会延迟，即第t个time step没有奖励，则为0.Agent的目标就是最大化奖励。

一个典型的强化学习过程：

：

：

：

以此类推

强化学习与监督学习和非监督学习的一个重要区别就是：监督学习和非监督学习不需要与环境互动，只要静态数据就可以。

注意：如果希望通过强化学习来解决现实中的问题，则需要制定环境规则，并且指定状态State，动作Action和奖励Reward。

阶段性任务（Episodic Task）：具有清晰结束点的任务，例如游戏胜利或失败等。

连续性任务（Continous Task）：没有结束时间点，一直持续的任务。例如金融市场上买卖股票的算法。

强化学习的一个重要假设：Agent的目标始终可以描述为最大化期望累积奖励，称之为奖励假设。

回报（Return）：第t个time step中，过去的奖励，不再受Agent控制，只有未来的奖励才受Agent控制。令，称为回报。

通常，Agent无法肯定未来的奖励，只能依赖预测或估计。

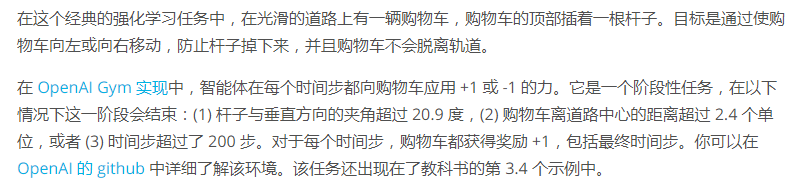
折扣回报（Discount Return）：第t个time step的回报，

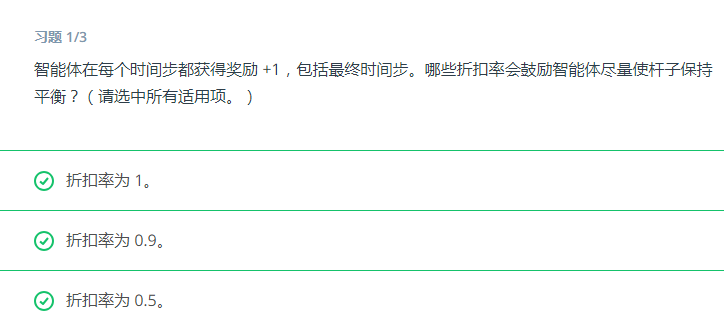


其中，称为折扣率(discount rate)

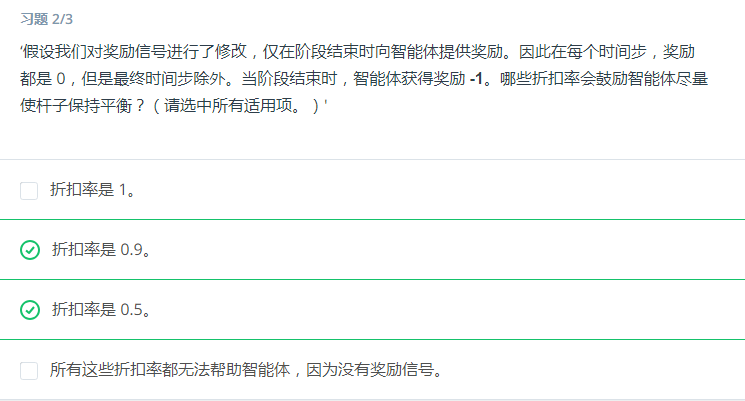
折扣回报让Agent更关心近期奖励，而不是遥远未来的奖励。

Udacity中折扣率与回报的例子：





原因：每个杆子保持平衡的时间步的奖励都是正面的，不管折扣率为多少，Agent都会尽量是杆子保持平衡。



原因：如果折扣率为1，则智能体始终获得负面奖励-1，有了折扣率后，智能体将尽量使杆子保持平衡，这样会有相对小的负面回报。



原因：如果折扣率为1，则智能体始终获得奖励+1，奖励信号将不会像智能体提供任何实用反馈，杆子是否平衡智能体不关心。

如果折扣率为0.5或0.9，最后1步的奖励为，k为time step个数，则智能体为了最大化奖励，就会尽快结束这一阶段（扔杆子等操作）。

所以需要重新设计奖励信号。

* 1. **马尔科夫决策过程(Markov Decision Process, MDP)**

定义：Agent处于环境E中，状态空间为S，其中每个状态是环境展示给Agent的环境描述。Agent能够采取的动作构造动作空间A，若某个动作作用在当前状态s上，则潜在的转移函数P将使得环境从当前状态按某种概率转移到另一个状态，在转移到另一个状态的同时，环境会根据潜在的奖励函数R反馈给Agent。

Udacity中的例子：机器人捡易拉罐

动作空间：机器人有3个动作，search(搜索),wait(等待)和recharge(充电)。



状态空间：机器人电量高high，电量低low



