彪哥带你学强化学习

3、强化学习基础概念(下)

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

讲师：韩路彪

回顾：基础模型——马尔科夫决策过程Markov decision process(MDP)

马尔科夫决策过程

马尔科夫奖励过程 + 策略

马尔科夫过程 + 奖励

单步状态依赖

S_t

$\pi(a_t|s_t)$

A_t

R_t

S_{t+1}

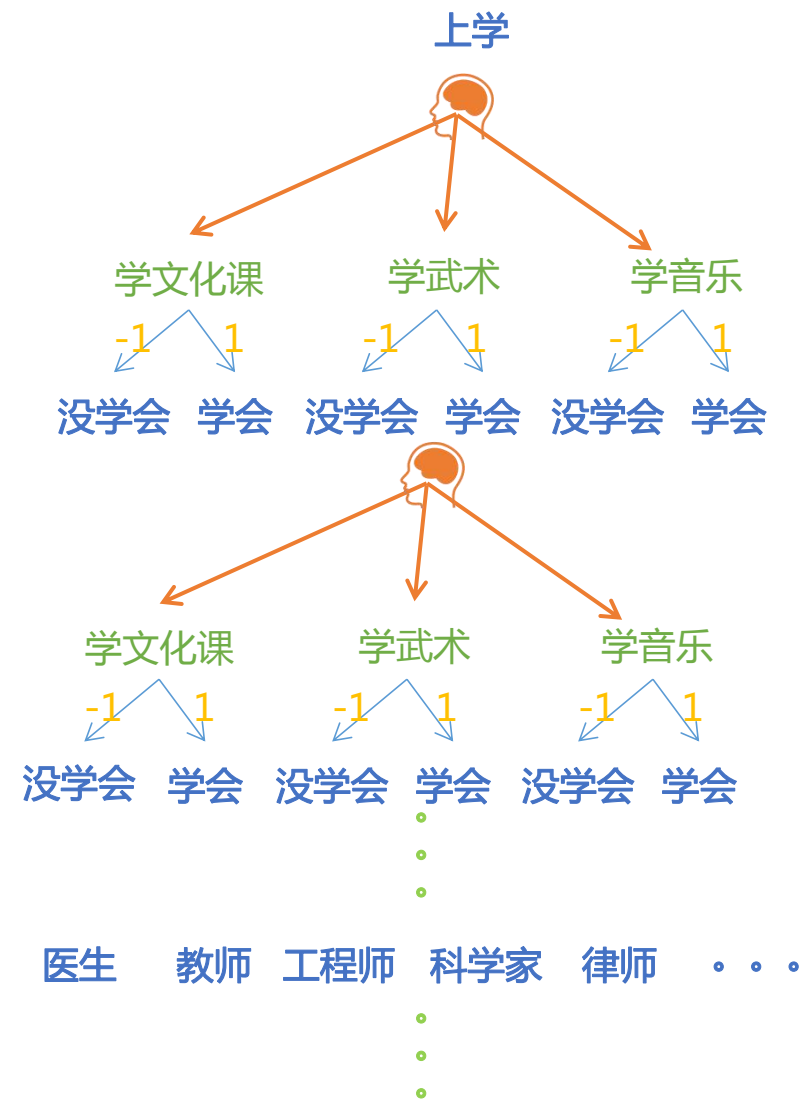
$\pi(a_{t+1}|s_{t+1})$

A_{t+1}

R_{t+1}

S_{t+2}

S_{t+m}



r_t ： $r(s_{t+1}|s_t, a_t)$

s_t $\pi(a_t|s_t)$ $p(s_{t+1}|s_t, a_t)$

折扣回报

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots$$

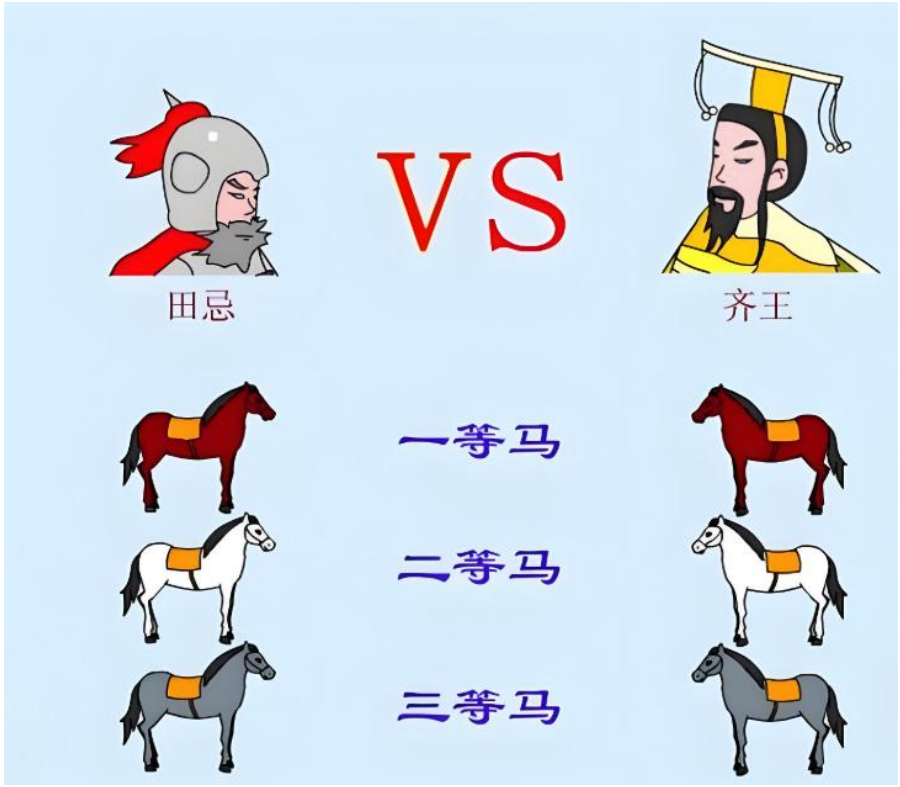
$$g_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots$$

回顾：强化学习是什么

研究对象　做一件事情的过程，核心是**序列**

研究目标　最优策略

监督学习	强化学习
单个值	序列
有明确结果	不容易给出每一步明确结果
需要标注好的数据	可以直接和环境交互获取数据
即时反馈	延迟反馈



价 值 函 数

为了可以比较和计算，引入奖励R/r

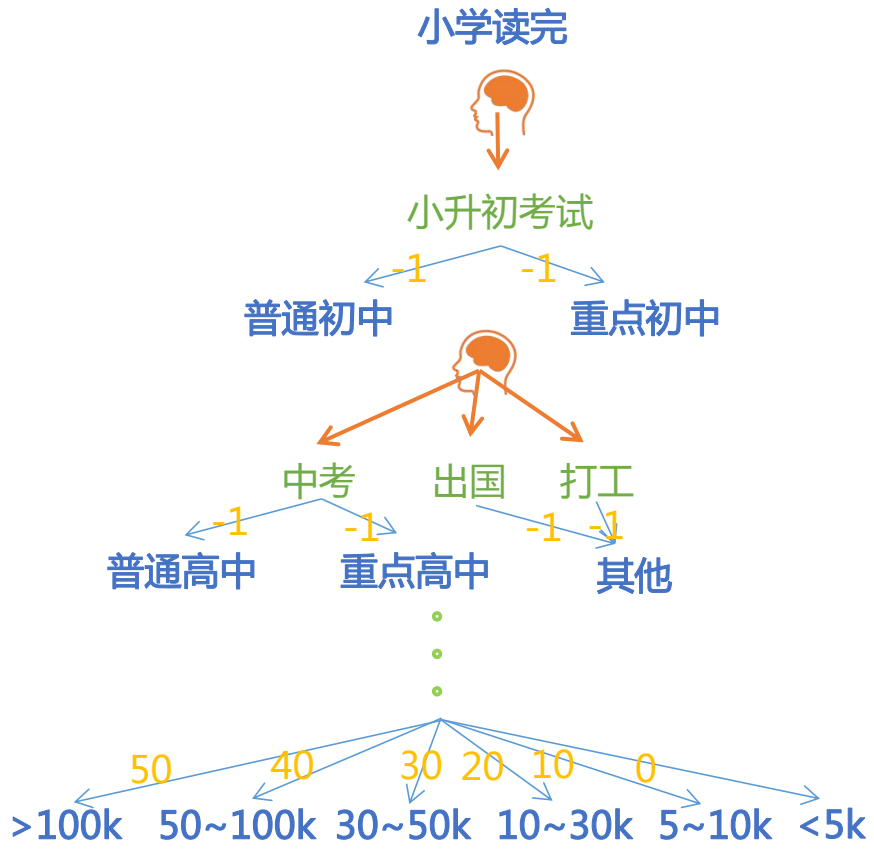
为了考虑长远奖励，引入回报G/g

$$g_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$$

回报存在的问题：

1. 针对单个样本
2. 事后才能知道结果，不容易指导策略
3. 方差大，不稳定

解决方法：引入价值函数



价 值 函 数

状态价值函数 $V^{\pi}(s_t) = \mathbb{E}_{\pi}(G_t|S_t = s_t)$

最优状态价值函数 $V^{*}(s_t) = \max_{\pi} V^{\pi}(s_t)$

状态动作价值函数 $Q^{\pi}(s_t, a_t) = \mathbb{E}_{\pi}(G_t|S_t = s_t, A_t = a_t)$

最优状态动作价值函数 $Q^{*}(s_t, a_t) = \max_{\pi} Q^{\pi}(s_t, a_t)$

最优策略 $\pi^{\star} = \arg\max_{\pi} Q^{\pi}(s_t, a_t)$

$$V^{\pi}(s_t) = \sum_{a \in \mathcal{A}} \pi(a|s_t) Q^{\pi}(s_t, a)$$
$$= \mathbb{E}_{a \sim \pi(\cdot|s_t)}\left[Q^{\pi}(s_t, a)\right]$$

$$V^{\star}(s_t) = \max_{a \in \mathcal{A}} Q^{\star}(s_t, a)$$