彪哥带你学强化学习

二、强化学习基础概念(上)

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

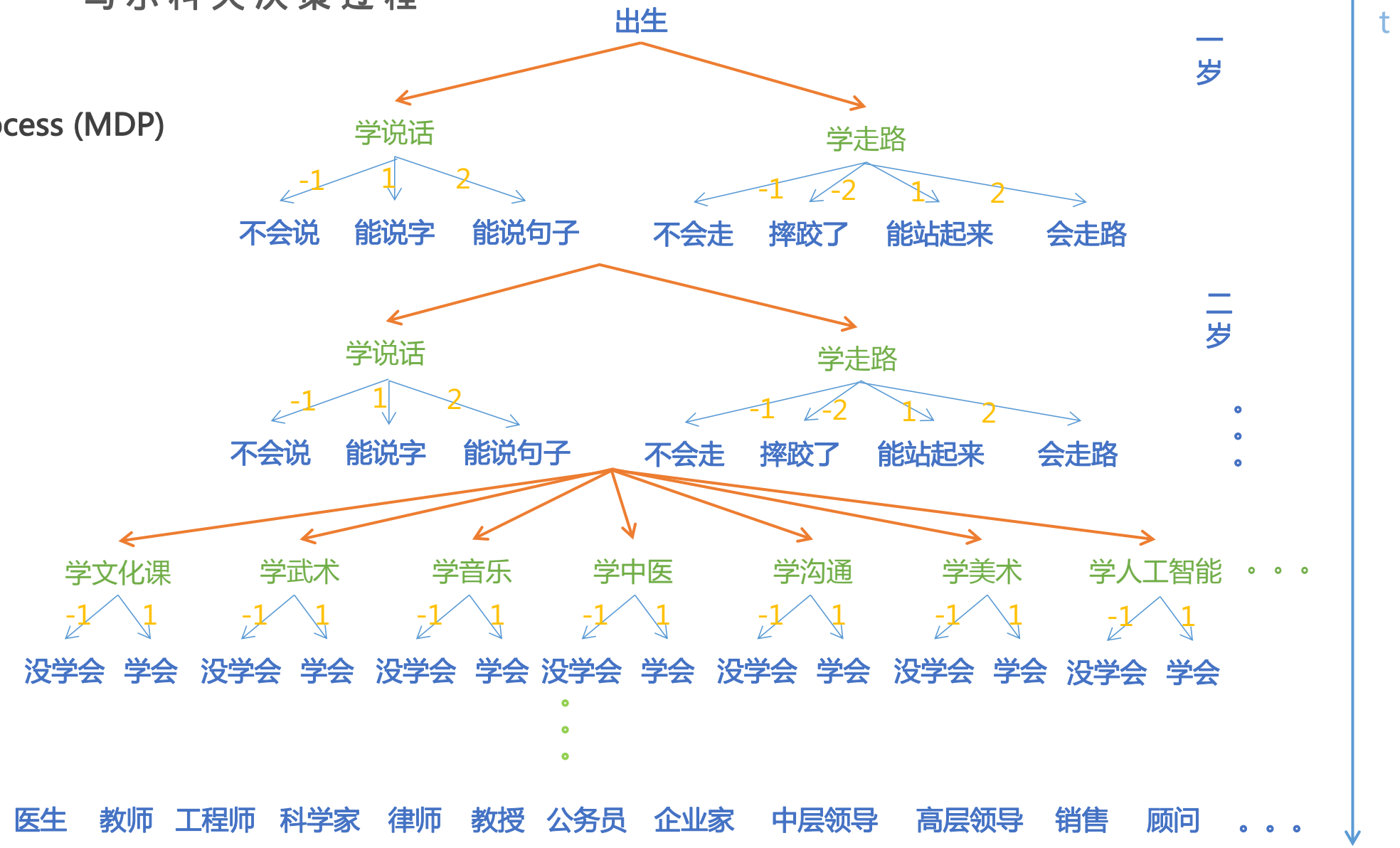
讲师：韩路彪

基础模型——马尔科夫决策过程

马尔科夫决策过程
Markov decision process (MDP)

状态(state) s
策略(policy) π
动作(action) a
奖励(reward) r
状态转移函数 p

状态空间 $\mathcal{S}$
动作空间 $\mathcal{A}$



基础模型——马尔科夫决策过程

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

基础模型——马尔科夫过程Markov Process(MP)

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

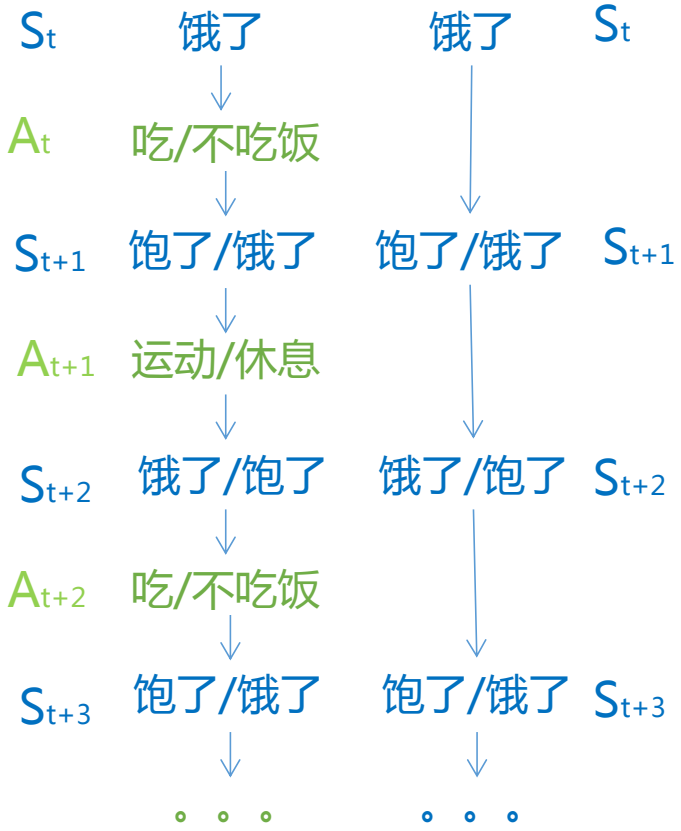
单步状态依赖

状态变量 S_t 今年年龄

具体状态 s_t 18

动作变量 A_t 象棋第一步棋

具体动作 a_t 单头炮



$$p(s_{t+1}|s_1,a_1,s_2,a_2,\ldots,s_t,a_t)=p(s_{t+1}|s_t,a_t)$$

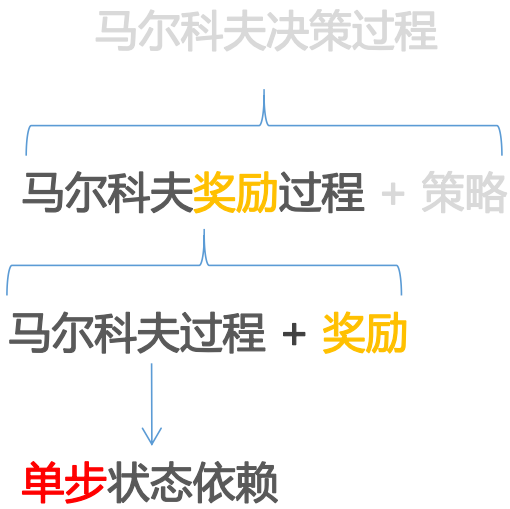
$$p(s_{t+1}|s_1,s_2,s_3,\ldots,s_t)=p(s_{t+1}|s_t)$$

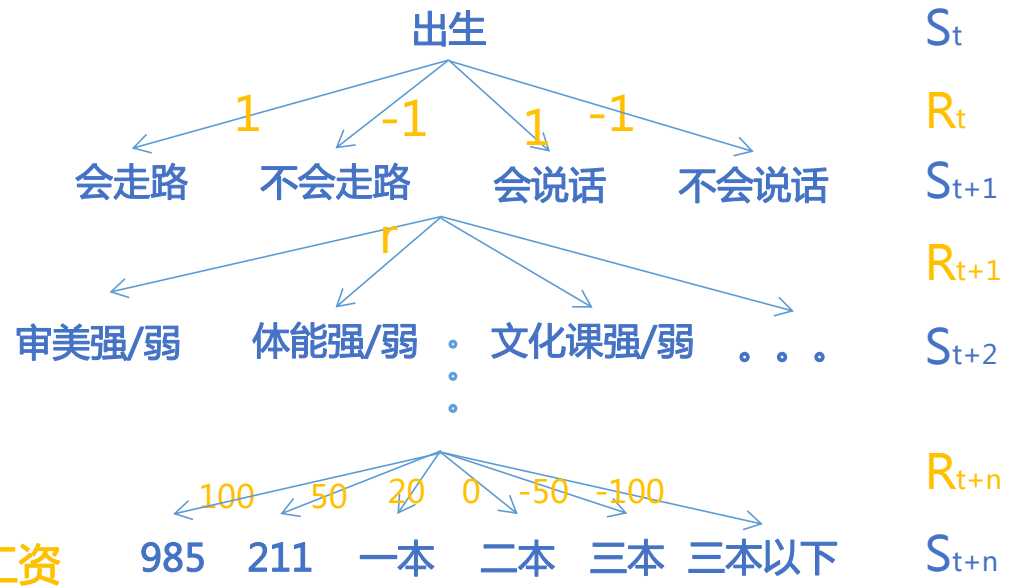
概率转移矩阵

$$\mathcal{P}=\begin{pmatrix} P(s_1|s_1) & \cdots & P(s_n|s_1) \\ \vdots & \ddots & \vdots \\ P(s_1|s_n) & \cdots & P(s_n|s_n) \end{pmatrix}$$

	饿了	饱了
饿了	0.1	0.9
饱了	0.9	0.1

基础模型——马尔科夫奖励过程Markov reward process(MRP)





回报

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

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

折扣回报

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

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

奖励变量	R_t	t时刻工资
具体奖励	r_t	5000
回报(return)	G_t	从t时刻到结束或终止的奖励和，把一个回合的奖励关联起来
回合(episodes)		一个采样，结果是一个具体的序列，比如张三完整的学习轨迹
具体回报	g_t	张三高考完了，反推他在刚上初中时候的回报
折扣回报	G_t	给未来的奖励打折扣
折扣率	γ	[0, 1]

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}

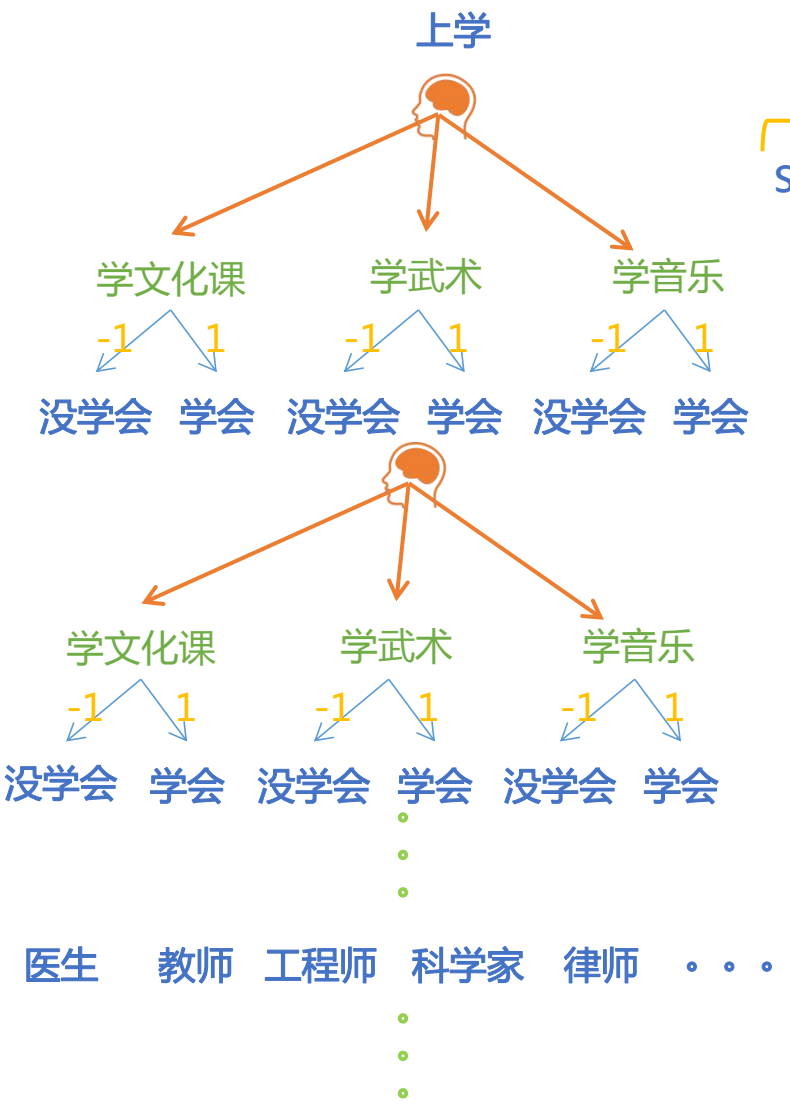
$\pi(a_{t+1}|s_{t+1})$

A_{t+1}

R_{t+1}

S_{t+2}

S_{t+m}



r_t : $r(s_{t+1}|s_t, a_t)$

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

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

$p(s_{t+2}|s_{t+1}, a_{t+1})$