

行为主义 \Rightarrow 进化主义；控制论学派 $\xrightarrow{\text{湘行}} \text{模拟论证}$

智能取决于感知和行动，提出智能行为的“感知-动作”模式

经典 RL: 输入: state-action space available action = f(state)

目的: 选择动作最大化预期未来奖励 \rightarrow 动态规划

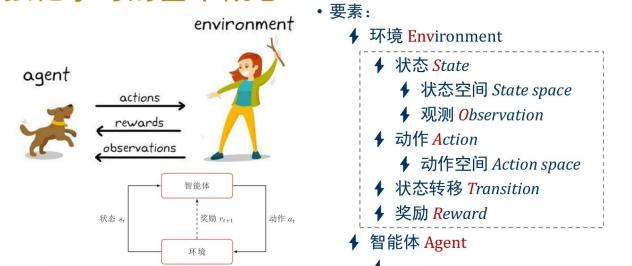
马尔可夫性: $P(S_{t+1} | S_t, A_t, \dots, S_0, A_0) = P(S_{t+1} | S_t, A_t)$

马尔可夫性: 模型已知，以行为序形成框架，求连续时间控制问题中的最优控制策略

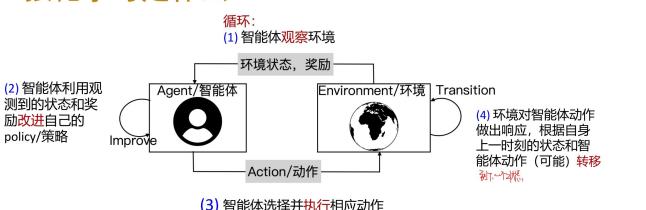
强化学习: Model-free, 通过与未知不确定环境直接交互学习一种行为使长期奖励最大化

概念: agent 与 env 交互作用的一种试错学习方式，以最大化预期(未来累积)奖励为目的

强化学习的基本概念



强化学习是什么



- 每个时刻 / time step, 都进行如上图所示的循环。因此, RL是用来解决序贯决策问题的。到下一
- 非常笼统地讲, RL包括给定状态预测奖励的预测算法和通过试错来学习好的动态控制问题(1)

启发: 反射; 操作反射; 拼布学习; 经验模型

RL基本流

下: 反向上有有关
马氏过程: 二元组 $\langle s, p \rangle$ 无记忆无后效

马氏决策过程: 五元组 $\langle S, A, P, R, T \rangle$

马尔可夫性: $P(S_{t+1} | S_t, A_t, \dots, S_0, A_0) = P(S_{t+1} | S_t, A_t)$

最优化控制: 模型已知, 以行为序形成框架, 求连续时间控

制问题中的最优控制策略

强化学习: Model-free, 通过与未知不确定环境直接交互学习一种行为使长期奖励最大化

概念: agent 与 env 交互作用的一种试错学习方式, 以最大化预期(未来累积)奖励为目的

要素:

环境 Environment

状态 State

动作 Action

状态转移 Transition

奖励 Reward

智能体 Agent

...

状态空间 State space

观测 Observation

动作空间 Action space

Transition

Reward

智能体 Agent

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...

...</p