Lecture Notes 1

2025 年 8 月 12 日

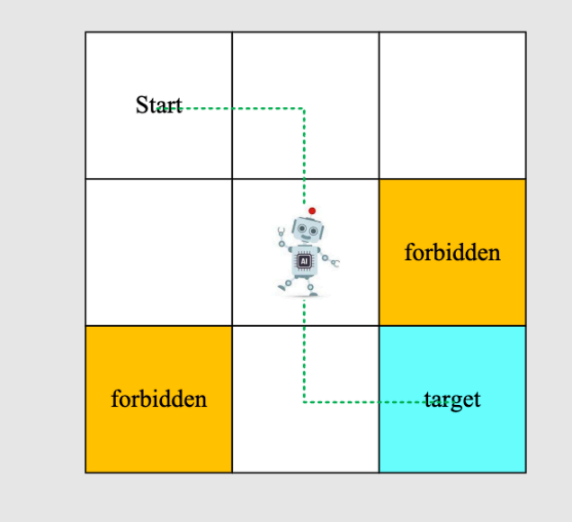
1 引言

强化学习的目标是求解最优策略。在此之前，我们需要一些工具来判断策略的好坏，以及什么是最优策略。这些工具包括状态值、贝尔曼公式（Bellman equation）以及贝尔曼最优公式。接下来需要求解具体的策略和状态值，有很多不同的方法。

2 第一章基本概念

2.1 Grid World example: find a good way to the target

state: 描述 agent 相对于 environment 的位置



state space: 状态的集合

action: 不同的状态可能有不同的 action。

action space: action 的集合。行动空间是状态空间的一个函数，因为不同的状态可能具有不同的行动。

state transition: 状态转移，在某状态时采取某行动会导致状态的转移。具体转移到哪个状态取决于状态空间

* 用概率描述状态转移是一种普适的方法

policy: 描述 agent 到了某个状态会实行某个策略，同样是以条件概率的形式

Mathematical representation: using conditional probability
For example, for state s_1:

$$\pi(a_1|s_1) = 0$$
$$\pi(a_2|s_1) = 1$$
$$\pi(a_3|s_1) = 0$$
$$\pi(a_4|s_1) = 0$$
$$\pi(a_5|s_1) = 0$$

It is a deterministic policy.

reward: 根据 agent 的行动结果作出奖励或者惩罚。agent 会尽量最大化奖励和最小化惩罚。同样可以使用条件概率进行表示

*reward 可以视为人类与 agent 交互的一种手段

trajectory: a state-action-reward transition

return: 计算沿着某个 trajectory 到达 target 时 agent 获得了多少奖励，可以用来评估 trajectory 的好坏

discount rate: 衰减率，$\gamma \in [0, 1)$。通过调整 γ，可以促使 agent 采取不同的策略。γ 越小，agent 就越近视。反之，则越远视。

episode: 一整个从开始到 target 的过程称为一个 episode。有限步数的任务称为 episodic task，而无限步数的任务称为 continuing task。

* 将 episodic task 转化为 continuing task 的方法：将 target 节点一般化，不视为最终节点，到达后可以进行移动