马尔可夫决策过程

惜日短

2025 年 9 月 19 日

摘要

学习是需要做一些总结的，尤其是人工智能相关，总是说学人工智能应该从实际项目入手，我也认可这个观点，但是基础理论知识也很重要。学习理论知识有一个弊端就是，人工智能涉及的知识面很广泛，即使是最简单的算法，可能用到的数学公式对于初学者而言有些摸不着头脑，起码对于我种数学基础比较差的来说，有很多数学公式很难看懂。不过一遍看不懂看两遍，看三遍……。最近也开始感觉到有些理解人工智能、机器学习、深度学习、强化学习了，似乎要打破瓶颈了吗？（芜湖……）目前在从强化学习入手。理论学习之初，会有大量的新概念，这些都需要理解与记忆，因此希望通过记笔记的形式进行巩固复习，加深记忆与理解。

强化学习的基本概念

强化学习包括两个部分：Agent(代理/智能体) 与 Environment(环境)

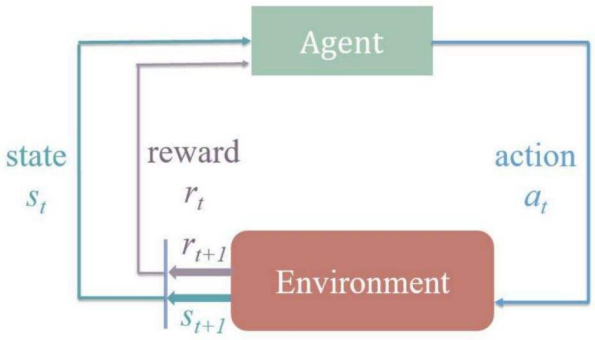

图 1: Agent-Environment

强化学习模型的基本流程是，Agent 通过与 Environment 交互，交互可以是 Agent 通过一些价值评估得到的 Action，Agent 通过 Action 与 Environment 进行交互，获得 Reward，

Reward 是一种标量反馈信号 (**scalar feedback signal**)，Reward 具有滞后性，同时我们希望最大化 Reward。Agent 的 Action 是离散的随机变量（也可以是连续的），那么 Agent 的 Action 就有取值范围，称为动作空间。在与 Environment 进行环境交互的过程中，Environment 也需要去描述，称为环境的状态（用 S^e 表示），Agent 对 Environment 的观测称为 Observation，Obervation 是对 S^e 的部分描述。

强化学习 Agent 的组成为**策略 (policy)**、**价值函数 (value function)**、**模型 (model)**。强化学习可以用元组来表示，一般包括环境状态 S，奖励 R，状态转移矩阵 P、动作 A、折扣因子 γ。

提示

自己写笔记的时候，我似乎有些明白为什么很多相关书籍都需要花费很大的篇幅来进行概述性的介绍了，上述只是一些重要的概念，还有很多内容需要补充描述，才能对强化学习有一个更好的了解。

1 马尔可夫决策过程 (Marcov Decision Process)

这里依旧是对马尔可夫决策过程的大量介绍。

1.1 马尔可夫过程 (Marcov Process)

1.1.1 马尔可夫性质 (Marcov Property)

一个随机过程中，给定现在的状态以及过去的所有状态，其未来状态的**条件概率 (conditional probability)** 分布只与当前状态有关。这个性质对于后面贝尔曼方程的推到非常重要，也就是条件概率的计算。

$$p(\mathcal{X}_{t+1}|\mathcal{X}_{0:t} = x_{0:t}) = p(\mathcal{X}_{t+1} = x_{t+1}|\mathcal{X}_t = x_t) \tag{1}$$

1.1.2 马尔可夫链 (Marcov Chains)

马尔可夫过程 (Marcov Process) 是一组具有马尔可夫性质的随机变量序列 $s_1, \ldots\ldots, s_t$。满足以下条件：

$$p(s_{t+1}|s_t) = p(S_{t+1}|h_t) \quad (h_t = s_1, s_2, \ldots\ldots, s_t) \tag{2}$$

▌ 注意： 离散时间的马尔可夫过程也称为马尔可夫链。

通过**状态转移方程 (state transition matrix)P** 来描述状态转移 $p(s_{t+1} = s^{'}|s_t = s$:

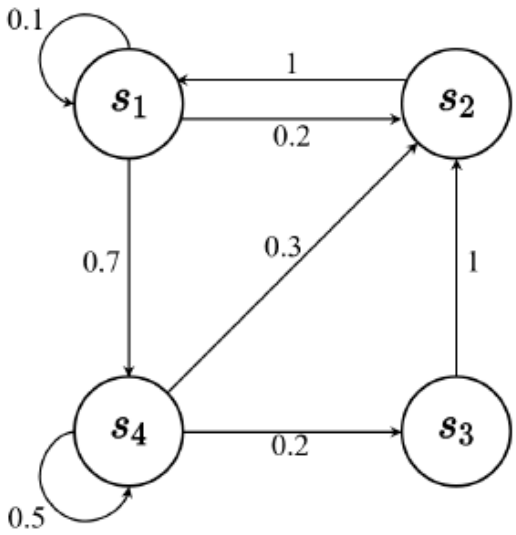

图 2: Marcov Chains

$$\mathbf{P} = \begin{bmatrix} p(s_1|s_1) & p(s_2|s_1) & \dots & p(s_n|s_1)) \\ p(s_1|s_2) & p(s_2|s_2) & \dots & p(s_n|s_2)) \\ \vdots & \vdots & \ddots & \vdots \\ p(s_1|s_n) & p(s_2|s_n) & \dots & p(s_n|s_n)) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.2 & 0 & 0.7 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.3 & 0.2 & 0.5 \end{bmatrix}$$

1.2 马尔可夫奖励过程 (Marcov Reward Process,MRP)

马尔可夫奖励过程 = 马尔可夫链 + 奖励函数 (reward function)

▌ 注意： 奖励函数是 **R** 的期望

1.2.1 回报 (Return/Gain) 与价值函数 (Value fuction)

范围 (horizon) 是指一个回合 (episode) 的长度，回报为奖励的逐步叠加，假设时刻 t 后的奖励序列为 $r_{t+1}, r_{t+2}, \ldots$，则回报的定义为：

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{T-t-1} r_T = \sum_{i=0}^{n} \gamma^i r_{t+1+i}$$

γ 是折扣因子 (discount factor)

有了回报，定义状态价值函数 (state-value function)，对于马尔可夫奖励过程，状态价值函数定义为对回报的期望，即：

$$\mathbf{V}^t(s) = \mathbb{E}[\mathbf{G}_t | s_t = s] \tag{3}$$

$$= \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{T-t-1} r_T | s_t = s] \tag{4}$$

使用折扣因子 γ(discount factor) 的原因

1. 避免环状马尔可夫过程无穷的奖励
2. 模拟环境的模型并不是准确的
3. 更倾向于即时奖励

> 注意： 折扣因子 γ 是一个超参数 (hyperprameter)

价值的计算通过采样获取随机过程，根据每个随机事件的奖励，通过 G 的计算公式计算回报。采样求和取平均值以计算价值函数，即蒙特卡洛 (Monte Carlo) 方法。

1.2.2 贝尔曼方程 (Bellman Equation)

从价值函数中推倒贝尔曼方程 (Bellman Equation):

$$\mathbf{V}(\mathbf{s}) = \underbrace{\mathbf{R}(s)}_{\text{即时奖励}} + \underbrace{\gamma \sum_{s' \in S} p(s_{'}|s) V(s_{'})}_{\text{未来奖励的折扣总和}}$$

全期望公式 (law of total expectation)

$$\mathbb{E}[V(s_{t+1})|s_t] = \mathbb{E}[\mathbb{E}[G_{t+1}|s_{t+1}]|s_t] = \mathbb{E}[G_{t+1}|s_t]$$

全期望公式的数学表达 $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$，即 X 关于条件 Y 期望的期望。

贝尔曼方程的推导

$$
\begin{aligned}
V(s) &= \mathbb{E}[G_t|s_t = s] && (5)\\
&= \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots |s_t = s] && (6)\\
&= \mathbb{E}[r_{t+1}|s_t = s] + \mathbb{E}[\gamma^2 r_{t+3} + \ldots |s_t = s] && (7)\\
&= R(s) + \gamma \mathbb{E}[G_{t+1}|s_t = s] && (8)\\
&= R(s) + \gamma \mathbb{E}[V(s_t + 1)|s_t = s] && (9)\\
&= R(s) + \gamma \sum_{s' \in S} p(s^{'}|s)V(s^{'}) && (10)
\end{aligned}
$$

贝尔曼方程的向量形式

$$\mathbf{v}_\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}_\pi$$

其中：

- $\mathbf{v}_\pi \in \mathbb{R}^{|\mathcal{S}|}$ 是状态价值向量，满足 $\mathbf{v}_\pi(s) = V^\pi(s)$；
- $\mathbf{r}_\pi \in \mathbb{R}^{|\mathcal{S}|}$ 是期望即时奖励向量，$\mathbf{r}_\pi(s) = \mathbb{E}_\pi[R_{t+1} \mid S_t = s]$；
- $\mathbf{P}_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ 是状态转移矩阵，$\mathbf{P}_\pi(s, s') = \sum_a \pi(a|s)P(s'|s, a)$；
- $\gamma \in [0, 1)$ 是折扣因子。

当矩阵 $\mathbf{I} - \gamma \mathbf{P}_\pi$ 可逆时，可得闭式解：

$$\mathbf{v}_\pi = (\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1} \mathbf{r}_\pi$$

计算马尔可夫奖励过程价值的迭代方法

1. 动态规划
2. 蒙特卡洛
3. 时序差分学习 (temporal-difference learning, TD learning)(动态规划与蒙特卡洛的结合)