2 - 马尔科夫决策过程

马尔科夫决策过程（Markov decision process, MDP），它能够一数学的形式来表达序列决策过程。

马尔科夫决策过程

马尔科夫决策过程是强化学习的基本问题模型之一，它能够以数学的形式来描述智能体在与环境交互的过程中学到一个目标的过程。

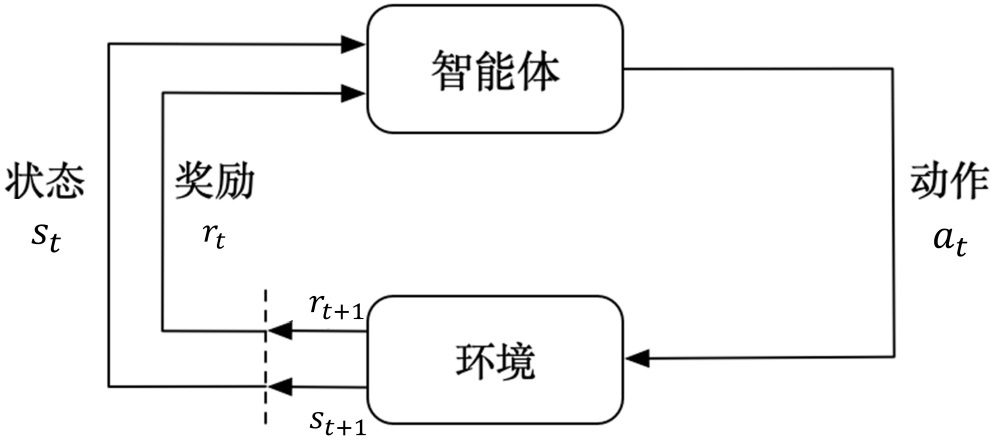

图 2-1 马尔可夫决策过程中智能体与环境的交互过程

这个决策过程是发生在离散的时步，并使用 t 来表示，$t=0,1,2,\dots$ 。在每一个时步，智能体会接受到当前环境的状态 s_t ，并根据状态 s_t 执行动作 a_t 。执行动作 a_t 对环境产生变化次那个给予奖励 r_t 。我们由此可以得出循环表示：

$$s_t, a_t, r_{t+1}$$

> 奖励包含正奖励和负奖励。马尔科夫决策过程就是智能体以最大化正奖励为目标的过程。

马尔科夫性质

马尔可夫决策过程的一个前提，即马尔可夫性质，如下式所示：

$$P(s_{t+1}|s_t) = P(s_{t+1}|s_0, s_1, \dots, s_t)$$

这个公式说明了马尔科夫决策过程会根据当前状态以做出对于下一个时步最有利的决定。

回报

前面讲到在马尔可夫决策过程中智能体的目标时最大化正奖励，一般累计的正奖励称为回报（Return），用 G_t 表示，最简单的回报公式可以表示为

$$G_t = r_1 + r_2 + \dots + r_t$$

其中 t 表示时步，意最大步数。这个公式其实只适用于有限步数的情景，也就是拥有终止状态或结束条件的情况。

当我们在持续性任务是 $t=\infty$ ，会导致上面的回报公式出问题，因为长此以往的计算会导致 $G_t=\infty$ 。为了解决这个问题，我们引入折扣因子（Discount factor）γ ，并将回报公式表示为：

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

其中 γ 取值范围在 $[0,1]$，它表示了我们对于未来奖励的重视程度，控制着当前奖励和未来奖励之间的权衡。

这样做能够让当前时步的回报 G_t 与下一个时步 G_{t+1} 的回报有所关联的，即：

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ldots \\ &= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \ldots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

状态转移矩阵

截至目前，我们讲的都是有限状态马尔可夫决策过程（Finite MDP），这指的是状态的数量必须是有限的过程。这个过程又被称为连续时间马尔可夫过程，它允许发生无限次事件。

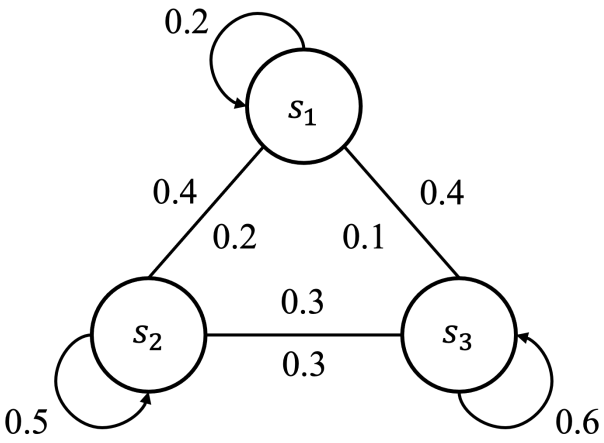

图 2.2 马尔可夫链

如图2.2所示，我们能将这个决策过程的有限数量状态以图表示，而其中数字是各个状态之间进行转移的概率，我们称之为状态转移概率，此概率能以公式表示为：

$$P_{ss'} = P(S_{t+1} = s' | S_t = s)$$

使用数学矩阵表示法可以表示为以下状态转移矩阵：

$$P_{ss'} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}$$

其中 n 表示了状态的数量。还有，状态转移矩阵是环境的一部分，智能体会根据状态转移矩阵来做出决定。

在此基础上增加奖励元素就会形成马尔科夫奖励过程（Markov reward process，MRP）

练习题

强化学习所解决的问题一定要严格满足马尔可夫性质吗？请举例说明。

不一定。
比如我们所熟知的棋类游戏，因为在我们决策的过程中不仅需要考虑到当前棋子的位置和对手的情况，还需要考虑历史走子的位置例如吃子等。
这意味着在特定情况如果想做出对于当前状态最有利的决定，历史状态也应纳入决策过程。

马尔可夫决策过程主要包含哪些要素？

时步，状态，奖励，动作。

马尔可夫决策过程与金融科学中的马尔可夫链有什么区别与联系？

- 区别：
  马尔可夫决策过程更关注于描述各个状态之间的转移。
  金融科学中的马尔可夫链更关注观察到的市场状态的演化。
- 联系：
  马尔科夫决策过程和金融中的马尔可夫链都涉及到随机过程，其中未来的状态依赖于当前状态。