Introduction to Reinforcement Learning

马尔可夫决策过程

何新卫

信息学院，
华中农业大学

2024 年 3 月 6 日

Part I

CHAPTER3: FINITE MARKOV DECISION PROCESSES (FINITE MDPS)

背景介绍

图. 贫富差距背后的哲学

如何建模强化学习中的环境？是本章关注的重点内容

随机过程

- 随机过程（stochastic process）是概率论的“动力学”部分。概率论的研究对象是静态的随机现象，而随机过程的研究对象是随时间演变的随机现象（例如天气随时间的变化、城市交通随时间的变化）。
- 在随机过程中，随机现象在某时刻 t 的取值是一个向量随机变量，用 S_t 表示（注：大写的为随机变量），所有可能的状态组成状态集合 $\mathcal{S}$。
- 随机现象便是状态的变化过程。在某时刻 t 的状态 S_t 通常取决于 t 时刻之前的状态。我们将已知历史信息 $(S_1, S_2, ..., S_t)$ 时下一个时刻状态 S_{t+1} 为的概率表示成 $P(S_{t+1}|S_1, S_2, ..., S_t)$

马尔可夫性质

- 当且仅当某时刻的状态只取决于上一时刻的状态时，一个随机过程被称为具有马尔可夫性质（Markov property），
$$P(S_{t+1}|S_t) = P(S_{t+1}|S_1, ..., S_t)$$
- 也就是说，当前状态是未来的充分统计量，即下一个状态只取决于当前状态，而不会受到过去状态的影响
- 需要明确的是，具有马尔可夫性并不代表这个随机过程就和历史完全没有关系。因为虽然时刻 $t+1$ 的状态只与时刻 t 的状态有关，但是时刻 t 的状态其实包含了 $t-1$ 时刻的状态的信息。通过这种链式的关系，历史的信息被传递到了现在。
- 马尔可夫性可以大大简化运算，因为只要当前状态可知，所有的历史信息都不再需要了，利用当前状态信息就可以决定未来。

马尔可夫过程

- 马尔可夫过程（Markov process）指具有马尔可夫性质的随机过程，也被称为马尔可夫链（Markov chain）
- 我们通常用元组 $< \mathcal{S}, \mathcal{P} >$ 描述一个马尔可夫过程，其中 $\mathcal{S}$ 是有限数量的状态集合，$\mathcal{P}$ 是是状态转移矩阵（state transition matrix）。
- 假设一共有 n 个状态，此时 $\mathcal{S} = s_1, s_2, ..., s_n$(注：小写的非随机变量)，则状态转移矩阵 P 定义了所有状态之间的转移概率，即: $P = \begin{bmatrix} p(s_1|s_1) & p(s_2|s_1) & p(s_3|s_1) & \cdots & p(s_n|s_1) \\ p(s_1|s_2) & p(s_2|s_2) & p(s_3|s_2) & \cdots & p(s_n|s_2) \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ p(s_1|s_n) & p(s_2|s_n) & p(s_3|s_n) & \cdots & p(s_n|s_n) \end{bmatrix}$
- 矩阵 P 中第 i 行第 j 列元素 $P(s_j|s_i) = P(S_{t+1} = s_j|S_t = s_i)$ 表示从状态 s_i 转移到状态 s_j 的概率（注意：S_{t+1} S_t 为随机变量，s_j 和 s_i 为随机变量取某个具体数值），我们称 $P(s'|s)$ 为状态转移函数。从某个状态出发，到达其他状态的概率和必须为 1，即状态转移矩阵的每一行的和为 1。

作业题 1

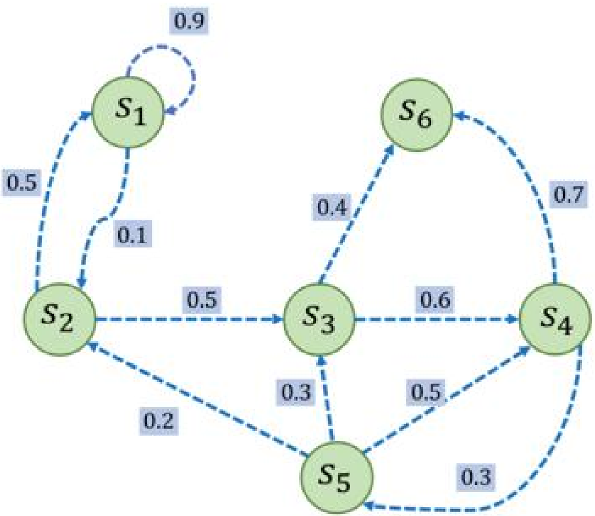

图. 上图是一个具有 6 个状态的马尔可夫过程的简单例子。其中每个绿色圆圈表示一个状态，每个状态都有一定概率（包括概率为 0）转移到其他状态，其中 s_6 通常被称为终止状态（terminal state），因为它不会再转移到其他状态，可以理解为它永远以概率 1 转移到自己

写出上述图中的状态转移矩阵

采样

- 给定一个马尔可夫过程，我们就可以从某个状态出发，根据它的状态转移矩阵生成一个状态序列（episode），这个步骤也被叫做采样（sampling）。
- 通过对状态的采样，我们可以生成很多这样的轨迹。
- 课堂练习题 从 s_1 出发，采样生成状态演化序列。

采样例题

有很多答案，比如 $s_1 \to s_2 \to s_3 \to s_6$ 或序列 $s_1 \to s_1 \to s_2 \to s_3 \to s_4 \to s_5 \to s_3 \to s_6$ 等。生成这些序列的概率和状态转移矩阵有关

马尔可夫奖励过程

在马尔可夫过程的基础上加入奖励函数 r 和折扣因子 γ，就可以得到马尔可夫奖励过程（Markov reward process）。一个马尔可夫奖励过程由 $<\mathcal{S}, \mathcal{P}, r, \gamma>$ 构成，各个组成元素的含义如下所示。

- $\mathcal{S}$ 是有限状态的集合。
- $\mathcal{P}$ 是状态转移矩阵。
- r 是奖励函数，某个状态的奖励指转移到该状态时可以获得奖励的期望。
- γ 是折扣因子（discount factor），的取值范围为 [0,1)。引入折扣因子的理由为远期利益具有一定不确定性，有时我们更希望能够尽快获得一些奖励，所以我们需要对远期利益打一些折扣。接近 1 的 γ 更关注长期的累计奖励，接近 0 的 γ 更考虑短期奖励。

回报

- 回报: 在一个马尔科夫奖励过程汇总，从 t 时刻，状态 S_t 开始，直至最终状态，所有奖励（$R_{t+1}, R_{t+2}, R_{t+3}, ...,$）的衰减之和统称为回报 G_t （Return），公式如下：

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

  其中，R_{t+1} 表示 t 时刻获得的奖励。
- 我们只需要沿用马尔可夫过程，在其基础上添加奖励函数，即可构成一个马尔可夫奖励过程。例如，进入状态 s_2 可以获得的奖励为-2，表面我们不希望进入这个状态，进入 s_4 后可获得最高奖励为 10，但是进入 s_6 后奖励为 0，此时终止序列也就终止了。

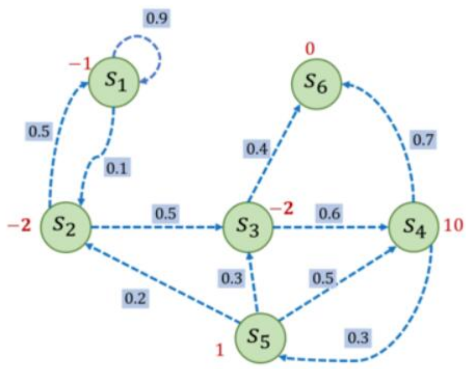

图. 马尔可夫奖励过程

马尔可夫奖励过程

作业题 2

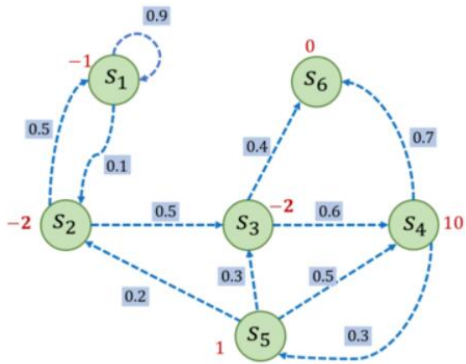

图. 马尔可夫奖励过程

对于上面马尔科夫奖励过程，假设从 s_1 为起始状态，设置 $\gamma = 0.5$, 采样得到一条状态序列为 $s_1 \to s_2 \to s_3 \to s_6$，请计算每个状态的回报。

马尔可夫奖励过程

作业题 2

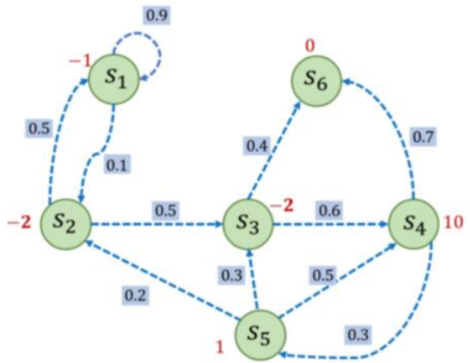

图. 马尔可夫奖励过程

对于上面马尔科夫奖励过程，假设从 s_1 为起始状态，设置 $\gamma = 0.5$, 采样得到一条状态序列为 $s_1 \to s_2 \to s_3 \to s_6$，请计算每个状态的回报。答案：$G_1 = -1 + 0.5 \times (-2) + 0.5^2 \times (-2) = -0.25....$

估算回报

- 可以采样很多轨迹，估算状态回报。