WDS

DataScience@Web

第1部分

马尔科夫决策过程

李蕾

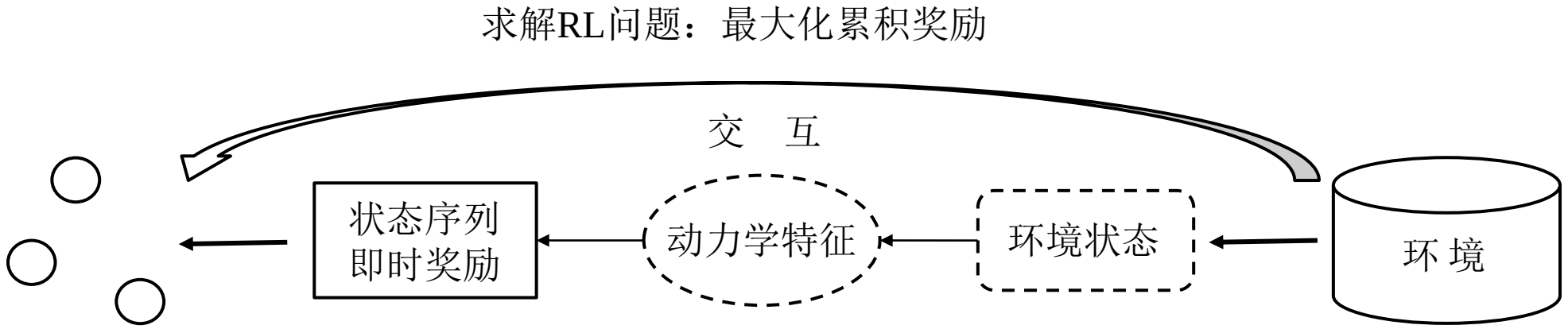
2021.07.06

2021 WDS暑期讨论班

目录

- 概述
- 马尔科夫过程
- 马尔科夫奖励过程
- 马尔科夫决策过程
- 最优策略

概述



环境状态完全可观测：构建马尔科夫决策过程来描述RL问题

环境状态不完全可观测：结合自身对环境的历史观测数据来构建一个近似完全可观测环境的描述

强化学习问题 →马尔科夫决策过程

目录

马尔科夫过程

- **马尔科夫性（Markov property）：**
  一个时序过程中，若t+1时刻的状态只取决于t时刻的状态S_t，而与之前的任何状态都无关时，认为t时刻的状态S_t具有马尔科夫性
- **马尔科夫过程（Markov process）**
  过程中的每一个状态都具有马尔科夫性，则这个过程具备马尔科夫性。
  又称马尔科夫链

描述：$< S, P >$， S是有限数量的状态集， P是状态转移概率矩阵

$$P_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$$

马尔科夫过程

$P_{ss'}$定义了从任意一个状态s到其所有后继状态s'的状态转移概率：

$$P = from \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \quad \text{to}$$

- **采样（sample）：**
  从符合马尔科夫过程给定的状态转移概率矩阵生成一个状态序列的过程
- **状态序列（episode）：**
  采样得到的一系列的状态转换过程
  状态序列最后一个状态是终止状态时，称为完整的状态序列（compete episode）

目录

马尔科夫奖励过程

马尔科夫奖励过程（Markov reward process，MRP）

把奖励反馈考虑进马尔科夫过程，则称为马尔科夫奖励过程

描述：$< S, P, R, \gamma >$

- S是一个有限状态集
- P是集合中状态转移概率矩阵：$P_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$
- R是一个奖励函数：$R_s = \mathbb{E}[R_{t+1} | S_t = s]$
- γ是一个衰减因子：$\gamma \in [0,1]$

马尔科夫奖励过程

- **收获/回报（return）**
  是一个MRP过程中从某一个状态S_t开始采样直到终止状态时所有奖励的有衰减的和

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

  - 计算某一状态到结束状态的累积奖励
  - 引入衰减系数γ使得后续某一状态对当前状态的贡献小于其奖励
    - γ取0：当前状态的收获即为当前状态获得的奖励，“短视”
    - γ取1：考虑后续所有状态，“长远眼光”

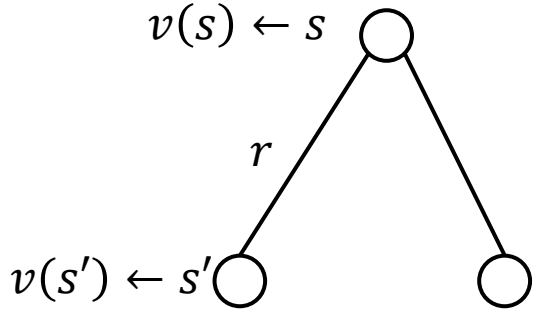
马尔科夫奖励过程

- **价值（value）**
  MRP中状态收获的期望

$$v(s) = \mathbb{E}[G_t | S_t = s]$$

  - 价值函数（value function）
    建立了从状态到价值的映射，给定一个状态能够得到该状态对应的价值
  - 将价值函数中的收获G_t按照定义展开：

$$\begin{aligned} v(s) &= \mathbb{E}[G_t | S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots) | S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) | S_t = s] \end{aligned}$$



用s'表示s状态下一时刻任一可能的状态：

$$v(s) = R_s + \gamma \sum_{s' \in S} P_{ss'} v(s')$$

马尔科夫奖励过程——贝尔曼方程

- **贝尔曼方程（Bellman equation）**
  一个状态的价值由该状态的奖励及后续状态价值按概率分布求和并衰减后联合而成

$$v(s) = R_s + \gamma \sum_{s' \in S} P_{ss'} v(s')$$

$$v = R + \gamma P v$$

$$\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} R(1) \\ \vdots \\ R(n) \end{bmatrix} + \gamma \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix}$$

理论上，该方程可直接求解：

$$v = (1 - \gamma P)^{-1} R$$

目录

马尔科夫决策过程

马尔科夫决策过程（Markov decision process, MDP）

将个体行为的选择考虑进马尔科夫奖励过程中

描述：$< S, A, P, R, \gamma >$

- S是一个有限状态集
- A是一个有限行为集
- P是集合中基于行为的状态转移概率矩阵：$P_{ss'}^{a} = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$
- R是基于状态和行为的奖励函数：$R_s^a = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$
- γ是一个衰减因子：$\gamma \in [0,1]$

马尔科夫决策过程

- **策略（policy）**
  个体在给定状态下从行为集中选择一个行为的依据称为策略，用π表示
  策略π是某一状态下基于行为集合的一个概率分布：

$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$$

  - 策略仅通过依靠当前状态就可以产生一个个体的行为，与历史状态无关
  - 策略描述的是个体的行为产生的机制，是不随状态变化而变化的，被认为是静态的

给定一个MDP：$M =< S, A, P, R, \gamma >$和一个策略$\pi$，对于一个符合MRP$< S, P_\pi, R_\pi, \gamma >$的采样$S_1, R_2, S_2, R_3, \ldots$序列来说，该奖励过程需要满足：

$$P^{\pi}_{s,s'} = \sum_{a \in A} \pi(a|s) P^{a}_{ss'} \qquad R^{\pi}_{s} = \sum_{a \in A} \pi(a|s) R^{a}_{s}$$

马尔科夫决策过程

- **状态价值函数**
  价值函数$v_\pi(s)$是在马尔科夫决策过程下基于策略π的状态价值函数，表示从状态s开始，遵循当前策略π所获得的收获的期望：
  $$v_\pi(s) = \mathbb{E}[G_t|S_t = s]$$

- **行为价值函数**
  基于策略π的行为价值函数$q_\pi(s,a)$，表示在遵循策略π时，对当前状态s执行某一具体行为a所能得到的收获的期望：
  $$q_\pi(s,a) = \mathbb{E}[G_t|S_t = s, A_t = a]$$

贝尔曼期望方程：

$$v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]$$
$$q_\pi(s,a) = \mathbb{E}[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a]$$

马尔科夫决策过程

- 状态的价值$v_\pi(s)$可以用该状态下所有行为价值$q_\pi(s,a)$来表达

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s,a)$$

- 行为价值$q_\pi(s,a)$可以用该行为所能到达的后续状态的价值$v_\pi(s)$来表达

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')$$

组合两式可以获得：

$$v_\pi(s) = \sum_{a \in A} \pi(a|s)(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s'))$$

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \sum_{a' \in A} \pi(a'|s') q_\pi(s',a')$$

目录

最优策略

强化学习问题的解 → 让个体与环境交互时获得最大收获的最优策略π^*

- **最优状态价值函数（optimal value function）**
  所有策略下产生的众多状态价值函数中的最大者：$v_* = \max_\pi v_\pi(s)$
- **最优行为价值函数（optimal action-value function）**
  所有策略下产生的众多行为价值函数中的最大者：$q_*(s,a) = \max_\pi q_\pi(s,a)$
- **最优策略（optimal policy）**
  若对于有限状态集中的任意一个状态s，不等式$v_\pi(s) \geq v_{\pi'}(s)$成立，策略$\pi$优于$\pi'$

对于任何MDP，存在一个最优策略π_*优于或至少不差于所有其他策略

最优策略

最优策略可以通过最大化最优行为价值函数$q_*(s,a)$来获得：

$$\pi_*(a|s)=\begin{cases}1 \text{ 如果} a=arg\max\limits_{a\in A} q_*(s,a)\\ 0 \qquad\qquad\qquad\qquad \text{其他情况}\end{cases}$$

- 在最优行为价值函数已知时，在某一状态s下，对于行为集里的每个行为a都对应一个最优行为价值$q_*(s,a)$
- 最优策略$\pi_*(a|s)$会给予所有最优行为价值中的最大值进行的行为100%的概率

强化学习问题 → 求解最优行为价值函数

最优策略

状态价值的最大值由以下贝尔曼最优方程得到：

$$v_*(s) = \max_a q_*(s,a)$$

- 状态的最优价值是该状态下所有行为对应的最优价值的最大值

行为价值的最大值由以下贝尔曼最优方程得到：

$$q_*(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s')$$

即时奖励　　状态价值×概率×衰减

最优策略

状态的最优价值可以通过其后续可能状态的最优价值计算得到：

$$v_*(s) = \max_a (R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s''))$$

最优行为价值函数可以由后续的最优行为价值函数来计算得到：

$$q_*(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \max_{a'} q_*(s',a')$$

迭代法：价值迭代、策略迭代、Q学习、Sarsa学习等
函数逼近：线性函数逼近/非线性函数逼近（深度神经网络）

总结

- 利用马尔科夫决策过程为强化学习问题建模
- 将决策的奖励反馈考虑进马尔科夫过程则构成MRP
  将个体行为考虑进MRP则构成MDP
- 最优策略中将强化学习问题转化成最优行为价值函数的求解

谢 谢