Lecture Notes 2

2025 年 8 月 12 日

1 第二章 贝尔曼公式

1.1 状态值及其定义

* 数学直观：return 为什么是重要的？

return 可以用来量化并评估某个 policy 是否是“好的”

v 为状态值，r 为当前状态的 reward，P 为与状态空间有关的矩阵，表明不同状态之间的关系

状态值：state value，是在采用某策略的条件下，某个状态出发得到的 return 的数学期望

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

* 状态值和 return 的区别：return 是根据某一条 trajectory 得到的确定的值，而状态值是一个符合概率分布的随机变量

1.2 贝尔曼公式的数学推导过程

贝尔曼公式的一般形式:

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

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[G_t|S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= \mathbb{E}[R_{t+1}|S_t = s] + \gamma\mathbb{E}[G_{t+1}|S_t = s] \end{aligned}$$

$$\begin{aligned}\mathbb{E}[R_{t+1}|S_t=s] &= \sum_a \pi(a|s)\mathbb{E}[R_{t+1}|S_t=s,A_t=a] \\ &= \sum_a \pi(a|s)\sum_r p(r|s,a)r\end{aligned}$$

$$\begin{aligned}\mathbb{E}[G_{t+1}|S_t=s] &= \sum_{s'} \mathbb{E}[G_{t+1}|S_t=s,S_{t+1}=s']p(s'|s) \\ &= \sum_{s'} \mathbb{E}[G_{t+1}|S_{t+1}=s']p(s'|s) \\ &= \sum_{s'} v_\pi(s')p(s'|s) \\ &= \sum_{s'} v_\pi(s')\sum_a p(s'|s,a)\pi(a|s)\end{aligned}$$

总结:

$$\begin{aligned}v_\pi(s) &= \mathbb{E}[R_{t+1}|S_t=s]+\gamma\mathbb{E}[G_{t+1}|S_t=s], \\ &= \underbrace{\sum_a \pi(a|s)\sum_r p(r|s,a)r}_{\text{mean of immediate rewards}} + \underbrace{\gamma\sum_a \pi(a|s)\sum_{s'} p(s'|s,a)v_\pi(s')}_{\text{mean of future rewards}}, \\ &= \sum_a \pi(a|s)\left[\sum_r p(r|s,a)r+\gamma\sum_{s'} p(s'|s,a)v_\pi(s')\right], \quad \forall s\in\mathcal{S}.\end{aligned}$$

贝尔曼公式用于计算某个状态的状态值。描述了不同状态的状态值之间的关系

* 贝尔曼公式对状态空间的所有状态都适用

* 通过求解状态值，可以用来评估 policy（即公式中的 π）

* 公式中的 $p(r|s,a)$ 和 $p(s'|s,a)$ 代表 dynamic model（或 environment model)

* 通过对比不同策略下同一个状态的状态值，就可以知道相对而言策略的好坏

1.3 贝尔曼公式的矩阵-向量形式

以大小为 4 的状态空间为例：

$$\underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi} = \underbrace{\begin{bmatrix} r_\pi(s_1) \\ r_\pi(s_2) \\ r_\pi(s_3) \\ r_\pi(s_4) \end{bmatrix}}_{r_\pi} + \gamma \underbrace{\begin{bmatrix} p_\pi(s_1|s_1) & p_\pi(s_2|s_1) & p_\pi(s_3|s_1) & p_\pi(s_4|s_1) \\ p_\pi(s_1|s_2) & p_\pi(s_2|s_2) & p_\pi(s_3|s_2) & p_\pi(s_4|s_2) \\ p_\pi(s_1|s_3) & p_\pi(s_2|s_3) & p_\pi(s_3|s_3) & p_\pi(s_4|s_3) \\ p_\pi(s_1|s_4) & p_\pi(s_2|s_4) & p_\pi(s_3|s_4) & p_\pi(s_4|s_4) \end{bmatrix}}_{P_\pi} \underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi}.$$