Lecture Notes 2

2025 年 8 月 14 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：

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

1.4 通过贝尔曼公式求解状态值

闭式解：$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$

* 需要求矩阵的逆，在维数较大的时候不易求解

迭代解法：

$$v_{k+1} = r_\pi + \gamma P_\pi v_k$$

$$v_k \to v_\pi = (I - \gamma P_\pi)^{-1} r_\pi, \quad k \to \infty$$

* 证明如下：

将误差定义为 $\delta_k = v_k - v_\pi$。我们只需证明 $\delta_k \to 0$。将 $v_{k+1} = \delta_{k+1} + v_\pi$ 和 $v_k = \delta_k + v_\pi$ 代入 $v_{k+1} = r_\pi + \gamma P_\pi v_k$ 可得：

$$\delta_{k+1} + v_\pi = r_\pi + \gamma P_\pi(\delta_k + v_\pi),$$

上式可以改写为：

$$\delta_{k+1} = -v_\pi + r_\pi + \gamma P_\pi \delta_k + \gamma P_\pi v_\pi = \gamma P_\pi \delta_k.$$

（注：这里的化简利用了贝尔曼方程 $v_\pi = r_\pi + \gamma P_\pi v_\pi$，所以 $-v_\pi + r_\pi + \gamma P_\pi v_\pi = 0$）

因此，通过迭代展开，我们得到：

$$\delta_{k+1} = \gamma P_\pi \delta_k = \gamma^2 P_\pi^2 \delta_{k-1} = \cdots = \gamma^{k+1} P_\pi^{k+1} \delta_0.$$

注意到 $0 \le P_\pi^k \le 1$，这意味着对于任意 $k = 0, 1, 2, \ldots$，P_π^k 矩阵中的每一个元素都不大于 1。这是因为 $P_\pi^k \mathbf{1} = \mathbf{1}$，其中 $\mathbf{1} = [1, \ldots, 1]^T$ 是一个全为 1 的向量。另一方面，由于折扣因子 $\gamma < 1$，我们知道当 $k \to \infty$ 时，$\gamma^k \to 0$。因此，$\delta_{k+1} = \gamma^{k+1} P_\pi^{k+1} \delta_0 \to 0$。

这就证明了 v_k 会收敛到 v_π。

1.5　行动值（action value）

定义：$q_\pi(s,a)=\mathbb{E}\left[G_t \mid S_t=s,\, A_t=a\right]$

与状态值的联系：$v_\pi(s)=\sum_a \pi(a \mid s)\, q_\pi(s,a)$ 根据状态值的计算推导出行动值：$q_\pi(s,a)=\sum_r p(r|s,a)r+\gamma\sum_{s'} p(s'|s,a)v_\pi(s')$