Reinforcement Learning: An Introduction notebook

黎雷蕾

2018 年 1 月 3 日

目录

Chapter 12

Eligibility Traces

12.1 The λ-return

首先给出 n-step 回报公式:

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{v}(S_{t+n}, \mathbf{w}_{t+n-1}), \quad (12.1)$$

在这里,平均回报可以由一半两步回报和一半四步回报构成,即 $G = \frac{1}{2}G_{t:t+2} + \frac{1}{2}G_{t:t+4}$。这种更新方式称为复合更新 (compound update), 由此引出的算法称为 TD(λ) 算法, 这个平均包含 n 步回报, 权重比例为 $\lambda^{n-1}, \lambda \in [0,1]$, 加上系数 $(1-\lambda)$ 保证和为 1(极限求和)。其公式可以写作:

$$G_t^\lambda \doteq (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n} \quad (12.2)$$

对其进行一定的分步计算:

$$G_t^\lambda \doteq (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t \quad (12.3)$$

由此我们可以引出离线 λ 回报算法 (off-line λ-return algorithm), 在该算法进行中, 它不会对权重向量进行改变, 根据半梯度 (semi-gradient) 思想, 有:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \left[G_t^{\lambda s} - \hat{v}(S_t, \mathbf{w}_t) \right] \nabla \hat{v}(S_t, \mathbf{w}_t), \quad (12.4)$$