Reinforcement Learning: An Introduction notebook

黎雷蕾

2017 年 12 月 11 日

目录

9 On-policy Prediction with Approximation 2

9.1 Value-function Approximation 2

9.2 The Prediction Objective (MSVE) 2

9.3 Stochastic-gradient and Semi-gradient Methods 3

Chapter 9

On-policy Prediction with Approximation

9.1 Value-function Approximation

明白:$s \mapsto g$,

- s:the state backed up;
- g:the backed-up value;

在不同的算法中:

- the Monte Carlo backup: $S_t \mapsto G_t$;
- the TD(0) backup: $S_t \mapsto R_{t+1} + \gamma\hat{v}(S_{t+1}, w_t)$;
- the n-step TD backup: $S_t \mapsto G_{t:t+n}$;
- the DP policy-evaluation backup: $s \mapsto \mathcal{E}_\pi[R_{t+1} + \gamma\hat{v}(S_{t+1}, w_t)|S_t = s]$;

9.2 The Prediction Objective (MSVE)

假设 $\mu(s) \geq 0$ 表示对于状态 s 的错误的重视程度，$\hat{v}(s, w)$ 表示近似值函数，$v_\pi(s)$ 表示真实值函数，那么均方值误差 (Mean Squared Value Error,

MSVE) 就可以写作：

$$MSVE(w) = \sum_{s \in \mathbb{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, w) \right]^2 \tag{9.1}$$

用上式的平方根来描述近似值与真实值之间的误差，这种方式叫做同策略分布 (on-policy distribution)。

9.3 Stochastic-gradient and Semi-gradient Methods

定义权重向量: $\mathbf{w} \doteq (w_1, w_2, \cdots, w_d)^T$; 用 $\mathbf{w}_t$ 代表再第 t 步的权重向量。近似值逼近的目的是通过有限的实例逼近所有的状态。

一个好的策略是尽量减少观察到的实例中的错误，随机梯度下降 (Stochastic gradient descent, (SGD)) 通过小幅度地调整权重向量，使其向最小化误差方向前进。

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{1}{2}\alpha \nabla \left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t) \right]^2 \\ &= \mathbf{w}_t + \alpha \left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t) \right] \nabla \hat{v}(S_t, \mathbf{w}_t) \end{aligned} \tag{9.2}$$

其中权重向量的偏导向量可以表示为：

$$\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \frac{\partial f(\mathbf{w})}{\partial w_2}, \cdots, \frac{\partial f(\mathbf{w})}{\partial w_d} \right)^T \tag{9.3}$$

假设第 t 次抽样的样本是 U_t，我们可以用 U_t 代替 $\hat{v}(S_t, \mathbf{w}_t)$，并用下面的公式更新权重向量:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha \left[U_t - \hat{v}(S_t, \mathbf{w}_t) \right] \nabla \hat{v}(S_t, \mathbf{w}_t) \tag{9.4}$$

该算法可以写为：

Gradient Monte Carlo Algorithm for Estimating $\hat{v} \approx v_\pi$

1. 输入：需要评估的策略 π;
2. 输入：一个误差函数 $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$;
3. 以适当的方式初始化价值权重向量 $\mathbf{W}$（如：$\mathbf{w} = 0$）;
4. 一直重复如下步骤
   (a) 采用策略 π 生成一条轨迹 $\langle S_0, A_0, R_1, S_1, A_1, \cdots, R_T, S_T$;
   (b) 对于轨迹中的每一步 $t \in [0, T-1]$:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha[G_t - \hat{v}(S_t, \mathbf{w})]\nabla\hat{v}(S_t, \mathbf{w})$$