Reinforcement Learning Notes

To Be Strong, To Be Gentle

Jiayu Song

2024 年 5 月 3 日

Motivation

Note my understanding and problems

2024 年 5 月 3 日

目录

-1.1 动态规划算法

-1.1.1 概要

基于动态规划的强化学习算法主要有两种：

策略迭代 policy iteration

策略评估 (policy evaluation)，使用贝尔曼期望方程得到一个策略的状态价值函数

策略提升 (policy improvement)，直接使用贝尔曼最优方程来进行动态规划，得到最终的最优状态价值

价值迭代 value iteration

-1.1.2 策略迭代算法

Policy Evaluation

贝尔曼期望方程

$$V^{\pi}(s)=\sum_{a\in A}\pi(a|s)(r(s,a)+\gamma\sum_{s'\in S}p(s'|s,a)V^{\pi}(s'))$$

动态规划 + 贝尔曼期望方程

$$V^{k+1}(s)=\sum_{a\in A}\pi(a|s)(r(s,a)+\gamma\sum_{s'\in S}p(s'|s,a)V^{k}(s'))$$

Policy Improvement

Q: 如果我们有 $Q^{\pi}(s,a)>V^{\pi}(s)$，则说明在状态 s 下采取动作 a 会比原来的策略 $\pi(a|s)$ 得到更高的期望回报

假设存在一个确定性策略 π' 在任意一个状态 s 下，都满足

$$Q^{\pi}(s,\pi'(s))\geq V^{\pi}(s)$$

于是在任意状态 s 下，都有

$$V^{\pi'}(s)\geq V^{\pi}(s)$$

$$\pi'(s)=\arg\max_{a}Q^{\pi}(s,a)=\arg\max_{a}\{r(s,a)+\gamma\sum_{s'}P(s'|s,a)V^{\pi}(s')\}$$

-1.1.3 价值迭代算法

$$V^*(s) = \max_{a \in A} r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s')$$

$$V^{k+1}(s) = \max_{a \in A} r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^k(s')$$

-1.2 时序差分算法

-1.2.1 概要

无模型的强化学习 model-free Reinforcement Learning

Sarsa 和 Q-Learning

在线策略学习

离线策略学习：更好地利用历史数据，并具有更小的样本复杂度

-1.2.2 时序差分方法

蒙特卡洛方法对价值函数的增量更新方式：

$$V(s_t) \leftarrow V(s_t) + \alpha[G_t - V(s_t)]$$

时序差分算法用当前获得的奖励加上下一个状态的价值估计来作为在当前状态会获得的回报

$$V(s_t) \leftarrow V(s_t) + \alpha[r_t + \gamma V(s_{t+1}) - V(s_t)]$$

其中 $R_t + \gamma V(s_{t+1}) - V(s_t)$ 被称为时序差分误差 (Temporal Difference Error)

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi[G_t|S_t = s] \\ &= \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k}|S_t = s] \\ &= \mathbb{E}_\pi[R_t + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s] \\ &= \mathbb{E}_\pi[R_t + \gamma V_\pi(S_{t+1})|S_t = s] \end{aligned}$$

蒙特卡洛方法将上式第一行作为更新的目标，时序差分算法将上式最后一行作为更新的目标

-1.2.3 Sarsa 算法

-1.2.4 Q-Learning 算法

-1.3 Paper Reading

-1.3.1 SCI 写作积累

引出自己的概念

“However, we believe that it is a distinct concept that is worth clarifying.”

“To disentangle the parameter count from the filter size,”

表示递进

“More narrowly,”

“Empirically”

同义词替换

解决“solve —> decouple = disentangle”

大量的“a lot of —> prohibitive amounts of; a great number of”

-1.3.2 Mamba: Linear-Time Sequence Modeling with Selective State Spaces

原文状态方程和离散化状态方程：

$$\begin{cases} h^{'}(t) = Ah(t) + Bx(t) \\ y(t) = Ch(t) \end{cases} \tag{1}$$

$$\begin{cases} h_t = \overline{A}h_{t-1} + \overline{B}x_t \\ y_t = Ch_t \end{cases} \tag{2}$$

$$\begin{cases} \overline{K} = (C\overline{B}, C\overline{AB}, ..., C\overline{A}^k\overline{B}) \\ y = x * \overline{K} \end{cases} \tag{3}$$

SSM 离散化过程推导

公式1到公式2是怎样实现的?

构造状态方程

构造新的函数 $\alpha(t)h(t)$，并对新函数求导

$$\frac{d[\alpha(t)h(t)]}{dt} = \frac{d\alpha(t)}{t}h(t) + \alpha(t)h^{'}(t) \tag{4}$$

将公式1代入公式4得到：

$$\begin{aligned}\frac{d[\alpha(t)h(t)]}{dt} &= \frac{d\alpha(t)}{dt}h(t) + \alpha(t)[Ah(t) + Bx(t)] \\ &= [A\alpha(t) + \frac{d\alpha(t)}{dt}]h(t) + B\alpha(t)x(t)\end{aligned} \tag{5}$$

为消除 $h(t)$，则有：

$$\begin{aligned}A\alpha(t) + \frac{d\alpha(t)}{dt} &= 0 \\ \frac{d\alpha(t)}{dt} &= -A\alpha(t) \\ \alpha(t) &= e^{-At}\underbrace{+C}_{overlook}\end{aligned} \tag{6}$$

将公式6代入公式5得到：

$$\frac{d[e^{-At}h(t)]}{dt} = Be^{-At}x(t) \tag{7}$$

两边求积分得到：

$$\begin{aligned}e^{-At}h(t) &= h(0) + \int_0^t Be^{-A\tau}x(\tau)\,d\tau \\ h(t) &= h(0)e^{At} + \int_0^t e^{A(t-\tau)}Bx(\tau)\,d\tau\end{aligned} \tag{8}$$

离散化后对 $\{t_k, t_{k+1}\}$，时间间隔为 $T = t_{k+1} - t_k$，做差分：

$$h(t_{k+1}) = h(t_k)e^{AT} + \int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}Bx(\tau)\,d\tau \tag{9}$$

令 $\eta = t_{k+1} - \tau$ 得到（"-" 改变上下界，和换元有两次变号）：

$$\begin{aligned}h(t_{k+1}) &= h(t_k)e^{AT} + \underbrace{\int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}}_{-}\underbrace{Bx(\tau)\,d\tau}_{-} \\ &= h(t_k)e^{AT} + \int_0^T e^{A\eta}Bx(\eta)\,d\eta\end{aligned} \tag{10}$$

对 $x(t)$ 应用零阶保持器，在 $\{t_k, t_{k+1}\}$ 区间数值不发生变化，可作为常数提出，于是则有：

$$\begin{aligned}h(t_{k+1}) &= h(t_k)e^{AT} + \int_0^T e^{A\eta}\,d\eta Bx(t_k) \\ &= h(t_k)e^{AT} + [A^{-1}e^{AT} - A^{-1}]Bx(t_k) \\ &= e^{AT}h(t_k) + A^{-1}(e^{AT} - I)Bx(t_k)\end{aligned} \tag{11}$$

时间差 $T = \Delta$(常数)，上式为：

$$h(t_{k+1}) = e^{A\Delta}h(t_k) + A^{-1}(e^{A\Delta} - I)Bx(t_k) \tag{12}$$

所以 Mamba 原文公式2中 $\overline{A}$ 和 $\overline{B}$ 对应的值为：

$$\begin{cases} \overline{A} & = e^{\Delta A} \\ \overline{B} & = A^{-1}(e^{\Delta A} - I)B \end{cases} \tag{13}$$

这一结果与原文中的结果相对应，但是让我想不明白的是：为什么要在 A^{-1} 中加入 Δ^{-1} 并在后面乘 Δ，虽然这样变化整式的结果是没有变化的。