机器学习复习14

2022年7月9日 星期六 00:24

Reinforcement Learning and Control

1. 马尔科夫决策过程

由五元构成 $(S, A, \{P_{sa}\}, \gamma, R)$

S：状态集

A：一组动作

P_{sa}：转移概率

$\gamma \in [0,1)$：折扣因子

R：$S \times A$：回报函数

eg: $s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \cdots$

回报函数：$R(s_0,a_0)+\gamma R(s_1,a_1)+\gamma^2 R(s_2,a_2)$

使 $E[R(s_1)+\gamma R(s_1)+\gamma^2 R(s_2)+\cdots]$

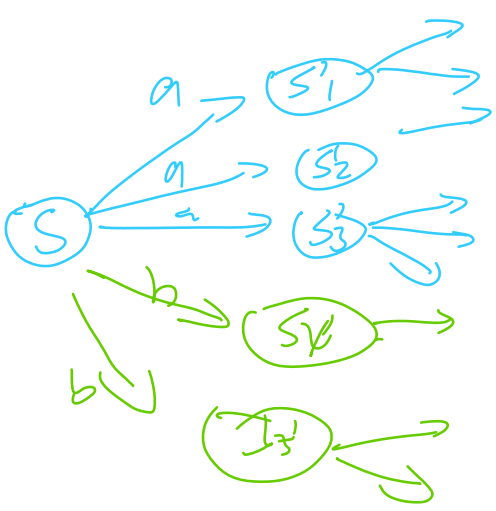
$$V^{\pi}(s)=R(s_0)+\gamma[E[R(s_1)+\gamma R(s_2)+\gamma^2 R(s_3)+\cdots]]$$

$$=R(s_0)+\gamma V^{\pi}(s')$$

$$V^{\pi}(s)=R(s)+\gamma\sum_{s'\in S}P_{s\pi(s)}(s')V^{\pi}(s')$$

（s' 表示下一个状态）

$$V^{*}(s)=\max V^{\pi}(s)$$



$$\pi^{*}(s)=\arg\max_{a\in A}\sum P_{sa}(s')V^{*}(s')$$

2、值迭代和策略迭代

① 值迭代：

1. $V(s)$ 初始化为0。
2. 循环至收敛{

$$V_s:=R(s)+\max_{a\in A}\gamma\sum_{s'}P_{sa}(s')V(s')$$

}

② 策略迭代：

1. 随机指定一个S到A的映射 π
2. 循环至收敛{

(a) 令 $V:=V^{\pi}$

(b) $\pi(s):=\arg\max_{a\in A}\sum P_{sa}(s')V(s')$

再由 Bellman $(V^{\pi}(s)=R(s)+\gamma\sum_{s'\in S}P_{s\pi}(s')V^{\pi}(s'))$

求出所有状态 S 的 $V^{\pi}(s)$