深度之眼
deepshare.net

强化学习专题:TD方法

导师：Alex

深度之眼
deepshare.net

目录

深度之眼
deepshare.net

TD方法

TD method

TD方法

Temporal Difference Method

在Monte Carlo的课程中，我们已经学习了incremental method

$$\hat{v} \longrightarrow \hat{v} + \alpha(v_{new} - \hat{v}) = \alpha v_{new} + (1 - \alpha)\hat{v}$$

在Monte Carlo Method中

- 用了什么值替换v_{new}?
- 如何估计那个值?

TD方法

Temporal Difference Method

在TD方法中，我们仍使用G_t作为v_{new}的替代，但使用另一种估计方法

$$G_t = R_{t+1} + \gamma G_{t+1} = R_{t+1} + \gamma v(s_{t+1})$$

从而写成

$$v(s_t) \longrightarrow v(S_t) + \alpha(R_{t+1} + \gamma v(S_{t+1}) - v(S_t))$$

TD方法

Temporal Difference Method

定义：TD error $\delta_t = R_{t+1} + \gamma v(S_{t+1}) - v(S_t)$，可以得到

$$
\begin{aligned}
G_t - v(S_t) &= R_{t+1} + \gamma G_{t+1} - v(S_t) + \gamma v(S_{t+1}) - \gamma v(S_{t+1}) \\
&= \delta_t + \gamma(G_{t+1} - v(S_{t+1})) \\
&= \delta_t + \delta_{t+1} + \gamma^2(G_{t+2} - v(S_{t+2})) \\
&= \delta_t + \gamma\delta_{t+1} + \cdots + \gamma^{T-t-1}\delta_{T-1} + \gamma^{T-t}(G_T - v(S_T)) \\
&= \sum_{k=t}^{T-1} \gamma^{k-t}\delta_k
\end{aligned}
$$

TD方法

Temporal Difference Method

很自然可以想到，我们可以使用一步之后的$v(S_{t+1})$，自然也可以使用n步之后的$v(S_{t+n})$

$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n v(S_{t+n})$$

$$v(S_t) \longrightarrow v(S_t) + \alpha(G_{t:t+n} - v(s_t))$$

这种方法叫做n-step TD

应该注意到，n越大，TD就越靠近Monte Carlo

深度之眼
deepshare.net

对比TD和MC

Comparison between TD and MC

对比TD和MC

Comparison between TD and MC

对比TD和Monte Carlo

联系：n-step TD可以接近甚至成为Monte Carlo

区别？

提示：n变大变小对于G_t的估计的影响

对比TD和MC

Comparison between TD and MC

答案：Bias-Variance trade off

使用实际reward R_t，R_{t+1}…的部分，没有bias，但是有较大的variance

使用已有的估计值$v(S_{t+n})$的部分，有bias，但variance较小甚至没有

何为更优？

大多数情况，TD收敛速度更快

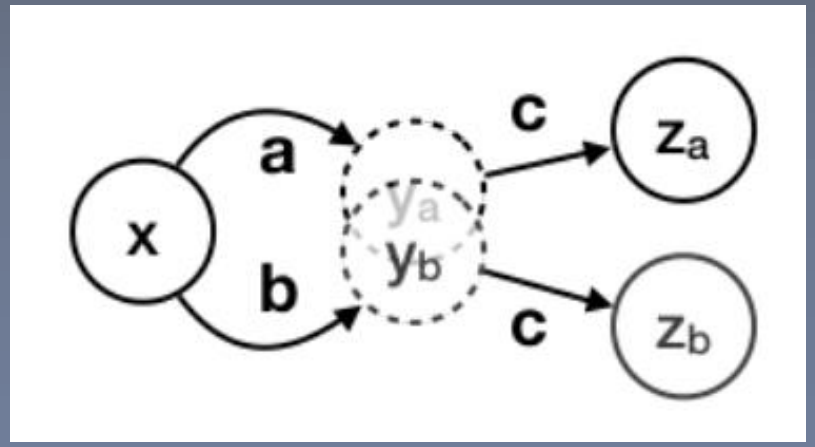
对比TD和MC

Comparison between TD and MC

但TD也有致命的缺陷，缺陷的来源就是$v(S_{t+n})$引入的bias

想象这个environment:

- agent无法区分y_a和y_b
- 到达z_a的reward远高于z_b
- b的reward稍稍高于a



TD方法无法学习到最优策略，但是Monte Carlo可以

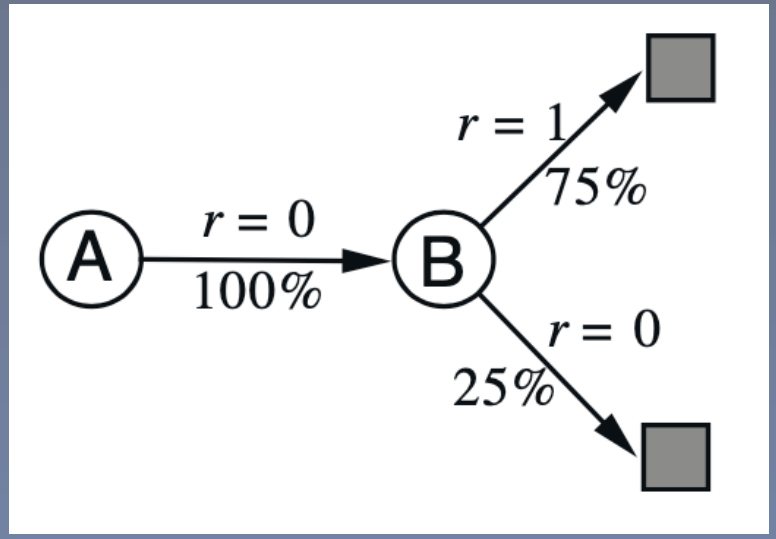
对比TD和MC

Comparison between TD and MC

这里我们也给出一个TD更“成功”的例子：

假设我们有8个episode，没有action，state和reward情况如下：

A,0,B,0	B,1
B,1	B,1
B,1	B,1
B,1	B,0



你会如何估计$v(A)$和$v(B)$?

$v(A) = v(B) = 3/4$

对比TD和MC

Comparison between TD and MC

- MC：所见即所得
- TD：带有一些“自然的”假设
- MC不需要MDP的假设，TD则受到MDP假设的影响

深度之眼
deepshare.net

Sarsa

Sarsa

Sarsa

Sarsa

我们已经讨论了对$v(S_t)$如何使用TD方法进行更新，如何拓展到$Q(S_t, A_t)$?

直接用$Q(S_t, A_t)$替换 $v(S_t)$

$$Q(S_t, A_t) \longrightarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

类似于TD，Sarsa也可以对R_{t+1}做n-step的拓展，替换为$G_{t:t+n}$

Sarsa

Sarsa

提问：Sarsa是online-learning还是offline-learning？

online-learning

提问：Sarsa的variance的来源有哪些？有没有可能降低？

$$Q(S_t, A_t) \longrightarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t))$$

这是Expected Sarsa

深度之眼
deepshare.net

Q-Learning

Q-Learning

Q-Learning

Q-Learning

Sarsa并不能直接收敛到最优策略的行动价值函数

但Sarsa的offline版本，Q-Learning可以

$$Q(S_t, A_t) \longrightarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t))$$

Q-Learning

Q-Learning

Q-Learning的一个严重缺陷是会“高估”一些值，原因是reward的随机性，我们通过一个例子去理解

解决方案：Double Q-Learning

$$Q_1(S_t, A_t) \longrightarrow Q_1(S_t, A_t) + \alpha(R_{t+1} + \gamma Q_2(S_{t+1}, \mathrm{argmax}_a Q_1(S_{t+1}, a)) - Q_1(S_t, A_t))$$
$$Q_2(S_t, A_t) \longrightarrow Q_2(S_t, A_t) + \alpha(R_{t+1} + \gamma Q_1(S_{t+1}, \mathrm{argmax}_a Q_2(S_{t+1}, a)) - Q_2(S_t, A_t))$$

结 语

本节课我们结束了强化学习基础部分的学习，并了解了部分代码，希望大家通过作业进一步巩固对于这些知识的掌握

从下一节课开始，我们将进入深度强化学习（DRL）的讲解，会带领大家一起阅读近几年的paper，进行进一步的学习

deepshare.net
深度之眼

联系我们:

电话: 18001992849

邮箱: service@deepshare.net

QQ: 2677693114

公众号

客服微信