Chapter 7 TD learning of state values.

一、TD learning

1. The data/experience required:

$(s_0, r_1, s_1, \ldots s_t\ r_{t+1}\ s_{t+1}, \ldots)$ or $\{(s_t, r_{t+1}, s_{t+1})\}_t$ generated following the given policy π.

2.
$$\begin{cases} V_t(s) = V_t(s_t) - \alpha_t(s_t)\left[V_t(s_t) - [r_{t+1} + \gamma V_t(s_{t+1})\right] \\ V_{t+1}(s) = V_t(s) \end{cases}$$

TD error δ_t

$$V_{t+1}(s_t) = V_t(s_t) - \alpha_t(s_t)\left[V_t(s_t) - [r_{t+1} + \gamma V_t(s_{t+1})]\right]$$

New estimate — old estimate — TD target $\bar{v}_t$

$$\delta_t = V_t(s_t) - [r_{t+1} + \gamma V_t(s_{t+1})]$$

$$\bar{v}_t = r_{t+1} - \gamma V_t(s_{t+1})$$

3. TD target:

$$V_{t+1}(s_t) = V_t(s_t) - \alpha_t(s_t)[V_t(s_t) - \bar{v}_t]$$

$$V_{t+1}(s_t) - \bar{v}_t = V_t(s_t) - \bar{v}_t - \alpha_t(s_t)[V_t(s_t) - \bar{v}_t]$$

$$V_{t+1}(s_t) - \bar{v}_t = [1 - \alpha_t(s_t)][V_t(s_t) - \bar{v}_t]$$

$$0 < 1 - \alpha_t(s_t) < 1$$

$$\therefore |V_{t+1}(s_t) - \bar{v}_t| \le |V_t(s_t) - \bar{v}_t|$$

4. TD error:

$$\delta_t = V(S_t) - [r_{t+1} + \gamma V(S_{t+1})]$$

$V_t = V_\pi \qquad \delta_t = 0$

$$\delta_{\pi,t} = V_\pi(S_t) - [r_{t+1} + \gamma V_\pi(S_{t+1})]$$

$$E[\delta_{\pi,t} | S_t = s_t] = 0$$

TD error: innovation 新信息.

5. other properties:

TD only estimates the value of a given policy

二. Cnvence and comparation.

TD algorithm solves the Bellman equation of a given policy π. without model.

$$V_\pi(s) = E[R + \gamma G | S = s]$$

$$E[G|S=s] = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) V_\pi(s') = E[V_\pi(S')|S=s]$$

$$V_\pi(s) = E[R + \gamma V_\pi(S') | S = s], \ s \in S \ ①$$

Using RM algorithm solve equation ①

$$g(V(s)) = V(s) - E[R + \gamma V_\pi(S') | s] = 0$$

$$g(V(s)) = 0$$

obtain samples: $r \to R$ and $s \to S'$

$$\tilde{g}(V(s)) = V(s) - [r + \gamma V_\pi(s')]$$

$$= (\underbrace{V(s) - E}_{g(V(s))}) + (\underbrace{E - [r + \gamma V_\pi(s')]}_{\eta})$$

RM algorithm for solving $g(V(s)) \equiv 0$

$$V_{k+1}(s) = V_k(s) - \alpha_k \tilde{g}(V_k(s))$$

$$V_{k+1}(s) = V_k(s) - \alpha_k (V(s) - [r_k + \gamma V_\pi(s')])$$

$$\{(s, r, s')\} \longrightarrow \{(s_t, r_{t+1}, s_{t+1})\}$$

方向问到什么状态更新什么状态，其余保持不动.

$$V_\pi(s') \Rightarrow V_\pi(s'_k)$$

二、TD learning of action value: Sarsa

1. a given π: $\{(S_t, a_t, r_{t+1}, S_{t+1}, a_{t+1})\}$

$$\begin{cases} q_{t+1}(s,a) = q_t(s,a), & \forall (s,a) \neq (S_t, a_t) \\ \text{where}, \quad t = 0, 1, 2 \cdots \end{cases}$$

✗ V→q

$q_t(S_t, a_t)$ is a estimate of $q_\pi(S_t, a_t)$

2. Sarsa = State – action – reward – state – action
S_t a_t r_{t+1} S_{t+1} a_{t+1}

3. TD solve

Sarsa solve: $q_\pi(s,a) = E[R + \gamma q_\pi(S', A') | s, a]$ $\forall s, a$

$V_\pi(s) = E[R + \gamma G | S = s]$

4. combine Sarsa with policy Improvement.

① Collect experience $(S_t, a_t, r_{t+1}, S_{t+1}, a_{t+1})$ following $\pi_t(S_t)$

② Update q-value.

$$q_{t+1}(s_t,a_t)=q_t(s_t,a_t)-\alpha_t(s_t,a_t)[q_t(s_t,a_t)-[r_{t+1}+\gamma q_t(s_{t+1},a_{t+1})]]$$

③ Update policy:

$$\pi_{t+1}(a|s_t)=1-\frac{\varepsilon}{|A|}(|A|-1)$$

if $a=\arg\max_a q_{t+1}(s_t,a)$

$$\pi_{t+1}(a|s_t)=\frac{\varepsilon}{|A|}$$ otherwise.

ε-greedy 策略.

三. Expected Sarsa / n-step Sarsa.

① Expected Sarsa:

$$q_{t+1}(s_t,a_t)=q_t(s_t,a_t)-\alpha_t(s_t,a_t)\left[q_t(s_t,a_t)-\left[r_{t+1}+\gamma q_t(s_{t+1},a_{t+1})\right]\right]$$

$$q_{t+1}(s_t,a_t)=q_t(s_t,a_t)-\alpha_t(s_t,a_t)\left[q_t(s_t,a_t)-\left[r_{t+1}+\gamma \mathbb{E}[q_t(s_{t+1},A)]\right]\right]$$

$$\mathbb{E}[q_t(s_{t+1},A)]=\sum_a \pi_t(a|s_{t+1})\,q_t(s_{t+1},a)=V_t(s_{t+1})$$

② n-step Sarsa:

$$q_\pi(s,a)=\mathbb{E}[G_t|S_t=s,A_t=a]$$

Sarsa $\leftarrow$ $G_t^{(1)}=R_{t+1}+\gamma q_\pi(s_{t+1},A_{t+1})$

$$G_t^{(2)}=R_{t+1}+\gamma R_{t+2}+\gamma^2 q_\pi(s_{t+1},A_{t+1})$$

$$\vdots$$

n Step Sarsa $\leftarrow$ $G_t^{(n)}=R_{t+1}+\cdots+\gamma^n q_\pi(s_{t+n},A_{t+n})$

$$\vdots$$

MC $\leftarrow$ $G_t^{(\infty)}=R_{t+1}+\gamma R_{t+2}+\cdots$

四 Q-Learning:

estimate. optimal action value

直接估计最优 action value.

1、

$$\begin{cases} q_{t+1}(S_t, a_t) = q_t(S_t, a_t) - \alpha(S_t, a_t)\left[q_t(S_t, a_t) - \left[r_{t+1} + \gamma \max_{a \in A} q_t(S_{t+1}, a)\right]\right] \\ q_{t+1}(s, a) = q_t(s, a) \end{cases}$$

TD target in Q-learning: $r_{t+1} + \gamma \max_{a \in A} q_t(S_{t+1}, a)$

TD target in Sarsa: $r_{t+1} + \gamma q_t(S_{t+1}, a_{t+1})$

2. Q-learning solve Bellman optimality equation

$$q(s,a) = E\left[R_{t+1} + \gamma \max_a q(S_{t+1}, a) \mid S_t = s, A_t = a\right]$$

3. off-policy & on-policy (对所有强化学习)

- behavior policy = generate experience
- target policy = constantly update towards optimal policy

on policy: behavior policy = target policy

off policy: behavior policy ≠ target policy

Sarsa — on policy. Q-learning — off policy

4. How to judge:

① Sarsa: on policy:

Solve: $q_\pi(s,a) = E[R + \gamma q_\pi(S', A')|s,a]$

where: $R \sim p(R|s,a)$ $S' \sim p(S'|s,a)$, $A' \sim \pi(A'|S')$

algorithm:

$$q_{t+1}(S_t,a_t) = q_t(S_t,a_t) - \alpha(S_t,a_t)\left[q_t(S_t,a_t) - [r_{t+1} + \gamma q_t(S_{t+1},a_{t+1})]\right]$$

requires: S_t, a_t, r_{t+1}, S_{t+1}, a_{t+1}

S_t, a_t: given; r_{t+1}: $p(r|s,a)$; S_{t+1}: $p(s'|s,a)$; a_{t+1}: $\pi_t(S_{t+1})$

π_t both target policy and behavior policy.

$\pi_t \rightarrow$ exp $\rightarrow q_{\pi(t)}$

② Monte Carlo learning: on policy.

$$q_\pi(s,a) = E[R_{t+1} + \gamma R_{t+2} + \cdots \mid S_t = s, A_t = a]$$

$$q(s,a) \approx r_{t+1} + \gamma r_{t+2} + \cdots$$

$\pi \rightarrow$ trajectory $\rightarrow q(\pi) \rightarrow \pi$.

③ Q-learning: off policy:

$$q(s,a) = E[R_{t+1} + \gamma \max_a q(S_{t+1}, a) \mid S_t = s, A_t = a], \quad \forall s, a$$

bellman optimal function.

$$q_{t+1}(s_t, a_t) = q_t(s_t, a_t) - \alpha_t(s_t, a_t)\left[q_t(s_t, a_t) - [r_{t+1} + \gamma \max_{a \in A} q_t(s_{t+1}, a)]\right]$$

s_t, a_t, s_{t+1}, a_{t+1}

given $p(r|s,a)$ $p(s'|s,a)$

Q-learning = behavior policy $s_t, a_t \rightarrow s_{t+1}, a_{t+1}$

target policy π optimal

5、Pseudocode: Q-learning

given: π_b —> episode: $\{S_0, a_0, r_1, S_1, a_1, r_2 \cdots\}$

For each step $t = 0, 1, 2, \cdots$ of the episode

Update q-value:

$$q_{t+1}(S_t, a_t) = q_t(S_t, a_t) - \alpha_{t+1}(S_t, a_t)\left[q_t(S_t, a_t) - [r_{t+1} + \gamma \max_{a \in A} q_t(S_{t+1}, a)]\right]$$

Update target policy:

greedy select $\pi_{T,t+1}(a|S_t) = 1$ 于 $a = \arg\max_a q_{t+1}$

不需要具有一定探索性.

① behavior policy selected. (探索性比较强)

② Update.

2.

unfied expression:

$$q_{t+1}(s_t,a_t) = q_t(s_t,a_t) - \alpha_t(s_t,a_t)[q_t(s_t,a_t) - \bar{q}_t]$$

All the algorithms we introduced in this lecture can be expressed in a unified expression:

$$q_{t+1}(s_t,a_t) = q_t(s_t,a_t) - \alpha_t(s_t,a_t)[q_t(s_t,a_t) - \bar{q}_t],$$

where $\bar{q}_t$ is the *TD target*.

Different TD algorithms have different $\bar{q}_t$.

Algorithm	**Expression of $\bar{q}_t$**
Sarsa	$\bar{q}_t = r_{t+1} + \gamma q_t(s_{t+1}, a_{t+1})$
n-step Sarsa	$\bar{q}_t = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^n q_t(s_{t+n}, a_{t+n})$
Expected Sarsa	$\bar{q}_t = r_{t+1} + \gamma \sum_a \pi_t(a\|s_{t+1}) q_t(s_{t+1}, a)$
Q-learning	$\bar{q}_t = r_{t+1} + \gamma \max_a q_t(s_{t+1}, a)$
Monte Carlo	$\bar{q}_t = r_{t+1} + \gamma r_{t+2} + \ldots$

The MC method can also be expressed in this unified expression by setting $\alpha_t(s_t, a_t) = 1$ and hence $q_{t+1}(s_t, a_t) = \bar{q}_t$.

所以这个是可以看成是n-step Sarsa的一个特殊情况

All the algorithms can be viewed as stochastic approximation algorithms solving the Bellman equation or Bellman optimality equation:

Algorithm	**Equation aimed to solve**
Sarsa	BE: $q_\pi(s,a) = \mathbb{E}\left[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \middle| S_t = s, A_t = a\right]$
n-step Sarsa	BE: $q_\pi(s,a) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^n q_\pi(s_{t+n}, a_{t+n}) \middle| S_t = s, A_t = a]$
Expected Sarsa	BE: $q_\pi(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \mathbb{E}_{A_{t+1}}[q_\pi(S_{t+1}, A_{t+1})] \middle| S_t = s, A_t = a\right]$
Q-learning	BOE: $q(s,a) = \mathbb{E}\left[R_{t+1} + \max_a q(S_{t+1}, a) \middle| S_t = s, A_t = a\right]$
Monte Carlo	BE: $q_\pi(s,a) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots \middle| S_t = s, A_t = a]$