第五课 monte carlo

model-free algorithm

1.Monte Carlo Basic

思路很简单，先随机给定一个初始的策略，然后根据这个策略，对每个(s,a)以它为出发点生成许多个episode，计算出discounted return，取平均作为action value。然后直接用这个action value更新策略。

The MC Basic algorithm

▷ Description of the algorithm:

Given an initial policy π_0, there are two steps at the kth iteration.

- **Step 1: policy evaluation.** This step is to obtain $q_{\pi_k}(s,a)$ for all (s,a). Specifically, for each action-state pair (s,a), run an infinite number of (or sufficiently many) episodes. The average of their returns is used to approximate $q_{\pi_k}(s,a)$.
- **Step 2: policy improvement.** This step is to solve $\pi_{k+1}(s) = \arg\max_\pi \sum_a \pi(a|s) q_{\pi_k}(s,a)$ for all $s \in \mathcal{S}$. The greedy optimal policy is $\pi_{k+1}(a_k^*|s) = 1$ where $a_k^* = \arg\max_a q_{\pi_k}(s,a)$.

Exactly the same as the policy iteration algorithm, except

- Estimate $q_{\pi_k}(s,a)$ directly, instead of solving $v_{\pi_k}(s)$.

2.MC Exploring Starts

主要有两点改进:

1. 在Monte Carlo Basic中，要收集完所有episode再更新action value。因此，我们收集完一个episode就直接更新action value。
2. 在Monte Carlo Basic中，收集一条episode却只更新episode开头的那一个(s,a)过于浪费数据。因此，我们将episode中经过的每个(s,a)，都把它们的discounted return作为样本估计它的action value

▷ If we use data and update estimate more efficiently, we get a new algorithm called MC Exploring Starts:

Pseudocode: MC Exploring Starts (a sample-efficient variant of MC Basic)

Initialization: Initial guess π_0.
Aim: Search for an optimal policy.

For each episode, do

Episode generation: Randomly select a starting state-action pair (s_0, a_0) and ensure that all pairs can be possibly selected. Following the current policy, generate an episode of length T: $s_0, a_0, r_1, \ldots, s_{T-1}, a_{T-1}, r_T$.

Policy evaluation and policy improvement:

Initialization: $g \leftarrow 0$

For each step of the episode, $t = T-1, T-2, \ldots, 0$, do

$g \leftarrow \gamma g + r_{t+1}$

Use the first-visit strategy:

If (s_t, a_t) does not appear in $(s_0, a_0, s_1, a_1, \ldots, s_{t-1}, a_{t-1})$, then

$Returns(s_t, a_t) \leftarrow Returns(s_t, a_t) + g$

$q(s_t, a_t) = \text{average}(Returns(s_t, a_t))$

$\pi(a|s_t) = 1$ if $a = \arg\max_a q(s_t, a)$

$s_1 a_1 \rightarrow s_2 a_2 \rightarrow s_1 a_3 \rightarrow s_4 a_4 \cdots$

$r+\gamma g$ $r+\gamma g$ g g

为了更新所有(s,a)，我们需要访问到所有(s,a)。访问有两种情况，start和visit。目前为止visit都只能根据策略访问，无法强制visit到所有(s,a)。因此我们只能在每个(s,a)上start一次episode，才能确保访问到所有(s,a)，而这也就是Exploring Starts的含义。

3.MC epsilon-greedy(without Exploring Starts)

相比MC Exploring Starts，MC epsilon-greedy只是把deterministic策略换成了epsilon-greedy策略。这样就能确保足够长的episode能visit到所有(s,a)，也就不需要Exploring Starts这个限制条件了。

ε-greedy policies

▷ What soft policies will we use? Answer: ε-greedy policies

- **What is an ε-greedy policy?**

$$\pi(a|s) = \begin{cases} 1 - \dfrac{\varepsilon}{|\mathcal{A}(s)|}(|\mathcal{A}(s)| - 1), & \text{for the greedy action,} \\[2ex] \dfrac{\varepsilon}{|\mathcal{A}(s)|}, & \text{for the other } |\mathcal{A}(s)| - 1 \text{ actions.} \end{cases}$$

where $\varepsilon \in [0, 1]$ and $|\mathcal{A}(s)|$ is the number of actions for s.

  - The chance to choose the greedy action is always greater than other actions, because $1 - \frac{\varepsilon}{|\mathcal{A}(s)|}(|\mathcal{A}(s)| - 1) = 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(s)|} \geq \frac{\varepsilon}{|\mathcal{A}(s)|}$.

- **Why use ε-greedy?** Balance between exploitation and exploration
  - When $\varepsilon = 0$, it becomes greedy! Less exploration but more exploitation!
  - When $\varepsilon = 1$, it becomes a uniform distribution. More exploration but less exploitation.

这里和MC Exploring Starts稍有不同的是，使用的every visit method，因为每个episode会很长，需要充分利用

Pseudocode: MC ϵ-Greedy (a variant of MC Exploring Starts)

Initialization: Initial guess π_0 and the value of $\epsilon \in [0, 1]$
Aim: Search for an optimal policy.

For each episode, do
Episode generation: Randomly select a starting state-action pair (s_0, a_0). Following the current policy, generate an episode of length T: $s_0, a_0, r_1, \ldots, s_{T-1}, a_{T-1}, r_T$.
Policy evaluation and policy improvement:
Initialization: $g \leftarrow 0$
For each step of the episode, $t = T-1, T-2, \ldots, 0$, do
$g \leftarrow \gamma g + r_{t+1}$
Use the every-visit method:

$Returns(s_t, a_t) \leftarrow Returns(s_t, a_t) + g$
$q(s_t, a_t) = \text{average}(Returns(s_t, a_t))$
Let $a^* = \arg\max_a q(s_t, a)$ and

$$\pi(a|s_t) = \begin{cases} 1 - \frac{|\mathcal{A}(s_t)|-1}{|\mathcal{A}(s_t)|}\epsilon, & a = a^* \\ \frac{1}{|\mathcal{A}(s_t)|}\epsilon, & a \neq a^* \end{cases}$$

当step很长时，一个episode就能访问遍所有(s,a)

最初epsilon可以比较大增加随机性，之后要逐渐减小epsilon以得到最优策略