

On policy 和 Off policy 的区别

LingTao HUANG(黄玲涛), YaNan LI(李亚男)

2017年3月9日

1 Q Learning—off policy

Q learning 的更新状态方程:

$$Q(S, A) \leftarrow Q(S, A) + \alpha \left[R_{t+1} + \gamma \max_a Q(S', a) - Q(S, A) \right]$$

每个 time step 就会更新一次Q值, 使用下一个state中Q值最大的action, 即 $\max_a Q(S', a)$, 但是下一个state到底执不执行这个action, 不一定。也就是说用来更新Q表的policy并不一定用来执行。

2 Sarsa—on policy

Sarsa 的更新状态方程:

$Q(S, A) \leftarrow Q(S, A) + \alpha [R_{t+1} + \gamma Q(S', A') - Q(S, A)]$ 每个 time step 就会更新一次Q值, 但是在更新Q表之前, 会先确定下一个state的采取的行动, 并用这个action所对应的Q值更新Q表。也就是说, 用来更新Q表的policy也是用来执行的。