Q-learning

2019年7月26日 9:43

一、算法思想

初始化$Q(s,a),\forall s \in S, a \in A(s)$,任意的数值，并且$Q(terminal-state,\cdot)=0$

重复（对每一节episode）：

初始化 状态S

重复（对episode中的每一步）：

使用某一个policy比如（$\epsilon-greedy$)根据状态S选取一个动作执行

执行完动作后，观察reward和新的状态S'

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

$$S \leftarrow S'$$

循环直到S终止

特点：

Q-learning是基于离散结点的。属于Value-based算法，输出的是action的value。

可以把Q值当成矩阵

单步更新

二、问题

1、维度灾难：状态太多

2、用函数去拟合Q表Q(s,a)=f(s,a,w)

3、高维输入低维输出

因此选择采用神经网络的方式