强化学习中值迭代算法的实现

重点：两个算法都是交替进行value update和policy update，来求解最优策略

Value迭代算法

迭代链路：

值迭代算法伪代码：

已知：p(r|s,a) agent在状态s采取动作a的奖励r的概率，也就是及时奖励分布函数

P(s’|s,a) agent在状态s采取动作a之后转移到状态s’的概率，也就是状态转移分布函数

初始化的V

While :

For s in S:

#遍历每一个状态

For a in Action(s):

#利用state value求action value



#对于状态s，求解action value中的最大值



#策略更新policy update，将策略指向刚才求出的最大值

π(a|s) = 1 if a=a\* else 0

#值更新 value update，利用刚才的策略，重新计算新的state value

V= max q(s,a)

里面每个变量在每轮迭代值都不一样，实际编程需要加上k，实际编程时求的是每个元素的差值的最大值

代码 见