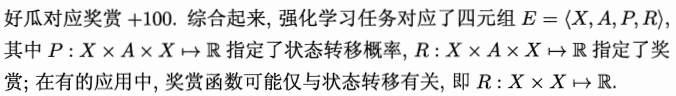
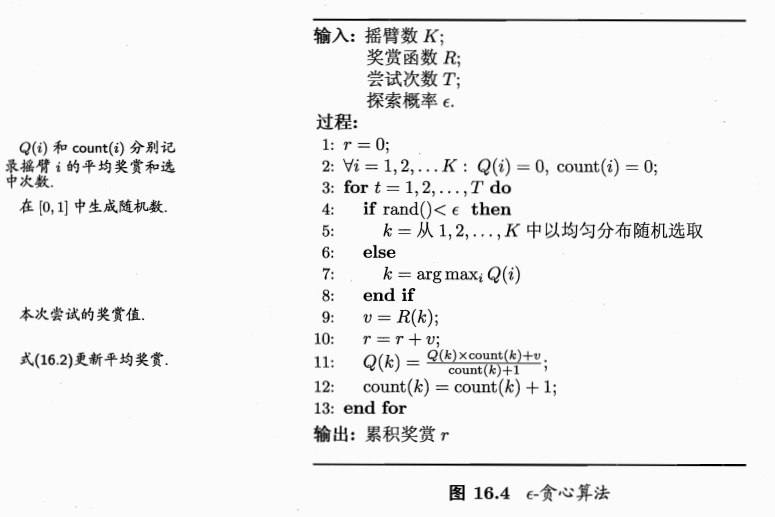
强化学习任务通常用马尔科夫决策过程来描述（Markov Decision Process,MDP）来描述：机器处于环境E中，状态空间为X,其中每个状态x属于X是机器感知到的环境的描述；机器能采取的动作构成了动作空间A.若某个动作a属于A作用在当前状态x上，则潜在的转移函数P将使得环境从当前状态按某种概率转移到另一个状态，在转移到另一个状态的同时，环境会根据潜在的奖赏函数R反馈给机器一个奖赏，如保持瓜苗健康对应奖赏+1，瓜苗凋零对应奖赏-10。



在强化学习任务中，学习的目的就是要找到能使长期累积奖赏最大化的策略。长期累积奖赏有多种计算方式，常用的有“T步累积奖赏” 和r折扣累积奖赏，其中rt表示第t步获得的奖赏值，E表示对所有随机变量求期望。

16.2 k –摇臂赌博机



Reference:

1.机器学习\_周志华