20337263 俞泽斌 理论作业部分

1、错误，MC方法直接从经历过的完整的状态序列中学习，不基于模型，某状态的价值等于多个状态序列中该状态算得到的所有return的平均。

2、正确，TD将当前获得的奖励加上下一个状态的价值估计来当作当前状态获得的回报

3、Sarsa算法首先定义一个五元组 $(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1})$

对于每个S，使用基于Q的策略选择A.

然后执行A，在此基础上，使于基于Q的策略选择A'

这就涉及到了在上一个猜测的基础上接着进行猜测

即自举的过程.

然后通过 $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S',A') - Q(S,A)]$

来更新值函数.

再进行下一个状态和动作的更新 $S \leftarrow S'$ $A \leftarrow A'$ 直到S是终止态.

不断重复猜测.