增强学习测试题参考答案

## 简答题（共10题，每题7分）

1. 一阶马尔可夫属性：P(St+1|St) = P(St+1|St, St-1, St-2,…,S0)

马尔可夫过程：多个符合马尔可夫属性的状态的序列

马尔可夫回报过程：在马尔可夫过程基础上，增加了回报R和衰减因子gamma。每个状态的变换伴随着一个回报，长程的回报受到衰减因子的惩罚。

马尔可夫决策过程：在马尔可夫回报过程基础上，增加了动作A。动作导致状态变化，收到回报。好的决策得到的回报更多。

1. 监督学习、无监督学习、增强学习
2. 增强学习想求解的是在连续的决策问题中，如何决策能都得到最大的收获。这些收获具有延迟性。
3. 贝尔曼方程使得value function可以写成迭代的形式，另一个好处是可以重用已经算好的value function，增加算法的效率。
4. Agent: 老鼠，reward: 电击、水滴、奶酪，state: 老鼠的位置
5. 对于任意policy，迭代地调用贝尔曼期望方程更新计算该policy的value function，直到收敛，收敛后的value function即是对该policy的评价。此时greedy式的选择能得到更大回报的决策，就得到一个新的policy。之后再迭代地调用贝尔曼期望方程。。。直到更新的policy与原来的policy的评价相同，此时的policy即为最优policy。
6. 一个改动是要用q value function替换v value function。另外要用ϵ-greedy的方式更新policy，而不是greedy式的。
7. MC算法使用的是整个episode作为学习样本，TD算法使用episode中的每一步作为学习样本；MC算法具有高variance, 无bias；TD算法具有低variance，有bias。
8. TD(λ)算法考虑每一个步长的TD算法，它以每个步长的TD算法的线性加权和作为返回值。λ决定了每个步长的权重，越往后的步长权重越小。
9. On-policy指的是指导采样的policy和算法优化的目标policy是同一个Policy。Off-policy中则是不同的policy。

## 应用题（共3题，每题10分）

1. 0.6\*（2+0.9\*10）+0.4\*（-2+0.9\*20）=13

则当时，可保证向减小的方向变化。

1. 可使用下式映射到[0,10]的值域空间