第三课 无模型价值函数估计和控制 上

1. MDP已知：即为R和P已知
2. $model-free$：针对MDP未知的情况，通过智能体与环境互动得到一组轨迹数据（$trajectory/episode$）$\{S_1A_1R_1S_2A_2R_2\cdots S_TA_TR_T\}$
3. 蒙特卡洛(MC)策略评估
    - 用经验平均（取几个轨迹然后取平均）而非期望，不要动态规划，不用$bootstrapping$，不用假设状态序列的马尔可夫性，适用于有终止的MDP
    - 具体算法：每个轨迹，每次状态s被访问，访问次数$N(s)\leftarrow N(s)+1$，总回报（$return$）$S(s)\leftarrow S(s)+G_t$，最后$v(s)\leftarrow\frac{S(s)}{N(s)}$。由大数定律，$v(s)\rightarrow v^{\pi}(s)(N(s)\rightarrow\infty)$
    - 算法写成累（$incremental$）加形式
        - 收集一个$episode(S_1A_1R_1\cdots S_t)$
        - 对每个状态S_t，计算回报G_t，$N(S)\leftarrow N(s)+1$，$v(s_t)\leftarrow v(s_t)+\frac{1}{N(s_t)}(G_t-v(s_t))$（或者写成滑动平均（$running\ mean$）形式$v(s_t)\leftarrow v(s_t)+\alpha(G_t-v(s_t))$）
    - MC相对于DP（$dynamic\ programming$动态规划）的优势
        - 可用于未知环境
        - 适用于状态转移概率计算复杂的情况
        - 可以从对解决问题有利的感兴趣的轨迹出发
4. $temporal\ difference\ learning(TD\ learning)$
    - 给定π通过经验在线学习v^{π}
    - 最简单的$TD\ learning$：$TD(0)$：$v(s_t)\leftarrow v(s_t)+\alpha(R_{t+1}+\gamma v(s_{t+1})-v(s_t))$
        - $R_{t+1}+\gamma v(s_{t+1})$：$TD\ target$
        - $R_{t+1}+\gamma v(s_{t+1})-v(s_t)$：$TD\ error$
        - 类似于$incremental+bootstrapping$，整体结构是$incremental$，但其中的$return\ G_t$使用的是$bootstrapping$中的写法$R_{t+1}+\gamma v(s_{t+1})$，类似于向前一步的$return$
    - $MC\ vs\ TD$
        - MC
            - 能在每一步在线学习
            - 能从未结束的序列中学习
            - 能从连续（无终止）的环境中学习
            - 用到了马尔可夫性，在马尔可夫的环境中更高效
        - TD
            - 要等一个轨迹结束，等直到了整体的$return$后才能开始学习
            - 只能从完结的序列中学习
            - 只能从$episodic\ environment$中学习
            - 不需要马尔可夫性，在非马尔可夫环境中更高效
    - $n-step\ TD$：$TD\ target$用$R_{t+1}+\gamma R_{t+2}+\cdots+\gamma^{n-1}R_{t+n}+\gamma^n v(s_{n+1})$