第二课 马尔可夫决策过程 下

1. 预测（$prediction$）和控制（$control$）
   - 预测：给定$MDP(S, A, P, R, \gamma)$和策略π，求解价值函数v^π
   - 控制：给定$MDP(S, A, P, R, \gamma)$，求最佳的价值函数v^*和最佳策略π
2. 动态规划的特点
   - 将问题分解为最佳子结构
     - 最优性原理适用
     - 最优解可被分解为子问题（最优性原理：多阶段决策过程的最优决策序列具有这样的性质：不论初始状态和初始决策如何，对前面决策所造成的某一状态而言，其后各阶段的决策序列必须构成最优决策）
   - 重叠子问题
     - 子问题多次重复出现
     - 解能被缓存和重新使用
3. MDP满足这两个性质：
   - $Bellman$等式给出了一个迭代的分解
   - 价值函数存储并能重复利用解
4. MDP的策略评估
   - 目标：评估MDP的一个给定策略
   - 输出：策略的价值函数v^π
   - 具体算法：$synchronous\ backup$

     $$V_{t+1}(s)=\sum_{a \in A} \pi(a \mid s)\left(R(s, a)+\gamma \sum_{s^{\prime} \in S} P\left(s^{\prime} \mid s, a\right) v_t\left(s^{\prime}\right)\right)$$

     或化为$MRP(S, P^\pi, R, \gamma)$

     $$v_{t+1}(s)=R^\pi(s)+\gamma \sum_{s^{\prime}\ inS} P^\pi(s^{\prime}|s) v_t(s^{\prime})$$
5. 最佳价值函数与最佳策略
   - 最佳价值函数为$v^*=\underset{\pi}{max}\ v^\pi(s)$
   - 最佳策略为$\pi^*(s)=\underset{\pi}{argmax}\ v^\pi(s)$
6. MDP控制
   - 算法1：$policy\ iteration + policy\ improvement$
     - $policy\ iteration$
       - 评估策略π
       - 改进策略：$\pi^{\prime}=greedy(v^\pi)$
       - 上述两步不断迭代最终$\pi^{\prime}$会收敛
     - $policy\ improvement$
       - 计算$q^{\pi_i}(s, a)=R(s, a)+\gamma \sum_{s^{\prime} \in S} P(s^{\prime}|s, a) v^{\pi_i}(s^{\prime})$
       - 计算$\pi_{i+1}(s)=\underset{a}{argmax}\ q^{\pi_i}(s, a)$
       - $Bellman$最优等式$v^\pi=\underset{a \in A}{max}\ q^\pi(s, a)$，满足后可有最优价值函数且

$$v^*(s) = \max_a R(s,a) + \gamma \sum_{s\prime \in S} P(s'|s,a) v^*(s')$$

$$q^*(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \max_{a'} q^*(s',a')$$

- 算法2：值迭代
  - 将$Bellman$最优等式作为更新规则 $v(s) \leftarrow \max_{a \in A} R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) v(s')$
  - 具体的：

$$\begin{aligned}
&k = 1, v_0(s) = 0 \\
&while\ k \leq H: \\
&\qquad for\ s: \\
&\qquad\qquad q_{k+1}(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) v_k(s') \\
&\qquad\qquad v_{k+1}(s) = \max_a q_{k+1}(s,a) \\
&\qquad k \leftarrow k+1
\end{aligned}$$

  - 得到v^*的同时可以得到$\pi(s) = \underset{a}{argmax}\ R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) v^*(s')$