第二课 马尔可夫决策过程 上

1. 马尔可夫过程
   - 状态的历史：$h_t = \{s_1, s_2, \ldots\ldots, s_t\}$,$s_t$是马尔可夫的当且仅当

     $$p(s_{t+1}|s_t) = p(s_{t+1}|h_t) \qquad p(s_{t+1}|s_t, a_t) = p(s_{t+1}|h_t, a_t)$$

   - 状态转移矩阵：

     $$P = \begin{bmatrix} p(s_1|s_1) & \cdots & p(s_1|s_N) \\ \vdots & & \vdots \\ p(s_N|s_1) & \cdots & p(s_N|s_N) \end{bmatrix}$$

2. 马尔可夫奖励过程（MRP）
   - MRP组成：一列状态S；状态转移矩阵P；奖励方程（$reward\ function$）R : $R(S_t = s) = E(r_t|s_t = s)$；折扣因子（$discount\ factor$）$\gamma$
   - $horizon$：每个$episode$中$agent$走过的步数
   - $return$：从时间t到$horizon$的奖励R的折扣加和

     $$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$$

   - MRP的价值函数

     $$V_t(s) = E\left[G_t \mid s_t = s\right] = E\left[R_{t+1} + \cdots + \gamma^{T-t-1} R_T | s_t = s\right]$$

     价值函数满足$bellman$等式

     $$V(s) = R(s) + \gamma \sum_{s' \in S} P\left(s' \mid s\right) V\left(s'\right)$$

     等式的矩阵形式

     $$\begin{bmatrix} V(s_1) \\ \vdots \\ V(s_N) \end{bmatrix} = \begin{bmatrix} R(s_1) \\ \vdots \\ R(s_N) \end{bmatrix} + \gamma \begin{bmatrix} p(s_1|s_1) & \cdots & p(s_N|s_1) \\ \vdots & & \vdots \\ p(s_1|s_N) & \cdots & p(s_N|s_N) \end{bmatrix} \begin{bmatrix} V(s_1) \\ \vdots \\ V(s_N) \end{bmatrix}$$

     即$V = R + \gamma PV$，可求得$V = (I - \gamma P)^{-1} R$

   - 用蒙特卡洛计算MRP的价值函数

     $$\begin{aligned} & i \leftarrow 0, G_t \leftarrow 0 \\ & while\ i \neq N\ do \\ & \qquad \text{生成一个}episode, \text{从状态}s, \text{时间}t\text{开始} \\ & \qquad \text{计算}return\ g = \sum_{i=t}^{H-1} \gamma^{i-t} r_i \\ & \qquad G_t \leftarrow G_t + g, i \leftarrow i + 1 \\ & end\ while \\ & V_t(s) \leftarrow \frac{G_t}{N} \end{aligned}$$

   - 用迭代法计算MRP的价值函数

$for\ all\ state\ s \in S$
$while\ \left\| V - V^{'} \right\| > \epsilon\ do$
$\quad V \leftarrow V^{'}$
$\quad for\ all\ state\ s \in S$
$$V^{'}(s) = R(s) + \gamma \sum_{s' \in S} P\left(s' \mid s\right) V\left(s'\right)$$
$end\ while$
$return\ V^{'}(s)\ for\ all\ s \in S$

3. 蒙特卡洛决策过程（MDP）
   - MDP组成：一列状态S；一列行为A（$action$）；状态转移矩阵P^a，矩阵中每个元素为$p(s_{t+1}|s_t = s, a_t = a)$；奖励方程（$reward\ function$）$R$：$R(S_t = s, a_t = a) = E(r_t|s_t = s, a_t = a)$；折扣因子（$discount\ factor$）$\gamma$
   - MDP的$policy$指明每个状态的$action$：$\pi(a|s) = p(a_t = a|s_t = s)$
   - MDP转MRP：给定$MDP(S, A, P, R, \gamma)$和$policy\ \pi$，序列$S_1 S_2 \cdots$是一个马尔可夫过程，序列$S_1 R_1 S_2 R_2 \cdots$是一个$MRP(S, P^{\pi}, R^{\pi}, \gamma)$，其中
$$p^{\pi}(s^{'}|s) = \sum_{a \in A} \pi(a|s) p(s^{'}|s, a)$$
$$R^{\pi}(s) = \sum_{a \in A} \pi(a|s) R(s, a)$$
   - MRP由$s^{'}$决定s，MDP由s决定a再决定s
   - MDP的价值函数：
     - $state\ value\ function$：$v^{\pi}(s) = E_{\pi}(G_t|S_t = s)$
     - $action\ value\ function$：$q^{\pi}(s, a) = E_{\pi}(G_t|S_t = s, A_t = a)$
     - 两者关系：$v^{\pi}(s) = \sum_{a \in A} \pi(a|s) q^{\pi}(s, a)$
     - 两个价值函数期望的下标意思是对π取样
   - $Bellman$等式：
$$v^{\pi}(s) = \sum_{a \in A} \pi(a \mid s) \left(R(s, a) + \gamma \sum_{s' \in S} P\left(s' \mid s, a\right) v^{\pi}\left(s'\right) \right)$$
$$q^{\pi}(s, a) = R(s, a) + \gamma \sum_{s' \in S} P\left(s' \mid s, a\right) \sum_{a' \in A} \pi\left(a' \mid s'\right) q^{\pi}(s'|a')$$