第一课 概括与基础 下

1. $RL\ agent$主要组成部分
   - $Policy$：$agent$的行为模型，有两类。一类是概率$policy$：$\pi(a|s)=P[A_t=a|S_t=s]$，另一类是已决定的$policy$：$a^*=\underset{a}{argmax}\ \pi(a|s)$
   - $Value\ function$：在特定$policy$下，未来的$reward$在一个$discount$下的加和的期望值，对应公式为

     $$v_\pi(s)=E_\pi[G_t|S_t=s]=E_\pi[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t=s]$$

     还有一种形式（$Q\ function$）

     $$q_\pi(s,a)=E_\pi[G_t|S_t=s,A_t=a]$$

     用于选$action$
   - $Model$：决定下一步的S和R，其中

     $$P^a_{s\,s'}=P[S_{t+1}=s^{'}|S_t=s,A_t=a]$$

     $$P^a_s=P[R_{t+1}|S_t=s,A_t=a]$$

     分别用于预测下一个S和下一个R

2. 根据$agent$学习的不同将$agent$分为三类
   - 基于$value\ function$的$agent$，学习$value\ function$，通过$value\ function$推$policy$
   - 基于$policy$的$agent$，学习$policy$，没有$value\ function$
   - $actor-criticagent$，同时学习$policy$和$value\ function$

3. 根据$agent$是否学习$model$分类
   - $model\ based$：学习$model$，不一定有$policy$或$value\ function$
   - $model\ free$：学习$policy$或$value\ function$，没有$model$

4. $exploration\ vs\ exploitation$
   - $exploration$：尝试新方法，这可能让$agent$有更好的决策
   - $exploitation$：用已知的可能有用的方法