It's a RL note created by YuXia.

第一章

Basic concept

name	description	Supplement
$state$(状态)	代理相对于环境的状态	无
$action$(动作)	代理在每个状态下的动作	每个状态下的所有动作构成动作空间
$state\_transition$ (状态转移)	代理在某个状态下采取某个动作	通常描述为状态->动作->新的状态。也可以用$p(s_{n+1} \mid s_n, a_n) = probability$ 来表示一个状态经过某个动作转移到另一个状态的条件概率
$policy$(策略)	策略表示的是一个从状态到动作的映射, 对于每个可能的状态,策略会选择一个相应的行动	策略同样可以用条件概率表示，表示为$\pi(a_n \mid s_n) = probability.$
$reward$(奖励)	采取某策略后的奖励称为$reward$	$reward$可以表示为$r_{state} = number.$ 由于代理在一个状态有多个策略,不同策略发生的概率不同。 获得的奖励也不同,所以可以用$p(r \mid s, a)$表示代理在状态s时采取行动a获得奖励的数量.
$Trajectory$(轨迹)	由一系列的$state - action - reward$ 构成了一条$Trajectory$	每个$state$的策略有多个,所以即使一个$state$也对应着多条$Trajectory.$
$return$(回报)	$return$指的是某个$Trajectory$获得的所有$reward$之和	显然如果$Trajectory$不同,获得的$return$也不同。$return$又区分为$immediate_reward$和$delay_reward$
$discounted\_return$ (折现率)	折现率的大小意味着代理对$immediate_reward$和$delay_reward$的关注程度	有时候你的选择不止影响到下一步的状态，也会影响到后续的状态。 例如你选择玩不仅影响你此时的心情，还会影响到期末考试成绩。 所以需要给奖励一定的折现率。取值越大， 说明你更加关注某件事的奖励对近期的影响。取值越小， 说明你更关注某件事的奖励对长期的影响。

Return与状态值

$return$的作用是衡量一个策略的良好程度，因此$return$十分重要。
$return$也可以定义为沿着轨迹的含贴现率的$reward$总和。

状态值

name	description	Supplement
$state\_value$	以某个状态为起点由它延申出的**所有状态序列**的加权平均定义为$state_value$	这里的状态序列也可以理解为**以某个状态为起点按照某种**

考虑一个时间序列$t = 0, 1, 2, ...$，假设在时间t，代理处于状态S_t且采取的行动是A_t,下一步的状态是S_{t+1}，获得的奖励记作R_{t+1}.那么后续构成的轨迹可以看作是:

$$S_t \xrightarrow{A_t} S_{t+1}, R_{t+1} \xrightarrow{A_t} S_{t+2}, R_{t+2} \xrightarrow{A_t} S_{t+3}, R_{t+3} ...$$

注意到这里的每一步的S_t、A_t、S_{t+1}、R_{t+1}都不是固定的，根据策略的不同它们会不同。如果把轨迹的含贴现率的总$return$记作G_t。那么就有

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...$$

其中$\gamma \in (0, 1)$.注意到G_t也是一个非固定值，根据策略的不同它会不同。

这里需要引入一个新的概念，以S_t为起点的一条轨迹，称作一条**状态序列(episode)**，显然对于任意的一个状态S_t，它可能拥有一条或多条状态序列。每条状态序列获得的总回报$(return)$就是我们的一个G_t.

那么我们可以把状态s的状态值定义为:

$$V_\pi(s) = \mathbb{E}[G_t \mid S_t = s]$$

这个公式的含义为,对所有以状态s为起点的所有可能的轨迹的总回报(也就是所有的G_t)求数学期望，即求它们的带权平均值。

> 再举一个例子说明什么是$V_\pi(s)$.
>
> 考虑一个2×2的网格世界.每个网格分别被记作状态$S_i (i \in [1, 4])$.
>
> 网格世界如下所示:

S_1	S_2
S_3	S_4

> 假设以S_1为起始状态，S_4为结束状态。把到达各个状态的奖励分别设置为R_1、R_2、R_3、R_4。
>
> 那么S_1的状态序列就有两条，分别是:
>
> $$S_1 \to S_2 \to S_4$$
>
> $$S_1 \to S_3 \to S_4$$
>
> 对于每条状态序列,获得的奖励分别是
>
> $$G_{t1} = R_2 + \gamma R_4$$
>
> $$G_{t2} = R_3 + \gamma R_4$$
>
> 那么S_1的状态值$V_\pi(s) = \mathbb{E}[G_t \mid S_t = s]$就是这二者的数学期望，也就是对它们乘以各自发生的概率求和。
> 总结以上，某个状态的状态值$V_\pi(s)$就是奖励过程中该状态所有$return$的期望。
>
> 需要注意的是，在实际应用中，我们通常**无法事先确定**每个状态序列出现的准确概率，因此需要使用一些估计方法来得到这些概率。例如，在蒙特卡罗方法中，我们可以通过采样多条状态序列，并统计它们的出现次数来估计每个状态序列的概率。

接着，我们将状态值函数$V_\pi(s) = \mathbb{E}[G_t \mid S_t = s]$进行展开，得到:

$$\begin{aligned}
v(s) &= \mathbb{E}[G_t \mid S_t = s] && (1)\\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... \mid S_t = s] && (2)\\
&= \mathbb{E}[R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3}) + ... \mid S_t = s] && (3)\\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] && (4)\\
&= \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s] && (5)
\end{aligned}$$

可以看出，某个状态的状态值，取决于当前状态进行某个行动获得的$immediate_reward$和下一步的状态值乘以折现率γ。

马尔科夫过程与贝尔曼方程

马尔科夫过程的定义

> 在一个时序过程中，如果$t+1$时刻的状态仅取决于t时刻状态S_t而与t时刻之前的任何状态都无关时，则认为t时刻的状态具有**马尔科夫性**$(Markov\ property)$。若过程中的每一个状态都具有马尔科夫性，则这个过程具备马尔科夫性。具备马尔科夫性的随机过程称为**马尔科夫过程**$(Markov\ process)$

采样与状态序列

从符合马尔科夫过程给定的状态转移概率矩阵生成一个状态序列的过程称为**采样(**$sample$**)**。采样得到的一系列的状态转换过程，称为**状态序列(**$episode$**)**。当状态序列的最后一个状态是终止状态时，称该状态序列是一个**完整的状态序列**。

状态转移概率矩阵

状态转移概率矩阵的形式为:

$$(P_{ss'}) = \mathbb{P}[S_{t+1} = s^{'} \mid S_t = s]$$

它定义了从任意一个状态s到其所有后继状态$s^{'}$的状态转移概率。

$$\begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \vdots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \tag{3}$$

其中矩阵的每一行数据表示从某一个状态到所有n个状态的转移概率值，每一行的这些值的和应该为1。说明针对某一状态所有可能发生的事情的总概率应该为1。
例如，$P_{13} = 0.5$说明了由状态1转移到状态3的概率为0.5。

马尔科夫奖励过程

马尔科夫奖励过程$(Markov\ reward\ process)$就是在马尔科夫过程的基础上，考虑了奖励这个元素。它是由$< S, P, R, \gamma >$构成的一个元组。其中:

- S是一个**有限个状态**构成的集合
- P是集合中的状态转移概率矩阵:$(P_{ss'}) = \mathbb{P}[S_{t+1} = s^{'} \mid S_t = s]$
- R是一个奖励函数:$R_s = \mathbb{E}[R_{t+1} \mid S_t = s]$
- γ 是折现率，也称为衰减因子。 $\gamma \in (0, 1)$

贝尔曼方程

根据刚才展开的状态值函数的结果，我们可以得到:

$$v(s) = R_s + \gamma \sum_{s' \in S} P_{ss'} v(s^{'})$$

这就是马尔科夫奖励过程的贝尔曼方程$(Bellman\ equation)$，它的含义是一个状态的状态值(或者说价值)由**该状态采取的第一步行动获得的奖励**和**后续状态的状态值按照概率分布求和并乘以一定的衰减比率**联合组成。
将该方程写成矩阵形式，则有:

$$v = R + \gamma P v \tag{6}$$
$$(E - \gamma P) v = R \tag{7}$$
$$v = (E - \gamma P)^{-1} R \tag{8}$$

显然如果状态转移概率矩阵P和奖励函数R确定的话，该方程即可求解。

马尔科夫决策过程

马尔科夫奖励过程并不涉及到个体行为的选择，因此有必要引入马尔科夫决策过程$(MarkovDecision\ Process)$。它是在马尔科夫奖励过程上新增了一个行为集合$A$。
具体的来说，它是由$< S, A, P, R, \gamma >$构成的一个元组。其中:

- S是一个**有限个状态**构成的集合

- A是一个**有限个行为**构成的集合
- P是集合中**基于行为**的状态转移概率矩阵:$(P_{ss'}) = \mathbb{P}[S_{t+1} = s^{'} \mid S_t = s，\ A_t = a]$
- R是一个**基于状态和行为**的奖励函数:$R_s = \mathbb{E}[R_{t+1} \mid S_t = s，\ A_t = a]$
- γ 是折现率，也称为衰减因子。 $\gamma \in (0, 1)$

什么是策略?

在马尔科夫决策过程中，个体有着**根据自身对当前状态的认识从行为集合中选择一个行为的权利**。个体在给定状态下从集合中选择一个**行为的依据**称为**策略**$(policy)$，以字母 π 来表示。
π 是**某一状态**下基于行为集合的一个**概率分布**:

$$\pi(a \mid s) = \mathbb{P}[A_t = a \mid S_t = s]$$

具体的来说，如果一个状态S_t具有两个行为a_1, a_2。它们各自发生的概率都为0.5，那么状态S_t的策略$\pi(a \mid S_t)$可以表示为:

$$\pi(a_1 \mid S_t) = 0.5$$
$$\pi(a_2 \mid S_t) = 0.5$$

在马尔科夫决策过程中，策略仅通过当前状态就可以产生一个个体的行为。可以认为，**策略仅与当前状态相关，而与历史状态无关**。对于**不同的状态**，个体依据**同一个策略**也可能产生**不同的行为**。对于同一个状态，个体依据**相同的策略**也可能产生**不同的行为**。策略描述的是个体行为产生的机制（多大概率会发生这个行为？），是不随着状态变化而变化的，被认为是静态的。

> $tips$: 随机策略是一个很常用的策略，个体使用随机策略时，个体在某一个状态下选择的行为并不确定。由此个体可以借助随机策略在同一状态下尝试不同的行为。

策略与马尔科夫决策过程

当给定一个马尔科夫决策过程：$M =< S, A, P, R, \gamma >$和一个策略$\pi$，状态序列$S_1, S_2, \ldots$是一个符合马尔科夫过程$< S, P_\pi >$的采样（$< S, P_\pi >$的含义是：采取了策略$\pi$的状态转移矩阵$P$与状态的联合）。
类似的，联合状态和奖励的序列$S_1 \to R_2 \to S_2 \to R_3 \ldots$是一个符合马尔科夫奖励过程$< S, P_\pi, R_\pi, \gamma >$的采样，并且这个奖励过程满足以下两个方程:

$$P^{\pi}_{s,s'} = \sum_{a \in A} \pi(a \mid s) P^{a}_{s,s'}$$
$$R^{\pi}_{s} = \sum_{a \in A} \pi(a \mid s) R^{a}_{s}$$

分别说明这两个方程:

> 方程1:描述了在给定策略π下，状态从s转移到$s^{'}$的概率。
> 具体的来说，$P^{\pi}_{s,s'}$表示在策略π的控制下，从状态s转移到$s^{'}$的概率。$P^{a}_{s,s'}$则表示采取了行动a的情况下，从状态s转移到$s^{'}$的概率。根据全概率公式，$P^{\pi}_{s,s'}$可以表示为所有可能的行动a下的加权平均值。$\pi(a \mid s)$表示策略π在状态s时选择行动a的概率。A是所有可能行动的集合。这个公式可以理解为，将所有可能的行动下从状态s转移到状态$s^{'}$的概率加权平均，得到在策略π的控制下从状态s转移到状态$s^{'}$的概率。
> 换个人话来说，左边说的是我**采取某个策略π的情况下今天中午吃三碗的概率**。假设我们有"跑步"、"看书"、"睡觉"三个行动，$P^{a}_{s,s'}$说的就是我在做了这三件事的某一件之后，**今天中午吃三碗的概率**。由于这三件事**都影响到我中午吃三碗的可能性**，而且各自**影响的程度不一样**（跑步的影响更大，可以说跑步这个策略更好），所以需要对这三件事影响程度的大小对$P^{a}_{s,s'}$做一个加权再求和，得到的就是**我在采取了策略π的情况下，今天中午吃三碗的概率**
> 总之，$P^{a}_{s,s'}$表示一种固定的行动下出现特定状态的概率，而$P^{\pi}_{s,s'}$则表示在采取某个策略的情况下出现特定状态的概率。因此，$P^{\pi}_{s,s'}$需要考虑策略对各种行动的影响程度，并将它们进行加权平均。

> 方程2:描述了在给定策略π下，从状态s开始采取行动的期望回报$(return)$。
> 具体的来说，R^{π}_{s}表示在策略π的控制下,从状态s开始采取行动能获得的期望回报。R^{a}_{s}表示在状态s时采取行动a的情况下获得的回报。
> 还是用人话来说，策略描述了我做某件事有多少可能性，假设我做不同的事对我的身体健康有不同的好处。那么对我的身体好处的数学期望就是**我做某件事的概率**乘以**我做这件事得到的好处**，把所有的都加起来就是身体好处的数学期望了。

再谈状态值函数与行为价值函数

考虑如下问题:

> 假设你在家里度过一个周末，可以选择做许多不同的事情来娱乐自己。这个周末的每个时刻都是一个状态，而你采取的行动则是策略。
> 假设现在是周六早上，你正在考虑应该如何度过这个周末。在这个状态下，有多种可选的策略。例如：
>
> - 策略1：去公园运动
> - 策略2：在家看电影
> - 策略3：约朋友出去聚餐
>
> 这三种策略都是可能的，并且在某些方面都有各自的优点和缺点。如果你喜欢运动和户外活动，那么策略1可能更适合你；如果你想要放松身心，享受电影的视听体验，那么策略2可能更适合你；如果你想和朋友一起共度美好时光，那么策略3可能更适合你。
> 换句话说，在相同的状态下，我们可以有不同的策略来选择行动。这些策略可能基于不同的目标、偏好、资源或限制，它们可能会导致不同的结果和后果。因此，在MDP中，我们需要仔细考虑每个状态下可选的策略，并尝试选择最优的策略，以便最大化累积奖励。

上述问题说明了同一个马尔科夫决策过程，不同的策略也会产生不同的马尔科夫（奖励）过程，进而针对当前状态会有不同的状态值。

定义：价值函数$v_\pi(s)$是在马尔科夫决策过程下基于策略π的状态值函数，表示从状态s开始，遵循策略π获得的收获的期望:

$$v_\pi(s) = \mathbb{E}[G_t \mid S_t = s]$$

由于引入了行为的概念，同一个状态下采取不同的行为也会产生不同的价值。因此定义基于策略的π的行为价值函数$q_\pi(s,a)$：

$$q_\pi(s,a) = \mathbb{E}[G_t \mid S_t = s, A_t = a]$$

行为价值函数与某一状态相关，毕竟针对不同的状态可能会有不同的行为。
总结以上，我们得到了两个贝尔曼期望方程：

$$v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s] \tag{9}$$
$$q_\pi(s,a) = \mathbb{E}[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a] \tag{10}$$

简述这两个方程，第一个自然是我们状态价值函数的定义。

> 方程1：一个状态的状态值等于该状态的即时奖励加上后续状态的状态值乘以一定的衰减比例γ后求期望，其中G_t表示特定策略下以该状态为起点的状态序列，根据策略的不同会产生不同的状态序列。因此需要求期望。
>
> 方程2：一个状态的行为价值，可以由该状态采取某一行动获得的即时奖励加上后续获得的行为奖励乘以一定的衰减比例求期望得到。

第二章

策略评估与迭代

预测$(prediction)$:对于已知的马尔科夫决策过程或奖励过程(已知指的是状态转移概率矩阵、奖励、策略等已知)，求解基于该策略下的状态集合中的所有状态的状态价值。
控制$(control)$:对于已知的马尔科夫决策过程求解最优状态价值v_*和最优策略π_*

策略评估$(policy_evaluation)$指的是在给定策略下迭代状态价值函数的过程。
公式为：

$$v_{k+1}(s) = \sum_{a \in A} \pi(a|s)(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_k(s'))$$

这个公式的含义为，遵循策略π的情况下，第$k+1$步时状态s的状态价值等于该状态下所有可能的行为价值的期望。

什么是确定性策略?

> $tips$:
> 是否需要进行策略评估，关键在于状态价值函数计算时使用的策略是否是确定性的。

> 确定性的策略指的是从某一个状态开始，后续所有状态的状态转移概率都已知，并且对于每个状态，都只存在一个确定的动作可以被执行。因此，在确定性的策略下，给定当前状态就可以唯一地确定接下来要采取的动作，不存在随机性。

对于一个4*4的格子世界中的智能体，书中给出的例子遵循的是基于**均一概率的随机策略**($uniform\_random\_policy$)，在这个策略下智能体向四周移动的概率均等。这样经过多次迭代，对于靠近终止状态的格子，可以计算出它的价值更大，对于远离终止状态的格子它的价值更小。

策略迭代

完成了一个策略的评估以后，将得到基于该策略下的每一个状态的价值。那么如果我们根据这里的状态价值，从起始状态出发不断选择状态价值最大的状态，会得到一个新的策略。

例如，我们对书上的格子世界一开始采用的是均一随机策略。在这样的策略下经过有限次的迭代后得到了各个状态的价值。此时我们在任意一个状态下可以选择状态价值最大的状态进行转移，这样会得到一个新的策略。

根据这个策略，我们再更新我们的状态价值函数，重复上面这个过程，就形成了策略迭代过程。

换句话说，假设一开始我们使用的是均一随机策略，那么在任何非边界的状态上我们上下左右移动的概率均等，都是0.25。而我们在进行了策略评估得到了所有状态的状态价值之后，我们就可以改善这个策略。假设我们在一个非边界的状态时，上下左右的状态价值分别是1，-1，-2，0。那么我们就可以增大在该状态下时向上走的概率，也就是我们所说的改善策略。

考虑一个确定性的策略π对任意状态s产生的行为$a = \pi(s)$。如果我们采用贪婪策略，在同样的状态s下会得到一个新的行为$a^{'}$，满足:

$$a^{'} = \arg\max_{a \in A} q_{\pi}(s, a)$$

也就是说，新的行为$a^{'}$满足在状态s下，该行为的行为价值最大。