马尔可夫决策过程

在强化学习中，马尔可夫决策过程(Markov decision process, MDP)是对完全可观测的环境进行描述的，也就是说观测到的状态内容完整的决定了决策需要的特征，几乎所有的强化学习问题都可以转换为MDP。

**马尔可夫过程 Markov Process**

* **马尔可夫性 Markov Property**

某一状态信息包含了所有相关的历史，只要当前状态可知，那么所有的历史信息都不在需要了，当前状态就可以决定未来，则认为该状态具有马尔可夫性。

可以用下面的状态转移概率公式来描述马尔可夫性：

下面的状态转移矩阵定义了所有状态的转移概率：

式中n为状态数量，矩阵中每一行元素之和为1.

* **马尔可夫过程 Markov Process**

马尔可夫过程也叫做马尔可夫链(Markov Chain)，它是一个无记忆的随机过程，可以用一个元组<S,P>表示，其中S是有限数量的状态集，P是状态转移概率矩阵。

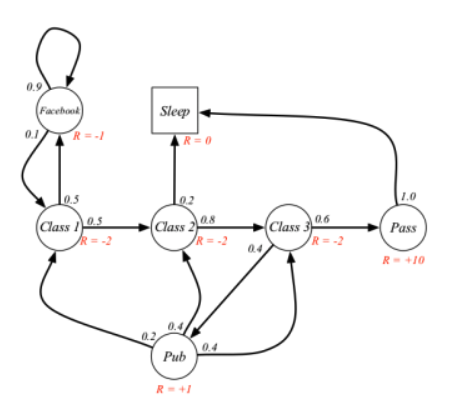
* **马尔可夫奖励过程 Markov Reward Process**

马尔可夫奖励过程在马尔可夫过程的基础上增加了奖励R和衰减系数：<S,P,R,>。R是一个奖励函数，**S状态下的奖励是某一时刻t处在状态s下，在下一时刻t+1时能获得的奖励期望：**

很多人纠结为什么奖励是t+1时刻的，**照此理解起来相当于离开这个状态才能获得奖励而不是进入这个状态即获得奖励。**强化学习的创始人David表示这仅仅是一个约定，为了在描述RL问题中涉及到的观测O、行为A、和奖励R时比较方便。他同时指出如果把奖励改成而不是，只要规定好，本质意义上是一样的。

**衰减系数Discount Factor：，**它的引入表明了远期利益具有一定的不确定性，符合人类对于眼前利益的追求。

下面是一个马尔可夫奖励的过程图示的例子，在马尔可夫过程的基础上增加了针对每一个状态的奖励，由于不涉及衰减系数相关的计算，这张图并没有特殊交代衰减系数值的大小。



* **收获Return/ Reward**

**收获：定义收获为在一个马尔可夫奖励链上从t时刻开始往后所有的奖励的衰减的总和。也有翻译为“收益”或“回报”的，公式如下：**

**其中衰减系数体现了未来的奖励在当前时刻价值比例，在k+1时刻获得的奖励R在t时刻体现出的价值是，接近0，则表明趋向于“近似性”评估，接近1则表明偏重考虑远期的利益。**

* **价值函数 Value Function**

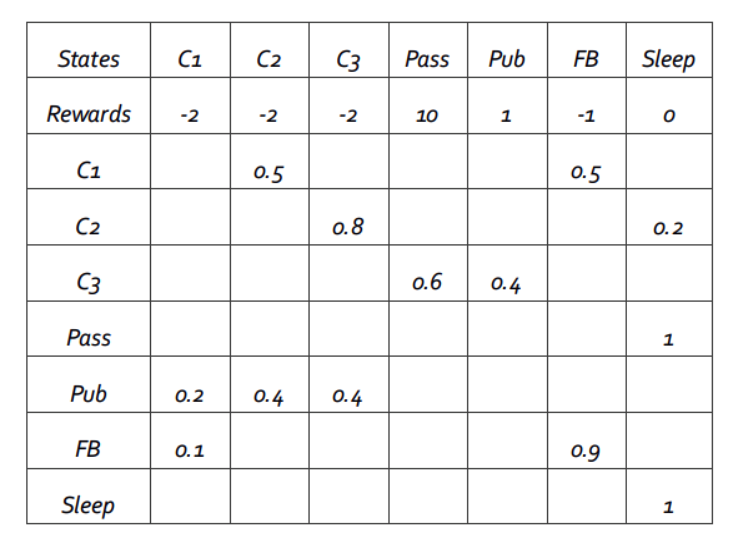
**价值函数给出了某一状态或某一行为的长期价值。**

**定义：一个马尔可夫奖励过程中某一状态的价值函数为从该状态开始的马尔可夫链收获的期望：**

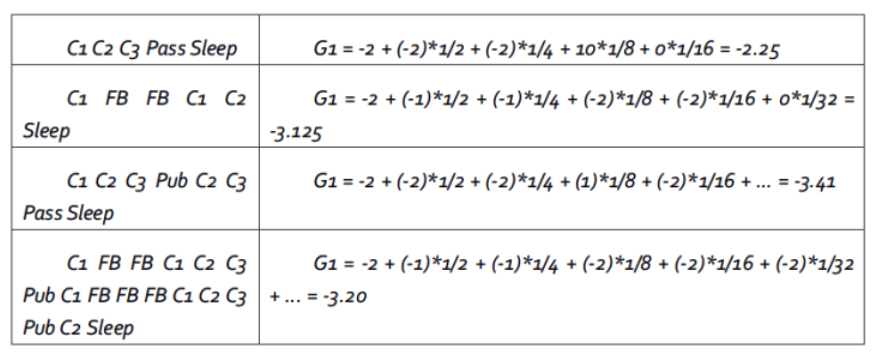
**注：价值可以仅描述状态，也可以描述某一状态下的某个行为，在一些特殊情况下还可以仅描述某个行为。另外还有更具体的一种说法是，用状态价值函数或价值函数来描述针对状态的价值；用行为价值函数来描述某一个状态下执行某一行为的价值。**

**举例说明收获和价值的计算**

**为方便计算，把“学生马尔可夫奖励过程”示例图表示成下表的形式。表中第二行对应各状态的即时奖励值，第三行一直到最后一行，表示状态转移概率，也就是从所在行状态转移到所在列状态的概率。**

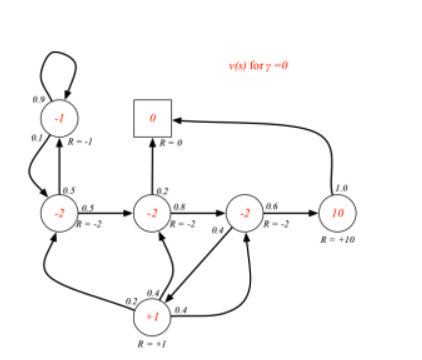
****

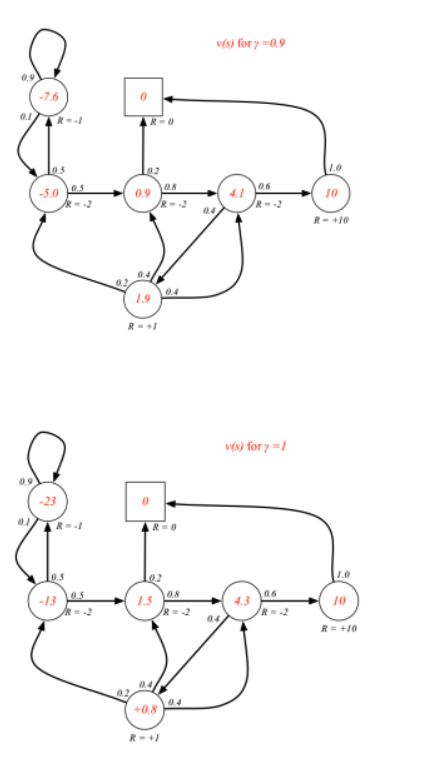
**加入有如下四个马尔可夫链，现在计算时，在t=1时刻()时状态的收获分别为【收获为从t时刻开始，往后的所有时刻的衰减的奖励的总和】：**

****

**从上表可以看到，收获是针对一个马尔可夫链中的某一个状态来说的。**

**当时，上表描述的MRP中，各状态的即时奖励就与该状态的价值相同。当时，各状态的价值需要通过计算得到，这里先给出分别为0，0.9和1三种情况下各状态的价值，如下图所示，其中，各状态圈内的数字表示该状态的价值，圈外的R=-2等表示的是该状态的即时奖励。**

****

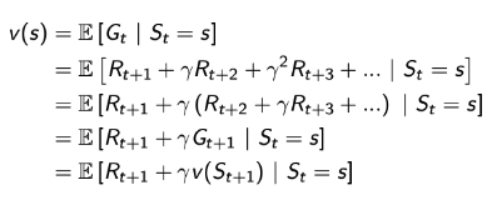
****

**各状态价值的确定是很重要的，RL的很多问题都可以归结为求状态的价值问题。因此如何求解各状态的价值，也就是寻找一个价值函数（从状态到价值的映射）就变得很重要了。**

**价值函数的推导**

* **Bellman方程(贝尔曼方程)—MRP**

**先尝试用价值的定义公式来推导看看能得到什么：**

****

**这个推导过程相对简单，仅在导出最后一行时，将变成了，理由则是收获的期望等于收获的期望的期望。下面是针对MRP的Bellman方程：**

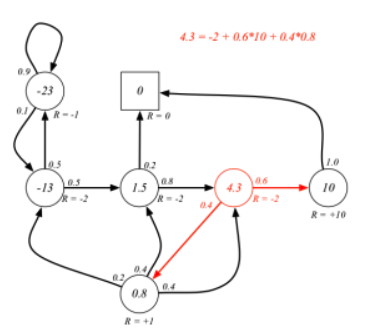
**通过方程可以看出v(s)由两部分组成，一是该状态的即时奖励期望，即时奖励期望等于即时奖励，因为根据即时奖励的定义，它与下一个状态无关；另一个是下一时刻状态的价值期望，可以根据下一时刻状态的概率分布得到其期望。如果用表示s状态下一时刻任一可能的状态，那么Bellman方程可以写成：**

**补充下，本文中对下标t和t+1的区分，其中和配对，也就是说时刻的回报reward用表示，不是，这是因为：**

* + **所有状态S都是从开始的，我们不会考虑s(-1)的情况。**
  + **s0的reward，是在agent离开s0之后得到的。**
  + **agent离开s0，就相当于t0时刻结束，而t1时刻开始了，因此，此时的R如果按照time-step来确定下标的话，应该是t1时刻获得的，记为R1。**
* **方程的解释**

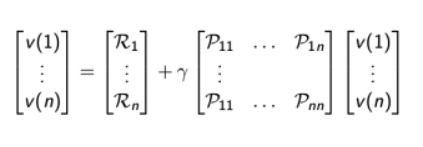
**下图已经给出了=1时各状态的价值，状态的价值可以根据状态Pub和Pass的价值以及他们之间的状态转移概率来计算：**

**4.3 = -2 + 1.0\*（0.6\*10 + 0.4\*0.8）**

****

* **Bellman方程的矩阵形式和求解**

**结合矩阵的具体表达形式可以按照如下来理解：**

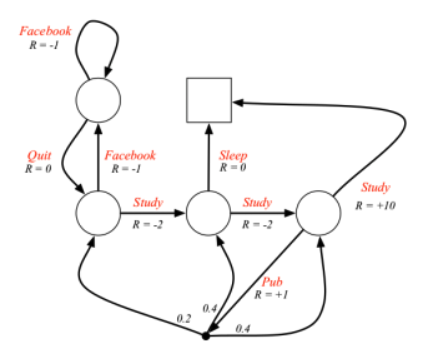
****

**马尔可夫决定过程 Markov Decision Process(MDP)**

**相较于马尔可夫奖励过程，马尔可夫决定过程多了一个行为集合A，它是这样的一个元组：<S,A,P,R,>。这里的P和R都与具体的行为a对应，而不像马尔可夫奖励过程那样仅对应于某个状态，A表示的是有限的行为的集合。具体的数学表达式如下：**

* **示例—学生MDP**

**下图给出了一个可能的MDP的状态转化图。图中红色的文字表示的是采取的行为，而不是先前的状态名，对比之前的学生MRP【马尔可夫奖励过程】示例可以发现，即时奖励与行为对应了，同一个状态下采取不同的行为得到的即时奖励是不一样的。由于引入了Action，容易与状态名混淆，因此此图没有给出各状态的名称；此图还把Pass和Sleep状态合并为一个终止状态；另外当选择Pub（去酒吧）这个动作时，主动进入了一个临时状态（图中小黑实点表示），随后被动的被环境按照其动力分配到另外三个状态，也就是说此时Agent没有选择决定权去哪一个状态。**

****

* **策略Policy**

**策略是概率的集合或者分布，其元素为对过程中的某一个状态s采取可能的行为a的概率，用表示：**

**一个策略完整的定义了个体的行为方式，也就是说定义了个体在各个状态下的各种可能的行为方式以及其概率的大小。Policy仅和当前的状态有关，与历史信息无关；同时某一确定的Policy是静态的，与时间无关；但是个体可以随着时间更新策略。**

**当给定一个MDP：M=<S,A,P,R,>和一个策略，那么状态序列是一个马尔可夫过程<S, >；同样，状态和奖励序列是一个马尔可夫奖励过程<S, , , >,并且在这个奖励过程中满足下面的两个方程：**

**用文字描述如下，在执行策略时，状态从s转移至s’时的概率等于一系列概率的和，这一系列概率指的是在执行当前策略时，执行某一个动作的概率与该行为能使状态从s转移到s’的概率的乘积。**

**奖励函数如下：**

**用文字表达如下：当前状态s下执行某一指定策略得到的即时奖励是该策略下所有可能行为得到的奖励与该行为发生的概率的乘积。**

**策略在MDP中的作用相当于agent可以在某一个状态时做出选择，进而有形成各种马尔可夫过程的可能，而且基于策略产生的每一个马尔可夫过程是一个马尔可夫奖励过程，各过程之间的差别是不同的选择产生了不同的后续状态以及对应的不同的奖励。**

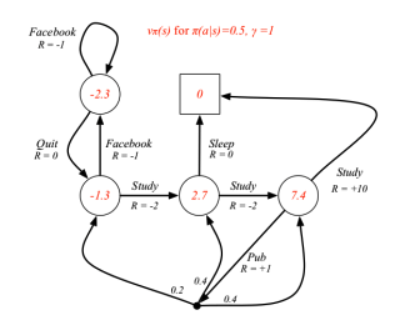
* **基于策略的价值函数**

**定义是在MDP下的基于策略的状态价值函数，表示从状态s开始，遵循当前策略时所获得的收获的期望；或者说在执行当前策略时，衡量个体处在状态s时的价值大小。数学表达如下：**

**注意策略是静态的，关于整体的概念，不随状态的改变而改变；变化的是在某一个状态时，依据策略可能产生的具体行为，因为具体的行为是有一定的概率的，策略就是用来描述各个不同状态下执行各个不同行为的概率。**

**定义为行为价值函数，表示在执行策略时，对当前状态s执行某一具体行为a所能得到的的收获reward的期望；或者说在遵循当前策略时，衡量对当前状态执行行为a的价值大小。行为价值函数一般都是与某一特定的状态对应的，更精细的描述是状态行为对价值函数。行为价值函数的公式描述如下：**

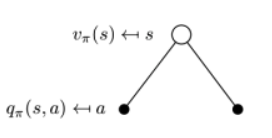
**下图用例子解释了行为价值函数：**

****

* **Bellman期望方程Bellman Expectation Equation**

**MDP下的状态价值函数和行为价值函数与MRP下的价值函数类似，可以改用下一时刻的状态价值函数或者行为价值函数来表达。具体方程如下：**

* **和的关系**

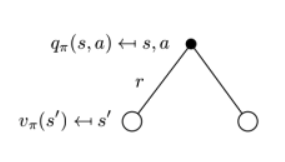
****

**上图中，空心较大圆圈表示状态，黑色实心小圆表示的是动作本身，连接状态和动作的线条仅仅把该状态以及该状态下可以采取的行为关联起来。可以看出，在遵循策略时，状态s的价值体现为在该状态下遵循某一策略而采取所有可能行为的价值与行为发生的概率的乘积之和。**

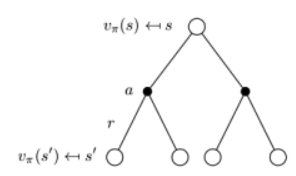
**类似的，一个行为价值函数也可以表示为状态价值函数的形式：**

**它表示，在某一状态下采取一个行为的价值，可以分为两部分：**

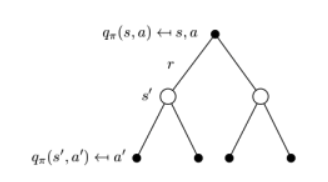
1. **离开这个状态的价值**
2. **进入所有新的状态的价值和它的转移概率的乘积的和**

****

**如果组合起来，可以得到下面的结果：**

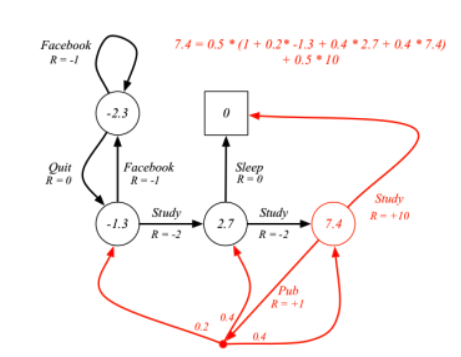
****

**也可以得到下面的结果：**

****

* **学生MDP示例**

**下图解释了红色空心圆圈状态的状态价值函数是如何计算的，遵循的策略随机策略，即所有可能的行为有相同的几率被选择执行。**

****

* **Bellman期望方程矩阵形式**
* **最优价值函数**

**最优状态价值函数指的是在从所有策略产生的状态价值函数中，选取使状态s价值最大的函数：**

**类似的，最优行为价值函数指的是从所有策略产生的行为价值函数中，选取是状态行为对<s,a>价值最大的函数：**

**最优价值函数明确了MDP的最优可能表现，当我们知道了最优价值函数，也就知道了每个状态的最优价值，这时便认为这个MDP获得了解决。**

* **最优策略**

**当对于任何状态s，遵循策略的价值不小于遵循策略下的价值，则策略优于策略：**

**定理：对于任何MDP，下面几点成立：**

1. **存在一个最优策略，比任何其他策略更好或至少相等。**
2. **所有的最优策略有相同的最优价值函数**
3. **所有的最优策略具有相同的行为价值函数**

* **寻找最优策略**

**可以通过最大化最优行为价值函数来找到最优策略：**

**对于任何MDP问题，总存在一个确定性的最优策略；同时，如果我们知道最优行为价值函数，则表明我们找到了最优策略。**