Kintoki

关注机器学习，数据挖掘，人工智能

增强学习（二）----- 马尔可夫决策过程MDP

1. 马尔可夫模型的几类子模型

大家应该还记得马尔科夫链(Markov Chain)，了解机器学习的也都知道隐马尔可夫模型(Hidden Markov Model，HMM)。它们具有的一个共同性质就是马尔可夫性(无后效性)，也就是指系统的下个状态只与当前状态信息有关，而与更早之前的状态无关。

马尔可夫决策过程(Markov Decision Process, MDP)也具有马尔可夫性，与上面不同的是MDP考虑了动作，即系统下个状态不仅和当前的状态有关，也和当前采取的动作有关。还是举下棋的例子，当我们在某个局面（状态s）走了一步(动作a)，这时对手的选择（导致下个状态s'）我们是不能确定的，但是他的选择只和s和a有关，而不用考虑更早之前的状态和动作，即s'是根据s和a随机生成的。

我们用一个二维表格表示一下，各种马尔可夫子模型的关系就很清楚了：

	不考虑动作	考虑动作
状态完全可见	马尔科夫链(MC)	马尔可夫决策过程(MDP)
状态不完全可见	隐马尔可夫模型(HMM)	不完全可观察马尔可夫决策过程(POMDP)

2. 马尔可夫决策过程

一个马尔可夫决策过程由一个四元组构成M = (S, A, P_{sa}, R) [注1]

- S: 表示状态集(states)，有s∈S，s_i表示第i步的状态。
- A:表示一组动作(actions)，有a∈A，a_i表示第i步的动作。
- P_{sa}: 表示状态转移概率。P_{sa} 表示的是在当前s ∈ S状态下，经过a ∈ A作用后，会转移到的其他状态的概率分布情况。比如，在状态s下执行动作a，转移到s'的概率可以表示为p(s'|s,a)。
- R: S×A⟼ℝ ，R是回报函数(reward function)。有些回报函数状态S的函数，可以简化为R: S⟼ℝ。如果一组(s,a)转移到了下个状态s'，那么回报函数可记为r(s'|s, a)。如果(s,a)对应的下

导航

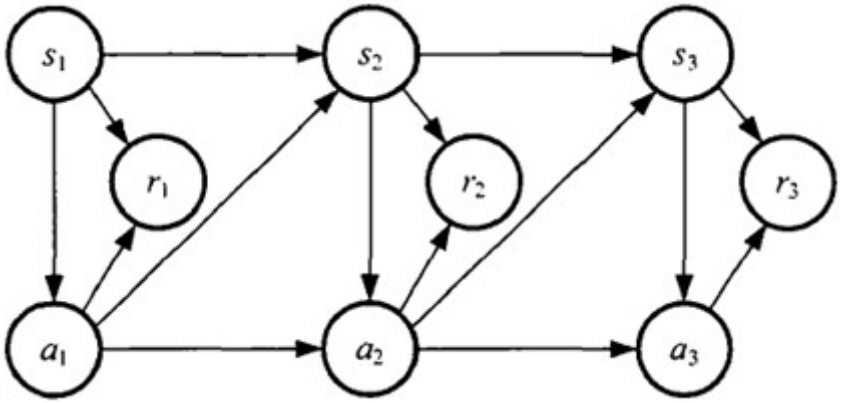
博客园 首页 联系 订阅 管理

个状态s'是唯一的，那么回报函数也可以记为r(s,a)。

MDP 的动态过程如下：某个智能体(agent)的初始状态为s_0，然后从 A 中挑选一个动作a_0执行，执行后，agent 按P_{sa}概率随机转移到了下一个s_1状态，$s_1 \in P_{s0a0}$。然后再执行一个动作a_1，就转移到了s_2，接下来再执行a_2…，我们可以用下面的图表示状态转移的过程。

$$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \xrightarrow{a_3} \cdots$$

如果回报r是根据状态s和动作a得到的，则MDP还可以表示成下图：



3. 值函数(value function)

上篇我们提到增强学习学到的是一个从环境状态到动作的映射（即行为策略），记为策略π: S→A。而增强学习往往又具有延迟回报的特点: 如果在第n步输掉了棋，那么只有状态s_n和动作a_n获得了立即回报$r(s_n,a_n)$=-1，前面的所有状态立即回报均为0。所以对于之前的任意状态s和动作a，立即回报函数r(s,a)无法说明策略的好坏。因而需要定义值函数(value function，又叫效用函数)来表明当前状态下策略π的长期影响。

用$V^{\pi}(s)$表示策略π下，状态s的值函数。r_i表示未来第i步的立即回报，常见的值函数有以下三种：

我的标签

我的标签

机器学习(6)

增强学习(5)

Machine learning(5)

Reinforcement learning(5)

机器人(4)

强化学习(3)

EM算法(1)

Q learning(1)

随笔分类(6)

机器学习(6)

随笔

随笔档案(6)

2016年1月 (1)

2014年2月 (1)

2014年1月 (3)

2013年5月 (1)

积分与排名

积分 - 15575

排名 - 17439

最新评论

1. Re:增强学习（二）----- 马尔可夫决策过程MDP

写的很好，很详细，读懂了。谢谢您~

--逍遥_叹

2. Re:增强学习（三）----- MDP的动态规划解法

a) $$V^{\pi}(s)=E_{\pi}\left[\sum_{i=0}^{h} r_i \middle| s_0=s\right]$$

b) $$V^{\pi}(s)=\lim_{h\to\infty} E_{\pi}\left[\frac{1}{h}\sum_{i=0}^{h} r_i \middle| s_0=s\right]$$

c) $$V^{\pi}(s)=E_{\pi}\left[\sum_{i=0}^{\infty} \gamma^i r_i \middle| s_0=s\right]$$

其中：

a)是采用策略π的情况下未来有限h步的期望立即回报总和；

b)是采用策略π的情况下期望的平均回报；

c)是值函数最常见的形式，式中γ∈[0,1]称为折合因子，表明了未来的回报相对于当前回报的重要程度。特别的，γ=0时，相当于只考虑立即不考虑长期回报，γ=1时，将长期回报和立即回报看得同等重要。接下来我们只讨论第三种形式，

现在将值函数的第三种形式展开，其中r_i表示未来第i步回报，s'表示下一步状态，则有：

$$V^{\pi}(s)=E_{\pi}\left[r_0+\gamma r_1+\gamma^2 r_2+\gamma^3 r_3+\ldots \middle| s_0=s\right]$$
$$=E_{\pi}\left[r_0+\gamma E\left[\gamma r_1+\gamma^2 r_2+\gamma^3 r_3+\ldots\right] \middle| s_0=s\right]$$
$$=E_{\pi}\left[r(s'|s,a)+\gamma V^{\pi}(s') \middle| s_0=s\right]$$

给定策略π和初始状态s，则动作a=π(s)，下个时刻将以概率p(s'|s,a)转向下个状态s'，那么上式的期望可以拆开，可以重写为：

$$V^{\pi}(s)=\sum_{s'\in S} p(s'|s,a)\left[r(s'|s,a)+\gamma V^{\pi}(s')\right]$$

你好，其实我最关系的不是值迭代，也不是策略迭代，而是算法经过很多次的学习之后，对于一个新的初始状态，算法从初始状态一直走到目标状态，也就是说，算法如何在不同的状态下去选择算法认为合适的action去达.....

--GoingMyWay

3. Re:增强学习（二）----- 马尔可夫决策过程MDP

你好，我想问一下算值函数时候为什么在里面加了个E[]?还有那个公式前面好像多乘了个γ

--Hiroki

4. Re:增强学习（二）----- 马尔可夫决策过程MDP

谢谢楼主，写得特别好。对了，部分可观察马尔可夫决策过程及其解法是怎样的？请楼主指教。。

--行僧

5. Re:增强学习（三）----- MDP的动态规划解法

请问:
1有什么靠谱的开源实现，最好python or c,cpp

2那个字母pi是什么意思，第二讲pi还是下表

--xman78

阅读排行榜

1. 增强学习（二）----- 马尔可夫决策过程MDP(12262)
2. 增强学习（五）----- 时间差分学习(Q learning, Sarsa learning)(7430)
3. 增强学习（三）----- MDP的动态规划解法(6801)
4. 增强学习（一）----- 基本概念(6388)
5. 增强学习（四）----- 蒙特卡罗方法(Monte Carlo Methods)(5553)

评论排行榜

1. 增强学习（二）----- 马尔可夫决策过程MDP(11)
2. 增强学习（三）----- MDP的动态规划解法(7)

上面提到的值函数称为状态值函数(state value function)，需要注意的是，在$V^{\pi}(s)$中，π和初始状态s是我们给定的，而初始动作a是由策略π和状态s决定的，即a=π(s)。

定义动作值函数(action value functionQ函数)如下：

$$Q^{\pi}(s,a)=E\left[\sum_{i=0}^{\infty}\gamma^{i}r_{i}\middle|s_{0}=s,a_{0}=a\right]$$

给定当前状态s和当前动作a，在未来遵循策略π，那么系统将以概率p(s'|s,a)转向下个状态s'，上式可以重写为：

$$Q^{\pi}(s,a)=\sum_{s'\in S}p(s'|s,a)\left[r(s'|s,a)+\gamma V^{\pi}(s')\right]$$

在$Q^{\pi}(s,a)$中，不仅策略π和初始状态s是我们给定的，当前的动作a也是我们给定的，这是$Q^{\pi}(s,a)$和$V^{\pi}(a)$的主要区别。

知道值函数的概念后，一个MDP的最优策略可以由下式表示：

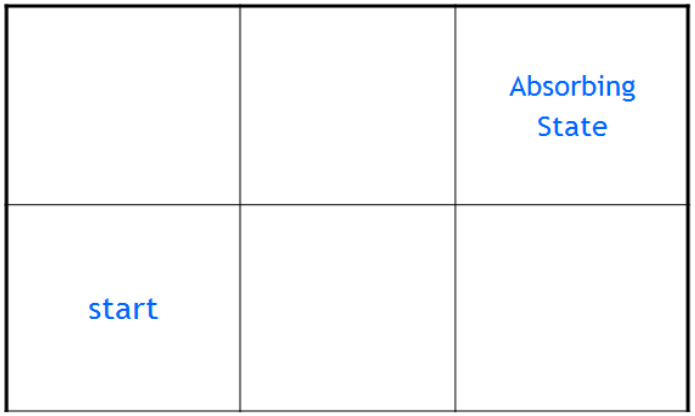
$$\pi^{*}=\arg\max_{\pi}V^{\pi}(s),(\forall s)$$

即我们寻找的是在任意初始条件s下，能够最大化值函数的策略π*。

4. 值函数与Q函数计算的例子

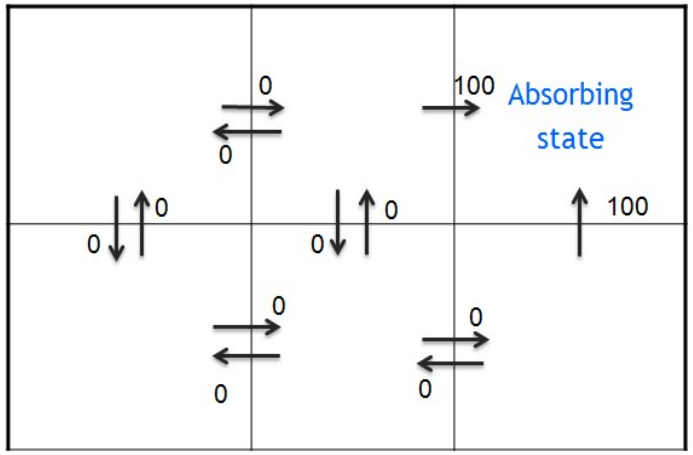
上面的概念可能描述得不够清晰，接下来我们实际计算一下，如图所示是一个格子世界，我们假设agent从左下角的start点出发，右上角为目标位置，称为吸收状态(Absorbing state)，对于进入吸收态的动作，我们给予立即回报100，对其他动作则给予0回报，折合因子γ的值我们选择0.9。

为了方便描述，记第i行，第j列的状态为s_{ij}, 在每个状态，有四种上下左右四种可选的动作，分别记为a_u,a_d,a_l,a_r。（up，down，left，right首字母），并认为状态按动作a选择的方向转移的概率为1。

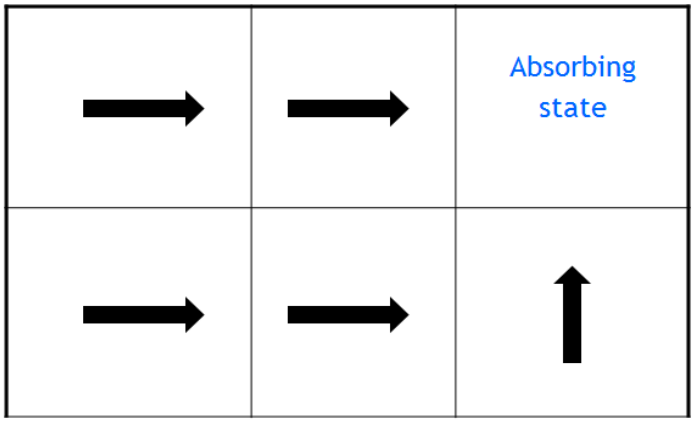


1.由于状态转移概率是1，每组(s,a)对应了唯一的s'。回报函数r(s'|s,a)可以简记为r(s,a)

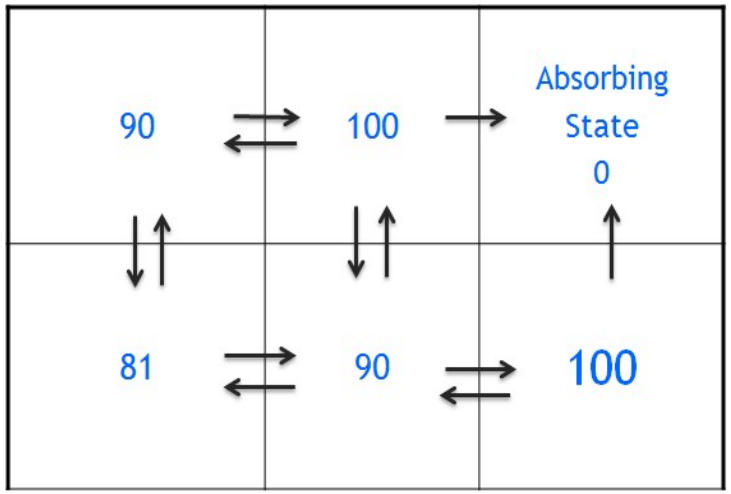
如下所示，每个格子代表一个状态s，箭头则代表动作a，旁边的数字代表立即回报，可以看到只有进入目标位置的动作获得了回报100，其他动作都获得了0回报。 即$r(s_{12},a_r) = r(s_{23},a_u) =100$。



2. 一个策略π如图所示：



3. 值函数$V^{\pi}(s)$如下所示



根据V^{π}的表达式，立即回报，和策略π，有

$V^{\pi}(s_{12}) = r(s_{12},a_r) = r(s_{13}|s_{12},a_r) = 100$

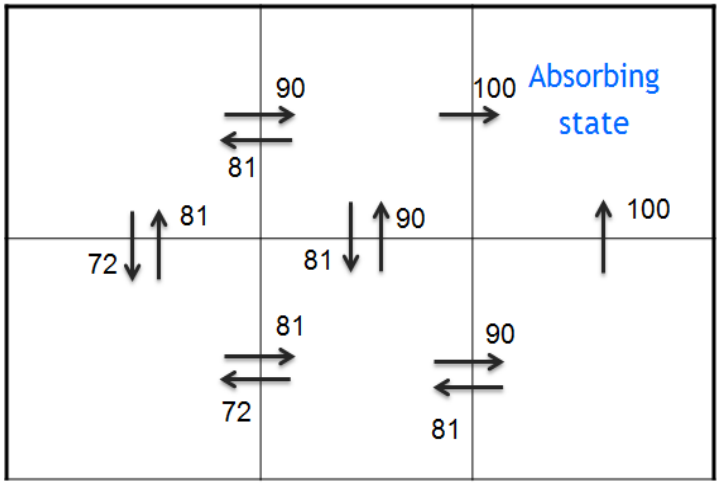
$V^{\pi}(s_{11})= r(s_{11},a_r)+\gamma*V^{\pi}(s_{12}) = 0+0.9*100 = 90$

$V^{\pi}(s_{23}) = r(s_{23},a_u) = 100$

$V^{\pi}(s_{22})= r(s_{22},a_r)+\gamma*V^{\pi}(s_{23}) = 90$

$V^{\pi}(s_{21})= r(s_{21},a_r)+\gamma*V^{\pi}(s_{22}) = 81$

4. Q(s,a)值如下所示



有了策略π和立即回报函数r(s,a), $Q^{\pi}(s,a)$如何得到的呢？

对s_{11}计算Q函数（用到了上面V^{π}的结果）如下：

$Q^{\pi}(s_{11},a_r)=r(s_{11},a_r)+ \gamma *V^{\pi}(s_{12}) =0+0.9*100 = 90$

$Q^{\pi}(s_{11},a_d)=r(s_{11},a_d)+ \gamma *V^{\pi}(s_{21}) = 72$

至此我们了解了马尔可夫决策过程的基本概念，知道了增强学习的目标（获得任意初始条件下，使V^{π}值最大的策略π^*），下一篇开始介绍求解最优策略的方法。

PS:发现写东西还是蛮辛苦的，希望对大家有帮助。另外自己也比较菜，没写对的地方欢迎指出~~

[注]采用折合因子作为值函数的MDP也可以定义为五元组M=(S, A, P, γ, R)。也有的书上把值函数作为一个因子定义五元组。还有定义为三元组的，不过MDP的基本组成元素是不变的。

参考资料：

[1] R.Sutton et al. Reinforcement learning: An introduction , 1998

[2] T.Mitchell. 《机器学习》，2003

[3] 金卓军，逆向增强学习和示教学习算法研究及其在智能机器人中的应用[D]，2011

[4] Oliver Sigaud et al，Markov Decision Process in Artificial Intelligence[M], 2010

分类: 机器学习

标签: 机器学习, 增强学习, Machine learning, Reinforcement learning, 机器人, 强化学习

« 上一篇：增强学习（一）----- 基本概念
» 下一篇：增强学习（三）----- MDP的动态规划解法

posted on 2014-01-14 00:21 金淑林 阅读(12262) 评论(11) 编辑 收藏

评论

#1楼 2014-01-14 15:04 钱吉

高深的算法啊。搞这些东西的都是将来的大牛

支持(0) 反对(0)

#2楼[楼主] 2014-01-14 15:53 金淑林

@ DarkHorse
过奖了，还很水~~

支持(0) 反对(0)

#3楼 2014-01-15 21:01 殁殇

支持，希望继续写下去！
没明白V和Q具体怎么计算的

支持(0) 反对(0)

#4楼[楼主] 2014-01-16 16:33 金淑林

@ 殁殇
谢谢！我写的比较慢~ 不过会坚持把基本的算法都写完的。
上面关于Q和V的计算，其实是在π确定下进行的，刚才我重新组织了下这篇，加入了计算过程。你可以重新看下3和4

支持(0) 反对(0)

#5楼 2014-01-23 22:43 代晏

楼主写得挺好的！第一遍看到中间时公式有点多，就跳了看了（脑力不够），过一会再回头又看一遍就懂了。网上很少有介绍增强学习的，楼主的排版看着也非常舒服，32个赞！

支持(0) 反对(0)

#6楼 2014-03-20 18:41 topone007

楼主写的很好，楼主辛苦了，非常赞!

支持(1) 反对(0)

#7楼 2014-11-15 21:58 super1Angle

楼主写得确实挺好的，给的例子有助于理解，赞一个，感谢了！

支持(1) 反对(0)

#8楼 2015-12-07 04:13 陈家小Q

多谢楼主，非常有用！

支持(1) 反对(0)

#9楼 2016-04-06 08:44 行僧

谢谢楼主，写得特别好。对了，部分可观察马尔可夫决策过程及其解法是怎样的？请楼主指教。。

支持(0) 反对(0)

#10楼 2016-04-14 18:48 Hiroki

你好，我想问一下算值函数时候为什么在里面加了个E[]?还有那个公式前面好像多乘了个γ

支持(2) 反对(0)

#11楼 2017-05-10 14:40 逍遥_叹

写的很好，很详细，读懂了。谢谢您~

支持(0) 反对(0)

最新IT新闻:

· 京东建首家电商生鲜检测中心：刘强东女儿只吃这里
· 乐视网CEO梁军新官上任之后强调了四点内容
· 视障女孩陈思颖的一天：感谢支付宝
· 贾跃亭：乐视无控制权之争，将触底反弹
· 李飞飞：我把今天AI所处的发展阶段称为“AI in vivo”

» 更多新闻...

最新知识库文章:

· 软件开发为什么很难
· 唱吧DevOps的落地，微服务CI/CD的范本技术解读
· 程序员，如何从平庸走向理想？
· 我为什么鼓励工程师写blog
· 怎么轻松学习JavaScript

» 更多知识库文章...