





听见下雨的声音

-  首页
-  分类
-  关于
-  归档
-  标签

【David Silver强化学习公开课之一】强化学习入门

 发表于 2016-06-06 |  分类于 [project experience](#) |  |  6893

本文是David Silver强化学习公开课第一课的总结笔记。第一课主要解释了强化学习在多领域的体现，主要解决什么问题，与监督学习算法的区别，完整的算法流程由哪几部分组成，其中的agent又包含什么内容，以及解释了强化学习涉及到的一些概念。

【转载请注明出处】chenrudan.github.io

本文是David Silver强化学习公开课第一课的总结笔记。第一课主要解释了强化学习在多领域的体现，主要解决什么问题，与监督学习算法的区别，完整的算法流程由哪几部分组成，其中的agent又包含什么内容，以及解释了强化学习涉及到的一些概念。

本课视频地址:[RL Course by David Silver - Lecture 1: Introduction to Reinforcement Learning](#)。

本课ppt地址:http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/intro_RL.pdf。

文章的内容是课程的一个总结和讨论，会按照自己的理解来组织。个人知识不足再加上英语听力不是那么好可能会有一些理解不准的地方，欢迎一起讨论。




建了一个强化学习讨论qq群，有兴趣的可以加一下群号595176373或者扫描下面的二维码。



1. 强化学习是什么

强化学习是多学科多领域交叉的一个产物，它的本质就是解决“decision making”问题，即学会自动进行决策。在computer science领域体现为机器学习算法。在Engineering领域体现在决定the sequence of actions来得到最好的结果。在Neuroscience领域体现在理解人类大脑如何做出决策，主要的研究是reward system。在Psychology领域，研究动物如何做出决策，动物的行为是由什么导致的。在Economics领域体现在博弈论的研究。这所有的问题最终都归结为一个问题，人为什么能够并且如何做出最优决策。

强化学习是一个Sequential Decision Making问题，它需要连续选择一些行为，从而这些行为完成后得到最大的收益最好的结果。它在没有任何label告诉算法应该怎么做的情况下，通过先尝试做出一些行为得到一个结果，通过判断这个结果是对还是错来对之前的行为进行反馈，然后由这个反馈来调整之前的行为，通过不断的调整，算法能够学习到在这样的情况下选择什么样的行为可以得到最好的结果。

© 2017  Rudan Chen
由Hexo强力驱动 | 主题 - NexT.Muse
 55280 |  114526

强化学习与监督学习有着不少区别，首先监督学习是有一个label的，这个label告诉算法什么样的输入对应什么样的输出，而强化学习没有label告诉它在某种情况下应该做出什么样的行为，只有一个做出一系列行为后最后反馈回来的reward signal，这个signal能判断当前选择的行为是好是坏。其次强化学习的结果反馈有延时，有可能需要走了很多步以后才知道以前的某一步的选择是好还是坏，而监督学习做了比较坏的选择会立刻反馈给算法。强化学习面对的输入总是在变化，输入不像监督学习是独立同分布的。而每当算法做出一个行为，它影响了下一决策的输入。

2. 强化学习组成

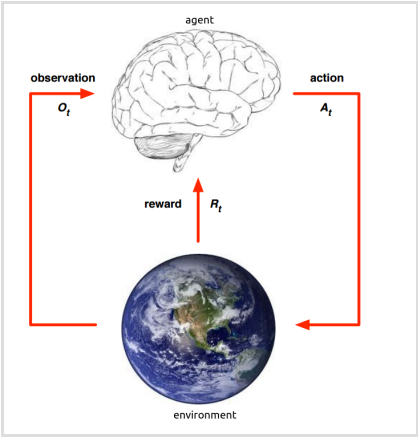


图1 强化学习组成部分(图片来源[1])

强化学习决策流程见上图。需要构造出一个agent(图中的大脑部分)，agent能够执行某个action，例如决定向左还是向右走，围棋棋子下在哪个位置。agent能够接收当前环境的一个observation，例如当前机器人的摄像头拍摄到的场景。agent还能接收当它执行某个action后的reward，即在第t步agent的工作流程是执行一个动作 A_t ，然后接收该动作之后的环境观测状况 O_t ，以及获得这个动作的反馈奖赏 R_t 。而环境environment则是agent交互的对象，它是一个行为不可控制的对象，agent一开始不知道环境会对不同action做出什么样的反应，而环境会通过observation告诉agent当前的环境状态，同时环境能够根据可能的最终结果反馈给agent一个reward，例如围棋棋面就是一个environment，它可以根据当前的棋面状况估计一下黑白双方输赢的比例。因而在第t步，environment的工作流程是接收一个 A_t ，对这个动作做出反应之后传递环境状况和评估的reward给agent。reward奖赏是一个反馈标量值，它表明了在第t步agent做出的决策有多好或者有多不好，整个强化学习优化的目标就是最大化累积reward。例如在射击游戏中，击中敌方的一架飞机，最后的得分会增加，那么这一步的reward就是正值。

3. 一些变量

history是所有动作、状态、奖赏的序列， $H_t = A_1, O_1, R_1, \dots, A_t, O_t, R_t$

environment state， S_t^e ，环境当前的状态，它反应了环境发生什么改变。这里需要明白的一点是环境自身的状态和环境反馈给agent的状态并不一定是相同的，例如机器人在走路时，当前的environment状态是一个确定的位置，但是它的camera只能拍到周围的景象，无法告诉agent具体的位置，而拍摄到的照片可以认为是对环境的observation，也就是说agent并不是总能知道环境是如何发生改变的，只能看到改变后的一个结果展示。

agent state， S_t^a ，是agent的现在所处状态的表示，它可以是history的任何函数。

information(Markov) state，它包含了history的所有有用信息。一个状态 S_t 有马尔可夫性质是指下一个时刻的状态仅由当前状态决定，与过去状态无关。这里定义可以看出environment state是有马尔可夫性质的(这个概念暂时不管)。

如果说environment是Fully Observable的，那么就是说agent能够直接看到环境当前的状态，在这种情况下state与environment state是相等的。而如果说environment是Partially Observable Environments，那么就是机器人的那个例子，agent能获取到的不是直接的环境状态。

4. Agent的组成

一个agent由三部分组成Policy、Value function、Model，但这三部分不是必须同时存在的。

[文章目录](#) [站点概览](#)

- [1.1. 强化学习是什么](#)
- [2.2. 强化学习组成](#)
- [3.3. 一些变量](#)
- [4.4. Agent的组成](#)
- [5.5. 探索和利用](#)
- [6.6. 引用](#)

Policy，它根据当前看到的observation来决定action，是从state到action的映射。有两种表达形式，一种是Deterministic policy即 $a = \pi(s)$ ，在某种状态s下，一定会执行某个动作a。一种是Stochastic policy即 $\pi(a|s) = p[A_t = a|S_t = s]$ ，它是在某种状态下执行某个动作的概率。

Value function，它预测了当前状态下未来可能获得的reward的期望。 $V_{\pi}(s) = E_{\pi}[R_{t+1} + \gamma V_{\pi}(S_{t+1}) | S_t = s]$ 。用于衡量当前状态的好坏。

Model，预测environment下一步会做出什么样的改变，从而预测agent接收到的状态或者reward是什么。分为两种类型的model，一种是预测下一个state的transition model即 $P_{ss'}^a = p[S_{t+1} = s' | S_t = s, A_t = a]$ ，一种是预测下一次reward的reward model即 $R_s^a = E[R_{t+1} | S_t = s, A_t = a]$

因而根据是否选取这三个部分agent可分为下图中红色字体标出来的五种类型(这里有一个迷宫的例子很好，见原视频 1:08:10起)。Model Free是指不需要去猜测environment的工作方式，而Model based则需要去猜测environment的工作方式。

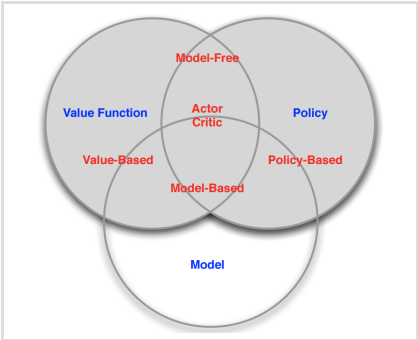


图2 Agent的分类(图片来源[1])

5. 探索和利用

强化学习是一种试错(trial-and-error)的学习方式，一开始不清楚environment的工作方式，不清楚执行什么动作是对的，什么样是错的。因而agent需要从不断尝试的经验中发现一个好的policy，从而在这个过程中获得最大的reward。

在这样的学习过程中，就会有一个在Exploration和Exploitation之间的权衡，前者是说会放弃一些已知的reward，而去尝试一些新的选择，即在某种状态下，算法也许已经学习到选择什么action让reward比较大，但是每次都做出同样的选择，也许另外一个没有尝试过的选择会让reward更大，即Exploration希望能够探索更多environment的信息。而后者是指根据已知的信息最大化reward。例如，在选择一个餐馆时，Exploitation会选择你最喜欢的餐馆，而Exploration会尝试选择一个新的餐馆。

以上是第一课的一些相关内容，主要是介绍了一些基础概念，从而对强化学习有一个基础的认识。

6. 引用

1. http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/intro_RL.pdf

1.1. 强化学习是什么

2.2. 强化学习组成

3.3. 一些变量

4.4. Agent的组成

5.5. 探索和利用

6.6. 引用

[文章目录](#) [站点概览](#)

- [1.1. 强化学习是什么](#)
- [2.2. 强化学习组成](#)
- [3.3. 一些变量](#)
- [4.4. Agent的组成](#)
- [5.5. 探索和利用](#)
- [6.6. 引用](#)