# 强化学习系列之七:在 @prerryAltu@yaman上实现w@utcelar/riting |

AlgorithmDog

主题 (http://www.tuicool.com/topics) 活动 (http://huodong.tuicool.com/)

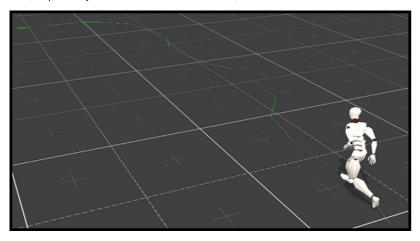
APP <sup>荐</sup> (http://www.tuicool.com/mobile)

时间 2016-06-14 00:24:56 @ AlgorithmDog (/sites/6736je2)

原文

主题 OpenAI (/topics/11120123)

这周试用了下 OpenAl Gym。OpenAl Gym是一款用于研发和比较强化学习算法的工具包。强化学习和有监督学习的评测不一样。有监督学习的评测工具是 数据。只要提供一批有标注的数据就能进行有监督学习的评测。强化学习的评测工具是环境。需要提供一个环境给 Agent 运行,才能评测 Agent 的策略的 优劣。OpenAl Gym 是提供各种环境的开源工具包。



# 1. OpenAl Gym 的基本知识

下面 OpenAI Gym 是一个示例。

```
import gym
env = gym.make('CartPole-v0') //实例化一个 CartPole 环境
for i_episode in range(20):
   observation = env.reset()
   for t in range(100):
        env.render() //更新动画
        action = env.action space.sample()
       observation, reward, done, info = env.step(action) //推进一步
        if done:
           break
```



(http://sa-summit.org/?chanr

(https://dami.ksyun.com/spe

IFF3J)

上的实现

ch=0006300014006666&hmsr=%E6%8E%A8%E9%85%B7&hmpl=special666



势 (/articles/AbeIreY) les/qau6Zfu)

(/articles/IrUNVnJ) (http://pm-summit.org/?chan

(http://click.aliyun.com/m/17039/)

OpenAI Gym 的最重要的功能就是提供各种强化学习环境。上面的代码 env = gym.make('CartPole-v0') 是实例化一个 CartPole 环境。CartPole 环境要求平衡一辆车上的一根棍子,如下图的第一个环境表示。下图是 安尼 提供的部分自动控制分值的环境。除此之外 OpenAI Gym 还提供了算法、文本和游戏方面的环境,具体可以查看 官方说明 (https://gym.openai.com/envs)。

Classic control problems from the RL literature.

文章 (http://www.tuicool.com/ah) 站点 (http://www.tuicool.com/sites/hot)



强化学习环境其实是马尔科夫决策过程,马尔科夫决策过程的四个基本元素:状态、动作、转移概率和奖励函数。

Acrobot-v0 Swing up a two-link robot.

#### 1.1 状态

代码中的 observation 就是马尔科夫决策过程的状态。更正确地说是,状态的特征。CartPole-v0 的状态特征是一维数组,比如array([-0.01377819, -0.01291427, 0.02268009, -0.0380999])。有些环境提供的状态特征是二维数组,比如 AirRaid-ram-v0 环境提供的是二维数组表示的游戏画面。

observation = env.reset() 是初始化环境,设置一个随机或者固定的初始状态。env.step(a1) 是环境接受动作 a1,返回的第一个结果是接受动作 a1 之后的状态特征。

#### 1.2 动作

代码中的 action 就是马尔科夫决策过程中的动作。CartPole-v0 的动作是离散型特征。在 OpenAl Gym 中,离散型动作是用从 0 开始的整数集合表示,比如 CartPole-v0 的动作有 0 和 1。另一种动作是连续型,用实数表示。

## 1.3 转移概率和奖励函数

在 OpenAI Gym 中,转移概率并没有显式表示出来,而是通过 env.step(a1) 的结果表示。env.step(a1) 返回的 observation 满足转移概率。

代码中的 reward 就是马尔科夫决策过程中的动作,用实数表示。在 OpenAl Gym 中,奖励函数也没有显式表示出来,也是通过 env.step(a1) 的结果表示。env.step(a1) 返回的 reward 满足奖励函数。

值得一提的是,env.step 返回的第四个结果 info 是系统信息,给开发人员调试用,不允许学习过程使用。本文只介绍在 OpenAl Gym 上实现 Q Learning 算法需要的知识。想了解更多 OpenAl Gym 知识,可以参考 OpenAl Gym 官方文档 (https://gym.openai.com/docs)。

# 2. 实验Q Learning 算法

我们在 OpenAI Gym 的 CartPole-v0 环境上实现Q Learning 算法。Q Learning 目标是学习状态动作价值。QLearning 让 Agent 按照策略进行探索,在探索每一步都进行状态价值的更新,更新公式如下。

(1)

由于 OpenAI Gym 提供的状态特征,因此我们要用价值函数近似,参数更新的代码如下所示。

```
def update(policy, s_fea, a, tvalue, alpha):
    pvalue = policy.qfunc(s_fea, a);
    error = pvalue - tvalue;
    s_a_fea = policy.get_state_action_fea(s_fea, a);
    policy.theta -= alpha * error * s_a_fea;
```



(http://sa-summit.org/?chanr

(https://dami.ksyun.com/spe

ch=00033.00018.6666&hmsr=%E6%8E%A8%E9%85%B7&hmpl=special666



(http://pm-summit.org/?chan

```
Q Learning 的代码如下。
```

## 推酷 (http://www.tuicool.com/)

```
def qlearning(env, policy, num_iter1, alpha, gamma):
   for i in xrange(len(policy.theta)):
                                      文章 (http://www.tuicool.com/ah)
                                                                  站点 (http://www.tuicool.com/sites/hot)
       policy.theta[i] = 0.1
                                      主题 (http://www.tuicool.com/topics)
                                                                    活动 (http://huodong.tuicool.com/)
   for iter1 in xrange(num_iter1):
       s_f
               = env.reset()
               = 0
       count
               = False
                                       搜索
                                                                                     登录 (http://www.tuicool
       while False == t and count < 10000:
```

```
while False == t and count < 10000:
    s_f1,r,t,i = env.step(a)
    qmax = policy.qfunc(s_f1,a) #random
    for a1 in policy.actions:
        pvalue = policy.qfunc(s_f1, a1);
        if qmax < pvalue:
            qmax = pvalue;
        update(policy, s_f, a, r + gamma * qmax, alpha);

    s_f = s_f1
        a = policy.epsilon_greedy(s_f)
        count += 1

return policy;</pre>
```



(http://sa-summit.org/?chanr

(https://dami.ksyun.com/spe

ch=00033.00018.6666&hmsr=%E6%8E%A8%E9%85%B7&hmpl=special666



(http://pm-summit.org/?chan

想了解更多代码,可以参见 Github

(https://github.com/algorithmdog/Reinforcement\_Learning\_Blog/tree/master/8.%E5%BC%BA%E5%8C%96%E5%AD%A6%E4%B9%A0%E7%B3%BB%E5%88%97%E4%B9%8B%E5%85%AB:OpenAi\_Gym\_%E8%AF%95%E7%94%A8)。实现的 Q Learning 算法的效果如下。

## 推酷 (http://www.tuicool.com/)





ch=00033.00018.6666&hmsr=%E6%8E%A8%E9%85%B7&hmpl=special666



(http://pm-summit.org/?chan

## 3. 总结

结果好烂啊。基本的强化学习算法还是无法解决 OpenAI Gym 里面的问题啊。本文的代码可以在 Github (https://github.com/algorithmdog/Reinforcement Learning Blog/tree/master/8.%E5%BC%BA%E5%8C%96%E5%AD%A6%E4%B9%A0%E7%B3%BB%E5%88%97%E4%B9%8B%E5%85%AB:OpenAi Gym %E8%AF%95%E7%94%A8) 上找到,欢迎有兴趣的同学帮我挑挑毛病。

最后欢迎关注我的公众号 AlgorithmDog,每周日的更新就会有提醒哦~



# 强化学习系列系列文章

- 强化学习系列之一:马尔科夫决策过程 (http://www.algorithmdog.com/%e5%bc%ba%e5%8c%96%e5%ad%a6%e4%b9%a0-%e9%a9%ac%e5%b0%94%e7%a7%91%e5%a4%ab%e5%86%b3%e7%ad%96%e8%bf%87%e7%a8%8b)
- 强化学习系列之二:模型相关的强化学习 (http://www.algorithmdog.com/%e5%bc%ba%e5%8c%96%e5%ad%a6%e4%b9%a0%e7%b3%bb%e5%88%97%e4%b9%8b%e4%ba%8c-%e6%a8%a1%e5%9e%8b%e7%9b%b8%e5%85%b3%e7%9a%84%e5%bc%ba%e5%8c%96%e5%ad%a6%e4%b9%a0)
- 强化学习系列之三:模型无关的策略评价 (http://www.algorithmdog.com/reinforcement-learning-model-free-evalution)
- 强化学习系列之四:模型无关的策略学习 (http://www.algorithmdog.com/reinforcement-learning-model-free-learning)
- 强化学习系列之五:价值函数近似 (http://www.algorithmdog.com/reinforcement-learning-value-function-approximation)
- 强化学习系列之六:策略梯度 (http://www.algorithmdog.com/rl-policy-gradient)
- 强化学习系列之七:逆向强化学习 (http://www.algorithmdog.com/rl-irl)
- 强化学习系列之七:在 OpenAI Gym 上实现 Q Learning



推酷 (http://www.tuicool.com/)

文章 (http://www.tuicool.com/ah) 站点 (http://www.tuicool.com/sites/hot)

金山云

主题 (http://www.tuicool.com/topics)

活动 (http://huodong.tuicool.com/)

618

▲ 纠错

2核4G100G 4克性 / Produce rem/mobiles

周刊 ▼ 更多 ▼ (https://dami.ksyun.com/special-

大米云主机 618 年中大促 限量抢购

登录 (http://www.tuicool

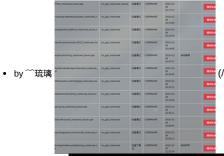
product/index.html?ch=00033.00018.6666&hmsr=%E6%8E%A8%E9%85%B7&hmpl=special6666&hmcu=&hmkw=&hmci=)

## 推荐文章

☆ 收藏

- 1. NEAT学习: 教机器自我编程 (/articles/EJNFF3J)
- 2. 详解Wassertein GAN:使用Keras在MNIST上的实现 (/articles/3InQBvb)
- 3. 从零开始:如何使用LSTM预测汇率变化趋势 (/articles/AbelreY)
- 4. arXiv Paper Daily: Wed, 4 Oct 2017 (/articles/gau6Zfu)
- 5. Theano停止更新之后,开发者们怎么说? (/articles/IrUNVnJ)
- 6. Deepmind "预测地图"论文背后:神经科学或将助力深度学习迎来新突破 (/articles/QJziQnI)

## 相关推刊



(/kans/2815762600) 《默认推刊》 (/kans/2815762600) 10



(/kans/3457356399) 《有用的》 (/kans/3457356399) 26



(http://sa-summit.org/?chanr

(https://dami.ksyun.com/spe

ch=00033.00018.6666&hmsr=%E6%8E%A8%E9%85%B7&hmpl=special666



(http://pm-summit.org/?chan

	文本加密为盲文	推酷 (http://www.tuicool.com/)				
	ien	文章 (http://www.tu	uicool.com/ah)	站点 (http://www.tuicoc	ol.com/sites/hot)	
• by jack_014581	SSERVOLETS**	(/kans/3083434343) 《默认推刊》 (/kans/3083434343) <i>267</i> <b>主题 (http://www.tuicool.com/topics)</b> 活动 (http://huodong.tuicool.com/)				
	1. 地程可的改变规模中设定—1 完好。点点处理机、万州县产加密和扩大。 1. 北海有空沙州在亚沙州亚沙州、北京军市场、万州县产业区区广大市、区域大学公司市场、沙河、海州市场、西南州市场、西南州市场和市场、沙河、海州市场、西南州市场、沙河、海州市场、西南州市、西南州市、西南州市、西南州市、西南州市、西南州市、西南州市、西南州市	APP <sup>荐</sup> (http://wwv	APP <sup>荐</sup> (http://www.tuicool.com/mobile)  周刊 ▼  更多 ▼			
我来评几句	<ul> <li>Ф. ВВИЗВИ В.Х. В.Х. В.Х. В.Х. В.Х. В.Х. В.Х. В.</li></ul>	搜索			登录 (http://www.tuicool	
请输入评论内容.			登录后评论			
					(	
已发表评论数(0)	)					

2017全球软件架构技术大会 (http://sa-summit.org/?chanr 点击免费领取大会PPT (https://dami.ksyun.com/spe 特价云主机强势来袭 ch=00033.00018.6666&hmsr=%E6%8E%A8%E9%85%B7&hmpl=special666 2017全球产品经理大会 12月07-09日 上海 (http://pm-summit.org/?chan 点击免费领取大会PPT

关于我们 (http://www.tuicool.com/about) 移动应用 (http://www.tuicool.com/mobile) 意见反馈 (http://www.tuicool.com/bbs/go/issues) 官方微博 (http://e.weibo.com/tuicool2012)