# [番外]如何从师生关系理解环境与智能体的互动？

[](https://www.zhihu.com/people/zhang-si-jun-52)

[张斯俊](https://www.zhihu.com/people/zhang-si-jun-52" \t "/home/admin/文档\\x/_blank)

愿成为一把梯子，助你跨过无数的坑。

在Qlearning开始前，我们先了解一下GYM。 其实和GYM相关的教程已经很多，这里只说最简单。并且希望让大家通过了解GYM,对智能体和环境互动有更深入的了解。

GYM是openAI出的一个模拟环境。除了简单的环境，还有atari游戏。我们日后会尝试把学到的算法，运用到atari游戏中。 之所以选择GYM有两个原因：

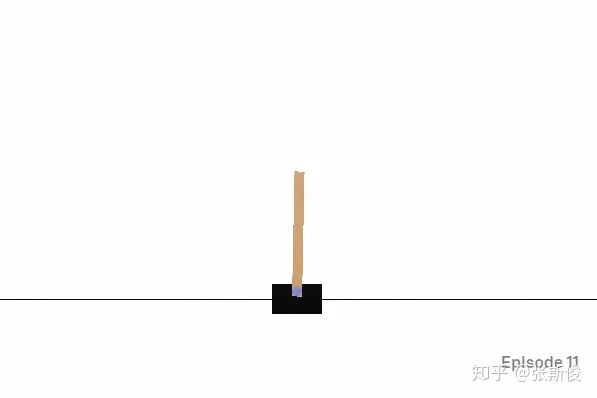
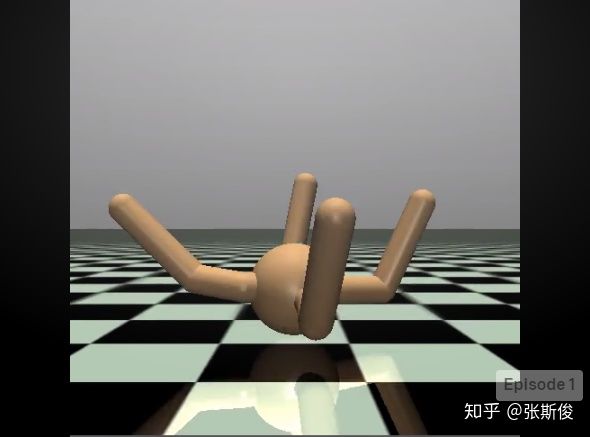
## **环境丰富**

官网:

[https://gym.openai.com/​](https://link.zhihu.com/?target=https://gym.openai.com/" \t "/home/admin/文档\\x/_blank)

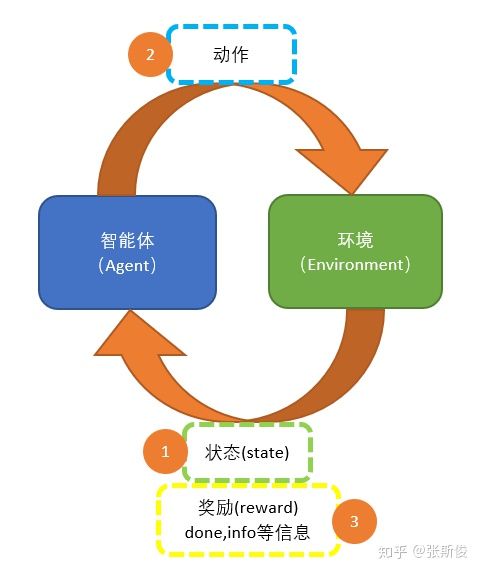
[gym.openai.com](https://link.zhihu.com/?target=https://gym.openai.com/" \t "/home/admin/文档\\x/_blank)

大家可以到官网上看一下GYM包含的环境。包含了从简单的Classic control、Toy text，到更复杂的MuJoCo、Robotics，当然包括Atari游戏。 环境还在不断扩展，现在也包括一些第三方的环境，例如围棋等。 所以我们经常用它来检验算法的有效性。



### 环境与智能体的互动

GYM环境对马可洛夫链(树)的抽象非常高。我们可以从下图看到智能体和环境的互动，结合我们之前说的用法，将会对环境有更深刻的理解。



其实我们可以理解，环境就是老师，负责出题和打分。智能体就是学生，负责做题。智能体的目标就是在老师的题海战术中不断学习，最终获得最多的分数。

我们从这个角度来看一下：

1. 老师首先出题，学生听题。环境初始化一个初始的状态S，智能体把S输入到自己算法。
2. 学生做题。智能体把S输入之后，经过算法进行一番计算。最终会输出action。
3. 老师改题。老师对学生做的题进行评价。也就是说对这个action进行评价，评价最终输出奖励。奖励越高，刚刚智能体的输出评价高：哟！不错哟！同时，老师会发下一题，就是下一个状态S。

虽然老子说：天地不仁以万物为刍狗。 但如果用老师-学生去理解环境和智能体的互动，是不是更简单了呢？

GYM对环境已经高度抽象，我们只需要执行对应的代码，就能让智能体和环境进行互动。减少了我们很多我们自己写环境的代码。我们以后工作一定会把算法用在新的环境，而新环境一般需要自己写的，这时候，参考GYM的这套框架就可以了。

### GYM的基本用法

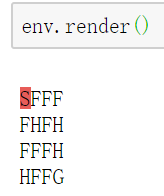
如果大家不会安装GYM环境，可以上网找一下，一般只需要pip一下即可。 gym的用法其实也挺简单的。基本需要用到这几个函数。

1. 导入：首先我们先import gym。导入gym包。
2. 创建环境：env = gym.make('环境名')。 其中环境名可以在这找到，还有相关的说明。其中，env就是environment的简称。 一般我们会顺手进行初始化。env.unwrapped 或者我们一起写。env = gym.make('环境名').unwrapped
3. 获取状态特征数和动作空间 N\_S = env.observation\_space N\_A = env.action\_space 还记得我们之前说的吗？状态特征，就是我们能观察到的环境。 例如自动驾驶汽车，交通灯算是一个状态特征，这个特征有不同的值，红灯/绿灯。距离最近的行人是一个特征，例如距离这个行人多远。自动驾驶汽车能观察的环境特征越多，感知能力越强，对抉择就越有帮助。所以不一般来说，工程师都会想不断利用各种算法和硬件，从真实环境提取环境特征。 GYM环境对于初学者方便的地方是，对于每个实验，环境都直接提供了环境特征。不需要我们自己去提取了。 如果是连续的状态空间，和动作空间，我们可以获取到边界值 s\_low\_bound = env.observation\_space.low s\_high\_bound = env.observation\_space.high 在这里大家可以理解为，老师出题的格式，和学生答案的格式。
4. 初始化环境，并返回一个初始状态。我们一般这样写。 s = env.reset() 老师出的第一道题！
5. 学生开始做题了。这里的f其实就是我们的算法。 现在我们就有了一个初始状态s。 假设有一个函数f（s），输入s，获得可以选择出动作a。我们的目标就是找出一个f，能让我们在这个环境中获取最多的reward。 a = f（s）
6. 学生提交答案，老师改卷了。把分返回给学生。并且开始出下一题。 我们选取a后。代入step函数，会根据 环境配置，会返回：进入的下一个状态，当前获得的奖励r。除此之外，还有一个标志位done，表示是否进入最终状态。另外还有一个包含信息的值(info)，但一般我们用不到。可以先忽略。 s\_,r,done,info = env.step(a)
7. 我们把s = s\_.并把4-6组成一个循环，那么游戏就能自动进行下去了。
8. render。如果有render可以把游戏给渲染出来。

### FrozenLake

FrozenLake是我们学习Qlearning要用到的环境。我们来熟悉一下。

我们索性地把环境渲染出来。长这样子。



S：start，出发点。 G：target，终点。 F：可以走的路，但偶尔你会因为滑到，自动往前滑一格。 H：hole：是陷阱。

FrozenLake-v0是一个4\*4的网络格子，每个格子可以是起始块，目标块、冻结块或者危险块。 我们的目标是让agent学习从开始块如何行动到目标块上，而不是移动到危险块上。 agent可以选择向上、向下、向左或者向右移动，同时游戏中还有可能吹来一阵风，将agent吹到任意的方块上。

我们的任务就是让我们的agent，从起点，走到终点。

好了，现在我们万事俱备，开始进入鼎鼎大名的Qlearning学习吧！

但在此之前，请大家复习好之前时序差分的介绍。因为Qlearning是基于时序差分的喔！

[张斯俊：如何用时序差分TD估算状态V值？​](https://zhuanlan.zhihu.com/p/110132710" \t "/home/admin/文档\\x/_blank)

===========你的支持，就是在下努力的原动力===========

如果专栏对你有用，请点赞并关注在下喔。如果发现有问题，也可以在文章下留言。

你的每一点关注，都是在下的继续努力的动力来源！感激！

[白话强化学习​](https://zhuanlan.zhihu.com/c_1215667894253830144" \t "/home/admin/文档\\x/_blank)zhuanlan.zhihu.com[](https://zhuanlan.zhihu.com/c_1215667894253830144)