# 怎样正确理解马尔科夫链？

[](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。

### 智能体和环境

首先需要明确，强化学习的任务是什么？

这用大白话说：就是希望用强化学习的方式，使某个**东东**获得**独立自主地完成某种任务**的能力。

而这个**东东，**称为**智能体**。而智能体学习和工作的地方，就称为**环境**。

注意！所谓独立自主，就是智能体一旦启动，就不需要人指挥了。

例如说扫地机器人打开开关，就不需要人类告诉机器人哪里有灰尘，哪里有垃圾，自己就会去清理了；

又例如：自动驾驶汽车在导航设置好之后，就不再需要司机去操作，汽车能够自己安全到达目的地。

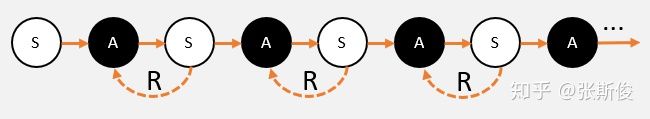
但应该怎样才能让智能体学会技能呢？

和研究其他的问题一样，首先需要把问题抽象为模型。才能在这个模型进行实验和探索后，再把结果运用到实际当中去。

**马尔科夫链**，就是对现实世界抽象的一个模型。因而马尔科夫链被广泛使用在各个领域，当然也包括讨论的强化学习领域。本专栏中提到的几乎所有的算法，都是基于马尔科夫链这个模型。必须了解问题是什么，才能更好解决问题。

### 马尔科夫链

看图说话，在强化学习中，马尔科夫链的经典表达图例是这样的。



马尔科夫链

在马尔科夫链中，有三个重要的元素：S，A，R。分别来看一下，他们代表的是什么。然后大家就会明白，为什么马尔科夫链是一个很好很常用的模型。

### **S（state）状态**，在图中用白色圈圈表示。

**状态**就是智能体**观察**到的当前环境的**部分或者全部特征**。

举个例子：无人驾驶汽车来到十字路口。和人类一样，它需要先“观察”这个路口的情况，再决定一下步的"动作"。

无人驾驶汽车通过摄像头，可以观察到交通标志、交通信号灯等情况；通过雷达，可以探测到与其他汽车、行人的距离；通过导航系统，了解前方的路段是否通畅等等。这些被观察到的环境特征，就是无人驾驶汽车的**状态特征**， **状态空间**就是智能体能够观察到的特征数量。

需要特别注意的是：环境的特征可能有许多，但只有智能体能够观察到的特征才算是状态。所以也用Observation（观察的英文）表示状态。

为什么要强调观察呢？

1. 这提醒要给智能体最有用的特征；因为在实际工作中，输入特征往往是很“贵”的，无人驾驶汽车的摄像头，雷达通常都是很昂贵的。而无用的特征，例如是否有乘客在唱歌之类的，输入到自动驾驶系统，这无疑加重了学习的负担。所以，必须非常慎重地选择状态特征。

2. 提醒要注意观察的角度。假设学有所成，希望做一个智能体学习如何打德州扑克。你就会突然发现，这个状态很微妙。因为对于每位玩家，都只看到自己的牌和公关牌，所以观察到的状态都是不一样的。

但在新手期，会把重点放在算法中，大家在实际工作中留个心眼就好。

### **A（action）动作**（用黑色圈圈表示）

动作其实不用解释，就是智能体做出的具体行为。例如扫地机器人会移动，吸尘，甚至喷水。无人驾驶汽车能够移动，加速，刹车，转弯等。

**动作空间**就是该智能体能够做出的动作数量。

举个例子：智能体身处十字路口。那么的方向就有4个。也就是说，能做的动作，就是4个。称能做的动作的集合，称为动作空间

### **R（reward）奖励**

当在某个状态下，完成动作。环境就会给反馈，告诉这个动作的效果如何。这种效果的数值表达，就是**奖励**。

其实这里的reward翻译为“反馈”可能更合适一点。因为反馈并不是完全正面的，也有负面。（但为了和其他文献统一，方便大家学习，这里仍然写作“奖励”） 当奖励可以是正数，表示鼓励当前的行为；如果是负数，表示惩罚这种行为。当然也可以是0。 而奖励值的大小，表示鼓励的和惩罚的力度不同。

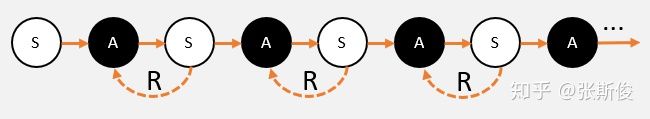
奖励在强化学习中有很关键的作用，会以奖励作为引导，让智能体学习做能获得**最多奖励**的动作。

例如：我需要训练机器人打乒乓球。机器人每次赢球，都可以加分；输球，就减分。这分数就表现了机器人的动作好坏。如果机器人希望获得更多的分数，就需要想办法赢球。

又例如：无人驾驶汽车如果成功到达目标地点，那么可以获得奖励；但如果闯红灯，那么就会被扣除大量的奖励作为惩罚。如果无人驾驶汽车希望获得更多的分数，那么就必须在遵守交通规则的情况下，成功到达目标地点。

敲黑板：奖励的设定是**主观**的，也就是说为了智能体更好地学习工作，自己定的。所以大家可以看到，很多时候会对奖励进行一定的修正，这也是加速智能体学习的方法之一。

现在来总结一下马尔科夫链，其中也包含了强化学习的一般步骤：



1. 智能体在环境中，观察到状态(S)；
2. 状态(S)被输入到智能体，智能体经过计算，选择动作(A);
3. 动作(A)使智能体进入另外一个状态(S1)，并返回奖励(R)给智能体。
4. 智能体根据返回，调整自己的策略。 重复以上步骤，一步一步创造马尔科夫链。

所以你看，强化学习跟教孩子是一个道理: 孩子做了好事，必须给奖励；孩子做错事了，必须惩罚。就这么简单！

### 马尔科夫‘链’？

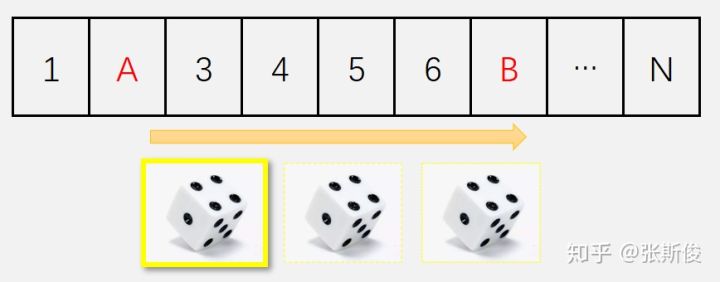
于是！于是就掉进第一个坑里了！ 在强化学习里，这根本不应该叫做‘**链**’而应该叫马尔科夫‘**树**’！

马尔科夫链之所以是现在看到的一条链条。是因为站在现在，往**后**看，所以是一条确定的路径。但如果往**前**看，就并不是一条路径，而是充满了各种“不确定性”。

这就像从家里到公司上班，中间有若干种上班的方式。现在你从家里出门，走过了两个路口，到了公交车站。 这时候往后看，从家到公交车站这一路，只能有一条路径。虽然你可以走其他路到公交车站，但这是你走过的路，已经确定下来了，所以路径只有一条；但如果往前看，从公交车坐车到公司，还有很多种方式到达，向前展开的是各种不确定性。

理解这一点很重要，这是理解强化学习理论的基础。为了更好理解，举个例子：

假设现在来玩这样一个游戏。这个游戏是简化版的大富翁，只考虑当前所处位置，也就是状态。智能体移动的时候，它可以选择投掷1-3个骰子，根据骰子点数的总和向前移动。

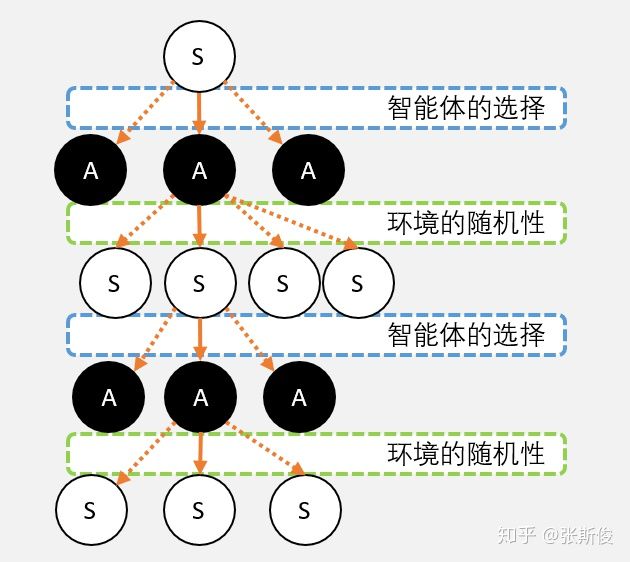


现在，智能体从格子A掷骰子，并移动到格子B。其实经历了两次不确定性。

第一次，是“选择”（策略）的过程。智能体主动选择骰子的个数。掷骰子的个数不同，到达格子B的概率也不同。所以“选择”会影响到下一个状态。这种不同动作之间的选择，称为智能体的**策略**。策略一般用**Pi**表示。的任务就是找到一个策略，能够获得最多的奖励。

第二次的不确定性，是环境的随机性，这是智能体无法控制的。在这个例子里就是骰子的随机性。注意，并不是所有环境都有随机性，有些环境是很确定的（例如把以上所有骰子每一面都涂成1点），但马尔科夫链允许有不确定性的存在。

从上面的例子知道，这种不确定性来自两个方面：1.智能体的行动选择（策略）。2.环境的不确定性。



马尔科夫树

所以，如果以后有人闯红灯，然后还大条道理地说：闯红灯又不一定会死，等绿灯也有可能死呀！ 这时候你要做的，除了离SB远点以外，心里还要清楚：闯红灯和等绿灯是我的选择，但不同选择后面的环境随机性给出的概率是不一样的。虽然我不能控制环境的随机性，但我能控制我的选择，让我避免高风险的低回报的情况出现。 类似地，总会有人说不读书也能成功的话。想想背后的逻辑，你将会更清楚自己的选择。

### 总结

1. 马尔科夫链是用来描述智能体和环境互动的过程。
2. 马尔科夫链包含三要素：state，action，reward

* state：只有智能体能够观察到的才是state。
* action：智能体的动作。
* reward：引导智能体工作学习的主观的值。
* 马尔科夫链的不确定性
* 核心思想：如果你不希望孩子有某种行为，那么当这种行为出现的时候就进行惩罚。如果你希望孩子坚持某种行为，那么就进行奖励。这也是整个强化学习的核心思想。

### 思考一下

了解了马尔科夫链是什么回事，那么就能更清楚了解他的适用范围了。 那么好，现在来想一下，是否能预测彩票号码呢？股票市场呢？如果可以，那么可以多大程度上预测呢？为了更准确的预测，可以往哪方面努力？

这跟教育孩子是一样的，如果你不希望孩子有某种行为，那么当这种行为出现的时候就进行惩罚。如果你希望孩子坚持某种行为，那么就进行奖励。这也是整个强化学习的核心思想。