# 15.强化学习

## 15.1.任务与奖赏

强化学习（Reinforcement Learning，简称RL）是机器学习的一个重要分支，人机大战的主角AlphaGo正是以强化学习为核心技术。在强化学习中，包含两种基本的元素：状态与动作，在某个状态下执行某种动作，这便是一种策略，学习器要做的就是通过不断地探索学习，从而获得一个好的策略。例如：在围棋中，一种落棋的局面就是一种状态，若能知道每种局面下的最优落子动作，那就攻无不克/百战不殆了。

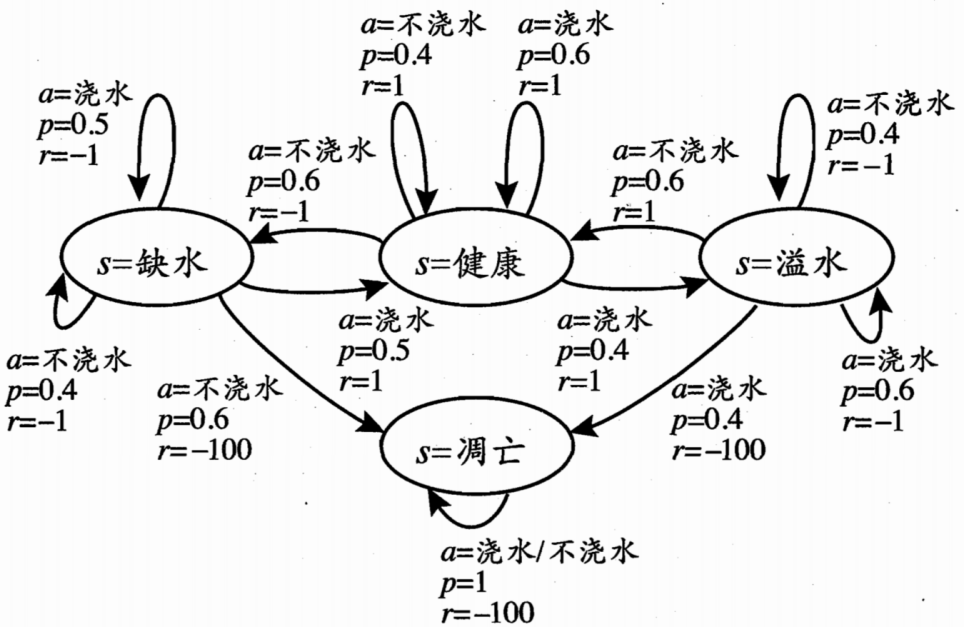
强化学习是在试图寻找一个映射，从已知状态推断出动作。但在实际问题中，强化学习通常都是在尝试动作后才能获得结果，因此强化学习是通过反馈的结果信息不断调整之前的策略，从而算法能够学习到：在什么样的状态下选择什么样的动作可以获得最好的结果。

例如在种瓜任务上，状态为当前瓜苗的长势，如健康、缺水、凋亡等；动作有浇水、施不同的肥、使用不同的农药等；如果瓜苗状态为缺水，若选择动作浇水，则瓜苗有一定的概率恢复健康，也有一定的概率无法恢复；定义如果保持瓜苗健康则奖赏+1，瓜苗凋零对应奖赏-10，最终种出了好瓜对应奖赏+100。这样就可以根据反馈信息学习到一个状态与动作的映射。

下面为一个给西瓜浇水的简单任务：

该任务只有四个状态（健康、缺水、溢水、凋亡）和两个动作（浇水、不浇水），在每一步转移后，若状态是保持瓜苗健康则获得奖赏1，瓜苗缺水或溢水奖赏为-1，这时通过浇水或不浇水可以恢复健康状态，当瓜苗凋亡时奖赏是最小值-100且无法恢复。

图中箭头表示状态转移，箭头旁的a，p，r分别表示导致状态转移的动作、转移概率以及返回的奖赏。



容易看出，最优策略在“健康”状态选择动作“浇水”、在“溢水”状态选择动作“不浇水”、在“缺水”状态选择动作“浇水”、在“凋亡”状态可选择任意动作。

一个策略的优劣取决于长期执行这一策略后的累积奖赏，换句话说：可以使用累积奖赏来评估策略的好坏，最优策略则表示在初始状态下一直执行该策略后，最后的累积奖赏值最高。

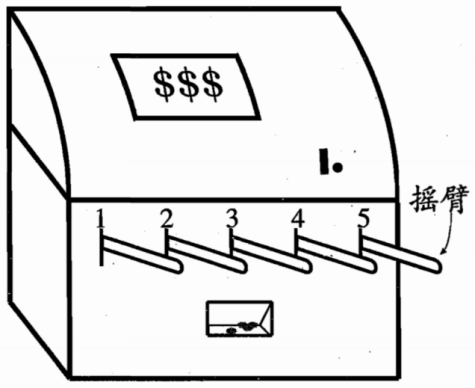
但是，通常没有人直接告诉机器在什么状态下应该做什么动作，只有等到最终结果揭晓，才能通过“反思”之前的动作是否正确来进行学习。因此，强化学习在某种意义上可看作具有“延迟标记信息”的监督学习问题。

## 15.2.k-摇臂赌博机

强化学习任务的最终奖赏是在多步动作之后才能观察到，这里我们不妨先考虑比较简单的情形：最大化单步奖赏，即仅考虑一步操作。需注意的是，即便在这样的简化情形下，强化学习仍与监督学习有显著不同，因为机器需通过尝试来发现各个动作产生的结果，而没有训练数据告诉机器应当做哪个动作。

欲最大化单步奖赏需考虑两个方面：一是需知道每个动作带来的奖赏，二是要执行奖赏最大的动作。若每个动作对应的奖赏是一个确定值，那么尝试一遍所有的动作便能找出奖赏最大的动作。然而，更一般的情形是，一个动作的奖赏值是来自于一个概率分布，仅通过一次尝试并不能确切地获得平均奖赏值。

“K-摇臂赌博机”（亦称K-摇臂老虎机）有K个摇臂，如图所示，每个摇臂以一定的概率吐出硬币，但这个概率赌徒并不知道。赌徒的目标是通过一定的策略最大化自己的奖赏，即获得最多的硬币。



若仅为获知每个摇臂的期望奖赏，则可采用“仅探索”法：将所有的尝试机会平均分配给每个摇臂（即轮流按下每个摇臂），最后以每个摇臂各自的平均吐币概率作为其奖赏期望的近似估计。

若仅为执行奖赏最大的动作，则可采用“仅利用”法：按下目前最优的（即到目前为止平均奖赏最大的）摇臂，若有多个摇臂同为最优，则从中随机选取一个。

显然，“仅探索”法能很好地估计每个摇臂的奖赏，却会失去很多选择最优摇臂的机会；“仅利用”法则相反，它没有很好地估计摇臂期望奖赏，很可能经常选不到最优摇臂。

“探索”和“利用”是矛盾的，因为尝试次数（即总投币数）有限，加强了一方则会自然削弱另一方，这就是强化学习所面临的“探索-利用窘境”。显然，欲累积奖赏最大，则必须在探索与利用之间达成较好的折中。

**-贪心法**

-贪心法基于一个概率来对探索和利用进行折中：每次尝试时，以的概率进行探索，即以均匀概率随机选取一个摇臂；以的概率进行利用，即选择当前平均奖赏最高的摇臂（若有多个，则随机选取一个）。

若摇臂奖赏的不确定性较大，例如概率分布较宽时，则需更多的探索，此时需要较大的值；若摇臂的不确定性较小，例如概率分布较集中时，则少量的尝试就能很好地近似真实奖赏，此时需要的较小。

**Softmax算法**

Softmax算法基于当前已知的摇臂平均奖赏来对探索和利用进行折中。

若各摇臂的平均奖赏相当，则选取各摇臂的概率也相当；若某些摇臂的平均奖赏明显高于其他摇臂，则它们被选取的概率也明显更高。概率公式为：



其中记录当前摇臂的平均奖赏，被称为温度，越小，概率的差距越大，越大，则概率的差距越小。