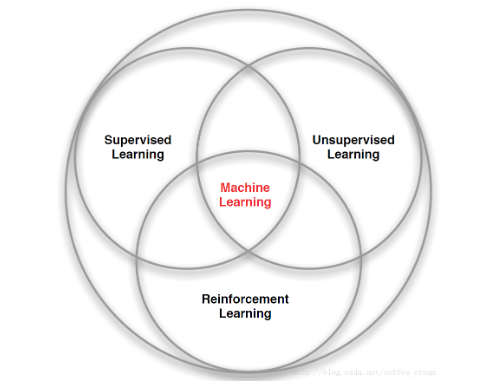
当前的机器学习算法可以分为3种：有监督的学习（Supervised Learning）、无监督的学习（Unsupervised Learning）和强化学习（Reinforcement Learning），结构图如下所示：



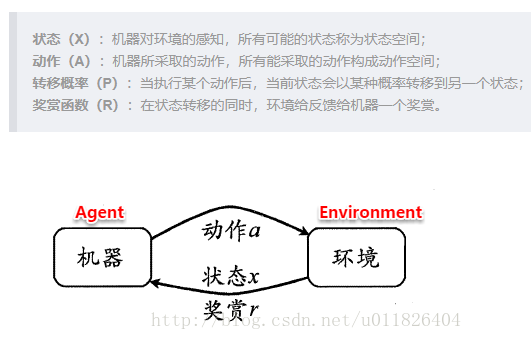
其他许多机器学习算法中学习器都是学得怎样做，而RL是在尝试的过程中学习到在特定的情境下选择哪种行动可以得到最大的回报。在很多场景中，当前的行动不仅会影响当前的rewards，还会影响之后的状态和一系列的rewards。RL最重要的3个特定在于：（1）基本是以一种闭环的形式；（2）不会直接指示选择哪种行动（actions）；（3）一系列的actions和奖励信号（reward signals）都会影响之后较长的时间。

**RL与有监督学习、无监督学习的比较：**   
　　（1）有监督的学习是从一个已经标记的训练集中进行学习，训练集中每一个样本的特征可以视为是对该situation的描述，而其label可以视为是应该执行的正确的action，但是有监督的学习不能学习交互的情景，因为在交互的问题中获得期望行为的样例是非常不实际的，agent只能从自己的经历（experience）中进行学习，而experience中采取的行为并不一定是最优的。这时利用RL就非常合适，因为RL不是利用正确的行为来指导，而是利用已有的训练信息来对行为进行评价。   
　　（2）因为RL利用的并不是采取正确行动的experience，从这一点来看和无监督的学习确实有点像，但是还是不一样的，无监督的学习的目的可以说是从一堆未标记样本中发现隐藏的结构，而RL的目的是最大化reward signal。   
　　（3）总的来说，RL与其他机器学习算法不同的地方在于：其中没有监督者，只有一个reward信号；反馈是延迟的，不是立即生成的；时间在RL中具有重要的意义；agent的行为会影响之后一系列的data。

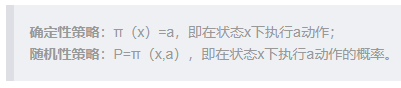
在强化学习中，包含两种基本的元素：状态与动作，在某个状态下执行某种动作，这便是一种策略，学习器要做的就是通过不断地探索学习，从而获得一个好的策略。例如：在围棋中，一种落棋的局面就是一种状态，若能知道每种局面下的最优落子动作，那就攻无不克/百战不殆了。

若将状态看作为属性，动作看作为标记，易知：监督学习和强化学习都是在试图寻找一个映射，从已知属性/状态推断出标记/动作，这样强化学习中的策略相当于监督学习中的分类/回归器。但在实际问题中，强化学习并没有监督学习那样的标记信息，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。

强化学习任务通常使用马尔可夫决策过程（Markov Decision Process，简称MDP）来描述，具体而言：机器处在一个环境中，每个状态为机器对当前环境的感知；机器只能通过动作来影响环境，当机器执行一个动作后，会使得环境按某种概率转移到另一个状态；同时，环境会根据潜在的奖赏函数反馈给机器一个奖赏。综合而言，强化学习主要包含四个要素：状态、动作、转移概率以及奖赏函数。



因此，**强化学习的主要任务就是通过在环境中不断地尝试，根据尝试获得的反馈信息调整策略，最终生成一个较好的策略π，机器根据这个策略便能知道在什么状态下应该执行什么动作**。常见的策略表示方法有以下两种：



**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。长期累积奖赏通常使用下述两种计算方法：

