1. 简介
2. Policy

策略是从当前环境状态到动作的一个映射。它以状态作为输入，决定下一步的动作。策略可以是确定的，也可以是随机的。

1. Reward

奖励定义强化学习中的问题。在每一个时间点，环境都会给代理（agent）返回一个数字，这个数字叫做奖励。代理唯一的目标是最大化奖励。同时奖励是影响策略选择的一个主要因素：当奖励较低时，代理会选择其他的动作。

1. Value function

价值函数是指从当前状态开始，代理可以获得的最大奖励的期望值。

The value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state.

有价值函数的目的是为了最大化奖励。奖励在一定程度上是主要的，价值函数是次要的。但是动作的选择与价值函数有关。

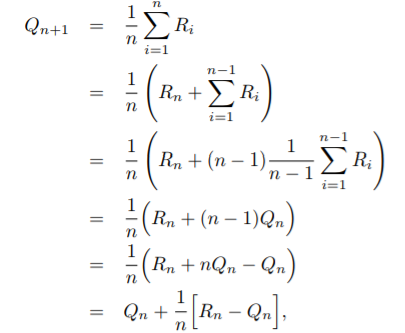
1. Model

对环境进行建模？

1. 多臂老虎机（Multi-armed Bandit）

由于每一个动作能提供的奖励是一个正态分布（是一个固定分布，不会改变），所以需要通过利用（exploitation）和探索（exploration）来估计出每一个动作所能提供奖励的均值sample-average method（大数定理）。同时在算法上采用greedy和两种方式进行探索。实验代码：<https://github.com/ShangtongZhang/reinforcement-learning-an-introduction/blob/master/chapter02/ten_armed_testbed.py>

在对每一个动作所能提供的奖励进行估计的时候，可以利用递推式的方式减少内存和计算量的使用：



1. 有限马尔可夫决策过程（Finite Markov Decision Processes）

MDP是一个经典的顺序决策过程，其当前动作不仅会影响当前的奖励也会影响后续的状态和之后的奖励。同时MDP包括延时奖励问题，必须权衡当前的奖励与延时奖励。在多臂老虎机中我们对动作的价值进行评估，而在MDP中我们对状态进行评估

1. 历史（history）是一系列的观察、动作和奖励组成的序列。
2. 状态（state）时代理（agent）用来决定下一步做什么的信息，是历史的一个简要概述。
3. 政策（Policy）:以状态作为输入，决定下一步的动作。

·A policy is the agent’s behavior.

·有确定性（Deterministic policy）和随机性（stochastic policy）

1. 评价函数（Value function）：在当前状态下或者是在当前动作下的好坏。

·Value function is a prediction of future reward.帮助选择action

1. 模型（Model）：agent感受环境的部分，这是agent眼中的环境。