蒙特卡洛方法

基本思想

用时间发生的频率代替时间发生的概率

蒙特卡洛策略评估

蒙特卡洛方法是基于采样的方法，给定策略π，让智能体与环境进行交互，得到多条轨迹，每个轨迹的回报如下：

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ...$$

求出所有轨迹回报的平均值，就可以知道一个策略对应状态的价值，如下：

$$V_\pi(s) = E_{r\sim\pi}[G_t|s_t = s]$$

蒙特卡罗仿真指的是我们可以采样大量的轨迹，计算所有轨迹的真实回报，然后计算均值。蒙特卡洛使用经验平均回报的方法来进行估计，不需要马尔可夫决策过程的转台转移函数和奖励函数。同时，不需要像动态规划那样使用自举的方法。

蒙特卡洛方法的局限性，它只能用在具有终止状态的马尔可夫决策过程中。