第一章　强化学习-机器人导航

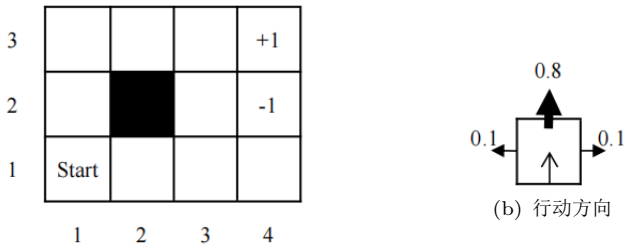
1.1　实验内容与任务

图1.1(a)是一个机器人导航问题的地图。机器人从起点Start出发，每一个时间点，它必须选择一个行动(上下左右)。在马尔可夫决策中实验中，机器人是根据环境模型中的转移矩阵 $P(x, a, x')$ 来进行价值函数和最优策略的计算，但是本次实验中，机器人并不知道这个转移矩阵。已知机器人行动之后，环境会告知机器人两件事情–新的实际位置以及到达新位置所得到的报酬。因此，如果机器人有一个策略$\pi(x)$，那么在与环境的交换中，机器人会具有这样一个数据序列：$(x_0, a_0, r_0, x_1, a_1, r_1, \ldots, a_{n-1}, r_{n-1}, x_n)$，其中$x_i, r_i$ 是环境告知的状态和该状态的报酬，$a_i = \pi(x_i)$是机器人自己的决策，x_0是Start, x_n 是一个终止状态。这个数据序列也称为一个样本路径。强化学习的任务是让机器人在环境中运行多次，得到多条样本路径，通过这些样本路径，来求解最优策略。

样本路径是与环境的交互中产生的，你先要实现一个环境模型。假设实际位置由环境按图1.1(b) 的方式决定：机器人每次移动的实际结果是机器人以0.8 的概率移向所选择方向，也可能是以0.1的概率移向垂直于所选方向。如果实际移动的方向上有障碍物，则机器人会停在原地。机器人移动到图中每个格子，会获得一个报酬, 图1.1(a) 中标有+1和-1 的格子中标记的就是该格子的报酬，其他格子的报酬是-0.04. 报酬会随着时间打折，假设折扣是1。

1.2　实验过程及要求

1. 实验环境要求：Windows/Linux操作系统，Python编译环境，numpy、scipy等程序库。



(a) 地图　(b) 行动方向

图 1.1: 机器人行动环境

2. 编写一个环境，它能跟机器人交互，主要提供行动的结果–下一步的状态及获得的报酬。

3. 已知机器人的策略$\pi(x)$, 通过与环境交互学习在该策略下的价值函数$U(x)$（或者叫效用函数）。

4. 机器的行动价值函数是$Q(x,a)$，且$\pi(x) = \arg\max Q(x,a)$是最优决策, 通过与环境交互学习这个行动价值函数$Q(x,a)$。

5. 已知机器人的策略$\pi(x)$, 用一个线性函数来逼近价值函数，通过与环境交互学习这个线性函数。

6. 撰写实验报告。

1.3　教学目标

1. 理解和掌握强化学习的原理。

2. 能够应用时序差分方法，计算状态的价值函数。

3. 能够应用时序差分方法，计算状态的行为价值函数，从而得到最优决策。

4. 能够提出一个新函数来逼近价值函数，并用梯度下降法来计算该新函数的参数。

5. 能够分析不同方案的优缺点，提高对复杂工程问题建模和分析的能力。

1.4 实验报告要求

实验报告需包含实验任务、实验平台、实验原理、实验步骤、实验数据记录、实验结果分析和实验结论等部分，特别是以下重点内容：

1. 建立机器人导航问题的马尔可夫决策模型，实现Env模块。
2. 用时序差分方法计算价值函数和行动价值函数。
3. 用线性函数逼近价值函数。
4. 利用环境模型，应用马尔可夫决策方法得到价值函数和最优决策，检验强化学习的结果。