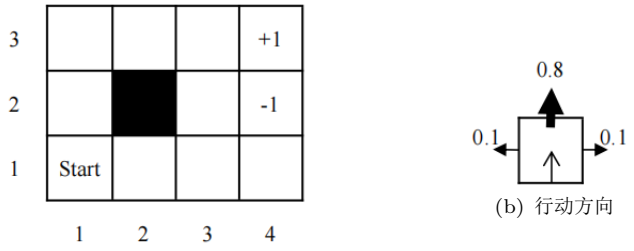
第一章　Markov决策-机器人导航

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，黑色格子是障碍物。机器人从起点Start出发进行连续移动，移动过程中机器人知道所在的格子。机器人每次移动一格，移动前必须在上下左右中选择一个方向，但是由于地板打滑的原因，实际移动的结果并不一定是在所选择的方向上。如图1.1(b) 所示，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，继续进行移动决策。如果机器人进入标有+1和-1的格子，则终止移动。机器人移动到图中每个格子，会获得一份报酬，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。用Markov决策的知识计算问题的价值函数，以及机器人的最佳策略。



(a) 地图　　(b) 行动方向

图 1.1: 机器人行动环境

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。
2. 建立环境模型，设置状态集、转移矩阵、报酬、折扣等环境参数。
3. 实现价值迭代算法，输出最优策略及其状态效用。
4. 实现策略迭代算法，输出最优策略及其状态效用。
5. 分别调整环境的报酬定义、折扣值、转移概率，比较相应的最优策略。
6. 撰写实验报告。

1.3　教学目标

1. 能够理解和掌握马尔可夫决策模型。
2. 能够应用价值迭代法和策略迭代法求解马尔可夫决策模型的最优策略及其状态效用。
3. 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　相关知识及背景

马尔可夫决策过程是Agent进行序列决策的模型。每次决策Agent会选取一个行动，该行动会改变状态，马尔可夫性规定Agent的状态变化只与前一时刻的状态相关，与更早的状态无关。Agent每进入一个状态会带来一定的报酬，因此序列决策会带来一个总报酬。为了获得多的总报酬，Agent需要采用最优的决策。

从不同的初始状态出发，获得的总报酬称为该初始状态的效用，状态的效用跟Agent的决策策略有关系。Bellman方程揭示了状态的报酬、最优策略、最优策略下的状态效用之间的关系。应用价值迭代法或者策略迭代法可以求解Bellman方程，获得最优策略及其状态效用。

表 1.1: 符号含义

符号	含义
n	状态个数，$n = 11$
X	状态集合，$X = 1, 2, \ldots, n$
x	一个状态，$x \in X$
A	行动集合，$A(x) = Right, Down, Left, Up$
a	一个行动，$a \in A$
P	三维的转移矩阵，$P(x, a, x')$定义机器人从状态x, 采取行动a后，转移到状态x'的概率$P(x'\|x, a)$
π	策略函数，$\pi(x)$是状态x应采取的行动
R	报酬函数，$R(x)$是状态x的回报，一个实数
γ	每个时间步上的折扣
$U^{\pi}(x)$	策略π下，状态x的效用
π^*	最优的策略

1.5　实验教学与指导

1.5.1　马尔可夫决策过程

模型的变量或者符号定义如表1.1。实验问题是一个马尔可夫过程，定义了状态集合X(初始状态s)，每个状态的行动集合A，转移模型$P(x, a, x')$，以及回报函数$R(x)$。目前这些元素在第1.1的描述中都已经给出，作为问题的已知条件。 $U^{\pi}(x)$是机器人从状态x出发，按照策略π，连续移动所获得的报酬总和的期望值：

$$U^{\pi}(x) = E(\sum_{t=0}^{t=+\infty} \gamma^t R(x_t)) \tag{1.1}$$

其中，$x_0 = x$。而最优策略则满足：

$$\pi^*(x) = \arg\max_{a \in A} \sum_{x'} P(x, a, x') U(x') \tag{1.2}$$

1.5.2　环境模型

```
class Env():
    def __init__(self,name):
        self.Name=name
        self.N=11
        self.A=np.arange(4) #{Right, Down, Left, Up}
        self.X=np.arange(self.N)
```

```
        self.makeP() #定义转移矩阵
        self.makeR() #定义报酬向量
        self.Gamma=1 #折扣
        self.StartState=0
        self.EndStates=[6,10]
```

根据第1.1的描述，转移矩阵P容易得到。如假设当前位置(1,1)格为状态0，(1,2)格为状态1，则根据图1.1(b)，机器人从状态0向右移动到达状态1的转移概率$P[0,0,1] = 0.8$，机器人从状态0向下移动到达状态1的转移$P[0,1,1] = 0.1$。

1.5.3　价值迭代算法

Bellman方程揭示了最优策略下，不同状态的效用之间的联系：

$$U(x) = R(x) + \gamma \max_{a \in A} \sum_{x'} P(x,a,x')U(x') \tag{1.3}$$

价值迭代算法根据Bellman方程，使用迭代算法来逼近最优策略下的状态效用

```
def ValueIter(E):
    U=np.zeros(E.N)
    U_=np.zeros(E.N)
    delta=1
    while delta>0.0001:
        U=np.copy(U_)
        U_=E.R+E.Gamma*np.max(np.dot(E.P[:,:,:],U),axis=1)
        delta = np.max(np.abs(U-U_))
    Pai=np.argmax(np.dot(E.P[:,:,:],U),axis=1)
    return U,Pai
```

1.5.4　策略评估

对一个策略π的评价可根据其产生的效用来进行:

$$U(x) = R(x) + \gamma \sum_{x'} P(x,\pi(x),x')U(x') \tag{1.4}$$

对所有的x，公式1.4代表了一个线性方程组，可以用代数方法求解。

```
def Eval(E,Pi):
    A=np.zeros((E.N,E.N))
    for i in range(E.N):
        A[i,:]=E.Gamma*E.P[i,Pi[i],:]
    A=A-np.identity(E.N)
    b=-E.R
    U=linalg.solve(A, b)
    return U
```

线性方程组的求解一般需要$O(N^3)$的计算时间。在状态数N过大时，方程组1.4也可以用一个迭代算法求解,时间为$O(k.N^2)$。

```
def Eval(E,Pi):
    U=np.zeros(E.N)
    k=20
    for k in range (k):
        for i in range(E.N):
            U[i]= E.R[i]+E.Gamma*np.dot(E.P[i,Pi[i],:],U)
    return U
```

1.5.5　策略迭代算法

价值迭代算法迭代到效用值收敛，一般需要较多的迭代次数。但最优策略可能比效用值更早收敛。策略迭代法通过更新策略来逼近最佳策略，如果某个行为a产生的效用值$U(x)$ 更好，则可以替代$\pi(x)$。这里需要使用第1.5.4节的策略评估函数。

```
def PolicyIter(E):
    Pai=np.zeros(E.N,dtype=np.int)          #初始策略
    change=True
    while change:
        U=Eval(E,Pai)          #计算该策略下的最大效用
        change=False
        for x in E.X:
            if np.max(np.dot(E.P[x,:,:],U))\
               >np.dot(E.P[x,Pai[x],:],U)+1E-5:
                Pai[x]=np.argmax(np.dot(E.P[x,:,:],U))
                change=True
    return U,Pai
```

1.6　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. 实现价值迭代算法和策略迭代算法。
3. 分析不同的报酬定义、折扣值、转移概率对最优策略的影响。
4. 比较价值迭代算法和策略迭代算法的优缺点。

1.7　考核要求与方法

实验总分100分，通过实验报告进行考核，标准如下：

1. 报告的规范性10分。报告中的术语、格式、图表、数据、公式、标注及参考文献是否符合规范要求。
2. 报告的严谨性40分。结构是否严谨，论述的层次是否清晰，逻辑是否合理，语言是否准确。
3. 实验的充分性50分。实验是否包含“实验报告要求”部分的4个重点内容，数据是否合理，是否有创新性成果或独立见解。

1.8　案例特色或创新

本实验的特色在于：通过对机器人导航问题的建模和求解，培养学生应用马尔可夫决策模型对时序问题进行决策和推理，要求学生能够完成价值迭代、策略迭代算法，并完成环境参数分析和比较。提高学生对复杂工程问题建模和分析的能力。