Artificial Intelligence

4. 状态空间决策与强化学习

罗晓鹏

xpluo@nju.edu.cn

工管 · 南京大学 · 2021 秋

目录

1. 状态空间搜索与序列决策

2. 状态价值迭代的直观例子

3. Bellman 方程与连续性运输方程

4. 状态价值迭代算法

5. 解的存在唯一性与迭代收敛性

6. 状态价值迭代简单实例

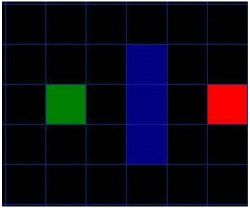
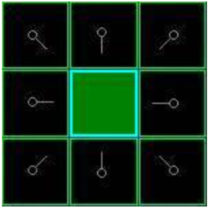
状态空间搜索与序列决策

- 状态空间模型
  - ▷ 状态 s 与状态集 $\{s\}$
  - ▷ 动作 a 与动作集 $\{a\}$
  - ▷ 目标与回报 $R(s, a)$
- 如何从“初始状态”转移到“目标状态”
  - ▷ (任意初始状态的)策略与最优策略
- 例：迷宫中的最短逃脱路径问题
  - ▷ 状态空间-迷宫地图，移动-动作
  - ▷ 逃脱点-目标状态
- 最优性原理、Bellman 方程与强化学习

状态价值迭代的直观例子

状态价值迭代算法的直观描述$_1$

- 最短路径查找问题

- 环境设定
  - ▷ 鹿群逃离迷宫，回归迷雾森林
  - ▷ 带透明自动门的格子迷宫
  - ▷ 强化学习状态价值迭代算法设置完成

状态价值迭代算法的直观描述$_2$

- 假设起点处鹿群的个体数量足够大
- “鹿群的逃离过程”等价于“训练过程”：
  - ▷ 初始的随机选择
  - ▷ 第一只小鹿达到逃离点
  - ▷ 逃离门开启，小鹿逃离，迷雾扩散，逃离门关闭
  - ▷ 第二只逃离的小鹿：两种情况
  - ▷ 迷雾反向扩散的概率与小鹿正向到达的几率
  - ▷ 逃离小鹿的逐渐增多
  - ▷ 迷宫格里的迷雾浓度近似地展示了最优策略

要点分析

- 环境：
  - ▷ 迷雾来源 ～ 迷雾森林
  - ▷ 迷雾反向扩散的存储 ～ 迷宫格
  - ▷ 迷雾反向扩散的原因 ～ 浓度差
- 训练：
  - ▷ 个体选择：迷雾浓度最大的迷宫格
  - ▷ 迷雾扩散：由个体路径选择引发的自动门开启
- 原理：
  - ▷ 迷雾的反向扩散过程
  - ▷ 个体的最大化选择

Bellman 方程与连续性运输方程

从直观例子导出 Bellman 方程

- 环境：
  - ▷ 迷雾来源 $\sim$ 奖励 $R(s,a)$
  - ▷ 迷雾反向扩散的存储 $\sim$ 状态值函数 $V(s,t)$
  - ▷ 迷雾反向扩散的原因 $\sim -\nabla_a V(s,t)$
- 迭代：
  - ▷ 个体选择：$a = \arg\max_{a\in\mathcal{A}(s)} \big(R(s,a) - \nabla_a V(s,t)\big)$
  - ▷ 迷雾扩散：$\frac{\Delta V(s,t)}{\Delta t} = \max_{a\in\mathcal{A}(s)} \big(R(s,a) - \nabla_a V(s,t)\big)$
- 令 $\Delta t \to 0$，得到连续型的 Bellman 方程：

$$\frac{\partial V(s,t)}{\partial t} = \max_{a\in\mathcal{A}(s)} \Big(R(s,a) - d(s,a)^{\mathrm{T}}\nabla_s V(s,t)\Big)$$

- 迷雾扩散所遵循的运输方程：$\frac{\partial}{\partial t}V(s,t) = -v^{\mathrm{T}}\nabla_s V(s,t)$

状态价值迭代算法

离散 Bellman 方程

- 离散的状态空间 $\mathcal{S} = \{s_i\}$
- 动作空间 $\mathcal{A} = \{a_i : s_i \to s_i'\}$
- 衰减因子 γ
- 确定性 Bellman 方程：
$$V(s) = \max_{a \in \mathcal{A}(s)} \Big(R(s) + \gamma V(s')\Big) = R(s) + \gamma \max_{a \in \mathcal{A}(s)} V(s')$$
- 概率性 Bellman 方程：
$$V(s) = R(s) + \gamma \max_{a \in \mathcal{A}(s)} \sum_{s'} P(s'|s,a) V(s')$$
- 状态价值迭代：
$$V_{i+1}(s) = R(s) + \gamma \max_{a \in \mathcal{A}(s)} \sum_{s'} P(s'|s,a) V_i(s')$$

状态价值迭代算法

Algorithm 2 状态价值迭代算法

定义状态空间 $\mathcal{S}$, 动作空间 $\mathcal{A}$, 回报 R, 衰减因子 γ.
for $i = 1 : n$ **do**
    for each $s \in \mathcal{S}$ **do**
        $V_{i+1}(s) = R(s) + \gamma \max_{a\in\mathcal{A}(s)} \sum_{s'} P(s'|s,a)V_i(s')$.
    end for
end for

- 状态价值迭代是 Bellman 方程对应的不动点迭代

解的存在唯一性与迭代收敛性

收敛性结论

定理 (状态价值迭代收敛性)

若 $\gamma < 1$，有限状态空间上 *Bellman* 方程的解存在且唯一，且相应状态价值迭代算法线性收敛于 *Bellman* 方程的唯一解.

Proof. (要点).

(0) Cauchy 序列与完备性

(1) 压缩映射不动点的存在唯一性

(2) 不动点迭代法线性收敛于压缩映射的不动点

(3) 极值不等式

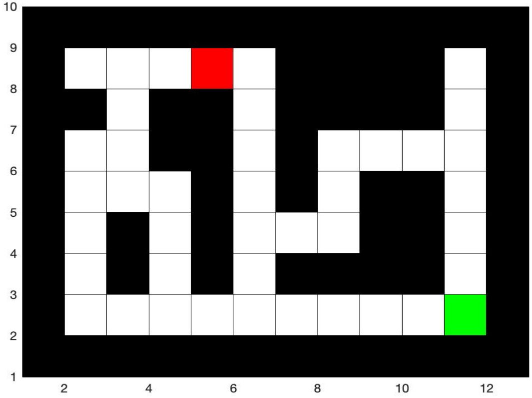
(4) 在题设条件下 Bellman 算子是压缩算子

(5) Bellman算子迭代在有限状态空间可实现

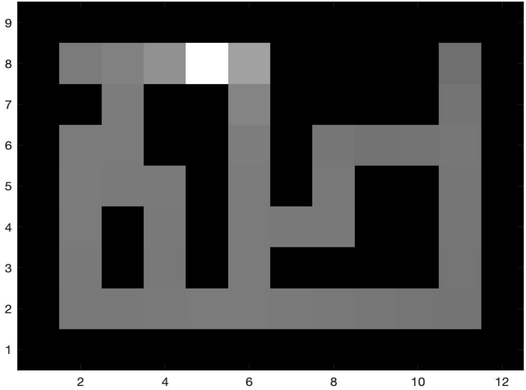
□

状态价值迭代简单实例

状态价值迭代实例



状态价值迭代实例



END