**问题的背景**

这里有一个会受到温度影响的智能体，我们假设它能够知道受温度的影响情况（如同人一样借助温度传感器）。此时它正处于一个极度寒冷的环境中，所幸前面有一个火堆，它将如何找到合适的距离来保持温度呢？（我们将状态空间和动作空间都做离散化处理以方便可视化）

**基于表格的Q学习算法**

这是一个有关状态和动作的表格，表格中间的数值为基于一定状态做出一定动作的最优价值（之后无论怎么移动，回报不会超过这个值），而最优价值与奖励的关系满足最优贝尔曼方程。

我们使用蒙特卡洛近似（一种经常用来估计pi值的方法）可以把左式近似为t时刻对最优价值做出的估计，右式近似为t+1时刻的估计（其中包含执行动作后的奖励）。

很明显右式更接近真实情况，所以我们鼓励估计用平滑的方式去接近

右式：



以上便是我们训练智能体的方法，智能体会依据衰减的ε-greedy策略（衰减的epsilon-greedy，即以epsilon的概率随机采样一个动作，反之根据价值表格中目前所处状态对应行中值最大的动作，并且epsilon的值会随着过程的进行不断缩小。如果不采用衰减的epsilon，表格将会无法收敛，而在部署时则会选用只参考价值网络的策略）来探索，并不断更新表格的值。

**定义环境和智能体**

源码见二维码对应的github仓库

**策略训练过程**

**证明有效性**

先将一个稍近的距离作为合适的状态，将奖励列表设置为如图

（在实际应用中，奖励应该和一些物理量直接关联起来。比如在这一背景中，可以用温度传感器来获取温度，再将不同的温度映射到不同的奖励）

（如果奖励列表设置的不合理，智能体的策略将难以收敛。奖励的设置一直是强化学习应用的一个关键点）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 距离 | 刚看到火堆 | 离火堆较远 | 一个稍远的距离 | 一个稍近的距离 | 一个更近的距离 | 一个过近的距离 | 零距离接触火堆 |
| 奖励 | -20 | -5 | -1 | 5 | -1 | -5 | -20 |

初始化智能体的价值表格



现在用嵌套的循环来模拟这一过程：外层循环表示初始化的100对智能体和环境，分别训练他们，以证明这一方法的稳定性；内层循环表示训练的过程（训练的过程持续20步，并得到最后的总奖励），在第一次循环中，我们会展示详细的采样，交互，反馈这一训练过程及价值表格更新的动态显示，而在后续的循环中，智能体会不断继承之前的价值表格来完成逐渐收敛的过程。

接下来是一个智能体训练的具体过程

第一次训练时

当前为第 0 步，当前状态： 0 ，选择的动作: -1 ，执行后的状态为: 0 ，执行获取的奖励为： -20 ，得分为： -20

更新后的价值表为：

[[ 0. -2. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]]

当前为第 1 步，当前状态： 0 ，选择的动作: 0 ，执行后的状态为: 0 ，执行获取的奖励为： -20 ，得分为： -40

更新后的价值表为：

[[ 0. -2. -2. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0.]]

当前为第 2 步，当前状态： 0 ，选择的动作: 2 ，执行后的状态为: 2 ，执行获取的奖励为： -1 ，得分为： -41

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 3 步，当前状态： 2 ，选择的动作: 2 ，执行后的状态为: 4 ，执行获取的奖励为： -1 ，得分为： -42

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 4 步，当前状态： 4 ，选择的动作: 0 ，执行后的状态为: 4 ，执行获取的奖励为： -1 ，得分为： -43

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. -0.1 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 5 步，当前状态： 4 ，选择的动作: -1 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： -38

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0.5 -0.1 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 6 步，当前状态： 3 ，选择的动作: 2 ，执行后的状态为: 5 ，执行获取的奖励为： -5 ，得分为： -43

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1]

[ 0. 0. 0. 0. -0.5]

[ 0. 0.5 -0.1 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 7 步，当前状态： 5 ，选择的动作: 0 ，执行后的状态为: 5 ，执行获取的奖励为： -5 ，得分为： -48

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1]

[ 0. 0. 0. 0. -0.5]

[ 0. 0.5 -0.1 0. 0. ]

[ 0. 0. -0.5 0. 0. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 8 步，当前状态： 5 ，选择的动作: 2 ，执行后的状态为: 6 ，执行获取的奖励为： -20 ，得分为： -68

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1]

[ 0. 0. 0. 0. -0.5]

[ 0. 0.5 -0.1 0. 0. ]

[ 0. 0. -0.5 0. -2. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 9 步，当前状态： 6 ，选择的动作: 0 ，执行后的状态为: 6 ，执行获取的奖励为： -20 ，得分为： -88

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1]

[ 0. 0. 0. 0. -0.5]

[ 0. 0.5 -0.1 0. 0. ]

[ 0. 0. -0.5 0. -2. ]

[ 0. 0. -2. 0. 0. ]]

当前为第 10 步，当前状态： 6 ，选择的动作: -1 ，执行后的状态为: 5 ，执行获取的奖励为： -5 ，得分为： -93

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1]

[ 0. 0. 0. 0. -0.5]

[ 0. 0.5 -0.1 0. 0. ]

[ 0. 0. -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

当前为第 11 步，当前状态： 5 ，选择的动作: -1 ，执行后的状态为: 4 ，执行获取的奖励为： -1 ，得分为： -94

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1 ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1 ]

[ 0. 0. 0. 0. -0.5 ]

[ 0. 0.5 -0.1 0. 0. ]

[ 0. -0.055 -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

当前为第 12 步，当前状态： 4 ，选择的动作: -1 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： -89

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1 ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1 ]

[ 0. 0. 0. 0. -0.5 ]

[ 0. 0.95 -0.1 0. 0. ]

[ 0. -0.055 -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

当前为第 13 步，当前状态： 3 ，选择的动作: -1 ，执行后的状态为: 2 ，执行获取的奖励为： -1 ，得分为： -90

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1 ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. -0.1 ]

[ 0. -0.1 0. 0. -0.5 ]

[ 0. 0.95 -0.1 0. 0. ]

[ 0. -0.055 -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

当前为第 14 步，当前状态： 2 ，选择的动作: -2 ，执行后的状态为: 0 ，执行获取的奖励为： -20 ，得分为： -110

更新后的价值表为：

[[ 0. -2. -2. 0. -0.1 ]

[ 0. 0. 0. 0. 0. ]

[-2. 0. 0. 0. -0.1 ]

[ 0. -0.1 0. 0. -0.5 ]

[ 0. 0.95 -0.1 0. 0. ]

[ 0. -0.055 -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

当前为第 15 步，当前状态： 0 ，选择的动作: -2 ，执行后的状态为: 0 ，执行获取的奖励为： -20 ，得分为： -130

更新后的价值表为：

[[-2. -2. -2. 0. -0.1 ]

[ 0. 0. 0. 0. 0. ]

[-2. 0. 0. 0. -0.1 ]

[ 0. -0.1 0. 0. -0.5 ]

[ 0. 0.95 -0.1 0. 0. ]

[ 0. -0.055 -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

当前为第 16 步，当前状态： 0 ，选择的动作: 1 ，执行后的状态为: 1 ，执行获取的奖励为： -5 ，得分为： -135

更新后的价值表为：

[[-2. -2. -2. -0.5 -0.1 ]

[ 0. 0. 0. 0. 0. ]

[-2. 0. 0. 0. -0.1 ]

[ 0. -0.1 0. 0. -0.5 ]

[ 0. 0.95 -0.1 0. 0. ]

[ 0. -0.055 -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

当前为第 17 步，当前状态： 1 ，选择的动作: 0 ，执行后的状态为: 1 ，执行获取的奖励为： -5 ，得分为： -140

更新后的价值表为：

[[-2. -2. -2. -0.5 -0.1 ]

[ 0. 0. -0.5 0. 0. ]

[-2. 0. 0. 0. -0.1 ]

[ 0. -0.1 0. 0. -0.5 ]

[ 0. 0.95 -0.1 0. 0. ]

[ 0. -0.055 -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

当前为第 18 步，当前状态： 1 ，选择的动作: 2 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： -135

更新后的价值表为：

[[-2. -2. -2. -0.5 -0.1 ]

[ 0. 0. -0.5 0. 0.5 ]

[-2. 0. 0. 0. -0.1 ]

[ 0. -0.1 0. 0. -0.5 ]

[ 0. 0.95 -0.1 0. 0. ]

[ 0. -0.055 -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

当前为第 19 步，当前状态： 3 ，选择的动作: 1 ，执行后的状态为: 4 ，执行获取的奖励为： -1 ，得分为： -136

更新后的价值表为：

[[-2. -2. -2. -0.5 -0.1 ]

[ 0. 0. -0.5 0. 0.5 ]

[-2. 0. 0. 0. -0.1 ]

[ 0. -0.1 0. -0.1 -0.5 ]

[ 0. 0.95 -0.1 0. 0. ]

[ 0. -0.055 -0.5 0. -2. ]

[ 0. -0.5 -2. 0. 0. ]]

可以发现此时还没有收敛

不断继承上一次训练的价值表格，再进行9999次训练后，最后一次训练过程如下：

当前为第 0 步，当前状态： 0 ，选择的动作: 2 ，执行后的状态为: 2 ，执行获取的奖励为： -1 ，得分为： -1

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.31036778e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.11825162e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 1 步，当前状态： 2 ，选择的动作: 1 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 4

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.11825162e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 2 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 9

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.13706911e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 3 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 14

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.15569842e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 4 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 19

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.17414143e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 5 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 24

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.19240002e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 6 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 29

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.21047602e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 7 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 34

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.22837126e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 8 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 39

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.24608754e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 9 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 44

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.26362667e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 10 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 49

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.28099040e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 11 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 54

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.29818050e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 12 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 59

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.31519869e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 13 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 64

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.33204671e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 14 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 69

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.34872624e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 15 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 74

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.36523898e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 16 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 79

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.38158659e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 17 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 84

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.39777072e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 18 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 89

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.41379301e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

当前为第 19 步，当前状态： 3 ，选择的动作: 0 ，执行后的状态为: 3 ，执行获取的奖励为： 5 ，得分为： 94

更新后的价值表为：

[[ 5.88165546e+00 5.88314980e+00 5.89130237e+00 2.47631106e+01

2.87746580e+01]

[ 0.00000000e+00 0.00000000e+00 -5.00000000e-01 2.88132607e+00

3.30893017e+01]

[-2.00000000e+00 0.00000000e+00 0.00000000e+00 3.30997364e+01

-1.00000000e-01]

[ 5.83935233e+00 5.38361093e+00 3.12241371e+01 4.51686282e+00

-1.41196788e+00]

[ 0.00000000e+00 1.85531020e+01 -1.00000000e-01 0.00000000e+00

0.00000000e+00]

[ 6.46175831e+00 -5.50000000e-02 -5.00000000e-01 -2.00000000e+00

-2.00000000e+00]

[-1.45000000e-02 -5.00000000e-01 -2.00000000e+00 -2.00000000e+00

-2.00000000e+00]]

可以看到，整个策略已经收敛，如果忽视epsilon-greedy策略的影响，直接将最后的价值表应用于实践，智能体将会采取如下行动：

初始处于状态0，观察这一行最大的值为第5列，所以执行动作：2，执行后的状态为2

现在处于状态2，观察这一行最大的值为第4列，所以执行动作：1，执行后的状态为3

现在处于状态3，观察这一行最大的值为第3列，所以执行动作：0，执行后的状态为3

不断重复上一步直到环境所规定的20步，这一策略将会积累最大的奖励和

**证明泛化性**

我们将环境做一下修改，将奖励列表设置为如图（即我们让火焰变得更大一些，使得状态2成为最合适的状态

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 距离 | 刚看到火堆 | 离火堆较远 | 一个稍远的距离 | 一个稍近的距离 | 一个更近的距离 | 一个过近的距离 | 零距离接触火堆 |
| 奖励 | -20 | -1 | 5 | -1 | -5 | -10 | -20 |

再次按照上面的方法训练智能体，同样展示第一次和最后一次训练的过程

当前为第 0 步，当前状态： 0 ，选择的动作: 2 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 5

更新后的价值表为：

[[0. 0. 0. 0. 0.5]

[0. 0. 0. 0. 0. ]

[0. 0. 0. 0. 0. ]

[0. 0. 0. 0. 0. ]

[0. 0. 0. 0. 0. ]

[0. 0. 0. 0. 0. ]

[0. 0. 0. 0. 0. ]]

当前为第 1 步，当前状态： 2 ，选择的动作: 1 ，执行后的状态为: 3 ，执行获取的奖励为： -1 ，得分为： 4

更新后的价值表为：

[[ 0. 0. 0. 0. 0.5]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 2 步，当前状态： 3 ，选择的动作: 2 ，执行后的状态为: 5 ，执行获取的奖励为： -10 ，得分为： -6

更新后的价值表为：

[[ 0. 0. 0. 0. 0.5]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 3 步，当前状态： 5 ，选择的动作: 0 ，执行后的状态为: 5 ，执行获取的奖励为： -10 ，得分为： -16

更新后的价值表为：

[[ 0. 0. 0. 0. 0.5]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. -1. 0. 0. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 4 步，当前状态： 5 ，选择的动作: 2 ，执行后的状态为: 6 ，执行获取的奖励为： -20 ，得分为： -36

更新后的价值表为：

[[ 0. 0. 0. 0. 0.5]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[ 0. 0. 0. 0. 0. ]]

当前为第 5 步，当前状态： 6 ，选择的动作: 0 ，执行后的状态为: 6 ，执行获取的奖励为： -20 ，得分为： -56

更新后的价值表为：

[[ 0. 0. 0. 0. 0.5]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[ 0. 0. -2. 0. 0. ]]

当前为第 6 步，当前状态： 6 ，选择的动作: -2 ，执行后的状态为: 4 ，执行获取的奖励为： -5 ，得分为： -61

更新后的价值表为：

[[ 0. 0. 0. 0. 0.5]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 7 步，当前状态： 4 ，选择的动作: 0 ，执行后的状态为: 4 ，执行获取的奖励为： -5 ，得分为： -66

更新后的价值表为：

[[ 0. 0. 0. 0. 0.5]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0. 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 8 步，当前状态： 4 ，选择的动作: -2 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： -61

更新后的价值表为：

[[ 0. 0. 0. 0. 0.5]

[ 0. 0. 0. 0. 0. ]

[ 0. 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 9 步，当前状态： 2 ，选择的动作: -2 ，执行后的状态为: 0 ，执行获取的奖励为： -20 ，得分为： -81

更新后的价值表为：

[[ 0. 0. 0. 0. 0.5 ]

[ 0. 0. 0. 0. 0. ]

[-1.955 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 10 步，当前状态： 0 ，选择的动作: -1 ，执行后的状态为: 0 ，执行获取的奖励为： -20 ，得分为： -101

更新后的价值表为：

[[ 0. -1.955 0. 0. 0.5 ]

[ 0. 0. 0. 0. 0. ]

[-1.955 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 11 步，当前状态： 0 ，选择的动作: 2 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： -96

更新后的价值表为：

[[ 0. -1.955 0. 0. 0.95 ]

[ 0. 0. 0. 0. 0. ]

[-1.955 0. 0. -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 12 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： -91

更新后的价值表为：

[[ 0. -1.955 0. 0. 0.95 ]

[ 0. 0. 0. 0. 0. ]

[-1.955 0. 0.5 -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 13 步，当前状态： 2 ，选择的动作: -2 ，执行后的状态为: 0 ，执行获取的奖励为： -20 ，得分为： -111

更新后的价值表为：

[[ 0. -1.955 0. 0. 0.95 ]

[ 0. 0. 0. 0. 0. ]

[-3.674 0. 0.5 -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 14 步，当前状态： 0 ，选择的动作: 1 ，执行后的状态为: 1 ，执行获取的奖励为： -1 ，得分为： -112

更新后的价值表为：

[[ 0. -1.955 0. -0.1 0.95 ]

[ 0. 0. 0. 0. 0. ]

[-3.674 0. 0.5 -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 15 步，当前状态： 1 ，选择的动作: 1 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： -107

更新后的价值表为：

[[ 0. -1.955 0. -0.1 0.95 ]

[ 0. 0. 0. 0.545 0. ]

[-3.674 0. 0.5 -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 16 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： -102

更新后的价值表为：

[[ 0. -1.955 0. -0.1 0.95 ]

[ 0. 0. 0. 0.545 0. ]

[-3.674 0. 0.995 -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 17 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： -97

更新后的价值表为：

[[ 0. -1.955 0. -0.1 0.95 ]

[ 0. 0. 0. 0.545 0. ]

[-3.674 0. 1.48505 -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 18 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： -92

更新后的价值表为：

[[ 0. -1.955 0. -0.1 0.95 ]

[ 0. 0. 0. 0.545 0. ]

[-3.674 0. 1.9701995 -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 19 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： -87

更新后的价值表为：

[[ 0. -1.955 0. -0.1 0.95 ]

[ 0. 0. 0. 0.545 0. ]

[-3.674 0. 2.27317955 -0.1 0. ]

[ 0. 0. 0. 0. -1. ]

[ 0.5 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

继承后不断训练9999次，最后一次：

当前为第 0 步，当前状态： 0 ，选择的动作: 2 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 5

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 31.67727347 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 1 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 10

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 31.86050073 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 2 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 15

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 32.04189573 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 3 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 20

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 32.22147677 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 4 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 25

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 32.399262 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 5 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 30

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 32.57526938 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 6 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 35

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 32.74951669 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 7 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 40

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 32.92202152 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 8 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 45

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 33.09280131 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 9 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 50

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 33.26187329 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 10 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 55

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 33.42925456 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 11 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 60

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 33.59496201 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 12 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 65

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 33.75901239 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 13 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 70

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 33.92142227 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 14 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 75

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 34.08220805 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 15 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 80

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 34.24138597 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 16 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 85

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 34.39897211 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 17 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 90

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 34.55498239 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 18 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 95

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 34.70943256 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

当前为第 19 步，当前状态： 2 ，选择的动作: 0 ，执行后的状态为: 2 ，执行获取的奖励为： 5 ，得分为： 100

更新后的价值表为：

[[10.17847514 10.19161017 10.19486559 29.18695558 33.54388891]

[ 0. 0. 0. 33.5427306 0. ]

[-1.05101335 10.03876014 31.73848931 2.0825882 -0.16188372]

[ 0. 14.33780525 0. 0. -1. ]

[12.52733409 0. -0.5 0. 0. ]

[ 0. 0. -1. 0. -2. ]

[-0.5 0. -2. 0. 0. ]]

如果将这一次的价值表应用于实践，智能体将会采取如下行动：

初始处于状态0，观察这一行最大的值为第5列，所以执行动作：2，执行后的状态为2

现在处于状态2，观察这一行最大的值为第3列，所以执行动作：0，执行后的状态为2

不断重复上一步直到环境所规定的20步，这一策略将会积累最大的奖励和

**总结**

在展示了两种情况下这一方法的表现，证明其泛化性后，让我们回到问题。如果我们能够直接得到温度和奖励对应的关系的话，是否一定有必要利用强化学习，以及强化学习到底智能在哪里。

首先我们能够想到的是，让智能体先移动一遍得到所有位置的奖励值，然后将智能体移动到奖励点最高的地方就是最好的方法。不过这其中涉及到一个问题——怎样移动是最好的，强化学习的答案是先2再1，如果运用动态规划的话，我们也能得到这样一个结果。事实上，强化学习的基础就一部分源自动态规划的研究。

那强化学习的亮点在哪呢？强化学习的核心是训练的数据来自于与环境交互的过程，进而能够解决连续动作空间，局部观测，多智能体，多任务等各系列的决策问题。而在我们这一背景中，这四个特点都没有，所以并没有展示出强化学习优势的地方，所以我们这一演示只能展示这一领域中最基础的q学习算法表格形式的训练过程，通过引入衰退的epsilon-greedy提高了训练速度和收敛速度。如果在强化学习算法的基础上再引入深度学习，就能使其处理更多维环境中的信息。