**國 立 中 山 大 學**

**機器學習**

第 一 次 作 業

GirdWorld – State-Value Function

老 師：黃國勝 教授

學 生：王佐軒

學 號：M023010014

系 別：電機所

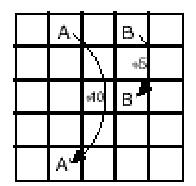
日 期：2013/11/05

1. **前言**

以agent的角度而言，State-value function可以計算每一個state的優劣程度， 而所謂的優劣程度則是定義為「未來能獲得的總獎賞(Reward)」，在Gridworld的實驗環境內，在Policy π下以Bellman equation推算出每個state的 value ，藉由改變疊代次數觀察數值的變化。

1. **實驗設計**

實驗環境：Gridworld



State s：空地(Reward = 0)、牆壁(Reward = -1)、state A(Reward = 10)、 state B(Reward = 5)

Action a：上下左右 (機率等同為1/4)

Discount rate γ = 90%

帶入Bellman公式計算

1. **模擬結果比較與分析 (請列出迭代1次5次10次結果(State-value function) )**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration times:1 | |  |  |  |
| -0.5 | 10 | -0.3 | 5 | -0.5 |
| -0.3 | 0 | 0 | 0 | -0.3 |
| -0.3 | 0 | 0 | 0 | -0.3 |
| -0.3 | 0 | 0 | 0 | -0.3 |
| -0.5 | -0.3 | -0.3 | -0.3 | -0.5 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration times:5 | |  |  |  |
| 3 | 9.3 | 4.3 | 5 | 1 |
| 1 | 2.7 | 1.9 | 1.5 | 0.1 |
| -0.3 | 0.5 | 0.4 | 0.2 | -0.6 |
| -0.9 | -0.4 | -0.2 | -0.4 | -0.9 |
| -1.5 | -0.9 | -0.8 | -1 | -1.5 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration times:10 | |  |  |  |
| 3.3 | 8.9 | 4.4 | 5.3 | 1.4 |
| 1.5 | 3 | 2.2 | 1.8 | 0.5 |
| 0 | 0.7 | 0.7 | 0.3 | -0.4 |
| -0.9 | -0.4 | -0.3 | -0.5 | -1.1 |
| -1.8 | -1.3 | -1.1 | -1.3 | -1.8 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration times:20 | |  |  |  |
| 3.3 | 8.8 | 4.4 | 5.3 | 1.5 |
| 1.5 | 3 | 2.3 | 1.9 | 0.6 |
| 0.1 | 0.8 | 0.7 | 0.4 | -0.4 |
| -1 | -0.4 | -0.3 | -0.6 | -1.2 |
| -1.8 | -1.3 | -1.2 | -1.4 | -2 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration times:40 | |  |  |  |
| 3.3 | 8.8 | 4.4 | 5.3 | 1.5 |
| 1.5 | 3 | 2.3 | 1.9 | 0.5 |
| 0.1 | 0.7 | 0.7 | 0.4 | -0.4 |
| -1 | -0.4 | -0.4 | -0.6 | -1.2 |
| -1.9 | -1.3 | -1.2 | -1.4 | -2 |

1. **結論**

根據上述實驗結果觀察，在State-value function疊代過10次左右後，其數值已經非常接近疊代至40次的數值，對於此小型的GridWorld學習的速度非常的快，藉由數值的分布明白地看出對於agent而言，處於每個state前進的方向為何。由驗證而得在特定Policy π下，Bellman Equation可以很簡單的將環境數值化。

**5. 程式原始碼**

