■ 專題報告:Gridworld 隨機策略下的 價值函數可視化系統

一、專題名稱

Gridworld 隨機策略的價值評估與視覺化分析系統

二、動機與背景

強化學習(Reinforcement Learning)是一種使智能體透過試誤方式學習最佳行動的技術。其中「價值函數」描述了智能體從某個狀態出發,根據特定策略所能獲得的期望報酬。在本專題中,我們以經典的 Gridworld 環境為例,探討在** 隨機策略(Random Policy)**下,如何計算並可視化各個狀態的價值函數,並透過網頁介面與互動操作使學習更直觀。

三、系統目標

- 實作 Gridworld 網格環境
- 使用固定的**隨機策略**(上下左右等機率)
- 計算各個狀態下的期望折扣報酬(即價值函數)
- 前端顯示價值表與隨機策略(四個方向箭頭)
- 支援動態調整網格大小、起點、終點、障礙物

四、系統架構

/project-root

| —— app.py # Flask 後端,負責策略評估與 API 處理 | —— templates/ | —— index.html # 前端主頁,含網格繪製與事件處理 | —— static/ | —— styles.css # 網格樣式與排版設計

• 後端框架: Flask

• **前端技術:** HTML、CSS、JavaScript (無需框架)

• **數值處理**: NumPy

五、核心方法

1. 隨機策略定義

每個狀態下,四個方向(↑↓←→)的選擇機率皆為 0.25。

2. 策略評估公式

使用以下公式對所有非終點與非障礙狀態反覆計算,直到收斂:

$$V(s) = \sum_{a \in A} \pi(a|s) \left[R(s,a) + \gamma V(s')
ight]$$

其中:

• R(s,a)=-1R(s,a) = -1R(s,a)=-1:每步懲罰

γ=0.9\gamma = 0.9γ=0.9: 折扣因子

• 終點與障礙物不參與計算

3. 視覺化策略

每個格子顯示隨機取樣的一個方向,表示智能體在此狀態下會隨機選擇四個方 向之一。

六、功能與操作介面

- 網格大小選擇:5x5~9x9
- 設定起點、終點、障礙物(點擊格子)
- 自動更新策略與價值函數
- 三個視圖:
 - 1. 初始網格(含起終點與障礙)
 - 2. 價值函數顯示(浮點數)
 - 3. 隨機策略方向(隨機取樣箭頭)

七、展示截圖

Gridworld Environment and Value evaluation

選擇網格大小 (5~9): 7~

請選擇設定模式:

○ 設定起點 (S) ○ 設定終點 (E) ® 設定障礙物 (X)

Þ	價值函數
X	價值函數

-9.96	-9.96	-9.94	-9.92	-9.90	-9.88	-9.86
-9.96	-9.95	-9.93	-9.90	-9.86	-9.82	-9.79
-9.94	-9.93	-9.89	-9.84	-9.76	-9.67	-9.60
-9.92	-9.90	-9.84	-9.73	-9.57	-9.35	-9.17
-9.90	-9.86	-9.76	-9.57	-9.23	-8.72	-8.19
-9.88	-9.82	-9.67	-9.35	-8.72	-7.56	-5.86
-9.86	-9.79	-9.60	-9.17	-8.19	-5.86	0.00

隨機策略

	1	1	→	1	1	1	1
	1	→	+	-	→	1	-
	-	+	Ť	→	→	1	1
	1	+	→	-	†	-	Ť
	-	1	†	1	1	→	→
	1	1	1	1	Ţ	1	+
	1	+	1	1	-	1	
-							

Gridworld Environment and Value evaluation

選擇網格大小 (5~9): 7~

請選擇設定模式: ○ 設定起點(S) ○ 設定終點(E) ⑩ 設定降顯物(X)

加松铜纹

TO ALIVOID						
s						
					х	
				x		
			×			
		х				
	×					
						E

海体系统

		1	貝田田安	•		
-9.99	-9.99	-9.99	-9.98	-9.97	-9.93	-9.86
-9.99	-9.99	-9.99	-9.98	-9.98	0.00	-9.74
-9.99	-9.99	-9.99	-9.99	0.00	-9.47	-9.50
-9.98	-9.98	-9.99	0.00	-9.29	-9.20	-9.08
-9.97	-9.98	0.00	-9.29	-9.06	-8.62	-8.12
-9.93	0.00	-9.47	-9.20	-8.62	-7.50	-5.83
-9.86	-9.74	-9.50	-9.08	-8.12	-5.83	0.00

隨機策略

				•		
1	1	1	1	1	1	-
-	→	†	+	+		†
1	1	1	→		†	→
-	1	1		+	→	-
-	1		1	+	-	-
-		Ţ	+	→	1	1
-	Ť	1	Ť	+	t	

八、實驗觀察與結果

- 價值函數數值呈現漸進遞減,終點附近的格子價值較高(懲罰少)
- 隨機策略下無法最短抵達終點,但平均距離可由價值觀察
- 增加障礙會影響周圍格子的值,顯示出環境變化敏感性

九、未來展望

- 支援更多策略(如 ε-greedy、手動策略)
- 加入路徑模擬(從某點模擬執行策略)
- 可視化策略機率(以箭頭透明度或長度表示機率)

十、結論

本系統透過策略評估實作與網頁視覺化,讓使用者能清楚理解在**非最佳策略 (隨機策略)**下,價值函數如何呈現各狀態的潛在長期回報。此專題不僅加 深了對強化學習基本概念的理解,也訓練了前後端整合與視覺設計的實作能 力。