利用強化學習訓練AI遊玩21點

108034555 蔡沛洹

背景

親朋好友們圍爐共度新春，是華人的傳統，而飯後不免俗的也會來點飯後娛樂，或是切磋牌技，小賭怡情。受此發想，此次練習希望使用強化學習，訓練AI學會玩21點，進而達到發大財的目標！

問題定義

What：要解決什麼問題？

解決打牌常常輸錢的問題

When：什麼時候進行？

準備大賺一筆前，比方說過年前

Who：由誰來參與？

打牌的新手或容易輸的玩家

Where：在何處進行？

電腦程式中

Why：為什麼要做這件事？

若訓練的好，可找到一最佳玩法，提升勝率

How：如何進行？

利用強化學習方法，並搭配參數的調整

As Is

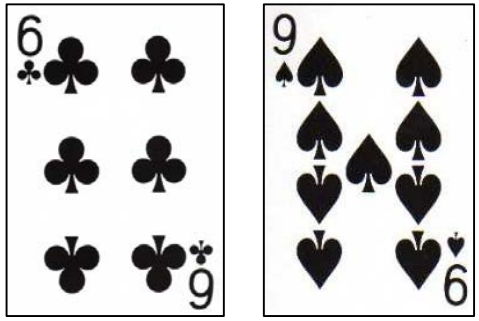
原本不知道怎麼玩21點，常常作出錯誤的決定導致賠錢

To Be

透過訓練AI，能夠找出各狀態下之最佳決策，提高勝率

玩法規則

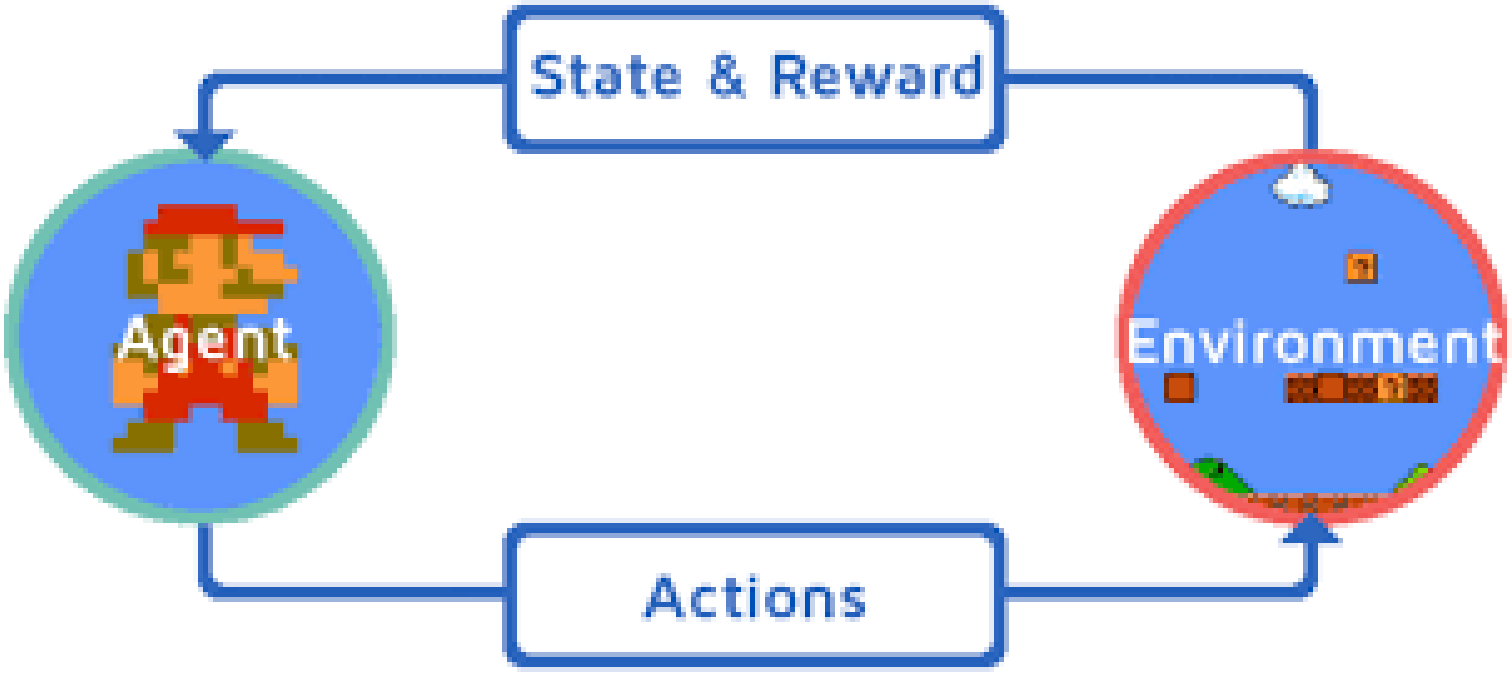
玩家



莊家

Hit： 要牌
Stand：不要牌

研究方法



研究方法

狀態(State)

包括玩家的點數總和及莊家的牌。舉例來說，State可能為(14, 5)，代表玩家點數加總為14，莊家明牌為5點。

決策(Action)

本練習假設較單純，玩家只有要牌和不要牌兩種決策

獎勵(Reward)

玩家獲勝reward為 +1，平手為0，而落敗reward為-1

研究方法

本練習中的強化學習是透過Q learning來進行，利用一Q table記錄各個state和action之Q值。初始時Q table為空的，每格Q值皆為零，此Q值會在每次做完決策後更新，最終便能得到一個最佳化後的Q table，也就是21點的遊玩策略。

Q值更新公式：

$$Q(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \left(\overbrace{\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}}}^{\text{learned value}} - \underbrace{Q(s_t, a_t)}_{\text{old value}} \right)$$

研究方法

α是學習率，α愈高代表比較相信當前這步的reward，愈低則代表比較相信舊的Q值。

γ是衰退率，愈高則代表考慮到許多步後，愈低則代表只考慮下一步。

Q值更新公式：

$$Q(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \left(\overbrace{\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}}}^{\text{learned value}} - \underbrace{Q(s_t, a_t)}_{\text{old value}} \right)$$

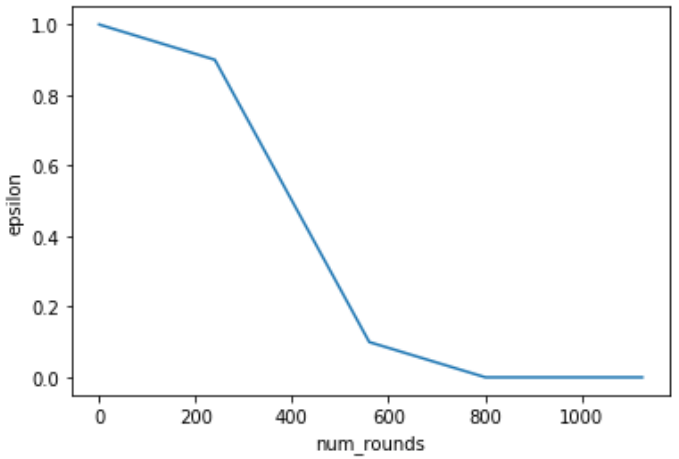
研究方法

參數調整：

參數調整主要調整α和γ這兩個參數，最終得到α為0.5，γ為0.2表現最佳。

探索率遞減：

在此練習中，採用探索率遞減，讓agent一開始可以充足探索環境，而隨著決策的回合數增加，agent探索環境愈來愈充足，探索率也逐漸下降，確保agent能夠開始增強學習，最後收斂至最佳的Q table值。



研究方法

```
(19, 10, True): {0: 0.5, 1: -0.04149999999999995},
(15, 10, False): {0: -1.0550000000000002, 1: -0.6816559075975512},
(20, 8, False): {0: 0.8528, 1: -0.5625},
(22, 8, False): {0: 0.0, 1: 0.0},
(18, 4, False): {0: 0.488, 1: -0.75},
(22, 4, False): {0: 0.0, 1: 0.0},
(14, 3, False): {0: 0.62, 1: -0.75},
(23, 3, False): {0: 0.0, 1: 0.0},
(13, 7, True): {0: -0.020000000000000018, 1: 0.0},
(16, 6, False): {0: -0.875, 1: -0.01999999999999997},
(20, 6, False): {0: 1.088, 1: 0.0},
(13, 10, False): {0: -0.889779763125, 1: -0.34343983928253813},
(17, 10, False): {0: -0.28174643712, 1: -0.91625},
(27, 10, False): {0: 0.0, 1: 0.0},
```

研究方法

Player's Hand	Dealer's upcard when ace is not usable										Dealer's upcard when ace is usable									
	['1',	'2',	'3',	'4',	'5',	'6',	'7',	'8',	'9',	'A']	['1',	'2',	'3',	'4',	'5',	'6',	'7',	'8',	'9',	'A']
1	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
2	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
3	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
4	['H',	'H',	'S',	'S',	'H',	'H',	'H',	'H',	'S',	'H']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
5	['H',	'H',	'H',	'H',	'H',	'H',	'H',	'S',	'S',	'H']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
6	['H',	'S',	'S',	'H',	'H',	'H',	'H',	'S',	'H',	'H']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
7	['S',	'S',	'H',	'H',	'H',	'H',	'H',	'H',	'S',	'H']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
8	['S',	'H',	'S',	'S',	'S',	'H',	'H',	'S',	'H',	'S']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
9	['H',	'H',	'H',	'H',	'H',	'S',	'H',	'H',	'S',	'H']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
10	['S',	'H',	'H',	'S',	'H',	'S',	'S',	'H',	'H',	'H']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
11	['H',	'H',	'H',	'H',	'H',	'H',	'H',	'S',	'S',	'H']	['-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-',	'-']
12	['H',	'H',	'S',	'S',	'H',	'S',	'S',	'S',	'H',	'H']	['H',	'S',	'S',	'S',	'H',	'S',	'H',	'S',	'H',	'H']
13	['H',	'S',	'H',	'S',	'H',	'H',	'H',	'S',	'H',	'H']	['H',	'H',	'H',	'H',	'H',	'H',	'H',	'H',	'H',	'H']
14	['S',	'H',	'H',	'H',	'H',	'S',	'H',	'H',	'H',	'H']	['H',	'H',	'S',	'S',	'S',	'S',	'H',	'H',	'H',	'S']
15	['H',	'H',	'H',	'S',	'H',	'S',	'H',	'H',	'H',	'S']	['S',	'H',	'S',	'S',	'H',	'S',	'H',	'S',	'H',	'H']
16	['S',	'S',	'H',	'H',	'S',	'H',	'S',	'H',	'S',	'S']	['S',	'S',	'H',	'S',	'S',	'H',	'S',	'S',	'S',	'H']
17	['S',	'S',	'S',	'S',	'S',	'S',	'S',	'H',	'H',	'S']	['H',	'S',	'H',	'S',	'S',	'H',	'H',	'S',	'H',	'S']
18	['H',	'S',	'S',	'S',	'S',	'S',	'S',	'H',	'H',	'S']	['S',	'H',	'S',	'H',	'H',	'S',	'S',	'H',	'S',	'H']
19	['S',	'H',	'S',	'S',	'S',	'S',	'S',	'S',	'S',	'S']	['S',	'H',	'S',	'H',	'H',	'S',	'S',	'S',	'S',	'S']
20	['S',	'S',	'S',	'S',	'S',	'S',	'S',	'S',	'S',	'S']	['H',	'H',	'H',	'S',	'S',	'S',	'S',	'H',	'S',	'H']
21	['S',	'S',	'S',	'H',	'S',	'S',	'S',	'S',	'S',	'S']	['S',	'H',	'S',	'H',	'H',	'S',	'S',	'S',	'S',	'S']

Average payout after 1000 rounds is -140.6

結果分析

為了檢驗此project訓練出的Q table表現如何，找兩組對照組一同比較，分別為隨機策略和基本策略。

隨機策略：

每次選擇策略時，皆隨機挑選action，即50%的機率選擇要牌，50%的機率選擇不要牌。

基本策略：

一般在玩21點時，其實是可以查表的，此表是透過數學和統計基礎得來，表上會列出手牌如何時，應該選擇要牌或是選擇不要牌。

結果分析

基本策略

Player's Hand	Dealer's Upcard									
	2	3	4	5	6	7	8	9	10	A
12	H	H	S	S	S	H	H	H	H	H
13	S	S	S	S	S	H	H	H	H	H
14	S	S	S	S	S	H	H	H	H	H
15	S	S	S	S	S	H	H	H	H	H
16	S	S	S	S	S	H	H	H	H	H
A2	H	H	H	D	D	H	H	H	H	H
A3	H	H	H	D	D	H	H	H	H	H
A4	H	H	D	D	D	H	H	H	H	H

結果分析

分別對強化學習模型和兩組對照組進行1000次的牌局模擬

結果如下表所示。觀察發現強化學習策略雖然已較隨機策略優異，但仍和基本策略有段落差，有改善的空間。

	隨機策略	基本策略	強化學習策略
報酬	-400	-100	-140

程式輸出呈現

```
玩家點數總和:  14
莊家的牌:  3
決策:  1
玩家點數總和:  20
莊家的牌:  3
決策:  0
You win!

玩家的牌:  [5, 9, 6]
莊家的牌:  [3, 10, 9]
```

```
玩家點數總和:  20
莊家的牌:  2
決策:  0
Draw

玩家的牌:  [10, 10]
莊家的牌:  [2, 8, 10]
```

後續研究方向

- 真實玩法較為複雜，除了選擇要牌或是不要牌，還有double和split的決策，應納入考量
- 除了使用強化學習外，也可以考慮使用神經網路來訓練
- 除了21點外，可進一步訓練麻將等更複雜的遊戲