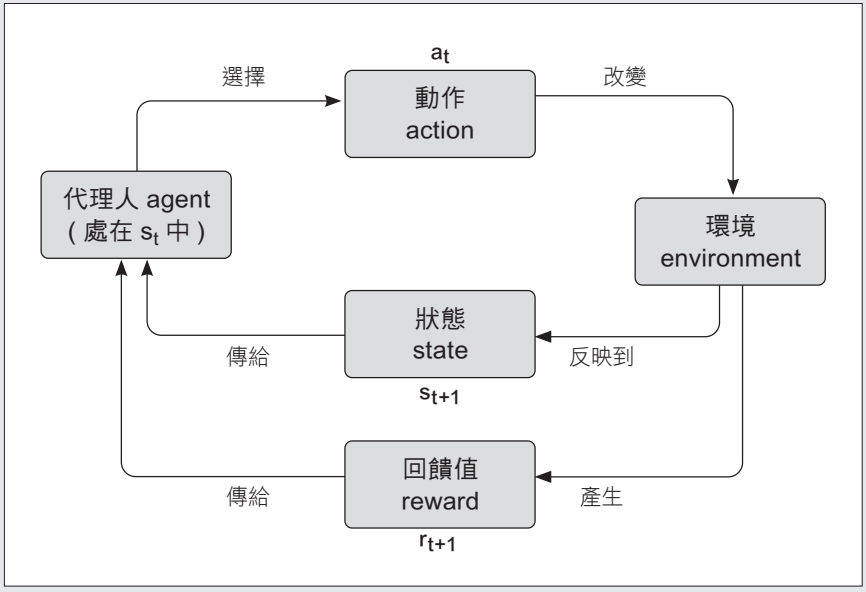
## 強化式學習之基礎架構

強化式學習中有幾個重要元素，即**代理人**（agent）、**動作**（action）、**環境**（environment）、**狀態**（state）、**回饋值**（reward），這幾個元素的關係可用下圖來表示：




強化式學習之基礎架構

## 強化式學習中之常見函數

| | |
|---|---|
| 策略函數 | 給定**狀態**及**策略**，該函數會傳回『各動作為最佳動作的**機率分佈**』，**如 softmax()** |
| 狀態價值函數 | 給定**狀態**及**策略**，會傳回『**該狀態的期望回饋值**』 |
| 動作價值函數 | 給定**狀態**及**策略**，會傳回『**各動作的期望回饋值**』 |

## 價值函數

價值函數分為兩類：

(i) 動作價值函數（也稱 Q 函數）

給定某個**狀態**及**動作**，該函數可傳回該動作的**價值**（即預計可得到的回饋值）。其中，**Q-Learning** 是實現動作價值函數的一種演算法。有了各動作的價值，我們便可遵循某種策略（如：$\varepsilon$ - 貪婪策略）來選擇動作。

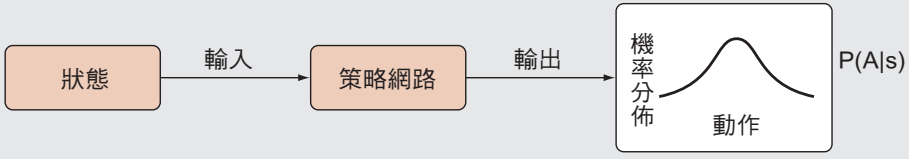

(ii) 狀態價值函數

給定特定**狀態**，該函數可傳回該**狀態的價值**（即該狀態開始至遊戲結束前，每一步期望回饋值之加權總和）。

$$V^{\pi}(s_0) = E^{\pi}[\sum_{i=0}^{t} w_i r_{i+1} | s_i] = w_0 E[r_1|s_0] + w_1 E[r_2|s_1] + \ldots + w_t E[r_{t+1}|s_t]$$

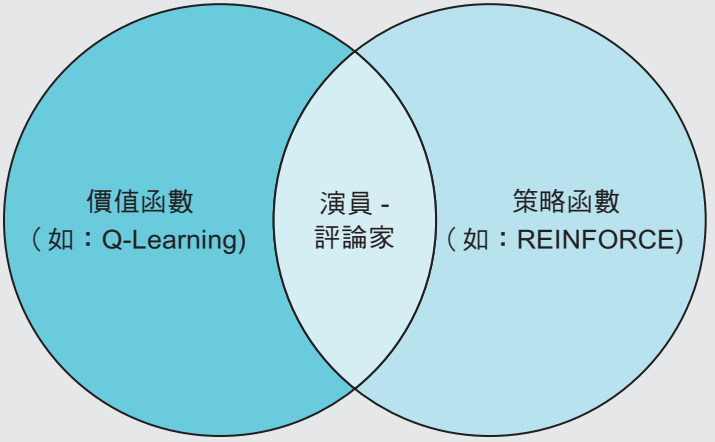
## 策略函數

給定某個**狀態**，該函數可傳回**各動作的機率分佈**，接著便可直接依照該機率分佈來隨機選擇動作。



## 演員 - 評論家

價值網路（DQN）及策略網路各有其優缺點，故可將它們結合在一起成為策略 - 價值演算法，即**演員 - 評論家**（Actor-Critic）。每當演員（策略網路）根據輸出的機率分佈隨機選擇一個動作，評論家（狀態價值網路）便會給予該動作評價（預測各動作的價值，並與真實收到的回饋值做比較），進而更新神經網路模型的參數。




在閱讀本書的第二部分前，讀者請務必搞懂以上內容，以便更好的銜接接下來的章節。