

Can AI Predict  
Stock Price?



# 「當AI遇上股票」 課程經驗分享

顏永進 / 20240222

from zero to error hero

# 大綱

1. FAQ
2. 課程背景
3. 課程發展
4. 課程簡介

# 1. FAQ

- Q1: 模型準確嗎(能賺錢嗎)?
- A1: 我還沒退休...QQ



資訊科技學科中心 (42)

會賺到錢嗎? 上午 9:46

好問題 上午 9:46

感覺應該要參加, 才能致富 上午 9:46

致富?? 上午 9:46

我有錢了, 我再請大家喝飲料吃披薩

輸入訊息

# 1. FAQ

- Q2: 不準還弄這個主題幹嘛?
- A2:
  1. 純粹希望學生能對AI與理財提高一點興趣，能找到AI在生活中可以帶來收益的實際應用(書中自有黃金屋)
  2. 學校教育太過缺乏理財知識
  3. 這是我學AI接觸到時間序列後第一個有強烈學習與實作動機的主題
  4. 不能只有我一個被騙...-\_-

## 2. 課程背景

- 1) 學校資訊相關課程
- 2) 授課對象與課程目標

# 1) 學校的資訊相關課程

2. 課程背景

年級	類別	學分	課程
一年級	必修	2	C++程式設計
一年級	選修	2	基礎程式設計
二年級	選修	2	APCS 與演算法
二年級	選修	2	資安實務
二年級	(上)選修	2	Python
二年級	(下)選修	2	AI選修
三年級	必修(電資學群)	4	進階程式設計(Python)。

## 先備知識(上學期PYTHON)

- 希望能具備的能力
  - 熟悉Python基本語法
  - 變數、輸入、輸出
  - 條件判斷、迴圈
  - 熟悉Python特殊的資料型別
  - Function
  - 第三方模組的安裝/管理
  - Numpy, Pandas, Matplotlib
  - 簡單的專題實作

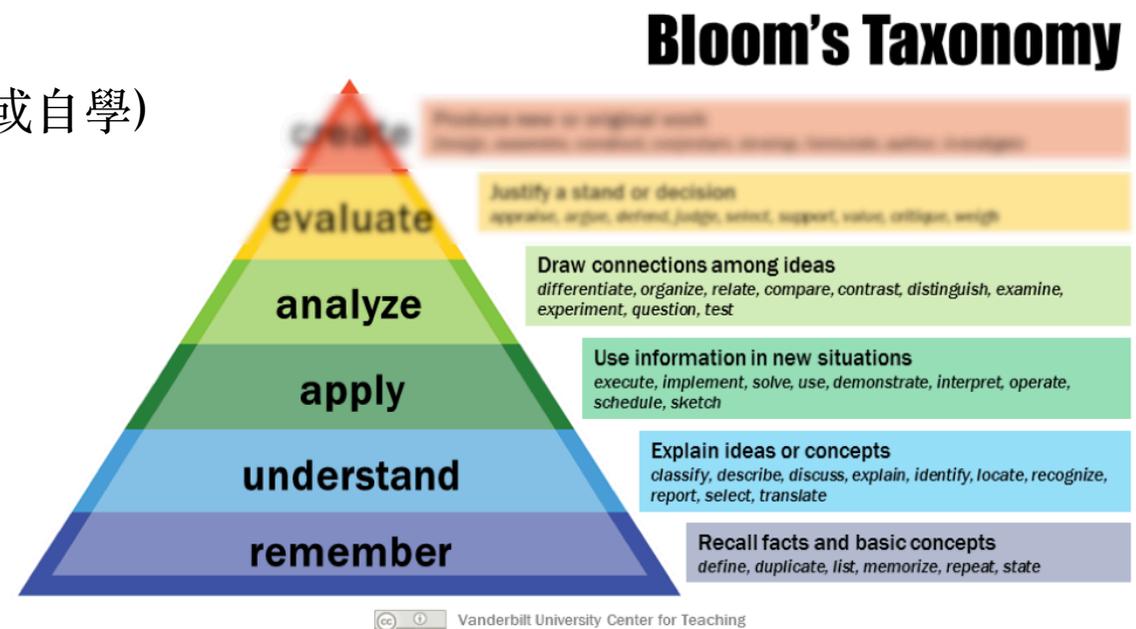
# 選修：PYTHON

			週五	
			307	多元選修
			1. 課程簡介 2. 變數、輸入與輸出	1. 課程簡介 2. IDE / 輸入輸出
中秋節				
			3. 運算式與運算子	3. 運算式與運算子
			4. Python內建函數	4. Python內建函數
			5. 判斷結構	5. 判斷結構
			6. 迴圈	6. 迴圈
			7. List / While	7. List / While
			8. 巢狀迴圈/List	9. 資料型別
			高三資訊科期中考	高二選修期中考
			8. 巢狀迴圈/List	10. Function
			9. 資料型別	12. Numpy / Matplotlib
			9. 資料型別	13. Pandas / Matplotlib
全校第2次定期考查				
			10. Function	期末Project
校慶運動會				
			高三資訊科期中考	期末Project
第3次定期考查				
				期末Project
				期末Project DEMO/onLine
			休業式	

## 2)授課對象與課程目標

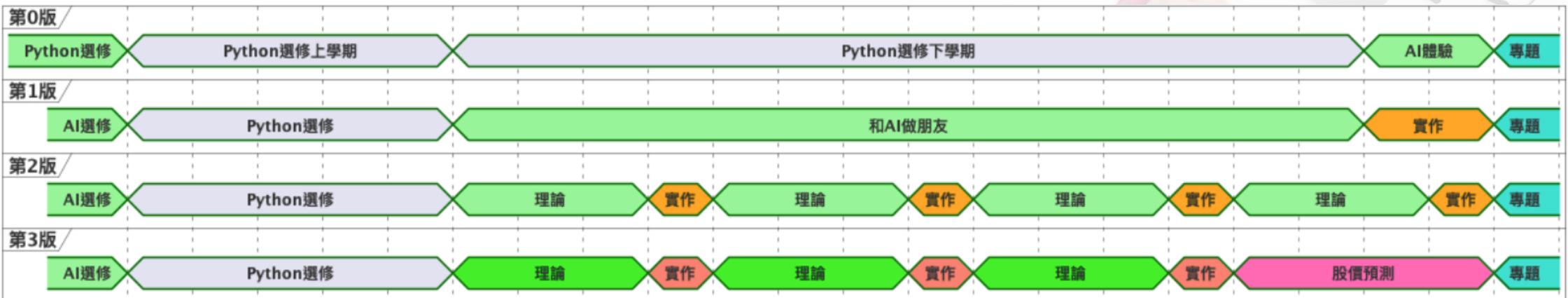
### 2. 課程背景

- 課程名稱: AI入門
- 對象: 高二下 (已修高二上Python選修或自學)
- 目的:
  1. 理解基本AI運作原理
  2. 進行資料預處理
  3. 挑選合適模型解決問題
  4. 評估效能
  5. 略能解釋模型

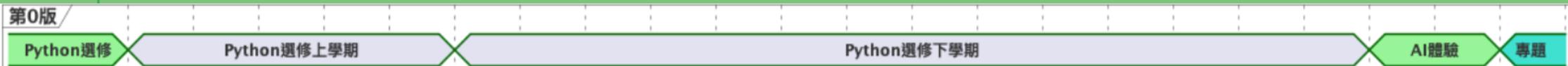


# 3. 課程發展

- 1) 版本0: Python選修附加活動
- 2) 版本1: 選修開設(教科書)
- 3) 版本2: 增加實作比重
- 4) 版本3: 加入股價預測活動



# 3. 課程發展



## 版本0

- Python選修附加活動
- AI概念
- AI模型體驗



## Say Hi to CNN



### TABLE OF CONTENTS

- |   |   |  |
|---|---|--|
| 1. READ THIS  | 2. 神經網路解題步驟   | 3. 收集資料<br>3.1. 資料類型<br>3.2. DEMO          |
| 4. 準備數據 (PREPARING THAT DATA)<br>4.1. 特徵 (FEATURE)<br>4.2. 資料正規化<br>4.3. 常態化<br>4.4. 標準化<br>4.5. DEMO | 5. 選擇模型 (CHOOSING A MODEL)<br>5.1. 語法<br>5.2. DEMO                      | 6. 訓練機器 (TRAINING)<br>6.1. 語法<br>6.2. DEMO |
| 7. 評估分析 (EVALUATION)  | 8. 調整參數 (HYPERPARAMETER TUNING)<br>8.1. MODEL參數<br>8.2. HYPERPARAMETERS | 9. 預測推論 (PREDICTION)                       |

# 3. 課程發展

第1版

AI選修

Python選修

和AI做朋友

實作

專題

## 版本1

- 二下選修
- 以教科書為主軸
- 期末加入實作



# 3. 課程

第1版

AI選修

Python選修

<b>第一章 人工智慧簡介</b>	<b>03</b>
1-1 思考如何避免思考	03
1-2 人工智慧能聰明到什麼程度	03
1-3 人工智慧的繁花盛開-人工智慧廣為應用於生活上的各領域	04
1-4 起伏不斷的人工智慧發展歷程	06
1-5 人工智慧的強弱之分	09
1-6 人工智慧與遊戲	10
1-7 人工智慧與機器學習	12
<b>第二章 背景知識</b>	<b>15</b>
2-1 資料收集	15
2-2 資料整理與儲存	18
2-3 特徵選擇	21
2-4 特徵距離的計算	23
2-5 資料的標準化	26
2-6 資料集分割	28
2-7 樹狀結構	31
<b>第三章 搜尋演算法</b>	<b>49</b>
3-1 搜尋是什麼？	49
3-2 如何搜尋？	49
3-3 當資料變得複雜時怎麼辦？	51
3-4 運用搜尋演算法處理複雜問題	55
3-5 推銷員問題	60
<b>第四章 資料探勘</b>	<b>64</b>
4-1 什麼是「資料探勘」？	64
4-2 資料探勘的方法-探索知識的程序	65

朋友

4-3 資料探勘的相關名詞	67
4-4 Apriori演算法-探勘常見項目集	68
4-5 關聯式法則	73
4-6 循序樣式探勘	75
4-7 SPAM演算法	76
<b>第五章 監督式學習</b>	<b>85</b>
5-1 監督式學習簡介	85
5-2 最短距離分類器	88
5-3 KNN分類器	91
5-4 決策樹	93
5-5 隨機森林	106
<b>第六章 非監督式學習</b>	<b>110</b>
6-1 非監督式學習簡介	110
6-2 K-means演算法	111
6-3 階層式分群法	117
<b>第七章 增強式學習</b>	<b>126</b>
7-1 增強式學習簡介	126
7-2 由代理人例子來了解如何進行最佳行動	127
7-3 Q-學習	138
<b>第八章 深度學習</b>	<b>146</b>
8-1 類神經網路	146
8-2 深度學習簡介	154
8-3 如何將待解問題數學化	156
8-4 卷積神經網路	159

專題



# 3. 課程發展

第1版

AI選修

Python選修

和AI做朋友

實作

專題



## 第五章 監督式學習

對應新課綱內容：十年級：演算法 | 資料表示、處理及分析

### 一、教案概述

科目 / 領域別	資訊科技 / 科技	單元名稱	監督式學習 Supervised Learning
教學對象	高二 / 高三	教學時數	4 節課
教學設備	電腦、投影機、廣播系統		
單元摘要	《特徵攻防戰》藉由圖片辨識，引領學生認識特徵值、分類器基本觀念，進而瞭解三種分類器（最近距離分類器、KNN 分類器、決策樹分類器）。在習得決策樹的概念後，再帶入「隨機森林」的概念，藉由個別決策樹的分類結果取得「眾數」來呈現其最終的分類結果。		
教學目標	<ol style="list-style-type: none"> <li>1. 學生能理解何謂人工智慧的監督式學習。</li> <li>2. 學生能理解特徵值觀念。</li> <li>3. 學生能理解運用特徵值進行分類的觀念。</li> <li>4. 學生能了解最近距離分類器、KNN 分類器、決策樹分類器之運作原理與相關應用。</li> <li>5. 學生能在瞭解決策樹的概念後，將其推展至隨機森林，瞭解其運作原理及應用。</li> </ol>		
先備知識	<ol style="list-style-type: none"> <li>1. 人工智慧基本簡介。</li> <li>2. 試算表軟體應用。</li> </ol>		
運算思維	<ol style="list-style-type: none"> <li>1. 運 t-IV-4 能應用運算思維解析問題。</li> <li>2. 運 r-V-1 能將問題以運算形式呈現。</li> </ol>		
與課程綱要的對應	學習內容	<ol style="list-style-type: none"> <li>1. 資 A-V-2 重要演算法的概念與應用。</li> <li>2. 資 D-V-1 巨量資料的概念。</li> <li>3. 資 D-V-2 資料探勘與機器學習的基本概念。</li> </ol>	
	學習表現	<ol style="list-style-type: none"> <li>1. 設 k-IV-1 能了解日常科技的意涵與設計製作的基本概念。</li> <li>2. 設 a-IV-2 能具有正確的科技價值觀，並適當的選用科技產品。</li> </ol>	

教學活動	活動內容	教材
引起動機 (7分鐘)	3. 訓練完分類器後，只要對著 Webcam 給予任何影像資料，就可以進行分類的判斷。該網站就會依據擷取到的影像資料，判定較可能屬於哪一種分類。	【搭配簡報 chap 03-3】 
	4. 在進行完上述活動後，教師引導學生思考幾個問題： <ul style="list-style-type: none"> <li>• 為什麼需要給予訓練的資料？</li> <li>• 從 Webcam 擷取的每一張圖片代表的意義是什麼？</li> <li>• 如果今天，給予的影像與原先訓練的分類結果有差異，是否一樣可以進行分類結果的判斷？</li> </ul>	
學習經驗回顧 (5分鐘)	1. 教師引導學生由簡報 ch-ap 03-4 中呈現的兩張動物圖片，左邊是什麼？右邊又是什麼？	【搭配簡報 chap 03-4】 
	2. 依據學生的回答，很快發現，學生能夠直接反應左邊鳥，及右邊是狗，甚至可以更精確的講出翠鳥與科基犬。教師在此時，需引導學生思考，究竟什麼時候，我們開始認識這些動物？甚至能進一步知道牠們的品種。	【搭配簡報 chap 03-5】 

監督式學習學習單 Part1 參考答案：



### 從0開始學AI 「非監督式學習」學習單Part 1

學生：.....年.....班.....號.....

**第一部份：最近距離分類器**

依據最近距離分類器的運作，請問若有一筆新的資料在座標軸上(5,5)，請問該筆資料應該屬於哪一個類別？

類別中心點座標	2點距離
紅色(3,6.25)	$d = \sqrt{(5-3)^2 + (5-6.25)^2} = \sqrt{5.5625}$
綠色(6.83,3.83)	$d = \sqrt{(5-6.83)^2 + (5-3.83)^2} = \sqrt{4.7178}$

**第二部分：KNN分類器K-Nearest Neighbor Classifier**

假設資料有A類別及B類別，座標上X軸與Y軸代表2種不同的特徵值，當要找出新的資料(5,5)是屬於A類別還是B類別的時候，就要視K值多寡而定。

A類別	距離	B類別	距離
A1 (4, 6)	$\sqrt{2}$	B1 (2, 2)	$\sqrt{18}$
A2 (5, 7)	2	B2 (4, 2)	$\sqrt{10}$
A3 (5, 8)	3	B3 (4, 4)	$\sqrt{2}$
A4 (5, 8, 6)	$\sqrt{1.64}$	B4 (5, 4)	1
A5 (6, 6)	$\sqrt{2}$	B5 (5, 3)	2
A6 (6, 7)	$\sqrt{5}$	B6 (6, 2)	$\sqrt{10}$
A7 (7, 5)	2		
A8 (7, 7)	$\sqrt{8}$		
A9 (8, 4)	$\sqrt{10}$		
A10 (9, 5)	4		

根據上表計算可得到K個與(5,5)最近的點是：(填代碼即可)

K值	A類別	B類別	判定類別
K = 1		B4	<input type="checkbox"/> A類別 <input checked="" type="checkbox"/> B類別
K = 3	A4、(A1)、(A5)	B4、(B3)	<input checked="" type="checkbox"/> A類別 <input checked="" type="checkbox"/> B類別
K = 5	A1、A4、A5	B3、B4	<input checked="" type="checkbox"/> A類別 <input type="checkbox"/> B類別
K = 9	A1、A2、A4、A5、A6、A7	B3、B4	<input checked="" type="checkbox"/> A類別 <input type="checkbox"/> B類別



## 和AI做朋友 從0開始學AI

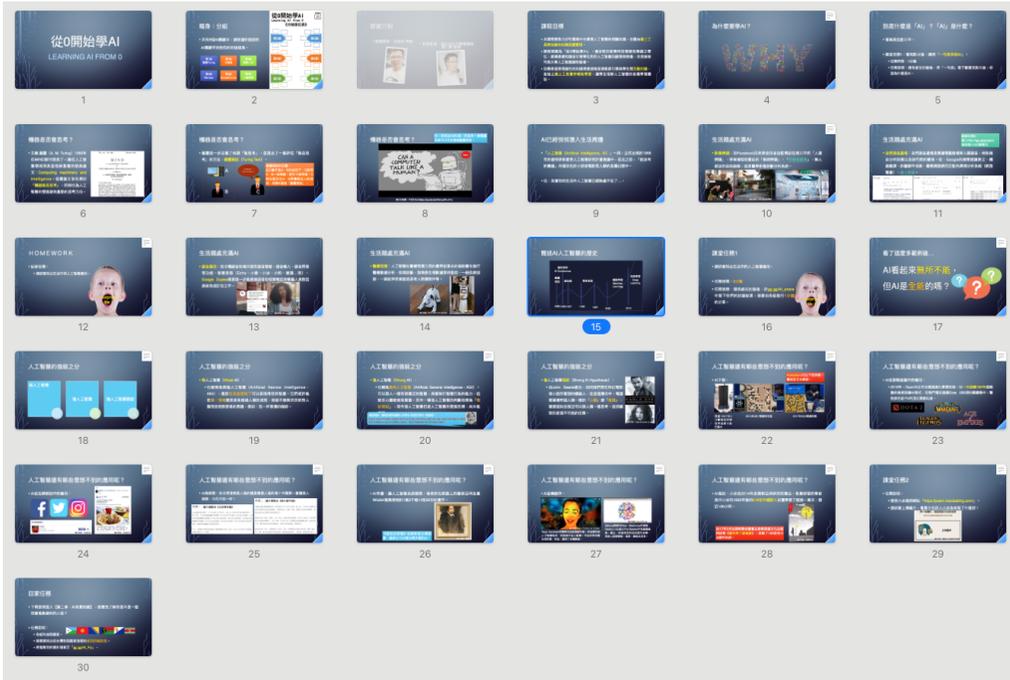
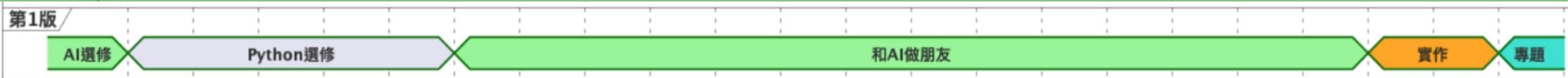
相知篇

中小學人工智慧教育教學示範系列

高中階段

教案

# 3. 課程發展



# 3. 課程發展

第1版

AI選修

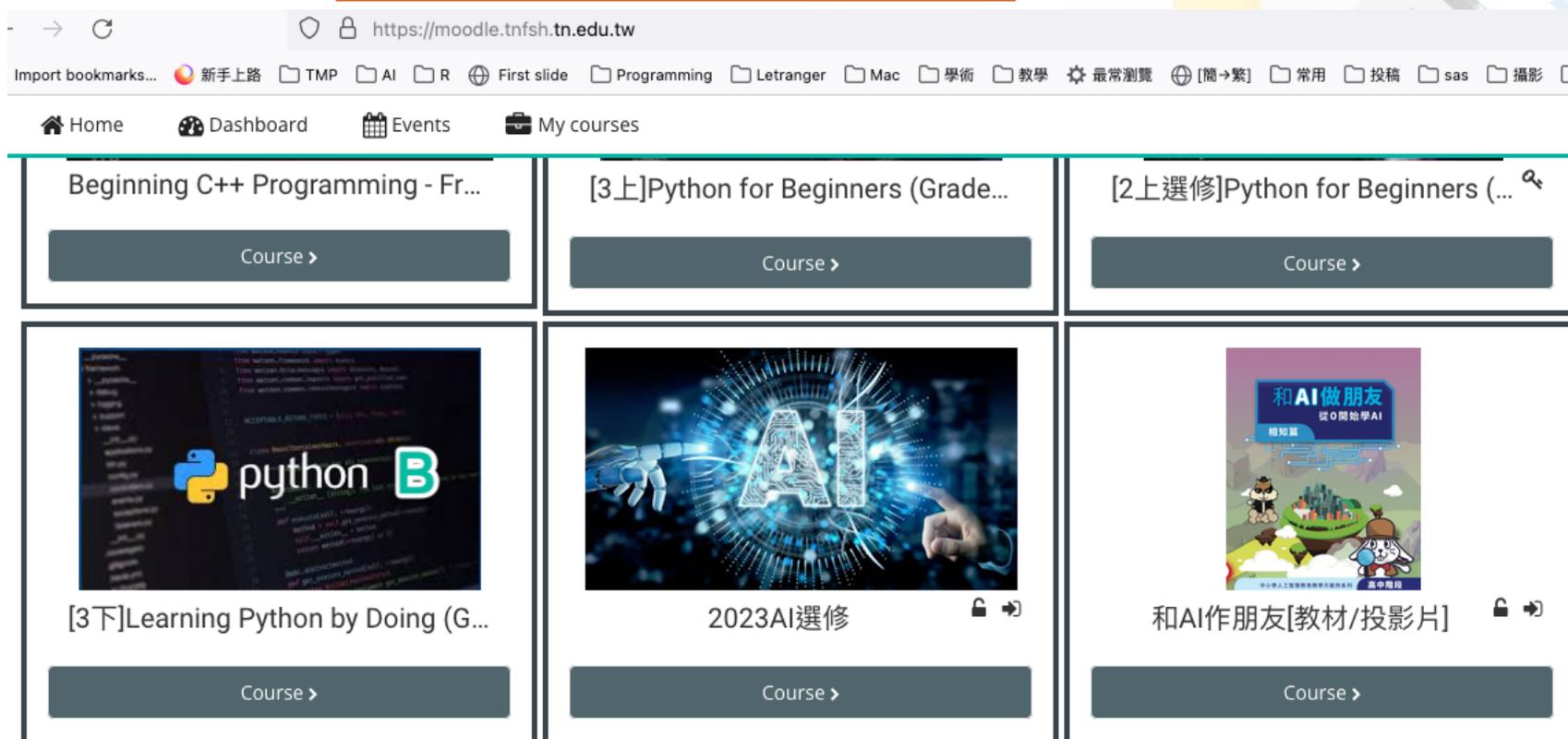
Python選修

和AI做朋友

實作

專題

相關資源下載：[moodle.tnfsh.tn.edu.tw](https://moodle.tnfsh.tn.edu.tw)



The screenshot displays the Moodle LMS interface for <https://moodle.tnfsh.tn.edu.tw>. The navigation bar includes Home, Dashboard, Events, and My courses. The main content area features a grid of course cards:

- Beginning C++ Programming - Fr...** with a "Course >" button.
- [3上]Python for Beginners (Grade...** with a "Course >" button.
- [2上選修]Python for Beginners (...** with a "Course >" button.
- [3下]Learning Python by Doing (G...** with a "Course >" button. The card features a Python logo and the letter 'B' on a dark background.
- 2023AI選修** with a "Course >" button. The card features a glowing 'AI' graphic with a hand pointing at it.
- 和AI做朋友[教材/投影片]** with a "Course >" button. The card features a colorful book cover titled "和AI做朋友 從0開始學AI" (Be Friends with AI: Learning AI from 0).

# 3. 課程發展



## 版本2

- 精簡內容
- 每個主題加入實作

https://letranger.github.io/AI-course/

## AI for TNFSH: 從入門到入土

hits 130

這是為 TNFSH 的 AI 入門 選修課程隨便編的講義，這門課假設你已修過 Python 入門 的選修課程或是已自學基本的 Python，內容持續編修中，編者不保證內容正確性，服用前請保持對一切內容懷疑的精神...

- AI 簡介
- 機器學習
- 監督式學習
- 非監督式學習
- 神經網路
- Say Hi to CNN
- CNN 實作
- 強化學習
- 深度學習
- Python 期末報告：用 deep learning 分類心電圖
- Tinker With a Neural Network Right Here in Your Browser.
- 期末報告

## 6-2 K-means演算法

俗話說「物以類聚」，分群的概念簡單來說就是將相近的資料彼此分在同一群體，接下來我們將以圖6-1來介紹K-means演算法的運作原理。在下圖6-1中，每一個點都視為一筆資料，而座標平面上將以X軸與Y軸來代表該筆資料的2種特徵值。

因此可以得到圖上共有16個點，座標值(X,Y)分別是(1,3)、(1,4)、(2,2)、(2,5)、(2,6)、(3,2)、(3,3)、(3,8)、(4,4)、(4,6)、(5,0)、(5,5)、(6,2)、(6,6)、(7,2)及(7,4)。

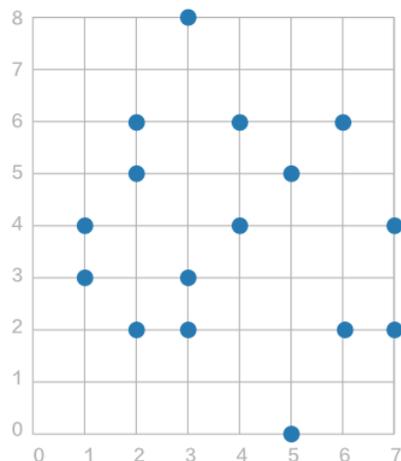


圖6-1 座標平面上共有16個點，X軸與Y軸代表2種特徵值

**K-means演算法的執行步驟如下：**

- **步驟1：**K-means演算法的第一步，需要先選定「K」的值，代表接下來要將資料分為K群。在此假設K為2。
- **步驟2：**上一步驟已經決定出K值，接著就任意在座標平面上指定K個點作為初始的分群中心點，在此隨機擇定以(2,2)及(6,6)此2點作為分群中心。
- **步驟3：**既已擇定K個(K=2)分群中心，座標上的各點可以分別計算與各分群中心間的「距離」，來決定歸屬於哪一個群體。在此距離計算方式採取「歐幾里得距離」，前一章節已有介紹，在此將直接進行計算各點與各分群中心的距離，如下表6-1。

## 3. K-Means

### 3.1. 簡介

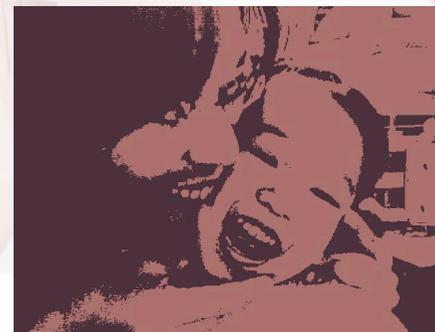
從資料中找出 K 個分類的非監督式機器學習演算法

```
sklearn.datasets.samples_generator import make_blobs
X, y_true = make_blobs(n_samples=300, centers=4, cluster_std=0.60, random_state=0)
plt.scatter(X[:, 0], X[:, 1], s=50);
plt.show()
```

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3)
kmeans.fit(X) cluster = kmeans.predict(X)
plt.scatter(X[:,0], X[:,1], c=cluster, cmap=plt.cm.Set1)
plt.show()
```

### 3.2. 應用：壓縮影像

```
1: import numpy as np
2: import matplotlib.pyplot as plt # 需安裝 pillow 才能讀 JPEG
3: from matplotlib import image
4: from sklearn.cluster import MiniBatchKMeans
```



# 3. 課程發展



## 7-2 由代理人例子來了解如何進行最佳行動

我們用一個簡單的例子來說解在增強式學習的過程中，機器也就是上述所指稱的「代理人」如何在環境中挑選最佳行動，以獲取最大的利益。假設今天在一個迷宮中，裡頭有著寶藏 (treasure) 和怪物 (monster)，而迷宮裡的路錯綜複雜，我們在入口處並不知道要怎麼走，而此時代理人的任務則是要最快地取得寶藏並避開怪物。如下圖7-2所示，代理人每次只能往上、下、左、右任一方向前進一格，灰色底的位置則是不能走的障礙物。

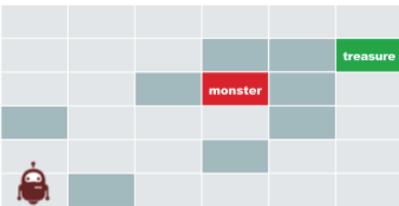


圖7-2 增強式學習中代理人所處環境之介紹

由於一開始代理人沒有任何的知識來教導它如何才能最快地到達寶藏，所以在迷宮內的某個點，它只能用嘗試各種可能性的方法選擇一個行動。代理人在不同的狀態下如何走下一步？又可以採取什麼策略呢？

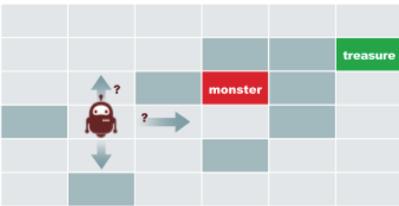


圖7-3 不同狀態下代理人如何挑選下一步

### 更新Q表中的Q值：

接下來將以範例來說明更新Q值的作法。以表5-2的Q表為例，若目前在S1這個狀態中，假設在此狀態做出任何行動後尚無法獲得寶藏，此時由Q表中S1選擇a3這個行動，因其獲得寶藏的可能性較高。若在S1這個狀態下執行a3會進入S2，則根據「回推」的原則，S2中Q值最高的a3就是更新S1中Q值之關鍵。依據上述定義，其Q值更新計算過程如下（假設 $\gamma=0.9$ ）：

$$Q(S1, a3) = R(S1, a3) + \gamma \max_{a'} (Q(S2, a')) = 0 + 0.9 \times 0.35 \approx 0.32$$

表7-2 搭配課程範例之Q表

	a <sub>1</sub>	a <sub>2</sub>	a <sub>3</sub>	a <sub>4</sub>
s <sub>1</sub>	0.35	-0.25	0.56	-0.05
s <sub>2</sub>	-0.54	0.22	0.35	-0.08

雖然已經知道更新Q值的方式，但我們通常不會直接以上述計算得到的結果來直接更新Q值，反而會透過「時值差異 (Temporal Difference, 簡稱TD)」的方式更新。時值差異TD的定義如下：

$$TD(S, a) = R(S, a) + \gamma \max_{a'} (Q(S', a')) - Q(S, a)$$

由時值差異TD的定義，可發現就是將上述計算更新後Q值的結果減掉原先的Q值，藉此找出兩者之間的落差。而得到時值差異TD後，就可以用來更新Q表中的Q值，如下所示：

$$Q_t(S, a) = Q_{t-1}(S, a) + \alpha TD(S, a)$$

因此在某個狀態S下執行行動a時，其Q值的變化在既有的Q值中，又參照一定比例的時值差異TD，來推估出新的Q值。其中， $\alpha$ 是一個介於0到1之間的值，當 $\alpha$ 越大，代表從時值差異TD學習的幅度越大，受下一步Q值的影響也越大。延續表7-2的範例，得到的Q值更新計算過程如下（假設 $\gamma=0.9$ 、 $\alpha=0.9$ ）：

$$TD(S1, a3) = 0 + 0.9 \times 0.35 - 0.56 = -0.245$$

$$Q_t(S1, a3) = 0.56 + 0.9 \times (-0.245) = 0.3395 \approx 0.34$$

## 3. CartPole

Gym上有許多可用工具，CartPole是其中較為簡單常見的一種，CartPole是一個桿子連在一個小車上，小車可以無摩擦的左右運動，桿子（倒立擺）一開始是豎直線向上的。小車通過左右運動使得桿子不倒。

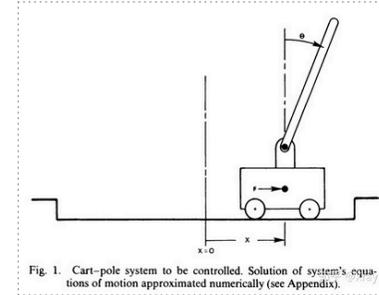
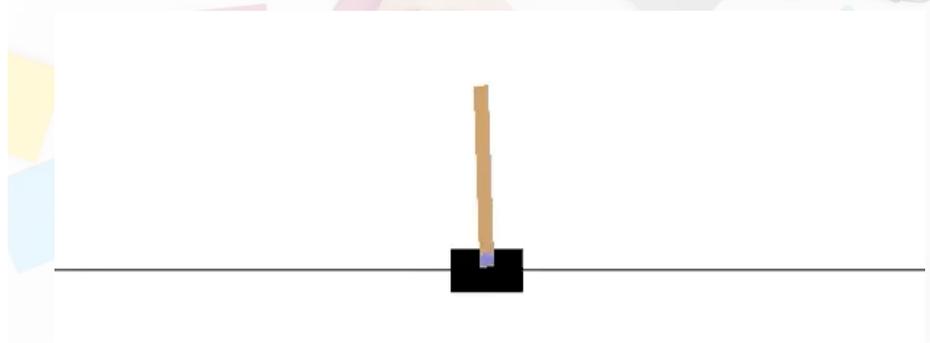


Fig. 1. Cart-pole system to be controlled. Solution of system's equations of motion approximated numerically (see Appendix).

Figure 2: Cart-pole system



# 3. 課程



這是為 TNFSH 的 AI 入門選修課程 隨便編的講義，這門課假設你已修過 Python 入門 的選修課程 或是已自學基本的 Python 語法及第三方函式庫 (如 Numpy, Matplotlib, Pandas...)。內容持續編修中，編者不保證內容正確性，服用前請保持對一切內容懷疑的精神...

### 1. 概論

- 人工智慧
- 機器學習
- CoLab
- 資料預處理

### 2. 監督/非監督

- 監督式學習
  - 分類
  - 迴歸
- 非監督式學習

### 3. 神經網路/深度學習

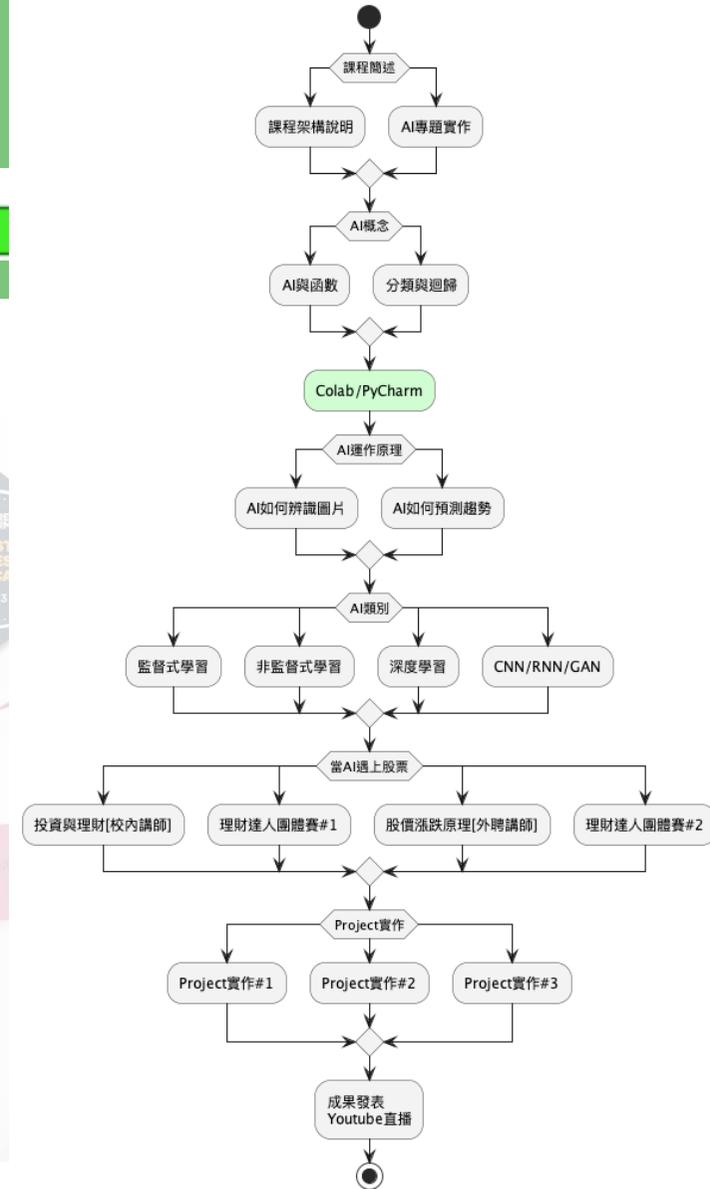
- 感知器
- 神經網路
- 卷積神經網路
- 深度學習

### 4. AI 應用實作

- CNN 實作
- 當 AI 遇上股票

## 版本3

- 加入股價預測活動
- 任務導向
- 自編教材為主軸



# 版本3起源

## 3. 課程發展

第3版

AI選修

Python選修

理論

實作

理論

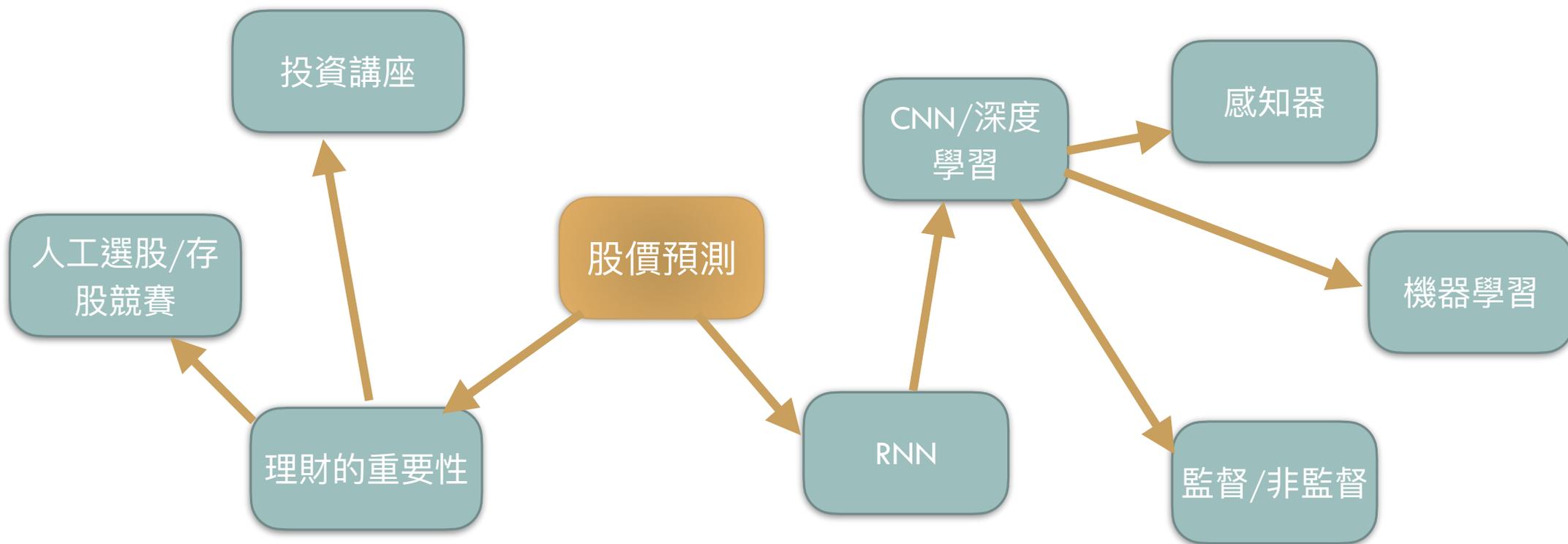
實作

理論

實作

股價預測

專題



# 版本3設計流程

## 3. 課程發展

第3版

AI選修

Python選修

理論

實作

理論

實作

理論

實作

股價預測

專題

### • 階段1: 預期結果

1. 學生能利用傳統工具規畫投資
2. 學生能將AI模型應用股票漲跌預測
3. 學生能修改模型、評估模型效能
4. 學生能將AI模型應用於其他與生活相關的實際事物中



# 版本3設計流程

## 3. 課程發展

第3版

AI選修

Python選修

理論

實作

理論

實作

理論

實作

股價預測

專題

### • 階段2: 評量方式

1. 理財達人賽獲利(依排名評分)
2. 股票走勢預測準備性(依排名評分)
3. 模型的進化歷程(各組修正不同版本的記錄)
4. 專題實作:
  - 專題創意、及實用價值
  - 模型應用、修訂以及效能評估
  - 對專題架構與內容的熟悉程度

# 版本3設計流程

## 3. 課程發展

第3版

AI選修

Python選修

理論

實作

理論

實作

理論

實作

股價預測

專題

### • 階段3: 學習活動

1. AI基本概念: 各種常見的AI學習類型及基本原理。
2. AI模型的實作練習
3. 投資與理財
4. 股票漲跌影響因素分析
5. CNN, RNN, LSTM簡介及股價預測實作
6. 資料處理: 整理、擴增



# UbD: Understanding by Design

3. 課程發展

第3版

AI選修

Python選修

理論

實作

理論

實作

理論

實作

股價預測

專題

- 逆向設計(Backward Design)：反向思考
  1. 先確定預期的學習成果(學習目標)
  2. 針對學生應該如何表現出對這門課的理解(學生該具備何種能力)，去設計測驗、評量方式，
  3. 最後才思考教學活動內容。



UbD的教學設計  
Backward Design 逆向設計





## 理財達人團體賽#1: 工人智慧

- 背景
  - 假設從六年前開始養成每月定期投資的習慣，現在的你損益情形如何？
  - 體驗經歷疫情與戰爭的投資成果
- 活動內容
  - 教師介紹幾種常見的投資商品
  - 示範如何由網路蒐集所需資訊
  - 示範如何以試算表計算六年投報率
  - 示範如何將資料視覺化
  - 學生實作、挑選投資商品、計算損益、繪製圖表

## 理財達人團體賽#1: 工人智慧

### 目標

- 成為賺最多的人，或
- 成為賠最多的人

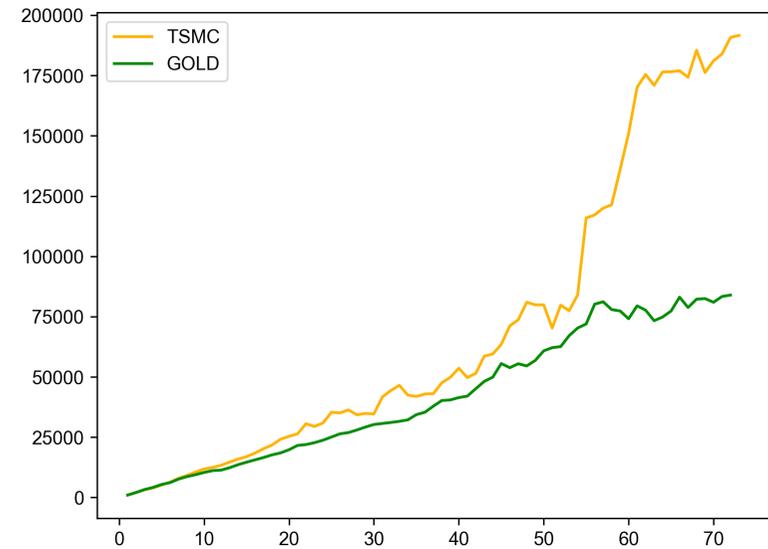
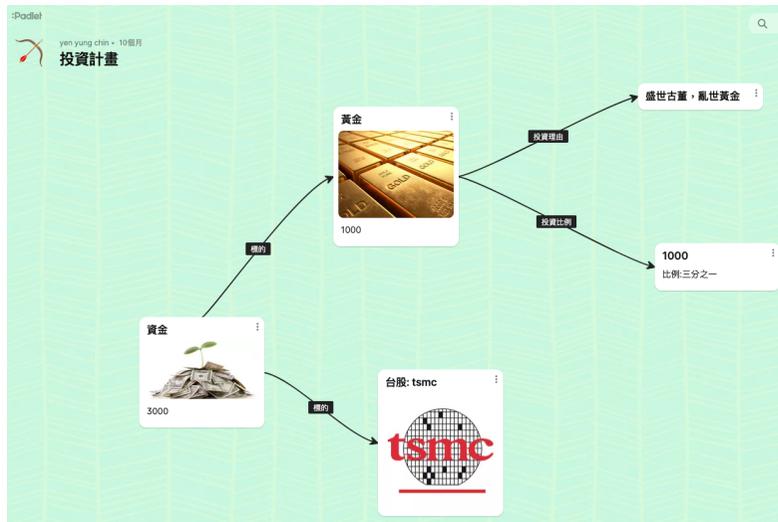
## Competition Rules

1. 投資期限：<2016-01-01 Fri> to <2021-12-31 Fri>
2. 分三種投資項目，每個項目每月投入1000元
3. 三種投資項目中需包含至少一檔台股
4. 股票配股配息併入下月購買資金，只要基於真實數據均可列入計算(各股配息資訊需自行搜尋)
5. 假設所有股票均提供零股買賣，且零股可以小數點買進(現實中股票最低買賣單位為1股)

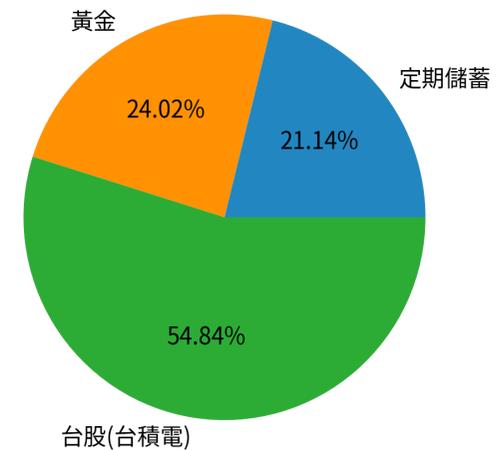
## Group Tasks

- 小組討論投資項目
- 收集六年中每月股價資訊
- 製表
- 計算每項獲利
- 合併組員數據計算全部獲利
- 繪圖

# 分組報告



總金額: 349482, 投報率: 61.80%



	A	B	C	D	E	F	G	H	I
1	月份	日期	時間	買入	賣出	資金	購入單位	累計單位	累計財富
2	1	0105-01-04	09:03:23	1,121	1,133	1000	0.8921	0.8921	1010.7047
3	2	0105-02-01	09:02:23	1,198	1,213	1000	0.8347	1.7268	2094.5904
4	3	0105-03-01	09:01:45	1,316	1,330	1000	0.7599	2.4867	3307.2626
5	4	0105-04-01	09:02:03	1,272	1,286	1000	0.7862	3.2728	4208.8557
6	5	0105-05-03	09:01:44	1,330	1,344	1000	0.7519	4.0247	5409.2060
7	6	0105-06-01	09:01:43	1,270	1,283	1000	0.7874	4.8121	6173.9351
8	7	0105-07-01	09:03:45	1,362	1,377	1000	0.7342	5.5463	7637.2865
9	8	0105-08-01	09:03:24	1,368	1,382	1000	0.7310	6.2773	8675.2520
10	9	0105-09-02	09:03:45	1,331	1,345	1000	0.7513	7.0286	9453.5097
11	10	0105-10-03	09:04:24	1,319	1,333	1000	0.7582	7.7868	10379.7802
12	11	0105-11-01	09:03:25	1,287	1,301	1000	0.7770	8.5638	11141.4812
13	12	0105-12-01	09:03:44	1,195	1,208	1000	0.8368	9.4006	11355.9281
14	13	0106-01-03	09:03:25	1,192	1,206	1000	0.8389	10.2395	12348.8719
15	14	0106-02-02	09:03:24	1,216	1,229	1000	0.8224	11.0619	13595.0718
16	15	0106-03-01	09:03:24	1,220	1,232	1000	0.8197	11.8816	14638.0936
17	16	0106-04-05	09:03:42	1,220	1,229	1000	0.8197	12.7012	15609.8259
18	17	0106-05-02	09:03:25	1,212	1,225	1000	0.8251	13.5263	16569.7470
19	18	0106-06-01	09:03:21	1,219	1,232	1000	0.8203	14.3467	17675.0958
20	19	0106-07-03	09:03:25	1,204	1,218	1000	0.8306	15.1777	18485.8703

## 理財達人團體賽#2: 人工智慧

- 活動一
  - 教師示範基本股價預測模型(每日收盤價)
- 活動二
  - 影響股價漲跌因素(外聘講師)
  - 小組修正股價預測模型
  - 教師指定股票代號，各組預測未來一週股價
  - 小組解釋模型預測失準原因，提出解決策略

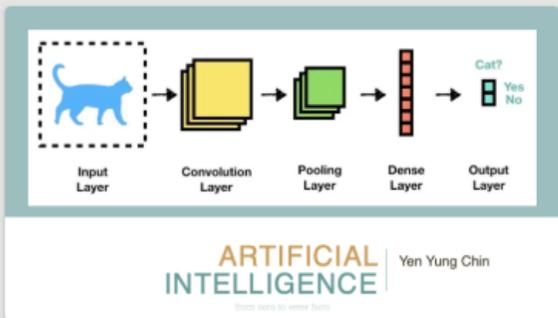
# 新興科技計畫

## 4. 課程簡介

1)主題樹連結：人工智慧預測股市動向為重心，為主題樹中的進階端課程。



# 4. 課程內容 #1



1

### OUTLINE

- AI做為研究主題的優勢與劣勢
- 從ChatGPT看AI的進展
- AI發展史
- AI類型
- AI在各領域的應用
- 各主題簡單理論/實作
- 期末專題發表

2

- ### AI做為研究主題的優勢與劣勢
- 優勢
    - AI的學習門檻低，只要能寫程式，幾個月就能上手
    - AI做為黑盒子，只玩應用算是做專題的捷徑，不懂也能用
  - 劣勢
    - 沒有深入的研究，只能玩皮毛
    - 要精通很困難

3

當我說我在研究深度學習

### 朋友以為我在做

The Triumph of Deep Learning!

AlphaGo

4

當我說我在研究深度學習

### 我媽以為我在做

5

當我說我在研究深度學習

### 鄰居以為我在做

6

當我說我在研究深度學習

### 其實我在做

7

### 從CHATBOT看AI的進展

1964: ELIZA    1995: ALICE    2013: Mitsuku    2022: ChatGPT    2023: Bard

8

### HISTORY OF AI DEVELOPMENT

- 1940: 人工神經網路
- 第一波(符號邏輯): 把人類的知識與思考放入電腦
- 第二波(專家系統): 讓電腦按照人類定義的規則做決策
- 第三波(機器學習): 讓電腦從資料中歸納規則

9

# 4. 課程內容 #1

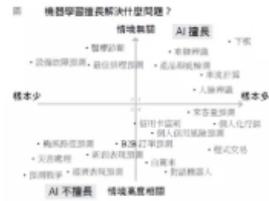
## AI在各領域的應用

- 人工智慧在台灣：產業轉型的契機與挑戰講AI在各個領域的實際應用
- 作者：陳昇璋, 溫怡玲
- 出版社：天下雜誌



49

## AI 擅長的解題領域



50

## 製造業

- 瑕疵檢測
- 自動流程控制: 生產流程中, 如馬達轉速、電流、電壓、環境溫度...等等會影響產品良率的因素可能高達上千個
- 預測性維護: 預測機器何時會出錯以提前進廠保養
- 原料組合最佳化: 每批原料可能來自不同供應商、品質、等級或特性可能有所差異, 如何在各原料、供應商、等級、成本的排列中找出最高 CP 值的組合即為重要工作

51

## 遊戲產業



- NPC的進化: 可以記住與玩家的互動, 產生不固定的對話內容、了解玩家的偏好招式、記住以往的恩怨情仇, 即, **更像玩家**

52

## 政府機構: 交通

- 高市交通局AI智慧號誌



53

## 政府機構: 預防犯罪

- 美國加州的PredPol公司



54

## 政府機構: 獨裁統治



- 中國利用 2 億個監視器監控 14 億公民, 並對公民的「社會信用評級」即時評分
- 中國新疆的天眼辨識系統嚴密監控當地的維吾爾少數民族
- 四川行政學院早於2017年就開發「智慧紅雲」, 來評估黨員學習教育的成效。
- 中共靠大數據對名列黑名單的總人口進行精準管控, 包括在兩會等特殊時間, 限制他們的人身自由。
- 河南省爆發建築糾紛, 部份人士銀行存款被莫名凍結, 民眾種臘抗議活動時, 地方政府直接對特定人發出「紅色」健康碼限制行動,

55

## 政府機構: 獨裁統治



56

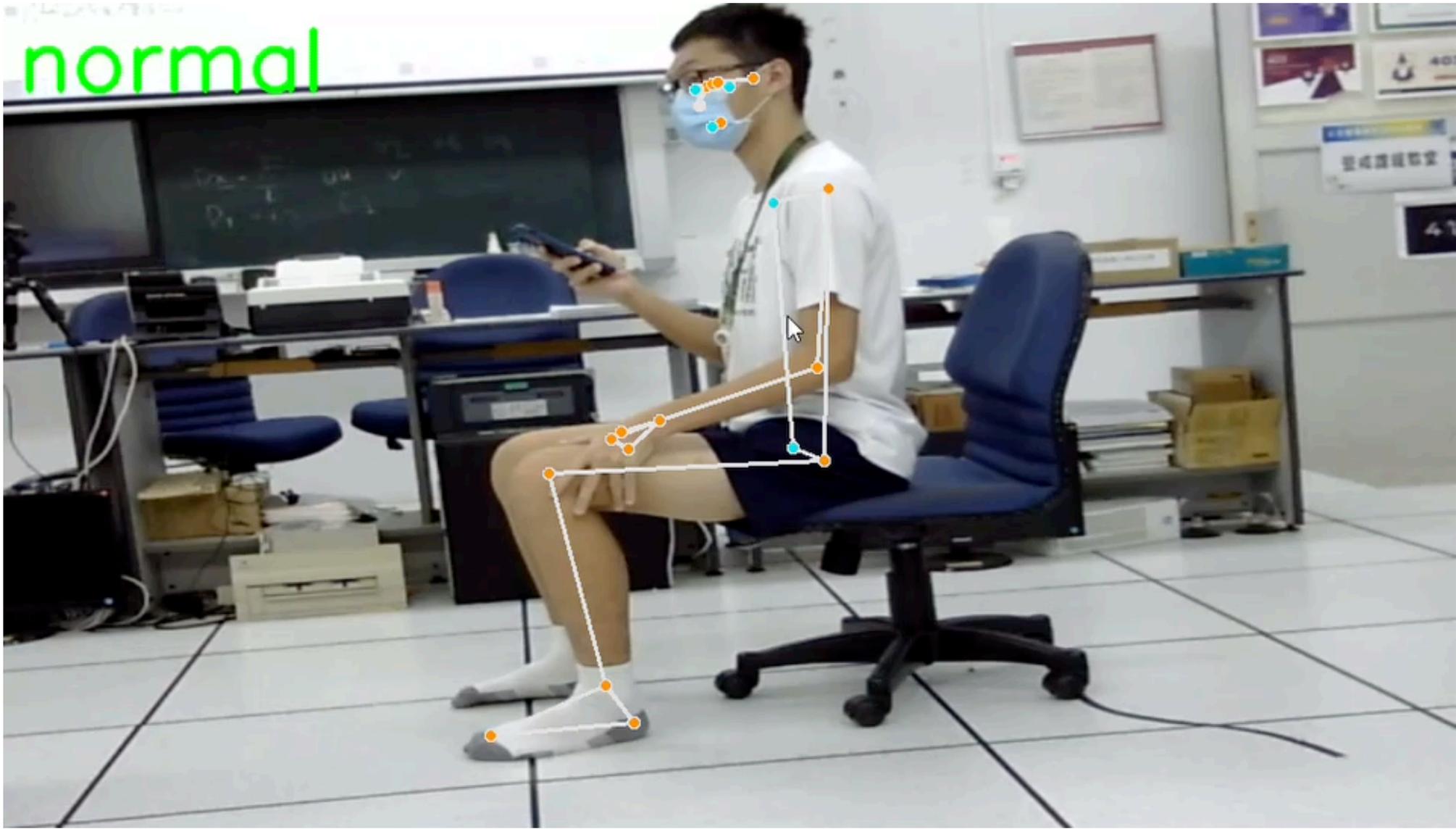
## 百貨零售業



- 美國連鎖百貨 Target 於 2002 年開始分析顧客的購買清單
- Target的資料分析師找出了 25種產品可以作為「懷孕預測指數」的指標, 還可以預測預產期, 讓行銷人員可以精準地在不同的孕期, 寄送專屬的折價券給消費者。

57

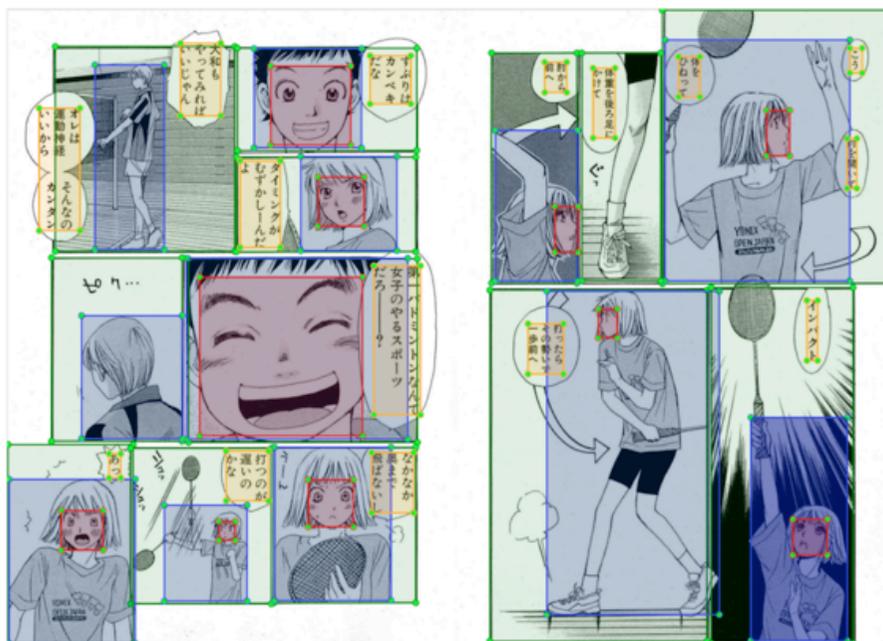
# 4. 課程內容 #1



# 外文漫畫之自動翻譯及製圖研究

## 肆、研究過程或方法

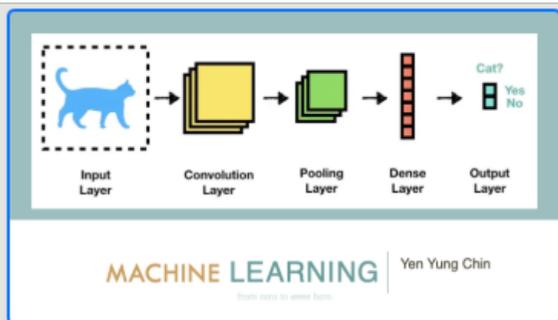
- 綠色：漫畫框
- 藍色：人物
- 紅色：臉部
- 橘色：對話



YamatoNoHane ©Saki Kaori

圖三、Manga109之漫畫資料及註解資料示範

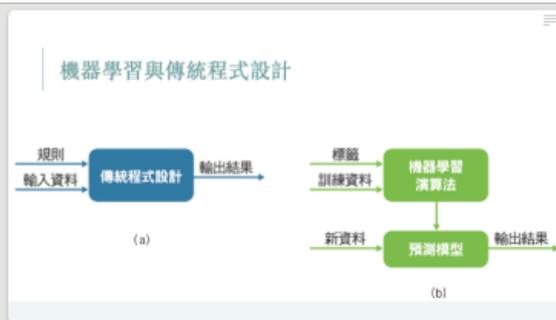
# 4. 課程內容 #2



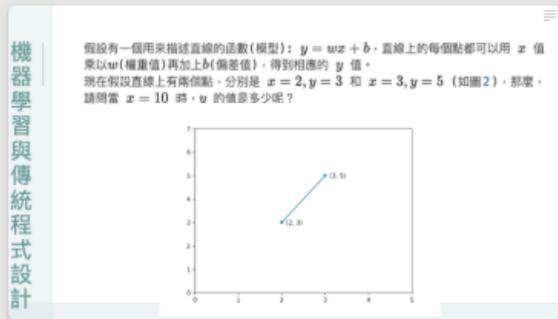
1

- ### OUTLINE
- 機器學習與傳統程式設計
  - 機器學習的類型
  - 機器學習的資料呈現
  - 資料預處理
  - 特徵值間的距離

2



3



4

### 機器學習與傳統程式設計

#### 傳統程式設計

```
def get_slope(p1, p2):
    p1x, p1y = p1 #取出slope中的(x, y)
    p2x, p2y = p2 #取出slope中的(x, y)
    w = (p2y - p1y) / (p2x - p1x)
    return w

def get_bias(p1, w):
    p1x, p1y = p1 #取出slope中的(x, y)
    b = p1y - (w*p1x)
    return b

def get_y(x, w, b):
    y = w*x + b

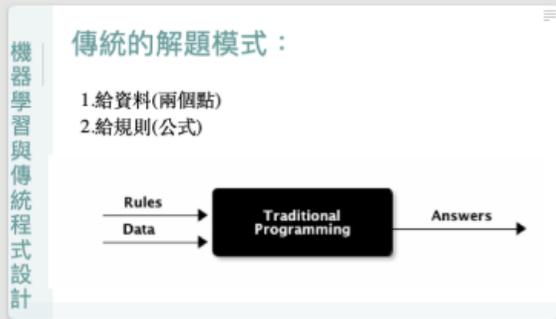
# 利用slope和bias求值(x, y)
p1 = (2, 3)
p2 = (3, 5)

w = get_slope(p1, p2)
b = get_bias(p1, w)

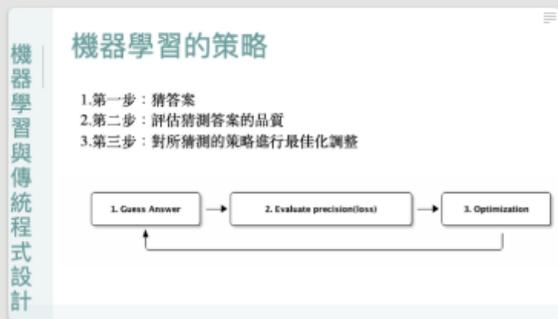
print("Slope:", w)
print("Bias:", b)
print(f"當x=10時, y={w}*10+{b}={w*10+b}")
```

Slope: 2.0  
Bias: -1.0  
當x=10時, y=2.0\*10+-1.0=19.0

5



6



### 機器學習與傳統程式設計

#### 機器學習的策略

```
import tensorflow as tf
import numpy as np
from tensorflow.keras import Sequential
from tensorflow.keras.layers import Dense

l0 = Dense(units=1, input_shape=(1))
model = Sequential([l0])
model.compile(optimizer='sgd', loss='mean_squared_error')

xs = np.array([2.0, 3.0], dtype=float)
ys = np.array([3.0, 5.0], dtype=float)

model.fit(xs, ys, epochs=500, verbose=0)

print(f'y的預測結果為: {model.predict([10.0])}')
print(f"模型的两个参数w, b: {}".format(10.get_weights()))

1/1 [=====] - 0s 172ms/step
y的預測結果為: 14.709847
模型的两个参数w, b: [array([[1.4212971]], dtype=float32),
array([[0.49687672]], dtype=float32)]
```

### 機器學習與傳統程式設計

#### 機器學習的策略

```
import tensorflow as tf
import numpy as np
from tensorflow.keras import Sequential
from tensorflow.keras.layers import Dense

l0 = Dense(units=1, input_shape=(1))
model = Sequential([l0])
model.compile(optimizer='sgd', loss='mean_squared_error')

xs = np.array([2.0, 3.0, 4.0], dtype=float)
ys = np.array([3.0, 5.0, 7.0], dtype=float)

model.fit(xs, ys, epochs=500, verbose=0)

print(f'y的預測結果為: {model.predict([10.0])}')
print(f"模型的两个参数w, b: {}".format(10.get_weights()))

y的預測結果為: [[17.07221]]
模型的两个参数w, b: [array([[1.7164488]], dtype=float32),
array([[0.09227743]], dtype=float32)]
```

# 4. 課程內容 #2

## 資料預處理

**插入空值：**在無法推測欄位資料時，也可以藉由插入空值/零的方式處理，亦即宣告該筆資料不含該項欄位具備之特徵。

很多欄位缺失	2	3	4	null	null	null	null	null	2	4
很多欄位缺失	2	3	4	0	0	0	0	0	2	4

圖2-6 無法推測欄位時，插入空值也是修正資料的方法之一

37

## 資料預處理

**插入平均值/中位數：**能計算資料各欄位的平均值或中位數，將之代入，即將該筆缺失的資料當作平均的情況看待。

38

## 資料預處理

**將名目值量化：**當資料欄位的尺屋為名目資料或次序資料時，可以透過量化的方式把它們轉化成屬於分析的數值資料。

欄位尺度不同	2	4	yes	0.2	n	3	Hot	2	2	3
			No=0 Yes=1	n=0 p=1			Hot=0.75 Cool=0.25			
欄位尺度不同	2	4	1	0.2	0	3	0.75	2	2	3

圖2-7 名目值資料透過量化轉為數值資料

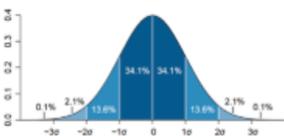
39

## 資料預處理

**移除異常資料：**在有異常值的情況下，如觀測儀器故障、問卷亂填等，我們大多會建議直接刪除該筆資料以保證資料的準確性。

如何找到異常值呢？最簡單的做法是利用標準差。在一般情況下，觀測值常會呈現常態分布：

由常態分布的特性可以知道95%的值都會分布在平均區正負3個標準差的範圍內。



40

## 資料儲存

在整理完資料後，要儲存整理完的資料以供未來使用。通常會將整理完的資料儲存成檔案或存放於資料庫。

左圖擷取自政府資料開放平臺存放之「一般高級中等學校名錄」檔案內容。

41

## 決定特徵

將原始資料整理完畢後，接下來要來檢視是否資料包含的所有特徵都是有用的。

為什麼「胖」跟「睡眠時間」有關係？



42

## 決定特徵-「皮爾森」相關係數

點選該網頁一下數學與商業數據分析

皮爾森相關係數 (Pearson Correlation Coefficient) 可用於度量兩個變數x和y之間的相關(線性相依)性，其值介於-1與1之間。

設兩特徵x、y各有n筆資料，形成(x1,y1),(x2,y2)...定義皮爾森相關係數為：

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

## 資料儲存

在整理完資料後，要儲存整理完的資料以供未來使用。通常會將整理完的資料儲存成檔案或存放於資料庫。

左圖擷取自政府資料開放平臺存放之「一般高級中等學校名錄」檔案內容。

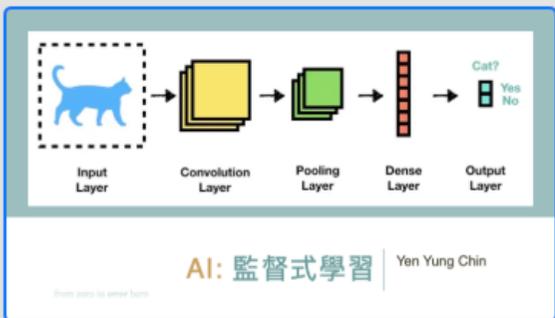
## 決定特徵

將原始資料整理完畢後，接下來要來檢視是否資料包含的所有特徵都是有用的。

為什麼「胖」跟「睡眠時間」有關係？



# 4. 課程內容 #3



1

## OUTLINE

- 監督式學習的類型
- 分類
- 迴歸

2

## 監督式學習 (SUPERVISED LEARNING)

- 目前九成以上的機器學習應用均屬此類
- 監督式學習指在訓練過程中直接告訴機器答案，也就是將資料進行標註 (Label)，例如，在 1000 張訓練集照片中標註「貓/狗」。
- 從給定的訓練資料集中學習出一個函式，當新的資料到來時，可以根據這個函式預測結果。
- 監督學習的訓練集要求是包括輸入和輸出，也可以說是特徵和目標。訓練集中的目標是由人標註的。

3

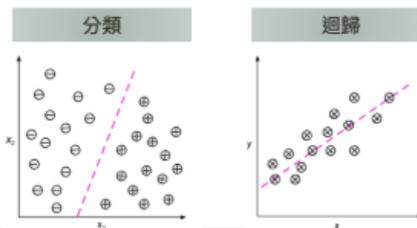
## 監督式學習演算法

- K-nearest neighbors (KNN)
- Methods based on tree (Decision tree and Random Forest)
- Boosting
- SVM (Support Vector Machines)
- 神經網路

4

## 監督式學習類型

機器學習的類型



5

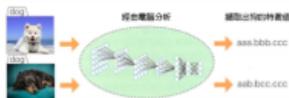
## 「監督式學習」的類型

- 當你預測的目標為連續的數值，我們稱其為迴歸；  
 $f(x) = \text{數值}$   $f(\text{前五次英文模考成績}) = \text{學測英文成績}$
- 若目標為離散的，稱其為分類。  
 $f(x) = \text{類別}$   
 $f(\text{貓}) = \text{貓: 89\%, 狗: 11\%}$   
 $f(\text{狗}) = \text{貓: 20\%, 狗: 80\%}$

6

## 分類

- 藉由特徵值的蒐集，將資料依照某些規則加以整理，然後產生一套區分該物件的原則，藉此建構出分類器 (Classifier)，新的事物藉由分類器去判斷是屬於哪一個分類。



7

## 分類

- 可再細分為：
  - Binary classification
  - Multiclass classification
- 演算法
  - 最短距離分類器
  - KNN分類器
  - 決策樹 Decision Tree

8

## 最短距離分類器

最短距離分類器是利用新進未知資料的特徵，與已知類別資料的特徵比較距離，看看新進資料與哪個類別的特徵最相似 (即資料相似度最高)，就將新進的資料預測為該類別。

Birds of the same kind are often seen to flock and fly together.

9

# 4. 課程內容 #3

## 最短距離分類器

### 找出特徵

- 頭部大小: 計算圖中動物的頭部與身體比例, 將比例值分為 10 個類別, 比例值越高, 代表頭部所佔面積越大。
- 尾巴長度: 計算圖中動物尾巴與身體長度比例, 將比例值分為 10 個類別, 比例值越高, 代表尾巴越長。



10

### 資料集範例

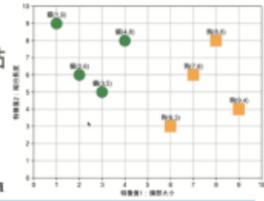
## 最短距離分類器

圖片內容	貓	狗
特徵值 1 (頭部大小)	3	6
特徵值 2 (尾巴長度)	5	3

11

## 最短距離分類器

### 原理：1. 找出中心點



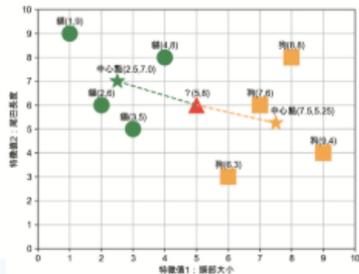
▼ 表 4-2 貓、狗兩類特徵值分佈及平均值

類別	特徵值 1	特徵值 2	類別	特徵值 1	特徵值 2
貓	1	9	狗	6	3
貓	2	6	狗	7	6
貓	3	5	狗	8	8
貓	4	8	狗	9	4
平均值	2.5	7.0	平均值	7.5	5.25

12

## 最短距離分類器

### 原理：2. 對新資料預測

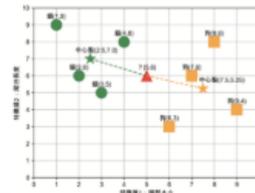


13

## 最短距離分類器

### 原理：2. 對新資料預測

- A與+的直線距離為  $\sqrt{(5-2.5)^2 + (6-7.0)^2} \approx 2.69$
- A與-的直線距離為  $\sqrt{(5-7.5)^2 + (6-5.25)^2} \approx 2.61$



14

## 最短距離分類器

### 其他計算距離的方式

- 除了使用中心點做為距離判斷依據, 也可以計算新進資料與所有資料集中各類的距離。
- 這裡的距離可以是曼哈頓距離或是歐幾里得距離。

15

## 作業

### [VPL 3.1] Create your own classifier (最短距離分類器)

Maximum number of files: 1

Type of work: Individual work

題目:

- 已知有兩類資料, 其特徵值以二維 List 表示如下: 請以 Python 撰寫一最短距離分類器, 輸入任一類資料的特徵數值, 判斷該資料應歸類為哪一類?
 

```

group0 = [(4, 6), (5, 7), (5, 8), (5, 9), (6, 4), (6, 7), (7, 5), (7, 7), (8, 4), (9, 5)]
group1 = [(2, 2), (4, 2), (4, 4), (5, 4), (5, 3), (6, 2)]
            
```

測資:

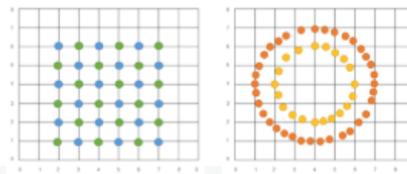
- 輸入: 6, 7
- 輸出: A

16

### 最短距離分類器

## 最短距離分類器

想一想：在甚麼樣的狀況下, 最短距離分類器不能發揮作用？

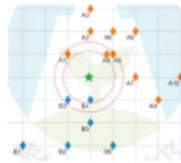


17

## 分類

### KNN 工作原理

1. 選定 k 的值和一個「距離度量」(distance metric)。
2. 找出 k 個想要分類的、最相近的鄰近樣本。
3. 以多數決的方式指定類別標籤。



18

## [VPL 3.1] Create your own classifier (最短距離分類器)

📁 Maximum number of files: 1

Type of work: 👤 Individual work

題目：

- 已知有兩類資料，其特徵值以二維list表示如下：請以Python撰寫一最短距離分類器，輸入任一新資料的特徵數據，判斷該資料應歸類為哪一組？

- `groupA = [[4, 6], [5, 7], [5, 8], [5.8, 6], [6, 6], [6, 7], [7, 5], [7, 7], [8, 4], [9, 5]]`

- `groupB = [[2, 2], [4, 2], [4, 4], [5, 4], [5, 3], [6, 2]]`

- 

測資：

- 輸入

6 7

- 輸出

A

## [VPL 3.2] Create your own classifier (KNN)

📁 Maximum number of files: 1

Type of work: 👤 Individual work

題目：

- 已知有兩類資料，其特徵值以二維list表示如下：請以Python撰寫一KNN分類器，輸入任一新資料的特徵數據，再輸入K值，判斷該資料應歸類為哪一組？

- `groupA = [[4, 6], [5, 7], [5, 8], [5.8, 6], [6, 6], [6, 7], [7, 5], [7, 7], [8, 4], [9, 5]]`

- `groupB = [[2, 2], [4, 2], [4, 4], [5, 4], [5, 3], [6, 2]]`

- 

測資：

- 輸入

1 1 7

- 輸出

B

# 以決策樹進行貸款核准分析作業

參考上述程式碼,完成以下任務

1. 以Pandas讀取線上csv
2. [<https://letranger.github.io/AI/Downloads/loantree.csv>]
3. 移除有缺失值的記錄
4. 將YES/NO、Male/Female等分類值改為0/1值
5. 依Gd,Md,Dd,Ed,SE這五個特徵值來決定是否核准貸款申請
6. 最終是否核准貸款的欄位為LS
7. 分別以Gini index, Entropy兩種策略來進行分類, 比較效能
8. 觀察其他特徵值, 你有其他的想法可以提高效能嗎?

# 4. 課程內容 #3

## 迴歸實作

迴歸

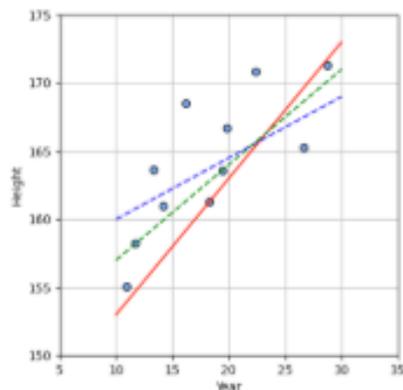
• Model:  $y=wx+b$

我們需要:

- loss function (誤差計算, 找出誤差)
- optimizer (最佳化方法, 做出修正)

98

年齡身高預測



查看資料

101

資料生成

```
import numpy as np
import matplotlib.pyplot as plt

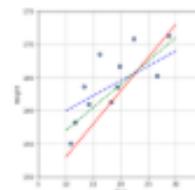
n = 10 # 資料筆數
year = 5 + 25 * np.random.rand(n) # 年紀
height = 170 - 108 * np.exp(-0.2 * year) + 4 * np.random.randn(n)
print(year)
print(height)
```

[13.3, 16.2, 10.9, 20.7, 19.8, 14.2, 11.7, 26.6, 22.4, 18.3, 19.4]  
[163.61, 168.53, 155.06, 171.3, 166.69, 160.98, 158.23, 165.27, 170.83, 161.31,

年齡身高預測

99

年齡身高預測



假設模型

$$y = ax + b$$

- $f_1(x) = 3x + 2$
- $f_2(x) = 4.5x - 3.12$
- ...
- $f_n(x) = -3x + 50$

$$f(x) = ax + b$$

- 迴歸的目的就是要從上述無限多個函數中找出一個最好的
- 為了從無限多個備選模型中找出最佳的, 我們需要有一個評估機制

102

# 依據期中考成績預測期末考成績

## 7.1. Data

- 線上資料: <https://letranger.github.io/AI/PythonScores.csv>
- 資料中有424筆記錄，每筆記錄分別為學生的
  1. `id`: 學號
  2. `class`: 平時成績
  3. `task`: 作業成績
  4. `mid`: 期中考成績
  5. `final`: 期末考成績

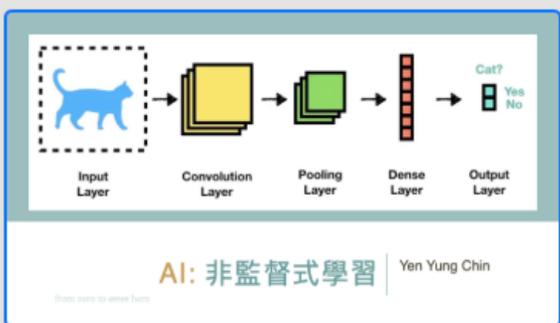
# 依據期中考成績預測期末考成績

## 7.2. Task

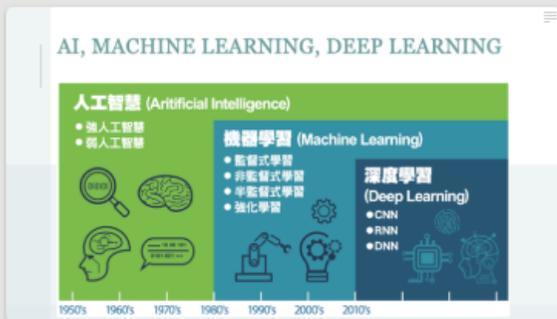
你的任務是建立一個模型，輸入一個或多個特徵值(class, task, mid)來預測期末考成績(final)，其他相關任務包括：

1. 部份學生的期中、期末考有缺考行為，請將這些缺考記錄填入0分
2. 畫出所有特徵資料的分佈狀況(直方圖)
3. 將所有分數間的相關以視覺化方式表現出來
4. 將資料集分割為訓練集(70%)及測試集(30%)
5. 請自行決定你要用多少個特徵值來預測，並以測試集來評估模型效能，輸出分數(R<sup>2</sup>-score)
6. 列出你找出的模型方程式

# 4. 課程內容 #4



1



2

### OUTLINE

- 非監督式學習的目的
- 非監督式學習的類型
- 分群: K-Means
- 聚合式階層分群
- 分裂式階層分群

3

### 回顧：監督式學習

運用監督式學習處理「分類」的問題。

KEY: 藉由原有的資料「建構、訓練分類模型」，用來針對新進資料進行「預測」。

最短距離分類器      KNN分類器      決策樹

4

### 非監督式學習

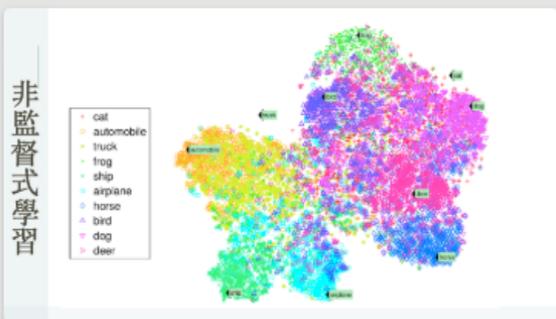
#### 目的

非監督式學習接收未被標記的數據，並通過演算法根據資料的基礎結構(如常見的模式、特色、或是其他因素)將數據分類，而非做出預測。

例如：

- 將網站訪客進行分類: 性別、喜好、上網時段
- 將一堆照片依類型分類: cat、automobile、truck、frog、ship...

5



6



7



8

### 非監督式學習

#### 常見演算法

- 分群(clustering)
  - 聚類(集結)
  - K-Means
  - DBSCAN
  - 層層式分群分析(Hierarchical Cluster Analysis, HCA)
- 異常檢測與新穎檢測
  - One-class SVM
  - 孤立森林(Isolation Forest)
- 降維
  - 線性投影
  - 流形學習

9

# 聚合式分群作業：資料

電子商務網站黃色鬼屋近日收集了200位VIP客戶資料，想將這些客戶依其同質性進行分類。

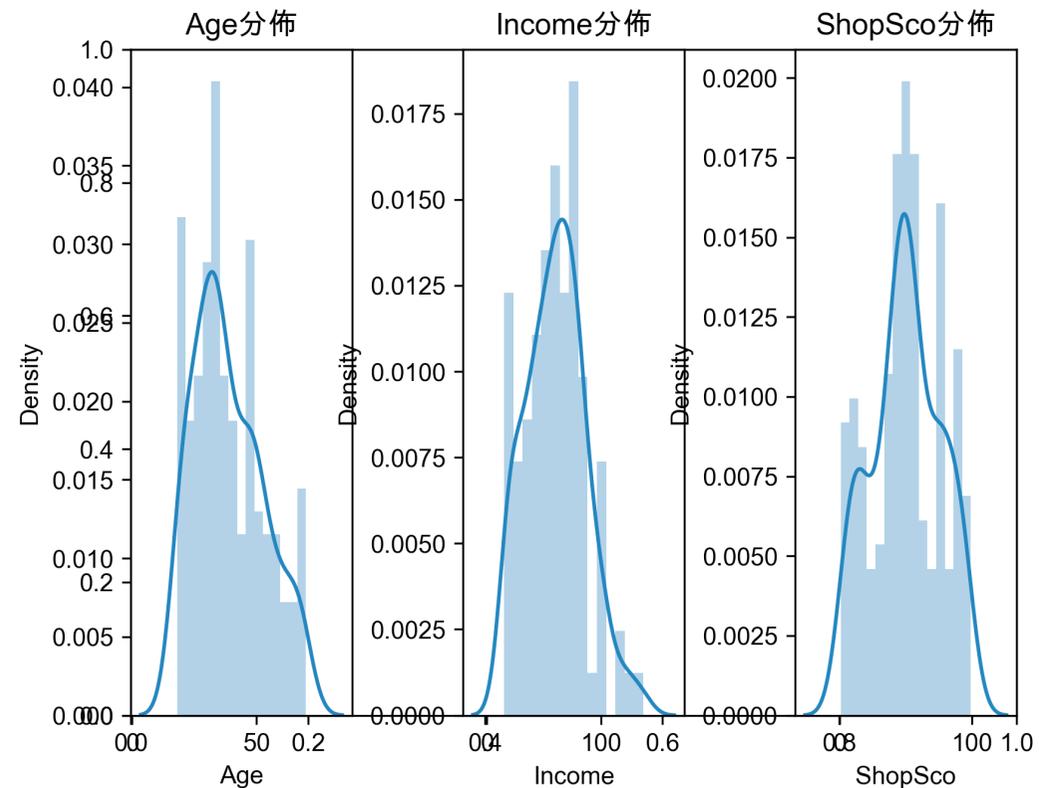
1	CID	Gd	Age	Income	ShopSco
2	1	Male	19	15	39
3	2	Male	21	15	81
4	3	Female	20	16	6
5	4	Female	23	16	77
6	5	Female	31	17	40
7	6	Female	22	17	76

# 聚合式分群作業: 資料

- 資料集URL: <https://github.com/letranger/AI/blob/gh-pages/Downloads/schopaholic.csv>
- CID: 客戶編號
- Gd: 性別(Male/Female)
- Age: 年齡
- Income: 月收入(單位為萬元)
- ShopSco: 這是黃色鬼屋自訂的敗家分數，範圍由0~100

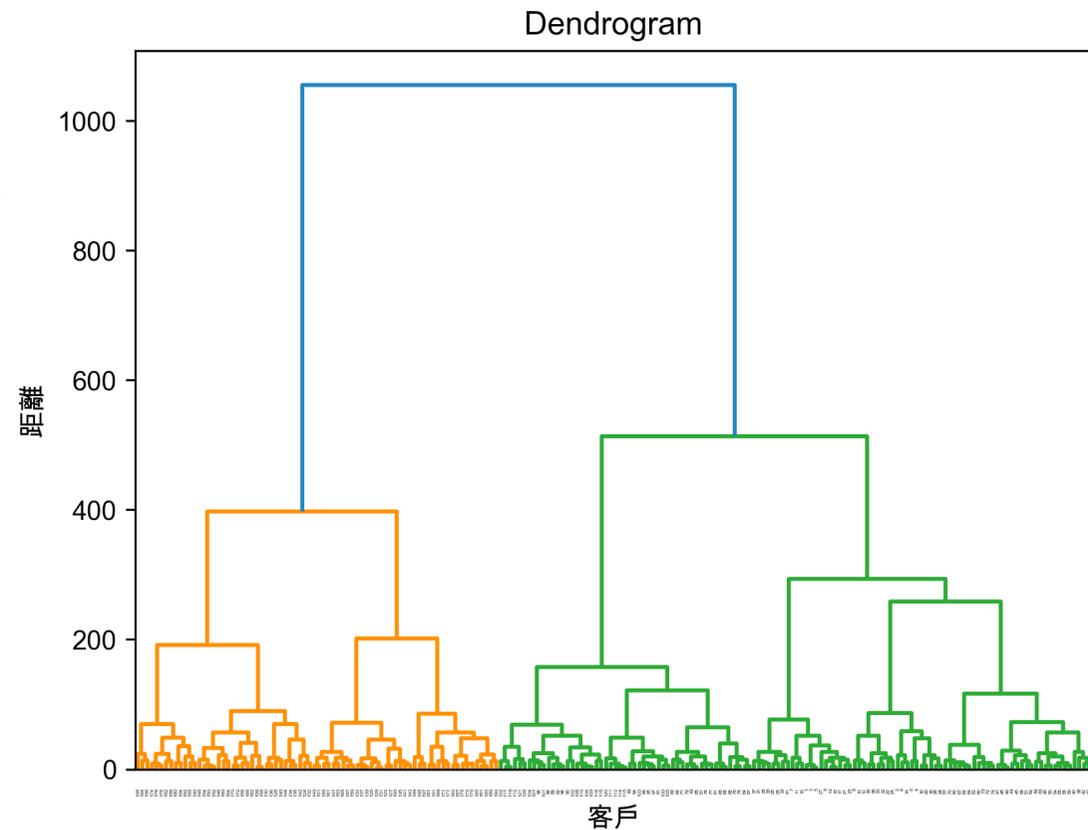
# 聚合式分群作業: 任務

1. 畫出200位VIP客戶的性別、年齡、月收入、敗家分數的分佈狀況，例如:



# 聚合式分群作業: 任務

2. 利用聚合式分群的模型幫黃色鬼屋將階層圖畫出來

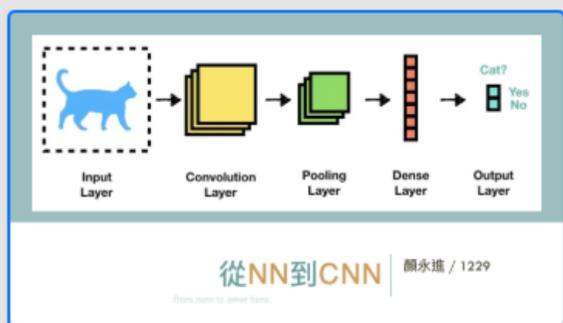


# 聚合式分群作業: 任務

3. 利用聚合式分群的模型幫黃色鬼屋輸出分成5群的結果

```
第1群客戶ID: 127 129 131 135 ...  
第2群客戶ID: 28 44 46 47 48 ...  
第3群客戶ID: 124 126 128 130 ...  
第4群客戶ID: 2 4 6 8 10 12 14 ...  
第5群客戶ID: 1 3 5 7 9 11 13 ...
```

# 4. 課程內容 #5



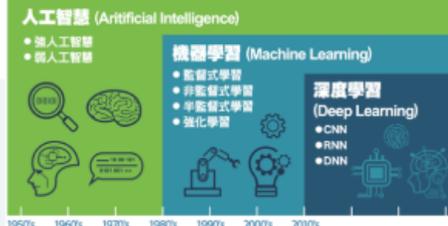
1

## CONTENT

1. AI, Machine Learning, Deep Learning
2. How machine recognize image?
3. Let's create an object detection machine

2

## 1 AI, MACHINE LEARNING, DEEP LEARNING



3

## 傳統程式設計V.S.機器學習



4

## 2. CNN架構與運作原理

### How machine recognize image?

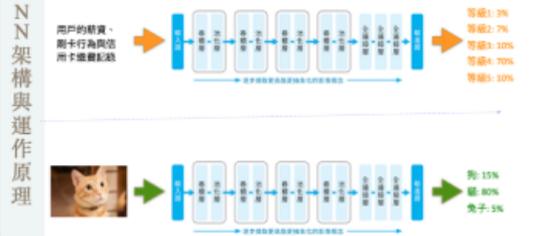
5

## 2. CNN架構與運作原理



6

## 2. CNN架構與運作原理



7

## 如何將待解問題數學化

- 深度學習是透過數學運算來進行，那麼要如何將一個所面臨的問題以數學形式表現出來呢？
- 例如一個圖像識別問題的問題要如何轉換為數學運算進行達成識別分類呢？讓我們先來看一個簡化的圖形識別問題：讓電腦來讀入「/」及「\」兩個圖像並進行識別。

8

## 如何將待解問題數學化

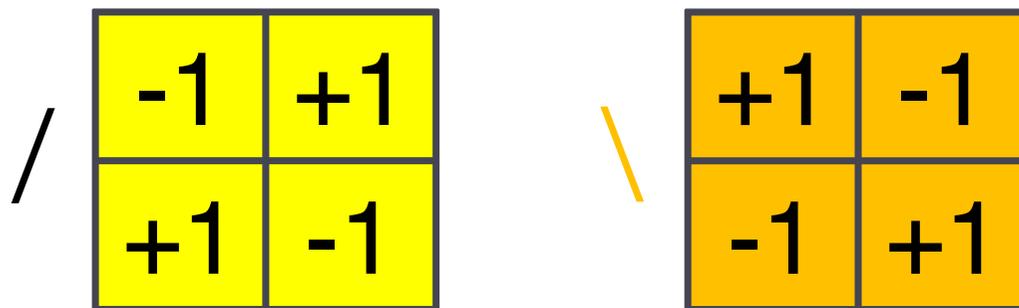
- 首先要理解，電腦眼中的圖像皆是由像素(pixel)組成，因而「/」及「\」兩個圖像在電腦眼中其實是如以下的矩陣所示，其中+1表示有像素存在，反之則以-1表示。



9

## 如何將待解問題數學化

- 首先要先理解，電腦眼中的圖像皆是由像素(pixel)組成，因而「/」及「\」兩個圖像在電腦眼中其實是如以下的矩陣所示，其中+1表示有像素存在，反之則以-1表示。



## 2. CNN 架構與運作原理

$$f\left( \begin{array}{|c|c|} \hline +1 & -1 \\ \hline -1 & +1 \\ \hline \end{array} \right) = \text{“ \ ”}$$

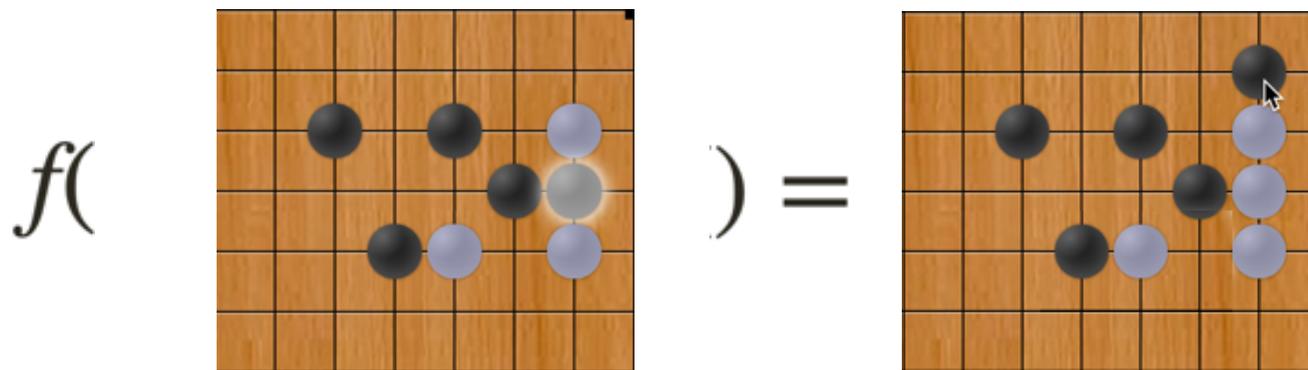
矩陣A

$$f\left( \begin{array}{|c|c|} \hline -1 & +1 \\ \hline +1 & -1 \\ \hline \end{array} \right) = \text{“ / ”}$$

矩陣B

$$f(\text{Image of a dog}) = \text{“dog”}$$

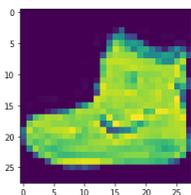
$$f(\text{Audio waveform}) = \text{“台南一中”}$$



$f($  前十日股價  $) =$  明日股價



## 2. 圖片識別與股價預測模型的差異

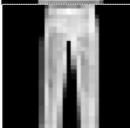


- 1) 準備訓練資料：商品圖片與標籤
- 2) 建立模型: 分類模型
- 3) 訓練模型
- 4) 以圖片預測類別

- 1) 準備訓練資料: 股價歷史資料
- 2) 建立模型: 數值預測模型
- 3) 訓練模型
- 4) 以前 $n$ 天的股價預測第 $n+1$ 天的股價

### 3. CNN 股價預測實作：建立測試資料

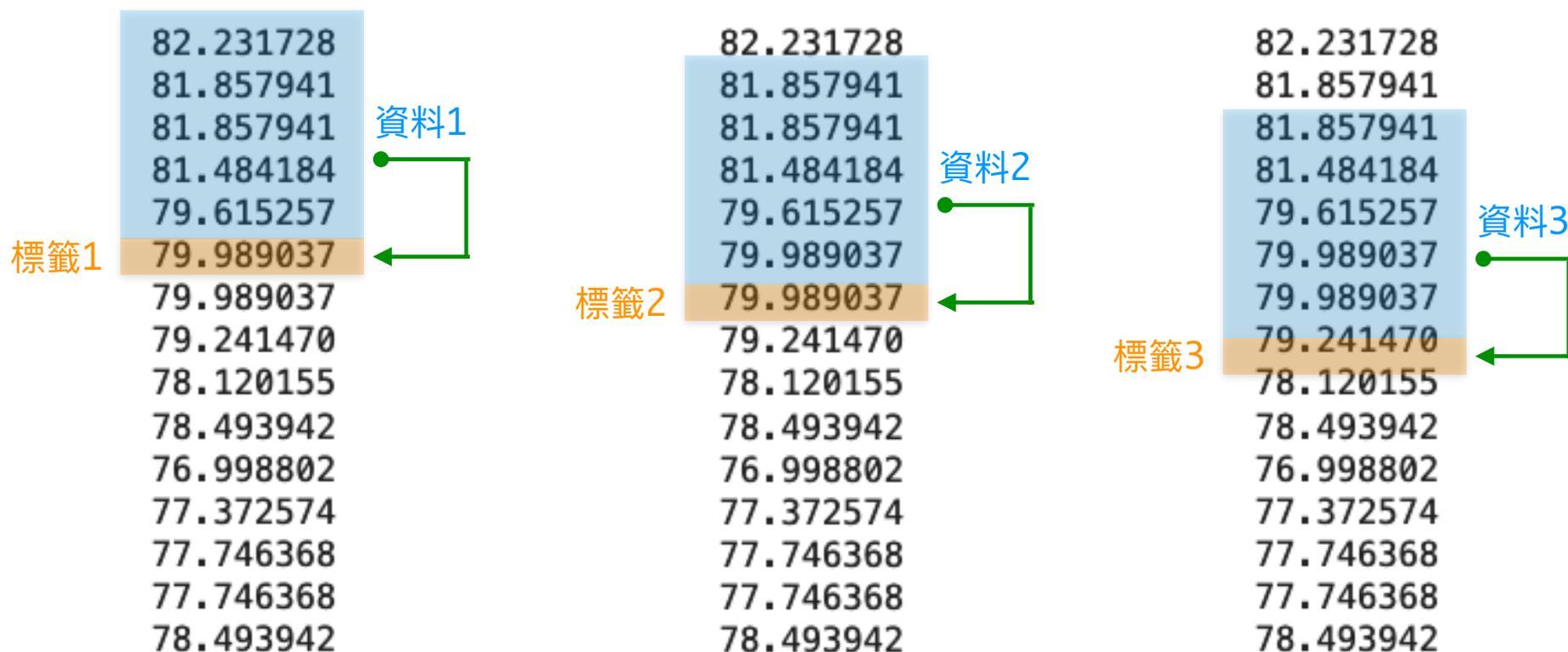
- 訓練AI模型所需**測試資料**：
  - 資料
  - 標籤
- 圖片識別模型：
  - 資料：60000萬張圖片
  - 標籤：每張圖片的類別答案
- 股票預測模型：
  - 資料：許多前n日的股價
  - 標籤：許多第n+1日的股價

資料(圖片)	標籤(圖片類別)
	9
	0
	1
	2
	9
...	...
	5

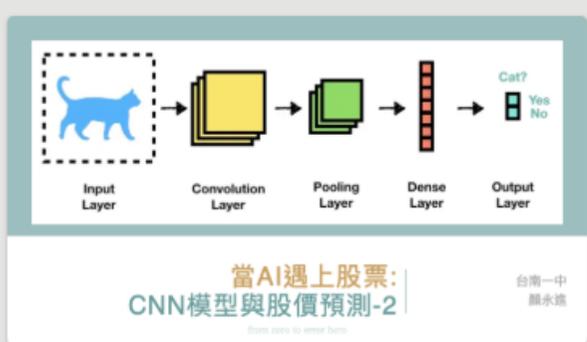
### 3. CNN 股價預測實作：建立測試資料

股票預測模型：

- 資料：前5日股價走勢
- 標籤：第6日股價



# 4. 課程內容 #6



1

## 大綱

1. 對於預測實作1的反省
2. 繼續動玩模型
3. 效能評估
4. 問題的核心

2

## 1. 對於預測實作1的反省

- 1) 使用更多特徵值
- 2) 資料標準化
- 3) 參考資料日數
- 4) 增加模型層數
- 5) 增加訓練epoch

3

## 2. 動手玩模型

- 1) 使用更多特徵值: 加入成交量
- 2) 資料標準化: 使用MinMax
- 3) 參考資料日數: 改為50日
- 4) 增加模型層數: 加入層數
- 5) 增加訓練epoch: 200

4

## 2. 動手玩模型: 使用更多特徵值

獲取特徵值

- 收盤價
- 成交量

取出需要的特徵值

```
data = df[['Date', 'Close', 'Volume']]
data
```

Date	Close	Volume
2013-10-30 00:00:00+08:00	80.231743	10582000
2013-10-31 00:00:00+08:00	81.857941	22472000
2013-11-01 00:00:00+08:00	81.857941	10837000
2013-11-04 00:00:00+08:00	81.484181	9601000
2013-11-05 00:00:00+08:00	79.615257	20304000
...	...	...
2023-10-24 00:00:00+08:00	544.000000	15007614
2023-10-25 00:00:00+08:00	544.000000	16526305
2023-10-26 00:00:00+08:00	531.000000	26188348
2023-10-27 00:00:00+08:00	533.000000	15679643
2023-10-30 00:00:00+08:00	532.000000	21817003

2641 rows x 2 columns

5

## 2. 動手玩模型: 使用更多特徵值

6

## 2. 動手玩模型: 資料標準化

```
from sklearn.preprocessing import MinMaxScaler
scalerX = MinMaxScaler(feature_range=(0, 1))
scalerY = MinMaxScaler(feature_range=(0, 1))
data = data[['Volume', 'Close']]
all_x = data[['Volume']]
all_y = data[['Close']]
scalerX.fit_transform(all_x.values)
scalerY.fit_transform(all_y.values.reshape(-1, 1))
```

Date	Close	Volume
2013-10-31 00:00:00+08:00	81.857948	81.857948
2013-11-01 00:00:00+08:00	81.857948	81.484169
2013-11-04 00:00:00+08:00	81.484169	79.615265
2013-11-05 00:00:00+08:00	79.615265	

```
[[0.01217892],
 [0.01217892],
 [0.01217931],
 [0.00833255]]
```

7

## 2. 動手玩模型: 參考資料日數

• 往前參考日數由5日增加至50日

```
import numpy as np
# 以前N天的收盤價來預測未來股價
previousNDays = 30
x_data, y_data = [], []
for i in range(len(scalerX.all_x) - previousNDays):
    x = scalerX.all_x[i:i+previousNDays]
    y = scalerY.all_y[i:i+previousNDays]
    x_data.append(x)
    y_data.append(y)
# 將資料型態轉換為模型所要求的numpy array
x_data, y_data = np.array(x_data), np.array(y_data)
```

8

## 2. 動手玩模型: 增加模型層數

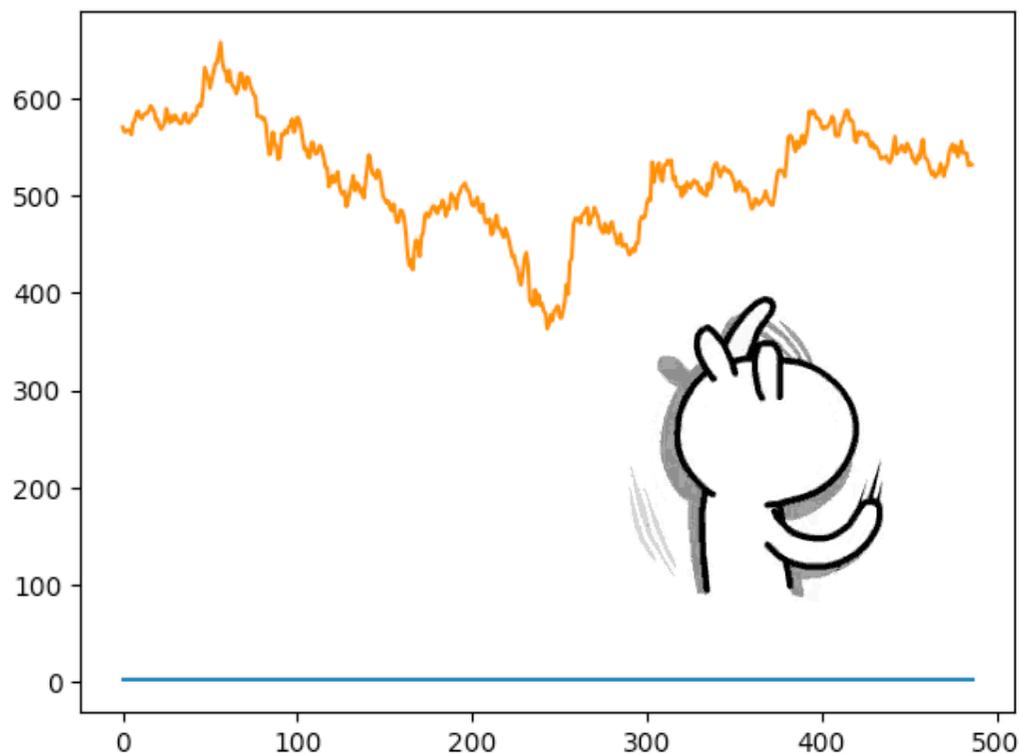
```
# 模型1
import tensorflow as tf
model = tf.keras.Sequential()
model.add(tf.keras.layers.Conv2D(filters=32, kernel_size=3, input_shape=(previousNDays, 1, 1)))
model.add(tf.keras.layers.Flatten())
model.add(tf.keras.layers.Dense(1, activation='sigmoid'))

# 模型2
import tensorflow as tf
model = tf.keras.Sequential()
model.add(tf.keras.layers.Conv2D(filters=32, kernel_size=3, input_shape=(previousNDays, 1, 1)))
model.add(tf.keras.layers.Conv2D(filters=32, kernel_size=3, input_shape=(previousNDays, 1, 1)))
model.add(tf.keras.layers.Flatten())
model.add(tf.keras.layers.Dense(1, activation='sigmoid'))
```

9

# 1. 對於預測實作1的反省

- 1) 使用更多特徵值
- 2) 資料標準化
- 3) 參考資料日數
- 4) 增加模型層數
- 5) 增加訓練epoch



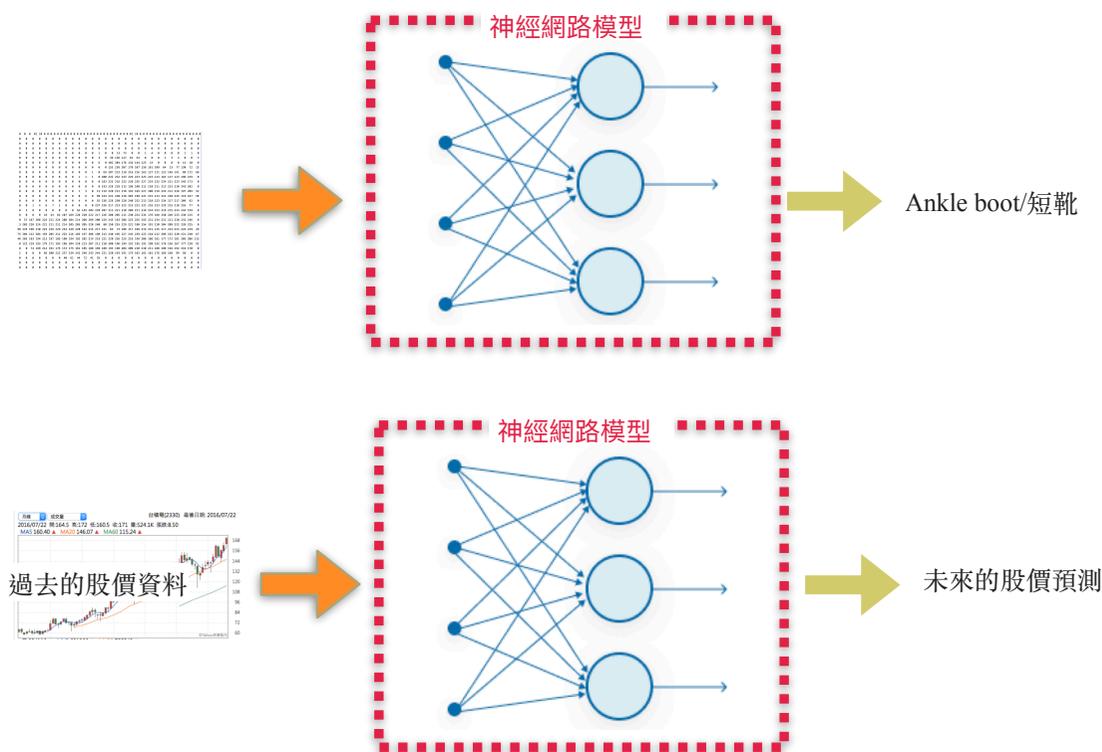
## 2. 動手玩模型

- 1) 使用更多特徵值: 加入成交量
- 2) 資料標準化: 使用MinMax
- 3) 參考資料日數: 改為30日
- 4) 增加模型層數: 加入層數
- 5) 增加訓練epoch: 200

## 4. 問題的核心

CNN的特性

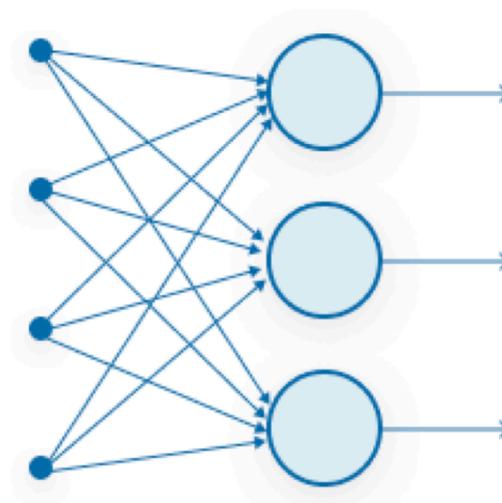
- 1) 模擬人類視覺功能
- 2) 以卷積核找出資料(圖形)的特徵
- 3) 不關心這些特徵出現的順序
- 4) 不適合處理時間事件序列資料(如文字生成、翻譯、股價)



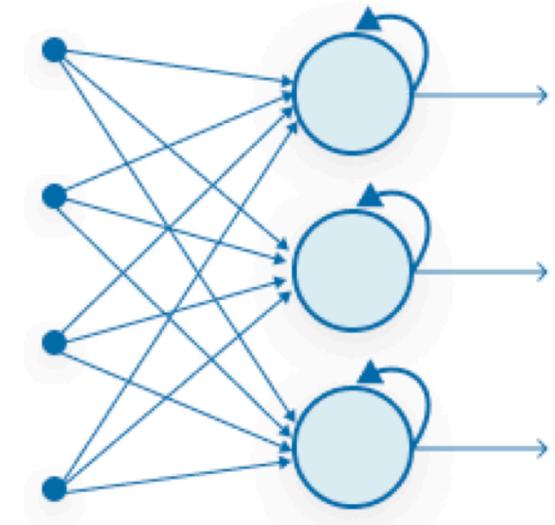
## 4. 問題的核心

適合處理時間序列資料的模型: RNN

- RNN: 遞迴類神經網路 (Recurrent Neural Network)
- 為一個環狀的神經網路，每個時間步驟都有一個輸入，並且可以記住之前的資料特徵。
- 主要用於處理時間序列數據，如語音識別、文本生成、股價預測



CNN



RNN

# 5) 專題成果發表

4. 課程簡介

- 1) padlet: 2023AI選修
- 2) YouTube: 線上成發

# 感謝聆聽



一本正經的胡說八道