

Machine Learning

Lecture 4 Machine Learning Concept

Chen-Kuo Chiang (江振國)
ckchiang@cs.ccu.edu.tw

中正大學 資訊工程學系

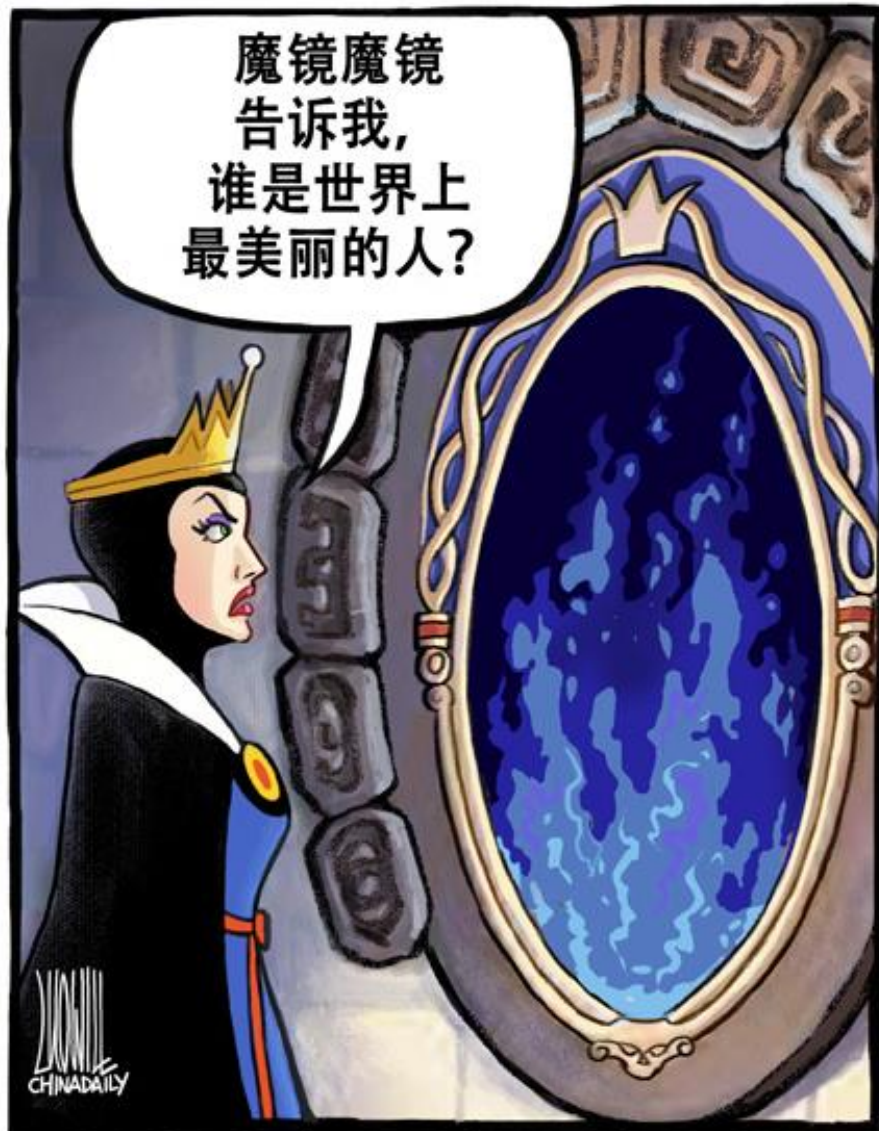
Section Summary

- 如何將目標問題轉為一個機器學習的問題？
- 如何定義訓練資料？
- 如何衡量模型的好壞？
- 模型如何學習？
- 如何產生資料的表示法或特徵值？
- 實際問題的試煉



從白雪公主之魔鏡認識機器學習

魔鏡的起源



魔鏡的學習歷程

- 什麼是美麗？



定義“美麗”

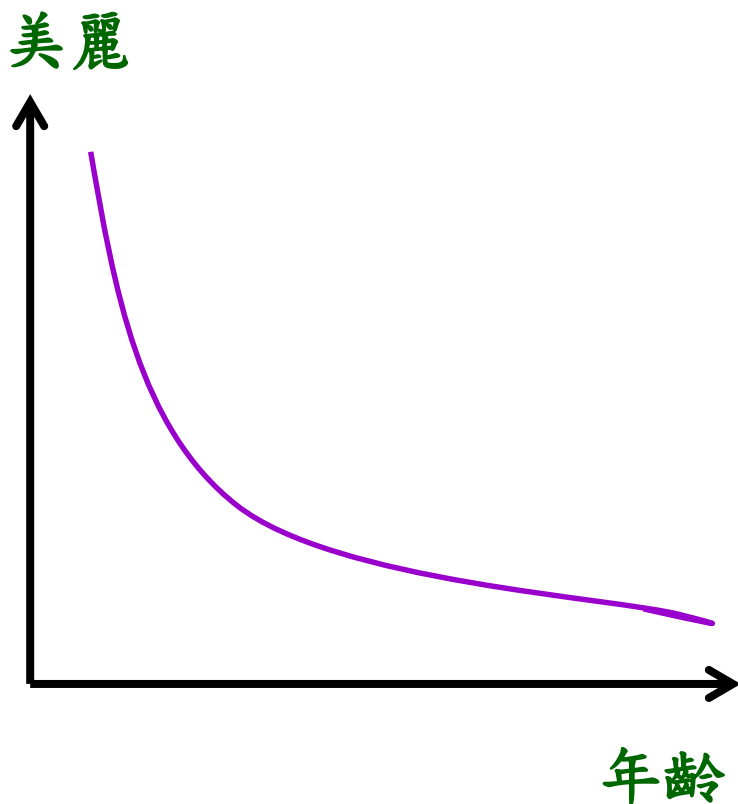
- 列出影響美麗的條件，並數值化
 - 眼睛的大小
 - 臉的大小形狀
 - 膚質
 - 氣質
 - 年齡
 - ...



2017年度百大美女評選

學習各種條件與美麗的關係

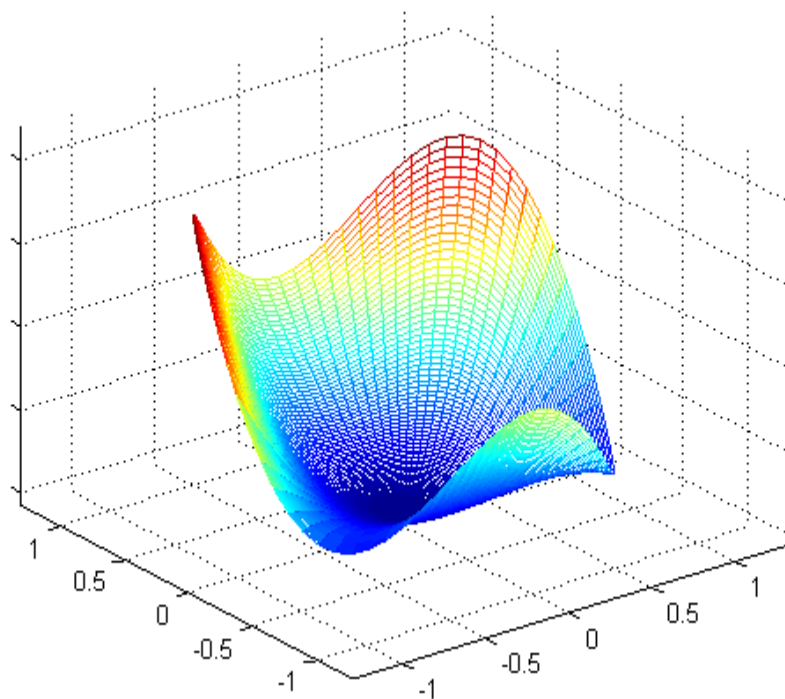
- 年齡 VS 美麗



- 所有特徵值 VS 美麗

特徵值

眼睛
臉型
膚質
氣質
年齡



找出特徵值與美麗的函數關係!

魔鏡的答案

- 「魔鏡啊魔鏡~世上最美的人是誰?」

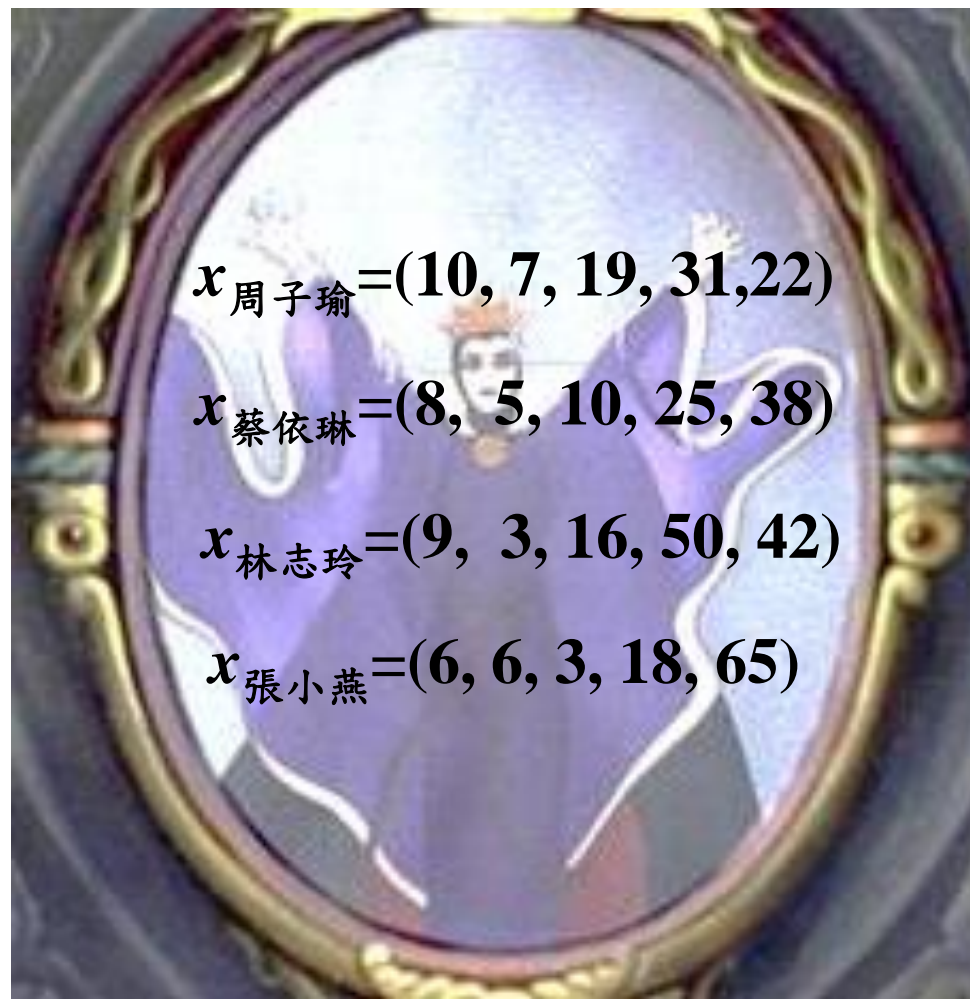


資料蒐集的重要性

- 需要先蒐集大量的資料，才能分析資料與答案的相關性

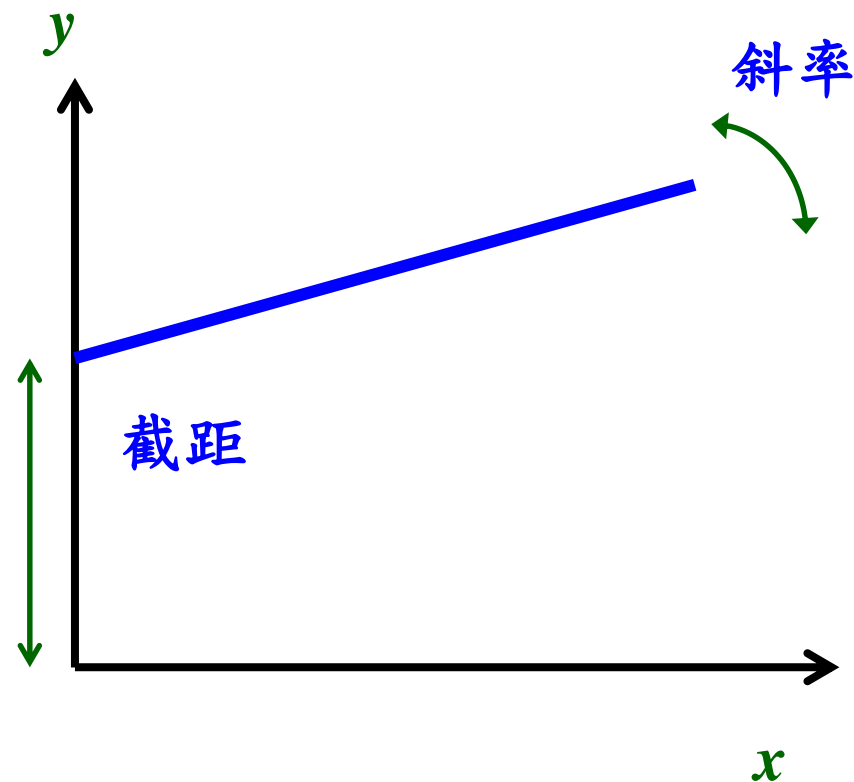
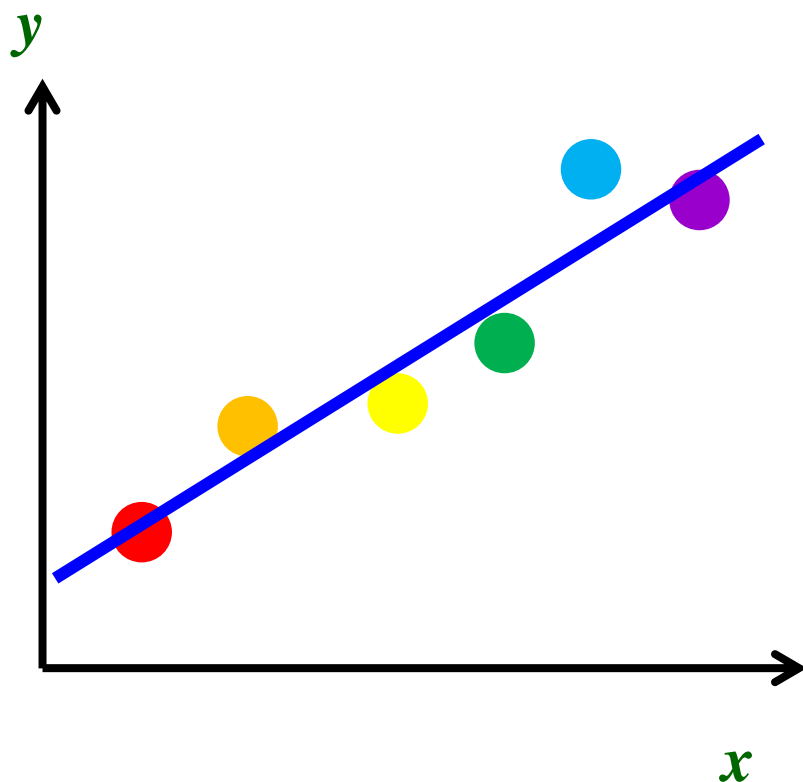
$$x_{\text{周子瑜}} = \begin{pmatrix} 10 \\ 7 \\ 19 \\ 31 \\ 22 \end{pmatrix} \begin{matrix} \leftarrow \text{眼睛大小} \\ \leftarrow \text{臉型} \\ \leftarrow \text{膚質} \\ \leftarrow \text{氣質} \\ \leftarrow \text{年齡} \end{matrix}$$

$$y_{\text{周子瑜}} = 95$$



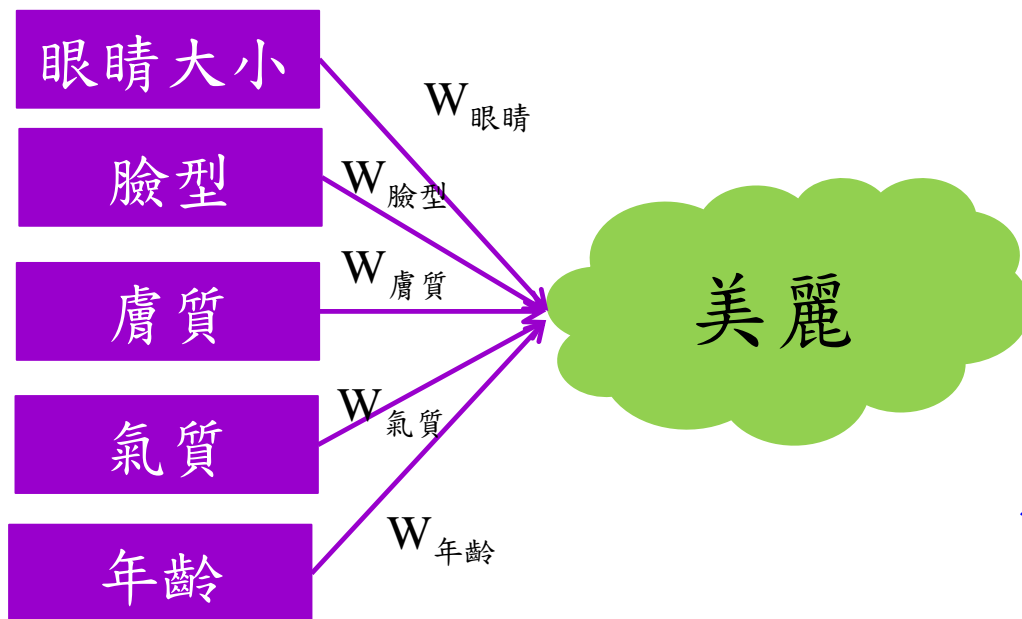
資料的函數表示

- 如何找到最符合資料分布的函數表示？
- 迴歸分析



美麗的關鍵

- 所有條件一樣重要嗎？
- 不一樣重要的話，加上權重



機器學習：推動推桿，每次調整權重，讓函數吻合資料分佈

魔鏡的學習

大量的資料特徵值



大量資料的答案

輸入

函數模型

輸出



調整參數，控制函數變化

觀察結果，吻合資料分佈




怎麼判斷函數學的好？

- 模型的答案要吻合真實的答案

	周子瑜	蔡依琳	林志玲	張小燕
真實答案	95	85	92	65
模型結果	98	83	87	66
誤差	-3	2	5	-1

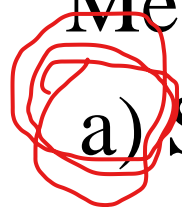
- 使用損失函數 (Loss Function)


$$\text{損失函數} = (-3)^2 + 2^2 + 5^2 + (-1)^2$$

- 常見的損失函數: SSE、SAD、MSE、MAE

Fun Time

- 下列四種損失函數的大小順序為何? Sum of Squared Error (**SSE**)、Sum of Absolute Error (**SAE**)、Mean of Squared Error (**MSE**)、Mean of Absolute Error (**MAE**)?



a) $SSE > SAE > MSE > MAE$

b) $SAE > SSE > MAE > MSE$

c) $MAE > MSE > SAE > SSE$

d) $MSE > MAE > SSE > SAE$

求得最佳解的方法??

- **最佳化方法**—盡可能縮小模型結果與真實答案的差距，得到一個最小誤差的一組權重值為解答
- **如何求得最佳解??**
 - 每次只推動拉桿一點點，看誤差有沒有縮小
 - 有縮小表示拉桿推動的方向對了，再推一點點
 - 如果誤差無法縮小，則推動其他根拉桿看看
 - 如此反覆的操作，直到誤差下降不了為止
 - 此方法稱為**坡度法**
- **梯度下降法**
 - 同時推動多根拉桿，讓誤差縮小

$$x + y + 1 = 0$$

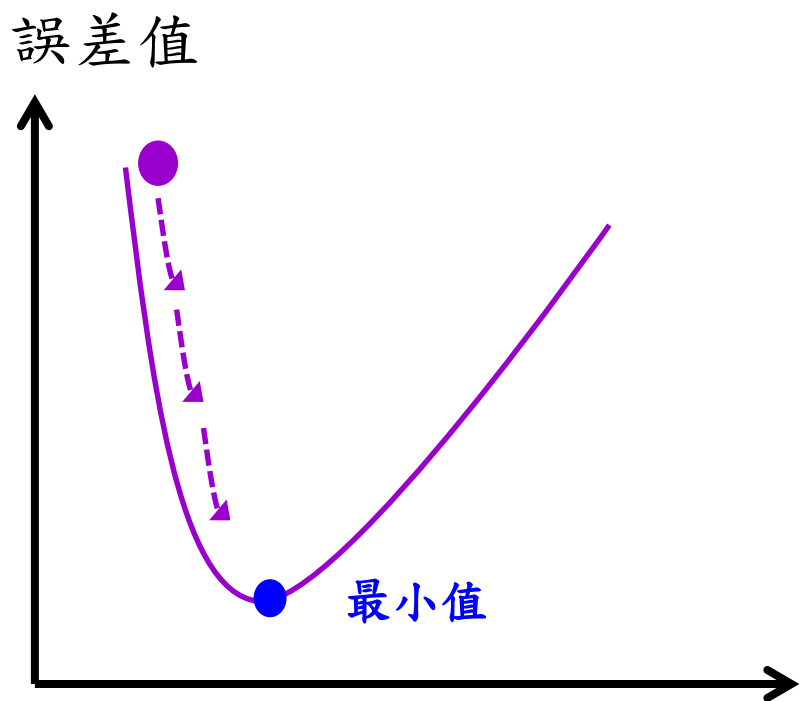
$$2x + y + 1 = 0$$

$$3x + y + 1 = 0$$

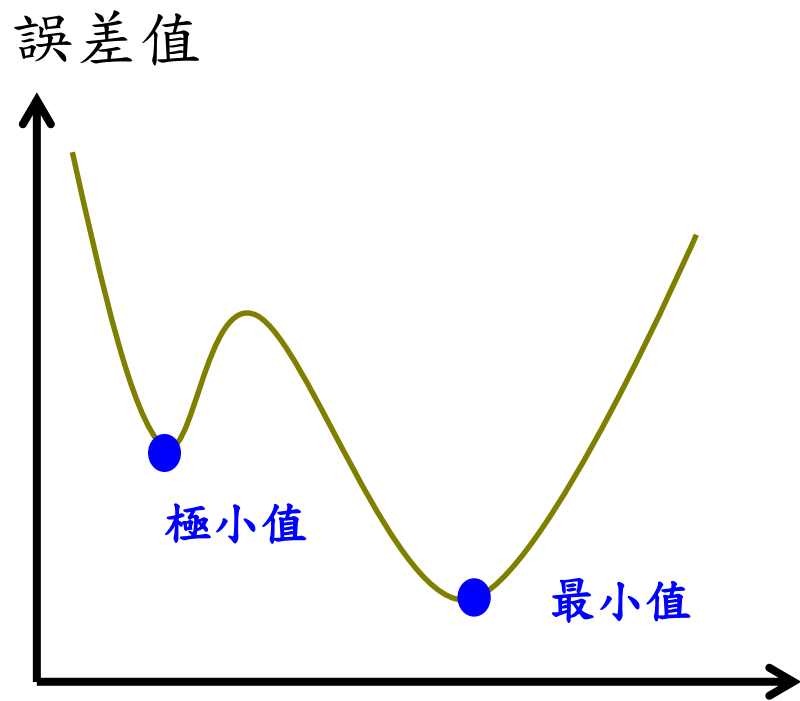
...

區域最佳解與全域最佳解

- 求到的最終解是否是最佳解？



模型參數



模型參數

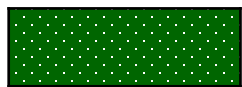
模型的一般性

- **訓練資料**-用於學習與決定模型參數的資料
- **模型的一般性**
 - 如果在看過的資料學習的答案很正確，在沒看過的資料也希望預測的很準確
- **測試資料**
 - 在模型訓練完之後，用模型沒看過的資料當測試資料，丟進去計算模型的準確性
- **機器學習的準則**
 - 手上的所有資料分成訓練資料&測試資料兩大塊
 - 訓練模型時，不可使用測試資料

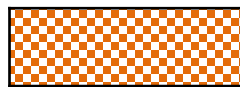
交叉驗證

- 有可能發生取得的訓練資料，剛好偏向某一個結果。
- 交叉驗證用於產生多種訓練/測試資料組合，求得公允的測試結果。

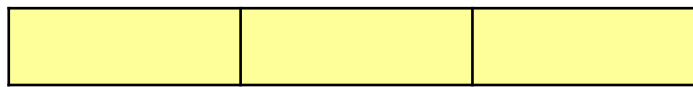
取得的所有資料



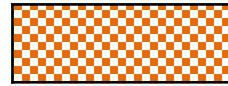
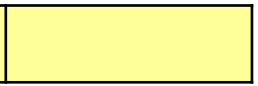
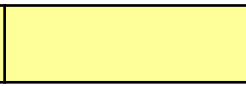
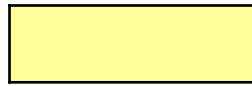
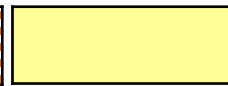
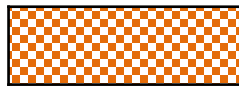
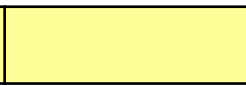
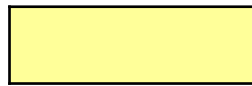
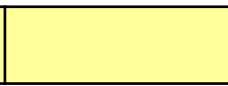
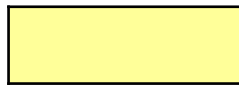
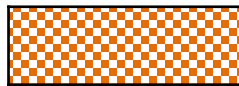
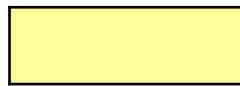
測試資料



驗證資料



訓練資料

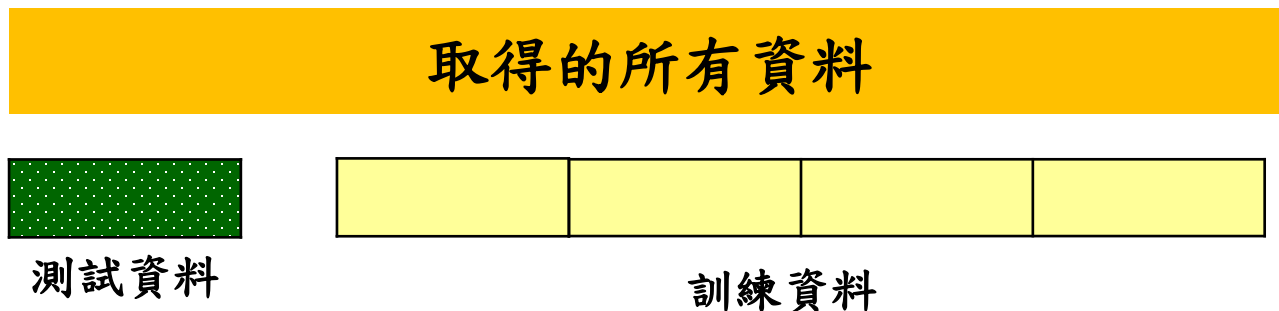


交叉驗證：

利用四種訓練資料分別計算預測誤差，再計算誤差平均值

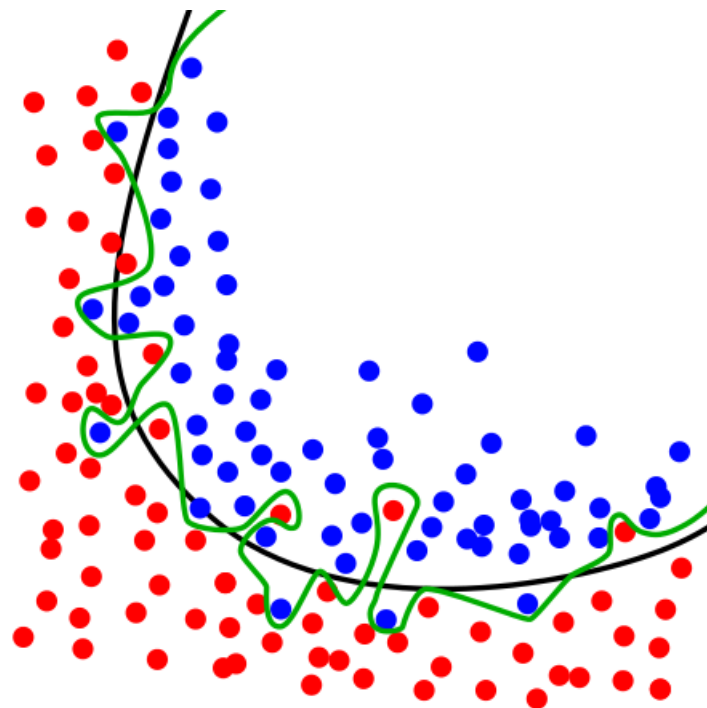
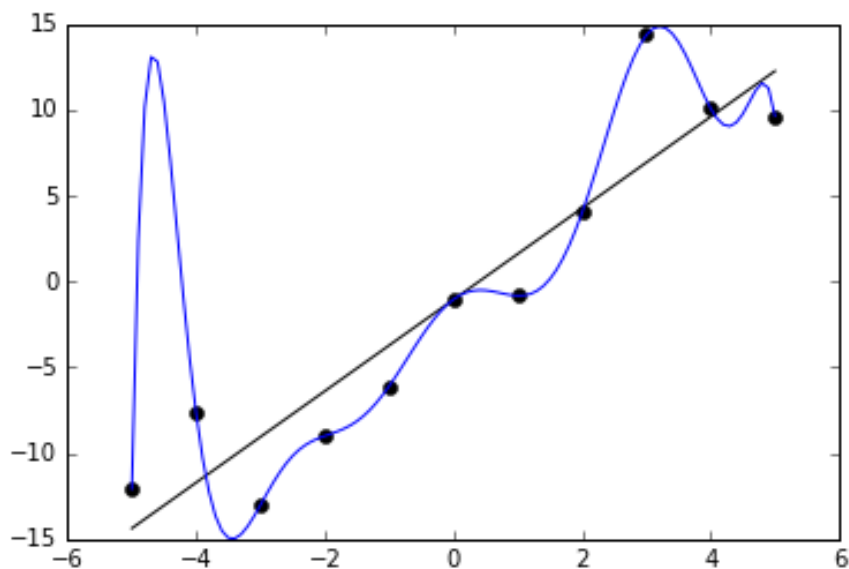
模型訓練與測試

- 經過交叉驗證效果良好的模型，再利用訓練資料+驗證資料，作為訓練資料重新訓練模型
 - 訓練資料量增加，模型訓練效果更好
- 模型結果評估
 - 訓練準確率-以訓練資料訓練模型、以訓練資料測試模型
 - 驗證準確率-以訓練資料訓練模型、以驗證資料測試模型
 - 測試準確率-以訓練資料訓練模型、以測試資料測試模型



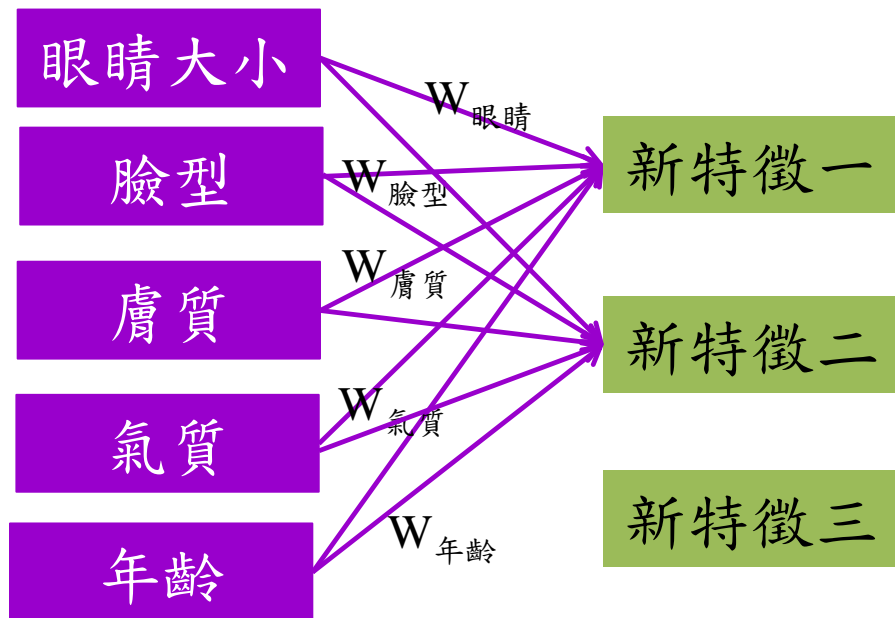
過度擬合(Overfitting)

- 當模型過度屈從訓練資料的分佈，導致測試資料的預測準確很差，稱為過度擬合



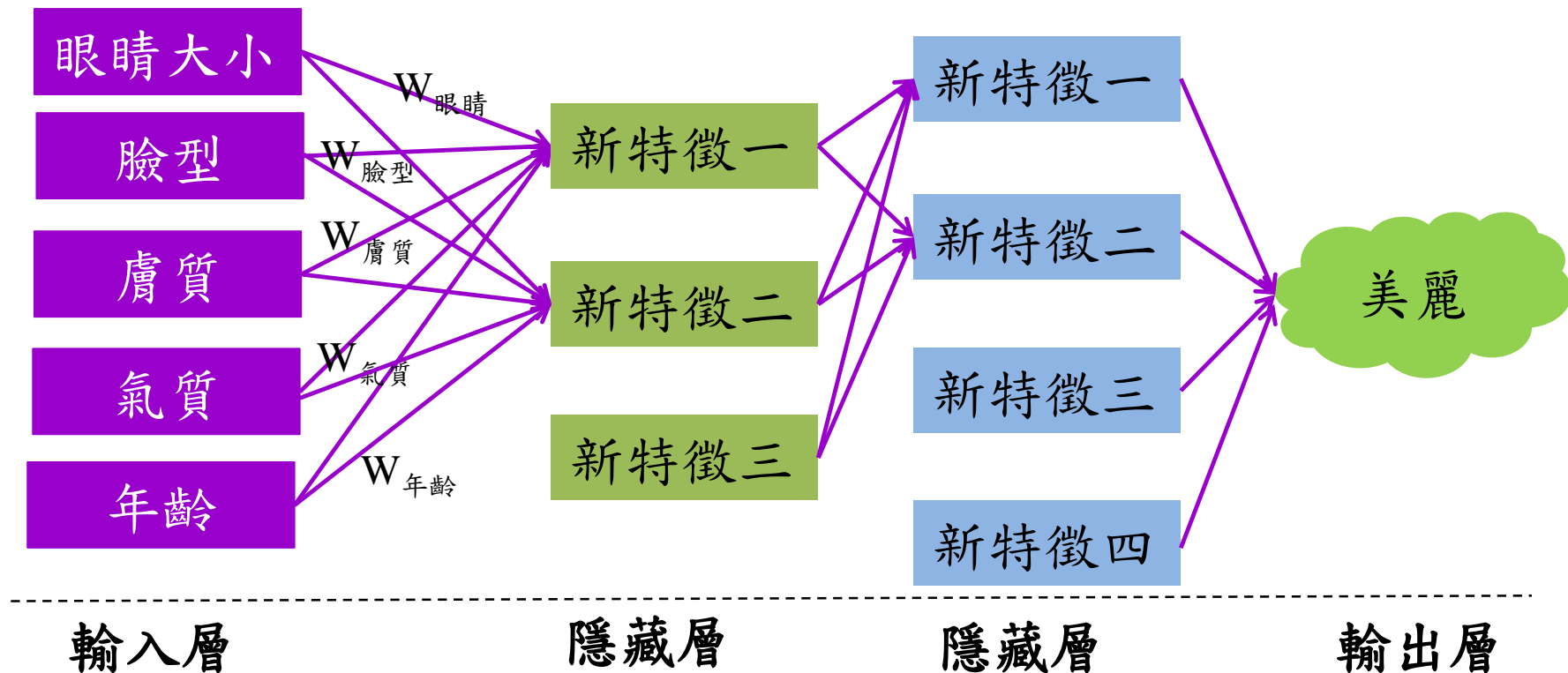
建立新的特徵值

- 利用特徵值與權重的加總，產生新的特徵



類神經網路

- 利用特徵值的組合，產生更多的新特徵組合，用來預估結果



新特徵產生與線性轉換

- 線性轉換

- 產生新的特徵時，是利用特徵乘上權重的結果再相
- 重複多次的乘法與加法還是只算一次的乘法與加法

$$\begin{bmatrix} s \end{bmatrix} = p \cdot \begin{bmatrix} b_1 \end{bmatrix} + q \cdot \begin{bmatrix} b_2 \end{bmatrix}, \begin{bmatrix} t \end{bmatrix} = u \cdot \begin{bmatrix} b_1 \end{bmatrix} + v \cdot \begin{bmatrix} b_2 \end{bmatrix}$$

$$\begin{bmatrix} v \end{bmatrix} = m \cdot \begin{bmatrix} s \end{bmatrix} + n \cdot \begin{bmatrix} t \end{bmatrix}$$

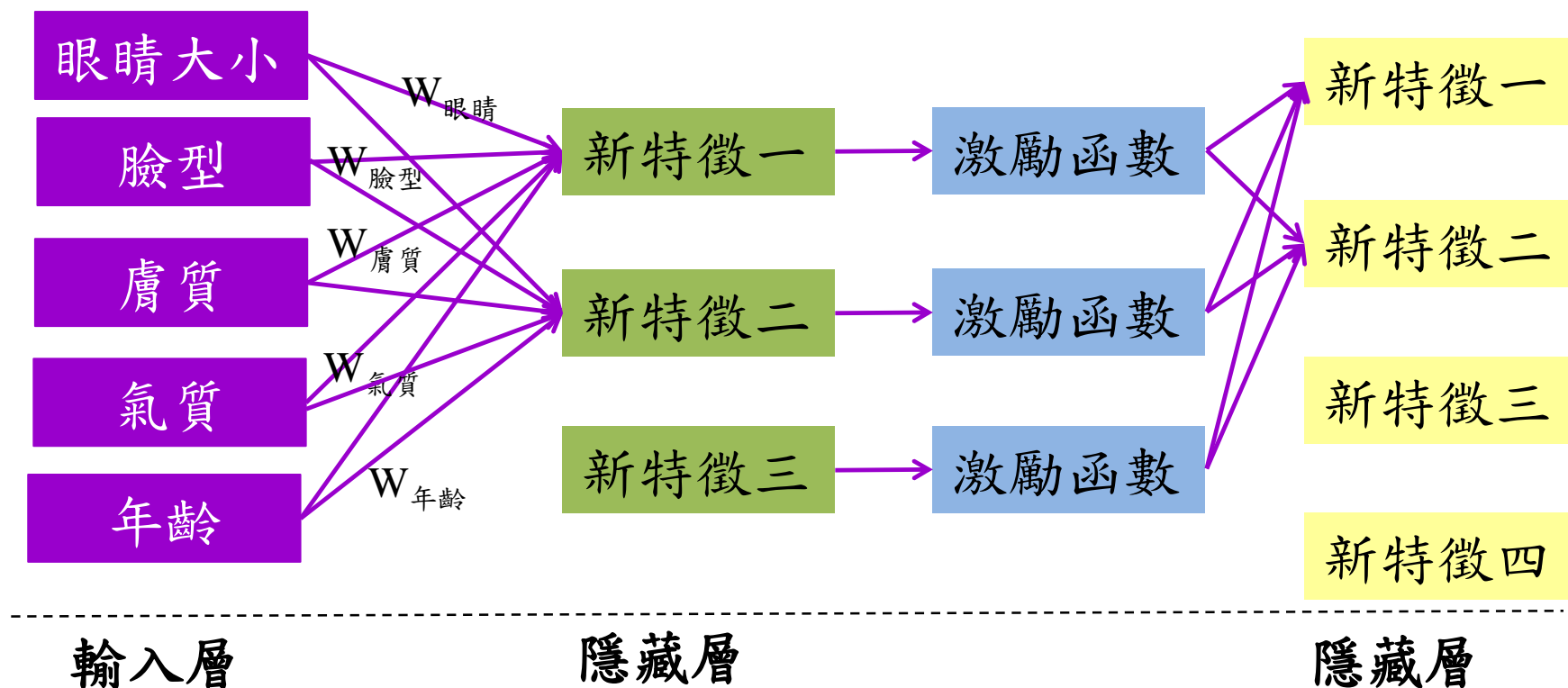
$$= m \cdot (p \cdot \begin{bmatrix} b_1 \end{bmatrix} + q \cdot \begin{bmatrix} b_2 \end{bmatrix}) + n \cdot (u \cdot \begin{bmatrix} b_1 \end{bmatrix} + v \cdot \begin{bmatrix} b_2 \end{bmatrix})$$

$$= (m \cdot p + n \cdot u) \cdot \begin{bmatrix} b_1 \end{bmatrix} + (m \cdot q + n \cdot v) \cdot \begin{bmatrix} b_2 \end{bmatrix}$$

$$= c \cdot \begin{bmatrix} b_1 \end{bmatrix} + d \cdot \begin{bmatrix} b_2 \end{bmatrix}$$

Activation Function (激勵函數)

- 利用 sigmoid 函數作為激勵函數
- 乘上權重值的結果，帶入激勵函數調整大小



非線性轉換

- 非線性轉換-不是只有乘法與加法的轉換方式
- Sigmoid function



$$\theta(-\infty) = 0;$$

$$\theta(0) = \frac{1}{2};$$

$$\theta(\infty) = 1$$

$$\theta(s) = \frac{e^s}{1 + e^s} = \frac{1}{1 + e^{-s}}$$

智慧魔鏡的實作



- 實作步驟
 - 收集決定[美麗]的資料
 - 將資料分成訓練/測試
 - 決定機器學習的模型
 - 將訓練資料丟入模型學習
 - 利用測試資料交叉驗證
 - 安裝攝影機與麥克風
 - 即時拍攝撥放與語音辨識
 - 顯示結果
 - 取特徵/線上更新模型

隨堂練習

- 思考一下，下列問題如何利用機器學習預測？

4105931 機器學習 (Machine Learning)

姓名: _____

學號: _____

Midterm

總分 105 分

中文作答

得分

1. → 如果我想訓練一個“預測碩班畢業會不會進台積電上班”的模型，試說明如何學習這樣的模型？請依序回答: a) (2%)請從這個問題，具體定義模型的輸入資料、輸出資料。b) (2%)請說明如何收集與標記 a)中所提到的輸入輸出資料。c) (2%)請說明應該用什麼模型比較適合解這個問題。d) (4%)試說明如何訓練這樣的模型？請從模型、損失函數、訓練、驗證到測試，說明完整步驟。