機器學習應用於股票市場之介紹

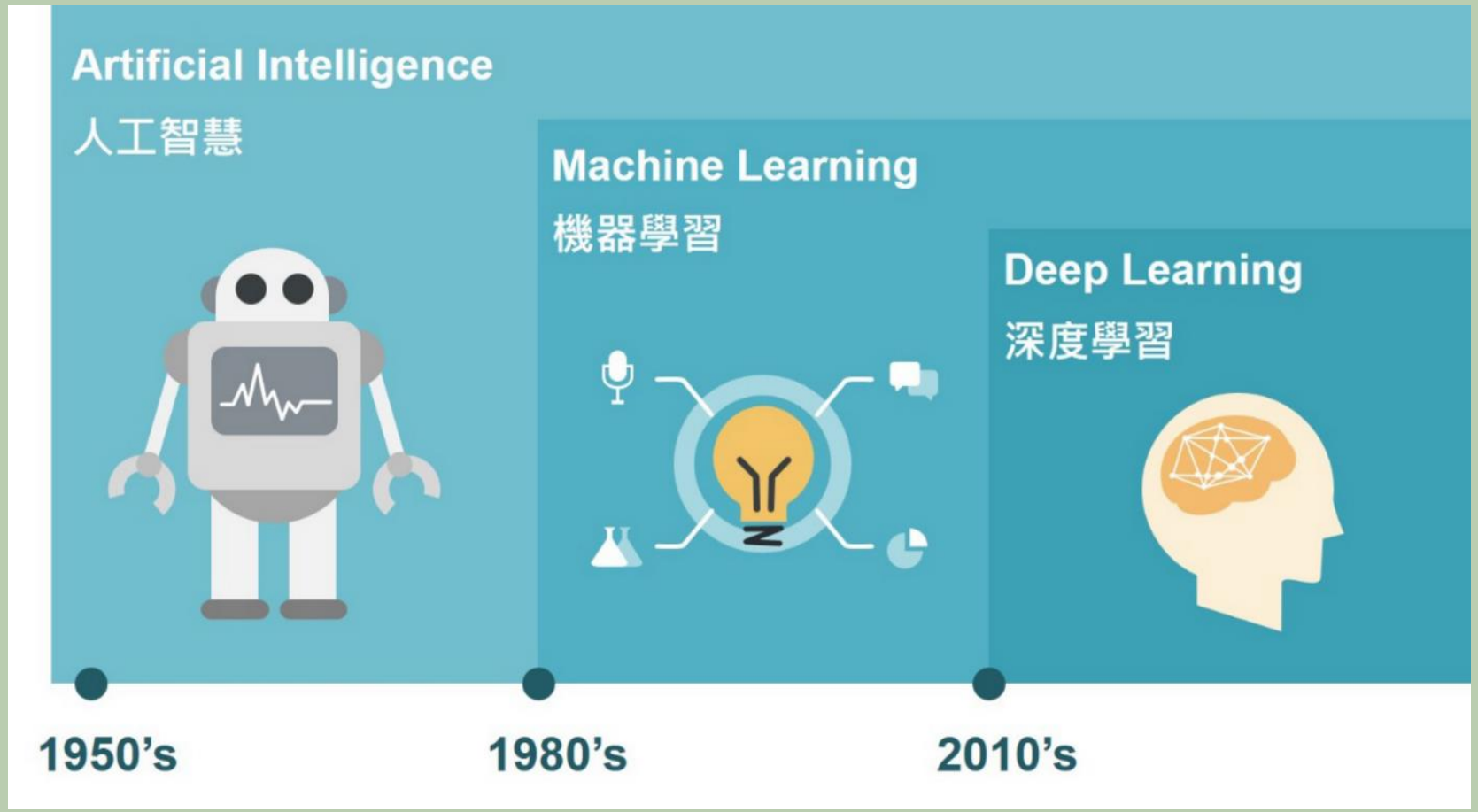
報告人 蘇禎佑

指導教授 李俊宏 教授

目錄

機器學習

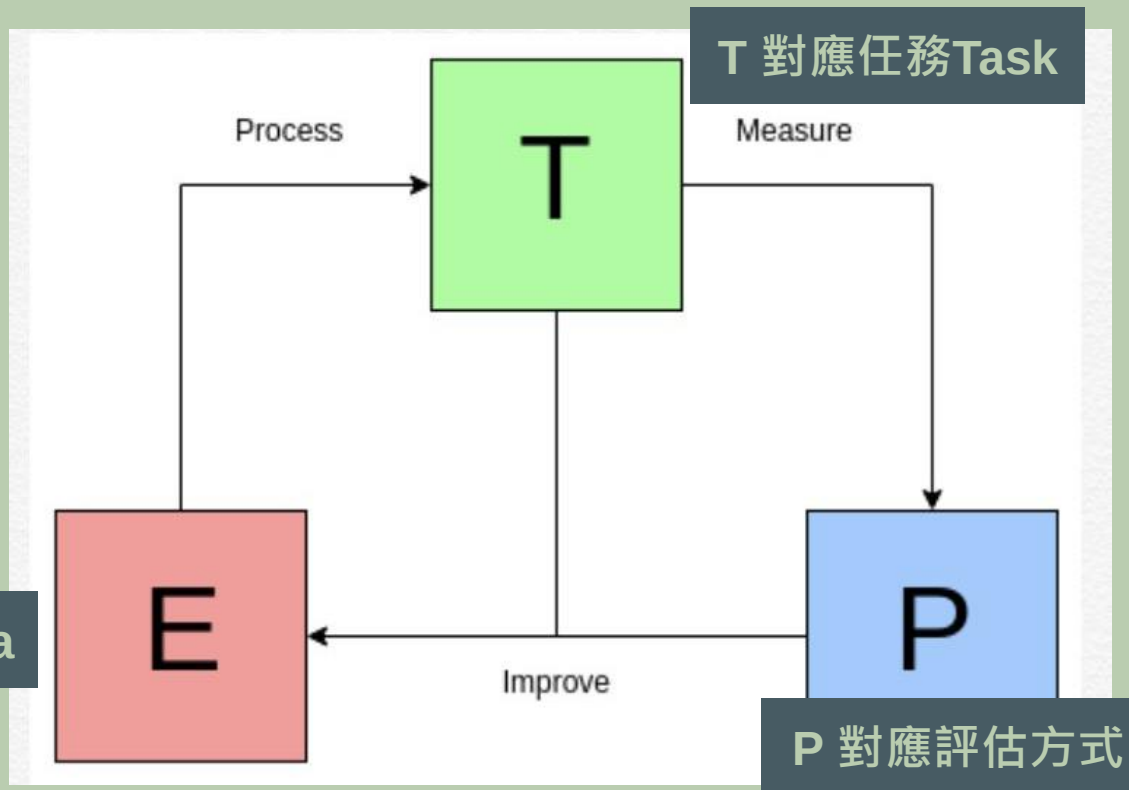
人工智慧、機器學習和深度學習的差別?



機器學習是什麼?

- 主流的定義有很多種，這邊採用 Tom M. Mitchell 在書中的定義

> ℹ A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.



資料來源: Kirk Borne's Twiter
https://twitter.com/kirkdborne/status/1079062765778669571?lang=zh-Hant

機器學習可以用在股票預測嗎?

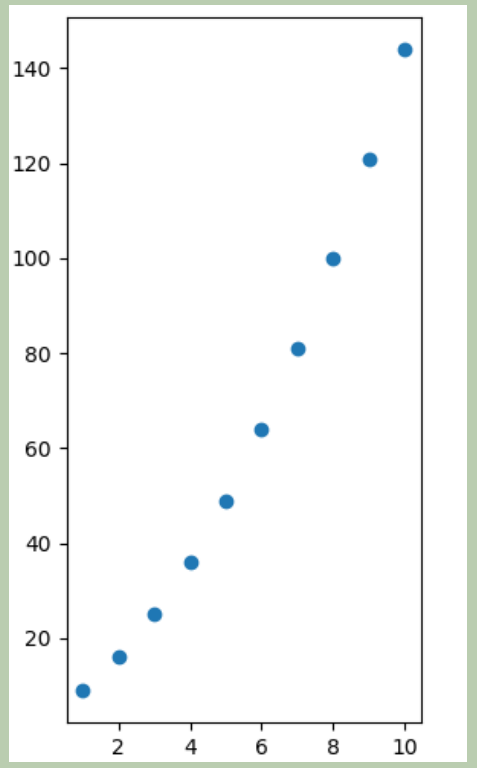
光靠

- 股市的歷史數據
- 技術分析

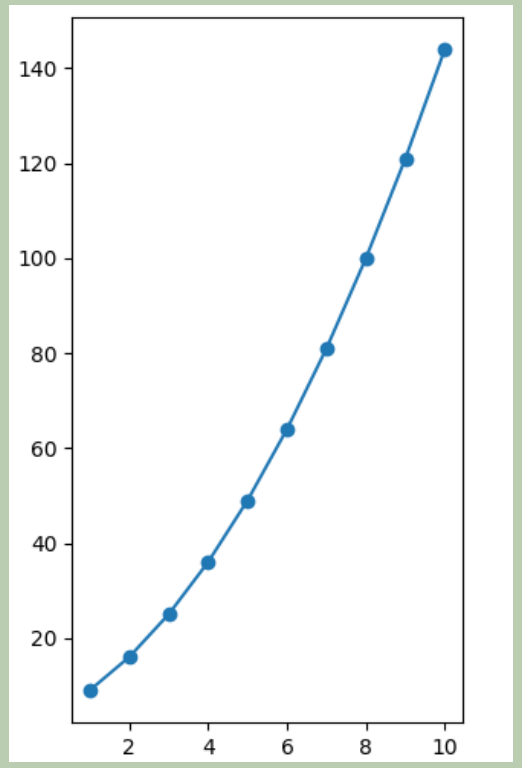
→ **很難**

機器學習可以用在股票預測嗎?

資料



機器學習



背後的規則

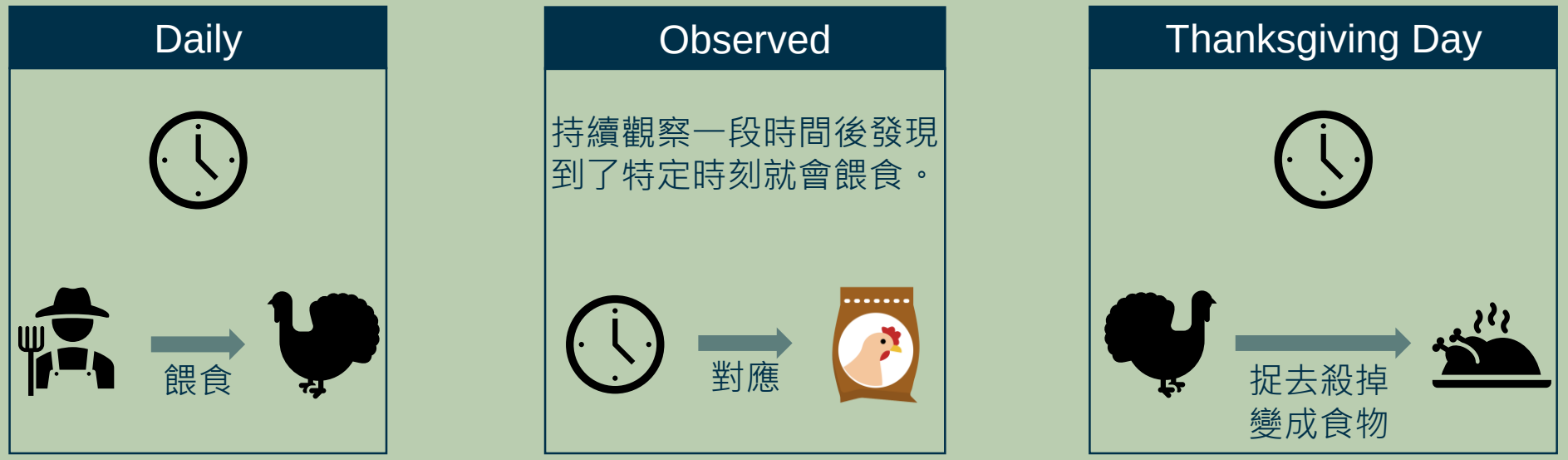
$$y = f(x) = x^2 + 4x + 4$$

透過歷史數據去得知資料背後的規則，
但是股市的規則並不是固定的。

機器學習可以用在股票預測嗎?

- 假設機器用過去十年的資料發現:
- 某個K線只要連續上升兩次，第三次也會上升的機率是80%
- 但有可能這個規則明天就失效了，甚至讓你賠大錢。

火雞問題 Russell's Turkey



機器學習可以用在股票預測嗎?

- 前面說了很多機器學習應用在股票市場的缺點
- 難道機器學習真的很糟嗎?

- 實際上也沒那麼誇張，會這樣描述的原因有二
  - AI 不是萬能的 (許多媒體都過分誇大了)
  - 預測股價(X)，了解股票特性(O)

機器學習的種類

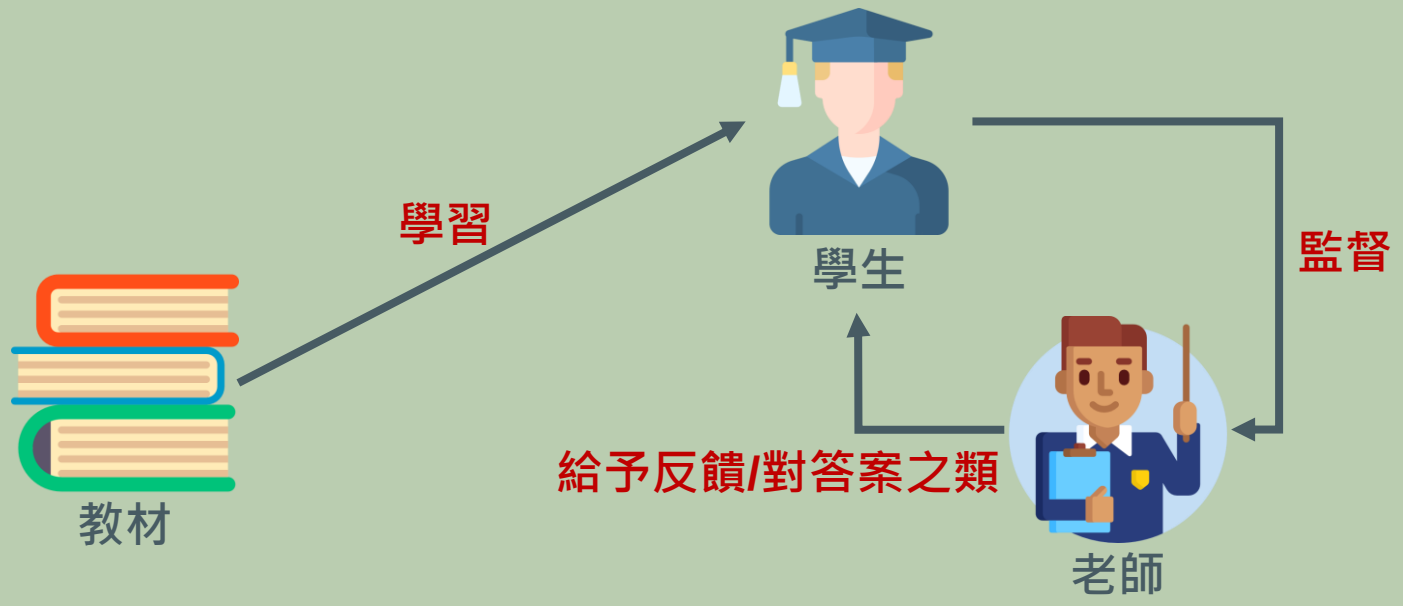
主流的分法大可區分為以下4種

- 監督式學習(supervised learning)
- 非監督式學習(unsupervised learning)
- 半監督式學習(semi-supervised learning)
- 強化式學習(reinforcement learning)

監督式學習

監督式學習

- 監督式學習可以把它想像成有個老師給你教材，並在旁邊監督。
- **教材**對應**大量具有關係的特徵資料**。
- **老師監督**則對應**相對應的標籤結果**，就像是教材的答案。



舉例

- 舉例來說:
- 氣壓、風速、雨量 -> 是否放颱風假 (Y/N)

教材

考卷題目

特徵 (independent variable, 自變量)

氣壓

風速

雨量

考卷答案

標籤 (dependent variable, 因變量)

迴歸問題

迴歸問題

- 監督式學習最常處理的兩個問題：迴歸、分類。
- 而當我們預測值為**連續的數值**，則稱為迴歸。

> 「畫出一條盡量通過這些點的線」

- 而今天會介紹線性迴歸、多元線性迴歸以及多項式迴歸

迴歸問題

- 什麼樣的資料我們會認為它是迴歸問題呢?

雨量

銷量

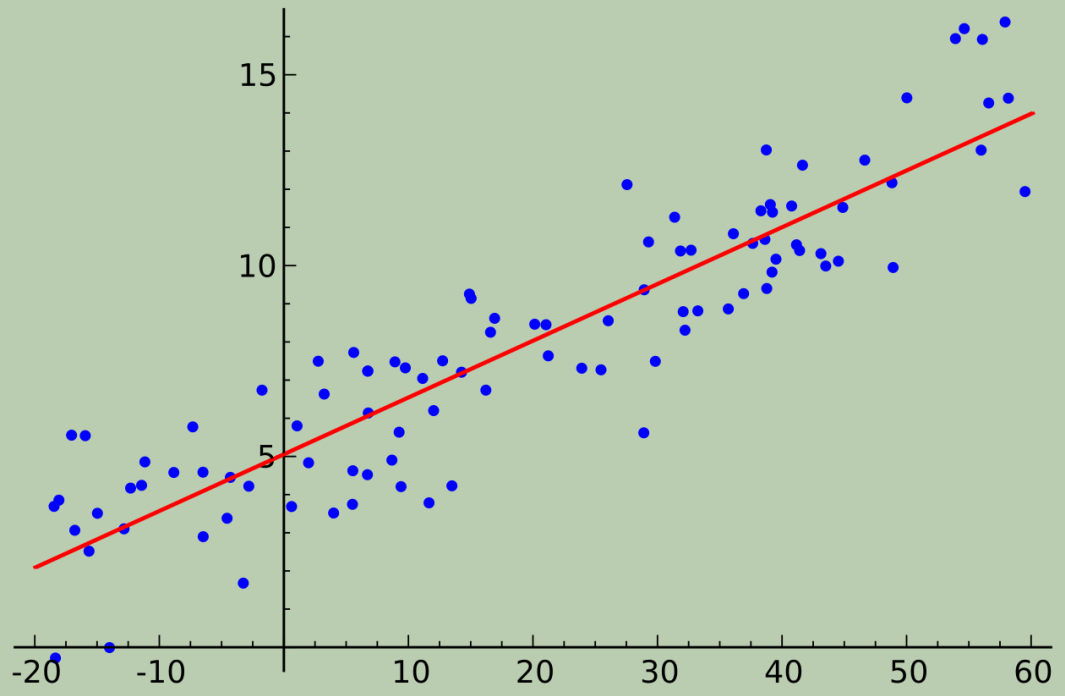
日照天數

股價

資料具有可量化、具有連續性的特性

線性迴歸 Linear Regression

- 機器學習的入門
- **使用一條線擬和所輸入的自變量資料 (就是二元一次方程式)**



$$y = f(x) = wx + b$$

y 因變量
x 自變量
w 係數
b 對應 bias，截距

資料來源: Wikipedia
https://en.wikipedia.org/wiki/Linear_regression

多元線性迴歸 Multiple Linear Regression

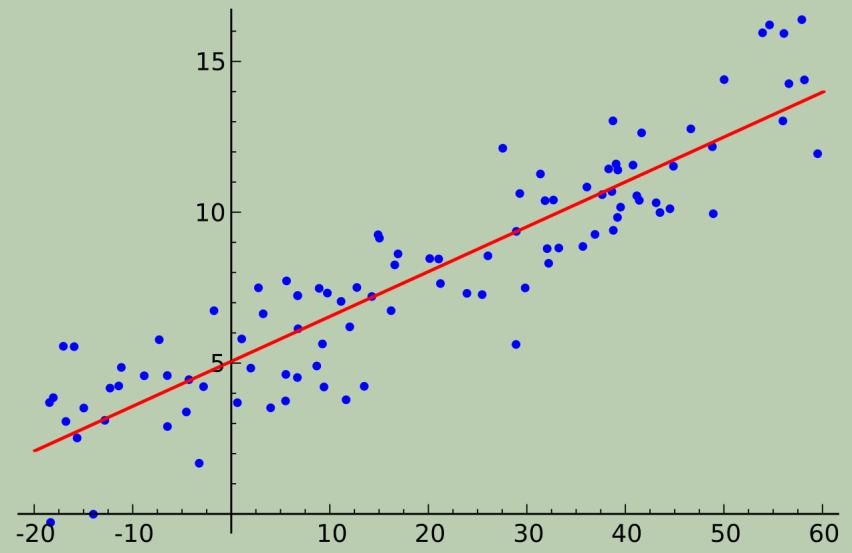
- 很多個二元一次方程式的組合

$$y = f(x_i) = w_1 x_1 + \cdots + w_n x_n + b$$

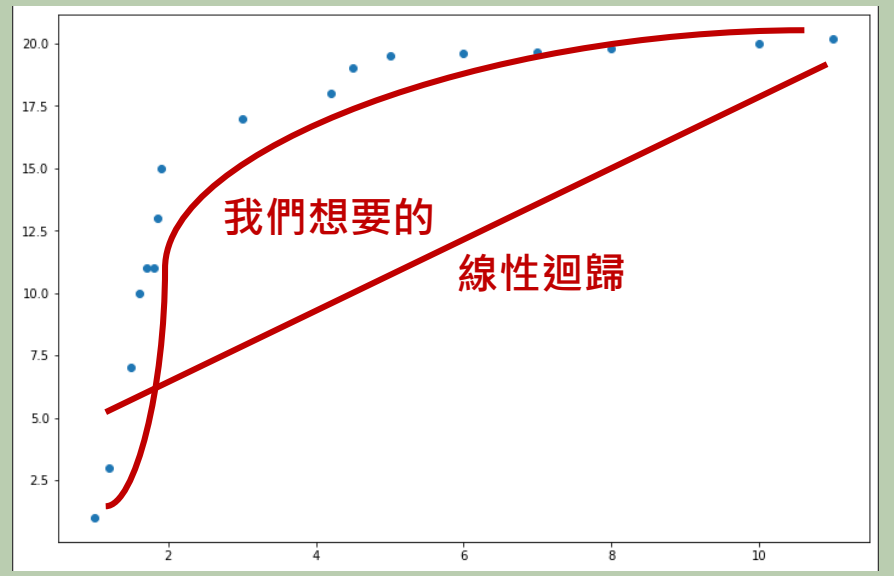
y 對應是否放假，也就是 0 和 1
x 對應不同種類的特徵值
w 則是相對應的特徵權重
b 對應 bias，沒有考慮到的其他因素

多項式迴歸 Polynomial Regression

- 對於線性回歸來說，資料都是很均勻地分布在一條直線上。
- 但現實的資料往往是**非線性的分佈**。



線性迴歸

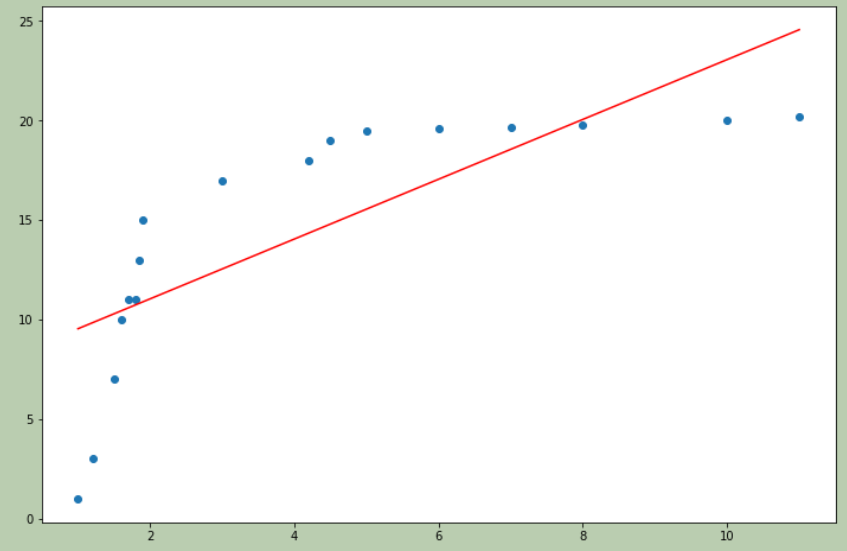


真實世界

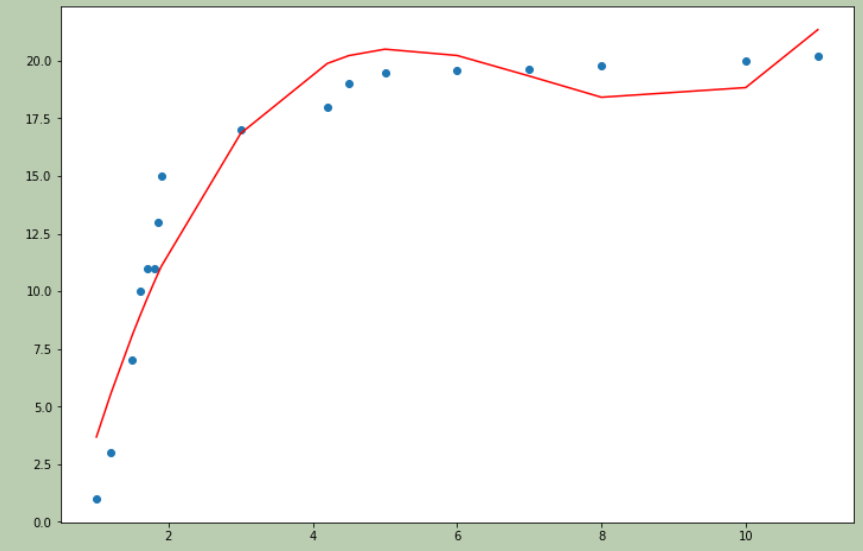
資料來源: Wikipedia https://en.wikipedia.org/wiki/Linear_regression

多項式迴歸 Polynomial Regression

- 而在多項式迴歸中，原始數據仍然不具有線性關係。
- 我們需要尋找一些非線性曲線去擬合。



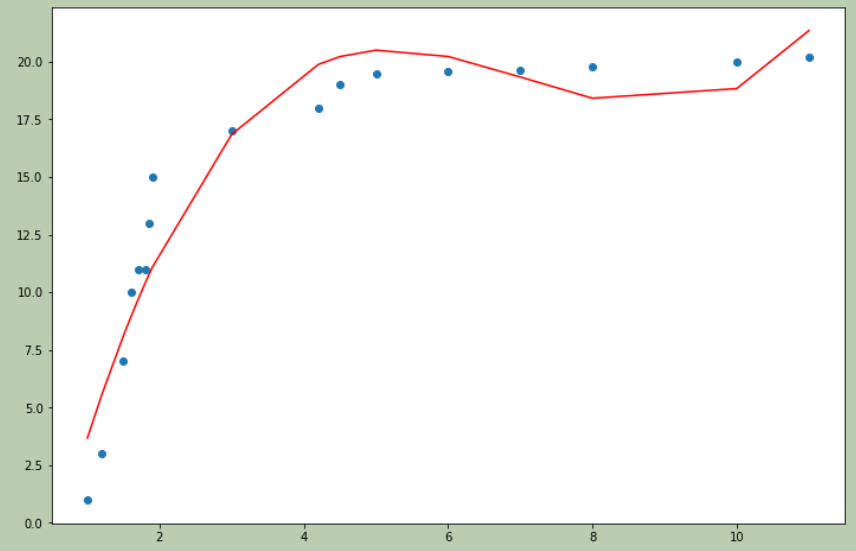
線性迴歸



多項式迴歸

多項式迴歸 Polynomial Regression

- 具體作法，就是將原本的 x 特徵作為基礎，用來建構許多新的特徵。



多項式迴歸

舉例來說:
左圖就是使用一條三次曲線去擬和數據。

$$y = ax^3 + bx^2 + cx + d$$

這樣一來，就變成在解多元線性回歸。
我們再找出 a、b、c、d 。

多項式迴歸 Polynomial Regression

- 實際運算過程，先將資料輸入轉換成如下所示，接著再進行線性迴歸

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

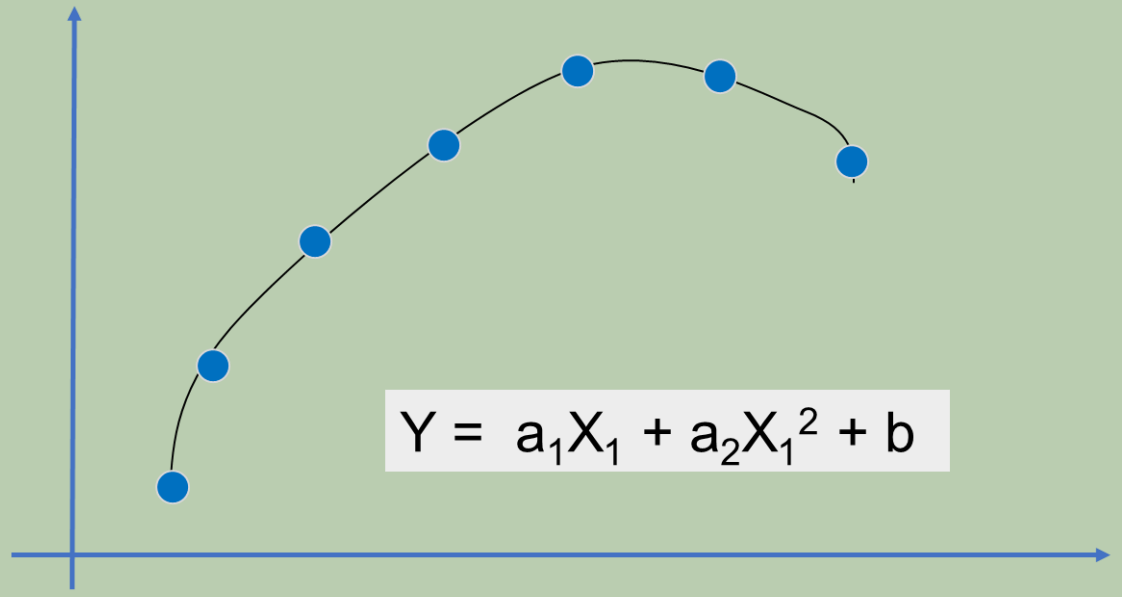
資料來源: Wikipedia
https://en.wikipedia.org/wiki/Polynomial_regression

```
x輸入
x = [ [1]
      [2]
      [3]
      [4] ]
經過處理後
x = [ [1 1  1]
      [1 2  4]
      [1 3  9]
      [1 4 16] ]
```

多項式迴歸 Polynomial Regression

- 也因此多項式迴歸是線性迴歸中的一種變體
- 專門用來針對非線性的問題

$$y = f(x_i) = w_1 {x_1}^1 + w_2 {x_1}^2 + w_3 {x_1}^3 + \cdots + w_n {x_1}^n + b$$



評估方法

- 如何評估迴歸模型的效能?
- 通常會使用迴歸問題的評估指標。

- 以下為列舉三種常用的指標:
- MSE (Mean Square Error): 均方差
- RMSE (Root Mean Square Error): 均方根差
- MAE (Mean Absolute Error): 絕對平均誤差

評估方法

- 這一類的方法都大同小異，我們以 MSE 做為範例來介紹。

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

預測值和實際值差距

- 簡單來說，迴歸方法的評估方式，就是計算預測值與實際值的差距。
- 從而了解這個模型本身的預測誤差為多少。

分類問題

分類問題

- 監督式學習最常處理的兩個問題：迴歸、分類。
- 而當我們預測值為**離散的數值**，則稱為分類。

「把資料分到我指定的幾個類別中」

- 而今天會介紹 Perceptron、Logistic Regression 和 SVM

分類問題

- 什麼樣的資料我們會認為它是分類問題呢?

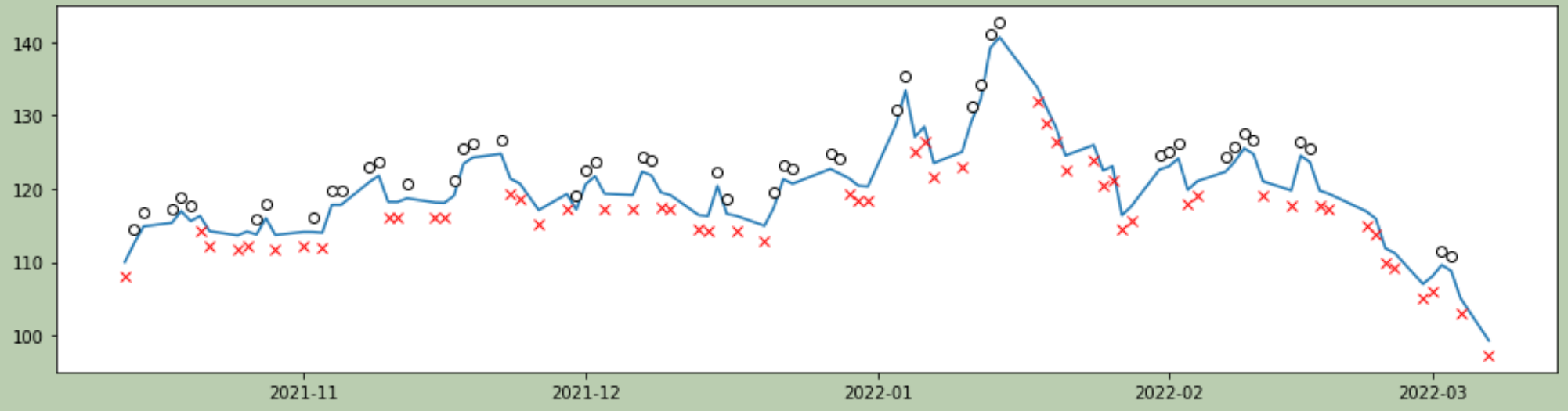
貓狗辨識

就算有著各種不同品種，牠們都會歸類成貓

資料非A即B，沒有模糊空間的中間值

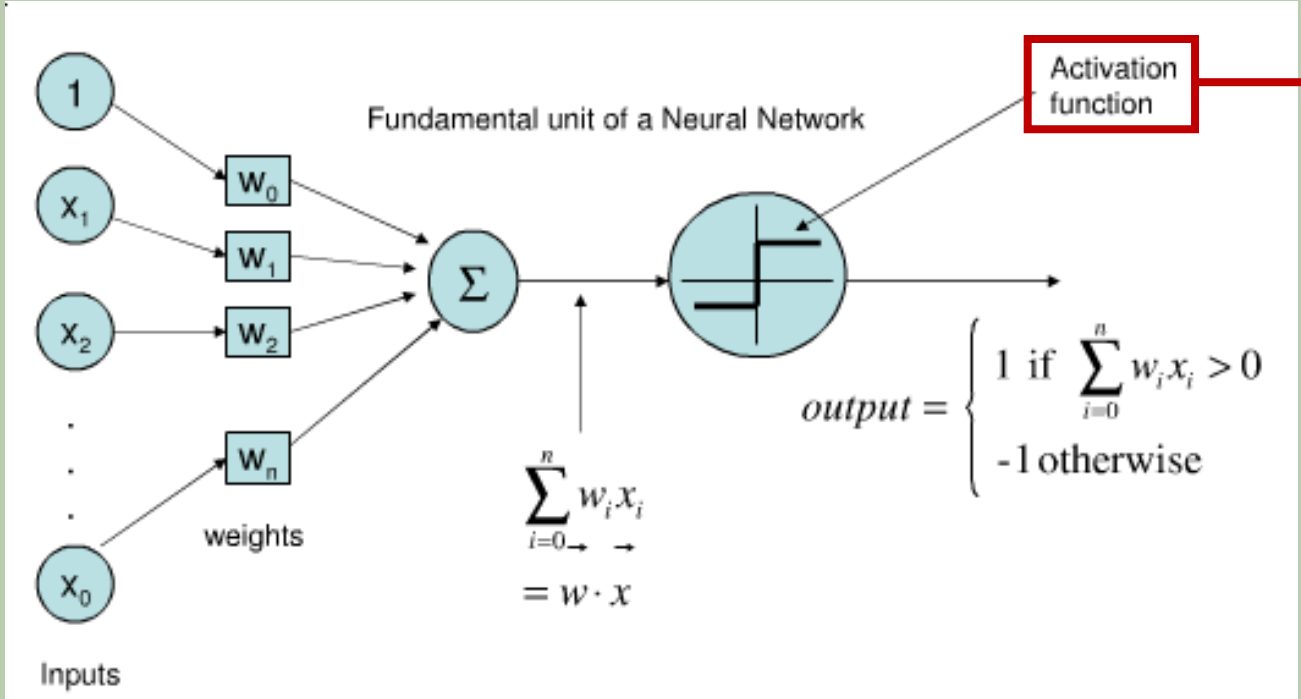
分類問題

- 而該如何從金融領域上，找到這樣的資料特性呢?
- 以股票漲跌為例，只要前一天的股價比今天低，即為漲；反之亦然。

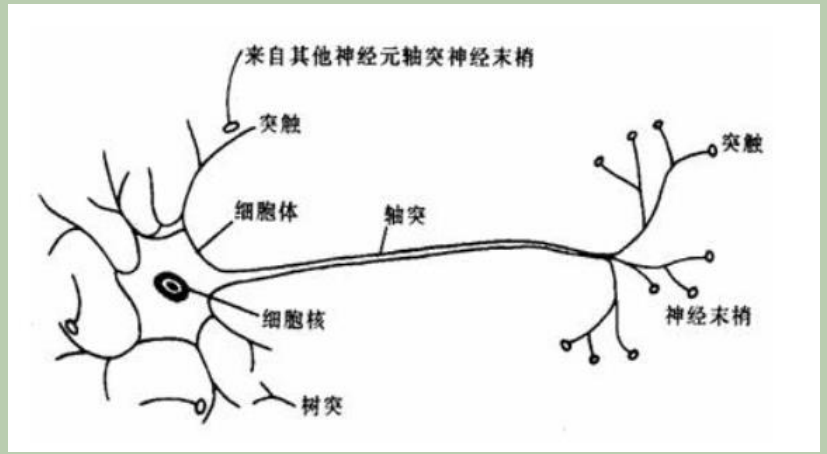


Perceptron

- 機器學習最早被開發出來的演算法
- Perceptron (Perceptron Learning Algorithm, PLA)

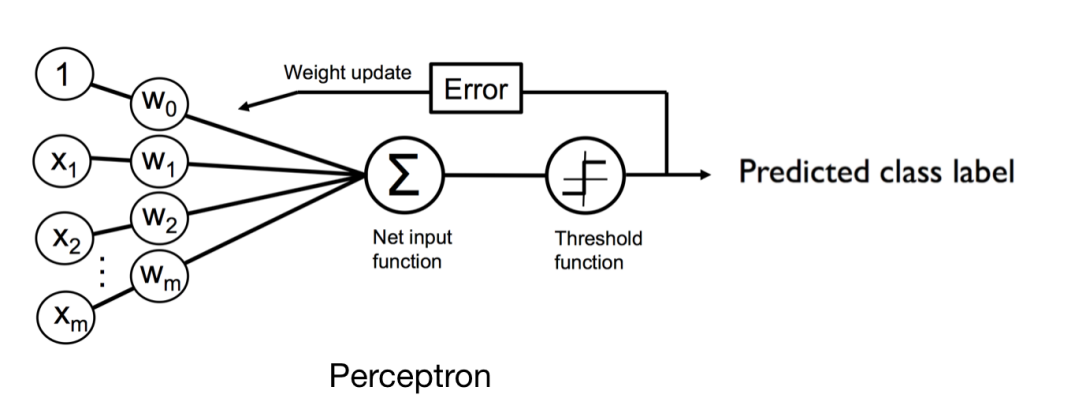


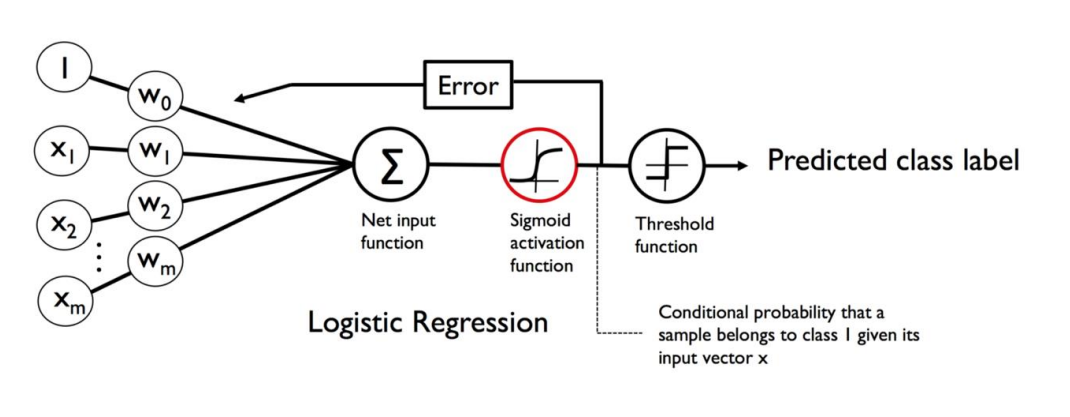
大於0就判斷1、小於等於0就判斷成是-1



Logistic Regression

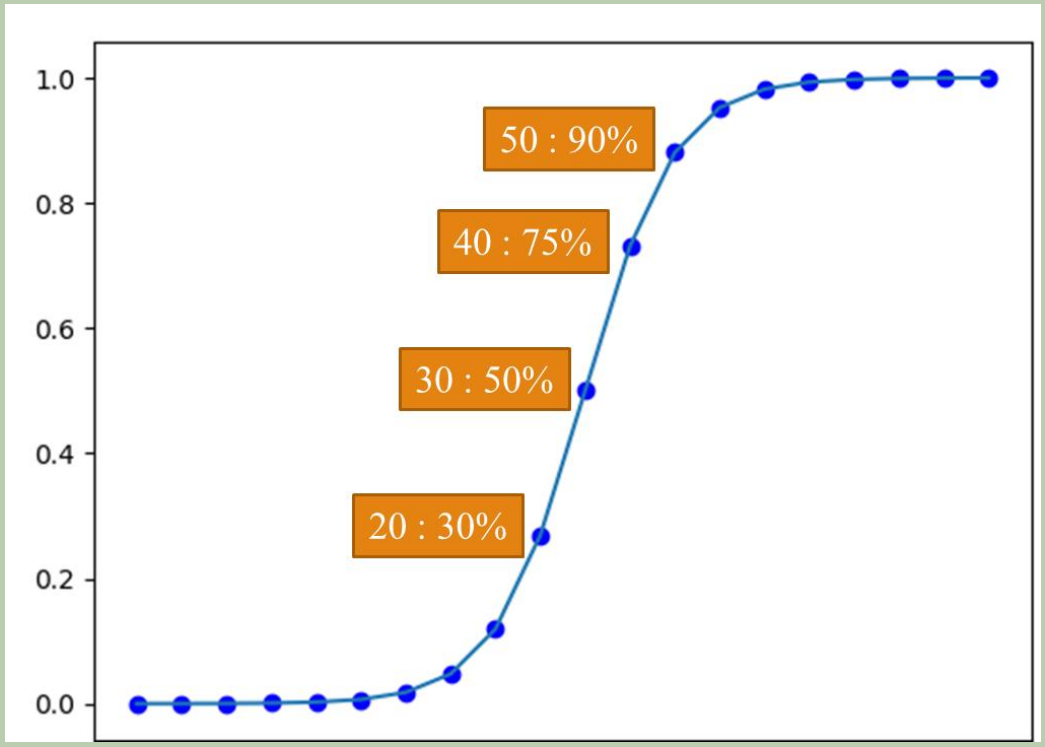
- 前面的 Perceptron 只能讓我們達成分類
- 我們只能得知預測結果，而不能知道機率
- 二者的差異主要是在激勵函數的部分，Logistic Regression 為 sigmoid 函數。



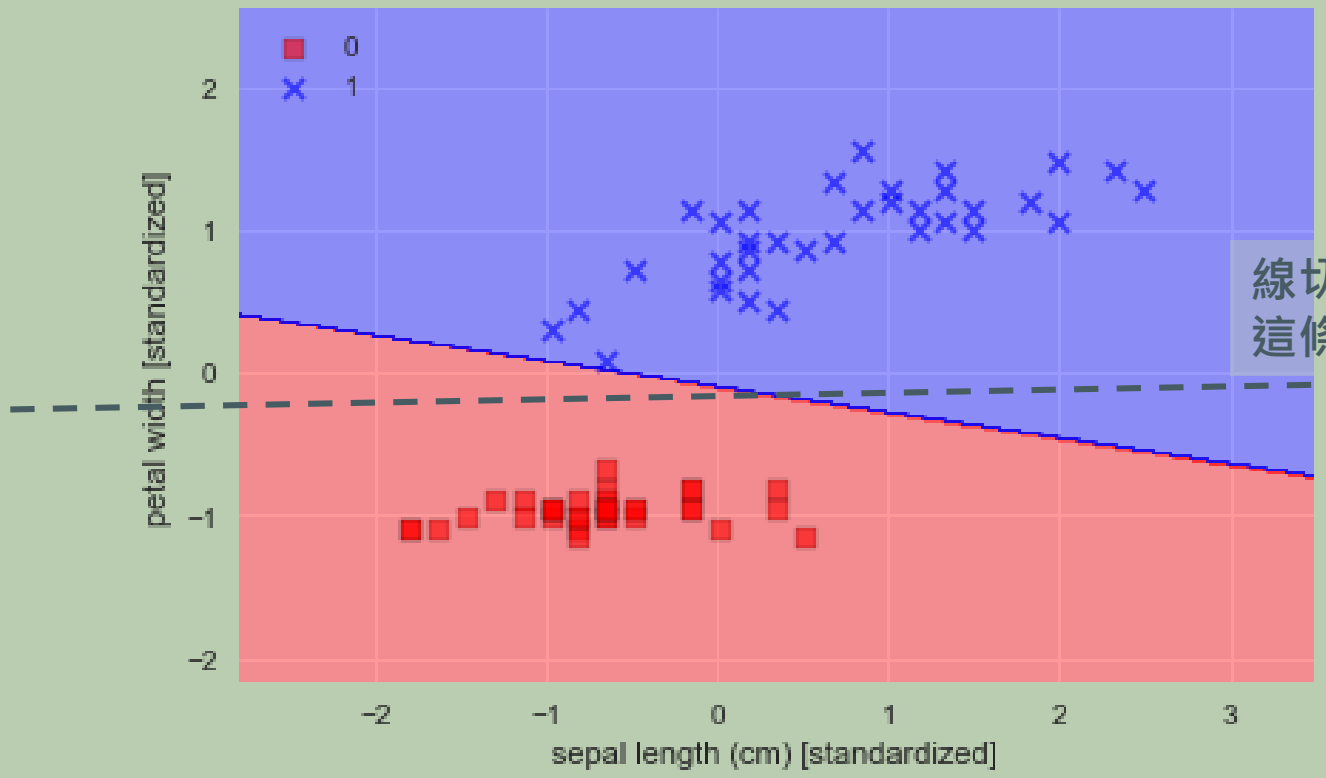


Logistic Regression

- Perceptron 是把所有輸入特徵加總起來，根據 ≦ 或 ≧ 0 來判斷其符合哪類。
- Logistic Regression 則透過 sigmoid function ，讓它能以機率的形式呈現。



Logistic Regression



線切的不夠好，
這條可能更加理想

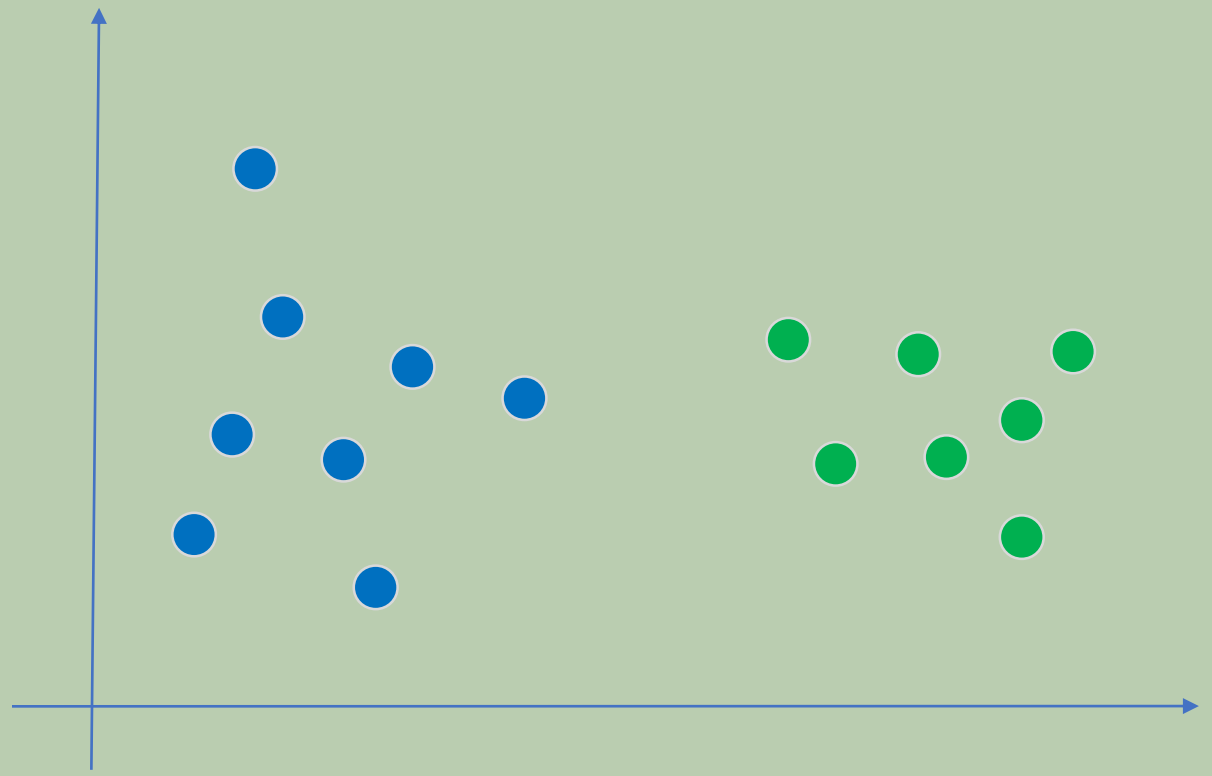
SVM

- 支持向量機 (support vector machine)
- 一種基於統計學的監督式演算法
- 透過找出一個超平面，使其兩個不同的集合分開
- SVM 的精神就是**找出一條最好的分隔線使所有在邊界上的點離得越遠越好**

- 另外 SVM 可以分為**線性可分**與**線性不可分**這兩種

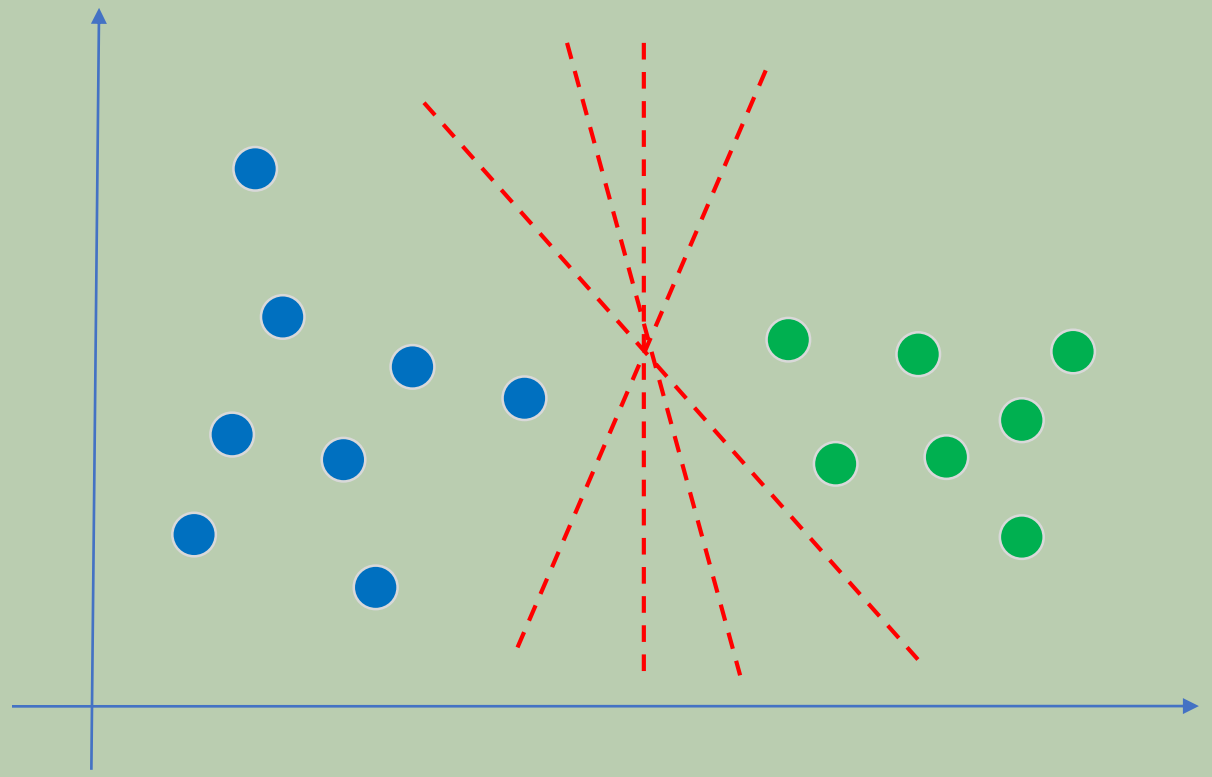
SVM 線性可分

- 如何區分這些資料



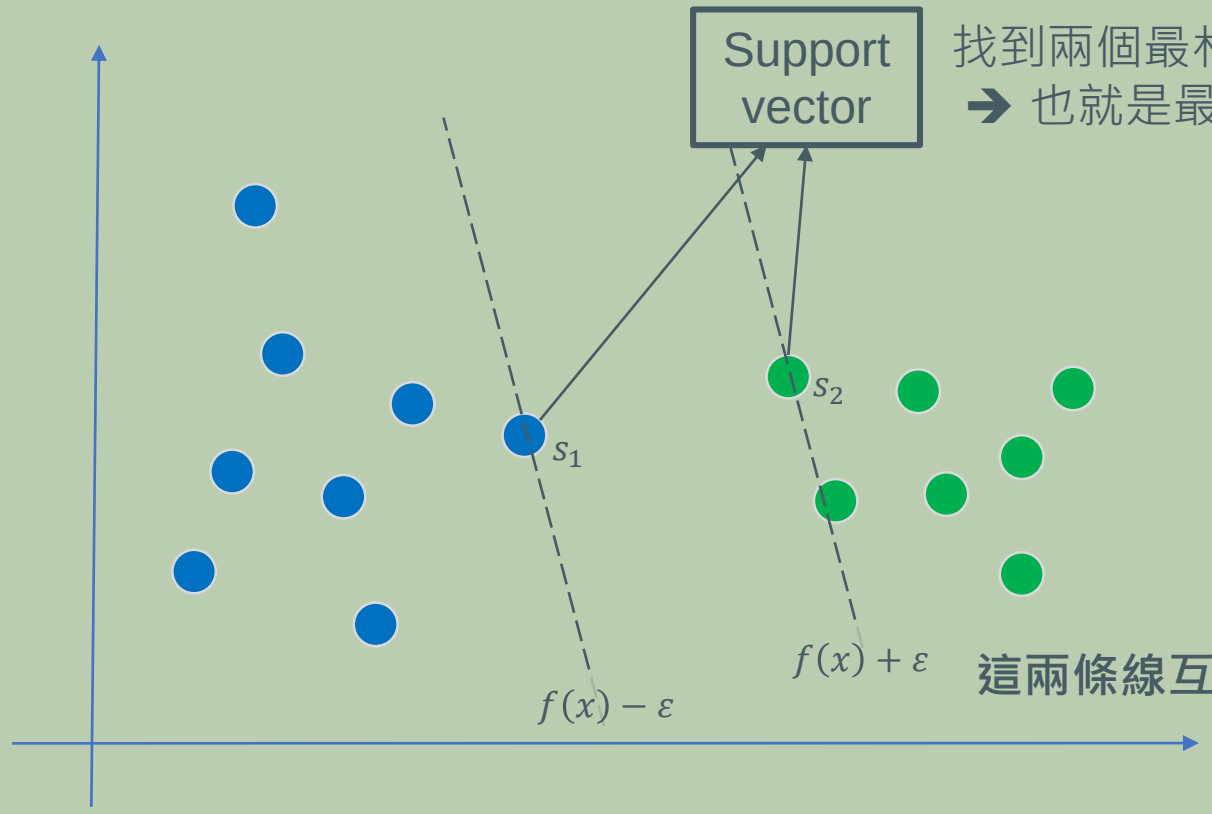
SVM 線性可分

- 也許可以透過這些線，但該如何知道哪一條是最好的?



SVM 線性可分

- 找出最近的兩點與超平面

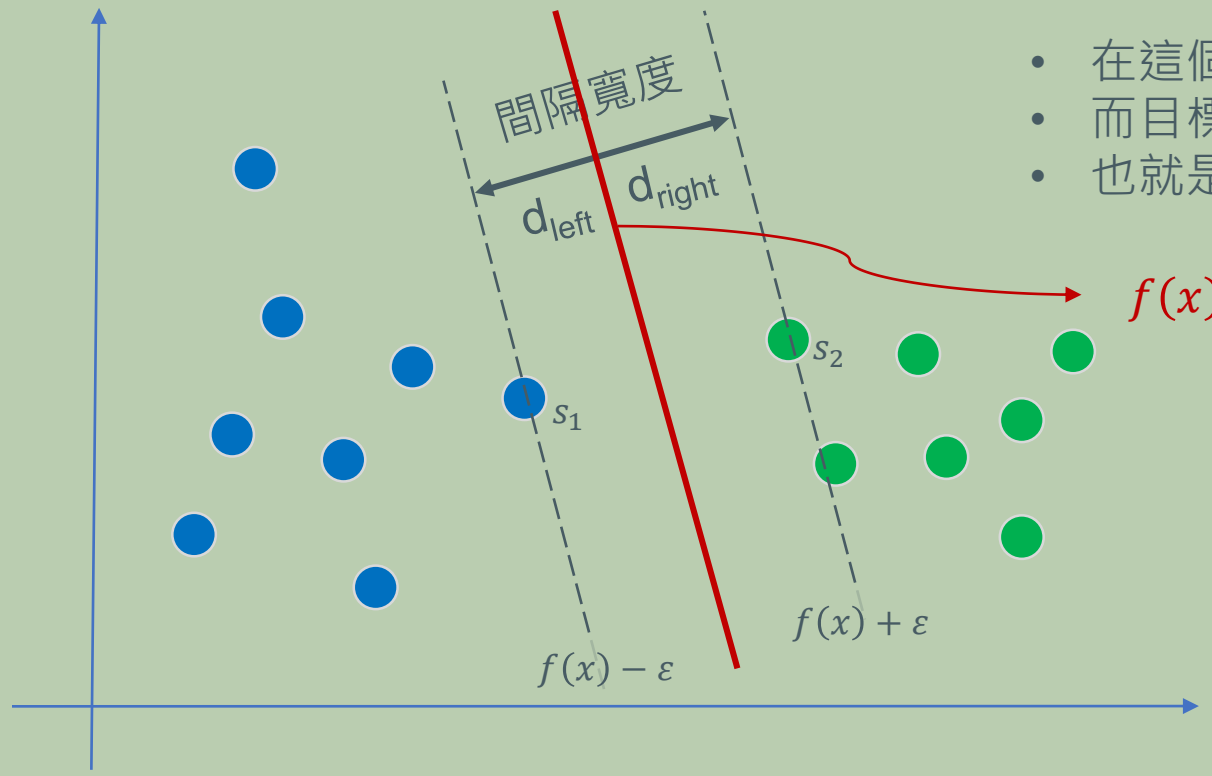


找到兩個最相鄰的點
➔ 也就是最近的兩個點

這兩條線互相平行，被稱為超平面

SVM 線性可分

- 透過計算得出最優解

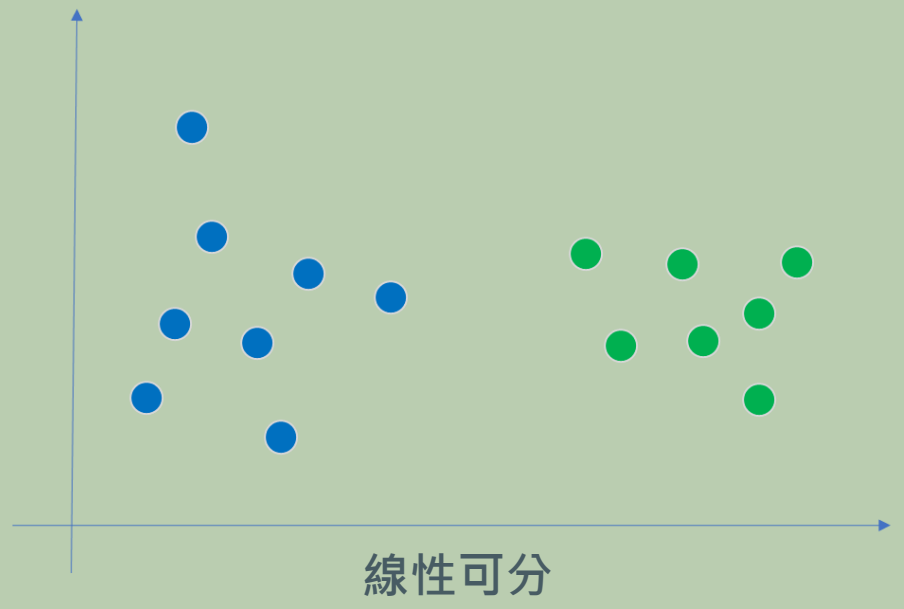


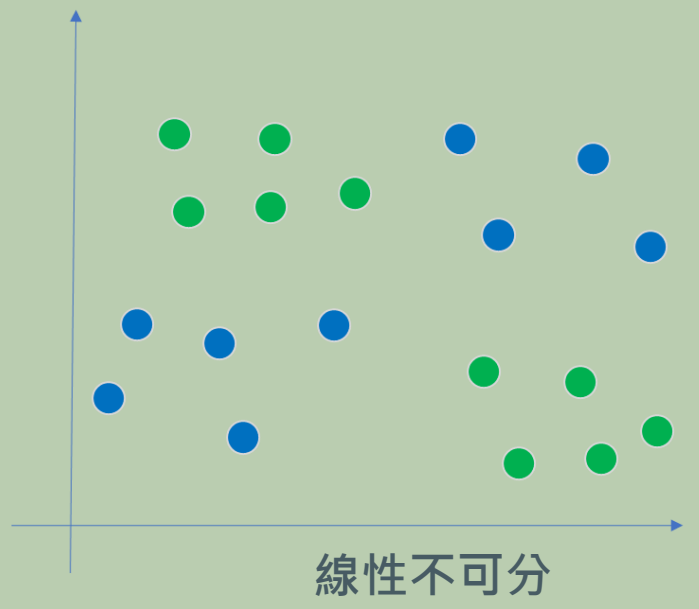
- 在這個間隔寬度內的所有平行線都屬於解
- 而目標是找到在這個間隔寬度中心的解
- 也就是最佳解 $f(x) = wTx + b$

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

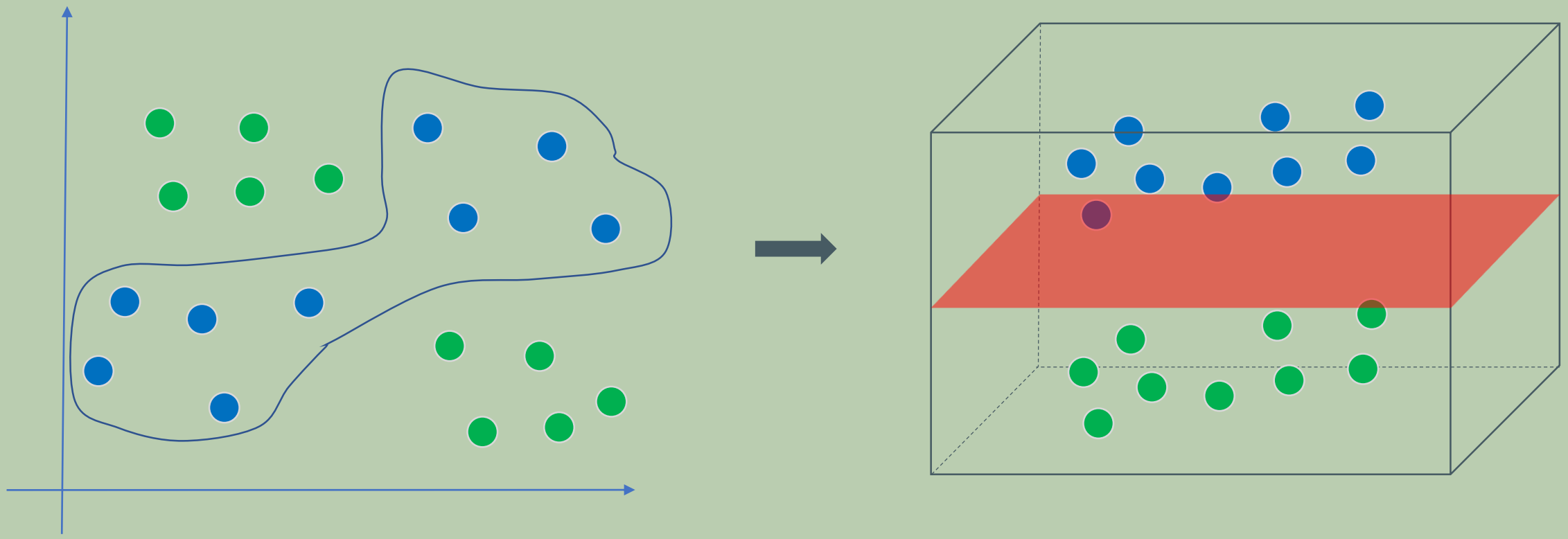
SVM 線性不可分

- 不過現實生活的資料往往稍微複雜，如果遇上線性不可分的集合該怎麼辦呢?
- 我們可以運用核函數(kernel function) 來造出分割平面
- 例如: Polynomial 高次方轉換、Radial Basis Function 高斯轉換





SVM 線性不可分



評估方法

- 如何評估分類模型的效能?
- 通常會計算準確率，並作為評估指標。

- 簡單來說，就是另外拿一筆資料，作為考題讓模型進行預測。
- 最後在計算模型一共猜對幾題，從而算出準確率有多少。

程式 demo

Thanks for your listening.