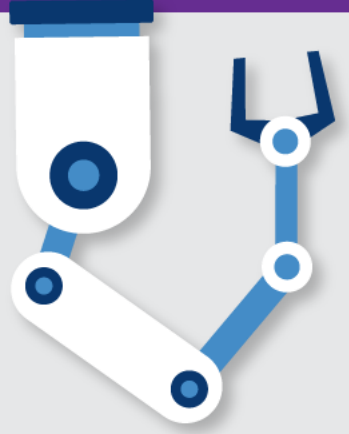
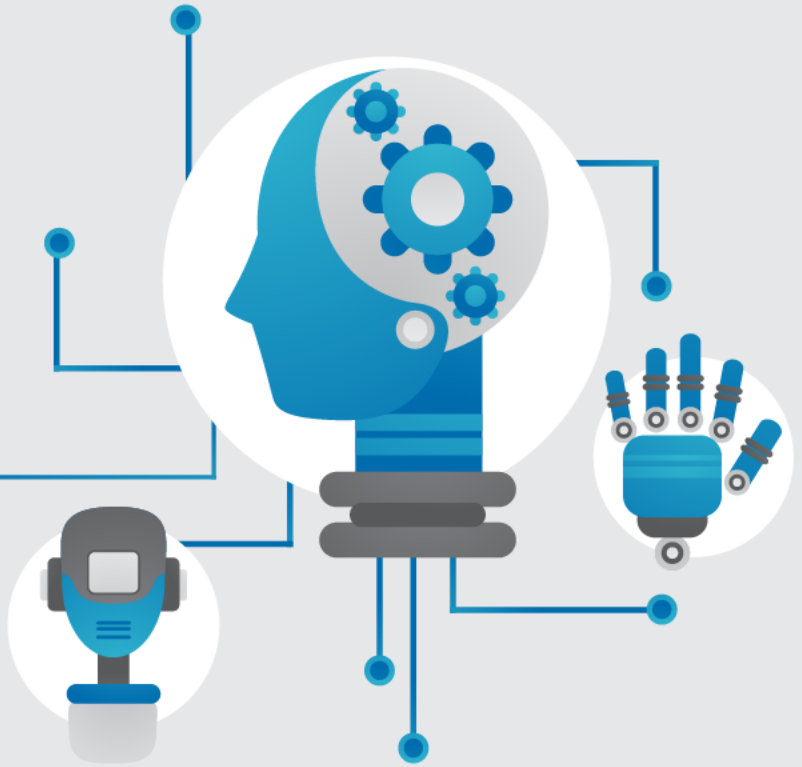


羅吉斯迴歸
Logistic Regression



迴歸分析

- 迴歸分析是一種統計學上分析數據的方法，目的在於了解變數間是否相關（相關方向、相關程度），並建立數學模型以便觀察特定變數來預測研究者感興趣的變數。

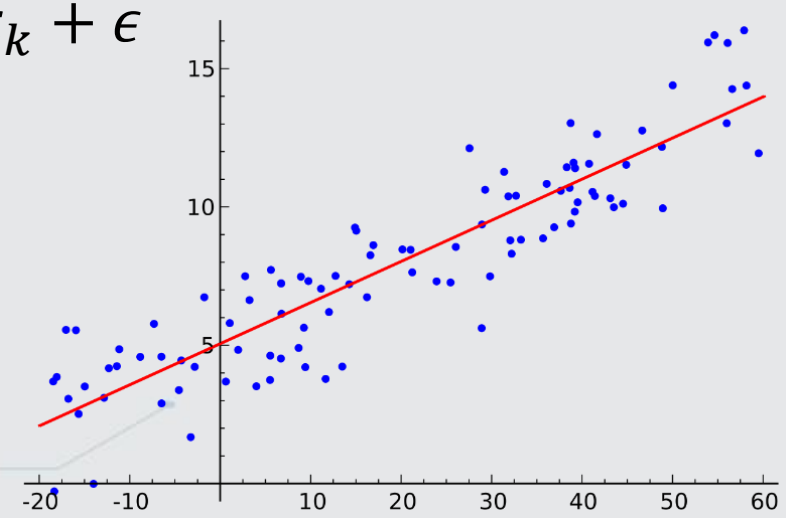
- 迴歸分析是用自變數（independent variable）來預測應變數（dependent variable）。

  例如：用父母身高來預測子女身高；
  用個人生理資訊和家族史，來預測是否會罹癌。

線性迴歸

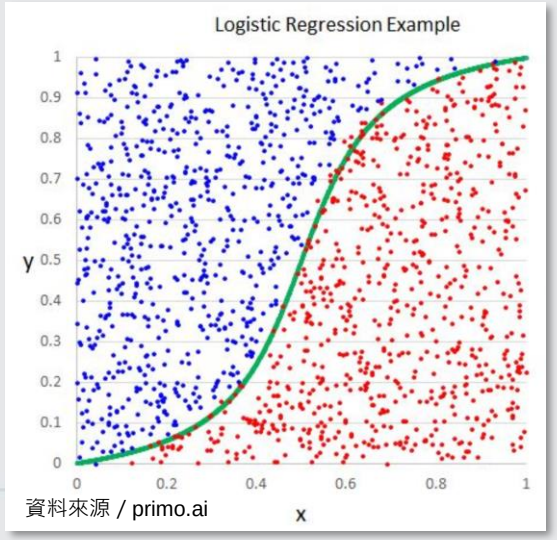
- 線性迴歸（Linear Regression）利用線性迴歸方程式的最小平方函數對一個或多個自變數（independent variable）和應變數（dependent variable）之間關係進行建模的一種迴歸分析。
- 給定一些點，找出一條直線方程式，使得所有點到這條直線的距離平方和最小。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_k x_k + \epsilon$$



羅吉斯迴歸

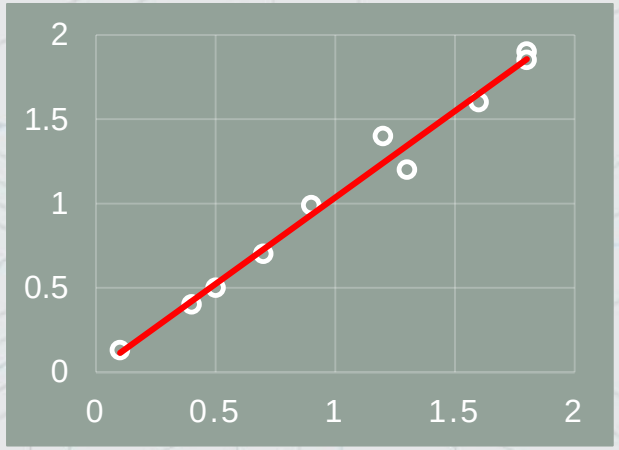
- 羅吉斯迴歸（Logistic Regression）類似線性迴歸，主要在探討**應變數**與**自變數**之間的關係，其中自變數對應變數的影響是以**指數**的方式做變動。
- 線性迴歸中的應變數(Y)通常為**連續型變數**，但羅吉斯迴歸所探討的應變數(Y)主要為**類別變數**，特別是分成兩類的變數（例如：是或否、有或無等）。



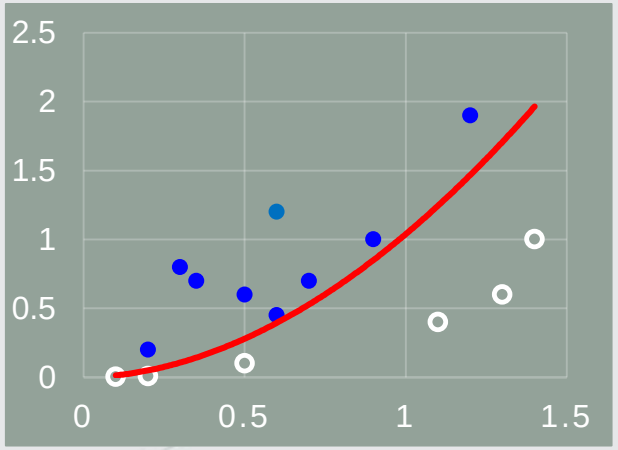
資料來源 / primo.ai

羅吉斯迴歸 vs 線性迴歸

- 在監督式學習中，常用的預測方法有
  - 對連續的資料使用迴歸模型進行資料預測
  - 對離散的資料進行分類預測
- 線性迴歸中的應變數(Y)通常為**連續型變數**，
  但羅吉斯迴歸所探討的應變數(Y)主要為**類別變數**



線性回歸希望資料集都能盡量的貼近紅線



邏輯斯回歸希望資料集能夠被紅線明顯分為兩類

羅吉斯迴歸的機制

- 羅吉斯迴歸與感知器的差異在於激活函數為**Sigmoid Function**，損失函數是**交叉熵誤差函數**（Cross entropy Error Function），也具有**正規化項**（Regularization Term），或稱懲罰項（Penalty Term），避免過度學習。

誤差函數

- 交叉熵誤差函數

$$E = -\sum_{i=1}^{N} t_i log y_i + (1 - t_i)\log(1 - y_i)$$

N : 資料數

y_i : 預測值

t_i : 正確值

正規化

- 正規化（Regulation）在學習時給予懲罰，讓決策分界線變得更平滑
- 加入正規化，目標函數 = 損失函數的總資料和 + 正規化項
- L2正規化：將權重參數的平方值當作損失函數的懲罰項

$\lambda \sum_{i=1}^{m} \varpi_i^2$, λ是控制懲罰項影響程度的參數

- L1正規化：將權重參數的絕對值當作損失函數的懲罰項

$\lambda \sum_{i=1}^{m} |\varpi_i|$, λ是控制懲罰項影響程度的參數

羅吉斯迴歸的優缺點

優

- 實現簡單，廣泛的應用於工業問題上
- 分類時計算量非常小，速度很快，存儲資源低
- 方便於觀測樣本概率分數
- 計算代價不高，易於理解和實現

羅吉斯迴歸的優缺點

缺

- 當特徵空間很大時，羅吉斯迴歸的性能不是很好
- 容易欠擬合，一般準確度不太高
- 不能很好地處理大量多類特徵或變數
- 只能處理兩分類問題且必須線性可分
- 對於非線性特徵，需要進行轉換