Python
資料科學與
人工智慧
應用實務 Journey to Data Scientists with Python

Artificial Intelligence

第15章
機器學習演算法
實作案例 – 迴歸

15-1 認識機器學習演算法

15-2 線性迴歸

15-3 複迴歸

15-4 Logistic迴歸

15-1 認識機器學習演算法

15-1-1 機器學習演算法的種類

15-1-2 Scikit-learn介紹

15-1-1 機器學習演算法的種類 – 監督式學習

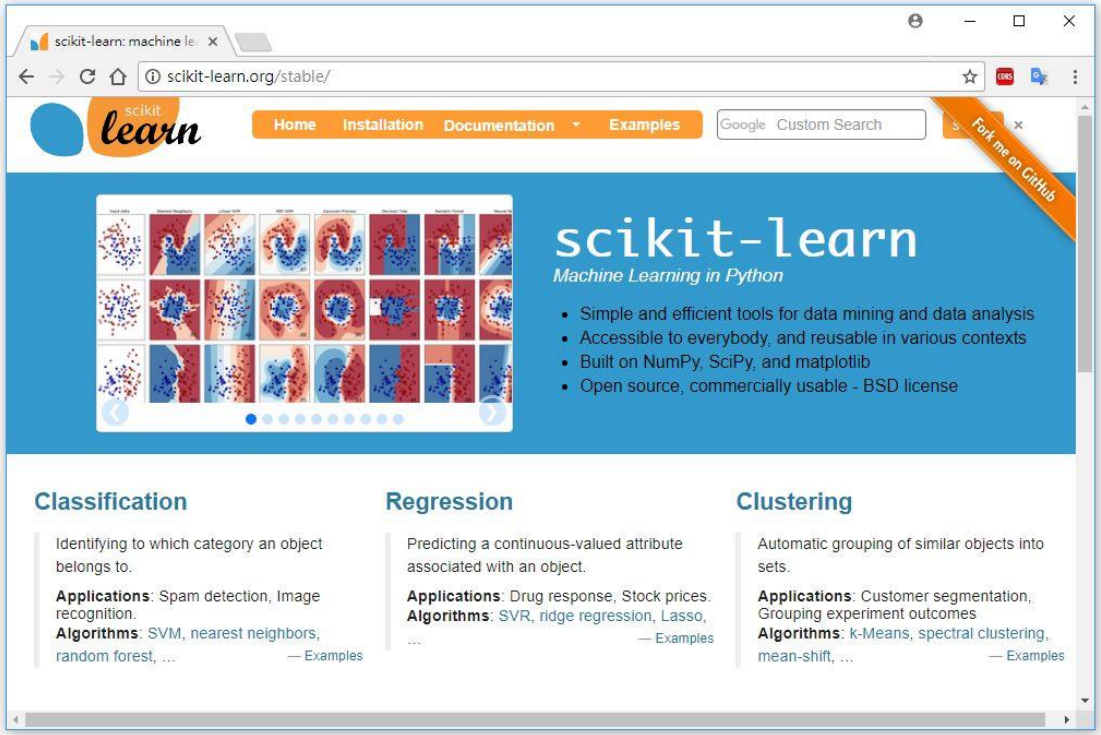
- 監督式學習的問題基本上分成兩類，如下所示：
  - **迴歸問題**：預測連續的回應資料，一種數值資料，我們可以預測商店的營業額、學生的身高和體重等。常用演算法有：線性迴歸、SVR等。
  - **分類問題**：預測可分類的回應資料，這是一些有限集合，我們可以分類成男與女、成功與失敗、癌症分成第1~4期等。常用演算法有：Logistic迴歸、決策樹、K鄰近演算法、CART、樸素貝葉斯等。

15-1-1 機器學習演算法的種類 – 非監督式學習

- 非監督式學習的問題基本上分成三類，如下所示：
  - **關聯**：找出各種現象同時出現的機率，稱為購物籃分析（Market-basket Analysis），當顧客購買米時，78%可能會同時購買雞蛋。常用演算法有：Apriori演算法等。
  - **分群**：將樣本分成相似的群組，這是資料如何組成的問題，可以幫助區分群出哪些喜歡同一類電影的觀眾。常用演算法有：K-means演算法等。
  - **降維**：減少資料集中變數的個數，但是仍然保留主要資訊而不失真，我們通常是使用特徵提取和選擇方法來實作。常用演算法有：主成分分析演算法等。

15-1-2 Scikit-learn介紹

- Scikit-learn是scikits.learn的正式名稱，一套支援Python 2和Python 3語言且完全免費的機器學習函數庫，內建多種迴歸、分類和分群等機器學習演算法，官方網址如下：
  http://scikit-learn.org/stable/



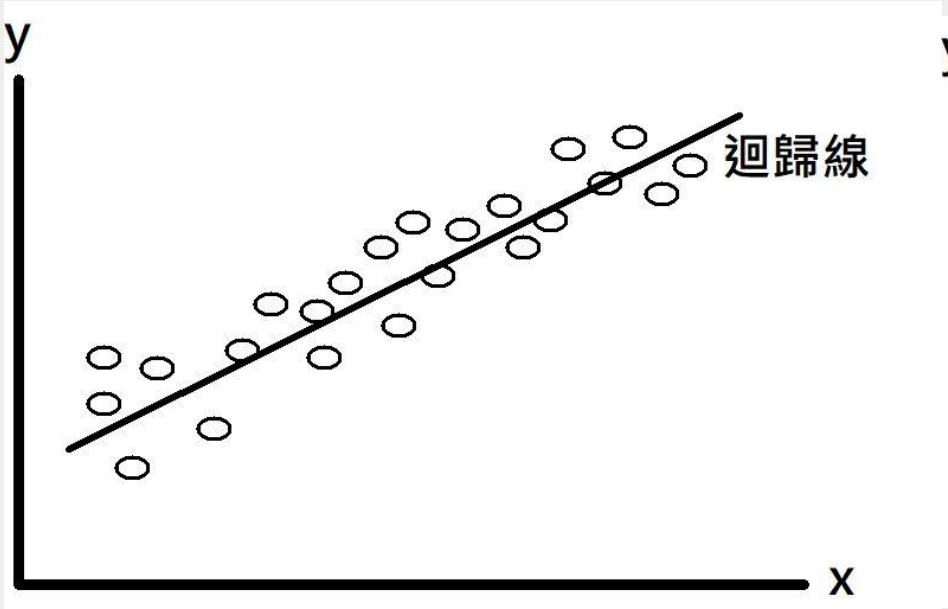
15-2 線性迴歸

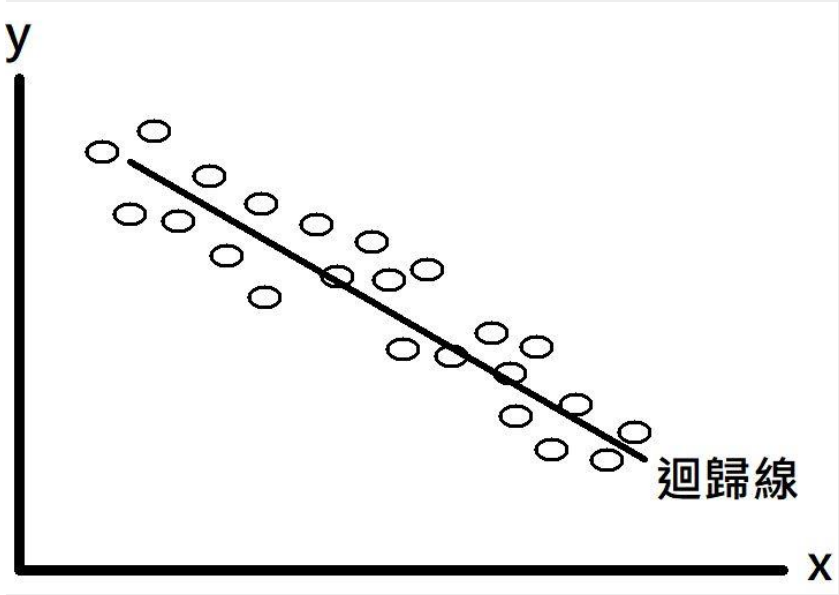
15-2-1 認識迴歸線

15-2-2 簡單線性迴歸

15-2-1 認識迴歸線

- 在說明線性迴歸之前，我們需要先認識什麼是迴歸線，基本上，當我們預測市場走向，例如：物價、股市、房市和車市等，都會使用散佈圖以圖形來呈現資料點，如下圖所示：





15-2-1 認識迴歸線

- 因為迴歸線是一條直線，其方向會往右斜向上，或往右斜向下，其說明如下所示：
  - **迴歸線的斜率是正值**：迴歸線往右斜向上的斜率是正值（見上述圖例），x和y的關係是正相關，x值增加；同時y值也會增加。
  - **迴歸線的斜率是負值**：迴歸線往右斜向下的斜率是負值，x和y的關係是負相關，x值減少；同時y值也會減少。

15-2-2 簡單線性迴歸 – 說明

- **簡單線性迴歸**（Simple Linear Regression）是一種最簡單的線性迴歸分析法，只有1個解釋變數，這條線可以使用數學的一次方程式來表示，也就是2個變數之間關係的數學公式，如下所示：

$$迴歸方程式 y = a + bX$$

- 公式的變數y是**反應變數**（Response，或稱應變數），X是**解釋變數**（Explanatory，或稱自變數），a是**截距**（Intercept），b是迴歸係數，當從訓練資料找出截距a和迴歸係數b的值後，就完成預測公式。我們只需使用新值X，即可透過公式來預測y值。

15-2-2 簡單線性迴歸 – 範例一：使用當日氣溫來預測當日的業積

- 在本市捷運站旁有一家飲料店，店長記錄下不同氣溫時的日營業額（千元），如下表所示：

氣溫	29	28	34	31	25	29	32	31	24	33	25	31	26	30
營業額	7.7	6.2	9.3	8.4	5.9	6.4	8.0	7.5	5.8	9.1	5.1	7.3	6.5	8.4

- 我們準備建立簡單線性迴歸的預測模型，讓店長提供當日氣溫，即可預測出當日的營業額（Python程式：Ch15_2_2.py）。

15-2-2 簡單線性迴歸 – 範例二：使用學生的身高來預測體重

- 在國內一所高中調查10位男學生的身高和體重資料，如下表所示：

身高	147.9	163.5	159.8	155.1	163.3	158.7	172.0	161.2	153.9	161.6
體重	41.7	60.2	47.0	53.2	48.3	55.2	58.5	49.0	46.7	52.5

- 我們準備建立簡單線性迴歸的預測模型，只需輸入男學生的身高，就可以預測學生的體重（Python程式：Ch15_2_2b.py）。

15-3 複迴歸

15-3-1 線性複迴歸

15-3-2 使用波士頓資料集預測房價

15-3-3 訓練和測試資料集

15-3-4 殘差圖

15-3-1 線性複迴歸 – 說明

- **複迴歸**（Multiple Regression）是第15-2-2節簡單線性迴歸的擴充，在預測模型的線性方程式不只1個解釋變數X，而是有多個解釋變數X1、X2...等。
- 在第15-2-2節的線性迴歸是研究「1因1果」的問題，線性複迴歸（Multiple Linear Regression）是一個反應變數y和多個解釋變數X1、X2、...、Xk的關係，換句話說，就是一種「多因1果」的問題。
- Python程式只需將原來解釋變數的DataFrame物件X，從1欄位擴充成多欄位，每一個欄位是一個解釋變數，即可使用和第15-2-2節相同的方式來建立複迴歸方程式。

15-3-1 線性複迴歸 – 範例一：使用身高和腰圍來預測體重

- 在國內大學調查10位大學生的腰圍、身高和體重資料，如下表所示：

腰圍	67	68	70	65	80	85	78	79	95	89
身高	160	165	167	170	165	167	178	182	175	172
體重	50	60	65	65	70	75	80	85	90	81

- 上表的解釋變數共有2個，即腰圍和身高，我們準備建立線性複迴歸的預測模型，只需輸入大學生的腰圍和身高，就可以預測其體重（Python程式：Ch15_3_1.py ）。

15-3-1 線性複迴歸 – 範例二：使用店面面積和車站距離來預測單月營業額

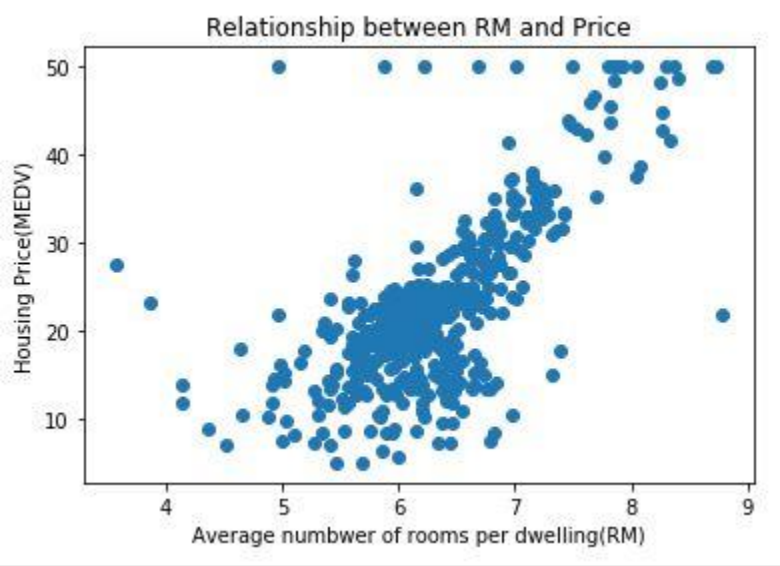
- 在捷運站附近開設的連鎖手搖飲料店準備再新開一間新分店，目前已知現有各分店的面積（坪）、距捷運站距離（公尺）和分店的單月營業額（萬元），如下表所示：

店面積	10	8	8	5	7	8	7	9	6	9
距捷運	80	0	200	200	300	230	40	0	330	180
月營收	46.9	36.6	37.1	20.8	24.6	29.7	36.6	43.6	19.8	36.4

- 上表因為解釋變數有2個，即分店面積和距離捷運站距離，我們準備建立線性複迴歸的預測模型，只需輸入新店的面積和距捷運站的距離，就可以預測新店的月營業額（Python程式：Ch15_3_1a.py）。

15-3-2 使用波士頓資料集預測房價

- 對於機器學習的初學者來說，除了使用第2篇的網頁爬蟲來搜集資料外，Scikit-learn套件本身的datasets物件內建有一些現成的資料集，可以讓我們用來學習如何訓練所需的預測模型，在這一節是使用波士頓資料集來預測波士頓近郊的房價。
  - 載入波士頓資料集
  - 建立DataFrame物件
  - 訓練預測模型
  - 使用預測模型預測房價

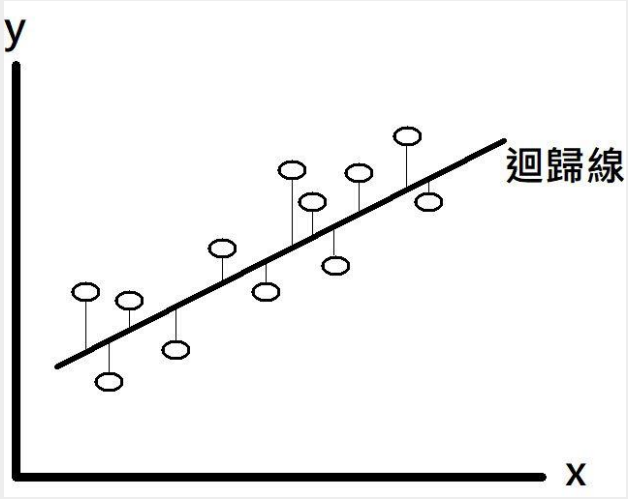


15-3-3 訓練和測試資料集 – 說明

- 在實務上，對於取得的資料集，我們並不會使用整個資料集來訓練預測模型，通常會使用隨機方式先切割成「訓練資料集」（Training Dataset）和「測試資料集」（Test Dataset），使用訓練資料集訓練預測模型後，使用測試資料集來驗證模型的績效。
  - 使用train_test_split()函數隨機分割資料集
  - 預測模型的績效

15-3-3 訓練和測試資料集 – 預測模型的績效

- 預測模型的績效是用來評量我們訓練出的模型是否是一個好的預測模型，如下圖所示：



- 上述圖例是使用簡單線性迴歸為例，一個好模型的迴歸方程式，應該最小化各資料點至迴歸線距離的總和，也就是說，觀察值和其模型的預測值差是最小的。

15-3-3 訓練和測試資料集 – 預測模型的績效

- 我們可以使用2種方式來呈現預測模型的績效，如下所示：
  - MSE（Mean Squared Error）：MSE可以告訴我們資料集的點是如何接近迴歸線，即測量各點至迴歸線的距離（這些距離稱為誤差）的平方和後，計算出平均值，因為是誤差，所以值越小；模型越好。
  - R-squared（R2）：R-squared也稱為決定係數（Coefficient of Determination），可以告訴我們資料集是如何符合迴歸線，R-squared的值是0~1，即反應變數的變異比例，我們可以使用Scikit-learn的score()函數計算R-squared，其值越大；模型就越好。

15-3-4 殘差圖 – 說明

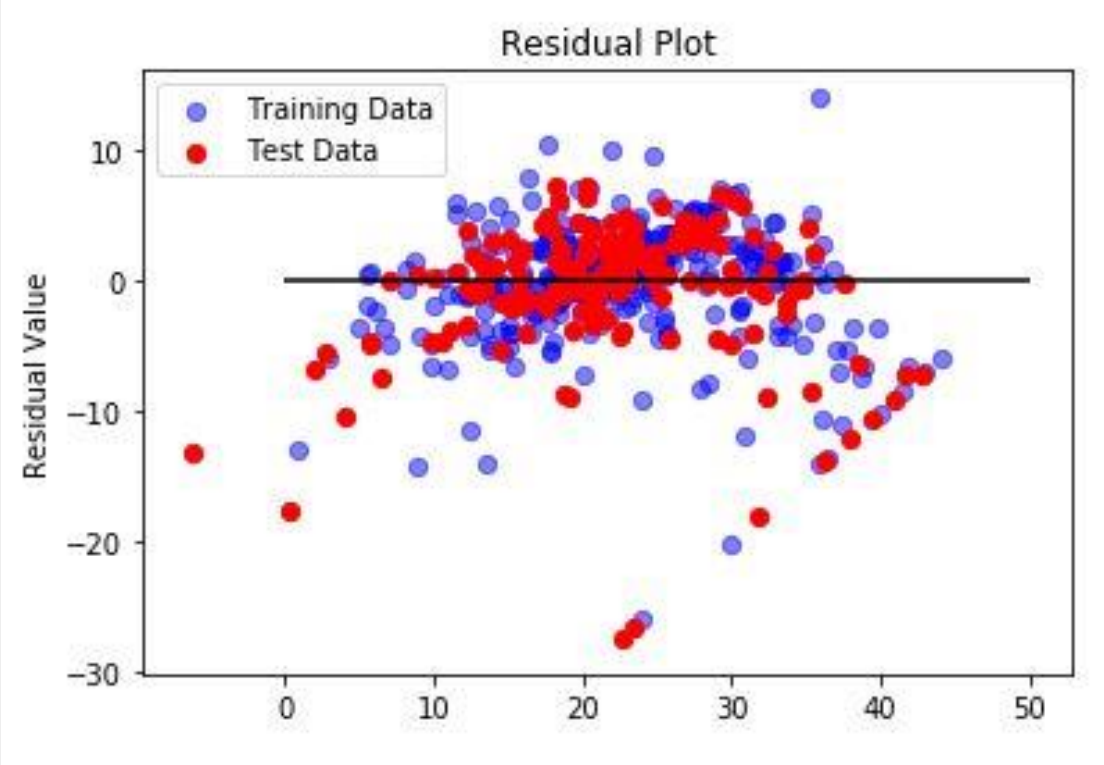
- 對於線性迴歸的預測模型來說，異常值（Outlier）會大幅影響模型的績效，我們可以使用「殘差圖」（Residual Plots）找出這些異常值。首先需要先計算出殘差值（Residual Value），其公式如下所示：

  殘差值=觀察值(Observed) - 預測值(Predicted)

- 上述公式的殘差值是原來測試資料和預測資料的差，最佳情況是等於0，即預測值符合測試資料，「>0」正值表示預測值太低；反之「<0」負值，表示預測值太高。我們可以使用殘差值作為Y軸，預設值是X軸來繪出散佈圖，這就是殘差圖（Python程式：Ch15_3_4.py ）。

15-3-4 殘差圖 – 說明

- 程式碼繪出殘差圖的散佈圖，hlines()函數可以在y=0位置繪出一條0~50的水平線，其執行結果如下圖所示：



15-4 Logistic迴歸

15-4-1 認識Logistic迴歸

15-4-2 鐵達尼號的生存預測

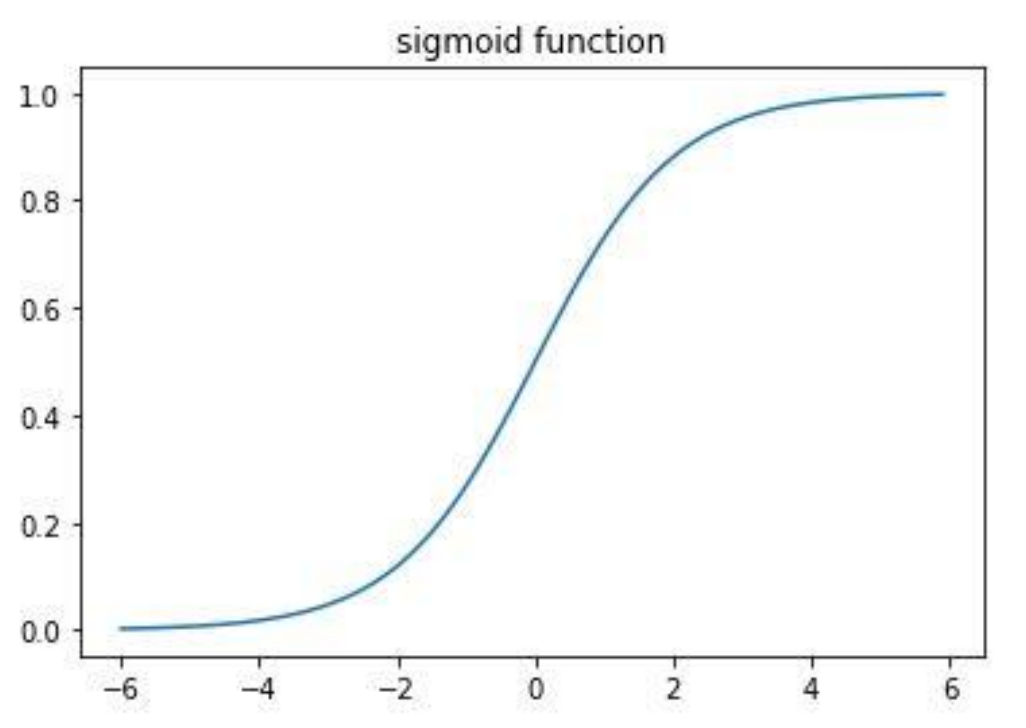
15-4-1 認識Logistic迴歸 – 說明

- **Logistic迴歸**（Logistic Regression，中文稱為邏輯迴歸）和**線性迴歸**是使用相同的觀念，不過其主要應用是二元性資料，例如：男或女、成功或失敗、真或假等，所以，Logistic迴歸和線性迴歸不同，它是在解決分類問題。
- 基本上，Logistic迴歸的作法和線性迴歸相同，只不過其結果需要使用logistic函數或稱sigmoid函數（即S函數）轉換成0~1之間的機率，其公式如下所示：

$$S(t) = \frac{1}{(1+e^{-t})}$$

15-4-1 認識Logistic迴歸 – 說明

- sigmoid函數可以使用Matplotlib套件繪出圖形（Python程式：Ch15_4_1.py），其執行結果可以看到sigmoid函數的圖形，如下圖所示：



15-4-2 鐵達尼號的生存預測

- 著名的鐵達尼號乘客資料是一份公開資訊（本書鐵達尼號資料集是取自R語言的內建資料集），在這一節我們準備使用Logistic迴歸進行鐵達尼號的生存預測。
  - 訓練Logistic迴歸預測模型
  - Logistic迴歸預測模型的準確度