監督式學習之迴歸學習器

林哲兆

January 9, 2024

機器學習的學習器有非常多種，其中又分為監督式學習與非監督式學習，差別在於是否需要資料訓練。而本文將介紹監督式學習中的三種學習器，三者皆與迴歸有關，分別是簡單線性迴歸模型、增廣性線性迴歸模型和邏輯斯迴歸模型，並用不同資料展示三種模型的分類情況，檢視不同資料在不同模型配適下的成果。

1 簡單線性迴歸模型

簡單線性迴歸對我們來說並不陌生，在很多領域都會使用到這項統計方法，因為簡單線性迴歸多適用在連續型資料，顯然把它放在分類器並不合適，但我們可以試著了解若將它作為分類器，是否真的會產生較大的誤差。

1.1 模型理論

- 學習原理：

  簡單線性回歸用於預測因變量 (y) 與單個自變量 (x) 之間的關係。它通過配適一條最適合資料的直線來建立模型，建立迴歸直線的方法則是透過最小化實際資料點與預測點的差距，來學習特徵和目標變量之間的關係。

- 預測原理：

  當模型訓練完成後，利用這條直線對新的自變量數值進行預測。將新的自變量輸入模型，通過直線方程來預測相應的因變量值。

以兩個連續型變數以及一個類別變數為例，迴歸模型將透過兩個特徵進行分類，首先將特徵 (x) 輸入之後會產生預測值 $\hat{y}$，並根據函數 G 的規則，以輸出值 0.5 為界進行分類：

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$G = \begin{cases} Group\,0, & \hat{y} \leq 0.5 \\ Group\,1, & \hat{y} > 0.5 \end{cases}$$

1.2 資料實作

第一份資料是利用 sklearn 內建套件 “make_ blob“ 生成的一個樣本，總樣本數 100，其中包含兩個連續型變數，與一個類別變數 (0、1 兩類)，這份資料兩種類別的離散程度較大，個別比較集中，是非常好進行分類的資料，簡單線性迴歸的分類結果如下：

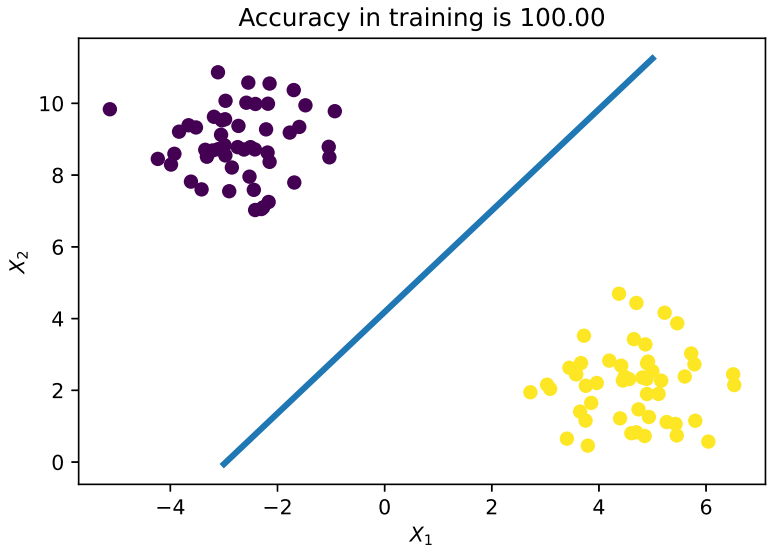

圖 1: 透過 make_ blob 生成的樣本分類

在容易分類的情況下，使用線性迴歸進行分類無傷大雅，顯然線性回歸仍具備非常基礎的分類能力，在這份樣本中達到了 100 的準確度。為此，本文將再測試資料混合程度更高的資料，以檢視其分類情況。

下面資料是來自 kaggle 的一份消費者購買行為的分類資料，其特徵包括兩個：觀測者的年齡與薪資，類別則是其購買與否 (0、1)，樣本數為 400，以下為分類狀況：

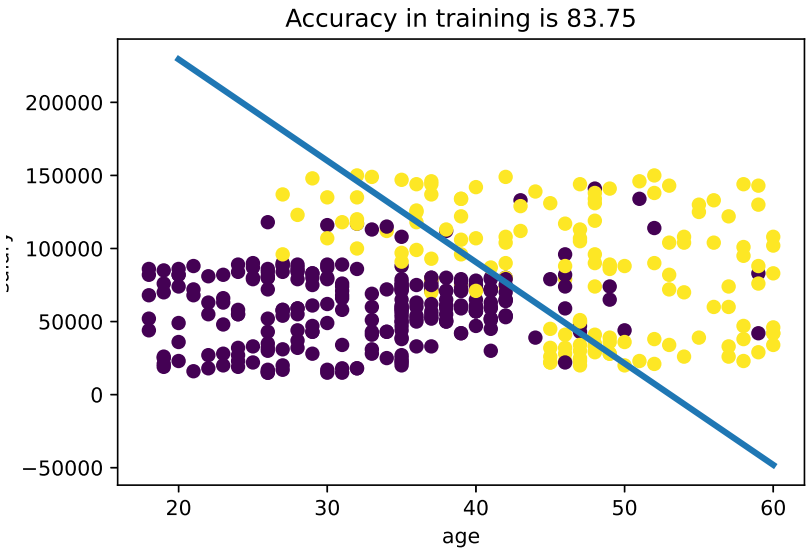

圖 2: 消費者購買行為的線性迴歸分類

由圖 2 可以看的出來當資料混合程度較高、邊界較複雜，有些資料點就會被分類錯誤，雖然如此，但其正確分類率還是有 83.75，這仍是一個滿高的數字，因此本文將再生成一個複雜程度高的資料，繼續觀察此模型分類的正確率。

下面是透過二維常態隨機亂數生成的一組樣本，其平均數與共變異數矩陣如下，並附上分類後的決策邊界與準確度：

$$\mu_0 = \begin{bmatrix} -2 \\ 3 \end{bmatrix} \mu_1 = \begin{bmatrix} -1 \\ -2 \end{bmatrix} \Sigma_0 = \begin{bmatrix} 15 & 1.2 \\ 1.2 & 20 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 5 & 2 \\ 2 & 18 \end{bmatrix}$$

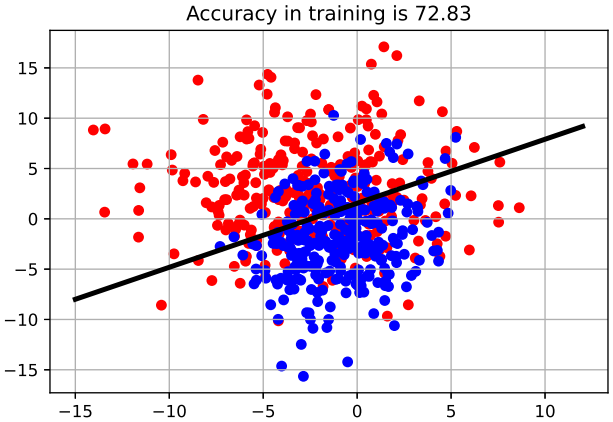

圖 3: 二維常態亂數樣本的線性迴歸分類

在這份較為複雜的資料中，雖然準確度還是有 72.83，但從圖 3 可以看見，以一條直線作為分群的界線已經不足夠，圖中可以也發現，如果要將這兩個組別完全盡量的分開，那麼兩組的決策邊界應該會是一條彎曲的線，使用直線作為邊界是一個比較粗糙的分類方法，我們也終於在這份資料看出了簡單線性迴歸的弊端。

2 加廣型線性回歸模型

加廣型線性回歸，顧名思義就是簡單線性迴歸的加深版本，在上一小節中，透過驗證發現一條直線並不能很好的解決分類的任務，因此為了改善這個問題，我們需要一條更貼合邊界的曲線，而加廣型線性回歸的引進，就稍微改善了這樣的問題。

2.1 模型理論

- 學習原理：

  增強回歸模型是對簡單線性回歸模型的擴展，它更考慮了自變量之間的關係，以及自變數的次方關係。與簡單線性迴歸相同，也用於預測因變量與多個自變量之間的關係。另外，這種模型可以處理多維度的資料，並通過擬合最適合資料的多維平面或超平面來建立模型。

- 預測原理：

  當模型訓練完成後，利用這條迴歸線對新的自變量數值進行預測。將新的自變量輸入模型，通過方程式來預測相應的因變量值。

2.2 資料實作

同樣以兩個特徵、兩種組別的資料為例，加廣型線性迴歸的模型如下，與簡單線性迴歸相比，增加了二次方項與交互作用項，分類方法同樣是將特徵 (x) 輸入之後產生預測值 $\hat{y}$，並以輸出值 0.5 為界進行分類：

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

$$G = \begin{cases} Group\,0, & \hat{y} \leq 0.5 \\ Group\,1, & \hat{y} > 0.5 \end{cases}$$

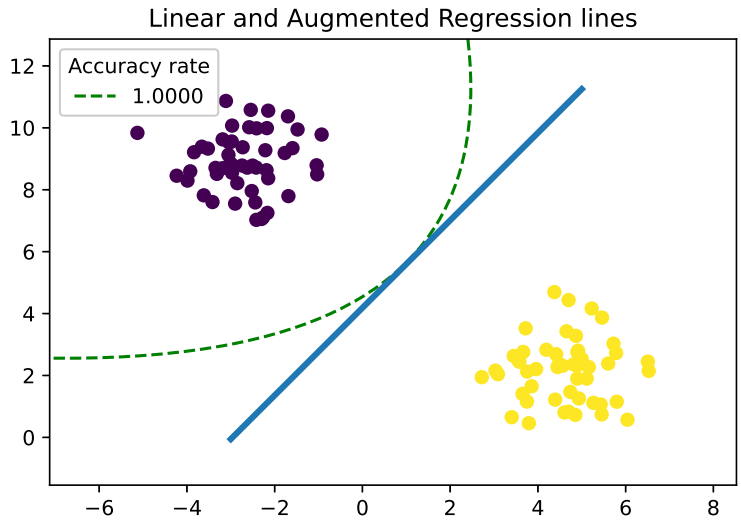

圖 4: 透過 make_ blob 生成樣本的線性迴歸分類

從圖 4 中可以看得出來，增廣型線性迴歸的決策邊界是一條曲線，雖然與簡單線性迴歸的決策邊界不同，但分類結果是相同的，正確率也是 100，這樣的結果應該是理所當然，因為更複雜的模型將會對結果有更高的適配性，所以如果較基礎的簡單線型迴歸都能做到完全分類，那更高階的增廣型線性迴歸也能做到。再來同樣使用第二份資料，也就是消費者行為分類的資料，其分類結果如下：

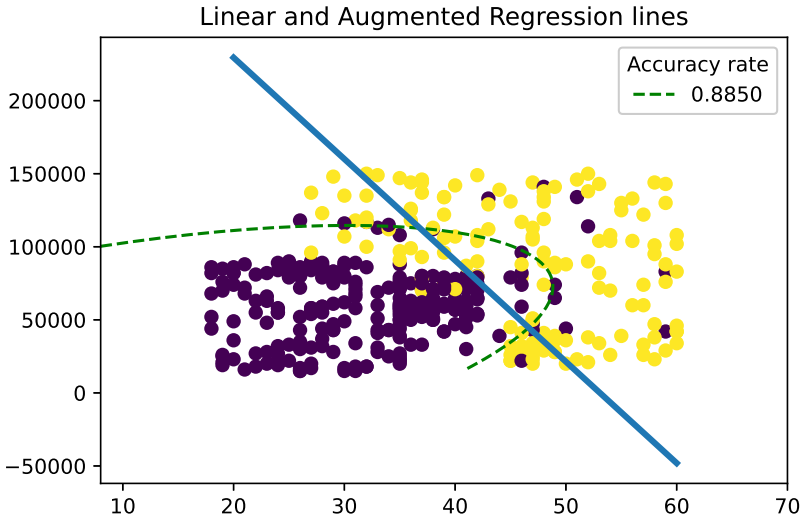

圖 5: 消費者購買行為的線性迴歸分類

在圖 5 中，簡單線性迴歸與增廣型線性迴歸的優劣一眼可見。這份資料的邊界類似一個直角，要正確的分類這份資料勢必需要一個非線性模型，因此增廣型線性迴歸是一個很好的選擇，它的邊界呈現如一個鉤子狀，相當近似這份資料的直角邊界，把組別 1、0 清楚地分別開來，正確率也達到了 88.5，相比先前的 83.75 有很大的提升。

最後看到第三份資料，也就是兩類別混合程度較高的資料，其分類結果如下：

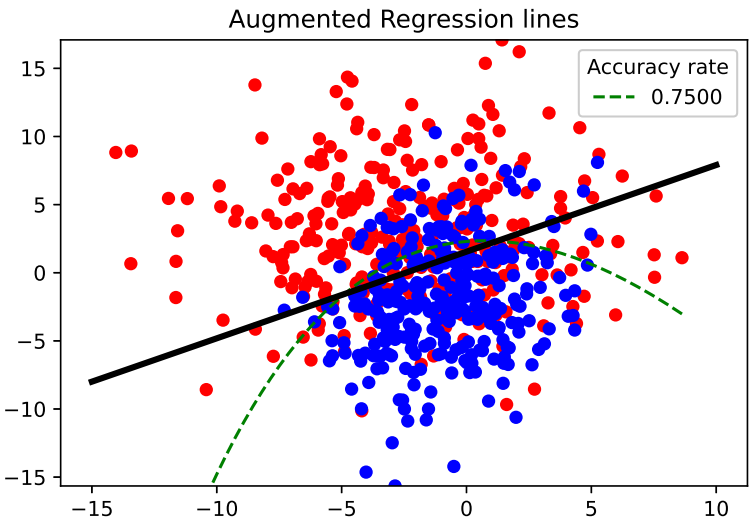

圖 6: 二維常態亂數樣本的線性迴歸分類

在這份資料中，增廣型線性迴歸的表現並不向上個資料那麼突出，其正確率僅僅增加 1.5，雖然決策邊界有更貼近，但由於資料混合程度較高，所以儘管為某部分資料調動邊界，仍會造成另一部分錯誤分類，但不可否認的，這樣的曲線邊界仍是比單單一條直線來得好。

3 邏輯斯回歸模型

以上兩種迴歸事實上多用於預測“連續資料“，在預測類別資料方面，有另一個比較適當的模型，也就是邏輯斯迴歸，其特點在於它的輸出本來就會界於 0 ～ 1 之間，因此將其作為分類器相當合理。

- 學習原理：

邏輯斯迴歸是一種分類模型，用於預測多元結果（兩個類別以上）的機率。它藉由將線性方程的結果通過一個邏輯函數 (sigmoid 函數，式 1）轉換成概率值。模型學習適應訓練資料的權重，使得預測結果能夠最好地符合實際觀察到的分類。

$$y = \frac{e^{\alpha x}}{1 + e^{\alpha x}} \tag{1}$$

- 預測原理：

  在新的資料上，模型會輸出一個介於 0 和 1 之間的概率值，通過設定閾值，可以將這個概率轉換成二元的分類結果（例如 0 或 1）。

同樣使用前面小節的三份資料來分類，首先第一份資料的分類情況如下：

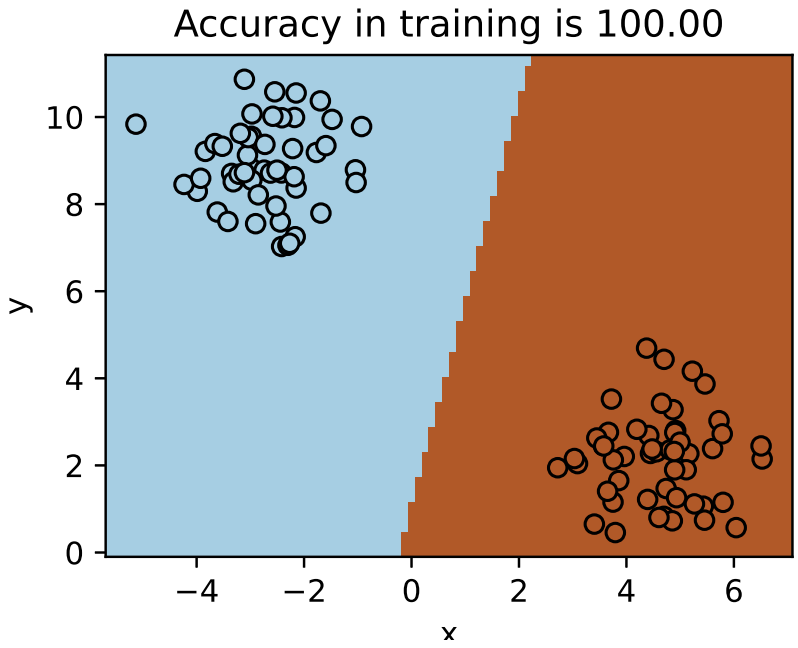

圖 7: 透過 make_ blob 生成樣本的邏輯斯迴歸分類

分類結果依然是完全正確，但值得注意的是，邏輯斯迴歸產生的決策邊界比簡單線性迴歸產生的更加接近垂直，雖然都是直線，但邏輯斯迴歸生成的邊界，離兩邊資料點的距離是較近一點的，也就是其邊界相對來說比較沒有麼貼合資料。接下來我們使用第二份資料：

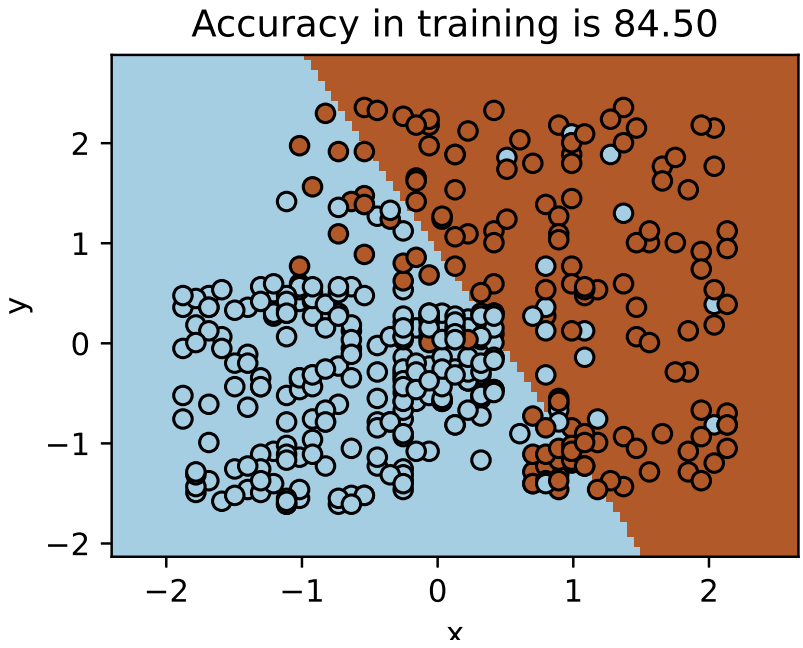

圖 8: 消費者購買行為的邏輯斯迴歸分類

從圖 8 可以看出邏輯斯分類的正確率高於線性迴歸 (83.5) 一些，從邊界圖可知其邊界也會是一條直線，與線性回歸不同的是其邊界較接近垂直，因此在此資料中使用邏輯斯迴歸與簡單線性迴歸的差異是不大的。最後是第三份資料的分類情況：

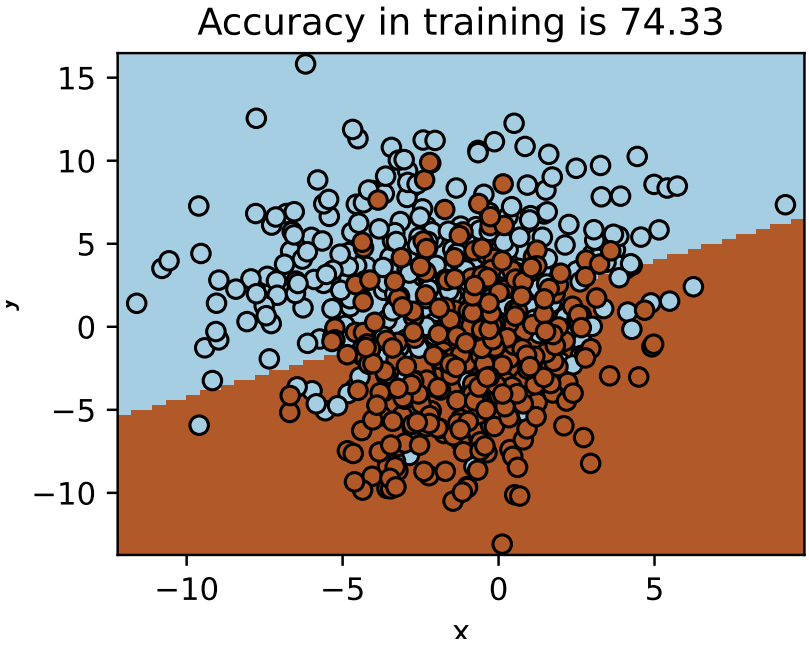

圖 9: 二維常態亂數樣本的邏輯斯迴歸分類

以上所討論的都是進行兩類分群的情況，當要分類三群或以上的時候，繼續使用線性迴歸將會變得複雜，例如分成三類的話，就得再額外去決定兩個閥值，隨著需要分類的類別增加，要設定的閥值也會越來越多，然而，羅吉斯回歸在操作這種多群分類就會較簡易，只需要將資料放進模型訓練，產生出的預測值就會是“群“，下面是使用邏輯斯迴歸進行三種類資料的分類結果，一樣是使用兩個特徵，在不需要設定閥值下，其能輕易做出決策邊界，並且準確度來到82。

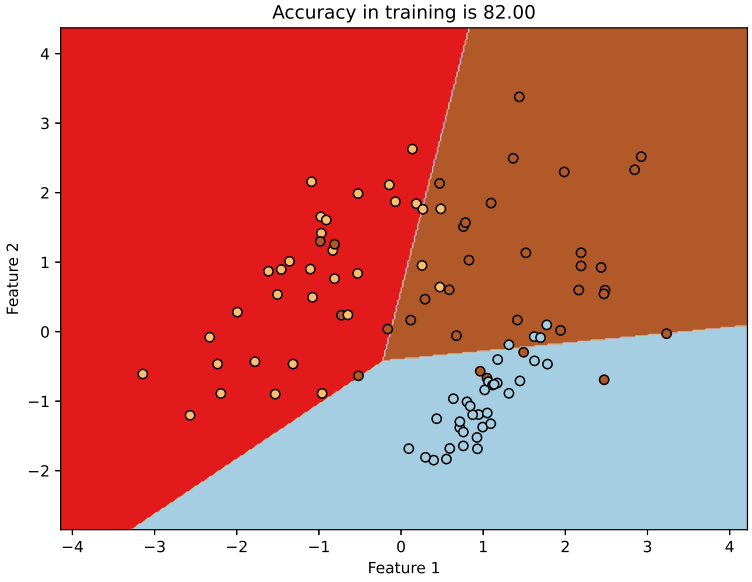

圖 10: 三個群組樣本的邏輯斯迴歸分類

4 結論

本文將三份不同分散程度的資料，分別放進三個迴歸學習器進行分類，可以知道在資料較為簡單的情況下，即使用簡單線性迴歸模型做分群，仍然可以正確分類大部份資料，但僅限在資料不複雜的情況；而遇到複雜情況時，增廣型線性模型提出了相對好的決策邊界，雖然如此，但在只有兩個特徵的情況下，增廣型線性模型就需要高達 6 個參數，很難保證其不會產生過度配適的問題，因為在過度配適的情況下，準確度很高是必然的。因此在分類問題方面，邏輯斯回歸還是最佳解，畢竟這就是一個為分類存在的模型，雖然在本文的例子中，邏輯斯迴歸的準確度都不算特別優秀 (因為資料較簡單)，但隨著資料複雜度增加，使用邏輯斯迴歸在解釋或是應用上都會更方便。