Linear and Quadratic Regression

Deterministic Approach

林哲緯

January 20, 2025

監督式學習是指使用帶有標籤的數據進行訓練，模型學習數據輸入 (input) 與標籤 (output) 之間的映射關係，以便能對新數據進行預測。學習的過程中，使得對輸入 x 的預測值 $\hat{y}$ 與真實值 y 盡可能接近。非監督式學習是指使用沒有標籤的數據進行訓練，模型試圖從數據中發現內部結構或模式。以下文章探討監督式學習中的線性迴歸和加廣型迴歸兩種模型之學習器，最後評比兩種學習器在模擬資料下的訓練表現。

1 線性模型

監督式學習中的輸出變數 Y 為類別型資料，像是 Group A 與 Group B，我們可以假設當輸入資料屬於群組 A 時，輸出變數以數字表示，如:$Y = 0$，另一個群組則為 $Y = 1$。類別資料量化之後的問題，可以套入線性迴歸模式 (Linear Regression Model) 來分析，利用某個數學關係式，譬如迴歸模型，去配適變數間的相關性，便是決定性 (Determinstic) 模式。

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

其中,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

假設共有 N 筆已知的輸入與輸出資料，假設反矩陣 $(\mathbf{X}^T\mathbf{X})^{-1}$ 存在，則迴歸係數 $\boldsymbol{\beta}$ 以最小平方法求得的最佳解為

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2}$$

每一個輸出值會根據類別分類，當給予一個新的輸入資料 $\mathbf{x}$，根據迴歸模型 (1)，其輸出擬合值為

$$\hat{\mathbf{Y}} = \mathbf{X}^T\hat{\boldsymbol{\beta}} \tag{3}$$

在迴歸模型下的擬合值 $\hat{Y}$ 不一定剛好是 0 或是 1，因此在作類別判斷時，會依下列規則判別: 以 G 代表判定的類別:

$$G = \begin{cases} \text{Group} \quad \text{if } \hat{Y} \leq 0.5 \\ \text{Group} \quad \text{if } \hat{Y} > 0.5 \end{cases}$$

上述規則以 $\hat{\mathbf{Y}} = \mathbf{X}^T\hat{\boldsymbol{\beta}}$ 做為平面空間中兩個群組的分界線，將 $\mathbb{R}^2$ 平面一分為二，線的一邊以集合 $\{\mathbf{X} \,|\, \mathbf{X}^T\hat{\boldsymbol{\beta}} \leq 0.5\}$ 代表 Group A，另一邊則為 Group B。

模型訓練 1.

在不使用 Python 套件下，根據輸出資料 Y 的類別，在二維平面上以不同顏色描繪出群組的散佈圖，利用估計出的迴歸模型參數 (2) 畫出式 (3) 中 $\hat{Y} = 0.5$ 的分界線，最後透過配適資料進行預測，利用估計值與原始值的誤差，計算訓練資料的準確率。Figure 1 為資料檔 `la_1.txt` 透過估計出的線性迴歸模型參數 (2) 畫出式 (3) 中 $\hat{Y} = 0.5$ 的分界線，其訓練之準確率為 94.00%。

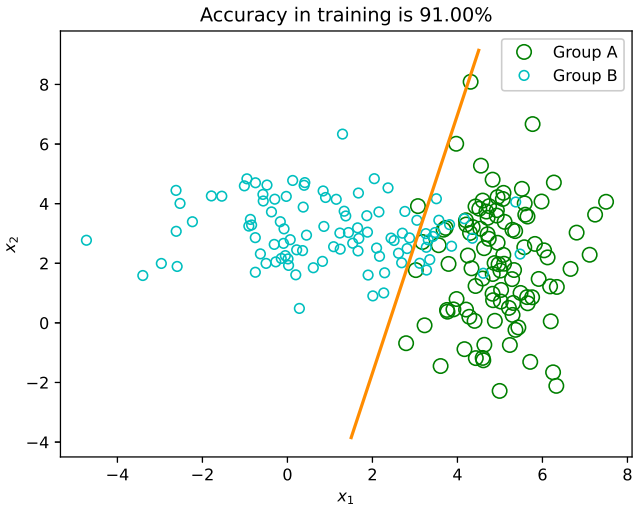

Figure 1: 線性迴歸模型分界線: 資料檔 la_1.txt

模型訓練 **2.**

使用 Scikit-Learn 套件中的`linear_model` 模組，其指令為 `LinearRegression` 建立線性迴歸模型，此訓練也將透過資料檔 `la_2.txt` 展示套件之使用方式、繪製分界線及計算訓練之準確率，準確率為 91.00%, 如 Figure 2 所示。

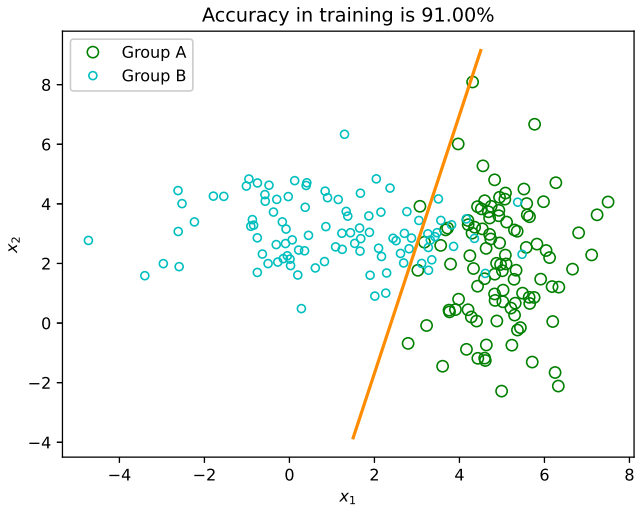

Figure 2: 線性迴歸模型的分界線: 資料檔 la_2.txt

模型訓練 **3.**

本次使用 `la_3.txt`, 對比有無使用套件下，分界線的繪製及計算訓練之準確率，如 Figure 3 所示，其中 Figure 3 左邊透過 `sklearn` 套件呈現，Figure 3 右邊則為不使用套件之呈現，兩者透過估計出的線性迴歸模型參數 (2) 畫出式 (3) 中 $\hat{Y} = 0.5$ 的分界線，其訓練之準確率皆為 73.00%。

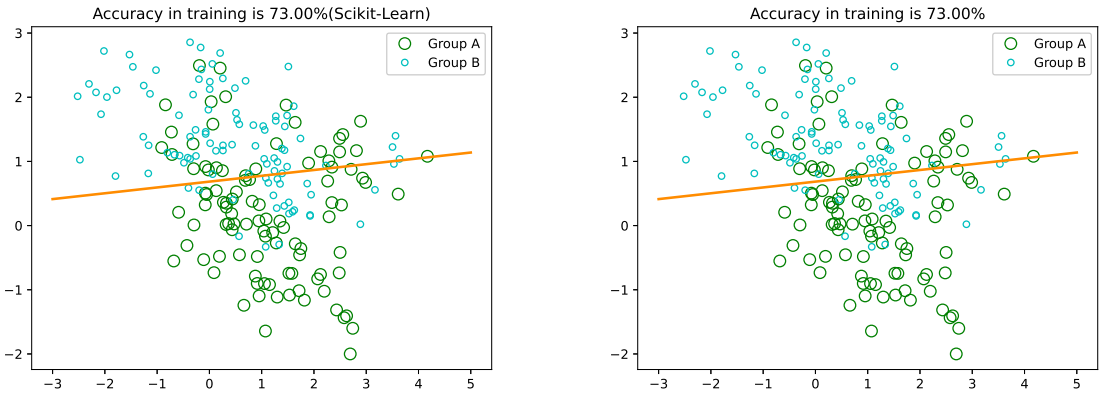

Figure 3: 比較有無使用套件下，線性迴歸模型的分界線: 資料檔 la_3.txt

2 加廣型迴歸模型

從 Figure 3 中資料分布的情況，可以發現這組資料的兩群組較密合，於是直線的分界線產生較大的判別誤差，這個誤差在機器學習的領域被稱為訓練誤差 (Training Error)。若要降低訓練誤差的方式很多，其一是變更模型，譬如改為加廣型迴歸模型 (Augmented Regression Model)，這是一條非線性的分界線，能提供更適切分隔效果。

假設輸入變數為 X_1, X_2，則 (X_1, X_2) 所有可能的值涵蓋二度空間。此時如果將兩個變數擴展為五個變數 $X_1, X_2, X_1X_2, X_1^2, X_2^2$，同樣利用迴歸模式與最小平方法建立一條分界線，當將此分界線投映回原來的空間時，它將呈現出一條曲線。這五個變數因其彼此相關的本質，並非將空間拓展為五度空間，實際仍在二度空間裡，這個所謂的加廣型迴歸模型寫成

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \epsilon \tag{4}$$

將式 (4) 以矩陣形式表示，會變為：

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} 1 & x_1(1) & x_2(1) & x_1(1)x_2(1) & x_1^2(1) & x_2^2(1) \\ 1 & x_1(2) & x_2(2) & x_1(2)x_2(2) & x_1^2(2) & x_2^2(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1(N) & x_2(N) & x_1(N)x_2(N) & x_1^2(N) & x_2^2(N) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

接著計算參數 $\hat{\boldsymbol{\beta}}$ 的最小平方估計：$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$，如同線性迴歸模式，加廣型迴歸模型的分界線表示為集合

$$\left\{(X_1, X_2) \,|\, \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 + \hat{\beta}_4 X_1^2 + \hat{\beta}_5 X_2^2 = 0.5\right\}$$

因此，我們可以將函數寫成

$$f(X_1, X_2) = \hat{\beta}_0 - 0.5 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 + \hat{\beta}_4 X_1^2 + \hat{\beta}_5 X_2^2 \tag{5}$$

模型訓練 **1.**

本次同樣以資料 `la_1.txt`作為模型訓練，以加廣型迴歸模型繪製的分界線，與 Figure 1分界線相同，且準確率亦為 94.00%.

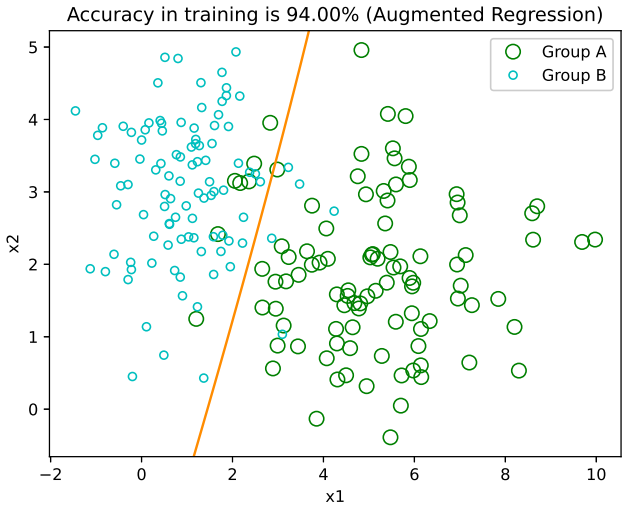

Figure 4: 加廣型線性迴歸模型分界線: 資料檔 la_1.txt

模型訓練 **2.**

此訓練以資料 `la_2.txt`作為模型訓練，繪製分界線與訓練準確率，如 Figure 5所示。可以觀察出使用加廣型迴歸模型，將資料切割得較好，且模型準確率為 93.50%，相對於線型迴歸模型的 91.00% 準確率，模型表現較好。

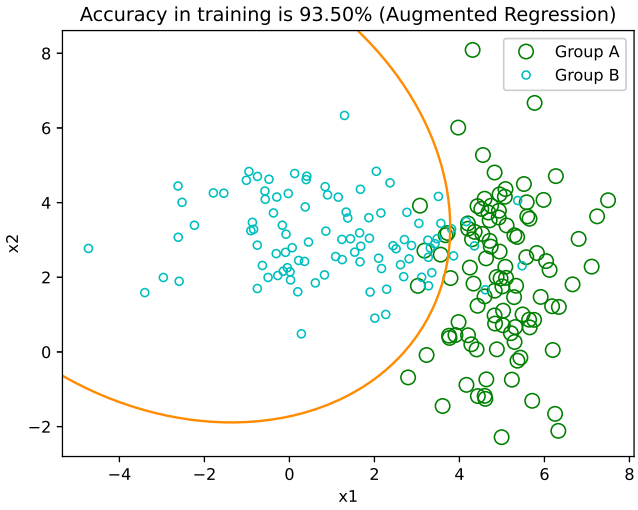

Figure 5: 加廣型線性迴歸模型分界線: 資料檔 la_2.txt

模型訓練 3.

此訓練以`la_3.txt`作為模型訓練，繪製分界線與訓練準確率，如 Figure 6所示。與 Figure 3比較後，可以觀察到使用加廣型迴歸模型，不一定能將同筆資料的分界線分科的較好，而透過計算模型準確率為 72.50%，也略低於線性模型訓練的準確率 73.00%.

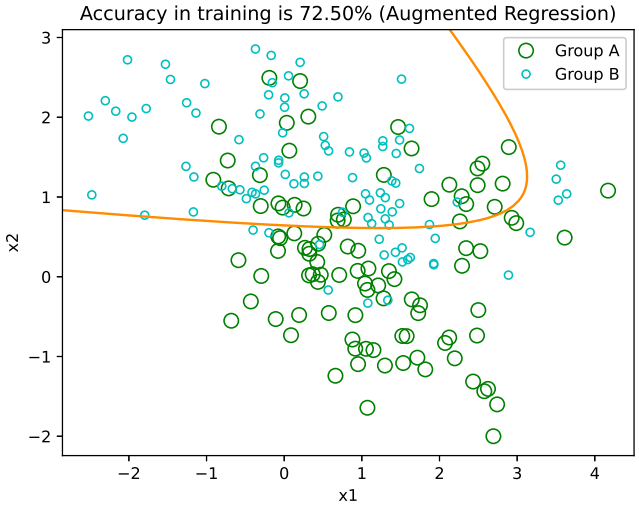

Figure 6: 加廣型線性迴歸模型分界線: 資料檔 la_3.txt

3 學習器之評比

在機器學習的領域，將不同的學習方法 (模型) 通稱為學習器，例如線性迴歸模型與加廣型迴歸模型都是學習器。先前幾次訓練可以觀察出，在相同資料及下，使用較複雜的加廣型迴歸模型，在分割資料上的表現，不一定優於線好迴歸模型。以下透過模擬不同資料，來比較不同學習器在雙變量常態母體資料下的各種表現。以下共有七種不同的模擬資料型態，其中，假設 Group A = 0、Group B = 1。

DATA 1.

Group A：

$$n_1 = 200，\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}，\boldsymbol{\Sigma_1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Group B：

$$n_2 = 200，\boldsymbol{\mu_2} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}，\boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Group A 和 Group B 皆為樣本數 $n = 200$，x_1、x_2 的變異數為 1，且相互獨立的二元常態母體。由 Figure 7 可以發現兩母體的距離是比較遠的，重合的點較少。而針對此種資料型態，由 Figure 7 可以看到加廣型迴歸的分界線很趨近於直線，並且它的判別正確率為 94.50 %，雖然很接近線性迴歸的判別正確率 94.75 %，但線性迴歸的模型較簡單，因此，使用線性迴歸進行預測會得到較佳的結果。

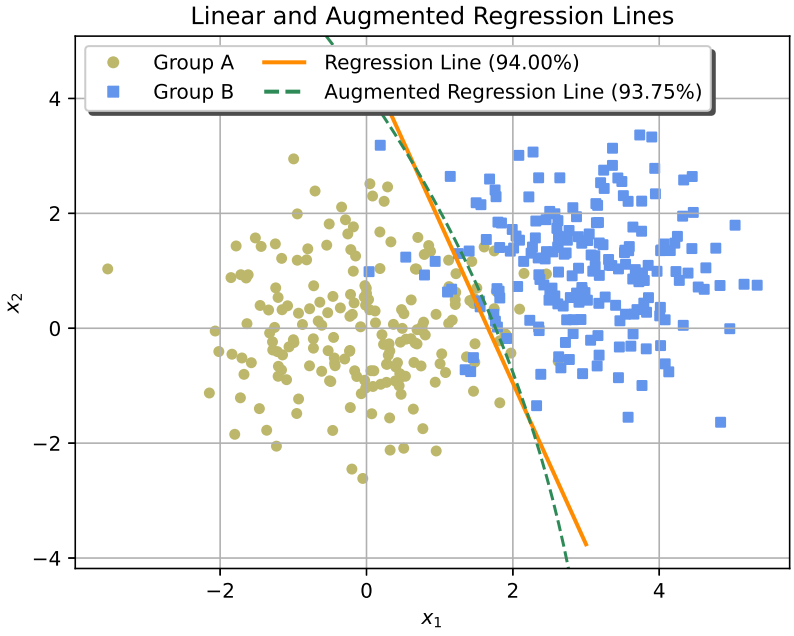

Figure 7: 兩種迴歸模型的分界線比較 (Data 1)

DATA 2.

Group A：

$$n_1 = 200，\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}，\boldsymbol{\Sigma_1} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Group B：

$$n_2 = 200，\boldsymbol{\mu_2} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}，\boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

Group A 為樣本數 $n_1 = 200$，x_1、x_2 的變異數為 2，且相互獨立的二元常態母體；Group B 為樣本數 $n_2 = 200$，x_1、x_2 的變異數為 1，且相依的二元常態母體。由 Figure 8 可以看出兩對母體的重和的資料點並不多, 且 Group A 較為分散，Group B 則較為集中且具有明顯的趨勢。而針對此種資料型態，由 Figure 8 可以看到兩種迴歸的分界線有明顯的不同。其中，加廣型迴歸的分界線的判別正確率為 92.25 %，線性迴歸的判別正確率為 91.25 %。因此，使用加廣型迴歸進行預測會得到較佳的結果。

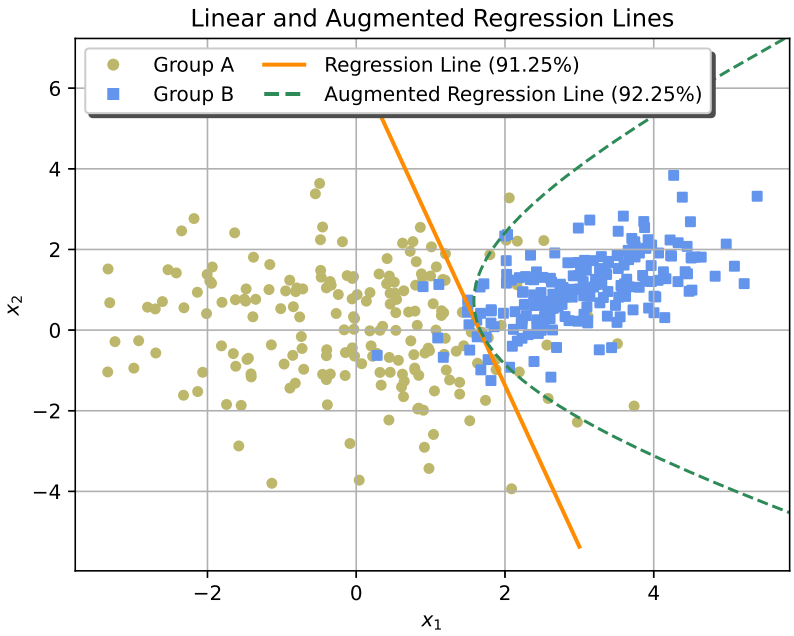

Figure 8: 兩種迴歸模型的分界線比較 (Data 2)

DATA 3.

Group A：

$$n_1 = 200，\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}，\boldsymbol{\Sigma_1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Group B：

$$n_2 = 200，\boldsymbol{\mu_2} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}，\boldsymbol{\Sigma_2} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$

Group A 為樣本數 $n_1 = 200$，x_1、x_2 的變異數為 1，且相互獨立的二元常態母體；Group B 為樣本數 $n_2 = 200$，x_1、x_2 的變異數為 5，且相互獨立的二元常態母體。由 Figure 9 可以發現兩對母體的重和部分並不多，距離較遠，且 Group B 的情況較為分散，使得有些許資料點位於 Group A 中。而針對此種資料型態，由 Figure 9 可以看到兩種迴歸的分界線有明顯的不同。其中，加廣型迴歸的分界線的判別正確率為 95.25 %，線性迴歸的判別正確率為 93.25 %。因此，使用加廣型迴歸進行預測會得到較佳的結果。

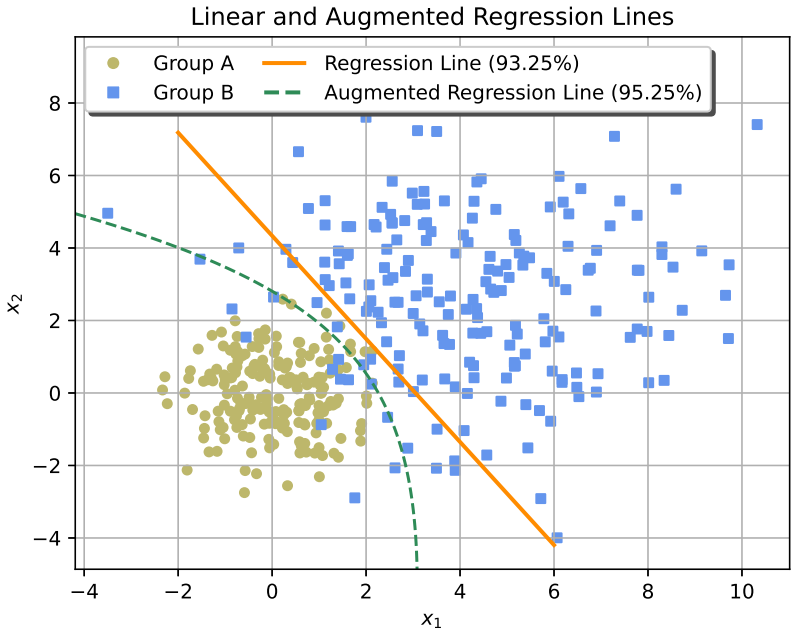

Figure 9: 兩種迴歸模型的分界線比較 (Data 3)

DATA 4.

Group A：

$$n_1 = 200，\boldsymbol{\mu_1} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}，\boldsymbol{\Sigma_1} = \begin{bmatrix} 1 & -0.6 \\ -0.6 & 1 \end{bmatrix}$$

Group B：

$$n_2 = 200，\boldsymbol{\mu_2} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}，\boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Group A 和 Group B 皆為樣本數 $n = 200$，x_1、x_2 的變異數為 1，且相依的二元常態母體。由 Figure 10 可以發現兩對母體皆具有些微的趨勢，且 Group B 資料較集中。而 Group A 的右端與 Group B 有部分重合的情形。而針對此種資料型態，由 Figure 10 可以看到兩種迴歸的分界線有些許的不同。其中，加廣型迴歸的分界線的判別正確率為 91.50%，線性迴歸的判別正確率為 90.00%。因此，使用加廣型迴歸進行兩群資料的分類會得到較佳的結果。

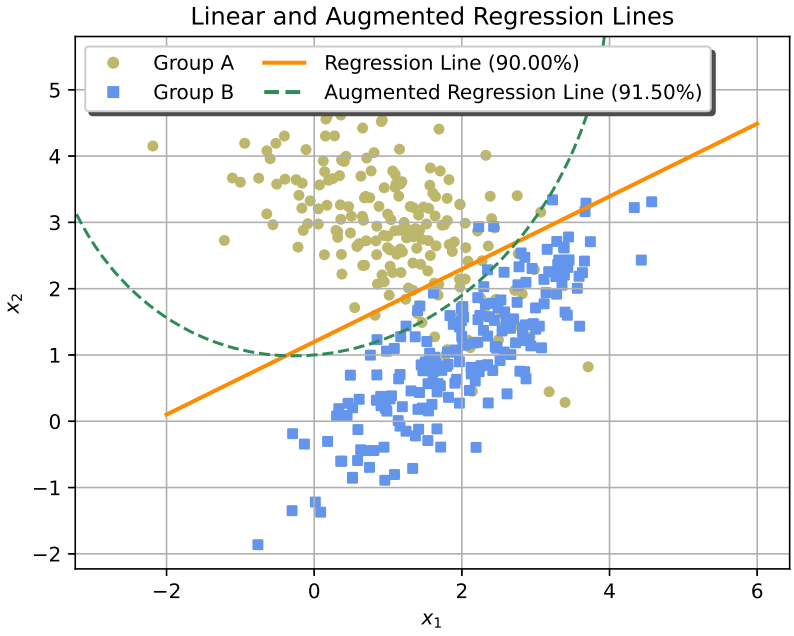

Figure 10: 兩種迴歸模型的分界線比較 (Data 4)

DATA 5.

Group A：

$$n_1 = 200，\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}，\boldsymbol{\Sigma_1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Group B：

$$n_2 = 200，\boldsymbol{\mu_2} = \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}，\boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Group A 和 Group B 皆為樣本數 $n = 200$，x_1、x_2 的變異數為 1，獨立的二元常態母體。由 Figure 11 可以發現兩對母體皆分散，且 Group A 和 Group B 有相當多的部分重合。而針對此種資料型態，由 Figure 11 可以看到兩種迴歸的分界線的表現明顯和前面幾組資料相比，分割準確率較低，且在兩組資料重疊部分，兩條分界線幾乎重合。其中，加廣型迴歸的分界線的判別正確率為 74.00%，線性迴歸的判別正確率為 74.25%。因此，使用線性迴歸進行兩群資料的分類會得到較佳的結果。

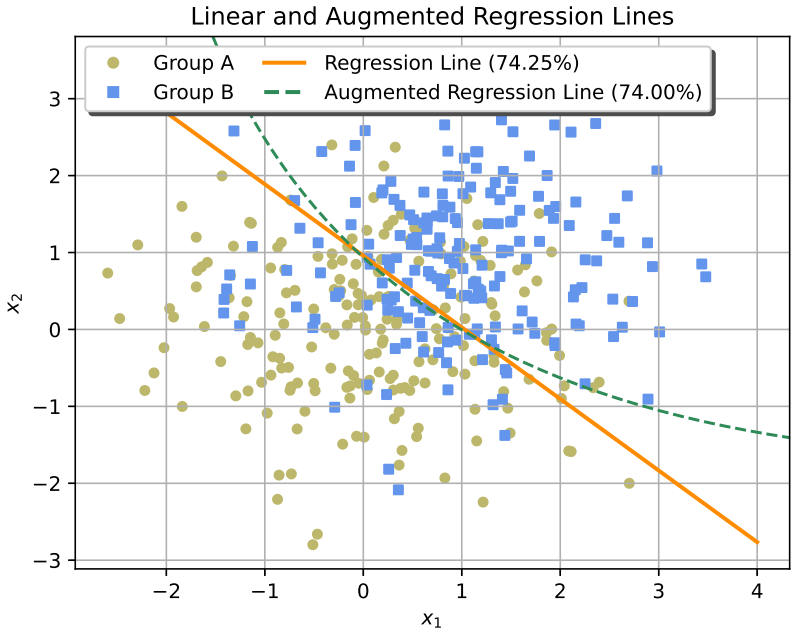

Figure 11: 兩種迴歸模型的分界線比較 (Data 5)

DATA 6.

Group A：

$$n_1 = 300，\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}，\boldsymbol{\Sigma_1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Group B：

$$n_2 = 100，\boldsymbol{\mu_2} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}，\boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Group A 為樣本數 $n_1 = 300$，x_1、x_2 的變異數為 1，且相互獨立的二元常態母體；Group B 為樣本數 $n_2 = 100$，x_1、x_2 的變異數為 1，且相互獨立的二元常態母體。由 Figure 12 可以發現兩對母體僅有少數資料點重和，兩群資料距離較遠，且 Group B 的樣本數相較於 Group A 明顯較少。而對於此種資料型態，由 Figure 12 可以看到兩種迴歸的分界線的表現交友不錯的效果，而加廣型迴歸的分界線沒有和線性迴歸的分界線重合，但也將大部分資料分割來。其中，加廣型迴歸的分界線的判別正確率為 95.50 %，線性迴歸的判別正確率為 96.25 %。因此，使用線性迴歸進行預測會得到較佳的結果。

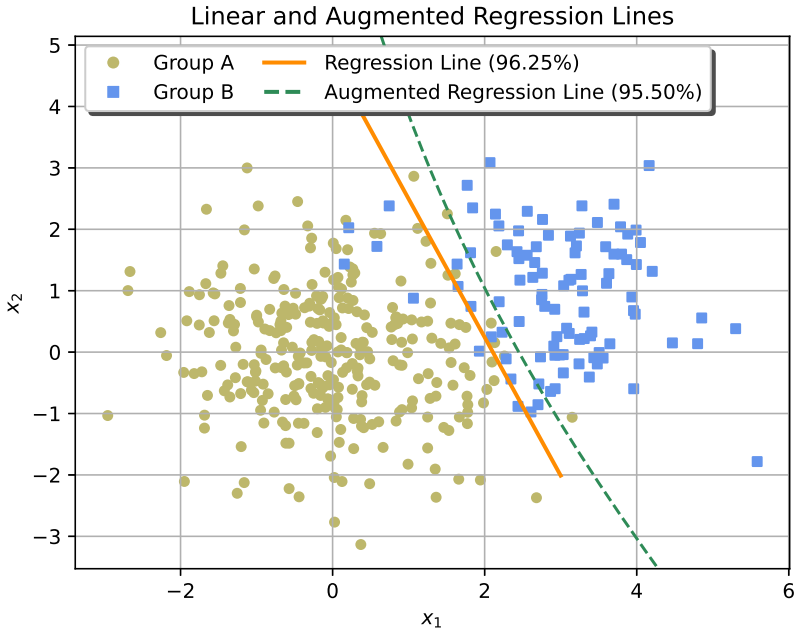

Figure 12: 兩種迴歸模型的分界線比較 (Data 6)

DATA 7.

Group A：

$$n_1 = 500，\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}，\boldsymbol{\Sigma_1} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$

Group B：

$$n_2 = 500，\boldsymbol{\mu_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}，\boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Group A 為樣本數 $n_1 = 500$，x_1、x_2 的變異數為 5，且相互獨立的二元常態母體；Group B 為樣本數 $n_2 = 500$，x_1、x_2 的變異數為 1，且相互獨立的二元常態母體。由 Figure 13 可以發現 Group B 完全被 Group A 圍繞，並且 Group B 的資料集中，而 Group A 較為分散。而對於此種資料型態，由 Figure 13 可以看到兩種迴歸的分界線是完全不一樣，加廣型迴歸的分界線為橢圓分界。其中，加廣型迴歸的分界線的判別正確率為 77.30 %，線性迴歸的判別正確率為 68.50 %，兩有較大的差異。因此，使用加廣型迴歸進行分類會得到較佳的結果。

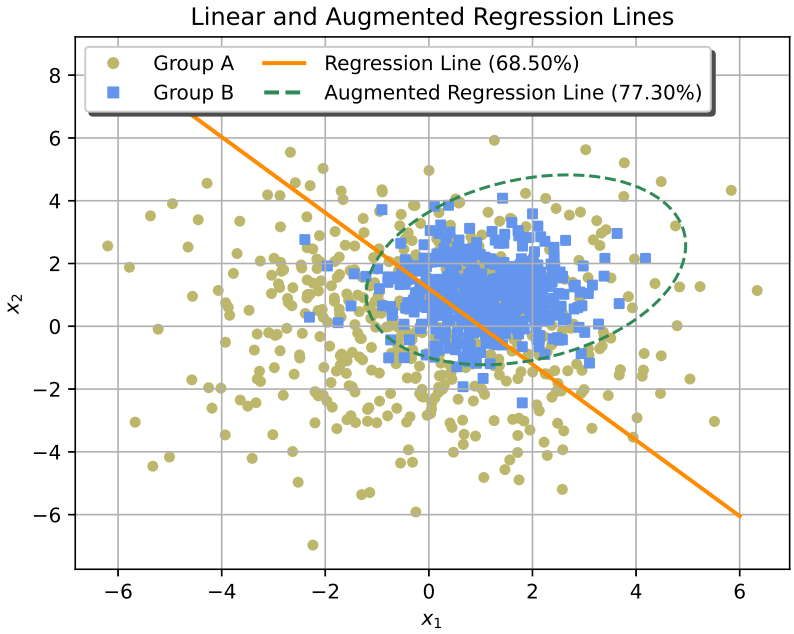

Figure 13: 兩種迴歸模型的分界線比較 (Data 7)

3.1 三個群組

Group A：

$$n_1 = 300，\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}，\boldsymbol{\Sigma_1} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Group B：

$$n_2 = 200，\boldsymbol{\mu_2} = \begin{bmatrix} -4 \\ -2 \end{bmatrix}，\boldsymbol{\Sigma_2} = \begin{bmatrix} 2 & -0.7 \\ -0.7 & 2 \end{bmatrix}$$

Group C：

$$n_3 = 100，\boldsymbol{\mu_3} = \begin{bmatrix} -4 \\ 2 \end{bmatrix}，\boldsymbol{\Sigma_3} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Group A 為樣本數 $n_1 = 300$，x_1、x_2 的變異數為 2，且相互獨立的二元常態母體；Group B 為樣本數 $n_2 = 200$，x_1、x_2 的變異數為 2，且相依的二元常態母體；Group C 為樣本數 $n_3 = 100$，x_1、x_2 的變異數為 1，且相互獨立的二元常態母體。Figure 14 為利用兩兩分群，繪製出三個群組資料的線性迴歸分界線，分別為 Group A & Group B、Group B & Group C 以及 Group A & Group C。

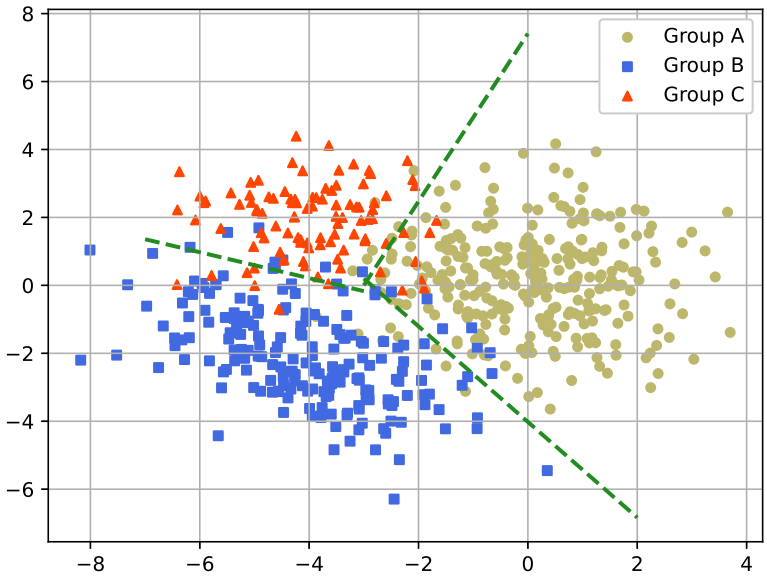

Figure 14: 三個群組的資料分群

4 迴歸模型比較

本節將整理以上兩種迴歸的分界線，在不同資料下的判別正確率，並推測出針對不同迴歸的分界線，較適用於何種資料的判別預測。

由表 1，可以發現 Data1、Data 5 以及 Data 6 都是線性迴歸的判別正確率較高；而 Data 2、Data 3、Data 4 以及 Data 7 則是加廣型迴歸的判別能力較好。並從表 2 中，可以大致歸納出在 Data1、Data 5 以及 Data 6 中，它們的兩個母體的變數 x_1、x_2 之變異皆相同，且變數 x_1、x_2 間相互獨立；而 Data 2、Data 3、Data 4 以及 Data 7 中，它們的兩個母體的變異可能不相同，且變數 x_1、x_2 間不一定獨立。另外，兩個母體的距離對於迴歸的分界線可能影響不大。

Table 1: 線性迴歸與加廣型迴歸分界線的判別正確率比較

判別正確率 (%)	DATA 1	DATA 2	DATA 3	DATA 4	DATA 5	DATA 6	DATA 7
線性迴歸	94.00	91.25	93.25	90.00	74.25	96.25	68.50
加廣型迴歸	93.75	92.25	95.25	91.50	74.00	95.50	77.30

Table 2: 資料型態

Group		DATA 1	DATA 2	DATA 3	DATA 4	DATA 5	DATA 6	DATA 7
A	樣本數	200	200	200	200	200	300	500
	變異數	1	2	1	1	1	1	5
	變數間的關係	獨立	獨立	獨立	相依	獨立	獨立	獨立
B	樣本數	200	200	200	200	200	100	500
	變異數	1	1	5	1	1	1	1
	變數間的關係	獨立	相依	獨立	相依	獨立	獨立	獨立
兩母體的距離		較遠	較遠	較遠	較遠	較近	較遠	較近