Supervised Learning: the Deterministic Approach

Linear and Augmented Regression Model

August 10, 2020

機器學習（Machine Learning）依資料的特性概分兩種型態：監督式學習（Supervised Learning）與非監督式學習（Unsupervised Learning）。從資料分析與學習的角度來看也分為兩種：決定性（Determinstic）與機率性（Probabilistic）。[1] 本章舉資料的群組分類（Classification）為例，討論採取決定性模型的監督式學習，並著重在最簡單的兩群組資料的判別。透過幾個簡單、典型的方法，實際去做群組的判別。過程中對 MATLAB 程式設計的技巧（依據理論來寫作程式及使用 MATLAB 的指令）、資料的產生及圖形的繪製都有進一步的延伸，也是本章真正的目的。

本章將學到關於程式設計

群組資料的繪製技巧、排序資料的索引技巧及最小平方法的矩陣計算方式。Machine Learning 套件使用、特殊線性方程式的繪圖、MATLAB 矩陣式的計算技巧。

〈本章關於 MATLAB 的指令與語法〉

指令：fitlm, mvnrnd, predict。

[1] 決定性與機率性的資料分析模式最大的差別在於對資料的假設。當研究者對資料來源掌握更多的資訊，譬如資料來自某個常態分配的母體，考慮類似此類機率型態假設在模型中，稱為機率性的方式（Probabilistic Approach），否則稱為決定性的方式（Determinsitic Approach），譬如本章討論的迴歸模型。

1 背景介紹

監督式學習應用在成對的資料 $(x_i, y_i)_{i=1}^N$，其中 x_i, y_i 分別代表變數 X 與 Y 的樣本資料。監督學習的目的是透過這些已知的資料，確立變數 X 與 Y 之間的相關性，通常以數學式 $Y = f(X)$ 表示，典型的案例常見於迴歸分析與時間序列。所謂「監督式」與「非監督」的差別在因變數 Y 是否已知；譬如，某種體積很小的昆蟲，不易辨別其性別，於是想從其張開的翅膀長度來探知其性別。假設翅膀長度變數為 X，性別變數為 Y，由於性別資料未知，只憑樣本資料 x_i 來推估未知資料 y_i，稱為非監督式學習，意即想從已知資料估計（揭露）未知資料（或稱隱藏資料）。另一方面，若透過某些精密儀器的檢測或長期觀察昆蟲的行為，確認了性別資料 y_i，此時想探索的問題變成翅膀長度是否與性別存在某種關係？如果找不到，也許還必須加入第二個因素，譬如體重。這種從成對的資料中找出其對應關係，稱為監督式學習，如圖 1 的示意圖，未知模型的輸入變數 X_1, X_2 代表翅膀長度與體重（稱為特徵值），輸出變數 Y 代表對應的性別（或稱組別）。

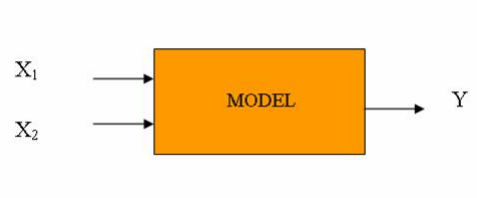

圖 1: 監督式學習式意圖，其中輸出變數 Y 的觀察資料已知

本章討論監督式學習中屬於類別資料的群組分辨（即輸出變數 Y 是類別型的資料），[2]並且著重在最簡單的「兩群組資料判別」。透過典型的迴歸分析方法，實際去做群組的判別。過程中依據理論來寫作程式並繪製相關圖形，讓讀者實際理解機器學習的背景。最後配合 MATLAB 強大的指令與 APP 產品，作為實際操作使用的工具。

為求簡單明瞭起見，本章假設輸入資料具兩個維度（變數），即具 X_1, X_2 的兩個特徵值，且每一筆已知資料的群組別 Y 也是已知。譬如圖 2 顯示 400 筆已知資料，包含輸入（X_1, X_2）與輸出（不同的圖示及顏色代表不同的組別），其關係亦如圖 1 所示。而面臨的問題是，如何從已知的 400 筆成對資料學習 X_1, X_2

[2]輸出資料概分兩種：Quantitative 及 Categorical，歸類問題的屬性時常以此為分別。當輸出是 Quantitative 型的資料，屬於迴歸分析（Regression）的範疇，當輸出是 Categorical，叫做分類（Classification）或分群。輸入資料當然也有不同的類型，不過應用的方法上差別比較小。Regression 與 Classification 在方法上也有許多類似之處，因為在 Categorical 資料的表達上，通常會以數字來代表類別，譬如 1 代表「成功」，0 代表「失敗」。這樣一來兩者的差距變模糊了，Regression 的方法也可以直接套用在 Categorical 的資料上。

與 Y 之間的關係，當給予一組未知群組別的資料時，如何預測其組別？[3]

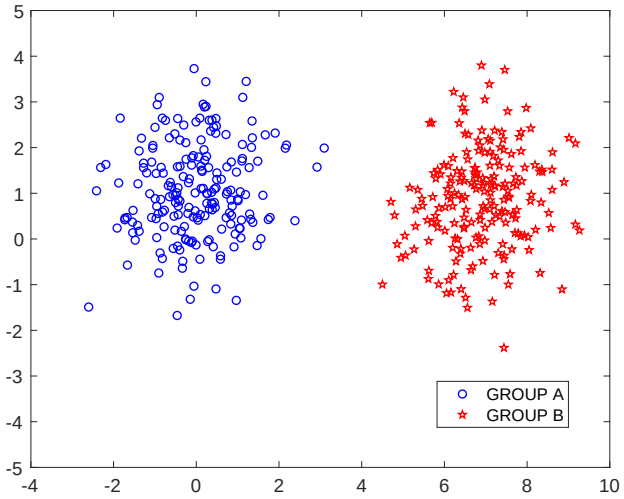

圖 2: 兩群體的輸入（特徵）資料與輸出（組別）資料，共 400 筆

圖 2 的 400 筆資料明顯的將所在的平面空間分成兩半，左半邊屬於群組 A，右半邊屬群組 B。當一筆新的資料需要判別其群組屬性時，只要看它落在平面上的哪一邊，即可判定。但問題是，分割平面空間的分界線如何界定？這條線將做為資料群組預測的根據，但從圖 3 來看，這條分界線似有無限可能，不同的方法形成的分隔線也不同，將如何判斷其優劣呢？[4]

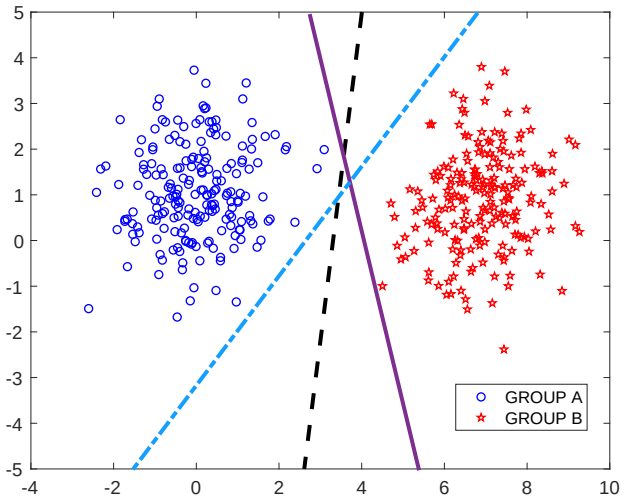

圖 3: 兩個群組的可能分界線

[3]從已知的資料學習，稱為模型的「訓練」階段，從訓練後的模型輸入資料以計算其所屬組別，一般稱為預測。

[4]本章內容僅介紹分隔線的建立，至於不同分隔線的優劣比較留到習題再做分析，請讀者自行寫程式做比較。

本章介紹線性與非線性迴歸模型與最小平方法在群組分析上的應用，試著在兩群組的資料空間劃上一條適當的分界線，並採理論與實作並進的方式逐步完成程式的設計，並介紹 MATLAB 在這方面提供的指令與做法。

1.1 線性迴歸模型

假設圖 1 的輸入輸出關係為「線性迴歸模式」。如果輸出資料屬於類別資料時，譬如，Group A 及 Group B，我們仍可以假設當輸入資料屬於群組 A 時，輸出變數以數字表示，譬如：$Y = 0$，另一個群組則為 $Y = 1$。將類別資料量化之後的問題，便可以直接套入以下的線性迴歸模式（Linear Regression Model）來分析，[5]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{1}$$

假設共有 N 筆已知的輸入與輸出資料（$[x_1(i)\ \ x_2(i)], y(i)$），則迴歸係數 $\beta_0, \beta_1, \beta_2$ 以最小平方法求得的最佳解為

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

其中

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1(1) & x_2(1) \\ 1 & x_1(2) & x_2(2) \\ \vdots & \vdots & \vdots \\ 1 & x_1(N) & x_2(N) \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} \tag{3}$$

分別代表迴歸模型模型的參數估計、輸入及輸出資料矩陣。式 (2) 假設反矩陣 $(X^T X)^{-1}$ 存在，而每個輸出值 $y(i)$ 根據其類別，非 0 即 1。一旦迴歸模型的參數 $\hat{\beta}$ 估計完成，我們說機器完成了學習。接下來便可以拿來對新資料做群組屬性的判別（預測）。規則如：

[5]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）」的模式。

群組判別：當給予一個新的輸入資料 $x = (x_1, x_2)$，根據迴歸模型 (1)，其輸出擬合值為：

$$\hat{y} = \mathbf{x}^T \hat{\beta} \tag{4}$$

其中 $\mathbf{x}^T = [1\ x_1\ x_2]$。在迴歸模型下的擬合值 $\hat{y}$ 不一定剛好是 0 或 1，它可以是任何數值，但作為類別判斷時，可以依下列規則判別：假設 G 代表判定的類別：

$$G = \begin{cases} \text{Group } A \ \text{ if } \ \hat{y} \leq 0.5 \\ \text{Group } B \ \text{ if } \ \hat{y} > 0.5 \end{cases}$$

換句話說，上述規則以 $\hat{y} = \mathbf{x}^T \hat{\beta} = 0.5$ 做為平面空間中兩個群組的分界線，將 $\Re^2$ 平面一分為二，線的一邊以集合 $\{\mathbf{x} | \mathbf{x}^T \hat{\beta} \leq 0.5\}$ 代表 Group A，另一邊則為 Group B。很明顯的，這條分界線的形成受到下列因素的影響：

- 已知資料 X 與 $\mathbf{y}$
- 迴歸模式 (1)
- $\hat{\beta}$ 的估計方法，即式 (2) 的最小平方法。

以下練習舉兩組資料為例（從網頁下載 la_1.txt, mix.mat 兩組資料），[6]如圖 4，協助初學者如何畫出群組分佈圖、計算 $\hat{\beta}$ 值與及分界線。

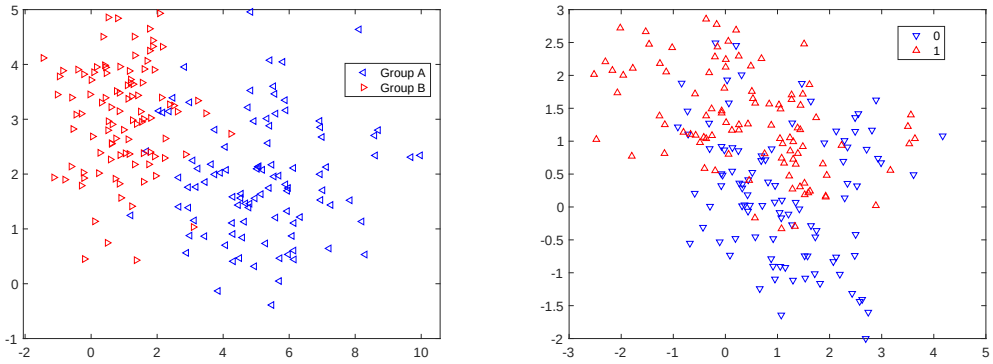

圖 4: 群組資料 la_1.txt（左）與 mix.mat（右）

在計算出分界線之前，通常會先將資料畫出來觀察其群組關係。當然這僅限於兩個輸入變數以下的情況。左邊的資料 (la_1.txt) 是模擬出來的，右邊 (mix.mat) 則來自參考文獻 [1] 的提供的資料。以下練習協助畫出圖 4。

[6]https://ntpuccw.blog/supplements/matlab-in-statistical-computing/

範例 1. 根據輸出資料 Y 的類別，在 X1-X2 平面上以不同顏色或符號描繪出群組的散佈圖。

MATLAB 畫散佈圖的指令有 plot, scatter 及 gscatter，其中 gscatter 適合繪製群組散佈圖。使用方式如下：

```
D=load('la_1.txt');
gscatter(D(:,1),D(:,2),D(:,3),'br','<>')% 群組散佈圖
```

資料檔 la_1.txt 的結構為一矩陣，前兩行代表 X 資料，第三行代表群組別的 0,1 資料。不難看出指令 gscatter 的參數順序代表的意思。後兩個參數代表顏色與符號。[7]

指令 gscatter 的符號選擇有限，如果想做出特殊的符號，如圖 5 所示，必須想想辦法。參考程式碼如下：

```
figure
D=load('la_1.txt');
y=D(:,3);
G1=D(y==0,1:2); G2=D(y==1,1:2);% 分出群組資料
axis([min(D(:,1))-1 max(D(:,1))+1 min(D(:,2))-1 ...
      max(D(:,2))+1]);
H=text(G1(:,1),G1(:,2),'A');
set(H,'Color','blue'); set(H,'fontsize',14)
E=text(G2(:,1),G2(:,2),'B');
set(H,'Color','red'); set(H,'fontsize',14)
```

指令 text 適用在圖形上做標記或文字說明，不能單獨使用，因此第一行的指令 figure 用來產生空白圖形，方便 text 的使用。指令 axis 的四個參數分別設定 X 軸與 Y 軸的範圍。有了空白圖形，text 根據前兩個參數代表的平面的座標位置，印出第三個參數的文字，譬如'A'。指令中的 H 代表圖形上的「物件」(Object Handle)，利用 set 指令可以改變其外觀（其他外觀選項可以參考該指令的說明）。

---

[7]顏色與符號的代碼，可以參考手冊或利用快速的查詢方式，作法是先輸入指令至 gscatter(D(:,1), D(:,2), D(:,3), '，然後按 TAB 鍵，旁邊將跳出一個小視窗，呈現所有顏色代碼供選擇。同樣的方式也能用在符號的選擇。

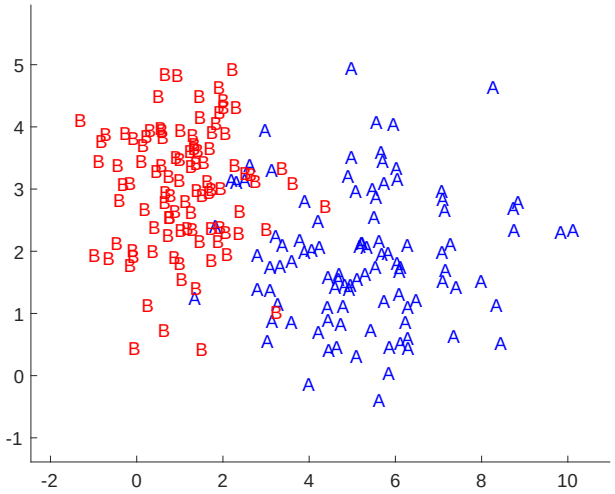

圖 5: 用特別符號繪製群組資料

範例 2. 利用前述範例的資料，估計迴歸模型的參數 (2) 並畫出式 (4) 中 $\hat{y} = 0.5$ 的分界線，也就是劃開兩群組空間的分界線。

計算式 (2) 的 $\hat{\beta}$ 比較簡單，先從原始資料建構資料矩陣 X 與 $\mathbf{y}$，再套入 (2) 的公式即可。接續之前的程式碼，$\hat{\beta}$ 的估計與分界線的呈現如下列程式碼與圖 6。

```
X=[ones(size(y)) D(:,1:2)];
b=(X'*X)\X'*y;% 迴歸係數 beta 的估計
f=@(x1,x2) b(1)-0.5+b(2)*x1+b(3)*x2;% 分界線函數
hold on
fimplicit(f,'LineWidth',3,'Color','black',...
          'LineStyle','--')
hold off
```

其中估計式 (2) 的反矩陣以反斜線取代，這是 MATLAB 的線上輔助系統的建議，用以加速計算。這個估計式也可以用 pseudo inverse 的方式，如[8]

```
b=pinv(X)*y;
```

要畫出兩群組間的分界線 $\hat{Y} = 0.5$，需要琢磨一下。這條分界線的方程式可以表示為

[8]當反矩陣 $(X'X)^{-1}$ 存在時，pinv(X)=(X'*X)\X'。當反矩陣 $(X'X)^{-1}$ 不存在時，pinv(X) 是 (X'*X)\X' 的估計值。

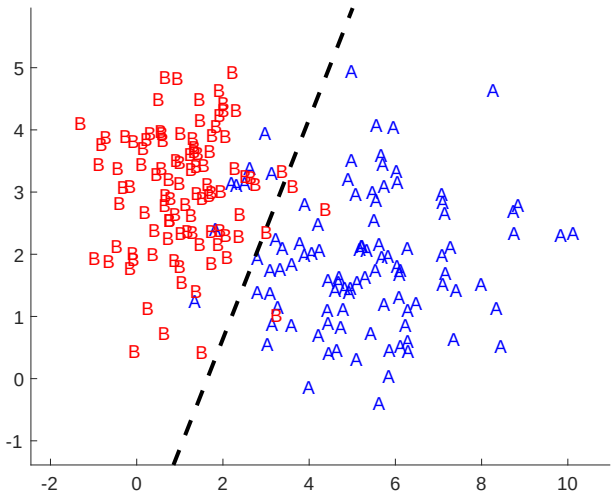

圖 6: 線性迴歸模型的分界線，資料來源 la_1.txt

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 0.5$$

這裏採用隱函數 fimplicit 的方式繪圖。

範例 3. MATLAB 的 Machine Learning 工具箱提供了指令 fitlm 建立了如式 (1) 的線性迴歸模型。本範例展示使用方式，並利用資料檔 mix.mat 示範繪製分界線，如圖 9。

指令 fitlm 可以針對表格型態的資料（table）或一般矩陣資料。本範例的資料檔 mix.mat 為一般矩陣資料，使用方式如下：

```
load mix % 內含資料變數 x,y
gscatter(x(:,1),x(:,2),y,'br','<>')
mdl=fitlm(x,y);% 輸出的資料型態為 LinearModel
TMP=mdl.Coefficients;%迴歸模型參數估計(資料型態為table)
a=TMP{:,{'Estimate'}};% 從 table 資料型態取得內容
f=@(x1,x2) a(1)-0.5+x1*a(2)+ x2*a(3);
hold on
fimplicit(f,'LineWidth',3,'Color','black',...
          'LineStyle','--')
hold off
```

上述程式碼中，指令 fitlm 將執行結果輸出到一個資料型態為 LinearModel 的變數資料 mdl，其內容如圖 7 所示。這是一個典型的迴歸分析報表，清楚地

呈現迴歸模型與所有的統計量。想要進一步取得裡面的數字，必須進入變數的結構裡面，如指令 mdl.Coefficients 取得線性迴歸模型的參數估計值（$\hat{\beta}$）。mdl.Coefficients 的結果是一個表格（table）資料，如圖 8 所示。欲取得表格資料的內容，可以參考上述程式碼 a=TMP{:, {'Estimate'}}，也就是取得表格欄位名稱為 Estimate 的那一欄所有資料。最後利用這一組參數估計值，畫出圖 9 那條分界線，也就是 $y = 0.32906 - 0.022636x_1 + 0.2496x_2$。

```
>> mdl

mdl =


Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate        SE        tStat        pValue
                   ________    ________    ________    __________

    (Intercept)      0.32906    0.047829      6.8799    7.7719e-11
    x1             -0.022636    0.025429    -0.89015       0.37447
    x2                0.2496    0.032147      7.7643    4.3807e-13


Number of observations: 200, Error degrees of freedom: 197
Root Mean Squared Error: 0.425
R-squared: 0.289,  Adjusted R-Squared 0.282
F-statistic vs. constant model: 40.1, p-value = 2.5e-15
```

圖 7: 資料型態為 LinearModel 的內容

```
>> TMP

TMP =

  3×4 table

                   Estimate        SE        tStat        pValue
                   ________    ________    ________    __________

    (Intercept)      0.32906    0.047829      6.8799    7.7719e-11
    x1             -0.022636    0.025429    -0.89015       0.37447
    x2                0.2496    0.032147      7.7643    4.3807e-13
```

圖 8: TMP=mdl.Coefficients 的表格內容

讀者可以比較前述兩個範例，若針對同一筆資料，是否得到相同結果？答案是肯定的。這也是從事研究工作者檢驗軟體提供的指令適用與否的方式，進而能更準確地掌握該指令的所有功能，而非盲目，甚至無知地使用。

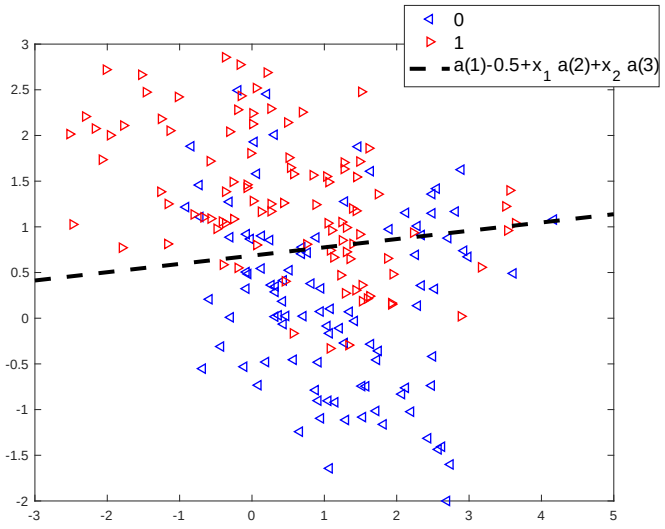

圖 9: 線性迴歸模型的分界線：資料檔 mix.mat

1.2 加廣型迴歸模型（Augmented Regression Model）

從圖 9 中資料分布的情況，很清楚的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）」。想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，這是一條非線性的分界線，也許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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 \tag{5}$$

其參數的估計與群組判別的方式與線性模式相同，在程式的編寫上只需在資料矩陣 **X** 再加入三欄分別來自 X_1X_2, X_1^2, X_2^2 的資料，即

$$X = \begin{bmatrix} 1 & x_1(1) & x_2(1) & x_1(1)x_2(1) & x_1^2(1) & x_2^2(1) \\ 1 & x_1(2) & x_2(2) & x_1(2)x_2(2) & x_1^2(2) & x_2^2(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1(N) & x_2(N) & x_1(N)x_2(N) & x_1^2(N) & x_2^2(N) \end{bmatrix}$$

接著計算參數 $\hat{\beta}$ 的最小平方估計：$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$，如同線性迴歸模式，則

式 (5) 的加廣型迴歸模型的分界線表示為集合

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

在平面上如圖 10 那條彎彎的線，方程式為

$$f(X_1, X_2) = \hat{\beta}_0 - 0.5 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 + \hat{\beta}_4 X_1^2 + \hat{\beta}_5 X_2^2 = 0$$

讀者可以自己寫程式計算上述分界線的係數 $\hat{\beta}$ 或直接採用 MATLAB 指令 fitlm 與隱函數的繪圖技巧，參考程式碼如：

```
load mix
gscatter(x(:,1),x(:,2),y,'br','<>')
mdl=fitlm(x, y, 'quadratic');% 選擇 quadratic 模型
TMP=mdl.Coefficients; % 迴歸模型參數估計結果
a=TMP{:,{'Estimate'}};% 從 table 資料型態取得內容
f=@(x1,x2) a(1)-0.5+x1*a(2)+ x2*a(3)+a(4)*x1.*x2+ ...
           a(5)*x1.^2+a(6)*x2.^2;
hold on
fimplicit(f,'LineWidth',3,'Color','black',...
            'LineStyle','--')
hold off
```

讀者可以從 TMP 的表格內容讀出迴歸係數 a 的順序與變數的關係。如此才能順利製作函數 f。迴歸模型指令 fitlm 的模式選項除了內定的'linear' 及上述程式碼使用的'quadratic' 外，還有其他四個選項，有興趣的讀者請參考使用手冊中關於 fitlm 的 model spec 選項。

1.3 學習器評比

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。而學習器的選擇、訓練與評比是機器學習的重要步驟。前兩節選擇兩種學習器並經過模擬或實務資料的訓練，接著我們想知道（一）學習器從訓練資料學習得有多好？（二）經過訓練後的學習器面對不同的資料時，表現如何？

圖 9 的直線與圖 10 的彎曲分界線，分別代表線性迴歸模型與加廣型迴歸模型兩種學習器的學習（訓練）成果。哪一條線的分群效果較好呢？或說，何者對於訓練資料的分群誤差比較小呢？其實直接將迴歸模型用於分群，有點突兀，理由是如式 (1) 與式 (5) 的迴歸模型，主要用在反應變數為連續型的資料，而非群

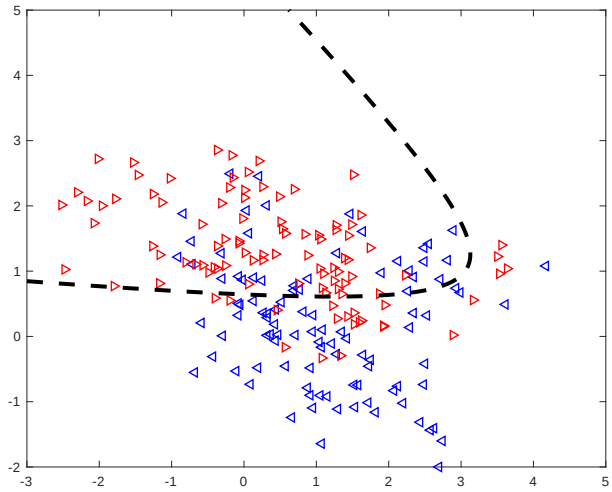

圖 10: 加廣型迴歸的非線性群組分界線

組型的類別資料。[9]當考慮迴歸模型對訓練資料的配適性時，不會參考如圖 7 所示關於迴歸模型的誤差指標，如 Root Mean Square Error、R-Squared 或 Adjusted R-Squared 等。反而是計算學習器對於訓練資料的群組錯判率，[10]才是觀察重點。以下程式碼針對資料檔 mix.mat 以線性迴歸模型（linear）與加廣型迴歸模型（quadratic）計算分界線的錯判率，執行結果同時顯示在圖 11。

```
n=length(y);% 總樣本數
mdl1=fitlm(x,y);                    % linear model
mdl2=fitlm(x,y,'quadratic');% quadratic model
y_linear=zeros(n,1);y_quad=y_linear;% 準備填裝分群結果
y_linear(mdl1.Fitted>0.5)=1;%從擬合值轉換成群組別
y_quad(mdl2.Fitted>0.5)=1;
Err_linear=sum(abs(y - y_linear))/n;
Err_quad=sum(abs(y - y_quad))/n;
```

從圖 11 顯示的結果發現，加廣型迴歸模式的誤差率反而稍大一些。也就是說，在訓練階段，加廣型模式的表現稍差一點。不過從機器學習的角度來看，這不代表面對新的資料時，也會表現比較差。本章後面的習題請讀者將圖 11 的資料檔 mix.mat 分成兩部分（比例 8:2)，佔八成的資料分別用來訓練兩個迴歸模型，並計算錯判率。之後，利用另外兩成資料當作新資料，做為測試之用。測試資料的錯判率往往才是好與壞的參考指標。[11]

[9]群組型的類別資料常使用羅吉斯迴歸模型（Logistic Regression）。

[10]學習器學得好不好，一般以訓練誤差來評比。以圖 9 的直線分界線為例，這條線是訓練後作為分群的依據。群組錯判率便是計算錯誤分群的資料比例。

[11]當資料必須按比例被分為訓練與測試兩分資料時，若資料量不夠大，將造成訓練資料不足，導致學習成效不彰。

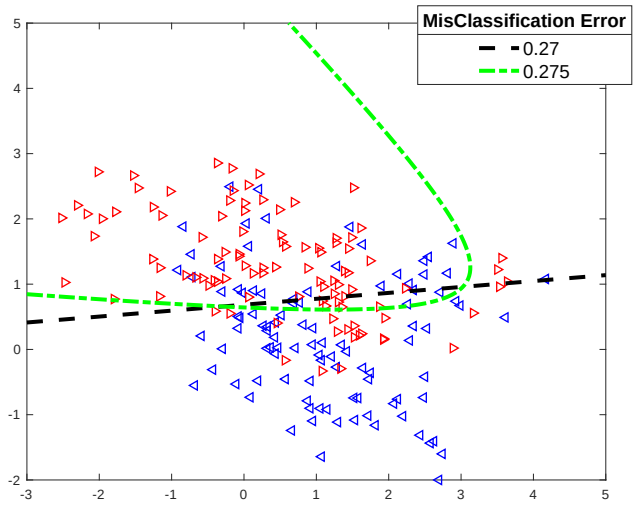

圖 11: 兩個迴歸模型的分群錯誤率比較

擬合值 $\hat{y}$ 是訓練資料所產生的輸出值，而且是連續數值，不是類別值，所以上述程式碼從擬合值是否大於 0.5 來分群。擬合值（mdl.Fitted 也在 fitlm 計算時一併被記錄了。如果面對測試的新資料，MATLAB 的 Machine Learning 套件一概使用指令 predict 來做預測。假設 x 是 n x 2 測試資料矩陣，y 是 n x 1 群組資料，則分群結果的錯判率計算如下：

```
mdl1=fitlm(x,y);       % linear model
M=predict(mdl,x);      % 擬合值或稱訓練資料的預測值
y_predict=zeros(n,1);% 準備填裝分群結果
y_predict(M>0.5)=1;    % 從擬合值轉換成群組別
Err_predict=sum(abs(y - y_predict))/n;
```

2 觀察與延伸

1. 當分界線劃上去之後，有多少資料被錯置組別呢？錯置的資料愈多，代表什麼意義？當兩個群組部分交錯時，資料的錯置是否不可避免？有更好的分界線可以讓錯置的情況降低嗎？這些都是我們考量使用何種學習器，必須思考的問題。

2. 使用已知的資料做出一條分界線，企圖將原母體在空間中的範圍切割出來。這個切割的好壞當然取決於已知資料的品質及分界線的決定方式。試試看給予一些新的資料（從原母體去產生），測試一下這條分割線能否對新的資料做出正確的組別判斷？譬如 100 個新資料有多少比率被正確辨別？

3. 由於資料的取得誤差或樣本數不夠，群組的區隔有時候不是很明顯，當然也可能是群組本身就非常靠近。圖 4 左圖的資料看起來分離的很好，直覺上比較容易作區域的切割，如中間的那一條分界線。而右圖的兩個群組相對緊密，即使能劃上一條分隔線，也可能必須選擇曲線比較能滿足現有資料能提供的訊息。而根據有限的資料做出最好的判斷，就是這門學問的精神所在。

4. 當群組數量大於 2 時，分界線將如何切割？想一想。手癢的話就動手做看看吧！

5. 本單元的資料模擬自雙變量常態的母體（Bivariate Normal Distribution），而且兩個變數是獨立的。如果變數間有相依性，本單元的方法還是可行嗎？如何去模擬具相依性的資料呢？模擬資料的產生可參考下列程式碼。

```
n1=200; n2=200; n=n1+n2;
mu1=[0 1]; mu2=[4 1];
Sigma=[1 0.2;0.2 1];
y=[zeros(n1,1);ones(n2,1)];% 製作群組標示
A=mvnrnd(mu1,Sigma,n1);    % 第一組資料
B=mvnrnd(mu2,Sigma,n2);    % 第二組資料
X=[A;B];                   % 資料矩陣：n x 2
gscatter(X(:,1),X(:,2),y,'br','op')% 群組散佈圖
save trainingData X y
```

最後一行將訓練資料儲存為檔名 trainingData 的 mat 檔案格式，內含兩個資料變數 X 與 y。讀者可以從這 n 筆資料撥出一部份當測試資料，或是再執行一次，產生所需的測試資料，檔名改為 testingData。

3 習題

1. 證明式 (2) 是迴歸模型 (1) 的最小平方法解，即

$$\hat{\beta} \doteq arg\min_{\beta} ||X\beta - \mathbf{y}||^2$$

2. 當資料來源為 la_2.txt，請畫出圖 12 的資料散佈圖與兩條分界線並計算分群的錯誤比率。

3. 圖 11 展示線性迴歸與加廣型迴歸模型在訓練階段的誤差。但依訓練資料所做的錯判率並不能當作判別式好壞的依據，尚須對訓練資料以外的資料做群組判別測試，才能論定。請將現有的資料檔 mix.mat 分成兩部分（比例 8:2)，佔八成的資料分別用來訓練兩個迴歸模型，並計算錯判率。再將

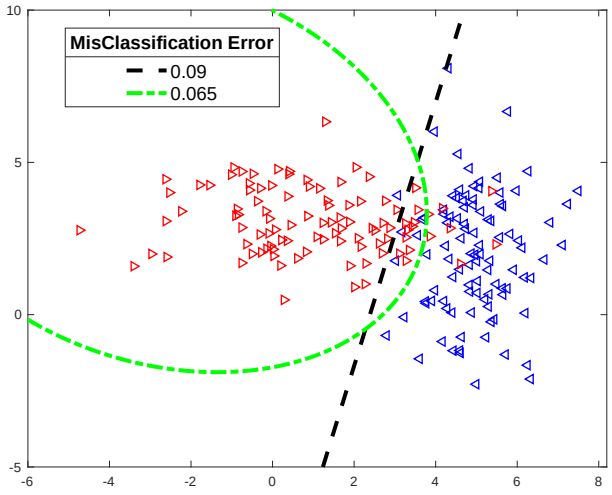

圖 12: 兩個迴歸模型的分群錯誤率比較：資料來源為 la_2.txt

訓練好的判別式測試另外 2 成的資料，再計算出兩個模型針對測試資料的錯判率。

4. 同前題，資料來源換成檔案 la_1.txt、la_2.tex。

5. 模擬資料能提供大量的訓練資料與測試資料，不受限於真實資料有限的樣本。於是在驗證模型分群效果時，常常需要大量製造模擬資料。請安排兩種群組分隔的模擬情境，其一使兩群組分隔較遠，另一種則是較近。資料皆來自雙變量常態的母體，平均數與標準差自訂（可參考上一節的模擬資料的程式碼）。

6. 本章的示範以兩群組資料為主，請讀者自行模擬三個群組的資料，試試指令 fitlm 能否幫忙畫出其間的兩兩的分隔線或是將空間區分為分屬三個群組的區塊。

References

[1] T. Hastie, R. Tibshirani, J. Friedman, ”The Elements of Statistical Learning:Data Mining, Inference, and Prediction,” Springer.

[2] A.G. Rencher, ”Multivariate Statistical Inference and Applications,” John Wiely & Sons, INC.