群組分類

線性迴歸與最小平方法

last modified July 22, 2008

本單元討論「Supervised Learning」中屬於類別 (即輸出變數 Y 是類別型的資料) 資料的群組分辨, 並且著重在最簡單的「兩群組 (two classes)」資料判別。透過幾個簡單、典型的方法, 實際去做群組的鑑別。過程中對 Matlab 程式設計的技巧、資料的產生及圖形的繪製都有進一步的延伸, 也是本課程眞正的目的。

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。

⟨本章關於 MATLAB 的指令與語法⟩

指令:sort, set, gscatter, mvnrnd

1 背景介紹

許多應用科學牽涉到從資料 (data) 中分析出所需要 (含) 的資訊 (information)。希望從已知的資料中瞭解問題的本質, 進而能控制或做出預測。這些資料通常有兩種型態; 其一, 包含特性 (features) 資料及結果 (outcome) 資料, 從收集或量測問題的特性 (或特徵) 資料及相對的結果資料分析出兩者的關係, 並進一步計算相關的參數, 最後確立模型。當給予新的特徵資料時, 便可以根據這個確立的模型產生結果做爲預測。由於有結果資料做爲模型建立的根據, 這些問題歸類爲「Supervised Learning」。譬如圖 1的示意圖, 未知模型的輸入變數X_1, X_2 代表特徵值, 輸出變數 Y 代表對應的組別。

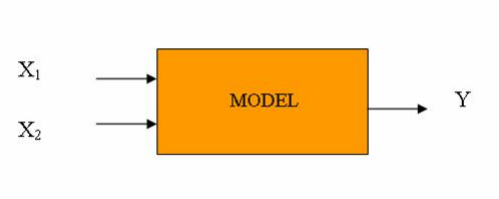


圖 1: supervised Learning

其二, 只有特徵資料, 不知其群組屬性。由於沒有明確的輸出結果做爲對照, 這類問題相對的困難, 稱爲「Unsupervised Learning」, 通常要先從特徵資料裡去找出隱藏的群組關係, 一般也稱爲「Clustering」。

本單元討論「Supervised Learning」中屬於類別 (即輸出變數 Y 是類別型的資料) 資料的群組分辨, [1] 並且著重在最簡單的「兩群組 (two classes)」資料判別。透過幾個簡單、典型的方法, 實際去做群組的鑑別。過程中對 Matlab 程式設計的技巧、資料的產生及圖形的繪製都有進一步的延伸, 也是本課程眞正的目的。

[1]輸出資料概分兩種:quantitative 及 qualitative, 歸類問題的屬性時常以此爲分別。當輸出是 quantitative 型的資料, 屬於迴歸分析 (regression) 的範疇, 當輸出是 qualitative, 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, 因爲在 qualitative 資料的表達上, 通常會以數字來代表, 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的方法也可以用在 qualitative 的資料上。

爲求簡單起見, 假設輸入資料具兩個維度, 即具X_1, X_2 的兩個特徵值, 且每一筆已知資料的群組別也是已知。譬如圖 2 顯示 200 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, 當給予一組未知群組別的資料時, 如何預測其組別?

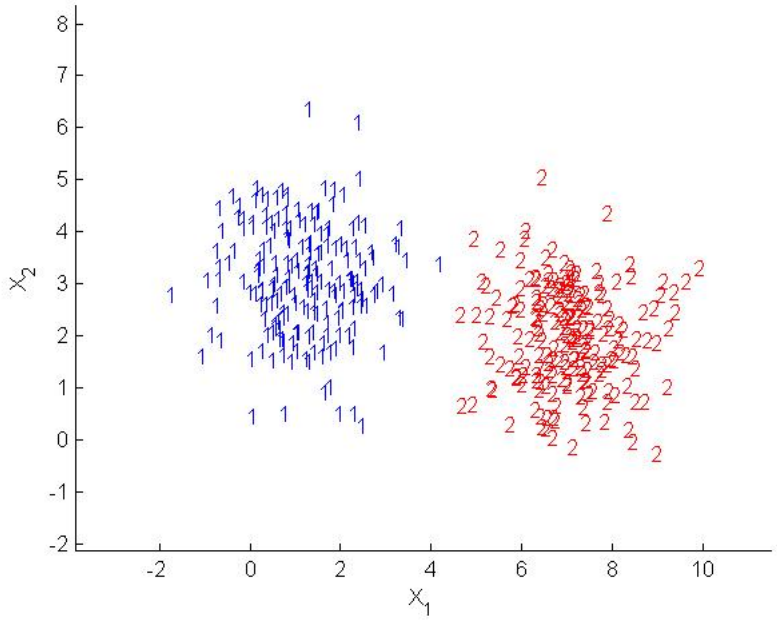


圖 2: 每群各有 100 筆資料的兩個群組

圖 2 的 200 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 1, 右半邊屬群組 2。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, 分割平面空間的界線 (separate line) 如何界定? 這條線將做爲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?

要在兩群組的資料間劃上一條適當的分界線, 有一些簡單的方法要在這個單元介紹, 並以理論與實作並進的方式逐步完成程式的設計。之後的單元接陸續介紹其他方法做群組的鑑別, 本單元介紹線性迴歸模型與最小平方法在群組分析上的應用。

假設圖 1 的輸入輸出關係爲「線性迴歸模式,」雖然輸出資料屬於類別資料 (Class 1 及

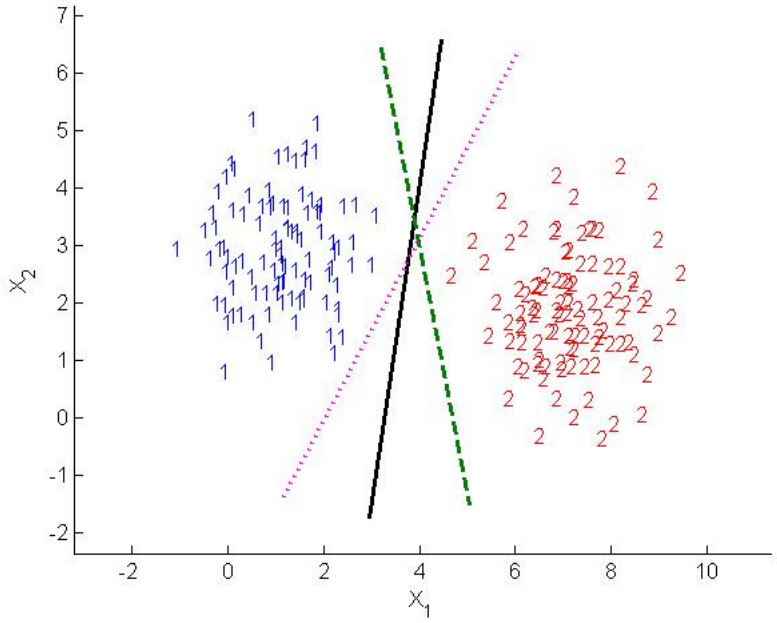


圖 3: 兩個群組的可能分界線

Class 2), 我們仍可以假設當輸入資料屬於群組 1 (Class 1) 時, 輸出變數以數字表示, 譬如:$Y = 0$, 另一個群組則爲 $Y = 1$。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,

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

根據 N 筆已知的輸入輸出資料, 迴歸係數 $\beta_0, \beta_1, \beta_2$ 以最小平方法求得的最佳解爲

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

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。式 (2) 假設 $(X^T X)^{-1}$ 存在, 而每個輸出值 $y(k)$ 根據其類別, 非 0 即 1。

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) 爲:

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

其中 $\mathbf{x}^T = [1\ x_1\ x_2]$。如何從這個輸出值判斷資料得群組屬性呢? 在迴歸模型下的擬合值不一定剛好是 0 或 1, 它可以是任何數值, 但作爲類別判斷時, 可以依下列規則判別: 假設 G 代表判定的類別:

$$G = \begin{cases} CLASS1 \ \text{ if } \ \hat{Y} \leq 0.5 \\ CLASS2 \ \text{ if } \ \hat{Y} > 0.5 \end{cases}$$

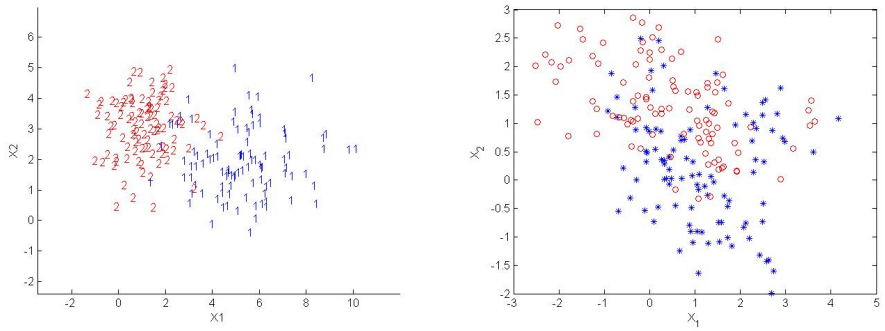
換句話說, 以 $\hat{Y} = \mathbf{x}^T \hat{\beta} = 0.5$ 做爲平面空間中兩個群組的分界線, 將 R^2 平面一分爲二, 線的一邊表示爲集合 $\{\mathbf{x} | \mathbf{x}^T \hat{\beta} \leq 0.5\}$ 爲 CLASS1, 另一邊則爲 CLASS2。很明顯的, 這條分界線的形成受到下列因素的影響:

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

以下練習協助初學者如何計算 $\hat{\beta}$ 值與畫出群組分佈圖及分界線。

2 練習

舉兩組資料爲例 (從網頁下載 la_1.txt, mix.mat 兩組資料), 如圖 4所示。



(a) (b)

圖 4: 群組資料

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。以下練習可以協助畫出上面的圖。

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「樣子,」如上面看到的兩張圖。

在實際的應用上, 資料的來源常不是自己可以控制的, 因此必要的時候必須做調整, 才能讓寫好的程式順利執行。這裡特別設計不同的資料 la_1.txt、mix.mat 對輸出類別資料的安排不一樣, 程式的寫作也因之有所不同。mix.mat的資料已經按類別排序好, 內含兩個變數資料共200筆資料, 分別是輸入的特徵資料 x 與輸出的群組資料 y, 其中 y 前 100 筆值爲 0 代表群組 1, 後 100 筆值爲 1 代表群組 2。下列指令以散佈圖的方式畫出圖 4(b)

```
load mix
x1=x(:,1);
x2=x(:,2);
plot(x1(1:100),x2(1:100),'*')
hold on
plot(x1(101:200),x2(101:200),'or')
hold off
```

第2個 plot 指令的第三個參數 'or'代表以紅色 (r) 英文字母'o'爲符號描點。

另外, 圖4(a) 呈現模擬資料 la_1.txt 的散佈圖。資料 la_1.txt 是一個 200×4 的矩陣, 前兩行代表兩個輸入的特徵值, 後兩行代表輸出的群組值。由於 la_1.txt 資料不按輸出類別資料排序, 作圖時可以

1. 先根據類別排序 (指令:[Y,I]=sort(y)), 記得輸入資料 x1,x2 也必須跟著排序 (程式第3行), 譬如

```
D =load('la_1.txt');
[ Y,I ]=sort(D(:,3)); %按第3行排序, 由小到大
D=D(I,:); %根據排序的索引值 I, 重新排列原資料矩陣
x1=D(:,1); %排列後的輸入資料
x2=D(:,2);
plot(x1(1:100),x2(1:100),'*')
hold on
plot(x1(101:200),x2(101:200),'or')
hold off
```

2. 完全不排序, 根據輸入的群組別, 直接寫一個迴圈將每一筆資料畫上去。

繪製如圖4的散佈圖, 方式很多, 主要是藉由不同的符號或顏色來區分群組。若要呈現如 (a) 圖有顏色的數字, 可以利用 text 指令取代 plot

```
axis([min(x1)-1  max(x1)+1  min(x2)-1  max(x2)-1])
H=text(x1(1:100),x2(1:100),'1');
set(H,'color','blue')
H=text(x1(101:200),x2(101:200),'2');
set(H,'color','red')
```

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 axis 指令用來產生空白圖形, 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 $(x1, x2)$ 代表的平面的座標位置, 印出第三個參數的文字, 譬如'1'。當 x1,x2 是 $1 \times N$ 或 $N \times 1$ 的向量變數時, 可以同時在 N 個座標位置根據第三個參數印上相同或不同的文字。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, 尚有更多的外觀選項可以參考該指令的說明。

MATLAB 也提供了一個方便的指令

```
gscatter(D(:,1),D(:,2),D(:,3),'br','*o')
```

這個指令畫出兩組資料的散佈圖, 其參數的順序與意義分別是: 前兩個參數代表資料、第三個參數以 0, 1 代表每筆資料不同的群組, 第四個參數代表兩個顏色, 第五個參數則是散佈圖的兩個符號。gscatter 簡單明瞭, 符合一般的需求, 是 MATLAB 使用者第一個選擇。

範例2: 根據範例 1 的資料, 計算迴歸模型的參數 (2) 並畫出式 (4) 中, $\hat{Y} = 0.5$ 的迴歸線, 也就是兩群組間的分界線。

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 $\mathbf{y}$, 再套入反矩陣的指令 inv 即可。接續之前的指令, $\hat{\beta}$ 的估計可以寫成

```
X=[ones(N,1);D(:,1:2)];
y=D(:,3);
beta_hat=inv(X'*X)*X'*y;
```

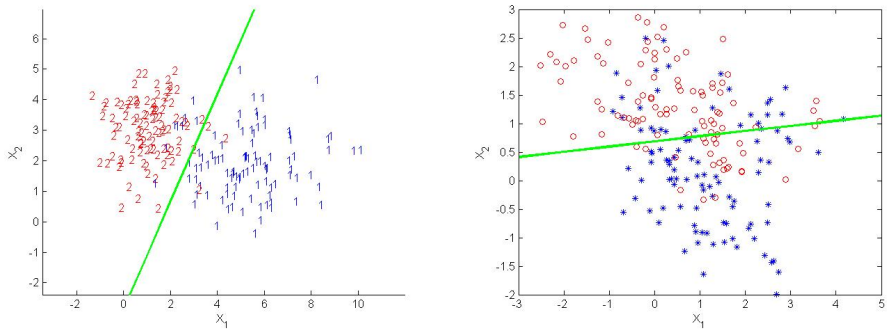
要畫出兩群組間的分界線 $\hat{Y} = 0.5$, 需要琢磨一下。這條分界線的方程式可以表示爲

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

爲繪圖方便, 可以轉換爲

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

再來就是直線繪圖的問題了。圖 5 展示這兩組資料的迴歸分界線。



(a) (b)

圖 5: 群組分界線

3 觀察

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?

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 5(a) 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。而圖5(b)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嗎。

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 Bivariate Normal Data, 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? 參考 MATLAB 關於多變量常態亂數產生器 mvnrnd 的使用方式。

4 作業

1. 證明式 (2) 是迴歸模型 (1) 的最小平方法解, 即

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

2. 畫出範例 2 的分界線。

3. 同上題, 決定出分界線後, 寫一段程式判斷一筆新的資料該屬於那個群組? 新資料的輸入方式以指令 input 在程式執行時取得。

4. 同上題, 寫一支程式計算分界線錯置群組的比例。譬如圖 5(a) 做出分界線後, 劃定右方爲群組 1的區域, 但仍有部分屬於群組 2 的資料落於分界線的右邊, 換句話說, 這條分界線並不能完全隔離這兩個群組的資料, 本題所謂「錯置群組的比例」便是計算這些原本該屬於群組 1(2) 的資料, 卻被分界線劃定在群組 2(1) 的區域, 佔原群組的比例。

5. 將式(1)的迴歸模型, 擴展爲所謂的 Augmented Regression Model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 \qquad (5)$$

同樣利用如式(2)的最小平方法解, 其分界線可以表示爲

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

在 $X_1 - X_2$ 平面上這是一條如圖 6所示的曲線。繪製這條曲線的技巧, 可以將下列的雙變量方程式, 以等高線圖繪製高度 0.5的這條線, 即

$$Z = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 + \hat{\beta}_4 X_1^2 + \hat{\beta}_5 X_2^2$$

繪製 $Z = 0.5$ 的等高線即爲分界線。指令如

```
contour(X1,X2,Z,[0.5 0.5])
```

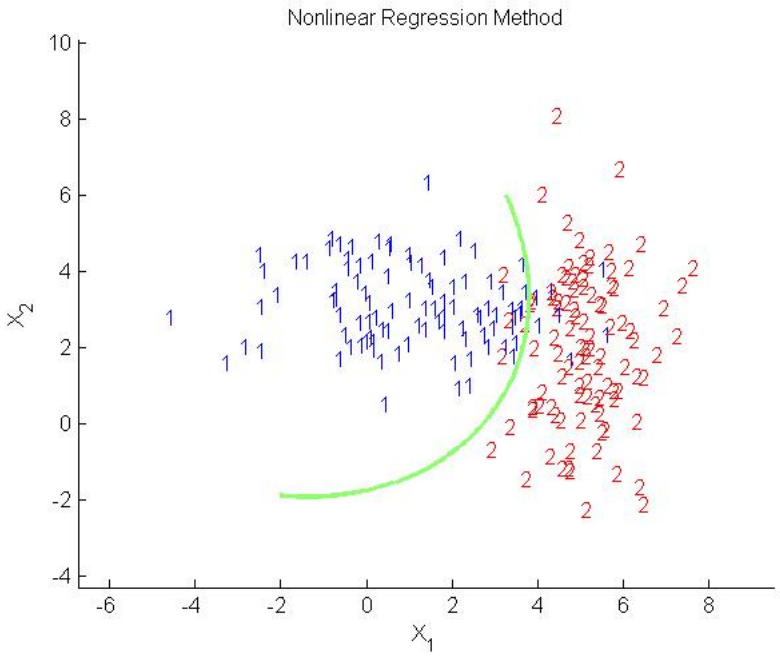


圖 6: Augmented Regression Model 群組分界線

參考文獻

[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.