群組分類: Discriminant Analysis

last modified August 24, 2006

群組分析 (Discriminant Analysis) 的用途非常廣, 譬如, 網路書店[1]想對其一百萬會員促銷一本新書。在成本考量下, 希望能針對可能購買該書的會員做促銷的動作。但如何得知哪些會員購買的可能性較高? 哪些會員根本不可能購買呢? 雖然該網站已有所有會員的資料及過去購買書籍的紀錄, 但對於特定的一本書仍缺乏足夠有利的資訊區分出購買群。

於是該網路書店先抽樣選擇1000名會員進行銷售, 結果有83名會員買了這本書。根據這個結果及這1000人的相關資料, 網路書店便可以進行群組分析, 建立群組分析模式, 最後對於其他大部分的會員 (999,000) 進行群組預測。當然促銷時, 只需針對可能購買的會員進行。如此可以集中「火力」, 節省大量成本。

1 背景介紹

1.1 Fisher's Approach

進行群組區別時, 通常是從自變數 $X_1, X_2, \cdots, X_N$ 的資料去判斷其所屬的群組。之前, 必須先進行群組分析, 以確立區別的準則。換句話說, 從已知的自變數資料及所屬的群組找出之間的關係。如果可以清楚的描述這項關係, 就等於掌握了群組的特性。群組與自變數的關係一般稱爲「區別函數」Discriminant Function。自變數間不同的組合可以構成不同的「區別函數」, 但區別的效果 (或稱區別率或鑑別率) 不同。什麼樣的組合才能達到最佳的鑑別程度呢?

首先必須瞭解自變數間線性組合的幾何意義; 假設 N 個自變數, 線性組合成一個變數, 等於從 N 度空間投射到 1 度空間上, 或說從 N 個座標軸簡化爲一個座標軸。理論上, 問題變簡單了, 但也同時因空間的縮減, 損失了若干訊息。即便如此, 在比較小的空間裡, 一樣可以找到最佳的「觀察角度」, 諸如此類的區別函數於焉產生。

舉一個簡單的例子, 有兩個群組, 彼此交錯, 其分佈及重疊情形如圖 17.1 所示。假設鑑別函數

$$f(X_1, X_2) = X_1 \tag{1}$$

即群組鑑別的依據完全取決於變數 X1 的值。其鑑別能力可以從圖右下角的兩個交錯的常態分配圖清楚看出。如果鑑別函數定義爲

$$f(X_1, X_2) = X_2 \tag{2}$$

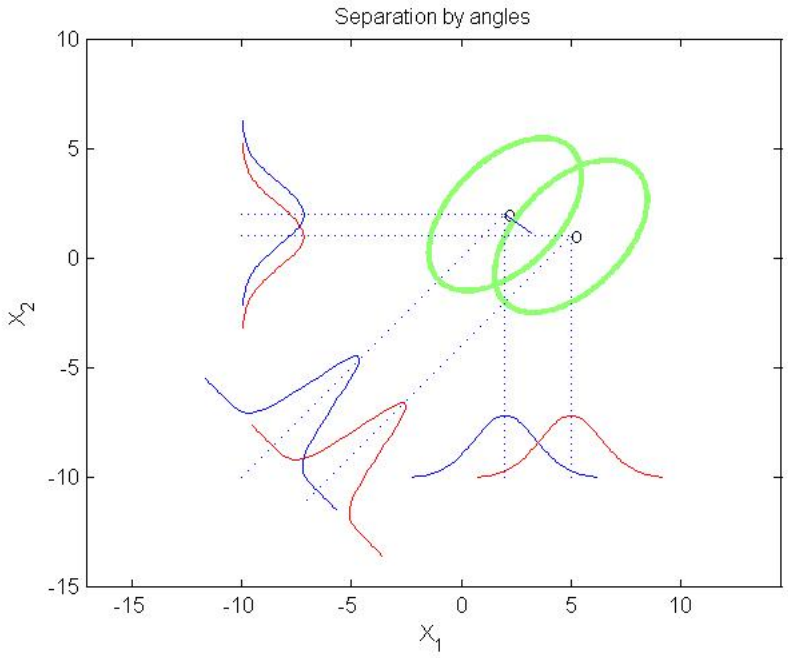

圖 1: 群組區別性與觀看角度的關係

即群組鑑別的依據完全取決於變數 X_2 的值, 則其鑑別能力可以從圖左上角的兩個交

錯更緊密的常態分配圖看出, 很明顯的, 其鑑別能力更差了。這也說明自變數間的不同組合可以決定群組的鑑別程度。如圖中交錯比較小的兩個常態分配圖, 不僅平均數的距離比較遠, 連變異數也比較小, 自然容易分辨。就好像遠看兩棟相鄰的建築物, 從不同的角度可以看到不同的形狀, 其間的差距也隨之不同。這也說明看待一件事情, 當從不同的角度觀察時, 常會呈現出不同的面貌。

Fisher[1]提出自變數的組合方式(幾何: 觀察的角度), 希望造成 (看到) 群組間的「距離」最大, 群組內的「分散」程度最小。若不能兩全, 則取其比例最大者。以專業的術語來描述「距離」與「分散」如;

> Maximize the ratio of the across-group sum of squares to the within-group sum of squares for the combination

或

$$\max_{\text{組合係數}} \frac{\text{across-group sum of squares}}{\text{within-group sum of squares}} \tag{3}$$

從一維的空間座標來看多維度的群組, 不同的座標角度會看到不同的群組分佈情況, 圖 17.1 便提供了三個角度。從那個角度可以看到群組間距最大? 同時組內的變異最小?Fisher 用了 across-group sum of squares 來量化群組間距, 以 within-group sum of squares 量化組內的變異。數學上的定義如下:

假設有兩個群組, 變數 t 爲 N 個自變數組合後的變數 (或稱爲鑑別函數), 即

$$t = k_1X_1 + k_2X_2 + \cdots + k_NX_N = \mathbf{x}^T\mathbf{k} \tag{4}$$

在幾何意義上, 稱爲投射 (mapping), 將處在 N 度空間上的點投射到 1 度空間上 (變數爲 t), 對 N 度空間上的群組的分辨能力, 只剩下一個角度, 就是與新的座標軸 t 垂直的方向, 這個幾何上的意義可以從圖 17.2 與 17.3 看出來。

從與向量 $\mathbf{k}$ 垂直的方向來看群組內的聚散情況, 稱爲 within-group sum of squares, 定義爲

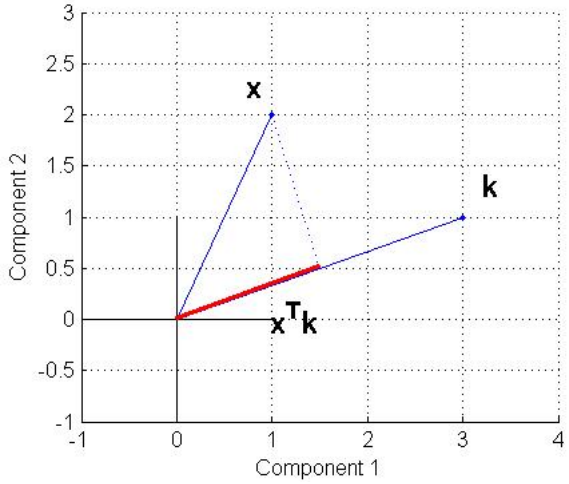

圖 2: 向量間的垂直投射:$\mathbf{x}^T\mathbf{k}$的幾何意義($||\mathbf{k}|| = 1$)。

$$SS_W = \sum_i (t_{i(1)} - \bar{t}_{(1)})^2 + \sum_i (t_{i(2)} - \bar{t}_{(2)})^2 \tag{5}$$

其中 $\bar{t}_{(1)}$ 及 $\bar{t}_{(2)}$ 分別代表組合變數中屬於群組1及群組2的平均值。$t_{i(1)}$ 及 $t_{i(2)}$ 則分別代表其第 i 個樣本值。假設群組1共有 n_1 個樣本, 群組2共有 n_2 個樣本。上式可以寫成

$$\begin{aligned} SS_W &= \sum_i (\mathbf{k}^T\mathbf{x}_{i(1)} - \mathbf{k}^T\overline{\mathbf{x}}_{(1)})^2 + \sum_i (\mathbf{k}^T\mathbf{x}_{i(2)} - \mathbf{k}^T\overline{\mathbf{x}}_{(2)})^2 \\ &= \mathbf{k}^T \left(\sum_i (\mathbf{x}_{i(1)} - \overline{\mathbf{x}}_{(1)})(\mathbf{x}_{i(1)} - \overline{\mathbf{x}}_{(1)})^T + \sum_i (\mathbf{x}_{i(2)} - \overline{\mathbf{x}}_{(2)})(\mathbf{x}_{i(2)} - \overline{\mathbf{x}}_{(2)})^T \right) \mathbf{k} \\ &= \mathbf{k}^T (W_1 + W_2)\mathbf{k} \\ &= \mathbf{k}^T \left((n_1 - 1)C_1 + (n_2 - 1)C_2 \right) \mathbf{k} \end{aligned} \tag{6}$$

其中 $\overline{\mathbf{x}}_{(1)}, \overline{\mathbf{x}}_{(2)}$ 分別代表群組1與群組2的樣本平均值, C_1 與 C_2 代表群組1與群組2的共變異矩陣 (又稱爲 within-group covariance matrix), 反映了群組內 (within-group) 樣本的分佈情況。在此使用了 unbiased 的估計式。

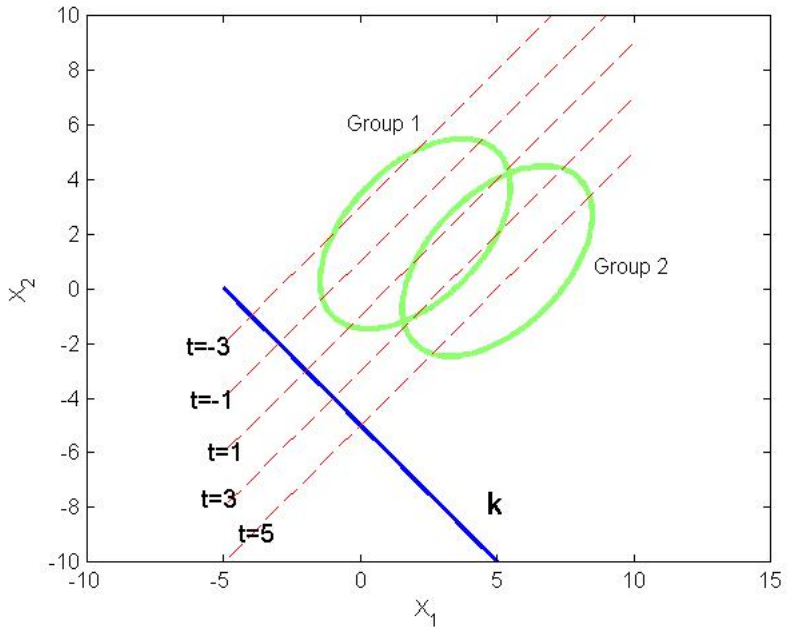

圖 3: 2度空間到1度空間的投射

另一方面, 式 (17.4) 代表群組間聚散的情況的 across-group sum of squares, 定義爲

$$SS_A = n_1(\bar{t}_{(1)} - \bar{t})^2 + n_2(\bar{t}_{(2)} - \bar{t})^2 \tag{7}$$

其中 $\bar{t}$ 代表組合變數的整體樣本平均值, 樣本大小 n_1 與 n_2 反應其比例上的權重。上式可以繼續推演爲 (作業1)

$$\begin{aligned} SS_A &= n_1\mathbf{k}^T(\overline{\mathbf{x}}_{(1)} - \overline{\mathbf{x}})(\overline{\mathbf{x}}_{(1)} - \overline{\mathbf{x}})^T\mathbf{k} + n_2\mathbf{k}^T(\overline{\mathbf{x}}_{(2)} - \overline{\mathbf{x}})(\overline{\mathbf{x}}_{(2)} - \overline{\mathbf{x}})^T\mathbf{k} \\ &= \mathbf{k}^T\left(n_1(\frac{n_1}{n_1+n_2})^2\mathbf{d}\mathbf{d}^T + n_2(\frac{n_2}{n_1+n_2})^2\mathbf{d}\mathbf{d}^T\right)\mathbf{k} \\ &= \lambda\mathbf{k}^T\mathbf{d}\mathbf{d}^T\mathbf{k} \end{aligned} \tag{8}$$

其中 $\overline{\mathbf{x}}$代表整體樣本的平均數, 而 $\mathbf{d} = \overline{\mathbf{x}}_{(1)} - \overline{\mathbf{x}}_{(2)}$ 。Fisher 對於自變數的最佳組合來自下列的最佳化問題:

$$\max_{\mathbf{k}} \frac{\mathbf{k}^T \mathbf{d}\mathbf{d}^T \mathbf{k}}{\mathbf{k}^T C_W \mathbf{k}} \tag{9}$$

目標函數爲 SS_A 與 SS_W 的比例, 其中 C_W 一般稱爲 Pooled within-group covariance matrix, 其不偏估計 (unbiased estimate) 定義爲

$$C_W = \frac{(n_1 - 1)C_1 + (n_2 - 1)C_2}{n_1 + n_2 - 2} \tag{10}$$

透過目標函數一次導數爲零的必要條件, 上式的最佳解爲 (作業2):

$$\mathbf{k}^o \propto C_W^{-1} \mathbf{d} \tag{11}$$

$\mathbf{k}^o$ 代表在所有可能的一度空間裡, 提供同時兼顧群組間距與群組內聚合性的最佳角度。值得注意的是,Fisher 的觀念只提出最佳的鑑別視野, 並未明確的指出群組的分界線在哪裡, 所以式 (17.11) 只是個方向, 還不能當作群組的鑑別條件。Mahalanobis 直接從群組分界線切入來看這個問題, 其結果不但與 Fisher 的鑑別觀念不謀而合, 並且也得到一組分界線的方程式。

1.2 Mahalanobis's Method

Mahalanobis對於自變數的組合方式有不一樣的看法; 找出兩群組間等距離的軌跡 (locus) 函數, 認爲這是對群組做最適當的切割。所謂「群組的距離」必須先定義。假設 $\mathbf{x}$ 爲任意一點, 其與第 k 個群組的距離定義爲:

$$D_k^2 = (\mathbf{x} - \overline{\mathbf{x}}_k)^T C_W^{-1} (\mathbf{x} - \overline{\mathbf{x}}_k) \tag{12}$$

其中 $\overline{\mathbf{x}}_k$ 代表第 k 個群組的中心點, 而 C_W^{-1} 的介入考慮了群組內的變異情形(在此假設所有群組的共變異矩陣相同, 都等於群組間的Pooled within-group covariance matrix C_W, 即 $C_1 = C_2 = \cdots = C_k = C_W$)。這也可以解釋爲「加權的歐幾理德距離 (Weighted Euclidian Distance)」。在只有兩個群組的條件下, 等距離的軌跡函數必須滿足以下的條件:

$$D_1^2 = D_2^2 \tag{13}$$

經過推導後, 上式變爲 (作業3)

$$2\mathbf{x}^T C_W^{-1}\mathbf{d} = \bar{\mathbf{x}}_{(1)}^T C_W^{-1}\mathbf{d} + \bar{\mathbf{x}}_{(2)}^T C_W^{-1}\mathbf{d} \tag{14}$$

其中 $C_W^{-1}\mathbf{d}$ 與Fisher 提出的組合係數 $\mathbf{k}$ 成正比(參考式(17.11)), 直接將 $\mathbf{k} = C_W^{-1}\mathbf{d}$代入, 上式變爲

$$\mathbf{x}^T\mathbf{k} = \frac{\bar{t}_{(1)} + \bar{t}_{(2)}}{2} \tag{15}$$

這就是分界線方程式。等式右邊提供了鑑別函數 $\mathbf{x}^T\mathbf{k}$ 的判斷準則。

2 練習

範例1: 式(17.6) 的 C_1 與 C_2 分別代表群組1與群組2的共變異矩陣, 計算式 (17.10) 的 Pooled within-group covariance matrix。從網站上下載資料 Book_1.txt[1,chap 12], 根據資料的描述計算 C_W。

通常資料處理前需經過前置作業 (Pre-processing), 將屬於同組的資料集合在一起, 方便後續的處理, 一方面減輕重複處裡的負擔, 一方面也讓程式簡潔些。譬如這個資料檔的第四欄是購買與 (1) 否 (0), 也就是群組的屬性。這時可以利用 find 指令, 如

```
data = load('BOOKS_1.txt');
b1 = find(data(:, 4) == 1);          % 購買者的索引
x1 = data(b1, 2 : 3);                %Group 1: 購買者的 input 資料
```

以上指令找出群組1的輸入 (自變數) 資料, 相同的方式可以找出另一個群組的資料。利用這些資料便可以順利的計算 C_W, 大約是

$$C_W = \begin{bmatrix} 63.2366 & 0.1644 \\ 0.1644 & 0.4308 \end{bmatrix}$$

範例2: 試著畫出圖1。這牽涉到下列的技巧:

- 畫出 Bivariate 常態的 pdf 空照圖 (contour plot), 當兩變數間具相關性, 看起來像是個傾斜的橢圓。
- 圖形轉向及位移 (可以利用 MATLAB 的指令 pol2cart 及 cart2pol 做角度的轉移)。
- 估計常態分配從不同角度觀察 (變數轉換) 的變異數。

圖形的轉向即座標位置的轉換, 可以利用轉置矩陣 T 來幫忙。譬如樣將向量 $a = [2 \;\; 1]^T$ 逆時針轉 $\theta = 30^o$, 可以這麼做

$$b = Ta, \qquad\qquad T = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

如圖17.4所示。向量圖形是利用 MATLAB(7.x 以上) 指令 biplot 繪製的。當做圖形轉換時, 向量 a 換成圖形的 $[x \; y]^T$ 矩陣即可, 如右圖將橢圓旋轉90度。程式碼如下

```
t = 0 : pi/20 : 2 * pi;
ellips_x = 1 * sin(t);
ellips_y = 0.5 * cos(t);
plot(ellips_x, ellips_y), hold on
axis([-22 -22])
theta = pi/2;                    %旋轉90度
T = [cos(theta)  -sin(theta); sin(theta)  cos(theta)];
R = T * [ellips_x; ellips_y];                    %針對所有的座標點做轉置
plot(R(1,:), R(2,:),'r'), hold off
```

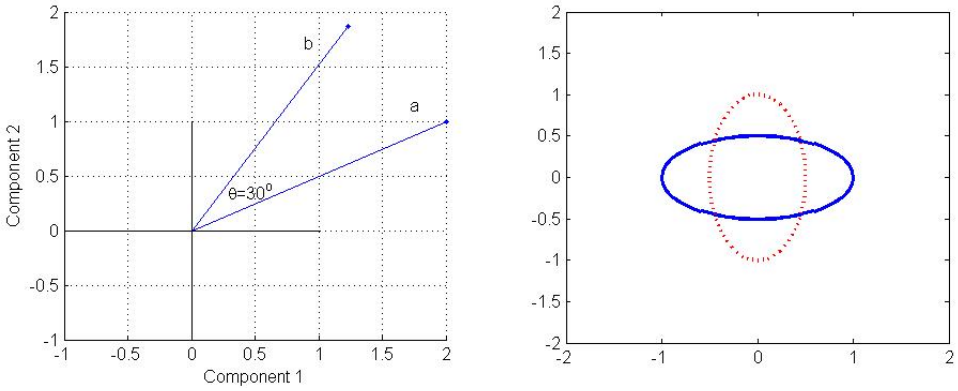

圖 4: 座標(圖形) 轉換

除使用轉置矩陣外,MATLAB 提供了兩個指令 $cart2pol, pol2cart$ 也可以達到圖形 (座標點) 轉置的目的。程式碼如下, 其中第一行指令 $cart2pol$ 將所有座標點的表示法從 Cartesian coordinate 改爲 polar coordinate, 第二行將原來每個座標點的角度加上欲旋轉的角度, 再利用 $pol2cart$ 轉回 Cartesian coordinate 的 X-Y 座標。

$[theta, rho] = cart2pol(ellips\_x, ellips\_y);$
$[x, y] = pol2cart(theta + pi/2, rho);$ %旋轉90度
$plot(x, y, 'r')$

3 觀察

1. 式 (17.5) 右邊第一項, 正比於轉換變數後屬第一群組的變異數。這個變異數愈小愈有利於分辨。這也是爲什麼圖 17.1 中間的常態圖變異數比較小的原因

2. Fisher 提出最佳的組合係數 k, 將自變數組合成一個單一變數, 如式 (17.4)。當代入樣本值時, 式 (17.4) 又稱爲 discriminant (function) score。本單元並未提及如何應用這個 score 來做群組區隔的判別。似乎需要定義 (或找出) 一個 score 來做爲群組判斷 (預測) 的關鍵值 (cut-off value)! Mahalanobis 提出的觀點補足了這個關鍵值。

4 作業

1. 推導式 (17.6) 與 (17.8)。
2. 推導式 (17.11)。
3. 推導式 (17.15)。
4. 以 Book_1.txt 的資料爲例, 利用練習一所估計的 C_W, 及式 (17.11), (17.15) 分別計算 Fisher 所提出的最佳組合係數 k 及 Mahalanobis 所定義的等距方程式, 並畫出如圖17.5所示的圖形。請注意 Book_1.txt 的資料都是整數, 繪圖前每筆資料都加上些許的變動值, 使得資料的分佈看起來更接近眞實。圖 17.5 中的虛線代表 Mahalanobis 所定義的等距方程式, 垂直於 Fisher 提出的 k vector。

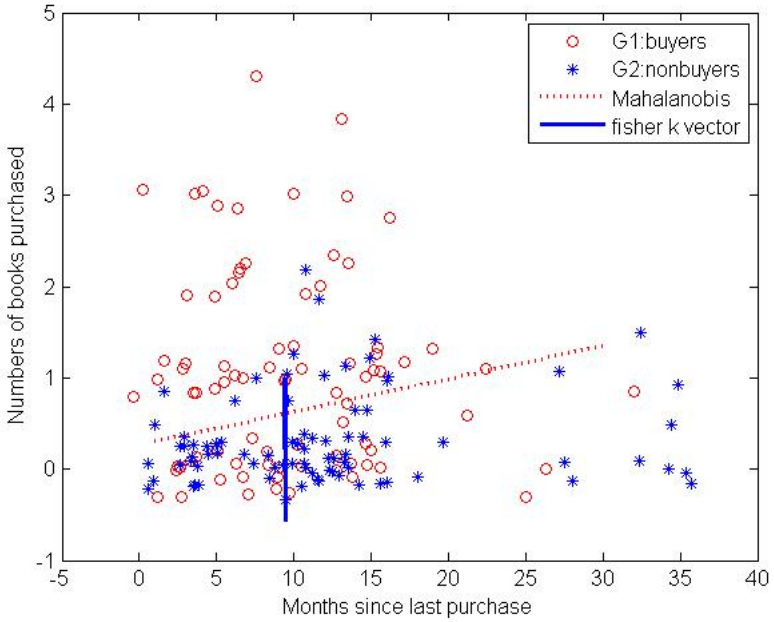

圖 5: Fisher 提出的最佳觀察角度與及 Mahalanobis 等距分界線

參考文獻

[1] J. Latin, D. Carroll, P. E. Green, "Analyzing Multivariate Data,"2003, Duxbruy.

[2] A. C. Rencher, "Multivariate Statistical Inference and Applications,"1998, John Wily and Sons.

[3] 黃俊英,"多變量分析 (第七版),"中國經濟企業研究所。