群組分類
Linear Discriminant Analysis

last modified July 22, 2008

前一個單元運用了機率的概念提供解決問題的另一個面相。理論上應該是比應用最小平方法的迴歸模式好, 因爲比較多的訊息被假設爲已知, 譬如後驗機率 (Posterior Probability) $Pr(G|X)$, Nearest Neighbor Method 利用平均數作爲後驗機率的估計值, 成功的將已知資料做一個很精準的群組分界。本單元介紹 Linear Discriminant Analysis, 從另一個角度來運用後驗機率密度函數, 對機率應用在實際問題的概念有很好的啓發。

本章將學到關於程式設計

較複查線性方程式的直線繪圖、MATLAB 矩陣式的計算技巧 (免迴圈)。

⟨本章關於 MATLAB 的指令與語法⟩

指令:mvnpdf, mvnrnd, hist3

1 背景介紹:Linear Discriminant Analysis

後驗機率是運用機率概念解決問題很直覺的出發點, 但卻是個很不友善的東西, 一般而言, 並沒有足夠的訊息寫出完整的函數。不過山不轉路轉, 有一個可愛又有一點討厭的貝氏定理撐腰,

$$P(G = k|X) = \frac{P(X|G = k)P(G = k)}{\sum_l P(X|G = l)P(G = l)} \tag{1}$$

等號右邊的兩個機率密度函數的假設反而比較可行, 其中 $P(X|G = k)$ 表示第 k 組資料發生的機率密度函數, 而 $P(G = k)$ 則提供每一個群組所佔的比例 (發生的機率)。[1] 這兩個機率函數比較合理的原因在於它們比較容易從已知的資料中估計出來。當然估計過程的假設與計算品質的好壞也間接影響了這個方案的準確度。

這個單元我們假設 $P(X = \mathbf{x}|G = k) = f_k(x)$ 服從 Multivariate Normal Distribution,

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{k})^T\Sigma_k(\mathbf{x}-\mu_\mathbf{k})} \tag{2}$$

其中 $\mathbf{x} \in R^p$, $\mu_\mathbf{k}$ 與 Σ_k 分別代表第 k 群資料常態假設的均值與共變異矩陣。這是對於母體的假設, 至於眞實的資料是否具備這個分配的特性是需要進一步檢驗的。爲簡化問題的複雜性, Linear Discriminant Analysis(LDA) 更進一步假設所有的共變異矩陣 (Covariance Matrix) 都相等, 即 $\Sigma_k = \Sigma, \forall k$。當然這個假設的合理性也是需要從實際的資料中做進一步的檢驗。

當給定已知資料($X = \mathbf{x}$), 要在空間中找出兩個群組間的分界線,[2]可以利用下面這個條件:

$$Pr(G = k|X = \mathbf{x}) = Pr(G = l|X = \mathbf{x}) \tag{3}$$

[1]一般稱 $P(X|G = k)$ 爲概似函數, 稱 $P(G = k)$ 爲先驗機率。

[2]這裡提到的「分界線」並非指平面空間上的一條線, 而是更廣泛的「Separating Hyperplanes」, 可能是線、面或更高維度的集合, 一般通稱爲 hyperplane 的幾何平面。

也就是, 在資料存在的空間裡, 群組 k 與群組 l 出現機率相同的地方。或者說, 分割群組 k 與群組 l 的線上的點必須滿足以上的條件。在後驗機率相等 (3) 的群組分界原則下, 結合由貝式定理 (1) 與資料的常態分配假設 (2), 分界線的函數可以寫成:

$$\begin{aligned}\log\frac{Pr(G=k|X=\mathbf{x})}{Pr(G=l|X=\mathbf{x})} &= \log\frac{f_k(\mathbf{x})}{f_l(\mathbf{x})}+\log\frac{Pr(G=k)}{Pr(G=l)} \\ &= \log\frac{Pr(G=k)}{Pr(G=l)}-\frac{1}{2}(\mu_k+\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l)+ \\ &\quad \mathbf{x}^T\Sigma^{-1}(\mu_k-\mu_l)=0 \qquad (4)\end{aligned}$$

這裡巧妙的運用 logit transformation 的技巧, 令 log-odds 爲零去除指數, 得到一組線性的方程式 (注意: 線性的關係來自「不同群組有相同共變異矩陣」的假設)。

群組判別: 上式從已知資料的後驗機率相等決定群組的分野, 當然也可供判斷一組新資料 $\mathbf{x}$ 的群組屬性:

$$\begin{aligned}G(\mathbf{x}) &= \arg\max_k \log Pr(G=k|X=\mathbf{x}) \\ &= \arg\max_k \log(Pr(X=\mathbf{x}|G=k)Pr(G=k)) \\ &= \arg\max_k \mathbf{x}^T\Sigma^{-1}\mu_k-\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k+\log Pr(G=k) \qquad (5)\end{aligned}$$

第一行的意思很直覺; 當 $X=\mathbf{x}$ 時, 哪一個群組出現的機率最高? 利用(1) 的關係, 去除與 k 無關的分母, 變成了第二行。再將 (2) 代入, 同樣去除與 k 無關的項目, 變成了第三行, 即所謂的 Linear Discriminant Function。

在式 (5) 中, 除了資料 $\mathbf{x}$ 已知外, 其餘都未知, 即便如此, 我們仍可以利用已知的資料來估計這些值: 譬如 μ_k 用第 k 組的資料的樣本平均值, Σ 可以用各組資料算出來的 Sample Covariance Matrices 的加權平均, $Pr(G=k)$則是已知資料中各組數量的比例, 即

- $Pr(G=k)\approx\hat{\pi}_k=N_k/N$, 其中 N_k 代表第 k 組的數量,N 代表所有資料的總數。

- $\hat{\mu}_k = \sum_{group=k} \mathbf{x}_i / N_k$
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{group=k} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T / (N - K)$, K 代表群組數。這個估計又稱爲 Pooled within-group covariance matrix。

2 練習

範例1: 取與前面單元相同的資料la_1.txt, 看看 LDA 的效果如何, 即畫出 LDA 的分界線。

首先估計出分界線函數 (4) 所需的 $\mu_1, \mu_2, \Sigma, Pr(G = 1), Pr(G = 2)$:

- 估計之前, 先檢視該資料的結構。爲方便上述估計量的計算, 有必要將資料依組別重新排列, 或將個別組別挑出來。這可以使用 find 指令來達到「挑選」的目的。

```
Y=load('la_1.txt')
I1=find(Y(:,3)==0);     %第3欄資料等於 0 的索引值
I2=find(Y(:,3)==1);     %第3欄資料等於 1 的索引值
data=[Y(I1,:);Y(I2,:)]; %根據索引值取得資料
```

- 當兩個群組的數量相當時, Σ 的估計可以採 $\hat{\Sigma} = (\hat{\Sigma}_1 + \hat{\Sigma}_2)/2$。
- 在只有兩個變數的情況下, 將式 (4) 中的 $\mathbf{x}^T$ 以 $[x_1 \ x_2]$代入, 仔細的推導出 $x_2 = f(x_1) = mx_1 + c$ 的形式, 即可以畫出分界線, 如圖 (2) 所示。這裡的斜率 m 與截距 c 要非常小心謹愼的推導, 正確與否可以從程式畫出來的圖察覺。

範例2: 方程式(5) 是所謂的 Linear Discriminant Function, 用來判斷新資料的組別。其判別方式必須針對每一個組別分別計算出一個數值, 產生最大數值的組別就是判斷的依據。從 Matlab 程式設計的角度, 最好是運用矩陣的特性, 同時計算出所有的數值並放在一個向量裡, 不需要採迴圈的方式一個個計算。不過練習之初, 倒可以先

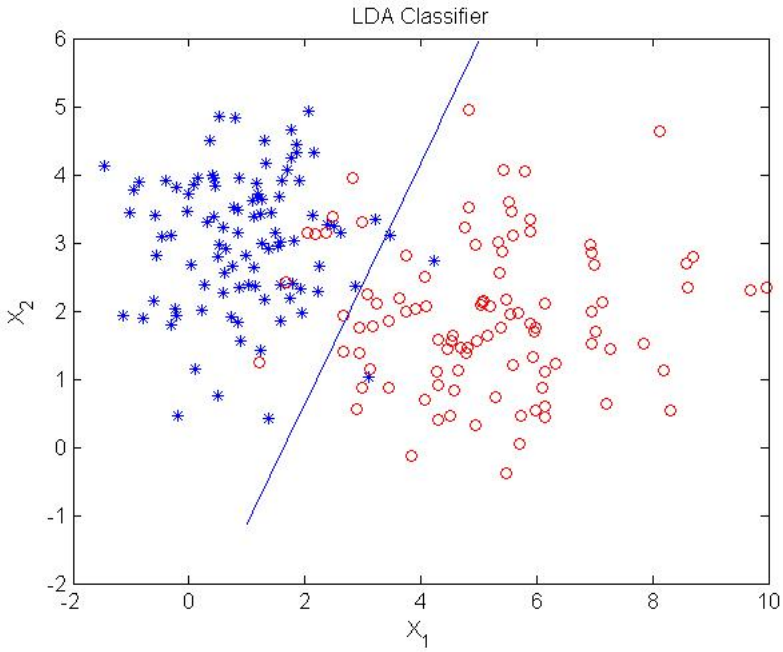

圖 1: LDA 群組分界線

以迴圈寫寫看, 再從迴圈的結構找出矩陣可以運用的地方。

假設有 k 個群組, 要同時爲每個群組計算式 (5) 的函數值的方式, 可以將 $p \times 1$ 的向量 μ_k 換成 $p \times k$ 的矩陣 $U = [\mu_1\ \mu_2\ \cdots\ \mu_k]$, 最後一項換成 $\pi = [\log Pr(G = 1)\ \log Pr(G = 2)\ \cdots\ \log Pr(G = k)]$, 直接套入式 (5) 計算,

$$\mathbf{g} = \mathbf{x}^T \Sigma^{-1} U - \frac{1}{2} U^T \Sigma^{-1} U + \pi$$

再利用 max 指令, 對向量 $\mathbf{g}$ 找出最大值出現的項次, 最爲群組的判斷。

```
[m,i]=max(g)   %向量 g 的第 i 項有最大值 m
```

當然也可以直接計算式 (5) 第二行的多變量常態機率 $Pr(X = \mathbf{x}|G = k)$, 對應的 MATLAB 指令爲 mvnpdf,

```
g=[mvnpdf(x,μ1,Σ1)   mvnpdf(x,μ2,Σ2)   ···   mvnpdf(x,μk,Σk)]
```

範例3: 前面的範例中使用的資料, 兩個群組具相同的數量。當群組大小不同時, 結果會有什麼差別呢? 不妨自己產生模擬資料, 自己決定兩個群組的位置與規模, 再利用這些人工資料看看群組間距與規模大小對 LDA 的表現的影響?

首先設定兩個群組的中心點與共同的共變異矩陣, 譬如:

$$\mu_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mu_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

兩群組大小分別爲 $N_1 = 500, N_2 = 250$。以下的指令模擬總共 750 筆資料:

```
mu1 = [1   -1]; mu2 = [-1   1];
Sigma = [1   0.4; 0.4   1];
r1 = mvnrnd(mu1, Sigma, 500);
r2 = mvnrnd(mu2, Sigma, 250);
Y=[r1 ; r2];
```

上述產生了一個 750×2 的模擬資料, 爲確定自己產生的資料是否合理或合乎原先的規劃, 可以依下列指令繪圖, 結果如圖所示:

```
subplot(2,1,1),plot(Y(:,1),Y(:,2),'*')
subplot(2,1,2),hist3(Y))
```

3 觀察

1. 在兩個群組的情況下, 如果訓練資料中兩群組的數量不同, 會產生什麼狀況? 譬如:$Pr(G = 1) = 2Pr(G = 2)$。

2. LDA 做出來的分界線與第一單元的迴歸模式 (最小平方法) 都是線性的, 在只有兩個群組的情況下, 有何相似之處嗎? 試著針對同一組資料把這兩條線同時畫出來, 比較看看。要仔細看喔!

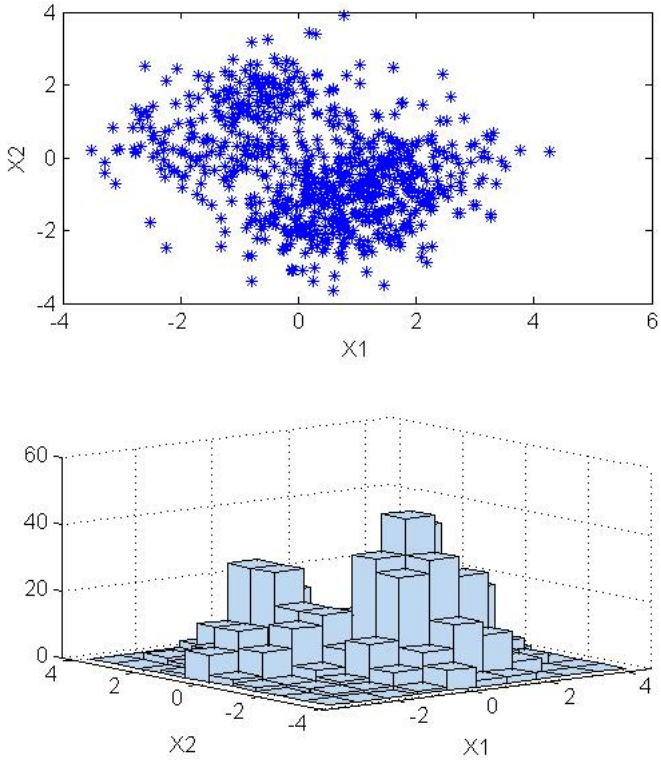

圖 2: 模擬資料的「樣子」

3. Logistic regression 是處理類別輸出資料一項很基本但也很重要的工具。運用在群組分析上可以避免如 LDA 對概似函數 (likelihood function)(2) 爲常態的假設。這算是對常態分配的解套, 不過另一方面, 它也做了對 posterior probability $P(G|X)$ 的假設

$$p = Pr(G = 1|X, \underline{\beta}) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{6}$$

這個假設對許多的資料來源而言, 具有相當程度的合理性。令其 log odds ratio 等於0 時, 求得最佳的分界線:

$$c = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x = 0 \tag{7}$$

參數 β_0, β_1 由已知的資料估計而得。前面的單元已經介紹過它的理論基礎與計算方法, 原來的程式只要稍作些微的修改即可應用在本單元的資料。

4 作業

1. 將你的程式運用在 mix.mat 這組資料, 畫出分界線, 並允許輸入新資料, 且立即輸出該資料經判別後的組別。

2. 請比較 logistic regression 與 LDA 在假設上、方法上及最後做出的分界線有何不同。Hint: 式 (4) 與 (7)。

3. 試著自己產生三組別的資料各100筆, 之間的距離自己拿捏。利用 LDA 的方法在不同的兩個群組間畫一條線, 共可畫出三條分界線以區隔三個群組。這三條線是否交於同一點?

參考文獻

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