群組分類
Goodness of Fit and Prediction

last modified May 3, 2006

前一個單元探討 Fisher 與 Mahalanobis 對於群組分析的觀念, 並推導出「鑑別函數」(Discriminant Function)。本單元則是討論該函數的配適性 (Goodness of Fit) 及如何應用該函數作為群組的預測 (Prediction)。

1 背景介紹

1.1 配適性 (Goodness of Fit)

Fisher提出兩群組間的鑑別函數

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{k} \tag{1}$$

其中最佳組合係數 $\mathbf{k}$ 與 $C_W^{-1}\mathbf{d}$ 成正比。這個組合根據群組的樣本資料, 提供了分辨兩個群組的最佳「視野」。但當兩群組交錯而存在模糊的界線時, 這個最佳的「角度」到底有多好? 配適已知的樣本的準確率有多高? 或說「誤判率」多低? 問題是,(1) 式如何配置樣本資料的群組呢? 有一個方式是提出函數的臨界值 (cutoff score) t_c, 也就是當個別資料代入函數 (1), 若其值大於 t_c, 便認定為某一群組, 否則為另一群組。

Mahalanobis提出兩群組的等距點, 是臨界值 (cutoff score) t_c 不錯的選擇, 即

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

其中 $\bar{t}_{(1)} = \overline{\mathbf{x}}_{(1)}^T \mathbf{k}$, $\bar{t}_{(2)} = \overline{\mathbf{x}}_{(2)}^T \mathbf{k}$爲兩群組中心點的鑑別函數值。不過當群組的大小不等時, 式(2) 需要做些修正以矯正因樣本數不均所造成的誤差; 譬如:

$$t_c = \frac{n_1 \bar{t}_{(1)} + n_2 \bar{t}_{(2)}}{n_1 + n_2} \tag{3}$$

其中 n_1, n_2分別代表群組1與群組2的樣本數。

1.2 預測 (Prediction)

當然鑑別函數對資料的配適性高低並不能保證期「預測」能力, 也就是對未知資料的鑑別能力。特別當我們考慮其他因素, 如判別錯誤的代價 (cost of misclassification)。誤判的代價或許因群組而異, 譬如將群組一誤判爲群組二, 其損失是將群組二誤判爲群組一的10倍。將誤判列入考慮是比較符合實際情況的作法。或者說, 將更多有力判斷的因素加入, 以提高預測的能力或降低預測錯誤的損失。

鑑別函數的好壞如何評斷呢? 有沒有理論上最佳的鑑別函數? 看一看這個條件式機率密度函數

$$P(Group\ k|\mathbf{x}) \qquad\qquad k = 1, 2, \cdots, C$$

這說明當資料 $\mathbf{x}$ 出現時, 它來自群組 k 的機率。在這個時候, 哪個群組的機率最高, 代表資料 $\mathbf{x}$ 來自該群組的可能性最高。這個函數稱爲最佳的鑑別函數, 也可以拿來作爲評斷其他鑑別函數的依據。但問題是, 這個條件式機率密度函數, 一般也稱爲posterior probability, 通常是不可知的。還好有個貝式定理可以緩和這個限制

$$P(Group\ k|\mathbf{x}) = \frac{P(\mathbf{x}|Group\ k)P(Group\ k)}{P(\mathbf{x})}$$

其中 $P(\mathbf{x}|Group\ k)$ 一般稱爲群組條件式機率密度函數 (Class conditional density function), $P(Group\ k)$ 稱爲群組的 prior probability。這兩個密度函數相對比較容易「取得 (透過估計或假設),」因此最佳的鑑別函數可以改寫爲

$$P(\mathbf{x}|Group\ k)P(Group\ k)$$

對於只有兩個群組而言, 最佳 (貝氏) 的鑑別函數可以寫成 (作業1)

$$f(\mathbf{x}) = \frac{P(\mathbf{x}|Group\ 1)}{P(\mathbf{x}|Group\ 2)} \tag{4}$$

其臨界值爲

$$t_c = \frac{P(Group\ 2)}{P(Group\ 1)}$$

當進一步假設 (a) 所有群組資料遵循常態分配,(b) 每個群組的共變異矩陣相同[1], 允許我們寫出群組1的條件機率密度函數:

$$P(\mathbf{x}|Group\ 1) = \frac{1}{det(C_W)\sqrt{2\pi}} exp(-\frac{1}{2}(\mathbf{x} - \overline{\mathbf{x}}_{(1)})^T C_W^{-1}(\mathbf{x} - \overline{\mathbf{x}}_{(1)})) \tag{5}$$

其中 $\overline{\mathbf{x}}_{(1)}$ 及 C_W 爲群組1的平均數與共變異矩陣的估計。觀察上式指數的部分恰是 Mahalanobis 對於「距離」的定義, 於是可以改寫爲

$$P(\mathbf{x}|Group\ 1) = \frac{1}{det(C_W)\sqrt{2\pi}} exp(-\frac{1}{2}D_1^2) \tag{6}$$

同理, 群組2的條件機率密度函數爲

$$P(\mathbf{x}|Group\ 2) = \frac{1}{det(C_W)\sqrt{2\pi}} exp(-\frac{1}{2}D_2^2) \tag{7}$$

根據式 (6)(7) 並假設群組的 prior probability 分別爲 q_1 及 q_2, 式 (4) 取對數後的群界分界線可以進一步寫成

$$\ln \frac{q_1}{q_2} - \frac{D_1^2 - D_2^2}{2} = 0 \tag{8}$$

其中 (作業2)

$$\frac{D_1^2 - D_2^2}{2} = \mathbf{x}^T\mathbf{k} - \frac{\bar{t}_{(1)} + \bar{t}_{(2)}}{2} = t - t_c \tag{9}$$

[1]即Linear Discriminant Analysis(LDA) 的假設

t 與 t_c 分別是 Fisher 的鑑別函數值及 Mahalanobis 提出的臨界值。式 (8) 做群組判別的預測時, 表示爲: 將觀察資料判斷爲群組1, 當

$$t < t_c + \ln \frac{q_1}{q_2} \tag{10}$$

式 (10) 看得出, 當群組大小一致時, 或說當 $q_1 = q_2$ 時, 這個群組的判斷與 Mahalanobis 提出的等距軌跡相同 (式 (2))。但當群組大小不一時, 式 (10) 的臨界值有別於前述的式 (3)。

當進一步考慮判斷錯誤的代價時, 譬如 $C(1|2)$ 代表將屬於群組2的資料誤判爲群組1 的代價, $C(2|1)$ 剛好相反。判斷式 (10) 可以修正爲

$$t < t_c + \ln \frac{q_1 C(2|1)}{q_2 C(1|2)} \tag{11}$$

2 練習

範例1:1. 利用 Fisher 的鑑別函數 (1) 及 Mahalanobis 提出的臨界值 t_c, 計算前一個單元使用的的資料 Book_1.txt, 計算其配適性, 並製作一張所謂的 hits-and-misses table(或稱 confusion matrix)。在這組資料裡, 兩個群組的大小相差很多, 因此可以朝兩方面去做:(a) 兩群組相同大小, 採式 (2) 的臨界值 (b) 兩群組不同大小, 但採式 (3) 的臨界值, 並仔細觀察這兩個結果。

這裡所謂的「配適性」可以簡單說是「命中率」。即將一組已知群組的資料以某種群組分界線 (譬如式 (9)(10)) 做群組判斷, 將判別結果與資料已知的群組做比較, 計算判別正確與錯誤的個數與比例, 製成一張所謂的 hits-and-misses table, 如表所示 (以兩群組爲例)
以 Book_1.txt 的資料爲例, 共有 1000筆, 其中群組1(非購買者) 有917筆, 群組2(購買者) 有83筆。

我們可以依群組大小的相同與否來做配適性計算, 測試式 (3) 的必要性。因爲群組 2 筆數較少, 爲使兩組大小相等, 我們自群組 1 隨意抽取 83 筆資料來做測試。其結果如下表

3 觀察

4 作業

1. 推導兩個群組的最佳鑑別函數。
2. 推導出式 (8) 及 (9)。
3. 利用練習 1 得到的組合係數與臨界值, 對另一組資料 Book_ 2.txt 作「預測」測試。同樣製作一張 hits-and-misses table, 比較看看這兩個結果的差別。
4. 同上, 但應用式 (10) 的結果。
5. 同上, 但應用式 (11) 的結果, 其中 $C(2|1) : C(1|2) = 6 : 1$。

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.