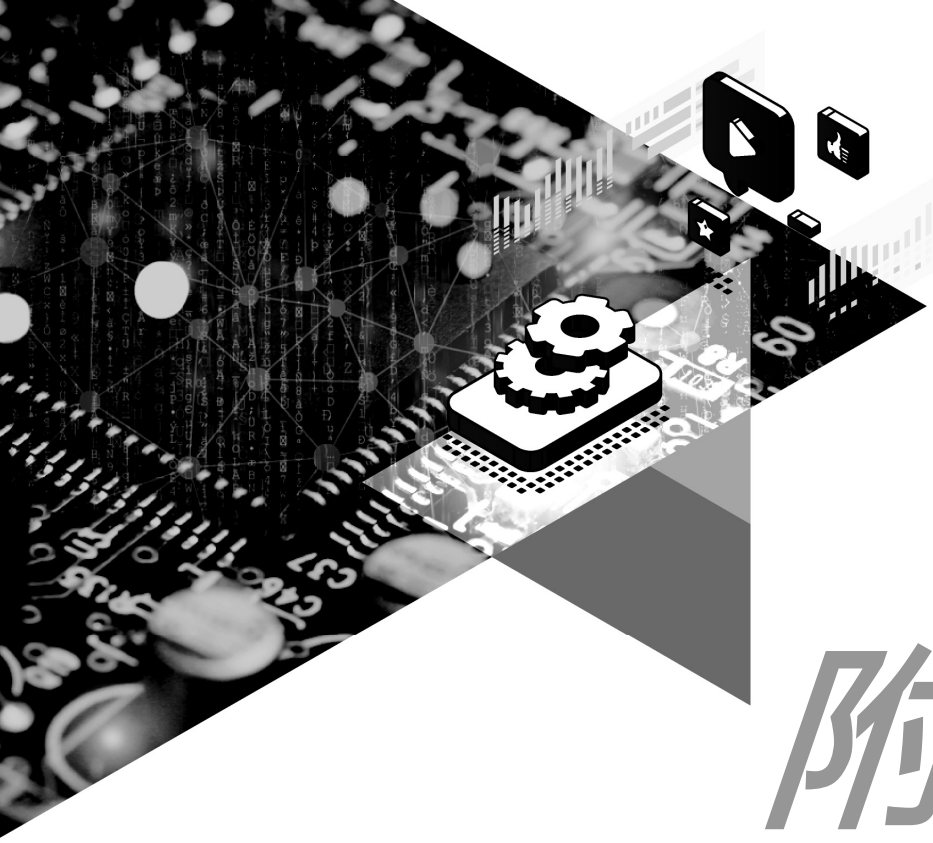


附錄B

線性分類器

Linear Classifiers

本章介紹的是「**線性分類器**」(linear classifier),屬於監督式學習中解決分類問題的基礎方法。「線性迴歸」解決的是迴歸(regression)問題,因此目標值為一個定量(quantitative)的連續型(continuous)變數。「線性分類器」解決的是分類(classification)問題,因此目標值為一個定性(qualitative)的類別型(categorical)變數,例如二元的良品與不良品,或是多類別的產品異常(bin code)類型。在充分掌握了線性迴歸模型的統計思維、模型解釋性以及建模步驟後,本章拓展至線性分類器中。

B.1 分類問題背景與觀念

B.1.1 鋼板製造缺陷

此章節以 UCI Machine Learning Repository 公開數據當中「鋼板缺陷」(steel plates faults dataset)的數據為例(https://archive.ics.uci.edu/ml/datasets/Steel+Plates+Faults),此數據一共包含了 27 個特徵以及 1,941 筆觀測值。作為範例,我們僅以兩種缺陷分別為「凹凸不平」(bumps)與「刮痕」(scratch)作為二元分類的兩目標變數,並僅探討製造的「鋼板面積」(pixel areas)與製程的「輸送帶長度」(length of conveyer)兩個特徵,挑選後的觀測值一共有 792 筆。假想我們在一座製造鋼板的工廠,我們想預測鋼板缺陷的類型以建立自動化程序減少人工判斷,從而快速重工(rework)或修補,或是找出造成缺陷的關鍵因子,探討製程中的這些因子與不同缺陷的因果關聯,以提升製程品質並協助故障排除(troubleshooting)。鋼板數據視覺化後的散佈圖如圖 B.1 所示,可看出兩種缺陷類型的特徵分布相當不同。若各別視覺化成箱型圖如圖 B.2 所示,可觀察到鋼板面積較大的樣本中刮痕缺陷的佔多數,而輸送帶長度較長的樣本中凹凸不平缺陷的佔多數。接著,如何建構線性分類器預測缺陷的類型,並且如何解釋特徵是怎麼影響缺陷類型的判別呢?

在分類的思維中對於分類器的建構,我們通常會有以下的疑問:(1)如何決定分類的界線或我們稱之為「決策邊界」(decision boundary)呢?(2)我們可否預測新數據屬於每一類的機率呢?(3)我們該如何解釋特徵與分類結果之間的關係呢?(4)除了二元分類的問題,我們是否能拓展至多類別的分類呢?在本章中,我們將介紹「**判別模型**」(discriminative model)與「**生成模型**」(generative model)兩大類的方法,以及「**二元與多類別分**

類」（multiclass or multinomial classification）的差異與策略。並依照分類思維介紹兩個最常見的線性分類器，分別是判別模型方法中的「**羅吉斯迴歸**」（logistic regression）與生成模型方法中的「**線性判別分析**」（linear discriminant analysis, LDA）。此外，補充介紹線性判別分析的前身「**樸素貝氏分類器**」（naïve Bayes classifier）以及延伸的「**二次判別分析**」（quadratic discriminant analysis, QDA），兩者皆屬於非線性的分類器。

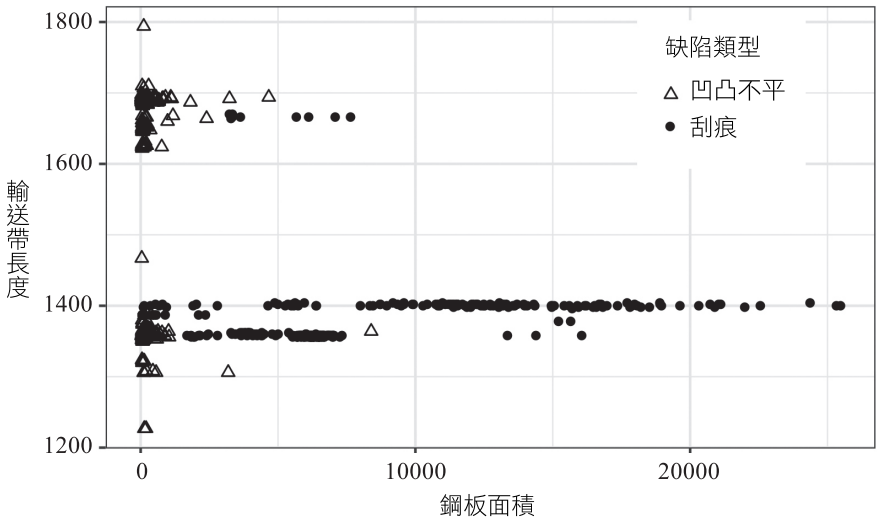


圖 B.1　鋼板數據視覺化

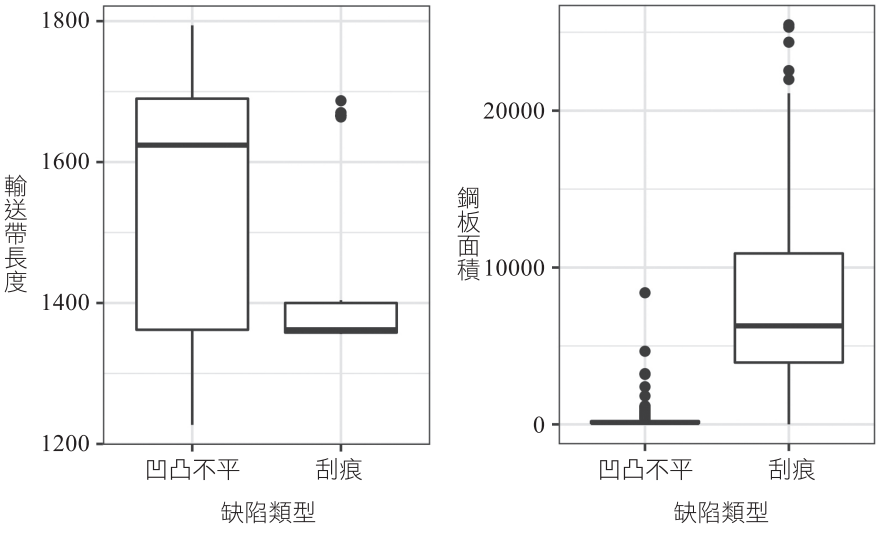


圖 B.2　鋼板數據箱型圖

B.1.2 判別模型與生成模型

在分類的思維中，一般我們會將分類器（模型）分為兩種類型，分別為「判別模型」與「生成模型」如圖 B.3 所示。「判別模型」的思維是「**最小化分類的誤差**」，以建構一個最佳判別的決策邊界，與迴歸模型最小化誤差，並直接建構特徵與目標關聯的概念相似。「生成模型」的思維則是基於「**目標的不同類別分別建構（生成）一個機率分配**」，而新數據可放入各個模型中，找出機率最大的那個，並判別新數據屬於該類別，屬於貝氏機率的概念；此外，生成模型也是可以求出決策邊界的（機率相等的地方即為決策邊界，但通常不等於最小化誤差的邊界）。

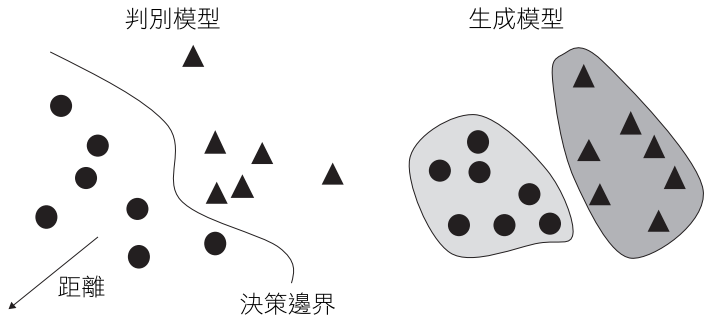


圖 B.3 判別模型與生成模型

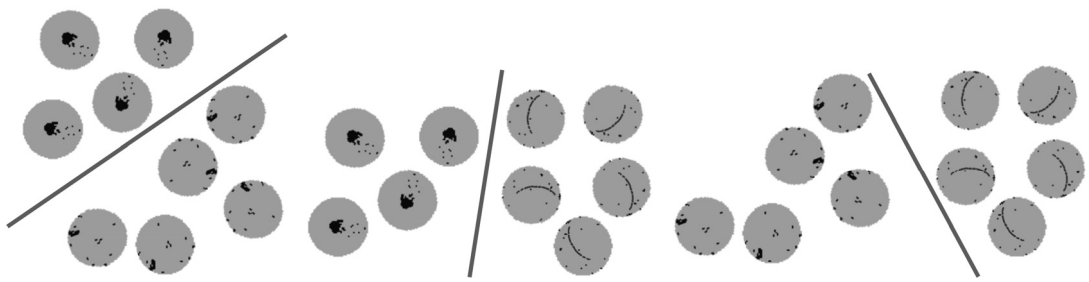
舉上述案例而言，若我們要區分凹凸不平與刮痕兩種缺陷類型，「判別模型」的方法是使用數據直接訓練一個判別模型，最小化分類誤差，並藉由特徵直接預測新的觀測值為凹凸不平或是刮痕的缺陷；而「生成模型」則是基於凹凸不平的數據與刮痕的數據「各自」建構一個機率模型，新的觀測值則分別放入兩模型中，若判斷為凹凸不平的機率大於刮痕的機率，則我們預測它為凹凸不平的缺陷。一般來說，判別模型在分類的績效表現與解釋能力通常比較直觀，這是由於它從本質上的最佳化了分類邊界（決策邊界）；而基於貝氏機率的生成模型則在數據貼近與滿足給定的特徵分配假設時，表現不遜色於判別模型，並且在多類別分類的模型建構更為容易。

B.1.3 二元與多類別分類問題

在分類問題中，依照目標變數的類別水準（level）可將其分為兩類的「二元問題」與兩類以上的「多類別問題」。對於一個「二元分類問題」，

無論是建構判別模型或生成模型以找出決策邊界與兩類別各自的分配都相當直觀。然而，對於一個「多類別分類問題」，生成模型同樣地可建構每個類別各自的分配；但在一般的判別模型中，由於一次只能決定一個決策邊界，因此需要經由多次的模型訓練才能決定出所有類別間彼此的決策邊界，也就是將多類別的問題轉為多個二元分類問題（分多個階段判別）。

多類別的問題轉二元分類的策略有兩種，分別為「一**對**一」（one-vs.-one）與「一**對剩餘**」（one-vs.-all or one-vs.-rest, OvA or OvR）的策略。舉例而言，以 Kaggle 數據庫的 WM-811K 晶圓圖（wafer map）為例，其中包含了中心（center）、邊緣（edge）、刮痕（scratch）三種缺陷。首先，「一對一策略」是將所有類別的兩兩組合各別建構一個判別模型，如圖 B.4 所示，也就是分別為中心與邊緣、中心與刮痕、邊緣與刮痕各自建構一個模型，一共三個模型。因此若推廣至K個類別時，則需建構$C_2^K = \frac{K(K-1)}{2}$個模型，並且對於每個模型而言，僅使用該兩類別的樣本，而在預測時我們會將樣本放入所有模型中，並挑選出在所有模型中機率最大的類別作為預測值。

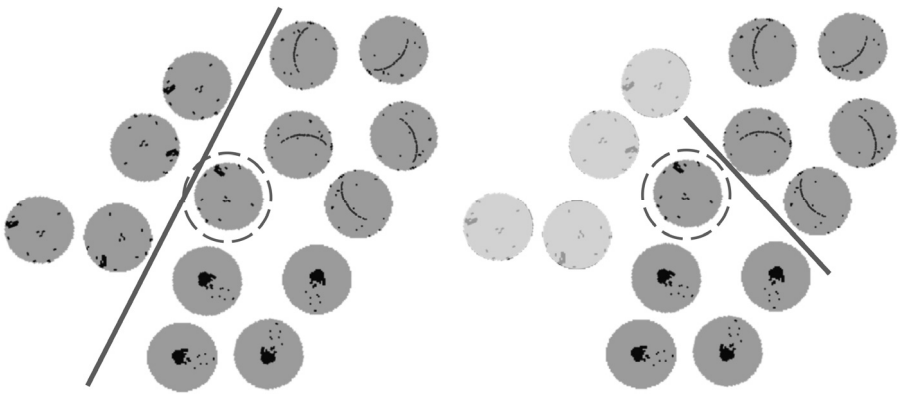


(a)中心與邊緣　(b)中心與刮痕　(c)邊緣與刮痕

圖 B.4　晶圓圖缺陷二元分類的一對一策略

其次，「一對剩餘策略」則是先固定某個特定的類別，並將其他其餘類別視為同一類，並建構第一個模型，接著從剩餘類別挑選某一類後，再將其餘類別再一次視為另一群剩餘類別，循序漸進地建構模型。如圖 B.5 晶圓圖分類所示，我們先為邊緣與非邊緣（中心與刮痕）建構依分類模型，接著再建構刮痕與非刮痕（中心與錯分的邊緣）的分類模型，一共兩個模型。因此若推廣至K個類別時，則需建構$K-1$個模型。然而雖需建構的模型個數相對一對一策略來得較少，但會因我們挑選類別的順序而產生

出不同的模型，也就是「**順序相依性問題**」（sequence-dependent problem）（詳見章節「特徵挑選與維度縮減」）。在預測時我們會將新數據依照建模順序放入模型中，新數據屬於某類別的機率較大時預測該樣本為某類。接續圖 B.5，我們可看到第一個分類邊緣缺陷的模型經分類後，左上角的四張晶圓圖被成功預測正確，而接著分類刮痕的模型則將錯分的邊緣與中心視為一類建模，因此最後錯分的邊緣將被誤判為中心，並且它的存在影響了分類刮痕的模型績效。在後續章節中，我們將提到部分非線性模型的演算法可以直接解多類別分類問題，然而相較於線性模型對於特徵的解釋能力則會較為薄弱。



(a)邊緣與非邊緣　　(b)刮痕與非刮痕

圖 B.5　晶圓圖缺陷二元分類的一對剩餘策略

接著以下將分別介紹線性分類器中屬於判別模型的「羅吉斯迴歸」；以及屬於生成模型的「線性判別分析」。

B.2　羅吉斯迴歸

首先，對於一個二元分類問題，我們可將目標值的兩個不同類別以一個「**虛擬變數**」（dummy variable）表示，舉上述凹凸不平（bumps）與刮痕（scratch）兩種缺陷為例，可表示如公式(B.1)。

$$Y = \begin{cases} 0，如果為凹凸不平 \\ 1，如果為刮痕 \end{cases} \tag{B.1}$$

定義這個二元的虛擬變數後，我們可以將此目標值視為每個觀測值為刮痕缺陷的機率。當此機率為 1 時認為它是刮痕缺陷，相反地，當此機率為 0 時則認為它是凹凸不平缺陷。然而有了目標值後，為何無法直接使用前章節「線性迴歸」模型，對目標值為機率的數據進行建模以最小化誤差呢？

在線性迴歸模型中，我們並未對目標值的範圍給予限制，因此它會是一個介於正負無窮大範圍的實數，然而對於一個機率而言，它的範圍應當介於 0 與 1 之間。對於一個分類器來說，「**羅吉斯迴歸**」期望其目標值不超出機率應當落於的範圍。如圖 B.6 所示，左圖為使用線性迴歸模型（斜直線）配適二元虛擬變數的結果，右圖則為使用羅吉斯迴歸模型（斜曲線）配適的結果，可以發現右圖對於目標值的配適相對合理。接著我們介紹羅吉斯迴歸並解釋其對目標值的範圍進行轉換。

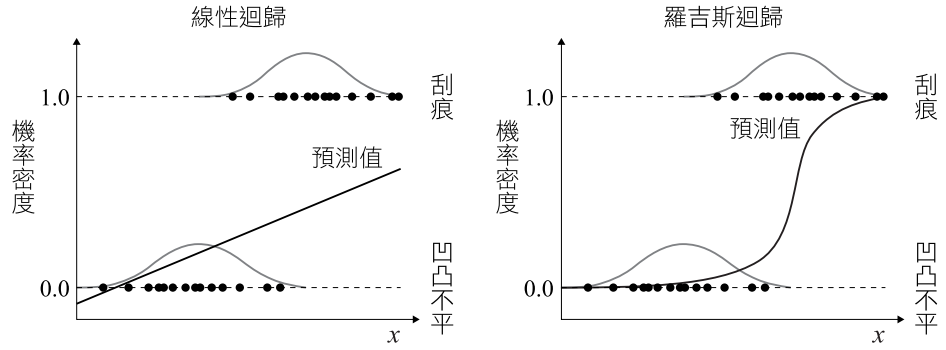


圖 B.6　線性迴歸與羅吉斯迴歸

B.2.1　模型假設與解釋

B.2.1.1　模型假設

若我們使用簡單線性迴歸配適一個目標值為機率的模型可表示如公式 (B.2)。

$$p(x) = \beta_0 + \beta_1 x \tag{B.2}$$

其中，$p(x)$為機率的目標值，x為特徵，迴歸係數β_0與β_1分別為截距項與

斜率項。然而，我們需要將上式右側的線性關係進行轉換，才能使得它的範圍介於 0 與 1 之間。因此，羅吉斯迴歸中使用的轉換函數即為「**羅吉斯函數**」(logistic function，又稱 sigmoid function)，如公式(B.3)所示。

$$f(x) = \frac{e^x}{1 + e^x} \tag{B.3}$$

上述的羅吉斯函數會將一個任意實數x的範圍轉換到 0 與 1 之間，並且此函數的形狀會是一個 S 型平滑曲線 (S-shaped) (如圖 B.6 右圖中的曲線)。因此，若我們將簡單線性迴歸模型公式(B.2)套入羅吉斯函數公式(B.3)中，則就成為一個簡單羅吉斯迴歸模型，如公式(B.4)。

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{B.4}$$

藉由估計上式的迴歸係數後，即可預測出一個合理的機率值$p(x)$。

B.2.1.2 模型解釋與勝算

若將公式(B.4)簡單代換後，如公式(B.5)所示。

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x} \tag{B.5}$$

可發現上式的左邊稱之為「**勝算**」(odds)，也就是計算成功機率是失敗機率的幾倍，而勝算會介於 0 與無限大之間，當成功失敗機率一半一半時，勝算會是$0.5/(1 - 0.5) = 1$。因此勝算大於 1 時成功機率較大，而勝算小於 1 時，則較易失敗。將羅吉斯迴歸模型替換成公式(B.5)的形式將有助於我們對特徵與目標關係的解釋。舉例而言，當x增加一單位時，則勝算會是原本的e^{β_1}倍 (由於是指數項，因此為相乘的「倍數關係」而非「加總關係」)。若接著將公式(B.5)的等式左右皆取對數後，如公式(B.6)所示，

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x \tag{B.6}$$

我們稱等號左邊為「**對數勝算**」(log-odds 或 logit)。上式我們可視為將機率透過「**羅吉斯鏈結函數**」(logit link function) 轉換到線性模型的形式，屬於廣義線性模型的其中一個案例 (後續將說明)，這是一個標準羅吉斯

迴歸表示的方式。

鋼板缺陷案例：模型假設

上述介紹了羅吉斯線性迴歸的模型假設，於是我們可以對鋼板缺陷數據提出一個羅吉斯迴歸的假設，如公式(B.7)表示，

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_{\text{pixel areas}} + \beta_2 x_{length\ \text{of conveyer}} \tag{B.7}$$

其中特徵$x_{\text{pixel areas}}$為鋼板面積，特徵$x_{length\ \text{of conveyer}}$為輸送帶長度，而目標值中的$p(x)$則是缺陷類型為刮痕缺陷的機率（公式(B.1)的虛擬變數）。上式對缺陷類型與鋼板面積關係的解釋為：當鋼板面積增加一單位時，缺陷類型為刮痕的勝算將會是原本的指數迴歸係數的e^{β_1}倍。因此當係數為正時，這個正相關代表著為鋼板面積越大時，刮痕類型缺陷的可能性更大；係數為負時的負相關，則刮痕類型缺陷的可能性降低。

B.2.2 迴歸係數的估計

在線性迴歸模型中，我們使用了最小平方法與最大概似法估計迴歸係數。在羅吉斯迴歸中，沒有誤差項難以使用最小平方法進行估計，若使用最大概似估計我們該如何假設數據的分配呢？

B.2.2.1 最大概似估計

在二元分類問題中，每個觀測值的實際目標y均為一個 0 或 1 值，於是我們可以合理假設它服從一個「**伯努利分布**」（Bernoulli distribution）。因此，我們可將概似函數表示如公式(B.8)所示。

$$\mathcal{L}(p_i) = \prod_{i=1}^{n} {p_i}^{y_i}(1-p_i)^{(1-y_i)} \tag{B.8}$$

將上式取對數後，如公式(B.9)所示。

$$
\begin{aligned}
l(p_i) &= \sum_{i=1}^{n} y_i \log p_i + (1-y_i)\log(1-p_i) \\
&= \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)
\end{aligned}
\tag{B.9}
$$

回顧上述假設的羅吉斯迴歸模型，若我們進一步將上式(B.9)的機率值p與特徵x連結，上式前部分的$\log\left(\frac{p_i}{1-p_i}\right)$可被替換為公式(B.6)的對數勝算，後部分的$p_i$則可替換為公式(B.4)的機率，因此對數概似函數可改寫成如公式(B.10)所示。

$$
\begin{aligned}
l(\beta_0,\beta_1) &= \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) + \log\left(1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right) \\
&= \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) - \log\left(1+e^{\beta_0+\beta_1 x_i}\right)
\end{aligned}
\tag{B.10}
$$

若改以多元羅吉斯迴歸表示（矩陣形式），則可以將上式(B.10)改寫成如公式(B.11)所示。

$$
l(\beta) = \sum_{i=1}^{n} y_i \mathbf{x}_i^T \beta - \log\left(1+e^{\mathbf{x}_i^T\beta}\right) \tag{B.11}
$$

$\mathbf{x}_i$代表特徵矩陣的第i個樣本，β為代表所有特徵迴歸係數的向量。接著，我們對β進行偏微分求極值，如公式(B.12)。

$$
\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta} &= \sum_{i=1}^{n} y_i \mathbf{x}_i^T - \frac{1}{1+e^{\mathbf{x}_i^T\beta}} \cdot e^{\mathbf{x}_i^T\beta} \cdot \mathbf{x}_i \overset{\text{set}}{=} 0 \\
&\Rightarrow \sum_{i=1}^{n} y_i \mathbf{x}_i^T - p_i \mathbf{x}_i \overset{\text{set}}{=} 0 \\
&\Rightarrow \mathbf{X}^T\left(\mathbf{y}-\mathbf{p}(\beta)\right) \overset{\text{set}}{=} 0
\end{aligned}
\tag{B.12}
$$

$\mathbf{X}^T$代表矩陣$[\mathbf{x}_1,\ldots,\mathbf{x}_n]$，$\mathbf{y}$代表向量（$y_1,\ldots,y_n$），$\mathbf{p}(\beta)$代表向量（$p_1,\ldots,p_n$）。然而，上式並無法直接求出函數根（root）以估計係數，這時需仰賴牛頓法的協助。

牛頓法

「牛頓法」（Newton-Raphson method, or Newton's method）是一種數值分析的方法，它使用了「泰勒展開」（Taylor expansion）對函數f進行求解，如公式(B.13)所示。

$$\underbrace{f(x)}_{\text{height of } x} \approx \underbrace{f(x_0)}_{\text{height of } x_0} + \underbrace{f'(x_0)}_{\text{slope of } x_0} \cdot \underbrace{(x - x_0)}_{\text{base from } x \text{ to } x_0} \tag{B.13}$$

其中x_0為一個任意點，上式的意思為函數$f(x)$可近似於任意點x_0的函數值加上任意點x_0斜率$f'(x_0)$乘上x到任意點x_0的距離$(x - x_0)$（可想像為一個在座標器上的直角三角形，從斜邊上比較高的點滑向低點）。因此一個函數的根（函數值等於零）可表示如公式(B.14)所示。

$$\begin{aligned} f(x) &\overset{\text{set}}{=} 0 \approx f(x_0) + f'(x_0)(x - x_0) \\ \Rightarrow x - x_0 &\approx -\frac{f(x_0)}{f'(x_0)} \\ \Rightarrow \quad x &\approx x_0 - \frac{f(x_0)}{f'(x_0)} \end{aligned} \tag{B.14}$$

依照上式近似關係可逐步迭代至最佳解，而這個過程便是牛頓法（演算法），如公式(B.15)所示。

$$x_t = x_{t-1} - \frac{f(x_{t-1})}{f'(x_{t-1})} \tag{B.15}$$

因此，在任意給定的初始x_0下，經由多次迭代後，x_t將漸進至最佳解（函數的根）。如圖 B.7 所示，左圖為第一次迭代，給定x_1經公式(B.15)迭代後得到x_2，右圖為第二次迭代，給定x_2同樣經迭代後得到x_3。依此類推，x_t將越來越趨近函式的根，牛頓法僅需要計算目標的函式與其一次微分，並不斷將現階段的最佳解代入公式(B.15)，不需要直接解複雜（可能非線性）函數的根。

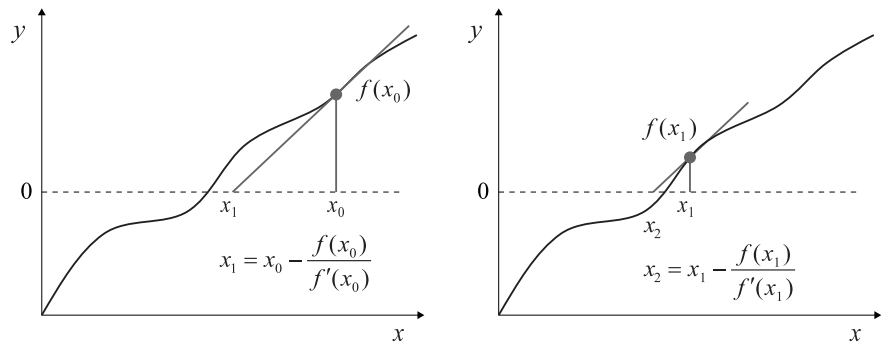


圖 B.7　牛頓法

B.2.2.2　牛頓法應用於羅吉斯迴歸估計

接續多元羅吉斯迴歸的最大概似估計，由於無法直接解出公式(B.12)的根，根據牛頓法，先求出此式的一次微分（原始最大概似函數的二次微分），如公式(B.16)所示。

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} &= \sum_{i=1}^{n} -\mathbf{x}_i p_i(1-p_i)\,\mathbf{x}_i^T \\ &= -\mathbf{X}^T\mathbf{V}(\beta)\mathbf{X}\end{aligned} \tag{B.16}$$

其中$\mathbf{V}(\beta)$代表對角矩陣其主對角元素為$diag\{p_1(1-p_1),\ldots,p_n(1-p_n)\}$。因此，以牛頓法表示的多元羅吉斯迴歸係數估計如公式(B.17)所示（代入公式(B.12)與公式(B.16)於公式(B.15)。

$$\beta_t = \beta_{t-1} + (\mathbf{X}^T\mathbf{V}(\beta)\mathbf{X})^{-1}\mathbf{X}^T\big(\mathbf{y}-\mathbf{p}(\beta)\big) \tag{B.17}$$

此外，為求迴歸係數的變異，我們基於費雪信息如公式(B.18)，以及最大概似估計的漸進常態性如公式(B.19)。

$$I(\theta) = -\mathrm{E}[l''(X|\theta)] = -\mathrm{E}\left[\frac{\partial^2 l(\theta)}{\partial\theta^2}\right] \tag{B.18}$$

$$\sqrt{n}(\hat{\theta}-\theta) \to N\left(0, \frac{1}{I(\theta)}\right) \tag{B.19}$$

我們可以推導出羅吉斯迴歸係數的變異，為費雪信息的倒數如公式(B.20)

所示。

$$\mathrm{Var}(\beta) = \frac{1}{I(\beta)} = \frac{1}{-\mathrm{E}\left[\frac{\partial^2 l(\beta)}{\partial \beta^2}\right]} = (\mathbf{X}^T\mathbf{V}(\beta)\mathbf{X})^{-1} \tag{B.20}$$

因此有了羅吉斯迴歸係數與變異的估計後，同樣地可以對各別的係數做統計檢定。

鋼板缺陷案例：估計與檢定

承續上述鋼板缺陷數據案例，我們對鋼板缺陷類型與兩特徵假設一個羅吉斯迴歸模型的線性分類器如公式(B.7)。接著，由上述最大概似估計法與牛頓法所得出的估計式(B.17)「估計迴歸係數」，並依照估計的變異公式(B.20)進行「統計檢定」，結果如表 B.1 所示。首先，在 95%的信心水準下，經檢定後三迴歸係數的 p-value 均遠小於 0.05，因此統計上相信他們均顯著不為零。其次，鋼板面積的迴歸係數為正，而輸送帶長度的則為負，代表著鋼板面積越大時越可能為刮痕缺陷，每增加一單位時，為刮痕缺陷的勝算原本的$e^{\beta_1} = 1.001$倍（因此圖 B.8 從 x 軸看決策邊界的斜率較陡）；而輸送帶長度越長時較可能凹凸不平的缺陷，每增加一單位時，為刮痕缺陷的勝算原本的$e^{\beta_2} = 0.9922$倍（因此圖 B.8 從 y 軸看決策邊界的斜率較平緩）。最後，若將羅吉斯迴歸模型視覺化後如圖 B.8 所示，圓點為刮痕缺陷，三角形為凹凸不平缺陷，因此依照上述迴歸係數的關係，最後得到的決策邊界為黑色的斜直線。

表 B.1　羅吉斯迴歸的估計結果

特徵（parameter）	迴歸係數（coefficient）	標準誤（standard error）	*t* 檢定統計量（t-statistic）	顯著程度（p-value）
截距項：β_0（intercept）	9.1956	1.8219	5.047	<0.0001***
鋼板面積：β_1（pixel areas）	0.0012	0.0001	8.767	<0.0001***
輸送帶長度：β_2（length of conveyer）	-0.0078	0.0013	-5.982	<0.0001***

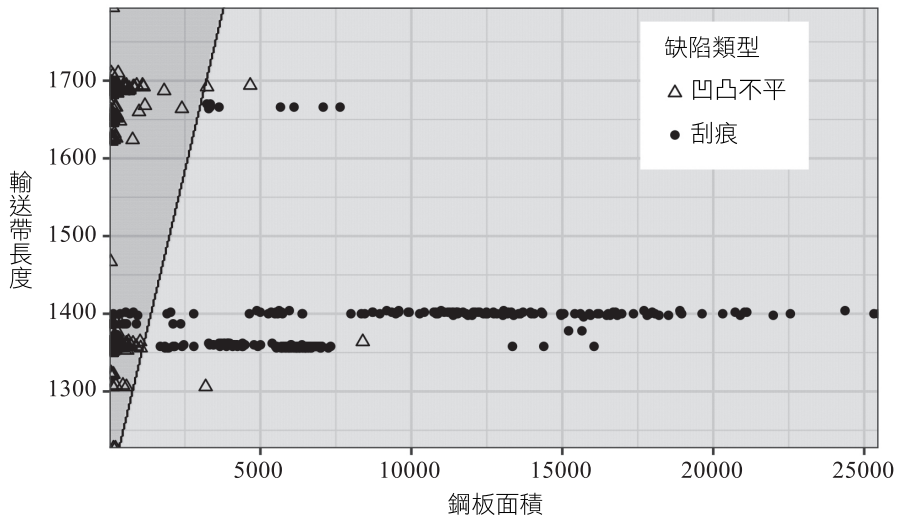


圖 B.8　羅吉斯迴歸於鋼板缺陷分類

B.2.3　多類別羅吉斯迴歸

上述介紹了多個特徵的多元羅吉斯迴歸於二元分類的問題，但倘若我們的目標值是多個水準的類別變數時（多類別分類問題），多類別羅吉斯迴歸應如何建構呢？多類別羅吉斯迴歸延續前述提到多類別問題，試圖將其轉為二元分類的一對一策略。若目標值的類別一共有K個時，首先我們先選定某一類別例如第K類為基底，接著建構所有其他類別與第K類之間的模型。若使用公式(B.6)中將迴歸表示為勝算對數的形式，則拓展至K個類別可表示如公式(B.21)所示。

$$\begin{aligned}
\log\left(\frac{\Pr(Y_i=1)}{\Pr(Y_i=K)}\right) &= \mathbf{x}_i^T\beta_1 \\
\log\left(\frac{\Pr(Y_i=2)}{\Pr(Y_i=K)}\right) &= \mathbf{x}_i^T\beta_2 \\
&\vdots \\
\log\left(\frac{\Pr(Y_i=K-1)}{\Pr(Y_i=K)}\right) &= \mathbf{x}_i^T\beta_{K-1}
\end{aligned} \tag{B.21}$$

上式需建構$K-1$個模型，因此有$K-1$組迴歸係數。然而，每個類別與第K類所建構的模型彼此是分開的，於是接著我們將公式(B.21)替換成 1 到

$K-1$類的機率，如公式(B.22)所示。

$$\begin{aligned}\Pr(Y_i=1)&=\Pr(Y_i=K)\,e^{\mathbf{x}_i^T\beta_1}\\ \Pr(Y_i=2)&=\Pr(Y_i=K)\,e^{\mathbf{x}_i^T\beta_2}\\ &\vdots\\ \Pr(Y_i=K-1)&=\Pr(Y_i=K)\,e^{\mathbf{x}_i^T\beta_{K-1}}\end{aligned}\tag{B.22}$$

因此，第K類的機率則會是 1 減去 1 到$K-1$類的機率加總，如公式(B.23)所示。

$$\begin{aligned}\Pr(Y_i=K)&=1-\sum_{k=1}^{K-1}\Pr(Y_i=k)=1-\sum_{k=1}^{K-1}\Pr(Y_i=K)\,e^{\mathbf{x}_i^T\beta_k}\\ \Longrightarrow \Pr(Y_i=K)&=\frac{1}{1+\sum_{k=1}^{K-1}e^{\mathbf{x}_i^T\beta_k}}\end{aligned}\tag{B.23}$$

有了第K類的機率後，我們即可反推回公式(B.22)中 1 到$K-1$類的機率，便可計算出新數據於每一類預測的機率，如公式(B.24)。

$$\begin{aligned}\Pr(Y_i=1)&=\frac{e^{\mathbf{x}_i^T\beta_1}}{1+\sum_{k=1}^{K-1}e^{\mathbf{x}_i^T\beta_k}}\\ \Pr(Y_i=2)&=\frac{e^{\mathbf{x}_i^T\beta_2}}{1+\sum_{k=1}^{K-1}e^{\mathbf{x}_i^T\beta_k}}\\ &\vdots\\ \Pr(Y_i=K-1)&=\frac{e^{\mathbf{x}_i^T\beta_{K-1}}}{1+\sum_{k=1}^{K-1}e^{\mathbf{x}_i^T\beta_k}}\end{aligned}\tag{B.24}$$

在多類別羅吉斯迴歸中所使用的模型個數僅需要$K-1$個，而前述提及的一對一策略需建構$C_2^K=\frac{K(K-1)}{2}$個模型，這是因為多類別羅吉斯迴歸可求出每個類別的機率，而非僅僅是決策邊界，體現了羅吉斯迴歸的另一項優點（但仍需建構多個模型）。

B.2.4 廣義線性模型

一般來說，「線性迴歸」假設了誤差服從「常態分配」；在本章的「羅

吉斯迴歸」則假設目標變數服從「二項分配」，這兩個模型皆屬於「**廣義線性模型**」（generalized linear model, GLM）的範疇。廣義的線性模型是如何定義與推廣的？可應用的情境為何？（廣義的線性模型並非侷限於分類問題）

B.2.4.1 鏈結函數

首先我們將廣義線性模型分為兩部分，第一部分為「**線性預測子**」（linear predictor）η，如公式(B.25)所示。

$$\eta = \mathbf{X}^T\beta \tag{B.25}$$

上式中的預測子η與迴歸係數β在這邊是未知的，由於特徵是固定的，因此我們視線性預測子為「**系統性成分**」（systematic component）。第二部分我們會對目標變數Y做出分配的假設，而這個分配必須屬於「**指數族**」（exponential family）（後續會詳細說明），而此隨機變數的條件期望值如公式(B.26)所示。

$$\mathrm{E}\left(Y|X=x\right) = \mu \tag{B.26}$$

這個期望值μ就是我們要估計的目標，由於假設目標變數是隨機變數，因此我們視指數族分配為「**隨機性成分**」（random component）。

廣義線性模型即是結合上述的線性預測子（公式(B.25)）與指數族分配的期望值（公式(B.26)）。也就是將線性預測子η以隨機變數期望值μ的函數表示（使模型同時包含系統性與隨機性成分），而這樣的函數稱為「**鏈結函數**」（link function）g，如公式(B.27)所示。

$$g(\mu) = \mathbf{X}^T\beta \tag{B.27}$$

而不同的分配假設則需不同的鏈結函數，例如常態分配的鏈結函數為「**恆等鏈結**」（identity link）：$g(\mu)=\mu$；二項分配的鏈結函數為「**羅吉斯鏈結**」（logit link）：$g(\mu)=\log\left(\frac{\mu}{1-\mu}\right)$；卜瓦松分配的鏈結函數為「**自然對數鏈結**」（log link）：$g(\mu)=\log(\mu)$。因此，當我們已知（或給定）目標變數的分配時，只要找到合適的鏈結函數轉換，即可建構出一個線性的模型。然而，這樣的關係必須建立在該分配屬於指數族的條件下，以下我們說明指數族的分配定義、指數族的特性以及如何決定指數族分配的鏈結函數。

B.2.4.2 指數族

一個隨機變數Y若屬於「指數族」，它的機率密度函數可表示如公式(B.28)所示。

$$f_Y(y;\theta,\tau)=\exp\left\{\frac{y\cdot\theta-b(\theta)}{\tau^2}-c(y,\tau)\right\} \tag{B.28}$$

其中θ為指數族的自然參數（natural parameter），τ為一個散度（離勢）參數（dispersion parameter），b與c為兩個已知的函數。舉例而言，常態分配的機率密度函數以指數族表示呈現如公式(B.29)所示。

$$\begin{aligned} & Y\sim N(\mu,\sigma^2) \\ \Longrightarrow f_Y(y) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}=\exp\left\{\frac{y\cdot\mu-\mu^2/2}{\sigma^2}+C\right\} \\ &= \exp\left\{\frac{y\cdot\theta-\theta^2/2}{\tau^2}+C\right\} \end{aligned} \tag{B.29}$$

其中 $\theta=\mu, b(\theta)=\theta^2/2$與 $\tau^2=\sigma^2$。伯努利分配的機率密度函數以指數族表示呈現如公式(B.30)所示。

$$\begin{aligned} & Y\sim Ber(p) \\ \Longrightarrow f_Y(y) &= p^y(1-p)^{(1-y)}=\exp\left\{y\log\left(\frac{p}{1-p}\right)+\log(1-p)\right\} \\ &= \exp\{y\cdot\theta-\log(1+e^\theta)\} \end{aligned} \tag{B.30}$$

其中 $\theta=\log\left(\frac{p}{1-p}\right)$，$b(\theta)=\log(1+e^\theta)$與$\tau^2=1$。卜瓦松分配的機率密度函數可以指數族表示呈現如公式(B.31)所示。

$$\begin{aligned} & Y\sim Poi(\lambda) \\ \Longrightarrow f_Y(y) &= \frac{\lambda^y e^{-\lambda}}{y!}=\exp\{y\log(\lambda)-\lambda+C\}=\exp\{y\cdot\theta-e^\theta+C\} \end{aligned} \tag{B.31}$$

其中 $\theta=\log(\lambda), b(\theta)=e^\theta$與$\tau^2=1$。

指數族有以下幾點重要特性：

- 第一個特性在於其指數族的自然參數所對應到原參數的函數，即為鏈結函數。經由鏈結函數的轉換便可連結線性預測子與指數族

分配的期望值，例如伯努利分布的鏈結函數為$\log\left(\frac{p}{1-p}\right)$，即是羅吉斯迴歸對目標的轉換。

- 指數族第二個特性則在於，對指數族呈現的機率密度函數若以「**動差法估計**」（method of moment estimation）可得到隨機變數的期望值與變異數分別如公式(B.32)所示。

$$\begin{aligned} \mathrm{E}(Y) &= \mu = \frac{\partial b(\theta)}{\partial \theta} \\ \mathrm{Var}(Y) &= \tau^2 \frac{\partial^2 b(\theta)}{\partial \theta^2} \end{aligned} \tag{B.32}$$

- 指數族第三個特性則是可用最大概似法對廣義線性模型進行估計。我們可分別得到概似函數的一次與二次微分後，使用牛頓法估計線性模型裡的迴歸係數與變異（詳細數理推導可參閱相關文獻）。

因此，當我們已知目標變數的分配（屬於指數族）時，即可建構一個廣義線性模型進行解釋與預測，其不限於常態分配的線性迴歸或二項分配的羅吉斯迴歸，也可以是服從卜瓦松分配的迴歸等。

B.3 判別分析

在羅吉斯迴歸中，我們直接建構了特徵X與目標值Y的關係$\Pr(Y = k|X = x)$，屬於判別模型的思維。而此節的「**判別分析**」（discriminant analysis）則是先建構在不同類別時特徵的分布$\Pr(X = x|Y = k)$，並且假設這些分布為常態分配，接著再使用貝氏定理將其轉換為$\Pr(Y = k|X = x)$的形式成為一個分類器，屬於生成模型的思維。線性判別分析與羅吉斯迴歸的相異之處如下：

- 當特徵於特定類別的分布近似於常態分配時，判別分析的效果可能比羅吉斯迴歸來得較好，然而與常態分配差異很大時，則效果可能較差。
- 如同前述提及的，在多類別分類的問題下，判別分析的使用會比羅吉斯迴歸來得容易。
- 當數據的分類結果相當完美時（誤判情況趨近於零），羅吉斯迴歸

的係數會非常大（或非常小）且不穩定，而判別分析則不受影響。

在本節中，首先我們將介紹判別分析的前身「**樸素貝氏分類器**」（naïve Bayes classifier），雖然它是一個非線性的分類器，但貝氏的思想對於後續的方法介紹與推廣相當重要。接著我們會提到屬於線性分類器的「**線性判別分析**」（linear discriminant analysis, LDA），並延伸介紹非線性的「**二次判別分析**」（quadratic discriminant analysis, QDA）。

B.3.1 樸素貝氏分類器

「樸素貝氏分類器」是一個基於「**貝氏定理**」（Bayes' Theorem）與假設特徵之間彼此獨立的機率分類器（通常為非線性方法）。基於貝氏定理，樸素貝氏分類器可表示如公式(B.33)。

$$\underbrace{\Pr(Y=k|X=x)}_{\text{posterior}}=\frac{\overbrace{\Pr(X=x|Y=k)}^{\text{likelihood}}\overbrace{\Pr(Y=k)}^{\text{prior}}}{\underbrace{\sum\Pr(X=x)}_{\text{margin}}} \tag{B.33}$$

其中$\Pr(Y=k|X=x)$為我們欲建構條件機率分類器，在貝氏中稱之為「**後驗分配**」（posterior）；$\Pr(Y=k)$為樣本屬於某一類別的機率，在貝氏中稱之為「**先驗分配**」（prior）；$\Pr(X=x|Y=k)$則是給定類別下該特徵的機率，在貝氏中稱之為「**概似函數**」（likelihood）；而分母$\sum\Pr(X=x)$則是將所有可能的加總，在貝氏中稱之為「**邊際和**」（margin）。

對於先驗分配$\Pr(Y=k)$來說，僅需計算該類別佔全部樣本的機率，這通常可由歷史數據得知。但概似函數$\Pr(X=x|Y=k)$的計算，則需要給定特徵X的機率密度函數（或分布）。因此，樸素貝氏分類器在處理離散的特徵時，可直接從數據中計算概似函數值；而在處理連續的特徵時，則需離散化或假設某個特定的分配（亦可使用無母數方法估計分配）。對於「線性判別分析」方法就屬於假設「常態分配」的特例。此外，在樸素貝氏分類器中，還會對特徵之間假設彼此是相互獨立的，因此我們可將概似函數拆解如公式(B.34)所示。

$$\begin{aligned}\underbrace{\Pr(X=x|Y=k)}_{\text{likelihood}}&=\Pr\left(X_1,X_2,\dots,X_p\middle|Y=k\right)\\&=\Pr(X_1|Y=k)\Pr(X_2|Y=k)\dots\Pr\left(X_p\middle|Y=k\right)\end{aligned} \tag{B.34}$$

當我們有每個特徵各自的分布時，即可計算所有獨立的概似函數相乘的結果。

鋼板缺陷案例：樸素貝氏分類器

承續上述鋼板缺陷數據案例，我們對鋼板缺陷類型與兩特徵使用樸素貝氏分類器（非線性模型）如公式(B.33)。由於我們選擇的特徵皆為連續型的變數，因此特徵的分布採用無母數的「核密度函數估計」(kernel density estimation)(後續將於章節「無母數迴歸與分類」說明)，結果如圖 B.9 所示。我們可以看到由凹凸不平缺陷所建構的無母數分配為非線性，雖這個非線性的樸素貝氏分類器在分類的精確度上表現不錯，但實際上，可能潛在過度配適的問題，還需對無母數分配中的超參數進行調整。

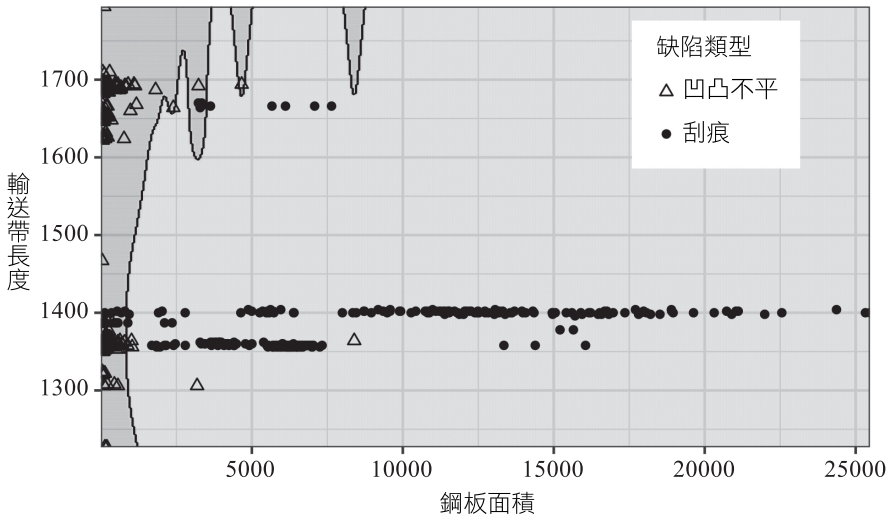


圖 B.9　樸素貝氏分類器於鋼板缺陷分類

B.3.2　線性判別分析

在前述樸素貝氏分類器中，我們需要在已知特徵X的機率密度函數下，才能計算概似函數$\Pr(X = x|Y = k)$，然而在線性判別分析中，通常我們會假設連續的特徵服從**常態分配**（或有相關領域知識下，可假設其他的分配）。若有多個特徵時，則為多變量常態分配（特徵間可不獨立）。基於

這樣的假設，我們便可以找出每個樣本於不同類的機率，以及類別之間的決策邊界。

B.3.2.1 一個特徵的線性判別分析

首先，若我們將樸素貝氏分類器公式(B.33)簡化如公式(B.35)所示。

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum \pi_k f_k(x)} \tag{B.35}$$

其中π_k為每個樣本屬於第k類的機率（先驗機率），$f_k(x)$為基於某類別下該特徵的機率（概似函數），$p_k(x)$為基於該特徵下每個樣本於不同類的機率（後驗機率）。其次，在一個維度的線性判別分析假設概似函數$f_k(x)$為常態分配的機率密度函數如公式(B.36)所示。

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{\frac{-(x-\mu_k)^2}{2\sigma_k^2}\right\} \tag{B.36}$$

其中μ_k與${\sigma_k}^2$為該特徵在第k類的平均數與變異數。接著，我們假設特徵在所有類別之間的變異數是相等的如$\sigma_1^2 = \cdots = \sigma_k^2 \overset{\text{set}}{=} \sigma^2$（線性判別分析的強假設），因此，把常態分配的機率密度函數公式(B.36)代入樸素貝氏分類器公式(B.35)的概似函數後，可得到如公式(B.37)所示。

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu_k)^2}{2\sigma^2}\right\}}{\sum \pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu_k)^2}{2\sigma^2}\right\}} \tag{B.37}$$

將公式(B.37)取對數整理後，可得到「判別函數」（discriminant function）$\delta_k(x)$如公式(B.38)所示。

$$\delta_k(x) = x^2 \cdot \frac{-1}{2\sigma^2} + x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k - C \tag{B.38}$$

其中C為一常數（公式(B.37)中的分母），當我們得到某個觀測值$X = x$代入上式後，便可得到每個類別k各別的機率（取對數後的）。接著，假設有兩個類別分別為k與l，若我們想找出兩類間的決策邊界，即是兩個類別機率相等的地方，如公式(B.39)所示。

$$\begin{aligned}
&\delta_k(x) \overset{\text{set}}{=} \delta_l(x) \\
\Longrightarrow &\left(x^2 \cdot \frac{-1}{2\sigma^2} + x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k - C\right) - \\
&\left(x^2 \cdot \frac{-1}{2\sigma^2} + x \cdot \frac{\mu_l}{\sigma^2} - \frac{\mu_l^2}{2\sigma^2} + \log \pi_l - C\right) \overset{\text{set}}{=} 0 \\
\Longrightarrow &\, x \cdot \frac{\mu_k - \mu_l}{\sigma^2} - \frac{\mu_k^2 - \mu_l^2}{2\sigma^2} + (\log \pi_k - \log \pi_l) \overset{\text{set}}{=} 0
\end{aligned} \tag{B.39}$$

因此在上式中，決定觀測值所屬類別的決策邊界為特徵x的一次線性函數，這就是我們稱之為線性判別分析中「線性」的原因。線性判別分析如同羅吉斯迴歸皆屬於線性分類器，而在一個維度的線性判別分析對應到的為邊界點（二維則為線）。若假設先驗機率相等$\pi_k = \pi_l$的情況下，決策邊界點x_{db}將如公式(B.40)所示。

$$x_{db} = \frac{\mu_k^2 - \mu_l^2}{2(\mu_k - \mu_l)} = \frac{\mu_k + \mu_l}{2} \tag{B.40}$$

舉例說明，如圖 B.10 所示，若我們隨機生成來自常態分配的兩個類別，其母體參數為平均數$\mu_1 = -1.25$、$\mu_2 = 1.25$，以及變異數$\sigma_1^2 = \sigma_2^2 = 1$。套用上述的決策邊界點公式(B.40)，則理論上邊界點為圖中虛線$x_{\text{boundary}} = (-1.25 + 1.25)/2 = 0$。若實際生成數據後，產生的樣本（圓點）所估計的決策邊界為圖中實線。

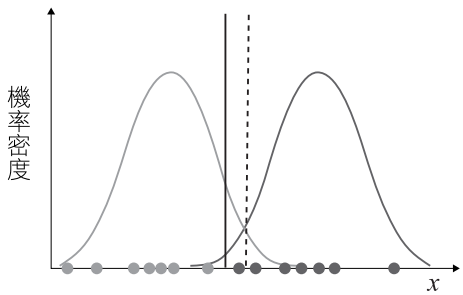


圖 B.10　線性判別分析

因此在實務上，我們需估計每個類別常態分配的平均數$\hat{\mu}_k$以及假設每類均相等的變異數$\hat{\sigma}^2$，估計式如公式(B.41)所示。

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i$$
$$\hat{\sigma}^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)^2 \tag{B.41}$$

其中n_k為第k類的樣本數，n為總樣本數。此外，在沒有任何的額外知識下，先驗機率π_k通常使用已知的數據直接估計，如公式(B.42)。

$$\hat{\pi}_k = \frac{n_k}{n} \tag{B.42}$$

在線性判別分析假設類別間特徵的變異數相等下，我們可將線性判別分析分類器$\hat{G}_{lda}$可由公式(B.38)引入估計值公式(B.41)與公式(B.42)，表示如公式(B.43)。

$$\begin{aligned}\hat{G}_{lda}(x) &= \underset{k}{\operatorname{argmax}}\,\hat{\delta}_k(x)\\ &= \underset{k}{\operatorname{argmax}}\, x\cdot\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log\hat{\pi}_k\end{aligned} \tag{B.43}$$

然而類別間「等變異」的假設與實務問題相去甚遠，因此在後續小節中介紹移除此假設的「二次判別分析」。

B.3.2.2 多個特徵與多個類別的線性判別分析

若我們將一維度的特徵X拓展到多維度的情境下，會將原本特徵服從一維常態分配拓展到多維度的多變量常態分配。多變量常態分配假設每一個維度皆服從常態分配，並且不同特徵間彼此存在某種相關性。在數學上，多變量常態分配通常可簡化表示為$X\sim N(\mu,\Sigma)$，平均數$\mu = \mathrm{E}(X)$為p個維度的向量，以及「**共變異矩陣**」（covariance matrix）$\Sigma = \mathrm{Cov}(X)$為$p\times p$的矩陣，因此不同類別的多變量常態分配機率密度函數表示如公式(B.44)所示。

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right\} \tag{B.44}$$

同樣地，我們假設不同類別之間的共變異矩陣是相等的如$\Sigma_1 = \cdots = \Sigma_k$

$\overset{\text{set}}{=}\Sigma$，接著，把多變量常態分配代入貝氏分類器後，所得到的的判別函數$\delta_k(x)$如公式(B.45)所示（為公式(B.38)的多變量形式）。

$$\delta_k(x) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k - C \tag{B.45}$$

當我們得到某個觀測值$X = x$代入上式後，便可得到每個類別k各別取對數後的機率，假設有兩個類別分別為k與l，若我們想找出決策邊界，即是兩個類別機率相等的地方，如公式(B.46)所示（為公式(B.39)的多變量形式）。

$$x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k \overset{\text{set}}{=} x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + \log\pi_l \tag{B.46}$$

因此，線性判別分析分類器$\hat{G}_{lda}$可表示如公式(B.47)所示（為公式(B.43)的多變量形式）。

$$\begin{aligned}\hat{G}_{lda}(x) &= \arg\max_k \hat{\delta}_k(x) \\ &= \arg\max_k x^T\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + \log\hat{\pi}_k\end{aligned} \tag{B.47}$$

此外，除了拓展至多個特徵外，若為多類別分類問題時，線性判別分析僅需估計k組多變量常態分配的參數即可。舉例說明，如圖 B.11 所示，若我們隨機生成來自多變量常態分配的三個類別，其母體參數分別為平均數$\mu_1 = (-1,-1)$、$\mu_2 = (3,1)$、$\mu_3 = (1,3)$，以及共變異矩陣於兩特徵間為一個大於零的常數且$\Sigma_1 = \Sigma_2 = \Sigma_3$。其中每個橢圓為各類別的多變量常態分配等高線，套用上述的決策邊界點公式(B.46)後，即可找出實際的決策邊界，為圖中虛線。若實際生成數據後，產生的樣本（圓點）所估計的決策邊界則為圖中實線。

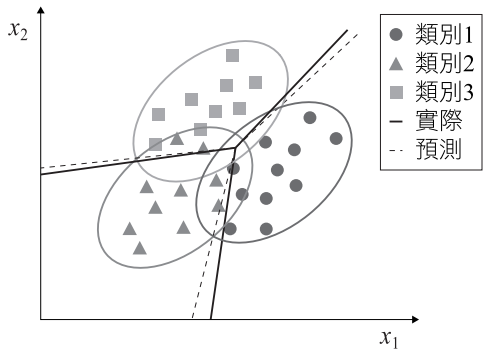


圖 B.11　三個類別的二元變量常態分配

鋼板缺陷案例：線性判別分析

承續上述鋼板缺陷數據案例，我們對鋼板缺陷類型與兩特徵使用線性判別分析分類器如公式(B.47)，結果如表 B.2、表 B.3、表 B.4 所示。首先，表 B.2 為兩種缺陷的先驗機率$\hat{\pi}_k$，為各類別樣本數的佔比如公式(B.42)。表 B.3 與表 B.4 為兩種缺陷各自多變量常態分配的所估計的平均數$\hat{\mu}_k$與共變異矩陣$\hat{\Sigma}$，估計的計算為公式(B.41)的多變量版本。有了這些估計值後，便可計算出線性判別分析的決策邊界如公式(B.46)與分類器如公式(B.47)，我們將結果視覺化後呈現如圖 B.12 所示。圓點為刮痕缺陷，三角形為凹凸不平缺陷，線性判別分析所建構的決策邊界為斜直線，而圖中的兩個橢圓則是分別描繪凹凸不平與刮痕缺陷各自估計的多變量常態分配，兩橢圓相交的地方即是決策邊界。此外，與羅吉斯迴歸模型（圖 B.9）比較，呈現了線性分類器中判別模型與生成模型的差異。

表 B.2　鋼板缺陷數據的先驗機率

	凹凸不平	刮痕
機率	0.508	0.492

表 B.3　不同缺陷類型在兩特徵上的平均值

	鋼板面積（X1）	輸送帶長度（X2）
凹凸不平	238.47	1522.94
刮痕	7250.78	1384.15

表 B.4　鋼板缺陷數據兩特徵上的共變異矩陣

	鋼板面積	輸送帶長度
鋼板面積	26639978	-229104
輸送帶長度	-229104	19297

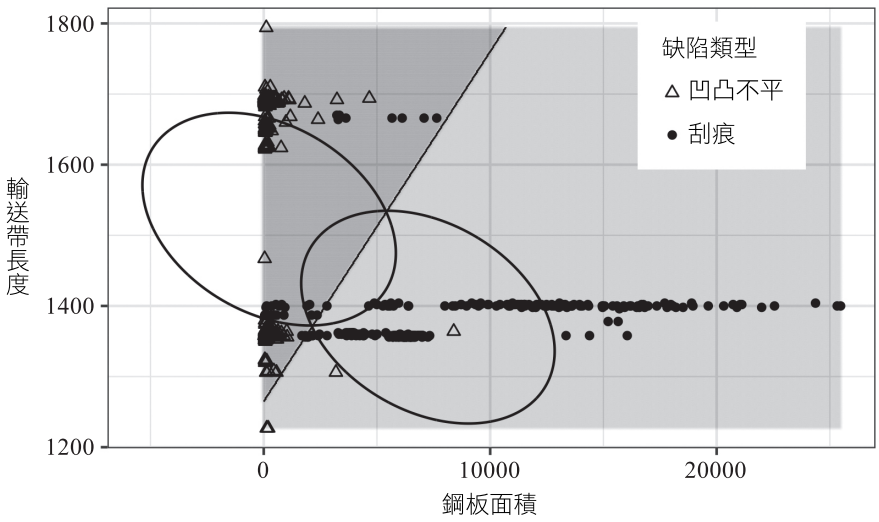


圖 B.12　線性判別分析於鋼板缺陷分類

B.3.3　二次判別分析

在前述的線性判別分析中，對類別間的變異（共變異矩陣）做出了相等的強假設。很顯然的，這與實際的數據經常相去甚遠。因此我們接著介紹的非線性生成模型「二次判別分析」打破了這個假設，並分別估計類別間各自多變量常態分配的共變異矩陣。

回顧以代入多變量常態分配為概似函數所得到的判別函數$\delta_k(x)$如公式(B.45)所示。若共變異矩陣並未假設相等時，則如同估計平均數一樣需各別估計每一類別的共變異矩陣Σ_k。因此，原判別函數$\delta_k(x)$公式(B.45)可改寫如公式(B.48)所示。

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k - C \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| \\ &\quad + \log\pi_k - C\end{aligned} \tag{B.48}$$

同樣地，假設有兩個類別分別為k與l，若我們想找出決策邊界，即是兩個類別機率相等的地方，如公式(B.49)。

$$\begin{aligned} &-\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \\ &\overset{\text{set}}{=} -\frac{1}{2}x^T\Sigma_l^{-1}x + x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l - \frac{1}{2}\log|\Sigma_l| + \log\pi_l \end{aligned} \tag{B.49}$$

上式與「線性判別分析」公式(B.46)的差異在於特徵的二次項（$-\frac{1}{2}x^T\Sigma_k^{-1}x$），由於不等變異的假設，無法相互抵消。也就是說，假設等變異時，特徵的二次項相消使得線性判別模型為特徵的一次關係，因而是線性模型，而去除了等變異假設後，特徵的二次項無法相消，因而二次判別分析為一個二次的非線性模型。因此，二次判別分析分類器$\hat{G}_{qda}$可表示如公式(B.50)。

$$\begin{aligned} \hat{G}_{lda}(x) &= \arg\max_k \hat{\delta}_k(x) \\ &= \arg\max_k -\frac{1}{2}x^T\hat{\Sigma}_k^{-1}x + x^T\hat{\Sigma}_k^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}_k^{-1}\hat{\mu}_k + \log\hat{\pi}_k \end{aligned} \tag{B.50}$$

以二維分類問題舉例說明，如圖 B.13 所示。真實的決策邊界為實線曲線，從圖中我們可以看到兩類別各自所展現的變異有所不同，左下樣本在兩特徵呈現較為正相關，而右上樣本在兩特徵呈現較為負相關，因此若使用等變異假設的線性判別分析實際上的分類效果不佳。若使用二次判別分析的結果如圖中虛線曲線，與真實的決策邊界較為相近。另一方面，若比較線性判別分析與二次判別分析，除了常態分配是否假設等變異外，因二次判別分析需各別估計不同類別間的共變異，因此所需估計的參數個數會是原本的k倍，一共為$kp(p+1)/2$個。從某個角度來說，雖然增加了模型的彈性，但同時也增加了過度配適的可能性。

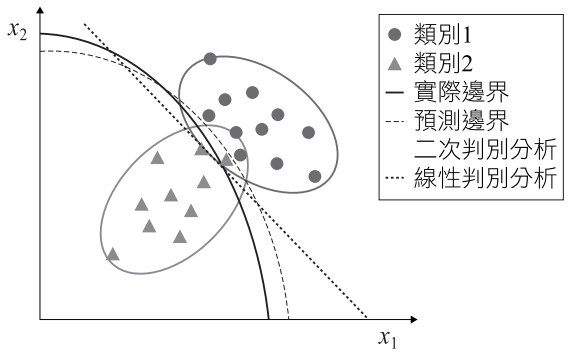


圖 B.13　線性判別分析與二次判別分析

鋼板缺陷案例：二次判別分析

承續上述鋼板缺陷數據案例，我們對鋼板缺陷類型與兩特徵使用二次判別分析分類器如公式(B.50)，結果如表 B.5、表 B.6 所示。首先，表 B.5 與表 B.6 分別為兩種缺陷的先驗機率$\hat{\pi}_k$與常態分配所估計的平均數$\hat{\mu}_k$，與線性判別分析的估計是相同的。其次，各缺陷的共變異矩陣$\hat{\Sigma}_k$則是分別估計，如表 B.5、表 B.6 所示。有了這些估計值後，便可計算出二次判別分析的決策邊界如公式(B.49)與分類器如公式(B.50)，我們將結果視覺化後呈現如圖 B.14 所示。圓點為刮痕缺陷，三角形為凹凸不平缺陷，二次判別分析所建構的決策邊界為曲線。而圖中的兩個橢圓則是分別描繪凹凸不平與刮痕缺陷各自估計的多變量常態分配，由於不等變異使得兩橢圓的方向性不相同。因此，兩橢圓相等的地方所形成的決策邊界才會呈現非線性的曲線。此外，我們可與線性判別模型（圖 B.12）比較，雖我們知道兩種缺陷的特徵分布並非近似於常態分配，但在此案例中二次判別分析的分類器於兩種缺陷中，所建構的不等變異常態分配與數據較為相符，因此分類效果也較線性判別模型來的好些，與羅吉斯迴歸相比差異不大（圖 B.8）。

表 B.5　鋼板缺陷數據於凹凸不平類別上的共變異矩陣

共變異矩陣（凹凸不平缺陷）	鋼板面積	輸送帶長度
鋼板面積	315222	-1194
刮痕	-1194	26101

表 B.6　鋼板缺陷數據於刮痕類別上的共變異矩陣

共變異矩陣（刮痕缺陷）	鋼板面積	輸送帶長度
鋼板面積	28822353	30606
刮痕	30606	2532

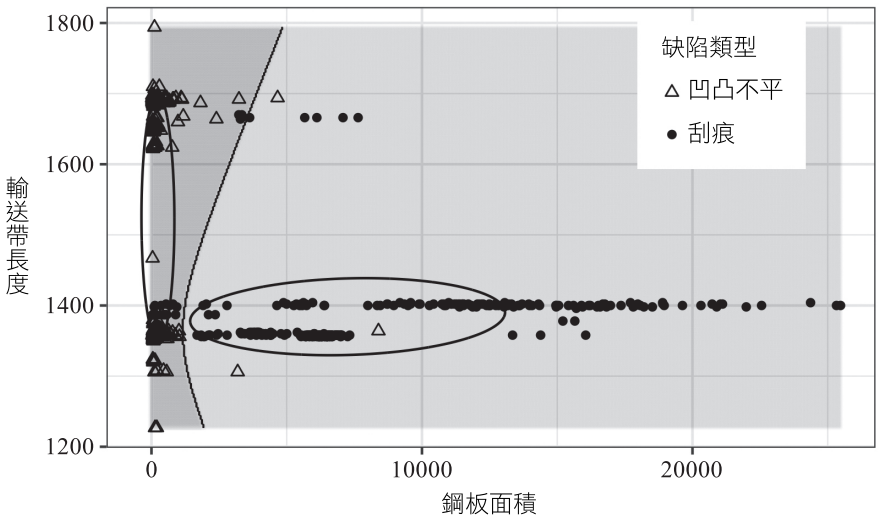


圖 B.14　二次判別分析於鋼板缺陷分類

B.4　結語

本章介紹線性分類器中的「**羅吉斯迴歸**」（判別模型）與「**線性判別分析**」（生成模型），同時說明二元與多類別分類的方法，並拓展到**廣義線性模型**、**樸素貝氏分類器**與**二次判別分析**。相似於線性迴歸，雖然分類效果表現欠佳，但仍有不錯的解釋能力，對於決策邊界的呈現相對清楚，且能提供數學式來說明其變數間影響的關係與解釋（例如羅吉斯迴歸的x變化如何影響機率與勝算），是相當有用的。線性分類器是機器學習與數據科學用於分類問題的基礎方法，從線性分類器出發，也延伸許多各式各樣的模型與應用，例如**支持向量機**（support vector machine）、**分類與迴歸樹**（classification and regression tree）、**隨機森林**（random forest）、**提升法**

（boosting）等，在處理分類問題上皆常用到線性分類器的觀念。事實上，近幾年熱門流行的**集成學習**（ensemble learning）中，通常也以弱分類器（weak classifier）（例如線性分類器）為基底模型（base model），用以建構強分類器，來提升預測與分類的正確率。另一方面，廣為人知的深度學習中，其生成模型的觀念與基礎，也可從線性判別分析的學習中窺知一二。因此，這些新技術新方法的發展，皆說明了線性分類器對問題本質的解析，以及其數理基礎的重要性。

參考文獻

[1] 陳順宇（2005），多變量分析，四版，華泰文化。

[2] Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. 1st edition, Springer.

[3] Johnson, R., and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. 6th edition, Pearson.

問題與討論

1. 試說明在分類的思維中，(a)判別模型（discriminative model）與生成模型（generative model）各別的概念為何？（提示：試以視覺化呈現與說明，若能以案例說明尤佳）(b)並說明建構多類別分類模型轉二元分類有哪些策略？(c)優劣與使用時機又為何？
2. 對於二元分類問題，我們可將目標值以啞變數的形式表示（並可將目標值視為機率）。(a)為什麼對目標值為連續的機率值無法直接建構線性迴歸模型呢？(b)羅吉斯迴歸對於目標值的啞變數如何進行轉換呢？(c)其迴歸係數如何估計呢？
3. 試著以一個二元的特徵（X）與二元的目標值（Y）舉例說明貝氏分類器是如何以貝氏定理進行推論的。（提示：先驗分配、概似函數以及後驗分配各別為何）
4. 試解釋為什麼線性判別分析與二次判別分析分別為線性與非線性模型。（提

示：兩者在假設有何差異，可以視覺化輔助說明，以及各別的決策邊界是如何求得的）

5. 試說明羅吉斯迴歸與線性判別分析各別的假設、優劣與使用時機。
6. 在 UCI Machine Learning Repository 開放數據中包含了一個鋼板缺陷數據（steel plates faults dataset，https://archive.ics.uci.edu/ml/datasets/steel+plates+faults），一共包含了 1,941 個觀測值，而每個觀測值具有 27 個特徵以及作為目標值的 7 種缺陷。試挑選出凹凸不平（Bumps）以及刮痕（K_Scratch）兩種缺陷進行分析。試著參考網路資源學習並撰寫程式，使用此數據回答下列問題：

(a) 試將羅吉斯迴歸分析的結果呈現如下表，並試著解釋任一特徵與目標值之間的關係。

	estimate	std. error	t value	p-value
intercept				
X_Minimum				
X_Maximum				
…				
Sigmoid Of Areas				

R-squared: 0.xxxx, Adjusted R-squared: 0.xxxx

(b) 基於上述(a)的結果，將上述特徵以 t value 進行排序後，哪些特徵的迴歸係數在統計上是顯著的呢（p-value<0.01）？

(c) 試問配適一個羅吉斯迴歸模型是否合適？試若配適不佳，試說明其可能的原因為何？

(d) 試問配適一個線性判別分析模型是否合適？若配適不佳，試說明其可能的原因為何？

(e) 試問配適一個二次判別分析模型是否合適？若配適不佳，試說明其可能的原因為何？