Linear Classifier
(第 6 章 線性分類器)

Chia-Yen Lee, Ph.D. (李家岩 博士)

Department of Information Management (資訊管理學系)
National Taiwan University (國立台灣大學)

目次

- 第一章 製造數據科學
- 第二章 製造系統分析與管理
- 第三章 數據科學基礎與模型評估
- 第四章 數據科學分析架構與系統運算決策
- 第五章 數據預處理與製造數據特性
- **第六章 線性分類器**
- 第七章 無母數迴歸與分類
- 第八章 決策樹與集成學習
- 第九章 特徵挑選與維度縮減
- 第十章 類神經網路與深度學習
- 第十一章 集群分析
- 第十二章 特徵工程、數據增強與數據平衡
- 第十三章 故障預測與健康管理
- 第十四章 可解釋人工智慧
- 第十五章 概念漂移
- 第十六章 元啟發式演算法
- 第十七章 強化學習

藍：老師課堂講授
綠：學生自學

- 附錄A 線性迴歸
- 附錄B 支持向量機
- 附錄C 統計製程管制與先進製程控制
- 附錄D 超參數最佳化

- 應用涵蓋

  產能規劃、瑕疵檢測、製程監控與診斷、機台保養、需求預測、 生產排程、電腦視覺、自動光學檢測、原料價格預測與採購等

導論

- 「線性分類器」(linear classifier)
  - 屬於監督式學習中解決分類問題的方法
  - 線性迴歸解決的是迴歸(regression)問題，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，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。

- 如何建構**線性分類器**預測缺陷的類型
- 如何解釋**特徵**怎麼影響缺陷類型的判別呢？
  - 如何決定分類的「決策邊界」(decision boundary)呢？
  - 可否預測新數據屬於每一類的機率呢？
  - 該如何解釋特徵與分類結果之間的關係呢？
  - 除了二元分類的問題，是否能拓展至多類別的分類呢？

分類問題背景與觀念

鋼板缺陷案例

- 目的：預測鋼板缺陷的類型以建立自動化程序減少人工判斷
- 以兩種缺陷分別為「凹凸不平」與「刮痕」作為二元分類的兩目標變數
- 探討製造兩特徵：「鋼板面積」與「輸送帶長度」

視覺化

- 鋼板面積較大的樣本中刮痕缺陷的佔多數
- 輸送帶長度長的樣本中凹凸不平缺陷的佔多數

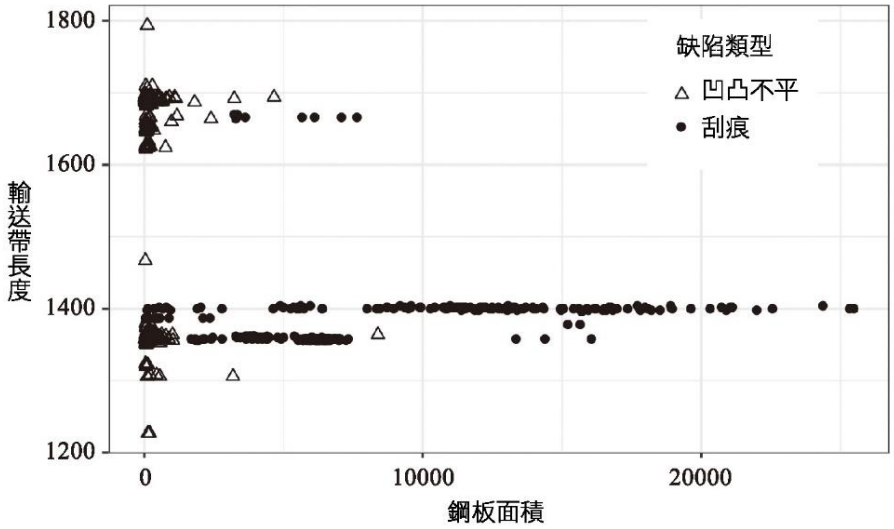


圖 6.1　鋼板數據視覺化

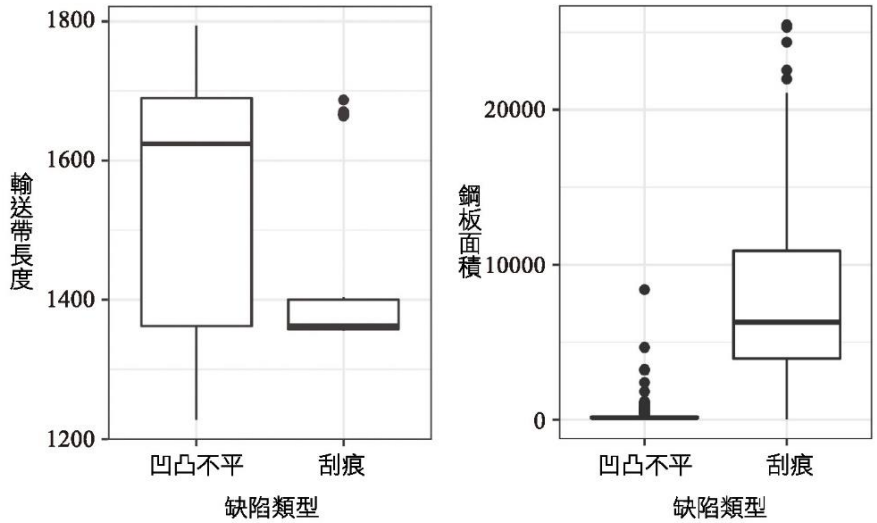


圖 6.2　鋼板數據箱型圖

分類問題背景與觀念

- 「判別模型」(discriminative model)
  - 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」(generative model)
  - 基於目標的不同類別分別建構一個機率分配
  - 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
  - 生成模型也可以求出決策邊界

分類問題背景與觀念

- 若要區分凹凸不平與刮痕兩種缺陷類型
  - 「判別模型」是使用數據直接訓練一個模型，最小化分類誤差，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
  - 「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刮痕的機 率，則我們預測它為凹凸不平的缺陷

- 判別模型v.s生成模型
  - 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，由於它從本質上的最佳化了決策邊界
  - 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。

分類問題背景與觀念

- 二元分類問題
  - 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
- 多類別分類問題(multiclass classification)
  - 生成模型可建構每個類別各自的分配
  - 但在一般的**判別模型**，一次只能決定一個決策邊界，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**

- 多類別的問題轉二元分類的策略有兩種
  - 一對一(one-vs.-one)
    - 將所有類別的兩兩組合各別建構一個判別模型
    - 推廣至K個類別時，需建構$C_2^k = \frac{K(K-1)}{2}$個模型

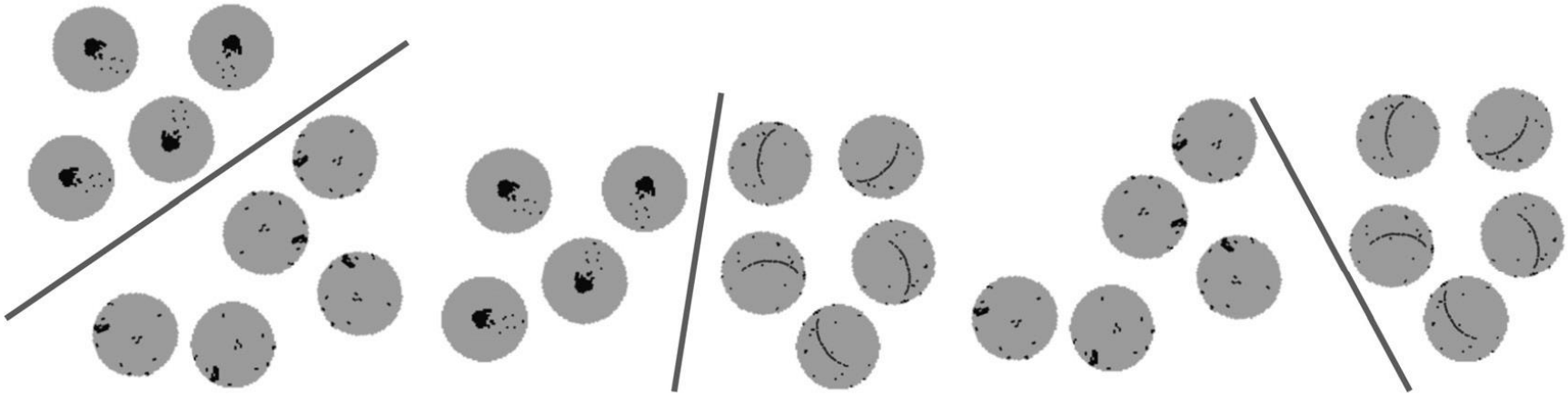
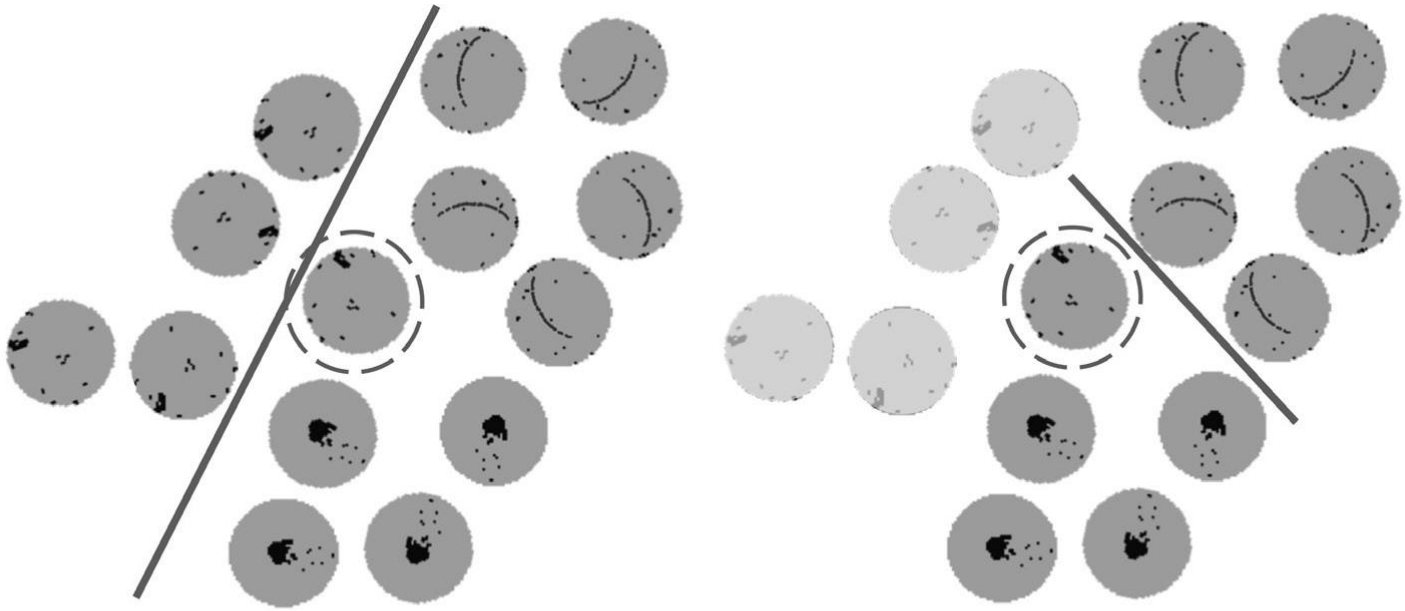


圖 6.4　晶圓圖缺陷二元分類的一對一策略

分類問題背景與觀念

- 多類別的問題轉二元分類的策略有兩種
  - 一對多(one-vs.-all, OvA)
    - 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
    - 因此若推廣至K個類別時，則需建構$K-1$個模型
    - 雖建構的模型個數較少，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)**
    - 預測時，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
    - 部分**非線性模型**的演算法可以直接解多類別分類問題，然而解釋能力較薄弱

- 多類別的問題轉二元分類的策略有兩種
  - 一對多(one-vs.-all, OvA)
    - 第一個分類凹痕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
    - 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



(a)邊緣與非邊緣　　(b)刮痕與非刮痕

圖 6.5　晶圓圖缺陷二元分類的一對剩餘策略

羅吉斯迴歸

- 首先，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
- 在線性迴歸模型中，目標值的範圍是一個介於正負無窮大範圍的實數
- 對於一個分類器來說，羅吉斯迴歸期望其目標值應當落於0與1之間。

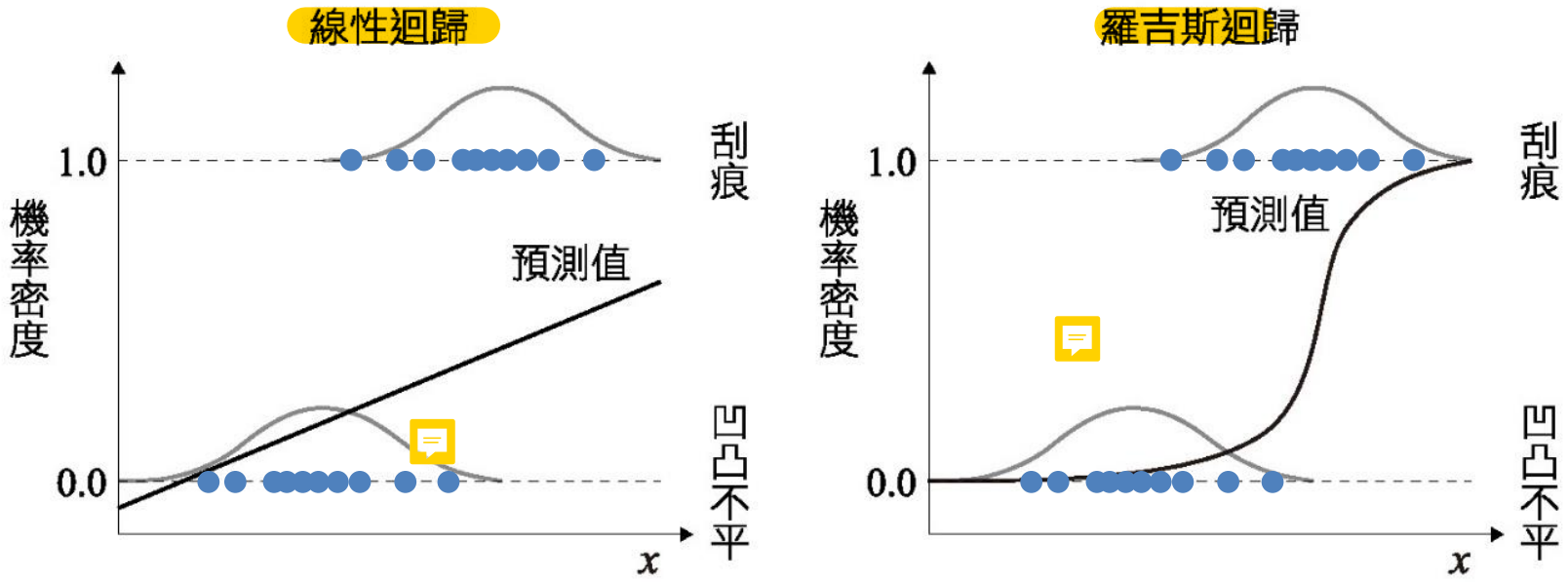


圖 6.6　線性迴歸與羅吉斯迴歸

模型假設與解釋

- 把簡單線性迴歸配適一個目標值為機率的模型用「羅吉斯函數」轉換，使得範圍介於0與1之間。

$$p(x) = \beta_0 + \beta_1 x \quad \Rightarrow \quad p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

- 估計上式的迴歸係數後，可預測出一個合理的機率值 $p(x)$

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

  - 左邊稱之為「勝算」(odds)，計算成功機率是失敗機率的幾倍，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$\frac{0.5}{1-0.5} = 1$。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以上公式將有助於我們對特徵與目標關係的解釋
    - 當 x 增加一單位時，則勝算會是原本的 e^{β_1} 倍

Odds and Odds Ratio (勝算比)

- Production Line
  - RecipeA shows 60 defects from 130 lots
  - RecipeB shows 120 defects from 200 lots

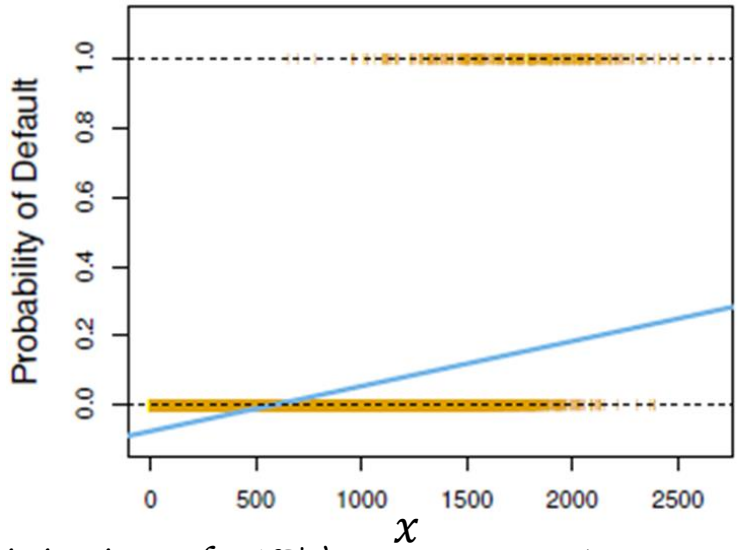
- Odds= $\frac{P}{1-P}$
  - 用RecipeA產生良品的勝算為$\text{Odds}_A = \frac{P}{1-P} = \frac{70/130}{60/130} = 1.167$
  - 用RecipeB產生良品的勝算為$\text{Odds}_B = \frac{P}{1-P} = \frac{80/200}{120/200} = 0.667$

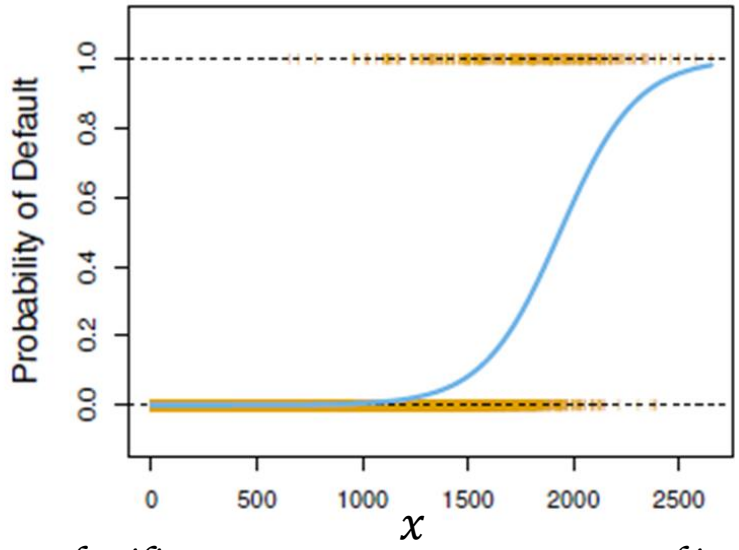
- Odds Ratio (OR, 勝算比)
  - RecipeA比RecipeB勝算比OR= $\frac{\text{Odds}_A}{\text{Odds}_B} = 1.75$

- 「對數勝算」(log-odds或logit)
  - 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，是標準羅吉斯迴歸表示的方式

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

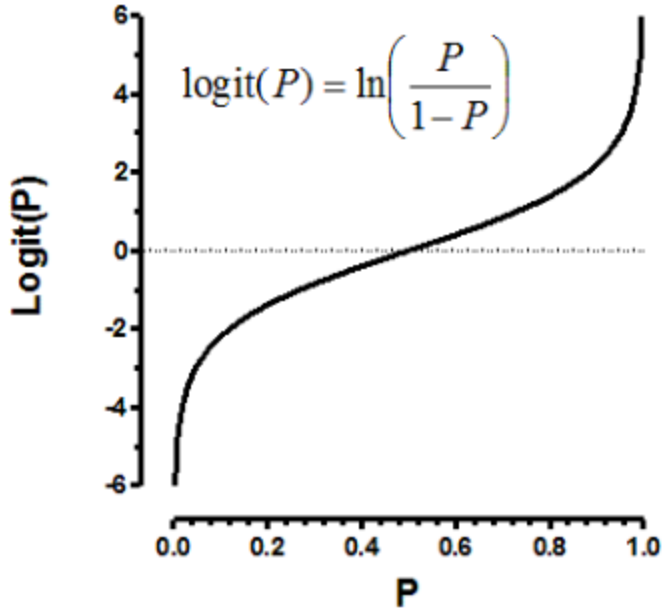
  - This monotone transformation is called the log odds or logit transformation of $p(X)$.
  - Logistic regression ensures that our estimate for $p(X)$ lies between 0 and 1.

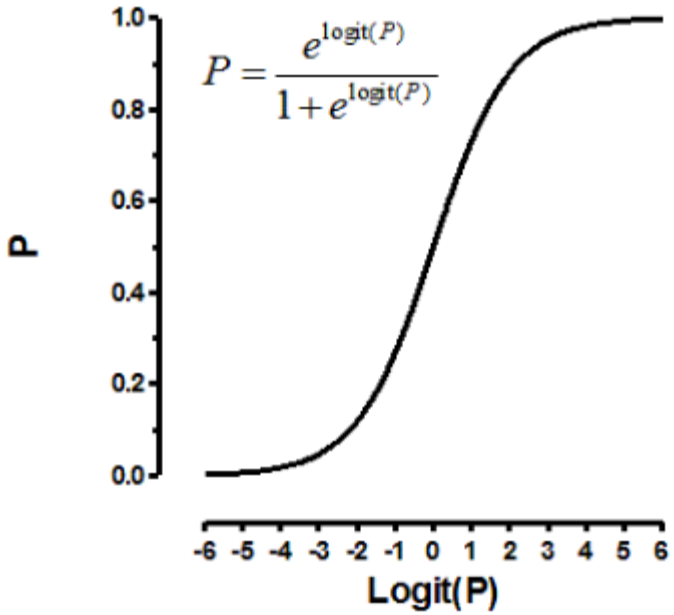


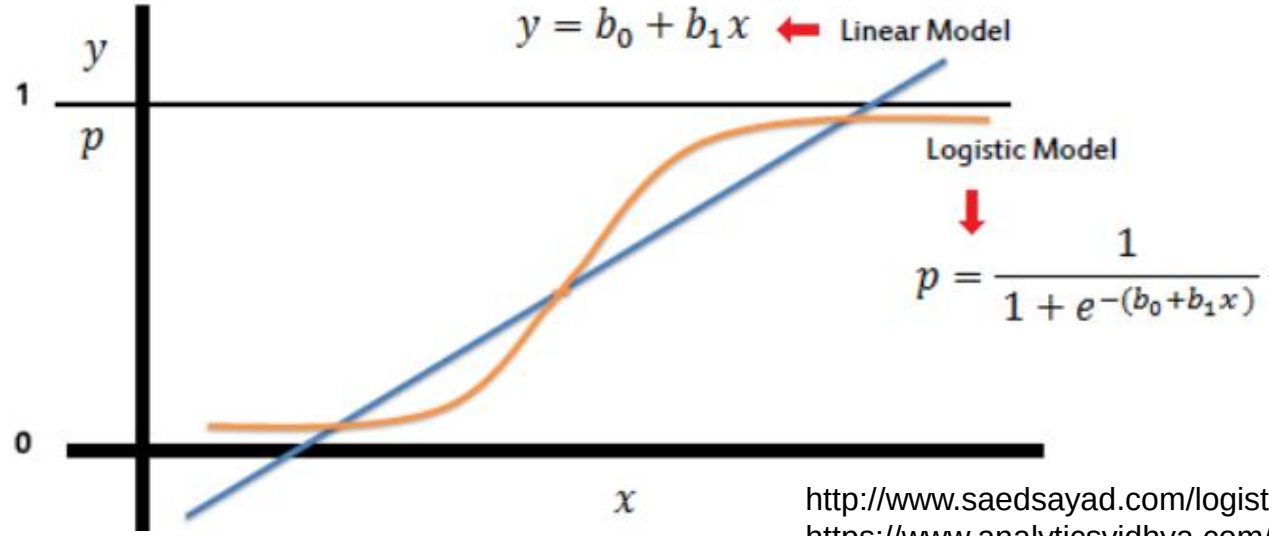


Logit Function

- Logistic regression is an estimate of a logit function.







http://www.saedsayad.com/logistic_regression.htm
https://www.analyticsvidhya.com/blog/2015/10/basics-logistic-regression/

鋼板缺陷案例：模型假設

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_{\text{pixel areas}} + \beta_2 x_{\text{length of conveyer}} \quad (6.7)$$

- 特徵 $x_{pixel\ areas}$ 為鋼板面積
- 特徵 $x_{length\ of\ conveyer}$ 為輸送帶長度
- 目標值中的 $p(x)$ 是缺陷類行為刮痕缺陷的機率
- 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 e^{β_1} 倍
- 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，則刮痕類型缺陷的可能性降低

迴歸係數的估計

- 觀測值的實際目標 y 均為一個0或1值，我們可假設他服從從一個「伯努力分佈」(Bernoulli distribution)

$$\mathcal{L}(p_i) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)} \tag{6.8}$$

- 取對數

$$\begin{aligned} l(p_i) &= \sum_{i=1}^{n} y_i \log p_i + (1-y_i)\log(1-p_i) \\ &= \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \end{aligned} \tag{6.9}$$

- 上式前部分的 $\log\left(\frac{p_i}{1-p_i}\right)$ 可被替換為對數勝算，後部分 p_i 可替換成機率

$$\begin{aligned} l(\beta_0,\beta_1) &= \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) + \log\left(1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right) \\ &= \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) - \log\left(1+e^{\beta_0+\beta_1 x_i}\right) \end{aligned} \tag{6.10}$$

- 改以多元羅吉斯迴歸表示(矩陣形式)
  - 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。

$$l(\beta) = \sum_{i=1}^{n} y_i \mathbf{x}_i^T \beta - \log\left(1 + e^{\mathbf{x}_i^T \beta}\right) \tag{6.11}$$

- 我們對 β 進行偏微分求極值

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta} &= \sum_{i=1}^{n} y_i \mathbf{x}_i^T - \frac{1}{1 + e^{\mathbf{x}_i^T \beta}} \cdot e^{\mathbf{x}_i^T \beta} \cdot \mathbf{x}_i \overset{\text{set}}{=} 0 \\ &\Rightarrow \sum_{i=1}^{n} y_i \mathbf{x}_i^T - p_i \mathbf{x}_i \overset{\text{set}}{=} 0 \\ &\Rightarrow \mathbf{X}^T\left(\mathbf{y} - \mathbf{p}(\beta)\right) \overset{\text{set}}{=} 0 \end{aligned} \tag{6.12}$$

- 無法直接求出函數根(root)以估計係數，需仰賴牛頓法的協助。

羅吉斯迴歸

- 根據牛頓法

$$\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} = \sum_{i=1}^{n} -\mathbf{x}_i p_i(1-p_i)\,\mathbf{x}_i^T \qquad (6.16)$$

$$= -\mathbf{X}^T\mathbf{V}(\beta)\mathbf{X}$$

  - $V(\beta)$ 代表對角矩陣，其主對角元素為 $diag\{p_1(1-p_1),\dots,p_n(1-p_n)\}$

- 以牛頓法表示的多元羅吉斯迴歸係數估計

$$\beta_t = \beta_{t-1} + (\mathbf{X}^T\mathbf{V}(\beta)\mathbf{X})^{-1}\mathbf{X}^T\big(\mathbf{y}-\mathbf{p}(\beta)\big) \qquad (6.17)$$

- 為求迴歸係數的變異，我們基於費雪信息公式以及最大概似估計的漸進常態性公式

$$I(\theta) = -\mathrm{E}[l''(X|\theta)] = -\mathrm{E}\left[\frac{\partial^2 l(\theta)}{\partial\theta^2}\right] \qquad (6.18)$$

$$\sqrt{n}(\hat{\theta}-\theta) \to N\left(0, \frac{1}{I(\theta)}\right) \qquad (6.19)$$

- 可推導出羅吉斯迴歸係數的變異，為費雪信息的倒數

$$\mathrm{Var}(\beta) = \frac{1}{I(\beta)} = \frac{1}{-\mathrm{E}\left[\frac{\partial^2 l(\beta)}{\partial\beta^2}\right]} = (\mathbf{X}^T\mathbf{V}(\beta)\mathbf{X})^{-1} \qquad (6.20)$$

鋼板缺陷案例：估計與檢定

- 由上述最大概似估計法與牛頓法所得出的估計式估計迴歸係數，並依照估計的變異公式進行「統計檢定」，結果如下表

表 6.1　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***

  - 在95%的信心水準下，三迴歸係數的p-value均遠小於0.05，統計上是顯著的可相信他們均不為零
  - 鋼板面積的迴歸係數為正，而輸送帶長度的則為負
  - 代表鋼板面積越大時越可能為刮痕缺陷，每增加一單位，為刮痕缺陷的勝算為原本的 $e^{\beta_1} = 1.001$倍(因此從x軸看決策邊界較陡)
  - 輸送帶長度越長時較可能凹凸不平的缺陷，每增加一單位，為刮痕缺陷的勝算為原本的 $e^{\beta_1} = 0.9922$倍 (因此從y軸看決策邊界較平)

羅吉斯迴歸

- 若將羅吉斯迴歸模型視覺化，圓點為刮痕缺陷，三角形為凹凸不平缺陷
- 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。

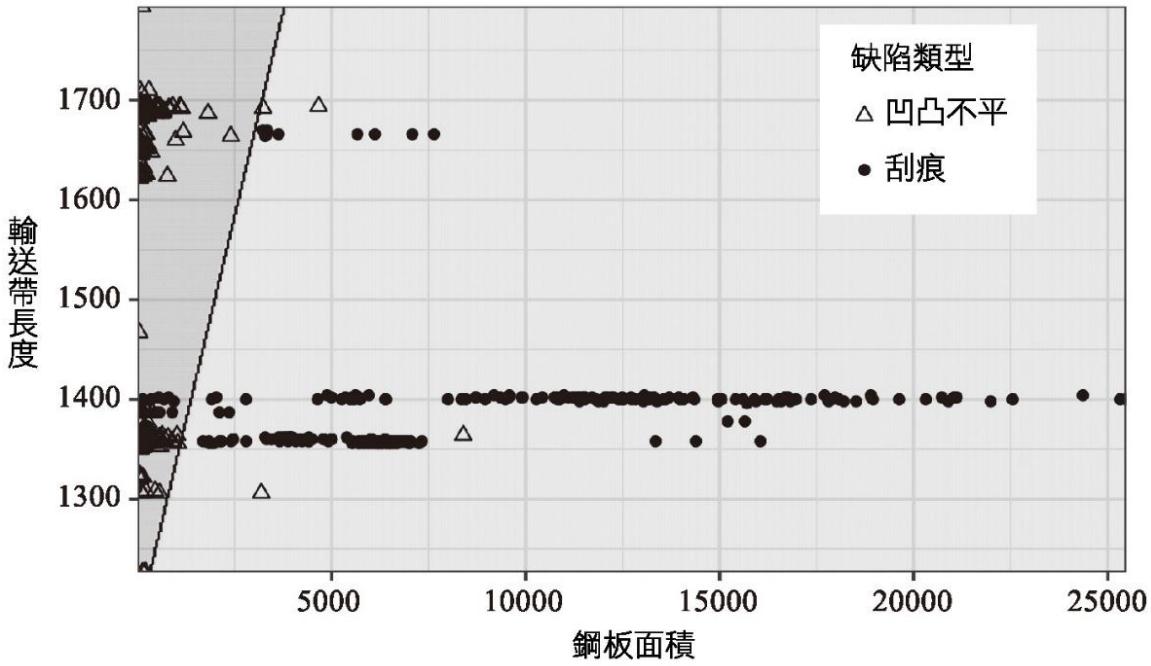


圖 6.8　羅吉斯迴歸於鋼板缺陷分類

拓展至K個類別

- 將迴歸表示為勝算對數的形式，並拓展至K個類別可為：

$$\begin{aligned}
\log\left(\frac{\Pr(Y_i=1)}{\Pr(Y_i=K)}\right) &= \mathbf{x}_i^T\beta_1 \\
\log\left(\frac{\Pr(Y_i=2)}{\Pr(Y_i=K)}\right) &= \mathbf{x}_i^T\beta_2 \\
&\vdots \\
\log\left(\frac{\Pr(Y_i=K-1)}{\Pr(Y_i=K)}\right) &= \mathbf{x}_i^T\beta_{K-1}
\end{aligned} \tag{6.21}$$

  - 需建構 $K-1$ 個模型，有 $K-1$ 組迴歸係數。
  - 每個類別與第K類所建構的模型彼此是分開的。

- 我們將公式替換成1到$K-1$類的機率：

$$\begin{aligned}
\Pr(Y_i=1) &= \Pr(Y_i=K)\,e^{\mathbf{x}_i^T\beta_1} \\
\Pr(Y_i=2) &= \Pr(Y_i=K)\,e^{\mathbf{x}_i^T\beta_2} \\
&\vdots \\
\Pr(Y_i=K-1) &= \Pr(Y_i=K)\,e^{\mathbf{x}_i^T\beta_{K-1}}
\end{aligned} \tag{6.22}$$

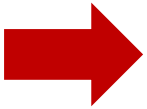
- 因此，第K類的機率會是1減去1到$K-1$類的機率加總：

$$\Pr(Y_i = K) = 1 - \sum_{k=1}^{K-1} \Pr(Y_i = k) = 1 - \sum_{k=1}^{K-1} \Pr(Y_i = K)\, e^{\mathbf{x}_i^T \beta_k} \tag{6.23}$$

$$\Rightarrow \Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\mathbf{x}_i^T \beta_k}}$$

- 有了第K類的機率後，我們可反推1到$K-1$類的機率，可計算出新數據於每一類預測的機率

$$\Pr(Y_i = 1) = \Pr(Y_i = K)\, e^{\mathbf{x}_i^T \beta_1}$$
$$\Pr(Y_i = 2) = \Pr(Y_i = K)\, e^{\mathbf{x}_i^T \beta_2}$$
$$\vdots$$
$$\Pr(Y_i = K-1) = \Pr(Y_i = K)\, e^{\mathbf{x}_i^T \beta_{K-1}}$$



$$\Pr(Y_i = 1) = \frac{e^{\mathbf{x}_i^T \beta_1}}{1 + \sum_{k=1}^{K-1} e^{\mathbf{x}_i^T \beta_k}}$$
$$\Pr(Y_i = 2) = \frac{e^{\mathbf{x}_i^T \beta_2}}{1 + \sum_{k=1}^{K-1} e^{\mathbf{x}_i^T \beta_k}}$$
$$\vdots$$
$$\Pr(Y_i = K-1) = \frac{e^{\mathbf{x}_i^T \beta_{K-1}}}{1 + \sum_{k=1}^{K-1} e^{\mathbf{x}_i^T \beta_k}}$$

- 多類別羅吉斯迴歸所使用的模型只需$K-1$個，因為他可求出每個類別的機率，而非僅僅是決策邊界，為一大優點

廣義線性模型

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**
- 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？

- 鏈結函數(link function)
- 指數族(exponential family)

廣義線性模型

鏈結函數(link function)

- 將廣義線性模型分為兩部分
- 第一部分為**線性預測子**(linear predictor)

$$\eta = \mathbf{X}^T\beta \qquad (6.25)$$

- 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，線性預測子為系統性成分(systematic component)

指數族(exponential family)

- 第二部分我們會對目標變數 Y 做出分配的假設，而這個分配須必須屬於**指數族**(exponential family)
  - 隨機變數的條件期望值為

$$\mathrm{E}\left(Y|X=x\right) = \mu \qquad (6.26)$$

  - 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

鏈結函數

- 廣義線性模型將線性預測子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
- 這種函數稱為**鏈結函數** g:

$$g(\mu) = \mathbf{X}^T\beta \qquad (6.27)$$

- 不同的分配假設則需不同的鏈結函數
- 因此當我們已知目標變數的分配（必須屬於指數族）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

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為

$$f_Y(y;\theta,\tau) = \exp\left\{\frac{y\cdot\theta - b(\theta)}{\tau^2} - c(y,\tau)\right\} \qquad (6.28)$$

  - 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如下：

$$Y \sim N(\mu, \sigma^2)$$

$$\Rightarrow f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} = \exp\left\{\frac{y \cdot \mu - \mu^2/2}{\sigma^2} + C\right\} \quad (6.29)$$

$$= \exp\left\{\frac{y \cdot \theta - \theta^2/2}{\tau^2} + C\right\}$$

  - 其中 $\theta = \mu,\ b(\theta) = \frac{\theta^2}{2},\ \tau^2 = \sigma^2$

- 伯努利分配的機率密度函數以指數族表示如下

$$Y \sim Ber(p)$$

$$\Rightarrow f_Y(y) = p^y(1-p)^{(1-y)} = \exp\left\{y \log\left(\frac{p}{1-p}\right) + \log(1-p)\right\} \quad (6.30)$$

$$= \exp\{y \cdot \theta - \log(1+e^\theta)\}$$

  - 其中 $\theta = \log\left(\frac{p}{1-p}\right), b(\theta) = \log(1+\mathrm{e}^\theta), \tau^2 = 1$

廣義線性模型

- 指數族有以下幾點重要特性
  - 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。
  - 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為

$$\mathrm{E}(Y) = \mu = \frac{\partial b(\theta)}{\partial \theta}$$

$$\mathrm{Var}(Y) = \tau^2 \frac{\partial^2 b(\theta)}{\partial \theta^2} \qquad (6.32)$$

  - 可用最大概似法對廣義線性模型進行估計。
    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

- 因此當我們已知目標變數的分配屬於指數族，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

判別分析(discriminant analysis)

- 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「判別分析」則是先建構在不同類別時特徵的分布 $\Pr(X = x|Y = k)$，假設分布為常態分配
  - 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

1. 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。
2. 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。
3. 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。

樸素貝氏分類器(naïve Bayes)

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
- 樸素貝氏分類器可表示如公式

$$\underbrace{\Pr(Y=k|X=x)}_{\text{posterior}} = \frac{\overbrace{\Pr(X=x|Y=k)}^{\text{likelihood}}\overbrace{\Pr(Y=k)}^{\text{prior}}}{\underbrace{\sum \Pr(X=x)}_{\text{margin}}} \qquad (6.33)$$

  - 其中 $\Pr(Y=k|X=x)$為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)
  - 分母$\sum \Pr(X=x)$是所有可能的加總，在貝氏中稱之為「邊際和」(margin)

樸素貝氏分類器(naïve Bayes)

- 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，通常可由歷史數據得知
- 但概似函數$\Pr(X=x|Y=k)$的計算，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(也可使用無母數方法估計分配)
- 在樸素貝氏分類器中，還會對特徵之間假設彼此相互獨立，故我們可將概似函數拆解為

$$\underbrace{\Pr(X=x|Y=k)}_{\text{likelihood}} = \Pr(X_1, X_2, \ldots, X_p|Y=k) = \Pr(X_1|Y=k)\Pr(X_2|Y=k)\ldots\Pr(X_p|Y=k) \quad (6.34)$$

- 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。

樸素貝氏分類器(naïve Bayes)

樸素貝氏分類器

- 是一條件機率模型。假設一個樣本$(x_1, \dots, x_p)$有p個特徵(隨機變數)
- 若有K個可能的類別C_k ，令機率$p(C_k|\mathbf{x} = x_1, \dots, x_p)$
- 貝氏定理：$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})}$；後驗機率$=\frac{\text{先驗機率}\times\text{概似函數}}{\text{證據}}$
- 樸素假設：所有的特徵**彼此獨立**(強假設)
  - i.e. $p(x_i|x_{i+1}, \dots, x_p, C_k) = p(x_i|C_k)$
- 則$p(C_k|x_1, \dots, x_p) \propto p(C_k, x_1, \dots, x_p) \propto p(C_k)p(x_1|C_k)p(x_2|C_k)p(x_3|C_k) \dots$
  $\propto p(C_k)\prod_{i=1}^{p} p((x_i|C_k)$
- $p(C_k|x_1, \dots, x_p) = \frac{1}{Z}p(C_k)\prod_{i=1}^{p} p((x_i|C_k)$，其中正規化因子$Z = p(\mathbf{x}) = \sum_k p(C_k)p(\mathbf{x}|C_k)$為已知的證據是一常數
- 分類器：class label $\hat{y} = C_k = arg \max_{k\in\{1,\dots,K\}} p(C_k)\prod_{i=1}^{p} p((x_i|C_k)$

https://mropengate.blogspot.com/2015/06/ai-ch14-3-naive-bayes-classifier.html

Application of Bayes' Theorem

- In NBA game, when the Lakers joins the final game, the fans of Lakers estimate
  - The probability of the Lakers wins (H_1) the champion: $P(H_1)=0.7$
  - The probability of the Lakers loses (H_2) the champion: $P(H_2)=0.3$
- Now, the Lakers loses the first game (the evidence you observed), what is the probability to win the champion?
- Based on the historical records, in the NBA final game…
  - The probability of losing-first-game (E) given it wins the champion is $P(E|H_1)=0.4$
  - The probability of losing-first-game (E) given it loses the champion is $P(E|H_2)=0.5$

Bayes' Theorem

Probability Correction

- What is the posterior probability $P(H_1|E)$

- P(wins champion | losing-first-game)

$$= \frac{P(E \mid H_1)P(H_1)}{P(E \mid H_1)P(H_1) + P(E \mid H_2)P(H_2)}$$

$$= \frac{2/5 * 0.7}{2/5 * 0.7 + 1/2 * 0.3}$$

$$= 0.65$$

- Thus, after you observe the "evidence", the probability is corrected from 0.7 to 0.65.

樸素貝氏分類器(naïve Bayes)

樸素貝氏分類器

- 根據過去歷史數據
- 若今有一新樣本(溫度=中、壓力=低、濃度=高、轉速=中)
  - P(良)P(溫度中|良)P(壓力低|良)P(濃度高|良)P(轉速中|良)
  - $= \frac{6}{10} \times \frac{1}{6} \times \frac{4}{6} \times \frac{3}{6} \times \frac{1}{6} = \frac{1}{180} = 0.00556$
  - P(不良)P(溫度中|不良)P(壓力低|不良)P(濃度高|不良)P(轉速中|不良)
  - $= \frac{4}{10} \times \frac{3}{4} \times \frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{3}{160} = 0.01875$
- 因此，預測會產生「不良品」。
- 但....
- 若以kNN的視角來說...
- 可能預測會產生「良品」。

	溫度	壓力	濃度	轉速	檢測
1	低	高	中	低	不良
2	低	低	中	低	良
3	低	高	高	低	良
4	高	高	高	高	良
5	中	高	低	低	不良
6	中	高	高	中	不良
7	高	低	低	低	良
8	中	低	高	中	良
9	中	低	高	高	不良
10	低	低	中	低	良

https://mropengate.blogspot.com/2015/06/ai-ch14-3-naive-bayes-classifier.html

鋼板缺陷案例：樸素貝氏分類器

- 對鋼板缺陷類型與兩特徵使用樸素貝氏分類器
- 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)
- 可以看到凹凸不平所建構的無母數分配為非線性，雖此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。

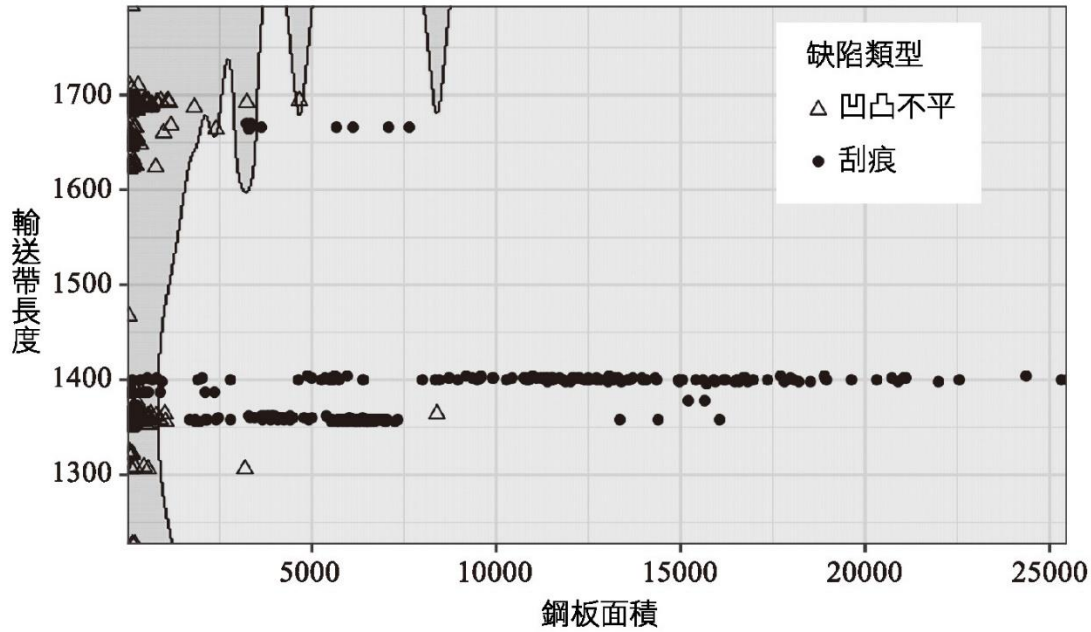


圖 6.9　樸素貝氏分類器於鋼板缺陷分類

線性判別分析

- 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
- 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，可找出每個樣本於不同類的機率，以及類別之間的決策邊界

一個特徵的線性判別分析公式

- 先將樸素貝氏分類器簡化為

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum \pi_k f_k(x)} \qquad (6.35)$$

- 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)
- $p_k(x)$ 為基於特徵下每個樣本於不同類的機率 (後驗機率)

線性判別分析

- 一個維度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

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{\frac{-(x-\mu_k)^2}{2\sigma_k^2}\right\} \qquad (6.36)$$

  - μ_k 與 σ_k^2 為該特徵在第 k 類的平均數與變異數。
  - 假設特徵在所有類別之間的變異數是相等的，如 $\sigma_1^2 = \cdots = \sigma_k^2 \overset{set}{=} \sigma^2$ (線性判別分析的強假設)

- 把常態分配pdf公式代入樸素貝氏分類器公式的概似函數後，可得到公式

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu_k)^2}{2\sigma^2}\right\}}{\sum \pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu_k)^2}{2\sigma^2}\right\}} \qquad (6.37)$$

  - 取對數後得到「**判別函數**」(discriminant function)

$$\delta_k(x) = x^2 \cdot \frac{-1}{2\sigma^2} + x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k - C \qquad (6.38)$$

  - 其中 C 為一常數，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 取對數後的機率

線性判別分析

- 假設有兩個類別分別為k與l，兩類間的決策邊界是兩個類別機率相等的地方

$$\delta_k(x) \overset{\text{set}}{=} \delta_l(x)$$

$$\Rightarrow \left(x^2 \cdot \frac{-1}{2\sigma^2} + x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k - C\right) -$$

$$\left(x^2 \cdot \frac{-1}{2\sigma^2} + x \cdot \frac{\mu_l}{\sigma^2} - \frac{\mu_l^2}{2\sigma^2} + \log \pi_l - C\right) \overset{\text{set}}{=} 0 \qquad (6.39)$$

$$\Rightarrow x \cdot \frac{\mu_k - \mu_l}{\sigma^2} - \frac{\mu_k^2 - \mu_l^2}{2\sigma^2} + (\log \pi_k - \log \pi_l) \overset{\text{set}}{=} 0$$

  - 決策邊界為特徵x的一次線性函數，故線性判別分析與羅吉斯迴歸同屬線性分類器
  - 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

$$x_{db} = \frac{\mu_k^2 - \mu_l^2}{2(\mu_k - \mu_l)} = \frac{\mu_k + \mu_l}{2} \qquad (6.40)$$

- 隨機生成來自常態分配的兩個類別
  - 母體參數為平均數 $\mu_1 = -1.25,\ \mu_2 = 1.25$
  - 變異數 $\sigma_1^2 = \sigma_2^2 = 1$。
  - 套用決策邊界點公式，理論上邊界點為圖中虛線 $x_{boundary} = \frac{-1.25+1.25}{2} = 0$

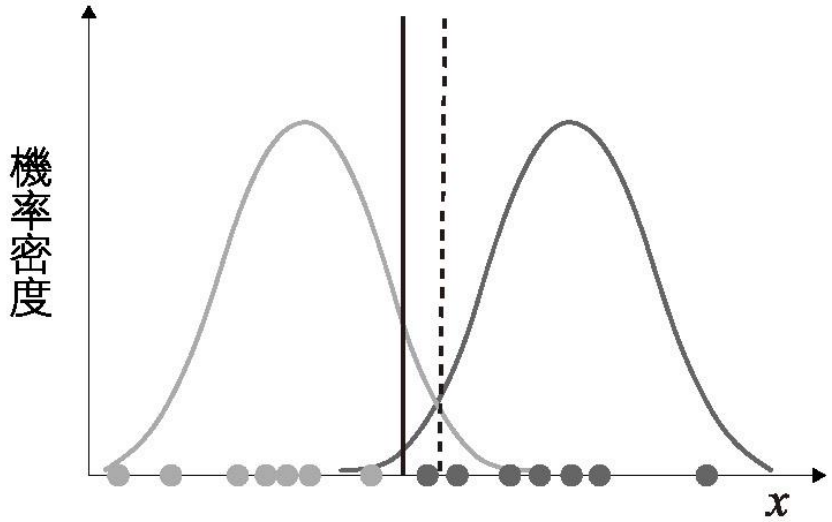


圖 6.10　線性判別分析

線性判別分析

- 在實務上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**為：

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \qquad (6.41)$$

  - n_k 為第 k 類的樣本數，n 為總樣本數
  - 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

$$\hat{\pi}_k = \frac{n_k}{n} \qquad (6.42)$$

- 線性判別分析假設類別間特徵的變異數相等下，可將線性判別分析分類器 $\hat{G}_{lda}(x)$ 可由判別函數公式引入估計值公式

$$\hat{G}_{lda}(x) = \underset{k}{\operatorname{argmax}}\, \hat{\delta}_k(x)$$

$$= \underset{k}{\operatorname{argmax}}\, x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k \qquad (6.43)$$

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

多個特徵與多個類別的線性判別分析

- 將一維度的特徵 X 拓展到多維度，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。
- 多變量常態分配可簡化表示為 $X \sim N(\mu, \Sigma)$
  - 平均數 $\mu = E(X)$ 為 p 個維度的向量
  - 共變異矩陣(covariance matrix) Σ=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為

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right\} \qquad (6.44)$$

- 假設不同類別之間的共變異矩陣是相等的 $\Sigma_1 = \ldots = \Sigma_k \overset{set}{=} \Sigma$
- 把多變量常態分配帶入貝氏分類器得到判別函數$\delta_k(x)$

$$\delta_k(x) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k - C \qquad (6.45)$$

  - 當我們得到某觀測值$X = x$代入上式，可得到每個類別k各別取對數後的機率

線性判別分析

- 假設有兩個類別分別為 k 與 l，決策邊界即是兩個類別機率相等的地方

$$x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k \overset{\text{set}}{=} x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + \log\pi_l \qquad (6.46)$$

- 因此，線性判別分析分類器 $\hat{G}_{lda}(x)$ 可表示為

$$\begin{aligned}\hat{G}_{lda}(x) &= \arg\max_k \hat{\delta}_k(x) \\ &= \arg\max_k x^T\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + \log\hat{\pi}_k\end{aligned} \qquad (6.47)$$

- 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。

- 若隨機生成來自多變量常態分配的三個類別
  - 其母體參數為平均數$\mu_1 = (-1 - 1),\ \mu_2 = (3,1),\ \mu_3 = (1,3)$
  - 共變異矩陣於兩特徵間為一個大於零的常數且$\Sigma_1 = \Sigma_2 = \Sigma_3$
  - 每個橢圓為各類別的多變量常態分配等高線，套用決策邊界點公式即可找出實際的決策邊界(虛線)
  - 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。

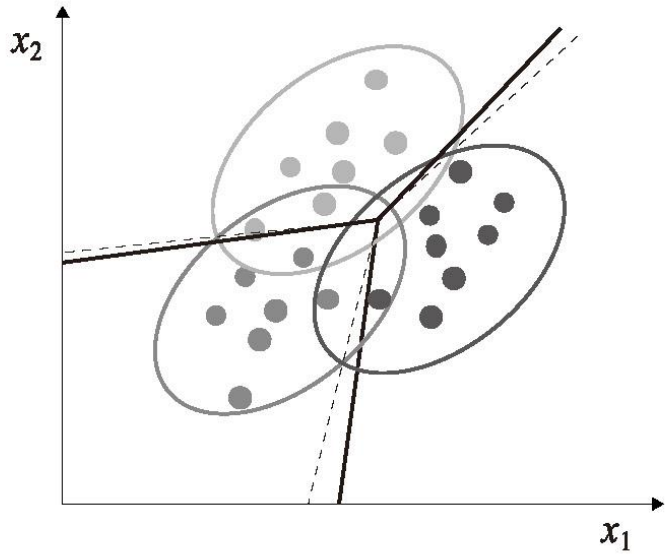


圖 6.11　三個類別的二元變量常態分配

鋼板缺陷案例：線性判別分析

- 對鋼板缺陷類型與兩特徵使用線性判別分析分類器，結果如下表

表 6.2　鋼板缺陷數據的先驗機率

	凹凸不平	刮痕
機率	0.508	0.492

表 6.3　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 6.4　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

- 表6.2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(6.42)。
- 表6.3與表6.4為兩種缺陷各自多變量常態分配所估計的平均數$\hat{\mu}_k$與共變異矩陣$\hat{\Sigma}_k$，估計的計算為公式(6.41)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(6.46)與分類器如公式(6.47)。

視覺化

- 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
- 橢圓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
- 與羅吉斯迴歸模型(圖6.9)比較，呈現了線性分類器中判別模型與生成模型的差異

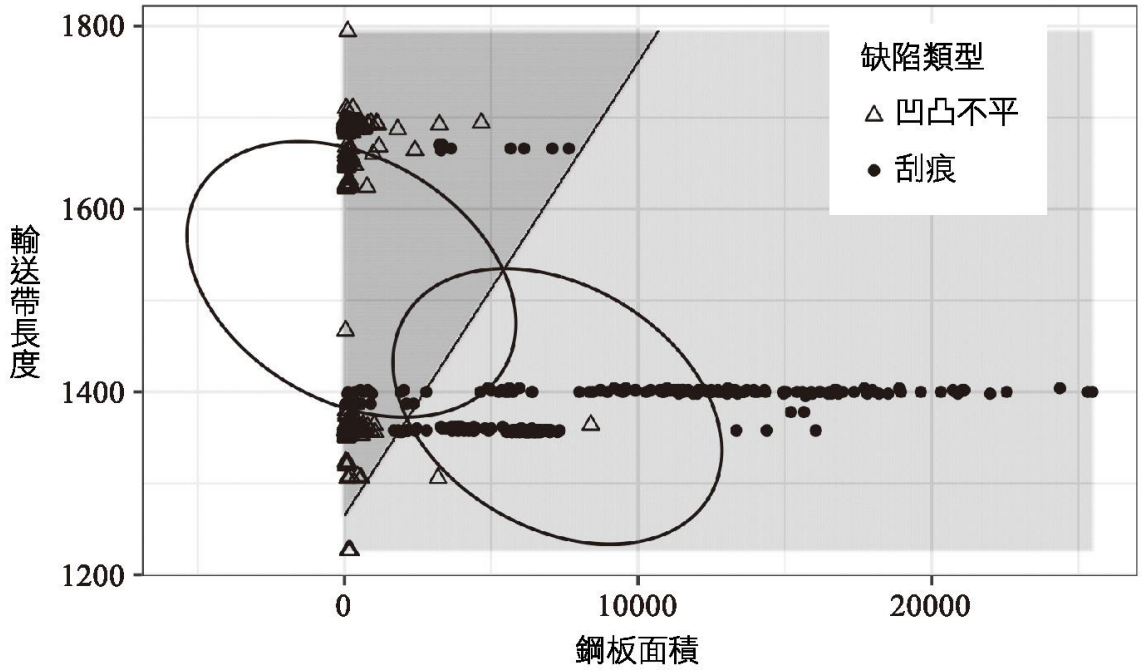


圖 6.12　線性判別分析於鋼板缺陷分類

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。
- 這與實際的數據經常相去甚遠。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。
- 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
- 原判別函數$\delta_k(x)$可改寫為

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k - C \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| \\ &\quad + \log\pi_k - C\end{aligned} \tag{6.48}$$

二次判別分析

- 假設有兩個類別分別為 k 與 l，決策邊界即是兩個類別機率相等的地方

$$\begin{aligned} & -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \\ \overset{\text{set}}{=} & -\frac{1}{2}x^T\Sigma_l^{-1}x + x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l - \frac{1}{2}\log|\Sigma_l| + \log\pi_l \end{aligned} \tag{6.49}$$

  - 與「線性判別分析」公式(6.46)的差異在特徵的二次項，由於不等變異的假設，無法相互抵消
  - 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

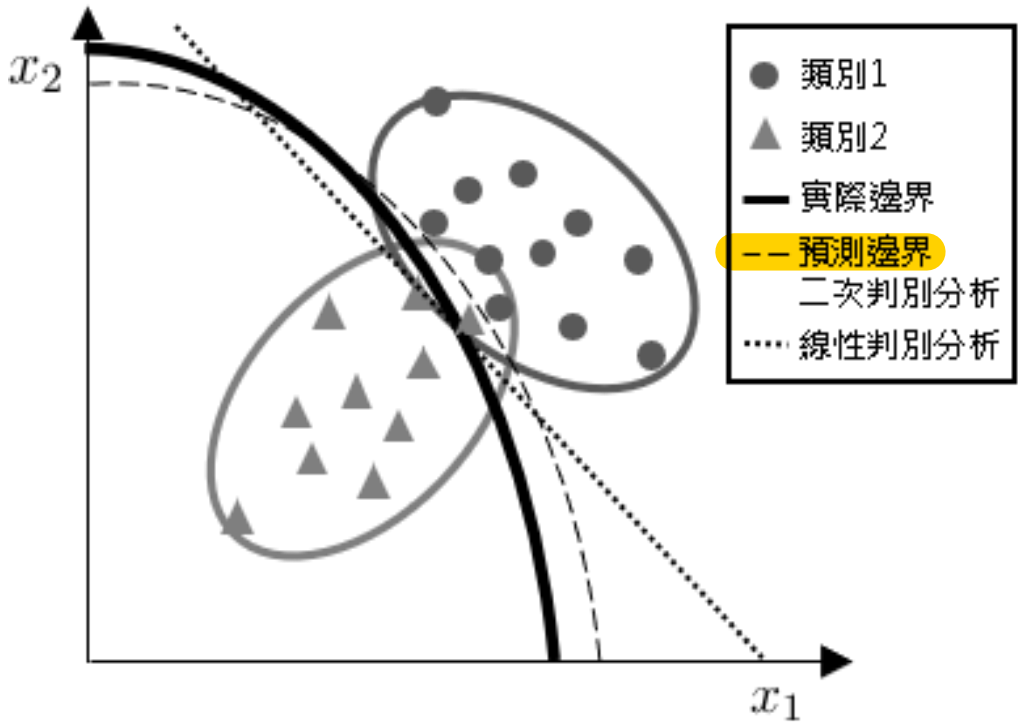
- 二次判別分析分類器 $\hat{G}_{lda}(x)$ 可表示為

$$\begin{aligned} \hat{G}_{lda}(x) &= \arg\max_k \hat{\delta}_k(x) \\ &= \arg\max_k -\frac{1}{2}x^T\hat{\Sigma}_k^{-1}x + x^T\hat{\Sigma}_k^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}_k^{-1}\hat{\mu}_k + \log\hat{\pi}_k \end{aligned} \tag{6.50}$$

二次判別分析

以二維分類問題舉例說明

- 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。



二次判別分析

鋼板缺陷案例

- 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6.12）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6.8）。

鋼板缺陷案例

表 6.5　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 6.6　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

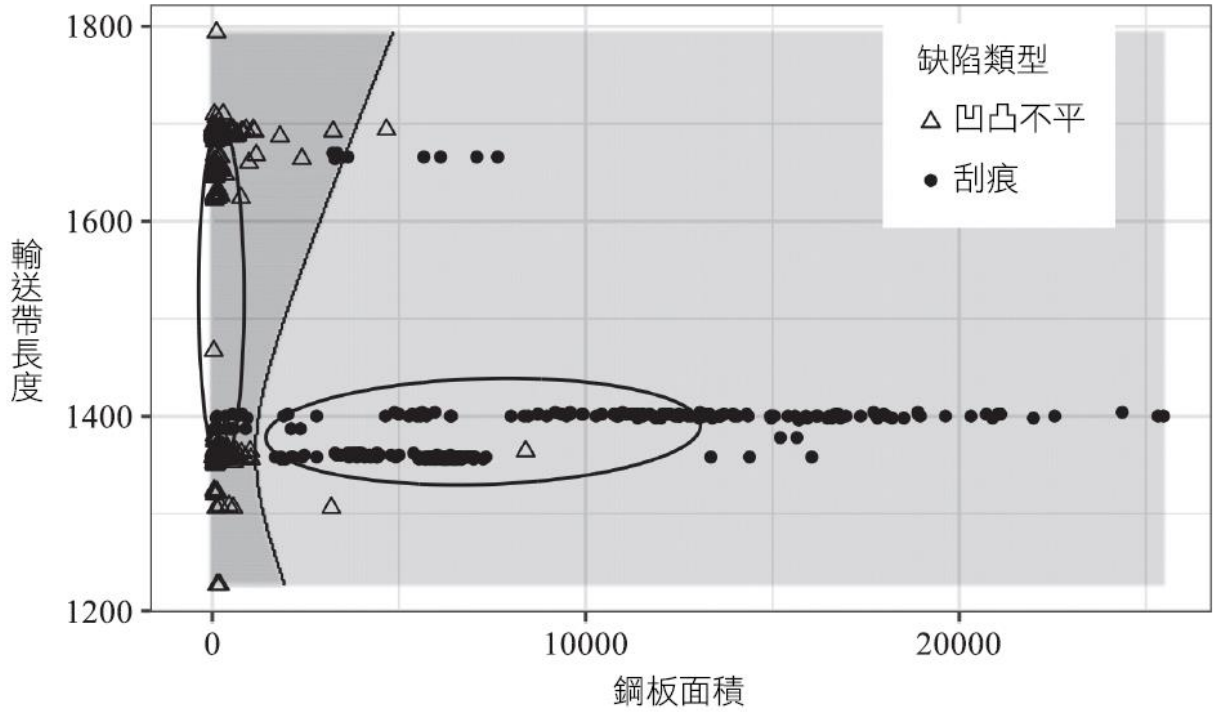


圖 6.14　二次判別分析於鋼板缺陷分類

結語

- 「羅吉斯迴歸」(判別模型)與「線性判別分析」(生成模型)
  - 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，但有不錯的解釋能力
  - 對於決策邊界的呈現相對清楚，能提供數學式來說明其變數間影響的關係與解釋
- 線性分類器是機器學習於分類問題的基礎方法，延伸出許多應用
  - 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)
  - 近幾年流行的集成學習(ensemble learning)，常以弱分類器(weak classifier)(如線性分類器)為基礎模型(base model)，用以建構強分類器
  - 深度學習中生成模型的觀念與基礎，也可從線性判別分析中窺知一二
- 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

Thanks for your attention

NTU Dept. of Information Management
name: 李家岩 (FB: Chia-Yen Lee)
phone: 886-2-33661206
email: chiayenlee@ntu.edu.tw
web: https://polab.im.ntu.edu.tw/