前言

本書的目標讀者包括：

- 將與資料科學家合作、管理資料科學導向專案，或是將投資於資料科學新興企業的商業人士
- 將實行資料科學解決方案的開發人員
- 積極求知的資料科學家

這不是一本介紹演算法的書，也不打算取代介紹演算法的書。我們刻意避開了以演算法為中心的做法。我們相信，在從資料中抽取有用知識的技術背後，存在有相對更精簡的一組基本概念或原則。這些概念是許多知名資料採礦演算法的基礎，也是以數據為中心之商業問題分析、資料科學解決方案之創立與評估，還有一般資料科學策略及提案之評估等的基礎。因此，我們的說明都圍繞著這些一般性的原則，而非特定的演算法。在需要解說程序細節時，本書採取以文字搭配圖表的方式呈現，因為我們認為這樣會比詳盡的演算步驟條列更容易理解。

閱讀本書不需要有高深的數學背景，但此書本身還是有一定程度的技術性質－目標是要讓讀者確實理解資料科學，而不是只提供整體概述。基本上我們已盡量努力縮減數學部分，並使說明內容盡可能「概念化」。

同行們都說，這本書非常寶貴，可協助經營、技術 / 開發及資料科學等團隊達成一致的理解。不過此觀察結果是基於相當小的樣本，所以我們很好奇地想知道這實際上有多普遍（請見第 5 章）。理想上，我們希望這是一本會讓每個資料科學家想遞給來自開發或經營團隊的合作夥伴的書，以藉此有效表達：若你真的想針對商業上的問題，設計 / 實行頂尖的資料科學解決方案，我們就需要對此題材有共通的理解。

同行們還說，此書有個意料之外的用途，那就是：可用於為面試資料科學工作應徵者做準備。企業對雇用資料科學家的需求相當強烈，且與日俱增。因應此現況，也有越來越多的求職者以資料科學家自居。而每個資料科學工作的應徵者都該了解本書所介紹的基礎知識（我們的業界同仁透露，他們很驚訝地發現其實很多人都不懂這些基礎。因此，我們還曾半開玩笑地討論要再出一本「資料科學工作面試手冊」呢！）。

我們的概念式資料科學教學

本書將介紹一系列最重要的資料科學基礎概念。這些概念有些就直接列成了「章名」，有些則是隨著內文討論自然而然地提到（因此不見得都歸類為基本概念）。這些概念橫跨從預想問題，到應用資料科學技術，再到運用結果來改善決策的整個程序。而這些概念也支援、鞏固了眾多商業分析的方法與技術，

它們可被分為三大類：

1. 關於資料科學如何融入組織與競爭環境的概念，包括吸引、建構、培育資料科學團隊的方法；思考資料科學如何能帶來競爭優勢的方法；還有做好資料科學專案的策略概念。
2. 一般的數據分析性思考方法。這些方法有助於確認合適的數據並考慮合適的方法。此類概念包括資料採礦程序，以及一系列不同的高層次資料採礦任務。
3. 實際從資料數據抽取知識的一般概念，而此類概念鞏固了眾多的資料科學任務及其演算法。

例如，有個基本概念是確定資料所描述的兩個實體的相似性，而此能力構成了各種具體任務的基礎，像是可直接用於找出與已知顧客類似的顧客群。這構成了幾種預測演算法（評估某一目標值，像是顧客的預期資源使用狀況，或是顧客對報價做出回應的機率）的核心，同時也是聚類技術（依實體的共同特徵來分組，沒有單一的焦點目標）的基礎。相似性是資訊檢索（與查詢內容有關的文件或網頁會被擷取出來）的根基，另外還支撐了幾種常見的推薦演算法。傳統的演算法導向書籍可能會將這些分別放在不同的章節、以不同的名稱來解說，使其共同點就此淹沒於演算法細節或數學命題之中。本書則是專注於統一的概念，將特定任務及演算法呈現為其自然的表現形式。

特徵來描述的每個既有顧客，將這些特徵輸入至模型 M 後，模型 M 便會產生一個流失得分或流失機率的評估值。這就是資料採礦結果的運用。而資料採礦是從某些其他的、通常是過去的資料來產生、建立出模型 M。

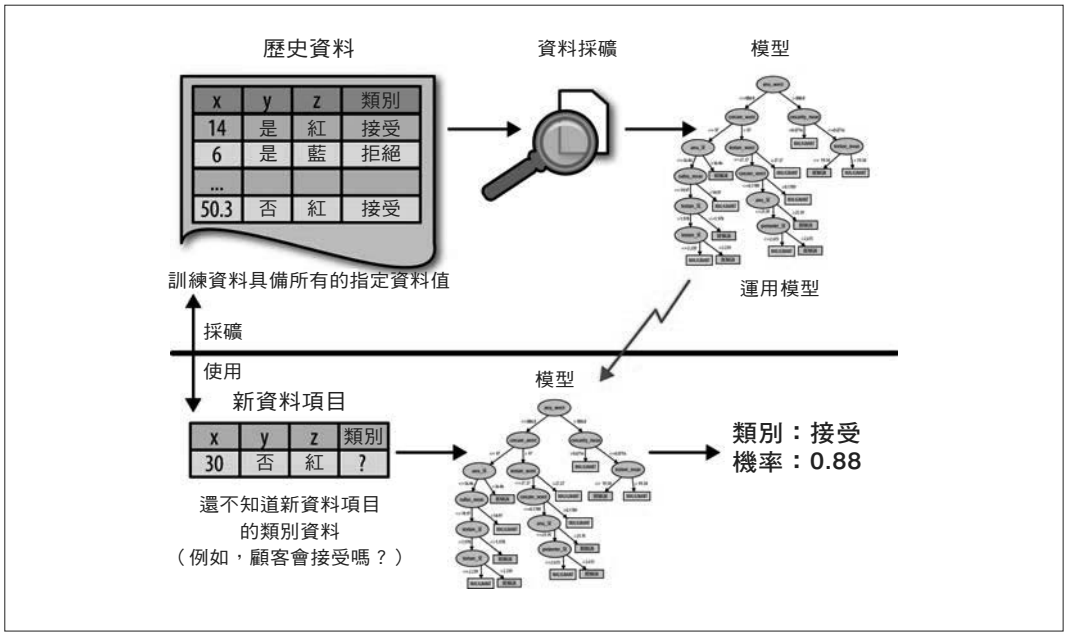

圖 2-1　資料採礦與資料採礦結果的運用。圖的上半部說明了挖掘歷史資料以產生模型的過程。其中重要的是，歷史資料擁有指定的目標值（「類別」）。圖的下半部則顯示了資料採礦結果的運用，將該結果模型用於還不知道類別值的新資料，然後模型便預測出其類別值，以及該類別變數成為該值的機率。

圖 2-1 說明了這兩個階段。資料採礦產生出機率評估模型，如圖的上半部所示。而在運用階段（圖的下半部），將該模型應用於新的、沒見過的例子，模型便會為此例生成機率評估。

資料採礦程序

資料採礦是一種工藝。它包含了科學與技術的大量應用，不過要運用得當仍需要一點藝術。就和許多成熟的工藝一樣，它有容易理解的程序可將問題結構化，達成合理的一致

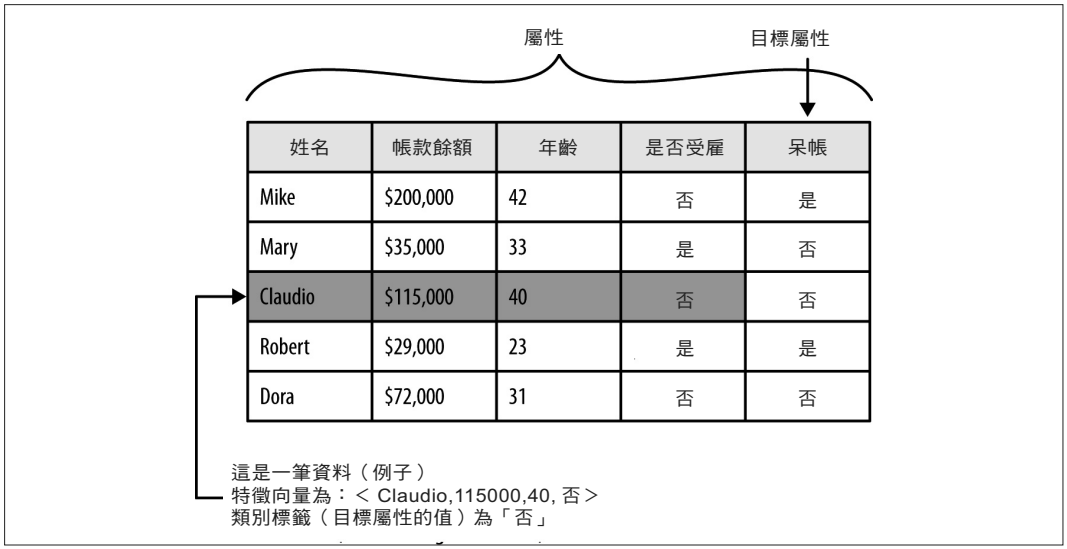

圖 3-1　監督式分類問題的資料採礦專用術語。由於此問題具有目標屬性，以及一些包含目標屬性值的「訓練」資料（training data），故屬於監督式問題。這是個分類（而非迴歸）型的問題，因為其目標是類別（是或否）而非數字。

在資料科學中，預測性模型是用來評估所關注之未知值（亦即目標）的公式。此公式可能是數學公式，也可能是一種邏輯敘述，例如某種規則。通常都是混合了兩者。由於我們把監督式資料採礦分成了分類和迴歸兩類，所以我們將分別考慮分類模型（還有類別機率評估模型）與迴歸模型。

專用術語：預測（Prediction）

在一般常見的用法裡，預測就是指預報未來事件。不過在資料科學中，預測更常用來指稱評估某個未知值。此值未來可能會是某個值（在一般常見的用法裡，真正的預測），但它在現在或過去也可能是某個值。實際上，由於資料採礦處理的通常都是歷史資料，故模型幾乎都是以過去的事件來構建並測試。信用評分的預測性模型，是用來評估潛在顧客違約（拖欠不繳帳單而變成呆帳）的可能性。垃圾郵件過濾的預測性模型，是用來評估特定電子郵件是否為垃圾郵件。詐欺檢測的預測性模型，是用來判斷某帳戶是否已被詐騙。因此關鍵就在於，此種模型是要用來評估某未知值的。

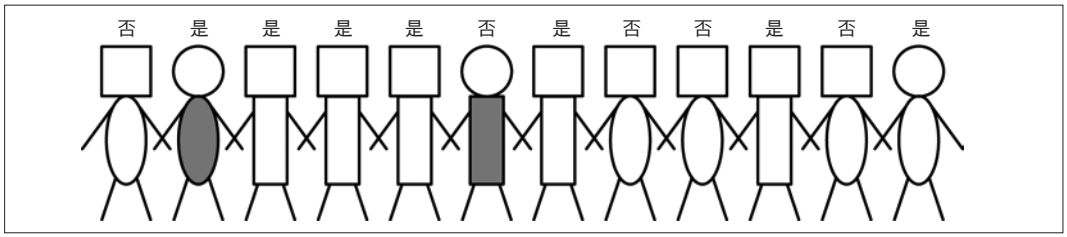

圖 3-2　一組待分類的人。每個人頭上的標籤（是 / 否）代表其目標變數值（是否變成呆帳），而顏色和形狀則代表不同的預測屬性。

我們現在要來仔細研究一種選擇有用變數的方法，然後會示範我們能夠如何反覆地運用這種技術來建立監督式區隔。雖然此方法非常實用又具示例性，但你還是要記住，直接且多變量的監督式區隔只是這個選擇有用變數之基礎觀念的一種應用而已。此觀念應成為你在思考資料科學問題時，更一般、普遍的概念工具之一。例如在接下來的討論過程中，我們也會深入探討其他的、不直接包含變數選擇的建模方法。在真實世界中面對非常大量的屬性時，回想這個早期觀念並選出一組有用的屬性，可能會對你非常有幫助。這麼做可大大縮減原本笨重、龐大的資料集，而且就如我們即將看到的，往往能夠提升成果模型的準確度。

選出有用的屬性

在有一大堆實例的情況下，我們如何選出一個屬性，好對它們做出有意義、有用的分割、分群呢？讓我們來考慮一個二分式（分成兩類）的分類問題，並且想想看我們希望從中得到什麼。舉個具體的例子，圖 3-2 是個簡單的區隔問題：以簡單的輪廓代表十二個人。其中頭有兩種類型，方的和圓的；身體也有兩種類型，長方形和橢圓形；另外有兩個人的身體是灰色的，其他都是白色的。

這些就是我們將用來描述這批人的屬性。寫在每個人頭上的是二分式的目標標籤，「是」或「否」，代表了此人是否會不繳帳單而變成呆帳。我們可將這些人的資料描述為：

- 屬性：
  - －頭的形狀：方形、圓形
  - －身體的形狀：長方形、橢圓形
  - －身體的顏色：灰色、白色
- 目標變數：
  - －呆帳：是、否

種建立的是所謂的「支援向量機（support vector machine）」，在以一個較簡單的目標函數為例具體解說後，我們會再回頭多介紹一下這個支援向量機。然後我們會簡短地討論一下迴歸（而非分類）的線性模型，最後則以最實用的資料採礦技術之一的邏輯迴歸（logistic regression）作結。邏輯迴歸這個名稱其實不是很恰當，因為它並沒有真的做我們所謂的迴歸，也就是數值目標變數的評估。邏輯迴歸將線性模型應用於類別機率評估，對許多方面的應用來說都特別有用。

線性迴歸、邏輯迴歸，以及支援向量機都是非常類似的、將（線性）模型配適於資料的基本技術實例。其關鍵差異在於，他們所用的目標函數各自不同。

從資料取得線性判別的範例

我們要利用一批改編過的鳶尾花資料集（*http://archive.ics.uci.edu/ml/datasets/Iris*）（取自加州大學爾灣分校的機器學習庫（*http://archive.ics.uci.edu/ml/*），Bache & Lichman，2013），來解說線性判別函數。這是個相當古老且單純的資料集，記錄各種類型的鳶尾花（鳶尾屬的開花植物）。原始的資料集包含呈現出四種屬性的三種鳶尾花，而其資料採礦問題是要依據這四種屬性，來將各實體分類至三種品種中的一種。

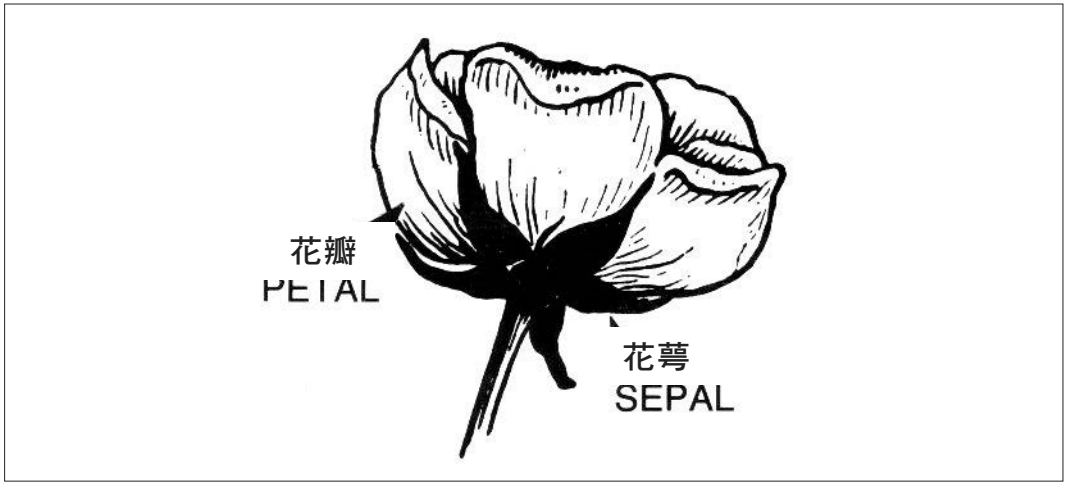

圖 4-6　花的兩個部分。鳶尾花資料集中包含了這些部分的寬度測量值。

在此我們只用兩種鳶尾花，山鳶尾（Iris Setosa）和變色鳶尾（Iris Versicolor），來說明。此資料集正是這兩個品種的花卉集合，每筆資料都描述了兩個測量值：花瓣的寬度

界的資料都是如此，有些資料點無可避免地會被模型錯誤分類。這對一般的線性判別觀念來說並不構成問題，因為它們不見得必須要正確分類每一個資料點。然而這樣一來，在將線性函數配適於資料時，我們就無法簡單地問：我們該選擇所有線條中哪一條能夠完美分隔資料的線？畢竟可能根本不存在這種能完美分隔的線！

這再次突顯了支援向量機解決方案是多麼地直覺又令人滿意。跳過數學部分，其觀念如下：在測量特定模型與訓練資料的配適（擬合）程度的目標函數中，對於落在決策邊界錯誤側的資料點，我們就是單純地予以懲罰。在資料確實可被線性分隔的情況下，我們無須做懲罰，只要最大化邊緣即可。而若資料無法被線性分隔，那麼最佳配適便是在最粗邊緣與較低的總誤差懲罰之間的某個平衡點。對錯誤分類資料點的懲罰，會與該資料點和決策邊界的距離成比例，故 SVM 只可能犯「小」錯。在技術上，這種誤差函數被稱做「合頁損失（hinge loss）」（詳見本章的「補充說明：損失函數（loss function）」，以及圖 4-9）。

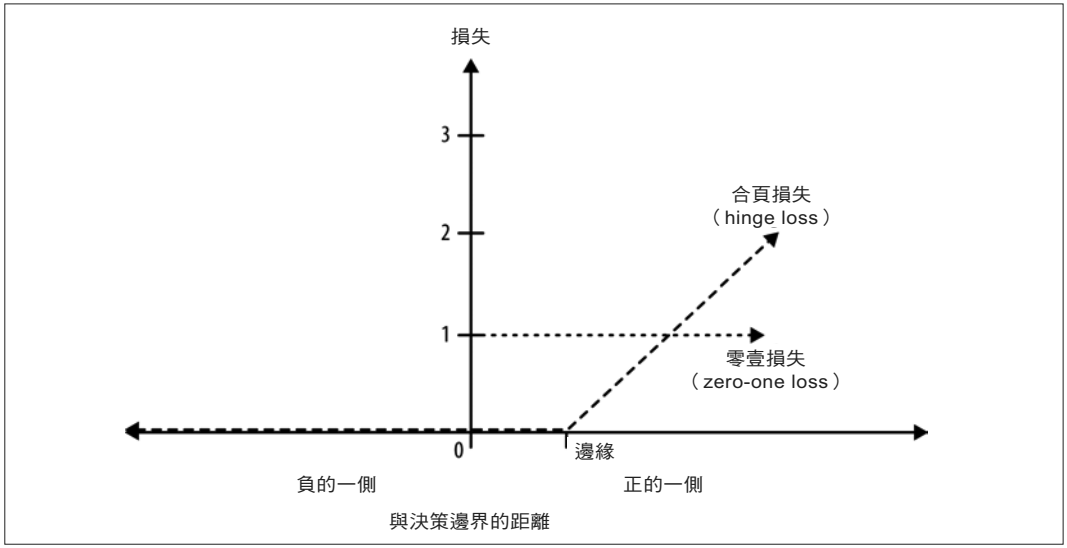

圖 4-9　圖中繪製了兩個損失函數。其中 x 軸代表與決策邊界的距離，y 軸則代表因負實體所造成的損失，其值是依據實體與決策邊界的距離函數而定（與正實體的情況呈對稱狀）。若負實體落在邊界的負的那一側，就沒有損失。但若它落在邊界的正的（錯誤的）那一側，不同的損失函數就會對此做出不同的懲罰（詳見本章的「補充說明：損失函數（loss function）」）。

$$\log\left(\frac{p_+(\mathbf{x})}{1 - p_+(\mathbf{x})}\right) = f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \cdots$$

方程式 4-3 具體說明了，針對一筆以特徵向量 x 描述的特定資料項目，該類別的對數比值會等於我們的線性函數 $f(\mathrm{x})$。由於我們要的通常是類別的評估機率，而非對數比值，故我們可解出方程式 4-3 裡的 $p_+(\mathrm{x})$。這樣便會產生不那麼漂亮的數量，如方程式 4-4 所示。

方程式 4-4　邏輯函數

$$p_+(\mathbf{x}) = \frac{1}{1 + e^{-f(\mathbf{x})}}$$

雖然方程式 4-4 的數量不是非常漂亮，但透過特定的方式將之繪製成圖表，我們就能看出它完全符合我們的直覺想法，也就是我們希望遠離決策邊界的類別評估是相對較確定的，而靠近決策邊界的是較不確定的。

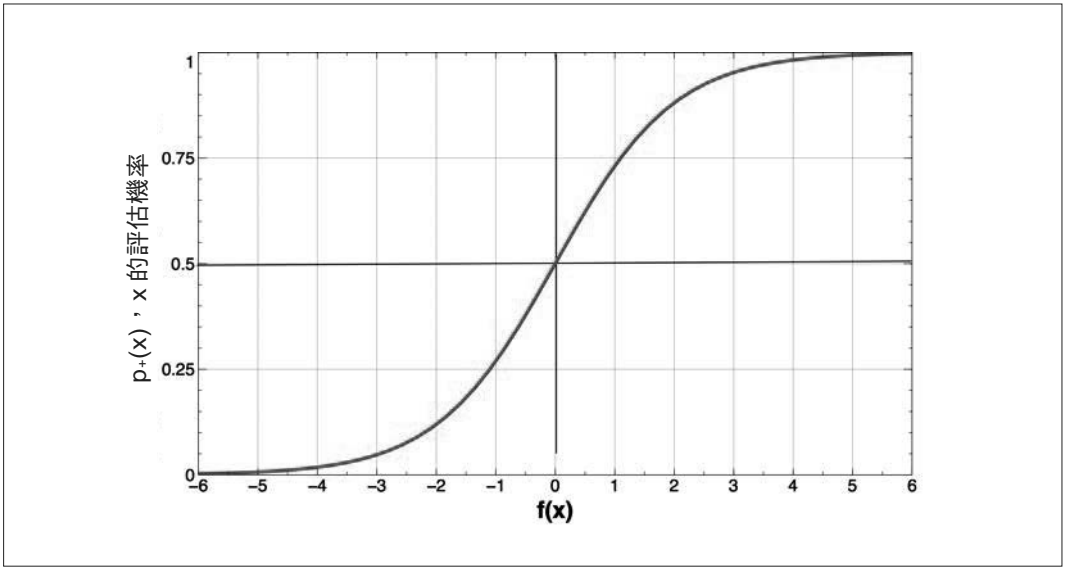

圖 4-10　邏輯迴歸的類別機率評估，以 $f(x)$ 的函數（亦即與分隔邊界的距離）表示。此曲線稱為「S 型曲線（sigmoid curve）」，因其形狀就像字母「S」，而它將機率壓縮在正確的範圍內（0 與 1 之間）。

這點為什麼重要？因為對於許多商業問題，資料科學團隊在模型的採用或實行上，並不具有最後的發言權。通常都至少會有一個管理者必須對模型的實際運用「畫押認可」，而且很多時候，還必須讓一堆利益相關者對該模型覺得滿意才行。例如若要實際運用新的模型，好在顧客打電話給電話公司後，派遣技術人員去修復問題，這時就必須讓來自營運支援、顧客服務及技術開發等部門的經理都相信，新的模型確實利大於弊－畢竟對這種問題來說，沒有所謂完美的模型。

接著讓我們用一個簡單但真實的資料集來嘗試邏輯迴歸，這個資料集是威斯康辛州的乳癌資料集（Wisconsin Breast Cancer Dataset（*http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)*））。就和前一章的蕈菇資料集一樣，這是在加州大學爾灣分校的機器學習庫中，另一個相當受歡迎的資料集。

其中的每個實例都描述了一張細胞核影像的多項特徵，且這些實例都已依據專家對細胞的診斷而標記為良性或惡性（即癌細胞）。圖 4-11 便是這種細胞影像的一個例子。

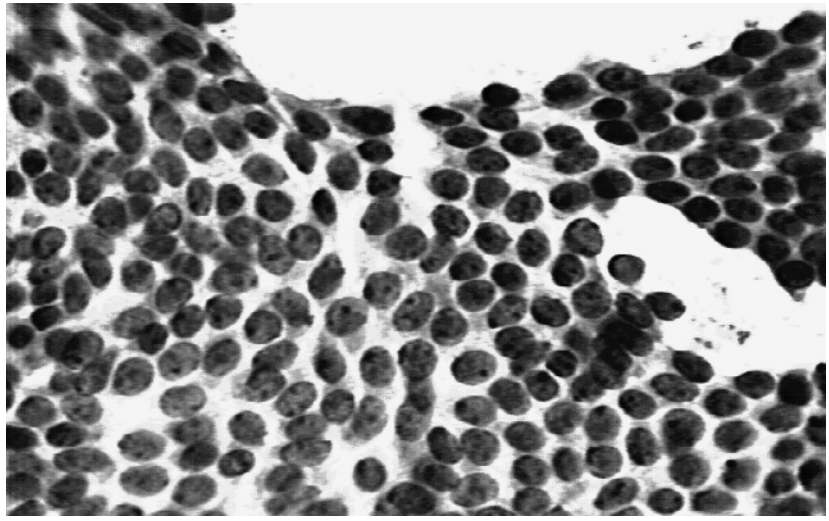

圖 4-11　取自威斯康辛州乳癌資料集的細胞影像之一（感謝 Nick Street 與 Bill Wolberg 提供影像）。

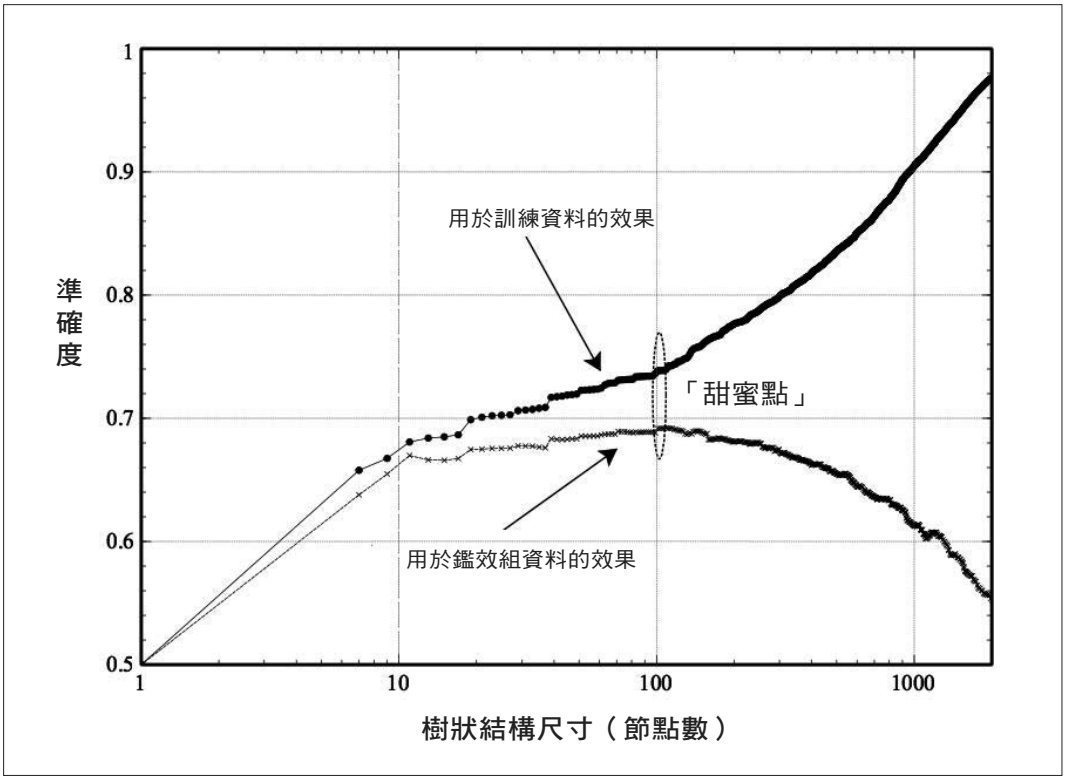

圖 5-3　一個典型的樹狀歸納配適圖。

從左側開始，樹狀結構非常小而且效果不好。隨著可擁有的節點數越來越多，其效果便迅速改善，不論是訓練資料的準確度還是鑑效組資料的準確度，都明顯提升。另外還可看到，訓練資料的準確度總是至少比鑑效組資料的準確度高一點，畢竟建立該模型時用的是訓練資料。不過到了某個點，樹狀結構就開始過適：它開始獲取訓練資料中的某些細節，而那些是不屬於由鑑效組資料集所呈現的一般性群體特徵。在此例中，過適約從 $x = 100$ 個節點處開始，即圖中標示了「甜蜜點」的位置。隨著樹狀結構的尺寸可以越來越大，訓練資料的準確度便持續提升－事實上，只要我們允許，它是能夠記憶整個訓練資料集的，而這會形成 1.0 的準確度（未呈現於圖中）。但鑑效組資料的準確度，卻是隨著樹狀結構的尺寸成長至超過「甜蜜點」時開始下降，這時位於葉節點的資料子集變得越來越小，於是模型的普遍化便是來自越來越少的資料。這樣的推測會越來越容易出錯，用於鑑效組資料的效果就會變糟。

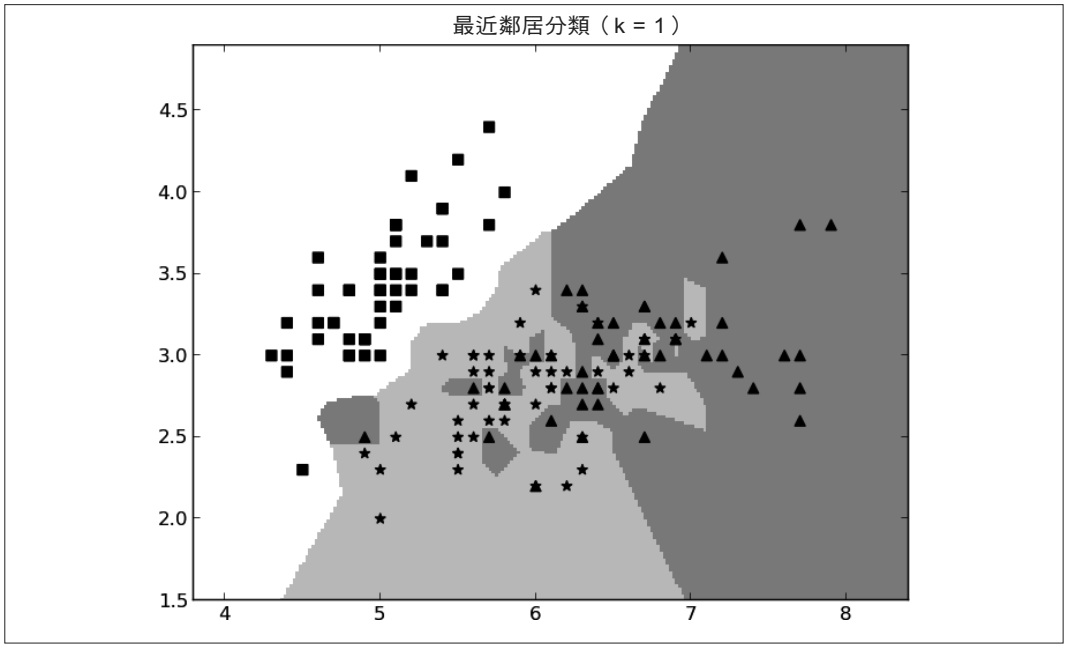

圖 6-4　針對三類別問題，以 1-NN（單一最近鄰居）建立的分類邊界。

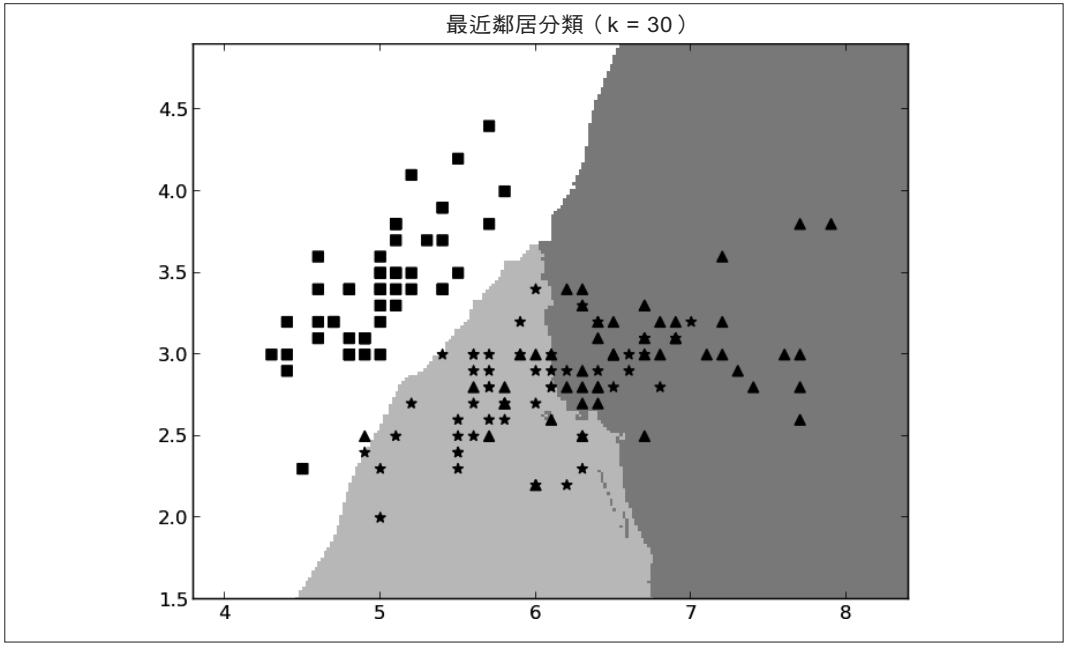

圖 6-5　針對三類別問題，以 30-NN（平均 30 個最近鄰居）建立的分類邊界。

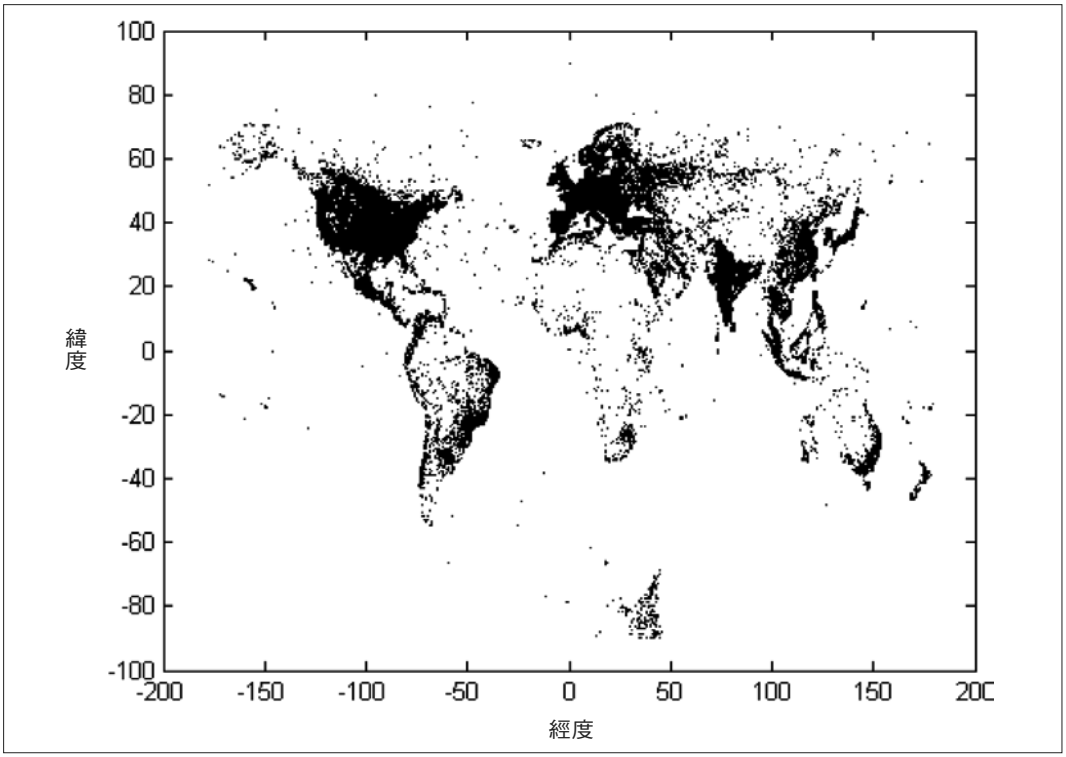

圖 14-1　擷取自行動裝置的 GPS 位置樣本的散佈圖。

附帶一提，有趣的是，這只是個以行動裝置所廣播出來的緯度和經度資料所描繪成的散佈圖；其中並沒有地圖存在！它提供了很清楚漂亮的世界人口密度分佈狀況。而這不禁讓我們懷疑起，那些在南極洲的行動裝置到底是怎麼一回事兒…。

我們可能會怎麼用這種資料呢？讓我們應用基礎概念來想。如果我們想超越探索性的數據分析（因為我們是從視覺化的圖 14-1 開始），就需要思考具體的商業問題。特定的一家公司可能有某些問題要解決，並聚焦於其中的一或兩個問題。企業家或投資者可能會逐一審視各種不同問題，而這些是她認為公司或消費者目前可能有的問題。讓我們來選一個與這些資料有關的吧。

廣告商在這個新世界所面臨到的問題是，我們看見各種不同的裝置，而一個特定消費者的行為可能會零碎地分佈於其中幾種裝置。在桌上型電腦的世界裡，廣告商一旦找到理想的潛在目標（可能是透過特定顧客瀏覽器裡的 cookie 或裝置 ID），就可開始採取相應