序

隨著科技的進步與資料分析技術的不斷演進，大數據分析已經成為企業決策與市場競爭中不可或缺的一環。自本書第一版問世以來，我們收到許多來自學術界與業界的回饋，這些寶貴的意見不僅讓我們更深刻地理解讀者的需求，也促使我們決定推出這本《商用大數據分析》第二版，以更完整、實用的內容來協助讀者掌握大數據時代的核心能力。

本書的初衷是讓非資訊背景的讀者——尤其是商管領域的學生與職場人士——能夠以淺顯易懂的方式學習大數據分析的理論與應用。因此，在本次改版中，我們不僅更新了部分章節內容，補充最新的資料分析技術與案例，也進一步優化了原有的解釋方式，使讀者能夠更直觀地理解資料科學與商業應用之間的關聯。此外，我們也特別針對生成式人工智慧（Generative AI）等新興技術進行介紹，讓讀者可以在大數據的框架下，掌握未來技術的發展趨勢。

在撰寫過程中，我們仍然堅持「理論與實務並重」的精神，透過管理面敘事與技術面實作相互搭配的方式，幫助讀者從概念學習到實際操作，真正應用於自己的專業領域。我們希望，無論是大數據初學者，還是希望在職場上提升競爭力的專業人士，都能透過本書，獲得扎實的數據分析能力，並將之轉化為管理決策的核心競爭力。

最後，我要感謝與我並肩努力的共同作者震耀、瑞益、惟元，以及所有在本書撰寫與編輯過程中提供寶貴建議的夥伴們。特別感謝許秉瑜院長的支持與指導，讓本書能夠更臻完善。同時，也感謝家人與朋友的鼓勵，讓我能夠堅持完成這項挑戰。

希望本書能為讀者提供更完整的學習體驗，並在數據驅動的世界裡，成為各位探索與應用大數據的起點。期待您在翻閱本書的過程中，能夠發現更多可能性，並將所學轉化為實際行動，為未來開創更多價值！

梁直青敬上

2025 年 2 月

本書旨在透過明確易懂的詞彙和介紹，協助商業管理類學生、讀者以及對大數據分析感興趣的人士迅速掌握大數據的理論和應用。為便於商業管理學生快速上手，本書選用了目前主流的 Python 作為主要分析工具。此外，為避免界面操作的困難以及版本更新造成程式無法運作的問題，本書的所有程式碼均可在 Google Colaboratory（簡稱 Colab）上執行。在撰寫內容方面，本書透過直接說明的方式，幫助學生了解大數據分析的原理並掌握應用方法。與市面上以程式碼介紹 Python 大數據分析的書籍不同，本書著重於協助讀者迅速上手，並以管理面議題作為主要論述起點。管理面的案例和介紹有助於商業管理（或非資訊科技）類學生產生共鳴。

本書前三章為大數據分析的基礎知識：第一章為基本介紹，對大數據的意義進行說明；第二章則探討大數據分析的前置作業和資料查找技巧，並對探究主題進行概要分類和說明。從第三章起，本書主要以第一章管理面的說明為基礎，逐章介紹分析方法，並根據大數據分析工具進行操作和詳細解釋。若希望在課堂中使用本書作為大數據分析教材，可遵循以下建議：以一週三小時的課程為例，首先在第一週進行基本介紹，簡單介紹可查找大數據分析資料的網站，幫助學生在網路上找到可用的公開免費資料檔案。接著引導學生理解資料需經過整理並挖掘管理意義，才能呈現大數據分析的真實價值。然後，運用一到兩週時間介紹 Colab 系統運作和基本語法或操作，以便學生了解如何應用 Python 進行大數據分析。在接下來的十三週課程（不含期中考與期末考）中，逐步引領讀者深入大數據分析領域。

另外，在上述的分析章節中，為了配合 Python 分析，本書提供的商業管理分析資料集源於大型超市百貨業的實際數據。然而，為了實現數據去識別化，資料已經過處理，不再顯示產品名稱、客戶名字等可識別的原始資訊。本書以專業的方式呈現大數據分析的理論和實踐，著重於協助讀者快速上手，並以管理面議題作為主要論述起點。透過提供清晰易懂的介紹和實用的案例，本書旨在幫助商業管理類學生、讀者及對大數據分析感興趣的人士深入理解大數據的學理與應用，並在實際操作中不斷提升分析能力。

1.1 認識商用大數據分析

大數據被定義為「極為龐大的資料集合」（Big Data）（De Mauro, Greco, & Grimaldi, 2016）。在此領域中，經常提及的大數據分析（Big Data Analysis）涵蓋了資料探勘或資料挖掘（英文皆為 Data Mining, DM）以及商業智慧（Business Intelligence, BI）。雖然商業管理人員對這些術語可能略知一二，但對於它們的具體應用、區別以及對商業管理領域的意義可能仍然不太了解。

實際上，資料探勘（Pujari, 2001）與商業智慧（Watson, 2007）都是大數據應用的重要組成部分，如圖 1.1.1 所示。從應用層面來看，商業智慧主要是幫助分析者可以快速整合企業內部各種資料，並以圖表形式呈現具有實際應用價值的分析結果。而資料探勘則主要是透過關聯（Association）、分類（Classification）、集群（Clustering）、趨勢預測（Trend）、以及序列樣式（sequence pattern）等多種分析方法，從企業內外部資料中挖掘出有助於公司發展的見解和結果（Pujari, 2001）。

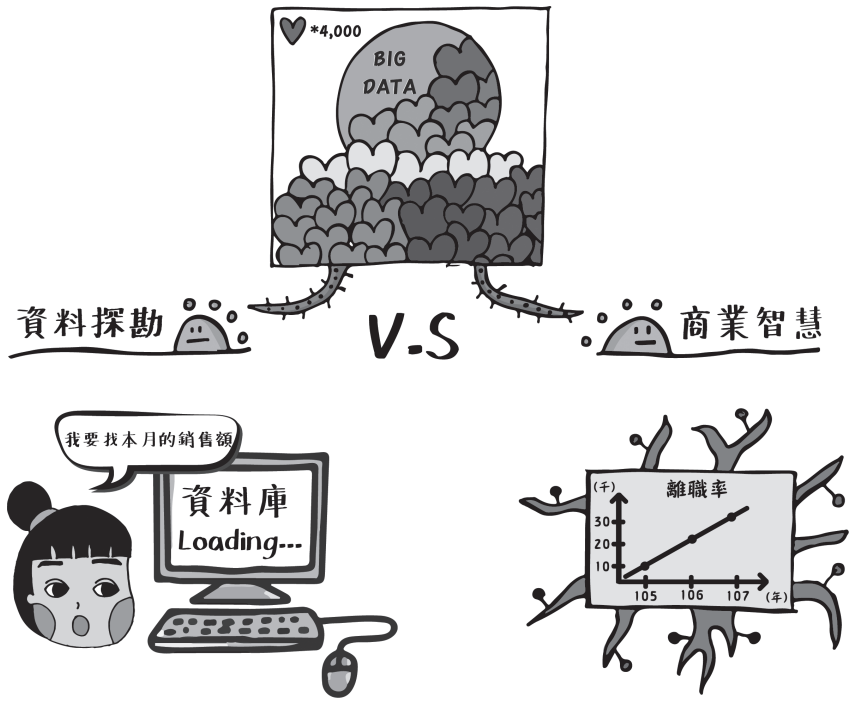


圖 1.1.1　大數據是在分析什麼

公司在透過資料探勘發現潛在結果後，便可依此制定行銷策略。例如，某公司可能在分析關聯規則時，發現部分顧客在購買尿布時同時購買啤酒。由於尿布與啤酒的關聯並非日常生活經驗中常見的組合，對於此發現產生疑惑的反應是正常的。然而，具有敏銳商業直覺的人應能從此例中衍生出以下想法：「這些顧客可能具有相似的行為模式，或許屬於特定族群，或在特定情境下展現此購買行為，這些顧客可能具有商品促銷的潛力」。

透過資料探勘手法及敏銳的商業直覺，公司能從客戶資料庫中找出這些潛在顧客。若已建立完善的會員制度，公司通常能獲得這些顧客的聯繫方式，並針對他們的交易資料進行精準行銷（Precision Marketing）（Hou, 2015）。在確立這些顧客的行為模式並理解可能的促銷策略後，可將資料探勘分析結果應用於後續具購買行為的新客戶。

本節心得

資料探勘是有從資料庫找尋有商業意義的資料。

回饋與作業

1. 請說明資料探勘與傳統統計分析的差異。

1.2 資料探勘（Data Mining）

資料探勘，顧名思義，是從一個完整的結構中發現有意義的資訊（如圖 1.2.1）（Fayyad, Piatetsky-Shapiro, & Smyth, 1996）。這個過程也可以被稱為挖掘或挖礦。重要的是，探勘的目的取決於所需求的資源。例如，在建築領域，目標可能是挖掘適合建築的沙土。然而，如果在挖掘過程中發現黃金，這些黃金對於建築可能並無直接價值。在這種情況下，黃金不是挖掘過程中的目標，而是尋找有助於實現特定目標的資源（如建築材料）。

圖 1.2.1　探勘就是在發現意涵

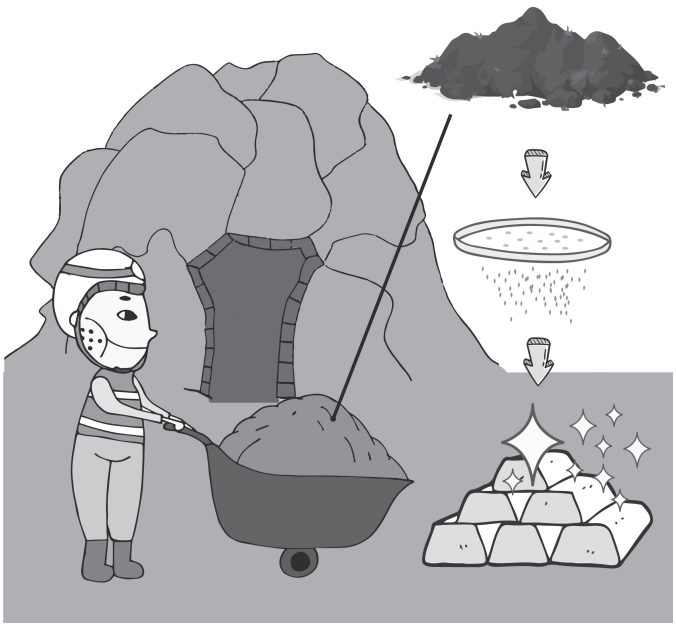
﹀ 註解

從挖礦角度來看，從礦坑裡挖出一堆泥土，不知道這堆泥土裡面有什麼，這堆泥土就叫做資料（「看不懂的東西」），將泥土（資料）過篩（資料整理）後，會變成想要的黃金，這黃金就叫做資訊也就是（「看得懂的東西」）

如果決策者靈活地運用策略，例如將挖掘到的黃金出售以購買建築所需的沙土，這也是一種可行的方法（Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, & Wirth, 2000）。這種策略依賴於決策者的判斷和商業嗅覺。在資料探勘過程中，意外發現的資訊可能成為具有管理價值的結果。

資料探勘並非像初學者所想的那樣容易（遍地黃金）。事實上，即使挖掘出黃金，也不一定對特定目標有用。因此，建議初學者在進行商業資料探勘時，不要隨意根據直覺作出結論。多數情況下，初學者的發現之所以難以理解，是因為對該領域不熟悉，而非發現本身無用。換句話說，培養商業直覺和基本嗅覺是學習資料探勘過程中，商管人員需要從職場實戰經驗中獲得的能力（Provost & Fawcett, 2013）。

此外，商管人員在進行資料探勘時，必須首先了解「資料」的本質。這將決定他們能否挖掘出有用的資訊，如圖 1.2.2 所示（Han, Kamber, & Pei, 2011）。理解資料的特性與結構對於確保資料探勘成果具有相應價值至關重要。



﹀ 圖 1.2.2 什麼是資料

∨ 註解

如果是蓋房子，圖上面的黃金是資料還是沙子是資料？如果要從資料（挖到的東西）中萃取資訊（可以用來建築的材料），那兩者都是資料。因為還看不懂挖到的東西。

在實際應用中，資料探勘與機器學習技術可以幫助商管人員從大量資料中提取有意義的見解和趨勢，進而做出更明智的商業決策（Witten, Frank, Hall, & Pal, 2016）。然而，商管人員在運用這些技術時，應該保持謹慎，確保他們了解所探勘的資料，並能夠判斷哪些資訊對解決問題具有實際價值。

資料探勘是一個複雜且具挑戰性的過程，商管人員需透過不斷學習和實踐來提高他們的技能和嗅覺。在探勘過程中，挖掘出的資訊可能對特定目標有用，也可能無用。因此，商管人員需要具備扎實的專業知識和經驗，以確保他們能夠在資料探勘過程中做出明智且負責任的決策。此外，了解資料的本質和特性對於確保資料探勘成果的有效性和可靠性至關重要。

何謂「資料」？與資訊相比，資料可以被視為「未經解析的原始內容」。當我們面對一個充滿文字、數據和符號的表格（如表 1.2.1 所示），通常我們可能無法立即理解其意義。在這種情況下，我們可能會根據過往的經驗和直覺，對表格內容進行初步的猜測。例如，當我們看到類似「日期」的記錄（如 "20201012"），並知道提供者來自採購部門時，我們可能會認為這是「進貨日期」、「出貨日期」等。然而，在確切了解欄位意義之前，我們不能僅憑直覺對這些資料進行解釋。事實上，我們甚至無法確定該記錄是否真的代表「日期」。

表 1.2.1　看不懂的就是資料

1201	1026	1	20020102	1321	32976	42113	NULL	NULL	9004061	4061	NULL
1201	1026	1	20020102	1322	32977	42113	NULL	NULL	4710543215040	10010491	NULL
1201	1026	1	20020102	1324	32978	42113	NULL	NULL	4710012241549	10016892	NULL
1201	1026	1	20020102	1324	32978	42113	NULL	NULL	4710583100016	10003407	NULL
1201	1026	1	20020102	1324	32978	42113	NULL	NULL	5010006100128	10013929	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	4710908110423	10031661	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	4710552062703	10020623	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	4715545051054	10033645	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	4710249001732	10082411	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	4710011408011	10004020	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	4710088411785	10008448	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	4710583100016	10003407	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	4711022100338	10003352	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	20000101286	10005055	NULL
1201	1026	1	20020102	1327	32979	42113	NULL	0200030083-00	4710315012358	10006847	NULL

為了確切理解資料的意義，我們需要對表格中的各個欄位進行側面對照。這意味著我們需要參考一些資料，以了解表格中各個欄位的確切定義。在資料庫管理和系統開發領域，這種對照被稱為資料字典（Data Dictionary）（Uhrowczik, 1973）。資料字典是一本解釋資料組成及其基本意義的參考工具。換句話說，我們需要一本字典 / 參考書，以便了解表格中各個欄位的含義。

當我們理解資料及其意義之後，這些資料將不再被稱為「資料」，而是變成了「資訊」。資料和資訊之間的差異在於它們是否可被理解。資訊是「已解析的資料」，而資料則是「未解析的資訊」。因此，理解資料的含義是將資料轉化為有價值的資訊的關鍵過程。

當我們理解資料內容的意義後，可以根據其意義選擇合適的欄位進行後續分析。然而，在進行資料探勘之前，我們還需考慮資料「品質」。資料品質是指資料需符合分析目的的要求。我們需要評估欄位選擇、資料完整性、數值正確性以及資料型態等方面，以確保資料品質達到標準。

對於商業管理人員來說，在進行大數據探勘之前，了解大數據探勘的組成是至關重要的（如圖 1.2.3 所示）。這包括「資料本身」以及大數據分析的「目的」。資料本身涵蓋資料品質和資料意義；「資料意義」即是資料字典（data dictionary）的存在原因（Rashid et al., 2020）。此外，資料品質和意義需與分析目的相互協調，以符合最初分析意圖和企業需求。

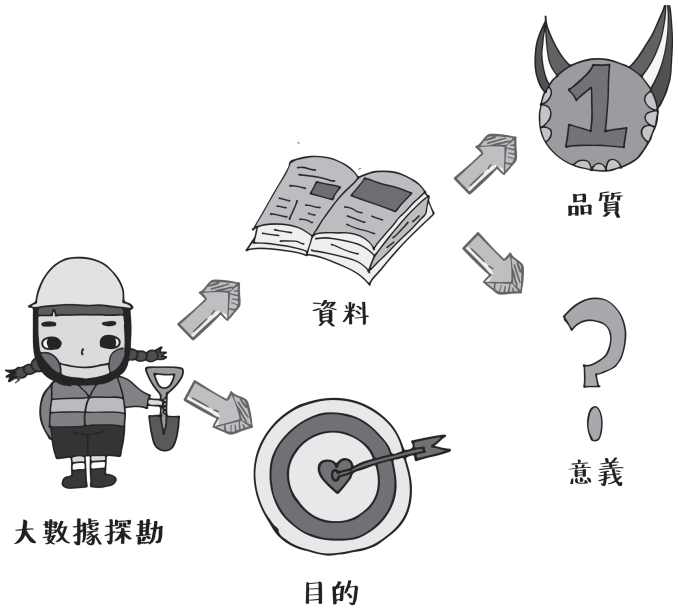


圖 1.2.3　大數據探勘的組成

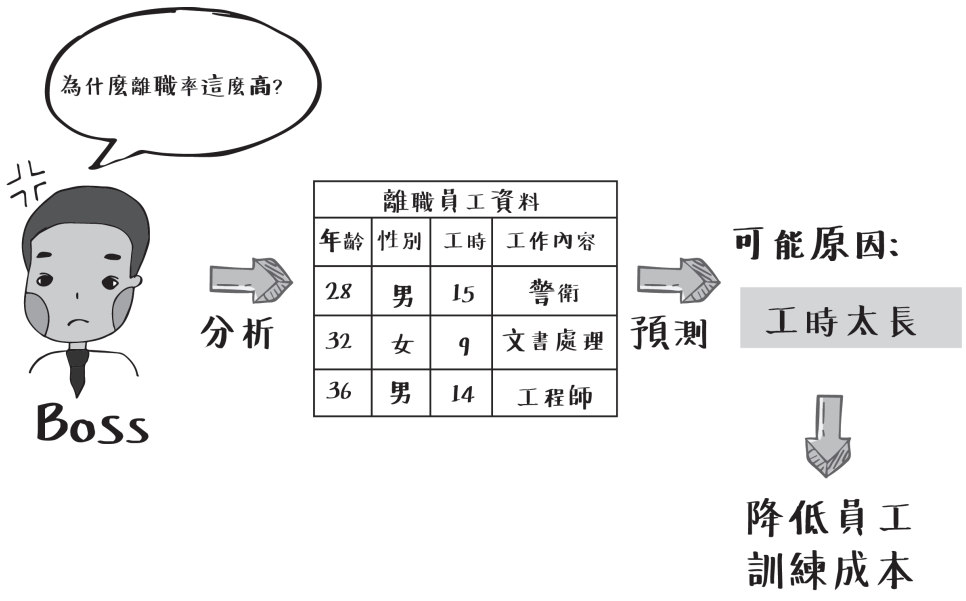


圖 12.1.1 離職率預測分析

12.2 預測的基本概念

在資料豐富的情況下，有更大的機會將所有可能面向納入預測模型進行考慮。意即，為了進行預測，需要根據對象（數據表中的每條記錄）的特性（即欄位，亦即上一節提及的考量面向）的觀察值來估算目標欄位（特性）的可能數據。例如，如果手頭有班級同學的性別、身高、體重、年齡等資料，在一個人走進班級時，想要預測該人的性別，就可以透過預測的方法分析上述欄位來進行計算。

傳統商業管理統計學中，預測多是透過迴歸分析來尋求可能的結果。迴歸分析旨在探索兩個或多個變量之間是否存在相關、相關的方向及其強度，並建立數學模型以觀察特定變量的變化，從而預測研究者感興趣的變量的可能結果。換言之，迴歸分析可以揭示因變量（Y）如何因自變量（X）的變化而變化。以下是一個典型的迴歸方程式，其中 Y 為因變量，X 為自變量。若應用於前述例子，X 可以分別代表身高、年齡、體重，而 Y 則是待預測的性別，ε 則表示誤差項。

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \varepsilon$$

在資料量龐大時，執行統計分析將耗用大量計算資源及時間。此外，迴歸預測在統計上存在限制，例如必須假設母體具常態分布、變量間需獨立、且變異數需恆定等嚴格條件（Longford, 1995）。對於大數據而言，因資料量過大，傳統的統計假設檢定常呈現顯著，但這反而可能失去原有檢定的意義。迴歸分析的另一問題是其假設性，假定預測可透過直線完成，這限制了考慮不同可能性的範圍，可能降低預測的準確性。

預測不僅限於迴歸分析。隨著電腦技術的進步，面對大量數據時，電腦演算法成為進行預測的有效工具。假設有一班級學生的資料，想要預測一個進入教室的人的性別，可以透過類神經網路（Artificial Neural Network, ANN）來計算性別預測。

1. 將「訓練資料集」（即班級學生的資料）如身高、年齡、體重等信息輸入到神經網路的輸入層。
2. 在隱藏層中，每個神經元對輸入的數據加權後進行計算，並透過激勵函數調整其輸出結果。
3. 若該結果未達到特定門檻值（即計算後的性別數值未達到設定的標準），則不向下一層輸出；若達到門檻值，則輸出該數據。
4. 經過層層計算，若數值符合預期，則進行到下一神經元並與其他數據整合，如圖 12.2.1 所示。
5. 經多層計算後，最終產生可能的性別，並由輸出層神經元提供答案。
6. 若最終輸出的性別與訓練資料的實際結果不符，則需對系統參數進行調整和修正，持續優化直至能準確預測性別為止。

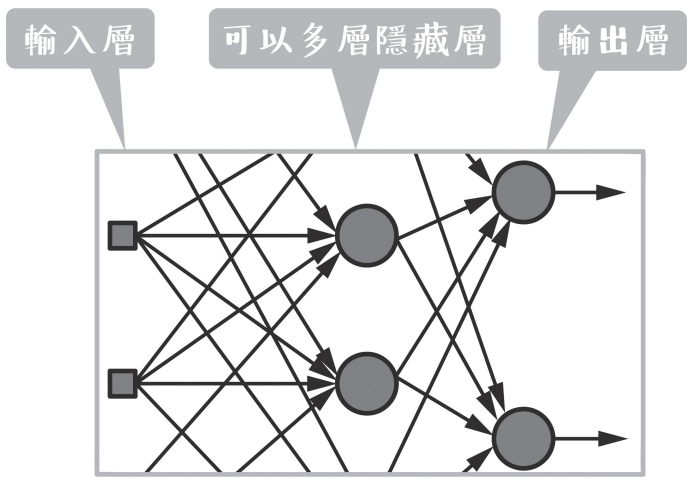


圖 12.2.1　類神經網路分層運作

ANN 預測分析通常提供的是質化而非量化的結果，即結果屬於某一類別，如男性或女性。為了確認預測結果的準確性，需要透過後續的驗證過程，利用測試數據與混淆矩陣來檢驗模型的準確度。

進行學生性別預測時，需設置訓練數據集與測試數據集兩部分。這涉及將班級學生資料透過 k-fold 方法劃分為**訓練組**（例如，75% 的數據）和**測試組**（剩餘的 25% 數據）。首先，訓練組數據被輸入至 ANN 進行模型訓練，隨後測試組數據用於檢驗訓練好的模型的預測效能。然而，不能假定訓練過程中確定的模型權重在測試階段能夠展現出良好的表現，即預測結果的準確性不是必然的。因此，應進行多次的隨機抽樣以及訓練和測試組的設定，透過反覆的演練，尋找到能產生最準確預測結果的模型參數。

類神經網路演算法的運作過程是頗為複雜的，因為電腦需要不斷地進行訓練與測試，以尋找最佳的解答。透過 ANN 演算法進行模型訓練時，可能會產生過度擬合的問題，這類似於給同一人反覆進行同樣測試，並允許其修正錯誤，最終必然能達到滿分的情況，但如果請這人做其他衍生題目可能就無法回應。

除了利用 k-fold 交叉驗證之外，如果條件允許（即有足夠的時間和預算來收集或獲取更多數據進行測試和驗證），使用新數據進行測試可以提高模型的準確度。然而，在現實中，往往受到成本的限制，公司無法等待新數據的測試結果來編制年度預算或進行業務規劃。

此外，由於隱藏層與神經元之間的複雜相互作用，類神經網路演算法產生的結果幾乎無法解釋，因此常被視為一種「黑箱」操作。這種不透明性使得分析人員難以向管理層解釋並說服他們接受基於這些模型的決策。

這種情況類似於 2002 年湯姆•克魯斯主演的電影電影關鍵報告（Minority Report），其中特警部隊能夠透過一種犯罪預知系統來預測並阻止犯罪發生。然而，在現實世界中，犯罪的動機和原因通常在犯罪發生後才能被推理出來，而無法提前預知。因此，當企業領導面對一個無法解釋原因的「黑箱」結果時，他們不太可能僅憑對演算法推理的信任來推動商業決策。

12.3 類神經如何運作

（ANN）的概念源自於模仿人類大腦神經元對刺激的處理和反應機制。在人類大腦中，當外部刺激被神經元接收後，會經過一系列的處理過程產生反應。例如，當看到一張照片時，大腦開始回想這人是誰，進而做出判斷。這一過程包括了接收刺激（看到照片）、進行思考（思考是否是某人），以及最終做出決定（判斷是否是某人）的連續步驟，顯示了腦神經如何逐步進行計算以得出結果。

人類大腦的運作機制極其複雜，科學家至今仍在不斷探索其工作原理。如圖 12.3.1 目前所了解的是，大腦透過神經元間的化學物質傳遞進行溝通，這些化學物質經由多個樹突從不同來源輸入（input）到神經元的細胞核中，細胞核會根據這些刺激作出相應的反應。當輸入訊號的總強度超過一定閾值（threshold）時，化學物質會通過軸突和突觸被傳遞到下一個神經元。此過程涵蓋了從接收刺激（例如看到照片的刺激或者是來自先前的神經元處理後的化學物質）、細胞核對這些刺激進行處理（神經細胞核針對目標與問題進行決策判決）並作出決策，到最終將處理後的訊號輸出到下一級神經元或直接作出行動決策（將新的化學物質送到下一層神經元或者輸出決策）的全過程。

在此過程中，神經元的工作可以分為接收輸入（例如，視覺刺激或來自其他神經元的訊號）、處理訊號（細胞核對刺激做出反應）以及產生輸出（將訊號傳遞至下一個神經元或產生決策）三個主要階段。神經元會盡可能產生一個合理的結果，即透過綜合考量多個因素後得到的邏輯上正確的輸出，這類似於人類在學習過程中參考內外部因素調整想法並做出正確決定的過程。

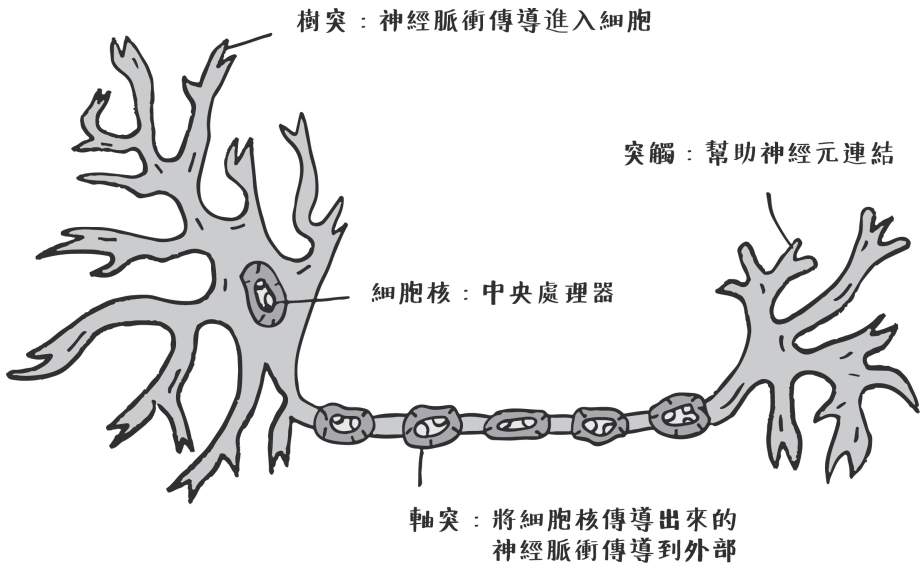


圖 12.3.1　腦神經元示意圖

類神經演算法就是嘗試以電腦模擬人類的思考歷程：從外部輸入刺激、神經元處理接受刺激後該如何回應、最後綜合神經元處理結果輸出決定。類神經網路演算法依據其複雜度，可以有一個到幾萬個神經元組成，但個別神經元運作原理會與上述的神經網路思考過程相似。

以下以一個最簡單類神經運作幫助各位了解類神經的處理過程，這神經網路以一個輸入、一顆神經元、一個輸出來說明類神經運作過程。假設看到一張照片，如圖 12.3.2 右邊所示要思考這是否是十年不見的阿娟。

支援向量機（SVM）的出現，很快地就成為當時機器學習模型的主流。在實務的應用上或是學術上都是很常看到使用的演算法。支援向量機（SVM）透過不同的間隔最大化方式，模型可以分成線性可分、近似線性可分和線性不可分等不同的支援向量機（SVM）。這個演算法實際上在 1963 年就由當時蘇聯的數學家 Vapnik 提出了支援向量的概念，隨著他移民到美國後，在這方面的研究受到當時學術界的重視，才讓支援向量機在二十世紀末變得炙手可熱。

整個來說，支援向量機（SVM）是一種強大的機器學習模型，能夠建立線性或非線性的決策邊界。透過決策邊界的建立，來將資料進行區分，進行分類的行為。在深度學習演算法還沒有像現在這麼風行的時候，支援向量機（SVM）可說是當時最閃亮的分類演算法，除了針對結構化數值資料，也有在當時應用到圖片的分類上，像是美國郵遞區號的圖片分類之類的問題處理，在分類的成果上，也有不錯的成果

15.1 跑一次支援向量機算法

支援向量機（SVM）就如同先前介紹的決策樹或是隨機森林等，是屬於機器學習領域中的一種監督式學習（Supervised learning）方法，主要用於解決二元分類問題（Classification），但也可以擴展到解決多元分類的問題上。除了分類問題的處理外，針對回歸問題，也可以透過 SVM 來獲得預測結果。

接下將利用 Python 來進行建模與預測，資料集與第八章是相同的資料集，資料內容是描述公司客戶價值的資料，依照貢獻的價值等級分為高價值客戶（H）、中價值客戶（M）及低價值客戶（L）三種。而評估客戶價值的欄位有三個，就是近一次消費（Recency）、消費頻率（Frequency）及消費金額（Monetary），資料集的檔案名稱為 **classification_svm_Ex1.csv**，第一到四個欄位都為數值資料，但嚴格說起來，第一個欄位是表示客戶的編號，基本上可算是 meta data。第五個欄位是類別資料儲存客戶類別值，分別是 H、M 以及 L 三種，資料如表 15.1.1 所示。

表 15.1.1 客戶價值資料集

cid	R	F	M	customer_value
1069	19	4	486	M
1113	54	4	557.5	M
1250	19	2	791.5	M
1359	87	1	364	L
1823	36	3	869	M

cid	R	F	M	customer_value
...	...	...		...
2179544	1	1	3753	M
2179568	1	1	406	L
2179605	1	1	6001	M
2179643	1	1	887	L
20002000	24	27	1814.63	H

資料分類的程序基本上同先前章節所敘述的有六個步驟，分別為資料載入與準備、訓練與測試資料切割、訓練（學習）與建立分類模型、產生測試資料集的預測結果、評估測試資料集在分類模型的績效表現以及預測未知類別的新資料，這一整套資料分類的程序與第八章的說明相同，如圖 15.1.1 所示。唯一不同就是使用的演算法不同。這次是使用支援向量機（SVM）作為建立分類模型的演算法，在實作上就會因為所選的不同演算法而使用不同的函數來建立模型。

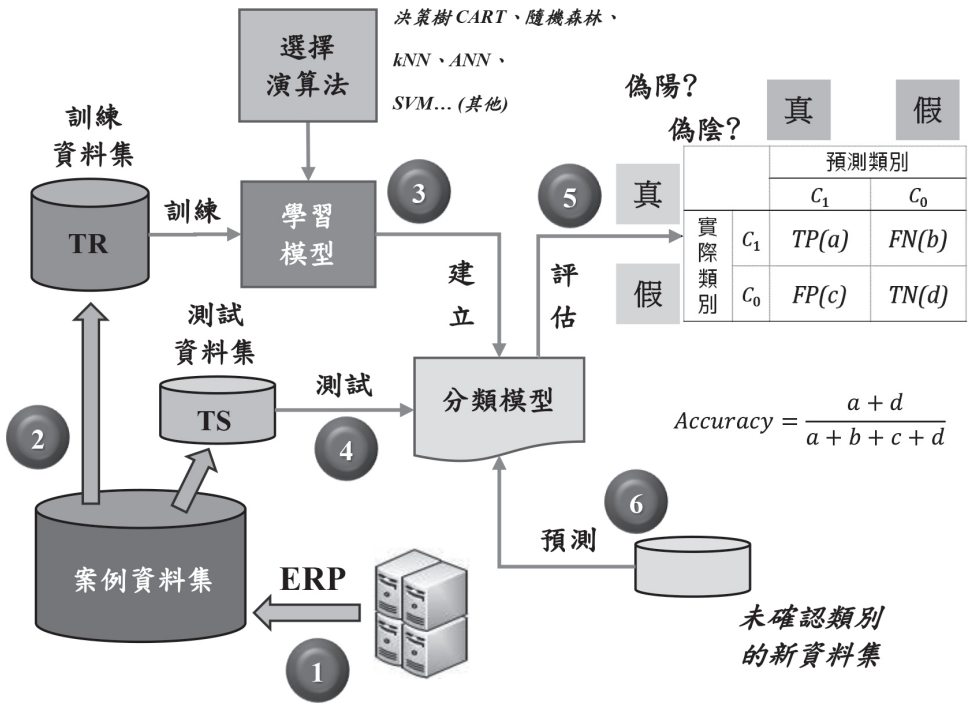


圖 15.1.1　模型建立步驟

在本章中，以客戶價值貢獻度資料帶領各位完成支援向量機（SVM）分類模型與預測分析。資料來源同前面的章節是以一個客戶價值資料集 **classification_SVM_Ex1.csv** 為主建立一個客戶資料分類模型。

首先開啟一個全新的空白的 Colab 筆記本，不管預設檔案名稱為何，先將檔案名稱修改為 **Classification_SVM1.ipynb**，並按「連線」按鈕讓筆記本連到 Google 雲端資源，接著執行「檔案→儲存」先將開啟的空白筆記本進行儲存。接下來請把經常使用的模組（pandas，以 pd 代稱，以及 numpy，以 np 代稱）引進到程式區塊中。操作說明請參考附錄一。

```
import pandas as pd
import numpy as np
```

緊接著就要來建立支援向量機（SVM）方法的預測模型。使用 Python 語言來建立分類模型時，其步驟有著非常高的標準化程序，通常可以使用六個步驟完成建立分類模型與預測的工作，如圖 15.1.1 所示，圖上有六個數字分別對應以下六個工作步驟，分述如下：

步驟一、資料載入與準備

我們先將表 15.1.1 的客戶價值資料集內容讀入 Colab 環境中。客戶價值資料集是透過 ERP 系統中下載客戶交易資料與客戶基本資料後經過 R、F、M 的計算所得到的客戶價值資料，並且也邀請行銷領域專家、學者評估每一位客戶價值資料後給予高（H）、中（M）、低（L）價值標籤處理，最後儲存成 csv（逗號分隔值，Comma-Separated Values）檔案資料，檔案名稱為 **classification_SVM_Ex1.csv**，總共有 32,266 筆資料，描述 32,266 位公司客戶的貢獻價值狀況。請先將 **classification_ SVM _Ex1.csv** 檔案上傳到 Colab 雲端環境中，並將下列斜體 Python 程式碼指令輸入到程式碼區塊中執行來查看整個客戶價值資料集。如下圖 15.1.2 所示。

```
df = pd.read_csv('/content/classification_SVM_Ex1.csv')
df
```

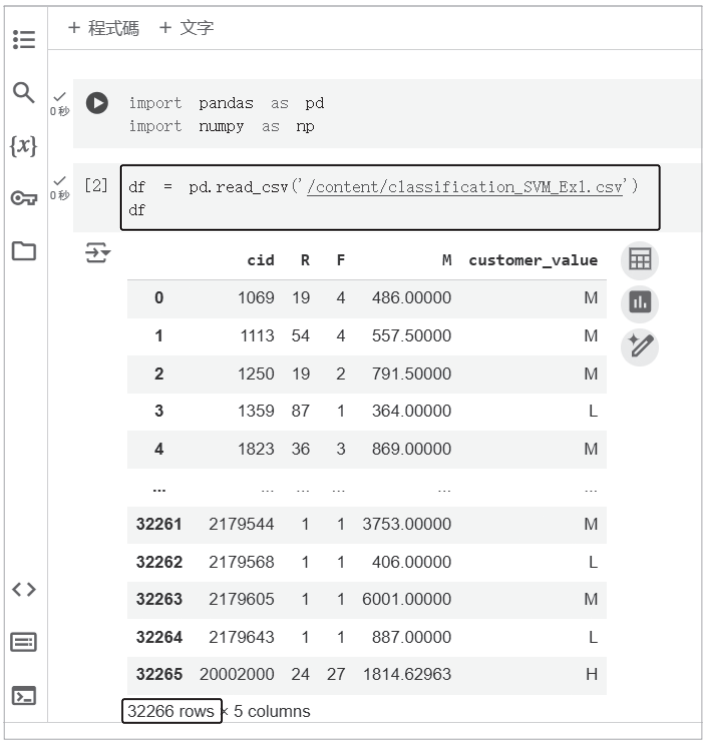


圖 15.1.2　顯示資料集檔案內容

因為支援向量機方法也是屬於監督式機器學習方法，就如同先前介紹的決策樹、隨機森林、kNN 的實作一樣的做法。同第八、十、十二章所言，在建立模型的過程中需要先行決定兩件事情：模型的**目標變數**與**自變數**。目標變數通常指的就是建立分類模型後，想要使用模型幫我們預測什麼事情。譬如我們想建立分類模型來預測新的客戶價值，因此 customer_value 欄位就會是目標變數欄位。接著就要去想是需要使用哪一些資料來預測新客戶的價值等級呢？這些會影響目標變數判定的欄位稱為自變數欄位，如圖 15.1.2 中的 R、F、M 欄位。與第八、十、十一章相同的目的，在客戶價值資料集的範例中，**目標變數**與**自變數**這兩個變數分別為：

- 目標變數欄位：customer_vlaue
- 自變數欄位：R、F、M

（請注意：通常編碼、代碼等欄位的資料值因為唯一性特質很強，所以不太會被納入自變數欄位，譬如 cid。）

在 Python 中為了方便後續建立模型以及模型的評估工作，我們需先將目標變數與自變數資料分別設定後使用。先將三個自變數 R、F、M 儲存在變數 CustomerData 中，而目標變數欄位 customer_value 就儲存在變數 CustomerTagret 內。先將下列斜體 Python 程式碼指令輸入到程式碼區塊中並且執行，如下圖 15.1.3 與 15.1.4 所示：

最後，我們也能夠透過評估模型的預測績效，如準確率和混淆矩陣，來衡量我們的分類結果。這有助於協助決策者判斷，在未來的行動中如何調整策略及資源配置。例如，如果模型顯示出高準確率，則可放心依賴這些預測結果來設計促銷活動；反之，則需進一步分析模型的不足之處，並根據結果進行調整。

本章心得

支援向量機是一個分類很強的演算法，在神經網路沒落的時期中，是一個代替神經網路的方法，不管在處理數值問題或是類別問題都是很有效的。

回饋與作業

1. 支援向量機（SVM）演算法為何是監督式演算法？試從實作中說明理由。

本章習題

1.（ ）下列哪一選項是一種強大的機器學習模型，能夠建立線性或非線性的決策邊界。透過決策邊界的建立，來將資料進行區分，進行分類的行為？

(A) 梯度降維　(B) 關聯規則
(C) 聚類分析　(D) 支援向量機

2.（ ）下列哪一選項是在深度學習演算法還沒有像現在這麼風行的時候，可說是當時最閃亮的分類演算法？

(A) 關聯式資料庫正規化方法　(B) 聚類分析
(C) 支援向量機　(D) 商業智慧

3.（ ）下列哪一選項是監督式學習（Supervised learning）方法？

(A) 支援向量機　(B) 聚類分析
(C) 關聯式資料庫正規化方法　(D) K-means

4.（ ）下列哪一選項是一種強大的機器學習模型，能夠建立線性或非線性的決策邊界來建立分類模型？

(A) 梯度升維　(B) kNN
(C) Apriori　(D) 支援向量機

5.（ ）下列哪一選項是透過不同的間隔最大化方式建立分類模型？

(A) FP-Tree　(B) 支援向量機 SVM
(C) 3NF　(D) Apriori