序言

1.寫作緣起

英國著名的唯物主義哲學家法蘭西斯．培根 (Francis Bacon)，在 1620 年的著作《偉大的復興》(Instauratio magna) 的第二部分中說道：「人的知識和人的力量結合為一」，又說「達到人的力量的道路和達到人的知識的道路是緊挨著的，而且幾乎是一樣的」這兩句話，被後人解釋成「知識就是力量」(knowledge is power)，這也是人類第一次提出知識投入與力量產出的概念。再者，1965 年彼得．杜拉克 (Peter F. Drucker) 針對當時科技的發展熱潮 (美國登陸計畫與核子武器發展成功) 提出他對知識的看法：「未來知識將占企業重要的地位，它將取代企業原本賴以為生的土地、勞力、資本以及設備等傳統生產要素。」這也是首位預測知識將是人類生產資源的第一位學者，從而改變了過去傳統經濟學中生產要素的法則。故知識的探勘與獲得，便成為未來人類生存與發展的重要因素。

資料探勘 (Data mining)，便是一種探索知識與發覺知識的方法與工具。換而言之，資料探勘是從大型資料集中探索有趣 (interesting) 及有價值 (valuable) 的問題，並可付諸行動之方案的一個過程。故資料探勘可以衍生 / 呈現存在於資料 (data) 中的某一種模式 (model) 和趨勢 (trend)。這些模式和趨勢可收集在一起，並定義為資料探勘的模型，藉著不同的模型，來協助人類去發現問題、定義問題、以及解決問題。人工智慧 (Artificial Intelligence, AI) 指由人類製造出來的機器所表現出來的智慧。人工智慧的研究可以分為幾個技術問題。其分支領域主要集中在解決具體問題，其中之一是，如何使用各種不同的工具完成特定的應用程式。機器學習 (Machine learning) 則是人工智慧的一個分支。人工智慧的研究歷史有著一條從以「推理」為重點，到以「知識」為重點，再到以「學習」為重點的自然、清晰的脈絡。顯然，機器學習是實現人工智慧的一個途徑，即以機器學習為手段解決人工智慧中的問題。

從法蘭西斯．培根「知識就是力量」的初始概念，人類知識的投入經過商業與市場活動的發展與驗證之後，知識的探勘與管理已然發展成一個兼具學術與實務的學術領域。隨著工業、科技、與管理環境的改變，知識的探勘與管理也不斷地在理論、方法、與工具方面，提供學術界與實務界研究與驗證的題材，因此，二十一世紀將是一個人類重視知識的世紀，而除了資料探勘在各個領域的蓬勃發展，人工智慧與機器學習也將是人類探索與運用知識的重要工具。本書以資料探勘的理論與方法為基礎，同時以資料探勘的應用與發展為導向，探討資料探勘在人工智慧與機器學習未來的發展，企望提供讀者對於相關主題的發展能夠一窺堂奧。

2.本書特色

資料探勘是一門結合統計學與資訊科學相關理論的方法學，藉由各種功能與模式的導入與實踐，使得資料探勘的應用遍及各個領域，成為研究與實務工作者重要的研究方法，尤其是運用在人工智慧及機器學習的未來發展。再者，隨著知識經濟的發展，以資料探勘為基礎，創造個人、組織競爭優勢、與經營績效的管理理論及工具，也就成為資料探勘發展及應用的趨勢。故資料探勘理論與工具方法的學習與導入於組織、企業，就成為知識探勘、運用與管理的重要工作。因此，我們也可以說資料探勘，對於學術界與實務界而言，是一門兼具問題、理論、與方法的學科。這本書所要提供給讀者的內容，即嘗試以不同資料探勘的理論為經，演算方法為緯，在經、緯的架構中，藉著個案實例，以及 SPSS Modeler 系統實際的操作，來說明資料探勘模式與功能所能提供問題解決的方法，以及在人工智慧及機器學習未來的發展。

3.本書的內容架構

本書共區分為十六個章節。第一章是資料探勘概論，將資料探勘的概念、定義、流程、與應用作說明。第二章是資料探勘的功能，將資料探勘的不同功能作介紹，包括分類、推估、預測、集群、關聯、順序等功能，以提供資料探勘不同分析模式的基礎。第三章是資料庫與資料探勘—大數據 I ，說明大數據與資料庫間的關係，將不同資料庫的類型做介紹，說明資料庫與資料探勘的關係。第四章是資料與資料探勘的方式與功能—大數據 II ，說明大數據與資料的關係，並探討的是資料庫架構與資料預處理，以提供資料格式、資料庫、與資料探勘系統聯結的基本概念。

接著的內容共十二章，為資料探勘分析功能，目的在於提供讀者資料探勘相關功能與模式的發展。因此，第五章是決策樹—C5.0，說明決策樹—C5.0 的基本概念與演算法。第六章是分類與迴歸樹—C&R Tree，說明分類與迴歸樹—C&R Tree 的基本概念與演算法。第七章是因數分析—PCA/Factor，說明因數分析—PCA/Factor 的基本概念與演算法。第八章是類神經網路—Neural Net，說明類神經網路—Neural Net 的基本概念與演算法。第九章是貝氏網路 (Bayesian Networks)，說明貝氏網路— Bayesian Networks 的基本概念與演算法。第十章是支援向量機 (Support Vector Machine)，說明貝氏網路—Support Vector Machine 的基本概念與演算法。第十一章是關聯法則—Apriori，說明關聯法則—Apriori 的基本概念與演算法。第十二章是次序分析—Sequence，說明次序分析—Sequence 的基本概念與演算法。第十三章是集群分析—K-Means，說明集群分析—K-Means 的基本概念與演算法。第十四章是類神經網路—Kohonen neural networks，說明類神經網路—Kohonen neural networks 的基本概念與演算法。第十五章是資料探勘與人工智慧發展。第十六章是資料探勘與機器學習發展。

上述十個資料探勘的分析功能除了基本概念與演算法的說明之外，每一個主題分別以 SPSS Modeler 的資料格式與設定，結合實際的例子作分析，並展示相關的分析步驟及系統功能，藉此；使得學習者能夠實際操作不同資料探勘的功能與模式，達到理論與實作兼具的學習目的。

4.本書教學配件

本書教學配件共有兩部份。一是資料探勘十個分析功能的課間實作的實作範例及資料，課間實作功能之一為提供教師各分析功能的範例教學，同時提供教師與學者 SPSS Modeler 資料格式、使用環境、與建模注意事項，因此本書

的課間實作也可以說是 SPSS Modeler 最佳的操作手冊，藉著課間實作的範例與資料，教師可以教授學生不同資料探勘實例的實習與實作。另一方面，第二部份本書配件包括各章教學投影片 PowerPoint 檔案的提供，根據每一章節的內容，本書製作投影片檔案供教師教學時使用。

5.致謝詞

一本教科書的製作，從綱要規劃、審核、資料蒐集、寫作、圖形處理、尋找實例、完成文字內容、校稿、定稿、到出版，實在是工程浩大，個人獨立是無法完成的。首先，感謝本書共同作者溫志皓博士的協助，尤其在資料探勘分析實作範例部分，分享他的系統操作知識與經驗，並提供不同資料探勘分析模式的實例說明，使得本書能夠兼具理論與實務功能。再者，感謝博碩文化總編輯陳錦輝先生，在前一本書：資料探勘理論與應用—以 IBM SPSS Modeler 為範例的基礎上，再度邀請我及溫志皓博士來負責本書的寫作，使得這本書能夠如期完成並重新上市。因此，這本書如果能夠對讀者有所助益，要歸功於上述的協助者。本書中的任何缺失以及錯誤，則是個人的疏忽與能力不足，尚祈各位先進不吝指導！

廖述賢

序言

本書是銜接由博碩出版社出版之「資料探勘理論與應用－以 IBM SPSS Modeler 為範例」之改版內容。由於前述書籍問世以來受到廣大的迴響，有讀者經由 email 寫信鼓勵及提問。因此，這次大幅反映了這些年來讀者的建議和指正，同時也因應 IBM SPSS Modeler 的軟體改版而做了必要的更新。

IBM SPSS Modeler 是一套非常適合進行資料探勘及數據科學的軟體，軟體設計之初，即考量使用者進行資料探勘時，需聚焦於相關的理論知識，因此，對於使用者來說，軟體的使用非常友善。當然，目前市場上有許多軟體是平台都能夠進行資料探勘的專業工作。不過，就筆者長年在學術研究與資料分析的經驗來看，IBM SPSS Modeler 時為一套具備教學、研究、實務的好工具。

為了切合讀者的需要，本書採取簡明易懂的敘述方式，並透過精心設計許多範例及插圖，讓使用者可以逐步完成不同領域資料的資料探勘。本書資料領域涵蓋了生物資訊、醫學診斷、學術量表分析、電力設備狀態監測、尼號乘客存活率分析、公共行政管理、零售業購物籃分析、零售業的需估、城市汙水處理廠的水質分析資料，以及天文星體辨識等領域。

這次改版承蒙廖述賢教授的大力協助與意見提供，讓本書在設計、撰編輯的過程中能夠更加順利。此外，尤其感謝內人羅月涓女士及宸澧寶貝均寶貝的配合，才能讓我在撰寫本書的過程中，無後顧之憂。謝謝晏伊。

本書在教學研究之餘寫成，恐難免疏漏。若讀者有任何指正，當在再更正。

溫志皓

Chapter 01 資料探勘概論

Chapter 02 資料探勘的功能

Chapter 03 資料庫與資料探勘 – 大資料 I

Chapter 04 資料與資料探勘 – 大數據 II

Chapter 05 決策樹：C5.0

Chapter 06 分類與迴歸樹: C&RT

Chapter 07 因數分析: FA/PCA

CHAPTE

資料探勘

學習目

- 瞭解資料探勘的概
- 瞭解何謂資料探勘
- 瞭解電腦資訊系統
- 瞭解資料探勘與統
- 瞭解何謂資料庫中
- 瞭解資料探勘的特
- 瞭解資料探勘的定
- 不同資料探勘定義
- 瞭解資料探勘的流
- 不同資料探勘流程
- 瞭解資料探勘的應
- 瞭解資料探勘的發