IBM SPSS Neural Networks 19

Note: Before using this information and the product it supports, read the general information under Notices第 84 頁.

序

IBM® SPSS® Statistics為分析資料的強大系統。神經網路 的選用性附加模組能提供其他本手冊所說明的分析技術。神經網路 的附加模組必須與 SPSS Statistics Core 系統搭配使用，而且是完全整合到系統中。

關於 SPSS Inc.，是一家 IBM 公司

SPSS Inc.，是一家 IBM 公司，為全球領先的預測分析軟體和解決方案供應商。該公司完整的系列產品 — 資料收集、統計量、模型製造與部署 — 捕捉人們的態度和意見，預測客戶未來的互動結果，然後將分析融入業務程序，以依照所得見解採取行動。SPSS Inc. 解決方案藉由著重於收斂性分析、IT 架構和業務程序，以達成整個組織相互關聯的業務目標。全球商業、政府和學界客戶均仰賴 SPSS Inc. 技術為競爭優勢，以吸引、留住和增加客戶人數，同時減少欺詐並降低風險。SPSS Inc. 在 2009 年 10 月由 IBM 收購。如需詳細資訊，請造訪 http://www.spss.com。

技術支援

技術支援可提供客戶維護的服務。客戶可以電洽技術支援以取得 SPSS Inc. 產品在使用上的協助，或是支援硬體環境的安裝說明。如果要聯絡技術支援，請參閱 SPSS Inc. 網站（網址是 http://support.spss.com），或是透過網站（網址是 http://support.spss.com/default.asp?refpage=contactus.asp）尋找當地的辦事處。請求協助時，請準備好的您個人、組織和支援合約的相關資訊。

客戶服務

如果您對於自己的貨品或帳號有任何疑問，請聯絡您的當地辦公室，列示於網站上：http://www.spss.com/worldwide。請備妥您的序號以供識別。

訓練研討會

SPSS Inc. 同時提供公開與線上訓練研討會。所有的研討會皆以傳達工作群為其特色。研討會將定期在各主要城市舉辦。如需有關這些研討會的更多資訊，請聯絡您的當地辦公室，列示於網站上：http://www.spss.com/worldwide。

其他出版品

SPSS Statistics：Guide to Data Analysis（資料分析指南）、SPSS Statistics：Statistical Procedures Companion（統計程序指南）以及 SPSS Statistics：Advanced Statistical Procedures Companion（進階統計程序指南）是由 Marija Norušis 撰寫，

由 Prentice Hall 發行，為推薦的輔助資料。這些出版品涵蓋 SPSS Statistics Base 模組、進階統計量模組和迴歸模組中的統計程序。不論您是資料分析的新手，還是已經準備使用高階應用程式，這些書籍都能幫助您善加利用 IBM® SPSS® Statistics 系列產品中的功能。如需其他資訊（包括出版品內容和章節樣本），請參閱作者的網站： http://www.norusis.com

內容

部 II: 範例

附錄

部 I:
使用手冊

章節
1

神經網路簡介

神經網路因為功能強大、彈性和易用，所以是許多預測性資料採礦應用程式較常使用的工具。預測性神經網路在基礎程序複雜的應用程式特別有用，例如：

- 預測客戶需求以便將製作模式和交付成本合理化。
- 預測直效郵遞行銷回應的機率，以決定郵寄名單中的哪些家庭應該寄送優惠訊息。
- 為申請人評分，以決定申請人之擴充信用的風險。
- 在保險請求給付資料庫中偵測詐欺交易。

用於預測性應用程式的神經網路，像是**多層認知 (MLP)** 和**半徑式函數 (RBF)** 網路，會因為模式預測結果可與目標變數的已知數值相比較而加以監督。神經網路選項可以讓您調適 MLP 和 RBF 網路，並儲存所產生的模式以供評分。

什麼是神經網路？

神經網路一詞源於大腦功能研究，適用於鬆散相關的模式系列，特色是大型參數空間與彈性結構。隨著這個系列的增長，大多數新模式的設計目的在於，透過反應其來源的相關詞彙，進行非生物應用。

神經網路的特定定義依所使用的領域而異。雖然沒有單一定義能正確涵蓋這整個模式系列，但現在可考慮下列說明(Haykin, 1998)：

神經網路是一個大規模的平行分散式處理程式，具有儲存經驗知識並加以運用的本性。在兩方面類似大腦：

- 這個網路透過學習程序取得知識。
- 名為突觸權重的中間神經元連接強度可用來儲存知識。

如需為何此定義或許限制過多的討論，請參閱 (Ripley, 1996)。

為了從使用此定義的傳統統計方法區分出神經網路，我們沒說的部分正如定義文字本身一般明顯。例如，傳統線性迴歸模式會從最小平方 方法獲得知識，並將知識儲存在迴歸係數。這在意義上而言就是神經網路了。事實上，您可爭論線性迴歸是否為特定神經網路的特殊案例。但線性迴歸從資料學習之前，就已經有穩固的模式結構和一組假設。

相反地，上述定義卻對模式結構和假設做出最低要求。因此，神經網路可以近似廣泛的統計模式，無須您預先假設依變數與自變數之間的特定關係。反之，這些關係形式會在學習過程中決定。若依變數與自變數之間的線性關係適當，神經網路結果應近似於線性迴歸模型的結果。若非線性關係較適當，神經網路會自動近似於「正確」模式結構。

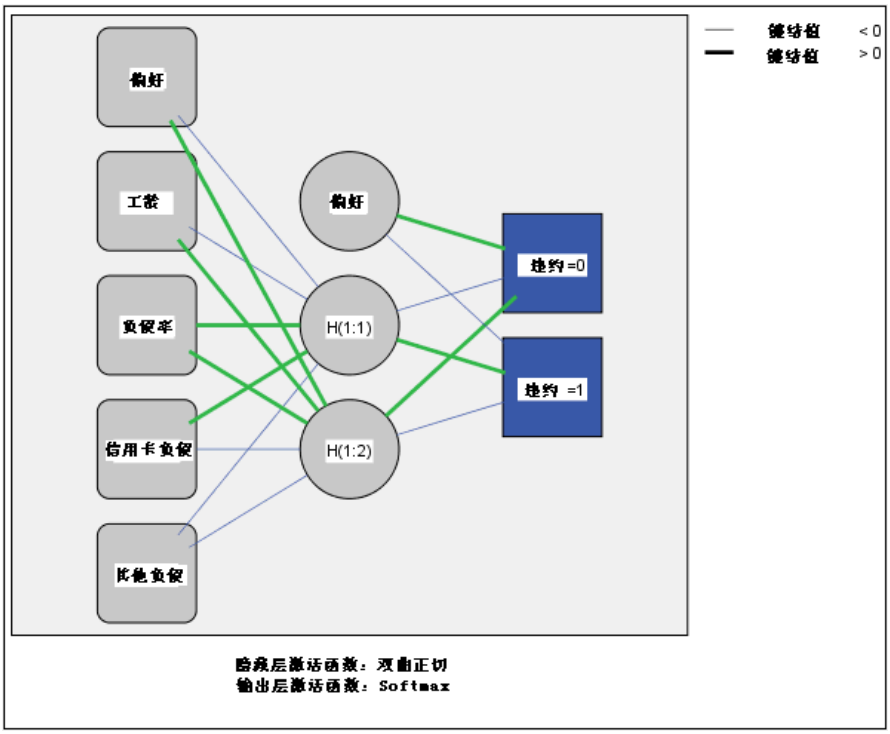
此彈性消長也就等同於神經網路的突觸權重，不容易解釋。因此，若您嘗試說明的基礎程序會產生依變數與自變數之間的關係，最好使用較傳統的統計模式。但若模式解釋能力不是很重要，則使用神經網路通常可較快取得不錯的模式結果。

神經網路結構

雖然神經網路在模式結構和假設上的要求很低， 但若能瞭解一般**網路架構**還是很有用。多層認知 (MLP) 或半徑式函數 (RBF) 網路是將目標變數 (也稱為輸出) 的預測誤差降至最低的預測變數 (也稱為輸入或自變數) 的函數。

請考慮在產品隨附的資料集 bankloan.sav 中，您想要從一堆貸款申請人找出可能的拖欠者。套用在這個問題上的 MLP 或 RBF 網路，就是可將預測預設值的誤差降至最低的測量函數。以下圖形有助於瞭解此函數的相關形式。

圖表 1-1
具有一個隱藏階層的前饋架構



此結構稱為**前饋架構**，因為在網路中的連線未經任何反饋迴圈便從輸入階層向前流至輸出階層。在本圖中：

- **輸入階層**包含預測變數。
- **隱藏階層**包含無法觀察的節點或單元。每一個隱藏單元的值都是預測變數的某種函數；此函數的精確形式部分取決於網路類型，部分取決於使用者可控制的規格。
- **輸出階層**包含反應值。由於預設值的歷程是含有兩個類別的類別變數，因此將它重新編碼為兩個指標變數。每一個輸出單元都是隱藏單元的某種函數。同樣地，此函數的精確形式部分取決於網路類型，部分取決於使用者可控制的規格。

MLP 網路可容許第二個隱藏階層；在此狀況下，第二個隱藏階層的每一個單元都是在第一個隱藏階層中該單元的函數，而每一個反應值都是在第二個隱藏階層中該單元的函數。

章節
2

多層感知器

「多層感知器」(MLP) 程序會依據預測變數值，為 1 個或多個依（目標）變數產生預測模式。

範例。 下列是兩個使用 MLP 程序的情況：

銀行放貸人員必須能辨識具有哪些特質的人可能會拖欠貸款，並使用這些特質來識別好和壞的信用風險。使用過去客戶作為樣本，她可以使用過去客戶的保留樣本來訓練多層感知器和驗證分析，然後使用網路來將準客戶分類為好的與壞的信用風險。

醫院系統著重於追蹤成本和入院接受心肌梗塞 (MI，或「心臟病」) 治療之病患的住院日數。取得這些量數的正確估計值可讓管理人員在病患接受治療的同時，正確地管理可用的空床。使用接受 MI 治療之病患的治療記錄樣本，管理人員可訓練網路來預測成本和住院日數。

依變數。 依變數可以是：

- **名義。** 當變數數值代表實質上並未等級化的類別時（例如，有員工工作的公司部門），則此變數可視為名義。名義變數的範例包括地區、郵遞區號以及宗教團體。
- **次序。** 當變數數值代表實質上已等級化的類別時（例如，服務滿意度從非常不滿意到非常滿意分級），則此變數可視為次序。次序變數的範例包括代表滿意度或信賴程度的態度分數、以及偏好等級分數。
- **尺度。** 若一變數可視為尺度（連續），表示它的的數值代表含有實際意義矩陣的已排列順序類別，因此適合比較數值之間的距離。尺度變數的範例包括以年份表示的年齡和以千元為單位的收入。

  本程序假設已指定給所有依變數適當的測量水準，但您可以在來源變數清單的變數上按一下滑鼠右鍵，並選取快顯功能表上的測量水準，暫時變更變數的測量水準。

變數清單中各變數旁的圖示可識別測量水準和資料類型：

測量水準(E)	**資料類型**			
	數字的	字串	日期	時間
尺度（連續）		無		
次序				
名義				

預測值變數。 可指定預測值為因子（類別）或共變量（尺度）。

類別變數編碼。本程序在整個程序期間，會使用 one-of-c 編碼來暫時記錄類別預測變數和依變數。如果一個變數有多個 c 類別，則變數會儲存為 c 向量，其中第一個類別標示為 (1,0,...,0)，第二個類別標示為 (0,1,0,...,0)，...，最後一個類別標示為 (0,0,...,0,1)。

此編碼架構會增加加權鍵結值的數目，這會導致訓練變慢；然而，更多「精簡」的編碼方法通常會產生不適合的神經網路。如果您的網路訓練進行的非常慢，您可以將類似的類別組合在一起，或捨棄具有極少類別的觀察值，以嘗試減少類別預測值中的類別個數。

所有的 one-of-c 編碼都是以訓練資料為基礎，即使已定義測試或保留樣本也是如此(請參閱分割 第 7 頁)。因此，如果測試或保留樣本所包含的觀察值之預測值類別不在訓練資料中，則這些觀察值不會用於程序或評定中。如果測試或保留樣本所包含的觀察值之依變數類別不在訓練資料中，則這些觀察值不會用於程序，但可能用於評分中。

調整。 尺度依變數和共變量會依照預設值進行調整，以改善網路訓練。所有的調整都是以訓練資料為基礎，即使已定義測試或保留樣本也是如此（請參閱分割 第 7 頁）。也就是視調整的類型而定，僅使用訓練資料來計算平均數、標準差、共變量或依變數的最小值或最大值。 如果您指定變數來定義分割，很重要的是這些共變量或依變數必須在訓練樣本、測試樣本和保留樣本之間有類似的分配。

次數加權。 此程序會忽略次數加權。

複製結果。 如果您要精確地複製結果，除了使用相同的程序設定外，請為亂數產生器使用相同的初始化值、相同的資料順序和相同的變數順序。此問題的詳細資料如下：

- **亂數產生器。** 程序會在隨機指派分割、將加權鍵結值初始化隨機分成次樣本、將自動架構選擇隨機分成次樣本的期間內使用亂數產生器，以及使用用於加權初始化和自動架構選擇的模擬退火演算法。未來若要重新產生相同的隨機化結果，請先使用與亂數產生器的相同初始化值，再執行每個「多層感知器」程序。請參閱準備進行分析所用的資料第 31 頁以取得逐步的指示。
- **觀察值順序。** 線上和小型批次訓練方法（請參閱訓練 第 10 頁）明確地依存於觀察值的順序；然而，即使是批次訓練也依存於觀察值的順序，因為加權鍵結值的初始化包含從資料集中抽出次樣本。

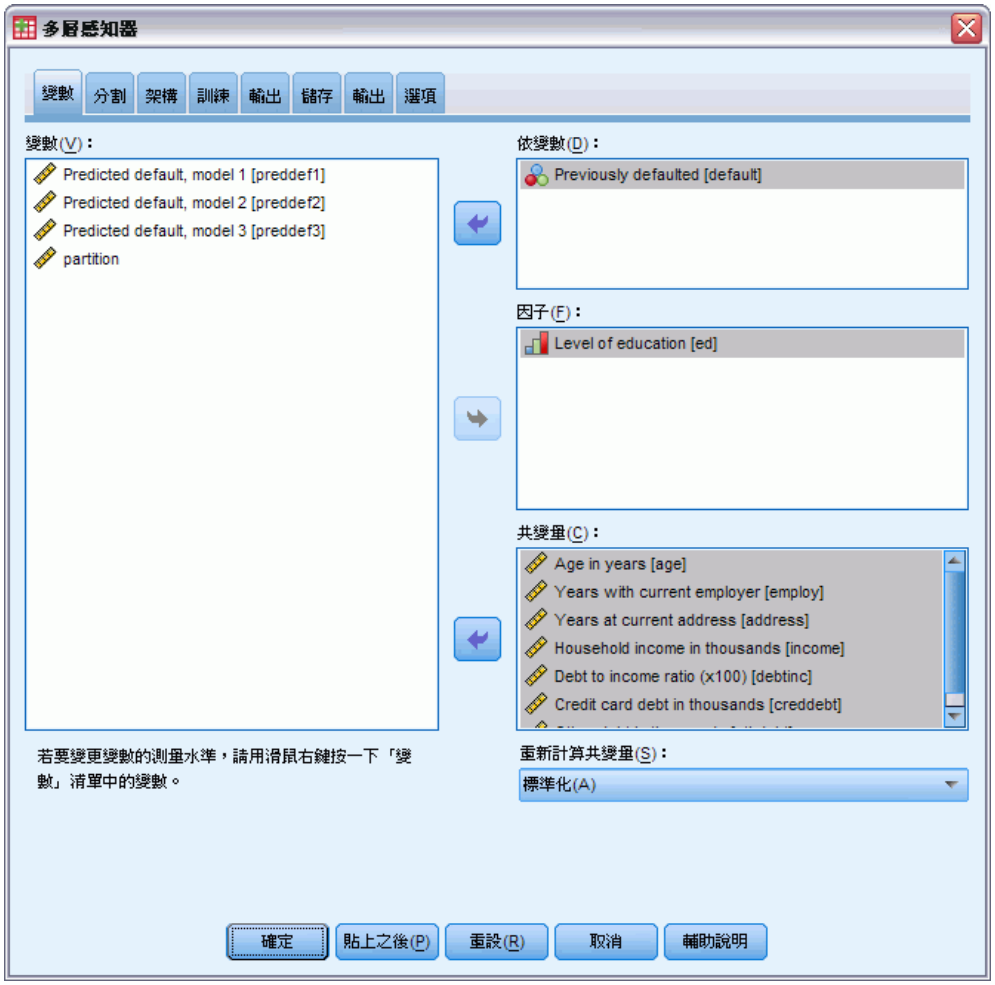
  若要將順序效應降到最低，請以隨機方式排列觀察值。若要驗證某個解決方案的穩定性，您也許會想要取得幾種不同的解決方案，其觀察值皆以不同的隨機順序排列。在檔案極大的情況下，可進行多次運算，以不同的隨機順序排列一個觀察值的樣本。
- **變數順序。** 結果可能會受到因子和共變量清單中的變數順序所影響，因為當變數順序變更時，會指派不同的初始值樣式。如同觀察值順序的影響，您可以嘗試不同的變數順序（只要在因子和共變量清單中進行拖放）以評估給定解答的穩定性。

建立多層感知器網路

從功能表選擇：

分析(A) > 神經網路 > 多層感知器...

圖表 2-1
多層感知器：「變數」索引標籤



- ▶ 選取至少一個依變數。
- ▶ 選取至少一個因子或共變量。

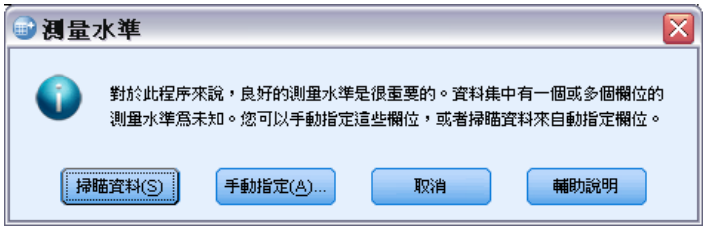
您可以選擇性地在「變數」索引標籤上變更調整共變量的方法。選項為：

- **標準化。** 減去平均數，然後除以標準差 (x−mean)/s。
- **常態化。** 減去最小值，然後除以範圍 (x−min)/(max−min)。常態化的值介於 0 和 1 之間。
- **調整後常態化。** 減去最小值，然後除以範圍的調整版本 [2*(x−min)/(max−min)]−1。調整後常態化的數值介於 −1 和 1 之間。
- **無。** 沒有調整共變量。

具有未知測量水準的欄位

若在資料集中出現一或多個未知的變數（欄位）測量水準，就會顯示「測量水準」警示。由於測量水準會影響此程序的結果計算，因此所有變數皆必須具有已定義的測量水準。

圖表 2-2
測量水準警示

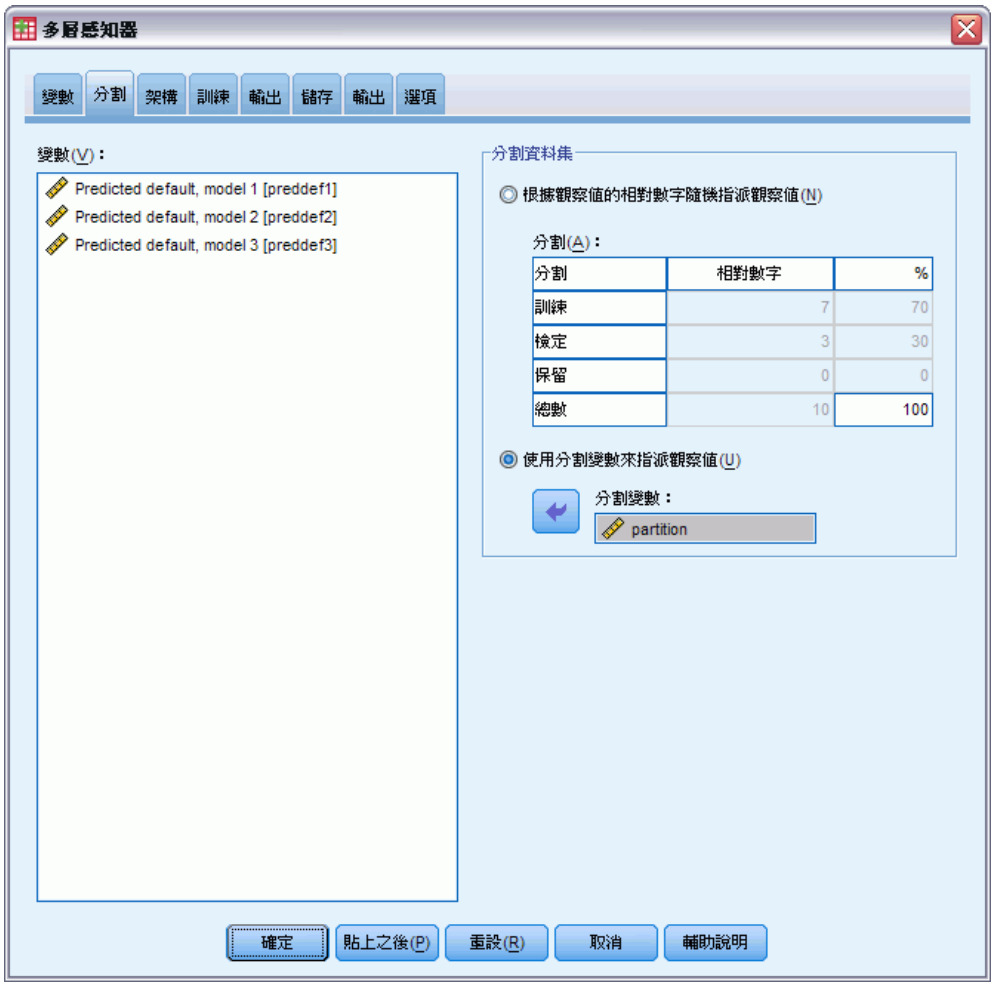


- **掃描資料。** 讀取作用中資料集的資料，並且針對目前具有未知測量水準的任何欄位指派預設的測量水準。若為大型資料集，則讀取時可能需要一些時間。
- **手動指派。** 開啟對話方塊，以列出具有未知測量水準所有欄位。您可以使用此對話方塊，來指派上述欄位的測量水準。您也可以在「資料編輯程式」的「變數檢視」中指派測量水準。

由於測量水準是此程序的重要項目，因此您在所有欄位皆擁有已定義的測量水準之前，無法存取對話方塊來執行此程序。

分割

圖表 2-3
多層感知器：「分割」索引標籤



區隔資料集。 此組別指定將作用中資料集區隔成訓練、測試和保留樣本的方法。 **訓練樣本** 由用來訓練神經網路的資料記錄所組成；在資料集中有某些比例的觀察值必須指定為訓練樣本以取得模式。 **測試樣本** 是一組獨立的資料記錄，用來追蹤訓練期間的錯誤以避免過度訓練。 強烈建議您建立一個訓練樣本，當測試樣本較訓練樣本小的時候，網路訓練通常會最有效率。 **保留樣本** 是另一組獨立的資料記錄，用來存取最後的神經網路；由於保留觀察值並未用來建立模式，保留樣本的錯誤為模式的預測能力提供了「誠實」的估計。

- **依據相對的觀察值個數隨機指定觀察值。** 說明相對的觀察值個數（比例）隨機指定給每一個樣本（訓練、測試和保留）。%> 欄會描述依據您所指定的相對數字，將觀察值百分比指定給每一個樣本。

例如，將訓練、測試和保留樣本的相對數字指定為7、3、0，對應到70%、30%和0%。將相對數字指定為2、1、1，對應到50%、25%和25%；1、1、1對應到將此資料集的訓練、測試和保留均分為三等分。

- **使用分割變數來指定觀察值。** 指定一數值變數將作用中資料集的每一個觀察值指定給訓練、測試或保留樣本。將變數中含有正值的觀察值指定給訓練樣本，將值為 0 的觀察值指定給測試樣本，而將負值的觀察值指定給保留樣本。含有系統遺漏值的觀察值會從分析中排除。任何分割變數的使用者遺漏值永遠視為有效。

  注意：使用分割變數將無法保證在程序連續執行時會產生相同的結果。請參閱主要多層感知器主題中的「複製結果」。

架構

圖表 2-4
多層感知器：「架構」索引標籤

「架構」索引標籤是用來指定網路的架構。程序可自動選取「最佳」的架構，或是您可以指定自訂的架構。

自動架構選擇會建立包含一個隱藏階層的網路。指定隱藏階層中所允許的單位最小和最大數目，則自動架構選擇會計算隱藏階層中「最佳」的單位數目。自動架構選擇會使用隱藏和輸出階層的預設啟動函數。

自訂架構選擇讓您可專業地控制隱藏和輸出階層，而且在您事先知道想要的架構或需要扭曲自動架構選擇的結果時最為有用。

隱藏階層

隱藏階層包含無法觀察的網路節點（單元）。每個隱藏單元皆為輸入加權總和的函數。此函數為啟動函數，而加權值是以估計演算法來決定。若網路包含第二個隱藏階層，則在此第二個階層中的每個隱藏單元皆為第一個隱藏階層中的單元加權總和函數。在兩個階層中會使用相同的啟動函數。

隱藏階層的數目。 一個多層感知器可以有一個或兩個隱藏階層。

啟動函數。 啟動函數會將階層中的單位加權總和「連結」至後續階層中的單位數值。

- **雙曲正切。** 此函數的形式為：$\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$。必須使用實值的引數並將它們轉換至範圍（-1、1）。當使用自動架構選擇時，此為隱藏階層中所有單位的啟動函數。
- **Sigmoid。** 此函數的形式為：$\gamma(c) = 1/(1+e^{-c})$。必須使用實值的引數並將它們轉換至範圍（0、1）。

單位數。 您可以明確指定每個隱藏階層中的單位個數，也可以讓估計演算法自動決定。

輸出階層

輸出階層包含目標變數（依變數）。

啟動函數。 啟動函數會將階層中的單位加權總和「連結」至後續階層中的單位數值。

- **單位。** 此函數的形式為：$\gamma(c) = c$。必須使用實值的引數並以原樣傳回。當使用自動架構選擇時，如果有任何尺度依變數，則此為輸出階層中所有單位的啟動函數。
- **Softmax。** 此函數的形式為：$\gamma(c_k) = \exp(c_k)/\Sigma_j \exp(c_j)$。必須使用實值引數的向量並將它轉換至其元素落在 (0, 1) 範圍且總和為 1 的向量。Softmax 只有在所有依變數為類別時才能使用。當使用自動架構選擇時，如果所有依變數為類別，則此為輸出階層中所有單位的啟動函數。
- **雙曲正切。** 此函數的形式為：$\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$。必須使用實值的引數並將它們轉換至範圍（-1、1）。
- **Sigmoid。** 此函數的形式為：$\gamma(c) = 1/(1+e^{-c})$。必須使用實值的引數並將它們轉換至範圍（0、1）。

調整尺度依變數。 只有在已選取至少一個尺度依變數時，才能使用這些控制項。

- **標準化。** 減去平均數，然後除以標準差 (x−mean)/s。
- **常態化。** 減去最小值，然後除以範圍 (x−min)/(max−min)。常態化的值介於 0 和 1 之間。如果輸出階層使用 Sigmoid 啟動函數，這會是尺度依變數所需的調整方法。修正選項指定可套用為修正調整公式的一個小數目 ε；此修正可確保所有調整的依變數值會在啟動函數的範圍內。尤其是數值 0 和 1 會定義 Sigmoid 函數的

範圍限制，但卻不在該範圍內，且當 x 為最小值和最大值時，數值 0 和 1 會出現在未修正的公式中。修正的公式為 [x−(min− ε)]/[(max+ ε)−(min− ε)]。請指定一個大於或等於 0 的數值。

- **調整後常態化。** 減去最小值，然後除以範圍的調整版本 [2*(x−min)/(max−min)]−1。調整後常態化的值介於 −1 和 1 之間。如果輸出階層使用雙曲正切啟動函數，這會是尺度依變數所需的調整方法。修正選項指定可套用為修正調整公式的一個小數目 ε；此修正可確保所有調整的依變數值會在啟動函數的範圍內。尤其是數值 − 1 和 1 會定義雙曲正切函數的範圍限制，但卻不在該範圍內，且當 x 為其最小值和最大值時，數值 1 和 1 會出現在未修正的公式中)。修正的公式為 {2*[(x−(min− ε))/((max+ ε)−(min− ε))]}−1。請指定一個大於或等於 0 的數值。
- **無。** 沒有調整尺度依變數。

訓練

圖表 2-5
多層感知器：「訓練」索引標籤

「訓練」索引標籤是用來指定訓練網路的方式。訓練類型和最佳化演算法會決定可用的訓練選項。

訓練類型。 練習類型會決定網路處理記錄的方式。選取以下任一訓練類型：

- **批次。** 僅在傳輸所有訓練資料記錄之後更新加權鍵結值；也就是，批次訓練使用訓練資料集中的所有記錄資訊。因為批次訓練可直接降低總錯誤數，因此為較常使用的選項；然而，批次訓練需要更新加權數次直到達到其中一個中止規則，因此可能需要進行多次資料傳輸。這對「較小」的資料集而言最為實用。
- **線上。** 在每一訓練資料記錄之後更新加權鍵結值；也就是，線上訓練一次使用一個記錄資訊。線上訓練會持續取得記錄並更新加權直到達到其中一個中止規則。若所有的記錄全都使用過，但沒有達到任何一個中止規則時，則程序會再次利用資料記錄持續進行。對於具有相關預測值的「較大」資料集而言，線上訓練優於批次訓練；也就是，若有許多記錄和輸入，且其數值並非彼此獨立，則比起批次訓練而言，線上訓練可較快速地取得合理的答案。
- **小型批次。** 將訓練資料記錄分成大約相等大小的組別，並在傳輸一個組別後更新加權鍵結值；也就是，小型批次會使用記錄組別的資料。如有需要，程序會再次利用該資料組別。小型批次是批次和線上訓練之間的折衷方式，且最適合用於「中型」資料集。程序可自動決定每一小型批次的訓練記錄數目，或是您可以指定一個大於 1 但小於或等於儲存在記憶體的觀察值最大數目的整數。您可以在「選項」索引標籤上，設定可儲存在記憶體的觀察值最大數目。

最佳化演算法。 這是用於估計加權鍵結值的方法。

- **尺度化共軛梯度。** 調整共軛梯度使用方法的假設僅適用於批次訓練類型，因此此方法不適用於線上或小型批次訓練。
- **梯度下降。** 此方法必須和線上或小型批次訓練一起使用；也可和批次訓練一起使用。

訓練選項。 訓練選項可讓您微調最佳化演算法。一般不太需要變更這些設定，除非網路在估計時遇到問題。

尺度化共軛梯度演算法的訓練選項包括：

- **初始 Lambda。** 尺度化共軛梯度演算法之 Lambda 參數的初始值。指定一個大於 0 且小於 0.000001 的數值。
- **初始 Sigma。** 尺度化共軛梯度演算法之 Sigma 參數的初始值。指定一個大於 0 且小於 0.0001 的數值。
- **區間中心和區間偏移。** 區間中心 (a_0) 和區間偏移 (a) 會定義區間 $[a_0-a, a_0+a]$，在此區間中使用模擬退火演算法時，會隨機產生加權向量。模擬退火演算法是在最佳化演算法應用期間，為達到尋找整體最小值之目的，作為破壞局部最小值之用。此方法適用於加權初始化和自動架構選擇。為區間中心指定一個數值，並為區間偏移指定一個大於 0 的數值。

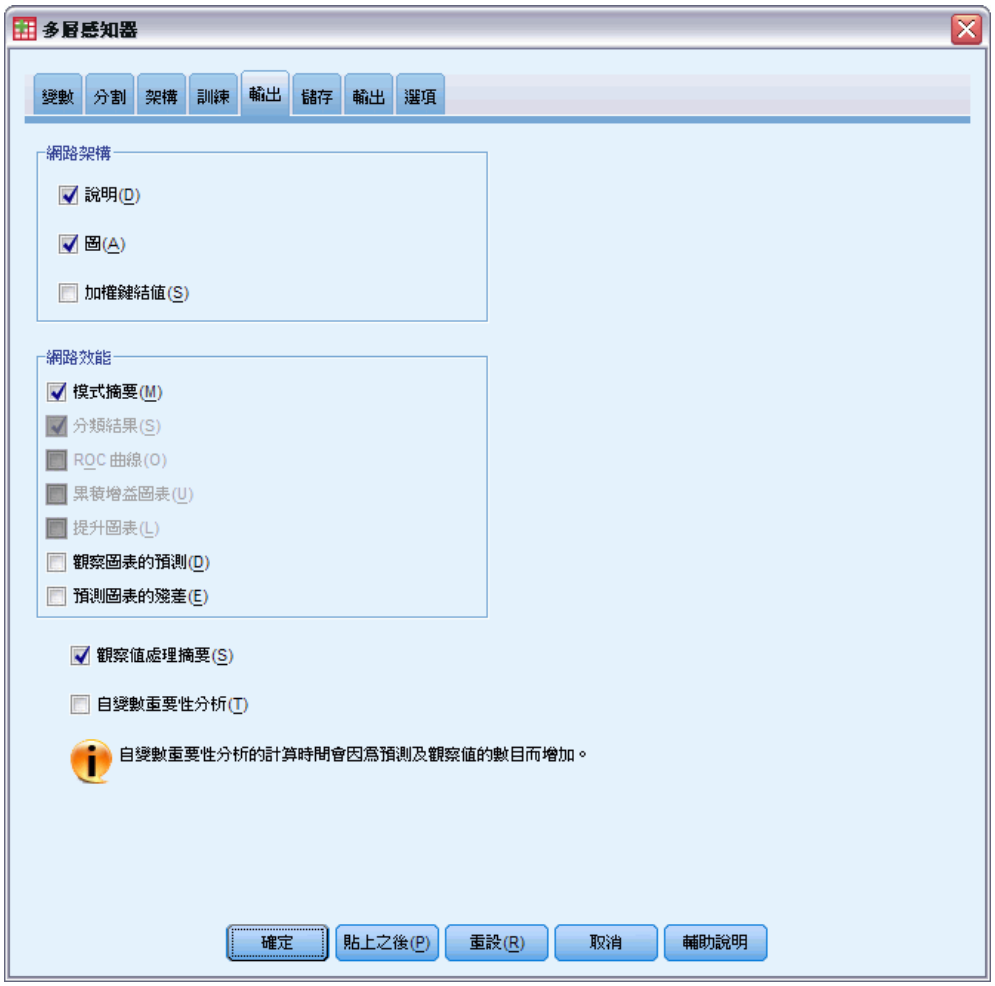
梯度下降演算法的訓練選項包括：

- **初始學習率。** 梯度下降演算法之學習率的初始值。較高的學習率表示網路將訓練更快，但可能造成系統不穩定。指定一個大於 0 的數值。
- **學習率下界。** 梯度下降演算法學習率下界。此設定僅適用於線上和小型批次訓練。指定一個大於 0 但小於初始學習率的數值。

- **動量。** 梯度下降演算法的初始動量參數。動量項目可有助於預防太高學習率所造成的系統不穩定。指定一個大於 0 的數值。
- **學習率縮減（以週期為單位）。** 當梯度下降和線上或小型批次訓練一起使用時，會需要週期的數目 (p) 或訓練樣本的資料傳輸以將初始學習率降低至學習率的下界。此項目可讓您控制學習率減少因子 $\beta = (1/pK)*\ln(\eta_0/\eta_{low})$，其中 η_0 為初始學習率、η_{low} 為學習率的下界，而 K 為訓練資料集中小型批次的總數目（或線上訓練的訓練記錄數目）。請指定一個大於 0 的整數。

輸出

圖表 2-6
多層感知器：「輸出」索引標籤



網路架構。 顯示有關神經網路的摘要資訊。

- **說明。** 顯示有關神經網路的資訊，包括依變數、輸入和輸出單位的數目、隱藏階層和單位的數目，以及啟動函數。

- **圖。** 以無法編輯的圖表來顯示網路結構圖。請注意，隨著共變量和因子水準的數目增加，結構圖也會越來越難解讀。
- **加權鍵結值。** 顯示係數估計值，此估計值會顯示已知階層中的單位和下一階層中的單位之間的關係。即使將作用中資料集分割成訓練、測試和保留資料，加權鍵結值仍會以訓練樣本為基礎。請注意，加權鍵結值的數目可以變得相當大，因此這些加權一般不適用於解讀網路結果。

網路效能。 顯示用於判斷模式是否「良好」的結果。注意：此組別中的圖表是以合併訓練和測試樣本為基礎，或是如果沒有測試樣本時，僅以訓練樣本為基礎。

- **模式摘要。** 以分割和整體方式顯示神經網路結果的摘要，包括錯誤、不正確預測的相對錯誤或百分比、用於中止訓練的中止規則和訓練時間。

  當識別函數、Sigmoid 函數或雙曲正切啟動函數套用在輸出階層時，錯誤為平方和錯誤。當 Softmax 啟動函數套用在輸出階層時，錯誤為交叉熵誤差。

  根據依變數測量水準來顯示不正確預測的相對錯誤或百分比。如果任一依變數具有尺度測量水準，則會顯示平均整體相對錯誤 (相對於平均數模式)。如果所有的依變數為類別，則會顯示不正確預測的平均百分比。也會顯示個別依變數之不正確預測的相對錯誤或百分比。
- **分類結果。** 以分割和整體方式來顯示每個類別依變數的分類表。每個表會提供每個依變數類別之正確或不正確分類的觀察值數目。也會報告已正確分類之總觀察值的百分比。
- **ROC 曲線。** 顯示每個類別依變數的 ROC (接收器作業特性) 曲線。也會顯示可提供每個曲線下之區域的表格。對一個特定的依變數來說，ROC 會為每個類別顯示一個曲線。如果依變數有兩種類別，則每個曲線會視討論中的類別為正向狀態和另一個類別進行比較。如果依變數有兩種以上的類別，則每個曲線會視討論中的類別為正向狀態和其他的類別整合進行比較。
- **累積增益圖表。** 顯示每個類別依變數的累積增益圖表。為每個依變數類別顯示一條曲線的方法和 ROC 曲線的方法相同。
- **提升圖表。** 顯示每個類別依變數的提升圖表。為每個依變數類別顯示一條曲線的方法和 ROC 曲線的方法相同。
- **觀察圖表的預測。** 顯示每個依變數的依觀察值預測圖表。如果是類別依變數，則會顯示每個回應值類別之預測虛擬機率的集群盒形圖，其中以觀察回應值類別作為集群變數。如果是尺度依變數，則會顯示散佈圖。
- **預測圖表的殘差。** 顯示每個尺度依變數的依預測值殘差圖表。殘差與預測值之間應該沒有可見的樣式。僅為尺度依變數產生此圖表。

觀察值處理摘要。 顯示觀察值處理摘要表，其中摘要出分析中包含和排除的觀察值數、總數，以及是依訓練、測試和保留樣本包含和排除。

自變數重要性分析。 執行敏感度分析，此分析會計算在決定神經網路時每個預測值的重要性。分析是以合併訓練和測試樣本為基礎，或是如果沒有測試樣本時，僅以訓練樣本為基礎。這會建立可顯示每個預測值之重要性和常態化重要性的表格和圖表。請注意，如果有大量的預測值或觀察值，則敏感度分析的計算會很昂貴且耗時。

儲存

圖表 2-7
多層感知器：「儲存」索引標籤

「儲存」索引標籤是用來將預測值儲存為資料集中的變數。

- **儲存每個依變數的預測值或類別。** 此動作會儲存尺度依變數的預測值和類別依變數的預測類別。
- **儲存每個依變數的預測虛擬機率或類別。** 此動作會儲存類別依變數的預測虛擬機率。對於前 n 個類別，系統會為每個類別儲存個別的變數，其中 n 是在「**要儲存的類別**」行中指定。

已儲存變數的名稱。 自動名稱產生會確保您能保留所有的工作。自訂名稱可讓您捨棄/取代先前執行的結果，而不必先刪除在「資料編輯程式」中儲存的變數。

機率和虛擬機率

具有 Softmax 啟動函數和交叉熵錯誤的類別依變數會有每個類別的預測值，而類別中的每個預測值為觀察值歸入類別的機率。

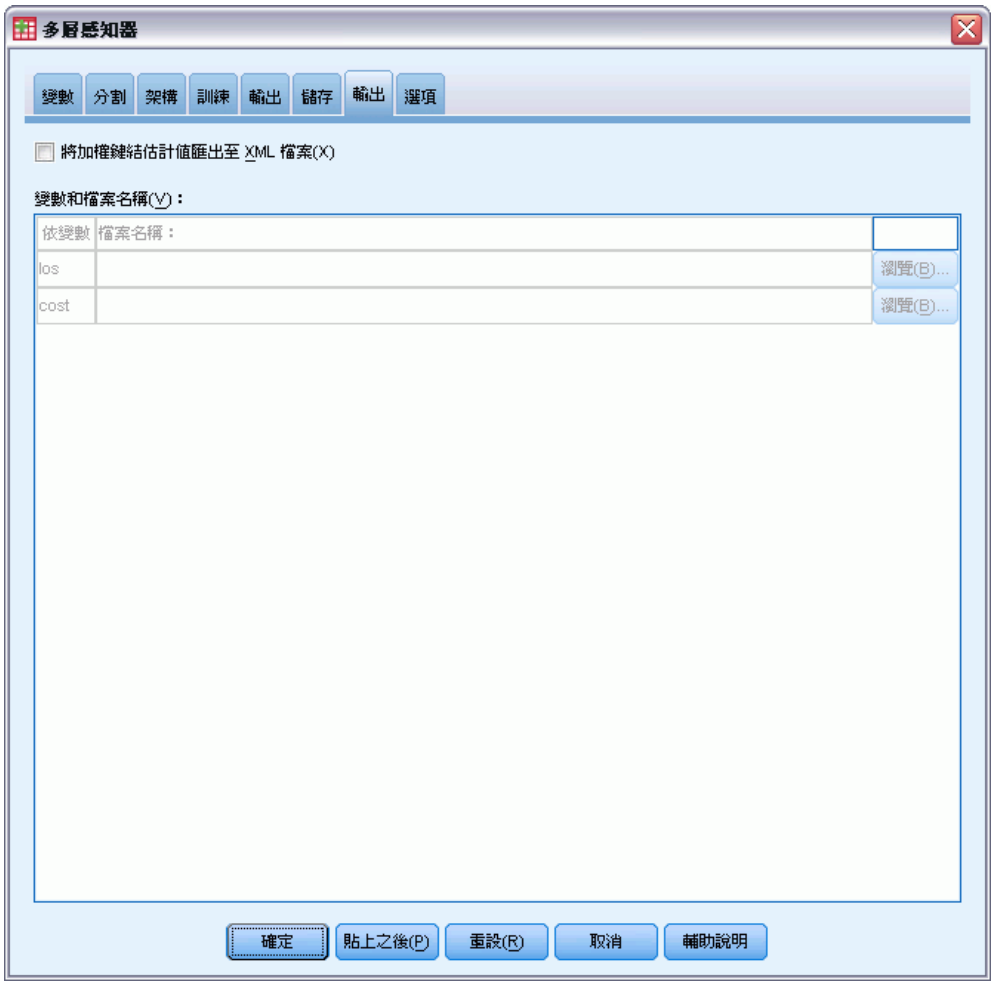
具有平方和錯誤的類別依變數會有每個類別的預測值，但無法將預測值解讀為機率。程序會儲存這些預測虛擬機率，即使任一機率小於 0 或大於 1，或是特定的依變數總和不是 1。

ROC、累積增益和提升圖表（請參閱輸出 第 12 頁）是根據虛擬機率而建立。當任一虛擬機率小於 0 或大於 1，或是特定變數的總和不是 1 時，首先會將機率調整在 0 和 1 之間，且總和為 1。虛擬機率是透過除以其總和來進行調整。例如，如果觀察值具有三類別依變數之 0.50、0.60 和 0.40 的預測虛擬機率，則每個虛擬機率會除以總和 1.50 以得到 0.33、0.40 和 0.27。

如果任一虛擬機率為負數，則在進行上述的調整前，會將最低的絕對值加到所有虛擬機率。例如，如果虛擬機率為 -0.30、0.50 和 1.30，則首先將 0.30 加到每個值以取得 0.00、0.80 和 1.60。接著，將每個新數值除以總和 2.40 以取得 0.00、0.33 和 0.67。

匯出

圖表 2-8
多層感知器：「匯出」索引標籤



「匯出」索引標籤是用來將每個依變數的加權鍵結估計值儲存到 XML (PMML) 檔案。您可以使用這個模式檔案，將模式資訊套用到其他資料檔案中以進行評分工作。 如果已定義分割檔，則此選項無法使用。

選項

圖表 2-9
多層感知器：「選項」索引標籤

使用者遺漏值。 因子必須有有效值，以將觀察值納入分析。這些控制項可讓您決定是否要在因子和類別依變數中，將使用者遺漏值視為有效值。

中止規則。 這些是決定何時停止訓練神經網路的規則。訓練會透過至少一個資料傳輸繼續執行。可依據下列準則來中止訓練，並依照下列順序檢查準則。在遵循的中止規則定義中，步驟會對應至線上和小型批次方法的資料傳輸，以及對應至批次方法的疊代。

- **最大步驟數目 (不包含錯誤縮減)。** 檢查錯誤縮減前所允許步驟數目。如果在指定的步驟數目後沒有出現錯誤縮減，則訓練會中止。指定一個大於 0 的整數。您也可以指定用於計算錯誤的資料樣本。**自動選擇**會使用測試樣本 (如果存在)，否則會使用訓

練樣本。請注意，批次訓練可保證在每個資料傳輸後訓練樣本錯誤會縮減；因此只在測試樣本存在時此選項才適用於批次訓練。**訓練及檢定資料兩者**會為這些樣本檢查錯誤；只在測試樣本存在時才可以使用此選項。

注意：在每個完整資料傳輸後，線上和小型批次訓練需要進行額外的資料傳輸，以計算訓練錯誤。這個額外的資料傳輸會讓訓練明顯變慢，因此通常會建議您提供測試樣本，並選取不論何種情況下「**自動選擇**」。

- **最大訓練時間。** 選擇是否要指定演算法執行的最大分鐘數。指定一個大於 0 的數值。
- **最大訓練週期。** 允許的最大週期數（資料傳輸）。如果超過最大週期數，訓練會中止。請指定一個大於 0 的整數。
- **訓練錯誤中的最小相對變更。** 如果與先前步驟相比時，訓練錯誤中的相對變更小於準則值時，則訓練會中止。指定一個大於 0 的數值。如果是線上或小型批次訓練，只有在使用測試資料計算錯誤時才會忽略此準則。
- **訓練錯誤比例中的最小相對變更。** 如果與零階模式錯誤率相比，訓練錯誤率小於準則值時，則訓練會中止。零階模式會預測所有依變數的平均值。指定一個大於 0 的數值。如果是線上或小型批次訓練，只有在使用測試資料計算錯誤時才會忽略此準則。

要儲存在記憶體中的最大觀察值數目。 此項目可控制下列多層感知器演算法中的設定。請指定一個大於 1 的整數。

- 在自動架構選擇中，用來判斷網路架構的樣本大小是 min(1000,memsize)，其中 memsize 是可儲存在記憶體中的最大觀察值數目。
- 在可自動計算小型批次數目的小型批次訓練中，小型批次的數目是 min(max(M/10,2),memsize)，其中 M 是訓練樣本中的觀察值數目。

章節
3

半徑式函數

「半徑式函數」((RBF) 程序會依據預測變數值，為 1 個或多個依（目標）變數產生預測模式。

範例。 某電信公司根據服務使用方式來切割客戶數量，並將客戶分成四個組別。使用人口資料來預測組別成員的 RBF 網路可讓公司替個別準客戶自訂報價。

依變數。 依變數可以是：

- **名義。** 當變數數值代表實質上並未等級化的類別時（例如，有員工工作的公司部門），則此變數可視為名義。名義變數的範例包括地區、郵遞區號以及宗教團體。
- **次序。** 當變數數值代表實質上已等級化的類別時（例如，服務滿意度從非常不滿意到非常滿意分級），則此變數可視為次序。次序變數的範例包括代表滿意度或信賴程度的態度分數、以及偏好等級分數。
- **尺度。** 若一變數可視為尺度（連續），表示它的的數值代表含有實際意義矩陣的已排列順序類別，因此適合比較數值之間的距離。尺度變數的範例包括以年份表示的年齡和以千元為單位的收入。

  本程序假設已指定給所有依變數適當的測量水準，但您可以在來源變數清單的變數上按一下滑鼠右鍵，並選取快顯功能表上的測量水準，暫時變更變數的測量水準。

變數清單中各變數旁的圖示可識別測量水準和資料類型：

測量水準(E)	資料類型			
	數字的	字串	日期	時間
尺度（連續）		無		
次序				
名義				

預測值變數。 可指定預測值為因子（類別）或共變量（尺度）。

類別變數編碼。本程序在整個程序期間，會使用 one-of-c 編碼來暫時記錄類別預測變數和依變數。如果一個變數有多個 c 類別，則變數會儲存為 c 向量，其中第一個類別標示為 (1,0,...,0)，第二個類別標示為 (0,1,0,...,0)，...，最後一個類別標示為 (0,0,...,0,1)。

此編碼架構會增加加權鍵結值的數目，這會導致訓練變慢；但是，更多「精簡」的編碼方法通常會產生不適合的神經網路。如果您的網路訓練進行的非常慢，您可以將類似的類別組合在一起，或捨棄具有極少類別的觀察值，以嘗試減少類別預測值中的類別個數。

所有的 one-of-c 編碼都是以訓練資料為基礎，即使已定義測試或保留樣本也是如此(請參閱分割 第 22 頁)。因此，如果測試或保留樣本所包含的觀察值之預測值類別不在訓練資料中，則這些觀察值不會用於程序或評定中。如果測試或保留樣本所包含的觀察值之依變數類別不在訓練資料中，則這些觀察值不會用於程序，但可能用於評分中。

調整。 尺度依變數和共變量會依照預設值進行調整，以改善網路訓練。所有的調整都是以訓練資料為基礎，即使已定義測試或保留樣本也是如此(請參閱分割 第 22 頁)。也就是視調整的類型而定，僅使用訓練資料來計算平均數、標準差、共變量或依變數的最小值或最大值。 如果您指定變數來定義分割，很重要的是這些共變量或依變數必須在訓練樣本、測試樣本和保留樣本之間有類似的分配。

次數加權。 此程序會忽略次數加權。

複製結果。 如果您要精確地複製結果，除了使用相同的程序設定外，請為亂數產生器使用相同的初始化值和相同的資料順序。此問題的詳細資料如下：

- **亂數產生器。** 在隨機指派分割期間，此程序會使用亂數產生器。未來若要重新產生相同的隨機化結果，請先使用與亂數產生器的相同初始化值，再執行每個「半徑式函數」程序。請參閱準備進行分析所用的資料第 64 頁以取得逐步的指示。
- **觀察值順序。** 結果也會依存於資料順序，因為使用了 TwoStep 集群演算法來判斷半徑式函數。

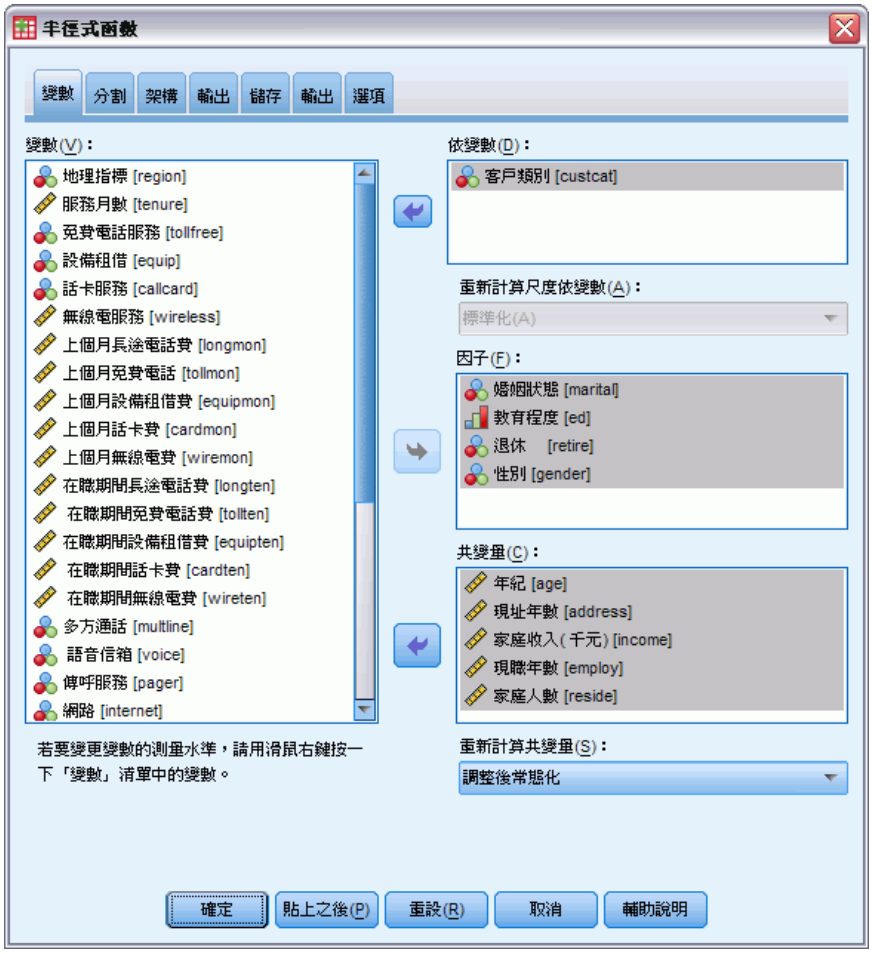
  若要將順序效應降到最低，請以隨機方式排列觀察值。若要驗證某個解決方案的穩定性，您也許會想要取得幾種不同的解決方案，其觀察值皆以不同的隨機順序排列。在檔案極大的情況下，可進行多次運算，以不同的隨機順序排列一個觀察值的樣本。

建立半徑式函數網路

從功能表選擇：

分析(A) > 神經網路 > 半徑式函數...

圖表 3-1
半徑式函數：「變數」索引標籤



- 選取至少一個依變數。
- 選取至少一個因子或共變量。

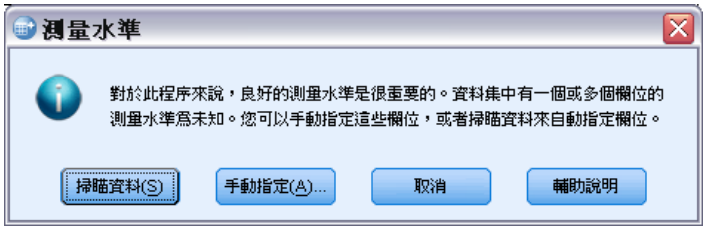
您可以選擇性地在「變數」索引標籤上變更調整共變量的方法。選項為：

- **標準化。** 減去平均數，然後除以標準差 (x−mean)/s。
- **常態化。** 減去最小值，然後除以範圍 (x−min)/(max−min)。常態化的值介於 0 和 1 之間。
- **調整後常態化。** 減去最小值，然後除以範圍的調整版本 [2*(x−min)/(max−min)]−1。調整後常態化的數值介於 −1 和 1 之間。
- **無。** 沒有調整共變量。

具有未知測量水準的欄位

若在資料集中出現一或多個未知的變數（欄位）測量水準，就會顯示「測量水準」警示。由於測量水準會影響此程序的結果計算，因此所有變數皆必須具有已定義的測量水準。

圖表 3-2
測量水準警示

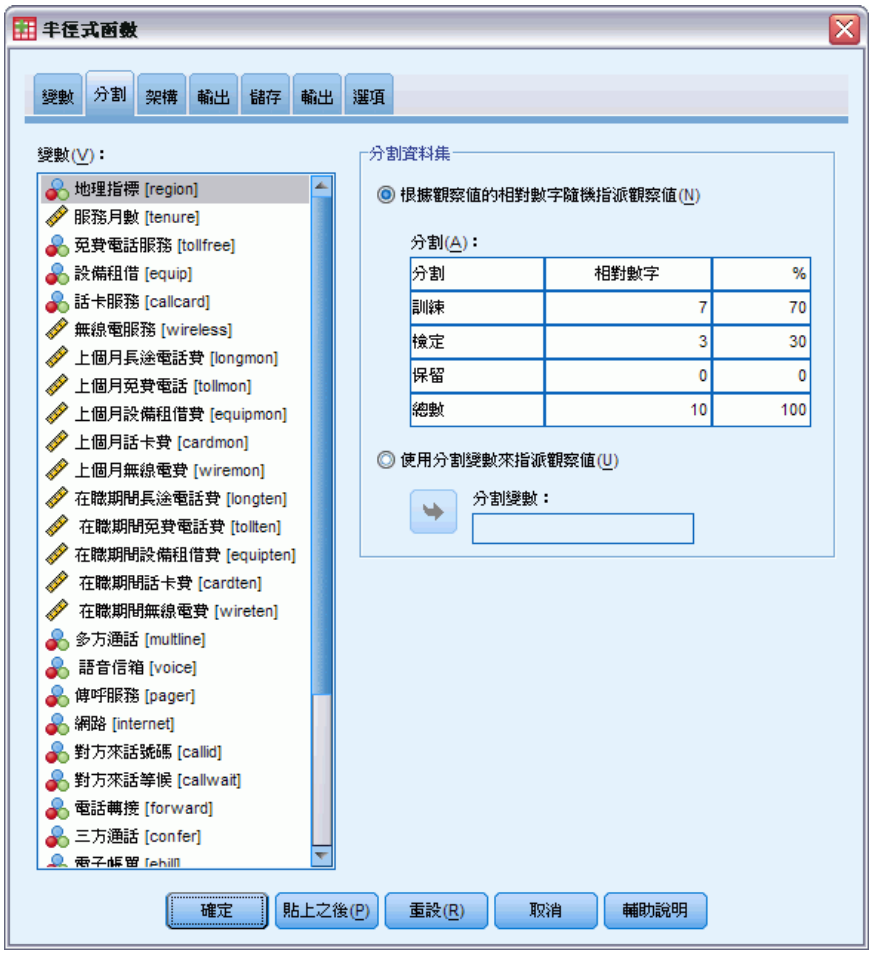


- **掃描資料。** 讀取作用中資料集的資料，並且針對目前具有未知測量水準的任何欄位指派預設的測量水準。若為大型資料集，則讀取時可能需要一些時間。
- **手動指派。** 開啟對話方塊，以列出具有未知測量水準所有欄位。您可以使用此對話方塊，來指派上述欄位的測量水準。您也可以在「資料編輯程式」的「變數檢視」中指派測量水準。

由於測量水準是此程序的重要項目，因此您在所有欄位皆擁有已定義的測量水準之前，無法存取對話方塊來執行此程序。

分割

圖表 3-3
半徑式函數：「分割」索引標籤



區隔資料集。 此組別指定將作用中資料集區隔成訓練、測試和保留樣本的方法。 **訓練樣本** 由用來訓練神經網路的資料記錄所組成；在資料集中有某些比例的觀察值必須指定為訓練樣本以取得模式。 **測試樣本** 是一組獨立的資料記錄，用來追蹤訓練期間的錯誤以避免過度訓練。 強烈建議您建立一個訓練樣本，當測試樣本較訓練樣本小的時候，網路訓練通常會最有效率。 **保留樣本** 是另一組獨立的資料記錄，用來存取最後的神經網路；由於保留觀察值並未用來建立模式，保留樣本的錯誤為模式的預測能力提供了「誠實」的估計。

- **依據相對的觀察值個數隨機指定觀察值。** 說明相對的觀察值個數（比例）隨機指定給每一個樣本（訓練、測試和保留）。%> 欄會描述依據您所指定的相對數字，將觀察值百分比指定給每一個樣本。

例如，將訓練、測試和保留樣本的相對數字指定為7、3、0，對應到70%、30%和0%。將相對數字指定為2、1、1，對應到50%、25%和25%；1、1、1對應到將此資料集的訓練、測試和保留均分為三等分。

- **使用分割變數來指定觀察值。** 指定一數值變數將作用中資料集的每一個觀察值指定給訓練、測試或保留樣本。將變數中含有正值的觀察值指定給訓練樣本，將值為 0 的觀察值指定給測試樣本，而將負值的觀察值指定給保留樣本。含有系統遺漏值的觀察值會從分析中排除。任何分割變數的使用者遺漏值永遠視為有效。

架構

圖表 3-4
半徑式函數：「架構」索引標籤

「架構」索引標籤是用來指定網路的架構。程序會建立包含一個隱藏「半徑式函數」階層的神經網路；一般而言，不需要變更這些設定。

隱藏階層的單位數。 選擇隱藏單位數的方法有三種。

1. **在自動計算的範圍內搜尋最佳單位數。** 程序會自動計算範圍的最小值和最大值，並且在該範圍內搜尋最佳的隱藏單位數。

如果已定義測試樣本，則此程序會使用測試資料準則：最佳的隱藏單位數即為在測試資料中產生最小錯誤的單位數。如果尚未定義測試樣本，則此程序會使用 Bayesian 資訊準則 (BIC)：最佳的隱藏單位數即為依據測試資料產生最小 BIC 的單位數。

2. **在指定範圍內搜尋最佳單位數。** 您可以提供自己的範圍，且程序將在該範圍內搜尋「最佳」的隱藏單位數。與先前的作法相同，範圍中最佳的隱藏單位數是使用測試資料準則或 BIC 來決定。

3. **使用指定的單位數。** 您可以覆寫使用的範圍，並直接指定特定的單位數。

隱藏階層的啟動函數。 隱藏階層的啟動函數為半徑式函數，其會將階層中的單位「連結」至後續階層中的單位數值。對輸出階層而言，啟動函數為識別函數；因此輸出單位僅會是隱藏單元的加權總和。

- **常態化半徑式函數。** 使用 Softmax 啟動函數，以讓所有隱藏單位的啟動常態化成總和為 1。
- **一般半徑式函數。** 使用指數啟動函數，以讓隱藏單位的啟動成為 Gaussian “突出物” 般的輸入函數。

隱藏單位中的重疊。 重疊因子是套用至半徑式函數寬度的乘數。自動計算的重疊因子數值是 1+0.1d，其中 d 是輸入單位數 (所有因子的類別數目與共變量數目的總和)。

輸出

圖表 3-5
半徑式函數：「輸出」索引標籤

網路架構。 顯示有關神經網路的摘要資訊。

- **說明。** 顯示有關神經網路的資訊，包括依變數、輸入和輸出單位的數目、隱藏階層和單位的數目，以及啟動函數。
- **圖。** 以無法編輯的圖表來顯示網路結構圖。請注意，隨著共變量和因子水準的數目增加，結構圖也會越來越難解讀。
- **加權鍵結值。** 顯示係數估計值，此估計值會顯示已知階層中的單位和下一階層中的單位之間的關係。即使將作用中資料集分割成訓練、測試和保留資料，加權鍵結值仍會以訓練樣本為基礎。請注意，加權鍵結值的數目可以變得相當大，因此這些加權一般不適用於解讀網路結果。

網路效能。 顯示用於判斷模式是否「良好」的結果。注意：此組別中的圖表是以合併訓練和測試樣本為基礎，或是如果沒有測試樣本時，僅以訓練樣本為基礎。

- **模式摘要。** 以分割和整體方式顯示神經網路結果的摘要，包括錯誤、不正確預測的相對錯誤或百分比和訓練時間。

錯誤為平方和錯誤。此外，根據依變數測量水準來顯示不正確預測的相對錯誤或百分比。如果任一依變數具有尺度測量水準，則會顯示平均整體相對錯誤（相對於平均數模式）。如果所有的依變數為類別，則會顯示不正確預測的平均百分比。也會顯示個別依變數之不正確預測的相對錯誤或百分比。

- **分類結果。** 顯示每個類別依變數的分類表。每個表會提供每個依變數類別之正確或不正確分類的觀察值數目。也會報告已正確分類之總觀察值的百分比。
- **ROC 曲線。** 顯示每個類別依變數的 ROC（接收器作業特性）曲線。也會顯示可提供每個曲線下之區域的表格。對一個特定的依變數來說，ROC 會為每個類別顯示一個曲線。如果依變數有兩種類別，則每個曲線會視討論中的類別為正向狀態和另一個類別進行比較。如果依變數有兩種以上的類別，則每個曲線會視討論中的類別為正向狀態和其他的類別整合進行比較。
- **累積增益圖表。** 顯示每個類別依變數的累積增益圖表。為每個依變數類別顯示一條曲線的方法和 ROC 曲線的方法相同。
- **提升圖表。** 顯示每個類別依變數的提升圖表。為每個依變數類別顯示一條曲線的方法和 ROC 曲線的方法相同。
- **觀察圖表的預測。** 顯示每個依變數的依觀察值預測圖表。如果是類別依變數，則會顯示每個回應值類別之預測虛擬機率的集群盒形圖，其中以觀察回應值類別作為集群變數。如果是尺度依變數，則會顯示散佈圖。
- **預測圖表的殘差。** 顯示每個尺度依變數的依預測值殘差圖表。殘差與預測值之間應該沒有可見的樣式。僅為尺度依變數產生此圖表。

觀察值處理摘要。 顯示觀察值處理摘要表，其中摘要出分析中包含和排除的觀察值數、總數，以及是依訓練、測試和保留樣本包含和排除。

自變數重要性分析。 執行敏感度分析，此分析會計算在決定神經網路時每個預測值的重要性。分析是以合併訓練和測試樣本為基礎，或是如果沒有測試樣本時，僅以訓練樣本為基礎。這會建立可顯示每個預測值之重要性和常態化重要性的表格和圖表。請注意，如果有大量的預測值或觀察值，則敏感度分析的需要大量計算且耗時。

儲存

圖表 3-6
半徑式函數：「儲存」索引標籤

「儲存」索引標籤是用來將預測值儲存為資料集中的變數。

- **儲存每個依變數的預測值或類別。** 此動作會儲存尺度依變數的預測值和類別依變數的預測類別。
- **儲存每個依變數的預測虛擬機率。** 此動作會儲存類別依變數的預測虛擬機率。對於前 n 個類別，系統會為每個類別儲存個別的變數，其中 n 是在「要儲存的類別」行中指定。

已儲存變數的名稱。 自動名稱產生會確保您能保留所有的工作。自訂名稱可讓您捨棄或取代先前執行的結果，而不必先刪除在「資料編輯程式」中儲存的變數。

機率和虛擬機率

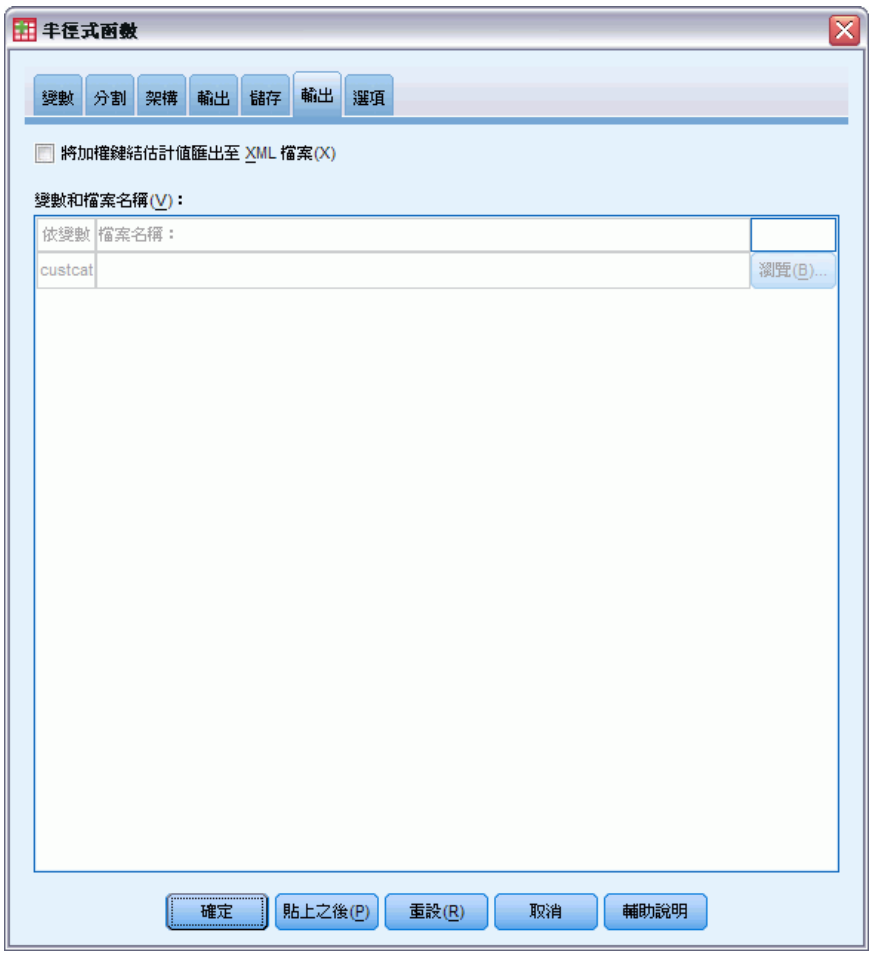
無法將預測虛擬機率解讀為機率，因為「半徑式函數」程序使用平方和錯誤與輸出階層的識別啟動函數。程序會儲存這些預測虛擬機率，即使任一機率小於 0 或大於 1，或是特定的依變數總和不是 1。

ROC、累積增益和提升圖表（請參閱輸出 第 25 頁）是根據虛擬機率而建立。當任一虛擬機率小於 0 或大於 1，或是特定變數的總和不是 1 時，首先會將機率調整在 0 和 1 之間，且總和為 1。虛擬機率是透過除以其總和來進行調整。例如，如果觀察值具有三類別依變數之 0.50、0.60 和 0.40 的預測虛擬機率，則每個虛擬機率會除以總和 1.50 以得到 0.33、0.40 和 0.27。

如果任一虛擬機率為負數，則在進行上述的調整前，會將最低的絕對值加到所有虛擬機率。例如，如果虛擬機率為 -0.30、.50 和 1.30，則首先將 0.30 加到每個值以取得 0.00、0.80 和 1.60。接著，將每個新數值除以總和 2.40 以取得 0.00、0.33 和 0.67。

匯出

圖表 3-7
半徑式函數：「匯出」索引標籤



「匯出」索引標籤是用來將每個依變數的加權鍵結估計值儲存到 XML (PMML) 檔案。您可以使用這個模式檔案，將模式資訊套用到其他資料檔案中以進行評分工作。 如果已定義分割檔，則此選項無法使用。

選項

圖表 3-8
半徑式函數：「選項」索引標籤

使用者遺漏值。 因子必須有有效值，以將觀察值納入分析。這些控制項可讓您決定是否要在因子和類別依變數中，將使用者遺漏值視為有效值。

部 II:
範例

章節
4

多層感知

多層感知（MLP）程序根據預測變數的值為一個或多個依（目標）變數產生預測模式。

使用多層認知評估信用風險

銀行放貸人員必須能辨識具有哪些特質的人可能會拖欠貸款，並使用這些特質來識別好和壞的信用風險。

假設有 850 位以前的客戶與現在的準客戶包含在 bankloan.sav 之中。 前 700 個觀察值為以前有借貸的客戶。使用這 700 位客戶的隨機樣本來建立多層認知，不理會其它的客戶以驗證這個分析。然後使用模式來將 150 位準客戶分類為好的與壞的信用風險。

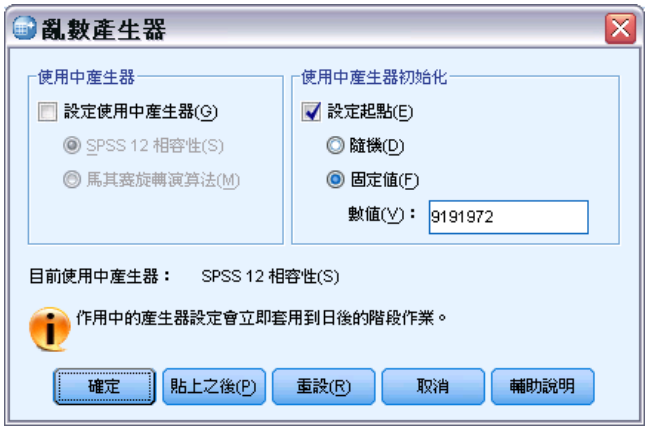
此外，放貸人員先前已使用 Logistic 迴歸（在「迴歸」選項中）進行資料分析，想瞭解多層認知作為分類工具的效用如何。

準備進行分析所用的資料

設定亂數種子可讓您複製出完全相同的分析。

▶ 若要設定亂數種子，從功能表選擇：
轉換 > 亂數產生器...

圖表 4-1
「亂數產生器」對話方塊



▶ 選取「設定起始點」。

▶ 選取「固定值，再輸入「9191972」作為值。

▶ 按一下「確定」。

在前一個 Logistic 迴歸分析中，將大約 70% 的過去客戶指定給訓練樣本，並且將 30% 指定給保留樣本。若要確實重新建立這些分析中使用的樣本，需要分割變數。

▶ 若要建立分割變數，請從功能表選擇：
轉換 > 計算變數...

圖表 4-2
計算變數對話方塊

▶ 在「目標變數」文字方塊中輸入 partition。

▶ 在「數值運算式」文字方塊中輸入 2*rv.bernoulli(0.7)-1。

這會將 partition 的值設定為含有 0.7 機率參數的隨機產生 Bernoulli 變量，並已經過修改，因此會採用值 1 或 −1，而非 1 或 0。叫回將分割變數正值指定給訓練樣本的觀察值、將負值指定給保留樣本的觀察值，以及將值 0 指定給測試樣本的觀察值。此時不會指定測試樣本。

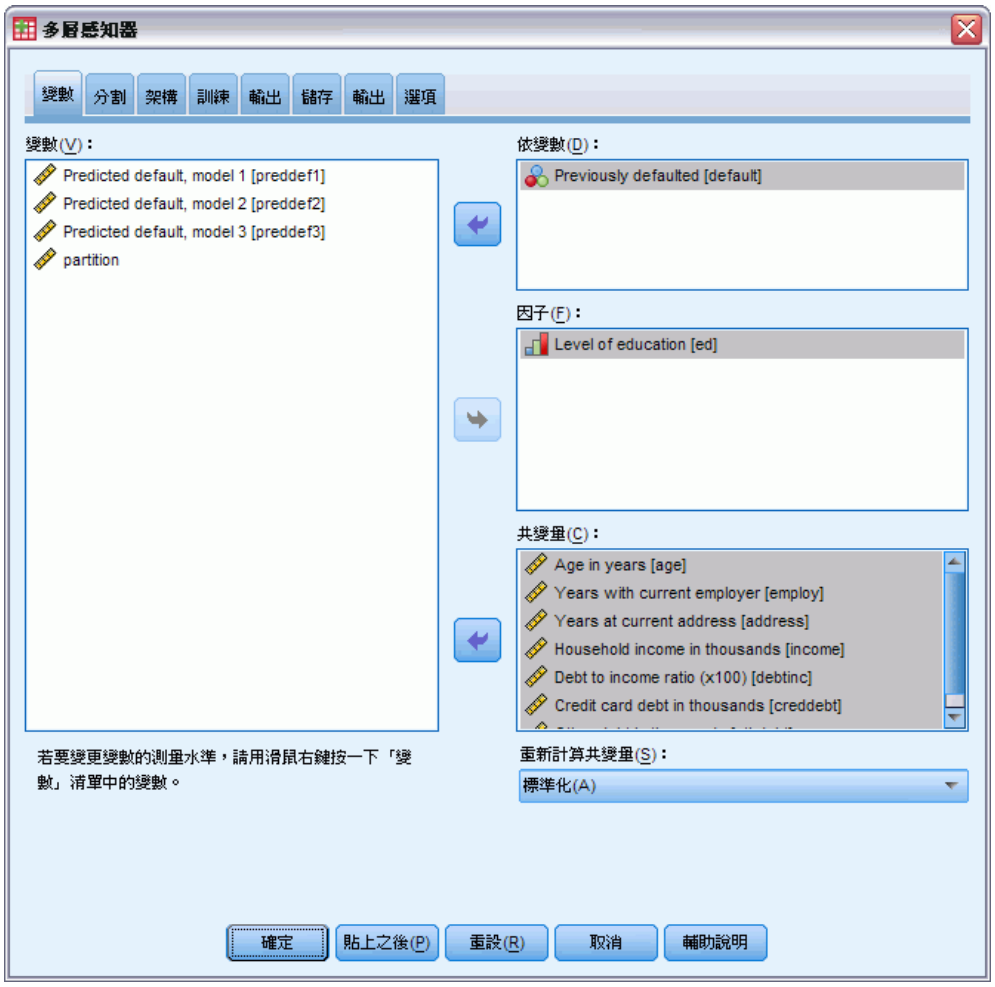
▶ 在「計算變數」對話方塊中，按一下「確定」。

在先前已取得貸款的客戶中，大約有 70% 的 partition 值為 1。這些客戶將用於建立模式。在其他先前已取得貸款的客戶中，如果 partition 值為 −1，則會用於驗證模式的結果。

執行分析

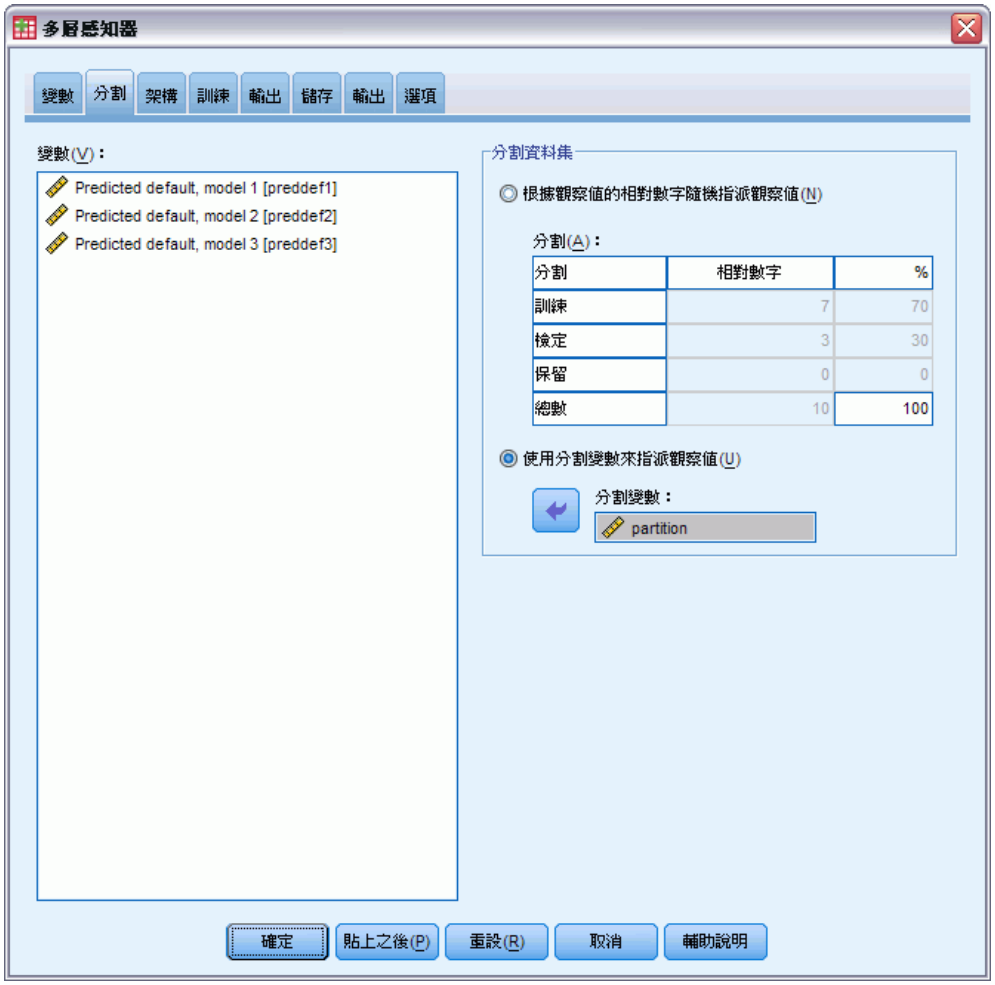
- 若要執行「多層認知」分析，從功能表選擇：
  分析 > 神經網路 > 多層認知...

圖表 4-3
多層認知：「變數」索引標籤



- 選擇「先前已拖欠 [default]」為依變數。
- 選取「教育程度 [ed]」為因子。
- 選擇「年齡 (以年為單位) [age]」到「其他負債 (以千為單位) [othdebt]」做為共變量。
- 按一下「分割」索引標籤。

圖表 4-4
多層認知：「分割」索引標籤



- 選擇「使用分割變數來指定觀察值」。
- 選擇 partition 作為分割變數。
- 按一下「輸出」索引標籤。

圖表 4-5
多層認知：「輸出」索引標籤

- 取消選擇「網路結構」組別中的「圖」。
- 選擇「網路效能」組別中的「ROC 曲線」、「累積增益圖表」、「提升圖表」和「觀察圖表的預測」。由於依變數不是尺度變數，因此無法使用預測圖表的殘差。
- 選擇「自變數重要性分析」。
- 按一下「確定」。

觀察值處理摘要

圖表 4-6
觀察值處理摘要

		個數	百分比
樣本	訓練	499	71.3%
	保留	201	28.7%
	有效	700	100.0%
	排除	150	
	總數	850	

觀察值處理摘要顯示 499 個觀察值指定給訓練樣本，201 個指定給保留樣本。150 個排除在分析之外的觀察值是現在的準客戶。

網路資訊

圖表 4-7
網路資訊

輸入階層	因子	1	Level of education
	共變量	1	Age in years
		2	Years with current employer
		3	Years at current address
		4	Household income in thousands
		5	Debt to income ratio (x100)
		6	Credit card debt in thousands
		7	Other debt in thousands
	因子	單位數	12
		共變量的重新計算方法	標準化
隱藏階層	因子	隱藏階層的數目	1
		隱藏階層 1 的單位數	4
		啓動函數	超正反切
輸出階層	依變數	1	Previously defaulted
	因子	單位數	2
		啓動函數	Softmax
		錯誤函數	交叉熵

a. 排除偏離單元

網路資訊表顯示神經網路的相關資訊，可用於確認規格是否正確。其中必須特別注意：

- 輸入階層的單位個數等於共變量個數加上因子水準總數；對於教育程度的各個類別，會建立個別的單位，而且不會比照許多模式化程序中，將任何類別視為「冗餘」單位。
- 同樣地，對於「先前已拖欠」的各個類別，會建立個別的輸出單位，輸出階層中總共會有兩個單位。
- 自動架構選擇已在隱藏階層中選擇四個單位。
- 其他所有網路資訊都預設用於程序中。

模式摘要

圖表 4-8
模式摘要

訓練	交叉熵錯誤	156.606
	百分比不正確預測	15.6%
	使用的中止規則	超過最大週期數目 (100)
	訓練時間	0:00:00.813
保留	百分比不正確預測	25.4%

依變數： Previously defaulted

模式摘要顯示訓練和套用最終網路於保留樣本的結果資訊。

- 由於輸出階層使用 Softmax 啟動函數，因此會顯示交叉熵錯誤。這是網路嘗試在訓練時最小化的錯誤函數。
- 不正確預測的百分比是從分類表中取得，後續會在該主題中進行討論。
- 由於已達到最大週期數目，因此估計演算法停止。最理想的狀況是，錯誤收斂時，訓練便停止。不過，這會有訓練期間是否發生錯誤與必須注意詳細檢視輸出結果的問題發生。

分類

圖表 4-9
分類

		預測次數		
樣本	觀察次數	No	Yes	百分比修正
訓練	No	347	28	92.5%
	Yes	50	74	59.7%
	整體百分比	79.6%	20.4%	84.4%
保留	No	123	19	86.6%
	Yes	32	27	45.8%
	整體百分比	77.1%	22.9%	74.6%

依變數： Previously defaulted

分類表顯示使用網路的實際結果。對於每個觀察值，如果觀察值的預測虛擬機率大於 0.5，則預測反應為「是」。對每一個樣本而言：

- 觀察值交叉分類對角線上的儲存格為正確預測。
- 不在觀察值交叉分類對角線上的儲存格為不正確預測。

用來建立模式的觀察值中，會正確分類曾經拖欠之 124 個人中的 74 人。會正確分類 375 個非拖欠者中的 347 個人。整體而言，84.4% 的訓練觀察值都經過正確分類，相對於 15.6% 的不正確分類觀察值顯示在模式摘要表中。較好的模式應該要能正確地識別較高百分比的觀察值。

根據使用來建立模式的觀察值來分類，會傾向於太「樂觀」，因為它們的分類比率是誇大的。保留樣本可協助驗證模式；其中 74.6% 的觀察值都經過模式的正確分類。這意味著，整體來說，您的模式事實上四次中有三次是正確的。

更正過度訓練

回想先前執行的 Logistic 迴歸分析時，放貸人員想起訓練樣本和保留樣本都正確預測近似的觀察值百分比，大約是 80%。相反的，神經網路的訓練樣本正確觀察值百分比偏高，其中的保留樣本在預測實際已拖欠的客戶時，表現相當不佳（保留樣本的 45.8% 正確度相較於訓練樣本的 59.7% 正確度）。結合模式摘要表中報告的停止規則後，這會讓您懷疑網路可能**過度訓練**；也就是說，網路採用隨機變異的方式求取訓練資料中不可信的形式。

所幸，解決方法比較簡單：指定測試樣本協助維持網路「穩定」。我們已經建立分割變數，因此這會正確重新建立 Logistic 迴歸分析中使用的訓練樣本和保留樣本；然而，Logistic 迴歸中不存在「測試」樣本的概念。此時必須將一部份的訓練樣本重新指定給測試樣本。

建立測試樣本

圖表 4-10
計算變數對話方塊

▶ 叫回「計算變數」對話方塊。

- 在「數值運算式」文字方塊中輸入 `partition - rv.bernoulli(0.2)`。
- 按一下「如果」。

圖表 4-11
計算變數：觀察值選擇條件對話方塊

- 選取「包含滿足條件的觀察值」。
- 在文字方塊中輸入 `partition>0`。
- 按一下「繼續」。
- 在「計算變數」對話方塊中，按一下「確定」。

這會重設大於 0 的 partition 值，因此大約 20% 會使用值 0，而 80% 則維持值 1。整體而言，大約 100*(0.7*0.8)=56% 先前已取得貸款的客戶會出現在訓練樣本中，而 14% 出現在測試樣本中。原先指定給保留樣本的客戶會保留不動。

執行分析

- 叫回「多層認知」對話方塊，並按一下「儲存」索引標籤。
- 選擇「儲存各依變數的預測虛擬機率」。
- 按一下「確定」。

觀察值處理摘要

圖表 4-12
測試樣本相關模式的觀察值處理摘要

		個數	百分比
樣本	訓練	398	56.9%
	檢驗	101	14.4%
	保留	201	28.7%
	有效	700	100.0%
	排除	150	
	總數	850	

在 499 個原先指定給訓練樣本的觀察值中，有 101 個重新指定給測試樣本。

網路資訊

圖表 4-13
網路資訊

輸入階層	因子	1	Level of education
	共變量	1	Age in years
		2	Years with current employer
		3	Years at current address
		4	Household income in thousands
		5	Debt to income ratio (x100)
		6	Credit card debt in thousands
		7	Other debt in thousands
	因子	單位數	12
		共變量的重新計算方法	標準化
隱藏階層	因子	隱藏階層的數目	1
		隱藏階層 1 的單位數	4
		啓動函數	超正反切
輸出階層	依變數	1	Previously defaulted
	因子	單位數	2
		啓動函數	Softmax
		錯誤函數	交叉熵

a. 排除偏離單元

網路資訊表唯一的變更是，自動架構選擇已經在隱藏階層中選擇七個單位。

模式摘要

圖表 4-14
模式摘要

訓練	交叉熵錯誤	159.870
	百分比不正確預測	20.1%
	使用的中止規則	超過最大週期數目 (100)
	訓練時間	0:00:01.013
檢驗	交叉熵錯誤	40.068
	百分比不正確預測	17.8%
保留	百分比不正確預測	20.4%

依變數： Previously defaulted

a. 錯誤計算基于測試樣本。

模式摘要顯示多個正數符號：

- 在訓練樣本、測試樣本和保留樣本之間，不正確預測的百分比大略相等。
- 經過演算法的一個步驟後，錯誤仍未減少，因此估計演算法已經停止。

這更表示原始模式事實上受到過度訓練，而且已新增測試樣本解決這項問題。當然，樣本大小也比較小，而我們可能不該過度解析有些百分比的變動範圍。

分類

圖表 4-15
分類

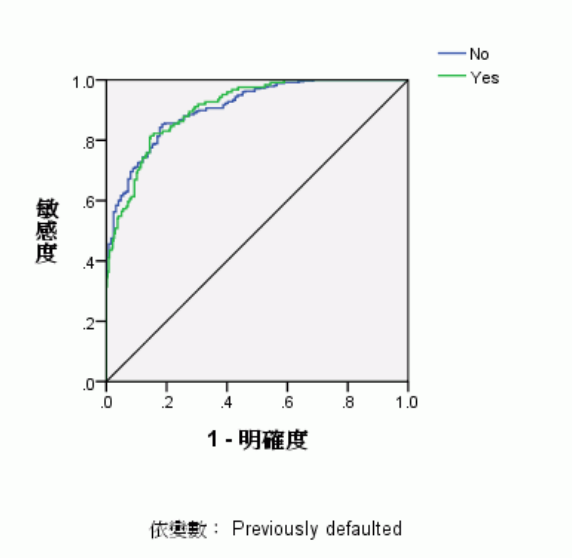
		預測次數		
樣本	觀察次數	No	Yes	百分比修正
訓練	No	263	34	88.6%
	Yes	46	55	54.5%
	整體百分比	77.6%	22.4%	79.9%
訓練	No	73	5	93.6%
	Yes	13	10	43.5%
	整體百分比	85.1%	14.9%	82.2%
保留	No	124	18	87.3%
	Yes	23	36	61.0%
	整體百分比	73.1%	26.9%	79.6%

依變數： Previously defaulted

分類表顯示，使用 0.5 作為分類的虛擬機率分割值時，網路預測非拖欠者的效果比預測拖欠者更好。然而，單一分割值只能發揮一部份的網路預測能力，因此未必能夠用於比較同性質的網路。請改為檢視 ROC 曲線。

ROC 曲線

圖表 4-16
ROC 曲線



ROC 曲線能夠以目視方式，在一張圖中呈現所有可能分割的**敏感度**和**明確性**，這比連續多個表格的呈現方式更為清楚完整。其中顯示的圖表有兩條曲線，一條表示類別「否」，另一條表示類別「是」。由於只有兩個類別，因此兩條曲線與從圖表左上角到右下角傾斜 45 度的線條（未顯示）保持對稱。

請注意，這張圖表是根據合併的訓練樣本和測試樣本而得。若要產生保留樣本的 ROC 圖表，請將分割變數上的檔案分割，然後在儲存的預測虛擬機率上執行 ROC 曲線程序。

圖表 4-17
曲線下的區域

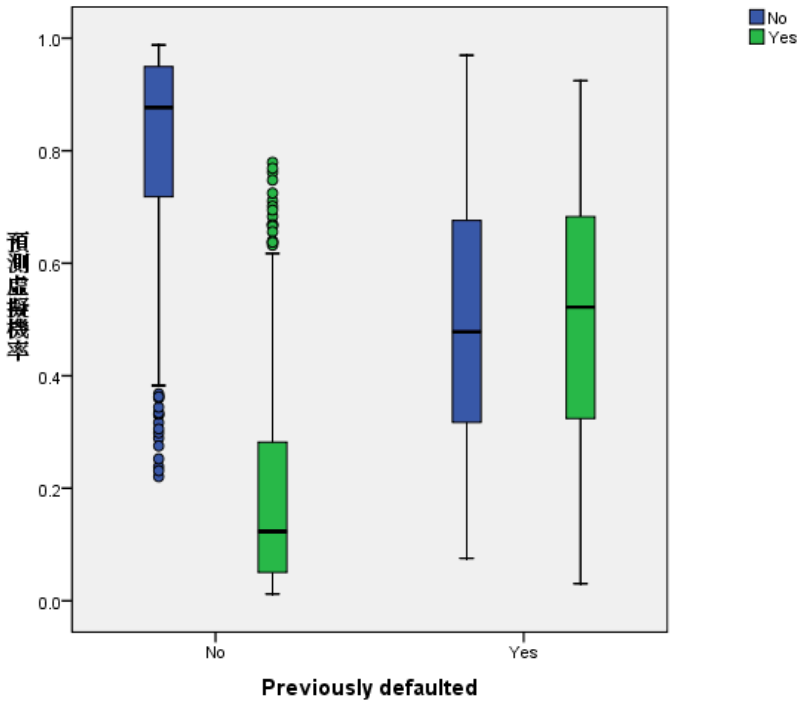
		區域
Previously defaulted	No	.853
	Yes	.853

曲線下的區域是 ROC 曲線的數值摘要，對於各個類別而言，表中的值表示：出現在該類別中的預測虛擬機率的機率，對於在該類別中隨機選擇的觀察值而言，高於不在該類別中隨機選擇的觀察值。例如，對於隨機選擇的拖欠者與隨機選擇的非拖欠者，拖欠者的拖欠模式預測虛擬機率比非拖欠者高出 0.853 的機率。

由於曲線下的區域是實用的網路準確性單一統計摘要，因此您必須能夠選擇分類客戶所依據的特定條件。在進行這項程序時，觀察值對預測值圖表能夠提供視覺參考。

觀察值對預測值圖表

圖表 4-18
觀察值對預測值圖表



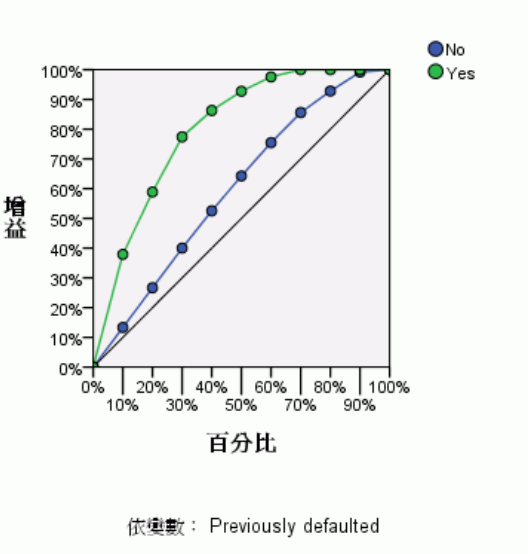
對於類別依變數，觀察值對預測值圖表會顯示結合訓練樣本和測試樣本的預測虛擬機率集群盒形圖。x 軸對應於觀察反應類別，而圖註對應於預測類別。

- 對於具有觀察類別「否」的觀察值，最左邊的盒形圖顯示屬於類別「否」的預測虛擬機率。y 軸 0.5 標記處以上的盒形圖部份表示分類表中顯示的正確預測。0.5 標記處以下的部份表示不正確預測。請記住，在分類表中網路使用 0.5 分割值準確預測含有「否」類別的觀察值，因此只有較低盒鬚其中一部份與某些偏離觀察值的分類錯誤。
- 對於具有觀察類別「否」的觀察值，右邊第二個盒形圖顯示屬於類別「是」的預測虛擬機率。由於目標變數中只有兩個類別，因此前兩個盒形圖與 0.5 處的水平線保持對稱。
- 對於具有觀察類別「是」的觀察值，第三個盒形圖顯示屬於類別「否」的預測虛擬機率。這和最後一個盒形圖與 0.5 處的水平線保持對稱。
- 對於具有觀察類別「是」的觀察值，最後一個盒形圖顯示屬於類別「是」的預測虛擬機率。y 軸 0.5 標記處以上的盒形圖部份表示分類表中顯示的正確預測。0.5 標記處以下的部份表示不正確預測。請記住，在分類表中網路使用 0.5 分割值預測的結果略高於含有「是」類別的觀察值一半，因此盒形良好部份的分類錯誤。

在盒形圖中，由於將觀察值分類為「是」的分割值從 0.5 降低至大約 0.3—這大約是第二個盒形頂端與第四個盒形底端之間的值—因此您可以更正確找出可能的拖欠者，而不流失許多潛在的良好客戶。也就是說，沿著第二個盒形從 0.5 移至 0.3，將相對較少的非拖欠客戶沿著盒鬚重新錯誤分類為預測拖欠者，然而，沿著第四個盒形移動時，將盒形中許多拖欠客戶重新正確分類為預測拖欠者。

累積增益圖表和提升圖表

圖表 4-19
累積增益圖表

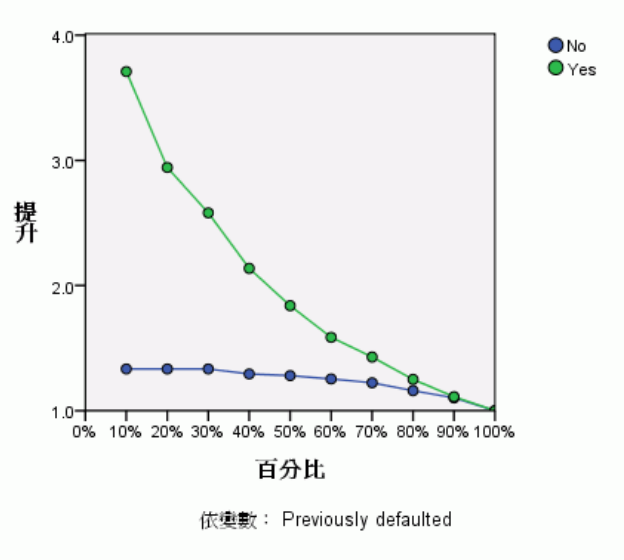


累積增益圖表以觀察值總數的百分比為目標，顯示指定類別「增益」中觀察值的總數百分比。例如，在「是」類別的曲線上，第一個點位在 (10%, 30%)，這表示，如果您使用網路評定資料集，並且依照預測虛擬機率「是」將所有觀察值排序，則前 10% 會包含大約 30% 實際具有類別「是」的觀察值 (拖欠者)。同樣地，前 20% 會包含大約 50% 的拖欠者，前 30% 的觀察值會包含 70% 的拖欠者，依此類推。如果您選取 100% 的評分資料集，便會取得資料集中所有的拖欠者。

對角線是「基準線」曲線；如果您從評分資料集隨機選擇 10% 的觀察值，則會「增益」大約 10% 實際具有類別「是」的所有觀察值。曲線在基準線上方的距離愈遠，則增益愈大。您可以使用累積增益圖表選擇對應於所需增益的百分比，然後將此百分比對應至適當的分割值，以協助選擇類別分割值。

「所需」增益的內容為何，需視類型一和類型二錯誤的成本而定。也就是將拖欠者歸類為非拖欠者 (類型一) 的成本是多少？將非拖欠者歸類為拖欠者 (類型二) 的成本是多少？如果呆帳是主要考量，則您會想要降低類型一錯誤；在累積增益圖表上，這會對應於拒絕放貸給前 40% 虛擬預測機率為「是」的申請人，這幾乎佔可能拖欠者的 90%，同時剔除將近一半的申請人。 如果以客戶數量的成長為優先考量，則您要降低「類型二」錯誤。在圖表上，這可能對應於拒絕前 10%，等於佔拖欠者的 30%，同時保留大部分的申請人。通常以上兩項都是主要考量，因此您必須選擇並決定一個分類的規則，此規則能讓敏感度與明確性有最佳的組合。

圖表 4-20
提升圖表



提升圖表衍生自累積增益圖表；y 軸上的值對應於各曲線與基準線的累積增益比例。因此，類別「是」提升 10% 便是 30%/10% = 3.0。這是另一種檢視累積增益圖表資料的方式。

注意：累積增益圖表和提升圖表是根據合併的訓練樣本和測試樣本而得。

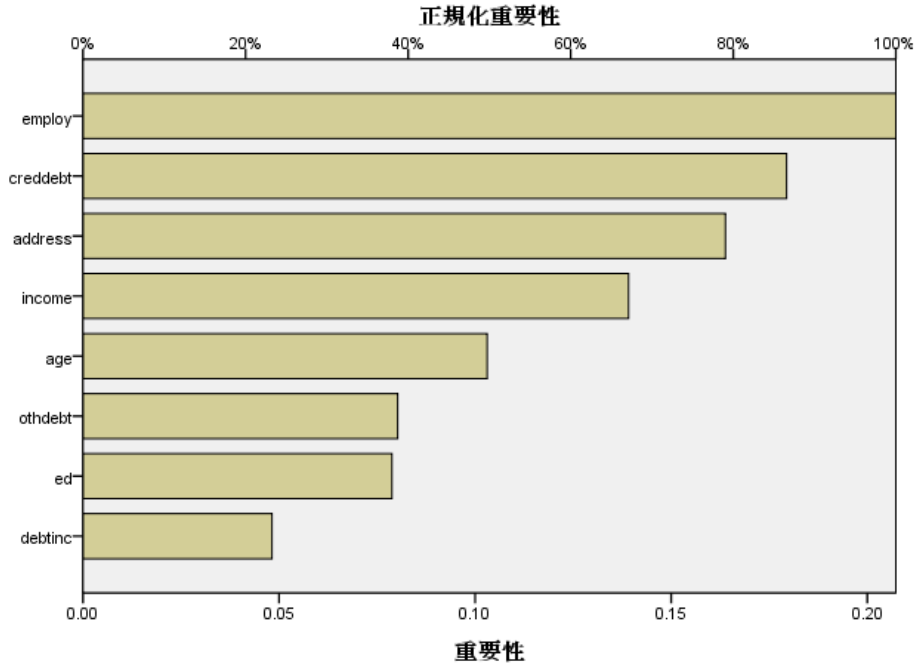
自變數的重要性

圖表 4-21
自變數重要性

	重要性	正規化重要性
Level of education	.032	11.9%
Age in years	.075	27.9%
Years with current employer	.268	100.0%
Years at current address	.166	61.8%
Household income in thousands	.033	12.2%
Debt to income ratio (x100)	.125	46.5%
Credit card debt in thousands	.213	79.3%
Other debt in thousands	.090	33.6%

自變數的重要性，在於能夠測量網路的模式預測值針對不同的自變數值發生多少變化。常態化重要性就是重要性值除以最高重要性值，並且以百分比表示。

圖表 4-22
自變數重要性圖表



重要性圖表就是重要性表格中各值的長條圖，並以遞減重要性值進行排序。與客戶穩定性 (employ、address) 和負債 (creddebt、debtinc) 相關的變數似乎對網路將客戶分類的方式影響最大；無法預測的是這些變數與拖欠預測機率之間關係的「方向」。您會猜測較多負債是否表示較可能拖欠，為了進行確認，您需要使用較容易解讀其中參數的模式。

摘要

您已經使用「多層認知」程序，建構預測指定客戶貸款拖欠之機率網路。模式結果能夠與使用「Logistic 迴歸」或「判別分析」取得的結果相互比較，所以您相當確信資料不會包含這些模式無法擷取的關係；因此，您可以使用這些結果進一步分析依變數與自變數之間關係的本質。

使用多層認知評估醫療保健成本與住院日數

醫院系統經常追蹤心肌梗塞 (MI 或「心臟病發作」) 治療相關的成本與病患住院日數。取得這些測量的精確估計值後，管理員便能夠正確管理治療病患的可用病床數。

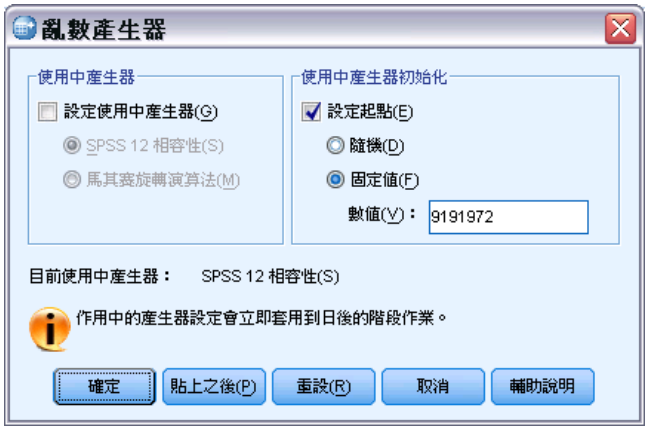
資料檔 patient_los.sav 含有接受 MI 治療的病患樣本治療記錄。 使用「多層認知」程序建立預測成本與住院日數的網路。

準備進行分析所用的資料

設定亂數種子可讓您複製出完全相同的分析。

- 若要設定亂數種子，從功能表選擇：
  轉換 > 亂數產生器...

圖表 4-23
「亂數產生器」對話方塊

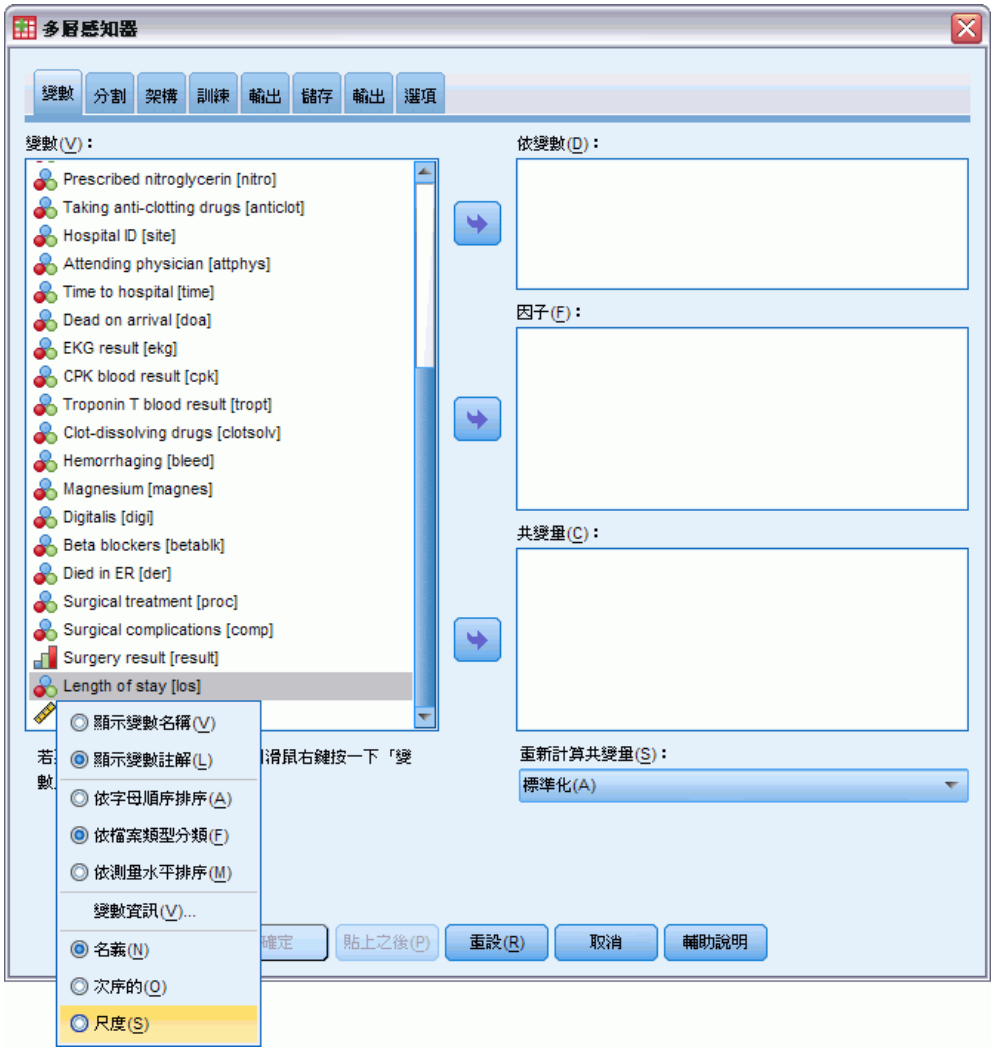


- 選取「設定起始點」。
- 選取「固定值，再輸入「9191972」作為值。
- 按一下「確定」。

執行分析

- 若要執行「多層認知」分析，從功能表選擇：
  分析 > 神經網路 > 多層認知...

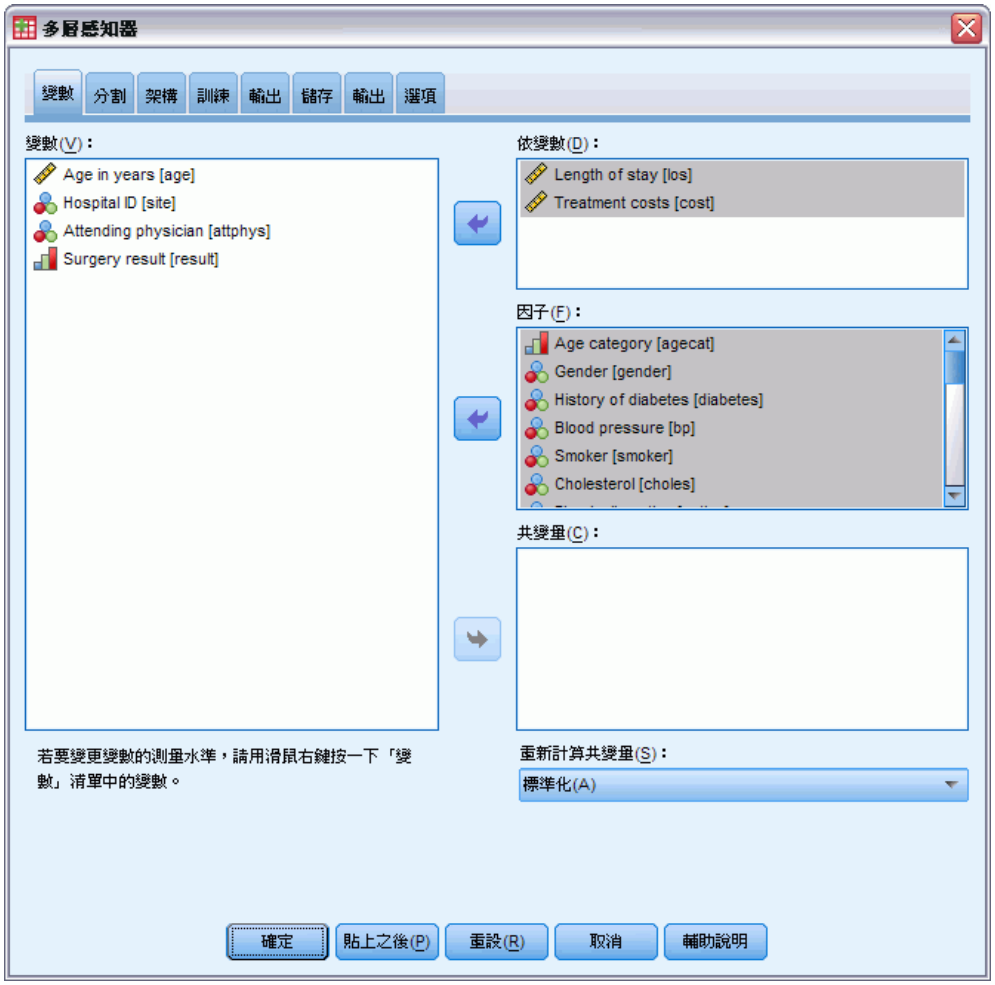
圖表 4-24
多層認知：住院日數的變數索引標籤和快顯功能表



「住院日數 [los]」具有次序測量水準，但是您想要讓網路將它視為尺度。

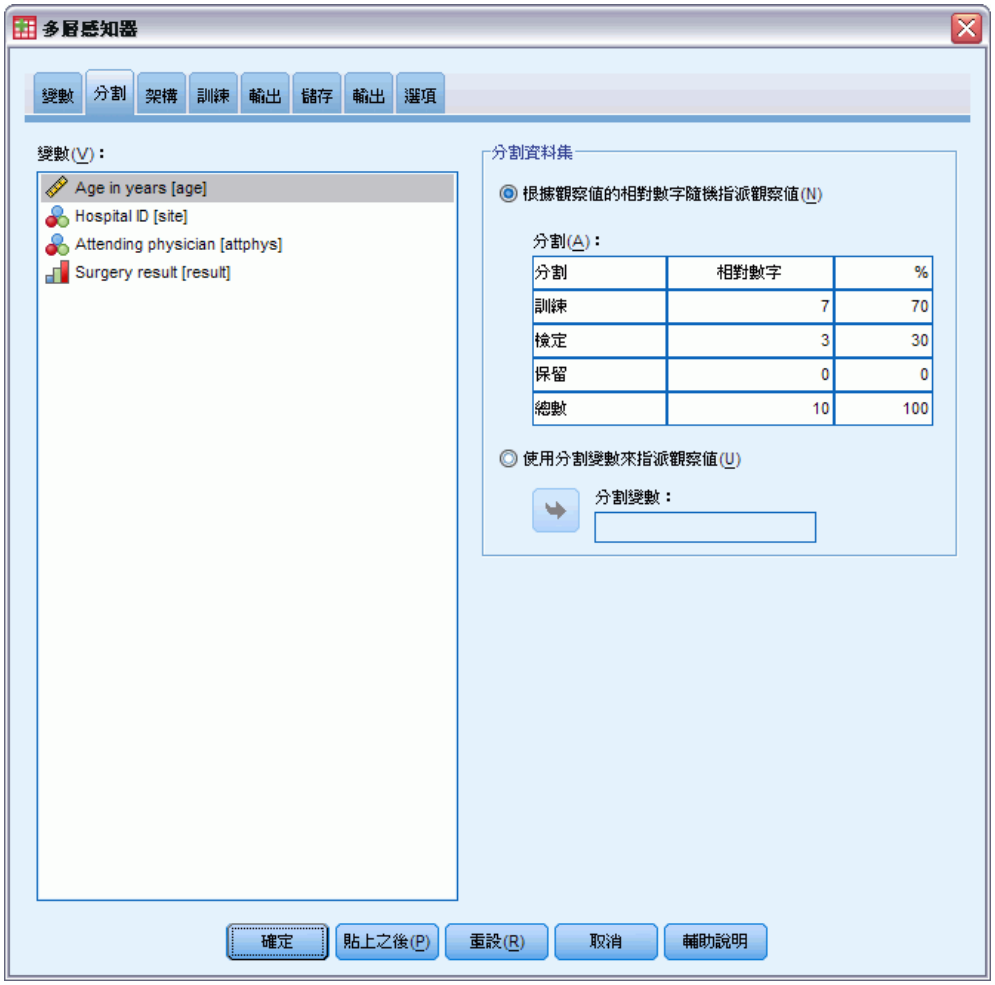
- 在「住院日數 [los]」上按一下滑鼠右鍵，然後選擇快顯功能表中的「**尺度**」。

圖表 4-25
多層認知：已選擇依變數與因子的變數索引標籤



- 選擇「住院日數 [los]」和「治療成本 [cost]」作為依變數。
- 選擇「年齡類別 [agecat]」到「服用抗凝血藥物 [anticlot]」，並選擇「到院時間 [time]」到「外科併發症 [comp]」作為因子。為確保確實重複以下的模式結果，務必維持因子清單中的變數順序。為了達到此目的，您會發現，選擇各組預測變數，並使用按鈕將這些預測變數移至因子清單，是相當便利的做法，而不需要使用拖曳的方式。另外，變更變數的順序可協助您評估解決辦法的穩定性。
- 按一下「**分割**」索引標籤。

圖表 4-26
多層認知：分割索引標籤



- 輸入 2 作為要指派給測試樣本的相對觀察值個數。
- 輸入 1 作為要指派給保留樣本的相對觀察值個數。
- 按一下「架構」索引標籤。

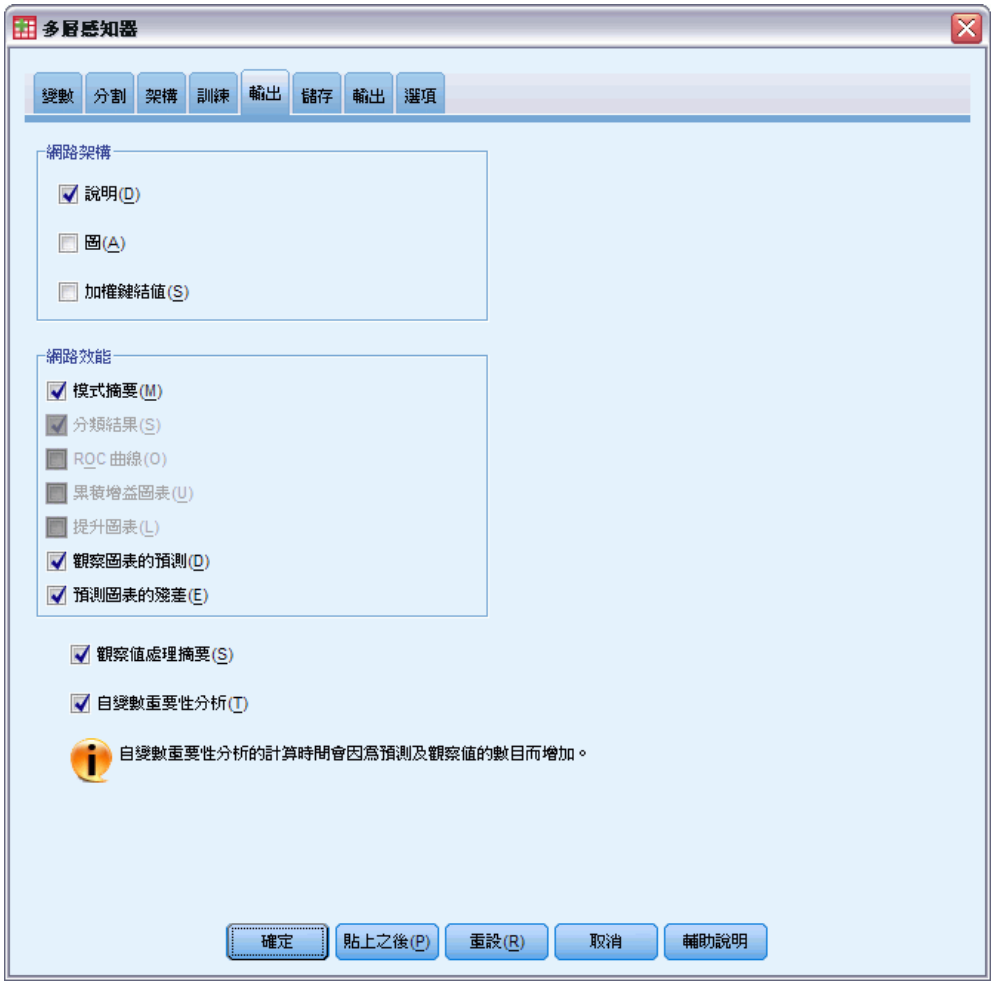
圖表 4-27
多層認知：「架構」索引標籤

- 選擇「自訂架構」。
- 選擇「2」作為隱藏階層數目。
- 選擇「雙曲線正切」作為輸出階層啟動函數。請注意，這會自動將依變數的調整方法設定為「調整後常態化」。
- 按一下「訓練」索引標籤。

圖表 4-28
多層認知：「訓練」索引標籤

- 選擇「線上」作為訓練類型。線上訓練應該能夠在包含相關預測變數的「較大型」資料集上正常執行。請注意，這會使用對應的預設選項，自動將「梯度坡降」設定為最佳化演算法。
- 按一下「輸出」索引標籤。

圖表 4-29
多層認知：「輸出」索引標籤



- 取消選擇「圖」；其中會有許多輸入，而且產生的圖不便使用。
- 選擇「網路效能」組別中的「**觀察值對預測值圖表**」和「**預測殘差圖表**」。由於未將任何依變數視為類別變數（名義變數或次序變數），因此不提供分類結果、ROC 曲線、累積增益圖表和提升圖表。
- 選擇「**自變數重要性分析**」。
- 按一下「**選項**」索引標籤。

圖表 4-30
「選項」索引標籤

- 選擇加入使用者遺漏值的變數。未進行外科程序的病患會有「外科併發症」變數的使用者遺漏值。這可確保這些病患會加入分析中。
- 按一下「確定」。

警告

圖表 4-31
警告

下列自變數在訓練樣本中爲常數，並且已從分析中排除: doa, der.

警告表格顯示變數 doa 和 der 是訓練樣本中的常數。到院時已死亡或在急診室中死亡的病患會有「住院日數」的使用者遺漏值。由於我們將「住院日數」視為此分析的尺度變數，而且已排除尺度變數中含有使用者遺漏值的觀察值，因此只會加入送出急診室仍存活的病患。

觀察值處理摘要

圖表 4-32
觀察值處理摘要

		個數	百分比
樣本	訓練	5647	70.6%
	測試	1570	19.6%
	保留	781	9.8%
	有效	7998	100.0%
	排除	2002	
	總數	10000	

觀察值處理摘要顯示 5647 個觀察值指定給訓練樣本，1570 個指定給測試樣本，781 個指定給保留樣本。2002 個排除在分析之外的觀察值是轉院途中死亡或在急診室中死亡的病患。

網路資訊

圖表 4-33
網路資訊

輸入階層	因子	1	Age category
		2	Gender
		3	History of diabetes
		4	Blood pressure
		5	Smoker
		6	Cholesterol
		7	Physically active
		8	Obesity
		9	History of angina
		10	History of myocardial infarction
		11	Prescribed nitroglycerin
		12	Taking anti-clotting drugs
		13	Time to hospital
		14	EKG result
		15	CPK blood result
		16	Troponin T blood result
		17	Clot-dissolving drugs
		18	Hemorrhaging
		19	Magnesium
		20	Digitalis
		21	Beta blockers
		22	Surgical treatment
		23	Surgical complications
	因子	單位數	63
隱藏階層	因子	隱藏階層的數目	2
		隱藏階層 1 的單位數	12
		Number of Units in Hidden Layer 2	9
		啓動函數	超正反切
輸出階層	依變數	1	Length of stay
		2	Treatment costs
	因子	單位數	2
		Rescaling Method for Scale Dependents	Adjusted Normalized
		啓動函數	超正反切
		錯誤函數	平方和

a. 排除偏離單元

網路資訊表顯示神經網路的資訊，可用於確認規格是否正確。其中必須特別注意：

- 輸入階層中的單位數是因子水準的總數（沒有任何共變量）。
- 已要求兩個隱藏階層，而且程序已選擇第一個隱藏階層的 12 的單位和第二個隱藏階層的 9 個單位。

- 已針對各個尺度依變數建立個別的輸出單位。這都會以調整後常態化方法進行調整，其中需要使用輸出階層的雙曲線正切啟動函數。
- 由於依變數是尺度，因此會回報平方和錯誤。

模式摘要

圖表 4-34
模式摘要

訓練	誤差平方和		91.812
	平均整體相對錯誤		.083
	尺度相依的相對錯誤	Length of stay	.131
		Treatment costs	.033
	使用的中止規則		不包含錯誤縮減的 1 連續步驟[a]
	訓練時間		0:00:26.422
測試	誤差平方和		26.798
	平均整體相對錯誤		.088
	尺度相依的相對錯誤	Length of stay	.141
		Treatment costs	.033
保留	平均整體相對錯誤		.099
	尺度相依的相對錯誤	Length of stay	.154
		Treatment costs	.041

a. 錯誤計算是根據測試樣本而定。

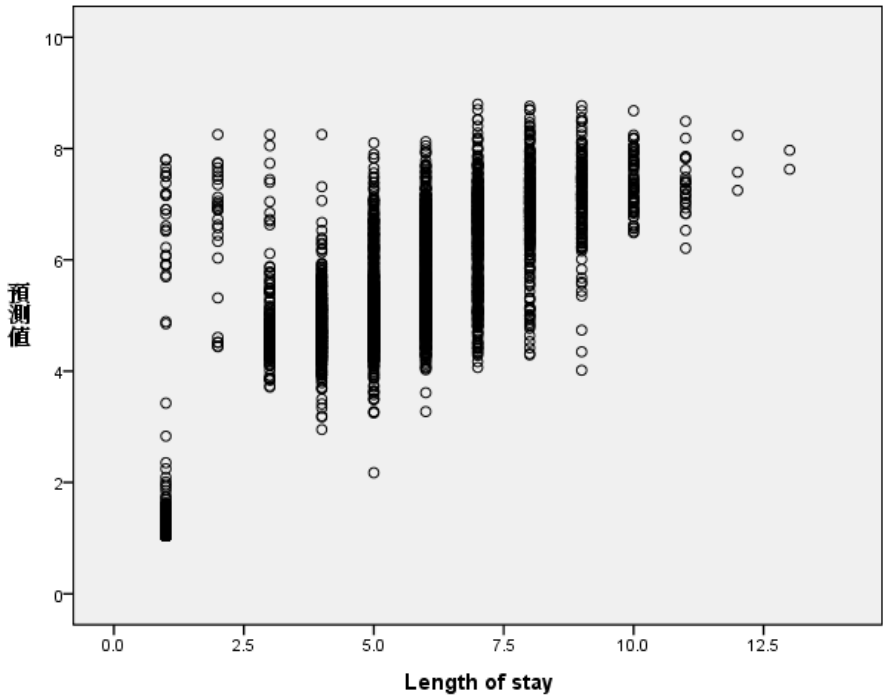
模式摘要顯示訓練和套用最終網路於保留樣本的結果資訊。

- 由於輸出階層有尺度依變數，因此會顯示平方和錯誤。這是網路嘗試在訓練時最小化的錯誤函數。請注意，針對依變數的調整值，會計算平方和及下列所有錯誤值。
- 各個尺度依變數的相對錯誤是依變數平方和錯誤與「虛無」模式平方和錯誤的比例，其中依變數的平均值是作為各個觀察值的預測值。預測「住院日數」的錯誤似乎比「治療成本」的錯誤更多。
- 平均整體錯誤是所有依變數平方和錯誤與「虛無」模式平方和錯誤的比例，其中依變數的平均值是作為各個觀察值的預測值。在此範例中，平均整體錯誤恰巧接近相對錯誤的平均，但是並非每次都會如此。

  在訓練樣本、測試樣本和保留樣本中，平均整體相對錯誤和相對錯誤都相當固定，您可以確信模式未受到過度訓練，而且由網路評分的未來觀察值中出現的錯誤也會接近表格中回報的錯誤。
- 經過演算法的一個步驟後，錯誤仍未減少，因此估計演算法已經停止。

觀察值對預測值圖表

圖表 4-35
住院日數的觀察值對預測值圖表

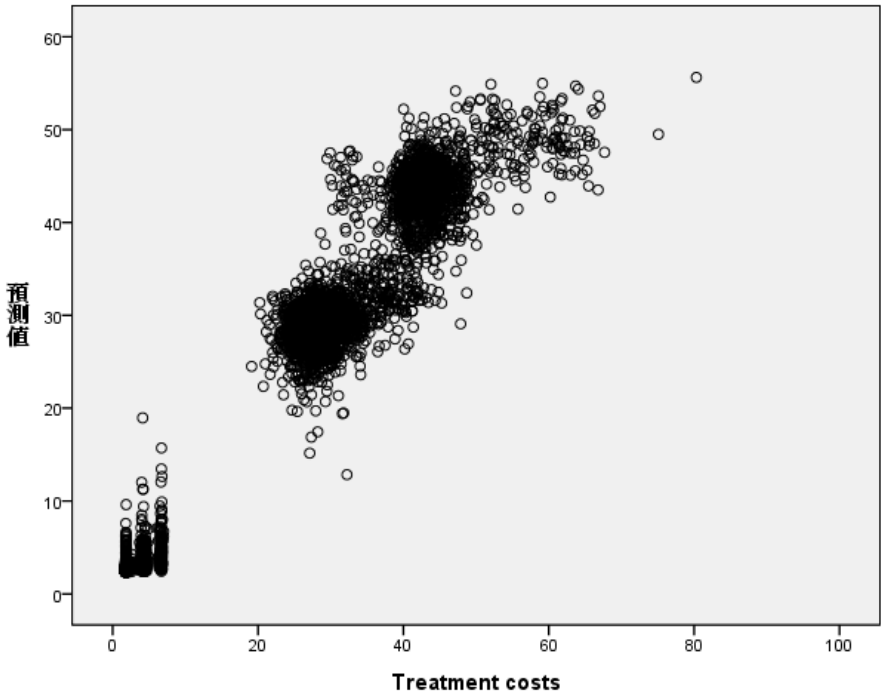


對於尺度依變數，觀察值對預測值圖表會針對結合訓練樣本和測試樣本，顯示 y 軸上預測值與 x 軸上觀察值的散佈圖。最理想的狀況是，值應該大略延著起點為原點的 45 度線散佈。此散佈圖中的各點會形成住院日數各觀察日數的垂直線。

在散佈圖中，網路似乎有效預測住院日數。散佈圖的一般傾向是偏離理想的 45 度線，觀察住院日數在 5 天以下的預測傾向高估住院日數，而觀察住院日數超過 6 天以上的預測則傾向低估住院日數。

散佈圖中左下角部份的一群病患有可能是未經過診療的病患。散佈圖中左上角出現另一群病患，他們的觀察住院日數為 1 至 3 天，而且預測值偏高。這些觀察值很可能是住院診療後死亡的病患。

圖表 4-36
治療成本的觀察值對預測值圖表



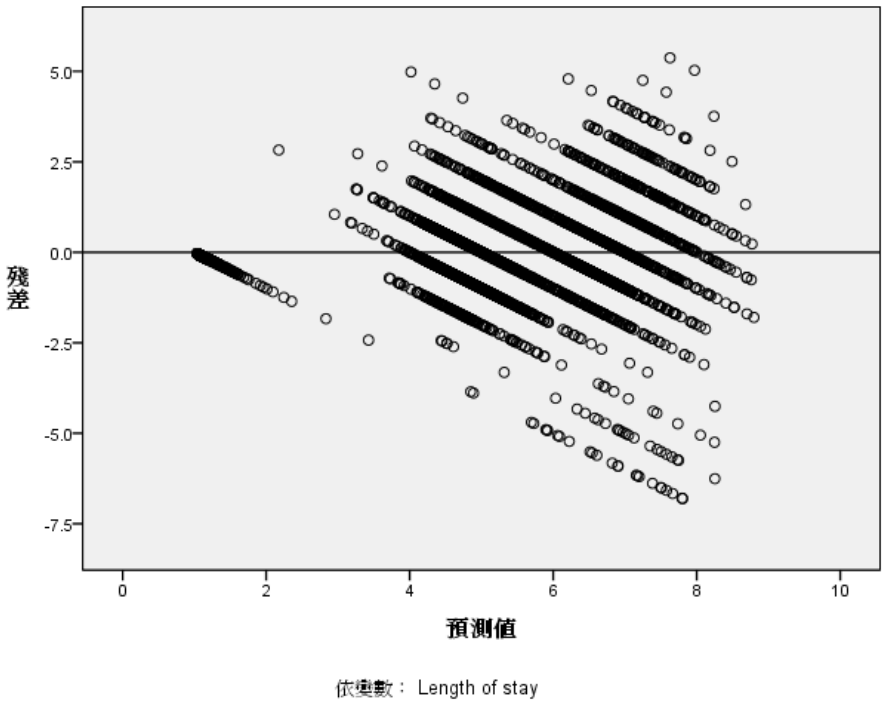
網路似乎也有效預測治療成本。其中主要分為三類病患：

- 右下方主要是未經過診療的病患。這些病患的成本較低，而且由急診室管理的「血塊溶解藥物 [clotsolv]」類型加以區分。
- 另一群病患的治療成本大約是 30,000 美元。這些病患曾經接受冠狀動脈血管擴張術 (PTCA) 的治療。
- 最後一群病患的治療成本超過 40,000 美元。這些病患曾經接受冠狀動脈繞道手術 (CABG) 的治療。這項手術的成本高於 PTCA，而且病患需要較長的住院休養時間，因此更導致成本增加。

還有一些觀察值的成本超過 50,000 美元，網路預測這些觀察值的效果不佳。這些病患在接受診療時出現併發症，因此增加診療成本與住院日數。

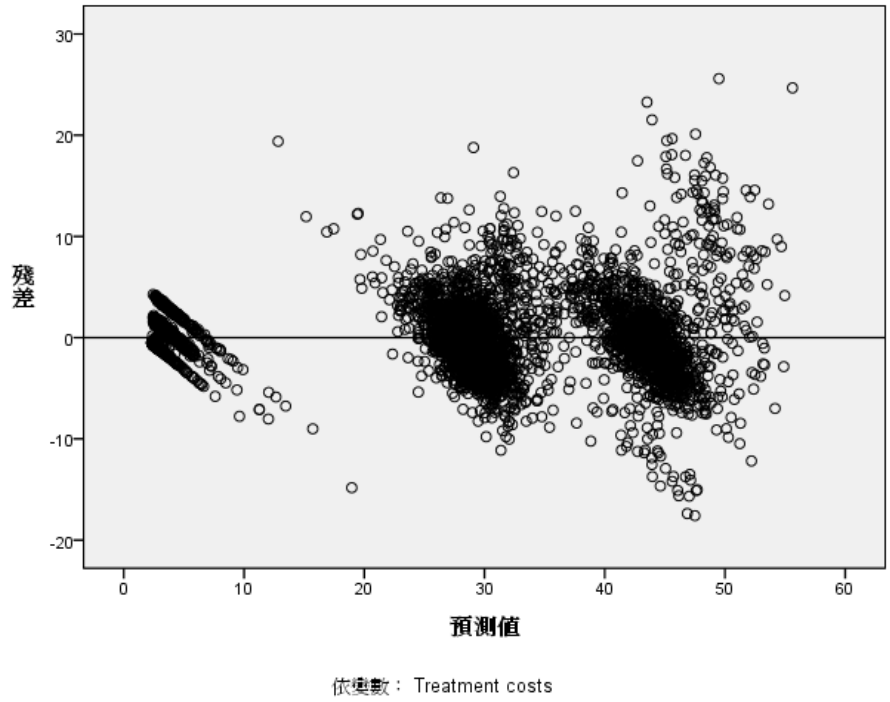
預測殘差圖表

圖表 4-37
住院日數的預測殘差圖表



預測殘差圖表顯示 y 軸上殘差（觀察值減去預測值）與 x 軸上預測值的散佈圖。散佈圖中的各條對角線對應於觀察值對預測值圖表中的垂直線，而且您可以在觀察的住院日數增加時，更明顯看出從住院日數過高預測到過低預測的進度變化。

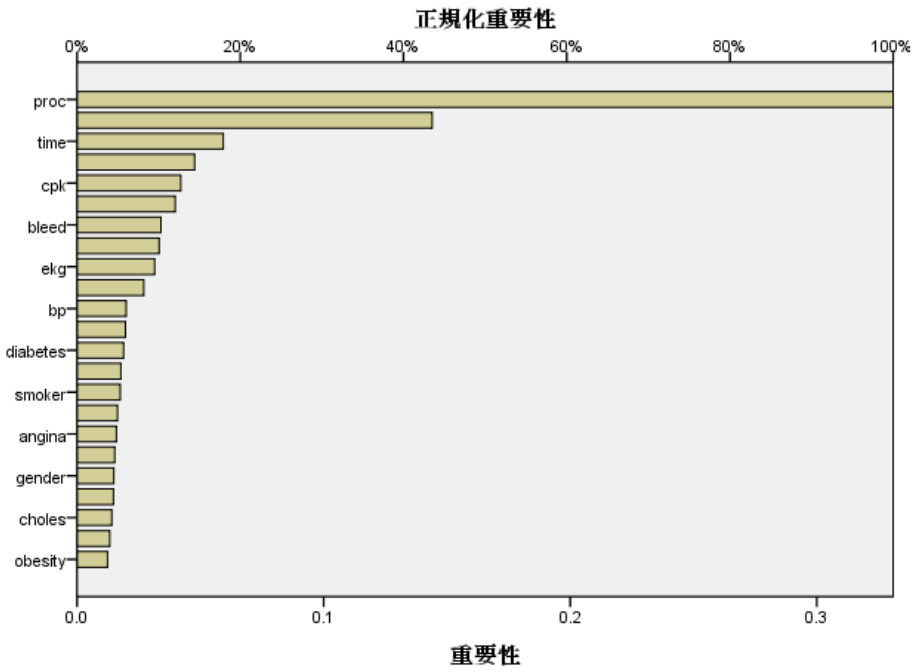
圖表 4-38
治療成本的預測殘差圖表



同樣地，對於從治療成本的預測殘差圖表中觀察而得的三類病患，在觀察的成本增加時，預測殘差圖表也會顯示從成本過高預測到過低預測的進度變化。進行 CABG 時出現併發症的病患仍然清楚可見，但是進行 PTCA 時出現併發症的病患更為明顯，這一小群病患出現在 x 軸 30,000 美元標記處附近一大群 PTCA 病患的偏右上方位置。

自變數的重要性

圖表 4-39
自變數重要性圖表



重要性圖表顯示，結果會受到執行的外科程序、後續是否發生併發症及其他相關預測變數等的影響。對於觀察住院日數最長的病患而言，雖然可以看出併發症對於住院日數的影響，但是，在治療成本散佈圖中，可以清楚看出外科程序的重要性，不過這在住院日數散佈圖中則較不明顯。

摘要

網路似乎有效預測「一般」病患的值，但是未擷取診療後死亡的病患。其中一種可解決這項問題的方式，是建立多個網路。其中一個網路可預測病患結果，例如病患是否存活，而另一個網路可預測病患是否存活的治療成本和 住院日數等狀況。您可以結合網路的結果，以取得較完整的預測。您可以使用類似的方法，解決接受手術時出現併發症的病患相關成本與住院日數過低預測的問題。

閱讀資料推薦

請參閱下列文字以瞭解神經網路和多層感知的詳細資訊：

Bishop, C. M. 1995. Neural Networks for Pattern Recognition (樣式辨識神經網路), 第三版 ed. Oxford (牛津): Oxford University Press (牛津大學出版部).

Fine, T. L. 1999. Feedforward Neural Network Methodology (前饋神經網路方法論), 第三版 ed. 紐約: Springer-Verlag.

Haykin, S. 1998. Neural Networks: (神經網路)A Comprehensive Foundation (全方位基礎), 第二版 ed. 紐約: Macmillan College Publishing (Macmillan 學院出版部).

Ripley, B. D. 1996. Pattern Recognition and Neural Networks (樣式辨識與神經網路). Cambridge (劍橋): Cambridge University Press (劍橋大學出版部).

章節
5

半徑式函數

半徑式函數（RBF）程序會產生根據預測變數值的一或多個依變數（目標變數）預測模式。

使用半徑式函數分類電信客戶

某電信公司根據服務使用方式來切割客戶數量，並將客戶分成四個組別。如果人口資料可用來預測組別成員關係，您可以替各個未來客戶自訂報價單。

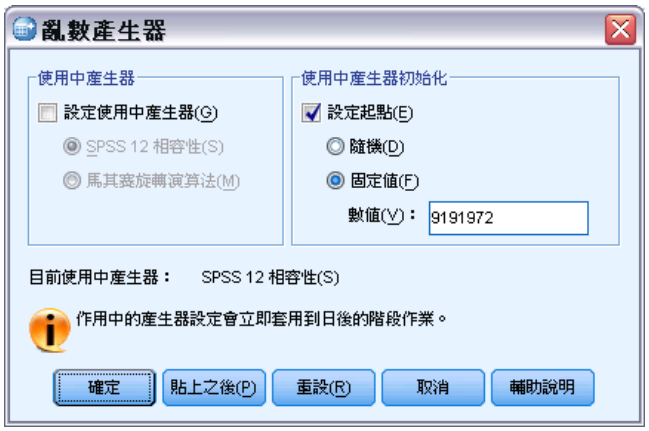
目前客戶的資訊存放在 telco.sav 中。 使用半徑式函數程序分類客戶

準備進行分析所用的資料

設定亂數種子可讓您複製出完全相同的分析。

- 若要設定亂數種子，從功能表選擇：
  轉換 > 亂數產生器...

圖表 5-1
「亂數產生器」對話方塊

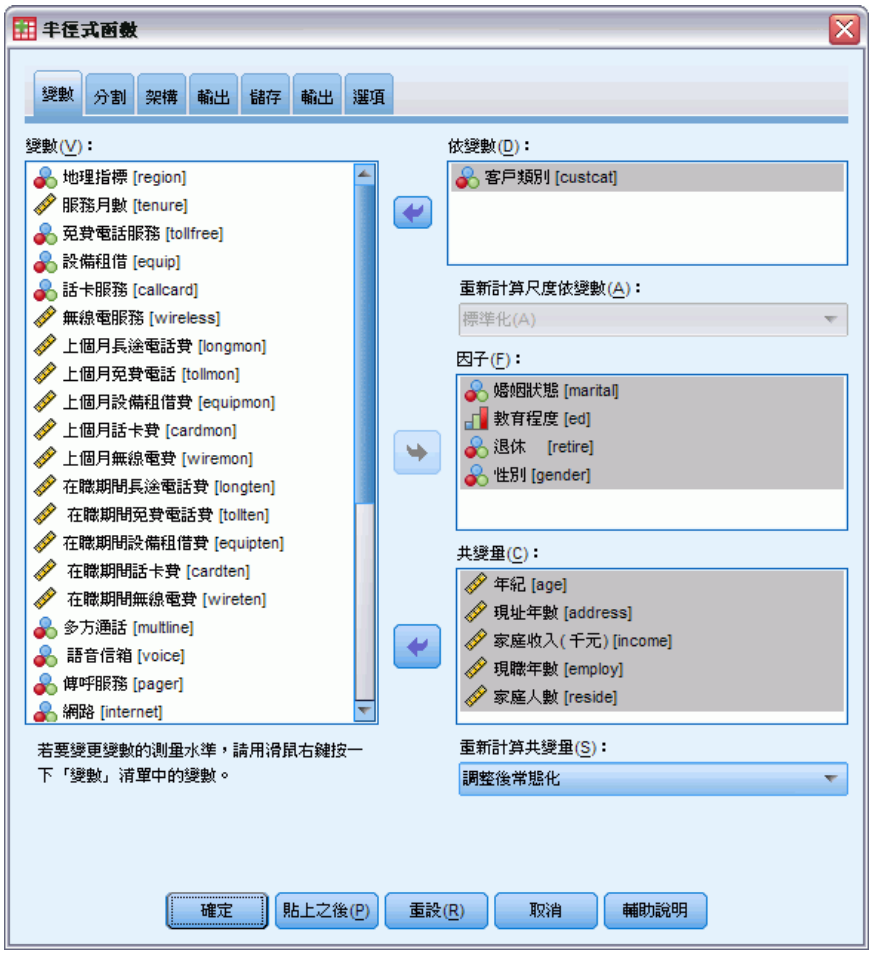


- 選取「設定起始點」。
- 選取「固定值，再輸入「9191972」作為值。
- 按一下「確定」。

執行分析

▶ 若要執行半徑式函數分析，請從功能表中選擇：
分析 > 神經網路 > 半徑式函數...

圖表 5-2
半徑式函數：「變數」索引標籤



▶ 選擇「客戶類別 [custcat]」作為依變數。

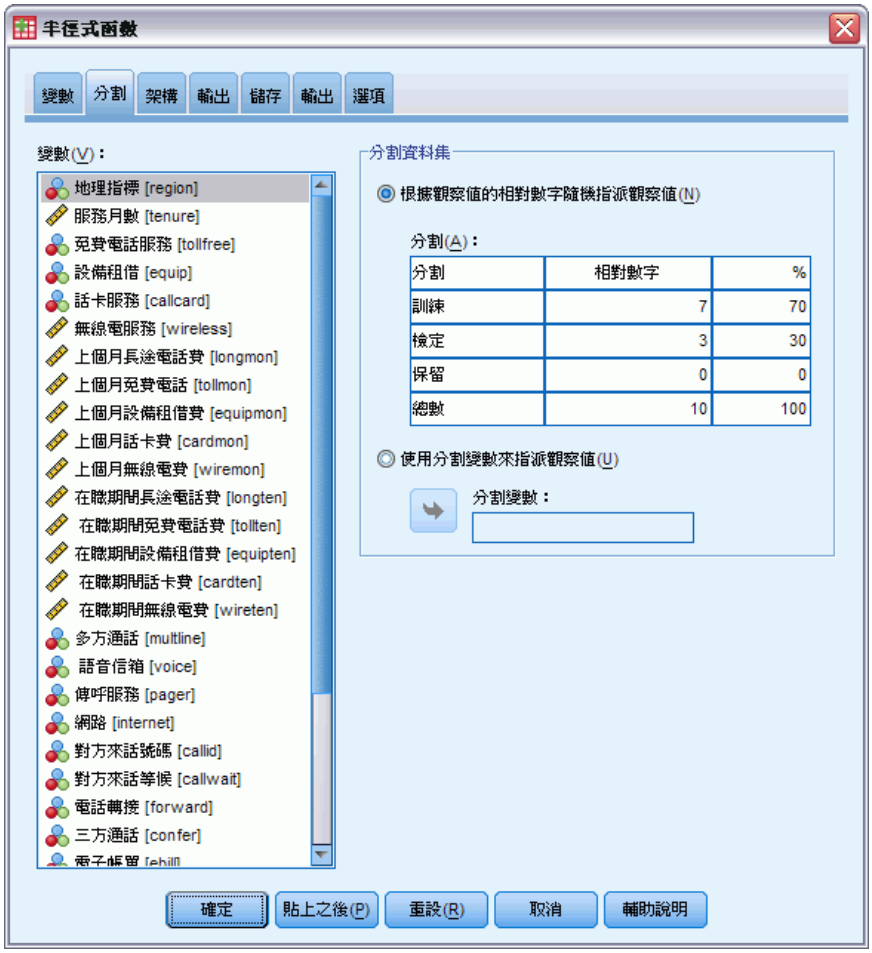
▶ 選擇「婚姻狀況 [marital]」、「教育程度 [ed]」、「退休 [retire]」和「性別 [gender]」作為因子。

▶ 選擇「年齡（以年為單位）[age]」到「家庭人數 [reside]」作為共變量。

▶ 選擇「**調整後常態化**」作為調整共變量的方法。

▶ 按一下「**分割**」索引標籤。

圖表 5-3
半徑式函數：「分割」索引標籤



指定觀察值的相對數字之後，便可以輕鬆建立不易指定百分比的分數分割。假設您要將 2/3 的資料集指定給訓練樣本，並且將剩餘觀察值的 2/3 指定給測試樣本。

- 輸入 **6** 做為訓練樣本的相對數字。
- 輸入 **2** 做為測試樣本的相對數字。
- 輸入 **1** 做為保留樣本的相對數字。

總共指定 9 個相對觀察值。6/9 = 2/3 或約 66.67% 是指定給訓練樣本；2/9 或約 22.22% 是指定給測試樣本；1/9 或約 11.11% 是指定給保留樣本。

- 按一下「**輸出**」索引標籤。

圖表 5-4
半徑式函數：「輸出」索引標籤

- 取消選擇「網路結構」組別中的「圖」。
- 選擇「網路效能」組別中的「ROC 曲線」、「累積增益圖表」、「提升圖表」和「觀察圖表的預測」。
- 按一下「儲存」索引標籤。

圖表 5-5
半徑式函數：「儲存」索引標籤

- 選擇「儲存各依變數的預測值或類別」和「儲存各依變數的預測虛擬機率」。
- 按一下「確定」。

觀察值處理摘要

圖表 5-6
觀察值處理摘要

		個數	百分比
樣本	訓練	665	66.5%
	測試	224	22.4%
	保留	111	11.1%
	有效	1000	100.0%
	排除	0	
	總數	1000	

觀察值處理摘要顯示 665 個觀察值指定給訓練樣本，224 個指定給測試樣本，111 個指定給保留樣本。沒有任何觀察值從分析中排除。

網路資訊

圖表 5-7
網路資訊

輸入階層	因子	1	婚姻狀態
		2	教育程度
		3	退休
		4	性別
	共變量	1	年紀
		2	現址年數
		3	家庭收入(千元)
		4	現職年數
		5	家庭人數
	因子	單位數	16
		共變量的重新計算方法	Adjusted Normalized
隱藏階層	因子	單位數	9[a]
		啓動函數	Softmax
輸出階層	依變數	1	客戶類別
	因子	單位數	4
		啓動函數	識別
		錯誤函數	平方和

a. 由測試資料準則決定：隱藏單位的「最佳」數目是在測試資料中產生最小錯誤的數目。

網路資訊表顯示神經網路的相關資訊，可用於確認規格是否正確。其中必須特別注意：

- 輸入階層的單位個數等於共變量個數加上因子水準總數；對於「婚姻狀況」、「教育程度」、「退休」和「性別」的各個類別，會建立個別的單位，而且不會比照許多模式化程序中，將任何類別視為「冗餘」單位。
- 同樣地，對於「客戶類別」的各個類別，會建立個別的輸出單位，輸出階層中總共會有 4 個單位。
- 共變量會以調整後常態化方法進行調整。
- 自動架構選擇已在隱藏階層中選擇 9 個單位。
- 其他所有網路資訊都預設用於程序中。

模式摘要

圖表 5-8
模式摘要

訓練	誤差平方和	235.969
	百分比不正確預測	61.8%
	訓練時間	0:00:03.078
測試	誤差平方和	80.851[a]
	百分比不正確預測	62.9%
保留	百分比不正確預測	59.5%

依變數：客戶類別

a. 隱藏單位的數目是由測試資料準則決定：隱藏單位的「最佳」數目是在測試資料中產生最小錯誤的數目。

模式摘要顯示訓練、測試和套用最終網路於保留樣本的結果資訊。

- 顯示平方和錯誤，因為這經常用於 RBF 網路。這是網路嘗試在訓練和測試時最小化的錯誤函數。
- 不正確預測的百分比是從分類表中取得，後續會在該主題中進行討論。

Classification（分類）

圖表 5-9
Classification（分類）

		預測次數				
樣本	觀察次數	基本服務	E化服務	加值服務	全方位服務	百分比修正
訓練	基本服務	64	0	66	45	36.6%
	E化服務	22	1	57	61	.7%
	加值服務	47	0	104	34	56.2%
	全方位服務	29	1	49	85	51.8%
	整體百分比	24.4%	.3%	41.5%	33.8%	38.2%
測試	基本服務	18	0	26	15	30.5%
	E化服務	15	0	16	22	.0%
	加值服務	11	0	39	15	60.0%
	全方位服務	4	0	17	26	55.3%
	整體百分比	21.4%	.0%	43.8%	34.8%	37.1%
保留	基本服務	11	0	11	10	34.4%
	E化服務	4	0	9	10	.0%
	加值服務	10	0	19	2	61.3%
	全方位服務	5	0	5	15	60.0%
	整體百分比	27.0%	.0%	39.6%	33.3%	40.5%

依變數：客戶類別

分類表顯示使用網路的實際結果。對於每個觀察值，預測反應是最高預測虛擬機率的類別。

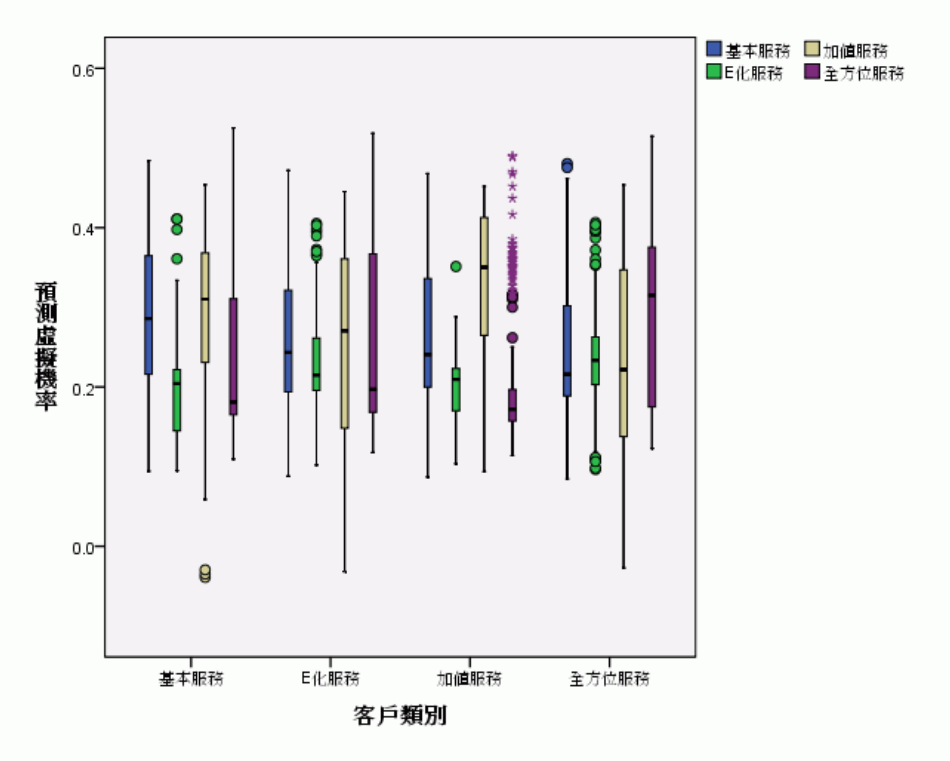
- 對角線上的儲存格為正確預測。
- 不在對角線上的儲存格為不正確預測。

如果有觀察的資料，「零階」模式（也就是沒有預測值的模式）會將所有的客戶分類到典型組別加值服務。因此，零階模式會是正確的 281/1000 = 28.1%。RBF 網路得到的結果是超過 10.1%，或是 38.2% 的客戶。事實上，您的模式的優點在於可以識別出 加值服務或總服務客戶。但是，它在分類電子服務客戶上的效果非常差。您可能需要找到另一個預測變數，才能區分這些客戶；或者，假設這些客戶最常被錯誤分類為加值服務和總服務的客戶，則公司只需要針對通常劃分為電子服務類別的潛在客戶嘗試向上促銷。

根據用於建立模式的觀察值來分類，會傾向於太「樂觀」，因為它們的分類比率是誇大的。保留樣本可協助驗證模式；其中 40.2% 的觀察值都經過模式的正確分類。雖然保留樣本較小，但是這意味著您的模式事實上五次中有兩次是正確的。

觀察值對預測值圖表

圖表 5-10
觀察值對預測值圖表



對於類別依變數，觀察值對預測值圖表會顯示結合訓練樣本和測試樣本的預測虛擬機率集群盒形圖。x 軸對應於觀察反應類別，而圖註對應於預測類別。因此：

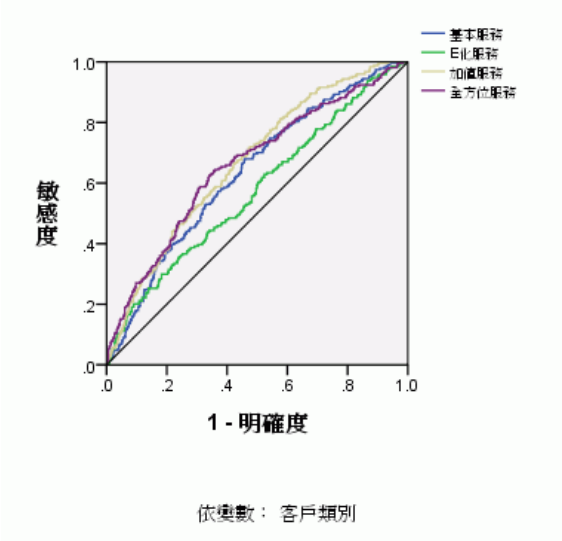
- 對於具有觀察類別「基本服務」的觀察值，最左邊的盒形圖顯示屬於類別「基本服務」的預測虛擬機率。
- 對於具有觀察類別「基本服務」的觀察值，右邊第二個盒形圖顯示屬於類別「電子服務」的預測虛擬機率。

- 對於具有觀察類別「基本服務」的觀察值，第三個盒形圖顯示屬於類別「加值服務」的預測虛擬機率。請記住，在分類表中錯誤分類為「加值服務」與正確分類為「基本服務」的「基本服務」客戶數目大致相同；因此，這個盒形圖概略相當於最左邊的盒形圖。
- 對於具有觀察類別「基本服務」的觀察值，第四個盒形圖顯示屬於類別「總服務」的預測虛擬機率。

由於目標變數中超過兩個類別，因此前四個盒形圖與 0.5 處或任一處的水平線都不保持對稱。因此，不容易針對含有兩個類別以上的目標解釋此圖，因為，檢視其中一個盒形圖的一部分觀察值時，無法判斷其他盒形圖的觀察值相對位置。

ROC 曲線

圖表 5-11
ROC 曲線



ROC 曲線能夠以目視方式，呈現所有可能分割的**敏感度**和**明確性**。其中顯示的圖表有四條曲線，分別表示目標變數的各個類別。

請注意，這張圖表是根據合併的訓練樣本和測試樣本而得。若要產生保留樣本的 ROC 圖表，請將分割變數上的檔案分割，然後在預測虛擬機率上執行 ROC 曲線程序。

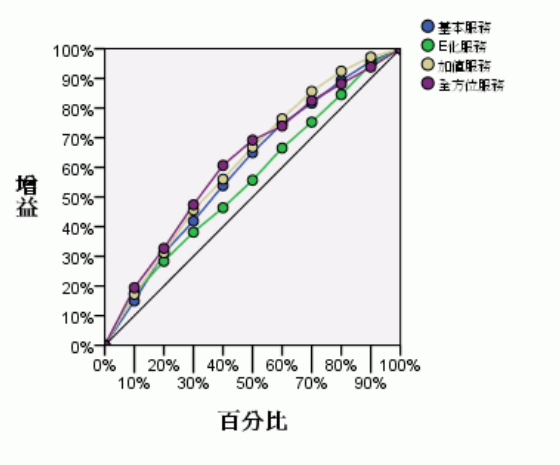
圖表 5-12
曲線下的區域

		區域
客戶類別	基本服務	.635
	E化服務	.573
	加值服務	.668
	全方位服務	.659

曲線下的區域是 ROC 曲線的數值摘要，對於各個類別而言，表中的值表示：出現在該類別中的預測虛擬機率的機率，對於在該類別中隨機選擇的觀察值而言，高於不在該類別中隨機選擇的觀察值。例如，對於在加值服務中隨機選擇的客戶，以及在基本服務、電子服務或總服務中隨機選擇的客戶，拖欠者的模式預測虛擬機率比加值服務客戶高出 0.668 的機率。

累積增益圖表和提升圖表

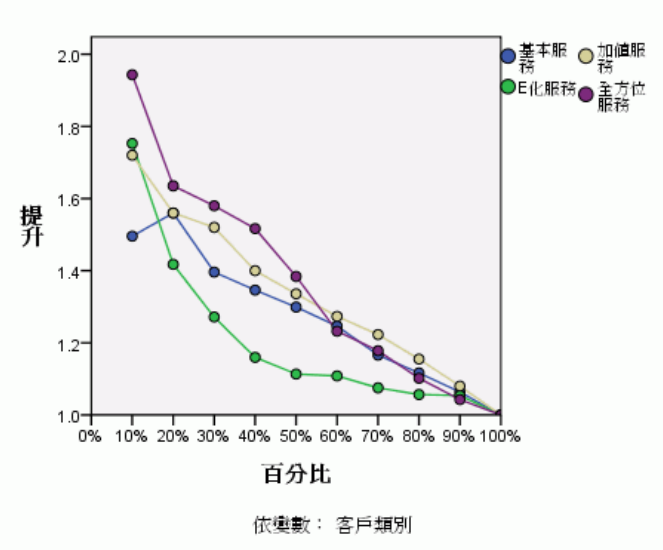
圖表 5-13
累積增益圖表



累積增益圖表以觀察值總數的百分比為目標，顯示指定類別「增益」中觀察值的總數百分比。例如，在「總服務」類別的曲線上，第一個點大約位在 (10%, 20%)，這表示，如果您使用網路評定資料集，並且依照預測虛擬機率「總服務」將所有觀察值排序，則前 10% 會包含大約 20% 實際具有類別「總服務」的觀察值。同樣地，前 20% 會包含大約 30% 的拖欠者，前 30% 的觀察值會包含 50% 的拖欠者，依此類推。如果您選取 100% 的評分資料集，便會取得資料集中所有的拖欠者。

對角線是「基準線」曲線；如果您從評分資料集隨機選擇 10% 的觀察值，則會「增益」10% 實際具有指定類別的所有觀察值。曲線在基準線上方的距離愈遠，則增益愈大。

圖表 5-14
提升圖表



提升圖表衍生自累積增益圖表；y 軸上的值對應於各曲線與基準線的累積增益比例。因此，類別「總服務」提升 10% 便大約是 20%/10% = 2.0。這是另一種檢視累積增益圖表資料的方式。

注意：累積增益圖表和提升圖表是根據合併的訓練樣本和測試樣本而得。

閱讀資料推薦

請參閱下列文字以取得「半徑式函數」的詳細資訊：

Bishop, C. M. 1995. Neural Networks for Pattern Recognition (樣式辨識神經網路), 第三版 ed. Oxford (牛津): Oxford University Press (牛津大學出版部).

Fine, T. L. 1999. Feedforward Neural Network Methodology (前饋神經網路方法論), 第三版 ed. 紐約: Springer-Verlag.

Haykin, S. 1998. Neural Networks: (神經網路)A Comprehensive Foundation (全方位基礎), 第二版 ed. 紐約: Macmillan College Publishing (Macmillan 學院出版部).

Ripley, B. D. 1996. Pattern Recognition and Neural Networks (樣式辨識與神經網路). Cambridge (劍橋): Cambridge University Press (劍橋大學出版部).

Tao, K. K. 1993. A closer look at the radial basis function (RBF) networks (近觀半徑式函數 (RBF) 網路). 於： Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers (第 27 屆 Asilomar 訊號、系統與電腦大會記錄), A. Singh, ed. Los Alamitos, Calif. (加州): IEEE Comput. Soc. Press (IEEE 電腦協會出版社).

Uykan, Z., C. Guzelis, M. E. Celebi, 和 H. N. Koivo. 2000. Analysis of input-output clustering for determining centers of RBFN (決定 RBFN 中心之輸入輸出集群分析). IEEE Transactions on Neural Networks (IEEE 神經網路彙刊), 11, .

附錄
A

範例檔案

與產品同時安裝的範例檔存放在安裝目錄的範例子目錄中。在下列每種語言的「範例」子目錄中存有個別資料夾：英文、法文、德文、義大利文、日文、韓文、波蘭文、俄文、簡體中文、西班牙文和繁體中文。

並非所有範例檔案皆提供各種語言。如果範例檔案沒提供您需要的語言，語言資料夾有英文版的範例檔案。

說明

以下是使用於本文件中不同範例的範例檔之簡要描述。

- **accidents.sav。** 這是有關某保險公司研究年齡和性別風險因子對給定地區汽車意外事件的假設資料檔。每一個觀察值對應至一個年齡類別和性別的交叉分類。
- **adl.sav。** 這是有關致力於確定一個建議中風病患治療類型之效益的假設資料檔。醫師隨機指定女性中風病患至兩個組別之一。第一組接受標準的物理治療，而第二組則接受額外的情緒治療。在治療了三個月後，將每一個病患進行日常活動的能力記分為次序變數。
- **advert.sav。** 這是有關一家零售商致力於調查廣告費與廣告後銷售情形之間的關係的假設資料檔。為了這個目的，他們收集了過往銷售數字和相關的廣告費用。
- **aflatoxin.sav。** 這是有關檢定玉米作物是否有黃麴毒素（一種毒物，其濃度在介於和處於作物產量中都有很大的差異）的假設資料檔。一名穀物加工者收到來自 8 個作物產量各 16 個樣本，並以十億當量 (PPB) 來測量黃麴毒素的水準。
- **aflatoxin20.sav。** 這個資料檔包含由 aflatoxin.sav 取得，來自 4 和 8 作物產量的 16 個樣本，每一個樣本的黃麴毒素測量。
- **anorectic.sav。** 在將厭食/暴食行為症狀學標準化的過程中，研究人員 (Van der Ham, Meulman, Van Strien, 和 Van Engeland, 1997) 研究了 55 個飲食失調的青少年。每個病患在四年之中被訪問四個回合，所以得到總數為 220 的觀察值。在每次觀察中，為病患在 16 種症狀上逐一評分。目前遺漏了第二次訪察的病患 71，第二次訪察的病患 76，以及第三次訪察的病患 47 的症狀分數，因此只剩下 217 個有效觀察值。
- **autoaccidents.sav。** 這是有關一位保險分析師致力於為每個駕駛的汽車意外事件次數建立模式，同時考量駕駛的年齡和性別的假設資料檔。每一個觀察值代表一位不同的駕駛，記錄了駕駛的性別、年齡、和近五年內的汽車意外事故次數。
- **band.sav。** 本資料檔包含某樂團音樂 CD 假設性的每週銷售數字。也包含三個可能預測變數的資料。
- **bankloan.sav。** 這是有關一家銀行致力於減少放款利率預設值的假設資料檔。本檔包含 850 位以前的客戶與現在的準客戶的財務和人口資料。前 700 個觀察值為以前有借貸的客戶。最後 150 個觀察值是銀行需要作信用風險優良與不良分類的準客戶。

- **bankloan_binning.sav。** 這是包含 500 位以前客戶的財務和人口資料的假設資料檔。
- **behavior.sav。**在典型範例 (Price 和 Bouffard, 1974) 中，52 名學生被要求為 15 種情境與 15 種行為組合評等，等級共分為 10 點，從 0 =「非常適當」到 9 =「 非常不適當」。平均值超過個別值，值會被視為相異性。
- **behavior_ini.sav。** 本資料檔包含 behavior.sav 之二維解的起始組態。
- **brakes.sav。** 這是有關一間生產高性能汽車碟型煞車片工廠中品質管制的假設資料檔。資料檔包含由 8 個生產機器分別取得 16 個碟片的直徑測量。煞車的目標直徑是 322 公釐。
- **breakfast.sav。**在經典研究中(Green 和 Rao, 1972)，21 名 Wharton 學院 MBA 學生及其配偶被要求為 15 項早餐食品按喜愛程度分出等級：從 1 = 「最喜愛」到 15 = 「最不喜愛」。他們的喜愛程度分六種不同情況記錄，從「整體喜愛」到「點心，僅配飲料」。
- **breakfast-overall.sav。** 本資料檔只包含第一種情況－「整體喜愛」－所喜愛的早餐項目。
- **broadband_1.sav。** 這是包含全國性寬頻服務地區用戶數目的假設資料檔。本資料檔包含四年期間 85 個地區每月的用戶數目。
- **broadband_2.sav。** 本資料檔與 broadband_1.sav 相同，但多了三個月的資料。
- **car_insurance_claims.sav。** 一個在別處 (McCullagh 和 Nelder, 1989) 出現和分析過，有關汽車損害理賠的資料集。理賠金額的平均數可建立模式為具有 gamma 分配，使用反連結函數將依變數的平均數相關至一被保險人年齡、車輛類型、和車齡的線性組合。提出理賠的數量可以用作尺度權重。
- **car_sales.sav。** 本資料檔包含假設性的銷售估計、定價、和不同的品牌與車輛型式的實體規格。定價和實體規格是由 edmunds.com 和製造商處輪流取得。
- **car_sales_uprepared.sav。** 這是 car_sales.sav 的修改版本，其中不包含任何欄位的轉換版本。
- **carpet.sav。**在一個普遍的範例 (Green 和 Wind, 1973) 中，計劃銷售全新地毯清潔機的公司想要檢驗影響消費者偏好的五個因子—包裝設計、品牌名稱、價格、「優秀家用品」獎章及退費保證。包裝設計有三個因子水準，每個水準中的清潔刷位置都不相同；三個品牌名稱 (K2R、Glory、及 Bissell)；三個價格水準；且最後兩個因子各有兩個水準 (無論無或有)。十名消費者將這些因子所定義的 22 種組合分級。「偏好」變數包含每個組合平均排名的等級。排名數值較小者會對應高偏好程度。這個變數反映每個組合偏好的整體量數。
- **carpet_prefs.sav。** 本資料檔是根據 carpet.sav 所描述的相同範例，但它包含 10 個消費者每一個人的實際等級。消費者被要求將 22 個產品組合從最喜歡排列到最不喜歡。變數「PREF1」到「PREF22」包含相關組合的識別碼，如 carpet_plan.sav 中所定義。
- **catalog.sav。** 本資料檔包含郵購公司銷售三項產品的每月假設銷售數字。也包含五個可能預測變數的資料。
- **catalog_seasfac.sav。** 本資料檔與 catalog.sav 相同，不過多了一組由「週期性分解」程序所計算的週期性因子以及隨附的資料變數。
- **cellular.sav。** 這是有關一家手機公司致力於減少顧客不忠的假設資料檔。顧客不忠傾向分數套用於帳戶，範圍由 0 至 100。帳戶分數 50 或以上有可能正尋求變更供應商。

- **ceramics.sav。** 這是有關一家製造商致力於確定一種新的優良合金是否較標準的合金有較大的耐熱性的假設資料檔。每一個觀察值代表對合金之一的不同檢定；記錄了讓軸承失效的溫度。
- **cereal.sav。** 這是有關對 880 人的早餐喜好進行訪談的假設資料檔，也記下他們的年齡、性別、婚姻狀況、和是否有活躍的生活型態 (根據他們是否一週運動兩次)。每一個觀察值代表一位不同的應答者。
- **clothing_defects.sav。** 這是有關一家服裝工廠品質管制過程的假設資料檔。由該工廠所生產的每一批產品中，檢查員取出一件服裝的樣本並計算不合格的服裝個數。
- **coffee.sav。** 本資料檔是關於六種冰咖啡品牌的感覺印象(Kennedy, Riquier, 和 Sharp, 1996)。對 23 種冰咖啡中每一種的印象屬性，由群眾來選取依其屬性描述的所有品牌。該六種品牌已標示為 AA、BB、CC、DD、EE、和 FF，以保持機密。
- **contacts.sav。** 這是有關一群公司電腦銷售代表聯絡清單的假設資料檔。每一個聯絡人依他們在公司所服務的部門及其公司的等級而分類。最後一次銷售的金額、到最後一次銷售的時間、和該聯絡人公司的規模也都被列入記錄。
- **creditpromo.sav。** 這是有關一家百貨公司致力於評估近期信用卡促銷活動效果的假設資料檔。為達此目標，隨機選取了 500 位持卡人。有半數收到廣告，促銷在未來三個月購買將獲得降低利率的優惠。半數收到標準的週期性廣告。
- **customer_dbase.sav。** 這是有關一家公司致力於使用其資料倉庫的資訊來對最有可能回應的客戶提供優惠的假設資料檔。隨機選取客戶庫的子集，提供優惠，再將他們的回應記錄下來。
- **customer_information.sav。** 本檔案是包含客戶郵寄資訊的假設資料檔，例如姓名和地址。
- **customer_subset.sav。** 80 個 customer_dbase.sav 的觀察值子集。
- **customers_model.sav。** 本檔案包含一市場行銷活動所鎖定之個人的假設資料。這些資料包含人口資訊、購買歷史摘要、和每一個人是否對該活動有回應。每一個觀察值代表一位不同的個人。
- **customers_new.sav。** 本檔案包含一市場行銷活動潛在候選人之個人的假設資料。這些資料包含每一位個人的人口資訊和購買歷史摘要。每一個觀察值代表一位不同的個人。
- **debate.sav。** 這是有關一項政治辯論會參與者辯論前和辯論後接受調查之成對反應的假設資料檔。每一個觀察值對應至一位不同的應答者。
- **debate_aggregate.sav。** 這是將 debate.sav 中之反應作整合的假設資料檔。每一個觀察值對應至辯論前和辯論後對偏好之交叉分類的反應。
- **demo.sav。** 這是有關提供郵寄每月優惠之購買客戶資料庫的假設資料檔。記錄了客戶是否對該優惠回應，以及各種的人口資訊。
- **demo_cs_1.sav。** 這是有關一家公司致力於匯編調查資訊資料庫之第一步的假設資料檔。每一個觀察值對應至一個不同的城市，也記錄了其地區、省、區、和城市識別。
- **demo_cs_2.sav。** 這是有關一家公司致力於匯編調查資訊資料庫之第二步的假設資料檔。每一個觀察值對應至在第一步中選取的城市中的一個不同的家庭單位，也記錄了其地區、省、區、分區、和單位識別。也納入了由該設計的前兩階段所得之取樣資訊。
- **demo_cs.sav。** 這是包含以複合取樣設計所收集之調查資訊的假設資料檔。每一個觀察值對應至一個不同的家庭單位，也記錄了各種的人口和取樣資訊。

- **dmdata.sav。** 這是包含直效行銷公司之人口和購買資訊的假設資料檔。dmdata2.sav 包含收到測試郵件的連絡人子集資訊，而 dmdata3.sav 則包含剩下未收到測試郵件的連絡人資訊。
- **dietstudy.sav。** 本假設資料檔包含對「Stillman 飲食法」(Rickman, Mitchell, Dingman, 和 Dalen, 1974) 研究的結果。每一個觀察值對應至一個不同的受試者，並記錄下他或她飲食法前、後之體重 (磅) 和三酸甘油酯水準 (毫克/100 毫升)。
- **dvdplayer.sav。** 這是有關新 DVD 播放器開發的假設資料檔。市場行銷團隊使用原型收集了焦點組別資料。每一個觀察值對應至不同調查到的使用者，並記錄下一些有關他們的人口資訊和他們對有關原型問題的回應。
- **german_credit.sav。** 本資料檔取自 (Blake 和 Merz, 1998) 艾文 (Irvine) 在加州大學機器學習資料庫儲存器的「德國信用」資料集。
- **grocery_1month.sav。** 本假設資料檔是將 grocery_coupons.sav 資料檔和每週購買的「彙總」，因此每一個觀察值對應至一個不同的客戶。結果部份每週變更的變數消失了，而目前所記錄的銷售量是在研究的四週期間銷售量之總和。
- **grocery_coupons.sav。** 這是包含某連鎖雜貨店想要知道他們客戶購買習慣所收集之調查資料的假設資料檔。每一個客戶被追蹤了四週，每一個觀察值對應至一個不同的客戶-週，並記錄有關客戶在何處及如何購物的資訊，包含那一週在雜貨店花了多少錢。
- **guttman.sav。**Bell(Bell, 1961) 以此表說明可能的社會團體。Guttman (Guttman 值, 1968) 過去曾使用此表的一部分，在這部分中有 5 個變數，分別說明 7 個理論社會團體的社會互動、團體歸屬感、成員實際接觸和關係正式性，而這 7 個群組包括：群眾 (例如，足球場上的人)、觀眾 (例如在戲院中和課堂上的人)、公眾 (例如，報紙讀者和電視觀眾)、暴民 (和群眾相似，但互動較為激烈)、原級團體 (親密性)、次級團體 (自願性) 和現代社群 (因親密的身體接近而導致鬆散的結盟和特殊服務的需求)。
- **health_funding.sav。** 這是包含醫療保健基金 (每 100 個人口的金額)、疾病率 (每 10,000 個人口的比率)、造訪醫療保健機構的比例 (每 10,000 個人口的比率) 的假設資料檔。每一個觀察值代表一個不同的城市。
- **hivassay.sav。** 這是有關一家製藥實驗室致力於開發一種偵測 HIV 感染快速檢驗的假設資料檔。檢驗結果是八個紅色加深的陰影，陰影愈深表示感染的可能性愈大。進行了一項實驗室的試驗，在 2,000 個血液樣本中，有半數遭到 HIV 的感染，而半數則未感染。
- **hourlywagedata.sav。** 這是有關在辦公室和醫院任職的護士依經驗水準不同之鐘點費的假設資料檔。
- **insurance_claims.sav.** 這是有關一家保險公司想要建立模式來標示可疑及可能的詐欺理賠之假設資料檔。每一個觀察值代表個不同的理賠。
- **insure.sav。** 這是有關一家保險公司正在研究表示客戶是否必定理賠 10 年壽險合約之風險因子的假設資料檔。在資料檔中的每一個觀察值代表二份合約，其一記錄了理賠而另一則否，二者的年齡和性別相符。
- **judges.sav。** 這是有關受過訓練的裁判 (加上一位熱心人士) 為 300 個體操表演評分的假設資料檔。每一列代表一個不同的表演；裁判們觀看相同的表演。
- **kinship_dat.sav。**Rosenberg 與 Kim (Rosenberg 和 Kim, 1975) 致力於分析 15 個親屬關係稱呼 (姑/姨、兄弟、堂/表兄弟姐妹、女兒、父親、孫女、祖父、祖母、孫子、母親、姪子/外甥、姪女/外甥女、姐妹、兒子、叔/舅父)。他們請四組大學生 (兩組女性、兩組男性) 根據其相似性來分類整理這些稱謂。他們請其中兩組 (一組

女性、一組男性）作兩次分類整理，第二次要根據與第一次不同的準則進行分類整理。因此，總共得到六個「來源」。每一個來源對應至一個 15 × 15 的相似性矩陣，其儲存格等於來源中人數減去物件在該來源中分為同組的次數。

- **kinship_ini.sav。** 本資料檔包含 kinship_dat.sav 之三維解的起始組態。
- **kinship_var.sav。** 本資料檔包含自變數「性別」、「世代」、和可用來解讀 kinship_dat.sav 解答維度的（分離）「度」。尤其，它們可用來將解答空間限制為這些變數的線性組合。
- **marketvalues.sav。** 本資料檔有關於一項在伊立諾州阿爾岡京 (Algonquin, Ill.) 的新屋開發案自 1999 年至 2000 年之房屋銷售情況。這些銷售與公共記錄有關。
- **nhis2000_subset.sav。** 「國民健康訪問調查 (NHIS)」為美國民間人口的一大型民眾調查。其以具全國代表性的家庭為樣本，面對面的完成訪問。而取得各家庭中成員的人口統計學資訊及健康行為、健康狀態方面等觀察報告。本資料檔包含一個 2000 年調查資訊的子集。國家衛生統計中心。2000 年「國民健康訪問調查 (NHIS)」。公用資料檔案和文件。ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/。2003 年曾存取。
- **ozone.sav。** 本資料包含對六個氣象變數所作的 330 個觀察值，以自其餘的變數中預測臭氧濃度。先前研究人員中，(Breiman 和 Friedman 檢定(F), 1985)、(Hastie 和 Tibshirani, 1990) 在這些會阻礙標準迴歸方式的變數中發現非線性。
- **pain_medication.sav。** 本假設資料檔包含治療慢性關節炎疼痛之消炎藥物臨床試驗的結果。特別關注於藥物發生作用的時間以及它是如何與現用藥物作比較。
- **patient_los.sav。** 本假設資料檔包含對因可能為心肌梗塞 (MI，或「心臟病」) 入院病患的治療記錄。每一個觀察值對應至一個不同的病患並記錄許多與其留院期間有關的變數。
- **patlos_sample.sav。** 本假設資料檔包含病患在為心肌梗塞 (MI，或「心臟病」) 治療期間接受血栓溶解治療的治療記錄樣本。每一個觀察值對應至一個不同的病患並記錄許多與其留院期間有關的變數。
- **polishing.sav。** 這是取自「資料和故事圖書館」的「Nambeware 打磨時間」資料檔。它是有關一家金屬餐具製造商 (Nambe Mills, 聖塔非, 新墨西哥州) 致力於規劃其生產排程。每一個觀察值代表生產線上一個不同的產品。每一個產品都記錄下直徑、打磨時間、價格、和產品類別。
- **poll_cs.sav。** 這是有關民意測驗專家致力於確定交付立法之前公眾對法案支持水準的假設資料檔。觀察值對應至登記選民。每一個觀察值記錄下選民的郡、鎮、和他居住的鄰近範圍。
- **poll_cs_sample.sav。** 本假設資料檔包含列於 poll_cs.sav 中的選民樣本。樣本是根據在 poll.csplan 計劃檔中指定的設計來取得，而本資料檔記錄了包含機率和樣本權重。不過，請注意，由於取樣計劃採用到機率 - 比例 - 大小 (PPS) 方法，也用到一個包含聯合選擇機率的檔案 (poll_jointprob.sav)。其他與選民人口及其對提議法案之意見有關的變數都在取樣後收集並加入資料檔中。
- **property_assess.sav。** 這是有關郡財產估價人員致力於對限定資源保持財產價值評估維持最新的假設資料檔。觀察值對應至郡內過去一年銷售的財產。資料檔中的每一個觀察值記錄了財產所在的鎮、上次訪查該財產的估價人員、自那次評估後經過的時間、當時定的估價、和該財產銷售價值。

- **property_assess_cs.sav。** 這是有關州財產估價人員致力於對限定資源保持財產價值評估維持最新的假設資料檔。觀察值對應至州中的財產。資料檔中的每一個觀察值記錄了郡、鎮、和財產所在的鄰近範圍、自最後一次評估後經過的時間、和當時定的估價。
- **property_assess_cs_sample.sav。** 本假設資料檔包含列於 property_assess_cs.sav 中的財產樣本。樣本是根據在 property_assess.csplan 計劃檔中指定的設計來取得，而本資料檔記錄了包含機率和樣本權重。另外的變數「目前價值」是在取樣後收集並加入資料檔中。
- **recidivism.sav。** 這是有關政府法令執行機構致力於瞭解其轄區內之再犯率的假設資料檔。每一個觀察值對應至一個先前的違法者並記錄其人口資訊、第一次犯罪的一些細節、然後是直到第二次被捕的時間（如果它發生在第一次被捕的兩年之內）。
- **recidivism_cs_sample.sav。** 這是有關政府法令執行機構致力於瞭解其轄區內之再犯率的假設資料檔。每一個觀察值對應到一個先前的違法者，在 2003 年六月第一次被捕後釋放，並記錄其人口資訊、第一次犯罪的一些細節、和第二次被捕日期（如果它發生在 2006 年六月之前）。違法者是根據在 recidivism_cs.csplan 中所指定的取樣計劃之樣本部門來選取；由於取樣計劃採用到機率 - 比例 - 大小 (PPS) 方法，也用到一個包含聯合選擇機率的檔案 (recidivism_cs_jointprob.sav)。
- **rfm_transactions.sav。** 本檔案是包含購買交易資料的假設資料檔，包括購買日期、購買項目及每一項交易的金額。
- **salesperformance.sav。** 這是有關評估兩個新售貨員訓練課程的假設資料檔。六十個員工，分成三個組別，全部接受標準訓練。此外，組別二得到技術訓練；組別三則是實務輔導簡介。每一個員工在訓練課程結束時接受測驗並記錄他們的分數。在資料檔中每一個觀察值代表一個不同的訓員，並記錄他們所分派的組別和他們在測驗中得到的分數。
- **satisf.sav。** 這是有關一家零售公司在 4 個商店位置所作之滿意度調查的假設資料檔。總共有 582 位客戶接受調查，每一個觀察值代表一位客戶的反應。
- **screws.sav。** 這個資料檔包含螺絲釘、螺栓、螺帽和圖釘之特色的資訊(Hartigan, 1975)。
- **shampoo_ph.sav。** 這是有關一家美髮產品工廠品質管制過程的假設資料檔。在固定的時間間隔，記錄下六個不同輸出批次的測量和它們的 pH 值。目標範圍是 4.5 - 5.5。
- **ships.sav。** 一個在別處 (McCullagh et al。, 1989) 出現和分析過，有關商船因風浪所造成損壞的資料集。事件次數可建立模式為以 Poisson 率發生，給定船型、建造期間、和服務期間。以因子交叉分類所形成的表格的每一個儲存格服務月數的整合，提供了暴露於風險之值。
- **site.sav。** 這是有關一家公司致力於為事業擴展選擇新地點的假設資料檔。您們僱請兩位顧問分別評估該地點，除了一份廣泛的報告之外，他們還要將每個地點摘要為前景「佳」、「可」、或「差」。
- **smokers.sav。** 本資料檔是由「1998 年全國家庭毒品濫用調查」中摘錄，且是美國家庭的機率樣本。(http://dx.doi.org/10.3886/ICPSR02934) 因此，在分析本資料檔的第一步應該是將資料加權以反映母群體傾向。
- **stroke_clean.sav。** 本假設資料檔包含一個醫療資料庫，其在以「資料準備」選項中的程序清理之後的狀態。
- **stroke_invalid.sav。** 本假設資料檔包含一個醫療資料庫的起始狀態並包含幾個資料輸入錯誤。

- **stroke_survival。** 本假設資料檔是有關缺血性中風的病患，其在結束康復計畫後存活時間方面，面臨許多挑戰。中風後，記載了心肌梗塞、缺血性中風、或出血性中風的發生，以及事件記錄的時間。由於它只包含在康復計劃所管制的中風存活的病患，此樣本的左側被截斷。
- **stroke_valid.sav。** 本假設資料檔包含一個醫療資料庫，在其值以「驗證資料」程序檢查之後的狀態。它仍包含可能的異常觀察值。
- **survey_sample.sav。** 本資料檔包含調查資料，包括人口資料和各種態度測量。雖然已修改一些資料數值，且為人口資料之目的新增了一些額外的虛構變數，但是資料仍是以「1998 NORC 基本社會調查」的變數子集為基礎。
- **telco.sav。** 這是有關一家電信公司致力於在客戶庫中減少顧客不忠的假設資料檔。每一個觀察值對應至一位不同的客戶並記錄不同的人口資料和服務使用方式資訊。
- **telco_extra.sav。** 本資料檔類似於 telco.sav 資料檔，但「任期」的對數轉換客戶花費變數已予刪除，並更換為標準的對數轉換客戶花費變數。
- **telco_missing.sav。** 本資料檔是 telco.sav 資料檔的子集，不過某些人口資料值已更換為遺漏值。
- **testmarket.sav。** 本假設資料檔有關於一家速食連鎖店計劃在菜單中加入新的項目。有三個可能的活動來促銷此新產品，所以該新項目在幾個隨機選取市場中的地點作介紹。在每一個地點使用不同的促銷，並記錄該新項目前四週的每週銷售量。每一個觀察值對應至一個不同的地點-週。
- **testmarket_1month.sav。** 本假設資料檔是將 testmarket.sav 資料檔和每週購買的「彙總」，因此每一個觀察值對應至一個不同的客戶。結果部份每週變更的變數消失了，而目前所記錄的銷售量是在研究的四週期間銷售量之總和。
- **tree_car.sav。** 這是包含人口資料和車輛購買價格資料的假設資料檔。
- **tree_credit.sav。** 這是包含人口資料和銀行放款歷史資料的假設資料檔。
- **tree_missing_data.sav** 這是包含有大量遺漏值的人口資料和銀行放款歷史資料的假設資料檔。
- **tree_score_car.sav。** 這是包含人口資料和車輛購買價格資料的假設資料檔。
- **tree_textdata.sav。** 一個只有兩個變數的簡單資料檔，主要目的在顯示變數預設狀態（在指定量測水準和數值標記之前）。
- **tv-survey.sav。** 這是有關一家電視製片廠考量是否要延長一個成功節目的播送所作之調查的假設資料檔。有 906 位應答者被問到在不同的狀況下他們是否願意觀看這個節目。每一列代表一個不同的應答者；每一行為一個不同的狀況。
- **ulcer_recurrence.sav。** 本檔案包含一項用來比較兩種防止潰瘍復發治療法功效之研究的部分資訊。它是很好的區間受限資料範例，且已在別處 (Collett, 2003) 出現和分析過。
- **ulcer_recurrence_recoded.sav。** 本檔案是將 ulcer_recurrence.sav 的資訊重新組織，以讓您為此研究的每一個區間事件機率而非只是研究目的事件機率建立模式。它已在別處 (Collett et al。, 2003) 出現和分析過。
- **verd1985.sav。** 本資料檔有關於一項調查 (Verdegaal, 1985)。在調查中記錄了來自 15 個受訪者對 8 個變數的回應。所需的變數被分成三組。集 1 包括 age 和 marital，集 2 包括 pet 和 news，集 3 包括 music 和 live。Pet 調整為多重名義量數，age 調整為次序量數，其他的變數調整為單一名義量數。

- **virus.sav。** 這是有關一家網際網路服務提供者致力於在其網路上判斷病毒之影響的假設資料檔。他們在其網路上追蹤從發現病毒直到控制威脅的這段時間，被病毒感染之電子郵件的流量（約略）百分比。
- **wheeze_steubenville.sav。** 這是空氣污染對兒童健康之影響 (Ware, Dockery, Spiro III, Speizer, 和 Ferris Jr., 1984) 縱向研究的子集。本資料包含來自俄亥俄州 Steubenville，年齡 7、8、9 和 10 歲兒童的氣喘聲狀態之重複二元測量，以及其母親在本研究的第一年是否抽煙的固定記錄。
- **workprog.sav。** 這是有關一項政府職業計劃，設法將弱勢民眾安置到較好之工作的假設資料檔。一個樣本的可能計劃參與者被追蹤，他們之中某些被選取加入本計劃，而其他的則否。每一個觀察值代表一位不同的計劃參與者。

附錄
B

Notices

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided “AS IS”, without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shmtl.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, http://www.winwrap.com.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

參考書目

Bell, E. H. 1961. Social foundations of human behavior: (人類行為之社會基礎)Introduction to the study of sociology (社會學研究簡介). 紐約: Harper & Row.

Bishop, C. M. 1995. Neural Networks for Pattern Recognition (樣式辨識神經網路), 第三版 ed. Oxford (牛津): Oxford University Press (牛津大學出版部).

Blake, C. L., 和 C. J. Merz. 1998. "UCI Repository of machine learning databases (機器學習資料庫 UCI 儲存器)." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., 和 J. H. Friedman 檢定(F). 1985. Estimating optimal transformations for multiple regression and correlation (估計多重迴歸與相關之最適轉換). Journal of the American Statistical Association (美國統計協會彙報), 80, .

Collett, D. 2003. Modelling survival data in medical research (模式化醫學研究中的存活資料), 2 ed. Boca Raton: Chapman & Hall/CRC.

Fine, T. L. 1999. Feedforward Neural Network Methodology (前饋神經網路方法論), 第三版 ed. 紐約: Springer-Verlag.

Green, P. E., 和 V. Rao. 1972. Applied multidimensional scaling (應用多元尺度方法). Hinsdale, Ill.: Dryden Press (Dryden 出版社).

Green, P. E., 和 Y. Wind. 1973. Multiattribute decisions in marketing: (市場行銷之多重屬性決策)A measurement approach (測量方法). Hinsdale, Ill.: Dryden Press (Dryden 出版社).

Guttman 值, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points (尋找組態點最小座標空間之一般非計量技巧). Psychometrika (心理學計量報導), 33, .

Hartigan, J. A. 1975. Clustering algorithms (集群演算法). 紐約: John Wiley and Sons.

Hastie, T., 和 R. Tibshirani. 1990. Generalized additive models (概化附加模式). 倫敦: Chapman and Hall.

Haykin, S. 1998. Neural Networks: (神經網路)A Comprehensive Foundation (全方位基礎), 第二版 ed. 紐約: Macmillan College Publishing (Macmillan 學院出版部).

Kennedy, R., C. Riquier, 和 B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research (市場研究類別資料之對應分析實際應用). Journal of Targeting, Measurement, and Analysis for Marketing (市場行銷之目標訂定、測量與分析雜誌), 5, .

McCullagh, P., 和 J. A. Nelder. 1989. 概化線性模式, 第二版 ed. 倫敦: Chapman & Hall.

Price, R. H., 和 D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior (行為適切性與情境限制作為社會行為維度). Journal of Personality and Social Psychology (人格與社會心理學雜誌), 30, .

Rickman, R., N. Mitchell, J. Dingman, 和 J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet (實行 Stillman 飲食法期間血膽固醇改變情形). Journal of the American Medical Association (美國醫學協會彙報), 228, .

Ripley, B. D. 1996. Pattern Recognition and Neural Networks (樣式辨識與神經網路). Cambridge (劍橋): Cambridge University Press (劍橋大學出版部).

Rosenberg, S., 和 M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research (多變量研究中作為資料收集程序之排序方法). Multivariate Behavioral Research (多變量行為研究), 10, .

Tao, K. K. 1993. A closer look at the radial basis function (RBF) networks (近觀半徑式函數 (RBF) 網路). 於： Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers (第 27 屆 Asilomar 訊號、系統與電腦大會記錄), A. Singh，ed. Los Alamitos, Calif. (加州): IEEE Comput. Soc. Press (IEEE 電腦協會出版社).

Uykan, Z., C. Guzelis, M. E. Celebi, 和 H. N. Koivo. 2000. Analysis of input-output clustering for determining centers of RBFN (決定 RBFN 中心之輸入輸出集群分析). IEEE Transactions on Neural Networks (IEEE 神經網路彙刊), 11, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, 和 H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: (依經驗法將青少年飲食異常次組別化)A longitudinal perspective (縱向觀點). British Journal of Psychiatry (英國心理學雜誌), 170, .

Verdegaal, R. 1985. Meer sets analyse voor kwalitatieve gegevens (in Dutch) (更多性質資料的集合分析 (荷蘭文)). Leiden (萊頓): Department of Data Theory, University of Leiden (萊頓大學資料理論系).

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, 和 B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities (六個城市中的兒童二手煙、氣體煮食與呼吸道健康). American Review of Respiratory Diseases (美國呼吸道疾病評論), 129, .

索引