IBM SPSS Neural Networks 21

注 ： この情報とサポートされている製品をご使用になる前に、「注意事項」(p.103)の一般情報をお読みください。

本版は IBM® SPSS® Statistics 21 ,および新版で指示されるまで後続するすべてのリリースおよび変更に対して適用されます。

Adobe 製品のスクリーンショットは Adobe Systems Incorporated の許可を得て転載しています。

Microsoft 製品のスクリーンショットは Microsoft 社の許可を得て転載しています。

はじめに

IBM® SPSS® Statistics は、データ分析の包括的システムです。Neural Networks は、このマニュアルで説明されている追加の分析手法を提供するオプションのアドオン モジュールです。Neural Networks アドオン モジュールは SPSS Statistics Core システムと組み合わせて使用し、Core システムに 完全に統合されます。

IBM Business Analytics について

IBM Business Analytics ソフトウェアは、意思決定者がビジネス パフォーマンスを向上させるために信頼する完全で、一貫した正確な情報を提供します。ビジネス インテリジェンス、予測分析、財務実績および戦略管理、および 分析アプリケーションの包括的なポートフォリオを利用することによって、現在の実績を明確、迅速に理解し、将来の結果を予測することができます。豊富な業界のソリューション、実績ある実例、専門サービスと組み合わせ、さまざまな規模の組織が、高い生産性を実現、意思決定を自信を持って自動化し、より良い決定をもたらします。

このポートフォリオの一部として、IBM SPSS Predictive Analytics ソフトウェアを使用する組織は、将来のイベントを予測し、その洞察に基づいて積極的に行動し、より優れた業績を実現することができます。全世界の企業、政府、学術分野のお客様が IBM SPSS の技術を活用し、不正行為を減少させ、リスクを軽減させながら、顧客の獲得、保持、成長において、競争優位を高めることができます。IBM SPSS ソフトウェアを日々の業務に取り入れることによって、組織は業務目標を達成し、大きな競争的優位を獲得することができるよう、意思決定を方向付け、自動化することができるようになります。お問い合わせは、http://www.ibm.com/spss を参照してください。

テクニカル サポート

テクニカル サポートのサービスをご利用いただけます。IBM Corp.製品の使用方法や、対応しているハードウェア環境へのインストールに関して問い合わせることもできます。テクニカル サポートの詳細については、IBM Corp. Web サイト（http://www.ibm.com/support）を参照してください。連絡の際は、所属団体名、サポート契約などを確認できるよう、あらかじめ手元にご用意ください。

学生向けテクニカル サポート

IBM SPSS ソフトウェア製品の Student 版、アカデミック版、Grad パック版を使用している学生の場合、学生用の特別オンライン ページ、Solutions for Education (http://www.ibm.com/spss/rd/students/) ページを参照してください。大学提供の IBM SPSS ソフトウェアのコピーを使用している場合、大学の IBM SPSS 製品コーディネータにお問い合わせください。

カスタマ サービス

配送やアカウントに関するご質問は、お近くの営業所にお問い合わせください。お問い合せの際には、シリアル番号をご用意ください。

トレーニング セミナー

IBM Corp. では一般公開およびオンサイトで トレーニング セミナーを実施しています。セミナーでは実践的な講習を行います。セミナーは主要都市で定期的に開催されます。セミナーに関する詳細については、http://www.ibm.com/software/analytics/spss/training を参照してください。

内容

パート I: ユーザー ガイド

パート II: 例

付録

パート I:
ユーザー ガイド

章

1

Neural Networks の概要

ニューラル ネットワークは、その能力、柔軟性、使いやすさで多くの予測データ マイニング アプリケーションにとって優れたツールです。予測ニューラル ネットワークは、基本的なプロセスが複雑なアプリケーションで特に役に立ちます。例を次に示します。

- 生産の合理化および配送コストの改善に対する顧客の需要を予測する。
- ダイレクト メールによるマーケティングに対する応答確率を予測し、メーリング リストに掲載されているどの世帯に割り引きを提供するか判断する。
- 申請者を点数付けし、貸し付け枠拡大のリスクを判断する。
- 保険金請求データベース内の不正な取り引きを検出する。

多層パーセプトロン (MLP) ネットワークや **放射基底関数 (RBF)** ネットワークなどの予測アプリケーションで使用されるニューラル ネットワークは、モデルで予測される結果が目標変数の既知の値と比較できるという意味で監視されています。Neural Networks オプションを使用すると、MLP ネットワークおよび RBF ネットワークを適合させ、その結果生じた得点付けのためのモデルを保存します。

ニューラル ネットワークとは

ニューラル ネットワークという用語は、大まかに関連付けられたモデルのファミリに適用され、脳機能に由来する大きなパラメータ領域、柔軟な構造が特徴です。関連する用語は元の用語を反映していますが、ファミリが大きくなるにつれ、新しいモデルの多くは関連のないアプリケーションに対して設計されていました。

ニューラル ネットワークの固有の定義は、使用されるフィールドと同様にさまざまなものがあります。モデルのファミリ全体を適切に表す単一の定義はありませんが、ここでは次の説明を考慮します。(Haykin, 1998)

ニューラル ネットワークは、経験に基づいた情報を保存して利用するための、自然な傾向を持つ大規模な並列分散プロセッサです。ニューラル ネットワークは、次の 2 点において脳と似ています。

- 情報は、学習プロセスを介してネットワークが取得します。
- シナプスの重みとして知られるインターニューロン接続の強度を使用し、情報を保存します。

この定義が限定的だと考えられる理由の詳細は、(Ripley, 1996)を参照してください。

この定義によってニューラル ネットワークと従来の統計方法とを区別するために、ここでは定義の実際の文と同じように重要なことが説明されているわけではありません。たとえば、従来の線型回帰モデルは最小 2 乗法 を使用して情報を取得し、回帰係数でその情報を保存します。この点で、これはニューラル ネットワークです。実際、線型回帰は特定のニューラル ネットワークの特別なケースであるといえます。ただし、線型回帰には固定されたモデル構造およびデータから学習する前に与えられた仮定のセットが含まれています。

それに対し、前述の定義は最小限のモデル構造および仮定を要求します。そのためニューラル ネットワークは、従属変数と独立変数の間の特定の関係を事前に仮定することなく、幅広い統計モデルを概算することができます。また、関係の形式は学習プロセス時に定義されます。従属変数および独立変数間の線型関係が適切な場合、ニューラル ネットワークの結果は線型回帰モデルの結果に近くなります。非線型関係がより適切である場合、ニューラル ネットワークは自動的に「適切な」モデル構造に近づきます。

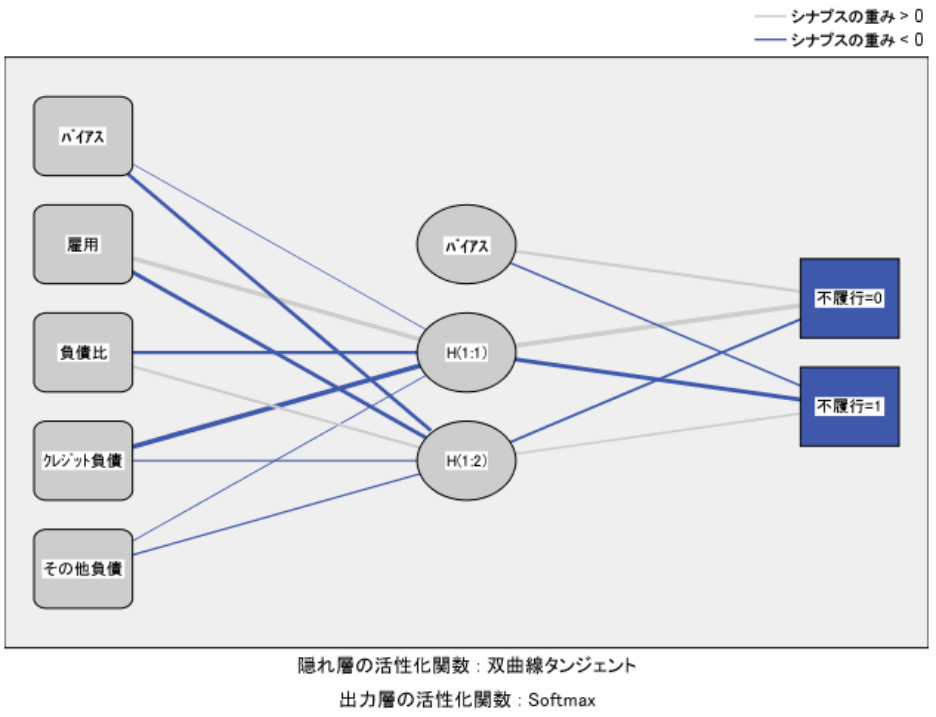
この柔軟性に対する代償として、ニューラル ネットワークのシナプスの重みを容易に解釈することはできません。つまり、従属変数および独立変数間の関係を生成する基本のプロセスを説明する場合、従来の統計モデルを使用するとより正確に説明できます。ただし、モデルの解釈のしやすさが重要でない場合、ニューラル ネットワークを使用して良いモデルの結果を迅速に取得することができます。

ニューラル ネットワークの構造

ニューラル ネットワークは最小限のモデル構造および仮定を要求し、一般的な **ネットワーク アーキテクチャ**を理解するのに役立ちます。多層パーセプトロン (MLP) ネットワークや放射基底関数 (RBF) ネットワークは目標変数（出力変数とも呼ばれる）の予測誤差を最小化する予測変数（入力変数または独立変数とも呼ばれる）の関数です。

製品に付属している bankloan.sav データセットを考えます。そこでローン申請者の中から潜在的な債務不履行者を識別する必要があります。こうした問題に適用される MLP ネットワークまたは RBF ネットワークは、債務不履行者を予測する場合の誤差を最小化する測定の関数です。次の図は、この関数形式の関連付けに役に立ちます。

図 1-1
隠れ層を含むフィードフォワード アーキテクチャ



ネットワーク フローはフィードバック ループなしで入力層から出力層へ接続するため、この構造は **フィードフォワード アーキテクチャ**として知られています。この図では、

- **入力層**には予測変数が含まれています。
- **隠れ層**には、観測不可能なノードまたは単位が含まれています。それぞれの隠れた単位の値は予測変数の関数です。正確な関数形式は、一部はネットワークの種類に、そしてまた一部はユーザーが管理可能な指定によって決まります。
- **出力層**には、応答が含まれます。債務不履行の履歴は 2 つのカテゴリを持つカテゴリ変数であるため、2 つの指示変数として記録されます。それぞれの出力単位の値は隠れた単位の関数です。正確な関数形式は、一部はネットワークの種類に、そしてまた一部はユーザーが管理可能な指定によって決まります。

MLP ネットワークによって、2 番目の隠れ層を作成します。その場合、2 番目の隠れ層の各単位は、最初の隠れ層の単位の関数で、それぞれの応答は 2 番目の隠れ層の単位の関数です。

章 2

多層パーセプトロン

多層パーセプトロン（MLP）手続きは、予測変数の値に基づいて、1 つ以上の従属（目標）変数に対する予測モデルを生成します。

例。 MLP 手続きを使用した 2 つのシナリオを次に示します。

銀行の融資担当者は、債務不履行になる可能性がある人物を示す特徴を特定し、その特徴を使用して信用リスクの良し悪しを識別する必要があります。過去の顧客のサンプルを使用して、多層パーセプトロンを学習し、過去の顧客のホールドアウト サンプルを検証し、ネットワークを使用して見込み客を信用リスクが良い客と悪い客に分類します。

病院のシステムの場合、心筋梗塞（MI または「心臓発作」）の治療で入院している患者の入院費用や期間を記録追跡できます。これらの測定の正確な推定値を取得することで、患者が治療を受けることができるベッド数を適切に管理することができます。心筋梗塞の治療を受けた患者のサンプルの治療記録を使用して、管理者はネットワークを学習し入院の費用および期間を予測することができます。

従属変数。 従属変数は次のものを使用できます。

- **名義データ.** 値がランキングなどを持たないカテゴリを表しているとき､名義（変数）として取り扱うことができます。たとえば、従業員の会社の所属などです。名義変数の例としては、地域やジップ コードや所属宗教などがあります。
- **順序データ.** 値がランキングをもったカテゴリを表しているとき、変数を順序として取り扱うことができます。たとえば、「かなり不満」から「かなり満足」までのようなサービス満足度のレベルなどです。順序変数の例としては、満足度や信頼度を表す得点や嗜好得点などです。
- **スケール データ.** 値が有意な基準を持った順序カテゴリを表しているとき、変数をスケール（連続型）として扱うことができます。値間の距離の比較などに適切です。スケール変数の例としては、年齢や、千ドル単位で表した所得があります。

  この手続きは適切な尺度がすべての従属変数に割り当てられると仮定します。ただし、ソース変数リスト内の変数を右クリックしコンテキスト メニューから尺度を選択して、変数の尺度を一時的に変更することができます。

変数リストで各変数の隣にあるアイコンは、次のような尺度とデータ型を表します。

	数値	String	Date	Time
スケール（連続）		利用不可		
順序				
名義				

予測変数。 予測変数は、因子（カテゴリ変数）または共分散（スケール変数）として指定することができます。

カテゴリ変数のコード化。 この手順では、手順の期間に対する one-of-c コード化を使用してカテゴリ予測変数および従属変数を一時的に記録します。変数の c カテゴリが存在する場合、変数は最初のカテゴリ (1, 0, . . . , 0)、次のカテゴリ (0, 1, 0, . . . , 0)、. . . そして最後のカテゴリ (0, 0, . . . , 0, 1) が表示され、c ベクトルとして格納されます。

このコード化方式ではシナプスの重みが増加し、より遅い学習となる場合があります。ただし、さらに「コンパクトな」コード化方式はニューラル ネットワークにあまり適合しなくなります。ネットワーク学習の速度が遅い場合、類似したカテゴリを結合するか極端にまれなカテゴリをもつケースを削除して、カテゴリ予測変数のカテゴリ数を削減します。

検定サンプルまたはホールドアウト サンプルが定義されている場合でも、one-of-c コード化はすべて学習データに基づいています（ p. 9 分割 を参照）。そのため、検定サンプルまたはホールドアウト サンプルに学習データに表示されない予測カテゴリを持つケースが含まれる場合、それらのケースは手順または得点付けでは使用されません。検定サンプルまたはホールドアウト サンプルに学習データに表示されない従属変数カテゴリを持つケースが含まれる場合、それらのケースは手続きでは使用されませんが、得点付けされる場合があります。

再調整。 スケール従属変数および共分散は、ネットワーク学習の向上のため、デフォルトで再調整されます。検定サンプルまたはホールドアウト サンプルが定義されている場合でも、再調整はすべて学習データに基づいて行われます（ p. 9 分割 を参照）。つまり再調整の種類に応じて、共分散または従属変数の平均値、標準偏差、最小値または最大値が学習データのみを使用して計算されます。変数を指定して分割を定義する場合、これらの共分散または従属変数に学習サンプル、検定サンプル、ホールドアウト サンプル全体の類似した分布が含まれていることが重要です。

度数による重み付け。 度数による重み付けは、この手続きによって無視されます。

結果の再現 結果を正確に再現する場合、同じ手続きの設定を行うだけでなく、乱数ジェネレータに同じ初期化の値、同じデータの順序、同じ変数の順序を使用します。この問題の詳細は、次の項目を参照してください。

- **乱数ジェネレータ。** この手続きでは分割の無作為割り当て時に乱数ジェネレータを、シナプスの重みの初期化およびアーキテクチャの自動選択ににランダムなサブサンプルを、そして重みの初期化とアーキテクチャの選択にシミュレートされたアニーリング アルゴリズムを使用します。今後同じランダム化された結果を再生成するには、［多層パーセプトロン］手続きをそれぞれ実行する前に乱数ジェネレータに同じ初期化の値を使用します。段階的な説明は、「分析用データの準備」（ p. 41 ）を参照してください。
- **ケースの並び順。** オンライン学習方法およびミニバッチ学習方法（学習（ p. 14 ）を参照）は、ケースの並び順に明示的に依存しますが、シナプスの重みの初期化ではデータセットからのサブサンプルが行われるため、バッチ学習方法もケースの並び順に依存します。

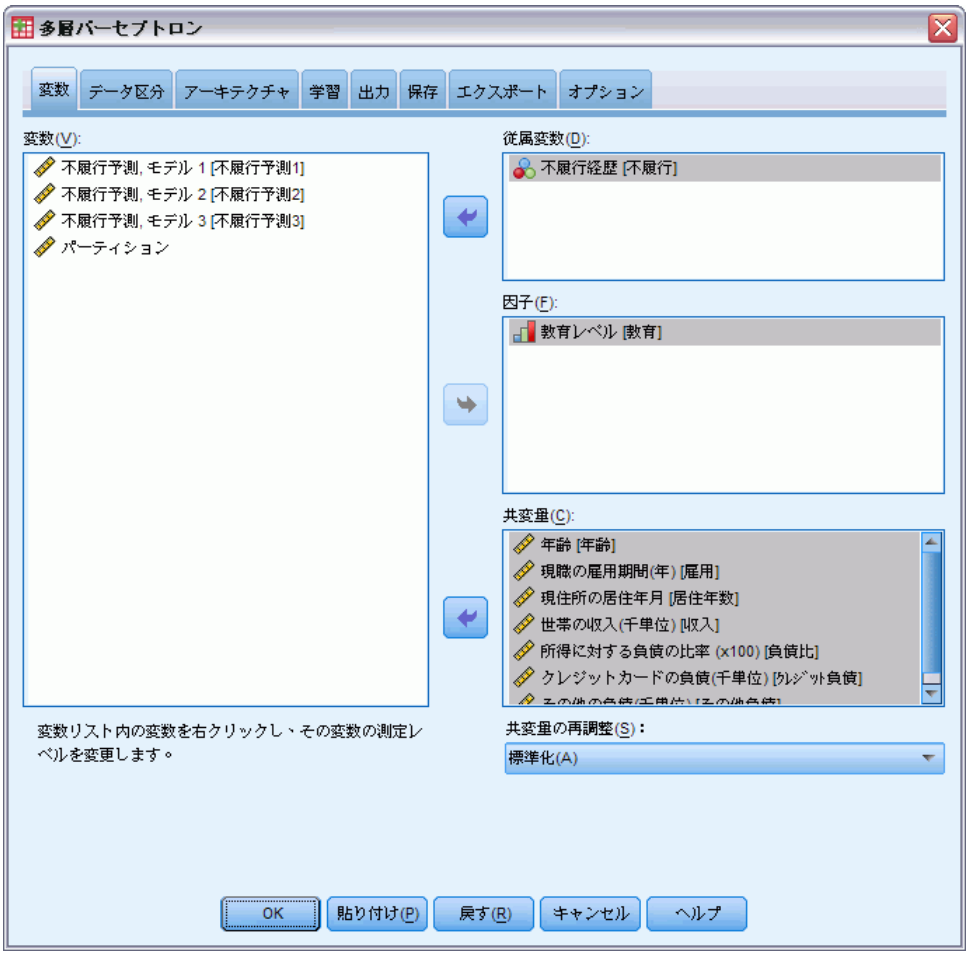
  並び順の影響を最小限に抑えるには、ケースを無作為に並べます。特定の解の安定性を確認するには、異なる無作為な順序で並べ替えられたケースを使用していくつかの異なる解を得てください。ファイル サイズが非常に大きい場合は、異なる無作為な順序で並べ替えられたケースのサンプルを使用し、複数回に分けて実行することができます。
- **変数の順序。**変数の順序が変更される場合、結果は割り当てられる初期値の異なるパターンによって因子および共分散リストの変数の順序に影響される場合があります。ケースの並び順の影響と同じように、異なる変数の順序（因子および共分散リスト内で簡単にドラッグ アンド ドロップする）を使用して、特定の解の安定性を評価します。

多層パーセプトロン ネットワークの作成

メニューから次の項目を選択します。

分析(A) > ニューラル ネットワーク > 多層パーセプトロン...

図 2-1
多層パーセプトロン: [変数] タブ



- 最低 1 つの従属変数を選択します。
- 少なくとも 1 つの因子または共変量を選択します。

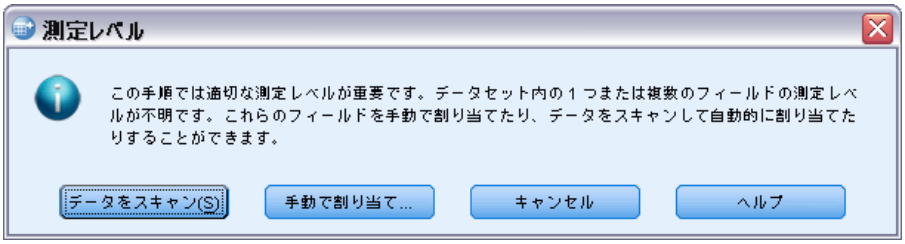
オプションで、[変数] タブで共分散を再調整する方法を変更します。次の項目から選択します。

- **標準化。** (x−mean)/s のように平均値を減算し標準偏差で分割します。
- **正規化。** (x−min)/(max−min) のように、最小値を減算し範囲で分割します。正規化された値は 0 ～ 1 です。
- **調整済み正規化。** [2*(x−min)/(max−min)]−1 のように、最小値を減算し範囲で分割した値を調整したものです。調整済み正規化の値は −1 ～ 1 です。
- **なし。** 共分散の再調整はありません。

測定レベルが不明なフィールドです。

データセットの 1 つまたは複数の変数（フィールド）の尺度が不明な場合、尺度の警告が表示されます。尺度はこの手順の結果の計算に影響を与えるため、すべての変数に尺度を定義する必要があります。

図 2-2
尺度の警告

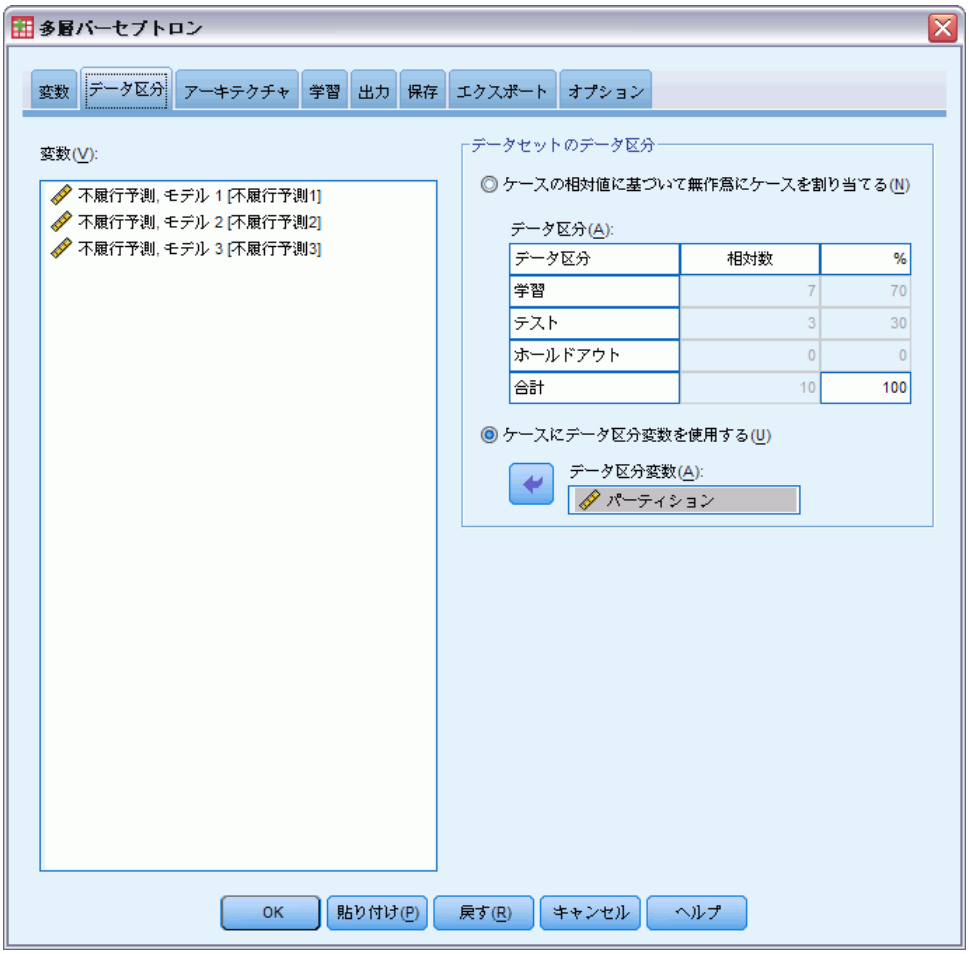


- **データをスキャン。** アクティブ データセットのデータを読み込み、デフォルトの尺度を尺度が現在不明なフィールドに割り当てます。データセットが大きい場合は時間がかかります。
- **手動で割り当てる。** 不明な尺度のフィールドをすべて表示するダイアログが開きます。このダイアログを使用して、尺度をこれらのフィールドに割り当てることができます。データ エディタの［変数ビュー］でも、尺度を割り当てることができます。

尺度がこの手順で重要であるため、すべてのフィールドに尺度が定義されるまで、ダイアログにアクセスしてこの手順を実行することはできません。

分割

図 2-3
多層パーセプトロン: [データ区分] タブ



データセットのデータ区分。 このグループは、アクティブなデータセットを分割する方法をサンプルの学習、検定およびホールドアウトに指定します。**学習サンプル**では、ニューラル ネットワークを学習するために使用するデータ レコードを判断します。データセット内のケースのいくらかの割合は、モデルを取得するために学習サンプルに割り当てる必要があります。**検定サンプル**は、過学習を防ぐため学習中の予測誤差の追跡に使用されるデータ レコードの独立したセットです。学習サンプルを作成することを強く推奨します。検定サンプルが学習サンプルより小さい場合、一般的にネットワーク学習が最も効果的です。**ホールドアウト サンプル**は、最終のニューラル ネットワークを評価するために使用するデータ レコードのもう 1 つの独立セットです。ホールドアウト ケースを使用してモデ

ルを構築できなかったため、ホールドアウト サンプルの誤差によってそのモデルの予測能力を「公正に」評価します。

- **ケースの相対的な数に基づいたケースの無作為割り当て。** 各サンプル（学習、検定、ホールドアウト）にランダムに割り当てられたケースの相対数（比率）を指定します。% 列は、指定した相対数に基づいて各サンプルに割り当てられるケースの割合を示します。

  たとえば、学習サンプル、検定サンプル、ホールドアウト サンプルの相対数として 7、3、0 を指定すると、それぞれ 70%、30%、0% となります。相対数として 2、1、1 を指定すると、それぞれ 50%、25%、25% となります。1、1、1 と指定するとデータセットは学習サンプル、検定サンプル、ホールドアウト サンプルにそれぞれ 3 分の 1 ずつ分けられます。

- **ケースにデータ区分変数を使用する。** アクティブなデータセットの各ケースを学習サンプル、検定サンプル、ホールドアウト サンプルに割り当てる数値型変数を指定します。変数に正の値を持つケースは学習サンプルに、0 の値を持つケースは検定サンプルに、負の値を持つケースはホールドアウト サンプルに割り当てられます。システム欠損値を持つケースは、分析から除外されます。分割変数のユーザー欠損値は、常に有効なものとして扱われます。

  注 :分割変数を使用すると、手続きを連続して実行しても同一の結果は保証されません。多層パーセプトロンのトピック内の「結果の再現」を参照してください。

アーキテクチャ

図 2-4
多層パーセプトロン: [アーキテクチャ] タブ

[アーキテクチャ] タブを使用して、ネットワークの構造を指定します。この手続きでは、自動的に「最良の」アーキテクチャを選択するか、ユーザー指定のアーキテクチャを指定できます。

自動的にアーキテクチャを選択すると、1 つの隠れ層を持つネットワークを構築します。自動的にアーキテクチャを選択して隠れ層に作成できる単位の最小数および最大数を指定すると、隠れ層に「最も適切な」単位数を計算します。自動的にアーキテクチャを選択する場合、隠れ層および出力層に対してデフォルトの活性化関数を使用します。

ユーザー指定でアーキテクチャを選択すると、隠れ層および出力層にエキスパート制御を行うことができ、必要なアーキテクチャが事前にわかっている場合、または自動的にアーキテクチャを選択して結果を調整する場合に役に立ちます。

隠れ層

隠れ層には、観測不可能なノード（単位）が含まれています。それぞれの隠れた単位は、入力層の重みの付いた合計の関数です。この関数は活性化関数で、重みの値は推定アルゴリズムによって決まります。ネットワークに 2 番目の隠れ層が含まれている場合、2 番目の層にあるそれぞれの隠れ層は最初の隠れ層の単位の重みの付いた合計の関数です。同じ活性化関数が、両方の層で使用されます。

隠れ層の数. 多層パーセプトロンは 1 個か 2 個の隠れ層を所持できます。

活性化関数. 活性化関数は、層の単位を後続の層の単位の値に「リンク」させます。

- **双曲線正接（ハイパボリック アークタンジェント）。** この関数の形式は、$\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$ です。これは実数の引数を取り、範囲 (-1, 1) に変換します。自動的にアーキテクチャを選択した場合、これは隠れ層のすべての単位に対する活性化関数となります。
- **シグモイド。** この関数の形式は、$\gamma(c) = 1/(1+e^{-c})$ です。これは実数の引数を取り、範囲 (0, 1) に変換します。

ユニットの数. 各隠れ層の単位数は明示的に指定されるか、予測アルゴリズムにより自動的に決定されます。

出力層

出力層には、目標（従属）変数が含まれます。

活性化関数. 活性化関数は、層の単位を後続の層の単位の値に「リンク」させます。

- **同一。** この関数の形式は、$\gamma(c) = c$ です。これは実数の引数を取り、変換せずに返します。自動的にアーキテクチャを選択した場合、スケール従属変数のある出力層の単位に対する活性化関数となります。
- **Softmax。** この関数の形式は、$\gamma(c_k) = \exp(c_k)/\Sigma_j \exp(c_j)$ です。これは実数の引数のベクトルを取り、要素が範囲 (0, 1) で合計が 1 となるベクトルに変換します。Softmax は、従属変数がカテゴリ型である場合にのみ使用できます。自動的にアーキテクチャを選択した場合、従属変数がカテゴリ変数である出力層の単位に対する活性化関数となります。
- **双曲線正接（ハイパボリック アークタンジェント）。** この関数の形式は、$\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$ です。これは実数の引数を取り、範囲 (-1, 1) に変換します。
- **シグモイド。** この関数の形式は、$\gamma(c) = 1/(1+e^{-c})$ です。これは実数の引数を取り、範囲 (0, 1) に変換します。

スケール従属変数の再調整。 これらの制御は、最低 1 つのスケール従属変数が選択されている場合にのみ適用されます。

- **標準化。** (x-mean)/s のように平均値を減算し標準偏差で分割します。
- **正規化。** (x-min)/(max-min) のように、最小値を減算し範囲で分割します。正規化された値は 0 ～ 1 です。出力層がシグモイドの活性化関数を使用している場合、スケール従属変数に対して必要な再調整方法です。修正オプションでは、再調整式に対する修正として適用される小さな数 ε として指定します。この修正により、再調整されたすべての従属変数は、活性化関数の範囲内の値となります。特に x が最小値および最大値を取る場合に未修正の式で発生する 0 と 1 の値は、S 字関数の範囲の限度を定義しますが、その範囲内ではありません。修正された式は [x-(min-ε)]/[(max+ ε) -(min-ε)] となります。0 以上の値を指定します。
- **調整済み正規化。** [2*(x-min)/(max-min)]-1 のように、最小値を減算し範囲で分割した値を調整したものです。調整済み正規化の値は -1 ～ 1 です。出力層が双曲線正接の活性化関数を使用している場合、スケール従属変数に対して必要な再調整方法です。修正オプションでは、再調整式に対する修正として適用される小さな数 ε として指定します。この修正により、再調整されたすべての従属変数は、活性化関数の範囲内の値となります。特に x が最小値および最大値を取る場合に未修正の式で発生する -1 と 1 の値は、双曲線正接関数の範囲の限度を定義しますが、その範囲内ではありません。修正された式は {2*[(x-(min-ε))/((max+ε)-(min-ε))]}-1 となります。0 以上の値を指定します。
- **なし。** スケール従属変数の再調整はありません。

学習

図 2-5
多層パーセプトロン: [学習] タブ

[学習] タブを使用して、ネットワークの学習方法を指定します。 b 学習の種類とアルゴリズムの最適化によって、使用できる学習オプションを決定します。

学習の種類。 学習の種類によって、ネットワークが記録をどのように処理するかが決まります。次の学習の種類からいずれかを選びます。

- **バッチ学習。** すべての学習データ記録を渡した後でシナプスの重みを更新します。つまり、バッチ学習では学習データセットのすべての記録の情報を使用します。全体の誤差を直接最小化するため、バッチ学習が好まれます。バッチ学習は、停止規則の 1 つが一致するまで重みを何度も更新する場合があり、そのため多くのデータを渡す必要があります。バッチ学習は「小さな」データセットで最も有用です。

- **オンライン学習。** 単一の学習データ記録をそれぞれ記録した後で、シナプスの重みを更新します。つまり、オンライン学習では 1 度に 1 つの記録を使用します。オンライン学習では継続的に記録を取得し、停止規則の 1 つが一致するまで重みを更新します。すべての記録がいったん使用され、一致する停止規則がない場合、データ記録を再利用して処理を継続します。オンライン学習は、関連する予測変数が含まれる「大きな」データセットに対してバッチ学習よりも優れています。つまり、多くの記録、入力があり、それらの値がお互いに独立していない場合、オンライン学習ではバッチ学習よりも迅速に適切な回答を取得できます。
- **ミニバッチ学習。** 学習データ記録をほぼ等しいサイズのグループに分割し、1 つのグループを渡した後でシナプスの重みを更新します。つまり、ミニバッチ学習では記録のグループの情報を使用します。この処理では、必要に応じてデータ グループを再利用します。ミニバッチ学習はバッチ学習とオンライン学習の中間に位置し、「中規模サイズの」データセットの場合に最適です。この手続きでは、ミニバッチ学習の学習記録数を自動的に決定したり、1 より大きな整数またはメモリに格納するケースの最大数以下の整数を指定することができます。［オプション］タブで、メモリー内に格納するケースの最大数を設定できます。

アルゴリズムの最適化。 この方法によって、シナプスの重みを推定します。

- **調整された共役勾配。** 共役勾配法の使用を指定する仮定は、バッチ学習にのみ適用されます。オンライン学習またはミニバッチ学習には適用されません。
- **勾配下降。** この方法は、オンライン学習またはミニバッチ学習で使用する必要があります。また、バッチ学習で使用することもできます。

学習オプション。 学習オプションを使用すると、アルゴリズムの最適化を調整できます。通常、ネットワークを推定の問題に実行しない限り、これらの設定を変更する必要はありません。

スケール化共役勾配アルゴリズムの学習オプションには次のものがあります。

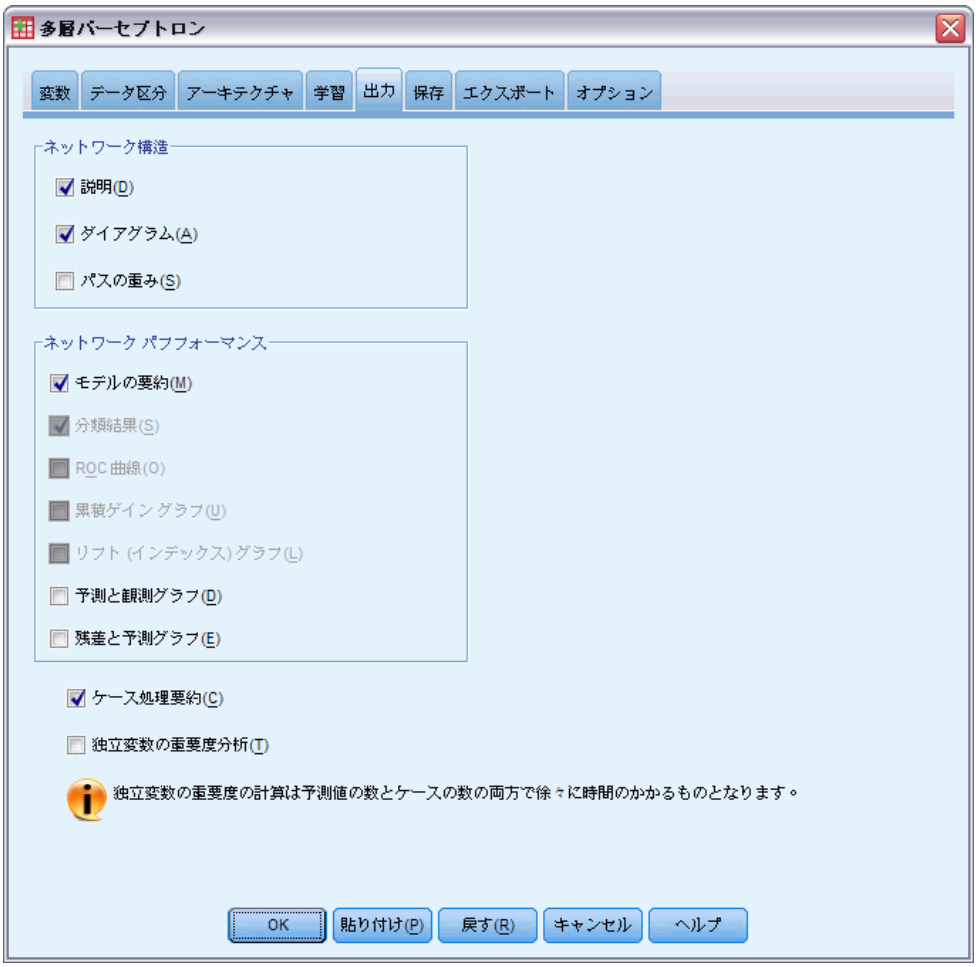
- **初期のラムダ。** スケール化共役勾配アルゴリズムのラムダ パラメータの初期値です。0 より大きく 0.000001 より小さい数を指定します。
- **初期のシグマ。** スケール化共役勾配アルゴリズムのシグマ パラメータの初期値です。0 より大きく 0.0001 より小さい数を指定します。
- **区間の中心と区間のオフセット。** 区間の中心（a_0）と区間のオフセット（a）は、シミュレートされたアニーリング アルゴリズムが使用された場合、重みのベクトルがランダムに生成される区間［a_0–a，a_0+a］を定義します。アルゴリズムの最適化の適用時にグローバルな最小値を検出する目的で、シミュレートされたアニーリング アルゴリズムを使用してローカルの最小値に割り込みます。このアプローチは、重みを初期化し自動的にアーキテクチャを選択する場合に使用します。区間の中心には数を指定し、区間のオフセットには 0 より大きな数を指定します。

勾配降下アルゴリズムの学習オプションには次のものがあります。

- **初期の学習率。** 勾配効果法アルゴリズムの学習率の初期値です。学習率が高い場合ネットワークの学習が早く、コストは安定しない場合があります。0 より大きい数字を指定します。
- **学習率の下限。** 勾配効果法アルゴリズムの学習率の下限です。この設定は、オンライン学習およびミニバッチ学習にのみ適用されます。0 より大きく、初期の学習率より小さな数を指定します。
- **推進力。** 勾配降下アルゴリズムの初期推進力パラメータです。推進力によって、高すぎる学習率による不安定性を回避できます。0 より大きい数字を指定します。
- **エポックの学習率の減衰。** オンライン学習またはミニバッチ学習で勾配下降が使用される場合、初期の学習率を学習率の下限まで減少させるために必要な Epoch（学習サンプルのデータ パス）数（ p ）です。これにより学習率の崩壊因子 $\beta = (1/pK)*\ln(\eta_0/\eta_{low})$ を制御できます。この場合、η_0 は初期の学習率を表し、η_{low} は学習率の下限、K は学習データセット内のミニバッチの総数（またはオンライン学習の場合学習記録数）です。0 より大きい整数を指定します。

出力

図 2-6
多層パーセプトロン: [出力] タブ



ネットワーク構造。 ニューラル ネットワークの概要を表示します。

- **説明:** 従属変数、入力単位および出力単位の数、隠れ層および隠れた単位の数、活性化関数の情報を表示します。
- **ダイアグラム。** ネットワークの図を編集不可能な図表として表示します。共変量の数と因子レベルが増加すると、図の理解がより困難になります。
- **シナプスの重み。** 次の層の単位にと特定の層の単位の関係を示す係数の推定値を表示します。アクティブなデータセットが学習データ、検定データ、ホールドアウト データに分割されている場合でも、シナプスの重みは学習サンプルに基づいています。シナプスの重みは大き

くなり、これらの重みは通常ネットワークの結果を理解するためには使用されません。

ネットワーク パフォーマンス。 モデルが「良い」かどうかを確認するために使用する結果を表示します。注 :このグループのグラフは、学習サンプルと検定サンプルの組み合わせに、また検定サンプルがない場合は学習サンプルのみに基づいています。

- **モデルの要約。** 誤差、相対誤差、または不正な予測変数の割合、学習を停止するために使用する停止規則および学習時間など、ニューラルネットワークの部分的または全体的な結果の要約を表示します。

  同一、シグモイド、双曲線正接の活性化関数が出力層に適用される場合、この誤差は平方和の誤差です。Softmax 活性化関数が出力層に適用される場合、クロスエントロピー誤差となります。

  相対誤差または不正な予測変数の割合は、従属変数尺度に応じて表示されます。従属変数にスケール尺度が含まれる場合、平均の全体的な相対誤差（平均値モデルに関連）が表示されます。すべての従属変数がカテゴリ変数の場合、不正な予測変数の平均割合が表示されます。相対誤差または不正な予測変数の割合は、それぞれの従属変数に対しても表示されます。

- **分類結果（距離と近接度）。** 各カテゴリ従属変数の分類テーブルを部分的または全体的に表示します。各テーブルは、各従属変数カテゴリに正しくまたは誤って分類されたケースの数を示します。正しく分類されたケース全体の割合も報告されます。

- **ROC 曲線。** 各カテゴリ従属変数の ROC（受信者動作特性）曲線を表示します。また、各曲線の下の領域を示すテーブルも表示します。特定の従属変数に対し、ROC 曲線のグラフは 各カテゴリに 1 つの曲線を表示します。従属変数に 2 つのカテゴリがある場合、それぞれの曲線は問題のカテゴリを もう 1 つのカテゴリに対し正の状態として扱います。従属変数に 2 つを超えるカテゴリがある場合、それぞれの曲線は問題のカテゴリを その他すべてのカテゴリの集計変数に対し正の状態として扱います。

- **累積ゲイン グラフ。** 各カテゴリ従属変数の累積ゲイン グラフを表示します。ROC 曲線と同様、各従属変数カテゴリに対し 1 つの曲線を表示します。

- **リフト グラフ。** 各カテゴリ従属変数のリフト グラフを表示します。ROC 曲線と同様、各従属変数カテゴリに対し 1 つの曲線を表示します。

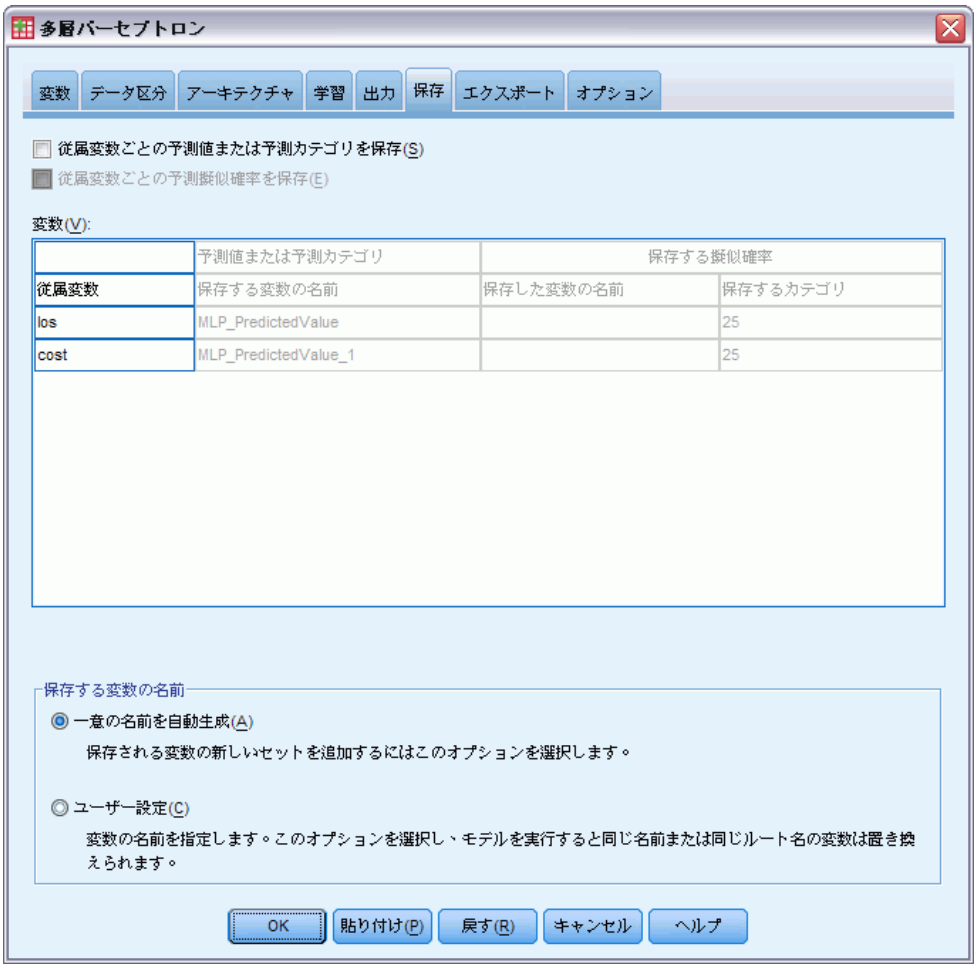
- **予測と観測のグラフ。**各従属変数の観測値により予測されるグラフを表示します。カテゴリ従属変数の場合、予測擬似確率のクラスタ箱ひげ図が各応答カテゴリに対し、各応答カテゴリをクラスタ変数として表示されます。スケール従属変数の場合、散布図が表示されます。
- **残差と予測のグラフ。**各スケール従属変数の予測値による残差グラフを表示します。残差と予測値の間にパターンが表示されることはありません。このグラフはスケール従属変数に対してのみ生成されます。

ケース処理の要約 分析に含まれたケースおよび除外されたケースの数を全体、学習サンプル、検定サンプルおよびホールドアウト サンプルごとに要約するケース処理の要約テーブルを表示します。

独立変数の重要度分析。感度分析を実行し、ニューラル ネットワークを定義する場合各予測変数の重要度を計算します。分析は、学習サンプルと検定サンプルの組み合わせに、また検定サンプルがない場合は学習サンプルのみに基づいています。この分析によって、各予測変数の重要度および正規化された重要度を表すテーブルおよびグラフを作成します。予測変数またはケースの数が多い場合、感度分析の計算は効率的でなく時間もかかります。

保存

図 2-7
多層パーセプトロン: [保存] タブ



[保存] タブを使用して、データセットの変数として予測変数を保存します。

- **従属変数ごとの予測値または予測カテゴリを保存。** これにより、スケール従属変数に予測された値を保存し、カテゴリ従属変数に予測カテゴリを保存します。
- **従属変数ごとの予測擬似確率を保存。** カテゴリ従属変数に予測擬似確率を保存します。各変数は、それぞれの最初の n カテゴリに対して保存されます。この場合、 n は **[保存するカテゴリ]** 列で指定されます。

保存する変数の名前 自動的な名前の生成によって、すべての作業を保存することができます。ユーザー指定の名前によって、Data Editor で保存された変数を最初に削除することなく、前回実行された結果を破棄または置き換えることができます。

確率および擬似確率

Softmax 活性化関数およびクロスエントロピー誤差を含むカテゴリ従属変数には、各カテゴリ変数の予測値が含まれます。この場合、各予測値はケースがカテゴリに属する確率です。

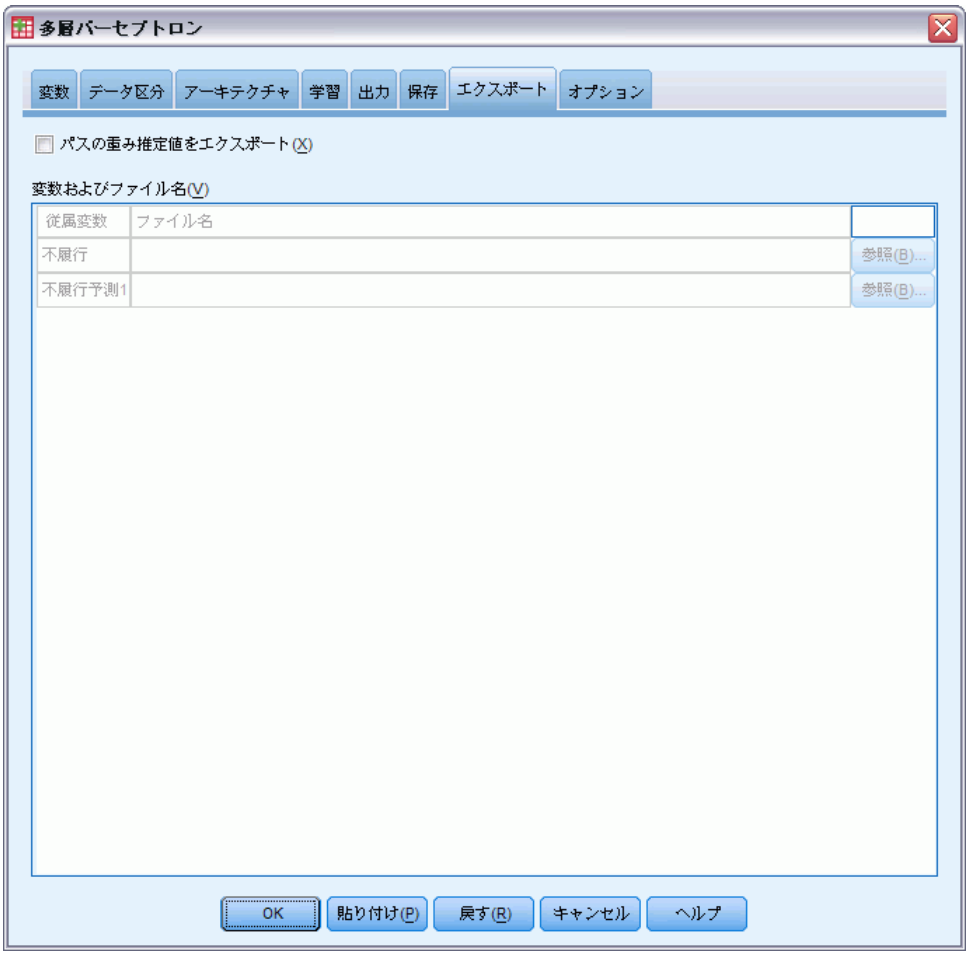
平方和の誤差を含むカテゴリ従属変数には各カテゴリ変数の予測値が含まれますが、予測値は確率として解釈されることはありません。値が 0 より小さいまたは 1 より大きい場合であっても、または特定の従属変数の合計が 1 でない場合でも、この手続きではこれらの予測擬似確率を保存します。

ROC 曲線、累積ゲイン グラフおよびリフト グラフ（ p.17 出力 を参照）は、擬似確率に基づいて作成されます。擬似確率が 0 より小さいまたは 1 より大きい、または特定の変数の合計が 1 でない場合、擬似確率は 0 ～ 1 に、変数の合計は 1 に再調整されます。擬似確率は合計で分割することによって再調整されます。たとえば、ケースに 3 つのカテゴリ従属変数に対し 0.50、0.60 および 0.40 の擬似確率が存在する場合、各擬似確率は合計の 1.50 で分割され、0.33、0.40、0.27 を取得します。

擬似確率のいずれかが負である場合、前述の再調整を行う前に最も低い確率の絶対値をすべての擬似確率に追加します。たとえば、擬似確率が -0.30、0.50、および 1.30 である場合、まず各値に 0.30 を追加して 0.00、0.80、および 1.60 を取得します。次に、それぞれの新しい値を、2.40 で分割して 0.00、0.33、および 0.67 を取得します。

エクスポート

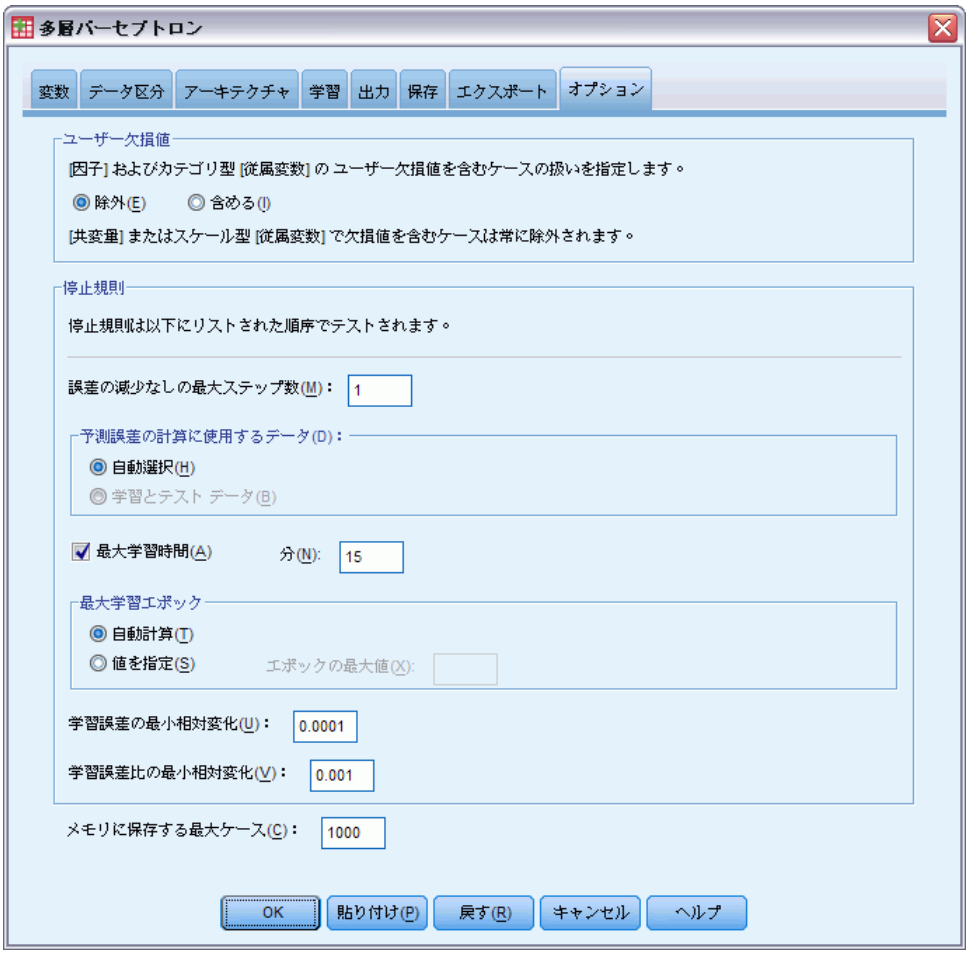
図 2-8
多層パーセプトロン: [エクスポート] タブ



[エクスポート] タブを使用して、各従属変数のシナプスの重みの推定値を XML (PMML) ファイルに保存します。このモデル ファイルを使用して、得点付けのために他のデータ ファイルにモデル情報を適用できます。分割ファイルが定義されている場合、このオプションは使用できません。

オプション

図 2-9
多層パーセプトロン: [オプション] タブ



ユーザー欠損値。 因子は、分析の対象となるケースに対して有効な値を取る必要があります。このオプションを使用すると、ユーザー欠損値を因子変数およびカテゴリ従属変数で有効な値として扱うかどうかを決定できます。

停止規則。 これらは、ニューラル ネットワークの学習をいつ停止するのかを決定する規則です。学習は、少なくとも 1 つのデータ パスで継続します。学習は、表示された順序で確認され、次の基準に応じて停止します。停止規則の定義において、ステップはオンライン学習およびミニバッチ学習方法ではデータ パスに、バッチ学習方法では反復に対応します。

- **誤差の減少なしの最大ステップ数。** 誤差の減少を確認する前に可能なステップ数です。指定されたステップ数を経て誤差の減少がない場合、学習が停止します。0 より大きい整数を指定します。誤差の計算にどのデータ

サンプルを使用するか指定することもできます。[**自動的に選択**] を選択すると、検定サンプルが存在する場合はそれを使用し、ない場合は学習サンプルを使用します。バッチ学習は各データ パスの後の学習サンプルの誤差の減少を保証するため、検定サンプルが存在する場合、このオプションはバッチ学習に対してのみ適用されます。[**学習データと検定データの両方**] を選択すると、これらのサンプルそれぞれの誤差を確認します。このオプションは検定サンプルが存在する場合にのみ適用されます。

注 :それぞれの完全なデータ パスの後、オンライン学習とミニバッチ学習には、学習誤差を計算するために追加のデータ パスが必要です。この追加データ パスでは、学習の速度がかなり遅くなる場合があります。その場合、通常検定サンプルを使用し [**自動的に選択**] を選択します。

- **最大学習時間。** 実行するアルゴリズムの最大時間（分）を指定するかどうかを選択します。0 より大きい数字を指定します。
- **最大学習エポック。** 可能な最大 Epoch（データ パス）数です。Epoch の最大数を超えた場合、学習が停止します。0 より大きい整数を指定します。
- **学習誤差の最小相対変化。** 以前のステップと比較した学習誤差の相対変化が基準の値を下回った場合、学習は停止します。0 より大きい数字を指定します。オンライン学習およびミニバッチ学習の場合、検定データのみを使用して誤差を計算するとこの基準は無視されます。
- **学習誤差比の最小相対変化。** ヌル モデルの誤差に対する学習誤差の割合が基準の値を下回った場合、学習は停止します。ヌル モデルはすべての従属変数に対する平均値を予測します。0 より大きい数字を指定します。オンライン学習およびミニバッチ学習の場合、検定データのみを使用して誤差を計算するとこの基準は無視されます。

メモリーに保存する最大ケース。 これにより、多層パーセプトロン アルゴリズムの次の設定を制御します。1 より大きい整数を指定します。

- 自動的にアーキテクチャを選択する場合、ネットワーク アーキテクチャの決定に使用されるサンプルのサイズは min(1000, memsize) です。ここで、memsize はメモリに格納するケースの最大数です。
- ミニバッチの数を自動的に計算するミニバッチ学習では、ミニバッチの数は min(max(M/10,2), memsize) です。ここで、M は学習サンプルのケース数です。

章 3

放射基底関数

放射基底関数（RBF）手続きは、予測変数の値に基づいて、1 つ以上の従属（目標）変数に対する予測モデルを生成します。

例: あるデータ通信プロバイダは、サービス利用パターンに基づいて顧客ベースを分類し、4 つのグループにカテゴリー化しました。人口統計データを使用して所属グループを予測する RBF ネットワークによって、会社は見込み客それぞれに対する提案をカスタマイズできます。

従属変数。 従属変数は次のものを使用できます。

- **名義データ.** 値がランキングなどを持たないカテゴリを表しているとき、名義（変数）として取り扱うことができます。たとえば、従業員の会社の所属などです。名義変数の例としては、地域やジップ コードや所属宗教などがあります。
- **順序データ.** 値がランキングをもったカテゴリを表しているとき、変数を順序として取り扱うことができます。たとえば、「かなり不満」から「かなり満足」までのようなサービス満足度のレベルなどです。順序変数の例としては、満足度や信頼度を表す得点や嗜好得点などです。
- **スケール データ.** 値が有意な基準を持った順序カテゴリを表しているとき、変数をスケール（連続型）として扱うことができます。値間の距離の比較などに適切です。スケール変数の例としては、年齢や、千ドル単位で表した所得があります。

  ソース変数リスト内の変数を右クリックしコンテキスト メニューから尺度を選択して変数の尺度を一時的に変更できますが、この手続きは適切な尺度がすべての従属変数に割り当てられると仮定します。

変数リストで各変数の隣にあるアイコンは、次のような尺度とデータ型を表します。

	数値	String	Date	Time
スケール（連続）		利用不可		
順序				
名義				

予測変数。 予測変数は、因子（カテゴリ変数）または共分散（スケール変数）として指定することができます。

カテゴリ変数のコード化。 この手順では、手順の期間に対する one-of-c コード化を使用してカテゴリ予測変数および従属変数を一時的に記録します。変数の c カテゴリが存在する場合、変数は最初のカテゴリ（1, 0, . . . , 0）、次のカテゴリ（0, 1, 0, . . . , 0）、. . . そして最後のカテゴリ（0, 0, . . . , 0, 1）が表示され、c ベクトルとして格納されます。

このコード化方式ではシナプスの重みが増加し、より遅い学習となる場合があります。さらに「コンパクトな」コード化方式によりニューラルネットワークにあまり適合しなくなります。ネットワーク学習の速度が遅い場合、類似したカテゴリを結合するか極端にまれなカテゴリをもつケースを削除して、カテゴリ予測変数のカテゴリ数を削減します。

検定サンプルまたはホールドアウト サンプルが定義されている場合でも、one-of-c コード化はすべて学習データに基づいています（ p. 29 分割を参照）。そのため、検定サンプルまたはホールドアウト サンプルに学習データに表示されない予測カテゴリを持つケースが含まれる場合、それらのケースは手順または得点付けでは使用されません。検定サンプルまたはホールドアウト サンプルに学習データに表示されない従属変数カテゴリを持つケースが含まれる場合、それらのケースは手順では使用されませんが、得点付けされる場合があります。

再調整。 スケール従属変数および共分散は、ネットワーク学習の向上のため、デフォルトで再調整されます。検定サンプルまたはホールドアウト サンプルが定義されている場合でも、再調整はすべて学習データに基づいて行われます（ p. 29 分割 を参照）。つまり再調整の種類に応じて、共分散または従属変数の平均値、標準偏差、最小値または最大値が学習データのみを使用して計算されます。変数を指定して分割を定義する場合、これらの共分散または従属変数に学習サンプル、検定サンプル、ホールドアウトサンプル全体の類似した分布が含まれていることが重要です。

度数による重み付け。 度数による重み付けは、この手続きによって無視されます。

結果の再現 結果を正確に再現する場合、同じ手続きの設定を行うだけでなく、乱数ジェネレータに同じ初期化の値を使用します。この問題の詳細は、次の項目を参照してください。

- **乱数ジェネレータ。** この手続きでは、分割の無作為割り当て時に乱数ジェネレータを使用します。今後同じランダム化された結果を再生成するには、[放射基底関数] 手続きをそれぞれ実行する前に乱数ジェネレータに同じ初期化の値を使用します。段階的な説明は、「分析用データの準備」（ p. 78 ）を参照してください。
- **ケースの並び順。** 2 段階ののクラスタ アルゴリズムを使用して放射基底関数を定義するため、結果はデータ順も依存します。

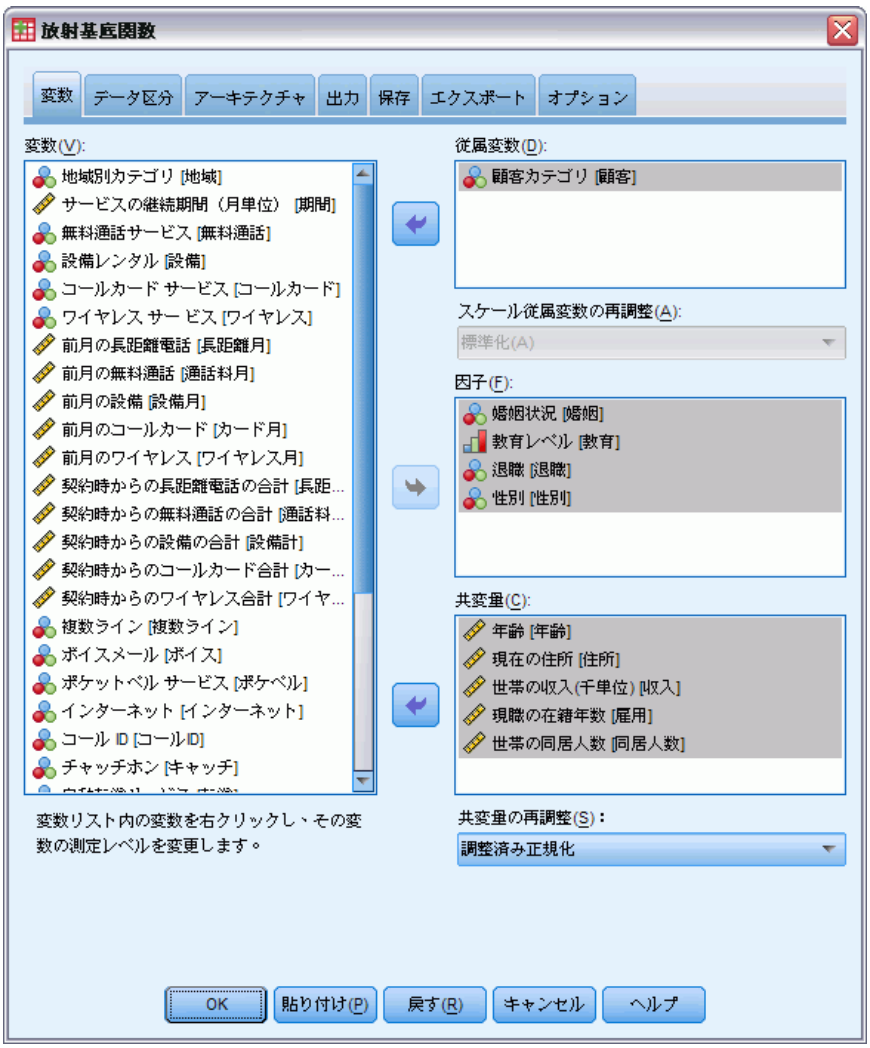
並び順の影響を最小限に抑えるには、ケースを無作為に並べます。特定の解の安定性を確認するには、異なる無作為な順序で並べ替えられたケースを使用していくつかの異なる解を得てください。ファイル サイズが非常に大きい場合は、異なる無作為な順序で並べ替えられたケースのサンプルを使用し、複数回に分けて実行することができます。

放射基底関数ネットワークの作成

メニューから次の項目を選択します。

分析(A) > ニューラル ネットワーク > 放射基底関数...

図 3-1
[放射基底関数: 変数] タブ



- ▶ 最低 1 つの従属変数を選択します。
- ▶ 少なくとも 1 つの因子または共変量を選択します。

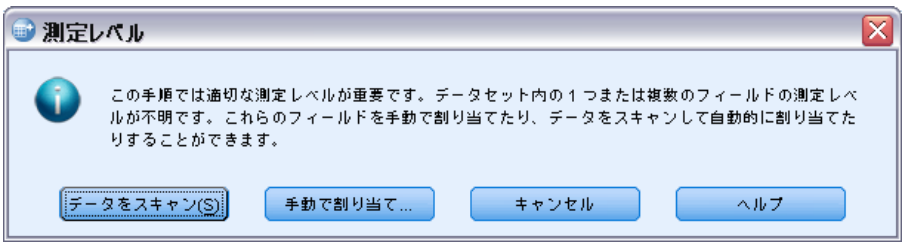
オプションで、[変数] タブで共分散を再調整する方法を変更します。次の項目から選択します。

- **標準化。** (x-mean)/s のように平均値を減算し標準偏差で分割します。
- **正規化。** (x-min)/(max-min) のように、最小値を減算し範囲で分割します。正規化された値は 0 ～ 1 です。
- **調整済み正規化。** [2*(x-min)/(max-min)]-1 のように、最小値を減算し範囲で分割した値を調整したものです。調整済み正規化の値は -1 ～ 1 です。
- **なし。** 共分散の再調整はありません。

測定レベルが不明なフィールドです。

データセットの 1 つまたは複数の変数（フィールド）の尺度が不明な場合、尺度の警告が表示されます。尺度はこの手順の結果の計算に影響を与えるため、すべての変数に尺度を定義する必要があります。

図 3-2
尺度の警告

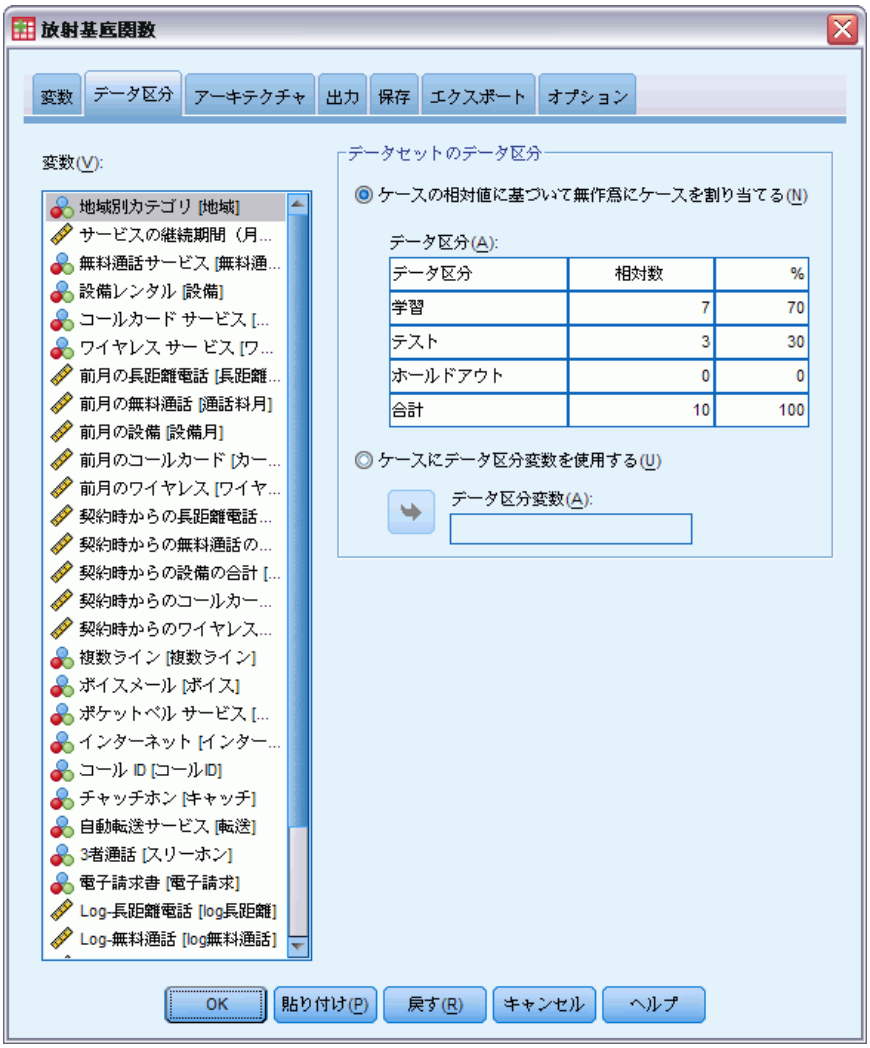


- **データをスキャン。** アクティブ データセットのデータを読み込み、デフォルトの尺度を尺度が現在不明なフィールドに割り当てます。データセットが大きい場合は時間がかかります。
- **手動で割り当てる。** 不明な尺度のフィールドをすべて表示するダイアログが開きます。このダイアログを使用して、尺度をこれらのフィールドに割り当てることができます。データ エディタの [変数ビュー] でも、尺度を割り当てることができます。

尺度がこの手順で重要であるため、すべてのフィールドに尺度が定義されるまで、ダイアログにアクセスしてこの手順を実行することはできません。

分割

図 3-3
[放射基底関数: 分割] タブ



データセットのデータ区分。 このグループは、アクティブなデータセットを分割する方法をサンプルの学習、検定およびホールドアウトに指定します。**学習サンプル**では、ニューラル ネットワークを学習するために使用するデータ レコードを判断します。データセット内のケースのいくらかの割合は、モデルを取得するために学習サンプルに割り当てる必要があります。**検定サンプル**は、過学習を防ぐため学習中の予測誤差の追跡に使用されるデータ レコードの独立したセットです。学習サンプルを作成することを強く推奨します。検定サンプルが学習サンプルより小さい場合、一般的にネットワーク学習が最も効果的です。**ホールドアウト サンプル**は、最終

のニューラル ネットワークを評価するために使用するデータ レコードのもう 1 つの独立セットです。ホールドアウト ケースを使用してモデルを構築できなかったため、ホールドアウト サンプルの誤差によってそのモデルの予測能力を「公正に」評価します。

- **ケースの相対的な数に基づいたケースの無作為割り当て。** 各サンプル（学習、検定、ホールドアウト）にランダムに割り当てられたケースの相対数（比率）を指定します。% 列は、指定した相対数に基づいて各サンプルに割り当てられるケースの割合を示します。

  たとえば、学習サンプル、検定サンプル、ホールドアウト サンプルの相対数として 7、3、0 を指定すると、それぞれ 70%、30%、0% となります。相対数として 2、1、1 を指定すると、それぞれ 50%、25%、25% となります。1、1、1 と指定するとデータセットは学習サンプル、検定サンプル、ホールドアウト サンプルにそれぞれ 3 分の 1 ずつ分けられます。

- **ケースにデータ区分変数を使用する。** アクティブなデータセットの各ケースを学習サンプル、検定サンプル、ホールドアウト サンプルに割り当てる数値型変数を指定します。変数に正の値を持つケースは学習サンプルに、0 の値を持つケースは検定サンプルに、負の値を持つケースはホールドアウト サンプルに割り当てられます。システム欠損値を持つケースは、分析から除外されます。分割変数のユーザー欠損値は、常に有効なものとして扱われます。

アーキテクチャ

図 3-4
[放射基底関数: アーキテクチャ] タブ

[アーキテクチャ] タブを使用して、ネットワークの構造を指定します。この手続きでは、隠れた「放射基底関数」層を持つニューラル ネットワークを作成します。通常、これらの設定を変更する必要はありません。

隠れ層の単位数。 隠れた単位の数を選択するには、3 つの方法があります。

1. **自動的に計算された範囲内の最適な単位数を検出します。** この手続きでは、範囲の最小値および最大値を自動的に計算し、範囲内で最適な隠れた単位の数を検出します。

検定サンプルが定義されている場合、この手続きでは、次の検定データ基準を使用します。最適な数の隠れた単位は、検定データでの誤差が最も小さい数です。検定サンプルが定義されていない場合、この手続きでは、次のベイズ情報量基準（BIC）を使用します。最適な数の隠れた単位は、学習データに基づく BIC が最も小さい数です。

2. **指定された範囲内の最適な単位数を検出します。** 独自の範囲を提供し、この手続きで範囲内の「最適な」隠れた単位の数を検出します。以前と同様、範囲内の最適な隠れた単位の数は、検定データ基準または BIC を使用して決定します。

3. **指定された単位の数を使用します。** 範囲の使用を無効にし、特定の単位の数を直接指定することができます。

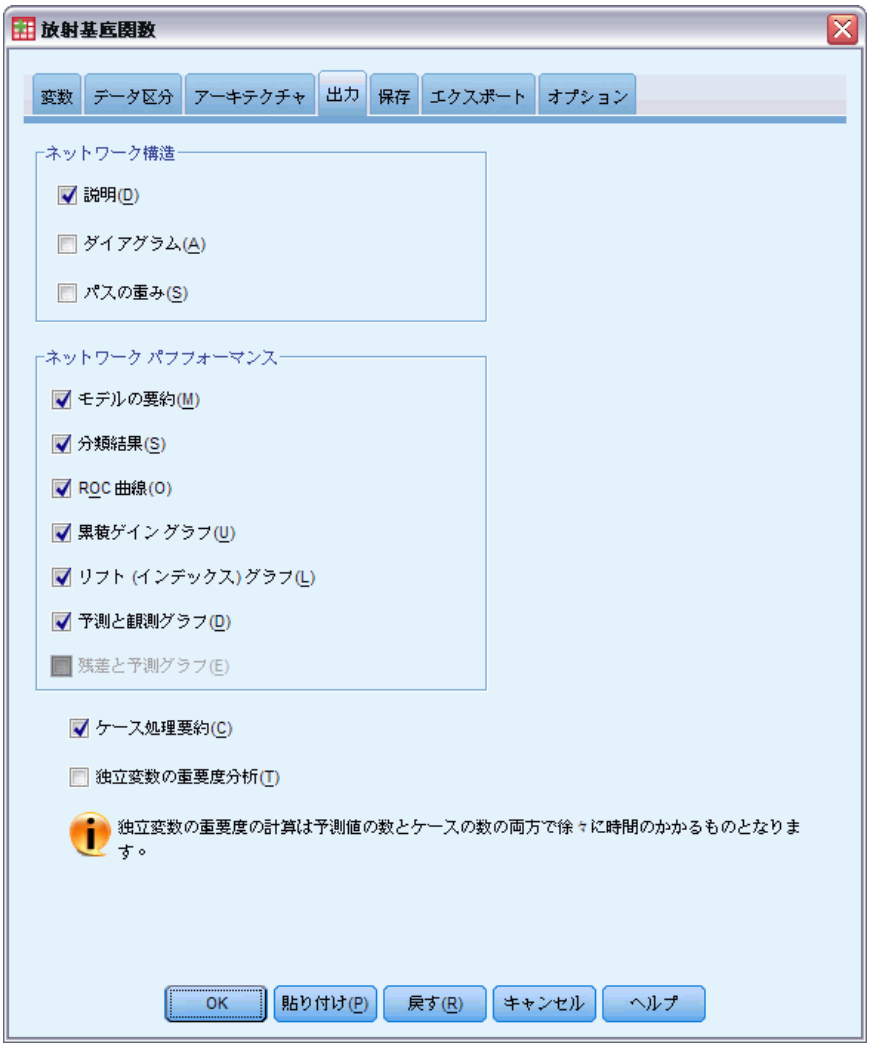
隠れ層の活性化関数。 隠れ層の活性化関数は放射基底関数で、層の単位を後続の層の単位の値に「リンク」させます。出力層の場合、活性化関数は同一関数です。そのため、出力単位は単純に重み付けされた隠れた単位の合計です。

- **正規化放射基底関数。** Softmax 活性化関数を使用すると、隠れた単位の活性化関数は合計が 1 に正規化されます。
- **通常の放射基底関数。** 指数活性化関数を使用すると、隠れた単位の活性化関数は入力単位の関数としてガウス分布の「バンプ」となります。

隠れ層の重複。 重複した因子は、放射基底関数の幅に適用される乗数です。重複した因子の自動的に計算された値は 1+0.1 d です。 d は入力単位の数を表します（因子全体のカテゴリ数の合計および共分散の数）。

出力

図 3-5
[放射基底関数: 出力] タブ



ネットワーク構造。 ニューラル ネットワークの概要を表示します。

- **説明:** 従属変数、入力単位および出力単位の数、隠れ層および隠れた単位の数、活性化関数の情報を表示します。

- **ダイアグラム。** ネットワークの図を編集不可能な図表として表示します。共変量の数と因子レベルが増加すると、図の理解がより困難になります。
- **シナプスの重み。** 次の層の単位にと特定の層の単位の関係を示す係数の推定値を表示します。アクティブなデータセットが学習データ、検定データ、ホールドアウト データに分割されている場合でも、シナプスの重みは学習サンプルに基づいています。シナプスの重みは大きくなり、これらの重みは通常ネットワークの結果を理解するためには使用されません。

ネットワーク パフォーマンス。 モデルが「良い」かどうかを確認するために使用する結果を表示します。注 :このグループのグラフは、学習サンプルと検定サンプルの組み合わせに、また検定サンプルがない場合は学習サンプルのみに基づいています。

- **モデルの要約。** 誤差、相対誤差、または不正な予測変数の割合、および学習時間など、ニューラル ネットワークの部分的または全体的な結果の要約を表示します。

  誤差は平方和の誤差です。また、相対誤差または不正な予測変数の割合は、従属変数尺度に応じて表示されます。従属変数にスケール尺度が含まれる場合、平均の全体的な相対誤差（平均値モデルに関連）が表示されます。すべての従属変数がカテゴリ変数の場合、不正な予測変数の平均割合が表示されます。相対誤差または不正な予測変数の割合は、それぞれの従属変数に対しても表示されます。
- **分類結果 (距離と近接度)。** 各カテゴリ従属変数の分類テーブルを表示します。各テーブルは、各従属変数カテゴリに正しくまたは誤って分類されたケースの数を示します。正しく分類されたケース全体の割合も報告されます。
- **ROC 曲線。** 各カテゴリ従属変数の ROC（受信者動作特性）曲線を表示します。また、各曲線の下の領域を示すテーブルも表示します。特定の従属変数に対し、ROC 曲線のグラフは 各カテゴリに 1 つの曲線を表示します。従属変数に 2 つのカテゴリがある場合、それぞれの曲線は問題のカテゴリを もう 1 つのカテゴリに対し正の状態として扱います。従属変数に 2 つを超えるカテゴリがある場合、それぞれの曲線は問題のカテゴリを その他すべてのカテゴリの集計変数に対し正の状態として扱います。
- **累積ゲイン グラフ。** 各カテゴリ従属変数の累積ゲイン グラフを表示します。ROC 曲線と同様、各従属変数カテゴリに対し 1 つの曲線を表示します。
- **リフト グラフ。** 各カテゴリ従属変数のリフト グラフを表示します。ROC 曲線と同様、各従属変数カテゴリに対し 1 つの曲線を表示します。

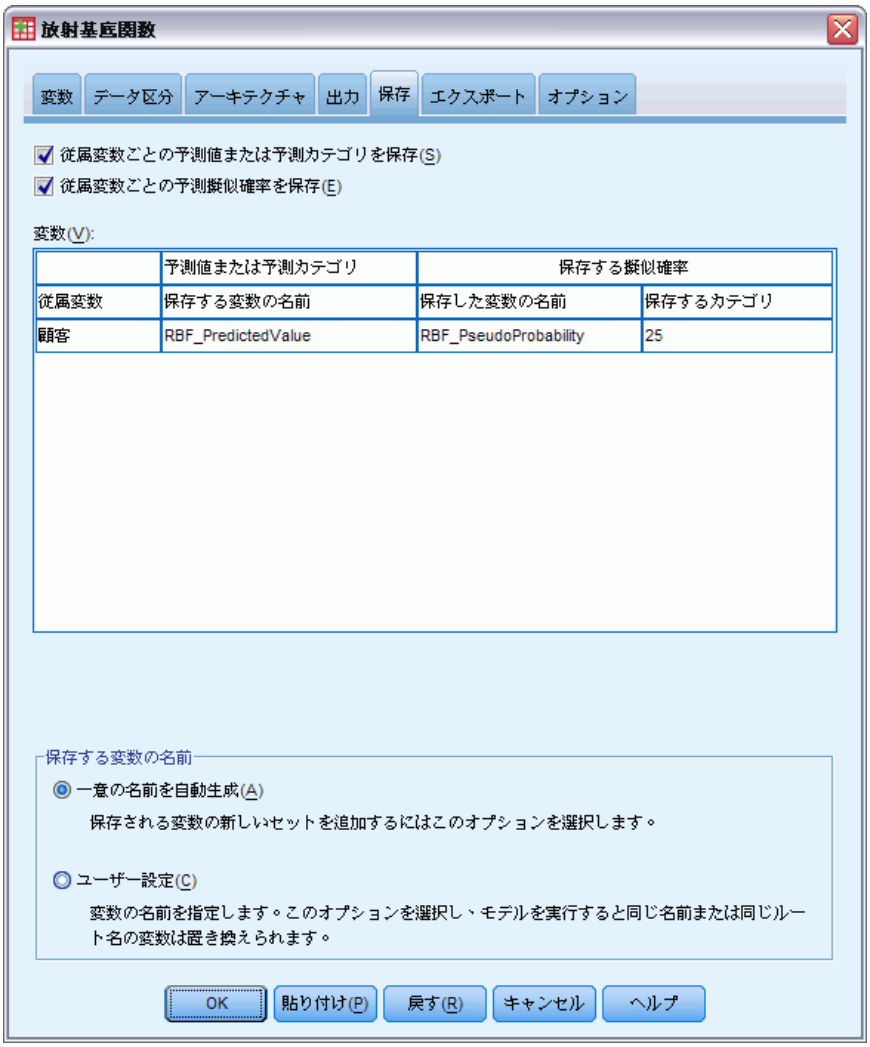
- **予測と観測のグラフ。**各従属変数の観測値により予測されるグラフを表示します。カテゴリ従属変数の場合、予測擬似確率のクラスタ箱ひげ図が各応答カテゴリに対し、各応答カテゴリをクラスタ変数として表示されます。スケール従属変数の場合、散布図が表示されます。
- **残差と予測のグラフ。**各スケール従属変数の予測値による残差グラフを表示します。残差と予測値の間にパターンが表示されることはありません。このグラフはスケール従属変数に対してのみ生成されます。

ケース処理の要約 分析に含まれたケースおよび除外されたケースの数を全体、学習サンプル、検定サンプルおよびホールドアウト サンプルごとに要約するケース処理の要約テーブルを表示します。

独立変数の重要度分析。感度分析を実行し、ニューラル ネットワークを定義する場合各予測変数の重要度を計算します。分析は、学習サンプルと検定サンプルの組み合わせに、また検定サンプルがない場合は学習サンプルのみに基づいています。この分析によって、各予測変数の重要度および正規化された重要度を表すテーブルおよびグラフを作成します。予測変数またはケースの数が多い場合、感度分析の計算は効率的でなく時間もかかります。

保存

図 3-6
[放射基底関数: 保存] タブ



[保存] タブを使用して、データセットの変数として予測変数を保存します。

- **従属変数ごとの予測値または予測カテゴリを保存。** これにより、スケール従属変数に予測された値を保存し、カテゴリ従属変数に予測カテゴリを保存します。
- **各従属変数の予測される疑似確率を保存する。** カテゴリ従属変数に予測擬似確率を保存します。各変数は、それぞれの最初の n カテゴリに対して保存されます。この場合、 n は [保存するカテゴリ] 列で指定されます。

保存する変数の名前 自動的な名前の生成によって、すべての作業を保存することができます。ユーザー指定の名前によって、データ エディタで保存された変数を最初に削除することなく、前回実行された結果を破棄または置き換えることができます。

確率および擬似確率

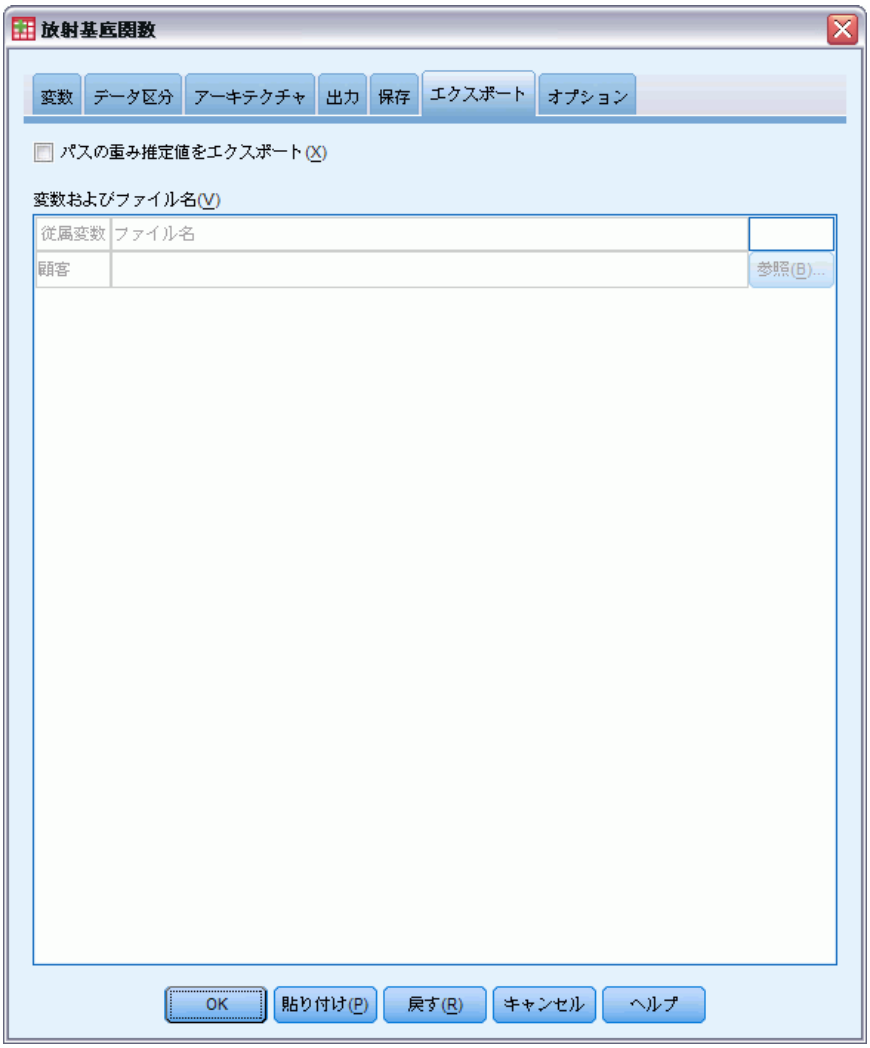
[放射基底関数] 手続きでは出力層に対し平方和の誤差および同一活性化関数を使用するため、予測擬似確率を確率として解釈することはできません。値が 0 より小さいまたは 1 より大きい場合であっても、または特定の従属変数の合計が 1 でない場合でも、この手続きではこれらの予測擬似確率を保存します。

ROC 曲線、累積ゲイン グラフおよびリフト グラフ（ p.33 出力 を参照）は、擬似確率に基づいて作成されます。擬似確率が 0 より小さいまたは 1 より大きい、または特定の変数の合計が 1 でない場合、擬似確率は 0 〜 1 に、変数の合計は 1 に再調整されます。擬似確率は合計で分割することによって再調整されます。たとえば、ケースに 3 つのカテゴリ従属変数に対し 0.50、0.60 および 0.40 の擬似確率が存在する場合、各擬似確率は合計の 1.50 で分割され、0.33、0.40、0.27 を取得します。

擬似確率のいずれかが負である場合、前述の再調整を行う前に最も低い確率の絶対値をすべての擬似確率に追加します。たとえば、擬似確率が -0.30、0.50、および 1.30 である場合、まず各値に 0.30 を追加して 0.00、0.80、および 1.60 を取得します。次に、それぞれの新しい値を、2.40 で分割して 0.00、0.33、および 0.67 を取得します。

エクスポート

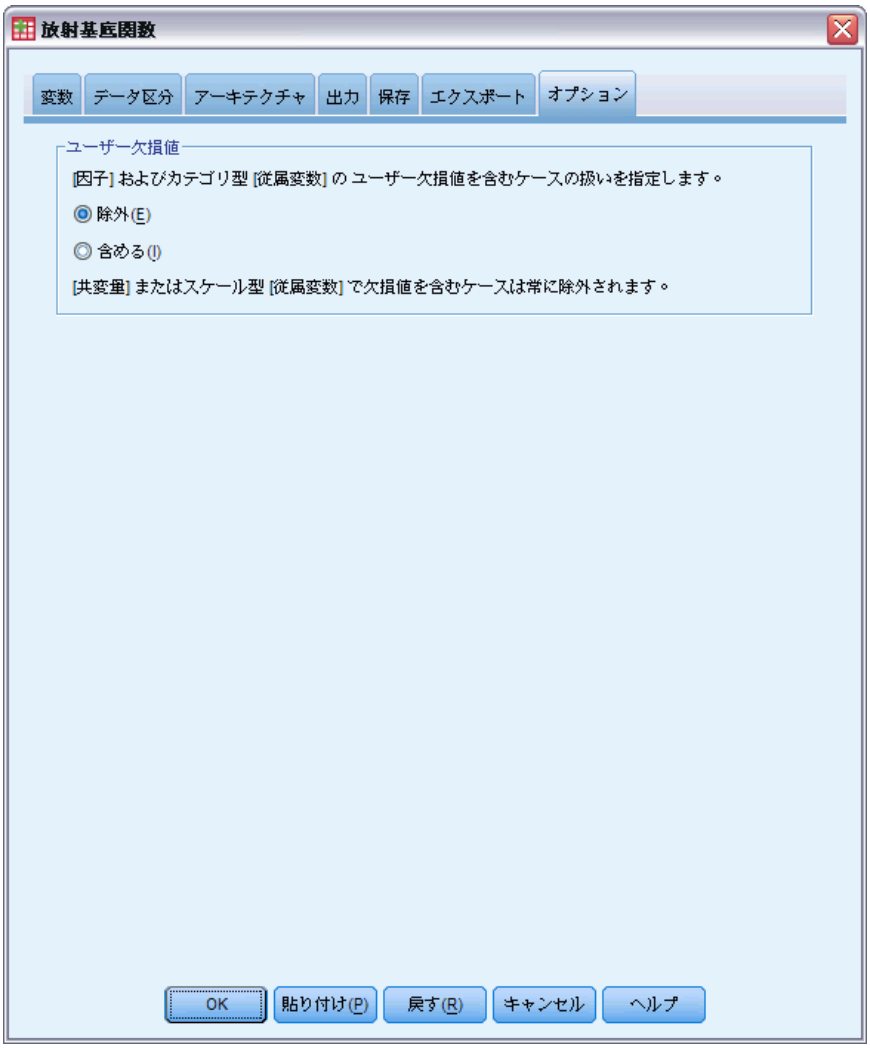
図 3-7
[放射基底関数: エクスポート] タブ



[エクスポート] タブを使用して、各従属変数のシナプスの重みの推定値を XML (PMML) ファイルに保存します。このモデル ファイルを使用して、得点付けのために他のデータ ファイルにモデル情報を適用できます。分割ファイルが定義されている場合、このオプションは使用できません。

オプション

図 3-8
[放射基底関数: オプション] タブ



ユーザー欠損値。 因子は、分析の対象となるケースに対して有効な値を取る必要があります。このオプションを使用すると、ユーザー欠損値を因子変数およびカテゴリ従属変数で有効な値として扱うかどうかを決定できます。

パート II:
例

章
4

多層パーセプトロン

多層パーセプトロン（MLP）手続きは、予測変数の値に基づいて、1 つ以上の従属（目標）変数に対する予測モデルを生成します。

多層パーセプトロンを使用した信用リスクの評価

銀行の融資担当者は、債務不履行になる可能性がある人物を示す特徴を特定し、その特徴を使用して信用リスクの良し悪しを識別する必要があります。

過去の客や見込み客 850 人の情報が、bankloan.sav に格納されているものとします。詳細は、 A 付録 p.92 サンプル ファイル を参照してください。 最初の 700 ケースは、以前に貸付を行った顧客です。分析の妥当性を検査するため、残りの顧客は除外して、この 700 人の顧客の無作為抽出を使って、多層パーセプトロンを作成します。それから、このモデルを利用して、見込み客 150 人を、信用リスクの高い人と低い人とに分類します。

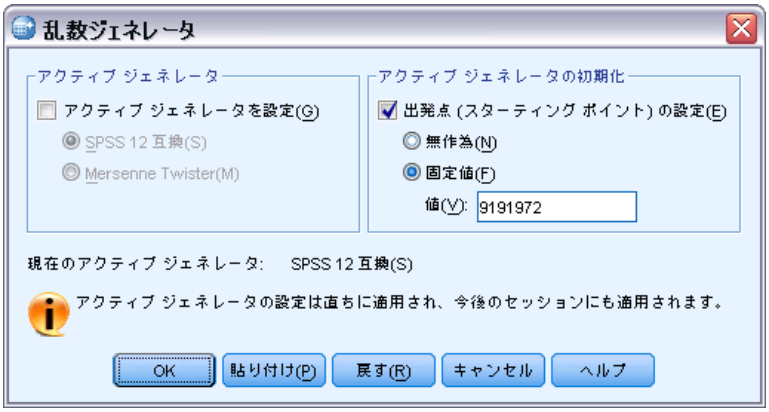
また、融資担当者は、ロジスティック回帰（Regression オプションの）を使用して以前にデータを分析したことがあり、それと比較して、分類ツールとしての多層パーセプトロンはどうなのかを知りたいと思っています。

分析用データの準備

乱数シードを設定すると、分析を正確に複製できます。

▶ 乱数シードを設定するには、メニューから次の項目を選択します。
変換 > 乱数ジェネレータ

図 4-1
[乱数ジェネレータ] ダイアログ ボックス



- [出発点 (スターティング ポイント) の設定] を選択します。
- [固定値] を選択し、値として「9191972」と入力します。
- [OK] をクリックします。

前のロジスティック回帰分析では、過去の顧客の約 70% が学習サンプルに割り当てられ、30% がホールドアウト サンプルに割り当てられました。これらの分析で使用するサンプルを正確に再作成するには、データ区分変数が必要です。

- データ区分変数を作成するには、メニューから次の項目を選択します。
  [変換] > [変数の計算...]

図 4-2
[変数の計算] ダイアログ ボックス

- ［目標変数］テキスト ボックスに「**データ区分**」と入力します。
- ［数式］テキスト ボックスに「**2*rv.bernoulli(0.7)-1**」と入力します。

  これにより、「データ区分」の値は、0.7 という確率パラメータで無作為に生成され、1 または 0 ではなく 1 または -1 という値になるように変更される**Bernoulli 変量**として設定されます。データ区分変数が正の値のケースは学習サンプルに割り当てられ、負の値のケースはホールドアウトサンプルに割り当てられ、値が 0 のケースは検証サンプルに割り当てられます。この場合は、検証サンプルは指定しません。

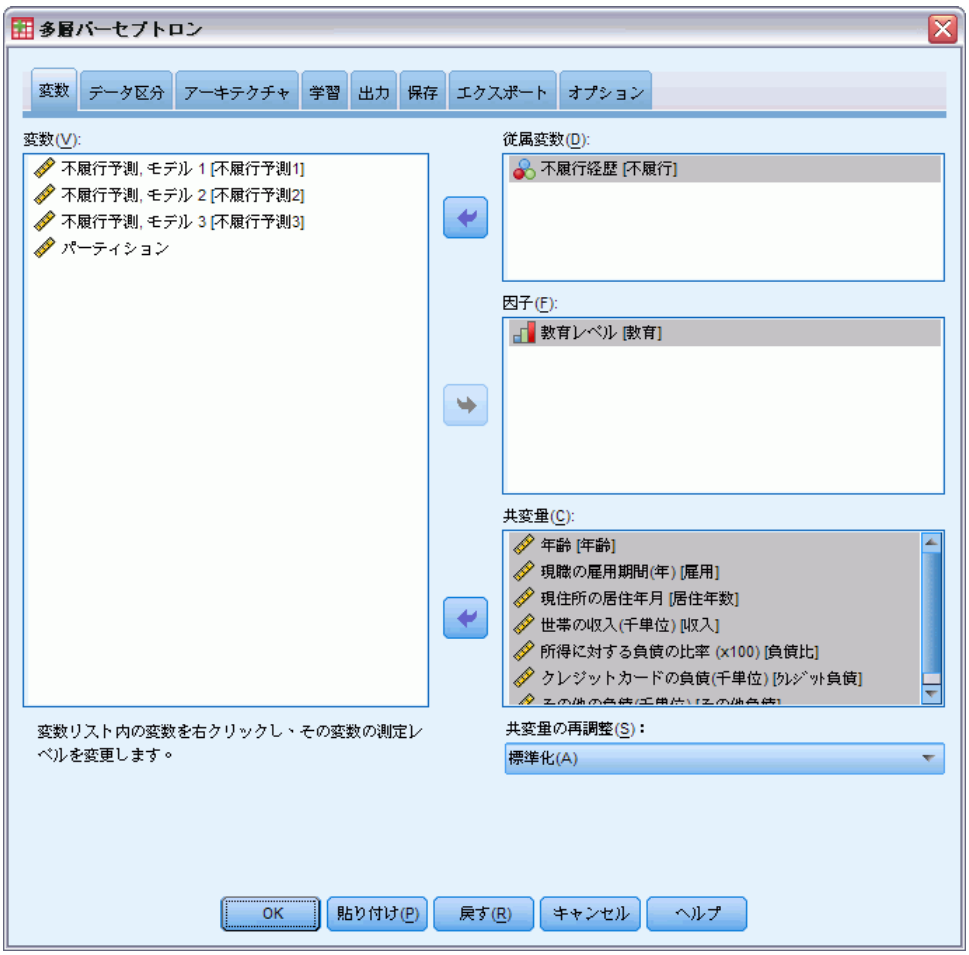
- ［変数の計算］ダイアログ ボックスで [OK] をクリックします。

  以前に融資を受けた顧客の約 70% は、[partition (分割)] の値が 1 になります。これらの顧客は、モデルの作成に使用します。以前に融資を受けた残りの顧客は、「データ区分」の値が -1 であり、モデルの結果の検証に使用されます。

分析の実行

▶ 多層パーセプトロン分析を実行するには、メニューから次の項目を選択します。
[分析] > [ニューラル ネットワーク] > [Multilayer Perceptron...（多層パーセプトロン...）]

図 4-3
[多層パーセプトロン: 変数] タブ



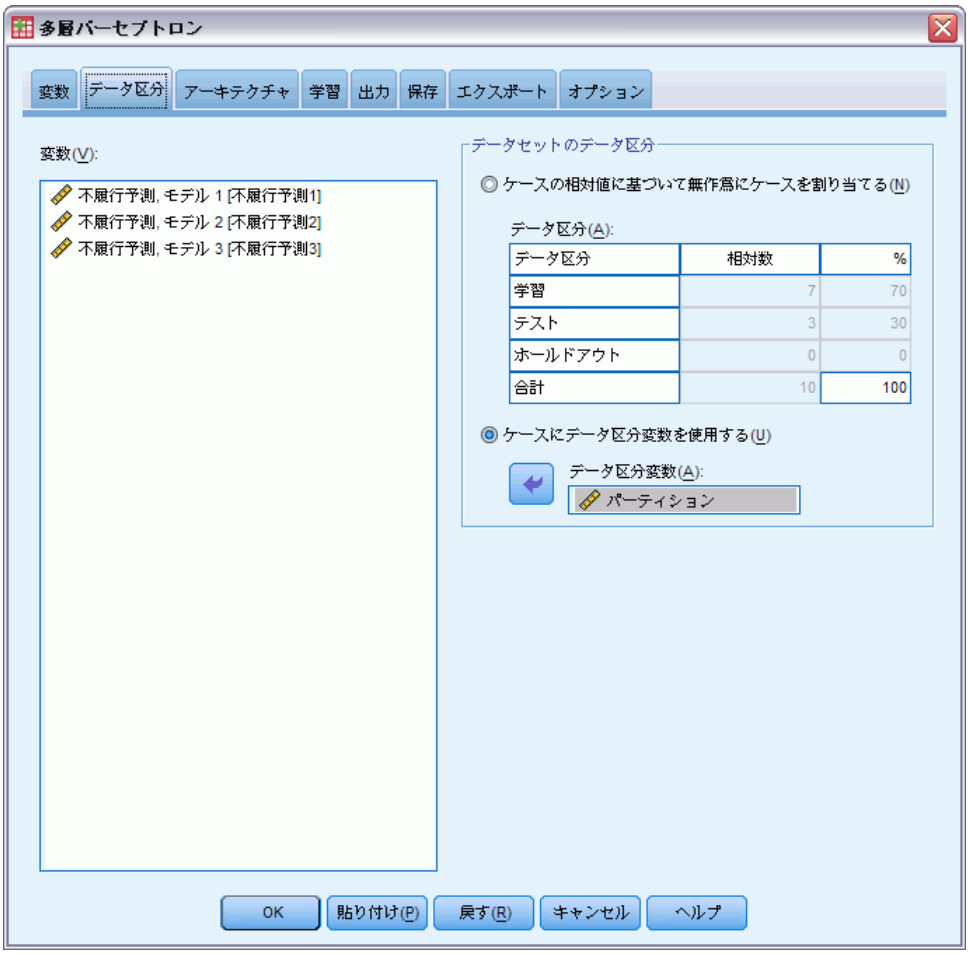
▶ 従属変数として［不履行履歴［不履行］］を選択します。

▶ 因子として［教育レベル［教育］］を選択します。

▶ 共変量として［年齢［年齢］］から［その他の負債（千単位）［その他の負債］］までを選択します。

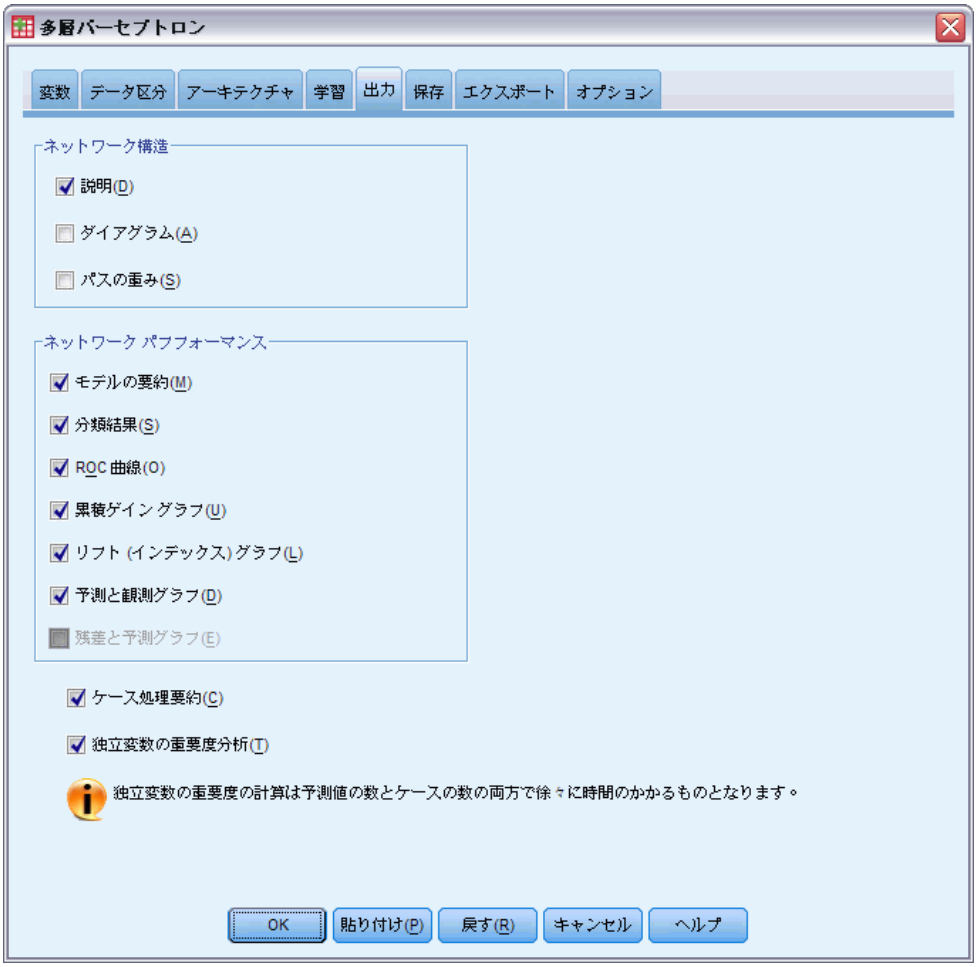
▶ [データ区分] タブをクリックします。

図 4-4
[多層パーセプトロン: データ区分] タブ



- [ケースの割り当てにデータ区分変数を使用する] を選択します。
- データ区分変数として「データ区分」を選択します。
- [出力] タブをクリックします。

図 4-5
[多層パーセプトロン: 出力] タブ



- ［ネットワーク構造］グループの［**ダイアグラム**］の選択を解除します。
- ［ネットワーク パフォーマンス］グループで、［**ROC 曲線**］、［**累積ゲイン グラフ**］、［**リフト (インデックス) グラフ**］、および［**残差と予測グラフ**］を選択します。［予測による残差］は従属変数がスケールではないため、使用できません。
- ［**独立変数の重要度分析**］を選択します。
- ［OK］をクリックします。

処理したケースの要約

図 4-6
ケース処理の要約

		度数	パーセント
サンプル	学習	499	71.3%
	ホールドアウト	201	28.7%
	有効数	700	100.0%
	除外数	150	
	合計	850	

ケース処理の要約は、499 個のケースが学習サンプルに割り当てられ、201 個のケースがホールドアウト サンプルに割り当てられたことを示しています。分析から除外された 150 個のケースは見込み客です。

ネットワーク情報

図 4-7
ネットワーク情報

入力層	Factors	1	教育レベル
	Covariates	1	年齢
		2	現職の雇用期間(年)
		3	現住所の居住年月
		4	世帯の収入(千単位)
		5	所得に対する負債の比率 (x100)
		6	クレジットカードの負債(千単位)
		7	その他の負債(千単位)
	単位数[a]		12
	Rescaling Method for Covariates		標準化
隠れ層	隠れ層の数		1
	隠れ層1のユニット数[a]		6
	活性化関数		双曲線正接
出力層	Dependent Variables	1	不履行経歴
	単位数		2
	活性化関数		Softmax
	誤差関数		ユニット数

a. バイアス ユニットは除外

ネットワーク情報テーブルには、ニューラル ネットワークに関する情報が表示され、指定が正しいことを確認するのに使用できます。ここで、特に次の点に注意する必要があります。

- 入力層の単位数は、共変量の数に因子レベルの合計数を加えたものです。「教育のレベル」のカテゴリごとに個別の単位が作成され、多くのモデル手続きで一般的であるように、「冗長」と見なされるカテゴリはありません。
- 同様に、「不履行履歴」のカテゴリごとに個別の出力単位が作成され、出力層の単位の数は全部で 2 です。

- 自動アーキテクチャ選択は、隠れ層で 4 単位を選択しています。
- 他のすべてのネットワーク情報は、手続きのデフォルトです。

モデルの要約 (ピボットテーブル 回帰)

図 4-8
モデルの要約

学習	交差エントロピの誤差	146.426
	誤った予測値の割合	11.8%
	停止規則の使用	エポックの最大数 (100) を超えています
	学習時間	00:00:00.640
ホールドアウト	誤った予測値の割合	24.4%

従属変数 : 不履行経歴

モデルの要約では、学習結果および最終ネットワークをホールドアウト サンプルに適用した結果に関する情報が表示されます。

- 出力層は softmax アクティブ化関数を使用するので、クロス エントロピー誤差が表示されます。これは、ネットワークが学習中に最小化しようとする誤差関数です。
- 不正な予測のパーセントは、分類表から取得されます。詳細については、該当するトピックで説明します。
- 推定アルゴリズムは、エポックの最大数に達したので停止しました。誤差が収束しているので、理想的には学習を停止する必要があります。学習中に問題が発生している可能性があり、さらに出力を検査するときに留意する必要があります。

分類

図 4-9
分類

		予測		
サンプル	観測	なし	あり	Percent Correct
学習	なし	347	28	92.5%
	あり	50	74	59.7%
	Overall Percent	79.6%	20.4%	84.4%
ホールドアウト	なし	123	19	86.6%
	あり	32	27	45.8%
	Overall Percent	77.1%	22.9%	74.6%

従属変数 : 不履行経歴

分類テーブルでは、実際にネットワークを使用した結果が表示されます。ケースごとの予測応答は、ケースの予測される擬似確率が 0.5 より大きい場合は［はい］です。サンプルごとに次のようになります。

- ケースのクロス分類の対角線上のセルは、正しい予測値です。
- ケースのクロス分類の対角線から外れたセルは、正しくない予測値です。

モデルを作成するために使われたケースのうちで、過去に債務不履行があった 124 人中 74 人は、適切に分類されます。また、債務不履行のなかった 375 人中 347 人も正しく分類されます。全体として、学習ケースの 84.4% は正しく分類され、15.6% は正しくないことが、モデルの要約表で示されます。モデルがよくなればなるほど、適切に分類されるケースのパーセントも高くなるはずです。

モデルを作成するために使われたケースに基づく分類は、分類率が誇張されるので、「楽観的」になりすぎる傾向があります。ホールドアウト サンプルは、モデルの検証に役立ちます。この場合は、ケースの 74.6% がモデルによって正しく分類されました。これは、実際には、モデルの分類の約 3/4 が正しいということを示しています。

過度な学習の修正

以前に行ったロジステック回帰分析では、学習サンプルとホールドアウト サンプルの両方がほぼ同じ割合（約 80%）のケースを正しく予測しました。一方、ニューラル ネットワークの場合は、学習サンプルでは正しいケースの割合が高く、ホールドアウト サンプルでは実際に債務不履行になった顧客の正しい予測が大幅に低下しました（正しい割合が、ホールドアウト サンプルでは 45.8%、学習サンプルでは 59.7%）。モデルの要約表で示された停止ルールと合わせて考えると、ネットワークが **過度な学習**を行っている可能性があります。つまり、ランダムな変動によって学習データに現れる誤ったパターンを追求しています。

そのような場合でも、これは比較的簡単に解決できます。つまり、ネットワークが「逸れない」ようにするための検証サンプルを指定します。ここでは、ロジステック回帰分析で使用された学習サンプルとホールドアウト サンプルが正確に再作成されるように、データ区分変数を作成しました。一方、ロジステック回帰には「検証」サンプルの概念はありません。学習サンプルの一部を取り出して、検証サンプルに割り当て直してみましょう。

検証サンプルの作成

図 4-10
[変数の計算] ダイアログ ボックス

- ▶ もう一度、[変数の計算] ダイアログ ボックスを表示します。
- ▶ [数式] テキスト ボックスに「**データ区分** = rv.bernoulli(0.2)」と入力します。
- ▶ [IF] をクリックします。

図 4-11
[変数の計算: IF 条件] ダイアログ ボックス

- [If 条件を満たしたケースを含む] チェック ボックスをオンにします。
- テキスト ボックスに「**データ区分>0**」と入力します。
- [続行] をクリックします。
- [変数の計算] ダイアログ ボックスで [OK] をクリックします。

これにより、0 より大きい [partition (分割)] の値は、約 20% が値 0 になり、残りの 80% が値 1 になるように再設定されます。全体として、以前に融資を受けた顧客の約 100*(0.7*0.8)=56% が学習サンプルに含まれ、14% が検証サンプルに含まれます。もともとホールドアウト サンプルに割り当てられていた顧客は、そのまま変わりません。

分析の実行

- [多層パーセプトロン] ダイアログ ボックスを再び開き、[保存] タブをクリックします。

► [各従属変数の予測される疑似確率を保存する] を選択します。

► [OK] をクリックします。

処理したケースの要約

図 4-12
検証サンプルでのモデルのケース処理の要約

		度数	パーセント
サンプル	学習	398	56.9%
	テスト	101	14.4%
	ホールドアウト	201	28.7%
	有効数	700	100.0%
	除外数	150	
	合計	850	

もともと学習サンプルに割り当てられていた 499 ケースのうち、101 が検証サンプルに再割り当てされました。

ネットワーク情報

図 4-13
ネットワーク情報

入力層	Factors	1	教育レベル
	Covariates	1	年齢
		2	現職の雇用期間(年)
		3	現住所の居住年月
		4	世帯の収入(千単位)
		5	所得に対する負債の比率 (x100)
		6	クレジットカードの負債(千単位)
		7	その他の負債(千単位)
	単位数[a]		12
	Rescaling Method for Covariates		標準化
隠れ層	隠れ層の数		1
	隠れ層1のユニット数[a]		6
	活性化関数		双曲線正接
出力層	Dependent Variables	1	不履行経歴
	単位数		2
	活性化関数		Softmax
	誤差関数		ユニット数

a. バイアス ユニットは除外

ネットワーク情報テーブルに対する唯一の変更は、自動アーキテクチャ選択が隠れ層で 7 単位を選択したことです。

モデルの要約 (ピボットテーブル 回帰)

図 4-14
モデルの要約

学習	交差エントロピの誤差	164.396
	誤った予測値の割合	18.8%
	停止規則の使用	減少のない 1 継続ステップがエラーです[a]
	学習時間	00:00:00.782
テスト	交差エントロピの誤差	45.152
	誤った予測値の割合	17.0%
ホールドアウト	誤った予測値の割合	20.9%

従属変数 : 不履行経歴

a. エラーの計算は、学習サンプルに基づいています。

モデルの要約では、よい兆候が 2 つ示されています。

- 正しくない予測のパーセントが学習、検証、およびホールドアウトの各サンプルでほぼ一定です。
- アルゴリズムのステップの後で誤差が減少しなかったので、推定アルゴリズムは停止しました。

このことはさらに、元のモデルでは実際には過度な学習が行われていた可能性があり、検証サンプルを追加することで問題が解決されたことを示唆しています。もちろん、サンプルのサイズは比較的小さく、わずかなパーセント ポイントの揺れを重視しすぎないようにする必要があります。

分類

図 4-15
分類

		予測		
サンプル	観測	なし	あり	Percent Correct
学習	なし	287	15	95.0%
	あり	60	37	38.1%
	Overall Percent	87.0%	13.0%	81.2%
テスト	なし	67	4	94.4%
	あり	13	16	55.2%
	Overall Percent	80.0%	20.0%	83.0%
ホールドアウト	なし	135	9	93.8%
	あり	33	24	42.1%
	Overall Percent	83.6%	16.4%	79.1%

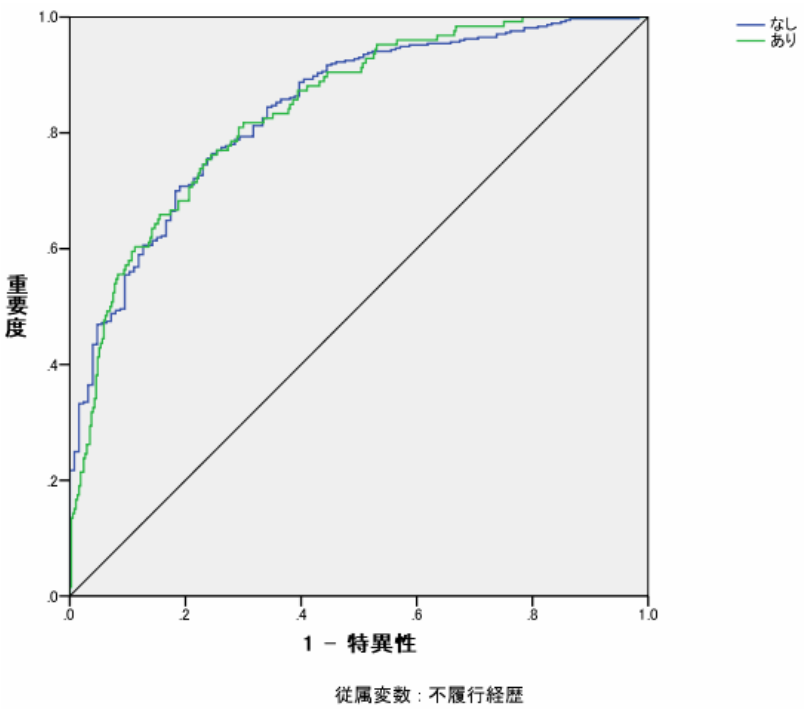
従属変数 : 不履行経歴

分類テーブルを見ると、分類の疑似確率の分割値として 0.5 を使用すると、ネットワークによる債務履行者の予測は不履行者の予測より大幅に向上することがわかります。残念ながら、単一の分割値ではネットワークの予測能力は非常に限られた範囲しかわからないので、競合するネッ

トワークを比較しても必ずしも役に立つとは限りません。代わりに、ROC 曲線を見ます。

ROC 曲線

図 4-16
ROC 曲線



ROC 曲線を見ると単一のプロットにおけるすべての可能な分割の**感度**と**特異性**がわかり、これは一連のテーブルよりはるかにわかりやすくて強力です。ここで示すグラフには 2 つの曲線が表示されており、1つはカテゴリ［いいえ］に対するもので、もう 1 つはカテゴリ［はい］に対するものです。2 つのカテゴリしかないので、曲線は、グラフの左上隅から右下への約 45 度の線（表示されてはいません）について対称になっています。

このグラフは、学習サンプルと検証サンプルの組み合わせに基づいていることに注意してください。ホールドアウト サンプルの ROC グラフを生成するには、データ区分変数でファイルを分割し、保存されている予測された疑似確率に対して ROC 曲線手続きを実行します。

図 4-17
曲線の下側の面積

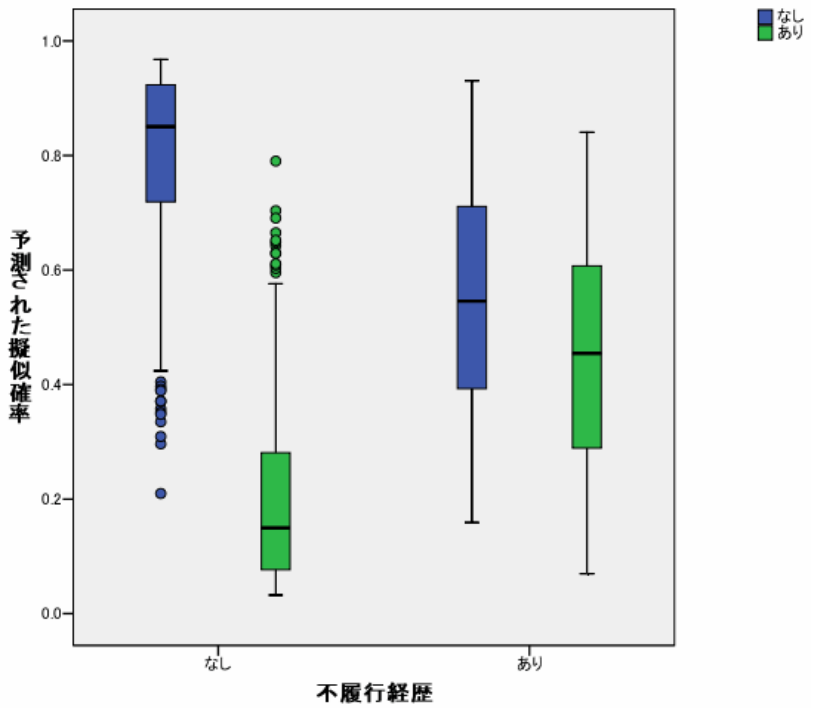
		面グラフ
不履行経歴	なし	.853
	あり	.853

曲線の下側の面積は ROC 曲線の数値的な要約であり、テーブルの値は、各カテゴリについて、そのカテゴリの予測された疑似確率が、そのカテゴリ以外で無作為に選択されたケースの場合より、そのカテゴリで無作為に選択されたケースの場合の方が高いことを表しています。たとえば、無作為に選択された債務不履行者と債務履行者の場合、債務不履行のモデル予測疑似確率が債務履行者より債務不履行者の方が高くなる確率は 0.853 です。

曲線の下側の面積はネットワークの精度についての役に立つ 1 つの統計的要約ですが、顧客を分類する具体的な条件を選択できる必要があります。観測により予測されるグラフは、このプロセスに対する目に見える起点を提供します。

観測により予測されるグラフ

図 4-18
観測により予測されるグラフ



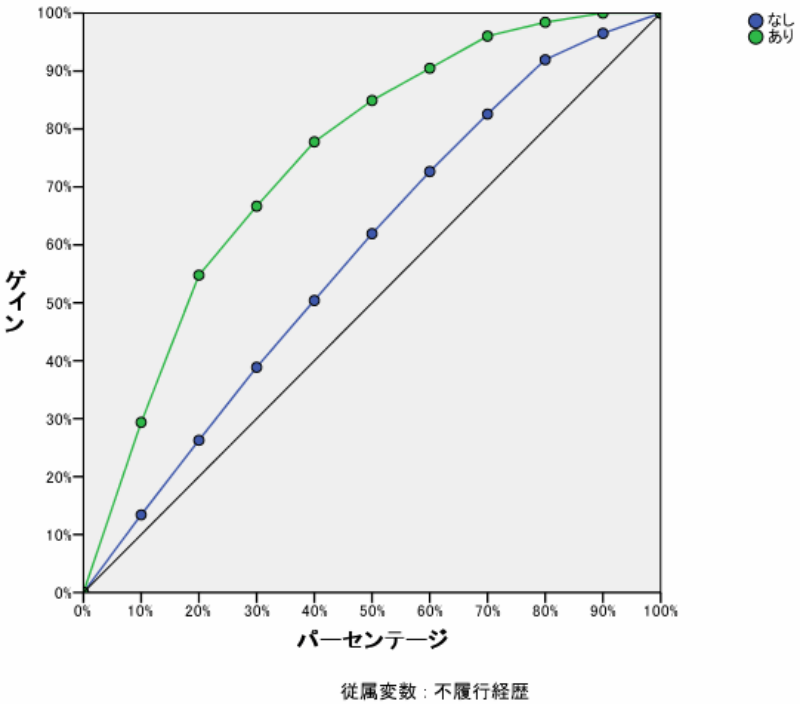
カテゴリ従属変数の場合、観測により予測されるグラフには、学習サンプルと検証サンプルの組み合わせに対する予測された疑似確率のクラスタ箱ひげ図が表示されます。x 軸は観測応答カテゴリに対応し、凡例は予測カテゴリに対応します。

- 左端の箱ひげ図は、観測されたカテゴリが **[いいえ]** であるケースの場合の、カテゴリ **[いいえ]** の予測疑似確率を表します。y 軸の 0.5 マークより上の箱ひげ図の部分は、分類テーブルで示されている正しい予測を表します。0.5 マークより下の部分は、正しくない予測を表します。分類テーブルから、0.5 分割を使用する **[いいえ]** カテゴリのケースをネットワークは非常によく予測するので、下側のひげの部分および若干の異常ケースのみが誤って分類されます。
- 右側の次の箱ひげ図は、観測されたカテゴリが **[いいえ]** であるケースの場合の、カテゴリ **[はい]** の予測疑似確率を表します。目標変数には 2 つのカテゴリしかないので、最初の 2 つの箱ひげ図は、0.5 の水平線について対称になっています。
- 3 番目の箱ひげ図は、観測されたカテゴリが **[はい]** であるケースの場合の、カテゴリ **[いいえ]** の予測疑似確率を表します。これと最後の箱ひげ図は、0.5 の水平線について対称になっています。
- 最後の箱ひげ図は、観測されたカテゴリが **[はい]** であるケースの場合の、カテゴリ **[はい]** の予測疑似確率を表します。y 軸の 0.5 マークより上の箱ひげ図の部分は、分類テーブルで示されている正しい予測を表します。0.5 マークより下の部分は、正しくない予測を表します。分類テーブルから、ネットワークは 0.5 分割を使用する **[はい]** カテゴリのケースの半分より少しだけ多くを予測するので、箱のよい部分は誤って分類されます。

プロットを見ると、**[はい]** としてケースを分類するための分割を 0.5 から約 0.3 – これは、2 番目の箱の上端と、4 番目の箱の下端になります – に下げることで、よい顧客を多く失うことなく、債務不履行の可能性のある顧客を正確にとらえる可能性が高くなることがわかります。つまり、2 番目の箱に沿って 0.5 から 0.3 に移動すると、ひげに沿った比較的少数の債務履行者は予測される債務不履行者として誤って再分類されますが、4 番目の箱に沿って移動すると、箱内の多くの債務不履行顧客が予測される債務不履行者として正しく再分類されます。

累積ゲイン グラフとリフト図表

図 4-19
累積ゲイン グラフ



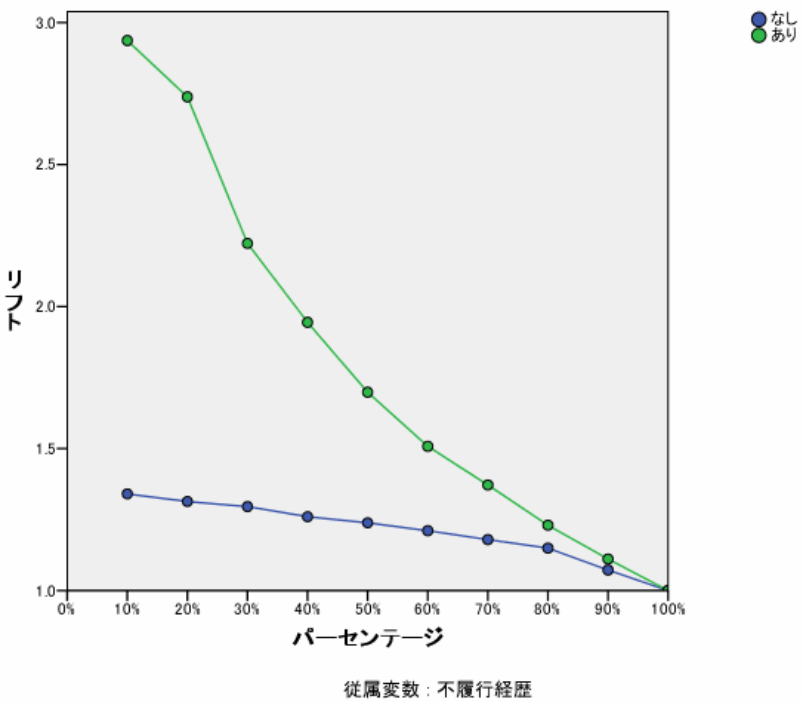
累積ゲイン グラフは、ケースの合計数のパーセントを目標にすることで、特定のカテゴリ「ゲイン」のケースの総数のパーセントを示します。たとえば、[はい] カテゴリの曲線の最初のポイントは（10%, 30%）であり、これは、ネットワークでデータセットをスコアリングし、[はい] の予測された疑似確率ですべてのケースをソートした場合、上位 10% が、実際にカテゴリ［はい］（債務不履行者）である全ケースの約 30% を含むと期待することを意味します。同様に、上位 20% は債務不履行者の約 50% を含み、上位 30% は債務不履行者の約 70% を含みます。スコアリングされたデータセットの 100% を選択すると、すべての債務不履行者がデータセットに含まれます。

対角線は「ベースライン」曲線です。スコアリングされたデータセットから無作為にケースの 10% を選択すると、実際にカテゴリが [はい] である全ケースの約 10% を「ゲインする」ことが期待されます。曲線がベースラインより上になるほど、ゲインが大きくなります。累積ゲイン グラフを使用すると、目標のゲインに対応するパーセントを選択し、パーセントを適切な分割値にマッピングすることで、分類の分割を選択できます。

何が「望ましい」ゲインを構成するかは、タイプ I およびタイプ II の誤りにかかるコストに依存します。つまり、債務の不履行者を履行者として分類する際のコスト（タイプ I）はどのようなもので、債務の履行者を

不履行者として分類する際のコスト（タイプ II）はどのようなものかが重要になります。不良債権に最も関心がある場合は、タイプ I の誤りを小さくします。累積ゲイン グラフでは、これは、[はい] の予測された疑似確率が上位 40% に入る申請者への融資を断ることに対応します。この部分には、可能性のある債務不履行者の約 90% が含まれますが、除去される申請者プールはほぼ半分です。 顧客基盤の拡大を優先する場合は、タイプ II の誤りを小さくします。累積ゲイン グラフでは、これは、債務不履行者の 30% が含まれる上位 10% を断り、申請者プールの大部分はそのままにすることに対応します。通常は、両方とも主要な問題なので、感度と特異性が適切に組み合わされた、顧客を分類するための決定規則を選択する必要があります。

図 4-20
リフト図表



リフト図表は、累積ゲイン グラフから導かれます。 y 軸の値は、各曲線の累積ゲインの、ベースラインに対する比率に対応します。したがって、カテゴリ [はい] の 10% におけるリフトは、30%/10% = 3.0 です。累積ゲイン グラフの情報を別の方法で見ることができます。

注: 累積ゲイン グラフとリフト図表は、学習サンプルと検証サンプルの組み合わせに基づいています。

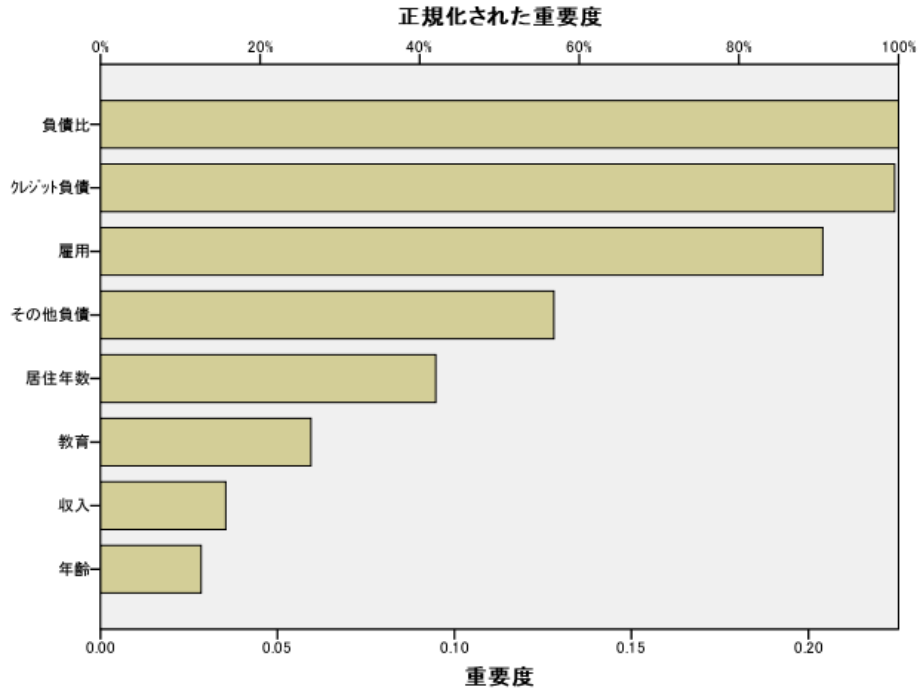
独立変数の重要度

図 4-21
独立変数の重要度

	重要度	正規化された重要度
教育レベル	.032	11.9%
年齢	.075	27.9%
現職の雇用期間(年)	.268	100.0%
現住所の居住年月	.166	61.8%
世帯の収入(千単位)	.033	12.2%
所得に対する負債の比率 (x100)	.125	46.5%
クレジットカードの負債 (千単位)	.213	79.3%
その他の負債(千単位)	.090	33.6%

独立変数の重要度は、独立変数の異なる値に対してネットワークのモデル予測値が変化する大きさの尺度です。正規化された重要度は、単に重要度の値を最大の重要度の値で割ったものであり、パーセントで表されます。

図 4-22
独立変数の重要度グラフ



重要度グラフは、重要度テーブルの値の棒グラフであり、重要度の値の降順にソートされています。顧客の安定性（雇用、居住年数）と負債（クレジット負債、負債比）に関連する変数が、ネットワークによる顧客の分類に最大の影響を持つことがわかります。わからないことは、これらの変数

と、債務不履行の予測される確率の間の関係の「向き」です。負債が大きいほど債務不履行の確率が大きいことを示すと考えられますが、もっと簡単に解釈できるパラメータのモデルを使用する必要があります。

要約

多層パーセプトロン手続きを使用して、特定の顧客が債務不履行になる確率を予測するためのネットワークを構築しました。モデルの結果は、ロジスティック回帰または判別分析を使用して得られる結果と比較できるので、これらのモデルでは獲得できない関係をデータが含まないことを確信し、これらを使用して、従属変数と独立変数の間の関係の特性をさらに調査できます。

多層パーセプトロンを使用した医療費および滞在期間の推定

病院のシステムの場合、心筋梗塞（MI または「心臓発作」）の治療で入院している患者の入院費用や期間を記録追跡できます。これらの測定値を正確に推定することにより、経営陣は、患者の治療に利用可能なベッド スペースを適切に管理することができます。

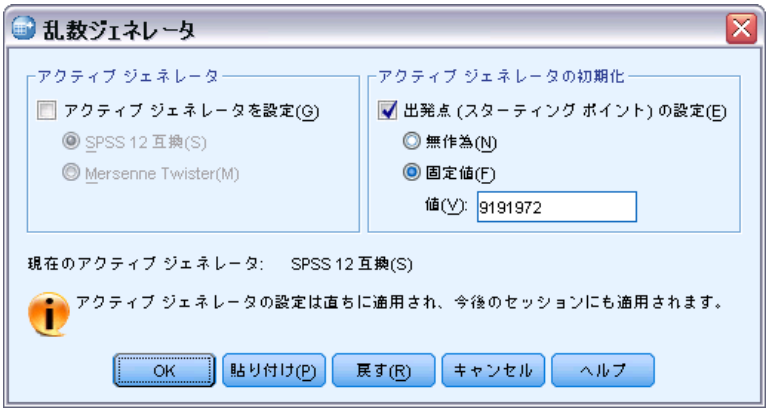
データ ファイル patient_los.sav には、MI の治療を受けた患者のサンプルの治療記録が含まれています。詳細は、 A 付録 p.92 サンプル ファイル を参照してください。多層パーセプトロン手続きを使用して、費用と滞在期間を予測するためのネットワークを作成します。

分析用データの準備

乱数シードを設定すると、分析を正確に複製できます。

▶ 乱数シードを設定するには、メニューから次の項目を選択します。
変換 > 乱数ジェネレータ

図 4-23
[乱数ジェネレータ] ダイアログ ボックス



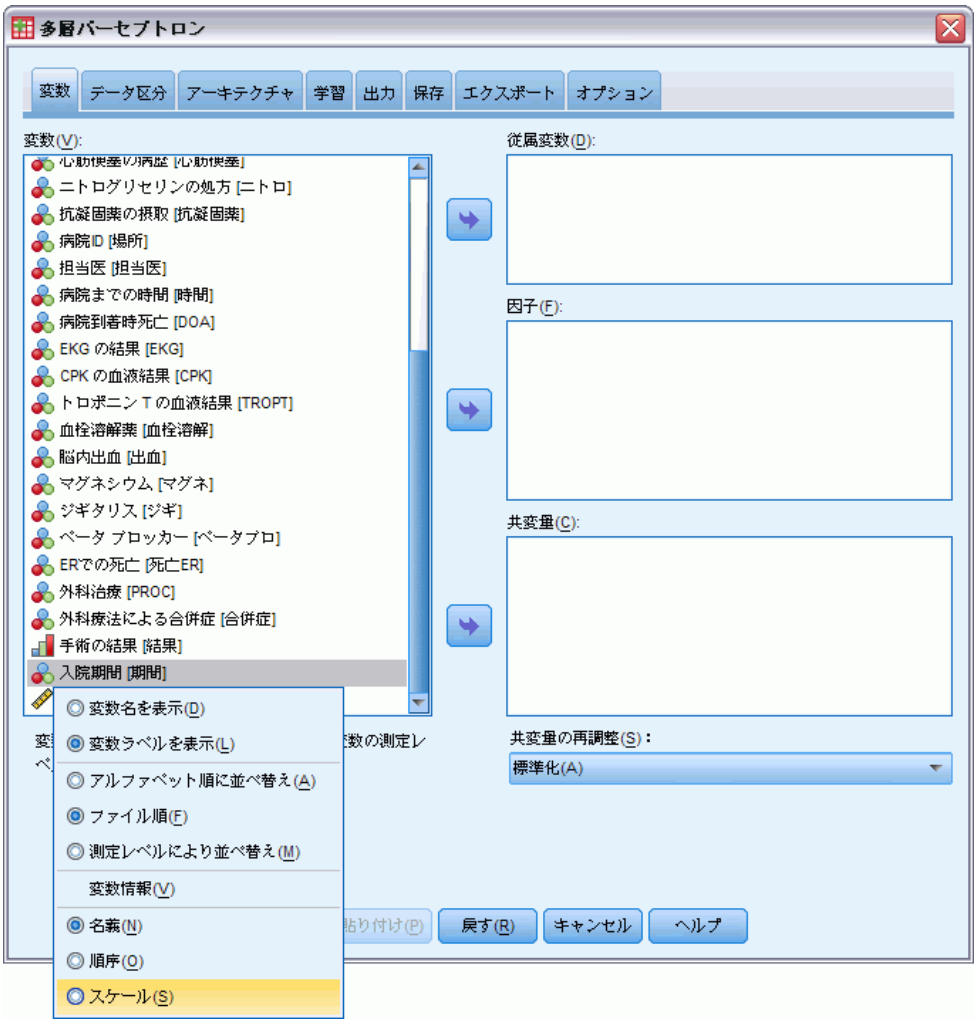
- [出発点 (スターティング ポイント) の設定] を選択します。
- [固定値] を選択し、値として「9191972」と入力します。
- [OK] をクリックします。

分析の実行

- 多層パーセプトロン分析を実行するには、メニューから次の項目を選択します。

  分析 > ニューラル ネットワーク > 多層パーセプトロン...

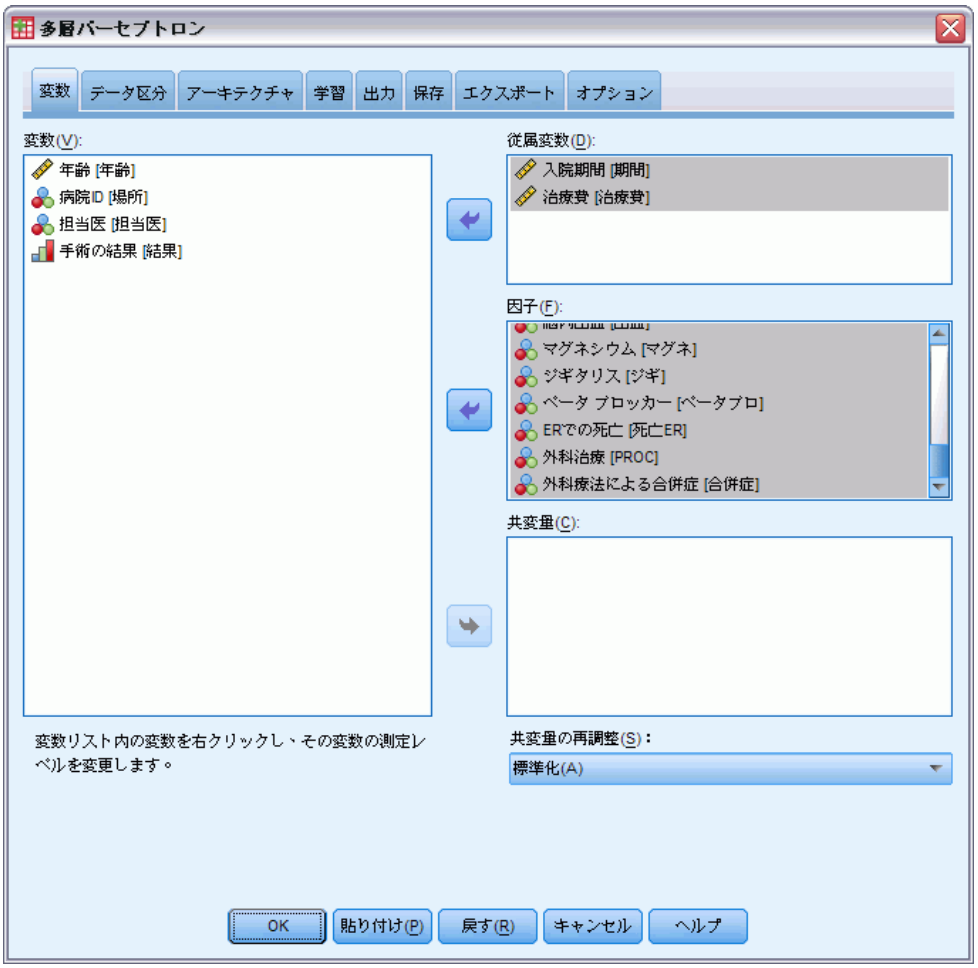
図 4-24
[多層パーセプトロン: 変数] タブと滞在期間のコンテキスト メニュー



「入院期間［期間］」には順序尺度がありますが、ネットワークにそれをスケールとして扱わせる必要があります。

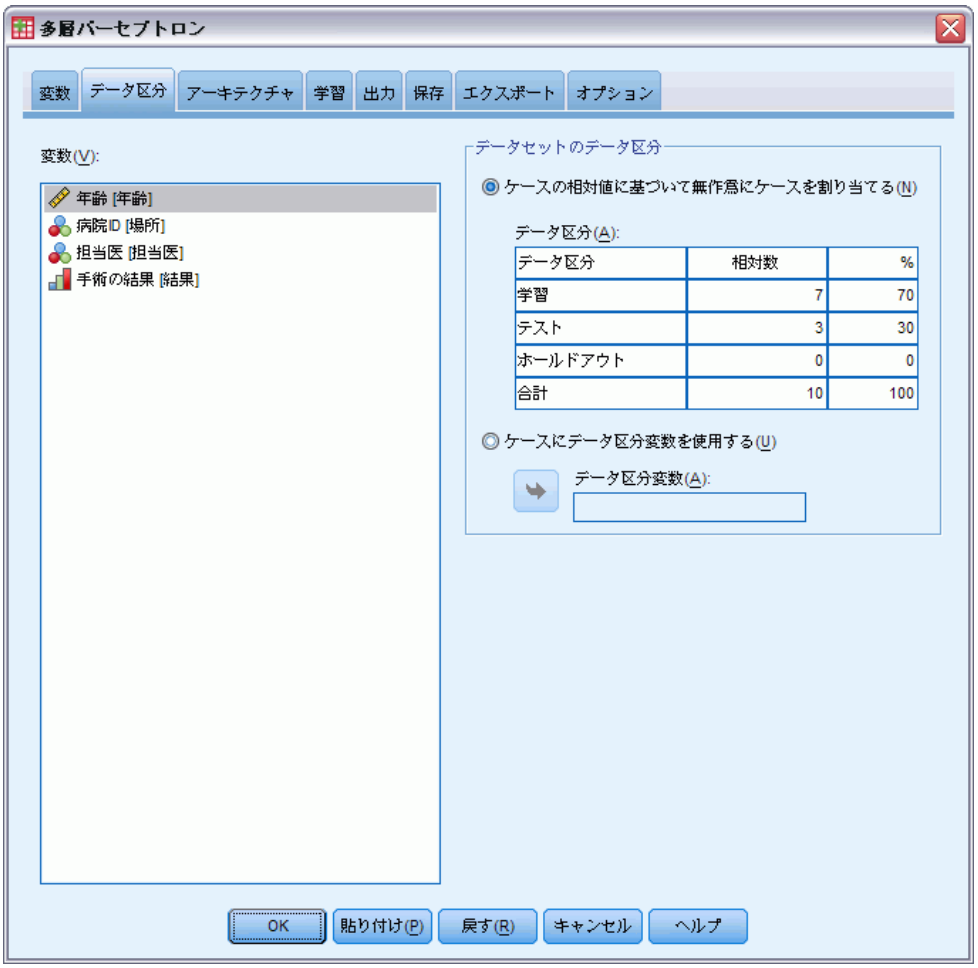
- 「入院期間［期間］」を右クリックし、コンテキスト メニューで [スケール] を選択します。

図 4-25
選択した従属変数と因子がある、[多層パーセプトロン: 変数] タブ



- 「入院期間［期間］」および「治療費［治療費］」を従属変数として選択します。
- 因子として「[年齢［年齢］」から「抗凝固薬の摂取［抗凝固薬］」までと「入院期間［時間］」から「外科合療法による合併症［合併症］」までを選択します。以下のモデルの結果を正確に複製するため、因子リストの変数の順序を保持してください。そのため、各予測変数セットを選択し、ドラッグ アンド ドロップではなくボタンを使用して因子リストに移動すると良いでしょう。または、変数の順序を変更すると、解の安定性を評価するのに役立ちます。
- **[データ区分]** タブをクリックします。

図 4-26
[多層パーセプトロン: データ区分] タブ



- 検証用サンプルに割り当てるケースの相対的な数として「2」を入力します。
- ホールドアウト サンプルに割り当てるケースの相対的な数として「1」を入力します。
- [アーキテクチャ] タブをクリックします。

図 4-27
[多層パーセプトロン: アーキテクチャ] タブ

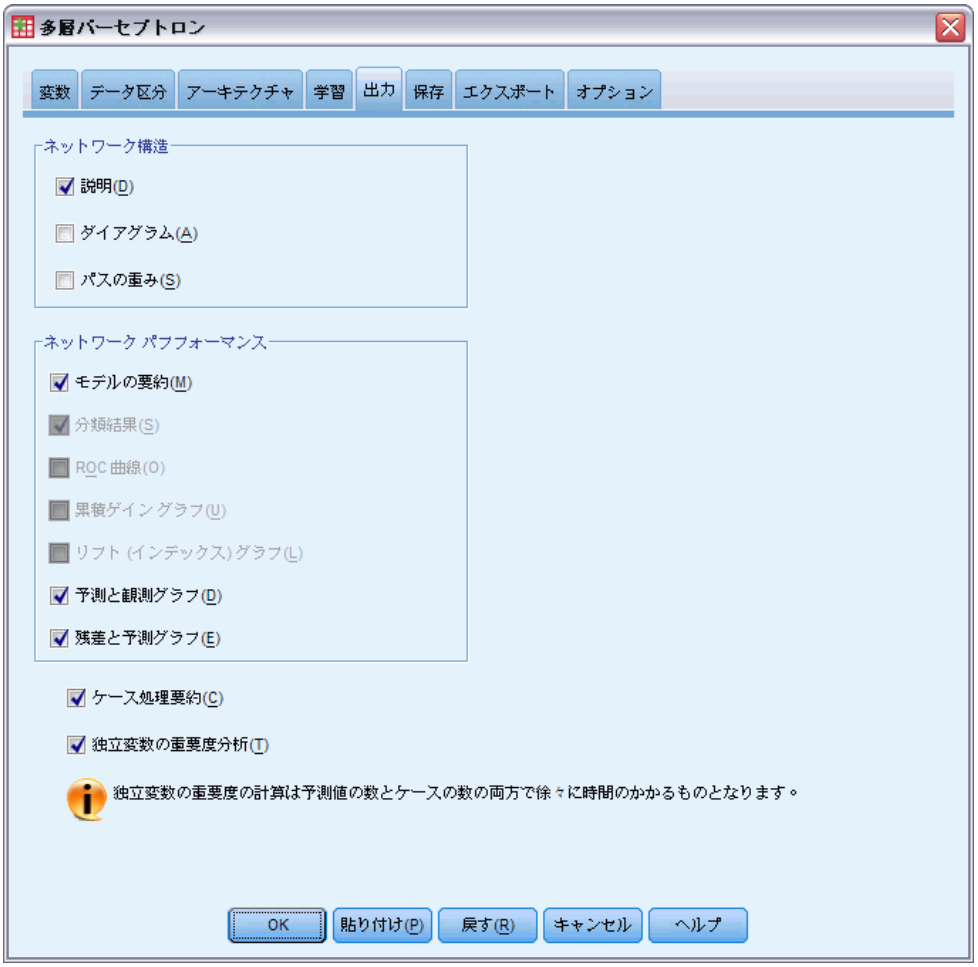
- ▶ [**カスタム アーキテクチャ**] を選択します。
- ▶ 隠れ層の数として、「2」を入力します。
- ▶ 層出力アクティブ化関数として [**双曲線正接**] を選択します。これによって、自動的に従属変数の尺度設定方法が [**調整済み正規化**] に設定されます。
- ▶ [**学習**] タブをクリックします。

図 4-28
[多層パーセプトロン: 学習] タブ

- 学習の種類として [**オンライン**] を選択します。オンライン学習は、相関関係のある予測変数のある「大きな」データセットでうまく機能するようにサポートされています。このため、アルゴリズムの最適化として、自動的に [**勾配下降**] が対応するデフォルトのオプションで設定されます。

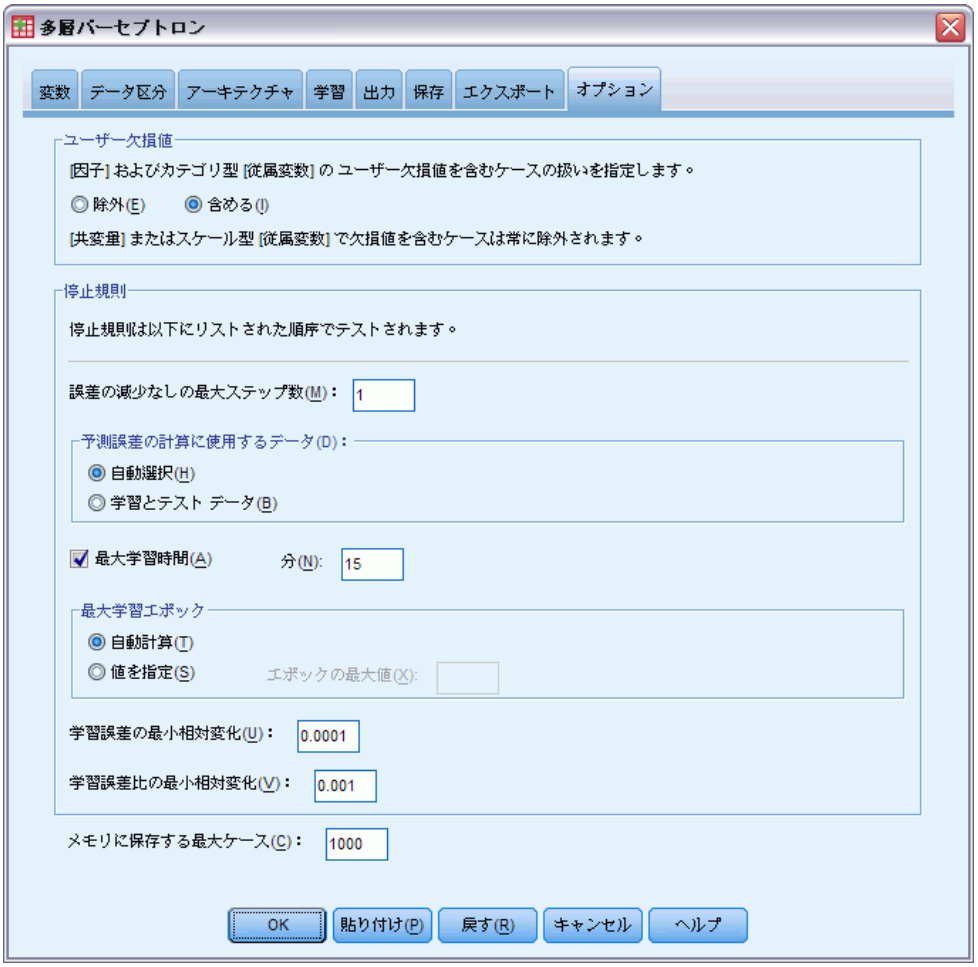
- [**出力**] タブをクリックします。

図 4-29
[多層パーセプトロン: 出力] タブ



- [**ダイアグラム**] の選択を解除します。多数の入力があるため、生成される図は扱いにくいものになります。
- [ネットワーク パフォーマンス] グループの [**予測と観測のグラフ**] および [**残差と予測のグラフ**] を選択します。分類結果、ROC 曲線、累積ゲイン グラフ、およびリフト図表は使用できません。いずれの従属変数もカテゴリ変数 (名義または順序変数) として扱われないためです。
- [**独立変数の重要度分析**] を選択します。
- [**オプション**] タブをクリックします。

図 4-30
[オプション] タブ



- ユーザー欠損変数を [含める] を選択します。外科的処置を受けていない患者は、「外科療法による合併症」変数にユーザー欠損値があります。これで、そのような患者が確実に分析に含まれます。

- [OK] をクリックします。

警告

図 4-31
警告

次の従属変数は学習用サンプルにおける定数であり、分析 : DOA, 死亡ER から除外されます。

警告表は、変数 doa および der が学習サンプル内では一定であることを示しています。到着時に死亡していた患者または緊急治療室で死亡した患者は、「入院期間」にユーザー欠損値があります。「入院期間」をこの分析のスケール変数として扱っており、スケール変数にユーザー欠損値のあるケースは除外されるため、緊急治療室の後も生存していた患者のみが含まれます。

処理したケースの要約

図 4-32
ケース処理の要約

		度数	パーセント
サンプル	学習	5647	70.6%
	テスト	1570	19.6%
	ホールドアウト	781	9.8%
	有効数	7998	100.0%
	除外数	2002	
	合計	10000	

ケース処理の要約は、5647 個のケースが学習サンプルに、1570 個が検証サンプルに、781 個がホールドアウト サンプルに、それぞれ割り当てられたことを示しています。分析から除外された 2002 個のケースは、病院への搬送中または緊急治療室で死亡した患者です。

ネットワーク情報

図 4-33
ネットワーク情報

入力層	Factors	1	年齢
		2	性別
		3	糖尿病の病歴
		4	血圧
		5	喫煙者
		6	コレステロール
		7	運動
		8	肥満
		9	狭心症の病歴
		10	心筋梗塞の病歴
		11	ニトログリセリンの処方
		12	抗凝固薬の摂取
		13	病院までの時間
		14	EKG の結果
		15	CPK の血液結果
		16	トロポニン T の血液結果
		17	血栓溶解薬
		18	脳内出血
		19	マグネシウム
		20	ジギタリス
		21	ベータ ブロッカー
		22	外科治療
		23	外科療法による合併症
	単位数[a]		63
隠れ層	隠れ層の数		2
	隠れ層1のユニット数[a]		12
	隠れ層2のユニット数[a]		9
	活性化関数		双曲線正接
出力層	Dependent Variables	1	入院期間
		2	治療費
	単位数		2
	Rescaling Method for Scale Dependents		Adjusted Normalized
	活性化関数		双曲線正接
	誤差関数		平方和

a. バイアス ユニットは除外

ネットワーク情報テーブルには、ニューラル ネットワークに関する情報が表示され、指定が正しいことを確認するのに使用できます。ここで、特に次の点に注意する必要があります。

- 入力層の単位数は、因子レベルの合計数です（共変量なし）。
- 2 つの隠れ層が要求され、手続きは最初の隠れ層で 12 単位、2 番目の隠れ層で 9 単位を選択しました。

- 各スケール従属変数に対して、個別の出力単位が作成されます。これらの単位は、出力層に対して双曲線正接のアクティブ化関数を使用する必要のある、調整済み正規化法によって再調整されます。
- 従属変数がスケール変数であるため、平方和の誤差が報告されます。

モデルの要約 (ピボットテーブル 回帰)

図 4-34
モデルの要約

学習	平方和のエラー		91.812
	平均値全体の相対エラー		.083
	スケール依存に対する相対エラー	入院期間	.131
		治療費	.033
	停止規則の使用		減少のない1継続ステップがエラーです[a]
	学習時間		00:00:18.055
テスト	平方和のエラー		26.798
	平均値全体の相対エラー		.088
	スケール依存に対する相対エラー	入院期間	.141
		治療費	.033
ホールドアウト	平均値全体の相対エラー		.099
	スケール依存に対する相対エラー	入院期間	.154
		治療費	.041

a. エラーの計算は、学習サンプルに基づいています。

モデルの要約では、学習結果および最終ネットワークをホールドアウト サンプルに適用した結果に関する情報が表示されます。

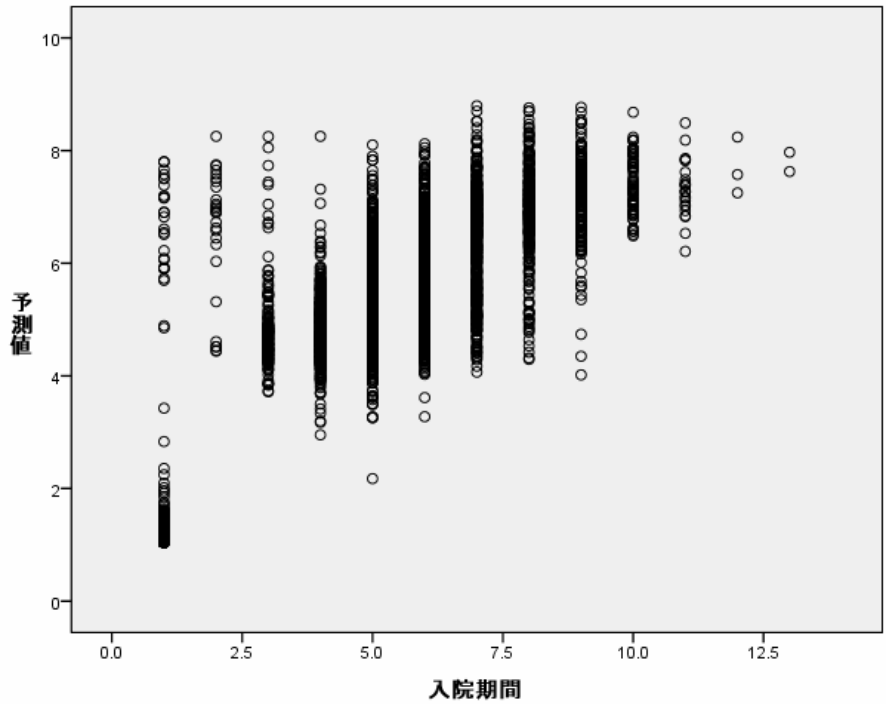
- 出力層にスケール従属変数があるため、平方和の誤差が表示されます。これは、ネットワークが学習中に最小化しようとする誤差関数です。平方和およびその後のすべての誤差の値は、従属変数の再調整された値に対して計算されます。
- 各スケール従属変数の相対誤差は、従属変数の平均値が各ケースの予測値として使用される「帰無仮説」モデルの平方和の誤差に対する、従属変数の平方和誤差の割合です。「入院期間」の予測の方が「治療費」の予測よりも多くの誤差があるように見えます。
- 平均全体誤差は、従属変数の平均値が各ケースの予測値として使用される「帰無仮説」モデルの平方和の誤差に対する、すべての従属変数の平方和の誤差の割合です。この例では、平均全体誤差は相対誤差の平均に近くなっていますが、必ずしもそうなるとは限りません。

  平均全体相対誤差および相対誤差は、学習サンプル、検証用サンプル、およびホールドアウト サンプルにわたってほぼ一定しており、モデルの学習が過度に行われていないこと、ネットワークによってスコアリングされる将来のケースの誤差は、このテーブルで報告された誤差に近くなる、という確信を与えてくれます。

■ アルゴリズムのステップの後で誤差が減少しなかったので、推定アルゴリズムは停止しました。

観測により予測されるグラフ

図 4-35
滞在期間に対する観測により予測されるグラフ

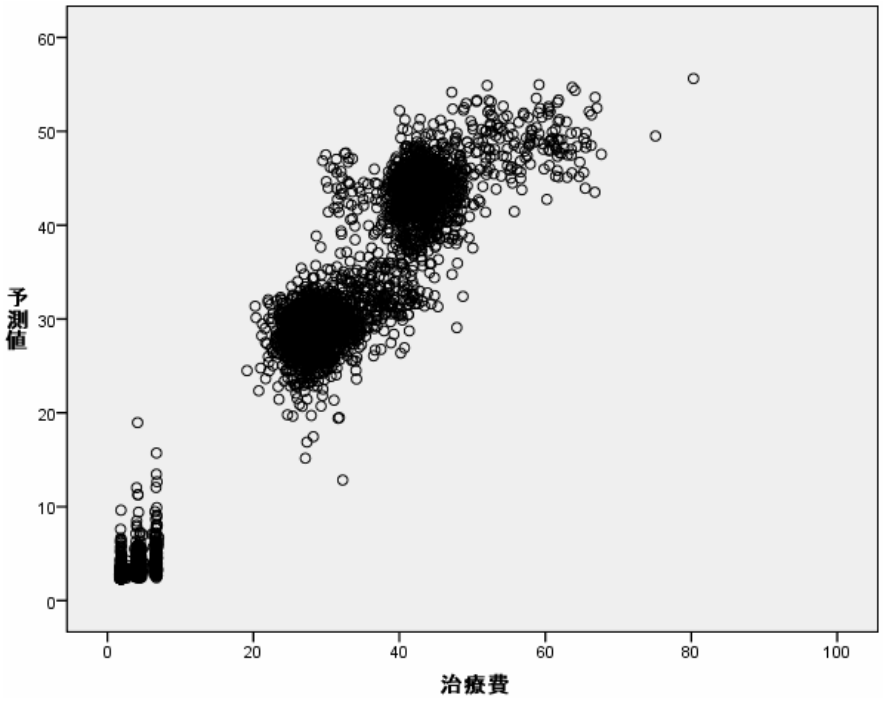


スケール従属変数の場合、観測により予測されるグラフには、学習用と検証用サンプルの組み合わせに対して、 y 軸が予測値で、 x 軸が観測値の散布図が表示されます。理想的には、原点から出発するおよそ 45 度の線に沿って値が点在します。このプロットの点は、「入院期間」の各観測日数で縦の線を形成します。

プロットを見ると、ネットワークは 滞在期間をかなり適切に予測しているように見えます。プロットの一般的な傾向は、5 日間以下の観測された滞在期間の予測は滞在期間を過大評価する傾向があり、6 日以上の観測された滞在期間の予測は滞在期間を過小評価する傾向にあるという点で、理想的な 45 度の線から外れています。

プロットの左下部分の患者のクラスタは、外科手術を受けなかった患者であると考えられます。プロットの左上部分にも患者クラスタがあり、観測された滞在期間は 1 ～ 3 日で、予測値ははるかに大きくなっています。これらのケースは、病院で外科手術後に死亡した患者であると考えられます。

図 4-36
治療費の観測により予測されるグラフ



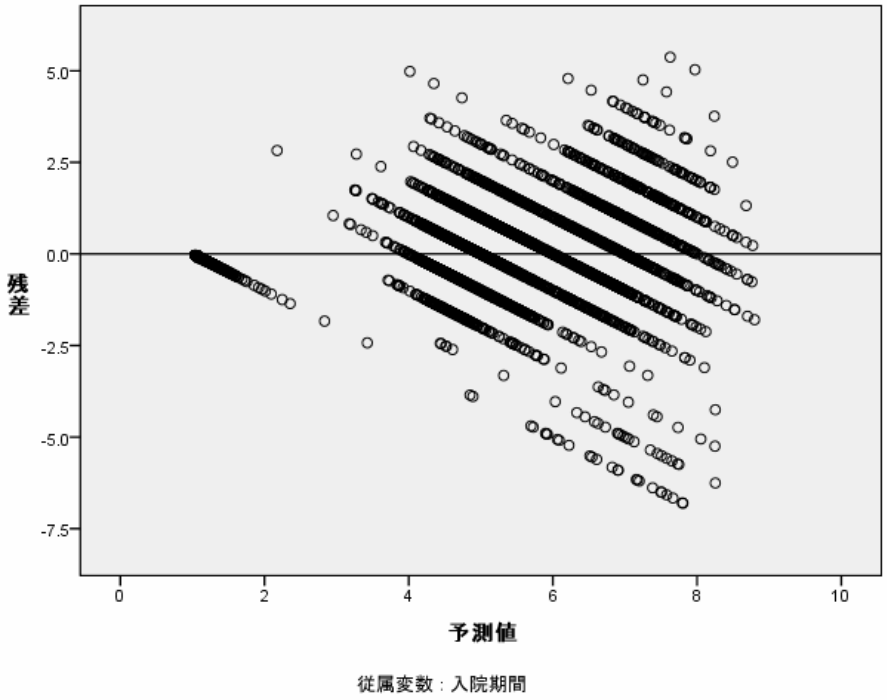
治療費についても、ネットワークは適切に予測しているように見えます。3つの主要な患者クラスタがあるようです。

- 左下は、主に外科手術を受けなかった患者です。これらの患者の治療費は比較的低く、緊急治療室で投与された「血栓熔解薬［血栓熔解薬］」の種類によって異なります。
- 次の患者クラスタの治療費は、約 30,000 ドルです。これらは、経皮的冠動脈形成術（PTCA）を受けた患者です。
- 最後のクラスタの治療費は、40,000 ドルを超えています。これらは、冠動脈バイパス手術（CABG）を受けた患者です。この手術は PTCA よりさらに高額であり、患者の病院での回復期間も長くなり、費用もさらに増加します。

また、ネットワークがあまり適切に予測できていない、治療費が 50,000 ドルを超えるケースも多数あります。これらは、手術中に合併症を併発した患者で、手術費および滞在期間が増加する可能性があります。

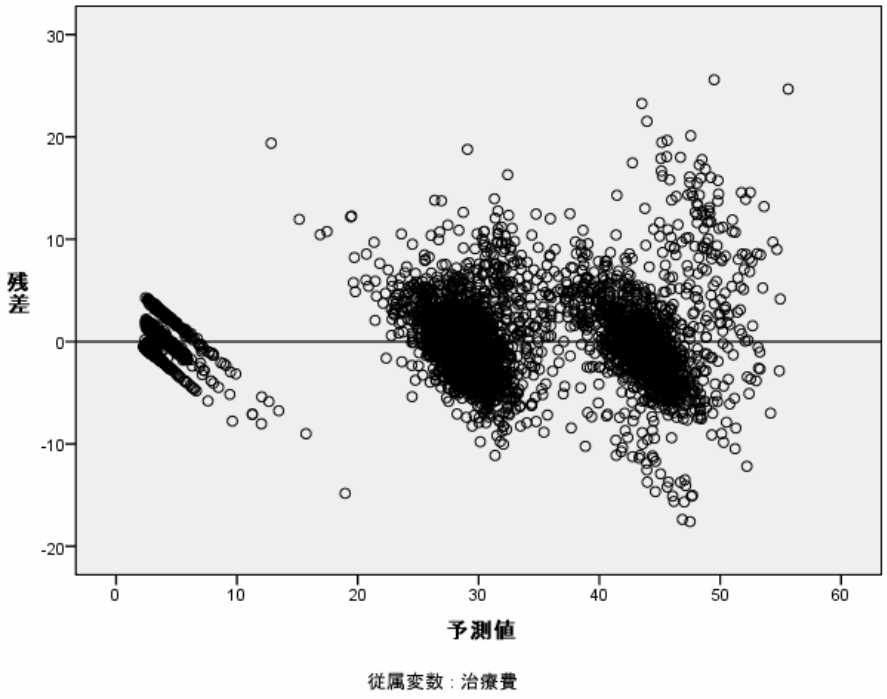
予測による残差グラフ

図 4-37
滞在期間の予測による残差グラフ



予測による残差グラフには、 y 軸が残差（観測値から予測値を引いた値）で x 軸が予測値の散布図が表示されます。このプロットの各対角線は、観測により予測されるグラフの縦線に対応しており、滞在期間の過剰予測から過少予測までの推移を、観測された滞在期間の増加として明確に確認することができます。

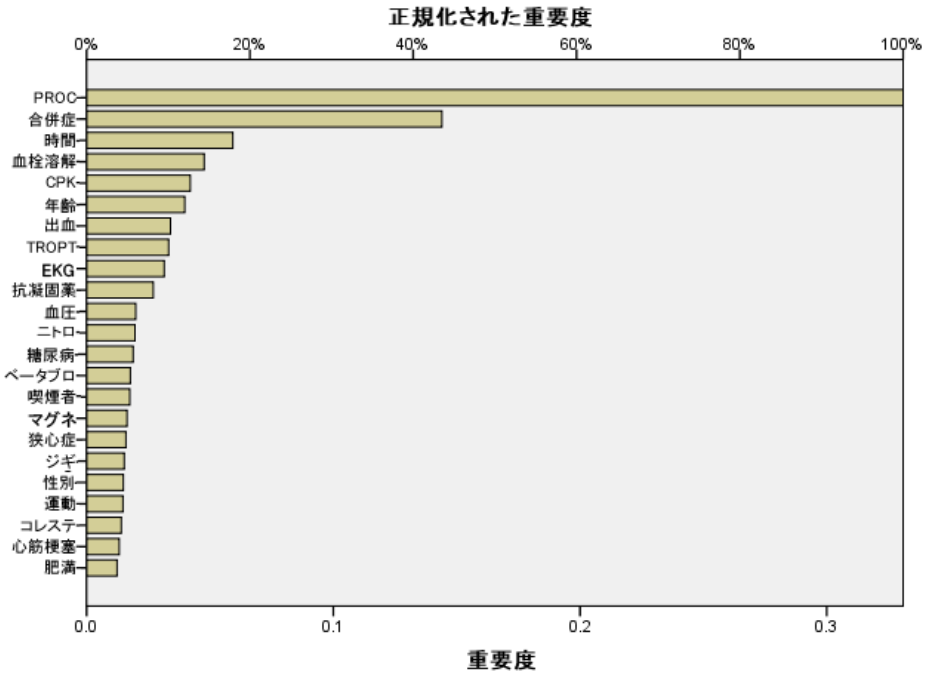
図 4-38
治療費の予測による残差グラフ



同様に、治療費の観測による予測プロットで観測された 3 つの各患者クラスタについても、予測による残差グラフには、治療費の過剰予測から過少予測までの推移が、観測された費用の増加として示されます。CABG 中に合併症を併発した患者も明確に表示されますが、PTCA 中に合併症を併発した患者も容易に確認できます。これらの患者は、PTCA 患者の主グループの上、わずかに右寄りの、 x 軸の 30,000 ドル マーク付近にサブクラスタとして表示されます。

独立変数の重要度

図 4-39
独立変数の重要度グラフ



重要度グラフは、結果の中心的位置を占めているのは施された外科的処置であり、その次に、合併症を併発したかどうか、さらにかなりの差でその他の予測変数であることを示しています。外科的処置の重要度は 治療費のプロットに明確に表示されます。滞在期間ではそれほど明確ではありませんが、滞在期間に及ぼす合併症の影響は、観測された滞在期間が最大の患者では表示されるようです。

集計 (報告書 データ列)

「一般的な」患者の値を予測する場合はネットワークはうまく機能しているように見えますが、手術後に死亡した患者は捕捉していません。これに対処する 1 つの方法として、複数のネットワークを作成することが考えられます。1 つのネットワークでは、患者が生存したかどうかにかかわらず、患者の結果を予測し、別のネットワークでは、患者が生存していることを条件として 治療費と 滞在期間を予測します。次に、これらのネットワークの結果を組み合わせると、より適切な予測が得られることが期待できます。手術中に合併症を併発した患者の、費用と滞在期間の過少予測の問題にも、同様の方法で対処することができます。

推奨参考文献

ニューラル ネットワークおよび多層パーセプトロンの詳細は、次のテキストを参照してください。

Bishop, C. M. 1995. Neural Networks for Pattern Recognition, 3rd ed. Oxford: Oxford University Press.

Fine, T. L. 1999. Feedforward Neural Network Methodology, 3rd ed. New York: Springer-Verlag.

Haykin, S. 1998. Neural Networks: A Comprehensive Foundation, 2nd ed. New York: Macmillan College Publishing.

Ripley, B. D. 1996. Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press.

章 5

放射基底関数

放射基底関数（RBF）手続きは、予測変数の値に基づいて、1 つ以上の従属（目標）変数に対する予測モデルを生成します。

放射基底関数を利用した通信サービス顧客の分類

あるデータ通信プロバイダは、サービス利用パターンに基づいて顧客ベースを分類し、4 つのグループにカテゴリー化しました。顧客がどのグループに属するかを、人口統計データを使って予測できれば、個々の見込み客にあわせてサービスをカスタマイズすることができます。

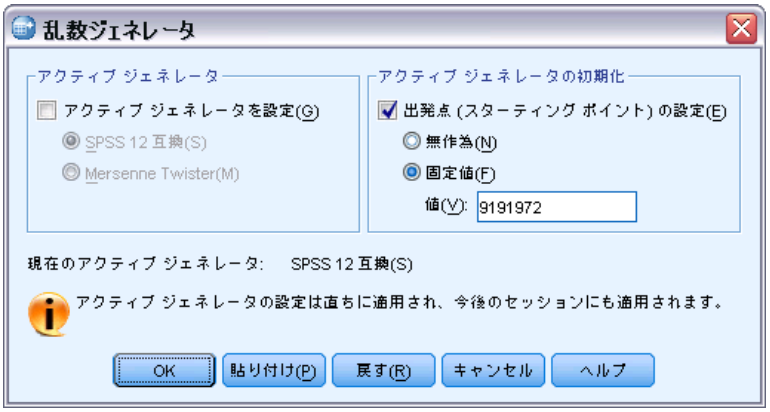
現在の顧客についての情報は、telco.sav に保存されています。詳細は、 A 付録 p.92 サンプル ファイル を参照してください。 この顧客を、放射基底関数手続きを利用して分類します。

分析用データの準備

乱数シードを設定すると、分析を正確に複製できます。

▶ 乱数シードを設定するには、メニューから次の項目を選択します。
変換 > 乱数ジェネレータ

図 5-1
[乱数ジェネレータ] ダイアログ ボックス



▶ [出発点 (スターティング ポイント) の設定] を選択します。

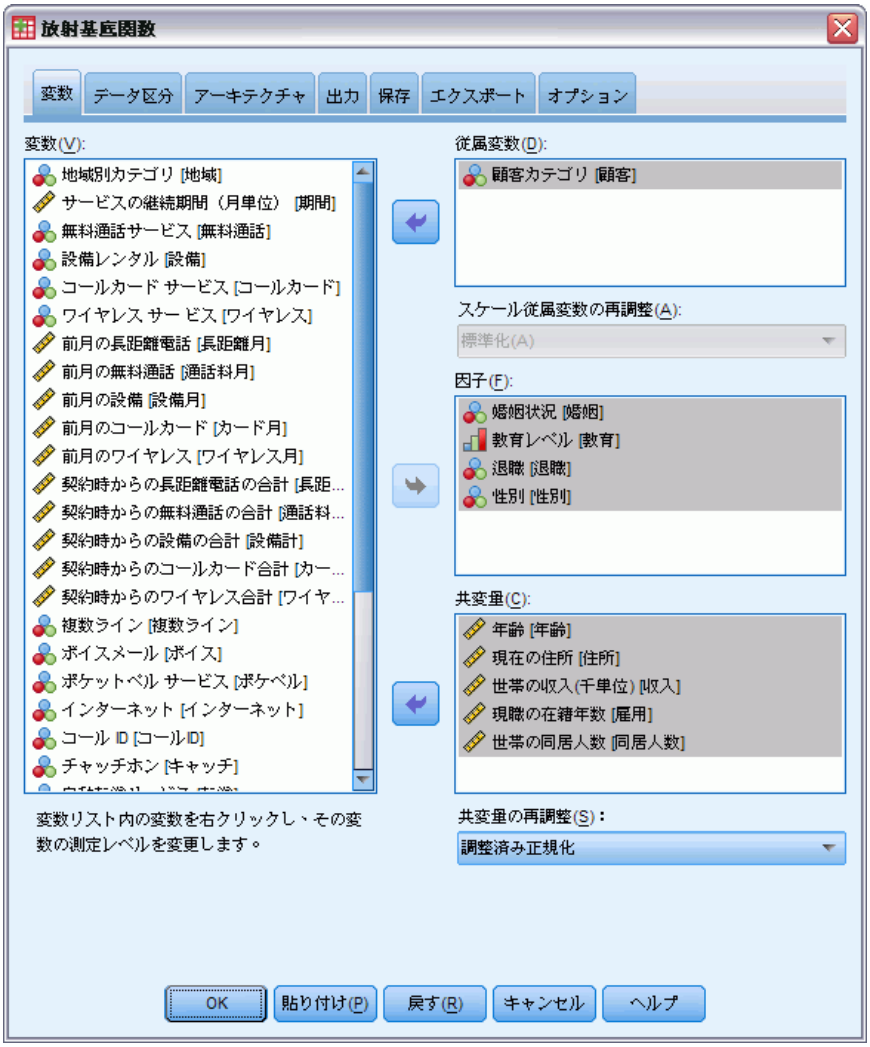
▶ [固定値] を選択し、値として「9191972」と入力します。

▶ [OK] をクリックします。

分析の実行

▶ 放射基底関数分析を実行するには、メニューから次の項目を選択します。
分析 > ニューラル ネットワーク > 放射基底関数...

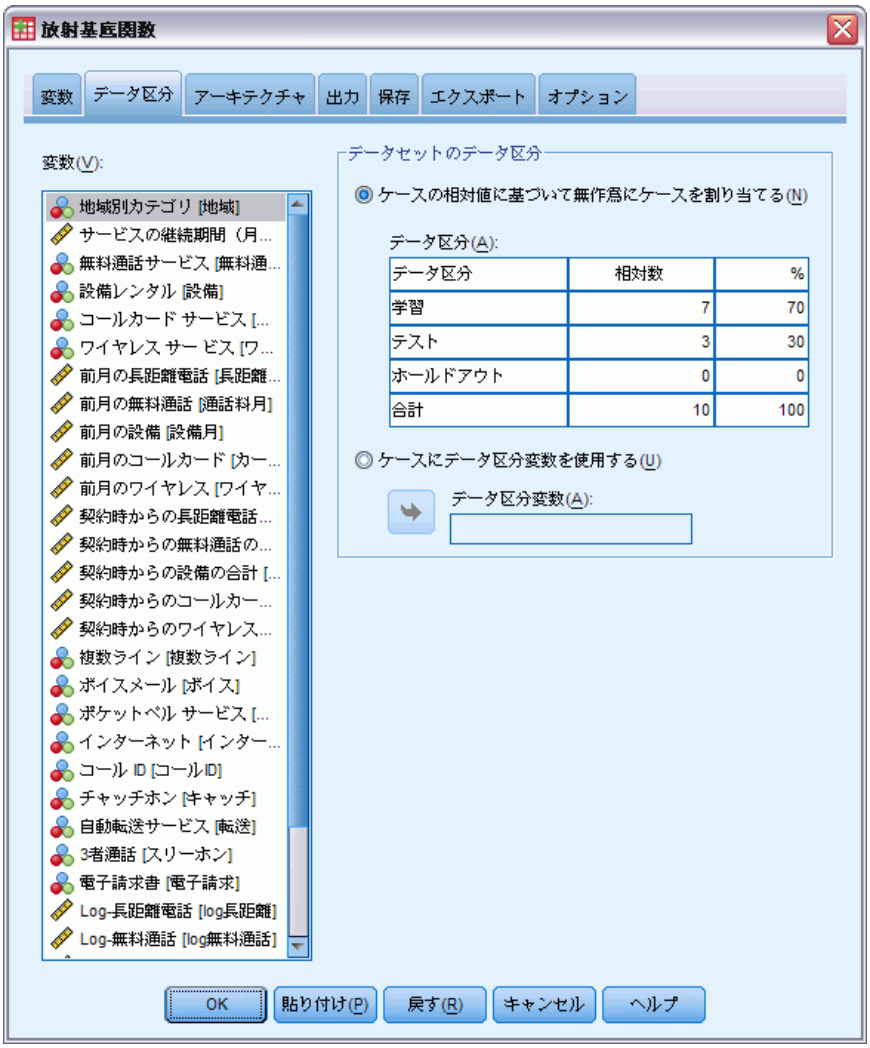
図 5-2
[放射基底関数: 変数] タブ



▶ 従属変数として「顧客カテゴリ［顧客］」を選択します。

- ▶ 因子として、「婚姻状況［婚姻状況］」、「教育のレベル［教育］」、「退職［退職］」、および「性別［性別］」を選択します。
- ▶ 共変量として「年齢［年齢］」から「世帯の人数［世帯人数］」までを選択します。
- ▶ 共変量の再調整の方法として「**調整済み正規化**」を選択します。
- ▶ **[データ区分]** タブをクリックします。

図 5-3
[放射基底関数: データ区分] タブ



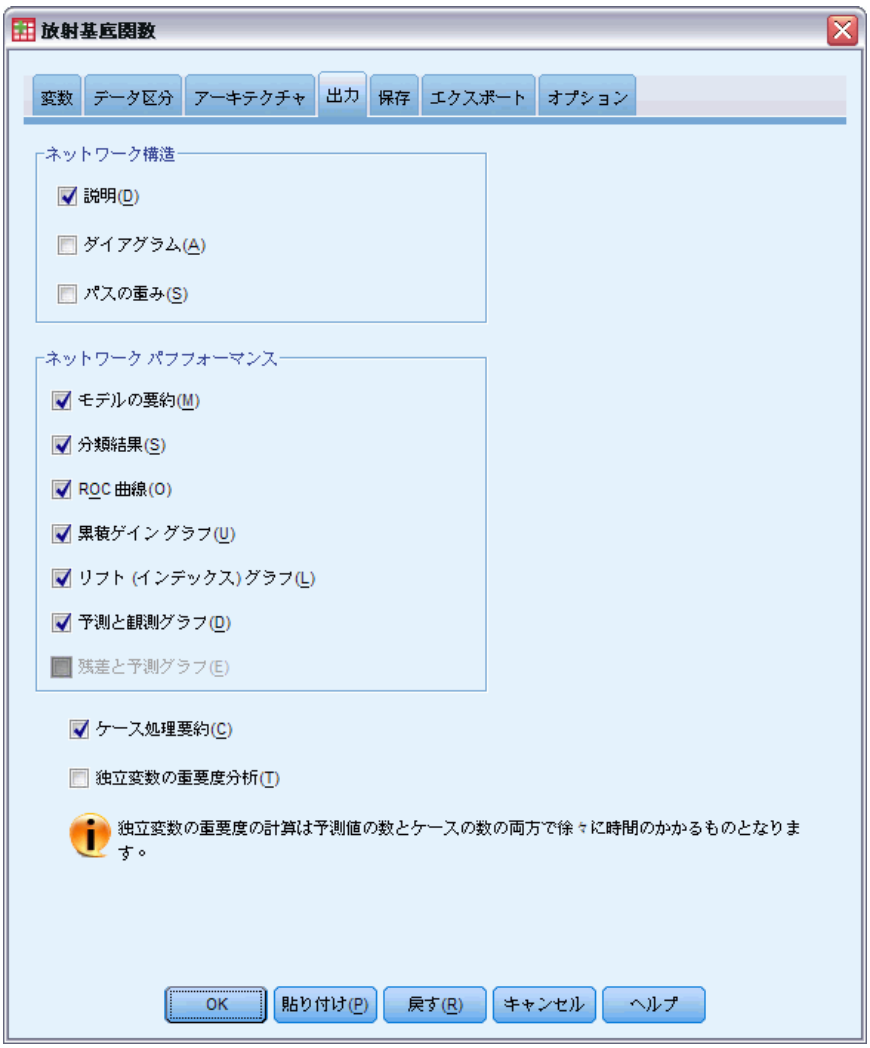
ケースの相対的な数を指定することで、パーセントを指定することが難しい部分的データ区分を簡単に作成できます。データセットの 2/3 を学習サンプルに割り当て、残りのケースの 2/3 を検定に割り当てるものとします。

- 学習サンプルの相対的な数として「6」と入力します。
- 検定サンプルの相対的な数として「2」と入力します。
- ホールドアウト サンプルの相対的な数として「1」と入力します。

  指定した相対ケースは全部で 9 です。6/9 = 2/3 つまり約 66.67% が学習サンプルに、2/9 つまり約 22.22% が検定に、1/9 つまり約 11.11% がホールドアウト サンプルに、それぞれ割り当てられます。

- [出力] タブをクリックします。

図 5-4
[放射基底関数: 出力] タブ

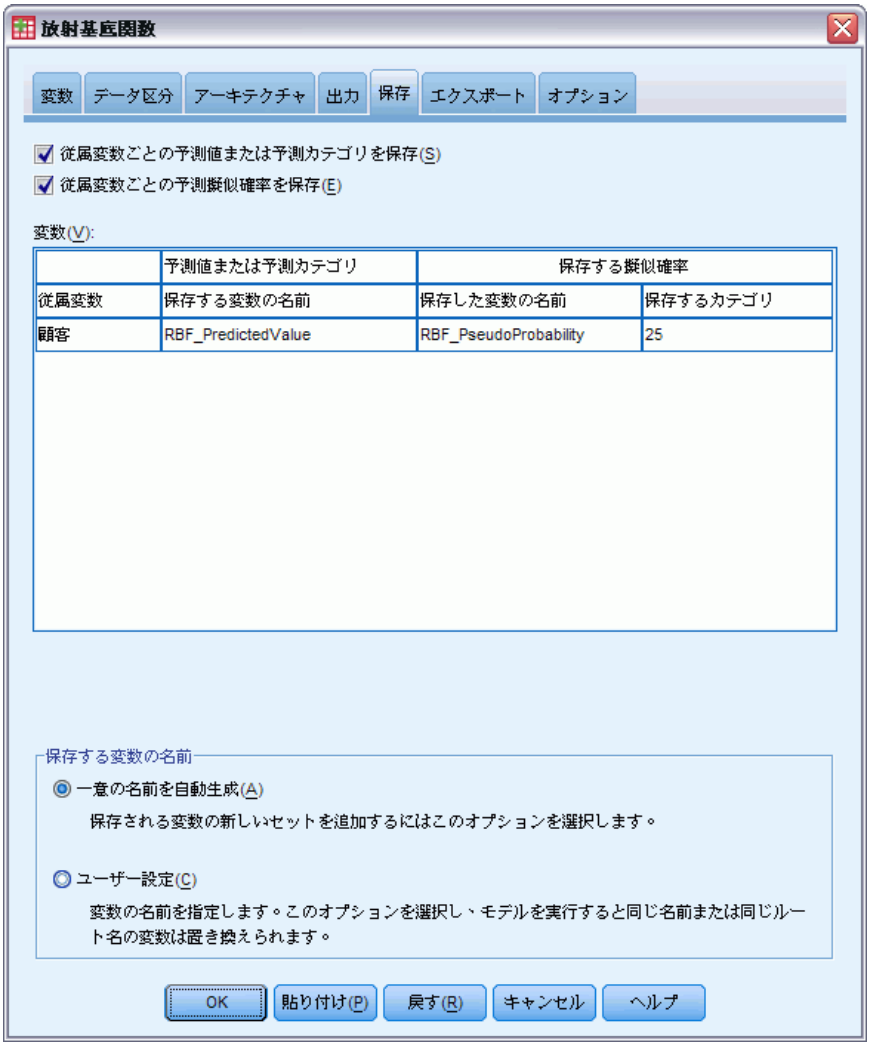


- [ネットワーク構造] グループの [ダイアグラム] の選択を解除します。

▶ ［ネットワーク パフォーマンス］グループで、[ROC 曲線]、[累積ゲイン グラフ]、[リフト (インデックス) グラフ]、および［残差と予測グラフ］を選択します。

▶ ［保存］タブをクリックします。

図 5-5
[放射基底関数: 保存] タブ



▶ ［従属変数ごとの予測値または予測カテゴリを保存］および［各従属変数の予測される疑似確率を保存する］を選択します。

▶ ［OK］をクリックします。

処理したケースの要約

図 5-6
ケース処理の要約

		度数	パーセント
サンプル	学習	665	66.5%
	テスト	224	22.4%
	ホールドアウト	111	11.1%
	有効数	1000	100.0%
	除外数	0	
	合計	1000	

ケース処理の要約は、665 個のケースが学習サンプルに、224 個が検証サンプルに、111 個がホールドアウト サンプルに、それぞれ割り当てられたことを示しています。分析から除外されたケースはありません。

ネットワーク情報

図 5-7
ネットワーク情報

入力層	Factors	1	婚姻状況
		2	教育レベル
		3	退職
		4	性別
	Covariates	1	年齢
		2	現在の住所
		3	世帯の収入(千単位)
		4	現職の在籍年数
		5	世帯の同居人数
	単位数		16
	Rescaling Method for Covariates		Adjusted Normalized
隠れ層	単位数		9[a]
	活性化関数		Softmax
出力層	Dependent Variables	1	顧客カテゴリ
	単位数		4
	活性化関数		単位
	誤差関数		平方和

a. テスト データの基準による決定 : 隠れ単位の「最高の」数値は、テスト データの最小エラーとなります。

ネットワーク情報テーブルには、ニューラル ネットワークに関する情報が表示され、指定が正しいことを確認するのに使用できます。ここで、特に次の点に注意する必要があります。

- 入力層の単位数は、共変量の数に因子レベルの合計数を加えたものです。「婚姻状況」、「教育のレベル」、「退職」、および「性別」のカテゴリごとに個別の単位が作成され、多くのモデル手続きで一般的であるように、「冗長」と見なされるカテゴリはありません。

- 同様に、「顧客カテゴリ」のカテゴリごとに個別の出力単位が作成され、出力層の単位の数は全部で 4 です。
- 共変量は、調整済み正規化法を使用して再調整されます。
- 自動アーキテクチャ選択は、隠れ層で 9 単位を選択しています。
- 他のすべてのネットワーク情報は、手続きのデフォルトです。

モデルの要約 (ピボットテーブル 回帰)

図 5-8
モデルの要約

学習	平方和のエラー	235.969
	誤った予測値の割合	61.8%
	学習時間	2.72
テスト	平方和のエラー	80851[a]
	誤った予測値の割合	62.9%
ホールドアウト	誤った予測値の割合	59.5%

従属変数 : 顧客カテゴリ

a. 隠れ単位の数は、学習データ基準に基づいて決定されます。隠れ単位の「最高の」数値は、テスト データの最小エラーとなります。

モデルの要約では、学習結果、検定結果、および最終ネットワークをホールドアウト サンプルに適用した結果に関する情報が表示されます。

- RBF ネットワークに対しては常に使用されるため、平方和の誤差が表示されます。これは、ネットワークが学習中および検定中に最小化しようとする誤差関数です。
- 不正な予測のパーセントは、分類表から取得されます。詳細については、該当するトピックで説明します。

分類

図 5-9
分類

		予測				
サンプル	観測	ベーシック サービス	E-サービス	プラス サービス	トータル サービス	Percent Correct
学習	ベーシック サービス	64	0	66	45	36.6%
	E-サービス	22	1	57	61	.7%
	プラス サービス	47	0	104	34	56.2%
	トータル サービス	29	1	49	85	51.8%
	Overall Percent	24.4%	.3%	41.5%	33.8%	38.2%
テスト	ベーシック サービス	18	0	26	15	30.5%
	E-サービス	15	0	16	22	.0%
	プラス サービス	11	0	39	15	60.0%
	トータル サービス	4	0	17	26	55.3%
	Overall Percent	21.4%	.0%	43.8%	34.8%	37.1%
ホールドアウト	ベーシック サービス	11	0	11	10	34.4%
	E-サービス	4	0	9	10	.0%
	プラス サービス	10	0	19	2	61.3%
	トータル サービス	5	0	5	15	60.0%
	Overall Percent	27.0%	.0%	39.6%	33.3%	40.5%

従属変数：顧客カテゴリ

分類テーブルでは、実際にネットワークを使用した結果が表示されます。ケースごとの予測応答は、予測される疑似確率が最も高いカテゴリです。

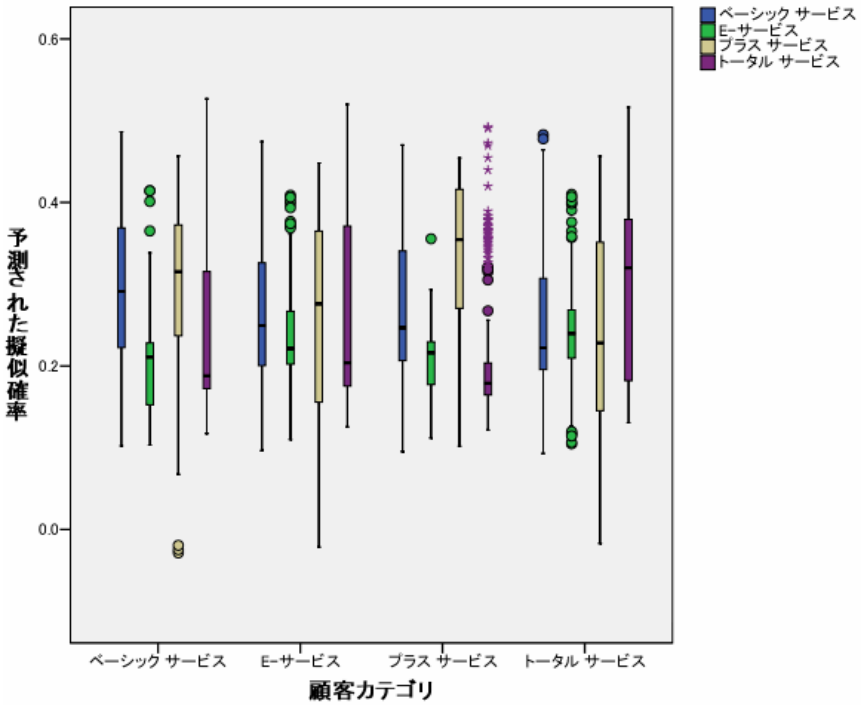
- 対角上のセルが正しい予測値です。
- 対角線から外れたセルは不正な予測値です。

「帰無仮説」モデル（つまり予測変数のないモデル）は、与えられた観測データの全顧客を、最頻グループ、つまり「プラスサービス」に分類します。したがって、「帰無仮説」モデルは 281/1000 = 28.1% 正しいということになります。RBF ネットワークは、10.1% の顧客増、つまり、38.2% の顧客を獲得します。このモデルは特に、「プラスサービス」 および「トータルサービス」の顧客の特定に優れています。しかし、このモデルでは、「E-サービス」顧客の分類がまったくうまくいきません。この種の顧客を分離するには、別の予測変数を見つける必要があります。あるいは、これらの顧客は「プラスサービス」および「トータルサービス」の顧客として誤って分類されることが最も多いので、通常であれば「E-サービス」カテゴリに分類される潜在顧客に対し、単により高いサービスの販売を試みることもできます。

モデルを作成するために使われたケースに基づく分類は、分類率が誇張されるので、「楽観的」になりすぎる傾向があります。ホールドアウト サンプルは、モデルの検証に役立ちます。この場合は、ケースの 40.2% がモデルによって正しく分類されました。ホールドアウト サンプルは比較的少数ですが、これは、実際にはモデルの分類の約 2/5 が正しいということを示しています。

観測により予測されるグラフ

図 5-10
観測により予測されるグラフ



カテゴリ従属変数の場合、観測により予測されるグラフには、学習サンプルと検証サンプルの組み合わせに対する予測された疑似確率のクラスタ箱ひげ図が表示されます。x 軸は観測応答カテゴリに対応し、凡例は予測カテゴリに対応します。したがって、次のようになります。

- 左端の箱ひげ図は、観測されたカテゴリが「ベーシックサービス」であるケースの場合の、カテゴリ「ベーシックサービス」の予測疑似確率を表します。
- 右側の次の箱ひげ図は、観測されたカテゴリが「ベーシックサービス」であるケースの場合の、カテゴリ「E-サービス」の予測疑似確率を表します。
- 3 番目の箱ひげ図は、観測されたカテゴリが「ベーシックサービス」であるケースの場合の、カテゴリ「プラスサービス」の予測疑似確率を表します。分類テーブルで見たように、「ベーシックサービス」の顧客は、「ベーシックサービス」として正しく分類されるのと同じくらい

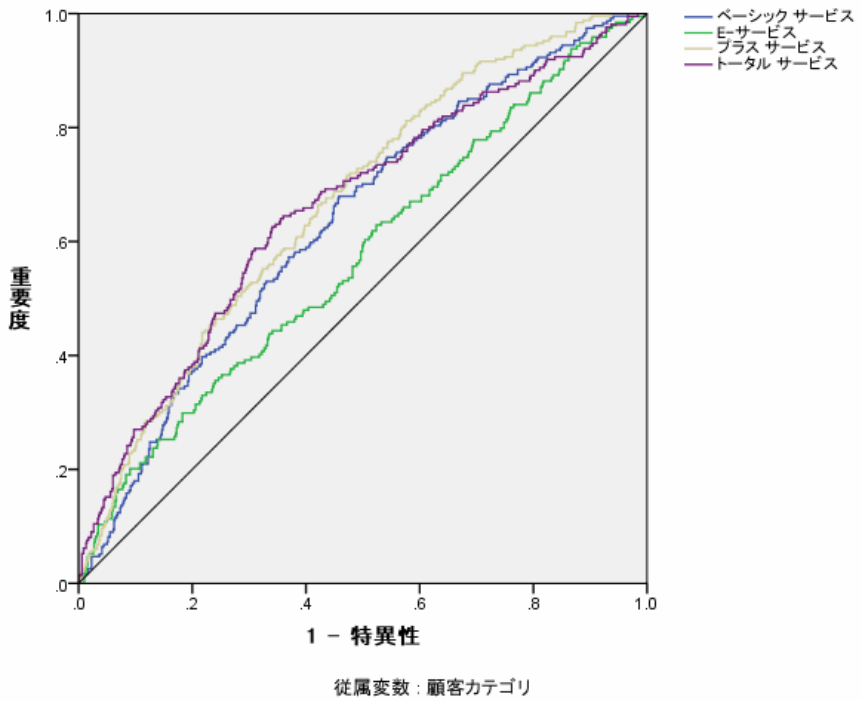
に、「プラスサービス」として誤って分類されました。したがって、この箱ひげ図は、左端のものとほぼ等しくなります。

- 4 番目の箱ひげ図は、観測されたカテゴリが「ベーシックサービス」であるケースの場合の、カテゴリ「トータルサービス」の予測疑似確率を表します。

目標変数にはカテゴリが 3 つ以上あるので、最初の 4 つの箱ひげ図は、0.5 の水平線についても、他のどのような線についても、対称にはなりません。結果として、ある箱ひげ図に含まれるケースの部分から、別の箱ひげ図でそのケースに対応する部分を判別するのは不可能なので、3 つ以上のカテゴリを含む目標のこのプロットを解釈するのは困難な場合があります。

ROC 曲線

図 5-11
ROC 曲線



ROC 曲線を見ると、可能なすべての分類打ち切りについての、**特異性**による**感度**がわかります。ここで示すグラフでは、目標変数のカテゴリごとに 1 つずつ、4 本の曲線が示されています。

このグラフは、学習サンプルと検証サンプルの組み合わせに基づいていることに注意してください。ホールドアウト サンプルの ROC グラフを生成するには、データ区分変数でファイルを分割し、予測された疑似確率に対して ROC 曲線手続きを実行します。

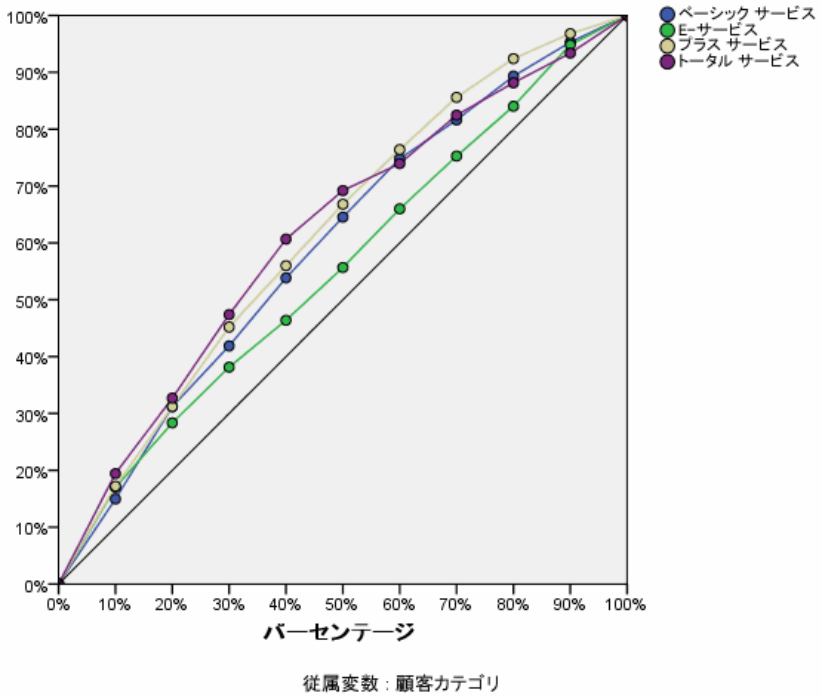
図 5-12
曲線の下側の面積

		面グラフ
顧客カテゴリ	ベーシック サービス	.635
	E-サービス	.573
	プラス サービス	.668
	トータル サービス	.659

曲線の下側の面積は ROC 曲線の数値的な要約であり、テーブルの値は、各カテゴリについて、そのカテゴリの予測された疑似確率が、そのカテゴリ以外で無作為に選択されたケースの場合より、そのカテゴリで無作為に選択されたケースの場合の方が高いことを表しています。たとえば、「プラスサービス」で無作為に選択した顧客と、「ベーシックサービス」、「E-サービス」、または「トータルサービス」で無作為に選択した顧客では、「プラスサービス」の顧客の方がデフォルトのモデル予測疑似確率が高い確率は 0.668 です。

累積ゲイン グラフとリフト図表

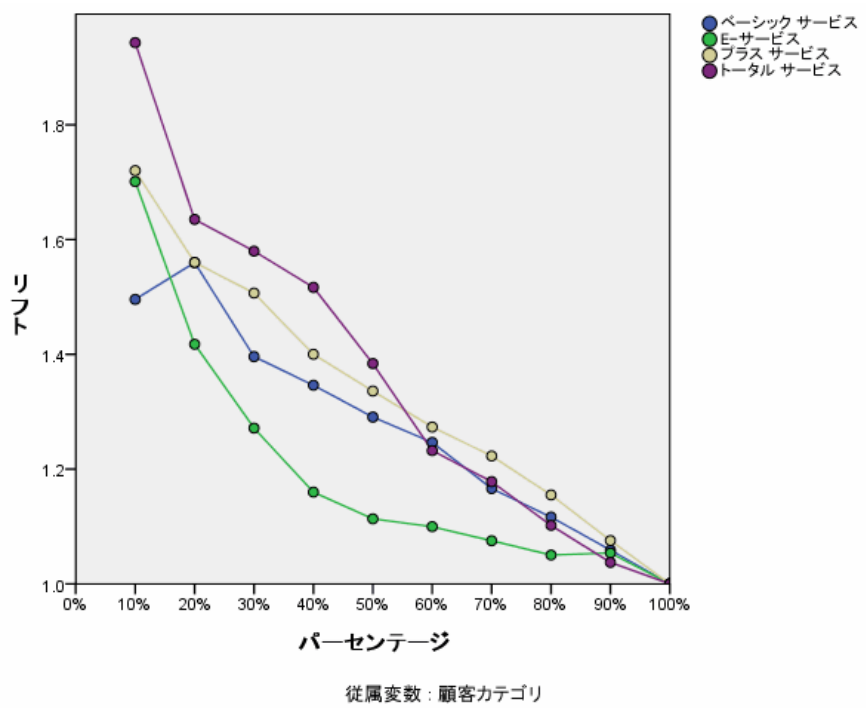
図 5-13
累積ゲイン グラフ



累積ゲイン グラフは、ケースの合計数のパーセントを目標にすることで、特定のカテゴリ「ゲイン」のケースの総数のパーセントを示します。たとえば、「トータルサービス」カテゴリの曲線の最初のポイントはだいたい (10%, 20%) であり、これは、ネットワークでデータセットをスコアリングし、「トータルサービス」の予測された疑似確率ですべてのケースをソートした場合、上位 10% が、実際にカテゴリ「トータルサービス」である全ケースの約 20% を含むと期待することを意味します。同様に、上位 20% は債務不履行者の約 30% を含み、上位 30% は債務不履行者の約 50% を含みます。スコアリングされたデータセットの 100% を選択すると、すべての債務不履行者がデータセットに含まれます。

対角線は「ベースライン」曲線です。スコアリングされたデータセットから無作為にケースの 10% を選択すると、実際に特定のカテゴリである全ケースの約 10% を「ゲインする」ことが期待されます。曲線がベースラインより上になるほど、ゲインが大きくなります。

図 5-14
リフト図表



リフト図表は、累積ゲイン グラフから導かれます。 y 軸の値は、各曲線の累積ゲインの、ベースラインに対する比率に対応します。したがって、カテゴリ「トータルサービス」の 10% におけるリフトは、約 20%/10% = 2.0 です。累積ゲイン グラフの情報を別の方法で見ることができます。

注: 累積ゲイン グラフとリフト図表は、学習サンプルと検証サンプルの組み合わせに基づいています。

推奨参考文献

放射基底関数 の詳細は、次の文献を参照してください。

Bishop, C. M. 1995. Neural Networks for Pattern Recognition, 3rd ed. Oxford: Oxford University Press.

Fine, T. L. 1999. Feedforward Neural Network Methodology, 3rd ed. New York: Springer-Verlag.

Haykin, S. 1998. Neural Networks: A Comprehensive Foundation, 2nd ed. New York: Macmillan College Publishing.

Ripley, B. D. 1996. Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press.

Tao, K. K. 1993. A closer look at the radial basis function (RBF) networks. In: Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers, A. Singh, ed. Los Alamitos, Calif.: IEEE Comput. Soc. Press, 401-405.

Uykan, Z., C. Guzelis, M. E. Celebi, および H. N. Koivo. 2000. Analysis of input-output clustering for determining centers of RBFN. IEEE Transactions on Neural Networks, 11, 851-858.

付録

A

サンプル ファイル

製品とともにインストールされるサンプル ファイルは、インストール ディレクトリの Samples サブディレクトリにあります。［サンプル］サブディレクトリ内に次の各言語の別のフォルダがあります。英語、フランス語、ドイツ語、イタリア語、日本語、韓国語、ポーランド語、ロシア語、簡体字中国語、スペイン語、そして繁体中国語です。

すべてのサンプル ファイルが、すべての言語で使用できるわけではありません。サンプル ファイルがある言語で使用できない場合、その言語のフォルダには、サンプル ファイルの英語バージョンが含まれています。

説明

以下は、このドキュメントのさまざまな例で使用されているサンプル ファイルの簡単な説明です。

- **accidents.sav。** 与えられた地域での自動車事故の危険因子を年齢および性別ごとに調べている保険会社に関する架空のデータ ファイルです。各ケースが、年齢カテゴリと性別のクロス分類に対応します。
- **adl.sav。**脳卒中患者に提案される治療の効果を特定するための取り組みに関する架空のデータ ファイルです。医師団は、女性の脳卒中患者たちを、2 つのグループのいずれかにランダムに割り当てました。一方のグループは標準的な理学療法を受け、もう一方のグループは感情面の治療も追加で受けました。治療の 3 か月後に、各患者が日常生活の一般的な行動をどの程度とることができるかを、順序変数として得点付けしました。
- **advert.sav。**広告費とその売上成果の関係を調べるための小売業者の取り組みに関する架空のデータ ファイルです。この小売業者は、そのために、過去の売上と、それに関係する広告費のデータを収集しました。
- **aflatoxin.sav。**収穫物によって濃度が大きく異なる毒物であるアフラトキシンを、トウモロコシの収穫物に関して検定することに関する架空のデータ ファイルです。ある穀物加工業者は、8 つそれぞれの収穫物から 16 のサンプルを受け取って、10 億分の 1 単位でアフラトキシン レベルを測定しました。
- **anorectic.sav。**拒食行動または過食行動の標準的な症状の特定を目指して、調査員 (Van der Ham, Meulman, Van Strien, および Van Engeland, 1997) が、摂食障害を持つ大人 55 人の調査を行いました。各患者が 4 年間で 4 回診察を受けたので、観測値は合計で 220 になりました。観

測値ごとに、16 種類の症状に関して患者の得点が記録されました。患者 71 (2 回目)、患者 76 (2 回目)、患者 47 (3 回目) の症状の得点が見つからなかったので、残っている 217 回分の観測値が有効です。

- **bankloan.sav。**債務不履行率を低減させるための銀行の取り組みに関する架空のデータ ファイルです。このファイルには、過去の顧客および見込み客 850 人に関する財務情報と人口統計情報が含まれています。最初の 700 ケースは、以前に貸付を行った顧客です。残りの 150 ケースは見込み顧客で、これらの顧客に関して銀行は信用リスクの良し悪しを分類する必要があります。
- **bankloan_binning.sav。**過去の顧客 5,000 人に関する財務情報と人口統計情報を含む架空のデータ ファイルです。
- **behavior.sav。**52 人の学生に 15 の状況と 15 の行動の組み合わせについて、0 = 「非常に適切」から 9 = 「非常に不適切」までの 10 段階でランク付けするよう依頼した研究があります (Price および Bouffard, 1974)。個人間の平均を取ったため、値は非類似度としてみなされます。
- **behavior_ini.sav。**このデータ ファイルには、behavior.sav の 2 次元の解の初期布置が含まれています。
- **brakes.sav。**高性能自動車のディスク ブレーキを生産している工場での品質管理に関する架空のデータ ファイルです。このデータ ファイルには、8 台の機械で生産した 16 個のディスクの直径測定値が含まれています。ブレーキの目標の直径は 322 ミリメートルです。
- **breakfast.sav。**21 人の Wharton School MBA の学生およびその配偶者に、15 種類の朝食を好みの順に (1 = 「最も好き」から 15 = 「最も嫌い」まで) ランク付けするよう依頼した研究があります (Green および Rao, 1972)。調査対象者の嗜好は、「すべて」から「スナックとドリンクのみ」まで、6 つの異なるシナリオに基づいて記録されました。
- **breakfast-overall.sav。**このデータ ファイルには、最初のシナリオ (「すべて」) のみの朝食の好みが含まれています。
- **broadband_1.sav。**全国規模のブロードバンド サービスの地域ごとの契約者数を含む架空のデータ ファイルです。このデータ ファイルには、85 地域の月々の契約者数が 4 年間分含まれています。
- **broadband_2.sav。**このデータ ファイルは broadband_1.sav と同じですが、データが 3 か月分追加されています。
- **car_insurance_claims.sav。**他の場所 (McCullagh および Nelder, 1989) で表示および分析される、自動車の損害請求に関するデータセットです。逆リンク関数を使用して従属変数の平均値を保険契約者の年齢、車種、製造年の線型結合と関連付けることにより、平均請求数はガンマ分布としてモデリングできます。申請された請求の数は、尺度重み付けとして使用できます。
- **car_sales.sav。**このデータ ファイルには、自動車のさまざまな車種やモデルの架空の売上推定値、定価、仕様が含まれています。定価と仕様はそれぞれ、edmunds.com と製造元のサイトから入手しました。

- **car_sales_uprepared.sav。** 変換したバージョンのフィールドを含まない car_sales.sav の修正したバージョンです。
- **carpet.sav。**一般的な例（Green および Wind, 1973）としては、新しいカーペット専用洗剤を市販することに関心のある企業が消費者の嗜好に関する 5 種類の因子（パッケージのデザイン、ブランド名、価格、サービスシール、料金の払い戻し）の影響について調べたい場合があります。パッケージのデザインには、3 つの因子レベルがあります。それぞれ塗布用ブラシの位置が異なります。また、3 つのブランド名（K2R、Glory、および Bissell）、3 つの価格水準があり、最後の 2 つの因子のそれぞれに対しては 2 つのレベル（「なし」または「あり」）があります。10 人の消費者が、これらの因子により定義された 22 個のプロファイルに順位を付けます。変数「嗜好」には、各プロファイルの平均順位の序列が含まれています。順位が低いほど、嗜好度は高くなります。この変数には、各プロファイルの嗜好測定値がすべて反映されます。
- **carpet_prefs.sav。**このデータ ファイルは carpet.sav と同じ例に基づいていますが、10 人の消費者それぞれから収集した実際のランキングが含まれています。消費者は、22 種類の製品プロファイルを、一番好きなものから一番嫌いなものまで順位付けすることを依頼されています。変数 PREF1 から PREF22 には、carpet_plan.sav で定義されている、関連するプロファイルの ID が含まれています。
- **catalog.sav。**このデータ ファイルには、あるカタログ会社が販売した 3 つの製品の、架空の月間売上高が含まれています。5 つの予測変数のデータも含まれています。
- **catalog_seasfac.sav。**このデータ ファイルは catalog.sav と同じですが、季節性の分解手続きとそれに付随する日付変数から計算した一連の季節因子が追加されています。
- **cellular.sav。**解約率を削減するための携帯電話会社の取り組みに関する架空のデータ ファイルです。解約の傾向スコアは、0 ～ 100 の範囲でアカウントに適用されます。スコアリングが 50 以上のアカウントはプロバイダの変更を考えている場合があります。
- **ceramics.sav。**新しい上質の合金に標準的な合金より高い耐熱性があるかどうかを特定するための、ある製造業者の取り組みに関する架空のデータ ファイルです。各ケースが 1 つの合金の別々のテストを表し、軸受けの耐熱温度が記録されます。
- **cereal.sav。**880 人を対象に、朝食の好みについて、年齢、性別、婚姻状況、ライフスタイルが活動的かどうか（週 2 回以上運動するか）を含めて調査した、架空のデータ ファイルです。各ケースが別々の回答者を表します。
- **clothing_defects.sav。**ある衣料品工場での品質管理工程に関する架空のデータ ファイルです。工場で生産される各ロットから、調査員が衣料品のサンプルを取り出し、不良品の数を数えます。

- **coffee.sav。**このデータ ファイルは、6 つのアイスコーヒー ブランド (Kennedy, Riquier, および Sharp, 1996) について受けた印象に関連しています。回答者は、アイス コーヒーに対する 23 の各印象属性に対して、その属性が言い表していると思われるすべてのブランドを選択しました。機密保持のため、6 つのブランドを AA、BB、CC、DD、EE、および FF で表しています。
- **contacts.sav。**企業のコンピュータ営業グループの担当者リストに関する架空のデータ ファイルです。各担当者は、所属する会社の部門および会社のランクによって分類されています。また、最新の販売金額、最後の販売以降の経過時間、担当者の会社の規模も記録されています。
- **creditpromo.sav。**最近のクレジット カード プロモーションの有効性を評価するための、あるデパートの取り組みに関する架空のデータ ファイルです。このために、500 人のカード所有者がランダムに選択されました。そのうち半分には、今後 3 か月間の買い物に関して利率を下げることをプロモーションする広告を送付しました。残り半分には、通常どおりの定期的な広告を送付しました。
- **customer_dbase.sav。**自社のデータ ウェアハウスにある情報を使用して、反応がありそうな顧客に対して特典を提供するための、ある会社の取り組みに関する架空のデータ ファイルです。顧客ベースのサブセットをランダムに選択して特典を提供し、顧客の反応が記録されています。
- **customer_information.sav。**名前や住所など、顧客の連絡先情報を含む架空のデータ ファイルです。
- **customer_subset.sav。**customer_dbase.sav の 80 件のケースのサブセット。
- **debate.sav。**政治討論の出席者に対して行った調査の、討論の前後それぞれの回答に関する架空のデータ ファイルです。各ケースが別々の回答者に対応します。
- **debate_aggregate.sav。**debate.sav 内の回答を集計する、架空のデータ ファイルです。各ケースが、討論前後の好みのクロス分類に対応しています。
- **demo.sav。**月々の特典を送付することを目的とした、購入顧客のデータベースに関する架空のデータ ファイルです。顧客が特典に反応したかどうかが、さまざまな人口統計情報と共に記録されています。
- **demo_cs_1.sav。**調査情報のデータベースをコンパイルするための、ある会社の取り組みの最初のステップに関する架空のデータ ファイルです。各ケースが別々の都市に対応し、地域、地方、地区、および都市の ID が記録されています。
- **demo_cs_2.sav。**調査情報のデータベースをコンパイルするための、ある会社の取り組みの第 2 のステップに関する架空のデータ ファイルです。各ケースが、最初のステップで選択した都市の別々の世帯単位に対応し、地域、地方、地区、都市、区画、および単位の ID が記録されます。計画の最初の 2 つの段階からの抽出情報も含まれています。

- **demo_cs.sav。**コンプレックス サンプル計画を使用して収集された調査情報を含む架空のデータ ファイルです。各ケースが別々の世帯単位に対応し、さまざまな人口統計情報および抽出情報が記録されています。
- **dmdata.sav。** これは、ダイレクト マーケティング企業の人口統計情報および購入情報を含む架空のデータです。dmdata2.sav には、テスト メールを受け取った連絡先のサブセットの情報を含み、dmdata3.sav には、テスト メールを受け取らなかった残りの連絡先に関する情報を含みます。
- **dietstudy.sav。**この架空のデータ ファイルには、”Stillman diet” (Rickman, Mitchell, Dingman, および Dalen, 1974) の研究結果が含まれています。各ケースが別々の被験者に対応し、被験者のダイエット前後の体重（ポンド単位）と、トリグルセリド レベル（mg/100 ml 単位）が記録されています。
- **dvdplayer.sav。**新しい DVD プレーヤーの開発に関する架空のデータ ファイルです。プロトタイプを使用して、マーケティング チームはフォーカス グループ データを収集しました。各ケースが別々の調査対象ユーザーに対応し、ユーザーの人口統計情報と、プロトタイプに関する質問への回答が記録されています。
- **german_credit.sav。**このデータ ファイルは、カリフォルニア大学アーバイン校の Repository of Machine Learning Databases (Blake および Merz, 1998) にある “German credit” データセットから取ったものです。
- **grocery_1month.sav。**この架空のデータ ファイルは、grocery_coupons.sav データ ファイルの週ごとの購入を「ロールアップ」して、各ケースが別々の顧客に対応するようにしたものです。その結果、週ごとに変わっていた変数の一部が表示されなくなり、買物の総額が、調査を行った 4 週間の買物額の合計になっています。
- **grocery_coupons.sav。**顧客の購買習慣に関心を持っている食料雑貨店チェーンが収集した調査データを含む架空のデータ ファイルです。各顧客を 4 週間に渡って追跡し、各ケースが別々の顧客の週に対応しています。その週に食料品に費やした金額も含め、顧客がいつどこで買物をするかに関する情報が記録されています。
- **guttman.sav。**Bell (Bell, 1961) は、予想される社会グループを示す表を作成しました。Guttman (Guttman, 1968)は、この表の一部を使用しました。この表では、社会交互作用、グループへの帰属感、メンバとの物理的な近接性、関係の形式化などを表す 5 個の変数が、理論上の 7 つの社会グループと交差しています。このグループには、観衆（例、フットボールの試合の観戦者)、視聴者（例、映画館または授業の参加者)、公衆（例、新聞やテレビの視聴者)、暴徒（観衆に似ているが、より強い交互作用がある)、第一次集団（親密な関係)、第二次集団（自発的な集団)、および近代コミュニティ（物理的により密接した近接性と特化されたサービスの必要性によるゆるい同盟関係）があります。

- **health_funding.sav。**医療用資金（人口 100 人あたりの金額）、罹患率（人口 10,000 人あたりの人数）、医療サービス機関への訪問率（人口 10,000 人あたりの人数）のデータを含む、架空のデータ ファイルです。各ケースが別々の都市を表します。
- **hivassay.sav。**HIV 感染を発見する迅速な分析方法を開発するための、ある製薬研究所の取り組みに関する架空のデータ ファイルです。分析の結果は、8 段階の濃さの赤で表現され、色が濃いほど感染の可能性が高くなります。研究所では 2,000 件の血液サンプルに関して試験を行い、その半数が HIV に感染しており、半分は感染していませんでした。
- **hourlywagedata.sav。**管理職から現場担当まで、またさまざまな経験レベルの看護師の時給に関する架空のデータ ファイルです。
- **insurance_claims.sav。**不正請求の恐れがある、疑いを区別するためにモデルを作成する必要がある保険会社の仮説データ ファイルです。各ケースがそれぞれの請求を表します。
- **insure.sav。**10 年満期の生命保険契約に対し、顧客が請求を行うかどうかを示す危険因子を調査している保険会社に関する架空のデータ ファイルです。データ ファイルの各ケースは、年齢と性別が一致する、請求を行った契約と行わなかった契約のペアを表します。
- **judges.sav。**訓練を受けた審判（および 1 人のファン）が 300 件の体操の演技に対して付けた得点に関する架空のデータ ファイルです。各行が別々の演技を表し、審判たちは同じ演技を見ました。
- **kinship_dat.sav。**Rosenberg と Kim (Rosenberg および Kim, 1975) は、15 種類の親族関係用語（祖父、祖母、父、母、叔父、叔母、兄弟、姉妹、いとこ、息子、娘、甥、姪、孫息子、孫娘）の分析を行いました。Rosenberg と Kim は、大学生の 4 つのグループ（女性 2 組、男性 2 組）に、類似性に基づいて上記の用語を並べ替えるよう依頼しました。2 つのグループ（女性 1 組、男性 1 組）には、1 回目と違う条件に基づいて、2 回目の並べ替えをするように頼みました。このようにして、合計で 6 つの「ソース」が取得できました。各ソースは、15×15 の近接行列に対応します。この近接行列のセルの数は、ソースの人数から、ソース内でオブジェクトを分割した回数を引いたものです。
- **kinship_ini.sav。**このデータ ファイルには、kinship_dat.sav の 3 次元の解の初期布置が含まれています。
- **kinship_var.sav。**このデータ ファイルには、kinship_dat.sav の解の次元の解釈に使用できる独立変数である性別、世代、および(ation), and 親等が含まれています。特に、解の空間をこれらの変数の線型結合に制限するために使用できます。
- **marketvalues.sav。**1999 ～ 2000 年の間の、イリノイ州アルゴンキンの新興住宅地での住宅売上に関するデータ ファイルです。これらの売上は、公開レコードの問題となります。

- **nhis2000_subset.sav。**National Health Interview Survey (NHIS) は、米国国民を対象とした人口ベースの大規模な調査です。全国の代表的な世帯サンプルについて対面式で調査が行われます。各世帯のメンバーに関して、人口統計情報、健康に関する行動および状態の観測値が得られます。このデータ ファイルには、2000 年の調査から得られた情報のサブセットが含まれています。National Center for Health Statistics。National Health Interview Survey, 2000。一般使用データおよびドキュメント。ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/。2003 年にアクセス。
- **ozone.sav。**データには、残りの変数からオゾン濃度を予測するための、6個の気象変数に対する 330 個の観測値が含まれています。それまでの研究者 (Breiman および Friedman, 1985)、(Hastie および Tibshirani, 1990)が、他の研究者と共に、これらの変数間に非線型性を確認しています。この場合、標準的な回帰アプローチは使用できません。
- **pain_medication.sav。**この架空のデータ ファイルには、慢性関節炎を治療する抗炎症薬の臨床試験の結果が含まれています。特に興味深いことは、薬の効果が出るまでの時間と、既存の薬剤との比較です。
- **patient_los.sav。**この架空のデータ ファイルには、心筋梗塞 (MI、または「心臓発作」) の疑いで入院した患者の治療記録が含まれています。各ケースが別々の患者に対応し、入院に関連する多くの変数が記録されています。
- **patlos_sample.sav。**この架空のデータ ファイルには、心筋梗塞 (MI、または「心臓発作」) の治療中に血栓溶解剤を投薬された患者のサンプルの治療記録が含まれています。各ケースが別々の患者に対応し、入院に関連する多くの変数が記録されています。
- **poll_cs.sav。**市民の法案支持率を議会開会前に特定するための、世論調査員の取り組みに関する架空のデータ ファイルです。各ケースは登録有権者に対応しています。ケースごとに、有権者が居住している郡、町、区域が記録されています。
- **poll_cs_sample.sav。**この架空のデータ ファイルには、poll_cs.sav の有権者のサンプルが含まれています。サンプルは、poll.csplan 計画ファイルで指定されている計画に従って抽出され、このデータ ファイルには包含確率およびサンプル重み付けが記録されています。ただし、抽出計画では確率比例 (PPS) 法を使用するため、結合選択確率を含むファイル (poll_jointprob.sav) もあります。サンプル抽出後、有権者の人口統計および法案に関する意見に対応する追加の変数が収集され、データ ファイルに追加されました。
- **property_assess.sav。**限られたリソースで資産価値評価を最新に保つための、郡の評価担当者の取り組みに関する架空のデータ ファイルです。各ケースは、前年に郡内で売却された資産に対応します。データ ファイル内の各ケースでは、資産が存在する町、最後に訪問した評価

担当者、その評価からの経過時間、当時行われた評価、および資産の売却価値が記録されています。

- **property_assess_cs.sav。**限られたリソースで資産価値評価を最新に保つための、州の評価担当者の取り組みに関する架空のデータ ファイルです。各ケースは州内の資産に対応します。データ ファイル内の各ケースでは、資産が存在する郡、町、および区域、最後の評価からの経過時間、および当時行われた評価が記録されています。
- **property_assess_cs_sample.sav。**この架空のデータ ファイルには、property_assess_cs.sav の資産のサンプルが含まれています。サンプルは、property_assess.csplan 計画ファイルで指定されている計画に従って抽出され、このデータ ファイルには包含確率およびサンプル重み付けが記録されています。サンプル抽出後、現在の価値変数が収集され、データ ファイルに追加されました。
- **recidivism.sav。**管轄地域での累犯率を把握するための、政府の法執行機関の取り組みに関する架空のデータ ファイルです。各ケースは元犯罪者に対応し、人口統計情報、最初の犯罪の詳細、初犯から 2 年以内の場合は 2 回目の逮捕までの期間が記録されています。
- **recidivism_cs_sample.sav。**管轄地域での累犯率を把握するための、政府の法執行機関の取り組みに関する架空のデータ ファイルです。各ケースは 2003 年の 7 月に最初の逮捕から釈放された元犯罪者に対応し、人口統計情報、最初の犯罪の詳細、2006 年 7 月までの 2 回目の逮捕のデータが記録されています。犯罪者は recidivism_cs.csplan で指定された抽出計画に従って抽出された部門から選択されます。 調査では確率比例 (PPS) 法を採用したため、結合選択確率を保持したファイル (recidivism_cs_jointprob.sav) も用意されています。
- **rfm_transactions.sav。**購入日、購入品目、各取引のマネタリー量など、購買取引データを含む架空のデータ ファイルです。
- **salesperformance.sav。**2 つの新しい販売トレーニング コースの評価に関する架空のデータ ファイルです。60 人の従業員が 3 つのグループに分けられ、全員が標準のトレーニングを受けます。さらに、グループ 2 は技術トレーニングを、グループ 3 は実践的なチュートリアルを受けます。トレーニング コースの最後に各従業員がテストを受け、得点が記録されました。データ ファイルの各ケースは別々の訓練生を表し、割り当てられたグループと、テストの得点が記録されています。
- **satisf.sav。**ある小売業者が 4 箇所の店舗で行った満足度調査に関する架空のデータ ファイルです。合計で 582 人の顧客を調査し、各ケースは 1 人の顧客からの回答を表します。
- **screws.sav。**このデータ ファイルには、ねじ、ボルト、ナット、鋲（びょう）(Hartigan, 1975) の特性に関する情報が含まれています。
- **shampoo_ph.sav。**あるヘアケア製品工場での品質管理に関する架空のデータ ファイルです。定期的に、6 つの異なる製品が測定され、pH が記録されます。目標範囲は 4.5 ～ 5.5 です。

- **ships.sav。**他の場所（McCullagh など，1989）で表示および分析される、波による貨物船への損害に関するデータセットです。件数は、船舶の種類、建造期間、およびサービス期間によって、ポワゾン率で発生するものとしてモデリングできます。因子のクロス分類によって形成されたテーブルの各セルのサービス月数の集計によって、危険にさらされる確率の値が得られます。
- **site.sav。**業務拡大に向けて新たな用地を選択するための、ある会社の取り組みに関する架空のデータ ファイルです。2 人のコンサルタントを雇って、用地を別々に評価させました。広範囲のレポートに加えて、各用地を「良い」、「普通」、「悪い」のいずれかで集計しました。
- **smokers.sav。**このデータ ファイルは、1998 年の National Household Survey of Drug Abuse から抜粋したものであり、アメリカの世帯の確率サンプルです。(http://dx.doi.org/10.3886/ICPSR02934) したがって、このデータ ファイルを分析する場合は、まず人口の傾向を反映させてデータを重み付けする必要があります。
- **stocks.sav** このデータ ファイルには、1 年あたりの在庫価格、量が含まれています。
- **stroke_clean.sav。**この架空のデータ ファイルには、[データの準備] オプションの手続きを使用して整理した後の、医療データベースの状態が含まれています。
- **stroke_invalid.sav。**この架空のデータ ファイルには、医療データベースの初期状態が含まれており、データ入力にいくつかエラーがあります。
- **stroke_survival。** この架空のデータ ファイルは、虚血性脳卒中で数回の困難に直面した後リハビリ プログラムを終えた患者の生存時間に関するものです。脳卒中後、心筋梗塞の発生、虚血性脳卒中、または出血性脳卒中が注意され、イベントの時間が記録されます。脳卒中後に実施されたリハビリ プログラムの最後まで生存した患者のみが含まれるため、サンプルは左側が切り捨てられます。
- **stroke_valid.sav。**この架空のデータ ファイルには、[データの検証] 手続きを使用して確認した後の、医療データベースの状態が含まれています。異常である可能性のあるケースが含まれています。
- **survey_sample.sav。** このデータ ファイルには、人口統計データおよびさまざまな態度指標などの調査データが含まれています。これは「1998 NORC General Social Survey」の変数のサブセットに基づいていますが、いくつかのデータ値が変更され、追加の架空変数がデモの目的で追加されています。
- **telco.sav。**顧客ベースにおける解約率を削減するための電気通信会社の取り組みに関する架空のデータ ファイルです。各ケースが別々の顧客に対応し、人口統計やサービス利用状況などのさまざまな情報が記録されています。

- **telco_extra.sav。**このデータ ファイルは telco.sav データ ファイルに似ていますが、「期間」および対数変換された顧客支出の属性が削除され、標準化された対数変換顧客支出の変数に置き換えられています。
- **telco_missing.sav。**このデータ ファイルは telco.sav データ ファイルのサブセットですが、一部の人口統計データ値が欠損値に置き換えられています。
- **testmarket.sav。**この架空のデータ ファイルは、新しいメニューを追加しようというファースト フード チェーンの計画に関連しています。新製品をプロモーションするためのキャンペーンには 3 つの候補があるため、新メニューはいくつかのランダムに選択した市場にある場所で紹介されます。場所ごとに別々のプロモーションを使用し、最初の 4 週間の新メニューの週間売上高が記録されます。各ケースが場所と週に対応します。
- **testmarket_1month.sav。**この架空のデータ ファイルは、testmarket.sav データ ファイルの週ごとの売上を「ロールアップ」して、各ケースが別々の場所に対応するようにしたものです。その結果、週ごとに変わっていた変数の一部が表示されなくなり、売上高が、調査を行った 4 週間の売上高の合計になっています。
- **tree_car.sav。**これは、人口統計および自動車購入価格のデータを含む架空のデータ ファイルです。
- **tree_credit.sav。**これは、人口統計および銀行ローン履歴のデータを含む架空のデータ ファイルです。
- **tree_missing_data.sav。**これは、人口統計および銀行ローン履歴のデータと、多数の欠損値を含む架空のデータ ファイルです。
- **tree_score_car.sav。**これは、人口統計および自動車購入価格のデータを含む架空のデータ ファイルです。
- **tree_textdata.sav。**尺度および値ラベルを割り当てる前の、変数のデフォルトの状態を示すことを主な目的とする、変数を 2 つだけ含む単純なデータ ファイルです。
- **tv-survey.sav。**テレビ スタジオで実施された、ヒットした番組の放送期間を延長するかどうかを検討する調査に関する架空のデータ ファイルです。906 人の回答者に、さまざまな条件下でこの番組を視聴するかどうかを質問しました。各行は別々の回答者を表し、各列は別々の条件を表します。
- **ulcer_recurrence.sav。**このファイルには、潰瘍の再発を防ぐための 2 つの治療の有効性を比較するように計画された調査の情報の一部が含まれています。これは区間調査の良い例であり、他の場所 (Collett, 2003) で表示および分析されています。
- **ulcer_recurrence_recoded.sav。**このファイルでは、ulcer_recurrence.sav の情報が、単に調査終了時のイベント確率ではなく調査の区間ごとのイベント確率をモデリングできるように再編成されています。これは他の場所 (Collett など, 2003) で表示および分析されています。

- **verd1985.sav。**このデータ ファイルは調査（Verdegaal, 1985）に関連しています。8 つの変数に対する 15 人の被験者の回答を記録しました。対象となる変数が 3 つのグループに分類されます。グループ 1 には「年齢」と「婚姻」、グループ 2 には「ペット」と「新聞」、グループ 3 には「音楽」と「居住地域」がそれぞれ含まれます。「ペット」は多重名義として尺度化され、「年齢」は順序として尺度化されます。また、その他のすべての変数は単一名義として尺度化されます。
- **virus.sav。**自社のネットワーク上のウィルスの影響を特定するための、インターネット サービス プロバイダ（ISP）の取り組みに関する架空のデータ ファイルです。この ISP は、ネットワーク上の感染した E メール トラフィックの（およその）パーセンテージを、発見の瞬間から脅威が阻止されるまで追跡しました。
- **wheeze_steubenville.sav。**これは、子供（Ware, Dockery, Spiro III, Speizer, および Ferris Jr., 1984）に対する大気汚染の健康上の影響の長期調査から得られたサブセットです。このデータには、オハイオ州スビューベンビルの 7 歳、8 歳、9 歳、10 歳の子供を対象に行った、喘鳴の状態の反復 2 値測定と、調査の初年に母親が喫煙していたかどうかの固定記録が含まれています。
- **workprog.sav。**体の不自由な人をより良い仕事に就かせようとする政府の事業プログラムに関する架空のデータ ファイルです。プログラムの参加者候補のサンプルが追跡されました。その中には、ランダムに選ばれてプログラムに登録された人と、そうでない人がいました。各ケースが別々のプログラム参加者を表します。
- **worldsales.sav** このデータ ファイルには、大陸および製品ごとの販売収益が含まれています。

付録

B

注意事項

この情報は、世界各国で提供される製品およびサービス向けに作成されています。

IBM はこのドキュメントで説明する製品、サービス、機能は他の国では提供していない場合があります。現在お住まいの地域で利用可能な製品、サービス、および、情報については、お近くの IBM の担当者にお問い合わせください。IBM 製品、プログラム、またはサービスに対する参照は、IBM 製品、プログラム、またはサービスのみが使用することができることを説明したり意味するものではありません。IBM の知的所有権を侵害しない機能的に同等の製品、プログラム、またはサービスを代わりに使用することができます。ただし、IBM 以外の製品、プログラム、またはサービスの動作を評価および確認するのはユーザーの責任によるものです。

IBMは、本ドキュメントに記載されている内容に関し、特許または特許出願中の可能性があります。本ドキュメントの提供によって、これらの特許に関するいかなる権利も使用者に付与するものではありません。ライセンスのお問い合わせは、書面にて、下記住所に送ることができます。

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

2 バイト文字セット (DBCS) 情報についてのライセンスに関するお問い合わせは、お住まいの国の IBM Intellectual Property Department に連絡するか、書面にて下記宛先にお送りください。

神奈川県大和市下鶴間1623番14号 日本アイ・ビー・エム株式会社 法務・知的財産 知的財産権ライセンス渉外

以下の条項はは、イギリスまたはこのような条項が法律に反する他の国では適用されません。 International Business Machines は、明示的または黙示的に関わらず、第三者の権利の侵害しない、商品性または特定の目的に対する適合性の暗黙の保証を含むがこれに限定されない、いかなる保証なく、本出版物を「そのまま」提供します一部の州では、特定の取引の明示的または暗示的な保証の免責を許可していないため、この文が適用されない場合があります。

この情報には、技術的に不適切な記述や誤植を含む場合があります。情報については変更が定期的に行われます。これらの変更は本書の新版に追加されます。IBM は、本書に記載されている製品およびプログラムについて、事前の告知なくいつでも改善および変更を行う場合があります。

IBM 以外の Web サイトに対するこの情報内のすべての参照は、便宜上提供されているものであり、決してそれらの Web サイトを推奨するものではありません。これらの Web サイトの資料はこの IBM 製品の資料に含まれるものではなく、これらの Web サイトの使用はお客様の責任によるものとします。

IBM はお客様に対する一切の義務を負うことなく、自ら適切と考える方法で、情報を使用または配布することができるものとします。

本プログラムのライセンス取得者が (i) 別途作成されたプログラムと他のプログラム (本プログラムを含む) との間の情報交換および (ii) 交換された情報の相互利用を目的とした本プログラムに関する情報の所有を希望する場合、下記住所にお問い合わせください。

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

上記のような情報は、該当する条項および条件に従い、有料で利用できるものとします。

本ドキュメントに記載されている許可されたプログラムおよびそのプログラムに使用できるすべてのライセンス認証された資料は、IBM Customer Agreement、IBM International Program License Agreement、および当社とかわした同等の契約の条件に基づき、IBM によって提供されます。

IBM 以外の製品に関する情報は、それらの製品の供給業者、公開済みの発表、または公開で使用できるソースから取得しています。IBM は、それらの製品のテストは行っておらず、IBM 以外の製品に関連する性能、互換性、またはその他の要求については確証できません。IBM 以外の製品の性能に関する質問は、それらの製品の供給業者に通知する必要があります。

この情報には、日常の業務処理で用いられるデータや報告書の例が含まれています。できる限り詳細に説明するため、例には、個人、企業、ブランド、製品などの名前が使用されています。これらの名称はすべて架空のものであり、実際の企業で使用される名称および住所とは一切関係ありません。

この情報をソフトコピーでご覧になっている場合は、写真やカラーのイラストが表示されない場合があります。

商標

IBM、IBM ロゴ、および ibm.com、SPSS は、世界の多くの国で登録された IBM Corporation の商標です。IBM の商標の現在のリストは、http://www.ibm.com/legal/copytrade.shtml を参照してください。

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 のロゴ、Intel Inside、Intel Inside のロゴ、Intel Centrino、Intel Centrino のロゴ、Celeron、Intel Xeon、Intel SpeedStep、Itanium、および Pentium は、米国およびその他の国の Intel Corporation または関連会社の商標または登録商標です。

Java およびすべての Java ベースの商標およびロゴは、米国およびその他の国の Sun Microsystems, Inc. の商標です。

Linux は、米国およびその他の国における Linus Torvalds の登録商標です。

Microsoft、Windows、Windows NT、および Windows のロゴは、米国およびその他の国における Microsoft 社の商標です。

UNIX は、米国およびその他の国における The Open Group の登録商標です。

この製品は、WinWrap Basic (Copyright 1993-2007, Polar Engineering and Consulting, http://www.winwrap.com) を使用します。

その他の製品名およびサービス名等は、IBM または他の会社の商標です。

Adobe 製品のスクリーンショットは Adobe Systems Incorporated の許可を得て転載しています。

Microsoft 製品のスクリーンショットは Microsoft 社の許可を得て転載しています。

参考文献

Bell, E. H. 1961. Social foundations of human behavior: Introduction to the study of sociology. New York: Harper & Row.

Bishop, C. M. 1995. Neural Networks for Pattern Recognition, 3rd ed. Oxford: Oxford University Press.

Blake, C. L., および C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., および J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80, 580-598.

Collett, D. 2003. Modelling survival data in medical research, 2 ed. Boca Raton: Chapman & Hall/CRC.

Fine, T. L. 1999. Feedforward Neural Network Methodology, 3rd ed. New York: Springer-Verlag.

Green, P. E., および V. Rao. 1972. Applied multidimensional scaling. Hinsdale, Ill.: Dryden Press.

Green, P. E., および Y. Wind. 1973. Multiattribute decisions in marketing: A measurement approach. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. Psychometrika, 33, 469-506.

Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley and Sons.

Hastie, T., および R. Tibshirani. 1990. Generalized additive models. London: Chapman and Hall.

Haykin, S. 1998. Neural Networks: A Comprehensive Foundation, 2nd ed. New York: Macmillan College Publishing.

Kennedy, R., C. Riquier, および B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. Journal of Targeting, Measurement, and Analysis for Marketing, 5, 56-70.

McCullagh, P., および J. A. Nelder. 1989. Generalized Linear Models, 2nd ed. London: Chapman & Hall.

Price, R. H., および D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. Journal of Personality and Social Psychology, 30, 579-586.

Rickman, R., N. Mitchell, J. Dingman, および J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. Journal of the American Medical Association, 228, 54-58.

Ripley, B. D. 1996. Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press.

Rosenberg, S., および M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. Multivariate Behavioral Research, 10, 489-502.

Tao, K. K. 1993. A closer look at the radial basis function (RBF) networks. In: Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers, A. Singh, ed. Los Alamitos, Calif.: IEEE Comput. Soc. Press, 401-405.

Uykan, Z., C. Guzelis, M. E. Celebi, および H. N. Koivo. 2000. Analysis of input-output clustering for determining centers of RBFN. IEEE Transactions on Neural Networks, 11, 851-858.

Van der Ham, T., J. J. Meulman, D. C. Van Strien, および H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. British Journal of Psychiatry, 170, 363-368.

Verdegaal, R. 1985. Meer sets analyse voor kwalitatieve gegevens (in Dutch). Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, および B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. American Review of Respiratory Diseases, 129, 366-374.

索引