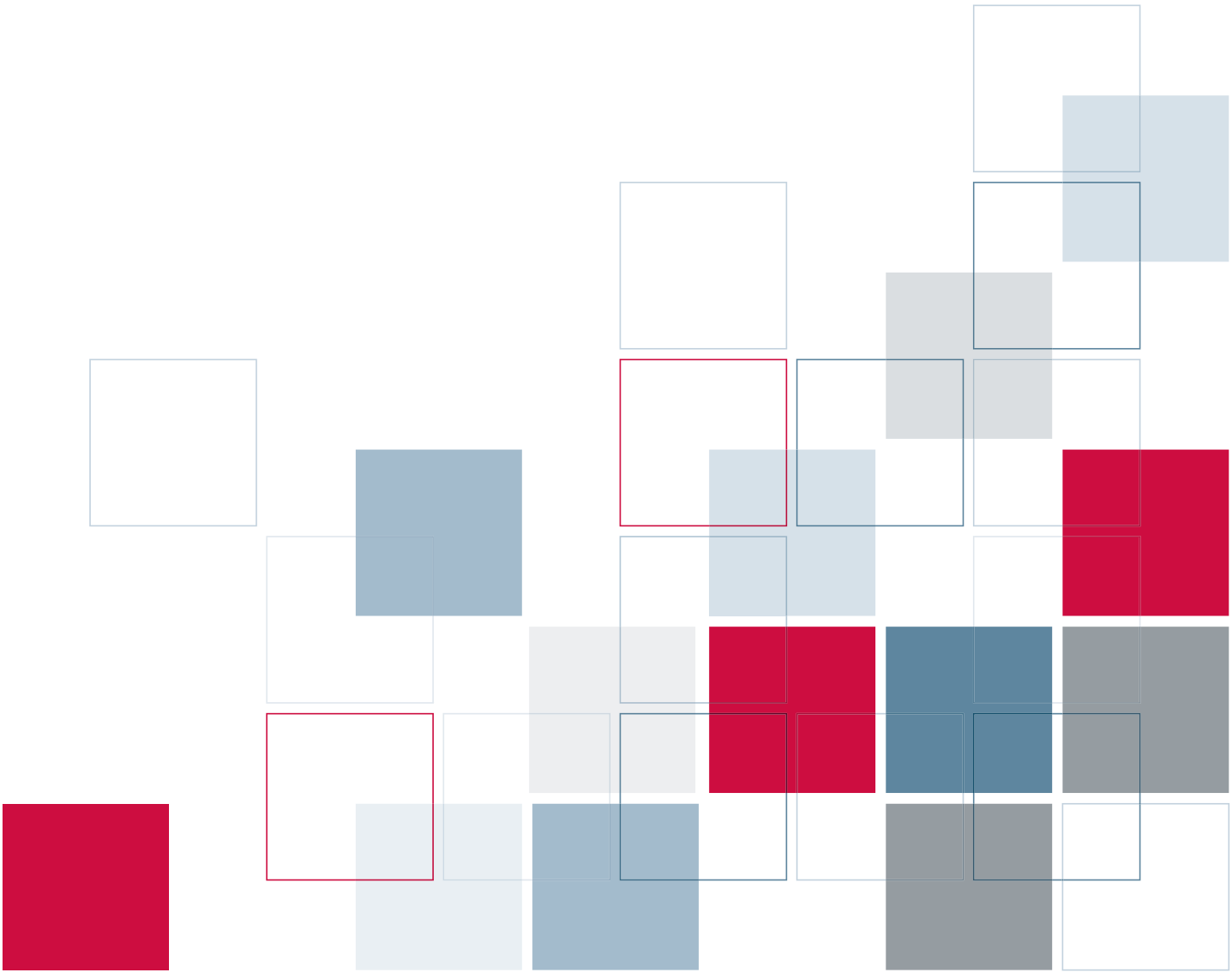
SPSS Neural Networks (Neuronale Netze)™ 16.0



Weitere Informationen zu SPSS®-Software-Produkten finden Sie auf unserer Website unter der Adresse *http://www.spss.com* oder wenden Sie sich an

SPSS Inc.
233 South Wacker Drive, 11th Floor
Chicago, IL 60606-6412, USA
Tel.: (312) 651-3000
Fax: (312) 651-3668

SPSS Neural Networks™ 16.0

ISBN-13: 978-1-56827-867-4
ISBN-10: 1-56827-867-5

1 2 3 4 5 6 7 8 9 0 10 09 08 07

Vorwort

SPSS 16.0 ist ein umfassendes System zum Analysieren von Daten. Das optionale Erweiterungsmodul SPSS Neural Networks (Neuronale Netze) bietet die zusätzlichen Analyseverfahren, die in diesem Handbuch beschrieben sind. Die Prozeduren im Erweiterungsmodul Neural Networks (Neuronale Netze) müssen zusammen mit SPSS 16.0 Base verwendet werden. Sie sind vollständig in dieses System integriert.

Installation

Zur Installation von SPSS Neural Networks (Neuronale Netze) Erweiterungsmodul führen Sie den Lizenzautorisierungsassistenten mit dem Autorisierungscode aus, den Sie von SPSS erhalten haben. Weitere Informationen finden Sie in den Installationsanweisungen im Lieferumfang von SPSS Neural Networks (Neuronale Netze) Erweiterungsmodul.

Kompatibilität

SPSS kann auf vielen Computersystemen ausgeführt werden. Mindestanforderungen an das System und Empfehlungen finden Sie in den Unterlagen, die mit Ihrem System geliefert werden.

Seriennummern

Die Seriennummer des Programms dient gleichzeitig als Identifikationsnummer bei SPSS. Sie benötigen diese Seriennummer, wenn Sie sich an SPSS wenden, um Informationen über Kundendienst, zu Zahlungen oder Aktualisierungen des Systems zu erhalten. Die Seriennummer wird mit dem Base-System ausgeliefert.

Kundendienst

Wenden Sie sich mit Fragen bezüglich der Lieferung oder Ihres Kundenkontos an Ihr regionales SPSS-Büro, das Sie auf der SPSS-Website unter *http://www.spss.com/worldwide* finden. Halten Sie bitte stets Ihre Seriennummer bereit.

Ausbildungsseminare

SPSS bietet öffentliche und unternehmensinterne Seminare an. Alle Seminare beinhalten auch praktische Übungen. Seminare finden in größeren Städten regelmäßig statt. Wenn Sie weitere Informationen zu diesen Schulungen wünschen, wenden Sie sich an Ihr regionales SPSS-Büro, das Sie auf der SPSS-Website unter *http://www.spss.com/worldwide* finden.

Technischer Support

Kunden von SPSS mit Wartungsvertrag können den Technischen Support in Anspruch nehmen. Kunden können sich an den Technischen Support wenden, wenn sie Hilfe bei der Arbeit mit SPSS oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Informationen über den Technischen Support finden Sie auf der Website von SPSS unter *http://www.spss.com* oder wenden Sie sich an Ihr regionales SPSS-Büro, das Sie auf der SPSS-Website unter *http://www.spss.com/worldwide* finden. Bei einem Anruf werden Sie nach Ihrem Namen, dem Namen Ihrer Organisation und Ihrer Seriennummer gefragt.

Weitere Veröffentlichungen

Weitere Exemplare von Produkthandbüchern können direkt bei SPSS Inc. bestellt werden. Besuchen Sie den SPSS Web Store unter *http://www.spss.com/estore* oder wenden Sie sich an Ihr regionales SPSS-Büro, das Sie auf der SPSS-Website unter *http://www.spss.com/worldwide* finden. Wenden Sie sich bei telefonischen Bestellungen in den USA und Kanada unter 800-543-2185 direkt an SPSS Inc. Wenden Sie sich bei telefonischen Bestellungen außerhalb von Nordamerika an Ihr regionales SPSS-Büro, das Sie auf der SPPS-Website finden.

Das Handbuch *SPSS Statistical Procedures Companion* von Marija Norušis wurde von Prentice Hall veröffentlicht. Eine neue Fassung dieses Buchs mit Aktualisierungen für SPSS 16.0 ist geplant. Das Handbuch *SPSS Advanced Statistical Procedures Companion*, bei dem auch SPSS 16.0 berücksichtigt wird, erscheint demnächst. Das Handbuch *SPSS Guide to Data Analysis* für SPSS 16.0 wird ebenfalls derzeit erstellt. Ankündigungen für Veröffentlichungen, die ausschließlich über Prentice Hall verfügbar sind, finden Sie auf der SPSS-Website unter *http://www.spss.com/estore* (wählen Sie Ihr Land aus und klicken Sie auf Books).

Kundenmeinungen

Ihre Meinung ist uns wichtig. Teilen Sie uns bitte Ihre Erfahrungen mit SPSS-Produkten mit. Insbesondere haben wir Interesse an neuen, interessanten Anwendungsgebieten von SPSS Neural Networks (Neuronale Netze) Erweiterungsmodul. Senden Sie uns eine E-Mail an *suggest@spss.com* oder schreiben Sie an: SPSS Inc., Attn: Director of Product Planning, 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.

Über dieses Handbuch

In diesem Handbuch wird die grafische Benutzeroberfläche für die in SPSS Neural Networks (Neuronale Netze) Erweiterungsmodul enthaltenen Prozeduren erläutert. Die Abbildungen der Dialogfelder stammen aus SPSS. Detaillierte Informationen zur Befehlssyntax für die Funktionen in SPSS Neural Networks (Neuronale Netze) Erweiterungsmodul sind auf zwei Arten verfügbar: als Bestandteil des umfassenden Hilfesystems und als separates Dokument im PDF-Format im Handbuch *SPSS 16.0 Command Syntax Reference*, das auch über das Menü “Hilfe” verfügbar ist.

Kontakt zu SPSS

Wenn Sie in unseren Verteiler aufgenommen werden möchten, wenden Sie sich an eines unserer Büros, die Sie auf unserer Website unter *http://www.spss.com/worldwide* finden.

Inhalt

Teil II: Beispiele

4 Mehrschichtiges Perzeptron 38

5 Radiale Basisfunktion 75

Anhang

Teil I: Benutzerhandbuch

Kapitel

1

Einführung in SPSS Neural Networks (Neuronale Netze)

Neuronale Netze sind aufgrund ihrer Leistungsfähigkeit, Flexibilität und Benutzerfreundlichkeit das bevorzugte Werkzeug für zahlreiche Anwendungen auf dem Gebiet des prädiktiven Data-Mining. Prädiktive neuronale Netze sind besonders nützlich bei Anwendungen, denen ein komplexer Prozess zugrunde liegt, wie beispielsweise:

- Prognose der Verbrauchernachfrage zur Rationalisierung von Produktions- und Lieferkosten.
- Vorhersage der Antwortwahrscheinlichkeit bei Marketingaktionen mit Postsendungen, um zu ermitteln, an welche Haushalte im Verteiler ein Angebot gesendet werden sollte.
- Scoring eines Antragstellers, um dessen Kreditrisiko zu ermitteln.
- Aufdecken betrügerischer Transaktionen in einer Datenbank mit Versicherungsforderungen.

Die in Prognoseanwendungen, wie Netzwerken vom Typ **Mehrschichtiges Perzeptron (MLP)** und **Radiale Basisfunktion (RBF)**, verwendeten Prognoseanwendungen werden dahingehend überwacht, dass die vom Modell vorhergesagten Ergebnisse mit bekannten Werten der Zielvariablen verglichen werden können. Mit der Option SPSS Neural Networks können Sie MLP- und RBF-Netzwerke anpassen und die so entstehenden Modelle für das Scoring speichern.

Was ist ein neuronales Netz?

Der Begriff **neuronales Netzwerk** bezieht sich auf eine locker miteinander verwandte Modellfamilie, die durch einen großen Parameterraum und eine flexible Struktur gekennzeichnet ist, die sich aus den Studien zur Funktionsweise des Gehirns herleitet. Als die Modellfamilie wuchs, wurden die meisten neuen Modelle für Anwendungen außerhalb der Biologie entwickelt, obwohl ein Großteil der zugehörigen Terminologie noch die Ursprünge erkennen lässt.

Die spezifischen Definitionen für neuronale Netze sind so vielfältig wie ihre Einsatzgebiete. Es gibt keine Definition, die die gesamte Modellfamilie richtig erfassen würde. Wir verwenden jedoch vorläufig folgende Beschreibung(Haykin, 1998):

Ein neuronales Netz ist ein verteilter massiv-paralleler Prozessor mit einer natürlichen Neigung zur Speicherung von experimentellem Wissen und seiner Bereitstellung. Es ähnelt dem Hirn in zwei Aspekten:

- Wissen wird vom Netzwerk durch einen Lernprozess erworben.
- Interneuronale Verbindungsstärken, auch als synaptische Gewichte bekannt, dienen zum Speichern des Wissens.

In (Ripley, 1996) finden Sie eine Diskussion darüber, warum diese Definition möglicherweise zu restriktiv ist.

Wenn wir neuronale Netze mit dieser Definition von traditionellen statistischen Methoden unterscheiden möchten, ist das, was *nicht* gesagt wurde, ebenso bedeutsam, wie der Text der Definition selbst. So kann beispielsweise das traditionelle lineare Regressionsmodell Wissen durch die Methode der kleinsten Quadrate erwerben und dieses Wissen in den Regressionskoeffizienten speichern. In dieser Hinsicht handelt es sich dabei um ein neuronales Netz. In der Tat lässt sich die Auffassung vertreten, dass die lineare Regression einen Sonderfall bestimmter neuronaler Netze darstellt. Allerdings weist die lineare Regression eine starre Modellstruktur und eine Menge von Annahmen auf, die angewendet werden, bevor aus den Daten "gelernt" wird.

Im Gegensatz dazu stellt die oben angegebene Definition nur minimale Anforderungen an Struktur und Annahmen. Daher kann ein neuronales Netz eine Annäherung an eine große Bandbreite statistischer Modelle bieten, ohne dass von vornherein Hypothesen über bestimmte Beziehungen zwischen den abhängigen und den unabhängigen Variablen erforderlich sind. Stattdessen wird die Form der Beziehungen im Laufe des Lernprozesses bestimmt. Wenn eine lineare Beziehung zwischen abhängigen und unabhängigen Variablen angemessen ist, sollten die Ergebnisse des neuronalen Netzwerks eine große Ähnlichkeit zu denen des linearen Regressionsmodells darstellen. Wenn eine nichtlineare Beziehung angemessener ist, ähnelt das neuronale Netzwerk automatisch der "richtigen" Modellstruktur.

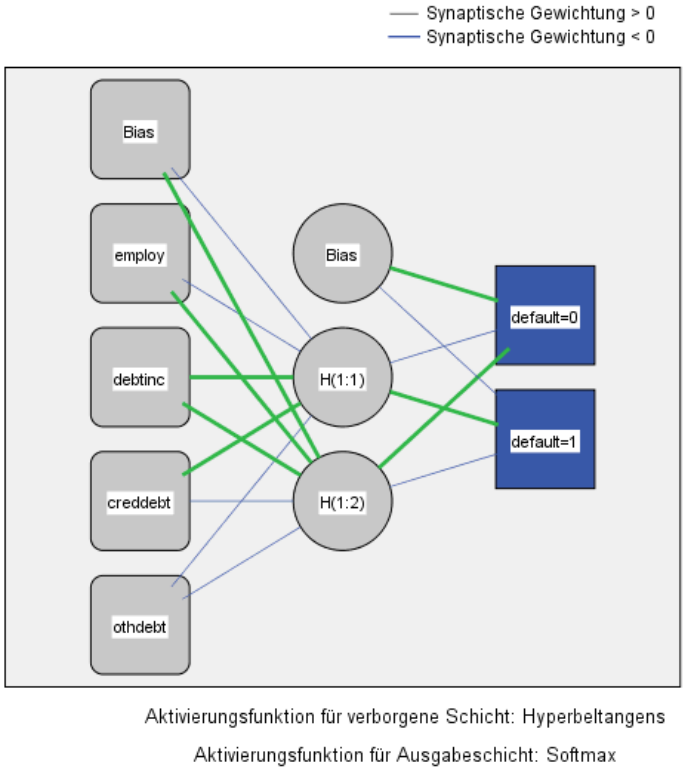
Der Preis für diese Flexibilität besteht darin, dass die synaptischen Gewichte eines neuronalen Netzwerks nicht leicht zu interpretieren sind. Wenn Sie also versuchen, den zugrunde liegenden Prozess zu erklären, der zu den Beziehungen zwischen den abhängigen und den unabhängigen Variablen führt, sollten Sie lieber ein traditionelleres statistisches Modell verwenden. Wenn jedoch die Interpretierbarkeit des Modells nicht von Belang ist, können Sie häufig schneller mithilfe eines neuronalen Netzwerks zu guten Modellergebnissen kommen.

Struktur neuronaler Netze

Auch wenn neuronale Netze nur minimale Anforderungen an die Modellstruktur und die geltenden Annahmen stellen, ist es dennoch nützlich, einen Einblick in die allgemeine **Netzwerkarchitektur** zu haben. Beim MLP- bzw. RBF-Netzwerk handelt es sich um eine Funktion von Einflussvariablen (auch als Prädiktoren, Eingaben oder unabhängige Variablen bezeichnet), die den Vorhersagefehler der Zielvariablen (auch als Ausgaben bezeichnet) minimiert.

Betrachten Sie das Daten-Set *bankloan.sav*, das im Lieferumfang des Produkts enthalten ist. In diesem Daten-Set sollen aus einem Pool von Kreditantragstellern die Personen ermittelt werden, die mit großer Wahrscheinlichkeit zahlungsunfähig werden. Bei einem auf dieses Problem angewendeten MLP- bzw. RBF-Netzwerk handelt es sich um eine Funktion von Messungen, die den Fehler bei der Vorhersage der Zahlungsunfähigkeit minimiert. Folgende Abbildung dient zur Angabe der Form dieser Funktion.

Abbildung 1-1
Feedforward-Architektur mit einer verborgenen Schicht



Diese Struktur ist als **Feedforward-Architektur** bekannt, da die Verbindungen im Netzwerk ohne Rückkopplungsschleifen vorwärts von der Eingabeschicht zur Ausgabeschicht verlaufen. In dieser Abbildung gilt:

- Die **Eingabeschicht** enthält die Einflussvariablen.
- Die **verborgene Schicht** enthält nicht sichtbare Knoten (Einheiten). Der Wert der verborgenen Einheiten ist jeweils eine Funktion der Einflussvariablen; die genaue Form der Funktion hängt zum Teil vom Netzwerktyp und zum Teil von den vom Benutzer festlegbaren Spezifikationen ab.
- Die **Ausgabeschicht** enthält die Antworten. Da es sich bei den früheren Fällen von Zahlungsverzug um eine kategoriale Variable mit zwei Kategorien handelt, wird sie als zwei Indikatorvariablen umkodiert. Jede Ausgabeeinheit ist jeweils eine Funktion der verborgenen Einheiten. Auch hier hängt die genaue Form der Funktion zum Teil vom Netzwerktyp und zum Teil von den vom Benutzer festlegbaren Spezifikationen ab.

Beim MLP-Netzwerk ist eine zweite verborgene Schicht zulässig; in diesem Fall ist jede Einheit der zweiten verborgenen Schicht eine Funktion der Einheiten in der ersten verborgenen Schicht, und jede Antwort ist eine Funktion der Einheiten in der zweiten verborgenen Schicht.

Kapitel
2

Mehrschichtiges Perzeptron

Die Prozedur "Mehrschichtiges Perzeptron" (Multilayer Perceptron, MLP) erstellt ein Vorhersagemodell für eine oder mehrere abhängige Variablen (Zielvariablen), das auf den Werten der Einflussvariablen beruht.

Beispiele. Im Folgenden finden Sie zwei Szenarien, die die Prozedur MLP verwenden:

Eine Kreditsachbearbeiterin in einer Bank muss in der Lage sein, Merkmale zu ermitteln, die auf Personen hindeuten, die mit hoher Wahrscheinlichkeit ihre Kredite nicht zurückzahlen, und diese Merkmale zur Feststellung eines guten bzw. schlechten Kreditrisikos einzusetzen. Mithilfe einer Stichprobe von früheren Kunden kann sie ein mehrschichtiges Perzeptron trainieren, die Analysen anhand einer Prüf-(Holdout-) Stichprobe früherer Kunden validieren und anschließend mit dem Netzwerk das Kreditrisiko potenzieller Kunden als gering oder hoch einstufen.

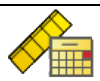
Ein Krankenhaussystem möchte die Kosten und die Aufenthaltsdauer für Patienten aufzeichnen, die zur Behandlung eines Herzinfarkts aufgenommen wurden. Durch genaue Schätzer dieser Messwerte kann die Krankenhausverwaltung die verfügbare Bettenkapazität während der Behandlung der Patienten besser verwalten. Mithilfe der Behandlungsakten einer Stichprobe von Patienten, die wegen eines Herzinfarkts behandelt wurden, kann die Verwaltung ein Netzwerk trainieren, mit dem sich die Kosten und die Dauer des Aufenthalts vorhersagen lassen.

Abhängige Variablen. Die abhängigen Variablen können wie folgt gestaltet sein:

- **Nominal.** Eine Variable kann als nominal behandelt werden, wenn ihre Kategorien sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- **Ordinal.** Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.
- **Metrisch.** Eine Variable kann als metrisch behandelt werden, wenn ihre Werte geordnete Kategorien mit einer sinnvollen Metrik darstellen, sodass man sinnvolle Aussagen über die Abstände zwischen den Werten machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

  Bei der Prozedur wird davon ausgegangen, dass allen abhängigen Variablen das richtige Messniveau zugewiesen wurde. Sie können das Messniveau für eine Variable jedoch vorübergehend ändern. Klicken Sie hierzu mit der rechten Maustaste auf die Variable in der Liste der Quellvariablen und wählen Sie das gewünschte Messniveau im Kontextmenü aus.

Messniveau und Datentyp sind durch ein Symbol neben der jeweiligen Variablen in der Variablenliste gekennzeichnet:

Messniveau	**Datentyp**			
	Numerisch	String	Datum	Zeit
Metrisch		entfällt		
Ordinal				
Nominal				

Einflussvariablen. Einflussvariablen können als Faktoren (kategorial) oder als Kovariaten (metrisch) angegeben werden.

Kodierung für kategoriale Variablen. Die Prozedur kodiert vorübergehend für die Dauer des Verfahrens kategoriale Einflussvariablen und abhängige Variablen mithilfe der "Eins-aus-*c*"-Kodierung neu. Wenn es *c* Kategorien für eine Variable gibt, wird die Variable als c Vektoren gespeichert. Dabei wird die erste Kategorie als (1,0,...,0) angegeben, die zweite Kategorie als (0,1,0,...,0), ... und die letzte Kategorie als (0,0,...,0,1).

Dieses Kodierungsschema erhöht die Anzahl der synaptischen Gewichtungen und kann zu einer Verlangsamung des Trainings führen, "kompaktere" Kodierungsmethoden führen jedoch in der Regel zu neuronalen Netzwerken mit geringer Anpassungsgüte. Wenn das Training des Netzwerks sehr langsam vorangeht, können Sie versuchen, die Anzahl der Kategorien der kategorialen Einflussvariablen zu verringern, indem Sie ähnliche Kategorien zusammenfassen oder Fälle ausschließen, die extrem seltene Kategorien aufweisen.

Jegliche "Eins-aus-*c*"-Kodierung beruht auf den Trainingsdaten, selbst wenn eine Test- bzw. Holdout-Stichprobe definiert wurde (siehe Partitionen auf S. 8). Wenn also die Test- bzw. Holdout-Stichproben Fälle mit Einflussvariablen-Kategorien enthalten, die in den Trainingsdaten nicht vorhanden sind, werden diese Fälle nicht in der Prozedur bzw. beim Scoring verwendet. Wenn die Test- bzw. Holdout-Stichproben Fälle mit Kategorien abhängiger Variablen enthalten, die in den Trainingsdaten nicht vorhanden sind, werden diese Fälle zwar nicht in der Prozedur, jedoch möglicherweise beim Scoring verwendet.

Neuskalierung. Metrische abhängige Variablen und Kovariaten werden standardmäßig neu skaliert, um das Training des Netzwerks zu verbessern. Jegliche Neuskalierung beruht auf den Trainingsdaten, selbst wenn eine Test- bzw. Holdout-Stichprobe definiert wurde (siehe Partitionen auf S. 8). Das bedeutet, dass je nach Neuskalierungstyp Mittelwert, Standardabweichung, Mindestwert bzw. Höchstwert einer Kovariaten oder abhängigen Variablen ausschließlich anhand der Trainingsdaten berechnet wird. Wenn Sie eine Variable zur Festlegung von Partitionen angeben, müssen diese Kovariaten bzw. abhängigen Variablen in der Trainings-, Test- und Holdout-Stichprobe ähnliche Verteilungen aufweisen.

Häufigkeitsgewichtungen. Häufigkeitsgewichtungen werden von dieser Prozedur ignoriert.

Reproduzieren der Ergebnisse. Wenn Sie Ihre Ergebnisse exakt reproduzieren möchten, müssen Sie nicht nur dieselben Einstellungen für die Prozedur, sondern auch denselben Initialisierungswert für den Zufallszahlengenerator, dieselbe Datenreihenfolge und dieselbe Variablenreihenfolge verwenden. Weitere Details zu diesem Problem folgen:

- **Generierung von Zufallszahlen.** Die Prozedur verwendet Zufallszahlengenerierung während der Zufallszuweisung von Partitionen, zufällige Ziehung von Teilstichproben für die Initialisierung der synaptischen Gewichtungen, zufällige Ziehung von Teilstichproben für die automatische Architekturauswahl und den Algorithmus der simulierten Abkühlung für die Initialisierung der Gewichtungen und die automatische Architekturauswahl. Um zu einem späteren Zeitpunkt dieselben randomisierten Ergebnisse zu reproduzieren, müssen Sie vor jeder Ausführung der Prozedur "Mehrschichtiges Perzeptron" denselben Initialisierungswert für den Zufallszahlengenerator verwenden. Einzelschrittanweisungen hierzu finden Sie unter Vorbereiten der Daten für die Analyse auf S. 38.
- **Fallreihenfolge.** Die Trainingsmethoden "Online" und "Mini-Batch" (siehe Training auf S. 13) sind explizit von der Fallreihenfolge abhängig; allerdings ist sogar Batch-Training von der Fallreihenfolge abhängig, da die Initialisierung der synaptischen Gewichtungen die Ziehung einer Teilstichprobe aus dem Daten-Set beinhaltet.

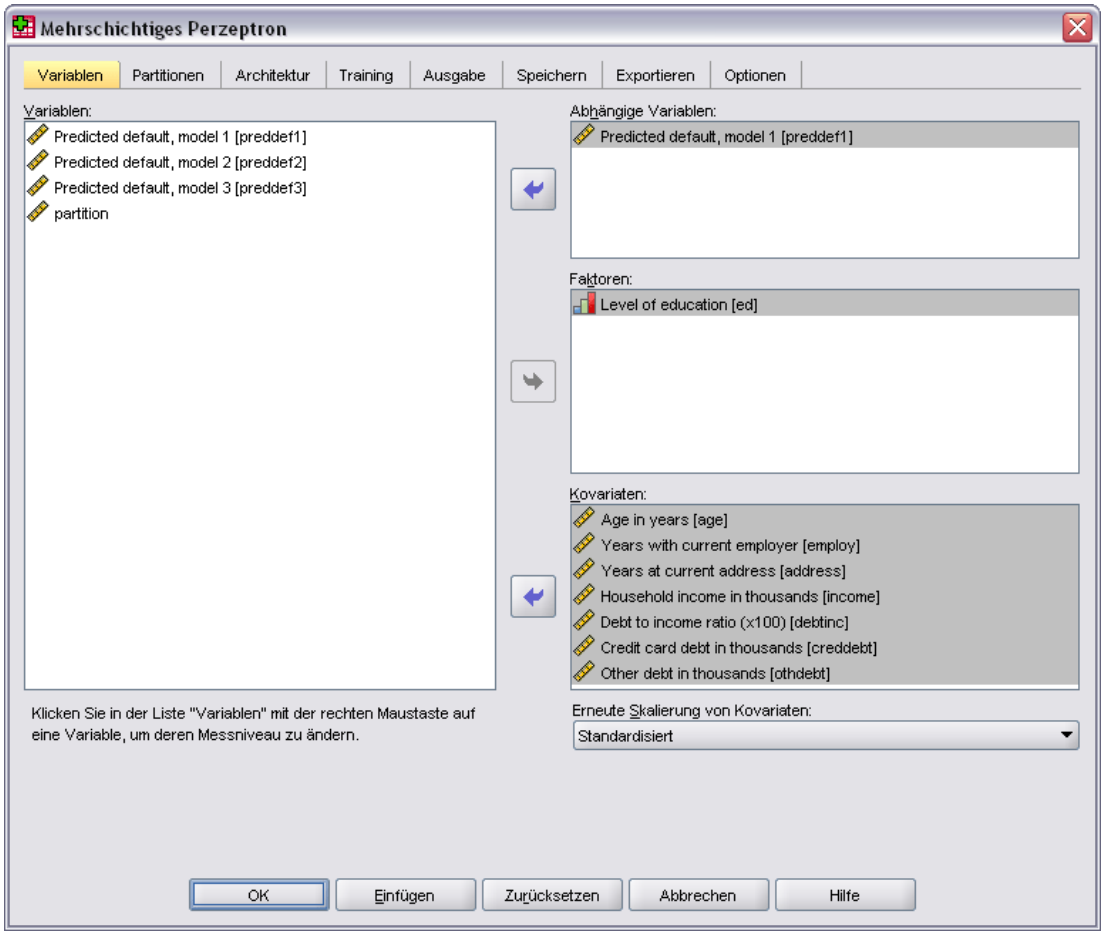
  Um die Auswirkungen der Reihenfolge zu minimieren, mischen Sie die Fälle in zufälliger Reihenfolge. Prüfen Sie daher die Stabilität einer bestimmten Lösung, indem Sie verschiedene Lösungen abrufen, bei denen die Fälle in einer unterschiedlichen, zufällig ausgewählten Reihenfolgen sortiert sind. In Situationen mit extrem umfangreichen Dateien können mehrere Durchgänge mit jeweils einer Stichprobe von Fällen durchgeführt werden, die in unterschiedlicher, zufällig ausgewählter Reihenfolge sortiert ist.
- **Reihenfolge der Variablen.** Die Ergebnisse können von der Reihenfolge der Variablen in der Faktorenliste und der Kovariatenliste beeinflusst werden, da die zugewiesenen Anfangswerte ein anderes Muster aufweisen, wenn die Reihenfolge der Variablen geändert wird. Wie bei den Effekten der Fallreihenfolge können Sie auch eine andere Reihenfolge der Variablen ausprobieren (durch Ziehen und Ablegen in der Liste der Faktoren bzw. Kovariaten), um die Stabilität einer bestimmten Lösung einzuschätzen.

Erstellen eines Netzwerks mit mehrschichtigen Perzeptronen

Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Neuronale Netze
    Mehrschichtiges Perzeptron...

Abbildung 2-1
Mehrschichtiges Perzeptron: Registerkarte "Variablen"



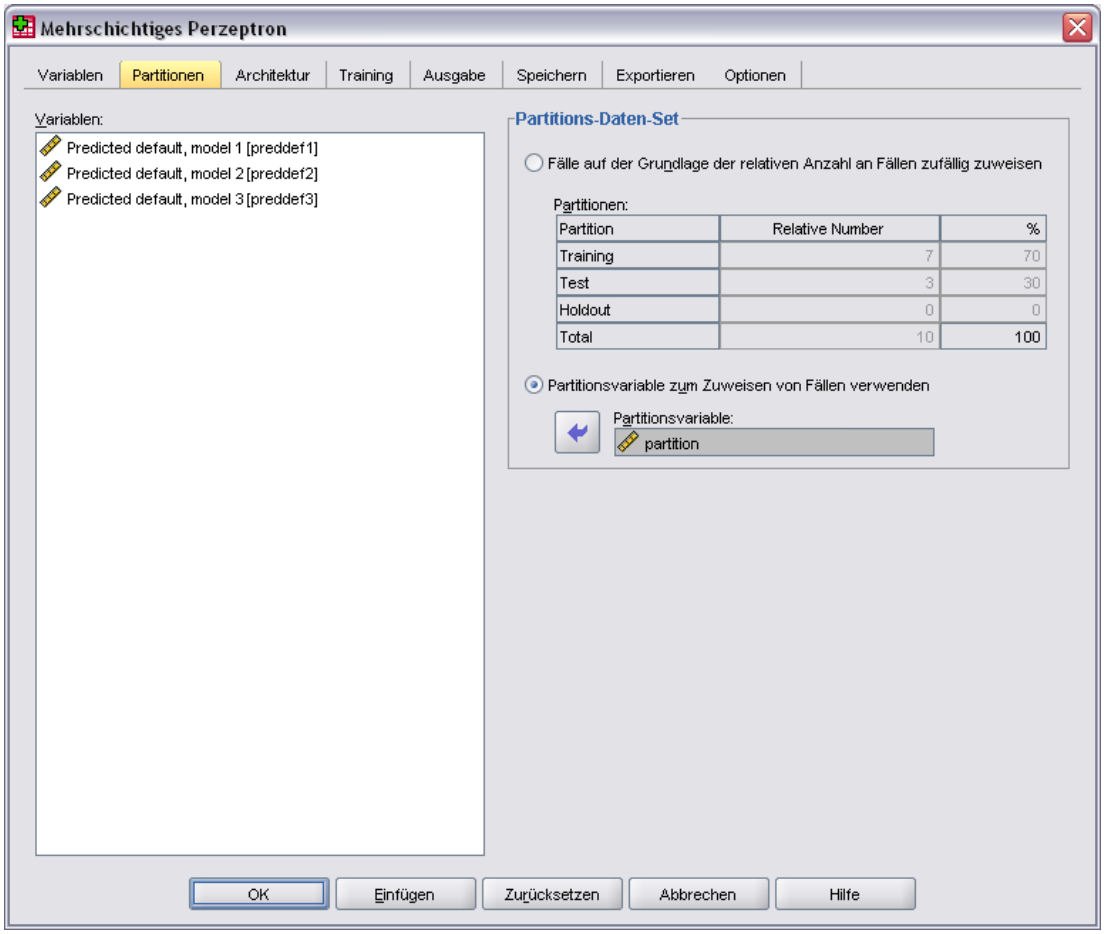
- ▶ Wählen Sie mindestens eine abhängige Variable aus.
- ▶ Wählen Sie mindestens einen Faktor oder eine Kovariate aus.

Optional können Sie auf der Registerkarte "Variablen" die Methode zur Neuskalierung der Kovariaten ändern. Folgende Optionen stehen zur Auswahl:

- **Standardisiert.** Subtraktion des Mittelwerts und Division durch die Standardabweichung, (x−Mittelwert)/s.
- **Normalisiert.** Subtraktion des Mittelwerts und Division durch den Bereich, (x−min)/(max−min). Normalisierte Werte liegen im Bereich zwischen 0 und 1.
- **Angepasst normalisiert.** Angepasste Version der Subtraktion des Mittelwerts und Division durch den Bereich, [2*(x−min)/(max−min)]−1. Angepasste normalisierte Werte liegen zwischen −1 und 1.
- **Keine.** Keine Neuskalierung der Kovariaten.

Partitionen

Abbildung 2-2
Mehrschichtiges Perzeptron: Registerkarte "Partitionen"



Partitions-Daten-Set. Diese Gruppe gibt die Methode zur Partitionierung der Arbeitsdatei in eine Trainings-, eine Test- und eine Holdout-Stichprobe an. Die **Trainingsstichprobe** umfasst die Datensätze, die zum Trainieren des neuronalen Netzwerks verwendet wurden; ein gewisser Prozentsatz der Fälle im Daten-Set muss der Trainingsstichprobe zugewiesen werden, um ein Modell zu erhalten. Die **Teststichprobe** ist ein unabhängiges Set von Datensätzen, die verwendet werden, um den Fehler während des Trainings aufzuzeichnen und dadurch ein Übertrainieren zu vermeiden. Es wird dringend empfohlen, eine Trainingsstichprobe zu erstellen. Das Netzwerktraining ist in der Regel am effizientesten, wenn die Teststichprobe kleiner ist als die Trainingsstichprobe. Die **Holdout-Stichprobe** ist ein weiterer unabhängiger Satz von Datensätzen, der zur Bewertung des endgültigen neuronalen Netzwerks verwendet wird; der Fehler für die Houldout-Stichprobe bietet eine "ehrliche" Schätzung der Vorhersagekraft des

Modells, da die Prüffälle (die Fälle in der Holdout-Stichprobe) nicht zur Erstellung des Modells verwendet wurden.

- **Fälle auf der Grundlage der relativen Anzahl an Fällen zufällig zuweisen.** Geben Sie die relative Anzahl (Verhältnis) der Fälle an, die den einzelnen Stichproben (Training, Test, und Holdout) nach dem Zufallsprinzip zugewiesen werden sollen. Die Spalte % gibt auf der Grundlage der von Ihnen angegebenen Werte für die relative Anzahl den Prozentsatz der Fälle an, die den einzelnen Stichproben zugewiesen werden.

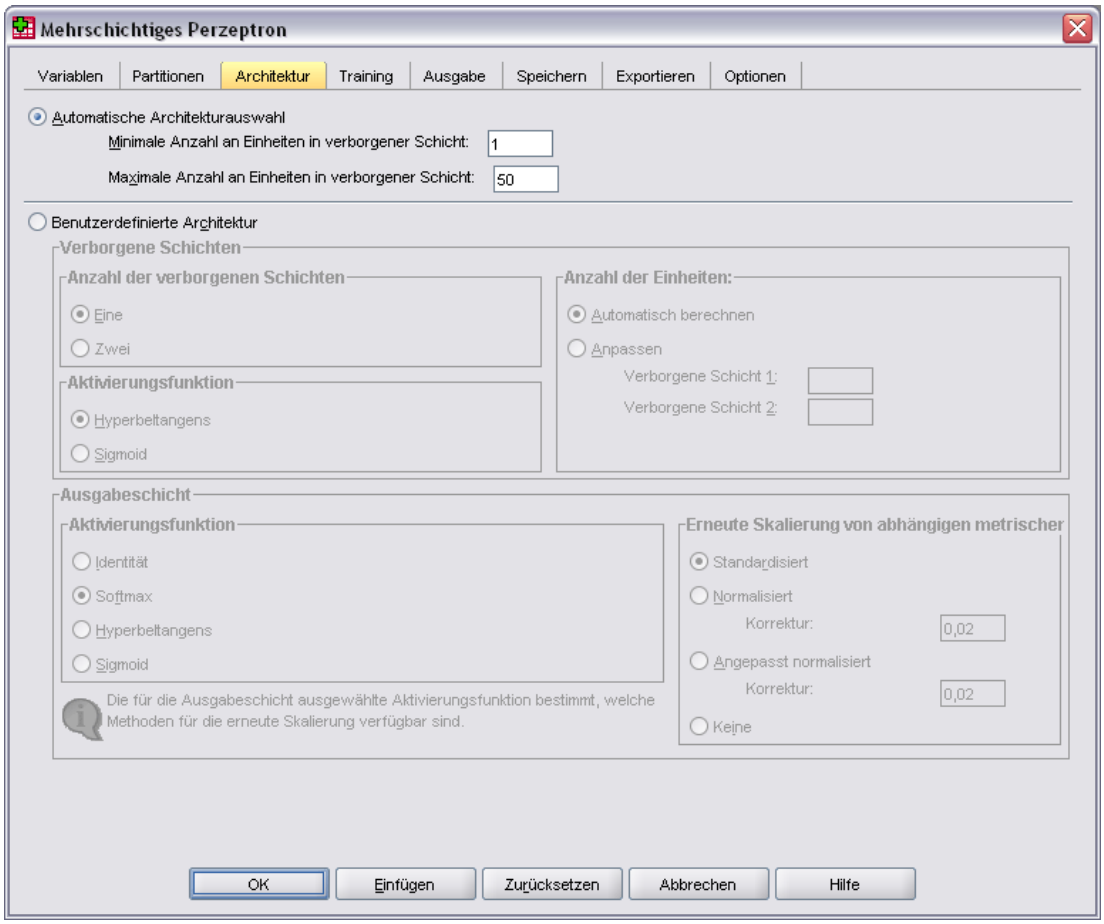
  Die Angabe von 7, 3, 0 als relative Anzahl für Training-, Test- und Holdout-Stichprobe entspricht 70 %, 30 % und 0 %. Die Angabe von 2, 1, 1 als Werte für die relative Anzahl entspricht 50 %, 25 % und 25 %; 1, 1, 1 entspricht der Aufteilung des Daten-Sets in drei gleich große Teile für Training, Test und Holdout.

- **Partitionsvariable zum Zuweisen von Fällen verwenden.** Geben Sie eine numerische Variable an, die jeden Fall in der Arbeitsdatei der Trainings-, Test bzw. Holdout-Stichprobe zuweist. Fälle mit einem positiven Wert für die Variable werden der Trainingsstichprobe zugewiesen, Fälle mit dem Wert 0 der Teststichprobe und Fälle mit einem negativen Wert der Holdout-Stichprobe. Fälle mit einem systemdefiniert fehlenden Wert werden aus der Analyse ausgeschlossen. Alle benutzerdefiniert fehlenden Werte für die Partitionsvariable werden immer als gültig behandelt.

  Anmerkung: Die Verwendung einer Partitionsvariablen garantiert keine identischen Ergebnisse bei späteren Ausführungen der Prozedur. Weitere Informationen finden Sie unter "Reproduzieren der Ergebnisse" im Thema Mehrschichtiges Perzeptron.

Architektur

Abbildung 2-3
Mehrschichtiges Perzeptron: Registerkarte "Architektur"



Auf der Registerkarte "Architektur" können Sie die Struktur des Netzwerks angeben. Die Prozedur kann automatisch die "beste" Architektur auswählen, Sie können aber auch eine benutzerdefinierte Architektur angeben.

Mit der automatischen Architekturauswahl wird ein Netzwerk mit genau einer verborgenen Schicht erstellt. Geben Sie die Mindest- und die Höchstzahl an Einheiten an, die in der verborgenen Schicht zulässig sein sollen. Die automatische Architekturauswahl berechnet daraus die "beste" Anzahl an Einheiten in der verborgenen Schicht. Die automatische Architekturauswahl verwendet die standardmäßigen Aktivierungsfunktionen für die verborgene Schichten und Ausgabeschichten.

Mit der benutzerdefinierten Architekturauswahl verfügen Sie über umfassende Kontrolle über die verborgenen Schichten und Ausgabeschichten. Dies ist insbesondere dann von Vorteil, wenn Sie im Voraus wissen, welche Architektur Sie wünschen, oder um eine Feinabstimmung der Ergebnisse der automatischen Architekturauswahl vorzunehmen.

Verborgene Schichten

Die verborgene Schicht enthält nicht sichtbare Netzwerkknoten (Einheiten). Jede verborgene Schicht ist eine Funktion der gewichteten Summe der Eingaben. Bei der Funktion handelt es sich um die Aktivierungsfunktion und die Werte der Gewichte richten sich nach dem Schätzungsalgorithmus. Wenn das Netzwerk eine zweite verborgene Schicht enthält, ist jede verborgene Einheit in der zweiten Schicht eine Funktion der gewichteten Summe der Einheiten in der ersten verborgenen Schicht. In beiden Schichten wird dieselbe Aktivierungsfunktion verwendet.

Anzahl der verborgenen Schichten. Ein mehrschichtiges Perzeptron kann eine oder zwei verborgene Schichten enthalten.

Aktivierungsfunktion. Die Aktivierungsfunktion "verknüpft" die gewichteten Summen der Einheiten in einer Schicht mit den Werten der Einheiten in der nachfolgenden Schicht.

- **Hyperbeltangens.** Diese Funktion weist folgende Form auf: $\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$. Sie verwendet Argumente mit reellen Werten und transformiert sie in den Bereich (–1, 1). Bei Verwendung der automatischen Architekturauswahl wird diese Aktivierungsfunktion für alle Einheiten in den verborgenen Schichten verwendet.
- **Sigmoid.** Diese Funktion weist folgende Form auf: $\gamma(c) = 1/(1+e^{-c})$. Sie verwendet Argumente mit reellen Werten und transformiert sie in den Bereich (0, 1).

Anzahl der Einheiten. Die Anzahl der Einheiten in den einzelnen verborgenen Schichten kann explizit festgelegt oder automatisch durch den Schätzalgorithmus bestimmt werden.

Ausgabeschicht

Die Ausgabeschicht enthält die Zielvariablen (abhängigen Variablen).

Aktivierungsfunktion. Die Aktivierungsfunktion "verknüpft" die gewichteten Summen der Einheiten in einer Schicht mit den Werten der Einheiten in der nachfolgenden Schicht.

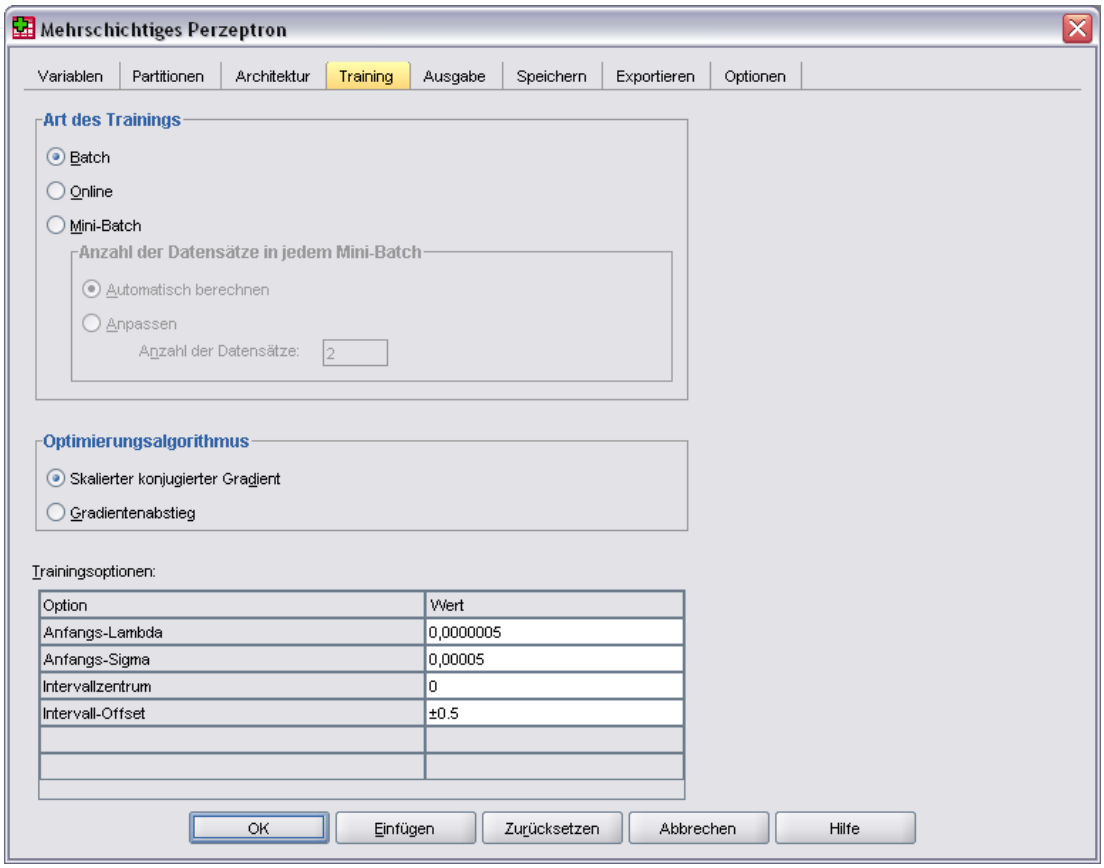
- **Identität.** Diese Funktion weist folgende Form auf: $\gamma(c) = c$. Sie verwendet Argumente mit reellen Werten und gibt sie unverändert wieder aus. Bei Verwendung der automatischen Architekturauswahl wird diese Aktivierungsfunktion für Einheiten in der Ausgabeschicht verwendet, sofern metrische abhängige Variablen vorliegen.
- **Softmax.** Diese Funktion weist folgende Form auf: $\gamma(c_k) = \exp(c_k)/\Sigma_j\exp(c_j)$. Sie verwendet einen Vektor von Argumenten mit reellen Werten und transformiert ihn in einen Vektor, dessen Elemente in den Bereich (0, 1) fallen und als Summe 1 ergeben. Softmax ist nur verfügbar, wenn alle abhängigen Variablen kategorial sind. Bei Verwendung der automatischen Architekturauswahl wird diese Aktivierungsfunktion für Einheiten in der Ausgabeschicht verwendet, sofern alle abhängigen Variablen kategorial sind.
- **Hyperbeltangens.** Diese Funktion weist folgende Form auf: $\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$. Sie verwendet Argumente mit reellen Werten und transformiert sie in den Bereich (–1, 1).
- **Sigmoid.** Diese Funktion weist folgende Form auf: $\gamma(c) = 1/(1+e^{-c})$. Sie verwendet Argumente mit reellen Werten und transformiert sie in den Bereich (0, 1).

Erneute Skalierung von abhängigen metrischen Variablen. Diese Steuerelemente sind nur verfügbar, wenn mindestens eine metrische abhängige Variable ausgewählt wurde.

- **Standardisiert.** Subtraktion des Mittelwerts und Division durch die Standardabweichung, (x−Mittelwert)/s.
- **Normalisiert.** Subtraktion des Mittelwerts und Division durch den Bereich, (x−min)/(max−min). Normalisierte Werte liegen zwischen 0 und 1. Dies ist die erforderliche Neuskalierungsmethode für metrische abhängige Variablen, wenn bei der Ausgabeschicht die Aktivierungsfunktion "Sigmoid" verwendet wird. Die Korrekturoption gibt eine kleinen ε-Wert an, der als Korrektur der Neuskalierungsformel verwendet wird. Durch diese Korrektur wird sichergestellt, dass alle neu skalierten Werte abhängiger Variablen innerhalb des Bereichs der Aktivierungsfunktion liegen. Insbesondere definieren die Werte 0 und 1, die in der unkorrigierten Formel vorkommen, wenn x den Mindest- bzw. Höchstwert annimmt, zwar die Grenzen des Bereichs der Sigmoid-Funktion, liegen jedoch nicht innerhalb dieses Bereichs. Die korrigierte Formel lautet [x−(min−ε)]/[(max+ε)−(min−ε)]. Geben Sie eine Zahl größer oder gleich 0 ein.
- **Angepasst normalisiert.** Angepasste Version der Subtraktion des Mittelwerts und Division durch den Bereich, [2*(x−min)/(max−min)]−1. Angepasste normalisierte Werte liegen zwischen −1 und 1. Dies ist die erforderliche Neuskalierungsmethode für metrische abhängige Variablen, wenn bei der Ausgabeschicht die Aktivierungsfunktion "Hyperbeltangens" verwendet wird. Die Korrekturoption gibt eine kleinen ε-Wert an, der als Korrektur der Neuskalierungsformel verwendet wird. Durch diese Korrektur wird sichergestellt, dass alle neu skalierten Werte abhängiger Variablen innerhalb des Bereichs der Aktivierungsfunktion liegen. Insbesondere definieren die Werte −1 und 1, die in der unkorrigierten Formel vorkommen, wenn x den Mindest- bzw. Höchstwert annimmt, zwar die Grenzen des Bereichs der Hyperbeltangens-Funktion, liegen jedoch nicht innerhalb dieses Bereichs. Die korrigierte Formel lautet {2*[(x−(min−ε))/((max+ε)−(min−ε))]}−1. Geben Sie eine Zahl größer oder gleich 0 an.
- **Keine.** Keine Neuskalierung metrischer abhängiger Variablen.

Training

Abbildung 2-4
Mehrschichtiges Perzeptron: Registerkarte "Training"



Auf der Registerkarte "Training" können Sie angeben, wie das Netzwerk trainiert werden sollte. Die Art des Trainings und der Optimierungsalgorithmus bestimmen, welche Trainingsoptionen verfügbar sind.

Art des Trainings. Die Art des Trainings bestimmt, wie das Netzwerk die Datensätze verarbeitet. Wählen Sie eine der folgenden Trainingsarten:

- **Batch.** Aktualisiert die synaptischen Gewichtungen erst nach dem Durchlauf sämtlicher Trainingsdatensätze. Beim Batch-Training werden also die Daten aus allen Datensätzen im Trainings-Daten-Set verwendet. Batch-Training wird häufig bevorzugt, da damit der Gesamtfehler unmittelbar minimiert wird. Allerdings kann beim Batch-Training eine sehr häufige Aktualisierung der Gewichtungen erforderlich sein, bis eine der Abbruchregeln erfüllt ist, sodass sehr viele Datendurchläufe notwendig sein können. Es eignet sich vor allem für "kleinere" Daten-Sets.
- **Online.** Aktualisiert die synaptischen Gewichtungen nach jedem einzelnen Trainingsdatensatz. Beim Online-Training werden also jeweils immer nur die Daten aus einem einzigen Datensatz verwendet. Das Online-Training ruft ständig einen Datensatz ab und aktualisiert die Gewichtungen, bis eine der Abbruchregeln erfüllt ist. Wenn alle Datensätze einmal verwendet

wurden und keine der Abbruchregeln erfüllt ist, wird der Prozess mit einem erneuten Durchlauf der Datensätze fortgesetzt. Online-Training ist dem Batch-Training bei "größeren" Daten-Sets mit zugeordneten Einflussvariablen vorzuziehen. Wenn also viele Datensätze und viele Eingaben vorliegen und ihre Werte nicht voneinander unabhängig sind, kann das Online-Training schneller zu einer brauchbaren Antwort führen als das Batch-Training.

- **Mini-Batch.** Unterteilt die Trainingsdatensätze in ungefähr gleich große Gruppen und aktualisiert dann die synaptischen Gewichtungen jeweils nach dem Durchlauf einer Gruppe. Beim Mini-Batch-Training werden also Informationen aus einer Gruppe von Datensätzen verwendet. Anschließend wird die Datengruppe, falls erforderlich, erneut verwendet. Mini-Batch-Training stellt einen Kompromiss zwischen Batch-Training und Online-Training dar und eignet sich am besten für "mittelgroße" Daten-Sets. Die Prozedur kann die Anzahl der Trainingsdatensätze pro Mini-Batch automatisch festlegen. Sie können jedoch auch eine ganze Zahl größer 1 und kleiner oder gleich der maximalen Anzahl der im Arbeitsspeicher zu speichernden Fälle angeben. Die maximale Anzahl der im Arbeitsspeicher zu speichernden Fälle können Sie auf der Registerkarte Optionen festlegen.

Optimierungsalgorithmus. Diese Methode wird zur Schätzung der synaptischen Gewichtungen verwendet.

- **Skalierter konjugierter Gradient.** Die Annahmen, die eine Verwendung von Methoden mit konjugiertem Gradienten rechtfertigen, gelten nur für das Batch-Training. Diese Methode steht also für Online- und Mini-Batch-Training nicht zur Verfügung.
- **Gradientenabstieg.** Diese Methode muss nur beim Online- bzw. Mini-Batch-Training verwendet werden. Auch beim Batch-Training kann sie eingesetzt werden.

Trainingsoptionen. Die Trainingsoptionen ermöglichen eine Feinabstimmung des Optimierungsalgorithmus. Im Allgemeinen müssen Sie diese Einstellungen nur ändern, wenn beim Netzwerk Probleme mit der Schätzung auftreten.

Folgende Trainingsoptionen stehen für den Algorithmus mit skaliertem konjugiertem Gradienten zur Verfügung:

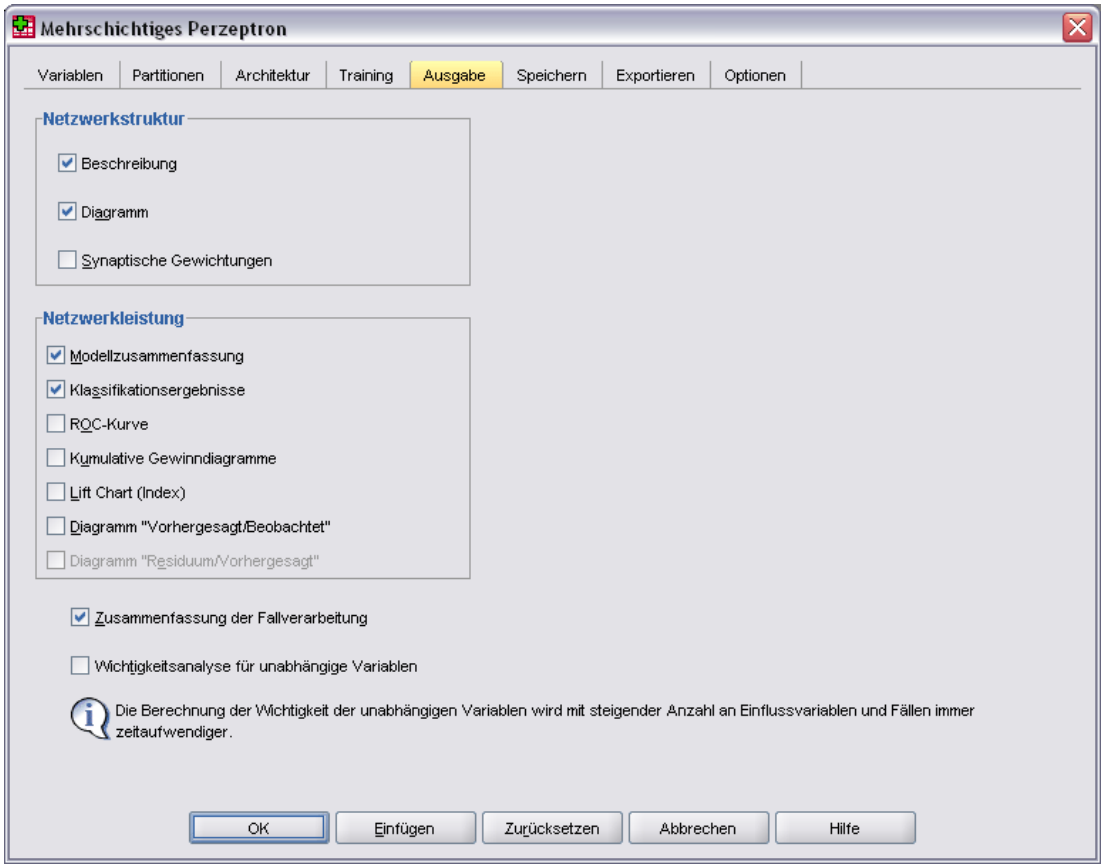
- **Anfangs-Lambda.** Der ursprüngliche Wert des Lambda-Parameters für den Algorithmus mit skaliertem konjugiertem Gradienten. Geben Sie einen Wert größer als 0 und kleiner als 0,000001 ein.
- **Anfangs-Sigma.** Der ursprüngliche Wert des Sigma-Parameters für den Algorithmus mit skaliertem konjugiertem Gradienten. Geben Sie einen Wert größer als 0 und kleiner als 0,0001 ein.
- **Intervallzentrum und Intervall-Offset.** Intervallzentrum (a_0) und Intervall-Offset (a) definieren das Intervall $[a_0-a, a_0+a]$, in dem bei Verwendung der simulierten Abkühlung Gewichtungsvektoren nach dem Zufallsprinzip erstellt werden. Die simulierte Abkühlung wird verwendet, um während der Anwendung des Optimierungsalgorithmus aus einem lokalen Minimum ausbrechen zu können, um das globale Minimum zu finden. Dieser Ansatz wird bei der Gewichtungsinitialisierung und bei der automatischen Architekturauswahl verwendet. Geben Sie den Wert für das Intervallzentrum und einen Wert größer 0 für das Intervall-Offset an.

Folgende Trainingsoptionen stehen für den Gradientenabstiegsalgorithmus zur Verfügung:

- **Anfängliche Lernrate.** Der ursprüngliche Wert der Lernrate für den Gradientenabstiegsalgorithmus. Bei einer höheren Lernrate erfolgt das Training des Netzwerks schneller, kann jedoch möglicherweise instabil werden. Geben Sie einen Wert größer 0 an.
- **Untergrenze der Lernrate.** Die Untergrenze der Lernrate für den Gradientenabstiegsalgorithmus. Diese Einstellung gilt nur für Online-Training und Mini-Batch-Training. Geben Sie einen Wert ein, der größer als 0 und kleiner als die anfängliche Lernrate ist.
- **Momentum.** Der ursprüngliche Momentum-Parameter für den Gradientenabstiegsalgorithmus. Der Momentum-Term (Impulsterm) trägt zur Vermeidung von Instabilitäten bei, die durch eine zu hohe Lernrate verursacht werden. Geben Sie einen Wert größer 0 an.
- **Lernratenreduzierung, in Epochen.** Die Anzahl der Epochen (p) bzw. Datendurchläufe der Trainingsstichprobe, die zur Reduzierung der anfänglichen Lernrate auf die Untergrenze der Lernrate erforderlich sind, wenn beim Online- oder Mini-Batch-Training Gradientenabstieg verwendet wird. Dadurch können Sie den Faktor für den Lernratenverfall $\beta = (1/pK)*\ln(\eta_0/\eta_{niedrig})$ steuern. Dabei ist η_0 die anfängliche Lernrate, $\eta_{niedrig}$ ist die Untergrenze der Lernrate und K ist die Gesamtzahl der Mini-Batches (bzw. beim Online-Training die Anzahl der Trainingsdatensätze) im Trainings-Daten-Set. Geben Sie eine ganze Zahl größer 0 an.

Ausgabe

Abbildung 2-5
Mehrschichtiges Perzeptron: Registerkarte "Ausgabe"



Netzwerkstruktur. Zeigt zusammenfassende Informationen über das neuronale Netzwerk an.

- **Beschreibung.** Zeigt Informationen zum neuronalen Netzwerk an, einschließlich der folgenden: abhängige Variablen, Anzahl von Eingabe- und Ausgabeeinheiten, Anzahl der verborgenen Schichten und Einheiten und Aktivierungsfunktionen.
- **Diagramm.** Zeigt das Netzwerkdiagramm als nicht bearbeitbares Diagramm an. Beachten Sie: Mit steigender Anzahl an Kovariaten und Faktorstufen wird das Diagramm schwerer zu interpretieren.
- **Synaptische Gewichtungen.** Zeigt die Koeffizientenschätzer an, die die Beziehung zwischen den Einheiten in einer bestimmten Schicht und den Einheiten in der nächsten Schicht anzeigen. Die synaptischen Gewichtungen beruhen auf der Trainingsstichprobe, selbst wenn die Arbeitsdatei in Trainings-, Test- und Holdout-Daten partitioniert ist. Beachten Sie, dass die Anzahl der synaptischen Gewichtungen recht groß werden kann und dass diese Gewichtungen im Allgemeinen nicht zur Interpretation der Netzwerkergebnisse verwendet werden.

Netzwerkleistung. Zeigt die Ergebnisse an, die verwendet werden, um zu bestimmen, ob das Modell "gut" ist. *Anmerkung*: Die Diagramme in dieser Gruppe beruhen auf der Kombination aus Trainings- und Teststichprobe bzw. nur auf der Trainingsstichprobe, wenn keine Teststichprobe vorhanden ist.

- **Modellzusammenfassung.** Zeigt eine Zusammenfassung der Ergebnisse des neuronalen Netzwerks nach Partition und insgesamt an, einschließlich der folgenden Werte: Fehler, Relativer Fehler bzw. Prozentsatz der falschen Vorhersagen, zum Beenden des Trainings verwendete Abbruchregel und Trainingszeit.

  Bei Anwendung der Aktivierungsfunktion "Identität", "Sigmoid" bzw. "Hyperbeltangens" auf die Ausgabeschicht handelt es sich um den Quadratsummenfehler. Bei Anwendung der Aktivierungsfunktion "Softmax" auf die Ausgabeschicht handelt es sich um den Kreuzentropiefehler.

  Die relativen Fehler bzw. Prozentsätze der falschen Vorhersagen werden in Abhängigkeit von den Messniveaus der abhängigen Variablen angezeigt. Wenn eine abhängige Variable ein metrisches Messniveau aufweist, wird der durchschnittliche relative Gesamtfehler (relativ zum Mittelwertmodell) angezeigt. Wenn alle abhängigen Variablen kategorial sind, wird der durchschnittliche Prozentsatz der falschen Vorhersagen angezeigt. Die relativen Fehler bzw. Prozentsätze der falschen Vorhersagen werden jeweils für die einzelnen abhängigen Variablen angezeigt.
- **Klassifikationsergebnisse.** Zeigt eine Klassifikationsmatrix für die einzelnen kategorialen abhängigen Variablen (nach Partition und insgesamt) an. Jede Tabelle gibt für jede Kategorie abhängiger Variablen die Anzahl der korrekt und nicht korrekt klassifizierten Fälle an. Der Prozentsatz der Gesamtzahl der Fälle, die korrekt klassifiziert wurden, wird ebenfalls angegeben.
- **ROC-Kurve.** Zeigt eine ROC-Kurve (Receiver Operating Characteristic) für jede kategoriale abhängige Variable an. Außerdem wird eine Tabelle angezeigt, die die Fläche unter den einzelnen Kurven angibt. Bei jeder abhängigen Variablen zeigt das ROC-Diagramm jeweils genau eine Kurve für jede Kategorie an. Wenn die abhängige Variable zwei Kategorien aufweist, behandelt jede Kurve die fragliche Kategorie als positiven Zustand gegenüber der anderen Kategorie. Wenn die abhängige Variable mehr als zwei Kategorien aufweist, behandelt jede Kurve die fragliche Kategorie als positiven Zustand gegenüber allen anderen Kategorien.
- **Kumulatives Gewinndiagramm.** Zeigt für jede kategoriale abhängige Variable ein kumulatives Gewinndiagramm an. Die Anzeige einer Kurve für jede Kategorie der abhängigen Variablen verhält sich wie bei ROC-Kurven.
- **Lift Chart (Index).** Zeigt für jede kategoriale abhängige Variable einen Lift Chart an. Die Anzeige einer Kurve für jede Kategorie der abhängigen Variablen verhält sich wie bei ROC-Kurven.
- **Diagramm "Vorhergesagt/Beobachtet".** Zeigt für jede abhängige Variable ein Diagramm an, das die vorhergesagten Werte in Abhängigkeit von den beobachteten Werten angibt. Bei kategorialen abhängigen Variablen werden für jede Antwortkategorie gruppierte Boxplots der vorhergesagten Pseudo-Wahrscheinlichkeiten angezeigt, wobei die Kategorie der

beobachteten Antworten als Klumpenvariable fungiert. Bei metrischen abhängigen Variablen wird ein Streudiagramm angezeigt.

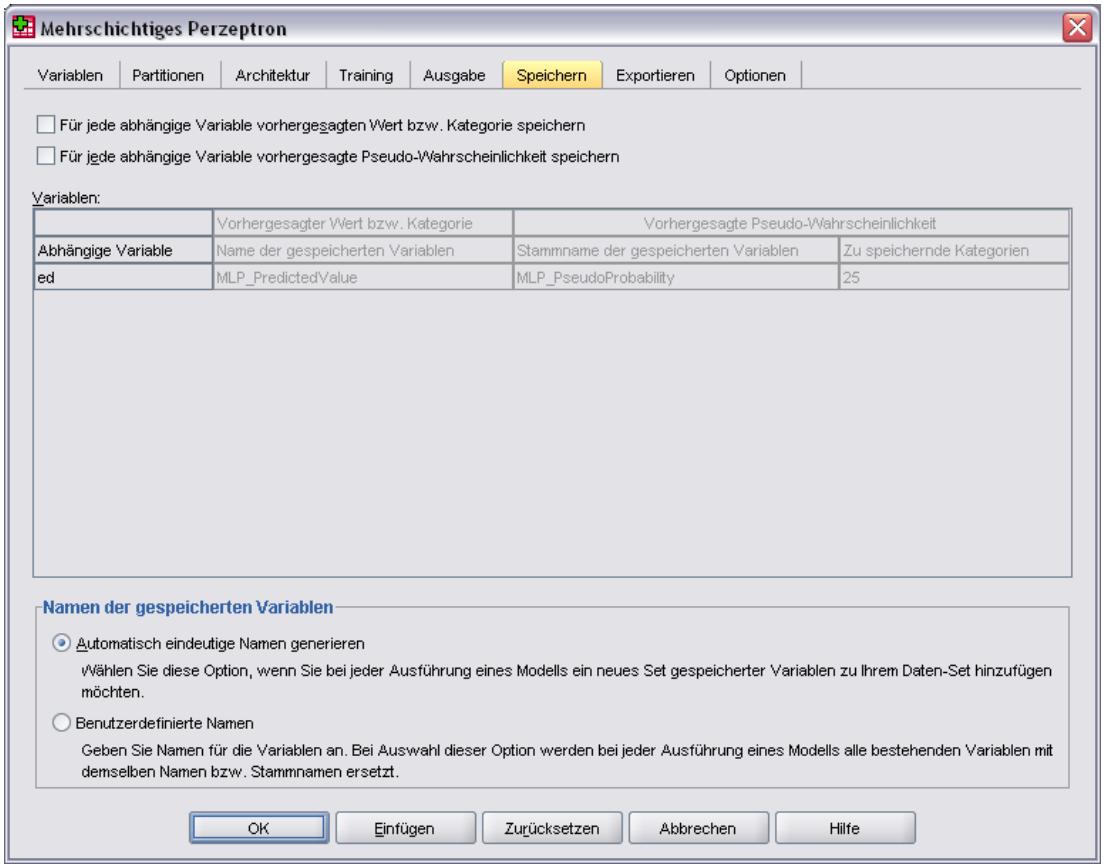
- **Diagramm "Residuum/Vorhergesagt".** Zeigt für jede metrische abhängige Variable ein Diagramm an, das die Residuen in Abhängigkeit von den vorhergesagten Werten angibt. Es sollte kein Muster zwischen Residuen und vorhergesagten Werten zu beobachten sein. Dieses Diagramm wird nur bei metrischen abhängigen Variablen erstellt.

Zusammenfassung der Fallverarbeitung. Zeigt die Tabelle mit der Zusammenfassung der Fallverarbeitung an, die die Anzahl der in der Analyse ein- und ausgeschlossenen Fälle zusammenfasst (insgesamt und nach Trainings-, Test- und Holdout-Stichprobe geordnet).

Wichtigkeitsanalyse für unabhängige Variablen. Führt eine Sensitivitätsanalyse durch, mit der die Wichtigkeit der einzelnen Einflussvariablen für die Bestimmung des neuronalen Netzwerks berechnet wird. Die Analyse beruht auf der Kombination aus Trainings- und Teststichprobe bzw. nur auf der Trainingsstichprobe, wenn keine Teststichprobe vorhanden ist. Dadurch werden eine Tabelle und ein Diagramm erstellt, die die Wichtigkeit und die normalisierte Wichtigkeit für die einzelnen Einflussvariablen anzeigen. Beachten Sie, dass die Sensitivitätsanalyse rechenintensiv und zeitaufwendig ist, wenn eine große Anzahl an Einflussvariablen oder Fällen vorliegt.

Speichern

Abbildung 2-6
Mehrschichtiges Perzeptron: Registerkarte "Speichern"



Auf der Registerkarte "Speichern" können Vorhersagen im Daten-Set als Variablen gespeichert werden.

- **Für jede abhängige Variable vorhergesagten Wert bzw. Kategorie speichern.** Damit wird bei metrischen abhängigen Variablen der vorhergesagte Wert und bei kategorialen abhängigen Variablen die vorhergesagte Kategorie gespeichert.
- **Für jede abhängige Variable vorhergesagte Pseudo-Wahrscheinlichkeit speichern.** Damit werden bei kategorialen abhängigen Variablen die vorhergesagten Pseudo-Wahrscheinlichkeiten gespeichert. Für die ersten *n* Kategorien wird eine separate Variable gespeichert. Dabei wird *n* in der Spalte Zu speichernde Kategorien angegeben.

Namen der gespeicherten Variablen. Durch eine automatische Generierung von Namen wird sichergestellt, dass Ihre Arbeit nicht verloren geht. Mit benutzerdefinierten Namen können Sie Ergebnisse aus früheren Durchgängen verwerfen/ersetzen, ohne zuerst die gespeicherten Variablen im Daten-Editor löschen zu müssen.

Wahrscheinlichkeiten und Pseudo-Wahrscheinlichkeiten

Kategoriale abhängige Variablen mit Softmax-Aktivierung und Kreuzentropiefehler weisen einen vorhergesagten Wert für jede Kategorie auf, wobei die einzelnen vorhergesagten Werte jeweils die Wahrscheinlichkeit angeben, dass der Fall zu der betreffenden Kategorie gehört.

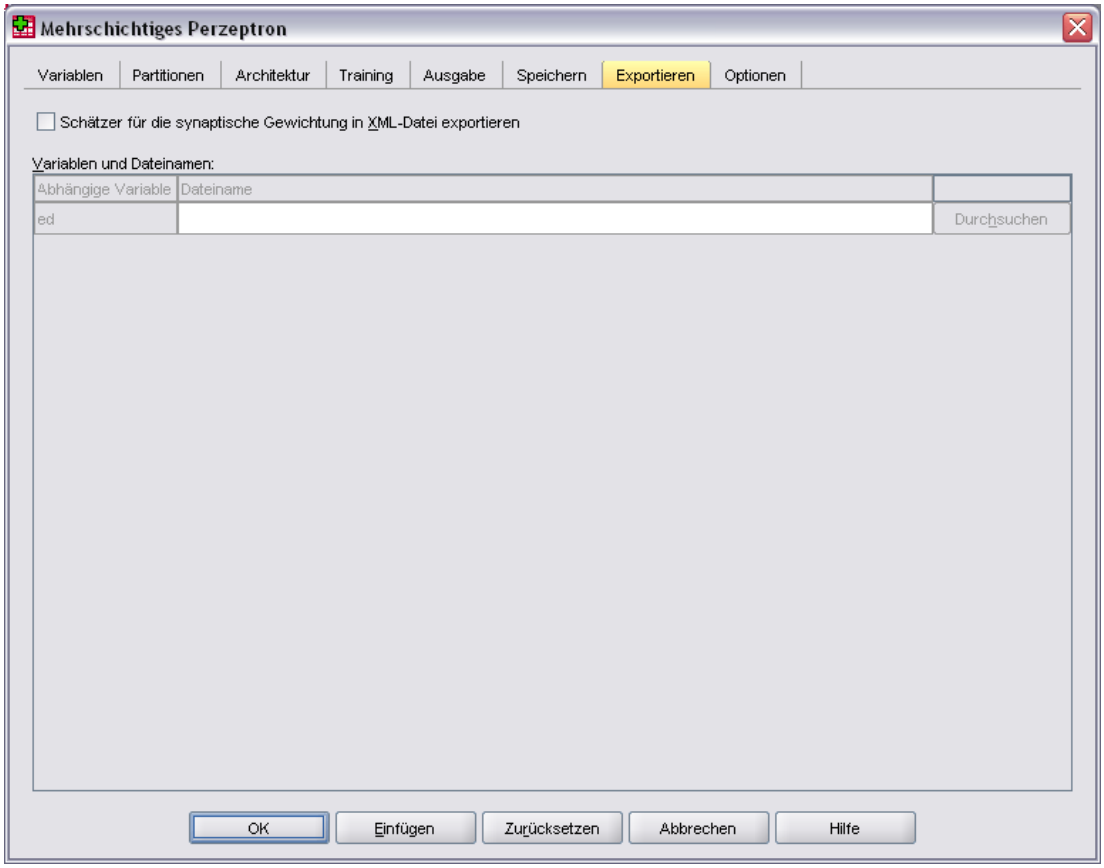
Kategoriale abhängige Variablen mit Quadratsummenfehler weisen einen vorhergesagten Wert für jede Kategorie auf, die vorhergesagten Werte können jedoch nicht als Wahrscheinlichkeiten interpretiert werden. Die Prozedur speichert diese vorhergesagten Pseudo-Wahrscheinlichkeiten, selbst wenn sie kleiner als 0 oder größer als 1 sind oder wenn die Summe für eine abhängige Variable nicht 1 ergibt.

ROC, kumulative Gewinne und Lift Charts (siehe Ausgabe auf S. 16) werden auf der Grundlage von Pseudo-Wahrscheinlichkeiten erstellt. Falls Pseudo-Wahrscheinlichkeiten kleiner als 0 oder größer als 1 sind oder die Summe für eine abhängige Variable nicht 1 ergibt, werden die Werte zunächst so neu skaliert, dass sie zwischen 0 und 1 liegen und als Summe 1 ergeben. Die Pseudo-Wahrscheinlichkeiten werden durch Division durch ihre Summe neu skaliert. Wenn ein Fall beispielsweise für eine abhängige Variable mit drei Kategorien vorhergesagte Pseudo-Wahrscheinlichkeiten von 0,50; 0,60 und 0,40 aufweist, wird jede Pseudo-Wahrscheinlichkeit durch die Summe 1,50 dividiert, woraus sich die Werte 0,33; 0,40 und 0,27 ergeben.

Wenn negative Pseudo-Wahrscheinlichkeiten vorliegen, werden vor der oben beschriebenen Neuskalierung allen Pseudo-Wahrscheinlichkeiten jeweils mit dem Betrag der niedrigsten Wahrscheinlichkeit addiert. Wenn die Pseudo-Wahrscheinlichkeiten beispielsweise -0,30, 0,50 und 1,30 betragen, müssen Sie zunächst 0,30 zu jedem Wert addieren und erhalten somit die Werte 0,00; 0,80 und 1,60. Als Nächstes dividieren Sie die einzelnen neuen Werte durch die Summe 2,40, wodurch sich die Werte 0,00; 0,33 und 0,67 ergeben.

Export

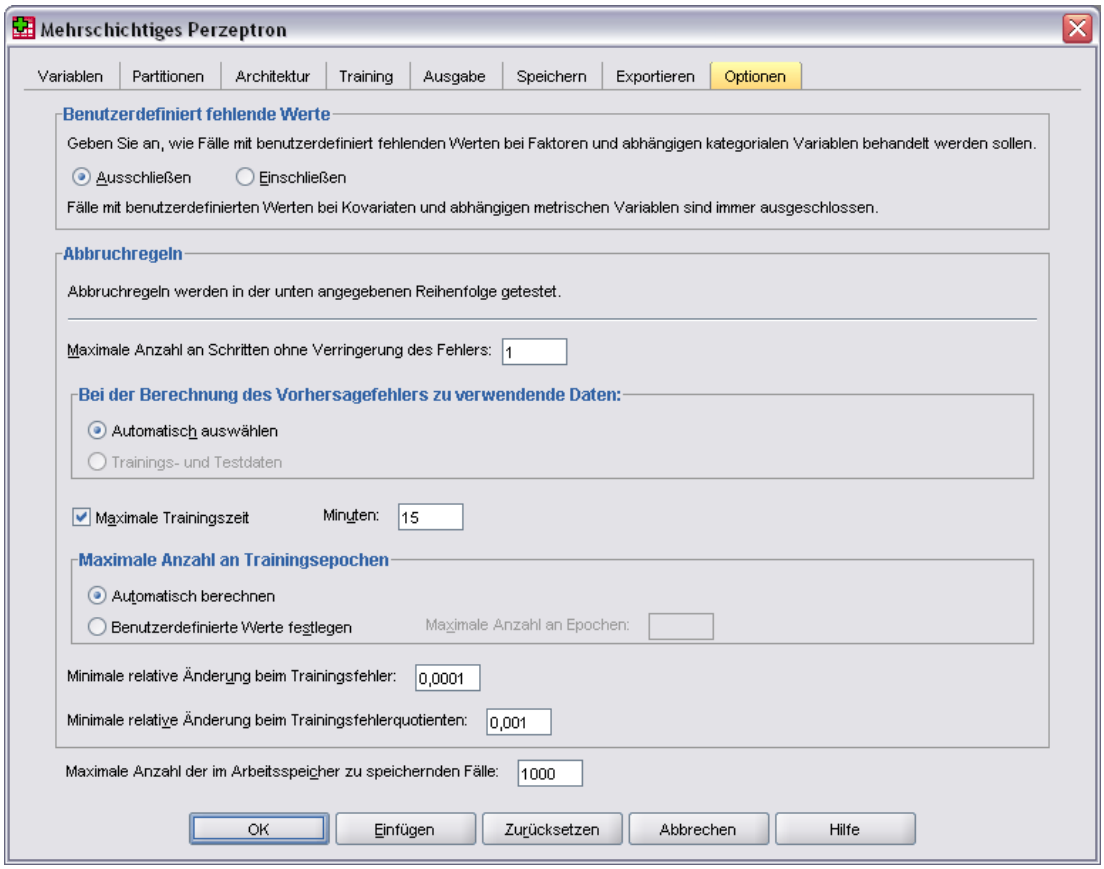
Abbildung 2-7
Mehrschichtiges Perzeptron: Registerkarte "Exportieren"



Die Registerkarte "Export" dient zum Speichern der Schätzer der synaptischen Gewichtungen für die einzelnen abhängigen Variablen in einer XML-Datei (PMML-Datei). SmartScore und SPSS Server (gesondertes Produkt) können anhand dieser Modelldatei die Modellinformationen zu Bewertungszwecken auf andere Datendateien anwenden. Diese Option ist nicht verfügbar, wenn aufgeteilte Dateien definiert wurden.

Optionen

Abbildung 2-8
Mehrschichtiges Perzeptron: Registerkarte "Optionen"



Benutzerdefinierte fehlende Werte. Faktoren müssen gültige Werte für einen Fall aufweisen, um in die Analyse aufgenommen zu werden. Mit diesen Steuerelementen legen Sie fest, ob benutzerdefiniert fehlende Werte bei den Faktoren und kategorialen abhängigen Variablen als gültige Werte behandelt werden sollen.

Abbruchregeln. Dies sind die Regeln, die festlegen, wann das Training des neuronalen Netzwerks abgebrochen werden soll. Das Training erfolgt über mindestens einen Datendurchlauf. Anschließend kann das Training gemäß den folgenden Kriterien beendet werden, die in der angegebenen Reihenfolge überprüft werden. In den folgenden Definitionen für Abbruchregeln entspricht ein Schritt bei den Methoden "Online" und "Mini-Batch" einem Datendurchlauf, bei der Batch-Methode einer Iteration.

- **Maximale Anzahl an Schritten ohne Verringerung des Fehlers.** Die Anzahl der Schritte, die zulässig sind, bevor eine Prüfung auf Verringerung des Fehlers erfolgt. Wenn nach der angegebenen Anzahl an Schritten keine Verringerung des Fehlers zu verzeichnen ist, wird das Training beendet. Geben Sie eine ganze Zahl größer 0 an. Außerdem können Sie angeben, welche Datenstichprobe zur Berechnung des Fehlers verwendet werden soll. Bei Automatisch auswählen wird die Teststichprobe verwendet, sofern vorhanden. Anderenfalls wird die

Trainingsstichprobe verwendet. Beachten Sie, dass beim Batch-Training der Fehler bei der Trainingsstichprobe garantiert nach jedem Datendurchlauf kleiner wird, daher kann diese Option nur auf das Batch-Training angewendet werden, wenn eine Teststichprobe vorhanden ist. Mit Trainings- und Testdaten wird der Fehler für jede dieser Stichproben geprüft; diese Option gilt nur, wenn eine Teststichprobe vorhanden ist.

Anmerkung: Nach jedem vollständigen Datendurchlauf ist beim Online- und Mini-Batch-Training ein zusätzlicher Datendurchlauf zur Berechnung des Trainingsfehlers erforderlich. Dieser zusätzliche Datendurchlauf kann das Training erheblich verlangsamen. Daher wird allgemein empfohlen, in jedem Fall eine Teststichprobe anzugeben und Automatisch auswählen zu verwenden.

- **Maximale Trainingszeit.** Wählen Sie aus, ob eine maximale Anzahl von Minuten für die Ausführung des Algorithmus angegeben werden soll. Geben Sie einen Wert größer 0 an.
- **Maximale Anzahl an Trainingsepochen.** Die maximal zulässige Anzahl an Epochen (Datendurchläufen). Wenn die maximale Anzahl an Epochen überschritten ist, wird das Training beendet. Geben Sie eine ganze Zahl größer 0 an.
- **Minimale relative Änderung beim Trainingsfehler.** Das Training wird beendet, wenn die relative Änderung beim Trainingfehler im Vergleich zum vorherigen Schritt kleiner ist als der Kriterienwert. Geben Sie eine Zahl größer 0 an. Beim Online- und Mini-Batch-Training wird dieses Kriterium ignoriert, wenn zur Berechnung des Fehlers ausschließlich Testdaten verwendet werden.
- **Minimale relative Änderung beim Trainingsfehlerquotienten.** Das Training wird beendet, wenn der Quotient aus Trainingfehler und Fehler des Nullmodells kleiner ist als der Kriterienwert. Das Nullmodell sagt den Durchschnittswert für alle abhängigen Variablen voraus. Geben Sie eine Zahl größer 0 an. Beim Online- und Mini-Batch-Training wird dieses Kriterium ignoriert, wenn zur Berechnung des Fehlers ausschließlich Testdaten verwendet werden.

Maximale Anzahl der im Arbeitsspeicher zu speichernden Fälle. Dadurch werden folgende Einstellungen innerhalb der Algorithmen mit mehrschichtigem Perzeptron gesteuert. Geben Sie eine ganze Zahl größer 1 an.

- Bei der automatischen Architekturauswahl beträgt die zur Bestimmung der Netzwerkarchitektur verwendete Stichprobe min(1000,*memsize*), wobei *memsize* die maximale Anzahl der im Arbeitsspeicher zu speichernden Fälle ist.
- Beim Mini-Batch-Training mit automatischer Berechnung der Anzahl an Mini-Batches, beträgt die Anzahl der Mini-Batches min(max(*M*/10,2),*memsize*), wobei *M* die Anzahl der Fälle in der Trainingsstichprobe ist.

Kapitel
3

Radiale Basisfunktion

Die Prozedur “Radiale Basisfunktion” (RBF) erstellt ein Vorhersagemodell für eine oder mehrere abhängige Variablen (Zielvariablen), das auf den Werten der Einflussvariablen beruht.

Beispiel. Ein Telekommunikationsanbieter hat seinen Kundenstamm nach Servicenutzungsmustern in vier Gruppen unterteilt hat. Mithilfe eines RBF-Netzwerks, das demografische Daten zur Vorhersage der Gruppenzugehörigkeit verwendet, kann das Unternehmen speziell angepasste Angebote für einzelne potenzielle Kunden entwickeln.

Abhängige Variablen. Die abhängigen Variablen können wie folgt gestaltet sein:

- **Nominal.** Eine Variable kann als nominal behandelt werden, wenn ihre Kategorien sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- **Ordinal.** Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.
- **Metrisch.** Eine Variable kann als metrisch behandelt werden, wenn ihre Werte geordnete Kategorien mit einer sinnvollen Metrik darstellen, sodass man sinnvolle Aussagen über die Abstände zwischen den Werten machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

  Bei der Prozedur wird davon ausgegangen, dass allen abhängigen Variablen das richtige Messniveau zugewiesen wurde. Sie können das Messniveau für eine Variable jedoch vorübergehend ändern. Klicken Sie hierzu mit der rechten Maustaste auf die Variable in der Liste der Quellvariablen und wählen Sie das gewünschte Messniveau im Kontextmenü aus.

Messniveau und Datentyp sind durch ein Symbol neben der jeweiligen Variablen in der Variablenliste gekennzeichnet:

Messniveau	**Datentyp**			
	Numerisch	String	Datum	Zeit
Metrisch		entfällt		

Ordinal				
Nominal				

Einflussvariablen. Einflussvariablen können als Faktoren (kategorial) oder als Kovariaten (metrisch) angegeben werden.

Kodierung für kategoriale Variablen. Die Prozedur kodiert vorübergehend für die Dauer des Verfahrens kategoriale Einflussvariablen und abhängige Variablen mithilfe der "Eins-aus-*c*"-Kodierung neu. Wenn es *c* Kategorien für eine Variable gibt, wird die Variable als c Vektoren gespeichert. Dabei wird die erste Kategorie als (1,0,...,0) angegeben, die zweite Kategorie als (0,1,0,...,0), ... und die letzte Kategorie als (0,0,...,0,1).

Dieses Kodierungsschema erhöht die Anzahl der synaptischen Gewichtungen und kann zu einer Verlangsamung des Trainings führen, "kompaktere" Kodierungsmethoden führen jedoch in der Regel zu neuronalen Netzwerken mit geringer Anpassungsgüte. Wenn das Training des Netzwerks sehr langsam vorangeht, können Sie versuchen, die Anzahl der Kategorien der kategorialen Einflussvariablen zu verringern, indem Sie ähnliche Kategorien zusammenfassen oder Fälle ausschließen, die extrem seltene Kategorien aufweisen.

Jegliche "Eins-aus-*c*"-Kodierung beruht auf den Trainingsdaten, selbst wenn eine Test- bzw. Holdout-Stichprobe definiert wurde (siehe Partitionen auf S. 27). Wenn also die Test- bzw. Holdout-Stichproben Fälle mit Einflussvariablen-Kategorien enthalten, die in den Trainingsdaten nicht vorhanden sind, werden diese Fälle nicht in der Prozedur bzw. beim Scoring verwendet. Wenn die Test- bzw. Holdout-Stichproben Fälle mit Kategorien abhängiger Variablen enthalten, die in den Trainingsdaten nicht vorhanden sind, werden diese Fälle zwar nicht in der Prozedur, jedoch möglicherweise beim Scoring verwendet.

Neuskalierung. Metrische abhängige Variablen und Kovariaten werden standardmäßig neu skaliert, um das Training des Netzwerks zu verbessern. Jegliche Neuskalierung beruht auf den Trainingsdaten, selbst wenn eine Test- bzw. Holdout-Stichprobe definiert wurde (siehe Partitionen auf S. 27). Das bedeutet, dass je nach Neuskalierungstyp Mittelwert, Standardabweichung, Mindestwert bzw. Höchstwert einer Kovariaten oder abhängigen Variablen ausschließlich anhand der Trainingsdaten berechnet werden. Wenn Sie eine Variable zur Festlegung von Partitionen angeben, müssen diese Kovariaten bzw. abhängigen Variablen in der Trainings-, Test- und Holdout-Stichprobe ähnliche Verteilungen aufweisen.

Häufigkeitsgewichtungen. Häufigkeitsgewichtungen werden von dieser Prozedur ignoriert.

Reproduzieren der Ergebnisse. Wenn Sie Ihre Ergebnisse exakt reproduzieren möchten, müssen Sie nicht nur dieselben Einstellungen für die Prozedur, sondern auch denselben Initialisierungswert für den Zufallszahlengenerator und dieselbe Datenreihenfolge verwenden. Weitere Details zu diesem Problem folgen:

- **Generierung von Zufallszahlen.** Die Prozedur verwendet Zufallszahlengenerierung während der Zufallszuweisung von Partitionen. Um zu einem späteren Zeitpunkt dieselben randomisierten Ergebnisse zu reproduzieren, müssen Sie vor jeder Ausführung der Prozedur "Radiale Basisfunktion" denselben Initialisierungswert für den Zufallszahlengenerator verwenden.

Einzelschrittanweisungen hierzu finden Sie unter Vorbereiten der Daten für die Analyse auf S. 75.

- **Fallreihenfolge.** Außerdem hängen die Ergebnisse von der Datenreihenfolge ab, da der Two-Step-Cluster-Algorithmus zur Ermittlung der radialen Basisfunktionen verwendet wird.

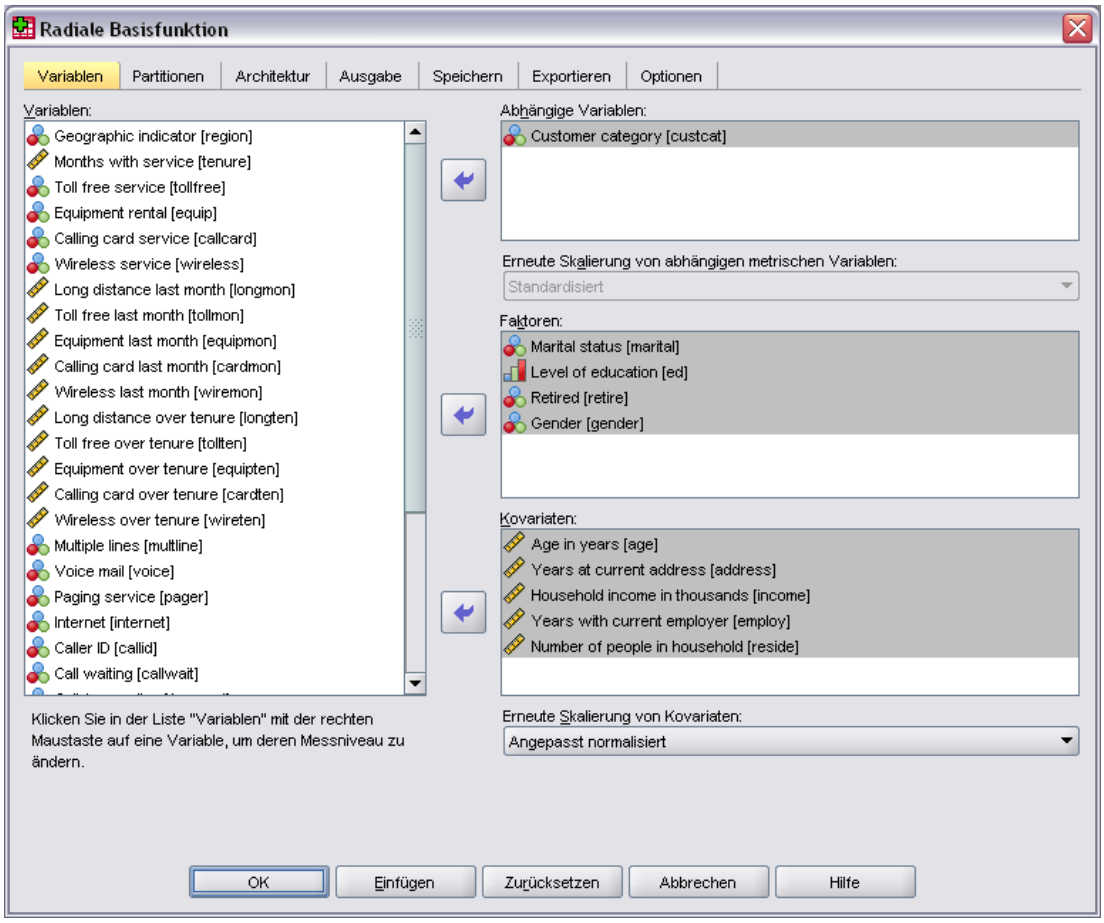
  Um die Auswirkungen der Reihenfolge zu minimieren, mischen Sie die Fälle in zufälliger Reihenfolge. Prüfen Sie daher die Stabilität einer bestimmten Lösung, indem Sie verschiedene Lösungen abrufen, bei denen die Fälle in einer unterschiedlichen, zufällig ausgewählten Reihenfolgen sortiert sind. In Situationen mit extrem umfangreichen Dateien können mehrere Durchgänge mit jeweils einer Stichprobe von Fällen durchgeführt werden, die in unterschiedlicher, zufällig ausgewählter Reihenfolge sortiert ist.

Erstellen eines Netzwerks mit radialen Basisifunktionen

Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Neuronale Netze
    Radiale Basisfunktion...

Abbildung 3-1
Radiale Basisfunktion: Registerkarte "Variablen"



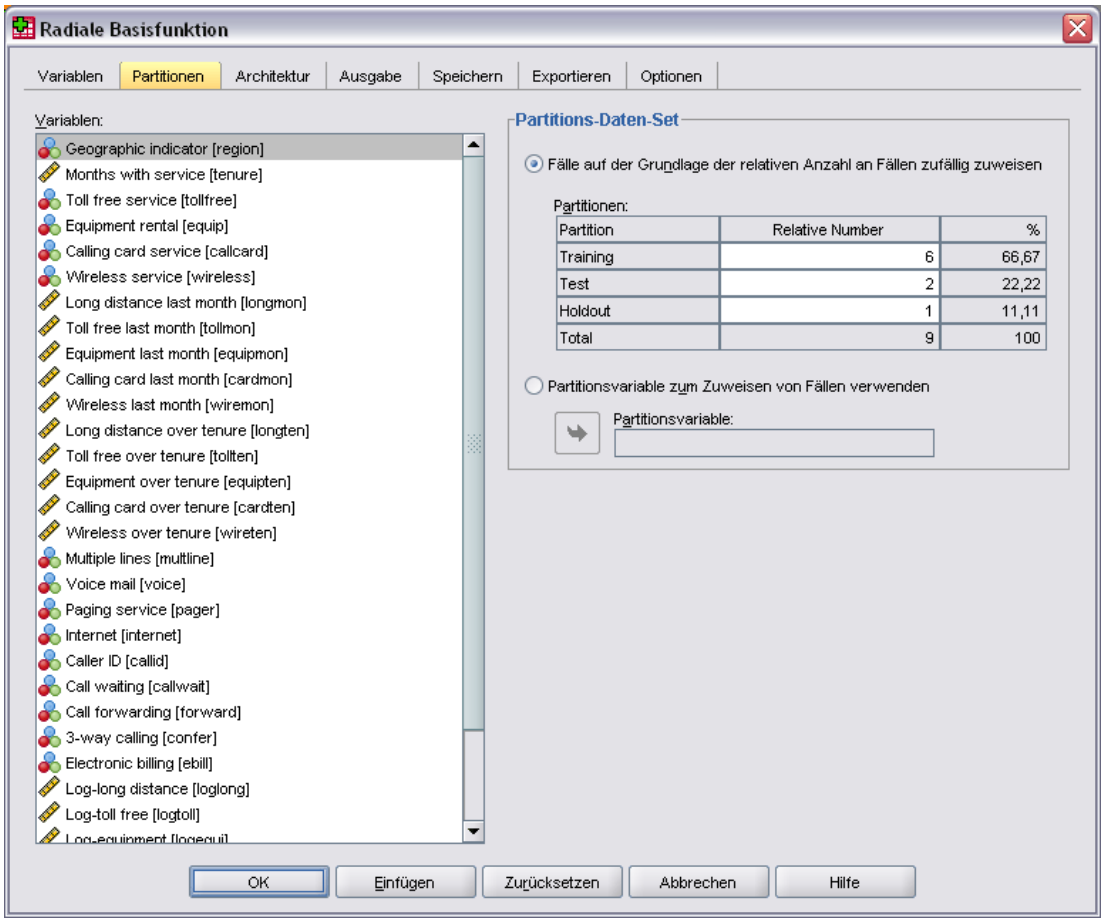
- Wählen Sie mindestens eine abhängige Variable aus.
- Wählen Sie mindestens einen Faktor oder eine Kovariate aus.

Optional können Sie auf der Registerkarte "Variablen" die Methode zur Neuskalierung der Kovariaten ändern. Folgende Optionen stehen zur Auswahl:

- **Standardisiert.** Subtraktion des Mittelwerts und Division durch die Standardabweichung, (*x*−Mittelwert)/*s*.
- **Normalisiert.** Subtraktion des Mittelwerts und Division durch den Bereich, (*x*−min)/(max−min). Normalisierte Werte liegen im Bereich zwischen 0 und 1.
- **Angepasst normalisiert.** Angepasste Version der Subtraktion des Mittelwerts und Division durch den Bereich, [2*(*x*−min)/(max−min)]−1. Angepasste normalisierte Werte liegen zwischen −1 und 1.
- **Keine.** Keine Neuskalierung der Kovariaten.

Partitionen

Abbildung 3-2
Radiale Basisfunktion: Registerkarte "Partitionen"



Partitions-Daten-Set. Diese Gruppe gibt die Methode zur Partitionierung der Arbeitsdatei in eine Trainings-, eine Test- und eine Holdout-Stichprobe an. Die **Trainingsstichprobe** umfasst die Datensätze, die zum Trainieren des neuronalen Netzwerks verwendet wurden; ein gewisser Prozentsatz der Fälle im Daten-Set muss der Trainingsstichprobe zugewiesen werden, um ein Modell zu erhalten. Die **Teststichprobe** ist ein unabhängiges Set von Datensätzen, die verwendet werden, um den Fehler während des Trainings aufzuzeichnen und dadurch ein Übertrainieren zu vermeiden. Es wird dringend empfohlen, eine Trainingsstichprobe zu erstellen. Das Netzwerktraining ist in der Regel am effizientesten, wenn die Teststichprobe kleiner ist als die Trainingsstichprobe. Die **Holdout-Stichprobe** ist ein weiterer unabhängiger Satz von Datensätzen, der zur Bewertung des endgültigen neuronalen Netzwerks verwendet wird; der Fehler für die Houldout-Stichprobe bietet eine "ehrliche" Schätzung der Vorhersagekraft des Modells, da die Prüffälle (die Fälle in der Holdout-Stichprobe) nicht zur Erstellung des Modells verwendet wurden.

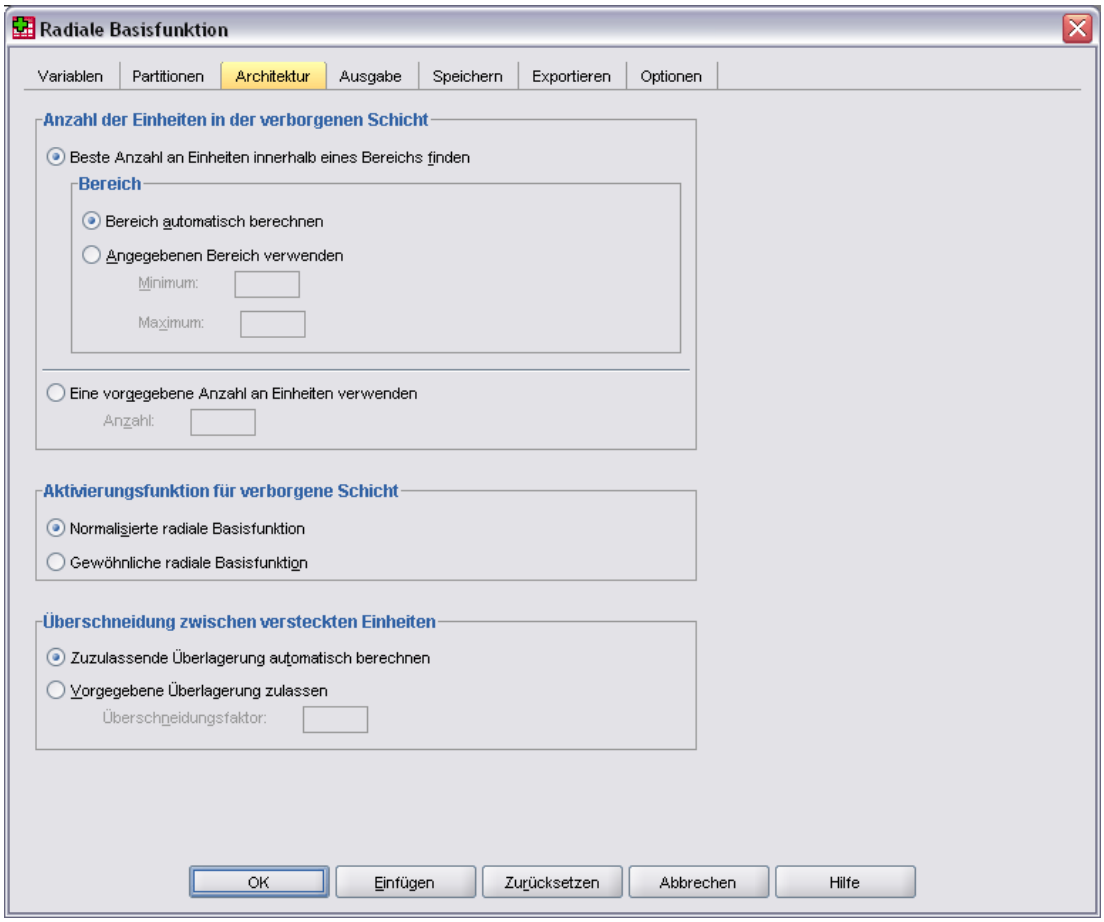
- **Fälle auf der Grundlage der relativen Anzahl an Fällen zufällig zuweisen.** Geben Sie die relative Anzahl (Verhältnis) der Fälle an, die den einzelnen Stichproben (Training, Test, und Holdout) nach dem Zufallsprinzip zugewiesen werden sollen. Die Spalte % gibt auf der Grundlage der von Ihnen angegebenen Werte für die relative Anzahl den Prozentsatz der Fälle an, die den einzelnen Stichproben zugewiesen werden.

  Die Angabe von 7, 3, 0 als relative Anzahl für Training-, Test- und Holdout-Stichprobe entspricht 70 %, 30 % und 0 %. Die Angabe von 2, 1, 1 als Werte für die relative Anzahl entspricht 50 %, 25 % und 25 %; 1, 1, 1 entspricht der Aufteilung des Daten-Sets in drei gleich große Teile für Training, Test und Holdout.

- **Partitionsvariable zum Zuweisen von Fällen verwenden.** Geben Sie eine numerische Variable an, die jeden Fall in der Arbeitsdatei der Trainings-, Test bzw. Holdout-Stichprobe zuweist. Fälle mit einem positiven Wert für die Variable werden der Trainingsstichprobe zugewiesen, Fälle mit dem Wert 0 der Teststichprobe und Fälle mit einem negativen Wert der Holdout-Stichprobe. Fälle mit einem systemdefiniert fehlenden Wert werden aus der Analyse ausgeschlossen. Alle benutzerdefiniert fehlenden Werte für die Partitionsvariable werden immer als gültig behandelt.

Architektur

Abbildung 3-3
Radiale Basisfunktion: Registerkarte "Architektur"



Auf der Registerkarte "Architektur" können Sie die Struktur des Netzwerks angeben. Diese Prozedur erstellt ein neuronales Netz mit genau einer verborgenen Schicht vom Typ "Radiale Basisfunktion". Normalerweise ist es nicht erforderlich, diese Einstellungen zu ändern.

Anzahl der Einheiten in der verborgenen Schicht. Es gibt drei Möglichkeiten zur Auswahl der Anzahl der verborgenen Einheiten.

1. **Beste Anzahl an Einheiten innerhalb eines automatisch berechneten Bereichs finden.** Die Prozedur berechnet automatisch den Mindest- und Höchstwert des Bereichs und ermittelt die beste Anzahl an verborgenen Einheiten innerhalb des Bereichs.

   Wenn eine Teststichprobe definiert wurde, verwendet die Prozedur das Testdatenkriterium: Die beste Anzahl an verborgenen Einheiten ist diejenige, die zum kleinsten Fehler bei den Testdaten führt. Wenn keine Teststichprobe definiert wurde, verwendet die Prozedur das Bayes-Informationskriterium (BIC): Die beste Anzahl an verborgenen Einheiten ist diejenige, die auf der Grundlage der Trainingsdaten zum kleinsten BIC führt.

2. **Beste Anzahl an Einheiten innerhalb eines angegebenen Bereichs finden.** Sie können selbst einen Bereich angeben und die Prozedur ermittelt die "beste" Anzahl an verborgenen Einheiten innerhalb dieses Bereichs. Wie zuvor wird auch hier die beste Anzahl an verborgenen Einheiten im Bereich mithilfe des Testdatenkriteriums bzw. des Bayes-Informationskriteriums (BIC) ermittelt.

3. **Eine vorgegebene Anzahl an Einheiten verwenden.** Sie können die Verwendung eines Bereichs außer Kraft setzen und stattdessen direkt eine bestimmte Anzahl an Einheiten eingeben.

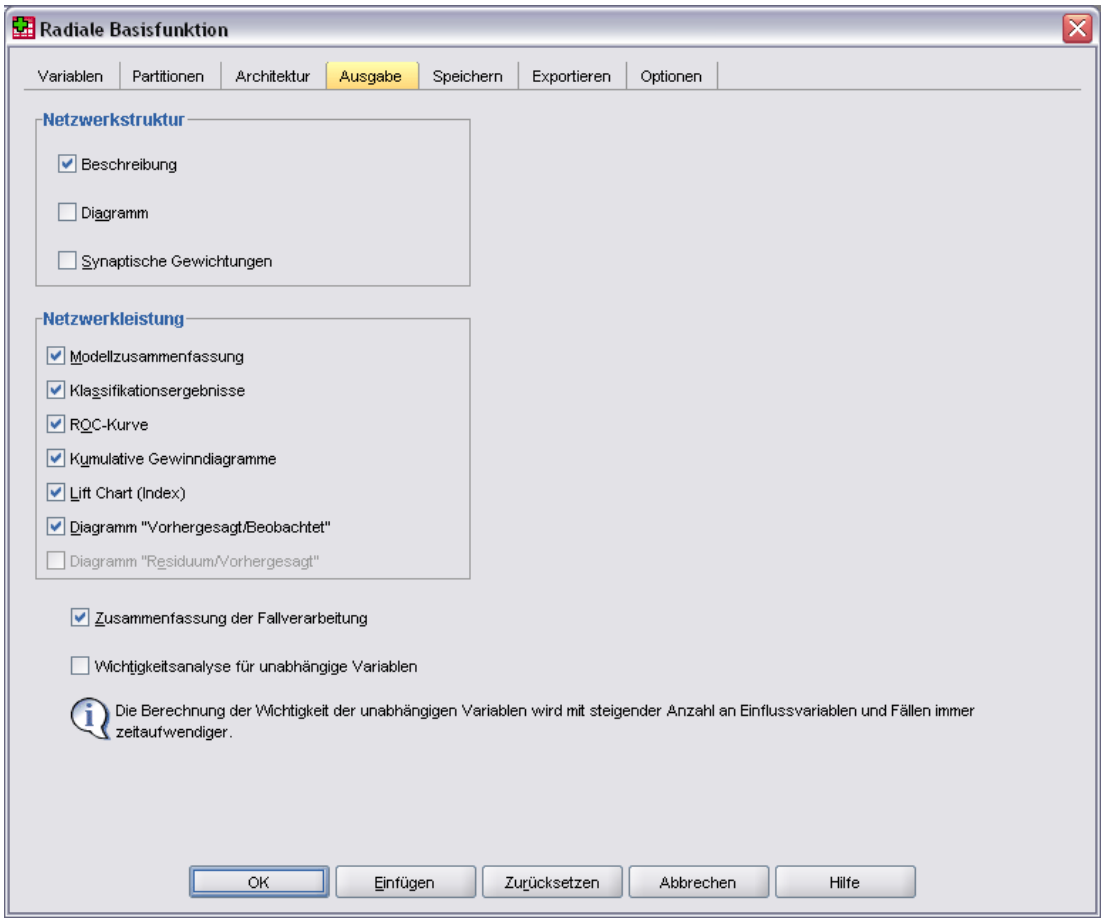
Aktivierungsfunktion für verborgene Schicht. Die Aktivierungsfunktion für die verborgene Schicht ist die radiale Basisfunktion, die die Einheiten in einer Schicht mit den Werten der Einheiten in der vorhergehenden Schicht "verknüpft". Bei der Ausgabeschicht dient die Identitätsfunktion als Aktivierungsfunktion. Die Ausgabeeinheiten sind also einfach gewichtete Summen der verborgenen Einheiten.

- **Normalisierte radiale Basisfunktion.** Verwendet die Aktivierungsfunktion "Softmax", sodass die Aktivierungen aller verborgenen Einheiten so normalisiert werden, dass ihre Summe 1 ergibt.
- **Gewöhnliche radiale Basisfunktion.** Verwendet die exponentielle Aktivierungsfunktion, sodass die Aktivierung der verborgenen Einheit eine Gaußglocke als Funktion der Eingaben darstellt.

Überschneidung zwischen versteckten Einheiten. Der Überschneidungsfaktor ist ein Multiplikator, der auf die Breite der radialen Basisfunktionen angewendet wird. Der automatisch berechnete Wert des Überschneidungsfaktors lautet 1+0,1*d*, wobei *d* die Anzahl der Eingabeeinheiten ist (die Summe aus der Anzahl an Kategorien in allen Faktoren und der Anzahl der Kovariaten).

Ausgabe

Abbildung 3-4
Radiale Basisfunktion: Registerkarte "Ausgabe"



Netzwerkstruktur. Zeigt zusammenfassende Informationen über das neuronale Netzwerk an.

- **Description (Beschreibung).** Zeigt Informationen zum neuronalen Netzwerk an, einschließlich der folgenden: abhängige Variablen, Anzahl von Eingabe- und Ausgabeeinheiten, Anzahl der verborgenen Schichten und Einheiten und Aktivierungsfunktionen.
- **Diagramm.** Zeigt das Netzwerkdiagramm als nicht bearbeitbares Diagramm an. Beachten Sie: Mit steigender Anzahl an Kovariaten und Faktorstufen wird das Diagramm schwerer zu interpretieren.
- **Synaptische Gewichtungen.** Zeigt die Koeffizientenschätzer an, die die Beziehung zwischen den Einheiten in einer bestimmten Schicht und den Einheiten in der nächsten Schicht anzeigen. Die synaptischen Gewichtungen beruhen auf der Trainingsstichprobe, selbst wenn die Arbeitsdatei in Trainings-, Test- und Holdout-Daten partitioniert ist. Beachten Sie, dass die Anzahl der synaptischen Gewichtungen recht groß werden kann und dass diese Gewichtungen im Allgemeinen nicht zur Interpretation der Netzwerkergebnisse verwendet werden.

Netzwerkleistung. Zeigt die Ergebnisse an, die verwendet werden, um zu bestimmen, ob das Modell "gut" ist. *Anmerkung*: Die Diagramme in dieser Gruppe beruhen auf der Kombination aus Trainings- und Teststichprobe bzw. nur auf der Trainingsstichprobe, wenn keine Teststichprobe vorhanden ist.

- **Modellzusammenfassung.** Zeigt eine Zusammenfassung der Ergebnisse des neuronalen Netzwerks nach Partition und insgesamt an, einschließlich der folgenden Werte: Fehler, Relativer Fehler bzw. Prozentsatz der falschen Vorhersagen und Trainingszeit.

  Der Fehler ist der Quadratsummenfehler. Außerdem werden die relativen Fehler bzw. Prozentsätze der falschen Vorhersagen in Abhängigkeit von den Messniveaus der abhängigen Variablen angezeigt. Wenn eine abhängige Variable ein metrisches Messniveau aufweist, wird der durchschnittliche relative Gesamtfehler (relativ zum Mittelwertmodell) angezeigt. Wenn alle abhängigen Variablen kategorial sind, wird der durchschnittliche Prozentsatz der falschen Vorhersagen angezeigt. Die relativen Fehler bzw. Prozentsätze der falschen Vorhersagen werden jeweils für die einzelnen abhängigen Variablen angezeigt.

- **Klassifikationsergebnisse.** Zeigt für jede kategoriale abhängige Variable eine Klassifikationsmatrix an. Jede Tabelle gibt für jede Kategorie abhängiger Variablen die Anzahl der korrekt und nicht korrekt klassifizierten Fälle an. Der Prozentsatz der Gesamtzahl der Fälle, die korrekt klassifiziert wurden, wird ebenfalls angegeben.

- **ROC-Kurve.** Zeigt eine ROC-Kurve (Receiver Operating Characteristic) für jede kategoriale abhängige Variable an. Außerdem wird eine Tabelle angezeigt, die die Fläche unter den einzelnen Kurven angibt. Bei jeder abhängigen Variablen zeigt das ROC-Diagramm jeweils genau eine Kurve für jede Kategorie an. Wenn die abhängige Variable zwei Kategorien aufweist, behandelt jede Kurve die fragliche Kategorie als positiven Zustand gegenüber der anderen Kategorie. Wenn die abhängige Variable mehr als zwei Kategorien aufweist, behandelt jede Kurve die fragliche Kategorie als positiven Zustand gegenüber allen anderen Kategorien.

- **Kumulatives Gewinndiagramm.** Zeigt für jede kategoriale abhängige Variable ein kumulatives Gewinndiagramm an. Die Anzeige einer Kurve für jede Kategorie der abhängigen Variablen verhält sich wie bei ROC-Kurven.

- **Lift Chart (Index).** Zeigt für jede kategoriale abhängige Variable einen Lift Chart an. Die Anzeige einer Kurve für jede Kategorie der abhängigen Variablen verhält sich wie bei ROC-Kurven.

- **Diagramm "Vorhergesagt/Beobachtet".** Zeigt für jede abhängige Variable ein Diagramm an, das die vorhergesagten Werte in Abhängigkeit von den beobachteten Werten angibt. Bei kategorialen abhängigen Variablen werden für jede Antwortkategorie gruppierte Boxplots der vorhergesagten Pseudo-Wahrscheinlichkeiten angezeigt, wobei die Kategorie der beobachteten Antworten als Klumpenvariable fungiert. Bei metrischen abhängigen Variablen wird ein Streudiagramm angezeigt.

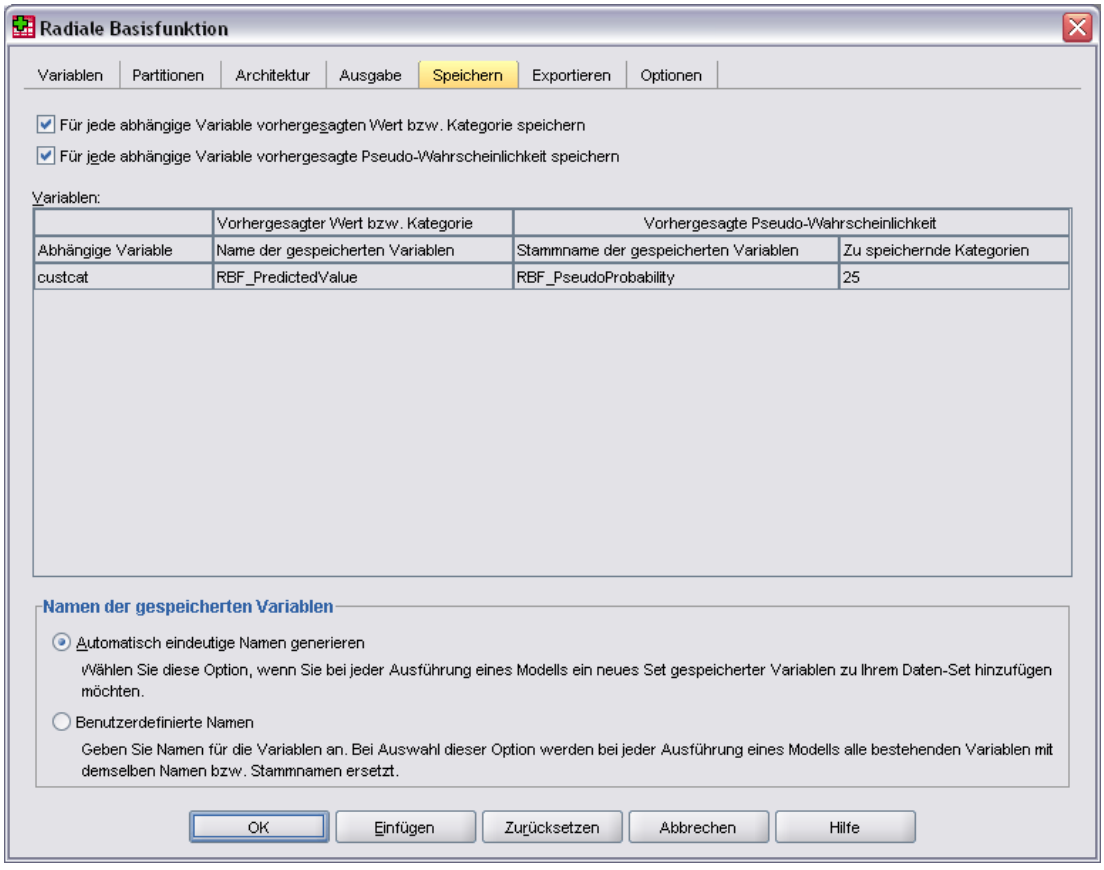
- **Diagramm "Residuum/Vorhergesagt".** Zeigt für jede metrische abhängige Variable ein Diagramm an, das die Residuen in Abhängigkeit von den vorhergesagten Werten angibt. Es sollte kein Muster zwischen Residuen und vorhergesagten Werten zu beobachten sein. Dieses Diagramm wird nur bei metrischen abhängigen Variablen erstellt.

Zusammenfassung der Fallverarbeitung. Zeigt die Tabelle mit der Zusammenfassung der Fallverarbeitung an, die die Anzahl der in der Analyse ein- und ausgeschlossenen Fälle zusammenfasst (insgesamt und nach Trainings-, Test- und Holdout-Stichprobe geordnet).

Wichtigkeitsanalyse für unabhängige Variablen. Führt eine Sensitivitätsanalyse durch, mit der die Wichtigkeit der einzelnen Einflussvariablen für die Bestimmung des neuronalen Netzwerks berechnet wird. Die Analyse beruht auf der Kombination aus Trainings- und Teststichprobe bzw. nur auf der Trainingsstichprobe, wenn keine Teststichprobe vorhanden ist. Dadurch werden eine Tabelle und ein Diagramm erstellt, die die Wichtigkeit und die normalisierte Wichtigkeit für die einzelnen Einflussvariablen anzeigen. Beachten Sie, dass die Sensitivitätsanalyse rechenintensiv und zeitaufwendig ist, wenn eine große Anzahl an Einflussvariablen oder Fällen vorliegt.

Speichern

Abbildung 3-5
Radiale Basisfunktion: Registerkarte "Speichern"



Auf der Registerkarte "Speichern" können Vorhersagen im Daten-Set als Variablen gespeichert werden.

- **Für jede abhängige Variable vorhergesagten Wert bzw. Kategorie speichern.** Damit wird bei metrischen abhängigen Variablen der vorhergesagte Wert und bei kategorialen abhängigen Variablen die vorhergesagte Kategorie gespeichert.
- **Für jede abhängige Variable vorhergesagte Pseudo-Wahrscheinlichkeit speichern.** Damit werden bei kategorialen abhängigen Variablen die vorhergesagten Pseudo-Wahrscheinlichkeiten gespeichert. Für die ersten *n* Kategorien wird eine separate Variable gespeichert. Dabei wird *n* in der Spalte *Zu speichernde Kategorien* angegeben.

Namen der gespeicherten Variablen. Durch eine automatische Generierung von Namen wird sichergestellt, dass Ihre Arbeit nicht verloren geht. Mit benutzerdefinierten Namen können Sie Ergebnisse aus früheren Durchgängen verwerfen bzw. ersetzen, ohne zuerst die gespeicherten Variablen im Daten-Editor löschen zu müssen.

Wahrscheinlichkeiten und Pseudo-Wahrscheinlichkeiten

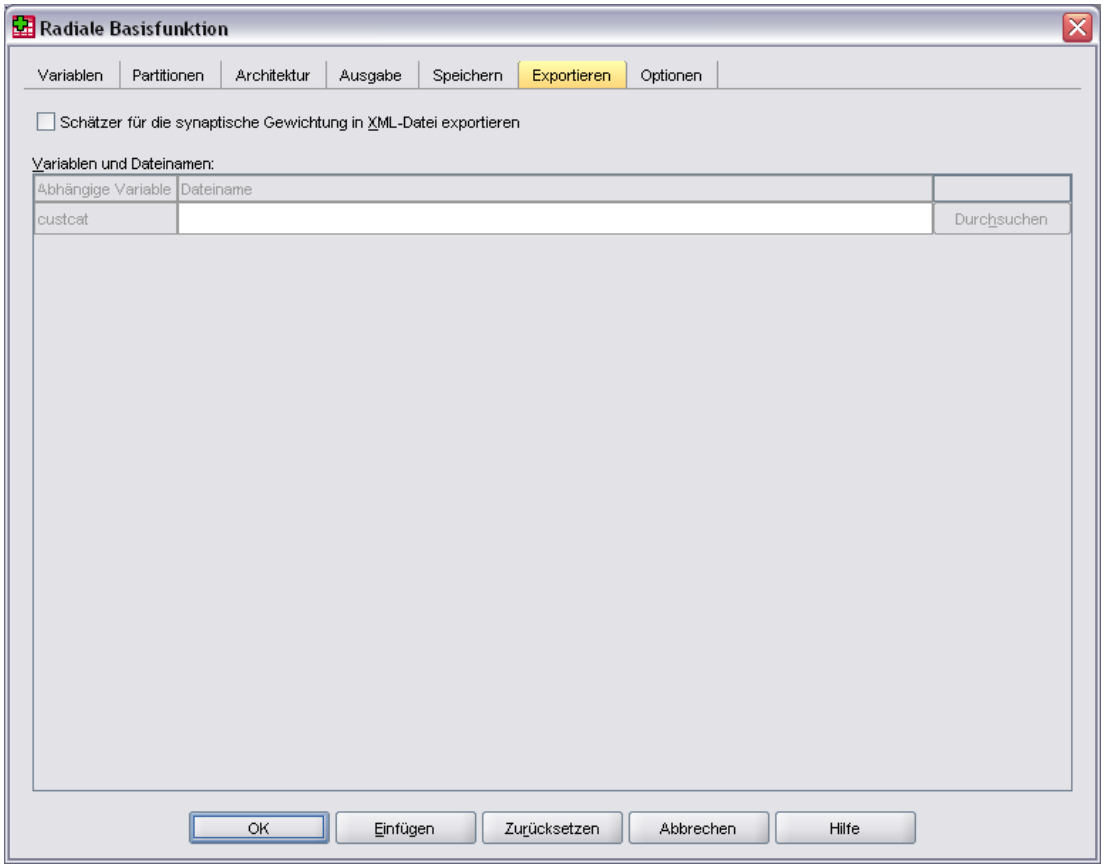
Vorhergesagte Pseudo-Wahrscheinlichkeiten können nicht als Wahrscheinlichkeiten interpretiert werden, da die Prozedur "Radiale Basisfunktion" für die Ausgabeschicht den Quadratsummenfehler und die Aktivierungsfunktion "Identität" verwendet. Die Prozedur speichert diese vorhergesagten Pseudo-Wahrscheinlichkeiten, selbst wenn sie kleiner als 0 oder größer als 1 sind oder wenn die Summe für eine abhängige Variable nicht 1 ergibt.

ROC, kumulative Gewinne und Lift Charts (siehe Ausgabe auf S. 31) werden auf der Grundlage von Pseudo-Wahrscheinlichkeiten erstellt. Falls Pseudo-Wahrscheinlichkeiten kleiner als 0 oder größer als 1 sind oder die Summe für eine abhängige Variable nicht 1 ergibt, werden die Werte zunächst so neu skaliert, dass sie zwischen 0 und 1 liegen und als Summe 1 ergeben. Die Pseudo-Wahrscheinlichkeiten werden durch Division durch ihre Summe neu skaliert. Wenn ein Fall beispielsweise für eine abhängige Variable mit drei Kategorien vorhergesagte Pseudo-Wahrscheinlichkeiten von 0,50; 0,60 und 0,40 aufweist, wird jede Pseudo-Wahrscheinlichkeit durch die Summe 1,50 dividiert, woraus sich die Werte 0,33; 0,40 und 0.27 ergeben.

Wenn negative Pseudo-Wahrscheinlichkeiten vorliegen, werden vor der oben beschriebenen Neuskalierung allen Pseudo-Wahrscheinlichkeiten jeweils mit dem Betrag der niedrigsten Wahrscheinlichkeit addiert. Wenn die Pseudo-Wahrscheinlichkeiten beispielsweise –0,30, 0,50 und 1,30 betragen, müssen Sie zunächst 0,30 zu jedem Wert addieren und erhalten somit die Werte 0,00; 0,80 und 1,60. Als Nächstes dividieren Sie die einzelnen neuen Werte durch die Summe 2,40, wodurch sich die Werte 0,00; 0,33 und 0,67 ergeben.

Export

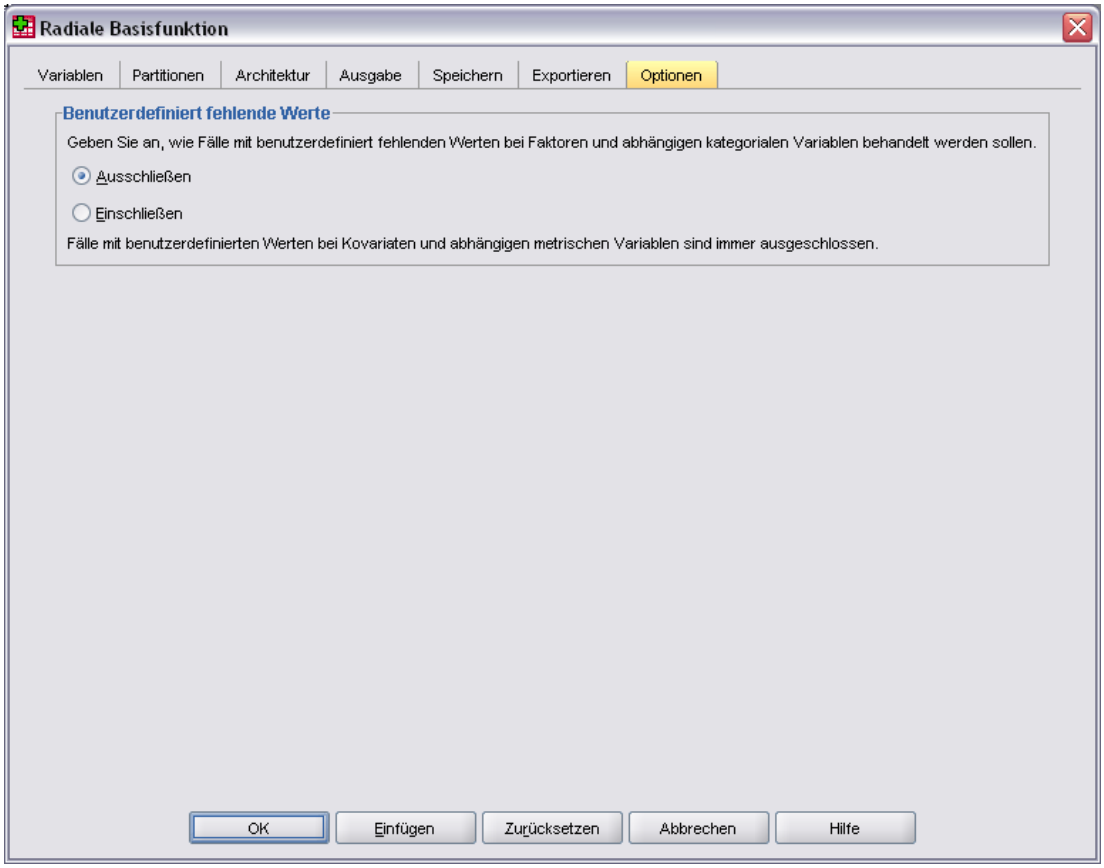
Abbildung 3-6
Radiale Basisfunktion: Registerkarte "Exportieren"



Die Registerkarte "Export" dient zum Speichern der Schätzer der synaptischen Gewichtungen für die einzelnen abhängigen Variablen in einer XML-Datei (PMML-Datei). SmartScore und SPSS Server (gesondertes Produkt) können anhand dieser Modelldatei die Modellinformationen zu Bewertungszwecken auf andere Datendateien anwenden. Diese Option ist nicht verfügbar, wenn aufgeteilte Dateien definiert wurden.

Optionen

Abbildung 3-7
Radiale Basisfunktion: Registerkarte "Optionen"



Benutzerdefinierte fehlende Werte. Faktoren müssen gültige Werte für einen Fall aufweisen, um in die Analyse aufgenommen zu werden. Mit diesen Steuerelementen legen Sie fest, ob benutzerdefiniert fehlende Werte bei den Faktoren und kategorialen abhängigen Variablen als gültige Werte behandelt werden sollen.

Teil II: Beispiele

Kapitel
4

Mehrschichtiges Perzeptron

Die Prozedur “Mehrschichtiges Perzeptron” (Multilayer Perceptron, MLP) erstellt ein Vorhersagemodell für eine oder mehrere abhängige Variablen (Zielvariablen), das auf den Werten der Einflussvariablen beruht.

Verwenden eines mehrschichtigen Perzeptrons zur Bewertung des Kreditrisikos

Eine Kreditsachbearbeiterin in einer Bank muss in der Lage sein, Merkmale zu ermitteln, die auf Personen hindeuten, die mit hoher Wahrscheinlichkeit ihre Kredite nicht zurückzahlen, und diese Merkmale zur Feststellung eines guten bzw. schlechten Kreditrisikos einzusetzen.

Angenommen, Informationen über 850 bisherige und potenzielle Kunden befinden sich in der Datei *bankloan.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 89. Bei den ersten 700 Fällen handelt es sich um Kunden, denen bereits ein Kredit gewährt wurde. Erstellen Sie anhand einer Zufallsstichprobe dieser 700 Kunden ein mehrschichtiges Perzeptron und lassen Sie die verbleibenden Kunden zunächst außen vor, um später damit die Analyse zu bewerten. Stufen Sie das Kreditrisiko der 150 zukünftigen Kunden dann mit diesem Modell als gering oder hoch ein.

Außerdem hat die Kreditsachbearbeiterin die Daten zuvor mithilfe einer logistischen Regression (in der Option “Regressionsmodelle”) analysiert und fragt sich, wie das mehrschichtige Perzeptron im Vergleich damit als Klassifizierungswerkzeug abschneidet.

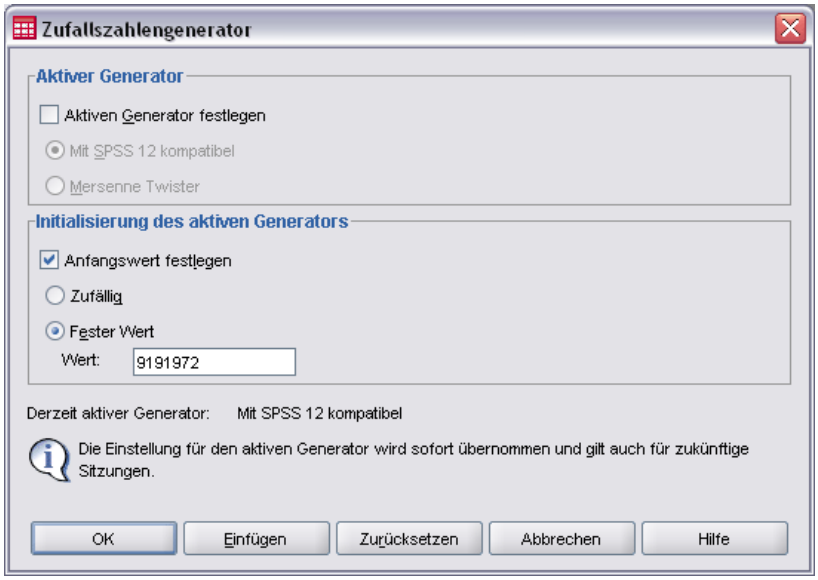
Vorbereiten der Daten für die Analyse

Durch die Festlegung des Startwerts können sie die Analyse exakt reproduzieren.

- Zur Festlegung des Startwerts wählen Sie die folgenden Menübefehle aus:

Transformieren
Zufallszahlengeneratoren...

Abbildung 4-1
Dialogfeld "Zufallszahlengenerator"



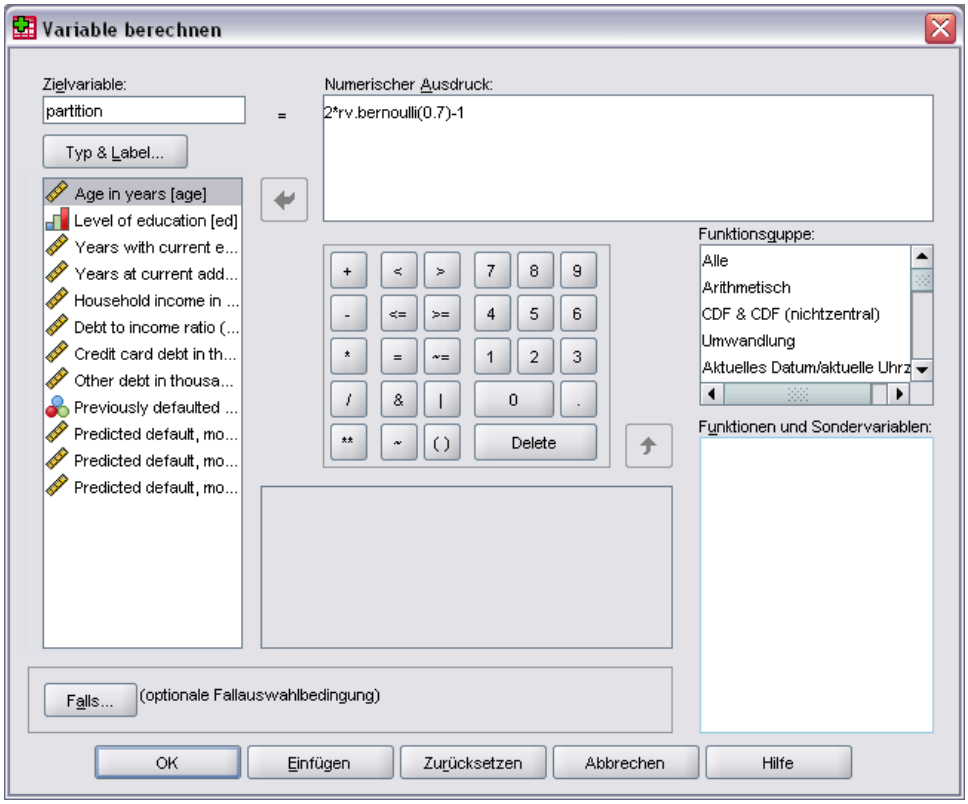
- Wählen Sie Anfangswert festlegen.
- Wählen Sie Fester Wert und geben Sie 9191972 als Wert ein.
- Klicken Sie auf OK.

In der vorangegangenen logistischen Regressionsanalyse wurden ungefähr 70 % der früheren Kunden der Trainingsstichprobe zugewiesen und 30 % einer Holdout-Stichprobe. Es ist eine Partitionsvariable erforderlich, um die in diesen Analysen verwendeten Stichproben exakt zu reproduzieren.

- Um die Partitionsvariable zu erstellen, wählen Sie folgende Optionen in den Menüs aus:

Transformieren
Variable berechnen...

Abbildung 4-2
Dialogfeld "Variable berechnen"



- Geben Sie partition in das Textfeld "Zielvariable" ein.
- Geben Sie 2*rv.bernoulli(0.7)-1 in das Textfeld "Numerischer Ausdruck" ein.

  Dadurch werden als Werte von *partition* **Bernoulli**-Zufallsvariablen mit einem Wahrscheinlichkeitsparameter von 0,7 verwendet, die so verändert werden, dass sie die Werte 1 oder −1 statt 1 bzw. 0 annehmen. Sie erinnern sich sicher, dass Fälle mit positiven Werten für die Partitionsvariable der Trainingsstichprobe zugewiesen werden, Fälle mit negativen Werten der Holdout-Stichprobe und Fälle mit dem Wert 0 der Teststichprobe. Im Moment geben wir keine Teststichprobe an.

- Klicken Sie im Dialogfeld "Variable berechnen" auf OK.

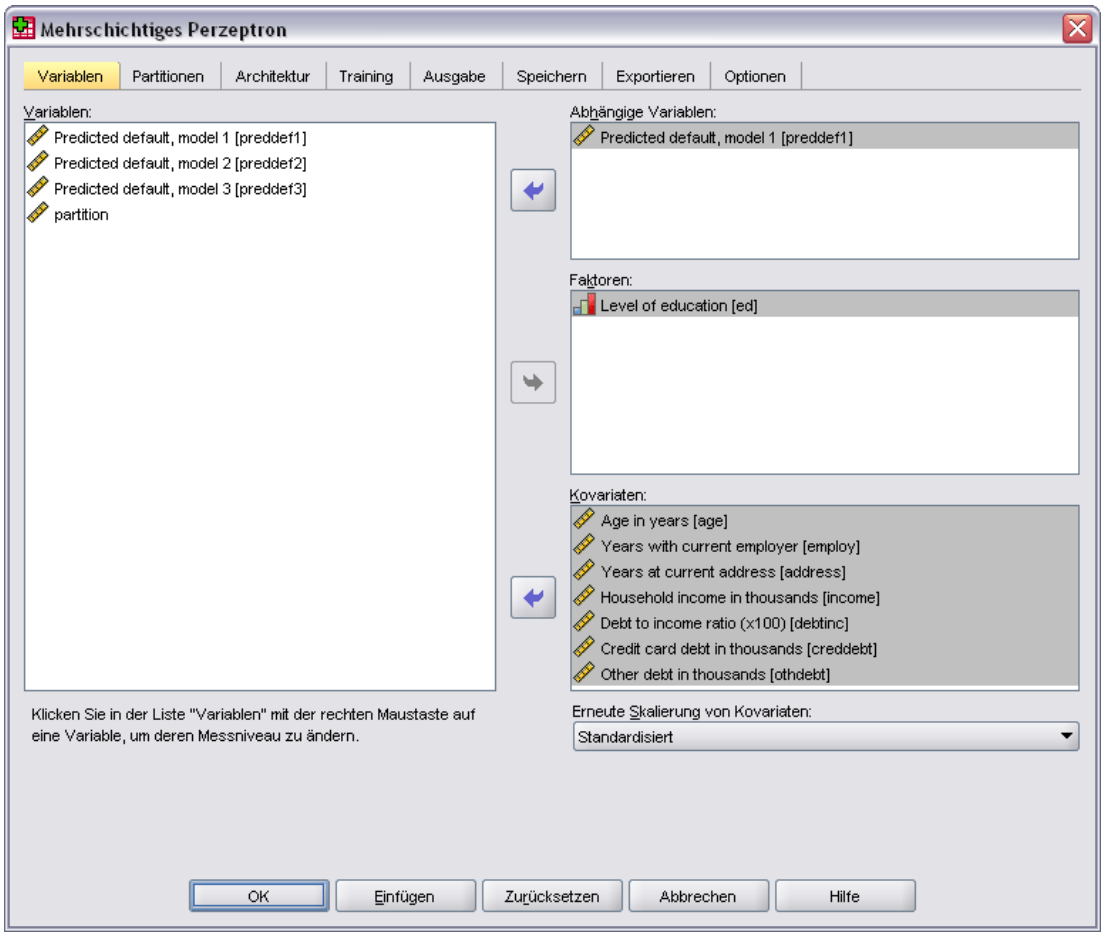
  Ungefähr 70 % der Kunden, die zuvor Kredite erhalten haben, weisen den Wert 1 für *partition* auf. Anhand dieser Kunden wird das Modell erstellt. Die restlichen Kunden, die zuvor Kredite erhalten haben, weisen den Wert −1 für *partition* auf und werden zur Validierung der Modellergebnisse verwendet.

Durchführung der Analyse

- Zum Ausführen einer Analyse vom Typ “Mehrschichtiges Perzeptron” wählen Sie die folgenden Menübefehle aus:

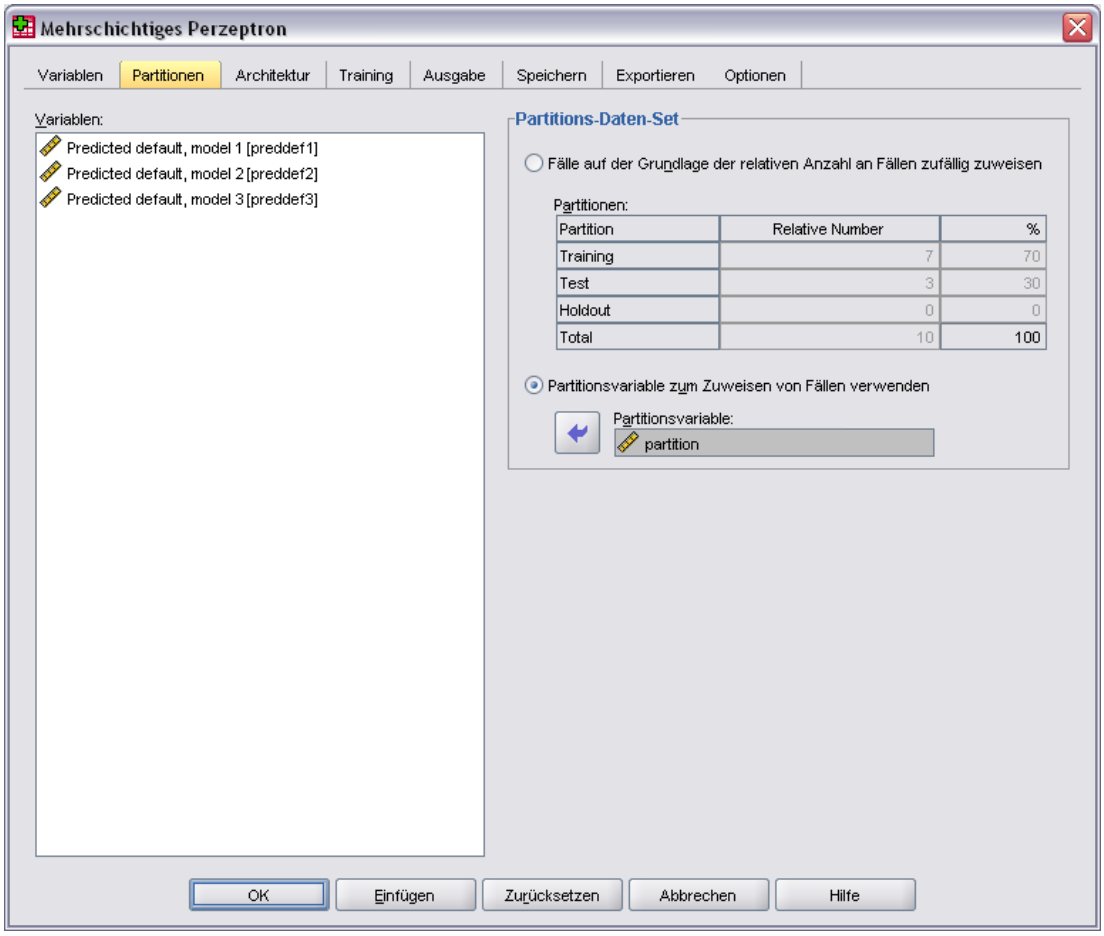
  Analysieren
  Neuronale Netze
  Mehrschichtiges Perzeptron...

Abbildung 4-3
Mehrschichtiges Perzeptron: Registerkarte “Variablen”



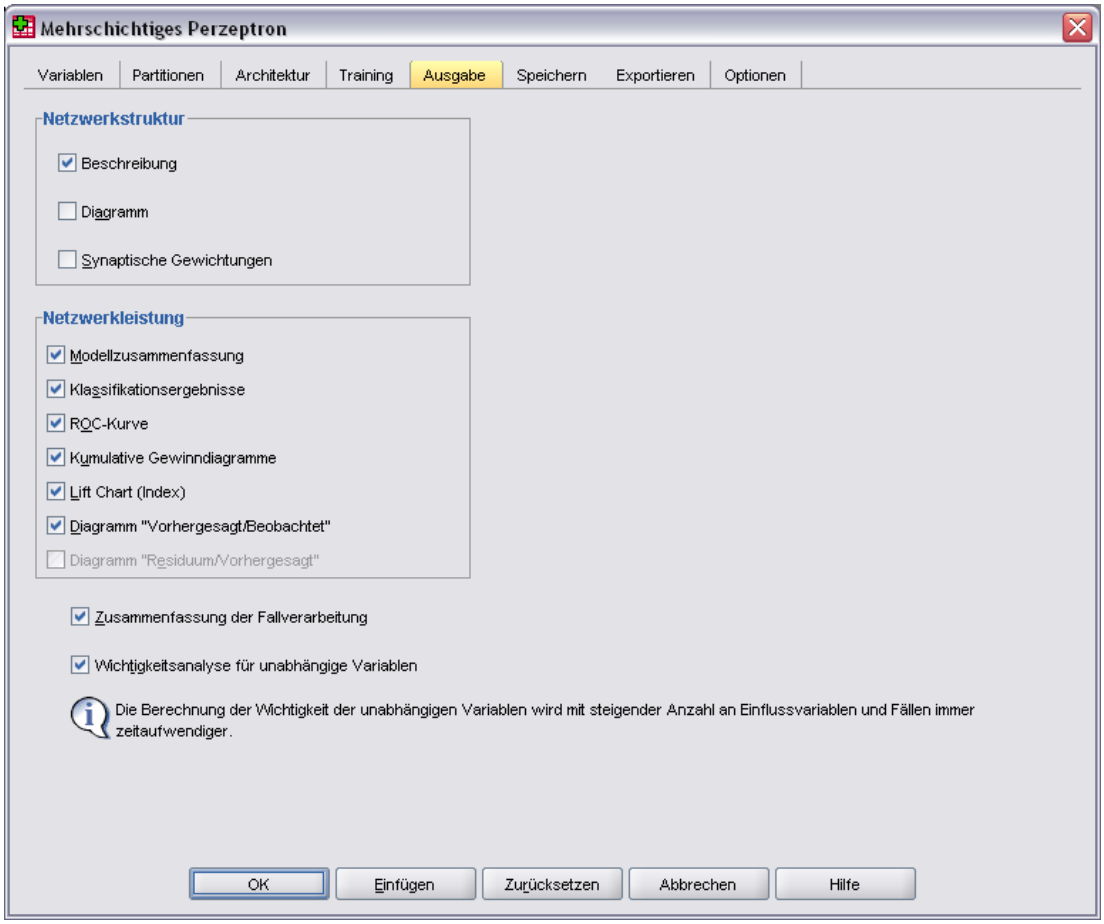
- Wählen Sie *Previously defaulted [default]* (vorherige Nichtzahlung) als abhängige Variable aus.
- Wählen Sie *Level of education [ed]* (Ausbildung) als Faktor aus.
- Wählen Sie *Age in years [age]* (Alter in Jahren) bis *Other debt in thousands [othdebt]* (Andere Schulden in Tausend) als Kovariaten aus.
- Klicken Sie auf die Registerkarte Partitionen.

Abbildung 4-4
Mehrschichtiges Perzeptron: Registerkarte "Partitionen"



- Wählen Sie die Option Partitionierungsvariable zum Zuweisen von Fällen verwenden aus.
- Wählen Sie *partition* als Partitionierungsvariable aus.
- Klicken Sie auf die Registerkarte Ausgabe.

Abbildung 4-5
Mehrschichtiges Perzeptron: Registerkarte "Ausgabe"



- Heben Sie im Gruppenfeld "Netzwerkstruktur" die Auswahl der Option Diagramm auf.
- Wählen Sie im Gruppenfeld "Netzwerkleistung" die Optionen ROC-Kurve, Kumulatives Gewinndiagramm, Lift Chart (Index) und Diagramm "Vorhergesagt/Beobachtet". Das Diagramm "Residuum/Vorhergesagt" ist nicht verfügbar, da die abhängige Variable nicht metrisch ist.
- Wählen Sie die Option Wichtigkeitsanalyse für unabhängige Variablen.
- Klicken Sie auf OK.

Zusammenfassung der Fallverarbeitung

Abbildung 4-6
Zusammenfassung der Fallverarbeitung

		N	Prozent
Beispiel	Training	499	71.3%
	Holdout	201	28.7%
Gültig		700	100,0%
Ausgeschlossen		150	
Gesamt		850	

Die Zusammenfassung der Fallverarbeitung zeigt, dass der Trainingsstichprobe 499 und der Holdout-Stichprobe 201 Fälle zugewiesen wurden. Bei den 150 aus der Analyse ausgeschlossenen Fällen handelt es sich um die potenziellen Kunden.

Netzwerkinformationen

Abbildung 4-7
Netzwerkinformationen

Eingabeschicht	Factors	1	Level of education
	Covariates	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
	Anzahl der Einheiten:[a]		12
	Rescaling Method for Covariates		Standardisiert
Verborgene Schicht(en):	Anzahl der verborgenen Schichten		1
	Anzahl der Einheiten in verborgener Schicht 1[a]		4
	Aktivierungsfunktion		Hyperbeltangens
Ausgabeschicht	Dependent Variables	1	Previously defaulted
	Anzahl der Einheiten:		2
	Aktivierungsfunktion		Softmax
	Fehlerfunktion		Kreuzentropie

a. Ohne die Verzerrungseinheit

In der Tabelle “Netzwerkinformationen” werden Informationen zum neuronalen Netzwerk angezeigt. Anhand dieser Tabelle können Sie sich vergewissern, dass die Spezifikationen korrekt sind. Beachten Sie hier insbesondere Folgendes:

- Die Anzahl der Einheiten in der Eingabeschicht ist die Anzahl der Kovariaten plus die Gesamtzahl der Faktorstufen; für jede Kategorie von *Level of education* (Ausbildung) wird eine gesonderte Einheit erstellt und keine der Kategorien wird als “redundante” Einheit betrachtet, wie dies bei vielen Modellierungsprozeduren üblich ist.

- Ebenso wird für jede Kategorie von *Previously defaulted* (vorherige Nichtzahlung) eine separate Ausgabeeinheit erstellt (für insgesamt zwei Einheiten in der Ausgabeschicht).
- Die automatische Architekturauswahl hat vier Einheiten in der verborgenen Schicht ausgewählt.
- Bei allen anderen Netzwerkinformationen werden die Standardwerte für die Prozedur verwendet.

Modellzusammenfassung

Abbildung 4-8
Modellzusammenfassung

Training	Kreuzentropiefehler	156,606
	Prozentsatz der falschen Vorhersagen	15,6%
	Verwendete Abbruchregel	Maximale Anzahl an Epochen (100) überschritten
	Trainingszeit	00:00:00.081
Holdout	Prozentsatz der falschen Vorhersagen	25,4%

Abhängige Variable: Previously defaulted

In der Modellzusammenfassung werden Informationen zu den Ergebnissen des Trainings und der Anwendung des endgültigen Netzwerks auf die Holdout-Stichprobe angezeigt.

- Der Kreuzentropiefehler wird angezeigt, da in der Ausgabeschicht die Aktivierungsfunktion "Softmax" verwendet wird. Dies ist die Fehlerfunktion, die das Netzwerk während des Trainings zu minimieren versucht.
- Der Prozentsatz der falschen Vorhersagen wird aus der Klassifikationsmatrix entnommen und in dem zugehörigen Thema eingehender erörtert.
- Der Schätzalgorithmus wurde angehalten, da die maximale Anzahl an Epochen erreicht war. Im Idealfall sollte das Training beendet werden, da der Fehler konvergiert hat. Dies wirft die Frage auf, ob während des Trainings etwas schief gelaufen ist, und sollte bei der weiteren Analyse der Daten im Hinterkopf behalten werden.

Klassifikation

Abbildung 4-9
Klassifikation

		Vorhergesagt		
Beispiel	Beobachtet	No	Yes	Percent Correct
Training	No	347	28	92,5%
	Yes	50	74	59,7%
	Overall Percent	79,6%	20,4%	84,4%
Test	No	123	19	86,6%
	Yes	32	27	45,8%
	Overall Percent	77,1%	22,9%	74,6%

Abhängige Variable: Previously defaulted

Die Klassifikationsmatrix zeigt die praktischen Ergebnisse der Verwendung des Netzwerks. In jedem Fall ist die vorhergesagte Antwort *Ja*, wenn die vorhergesagte Pseudo-Wahrscheinlichkeit der Fälle größer als 0,5 ist. Für jede Stichprobe gilt:

- Zellen auf der Diagonale der Kreuzklassifikation der Fälle stellen korrekte Vorhersagen dar.
- Zellen abseits der Diagonale der Kreuzklassifikation der Fälle stellen falsche Vorhersagen dar.

Von den für die Modellerstellung verwendeten Fällen wurden 74 von 124 Personen, die zuvor Zahlungsunfähig waren, korrekt klassifiziert. 347 der 375 zahlungsfähigen Personen wurden korrekt klassifiziert. Insgesamt wurden 84,4 % der Fälle korrekt klassifiziert. Dies entspricht den 15,6 % der falsch klassifizierten Fälle, die aus der Modellzusammenfassungstabelle ersichtlich sind. Das Modell ist umso besser, je höher der Prozentsatz der korrekt klassifizierten Fälle ist.

Die Klassifizierung anhand der Fälle, mit denen das Modell erstellt wurde, gerät jedoch leicht zu "optimistisch", da die Klassifizierungsrate aufgebläht ist. Die Holdout-Stichprobe erleichtert die Validierung der Modells; hier wurden 74,6 % der Fälle korrekt vom Modell klassifiziert. Dies deutet darauf hin, dass das Modell insgesamt in ungefähr drei von vier Fällen richtig liegt.

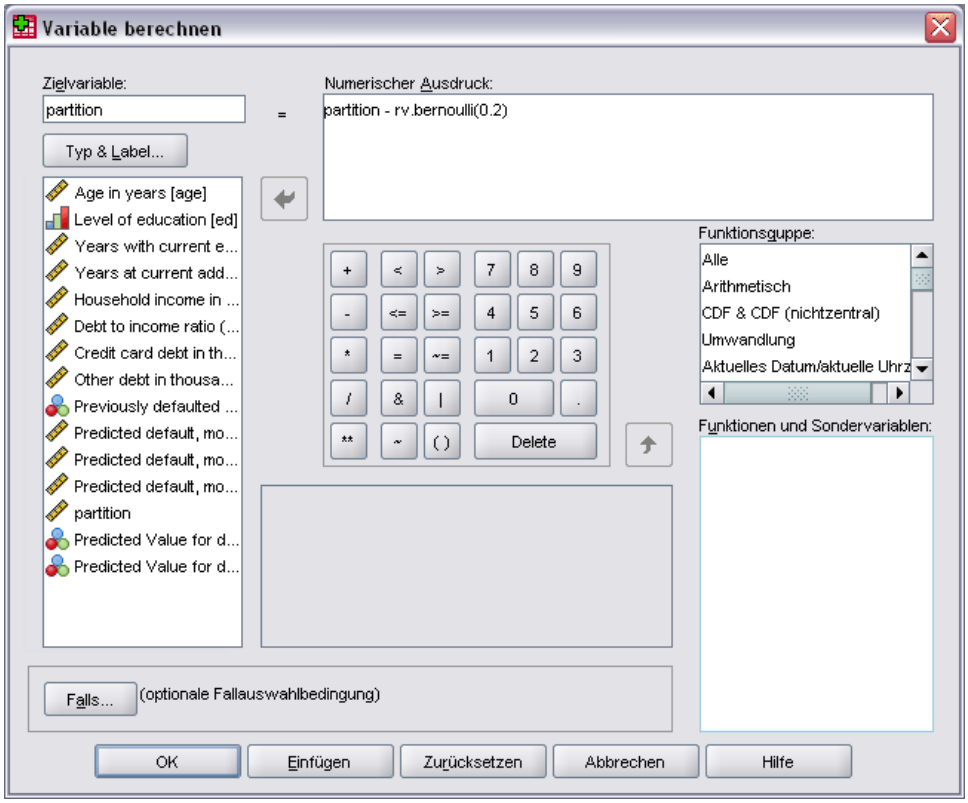
Korrigieren von Übertraining

Beim Rückblick auf die zuvor durchgeführte logistische Regressionsanalyse erinnert sich die Kreditsachbearbeiterin, dass die Trainings- und die Holdout-Stichprobe einen ähnlich hohen Prozentsatz der Fälle korrekt vorhersagte, nämlich ungefähr 80 %. Im Vergleich lag beim neuronalen Netzwerk ein höherer Prozentsatz korrekter Fälle in der Trainingsstichprobe vor, während die Holdout-Stichprobe bei der Vorhersage der Kunden, die tatsächlich zahlungsunfähig wurden, wesentlich schlechter abschnitt (45,8 % korrekt bei der Holdout-Stichprobe gegenüber 59,7 % bei der Trainingsstichprobe). In Verbindung mit der in der Modellzusammenfassungstabelle angegebenen Abbruchregel lässt dies darauf schließen, dass das Netzwerk möglicherweise **übertrainiert**, dass es also scheinbare Muster verfolgt, die durch zufällige Variation in den Trainingsdaten auftreten.

Glücklicherweise ist die Lösung für dieses Problem relativ einfach: Wir geben eine Teststichprobe an, damit das Netzwerk nicht "den Faden verliert". Wir haben die Partitionsvariable so erstellt, dass sie eine exakte Reproduktion der Trainings- und der Holdout-Stichprobe erstellt, die in der logistischen Regressionsanalyse erstellt wurden; bei der logistischen Regression gibt es jedoch keine Teststichproben. Wir nehmen daher einen Teil der Trainingsstichprobe und weisen ihn einer Teststichprobe zu.

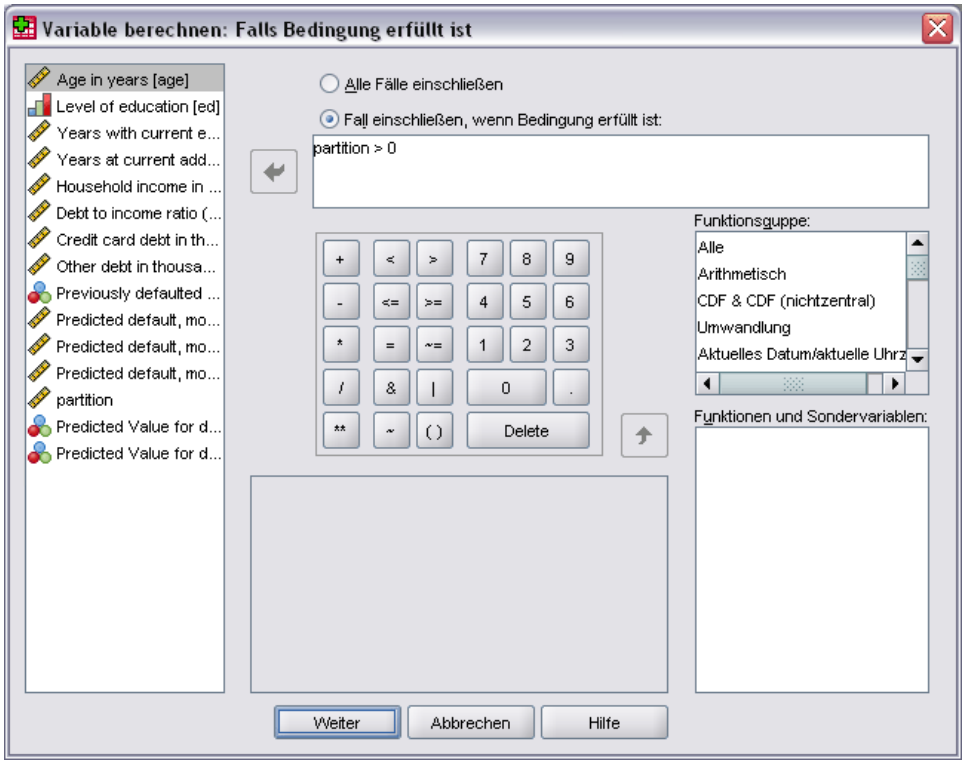
Erstellen der Teststichprobe

Abbildung 4-10
Dialogfeld "Variable berechnen"



- Rufen Sie das Dialogfeld "Variable berechnen" auf.
- Geben Sie partition - rv.bernoulli(0.2) in das Textfeld "Numerischer Ausdruck" ein.
- Klicken Sie auf Falls.

Abbildung 4-11
Variable berechnen: Dialogfeld "Variable berechnen: Falls Bedingung erfüllt ist"



- Wählen Sie Fall einschließen, wenn Bedingung erfüllt ist aus.
- Geben Sie partition>0 im Textfeld ein.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Variable berechnen" auf OK.

Dadurch werden die Werte von *partition*, die größer waren als 0, zurückgesetzt, sodass ungefähr 20 % den Wert 0 annehmen und 80 % weiterhin den Wert 1 aufweisen. Insgesamt befinden sich nun 100*(0,7*0,8)=56 % der Kunden, die zuvor Kredite erhalten haben, in der Trainings-Stichprobe und 14 % in der Teststichprobe. Kunden, die ursprünglich der Holdout-Stichprobe zugewiesen wurden, verbleiben dort.

Durchführung der Analyse

- Rufen Sie das Dialogfeld "Mehrschichtiges Perzeptron" erneut auf und klicken Sie auf die Registerkarte Speichern.
- Wählen Sie die Option Für jede abhängige Variable vorhergesagte Pseudo-Wahrscheinlichkeit speichern.
- Klicken Sie auf OK.

Zusammenfassung der Fallverarbeitung

Abbildung 4-12
Zusammenfassung der Fallverarbeitung für Modell mit Teststichprobe

		N	Prozent
Beispiel	Training	398	56,9%
	Testing	101	14,4%
	Holdout	201	28,7%
Gültig		700	100,0%
Ausgeschlossen		150	
Gesamt		850	

Von den 499 Fällen, die ursprünglich der Trainingsstichprobe zugewiesen wurden, wurden 101 nun der Teststichprobe zugewiesen.

Netzwerkinformationen

Abbildung 4-13
Netzwerkinformationen

Eingabeschicht	Factors	1	Level of education
	Covariates	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
	Anzahl der Einheiten:[a]		12
	Rescaling Method for Covariates		Standardisiert
Verborgene Schicht(en):	Anzahl der verborgenen Schichten		1
	Anzahl der Einheiten in verborgener Schicht 1[a]		7
	Aktivierungsfunktion		Hyperbeltangens
Ausgabeschicht	Dependent Variables	1	Previously defaulted
	Anzahl der Einheiten:		2
	Aktivierungsfunktion		Softmax
	Fehlerfunktion		Kreuzentropie

a. Ohne die Verzerrungseinheit

Die einzige Veränderung an der Tabelle der Netzwerkinformationen besteht darin, dass die automatische Architekturauswahl sieben Einheiten in der verborgenen Schicht ausgewählt hat.

Modellzusammenfassung

Abbildung 4-14
Modellzusammenfassung

Training	Kreuzentropiefehler	159,870
	Prozentsatz der falschen Vorhersagen	20,1%
	Verwendete Abbruchregel	1 aufeinander folgende(r) Schritt(e) ohne Verringerung des Fehlers[a]
	Trainingszeit	00:00:00,905
Test	Kreuzentropiefehler	40,068
	Prozentsatz der falschen Vorhersagen	17,8%
Prüfung (Holdout)	Prozentsatz der falschen Vorhersagen	20,4%

Abhängige Variable: Previously defaulted

a. Fehlerberechnungen beruhen auf der Teststichprobe.

Die Modellzusammenfassung weist eine Reihe positiver Merkmale auf:

- Der Prozentsatz falscher Vorhersagen ist in der Training-, Test- und Holdout-Stichprobe jeweils ungefähr gleich groß.
- Der Schätzalgorithmus wurde angehalten, da der Fehler nach einem Schritt im Algorithmus nicht kleiner wurde.

Dies ist ein weiterer Hinweis darauf, dass das ursprüngliche Modell tatsächlich übertrainiert war und das Problem durch das Hinzufügen einer Teststichprobe gelöst wurde. Freilich sind die Stichprobengrößen relativ klein und wir sollten vielleicht die Verlagerung um einige wenige Prozentpunkte nicht überinterpretieren.

Klassifikation

Abbildung 4-15
Klassifikation

		Vorhergesagt		
Beispiel	Beobachtet	No	Yes	Percent Correct
Training	No	263	34	88,6% %
	Yes	46	55	54,5% %
	Overall Percent	77,6%	22,4%	79,9% %
Test	No	73	5	93,6% %
	Yes	13	10	43,5% %
	Overall Percent	85,1%	14,9%	82,2% %
Holdout	No	124	18	87,3% %
	Yes	23	36	61,0% %
	Overall Percent	73,1%	26,9%	79,6% %

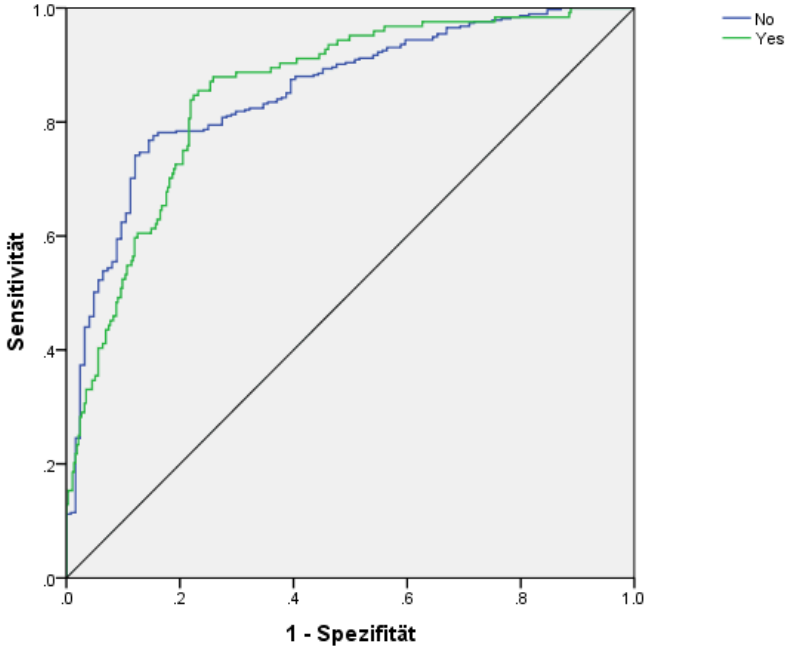
Abhängige Variable: Previously defaulted

Die Klassifikationsmatrix zeigt, dass das Netzwerk bei Verwendung von 0,5 als Pseudo-Wahrscheinlichkeits-Trennwert für die Klassifikation wesentlich bessere Ergebnisse bei der Vorhersage von Personen erzielt, die nicht in Zahlungsschwierigkeiten geraten, als bei der Vorhersage von zahlungsunfähigen Personen. Leider bietet der Trennwert als Einzelwert nur

einen sehr begrenzten Einblick in die Vorhersagekraft des Netzwerks, sodass er nicht unbedingt übermäßig nützlich für den Vergleich konkurrierender Netzwerke ist. Stattdessen sollten wir lieber einen Blick auf die ROC-Kurve werfen.

ROC-Kurve

Abbildung 4-16
ROC-Kurve



Abhängige Variable: Previously defaulted

Die ROC-Kurve bietet eine grafische Darstellung der **Sensitivität** und **Spezifität** für alle möglichen Trennwerte in einem einzelnen Diagramm. Diese Darstellungsweise ist wesentlich übersichtlicher und aussagekräftiger als eine Reihe von Tabellen. Das hier gezeigte Diagramm enthält zwei Kurven, eine für die Kategorie *Nein* und eine für die Kategorie *Ja*. Da es nur zwei Kategorien gibt, sind die Kurven bezüglich einer Linie im 45-Grad-Winkel (nicht angezeigt) symmetrisch, die von der linken oberen Ecke des Diagramms zur rechten unteren Ecke verläuft.

Beachten Sie, dass dieses Diagramm auf der Kombination aus Trainings- und Teststichprobe beruht. Um ein ROC-Diagramm für die Holdout-Stichprobe zu erstellen, müssen Sie die Datei an der Partitionsvariablen aufteilen und die Prozedur “ROC-Kurve” für die gespeicherten vorhergesagten Pseudo-Wahrscheinlichkeiten ausführen.

Abbildung 4-17
Fläche unter der Kurve

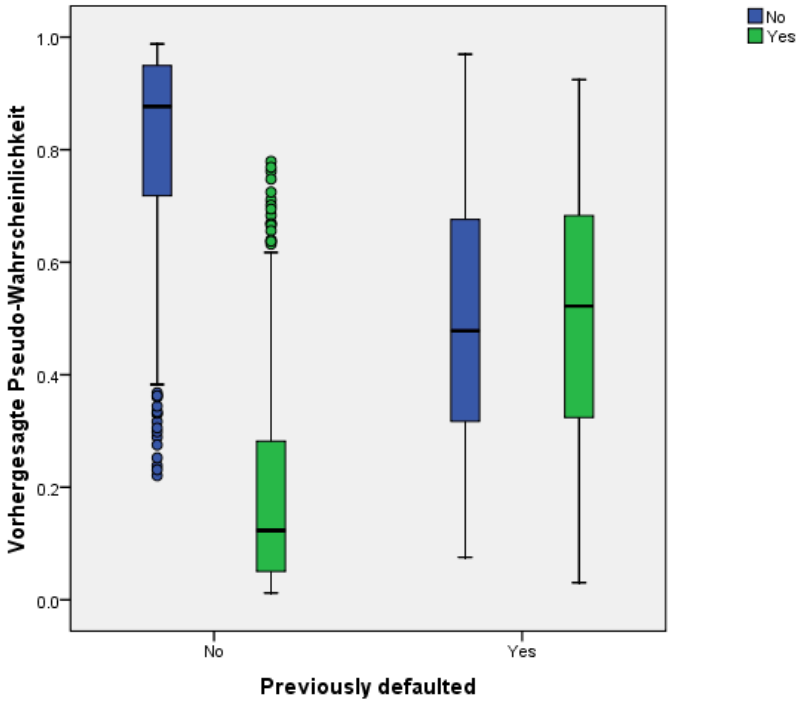
		Fläche
Previously defaulted	No	,858
	Yes	,858

Die Fläche unter der Kurve ist eine numerische Zusammenfassung der ROC-Kurve und die Werte in der Tabelle stellen für jede Kategorie die Wahrscheinlichkeit dar, dass die vorhergesagte Wahrscheinlichkeit, in diese Kategorie zu gehören, für einen zufällig ausgewählten Fall in der betreffenden Kategorie größer ist als für einen zufällig ausgewählten Fall, der nicht in diese Kategorie eingeteilt wurde. Wenn beispielsweise nach dem Zufallsprinzip eine zahlungsunfähige Person und eine zahlungsfähige Person ausgewählt werden, liegt die Wahrscheinlichkeit, dass die vom Modell vorhergesagte Pseudo-Wahrscheinlichkeit für Zahlungsunfähigkeit für die zahlungsunfähige Person höher ist als für die zahlungsfähige Person bei 0,853.

Die Fläche unter der Kurve ist zwar eine nützliche, aus einem einzigen statistischen Wert bestehende Zusammenfassung für die Genauigkeit des Netzwerks, aber Sie müssen in der Lage sein, ein bestimmtes Kriterium auszuwählen, nach dem die Kunden klassifiziert werden sollen. Das Diagramm "Vorhergesagt/Beobachtet" bietet einen visuellen Ausgangspunkt für diesen Vorgang.

Diagramm "Vorhergesagt/Beobachtet"

Abbildung 4-18
Vorhergesagt/Beobachtet, Diagramm



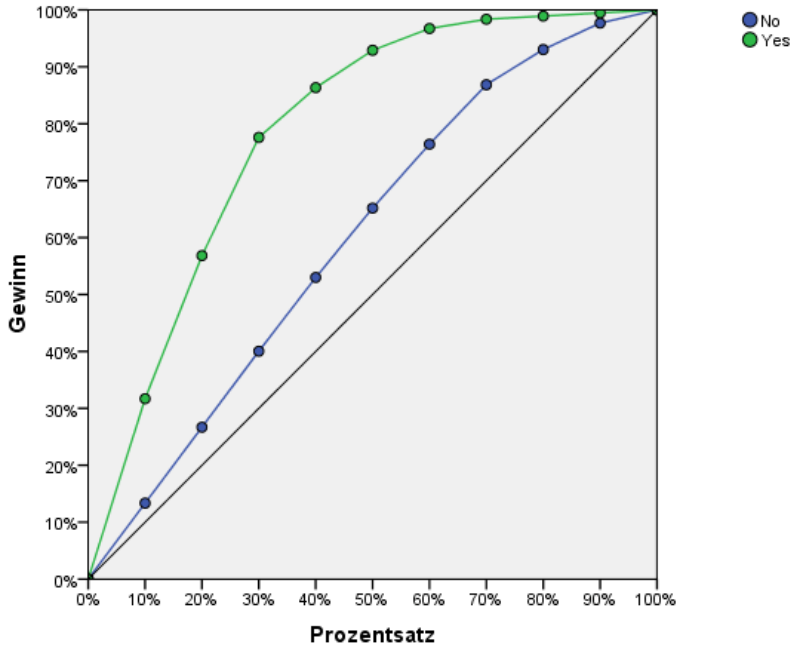
Für kategoriale abhängige Variablen zeigt das Diagramm "Vorhergesagt/Beobachtet" gruppierte Boxplots vorhergesagter Pseudo-Wahrscheinlichkeiten für die Kombination aus Trainings- und Teststichprobe an. Die *x*-Achse entspricht den beobachteten Antwortkategorien und die Legende entspricht vorhergesagten Kategorien.

- Der Boxplot ganz links zeigt für Fälle mit der beobachteten Kategorie Nein die vorhergesagte Pseudo-Wahrscheinlichkeit für die Kategorie Nein. Der Bereich des Boxplots über der Marke 0,5 auf der *y*-Achse stellt die korrekten Vorhersagen in der Klassifikationsmatrix dar. Der Bereich unterhalb der Marke von 0,5 stellt die falschen Vorhersagen dar. Wir erinnern uns aus der Klassifikationsmatrix, dass das Netzwerk unter Verwendung eines Trennwerts von 0,5 sehr gute Ergebnisse bei der Vorhersage von Fällen mit der Kategorie Nein erzielt, sodass nur ein Teil des unteren Whiskers und einige Ausreißer falsch klassifiziert sind.
- Der nächste Boxplot zeigt für Fälle mit der beobachteten Kategorie Nein die vorhergesagte Pseudo-Wahrscheinlichkeit für die Kategorie Ja. Da die Zielvariable nur zwei Kategorien enthält, sind die ersten beiden Boxplots bezüglich der horizontalen Linie bei 0,5 symmetrisch.
- Der dritte Boxplot zeigt für Fälle mit der beobachteten Kategorie Ja die vorhergesagte Pseudo-Wahrscheinlichkeit für die Kategorie Nein. Dieser Boxplot und der letzte Boxplot sind bezüglich der horizontalen Linie bei 0,5 symmetrisch.
- Der letzte Boxplot zeigt für Fälle mit der beobachteten Kategorie Ja die vorhergesagte Pseudo-Wahrscheinlichkeit für die Kategorie Ja. Der Bereich des Boxplots über der Marke 0,5 auf der *y*-Achse stellt die korrekten Vorhersagen in der Klassifikationsmatrix dar. Der Bereich unterhalb der Marke von 0,5 stellt die falschen Vorhersagen dar. Wir erinnern uns aus der Klassifikationsmatrix, dass das Netzwerk unter Verwendung eines Trennwerts von 0,5 etwas mehr als die Hälfte der Fälle mit der Kategorie Ja vorhersagt, sodass ein relativ großer Teil der Box falsch klassifiziert ist.

Eine Betrachtung des Plots ergibt, dass durch eine Senkung des Trennwerts zur Klassifizierung eines Falls als Ja von 0,5 auf ungefähr 0,3 — dies ist ungefähr der Wert, bei dem die Oberkante der zweiten Box und die Unterkante der vierten Box liegen — die Wahrscheinlichkeit, Personen, die später zahlungsunfähig werden, korrekt zu erfassen, erhöht werden kann, ohne dass dabei viele potenzielle gute Kunden verloren gehen. Durch das Verschieben von 0,5 auf 0,3 entlang der zweiten Box werden also nur relativ wenige zahlungskräftige Kunden entlang dem Whisker nun fälschlicherweise als vorhergesagte zahlungsunfähige Kunden klassifiziert, während durch diese Verschiebung entlang der vierten Box nun viele zahlungsunfähige Kunden innerhalb der Box korrekt als vorhergesagte zahlungsunfähige Kunden klassifiziert werden.

Kumulatives Gewinndiagramm und Lift Chart

Abbildung 4-19
Kumulatives Gewinndiagramm



Abhängige Variable: Previously defaulted

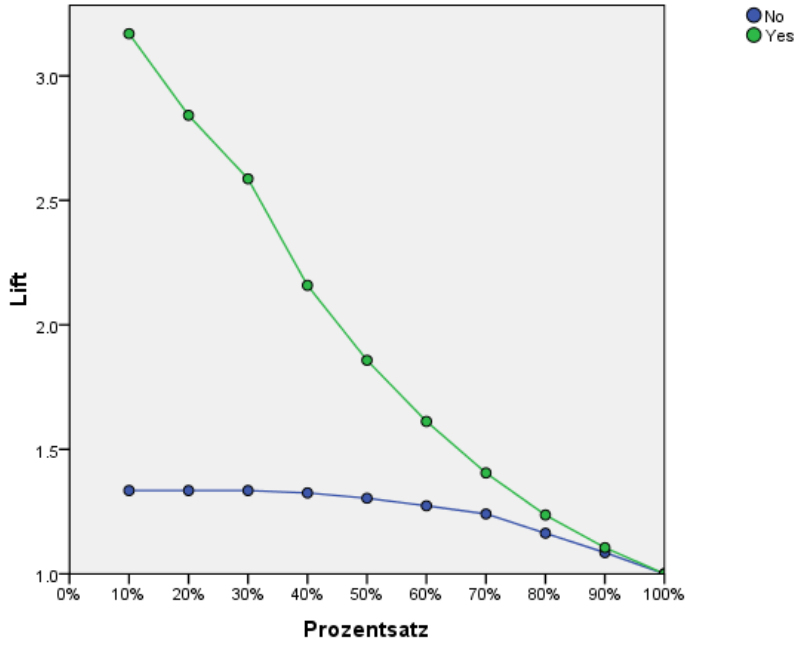
Das kumulative Gewinndiagramm zeigt den Prozentsatz der Fälle in einer bestimmten Kategorie, die "gewonnen" werden, indem ein bestimmter Prozentsatz der Gesamtzahl der Fälle anvisiert wird. Beispiel: Der erste Punkt auf der Kurve für die Kategorie *Ja* liegt bei (10%, 30%). Dies bedeutet Folgendes: Wenn Sie ein Daten-Set mit dem Netzwerk scoren und alle Fälle nach der vorhergesagten Pseudo-Wahrscheinlichkeit von *Ja* sortieren, ist zu erwarten, dass die obersten 10 % ungefähr 30 % aller Fälle enthalten, die tatsächlich in die Kategorie *Ja* (zahlungsunfähige Personen) fallen. Ebenso enthalten die obersten 20 % ungefähr 50 % der zahlungsunfähigen Personen, die obersten 30 % der Fälle 70 % der zahlungsunfähigen Personen usw. Bei Auswahl von 100 % des gescorten Daten-Sets erfassen Sie alle zahlungsunfähigen Personen im Daten-Set.

Die diagonale Linie ist die "Basis"-Kurve. Wenn Sie nach dem Zufallsprinzip 10 % der Fälle aus dem gescorten Daten-Set auswählen, ist zu erwarten, dass Sie ungefähr 10 % der Fälle "gewinnen", die tatsächlich in die Kategorie Ja fallen. Je höher über der Basis eine Kurve liegt, desto größer ist der Gewinn. Das kumulative Gewinndiagramm erleichtert die Auswahl eines Trennwerts für die Klassifizierung: Wählen Sie einen Prozentsatz aus, der dem angestrebten Gewinn entspricht, und ordnen Sie dann diesen Prozentsatz dem entsprechenden Trennwert zu.

Welcher Gewinn angestrebt wird, hängt von den Kosten für Fehler erster und zweiter Art (Typ I und Typ II) ab. Wie hoch sind die Kosten der Einstufung einer zahlungsunfähigen Person in die Gruppe der nicht zahlungsunfähigen Personen (Fehler erster Art)? Wie hoch sind die Kosten der Einstufung einer nicht zahlungsunfähigen Person in die Gruppe der zahlungsunfähigen Personen (Fehler zweiter Art)? Wenn die Vermeidung uneinbringlicher Forderungen das Hauptanliegen ist, sollte der Fehler erster Art (Typ I) möglichst niedrig gehalten werden. Beim kumulativen

Gewinndiagramm könnte dies damit erreicht werden, dass Antragstellern aus den obersten 40 % der Pseudo-Wahrscheinlichkeit von *Ja* keine Kredite gewährt werden. Damit sind fast 90 % der Personen, die voraussichtlich zahlungsunfähig werden, erfasst. Allerdings wird damit auch fast die Hälfte der Antragsteller abgelehnt. Wenn die Erweiterung des Kundenstamms oberste Priorität hat, sollte der Fehler zweiter Art (Typ II) minimiert werden. in diesem Diagramm entspricht dies einer Ablehnung der obersten 10 %, wodurch 30 % der zahlungsunfähigen Personen erfasst werden und die Menge der Antragsteller nahezu gleich bleibt. Normalerweise sind beide Punkte von großer Bedeutung, sodass Sie eine Entscheidungsregel für die Klassifizierung von Kunden aufstellen müssen, die die beste Mischung aus Sensitivität und Spezifität bietet.

Abbildung 4-20
Lift Chart (Index)



Der Lift Chart wird aus dem kumulativen Gewinndiagramm abgeleitet; die Werte auf der *y*-Achse entspricht dem Quotienten aus dem kumulativen Gewinn für jede Kurve und der Basis. Der Lift bei 10 % für die Kategorie Ja beträgt somit 30 % / 10 % = 3,0. Er bietet eine alternative Möglichkeit zur Analyse der Informationen im kumulativen Gewinndiagramm.

Anmerkung: Das kumulative Gewinndiagramm und der Lift Chart beruhen auf der Kombination aus Trainings- und Teststichprobe.

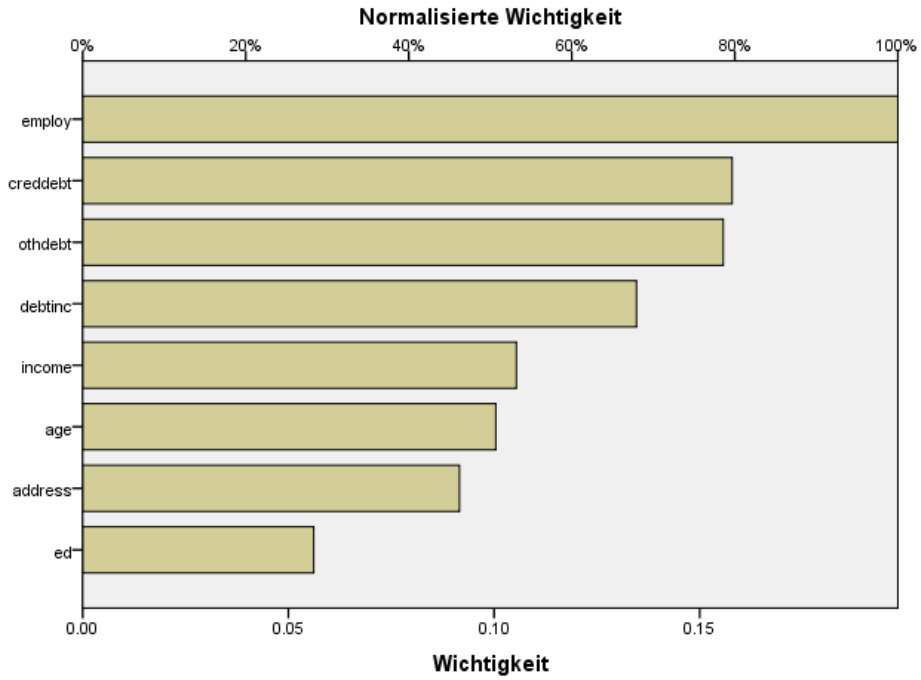
Wichtigkeit der unabhängigen Variablen

Abbildung 4-21
Wichtigkeit der unabhängigen Variablen

	Wichtigkeit	Normalisierte Wichtigkeit
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%

Die Wichtigkeit einer unabhängigen Variablen ist ein Maß dafür, wie stark sich der vom Modell vorhergesagte Wert des Netzwerks für verschiedene Werte der unabhängigen Variablen ändert. Die normalisierte Wichtigkeit berechnet sich einfach, indem die Wichtigkeitswerte durch die größten Wichtigkeitswerte dividiert und als Prozentsätze ausgedrückt werden.

Abbildung 4-22
Wichtigkeitsdiagramm für die unabhängigen Variablen



Das Wichtigkeitsdiagramm ist einfach ein Balkendiagramm der Werte in der Wichtigkeitstabelle, nach absteigender Wichtigkeit sortiert. Es sieht so aus, dass Variablen, die mit der Stabilität eines Kunden (*employ* (Jahre der Beschäftigung beim derzeitigen Arbeitgeber), *address* (wohnhaft an gleicher Adresse (in Jahren)) und Schulden (*creddebt* (Schulden auf Kreditkarte in Tausend),

debtinc (Relation Schulden zu Einkommen)) zu tun haben, den größten Effekt darauf haben, wie das Netzwerk Kunden klassifiziert; was nicht abgelesen werden kann, ist die "Richtung" der Beziehung zwischen diesen Variablen und der vorhergesagten Wahrscheinlichkeit der Nichtzurückzahlung. Man würde annehmen, dass ein höherer Schuldenstand auf eine größere Wahrscheinlichkeit der Nichtrückzahlung hinweist, aber um sicher zu sein, müsste ein Modell mit leichter interpretierbaren Parametern verwendet werden.

Zusammenfassung

Mit der Prozedur "Mehrschichtiges Perzeptron" haben Sie ein Netzwerk für die Vorhersage der Wahrscheinlichkeit erstellt, mit der ein bestimmter Kunde seinen Kredit nicht zurückzahlen wird. Die Modellergebnisse sind mit den Ergebnissen vergleichbar, die mithilfe der logistischen Regression oder der Diskriminanzanalyse gewonnen werden. Sie können also recht zuversichtlich sein, dass die Daten keine Beziehungen enthalten, die sich nicht durch diese Modelle erfassen lassen. Daher können Sie diese Modelle für die weitere Analyse der Eigenschaften der Beziehung zwischen abhängigen und unabhängigen Variablen verwenden.

Verwenden eines mehrschichtigen Perzeptrons zur Abschätzung von Behandlungskosten und Aufenthaltsdauer

Ein Krankenhaussystem möchte die Kosten und die Aufenthaltsdauer für Patienten aufzeichnen, die zur Behandlung eines Herzinfarkts aufgenommen wurden. Durch genaue Schätzer dieser Messwerte kann die Krankenhausverwaltung die verfügbare Bettenkapazität während der Behandlung der Patienten besser verwalten.

Die Datendatei *patient_los.sav* enthält die Behandlungsaufzeichnungen zu Patienten, die wegen eines Herzinfarkts behandelt wurden.Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 89. Erstellen Sie mithilfe von "Mehrschichtiges Perzeptron" ein Netzwerk zur Vorhersage der Kosten und der Aufenthaltsdauer im Krankenhaus.

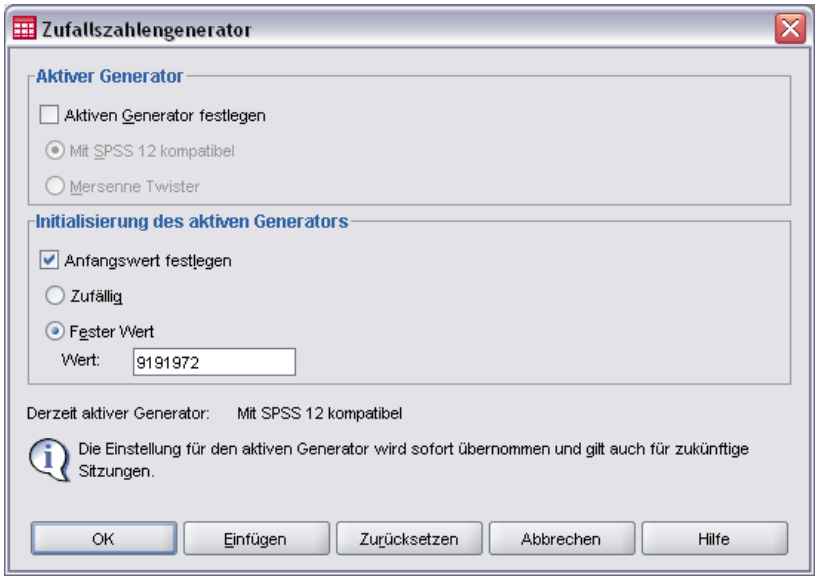
Vorbereiten der Daten für die Analyse

Durch die Festlegung des Startwerts können sie die Analyse exakt reproduzieren.

- Zur Festlegung des Startwerts wählen Sie die folgenden Menübefehle aus:

Transformieren
  Zufallszahlengeneratoren...

Abbildung 4-23
Dialogfeld "Zufallszahlengenerator"



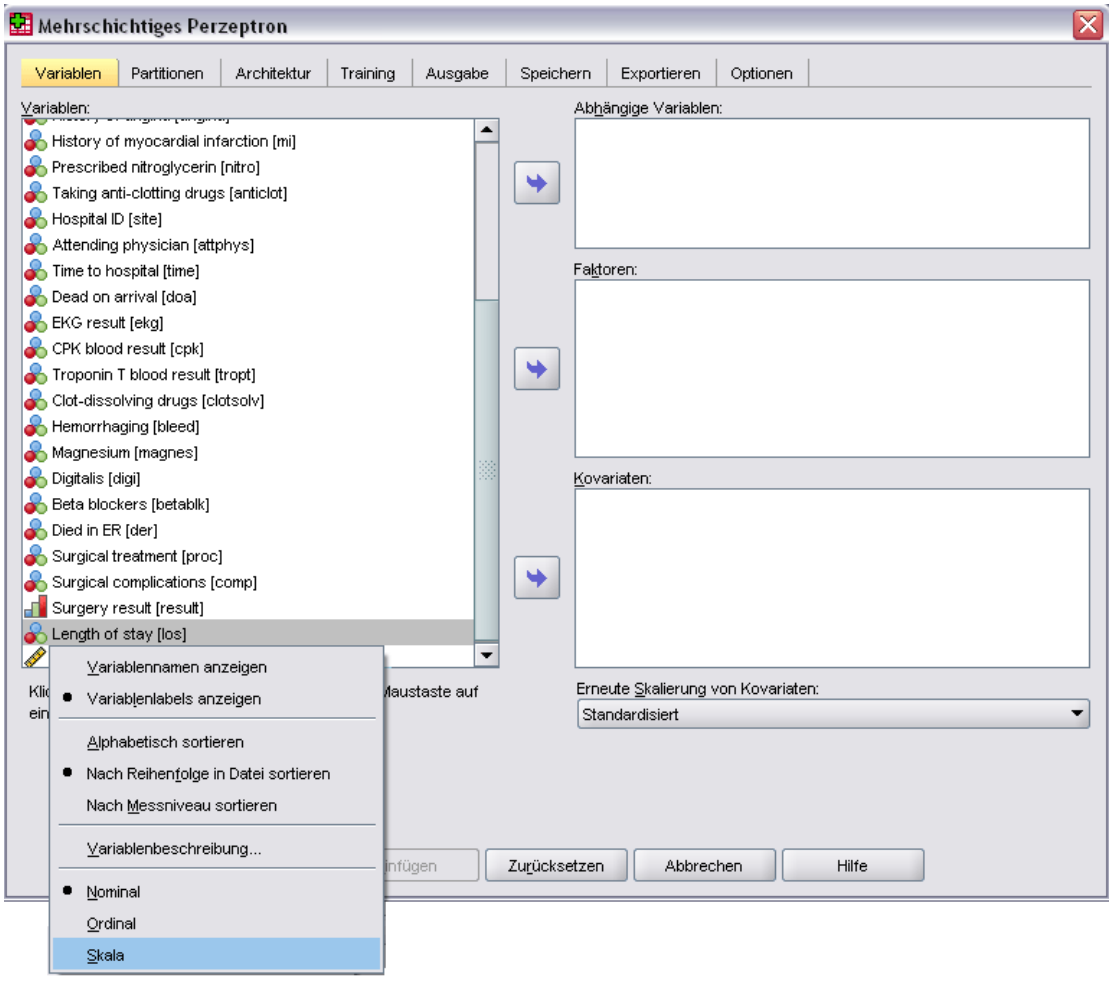
- Wählen Sie Anfangswert festlegen.
- Wählen Sie Fester Wert und geben Sie 9191972 als Wert ein.
- Klicken Sie auf OK.

Durchführung der Analyse

- Zum Ausführen einer Analyse vom Typ "Mehrschichtiges Perzeptron" wählen Sie die folgenden Menübefehle aus:

Analysieren
  Neuronale Netze
    Mehrschichtiges Perzeptron...

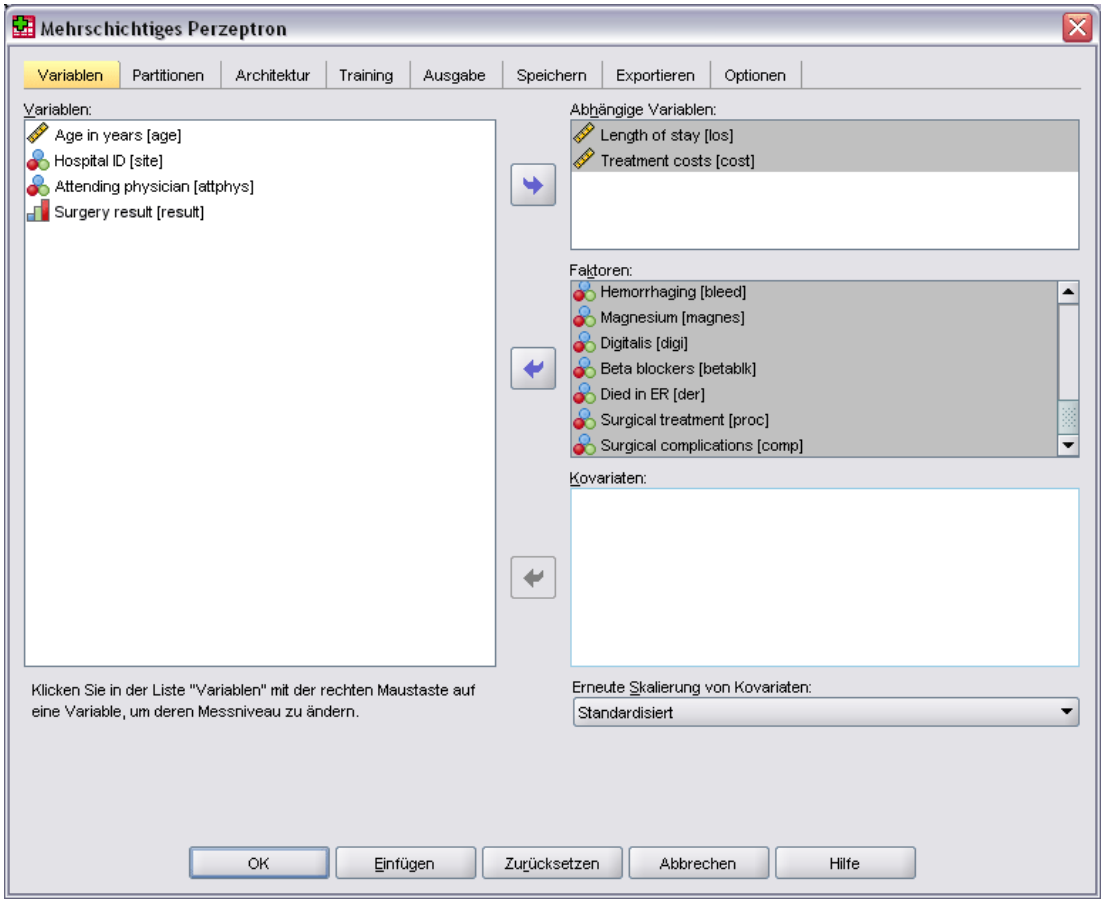
Abbildung 4-24
Mehrschichtiges Perzeptron: Registerkarte "Variablen" und Kontextmenü für "Length of stay" (Aufenthaltsdauer)



Length of stay [los] (Aufenthaltsdauer) weist ein ordinales Messniveau auf, Sie möchten jedoch, dass das Netzwerk diese Variable als metrisch behandelt.

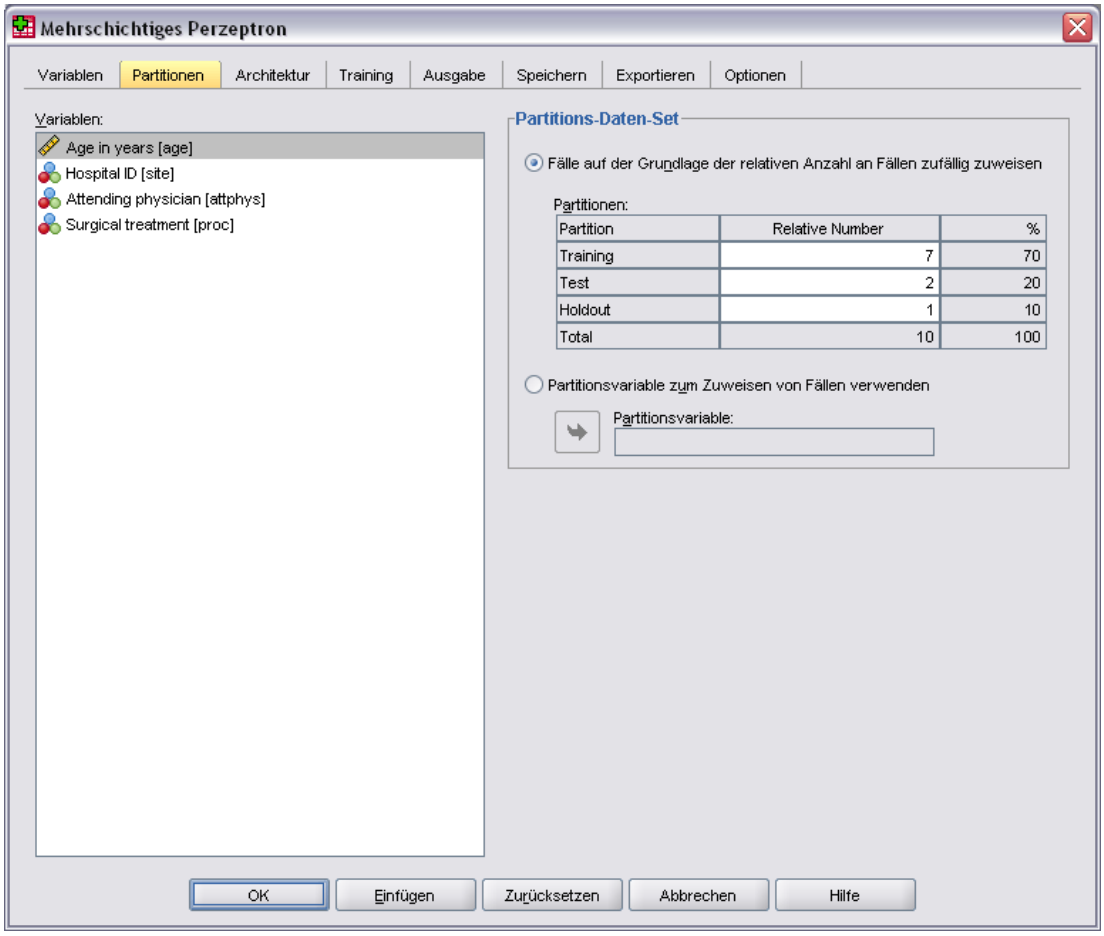
- Klicken Sie mit der rechten Maustaste auf *Length of stay [los]* (Aufenthaltsdauer) und wählen Sie im Kontextmenü die Option Skala (Metrisch) aus.

Abbildung 4-25
Mehrschichtiges Perzeptron: Registerkarte "Variablen" mit ausgewählten Variablen



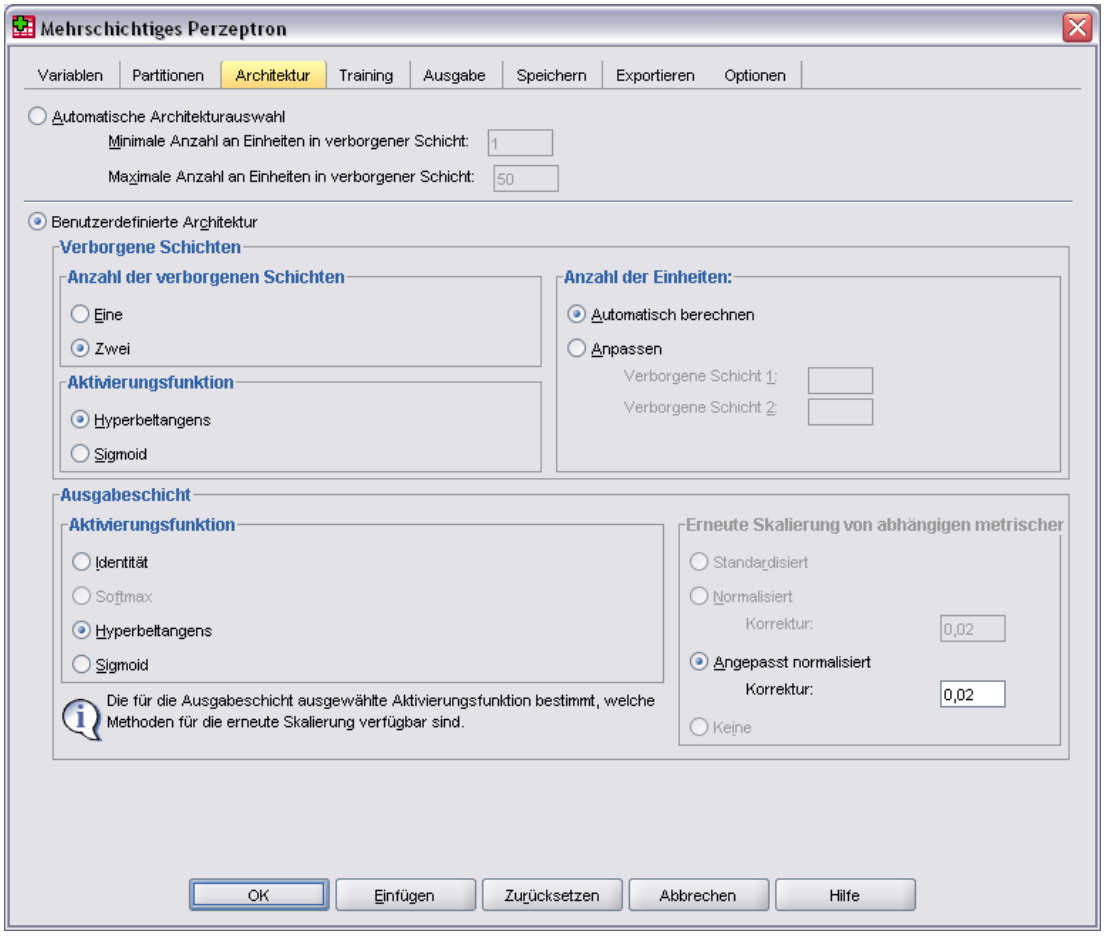
- Wählen Sie *Length of stay [los]* (Aufenthaltsdauer) und *Treatment costs [cost]* (Behandlungskosten) als abhängige Variablen aus.
- Wählen Sie *Age category [agecat]* (Alterskategorie) bis *Taking anti-clotting drugs [anticlot]* (Einnahme von Gerinnungshemmern) und *Time to hospital [time]* (Zeit bis Krankenhaus) und *Surgical complications [comp]* (chirurgische Komplikationen) als Faktoren aus. Um die unten angegebenen Modellergebnisse exakt zu reproduzieren, müssen Sie unbedingt die Reihenfolge der Variablen in der Faktorenliste beibehalten. Dazu kann es hilfreich sein, die einzelnen Einflussvariablen-Sets auszuwählen und sie mithilfe der Schaltfläche (also nicht durch Ziehen und Ablegen) in die Faktorenliste zu verschieben. Alternativ lässt sich durch eine Änderung der Reihenfolge der Variablen leichter die Stabilität der Lösung einschätzen.
- Klicken Sie auf die Registerkarte Partitionen.

Abbildung 4-26
Mehrschichtiges Perzeptron: Registerkarte "Partitionen"



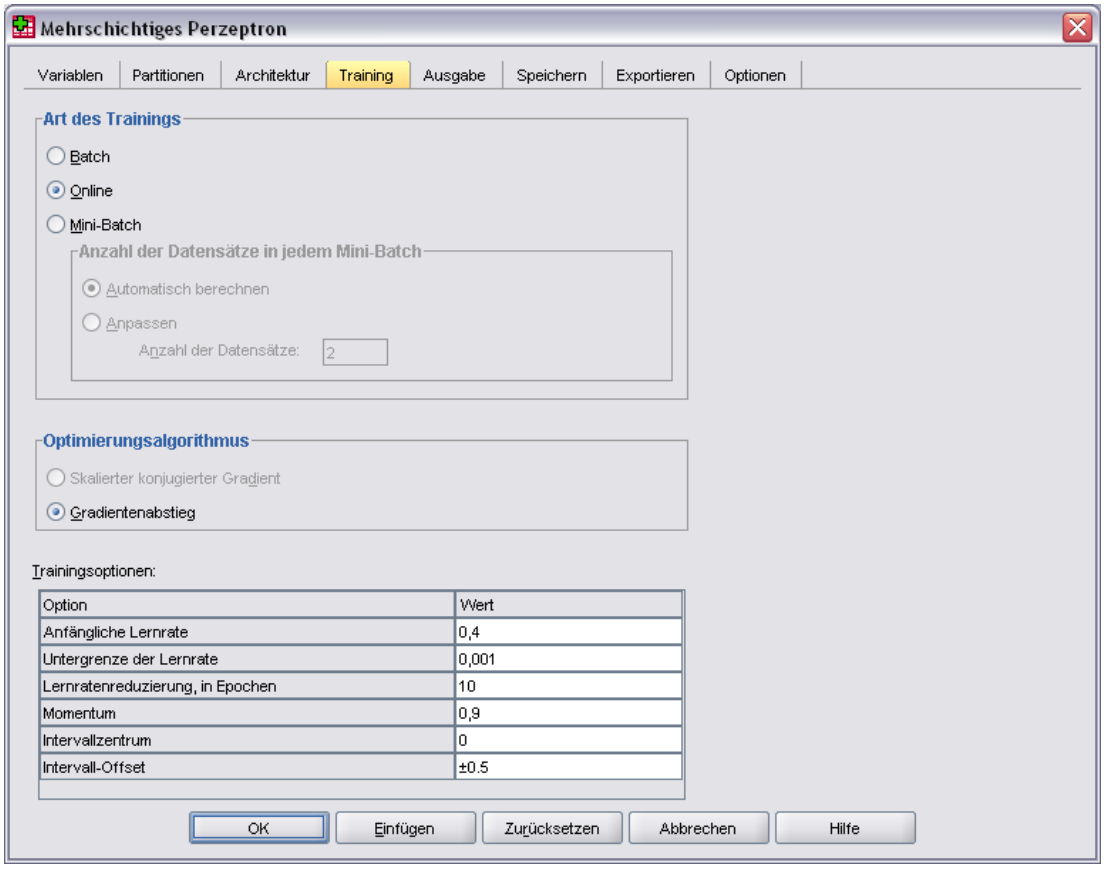
- Geben Sie 2 als relative Anzahl der Fälle ein, die der Teststichprobe zugewiesen werden sollen.
- Geben Sie 1 als relative Anzahl der Fälle ein, die der Holdout-Stichprobe zugewiesen werden sollen.
- Klicken Sie auf die Registerkarte Architektur.

Abbildung 4-27
Mehrschichtiges Perzeptron: Registerkarte "Architektur"



- Wählen Sie Benutzerdefinierte Architektur.
- Wählen Sie Zwei als Anzahl der verborgenen Schichten aus.
- Wählen Sie Hyperbeltangens als Aktivierungsfunktion für die Ausgabeschicht aus. Beachten Sie, dass dadurch die Methode für die erneute Skalierung der abhängigen Variablen automatisch auf Angepasst normalisiert gesetzt wird.
- Klicken Sie auf die Registerkarte Training.

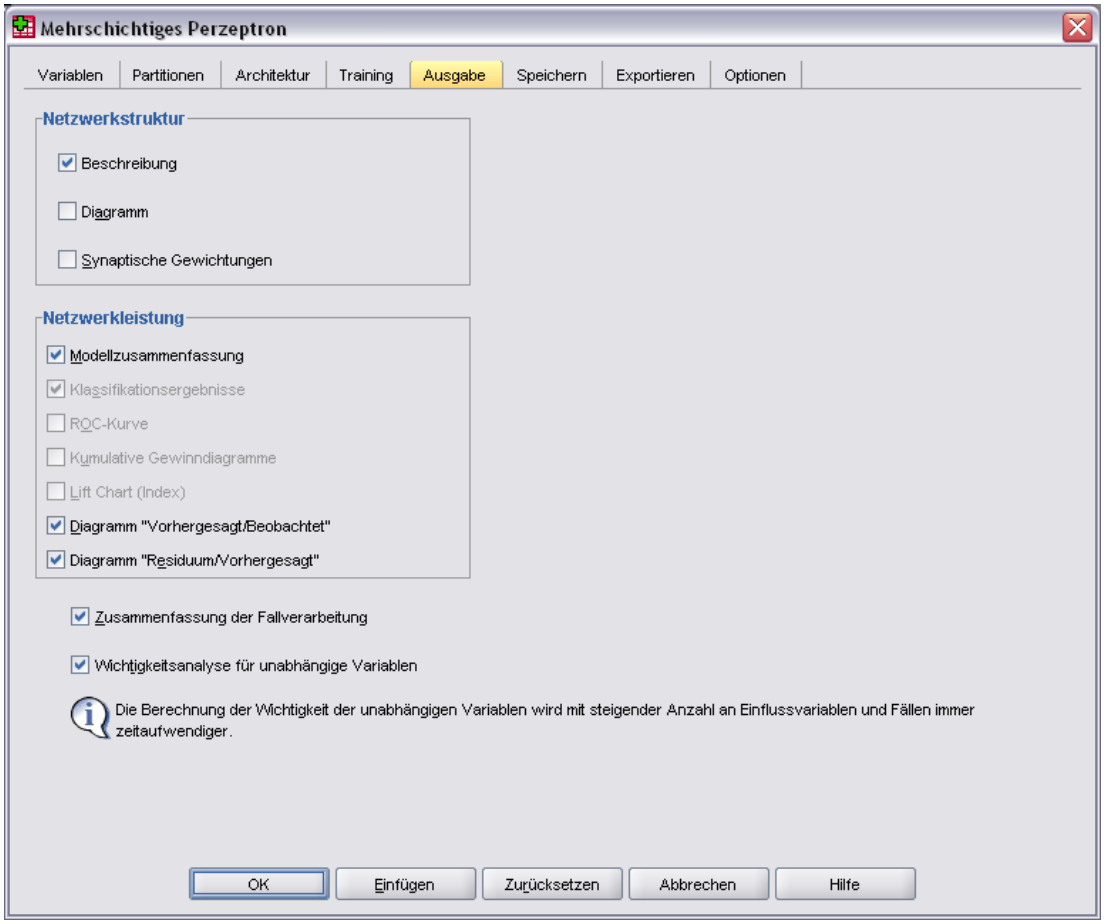
Abbildung 4-28
Mehrschichtiges Perzeptron: Registerkarte "Training"



- Wählen Sie Online als Art des Trainings. Online-Training gilt als leistungsstark für "größere" Daten-Sets mit korrelierten Einflussvariablen. Beachten Sie, dass dadurch Gradientenabstieg automatisch als Optimierungsalgorithmus mit den entsprechenden Standardoptionen festgelegt wird.

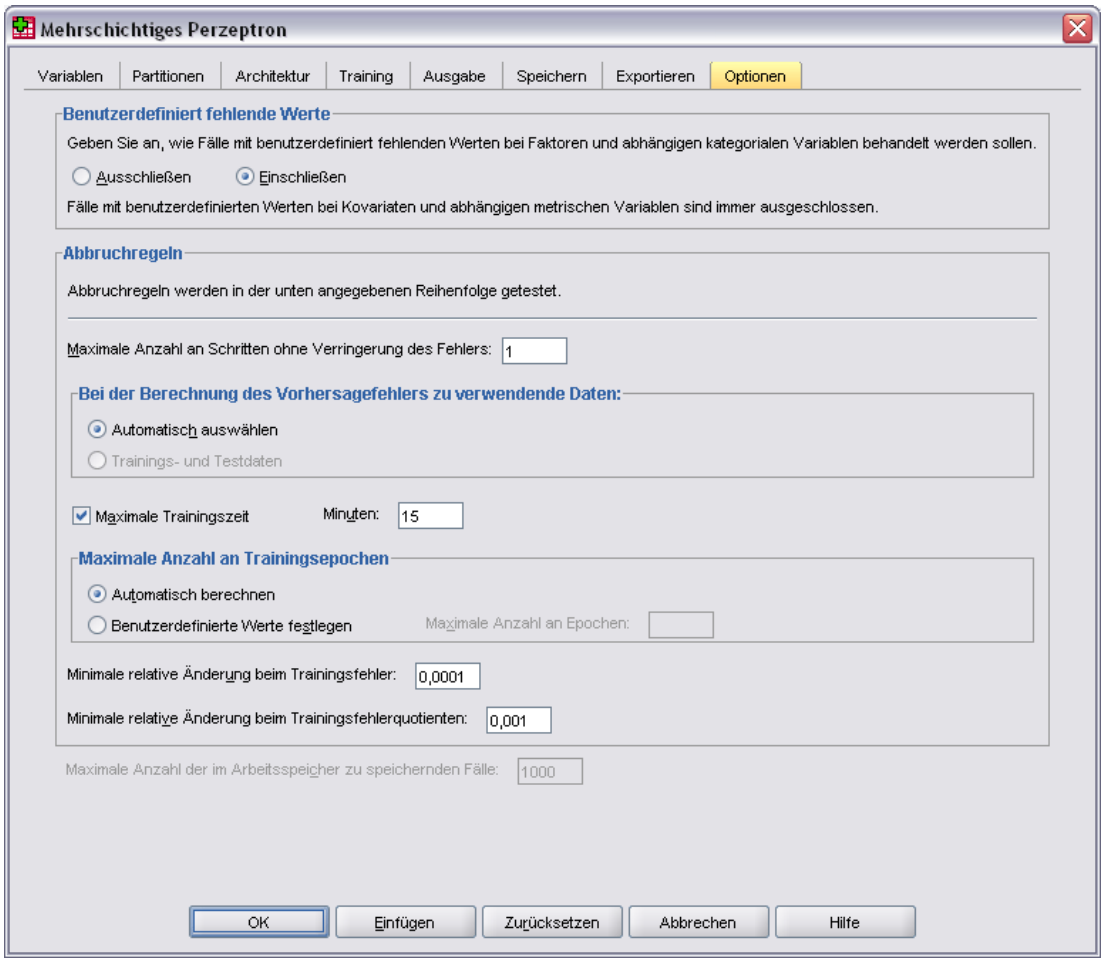
- Klicken Sie auf die Registerkarte Ausgabe.

Abbildung 4-29
Mehrschichtiges Perzeptron: Registerkarte "Ausgabe"



- Deaktivieren Sie die Option Diagramm. Durch die vielen Eingaben würde das Diagramm unüberschaubar.
- Wählen Sie im Gruppenfeld "Netzwerkleistung" die Optionen Diagramm "Vorhergesagt/Beobachtet" und Diagramm "Residuum/Vorhergesagt" aus. Die Klassifikationsergebnisse, die ROC-Kurve, das kumulative Gewinndiagramm und der Lift Chart sind nicht verfügbar, da keine der abhängigen Variablen als kategorial (nominal oder ordinal) behandelt wird.
- Wählen Sie die Option Wichtigkeitsanalyse für unabhängige Variablen.
- Klicken Sie auf die Registerkarte Optionen.

Abbildung 4-30
Mehrschichtiges Perzeptron: Registerkarte "Optionen"

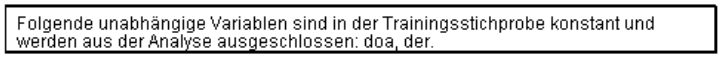


- Wählen Sie Einschließen für benutzerdefinierte Variablen aus. Patienten, bei denen kein chirurgischer Eingriff vorgenommen wurde, weisen benutzerdefiniert fehlende Werte bei der Variablen *Surgical complications* (chirurgische Komplikationen) auf. Dadurch wird sichergestellt, dass die betreffenden Patienten in die Analyse aufgenommen werden.

- Klicken Sie auf OK.

Warnungen

Abbildung 4-31
Warnungen



In der Warnungstabelle ist vermerkt, dass die Variablen *doa* (bereits tot bei Ankunft) und *der* (in Notaufnahme verstorben) in der Trainingsstichprobe konstant sind. Patienten, die bereits beim Eintreffen tot waren oder in der Notaufnahme verstarben, weisen benutzerdefiniert fehlende Werte für *Length of stay* (Aufenthaltsdauer) auf. Da wir *Length of stay* (Aufenthaltsdauer) als metrische Variable für diese Analyse behandeln und Fälle mit benutzerdefiniert fehlenden Werten bei metrischen Variablen ausgeschlossen werden, werden nur Patienten, die nach Verlassen der Notaufnahme noch am Leben waren, in die Stichprobe aufgenommen.

Zusammenfassung der Fallverarbeitung

Abbildung 4-32
Zusammenfassung der Fallverarbeitung

		N	Prozent
Beispiel	Training	5647	70,6%
	Test	1570	19,6%
	Holdout	781	9,8%
Gültig		7998	100,0%
Ausgeschlossen		2002	
Gesamt		10000	

Die Zusammenfassung der Fallverarbeitung zeigt, dass der Trainingsstichprobe 5647, der Teststichprobe 1570 und der Holdout-Stichprobe 781 Fälle zugewiesen wurden. Bei den 2002 Fällen, die aus der Analyse ausgeschlossen wurden, handelt es sich um Patienten, die auf dem Weg ins Krankenhaus oder in der Notaufnahme verstarben.

Netzwerkinformationen

Abbildung 4-33
Netzwerkinformationen

Eingabeschicht	Factors	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
	Anzahl der Einheiten:[a]		63
Verborgene Schicht(en):	Anzahl der verborgenen Schichten		2
	Anzahl der Einheiten in verborgener Schicht 1[a]		12
	Number of Units in Hidden Layer 2[a]		9
	Aktivierungsfunktion		Hyperbeltangens
Ausgabeschicht	Dependent Variables	1	Length of stay
		2	Treatment costs
	Anzahl der Einheiten:		2
	Rescaling Method for Scale Dependents		Adjusted Normalized
	Aktivierungsfunktion		Hyperbeltangens
	Fehlerfunktion		Quadratsumme

a. Ohne die Verzerrungseinheit

In der Tabelle “Netzwerkinformationen” werden Informationen zum neuronalen Netzwerk angezeigt. Anhand dieser Tabelle können Sie sich vergewissern, dass die Spezifikationen korrekt sind. Beachten Sie hier insbesondere Folgendes:

- Die Anzahl der Einheiten in der Eingabeschicht ist die Gesamtzahl der Faktorstufen (es gibt keine Kovariaten).
- Es wurden zwei verborgene Schichten angefordert und die Prozedur hat 12 Einheiten in der ersten verborgenen Schicht und 9 in der zweiten verborgenen Schicht ausgewählt.

- Für jede der metrischen abhängigen Variablen wurde eine separate Ausgabeeinheit erstellt. Diese werden mit der Methode “Angepasst normalisiert” erneut skaliert. Dazu muss die Aktivierungsfunktion “Hyperbeltangens” für die Ausgabeschicht verwendet werden.
- Ein Quadratsummenfehler wird gemeldet, da die abhängigen Variablen metrisch sind.

Modellzusammenfassung

Abbildung 4-34
Modellzusammenfassung

Training	Quadratsummenfehler		91,812
	Durchschnittlicher relativer Gesamtfehler		,083
	Relativer Fehler für abhängige metrische Variablen	Length of stay	,131
		Treatment costs	,033
	Verwendete Abbruchregel		1 aufeinander folgende(r) Schritt(e) ohne Verringerung des Fehlers[a]
	Trainingszeit		00:00:18.055
Test	Quadratsummenfehler		26,798
	Durchschnittlicher relativer Gesamtfehler		,088
	Relativer Fehler für abhängige metrische Variablen	Length of stay	,141
		Treatment costs	,033
Prüfung (Holdout)	Durchschnittlicher relativer Gesamtfehler		,099
	Relativer Fehler für abhängige metrische Variablen	Length of stay	,154
		Treatment costs	,041

a. Fehlerberechnungen beruhen auf der Teststichprobe.

In der Modellzusammenfassung werden Informationen zu den Ergebnissen des Trainings und der Anwendung des endgültigen Netzwerks auf die Holdout-Stichprobe angezeigt.

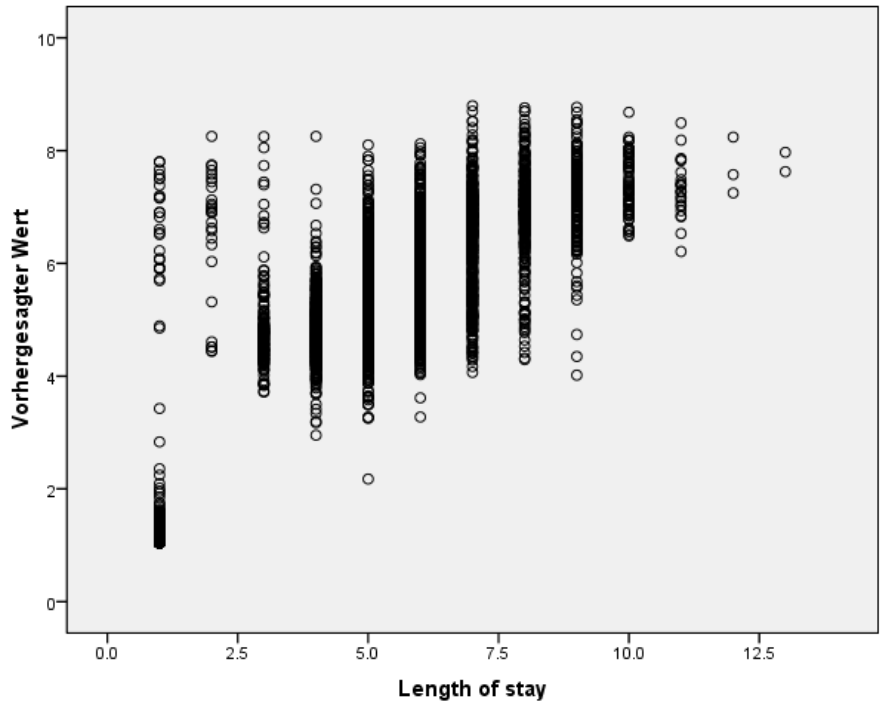
- Ein Quadratsummenfehler wird angezeigt, da die Ausgabeschicht metrische abhängige Variablen aufweist. Dies ist die Fehlerfunktion, die das Netzwerk während des Trainings zu minimieren versucht. Beachten Sie, dass die Quadratsummen und alle folgenden Fehlerwerte für die neu skalierten Werte der abhängigen Variablen berechnet werden.
- Der relative Fehler für die einzelnen metrischen abhängigen Variablen ist jeweils der Quotient aus dem Quadratsummenfehler für die abhängige Variable und dem Quadratsummenfehler für das “Null”-Modell, in dem der Mittelwert der abhängigen Variablen als vorhergesagter Wert für die einzelnen Fälle verwendet wird. In den Vorhersagen von *Length of stay* (Aufenthaltsdauer) scheint der Fehler größer zu sein als in *Treatment costs* (Behandlungskosten).
- Der durchschnittliche Gesamtfehler ist der Quotient aus dem Quadratsummenfehler für alle abhängigen Variablen und dem Quadratsummenfehler für das “Null”-Modell, in dem die Mittelwerte der abhängigen Variablen als vorhergesagte Werte für die einzelnen Fälle verwendet werden. In diesem Beispiel liegt der durchschnittliche Gesamtfehler zufälligerweise nahe bei dem Durchschnitt der relativen Fehler. Dies ist jedoch keineswegs immer der Fall.

Der durchschnittliche relative Gesamtfehler und die relativen Fehler sind in der Trainigs-, Test- und Holdout-Stichprobe relativ konstant, wodurch Sie mit einer gewissen Zuversicht davon ausgehen können, dass das Modell nicht übertrainiert ist und der Fehler in zukünftigen Fällen, die vom Netzwerk gescort werden, im Bereich des in dieser Tabelle angegebenen Fehlers liegt.

- Der Schätzalgorithmus wurde angehalten, da der Fehler nach einem Schritt im Algorithmus nicht kleiner wurde.

Diagramme vom Typ "Vorhergesagt/Beobachtet"

Abbildung 4-35
Diagramm "Vorhergesagt/Beobachtet" für "Length of stay" (Aufenthaltsdauer)

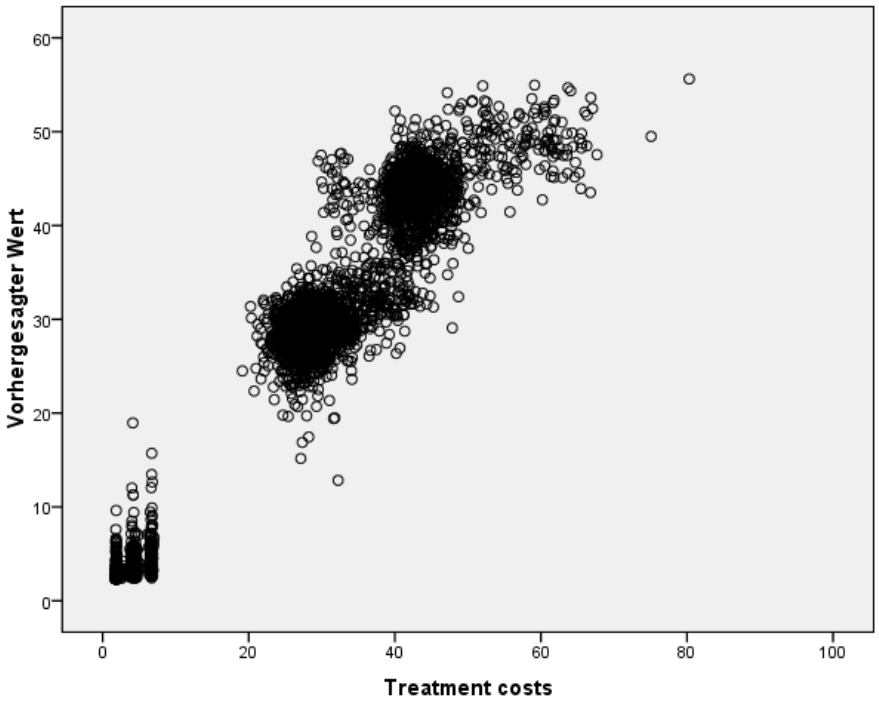


Bei metrischen abhängigen Variablen zeigt das Diagramm "Vorhergesagt/Beobachtet" für die Kombination aus Trainings- und Teststichprobe ein Streudiagramm der vorhergesagten Werte auf der *y*-Achse in Abhängigkeit von den beobachteten Werten auf der *x*-Achse an. Idealerweise sollten die Werte ungefähr entlang einer 45-Grad-Linie liegen, die im Ursprung beginnt. Die Punkte in diesem Diagramm bilden vertikale Linien an jeder beobachteten Anzahl von Tagen der Variablen *Length of stay* (Aufenthaltsdauer).

Das Diagramm erweckt den Eindruck, dass das Netzwerk recht gute Arbeit bei der Vorhersage von *Length of stay* (Aufenthaltsdauer) leistet. Der allgemeine Trend des Streudiagramms liegt abseits der idealen 45-Grad-Linie, dahingehend, dass die Vorhersagen für eine beobachtete Aufenthaltsdauer von unter fünf Tagen dazu neigen, die Aufenthaltsdauer zu überschätzen, wohingegen die Prognosen für eine beobachtete Aufenthaltsdauer von mehr als sechs Tagen die Aufenthaltsdauer tendenziell unterschätzen.

Bei dem Patientencluster im linken unteren Bereich des Diagramms handelt es sich vermutlich um Patienten, die nicht operiert wurden. Außerdem befindet sich ein Cluster von Patienten im linken oberen Bereich des Diagramms, wo die beobachtete Aufenthaltsdauer ein bis drei Tage beträgt, die vorhergesagten Werte jedoch wesentlich höher liegen. Bei diesen Fällen handelt es sich wahrscheinlich um Patienten, die nach der Operation im Krankenhaus verstarben.

Abbildung 4-36
Diagramm "Vorhergesagt/Beobachtet" für "Treatment costs" (Behandlungskosten)



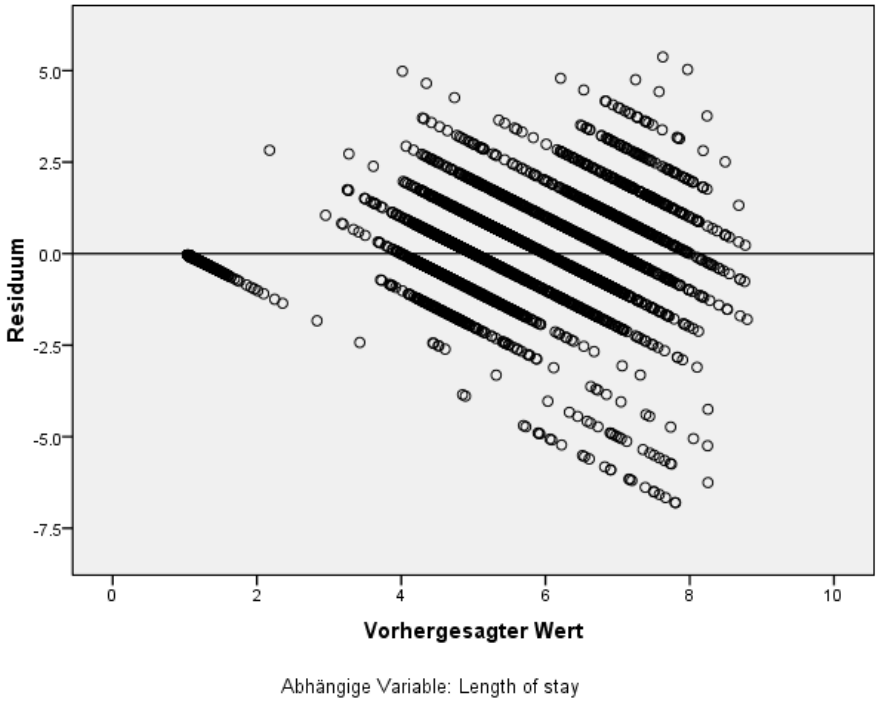
Das Netzwerk scheint auch recht gute Arbeit bei der Vorhersage der *Treatment costs* (Behandlungskosten) zu leisten. Es scheint drei wichtige Patientencluster zu geben:

- Links unten befinden sich vor allem Patienten, die nicht operiert wurden. Die für diese Patienten anfallenden Kosten sind relativ niedrig und sind nach der Art der in der Notaufnahme verabreichten *Clot-dissolving drugs [clotsolv]* (Gerinnungshemmer) differenziert.
- Der nächste Patientencluster weist Behandlungskosten von ungefähr 30.000 Dollar auf. Hierbei handelt es sich um Patienten, die einer Ballondilatation (Perkutane transluminale Coronarangioplastie (PTCA)) unterzogen wurden.
- Der letzte Cluster schließlich weist Behandlungskosten von mehr als 40.000 Dollar auf. Hierbei handelt es sich um Patienten, die einen Koronararterien-Bypass (CABG) erhielten. Diese Operation ist etwas teurer als PTCA und die Patienten müssen nach der Operation länger stationär im Krankenhaus behandelt werden, was die Kosten weiter in die Höhe treibt.

Außerdem gibt es eine Reihe von Fällen mit Kosten von über 50.000 Doller, die das Netzwerk nicht gut vorhersagt. Hierbei handelt es sich um Patienten, bei denen während der OP Komplikationen auftraten, was zu höheren Operationskosten und längerer Aufenthaltsdauer führen kann.

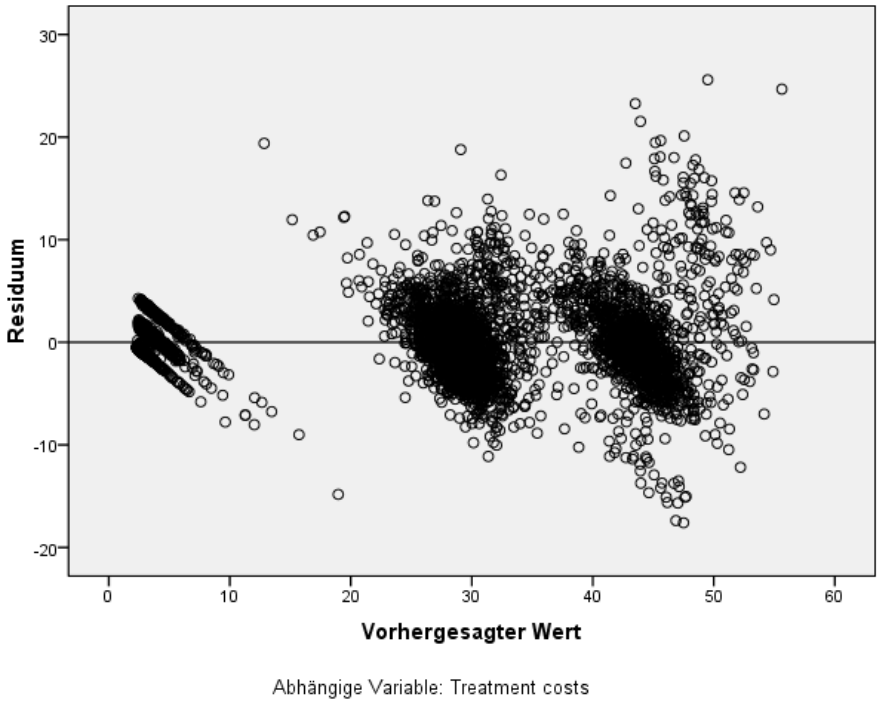
Diagramme vom Typ "Residuum/Vorhergesagt"

Abbildung 4-37
Diagramm "Residuum/Vorhergesagt" für "Length of stay" (Aufenthaltsdauer)



Das Diagramm "Residuum/Vorhergesagt" zeigt ein Streudiagramm des Residuums (beobachteter Wert minus vorhergesagter Wert) auf der *y*-Achse in Abhängigkeit vom vorhergesagten Wert auf der *x*-Achse an. Jede diagonale Linie in diesem Diagramm entspricht einer vertikalen Linie im Diagramm "Vorhergesagt/Beobachtet", und der Verlauf von Übervorhersage zu Untervorhersage der Aufenthaltsdauer mit zunehmender beobachteter Aufenthaltsdauer wird deutlicher erkennbar.

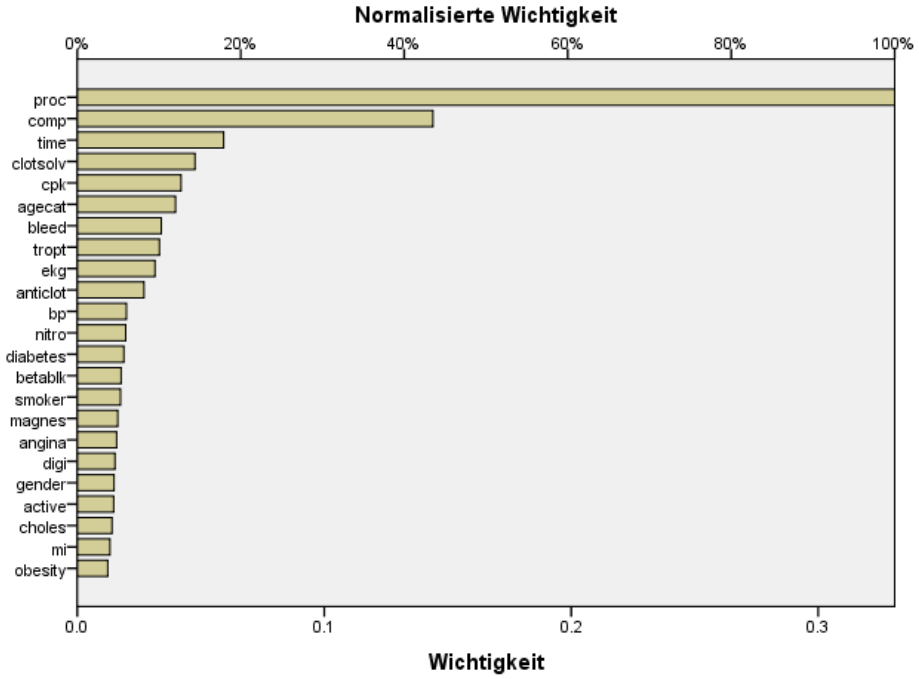
Abbildung 4-38
Diagramm "Vorhergesagt/Beobachtet" für "Treatment costs" (Behandlungskosten)



Ebenso zeigt das Diagramm "Residuum/Vorhergesagt" für jeden der drei im Diagramm "Vorhergesagt/Beobachtet" für *Treatment costs* (Behandlungskosten) ersichtlichen Patientencluster bei zunehmenden beobachteten Kosten einen Verlauf von Übervorhersage zu Untervorhersage. Die Patienten, bei denen während der CABG Komplikationen auftraten, sind immer noch deutlich sichtbar, aber nun lassen sich auch leichter die Patienten erkennen, bei denen während der PTCA Komplikationen auftreten; sie erscheinen als Untercluster ein wenig rechts und oberhalb der Hauptgruppe der PTCA-Patienten um die 30.000-Dollar-Marke auf der *x*-Achse.

Wichtigkeit der unabhängigen Variablen

Abbildung 4-39
Wichtigkeitsdiagramm für die unabhängigen Variablen



Das Wichtigkeitsdiagramm zeigt, dass die Ergebnisse vor allem vom durchgeführten Operationsverfahren abhängen, gefolgt davon, ob Komplikationen auftraten. Die anderen Einflussvariablen folgen in weitem Abstand. Die Bedeutung des Operationsverfahrens ist deutlich in den Diagrammen für *Treatment costs* (Behandlungskosten) erkennbar und etwas weniger deutlich bei *Length of stay* (Aufenthaltsdauer), während der Effekt von Komplikationen auf *Length of stay* (Aufenthaltsdauer) bei den Patienten mit den höchsten Werten für die beobachtete Aufenthaltsdauer sichtbar zu sein scheint.

Zusammenfassung

Das Netzwerk scheint gute Arbeit bei der Vorhersage von Werten für “typische” Patienten zu leisten, erfasst jedoch keine Patienten, die nach der Operation verstarben. Eine Möglichkeit, dieses Problem anzugehen, besteht darin, mehrere Netzwerke zu erstellen. Ein Netzwerk zur Vorhersage des Patientenergebnisses, vielleicht einfach nur, ob der Patient überlebte oder nicht, und dann separate Netzwerke, die abhängig davon, ob der Patient überlebte, *Treatment costs* (Behandlungskosten) und *Length of stay* (Aufenthaltsdauer) vorhersagen. Anschließend können Sie die Netzwerkergebnisse kombinieren, um vermutlich bessere Vorhersagen zu erzielen. Ein ähnlicher Ansatz könnte zur Lösung des Problems der Untervorhersage von Kosten und Aufenthaltsdauer von Patienten mit Komplikationen während der Operation verfolgt werden.

Empfohlene Literatur

In folgenden Texten finden Sie weitere Informationen zu neuronalen Netzwerken und mehrschichtigen Perzeptronen:

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*, 3rd (Hg.). Oxford: Oxford University Press.

Fine, T. L. 1999. *Feedforward Neural Network Methodology*, 3rd (Hg.). New York: Springer-Verlag.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*, 2nd (Hg.). New York: Macmillan College Publishing.

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

Kapitel
5

Radiale Basisfunktion

Die Prozedur “Radiale Basisfunktion” (RBF) erstellt ein Vorhersagemodell für eine oder mehrere abhängige Variablen (Zielvariablen), das auf den Werten der Einflussvariablen beruht.

Verwenden der radialen Basisfunktion zum Klassifizieren von Telekommunikationskunden

Ein Telekommunikationsanbieter hat seinen Kundenstamm in Muster der Servicenutzung eingeteilt und die Kunden in vier Gruppen kategorisiert. Wenn demografische Daten zum Vorhersagen der Gruppenzugehörigkeit verwendet werden können, sind angepasste Angebote für die einzelnen potenziellen Kunden möglich.

Angenommen, die Informationen über die derzeitigen Kunden befinden sich in der Datei *telco.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 89. Verwenden Sie die radiale Basisfunktion zum Klassifizieren von Kunden.

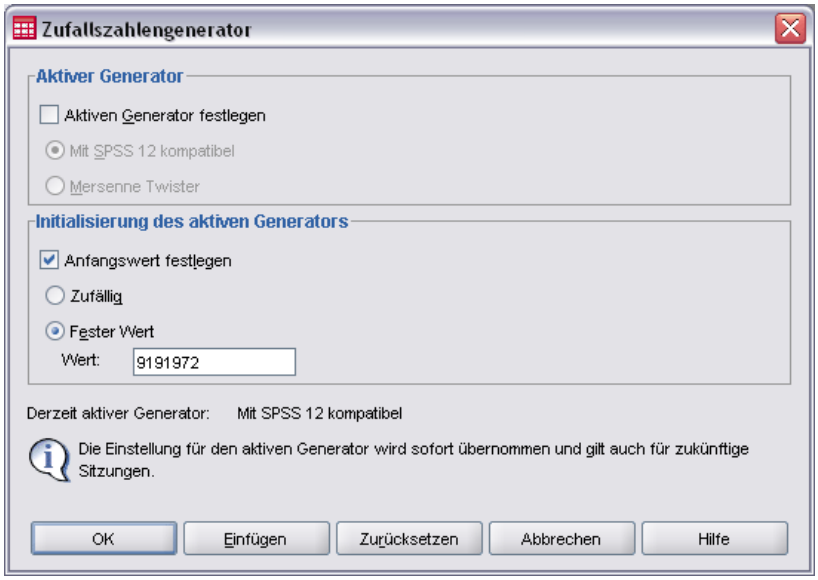
Vorbereiten der Daten für die Analyse

Durch die Festlegung des Startwerts können sie die Analyse exakt reproduzieren.

- Zur Festlegung des Startwerts wählen Sie die folgenden Menübefehle aus:

Transformieren
Zufallszahlengeneratoren...

Abbildung 5-1
Dialogfeld "Zufallszahlengenerator"



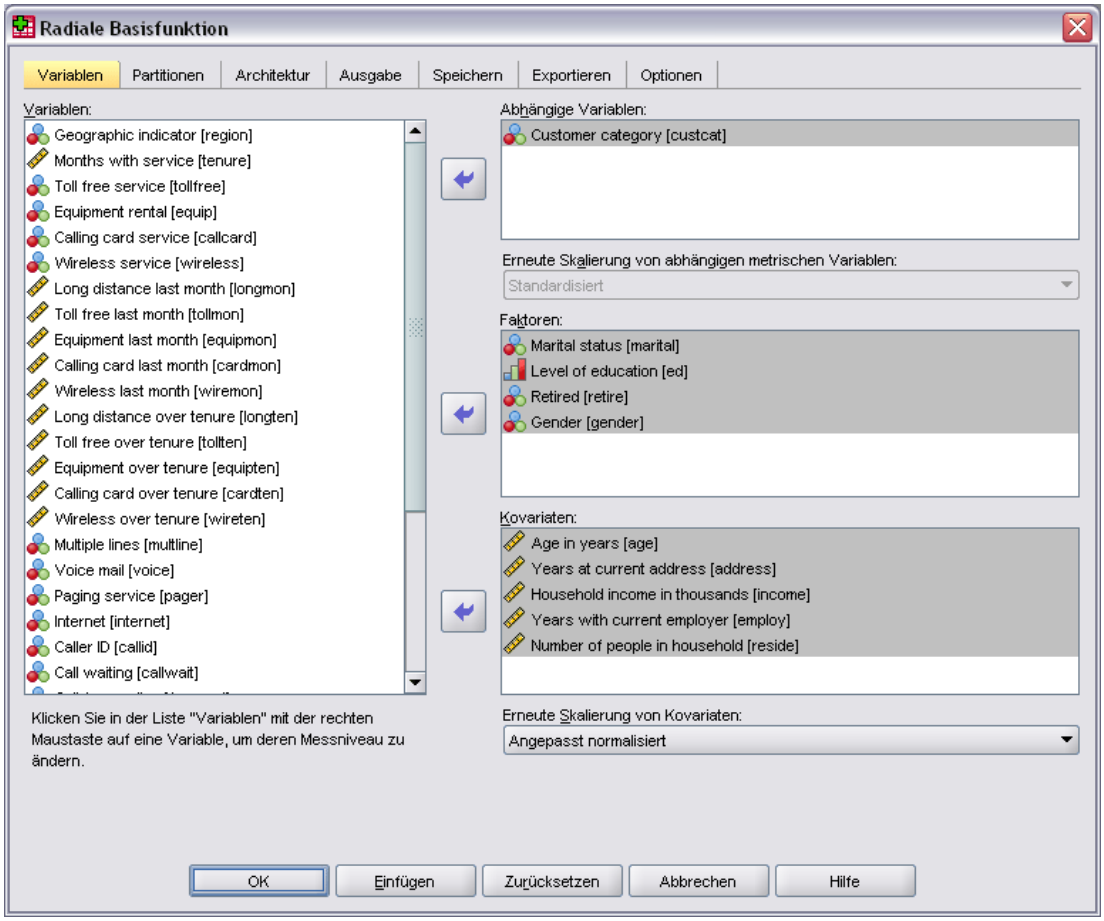
- Wählen Sie Anfangswert festlegen.
- Wählen Sie Fester Wert und geben Sie 9191972 als Wert ein.
- Klicken Sie auf OK.

Durchführung der Analyse

- Um eine Analyse vom Typ "Radiale Basisfunktion" durchzuführen, wählen Sie folgende Optionen aus den Menüs aus:

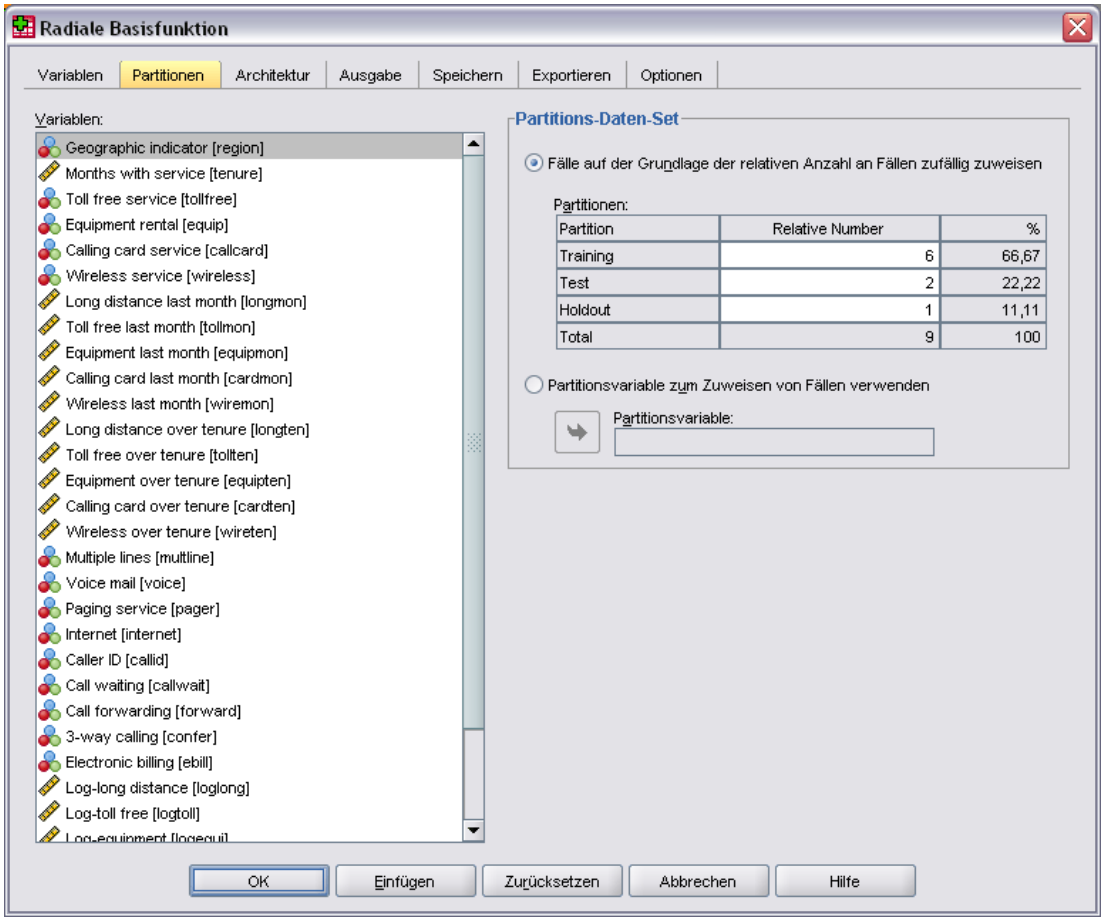
  Analysieren
  Neuronale Netze
  Radiale Basisfunktion...

Abbildung 5-2
Radiale Basisfunktion: Registerkarte "Variablen"



- Wählen Sie *Customer category [custcat]* (Kundenkategorie) als abhängige Variable aus.
- Wählen Sie *Marital status [marital]* (Familienstand), *Level of education [ed]* (Bildungsniveau), *Retired [retire]* (Im Ruhestand) und *Gender [gender]* (Geschlecht) als Faktoren aus.
- Wählen Sie *Age in years [age]* (Alter in Jahren) bis *Number of people in household [reside]* (Haushaltsgröße) als Kovariaten aus.
- Wählen Sie Angepasst normalisiert als Methode für die Neuskalierung von Kovariaten aus.
- Klicken Sie auf die Registerkarte Partitionen.

Abbildung 5-3
Radiale Basisfunktion: Registerkarte "Partitionen"



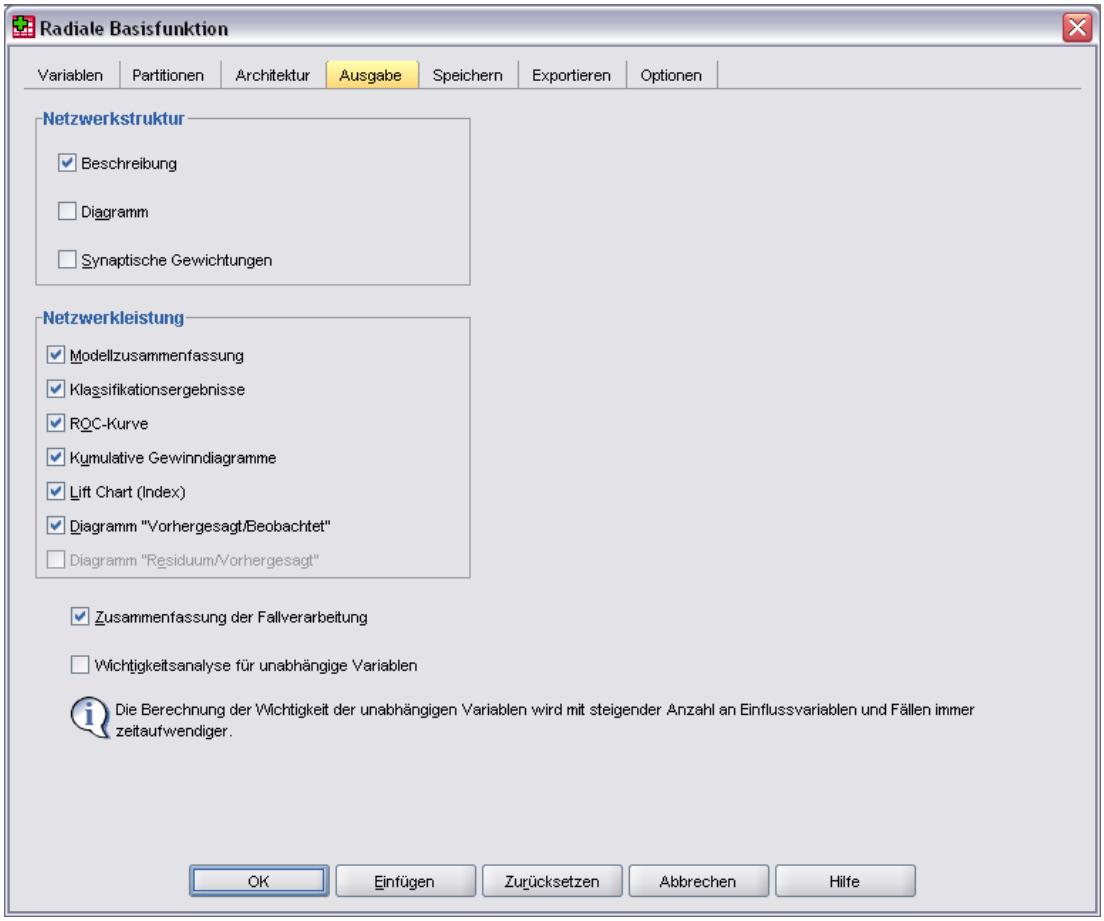
Durch die Angabe von Werten für die relative Anzahl der Fälle ist es einfach, fraktionale Partitionen zu erstellen, für die die Angabe von Prozentsätzen schwierig wäre. Angenommen, Sie möchten 2/3 des Daten-Sets der Trainingsstichprobe zuweisen und 2/3 der übrigen Fälle der Teststichprobe.

- Geben Sie 6 als relative Zahl für die Trainingsstichprobe ein.
- Geben Sie 2 als relative Zahl für die Teststichprobe ein.
- Geben Sie 1 als relative Zahl für die Holdout-Stichprobe ein.

Insgesamt wurden 9 relative Fälle angegeben. 6/9 = 2/3, also ca. 66,67 %, werden der Trainingsstichprobe zugewiesen; 2/9, also ca. 22,22 %, der Teststichprobe, 1/9, also ca. 11,11 % der Holdout-Stichprobe.

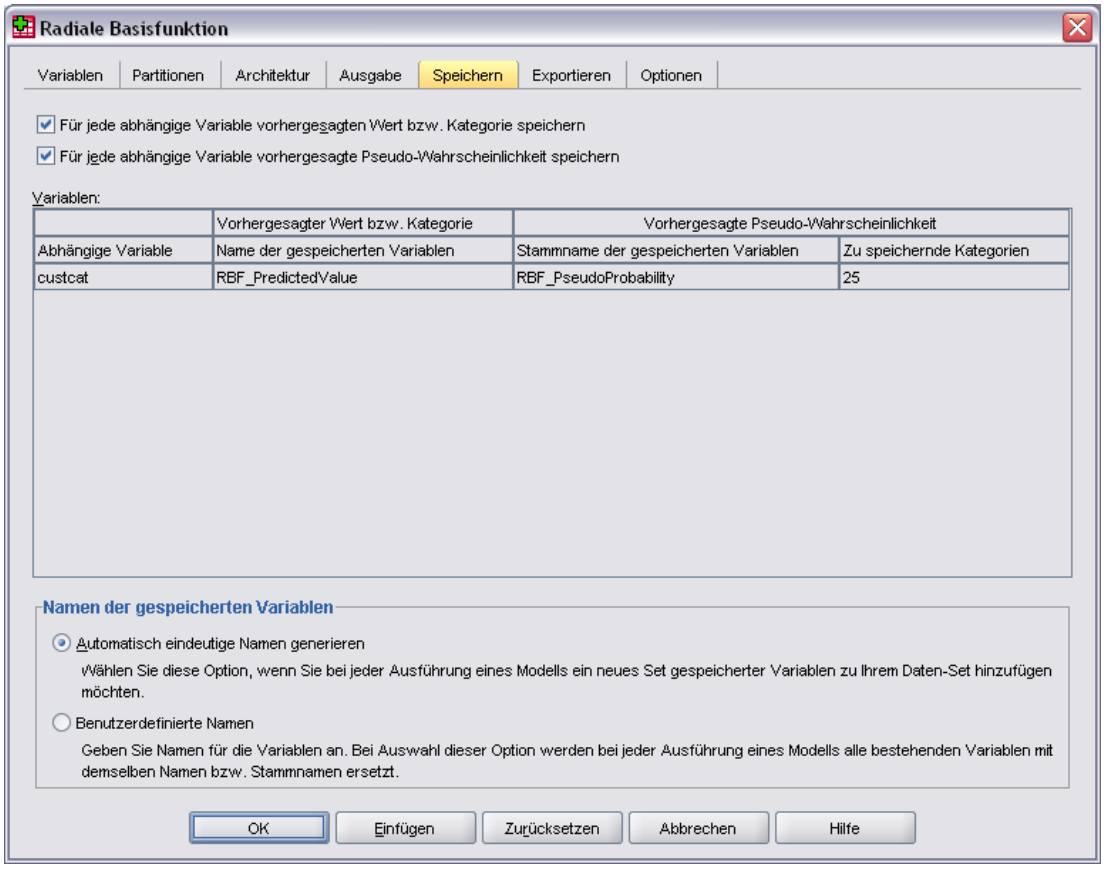
- Klicken Sie auf die Registerkarte Ausgabe.

Abbildung 5-4
Radiale Basisfunktion: Registerkarte "Ausgabe"



- Heben Sie im Gruppenfeld "Netzwerkstruktur" die Auswahl der Option Diagramm auf.
- Wählen Sie im Gruppenfeld "Netzwerkleistung" die Optionen ROC-Kurve, Kumulatives Gewinndiagramm, Lift Chart (Index) und Diagramm "Vorhergesagt/Beobachtet".
- Klicken Sie auf die Registerkarte Speichern.

Abbildung 5-5
Radiale Basisfunktion: Registerkarte "Speichern"



- Aktivieren Sie Für jede abhängige Variable vorhergesagten Wert bzw. Kategorie speichern und Für jede abhängige Variable vorhergesagte Pseudo-Wahrscheinlichkeit speichern.
- Klicken Sie auf OK.

Zusammenfassung der Fallverarbeitung

Abbildung 5-6
Zusammenfassung der Fallverarbeitung

		N	Prozent
Beispiel	Training	665	66,5%
	Test	224	22,4%
	Prüfung (Holdout)	111	11,1%
Gültig		1000	100,0%
Ausgeschlossen		0	
Gesamt		1000	

Die Zusammenfassung der Fallverarbeitung zeigt, dass der Trainingsstichprobe 665, der Teststichprobe 224 und der Holdout-Stichprobe 111 Fälle zugewiesen wurden. Es wurden keine Fälle aus der Analyse ausgeschlossen.

Netzwerkinformationen

Abbildung 5-7
Netzwerkinformationen

Eingabeschicht	Factors	1	Marital status
		2	Level of education
		3	Retired
		4	Gender
	Covariates	1	Age in years
		2	Years at current address
		3	Household income in thousands
		4	Years with current employer
		5	Number of people in household
	Anzahl der Einheiten:[a]		16
	Rescaling Method for Covariates		Adjusted Normalized
Verborgene Schicht(en):	Anzahl der verborgenen Schichten		1
	Anzahl der Einheiten in verborgener Schicht 1[a]		5
	Aktivierungsfunktion		Hyperbeltangens
Ausgabeschicht	Dependent Variables	1	Customer category
	Anzahl der Einheiten:		4
	Aktivierungsfunktion		Softmax
	Fehlerfunktion		Kreuzentropie

a. Ohne die Verzerrungseinheit

In der Tabelle "Netzwerkinformationen" werden Informationen zum neuronalen Netzwerk angezeigt. Anhand dieser Tabelle können Sie sich vergewissern, dass die Spezifikationen korrekt sind. Beachten Sie hier insbesondere Folgendes:

- Die Anzahl der Einheiten in der Eingabeschicht ist die Anzahl der Kovariaten plus die Gesamtzahl der Faktorstufen; für jede Kategorie von *Marital status* (Familienstand), *Level of education* (Bildungsniveau), *Retired* (Ruhestand) und *Gender* (Geschlecht) wird eine gesonderte Einheit erstellt und keine der Kategorien wird als "redundante" Einheit betrachtet, wie dies bei vielen Modellierungsprozeduren üblich ist.
- Ebenso wird für jede Kategorie von *Customer category* (Kundenkategorie) eine separate Ausgabeeinheit erstellt (für insgesamt 4 Einheiten in der Ausgabeschicht).
- Die Kovariaten werden mit der Methode "Angepasst normalisiert" neu skaliert.
- Die automatische Architekturauswahl hat 9 Einheiten in der verborgenen Schicht ausgewählt.
- Bei allen anderen Netzwerkinformationen werden die Standardwerte für die Prozedur verwendet.

Modellzusammenfassung

Abbildung 5-8
Modellzusammenfassung

Training	Quadratsummenfehler	235.969
	Prozentsatz der falschen Vorhersagen	61.8%
	Trainingszeit	00:00:04.297
Test	Quadratsummenfehler	80.851[a]
	Prozentsatz der falschen Vorhersagen	62.9%
Prüfung (Holdout)	Prozentsatz der falschen Vorhersagen	59.5%

Abhängige Variable: Customer category

a. Die Anzahl der verborgenen Einheiten wird durch das Testdatenkriterium bestimmt: Die "beste" Anzahl verborgener Einheiten ist diejenige, die den kleinsten Fehler in den Testdaten ergibt.

In der Modellzusammenfassung werden Informationen zu den Ergebnissen des Trainings, des Tests und der Anwendung des endgültigen Netzwerks auf die Holdout-Stichprobe angezeigt.

- Der Quadratsummenfehler wird angezeigt, da dieser immer für RBF-Netzwerke verwendet wird. Dies ist die Fehlerfunktion, die das Netzwerk während des Training und Tests zu minimieren versucht.
- Der Prozentsatz der falschen Vorhersagen wird aus der Klassifikationsmatrix entnommen und in dem zugehörigen Thema eingehender erörtert.

Klassifikation

Abbildung 5-9
Klassifikation

		Vorhergesagt				
Beispiel	Beobachtet	Basic service	E-service	Plus service	Total service	Percent Correct
Training	Basic service	64	0	66	45	36,6%
	E-service	22	1	57	61	,7%
	Plus service	47	0	104	34	56,2%
	Total service	29	1	49	85	51,8%
	Overall Percent	24,4%	,3%	41,5%	33,8%	38,2%
Testing	Basic service	18	0	26	15	30,5%
	E-service	15	0	16	22	,0%
	Plus service	11	0	39	15	60,0%
	Total service	4	0	17	26	55,3%
	Overall Percent	21,4%	,0%	43,8%	34,8%	37,1%
Holdout	Basic service	11	0	11	10	34,4%
	E-service	4	0	9	10	,0%
	Plus service	10	0	19	2	61,3%
	Total service	5	0	5	15	60,0%
	Overall Percent	27,0%	,0%	39,6%	33,3%	40,5%

Abhängige Variable: Customer category

Die Klassifikationsmatrix zeigt die praktischen Ergebnisse der Verwendung des Netzwerks. Für jeden Fall ist die vorhergesagte Antwort die Kategorie mit der höchsten vorhergesagten Pseudo-Wahrscheinlichkeit.

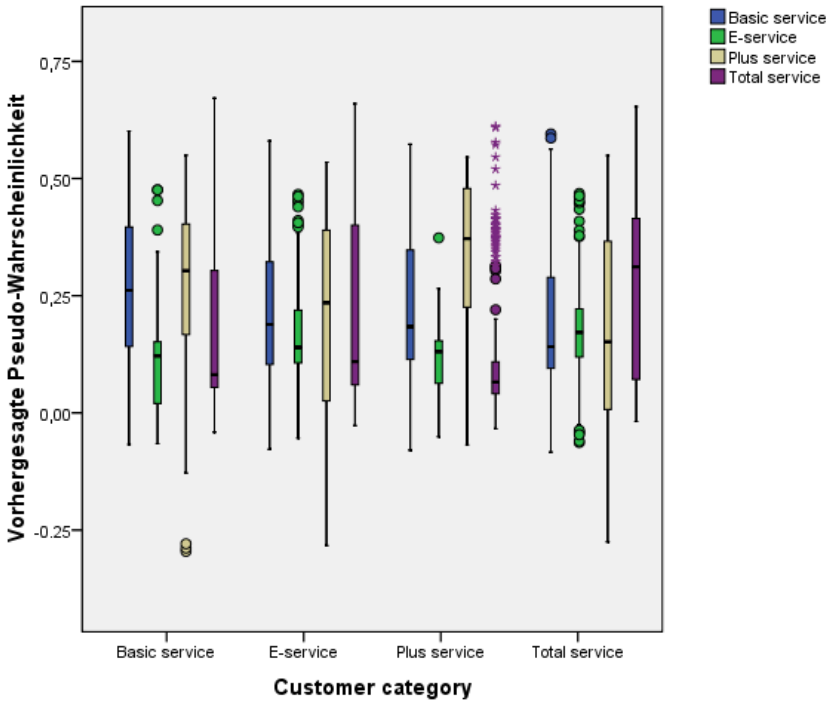
- Die Zellen auf der Diagonale stellen korrekte Vorhersagen dar.
- Die Zellen abseits der Diagonale stellen falsche Vorhersagen dar.

Mit den beobachteten Daten würde das "Nullmodell" (d. h. ein Modell ohne Einflussvariablen) alle Kunden in die Modalgruppe, *Plus service*, einordnen. Das Nullmodell wäre daher in 281/1000 = 28,1 % der Fälle richtig. Das RBF-Netzwerk erfasst weitere 10,1 % der Kunden, also 38,2 %. Insbesondere ist das Modell beim Ermitteln von *Plus service*- und *Total service*-Kunden überlegen. Bei der Klassifikation von *E-service*-Kunden liegt dagegen ein außerordentlich schlechter Wert vor. Möglicherweise müssen Sie eine weitere Einflussvariable finden, um diese Kunden auseinanderzuhalten. In Anbetracht der Tatsache, dass diese Kunden am häufigsten als *Plus service*- und *Total service*-Kunden fehlklassifiziert werden, besteht eine weitere Alternative darin, dass das Unternehmen einfach versucht, potenziellen Kunden, die normalerweise in die Kategorie *E-service* fallen würden, höherwertige Dienstleistungen zu verkaufen.

Die Klassifizierung anhand der Fälle, mit denen das Modell erstellt wurde, gerät jedoch leicht zu "optimistisch", da die Klassifizierungsrate aufgebläht ist. Die Holdout-Stichprobe erleichtert die Validierung der Modells; hier wurden 40,2 % der Fälle korrekt vom Modell klassifiziert. Obwohl die Holdout-Stichprobe relativ klein ist, legt dies nahe, dass Ihr Modell in der Tat in ungefähr zwei von fünf Fällen korrekt ist.

Diagramm "Vorhergesagt/Beobachtet"

Abbildung 5-10
Vorhergesagt/Beobachtet, Diagramm



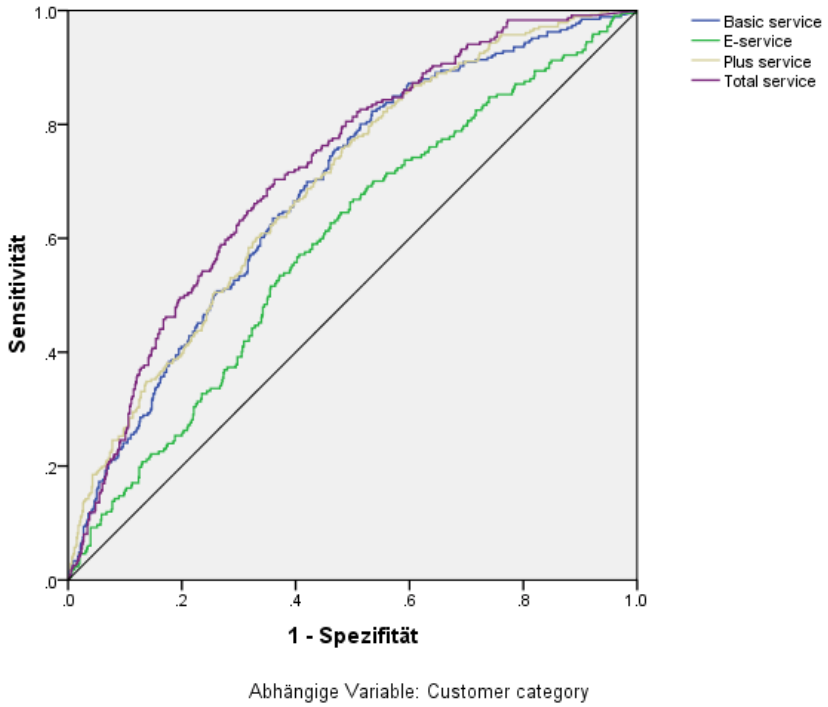
Für kategoriale abhängige Variablen zeigt das Diagramm “Vorhergesagt/Beobachtet” gruppierte Boxplots vorhergesagter Pseudo-Wahrscheinlichkeiten für die Kombination aus Trainings- und Teststichprobe an. Die *x*-Achse entspricht den beobachteten Antwortkategorien und die Legende entspricht vorhergesagten Kategorien. Somit gilt:

- Der Boxplot ganz links zeigt für Fälle mit der beobachteten Kategorie *Basic service* die vorhergesagte Pseudo-Wahrscheinlichkeit für die Kategorie *Basic service*.
- Der nächste Boxplot zeigt für Fälle mit der beobachteten Kategorie *Basic service* die vorhergesagte Pseudo-Wahrscheinlichkeit für die Kategorie *E-service*.
- Der dritte Boxplot zeigt für Fälle mit der beobachteten Kategorie *Basic service* die vorhergesagte Pseudo-Wahrscheinlichkeit für die Kategorie *Plus service*. Aus der Klassifikationsmatrix wissen wir, dass ungefähr so viele *Basic service*-Kunden als *Plus service* fehlklassifiziert wurden, wie korrekt als *Basic service*-Kunden klassifiziert wurden; daher entspricht dieser Boxplot ungefähr dem Boxplot ganz links.
- Der vierte Boxplot zeigt für Fälle mit der beobachteten Kategorie *Basic service* die vorhergesagte Pseudo-Wahrscheinlichkeit für die Kategorie *Total service*.

Da die Zielvariable mehr als zwei Kategorien enthält, sind die ersten vier Boxplots weder bezüglich der horizontalen Linie bei 0,5 noch auf irgendeine andere Weise symmetrisch. Daher kann die Interpretation dieses Plots für Ziele mit mehr als zwei Kategorien schwierig sein, da es unmöglich ist, aus der Betrachtung eines Teils der Fälle in einem Boxplot die entsprechende Lage dieser Fälle in einem anderen Poxplot zu bestimmen.

ROC-Kurve

Abbildung 5-11
ROC-Kurve



Eine ROC-Kurve bietet eine grafische Anzeige von **Sensitivität** gegenüber **Spezifität** für alle möglichen Klassifikationstrennwerte. Das hier dargestellte Diagramm enthält vier Kurven, eine für jede Kategorie der Zielvariablen.

Beachten Sie, dass dieses Diagramm auf der Kombination aus Trainings- und Teststichprobe beruht. Um ein ROC-Diagramm für die Holdout-Stichprobe zu erstellen, müssen Sie die Datei an der Partitionsvariablen aufteilen und die Prozedur "ROC-Kurve" für die vorhergesagten Pseudo-Wahrscheinlichkeiten ausführen.

Abbildung 5-12
Fläche unter der Kurve

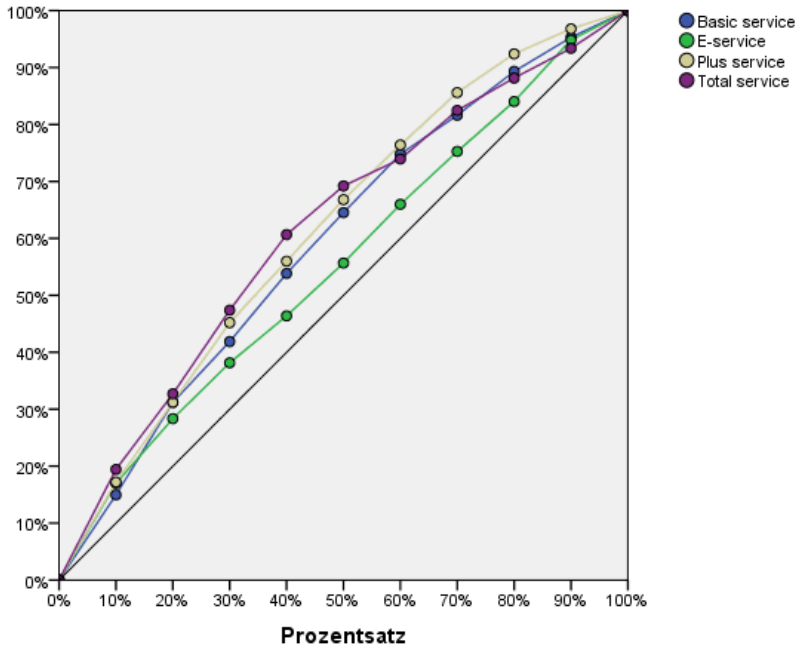
		Fläche
Customer category	Basic service	,635
	E-service	,573
	Plus service	,668
	Total service	,659

Die Fläche unter der Kurve ist eine numerische Zusammenfassung der ROC-Kurve und die Werte in der Tabelle stellen für jede Kategorie die Wahrscheinlichkeit dar, dass die vorhergesagte Wahrscheinlichkeit, in diese Kategorie zu gehören, für einen zufällig ausgewählten Fall in der betreffenden Kategorie größer ist als für einen zufällig ausgewählten Fall, der nicht in diese Kategorie eingeteilt wurde. So besteht beispielsweise bei einem zufällig ausgewählten Kunden in *Plus service* und einem zufällig ausgewählten Kunden in *Basic service*, *E-Service*

oder *Total service* eine Wahrscheinlichkeit von 0,668, dass die vom Modell vorhergesagte Pseudo-Wahrscheinlichkeit der Zahlungsunfähigkeit für den Kunden in *Plus service* höher ist.

Kumulatives Gewinndiagramm und Lift Chart

Abbildung 5-13
Kumulatives Gewinndiagramm

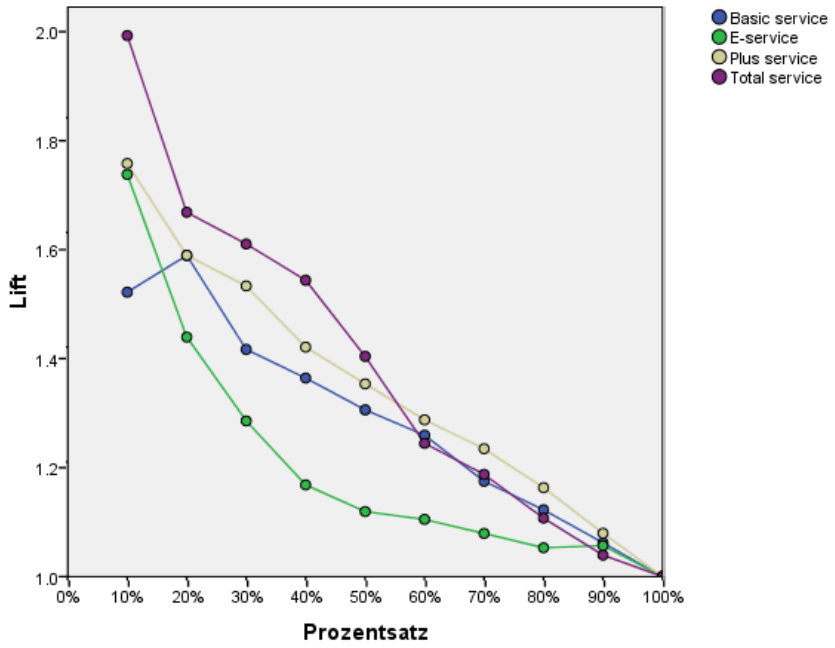


Abhängige Variable: Customer category

Das kumulative Gewinndiagramm zeigt den Prozentsatz der Fälle in einer bestimmten Kategorie, die "gewonnen" werden, indem ein bestimmter Prozentsatz der Gesamtzahl der Fälle anvisiert wird. Beispiel: Der erste Punkt auf der Kurve für die Kategorie *Total service* liegt ungefähr bei (10 %, 20 %). Dies bedeutet Folgendes: Wenn Sie ein Daten-Set mit dem Netzwerk scoren und alle Fälle nach der vorhergesagten Pseudo-Wahrscheinlichkeit von *Total service* sortieren, ist zu erwarten, dass die obersten 10 % ungefähr 20 % aller Fälle enthalten, die tatsächlich in die Kategorie *Total service* fallen. Ebenso enthalten die obersten 20 % ungefähr 30 % der zahlungsunfähigen Personen, die obersten 30 % der Fälle 50 % der zahlungsunfähigen Personen usw. Bei Auswahl von 100 % des gescorten Daten-Sets erfassen Sie alle zahlungsunfähigen Personen im Daten-Set.

Die diagonale Linie ist die "Basis"-Kurve. Wenn Sie nach dem Zufallsprinzip 10 % der Fälle aus dem gescorten Daten-Set auswählen, ist zu erwarten, dass Sie ungefähr 10 % der Fälle "gewinnen", die tatsächlich in eine bestimmte Kategorie fallen. Je höher über der Basis eine Kurve liegt, desto größer ist der Gewinn.

Abbildung 5-14
Lift Chart (Index)



Abhängige Variable: Customer category

Der Lift Chart wird aus dem kumulativen Gewinndiagramm abgeleitet; die Werte auf der *y*-Achse entsprechen dem Quotienten aus dem kumulativen Gewinn für jede Kurve und der Basis. Der Lift bei 10 % für die Kategorie *Total service* beträgt somit 20 % / 10 % = 2,0. Er bietet eine alternative Möglichkeit zur Analyse der Informationen im kumulativen Gewinndiagramm.

Anmerkung: Das kumulative Gewinndiagramm und der Lift Chart beruhen auf der Kombination aus Trainings- und Teststichprobe.

Empfohlene Literatur

In folgenden Texten finden Sie weitere Informationen zu “Radiale Basisfunktion”:

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*, 3rd (Hg.). Oxford: Oxford University Press.

Fine, T. L. 1999. *Feedforward Neural Network Methodology*, 3rd (Hg.). New York: Springer-Verlag.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*, 2nd (Hg.). 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 (Hg.). Los Alamitos, Kalifornien: IEEE Comput. Soc. Press, 401–405.

Uykan, Z., C. Guzelis, M. E. Celebi, als auch H. N. Koivo. 2000. Analysis of input-output clustering for determining centers of RBFN. *IEEE transactions on neural networks*, 11, 851–858.

Anhang

A

Beispieldateien

Die zusammen mit dem Produkt installierten Beispieldateien finden Sie im Unterverzeichnis *Samples* des Installationsverzeichnisses.

Beschreibungen

Im Folgenden finden Sie Kurzbeschreibungen der in den verschiedenen Beispielen in der Dokumentation verwendeten Beispieldateien:

- **accidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die alters- und geschlechtsabhängige Risikofaktoren für Autounfälle in einer bestimmten Region untersucht. Jeder Fall entspricht einer Kreuzklassifikation von Alterskategorie und Geschlecht.
- **adl.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die Vorteile einer vorgeschlagenen Therapieform für Schlaganfallpatienten zu ermitteln. Ärzte teilten weibliche Schlaganfallpatienten nach dem Zufallsprinzip jeweils einer von zwei Gruppen zu. Die erste Gruppe erhielt die physische Standardtherapie, die zweite erhielt eine zusätzliche Emotionaltherapie. Drei Monate nach den Behandlungen wurden die Fähigkeiten der einzelnen Patienten, übliche Alltagsaktivitäten auszuführen, als ordinale Variablen bewertet.
- **advert.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Einzelhändlers geht, die Beziehungen zwischen den in Werbung investierten Beträgen und den daraus resultierenden Umsätzen zu untersuchen. Zu diesem Zweck hat er die Umsätze vergangener Jahre und die zugehörigen Werbeausgaben zusammengestellt.
- **aflatoxin.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests von Maisernten auf Aflatoxin geht, ein Gift, dessen Konzentration stark zwischen und innerhalb von Ernteerträgen schwankt. Ein Kornverarbeitungsbetrieb hat aus 8 Ernteerträgen je 16 Proben erhalten und das Aflatoxinniveau in Teilen pro Milliarde (parts per billion, PPB) gemessen.
- **aflatoxin20.sav.** Diese Datendatei enthält die Aflatoxinmessungen aus jeder der 16 Stichproben aus den Erträgen 4 und 8 der Datendatei *aflatoxin.sav*.
- **anorectic.sav.** Bei der Ausarbeitung einer standardisierten Symptomatologie anorektischen/bulimischen Verhaltens führten Forscher (Van der Ham, Meulman, Van Strien, als auch Van Engeland, 1997) eine Studie mit 55 Jugendlichen mit bekannten Ess-Störungen durch. Jeder Patient wurde vier Mal über einen Zeitraum von vier Jahren untersucht, es fanden also insgesamt 220 Beobachtungen statt. Bei jeder Beobachtung erhielten die Patienten Scores für jedes von 16 Symptomen. Die Symptomwerte fehlen für Patient 71

zum Zeitpunkt 2, Patient 76 zum Zeitpunkt 2 und Patient 47 zum Zeitpunkt 3, wodurch 217 gültige Beobachtungen verbleiben.

- **autoaccidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Versicherungsanalysten geht, ein Modell zur Anzahl der Autounfälle pro Fahrer unter Berücksichtigung von Alter und Geschlecht zu erstellen. Jeder Fall stellt einen Fahrer dar und erfasst das Geschlecht des Fahrers, sein Alter in Jahren und die Anzahl der Autounfälle in den letzten fünf Jahren.
- **band.sav.** Diese Datendatei enthält die hypothetischen wöchentlichen Verkaufszahlen von CDs für eine Musikgruppe. Daten für drei mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **bankloan.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Bank geht, den Anteil der nicht zurückgezahlten Kredite zu reduzieren. Die Datei enthält Informationen zum Finanzstatus und demografischen Hintergrund von 850 früheren und potenziellen Kunden. Bei den ersten 700 Fällen handelt es sich um Kunden, denen bereits ein Kredit gewährt wurde. Bei den letzten 150 Fällen handelt es sich um potenzielle Kunden, deren Kreditrisiko die Bank als gering oder hoch einstufen möchte.
- **bankloan_binning.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen zum Finanzstatus und demografischen Hintergrund von 5.000 früheren Kunden enthält.
- **behavior.sav.** In einem klassischen Beispiel (Price als auch Bouffard, 1974) wurden 52 Schüler/Studenten gebeten, die Kombinationen aus 15 Situationen und 15 Verhaltensweisen auf einer 10-Punkte-Skala von 0 = "ausgesprochen angemessen" bis 9 = "ausgesprochen unangemessen" zu bewerten. Die Werte werden über die einzelnen Personen gemittelt und als Unähnlichkeiten verwendet.
- **behavior_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine zweidimensionale Lösung für *behavior.sav*.
- **brakes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik geht, die Scheibenbremsen für Hochleistungsautomobile herstellt. Die Datendatei enthält Messungen des Durchmessers von 16 Scheiben aus 8 Produktionsmaschinen. Der Zieldurchmesser für die Scheiben ist 322 Millimeter.
- **breakfast.sav.** In einer klassischen Studie (Green als auch Rao, 1972) wurden 21 MBA-Studenten der Wharton School mit ihren Lebensgefährten darum gebeten, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 = "am meisten bevorzugt" bis 15 = "am wenigsten bevorzugt" zu ordnen. Die Bevorzugungen wurden in sechs unterschiedlichen Szenarien erfasst, von "Overall preference" (Allgemein bevorzugt) bis "Snack, with beverage only" (Imbiss, nur mit Getränk).
- **breakfast-overall.sav.** Diese Datei enthält die Daten zu den bevorzugten Frühstücksartikeln, allerdings nur für das erste Szenario, "Overall preference" (Allgemein bevorzugt).
- **broadband_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die die Anzahl der Abonnenten eines Breitband-Service, nach Region geordnet, enthält. Die Datendatei enthält die monatlichen Abonnentenzahlen für 85 Regionen über einen Zeitraum von vier Jahren.
- **broadband_2.sav.** Diese Datendatei stimmt mit *broadband_1.sav* überein, enthält jedoch Daten für weitere drei Monate.

- **car_insurance_claims.sav.** Ein an anderer Stelle (McCullagh als auch Nelder, 1989) vorgestelltes und analysiertes Daten-Set bezieht sich auf Schadensansprüche für Autos. Die durchschnittliche Höhe der Schadensansprüche lässt sich mit Gamma-Verteilung modellieren. Dazu wird eine inverse Verknüpfungsfunktion verwendet, um den Mittelwert der abhängigen Variablen mit einer linearen Kombination aus Alter des Versicherungsnehmers, Fahrzeugtyp und Fahrzeugalter in Bezug zu setzen. Die Anzahl der eingereichten Schadensansprüche kann als Skalierungsgewicht verwendet werden.
- **car_sales.sav.** Diese Datendatei enthält hypothetische Verkaufsschätzer, Listenpreise und physische Spezifikationen für verschiedene Fahrzeugfabrikate und -modelle. Die Listenpreise und physischen Spezifikationen wurden von *edmunds.com* und Hersteller-Websites entnommen.
- **carpet.sav.** In einem beliebten Beispiel möchte (Green als auch Wind, 1973) einen neuen Teppichreiniger vermarkten und dazu den Einfluss von fünf Faktoren auf die Bevorzugung durch den Verbraucher untersuchen: Verpackungsgestaltung, Markenname, Preis, Gütesiegel, *Good Housekeeping* und Geld-zurück-Garantie. Die Verpackungsgestaltung liegt in drei Faktorstufen vor, die sich durch die Position der Auftragebürste unterscheiden. Außerdem gibt es drei Markennamen (*K2R*, *Glory* und *Bissell*), drei Preisstufen sowie je zwei Stufen (Nein oder Ja) für die letzten beiden Faktoren. 10 Kunden stufen 22 Profile ein, die durch diese Faktoren definiert sind. Die Variable *Preference* enthält den Rang der durchschnittlichen Einstufung für die verschiedenen Profile. Ein niedriger Rang bedeutet eine starke Bevorzugung. Diese Variable gibt ein Gesamtmaß der Bevorzugung für die Profile an.
- **carpet_prefs.sav.** Diese Datendatei beruht auf denselben Beispielen, wie für *carpet.sav* beschrieben, enthält jedoch die tatsächlichen Einstufungen durch jeden der 10 Kunden. Die Kunden wurden gebeten, die 22 Produktprofile in der Reihenfolge ihrer Präferenzen einzustufen. Die Variablen *PREF1* bis *PREF22* enthalten die IDs der zugeordneten Profile, wie in *carpet_plan.sav* definiert.
- **catalog.sav.** Diese Datendatei enthält hypothetische monatliche Verkaufszahlen für drei Produkte, die von einem Versandhaus verkauft werden. Daten für fünf mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **catalog_seasfac.sav.** Diese Datendatei ist mit *catalog.sav* identisch, außer, dass ein Set von saisonalen Faktoren, die mithilfe der Prozedur "Saisonale Zerlegung" berechnet wurden, sowie die zugehörigen Datumsvariablen hinzugefügt wurden.
- **cellular.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Mobiltelefonunternehmens geht, die Kundenabwanderung zu verringern. Scores für die Abwanderungsneigung (von 0 bis 100) werden auf die Kunden angewendet. Kunden mit einem Score von 50 oder höher streben vermutlich einen Anbieterwechsel an.
- **ceramics.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Herstellers geht, der ermitteln möchte, ob ein neue, hochwertige Keramiklegierung eine größere Hitzebeständigkeit aufweist als eine Standardlegierung. Jeder Fall entspricht einem Test einer der Legierungen; die Temperatur, bei der das Keramikwälzlager versagte, wurde erfasst.
- **cereal.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Umfrage geht, bei der 880 Personen nach ihren Frühstückgewohnheiten befragt wurden. Außerdem wurden Alter, Geschlecht, Familienstand und Vorliegen bzw. Nichtvorliegen eines aktiven Lebensstils (auf der Grundlage von mindestens zwei Trainingseinheiten pro Woche) erfasst. Jeder Fall entspricht einem Teilnehmer.

- **clothing_defects.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Bekleidungsfabrik geht. Aus jeder in der Fabrik produzierten Charge entnehmen die Kontrolleure eine Stichprobe an Bekleidungsartikeln und zählen die Anzahl der Bekleidungsartikel die inakzeptabel sind.
- **coffee.sav.** Diese Datendatei enthält Daten zum wahrgenommenen Image von sechs Eiskaffeemarken (Kennedy, Riquier, als auch Sharp, 1996). Bei den 23 Attributen des Eiskaffee-Image sollten die Teilnehmer jeweils alle Marken auswählen, die durch dieses Attribut beschrieben werden. Die sechs Marken werden als "AA", "BB", "CC", "DD", "EE" und "FF" bezeichnet, um Vertraulichkeit zu gewährleisten.
- **contacts.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Kontaktlisten einer Gruppe von Vertretern geht, die Computer an Unternehmen verkaufen. Die einzelnen Kontaktpersonen werden anhand der Abteilung, in der sie in ihrem Unternehmen arbeiten und anhand ihrer Stellung in der Unternehmenshierarchie in Kategorien eingeteilt. Außerdem werden der Betrag des letzten Verkaufs, die Zeit seit dem letzten Verkauf und die Größe des Unternehmens, in dem die Kontaktperson arbeitet, aufgezeichnet.
- **creditpromo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Kaufhauses geht, die Wirksamkeit einer kürzlich durchgeführten Kreditkarten-Werbeaktion einzuschätzen. Dazu wurden 500 Karteninhaber nach dem Zufallsprinzip ausgewählt. Die Hälfte erhielt eine Werbebeilage, die einen reduzierten Zinssatz für Einkäufe in den nächsten drei Monaten ankündigte. Die andere Hälfte erhielt eine Standard-Werbebeilage.
- **customer_dbase.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, das die Informationen in seinem Data Warehouse nutzen möchte, um spezielle Angebote für Kunden zu erstellen, die mit der größten Wahrscheinlichkeit darauf ansprechen. Nach dem Zufallsprinzip wurde eine Untergruppe des Kundenstamms ausgewählt. Diese Gruppe erhielt die speziellen Angebote und die Reaktionen wurden aufgezeichnet.
- **customers_model.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, auf die sich eine Marketingkampagne richtete. Zu diesen Daten gehören demografische Informationen, eine Übersicht über die bisherigen Einkäufe und die Angabe ob die einzelnen Personen auf die Kampagne ansprachen oder nicht. Jeder Fall entspricht einer Einzelperson.
- **customers_new.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, die potenzielle Kandidaten für Marketingkampagnen sind. Zu diesen Daten gehören demografische Informationen und eine Übersicht über die bisherigen Einkäufe für jede Person. Jeder Fall entspricht einer Einzelperson.
- **debate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die gepaarte Antworten auf eine Umfrage unter den Zuhörern einer politischen Debatte enthält (Antworten vor und nach der Debatte). Jeder Fall entspricht einem Befragten.
- **debate_aggregate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der die Antworten aus *debate.sav* aggregiert wurden. Jeder Fall entspricht einer Kreuzklassifikation der bevorzugten Politiker vor und nach der Debatte.
- **demo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Kundendatenbank geht, die zum Zwecke der Zusendung monatlicher Angebote erworben wurde. Neben verschiedenen demografischen Informationen ist erfasst, ob der Kunde auf das Angebot geantwortet hat.

- **demo_cs_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den ersten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einer anderen Stadt. Außerdem sind IDs für Region, Provinz, Landkreis und Stadt erfasst.
- **demo_cs_2.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den zweiten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einem anderen Stadtteil aus den im ersten Schritt ausgewählten Städten. Außerdem sind IDs für Region, Provinz, Landkreis, Stadt, Stadtteil und Wohneinheit erfasst. Die Informationen zur Stichprobenziehung aus den ersten beiden Stufen des Stichprobenplans sind ebenfalls enthalten.
- **demo_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfrageinformationen enthält die mit einem komplexen Stichprobenplan erfasst wurden. Jeder Fall entspricht einer anderen Wohneinheit. Es sind verschiedene Informationen zum demografischen Hintergrund und zur Stichprobenziehung erfasst.
- **dietstudy.sav.** Diese hypothetische Datendatei enthält die Ergebnisse einer Studie der "Stillman-Diät" (Rickman, Mitchell, Dingman, als auch Dalen, 1974). Jeder Fall entspricht einem Teilnehmer und enthält dessen Gewicht vor und nach der Diät in amerikanischen Pfund sowie mehrere Messungen des Triglyceridspiegels (in mg/100 ml).
- **dischargedata.sav.** Hierbei handelt es sich um eine Datendatei zum Thema *Seasonal Patterns of Winnipeg Hospital Use*, (Menec , Roos, Nowicki, MacWilliam, Finlayson , als auch Black, 1999) (Saisonale Muster der Belegung im Krankenhaus von Winnipeg) vom Manitoba Centre for Health Policy.
- **dvdplayer.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Entwicklung eines neuen DVD-Spielers geht. Mithilfe eines Prototyps hat das Marketing-Team Zielgruppendaten erfasst. Jeder Fall entspricht einem befragten Benutzer und enthält demografische Daten zu dem Benutzer sowie dessen Antworten auf Fragen zum Prototyp.
- **flying.sav.** Diese Datendatei enthält die Flugmeilen zwischen zehn Städten in den USA.
- **german_credit.sav.** Diese Daten sind aus dem Daten-Set "German credit" im Repository of Machine Learning Databases (Blake als auch Merz, 1998) an der Universität von Kalifornien in Irvine entnommen.
- **grocery_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *grocery_coupons.sav*, wobei die wöchentlichen Einkäufe zusammengefasst sind, sodass jeder Fall einem anderen Kunden entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und der verzeichnete ausgegebene Betrag ist nun die Summe der Beträge, die in den vier Wochen der Studie ausgegeben wurden.
- **grocery_coupons.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfragedaten enthält, die von einer Lebensmittelkette erfasst wurden, die sich für die Kaufgewohnheiten ihrer Kunden interessiert. Jeder Kunde wird über vier Wochen beobachtet, und jeder Fall entspricht einer Kundenwoche und enthält Informationen zu den Geschäften, in denen der Kunde einkauft sowie zu anderen Merkmalen, beispielsweise welcher Betrag in der betreffenden Woche für Lebensmittel ausgegeben wurde.
- **guttman.sav.** Bell (Bell, 1961) legte eine Tabelle zur Darstellung möglicher sozialer Gruppen vor. Guttman (Guttman, 1968) verwendete einen Teil dieser Tabelle, bei der fünf Variablen, die Aspekte beschreiben, wie soziale Interaktion, das Gefühl der Gruppenzugehörigkeit, die

physische Nähe der Mitglieder und die Formalität der Beziehung, mit sieben theoretischen sozialen Gruppen gekreuzt wurden: “crowds” (Menschenmassen, beispielsweise die Zuschauer eines Fußballspiels), “audience” (Zuhörerschaften, beispielsweise die Personen im Theater oder bei einer Vorlesung), “public” (Öffentlichkeit, beispielsweise Zeitungsleser oder Fernsehzuschauer), “mobs” (Mobs, wie Menschenmassen, jedoch mit wesentlich stärkerer Interaktion), “primary groups” (Primärgruppen, vertraulich), “secondary groups” (Sekundärgruppen, freiwillig) und “modern community” (die moderne Gesellschaft, ein lockerer Zusammenschluss, der aus einer engen physischen Nähe und dem Bedarf an spezialisierten Dienstleistungen entsteht).

- **healthplans.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Versicherungsgruppe geht, vier verschiedene Pläne zur Gesundheitsvorsorge für Kleinbetriebe zu evaluieren. Zwölf Inhaber von Kleinbetrieben (Arbeitgeber) wurden gebeten, die Pläne danach in eine Rangfolge zu bringen, wie gern sie sie ihren Mitarbeitern anbieten würden. Jeder Fall entspricht einem Arbeitgeber und enthält die Reaktionen auf die einzelnen Pläne.
- **health_funding.sav.** Hierbei handelt es sich um eine hypothetische Datei, die Daten zur Finanzierung des Gesundheitswesens (Betrag pro 100 Personen), Krankheitsraten (Rate pro 10.000 Personen der Bevölkerung) und Besuche bei medizinischen Einrichtungen/Ärzten (Rate pro 10.000 Personen der Bevölkerung) enthält. Jeder Fall entspricht einer anderen Stadt.
- **hivassay.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Bemühungen eines pharmazeutischen Labors, einen Schnelltest zur Erkennung von HIV-Infektionen zu entwickeln. Die Ergebnisse des Tests sind acht kräftiger werdende Rotschattierungen, wobei kräftigeren Schattierungen auf eine höhere Infektionswahrscheinlichkeit hindeuten. Bei 2.000 Blutproben, von denen die Hälfte mit HIV infiziert war, wurde ein Labortest durchgeführt.
- **hourlywagedata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zum Stundenlohn von Pflegepersonal in Praxen und Krankenhäusern mit unterschiedlich langer Berufserfahrung.
- **insure.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die die Risikofaktoren untersucht, die darauf hinweisen, ob ein Kunde die Leistungen einer mit einer Laufzeit von 10 Jahren abgeschlossenen Lebensversicherung in Anspruch nehmen wird. Jeder Fall in der Datendatei entspricht einem Paar von Verträgen, je einer mit Leistungsforderung und der andere ohne, wobei die beiden Versicherungsnehmer in Alter und Geschlecht übereinstimmen.
- **judges.sav.** Hierbei handelt es sich um eine hypothetische Datendatei mit den Wertungen von ausgebildeten Kampfrichtern (sowie eines Sportliebhabers) zu 300 Kunstturnleistungen. Jede Zeile stellt eine Leistung dar; die Kampfrichter bewerteten jeweils dieselben Leistungen.
- **kinship_dat.sav.** Rosenberg und Kim (Rosenberg als auch Kim, 1975) haben 15 Bezeichnungen für den Verwandtschaftsgrad untersucht (Tante, Bruder, Cousin, Tochter, Vater, Enkelin, Großvater, Großmutter, Enkel, Mutter, Neffe, Nichte, Schwester, Sohn, Onkel). Die beiden Analytiker baten vier Gruppen von College-Studenten (zwei weibliche und zwei männliche Gruppen), diese Bezeichnungen auf der Grundlage der Ähnlichkeiten zu sortieren. Zwei Gruppen (eine weibliche und eine männliche Gruppe) wurden gebeten, die Bezeichnungen zweimal zu sortieren; die zweite Sortierung sollte dabei nach einem anderen Kriterium erfolgen als die erste. So wurden insgesamt sechs “Quellen” erzielt. Jede Quelle entspricht einer Ähnlichkeitsmatrix mit 15×15 Elementen. Die Anzahl der Zellen ist dabei gleich der

Anzahl der Personen in einer Quelle minus der Anzahl der gemeinsamen Platzierungen der Objekte in dieser Quelle.

- **kinship_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine dreidimensionale Lösung für *kinship_dat.sav*.
- **kinship_var.sav.** Diese Datendatei enthält die unabhängigen Variablen *gender* (Geschlecht), *gener* (Generation) und *degree* (Verwandtschaftsgrad), die zur Interpretation der Dimensionen einer Lösung für *kinship_dat.sav* verwendet werden können. Insbesondere können sie verwendet werden, um den Lösungsraum auf eine lineare Kombination dieser Variablen zu beschränken.
- **mailresponse.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines Bekleidungsherstellers geht, der ermitteln möchte, ob die Verwendung von Briefsendungen für das Direktmarketing zu schnelleren Antworten führt als Postwurfsendungen. Die Mitarbeiter in der Bestellannahme erfassen, wie vielen Wochen nach der Postsendung die einzelnen Bestellungen aufgegeben wurden.
- **marketvalues.sav.** Diese Datendatei betrifft Hausverkäufe in einem Neubaugebiet in Algonquin, Illinois, in den Jahren 1999–2000. Diese Verkäufe sind in Grundbucheinträgen dokumentiert.
- **mutualfund.sav.** Diese Datendatei betrifft Aktienmarktdaten für verschiedene Technologieaktien, die in im Index S&P 500 verzeichnet sind. Jeder Fall entspricht einem Unternehmen.
- **nhis2000_subset.sav.** Die "National Health Interview Survey (NHIS)" ist eine große, bevölkerungsbezogene Umfrage in unter der US-amerikanischen Zivilbevölkerung. Es werden persönliche Interviews in einer landesweit repräsentativen Stichprobe von Haushalten durchgeführt. Für die Mitglieder jedes Haushalts werden demografische Informationen und Beobachtungen zum Gesundheitsverhalten und Gesundheitsstatus eingeholt. Diese Datendatei enthält eine Teilmenge der Informationen aus der Umfrage des Jahres 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Datendatei und Dokumentation öffentlich zugänglich. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Zugriff erfolgte 2003.
- **ozone.sav.** Die Daten enthalten 330 Beobachtungen zu sechs meteorologischen Variablen zur Vorhersage der Ozonkonzentration aus den übrigen Variablen. Bei früheren Untersuchungen (Breiman als auch Friedman, 1985), (Hastie als auch Tibshirani, 1990) fanden Wissenschaftler einige Nichtlinearitäten unter diesen Variablen, die die Standardverfahren bei der Regression behindern.
- **pain_medication.sav.** Diese hypothetische Datendatei enthält die Ergebnisse eines klinischen Tests für ein entzündungshemmendes Medikament zur Schmerzbehandlung bei chronischer Arthritis. Von besonderem Interesse ist die Zeitdauer, bis die Wirkung des Medikaments einsetzt und wie es im Vergleich mit bestehenden Medikamenten abschneidet.
- **patient_los.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen zu Patienten, die wegen des Verdachts auf Herzinfarkt in das Krankenhaus eingeliefert wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.

- **patlos_sample.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen für eine Stichprobe von Patienten, denen während der Behandlung eines Herzinfarkts Thrombolytika verabreicht wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **polishing.sav.** Hierbei handelt es sich um die Datendatei "Nambeware Polishing Times" aus der Data and Story Library. Sie bezieht sich auf die Bemühungen eines Herstellers von Metallgeschirr (Nambe Mills, Santa Fe, New Mexico) zur zeitlichen Planung seiner Produktion. Jeder Fall entspricht einem anderen Artikel in der Produktpalette. Für jeden Artikel sind Durchmesser, Polierzeit, Preis und Produkttyp erfasst.
- **poll_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die öffentliche Unterstützung für einen Gesetzentwurf zu ermitteln, bevor er im Parlament eingebracht wird. Die Fälle entsprechen registrierten Wählern. Für jeden Fall sind County, Gemeinde und Wohnviertel des Wählers erfasst.
- **poll_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *poll_cs.sav* aufgeführten Wähler. Die Stichprobe wurde gemäß dem in der Plandatei *poll.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Beachten Sie jedoch Folgendes: Da im Stichprobenplan die PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*poll_jointprob.sav*). Die zusätzlichen Variablen zum demografischen Hintergrund der Wähler und ihrer Meinung zum vorgeschlagenen Gesetzentwurf wurden nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **property_assess.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen Bezirk (County) zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien, die im vergangenen Jahr in dem betreffenden County verkauft wurden. Jeder Fall in der Datendatei enthält die Gemeinde, in der sich die Immobilie befindet, den Bewerter, der die Immobilie besichtigt hat, die seit dieser Bewertung verstrichene Zeit, den zu diesem Zeitpunkt ermittelten Wert sowie den Verkaufswert der Immobilie.
- **property_assess_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen US-Bundesstaat zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien in dem betreffenden Bundesstaat. Jeder Fall in der Datendatei enthält das County, die Gemeinde und das Wohnviertel, in dem sich die Immobilie befindet, die seit der letzten Bewertung verstrichene Zeit sowie zu diesem Zeitpunkt ermittelten Wert.
- **property_assess_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *property_assess_cs.sav* aufgeführten Immobilien. Die Stichprobe wurde gemäß dem in der Plandatei *property_assess.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Die zusätzliche Variable *Current value* (Aktueller Wert) wurde nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.

- **recidivism.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem frühren Straftäter und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Zeit bis zu seiner zweiten Festnahme, sofern diese innerhalb von zwei Jahren nach der ersten Festnahme erfolgte.
- **recidivism_cs_sample.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter, der im Juni 2003 erstmals aus der Haft entlassen wurde, und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Daten zu seiner zweiten Festnahme, sofern diese bis Ende Juni 2006 erfolgte. Die Straftäter wurden aus per Stichprobenziehung ermittelten Polizeidirektionen ausgewählt (gemäß dem in *recidivism_cs.csplan* angegebenen Stichprobenplan). Da hierbei eine PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*recidivism_cs_jointprob.sav*).
- **salesperformance.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung von zwei neuen Verkaufsschulungen geht. 60 Mitarbeiter, die in drei Gruppen unterteilt sind, erhalten jeweils eine Standardschulung. Zusätzlich erhält Gruppe 2 eine technische Schulung und Gruppe 3 eine Praxisschulung. Die einzelnen Mitarbeiter wurden am Ende der Schulung einem Test unterzogen und die erzielten Punkte wurden erfasst. Jeder Fall in der Datendatei stellt einen Lehrgangsteilnehmer dar und enthält die Gruppe, der der Lehrgangsteilnehmer zugeteilt wurde sowie die von ihm in der Prüfung erreichte Punktzahl.
- **satisf.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Zufriedenheitsumfrage, die von einem Einzelhandelsunternehmen in 4 Filialen durchgeführt wurde. Insgesamt wurden 582 Kunden befragt. Jeder Fall gibt die Antworten eines einzelnen Kunden wieder.
- **screws.sav.** Diese Datendatei enthält Informationen zu den Eigenschaften von Schrauben, Bolzen, Muttern und Reißnägeln (Hartigan, 1975).
- **shampoo_ph.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik für Haarpflegeprodukte geht. In regelmäßigen Zeitabständen werden Messwerte von sechs separaten Ausgangschargen erhoben und ihr pH-Wert erfasst. Der Zielbereich ist 4,5–5,5.
- **ships.sav.** Ein an anderer Stelle (McCullagh et al., 1989) vorgestelltes und analysiertes Daten-Set bezieht sich auf die durch Wellen verursachten Schäden an Frachtschiffen. Die Vorfallshäufigkeiten können unter Angabe von Schiffstyp, Konstruktionszeitraum und Betriebszeitraum gemäß einer Poisson-Rate modelliert werden. Das Aggregat der Betriebsmonate für jede Zelle der durch die Kreuzklassifizierung der Faktoren gebildeten Tabelle gibt die Werte für die Risikoanfälligkeit an.
- **site.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, neue Standorte für die betriebliche Expansion auszuwählen. Das Unternehmen beauftragte zwei Berater unabhängig voneinander mit der Bewertung der Standorte. Neben einem umfassenden Bericht gaben die Berater auch eine zusammenfassende Wertung für jeden Standort als “good” (gut) “fair” (mittelmäßig) oder “poor” (schlecht) ab.

- **siteratings.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Betatests der neuen Website eines E-Commerce-Unternehmens geht. Jeder Fall entspricht einem Beta-Tester, der die Brauchbarkeit der Website auf einer Skala von 0 bis 20 bewertete.
- **smokers.sav.** Diese Datendatei wurde aus der Umfrage "National Household Survey of Drug Abuse" aus dem Jahr 1998 abstrahiert und stellt eine Wahrscheinlichkeitsstichprobe US-amerikanischer Haushalte dar. Daher sollte der erste Schritt bei der Analyse dieser Datendatei darin bestehen, die Daten entsprechend den Bevölkerungstrends zu gewichten.
- **smoking.sav.** Hierbei handelt es sich um eine von Greenacre (Greenacre , 1984) vorgestellte hypothetische Tabelle. Die relevante Tabelle wird durch eine Kreuztabelle der Rauchgewohnheiten und der Berufskategorie gebildet. Die Variable *Berufsgruppe* enthält die Berufskategorien *Senior Manager*, *Junior Manager*, *Angestellter mit Erfahrung*, *Angestellter ohne Erfahrung* und *Sekretariat* sowie die Kategorie *National Average*, die als Ergänzung der Analyse dienen kann. Die Variable *Rauchen* enthält die Rauchgewohnheiten *Nichtraucher*, *Leicht*, *Mittel* und *Stark* sowie die Kategorien *No Alcohol* und *Alcohol*, die als Ergänzung der Analyse dienen können.
- **storebrand.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Verkaufsleiterin in einem Lebensmittelmarkt geht, die die Verkaufszahlen des Waschmittels der Eigenmarke gegenüber den anderen Marken steigern möchte. Sie erarbeitet eine Werbeaktion im Geschäft und spricht an der Kasse mit Kunden. Jeder Fall entspricht einem Kunden.
- **stores.sav.** Diese Datendatei enthält hypothetische monatliche Marktanteilsdaten für zwei konkurrierende Lebensmittelgeschäfte. Jeder Fall entspricht den Marktanteilsdaten für einen bestimmten Monat.
- **stroke_clean.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozeduren in der Option "Data Preparation" bereinigt wurde.
- **stroke_invalid.sav.** Diese hypothetische Datendatei enthält den ursprünglichen Zustand einer medizinischen Datenbank, der mehrere Dateneingabefehler aufweist.
- **stroke_survival.** In dieser hypothetischen Datendatei geht es um die Überlebenszeiten von Patienten, die nach einem Rehabilitationsprogramm wegen eines ischämischen Schlaganfalls mit einer Reihe von Problemen zu kämpfen haben. Nach dem Schlaganfall werden das Auftreten von Herzinfarkt, ischämischem Schlaganfall und hämorrhagischem Schlaganfall sowie der Zeitpunkt des Ereignisses aufgezeichnet. Die Stichprobe ist auf der linken Seite abgeschnitten, da sie nur Patienten enthält, die bis zum Ende des Rehabilitationprogramms, das nach dem Schlaganfall durchgeführt wurde, überlebten.
- **stroke_valid.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozedur "Daten validieren" überprüft wurde. Sie enthält immer noch potenziell anomale Fälle.
- **tastetest.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung der Auswirkungen der Mulchfarbe auf den Geschmack von Pflanzenprodukten geht. Der Geschmack von Erdbeeren, die in rotem, blauem und schwarzem Rindenmulch gezogen wurden, wurde von Testpersonen auf einer ordinalen Skala (weit unter bis weit über dem Durchschnitt) bewertet. Jeder Fall entspricht einem Geschmackstester.

- **telco.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Telekommunikationsunternehmens geht, die Kundenabwanderung zu verringern. Jeder Fall entspricht einem Kunden und enthält verschiedene Informationen zum demografischen Hintergrund und zur Servicenutzung.
- **telco_extra.sav.** Diese Datendatei ähnelt der Datei *telco.sav*, allerdings wurden die Variablen "tenure" und die Log-transformierten Variablen zu den Kundenausgaben entfernt und durch standardisierte Log-transformierte Variablen ersetzt.
- **telco_missing.sav.** Diese Datendatei entspricht der Datei *telco_mva_complete.sav*, allerdings wurde ein Teil der Daten durch fehlende Werte ersetzt.
- **telco_mva_complete.sav.** Bei dieser Datendatei handelt es sich um eine Teilmenge der Datendatei *telco.sav*, allerdings mit anderen Variablennamen.
- **testmarket.sav.** Diese hypothetische Datendatei bezieht sich auf die Pläne einer Fast-Food-Kette, einen neuen Artikel in ihr Menü aufzunehmen. Es gibt drei mögliche Kampagnen zur Verkaufsförderung für das neue Produkt. Daher wird der neue Artikel in Filialen in mehreren zufällig ausgewählten Märkten eingeführt. An jedem Standort wird eine andere Form der Verkaufsförderung verwendet und die wöchentlichen Verkaufszahlen für das neue Produkt werden für die ersten vier Wochen aufgezeichnet. Jeder Fall entspricht einer Standort-Woche.
- **testmarket_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *testmarket.sav*, wobei die wöchentlichen Verkaufszahlen zusammengefasst sind, sodass jeder Fall einem Standort entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und die verzeichneten Verkaufszahlen sind nun die Summe der Verkaufszahlen während der vier Wochen der Studie.
- **tree_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_credit.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält.
- **tree_missing_data.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält und eine große Anzahl fehlender Werte aufweist.
- **tree_score_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_textdata.sav.** Eine einfache Datendatei mit nur zwei Variablen, die vor allem den Standardzustand von Variablen vor der Zuweisung von Messniveau und Wertelabels zeigen soll.
- **tv-survey.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Studie, die von einem Fernsehstudio durchgeführt wurde, das überlegt, ob die Laufzeit eines erfolgreichen Programms verlängert werden soll. 906 Personen wurden gefragt, ob sie das Programm unter verschiedenen Bedingungen ansehen würden. Jede Zeile entspricht einem Befragten; jede Spalte entspricht einer Bedingung.
- **ulcer_recurrence.sav.** Diese Datei enthält Teilinformationen aus einer Studie zum Vergleich der Wirksamkeit zweier Therapien zur Vermeidung des Wiederauftretens von Geschwüren. Es stellt ein gutes Beispiel für intervallzensierte Daten dar und wurde an anderer Stelle (Collett, 2003) vorgestellt und analysiert.

- **ulcer_recurrence_recoded.sav.** In dieser Datei sind die Daten aus *ulcer_recurrence.sav* so umstrukturiert, dass das Modell der Ereigniswahrscheinlichkeit für jedes Intervall der Studie berechnet werden kann und nicht nur die Ereigniswahrscheinlichkeit am Ende der Studie. Sie wurde an anderer Stelle (Collett et al., 2003) vorgestellt und analysiert.
- **verd1985.sav.** Diese Datendatei enthält eine Umfrage (Verdegaal, 1985). Die Antworten von 15 Subjekten auf 8 Variablen wurden aufgezeichnet. Die relevanten Variablen sind in drei Sets unterteilt. Set 1 umfasst *alter* und *heirat*, Set 2 besteht aus *pet* und *news* und in Set 3 finden sich *music* und *live*. Die Variable *pet* wird mehrfach nominal skaliert und die Variable *Alter* ordinal. Alle anderen Variablen werden einzeln nominal skaliert.
- **virus.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Internet-Dienstanbieters geht, der die Auswirkungen eines Virus auf seine Netzwerke ermitteln möchte. Dabei wurde vom Moment der Virusentdeckung bis zu dem Zeitpunkt, zu dem die Virusinfektion unter Kontrolle war, der (ungefähre) prozentuale Anteil infizierter E-Mail in den Netzwerken erfasst.
- **waittimes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Wartezeiten für Kunden bei drei verschiedenen Filialen einer Bank. Jeder Fall entspricht einem Kunden und zeichnet die Wartezeit und die Filiale.
- **webusability.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests zur Benutzerfreundlichkeit eines neuen Internetgeschäfts geht. Jeder Fall entspricht einer von fünf Testpersonen, die die Benutzerfreundlichkeit bewerten und gibt für sechs separate Aufgaben an, ob die Testperson sie erfolgreich ausführen könnte.
- **wheeze_steubenville.sav.** Hierbei handelt es sich um eine Teilmenge der Daten aus einer Langzeitstudie zu den gesundheitlichen Auswirkungen der Luftverschmutzung auf Kinder (Ware, Dockery, Spiro III, Speizer, als auch Ferris Jr., 1984). Die Daten enthalten wiederholte binäre Messungen des Keuchens von Kindern aus Steubenville, Ohio, im Alter von 7, 8, 9 und 10 Jahren sowie eine unveränderlichen Angabe, ob die Mutter im ersten Jahr der Studie rauchte oder nicht.
- **workprog.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einem Arbeitsprogramm der Regierung, das versucht, benachteiligten Personen bessere Arbeitsplätze zu verschaffen. Eine Stichprobe potenzieller Programmteilnehmer wurde beobachtet. Von diesen Personen wurden nach dem Zufallsprinzip einige für die Teilnahme an dem Programm ausgewählt. Jeder Fall entspricht einem Programmteilnehmer.

Bibliografie

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 (Hg.). Oxford: Oxford University Press.

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

Breiman, L., als auch 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 (Hg.). Boca Raton: Chapman & Hall/CRC.

Fine, T. L. 1999. *Feedforward Neural Network Methodology*, 3rd (Hg.). New York: Springer-Verlag.

Green, P. E., als auch V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

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

Greenacre , M. J. 1984. *Theory and applications of correspondence analysis*. London: Academic 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., als auch R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*, 2nd (Hg.). New York: Macmillan College Publishing.

Kennedy, R., C. Riquier, als auch 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., als auch J. A. Nelder. 1989. *Generalized Linear Models*, 2nd (Hg.). London: Chapman & Hall.

Menec , V., N. Roos, D. Nowicki, L. MacWilliam, G. Finlayson , als auch C. Black. 1999. *Seasonal Patterns of Winnipeg Hospital Use*. : Manitoba Centre for Health Policy.

Price, R. H., als auch 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, als auch 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., als auch 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 (Hg.). Los Alamitos, Kalifornien: IEEE Comput. Soc. Press, 401–405.

Uykan, Z., C. Guzelis, M. E. Celebi, als auch 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, als auch 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 niederländischer Sprache)*. Leiden: Department of Data Theory, Universität Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, als auch 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.

Index