IBM SPSS Neural Networks 20

Nota: Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni generali disponibili in Note legali a pag. 99.

Questa versione si applica a IBM® SPSS® Statistics 20 e a tutte le successive versioni e modifiche fino a eventuali disposizioni contrarie indicate in nuove versioni.

Le schermate dei prodotti Adobe sono state ristampate su autorizzazione di Adobe Systems Incorporated.

Le schermate dei prodotti Microsoft sono state ristampate su autorizzazione di Microsoft Corporation.

Prefazione

IBM® SPSS® Statistics è un sistema completo per l'analisi dei dati. Il modulo aggiuntivo opzionale Neural Networks include le tecniche di analisi aggiuntive descritte nel presente manuale. Il modulo aggiuntivo Neural Networks deve essere usato con il modulo Core SPSS Statistics in cui è completamente integrato.

Informazioni su Business Analytics di IBM

Il software IBM Business Analytics fornisce informazioni complete, coerenti e accurate a cui i responsabili delle decisioni possono affidarsi per ottimizzare le prestazioni dell'azienda. Un ampio portafoglio di applicazioni di business intelligence, analisi predittiva, gestione delle prestazioni e delle strategie finanziarie e analisi offre una panoramica chiara, istantanea e interattiva delle prestazioni attuali e la possibilità di prevedere i risultati futuri. Utilizzato in combinazione con potenti soluzioni di settore, prassi consolidate e servizi professionali, questo software consente alle aziende di tutte le dimensioni di ottimizzare la produttività, automatizzare le decisioni senza problemi e fornire risultati migliori.

Come parte di questo portafoglio, il software IBM SPSS Predictive Analytics consente alle aziende di prevedere gli eventi futuri e di agire tempestivamente in modo da migliorare i risultati delle attività aziendali. Le aziende, gli enti governativi e le università di tutto il mondo si affidano alla tecnologia IBM SPSS perché rappresenta un vantaggio concorrenziale in termini di attrazione, retention e aumento dei clienti, riducendo al tempo stesso le frodi e limitando i rischi. Incorporando il software IBM SPSS nelle attività quotidiane, le aziende diventano imprese in grado di effettuare previsioni e di gestire e automatizzare le decisioni, per raggiungere gli obiettivi aziendali e vantaggi tangibili sulla concorrenza. Per ulteriori informazioni o per contattare un rappresentante, visitare il sito *http://www.ibm.com/spss*.

Supporto tecnico

Ai clienti che richiedono la manutenzione, viene messo a disposizione un servizio di supporto tecnico. I clienti possono contattare il supporto tecnico per richiedere assistenza per l'utilizzo dei prodotti IBM Corp. o per l'installazione di uno degli ambienti hardware supportati. Per contattare il supporto tecnico, visitare il sito Web di IBM Corp. all'indirizzo *http://www.ibm.com/support*. Ricordare che durante la richiesta di assistenza sarà necessario fornire i dati di identificazione personali, i dati relativi alla propria società e il numero del contratto di manutenzione.

Supporto tecnico per studenti

Gli studenti che utilizzano una versione accademica o grad pack di qualsiasi prodotto software IBM SPSS sono pregati di utilizzare le apposite pagine online per studenti Solutions for Education (*http://www.ibm.com/spss/rd/students/*). Gli studenti che utilizzano una copia del software IBM SPSS fornita dall'università, sono pregati di contattare il coordinatore del prodotto IBM SPSS presso l'università.

Servizio clienti

Per eventuali chiarimenti in merito alla spedizione o al proprio conto, rivolgersi alla sede locale. Tenere presente che sarà necessario fornire il numero di serie.

Corsi di formazione

IBM Corp. organizza corsi di formazione pubblici e onsite che includono esercitazioni pratiche. Tali corsi si terranno periodicamente nelle principali città. Per ulteriori informazioni su questi seminari, andare a *http://www.ibm.com/software/analytics/spss/training*.

Pubblicazioni aggiuntive

I documenti *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion* e *SPSS Statistics: Advanced Statistical Procedures Companion*, scritti da Marija Norušis e pubblicati da Prentice Hall sono disponibili come materiale supplementare consigliato. Queste pubblicazioni descrivono le procedure statistiche nei moduli SPSS Statistics Base, Advanced Statistics e Regression. Utili sia come guida iniziale all'analisi dei dati che per applicazioni avanzate, questi manuali consentono di ottimizzare l'utilizzo delle funzionalità presenti nell'offerta IBM® SPSS® Statistics. Per ulteriori informazioni, inclusi contenuti delle pubblicazioni e capitoli di esempio, visitare il sito Web dell'autrice: *http://www.norusis.com*

Contenuto

Appendici

Parte I:
Manuale dell'utente

Capitolo
1

Introduzione a Neural Networks

Le reti neurali rappresentano lo strumento preferito per molte applicazioni predittive di data mining, grazie alla loro potenza, flessibilità e facilità di utilizzo. Le reti neurali predittive sono particolarmente utili nelle applicazioni in cui il processo sottostante è complesso, ad esempio:

- Previsione della domanda del cliente per ottimizzare i costi di produzione e di consegna.
- Previsione della probabilità di risposta a una campagna di marketing tramite posta diretta per determinare a quali abitazioni di una lista di distribuzione inviare un'offerta.
- Assegnazione dei punteggi a un richiedente per determinare il rischio nel concedergli credito.
- Individuazione delle transazioni fraudolente in un database di richieste di risarcimento.

Le reti neurali utilizzate nelle applicazioni predittive quali le reti **Perceptron a più strati (MLP)** e **Funzione a base radiale (RBF)** sono supervisionate, ovvero i risultati previsti per il modello possono essere confrontati con i valori noti delle variabili di destinazione. L'opzione Neural Networks consente di adattare le reti MLP e RBF e salvare i modelli risultanti per l'assegnazione del punteggio.

Definizione di rete neurale

Il termine **rete neurale** si riferisce a una famiglia di modelli non strettamente correlata, caratterizzata da un grande spazio per i parametri e da una struttura flessibile, derivante da studi sul funzionamento del cervello. Con il crescere della famiglia, la maggior parte dei nuovi modelli è stata progettata per applicazioni non biologiche, sebbene molta della terminologia associata rifletta la sua origine.

Le definizioni specifiche delle reti neurali sono tanto svariate quanto i campi in cui vengono utilizzate. Sebbene nessuna singola definizione ricopra l'intera famiglia di modelli, si prenda in considerazione, per il momento, la seguente descrizione(Haykin, 1998):

Una rete neurale è un processore distribuito e parallelo che ha una naturale propensione a memorizzare la conoscenza sperimentale e a renderla disponibile per l'uso. Assomiglia al cervello sotto due aspetti:

- La conoscenza viene acquisita dalla rete tramite un processo di apprendimento.
- Le forze di connessione interneurale conosciute come pesi sinattici vengono utilizzate per memorizzare la conoscenza.

Vedere (Ripley, 1996) per informazioni dettagliate su questa definizione che risulta forse essere troppo restrittiva.

Al fine di differenziare le reti neurali dai tradizionali metodi statistici utilizzando questa definizione, ciò che *non* viene detto è tanto significativo quanto il testo vero e proprio della definizione. Ad esempio, il modello di regressione lineare tradizionale può acquisire la conoscenza tramite il metodo dei minimi quadrati e memorizzare tale conoscenza nei coefficienti di regressione. In tal senso, si tratta di una rete neurale. Infatti, è possibile sostenere che la

regressione lineare è un caso speciale di determinate reti neurali. Tuttavia, la regressione lineare comporta una rigida struttura dei modelli e un rigido insieme di ipotesi che vengono imposti prima di apprendere dai dati.

In contrasto, la definizione citata precedentemente comporta richieste minime sulla struttura dei modelli e le ipotesi. Pertanto, una rete neurale può avvicinarsi a un'ampia gamma di modelli statistici senza richiedere di ipotizzare preventivamente determinate relazioni tra le variabili dipendenti e indipendenti. Invece, la forma delle relazioni viene stabilita durante il processo di apprendimento. Se una relazione lineare tra le variabili dipendenti e indipendenti è appropriata, i risultati della rete neurale dovrebbero avvicinarsi strettamente a quelli del modello di regressione lineare. Se una relazione non lineare è più appropriata, la rete neurale si avvicinerà automaticamente alla struttura di modelli "corretta".

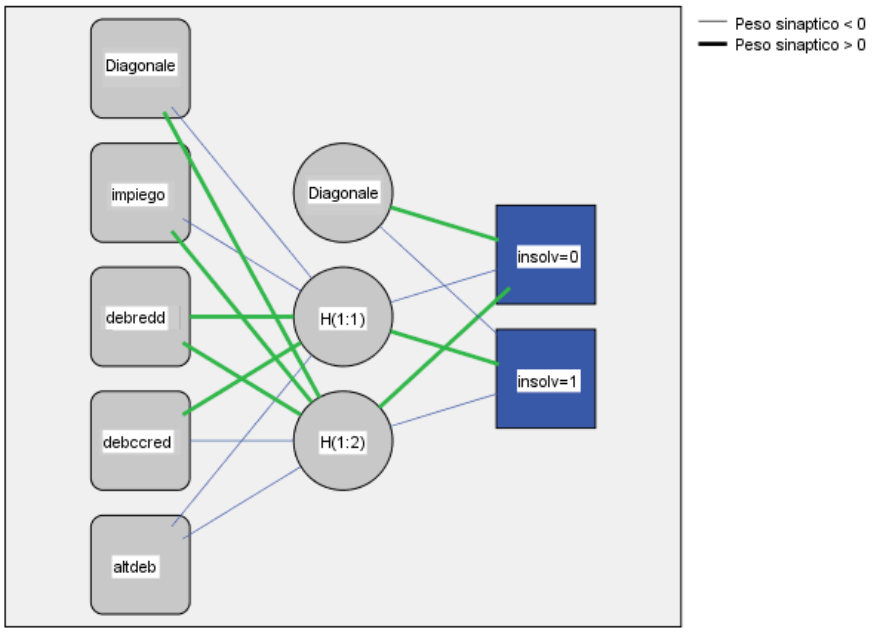
Il compromesso di questa flessibilità è che i pesi sinattici di una rete neurale non sono facilmente interpretabili. Pertanto, qualora si tenti di spiegare un processo sottostante che produca le relazioni tra le variabili dipendenti e indipendenti, sarebbe meglio utilizzare un modello statistico più tradizionale. Tuttavia, se l'interpretabilità del modello non è importante, è possibile ottenere spesso più rapidamente buoni risultati per i modelli tramite una rete neurale.

Struttura della rete neurale

Sebbene le reti neurali applichino richieste minime sulla struttura dei modelli e le ipotesi, è utile per comprendere l'**architettura di rete** generale. La rete Perceptron a più strati (MLP) o Funzione a base radiale (RBF) è una funzione di predittori (noti anche come input o variabili indipendenti) che riducono l'errore di previsione delle variabili di destinazione (note anche come output).

Considerare l'insieme di dati *bankloan.sav*, fornito con il prodotto, nel quale si possono identificare possibili inadempienti tra un gruppo di richiedenti di un prestito. Una rete MLP o RBF applicata a questo problema è una funzione delle misure che riducono l'errore nell'impostazione predefinita di previsione. La figura seguente è utile per illustrare il formato di questa funzione.

Figura 1-1
Architettura feedforward con uno strato nascosto



Questa struttura è conosciuta come **architettura feedforward** poiché le connessioni all'interno della rete passano dallo strato di input a quello di output senza alcun ciclo di feedback. In questa figura:

- Lo **strato di input** contiene i predittori.
- Lo **strato nascosto** contiene unità o nodi non osservabili. Il valore di ogni unità nascosta è una funzione dei predittori; il formato esatto della funzione dipende in parte dal tipo di rete e in parte dalle specifiche controllabili dall'utente.
- Lo **strato di output** contiene le risposte. Poiché la cronologia di inadempienza è una variabile categoriale con due categorie, viene registrata come due variabili indicatore. Ogni unità di output è una funzione di unità nascoste. Anche in questo caso il formato esatto della funzione dipende in parte dal tipo di rete e in parte dalle specifiche controllabili dall'utente.

La rete MLP consente un secondo strato nascosto; in tal caso, ogni unità nel secondo strato nascosto è una funzione delle unità nel primo strato nascosto e ogni risposta è una funzione delle unità nel secondo strato nascosto.

Capitolo
2

Perceptron a più strati

La procedura Perceptron a più strati (MLP) produce un modello predittivo per una o più variabili dipendenti (di destinazione) basato sui valori delle variabili predittori.

Esempi. Di seguito vengono illustrati due scenari di utilizzo della procedura MLP.

Un funzionario mutui presso una banca deve essere in grado di identificare le caratteristiche indicative delle persone che tendenzialmente saranno inadempimenti per quanto riguarda il rimborso dei prestiti e utilizzare tali caratteristiche per identificare i rischi di credito positivi e negativi. In base a un campione di clienti precedenti, il funzionario può creare una rete Perceptron a più strati, convalidare l'analisi utilizzando un campione di controllo composto da clienti precedenti, quindi utilizzare la rete per classificare i potenziali clienti come rischi di credito positivi o negativi.

Un ospedale è interessato a registrare i costi e la durata delle degenze per i pazienti ricoverati per infarto del miocardio. Stime accurate relative a tali misure consentono all'amministrazione di gestire correttamente il numero di posti letto disponibili per la degenza dei pazienti. Sulla base dell'archivio sanitario di un campione di pazienti curati per infarto, l'amministratore può creare una rete di previsione sia dei costi sia della durata della degenza.

Variabili dipendenti. Le variabili dipendenti possono essere:

- **Nominale.** Una variabile può essere considerata nominale quando i relativi valori rappresentano categorie prive di ordinamento intrinseco, per esempio l'ufficio di una società, Tra gli esempi di variabili nominali troviamo la regione, il codice postale e la religione.
- **Ordinale.** Una variabile può essere considerata ordinale quando i relativi valori rappresentano categorie con qualche ordinamento intrinseco, per esempio i gradi di soddisfazione per un servizio, da molto insoddisfatto a molto soddisfatto, i punteggi di atteggiamento corrispondenti a gradi di soddisfazione o fiducia e i punteggi di preferenza.
- **Scala.** Una variabile può essere considerata di scala (continua) quando i relativi valori rappresentano categorie ordinate con una metrica significativa, tale che i confronti fra le distanze dei relativi valori siano appropriati. Esempi di variabili di scala sono l'età espressa in anni o il reddito espresso in migliaia di Euro.

  La procedura presume che il livello di misurazione sia stato assegnato a tutte le variabili dipendenti, tuttavia è possibile modificare temporaneamente il livello di misurazione di una variabile facendo clic con il pulsante destro del mouse sulla variabile nell'elenco delle variabili sorgente e scegliere un livello di misurazione dal menu di scelta rapida.

L'icona accanto a ciascuna variabile nell'elenco delle variabili identifica il livello di misurazione e il tipo di dati.

	Numerico	Stringa	Data	Ora
Scala (continuo)		n/d		
Ordinale				
Nominale				

Variabili indipendenti. Possono essere specificate come fattori (categoriali) o covariate (scala).

Codifiche variabili categoriali. La procedura ricodifica temporaneamente le variabili dipendenti e indipendenti categoriali utilizzando le codifiche one-of-*c* per la durata della procedura. Se esistono categorie *c* di una variabile, la variabile viene archiviata come vettori c, con la prima categoria indicata (1,0,...,0), la categoria successiva (0,1,0,...,0), ..., e la categoria finale (0,0,...,0,1).

Questo schema di codifica aumenta il numero di pesi sinaptici e può generare un addestramento più lento; tuttavia, metodi di codifica più "compatti" generano di solito reti neurali con adattamento peggiore. Se l'addestramento della rete sta procedendo molto lentamente, cercare di ridurre il numero di categorie nei predittori di categorie mediante la combinazione di categorie simili o casi di rilascio con categorie estremamente rare.

Tutte la codifica one-of-*c* è basata sui dati di addestramento, anche se viene definito un campione di verifica o di controllo (vedere Partizioni a pag. 9). Pertanto, se i campioni di verifica o controllo contengono casi con categorie predittive che non sono presenti nei dati di addestramento, tali casi non vengono utilizzati dalla procedura o nel punteggio. Se i campioni di verifica o di controllo contengono casi con categorie di variabili dipendenti che non sono presenti nei dati di addestramento, tali casi non vengono utilizzati dalla procedura ma possono essere inclusi nel punteggio.

Modifica della scala. Per impostazione predefinita viene modificata la scala delle covariate e delle variabili dipendenti dalla scala per migliorare l'addestramento della rete. Tale modifica viene eseguita interamente sulla base dei dati di addestramento, anche se viene definito un campione di verifica o di controllo (vedere Partizioni a pag. 9), ovvero in base al tipo di modifica di scala, la media, la deviazione standard, il valore minimo o massimo di una variabile dipendente o covariata vengono elaborati utilizzando soltanto i dati di addestramento. Se si specifica una variabile per definire le partizioni, è importante che tali covariate o variabili dipendenti abbiano distribuzioni simili nei campioni di addestramento, test e controllo.

Ponderazione. La ponderazione viene ignorata da questa procedura.

Replica dei risultati. Se si desidera replicare esattamente i risultati ottenuti, utilizzare lo stesso valore di inizializzazione per il generatore di numeri casuali, lo stesso ordine dei dati e delle variabili, oltre alle stesse impostazioni della procedura. Di seguito vengono riportati ulteriori dettagli su questo argomento.

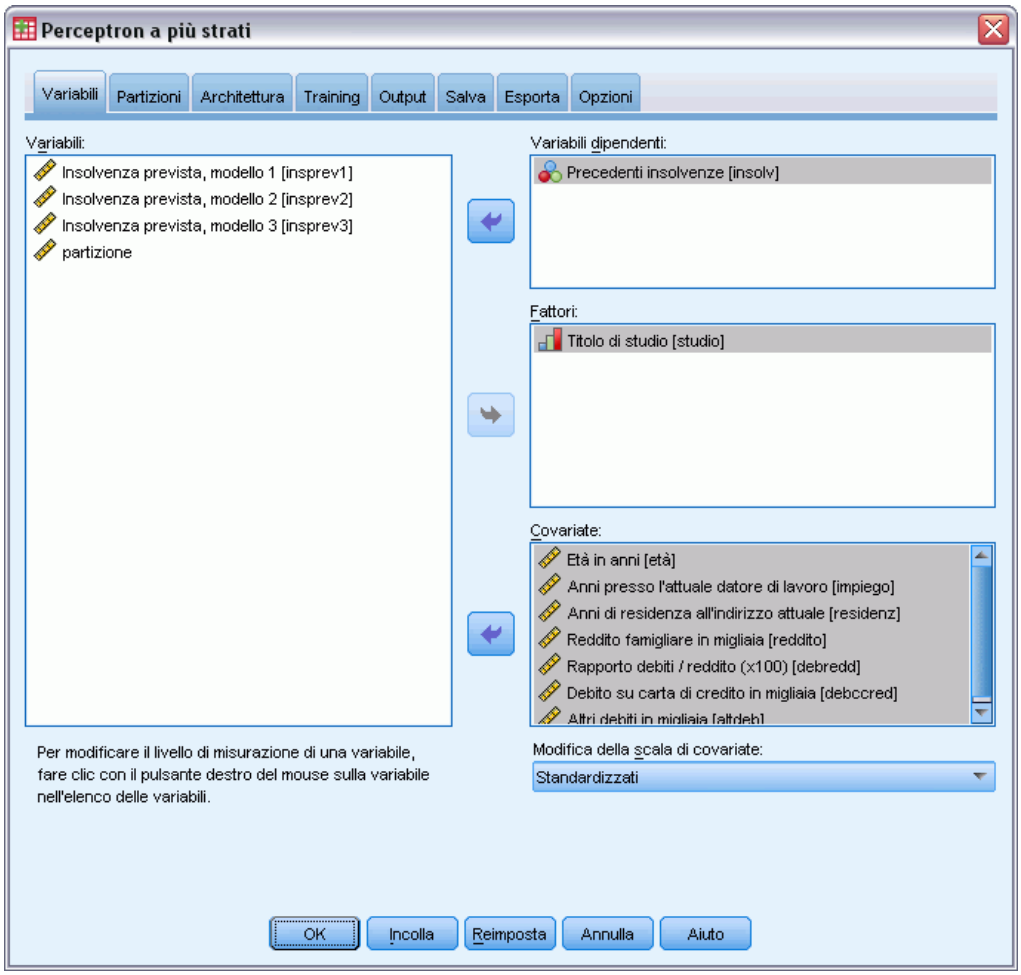
- **Generazione di numeri casuali.** La procedura utilizza la generazione di numeri casuali durante l'assegnazione casuale delle partizioni, il sottocampionamento casuale per l'inizializzazione dei pesi sinaptici, il sottocampionamento casuale per la selezione automatica dell'architettura e l'algoritmo di Simulated Annealing utilizzato nella inizializzazione ponderale e nella selezione automatica dell'architettura. Per riprodurre gli stessi risultati randomizzati in futuro, utilizzare lo stesso valore di inizializzazione per il generatore di numeri casuali prima di ogni successione della procedura Perceptron a più strati. Vedere Preparazione dei dati per l'analisi a pag. 39 per istruzioni passo passo.
- **Ordine dei casi.** I metodi di training in linea e mini-batch (vedere Training a pag. 14) dipendono in modo esplicito dall'ordine dei casi; tuttavia, anche il training batch dipende da tale ordine poiché l'inizializzazione dei pesi sinaptici comporta il sottocampionamento dall'insieme di dati.

  Per ridurre al minimo gli effetti dell'ordine, disporre i casi in ordine casuale. Per verificare la stabilità di una data soluzione, può essere utile ottenere più soluzioni diverse con casi disposti in ordini casuali diversi. Se i file sono particolarmente grandi, è possibile eseguire più analisi con un campione di casi disposti in più ordini casuali.
- **Ordine delle variabili.** I risultati possono essere influenzati dall'ordine delle variabili negli elenchi dei fattori e delle covariate a causa del modello differente dei valori iniziali assegnati quando viene modificato l'ordine delle variabili. Come per gli effetti conseguenti all'ordine dei casi, è possibile provare a utilizzare ordini differenti (è sufficiente effettuare il trascinamento negli elenchi di fattori e di covariate) per stimare la stabilità di una determinata soluzione.

Creazione di una rete Perceptron a più strati

Dai menu, scegliere:
Analizza > Neural Networks > Perceptron a più strati...

Figura 2-1
Perceptron a più strati: scheda Variabili



- Selezionare almeno una variabile dipendente.
- Selezionare almeno un fattore o una covarianza.

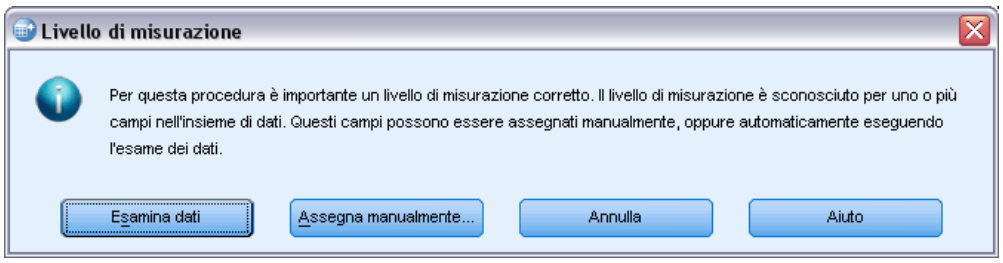
Opzionalmente, nella scheda Variabili è possibile cambiare il metodo per la modifica della scala delle covariate. Le scelte sono:

- **Standardizzati.** Sottrarre la media e dividere per la deviazione standard (x−media)/s.
- **Normalizzati.** Sottratte il minimo e dividere per l'intervallo, (x−min)/(max−min). Valori normalizzati compresi tra 0 e 1.
- **Normalizzati corretti.** La versione corretta della sottrazione del minimo e della divisione per l'intervallo [2*(x−min)/(max−min)]−1. I valori normalizzati corretti rientrano nell'intervallo compreso tra −1 e 1.
- **Nessuna.** Non viene eseguita la modifica delle scala delle covariate.

Campi con livello di misurazione sconosciuto

L'avviso Livello di misurazione viene visualizzato quando il livello di misurazione di una o più variabili (campi) dell'insieme di dati è sconosciuto. Poiché influisce sul calcolo dei risultati di questa procedura, il livello di misurazione deve essere definito per tutte le variabili.

Figura 2-2
Avviso Livello di misurazione

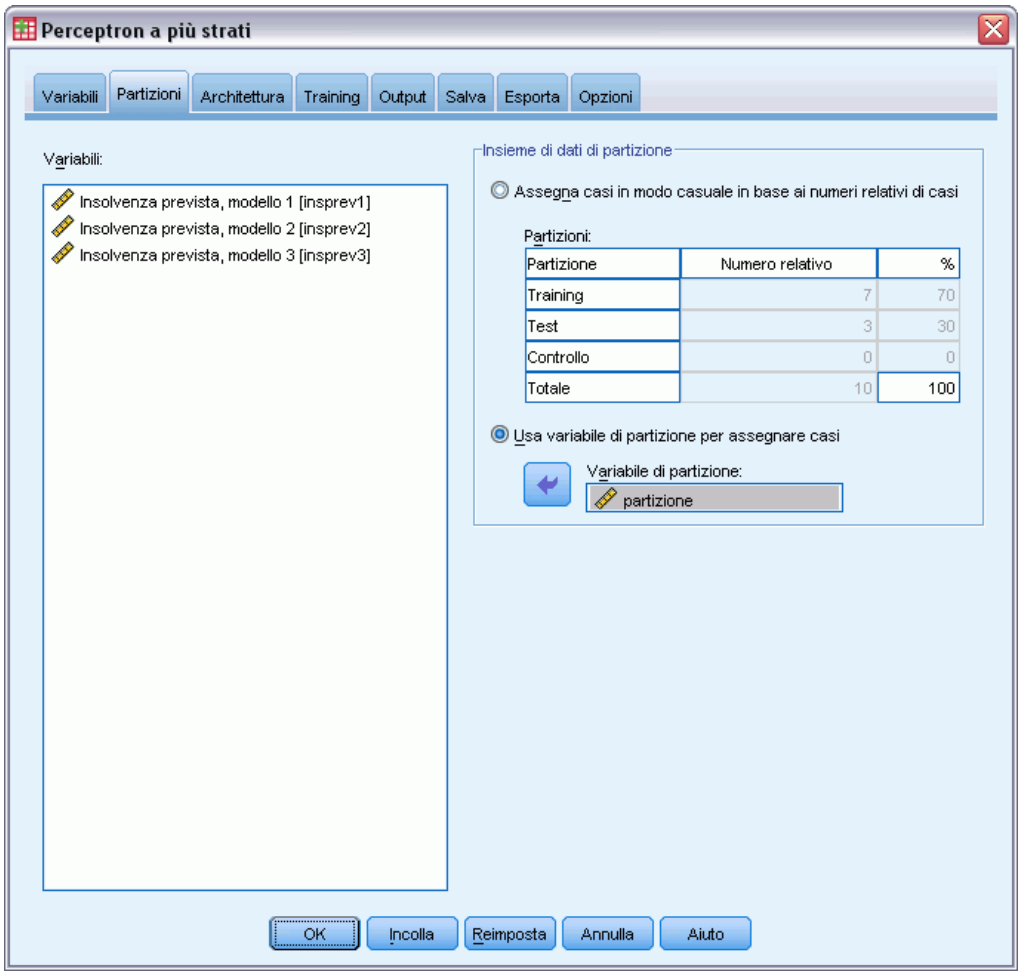


- **Esamina dati.** Legge i dati dell'insieme di dati attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con insiemi di dati di grandi dimensioni, questa operazione può richiedere del tempo.
- **Assegna manualmente.** Apre una finestra di dialogo che elenca tutti i campi con livello di misurazione sconosciuto, mediante la quale è possibile assegnare un livello di misurazione a questi campi. Il livello di misurazione si può assegnare anche nella Visualizzazione variabili dell'Editor dei dati.

Dal momento che il livello di misurazione è importante per questa procedura, è possibile accedere alla finestra di dialogo per la sua esecuzione solo quando per tutti i campi è stato definito un livello di misurazione.

Partizioni

Figura 2-3
Perceptron a più strati: Scheda Partizioni



Insieme di dati di partizione. Questo gruppo specifica il metodo di partizionamento dell'insieme di dati attivo in campioni di addestramento, test e controllo. Il **campione di addestramento** include i record di dati utilizzati per formare una rete neurale; una percentuale di casi nell'insieme di dati deve essere assegnata al campione di addestramento per ottenere un modello. Il **campione di verifica** è un insieme indipendente di record di dati utilizzato per tenere traccia degli errori durante l'addestramento per evitare un eccesso di addestramento. Si consiglia di creare un campione di addestramento; l'addestramento della rete sarà solitamente più efficace se il campione di verifica è più piccolo del campione di addestramento. Il **campione di controllo** è un altro insieme indipendente di record di dati utilizzato per valutare la rete neurale finale; l'errore per il campione

di controllo fornisce una stima “attendibile” della capacità predittiva del modello poiché i casi di controllo non sono stati utilizzati per generare il modello.

- **Assegna casi in modo casuale in base ai numeri relativi di casi.** Specificare il numero relativo (rapporto) di casi assegnati casualmente per ogni campione (addestramento, verifica e controllo). La colonna % indica la percentuale di casi che verranno assegnati a ogni campione in base ai numeri relativi specificati.

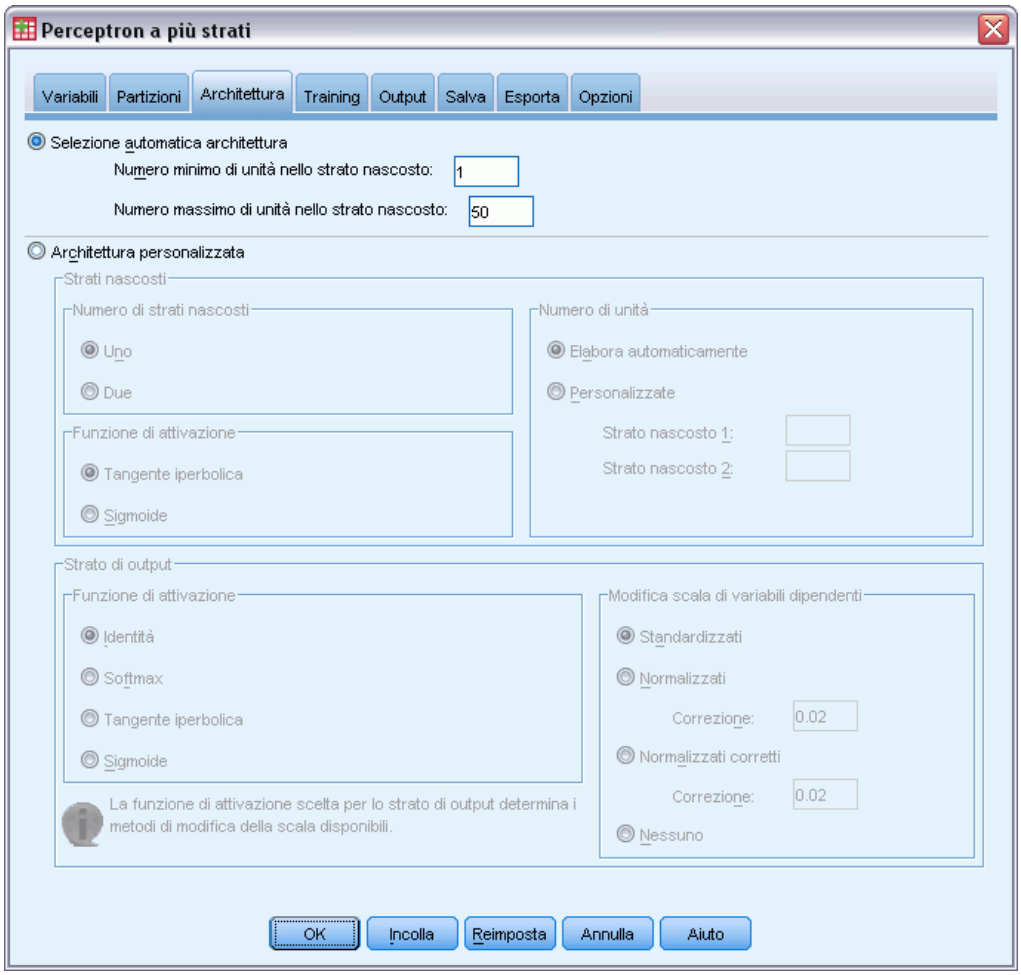
  Ad esempio, specificare 7, 3, 0 come numeri relativi per i campioni di addestramento, test e controllo, corrisponde a 70%, 30% e 0%. Specificare 2, 1, 1 come numeri relativi, corrisponde a 50%, 25% e 25%; 1, 1, 1 corrisponde a dividere l’insieme di dati in tre parti uguali di addestramento, test e controllo.

- **Usa variabile di partizione per assegnare casi.** Specificare una variabile numerica che assegni ogni caso nell’insieme di dati attivo al campione di addestramento, test e controllo. I casi con valore positivo nella variabile vengono assegnati al campione di addestramento, i casi con valore 0 al campione di verifica e i casi con un valore negativo al campione di controllo. I casi con un valore di sistema mancante vengono esclusi dall’analisi. I valori mancanti definiti dall’utente per la variabile di partizione sono sempre considerati validi.

  Nota: l’utilizzo di una variabile di partizionamento non garantisce risultati identici nelle esecuzioni successive della procedura. Vedere “Replica dei risultati” nell’argomento principale di Perceptron a più strati.

Architettura

Figura 2-4
Perceptron a più strati: Scheda Architettura



La scheda Architettura viene utilizzata per specificare la struttura della rete. La procedura consente di selezionare automaticamente l'architettura "migliore" oppure di specificare un'architettura personalizzata.

La selezione automatica dell'architettura consente di creare una rete con uno strato nascosto. Specificare i valori minimo e massimo di unità consentite nello strato nascosto; la selezione automatica dell'architettura elabora il numero "migliore" di unità. La selezione automatica dell'architettura utilizza le funzioni di attivazione predefinite per gli strati nascosti e di output.

La selezione personalizzata dell'architettura consente un controllo avanzato degli strati nascosti e di output e può risultare notevolmente utile quando si conosce in anticipo l'architettura desiderata oppure quando si necessita di modificare leggermente i risultati della selezione automatica dell'architettura.

Strati nascosti

Lo strato nascosto contiene nodi di rete (unità) invisibili. Ogni unità nascosta è una funzione della somma ponderata degli input. La funzione è quella di attivazione e i valori dei pesi vengono determinati dall'algoritmo di stima. Se la rete contiene un secondo strato nascosto, ogni unità nascosta nel secondo strato è una funzione della somma ponderata delle unità nel primo strato nascosto. La stessa funzione di attivazione viene utilizzata in entrambi gli strati.

Numero di strati nascosti. Un perceptron a più strati può avere uno o due strati nascosti.

Funzione di attivazione. La funzione di attivazione "collega" le somme ponderate delle unità in uno strato ai valori delle unità nello strato successivo.

- **Tangente iperbolica.** Questa funzione ha la seguente forma: $\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$. Prende gli argomenti a valori reali e li trasforma nell'intervallo (–1, 1). Quando viene utilizzata la selezione automatica dell'architettura, questa è la funzione di attivazione per tutte le unità negli strati nascosti.
- **Sigmoide.** Questa funzione ha la seguente forma: $\gamma(c) = 1/(1+e^{-c})$. Prende gli argomenti a valori reali e li trasforma nell'intervallo (0, 1).

Numero di unità. Il numero di unità in ogni strato nascosto può essere specificato esplicitamente oppure può essere determinato automaticamente dall'algoritmo di stima.

Strato di output

Lo strato di output contiene le variabili di destinazione (dipendenti).

Funzione di attivazione. La funzione di attivazione "collega" le somme ponderate delle unità in uno strato ai valori delle unità nello strato successivo.

- **Identità.** Questa funzione ha la seguente forma: $\gamma(c) = c$. Prende gli argomenti a valori reali e li restituisce invariati. Quando viene utilizzata la selezione automatica dell'architettura, questa è la funzione di attivazione per le unità nello strato di output se sono presenti variabili dipendenti di scala.
- **Softmax.** Questa funzione ha la seguente forma: $\gamma(c_k) = \exp(c_k)/\Sigma_j\exp(c_j)$. Prende un vettore di argomenti a valori reali, lo trasforma in un vettore i cui elementi rientrano nell'intervallo (0, 1) e lo somma a 1. Softmax è disponibile solo se tutte le variabili dipendenti sono categoriali. Quando viene utilizzata la selezione automatica dell'architettura, questa è la funzione di attivazione per le unità nello strato di output se tutte le variabili dipendenti sono categoriali.
- **Tangente iperbolica.** Questa funzione ha la seguente forma: $\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$. Prende gli argomenti a valori reali e li trasforma nell'intervallo (–1, 1).
- **Sigmoide.** Questa funzione ha la seguente forma: $\gamma(c) = 1/(1+e^{-c})$. Prende gli argomenti a valori reali e li trasforma nell'intervallo (0, 1).

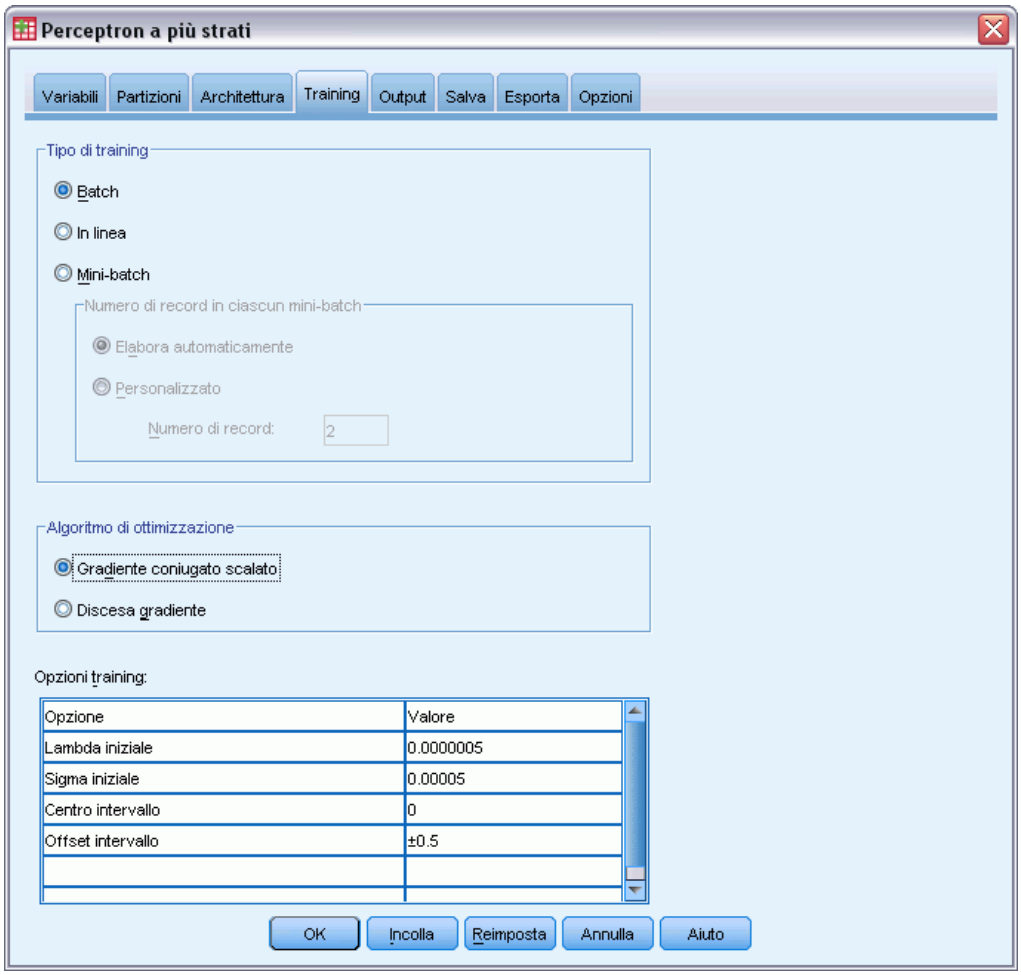
Modifica scala di variabili dipendenti. Questi controlli sono disponibili solo se è stata selezionata almeno una variabile dipendente di scala.

- **Standardizzati.** Sottrarre la media e dividere per la deviazione standard (x−media)/s.

- **Normalizzati.** Sottratte il minimo e dividere per l'intervallo, (x−min)/(max−min). I valori normalizzati sono compresi tra 0 e 1. Questo è il metodo di modifica della scala per le variabili dipendenti di scala se lo strato di output utilizza la funzione di attivazione sigmoide. L'opzione di correzione specifica un numero piccolo, ε, che viene applicato come correzione alla formula di modifica della scala; tale correzione garantisce che tutti i valori delle variabili dipendenti di scala di cui è stata modificata la scala saranno inclusi nell'intervallo della funzione di attivazione. In particolare, i valori 0 e 1, che sono presenti nella formula non corretta quando x assume i relativi valori minimo e massimo, definiscono i limiti dell'intervallo della funzione sigmoide ma non sono inclusi in tale intervallo. La formula corretta è [x−(min−ε)]/[(max+ε)−(min−ε)]. Specificare un numero maggiore o uguale a 0.
- **Normalizzati corretti.** La versione corretta della sottrazione del valore minimo e della divisione per l'intervallo, [2*(x−min)/(max−min)]−1. I valori normalizzati corretti rientrano nell'intervallo compreso tra −1 e 1. Questo è il metodo di modifica della scala richiesto per le variabili dipendenti di scala se lo strato di output utilizza la funzione di attivazione tangente iperbolica. L'opzione di correzione specifica un numero piccolo, ε, che viene applicato come correzione alla formula di modifica della scala; tale correzione garantisce che tutti i valori delle variabili dipendenti di scala di cui è stata modificata la scala saranno inclusi nell'intervallo della funzione di attivazione. In particolare, i valori −1 e 1, che sono presenti nella formula non corretta quando x assume il valore minimo e massimo, definiscono i limiti dell'intervallo della funzione tangente iperbolica ma non sono inclusi in tale intervallo. La formula corretta è {2*[(x−(min−ε))/((max+ε)−(min−ε))]}−1. Specificare un numero maggiore o uguale a 0.
- **Nessuna.** La scala delle variabili dipendenti non viene modificata.

Training

Figura 2-5
Perceptron a più strati: Scheda Training



La scheda Training viene utilizzata per specificare la modalità di training della rete. Il tipo di training e l'algoritmo di ottimizzazione determinano le opzioni di training disponibili.

Tipo di training. Il tipo di training determina la modalità in cui la rete elabora i record. Selezionare uno dei tipi di training seguenti:

- **Batch.** Aggiorna i pesi sinaptici solo al termine dei cicli relativi a tutti i record contenenti dati di addestramento, ossia vengono utilizzate le informazioni contenute in tutti i record nell'insieme di dati di addestramento. Il training batch viene utilizzato spesso perché riduce al minimo direttamente l'errore totale; tuttavia potrebbe essere necessario aggiornare i pesi numerose volte finché non viene soddisfatta una delle regole di interruzione e pertanto potrebbero essere necessari numerosi cicli di dati. È l'addestramento più utile per gli insiemi di dati "più piccoli".

- **In linea.** Aggiorna i pesi sinaptici dopo ogni singolo record contenente i dati di training, ossia vengono utilizzate le informazioni contenute in un record alla volta. Il training in linea preleva in modo continuo un record e aggiorna i pesi finché non viene soddisfatta una delle regole di interruzione. Se tutti i record vengono utilizzati una volta e nessuna delle regole di interruzione viene soddisfatta, il processo continua riciclando i record dei dati. L'addestramento in linea garantisce risultati migliori rispetto a quello batch per gli insiemi di dati "più grandi" con predittori associati, ossia se esistono numerosi record e numerosi input e i relativi valori non sono indipendenti tra loro, questo tipo di addestramento può ottenere più rapidamente una risposta accettabile rispetto all'addestramento batch.
- **Mini-batch.** Divide i record contenenti dati di training in gruppi di dimensione approssimativamente uguale, quindi aggiorna i pesi sinaptici dopo avere concluso il ciclo di un gruppo, ossia vengono utilizzate le informazioni contenute in un gruppo di record, quindi il processo ricicla il gruppo di dati, se necessario. L'addestramento mini-batch offre un compromesso tra i tipi batch e in linea e può risultare particolarmente adatto per gli insiemi di dati di "dimensione intermedia". La procedura può determinare automaticamente il numero di record di training per mini-batch oppure è possibile specificare un numero intero superiore a 1 e inferiore o uguale al numero massimo di casi da archiviare in memoria. È possibile impostare il numero massimo di casi da archiviare in memoria nella scheda Opzioni.

Algoritmo di ottimizzazione. Questo metodo viene utilizzato per stimare i pesi sinaptici.

- **Gradiente coniugato scalato.** Le ipotesi che giustificano l'utilizzo di metodi a gradiente coniugate si applicano solo ai tipi di addestramento batch; pertanto questo metodo non è disponibile per l'addestramento in linea o mini-batch.
- **Discesa gradiente.** Questo metodo deve essere utilizzato con il training in linea o mini-batch e può essere utilizzato anche con il tipo batch.

Opzioni training. Le opzioni di training consentono di perfezionare l'algoritmo di ottimizzazione. In genere, non è necessario cambiare queste impostazioni a meno che non si verifichino problemi di rete relativi alla stima.

Sono disponibili le seguenti opzioni di addestramento per l'algoritmo gradiente coniugato scalato:

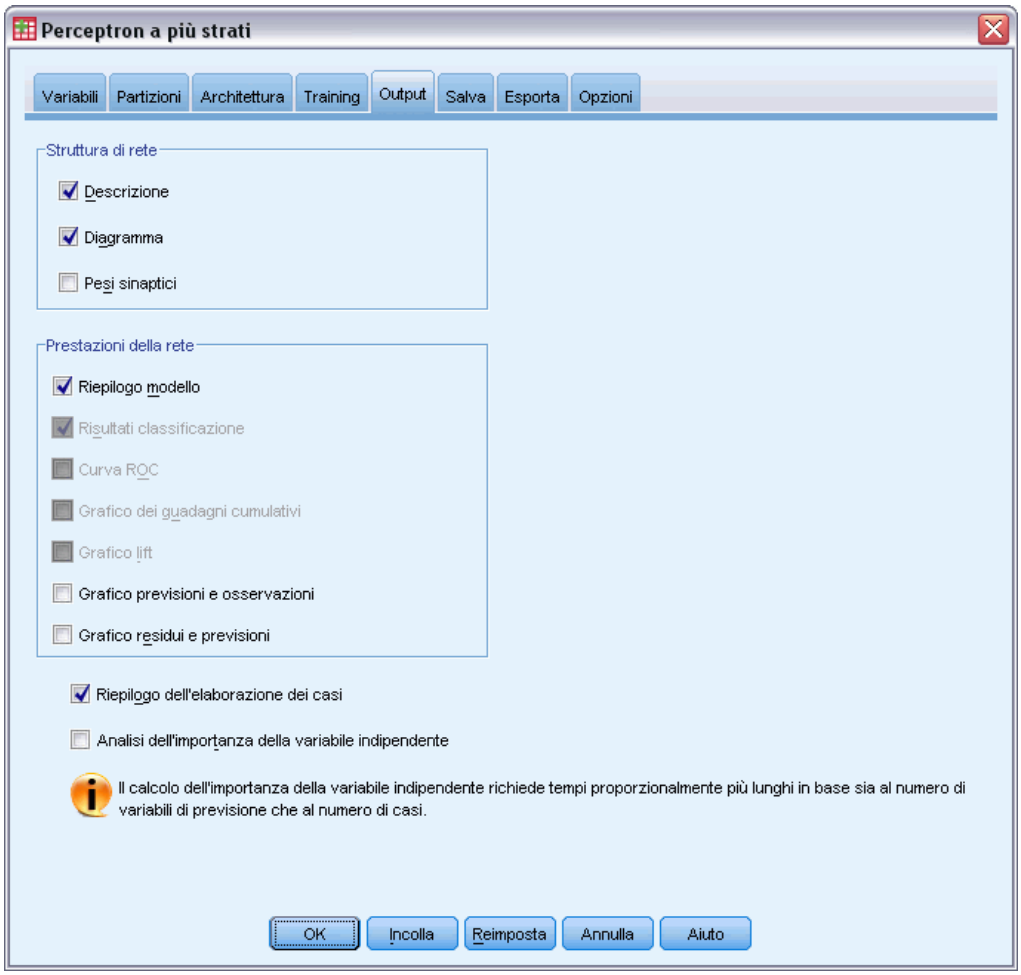
- **Lambda iniziale.** Il valore iniziale del parametro lambda per l'algoritmo gradiente coniugato scalato. Specificare un numero maggiore di 0 e inferiore a 0,000001.
- **Sigma iniziale.** Il valore iniziale del parametro sigma per l'algoritmo gradiente coniugato scalato. Specificare un numero maggiore di 0 e inferiore a 0.0001.
- **Centro intervallo e Offset intervallo.** Il centro intervallo (a_0) e l'offset intervallo (a) definiscono l'intervallo $[a_0-a, a_0+a]$, in cui i vettori ponderali vengono generati in modo casuale quando viene utilizzato l'algoritmo di Simulated Annealing. L'algoritmo di Simulated Annealing viene utilizzato per superare un valore minimo locale con l'obiettivo di individuare il valore minimo globale durante l'applicazione dell'algoritmo di ottimizzazione. Questo approccio viene utilizzato nell'inizializzazione ponderale e nella selezione automatica dell'architettura. Specificare un numero per il centro dell'intervallo e un numero maggiore di 0 per l'offset dell'intervallo.

Sono disponibili le seguenti opzioni di addestramento per l'algoritmo gradiente coniugato scalato:

- **Tasso di apprendimento iniziale.** Il valore iniziale del tasso di apprendimento per l'algoritmo di discesa del gradiente. Un tasso di apprendimento superiore implica un training della rete più veloce ma c'è la possibilità di un'eventuale instabilità. Specificare un numero maggiore di 0.
- **Limite inferiore del tasso di apprendimento.** Il limite inferiore del tasso di apprendimento per l'algoritmo di discesa del gradiente. Questa impostazione si applica solo all'addestramento in linea e mini-batch. Specificare un numero maggiore di 0 e minore del tasso di apprendimento iniziale.
- **Momento.** Il parametro del momento iniziale per l'algoritmo di discesa del gradiente, che consente di impedire instabilità causate da un tasso di apprendimento troppo alto. Specificare un numero maggiore di 0.
- **Riduzione del tasso di apprendimento, espresso in epoche.** Il numero di epoche (p) o cicli di dati del campione di addestramento per ridurre il tasso di apprendimento iniziale al limite inferiore del tasso di apprendimento quando la discesa del gradiente viene utilizzata con l'addestramento in linea o mini-batch. In questo modo è possibile controllare il fattore di decremento del tasso di apprendimento $\beta = (1/pK)*\ln(\eta_0/\eta_{low})$, in cui η_0 è il tasso di apprendimento iniziale, η_{low} è il limite inferiore del tasso di apprendimento e K è il numero totale di mini-batch (o il numero di record per l'addestramento in linea) nell'insieme di dati di addestramento. Specificare un numero intero maggiore di 0.

Output

Figura 2-6
Perceptron a più strati: scheda Output



Struttura di rete. Visualizza informazioni di riepilogo sulla rete neurale.

- **Descrizione.** Visualizza informazioni sulla rete neurale, incluse le variabili dipendenti, numero di unità di input e output, numero di unità e di strati nascosti e funzioni di attivazione.
- **Diagramma.** Visualizza il diagramma di rete come un grafico non modificabile. Tenere presente che con l'aumento del numero di livelli dei fattori e di covariate, il diagramma diventa più difficile da interpretare.
- **Pesi sinaptici.** Visualizza le stime dei coefficienti che mostrano la relazione tra le unità in un determinato strato e le unità nello strato seguente. I pesi sinaptici si basano sul campione di addestramento se l'insieme di dati attivo è diviso in dati di addestramento, test e controllo. Tenere presente che il numero di pesi sinaptici può essere piuttosto grande e che tali valori non vengono solitamente utilizzati per interpretare i risultati della rete.

Prestazioni della rete. Visualizza i risultati utilizzati per determinare se il modello è "valido". *Nota*: i grafici in questo gruppo si basano sui campioni di addestramento e verifica oppure solo sul campione di addestramento se non è presente un campione di verifica.

- **Riepilogo del modello.** Visualizza un riepilogo dei risultati della rete neurale per partizione e valore globale, inclusi l'errore, l'errore relativo o la percentuale di previsioni errate, la regola di interruzione utilizzata per interrompere l'addestramento e il tempo di addestramento.

  L'errore è quello relativo alla somma dei quadrati quando allo strato di output viene applicata la funzione di attivazione di identità, sigmoide o tangente iperbolica. L'errore è quello relativo all'entropia incrociata quando allo strato di output viene applicata la funzione di attivazione Softmax.

  Gli errori relativi o le percentuali di previsioni errate vengono visualizzati a seconda dei livelli di misurazione delle variabili dipendenti. Se una variabile dipendente ha un livello di misurazione di scala, viene visualizzato l'errore relativo complessivo medio (relativo al modello di media). Se tutte le variabili dipendenti sono categoriali, viene visualizzata la percentuale media delle previsioni non corrette. Vengono inoltre visualizzati gli errori relativi o le percentuali delle previsioni non corrette per le singole variabili dipendenti.

- **Risultati della classificazione (Analisi discriminante).** Visualizza una tabella di classificazione per ogni variabile dipendente categoriale per partizione e valore globale. Ogni tabella fornisce il numero di casi classificati correttamente e non correttamente per ogni categoria di variabile dipendente. Viene inoltre indicata la percentuale dei casi totali classificati correttamente.

- **Curva ROC.** Visualizza una curva ROC (Receiver Operating Characteristic) per ogni variabile dipendente categoriale. Visualizza inoltre una tabella che fornisce l'area sotto ogni curva. Per una specifica variabile dipendente, il grafico ROC visualizza una curva per ogni categoria. Se la variabile dipendente ha due categorie, ogni curva considera la categoria in questione come stato positivo rispetto alle altre categorie. Se la variabile dipendente ha più di due categorie, ogni curva considera la categoria in questione come stato positivo rispetto a tutte le altre categorie.

- **Grafico dei guadagni cumulativi.** Visualizza un grafico dei guadagni cumulativi per ogni variabile dipendente categoriale. La vista di una curva per ogni categoria di variabile dipendente è uguale alle curve ROC.

- **Grafico lift.** Visualizza un grafico lift per ogni variabile dipendente categoriale. La vista di una curva per ogni categoria di variabile dipendente è uguale alle curve ROC.

- **Grafico previsioni e osservazioni.** Visualizza un grafico dei valori delle previsioni e osservazioni per ogni variabile dipendente. Per le variabili dipendenti categoriali, vengono visualizzati grafici a scatole raggruppati delle pseudo-probabilità previste per ogni categoria di risposta, con la categoria di risposta osservata come variabile di raggruppamento. Per le variabili dipendenti di scala viene visualizzato un grafico a dispersione.

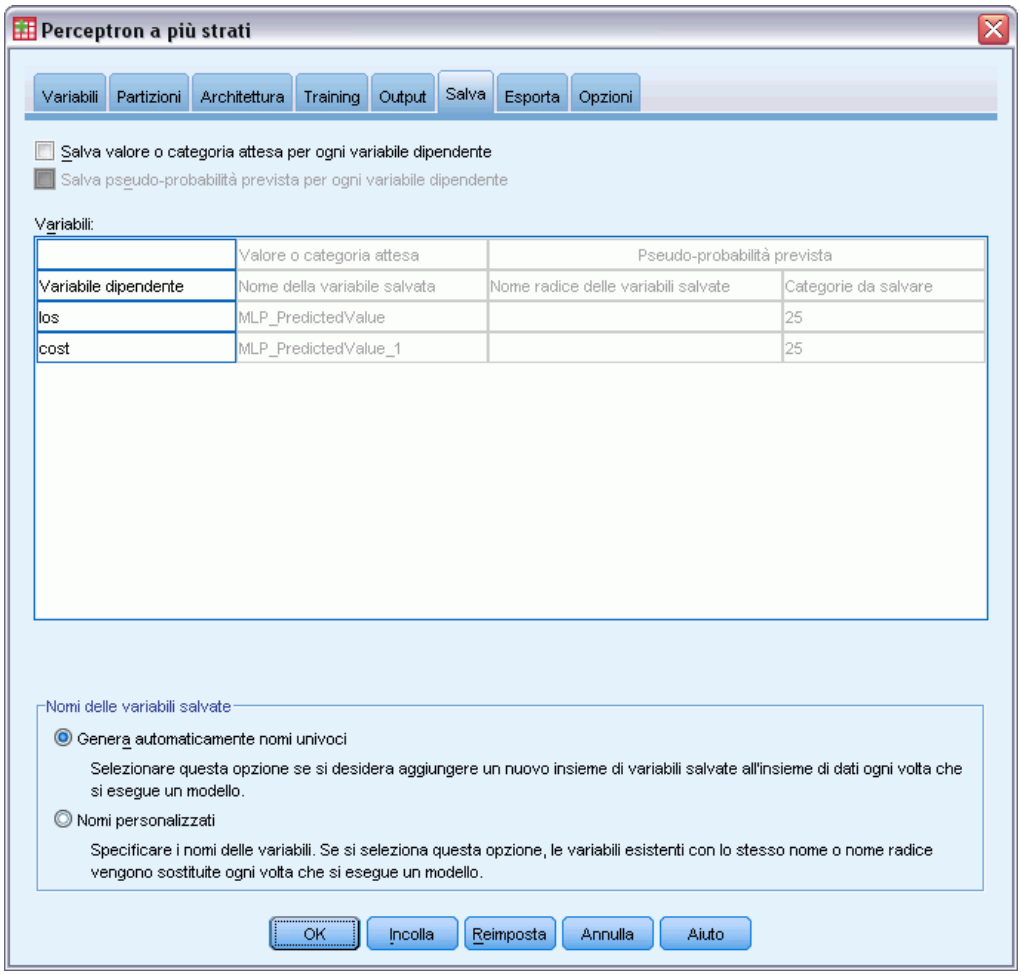
- **Grafico residui e previsioni.** Visualizza un grafico dei residui e delle previsioni per ogni variabile dipendente di scala. Tra i valori residui e attesi non dovrebbero essere presenti modelli visibili. Questo grafico viene generato solo per le variabili dipendenti di scala.

Riepilogo dell'elaborazione dei casi. Visualizza la tabella di riepilogo di elaborazione dei casi, che riepiloga il numero di casi inclusi ed esclusi dall'analisi, in totale e per campioni di addestramento, test e controllo.

Analisi dell'importanza della variabile indipendente. Esegue un'analisi della sensibilità, che calcola l'importanza di ogni predittore nel processo di determinazione della rete neurale. L'analisi si basa sui campioni di addestramento e verifica oppure solo sul campione di addestramento se non è presente un campione di verifica. Questo crea una tabella e un grafico che mostrano l'importanza e l'importanza normalizzata di ogni predittore. Tenere presente che l'analisi della sensibilità è impegnativa a livello di calcolo e richiede parecchio tempo se sono presenti molti predittori o casi.

Salva

Figura 2-7
Perceptron a più strati: scheda Salva



La scheda Salva viene utilizzata per salvare le previsioni come variabili nell'insieme di dati.

- **Salva valore atteso o categoria per ogni variabile dipendente.** Vengono salvati il valore atteso per le variabili dipendenti di scala e la categoria prevista per le variabili dipendenti categoriali.
- **Salva pseudo-probabilità prevista o categoria per ciascuna variabile dipendente.** Salva le pseudo-probabilità previste per le variabili dipendenti categoriali. Una variabile separata viene salvata per ognuna delle prime *n* categorie, dove *n* viene specificato nella colonna Categorie da salvare.

Nomi delle variabili salvate. La generazione automatica del nome assicura il mantenimento di tutto il lavoro. I nomi personalizzati consentono di eliminare/sostituire i risultati di precedenti esecuzioni senza dover prima eliminare le variabili salvate nell'Editor dei dati.

Probabilità e pseudo-probabilità

Le variabili dipendenti categoriali con l'attivazione Softmax e l'errore di entropia incrociata avranno un valore previsto per ciascuna categoria, in cui ciascun valore atteso indica la probabilità che il caso appartenga alla categoria.

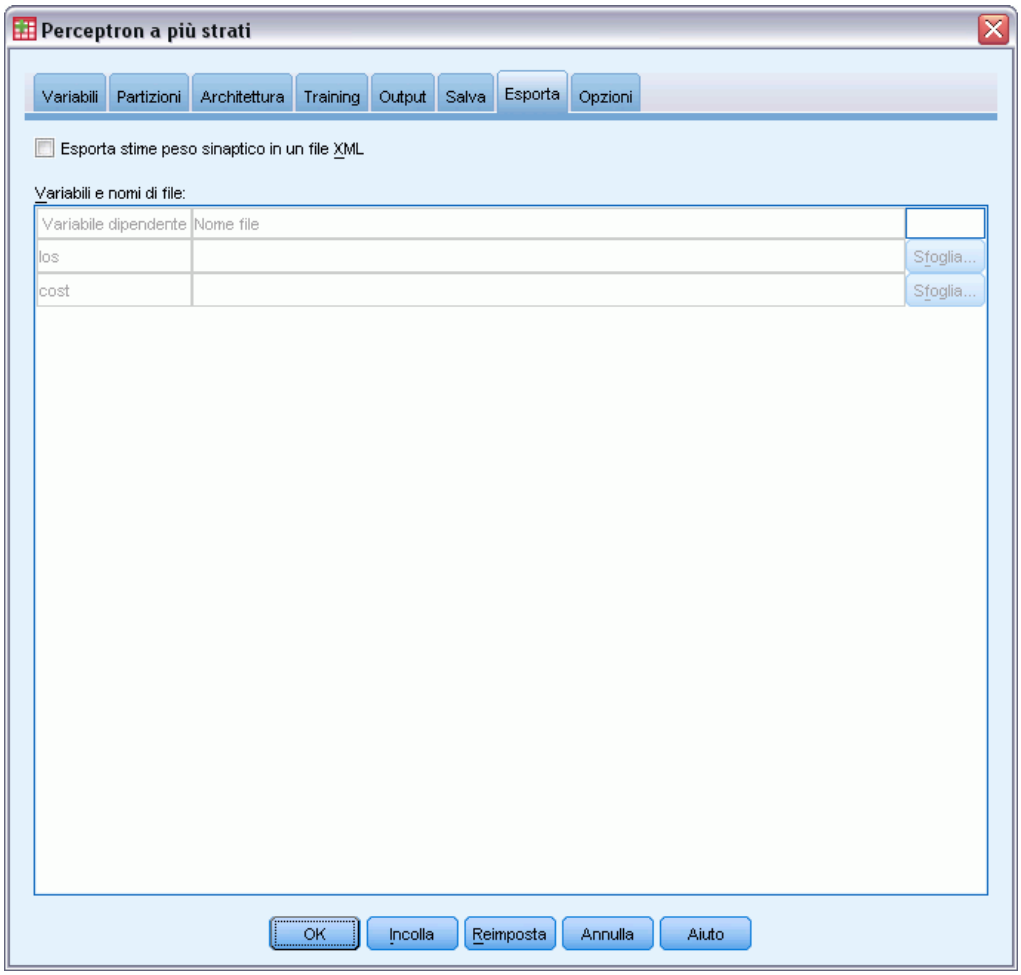
Le variabili dipendenti categoriali con l'errore somma dei quadrati avranno un valore per ogni categoria, ma i valori attesi non possono essere interpretati come probabilità. La procedura salva queste pseudo-probabilità previste anche se nessuna è minore di 0 o maggiore di 1 o la somma di una specifica variabile dipendente non è 1.

I grafici ROC, dei guadagni cumulativi e lift (vedere Output a pag. 17) vengono creati in base alle pseudo-probabilità. Nel caso in cui nessuna pseudo-probabilità sia minore di 0 o maggiore di 1 o la somma di una specifica variabile non sia 1, queste vengono riscalate per essere comprese tra 0 e 1 e in modo che la somma dia come risultato 1. Le pseudo-probabilità vengono riscalate eseguendo la divisione per la relativa somma. Ad esempio, se un caso ha pseudo-probabilità previste di 0,50, 0,60 e 0,40 per una variabile dipendente a tre categorie, ogni pseudo-probabilità viene divisa per la somma 1,50 per ottenere 0,33, 0,40 e 0.27.

Se nessuna delle pseudo-probabilità è negativa, il valore assoluto del valore più basso viene aggiunto a tutte le pseudo-probabilità prima della modifica della scala. Ad esempio, se le pseudo-probabilità sono -0,30, 0.50 e 1,30, aggiungere 0,30 a ogni valore per ottenere 0,00, 0,80 e 1,60. Quindi, dividere ogni nuovo valore per la somma 2,40 per ottenere 0,00, 0,33 e 0,67.

Esporta

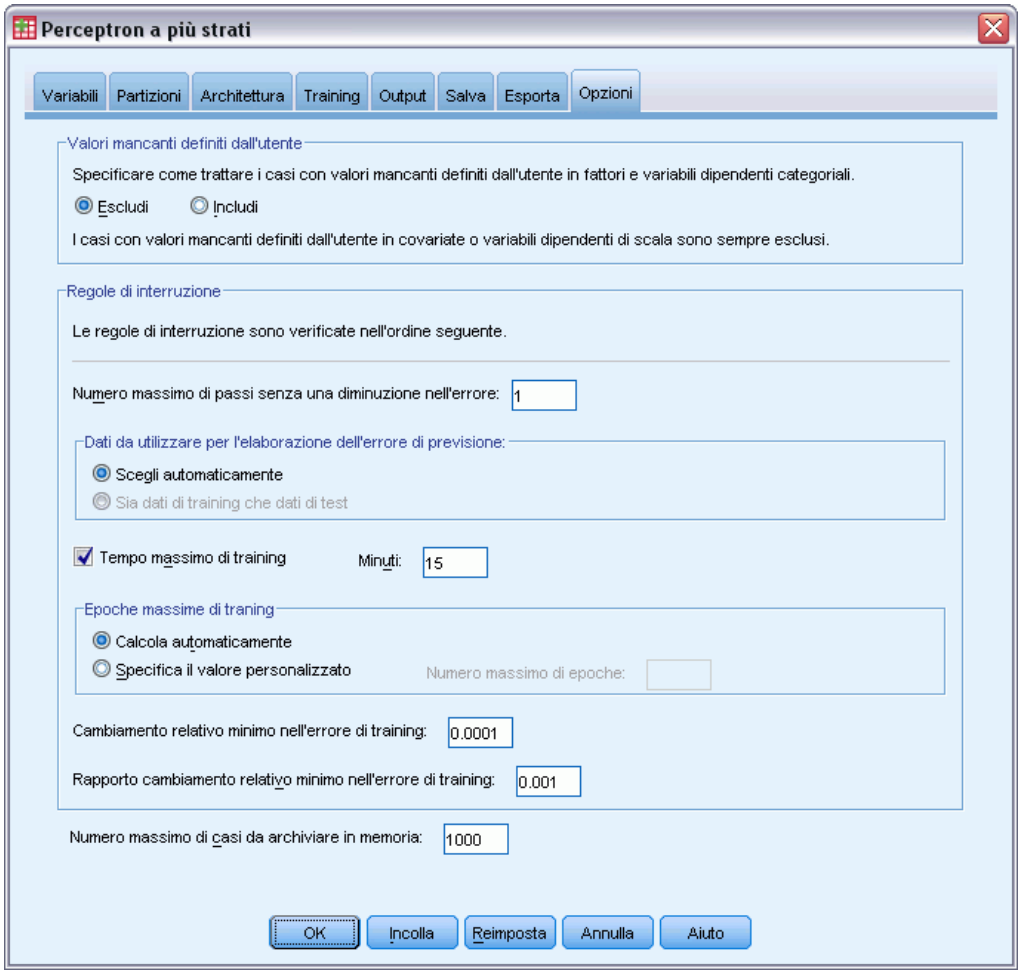
Figura 2-8
Perceptron a più strati: scheda Esporta



La scheda Esporta viene utilizzata per salvare le stime del peso sinaptico per ogni variabile dipendente in un file XML (PMML). È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio. Questa opzione non è disponibile se sono stati definiti file suddivisi.

Opzioni

Figura 2-9
Perceptron a più strati: Scheda Opzioni



Valori mancanti definiti dall'utente. Affinché un caso possa essere incluso nell'analisi, è necessario che i fattori abbiano valori validi. Questi controlli consentono di decidere se i valori mancanti definiti dall'utente devono essere considerati validi tra fattori e variabili dipendenti categoriali.

Regole di interruzione. Queste sono le regole che determinano quando interrompere il training della rete neurale. Il training procede almeno di un ciclo di dati e può essere interrotto in base ai seguenti criteri, che vengono selezionati nell'ordine elencato. Nelle definizioni delle regole di interruzione che seguono, un passo corrisponde a un ciclo di dati per i metodi in linea e mini-batch e a un'iterazione per il metodo batch.

- **Numero massimo di passi senza una diminuzione nell'errore.** Il numero di passi da consentire prima di controllare una diminuzione nell'errore. Se non si verifica una diminuzione nell'errore dopo il numero specificato di passi, il training si interrompe. Specificare un numero intero maggiore di 0. È possibile anche specificare il campione di dati che viene utilizzato per calcolare l'errore. Scegli automaticamente utilizza il campione di verifica, se

esistente, e in caso contrario il campione di addestramento. L'addestramento batch garantisce una diminuzione nell'errore del campione di addestramento dopo ogni ciclo di dati, perciò questa opzione si applica solo all'addestramento batch se esiste un campione di verifica. Sia dati di training che dati di test controlla l'errore per ognuno di questi campioni; questa opzione si applica solo se esiste un campione di verifica.

Nota: Dopo ogni ciclo di dati completo, l'addestramento in linea e mini-batch richiede un ciclo di dati extra per calcolare l'errore di addestramento. Questo ciclo di dati extra può rallentare notevolmente il training ed è pertanto consigliabile fornire un campione di verifica e selezionare Scegli automaticamente in qualsiasi caso.

- **Tempo massimo di training.** Scegliere se specificare un numero massimo di minuti per l'esecuzione dell'algoritmo. Specificare un numero maggiore di 0.
- **Epoche massime di training.** Il numero massimo di epoche (cicli di dati) consentito. Se tale numero viene superato, il training si interrompe. Specificare un numero intero maggiore di 0.
- **Cambiamento relativo minimo nell'errore di addestramento.** Il training si interrompe se il cambiamento relativo nell'errore di training confrontato al passo precedente è inferiore al valore del criterio. Specificare un numero maggiore di 0. Per il training in linea e mini-batch, questo criterio viene ignorato solo se i dati di test vengono utilizzati per calcolare l'errore.
- **Cambiamento relativo minimo nel rapporto dell'errore di addestramento.** Il training si interrompe se il rapporto tra l'errore di training e l'errore del modello null è inferiore al valore del criterio. Il modello null prevede il valore medio per tutte le variabili dipendenti. Specificare un numero maggiore di 0. Per il training in linea e mini-batch, questo criterio viene ignorato solo se i dati di test vengono utilizzati per calcolare l'errore.

Numero massimo di casi da archiviare in memoria. Controlla le impostazioni seguenti entro gli algoritmi di Perceptron a più strati. Specificare un numero intero maggiore di 1.

- Nella selezione automatica dell'architettura, la dimensione del campione usato per determinare la struttura della rete è min(1000,*dimmem*), in cui *dimmem* è il numero massimo di casi da archiviare in memoria.
- Nell'addestramento mini-batch con calcolo automatico del numero di mini-batch, il numero di mini-batch è min(max(*M*/10,2),*dimmem*), in cui *M* è il numero di casi nel campione di addestramento.

Capitolo
3

Funzione a base radiale

La procedura Funzione a base radiale (RBF) produce un modello predittivo per una o più variabili dipendenti (di destinazione) basato sui valori delle variabili predittore.

Esempio. Un fornitore di telecomunicazioni ha segmentato la base clienti per modelli di utilizzo del servizio, suddividendo i clienti in quattro categorie. Una rete RBF che utilizza i dati demografici per stimare il gruppo di appartenenza consente alla società di personalizzare le offerte per singoli potenziali clienti.

Variabili dipendenti. Le variabili dipendenti possono essere:

- **Nominale.** Una variabile può essere considerata nominale quando i relativi valori rappresentano categorie prive di ordinamento intrinseco, per esempio l'ufficio di una società, Tra gli esempi di variabili nominali troviamo la regione, il codice postale e la religione.
- **Ordinale.** Una variabile può essere considerata ordinale quando i relativi valori rappresentano categorie con qualche ordinamento intrinseco, per esempio i gradi di soddisfazione per un servizio, da molto insoddisfatto a molto soddisfatto, i punteggi di atteggiamento corrispondenti a gradi di soddisfazione o fiducia e i punteggi di preferenza.
- **Scala.** Una variabile può essere considerata di scala (continua) quando i relativi valori rappresentano categorie ordinate con una metrica significativa, tale che i confronti fra le distanze dei relativi valori siano appropriati. Esempi di variabili di scala sono l'età espressa in anni o il reddito espresso in migliaia di Euro.

  La procedura presume che il livello di misurazione appropriato sia stato assegnato a tutte le variabili dipendenti, sebbene sia possibile modificare temporaneamente il livello di misurazione di una variabile facendo clic con il pulsante destro del mouse sulla variabile nell'elenco delle variabili sorgente e scegliendo un livello di misurazione dal menu di scelta rapida.

L'icona accanto a ciascuna variabile nell'elenco delle variabili identifica il livello di misurazione e il tipo di dati.

	Numerico	Stringa	Data	Ora
Scala (continuo)		n/d		
Ordinale				
Nominale				

Variabili indipendenti. Possono essere specificate come fattori (categoriali) o covariate (scala).

Codifiche variabili categoriali. La procedura ricodifica temporaneamente le variabili dipendenti e indipendenti categoriali utilizzando le codifiche one-of-*c* per la durata della procedura. Se esistono categorie *c* di una variabile, la variabile viene archiviata come vettori c, con la prima categoria indicata (1,0,...,0), la categoria successiva (0,1,0,...,0), ..., e la categoria finale (0,0,...,0,1).

Questo schema di codifica aumenta il numero di pesi sinaptici e può generare un addestramento più lento, ma metodi di codifica più "compatti" generano di solito reti neurali con adattamento peggiore. Se l'addestramento della rete sta procedendo molto lentamente, cercare di ridurre il numero di categorie nei predittori di categorie mediante la combinazione di categorie simili o casi di rilascio con categorie estremamente rare.

Tutte la codifica one-of-*c* è basata sui dati di addestramento, anche se viene definito un campione di verifica o di controllo (vedere Partizioni a pag. 28). Pertanto, se i campioni di verifica o controllo contengono casi con categorie predittive che non sono presenti nei dati di addestramento, tali casi non vengono utilizzati dalla procedura o nel punteggio. Se i campioni di verifica o di controllo contengono casi con categorie di variabili dipendenti che non sono presenti nei dati di addestramento, tali casi non vengono utilizzati dalla procedura ma possono essere inclusi nel punteggio.

Modifica della scala. Per impostazione predefinita viene modificata la scala delle covariate e delle variabili dipendenti dalla scala per migliorare l'addestramento della rete. Tale modifica viene eseguita interamente sulla base dei dati di addestramento, anche se viene definito un campione di verifica o di controllo (vedere Partizioni a pag. 28), ovvero in base al tipo di modifica di scala, la media, la deviazione standard, il valore minimo o massimo di una variabile dipendente o covariata vengono elaborati utilizzando soltanto i dati di addestramento. Se si specifica una variabile per definire le partizioni, è importante che tali covariate o variabili dipendenti abbiano distribuzioni simili nei campioni di addestramento, test e controllo.

Ponderazione. La ponderazione viene ignorata da questa procedura.

Replica dei risultati. Se si desidera replicare esattamente i risultati, utilizzare lo stesso valore di inizializzazione per il generatore di numeri casuali e lo stesso ordine di dati, in aggiunta all'utilizzo delle stesse impostazioni della procedura. Di seguito vengono riportati ulteriori dettagli su questo argomento.

- **Generazione di numeri casuali.** La procedura utilizza la generazione di numeri casuali durante l'assegnazione causale delle partizioni. Per riprodurre gli stessi risultati randomizzati in futuro, utilizzare lo stesso valore di inizializzazione per il generatore di numeri casuali per ciascuna esecuzione della procedura Funzione a base radiale. Vedere Preparazione dei dati per l'analisi a pag. 76 per istruzioni passo passo.
- **Ordine dei casi.** I risultati dipendono anche dall'ordine dei dati poiché l'algoritmo cluster a due fasi viene utilizzato per determinare le funzioni a base radiale.

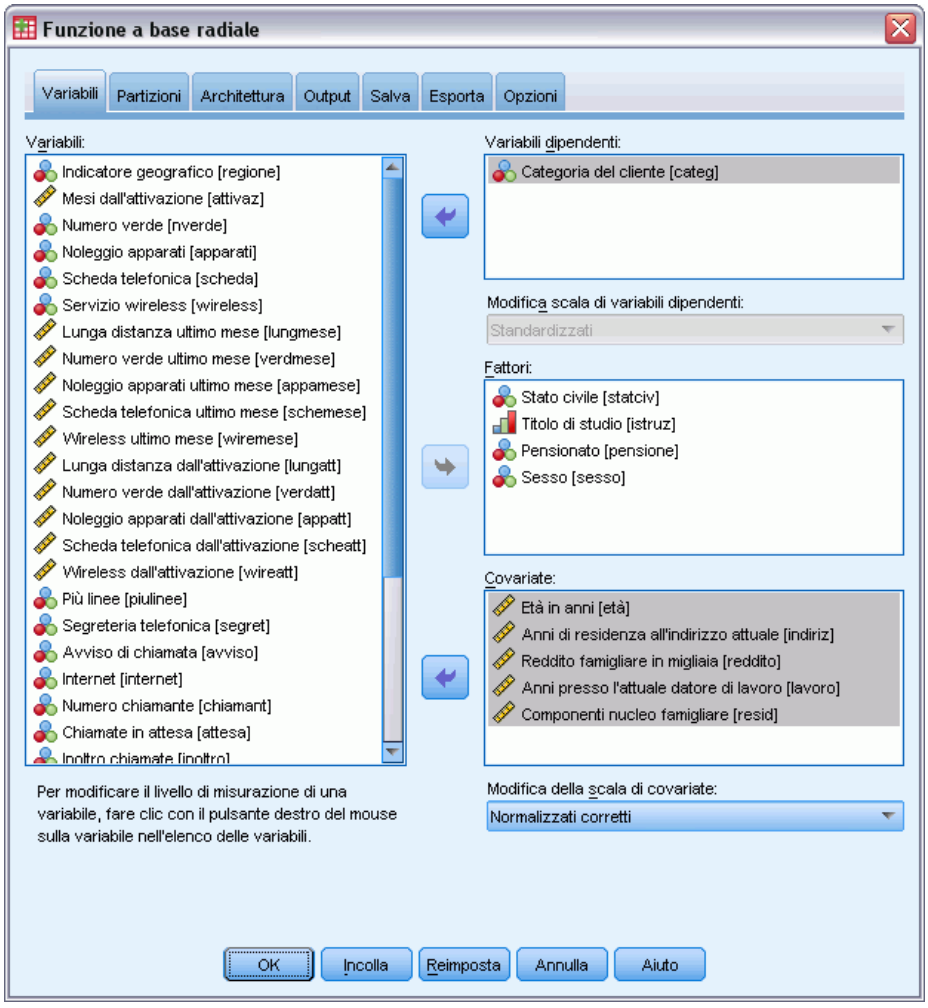
  Per ridurre al minimo gli effetti dell'ordine, disporre i casi in ordine casuale. Per verificare la stabilità di una data soluzione, può essere utile ottenere più soluzioni diverse con casi disposti in ordini casuali diversi. Se i file sono particolarmente grandi, è possibile eseguire più analisi con un campione di casi disposti in più ordini casuali.

Creazione di una rete Funzione a base radiale

Dai menu, scegliere:

Analizza > Neural Networks > Funzione a base radiale...

Figura 3-1
Funzione a base radiale: scheda Variabili



- ▶ Selezionare almeno una variabile dipendente.
- ▶ Selezionare almeno un fattore o una covarianza.

Opzionalmente, nella scheda Variabili è possibile cambiare il metodo per la modifica della scala delle covariate. Le scelte sono:

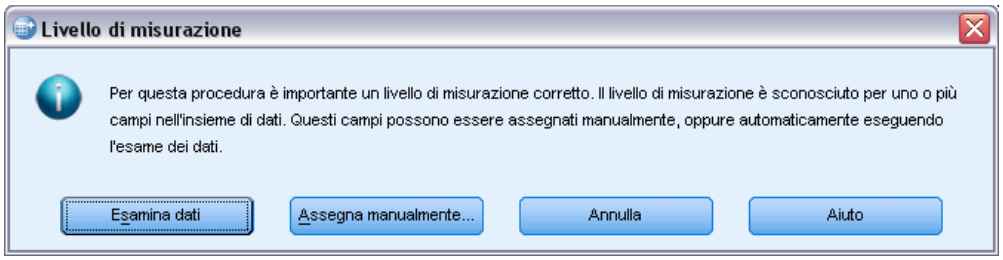
- **Standardizzati.** Sottrarre la media e dividere per la deviazione standard $(x-\text{media})/s$.
- **Normalizzati.** Sottratte il minimo e dividere per l'intervallo, $(x-\text{min})/(\text{max}-\text{min})$. Valori normalizzati compresi tra 0 e 1.

- **Normalizzati corretti.** La versione corretta della sottrazione del minimo e della divisione per l'intervallo [2*(*x*−min)/(max−min)]−1. I valori normalizzati corretti rientrano nell'intervallo compreso tra −1 e 1.
- **Nessuna.** Non viene eseguita la modifica delle scala delle covariate.

Campi con livello di misurazione sconosciuto

L'avviso Livello di misurazione viene visualizzato quando il livello di misurazione di una o più variabili (campi) dell'insieme di dati è sconosciuto. Poiché influisce sul calcolo dei risultati di questa procedura, il livello di misurazione deve essere definito per tutte le variabili.

Figura 3-2
Avviso Livello di misurazione

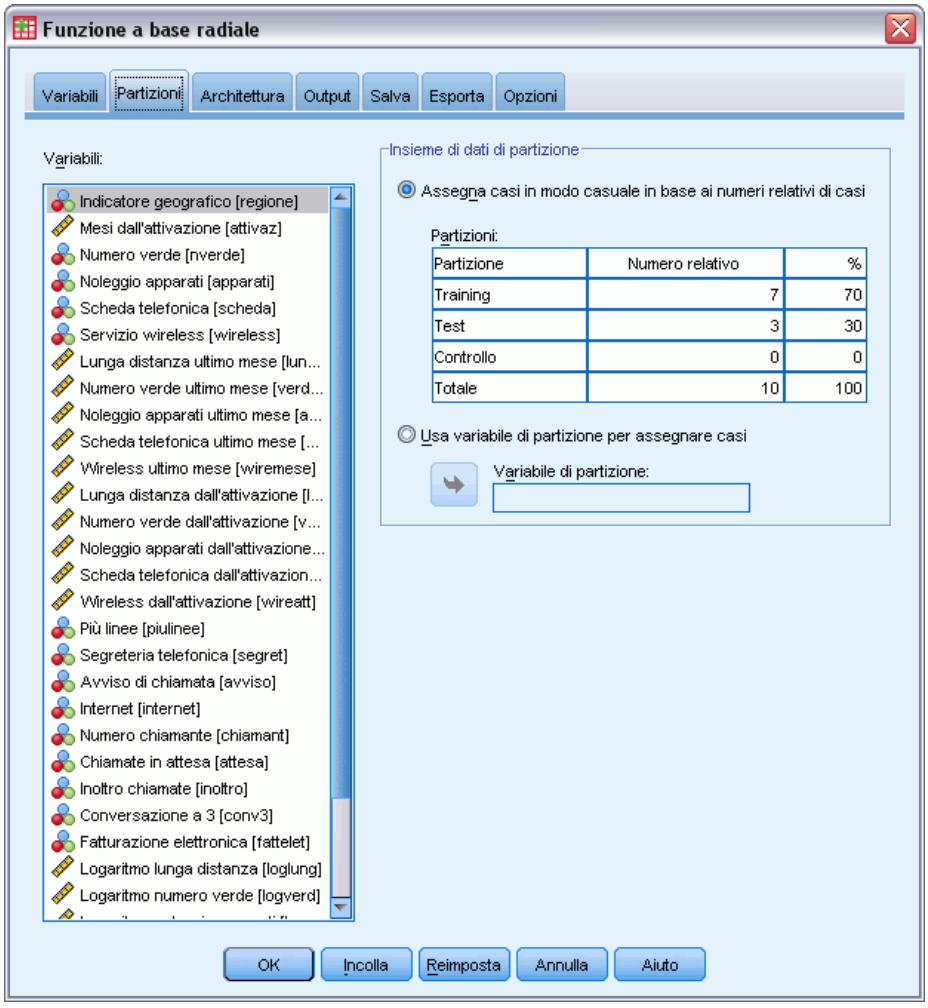


- **Esamina dati.** Legge i dati dell'insieme di dati attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con insiemi di dati di grandi dimensioni, questa operazione può richiedere del tempo.
- **Assegna manualmente.** Apre una finestra di dialogo che elenca tutti i campi con livello di misurazione sconosciuto, mediante la quale è possibile assegnare un livello di misurazione a questi campi. Il livello di misurazione si può assegnare anche nella Visualizzazione variabili dell'Editor dei dati.

Dal momento che il livello di misurazione è importante per questa procedura, è possibile accedere alla finestra di dialogo per la sua esecuzione solo quando per tutti i campi è stato definito un livello di misurazione.

Partizioni

Figura 3-3
Funzione a base radiale: Scheda Partizioni



Insieme di dati di partizione. Questo gruppo specifica il metodo di partizionamento dell'insieme di dati attivo in campioni di addestramento, test e controllo. Il **campione di addestramento** include i record di dati utilizzati per formare una rete neurale; una percentuale di casi nell'insieme di dati deve essere assegnata al campione di addestramento per ottenere un modello. Il **campione di verifica** è un insieme indipendente di record di dati utilizzato per tenere traccia degli errori durante l'addestramento per evitare un eccesso di addestramento. Si consiglia di creare un campione di addestramento; l'addestramento della rete sarà solitamente più efficace se il campione di verifica è più piccolo del campione di addestramento. Il **campione di controllo** è un altro insieme indipendente di record di dati utilizzato per valutare la rete neurale finale; l'errore per il campione

di controllo fornisce una stima "attendibile" della capacità predittiva del modello poiché i casi di controllo non sono stati utilizzati per generare il modello.

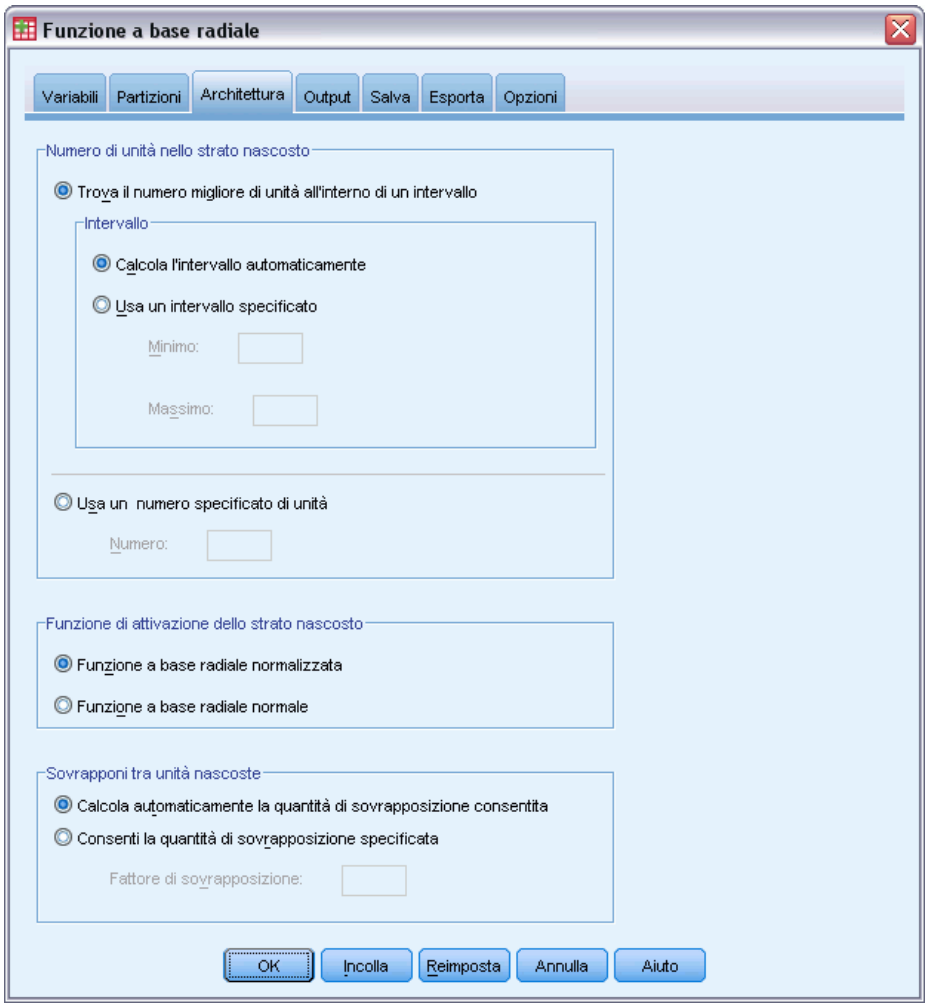
- **Assegna casi in modo casuale in base ai numeri relativi di casi.** Specificare il numero relativo (rapporto) di casi assegnati casualmente per ogni campione (addestramento, verifica e controllo). La colonna % indica la percentuale di casi che verranno assegnati a ogni campione in base ai numeri relativi specificati.

  Ad esempio, specificare 7, 3, 0 come numeri relativi per i campioni di addestramento, test e controllo, corrisponde a 70%, 30% e 0%. Specificare 2, 1, 1 come numeri relativi, corrisponde a 50%, 25% e 25%; 1, 1, 1 corrisponde a dividere l'insieme di dati in tre parti uguali di addestramento, test e controllo.

- **Usa variabile di partizione per assegnare casi.** Specificare una variabile numerica che assegni ogni caso nell'insieme di dati attivo al campione di addestramento, test e controllo. I casi con valore positivo nella variabile vengono assegnati al campione di addestramento, i casi con valore 0 al campione di verifica e i casi con un valore negativo al campione di controllo. I casi con un valore di sistema mancante vengono esclusi dall'analisi. I valori mancanti definiti dall'utente per la variabile di partizione sono sempre considerati validi.

Architettura

Figura 3-4
Funzione a base radiale: Scheda Architettura



La scheda Architettura viene utilizzata per specificare la struttura della rete. La procedura crea una rete neurale con uno strato nascosto della "funzione a base radiale"; in generale, non sarà necessario cambiare tali impostazioni.

Numero di unità nello strato nascosto. Esistono tre metodi per scegliere il numero di unità nascoste.

1. **Trova il numero migliore di unità all'interno di un intervallo calcolato automaticamente.** La procedura calcola automaticamente i valori minimo e massimo dell'intervallo e trova il numero migliore di unità nascoste nell'intervallo.

   Se è definito un campione di verifica, la procedura utilizza il criterio dei dati di test: Il numero migliore di unità nascoste corrisponde a quello che restituisce il numero minimo di errori nei dati di test. Se non è definito un campione di verifica, la procedura utilizza il criterio di informazione

bayesiano (BIC): Il numero migliore di unità nascoste corrisponde a quello che restituisce il numero minimo di criteri informativi di Bayes in base ai dati di test.

2. **Trova il numero migliore di unità all'interno di un intervallo specificato.** È possibile specificare un intervallo e la procedura troverà il "miglior" numero di unità nascoste in tale intervallo. Come detto in precedenza, il numero migliore di unità nascoste dall'intervallo viene determinato utilizzando il criterio dei dati di verifica o il BIC.

3. **Usa un numero specificato di unità.** In alternativa all'utilizzo di un intervallo, è possibile specificare direttamente uno specifico numero di unità.

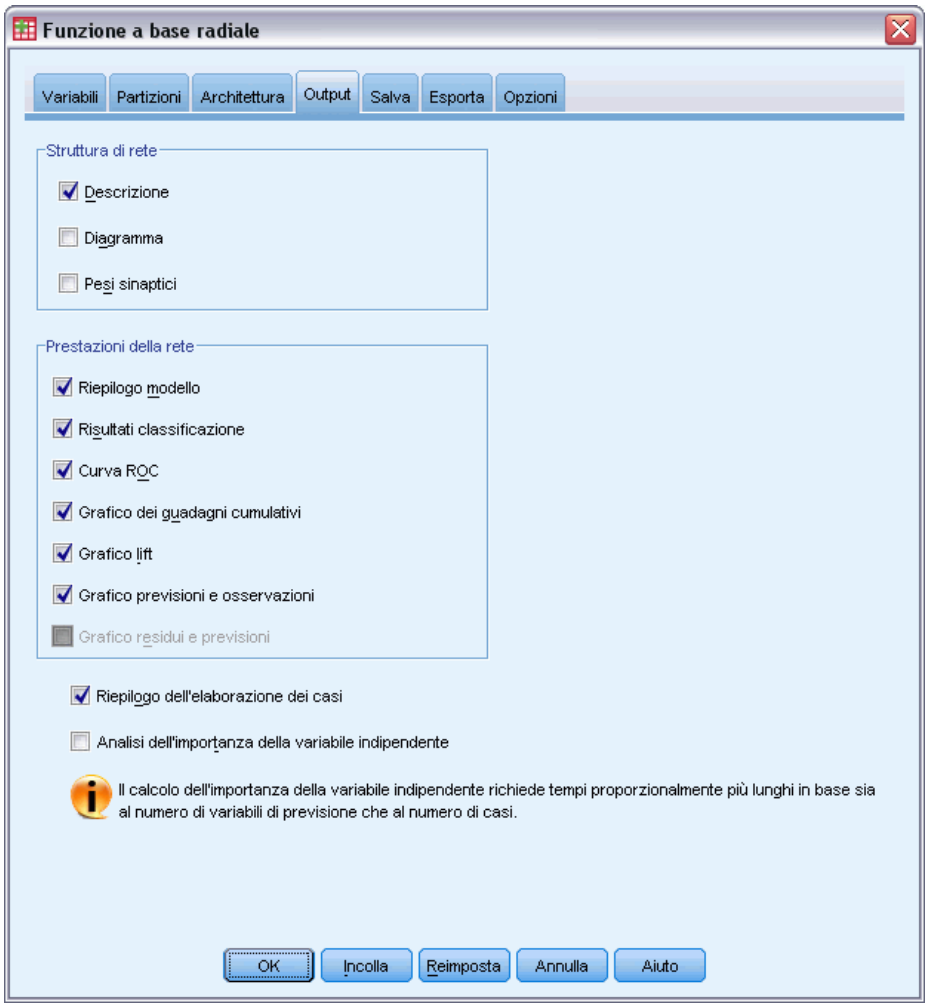
Funzione di attivazione dello strato nascosto. La funzione di attivazione dello strato nascosto è la funzione a base radiale, che "collega" le unità in uno strato ai valori delle unità nello strato successivo. Per lo strato di output, la funzione di attivazione è la funzione di identità; quindi le unità di output sono semplicemente somme ponderate delle unità nascoste.

- **Funzione a base radiale normalizzata.** Utilizza la funzione di attivazione Softmax e pertanto le attivazioni di tutte le unità nascoste sono normalizzate per la somma pari a 1.
- **Funzione a base radiale normale.** Utilizza la funzione di attivazione esponenziale e pertanto l'attivazione dell'unità nascosta è un "disturbo" gaussiano come una funzione di input.

Sovrapponi tra unità nascoste. Il fattore di sovrapposizione è un moltiplicatore applicato all'ampiezza delle funzioni a base radiale. Il valore calcolato automaticamente del fattore di sovrapposizione è $1+0,1d$, dove d è il numero di unità di input (la somma del numero di categorie su tutti i fattori e il numero di covariate).

Output

Figura 3-5
Funzione a base radiale: scheda Output



Struttura di rete. Visualizza informazioni di riepilogo sulla rete neurale.

- **Descrizione.** Visualizza informazioni sulla rete neurale, incluse le variabili dipendenti, numero di unità di input e output, numero di unità e di strati nascosti e funzioni di attivazione.
- **Diagramma.** Visualizza il diagramma di rete come un grafico non modificabile. Tenere presente che con l'aumento del numero di livelli dei fattori e di covariate, il diagramma diventa più difficile da interpretare.
- **Pesi sinaptici.** Visualizza le stime dei coefficienti che mostrano la relazione tra le unità in un determinato strato e le unità nello strato seguente. I pesi sinaptici si basano sul campione di addestramento se l'insieme di dati attivo è diviso in dati di addestramento, test e controllo. Tenere presente che il numero di pesi sinaptici può essere piuttosto grande e tali pesi non vengono solitamente utilizzati per interpretare i risultati della rete.

Prestazioni della rete. Visualizza i risultati utilizzati per determinare se il modello è "valido". *Nota*: i grafici in questo gruppo si basano sui campioni di addestramento e verifica oppure solo sul campione di addestramento se non è presente un campione di verifica.

- **Riepilogo del modello.** Visualizza un riepilogo dei risultati della rete neurale per partizione e valore globale, incluso l'errore, l'errore relativo o la percentuale di previsioni errate e il tempo di addestramento.

  L'errore è l'errore della somma dei quadrati. Inoltre, gli errori relativi o le percentuali di previsioni errate vengono visualizzati a seconda dei livelli di misurazione delle variabili dipendenti. Se una variabile dipendente ha un livello di misurazione di scala, viene visualizzato l'errore relativo complessivo medio (relativo al modello di media). Se tutte le variabili dipendenti sono categoriali, viene visualizzata la percentuale media delle previsioni non corrette. Vengono inoltre visualizzati gli errori relativi o le percentuali delle previsioni non corrette per le singole variabili dipendenti.

- **Risultati della classificazione (Analisi discriminante).** Visualizza una tabella di classificazione per ogni variabile dipendente. Ogni tabella fornisce il numero di casi classificati correttamente e non correttamente per ogni categoria di variabile dipendente. Viene inoltre indicata la percentuale dei casi totali classificati correttamente.

- **Curva ROC.** Visualizza una curva ROC (Receiver Operating Characteristic) per ogni variabile dipendente categoriale. Visualizza inoltre una tabella che fornisce l'area sotto ogni curva. Per una specifica variabile dipendente, il grafico ROC visualizza una curva per ogni categoria. Se la variabile dipendente ha due categorie, ogni curva considera la categoria in questione come stato positivo rispetto alle altre categorie. Se la variabile dipendente ha più di due categorie, ogni curva considera la categoria in questione come stato positivo rispetto a tutte le altre categorie.

- **Grafico dei guadagni cumulativi.** Visualizza un grafico dei guadagni cumulativi per ogni variabile dipendente categoriale. La vista di una curva per ogni categoria di variabile dipendente è uguale alle curve ROC.

- **Grafico lift.** Visualizza un grafico lift per ogni variabile dipendente categoriale. La vista di una curva per ogni categoria di variabile dipendente è uguale alle curve ROC.

- **Grafico previsioni e osservazioni.** Visualizza un grafico dei valori delle previsioni e osservazioni per ogni variabile dipendente. Per le variabili dipendenti categoriali, vengono visualizzati grafici a scatole raggruppati delle pseudo-probabilità previste per ogni categoria di risposta, con la categoria di risposta osservata come variabile di raggruppamento. Per le variabili dipendenti di scala, viene visualizzato un grafico a dispersione.

- **Grafico residui e previsioni.** Visualizza un grafico dei residui e delle previsioni per ogni variabile dipendente di scala. Tra i valori residui e attesi non dovrebbero essere presenti modelli visibili. Questo grafico viene generato solo per le variabili dipendenti di scala.

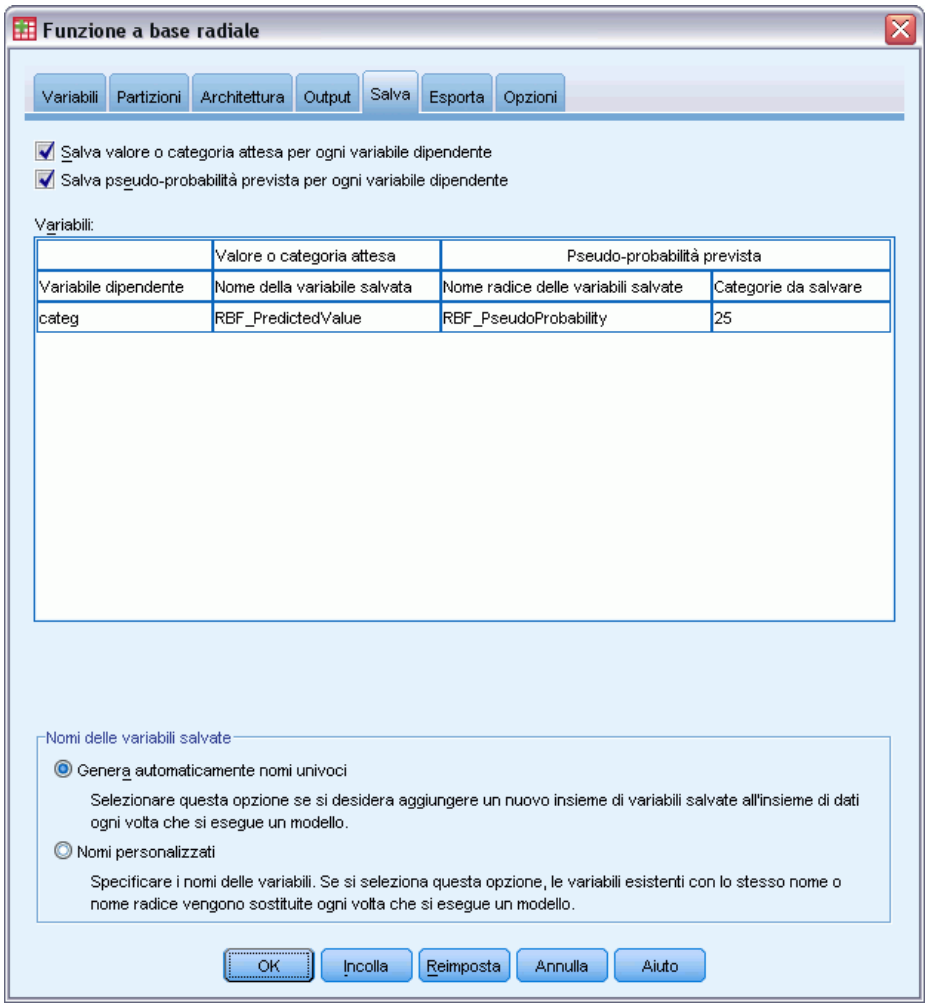
Riepilogo dell'elaborazione dei casi. Visualizza la tabella di riepilogo di elaborazione dei casi, che riepiloga il numero di casi inclusi ed esclusi dall'analisi, in totale e per campioni di addestramento, test e controllo.

Analisi dell'importanza della variabile indipendente. Esegue un'analisi della sensibilità, che calcola l'importanza di ogni predittore nel processo di determinazione della rete neurale. L'analisi si basa sui campioni di addestramento e verifica oppure solo sul campione di addestramento se non è

presente un campione di verifica. Questo crea una tabella e un grafico che mostrano l'importanza e l'importanza normalizzata di ogni predittore. Tenere presente che l'analisi della sensibilità è impegnativa a livello di calcolo e richiede parecchio tempo se sono presenti molti predittori o casi.

Salva

Figura 3-6
Funzione a base radiale: scheda Salva



La scheda Salva viene utilizzata per salvare le previsioni come variabili nell'insieme di dati.

- **Salva valore atteso o categoria per ogni variabile dipendente.** Vengono salvati il valore atteso per le variabili dipendenti di scala e la categoria prevista per le variabili dipendenti categoriali.
- **Salva la pseudo-probabilità prevista per ogni variabile dipendente.** Salva le pseudo-probabilità previste per le variabili dipendenti categoriali. Una variabile separata viene salvata per ognuna delle prime *n* categorie, dove *n* viene specificato nella colonna *Categorie da salvare*.

Nomi delle variabili salvate. La generazione automatica del nome assicura il mantenimento di tutto il lavoro. I nomi personalizzati consentono di eliminare o sostituire i risultati di precedenti esecuzioni senza dover prima eliminare le variabili salvate nell'Editor dei dati.

Probabilità e pseudo-probabilità

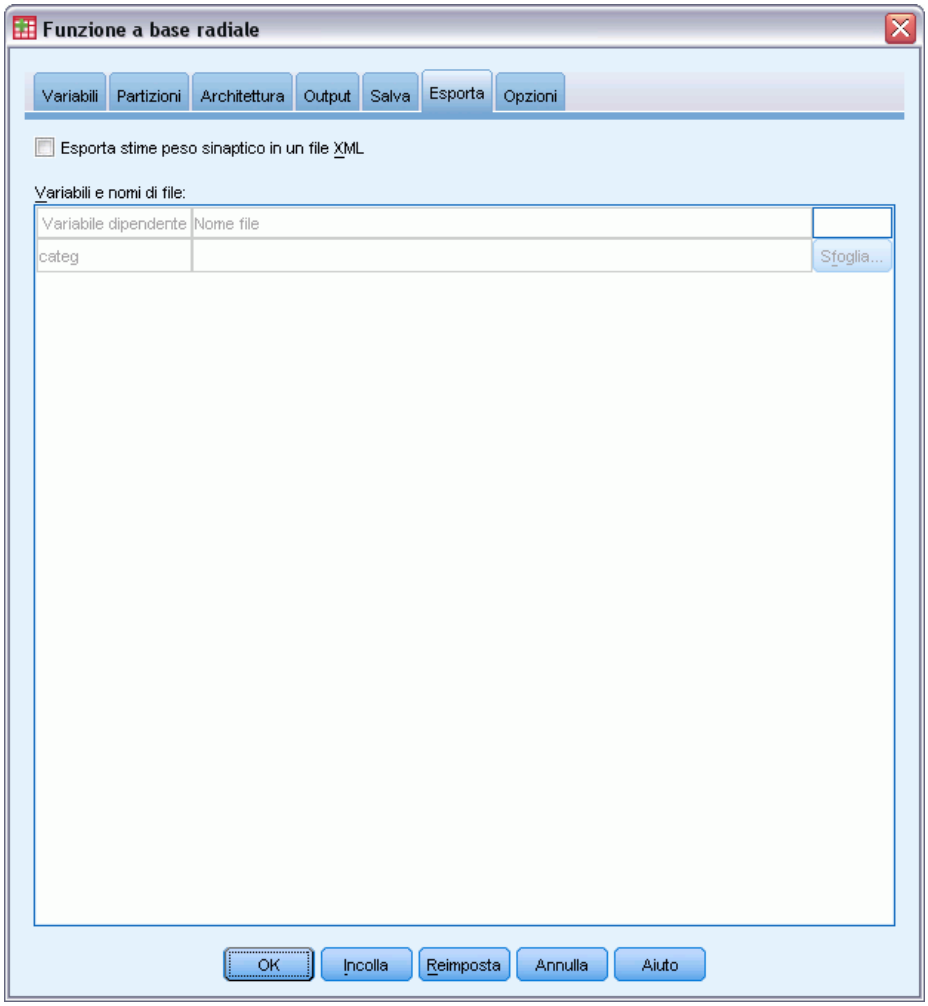
Le pseudo-probabilità previste non possono essere interpretate come probabilità poiché la procedura Funzione a base radiale utilizza l'errore della somma dei quadrati e la funzione di attivazione dell'identità per lo strato di output. La procedura salva queste pseudo-probabilità previste anche se nessuna è minore di 0 o maggiore di 1 o la somma di una specifica variabile dipendente non è 1.

I grafici ROC, dei guadagni cumulativi e lift (vedere Output a pag. 32) vengono creati in base alle pseudo-probabilità. Nel caso in cui nessuna pseudo-probabilità sia minore di 0 o maggiore di 1 o la somma di una specifica variabile non sia 1, queste vengono riscalate per essere comprese tra 0 e 1 e in modo che la somma dia come risultato 1. Le pseudo-probabilità vengono riscalate eseguendo la divisione per la relativa somma. Ad esempio, se un caso ha pseudo-probabilità previste di 0,50, 0,60 e 0,40 per una variabile dipendente a tre categorie, ogni pseudo-probabilità viene divisa per la somma 1,50 per ottenere 0,33, 0,40 e 0.27.

Se nessuna delle pseudo-probabilità è negativa, il valore assoluto del valore più basso viene aggiunto a tutte le pseudo-probabilità prima della modifica della scala. Ad esempio, se le pseudo-probabilità sono –0,30, ,50 e 1,30, aggiungere 0,30 a ogni valore per ottenere 0,00, 0,80 e 1,60. Quindi, dividere ogni nuovo valore per la somma 2,40 per ottenere 0,00, 0,33 e 0,67.

Esporta

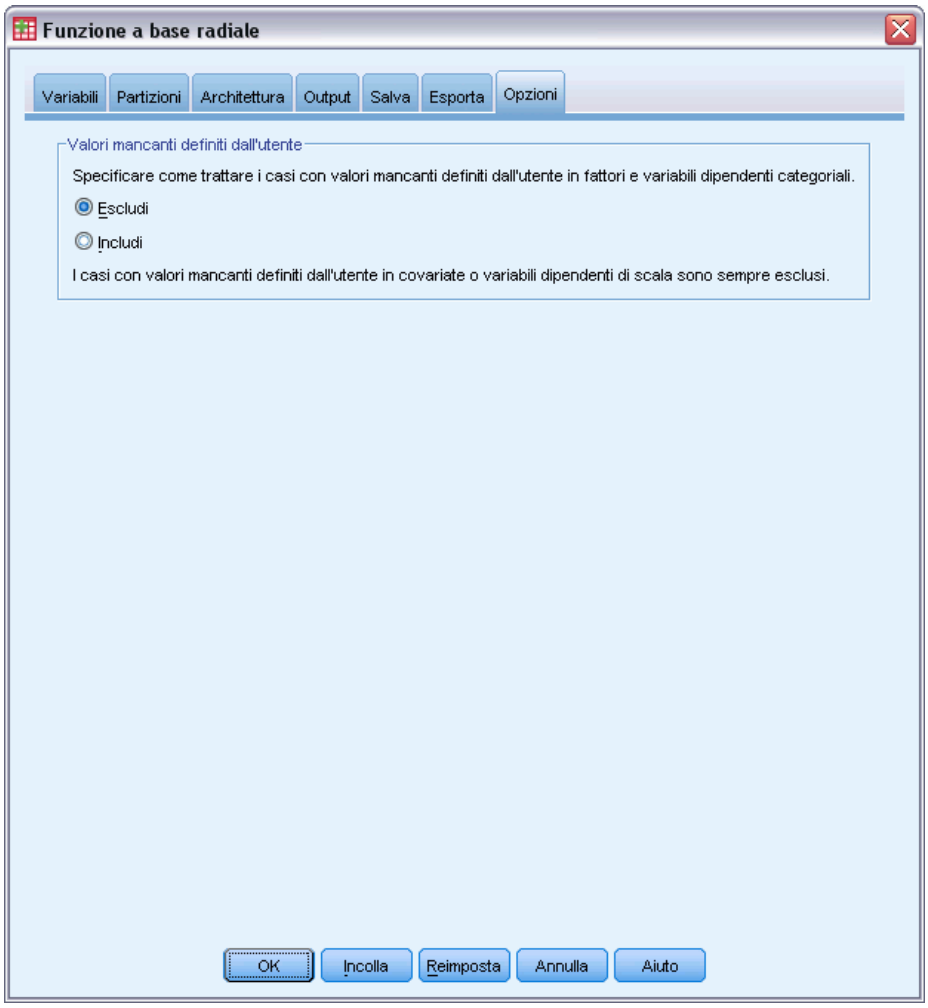
Figura 3-7
Funzione a base radiale: scheda Esporta



La scheda Esporta viene utilizzata per salvare le stime del peso sinaptico per ogni variabile dipendente in un file XML (PMML). È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio. Questa opzione non è disponibile se sono stati definiti file suddivisi.

Opzioni

Figura 3-8
Funzione a base radiale: Scheda Opzioni



Valori mancanti definiti dall'utente. Affinché un caso possa essere incluso nell'analisi, è necessario che i fattori abbiano valori validi. Questi controlli consentono di decidere se i valori mancanti definiti dall'utente devono essere considerati validi tra fattori e variabili dipendenti categoriali.

Parte II: Esempi

Capitolo
4

Perceptron a più strati

La procedura Perceptron a più strati (MLP) genera un modello predittivo per una o più variabili dipendenti (di destinazione) basato sui valori delle variabili predittori.

Utilizzo di Perceptron a più strati per la valutazione del rischio di credito

Un funzionario mutui presso una banca deve essere in grado di identificare le caratteristiche indicative delle persone che tendenzialmente saranno inadempimenti per quanto riguarda il rimborso dei prestiti e utilizzare tali caratteristiche per identificare i rischi di credito positivi e negativi.

Si supponga che le informazioni su 850 clienti già acquisiti e potenziali siano contenute nel file *bankloan.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 89. I primi 700 casi riguardano clienti a cui in precedenza sono stati concessi prestiti. Utilizzare un campione casuale di questi 700 clienti per creare una procedura Perceptron a più strati, tenendo da parte i clienti restanti per convalidare l'analisi. Utilizzare quindi il modello per classificare i 150 potenziali clienti come rischi di credito positivi o negativi.

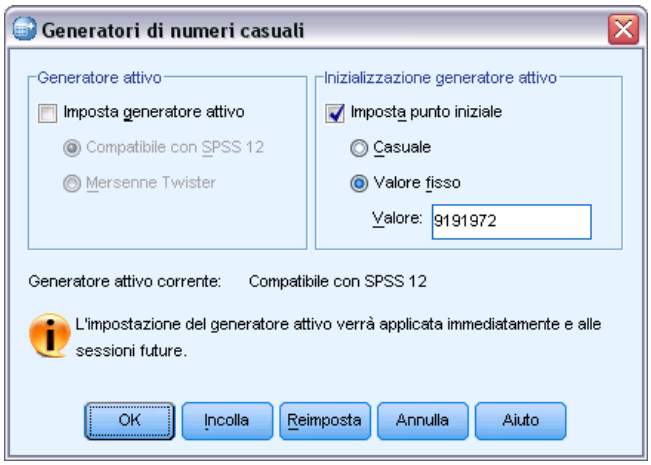
Inoltre, in precedenza il funzionario mutui ha analizzato i dati utilizzando la regressione logistica (nell'opzione Regressione) e si interroga sulla modalità di confronto della procedura Perceptron a più strati come strumento di classificazione.

Preparazione dei dati per l'analisi

L'impostazione del seme casuale consente di replicare esattamente l'analisi.

- Per impostare il seme casuale, dai menu scegliere:
  Trasforma > Generatori numeri casuali...

Figura 4-1
Finestra di dialogo Generatori di numeri casuali



- Selezionare Imposta punto inziale.
- Selezionare Valore fisso e digitare il valore 9191972.
- Fare clic su OK.

Nella precedente analisi di regressione logistica, il 70% dei clienti precedenti è stato assegnato al campione di training e il 30% a un campione di controllo. Per ricreare esattamente i campioni utilizzati in tali analisi, sarà necessaria una variabile di partizione.

- Per creare una variabile di questo tipo, dai menu scegliere:
  Trasforma > Calcola variabile...

Figura 4-2
Finestra di dialogo Calcola variabile

- Digitare partizione nella casella di testo Variabile di destinazione.
- Digitare 2*rv.bernoulli(0.7)-1 nella casella di testo Espressione numerica.

  In questo modo i valori di *partizione* verranno impostati come variate di **Bernoulli** generate in modo casuale con il parametro di probabilità 0,7 modificato in modo che assuma i valori 1 o −1, anziché 1 o 0. Tenere presente che i casi con valori positivi nella variabile di partizione vengono assegnati al campione di addestramento, i casi con valori negativi vengono assegnati al campione di controllo e i casi con valori pari a 0 vengono assegnati al campione di verifica. Al momento non verrà specificato un campione di verifica.

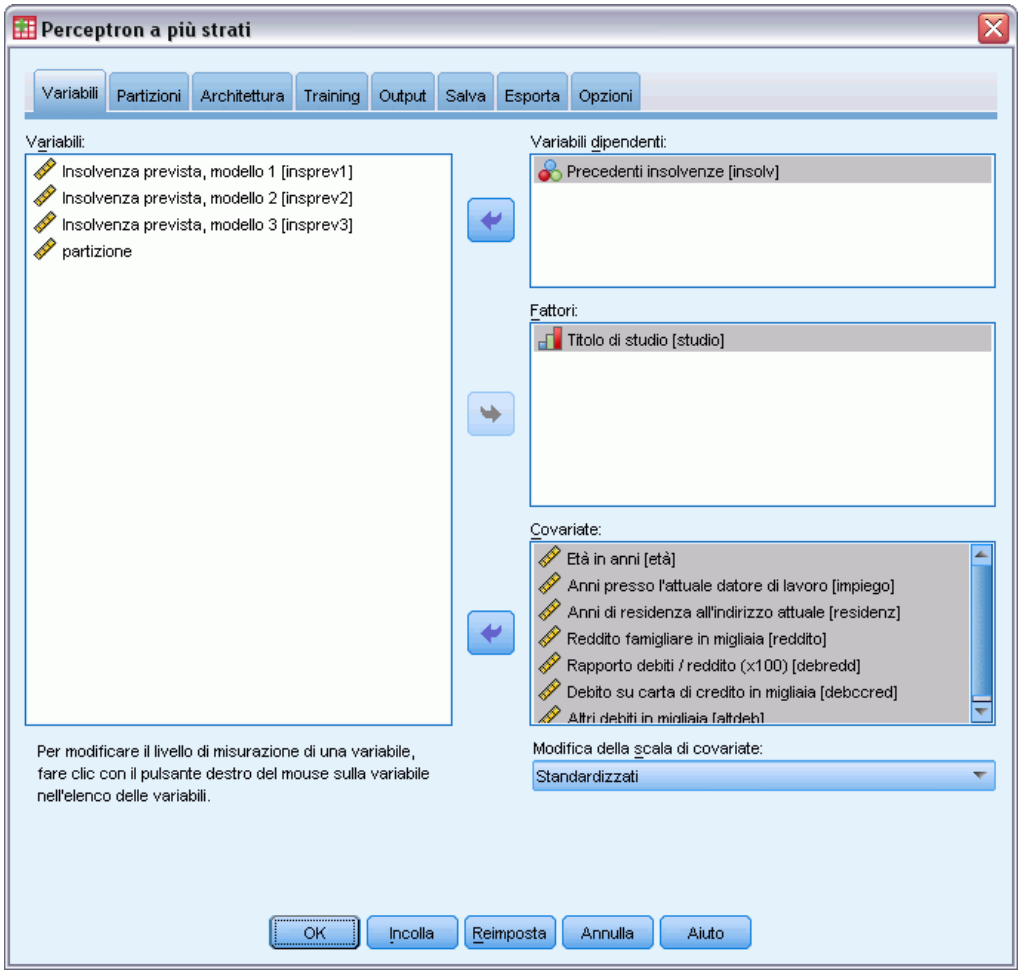
- Fare clic su OK nella finestra di dialogo Calcola variabile.

  Per il 70% circa dei clienti ai quali in precedenza è stato concesso un prestito verrà calcolato il valore *partizione* uguale a 1. La creazione del modello verrà basata su questi clienti. I clienti rimanenti ai quali era stato concesso un prestito avranno un valore della *partizione* uguale a −1 e verranno utilizzati per convalidare i risultati del modello.

Esecuzione dell'analisi

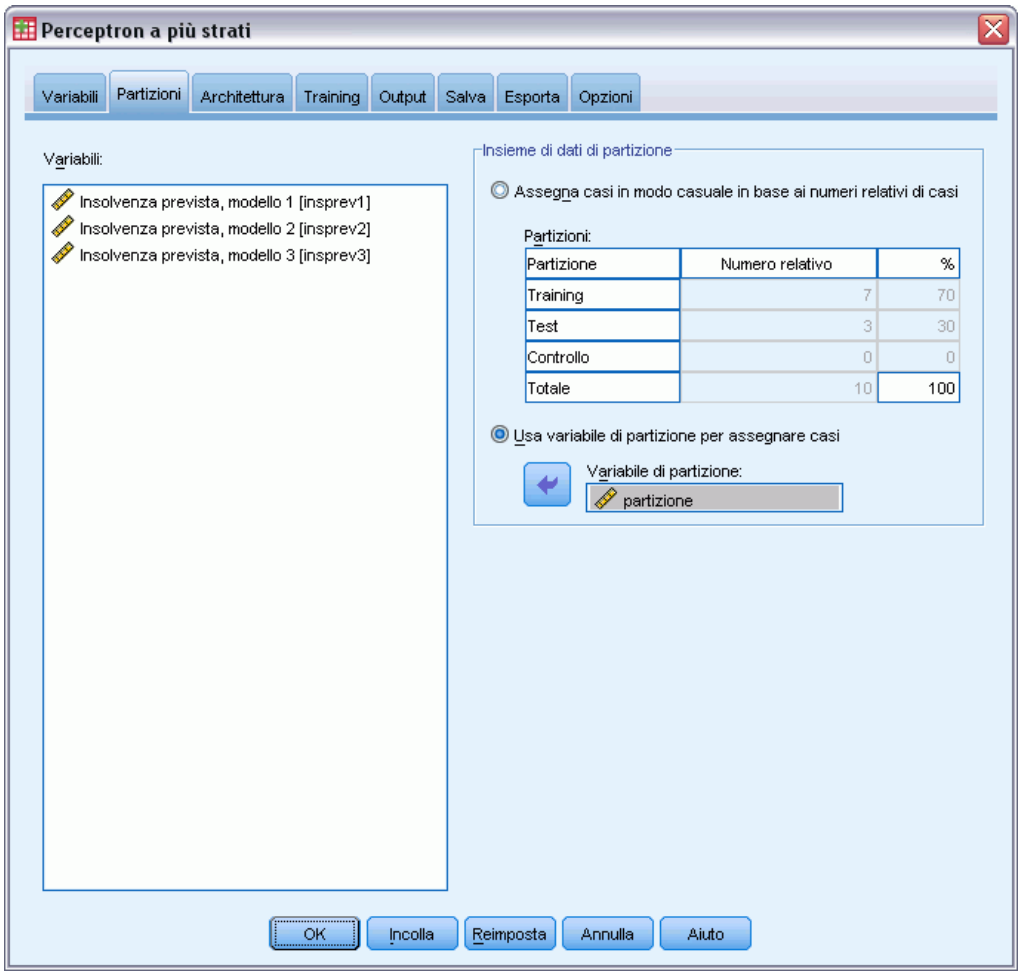
- Per eseguire un'analisi Perceptron a più strati, dai menu scegliere:
  Analizza > Neural Networks > Perceptron a più strati...

Figura 4-3
Perceptron a più strati: scheda Variabili



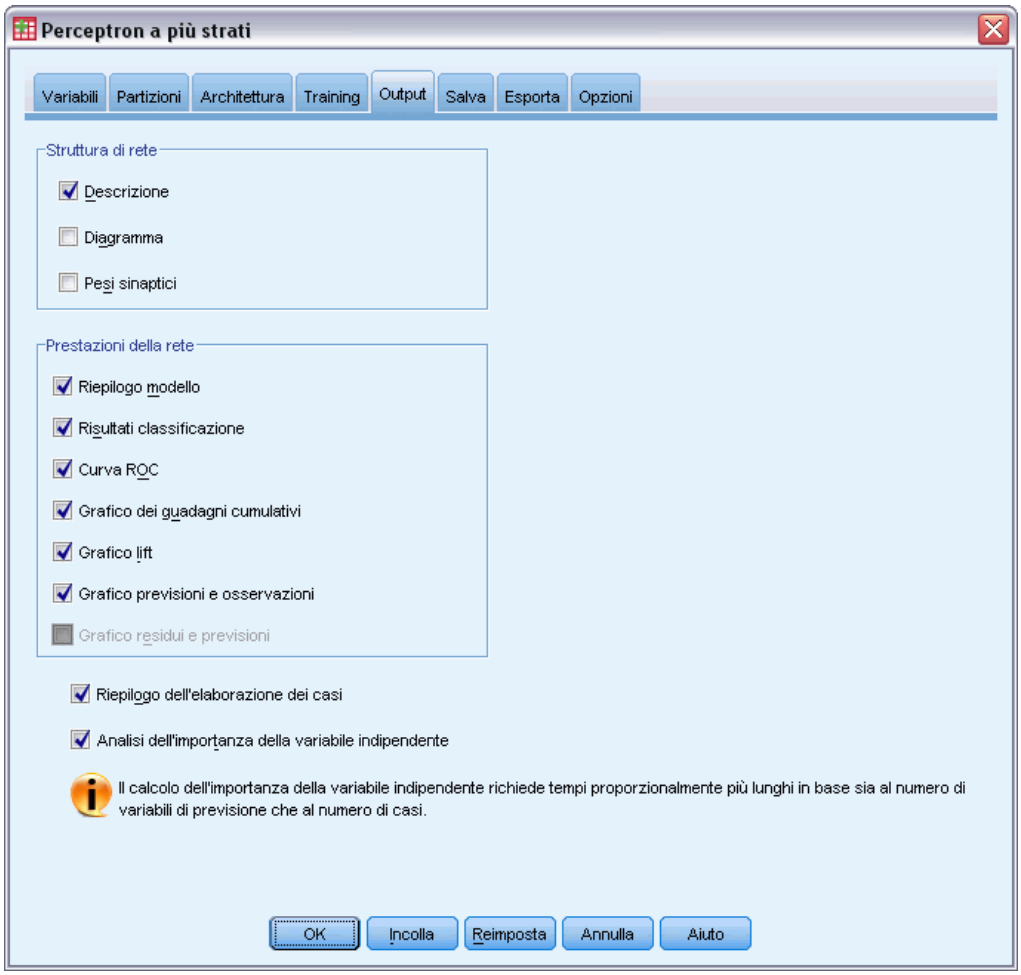
- Selezionare *Precedenti insolvenze [insolv]* come variabile dipendente.
- Selezionare*Titolo di studio [studio]* come fattore.
- Selezionare da *Età in anni [età]* ad *Altri indebitamenti in migliaia [altrideb]* come covariate.
- Fare clic sulla scheda Partizioni.

Figura 4-4
Perceptron a più strati: Scheda Partizioni



- Selezionare Usa variabile di partizione per assegnare casi.
- Selezionare *partizione* come variabile di partizionamento.
- Fare clic sulla scheda Output.

Figura 4-5
Perceptron a più strati: scheda Output



- Deselezionare Diagramma nel gruppo Struttura di rete.
- Selezionare Curva ROC, Grafico dei guadagni cumulativi, Grafico lift e Grafico previsioni e osservazioni nel gruppo Prestazioni della rete. Il grafico dei residui e delle previsioni non è disponibile poiché la variabile dipendente non è una scala.
- Selezionare Analisi dell'importanza della variabile indipendente.
- Fare clic su OK.

Riepilogo dei casi

Figura 4-6
Riepilogo dell'elaborazione dei casi

		N	Percentuale
Campione	Training	499	71,3%
	Test	201	28,7%
	Valida	700	100,0%
	Esclusa	150	
	Totale	850	

Il riepilogo di elaborazione dei casi mostra che 499 casi sono stati assegnati al campione di training e 201 al campione di controllo. I 150 casi esclusi dall'analisi sono i potenziali clienti.

Informazioni di rete

Figura 4-7
informazioni di rete

Strato di input	Fattori	1	Titolo di studio
	Covariates	1	Età in anni
		2	Anni presso l'attuale datore di lavoro
		3	Anni di residenza all'indirizzo attuale
		4	Reddito famigliare in migliaia
		5	Rapporto debiti / reddito (x100)
		6	Debito su carta di credito in migliaia
		7	Altri debiti in migliaia
	Numero di unità[a]		12
	Rescaling Method for Covariates		Standardizzata
Strato/i nascosto/i	Numero di strati nascosti		1
	Numero di unità nello strato nascosto 1[a]		4
	Funzione di attivazione		Tangente iperbolica
Strato di output	Dependent Variables	1	Precedenti insolvenze
	Numero di unità		2
	Funzione di attivazione		Softmax
	Funzione di errore		Entropia incrociata

a. Senza unità diagonale

La tabella delle informazioni sulla rete visualizza le informazioni sulla rete neurale ed è utile per garantire che le specifiche siano corrette. Si noti in particolare che:

- Il numero di unità nello strato di input corrisponde al numero di covariate più il numero totale di livelli di fattore; viene creata un'unità separata per ogni categoria di *Titolo di studio* e nessuna delle categorie viene considerata unità "ridondante" come avviene tipicamente in numerose procedure di creazione di modelli.

- Analogamente, viene creata un'unità di output separata per ogni categoria di *Precedenti insolvenze* per un totale di 2 unità nello strato di output.
- La selezione automatica dell'architettura ha scelto quattro unità nello strato nascosto.
- Tutte le altre informazioni sulla rete sono predefinite per la procedura.

Riepilogo del modello (Regressione output)

Figura 4-8
Riepilogo modello

Training	Errore di entropia incrociata	156,606
	Percentuale di previsioni errate	15,6%
	Regola di interruzione utilizzata	1 passi consecutivi senza diminuzione degli errori
	Tempo di training	00:00:01.015
Test	Percentuale di previsioni errate	25,4%

Variabile dipendente: Precedenti insolvenze

Il riepilogo del modello visualizza le informazioni sui risultati del training e l'applicazione della rete finale al campione di controllo.

- Viene visualizzato l'errore di entropia incrociata poiché lo strato di output utilizza la funzione di attivazione Softmax, che è la funzione di errore che la rete cerca di minimizzare durante il training.
- La percentuale di previsioni non corrette viene prelevata dalla tabella di classificazione e verrà discussa più avanti in questo argomento.
- L'algoritmo di stima viene interrotto poiché è stato raggiunto il numero massimo di epoche. Idealmente l'addestramento si interrompe perché è stata ottenuta la convergenza dell'errore. In questo modo si pongono questioni, ad esempio se si è verificato un errore durante l'addestramento e se è necessario valutare qualche aspetto durante le successive ispezioni dell'output.

Classification

Figura 4-9
Classification

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

Variabile dipendente: Precedenti insolvenze

La tabella di classificazione mostra i risultati pratici dell'utilizzo della rete. Per ogni caso, la risposta prevista è *Sì* se la pseudo-probabilità prevista dei casi è maggiore di 0,5. Per ogni campione:

- Le celle sulla diagonale della classificazione incrociata dei casi sono le previsioni corrette.
- Le celle fuori dalla diagonale della classificazione incrociata dei casi sono le previsioni non corrette.

Nei casi utilizzati per sviluppare il modello, 74 su 124 persone che precedentemente sono risultate inadempienti, vengono classificate correttamente. 347 dei 375 non inadempienti vengono classificati correttamente. In totale, l'84,4% dei casi di training vengono classificati correttamente, corrispondente al 15,6% dei casi non classificati correttamente indicati nella tabella di riepilogo del modello. Un modello migliore deve identificare correttamente una percentuale maggiore di casi.

Le classificazioni basate sui casi utilizzati per creare il modello tendono a essere troppo "ottimistiche" in quanto il tasso di classificazione è gonfiato. Il campione di controllo consente di convalidare il modello; in questo caso il 74,6% di questi casi sono stati classificati correttamente dal modello. Ciò indica che, complessivamente, il modello è corretto circa tre volte su quattro.

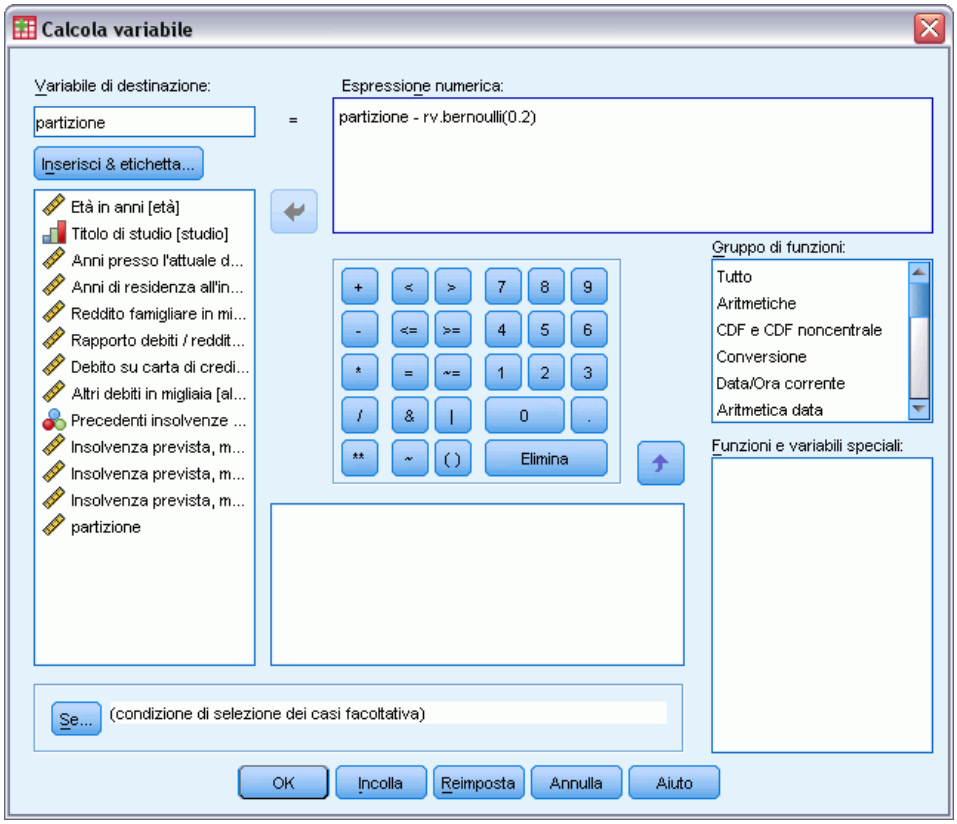
Correzione dell'eccesso di training

Sulla base dell'analisi di regressione logistica eseguita in precedenza, il funzionario mutui ricorda che i campioni di training e di controllo prevedevano correttamente una percentuale di casi simile, circa l'80%. Al contrario, la rete neurale registrava una percentuale superiore di casi corretti nel campione di addestramento, mentre il campione di controllo offriva una previsione notevolmente peggiore dei clienti effettivamente inadempienti (il 45,8% di casi corretti per il campione di controllo rispetto al 59,7% del campione di addestramento). In combinazione con la regola di interruzione riportata nella tabella di riepilogo del modello, si può supporre che si sia verificato un **eccesso di addestramento** della rete, ovvero vengono cercati i modelli spuri che appaiono nei dati di addestramento per variazione casuale.

La soluzione è relativamente semplice: specificare un campione di verifica per mantenere la rete entro i limiti corretti. È stata creata la variabile di partizione in modo che sia possibile ricreare i campioni di training e di controllo utilizzati nell'analisi della regressione logistica; tuttavia, la regressione logistica non implica il concetto di campione di "test". Prelevare una porzione del campione di training e riassegnarla a un campione di verifica.

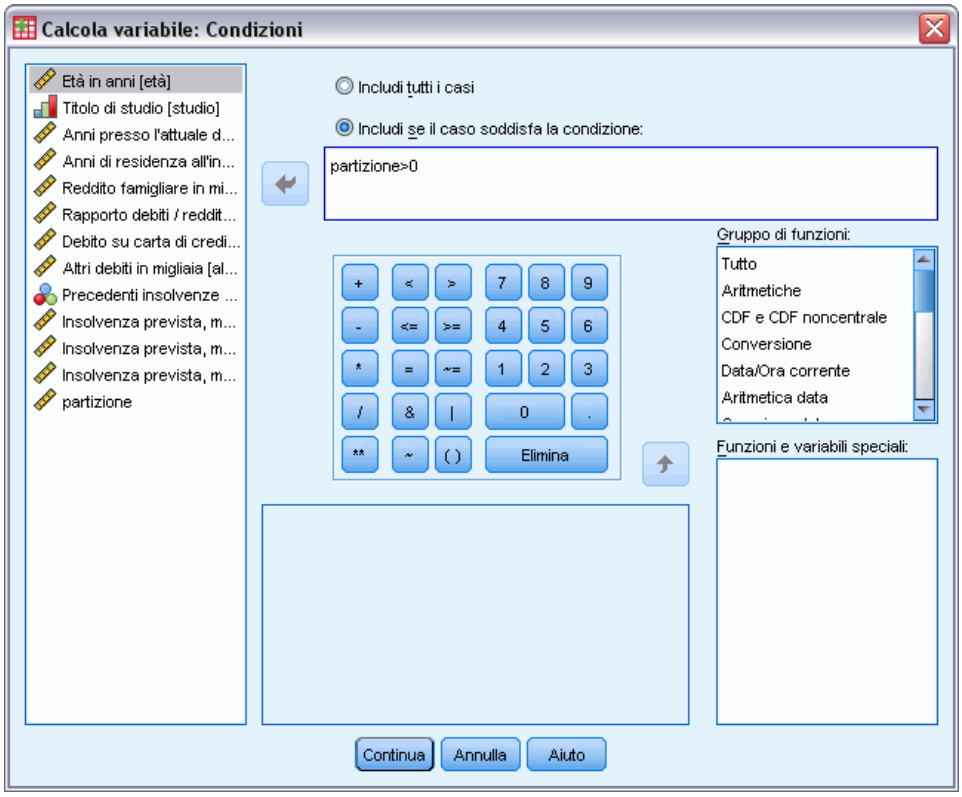
Creazione del campione di verifica

Figura 4-10
Finestra di dialogo Calcola variabile



- ▶ Richiamare la finestra di dialogo Calcola variabile.
- ▶ Digitare partizione - rv.bernoulli(0.2) nella casella di testo Espressione numerica.
- ▶ Fare clic su Se.

Figura 4-11
Calcola variabile: Finestra di dialogo Condizioni



- Selezionare Includi se il caso soddisfa la condizione.
- Digitare partizione>0 nella casella di testo.
- Fare clic su Continua.
- Fare clic su OK nella finestra di dialogo Calcola variabile.

In questo modo i valori della *partizione* che erano maggiori di 0 vengono reimpostati in modo che circa il 20% assuma il valore 0 e l'80% rimanga con il valore 1. Complessivamente, circa 100*(0,7*0,8)=56% dei clienti a cui in precedenza sono stati concessi prestiti verranno assegnati al campione di addestramento e il 14% al campione di verifica. L'assegnazione dei clienti che sono stati assegnati originariamente al campione di controllo rimane invariata.

Esecuzione dell'analisi

- Richiamare la finestra di dialogo Perceptron a più strati e fare clic sulla scheda Salva.
- Selezionare Salva pseudo-probabilità prevista per ogni variabile dipendente.
- Fare clic su OK.

Riepilogo dei casi

Figura 4-12
Riepilogo dei casi per il modello con il campione di verifica

		N	Percentuale
Campione	Training	398	56,9%
	Testing	101	14,4%
	Controllo	201	28,7%
Valida		700	100,0%
Esclusa		150	
Totale		850	

Dei 499 casi assegnati in origine al campione di training, 101 sono stati riassegnati al campione di verifica.

Informazioni di rete

Figura 4-13
informazioni di rete

Strato di input	Fattori	1	Titolo di studio
	Covariates	1	Età in anni
		2	Anni presso l'attuale datore di lavoro
		3	Anni di residenza all'indirizzo attuale
		4	Reddito famigliare in migliaia
		5	Rapporto debiti / reddito (x100)
		6	Debito su carta di credito in migliaia
		7	Altri debiti in migliaia
	Numero di unità[a]		12
	Rescaling Method for Covariates		Standardizzata
Strato/i nascosto/i	Numero di strati nascosti		1
	Numero di unità nello strato nascosto 1[a]		7
	Funzione di attivazione		Tangente iperbolica
Strato di output	Dependent Variables	1	Precedenti insolvenze
	Numero di unità		2
	Funzione di attivazione		Softmax
	Funzione di errore		Entropia incrociata

a. Senza unità diagonale

L'unica modifica apportata alla tabella delle informazioni sulla rete è che la selezione automatica dell'architettura ha definito sette unità nello strato nascosto.

Riepilogo del modello (Regressione output)

Figura 4-14
Riepilogo modello

Training	Errore di entropia incrociata	159,870
	Percentuale di previsioni errate	20,1%
	Regola di interruzione utilizzata	1 passi consecutivi senza diminuzione degli errori[a]
	Tempo di training	00:00:00.889
Test	Errore di entropia incrociata	40,068
	Percentuale di previsioni errate	17,8%
Controllo	Percentuale di previsioni errate	20,4%

Variabile dipendente: Precedenti insolvenze

a. I calcoli degli errori si basano sul campione di verifica.

Il riepilogo del modello rappresenta una coppia di segni positivi:

- La percentuale di previsioni non corrette è approssimativamente uguale tra i campioni di addestramento, di verifica e di controllo.
- L'algoritmo di stima si è interrotto poiché l'errore non è diminuito dopo un passo dell'algoritmo.

Questo aspetto conferma che potrebbe essersi verificato un eccesso di addestramento nel modello originale e che il problema è stato risolto aggiungendo un campione di verifica. Naturalmente le dimensioni del campione sono relativamente ridotte e non è opportuno interpretare ulteriormente la lettura dello spostamento di pochi punti percentuali.

Classification

Figura 4-15
Classification

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

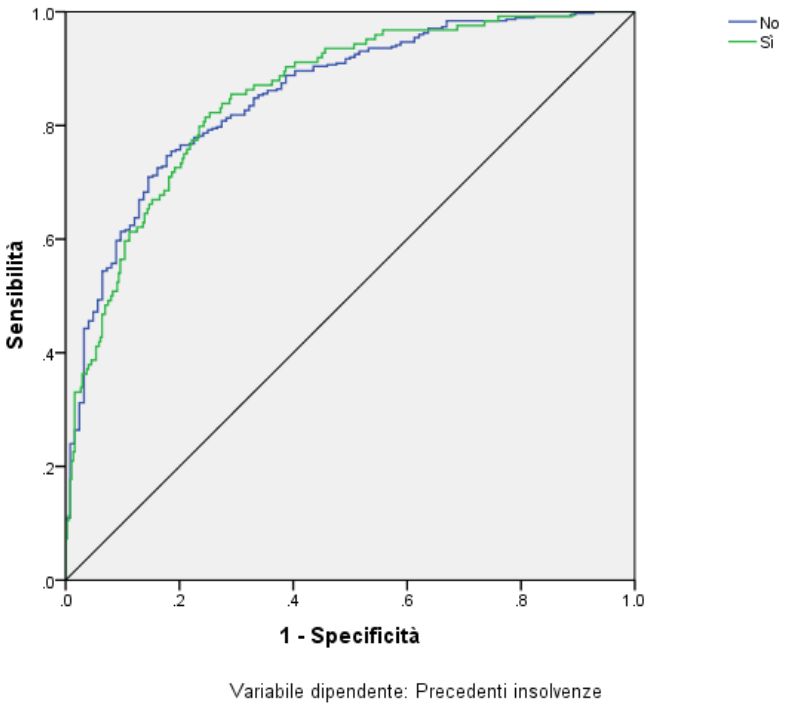
Variabile dipendente: Precedenti insolvenze

La tabella di classificazione mostra che, utilizzando 0,5 come riferimento della pseudo-probabilità per la classificazione, la rete ottiene risultati notevolmente migliori nella previsione dei non inadempienti rispetto a quella degli inadempienti. Tuttavia, il singolo valore di riferimento

consente una visione molto limitata della capacità di previsione della rete; pertanto non è necessariamente molto utile per il confronto di reti competitive. Esaminare piuttosto la curva ROC.

Curva ROC

Figura 4-16
Curva ROC



La curva ROC rappresenta visivamente la **sensibilità** e la **specificità** per tutti i possibili riferimenti in un singolo diagramma, consentendo una rappresentazione più chiara ed efficace rispetto a una serie di tabelle. Il grafico mostrato visualizza due curve, una per la categoria *No* e una per la categoria *Sì*. Poiché esistono solo due categorie, le curve sono simmetriche rispetto a una linea di 45° (non visualizzata) dall'angolo superiore sinistro del grafico all'angolo inferiore destro.

Tenere presente che questo grafico si basa sui campioni di addestramento e test. Per generare un grafico ROC per il campione di controllo, suddividere il file in base alla variabile di partizione ed eseguire la procedura relativa alla curva ROC per le pseudo-probabilità previste salvate.

Figura 4-17
Area sotto la curva

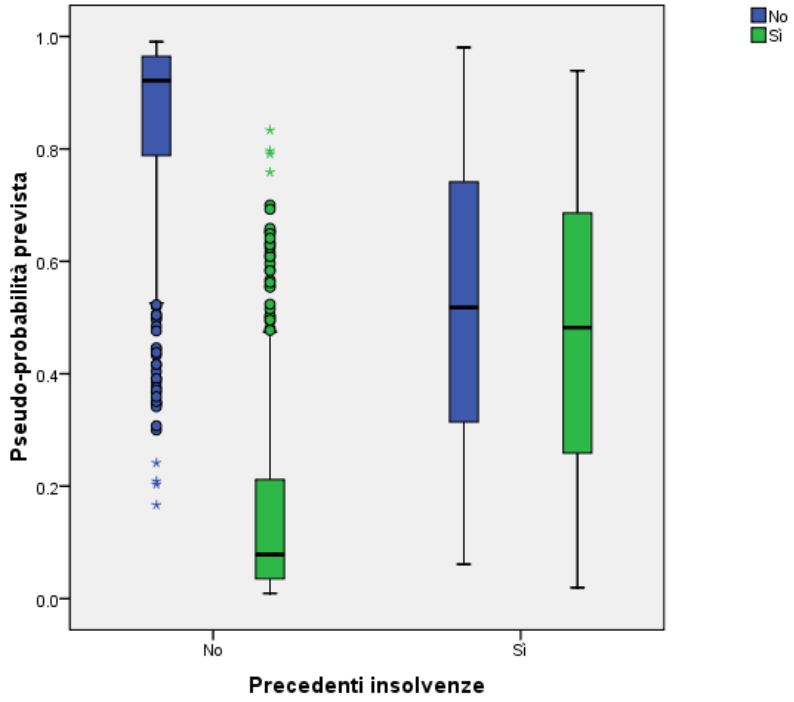
		Area
Precedenti insolvenze	No	,853
	Sì	,853

L'area sotto la curva è un riepilogo numerico della curva ROC e i valori nella tabella rappresentano, per ogni categoria, la probabilità che la pseudo-probabilità prevista si trovi in tale categoria è maggiore per un caso scelto in modo casuale in tale categoria rispetto a un caso scelto in modo casuale ma che non si trova in tale categoria. Ad esempio, per un cliente inadempiente selezionato in modo casuale e un cliente non inadempiente selezionato in modo casuale, esiste una probabilità pari a 0,853 che la pseudo-probabilità prevista dal modello sia superiore per il cliente inadempiente rispetto al cliente non inadempiente.

Mentre l'area sotto la curva è un utile riepilogo a una statistica dell'accuratezza della rete, è necessario essere in grado di scegliere un criterio specifico in base al quale vengano classificati i clienti. Il grafico delle previsioni e osservazioni consente una rappresentazione visiva di questo processo.

Grafico previsioni e osservazioni

Figura 4-18
grafico previsioni e osservazioni



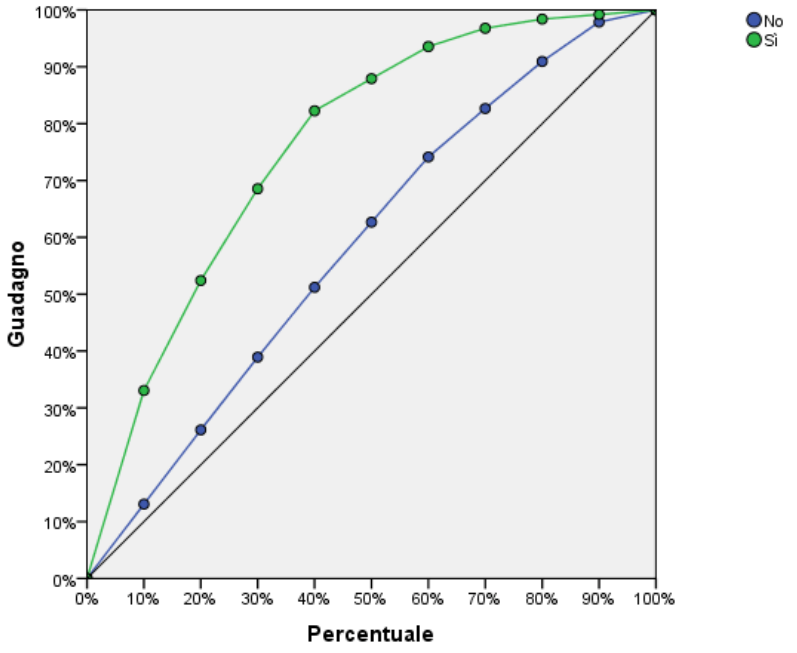
Per le variabili dipendenti categoriali, il Grafico previsioni e osservazioni visualizza grafici a scatole delle pseudo-probabilità previste dei campioni di verifica e di addestramento combinati. L'asse *X* corrisponde alle categorie di risposta osservate e la legenda corrisponde alle categorie previste.

- Il grafico a scatole all'estremità sinistra mostra, per i casi con la categoria osservata No, la pseudo-probabilità prevista della categoria No. La porzione del grafico a scatole sopra l'indicazione 0,5 sull'asse *Y* rappresenta le previsioni corrette indicate nella tabella di classificazione. La porzione sotto l'indicazione 0,5 rappresenta le previsioni non corrette. Tenere presente in base alla tabella di classificazione che le rete è molto efficace nella previsione dei casi con la categoria No utilizzando il valore di riferimento 0,5; pertanto solo una porzione del baffo e alcuni casi anomali non vengono classificati correttamente.
- Il successivo grafico a scatole all'estremità destra mostra, per i casi con la categoria osservata No, la pseudo-probabilità prevista della categoria Sì. Poiché esistono solo due categorie nella variabile di destinazione, i primi due grafici a scatole sono simmetrici rispetto alla linea orizzontale in corrispondenza di 0,5.
- Il terzo grafico a scatole mostra, per i casi con la categoria osservata Sì, la pseudo-probabilità prevista della categoria No. Questo e l'ultimo grafico a scatole sono simmetrici rispetto alla linea orizzontale in corrispondenza di 0,5.
- L'ultimo grafico a scatole mostra, per i casi con la categoria osservata Sì, la pseudo-probabilità prevista della categoria Sì. La porzione del grafico a scatole sopra l'indicazione 0,5 sull'asse *Y* rappresenta le previsioni corrette indicate nella tabella di classificazione. La porzione sotto l'indicazione 0,5 rappresenta le previsioni non corrette. Tenere presente in base alla tabella di classificazione che le rete prevede leggermente più della metà dei casi con la categoria Sì utilizzando il valore di riferimento 0,5; pertanto una buona parte della scatola non viene classificata correttamente.

Esaminando il grafico è possibile notare che la diminuzione del riferimento per la classificazione di un caso come Sì da 0,5 a circa 0,3, questo è il valore in cui si trovano sia la parte superiore della seconda scatola che la parte inferiore della quarta scatola, permette di individuare correttamente i possibili clienti inadempienti senza perderne numerosi di potenzialmente buoni. In altre parole, lo spostamento da 0,5 a 0,3 lungo la seconda scatola riclassifica non correttamente i relativamente pochi clienti non inadempienti lungo il baffo come clienti inadempienti previsti, mentre lungo la quarta scatola questo spostamento riclassifica correttamente i numerosi clienti inadempienti all'interno della scatola come clienti inadempienti previsti.

Grafici dei guadagni cumulativi e lift

Figura 4-19
Grafico dei guadagni cumulativi



Variabile dipendente: Precedenti insolvenze

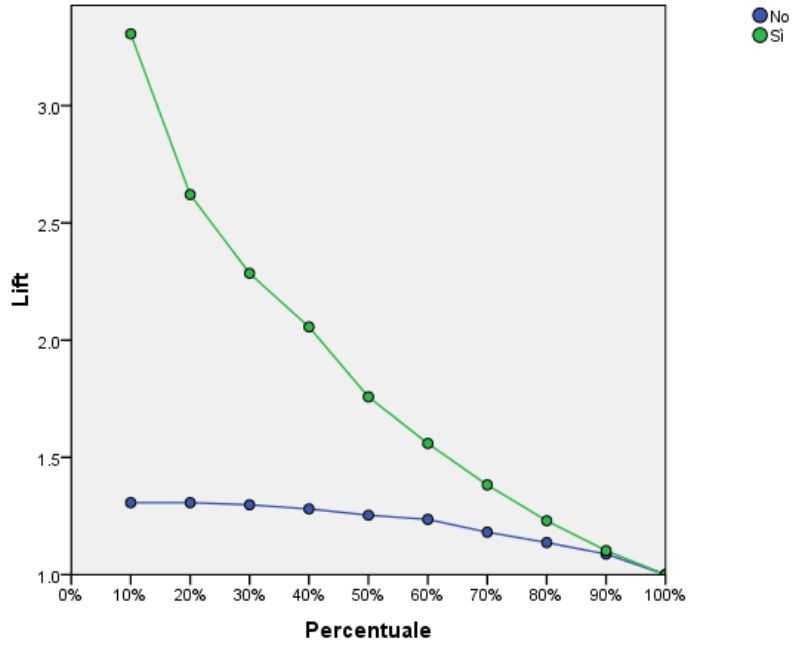
Il grafico dei guadagni cumulativi mostra la percentuale del numero totale dei casi in una determinata categoria ottenuta definendo come target una percentuale del numero totale dei casi. Ad esempio, il primo punto sulla curva per la categoria *Sì* si trova in corrispondenza di (10%, 30%), indicando che se si assegna un punteggio a un insieme di dati con la rete e si ordinano tutti i casi per la pseudo-probabilità prevista *Sì*, si prevede che il primo 10% contenga circa il 30% di tutti i casi che effettivamente fanno parte della categoria *Sì* (inadempienti). Analogamente, il primo 20% dovrebbe contenere circa il 50% degli inadempienti, il primo 30% dei casi dovrebbe contenere il 70% degli inadempienti e così via. Se si seleziona il 100% dell'insieme dei dati valutato, si ottengono tutti gli inadempienti nell'insieme di dati.

La linea diagonale è la curva "di base"; se si seleziona il 10% dei casi dall'insieme di dati nel punteggio in modo causale, si prevede di "guadagnare" circa il 10% di tutti i casi che effettivamente fanno parte della categoria Sì. Più in alto si trova la linea di base in cui si trova una curva, maggiore è il guadagno. È possibile utilizzare il grafico dei guadagni cumulativi per consentire di scegliere un riferimento della classificazione scegliendo una percentuale che corrisponda a un guadagno preferibile, quindi associando tale percentuale al valore di riferimento appropriato.

La definizione di guadagno "preferibile" dipende dal costo degli errori di Tipo I e di Tipo II. In altre parole, qual è il costo della classificazione di un soggetto inadempiente come non inadempiente (Tipo I)? Qual è il costo della classificazione di un soggetto non inadempiente come inadempiente (Tipo II)? Se l'interesse principale è l'inadempienza, è utile ridurre l'errore di Tipo I; nel grafico dei guadagni cumulativi, ciò potrebbe corrispondere alla mancata concessione di

mutui ai richiedenti nel primo 40% della probabilità pseudo-prevista di *Sì*, che cattura quasi il 90% dei possibili inadempienti ma rimuove quasi la metà del gruppo di richiedenti. Se la priorità è ampliare la base clienti, è utile ridurre l'errore di Tipo II. Nel grafico ciò potrebbe corrispondere a rifiutare il primo 10%, che cattura il 30% degli inadempienti e lascia intatta la maggior parte del gruppo di richiedenti. Normalmente, entrambi questi elementi hanno importanza, quindi diventa necessario scegliere una regola decisionale per la classificazione della clientela che offra una combinazione ottimale di sensibilità e specificità.

Figura 4-20
Grafico lift



Il grafico lift deriva dal grafico dei guadagni cumulativi; i valori sull'asse *Y* corrispondono al rapporto del guadagno cumulativo per ogni curva con la linea di base. Pertanto, il lift in corrispondenza del 10% per la categoria Sì è 30%/10% = 3,0. Consente di esaminare in modo differente le informazioni nel grafico dei guadagni cumulativi.

Nota: il grafico dei guadagni cumulativi e il grafico lift si basano sui campioni di verifica e di addestramento combinati.

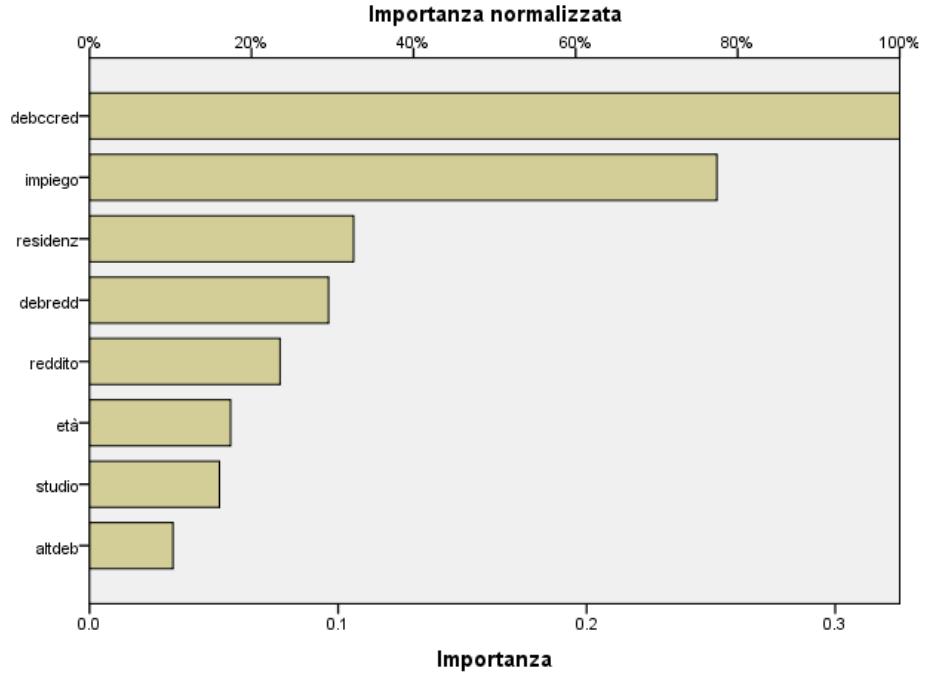
Importanza della variabile indipendente

Figura 4-21
Importanza della variabile indipendente

	Importanza	Importanza normalizzata
Titolo di studio	,032	11,9%
Età in anni	,075	27,9%
Anni presso l'attuale datore di lavoro	,268	100,0%
Anni di residenza all'indirizzo attuale	,166	61,8%
Reddito famigliare in migliaia	,033	12,2%
Rapporto debiti / reddito (x100)	,125	46,5%
Debito su carta di credito in migliaia	,213	79,3%
Altri debiti in migliaia	,090	33,6%

L'importanza di una variabile indipendente è una misura della variazione del valore atteso del modello della rete per valori differenti della variabile indipendente. L'importanza normalizzata indica semplicemente i valori dell'importanza divisi per i valori dell'importanza più grandi ed espressi come percentuale.

Figura 4-22
Grafico dell'importanza delle variabili indipendenti



Il grafico di importanza è semplicemente un grafico a barre dei valori contenuti nella tabella dell'importanza, ordinati in base al valore decrescente dell'importanza. Sembra che le variabili correlate alla stabilità del cliente (*impiego*, *indirizzo*) e all'indebitamento (*creddeb*, *debred*) esercitino il maggiore influsso sulla modalità in cui la rete classifica i clienti; ciò che non può

essere indicato è la “direzione” della relazione tra queste variabili e la probabilità prevista di inadempienza. Si può ipotizzare che un indebitamento maggiore indichi una maggiore probabilità di inadempienza, ma per verificare questa condizione è necessario utilizzare un modello con parametri più facili da interpretare.

Riepilogo

Utilizzando la procedura Perceptron a più strati è stata creata una rete per la previsione della probabilità che un determinato cliente sia inadempiente rispetto a un mutuo. I risultati del modello sono paragonabili con quelli ottenuti tramite la regressione logistica o l’analisi discriminante; è quindi possibile ritenere che i dati non contengano relazioni che non possano essere catturate da questi modelli ed è pertanto possibile utilizzarli per esplorare ulteriormente la natura della relazione tra le variabili dipendenti e indipendenti.

Utilizzo di Perceptron a più strati per valutare i costi del sistema sanitario e la durata della degenza

Un ospedale è interessato a registrare i costi e la durata delle degenze per i pazienti ricoverati per infarto del miocardio. Stime accurate relative a tali misure consentono all’amministrazione di gestire correttamente il numero di posti letto disponibili per la degenza dei pazienti.

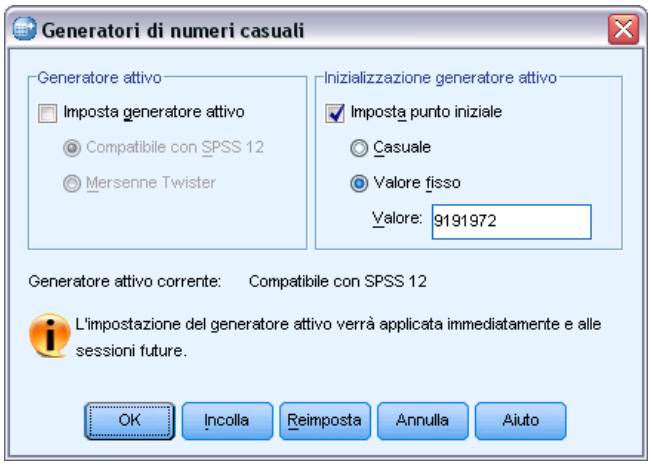
Il file di dati *patient_los.sav* contiene l’archivio sanitario di un campione di pazienti curati per infarto. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 89. Utilizzare la procedura Perceptron a più strati per creare una rete di previsione dei costi e della durata della degenza.

Preparazione dei dati per l’analisi

L’impostazione del seme casuale consente di replicare esattamente l’analisi.

- Per impostare il seme casuale, dai menu scegliere:
  Trasforma > Generatori numeri casuali...

Figura 4-23
Finestra di dialogo Generatori di numeri casuali

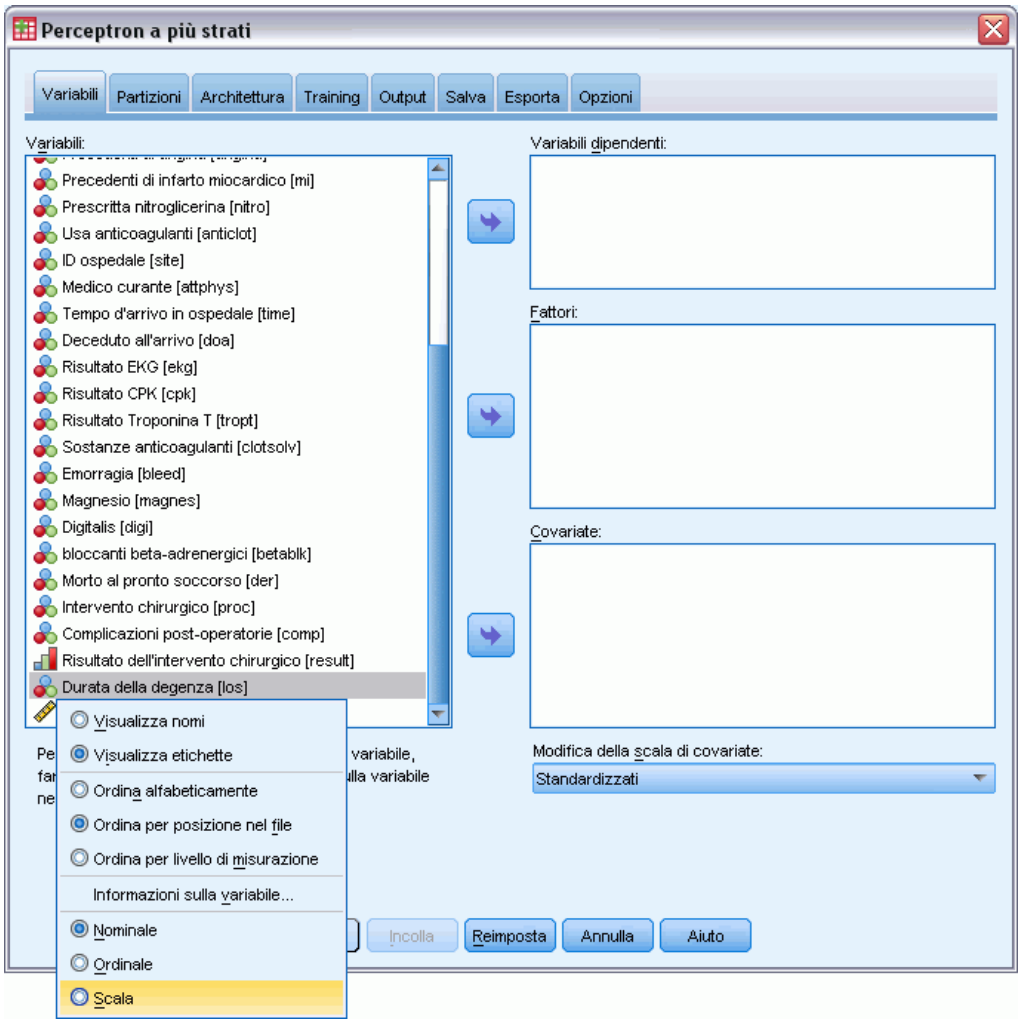


- Selezionare Imposta punto inziale.
- Selezionare Valore fisso e digitare il valore 9191972.
- Fare clic su OK.

Esecuzione dell'analisi

- Per eseguire un'analisi Perceptron a più strati, dai menu scegliere:
  Analizza > Neural Networks > Perceptron a più strati...

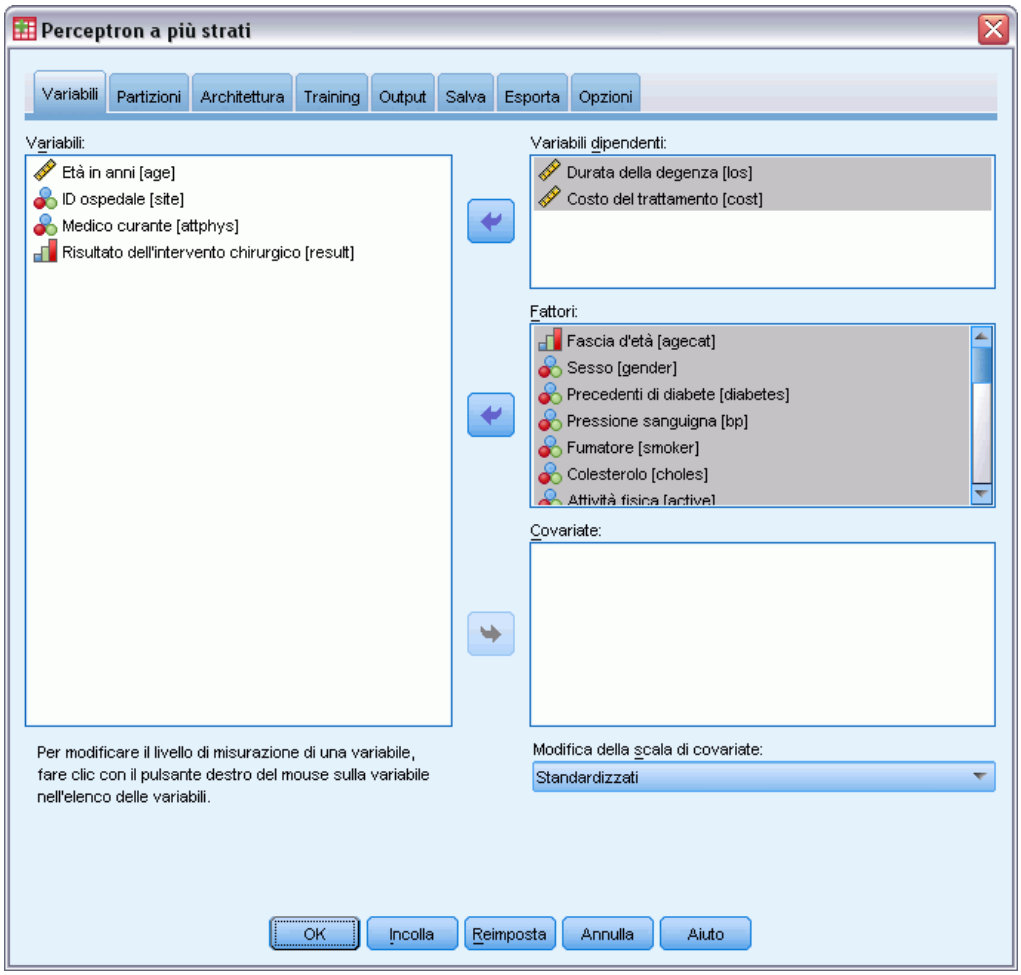
Figura 4-24
Perceptron a più strati: Scheda Variabili e menu di scelta rapida per la durata della degenza



Durata della degenza [los] ha un livello di misurazione ordinale, ma si desidera che la rete tratti questo valore come scala.

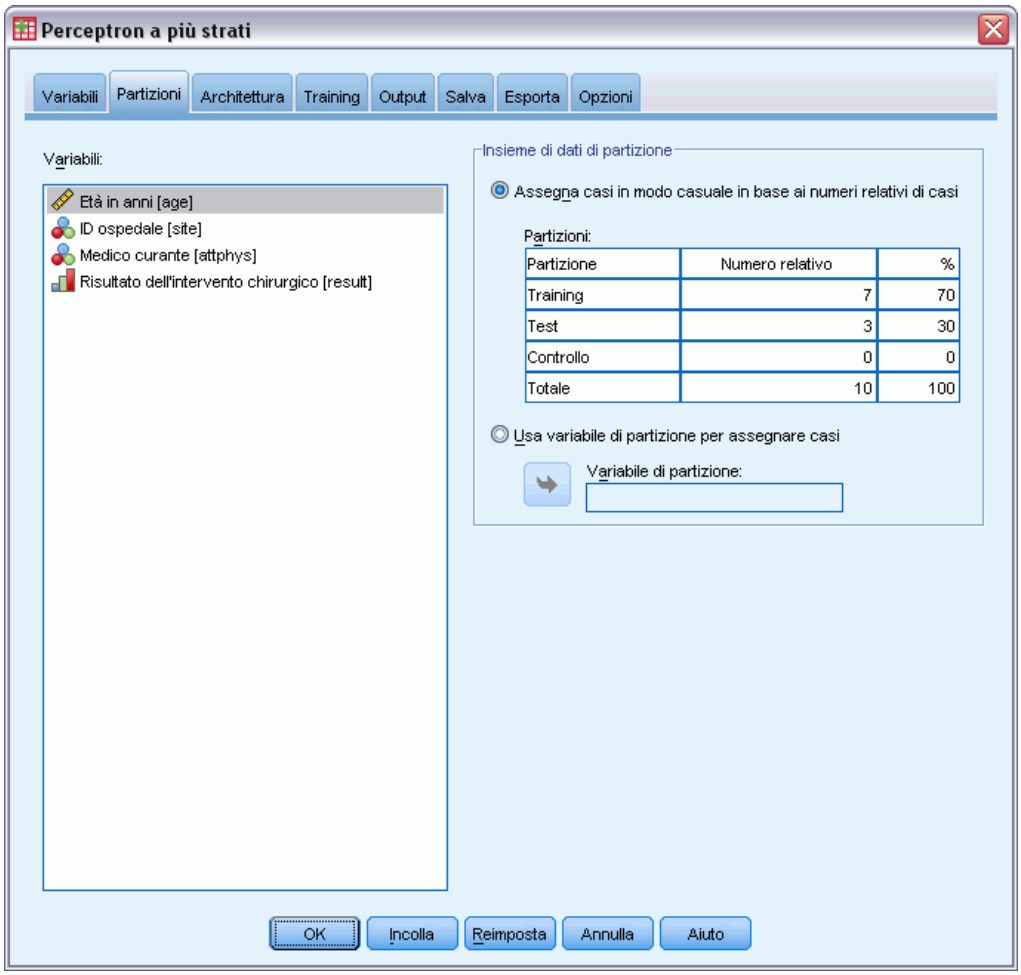
- Fare clic con il pulsante destro del mouse su *Durata della degenza [los]* e selezionare Scala nel menu di scelta rapida.

Figura 4-25
Perceptron a più strati: scheda Variabili con variabili e fattori dipendenti selezionati



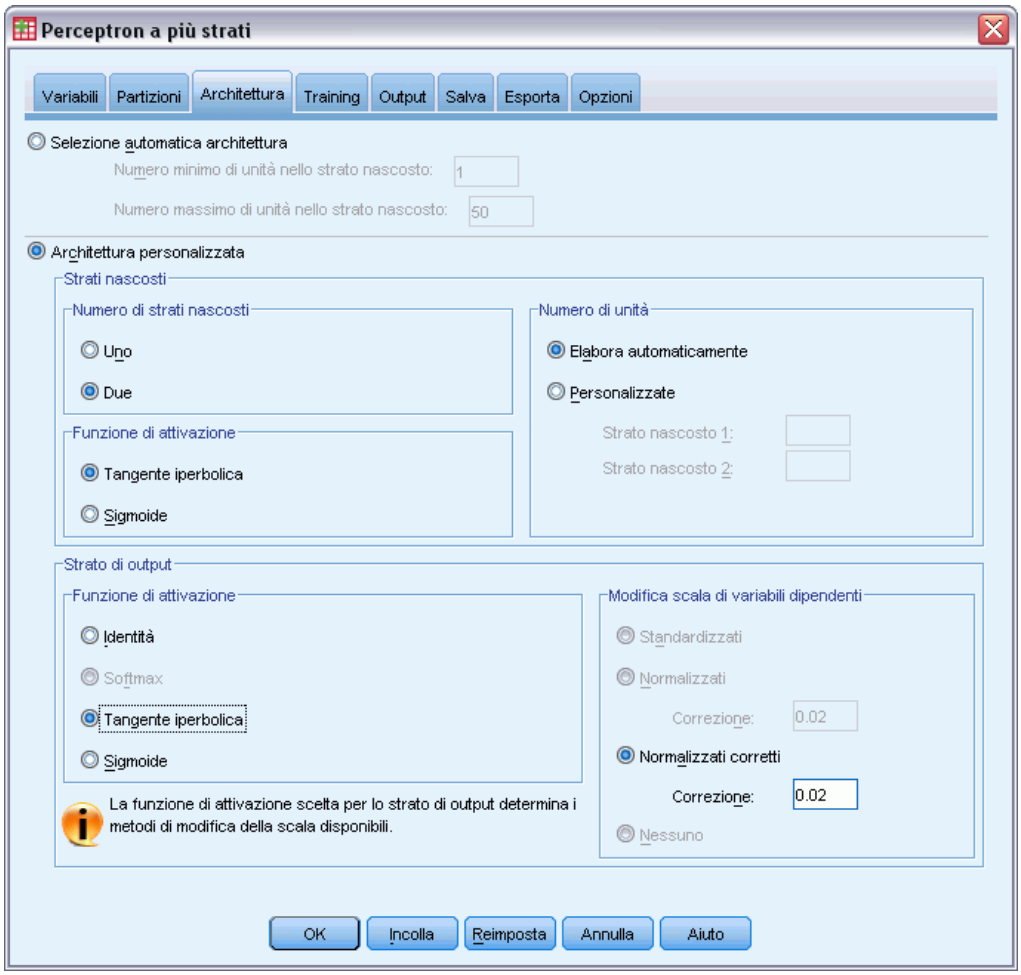
- Selezionare *Durata della degenza [los]* e *Costo del trattamento [cost]* come variabili dipendenti.
- Selezionare da *Fascia d'età [agecat]* a *Usa anticoagulanti [anticlot]* e da *Tempo d'arrivo in ospedale [time]* a *Complicazioni post-operatorie [comp]* come fattori. Per assicurare la replica esatta dei risultati del modello riportati di seguito, assicurarsi di mantenere l'ordine delle variabili nell'elenco dei fattori. A tale scopo, può essere utile selezionare ciascun insieme di predittori e utilizzare il pulsante per spostarli nell'elenco dei fattori, piuttosto che trascinarli. In alternativa, cambiare l'ordine delle variabili può essere di aiuto nel valutare la stabilità della soluzione.
- Fare clic sulla scheda Partizioni.

Figura 4-26
Perceptron a più strati: Scheda Partizioni



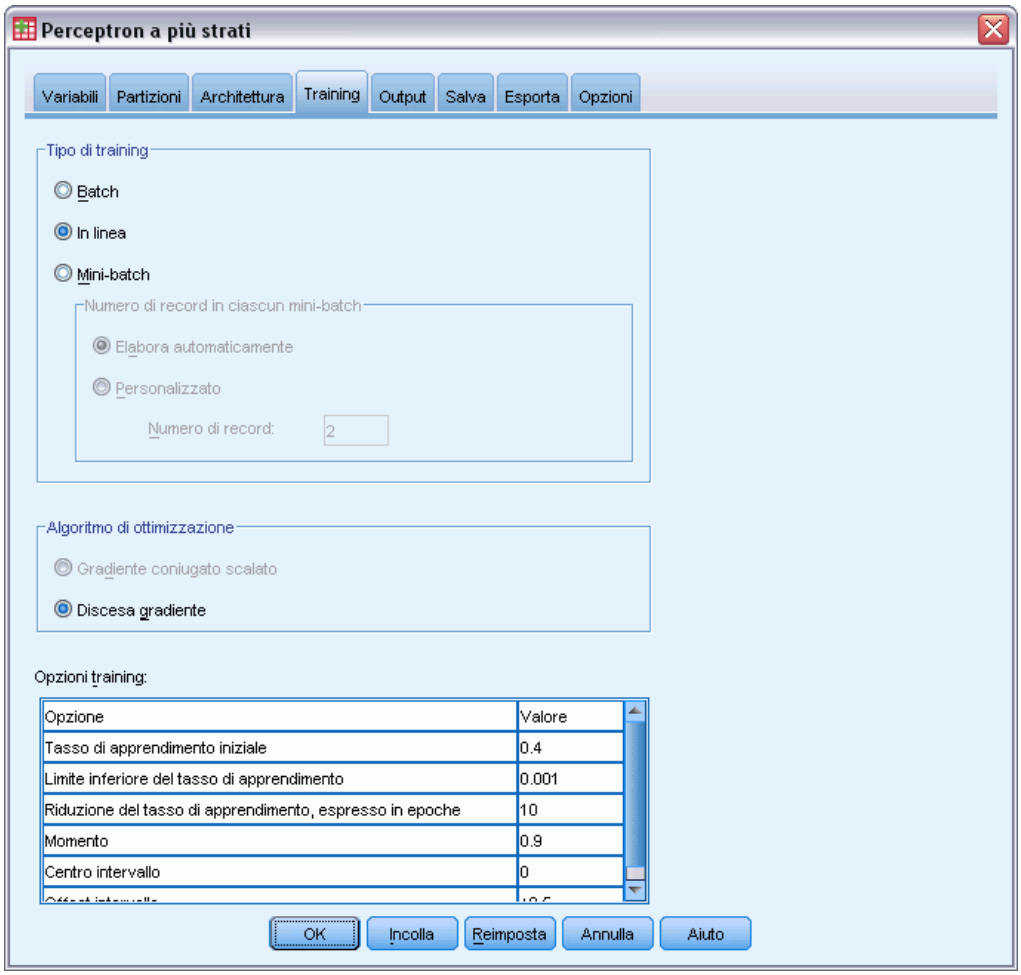
- Digitare 2 come numero relativo di casi da assegnare al campione di verifica.
- Digitare 1 come numero relativo di casi da assegnare al campione di controllo.
- Fare clic sulla scheda Architettura.

Figura 4-27
Perceptron a più strati: Scheda Architettura



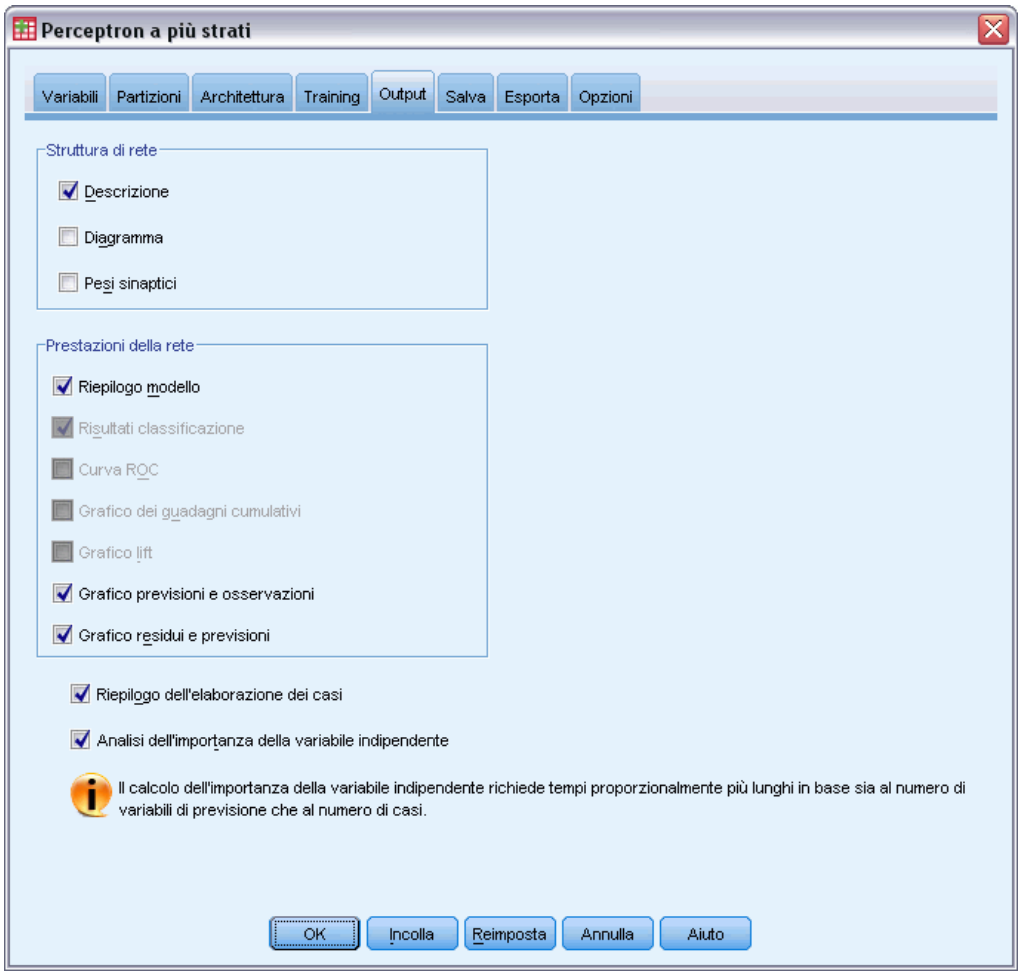
- Selezionare Architettura personalizzata.
- Selezionare Due come numero di strati nascosti.
- Selezionare Tangente iperbolica come funzione di attivazione per lo strato di output. Notare che in questo modo si imposta automaticamente il metodo di modifica della scala per le variabili dipendenti sul valore Normalizzati corretti.
- Fare clic sulla scheda Training.

Figura 4-28
Perceptron a più strati: Scheda Training



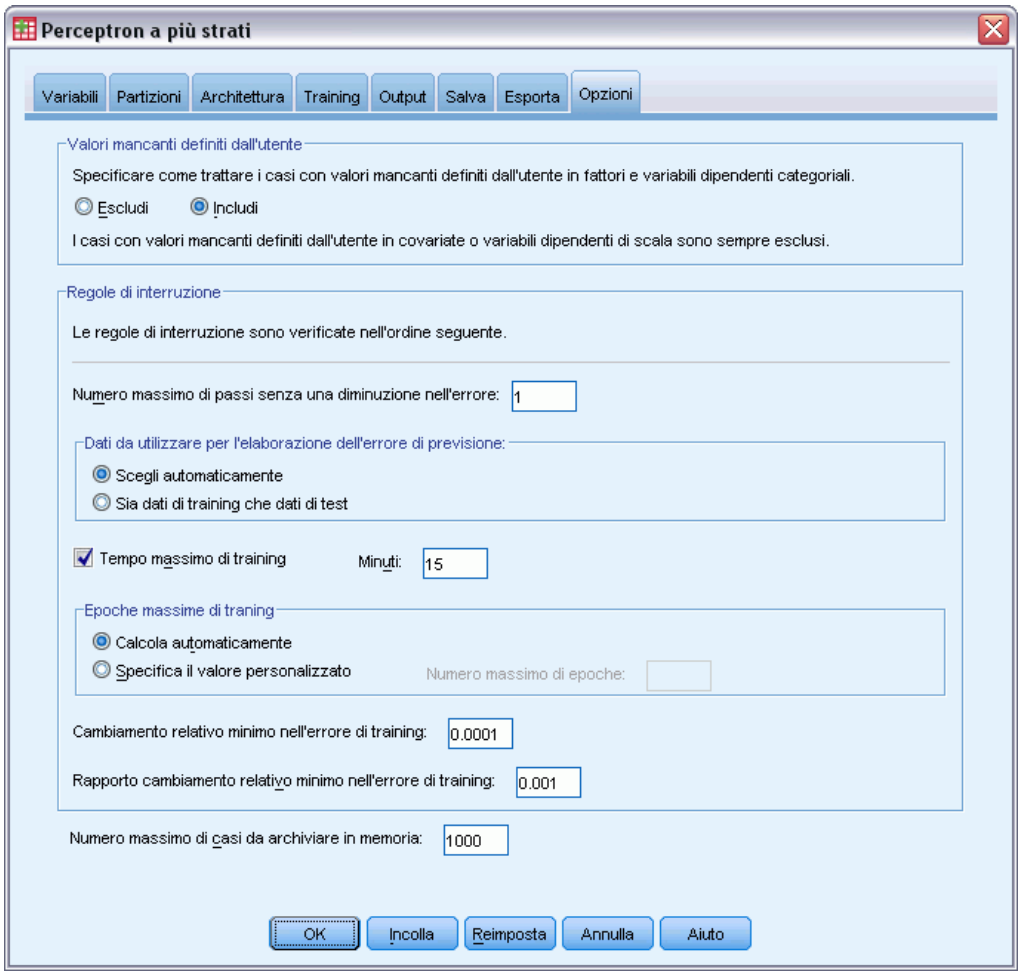
- Selezionare In linea come tipo di addestramento. L'addestramento in linea è adatto per gli insiemi di dati di grandi dimensioni con predittori correlati. Notare che viene automaticamente impostato Discesa gradiente come algoritmo di ottimizzazione con le corrispondenti opzioni predefinite.
- Fare clic sulla scheda Output.

Figura 4-29
Perceptron a più strati: scheda Output



- Deselezionare Diagramma; poiché ci sono numerosi dati di input, il diagramma risultante potrebbe essere poco gestibile.
- Selezionare Grafico previsioni e osservazioni e Grafico residui e previsioni nel gruppo Prestazioni della rete. I risultati della classificazione, la curva ROC, il grafico dei guadagni cumulativi e il grafico lift non sono disponibili poiché nessuna variabile dipendente viene trattata come categorica (nominale oppure ordinale).
- Selezionare Analisi dell'importanza della variabile indipendente.
- Fare clic sulla scheda Opzioni.

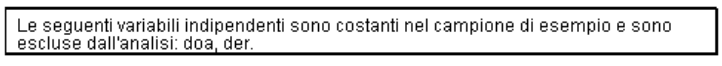
Figura 4-30
Scheda Opzioni



- Scegliere di includere le variabili utente mancanti. Per i pazienti che non sono stati sottoposti a interventi chirurgici vi sono dei valori definiti dall'utente mancanti nella variabile *Complicazioni post-operatorie*. Ciò assicura che tali pazienti vengano sempre inclusi nell'analisi.
- Fare clic su OK.

Avvisi

Figura 4-31
Avvisi



La tabella degli avvisi indica che le variabili *doa* e *der* sono costanti nel campione dell'addestramento. Per i pazienti che sono giunti deceduti all'ospedale oppure il cui decesso è avvenuto al pronto soccorso vi sono dei valori definiti dall'utente mancanti in *Durata della degenza*. Poiché la variabile *Durata della degenza* viene trattata come variabile di scala per questa analisi e i casi con valori definiti dall'utente mancanti sulle scale variabili sono esclusi, vengono inclusi solo i pazienti usciti in vita dal pronto soccorso.

Riepilogo dei casi

Figura 4-32
Riepilogo dell'elaborazione dei casi

		N	Percentuale
Campione	Training	5647	70,6%
	Test	1570	19,6%
	Controllo	781	9,8%
	Valida	7998	100,0%
	Esclusa	2002	
	Totale	10000	

Il riepilogo di elaborazione dei casi mostra che 5647 casi sono stati assegnati al campione di addestramento, 1570 al campione di verifica e 781 al campione di controllo. I 2002 casi esclusi dall'analisi riguardano pazienti che sono deceduti durante il trasporto all'ospedale oppure al pronto soccorso.

Informazioni di rete

Figura 4-33
informazioni di rete

Strato di input	Fattori	1	time
		2	ekg
		3	cpk
		4	tropt
		5	clotsolv
		6	bleed
		7	magnes
		8	digi
		9	betablk
		10	proc
		11	comp
		12	gender
		13	diabetes
		14	bp
		15	smoker
		16	choles
		17	active
		18	obesity
		19	angina
		20	mi
		21	nitro
		22	anticlot
	Numero di unità[a]		63
Strato/i nascosto/i	Numero di strati nascosti		2
	Numero di unità nello strato nascosto 1[a]		12
	Number of Units in Hidden Layer 2[a]		9
	Funzione di attivazione		Tangente iperbolica
Strato di output	Dependent Variables	1	los
		2	cost
	Numero di unità		2
	Rescaling Method for Scale Dependents		Adjusted Normalized
	Funzione di attivazione		Tangente iperbolica
	Funzione di errore		Somma dei quadrati

a. Senza unità diagonale

La tabella delle informazioni sulla rete visualizza le informazioni sulla rete neurale ed è utile per garantire che le specifiche siano corrette. Si noti in particolare che:

- Il numero di unità nello strato di input rappresenta il numero totale di livelli di fattori (non sono presenti covariate).
- Sono stati richiesti due strati nascosti e la procedura ha scelto 12 unità nel primo strato nascosto e 9 nel secondo.
- Viene creata un'unità di output per ciascuna delle variabili dipendenti di scala. La scala di tali unità viene modificata in base al metodo normalizzato corretto, che richiede l'uso della funzione di attivazione della tangente iperbolica per lo strato di output.
- Viene riportato l'errore relativo alla somma dei quadrati, poiché le variabili dipendenti sono variabili di scala.

Riepilogo del modello (Regressione output)

Figura 4-34
Riepilogo modello

Training	Errore della somma dei quadrati		91,812
	Errore relativo complessivo medio		,083
	Errore relativo per variabili dipendenti di scala	los	,131
		cost	,033
	Regola di interruzione utilizzata		1 passi consecutivi senza diminuzione degli errori[a]
	Tempo di training		00:00:18.055
Test	Errore della somma dei quadrati		26,798
	Errore relativo complessivo medio		,088
	Errore relativo per variabili dipendenti di scala	los	,141
		cost	,033
Controllo	Errore relativo complessivo medio		,099
	Errore relativo per variabili dipendenti di scala	los	,154
		cost	,041

a. I calcoli degli errori si basano sul campione di verifica.

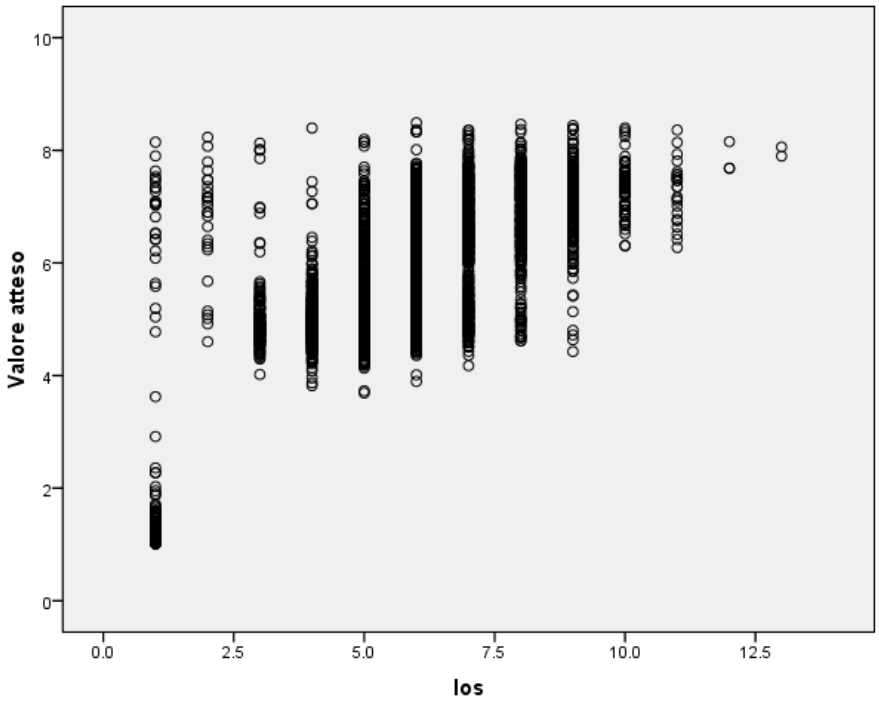
Il riepilogo del modello visualizza le informazioni sui risultati del training e l'applicazione della rete finale al campione di controllo.

- Viene riportato l'errore relativo alla somma dei quadrati, poiché lo strato di output utilizza variabili dipendenti di scala, che è la funzione di errore che la rete cerca di minimizzare durante il training. Notare che la somma dei quadrati e i valori degli errori seguenti vengono calcolati per i valori delle variabili dipendenti di cui è stata modificata la scala.
- L'errore relativo di ciascuna variabile dipendente di scala è il rapporto tra l'errore della somma dei quadrati della variabile dipendente e l'errore della somma dei quadrati del modello "nullo" in cui il valore medio della variabile dipendente viene utilizzato come valore previsto per ciascun caso. Il risultato indica la presenza di più errori nelle previsioni della *Durata della degenza* rispetto al *Costo del trattamento*.
- L'errore globale medio è il rapporto tra l'errore della somma dei quadrati di tutte le variabili dipendenti e la somma dell'errore dei quadrati del modello "nullo" in cui i valori medi delle variabili dipendenti vengono utilizzati come valori previsti per ciascun caso. In questo esempio, l'errore globale medio si avvicina alla media degli errori relativi, ma questo non è sempre il caso.

  L'errore relativo globale medio e gli errori relativi rimangono abbastanza costanti nei campioni di addestramento, di verifica e di controllo. Ciò indica che non si è verificato un eccesso di addestramento del modello e l'errore nei casi futuri rilevati dalla rete sarà vicino all'errore riportato in questa tabella.
- L'algoritmo di stima si è interrotto poiché l'errore non è diminuito dopo un passo dell'algoritmo.

Grafici previsioni e osservazioni

Figura 4-35
Grafico previsioni e osservazioni per la durata della degenza

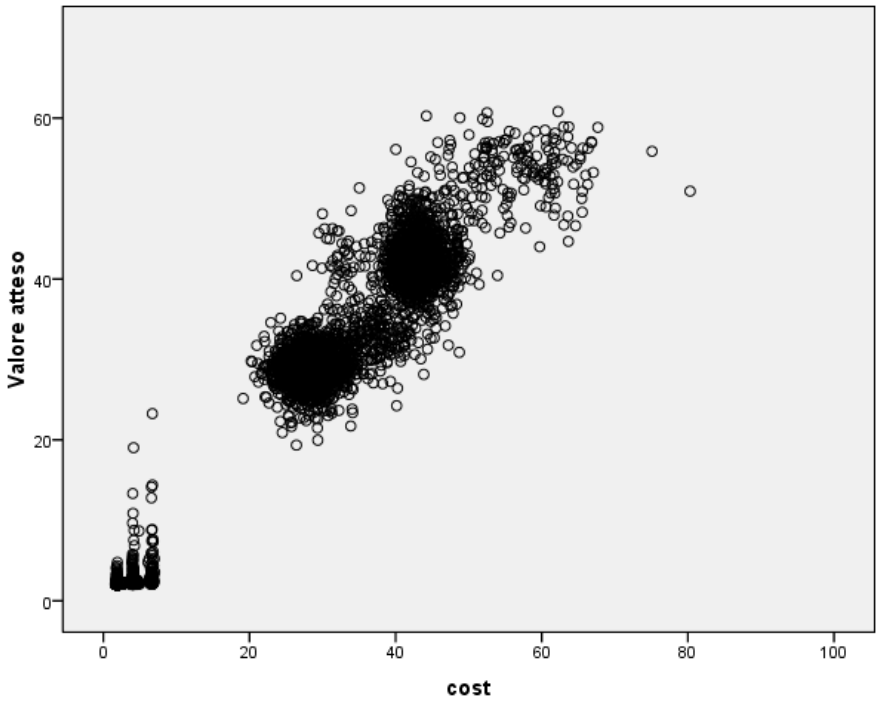


Per le variabili dipendenti di scala, il Grafico previsioni e osservazioni visualizza un grafico a dispersione dei valori attesi sull'asse *Y* e i valori osservati sull'asse *X* relativi ai campioni di addestramento e verifica combinati. Idealmente, i valori dovrebbero trovarsi approssimativamente lungo una linea a 45 gradi a partire dal punto di origine. I punti di questo grafico formano linee verticali in corrispondenza di ciascun numero di giorni osservato indicato in *Durata della degenza*.

Se si osserva il grafico, è possibile notare che la rete è abbastanza affidabile per quanto riguarda la previsione della *Durata della degenza*. La tendenza generale del grafico non segue la linea ideale a 45 gradi, nel senso che le previsioni della durata delle degenze osservata nei primi cinque giorni tendono a un valore sovrastimato, mentre le previsioni oltre il sesto giorno tendono a un valore sottostimato.

Il gruppo di pazienti nella parte in basso a sinistra del grafico indica pazienti che probabilmente non sono stati sottoposti a intervento chirurgico. Si nota anche un gruppo di pazienti in alto a sinistra del grafico, la cui durata del ricovero osservata è da 1 a 3 giorni ma i valori previsti sono di gran lunga maggiori. Si tratta probabilmente di casi di pazienti che sono deceduti nell'ospedale dopo essere stati sottoposti a intervento chirurgico.

Figura 4-36
Grafico previsioni e osservazioni per il costo del trattamento



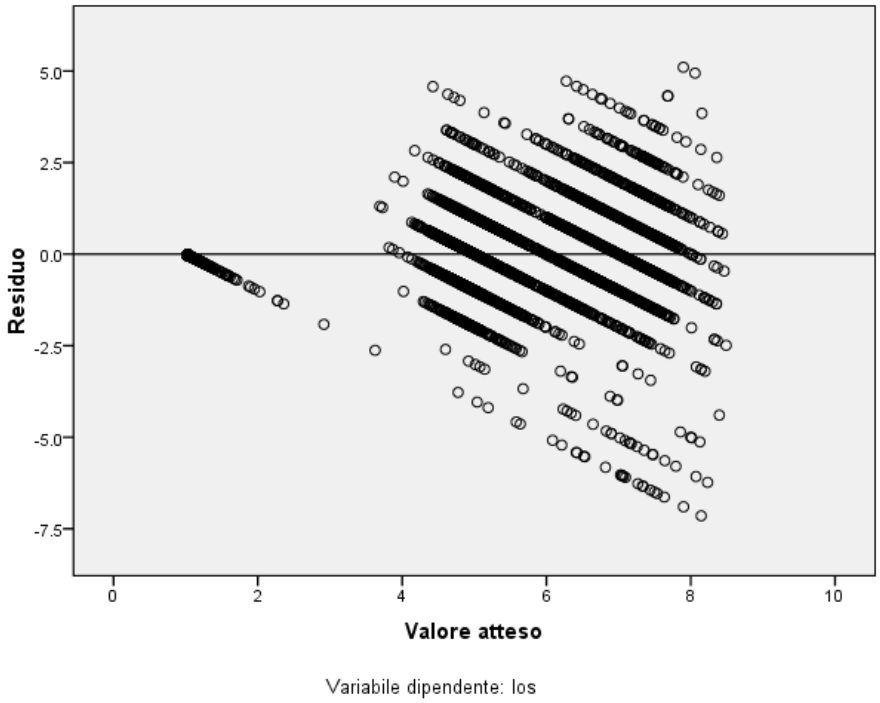
Anche in questo caso la rete si dimostra abbastanza affidabile nell'ambito del *Costo del trattamento*. Il grafico mostra tre gruppi principali di pazienti.

- In basso a sinistra è riportato un gruppo che non è stato sottoposto a intervento chirurgico. I costi ad esso associati sono relativamente bassi e differenziati in base al tipo di trattamento con anticoagulanti (variabile *Sostanze anticoagulanti [clotsolv]*) effettuato al pronto soccorso.
- Il gruppo di pazienti successivo presenta costi pari a circa $ 30.000. Si tratta di pazienti sottoposti ad angioplastica coronarica percutanea transluminale (PTCA).
- Il gruppo di pazienti finale presenta costi di trattamento che superano $ 40.000. Si tratta di pazienti sottoposti ad innesto di bypass aortocoronarico (CABG). Questo intervento chirurgico è piuttosto più costoso del PTCA e prevede una degenza in ospedale più lunga (che aumenta ulteriormente i costi).

Sono presenti anche altri casi con costi superiori a $ 50.000, ma per questi la previsione della rete non è affidabile. Si tratta di pazienti che hanno avuto complicazioni durante l'intervento chirurgico, con conseguente aumento del costo dell'intervento e della durata del ricovero.

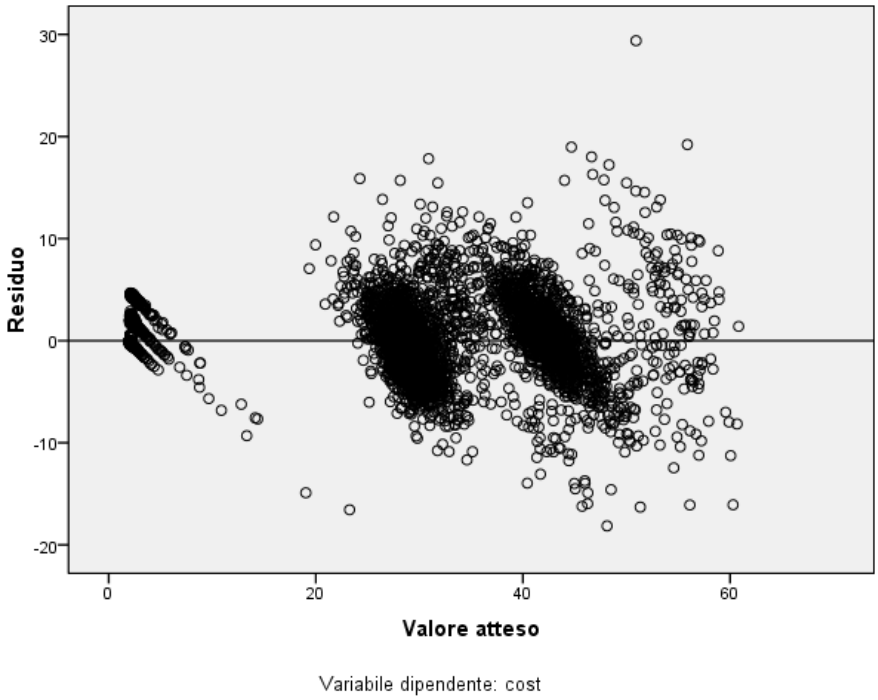
Grafici residui e previsioni

Figura 4-37
Grafico residui e previsioni per Durata della degenza



Il Grafico residui e previsioni visualizza un grafico a dispersione dei valori residui (valore osservato meno valore atteso) sull'asse *Y* in base ai valori previsti indicati sull'asse *X*. Ciascuna linea diagonale in questo grafico corrisponde a una linea verticale del Grafico previsioni e osservazioni. Il grafico mostra più chiaramente la progressione dalla previsione sovrastimata alla previsione sottostimata della durata del ricovero man mano che aumentano i giorni di ricovero osservati.

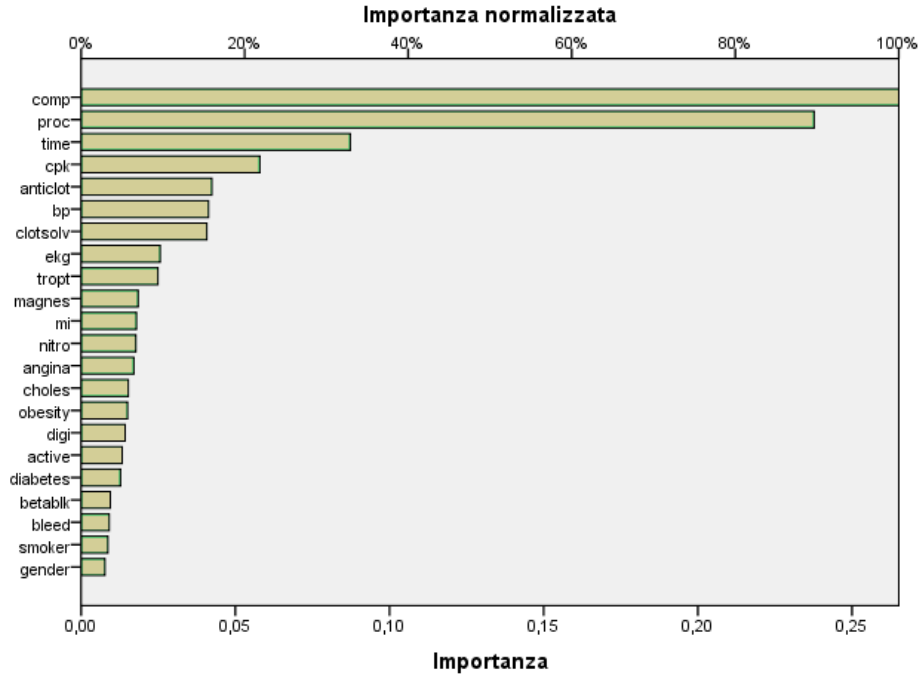
Figura 4-38
Grafico previsioni e osservazioni per il costo del trattamento



Allo stesso modo, per ciascuno dei tre gruppi di pazienti osservati nel grafico previsioni e osservazioni relativamente al *Costo del trattamento*, i valori residui del grafico delle previsioni mostrano una previsione che passa da valori soprastimati a valori sottostimati dei costi man mano che aumentano i costi osservati. I pazienti che hanno avuto complicazioni durante l'intervento CABG sono ancora chiaramente visibili, ma è anche facile notare i pazienti che hanno avuto complicazioni durante l'intervento PTCA. Questi pazienti si trovano nel sottogruppo a destra e sopra il gruppo principale di pazienti PTCA intorno al valore $ 30.000 sull'asse *X*.

Importanza della variabile indipendente

Figura 4-39
Grafico dell'importanza delle variabili indipendenti



Il grafico dell'importanza indica che nei risultati ottenuti prevale l'intervento chirurgico, seguito dalle complicazioni che si sono incontrate e, a distanza, da altri predittori. L'importanza dell'intervento chirurgico è chiaramente visibile nei grafici che prendono in esame il *Costo del trattamento*; un po' meno in quelli che prendono in esame la *Durata della degenza*, anche se l'effetto delle complicazioni sulla *Durata della degenza* è visibile nei pazienti che hanno avuto il periodo di osservazione più lungo.

Riepilogo

La rete sembra essere affidabile nel prevedere i valori per i pazienti "tipici", ma non riesce a valutare i pazienti che sono deceduti dopo l'intervento chirurgico. Un modo per gestire questa situazione potrebbe essere creare più reti: una rete per prevedere i risultati dei pazienti, ad esempio quelli che sono sopravvissuti e quelli deceduti, e altre reti separate per prevedere il *Costo del trattamento* e la *Durata della degenza* per i pazienti che hanno superato l'intervento chirurgico. I risultati ricavati dalle reti possono essere utilizzati insieme per cercare di ottenere previsioni migliori. È possibile intraprendere un approccio simile al problema della previsione sottostimata dei costi e delle durate dei ricoveri dei pazienti che hanno incontrato complicazioni durante l'intervento chirurgico.

Letture consigliate

Consultare i testi seguenti per ulteriori informazioni sulle reti neurali e sui perceptron a più strati:

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

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

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

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

Capitolo
5

Funzione a base radiale

La procedura Funzione a base radiale (RBF) produce un modello predittivo per una o più variabili dipendenti (di destinazione) basato sui valori delle variabili predittore.

Utilizzo della Funzione a base radiale per classificare i clienti delle telecomunicazioni

Un fornitore di telecomunicazioni ha segmentato la base clienti per modelli di utilizzo del servizio, suddividendo i clienti in quattro categorie. Se si utilizzano i dati demografici per prevedere l'appartenenza al gruppo, è possibile personalizzare le offerte per potenziali clienti individuali.

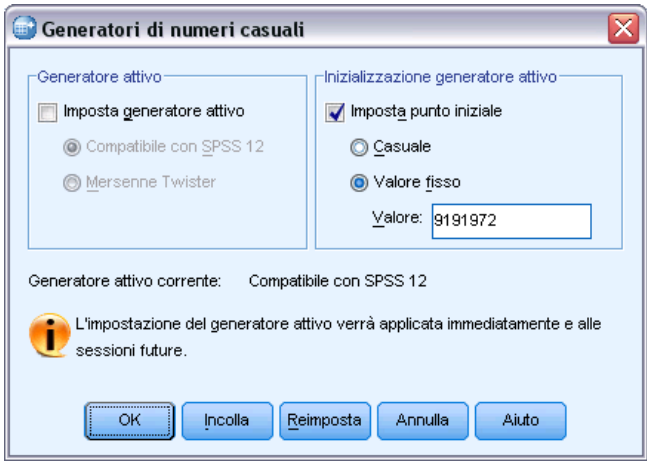
Si supponga che le informazioni sugli attuali clienti siano contenute in *telco.sav*. Per ulteriori informazioni, vedere l'argomento File di esempio in l'appendice A a pag. 89. Utilizzare la procedura Funzione a base radiale per classificare i clienti.

Preparazione dei dati per l'analisi

L'impostazione del seme casuale consente di replicare esattamente l'analisi.

- Per impostare il seme casuale, dai menu scegliere:
  Trasforma > Generatori numeri casuali...

Figura 5-1
Finestra di dialogo Generatori di numeri casuali



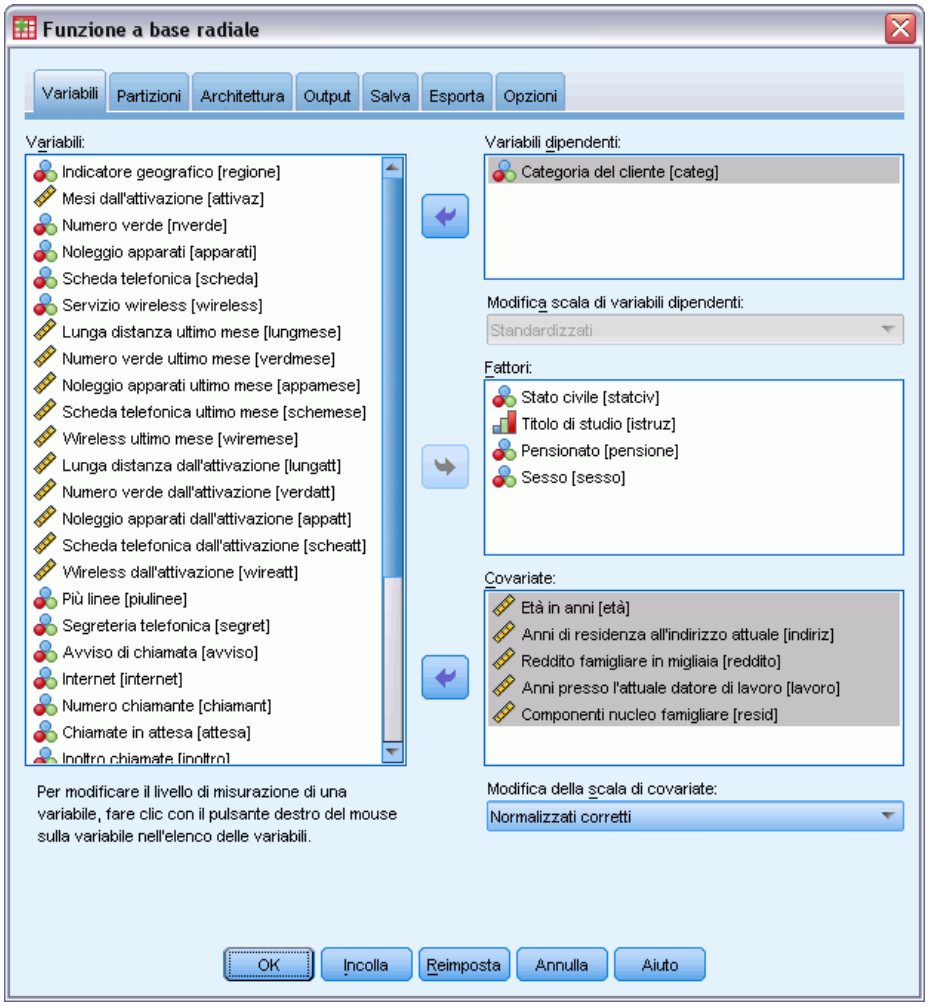
- Selezionare Imposta punto inziale.
- Selezionare Valore fisso e digitare il valore 9191972.

- Fare clic su OK.

Esecuzione dell'analisi

- Per eseguire un'analisi di tipo Funzione a base radiale, dai menu, scegliere:
  Analizza > Reti neurali > Funzione a base radiale...

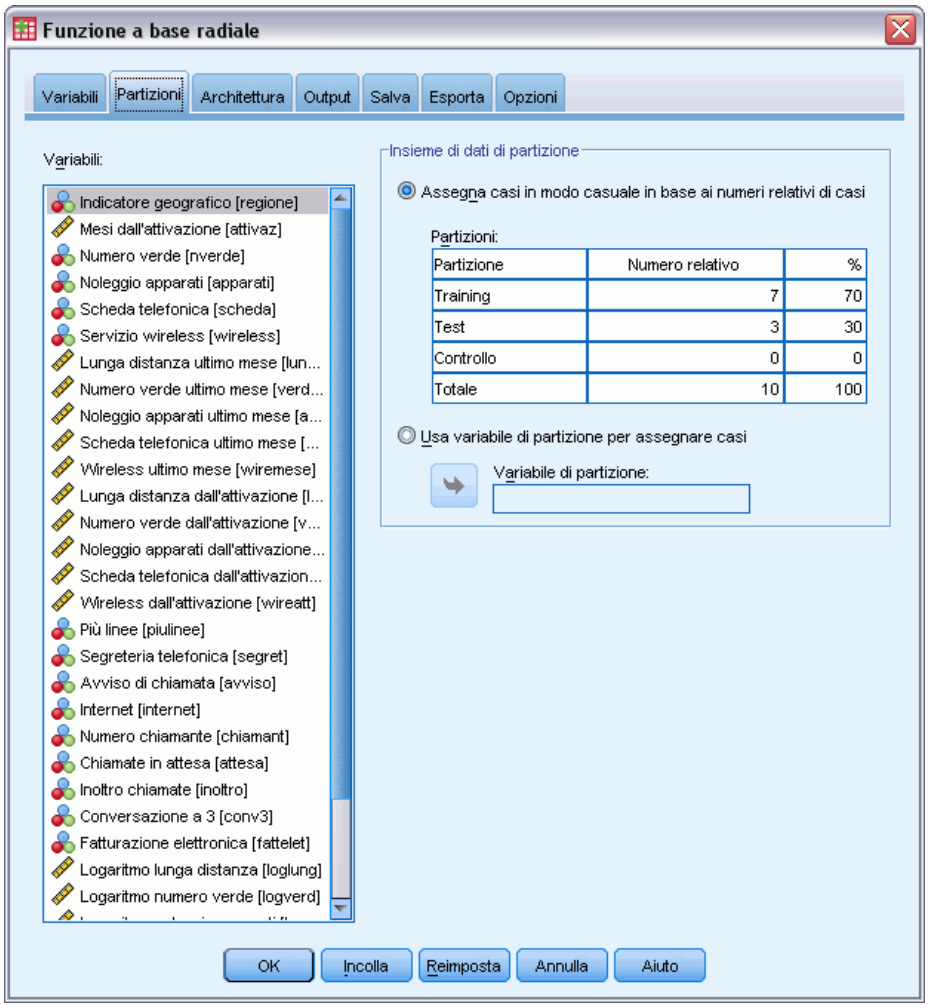
Figura 5-2
Funzione a base radiale: scheda Variabili



- Selezionare *Categoria del cliente [categ]* come variabile dipendente.
- Selezionare *Stato civile [statociv]*, *Titolo di studio [istruz]*, *Pensionato [pensione]* e *Sesso [sesso]* come fattori.
- Selezionare da *Età in anni [età]* a *Componenti nucleo famigliare [resid]* come covariate.
- Selezionare Normalizzati corretti come metodo per la modifica della scala delle covariate.

▶ Fare clic sulla scheda Partizioni.

Figura 5-3
Funzione a base radiale: Scheda Partizioni



Specificando i numeri relativi di casi, è facile creare partizioni frazionarie di cui sarebbe difficile specificare le percentuali. Si supponga di voler assegnare 2/3 dell'insieme di dati al campione di addestramento e 2/3 dei casi rimanenti al campione di verifica.

▶ Digitare 6 come numero relativo per il campione di addestramento.

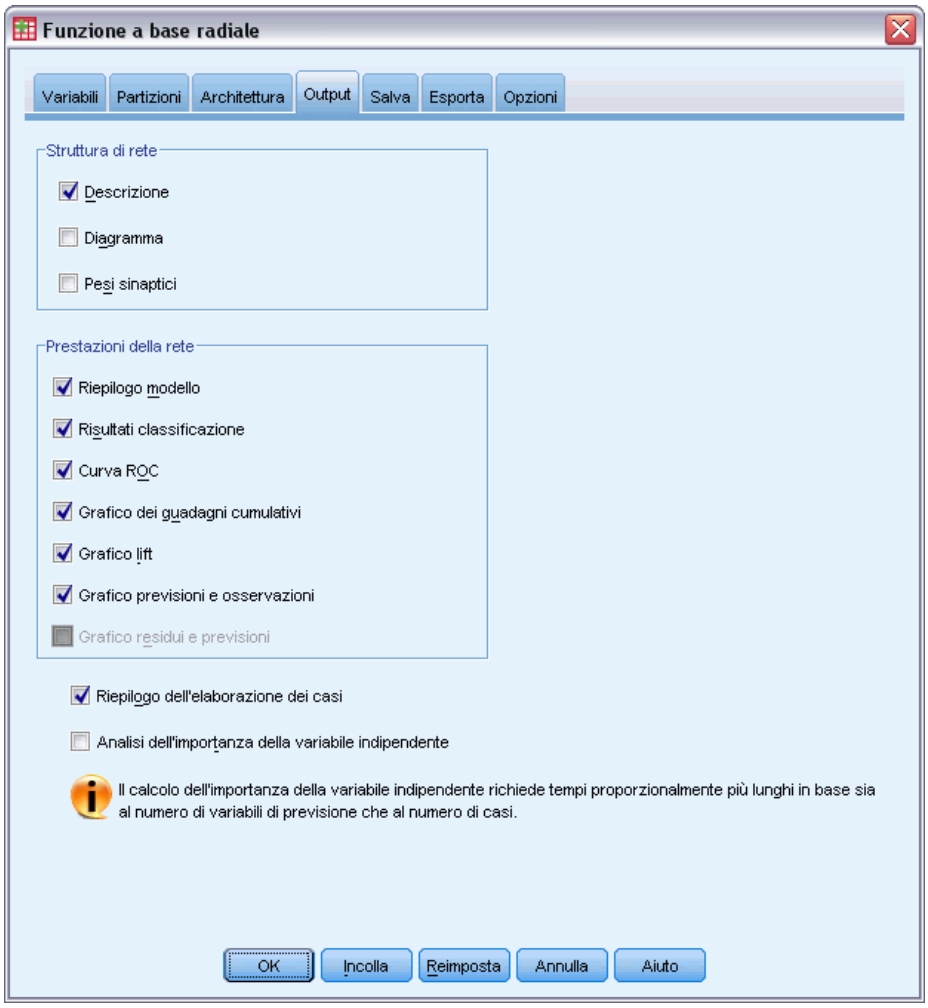
▶ Digitare 2 come numero relativo per il campione di verifica.

▶ Digitare 1 come numero relativo per il campione di controllo.

È stato specificato un totale di 9 casi relativi. 6/9 = 2/3, o il 66,67% circa, sono assegnati al campione di addestramento; 2/9 o il 22,22% circa, sono assegnati al campione di verifica; 1/9 o l'11,11% circa sono assegnati al campione di controllo.

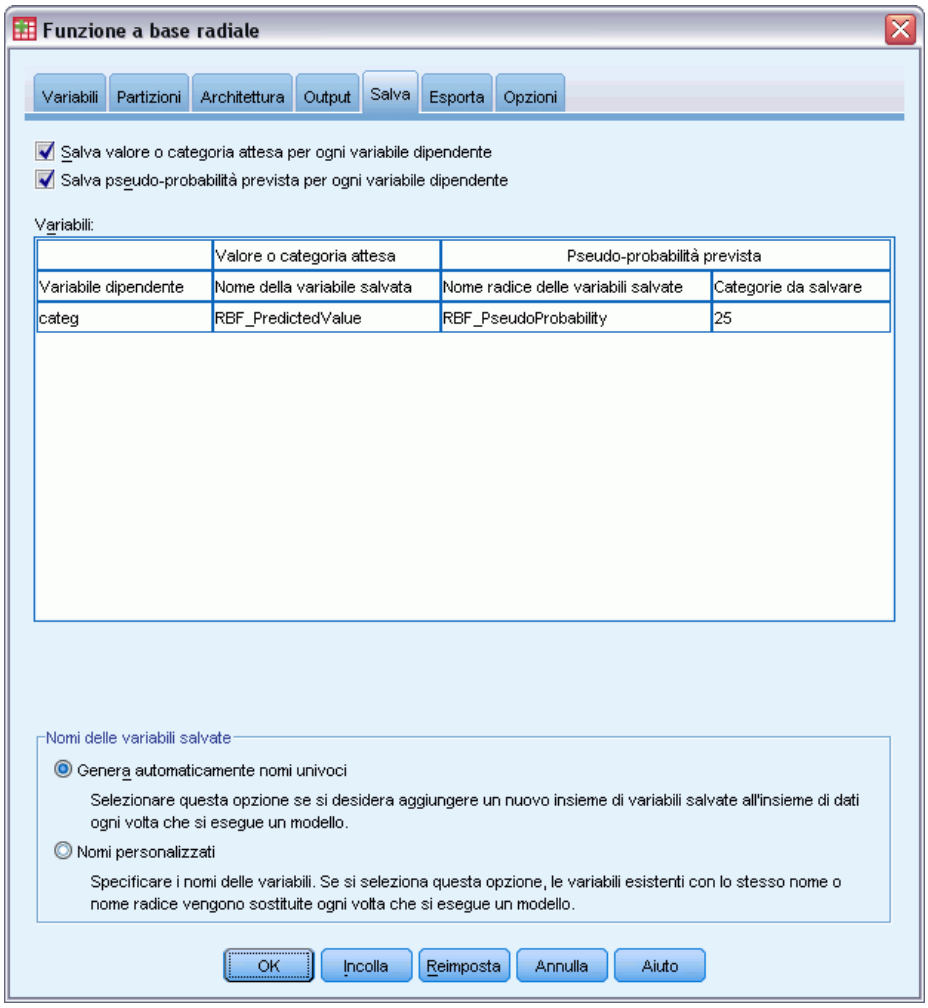
▶ Fare clic sulla scheda Output.

Figura 5-4
Funzione a base radiale: scheda Output



- Deselezionare Diagramma nel gruppo Struttura di rete.
- Selezionare Curva ROC, Grafico dei guadagni cumulativi, Grafico lift e Grafico previsioni e osservazioni nel gruppo Prestazioni della rete.
- Fare clic sulla scheda Salva.

Figura 5-5
Funzione a base radiale: scheda Salva



- Selezionare Salva valore o categoria attesa per ogni variabile dipendente e Salva pseudo-probabilità prevista per ogni variabile dipendente.
- Fare clic su OK.

Riepilogo dei casi

Figura 5-6
Riepilogo dell'elaborazione dei casi

		N	Percentuale
Campione	Training	665	66,5%
	Test	224	22,4%
	Controllo	111	11,1%
Valida		1000	100,0%
Esclusa		0	
Totale		1000	

Il riepilogo di elaborazione dei casi mostra che 665 casi sono stati assegnati al campione di addestramento, 224 al campione di verifica e 111 al campione di controllo. Da questa analisi non sono stati esclusi casi.

Informazioni di rete

Figura 5-7
informazioni di rete

Strato di input	Fattori	1	Stato civile
		2	Titolo di studio
		3	Pensionato
		4	Sesso
	Covariates	1	Età in anni
		2	Anni di residenza all'indirizzo attuale
		3	Reddito famigliare in migliaia
		4	Anni presso l'attuale datore di lavoro
		5	Componenti nucleo famigliare
	Numero di unità		16
	Rescaling Method for Covariates		Adjusted Normalized
Strato nascosto	Numero di unità		9[a]
	Funzione di attivazione		Softmax
Strato di output	Dependent Variables	1	Categoria del cliente
	Numero di unità		4
	Funzione di attivazione		Identità
	Funzione di errore		Somma dei quadrati

a. Determinato dal criterio dei dati di test: il numero "migliore" di unità nascoste corrisponde a quello che restituisce il numero minimo di errori nei dati di test.

La tabella delle informazioni sulla rete visualizza le informazioni sulla rete neurale ed è utile per garantire che le specifiche siano corrette. Si noti in particolare che:

- Il numero di unità nello strato di input è il numero di covariate più il numero totale di livelli di fattore; viene creata un'unità separata per ciascuna categoria di *Stato civile*, *Titolo di studio*, *Pensionato* e *Sesso*, nessuna delle categorie viene considerata unità "ridondante" come è tipico in numerose procedure di creazione di modelli.

- Analogamente, viene creata un'unità di output separata per ciascuna categoria di *Categoria del cliente* per un totale di 4 unità nello strato di output.
- Le covariate vengono riscalate utilizzando il metodo normalizzato corretto.
- La selezione automatica dell'architettura ha scelto 9 unità nello strato nascosto.
- Tutte le altre informazioni sulla rete sono predefinite per la procedura.

Riepilogo del modello (Regressione output)

Figura 5-8
Riepilogo modello

Training	Errore della somma dei quadrati	235,969
	Percentuale di previsioni errate	61,8%
	Tempo di training	00:00:04.563
Test	Errore della somma dei quadrati	80,851[a]
	Percentuale di previsioni errate	62,9%
Controllo	Percentuale di previsioni errate	59,5%

Variabile dipendente: Categoria del cliente

a. Il numero di unità nascoste è determinato dal criterio dei dati di test: il numero "migliore" di unità nascoste corrisponde a quello che restituisce il numero minimo di errori nei dati di test.

Il riepilogo del modello visualizza le informazioni sui risultati dell'addestramento, del test e l'applicazione della rete finale al campione di controllo.

- Viene visualizzato l'errore della somma dei quadrati che viene sempre utilizzato per le reti RBF. Questa è la funzione di errore che la rete tenta di minimizzare durante l'addestramento e il test.
- La percentuale di previsioni non corrette viene prelevata dalla tabella di classificazione e verrà discussa più avanti in questo argomento.

Classificazione

Figura 5-9
Classificazione

Campione	Osservati	Previsto				
		Servizio base	E-service	Servizio plus	Servizio completo	Percent Correct
Training	Servizio base	64	0	66	45	36,6%
	E-service	22	1	57	61	,7%
	Servizio plus	47	0	104	34	56,2%
	Servizio completo	29	1	49	85	51,8%
	Overall Percent	24,4%	,3%	41,5%	33,8%	38,2%
Test	Servizio base	18	0	26	15	30,5%
	E-service	15	0	16	22	,0%
	Servizio plus	11	0	39	15	60,0%
	Servizio completo	4	0	17	26	55,3%
	Overall Percent	21,4%	,0%	43,8%	34,8%	37,1%
Controllo	Servizio base	11	0	11	10	34,4%
	E-service	4	0	9	10	,0%
	Servizio plus	10	0	19	2	61,3%
	Servizio completo	5	0	5	15	60,0%
	Overall Percent	27,0%	,0%	39,6%	33,3%	40,5%

Variabile dipendente: Categoria del cliente

La tabella di classificazione mostra i risultati pratici dell'utilizzo della rete. Per ogni caso, la risposta prevista è la categoria con la maggiore pseudo-probabilità prevista.

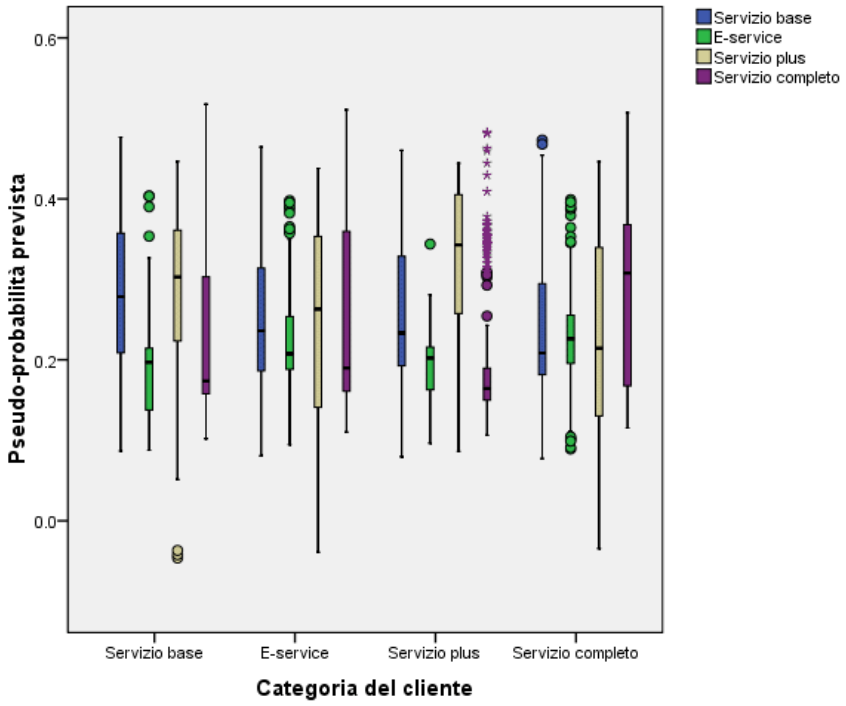
- Le celle sulla diagonale rappresentano previsioni corrette.
- Le celle fuori dalla diagonale rappresentano previsioni non corrette.

Dati i dati osservati, il modello "nullo" (ovvero quello privo di predittori), classificherebbe tutti i clienti nel gruppo modale *Servizio Plus*. Quindi, il modello nullo sarebbe corretto 281/1000 = 28,1% delle volte. La rete RBF ottiene il 10,1% in più o il 38,2% dei clienti. Il modello si rivela superiore soprattutto nell'identificazione dei clienti *Servizio plus* e *Servizio totale*. ma esegue in modo poco soddisfacente la classificazione dei clienti del gruppo *Servizi elettronici*. Al fine di separare questi clienti, può essere necessario individuare un altro predittore; in alternativa, dato che tali clienti sono molto spesso classificati in modo errato come *Servizio plus* e *Servizio totale*, la società può semplicemente tentare di eseguire l'upsell dei potenziali clienti che solitamente dovrebbero rientrare nella categoria *Servizio elettronico*.

Le classificazioni basate sui casi utilizzati per creare il modello tendono a essere troppo "ottimistiche" in quanto il tasso di classificazione è gonfiato. Il campione di controllo consente di convalidare il modello; in questo caso il 40,2% di questi casi sono stati classificati correttamente dal modello. Sebbene il campione di controllo sia limitato, ciò lascia intendere che complessivamente il modello è corretto circa due volte su cinque.

Grafico previsioni e osservazioni

Figura 5-10
grafico previsioni e osservazioni



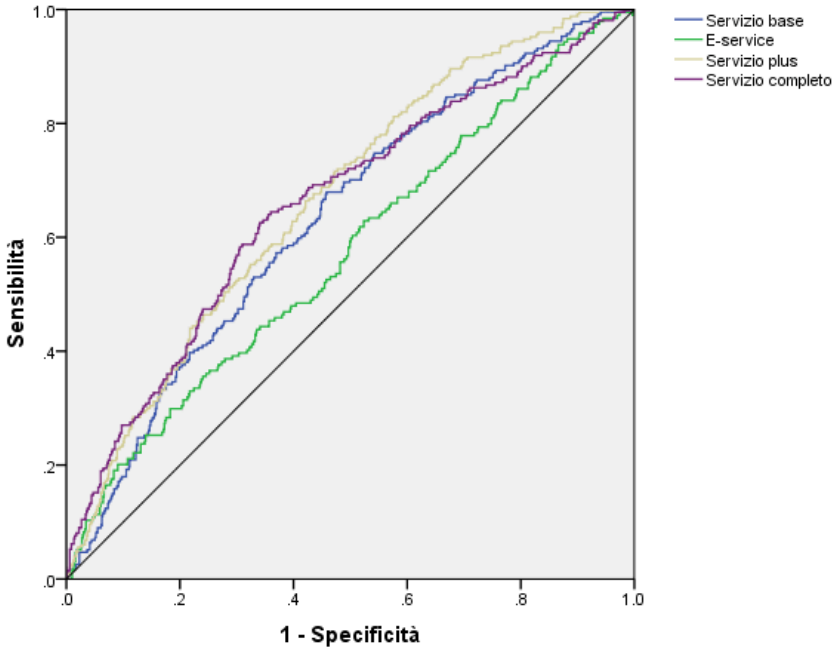
Per le variabili dipendenti categoriali, il Grafico previsioni e osservazioni visualizza grafici a scatole delle pseudo-probabilità previste dei campioni di verifica e di addestramento combinati. L'asse *X* corrisponde alle categorie di risposta osservate e la legenda corrisponde alle categorie previste. Quindi:

- Il grafico a scatole di sinistra mostra, per i casi con la categoria osservata *Servizio base*, la pseudo-probabilità prevista della categoria *Servizio base*.
- Il grafico a scatole successivo mostra, per i casi con la categoria osservata *Servizio base*, la pseudo-probabilità prevista della categoria *Servizio elettronico*.
- Il terzo grafico a scatole mostra, per i casi con la categoria osservata *Servizio base*, la pseudo-probabilità prevista della categoria *Servizio plus*. Dalla tabella di classificazione si evince che il numero di clienti *Servizio base* che è stato classificato in modo errato come *Servizio plus* corrisponde al numero di clienti correttamente classificati come *Servizio base*; quindi, questo grafico a scatole è abbastanza equivalente a quello di sinistra.
- Il quarto grafico a scatole mostra, per i casi con la categoria osservata *Servizio base*, la pseudo-probabilità prevista della categoria *Servizio totale*.

Poiché nella variabile di destinazione sono presenti più di due categorie, i primi quattro grafici a scatole non sono simmetrici rispetto alla linea orizzontale in 0,5, né in nessun altro modo. L'interpretazione di questo grafico per le destinazioni con più di due categorie può essere pertanto difficile, poiché è impossibile determinare, osservando una parte dei casi in un grafico a scatole, la posizione corrispondente di tali casi in un altro grafico a scatole.

Curva ROC

Figura 5-11
Curva ROC



Variabile dipendente: Categoria del cliente

Una curva ROC fornisce una visualizzazione della **sensibilità** per **specificità** per tutti i valori di riferimento possibili per la classificazione. Il grafico mostrato di seguito mostra quattro curve, una per ogni categoria della variabile di destinazione.

Tenere presente che questo grafico si basa sui campioni di addestramento e test. Per generare un grafico ROC per il campione di controllo, suddividere il file in base alla variabile di partizione ed eseguire la procedura relativa alla curva ROC per le pseudo-probabilità previste.

Figura 5-12
Area sotto la curva

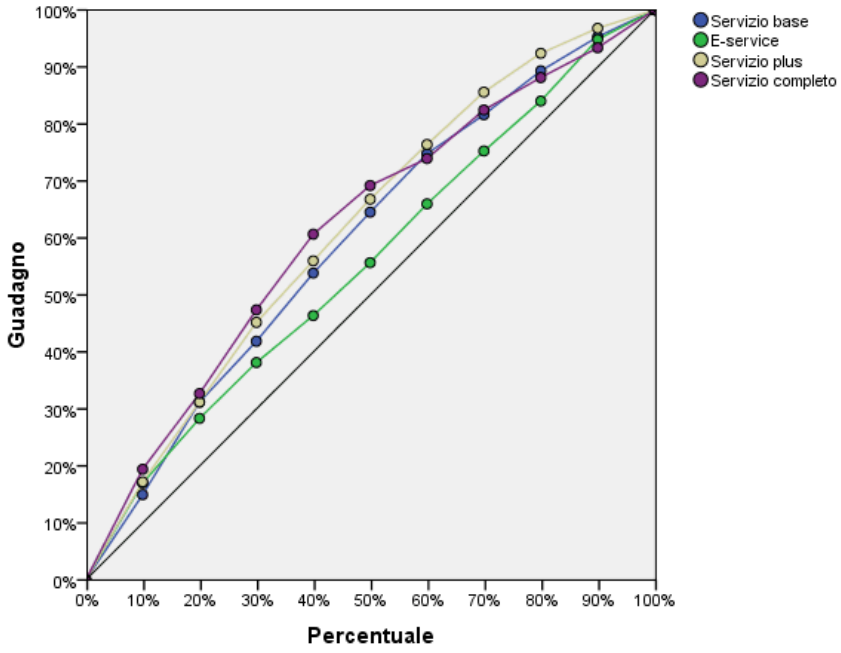
		Area
Categoria del cliente	Servizio base	,645
	E-service	,549
	Servizio plus	,696
	Servizio completo	,691

L'area sotto la curva è un riepilogo numerico della curva ROC e i valori nella tabella rappresentano, per ogni categoria, la probabilità che la pseudo-probabilità prevista si trovi in tale categoria è maggiore per un caso scelto in modo casuale in tale categoria rispetto a un caso scelto in modo casuale ma che non si trova in tale categoria. Ad esempio, per un cliente scelto in modo casuale in *Servizio plus* e in *Servizio base*, *Servizio elettronico* o *Servizio totale*, esiste

una probabilità di 0,668 che la pseudo-probabilità prevista per il modello di inadempienza sia maggiore per il cliente in *Servizio plus*.

Grafici dei guadagni cumulativi e lift

Figura 5-13
Grafico dei guadagni cumulativi

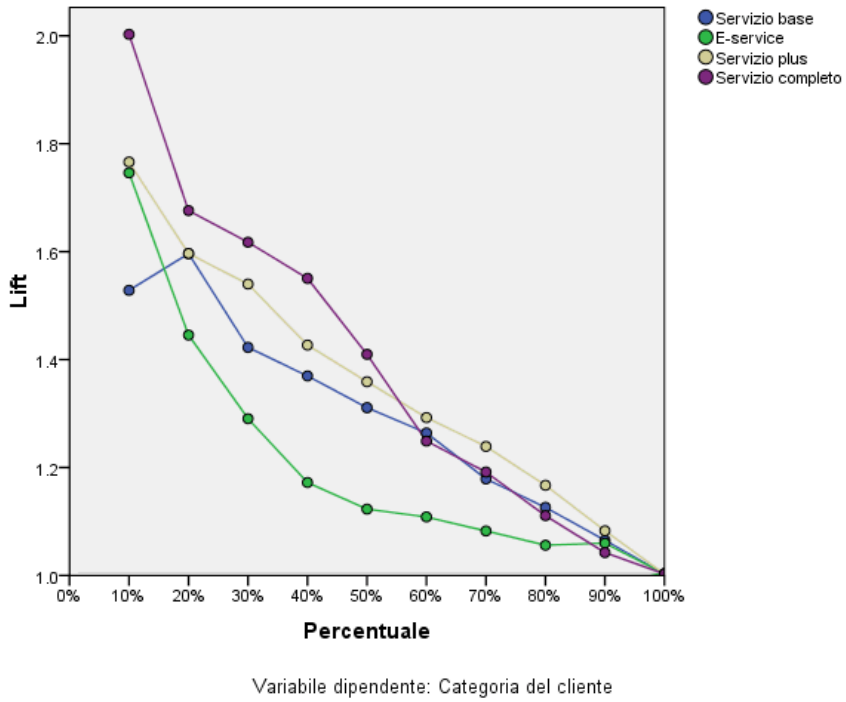


Variabile dipendente: Categoria del cliente

Il grafico dei guadagni cumulativi mostra la percentuale del numero totale dei casi in una determinata categoria ottenuta definendo come target una percentuale del numero totale dei casi. Ad esempio, il primo punto nella curva per la categoria *Servizio totale* è circa al 10%, 20% e questo significa che se si valuta un insieme di dati con la rete e si ordinano tutti i casi per la pseudo-probabilità prevista di *Servizio totale*, ci si dovrebbe aspettare che il primo 10% contenga circa il 20% di tutti i casi che fanno effettivamente parte della categoria *Servizio totale*. Analogamente, i primo 20% dovrebbe contenere circa il 30% degli inadempienti, il primo 30% dei casi il 50% degli inadempienti e così via. Se si seleziona il 100% dell'insieme dei dati valutato, si ottengono tutti gli inadempienti nell'insieme di dati.

La linea diagonale è la curva di base; se si seleziona il 10% dei casi dall'insieme di dati valutato in modo casuale, ci si dovrebbe aspettare di ottenere circa il 10% di tutti i casi che effettivamente fanno parte di qualsiasi categoria. Più in alto si trova la linea di base in cui si trova una curva, maggiore è il guadagno.

Figura 5-14
Grafico lift



Variabile dipendente: Categoria del cliente

Il grafico lift deriva dal grafico dei guadagni cumulativi; i valori sull'asse *Y* corrispondono al rapporto del guadagno cumulativo per ogni curva con la linea di base. Quindi, il passaggio al 10% per la categoria *Servizio totale* è circa il 20%/10% = 2. Consente di esaminare in modo differente le informazioni nel grafico dei guadagni cumulativi.

Nota: Il grafico dei guadagni cumulativi e il grafico lift si basano sui campioni di verifica e di addestramento combinati.

Letture consigliate

Consultare i testi seguenti per ulteriori informazioni sulle Funzioni a base radiale:

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

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

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

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

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

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

Appendice

A

File di esempio

Il file di esempio installato con il prodotto si trova nella sottodirectory *Samples* della directory di installazione. La sottodirectory Samples contiene cartelle separate per ciascuna delle seguenti lingue: Inglese, Francese, Tedesco, Italiano, Giapponese, Coreano, Polacco, Russo, Cinese semplificato, Spagnolo e Cinese tradizionale.

Non tutti i file di esempio sono disponibili in tutte le lingue. Se un file di esempio non è disponibile in una lingua, la cartella di tale lingua contiene una versione inglese del file.

Descrizioni

Questa sezione contiene brevi descrizioni dei file di esempio utilizzati negli esempi riportati in tutta la documentazione.

- **accidents.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nello studio dei fattori di rischio correlati all'età e al sesso per gli incidenti automobilistici che si verificano in una determinata regione. Ciascun caso corrisponde a una classificazione incrociata della categoria relativa età e del sesso.
- **adl.sav.** File di dati ipotetici che prende in esame l'impegno richiesto per determinare i vantaggi di un tipo di terapia proposto per i pazienti con problemi di cuore. I medici hanno assegnato in modo casuale i pazienti con problemi di cuore di sesso femminile a uno di due gruppi. Al primo gruppo è stata assegnata la terapia fisica standard; al secondo gruppo, un'ulteriore terapia di supporto psicologico. Dopo tre mesi di trattamenti, a ciascuna capacità dei pazienti che consente di riprendere le normali attività giornaliere è stato assegnato un punteggio come variabile ordinale.
- **advert.sav.** File di dati ipotetici che prende in esame l'impegno di un rivenditore al dettaglio che desidera esaminare la relazione tra il denaro speso per la pubblicità e le vendite risultanti. Finora sono stati raccolti i dati delle vendite precedenti e i relativi costi pubblicitari.
- **aflatoxin.sav.** File di dati ipotetici che prende in esame il test di raccolti di mais con presenza di Aflatossina, un veleno la cui concentrazione varia notevolmente nei raccolti. Una macchina per la lavorazione dei cereali ha ricevuto 16 campioni da ciascuno degli otto raccolti di mais e ha misurato i livelli di Aflatossina in parti per miliardo (PPB).
- **anorectic.sav.** Per trovare una sintomatologia standardizzata del comportamento anoressico/bulimico, i ricercatori (Van der Ham, Meulman, Van Strien, e Van Engeland, 1997) hanno condotto uno studio basato su 55 adolescenti affetti da disordini alimentari conosciuti. Ogni paziente è stato visitato quattro volte in quattro anni, per un totale di 220 visite. Durante ogni visita, ai pazienti sono stati assegnati punteggi per ciascuno dei 16 sintomi. I punteggi relativi ai sintomi sono assenti per il paziente 71 alla visita 2, il paziente 76 alla visita 2 e il paziente 47 alla visita 3, con 217 osservazioni valide.
- **bankloan.sav.** File di dati ipotetici che prende in esame l'impegno di una banca nel tentativo di ridurre il tasso di inadempienza nel rimborso di un prestito. Il file contiene informazioni finanziarie e demografiche su 850 vecchi e potenziali clienti. I primi 700 casi riguardano i

clienti a cui sono stati concessi dei prestiti precedentemente. Gli ultimi 150 casi riguardano i potenziali clienti che la banca deve classificare come rischi di credito positivi o negativi.

- **bankloan_binning.sav.** File di dati ipotetici che contiene informazioni finanziarie e demografiche su 5000 vecchi clienti.
- **behavior.sav.** In un classico esempio (Prezzo e Bouffard, 1974), è stato chiesto a 52 studenti di classificare una combinazione di 15 situazioni e 15 comportamenti utilizzando una scala da 0="molto appropriato" a 9="molto inadeguato". I valori medi riferiti ai partecipanti sono stati considerati dissimilarità.
- **behavior_ini.sav.** Questo file di dati contiene la configurazione iniziale di una soluzione a due dimensioni per *behavior.sav*.
- **brakes.sav.** File di dati ipotetici che prende in esame il controllo di qualità di un'industria che produce freni a disco per automobili con elevate prestazioni. Il file di dati contiene le misurazioni del diametro di 16 dischi da ciascuna delle otto macchine di produzione. L'obiettivo finale è ottenere un diametro dei dischi pari a 322 millimetri.
- **breakfast.sav.** In uno studio classico (Green e Rao, 1972), è stato chiesto a 21 studenti MBA della Wharton School e ai loro consorti di classificare 15 cibi da colazione in ordine di preferenza, dove il valore 1 corrispondeva all'alimento preferito in assoluto e il valore 15 a quello meno preferito. Le loro preferenze sono state registrate per sei diversi scenari, che comprendevano tutti gli scenari compresi tra "Preferenza generale" e "Solo snack con bibita".
- **breakfast-overall.sav.** Questo file contiene le preferenze degli alimenti della colazione solo per il primo scenario, "Preferenza generale".
- **broadband_1.sav.** File di dati ipotetici che contiene il numero di sottoscrittori, per area, di un provider di servizi a banda larga nazionale. Il file di dati contiene il numero dei sottoscrittori mensili di 85 aree in un periodo di quattro anni.
- **broadband_2.sav.** Questo file è identico al file *broadband_1.sav*, ma contiene i dati per ulteriori tre mesi.
- **car_insurance_claims.sav.** Un insieme di dati presentato e analizzato altrove (McCullagh e Nelder, 1989) riguarda le richieste di risarcimento auto. La quantità media di richieste di risarcimento può essere adattata come avente una distribuzione gamma, utilizzando una funzione di collegamento inverso per correlare la media della variabile dipendente a una combinazione lineare di età del contraente della polizza e tipo e anni del veicolo. Il numero delle richieste di risarcimento specificato può essere utilizzato come peso scalato.
- **car_sales.sav.** Questo file di dati ipotetici contiene le stime sulle vendite, i prezzi di listino e le specifiche fisiche di numerose marche e modelli di veicoli. I prezzi di listino e le specifiche fisiche sono state ottenute dal sito *edmunds.com* e dai siti dei produttori.
- **car_sales_uprepared.sav.** Questa è una versione modificata di *car_sales.sav* che non comprende versioni trasformate dei campi.
- **carpet.sav.** Come esempio tipico (Green e Wind, 1973), un'azienda interessata alla commercializzazione di un nuovo battitappeto desidera esaminare l'influenza di cinque fattori sulle preferenze del consumatore, ovvero design della confezione, marca, prezzo, la presenza di un *marchio di qualità* e una garanzia "Soddisfatti o rimborsati". Esistono tre livelli di fattore per il design della confezione, che differiscono per la posizione della spazzola dell'applicatore; tre marchi (*K2R*, *Glory* e *Bissell*); tre livelli di prezzo e due livelli (no o sì) per ciascuno degli ultimi due fattori. Dieci consumatori sono classificati in 22 profili definiti

da questi fattori. La variabile *Preferenza* include il rango delle classificazioni medie per ogni profilo. Classificazioni basse corrispondono a una preferenza elevata. La variabile riflette una misura globale della preferenza per ogni profilo.

- **carpet_prefs.sav.** Questo file di dati si basa sullo stesso esempio del file *carpet.sav*, ma contiene le classificazioni effettive raccolte da ciascuno dei 10 clienti. Ai clienti è stato chiesto di classificare 22 profili di prodotti in ordine di preferenza. Le variabili da *PREF1* a *PREF22* contengono gli ID dei profili associati, come definito nel file *carpet_plan.sav*.
- **catalog.sav.** File di dati ipotetico che contiene le cifre sulle vendite mensili di tre prodotti venduti da una società di vendita per corrispondenza. Il file include anche i dati di cinque possibili variabili predittore.
- **catalog_seasfac.sav.** Questo file di dati è uguale al file *catalog.sav* con l'eccezione che contiene un insieme di fattori stagionali calcolati dalla procedura Decomposizionale stagionale insieme a variabili di dati.
- **cellular.sav.** File di dati ipotetici che prende in esame l'impegno di un'azienda di telefonia cellulare nel tentativo di ridurre il churn, ovvero l'abbandono dei clienti. Agli account vengono applicati i punteggi relativi alla propensione al churn, con valori compresi tra 0 e 100. Gli account con punteggio pari a 50 o superiore è probabile che stiano cercando nuovi provider.
- **ceramics.sav.** File di dati ipotetici che prende in esame l'impegno di un produttore che desidera stabilire se una nuova lega premium ha una maggiore resistenza al calore rispetto alla lega standard. Ciascun caso rappresenta il test separato di una delle leghe. È indicata la temperatura massima alla quale può essere sottoposto il cuscinetto.
- **cereal.sav.** File di dati ipotetici che prende in esame le preferenze relative agli alimenti della colazione di un campione di 880 persone. Il file riporta anche l'età, il sesso e lo stato civile del campione e se le persone conducono uno stile di vita attivo (in base a un'attività sportiva con frequenza di due volte alla settimana). Ogni caso rappresenta un rispondente separato.
- **clothing_defects.sav.** File di dati ipotetici che prende in esame il processo di controllo di qualità di un'industria di abbigliamento. Per ciascun lotto prodotto nella fabbrica, gli ispettori prelevano un campione di abiti per contare il numero dei capi che non sono accettabili per la vendita.
- **coffee.sav.** Questo file di dati contiene informazioni sulle immagini percepite di sei marche di caffè freddo (Kennedy, Riquier, e Sharp, 1996). Per ciascuno dei 23 attributi dell'immagine del caffè freddo, sono state selezionate tutte le marche descritte da tale attributo. Le sei marche sono indicate dalle sigle AA, BB, CC, DD, EE e FF per tutelare la confidenzialità dei dati.
- **contacts.sav.** File di dati ipotetici che prende in esame l'elenco dei contatti di un gruppo di rappresentanti di vendita di computer aziendali. Ciascun contatto è classificato in base al reparto della società in cui lavora e dalle relative categorie aziendali. Il file riporta anche l'importo dell'ultima vendita effettuata, il tempo trascorso dall'ultima vendita e le dimensioni della società del contatto.
- **creditpromo.sav.** File di dati ipotetici che prende in esame l'impegno di un grande magazzino nel tentativo di valutare l'efficacia di una recente promozione con carta di credito. A tale scopo, sono stati selezionati 500 titolari di carta in modo casuale. Alla metà di questi è stato inviato un annuncio promozionale che comunica la riduzione del tasso d'interesse nel caso di acquisti effettuati entro i tre mesi successivi. All'altra metà è stato inviato un annuncio stagionale standard.

- **customer_dbase.sav.** File di dati ipotetico che prende in esame l'impegno di una società nel tentativo di utilizzare le informazioni contenute nel proprio database dei dati per creare offerte speciali per i clienti che più probabilmente risponderanno all'offerta. È stato selezionato in modo casuale un sottoinsieme della base dei clienti a cui è stata inviata l'offerta speciale e sono state registrate le risposte ricevute.
- **customer_information.sav.** File di dati ipotetici contenente le informazioni postali del cliente, ad esempio il nome e l'indirizzo.
- **customer_subset.sav.** Un sottoinsieme di 80 casi da *customer_dbase.sav*.
- **debate.sav.** File di dati ipotetici che prende in esame le risposte appaiate a un'indagine da parte dei partecipanti a un dibattito politico prima e dopo il dibattito. Ogni caso rappresenta un rispondente separato.
- **debate_aggregate.sav.** File di dati ipotetici che aggrega le risposte contenute nel file *debate.sav*. Ciascun caso corrisponde a una classificazione incrociata della preferenza prima e dopo il dibattito.
- **demo.sav.** File di dati ipotetici che prende in esame un database di clienti che hanno fatto acquisti al fine di inviare offerte mensili tramite il metodo del direct mailing. Viene registrata la risposta dei clienti, sia che abbiano aderito all'offerta o meno, insieme a diverse informazioni demografiche.
- **demo_cs_1.sav.** File di dati ipotetici che prende in esame il primo passo che una società intraprende per compilare un database con informazioni ricavate dai sondaggi. Ogni caso rappresenta una diversa città. Sono registrate anche le informazioni sulla regione, provincia, distretto e città.
- **demo_cs_2.sav.** File di dati ipotetici che prende in esame il secondo passo che una società intraprende per compilare un database con informazioni ricavate dai sondaggi. Ogni caso rappresenta una diversa unità di abitazione, ricavata dalle città selezionate nel primo passo. Sono registrate anche le informazioni sulla regione, provincia, distretto, città, suddivisione e unità. Il file include inoltre informazioni sul campionamento ottenute dai primi due stadi del disegno.
- **demo_cs.sav.** File di dati ipotetici che contiene informazioni sulle indagini raccolte utilizzando un disegno di campionamento complesso. Ogni caso rappresenta una diversa unità di abitazione. Sono registrate diverse informazioni demografiche e sul campionamento.
- **dmdata.sav.** File di dati ipotetici che contiene informazioni demografiche e di acquisto di una società di direct marketing. *dmdata2.sav* contiene informazioni su un sottoinsieme di contatti che hanno ricevuto un mailing di prova e *dmdata3.sav* contiene informazioni sui contatti rimanenti che non hanno ricevuto il mailing di prova.
- **dietstudy.sav.** File di dati ipotetici che contiene il risultato di uno studio ipotetico sulla dieta chiamato "Stillman diet" (Rickman, Mitchell, Dingman, e Dalen, 1974). Ogni caso rappresenta un diverso soggetto e ne riporta il peso prima e dopo la dieta in libbre e i livelli dei trigliceridi in mg/100 ml.
- **dvdplayer.sav.** File di dati ipotetici che prende in esame lo sviluppo di un nuovo lettore DVD. Utilizzando un prototipo, il personale addetto al marketing ha raccolto dati sui gruppi di interesse. Ogni caso rappresenta un diverso utente che è stato sottoposto all'indagine e include informazioni demografiche personali dell'utente e sulle risposte che ha fornito riguardo al prototipo.

- **german_credit.sav.** Questo file di dati contiene informazioni ricavate dall'insieme di dati "German Credit" del Repository of Machine Learning Databases (Blake e Merz, 1998) presso la University of California, Irvine.
- **grocery_1month.sav.** Questo file di dati ipotetici corrisponde al file di dati *grocery_coupons.sav* con gli acquisti settimanali organizzati in modo che ogni caso corrisponda a un cliente separato. Alcune delle variabili che cambiano settimanalmente non vengono riportate nei risultati; l'importo speso registrato corrisponde ora alla somma degli importi spesi durante le quattro settimane dello studio.
- **grocery_coupons.sav.** File di dati ipotetici che contiene i dati sui sondaggi raccolti da una catena di drogherie interessata alle abitudini di acquisto dei suoi clienti. Ciascun cliente viene seguito per quattro settimane e ciascun caso corrisponde a una settimana per cliente con informazioni sul luogo degli acquisti e i tipi di acquisti, incluso l'importo speso nelle drogherie durante la settimana.
- **guttman.sav.** Bell (Bell, 1961) ha presentato una tabella per illustrare i possibili gruppi sociali. Guttman (Guttman, 1968) ha utilizzato una parte di tale tabella, in cui cinque variabili che descrivono elementi come l'interazione sociale, i sentimenti di appartenenza a un gruppo, la vicinanza fisica dei membri e il grado di formalità della relazione, sono state incrociate con cinque gruppi sociali teorici, compresi folla (ad esempio, le persone presenti a una partita di calcio), uditorio (ad esempio, di uno spettacolo teatrale o di una lezione universitaria), pubblico (ad esempio televisivo), calca (come una folla, ma con un'interazione molto maggiore), gruppi primari (intimi), gruppi secondari (volontari) e la comunità moderna (unione non stretta derivante da una vicinanza fisica elevata e dall'esigenza di servizi specializzati).
- **health_funding.sav.** File di dati ipotetici che contiene i dati sui fondi di assistenza sanitaria (importo per 100 persone), sui tassi di malattie (tasso per 10.000 persone) e sulle visite ai fornitori di assistenza sanitaria (tasso per 10.000 persone). Ogni caso rappresenta una diversa città.
- **hivassay.sav.** File di dati ipotetici che prende in esame l'impegno di un'industria farmaceutica nel tentativo di sviluppare un'analisi che riesca a rilevare in tempi brevi l'infezione da virus HIV. I risultati dell'analisi sono otto sfumature di colore rosso sempre più intenso; le sfumature più intense indicano la maggiore probabilità di infezione. Un esperimento di laboratorio è stato condotto su 2000 campioni di sangue. La metà di questi è risultata infetta al virus HIV, l'altra metà non è risultata infetta.
- **hourlywagedata.sav.** File di dati ipotetici che prende in esame la paga oraria degli infermieri occupati presso uffici e ospedali e in base ai diversi livelli di esperienza.
- **insurance_claims.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nella creazione di un modello per contrassegnare le richieste di risarcimento sospette e potenzialmente fraudolente. Ogni caso rappresenta una richiesta di risarcimento separata.
- **insure.sav.** File di dati ipotetici che prende in esame una compagnia di assicurazioni impegnata nello studio dei fattori di rischio, che indicano l'eventualità che un cliente presenti una domanda di indennizzo in un contratto assicurativo sulla vita della durata di dieci anni. Ogni caso nel file di dati rappresenta una coppia di contratti. In un contratto sono contenute informazioni su una richiesta di risarcimento, l'altro sull'età e sul sesso.

- **judges.sav.** File di dati ipotetici che prende in esame il punteggio assegnato, da giurie qualificate (più un appassionato) a 300 prestazioni sportive. Ciascuna riga rappresenta una diversa prestazione; i giudici hanno esaminato le stesse prestazioni.
- **kinship_dat.sav.** Rosenberg e Kim (Rosenberg e Kim, 1975) si prefiggono di analizzare 15 termini indicanti parentela (zia, fratello, cugino, padre, nipote femmina, di nonni, nonno, nonna, nipote maschio di nonni, madre, nipote maschio di zii), nipote femmina di zii, sorella, figlio, zio). Hanno richiesto a quattro gruppi di studenti universitari (due composti da femmine e due da maschi) di ordinare questi termini in base alla similiarità. A due gruppi (uno femminile e uno maschile) è stato richiesto di effettuare l'ordinamento due volte, con il secondo ordinamento basato su un criterio diverso rispetto al primo. Di conseguenza, sono state ottenute sei "sorgenti" in totale. Ogni sorgente corrisponde a una matrice di prossimità 15×15, le cui celle sono uguali al numero delle persone in una sorgente meno il numero di volte in cui gli oggetti sono stati ripartiti insieme nella sorgente.
- **kinship_ini.sav.** Questo file di dati contiene la configurazione iniziale di una soluzione a tre dimensioni per *kinship_dat.sav*.
- **kinship_var.sav.** Questo file di dati contiene variabili indipendenti relative a *sesso*, *generazione* e *grado* di separazione che possono essere utilizzate per interpretare le dimensioni di una soluzione per *kinship_dat.sav*. In modo specifico, tali variabili possono essere utilizzate per limitare lo spazio della soluzione a una combinazione lineare di tali variabili.
- **marketvalues.sav.** File di dati che prende in esame le vendite di abitazioni in un nuovo centro abitato in Algonquin, Ill., durate gli anni 1999–2000. Tali vendite sono una questione di dominio pubblico.
- **nhis2000_subset.sav.** Il National Health Interview Survey (NHIS) è un sondaggio di grandi dimensioni condotto sulla popolazione civile americana. Le interviste vengono realizzate di persona e si basano su un campione rappresentativo di famiglie a livello nazionale. Per ogni membro di una famiglia vengono raccolte osservazioni e informazioni di carattere demografico relative allo stato di salute. Questo file di dati contiene un sottoinsieme delle informazioni ottenute dall'indagine del 2000. National Center for Health Statistics. National Health Interview Survey, 2000. File di dati e documentazione di dominio pubblico. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accesso 2003.
- **ozone.sav** I dati includono 330 osservazioni basate su sei variabili meteorologiche per quantificare la concentrazione dell'ozono dalle variabili rimanenti. I precedenti ricercatori, (Breiman e Friedman, 1985) e (Hastie e Tibshirani, 1990), hanno rilevato non linearità tra queste variabili, che impediscono un approccio di regressione standard.
- **pain_medication.sav.** File di dati ipotetici che contiene i risultati di un test clinico per stabilire la cura antinfiammatoria per il trattamento del dolore generato dall'artrite cronica. Di particolare interesse, il test ha evidenziato il tempo che impiega il farmaco ad avere effetto e il confronto con altri farmaci esistenti.
- **patient_los.sav.** File di dati ipotetici che contiene informazioni sul trattamento dei pazienti ricoverati per sospetto di infarto del miocardio. Ogni caso corrisponde a un diverso paziente e contiene diverse variabili correlate alla degenza nell'ospedale.
- **patlos_sample.sav.** File di dati ipotetici che contiene informazioni sul trattamento di un campione di pazienti curato con trombolitici durante la degenza per infarto del miocardio. Ogni caso corrisponde a un diverso paziente e contiene diverse variabili correlate alla degenza nell'ospedale.

- **poll_cs.sav.** File di dati ipotetici che prende in esame i sondaggi per stabilire il livello di sostegno pubblico nei confronti di un disegno di legge prima che diventi una legge vera e propria. I casi corrispondono ai votanti registrati. Ciascun caso riporta informazioni sulla contea, sul comune e sul quartiere in cui vive il votante.
- **poll_cs_sample.sav.** File di dati ipotetici che contiene un campione dei votanti elencati nel file *poll_cs.sav*. Il campione è stato selezionato in base al disegno specificato nel file di piano *poll.csplan* e questo file di dati contiene le probabilità di inclusione e i pesi del campione. Tuttavia, notare che poiché fa uso del metodo PPS (probability-proportional-to-size, probabilità proporzionale alla dimensione), esiste anche un file contenente le probabilità di selezione congiunte (*poll_jointprob.sav*). Le ulteriori variabili corrispondenti ai dati demografici dei votanti e alla loro opinione sul disegno di legge, sono state raccolte e aggiunte al file di dati dopo aver acquisito il campione.
- **property_assess.sav.** File di dati ipotetici che prende in esame l'impegno di un perito di una contea nel tentativo di mantenere gli accertamenti sui valori delle proprietà aggiornati in base alle risorse limitate. I casi rappresentano le proprietà vendute nella contea nello scorso anno. Ogni caso nel file di dati contiene informazioni sul comune in cui si trova la proprietà, il perito che per ultimo ha visitato la proprietà, il tempo trascorso dall'accertamento, la valutazione fatta in tale momento e il valore di vendita della proprietà.
- **property_assess_cs.sav.** File di dati ipotetici che prende in esame l'impegno di un perito di uno stato nel tentativo di mantenere aggiornati gli accertamenti sui valori delle proprietà in base alle risorse limitate. I casi corrispondono alle proprietà nello stato. Ogni caso nel file di dati include informazioni sulla contea, il comune e il quartiere in cui risiede la proprietà, la data dell'ultimo accertamento e la valutazione fatta in tale data.
- **property_assess_cs_sample.sav.** File di dati ipotetici che contiene un campione delle proprietà elencate nel file *property_assess_cs.sav*. Il campione è stato selezionato in base al disegno specificato nel file di piano *property_assess.csplan* e questo file di dati contiene le probabilità di inclusione e i pesi del campione. L'ulteriore variabile *Valore corrente* è stata raccolta e aggiunta al file di dati dopo aver acquisito il campione.
- **recidivism.sav.** File di dati ipotetici che prende in esame l'impegno delle Forze dell'Ordine nel tentativo di valutare il tasso di recidività nella propria area di giurisdizione. Ogni caso corrisponde a un precedente trasgressore e include le informazioni demografiche, alcuni dettagli sul primo crimine, il tempo trascorso fino al secondo arresto e se tale arresto è avvenuto entro due anni dal primo.
- **recidivism_cs_sample.sav.** File di dati ipotetici che prende in esame l'impegno delle Forze dell'Ordine nel tentativo di valutare il tasso di recidività nella propria area di giurisdizione. Ogni caso corrisponde a un trasgressore precedente, rilasciato dopo il primo arresto durante il mese di giugno del 2003 e registra le relative informazioni demografiche, alcuni dettagli sul primo crimine commesso e i dati del secondo arresto, se si è verificato prima della fine di giugno del 2006. I trasgressori sono stati selezionati dai dipartimenti sottoposti a campione in base al piano di campionamento specificato nel file *recidivism_cs.csplan*. Poiché viene utilizzato un metodo PPS (Probability-Proportional-to-Size, probabilità proporzionale alla dimensione), esiste anche un file contenente le probabilità di selezione congiunte (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** File di dati ipotetici contenente i dati delle transazioni di acquisto, inclusa la data di acquisto, gli articoli acquistati e il valore monetario di ciascuna transazione.

- **salesperformance.sav.** File di dati ipotetici che prende in esame la valutazione di due nuovi corsi di formazione alle vendite. Sessanta dipendenti, divisi in tre gruppi, ricevono tutti la formazione standard. In più, al gruppo 2 viene assegnato un corso di formazione tecnica e al gruppo 3 un'esercitazione pratica. Alla fine del corso di formazione, ciascun dipendente viene sottoposto a un esame e il punteggio conseguito viene registrato. Ciascun caso nel file di dati rappresenta un diverso partecipante. Il file di dati include il gruppo a cui è assegnato il partecipante e il punteggio conseguito all'esame finale.
- **satisf.sav.** File di dati ipotetico che prende in esame un'indagine sulla soddisfazione dei clienti condotta da una società di vendita al dettaglio presso 4 negozi. Sono stati intervistati 582 clienti e ciascun caso rappresenta le risposte ottenute da un singolo cliente.
- **screws.sav.** Questo file di dati contiene informazioni sulle caratteristiche di viti, bulloni, dadi e puntine (Hartigan, 1975).
- **shampoo_ph.sav.** File di dati ipotetici che prende in esame il processo di controllo di qualità di un'industria di prodotti per capelli. A intervalli di tempo regolari, vengono misurati sei diversi lotti prodotti e ne viene registrato il relativo pH. I valori accettati sono compresi tra 4,5 e 5,5.
- **ships.sav.** Ad esempio, un insieme di dati presentato e analizzato altrove (McCullagh et al., 1989) riguarda i danni subiti dalle navi da carico a causa delle onde. I conteggi degli incidenti possono essere presentati con un tasso di Poisson in base al tipo di nave, al periodo di costruzione e al periodo di servizio. I mesi di servizio aggregati di ciascuna cella della tabella generata dalla classificazione incrociata dei fattori fornisce i valori di esposizione al rischio.
- **site.sav.** File di dati ipotetici che prende in esame l'impegno di una società nella scelta di nuovi siti in cui espandere la propria presenza. La società ha incaricato due consulenti separati che, oltre a valutare i siti e presentare un report completo, devono classificarli come potenzialmente "molto adatti", "adatti" o "poco adatti".
- **smokers.sav.** Questo file di dati è un estratto del 1998 National Household Survey of Drug Abuse e rappresenta un campione probabile di famiglie americane. (*http://dx.doi.org/10.3886/ICPSR02934*) Il primo passo nell'analisi di questo file di dati consiste quindi nel pesare i dati per rispecchiare le tendenze della popolazione.
- **stocks.sav** Questo file di dati ipotetici contiene i prezzi e i volumi delle scorte riferiti a un anno.
- **stroke_clean.sav.** File di dati ipotetici che riporta lo stato di un database medico dopo averne eseguito la pulizia utilizzando le procedure del modulo Data Preparation.
- **stroke_invalid.sav.** File di dati ipotetici che riporta lo stato iniziale di un database medico e contiene numerosi errori di immissione dati.
- **stroke_survival.** Questo file di dati ipotetici riguarda i tempi di sopravvivenza per i pazienti che, dopo avere completato un programma riabilitativo in seguito a un ictus postischemico, affrontano alcune sfide. Dopo l'attacco, viene annotata l'occorrenza dell'infarto miocardiaco, dell'ictus ischemico o emorragico e viene registrata l'ora dell'evento. Questo campione viene troncato a sinistra perché include solo i pazienti che sono sopravvissuti fino alla fine del programma riabilitativo post-ictus.
- **stroke_valid.sav.** File di dati ipotetici che riporta lo stato di un database medico dopo il controllo dei valori eseguito con la procedura Convalida i dati. Il database contiene comunque casi potenzialmente anomali.

- **survey_sample.sav.** File di dati che contiene i dati dell'indagine, compresi i dati demografici e varie misure dell'atteggiamento. Si basa su un sottoinsieme di variabili tratte dal 1998 NORC General Social Survey, benché i valori di alcuni dati siano stati modificati e siano state aggiunte variabili fittizie a scopo dimostrativo.
- **telco.sav.** File di dati ipotetici che prende in esame l'impegno di un'azienda di telecomunicazioni nel tentativo di ridurre il churn, ovvero l'abbandono dei propri clienti. Ciascun caso rappresenta un cliente separato e riporta diverse informazioni demografiche e sull'uso del servizio.
- **telco_extra.sav.** Questo file di dati è simile al file *telco.sav*, ma le variabili "tenure" e spesa del cliente trasformata tramite logaritmo sono state sostituite dalle variabili di spesa del cliente trasformata tramite logaritmo standardizzate.
- **telco_missing.sav.** Questo file di dati è un sottoinsieme del file di dati *telco.sav*, ma alcuni dei valori di dati demografici sono stati sostituiti con valori mancanti.
- **testmarket.sav.** File di dati ipotetici che prende in esame i piani di una catena di fast food per aggiungere un nuovo prodotto al proprio menu. Sono previste tre campagne promozionali del nuovo prodotto. Il prodotto viene introdotto in diversi mercati selezionati in modo casuale. Per ogni sede viene utilizzata una promozione differente registrando le vendite settimanali della nuova voce per le prime quattro settimane. Ogni caso rappresenta un luogo e una settimana diversi.
- **testmarket_1month.sav.** Questo file di dati ipotetici corrisponde al file *testmarket.sav* con le vendite settimanali organizzate in modo che ogni caso corrisponda a un luogo separato. Alcune delle variabili che cambiano settimanalmente non vengono riportate nei risultati; le vendite registrate corrispondono ora alla somma delle vendite conseguite durante le quattro settimane dello studio.
- **tree_car.sav.** File di dati ipotetici che contiene dati demografici e sul prezzo di acquisto dei veicoli.
- **tree_credit.sav.** File di dati ipotetici che contiene dati demografici e sulla cronologia dei mutui di una banca.
- **tree_missing_data.sav.** File di dati ipotetici che contiene dati demografici e sulla cronologia dei mutui di una banca con un numero elevato di valori mancanti.
- **tree_score_car.sav.** File di dati ipotetici che contiene dati demografici e sul prezzo di acquisto dei veicoli.
- **tree_textdata.sav.** File di dati semplice con due variabili destinato principalmente per mostrare lo stato predefinito delle variabili prima dell'assegnazione dei livelli di misurazione e delle etichette dei valori.
- **tv-survey.sav.** File di dati ipotetici che prende in esame un sondaggio condotto da una emittente televisiva che deve stabilire se estendere la durata di un programma di successo. A un campione di 906 intervistati è stato chiesto se preferisce guardare il programma con diverse condizioni. Ciascuna riga rappresenta un diverso intervistato e ciascuna colonna una diversa condizione.
- **ulcer_recurrence.sav.** Questo file contiene informazioni parziali su uno studio svolto per mettere a confronto l'efficacia di due terapie preventive per la recidiva delle ulcere. Fornisce un ottimo esempio di dati acquisiti a intervalli ed è stato presentato e analizzato in altri luoghi (Collett, 2003).

- **ulcer_recurrence_recoded.sav.** In questo file sono contenute le informazioni del file *ulcer_recurrence.sav* riorganizzate per consentire di presentare la probabilità degli eventi per ciascun intervallo dello studio, anziché solo alla fine. È stato presentato e analizzato in altri luoghi (Collett et al., 2003).
- **verd1985.sav.** Questo file di dati prende in esame un'indagine (Verdegaal, 1985). Sono state registrate le risposte di quindici soggetti a otto variabili. Le variabili di interesse sono suddivise in tre insiemi. L'insieme 1 include *età* e *statociv*, l'insieme 2 include *andom* e *giornale* e l'insieme 3 include *musica* e *vicinato*. *Andom* viene scalata come nominale multipla ed *età* come ordinale; tutte le altre variabili vengono scalate come nominali singole.
- **virus.sav.** File di dati ipotetici che prende in esame l'impegno di un ISP (Internet Service Provider) nel tentativo di determinare gli effetti che un virus può generare nelle sue reti. Si è tenuta traccia della percentuale (approssimativa) di traffico e-mail infettato da virus sulla rete in un lasso di tempo, dal momento dell'individuazione fino alla soppressione della minaccia.
- **wheeze_steubenville.sav.** Questo file è un sottoinsieme di uno studio longitudinale degli effetti che l'inquinamento provoca sulla salute dei bambini (Ware, Dockery, Spiro III, Speizer, e Ferris Jr., 1984). I dati contengono misure binarie ripetute del livello di asma dei bambini della città di Steubenville, Ohio, di 7, 8, 9 e 10 anni. I dati indicano anche se la mamma dei bambini era fumatrice durante il primo anno dello studio.
- **workprog.sav.** File di dati ipotetici che prende in esame un programma di lavoro governativo il cui obiettivo è fornire attività più adatte alle persone diversamente abili. È stato seguito un campione di potenziali partecipanti al programma, alcuni dei quali sono stati selezionai in modo casuale e altri no. Ogni caso rappresenta un diverso partecipante al programma.
- **worldsales.sav** Questo file di dati ipotetici contiene i ricavi suddivisi per continenti e prodotti.

Appendice

B

Note legali

Queste informazioni sono state preparate per prodotti e servizi offerti in tutto il mondo.

IBM potrebbe non offrire i prodotti, i servizi o le funzionalità di cui si tratta nel presente documento in altri paesi. Contattare il rappresentante IBM locale per informazioni sui prodotti e i servizi attualmente disponibili nella propria zona. Qualsiasi riferimento a un prodotto, programma o servizio IBM non intende dichiarare o implicare che sia possibile utilizzare esclusivamente tale prodotto, programma o servizio IBM. Potrà invece essere utilizzato qualsiasi prodotto, programma o servizio con funzionalità equivalente e che non violi i diritti di proprietà intellettuale di IBM. Tuttavia, è responsabilità dell'utente valutare e verificare il funzionamento di qualsiasi prodotto, programma o servizio non IBM.

IBM può essere titolare di brevetti o domande di brevetto relativi alla materia oggetto del presente documento. La consegna del presente documento non conferisce alcuna licenza rispetto a questi brevetti. Rivolgere per iscritto i quesiti sulle licenze a:

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

Per richieste di informazioni sulle licenze riguardanti il set di caratteri a byte doppio (DBCS), contattare l'Intellectual Property Department di IBM del proprio paese, oppure inviare le richieste in forma scritta all'indirizzo:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Giappone.

Il seguente paragrafo non si applica per il Regno Unito o altri paesi in cui le presenti disposizioni non sono conformi alle leggi locali: INTERNATIONAL BUSINESS MACHINES FORNISCE QUESTA PUBBLICAZIONE "COSÌ COM'È" SENZA GARANZIA DI ALCUN TIPO, SIA ESSA ESPRESSA O IMPLICITA, INCLUSE, MA NON LIMITATE A, LE GARANZIE IMPLICITE DI NON VIOLAZIONE, COMMERCIABILITÀ O IDONEITÀ A UNO SCOPO SPECIFICO. Alcuni stati non consentono limitazioni di garanzie espresse o implicite in determinate transazioni, pertanto quanto sopra potrebbe non essere applicabile.

Le presenti informazioni possono includere imprecisioni tecniche o errori tipografici. Le modifiche periodiche apportate alle informazioni contenute in questa pubblicazione verranno inserite nelle nuove edizioni della pubblicazione. IBM può apportare miglioramenti e/o modifiche al/ai prodotto/i e/o al/ai programma/i descritti nella presente pubblicazione in qualsiasi momento senza preavviso.

Qualsiasi riferimento nelle presenti informazioni a siti Web non IBM viene fornito esclusivamente per facilitare la consultazione e non rappresenta in alcun modo un'approvazione o sostegno da parte nostra di tali siti Web. I materiali contenuti in tali siti Web non fanno parte dei materiali di questo prodotto IBM e il loro utilizzo è esclusivamente a rischio dell'utente.

IBM può utilizzare o distribuire eventuali informazioni fornite dall'utente nei modi che ritiene appropriati senza incorrere in alcun obbligo nei confronti dell'utente.

I licenziatari del programma che desiderassero informazioni su di esso allo scopo di abilitare: (i) lo scambio di informazioni tra programmi creati indipendentemente e altri programmi (questo compreso) e (ii) l'utilizzo in comune delle informazioni scambiate, dovranno rivolgersi a:

IBM Software Group, All'attenzione di: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Tali informazioni saranno fornite in conformità ai termini e alle condizioni in vigore e, in alcuni casi, dietro pagamento.

Il programma concesso in licenza descritto nel presente documento e tutto il materiale correlato disponibile sono forniti da IBM in base ai termini del contratto di licenza cliente IBM, del contratto di licenza internazionale IBM o del contratto equivalente esistente tra le parti.

le informazioni relative a prodotti non IBM sono state ottenute dai fornitori di tali prodotti, da loro annunci pubblicati e da altre fonti disponibili al pubblico. IBM non ha verificato tali prodotti e non può confermare l'accuratezza delle prestazioni, la compatibilità o qualsiasi altra dichiarazione relativa a prodotti non IBM. Eventuali domande in merito alle funzionalità dei prodotti non IBM vanno indirizzate ai fornitori di tali prodotti.

Le presenti informazioni includono esempi di dati e report utilizzati in operazioni aziendali quotidiane. Per fornire una descrizione il più possibile esaustiva, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti questi nomi sono fittizi e ogni somiglianza a nomi e indirizzi utilizzati da aziende reali è puramente casuale.

Per chi visualizza queste informazioni a video: le fotografie e le illustrazioni a colori potrebbero non essere disponibili.

Marchi commerciali

IBM, il logo IBM, ibm.com e SPSS sono marchi di IBM Corporation, registrati in numerose giurisdizioni nel mondo. Un elenco aggiornato dei marchi IBM è disponibile sul Web all'indirizzo *http://www.ibm.com/legal/copytrade.shtml*.

Adobe, il logo Adobe, PostScript e il logo PostScript sono marchi o marchi registrati di Adobe Systems Incorporated negli Stati Uniti e/o in altri paesi.

Intel, il logo Intel, Intel Inside, il logo Intel Inside, Intel Centrino, il logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium sono marchi o marchi registrati di Intel Corporation o delle sue consociate negli Stati Uniti e in altri paesi.

Java e tutti i marchi e i logo basati su Java sono marchi di Sun Microsystems, Inc. negli Stati Uniti e/o negli altri paesi.

Linux è un marchio registrato di Linus Torvalds negli Stati Uniti e/o negli altri paesi.

Microsoft, Windows, Windows NT e il logo Windows sono marchi di Microsoft Corporation negli Stati Uniti e/o negli altri paesi.

UNIX è un marchio registrato di The Open Group negli Stati Uniti e in altri paesi.

Questo prodotto utilizza WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Altri nomi di prodotti e servizi possono essere marchi commerciali di IBM o di altre aziende.

Le schermate dei prodotti Adobe sono state ristampate su autorizzazione di Adobe Systems Incorporated.

Le schermate dei prodotti Microsoft sono state ristampate su autorizzazione di Microsoft Corporation.

Bibliografia

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

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

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

Breiman, L., e J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

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

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

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

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

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

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

Hastie, T., e R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

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

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

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

Prezzo, R. H., e D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Rickman, R., N. Mitchell, J. Dingman, e J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

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

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

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

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

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

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

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

Indice