TPM®

CONFRONTO TRA UN MODELLO DI REGRESSIONE MULTIPLA E UN MODELLO CONNESSIONISTA PER LA PREVISIONE DELLA PRESTAZIONE UNIVERSITARIA

MARGHERITA PASINI
UNIVERSITÀ DI VERONA
MASSIMILIANO PASTORE
UNIVERSITÀ DI PADOVA

A comparison between a multiple regression and a connectionist model of forecasting the university performance.

The paper presents a comparison between two forecasting methods, the first based on a multiple regression model and the second based on a connectionist model applied to the problem of choosing bachelor course. The sample is a set of 250 students enrolled at the faculty of Economics of University of Padova in the academic year 1995/96. We investigated the relation between some explanatory variables (scores obtained at the admission test to the bachelor Course, school-leaving votes and others) and the number of passed exams together with exams mark mean accorpate in one variable. This variable allows to characterize three levels of performance. The comparison between the observed and estimated levels, carried out by making use of distribution free tests, has highlighted both substantial homogeneity of results and interesting peculiarities.

Il presente lavoro si propone di confrontare due metodi di previsione, uno che si basa su un modello di regressione multipla e uno che considera un modello connessionista, applicati in un contesto di orientamento alla scelta universitaria. Il campione è costituito da 250 studenti iscritti alla facoltà di Economia e Commercio di Padova nell'A.A. 1995/96. È stata studiata la relazione tra alcune variabili considerate come predittori (punteggi ottenuti al test di ammissione al Corso di laurea, voto di maturità ed altre) e il numero di esami sostenuti e la media dei voti accorpate in un'unica variabile tramite la quale sono stati individuati tre livelli di prestazione. Il confronto tra i livelli osservati e quelli attesi, effettuato con statistiche non parametriche, ha evidenziato una sostanziale omogeneità di risultati, ma ha anche messo in luce alcune interessanti peculiarità.

Key words: Multiple regression; Connectionist model; Contingency coefficient; Goodman-Kruskal's gamma; Performance.

INTRODUZIONE

Il presente lavoro si propone di confrontare due metodi di previsione, uno che si basa su un modello di regressione multipla e uno che considera un modello connessionista. Tali metodi vengono applicati per prevedere la prestazione di un soggetto iscritto all'università una volta note alcune caratteristiche. Obiettivo del lavoro non è valutare la "bontà" dei risultati ottenuti con i due metodi, quanto piuttosto individuare una modalità di confronto tra due procedure che si basano su approcci diversi. Nel modello di regressione vengono considerate poche variabili predittive e si suppone una relazione di tipo lineare tra queste e le dipendenti. Nel modello connessio-

nista invece si considerano molte variabili indipendenti e non si è vincolati ad un assunto di linearità.

Il campione analizzato si compone di 250 studenti iscritti alla facoltà di economia e commercio di Padova, immatricolati nell'anno accademico 1995/96. Questi studenti, per essere ammessi al Corso di laurea, hanno dovuto sostenere un test di selezione. In base alle domande su tale test possono essere individuate 10 aree (competenza semantica, cultura generale, ragionamento, calcolo, somiglianza lettere, completamento serie lettere o numeri, capacità visuospaziale, logica); il punteggio totale dipende dal risultato ottenuto su tali aree. Inoltre, sono note altre caratteristiche sia di tipo qualitativo (sesso, luogo di residenza, tipo di diploma), che di tipo quantitativo (voto di diploma, numero di esami sostenuti e media dei voti ai suddetti esami).

Gli indicatori scelti per valutare la prestazione sono due: la media dei voti ed il numero di esami effettivamente sostenuti. Le due variabili sono state sintetizzate in un unico indice, chiamato *prestazione*, che varia tra zero, per chi non ha sostenuto alcun esame, e uno, per chi invece ha superato tutti gli esami previsti con il massimo della media. Tale indice viene calcolato utilizzando il numero massimo di esami sostenuti nel campione considerato (14[1]) nel seguente modo:

$$prestazione = (\text{media} \times \text{numero di esami})/(30 \times 14) \quad (1)$$

In questo modo è possibile discriminare due soggetti che, a parità di media, hanno sostenuto un numero diverso di esami e di conseguenza hanno ottenuto un differente livello di prestazione. La Figura 1 mostra come si modificano i valori della variabile *prestazione* a seconda del numero di esami sostenuti, nel caso di media 18, 24 e 30.

Facendo riferimento alla distribuzione di frequenza della variabile *prestazione* sui 250 soggetti, sono state individuate tre fasce: i soggetti con punteggio fino al 33° percentile (.3976)

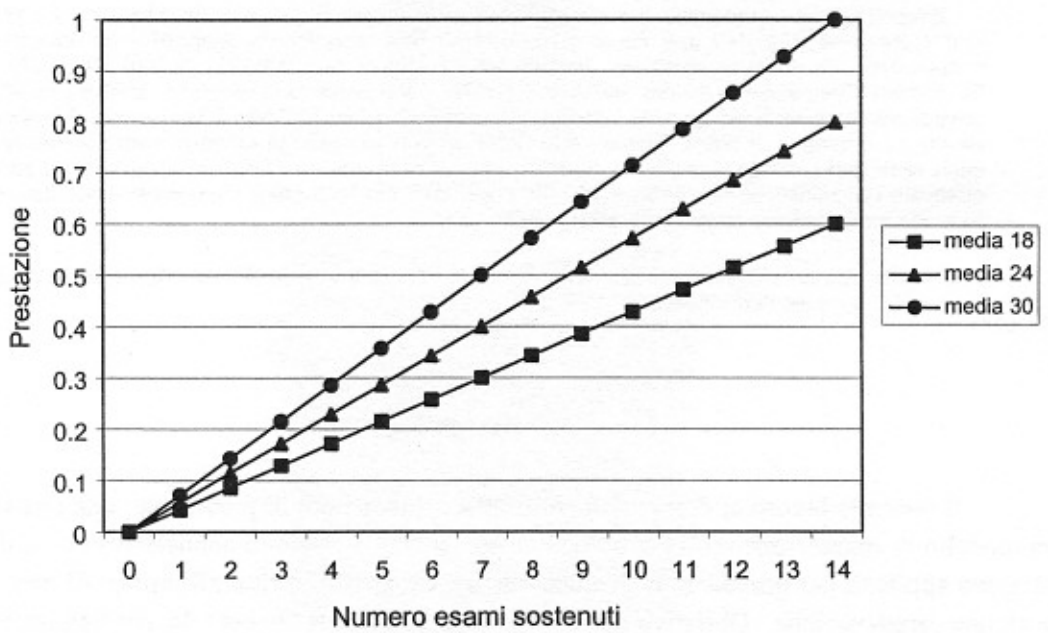

FIGURA 1
Andamento della variabile *prestazione* in funzione del numero di esami.

rappresentano un livello di prestazione basso, quelli dal 33° al 66° (.619) un livello di prestazione medio, quelli oltre il 66° percentile un livello di prestazione alto. Questa suddivisione ha permesso di giungere a dei risultati confrontabili tra i due modelli; per ciascun soggetto infatti, a fronte di un livello osservato di prestazione, vengono stimati due livelli attesi, uno secondo il modello di regressione ed uno secondo la rete neurale.

In relazione alle variabili sesso e diploma di maturità, il campione dei 250 studenti risulta composto nel modo riportato in Tabella 1.

Tabella 1
Soggetti del campione per sesso e diploma.

		Sesso		Totale
		Maschi	Femmine	
Maturità	Classica	6	9	15
	Scientifica	66	52	118
	Ragioneria	46	55	101
	Altro	9	7	16
Totale		127	123	250

Tale campione è stato suddiviso in tre gruppi, due dei quali sono stati usati per la stima dei parametri, lasciando il terzo per testare i modelli ottenuti. In particolare, nel modello connessionista, il primo e il secondo gruppo sono stati usati rispettivamente per l'addestramento (151 casi) e la validazione (49 casi), mentre nel modello di regressione questi due gruppi sono stati considerati insieme (200 casi) [2]. In questo modo le condizioni di partenza erano identiche sia per il modello neurale sia per quello di regressione. La valutazione complessiva dell'efficacia di ciascun modello è stata verificata sul terzo sottocampione, quello di test (50 casi), sempre tratto dalla precedente distribuzione di dati, mai utilizzato precedentemente.

La suddivisione del campione totale nei tre sottogruppi, è stata pilotata in modo da mantenere costante la percentuale per sesso e tipo di diploma di maturità; anche il livello di prestazione è stato utilizzato nella stratificazione inserendo in ciascun gruppo circa un terzo di soggetti per ogni livello.

Il Modello di Regressione

L'obiettivo di un modello di regressione consiste nello stimare la forza del legame tra una o più variabili indipendenti (predittori) ed una o più variabili dipendenti. Nella situazione più semplice il legame è di tipo lineare e ciascuna variabile dipendente può essere espressa con una equazione del tipo:

$$y_j = a_j + b_{j1}x_1 + b_{j2}x_2 + \ldots + b_{jI}x_I \qquad (2)$$

in cui *I* rappresenta il numero di predittori, *a* è la costante additiva o intercetta ed i vari *b* rappresentano i coefficienti di regressione.

Essendo il nostro interesse principale un confronto con i risultati ottenuti dalla rete neurale, abbiamo costruito un modello di regressione molto semplice, in cui abbiamo ipotizzato solo relazioni lineari e preso in considerazione poche variabili, tutte di tipo quantitativo.

Come già spiegato precedentemente, i soggetti considerati sono un campione di 250, equamente ripartiti per sesso e per livello di prestazione osservato. In prima battuta è stata eseguita una analisi di varianza sulle variabili che costituiscono la prestazione (media dei voti ed esami sostenuti), in modo da valutare se ci fossero degli effetti legati alle variabili qualitative sesso e diploma di maturità.

I grafici in Figura 2 e 3 rappresentano le medie dei voti e del numero di esami per anno suddivisi per sesso e tipo di diploma di maturità.

Si può notare che le prestazioni delle femmine sono migliori di quelle dei maschi in entrambe le variabili. Tale differenza risulta statisticamente significativa sia nel confronto tra le medie dei voti ($F_{(1,242)} = 4.68$; $p < .05$) sia nel confronto sul numero di esami ($F_{(1,242)} = 1.4$; $p < .01$). Non risultano significativi gli altri effetti legati al tipo di diploma e all'interazione tra le due variabili. A fronte di una tale differenza significativa si è pensato di costruire due modelli di regressione, uno per i maschi ed uno per le femmine[3].

I parametri delle regressioni sono stati stimati utilizzando un sottocampione di 200 soggetti, selezionati casualmente, come già spiegato. I predittori considerati sono quattro[4]: il voto di maturità (VOTODIP), il punteggio totale al test (TOTTEST), il numero di risposte omesse (TOTMISS) ed il numero di errori commessi (TESTERR). Tali variabili, che non risultavano distribuite normalmente, sono state normalizzate. Le variabili dipendenti sono quelle che vanno a costituire la prestazione e cioè la media dei voti (MEDIA) ed il numero medio di esami sostenuti per anno (ES_ANNO). Nella Tabella 2 sono riportati i coefficienti di determinazione e i valori di *F* calcolati sulla regressione.

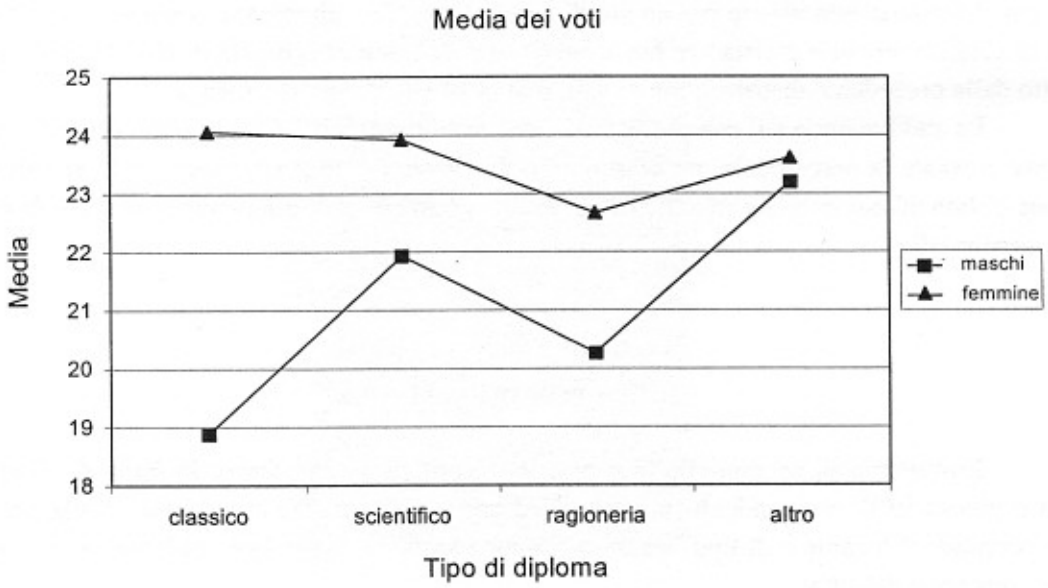

FIGURA 2
Media dei voti in relazione al titolo di studio e al sesso.

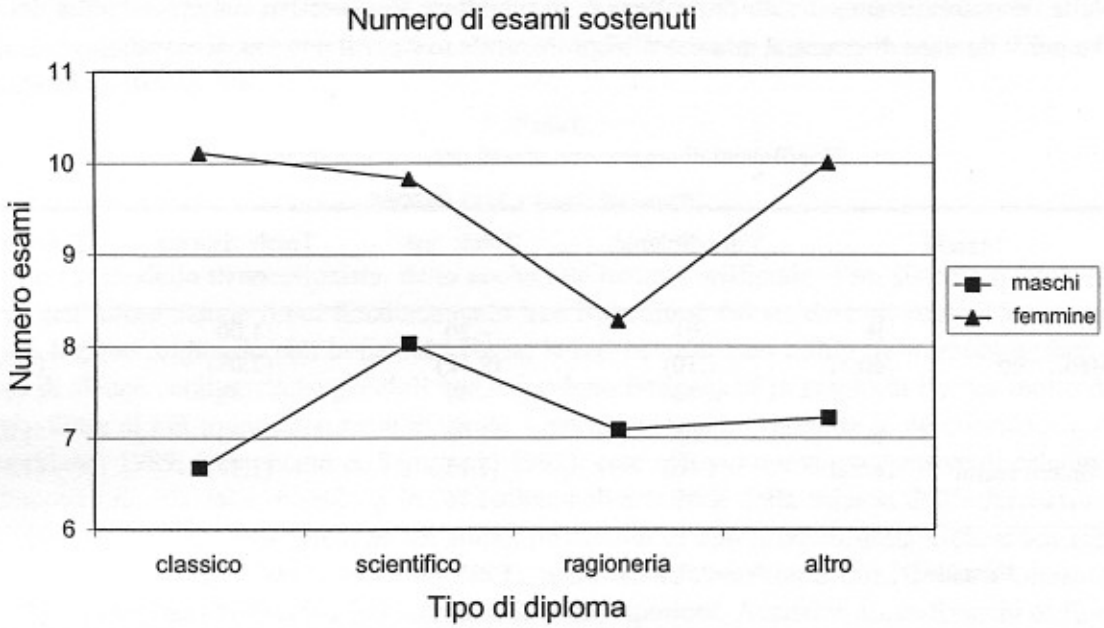

FIGURA 3
Numero medio di esami in relazione al titolo di studio e al sesso.

Tabella 2
Coefficienti di determinazione e valori di F calcolati sulle regressioni.

		R^2	Q	N	F	*sig.*
Maschi	Media voti	.03	4	102	.75	*n.s.*
	N. esami	.11	4	102	3.00	$p < .05$
Femmine	Media voti	.11	4	98	2.87	$p < .05$
	N. esami	.19	4	98	5.45	$p < .01$

I valori di F sono calcolati con la seguente formula (Jöreskog & Sörbom, 1996):

$$F = \frac{R^2 / q}{(1 - R^2)/(N - q - 1)} \tag{3}$$

in cui R^2 è il coefficiente di determinazione, in altre parole la frazione di devianza spiegata dalla combinazione lineare dei predittori, q il numero di predittori e N il numero di osservazioni; questa statistica si distribuisce con q e N-q-1 gradi di libertà.

Dalla lettura della Tabella 2 emerge che solamente la regressione sulla media dei voti nei maschi non risulta statisticamente significativa. Nella Tabella 3 si possono leggere i coefficienti di regressione stimati per i maschi e per le femmine con i relativi errori standard e valori t calcolati come rapporto B/err.st. Nel gruppo dei maschi nessun predittore risulta essere significativo.

Nelle femmine invece, il voto di diploma è un predittore significativo sia per la media dei voti che per il numero di esami, il numero di risposte errate lo è per il numero di esami.

Tabella 3
Coefficienti di regressione stimati per maschi e femmine.

Maschi		Voto diploma	Totale test	Totale risposte errate	Totale risposte omesse
	B	–.01	–.20	1.00	1.43
Media voti	err.st	(.10)	(2.12)	(2.09)	(2.90)
	t	–.14	–.09	.48	.49
	B	.05	–1.99	–.83	–1.44
Numero esami	err.st	(.06)	(1.13)	(1.12)	(1.55)
	t	.89	–1.76	–.74	–.93
Femmine		**Voto diploma**	**Totale test**	**Totale risposte errate**	**Totale risposte omesse**
	B	.19*	–1.50	–2.48	–2.62
Media voti	err.st	(.07)	(1.39)	(1.57)	(2.00)
	t	2.66	–1.08	–1.58	–1.31
	B	.22*	–2.08	–2.58*	–3.07
Numero esami	err.st	(.06)	(1.12)	(1.27)	(1.62)
	t	3.84	–1.85	–2.03	–1.90

*significativo con $p < .05$

Tenendo conto anche delle costanti additive, non riportate in Tabella 3, le equazioni di regressione per la stima della prestazione attesa risultano le seguenti; nel gruppo dei maschi:

$$\begin{cases} media = 22.12 - .01 * votodip - .20 * tottest + 1.00 * testerr + 1.43 * totmiss \\ n\_es = 5.38 + .05 * votodip - 1.99 * tottest - .83 * testerr - 1.44 * totmiss \end{cases} \quad (4)$$

e nel gruppo delle femmine:

$$\begin{cases} media = 14.07 + .19 * votodip - 1.50 * tottest - 2.48 * testerr - 2.62 * totmiss \\ n\_es = -2.12 + 0.22 * votodip - 2.08 * tottest - 2.58 * testerr - 3.07 * totmiss \end{cases} \quad (5)$$

Utilizzando queste equazioni abbiamo stimato la prestazione attesa sul gruppo di test, composto dai 50 soggetti, non considerati prima. In pratica, per ciascun soggetto vengono calcolati i valori attesi della media voti e del numero di esami. Combinando i valori attesi di media voti e numero di esami con la formula (1) abbiamo ottenuto il valore atteso di prestazione.

Ad esempio, supponiamo che un soggetto abbia una media voti attesa pari a 25.28 ed un numero di esami attesi pari a 11.13, applicando la (1) avremo il valore atteso di prestazione:

prestazione attesa = (25.28 * 11.13) / (30 * 14) = .67.

Sulla base del valore atteso di prestazione i soggetti sono stati assegnati ad uno dei tre livelli individuati precedentemente. Per cui, sempre dall'esempio, una prestazione di .67 corrisponde ad un livello alto.

Modello Connessionista

Un modello connessionista, detto anche rete neurale artificiale, è un sistema di elaborazione dell'informazione il cui funzionamento trae ispirazione dal sistema nervoso. Nonostante questo legame originario con la neurobiologia, le reti neurali sono utilizzate in molti ambiti, a causa di alcune caratteristiche generali che le rendono interessanti in campi di ricerca molto diversi. Oltre al più immediatamente evidente apporto per quanto riguarda le neuroscienze (cfr. Churchland, 1989; Churchland & Sejnowski 1992), esse offrono nuove prospettive di calcolo e una nuova impostazione di molti principi tradizionali alla base della scienza dell'informazione (cfr. Hecht-Nielsen, 1990), trovano un'ampia possibilità di applicazioni nella cibernetica (cfr. Fuminori & Fukuda, 1997; Pomerleau, 1993), nelle analisi finanziarie (cfr. Collins, Ghosh & Scofield, 1988), nella medicina (cfr. Anderson, 1986; Apolloni, Avanzini, Cesa-Bianchi & Ronchini, 1990), nella psicologia (Parisi, 1989) e in generale nello studio e nella simulazione di sistemi dinamici complessi.

Da un punto di vista statistico le reti neurali multistrato costituiscono una classe di stimatori parametrici di una funzione che realizza una assegnata dipendenza fra i dati (Vapnick, 1982), e la letteratura sulle relazioni tra i modelli connessionisti e la statistica comincia ad essere ormai abbastanza vasta (cfr. Bellacicco & Lauro, 1997; Bishop, 1995; Cheng & Titterington, 1994).

Una rete neurale è costituita da un insieme di unità interconnesse tra loro tramite connessioni. Il suo funzionamento consiste nel fatto che nella rete, attraverso le connessioni, si propaga attivazione o inibizione. In questo modo il sistema produce una certa risposta quando il propagarsi dell'attivazione della rete si stabilizza e il sistema trova un punto di equilibrio, oppure quando l'attivazione arriva sulle unità di uscita. Alcune di queste unità ricevono informazioni dall'ambiente esterno e altre emettono risposte, altre ancora hanno collegamenti solo con altre unità. Le unità che ricevono informazioni dall'ambiente sono dette *unità di input*, quelle intermedie *unità nascoste*, e quelle che emettono il loro segnale all'esterno *unità di output*. L'attivazione delle unità di input viene propagata, attraverso le connessioni, alle altre unità. Queste connessioni agiscono come dei filtri che trasformano il messaggio ricevuto aumentandone o diminuendone l'intensità a seconda delle loro caratteristiche, che sono definite *pesi*. Sia le attivazioni delle unità che i pesi sulle connessioni sono valori numerici. Nella maggior parte dei modelli i pesi possono assumere valori positivi o negativi continui e sono modificabili durante la fase di apprendimento: la rete neurale impara a fornire le risposte appropriate modificando i pesi sulle connessioni, in base a delle regole di apprendimento. In pratica ogni connessione manda all'unità con cui è collegata un segnale che non è altro che il prodotto del valore di attivazione dell'unità da cui parte la connessione per il peso sulla connessione stessa. Ciascuna unità, che può ricevere più connessioni, ha un valore di attivazione che è funzione della sommatoria di questi prodotti. La funzione matematica che trasforma il segnale ricevuto nell'attivazione dell'unità può assumere forme diverse, ma frequentemente appartiene alla famiglia delle funzioni continue non lineari, ad esempio la logistica. La risposta della rete è il valore di attivazione del-

le unità di output. Ogni particolare evento è un pattern di attivazione delle unità di input e la conoscenza della rete consiste nei pesi sulle connessioni.

Alla famiglia delle reti *etero-associative*, nelle quali le unità di ingresso che ricevono l'input dall'ambiente esterno sono distinte dalle unità di uscita che forniscono la risposta, appartengono le reti multistrato, che possiedono unità nascoste. Queste vengono dette *feedforward* quando ciascun nodo riceve connessioni soltanto dagli strati precedenti e il flusso di informazioni procede solo in una direzione.

Uno dei compiti fondamentali risolti dalle reti neurali è quello della classificazione. Una rete neurale impara a classificare dei pattern e, una volta conclusa questa fase (apprendimento), può rispondere in modo corretto a pattern nuovi che non facevano parte dei pattern con cui è stata addestrata (generalizzazione).

L'apprendimento è possibile grazie alla modifica ottimale dei pesi sulle connessioni attraverso delle regole, dette *algoritmi di apprendimento*. Una regola di apprendimento con supervisore esterno, utilizzata in numerose ricerche e applicazioni, tra cui anche la presente ricerca, è quella della "propagazione all'indietro" (*back-propagation*) dell'errore (Rumelhart, Hinton & Williams, 1986). Le unità di input vengono attivate dall'esterno, e questo input viene propagato fino alle unità di output per le quali si ottiene quindi un'attivazione. Questa è confrontata con l'output desiderato, chiamato anche *input di addestramento* (*teaching input*) e viene calcolato l'errore in base alla differenza tra le due attivazioni. L'errore viene poi propagato all'indietro con lo scopo di modificare i pesi sulle connessioni in modo tale da renderlo minimo. Data una rete con n unità di input, uno strato di unità nascoste e m unità di output, si indica con w_{ij} il peso sulla connessione tra la unità j-esima e la i-esima. Il modello opera nel seguente modo: in base al vettore di dati fornito dalle unità di input ogni unità nascosta determina la propria attivazione secondo la seguente:

$$y_i = \Phi\left(\sum_j^N w_{ij} x_j\right) \tag{6}$$

dove x_j è il valore di attivazione della j-esima unità che invia il segnale, e Φ una opportuna funzione di attivazione, tipicamente non lineare. Il risultato così ottenuto diventa l'input per le unità dello strato di output; ognuna di queste, ripetendo le stesse operazioni, produce il risultato finale. Per poter realizzare l'apprendimento utilizzando la *backpropagation*, è necessario disporre di un insieme di addestramento costituito un numero p di coppie di vettori di input x^γ e dei corrispondenti output desiderati y^γ. L'apprendimento consiste nel determinare quei valori dei pesi per i quali, per ogni $\gamma = 1, \ldots, p$ l'output fornito dalla rete O^γ in corrispondenza dei dati di input x^γ sia il più vicino possibile all'output desiderato y^γ. Formalmente, esso consiste nel rendere minima la funzione errore E così definita:

$$E = \frac{1}{2}\sum_{\gamma=1}^{p}\sum_{k=1}^{m}\left(O_k^\gamma - y_k^\gamma\right)^2 \tag{7}$$

La funzione E dipende solo dai pesi w_{ij} e per minimizzare tale funzione si utilizza la tecnica del gradiente discendente. La formula della modifica dei pesi contiene due parametri, scelti dal ricercatore: il parametro η è il *tasso di apprendimento* (*learning rate*), cioè una costante di proporzionalità con l'aumentare della quale aumentano anche i cambiamenti dei pesi, e il parametro α, ovvero la costante *momento*, solitamente scelta nell'intervallo [0-1], che determina quale frazione della modifica dei pesi ottenuta all'istante precedente (t-1) va aggiunta nel calcolo della modifica del peso sulla connessione. Il *momento* controlla la quantità di inerzia fornita dal-

la modifica dei pesi, e permette di utilizzare un tasso di apprendimento più grande che consente di rendere più veloce l'apprendimento.

Oltre all'algoritmo di apprendimento è importante anche definire un criterio di presentazione degli esempi, un criterio di aggiornamento dei pesi e un criterio per finire il processo di apprendimento. Alla fine del processo di apprendimento, infatti, i pesi non vengono più modificati e la rete così addestrata può essere testata per vedere se è in grado di rispondere correttamente a esempi mai visti in precedenza. La capacità di generalizzazione della rete dipende dagli esempi dati in fase di apprendimento. Se gli esempi sono rappresentativi di tutto lo spazio del problema, allora la rete potrà essere un generalizzatore efficiente; nel caso di problemi complessi, la soluzione migliore sarebbe quella di avere più esempi possibile, così da evitare che un evento nuovo non sia paragonabile a nessun evento presentato durante l'apprendimento. Come è prevedibile, però, non sempre questo è possibile nella realtà.

Per ottenere una buona generalizzazione è necessario scegliere l'architettura migliore cioè il numero ottimale di unità nascoste, e questo in genere si ottiene attraverso un criterio di tipo empirico, procedendo cioè per tentativi. Le connessioni di una rete neurale sono i parametri che devono essere stimati durante l'apprendimento dei pattern di input e una soluzione stabile non può essere individuata se il numero di parametri da stimare è superiore al numero di esemplari che compongono il campione di osservazione. Se il numero di unità nascoste è troppo piccolo, la rete non è in grado di separare gli input nelle classi di risposta richieste; al contrario, un numero troppo grande di pesi e unità nascoste rischia di condurre la rete ad una corrispondenza esatta tra esempi in input e risposte da apprendere, diminuendo la probabilità che vengano trovati i parametri della funzione che descrive l'intero dominio del problema. In questo caso la risposta sarà molto buona per i pattern della fase di apprendimento, ma la generalizzazione sarà inadeguata.

Un'altra scelta importante per ottenere una buona generalizzazione riguarda il numero ottimale dei cicli di apprendimento: è conveniente fermare l'apprendimento sull'insieme di addestramento prima che si raggiunga una situazione di *overfitting*, ovvero prima che la rete individui una soluzione al problema troppo specializzata sull'insieme di dati utilizzato e quindi non buona al di fuori degli esempi presentati durante l'addestramento. Una strategia empirica frequentemente messa in atto per decidere quando interrompere l'addestramento è quella di utilizzare un insieme di validazione, tratto dalla stessa popolazione dell'insieme usato per l'addestramento. Viene considerata ottimale quella rete che, pur addestrata in base a quest'ultimo insieme, minimizza l'errore sull'insieme di validazione (cfr. Floreano, 1996). Si verifica infine la generalizzazione su un terzo sottoinsieme del campione di dati a disposizione, quello di test.

Le Simulazioni

La Scelta dell'Architettura e della Funzione di Attivazione

Nelle simulazioni presentate in questo studio sono state utilizzate reti neurali *feedforward* a tre strati. Il numero di unità di ingresso è stato scelto in base alle variabili indipendenti a disposizione; alle variabili di tipo dicotomico corrispondono unità la cui attivazione può essere 1 o 0, mentre le variabili di tipo continuo hanno una attivazione che varia in modo continuo da 0 a 1.

Sono stati provati nove tipi di architetture, che differivano per il numero di unità di ingresso e per il numero di unità nascoste. Per tutte le architetture lo strato di uscita era composto di tre unità, una per ogni livello di prestazione; i livelli corrispondono a quelli descritti nell'introduzione e sono individuati attraverso la variabile *prestazione*. Si tratta cioè delle seguenti tre classi di prestazione: "bassa", "media" e "alta".

Il numero di unità di ingresso variava a seconda di quante variabili, tra quelle a disposizione, sono state effettivamente utilizzate. Tre prime simulazioni sono state compiute utilizzando tutte le variabili a disposizione. Le unità di ingresso erano dunque 18: 6 unità per le seguenti variabili dicotomiche, in cui il valore 1 indicava la presenza della caratteristica e il valore 0 la sua assenza: "genere maschile" (1 unità), residenza nel comune di Padova (1 unità), tipo di diploma di maturità (4 unità: classico, scientifico, ragioneria, altro diploma), e 12 unità per le seguenti variabili continue, tutte ricondotte mediante trasformazioni lineari entro l'intervallo 0-1: voto di maturità (1 unità), prestazione al test di ammissione alla facoltà di Economia e Commercio (3 unità: risposte corrette, risposte errate, risposte omesse), e infine risposte corrette nelle otto aree distinte individuate nel test di selezione, descritte all'inizio (8 unità). Le tre architetture differivano per numero di unità nascoste.

In un secondo momento, allo scopo di semplificare il modello, è stato ridotto il numero di unità di ingresso, considerando, tra le otto aree del test di selezione, solo le due che erano composte da un numero di item rilevante, ovvero "cultura generale" (20 item) e "logica" (30 item), ed eliminando le altre sei. Le unità di ingresso sono state quindi ridotte da 18 a 12, e anche in questo caso sono state compiute simulazioni con tre diverse architetture variando il numero di unità nascoste.

Infine, le ultime tre architetture sono state costruite con lo scopo di utilizzare un sistema di variabili più vicino a quello usato per la regressione, utilizzando soltanto le variabili genere, voto di diploma e prestazione al test di ammissione (risposte corrette, omesse e errate) per un totale di 4 unità di ingresso.

Per quanto riguarda il numero di unità nascoste, sono stati fatti tre tentativi per ognuna delle tre architetture sopra descritte, quella con 18, oppure con 12 o infine con 4 unità di ingresso, con lo scopo di scegliere la simulazione che garantiva la migliore previsione. Le connessioni di una rete neurale sono, come si è detto, i parametri che devono essere stimati durante l'apprendimento dei pattern del campione di addestramento. Per questo motivo, per poter arrivare ad una soluzione, è opportuno che il numero di connessioni non sia più elevato del numero di casi diversi presentati alla rete in fase di addestramento, che nella ricerca qui descritta corrisponde ai 151 studenti. Il numero totale di connessioni di una rete *feedforward* con uno strato di unità nascoste si ottiene moltiplicando il numero di unità nascoste per la somma delle unità di ingresso e di uscita; quindi, poiché le unità di uscita sono sempre tre, il numero di unità nascoste variava in base al numero di unità di ingresso. In particolare, nei primi tre casi, in cui le unità di ingresso erano 18, sono state fatte simulazioni con strati intermedi di 6, 7 e 8 unità, che corrispondono rispettivamente ad un totale di 126, 147 e 168 connessioni. Nelle tre simulazioni con 12 unità di ingresso le unità nascoste erano 9, 10 e 11, corrispondenti ad un minimo di 135 connessioni nel primo caso e un massimo di 165 nel terzo. Le ultime tre simulazioni sono state fatte con reti nelle quali lo strato nascosto comprendeva 13, 17 e 21 unità, cioè rispettivamente con 91, 119 e 147 connessioni totali.

L'attivazione delle unità di ingresso era determinata dal valore delle corrispondenti variabili utilizzate, descritte sopra, mentre come funzione di attivazione delle unità degli altri due strati è stata utilizzata la logistica.

La Tabella 4 riassume tutte le caratteristiche delle nove diverse architetture.

Tabella 4
Numero di unità e di connessioni per le nove architetture utilizzate.

Architettura	Numero di unità di ingresso	Numero di unità nascoste	Numero di unità di uscita	Numero totale di connessioni
1	18	6	3	126
2	18	7	3	147
3	18	8	3	168
4	12	9	3	135
5	12	10	3	150
6	12	11	3	165
7	4	13	3	91
8	4	17	3	119
9	4	21	3	147

La Fase di Addestramento e Validazione

L'addestramento delle reti è stato condotto utilizzando l'algoritmo della *back-propagation,* con i due parametri fissi (momento $\alpha = .1$ e tasso di apprendimento $\eta = .7$), con inizializzazione casuale dei pesi in un intervallo [–.3, +.3], e modificando i pesi sulle connessioni dopo la presentazione di ogni singolo pattern di ingresso. La presentazione dei 151 pattern di ingresso, in ogni epoca[5], è stata fatta in modo casuale senza ripetizione, per un numero di epoche. Il campione di addestramento è stato utilizzato per individuare la soluzione mentre il campione di validazione (51 casi) è servito per decidere dopo quante epoche interrompere l'addestramento e "congelare" i pesi. La misura di errore cui si è fatto riferimento è l'Errore Quadratico Medio (EQM) definito come:

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(t_i - y_i)^2 \qquad (8)$$

dove n indica il numero dei pattern, nel nostro caso i soggetti, t_i indica l'output desiderato e y_i quello realmente ottenuto dalla rete per l'i-esimo pattern.

Dati i valori in ingresso, venivano richiesti per le tre unità di uscita i seguenti valori di attivazione: 1 0 0 per ogni output che corrispondeva al primo livello di prestazione (basso), 0 1 0 per il secondo livello (medio) e 0 0 1 per il terzo (alto). La rete era cioè addestrata ad individuare il livello di prestazione dei casi presentati nel campione di addestramento. Alla fine del processo di addestramento, per eliminare ogni risposta ambigua, è stato scelto di trasformare i valori di uscita nel seguente modo: l'unità di uscita che presentava il valore di attivazione più alto ve-

niva considerata attiva (attivazione 1) mentre le altre due erano considerate non attive (valore di attivazione 0). In questo modo ogni risposta della rete individuava un livello di prestazione.

Per ciascuna delle nove diverse architetture si è partiti da un numero minimo di 10 epoche (1.510 cicli) di addestramento per arrivare ad un massimo di 10.000 epoche (1.510.000 cicli).

È interessante notare che il numero di variabili prese in considerazione in ingresso ha comportato notevoli differenze nella possibilità di raggiungere delle soluzioni stabili. Infatti, mentre nelle prime sei architetture il numero di casi correttamente classificati continuava a crescere rapidamente, via via che il numero di epoche di addestramento aumentava, fino ad arrivare a riconoscere più del 90% dei casi, nelle ultime tre architetture, quelle in cui le unità di ingresso erano soltanto 4, le reti sono riuscite a classificare correttamente al massimo il 64 % dei casi nel campione di addestramento. Tuttavia proprio queste ultime tre architetture si sono dimostrate quelle con maggiori capacità di generalizzazione, arrivando spesso a classificare correttamente più del 50% dei casi del campione di validazione. È importante ribadire che i pesi sulle connessioni sono stimati in base al campione di addestramento, e non su quello di validazione. Quest'ultimo entra in gioco soltanto per indicare quando è il caso di interrompere l'addestramento, per evitare che la rete si "specializzi" sui casi presentati in questa fase, visto che lo scopo è piuttosto quello di individuare un modello generale di previsione della prestazione a partire da alcune caratteristiche date (le variabili in ingresso).

La simulazione che ha fornito i risultati complessivamente migliori sui due campioni di addestramento e di validazione è stata quella con 4 unità di ingresso, 21 unità nascoste, addestrata interrompendo l'addestramento dopo 10.000 epoche, con il 55.2% di casi correttamente classificati per il campione di addestramento e il 49% di corrette classificazioni nel campione di validazione.

A questo punto, per verificare la bontà del modello, i pesi ottenuti sulle connessioni in questa simulazione, ovvero i parametri stimati dalla rete sono stati tenuti invariati e la rete è stata fatta girare, senza apprendimento, con i 50 casi del campione test in ingresso, registrandone il relativo output. Questi 50 casi sono stati utilizzati per la prima volta in questo momento e quindi risultavano completamente nuovi per la rete.

CONFRONTO TRA I MODELLI

I livelli di prestazione, individuati rispettivamente dal modello di regressione e dalla rete neurale sul campione di test, sono riportati nelle Tabelle 5 e 6 incrociando le frequenze dei livelli di prestazione osservati con quelli attesi dai due modelli. I valori riportati sulle diagonali principali indicano le previsioni eseguite correttamente.

La regressione (Tabella 5) tende a stimare essenzialmente prestazioni medie (il 58% dei valori attesi cade in questa fascia di prestazione). In totale ci sono 22 soggetti (pari al 44% del totale) per i quali la prestazione osservata e quella attesa coincidono; 24 soggetti (pari al 48%) hanno una prestazione attesa che differisce di una classe dalla prestazione osservata. I restanti 4 soggetti presentano uno scarto di due classi tra prestazione osservata ed attesa.

Il modello connessionista (Tabella 6) fornisce in totale 26 previsioni corrette (52%), con 17 soggetti per i quali la prestazione attesa differisce di una classe (34%) e i rimanenti 7 soggetti

Tabella 5
Confronto tra il livello di prestazione osservato e quello teorico previsto dal modello di regressione.

			Livello di prestazione atteso			
			Basso	Medio	Alto	Totale
Livello di prestazione osservato	Basso	Conteggio	7	8	2	17
		% per riga	41.2%	47.1%	11.8%	100%
	Medio	Conteggio	7	12		19
		% per riga	36.8%	63.2 %		100 %
	Alto	Conteggio	2	9	3	14
		% per riga	14.3%	64.3%	21.4%	100%
Totale		Conteggio	16	29	5	50
		% per riga	32.0%	58.0%	10.0%	100%

di due (14%). Si nota che vengono classificati correttamente 12 dei 14 casi di alto livello di prestazione e circa la metà dei casi di bassa prestazione, mentre i risultati peggiori in termini di corretta classificazione sono quelli che riguardano i casi di prestazione intermedia, in cui le classificazioni corrette sono soltanto il 26.3%. In generale il modello connessionista tende a stimare essenzialmente prestazioni alte o basse, escludendo le prestazioni medie (vengono classificati come appartenenti a questa classe soltanto il 16% dei casi). Per un confronto tra i modelli, abbiamo scelto due indici di associazione per tabelle di frequenza: il coefficiente di contingenza (Kendall, 1970; McNemar, 1969) e il Gamma (Goodman & Kruskal, 1954, 1959).

Tabella 6
Confronto tra il livello di prestazione osservato e quello teorico previsto dal modello connessionista.

			Livello di prestazione atteso			
			Basso	Medio	Alto	Totale
Livello di prestazione osservato	Basso	Conteggio	9	3	5	17
		% per riga	52.9%	17.6%	29.4%	100%
	Medio	Conteggio	5	5	9	19
		% per riga	26.3%	26.3%	47.4%	100%
	Alto	Conteggio	2		12	14
		% per riga	14.3%		85.7%	100%
Totale		Conteggio	16	8	26	50
		% per riga	32.0%	16.0%	52.0%	100%

Nella Tabella 7 riportiamo i valori di tali indici, ottenuti rispettivamente nel modello di regressione e in quello connessionista.

Tabella 7
Misure dei coefficienti di contingenza e Gamma per le tabelle prodotte dai due modelli.

	Regressione		Rete neurale	
	Valori	*p.*	Valori	*p.*
Contingenza	.334	*n.s.*	.443	< .05
Gamma	.333	*n.s.*	.586	< .01

Entrambi questi coefficienti esprimono indipendenza tra le variabili quando valgono zero; quanto più si avvicinano ad uno tanto maggiore sarà la relazione tra le variabili. Il coefficiente di contingenza esprime in particolare quanto le frequenze tendano a concentrarsi in una cella per ogni riga e colonna; in pratica il valore massimo si ha quando c'è un'unica cella non vuota in ogni riga ed in ogni colonna della tabella. Dato che la variabile è ordinale, ci attendiamo che, quanto migliore sarà il modello, tanto più le frequenze si concentreranno sulle celle della diagonale, dalla prima in alto a sinistra all'ultima in basso a destra. Valori di Gamma vicini ad uno indicano una tale distribuzione delle frequenze.

I valori di contingenza e di Gamma non risultano statisticamente significativi per la regressione mentre lo sono per la rete neurale, questo sembrerebbe deporre a favore di una migliore capacità predittiva di questo modello, almeno a livello generale. Tale "superiorità" si osserva anche in Tabella 8, ove sono riportati il numero e la tipologia di errori commessi dai due modelli di previsione; la rete ha una percentuale di previsioni corrette (52%) superiore a quella della regressione (44%), anche se quest'ultima commette meno errori estremi, stimando cioè come alti dei livelli di prestazione bassi e viceversa.

Tabella 8
Previsioni corrette e sbagliate nei due modelli (tra parentesi il percentuale sul totale)

Differenza prestazione attesa — osservata	Regressione	Rete
Corrette	22 (44%)	26 (52%)
± 1 classe	24 (48%)	17 (34%)
± 2 classi	4 (8%)	7 (14%)

In sintesi, il modello connessionista offre risultati migliori in termini di previsione quando deve stimare la prestazione di soggetti che si collocano nelle due classi estreme, mentre la regressione va meglio nella classe centrale (vedi Tabelle 5 e 6).

In Tabella 9 possiamo leggere la concordanza dei due modelli nei tre livelli di prestazione osservati. Risulta che per un livello osservato basso solo cinque soggetti su 17 sono correttamente classificati da entrambi i modelli; se il livello osservato è medio c'è un solo soggetto su 19 classificato correttamente; per il livello alto ci sono tre soggetti su 14.

Tabella 9
Confronto tra rete e regressione nei tre livelli di prestazione osservati.

		Rete			
Livello osservato basso		Basso	Medio	Alto	Totale
Regressione	Basso	5	2	0	7
	Medio	4	1	3	8
	Alto	0	0	2	2
Totale		9	3	5	17
Livello osservato medio		Basso	Medio	Alto	Totale
Regressione	Basso	3	4	0	7
	Medio	2	1	9	12
	Alto	0	0	0	0
Totale		5	5	9	19
Livello osservato alto		Basso	Medio	Alto	Totale
Regressione	Basso	0	0	2	2
	Medio	2	0	7	9
	Alto	0	0	3	3
Totale		2	0	12	14

Per ciascun livello di prestazione osservato, abbiamo calcolato i valori di Gamma (sulle tre parti della Tabella 9); sia per il livello basso sia per quello medio il comportamento dei modelli tende ad essere significativamente concorde (Gamma (liv. basso) = .724 $p < .01$; Gamma (liv. medio) = .784 $p < .001$), non così quando si considerano i soggetti con livello di prestazione osservato alto (Gamma (liv: alto) = .2 *n.s.*).

In conclusione, sembra che la modalità scelta per confrontare i due modelli privilegi i risultati della rete e ciò accade probabilmente per due ordini di ragioni. In primo luogo, trasformando i valori stimati dalla regressione, che sono continui, in classi discrete, si perde in "accuratezza" e di conseguenza anche in precisione della stima; da questo punto di vista la rete, che ottiene in output valori già disposti in classi discrete gioca praticamente in casa. Un'altra ragione può essere legata agli assunti che regolano i due modelli; nella regressione un solo predittore risulta statisticamente significativo (vedi Tabella 3), ma ciò implica solo che il suo legame con le variabili dipendenti è di tipo lineare, nulla possiamo dire su eventuali relazioni di altro ordine. Nella rete non abbiamo significatività statistica ma, in compenso, non siamo vincolati ad assunti di linearità e questo potrebbe spiegare la migliore predittività generale di tale modello. In accordo con questa riflessione un possibile sviluppo della ricerca potrebbe introdurre le seguenti modifiche: da una parte avvicinare la rete alla regressione aumentando il numero di categorie in output, dall'altra avvicinare la regressione alla rete considerando un modello di tipo non lineare.

Un altro miglioramento si otterrebbe, a nostro avviso, eliminando dal campione i soggetti che non hanno sostenuto esami. Tali soggetti, infatti, creano una specie di salto nella variabile prestazione, salto che ha degli indubbi effetti sulla stima dei parametri di regressione. Per trattare tali soggetti è necessario assumere che essi si distribuiscano linearmente nell'intervallo di voti che va da zero a 18, assunzione piuttosto difficile da sostenere.

NOTE

1. Non è stato possibile considerare il numero di esami attesi dal piano di studi in quanto i dati del campione sono stati rilevati nel corso di svolgimento dell'anno accademico. Per questo motivo è stato considerato il numero massimo di esami realmente sostenuti dagli studenti del campione.
2. La necessità di avere tre gruppi è propria del modello connessionista, per motivi che verranno spiegati più avanti; nel modello di regressione, invece, è sufficiente disporre di un gruppo per la stima dei coefficienti ed un secondo per applicarli.
3. In alternativa, avremmo potuto inserire nel modello di regressione anche sesso e diploma utilizzando variabili di tipo dummy, ciò non è stato fatto per mantenere il modello il più semplice possibile.
4. Tra parentesi riportiamo i nomi delle variabili utilizzati nelle elaborazioni e che verranno poi usati per la scrittura delle equazioni di regressione.
5. Dal momento che si definisce "epoca" la presentazione di tutti i pattern di addestramento, in queste simulazioni un'epoca comprende 151 cicli, con modifica dei pesi dopo la presentazione di ogni singolo pattern di ingresso.

BIBLIOGRAFIA

Anderson, J. A. (1986). Cognitive capabilities of a parallel system. In E. Bienenstock, F. Fogelman-Souli & G. Weisbuch (Eds:), *Disordered systems and biological organization* (Vol. F20) (pp. 209-226). Berlin: Springer-Verlag, NATO-ASI Series.

Apolloni, B., Avanzini, G., Cesa-Bianchi, N., & Ronchini, G. (1990). Diagnosis of epilepsy via backpropagation. In *Proceedings of the 1990 International Joint Conference on Neural Networks* (Vol. II, pp. 571-574). Washington, DC.

Bellacicco, A., & Lauro, N. C. (1997). *Reti neurali e statistica.* Milano: Franco Angeli.

Bishop, C. M. (1995). *Neural Networks for pattern recognition.* Oxford: University Press.

Cheng, B., & Titterington, D. M. (1994). Neural networks: a review from a statistical perspective. *Statistical Science, 9,* 2-54.

Churchland, P. M., (1989). *The neurocomputational perspective.* Cambridge, MA: MIT Press. (Trad. it. parziale *La natura della mente e la struttura della scienza*, Bologna, Il Mulino, 1992).

Churchland, P. M., & Sejnowski, T. J. (1992). *The computational brain.* Cambridge, MA: MIT Press. (Trad. it. *Il cervello computazionale.* Bologna: Il Mulino, 1995).

Collins, E., Ghosh, S., & Scofield, C. L. (1988). An application of multiple neural network learning system to the emulation of mortgage underwriting judgements. In M. Caudill & C. Butle (Eds.), *Proceedings of the IEEE International Conference on Neural Networks* (Vol. II, pp. 459-466). New York: IEEE Press.

Dutta, S., & Shekkar, S. (1988). Bond rating: a non-conservative application of neural networks. In M. Caudill & C. Butler (Eds.), *Proceedings of the IEEE International Conference on Neural Networks* (Vol. II, pp. 443-450). New York: IEEE Press.

Ercolani, A. P., Areni, A., & Mannetti, L. (1990). *La ricerca in psicologia.* Roma: La Nuova Italia Scientifica.

Floreano, D. (1996). *Manuale sulle reti neurali.* Il Mulino: Bologna.

Fuminori, S., & Fukuda, T. (1997). A first result of the Brachiator III. A new Brachiation Robot Modeled on a Siamang. In C. Langton & K. Shimohara (Eds.), *Artificial life V* (pp. 354-361). Cambridge, MA: MIT Press.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732-764.

Goodman, L. A., & Kruskal, W. H. (1959). Measures of association for cross classifications. II: further discussion and references. *Journal of the American Statistical Association, 54,* 123-163.

Hecht-Nielsen, R. (1990). *Neurocomputing.* Reading, MA: Addison-Wesley.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: user's reference guide.* Chicago, IL: Scientific Software International, Inc.

Kendall, M. G. (1970). *Rank correlation methods* (fourth edition). London: Griffin.

McNemar, Q. (1969). *Psychological statistics* (fourth edition). New York: J. Wiley.

Parisi, D. (1989). *Intervista sulle reti neurali.* Bologna: Il Mulino.

Pomerleau, D. A. (1993). *Neural network perception for mobile robot guidance.* Boston: Kluwer Academic Publishing.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations through error propagation. In D. E. Rumelhart, J. L. McClelland & The PDP Research Group (Eds.), *Parallel distributed processing: experiments in the microstructure of cognition* (Vol. I, pp. 318-362). Cambridge: MIT Press.

Siegel, S. N., & Castellan J. J. (1992). *Statistica non parametrica* (trad.it.). McGraw-Hill: Milano.

Vapnick, V. (1982). *Estimation of dependencies based on empirical data.* New York: Springer-Verlag.