

Un criterio di inizializzazione per reti neurali nell'analisi del trend

Francesco Giordano

Dipartimento di Scienze Economiche e Statistiche, Università di Salerno
E-mail: giordano@unisa.it

Summary: This paper deals with the problem of the weights initialisation for feed-forward neural networks in order to estimate the Trend component. The proposed procedure gives a starting point for the minimisation of the objective function. The aim is to grant for this starting point to be the closest to an optimal point.

Keywords: Feed-forward Neural Networks, Weights initialisation, Trend Analysis

1. Introduzione

Si consideri un modello del tipo “Segnale + errore “ formulato come:

$$Y_t = f(t) + \mathbf{e}_t \quad (1)$$

dove Y_t è la serie storica osservata, $f(\cdot) \in C^1[K]$ dove K un compatto su \mathfrak{R} e $\mathbf{e}_t \sim \text{WN}(0, \mathbf{S}^2)$.

Il Segnale, in questo contesto, è un Trend, non necessariamente, lineare, di cui si vuole ottenere una stima.

In letteratura esistono diversi approcci per la stima del Trend in una serie storica. In termini parametrici il problema può essere risolto fissando una forma funzionale ad $f(\cdot)$ e stimando i parametri da cui essa dipende.

In ambito non parametrico non è fissata alcuna forma funzionale per $f(\cdot)$ ed è possibile utilizzare vari stimatori come i Kernel, le Reti Neurali ecc. La scelta tra le diverse tipologie di stimatori può essere fatta (a) rispetto alla natura della stima: globale o locale; (b) relativamente alle diverse condizioni di completezza sugli spazi funzionali e così via.

L'uso delle reti neurali come stimatori non parametrici si pone nella classe degli stimatori globali relativamente a spazi di tipo $C^1[K]$. Tralasciando il problema della specificazione della topologia della rete, in questo lavoro focalizziamo l'attenzione su uno dei problemi più cruciali nell'uso delle reti neurali, cioè l'inizializzazione dei pesi.

Lo scopo del lavoro è quello di proporre un metodo per inizializzare tali pesi confrontando i risultati con gli altri metodi introdotti in letteratura.

Il paragrafo 2 introduce definisce la topologia di rete neurale usata. La metodologia proposta è spiegata nel paragrafo 3, rimandando all'appendice per la dimostrazione della consistenza della procedura. Nel paragrafo 4 si riporta una simulazione volta a comparare i risultati della metodologia proposta con quella più usata (inizializzazione casuale) e con quella suggerita da Yam e Chow, 2000. Infine seguono alcune considerazioni e sviluppi futuri.

2. Le reti neurali artificiali

La rete neurale considerata per l'analisi del Trend è una rete di tipo *feed-forward* con una topologia costituita da un solo neurone di input (t), un solo strato intermedio di neuroni ed, infine, un solo neurone di output (Y_t). Formalmente si ha:

$$Y_t^* = \sum_{j=1}^m c_j \mathcal{Y}(a_j t) \quad (2)$$

dove m indica il numero dei neuroni dello strato intermedio, $\mathbf{y}(\cdot)$ è una funzione di attivazione definita come $\mathbf{y}(z) = \frac{1}{1 + \exp(-z)}$, c_j e a_j sono $2*m$ pesi incogniti da stimare.

La stima dei parametri a_j e c_j , $j=1,2,\dots,m$, nella (2) implica la soluzione di un problema di minimizzazione non lineare, cioè

$$\min_{a_j, c_j} \frac{1}{N} \sum_{k=1}^N \left[Y_k - \sum_{j=1}^m c_j \mathbf{y}(a_j t_k) \right]^2$$

dove t_k , per $k=1,2,\dots,N$, è una sequenza deterministica che può essere assunta, senza perdere in generalità, come $t_k \equiv k$.

La variabile indipendente t_k può essere definita come una qualsiasi sequenza uniforme (in progressione aritmetica) su un compatto di \mathfrak{R} .

3. Metodologia per inizializzare i pesi della rete neurale.

Il problema è di scegliere il valore iniziale dei $2*m$ pesi. L'algoritmo proposto parte dalla constatazione che $\mathbf{y}(\cdot)$ può essere considerata una variabile casuale Uniforme su $[0,1]$ e che $\mathbf{y}_{ij} \equiv \mathbf{y}(a_j t)$, $j=1,2,\dots,m$, deve essere una base dello spazio $C^1[K]$ su cui è definita $f(\cdot)$.

La procedura proposta è strutturata nei seguenti passi:

1. Si fissa un $t_0 \in \{t_k\}$ con $k=1,2,\dots,N$,
2. Si generano m numeri casuali da una variabile casuale Uniforme su $[0,1]$, u_j , $j=1,2,\dots,m$.
3. Si calcolano $a_{j0} = \mathbf{y}^{-1}(u_j)/t_0$, $\forall j=1,2,\dots,m$.
4. Si può così costruire una matrice

$$X_0 = \begin{pmatrix} \mathbf{y}(a_{10}1) & \dots & \mathbf{y}(a_{m0}1) \\ \dots & \dots & \dots \\ \mathbf{y}(a_{10}N) & \dots & \mathbf{y}(a_{m0}N) \end{pmatrix} \text{ di dimensioni } N \times m.$$

5. Per $N > m$, il rango di X_0 deve essere m in quanto gli \mathbf{y}_{ij} devono essere una base. Per cui
 - a. rango di $X_0 < m$;
 - b. rango di $X_0 = m$.
 - 5a Dato che il sistema non è una base allora la procedura inizializza tutti i pesi da una variabile casuale Uniforme in $[-1,1]$
 - 5b In questo caso i pesi a_{j0} sono inizializzati come al passo 3 mentre i pesi c_{j0} sono inizializzati risolvendo un problema di regressione lineare in cui Y_t è la variabile dipendente e X_0 è la matrice dei regressori.
- 6 Sia v_0 il relativo valore della funzione obiettivo.
- 7 I passi da 1 a 6 sono ripetuti per un certo numero h di punti di $\{t_k\}$ generando ad ogni iterazione g , a_{jg} , c_{jg} e v_g , con $g=0,1,\dots,h-1$. La scelta, quindi, dell'inizializzazione sarà per a_{jg^*} e c_{jg^*} tale che $g^* = \arg \min_g v_g$.

Tale algoritmo di inizializzazione, essendo condizionato dal fatto che le funzioni \mathbf{y}_{ij} con $j=1,2,\dots,m$ e $t=1,2,\dots,N$, devono essere una base, si riconduce alla classica inizializzazione casuale se m è sovradimensionato rispetto al rapporto Segnale/Errore cioè la stima neurale cattura anche parte dell'Errore. Per quanto riguarda il valore di h , è plausibile legarlo al numero dei parametri da stimare e fissare, per esempio, $h=2*m$. In effetti tale algoritmo si basa sull'estrazione casuale da una variabile casuale Uniforme su $[0,1]$. Questo intervallo è sicuramente limitato ma ha la potenza del continuo. Essendo l'estrazione da tale Uniforme necessariamente finita, allora la procedura, sopra descritta, è consistente se per h (il numero dei punti del supporto di $f(\cdot)$) che tende all'infinito si ha direttamente un minimo della funzione obiettivo. Quando h tende all'infinito si ha una successione di estrazioni dalla variabile casuale Uniforme e quindi con una potenza inferiore al quella del continuo. Si rimanda all'appendice per la dimostrazione della consistenza di tale procedura di inizializzazione.

4. Un esperimento di simulazione e risultati empirici

La procedura descritta al paragrafo precedente è stata comparata con quella di Yam e Chow, (2000), la quale, come verificato, ha performance sicuramente superiori a tutte le tecniche proposte in letteratura. L'esperimento di simulazione è stato effettuato generando le osservazioni delle serie storiche modellando il Trend per mezzo della funzione di Wahba, cioè:

$$Y_t = 4.26 (\exp(-t) - 4\exp(-2t) + 3\exp(-3t)) + \mathbf{e}_t$$

con $t \in [0, 2.5]$ e dove $\mathbf{e}_t \sim N(0, 0.04)$

La lunghezza della serie storica è di $N=200$ osservazioni e per ogni serie generata si è proceduto ad analizzare tre tipi di inizializzazioni.

- 1 Inizializzazione dei pesi secondo una variabile casuale Uniforme su $[-1, 1]$ (IU).
- 2 Inizializzazione secondo Yam e Chow (IYC).
- 3 Inizializzazione secondo la procedura del paragrafo 3 (IG).

Nel caso dell'analisi del Trend e tenendo conto della rete neurale definita nella (2), l'inizializzazione (IYC) è strutturata come segue.

Si fissa $\bar{s} = 4.59$ avendo posto la funzione di attivazione $\mathbf{y}(\cdot)$ come la logistica. Per ogni osservazione della serie storica si calcolano

$$b_t = \frac{\bar{s}}{t} \sqrt{3} \text{ e } b = \min_t b_t.$$

I pesi a_j , $j=1, 2, \dots, m$, vengono generati da una variabile casuale Uniforme su $[-b, b]$. L'inizializzazione dei pesi c_j , $j=1, 2, \dots, m$, avviene invece analogamente alla procedura definita in (IG).

La simulazione è condotta su 200 repliche. Su ogni serie si calcolano i pesi per l'inizializzazione per i casi 1-3 e quindi, si lancia l'algoritmo di minimizzazione per ottenere i pesi ottimali. Nel punto di ottimo si valuta la funzione obiettivo per i tre casi di inizializzazione. I risultati sono allora tre serie, di lunghezza 200, relative al valore della funzione obiettivo nel punto di ottimo in relazione ai tre tipi di inizializzazione.

Data una replica $j, j=1,2,\dots,200$, sia

- $RS1_j$ il valore della funzione obiettivo nel punto di minimo partendo dalla procedura di inizializzazione (IU).
- $RS2_j$ il valore della funzione obiettivo nel punto di minimo partendo dalla procedura di inizializzazione (IYC).
- $RS3_j$ il valore della funzione obiettivo nel punto di minimo partendo dalla procedura di inizializzazione (IG).

Inoltre, essendo ogni replica indipendente dalle altre, è possibile definire con

- ◆ N_1 il numero delle volte che $RS2_j < RS1_j$;
- ◆ N_2 il numero delle volte che $RS3_j < RS1_j$.

Siano $\hat{P}_1 = \frac{N_1}{200}$ e $\hat{P}_2 = \frac{N_2}{200}$,

le probabilità empiriche di successo, rispettivamente della procedura (IYC) e di (IG). Bisogna tener presente che sia la procedura (IG) che la (IYC) tendono, per costruzione, al caso (IU) se la base generata ha dimensione minore di m .

L'ipotesi che i neuroni dello strato intermedio devono essere una base risulta cruciale per le procedure (IYC) e (IG). Infatti dalla simulazione risulta che la procedura (IYC) riesce a definire una base con un $m < 6$ mentre con la procedura (IG) risulta $m < 10$. Questa differenza si può spiegare in quanto la (IYC) tende a sottostimare i pesi a_j . Tale risultato ha una conseguenza immediata poiché, per costruzione, l'inizializzazione (IG) e (IYC) si riconducono alla (IU) se i pesi a_j non implicano una base di dimensione m .

Per valutare in modo globale i risultati della simulazione, è possibile costruire il test

$$\begin{aligned} H_0 : P_i &= 1/2 \\ H_1 : P_i &\neq 1/2 \end{aligned} \quad \text{con } i=1,2$$

Tabella 1. Test sulle proporzioni. m è il numero dei neuroni dello strato intermedio, \hat{P}_1 e \hat{P}_2 sono le proporzioni stimate con la simulazione e $Prob$ è il corrispondente p-value.

m	\hat{P}_1	$Prob$	\hat{P}_2	$Prob$
4	0.505	0.887	0.685	0
6	0.52	0.57	0.71	0
10	0.475	0.4795	0.575	0.033

Dalla Tabella 1, per il caso $m=10$ è possibile ritenere che sia vera l'ipotesi nulla sia per la procedura (IYC) ($i=1$) sia per quella (IG) ($i=2$) con un livello di significatività del 1%. In effetti in questo caso, come è già stato accennato, le due procedure non generano una base relativamente ai pesi a_j . Quindi sono, praticamente, uguali alla inizializzazione (IU), come risulta dalla tabella. Il caso $m=6$, invece, conduce ad accettare l'ipotesi nulla per la procedura (IYC) ma a rifiutarla per la (IG). Questa situazione è caratterizzata dal fatto che la procedura (IYC), non generando una base, si confonde con l'inizializzazione casuale mentre quella proposta genera quasi sempre una base e risulta, quindi, significativamente migliore. Infine, il caso per $m=4$ è importante perché sia la procedura (IYC) che la (IG) generano una base. Tuttavia la prima non sembra essere significativamente migliore rispetto a quella casuale mentre lo è la (IG).

I test presentati nella tabella 1 sono un risultato globale sulla simulazione e perciò sintetico sulla performance delle varie procedure di inizializzazione. Allo scopo di valutare meglio il risultato della simulazione nella figura 1 sono riportati i boxplot del valore della funzione obiettivo nel punto di ottimo, relativi alle 200 repliche della simulazione stessa, nel caso di $m=4$ neuroni nello strato intermedio e rispettivamente per la procedura IU (1), IYC (2) e IG (3).

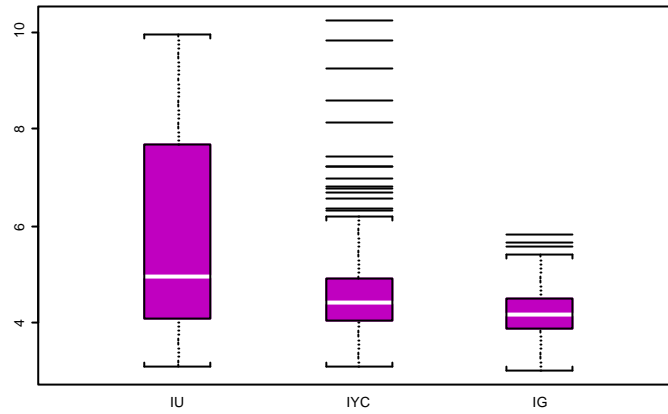


Figura 1. Boxplot del valore della funzione obiettivo nel punto di ottimo; $m=4$ neuroni

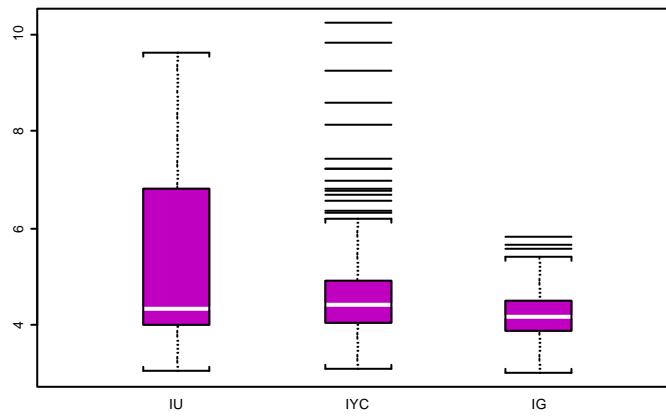


Figura 2. Boxplot del valore della funzione obiettivo nel punto di ottimo; $m=6$ neuroni

Si può notare come la procedura (IYC) e quella (IG) sono relativamente meno variabili rispetto alla inizializzazione casuale (IU). Tuttavia la procedura (IYC) produce diversi valori eccezionali. Sicuramente la procedura (IG) è quella meno variabile. Rispetto al risultato sintetico del test riportato nella tabella 1 non si può ritenere che la procedura (IYC) sia equivalente a quella casuale. Infatti dal test non si percepisce la variabilità nella simulazione del risultato delle diverse procedure di inizializzazione. Anche nel caso di $m=6$ neuroni nello strato intermedio, è confermata la migliore performance della procedura (IG), come è evidente dalla figura 2. Da essa emerge anche una notevole differenza tra le procedure (IYC) e (IU). Infatti la procedura (IYC) presenta diversi valori eccezionali che dimostrano, in alcune iterazioni, risultati peggiori rispetto alla inizializzazione casuale (IU). In effetti questo caso per circa il 95% delle iterazioni la procedura (IYC) è la stessa di quella casuale. Questo dimostra la notevole instabilità nella stima dei pesi nell'ambito delle reti neurali.

I boxplot relativi al caso $m=10$ sono del tutto equivalenti per le tre procedure sia come variabilità che per presenza di valori eccezionali come si può vedere dalla figura 3.

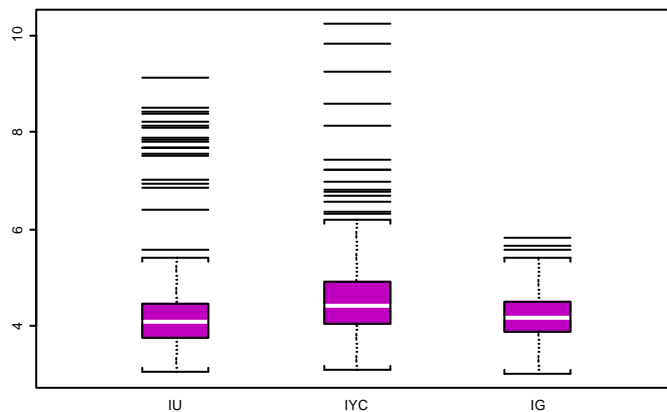


Figura 3. Boxplot del valore della funzione obiettivo nel punto di ottimo; $m=10$ neuroni

5. Conclusioni

In questo caso, relativo al problema della stima del Trend e con l'uso di una topologia di rete neurale abbastanza particolare, la procedura di inizializzazione dei pesi proposta (IG) dà un risultato migliore rispetto a quella di Yam e Chow (IYC) e risulta anche consistente (vedi Appendice). Il lavoro di Yam e Chow si pone in un'ottica prettamente deterministica mentre la simulazione effettuata considera una struttura stocastica per le serie generate. Tuttavia anche in un contesto deterministico i risultati sono del tutto simili a quelli presentati nel paragrafo precedente. La procedura proposta è sicuramente più complessa a livello computazionale rispetto a quella di Yam e Chow in quanto implica il calcolo del rango di una matrice rettangolare e la soluzione di un problema di regressione lineare. In effetti non è una complessità eccessiva tenendo conto della particolare topologia della rete per l'analisi del Trend.

Uno dei possibili sviluppi futuri è, appunto, quello di valutare la procedura proposta nell'ambito di una topologia di rete neurale più generale.

Riferimenti bibliografici

Barron, A. R. (1993) Universal Approximation Bounds for Superpositions of a sigmoidal function, *IEEE Transactions on Information Theory*, 39, 3, 930-945.

Yam, Y. F. e Chow, T. W. S. (2000) A weight initialization method for improving training speed in feedforward neural network, *Neurocomputing*, 30, 219-232.

Tong, H. (1990) *Nonlinear Time Series Analysis: a dynamic approach*, Oxford University Press.

White, A. (1989) Some Asymptotic Results for Learning in Single Hidden-Layer Feed-Forward Network Models *JASA*, 84, 1003-1013.

Appendice

La procedura di inizializzazione descritta al paragrafo 3 si può definire consistente se essa garantisce, “in un certo senso”, il minimo assoluto della funzione obiettivo quando le osservazioni diventano sufficientemente numerose.

Sia $I_h = \{1, 2, \dots, h\}$ un insieme finito costituito da h elementi. $\forall s \in I_h$ si ha la seguente funzione obiettivo

$$v_s(N) = \frac{1}{N} \sum_{k=1}^N \left[Y_k - \sum_{j=1}^m c_{js} \mathcal{Y}(a_{js} t_k) \right]^2 \quad (\text{A1})$$

dove c_{js} e a_{js} sono i pesi inizializzati con centro sul punto t_s del supporto.

N rappresenta il numero delle osservazioni e $t_s \in K$, un compatto in \mathfrak{R} .

Dato che la funzione obiettivo è valutabile solo per N punti del supporto di cui si dispongono le osservazioni Y_i , $i=1, 2, \dots, N$, allora $h \leq N$ e quindi se $h \rightarrow \infty \Rightarrow N \rightarrow \infty$.

Inoltre, è noto (Barron, 1993) che:

$$\exists \underline{c}^*, \underline{a}^* : \left\| f(t) - \sum_{j=1}^m c_j^* \mathcal{Y}(a_j^* t) \right\|_{L^2(K)}^2 = O(1/m)$$

dove $\underline{c}^* = (c_1^*, \dots, c_m^*)$ e $\underline{a}^* = (a_1^*, \dots, a_m^*)$ sono i vettori dei parametri della rete neurale tali da minimizzare la norma.

Sia

$$Y(t; \underline{c}_N^*, \underline{a}_N^*) = \sum_{j=1}^m c_{j(N)}^* \mathcal{Y}(a_{j(N)}^* t)$$

dove $\underline{c}_N^* = (c_{1(N)}^*, \dots, c_{m(N)}^*)$ e $\underline{a}_N^* = (a_{1(N)}^*, \dots, a_{m(N)}^*)$ sono i vettori dei pesi della rete neurale calcolati minimizzando la funzione obiettivo su N osservazioni della funzione incognita $f(\cdot)$.

Essendo la funzione di attivazione la logistica, $\mathbf{y}(\cdot)$ è continua secondo Lipschitz e inoltre (White, 1989) è

$$\underline{c}_N^* \xrightarrow{q.c.} \underline{c}^* \quad \text{e} \quad \underline{a}_N^* \xrightarrow{q.c.} \underline{a}^*$$

per cui è facile verificare che:

$$Y(t; \underline{c}_N^*, \underline{a}_N^*) \xrightarrow{L^2(K)} Y(t; \underline{c}^*, \underline{a}^*).$$

Dalla (1) si evince che l'errore $|Y_i - f(\cdot)|$ tende ad una costante per N che tende all'infinito. Dato che la variabile t è deterministica e uniformemente distribuita in K , e posto:

$$\|f(t) - Y(t; \underline{c}^*, \underline{a}^*)\|_{L^2(K)}^2 \equiv v^*,$$

la procedura di inizializzazione proposta è *consistente* se

$$\min_{s \in I_h} v_s(N) \xrightarrow{h \rightarrow \infty} v^* \quad (\text{A2})$$

a meno di una costante additiva.

Si può inoltre dimostrare il seguente teorema.

Teorema. Se $\mathbf{y}(a_j t)$, $j=1,2,\dots,m$, è una base, allora la procedura proposta al paragrafo 3 è consistente nel senso della (A2).

Dim.

La dimostrazione può essere divisa in tre parti. La prima parte riguarda la consistenza della procedura relativamente ai pesi c_j . La seconda parte affronta il problema della consistenza sui pesi a_j in un

caso particolare. Infine la terza parte completa la dimostrazione nel caso generale.

Parte I

Sia

$$F_N(t; \underline{c}, \underline{a}) = \frac{1}{N} \sum_{i=1}^N (Y_i - Y(t; \underline{c}, \underline{a}))^2$$

e siano \underline{a}_N^* e \underline{c}_N^* i vettori che minimizzano la funzione $F_N(t; \underline{c}_N^*, \underline{a}_N^*)$

Fissato N e \underline{a}_N^* , si ha che:

$$X_N = \begin{pmatrix} \mathbf{y}(a_1^*1) & \dots & \mathbf{y}(a_m^*1) \\ \dots & \dots & \dots \\ \mathbf{y}(a_1^*N) & \dots & \mathbf{y}(a_m^*N) \end{pmatrix}$$

$$\hat{\underline{c}}_N = \arg \min_{\underline{c}} F_N(t; \underline{c}, \underline{a}_N^*).$$

Dato che $F_N(t; \underline{c}, \underline{a}_N^*)$ è lineare in \underline{c} e ammette un'unica soluzione,

essendo X_N di rango m per ipotesi, deve essere $\hat{\underline{c}}_N = \underline{c}_N^*$.

Seguendo le stesse argomentazioni presentate per introdurre la consistenza della procedura risulta inoltre che

$$F_N(t; \underline{c}_N^*, \underline{a}_N^*) \xrightarrow{L(K)} F(t; \underline{c}^*, \underline{a}^*)$$

Parte II

Per l'inizializzazione relativa ai pesi a_j si estraggono dei numeri casuali da una Uniforme su $[0,1]$, cioè

$$\mathbf{m}_j \sim U[0,1] \quad i=1,2,\dots,h \quad \text{e} \quad j=1,2,\dots,m$$

Siano:

$$\underline{\mathbf{m}}_1 \in \mathfrak{R}^m(0,1), \dots, \underline{\mathbf{m}}_h \in \mathfrak{R}^m(0,1).$$

Allora è $P(\underline{\mathbf{m}}_i = \underline{\mathbf{m}}_0) = 0$ con $\underline{\mathbf{m}}_0$ vettore di costanti.

Sia Φ_1 un insieme che contiene l'elemento $\underline{\mathbf{m}}_1$, Φ_2 l'insieme che contiene Φ_1 e l'elemento $\underline{\mathbf{m}}_2$, cioè

$$\Phi_1 = \{\underline{\mathbf{m}}_1\}, \Phi_2 = \{\underline{\mathbf{m}}_1, \underline{\mathbf{m}}_2\}, \dots, \Phi_h = \{\underline{\mathbf{m}}_1, \underline{\mathbf{m}}_2, \dots, \underline{\mathbf{m}}_h\}.$$

Si ha così una successione crescente di insiemi con

$$\Phi_h = \Phi \text{ per } h \rightarrow \infty.$$

L'insieme Φ può essere considerato isomorfo a $Q^m(0,1)$; ma essendo la distribuzione continua, Φ è anche isomorfo a $Q^m[0,1]$ q.o.

Poiché, inoltre, $Q^m[0,1]$ è denso in $\mathfrak{R}^m[0,1]$ allora

$$\forall \underline{\mathbf{m}} \in \mathfrak{R}^m[0,1] \exists \underline{\mathbf{m}}' \in Q^m[0,1]: \|\underline{\mathbf{m}} - \underline{\mathbf{m}}'\| < \mathbf{e} / u_1$$

dove $\|\cdot\|$ indica la norma euclidea su \mathfrak{R}^m e u_1 è una costante positiva.

La funzione obiettivo è sicuramente derivabile rispetto a t ed ai pesi a_j e c_j , per cui è continua secondo Lipschitz, cioè

$$\left| F(\underline{\mathbf{m}}; t, \underline{\mathbf{c}}) - F(\underline{\mathbf{m}}'; t, \underline{\mathbf{c}}) \right| \leq u_1 \|\underline{\mathbf{m}} - \underline{\mathbf{m}}'\| < \mathbf{e} \text{ q.o.}$$

dove $\mathbf{m}_i = \mathbf{y}(a_i; t)$, $i=1,2,\dots,m$.

Parte III

La parte I e II della dimostrazione permettono di verificare la consistenza della procedura di inizializzazione nel caso che l'ottimo $\mathbf{y}(a_j^*; t)$, $j=1,2,\dots,m$, sia definito per valori razionali in $[0,1]$.

Per completare la dimostrazione è necessario estendere tale risultato ai numeri reali in $[0,1]$. Infatti risulta chiaro che la densità dell'insieme dei numeri razionali su quelli reali definiti su un compatto $[0,1]$ permette di accettare la consistenza della procedura per i pesi a_j . Basta, quindi, verificare che tale proprietà si estende anche ai pesi c_j .

Siano $\underline{\mathbf{m}}' \in Q^m [0,1]$ e $\underline{\mathbf{m}} \in \mathcal{R}^m [0,1]$. La procedura di inizializzazione produce i valori di ottimo $\underline{\mathbf{m}}'$ e $\hat{\underline{\mathbf{c}}}_{\mathbf{m}'}$. Per quanto precedentemente dimostrato (parte II) risulta:

$$\|\underline{\mathbf{m}}' - \underline{\mathbf{m}}\| < \mathbf{e} \Rightarrow F(\underline{\mathbf{m}}'; t, \hat{\underline{\mathbf{c}}}_{\mathbf{m}'}) = F(\underline{\mathbf{m}}; t, \hat{\underline{\mathbf{c}}}_{\mathbf{m}}).$$

Sia $F'_{c_j}(\underline{\mathbf{m}}'; t, \hat{\underline{\mathbf{c}}}_{\mathbf{m}'})$ la derivata parziale della funzione obiettivo rispetto ad un generico $c_j, j=1,2,\dots,m$. Dato che la funzione obiettivo è lineare rispetto a tali parametri, risultano continue e derivabili anche le derivate parziali. Inoltre, per costruzione, $F'_{c_j}(\underline{\mathbf{m}}'; t, \hat{\underline{\mathbf{c}}}_{\mathbf{m}'}) = 0$.

Per la condizione di Lipschitz, u_2 positivo, risulta:

$$\begin{aligned} \left| F'_{c_j}(\underline{\mathbf{m}}; t, \hat{\underline{\mathbf{c}}}_{\mathbf{m}}) - F'_{c_j}(\underline{\mathbf{m}}'; t, \hat{\underline{\mathbf{c}}}_{\mathbf{m}'}) \right| \leq u_2 \|\underline{\mathbf{m}} - \underline{\mathbf{m}}'\| < \mathbf{e} * u_2 / u_1 \Rightarrow F'_{c_j}(\underline{\mathbf{m}}; t, \hat{\underline{\mathbf{c}}}_{\mathbf{m}}) = 0 \\ \Rightarrow \hat{\underline{\mathbf{c}}}_{\mathbf{m}'} = \hat{\underline{\mathbf{c}}}_{\mathbf{m}} = \underline{\mathbf{c}}^* . \end{aligned}$$

Questo conclude la dimostrazione.