Reti Neurali Artificiali per lo studio del mercato

Università degli studi di Brescia - Dipartimento di metodi quantitativi
Marco Sandri (sandri.marco@gmail.com)

Regressione: 4 istruttivi esempi

Obiettivi:

(1) Vogliamo qui analizzare con Matlab 4 diversi insiemi di dati (dati_reg1.txt, dati_reg2.txt, dati_reg3.txt e dati_reg4.txt) mediante la stima di alcuni modelli di regressione lineare e polinomiale, valutando poi la bontà del modello sulla base delle caratteristiche dei residui.
Qui non vengono ancora utilizzate le reti neurali. Questi esempi sono però molto utili per comprendere il problema della regressione nei suoi diversi aspetti, problema che verrà poi affrontato anche con le reti neurali.
I 4 casi che qui vengono presentati rispecchiano situazioni che si possono presentare di frequente nella realtà.

(A) Nel primo esempio ci troviamo a modellizzare una variabile y sulla base dell'informazione contenuta in una variabile indipendente x (covariata). Il campione contiene 500 osservazioni. La prima cosa che facciamo è quella di andare a stimare un modello di regressione lineare:

$$y = a_0 + a_1 * x$$

Dall'analisi dei residui di questo modello ci rendiamo però conto che esso non riesce a cogliere completamente la componente deterministica presente nei dati. I residui mostrano infatti una componente sistematica non lineare che varia al variare di x, componente che non è stata catturata dal modello lineare. Nella fase successiva stimiamo il modello quadratico:

$$y = a_0 + a_1 * x + a_2 * x^2$$

Questo modello mostra una migliore bontà di adattamento ai dati e mostra di essere stato in grado di catturare in modo completo la componente sistematica presente nei dati (cioè il meccanismo generatore dei dati).

(B) Anche nel secondo esempio ci troviamo a modellizzare una variabile y sulla base dell'informazione contenuta in una variabile indipendente x (covariata). Il campione contiene 1500 osservazioni. Per prima cosa andiamo di nuovo a stimare un modello di regressione lineare:

$$y = a_0 + a_1 * x$$

Dall'analisi dei residui ci rendiamo però conto che esso non riesce a spiegare gran parte della componente deterministica presente nei dati. I residui mostrano infatti una componente sistematica non lineare che varia al variare di x. Nella fase successiva stimiamo un modello polinomiale di ordine n:

$$y = a_0 + a_1 * x + a_2 * x^2 + \ldots + a_n * x^n$$

Il parametro n non è noto e possiamo solo cercare di stimarlo mediante una serie di tentativi. Il modello che otteniamo mostra una migliore bontà di adattamento ai dati e mostra di essere stato in grado di catturare buona parte della componente sistematica presente nei dati (cioè il meccanismo generatore dei dati). Esso presenta però anche alcuni problemi: ha un elevato numero di parametri (n+1) ed è difficile da interpretare (che significato si può infatti attribuire alle potenze di una data variabile?).

(C) Nel terzo esempio ci troviamo a modellizzare una variabile y sulla base dell'informazione contenuta in due variabili indipendenti x_1 ed x_2, dove la seconda variabile è binaria (0/1). Il campione contiene 1000 osservazioni. Per prima cosa andiamo a stimare un modello di regressione lineare con la sola variabile x_1:

$$y = a_0 + a_1 * x_1$$

Dall'analisi dei residui ci rendiamo però conto che essi mostrano una distribuzione bimodale, lontana dall'ipotesi di guassianità. Nella fase successiva stimiamo un nuovo modello lineare considerando ora anche la variabile x_2:

$$y = a_0 + a_1 * x_1 + a_2 * x_2$$

Il modello che otteniamo mostra un'ottima bontà di adattamento ai dati e risulta aver catturato interamente il meccanismo generatore dei dati.
Si osservi che il modello qui sopra scritto equivale a:

$$y = a_0 + a_1 * x_1 \quad \text{se } x_2=0$$
$$y = (a_0 + a_2) + a_1 * x_1 \quad \text{se } x_2=1$$

Equivale cioè a due modelli lineari aventi diversa intercetta nei due sottogruppi $x_2=0$ e $x_2=1$.

(D) Nel quarto ed ultimo esempio ci troviamo ancora a modellizzare una variabile y sulla base dell'informazione contenuta in due variabili indipendenti x_1 ed x_2, dove la seconda variabile è binaria (0/1). Il campione contiene 1000 osservazioni. Per prima cosa andiamo a stimare un modello di regressione lineare con le due variabili:

$$y = a_0 + a_1 * x_1 + a_2 * x_2$$

Dall'analisi dei residui ci rendiamo però conto che essi mostrano una evidente eteroschedasticità, cioè la variabilità dei residui cambia al variare di x_1, è ampia vicino ai valori 0 e 3 e si riduce avvicinandosi a 1.5. Nella fase successiva stimiamo un nuovo modello lineare considerando ora anche l'interazione fra la variabile x_1 e la variabile x_2, cioè il loro prodotto:

$$y = a_0 + a_1 * x_1 + a_2 * x_2 + a_3 * x_1 * x_2$$

Il modello che otteniamo mostra un'ottima bontà di adattamento ai dati e risulta aver catturato interamente il meccanismo generatore dei dati.
Si osservi che il modello qui sopra scritto equivale a:

$$y = a_0 + a_1 * x_1 \quad \text{se } x_2=0$$
$$y = (a_0 + a_2) + (a_1+a_3) * x_1 \quad \text{se } x_2=1$$

Equivale cioè a due modelli lineari aventi diversa intercetta e diverso coefficiente angolare nei due sottogruppi $x_2=0$ e $x_2=1$.