Data mining

Il data mining rappresenta l'attività di elaborazione in forma grafica o numerica di grandi raccolte o flussi continui di dati con lo scopo di estrarre 'informazione utile' a chi detiene i dati stessi.

E' cioè un processo di selezione, esplorazione e modellazione di grande masse di dati al fine di scoprire regolarità o relazioni non note a priori e allo scopo di ottenere un risultato chiaro e utile al proprietario del data base.

Data mining

L'oggetto di interesse del data mining in generale non è specificato a priori, ma si cerca di individuarlo 'scavando' tra i dati. Il data mining è anche noto come: knowledge-discovery in databases (KDD). L'approccio è diametralmente opposto a quello di altre indagini in cui l'obiettivo è invece specificato a priori (ad es. studi clinici).

"*The most fundamental difference between classical statistical applications and data mining is the size of the data*". (Hand et al. 2001)

Data mining

E’ una disciplina recente che si colloca al punto di intersezione di diverse aree scientifiche: gestione dei data base, intelligenza artificiale (machine learning, pattern recognition, etc.) e statistica.

Il data mining ha incontrato una certa diffidenza da parte degli statistici per varie ragioni, fra cui: i dati non sono quasi mai raccolti in base ad un dato piano di campionamento, l’obiettivo dello studio spesso non è definito a priori.

Data mining

"*If you torture the data long enough, Nature will always confess*".
(R.H.Coase, premio Nobel 1991 per l'economia)

"*Data-Snooping occurs when a given set of data is used more than once for purposes of inference or model selection. This leads to the possibility that any results obtained in a statistical study may simply be due to chance rather than to any merit inherent in the method yielding the results*."

Data mining

"Data dredging is the term used to refer to the unscrupulous search for 'statistically significant' relationships in large quantities of data. Conventional statistical procedure is to formulate a research hypothesis, then collect relevant data, then carry out a statistical significance test to see whether the results could be due to the effects of chance.

A key point is that one is not allowed to formulate the hypothesis as a result of seeing the data. If you want to work this way, you have to collect a data set, then partition it into two subsets, A and B.

Subset A is held back and subset B is examined for interesting hypotheses. Once a hypothesis has been formulated it can be tested on subset A, since it was not used to construct the hypothesis.".

Data mining

Un modello è una rappresentazione semplificata del fenomeno di interesse, funzionale ad un obiettivo specifico.

- Semplicità: descrivere gli aspetti rilevanti ed eliminare gli aspetti inessenziali.
- Dipendenza dall'obiettivo: esistono modelli diversi dello stesso fenomeno, a seconda dell'obiettivo.
- Non esiste un modello 'vero', ma modelli più o meno utili per quell'obiettivo.

Data mining

Pensiamo ai nostri dati come se fossero generati da una sorta di scatola nera in cui un vettore **x** di variabili di input (variabili indipendenti) entra nella scatola e dall'altra parte esce la variabile **y** di risposta (variabile dipendente).
Dentro questa scatola, la natura crea una associazione (a noi ignota) fra variabile di risposta e variabili predittive.

Data mining

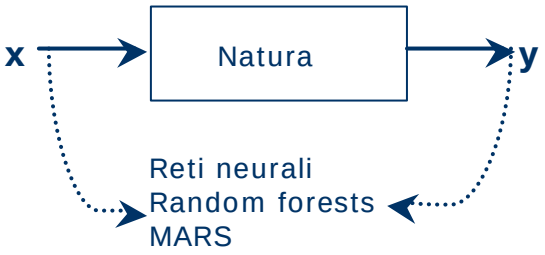
In generale, nell'analisi dei dati, ci si pone uno di questi obiettivi:
- previsione: conoscere quale sara' la risposta y in corrispondenza di determinati input x;
- informazione: estrarre qualche informazione sul legame esistente fra la variabile dipendente **y** e le variabili **x**.

Data mining

Due diversi approcci al problema:
- 'data modeling': questo tipo di analisi parte con l'assumere un modello stocastico generatore dei dati del tipo **y** = f(**x**, rumore, parametri)
Il valore dei parametri viene stimato mediante i dati e il modello (cioe' f) viene usato per l'informazione e/o la previsione. La scatola nera viene cio riempita con un qualche tipo di modello:

Data mining

Due diversi approcci al problema:
- ‘algorithmic modeling’: questo tipo di analisi considera cio’ che sta dentro la scatola come una relazione complessa e ignota; gli sforzi vengono concentrati sull’individuazione di una f(x), cioe di un algoritmo che partendo da x predice la risposta y.



Riferimenti bibliografici

- Azzalini A. e Scarpa B. (2004), “Analisi dei dati e data mining”, Springer-Verlag Italia, Milano
- Giudici P. (2001), “Data mining. Metodi statistici per le applicazioni aziendali”, McGraw-Hill Co.
- Breiman L. (2001), “Statistical modelling: the two cultures”, Statistical Science, 16 (3), 199-231

Introduzione

La tecnica delle **Reti Neurali Artificiali** (RNA) è stata sviluppata inizialmente nell'ambito degli studi sull'intelligenza artificiale.

In seguito si è constatato che le RNA sono in grado di affrontare molti problemi trattati generalmente attraverso tecniche statistiche e sono quindi divenute oggetto di studio anche da parte di molti statistici.

Introduzione

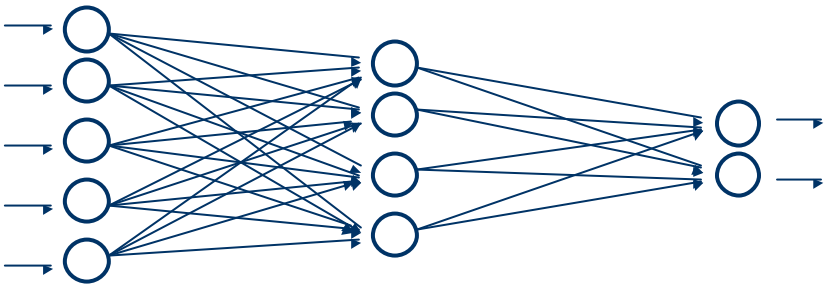
Obiettivo è quello di delineare sia da un punto di vista teorico che applicativo la tecnica delle RNA, osservando le varie problematiche ***in un'ottica statistica di Analisi dei Dati***.

Le RNA trovano utilizzo in una grande varietà di campi di studio, dalla fisica all'ingegneria, dalla medicina all'economia, dalla biologia alle scienze sociali.

In questo Corso sarà data particolare attenzione all'***utilizzo delle RNA per lo studio del mercato***.

Definizione

Le Reti Neurali Artificiali sono un insieme di unità computazionali semplici (neuroni artificiali, nodi, unità) disposte su vari livelli ed interconnesse tra loro attraverso un definito sistema di legami (architettura).

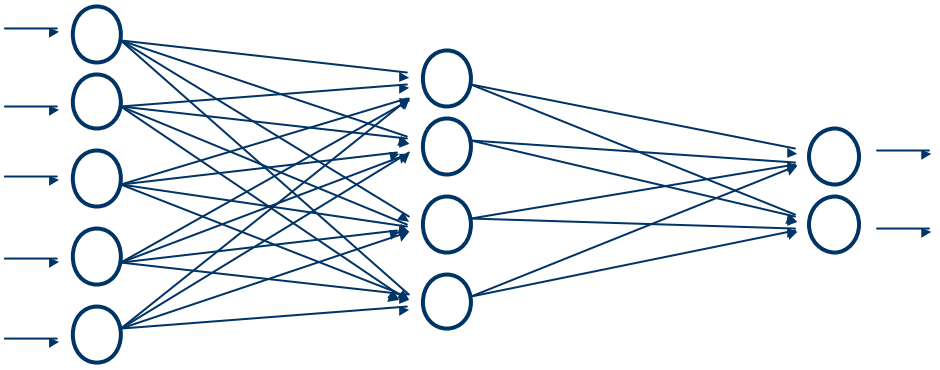


LIVELLO DI INPUT | LIVELLO NASCOSTO | LIVELLO DI OUTPUT

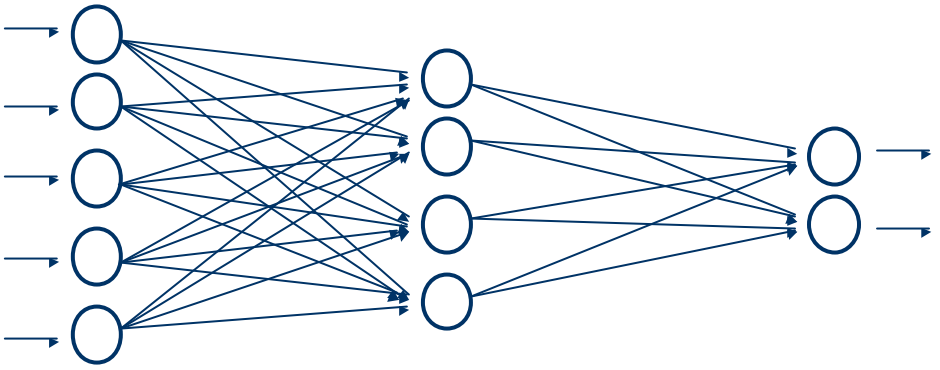
Definizione

La filosofia che sta alla base delle RNA è mutuata dall'osservazione del funzionamento del cervello umano: ad ogni singolo neurone artificiale sono assegnati dei compiti molto semplici, ma la sua interazione con gli altri neuroni crea un flusso di segnali in grado di elaborare l'informazione attraverso funzioni altamente flessibili e complesse.

Definizione



Definizione



Definizione

Le RNA si utilizzano per gli scopi più diversi. Il contesto generale di lavoro è costituito dallo studio di un fenomeno in cui alcune variabili Y_1, Y_2, …, Y_c (**qualitative o quantitative**) possono considerarsi spiegabili attraverso un certo numero di variabili esplicative X_1, X_2, …, X_s (anch'esse qualitative o quantitative).

L'*input* della rete è costituito dalle osservazioni x_{ki} delle variabili X_k, mentre le previsioni y_{ki} per le variabili Y_k sono fornite come *output*.

Architettura della rete

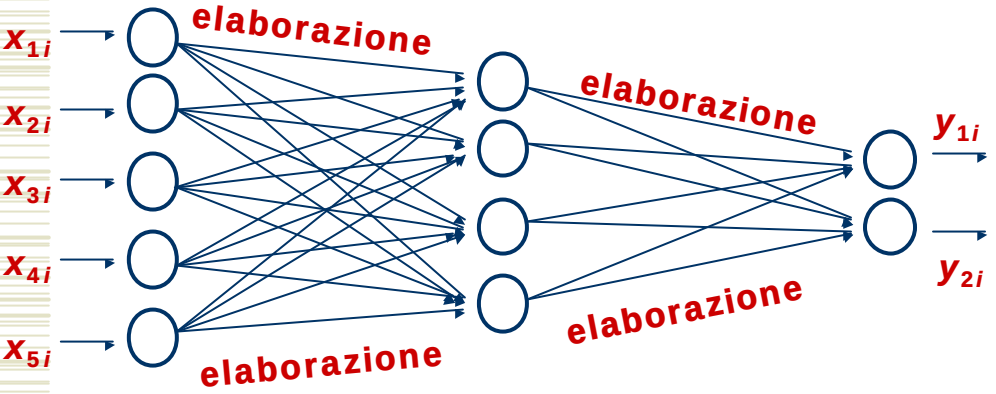
Il numero di nodi nel livello di *input* è determinato dal numero ***s*** di variabili esogene ed equivalentemente il numero di nodi nel livello di *output* dal numero ***c*** di variabili dipendenti.

Il numero ***l*** di livelli nascosti ed il numero ***m*** di nodi in ogni livello è determinato dal ricercatore (in genere RNA con un solo livello nascosto sono in grado di fornire ottimi risultati).

L'insieme dei valori {***s***, ***l***, ***m***, ***c***} e le connessioni che si stabiliscono tra i neuroni costituiscono la cosiddetta **ARCHITETTURA DELLA RETE**

Architettura della rete

$s = 5$ $l = 1$ $m = 4$ $c = 2$



Architettura della rete

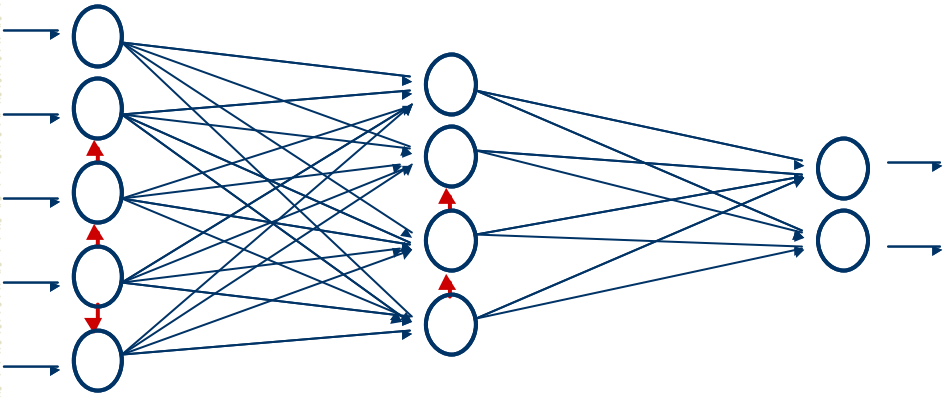
Ci occuperemo soltanto delle architetture di RNA più utilizzate:

TOTALMENTE CONNESSE (ogni unità è connessa con tutte quelle del livello successivo, ma mai con quelle dello stesso livello)

NON RICORRENTI o *FEED-FORWARD* (le connessioni operano in un solo senso, dal livello di input ai successivi)

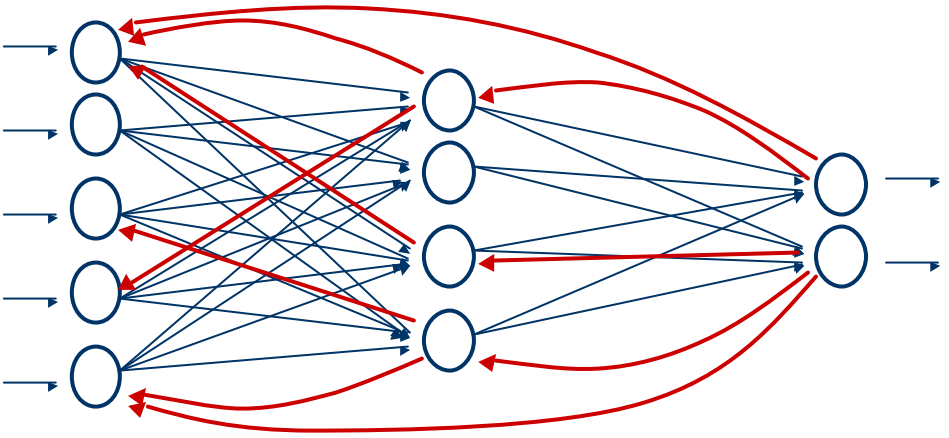
Architettura della rete

RETI NON TOTALMENTE CONNESSE



Architettura della rete

RETI RICORRENTI



Esercizi

Esercizio 1: rappresentare l'architettura di una RNA totalmente connessa, non ricorrente con $s = 7$ nodi nel livello di *input*, $l = 1$ livello nascosto composto da $m = 4$ nodi e $c = 3$ nodi nel livello di *output*.

Esercizio 2: rappresentare l'architettura di una RNA totalmente connessa, non ricorrente con $s = 6$ nodi nel livello di *input*, $l = 2$ livelli nascosti composti rispettivamente da $m_1 = 4$ e $m_2 = 2$ nodi e $c = 1$ nodo nel livello di *output*.

Esempio

Esempio: una società proprietaria di una catena di ipermercati vuole stimare il fatturato di un punto vendita di prossima apertura, sulla base di tre variabili esogene (due quantitative ed una qualitativa).

Formalmente si ha $s = 3$, $c = 1$:

X_1 = numero di abitanti del bacino di utenza

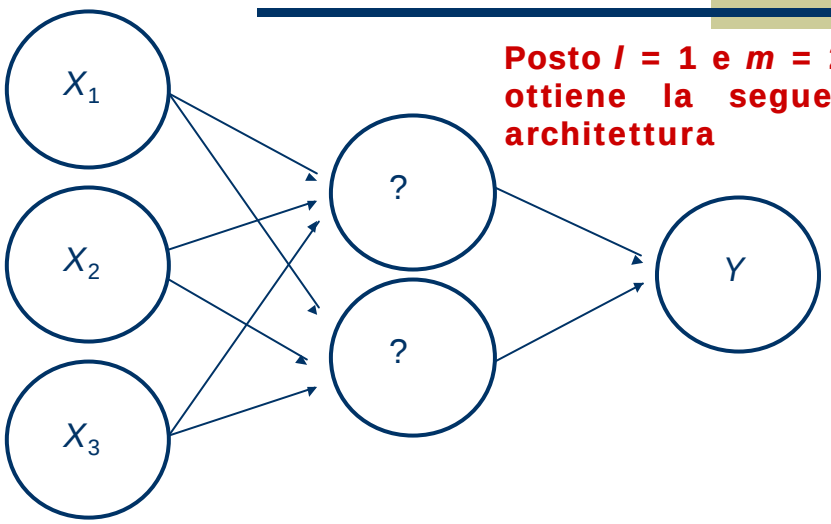
X_2 = reddito medio degli abitanti del bacino di utenza

X_3 = {0, 1} (presenza o meno di ipermercati concorrenti entro un certo raggio)

Y = fatturato

Esempio

Posto *l* = 1 e *m* = 2 si ottiene la seguente architettura



Esercizi

Esercizio 3: un'industria produttrice di yogurt "da passeggio" che distribuisce attraverso una rete di punti vendita in *franchising* vuole stimare la quantità ottimale da fornire mensilmente ad un nuovo punto vendita, sulla base di 10 variabili esogene riguardanti la collocazione del punto vendita e la sua dimensione e dei dati relativi ai punti vendita già esistenti. Rappresentare l'architettura della RNA con $l = 1$ e $m = 3$ utilizzata per risolvere il problema.

Esercizi

Esercizio 4: una società di noleggio di automobili vuole aprire una nuova filiale e stimare per questa la domanda settimanale di

- automobili utilitarie
- automobili di dimensioni medio-grandi
- automobili di lusso

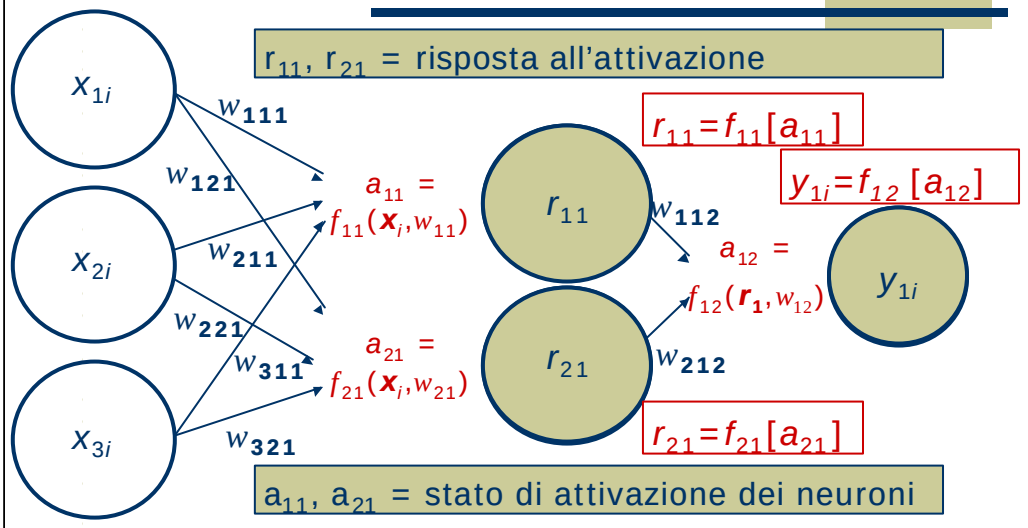
sulla base di 15 variabili esogene riguardanti la struttura socio-economica del bacino di utenza della nuova filiale e disponendo dei dati relativi alle filiali già esistenti. Rappresentare l'architettura della RNA con $l = 1$ e $m = 7$ utilizzata per risolvere il problema.

Formalizzazione matematica

Il sistema di interconnessioni tra i neuroni e il livello di risposta al loro stato di attivazione vengono specificati attraverso:

- Un insieme di ***pesi*** w
- Un insieme di funzioni f (***propagation rule***)
- Un insieme di ***funzioni di attivazione*** (o ***di trasferimento***) f

Formalizzazione matematica



Formalizzazione matematica

$\boldsymbol{x}_i = \{x_{1i}, \ldots, x_{si}\}$ vettore degli *input* per l'osservazione *i*-esima

w_{jkh} peso della connessione tra il *j*-esimo neurone del livello *h*-1 e il *k*-esimo neurone del livello *h*.

$w_{kh} = \{w_{1kh}, \ldots, w_{mkh}\}$ insieme dei pesi che agiscono sul *k*-esimo neurone del livello *h*-esimo

f_{kh} *propagation rule* del *k*-esimo neurone del livello *h*-esimo

Formalizzazione matematica

a_{kh} stato di attivazione del k-esimo neurone del livello h-esimo

f_{kh} funzione di attivazione del k-esimo neurone del livello h-esimo

r_{kh} risposta del k-esimo neurone del livello h-esimo

$\mathbf{r}_h = \{r_{1h}, \ldots, r_{mh}\}$ insieme dei livelli di risposta del livello h-esimo

y_{ki} k-esimo *output* della rete per l'osservazione i-esima

Formalizzazione matematica

Le funzioni f **trasformano l'*output* dei nodi di un dato livello nell'*input* dei nodi del livello successivo (stato di attivazione)**, attraverso i pesi w specificati nelle interconnessioni tra le unità.

Le funzioni f sono dette **funzioni di attivazione perché specificano la reazione del neurone allo stato di attivazione, fornendo la risposta del neurone alla sollecitazione**.

Lo stato di attivazione

Lo **stato di attivazione** a_{kh} del k-esimo neurone nel livello h-esimo è calcolato come funzione degli *input* provenienti dal livello precedente e dei pesi delle interconnessioni attraverso la *propagation rule* f_{kh}:

$a_{k1} = f_{k1}(\boldsymbol{x}, w_{k1})$ per il primo livello nascosto

$a_{kh} = f_{kh}(\boldsymbol{r}_{h-1}, w_{kh})$ per i livelli nascosti successivi

$a_{k(l+1)} = f_{k(l+1)}(\boldsymbol{r}_l, w_{k(l+1)})$ per il livello di *output*

La propagation rule f

Le funzioni f sono determinanti per il calcolo dello stato di attivazione.

E' compito del ricercatore stabilire quale forma attribuire alle funzioni f.

Tale forma può essere definita in un modo qualunque, tuttavia esistono alcune forme tipiche.

La propagation rule f

• Combinazioni lineari degli *input* • ***sigma units***

In questo caso la la *propagation rule* f_{kh} è data da:

per il primo livello nascosto

$$f_{k1}(\boldsymbol{x}, \boldsymbol{?}_{k1}) = w_{0k1} + \sum_{j=1}^{s} w_{jk1} x_j$$

La propagation rule f

per i livelli nascosti successivi

$$f_{kh}(\mathbf{r}_{h-1}, \boldsymbol{?}_{kh}) = w_{0kh} + \sum_{j=1}^{m_{h-1}} w_{jkh} r_{j(h-1)}$$

per il livello di *output*

$$f_{k(l+1)}(\mathbf{r}_l, \boldsymbol{?}_{k(l+1)}) = w_{0k(l+1)} + \sum_{j=1}^{m_l} w_{jk(l+1)} r_{jl}$$

I neuroni con funzione di attivazione di questo tipo sono detti **neuroni sigma** (***sigma units***).

La propagation rule f

2. Combinazioni lineari del prodotto tra un insieme di *input* ? ***sigma-pi units***
3. Misure di distanza da un centro assegnato ad ogni neurone • **reti *Radial Basis Function***

Il livello di risposta

Il **livello di risposta** r_{kh} del k-esimo neurone nel livello h-esimo è calcolato come funzione dello stato di attivazione a_{kh}:

$r_{kh} = f_{kh}[a_{kh}] = f_{kh}[f_{kh}(\boldsymbol{r}_{h-1}, w_{kh})]$ per i livelli nascosti

$y_{ki} = f_{k(l+1)}\,[f_{k(l+1)}\,(\boldsymbol{r}_l, w_{k(l+1)})]$ per il livello di *output*

La funzione di attivazione o di trasferimento

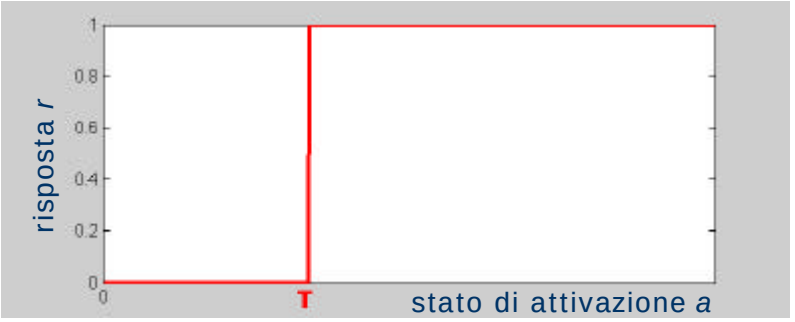
Le funzioni di attivazione *f* possono venire definite in molti modi, a seconda del tipo di reazione che si vuole imporre ai neuroni artificiali.

Esaminiamo ora alcuni esempi, ma vi sono molte altre possibilità.

La funzione di attivazione o di trasferimento

$$f(a_k) = \begin{cases} 0 & se \quad a_k \leq T \\ 1 & se \quad a_k > T \end{cases}$$

funzione *threshold* o a soglia



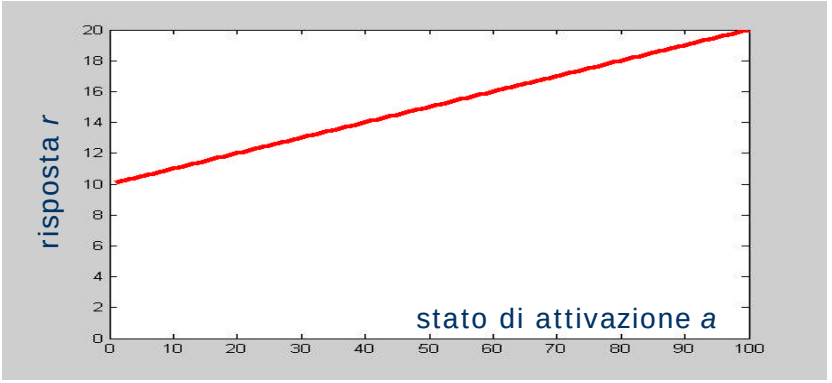
La funzione di attivazione o di trasferimento

$f(a_k) = a_k$ **funzione identità**

La funzione di attivazione o di trasferimento

$$f(a_k) = p_0 + p_1 a_k$$

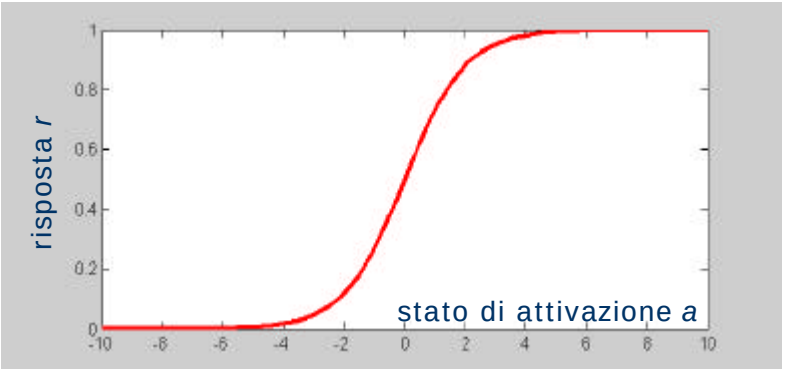
funzione lineare



La funzione di attivazione o di trasferimento

$$f(a_k) = \frac{1}{1 + e^{-a_k}}$$

funzione logistica o sigmoide



Semplificazione della simbologia

La formalizzazione matematica ora introdotta si avvale di una simbologia molto pesante e per molti versi sovrabbondante.

Infatti è definita in modo da consentire la specificazione di una funzione f e di una funzione f differente *per ogni nodo*.

Nella pratica ciò non viene mai fatto anche perché risulta sostanzialmente inutile.

Semplificazione della simbologia

La simbologia si semplifica notevolmente se si ipotizza che le funzioni f e f siano **tutte uguali** in ogni livello o addirittura tutte uguali nell'intera rete.

Nelle applicazioni pratiche, infatti, le funzioni f e f vengono *al più* differenziate per livello.

Fra breve introdurremo la simbologia semplificata, sotto l'ipotesi di funzioni f e f tutte uguali per l'intera rete (riservandoci di introdurre nuovamente il riferimento al livello qualora si rendesse necessario).

Semplificazione della simbologia

Un'ulteriore semplificazione della simbologia si può ottenere eliminando dai pesi w_{jkh}, dagli stati di attivazione a_{kh} e dalle risposte r_{kh}, **il pedice che fa riferimento al livello *h*-esimo**.

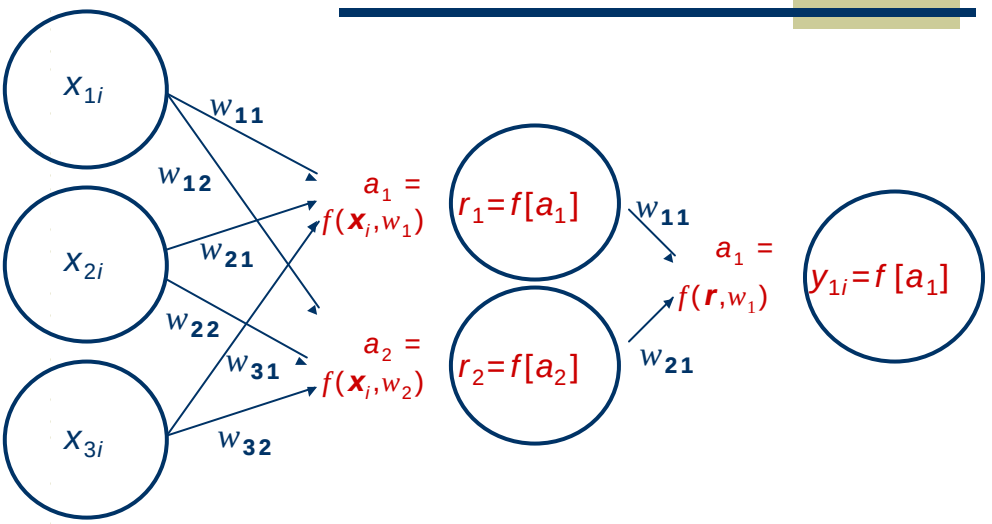
Questa imprecisione è "perdonabile" perché come sarà chiaro più avanti, **non interessa conoscere il valore effettivamente assunto dai parametri**, dagli stati di attivazione dalle risposte: tali grandezze restano tutte all'interno del meccanismo della rete e non hanno alcun interesse pratico.

Semplificazione della simbologia

Ora introdurremo anche questa semplificazione nella simbologia: resta chiaro che in livelli differenti i pesi, gli stati di attivazione e le risposte assumono valori differenti anche qualora siano indicati con lo stesso simbolo che già compare in altri livelli.

Ultima semplificazione: qualora non specificato diversamente intenderemo che la RNA abbia **un solo livello nascosto**. Questa soluzione è infatti la più utilizzata, in quanto in genere più livelli nascosti tendono a risultare sovrabbondanti.

Semplificazione della simbologia



Esercizi

Esercizio 5: dato un neurone con funzione di attivazione a soglia con T = 0, calcolare il livello di risposta ai seguenti stati di attivazione:

- $a_k = -5$
- $a_k = 5$
- $a_k = 15$
- $a_k = -10.5$
- $a_k = -9$
- $a_k = -7$
- $a_k = 1$

Esercizi

Esercizio 6: dato un neurone con funzione di attivazione identità, calcolare il livello di risposta ai seguenti stati di attivazione:

- $a_k = -5$
- $a_k = 5$
- $a_k = 15$
- $a_k = -10.5$
- $a_k = -9$
- $a_k = -7$
- $a_k = 1$

Esercizi

Esercizio 7: dato un neurone con funzione di attivazione lineare con $p_0 = 2$ e $p_1 = 0.5$, calcolare il livello di risposta ai seguenti stati di attivazione:

- $a_k = -5$
- $a_k = 5$
- $a_k = 15$
- $a_k = -10.5$
- $a_k = -9$
- $a_k = -7$
- $a_k = 1$

Esercizi

Esercizio 8: dato un neurone con funzione di attivazione sigmoide, calcolare il livello di risposta ai seguenti stati di attivazione:

- $a_k = -5$
- $a_k = 1.3$
- $a_k = 15$
- $a_k = 0$
- $a_k = -9$
- $a_k = -2$
- $a_k = 1$

Esercizi

Esercizio 9: data una RNA con

- $s = 3$, $l = 1$, $m = 2$, $c = 1$,
- pesi dati da $w_{11} = 1.5$, $w_{21} = 0.7$, $w_{31} = 1$, $w_{12} = 0.5$, $w_{22} = 1.7$, $w_{32} = 2$ nel livello nascosto e $w_{11} = 0.4$, $w_{21} = 0.6$ nel livello di *output*,
- *propagation rule* lineari a tutti i livelli senza pesi w_{0k},
- funzione di attivazione a soglia (T = 2) nel livello nascosto e identità nel livello di *output*,

rappresentare l'architettura della rete e fornire *l'output* relativo *all'input* (0, 1, 1.1)

Esercizi

Esercizio 10: data una RNA con

- $s = 3$, $l = 1$, $m = 2$, $c = 1$,
- pesi dati da $w_{11} = 1.5$, $w_{21} = 0.7$, $w_{31} = 1$, $w_{12} = 0.5$, $w_{22} = 1.7$, $w_{32} = 2$ nel livello nascosto e $w_{11} = 0.4$, $w_{21} = 0.6$ nel livello di *output*,
- *propagation rule* lineari a tutti i livelli senza pesi w_{0k},
- funzione di attivazione identità nel livello nascosto e a soglia (T = 0) nel livello di *output*,

rappresentare l'architettura della rete e fornire *l'output* relativo *all'input* (0.5, 2, 1)

Esercizi

Esercizio 11: data una RNA con

- $s = 3$, $l = 1$, $m = 2$, $c = 1$,
- pesi dati da $w_{11} = 1.5$, $w_{21} = 0.7$, $w_{31} = 1$, $w_{12} = 0.5$, $w_{22} = 1.7$, $w_{32} = 2$ nel livello nascosto e $w_{11} = 0.4$, $w_{21} = 0.6$ nel livello di *output*,
- *propagation rule* lineari a tutti i livelli senza pesi w_{0k},
- funzione di attivazione lineare con $p_0 = 2$ e $p_1 = 0.5$ nel livello nascosto e identità nel livello di *output*,

rappresentare l'architettura della rete e fornire *l'output* relativo *all'input* (0.1, 1, 0)

Esercizi

Esercizio 12: data una RNA con

- $s = 3$, $l = 1$, $m = 2$, $c = 1$,
- pesi dati da $w_{11} = 1.5$, $w_{21} = 0.7$, $w_{31} = 1$, $w_{12} = 0.5$, $w_{22} = 1.7$, $w_{32} = 2$ nel livello nascosto e $w_{11} = 0.4$, $w_{21} = 0.6$ nel livello di *output*,
- *propagation rule* lineari a tutti i livelli senza pesi w_{0k},
- funzione di attivazione sigmoide nel livello nascosto e identità nel livello di *output*,

rappresentare l'architettura della rete e fornire *l'output* relativo *all'input* (0, 1, 1.1)

Metodo di studio

Il testo di riferimento principale per lo studio teorico è:

Zani S., Analisi dei Dati Statistici, Giuffrè Editore, 2000 (Cap. VIII)

Bibliografia di integrazione/approfondimento ed altro materiale didattico sono segnalati sul sito:

http://www.msandri.it/reti_neurali/

Per le applicazioni sarà utilizzato il modulo ***Neural Networks*** di ***MatLab***.