

Le reti neurali

Alberto Borghese

Università degli Studi di Milano
Laboratory of Applied Intelligent Systems (AIS-Lab)
Dipartimento di Informatica
alberto.borghese@unimi.it



A.A. 2021-2022

1/55

<http://borghese.di.unimi.it>



Sommario



Dal neurone artificiale alle reti neurali

L'apprendimento in reti di perceptroni

Esempio con unità lineari ed accenno ad unità non-lineari

A.A. 2021-2022

2/55

<http://borghese.di.unimi.it>



Brains cause minds (J. Searle)



Le reti neurali

Se il neurone biologico consente l'intelligenza, perché non dovrebbe consentire l'intelligenza artificiale un neurone sintetico?

“.. a neural network is a system composed of *many simple processing elements* operating in *parallel* whose function is determined by *network structure, connection strengths*, and the *processing performed at computing elements* or nodes. ... Neural network architectures are inspired by the architecture of biological nervous systems, which use many simple processing elements operating in parallel to obtain high computation rates”. (DARPA, 1988)....

Now, this is called learning with Kernels



A cosa servono?



Le reti neurali offrono i seguenti specifici vantaggi nell'elaborazione dell'informazione:

- Apprendimento basato su esempi (non è richiesta l'elaborazione di un modello aderente alla realtà)
- Autoorganizzazione dell'informazione nella rete
- Robustezza ai guasti (codifica ridondante dell'informazione)
- Funzionamento in tempo reale (realizzazione HW)
- Basso consumo (0.5nW ÷ 4nW per neurone, 20W per il SN) -> calcolatori chimici.
- Modelli di reti neurali per il calcolo.



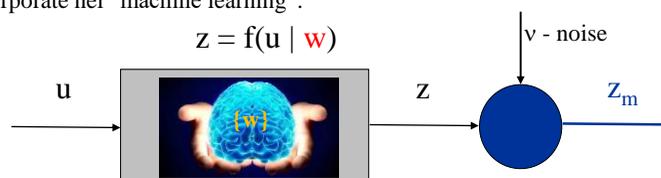
nimi.it



Cosa sono le reti neurali artificiali?



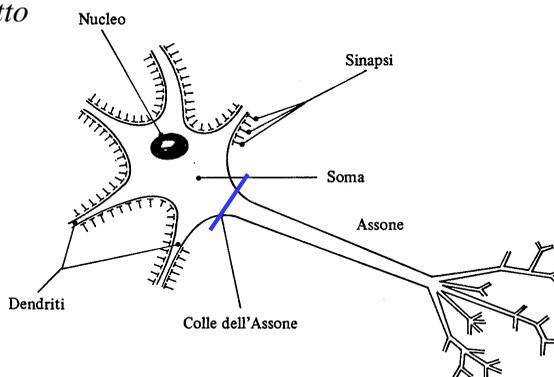
- Le reti neurali sono modelli non lineari per l'**approssimazione** della soluzione di problemi dei quali non esiste un modello con parametri semantici (o se esiste è troppo oneroso computazionalmente). I parametri dei modelli risultanti (semiparametrici) vengono calcolati mediante l'utilizzo di esempi (dati di ingresso e uscita desiderata). Connessioni con il dominio della statistica.
- Vengono utilizzate soprattutto per la classificazione e la regressione.
- Sono un capitolo importante negli argomenti di intelligenza artificiale.
- Da un altro punto di vista possono essere utilizzate per lo studio delle reti neurali naturali, ovvero dei processi cognitivi.
- Sono state incorporate nel "machine learning".





Il neurone artificiale

- *Potenziale di azione (tutto o nulla).*
- *Integrazione nel soma.*
- *Soglia di attivazione.*



Neurone come elemento di calcolo universale: in grado di calcolare qualsiasi funzione logica (cioè implementabile in un computer).

A.A. 2021-2022

7/55

<http://borgnese.di.unimi.it>



Il modello di McCulloch-Pitts

• La variazione delle forma d'onda del potenziale di membrana lungo il dendrita non viene considerata.

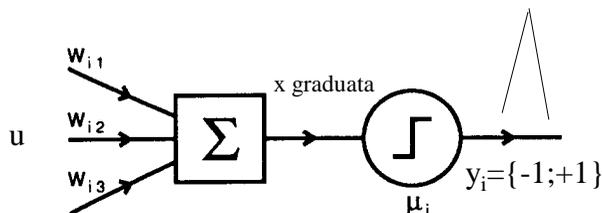
• Gli input non sono sincroni.
$$y_i(t+1) = \Theta \left(\sum (w_{ij}u_j(t)) - \mu_i \right)$$

• Le interazioni tra input non sono lineari.

• I pesi sono supposti costanti.

$$\Theta(x) = \begin{cases} 1 & \text{se } x \geq 0 \\ -1 & \text{altrimenti} \end{cases}$$

Sono state pensate per calcolare **funzioni logiche (V o F)**.
Usate oggi per i **classificatori binari**.



A.A. 2021-2022

McCulloch-Pitts (1943)

it



Rappresentazione della retta

$$x_0 = -1$$

$$w_0 = \mu$$

Se:

$$x_1 = 0$$

$$x_2 = 0$$

$$y = +1$$

x_2

La retta taglia in due il piano

True
 $x_2 > x_1 - 0.5$

$$w_1x_1 + w_2x_2 + w_0x_0 = 0$$

$$\mathbf{w} \cdot \mathbf{x} = 0$$

False
 $x_2 < x_1 - 0.5$

$$x_2 = x_1 - 0.5$$

Se $x_1 = 1, x_2 = 0, y = -1$ - OK



x_1

$$w_1x_1 + w_2x_2 - \mu = 0$$

$$w_1x_1 + w_2x_2 + w_0x_0 = 0$$

$$y_i(t+1) = \Theta \left(\sum (w_{ij}u_j(t)) - \mu_i \right)$$

$$y_i(t+1) = \Theta \left(\sum w_{ij}u_j(t) \right)$$

A.A. 2021-2022

9/55

<http://borghese.di.unimi.it>

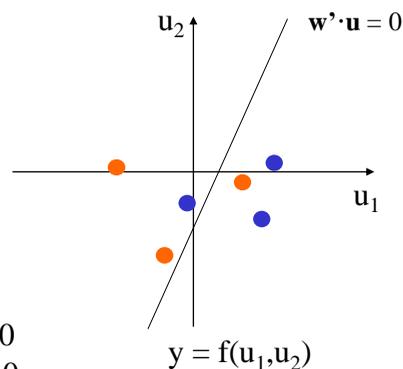
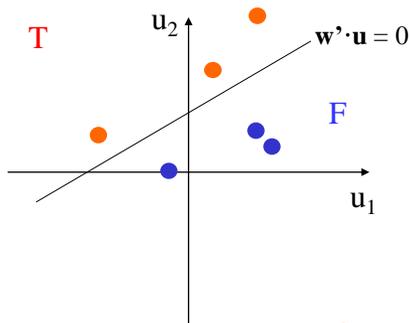


Funzioni linearmente separabili

$$w_1u_1 + w_2u_2 - \mu = 0$$

Linearmente separabile

Non linearmente separabile



● $y > 0$
● $y < 0$

Rilevante per i **classificatori** (boosting, SVM)

Classificare vuol dire trovare la linea di separazione tra i dati

A.A. 2021-2022

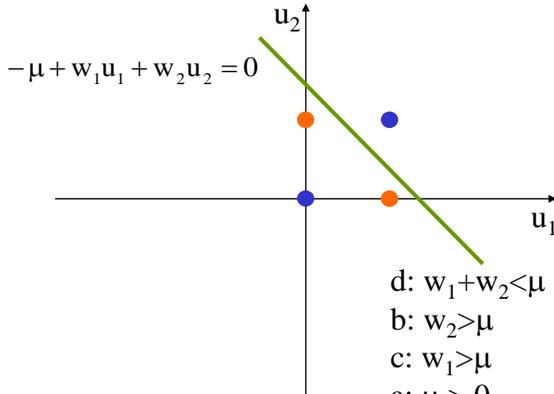
<http://borghese.di.unimi.it>



La "morte" del neurone di McCulloch-Pitts (Minsky, 1969): XOR



$$w_1 u_1 + w_2 u_2 - \mu = 0 \quad \mathbf{w} \cdot \mathbf{u} = 0 \quad u_0 = -1 \quad w_0 = -\mu$$



u_1	u_2	y
0	0	-1
0	1	1
1	0	1
1	1	-1

a
b
c
d

- $y(u_1, u_2, 1) = 1$
- $y(u_1, u_2, 1) = -1$

- d: $w_1 + w_2 < \mu$
- b: $w_2 > \mu$
- c: $w_1 > \mu$
- a: $\mu > 0$

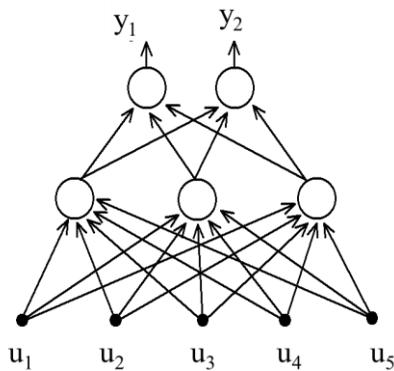
Il sistema di 4 equazioni non è risolvibile.

$w_1, w_2 > \mu$ e $w_1 + w_2 < \mu$ Impossibile!!

A.A. 202 Si possono imparare solamente funzioni linearmente separabili nimi.it



Spiking neurons

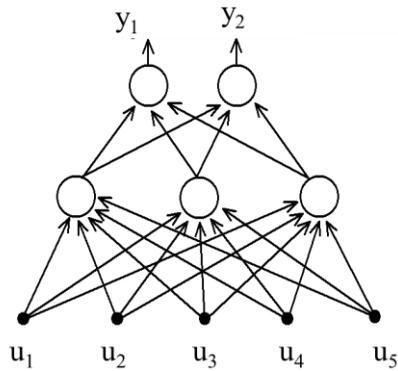


Spiking neurons. Sono neuroni la cui uscita è il singolo spike. Modellazione realistica (e.g. McCullochPitts). **Spike del neurone.**

Segnale binario: spike / non spike + elaborazione continua all'interno del «soma».



Neuroni artificiali «moderni»



Connessionismo classico.

Uscita compresa tra min – Max.

Frequenza di scarica.

Uscita di ogni neurone compresa tra 0 e 1 (o tra -1 e 1):
 $y_i = [0 \ 1]$.

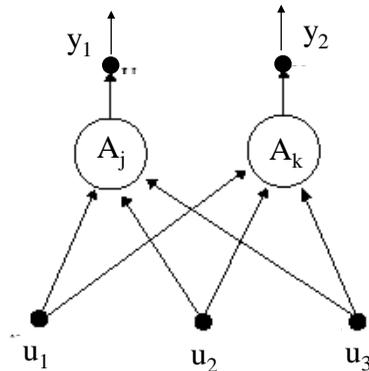
L'uscita può essere binarizzata definendo una soglia.



La rete neurale ad un livello



La rete opera una trasformazione dallo spazio di input allo spazio di output.



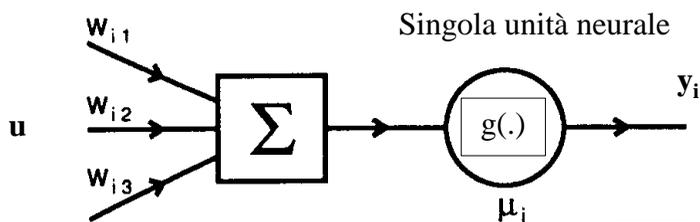
$$y_i = g(\sum_j (w_{ij} u_j) - \mu_i)$$

La trasformazione o mappatura dipende dai parametri $\{w_{ij}\}$ e $\{\mu_i\}$ in modo tale che la rete neurale approssimi la trasformazione tra i pattern di input e di output.

Se $g(\cdot) = 1$, la rete diventa un modello lineare: $y_i = \sum_j (w_{ij} u_j) - \mu_i$



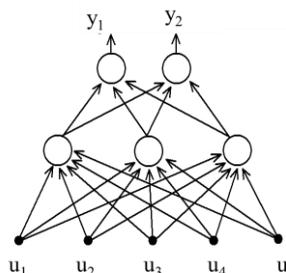
Una rete neurale a più livelli



$$y_i = g(\sum_j (w_{ij}u_j) - \mu_i)$$

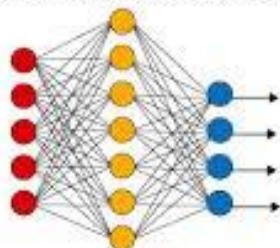
Unità nascoste – Hidden layer

Deep neural networks -> multi-livelli
(+ operazioni di sotto-campionamento, gather-scatter...)

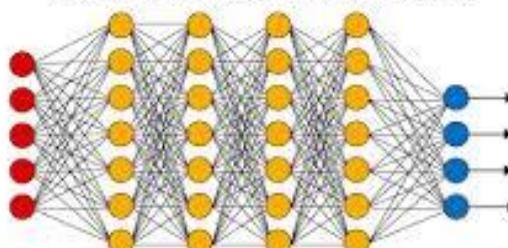


Deep Neural Networks

Simple Neural Network



Deep Learning Neural Network

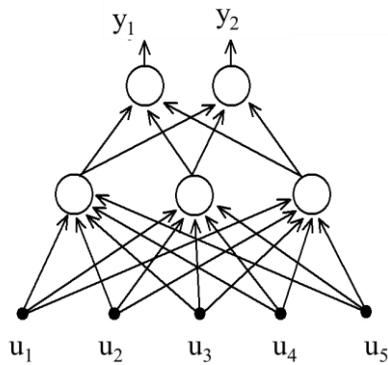


● Input Layer ● Hidden Layer ● Output Layer

Convolutional layers (e.g. RBF networks) - filters
Pooling
Subsampling



Caratteristiche



Livelli di unità di attivazione

Collegamento in cascata

Input convergenti, output divergenti.

Capacità di approssimazione universale

Perceptrone: layered networks, flusso unidirezionale dell'elaborazione.

L'output viene interpretato come frequenza di scarica del neurone d'uscita della rete.

A.A. 2021-2022

17/55

<http://borghese.di.unimi.it>



Complessità della funzione realizzabile

Quanti più neuroni artificiali vengono connessi tanto più la funzione complessiva approssimabile diviene più complessa

$$Y = |y_1, y_2, y_3, \dots, y_n|^T$$

$$X = |x_1, x_2, x_3, \dots, x_m|^T$$

$$y = F(X)$$

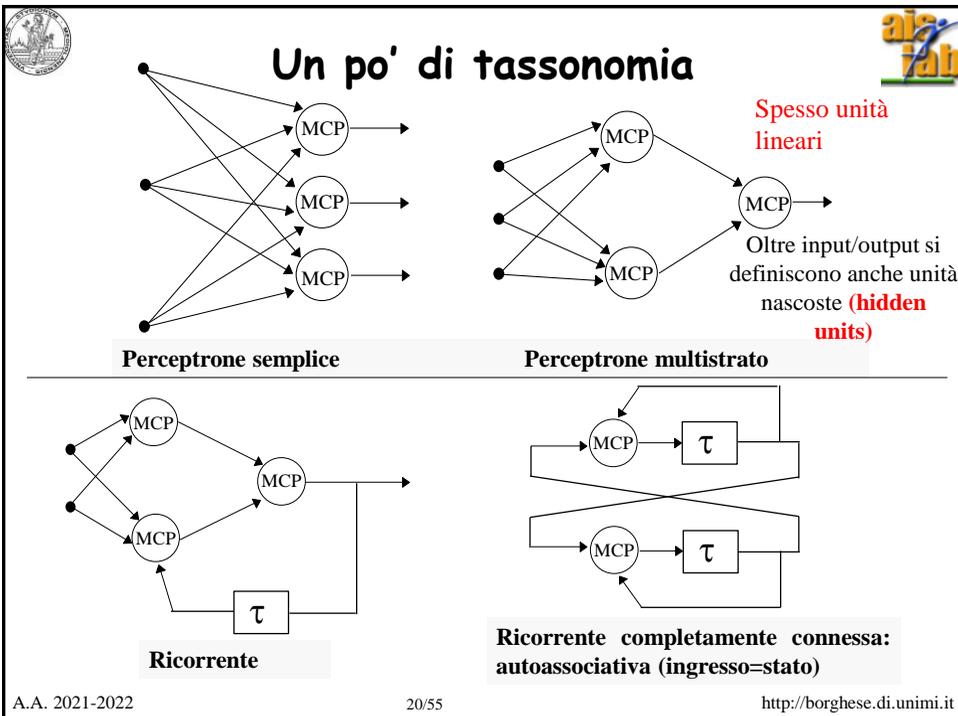
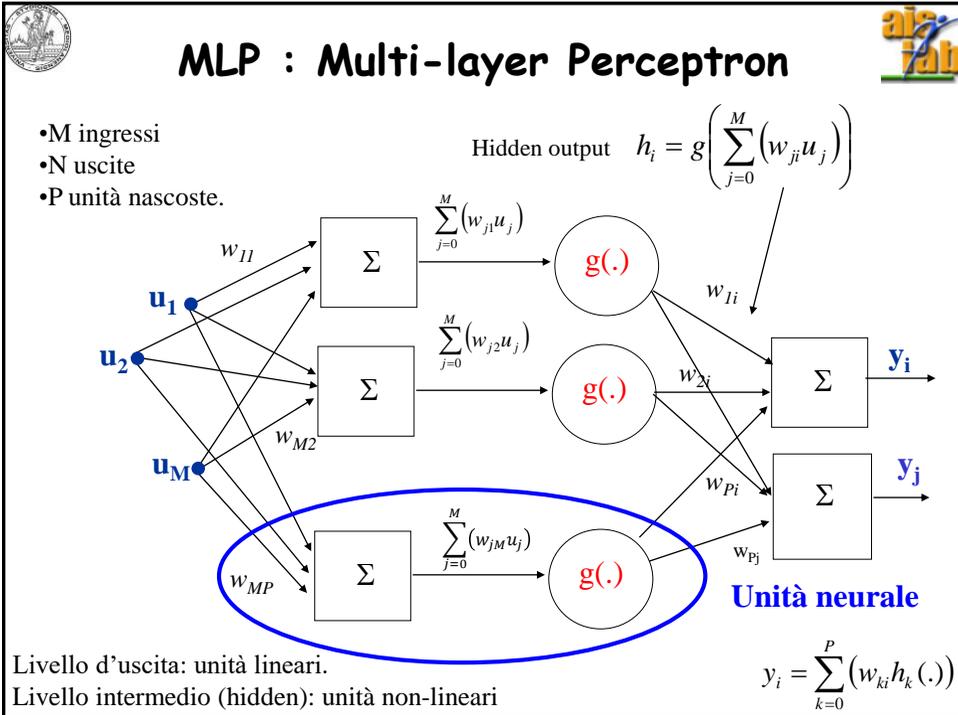


Reti neurali = approssimatori universali.

A.A. 2021-2022

18/55

<http://borghese.di.unimi.it>





Sommario



Dal neurone artificiale alle reti neurali

L'apprendimento in reti di perceptroni

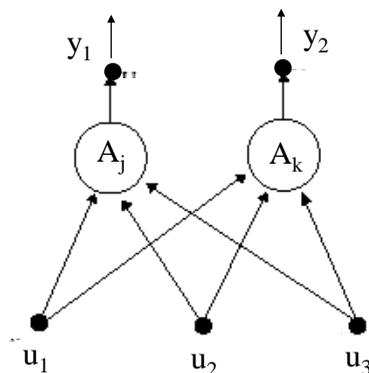
Esempio con unità lineari ed accenno ad unità non-lineari



Lo spirito dell'apprendimento supervisionato



La rete opera una trasformazione dallo spazio di input allo spazio di output.



Apprendimento è la modifica dei parametri $\{w_{ij}\}$ e $\{\mu_i\}$ in modo tale che la rete neurale approssimi la trasformazione tra i pattern di input e di output.

$$y_i = g(\sum_j (w_{ij} u_j) - \mu_i)$$

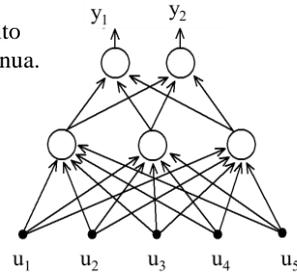


Funzione costo per unità di attivazione continue



Possiamo derivare una regola di apprendimento di spirito **Hebbiano** per una qualsiasi funzione di attivazione continua. Consideriamo un perceptrone ad un livello.

$$y = g \left(\sum_{j=1} w_{ij} u_j - \mu_i \right) = g \left(\sum_{j=0} (w_{ij} u_j) \right)$$



Si tratta di un problema di minimizzazione di una cifra di merito, J , sullo spazio dei parametri W :

$$E(w) = \underbrace{\|y^D - g(W^{nuovo}U)\|}_{\text{Errore}} \leq \|y^D - g(W^{vecchio}U)\|$$

Devo trovare $\{w\}$: $E(w)$ è minimo.

$$E(w) = \frac{1}{2} \sum_p \left[\sum_j (y_{jp}^D - y_{jp})^2 \right] = \frac{1}{2} \sum_p \left[\sum_j \left(y_{jp}^D - g \left(\sum_i w_{ij} u_{ip} \right) \right)^2 \right]$$

su tutti i pattern p



Apprendimento supervisionato



$$\min_{\{w\}} E(.) \quad E = \|Y^D - g(W^{nuovo}U)\| \leq \|Y^D - g(W^{vecchio}U)\|$$

Y^D è l'uscita desiderata nota.

Si tratta di un problema di minimizzazione di una cifra di merito (E) sullo spazio dei parametri W .

Soluzione iterativa (gradiente):

Obiettivo: se esiste una soluzione, trovare ΔW in modo iterativo tale che l'insieme dei pesi W^{nuovo} ottenuto come:

$$W^{nuovo} = W^{vecchio} + \Delta W$$

dia luogo a un errore sulle uscite di norma minore che con $W^{vecchio}$ (si parte da un W_0 iniziale, arbitrario).



Minimizzazione di funzioni di più variabili



$\min(E\{\mathbf{w}\} | \dots)$ funzione costo o errore

$$\text{Gradiente: } \nabla E(\mathbf{w}) = \frac{\delta E(\{\mathbf{w}\} | \dots)}{dw_1} + \frac{\delta E(\{\mathbf{w}\} | \dots)}{dw_2} + \frac{\delta E(\{\mathbf{w}\} | \dots)}{dw_3} + \frac{\delta E(\{\mathbf{w}\} | \dots)}{dw_4} + \dots$$

Modifico il valore dei pesi di una quantità proporzionale alla pendenza della funzione costo rispetto a quel parametro.

Estensione della tecnica del gradiente a più variabili.

$$\Delta \mathbf{w} = -\eta \nabla E(\mathbf{w}) \Leftrightarrow \Delta w_{ij} = -\eta \frac{\delta E(\{\mathbf{w}\} | \dots)}{dw_{ij}}$$

Serve un' **approssimazione iniziale** per i pesi $W_{ini} = \{w_j\}_{ini}$.



La pratica dell'apprendimento supervisionato



Fino a quando l'apprendimento non è stato completato:

Forward pass;

1. Presentazione di un pattern di input / output (**dati**).
2. Calcolo dell'output della rete con il pattern corrente (**modello**).
3. Calcolo di una misura di errore (**distanza, discrepanza, entropia**).

Backwards pass;

4. Calcolo dei gradienti (**apprendimento**).
5. Calcolo dell'incremento dei pesi (**apprendimento**).

Aggiornamento dei pesi:

- Per trial (ogni pattern)
- Per epoca (ogni insieme di pattern).



Apprendimento supervisionato tramite gradiente



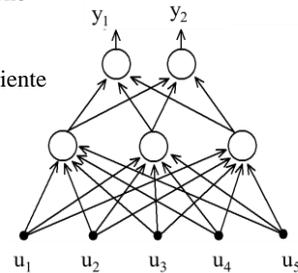
Coppie input/output note.

Definizione di una **funzione costo** che misuri l'errore sull'uscita (abbiamo visto la norma).

Modifica dei valori dei pesi in modo tale che la funzione costo sia minimizzata.

Reti multi-strato hanno elevata capacità computazionale, ma anche elevata complessità.

La funzione di uscita è ottenuta come funzione di funzioni, gradiente complesso.



A.A. 2021-2022

27/55

<http://borgnese.di.unimi.it>

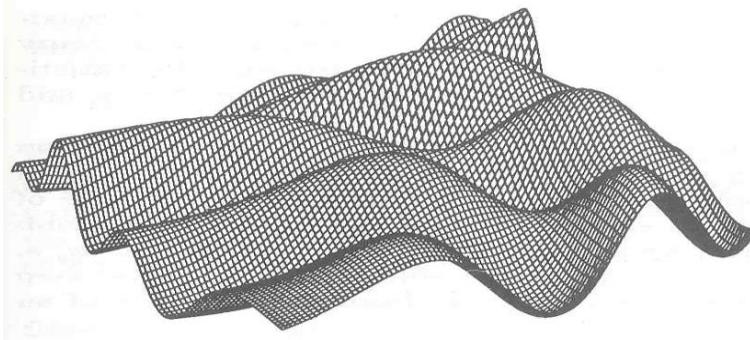


Problemi nell'apprendimento supervisionato tramite gradiente



- Nota: W_{mi} è generalmente casuale e può condizionare la convergenza degli algoritmi iterativi. Solitamente si sceglie molto piccolo e random.

- I problemi di convergenza sono legati all'esistenza di minimi locali del funzionale $J(w | \dots)$



A.A. 2021-2022

28/55

<http://borgnese.di.unimi.it>



Sommario



Dal neurone artificiale alle reti neurali

L'apprendimento in reti di perceptroni

Esempio con unità lineari ed accenno ad unità non-lineari



Modelli lineari e non lineari



Classificazione alternativa dei modelli. Vengono utilizzate classi molto diversi di algoritmi per stimare i parametri di questi due tipi di modelli.

$$z(p(x, y)) = f(x) = \sum_i w_i x$$

$$z(p(x, y)) = \sum_i f_i(p; w)$$

$f(\cdot)$ è funzione lineare nei $\{w_i\}$

$f(\cdot)$ è funzione non lineare

e.g. $f(\cdot) = e^{w \cdot x}$
 $f(\cdot) = x \ln(w \cdot x)$
....

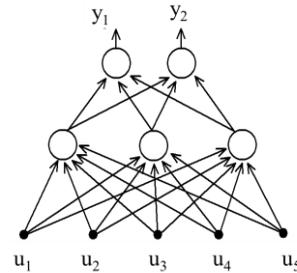


Unità lineari, soluzione iterativa



$$J = E(\mathbf{w}) = \frac{1}{2} \sum_p \left[\sum_j (y_{jp}^D - y_{jp})^2 \right] = \frac{1}{2} \sum_p \left[\sum_j \left(y_{jp}^D - \left(\sum_i w_{ij} u_{ip} \right) \right)^2 \right]$$

$$\Delta w_{ij} = -\eta \frac{\partial}{\partial w_{ij}} \frac{1}{2} \sum_j \left(y_j^D - \left(\sum_i w_{ij} u_i \right) \right)^2$$



$$\Delta w_{ij} = +\eta \sum_j \left(y_j^D - \sum_k w_{kj} u_k \right) u_i = +\eta \sum_j (y_j^D - y_j) u_i$$

Hebbian learning

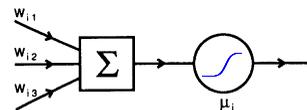
↗ δ rule (Hoff, 1960)



Unità di attivazione non-lineari



$$y_j = g \left(\sum_{i=1}^M w_{ij} u_i - \mu_j \right) = g \left(\sum_{i=0}^M (w_{ij} u_i) \right)$$



$\sum_{i=0}^M (w_{ij} u_i)$ è l'argomento della funzione di attivazione $g(\cdot)$

Apprendimento: minimizzazione dell'errore, $E(y^D, x^D | \mathbf{w})$, sui pattern tra l'uscita desiderata prescritta e quella fornita dal modello.

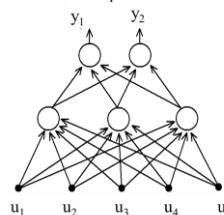


Unità non-lineari, soluzione iterativa



$$J = E(y^D, x^D | \mathbf{w}) = \frac{1}{2} \sum_p \left[\sum_j (y_{jp}^D - y_{jp})^2 \right] = \frac{1}{2} \sum_p \left[\sum_j \left(y_{jp}^D - g \left(\sum_i w_{ij} u_{ip} \right) \right)^2 \right]$$

$$\Delta w_{ijp} = -\eta \frac{\partial}{\partial w_{ij}} \frac{1}{2} \sum_j \left(y_{jp}^D - g \left(\sum_i w_{ij} u_{ip} \right) \right)^2 =$$



$$\Delta w_{ij} = +\eta \sum_j \left(y_{jp}^D - g \left(\sum_k w_{kj} u_{kp} \right) \right) g' \left(\sum_k w_{kj} u_{kp} \right) u_i =$$

$$+\eta \underbrace{\sum_j (y_j^D - y_j)}_{\delta \text{ rule}} g' \left(\sum_k w_{kj} u_{kp} \right) u_i$$

A.A. 2021-2022

33/55

<http://borghese.di.unimi.it>



Perceptrone con unità di attivazione logistiche

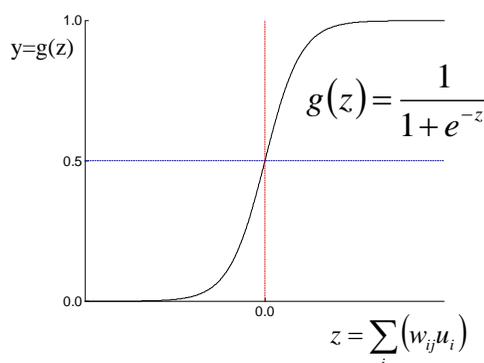


$$y_j = g \left(\sum_i w_{ij} u_i \right) \quad \left(\sum_i w_{ij} u_i \right) = z$$

$$g'(z) = \frac{e^{-z}}{(1+e^{-z})^2} =$$

$$= \frac{1}{1+e^{-z}} \left(1 - \frac{1}{1+e^{-z}} \right) =$$

$$g'(z) = g(z) \cdot (1 - g(z))$$



A.A. 2021-2022

34/55

<http://borghese.di.unimi.it>



Update dei pesi per funzione logistica



$$J = E(\mathbf{w}) = \frac{1}{2} \sum_p \left[\sum_j (y_{jp}^D - y_{jp})^2 = \frac{1}{2} \sum_j \left(y_{jp}^D - g\left(\sum_i w_{ij} u_{ip}\right) \right)^2 \right]$$

$$\Delta w_{ijp} = +\eta \sum_j (y_{jp}^D - g(\cdot)) g'(\cdot) u_i = +\eta \underbrace{(y_{jp}^D - y_j)}_{\delta \text{ rule}} \underbrace{u_{ip} y_j (1 - y_j)}_{\text{derivata}}$$

NB $y_i \in [0, 1]$. Per $y_i = 0$ o $y_i = 1$ non c'è apprendimento anche se l'uscita è sbagliata. Quando si verifica questa situazione?

Si cerca di mantenere le unità lontane della saturazione.



Funzioni costo

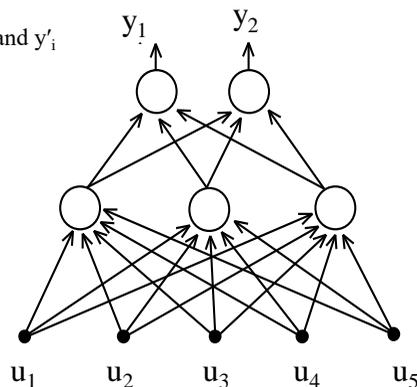


$$\| \mathbf{Y}^D - g(\mathbf{W} \mathbf{U}) \|^2 \quad \text{Squared loss function. Gaussian error.}$$

For classification

y_i is the predicted probability value for class i and y'_i is the true probability for that class.

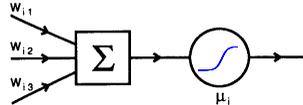
$$H_{y'}(y) = - \sum_i y'_i \log(y_i) \quad \text{cross-entropy loss}$$





Unità di attivazione lineari



$$y_j = g\left(\sum_{i=1}^M w_{ij}u_i - \mu_j\right) = g\left(\sum_{i=0}^M (w_{ij}u_i)\right)$$


Caso lineare ($g(\cdot) = I$):

$$y_j = \sum_{i=1} w_{ij}u_i - \mu_j = \sum_{i=0} (w_{ij}u_i) \quad \implies \quad \mathbf{Y} = \mathbf{W} \mathbf{U}$$

Soluzione di un sistema lineare nei pesi!!

Condizione di risolubilità: \mathbf{W} di rango massimo \rightarrow
 $\{w\}$ sono linearmente indipendenti.



Unità lineari, soluzione iterativa



$$J = E(\mathbf{w}) = \frac{1}{2} \sum_p \left[\sum_j (y_{jp}^D - y_{jp})^2 \right] = \frac{1}{2} \sum_p \left[\sum_j \left(y_{jp}^D - \left(\sum_i w_{ij}u_{ip} \right) \right)^2 \right]$$

$$\Delta w_{ij} = -\eta \frac{\partial}{\partial w_{ij}} \frac{1}{2} \sum_j \left(y_j^D - \left(\sum_i w_{ij}u_i \right) \right)^2$$

$$\Delta w_{ij} = +\eta \sum_j \left(y_j^D - \left(\sum_j w_{ij}u_i \right) \right) u_i = +\eta \sum_j (y_j^D - y_j) u_i$$

Hebbian learning

δ rule (Hoff, 1960)

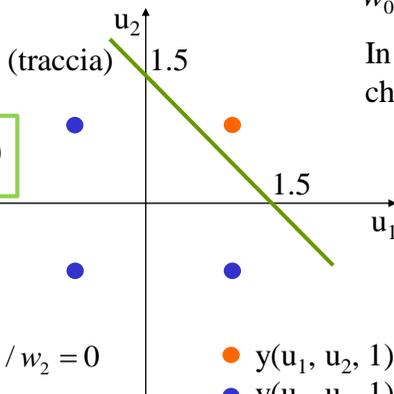


Esempio - AND



Troviamo la soluzione graficamente

$$u_2 + u_1 - 1.5 = 0$$



$$w_0 u_0 + w_1 u_1 + w_2 u_2 = 0$$

In verde la retta $w' \cdot u = 0$ che taglia il piano $u_1 u_2$.

$$u_0 = -1 \\ w_0 = 1.5$$

$$w_1 = 1 \\ w_2 = 1$$

$$u_2 + (w_1 / w_2) u_1 + w_0 / w_2 = 0$$

$$u_2 + u_1 - 1.5 = 0$$

⇓

$$w_1 / w_2 = 1 \quad w_0 / w_2 = -1.5 \quad \Rightarrow \quad w_2 = k \quad w_1 = k \quad w_0 = 1.5 * k$$

Esistono più soluzioni
Separabilità lineare.

A.A. 2021-2022

39/55

<http://borghese.di.unimi.it>



$$w_0 u_0 + w_1 u_1 + w_2 u_2 = 0$$

Problema lineare



Matrice dei termini noti: $b = y^D$

$$b = \begin{bmatrix} -1 \\ -1 \\ -1 \\ +1 \end{bmatrix}$$

u_1	u_2	y	y^D
-1	-1	-1	-1
-1	1	+1	-1
1	-1	-1	-1
1	1	-1	+1

Matrice dei coefficienti: A

$$A = \begin{bmatrix} +1 & -1 & -1 \\ +1 & +1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & +1 \end{bmatrix}$$

$u_0 \quad u_2 \quad u_1$

Vettore delle incognite: $x = w$

$$Ax = b \\ x = (A' * A)^{-1} * A' * b \quad \rightarrow \quad w = \begin{bmatrix} +0.5 \\ +0.5 \\ -0.5 \end{bmatrix}$$

Soluzione ottima: minimizzo implicitamente la distanza tra la retta ed i 4 punti.

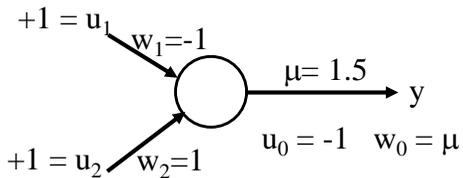
i.it



Calcolo dell'uscita



Inizializzo i pesi: $w_1 = -1$, $w_2 = 1$, $w_0 = 1.5$



u_1	u_2	y	y^D
-1	-1		-1
-1	1		-1
1	-1		-1
1	1	1.5	+1

$U = |1, 1|$ $y^D = +1$

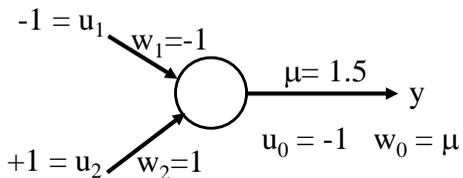
$$y = \sum_{i=1} w_i u_i - \mu = \sum_{i=0} (w_i u_i) = (-1)(1) + (1)(1) + (1.5)(-1) = -1.5 \ll +1$$



Soluzione iterativa Delta rule I: calcolo dell'uscita



Inizializzo i pesi: $w_1 = -1$, $w_2 = 1$, $w_0 = 1.5$



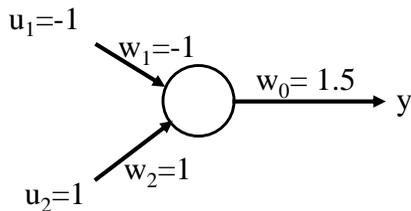
u_1	u_2	y	y^D
-1	-1		-1
-1	1	0.5	-1
1	-1		-1
1	1		+1

$U = |-1, 1|$ $y^D = -1$

$$y = \sum_{i=1} w_i u_i - \mu = \sum_{i=0} (w_i u_i) = (-1)(-1) + (1)(1) + (-1.5)(+1) = 0.5 \gg -1$$



Delta rule II: Calcolo dell'errore



u_1	u_2	y	y^D
-1	-1		-1
-1	1	0,5	-1
1	-1		-1
1	1		+1

$$y = \sum_{i=0} (w_i u_i) = 0.5$$

$$\text{Errore} = 1/2(y^D - y)^2 = 1/2(-1 - (0.5))^2 = 0.5 * (-1.5^2)$$



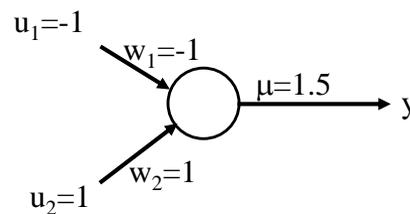
Delta rule III: calcolo del gradiente



$$\frac{d\text{Errore}}{dw_1} = (y_i^D - y_i) u_1 = (-1 - 0.5)(-1) = 1.50$$

$$\frac{d\text{Errore}}{dw_2} = (y_i^D - y_i) u_2 = (-1 - 0.5)(+1) = -1.50$$

$$\frac{d\text{Errore}}{d\mu} = -\frac{d\text{Errore}}{dw_0} = +1.50$$





Delta rule IV: aggiornamento pesi



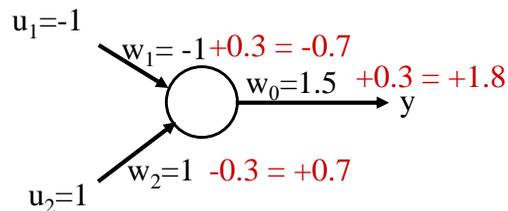
$$\Delta w_{ij} = +\eta(y_j^D - y_j)u_i \quad U = |-1, 1| \quad y^D = -1$$

$$\eta = 0.2$$

$$\Delta w_1 = \eta(y_i^D - y_i)u_1 = \eta(-1 - 0.5)(-1) = +0.30$$

$$\Delta w_2 = \eta(y_i^D - y_i)u_2 = \eta(-1 - 0.5)(+1) = -0.30$$

$$\Delta w_0 = \eta(y_i^D - y_i)u_0 = \eta(-1 - 0.5)(-1) = +0.30$$



A.A. 2021-2022

45/55

<http://borghese.di.unimi.it>



Delta rule V: Nuovo valore di uscita

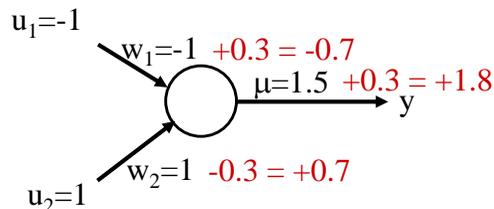


$$U = |-1, 1| \quad y^D = -1$$

$$\eta = 0.2$$

$$y = \sum_{i=1} (w_i u_i - \mu) =$$

$$\sum_{i=0} (w_i u_i) = -0.4 > -1 \quad \text{ma negativo} \rightarrow \text{Ok}$$



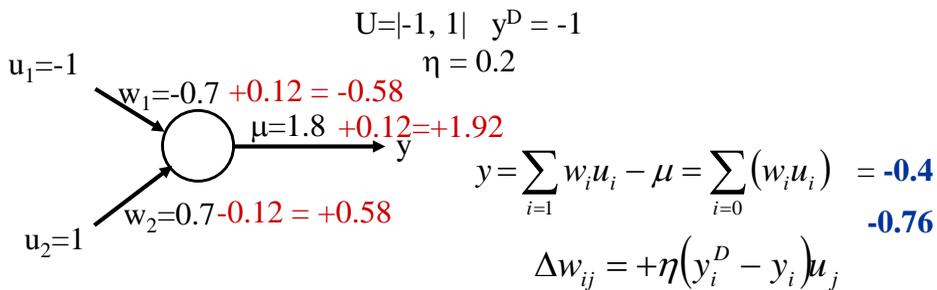
A.A. 2021-2022

46/55

<http://borghese.di.unimi.it>



Delta rule VI- Nuovo aggiornamento



$$\Delta w_1 = \eta (y_i^D - y_i) u_1 = \eta (-1 - 0.4) (-1) = +0.12$$

$$\Delta w_2 = \eta (y_i^D - y_i) u_2 = \eta (-1 - 0.4) (+1) = -0.12$$

$$\Delta w_0 = \eta (y_i^D - y_i) u_0 = \eta (-1 - 0.4) (-1) = +0.12$$

Che relazione c'è tra i pesi e la retta che separa le uscite positive da quelle negative?

A.A. 20

ni.it



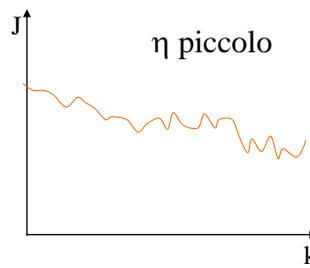
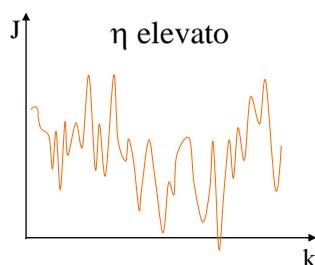
Ruolo di η – learning rate



$$\Delta w_{ij} = +\eta (y_j^D - y_j) u_i$$

Calmiera il Δw_{ij} per evitare che :

- Un peso sia specifico di un'unità ingresso-uscita.
- Oscillazioni durante l'apprendimento senza convergenza.



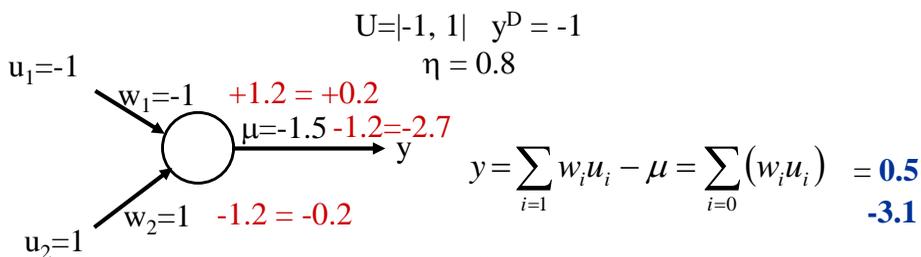
η può variare durante l'addestramento.

A.A. 2021-2022

<http://borghese.di.unimi.it>



Esempio di delta rule - Cattiva scelta di η



$$\Delta w_{ij} = +\eta (y_j^D - y_j) \mu_i$$

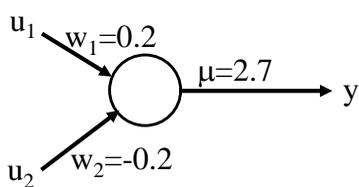
$$\Delta w_1 = \eta (y_1^D - y_1) u_1 = \eta (-1 - 0.5) (-1) = +1.2$$

$$\Delta w_2 = \eta (y_1^D - y_1) u_2 = \eta (-1 - 0.5) (+1) = -1.2$$

$$\Delta w_0 = \eta (y_1^D - y_1) u_0 = \eta (-1 - 0.5) (-1) = +1.2$$



Esempio di specializzazione sul pattern b



u_1	u_2	y^D
-1	-1	-1
-1	1	-1
1	-1	-1
1	1	1

a
b
c
d

a $y = \sum w_i u_i - \mu = \sum (w_i u_i) = (0.2)(-1) + (-0.2)(1) - 2.7 = -3.1$

b $y = \sum_{i=1} w_i u_i - \mu = \sum_{i=0} (w_i u_i) = (0.2)(-1) + (-0.2)(1) - 2.7 = -2.9$

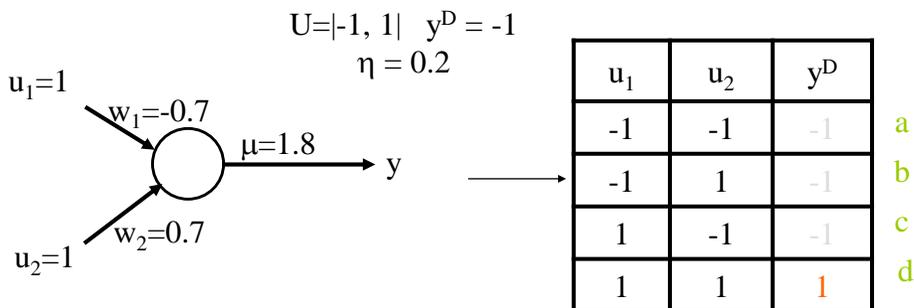
c $y = \sum_{i=1} w_i u_i - \mu = \sum_{i=0} (w_i u_i) = (0.2)(1) + (-0.2)(-1) - 2.7 = -2.3$

d $y = \sum_{i=1} w_i u_i - \mu = \sum_{i=0} (w_i u_i) = (0.2)(1) + (-0.2)(1) - 2.7 = -2.7$

Errato su d. Specializzazione su a, b, c



Uscita per il pattern d



$$y = \sum_{i=1} w_i u_i - \mu = \sum_{i=0} (w_i u_i) = -1.8 \ll +1$$



Riassunto - topologia



I neuroni connessionisti sono basati su:

- Ricevere una somma pesata degli ingressi.
- Trasformarla secondo una funzione non-lineare (scalino o logistica)
- Inviare il risultato di questa funzione all'uscita o ad altre unita'.

Le reti neurali sono topologie ottenute connettendo tra loro i neuroni in modo opportuno e riescono a calcolare funzioni molto complesse.



Riassunto - Apprendimento



Algoritmi iterativi per adattare il valore dei parametri (pesi).

Definizione di una funzione costo che misura la differenza tra valore fornito e quello desiderato.

Algoritmo (gradiente) che consente di aggiornare i pesi in modo da minimizzare la funzione costo.

Training per pattern (specializzazione) o per epoche.



Direzioni di sviluppo della ricerca



1) **Spiking neurons models** – Modelli computazionali a neurone singolo. Evoluzione temporale.

2) **Continuous activation models** – Apprendimento di pattern.

Deep learning is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers with complex structures, or otherwise composed of multiple non-linear transformations (e.g. tensor flow in Keras <https://keras.io/>)

Capsule networks. Modules connected.



Sommario



Dal neurone artificiale alle reti neurali

L'apprendimento in reti di perceptroni

Esempio con unità lineari ed accenno ad unità non-lineari