

Metodi Matematici per il Machine Learning

Davide Peccioli

a.a. 2024-25

Indice

I	De Rossi	5
1	Reti Neurali	7
1.1	Neurone Artificiale	7
1.1.1	Neurone Sigma-Heaviside	8
1.1.2	Regressione Lineare	9
1.2	Rete Neurale	10
1.3	Funzioni Costo (Machine Learning)	10
1.3.1	La Funzione Errore Supremum	10
1.3.2	La Funzione Errore Norma L2	10
2	Cenni di Teoria della Misura	13
2.1	Funzioni sigmoidali e funzioni discriminatorie	13
II	Cordero	15
III	Sirovich	17

Parte I

De Rossi

Capitolo 1

Reti Neurali

1.1 Neurone Artificiale

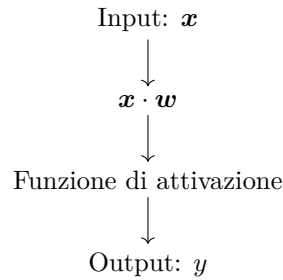
Un neurone è una cellula del corpo umano che può essere schematizzata come segue:

A neuron is a cell which consists of the following parts: dendrites, axon, and body-cell. The synapse is the connection between the axon of one neuron and the dendrite of another. The functions of each part is briefly described below:

- Dendrites are transmission channels that collect information from the axons of other neurons. The signal traveling through an axon reaches its terminal end and produces some chemicals x_i which are liberated in the synaptic gap. These chemicals are acting on the dendrites of the next neuron either in a strong or a weak way. The connection strength is described by the weight system w_i .
- The body-cell collects all signals from dendrites. Here the dendrites activity adds up into a total potential and if a certain threshold is reached, the neuron fires a signal through the axon. The threshold depends on the sensitivity of the neuron and measures how easy is to get the neuron to fire.
- The axon is the channel for signal propagation. The signal consists in the movement of ions from the body-cell towards the end of the axon. The signal is transmitted electrochemically to the dendrites of the next neuron.

Matematicamente, quindi, si considera un neurone come una unità che riceve degli input (un vettore \mathbf{x}), lo [moltiplica](#) per un vettore di pesi $\mathbf{w} = (w_0, \dots, w_n)$, e produce un output processando il

prodotto scalare tramite una funzione di attivazione:

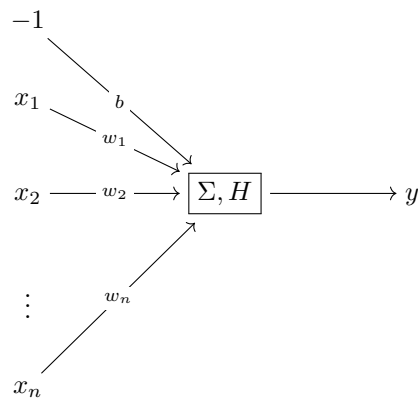


1.1.1 Neurone Sigma-Heaviside

Il modello più semplice è quello che riceve degli input, li somma dopo averli moltiplicati per dei pesi, e:

- restituisce 0 se la somma così ottenuta non supera un threshold b ;
- restituisce 1 se la somma così ottenuta è maggiore o uguale a b .

Questo viene schematizzato in questo modo:



e l'output y è dato da¹

$$y = H \left(\sum_{i=0}^n x_i w_i \right)$$

dove per convenzione si è posto $w_0 = b$ e $x_0 = -1$.

La convenzione è per semplicità di notazione, infatti

$$H \left(\sum_{i=0}^n x_i w_i \right) = \begin{cases} 1 & \sum_{i=1}^n x_i w_i \geq b \\ 0 & \sum_{i=1}^n x_i w_i < b \end{cases}$$

¹La funzione $H : \mathbb{R} \rightarrow \mathbb{R}$ è la Funzione di Heaviside:

$$H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

1.1.2 Regressione Lineare

Un altro esempio di neurone è quello che approssima² una [funzione continua](#)

$$f : K \rightarrow \mathbb{R}$$

con $K \subseteq \mathbb{R}^n$ [compatto](#).

L'input del neurone sarà una [n-upla](#) $X = (x_1, \dots, x_n) \in K$, mentre l'output sarà la funzione lineare

$$L(X) = b + \sum_{i=1}^n a_i x_i$$

Per semplicità si considera l'approssimazione vicino allo zero, e si suppone che

$$L(0) = f(0) = 0$$

(a meno di traslazione verticale per $f(0)$).

Si vuole quindi minimizzare

$$C(a_1, \dots, a_n) = \frac{1}{2} \|f - L\|_{L^2}^2 = \frac{1}{2} \int_K \left(\sum_{i=1}^n a_i x_i - f(X) \right)^2 dx_1 \cdots dx_n$$

calcolandone il [gradiente](#)

$$\begin{aligned} \frac{\partial C}{\partial a_k} &= \int_K x_k \left(\sum_{i=1}^n a_i x_i - f(X) \right) dx_1 \cdots dx_n \\ &= \sum_{i=1}^n a_i \int_K x_i x_k dx_1 \cdots dx_n - \int_K x_k f(X) dx_1 \cdots dx_n \end{aligned}$$

e dunque, posti

$$\rho_{ij} := \int_K x_i x_j dx_1 \cdots dx_n, \quad m_k := \int_K x_k f(X) dx_1 \cdots dx_n$$

si ha che

$$\frac{\partial C}{\partial a_k} = \sum_{i=1}^n a_i \rho_{ik} - m_k$$

ovvero, in forma matriciale, posta³ $\rho = (\rho_{ij})$, $\mathbf{a} = (a_1, \dots, a_n)$ e $\mathbf{m} = (m_1, \dots, m_n)$:

$$\nabla C = \rho \mathbf{a} - \mathbf{m}$$

²Approssima in senso L^2 (ovvero minimizza la [norma \$L^2\$](#) della differenza).

³Vedi:

- [Spazio delle matrici](#)
- [Matrice Trasposta](#)
- [Gradiente di una funzione](#)

Dunque, posto che ρ sia [invertibile](#), si ottiene che i valori ottimali per L siano

$$\mathbf{a} = \rho^{-1} \mathbf{m}.$$

Nel caso di funzioni a valori in \mathbb{R}^m il problema si scompone nelle diverse coordinate.

1.2 Rete Neurale

Una rete neurale è [...]

Questa riceve degli input e produce un output, in base a certi parametri \mathbf{w} , per simulare la FUNZIONE TARGET; quest'ultima è l'obiettivo finale della Rete Neurale (ovvero si vuole far sì che l'output della rete neurale sia il più vicino possibile al risultato della funzione target).

Per misurare la distanza tra l'output di una rete e la funzione target si utilizza una funzione errore (o funzione costo), che deve essere scelta in base all'applicazione specifica.

Il processo di apprendimento è quello che, partendo dai parametri \mathbf{w} , li modifica (iterativamente), fino a dei parametri \mathbf{w}^* , che sono ottimali, nel senso che minimizzano la funzione errore. Dunque il processo di apprendimento comporta la minimizzazione della funzione costo.

1.3 Funzioni Costo (Machine Learning)

1.3.1 La Funzione Errore Supremum

Una rete neurale prende input $x \in [0, 1]$ e deve imparare una data funzione continua $\phi : [0, 1] \rightarrow [0, 1]$.

La funzione della rete neurale, dipendete dai parametri \mathbf{w}, b , è $f_{\mathbf{w}, b}(x)$.

La funzione costo Supremum è

$$C(\mathbf{w}, b) := \sup_{x \in [0, 1]} |f_{\mathbf{w}, b}(x) - \phi(x)|.$$

Se la funzione ϕ è conosciuta solo per N valori x_1, \dots, x_N , allora la funzione costo diventa

$$C(\mathbf{w}, b) := \max_{i=1, \dots, N} |f_{\mathbf{w}, b}(x_i) - \phi(x_i)|.$$

1.3.2 La Funzione Errore Norma L2

Una rete neurale prende input $x \in [0, 1]$ e deve imparare una data funzione $\phi : [0, 1] \rightarrow \mathbb{R}$ tale che

$$\int_0^1 (\phi(x))^2 dx < \infty$$

La funzione della rete neurale, dipendete dai parametri \mathbf{w}, b , è $f_{\mathbf{w}, b}(x)$. La funzione costo associata a questo tipo di problema è quella che misura la distanza nella [norma \$L^2\$](#) :

$$C(\mathbf{w}, b) := \int_{[0, 1]} (f_{\mathbf{w}, b}(x) - \phi(x))^2 dx.$$

Se la funzione ϕ è conosciuta soltanto in N punti

$$z_1 = \phi(x_1), \quad z_2 = \phi(x_2), \quad \dots, \quad z_N = \phi(x_N)$$

allora, posti $\mathbf{z} = (z_1, \dots, z_N)$ e $\mathbf{x} = (x_1, \dots, x_N)$, la funzione costo diventa la [distanza](#) in \mathbb{R}^N tra \mathbf{z} e $f_{\mathbf{w}, \mathbf{b}}(\mathbf{x}) := (f_{\mathbf{w}, \mathbf{b}}(x_1), \dots, f_{\mathbf{w}, \mathbf{b}}(x_N))$:

$$C(\mathbf{w}, \mathbf{b}) = \|\mathbf{z} - f_{\mathbf{w}, \mathbf{b}}(\mathbf{x})\|^2 = \sum_{i=1}^N |z_i - f_{\mathbf{w}, \mathbf{b}}(x_i)|^2$$

Interpretazione Geometrica

Fissati \mathbf{x} e \mathbf{z} , la mappa $(\mathbf{w}, \mathbf{b}) \mapsto f_{\mathbf{w}, \mathbf{b}}(\mathbf{x})$ rappresenta una ipersuperficie in \mathbb{R}^N :

$$\Phi(\mathbf{w}, \mathbf{b}) = \begin{pmatrix} \Phi_1(\mathbf{w}, \mathbf{b}) \\ \vdots \\ \Phi_N(\mathbf{w}, \mathbf{b}) \end{pmatrix} = \begin{pmatrix} f_{\mathbf{w}, \mathbf{b}}(x_1) \\ \vdots \\ f_{\mathbf{w}, \mathbf{b}}(x_N) \end{pmatrix}$$

e la funzione costo $C(\mathbf{w}, \mathbf{b})$ è la [distanza](#) euclidea in \mathbb{R}^N di un punto sulla ipersuperficie dal punto \mathbf{z} . Si suppongano appropriate ipotesi di differenziabilità della ipersuperficie.

Il costo è minimizzato in $(\mathbf{w}^*, \mathbf{b}^*)$ quando la distanza è minima, ovvero quando $\Phi(\mathbf{w}^*, \mathbf{b}^*)$ è la proiezione ortogonale di \mathbf{z} sulla ipersuperficie: questo significa che il vettore $\Phi(\mathbf{w}^*, \mathbf{b}^*) - \mathbf{z}$ è ortogonale al [piano tangente](#) alla ipersuperficie in $\Phi(\mathbf{w}^*, \mathbf{b}^*)$: quest'ultimo è generato dai vettori⁴

$$\partial_{w_k} \Phi(\mathbf{w}, \mathbf{b})|_{(\mathbf{w}^*, \mathbf{b}^*)}; \quad \partial_{b_j} \Phi(\mathbf{w}, \mathbf{b})|_{(\mathbf{w}^*, \mathbf{b}^*)}$$

Richiedere l'ortogonalità, quindi, significa richiedere che i prodotti scalari:

$$\begin{aligned} (\partial_{w_k} \Phi(\mathbf{w}, \mathbf{b})|_{(\mathbf{w}^*, \mathbf{b}^*)}) \cdot (\Phi(\mathbf{w}^*, \mathbf{b}^*) - \mathbf{z}) &= 0 \\ (\partial_{b_j} \Phi(\mathbf{w}, \mathbf{b})|_{(\mathbf{w}^*, \mathbf{b}^*)}) \cdot (\Phi(\mathbf{w}^*, \mathbf{b}^*) - \mathbf{z}) &= 0. \end{aligned}$$

Queste sono le equazioni normali, che operativamente diventano

$$\begin{aligned} \sum_{i=1}^N (f_{\mathbf{w}, \mathbf{b}}(x_i) - z_i) \cdot \partial_{w_k} f_{\mathbf{w}, \mathbf{b}}(x_i)|_{(\mathbf{w}, \mathbf{b})=(\mathbf{w}^*, \mathbf{b}^*)} &= 0 \\ \sum_{i=1}^N (f_{\mathbf{w}, \mathbf{b}}(x_i) - z_i) \cdot \partial_{b_j} f_{\mathbf{w}, \mathbf{b}}(x_i)|_{(\mathbf{w}, \mathbf{b})=(\mathbf{w}^*, \mathbf{b}^*)} &= 0 \end{aligned}$$

⁴Vedi: [“Derivata parziale”](#)

Capitolo 2

Cenni di Teoria della Misura

2.1 Funzioni sigmoidali e funzioni discriminatorie

Definizione 2.1.1. Una funzione $\sigma : \mathbb{R} \rightarrow [0, 1]$ si dice sigmoidale se¹

$$\lim_{x \rightarrow -\infty} \sigma(x) = -1, \quad \lim_{x \rightarrow +\infty} \sigma(x) = +1.$$

Definizione 2.1.2. Sia \mathcal{M} la famiglia delle *misure di Baire* per \mathbb{R}^n sul cubo $I^n := [0, 1]^n \subseteq \mathbb{R}$, *finite, con segno e regolari*.

Una funzione $f : \mathbb{R} \rightarrow \mathbb{R}$ si dice discriminatoria per \mathcal{M} se per ogni $\mu \in \mathcal{M}$:

$$\left(\forall \mathbf{w} \in \mathbb{R}^n, \forall \theta \in \mathbb{R} \quad \int_{I^n} f(\mathbf{w} \cdot \mathbf{x}) d\mu(\mathbf{x}) = 0 \right) \implies \mu = 0$$

Proposizione 2.1.3. Ogni funzione *sigmoidale* $\sigma : \mathbb{R} \rightarrow [0, 1]$ è *discriminatoria per \mathcal{M}* , dove \mathcal{M} è l'insieme *misure di Baire* per \mathbb{R}^n sul cubo $I^n := [0, 1]^n \subseteq \mathbb{R}$, *finite, con segno e regolare*.

Ovvero, se $\sigma : \mathbb{R} \rightarrow [0, 1]$ è tale che

$$\lim_{x \rightarrow -\infty} \sigma(t) = 0; \quad \lim_{x \rightarrow +\infty} \sigma(t) = 1$$

allora, per ogni $\mu \in \mathcal{M}$,

$$\left(\forall \mathbf{w} \in \mathbb{R}^n, \forall \theta \in \mathbb{R} \quad \int_{I^n} f(\mathbf{w} \cdot \mathbf{x}) d\mu(\mathbf{x}) = 0 \right) \implies \mu = 0$$

¹Vedi “[Limite \(Analisi Matematica\)](#)”

Parte II

Cordero

Parte III

Sirovich

