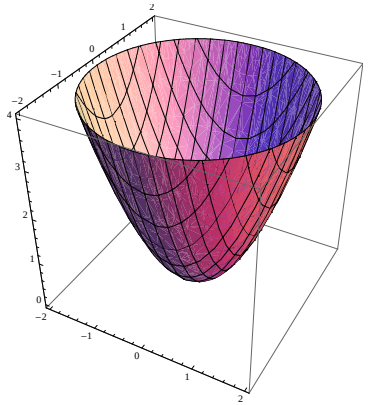
Introduzione alle Reti Neurali Artificiali
Richiami di Matematica



Prof. Crescenzio Gallo

c.gallo@unifg.it

Università degli Studi di Foggia

Dipartimento di Scienze Biomediche

Indice

Prefazione **v**

1 Richiami di Algebra **1**
1.1 Cenni sui vettori . . . 1
1.1.1 Prodotto di un vettore per uno scalare . . . 3
1.1.2 Somma di due vettori . . . 4
1.1.3 Prodotto interno (scalare) . . . 4
1.1.4 Norma . . . 4
1.1.5 Disuguaglianza di Cauchy-Schwartz . . . 6
1.2 Cenni sulle matrici . . . 7
1.2.1 Trasposta di una matrice. . . . 8
1.2.2 Nozioni di algebra lineare . . . 9
1.2.3 Metodo di eliminazione di Gauss (rango della matrice) . . . 21

2 Richiami di Analisi **23**
2.1 La Derivata . . . 23
2.2 L'Integrale . . . 26

Prefazione

Quando si lavora con le reti neurali è utile fare uso di alcuni costrutti matematici di norma utilizzati nella risoluzione di molti problemi in ambito scientifico. Nelle reti neurali è infatti possibile incappare in questioni che, dal punto di vista matematico, possono divenire anche abbastanza complesse.

In configurazioni semplici, dove i neuroni avranno al massimo funzioni di trasferimento binarie, non vi è molta necessità di ricorrere a concetti matematici sofisticati. Diversa è la questione quando utilizziamo algoritmi più complessi con back-propagation o una delle sue varianti. Inoltre, alcune operazioni sui pesi possono divenire molto più veloci quando si fa uso di alcune delle regole offerte dall'algebra lineare per quanto riguarda le operazioni sulle matrici.

Infatti il miglior modo di lavorare con le reti è di definirle come matrici: ciò rende meno dispendioso il processo dal punto di vista computazionale.

Partendo quindi da queste motivazioni, vediamo alcuni concetti di analisi ed algebra lineare che possono risultare molto utili nel lavoro con le reti neurali.

Capitolo 1

Richiami di Algebra

1.1 Cenni sui vettori

Ogni componente di una rete neurale artificiale può in effetti essere intesa come un vettore di elementi. Ad esempio i valori potenziali dei neuroni, i valori dei pesi possono essere considerati come vettori e matrici.

Possiamo descrivere un vettore nel seguente modo:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{1.1}$$

Il seguente è il vettore trasposto di $\mathbf{x}$:

$$\mathbf{x}^T = [x_1, x_2, \ldots, x_n] \tag{1.2}$$

Gli elementi del vettore sono chiamati *componenti*. Il fatto di usare una notazione che si serve di indici consente di denotare in modo molto sintetico certi tipi di operazioni aritmetiche sulle componenti di un vettore. Ad esempio, la somma dei valori di tutte le n componenti di un vettore si può indicare con:

$$\sum_{i=1}^{n} v_i \tag{1.3}$$

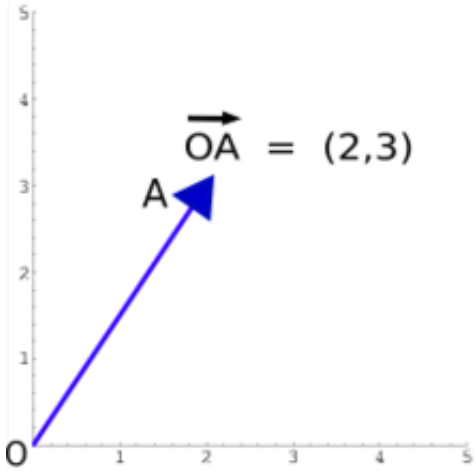
che si legge "sommatoria per i che va da 1 a n di v_i", o più genericamente con:

$$\sum_{i} v_i \tag{1.4}$$

Un'analoga notazione si adopera per i prodotti delle componenti, come in:

$$\prod_{i=1}^{n} v_i \quad (\text{ o più genericamente } \prod_{i} v_i) \tag{1.5}$$

Un vettore può essere rappresentato graficamente come una freccia (segmento orientato) in uno spazio ad n dimensioni, dove n è la dimensione (numero di componenti) del vettore. La "punta" della freccia ha coordinate corrispondenti alle componenti del vettore, mentre la "lunghezza" della freccia a partire dall'origine è la *norma euclidea* $\|\mathbf{x}\|$ del vettore.



1.1.1 Prodotto di un vettore per uno scalare

Un vettore $\mathbf{x}$ può essere moltiplicato, elemento per elemento, per uno scalare λ nel modo seguente:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} ; \lambda\mathbf{x} = \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{bmatrix} \tag{1.6}$$

Graficamente questa operazione può corrispondere ad un "annullamento", "allungamento" o "contrazione" del vettore a seconda del risultato del prodotto dello scalare per le sue componenti.

1.1.2 Somma di due vettori

La stessa cosa vale per la somma di due vettori:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix} \tag{1.7}$$

con la differenza che dal punto di vista geometrico questa corrisponde alla diagonale del parallelogramma con i due lati individuati dai vettori.

1.1.3 Prodotto interno (scalare)

Un caso più interessante è invece il prodotto interno. Questo corrisponde esattamente al calcolo del potenziale del neurone.

$$\langle \mathbf{x}, \mathbf{w} \rangle = \mathbf{x} \cdot \mathbf{w} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = x_1 w_1 + x_2 w_2 + \cdots + x_n w_n = \sum_{i=1}^{n} x_i w_i \tag{1.8}$$

1.1.4 Norma

La norma euclidea di un vettore corrisponde alla sua lunghezza:

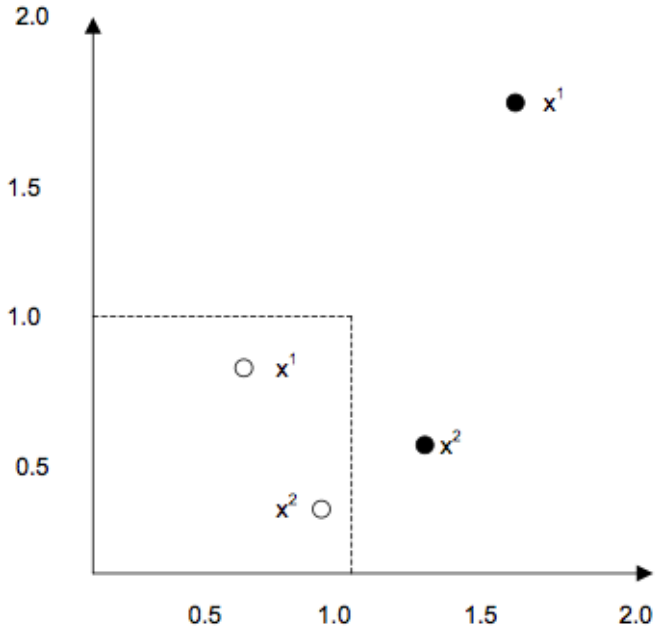
$$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{1.9}$$

La *normalizzazione* di un vettore consiste nel dividere il vettore per la propria norma. Cosa significa? A volte gli input di una rete neurale possono arrivare da dispositivi, che per loro natura, possono essere affetti da oscillazioni.

Si pensi ad esempio a dei sensori esterni che hanno una qualche funzione di trasduzione: questi possono fornire input scorretti con intensità che potrebbero portare la rete neurale a errori. Per questo motivo gli input vengono in certi casi normalizzati. Questo processo avviene dividendo ogni componente per la norma vettoriale:

$$x'_i = \frac{x_i}{\|\mathbf{x}\|}, \ \forall i \in \{1, 2, \ldots, n\} \tag{1.10}$$

Il seguente grafico mostra il processo di normalizzazione su due input, i quali vengono normalizzati per rientrare tra i valori 0 e 1:



Nel campo delle reti, la norma viene utilizzata per calcolare la distanza di due vettori. Ad esempio la distanza tra il vettore degli input e il vettore dei pesi, oppure il vettore di output rispetto quello della risposta desiderata (target). Per questo si ricorre alla norma della differenza:

$$\|a - b\| = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \tag{1.11}$$

Questo calcolo, nel caso di vettori binari (ovvero con componenti 0 e 1) dà in output la distanza di Hamming tra i due vettori, definita come il numero di componenti corrispondenti differenti tra due vettori. Invece, per un solo vettore binario tale distanza può essere intesa come la distanza tra questo e il vettore nullo, un vettore ove tutte le componenti sono poste uguali a 0.

1.1.5 Disuguaglianza di Cauchy-Schwartz

Questa disuguaglianza dice che il valore assoluto del prodotto interno di due vettori $\mathbf{x}, \mathbf{w}$ è minore o uguale al prodotto tra le loro norme:

$$|\mathbf{x} \cdot \mathbf{w}| \leq \|\mathbf{x}\| \cdot \|\mathbf{w}\| \tag{1.12}$$

Quindi esisterà un angolo θ tra i due vettori ($0 \leq \theta \leq \pi$) per il quale:

$$\cos\theta = \frac{\mathbf{x} \cdot \mathbf{w}}{\|\mathbf{x}\| \cdot \|\mathbf{w}\|} \tag{1.13}$$

Il prodotto interno dei due vettori può quindi anche essere definito con:

$$\mathbf{x} \cdot \mathbf{w} = \|\mathbf{x}\| \cdot \|\mathbf{w}\| \cdot \cos\theta \tag{1.14}$$

Questo indica che il prodotto interno dei due vettori sarà proporzionale al coseno dell'angolo che viene a formarsi. Da:

$$\cos 0 = 1,\ \cos \pi/2 = 0,\ \cos \pi = -1 \tag{1.15}$$

si deduce che il prodotto interno, quindi la risposta del neurone, sarà tanto maggiore quanto minore è la distanza (angolare) tra il vettore di input $\mathbf{x}$ e il vettore dei pesi $\mathbf{w}$.

1.2 Cenni sulle matrici

Una matrice reale A è definibile come una raccolta di m righe per n colonne di numeri reali. Ad esempio:

$$A = \begin{bmatrix} 1 & 5 & 6 \\ 4 & 6 & -7 \\ 1 & 3 & -1 \end{bmatrix}$$

Solitamente le sue componenti si indicano con a_{ij} in modo da identificare una componente della matrice mediante due indici, rispettivamente per la riga e la colonna.

Il fatto di utilizzare due indici consente di aumentare il numero di possibili operazioni sugli elementi di una matrice nonché quello dei tipi di scritture sintetiche adottate per designarle. Ad esempio:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \tag{1.16}$$

indica la somma di tutti gli elementi di una matrice $m \times n$, mentre:

$$\sum_{j=1}^{n} a_{ij} \tag{1.17}$$

indica la somma degli elementi della $i-$esima riga della matrice.

Molto spesso scritture simboliche come $\sum_i \sum_j$ vengono ulteriormente abbreviate in $\sum_{ij}$.

1.2.1 Trasposta di una matrice.

Come per i vettori, anche per la matrice è possibile fare la *trasposta*:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}; \quad A^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

È possibile calcolare anche i prodotti $A\,A^T$ e $A^T A$, che naturalmente sono diversi (verificare!).

Questa operazione si applica a qualunque tipo di matrice. Essa dà quindi luogo ad un'altra matrice le cui righe coincidono con le colonne della matrice di partenza e le cui colonne coincidono con le righe della matrice di partenza.

Data una matrice A, la sua trasposta si indica con A^T. Se A è una matrice $m \times n$, A^T sarà una matrice $n \times m$. Facendo a sua volta la trasposta di una matrice trasposta si ottiene la matrice di partenza:

$$(A^T)^T = A \tag{1.18}$$

Quando una matrice A è quadrata l'operazione di trasposizione non altera né la traccia né il determinante della matrice stessa:

$$\mathrm{Tr}(A^T) = \mathrm{Tr}(A); \qquad \det(A^T) = \det(A) \tag{1.19}$$

Ecco un esempio. Data la matrice rettangolare 2×3:

$$A = \begin{bmatrix} 5 & -3 & 2 \\ -2 & 1 & 4 \end{bmatrix}$$

la sua trasposta A^T sarà la matrice 3×2:

$$A = \begin{bmatrix} 5 & -2 \\ -3 & 1 \\ 2 & 4 \end{bmatrix}$$

1.2.2 Nozioni di algebra lineare

L'algebra lineare studia le operazioni con le matrici e le loro proprietà. Alcune di queste operazioni riguardano i vettori, che sono casi particolari di matrici.

Minore. Data una matrice A, la sottomatrice che si ottiene eliminando alcune righe e/o colonne viene chiamata *minore* di A.

Ad esempio, se abbiamo $A = \left[\begin{smallmatrix} 6 & -8 & 0 \\ 2 & 4 & -3 \\ -1 & 5 & -2 \end{smallmatrix}\right]$, eliminando la prima riga e la seconda colonna da A si ottiene il minore $\left[\begin{smallmatrix} 2 & -3 \\ -1 & -2 \end{smallmatrix}\right]$.

Minore complementare. Dato un elemento a_{ij} di una matrice A, il minore che si ottiene eliminando da A la riga i e la colonna j viene chiamato *minore complementare* di a_{ij} (relativamente ad A).

Ad esempio, nella matrice A mostrata sopra, il minore complementare dell'elemento $a_{21} = 2$ si ottiene da A cancellando appunto la seconda riga e la prima colonna. Il risultato è $\left[\begin{smallmatrix} -8 & 0 \\ 5 & -2 \end{smallmatrix}\right]$

Traccia. Si definisce *traccia* di una matrice quadrata A il numero $\mathrm{Tr}(A)$ che si ottiene sommando tutti i valori degli elementi posti sulla *diagonale principale* della matrice, ovvero degli elementi in cui l'indice di riga è uguale all'indice di colonna.

Data la matrice $A = \begin{bmatrix} 2 & -5 & 1 \\ -4 & 3 & 0 \\ -2 & 1 & 4 \end{bmatrix}$ gli elementi posti sulla diagnale principale hanno, rispettivamente, i valori 2, 3 e 4; quindi la traccia è data da $\mathrm{Tr}(A) = 2 + 3 + 4 = 9$.

Determinante di una matrice. Il *determinante* di una matrice quadrata è un numero che, in generale, si ottiene tramite un procedimento iterativo (anzi, ricorsivo) basato su due regole:

1) quella che permette di esprimere il determinante di una matrice di ordine n (cioè $n \times n$) in funzione di quello di matrici di ordine $n - 1$;

2) quella che fissa quanto vale il determinante di una matrice di ordine 1, e cioè il numero stesso.

La regola che consente di esprimere il determinante di una matrice di ordine n in funzione di quelli di matrici di ordine $n - 1$ è basata sul concetto di *complemento algebrico*. Più precisamente, dato un elemento a_{ij} di una matrice di ordine n, posto sulla $i-$esima riga e sulla $j-$esima colonna, il suo complemento algebrico è dato dal determinante del suo minore complementare (che è una matrice quadrata di ordine $n-1$), moltiplicato per il numero $(-1)^{i+j}$.

La regola 1) si può ora esprimere così:

> il determinante di una matrice di ordine n è dato dalla somma dei prodotti degli elementi di una qualsiasi riga o colonna per i rispettivi complementi algebrici (che dipendono da determinanti di matrici di ordine $n - 1$).

Menzioniamo qui brevemente alcuni casi particolari dell'applicazione delle regole 1) e 2). Se si ha una matrice di ordine 2 (qui gli elementi sono genericamente indicati con lettere dell'alfabeto):

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

si vede che queste regole implicano che il suo determinante, denotato con $\det(A)$, è dato da: $\det(A) = ad - bc$. Nel caso di una matrice di ordine 3:

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

si ottiene invece:

$$\det(A) = aei - ahf - bdi + bfg + cdh - ceg$$

Moltiplicazione di una matrice per uno scalare Anche questa operazione si applica a qualunque tipo di matrici. Il risultato della moltiplicazione di una matrice per uno scalare (cioè per un numero) è una nuova matrice i cui elementi sono dati dal risultato della moltiplicazione degli elementi della matrice di partenza per lo scalare in questione.

La nuova matrice si indica semplicemente premettendo il valore dello scalare al nome della matrice di partenza. Così, se la matrice A viene moltiplicata per lo scalare 3, la matrice risultante si indica con $3A$. Questa operazione non altera le dimensioni della matrice di partenza. Se quest'ultima è quadrata, invece, vengono alterati sia la traccia che il determinante.

Più precisamente, moltiplicando una matrice quadrata A di ordine n per lo scalare λ, la traccia viene moltiplicata per λ e il determinante per λ^n. Ecco qui un banale esempio. Sia data la matrice quadrata:

$$A = \begin{bmatrix} 2 & 3 \\ -2 & -1 \end{bmatrix}$$

che ha:

$$\mathrm{Tr}(A) = 1, \quad \det(A) = 4$$

Moltiplichiamo A per lo scalare 3, ottenendo:

$$A = \begin{bmatrix} 6 & 9 \\ -6 & -3 \end{bmatrix}$$

Si ha subito:

$$\mathrm{Tr}(3A) = 3 = 3\mathrm{Tr}(A), \quad \det(3A) = 36 = 3^2 \det(A)$$

Addizione di matrici. L'*addizione* tra due matrici è possibile solo se hanno uguale numero di righe e colonne. Così una matrice 2×3 si può sommare ad un'altra matrice 2×3, ma non ad una matrice — poniamo — 3×5. Quando l'operazione è eseguibile la matrice risultato ha gli stessi numeri di righe e di colonne delle matrici che sono state sommate. Ogni elemento della matrice risultato è fornito dalla somma degli elementi di posto corrispondente in ciascuna delle due matrici che verranno addizionate.

Riportiamo qui un semplice esempio che fa uso di due matrici 3×2:

$$A = \begin{bmatrix} 1 & -4 \\ -6 & 2 \\ -3 & 5 \end{bmatrix} \qquad B = \begin{bmatrix} 3 & 2 \\ 1 & -5 \\ 4 & 7 \end{bmatrix}$$

Si ha che:

$$A + B = \begin{bmatrix} 1+3 & -4+2 \\ -6+1 & 2+(-5) \\ -3+4 & 5+7 \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ -5 & -3 \\ 1 & 12 \end{bmatrix}$$

L'operazione di somma tra matrici gode delle stesse proprietà dell'operazione di somma tra numeri, cioè è commutativa e associativa:

$$A + B = B + A; \quad (A + B) + C = A + (B + C) \tag{1.20}$$

Inoltre l'operazione di moltiplicazione per uno scalare è distributiva rispetto alla somma:

$$\lambda(A + B) = \lambda A + \lambda B \tag{1.21}$$

Per ogni particolare scelta del numero di righe e del numero di colonne esiste una matrice O (la matrice nulla) i cui elementi sono tutti uguali a zero che svolge lo stesso ruolo dello zero nell'addizione tra numeri:

$$A + O = O + A = A \tag{1.22}$$

Moltiplicazione di matrici. È possibile *moltiplicare una matrice per un vettore* purché il vettore abbia lo stesso numero di componenti del numero di colonne della matrice. Ad esempio data la matrice W seguente:

$$W = \begin{bmatrix} 0.6 & 0.5 & -0.1 \\ -1 & 0.3 & 1 \end{bmatrix}$$

ed il vettore X:

$$X = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

è possibile definire un nuovo vettore WX moltiplicando ogni componente della matrice con il componente del vettore corrispondente di colonna. Questo per ogni riga:

$$WX = \begin{bmatrix} 0.6 & 0.5 & -0.1 \\ -1 & 0.3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0.6 \cdot 1 + 0.5 \cdot 0 + (-0.1) \cdot (-1) \\ -1 \cdot 1 + 0.3 \cdot 0 + 1 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 0.5 \\ -2 \end{bmatrix}$$

Se ad esempio la matrice W corrisponde a due vettori di pesi sinaptici, quindi per due neuroni, ed il vettore X al vettore di input, il prodotto tra matrici consente di calcolare in un solo passo tutti i nuovi valori per i neuroni di output. Infatti abbiamo due prodotti interni nel vettore finale WX.

Si osservi che la *moltiplicazione tra due matrici* AB è possibile solo se la matrice A ha numero di colonne uguale alle righe della matrice B: in questo caso, il prodotto finale corrisponde al prodotto del vettore della riga i della prima matrice con il vettore della colonna j della seconda matrice. Ad esempio:

$$A\,B = \begin{bmatrix} 3 & 1 & -1 \\ 2 & 0 & 4 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ 3 & -1 \\ -2 & 2 \end{bmatrix} = \begin{bmatrix} \big(3\cdot 1 + 1\cdot 3 + (-1)\cdot(-2)\big) & + & \big(3\cdot 5 + 1\cdot(-1) + (-1)\cdot 2\big) \\ \big(2\cdot 1 + 0\cdot 3 + 4\cdot(-2)\big) & + & \big(2\cdot 5 + 0\cdot(-1) + 4\cdot 2\big) \end{bmatrix} = \begin{bmatrix} 8 & 12 \\ 6 & 18 \end{bmatrix}$$

Naturalmente la moltiplicazione tra matrici non è commutativa, ovvero $AB \neq BA$, perché il risultato dei due prodotti sarebbe diverso.

Soffermiamo la nostra attenzione su alcuni casi particolari del prodotto tra matrici, riesaminando le operazioni sui vettori visti ora come matrici.

Il primo riguarda il prodotto tra un vettore riga e un vettore colonna. Se, ad esempio, si ha un vettore riga X con 1 riga e m colonne e un vettore colonna Y con m righe e 1 colonna, il prodotto XY è eseguibile e dà origine ad una matrice di una riga e di una colonna (infatti le dimensioni della prima matrice sono $1 \times m$, quelle della seconda $m \times 1$ con una matrice risultante 1×1), ovvero ad un singolo numero. Quest'ultimo viene chiamato anche *prodotto scalare* dei due vettori. Per fare un esempio, i due vettori:

$$X = \begin{bmatrix} 2 & -3 & 4 \end{bmatrix}, \quad Y = \begin{bmatrix} -6 \\ 5 \\ -7 \end{bmatrix}$$

danno origine al prodotto scalare:

$$XY = 2\cdot(-6) + (-3)\cdot 5 + 4\cdot(-7) = -55$$

È da notare che anche il prodotto YX è possibile (infatti le dimensioni in gioco sono $3 \times 1 \leftrightarrow 1 \times 3$). Tuttavia, esso dà origine ad una matrice quadrata che viene talvolta chiamata *prodotto esterno* dei due vettori. Nel caso dell'esempio precedente, tale matrice ha dimensioni 3×3 ed è data da:

$$YX = \begin{bmatrix} -12 & 18 & -24 \\ 10 & -15 & 20 \\ -14 & 21 & -28 \end{bmatrix}$$

Se il prodotto scalare di due vettori è nullo, si usa dire che i due vettori sono reciprocamente *ortogonali*. Per un dato vettore colonna V la sua trasposta è costituita da un vettore riga V^T. La radice quadrata del prodotto scalare $V^T V$ definisce un numero, chiamato *norma* o *lunghezza* del vettore V. Così, dato il vettore colonna:

$$V = \begin{bmatrix} 4 \\ -2 \\ 1 \end{bmatrix}$$

la corrispondente trasposta è il vettore riga:

$$V^T = \begin{bmatrix} 4 & -2 & 1 \end{bmatrix}$$

e la norma di V è data da:

$$\sqrt{V^T V} = \sqrt{(4)^2 + (-2)^2 + (1)^2} = \sqrt{21} = 4.58\ldots$$

In modo analogo, dato un vettore riga W, la sua norma sarà data dalla radice quadrata del prodotto scalare WW^T. Notare che in questo caso abbiamo dovuto moltiplicare a destra per la trasposta, anziché a sinistra, come nel caso precedente.

Un altro caso particolare molto importante del prodotto tra matrici è dato dal prodotto di una matrice quadrata A di dimensioni $m \times m$ (che funge da primo fattore) per un vettore colonna V avente m righe e 1 colonna. In base alle

definizioni date in precedenza, questo prodotto dà luogo ad un nuovo vettore W avente ancora m righe e 1 colonna, cioè le stesse dimensioni di V. Questa circostanza permette di interpretare la matrice A come un operatore che, agendo tramite la regola del prodotto sul vettore V, produce la trasformazione di V in un nuovo vettore W.

Come è facile intuire, una circostanza analoga vale per i vettori riga, salvo lo scambio della posizione del prodotto. Infatti moltiplicando il vettore riga W, avente una riga e m colonne, per la matrice A di dimensioni m (stavolta W funge da primo fattore) si ottiene ancora un vettore riga U di 1 riga e m colonne.

Matrice unitaria. La *matrice unitaria* I è una matrice quadrata in cui ciascun componente è dato dal cosiddetto "delta di Kronecker"

$$\delta(ij) = \begin{cases} 0 & \text{se } i \neq j \\ 1 & \text{se } i = j \end{cases} \tag{1.23}$$

Ad esempio la matrice unitaria di ordine 3 è la seguente:

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Il prodotto tra una matrice A e la matrice I corrispondente restituisce sempre la matrice originaria:

$$A\,I = I\,A = A \tag{1.24}$$

Inversa di una matrice. Osserviamo che l'esistenza della matrice unitaria I consente di definire un'operazione unaria chiamata *inversione*. Più precisamente, la matrice inversa di una matrice quadrata A, indicata col simbolo A^{-1}, deve soddisfare le condizioni:

$$AA^{-1} = A^{-1}A = I \tag{1.25}$$

L'inversione di una matrice A è possibile se e solo se il determinante di A è diverso da zero. Il calcolo diretto della matrice inversa è assai oneroso dal punto di vista computazionale. Esistono molti metodi (di cui non ci occuperemo in questa sede) per effettuarlo, in modo esatto o approssimato, tramite un computer.

Autovalori di una matrice quadrata. Abbiamo visto che, moltiplicando una matrice quadrata A di dimensioni $m \times m$ per un vettore colonna V di m elementi, si ottiene un nuovo vettore colonna W, sempre avente m righe. Tuttavia, in generale, W è completamente diverso da V.

Ci si può ora chiedere se esistono dei vettori V (le cui componenti non siano tutte nulle) fatti in modo tale che, moltiplicando A per V, si ottenga un nuovo vettore le cui componenti siano tutte direttamente proporzionali a quelle di V tramite lo stesso fattore di proporzionalità. In sostanza, il nuovo vettore dovrebbe essere pari a V moltiplicato per un opportuno scalare. Se esistono vettori V del genere, essi debbono quindi soddisfare la condizione:

$$AV = \lambda V \tag{1.26}$$

dove λ indica un opportuno scalare. Si capisce subito che sia A che V, ammesso che esistano, dipendono in modo critico da come è fatta la matrice A. Per questa ragione, quando la condizione sopra scritta è soddisfatta, si usa dire che λ è un *autovalore* di A associato all'*autovettore* V.

I concetti di autovalore e di autovettore hanno una grande importanza in molti domini della scienza, compresa la Statistica e le Reti Neurali. Perciò qui accenneremo brevemente ad alcuni aspetti della teoria ad essi relativa, che costituisce uno dei capitoli fondamentali dell'algebra lineare.

La prima circostanza da tener presente è che si può dimostrare che la definizione data sopra di autovalore e di autovettore implica che gli autovalori λ di una matrice A di dimensioni $m \times m$ debbano soddisfare la condizione:

$$\det(A - \lambda I) = 0 \tag{1.27}$$

dove I è la matrice unitaria. Questa condizione può essere vista come un'equazione algebrica in λ, le cui soluzioni forniscono gli autovalori cercati di A. Si può facilmente mostrare che questa equazione è di grado m (pari al numero

delle righe e delle colonne di A) e quindi ammette m soluzioni. Questo permette intanto di concludere che una matrice $m \times m$ ammette m autovalori.

Per quanto riguarda però la natura di queste soluzioni, la casistica può essere molto varia: queste possono essere numeri reali oppure numeri complessi, possono o no avere coppie di valori coincidenti, il tutto in funzione di come è concretamente fatta la matrice A.

Vi è inoltre da considerare che, quando il valore di m è molto grande, date le difficoltà pratiche di calcolo del determinante di cui abbiamo detto in precedenza, la ricerca degli autovalori deve necessariamente essere basata su metodi numerici approssimati. Nel caso particolare, invece, in cui $m = 2$ la condizione precedente si può scrivere sotto forma di una semplice equazione di secondo grado in λ data da:

$$\lambda^2 - \mathrm{Tr}(A) \cdot \lambda + \det(A) = 0 \tag{1.28}$$

Questo permette un calcolo diretto degli autovalori.

Si abbia, ad esempio, la matrice A di ordine 2 definita da:

$$A = \begin{bmatrix} 6 & -3 \\ 1 & 2 \end{bmatrix}$$

per la quale si ha:

$$\mathrm{Tr}(A) = 6 + 2 = 8, \quad \det(A) = 6 \cdot 2 - (-3) \cdot 1 = 15$$

Dunque, l'equazione algebrica per gli autovalori è:

$$\lambda^2 - 8\lambda + 15 = 0$$

Essa ammette le due soluzioni reali e distinte $\lambda_1 = 3$, $\lambda_2 = 5$, che rappresentano gli autovalori cercati di A. Una volta noti gli autovalori di una matrice, in corrispondenza ad ognuno di essi gli autovettori associati si ottengono

sfruttando la definizione fondamentale (1.26), che può essere vista come un sistema di equazioni algebriche di primo grado nelle incognite costituite dalle componenti dell'autovettore V.

Si può dimostrare che tale sistema è indeterminato, cioè ammette infinite soluzioni e questo permette di concludere che ad ogni autovalore sono associati infiniti autovettori. Questo implica che, per sceglierne uno, occorre imporre ulteriori condizioni aggiuntive che dipendono dagli scopi che si hanno nel contesto in cui si applicano questi concetti. Tra queste condizioni una delle più usate consiste nell'imporre che gli autovettori siano normalizzati, ovvero che siano dei vettori la cui norma è pari a 1.

Menzioniamo ora due ulteriori importanti teoremi riguardanti gli autovalori delle matrici quadrate simmetriche, cioè delle matrici in cui l'elemento posto sulla $i-$esima riga e sulla $j-$esima colonna è uguale all'elemento posto sulla $j-$esima riga e sulla $i-$esima colonna. Molte delle matrici utilizzate in Statistica rientrano in questa categoria. Gli enunciati dei teoremi in questione sono:

i) Gli autovalori di una matrice simmetrica, i cui elementi sono tutti numeri reali, sono anch'essi tutti costituiti da numeri reali;

ii) Due autovettori di una matrice simmetrica associati a due autovalori distinti sono reciprocamente ortogonali.

Matrici definite positive e non negative. Prima di terminare questo paragrafo introduciamo un ulteriore concetto relativo alle matrici quadrate: quello di *matrice definita positiva*. Più precisamente, data una matrice quadrata A di dimensioni $m \times m$, se, per qualsiasi scelta del vettore colonna V con m righe (con elementi non tutti nulli), si ha che il numero q definito da:

$$q = V^T A V \tag{1.29}$$

è *sempre positivo*, allora A è una matrice definita positiva.

Se invece il numero q può essere positivo alcune volte e uguale a zero altre volte, ma non può mai essere negativo, si dice che la matrice A è *definita non negativa*. Facciamo un esempio, basato sulla matrice 2×2 definita da:

$$A = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix} \tag{1.30}$$

dove il simbolo a denota un numero il cui valore assoluto può essere compreso tra 1 e -1 (estremi esclusi). Se scegliamo un vettore colonna V assolutamente generico definito da:

$$V = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \qquad V^T = \begin{bmatrix} v_1 & v_2 \end{bmatrix} \tag{1.31}$$

alcuni calcoli mostrano che:

$$q = V^T A V = v_1^2 + v_2^2 + 2a v_1 v_2 \tag{1.32}$$

Si vede subito che, se si avesse $a = 1$, q si potrebbe scrivere come:

$$q = (v_1 + v_2)^2 \tag{1.33}$$

e quindi, se v_1 e v_2 non sono entrambi nulli, sarebbe una quantità positiva. D'altra parte, se fosse $a = -1$, si avrebbe:

$$q = (v_1 - v_2)^2 \tag{1.34}$$

che è una quantità che è positiva o nulla (nel caso in cui $v_1 = v_2$). Se a ha, come nel nostro caso, valori intermedi tra 1 e -1, il valore di q è compreso tra $(v_1 - v_2)^2$ e $(v_1 + v_2)^2$ e quindi, in base alle considerazioni precedenti, è necessariamente una quantità positiva. Si deve perciò concludere che la matrice A è definita positiva.

Le matrici definite positive sono caratterizzate da un importante teorema, che ci limitiamo ad enunciare:

iii) Gli autovalori di una matrice definita positiva sono tutti numeri positivi.

Per le matrici definite non negative vale un teorema analogo:

iv) Gli autovalori di una matrice definita non negativa sono tutti numeri positivi o nulli.

1.2.3 Metodo di eliminazione di Gauss (rango della matrice)

Questo procedimento possiede un vasto numero di applicazioni, tra cui la soluzione di un sistema di equazioni lineari omogenee ad n incognite disposte in forma matriciale.

Il metodo di eliminazione di Gauss permette di ottenere una matrice (triangolare superiore) equivalente che possiede un numero crescente di zero iniziali sulle righe ed 1 sulla diagonale principale:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & * & \cdots & * \\ 0 & 1 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{1.35}$$

dove il simbolo $*$ rappresenta qualsiasi valore risultante dalla trasformazione.

Vi sono solo tre operazioni che è possibile fare sulla matrice (in tal modo il sottostante sistema di equazioni lineari non cambia, con le relative soluzioni):

1. Moltiplicazione di una riga per uno scalare diverso da 0, ad esempio: $R_i \Rightarrow \lambda R_i$
2. Scambio di due righe: $R_i \Leftrightarrow R_j$
3. Somma (algebrica) del multiplo di una riga a un'altra: $R_i \Rightarrow R_i + \lambda R_j$

Operando in serie in questo modo e ripetendo da capo il procedimento si arriva ad ottenere una matrice che non può essere più ridotta. A questo punto il numero delle righe che non hanno componenti uguali a zero viene detto *rango* della matrice (corrisponde al numero di soluzioni indipendenti del sottostante sistema).

Ad esempio:

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 2 & 0 \\ 0 & 6 & -4 \end{bmatrix} \cdot(-1) \Rightarrow \begin{bmatrix} \mathbf{-1} & \mathbf{1} & \mathbf{0} \\ 1 & 2 & 0 \\ 0 & 6 & -4 \end{bmatrix} \begin{matrix} R_1+ \\ R_2 \end{matrix} \Rightarrow \begin{bmatrix} -1 & 1 & 0 \\ \mathbf{0} & \mathbf{3} & \mathbf{0} \\ 0 & 6 & -4 \end{bmatrix} \cdot(-2) \Rightarrow \begin{bmatrix} -1 & 1 & 0 \\ \mathbf{0} & \mathbf{-6} & \mathbf{0} \\ 0 & 6 & -4 \end{bmatrix} \begin{matrix} \\ R_2+ \\ R_3 \end{matrix}$$

$$\Rightarrow \begin{bmatrix} -1 & 1 & 0 \\ 0 & -6 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{-4} \end{bmatrix} \begin{matrix} \cdot(-1) \\ \cdot(-1/6) \\ \cdot(-1/4) \end{matrix} \Rightarrow \begin{bmatrix} \mathbf{1} & \mathbf{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}$$

Capitolo 2

Richiami di Analisi

2.1 La Derivata

Solitamente quando si lavora con le reti neurali si può arrivare ad utilizzare concetti di analisi matematica abbastanza frequentemente. La derivata in primis è molto utilizzata, dato che è necessaria per l'algoritmo di apprendimento supervisionato *backpropagation*.

In questa sezione riassumeremo brevemente il concetto di derivata e come si calcola.

Dal punto di vista matematico la derivata si definisce come limite del rapporto incrementale fra la variazione dei valori della funzione e quella della variabile indipendente, al tendere a zero di quest'ultima.

Supponiamo di voler descrivere il movimento di un'auto che si muove lungo una guida rettilinea; sia $p(t)$ la posizione (rispetto ad un punto fissato) dell'auto al tempo t. Il tachimetro segna, al tempo iniziale t_0, una velocità $v(t_0)$. La velocità media nell'intervallo $[t_0, t_1]$ è data dal rapporto fra lo spazio percorso ed il tempo impiegato a percorrerlo,

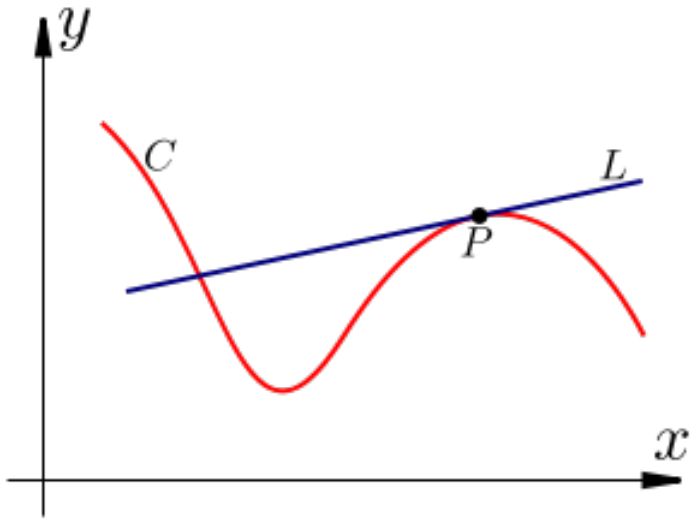
quindi:

$$v[t_0, t_1] = \frac{p(t_1) - p(t_0)}{t_1 - t_0} \tag{2.1}$$

Quindi la velocità al tempo t_0 segnata sul tachimetro è il limite della velocità media quando l'ampiezza dell'intervallo tende a zero:

$$v(t_0) = \lim_{t_1 \to t_0} \frac{p(t_1) - p(t_0)}{t_1 - t_0} \tag{2.2}$$

Graficamente la derivata $f'(x_0)$ nel punto x_0 di una funzione f esprime il coefficiente angolare (pendenza) della retta tangente nel punto x_0 al grafico della funzione. Nella figura seguente, la retta L tangente in P al grafico C della funzione ha pendenza data dalla derivata della funzione in P:



Il tutto si può sintetizzare come segue. Sia $f : (a, b) \to \mathbb{R}$ una funzione reale, ed $x_0 \in (a, b)$. Consideriamo una piccola variazione $x_1 = x_0 + h$ e definiamo il rapporto incrementale della funzione nel punto x_0 con incremento h

come:

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0 + h) - f(x_0)}{h} \tag{2.3}$$

Il rapporto incrementale rappresenta la velocità media dell'esempio precedente, oppure un tasso medio di crescita, e così via. Ci interessa studiare il limite per h che tende a zero:

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{2.4}$$

o, meglio, i due limiti per h che tende a zero da "sinistra" (cioè per valori negativi crescenti verso lo zero):

$$\lim_{h \to 0^-} \frac{f(x_0 + h) - f(x_0)}{h} \tag{2.5}$$

e da "destra" (cioè per valori positivi decrescenti verso lo zero):

$$\lim_{h \to 0^+} \frac{f(x_0 + h) - f(x_0)}{h} \tag{2.6}$$

Se i due limiti esistono e coincidono, la funzione f è derivabile in x_0 con:

$$f'(x_0) = \lim_{h \to 0^-} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{h \to 0^+} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{2.7}$$

La derivata consente di trovare nella funzione eventuali minimi. Nell'algoritmo Error back-propagation (EBP) si ha bisogno di calcolare la derivata della funzione di trasferimento utilizzata, dato che viene impiegata per calcolare l'errore di ciascun nodo dello strato di output e degli strati nascosti.

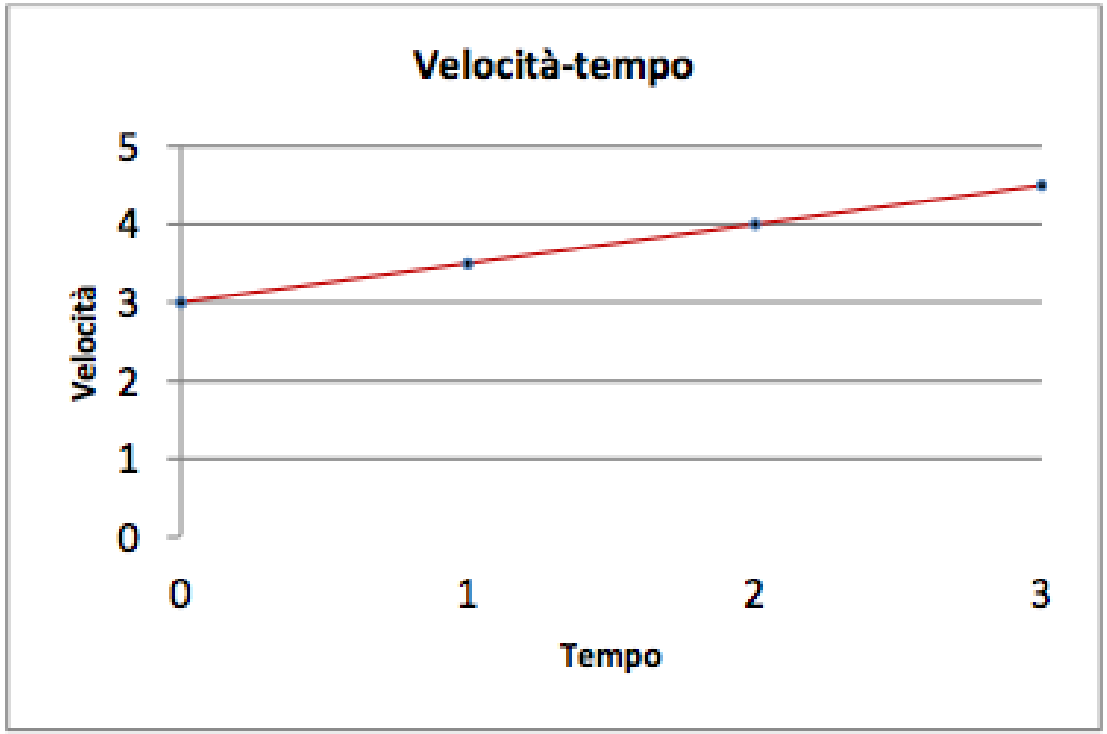
È anche importante il concetto di **gradiente** come estensione del concetto di derivata. Il gradiente di una funzione $f(x)$ vettoriale (ovvero di più variabili indipendenti) può essere inteso come un vettore composto dalle derivate dei suoi componenti. Lo si incontra nell'algoritmo EBP.

2.2 L'Integrale

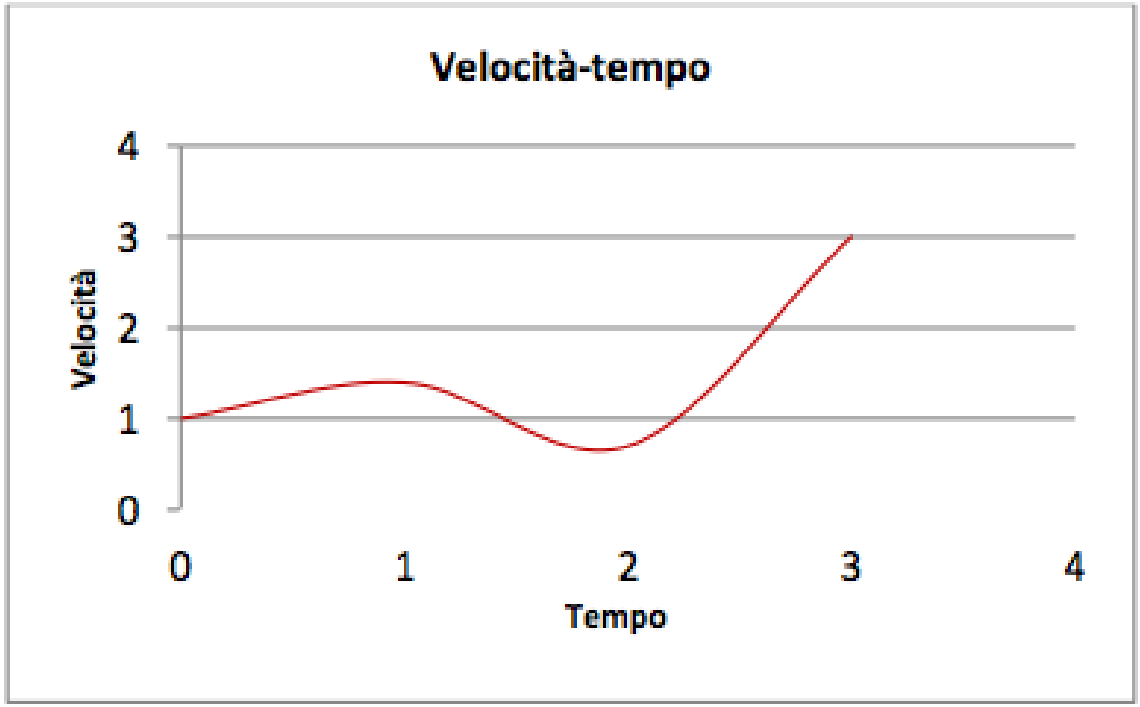
A parte il suo utilizzo nelle reti neurali, la definizione di integrale è bene conoscerla, dato il suo impiego nella maggior parte dei campi scientifici (si pensi ad esempio all'analisi di Fourier).

Il concetto di integrale è piuttosto semplice in sé.

Si consideri la relazione tra velocità e tempo. Se il moto è anche uniformemente accelerato, avremo un grafico come il seguente:

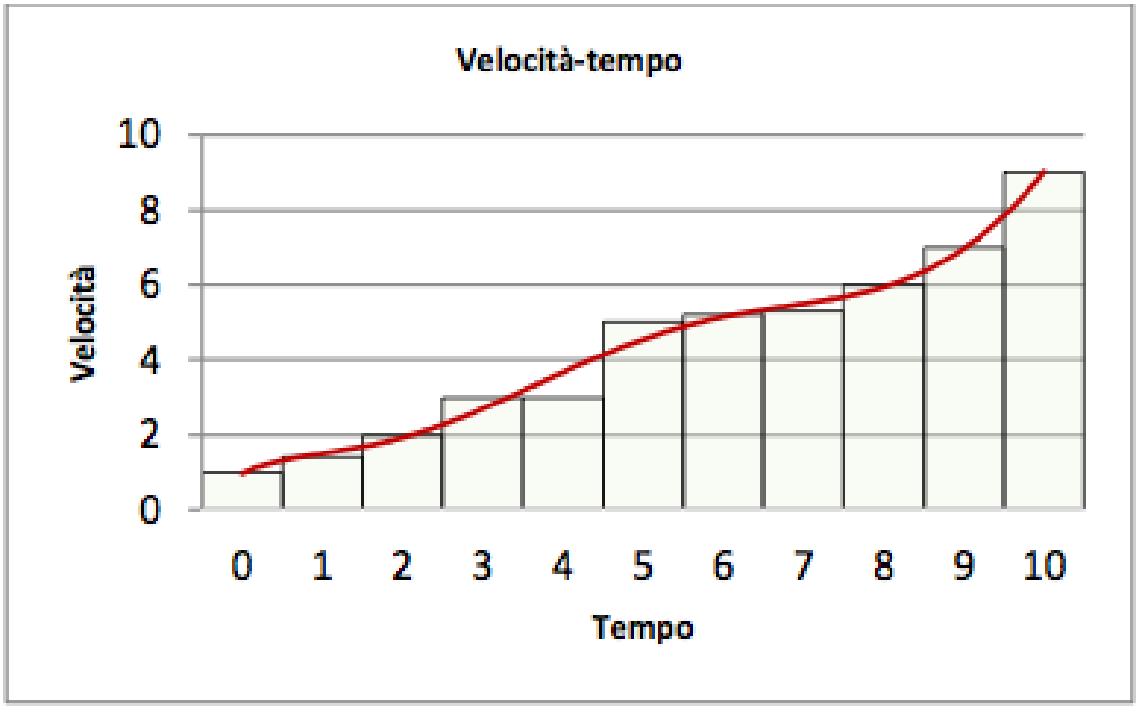


Trovare lo spazio percorso non è complicato dato che basta trovare l'area del trapezio sottostante. Ma come possiamo fare nel caso avessimo un moto non uniforme?



In questo caso il calcolo dell'area tra la funzione e l'asse x non può essere fatto con una semplice moltiplicazione. Si può pensare allora di costruire nel grafico una serie di rettangoli aventi per base un segmento ricavato sull'asse delle x e come valore massimo un valore del tratto di funzione y rispettivo. In questo modo tentiamo di avvicinarci all'area reale facendo la somma di tutte le aree dei rettangoli.

Graficamente:



Naturalmente non riusciamo a riprodurre esattamente l'area. Possiamo solo avere una delle due situazioni:

$$\sum_{k=1}^{n} f(x_k) \cdot (x_k - x_{k-1}) \leq \text{Area trapezoide} \tag{2.8}$$

oppure

$$\sum_{k=1}^{n} F(x_k) \cdot (x_k - x_{k-1}) \geq \text{Area trapezoide} \tag{2.9}$$

dove $(x_k - x_{k-1})$ definisce un segmento sull'asse delle x, ovvero la base del rettangolo, $f(x_k)$ identifica il valore minore della funzione nell'intervallo ed $F(x_k)$ identifica invece il valore più grande.

Il "trucco" per trovare l'area giusta sta nello scegliere rettangoli dalla base molto piccola. Se facciamo tendere il numero di rettangoli ad infinito otteniamo l'area corretta. Quindi:

$$\lim_{n\to\infty}\sum_{k=1}^{n}F(x_k)\cdot(x_k-x_{k-1})=\int_a^b f(x)\mathrm{d}x=\lim_{n\to\infty}\sum_{k=1}^{n}f(x_k)\cdot(x_k-x_{k-1}) \tag{2.10}$$

Detto questo, è facile definire la funzione integrale e quindi l'integrale indefinito.

L'integrale è detto *definito* quando è considerato entro un dato intervallo; di conseguenza, se rendiamo variabile uno degli estremi, abbiamo un funzione (la funzione integrale) che esprime l'integrale *indefinito*.

Al variare dell'estremo, il valore integrale si avvicina o si allontana dall'ipotetica area vista prima:

$$F(x)=\int_a^x f(t)\mathrm{d}t \tag{2.11}$$

Questa è una funzione crescente, dato che x è un punto sull'asse compreso fra a e b.

Il teorema fondamentale del calcolo integrale dice che la derivata della funzione integrale è uguale alla funzione di partenza. Indicando la derivata rispetto ad x con D avremo:

$$F'(x)=\mathrm{D}\left[\int_a^x f(t)\mathrm{d}t\right]=f(x) \tag{2.12}$$

Indice analitico

addizione tra due matrici, 12
autovalore, 17
autovettore, 17

complemento algebrico, 10

derivata, 23
determinante, 10
distanza di due vettori, 6
disuguaglianza di Cauchy-Schwartz, 6

funzione integrale, 29

gradiente, 25

integrale, 26

matrice, 7
matrice definita non negativa, 19
matrice definita positiva, 19
matrice inversa, 16
matrice unitaria, 16
metodo di eliminazione di Gauss, 21
minore, 9
minore complementare, 9
moltiplicare una matrice per un vettore, 13
Moltiplicazione di una matrice per uno scalare, 11
moltiplicazione tra due matrici, 14

norma (euclidea), 4
normalizzazione, 5

prodotto esterno di due vettori, 14
prodotto interno, 4

scalare, 3
somma di due vettori, 4

traccia, 9
trasposta di una matrice, 8

vettore, 1
vettori ortogonali, 15