STATISTICA E ANALISI DEI DATI

Capitolo 6 – AI, Funzione distanza, similarità e metriche

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2024-2025

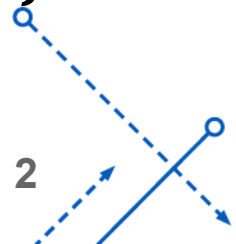
INTELLIGENZA ARTIFICIALE

Cosa si intende con «**intelligenza**»?

- Il termine «intelligente» deriva dal Latino «intelliggere» ovvero "intendere", "concepire", "comprendere" con l'Intelletto.

- L' intelligenza **è la capacità di cogliere le relazioni tra le cose**, di leggere in profondità **e di comprendere il significato di ciò che ci circonda**.

Un problema che si verifica spesso quando si parla di intelligenza artificiale è che si confondono o si utilizzano come sinonimi i termini di: ***SEZIENTE, INTELLIGENTE, COSCIENTE***.



INTELLIGENZA ARTIFICIALE

- **Non esiste una definizione unica di intelligenza** (anche perché sembra che non possa esistere un un'unica forma di intelligenza).
  - Intelligenza **fluida**: La capacità di ragionare e di risolvere problemi in modo nuovo.
  - Intelligenza **cristallizzata**: Il bagaglio di conoscenze e abilità acquisite nel corso dell'esperienza.
  - Intelligenza **linguistica**: La capacità di comprendere e utilizzare il linguaggio.
  - Intelligenza **logico-matematica:** La capacità di ragionare in modo logico e di risolvere problemi matematici.
  - Intelligenza **visuo-spaziale**: La capacità di percepire e di manipolare informazioni visive e spaziali.

INTELLIGENZA ARTIFICIALE

- Inoltre, noi definiamo l'intelligenza in base alla «nostra conoscenza dell'intelligenza», pertanto non siamo in grado di conoscere altri tipi di intelligenza se prima non li incontriamo!

  (Ecco perché potrebbe essere difficile trovare intelligenza Aliena?)

mbra che non possa esistere un

n modo nuovo.

site nel corso dell'esperienza.

nguaggio.

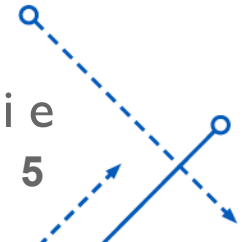
logico e di risolvere problemi

e informazioni visive e spaziali.

INTELLIGENZA ARTIFICIALE

Possiamo **distinguere i tratti comuni** a più tipi di intelligenza!

- Capacità di apprendimento: L'intelligenza implica la capacità di acquisire nuove conoscenze e abilità, di adattarsi a nuove situazioni e di risolvere problemi.
- Capacità di ragionamento: L'intelligenza permette di riflettere, analizzare informazioni, trarre conclusioni e fare inferenze logiche.
- Capacità di astrazione: L'intelligenza consente di cogliere concetti generali e di applicare principi a situazioni diverse.
- Capacità di problem solving.
- Capacità di pensiero critico: L'intelligenza implica la capacità di valutare criticamente informazioni e argomenti, di individuare bias e di formulare giudizi ponderati.



SEZIENTE, COSCIENTE, INTELLIGENTE

Cosa si intende con «**seziente**»?

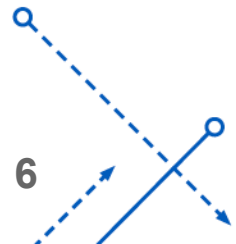
La differenza principale tra "seziente" e "intelligente" è che:

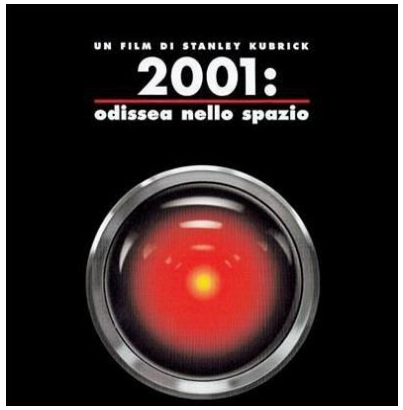
- la senzientia si riferisce alla capacità di provare sensazioni e percezioni;
- l 'intelligenza implica la capacità di ragionare, comprendere e risolvere problemi.

E cosa si intende con «**cosciente**»?

- La coscienza di sé è la capacità di essere consapevoli **di se stessi come individuo separato dal mondo esterno.** Essa implica la consapevolezza dei propri pensieri, emozioni, sensazioni, ricordi e desideri, nonché la comprensione del proprio ruolo nel mondo e delle proprie relazioni con gli altri.

L'intelligenza artificiale riesce ad emulare la senzientia (sensori, emotion detection, face recognition…)
Inoltre, l'intelligenza artificiale riesce ad emulare l'intelligenza ☺ (le auto guidano da sole!)
Ciò che manca alla AI al momento (non sappiamo se la abbia) è la «coscienza di se»




UN FILM DI STANLEY KUBRICK
2001:
odissea nello spazio


TERMINATOR 3
RISE OF THE MACHINES
OFFICIAL MAGAZINE
TERMINATOR 3
RISE OF THE MACHINES
OFFICIAL MAGAZINE


NETFLIX
T A U


SHARLTO COPLEY
DEV PATEL
con SIGOURNEY WEAVER
HUGH JACKMAN
HUMANDROiD

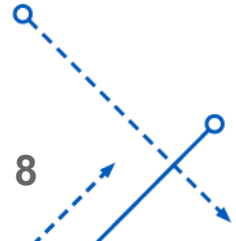

28 DAYS LATER
EX_MACHINA
THERE IS NOTHING MORE HUMAN
THAN THE WILL TO SURVIVE
COMING SOON


hulu
MOTHER/
ANDROID

Distanza e Similarità /1

Per risolvere il problema di clustering:

- Ci servono **strumenti** per capire **«chi si appaia con chi»** e in quale cluster;
- Dobbiamo **definire i termini somiglianza o differenza** in modo quantitativo;
- Occorre precisare cosa significa la **somiglianza di due individui Ii e Ij assegnati allo stesso cluster**;
- Occorre precisare cosa significa la **differenza di due individui assegnati a differenti cluster**.



Distanza e Similarità /2

Dati due individui I_i e I_j $(i \neq j)$.

Cosa significa **somiglianza** di due individui $\mathbf{I_i}$ e $\mathbf{I_j}$ assegnati allo stesso cluster?

$coefficiente\ di\ similarità\ s_{ij} = s(X_i, X_j)$ ⟵ **E' quanto più vicino al massimo possibile.** Spostare $\mathbf{X_i}$ o $\mathbf{X_j}$ in un altro cluster ridurrebbe questo valore!

Cosa significa **differenza** di due individui assegnati a differenti cluster?

$d_{ij} = d(X_i, X_j)$ ⟵ **E' massima!** Spostare $\mathbf{X_i}$ o $\mathbf{X_j}$ in un altro cluster ridurrebbe questo valore e sarebbe possibile trovare un altro cluster nel quale questa distanza sia più grande.

Proprietà importanti di s_{ij} e d_{ij}

$$s_{ij} = s(X_i, X_j)$$

Il **coefficiente** di **similarità** assume valori **nell'intervallo** [0,1]

$$d_{ij} = d(X_i, X_j)$$

Le misure di distanza possono assumere **qualsiasi valore reale maggiore o uguale a zero**.

Importante: è possibile clusterizzare usando il coefficiente di similarità oppure la funzione di distanza. Entrambi sono indicatori che ci dicono «chi è simile a chi» e «chi dovrebbe appaiarsi con chi».

Clustering “naïve”

Un criterio per risolvere il problema di clustering potrebbe essere quello di assegnare due individui **Ii** e **Ij** ($i \neq j$):

- allo stesso cluster se il coefficiente di similarità tra i punti $\mathbf{X}_i$ e $\mathbf{X}_j$ è prossimo ad 1 **oppure** se la distanza tra i punti $\mathbf{X}_i$ e $\mathbf{X}_j$ è sufficientemente piccolo;

- a differenti cluster se il coefficiente di similarità tra i punti è prossimo ad 0 **oppure** se la distanza tra i punti è sufficientemente grande.

Ricordate i melanomi?

Come calcoliamo s_{ij} d_{ij}

Funzione di distanza /1

Le misure metriche di somiglianza sono soprattutto basate sulle funzioni distanza tra i vettori delle caratteristiche. Occorre quindi definire tale funzione.

Una funzione a valori reali d(Xi, Xj) `e detta funzione distanza **se e soltanto se essa soddisfa** le seguenti condizioni:

$$(i)\ d(X_i, X_j) = 0\ se\ e\ solo\ se\ X_i = X_j,\ con\ X_i\ e\ X_j\ in\ E_p;$$

La distanza tra un elemento e se stesso è zero.

$$(ii)\ d(X_i, X_j) \geq 0\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$$

La distanza è una funzione NON negativa

$$(iii)\ d(X_i, X_j) = d(X_j, X_i)\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$$

La distanza tra Xi e Xj è simmetrica (e la stessa tra Xj e Xi)

$$(iv)\ d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)\ per\ ogni\ X_i,\ X_j\ e\ X_k\ in\ E_p.$$

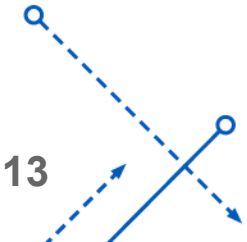
Funzione di distanza /2

$$(iv)\ d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)\ \ per\ ogni\ X_i,\ X_j\ \ e\ X_k\ \ in\ E_p.$$

Diseguaglianza triangolare

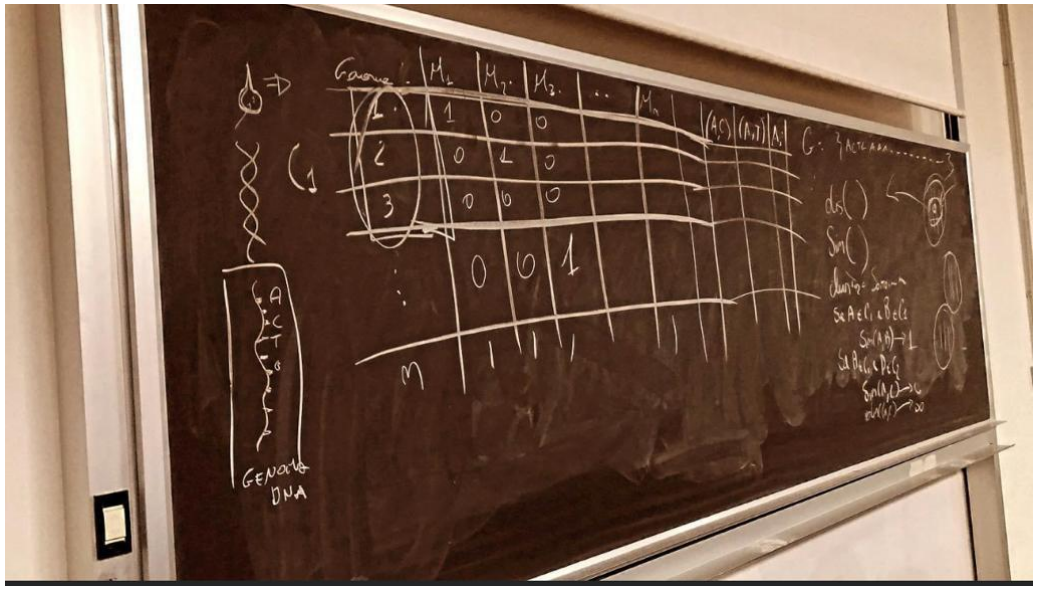
La distanza tra Xi e Xj deve essere sempre minore o uguale della somma delle distanze di ognuno dei due vettori considerati da qualunque altro terzo vettore Xk.

In parole povere: se ci sta «un vettore in mezzo» tra Xi e Xj, allora quel vettore deve essere o nullo (0) oppure darà un contributo e aumenterà la distanza tra Xi e Xj



Distance Matrix (matrice delle distanze)

Negli esempi visti alla lavagna abbiamo avuto modo di apprezzare la forma di una **matrice di similarità.**



$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix},$$

La matrice delle distanze contiene i valori di dij per le varie coppie anziché sij

Chiarimento!

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix},$$

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix},$$

	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}
mean	$\overline{x}_1$	$\overline{x}_2$	$\dots$	$\overline{x}_j$	$\dots$	$\overline{x}_p$
var	s_1^2	s_2^2	$\dots$	s_j^2	$\dots$	s_p^2

Distance Matrix (matrice delle distanze) /1

Le **distanze** tra tutte le possibili coppie di unità sono inserite in **una matrice simmetrica D** di cardinalità n × n, ossia

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix},$$

$$d_{ij} = d(X_i, X_j)$$

$(i)\ d(X_i, X_j) = 0\ se\ e\ solo\ se\ X_i = X_j,\ con\ X_i\ e\ X_j\ in\ E_p;$

I termini sulla diagonale principale sono tutti uguali a zero mentre

$(iii)\ d(X_i, X_j) = d(X_j, X_i)\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$

I termini simmetrici sono uguali a due a due

Distance Matrix (matrice delle distanze) /2

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}, \quad d_{ij} = d(X_i, X_j)$$

OTTIMIZZAZIONE INFORMATICA!

E sufficiente considerare la matrice triangolare al di sopra o al di sotto della diagonale principale di D.
Risparmiamo RAM!

Distance Matrix (matrice delle distanze)

Non esiste una sola funzione distanza, ma esiste un'intera famiglia di funzioni che rispettano almeno le quattro proprietà precedenti. Abbiamo inoltre che:

(a) se d e d' sono due metriche anche $d + d'$ è una metrica;

(b) se d è una metrica e c un numero reale positivo allora anche $c\,d$ è una metrica;

(c) se d è una metrica e c un numero reale positivo allora anche $d' = d/(c + d)$ è una metrica.

ATTENZIONE: Il prodotto di due metriche (in particolare il quadrato di una metrica) **non necessariamente soddisfa la disuguaglianza triangolare** e quindi può non essere una misura di distanza.

Metrica Euclidea (d_2)

Consideriamo due entità $\mathbf{I_i}$ e $\mathbf{I_j}$ aventi **p** caratteristiche quantitative.

Possiamo definire la metrica di distanza euclidea come:

$$d_2(X_i, X_j) = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2}$$

$\mathbf{x_{jk}}$ è il valore della k–esima caratteristica dell'individuo **Ij**.

$\mathbf{x_{ik}}$ è il valore della k–esima caratteristica dell'individuo **Ii**.

Curiosità: Se si considerano due caratteristiche, ossia p = 2, l'espressione **corrisponde alla radice quadrata della somma dei quadrati costruiti sui cateti di un triangolo rettangolo**

Per il teorema di Pitagora tale radice fornisce la misura dell'ipotenusa del triangolo stesso.

IMPORTANTE: La distanza Euclidea usata su tutti i dati è **fortemente influenzata dall'unità di misura** in base alla quale è valutata ciascuna delle p caratteristiche. → Ora lo vediamo!

Metrica Euclidea (d_2) / 2

$$d_2(X_i, X_j) = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right]^{1/2}$$

ESEMPIO

Consideriamo un dataset (X) contenente peso in Kg (C_1) e Altezza in cm (C_2) di tre persone. In questo caso, abbiamo due caratteristiche (dunque p=2)

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 60 & 160 \\ 65 & 165 \\ 63 & 170 \end{pmatrix} \end{array} \begin{array}{l} \\ X_1 \\ X_2 \\ X_3 \end{array} \quad \leftrightarrow \quad X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

Esercizio 1

Calcoliamo **d2** in base al peso

d_2(X1,X2) = sqrt((60-65)^2 + (160-165)^2) → sqrt (50) = 7,071

d_2(X1,X3) = sqrt((60-63)^2 + (160-170)^2) → sqrt (109) = 10.44

d_2(X2,X3) = sqrt((65-63)^2 + (165-170)^2) → sqrt (29) = 5.38

Metrica Euclidea (d_2) / 2

```
> X<-data.frame(peso=c(60,65,63),altezza=c(160,165,170))
> row.names(X)<-c("I1","I2","I3")
> X #visualizza il data frame X
   peso altezza
I1   60      160
I2   65      165
I3   63      170
>
> dist(X,method="euclidean",diag=TRUE,upper=TRUE)
           I1         I2         I3
I1  0.000000   7.071068 10.440307
I2  7.071068   0.000000  5.385165
I3 10.440307   5.385165  0.000000
```

Metrica Euclidea (d_2) / 3

```
> Y<-data.frame(peso=c(60,65,63),altezza=c(1.60,1.65,1.70))
> row.names(Y)<-c("I1","I2","I3")
> Y #visualizza il data frame Y
   peso altezza
I1   60     1.60
I2   65     1.65
I3   63     1.70
>
> dist(Y,method="euclidean",diag=TRUE,upper=TRUE)
          I1        I2        I3
I1 0.000000 5.000250 3.001666
I2 5.000250 0.000000 2.000625
I3 3.001666 2.000625 0.000000
```

Metrica Euclidea (d_2) / 4

Cosa notate, osservando le due matrici di distanza calcolate usando prima i centimetri e poi I metri? **Notate che stiamo parlando delle stesse tre persone.**

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 60 & 160 \\ 65 & 165 \\ 63 & 170 \end{pmatrix} \end{array}$$

$$Y = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 60 & 1.60 \\ 65 & 1.65 \\ 63 & 1.70 \end{pmatrix} \end{array}$$

	I1	I2	I3
I1	0.000000	7.071068	10.440307
I2	7.071068	0.000000	5.385165
I3	10.440307	5.385165	0.000000

	I1	I2	I3
I1	0.000000	5.000250	3.001666
I2	5.000250	0.000000	2.000625
I3	3.001666	2.000625	0.000000

Metrica Euclidea (d_2) / 4

Cosa notate, osservando le due matrici di distanza calcolate usando prima i centimetri e poi I metri? **Notate che stiamo parlando delle stesse tre persone.**

Nel primo caso l'individuo I1 e più simile a I2
Nel secondo caso lo stesso I1 è più simile I3

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 60 & 160 \\ 65 & 165 \\ 63 & 170 \end{pmatrix} \end{array}$$

$$Y = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 60 & 1.60 \\ 65 & 1.65 \\ 63 & 1.70 \end{pmatrix} \end{array}$$

	I1	I2	I3
I1	0.000000	7.071068	10.440307
I2	7.071068	0.000000	5.385165
I3	10.440307	5.385165	0.000000

	I1	I2	I3
I1	0.000000	5.000250	3.001666
I2	5.000250	0.000000	2.000625
I3	3.001666	2.000625	0.000000

Metrica Euclidea (d_2) / 5

Purtroppo non esiste una trasformazione che permetta di passare dai primi valori della distanza ai secondi valori, il che significa che **la funzione distanza è legata alle unità di misura in maniera non invariante.**

Il metodo più utilizzato per ovviare a questo inconveniente è quello di **scalare e standardizzare** inizialmente le misure.

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{s_j} \qquad (i = 1, 2, \ldots, n; j = 1, 2, \ldots, p)$$

In questo modo si ha la possibilità di un confronto tra le misure, ossia considerare delle nuove variabili

Scalare e standardizzare le misure

Purtroppo non esiste una trasformazione che permetta di passare dai primi valori della distanza ai secondi valori, il che significa che **la funzione distanza è legata alle unità di misura in maniera non invariante.**

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{s_j} \qquad (i = 1, 2, \ldots, n; j = 1, 2, \ldots, p)$$

Media campionaria j-esima caratteristica

Deviazione standard

$$\text{scale}(\text{X}, \text{center} = \text{TRUE}, \text{scale} = \text{TRUE})$$

Scalare e standardizzare le misure /2

	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}
mean	$\overline{x}_1$	$\overline{x}_2$	$\dots$	$\overline{x}_j$	$\dots$	$\overline{x}_p$
var	s_1^2	s_2^2	$\dots$	s_j^2	$\dots$	s_p^2

	x_{11}^*	x_{12}^*	$\dots$	x_{1j}^*	$\dots$	x_{1p}^*
	x_{21}^*	x_{22}^*	$\dots$	x_{2j}^*	$\dots$	x_{2p}^*
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{i1}^*	x_{i2}^*	$\dots$	x_{ij}^*	$\dots$	x_{ip}^*
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{n1}^*	x_{n2}^*	$\dots$	x_{nj}^*	$\dots$	x_{np}^*
mean	0	0	$\dots$	0	$\dots$	0
var	1	1	$\dots$	1	$\dots$	1

Scalare e standardizzare le misure /2

$$x^*_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \qquad (i = 1, 2, \ldots, n; j = 1, 2, \ldots, p)$$

	x^*_{11}	x^*_{12}	$\ldots$	x^*_{1j}	$\ldots$	x^*_{1p}
	x^*_{21}	x^*_{22}	$\ldots$	x^*_{2j}	$\ldots$	x^*_{2p}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x^*_{i1}	x^*_{i2}	$\ldots$	x^*_{ij}	$\ldots$	x^*_{ip}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x^*_{n1}	x^*_{n2}	$\ldots$	x^*_{nj}	$\ldots$	x^*_{np}
mean	0	0	$\ldots$	0	$\ldots$	0
var	1	1	$\ldots$	1	$\ldots$	1

$$\overline{x}^*_j = \frac{1}{n}\sum_{i=1}^{n} x^*_{ij} = \frac{1}{n}\sum_{i=1}^{n} \frac{x_{ij} - \overline{x}_j}{s_j} = \frac{1}{n\, s_j}\left[\sum_{i=1}^{n} x_{ij} - n\overline{x}_j\right] = 0,$$

$$(s^*_j)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x^*_{ij} - \overline{x}^*_j\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{ij} - \overline{x}_j}{s_j}\right)^2 = 1.$$

scale(X, center = TRUE, scale = TRUE)

Metrica Manhattan L_1 (aka del valore assoluto)

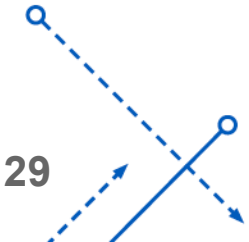
La metrica Euclidea è molto usata nelle tecniche di clustering. Tuttavia, esistono molte altre possibili metriche.

$$d_1(X_i, X_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

```
dist(Z,method="manhattan",diag=TRUE,upper=TRUE)
```

Se si considerano due caratteristiche, ossia p = 2, la metrica del valore assoluto corrisponde alla somma delle misure dei due cateti di un triangolo rettangolo.

Curiosità: Il suo nome “Manhattan” deriva proprio dal fatto che essa corrisponde alla lunghezza che si deve percorrere qualora sia consentito di muoversi solo nelle direzioni parallele agli assi, come avviene in una città una griglia regolare di strade che si intersecano ad angolo retto.



Metrica Cebycev (aka del Massimo)

$$d_{\infty}(X_i, X_j) = \max_{k=1,2...,p} |x_{ik} - x_{jk}|.$$

Entrambe le metriche del valore assoluto e del massimo sono computazionalmente semplici da calcolare con l'unica differenza che la metrica di Chebycev coinvolge anche una procedura di ordinamento

```
dist(Z,method="maximum",diag=TRUE,upper=TRUE)
```

Metrica di Minkowski (L_r)

Una misura di distanza che include come caso particolare la distanza Euclidea, la metrica del valore assoluto e la metrica di Chebycev risulta essere la metrica di Minkowski, detta anche metrica Lr, così definite:

$$d_r(X_i, X_j) = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r\right]^{1/r},$$

Curiosità:
La distanza di Minkowski (o city-block) trova applicazioni nell'urbanistica ed in particolare nella costruzione di strade in una città in cui sono già presenti degli edifici.

NOTE:

- Se r = 2 si ottiene la metrica Euclidea
- Se r = 1 si ottiene la metrica del valore assoluto
- Se r = ∞ si ottiene la metrica di Chebycev.

$$d_\infty(X_i, X_j) \leq d_2(X_i, X_j) \leq d_1(X_i, X_j).$$

```
dist(Z,method="minkowski",4,diag=TRUE,upper=TRUE)
```

Metrica di Canberra (d_c) /1

Un'altra possibile metrica è la metrica di Canberra.

$$d_c(X_i, X_j) = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|},$$

La metrica di Canberra è definita per variabili non negative ed ha la caratteristica di essere **sensibile alle differenze relative** piuttosto che a quelle assolute.

Questa distanza assegna alla differenza fra i valori relativi ai vettori **Xi** e **Xj** un peso inversamente proporzionale alla somma dei valori stessi.

Uno dei problemi di questa distanza è che, se uno dei due valori $\mathbf{x_{ik}}$ o $\mathbf{x_{jk}}$ uguale a zero, allora il contributo $|x_{ik} - x_{jk}|/|x_{ik} + x_{jk}|$ alla distanza totale è uguale a 1, ossia il massimo possibile.

Metrica di Canberra (d_c) /2

Alcune proprietà important della metrica di Canberra

$$d_c(X_i, X_j) = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|},$$

```
dist(X,method="canberra",diag=TRUE,upper=TRUE)
```

Se si utilizza tale metrica non è necessario scalare la matrice dei dati, poiché i contributi alla somma sono adimensionali.

La metrica di Canberra è poco sensibile all'asimmetria delle distribuzioni delle variabili (caratteristiche) e **alla presenza di eventuali valori anomali (outlier)**

Distanza di Jaccard /1

$$d(X_i, X_j) = 1 - \frac{\sum_{k=1}^{p} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{p} \max(x_{ik}, x_{jk})}$$

```
dist(U,method="binary",diag=TRUE,upper=TRUE)
```

Ricordate che in R questa metrica è soltanto per vettori binari.

$$J_\delta(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$