Analisi dei Dati

Variabilità e distanze

Domenico De Stefano

a.a. 2021/2022

Guardando oltre al centro della distribuzione

Ci interessa avere anche un'idea di quanto diversi siano i valori assunti dalla variabile, ossia ci interessa avere un'idea della variabilità di un carattere

Guardando oltre al centro della distribuzione

Ci interessa avere anche un'idea di quanto diversi siano i valori assunti dalla variabile, ossia ci interessa avere un'idea della variabilità di un carattere

Per farlo, possiamo vedere come si muovono le osservazioni intorno al centro della distribuzione.

Guardando oltre al centro della distribuzione

Ci interessa avere anche un'idea di quanto diversi siano i valori assunti dalla variabile, ossia ci interessa avere un'idea della variabilità di un carattere

Per farlo, possiamo vedere come si muovono le osservazioni intorno al centro della distribuzione.

E per fare ciò, possiamo usare l'idea di "distanza".

Esempio: assenza di variabilità

Se non c'è variabilità, tutte le unità statistiche mostrano la stessa modalità del carattere.

$$1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1$$

Abbiamo

- $x_{(1)} = 1{,}1 \;\; x_{(N)} = 1{,}1$
- $q_{0.25} = 1{,}1 \;\; m = 1{,}1 \;\; q_{0.75} = 1{,}1$

Misurando distanze dal centro della distribuzione, possiamo costruire indicatori che valgono 0 in assenza di variabilità.

- $|x_i - m| = |x_i - x_j| = 0, \;\; i, j = 1, \ldots, N,$
- $x_{(N)} - m = m - x_{(1)} = 0,$
- $q_{0.75} - m = m - q_{0.25} = 0.$

Indici elementari di variabilità

- $x_{(N)} - x_{(1)}$ è il *campo di variazione* (range).
- $q_{0.75} - q_{0.25}$ è la *distanza interquartilica* (IQR).

Indici elementari di variabilità

- $x_{(N)} - x_{(1)}$ è il *campo di variazione* (range).
- $q_{0.75} - q_{0.25}$ è la *distanza interquartilica* (IQR).

Ovviamente, in presenza di variabilità, sia il campo di variazione che la distanza interquartilica assumono un valore maggiore di zero.

E, in presenza di variabilità, possiamo cercare di rappresentare come variano le modalità.

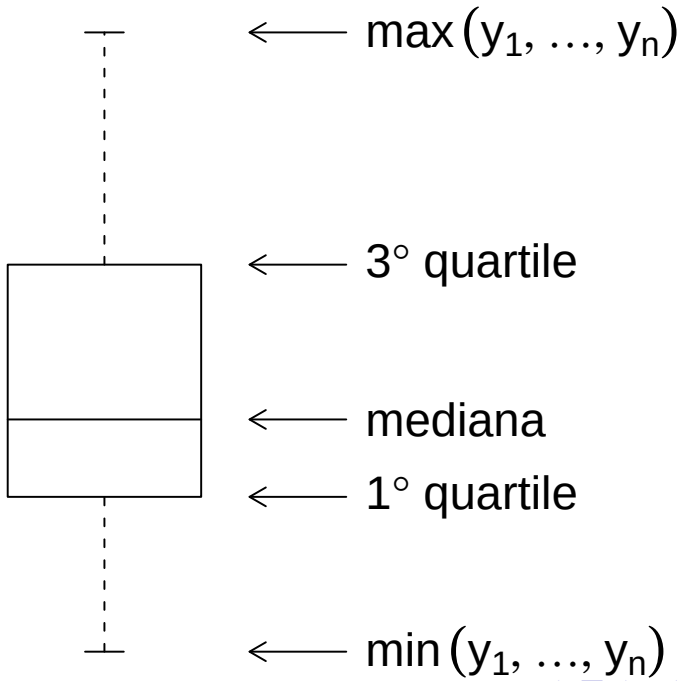
Diagramma a scatola con baffi (*box and whiskers plot o boxplot*)

Il boxplot è un grafico molto utilizzato in statistica. Esso fornisce un'idea schematica di un insieme di dati (di una distribuzione) basata sui quartili.

Diagramma a scatola con baffi (*box and whiskers plot o boxplot*)

Il boxplot è un grafico molto utilizzato in statistica. Esso fornisce un'idea schematica di un insieme di dati (di una distribuzione) basata sui quartili.

Sono costituiti, come dice il nome, da una *scatola* e da due *baffi* costruiti in accordo al disegno sottostante.



Ancora sulla variabilità

Abbiamo detto che per misurare la variabilità, possiamo utilizzare la "distanza" delle osservazioni dal centro della distribuzione.

Ancora sulla variabilità

Abbiamo detto che per misurare la variabilità, possiamo utilizzare la "distanza" delle osservazioni dal centro della distribuzione.

Proviamo a utilizzare la media per caratterizzare il centro della distribuzione.

Ancora sulla variabilità

Abbiamo detto che per misurare la variabilità, possiamo utilizzare la "distanza" delle osservazioni dal centro della distribuzione.

Proviamo a utilizzare la media per caratterizzare il centro della distribuzione.

Siano $x = (x_1, \ldots, x_N)$ i dati osservati, N il loro numero e $\overline{x}$ la loro media aritmetica, ovvero $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

La distanza di ogni osservazione x_i dalla media $\overline{x}$, il cosidetto scarto dalla media, può essere misurata così:

$$|x_i - \overline{x}|.$$

Ancora sulla variabilità

Abbiamo detto che per misurare la variabilità, possiamo utilizzare la "distanza" delle osservazioni dal centro della distribuzione.

Proviamo a utilizzare la media per caratterizzare il centro della distribuzione.

Siano $x = (x_1, \ldots, x_N)$ i dati osservati, N il loro numero e $\overline{x}$ la loro media aritmetica, ovvero $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

La distanza di ogni osservazione x_i dalla media $\overline{x}$, il cosidetto scarto dalla media, può essere misurata così:

$$|x_i - \overline{x}|.$$

Perché abbiamo bisogno del valore assoluto?

Ancora sulla variabilità (cont)

È ancora meglio se consideriamo lo scarto al quadrato:

$$(x_i - \overline{x})^2.$$

Perché il quadrato?

Ancora sulla variabilità (cont)

È ancora meglio se consideriamo lo scarto al quadrato:

$$(x_i - \overline{x})^2.$$

Perché il quadrato?

Perché il quadrato "amplifica" le distanze grandi e "attenua" quelle piccole.

Esempio: $10^2 = 100$, $0.1^2 = 0.01$

Ancora sulla variabilità (cont)

È ancora meglio se consideriamo lo scarto al quadrato:

$$(x_i - \overline{x})^2.$$

Perché il quadrato?

Perché il quadrato "amplifica" le distanze grandi e "attenua" quelle piccole.

Esempio: $10^2 = 100$, $0.1^2 = 0.01$
Quindi, per costruire un indice di variabilità, possiamo costruire queste N quantità (per $i = 1, \ldots, N$) e farne una media.

Varianza

La varianza è la media dei quadrati degli scarti di ogni osservazione dalla media arjtmetica.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

Varianza

La varianza è la media dei quadrati degli scarti di ogni osservazione dalla media arjtmetica.
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$ Esempio: ore di studio per settimana

- La media è $\bar{x} = 18.58$.
- La varianza è calcolata come:
  $\sigma^2 = \frac{(2-18.58)^2+(30-18.58)^2+...+(42-18.58)^2}{51} = 183.12$

Deviazione standard

La deviazione standard è la radice quadrata della varianza ed è espressa nella stessa unità di misura del carattere.
$\sigma = \sqrt{\sigma^2}$

La deviazione standard per le ore di studio/settimana degli studenti si calcola come: $\sigma = \sqrt{183.12} = 13.53$

Devianza

La deviazione standard non deve essere confusa con la devianza, che è la quantità al numeratore della varianza.
$\sum_{i=1}^{N}(x_i - \bar{x})^2$
La devianza rappresenta quindi la somma dei quadrati degli scarti delle osservazioni dalla propria media.

Varianza campionaria corretta

Quando si lavora con un campione (quindi nella stragrande maggioranza dei casi...), si utilizza spesso la varianza campionaria corretta, che differisce dalla varianza campionaria solo per il denominatore (che anzichè N è uguale a $N-1$):

$s^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}$

La ragione della modifica del denominatore è legata a proprietà teoriche di s^2 che la rendono una misura di variabilità più comoda quando farete inferenza.

Varianza: una formula operativa (cont)

Una formula operativa è la seguente:

$$\sigma^2 = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^2 \right) - \overline{x}^2$$

ovvero

$$(\text{varianza}) = \left(\begin{array}{c} \text{media dei} \\ \text{quadrati} \end{array} \right) - \left(\begin{array}{c} \text{quadrato della} \\ \text{media} \end{array} \right).$$

Indice

Forma di una distribuzione

Oltre alla media e alla varianza (e deviazione standard), ci sono altri aspetti da valutare per "descrivere" una distribuzione.

Quanti picchi mostra l'istogramma: uno (distribuzione *unimodale*), molti (distribuzione *bimodale/multimodale*), o nessuno (distribuzione *uniforme*)?

unimodale

bimodale

multimodale

uniforme

Indice

Distanze e similarità

Le misure di distanza o similarità sono essenziali in analisi dei dati, in particolare per metodi di classificazione e raggrupamento (clustering). Esistono diverse misure di distanza o similarità che è possibile usare per confrontare due insiemi di dati.

In generale:

- Le misure di similarità restituiscono un valore numerico che esprime il grado di "somiglianza" tra due "vettori". Spesso tale valore varia tra 0 (nessuna similarità) a 1 (massima similarità)
- Le misure di dissimilarità restituiscono un valore numerico che esprime quanto sono differenti due insiemi di dati. Spesso tale valore varia tra 0 (massima similarità) a inf (massima dissimilarità). Le distanze sono esempi di misure di dissimilarità
- Misure di prossimità: termine che si riferisce in generale a misure di similarità o dissimilarità

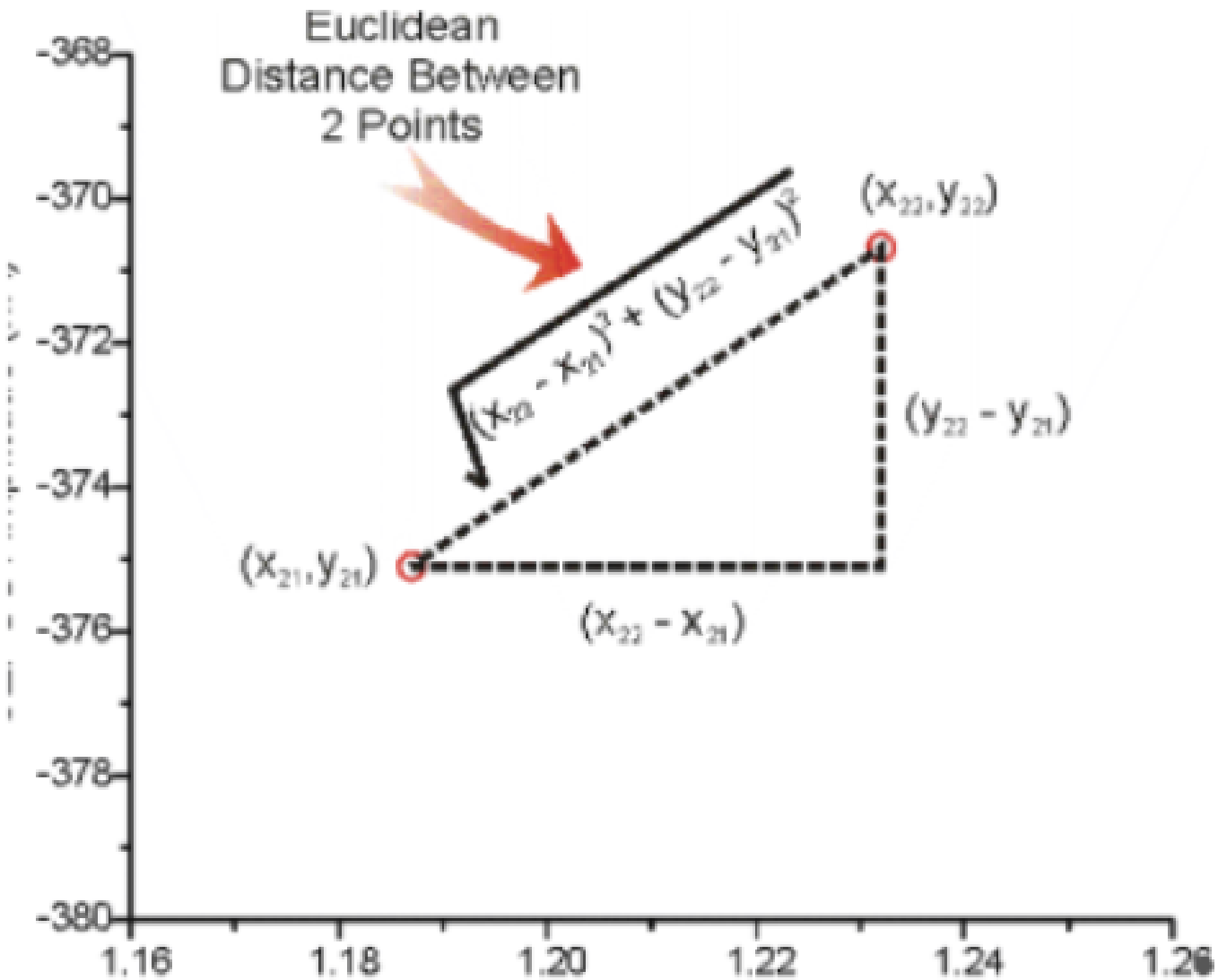
Distanza euclidea

La misura di dissimilarità (distanza) più comunemente usata è la distanza euclidea. Siano $\mathbf{p}$, $\mathbf{q}$ due vettori $\mathbf{p} = (p_1, p_2, \ldots, p_i, \ldots, p_k)$ e $\mathbf{q} = (q_1, q_2, \ldots, q_i, \ldots, q_k)$, dove k è il numero di dimensioni e p_i e q_i rappresentano l'i-mo elemento (o componente) di $\mathbf{p}$ e $\mathbf{q}$, rispettivamente

$$d_{euclidea}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{k}(p_i - q_i)^2}$$

- Si usa solo per vettori numerici (cioè quando abbiamo variabili quantitative)
- Rappresenta la distanza più "breve" tra due punti nel modno fisico
- Computazionalmente costosa

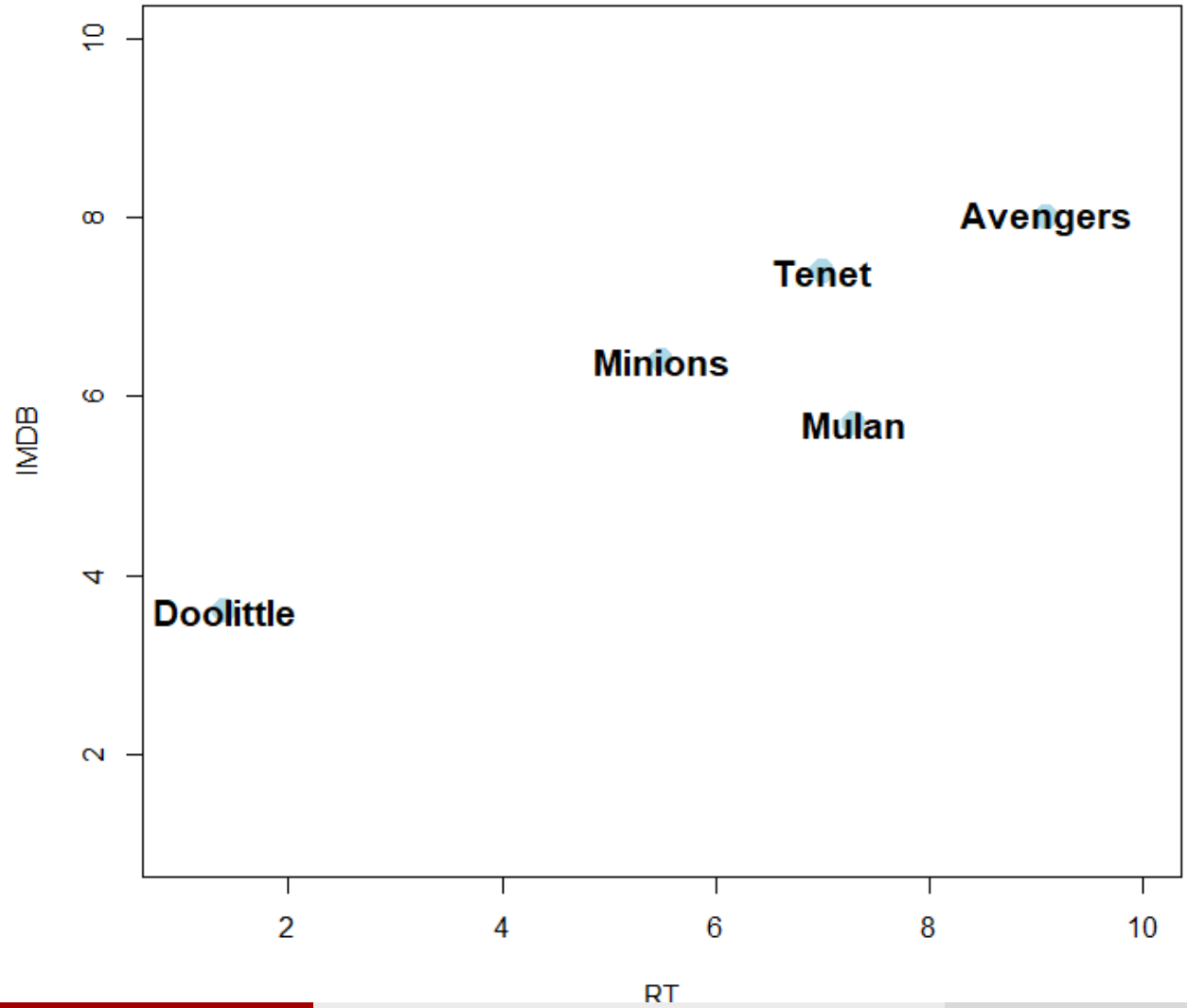
Distanza euclidea



Distanza euclidea: esempio

film	IMDB	RT
Avengers	8	9.1
Minions	6.4	5.5
Tenet	7.4	7
Mulan	5.7	7.3
Dolittle	3.6	1.4

Distanza euclidea: esempio



Distanza euclidea: esempio

Calcoliamo la distanza tra i film Doolittle e Avengers sulla base delle valutazioni su IMDB e RT (2 dimensioni)

$$d_{euclidea}(\textbf{Doolittle}, \textbf{Avengers}) =$$

$$= \sqrt{(9.1 - 1.4)^2 + (8 - 3.6)^2} = \sqrt{7.7^2 + 4.4^2} =$$

$$= \sqrt{59.3 + 19.4} = \sqrt{78.7} = 8.87$$

Distanza euclidea: esempio

Matrice di distanze (euclidee)

	Avengers	Minions	Tenet	Mulan	Doolittle
Avengers	0.00	3.94	2.18	2.92	8.87
Minions	3.94	0.00	1.80	1.93	4.96
Tenet	2.18	1.80	0.00	1.73	6.77
Mulan	2.92	1.93	1.73	0.00	6.26
Doolittle	8.87	4.96	6.77	6.26	0.00

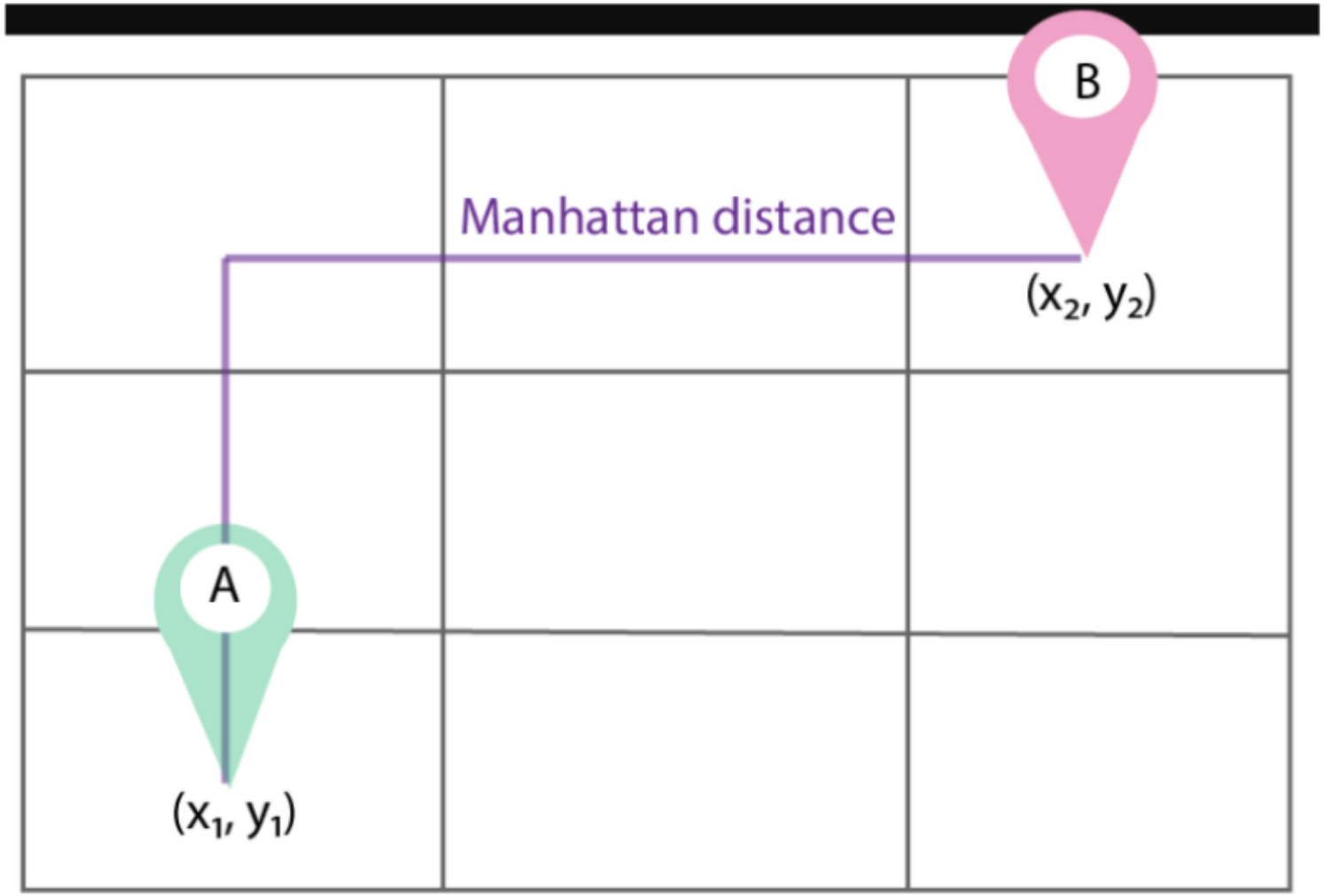
Altre distanze: Distanza di Manhattan

La distanza di manhattan (o taxicab distance) è un altro tipo di dissimilarità spesso usata in caso di variabili discrete. Siano $\mathbf{p}$, $\mathbf{q}$ due vettori $\mathbf{p} = (p_1, p_2, \ldots, p_i, \ldots, p_k)$ e $\mathbf{q} = (q_1, q_2, \ldots, q_i, \ldots, q_k)$, dove k è il numero di dimensioni e p_i e q_i rappresentano l'i-mo elemento (o componente) di $\mathbf{p}$ e $\mathbf{q}$, rispettivamente

$$d_{manhattan}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{k} |p_i - q_i|$$

- può essere definita come il percorso più breve che un taxi farebbe in una città come Manhattan (con struttura a griglia)
- Si usa principalmente quando i punti sono disposti a scacchiera (cioè in caso di variabili quantitative discrete)
- è computazionalmente meno costosa rispetto alla distanza euclidea ed è a questa preferita quando le dimensioni sono tante (es k=1000)

Altre distanze: Distanza di Manhattan



Distanza di Manhattan: esempio

$$d_{manhattan}(\textbf{Avengers}, \textbf{Doolittle}) = |9.1 - 1.4| + |8 - 3.6| = 7.7 + 4.4 = 12.1$$

Matrice di distanze (manhattan)

	Avengers	Minions	Tenet	Mulan	Doolittle
Avengers	0.00	5.20	2.70	4.10	12.10
Minions	5.20	0.00	2.50	2.50	6.90
Tenet	2.70	2.50	0.00	2.00	9.40
Mulan	4.10	2.50	2.00	0.00	8.00
Doolittle	12.10	6.90	9.40	8.00	0.00

Similarità del coseno

La misura di similarità più comunemente usata è la similarità del coseno (cosine similarity). Siano $\mathbf{p}$, $\mathbf{q}$ due vettori $\mathbf{p} = (p_1, p_2, \ldots, p_i, \ldots, p_k)$ e $\mathbf{q} = (q_1, q_2, \ldots, q_i, \ldots, q_k)$, dove k è il numero di dimensioni e p_i e q_i rappresentano l'i-mo elemento (o componente) di $\mathbf{p}$ e $\mathbf{q}$, rispettivamente

$$sim_{coseno}(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p} \bullet \mathbf{q}}{\|\mathbf{p}\| \|\mathbf{q}\|}$$

- è il coseno dell'angolo formato dai due vettori che si calcola facendo il rapporto tra il prodotto scalare tra i due vettori diviso il prodotto tra le loro norme (lunghezze dei vettori)
- è una similarità molto semplice che tiene solo conto della direzione dei vettori
- Assume valori tra -1 e 1 (massima similarità)
- -1 si ha quando tra i due vettori si ha un angolo di 180^o e 1 quando si ha un angolo di 0^o