

Analisi Esplorativa

Aldo Solari



- ① Aspetti organizzativi
- ② L'analisi multivariata
- ③ Riduzione della dimensionalità
- ④ Raggruppamento delle unità statistiche



Outline

- ➊ Aspetti organizzativi
- ➋ L'analisi multivariata
- ➌ Riduzione della dimensionalità
- ➍ Raggruppamento delle unità statistiche



Docente

E-mail : aldo.solari@unimib.it
Ricevimento : Mercoledì, 17:00-18:00, stanza 2030, edificio U7
Pagina personale : <https://aldosolari.github.io/>

Tutor : Dott.ssa Laura Belloni



Pagina web del corso

<https://aldosolari.github.io/AE/>

dove potete trovare:

- Ultimi avvisi
- Calendario delle lezioni
- Materiale didattico da scaricare
- Modalità d'esame
- Etc.



Modalità d'esame

La prova parziale di AE si svolgerà in Laboratorio e consiste in:

- ① 10 domande a risposta multipla: si accede alla prova successiva solo se si risponde correttamente ad almeno 6 domande
- ② Prova scritta e di laboratorio con R

Il risultato della prova parziale di AE **rimane valido** per 4 anni



Appelli d'esame

- Gennaio-Febbraio: 2 appelli
- Aprile: 1 appello
- Giugno-Luglio: 2 appelli
- Settembre: 1 appello

Per sostenere la prova parziale di AE è necessario:

- ① essere iscritti all'appello (tramite piattaforma Esse3)
- ② avere con sé un documento di riconoscimento valido

Visione dei compiti:

- ① Verrà stabilita un'unica data per la visione della prova scritta e di laboratorio con R
- ② Non è prevista la visione della prima prova (questionario)



Propedeuticità

- Algebra lineare
- Analisi Matematica I
- Calcolo delle Probabilità
- Statistica I

Si consiglia inoltre la conoscenza degli argomenti trattati nel modulo R per l'Analisi Statistica Multivariata



Outline

- ① Aspetti organizzativi
- ② L'analisi multivariata
- ③ Riduzione della dimensionalità
- ④ Raggruppamento delle unità statistiche



L'analisi multivariata

- Riguarda l'*analisi congiunta di più variabili* misurate sul medesimo insieme di unità statistiche.
- In qualche caso ha senso l'analisi delle singole variabili raccolte, molto più spesso le variabili sono legate in modo tale che solo un'analisi congiunta di esse permette di rilevare pienamente la struttura dei dati
- Le tecniche per l'analisi di dati multivariati possono avere una natura *descrittiva/esplorativa* oppure *inferenziale*
- Per gli scopi di questo corso, ci occuperemo principalmente delle tecniche descrittive/esplorative, lasciando gli aspetti inferenziali a corsi più avanzati



Obiettivi

Fra i molteplici obiettivi dell'analisi multivariata considereremo:

- ❶ Esplorazione di dati multidimensionali
(*exploratory analysis*)
- ❷ Riduzione della dimensionalità dei dati
(*dimensionality reduction*)
 - Analisi delle componenti principali
(*principal component analysis*)
 - Analisi fattoriale
(*factor analysis*)
- ❸ Raggruppamento delle unità statistiche
(*cluster analysis*)
 - *k*-medie (*k-means*)
 - analisi dei gruppi gerarchica (*hierarchical clustering*)



Unsupervised learning

Nella nomenclatura della letteratura *machine learning* questi temi vanno sotto il nome di *unsupervised learning*

Significa che l'apprendimento non è guidato da una variabile risposta, come invece accade nei problemi di *supervised learning*

	<i>Output</i> discreto	<i>Output</i> continuo
<i>Supervised learning</i>	Classificazione	Regressione
<i>Unsupervised learning</i>	Raggruppamento	Riduzione dimensionalità



Outline

- ① Aspetti organizzativi
- ② L'analisi multivariata
- ③ Riduzione della dimensionalità**
- ④ Raggruppamento delle unità statistiche



Riduzione della dimensionalità

$$\underset{n \times p}{X} \mapsto \underset{n \times q}{Y}$$

Input

matrice $\underset{n \times p}{X}$ con p variabili quantitative

Output

matrice $\underset{n \times q}{Y}$ con $q < p$ variabili quantitative

Obiettivo

Ridurre la dimensione perdendo meno informazione possibile



Dati heptathlon

L'eptathlon è una specialità dell'atletica leggera che contempla $p = 7$ gare di discipline diverse:

- 100 metri ostacoli
- salto in alto
- getto del peso
- 200 metri piani
- salto in lungo
- tiro del giavellotto
- 800 metri piani

I dati che abbiamo a disposizione riguardano i risultati di $n = 25$ atlete alle Olimpiadi di Seul del 1988



	hurdles	highjump	shot	run200m	longjump	javelin	run800m
Joyner-Kersey (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20
Sablovskaitė (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90
Schulz (GDR)	13.75	1.83	13.50	24.65	6.33	42.82	125.79
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74
Wijnsma (HOL)	13.75	1.86	13.01	25.03	6.34	37.86	131.49
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54
Scheider (SWI)	13.85	1.86	11.58	24.87	6.05	47.50	134.93
Braun (FRG)	13.71	1.83	13.16	24.78	6.12	44.58	142.82
Ruotsalainen (FIN)	13.79	1.80	12.32	24.61	6.08	45.44	137.06
Yuping (CHN)	13.93	1.86	14.21	25.00	6.40	38.60	146.67
Hagger (GB)	13.47	1.80	12.75	25.47	6.34	35.76	138.48
Brown (USA)	14.07	1.83	12.69	24.83	6.13	44.34	146.43
Mulliner (GB)	14.39	1.71	12.68	24.92	6.10	37.76	138.02
Hautenaue (BEL)	14.04	1.77	11.81	25.61	5.99	35.68	133.90
Kytola (FIN)	14.31	1.77	11.66	25.69	5.75	39.48	133.35
Geremias (BRA)	14.23	1.71	12.95	25.50	5.50	39.64	144.02
Hui-Ing (TAI)	14.85	1.68	10.00	25.23	5.47	39.14	137.30
Jeong-Mi (KOR)	14.53	1.71	10.83	26.61	5.50	39.26	139.17
Launa (PNG)	16.42	1.50	11.78	26.16	4.88	46.38	163.43



Obiettivo

Determinare un punteggio da attribuire a ciascun atleta che sintetizzi le *performance* nelle sette gare al fine di ottenere la classifica finale

ovvero ridurre la dimensionalità da $p = 7$ a $q = 1$:

$$\underset{25 \times 7}{X} \mapsto \underset{25 \times 1}{y}$$



Punteggio finale

	score
Joyner-Kersee (USA)	7291
John (GDR)	6897
Behmer (GDR)	6858
Sablovskaitė (URS)	6540
Choubenkova (URS)	6540
Schulz (GDR)	6411
Fleming (AUS)	6351
Greiner (USA)	6297
Lajbnerova (CZE)	6252
Bouraga (URS)	6252
Wijnsma (HOL)	6205
Dimitrova (BUL)	6171
Scheider (SWI)	6137
Braun (FRG)	6109
Ruotsalainen (FIN)	6101
Yuping (CHN)	6087
Hagger (GB)	5975
Brown (USA)	5972
Mulliner (GB)	5746
Hautenauve (BEL)	5734
Kytola (FIN)	5686
Geremias (BRA)	5508
Hui-Ing (TAI)	5290
Jeong-Mi (KOR)	5289
Launa (PNG)	4566



Dati face



X
 243×220



Immagine = dati

- Una immagine (in bianco e nero), può essere rappresentata come una matrice di dati, dove l'intensità di grigio di ogni pixel viene rappresentata nella corrispondente cella della matrice
- I colori più chiari sono associati valori più alti, colori più scuri sono associati valori più bassi (nel range $[0,1]$).

r/c	...	110	111	112	113	114	...
...
110	...	0.96	0.93	0.92	0.93	0.90	...
111	...	0.97	0.96	0.95	0.95	0.93	...
112	...	0.95	0.96	0.94	0.93	0.90	...
113	...	0.87	0.90	0.90	0.87	0.82	...
114	...	0.85	0.86	0.87	0.85	0.82	...
...



Immagine compressa



$$Y_{n \times q} V'_{q \times p} + \frac{1}{n \times 11 \times p} \bar{x}'$$

con $q = 10$



Pixels e bytes

Immagine originale

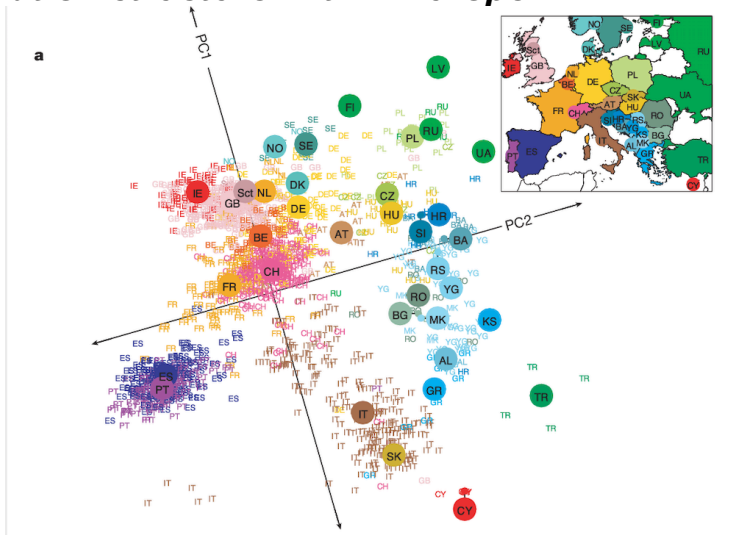
- $X_{243 \times 220}$: $243 \times 220 = 53460$ pixels
- Memoria richiesta: 427880 bytes

Immagine compressa

- $Y_{243 \times 10}, V_{220 \times 10}, \bar{x}_{220 \times 1}$: $243 \times 10 + 220 \times 10 + 220 = 4850$ pixels
- Memoria richiesta: 40872 bytes
- Fattore di riduzione = $427880 \text{ bytes} / 40872 \text{ bytes} = 10.47$



Population structure within Europe



$$n \approx 1300, p \approx 200000, q = 2$$

From: *Genes mirror geography within Europe*

<https://www.nature.com/articles/nature07331/figures/1>



Outline

- ① Aspetti organizzativi
- ② L'analisi multivariata
- ③ Riduzione della dimensionalità
- ④ Raggruppamento delle unità statistiche



Raggruppamento delle unità statistiche

$$\underset{n \times p}{X} \mapsto \underset{n \times 1}{y}$$

Input

matrice $\underset{n \times p}{X}$ con p variabili quantitative e/o qualitative

Output

$$\text{vettore } \underset{n \times 1}{y} = \begin{bmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_n \end{bmatrix} \text{ con } y_i \in \{G_1, G_2, \dots, G_k\}$$

dove G_1, G_2, \dots, G_k rappresenta il primo, \dots , il k -simo gruppo

Obiettivo

Formare k gruppi omogenei al loro interno e disomogenei tra di loro



Dati movielens

I dati che abbiamo a disposizione riguardano il *rating* (da 0.5 a 5) attribuito a $n = 9125$ film da parte di $p = 671$ utenti tra il 09 gennaio 1995 e il 16 ottobre 2016

Per l'esempio che segue considereremo $n = 50$ film e $p = 139$ utenti



	U8	U15	U17	U19	U20	U21	U22	U23	U24
Ace Ventura		2.0		3.0	1.0	3.0		2.0	0.0
Aladdin		0.5		3.0	3.5		2.0	4.0	
American Beauty	4.5	4.0	4.5				4.0	3.5	4.0
Apollo 13		3.0		3.0	3.0			3.5	
Back to the Future	4.0	5.0	4.5	5.0	3.5	4.0	4.0	4.5	
Batman		4.0		4.0	4.0	3.0	4.5	3.5	
Beauty and the Beast				5.0	4.0	3.0		4.5	
Braveheart	4.0	3.0		3.0	2.0			3.5	
Dances with Wolves		3.0	3.0	3.0	2.0	4.0		2.5	
Dumb & Dumber		3.5		3.0	1.0		2.5		
E.T.		4.0		5.0	1.5	3.0	2.5	5.0	
Fargo		5.0	3.5	5.0	2.0			4.5	3.0
Fight Club	4.0	5.0	5.0		0.5		4.0	3.5	4.0
Forrest Gump	4.0	1.0	2.5	5.0	2.0	4.0	3.5	4.5	4.0
Fugitive, The	4.5	5.0		4.0	4.5	3.0	4.5	3.5	3.0
Gladiator	5.0	2.0	4.0				3.0	4.0	2.0
Godfather, The	5.0	5.0	5.0	5.0	2.0	4.0	4.0	5.0	4.0
Good Will Hunting	4.0	4.0	4.0					3.5	
⋮									



Obiettivo

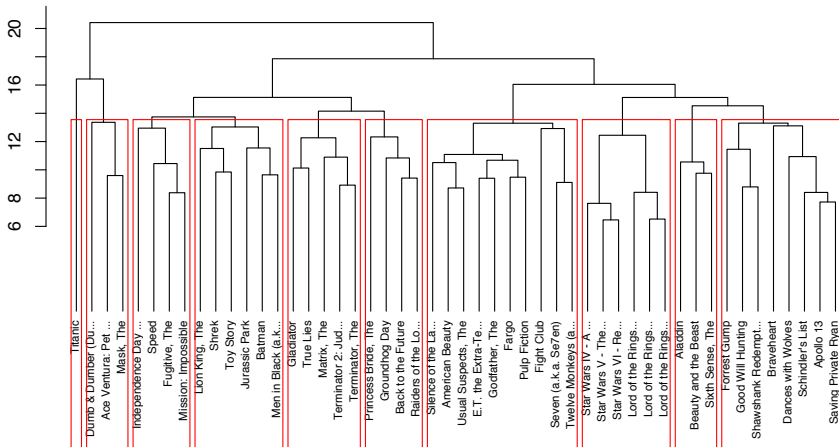
Raggruppare i film in gruppi omogenei al loro interno e disomogenei tra di loro rispetto al *rating* che hanno ottenuto dagli utenti

Ad esempio, se decidiamo di raggruppare i $n = 50$ film in $k = 10$ gruppi A, B, C, D, E, F, G, H, I, L

$$X_{50 \times 139} \mapsto y_{50 \times 1} = \begin{bmatrix} B \\ A \\ \dots \\ A \\ \dots \\ C \\ D \end{bmatrix}$$



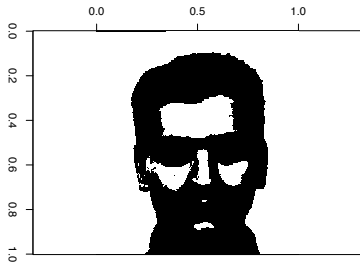
Height



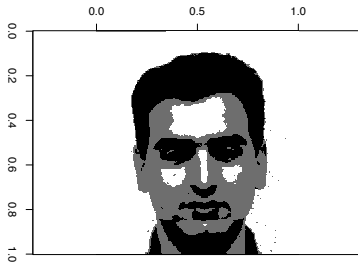
hclust (*, "complete")



Vector quantization



$$k = 2$$



$$k = 3$$

