Analisi Esplorativa

Aldo Solari

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

Outline

Docente

E-mail : aldo.solari@unimib.it
Pagina personale : https://aldosolari.github.io/

Pagina MOODLE

https://elearning.unimib.it/course/view.php?id=30594

dove potete trovare:

- Forum di discussione
- Registrazioni delle lezioni
- Link alla pagina WEB

Pagina WEB

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

dove potete trovare:

- Calendario delle lezioni
- Materiale didattico da scaricare
- Calendario degli esami
- Modalità d'esame
- Ecc.

Outline

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:

1. Esplorazione di dati multidimensionali (*exploratory analysis*)
2. Riduzione della dimensionalità dei dati (*dimensionality reduction*)
   - Analisi delle componenti principali (*principal component analysis*)
   - Analisi fattoriale (*factor analysis*)
3. 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*

	Output discreto	*Output* continuo
Supervised learning	Classificazione	Regressione
Unsupervised learning	Raggruppamento	Riduzione dimensionalità

Outline

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
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54
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
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54
Sablovskaite (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05
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
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74
Joyner-Kersee (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51
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
Hautenauve (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
Sablovskaite (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`

$$\underset{243\times220}{X}$$

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

$$\underset{n\times q}{Y}\ \underset{q\times p}{V'} + \underset{n\times 1}{1}\ \underset{1\times p}{\bar{x}'}$$

con $q = 10$

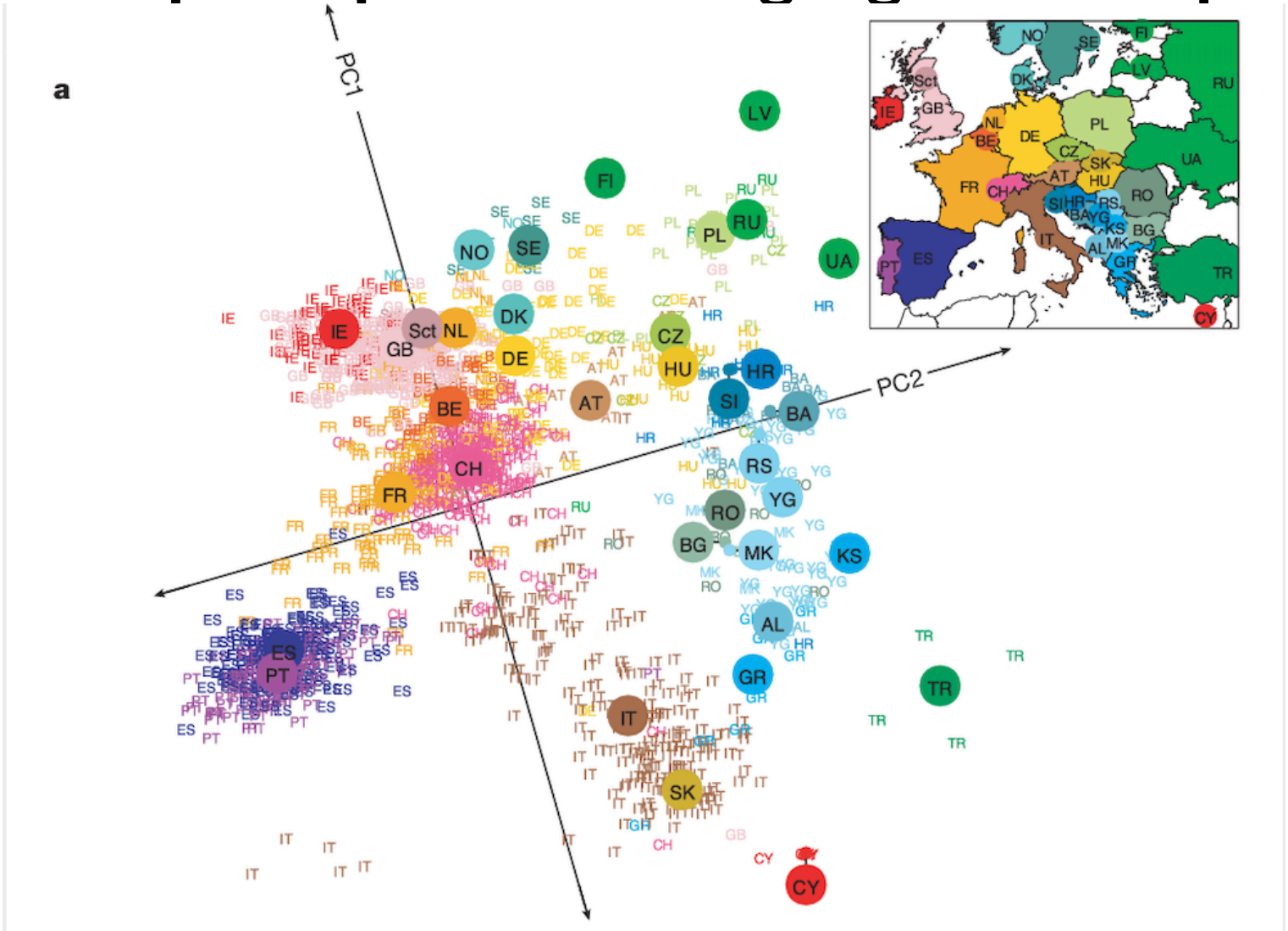
Immagine originale

$\underset{243\times220}{X}$: $243 \times 220 = 53460$ numeri

Immagine compressa

$\underset{243\times10}{Y}$, $\underset{220\times10}{V}$, $\underset{220\times1}{\bar{x}}$: $243 \times 10 + 220 \times 10 + 220 = 4850$ numeri

I geni europei rispecchiano la geografia europea?



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

Source: *Genes mirror geography within Europe*
https://www.nature.com/articles/nature07331/figures/1

Outline

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

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

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

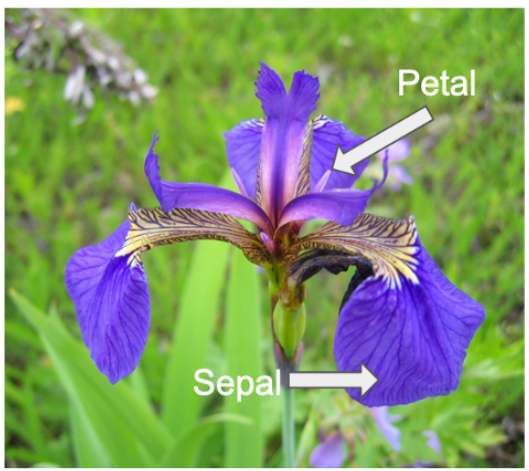
Obiettivo

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

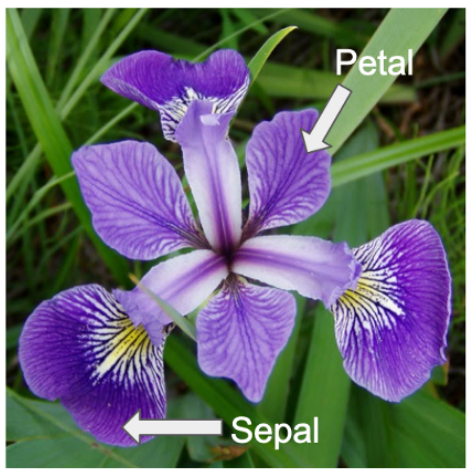
Dati `iris`

- Il dati `iris` sono stati analizzati da Ronald Fisher nel 1936
- Il dataset consiste in $n = 150$ fiori di genere Iris (dalla parola greca iris che significa arcobaleno) misurate da Edgar Anderson e classificate secondo tre specie: Iris setosa, Iris virginica e Iris versicolor

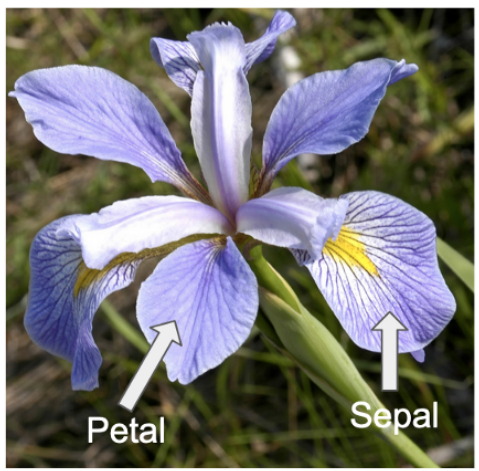
Iris setosa



Iris versicolor



Iris virginica



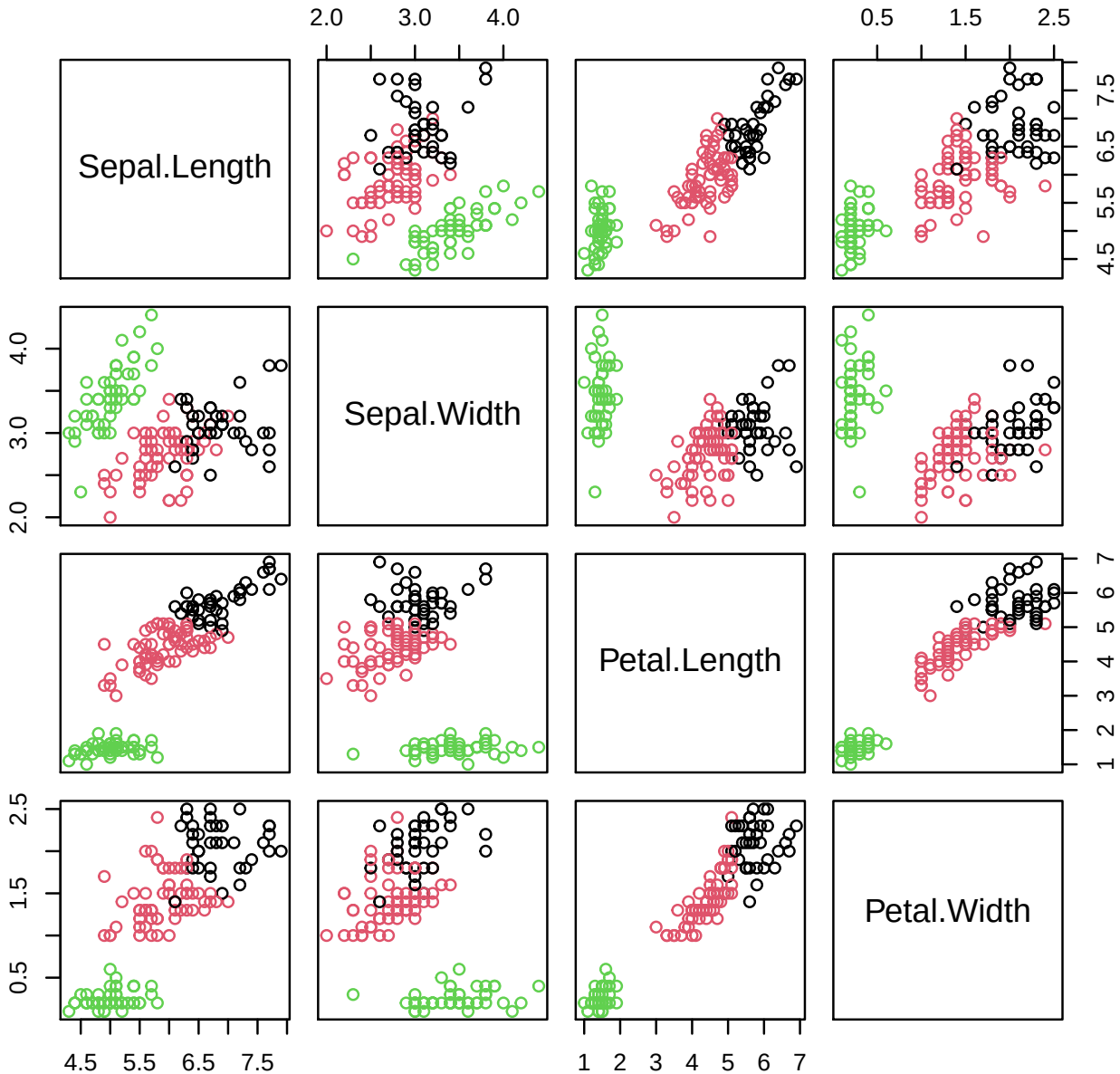
Dati `iris`

Le quattro variabili considerate sono la lunghezza e la larghezza del sepalo e del petalo

Table I

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8

Dati `iris`


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
2.0
3.0
4.0
0.5
1.5
2.5
4.5
5.5
6.5
7.5
1
2
3
4
5
6
7

Dati `iris`

L'analisi di raggruppamento fornisce circa il 90% di osservazioni classificate correttamente:

	setosa	versicolor	virginica
gruppo A	0	2	36
grippo B	0	48	14
gruppo C	50	0	0

Dati `movielens`

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

L'esempio che segue considera $n = 50$ film e $p = 139$ utenti

	U8	U15	U17	U19	U20	U21	U22	U23	U2
Ace Ventura		2.0		3.0	1.0	3.0		2.0	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.
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.
Fight Club	4.0	5.0	5.0		0.5		4.0	3.5	4.
Forrest Gump	4.0	1.0	2.5	5.0	2.0	4.0	3.5	4.5	4.
Fugitive, The	4.5	5.0		4.0	4.5	3.0	4.5	3.5	3.
Gladiator	5.0	2.0	4.0				3.0	4.0	2.
Godfather, The	5.0	5.0	5.0	5.0	2.0	4.0	4.0	5.0	4.
Good Will Hunting	4.0	4.0	4.0					3.5	
⋮									

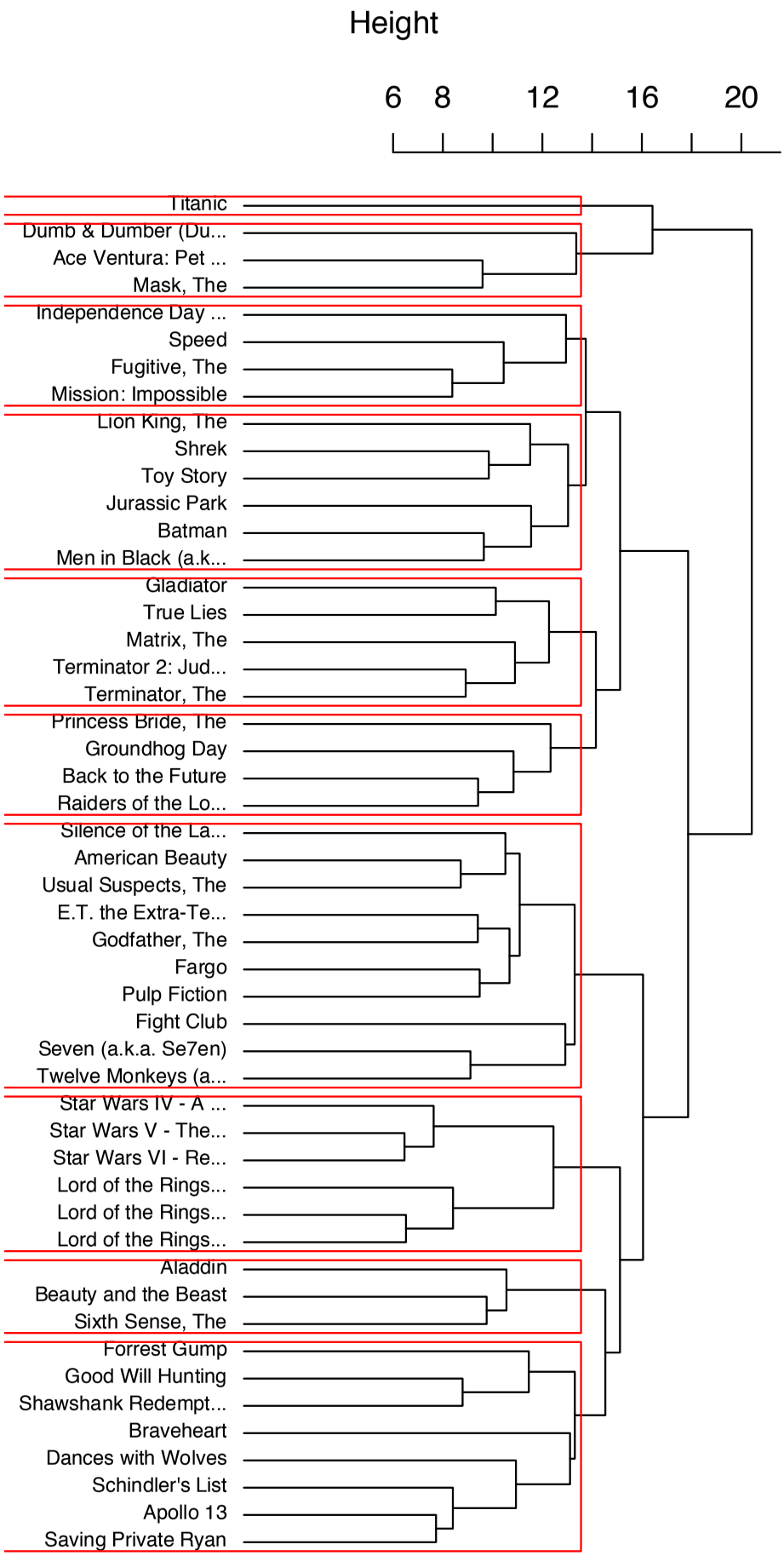
Obiettivo

Uno delle sfide da affrontare è il problema dei valori mancanti (missing values). Cosa fare quando il nostro dataset presenta dei buchi?

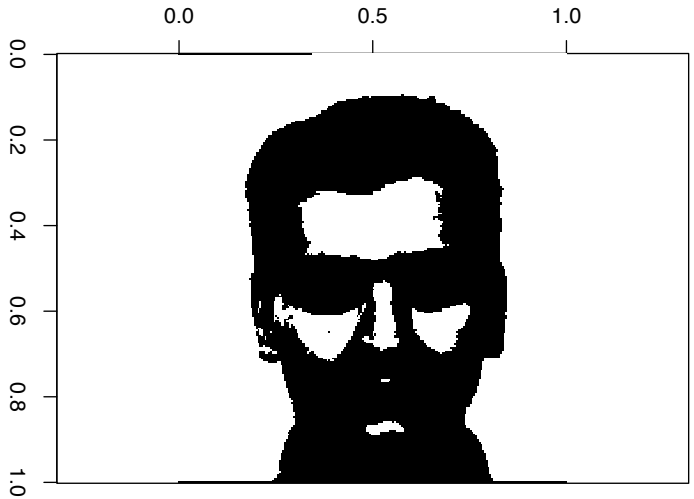
Una volta affrontato il problema dei dati mancanti, si può procedere raggruppando i film in gruppi omogenei al loro interno e disomogenei tra di loro rispetto al *rating* che hanno ottenuto dagli utenti

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

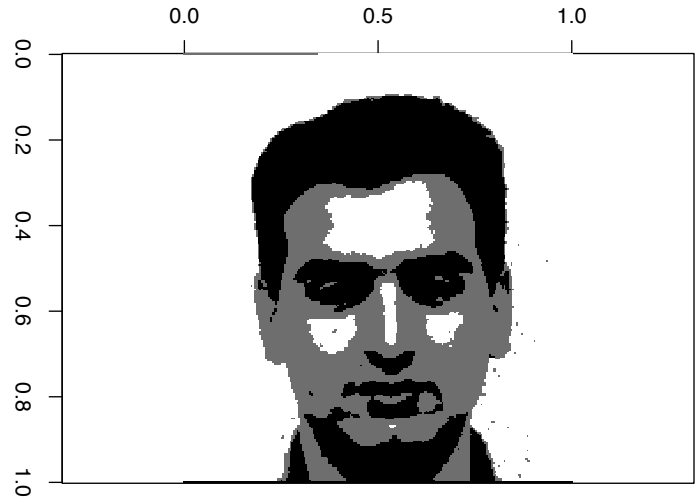
$$\underset{50\times 139}{X} \mapsto \underset{50\times 1}{y} = \begin{bmatrix} \mathrm{B} \\ \mathrm{A} \\ \dots \\ \mathrm{A} \\ \dots \\ \mathrm{C} \\ \mathrm{D} \end{bmatrix}$$


Height
6
8
12
16
20
Titanic
Dumb & Dumber (Du...
Ace Ventura: Pet ...
Mask, The
Independence Day ...
Speed
Fugitive, The
Mission: Impossible
Lion King, The
Shrek
Toy Story
Jurassic Park
Batman
Men in Black (a.k...
Gladiator
True Lies
Matrix, The
Terminator 2: Jud...
Terminator, The
Princess Bride, The
Groundhog Day
Back to the Future
Raiders of the Lo...
Silence of the La...
American Beauty
Usual Suspects, The
E.T. the Extra-Te...
Godfather, The
Fargo
Pulp Fiction
Fight Club
Seven (a.k.a. Se7en)
Twelve Monkeys (a...
Star Wars IV - A ...
Star Wars V - The...
Star Wars VI - Re...
Lord of the Rings...
Lord of the Rings...
Lord of the Rings...
Aladdin
Beauty and the Beast
Sixth Sense, The
Forrest Gump
Good Will Hunting
Shawshank Redempt...
Braveheart
Dances with Wolves
Schindler's List
Apollo 13
Saving Private Ryan
hclust (*, "complete")

Vector quantization



$k = 2$



$k = 3$