L'analisi multivariata

Analisi Esplorativa

Aldo Solari

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

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 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*)
3. Raggruppamento delle unità statistiche (*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

Perdere meno informazione possibile

Riduzione della dimensionalità: dati heptathlon

L'eptathlon (anche eptatlon o heptathlon) è una specialità dell'atletica leggera che contempla 7 gare di discipline diverse.

Nella prima giornata dell'eptathlon femminile outdoor si svolgono:

- 100 metri ostacoli,
- salto in alto,
- getto del peso,
- 200 metri piani.

Nella seconda:

- salto in lungo,
- tiro del giavellotto,
- 800 metri piani.

L'eptathlon fu inserito nel programma olimpico a partire dalle Olimpiadi di Los Angeles del 1984, in sostituzione del pentathlon.

Olimpiadi di Seul del 1988

In the 1988 Olympics held in Seoul, the heptathlon was won by one of the stars of women's athletics in the USA, Jackie Joyner-Kersee. The results for all 25 competitors are given here:

Dati

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
Joyner-Kersee (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
Sablovskaite (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
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

Domanda di interesse

La domanda di interesse è determinare un punteggio da attribuire a ciascun atleta che sintetizzi le *performances* nelle sette gare al fine di ottenere la classifica finale.

Vogliamo ridurre la dimensionalità $p = 7$ a $q = 1$:

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

Punteggio attribuito dalla manifestazione

	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

Riduzione della dimensionalità: 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 $q = 10$

Pixels e bytes

Immagine originale

- $\underset{243\times220}{X}$: $243 \times 220 = 53460$ pixels
- Memoria richiesta: 427880 bytes

Immagine compressa

- $\underset{243\times10}{Y}$, $\underset{220\times10}{V}$, $\underset{220\times1}{\bar{x}}$: $243 \times 10 + 220 \times 10 + 220 = 4850$ pixels
- Memoria richiesta: 40872 bytes
- Fattore di riduzione = 427880 bytes / 40872 bytes = 10.47

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 $y_i \in \{1, 2, \dots, k\}$,

dove $1, 2, \dots, k$ rappresenta il primo, ..., il k-simo gruppo

Obiettivo

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

Classificazione delle unità: dati Whisky

A 86 whisky di malto prodotti in Scozia è stato assegnato un punteggio da 0 a 4 su 12 categorie

- Body
- Sweetness
- Smoky
- Medicinal
- Tobacco
- Honey
- Spicy
- Winey
- Nutty
- Malty
- Fruity
- Floral

Inoltre è disponibile la latitudine e la longitudine delle distillerie.

Dati

	Distillery	Latitude	Longitude	Body	Sweet	Smoky	Medicinal	Tobacco	Honey	...
1	Aberfeldy	286580	749680	2	2	2	0	0	2	...
2	Aberlour	326340	842570	3	3	1	0	0	4	...
3	AnCnoc	352960	839320	1	3	2	0	0	2	...
4	Ardbeg	141560	646220	4	1	4	4	0	0	...
5	Ardmore	355350	829140	2	2	2	0	0	1	...
6	ArranIsleOf	194050	649950	2	3	1	1	0	1	...
7	Auchentoshan	247670	672610	0	2	0	0	0	1	...
.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	
85	Tormore	315180	834960	2	2	1	0	0	1	...
86	Tullibardine	289690	708850	2	3	0	0	1	0	...

Domanda di interesse

La domanda di interesse è raggruppare le diverse distillerie in k gruppi omogenei al loro interno e disomogenei tra di loro relativamente alle 12 categorie.

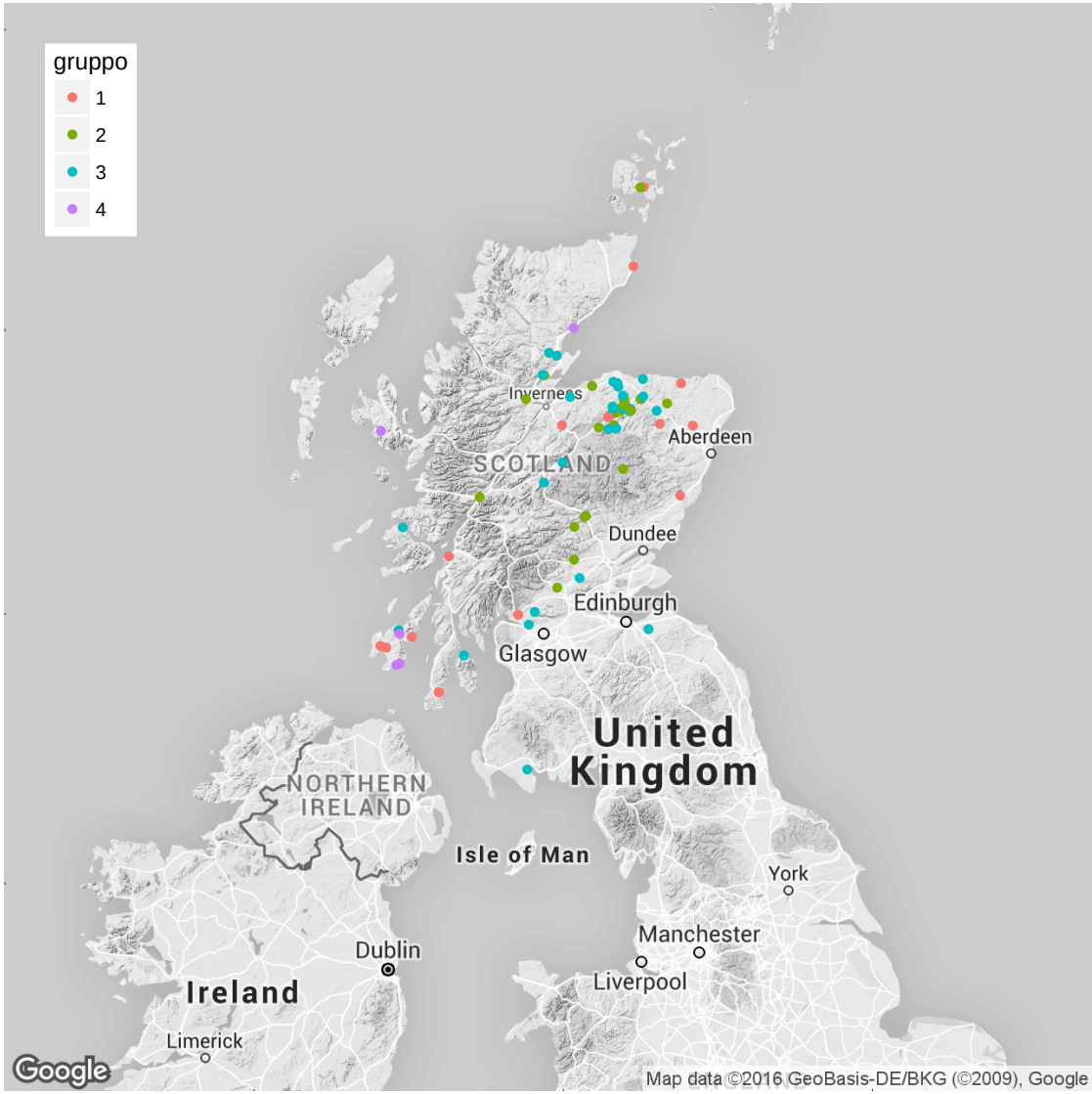
Ad esempio, se decidiamo di raggruppare le $n = 86$ osservazioni in $k = 4$ gruppi A, B, C, D

$$\underset{86\times 12}{X} \mapsto \underset{86\times 1}{y} = \begin{bmatrix} B \\ A \\ \dots \\ A \\ \dots \\ D \\ C \end{bmatrix}$$

Gruppo D

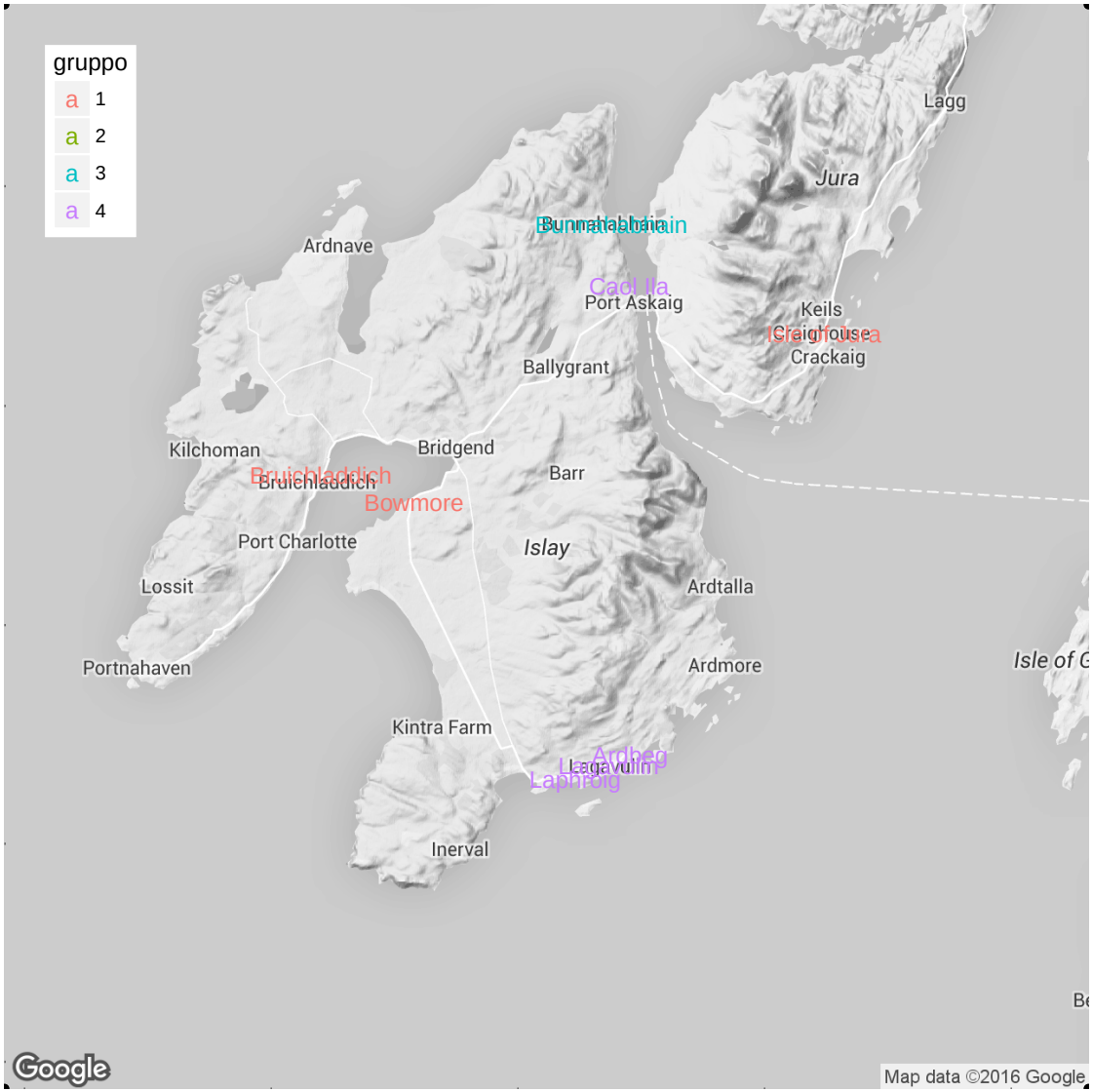
	Distillery	Body	Sweetness	Smoky	Medicinal	Tobacco	Honey	...
4	Ardbeg	4	1	4	4	0	0	...
22	Caol Ila	3	1	4	2	1	0	...
24	Clynelish	3	2	3	3	1	0	...
58	Lagavulin	4	1	4	4	1	0	...
59	Laphroig	4	2	4	4	1	0	...
78	Talisker	4	2	3	3	0	1	...

Scozia



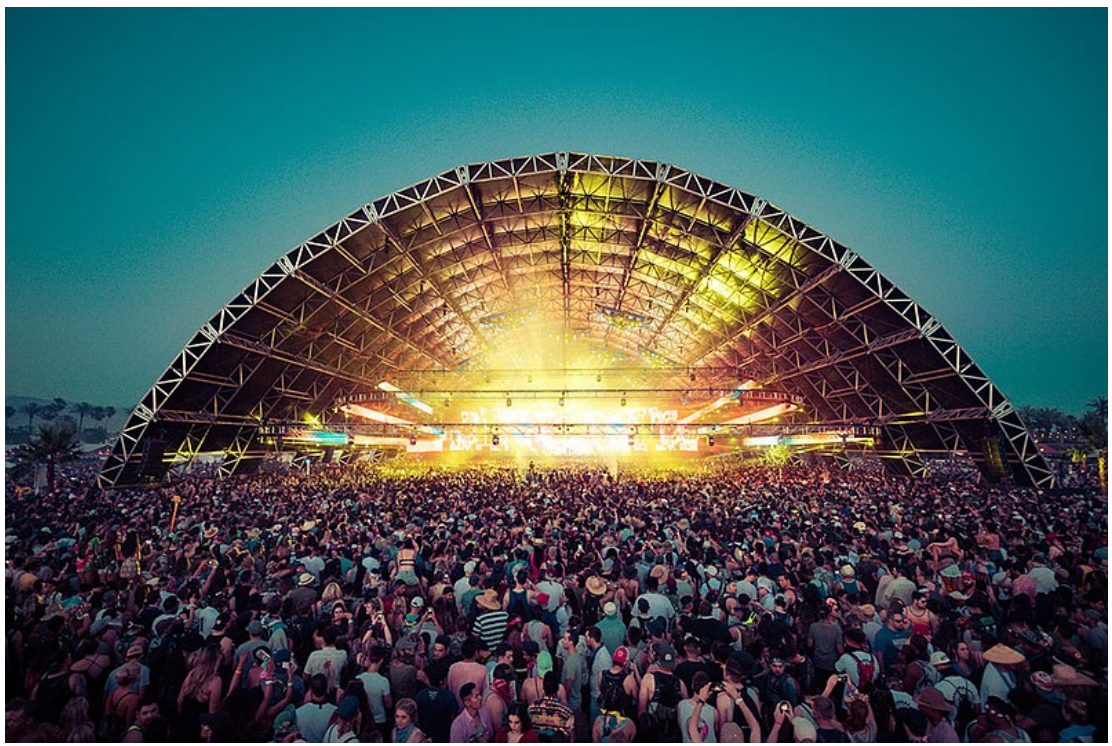
Gruppi ottenuti rispetto alle coordinate geografiche.

Islay


gruppo
a 1
a 2
a 3
a 4
Lagg
Jura
Ardnave
Port Askaig
Keils
Crackaig
Ballygrant
Kilchoman
Bridgend
Barr
Bowmore
Port Charlotte
Islay
Lossit
Ardtalla
Portnahaven
Ardmore
Kintra Farm
Laphroig
Inerval
Google
Map data ©2016 Google

Coachella festival

Il *Coachella Music and Arts Festival* è un festival musicale che si svolge annualmente nell'arco di due o tre giorni intorno alla fine di aprile negli Stati Uniti d'America, negli Empire Polo Fields di Indio in California



Edizione 2017

Osservazioni: artisti

Radiohead, Lady Gaga, Iggy Azalea, Kendrick Lamar, The xx, Travis Scott, Father John Misty, Empire of the Sun, Dillon Francis, Mac Miller, Ariana Grande, Bon Iver, Future, DJ Snake, Martin Garrix, ScHoolboy Q, Gucci Mane, Two Door Cinema Club, Lorde, Victoria Justice, New Order, Dreamcar, Porter Robinson & Madeon, Future Islands, Hans Zimmer, PNL e DJ Khaled, Vanessa Hudgens, Bastille, etc.

Variabili: Spotify track audio features

danceability, *energy*, *loudness*, *speechiness*, *acousticness*, *instrumentalness*, *liveness*, *valence*, *tempo*, *duration*

Analisi

Nell'analisi svolta da RCharlie, vengono identificati 3 gruppi di artisti:

- Hip-hop/Rock
- EDM/Experimental
- Alternative/Acoustic

Risultati dell'analisi:

- Exploratory plot
- Cluster analysis + Principal Component Analysis
- 3d plot