Slide Set 1
Nozioni Introduttive

Pietro Coretto
pcoretto@unisa.it

Corso di
Analisi e Visualizzazione dei Dati (Parte I)

Corso di Laurea in "Statistica per i Big Data" (L-41)
Università degli Studi di Salerno

Versione: 22 febbraio 2022 (h10:14)

Notes

Definizioni primitive

Dato
dal latino *datum* (dono, cosa data). È una descrizione elementare, generalmente codificata, dello stato di un oggetto fisico o astratto.

Informazione
Dal latino *informatio(-nis)* ("dare forma alla mente", "insegnare"). Si ottiene dal *significato* generato mettendo in relazione i *dati*.

Statistica
Branca della matematica (moderna) che si occupa della raccolta dei dati, e dell'estrazione dell'informazione

Informatica
Scienza che si occupa del trattamento dell'informazione mediante procedure automatizzate

Notes

Popolazione

Insieme dei possibili eventi oggetto di studio. Si tratta spesso di insiemi astratti e di natura infinita. Esempi:

- insieme dei possibili colori di un pixel in una una foto digitale
- insieme dei possibili livelli del consumo nazionale
- insieme delle possibilità per il numero di esami superati da uno studente di UniSA

Campione

Un insieme finito di casi osservati dell'oggetto di studio. I *dati sono la descrizione dei casi osservati*, e vengono usati per estrarre informazione sulla popolazione. Esempi:

- su 4 studenti di UniSA rilevo il numero di esami superati
- misuro il consumo di beni in Italia dal 1980 ad oggi
- In una foto digitale scelgo 10 pixels e rilevo l'RGB

⚠ Popolazione e campione non vanno confusi con l'insieme delle unità statistiche (vedi dopo)

Notes

Modello

Astrazione, mediante il linguaggio matematico, del funzionamento della popolazione.

Esempio 1: sia $C_t =$ consumo di beni nell'anno t, $Y_t =$ reddito nell'anno t, $R_t =$ risparmio nell'anno t, si consideri il *modello*

$$C_t = \alpha + \beta Y_t + \gamma R_{t-1} + \ldots, + \text{altri fattori}$$

Esempio 2: $C_t = f(Y_t, R_{t-1}, \ldots)$, dove f è una funziona continua con derivate parziali positive

⚠ Qualunque "*preconcetto*" sul funzionamento della popolazione è un modello. La sorgente di errore più comune è quello di analizzare i dati avendo un modello in mente senza saperlo.

I modelli vengono usati entro due paradigmi scientifici

- Predittivo
- Esplicativo

Notes

Campionamento
Metodi per decidere come scegliere i campioni in modo ottimale in funzione dell'informazione che si desidera estrarre

Inferenza
È un insieme di metodi *induttivi* attraverso i quali si usa l'informazione campionaria per *generalizzare* le caratteristiche della popolazione.
L'inferenza si basa quasi sempre sulla formulazione di un *modello* (spesso anche quando non sappiamo di averne uno)

Analisi esplorativa dei dati (EDA)
Metodi numerici e grafici per estrarre informazione dai dati a prescindere dalla formalizzazione di modelli. L'EDA, nella maggior parte, dei casi precede la formalizzazione di modelli

Notes

John W. Tukey (1977)
"*Exploratory Data Analysis*", Addison-Wesley

Notes

Data set

Indicatori di Bilancio
Imprese Italiane Settore Food & Beverage

#	INCORPORATION_DATE	STATUS	NUTS1	ROE	EMPL	ROA	CREDIT_DAYS	RISK
1	1961-01-27	Active	ITC - Northwest	3.21	26,234	1.99	46	H
2	1988-10-12	Active	ITH - Northeast	13.62	4,130	9.46	71	H
3	1983-10-17	Active	ITH - Northeast	0.32	6,842	0.02	30	M
4	1948-12-03	Active	ITC - Northwest	2.07	2,957	1.38	55	H
5	1988-02-24	Active	ITH - Northeast	3.15	5,240	1.22	42	H
6	1982-05-04	Active	ITC - Northwest	18.35	730	8.07	45	L
7	1980-11-06	Active	ITC - Northwest	12.83	48	10.48	1	M
8	1998-11-27	Unknown	ITI - Centre	-0.36	**NA**	-0.23	54	L
9	1988-03-09	Active	ITH - Northeast	4.21	2,899	0.93	98	M
10	1913-12-04	Active	ITC - Northwest	9.08	3,485	1.90	95	M
11	1970-10-12	Active	ITH - Northeast	5.96	3,305	2.25	52	M
12	1998-01-29	Active	ITC - Northwest	10.35	1,895	6.40	38	H
13	***NA***	liquidation	***NA***	***NaN***	18,081	-42.68	83	H
14	1961-10-06	Active	ITC - Northwest	14.22	1,076	9.95	50	L
15	2002-01-25	Active	ITC - Northwest	10.15	1,664	7.08	36	H
16	1961-04-15	Dissolved	ITH - Northeast	6.77	2,415	1.70	67	H
17	1976-05-24	Active	ITH - Northeast	0.18	2,772	0.05	101	H
18	1951-04-10	Active	ITC - Northwest	11.64	820	8.17	56	L
19	1977-12-13	Active	ITH - Northeast	21.02	1,077	9.93	48	M
20	1991-07-18	Active	ITH - Northeast	4.21	170	0.86	70	M

Notes

Unità campionarie e variabili

Unità statistica / Sample unit / Sample point

- è l'elemento reale/astratto sul quale rileviamo i dati
  Es: l'impresa XYZ selezionata per la rilevazione
- sia $i = 1, 2, \ldots, n$ un intero usato per indicizzare le unità statistiche
  Es: l'azienda $i = 10$ riporta `STATUS`=`'active'`, `ROE`= 9.08, etc.

Variabili / Features

- fenomeno/caratteristica di interessse
  Es: `STATUS`=stato economico/giuridico dell'azienda, `ROE`=redditività del capitale proprio, etc
- solitamente si misurano più variabili/features su ogni unità statistica
  Sia X_k la k-ma variabile, per $k = 1, 2, \ldots, K$
  Es. dati di bilancio: $X_1 =$ `INCORPORATION DATE`, $X_2 =$ `STATUS` $\ldots X_4 =$ `ROE`

Notes

Espressione di una variabile

Livello

Quando una variabile ha una rappresentazione *ordinabile*, si chiama *livello* la sua espressione

- Es: i livelli di X_4=`ROE` sono -0.36, 0.18, 0.32,...
- Es: il livelli di X_8=`RISK` sono `L`="low", `M`="medium", `H`="high"
- X_7=`CREDIT_DAYS` ha un *range* di livelli [1, 101]

Label

Quando una variabile non ha una rappresentazione numerica, la sua espressione ha la forma di un attributo. Quest'ultimo si chiama *label*

- Es: i labels di X_3 = `NUTS1` sono `ITC-Northwest, ITH-NOrtheast`, ...
- Es: i labels di X_2=`STATUS` sono `Active`, `Liquidation`, ...

Notes

Elementi di un data set

Data set univariato

Si osserva una sola variabile, sia essa X. Tipicamente indichiamo con x_i l'espressione del livello/label di X sulla i-ma unità campionaria

Es: si rileva X="altezza in cm", x_2=altezza espressa in cm misurata sul individuo $i = 2$ del campione

Data set multivariato

Si misurano K variabili $(X_1, X_2, \ldots, X_k)$ su ciascun individuo del campione. Sia x_{ik} l'espressione del livello/label della variabile X_k rilevato sull'i-ma unità campionaria

Es. Indicatori di Bilancio: si misurano $K = 8$ variabili su $n = 20$ imprese.
In questo caso X_1 = `INCORPORATION DATE`, X_2 = `STATUS` ....
In questo caso $x_{14,2}$="`active`" è il label dello `STATUS` per l'impresa $i = 14$ nel campione

Notes

Classificazione: variabili vs espressione

Quantitative (numeriche)

I livelli sono numerici, ordinabili, ed hanno un'interpretazione "*metrica*".
Una variabile numerica può essere

- discreta: i livelli sono numeri interi (Es: X_7= `CREDIT_DAYS`)
- continua: i livelli sono numeri reali (Es: X_4=`ROE`)

Ordinali

I livelli possono essere alfanumerici, e sono ordinabili, tuttavia non hanno un'interpretazione "*metrica*".

- Es 1: X_8 = `RISK`, si puo' dire che: `L` precede `M`, che a sua volta precede `H`. Tuttavia possiamo dire se `H` − `M` sia più o meno grande di `M` − `L`? Cosa significa $2 \times$ `L`?
- Es 2: anche ricorrendo alla *codifica* { `L`=1, `M`=2, `H`=3}, non cambierebbe niente.

Notes

Categoriali (nominali)

I livelli possono essere alfanumerici, non sono ordinabili, e non hanno un'interpretazione "*metrica*".

- Es 1: X_3 = `NUST1`, si puo' dire che: `ITC-Northwest` precede `ITH-Northeast`?
- Es 2: X_2 = `STATUS` la differenza `Active` - `Dissolved` avrebbe un significato?

Notes

Classificazione rispetto al ruolo delle unità statistiche

Cross-section

Si misura l'espressione di una o più variabili su un insieme di unità di rilevazione nello stesso instante temporale. Consentono di catturare variazioni statiche, ovvero variazioni tra unità nello stesso instante temporale.

Es: si rilevano le variabili (R_i, C_i, I_i)=(Reddito, Consumo, Investimenti) riferite all'individuo i al 2019, fissato $i = 6$, c_6 =consumi dell'individuo $i = 6$ al 2019.

Nota bene: gli indici delle unità statistiche si possono *scambiare*, senza produrre alterazione dell'informazione campionaria. Infatti, nell'esempio precedente scambiando (r_6, c_6, i_6) con $(r_{100}, c_{100}, i_{100})$, l'informazione sullo stato dell'economia al 2019 non cambierebbe

Notes

Serie Storiche

Si misura l'espressione di una o più variabili su una data unità in istanti temporali successivi. Consentono di catturare variazioni dinamiche, ovvero variazioni intervenute sulla stessa unità per effetto del trascorrere del tempo.

Es: si rilevano le variabili (R_t, C_t, I_t)=(Reddito, Consumo, Investimenti) per l'economia italiana misurati nell'anno t.

Nota bene: gli indici temporali delle unità statistiche non si possono *scambiare*, altrimenti vi sarebbe la perdita di informazione campionaria. Infatti, nell'esempio precedente scambiando $(r_{1900}, c_{1900}, i_{1900})$ con $(r_{2000}, c_{2000}, i_{2000})$, l'informazione sulle trasformazioni dello stato dell'economia tra il 1900 ed il 2000 andrebbe completamente persa

Notes

Serie Spaziali

Si misura l'espressione di una o più variabili su una data unità in locazioni diverse allo stesso instante temporale. Consentono di catturare variazioni spaziali, ovvero variazioni intervenute sulla stessa unità per effetto della dislocazione geografica

Es: si rilevano le variabili $(P_{\text{lat,lon}}, T_{\text{lat,lon}}, R_{\text{lat,lon}})$=(Pressione, Temperatura, Pioggia), dove $P_{\text{lat,lon}}$ è la pressione misurata alla longitudine=lon e latitudine=lat.

Nota bene: gli indici spaziali [nell'esempio (lat,lon)] delle unità statistiche non si possono *scambiare*, altrimenti vi sarebbe la perdita di informazione campionaria

Notes

Serie spazio-temporali

Si misura l'espressione di una o più variabili in locazioni diverse ed in instanti temporali diversi. Consentono di catturare contemporaneamente variazioni

- spaziali, dovute alla dislocazione geografica
- dinamiche, dovute al trascorrere del tempo

Es: si rileva $P_{\text{lat,lon,t}}$, ovvero è la pressione misurata alla longitudine=lon, latitudine=lat, e riferita all'istante t

Nota bene: gli indici spazio-temporali [nell'esempio (lat,lon,t)] non si possono *scambiare*, altrimenti vi sarebbe la perdita di informazione campionaria

Notes

Dati panel / longitudinali

Si misura l'espressione di una o più variabili sullo stesso insieme di unità di rilevazione in istanti temporali successivi. Consentono di catturare contemporaneamente variazioni

- statiche: tra individui allo stesso istante
- dinamiche: nel tempo riferite ad uno stesso individuo

Es: si rilevano le variabili $(R_{i,t}, C_{i,t}, I_{i,t})$=(Reddito, Consumo, Investimenti) misurati su $n = 100$ individui tra il 1980 ed il 2019. $r_{6,1985} =$ reddito dell'individuo $i = 6$ al 1985, $r_{6,2000} =$ reddito dello stesso individuo $i = 6$ al 2000

Nota bene: lo scambio degli indici *temporali* delle unità statistiche produrrebbe una notevole perdita di informazione.

Notes

Campionamento vs fattori di disturbo

Campioni sperimentali

Il campionamento è effettuato in condizioni piuttosto controllate, in tal modo è possibile eliminare fattori di *disturbo* per l'analisi. Questo tipo di studi è necessario quando si vogliono stabilire relazioni di causa-effetto

Es: si somministra una medicina a 10 topi sani scelti a caso, e 10 topi scelti a caso tra quelli affetti da una certa malattia. Durante il periodo di osservazione i topi sono tenuti tutti nelle stesse condizioni

Campioni osservazionali

Il campionamento è effettuato in condizioni poco controllate. Le variabili di interesse vengono misurate senza che lo sperimentatore possa interagire con le unità. Non è possibile eliminare tutti i fattori di *disturbo*. È difficile poter trovare relazioni di causa-effetto

Es: si scelgono a caso n individui e si misurano una serie di variabili di salute fisica, ed il consumo di farmaci per trattamenti cardiovascolari.

Notes

Componenti dell'EDA

- Data pre-processing: formattazione del data set
  Talvolta richiede abilità informatiche piuttosto sofisticate
- Caratterizzazione della distribuzione dei dati
  (tendenza centrale, dispersione, etc)
- Individuazione di dati contaminati
  (outliers, robustezza)
- Ricerca di *patterns*
  (correlazione, regressione, etc)
- Riduzione della dimensionalità
  (componenti principali, MDS, etc.)
- Riduzione della eterogeneità
  (classificazione, clustering, etc.)
- . . . etc.

Notes

Notes