STATISTICA 2

AA 2017/18

Prof. Marco Romito

marco.romito@unipi.it

http://people.dm.unipi.it/romito

→ DIDATTICA

→ STAT 2 2018

PROGRAMMA DEL CORSO

1. Regressione lineare
2. Analisi delle Componenti principali
3. Problemi di classificazione e clustering

\+ Implementazione di tali argomenti su un software statistico (GNU **R**)

R STUDIO

4. [redacted]

MATERIALE DIDATTICO : dispense del corso (DA INTERNET)
libro di Frascella C.

MODALITÀ DI SVOLGIMENTO ESAME

(A) Scritto + Orale (nello stesso appello)

(B) Progetto (da realizzare durante il corso) + Orale (nel primo appello o nella data previste preappello)
- ↳ Reperire una serie di dati
- ↳ Esaminarsi
- ↳ Trarre le conclusioni
- ↳ Breve relazione

SVOLGERE SINGOLARMENTE

Ricevimento → per ora su appuntamento / LUN 14 - 16

Controllare nella pagina del corso negli anni precedenti

S. Ross Probabilità e Statistica per l'ingegneria e le scienze

NB La verbalizzazione dell'esame avviene con il superamento di R02 (ENTRO UN A.A.)

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

master COPY COPISTERIA
050/8312126 388/9837745
MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

① REGRESSIONE LINEARE

27-09-17

Supponiamo di avere un campione, una popolazione di una serie di individui, vogliamo capire se ci sono delle relazioni e/o dipendenze tra le caratteristiche degli individui.
Analizzando grandezze fisiche trovare dipendenze è "più semplice", mentre con le attività umane c'è molta più variabilità nei dati che si osservano e le conseguenze sono meno ovvie.
Capire le dipendenze funzionali consente di individuare quali sono i fattori preponderanti e soprattutto mostrare quantitativamente l'impatto di tali fattori.

Consideriamo innanzitutto 2 GRANDEZZE e la DIPENDENZA LINEARE

Richiami

X, Y GRANDEZZE (variabili aleatorie)

$\mathrm{cov}(X,Y) = \mathbb{E}\left[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right]$ COVARIANZA ($\mathbb{E}$ VALORE ATTESO) SI CALCOLANO LE VARIAZIONI CONGIUNTE DEI DATI RISPETTO AI VALORI MEDI

$\rho(X,Y) = \dfrac{\mathrm{cov}(X,Y)}{\mathrm{sd}(X)\,\mathrm{sd}(Y)}$ COEFFICIENTE DI CORRELAZIONE (MISURA LE MUTUE VARIABILITÀ)

$\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$ DEVIAZIONE STANDARD (σ^2) INDICE DI DISPERSIONE OVVERO SUNA DELLA VARIABILITÀ

$\mathrm{Var}(X) = \mathrm{Cov}(X,X)$ VARIANZA MISURA DELLA VARIABILITÀ (cov generalizza var)

PROPRIETÀ:

1. $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$ SIMMETRIA
2. $\mathrm{cov}(aX, bY) = ab\,\mathrm{cov}(X,Y)$
3. $\mathrm{cov}(aX+bY+c, Z) = a\,\mathrm{cov}(X,Z) + b\,\mathrm{cov}(Y,Z)$
4. $\rho(aX, bY) = \rho(X,Y)$ ρ NON DIPENDE DALL'UNITÀ DI MISURA DEI DATI
5. $|\rho(X,Y)| \le 1$
6. $\rho(X,Y) = \pm 1$ solo se (e viceversa) $Y = aX+b$ $\rho = 1 \Leftrightarrow$ REL. LIN. TRA X E Y
   inoltre $a, \rho(X,Y)$ sono concordi
7. X, Y indipendenti, allora $\mathrm{cov}(X,Y) = \rho(X,Y) = 0$ (ma non il viceversa) discende da $\mathbb{E}[XY] = \mu_X \mu_Y$

INDIPENDENZA IMPLICA SCORRELAZIONE ←

⇨ A livello empirico supponiamo di avere una serie di dati in tabella

x_1	y_1
⋮	⋮
x_n	y_n

RIGHE = fattori relativi al singolo individuo
COLONNE = fattori presi in esame

STIME DEGLI INDICI SOPRA DEFINITI

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

$$\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$$

$$\sigma_x^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2 \qquad \sigma_y^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k-\bar{y})^2$$

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}$$

MASTER COPY Tel. 050 8312126 Cell. 388 9837745
FABER-CASTELL

Il coefficiente di correlazione consente di valutare QUANTITATIVAMENTE la dipendenza tra le grandezze.

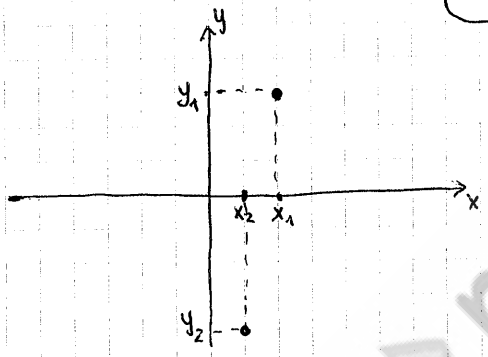
CASI LIMITE → $\rho = \pm 1$ DIPENDENZA LINEARE
↘ $\rho = 0$ NESSUNA DIPENDENZA

IN REALTÀ CIÒ NON SUCCEDE QUASI MAI! MA PIUTTOSTO SI È VICINO A 1 OPPURE VICINO A 0

se molto vicino a 0 indipendenza o meglio scorrelazione

↳ Allora come capire se correlazione alta oppure correlazione bassa ?
↳ Con l'esperienza e le prove empiriche!

NOTA: sito guessthecorrelation.com UTILE PER PROVARE A DEFINIRE LA CORRELAZIONE DI UNA SERIE DI DATI PROPOSTI

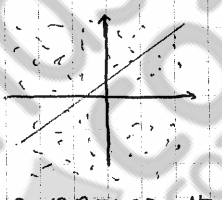


(1) INDIPENDENZA o BASSA DIPENDENZA

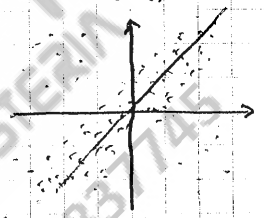
(2) ELEVATO LEGAME POSITIVO

(3) ELEVATO LEGAME NEGATIVO

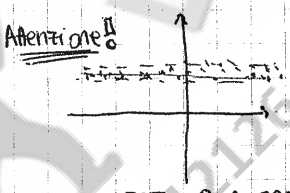
Diagramma di dispersione
↳ RAPPRESENTAZIONE GRAFICA DEI DATI PRESENTI IN TABELLA



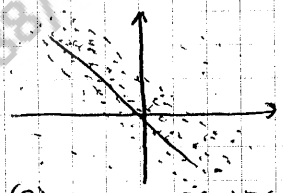
(1) CORRELAZIONE BASSA



(2) CORR. ALTA POSITIVA



DIFETTO PER ORDINI DI GRANDEZZA TROPPO DIFFERENTI



(3) CORR. ALTA NEGATIVA

Definiamo (X) FATTORE DI INPUT → dato in ingresso ↝ si PUÒ CONTROLLARE

(Y) FATTORE DI OUTPUT → dato che si vuole studiare

Vogliamo capire come il fattore di input influenza il fattore di output.

Modello di Regressione

$Y = aX + b + \epsilon$ DIPENDENZA LINEARE (considerando anche un errore ϵ)

È importante cercare di definire bene i termini **a** e **b** e l'errore in modo che il modello possa essere anche PREDITTIVO.

Ipotizziamo un legame lineare perturbato da un'azione casuale

FABER-CASTELL

Vediamo come definire i termini della relazione scritta.

MODELLO TEORICO

Supponiamo
- $\mathbb{E}(\epsilon) = 0$ errore in media nullo → {IPOTESI CHE VA BENE PERCHÉ SE CI FOSSERO VARIAZIONI NOTEVOLI SAREBBERO COMPRESI IN b}
- $\text{cov}(X, \epsilon) = 0$ supposizione forte perché così diciamo che il modello è perfetto → {SE C'È UNA PARTE DI X DENTRO L'ERRORE SIGNIFICA CHE NON È STATO COMPRESO BENE IL FATTORE X!}

L'errore è importante perché non è detto che si riescano a comprendere SEMPRE tutti i fattori che incidono sul fenomeno.

TROVARE a E b

$$\text{Cov}(X,Y) = \text{cov}(X, aX+b+\epsilon) = a\,\text{Var}(X) + \underbrace{\text{cov}(X,\epsilon)}_{=0 \text{ per ipotesi}} = a\,\text{Var}(X)$$

$$\Rightarrow \boxed{a = \frac{\text{cov}(X,Y)}{\text{Var}(X)}}$$

$$\mathbb{E}(Y) = \mathbb{E}(aX+b+\epsilon) = a\,\mathbb{E}(X) + b + \underbrace{\mathbb{E}(\epsilon)}_{=0 \text{ per ipotesi}} = a\,\mathbb{E}(X) + b$$

$$\Rightarrow \boxed{b = \mathbb{E}(Y) - a\,\mathbb{E}(X)}$$

Per valutare la bontà del modello bisogna analizzare la variabilità dell'errore.

TROVARE VARIANZA DI ϵ

$$\text{Var}(Y) = \text{Var}(aX+b+\epsilon) = a^2\,\text{Var}(X) + \text{Var}(\epsilon)$$

$$\Rightarrow \boxed{\text{Var}(\epsilon) = \text{Var}(Y) - a^2\,\text{Var}(X)}$$

Riprendiamo i dati scritti prima e consideriamo le **grandezze empiriche**:

$x_1 \; y_1$ … $x_n \; y_n$ $\qquad Y = aX + b + \epsilon$ $\qquad \hat{a} = \dfrac{\hat{\text{cov}}(X,Y)}{\sigma_x^2}$ $\qquad \hat{b} = \bar{y} - a\bar{x}$

$y = \hat{a}x + \hat{b}$ **RETTA DI REGRESSIONE**

Vediamo di studiare l'errore: $\sigma_\epsilon^2 = \sigma_y^2 - \hat{a}^2\sigma_x^2 = \sigma_y^2 - \dfrac{(\hat{\text{cov}}(X,Y))^2}{\sigma_x^4}\sigma_x^2 =$

VARIABILITÀ ERRORE = parte della variab. del fattore in uscita non spiegata dalla variabilità del fattore in ingresso

$$= \sigma_y^2 - \frac{(\hat{\text{cov}}(X,Y))^2}{\sigma_x^2}$$

↳ termine presente per la dipendenza tra x e y

$$\hat{\rho}(X,Y) = \frac{\hat{\text{cov}}(X,Y)}{\sigma_x \sigma_y}$$

$$\sigma_\epsilon^2 = \sigma_y^2\left(1 - \frac{(\hat{\text{cov}}(X,Y))^2}{\sigma_x^2\sigma_y^2}\right) = \sigma_y^2\left(1 - (\hat{\rho}(X,Y))^2\right)$$

→ CI DICE IN CHE MISURA IL COEFF. DI CORRELAZIONE STIMA CON PRECISIONE IL MODELLO, OVVERO QUANTO È PRECISO IL LEGAME LINEARE TRA X E Y.

$\hat{\rho}_{XY} \sim 0 \Rightarrow \sigma_\epsilon^2$ elevato $\Rightarrow$ CATTIVO LEGAME LINEARE

$\hat{\rho}_{XY} \sim \pm 1 \Rightarrow \sigma_\epsilon^2 \sim 0 \Rightarrow$ LEGAME LINEARE STRETTO

FABER-CASTELL