Statistica 2

Gli appunti sono formati da più parti: una prima parte sono tutte le lezioni di teoria, prese in aula nell'anno 2017-2018, una seconda parte parte di lezioni svolte sul software R, con tutti gli appunti presi in aula al pc, una terza parte con la soluzione di alcuni esami che ho svolto, di cui ovviamente non garantisco la correttezza, tranne di alcuni che ho riguardato a ricevimento, e che possono essere utili come base per come svolgere gli esercizi, e infine una parte di 10 fogli dove ho raccolto tutta la teoria per l'orale e in particolare ho svolto le dimostrazioni, necessarie da sapere, in maniera più lineare.




master
COPY
COPISTERIA
050/8312126 388/9837745

STATISTICA 2

27/9

MARCO ROMITO

MARCO.ROMITO @ UNIPI.IT

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

→ didattica
→ stat 2 2018

- diapositive corso
- Fioscello C. (libro)

Argomenti trattati:

1) regressione lineare
2) analisi delle componenti principali
3) classificazione e clustering
4) serie storiche

MASTER COPY
Tel. 050 [illegible]
Cell. 393 [illegible]

Importante del corso → applicazione della teoria alla pratica.

-Useremo un SOFTWARE STATISTICO: GNUR

ESAME: - Standard scritto/orale (stesso sess.)

- Pz da colorare progetto/orale
ev. darà dei compiti da fare:

progetto: - reperire dati
- esaminarli con i metodi dell'impatto
- stiliamo una BREVE relazione

progetto singolo!

giungere al risultato in modo diretto

dunque è utile farli

Nel progetto sono i dati al centro → le conclusioni devono essere basate sulla teoria fatta a lezione e devono essere robuste

Argomenti base: Algebra, Algebra lineare, analisi dati, analisi matematica

Esame di statistica è data x scontato

Consiglio: date la data comunque 1 sett dopo la fine del corso

→ A COMPUTER

installare il SOFTWARE R

→ RSTUDIO quello che usa il prof

REGRESSIONE LINEARE

obiettivo del corso → capire le dipendenze tra un campione di individui

Si ha un campione di individui con diverse caratteristiche. Può esserci dipendenza funzionale tra 2 grandezze (come modello quantitativo) → ci concentreremo sulla dipendenza lineare (senza tenere in causa le altre dipendenze) capire se c'è una relazione tra due grandezze

e nella fisica funziona sempre perché sono grandezze macroscopiche che mettiamo sempre nella media.

Quando si tratta di grandezze che hanno a che fare con l'uomo sono affette da grande variabilità.

Es. Andamento delle vendite

quali sono i fattori che intervengono nell'andamento?

Individuare i fattori è una cosa a cui studieremo in corsi specifici, poi una volta trovati bisogna capire quali sono quelli che contano quantitativamente → questo è lo scopo del corso

DIPENDENZA LINEARE → è la + semplice

cosa serve x fare questo tipo di analisi?

X, Y

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

ci dà una misura della dipendenza, ma è scarsa in come misuriamo X e Y la cov cambia in quella di ordini di grandezza + grande spesso il risultato quindi sarebbe bene rendere x e y essendo tramite:

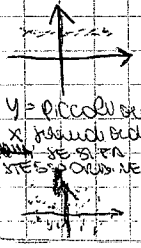
sd = deviazione standard

$$\rho(X,Y) = \frac{Cov(X,Y)}{sd(X)\ sd(Y)}$$

Var = varianza

$$sd(X) = \sqrt{Var(X)} \qquad sd(Y) = \sqrt{Var(Y)}$$

$$Var(X) = Cov(X,X)$$



ρ = coeff di relazione → misura le mutue variabilità (non ha unità di misura) ed è necessario che X e Y abbiano ordine di grand. paragonabili

Proprietà

1) $cov(x,y) = cov(y,x)$
2) $cov(ax, by) = ab\, cov(x,y)$
3) $cov(ax+by+c, z) = a\, cov(x,z) + b\, cov(y,z)$
4) $\rho(ax, by) = \rho(x,y)$
5) $|\rho(x,y)| \leq 1$
6) $\rho(x,y) = \pm 1 \iff y = ax+b$ } dipendenza lineare tra x e y
   → se a↑ allora y^A
   nostra a, $\rho(x,y)$ sono concordi
7) x, y sono indipendenti, allora
   $cov(x,y) = \rho(x,y) = 0$

A livello empirico sappiamo di avere una tabella di dati con due tabelle di numeri dove le righe rappresentano

x_1	y_1
⋮	⋮
x_m	y_m

cov > 0 sono entrambi > 0 < del valore atteso dipendenza così
cov < 0 succede quando uno cresce l'altro ↓ PENSO NON CI DA DELLE INFO

media: $\bar{x} = \frac{1}{m}\sum_{k=1}^{m} x_k$

$\bar{y} = \frac{1}{m}\sum_{k=1}^{m} x_k$

covarianza: $cov(x,y) = \frac{1}{m}\sum_{k=1}^{m} \frac{(x_k-\bar{x})(y_k-\bar{y})}{\sigma(x)\,\sigma(y)}$

varianza: $\sigma_x^2 = \frac{1}{m}\sum_{k=1}^{m}(x_k-\bar{x})^2$

$\sigma_y^2 = \frac{1}{m}\sum_{k=1}^{m}(y_k-\bar{y})^2$

coefficiente di correlazione: $\hat{\rho}(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$

↓ ↑ ha il cappello perché è un dato empirico e non esatto

è permette di misurare la dipendenza quantitativa tra x e y

se $\hat{\rho} \approx 1$ allora ci può essere una correlazione e vado a fondo

$\hat{\rho} \approx 0$ allora può non esserci corr.

MASTER COPY
Tel. 050 83121 26
Cell. 333 9937145
Questione correlation.com per Mara e di fare in occhio alla corrispondenza