Probabilità e Statistica per l'Informatica

UniShare

Davide Cozzi
@dlcgold

Gabriele De Rosa
@derogab

Federica Di Lauro
@f_dila

Indice

Capitolo 1

Introduzione

Questi appunti sono presi a lezione. Per quanto sia stata fatta una revisione è altamente probabile (praticamente certo) che possano contenere errori, sia di stampa che di vero e proprio contenuto. Per eventuali proposte di correzione effettuare una pull request. Link: `https://github.com/dlcgold/Appunti`.
Grazie mille e buono studio!

Capitolo 2

Breve Introduzione

Ormai i dati sono pervasivi e un loro studio è diventato necessario. Inoltre si parla spesso di target marketing, con una selezione dei possibili clienti. Inoltre la statistica è usata in maniera massiccia nel mondo dello shopping online. Si ha l'*A-B testing*, per decidere tra due scelte la migliore. Per scegliere si analizzano i dati presi da campioni di popolazione. Si ha il *tasso di conversione*. Un altro ambito importante è la robotica e la domotica. Ovviamente la statistica è importante anche nel mondo dei videogames. Anche genetica e biologia fanno uso abbondante dei dati statistici.
Si hanno i seguenti 8 argomenti:

1. statistica descrittiva
2. calcolo delle probabilità
3. distribuzioni notevoli
4. teoremi di convergenza
5. stima dei parametri
6. test di ipotesi parametrici
7. test di ipotesi non parametrici
8. regressione lineare

Capitolo 3

Statistica Descrittiva

La statistica descrittiva è una raccolta di metodi e strumenti matematici usati per organizzare una o più serie di dati al fine di trovare:

- simmetrie
- periodicità
- leggi varie

Si descrivono quindi le informazioni implicite ai dati. Solitamente la serie di dati di cui si dispone è costituita da un numero limitato di **osservazioni** che devono essere rappresentative di un'ampia **popolazione**, che è quindi l'insieme a cui si riferisce l'indagine statistica. Un **campione** è un sottoinsieme selezionato della popolazione, che viene analizzato per dire qualcosa sulla popolazione da cui è stato prelevato. Non esiste un solo campione di una popolazione ma si hanno molti modi diversi di scegliere un campione. Si vuole affermare qualcosa riguardo i **caratteri/caratteristiche** della popolazione. Si hanno:

- **caratteri qualitativi**, che rappresentano qualità (colori, stili, materiali etc...) e non dati numerici e solitamente non hanno una *relazione d'ordine*
- **caratteri quantitativi**, maggiormente studiati dal corso, hanno una *relazione d'ordine* e sono divisi in:
  - **discreti**, come i lanci di un dado, rappresentanti valori in $\mathbb{Z}$
  - **continui**, che assumono valori reali, come la temperatura, i $\mathbb{R}$

Supponiamo di considerare n elementi della popolazione e di rilevare, per ognuno di essi, il dato relativo al carattere quantitativo da esaminare. Si ha un insieme di dati:

$$E = \{x_1, ..., x_n\}$$

la numerosità è il numero di elementi considerati, n. Se il carattere è discreto è comodo raggruppare i dati considerando l'insieme di tutti i valori assumibili, **modalità del carattere** ed associare ad ognuno di tali valori il numero di volte che esso compare in E. Si quindi N che è il numero di totalità del carattere e si ha l'insieme delle modalità:

$$S = \{s_1, ..., s_N\}$$

Si chiama **frequenza assoluta** s_j. f_j è il numero di volte che si presenta un elemento di un campione ovvero è il numero di elementi di E che hanno lo stesso valore s_j.
Si ha la **distribuzione di frequenza assoluta dei dati** unzione che associa ad ogni modalità la corrispondente frequenza assoluta:

$$F : S \to N$$

Si definisce **frequenza cumulata assoluta** per la modalità s_j la somma delle frequenze assolute di tutte le modalità:

$$s : k \in S : s_k \leq s_j$$

$$F_j = \sum_{k:s_k \leq s_j} f_k$$

frequenza relativa:

$$p_j = \frac{f_j}{n}$$

frequenza cumulativa relativa:

$$P_j = \sum_{k:s_k \leq s_j} p_k$$

Si dicono distribuzione di frequenza cumulata assoluta, relativa e cumulata relativa dei dati osservati, le funzioni F, p, P che associano ad ogni modalità frequenze s_j le relative frequenze F_j, p_j, P_j, con n numero di osservazioni Quando il carattere da studiare è continuo (o discreto con un gran numero di valori) è conveniente ricondursi a raggruppamenti come quelli appena trattati. Si suddivide S, l'insieme delle modalità, in alcune classi (sottoinsiemi di S) che formano una partizione (ogni famiglia di classi tra loro disgiunte la cui

unione dà tutto S). La scelta delle classi con cui si suddivide l'insieme S è del tutto arbitraria anche se è necessario che esse formino una partizione di S. Le partizioni devono essere significative e sufficientemente numerose. Ad ogni classe si asscoiano delle grandezze:

- confine superiore e inferiore (valori estremi della classe)
- ampiezza (differenza tra confine superiore ed inferiore)
- valore centrale (semi-somma tra i confini)

Nel caso in cui il carattere esaminato sia continuo occorre specificare quando le classi sono chiuse, a destra o a sinistra, ovvero specificare se gli elementi dell'indagine il cui dato coincide con il confine della classe sono da raggruppare all'interno della classe stessa oppure no.

3.0.1 Indici di tendenza Generale

Si cerca un modo di rappresentare una qualche serie di dati con un solo valore. Si usano gli **indici di tendenza generale** che sono quantità in grado di sintetizzare con un solo valore numerico i valori assunti dai dati. Uno di questi è la **media aritmetica** $\overline{x}$. Ho un campione $x_1, ..., x_n$. Si ha la media:

$$\overline{x} = \frac{1}{n} \sum x_i = \frac{x_1 + \cdots + x_n}{n}$$

Nel caso in cui i dati siano di tipo quantitativo discreto allora avremo:

$$\overline{x} = \frac{1}{n} \sum s_j f_j = \frac{s_1 f_j + \cdots + s_n f_n}{n}$$

$$\overline{x} = \sum s_j p_j = s_1 p_1 + \cdots + s_N p_N$$

si ha il momento k-esimo rispetto ad y:

$$M_{k,y} = \frac{1}{n} \sum (x_i - y)^k$$

la media è anche il **momento primo rispetto all'origine**, con $y = 0$.

Un secondo indice di tendenza è rappresentato dalla **mediana** definita come *quel numero reale che precede tanti elementi della serie di dati quanti ne segue.*

Se ordino $x_1, ..., x_n$ in ordine crescente ottengo la serie:

$$x_{(1)}, ..., x_{(n)}$$

e la **mediana** $\hat{x}$ è:

- $\frac{n+1}{2}$ se n è dispari
- dalla media aritmetica tra l'elemento di posto $\frac{n}{2}$ e quello di posto $\frac{n}{2}+1$ se n è pari

la **moda** $\widetilde{x}$ è quel valore o classe cui corrisponde la massima frequenza assoluta. La moda viene spesso utilizzata nel caso di dati qualitativi ovvero quando risulti impossibile definire media e mediana. Si ha che non è garantita l'unicità della moda infatti parleremo di:

- **distribuzione uni-modale** nel caso in cui vi sia un unica moda
- **distribuzione multi-modale** nel caso in cui vi siano più mode