

UNIVERSITÀ DEGLI STUDI DI SALERNO



**Dipartimento di Ingegneria dell'Informazione ed
Elettrica e Matematica applicata**

Corso di Laurea Magistrale in Ingegneria Informatica

**APPUNTI DI DATA SCIENCE
DI FRANCESCO PIO CIRILLO**

<https://github.com/francescopiocirillo>



"Sii sempre forte"

 Ehi, un attimo prima di iniziare!

Hai appena aperto una raccolta di appunti che ho deciso di condividere **gratuitamente** su GitHub, se ti sono utili fai **una buona azione digitale**:

-  **Lascia una stellina alla repo:** è gratis, indolore e fa super piacere!
-  **Condividerla con amici**, compagni di corso, o chiunque possa averne bisogno.

Insomma, se questi appunti ti salvano anche solo una giornata di studio... fammelo sapere con una **stellina!**

Grazie di cuore 

Ricapitolazione su MLE - 07/10 (Stima parametro deterministico θ)

Stima del parametro deterministico θ

(corrispondenza con Lab 2 - MLE CI MonteCarlo heterData MMSE.R - line 10)

$$Y_i = \theta + w_i$$

\hookrightarrow deterministico

L'obiettivo è stimare θ .

$$\begin{aligned} \mathbb{E}[w_i] &= 0 && \text{iid} \\ \mathbb{E}[w_i^2] &= \text{Var}[w_i] = \sigma^2 \end{aligned}$$

L'errore w è a media 0, visto che la varianza si calcola:

$$\sigma^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Con media 0 è uguale al momento del secondo ordine.

Sappiamo che θ è deterministico, quindi la sua media è se stesso. La media di Y_i è proprio θ .

$$\mathbb{E}[Y_i] = \mathbb{E}[\theta + w_i] = \mathbb{E}[\theta] + \mathbb{E}[w_i] = \theta$$

Possiamo quindi stimare θ per mezzo della media di Y_i e un buon stimatore della media è la media campionaria.

$$\begin{aligned} \{Y_i\}_{i=1}^n &\xrightarrow{\text{random sample}} \text{FACCIO UNA STIMA con la MEDIA CAMPIONARIA} \\ \text{STIMATORE} &\Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^n Y_i = \hat{\Theta} && \hookrightarrow \text{theta maiuscola} \\ \text{STIMA} &\Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^n y_i && \text{sample mean} \\ \{y_i\}_{i=1}^n &\xrightarrow{\text{observed sample}} \end{aligned}$$

Se è lo stimatore quindi variabile aleatoria è maiuscolo, la specifica osservazione invece è minuscola.
Tutto questo era per capire in che campo ci muoviamo, ora mettiamo il modello.

Aggiungiamo il modello

Assegniamo un modello a w_i , che saranno gaussiane a media 0 e varianza σ^2 (che in questa fase assumiamo nota).

Vogliamo ottenere lo stimatore di θ .

CALCOLI : $Y_i = \theta + w_i$ $\theta ?$	$w_i \sim N(0, \sigma^2)$ σ^2 nota $Y_i \sim N(\theta, \sigma^2)$ $\{Y_i\}_{i=1}^n$	$\{y_i\}_{i=1}^n$ controlloiamo se lo stimatore è a massima verosimiglianza
--	---	---

Le w_i saranno iid e quindi anche le y_i .

Usando il modello possiamo usare una strategia più avanzata della media campionaria. Sappiamo che anche Y_i saranno gaussiane.

Quando abbiamo il modello (ora stiamo facendo tutto model based) possiamo fare lo stimatore a massima verosimiglianza oppure Least Squares (è una strategia per la stima di quantità deterministiche ma ignote, poi cambieremo paradigma passando allo Bayesiano quando bisognerà stimare variabili aleatorie).

MLE	$L(\theta) = L(\theta; \underline{y}) = \ell(\theta; \underline{y})$ produttoria delle dist. marginali	$\underline{y} = (y_1, \dots, y_n)^T$
------------	---	---------------------------------------

La verosimiglianza è la congiunta delle osservazioni vista come congiunta del parametro theta.

▼ GPT definizione formale

Definizione formale

Supponiamo di avere un insieme di dati osservati $\mathcal{D} = \{x_1, x_2, \dots, x_n\}$, che si suppone siano realizzati da una variabile casuale con una funzione di densità di probabilità (o una funzione di massa di probabilità, nel caso discreto) dipendente da un parametro θ . La funzione di verosimiglianza $\mathcal{L}(\theta; \mathcal{D})$ è definita come:

$$\mathcal{L}(\theta; \mathcal{D}) = f(x_1, x_2, \dots, x_n | \theta)$$

dove $f(x_1, x_2, \dots, x_n | \theta)$ è la funzione di probabilità o densità congiunta dei dati osservati \mathcal{D} , data la distribuzione parametrizzata da θ .

Per massimizzare la verosimiglianza dobbiamo avere una verosimiglianza da massimizzare, quindi per prima cosa bisogna trovare questa funzione, spesso chiamata $L(\theta, \mathbf{y} vettore)$.

$$\Rightarrow l(\theta; \mathbf{y}) = \prod_{i=1}^n f(y_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \theta)^2}{2\sigma^2}}$$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y - \theta)^2}{2\sigma^2}}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2\right\}$$

Alcuni la chiamano log verosimiglianza con la l piccola.

Visto che tutto è gaussiano la nostra funzione di verosimiglianza, cioè il nostro l , sarà la congiunta delle pdf.

La congiunta delle pdf è la produttoria delle distribuzioni marginali.

$\hat{\theta}$ sarà l'argmax, per tutte le theta appartenenti al suo spazio, della funzione di likelihood. A noi interessa il punto in cui si massimizza la funzione, non il valore esatto del punto di massimo (ci saranno poi anche misure relative al punto di massimo).

La massimizzazione non è modificata da una costante moltiplicativa in quanto una costante moltiplicativa implica uno shift verticale pari per tutta la funzione.

Infatti la verosimiglianza la definiamo sempre a meno di costanti.

Ignoriamo quindi la costante e calcoliamo l'argmax solo dell'esponenziale, addirittura con la log verosimiglianza togliamo anche l'esponenziale.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} l(\theta; \mathbf{y}) = \underset{\theta}{\operatorname{argmax}} \log l(\theta; \mathbf{y})$$

$$= \underset{\theta}{\operatorname{argmax}} \left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \theta)^2 \right\}$$

Si chiamerà maximum log likelihood ma effettivamente è sempre la likelihood.

?

La introduciamo perché è una trasformazione monotona e crescente, quindi il punto di massimo si trova anche con il punto di massimo e minimo della funzione.

C'è un meno, l'argmax diventa l'argmin ma la cosa non rappresenta un problema.

$$\frac{d}{d\theta} -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 = 0 = -\frac{1}{2} \sum_{i=1}^n (y_i - \theta)$$

le cost le possiamo togliere

torna con la derivata

Di nuovo $1/\sigma^2$ essendo costante può essere rimosso.

Visto che ci interessa il massimo occorre utilizzare derivata prima/gradiente e derivata seconda/matrice hessiana.

$$\Rightarrow \sum_1^n y_i = n \theta \Rightarrow \hat{\theta}_{ML} = \frac{1}{n} \sum_1^n y_i$$

Abbiamo dimostrato che lo stimatore è la media campionaria.

Abbiamo trovato un punto stazionario, per decidere che è il massimo serve la derivata seconda, sicuramente $-1/\sigma^2$ ci darà una derivata negativa quindi tutto si trova plausibilmente, non ci siamo dedicati a fare anche i calcoli.

La stima a massima verosimiglianza in questo caso ci ha restituito proprio quello che suggeriva anche l'intuito, cioè che per stimare una media basta la media campionaria.



La media campionaria esce come stimatore per la massima verosimiglianza di molte distribuzioni utili praticamente che si dicono appartenenti alla famiglia esponenziale.

Valutazioni sullo stimatore trovato

Bias

$$\hat{\theta} = \frac{1}{n} \sum_1^n y_i \rightarrow \text{stima} \quad \hat{\theta} = \frac{1}{n} \sum_1^n Y_i \rightarrow \text{stimatore}$$

Perché y minuscole
quindi osservazioni

Perché Y maiuscole
quindi variabili aleatorie

Abbiamo detto che $\hat{\theta}$ è la media campionaria.

Ma qual è la media di $\hat{\theta}$? Come dimostrato in passato la media campionaria è uno stimatore unbiased della media, e lo possiamo dimostrare.

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}\left[\frac{1}{n} \sum_1^n Y_i\right] = \frac{1}{n} \sum_1^n \mathbb{E}[Y_i] = \frac{1}{n} \sum_1^n \theta =$$

|

$= \theta \rightarrow \text{stimatore unbiased}$

La media di una combinazione lineare è la combinazione lineare delle medie.

Ma il MLE è sempre unbiased? La risposta è no, in questo caso specifico lo è.

Si può dimostrare che lo stimatore ML è asintoticamente unbiased.

STIMATORE ML asintoticamente unbiased
 ↳ per $n \rightarrow \infty$

Varianza

Calcoliamo la varianza del nostro stimatore (abbiamo cambiato il nome in $\hat{\mu}$ perché è uno stimatore della media ma comunque è lo stesso di prima).

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_1^n Y_i \quad \text{Var}[\hat{\mu}_{ML}] = \frac{1}{n^2} \text{Var}\left[\sum_1^n Y_i\right]$$

$\begin{matrix} \text{incorr.} \\ \text{ind} \end{matrix} \quad \begin{matrix} \sum_1^n \\ \text{Var}[Y_i] = \sigma^2 \end{matrix} \quad \text{Var}[\hat{\mu}_{ML}] = \frac{\sigma^2}{n}$

Il fatto che la varianza di una sommatoria sia equivalente alla sommatoria delle varianze è implicato dalla incorrelazione perché la covarianza è 0 se sono incorrelate, significa che non c'è legame lineare.

Legami non lineari potrebbero esserci e solo se non ci sono si parla di indipendenza. L'indipendenza implica l'incorrelazione. In questo caso sappiamo che sono indipendenti.

Notiamo che lo stimatore della media per n che tende ad infinito diventa una costante (la varianza va a 0), la sua distribuzione diventa una delta, la pdf con un solo valore, è una degenerazione.

MSE

Proviamo a calcolare l'MSE del nostro stimatore.

$$\text{MSE}(\hat{\mu}_{ML}) = \mathbb{E}[(\hat{\mu}_{ML} - \mu)^2] = \text{Var}[\hat{\mu}_{ML}] + b^2(\underbrace{\mu_{ML}}_0) = \frac{\sigma^2}{n}$$

Abbiamo fatto i conti ma tutto questo potevamo dirlo senza calcoli, la media aritmetica di osservazioni tende al valore della media all'aumentare del campione per la Legge dei Grandi Numeri nella sua forma debole o forte a seconda della convergenza ma comunque il concetto è lo stesso.

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_1^n Y_i \xrightarrow{n \rightarrow \infty} \mu = \mathbb{E}[Y_i]$$

Law of Large Numbers (LLN)

In statistica il fatto che all'aumentare del campione lo stimatore converge al valore vero prende il nome di consistenza.

$\hat{\theta}_{ML}$ è uno stimatore consistente



Per ogni stimatore costruito con ML sarà vero che lo stimatore ottenuto è consistente. Cioè converge al valore vero TUTTA la distribuzione dello stimatore che diventa una delta di dirac.

MLE è quindi efficiente, non è unbiased ma lo è asintoticamente, è consistente, ha in generale bassa varianza. Non è detto che sia lo stimatore a minima varianza ma converge ad essere il migliore possibile. Ovviamente il limite di MLE è che bisogna conoscere il modello, **è model based**.



Per la varianza c'è un limite invalicabile, non possiamo arrivare a 0, c'è un lower bound, Cramér-Rao Lower Bound (CRLB).

$$\text{MLE} \quad X \sim \bar{E}_{x_p}(\mu)$$

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x \geq 0$$

$$\mu = \mathbb{E}[x]$$

Digressione su MLE



Lo stimatore a massima verosimiglianza è molto potente e la procedura per trovarlo è proprio quella che abbiamo visto, cioè si ricava la funzione di verosimiglianza e poi la si massimizza.

Ovviamente se tutto analitico è facile il massimo lo troviamo carta e penna, altrimenti se è più complesso perché si vanno a derivare funzioni complesse si segue una strada con algoritmi numerici. Essenzialmente o c'è l'espressione analitica o si implementano algoritmi numerici.

Differenze tra MLE e LS

Per stimare il parametro deterministico nel primo esempio abbiamo usato MLE perché avevamo il modello di Y e w , è necessario avere la pdf e avere l'ipotesi di sapere come combinarle (se sono indipendenti basta il prodotto, altrimenti si può fare ma è più complesso).

Nella regressione però non sempre abbiamo il modello dell'errore, lo usiamo sicuramente per test di ipotesi e stime intervallari ma in linea di principio la stima delle beta con Least Squares **NON** richiede il

modello dell'errore w (o epsilon).



Se tutto è Gaussiano le stime ML e LS coincidono, questo ci fa capire che anche least squares gode delle proprietà matematiche importanti di ML.

MLE quindi è model based, ci serve il modello probabilistico o comunque lo dobbiamo stimare per poter scrivere la funzione di verosimiglianza che poi dobbiamo massimizzare.

ESEMPIO STIMATORE NON UNBIASED - Stima a ML della varianza di una Gaussiana a media μ e varianza σ^2

(corrispondenza con Lab 2 - MLE CI MonteCarlo heterData MMSE.R - line 57)

$$\text{MLE varianza di } N(\mu, \sigma^2) \quad \mu \text{ known} \\ \sigma^2 ?$$

Supponiamo μ noto, per prima cosa raccogliamo i campioni.

$$x_1, \dots, x_n \stackrel{iid}{\sim} N(\mu, \sigma^2) \quad f_{x_i}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\ \sigma^2 = \theta$$

Procediamo alla ricerca della funzione di verosimiglianza.

$$\begin{aligned} L(\theta) &\stackrel{iid}{=} \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x_i - \mu)^2}{2\theta}} \\ &= (2\pi\theta)^{-\frac{n}{2}} \exp \left\{ -\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\theta} \right\} \end{aligned}$$

Passiamo alla log likelihood al fine di semplificare i calcoli.

$$\Rightarrow \log L(\theta) = -\frac{n}{2} \log(2\pi\theta) - \frac{\sum_i (x_i - \mu)^2}{2\theta}$$

Poniamo la derivata della log likelihood a 0 per cercare il massimo.

$$\frac{d}{d\theta} \log L(\theta) = 0$$

Quindi:

$$-\frac{n}{2} \frac{2\hat{\sigma}}{2n\theta} - \frac{-2\sum_i (x_i - \mu)^2}{2\theta^2} = 0$$

$$-\frac{n}{2\theta} + \frac{\sum_i (x_i - \mu)^2}{2\theta^2} = 0$$

E così troviamo uno stimatore che è **asintoticamente unbiased** cioè diventa unbiased per n che va a infinito:

$$\hat{\sigma}_{ML}^2 = \hat{\sigma}_{ML}^2 = \frac{\sum_i (x_i - \mu)^2}{n}$$

asintoticamente unbiased

Notiamo la differenza dallo stimatore unbiased della varianza già visto in passato:

$$\hat{\sigma}_{unb.}^2 = \frac{1}{n-1} \sum_1^n (x_i - \mu)^2$$



IMPORTANTE:

In questo caso abbiamo supposto μ noto, ma se anche questo fosse stato ignoto comunque sarebbe stato possibile utilizzare Maximum Likelihood visto che conosciamo la distribuzione, semplicemente nella Likelihood function ci sarebbero stati due parametri e quindi invece di utilizzare la derivata prima sarebbe stato necessario il gradiente.

ESERCIZIO PER CASA - Stimare l'MLE per un modello esponenziale di parametro μ

Vogliamo trovare il Maximum Likelihood Estimator per un modello esponenziale di parametro μ .

Il modello esponenziale può essere scritto in due modi:

- il parametro è il parametro della media;
- il parametro è il reciproco della media.

Quindi è meglio esplicitare la pdf per essere chiari.

Semplicemente verificandolo sarà possibile constatare che μ è proprio la media.

I dati sono raccolti in un vettore x , come sempre un campione di n osservazioni iid.

Ovviamente ogni x_i è estrazione casuale della variabile aleatoria X_i distribuita in accordo al modello di popolazione X .

$$\begin{array}{l} \underline{x} = (x_1, \dots, x_n) \quad \text{dati } x_i \stackrel{\text{iid}}{\sim} X \\ \hat{\mu}_{\text{ML}} = ? \end{array}$$

↳ mod di pop.

L'obiettivo è stimare a massima verosimiglianza la μ .

tl;dr

▼ tl;dr Bignami per "Stima del parametro deterministico θ "

Introduzione al problema

Abbiamo un fenomeno di interesse che è modellabile come una variabile aleatoria Y .

Sappiamo che $Y_i = \theta + W_i$ dove θ è un parametro deterministico, cioè un numero, e W_i è una variabile aleatoria della quale NON conosciamo la distribuzione (Gaussiana, Chi Squared, t di student ecc...) MA sappiamo che ha media 0 e varianza σ^2 .

L'obiettivo è stimare θ , come sempre abbiamo un dataset di n osservazioni y_i .

Soluzione semplice

Dimostriamo che θ è uguale alla media di Y_i , quindi abbiamo risolto il problema perché invece di stimare θ stimiamo la media di Y_i e lo stimatore per eccellenza della media è la media campionaria.

Soluzione difficile

Con i dati che avevamo prima non si poteva fare niente di particolare MA immaginiamo invece che oltre a quello che sapevamo prima sappiamo anche qual è il modello di W_i (e conseguentemente anche di Y_i) e nello specifico immaginiamo che la traccia ci dica che è Gaussiano, sempre a media 0 e varianza σ^2 come prima.

Conoscendo il modello possiamo usare un approccio più sofisticato rispetto alla media campionaria: Maximum Likelihood Estimator.

Essenzialmente calcoliamo una funzione densità di probabilità speciale che si chiama Likelihood dopodiché la massimizziamo facendo la derivata prima, o gradiente, posta a 0 e poi controllando di aver trovato il massimo e non un minimo per mezzo della derivata seconda, o matrice Hessiana, che deve essere definita negativa.

La cosa interessante è che ciò che otteniamo è, di nuovo, la media campionaria.

Considerazioni

Calcolando la media dello stimatore trovato $\hat{\theta}$ scopriamo che è pari a θ cioè al valore vero, ergo lo stimatore è unbiased, in generale MSE non è per forza unbiased ma è asintoticamente unbiased.

La varianza di $\hat{\theta}$ è $\frac{\sigma^2}{n}$, il che significa che per n che va a infinito la varianza tende a 0, in altre parole la distribuzione del nostro stimatore diventa un numero, la media, e visto che la media è il valore vero (unbiased) il nostro stimatore diventa il valore vero.

Visto che il nostro stimatore è la media campionaria questa dimostrazione è coerente alla Law of Large Numbers.

Visto che l'MSE è la somma di bias e varianza in questo caso l'MSE è pari alla varianza, il fatto che l'errore vada a 0 per n che va a infinito rende lo stimatore trovato consistente, MLE produce sempre stimatori consistenti.

▼ tl;dr Bignami per "ESEMPIO STIMATORE NON UNBIASED - Stima a ML della varianza di una Gaussiana a media μ e varianza σ^2 "

Introduzione al problema

Supponiamo di avere un fenomeno di interesse che è modellabile come una Gaussiana a media μ (supposta nota) e varianza σ^2 .

Il nostro obiettivo è stimare la varianza, visto che abbiamo il modello (Gaussiana) possiamo usare un approccio model-based come MLE.

Soluzione

Individuando la funzione Likelihood e massimizzandola (sempre con tutto il discorso delle derivate) otteniamo uno stimatore per la varianza.

Lo stimatore risulta non essere quello unbiased che abbiamo usato in precedenti lezioni ma un altro che è asintoticamente unbiased il che è in linea con il fatto che MLE è sempre asintoticamente unbiased ma non sempre unbiased.

La differenza tra i due stimatori è che quello unbiased ha $(n-1)$ a denominatore mentre quello trovato ora ha solo n .

Se non avessimo conosciuto μ ?

Per usare MLE serve il modello, cioè sapere se è una Gaussiana o una Chi squared ecc...

In questo caso sappiamo che è una Gaussiana quindi anche se non conosciamo neanche μ possiamo comunque usare MLE e semplicemente invece della derivata su σ^2 si fa il gradiente su μ e σ^2 e così si ottengono due stimatori.

DISCLAIMER

Questi appunti sono stati realizzati a scopo puramente educativo e di condivisione della conoscenza. Non hanno alcun fine commerciale e non intendono violare alcun diritto d'autore o di proprietà intellettuale.

I contenuti di questo documento sono una rielaborazione personale di lezioni universitarie, materiali di studio e concetti appresi, espressi in modo originale ove possibile. Tuttavia, potrebbero includere riferimenti a fonti esterne, concetti accademici o traduzioni di materiale didattico fornito dai docenti o presente in libri di testo.

Se ritieni che questo documento contenga materiale di tua proprietà intellettuale e desideri richiederne la modifica o la rimozione, ti invito a contattarmi. Sarò disponibile a risolvere la questione nel minor tempo possibile.

In quanto autore di questi appunti non posso garantire l'accuratezza, la completezza o l'aggiornamento dei contenuti e non mi assumo alcuna responsabilità per eventuali errori, omissioni o danni derivanti dall'uso di queste informazioni. L'uso di questo materiale è a totale discrezione e responsabilità dell'utente.