

APPUNTI DEL CORSO

IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI

A.A. 2017/2018 - UNIVERSITY OF BERGAMO

PARTE I: SISTEMI STATICI

AUTORE: MIRKO MAZZOLENI



Control Automation Lab

L'uso e la distribuzione di questi appunti è consentita previa citazione dell'autore e della fonte originari

Corsi di IDENTIFICAZIONE DEI MODELLI & ANALISI DEI DATI

I) IDENTIFICAZIONE DEI MODELLI

trova il legame tra queste
grandezze e descrivere
matematicamente

MODELLO: Descrizione matematica di un fenomeno o di un sistema

- L'economico: relazione tra reddito ed educazione
- L'sociale: relazione tra luoghi di abitazione e criminalità
- L'fisico: relazione tra massa e peso di una persona

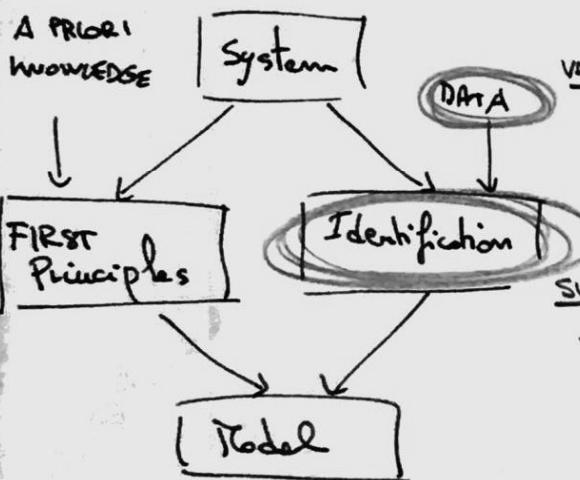
SISTEMA: Meccanismo costituito che trasforma input (causa) in output (effetto)

$$u \rightarrow [S] \rightarrow y$$

CONOSCENZA
A PRIORI

Due approcci fondamentali:

a) WHITE BOX MODELING: - approccio basato su leggi e principi base delle FISICA o
L'es. modello di un condensatore $I(t) = C \cdot \frac{dV(t)}{dt}$



VANTAGGI

- conoscenza del significato delle variabili (C : capacità di conservare)
- generalizzabile: se conosco C , il modello vale anche per altri sistemi
- perché conoscerne le relazioni CAUSALI tra C ed $I(t)$

Svantaggi

- Richiede conoscenze avanzate delle leggi del problema SPECIFICO
 - L'es. costo, costi alti
- per sistemi complessi la scrittura di molte equazioni diventa impossibile
- limitato a campi in cui esistono leggi causali

b) BLACK BOX MODELING: - approccio basato su DATI Sperimentali

VANTAGGI

- presuppone del particolare tipo di problema, limitandosi a costruire il legame tra le variabili $y = f(u)$

SVANTAGGI

- non interpretabile fisicamente
- non generali, dipendono dal tipo di dati acquisiti. Per ogni modifica del sistema, bisogna ripetere l'esperimento

IDENTIFICAZIONE DEI MODELLI

↓
deve trovare un
modello che
me li descriva
i dati

PROBLEMA DI STIMA

↓
stima le caratteristiche
di questi dati

2) ANALISI DEI DATI

- Determinare le caratteristiche statistiche dei dati e delle variabili misurate
 - L'essi infatti sono affetti da RUMORE ed INCERTITUDINE
 - L'media L'correlazione tra variabili
 - L'varianza L'distribuzione probabilistica

STATISTICA
DESCRITTIVA

- Individuare la STRUTTURA, delle regolarità (se ci sono)

L'i dati presentano dei "PATTERN" riconoscibili o sono RANDOM?

L'osservare ALLENARE algoritmi che NA SOLI individuino pattern? HL

LE PROCEDURE 1) ed 2)

1) e 2) sono strettamente interconnesse:

- i) Spesso l'analisi preliminare dei dati dà indicazioni sul modello migliore per descriverli
- ii) Tecniche di analisi dei dati sono usate per descrivere le tendenze del modello
- iii) Una rappresentazione probabilistica dei dati fa legge ad un modello capace di gestire l'incertezza

L' sia velle misure

L' sia velle conoscenze della realtà

L' quantificare quello che va su

DECLINERETE le due procedure sia per sistemi statici che per sist. DINAMICI

Lsistemi statici: la sola conoscenza delle variabili e eg è sufficiente a calcolare $y \Rightarrow V(t) = R \cdot I(t)$

Lsistemi DINAMICI: bisogna di sapere le condizioni iniziali: $\frac{dV(t)}{dt} = \frac{1}{C} \cdot I(t)$
per conoscere $V(t)$ deve sapere $V(t_0)$

L concetto di STATO: $V(t) = x_1(t) \Rightarrow \begin{cases} \dot{x}_1(t) = \frac{1}{C} \cdot u(t) \\ I(t) = u(t) \\ V(t) = y(t) \end{cases}$

L Fondamentali di automotrice !! \Rightarrow la $G(z)$ era DATI.
Come trovarla?: -WHITE BOX

-BLACK BOX

$$\begin{array}{c} U \\ | G(z) | \end{array}$$

(2)

RICHIATI DI STATISTICA

- Una variabile casuale V è una variabile definita a partire dall'esito S di un esperimento casuale \rightarrow Es. L'esperimento è il lancio di una moneta. A seconda se essa teste o croce, V assume un valore
L'indichiamo con v.c come $V(S)$
L'el valore assunto da v a seguito di un particolare esito S è $v(S)$

Se v può assumere diversi valori, come li descriviamo? \Rightarrow Assegnare una probabilità che ogni esito occorra \Rightarrow questo influenza sulla probabilità dei valori che v può assumere

- Se v assume valori DISCRETI (v è una variabile casuale discreta)
 - L'funzione di probabilità di massa $p(x) = P(v=x)$ associa ad ogni valore x di v una probabilità (pmf)

Indichiamo con x_i : valori di v . Se v può assumere m diversi valori, allora $\sum_{i=1}^m p(x_i) = 1$

Esempio: TIRARO

$$\begin{array}{ll} x_1=1 & p(x_1) = P(v=x_1) = P(v=1) = \frac{1}{6} \\ x_2=2 & p(x_2) = P(v=2) = \frac{1}{6} \\ | & | \\ x_6=6 & p(x_6) = P(v=6) = \frac{1}{6} \\ m=6 & \sum_{i=1}^6 p(x_i) = 6 \cdot \frac{1}{6} = 1 \end{array}$$

- Se v assume valori CONTINUI (v è una v.c. continua)

L'funzione di densità di probabilità $f(x)$ (pdf)

- ~~$P(v=x)$~~ non ha senso

fossero equiprobabili anche se i valori fossero equiprobabili come nel caso, la prob. di un valore sarebbe $\frac{1}{\infty} = 0$ ci sono infiniti possibili valori: la prob. di uno esattamente uno di quelli è zero

$$P(v \in [a, b]) = \int_a^b f(x) dx$$

Es: v è l'altezza di un uomo adulto

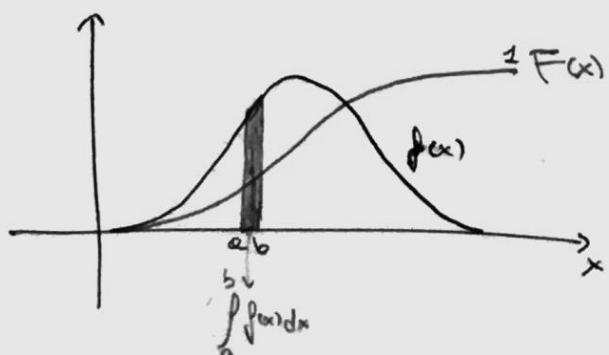
\hookrightarrow non ha senso chiedersi la probabilità che un uomo sia alto ESATTAMENTE 1,7235142... metri

$$\begin{aligned} & f(x) \geq 0 \\ & \int_{-\infty}^{+\infty} f(x) dx = 1 \end{aligned}$$

- Funzione di densità cumulata (cdf) o distribuzione di probabilità

$$F(z) = \int_{-\infty}^z f(x) dx = P(x \leq z)$$

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



A FINI PRATICI x È DISCRETO
 $f(x) \propto P(v=x)$

- Il valore atteso di una v.c. continua è:

$$E[v] = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

Somma pesata dei valori x da v può assumere. I pesi sono le prob. $f(x)$. Peso ogni valore per le sue probabilità di manifestarsi

- LINEARITÀ: $E[\alpha v_1 + \beta v_2 + \gamma] = \alpha E[v_1] + \beta E[v_2] + \gamma \quad \forall \alpha, \beta, \gamma \in \mathbb{R}$

- La varianza di una v.c. continua è:

$$\text{Var}[v] = \int_{-\infty}^{+\infty} (x - E[v])^2 \cdot f(x) dx$$

$$= E[(x - E[x])^2]$$

- di quanto i valori x si scostano dalla loro media
- se piccoli, v assume valori molto vicini fra loro

Osservazione

- $\text{Var}[v] \geq 0$. Se $\text{Var}[v] = 0$, la variabile v è deterministica (assume sempre un solo valore)

- Deviazione standard: $\sigma[v] = \sqrt{\text{Var}[v]}$

$$\begin{aligned} \boxed{\text{Var}[v]} &= \boxed{E[(v - E[v])^2]} = E[v^2 - 2E[v]v + E[v]^2] = E[v^2] - 2E[v \cdot E[v]] + E[E[v]^2] \\ &= E[v^2] - 2E[v] \cdot E[v] + E[v]^2 = \boxed{E[v^2] - E[v]^2} \end{aligned}$$

- $\text{Var}[\alpha \cdot v_1 + \beta] = \alpha^2 \cdot \text{Var}[v_1] \quad \forall \alpha \in \mathbb{R}$

- Date due v.c. v_1 e v_2 si definisce il coefficiente di correlazione come:

$$\rho = \frac{E[(v_1 - E[v_1]) \cdot (v_2 - E[v_2])]}{\sigma[v_1] \cdot \sigma[v_2]}$$

- ρ indica il grado di dipendenza lineare tra v_1 e v_2 . Infatti se $v_2 = \alpha v_1 + \beta \Rightarrow \rho = 1$

- Se $\rho = 0$ le due variabili si dicono sconelte

- Date v_1 e v_2 si definisce covarianza la varianza come

$$\text{Cov}(v_1, v_2) = E[(v_1 - E[v_1]) \cdot (v_2 - E[v_2])]$$

e quindi:

$$\rho = \frac{\text{Cov}(v_1, v_2)}{\sigma[v_1] \cdot \sigma[v_2]}$$

- v_1 e v_2 sono sconelte se $\text{Cov}(v_1, v_2) = 0$

- Le precedenti definizioni si possono estendere al caso di vettore di variabili casuali $\bar{v} = [v_1, v_2, \dots, v_d]^T$

- distribuzione di probabilità

$$\begin{aligned} F(x_1, x_2, x_3, \dots, x_d) &= P(v_1 \leq x_1, v_2 \leq x_2, \dots, v_d \leq x_d) \\ &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_d} f(x_1, x_2, \dots, x_d) dx_1 dx_2 \dots dx_d \end{aligned}$$

pdf congiunta

- vettore atteso è un vettore colonna di d componenti

$$E[\bar{v}] = [E[v_1], E[v_2], \dots, E[v_d]]^T \in \mathbb{R}^{d \times 1}$$

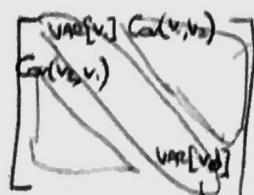
- La varianza è una matrice $m \times m$ semidefinita positiva:

$$\text{Var}[\bar{v}] = \int_{\mathbb{R}^d} ((x - E[\bar{v}]) (x - E[\bar{v}])^T f(x)) dx$$

L'insieme $\{x \in \mathbb{R}^d : x \geq 0 \text{ per numeri reali}$

M reale è semidefinita positiva se $x^T M x \geq 0$, $\forall x \in \mathbb{R}^d$

L'autosvalo > 0 tranne se $= 0$



\rightarrow VARIANZE DI v_1, v_2, \dots, v_d

→ COVARIANZE TDI

$v_1, v_2, v_1, v_3, \dots$

- SOTTOCASO: la covariante tra v_1 e v_2 è la stessa che tra v_2 e v_1 .

MATRICE DI VARIANZE - COVARIANZE

- Due variabili casuali v_1 e v_2 con funzione di probabilità composta f si dicono indipendenti se e solo se:

$$f(v_1, v_2) = f(v_1) \cdot f(v_2)$$

Mirko Mazzoleni - University of Bergamo

Teorema

Se v_1 e v_2 sono indipendenti, allora sono scorrelate

aviamente μ e σ^2 determinano
cose sono le caratteristiche dei miei
dati

Es la densità di probabilità di y ha una
buona funzione che dipende da θ \rightarrow VER^{xx}

ESTIMA $\Rightarrow \hat{\theta}$

Observe detto che ci concentriamo sulla STIMA PARAMETRICA. Vogliamo quindi trovare il parametro θ^* che le genera: dati $D = \{y(1), \dots, y(n)\}$

Interpretazione: dati come v.c.
per vedere le loro incertezze, $D = D(\bar{s}, \theta^*)$

l'incertezza i dati sono:
nelle loro variabili casuali, quelli che noi osserviamo,
misurare

\hookrightarrow l'incertezza dello specifico esito $\bar{s} \Rightarrow D = D(\bar{s}, \theta^*)$

che associa ai dati un
valore del parametro da stimare

Uno STIMATORE è una funzione $T(D(s, \theta^*))$. La stima è il risultato di un stimatore
su una specifica $\hat{\theta} = T(D(\bar{s}, \theta^*)) \rightarrow$ poiché il risultato di T dipende dall'esito s (da
realizzazione dei dati)

ci dipende i dati), allora lo stimatore è una variabile casuale dipendente da s

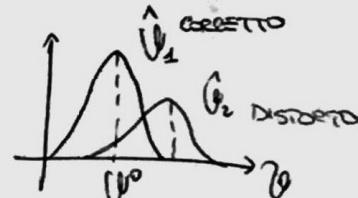
Es misura il peso V degli studenti $\Rightarrow \hat{\theta}_1 = T(D(s_1, \theta^*))$

\hookrightarrow posso misurare solo 5 studenti $D = \{y(1) - y(5)\} \Rightarrow \hat{\theta}_2 = T(D(s_2, \theta^*))$

Ma se ne quindi calcolare valore atteso e varianza di questa variabile casuale; in base a queste, valuteremo la bontà di un stimatore.

PROPRIETÀ DI UNO STIMATORE

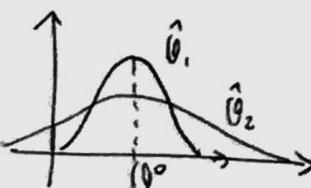
- Un stimatore si dice corretto se e solo se: $E[\hat{\theta}] = \theta^*$
L'errore cioè un errore sistematico di stima



- Un stimatore si dice asintoticamente corretto se e solo se: $\lim_{N \rightarrow \infty} E[\hat{\theta}] = \theta^*$
- è una proprietà più debole

PRIMA DEF. CONSISTENTE

Se due stimatori sono entrambi corretti, qual è il migliore? Quello è minima varianza



è la maggiore probabilità di ritrovare una stima vicina al valore reale!

- Un stimatore si definisce consistente se: $\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}] = 0$



Mirko Mazzoleni - University of Bergamo

- La consistenza grafica è all'aumentare del numero dei campioni la qualità delle stime aumenta

- Se $\hat{\theta}$ è corretto si ha che $\text{Var}[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2] = E[(\hat{\theta} - \theta^0)^2] = \text{Var}[\epsilon^2]$

$$\epsilon = \hat{\theta} - \theta^0$$

- ERRORE DI STIMA

corretto

- Un stimatore si dice ottimale se la sua varianza è la più piccola per una serie N di dati $D = \{y_1, y_2, \dots, y_N\}$

Es STIMATORE MEDIA

Sia $y = \frac{1}{N} \sum_{i=1}^N y_i$ la stimatore media con varianza σ^2

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N y_i \text{ è corretto. Infatti } E[\hat{y}] = E\left[\frac{1}{N} \sum_{i=1}^N y_i\right] = \frac{1}{N} \sum_{i=1}^N E[y_i] = \frac{1}{N} \cdot N \cdot \mu = \mu$$

Si dimostra che è consistente, $\text{Var}[\hat{y}] = \frac{\text{Var}[y]}{N} = \frac{\sigma^2}{N}$

$$\text{Es STIMATORE VARIANZA} \quad \text{varianza} = \text{Var}\left[\frac{1}{N} \sum_{i=1}^N y_i\right] = \frac{1}{N^2} \sum_{i=1}^N \text{Var}[y_i] = \frac{1}{N^2} \cdot N \cdot \sigma^2 = \frac{\sigma^2}{N}$$

$D = \{y_1, \dots, y_N\}$ con varianza σ^2 la stimatore s_{N-1}^2 è corretto, $s_{N-1}^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N-1}$

$$s_{N-1}^2 = E\left[\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2\right] = E\left[\frac{1}{N-1} \sum_{i=1}^N (y_i^2 + \bar{y}^2 - 2y_i\bar{y})\right] = E\left[\frac{1}{N-1} \left(\sum_{i=1}^N y_i^2 + N\bar{y}^2 - 2\bar{y} \sum_{i=1}^N y_i\right)\right] = E\left[\frac{1}{N-1} \left(\sum_{i=1}^N y_i^2 + N\bar{y}^2 - 2\bar{y} \cdot N\bar{y}\right)\right]$$

$$= E\left[\frac{1}{N-1} \left(\sum_{i=1}^N y_i^2 + N\bar{y}^2 - 2\bar{y} \cdot N\bar{y}\right)\right] = E\left[\frac{1}{N-1} \left(\sum_{i=1}^N y_i^2 - N\bar{y}^2\right)\right] = \frac{1}{N-1} \left(\sum_{i=1}^N y_i^2 - N\bar{y}^2\right)$$

$$\text{Var}[y] - E[y]^2 = \frac{1}{N-1} \left(\sum_{i=1}^N E[y_i^2] - N \cdot E[\bar{y}^2] \right) = \frac{1}{N-1} \left(N \cdot E[y^2] - N \cdot E[\bar{y}^2] \right) = \frac{N}{N-1} \left(E[y^2] - E[\bar{y}^2] \right)$$

$$= \frac{N}{N-1} \left(E[y^2] + E[y]^2 - E[\bar{y}]^2 - E[\bar{y}]^2 \right) = \frac{N}{N-1} \left(\text{Var}[y] + \bar{y}^2 - \text{Var}[\bar{y}] - \bar{y}^2 \right)$$

$$= \frac{N}{N-1} \left(\text{Var}[y] - \frac{\text{Var}[y]}{N} \right) = \frac{N}{N-1} \left(\frac{(N-1)}{N} \text{Var}[y] \right) = \text{Var}[y] = \sigma^2$$

CORRETTO!

Stabilisce un limite inferiore per la varianza di un qualsiasi stimatore

non possa essere più preciso di un altro.

L'è questo perché i dati sono offetti da un errore di misura che non posso rimuovere con le mie stime.

Nel caso di stimatori corretti abbiamo che: $\text{Var}[\hat{\theta}] \geq m^{-1}$ m: quantità di informazione di Fisher

L'è se $\hat{\theta}$ è un vettore: $\text{Var}[\hat{\theta}] - M^{-1} \geq 0$

- Un stimatore ~~è~~ si dice efficiente se $\text{Var}[\hat{\theta}] = m^{-1}$
- Un stimatore si dice asintoticamente efficiente se $\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}] = m^{-1}$ infatti se un $\hat{\theta}$ è grande la varianza è piccola

STIMA DI PESOLOLINEA:

STIMA A MINIMI QUADRATI (LEAST SQUARES)

Obbiamo finire descritto i dati $y_{(1)}, \dots, y_{(N)}$ in termini della loro media e varianza, dando degli stimatori per queste due quantità $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N y_{(i)}$, $S_{n-1}^2 = \frac{\sum_{i=1}^N (y_{(i)} - \hat{\mu})^2}{N-1}$

Supponiamo ora che vogliano descrivere i dati D tramite una relazione lineare:

Suppongo che i dati abbiano questa struttura, le imposto che x_1, x_2, \dots, x_d siano variabili di cui ci dispongono misure. Questo modello prende il nome di regressione lineare

Ese

$$y = \text{Peso } [kg]$$

$x_1 = \text{altezza } [m] \Rightarrow$ variabile numerica (c'è ordinamento) e quantificare la distanza

$x_2 = \text{sess} [M/F] \Rightarrow$ variabile categoreale (non c'è ordinamento)

$x_d = \text{lavoro di nascita } [\equiv]$

Vogliamo esprimere il peso in funzione delle varie x_1, x_2, \dots, x_d

N persone misurate

Definiamo i vettori:

$$\begin{aligned} \vartheta &= \begin{bmatrix} \vartheta_0 \\ \vartheta_1 \\ \vdots \\ \vartheta_d \end{bmatrix} & \varphi(i) &= \begin{bmatrix} 1 \\ x_1(i) \\ x_2(i) \\ \vdots \\ x_d(i) \end{bmatrix} & \Rightarrow & \boxed{y(i) = \varphi(i)^T \cdot \vartheta + e(i), i=1 \dots N} \\ dx_1 & & & & & \boxed{\boxed{\quad}} \end{aligned}$$

Approfondimenti

Possiamo misurare le stesse entità con diversi tipi di variabili. Ad esempio, i partecipanti ad una maratona possono essere rappresentati come:

- 1) Tempo impiegato per raffigurare il traguardo \Rightarrow **VARIABILE METRICA (NUMERICA)**
- 2) Posizione di arrivata (primo, secondo, terzo, ...) \Rightarrow **VARIABILE ORDINALE**
- 3) Nome del team di appartenenza \Rightarrow **VARIABILE NOMINALE (CATEGORICA)**

VARIABILE METRICA (METRIC VARIABLE)

- Descrivono una quantità (es. tempo, altezza, temperatura, peso)
- È definito un ordinamento (si dice quale valore è "più grande" di un altro)
- È definita una distanza (si chiede "di quanto" un numero è più grande di un altro)

Un caso speciale di variabile metrica è una **VARIABILE CONTEGGIO (COUNT VARIABLE)**

L'esprime il numero di eventi occorsi (nel tempo o nello spazio)

L'Es. numero di macchine traslate al cesso in un'ora

VARIABILE ORDINALE

- Descrivono oggetti sui quali la scissione impone un ordine
- Non ha senso chiedersi "di quanto" un valore è più grande di un altro
- Es. Posizionamento in una corsa (primo, secondo, ...)

L'ordine di confidenza su un argomento ($0 = \text{non confidante}$; $1 = \text{per confidante}$, $-s = \text{molte confidante}$)

L'affidabilità di un voto (x_S, s, M, L, x_L)

VEDI RETRO

VARIABILE CATEGORICA

- Descrivono delle categorie di appartenenza
- Non ha senso impostare un ordinamento
- " " " chiedersi di quanto un valore è più grande di un altro
- Es

L'sessualità (M/F)

L'affiliazione politica (REPUBBLICANA/DEMOCRATICA)

L'odore degli occhi (BLU, VERDE, NERONE)

8.5

PERCHÉ A INTERESSA IL TPO DI VARIABILE?

Ci interessa perché, a seconda del tipo di variabile, utilizziamo un modello appropriato per quel tipo

Es

- 1) VARIABILE METRICA: $y \sim N(\mu, \sigma^2)$
- 2) VARIABILE COUNT: $y \sim \text{Poisson}(\lambda)$ λ : # modi eventi nell'unità di tempo
- 3) VARIABILE CATEGORICA
DICOTOMICA: $y \sim \text{Bernoulli}(\pi)$ π : probabilità che $y=1$
 $(y=0, y=1)$

Esistono modelli più complessi per dati ordinali.

Consiglio

Uno degli step iniziali per sviluppare un modello dei dati è DETERMINARE LA TIPLOGIA DELLE MISURE in gioco

Il metodo dei minimi quadrati mimimizza le somme quadratiche tra i dati ed il modello

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N (y_{(i)} - \varphi_{(i)}^T \theta)^2 = \sum_{i=1}^N e_i^2$$

Vogliamo il $\hat{\theta}$ che minimizza queste quantità

$$\nabla J(\theta) = \frac{dJ(\theta)}{d\theta} = 0 \Rightarrow \frac{2}{N} \sum_{i=1}^N \varphi_{(i)} \cdot (y_{(i)} - \varphi_{(i)}^T \theta) = 0 \Rightarrow \sum_{i=1}^N \varphi_{(i)} y_{(i)} - \sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^T \theta = 0$$

$$\Rightarrow \left[\sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^T \right] \theta = \sum_{i=1}^N \varphi_{(i)} y_{(i)} \Rightarrow \hat{\theta} = \left[\sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^T \right]^{-1} \left[\sum_{i=1}^N \varphi_{(i)} y_{(i)} \right]$$

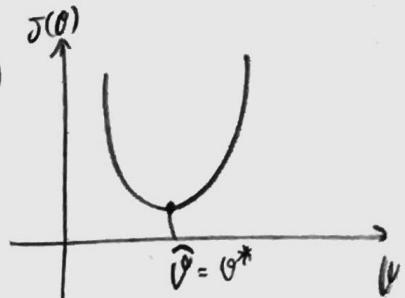
Osservazioni

- Se $\det \left[\sum_{i=1}^N \varphi_{(i)} \varphi_{(i)}^T \right] \neq 0$ la soluzione è unica!

- - - - - $= 0$, \exists INFINITE SOLUZIONI

- Dato che il modello è lineare e le funzioni di costi quadratiche, essa sono una forma quadratiche di $J(\theta)$.

L si dimostra che $\hat{\theta}$ è MINIMO GLOBALE di $J(\theta)$



MINIMI QUADRATI - NOTAZIONE MATRICIALE

$$X = \begin{bmatrix} 1 & x_1(1) & x_2(1) & \cdots & x_d(1) \\ 1 & x_1(2) & x_2(2) & \cdots & x_d(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1(N) & x_2(N) & \cdots & x_d(N) \end{bmatrix}$$

$\varphi(1)^T$

ogni colonna è un regressore / features

$$X = \begin{bmatrix} \varphi(1)^T \\ \varphi(2)^T \\ \vdots \\ \varphi(N)^T \end{bmatrix}$$

$$Y = X\theta + E \Rightarrow J(\theta) = \frac{1}{N} \| Y - X\theta \|^2 = \frac{1}{N} (Y - X\theta)^T (Y - X\theta)$$

$$\|X\|_F = \sqrt{x_1^2 + x_2^2 + \dots + x_d^2}$$

$$Y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} \quad E = \begin{bmatrix} e(1) \\ e(2) \\ \vdots \\ e(N) \end{bmatrix}$$

$$\nabla_{\theta} (J(\theta)) = (A + A^T)\theta \quad \nabla_{\theta} (J(\theta)) = \frac{1}{N} (Y^T Y - Y^T X \theta) - (X^T X^T \cdot Y) + (X^T X^T \cdot \theta)$$

$$\nabla J(\theta) = 0$$

$$\Rightarrow \frac{1}{N} \left(-2X^T Y + 2X^T X \theta \right) = 0 \Rightarrow$$

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

$$\nabla_{\theta} (-) = (X^T X + (X^T X)^T) / N - 2X^T X \theta$$

Come si computa lo stimatore a tenere conto delle opere nello (modello lineare) nel caso in cui il sistema vero sia effettivamente lineare?

$$y(i) = \varphi(i)^T \theta^* + v(i) \quad (\theta^*: \text{valore vero dei parametri})$$

- Supponendo $v(i)$ un numero costante di valori nulli $E[v(i)] = 0$

$$\downarrow \quad E[\hat{\theta}] = \theta^* \quad \text{CORRETTO}$$

- Supponendo inoltre che i numeri siano indipendenti: $E[v(i)v(j)] = 0 \quad \forall i \neq j$
e varianza $\sigma^2 \rightarrow \text{Var}[v(i)] = \sigma^2$

$$\text{Var}[\hat{\theta}] = \sigma^2 \cdot \left[\sum_{i=1}^N \varphi(i) \varphi(i)^T \right]^{-1} \quad \text{CONSISTENTE}$$

Es θ^* scalare

$$\begin{aligned} S: y(i) &= x(i)\theta^* + v(i) \\ T: y(i) &= x(i)\theta + e(i) \Rightarrow J(\theta) = \frac{1}{N} \sum_{i=1}^N (y(i) - x(i)\theta)^2 \\ \frac{dJ(\theta)}{d\theta} &= 0 \rightarrow -\frac{2}{N} \sum_{i=1}^N (y(i) - x(i)\theta) x(i) = 0 \\ &\Rightarrow \sum_{i=1}^N (y(i)x(i) - x(i)^2\theta) = 0 \Rightarrow \sum_{i=1}^N y(i)x(i) - \sum_{i=1}^N x(i)^2\theta = 0 \\ &\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^N y(i)x(i)}{\sum_{i=1}^N x(i)^2} \\ E[\hat{\theta}] &= E\left[\frac{\sum_{i=1}^N (y(i)x(i))}{\sum_{i=1}^N x(i)^2}\right] = \frac{\sum_{i=1}^N E[y(i)x(i)]}{\sum_{i=1}^N x(i)^2} = \frac{\sum_{i=1}^N E[x(i)\theta^* + v(i)x(i)]}{\sum_{i=1}^N x(i)^2} \\ &= \frac{\sum_{i=1}^N (x(i)\theta^* + 0)x(i)}{\sum_{i=1}^N x(i)^2} = \boxed{\theta^*} \end{aligned}$$

$$\text{Var}[\hat{\theta}] = \frac{\lambda^2}{\sum_{i=1}^N x(i)^2}$$

STIMA A MASSIMA VEROSSIGLIANZA

Ottavans presentato finora due tipi di stimatori,

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N y(i) \quad \text{media campionaria} \Rightarrow \hat{\theta} = \mu \in \mathbb{R}$$

$$L S^2 = \frac{1}{N-1} \sum (y(i) - \hat{\mu})^2 \quad \text{varianza campionaria} \Rightarrow \hat{\sigma}^2 = S^2 \in \mathbb{R}$$

$$\begin{aligned} L \text{ STIMA MIN. QUADRATI } y(i) &= \theta_0 + \theta_1 x_1(i) + \theta_2 x_2(i) + \dots + \theta_d x_d(i) + \epsilon(i) \\ &\Rightarrow \hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_d]^T \in \mathbb{R}^{d+1} \end{aligned}$$

Ottavans presentato stimatori PARAMETRICI, avendo rappresentato i dati tramite un modello parametrico (es. modello lineare)

L'Non ottavans mai fanno assunzioni sulle pdf dei dati $D = \{y(i), \dots, y(n)\}$

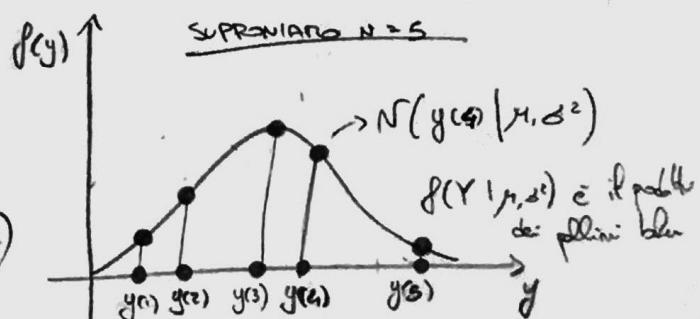
Il metodo della MASSIMA VEROSSIGLIANZA è una procedura di stima che, dato un modello probabilistico, stima i suoi parametri in modo che siano il più possibile consistenti con quanto osservato.

Supponiamo di avere $\mathbf{Y} = [y(1), \dots, y(N)]^T$: N osservazioni della variabile scobie y

L' $y(i) \sim N(\mu, \sigma^2)$ i.i.d.

La pdf del vettore dei dati è:

$$f(y(1), y(2), \dots, y(N) | \mu, \sigma^2) = \prod_{i=1}^N N(y(i) | \mu, \sigma^2)$$



- È la prob. che si realisi il vettore di dati osservato

L'siccome $y(i) \text{i.i.d.}$, la prob. di osservare $y(1)$ AND $y(2)$ AND ... è il prodotto delle varie pdf delle singole voci

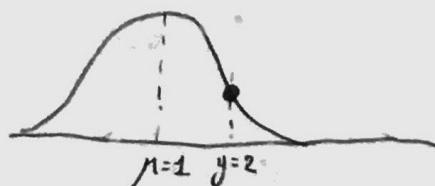
L' se in funzione della \mathbf{Y} , è una pdf N-dimensionale

però io so il valore vero della $y(i) \Rightarrow$ se conosco anche μ e σ^2 , posso calcolare il valore osservato della pdf

- Quand' questa funzione è vista in funzione di μ e σ^2 (conoscendo le \mathbf{Y}), allora prende il nome di VEROSIMIGLIANZA (LIKELIHOOD)

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{(y-\mu)}{\sigma} \right)^2} = f(y | \mu, \sigma^2)$$

NUMERO NOTO FUNZIONE DI
y

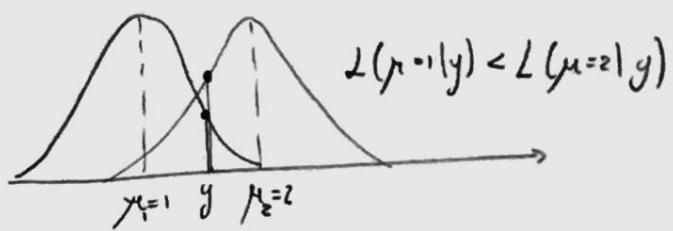


$$\Rightarrow L(\mu, \sigma^2 | y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{(y-\mu)}{\sigma} \right)^2}$$

(11)

$$L(\mu, \sigma^2 | y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2}$$

SUPPONIAMO σ^2 NOTO $\rightarrow L(\mu | y)$



Lo stimatore Massima Verosimiglianza è quel valore del parametru θ che massimizza $L(\theta | y)$

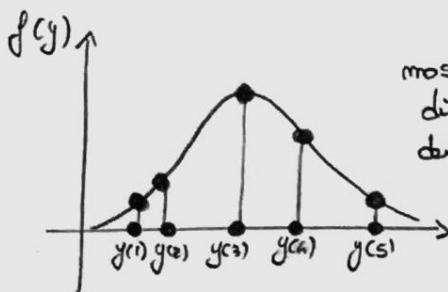
L'ad esempio, $\theta = \mu$ (σ^2 noto) $\Rightarrow \mu = 2$ è più verosimile di $\mu = 1$ perché $f(y | \mu = 1) < f(y | \mu = 2)$

In questo caso, lo stima più verosimile sarà $\mu = y$



Quindi, nel caso di più osservazioni $y^{(1)}, \dots, y^{(N)}$ i.i.d., $Y = [y^{(1)}, \dots, y^{(N)}]^T$, deve massimizzare $L(y^{(1)}, \dots, y^{(N)} | \mu, \sigma^2) = L(\mu, \sigma^2 | Y) = \prod_{i=1}^N N(y^{(i)} | \mu, \sigma^2)$

SUPPONIAMO $N = 5$



massimizzare le verosimiglianze vuol dire approssimare μ e σ^2 t.c. il prodotto dei pallini belli è max

$$\hat{\theta}_m = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \underset{\theta}{\operatorname{argmax}} L(\theta | Y) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^N N(y^{(i)} | \theta)$$

In genere per attribuire ai dati qualsiasi pdf $f(Y | \theta)$

$$\boxed{\hat{\theta}_m = \underset{\theta}{\operatorname{argmax}} L(\theta | Y) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^N f(y^{(i)} | \theta)}$$

Spesso, anziché massimizzare $L(\theta | Y)$, si massimizza il suo logaritmo naturale

L'atto de il logaritmo è una funzione monotona crescente, ha lo stesso massimo di $L(\theta | Y)$

L'è efficiente del punto di vista implementativo, perché evita l'indebolire del prodotto di piccole probabilità (sostituendolo con la somma delle log-probabilità)

$$\boxed{\hat{\theta}_m = \underset{\theta}{\operatorname{argmax}} \ln [L(\theta | Y)]}$$

Soltamente queste stime hanno viene effettuate con metodi numerici iterativi



In casi così si può fare analiticamente (Gaussiano, ...)

STIMA DI PARAMETRI DI UNA POPOLAZIONE:

ES: Supponiamo che siano disponibili i punti delle popolazione delle y

Siamo $y(i) \sim N(\mu, \sigma^2)$ i.i.d. \Rightarrow Trovare la stima max verosimiglianza di $\theta = [\mu, \sigma^2]$

$$f(y(i)|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y(i)-\mu}{\sigma}\right)^2} \Rightarrow \text{i.i.d.} \Rightarrow L(\underbrace{\mu, \sigma^2}_{\theta} | y(1), \dots, y(N)) = \prod_{i=1}^N f(y(i)|\mu, \sigma^2)$$

$$L(\theta|Y) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y(i)-\mu}{\sigma}\right)^2} \xrightarrow{\text{Log}} \ln[L(\theta|Y)] = \sum_{i=1}^N \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y(i)-\mu}{\sigma}\right)^2} \right].$$

$$= \sum_{i=1}^N \left(\ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] + \ln \left[e^{-\frac{1}{2}\left(\frac{y(i)-\mu}{\sigma}\right)^2} \right] \right) = \sum_{i=1}^N \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] + \sum_{i=1}^N \ln \left[e^{-\frac{1}{2}\left(\frac{y(i)-\mu}{\sigma}\right)^2} \right].$$

$$= N \cdot \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] + \sum_{i=1}^N -\frac{1}{2} \left(\frac{y(i)-\mu}{\sigma} \right)^2 (\ln e) = N \cdot \ln \left[2\pi\sigma^2 \right]^{\frac{1}{2}} - \frac{1}{2} \sum_{i=1}^N \left(\frac{y(i)-\mu}{\sigma} \right)^2 =$$

$$-\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^N \left(\frac{y(i)-\mu}{\sigma} \right)^2 = \boxed{-\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^N (y(i)-\mu)^2}$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta|Y) \xrightarrow{\downarrow} \begin{cases} \frac{\partial L(\mu, \sigma^2|Y)}{\partial \mu} = 0 \\ \frac{\partial L(\mu, \sigma^2|Y)}{\partial \sigma^2} = 0 \end{cases} \Rightarrow \begin{cases} +\frac{1}{\sigma^2} \sum_{i=1}^N (y(i)-\mu) = 0 \\ -\frac{N}{2} \cdot \frac{1}{\sigma^2} - \frac{1}{2} \sum_{i=1}^N (y(i)-\mu)^2 \cdot \left(-\frac{1}{\sigma^4}\right) = 0 \end{cases}$$

$$\frac{1}{\sigma^2} \sum_{i=1}^N (y(i)-\mu) = 0 \Rightarrow \sum_{i=1}^N (y(i)-\mu) = 0 \Rightarrow \sum_{i=1}^N y(i) - \sum_{i=1}^N \mu = 0 \Rightarrow \sum_{i=1}^N y(i) - N\mu = 0$$

$$\text{CORRETTO!} \Rightarrow \hat{\mu} = \frac{1}{N} \sum y(i)$$

$$\left\{ -\frac{N}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^N (y(i)-\mu)^2 = 0 \right. \xrightarrow{\text{sostituisci } \hat{\mu}} \left. \frac{1}{2\sigma^2} \sum_{i=1}^N (y(i)-\hat{\mu})^2 = \frac{N}{2} \cdot \frac{1}{\sigma^2} \right.$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{N} \sum (y(i)-\hat{\mu})^2$$

VARIANZA CAPITONARIA

DISTORTO!

Lo stimatore è massima verosimiglianza per essere dotato!

↓ In genere però, esser gode di buone proprietà

PROPRIETÀ STIMA MASSIMA VEROSIMIGLIANZA

- 1) Assintoticamente corretto: $\lim_{N \rightarrow +\infty} E[\hat{\theta}_n] = \theta^*$ Es. STIMATORE VARIANZA
 $\hat{\sigma}_{re}^2 = \frac{1}{N} \sum_{i=1}^N (y^{(i)} - \mu)^2$
 quando $N \rightarrow +\infty$, divida per N e per N non cambia
- 2) Costante: più N grande, + stime precise
- 3) Assintoticamente efficiente: $\lim_{N \rightarrow +\infty} \text{Var}[\hat{\theta}_n] = H^{-1}$ H : matrice di informazione di Fisher
- 4) Assintoticamente normale: $\hat{\theta}_n \sim N(\theta^*, \frac{1}{H})$ se $N \rightarrow +\infty$

L' $\hat{\theta}_n$ è centrato sul valore vero e fa volerla più ~~lontano~~ vicina all'informazione di Fisher

Esempio 1 - con numeri

Sia $y^{(i)} \sim N(\mu, \sigma^2 = 1)$, $i = 1, 2$, iid. Calcolare le stime di μ nel caso in cui i dati osservati sono:

$$y^{(1)} = 4 \quad y^{(2)} = 6$$

$$f(y|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (y-\mu)^2}$$

La densità imponendone delle due osservazioni è:

$$\begin{aligned} f(y^{(1)}=4, y^{(2)}=6 | \mu, \sigma^2 = 1) &= f(4 | \mu, \sigma^2 = 1) \cdot f(6 | \mu, \sigma^2 = 1) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (4-\mu)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (6-\mu)^2} \end{aligned}$$

La pdf condizionata (icit) è:

$$f(y^{(1)}=4, y^{(2)}=6 | \mu, \sigma^2 = 1) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (4-\mu)^2} \right) \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (6-\mu)^2} \right)$$

↓
È FUNZIONE SOLO DI μ !

Interpretando $\mathcal{L}(\mu | y_{(1)}=4, y_{(2)}=6 | \eta, \sigma^2 = 1)$ come funzione di μ , otteniamo
la verosimiglianza

$$\mathcal{L}(\mu | \underbrace{y_{(1)}=4}_{\Theta}, \underbrace{y_{(2)}=6}_Y) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right) \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right)$$

$$\hat{\mu} = \underset{\mu}{\operatorname{arg\ max}} \mathcal{L}(\mu | y_{(1)}=4, y_{(2)}=6)$$

Calcolare la log-verosimiglianza:

$$\begin{aligned} \ln[\mathcal{L}] &= \ln \left[\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right) \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right) \right] \\ &= \ln \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right] + \ln \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right] = \\ &= \ln \frac{1}{\sqrt{2\pi}} + \ln \left[e^{-\frac{1}{2}(4-\mu)^2} \right] + \ln \frac{1}{\sqrt{2\pi}} + \ln \left[e^{-\frac{1}{2}(6-\mu)^2} \right] \\ &= \cancel{\ln \frac{1}{\sqrt{2\pi}}} + -\frac{1}{2}(4-\mu)^2 \cancel{\ln e} - \frac{1}{2}(6-\mu)^2 \cancel{\ln e} \\ &= \overline{2 \ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(4-\mu)^2 - \frac{1}{2}(6-\mu)^2} \end{aligned}$$

Trovare il massimo:

$$\begin{aligned} \frac{\partial \ln[\mathcal{L}]}{\partial \mu} = 0 &\Rightarrow +\frac{2}{2}(4-\mu) + \frac{2}{2}(6-\mu) = 0 \Rightarrow \frac{4+6}{2} = 2\mu \\ &\Rightarrow \boxed{\hat{\mu} = \frac{4+6}{2} = 5} \end{aligned}$$

MEDIA
CAMPIONARIA!

E_s

• Colobae b. triunfo massiva verosimilmente nel corso in cui Notti e
pomeriggi devo le sue distribuzioni di Bernoulli con parametri

$$P(y| \pi) = \pi^y \cdot (1-\pi)^{1-y} \quad y=0,1 \quad \text{hence we denote } \pi \text{ as the probability of success}$$

$$\begin{aligned} \mathcal{L}(\pi | Y) &= \prod_{i=1}^N \pi^{y(i)} \cdot (1-\pi)^{1-y(i)} = \pi^{\sum_{i=1}^N y(i)} \cdot (1-\pi)^{\sum_{i=1}^N (1-y(i))} \\ &= \pi^{\text{# successes}} \cdot (1-\pi)^{\text{# failures}} \end{aligned}$$

Color be by-removing phrase

$$\begin{aligned}
 \ln L &= \ln \left[\pi \sum_{i=1}^N y(i) \cdot (1-\pi) \sum_{i=1}^N (1-y(i)) \right] = \ln \pi \sum_{i=1}^N y(i) + \ln (1-\pi) \sum_{i=1}^N (1-y(i)) \\
 &= \underbrace{\sum_{i=1}^N y(i)}_{\text{no. of ones}} \cdot \ln \pi + \underbrace{\sum_{i=1}^N (1-y(i))}_{\sum_{i=1}^N 1 - \sum_{i=1}^N y(i) = N - \gamma} \ln (1-\pi) = \gamma \ln \pi + (N-\gamma) \ln (1-\pi)
 \end{aligned}$$

Now f losses

$$\frac{\partial \ln[L]}{\partial \pi} = 0 \Rightarrow \frac{\gamma}{\pi} - \frac{(N-\gamma)}{1-\pi} = 0 \Rightarrow \frac{(\pi-\gamma)\gamma - \pi(N-\gamma)}{\pi(1-\pi)} = 0$$

$$\Rightarrow \gamma - \bar{y}f - \pi N + y\kappa = 0 \Rightarrow \boxed{\bar{c} = \frac{\gamma}{N} = \frac{1}{N} \sum_{i=1}^N y(i)}$$

~~MEDIA
CAMPIONADA~~

Osservazioni

e le % di successi

La distribuzione di Bernoulli $p(y|\pi)$ è una distribuzione DISCRETA. Infatti π è fisso AD UN VALORE ed il dato y è la variabile che assume solo 2 valori discreti: 0 e 1

La likelihood $L(\pi; Y) = \pi^y \cdot (1-\pi)^{n-y}$ è una funzione continua del parametro π che è continua tra $[0, 1]$. Non è una distribuzione perché non integrale a 1.

Osservazione

È importante notare che massimizzazione la log-likelihood equivale a minimizzazione la negativa log-likelihood

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{arg\,max}} \ln [\mathcal{L}(\theta | Y)] \\ = \underset{\theta}{\operatorname{arg\,min}} - \ln [\mathcal{L}(\theta | Y)]$$

In questo modo, abbiamo un problema di minimizzazione come con le regressioni lineari, dove minimizziamo (tramite il metodo dei minimi quadrati):

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N (y(i) - \varphi(i)^T \cdot \theta)^2$$

$$\hat{\theta} = \underset{\theta}{\operatorname{arg\,min}} J(\theta)$$

STIMAMASSIMA VEROSSIMILANZA DI MODELLI LINEARI

Come nel caso in cui non vi erano osservazioni sulla pdf dei dati, dobbiamo cercare degli stimatori per descrivere i dati con dei parametri delle loro popolazioni.

↳ possiamo usare il metodo ML anche nel caso in cui vogliamo descrivere i dati attraverso un modello lineare

$$y(i) = \theta_0 + \theta_1 x_1(i) + \theta_2 x_2(i) + \dots + (\theta_d x_d(i)) + e(i)$$

$$= \varphi(i)^T \theta + e(i)$$

$$\varphi(i) = \begin{bmatrix} 1 & x_1(i) & x_2(i) & \dots & x_d(i) \end{bmatrix}^T$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix}$$

$$e(i) \sim N(0, \lambda^2) \quad \text{(i.i.d.)}, \quad e(i) \perp \theta$$

$$\boxed{y(i) \sim N(\varphi(i)^T \theta, \lambda^2)}$$

La modellazione è esattamente la stessa funzione lineare dei regressori!

Le probabilità di osservare i dati misurati è data dalle probabilità compiate delle $y^{(i)}$:

$$f(\underbrace{y^{(1)}, \dots, y^{(N)}}_Y | X, \theta, \lambda^2) = \prod_{i=1}^N f(y^{(i)} | \varphi^{(i)}, \theta, \lambda^2) =$$

$$\begin{aligned} X &= \begin{bmatrix} \varphi^{(1)\top} \\ \varphi^{(2)\top} \\ \vdots \\ \varphi^{(N)\top} \end{bmatrix} = \prod_{i=1}^N N(\varphi^{(i)\top} \theta, \lambda^2) = \\ &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)\top} \theta}{\lambda} \right)^2} = L(\theta, \lambda^2 | Y, X) \end{aligned}$$

Supponiamo λ^2 noto per semplicità.

L'è verosimiglianza è funzione del vettore dei coefficienti $\theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_{d-1} \end{bmatrix}$

Calcola la log-verosimiglianza

$$\begin{aligned} \ln[L(\theta | X, Y)] &= \ln \left[\prod_{i=1}^N \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)\top} \theta}{\lambda} \right)^2} \right] \\ &= \sum_{i=1}^N \ln \left[\frac{1}{\sqrt{2\pi\lambda^2}} \cdot e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)\top} \theta}{\lambda} \right)^2} \right] = \sum_{i=1}^N \left(\ln \frac{1}{\sqrt{2\pi\lambda^2}} + \ln \left[e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)\top} \theta}{\lambda} \right)^2} \right] \right) \\ &= \sum_{i=1}^N \ln \frac{1}{\sqrt{2\pi\lambda^2}} + \sum_{i=1}^N \ln \left[e^{-\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)\top} \theta}{\lambda} \right)^2} \right] = N \cdot \ln(2\pi\lambda^2)^{-\frac{1}{2}} + \sum_{i=1}^N -\frac{1}{2} \left(\frac{y^{(i)} - \varphi^{(i)\top} \theta}{\lambda} \right)^2 \\ &= -\frac{1}{2} N \cdot \ln 2\pi\lambda^2 - \frac{1}{2\lambda^2} \sum_{i=1}^N (y^{(i)} - \varphi^{(i)\top} \theta)^2 = \boxed{-\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \lambda^2 - \frac{1}{2\lambda^2} \sum_{i=1}^N (y^{(i)} - \varphi^{(i)\top} \theta)^2} \end{aligned}$$

Calcola il massimo di $\ln[L(\theta | X, Y)]$ è uguale al calcolare il minimo di $-\ln[L(\theta | X, Y)]$

$$-\ln[L(\theta | X, Y)] = +\frac{N}{2} \ln 2\pi + \frac{N}{2} \ln \lambda^2 + \frac{1}{2\lambda^2} \sum_{i=1}^N (y^{(i)} - \varphi^{(i)\top} \theta)^2$$

NON DIPENDONO DA θ

$$\Rightarrow \boxed{\hat{\theta}_{ML} = \underset{\theta}{\operatorname{arg\,min}} \frac{1}{2\lambda^2} \sum_{i=1}^N (y^{(i)} - \varphi^{(i)\top} \theta)^2}$$

Osservazione

Le stime ML così ottenuta ha lo stesso minimo (è equivalente) che stima ottenuta con i minimi quadrati (in presenza di osservazioni possibili)

$$\hat{\theta}_{ML} = \operatorname{arg\min}_{\theta} \frac{1}{2N} \sum_{i=1}^N (y_{(i)} - p_{(i)}^T \theta)^2$$

$$\hat{\theta}_{LS} = \operatorname{arg\min}_{\theta} \frac{1}{N} \sum_{i=1}^N (y_{(i)} - p_{(i)}^T \theta)^2$$

\Rightarrow scarto per una costante (che esse sia $\frac{1}{2N}$ o $\frac{1}{N}$) non cambia il minimo delle funz. di cost.



Le stime ML del modello $y_{(i)} = p_{(i)}^T \theta + e_{(i)}$, dove $e_{(i)} \sim N(0, \sigma^2)$ iid, è equivalente alle stime LS.

↳ queste osservazioni di modello danno origine al modello di REGRESSIONE LINEARE

Osservazione

Considerando le ipotesi sulla distribuzione del rumore, si ottiene che funzione di cost. è quindi altri algoritmi, che modellano i dati in modo diverso delle regressioni lineare.

• REGRESSIONE LOGISTICA •

Il procedimento delle regressione lineare modellizza dati metrici attraverso un modello lineare, tramite l'ausilio di regressori (features)



Un problema frequente è la modellizzazione di dati CATEGORICI DICOTOMICI, in cui y assume valore 0 o 1. \Rightarrow Es:

- predire se una persona in un studio demografico sia maschio o femmina in base a peso e altezza
- predire cosa voterà una persona fra due candidati in base al reddito
- predire se un giocatore di baseball colpirà la pallina in base al suo ruolo



In questo caso, NON HA SENSO utilizzare il modello lineare $y_{(i)} = p_{(i)}^T \theta + e_{(i)}$

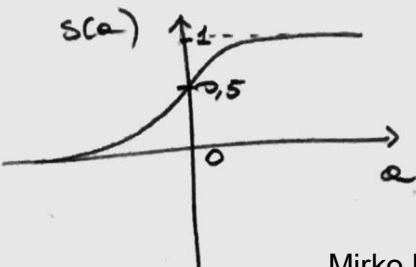
↳ non ha senso sommare un entità continuo (EPR) ad una variabile y che può assumere solo valori come 0 e 1, e non 0,98 o 1,01

↳ il modello potrebbe prendere anche valori <0 o >1! Non c'è niente che "limita" l'uscita \hat{y} tra 0 ed 1



quello che si fa è utilizzare la **FUNZIONE LOGISTICA (SIGMOIDE)**

$$s(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{1+e^a}$$



- se $a \gg 0 \Rightarrow s(a) = 1$
- se $a \ll 0 \Rightarrow s(a) = 0$

Mirko Mazzoleni - University of Bergamo

L'obiettivo di questi modelli è modellare la probabilità che $y=1$ tramite un modello lineare

Probabilità ↓

$$\leftarrow P(y=1 | \varphi) = s(\varphi^T \cdot \theta) = \frac{1}{1+e^{-(\varphi^T \cdot \theta)}}$$

L'output di $s(\varphi^T \theta)$ è interpretato come una probabilità

- se $\varphi^T \cdot \theta \gg 0 \Rightarrow P(y=1 | \varphi) = 1$
- se $\varphi^T \cdot \theta \ll 0 \Rightarrow P(y=1 | \varphi) = 0$

REGRESSIONE LINEARE

$$\mu = \varphi^T \theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d$$

$$y \sim N(\mu, \sigma^2)$$

REGRESSIONE LOGISTICA

$$\pi = s(\varphi^T \theta) = s(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d)$$

$$y \sim \text{Bernoulli}(\pi)$$

Sia la regressione lineare e la regressione logistica fanno parte dei cosiddetti GLM (Generalized Linear Model) in cui:

L'idea di un modello lineare è usata per modellare un parametruo di "tendenza centrale" delle distribuzioni dei dati.

Il termine "lineare" dice che modello il parametruo delle distribuzioni de mi indica il valore medio dei dati \Rightarrow non è sempre lo stesso! I dati y sono modellati tramite una distribuzione di probabilità in cui c'è il parametruo μ .

In generale: μ generiche funzione

$$\mu = f(\theta_0 + x_1 \theta_1 + x_2 \theta_2 + \dots + x_d \theta_d) = f(\varphi^T \theta)$$

$$y \sim \text{pdf}(\mu, [\text{altri parametri}])$$

REGRESSIONE LINEARE

$$\begin{aligned} \mu &= f(\varphi^T \theta) = \varphi^T \theta \quad (\text{funzione idoneità}) \\ \mu &= \mu \Rightarrow \mu = \varphi^T \theta \\ y &\sim N(\mu, \sigma^2) \end{aligned}$$

REG. LOGISTICA

$$\begin{aligned} \pi &= f(\varphi^T \theta) = s(\varphi^T \theta) \quad (\text{funzione logistica}) \\ \pi &= \pi \Rightarrow \pi = s(\varphi^T \theta) \\ y &\sim \text{Bernoulli}(\pi) \end{aligned}$$

STIMA MAXIMUM LIKELIHOOD DI UN MODELLO DI REGRESSIONE LOGISTICA

Sia dato un dataset $D = \{(\varphi(1), y(1)), (\varphi(2), y(2)), \dots, (\varphi(N), y(N))\}$
 $\varphi \in \mathbb{R}^{d_{\text{var}}}$, dove $y \in \{0, 1\}$.

Stimare un modello di regressione logistica $P(y=1 | \varphi) = \frac{1}{1+e^{-(\varphi^T \vartheta)}} = \hat{\pi}$

Interpretazione: i dati come $y \sim \text{Bernoulli}(\hat{\pi})$

Calcoliamo la verosimiglianza dei dati

$$P(y_{(i)}=1 | \varphi_{(i)}) = \frac{1}{1+e^{-(\varphi_{(i)}^T \vartheta)}} = \hat{\pi}_{(i)}$$

$$\mathcal{L}(\hat{\pi} | Y) = \prod_{i=1}^N \hat{\pi}_{(i)}^{y_{(i)}} \cdot (1-\hat{\pi}_{(i)})^{1-y_{(i)}} \Rightarrow \text{calcola la log-likelihood} \rightarrow \text{diventa una funzione da minimizzare}$$

$$Y = \begin{bmatrix} y_{(1)} \\ \vdots \\ y_{(N)} \end{bmatrix}$$

dipende dai parametri ϑ !!

$$\mathcal{L}(\hat{\pi} | Y) = \mathcal{L}(\vartheta | Y)$$

i veri parametri sono questi

$$\begin{aligned} -\ln[\mathcal{L}(\hat{\pi} | Y)] &= -\ln \left[\prod_{i=1}^N \hat{\pi}_{(i)}^{y_{(i)}} \cdot (1-\hat{\pi}_{(i)})^{1-y_{(i)}} \right] = \\ &= -\sum_{i=1}^N \ln \left[\hat{\pi}_{(i)}^{y_{(i)}} \cdot (1-\hat{\pi}_{(i)})^{1-y_{(i)}} \right] = -\sum_{i=1}^N \left(\ln \hat{\pi}_{(i)}^{y_{(i)}} + \ln [1-\hat{\pi}_{(i)}]^{1-y_{(i)}} \right) = \\ &= -\sum_{i=1}^N \left(y_{(i)} \ln \hat{\pi}_{(i)} + (1-y_{(i)}) \ln [1-\hat{\pi}_{(i)}] \right) = J(\vartheta) \end{aligned}$$

Interpretazione della funzione di costo

Supponiamo di avere un solo dato $D = \{(\varphi, y)\}$:

$$J(\vartheta) = \begin{cases} -\ln \hat{\pi} & \text{se } y=1 \\ -\ln [1-\hat{\pi}] & \text{se } y=0 \end{cases}$$

CASO $y=1$

$$J(\vartheta) = -\ln \hat{\pi} \Rightarrow$$

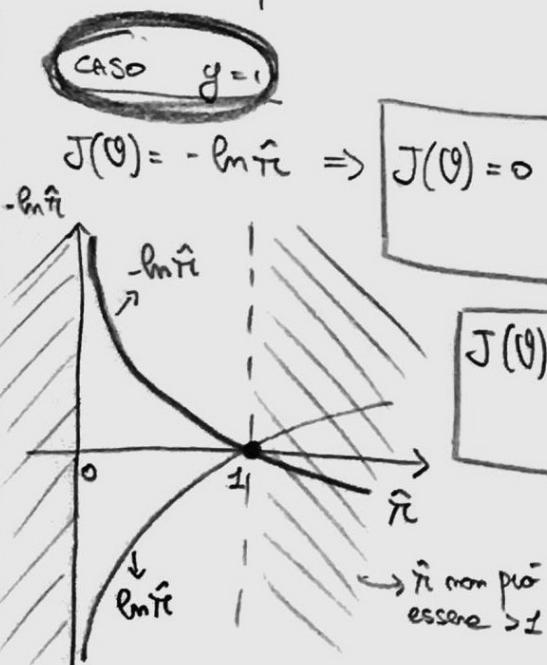
$$J(\vartheta) = 0 \quad \text{SE } \hat{\pi} = 1$$

Costo = 0 se predice giusto

Cuttura l'intuizione che se $y=1$, ma si predice una bassa probabilità che $y=1$, ovvero predice $\hat{\pi} \ll 1$ ($P(y=1|\vartheta) \ll 1$) allora commetto un grande sbaglio e $J(\vartheta) \rightarrow +\infty$ (perdita molto)

is vogli minimizzare questo sbaglio!

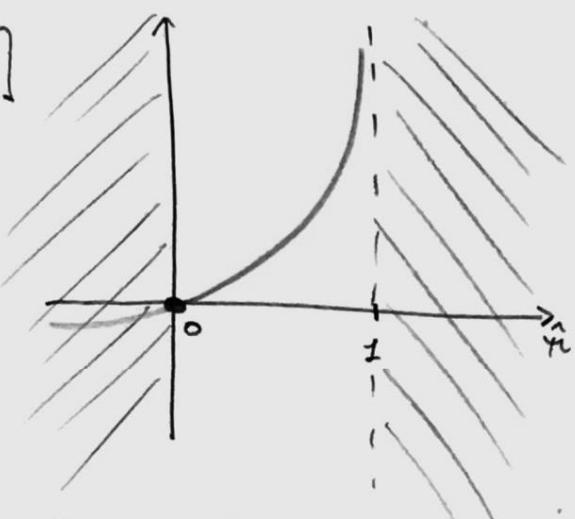
(21)



CASO $y=0$

$$J(\theta) = -\ln[1 - \hat{p}]$$

$$-\ln[1 - \hat{p}]$$



$$\boxed{J(\theta) = 0 \text{ SE } y=0 \text{ & } \hat{p}=0}$$

$$\boxed{J(\theta) = +\infty \text{ SE } y=0 \text{ & } \hat{p}=1}$$

Se $y=0$ ma si predice con alta probabilità che $y=1$, ovvero $\hat{p} \rightarrow 0$ ($P(y=1|p) \gg 0$) allora $J(\theta)$ molto e $J(\theta) \rightarrow +\infty$

CALCULO DEL MINIMO

Calcolare il gradiente di $J(\theta)$ rispetto al vettore di parametri $\theta \in \mathbb{R}^{d_k}$

Per prima cosa, calcolare la derivata di $s(a) = \frac{1}{1+e^{-a}}$

$$\begin{aligned} \frac{ds(a)}{a} &= \frac{d}{da} \left[\frac{1}{1+e^{-a}} \right] = \frac{d}{da} \left[(1+e^{-a})^{-1} \right] = -(1+e^{-a})^{-2} \cdot (e^{-a})(-1) = -(1+e^{-a})^{-2}(-e^{-a}) \\ &= \frac{-e^{-a}}{(1+e^{-a})^2} = \frac{e^{-a}}{(1+e^{-a})^2} = \frac{1}{(1+e^{-a})} \cdot \frac{e^{-a}}{(1+e^{-a})} = \frac{1}{(1+e^{-a})} \cdot \frac{(1+e^{-a})-1}{1+e^{-a}} \\ &= \underbrace{\frac{1}{1+e^{-a}}}_{s(a)} \cdot \left(\underbrace{\frac{1+e^{-a}}{1+e^{-a}}}_{1} - \underbrace{\frac{1}{1+e^{-a}}}_{s(a)} \right) = \boxed{s(a) \cdot [1 - s(a)]} \end{aligned}$$

uguale formazione

Nel caso in cui $a = \varphi^\top \theta \Rightarrow s(a) = s(\varphi^\top \theta) = \frac{1}{1+e^{-\varphi^\top \theta}}$

$$\begin{aligned} \frac{ds(\varphi^\top \theta)}{d\theta} &= \frac{d}{d\theta} \left[\frac{1}{1+e^{-\varphi^\top \theta}} \right] = \frac{d}{d\theta} \left[(1+e^{-\varphi^\top \theta})^{-1} \right] = \underset{dx_1}{\varphi_1} \cdot \underset{dx_1}{(-1)} \underset{dx_1}{(1+e^{-\varphi^\top \theta})^{-2}} \underset{dx_1}{(e^{-\varphi^\top \theta})} \\ &= -\varphi \cdot (1+e^{-\varphi^\top \theta})^{-2} (e^{-\varphi^\top \theta}) \end{aligned}$$

stesso passaggio:

$$\begin{aligned} &= \underset{dx_1}{\varphi} \cdot \underset{dx_1}{s(\varphi^\top \theta)} \underset{dx_1}{[1 - s(\varphi^\top \theta)]} \\ &= \boxed{\varphi \cdot \hat{p} \cdot (1 - \hat{p})} \end{aligned}$$

Possiamo ora calcolare il gradiente della $J(\theta)$

Mirko Mazzoleni - University of Bergamo

$$J(\theta) = \sum_{i=1}^N \left(y_{(i)} \ln \hat{p}_{(i)} + (1-y_{(i)}) \ln [1-\hat{p}_{(i)}] \right)$$

$$\hat{p}_{(i)} = \frac{1}{1+e^{-\varphi_{(i)}(\theta)}}$$

$$\begin{aligned}\nabla J(\theta) &= - \sum_{i=1}^N \left(y_{(i)} \frac{\hat{p}'_{(i)}}{\hat{p}_{(i)}} + (1-y_{(i)}) \frac{-\hat{p}'_{(i)}}{1-\hat{p}_{(i)}} \right) = \\ &= - \sum_{i=1}^N \left(y_{(i)} \cdot \cancel{\frac{\varphi'_{(i)} \cdot \hat{p}_{(i)} [1-\hat{p}_{(i)}]}{\hat{p}_{(i)}}} + (1-y_{(i)}) \cancel{\frac{\varphi'_{(i)} \hat{p}_{(i)} [1-\hat{p}_{(i)}]}{1-\hat{p}_{(i)}}} \right) \\ &= \sum_{i=1}^N \left(-y_{(i)} \varphi'_{(i)} [1-\hat{p}_{(i)}] + (1-y_{(i)}) (\varphi'_{(i)} \cdot \hat{p}_{(i)}) \right) \\ &= \sum_{i=1}^N \left(\varphi'_{(i)} [-y_{(i)} + y_{(i)} \hat{p}_{(i)}] + \varphi'_{(i)} [\hat{p}_{(i)} - y_{(i)} \hat{p}_{(i)}] \right) \\ &= \sum_{i=1}^N \left(\varphi'_{(i)} [-y_{(i)} + y_{(i)} \hat{p}_{(i)} - y_{(i)} \hat{p}_{(i)} + \hat{p}_{(i)}] \right) \\ &\boxed{= \sum_{i=1}^N \varphi'_{(i)} (\hat{p}_{(i)} - y_{(i)})}\end{aligned}$$

Osservazione

Le derivate $\sum_{i=1}^N \varphi'_{(i)} (\hat{p}_{(i)} - y_{(i)}) = 0$ sono un sistema di $|d|$ equazioni non lineari in θ

↳ Non si mettono mai soluzioni in forma chiusa come per la regressione lineare \Rightarrow per via delle non linearità della sigmoida

↳ si dimostra però che $J(\theta)$ è convessa, quindi ha un unico minimo

L'ottimizzazione è quindi solitamente algoritmi iterativi di ottimizzazione.
Uno di questi è il GRADIENT DESCENT:



il valore attuale dei parametri all'iterazione $j+1$ è:

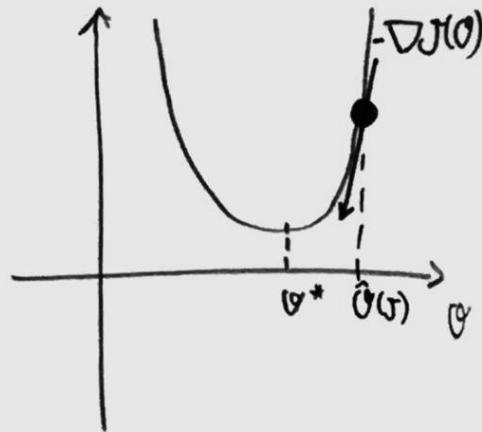
α è la LEARNING RATE (determina il passo con cui abbassano il valore)

↳ $\hat{\theta}(0)$ è inizialmente RANDOM

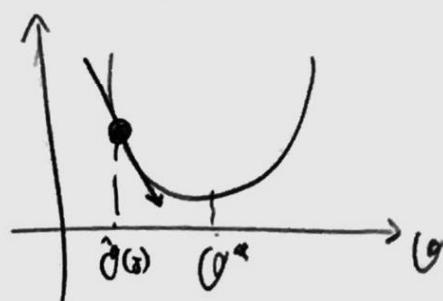
$$\hat{\theta}(j+1) = \hat{\theta}(j) - \alpha \nabla J(\theta) \Big|_{\theta=\hat{\theta}(j)}$$

$$\hat{\theta}(J+1) = \hat{\theta}(J) - \alpha \nabla J(\theta) \quad |_{\theta = \hat{\theta}(J)}$$

- se $\nabla J(\theta)|_{\theta=\hat{\theta}(J)} > 0 \Rightarrow \hat{\theta}(J+1) < \hat{\theta}(J)$



Mirko Mazzoleni - University of Bergamo



STIMA BAYESIANA (BAYESIAN INFERENCE)

PROBABILITÀ CONGIUNTE, CONDIZIONATE, MARGINALI

Supponiamo di avere 2 variabili casuali a e b , discrete bimode, con le seguenti distribuzioni di probabilità congiunte:

DISTRIBUZIONE CONGIUNTA

$P(a,b)$

	a	
b	0	1
0	0,06	0,24
1	0,28	0,42

$$\sum_{a,b} p(a,b) = 1$$

probabilità che si verifichi sia a che b , contemporaneamente

Le distribuzioni MARGINALI sono le distribuzioni di probabilità di un stato insieme di variabili casuali

↳ nel nostro caso, dato che abbiamo 2 variabili casuali, vi saranno 2 prob. marginali, ovvero $p(a)$ e $p(b)$

- È ottenuta "marginalmente", ovvero sommando, rispetto alle variabili che non sono di interesse

DISTRIBUZIONE TARGNALE

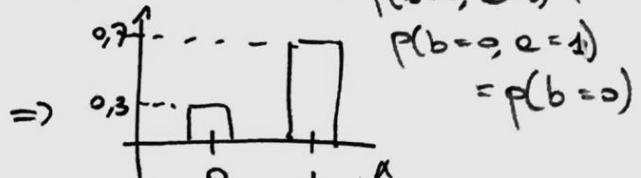
		$a=0$	$a=1$
		0	1
b	0	0,06	0,24
	1	0,28	0,42

$$P(b=0) = 0,3$$

$$P(b=1) = 0,7$$

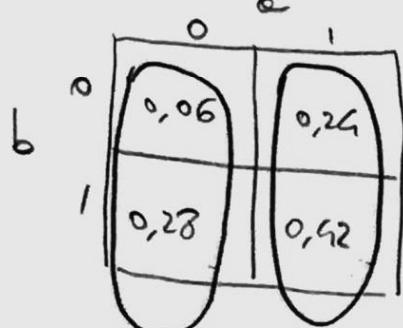
non mi interessa se $a=0$ o $a=1$.
 → interesse solo che $b=0$. Quindi la probabilità di $b=0$ è la somma delle probabilità quando

$$P(b) \quad b=0 \Rightarrow P(b=0, a=0) + P(b=0, a=1)$$



$$\Rightarrow \sum_b P(b) = 1$$

$$- P(b=1) = P(b=1, a=0) + P(b=1, a=1)$$

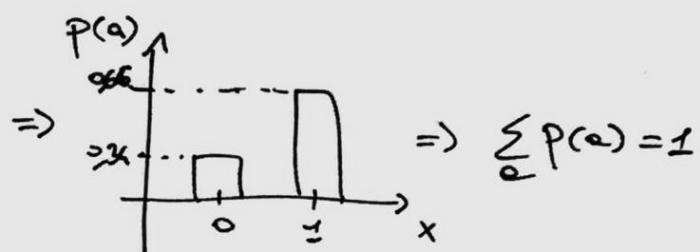


$$P(a=0) \quad P(a=1) = 0,66$$

||
0,34

$$- P(a=0) = P(a=0, b=0) + P(a=0, b=1)$$

$$- P(a=1) = P(a=1, b=0) + P(a=1, b=1)$$



$$\Rightarrow \sum_a P(a) = 1$$

La distribuzione condizionata indica come le probabilità si redistribuiscono dato che restringono la popolazione ad un particolare sottosinsieme.

Es

Siamo date N persone, dove N_A è il numero di persone con capelli lunghi e N_B è il numero di persone di sesso femminile. Siano gli eventi:

A : persone con capelli lunghi

B : persone di sesso femminile

$$P(A) = \frac{N_A}{N} = \frac{\# \text{di persone con capelli lunghi}}{\# \text{totale di persone}}$$

$$P(B) = \frac{N_B}{N} = \frac{\# \text{di donne}}{\# \text{totale di persone}}$$

Consideriamo la sotto popolazione femminile:

L'probabilità che una persona scelta a caso da questa popolazione abbia i capelli lunghi è $\frac{N_{AB}}{N_B}$, dove N_{AB} è il numero di donne con capelli lunghi.

L'questa probabilità è chiamata probabilità condizionata (al fatto che le persone sia di sesso femminile)

$$P(A|B) = \frac{N_{AB}}{N_B}$$

1/ La popolazione considerata è N_B , non N

25

- La probabilità di selezionare una donna ^{con} capelli biondi è $P(A, B) = \frac{N_{AB}}{N}$
- = $\frac{\text{numero di donne con capelli biondi}}{\text{totale di persone}}$
- ↓

- Posso esprimere $P(A|B)$ come: $P(A|B) = \frac{N_{AB}/N}{N_B/N} = \frac{P(A, B)}{P(B)}$
- ↓

Quindi: $P(A|B) = \frac{P(A, B)}{P(B)} \Rightarrow \boxed{P(A, B) = P(A|B)P(B)}$

Osservazione

- La probabilità che accade sia A che B è la probabilità che si verifichino entrambi moltiplicata per la probabilità che si verifichi A dato che B si è verificato
- $P(A, B) = P(A) \cdot P(B)$ se e solo se $P(A|B) = P(A)$. Questo vuol dire che A e B sono indipendenti, ovvero il verificarsi di B non modifica la probabilità che A si verifichi.

↳ Esempio: A: lancio un dadi ed esce 4
 B: lancio una moneta ed esce TESTA \Rightarrow anche se usciva croce, il dadi ha la stessa probabilità ($\frac{1}{6}$) di risultare impari

$$\therefore P(A, B) = P(A) \cdot P(B)$$

- Sappiamo che $P(A, B) = P(B, A)$. Quindi: $P(B|A) = \frac{P(B|A)P(A)}{P(A)}$, e di conseguenza:

$$P(A|B)P(B) = P(B|A)P(A) \Rightarrow P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

TEOREMA DI BAYES

- Il teorema di Bayes permette di ridistribuire le probabilità: prima conoscere $P(B)$, adesso $P(B|A)$ \Rightarrow la probabilità di B è cambiata in seguito alla conoscenza di A
- $P(A) = \sum_B P(A|B)P(B)$ è la marginal di A, ovvero somma rispetto a tutti i valori di B

(26)

Riprendendo l'esempio della tabella; calcoliamo la distribuzione $p(a|b)$

a	0	1
0	0,06	0,8
1	0,4	0,92

$$p(a=1|b=0) = \frac{p(a=1, b=0)}{p(b=0)} = \frac{0,24}{0,3} = 0,8$$

Mirko Mazzoleni - University of Bergamo

$$p(a|b) = \frac{p(a, b)}{p(b)}$$

$$p(a=0|b=1) = \frac{p(a=0, b=1)}{p(b=1)} = \frac{0,28}{0,7} = 0,4$$

Oltre stessa modo possi calcolare $p(b|a)$.

$$p(b=1) = p(b=1|a=0)p(a=0) + p(b=1|a=1)p(a=1)$$

Esempio : Interpretazione delle probabilità condizionate come ridistribuzione delle probabilità

Supponi di tirare bendato una freccette contro un bersaglio con 20 cerchi concentrici:



Le probabilità di beccare un cerchio qualsiasi, senza vedere, supponiamo sia $\frac{1}{20}$ (ogni cerchio è equiprobabile)

- Quel è la probabilità di aver beccato il cerchio numero 5?

$$P(\text{cerchio } \#5) = \frac{1}{20}$$

Supponiamo che si obietti b dice che NON HA PREZZO il cerchio $\#20$.

- Quel è obietto la probabilità di aver beccato il cerchio $\#5$?

Dato che non lo sicuramente per il $\#20$, la probabilità di aver per il $\#5$ è $P(\#5 | \text{NOT } \#20) = \frac{1}{19}$, perché le 19 valori possibili, escludere escluso 1



Le probabilità si è quindi ridistribuita sui 19 esiti restanti dei 20 esiti.

$$P(\#5 | \text{NOT } \#20) = \frac{P(\#5, \text{NOT } \#20)}{P(\text{NOT } \#20)} = \frac{P(\#5)}{P(\text{NOT } \#20)} = \frac{\frac{1}{20}}{\frac{19}{20}} = \boxed{\frac{1}{19}}$$

INTRODUZIONE ALLA STIMA BAYESIANA

Ottaviano finora considerato il parametro ignoto θ come una variabile deterministica.
Spero però, ottavano delle informazioni, delle credenze, sui possibili valori che potrebbe avere θ .

↳ Esempio:

↳ stima delle concentrazioni di anidride solforosa nell'aria: si ha un'idea dell'ordine di grandezza, in base anche a studi precedenti.

↳ stima della probabilità di una moneta non truccata risulti TESTA dopo un buco: si sa che non può essere 0,1 o 0,9 ma sarà attorno agli 0,5

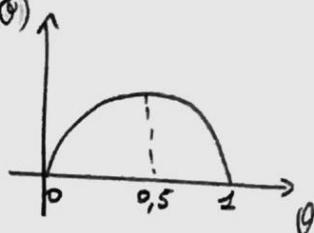
Ha quindi senso considerare θ come una variabile casuale anche come variabile deterministica.

↳ In questo modo posso specificare una distribuzione di probabilità per θ (dato che è una v.c.), assegnando una probabilità maggiore a valori di θ che io credo siano più verosimili che θ assuma, e minor probabilità a valori di θ che io credo non potranno accadere.

Esempio

Sia θ la probabilità di il lato di una moneta non truccata risulti in TESTA.
Una possibile distribuzione per θ è: $P(\theta)$

Osservazioni



- $P(\theta)$ ha dominio $[0, 1]$, poiché θ , modello di una probabilità, deve stare tra 0 ed 1
- Dato che la moneta è non truccata, $\theta=0,5$ sarà il valore più probabile di θ , e $\theta=0$ o $\theta=1$ sono praticamente impossibili (la probabilità che θ sia 0 o 1 è vicina a 0)
- Dato che distribuzione su θ , ottavo già una stima di θ (STIMA APRIORI). Ad esempio posso prendere come valore centrale per la stima di θ il valore ottenuto da $P(\theta)$. L'incertezza sulla stima sarà allora la varianza di $P(\theta)$ (INCERTITÀ A PRIORI)
- Con l'arrivo di dati osservati, ci si aspetta che:
 - 1) Il valore ottenuto cambi
 - 2) L'incertezza decrese (cioè più informazioni!)

Obbligatori quindi due elementi da poter informazione:

- 1) La distribuzione sui possibili valori di θ , ovvero $P(\theta)$
- 2) L'informazione da poter: i dati sui possibili valori di θ , ovvero le likelihood $P(Y|\theta)$

Quello che vogliere veramente è sapere quanto θ dato le osservazioni dati: $P(\theta|Y)$

Usando il Teorema di Bayes posso avere: due elementi di informazione:

$$P(\theta|Y) = \frac{P(Y|\theta) P(\theta)}{P(Y)}$$

- $P(\theta)$: PRIOR
- $P(Y|\theta)$: LIKELIHOOD
- $P(Y)$: MARGINAL LIKELIHOOD
- $P(\theta|Y)$: POSTERIOR

Osservazione

- $P(\theta|Y)$ è una distribuzione di possibile valori di θ , le cui probabilità sono modificate (riallocate, ridistribuite), rispetto a $P(\theta)$, dell'aver osservato: dati Y
- Nel caso in cui $P(Y|\theta)$ e $P(\theta)$ siano pdf continue (es. Gaussiane), allora $P(Y)$ sarà $P(Y) = \int_{-\infty}^{\infty} P(Y|\theta) P(\theta) d\theta$
- Conoscere le forme funzionali di $P(\theta)$ e $P(Y|\theta)$ perché le imposto io. Come posso dire su che distribuzione sono $P(\theta|Y)$?
 - 1) In genere, nullo. Solitamente in tali casi l'integrale di $P(\theta|Y)$ è in forme funzionali note
 - 2) Questo avviene se, ad esempio, $P(\theta)$ è Gaussiana e $P(Y|\theta)$ è Gaussiana. Allora anche $P(\theta|Y)$ sarà Gaussiana
 - 3) Un altro problema è che $P(Y)$ è un integrale da potremmo non sapere come risolvere.

Per far fronte a questi problemi, si usano metodi numerici e di compromesso che evitano R calcoli analitici. Questi metodi si chiamano MARKOV CHAIN MONTE CARLO (MCMC)

Un modo per calcolare $P(\Theta|Y)$ da cui si basa nei sui metodi MCMC è quello di discretizzare il range di valori del parametro Θ tramite una griglia di valori θ → valori $P(Y|\theta)$ e $P(\theta)$ solo in quei valori di θ

Esempio

Stimare la probabilità che il buco si sia moneta risulti in TESTA. Supponiamo di lanciare una moneta N volte. Osserviamo i dati $y_{(i)}$:

$$y_{(i)} = \begin{cases} 1 & \text{se TESTA} \\ 0 & \text{se CROCE} \end{cases} \quad i = 1, \dots, N$$

Modelliamo i dati, categordici e dicotomici, con una distribuzione di Bernoulli:

$y_{(i)} \sim \text{Bernoulli}(\pi)$, iid., π : Prob. TESTA (parametro ignoto)

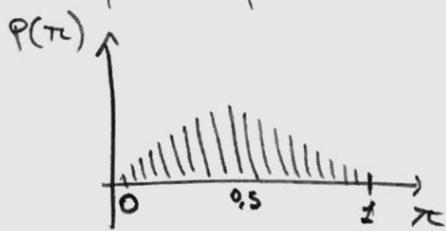
$$P(Y|n) = \pi^y \cdot (1-\pi)^{1-y}$$

- Se $y=1 \Rightarrow P(Y=1|n) = \pi$

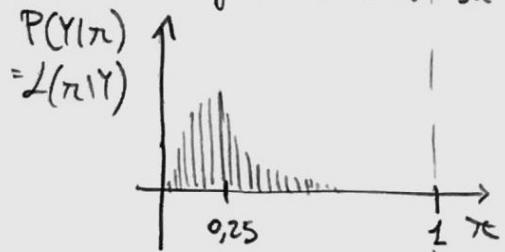
- se $y=0 \Rightarrow P(Y=0|n) = 1-\pi$

$$\begin{aligned} L(\pi|Y) &= \prod_{i=1}^N \pi^{y_{(i)}} \cdot (1-\pi)^{1-y_{(i)}} \\ &\stackrel{\Psi[y_{(i)}, \dots, y_{(N)}]}{=} \frac{\pi^{\sum y_{(i)}}}{\pi^N \cdot (1-\pi)^{\sum 1-y_{(i)}}} = \pi^{\text{#successi}} \cdot (1-\pi)^{\text{#fallimenti}} \end{aligned}$$

- Supponiamo una prior di questo tipo



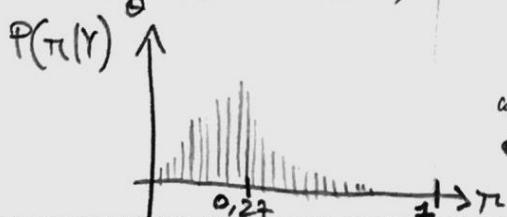
- Supponiamo di aver osservato $y=10$ successi su $N=40$ buchi. La likelihood ha la forma:



Il valore più probabile del parametro π è la stima ML. Nel caso di likelihood Bernoulli, lo che $\hat{\pi} = \frac{1}{N} \sum_{i=1}^N y_{(i)}$, ovvero le % di successi.

$$\text{In questo caso } \hat{\pi} = \frac{10}{40} = 0,25$$

- Per calcolare la posterior faccio il quoziente di $P(Y|\theta)$ e $P(\theta)$ per ogni valore di θ e dividere per $P(Y) = \sum_{\theta} P(Y|\theta)P(\theta)$, che somma su ogni valore di θ (della griglia)



La MODA è un compromesso tra 0,25 e 0,5

DISTRO
=>

- L'opera a griglia con ϵ generalizzabile nel caso in cui θ sia un vettore con molte componenti



Il PC si impiegherebbe troppo a fare tutte le combinazioni di parametri

- Se si ha invece o usare prior e likelihood tali che le posteriori hanno la stessa forma che si può ricavare analiticamente (se le posteriori hanno le stesse forme delle priori, prior e likelihood si dicono conjugate)
L'opera us metodo MCMC

Supponiamo di avere $P(\theta|Y)$. Ossiamo una distribuzione di valori del parametro θ ignoto. Ci sono però un valore sol, un valore puntuale

de valore puntuale per la nostra stima $\hat{\theta}$?

Ci sono varie possibilità:

1) $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(\theta|Y)$, ovvero prendere il valore θ di misulta essere più probabile

Questa stima è nota come stima MAXIMUM A POSTERIORI (MAP)

2) $\hat{\theta} = E[P(\theta|Y)] = E[\theta|Y]$, la MEDIA delle distribuzioni a posteriori

3) Altre quantità come la MEDIANA, ecc.

Ricordiamo che in genere individuiamo un stimatore come una funzione T dei dati D :

$$\hat{\theta} = T(D)$$

Vogliamo che la variabile casuale $\hat{\theta}$ sia vicina alla variabile casuale θ . Usiamo quindi le frequenze di costo:

$$J(T(\cdot)) = E[\|\theta - T(D)\|^2] \quad (*) \quad \text{MEAN SQUARED ERROR}$$

Lo stimatore ottimo di Bayes è quella funzione $T^*(\cdot)$ tale che:

$$E[\|\theta - T^*(D)\|^2] \leq E[\|\theta - T(D)\|^2] \quad \forall T(\cdot)$$

cioè che minimizza le frequenze di costo rispetto a $T(\cdot)$

Si dimostra che $T^*(Y) = E[\theta | D=Y]$, ovvero il valore ottenuto dalla

distribuzione $P(\theta|Y)$, cioè il valore ottenuto considerando di fatto che i dati D abbiano assunto valore Y



Considereremo quindi $E[\theta|Y]$ come stima puntuale di $\hat{\theta}$, soprattutto in che senso essa è una stima ottima

nel senso che

minimizza $(*)$

Supponiamo ora che sia: θ che il parmetro θ siano delle v.c. Gaussiane,
Quindi la loro pdf conjunta è Gaussiana.

↓

Supponiamo per semplicità di avere un dato scobe y e che θ sia scobe, tali che
 $E[y] = 0$ e $E[\theta] = 0$

Vogliamo calcolare $P(\theta | y)$. Essendo θ e y conjuntamente Gaussiane si ha che:

$$\begin{bmatrix} y \\ \theta \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_{yy} & \lambda_{y\theta} \\ \lambda_{\theta y} & \lambda_{\theta\theta} \end{bmatrix}\right) \quad \mu: \text{vettore media}$$

$z \in \mathbb{R}^{2 \times 1}$ $\mu \in \mathbb{R}^{2 \times 1}$ $\Sigma \in \mathbb{R}^{2 \times 2}$

Σ : matrice varianze-covarianze

Le pdf conjunta $P(\theta, y)$ ha quindi la forma:

$$P(\theta, y) = \frac{1}{\sqrt{(2\pi)^2 \det \Sigma}} e^{-\frac{1}{2} (z - \mu)^T \Sigma^{-1} (z - \mu)}$$

Le pdf dei dati è $P(y) = \frac{1}{\sqrt{2\pi \lambda_{yy}}} e^{-\frac{1}{2\lambda_{yy}} (y - 0)^2}$

Si dimostra che $P(\theta | y) = \frac{P(\theta, y)}{P(y)}$ è una Gaussiana, $P(\theta | y) = N(\mu_{\theta | y}, \lambda_{\theta | y})$, con:

- VALORE ATTESO:

$$\boxed{\mu_{\theta | y} = \frac{\lambda_{\theta y}}{\lambda_{yy}} \cdot y}$$

- VARIANZA:

$$\boxed{\lambda_{\theta | y} = \lambda_{\theta\theta} - \frac{\lambda_{\theta y}^2}{\lambda_{yy}}}$$

Il valore $\frac{\lambda_{\theta y}^2}{\lambda_{yy}}$ è > 0 . Quindi l'incertezza a posteriori è minore di quella a priori

Ora se osserviamo il dato $y = y_{(1)}$, le stime ottime di Bayes sono:

$$\boxed{\hat{\theta} = E[\theta | y = y_{(1)}] = \frac{\lambda_{\theta y}}{\lambda_{yy}} y_{(1)}}$$

Si può calcolare le varianze dell'errore di stima, ovvero:

$$\text{Var}[\theta - \hat{\theta}] = E\left[\left((\theta - \hat{\theta}) - E[\theta - \hat{\theta}]\right)^2\right]$$

$$\rightarrow E[\theta - \hat{\theta}] = E[\theta] - E[\hat{\theta}] = 0 - E\left[\frac{\lambda_{0y}}{\lambda_{yy}} y\right] = 0 - 0$$

\downarrow
= 0 per ipotesi

$$\begin{aligned} \Rightarrow \text{Var}[\theta - \hat{\theta}] &= E[(\theta - \hat{\theta})^2] = E\left[\left(\theta - \frac{\lambda_{0y}}{\lambda_{yy}} y\right)^2\right] = E\left[\theta^2 - 2\frac{\lambda_{0y}}{\lambda_{yy}} \theta y + \frac{\lambda_{0y}^2}{\lambda_{yy}^2} y^2\right] \\ &= E[\theta^2] - 2\frac{\lambda_{0y}}{\lambda_{yy}} E[\theta y] + \frac{\lambda_{0y}^2}{\lambda_{yy}^2} E[y^2] \\ &= \lambda_{00} - 2\frac{\lambda_{0y}}{\lambda_{yy}} \cdot \lambda_{0y} + \frac{\lambda_{0y}^2}{\lambda_{yy}^2} \lambda_{yy} = \boxed{\lambda_{00} - \frac{\lambda_{0y}^2}{\lambda_{yy}^2}} \end{aligned}$$

VARIANZA
DELLA POSTERIORI

«STIMA LINEARE»

Non è sempre il caso che θ e y siano compiutamente Gaussiane. Vogliamo quindi trovare un stimatore di non base istero sulle pdf compiute di θ ed y .

Supponiamo θ e y r.c. solni con valore atteso nullo e varianza λ_{00} e λ_{yy} rispettivamente.

$$E[\theta] = 0 \quad E[y] = 0 \quad E[\theta^2] = \lambda_{00} \quad E[y^2] = \lambda_{yy} \quad E[\theta y] = \lambda_{0y}$$

Vogliamo stimare θ dato y tenendo un stimatore lineare, t.c.:

$$\hat{\theta} = \alpha y + \beta \quad \alpha, \beta \in \mathbb{R} \text{ parametri reali}$$

Per trovare α e β , impostiamo la cifra di merito da mimimizzare

$$J(\alpha, \beta) = E\left[(\theta - \hat{\theta})^2\right] = E\left[(\theta - \alpha y - \beta)^2\right]$$

$$\begin{aligned} \bullet \frac{\partial J(\alpha, \beta)}{\partial \alpha} = 0 \Rightarrow 2 \cdot E[(\theta - \alpha y - \beta) \cdot (-y)] &= 2(E[-\theta y] + E[\alpha y^2] + E[\beta y]) = \\ &= 2(-\lambda_{0y} + \alpha \lambda_{yy} + \beta \cdot 0) = 2(-\lambda_{0y} + \alpha \lambda_{yy}) = 0 \end{aligned}$$

$$\bullet \frac{\partial J(\alpha, \beta)}{\partial \beta} = 0 \Rightarrow 2E[(\alpha - \gamma y - \beta) \cdot (-1)] = 2E[-\gamma + \alpha y + \beta] = 2(E[\gamma] + \alpha E[y] + E[\beta]) = 2\beta = 0$$

$$\begin{cases} 2(-\gamma y + \alpha \gamma y) = 0 \\ 2\beta = 0 \end{cases} \quad \begin{cases} \alpha = \frac{\gamma y}{\gamma y} \\ \beta = 0 \end{cases}$$

Mirko Mazzoleni - University of Bergamo

Lo stimatore bivariato ottimale è dato quindi da:

$$\hat{\theta} = \alpha y + \beta = \frac{\gamma y}{\gamma y} \cdot y + 0 = \frac{\gamma y}{\gamma y} \cdot y$$

CONCIDE CON LO STIMATORE DI RAYES!!
NEL CASO GAUSSIANO

La varianza dell'errore di stima si ricava essere:

$$\text{Var}[V - \hat{\theta}] = \gamma_{yy} - \frac{\gamma^2}{\gamma y}$$

COME RAYES NEL CASO GAUSSIANO!

↓
l'incertezza dell'errore di stima è minore rispetto a quella a priori

Osservazione

Lo stimatore bivariato fa nessuna ipotesi sulla distribuzione congiunta delle variabili. Infatti gli basta conoscere γ_{yy} e γ_{yy} .



Potrebbe dunque esserci un stimatore migliore di quello bivariato ottimale, cioè con varianza dell'errore di stima minore



Se però incognita e dati fossero distribuiti congiuntamente gaussiani, esso è anche migliore stimatore di quello bivariato ottimale.

Osservazioni

- 1) Se $\gamma_{yy} = 0$, cioè θ e y siano incollati, ovvero il dato y non porta informazioni su θ , allora le stime a priori non viene modificata dal dato. Infatti $\hat{\theta} = 0$
se $\gamma_{yy} = 0$, e $\text{Var}[\theta - \hat{\theta}] = \text{Var}[\theta] = \gamma_{yy}$
- 2) A parità di γ_{yy} , più elevato è γ_{yy} , e più piccola sarà la diminuzione di $\text{Var}[\theta - \hat{\theta}]$ causata dal dato y . Un valore elevato di γ_{yy} significa che il dato y è offerto da elevata incertezza (quindi porta poca informazione)

GENERALIZZAZIONE 1: valore ottimo non nullo, θ e y scalari

Se $E[\theta] = \mu_\theta \neq 0 \Rightarrow$ lo stimatore di Bayes nel caso gaussiano e lo stimatore lineare ottimo sono:

$$\boxed{\begin{aligned}\hat{\theta} &= \mu_\theta + \frac{\lambda_{\theta y}}{\lambda_{yy}} (y - \mu_y) \\ \text{Var}[\theta - \hat{\theta}] &= \lambda_{\theta\theta} - \frac{\lambda_{\theta y}^2}{\lambda_{yy}}\end{aligned}}$$

GENERALIZZAZIONE 2: y e θ sono vettori, $y \in \mathbb{R}^{m_y \times 1}$, $\theta \in \mathbb{R}^{m_\theta \times 1}$

Se $E[\theta] = \mu_\theta \neq 0$

$E[y] = \mu_y \neq 0$

$$\text{Var}\begin{bmatrix} y \\ \theta \end{bmatrix} = \begin{bmatrix} \Lambda_{yy} & \Lambda_{y\theta} \\ \Lambda_{\theta y} & \Lambda_{\theta\theta} \end{bmatrix} \quad \text{con } \Lambda_{y\theta} = \Lambda_{\theta y}^T$$

$$\begin{aligned}\mu_\theta &\in \mathbb{R}^{m_\theta \times 1} \\ \mu_y &\in \mathbb{R}^{m_y \times 1}\end{aligned}$$

$$\begin{aligned}\Lambda_{yy} &\in \mathbb{R}^{m_y \times m_y}, \quad \Lambda_{\theta\theta} \in \mathbb{R}^{m_\theta \times m_\theta} \\ \text{Var}\begin{bmatrix} y \\ \theta \end{bmatrix} &\in \mathbb{R}^{(m_y + m_\theta) \times (m_y + m_\theta)}\end{aligned}$$

Ora lo stimatore di Bayes nel caso Gaussiano e lo stimatore lineare ottimo sono dati da:

$$\boxed{\begin{aligned}\hat{\theta} &= \mu_\theta + \Lambda_{\theta y} \Lambda_{yy}^{-1} (y - \mu_y) \\ \text{Var}[\theta - \hat{\theta}] &= \Lambda_{\theta\theta} - \Lambda_{\theta y} \Lambda_{yy}^{-1} \Lambda_{y\theta}\end{aligned}}$$

Note

Le formule appena viste omaggiano alle forme ricorsive: si ottiene così lo stimatore $\hat{\theta}$ con l'ormai di nuovi dati, partendo dalle stime precedente



Queste equazioni ricorsive sono alla base del FILTO DI KALMAN, in cui lo stato $x(t)$ e l'uscita $y(t)$ sono visti come variabili casuali, e si vuol stimare lo stato $x(t)$ (l'incognita) dato l'osservazione dei dati $y(t)$

STIMA PATESIANA DEL VALORE ATTESO DI VARIABILI GAUSSIANE

Siano $y_{ij} \sim N(\theta, \sigma_{yy})$, iid, NOTA. Si vuole stimare il parametro θ tramite stima Bayesiana. Supponiamo $N=1$

Si definisce quindi una prior sul parametro θ , ovvero $P(\theta)$. Osserviamo che il parametro ignoto in questo caso è $\theta = E[y]$, ovvero il valore atteso di y .

Imponiamo una prior Gaussiana su θ , ovvero $P(\theta) = N(\mu_\theta, \sigma_{\theta\theta})$

Un modo per descrivere i dati y è: $y_{ij} = \theta + e_{ij}$ con $N(0, \sigma_{ee})$, iid, $e \perp \theta$, ovvero, i dati formano medie date dal valore di θ e disturbi dati da e_{ij}

↓
possiamo definire la likelihood $P(y|\theta) = N(\theta, \sigma_{ee})$, in cui θ è la variabile indipendente

Siamo quindi nel tipico caso di inferenza bayesiana in cui ho un prior su un parametro, $P(\theta)$, e ho una likelihood in funzione del quel parametro, $P(y|\theta)$. Possiamo calcolare la posterior come:

$$P(\theta|y) = \frac{P(y|\theta) \cdot P(\theta)}{P(y)} \rightarrow \text{se } e \text{ è Gaussiana}$$

dove $P(y) = \int_{-\infty}^{+\infty} P(y|\theta) P(\theta) d\theta$, e usare come $\hat{\theta}$ il valore atteso condizionato di $P(\theta|y)$.

↓
Osserviamo che, dato che $P(y|\theta)$ e $P(\theta)$ sono Gaussiane, allora anche $P(y|\theta)$ è Gaussiana

Quindi useremo le formule parziali per le distribuzioni condizionate sul caso Gaussiano, e useremo $\hat{\theta} = E[\theta|y]$ (che coincide con la stima da mea ottima)

La stima attesa è quindi:

$$\hat{\theta} = \mu_\theta + \frac{\partial \theta}{\partial y} (y - E[y])$$

Dove calcolare $E[y]$, $\frac{\partial \theta}{\partial y}$, $\frac{\partial^2 \theta}{\partial y^2}$

$$\bullet E[y] = E[\theta + e] = E[\theta] + E[e] = \mu_\theta + 0 = \boxed{\mu_\theta}$$

$$\bullet \sigma_{\theta y} = E[(\theta - \mu_\theta) \cdot (y - \mu_y)] = E[\theta y - \theta \mu_y - \mu_\theta y + \mu_\theta^2]$$

$$= E[\theta y] - E[\theta \mu_y] - E[\mu_\theta y] + E[\mu_\theta^2]$$

$$= E[\theta(\theta + e)] - \cancel{\mu_\theta \mu_\theta} - \cancel{\mu_\theta \mu_\theta} + \cancel{\mu_\theta^2}$$

$$= E[\theta^2] + E[\theta e] - \mu_\theta^2 = E[\theta^2] - E[\theta]^2 = \text{Var}[\theta] = \boxed{\lambda_{\theta\theta}}$$

$$\bullet \sigma_{yy} = E[(y - E[y])^2] = E[(y - \mu_y)^2] = E[y^2 - 2y\mu_y + \mu_y^2]$$

$$= E[y^2] - 2\mu_y E[y] + E[\mu_y^2]$$

$$= E[(\theta + e)^2] - 2\mu_\theta \mu_y + \mu_\theta^2$$

$$= E[\theta^2 + 2\theta e + e^2] - \mu_\theta^2 = E[\theta^2] + 2E[\theta e] + E[e^2] - \mu_\theta^2$$

$$= \underbrace{E[\theta^2] - E[\theta]^2}_{\text{Var}[\theta]} + E[e^2]$$

$$= \text{Var}[\theta] + \text{Var}[e] = \boxed{\lambda_{\theta\theta} + \lambda_{ee}}$$

Oggi:

$$\hat{\theta} = \mu_\theta + \frac{\lambda_{\theta y}}{\lambda_{yy}} (y - E[y]) = \mu_\theta + \frac{\lambda_{\theta\theta}}{\lambda_{\theta\theta} + \lambda_{ee}} (y - \mu_y)$$

$$= \mu_\theta + \frac{\lambda_{\theta\theta}}{\lambda_{\theta\theta} + \lambda_{ee}} y - \frac{\lambda_{\theta\theta}}{\lambda_{\theta\theta} + \lambda_{ee}} \mu_y = \frac{\mu_\theta (\lambda_{ee} + \lambda_{\theta\theta}) + \lambda_{\theta\theta} y - \lambda_{\theta\theta} \mu_y}{\lambda_{\theta\theta} + \lambda_{ee}}$$

$$= \boxed{\frac{\lambda_{ee}}{\lambda_{\theta\theta} + \lambda_{ee}} \mu_\theta + \frac{\lambda_{\theta\theta}}{\lambda_{\theta\theta} + \lambda_{ee}} y}$$

È IL VALORE ATTESO DELLA
POSTERIORI $P(\theta|y)$

↓
la distribuzione a posteriori delle
media ha questo valore atteso

Osservazioni

- La ~~stima~~ stime a posteriori del valore ottenere è una via di mettere tra le stime a priori μ_0 e le stime date dai dati, ovvero il valore y
- Nel caso in cui osserviamo N dati, ha che:

$$\hat{\theta} = \frac{\lambda_{\text{ee}}}{N\lambda_{\text{ee}} + \lambda_{\text{prior}}} \mu_0 + \frac{N\lambda_{\text{ee}}}{N\lambda_{\text{ee}} + \lambda_{\text{prior}}} \hat{\mu}_{\text{ML}}$$

STIMA ML della media
di una Gaussiana

$$\hat{\mu}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N y^{(i)}$$

L' se $N \rightarrow \infty$, allora $\hat{\theta} = \hat{\mu}_{\text{ML}} \Rightarrow$ forza sacer di evidenze!!

L' se $\lambda_{\text{ee}} \gg N\lambda_{\text{prior}}$, allora i dati hanno molta incertezza e non forza conciliano le stime a priori