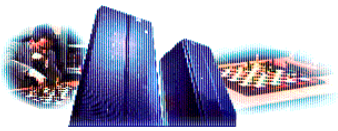
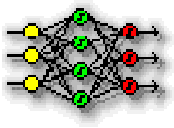




Sistemi Intelligenti
Relazione tra ottimizzazione e statistica - III

Alberto Borghese

Università degli Studi di Milano
Laboratory of Applied Intelligent Systems (AIS-Lab)
Dipartimento di Scienze dell'Informazione
borghese@di.unimi.it

Sommario

Analisi dell'affidabilità della stima

Distribuzione di Poisson e maximum likelihood

Sommario

Matrici e Sistemi lineari

Esempio di sistema linearizzato

Soluzione di un sistema lineare

Analisi dell'affidabilità della stima

Determinazione dei parametri di un modello non-lineare

Sistema lineare con misure affette da rumore

$$a_{11}x_1 + a_{12}x_2 + \ldots\ldots a_{1N}x_N = b_1 + \nu_1$$
$$a_{21}x_1 + a_{22}x_2 + \ldots\ldots a_{2N}x_N = b_2 + \nu_2$$
$$\ldots\ldots\ldots\ldots$$
$$a_{M1}x_1 + a_{M2}x_2 + \ldots\ldots a_{MN}x_N = b_M + \nu_M$$

Modello **Misure**

Errore di modello (sistematico, randomico). M x 1 => **Residuo.**

A x = b + N

M x 1

Vettore dei termini noti

M x N

(Matrice di disegno)

N x 1

Vettore delle incognite

Quale criterio viene soddisfatto da **x**?

Soluzione come problema di ottimizzazione

Funzione costo: $(Ax - b)^2 = \sum_k \nu_k^{\,2} = \|Ax - b\|^2$

Assegno un costo al fatto che la soluzione x, non soddisfi tutte le equazioni, la somma dei residui associati ad ogni equazioni viene minimizzata. Geometricamente: viene trovato il punto a distanza (verticale) minima da tutte le rette.

$$\min_{\mathbf{x}} \sum_k \nu_k^{\,2} = \min_{\mathbf{x}} (\mathbf{Ax} - \mathbf{b})^2$$

$$\frac{d}{d\mathbf{x}} (\mathbf{Ax} - \mathbf{b})^2 = 2\mathbf{A}^{\mathbf{T}} (\mathbf{Ax} - \mathbf{b}) = 0$$

$$\mathbf{A^T Ax = A^T b}$$

$$\mathbf{x} = (\mathbf{A^T A})^{-1} \mathbf{A^T b}$$

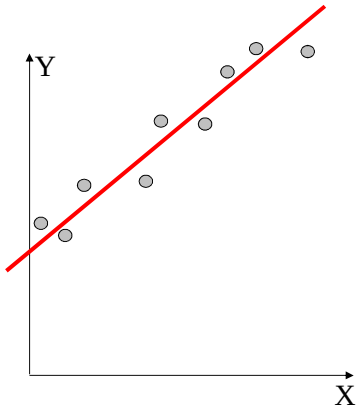
NB le funzioni costo sono spesso quadratiche (problemi di minimizzazione convessi) perchè il costo cresce sia che il modello sovrastimi che sottostimi le misure.
Inoltre, le derivate calcolate per imporre le condizioni di stazionarietà (minimo), sono relativamente semplici.

Stima alla massima verosimiglianza

Per ogni punto, dovrebbe valere $y_i = mx_i + q$.

Ma c'è l'errore di misura, misuriamo in realtà $y_i + \nu_i$.

Cerchiamo i parametri m e q che sono più verosimili.



Fitting di una retta

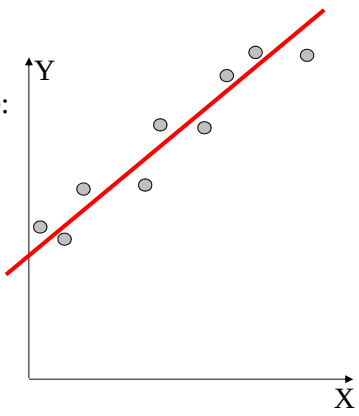
Vogliamo stimare i parametri di una retta: y = mz + q, con m e q incogniti: X = {m, q}

Abbiamo a disposizione N misure effettuate:

$Y = \{y_i ; x_i\}$

Sappiamo che le y_i sono affette da rumore Gaussiano a media nulla. In pratica:

$y_i = y_i + \nu_i$ dove ν_i è il rumore di misura.



Possiamo anche scrivere che:

$y_i = G(mx_i + b, \sigma^2)$, dove $G(\mu, \sigma^2)$ indica una distribuzione monodimensionale gaussiana a media μ e varianza σ^2.

Stima alla massima verosimiglianza

- Impostiamo il problema scrivendo la funzione di verosimiglianza e massimizzando tale funzione rispetto a m e q…
- Scriviamo prima di tutto la densità di probabilità di ottenere y_i per ciascun dato:

$$p(y_i \mid m, q; x_i) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - q}{\sigma} \right)^2 \right]$$

Stima alla massima verosimiglianza

Scriviamo il logaritmo negativo della verosimiglianza:

$$f(y_1, y_2 .... y_N; m, b; x_1, x_2 ..... x_N) = -\sum_{i=1}^{N} \ln\left\{ \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - q}{\sigma} \right)^2 \right] \right\} =$$

$$= -\sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \sum_{i=1}^{N} \left[-\frac{1}{2}\left(\frac{y_i - mx_i - q}{\sigma} \right)^2 \right] =$$

$$= -\sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} \right) + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - mx_i - q)^2$$

Minimizzo f(.) rispetto a {m.q}

Stima massima verosimiglianza

$$\left[\sum_{i=1}^{N} \left(x_i^{\,2} \right) \right] \cdot m + \left[\sum_{i=1}^{N} (x_i) \right] \cdot b = \left[\sum_{i=1}^{N} (y_i \cdot x_i) \right]$$ 1° equazione

$$\left[\sum_{i=1}^{N} (x_i) \right] \cdot m + \left[\sum_{i=1}^{N} (1) \right] \cdot b = \left[\sum_{i=1}^{N} (y_i) \right]$$ 2° equazione

Le incognite, m e b, compaiono con esponente 1 => equazioni lineari in m e b: X={m,q}.

Posso risolvere come:

$$\mathbf{A^T A x = A^T b}$$

$$\mathbf{x = (A^T A)^{-1} A^T b}$$

Stima alla massima verosimiglianza: caso 2D

- Ottengo un sistema di 2 equazioni in 2 incognite

$$\begin{bmatrix} \sum_{i=1}^{N}(x_i^2) & \sum_{i=1}^{N}(x_i) \\ \sum_{i=1}^{N}(x_i) & \sum_{i=1}^{N}(1) \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N}(y_i \cdot x_i) \\ \sum_{i=1}^{N}(y_i) \end{bmatrix}$$

A x = b →

X = (A'A)$^{-1}$A'B = CA'B - C è matrice di covarianza.

NB Le y_i sono solo ai termini noti

Stima ai minimi quadrati caso 2D

- Scriviamo l'equazione della retta per tutti i punti in forma matriciale (sistema lineare Ax=b, N equazioni, 2 incognite):

$$\begin{bmatrix} x1 & 1 \\ x2 & 1 \\ \dots & \dots \\ xN & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} y1 \\ y2 \\ \dots \\ yN \end{bmatrix}$$

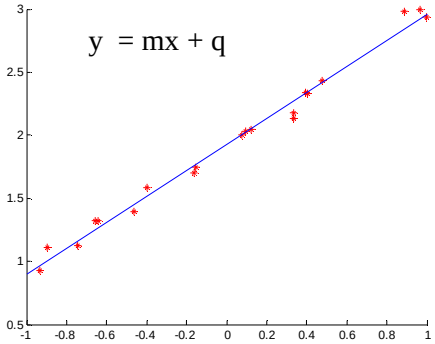
- Vogliamo trovare x t.c. $(Ax-b)^T(Ax-b)$ è minima (minimizzazione dei quadrati dei residui).

X = (A'A)$^{-1}$A'B = CA'B - C è matrice di covarianza.

Esempio - Caso 2D (2 parametri)

N = 20 punti
$\sigma_o^2 = 0.01$
m reale = 1
q reale = 2

m stimato = 0.9931
q stimato = 2.0106



Cosa vuol dire che {m,q} sono i più verosimili?
Quanto sono più verosimili?

Stima a massima verosimiglianza e minimi quadrati

$$\mathbf{A^T A x = A^T b}$$
$$\mathbf{x = (A^T A)^{-1} A^T b}$$

La soluzione a massima verosimiglianza, quando il rumore è Gaussiano a media nulla, coincide con la soluzione ai minimi quadrati del sistema lineare associato (la soluzione ai minimi quadrati è un caso particolare della stima alla massima verosimiglianza).

La soluzione è quella che minimizza lo scarto quadratico medio dei residui, ovverosia è a minima varianza.

La stima a massima verosimiglianza è un approccio generale, e si presta a p(x) di qualsiasi forma. La Gaussiana consente di ottenere una formulazione lineare del problema.

Stima ai minimi quadrati e verosimiglianza

- Nella soluzione ai minimi quadrati del sistema lineare Ax=b si definisce un vettore errore v = Ax - b;
- Nel caso di soluzione "perfetta" v = 0;
- Dal momento che abbiamo un numero di equazioni maggiore rispetto al numero di incognite, cerchiamo il vettore e a norma minima;
- In pratica cerchiamo x t.c. $v^T v = \Sigma_i v_i^2$ è minimo.

Giustificazione statistica

- **C'è un solo insieme vero dei parametri**, mentre ci **possono essere infiniti universi di dati per effetto dell'errore di misura**.
- La domanda quindi più corretta sarebbe: "Dato un certo insieme di parametri, qual'è la probabilità che questo insieme di dati sia estratto?" (più correttamente si parla di densità di probabilità?)
- Cioè, **per ogni insieme di parametri, calcoliamo la probabilità che i dati siano estratti. Ovverosia la likelihood (verosimiglianza) dei dati, dato un certo insieme di parametri.**

La stima ai minimi quadrati dei parametri è equivalente a determinare i parametri che massimizzano la funzione di **verosimiglianza** sotto l'ipotesi di errore **Gaussiano a media nulla**.

Stima ai minimi quadrati pesata

$\min \| P(Ax - b) \|$ **P di dimensioni m x m – matrice dei pesi, diagonale**

$$p_1a_{11}x_1 + p_1a_{12}x_2 - p_1b_1 = p_1v_1$$
$$p_2a_{21}x_1 + p_2a_{22}x_2 - p_2b_2 = p_2v_2$$
$$p_3a_{31}x_1 + p_3a_{32}x_2 - p_3b_3 = p_3v_3$$

Residuo pesato $\min \sum_k (p_k v_k)^2$

$$\mathbf{A^T PA\, x = A^T P\, b}$$

$$\mathbf{x = (A^T P A)^{-1} A^T P\, b}$$

Rank(A) = Rank(C)

$C = (A^T * P * A)^{-1}$ è la matrice di covarianza (matrice quadrata n x n)

Sistema lineare: soluzione robusta

A x = b ⇒ **A' A x = A' b** ⇒ **x = (A'A)$^{-1}$A'b**

Numero di condizionamento varia circa con (A'*A).

Soluzione tramite Singular Value Decomposition (diagonalizzazione)

Numero di condizionamento varia circa con det(A).

A x = b

U W V X = B

x = V'W^{-1}U'b

Ortonormale M x N — Diagonale (N x N) — Ortonormale N x N

$\mathbf{V^T W^{-1} U^T U W V X = V^T W^{-1} U^T b}$ ➔ $\mathbf{X = V^T W^{-1} U^T b}$

- La matrice C non viene formata.
- W^{-1} contiene i reciproci degli elementi di W.

W^{-1} è diagonale. $w_{ii}^{-1} = 1/w_{ii}$

Rank-deficiency nella matrice dei coefficienti

Quando (A'*A) è singolare?

$x = (A'*A)^{-1}A'*b$ $\quad$ **$x = V'W^{-1}U'b$**

Se A è rank-deficient, A'*A è singolare.

Si può facilmente osservare valutando il valore singolare più piccolo della matrice W che risulta uguale a 0.

In questo caso il problema è sovraparametrizzato.

Si può anche valutare il grado di condizionamento di A, analizzando $\frac{w_{\min}}{w_{Max}}$

Soluzione come problema di ottimizzazione

Funzione costo: $(Ax - b)^2 = \sum_k v_k^2 = \|Ax - b\|^2$

Assegno un costo al fatto che la soluzione x, non soddisfi tutte le equazioni, la somma dei residui associati ad ogni equazioni viene minimizzata. Geometricamente: viene trovato il punto a distanza (verticale) minima da tutte le rette.

$$\min_{\mathbf{x}} \sum_k v_k^2 = \min_{\mathbf{x}} (\mathbf{Ax} - \mathbf{b})^2$$

$$\frac{d}{d\mathbf{x}}(\mathbf{Ax} - \mathbf{b})^2 = 2\mathbf{A}^T(\mathbf{Ax} - \mathbf{b}) = 0$$

$$\mathbf{A^TAx = A^Tb}$$

$$\mathbf{x = (A^TA)^{-1}A^Tb}$$

NB le funzioni costo sono spesso quadratiche (problemi di minimizzazione convessi) perchè il costo cresce sia che il modello sovrastimi che sottostimi le misure.
Inoltre, le derivate calcolate per imporre le condizioni di stazionarietà (minimo), sono relativamente semplici.

Sommario

Matrici e Sistemi lineari

Esempio di sistema linearizzato

Soluzione di un sistema lineare

Analisi dell'affidabilità della stima

Determinazione dei parametri di un modello non-lineare

Valutazione della bontà della stima

$x = (A'^*A)^{-1}A'^* b \quad <==> \quad \min_{\mathbf{x}} \sum_k v_k^2 = \min_{\mathbf{x}} (\mathbf{Ax}-\mathbf{b})^2$

Errore di modellizzazione Gaussiano a media nulla $N(0,\sigma^2)$

$$<v_k> = 0$$

$$\hat{\sigma}_0^2 = \sum_{k=1}^{M} \left(v_k^2\right) = |v|^2$$

Varianza della stima = varianza dell'errore di misura

.

Valutazione della bontà della stima del singolo parametro

$x = (A'^*A)^{-1}A'^* b$

$x = CA'^* b$

$$\hat{\sigma}_0^2 = \sum_{m=1}^{M} \left(\nu_m^2\right)$$

Chiamiamo u e v le variabili casuali associate all'errore sui parametri e all'errore di modellizzazione, rispettivamente. Si suppone errore a media nulla e Gaussianamente distribuito.

$u = \Delta x$ $\quad (x + u) = C A' (b + v)$

$\mathbf{x = CA'b}$ $\quad u = C A'^* v$ $\quad E[u] = 0$

C è la matrice di covarianza

Impostazione del calcolo della correlazione tra i parametri

$u = C A' v$

Vogliamo individuare la correlazione tra due parametri i e j. Devo quindi determinare la loro correlazione:

$$\begin{bmatrix} u_1^2 & u_1u_2 & \dots & u_1u_W \\ u_2u_1 & u_2^2 & \dots & u_2u_W \\ \dots & \dots & \dots & \dots \\ u_Wu_1 & u_Wu_2 & \dots & u_W^2 \end{bmatrix}$$

$\langle u_i, u_j \rangle$

$u = C A' v$ $\quad \Rightarrow \quad u' = v'A (C)'$

$uu' = C A' vv'A C'$ => Applicando l'operatore di media, si ottiene:

$\langle uu' \rangle = C A' \langle vv' \rangle A C'$

Dato che v sono i residui, e sono indipendenti, e tutte i punti di controllo hanno lo stesso tipo di errore di misura, si avrà che $\langle vv' \rangle = I\sigma_0^2$.

Incertezza sulla stima dei parametri

$\langle uu' \rangle = CA'\, IA\, C'\sigma_0^2 = C'\, \sigma_0^2$ $\qquad$ $\boxed{\langle u'u \rangle = C\, \sigma_0^2}$

Segue che: $\sigma^2(u_{ij}) = c_{ij}\, \sigma_0^2$ Varianza sulla stima del parametro.

Spiegazione intuitiva:

a x + 3 = y +noise

Calcolo x come: $x = y * a^{-1} - 3$

Quanto è sensibile questa stima? Cosa succede se, per effetto del noise, invece di misurare y, misuro y + v?

x varierà di $v * a^{-1}$. Il rumore viene cioè moltiplicato per a^{-1}.

Matrice di covarianza

Date N variabili casuali: $x = [x_1, x_2, \ldots x_N]$ si può misurare la correlazione tra coppie di variabili. E' comodo rappresentare la correlazione tra variabili casuali in un'unica matrice detta **matrice di covarianza** come:

$$C = \begin{bmatrix} \sigma_{x_1x_1} & \sigma_{x_1x_2} & . & \sigma_{x_1x_N} \\ \sigma_{x_2x_1} & \sigma_{x_2x_2} & . & \sigma_{x_2x_N} \\ . & . & . & . \\ \sigma_{x_Nx_1} & \sigma_{x_Nx_2} & . & \sigma_{x_Nx_N} \end{bmatrix}$$

Varianza: $\sigma_{x_1x_1} = \sigma^2_{x_1}$ N parametri

Covarianza: $\sigma_{x_ix_j} = \sigma_{x_jx_i}$ $i \neq j$ $(N-1)^2/2$ parametri

Correlazione tra coppie di parametri

Date due variabili casuali: x_i, x_j, l'indice di correlazione misura quanto le coppie di variabili estratte: $p(x_i, x_j)$ stanno su una retta:

$$r = \frac{M_{x_i x_j} - M_{x_i} M_{x_j}}{\sigma_{x_i} \sigma_{x_j}} \qquad -1 \le r \le +1$$

Definendo la covarianza tra x_i ed x_j come:

$$\sigma_{x_i x_j} = \frac{1}{N} \sum_i \sum_j \left(x_i - M_{x_i}\right)\left(x_j - M_{x_j}\right)$$

Dalla definizione di deviazione standard risulta:

$$r = \frac{\sigma_{x_i x_j}}{\sigma_{x_i} \sigma_{x_j}}$$

Correlazione tra i parametri

$\langle uu' \rangle = CA' \, IA \, C' \sigma_0^2 = C' \, \sigma_0^2$ $\qquad$ $\langle u'u \rangle = C \, \sigma_0^2$

Da cui si giustifica il nome di matrice di covarianza per C.

Segue che: $\sigma^2(u_{ij}) = c_{ij} \, \sigma_0^2$ Varianza sulla stima del parametro.

$$-1 \le r_{ij} = \frac{\langle u_i u_j \rangle}{\sqrt{\langle u_i \rangle^2 \langle u_j \rangle^2}} = \frac{c_{ij}}{\sqrt{c_i c_j}} \le +1$$

Indice di correlazione tra il parametro i ed il parametro j
(empiricamente si scartano parametri quando la correlazione è superiore al 95%)

Vanno rapportati alle dimensioni dei parametri coinvolti.

La covarianza: momenti di 2 variabili statistiche

Covarianza = $E[(x - \mu_x)(y - \mu_y)]$

Varianza = $E[(x - \mu_x)(x - \mu_x)]$

Per due variabili indipendenti, la covarianza = 0, non variano assieme (covariano)

$$C = \begin{bmatrix} \sigma_x^2 & \sigma_x\sigma_y \\ \sigma_y\sigma_x & \sigma_y^2 \end{bmatrix}$$

```
>> x = randn(N,1);
>> y = randn(N,1);
>> temp = x.*y;
>> covarianza = mean(temp)
```


Misura di correlazione su 2 parametri

Misura la inter-dipendenza tra 2 variabili statistiche:

$$-1 \le \frac{\sigma_{xy}}{\sigma_x\sigma_y} = c = \lim_{N->\infty} \frac{\sum_k (x_k - \mu_x)(y_k - \mu_y)}{\sqrt{\sum_k (x_k - \mu_x)^2}\sqrt{\sum_k (y_k - \mu_y)^2}} \le +1$$

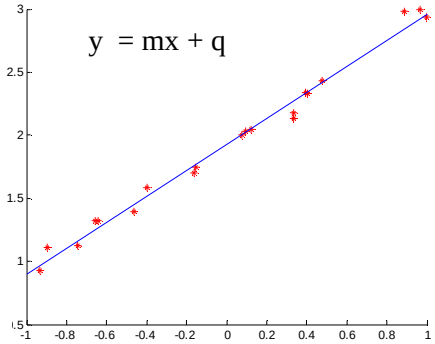
```
>> x = randn(N,1);
>> y1 = randn(N,1);
>> y2 = x;
>> temp1 = x.*y1;
>> temp2 = x.*y2;
>> covarianza1 = mean(temp1)% Uncorrelated variables(c ->
1)
>> covarianza2 = mean(temp2)% Correlated variables (c = 0)
```

Caso 2D

N = 20 punti $\sigma_o^2 = 0.01$
m reale = 1 q reale = 2

C =
0.0494 0.0087
0.0087 0.0515

m stimato = 0.9931
q stimato = 2.0106



Altra realizzazione del rumore:

C=
0.1702 0.0124
0.0124 0.0509

m stimato = 0.9937
q stimato = 1.9522

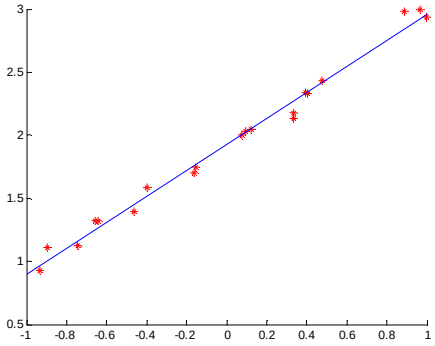
$\sigma^2(m) = c_{11}\,\sigma_0^2 = 0.1702 * 0.01 = 0.0017 \Rightarrow \sigma = 0.04$
$\sigma^2(q) = c_{22}\,\sigma_0^2 = 0.0515 * 0.01 = 0.0005 \Rightarrow \sigma = 0.022$

Caso 2D – less points

N = 10 punti $\sigma_o^2 = 0.01$
m reale = 1 q reale = 2

C =
0.5927 -0.0030
-0.0030 0.1000
m_stimato =
1.0081
q_stimato =
1.9616

C =
0.2514 -0.0360
-0.0360 0.1051
m_stimato =
1.0012
q_stimato =
1.9107



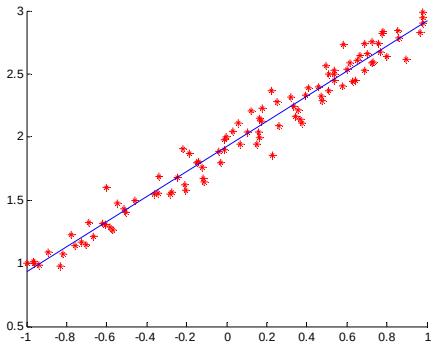
y = mx + q

Diminuisce la confidenza nella stima

Caso 2D – more points

N = 100 punti $\sigma_o^2 = 0.01$
m reale = 1 q reale = 2

C =
0.0327 -0.0034
-0.0034 0.0103
m_stimato =
0.9942
q_stimato =
1.9978



y = mx + q

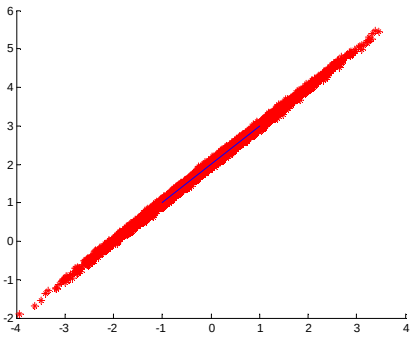
Aumenta la confidenza nella stima

Caso 2D – even more points

N = 10000 punti $\sigma_o^2 = 0.01$
m reale = 1 q reale = 2

C =
0.0000993 0.0000004
0.0000004 0.0001000

m_stimato =
1.000039
q_stimato =
2.0012



y = mx + q

Aumenta la confidenza nella stima

Sommario

Matrici e Sistemi lineari

Esempio di sistema linearizzato

Soluzione di un sistema lineare

Analisi dell'affidabilità della stima

Determinazione dei parametri di un modello non-lineare

Stima di parametri in insiemi di equazioni non lineari - linearizzazione

y = f(x) viene linearizzata utilizzando il differenziale (retta tangente):

$$dy = f(x_o) + \left.\frac{df(x)}{dx}\right|_{x=x_o} dx \quad = y_o + \left.\frac{df(x)}{dx}\right|_{x=x_o} dx$$

Si può vedere come sviluppo di Taylor arrestato al 1° ordine
E' un'equazione lineare.

Per funzioni di più variabili, f(**P;W**) = 0, la linearizzazione nell'intorno di **P**, si può scrivere come:

$$F(\mathbf{P};\mathbf{W}) = F(\mathbf{P_o};\mathbf{W_o}) + \sum_{j=1}^{W} \left.\frac{\partial F(.)}{\partial w_j}\right|_{P_o,W_o} * dw_j = \quad k - \sum_{j=1}^{W} a_j * dw_j$$

E' un'equazione lineare che descrive il comportamento della funzione F(.) nell'intorno del punto **Po** con i parametri **Wo**.

Metodo di Gauss-Newton

- L'idea:

Inizializzazione:

- Inizializzo i parametri ad un valore iniziale.

Iterazioni:

- 1) Linearizzazione delle equazioni.
- 2) Stima dell'aggiornamento dei parametri nel modello linearizzato ai minimi quadrati (soluzione ottimale, minimo del problema linearizzato).
- 3) Correzione dei parametri.

Può essere pesante perchè richiede l'inversione della matrice di covarianza. Spesso si preferiscono utilizzare metodi di ottimizzazione del primo ordine.

In pratica

$\mathbf{y} = f(\mathbf{x})$ x, y vettori di N ed M elementi rispettivamente

$\mathbf{y_0} = f(\mathbf{x_0})$ x_0, y_0 valore iniziale

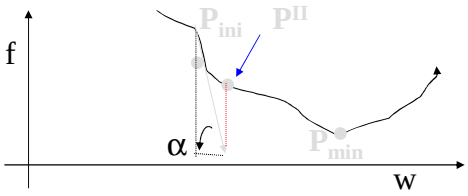
Iterazione di (nella prima iterazione k = 0):

- $\mathbf{dy_k} + \mathbf{y_k} = (\Sigma \delta f(x)/dx)_{xk}\ \mathbf{dx} + \mathbf{f(x_k)}$ $(\Sigma \delta f(x)/dx)_{xk}$ **are numbers!**
- Si ottiene un sistema lineare
- Viene risolto come $\mathbf{dx_k} = (AA^T)^{-1}A^T\mathbf{dy_k}$
- **Si aggiorna il valore di x come $\mathbf{x_{k+1}} = \mathbf{x_k} + \mathbf{dx_k}$**

Fino a convergenza

Minimizzazione tramite gradiente (metodo del primo ordine): 1 variabile

Tecnica del gradiente applicata alla minimizzazione di funzioni non-lineari di **una variabile, x,** e di **un parametro, w**: f = f (**x** | **w**).



La derivata, mi dà due informazioni:

1) In quale direzione di **w**, la funzione decresce.
2) Quanto rapidamente decresce.

Definisco uno spostamento arbitrario lungo la pendenza: maggiore la pendenza maggiore lo spostamento.

dw ∝ -f'(w;P) dati P, w. La derivata viene calcolata rispetto a w.

Occorre un'inizializzazione.

Metodo iterativo.

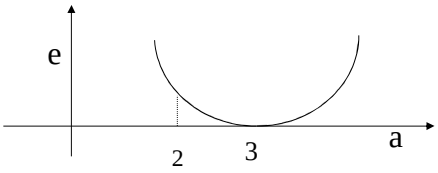
Esempio di applicazione tecnica del gradiente per funzioni di 1 variabile

Supponiamo che il modello da noi considerato sia semplice: $y = ax^2$

Abbiamo un unico parametro da determinare: a. La funzione è lineare in a.

Misuriamo un punto sulla parabola: x = 1, y =3.
Vogliamo modificare a in modo che la parabola passi per P(x,y).
La funzione costo da minimizzare sarà: $e = f(a \mid x,y) = (y-ax^2)^2$
La soluzione è a = 3

Partiamo da $a_{ini} = 2$.
$err = (3 – 2*1)*1 = 1$



Utilizziamo il metodo del gradiente:

Calcoliamo la derivata di $f(a \mid x,y)$ ➔ $f'(a) = -2 (y – a x^2) x^2$

Minimizzazione – underdamping

Consideriamo $\alpha = 1$
Calcoliamo la derivata di f(.) ➔ $f'(.) = -2 (y – a x^2) x^2$

Utilizziamo il metodo del gradiente:

Passo 1:
Calcoliamo l'incremento da dare al parametro a:
da =-[-2 (3 – 2 1) 1] = -[-6 +4] = 2 a' = 2 + 2 = 4

Passo 2:
Calcoliamo l'incremento da dare al parametro a:
da = -[-2 (3 - 4 1) 1] = -[-6 +8] = -2 a'' = 4 – 2 = 2
Oscillazioni!!!

Mi sposto troppo velocemente da una parte all'altra del minimo.

Minimizzazione -2 passi

Consideriamo $\alpha = 0.4$
Calcoliamo la derivata di f(.) ➔ $f'(.) = -2 (y – a x^2) x^2$

Utilizziamo il metodo del gradiente:

Passo 1:
Calcoliamo l'incremento da dare al parametro a:
da = -0.4 [-2 (3 – 2 1) 1] = -[-6 +4] = 0.8 a' = 2 + 0.8 = 2.8

Passo 2:
Calcoliamo l'incremento da dare al parametro a:
da = - 0.4 [-2 (3 – 2.8 1) 1] = -[-6 + 5.6] = 0.16 a'' = 2.8 + 0.16 = 2.96
Converge ad a = 3.

Posso correre il rischio di spostarmi troppo lentamente

Minimizzazione di funzioni di più variabili

min(f(**x**, **w**)) funzione costo od errore, **w** vettore.

Modifico il valore dei pesi di una quantità proporzionale alla pendenza della funzione costo rispetto a quel parametro.
La pendenza è una direzione nello spazio, non è più solamente destra / sinistra. Devo calcolare la derivata spaziale = **gradiente** della funzione costo, f(.).
Estensione della tecnica del gradiente a più variabili.

$$dw = -\alpha \nabla f(x;w), \quad \text{dato P, W.}$$

Serve un'**approssimazione iniziale** per i pesi $W_{ini} = \{w_j\}_{ini}$.

Evoluzione dei metodi del primo ordine

- α è un parametro critico. Se è troppo piccolo convergenza molto lenta, se è troppo grande overshooting.
- Ottimizzazione di α. Ad ogni passo viene calcolato a ottimale, per cui la funzione è decrescente (line search).

Sommario

Matrici e Sistemi lineari

Esempio di sistema linearizzato

Soluzione di un sistema lineare

Analisi dell'affidabilità della stima

Determinazione dei parametri di un modello non-lineare

Stima di parametri in insiemi di equazioni non lineari - linearizzazione

y = f(x) viene linearizzata utilizzando il differenziale:

$$y = f(x_o) + \left.\frac{df(x)}{dx}\right|_{x=x_o} dx = y_o + \left.\frac{df(x)}{dx}\right|_{x=x_o} dx$$

Si può vedere come sviluppo di Taylor arrestato al 1° ordine
E' un'equazione lineare in dx.

Per funzioni di più variabili, f(**P;W**) = 0, la linearizzazione si può scrivere come:

$$F(\mathbf{P};\mathbf{W}) = F(\mathbf{P_o};\mathbf{W_o}) + \sum_{j=1}^{W} \left.\frac{\partial F(.)}{\partial w_j}\right|_{P_o,W_o} * dw_j = \quad k - \sum_{j=1}^{W} a_j * dw_j$$

E' un'equazione lineare nei dw che descrive il comportamento della funzione F(.) nell'intorno del punto Po con i parametri Wo.

Sommario

Matrici e Sistemi lineari

Esempio di sistema linearizzato

Soluzione di un sistema lineare

Analisi dell'affidabilità della stima

Determinazione dei parametri di un modello non-lineare

Sommario

Analisi dell'affidabilità della stima

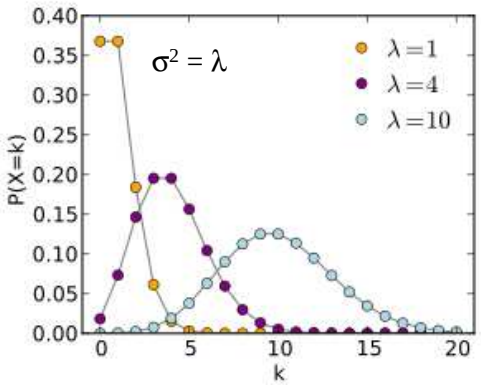
Distribuzione di Poisson e maximum likelihood

Il caso poissoniano

- La formulazione di un problema di verosimiglianza permette di trattare casi in cui la variabili misurate abbiano distribuzione diversa da quella gaussiana.
- Consideriamo ad esempio una variabile poissoniana (es. conteggio di un numero limitato di fotoni, es. radiografia).
- Media della variabile casuale (=varianza nel caso della poisson) = λ.

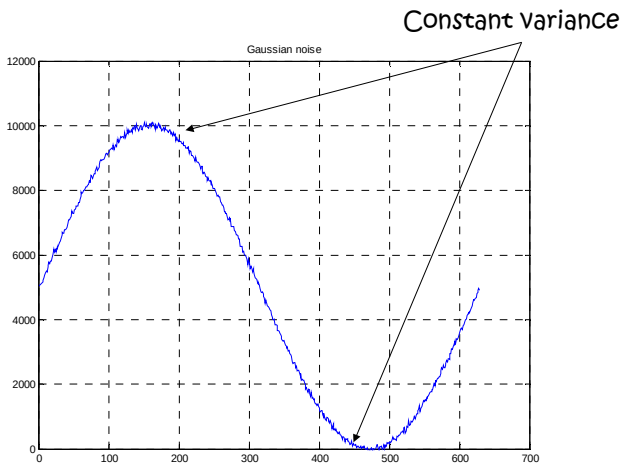
$$p(n_i \mid \lambda) = \frac{\lambda^{n_i} \cdot e^{-\lambda}}{n_i!}$$

Distribuzioni notevoli: la Poisson

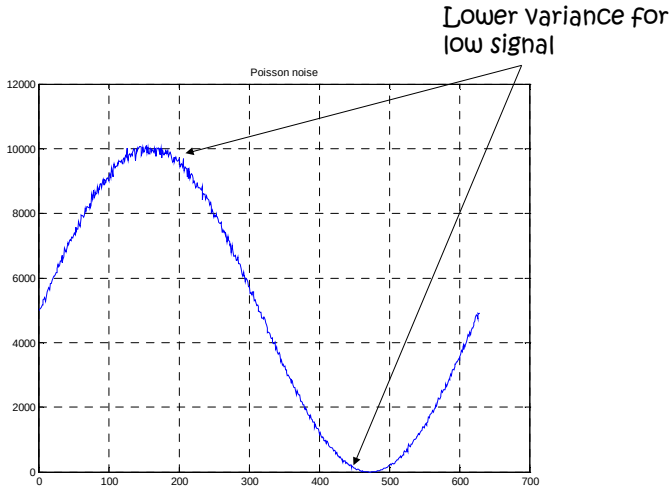


$$p(n_i \mid \lambda) = \frac{\lambda^{n_i} \cdot e^{-\lambda}}{n_i!}$$

Gaussian noise: example



Poisson noise: example



Il caso poissoniano

- Sia y un vettore di misure di una variabile poissoniana. Si vuole stimare la media della variabile.
- Scriviamo il logaritmo negativo della funzione di verosimiglianza:

$$f(y_1, y_2, y_3 ..... y_N \mid \lambda) = (-\ln(L(.)) = -\sum_{i=1}^{N} \ln[p(y_i \mid \lambda)] = -\sum_{i=1}^{N} \ln\left[\frac{\lambda^{y_i} \cdot e^{-\lambda}}{y_i!}\right] =$$

$$= -\sum_{i=1}^{N} \ln[\lambda^{y_i}] - \sum_{i=1}^{N} \ln[e^{-\lambda}] - \sum_{i=1}^{N} \ln\left[\frac{1}{y_i!}\right] =$$

$$-\sum_{i=1}^{N} y_i \cdot \ln[\lambda] + \sum_{i=1}^{N} \lambda + \sum_{i=1}^{N} \ln[y_i!] =$$

$$= -\ln(\lambda) \cdot \sum_{i=1}^{N} y_i + N \cdot \lambda + \sum_{i=1}^{N} \ln[y_i!]$$

Il caso poissoniano

- Massimizziamo la verosimiglianza rispetto a λ (ponendo a zero la derivata):

$$\frac{\partial}{\partial \lambda}[f(.)] = \frac{\partial}{\partial \lambda}\left[-\ln(\lambda) \cdot \sum_{i=1}^{N} y_i + N \cdot \lambda + \sum_{i=1}^{N} \ln[y_i!]\right] =$$

$$= -\frac{1}{\lambda} \cdot \sum_{i=1}^{N} y_i + N + 0 = 0 \Rightarrow \lambda = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Otteniamo anche in questo caso la media campionaria come stima della media della distribuzione.

Il caso poissoniano (riassunto)

- L'approccio alla massima verosimiglianza permette di effettuare stime di parametri, data una qualsiasi densità di probabilità dei dati misurati (non solo gaussiana, poisson!)

Sommario

Analisi dell'affidabilità della stima

Distribuzione di Poisson e maximum likelihood