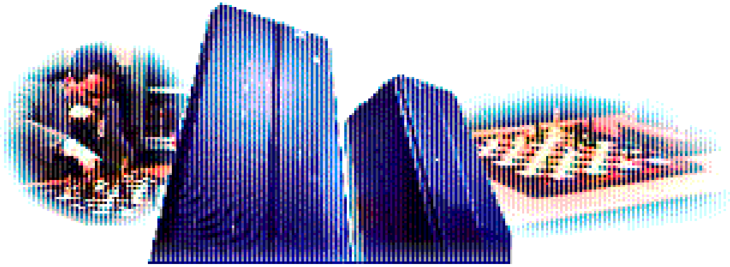
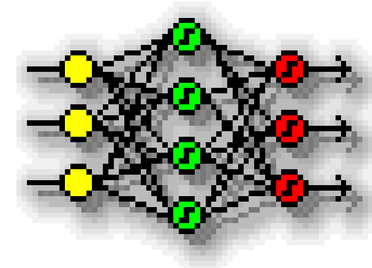




Sistemi Intelligenti
Stimatori e identificazione - II

Alberto Borghese

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

Overview

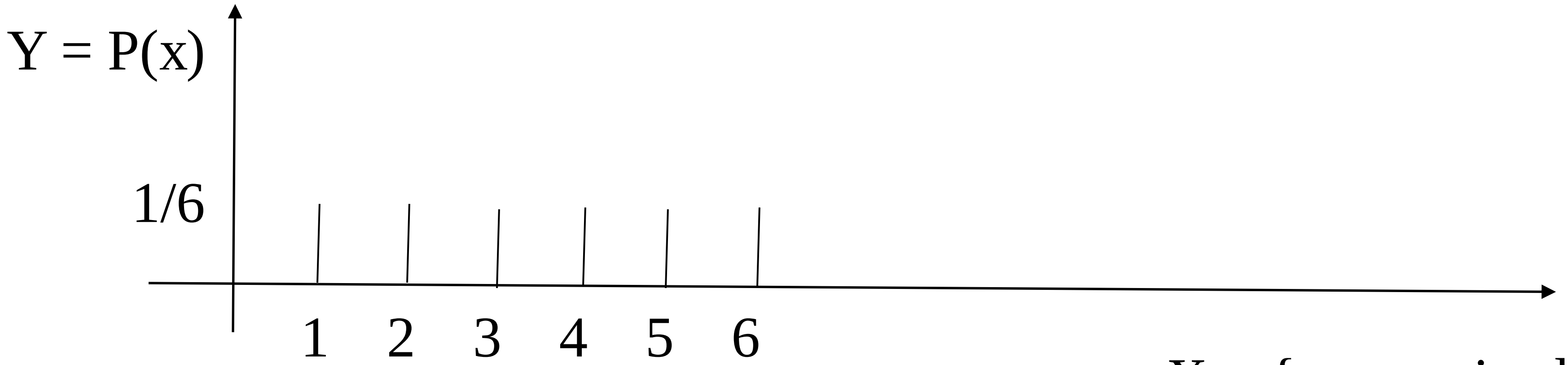
Densità di probabilità

Stimatori

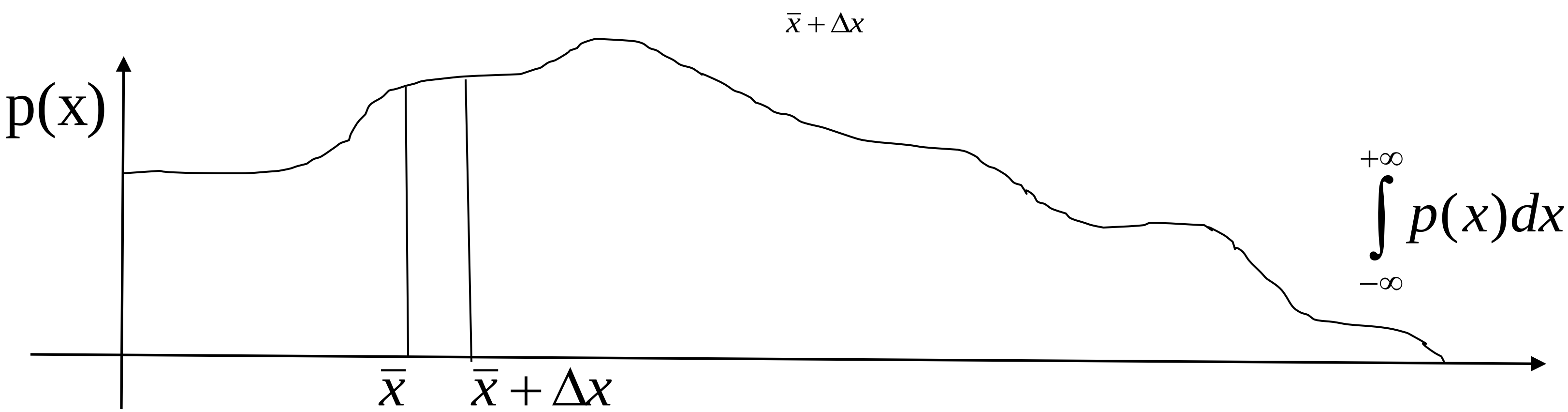
Modelli

Stima alla massima verosimiglianza di parametri di un modello

La probabilità nel caso continuo



X = {punteggio dado}



$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

$X \in C^0$

$$P(x \in [\bar{x}, \bar{x} + \Delta x]) \int_{\bar{x}}^{\bar{x}+\Delta x} p(x)dx$$

Definizione di p(x)

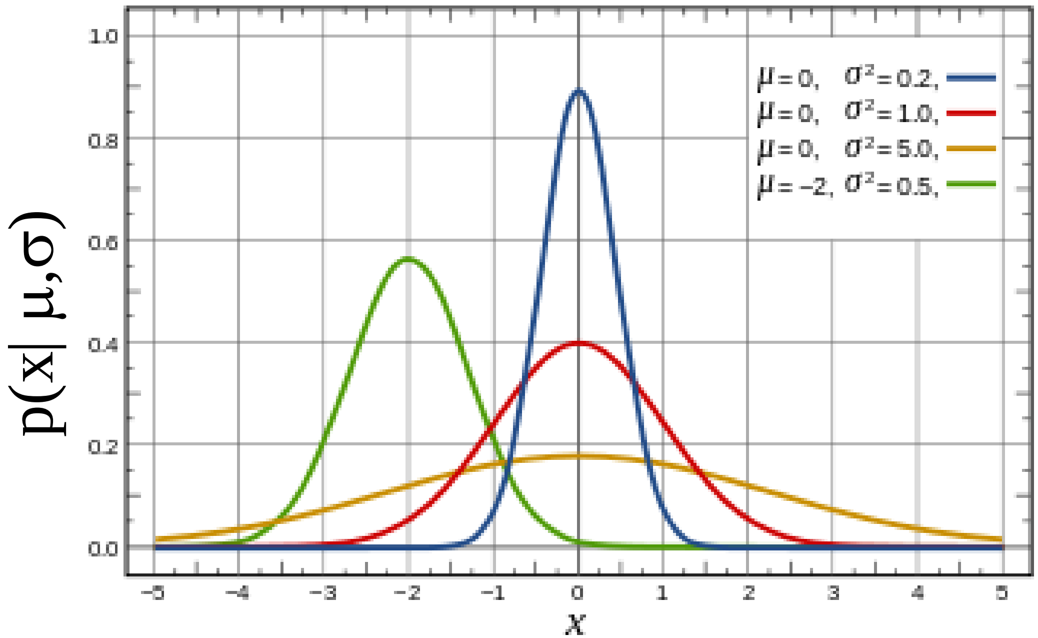
Caso discreto: prescrizione della probabilità per ognuno dei finiti valori che la variabile X può assumere: P(X).

Caso continuo: i valori che X può assumere sono infiniti. Devo trovare un modo per definirne la probabilità. Descrizione **analitica** mediante la funzione densità di probabilità. Si considera la probabilità che x cada in un certo intervallo.

Valgono le stesse relazioni del caso discreto, dove alla somma si sostituisce l'integrale.

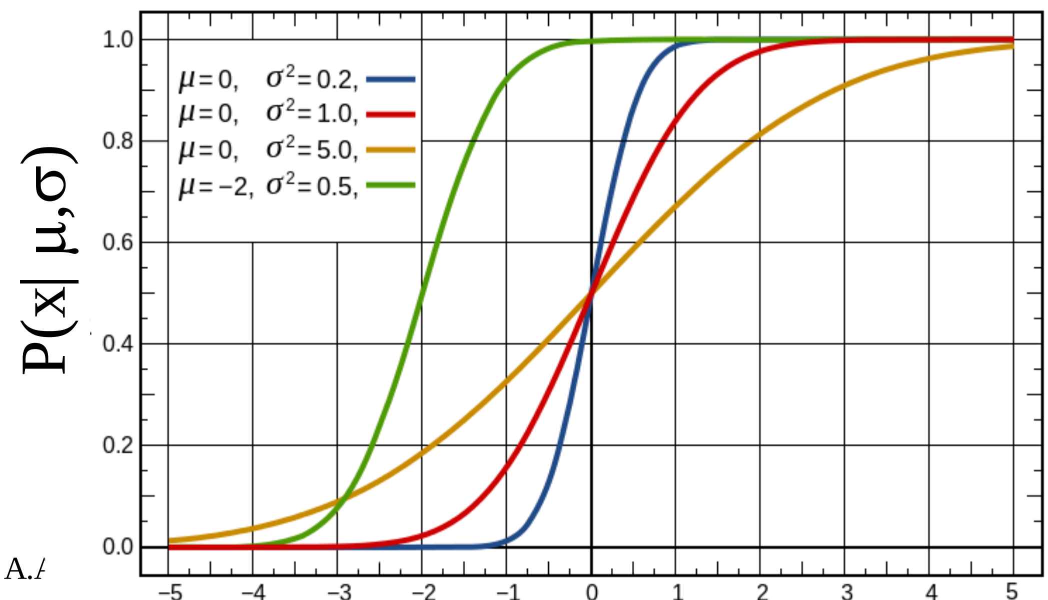
$$P(X = x \in [\bar{x}, \bar{x} + \Delta x]) \int_{\bar{x}}^{\bar{x}+\Delta x} \int_{-\infty}^{+\infty} p(x, y) dxdy$$

Distribuzioni notevoli: la Gaussiana



$$p(x \mid \mu,\sigma) = \frac{1}{\left(\sqrt{2\pi}\right)\Sigma^{D}} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\Sigma}\right)^{2}\right]$$

D = dimensione, in questo caso D = 1



$\Pr(|X-\mu| < \sigma) = 0.68268$

$\Pr(|X-\mu| < 2\sigma) = 0.95452$

$\Pr(|X-\mu| < 3\sigma) = 0.9973$

I momenti di una variabile statistica

$$\mu^k(X) = \int_{-\infty}^{+\infty} (x-a)^k \, p(x) dx$$ Momento rispetto ad a, solitamente alla media

$$E[X] = \int_{-\infty}^{+\infty} (x-\mu) \, p(x)$$ Valore atteso (Expected value) di X = media distribuzione

$$E\left[(X-\mu)^2\right] = \int_{-\infty}^{+\infty} (x-\mu)^2 \, p(x)$$ Varianza (σ^2)

$$E\left[(X-\mu)^3\right] = \int_{-\infty}^{+\infty} (x-\mu)^3 \, p(x)$$ Asimmetria

$$E\left[(X-\mu)^4\right] = \int_{-\infty}^{+\infty} (x-\mu)^4 \, p(x)$$ Kurtosi – peso delle code di p(x)

Overview

Densità di probabilità

Stimatori

Modelli

Stima alla massima verosimiglianza di parametri di un modello

Stimatori

Variabili indipendenti: $p(y_1,y_2) = p(y_1)p(y_2)$

Gaussiana: siano date due realizzazioni **indipendenti** della stessa variabile casuale Y… Quale è la probabilità di misurare y_1 nella prima realizzazione e y_2 nella seconda realizzazione (estrazione con ripetizione)?

$$p(y_1 \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_1-\mu}{\sigma}\right)^2\right]$$

$$p(y_2 \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_2-\mu}{\sigma}\right)^2\right]$$

$$p(y_1, y_2 \mid \mu,\sigma) = p(y_1 \mid \mu,\sigma) \cdot p(y_2 \mid \mu,\sigma) =$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_1-\mu}{\sigma}\right)^2\right] \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_2-\mu}{\sigma}\right)^2\right]$$

Il risultato dipende dai parametri della Gaussiana che descrive l'errore di misura: μ e σ.
Qual'è il valore più ragionevole per μ e σ?

Funzione di verosimiglianza

- Siano date **N variabili casuali <u>indipendenti</u>**… Quale è la **probabilità di misurare il vettore [y_1, …, y_N]**?

$$p(y_1, y_2, ..., y_N) = p(y_1) \cdot p(y_2) \cdot ... p(y_N) = L(y_1, y_2, ..., y_N)$$

- E' il prodotto delle probabilità semplici.

- Questa è la **Funzione di verosimiglianza** o **funzione di Likelihood**, L(.)

Funzione di verosimiglianza (riassunto)

- Data una serie di misure y_i $i=1\ldots N$ di variabili casuali…
- … Nota la distribuzione statistica dell'errore sulla misura espressa come densità di probabilità di ciascuna variabile casuale…
- … Sotto l'ipotesi che le variabili siano tra loro indipendenti…
- … E' possibile scrivere la funzione di verosimiglianza come il prodotto delle probabilità di ciascuna misura y_i $i=1\ldots N$.

Stima alla massima verosimiglianza caso Gaussiano

- Supponiamo il vettore **y** corrisponda a N realizzazioni di una variabile Gaussiana a media μ, deviazione standard σ (N misure indipendenti di una stessa quantità)
- La funzione verosimiglianza dipende da μ e σ che non sono note a-priori.

$$L(y_1, y_2, ..., y_N \mid \mu, \sigma) = p(y_1) \cdot p(y_2) \cdot ... p(y_N) =$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_1 - \mu}{\sigma} \right)^2 \right] \cdot ... \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_N - \mu}{\sigma} \right)^2 \right] =$$

- Quale sarà il valore più verosimile per μ e σ?

Stima alla massima verosimiglianza

- Se **massimizziamo** $L=L(\mathbf{y} \mid \mu,\sigma)$ rispetto a μ e σ

- troviamo i parametri μ,σ tali per cui è massima la probabilità di misurare il vettore di dati $\mathbf{y} = \{y_i, i=1\ldots N\}$.

- **Stima alla massima verosimiglianza**.

- Più in generale, le variabili possono avere densità di probabilità diverse, ciascuna descritta da un set di parametri. I parametri delle diverse densità di probabilità possono essere calcolati utilizzando l'approccio alla massima verosimigianza…

- La funzione di verosimiglianza dipende dai parametri che definiscono le densità di probabilità delle variabili casuali che entrano nella verosimiglianza…

- Massimizzando la funzione di verosimiglianza rispetto a tali parametri se ne effettua la stima in modo tale che il vettore osservato $y = \{y_i\}$ $i = 1\ldots N$ sia massimamente probabile (massima verosimiglianza).

- E' una forma di problema inverso.

Stima alla massima verosimiglianza Il caso gaussiano

- E' solitamente più facile minimizzare il logaritmo negativo della verosimiglianza, f(.), (prodotto → sommatoria)

$$f\left(y_1, y_2, ..., y_N \mid \mu, \sigma\right) = -\ln\left[L\left(y_1, y_2, ..., y_N \mid \mu, \sigma\right)\right] =$$

$$= -\ln\left\{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right]\right\} =$$

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

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

Determino μ

- Per trovare il minimo, poniamo a zero le derivate:

$$\frac{\partial f\left(y_1, y_2 .... y_N \mid \mu, \sigma\right)}{\partial \mu} = \frac{\partial}{\partial \mu}\left\{N \cdot \ln\left(\sqrt{2\pi}\right) + N \cdot \ln(\sigma) + \sum_{i=1}^{N}\left[\frac{1}{2} \cdot \left(\frac{y_i - \mu}{\sigma}\right)^2\right]\right\} =$$

$$= 0 + 0 + \sum_{i=1}^{N} \frac{1}{2} \cdot \left(\frac{y_i - \mu}{\sigma}\right) \cdot 2 \cdot \left(-\frac{1}{\sigma}\right) = -\frac{1}{\sigma^2} \cdot \sum_{i=1}^{N}\left(y_i - \mu\right) = 0 \Rightarrow$$

$$\sum_{i=1}^{N}\left(y_i - \mu\right) = 0 \Rightarrow \sum_{i=1}^{N} y_i = N \cdot \mu \Rightarrow \mu = \frac{\sum_{i=1}^{N} y_i}{N}$$

Media campionaria!

Determino σ

- Per trovare il minimo, poniamo a zero le derivate:

$$\frac{\partial f\left(y_1, y_2, \ldots, y_N \mid \mu, \sigma\right)}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left\{N \cdot \ln\left(\sqrt{2\pi}\right) + N \cdot \ln(\sigma) + \sum_{i=1}^{N}\left[\frac{1}{2} \cdot \left(\frac{y_i - \mu}{\sigma}\right)^2\right]\right\} =$$

$$= 0 + \frac{N}{\sigma} + \sum_{i=1}^{N} 2\frac{1}{2} \cdot \left(\frac{y_i - \mu}{\sigma}\right) \cdot (y_i - \mu) \cdot \left(-\frac{1}{\sigma^2}\right) =$$

$$= \frac{N}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{N}(y_i - \mu)^2 = 0 \Rightarrow N - \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - \mu)^2 = 0 \Rightarrow$$

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

Varianza campionaria!

Overview

Densità di probabilità

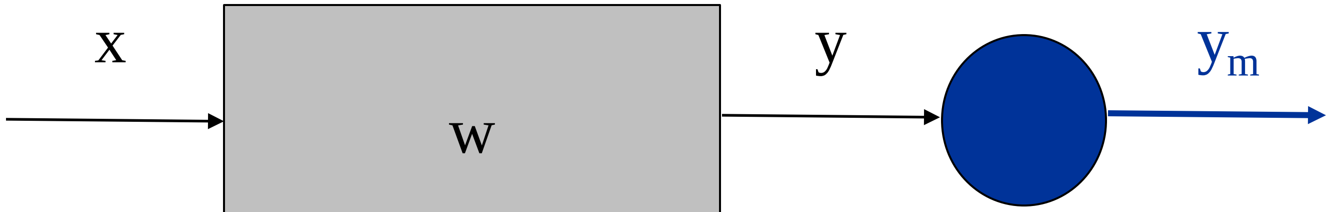
Stimatori

Modelli

Stima alla massima verosimiglianza di parametri di un modello

Modello

$$y = f(x \mid w)$$



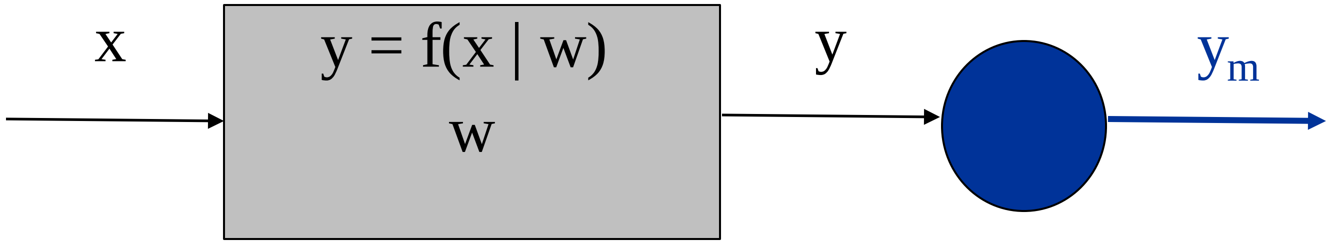
x – causa => y_m – effetto (misurato)

Control / Prediction: determine {y} from {x},{w}

Inverse problem: determine cause {x} from {y_m},{w}

Inverse problem: Identification: determine {w} from {x},{y_m} - ***Supervised learning***

Controllo/Predizione



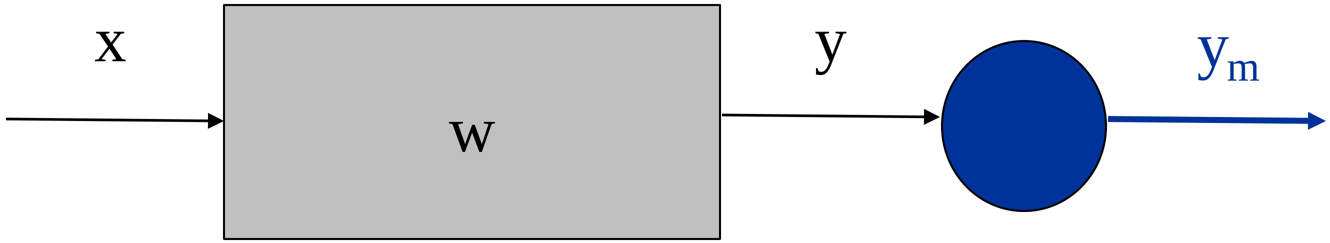
x – causa => y – effetto

Control / Prediction: determine {y} from {x},{w}

$$p\left(y_{m1} \mid \mu,\sigma\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\exp\left[-\frac{1}{2}\left(\frac{y_1-\mu}{\sigma}\right)^2\right]$$

$$p\left(y_1 \mid \mu,\sigma\right)= f\left(\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\exp\left[-\frac{1}{2}\left(\frac{f\,(x_1 \mid w)-\mu}{\sigma}\right)^2\right]\right)$$

Controllo/Predizione



$y = f(x \mid w)$

$$p(y_1 \mid \mu, \sigma) = f\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{f(x_1 \mid w) - \mu}{\sigma}\right)^2\right]\right)$$

$$p(y_2 \mid \mu, \sigma) = f\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{f(x_2 \mid w) - \mu}{\sigma}\right)^2\right]\right)$$

$$p(y_3 \mid \mu, \sigma) = f\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{f(x_3 \mid w) - \mu}{\sigma}\right)^2\right]\right)$$

Siamo in grado di predirre y: p(y) per ogni x (dato il modello f(.))

Funzione di verosimiglianza

- Siano date **N variabili casuali <u>indipendenti</u>**… Quale è la **probabilità di misurare il vettore [y_1, …, y_N]**?

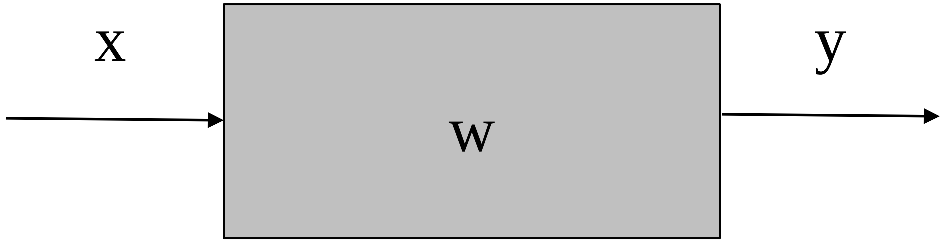
$$p(y_1, y_2, \ldots, y_N) = p(y_1) \cdot p(y_2) \cdot \ldots p(y_N) = L(y_1, y_2, \ldots, y_N)$$

- E' il prodotto delle probabilità semplici.

- Questa è la **Funzione di verosimiglianza** o **funzione di Likelihood**, L(.)

- In questo caso le y sono legate alle x da f(x).

I due problemi inversi con i modelli



$y = f(x \mid w)$

x – causa => y – effetto (misurato)

Inverse problem: determine cause {x} from {y} {w}

Inverse problem: Identification: determine {w} from {x};{y} - *Supervised learning*

$$p(y_1, y_2, \ldots, y_N) = p(y_1) \cdot p(y_2) \cdot \ldots p(y_N) = L(y_1, y_2, \ldots, y_N)$$

$$p(y_1, y_2, \ldots, y_N \mid x, w) = p(y_1 \mid x_1, w) \cdot p(y_2 \mid x_2, w) \cdot \ldots p(y_N \mid x_N, w) = L(y_1, y_2, \ldots, y_N \mid x, w)$$

Overview

Densità di probabilità

Stimatori

Modelli

Stima alla massima verosimiglianza di parametri di un modello

Stima alla massima verosimiglianza di un modello

- Se **massimizziamo** L=L(**y** | x, w) rispetto a w

- troviamo i parametri w tali per cui è massima la probabilità di misurare il vettore di dati **y = {**y_i**,** i=1…N}.

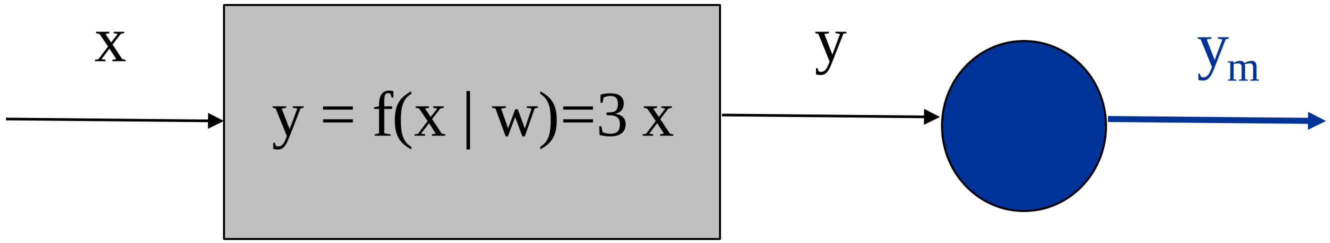
- **Stima alla massima verosimiglianza**.

- Più in generale, le variabili possono avere densità di probabilità diverse, ciascuna descritta da un set di parametri. I parametri delle diverse densità di probabilità possono essere calcolati utilizzando l'approccio alla massima verosimigianza…

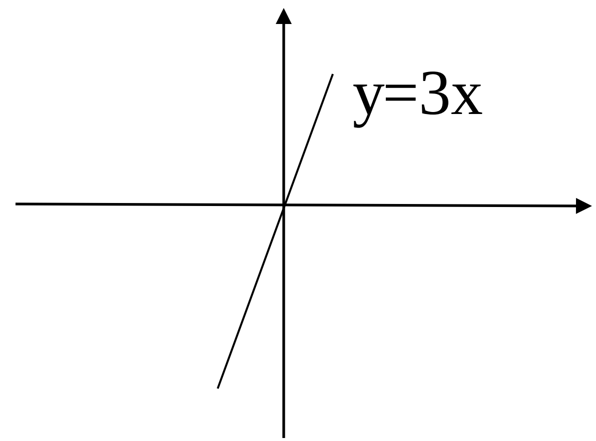
- La funzione di verosimiglianza dipende dai parametri che definiscono le densità di probabilità delle variabili casuali che entrano nella verosimiglianza…

- Massimizzando la funzione di verosimiglianza rispetto a tali parametri se ne effettua la stima in modo tale che il vettore osservato y = {y_i} i =1…N sia massimamente probabile (massima verosimiglianza).

Modello lineare



$x = 0, \quad y = 0, \quad y_m = 0.0001$
$x = +1, \; y = +3, \; y_m = 2.997$
$x = -1, \; y = -3, \; y_m = -3.011$



$$p(y_1, y_2, \ldots, y_N) = p(y_1) \cdot p(y_2) \cdot \ldots p(y_N) = L(y_1, y_2, \ldots, y_N \mid k, \mu, \sigma)$$

Unica incognita è k

Derivazione del parametro k

$$\underset{k}{Max}\left(L\left(y_1, y_2, ..., y_N \mid k, \mu, \sigma\right)\right) = -\min_k\left(\log\left(L\left(y_1, y_2, ..., y_N \mid k, \mu, \sigma\right)\right)\right) =$$

$$-\min_k\left(\log \Pi_i \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_{mi}-y_i)^2}{\sigma^2}}\right)\right) =$$

NB $\quad y_i = kx_i$

$<y_m - y_i> = \mu$

$$-\min_k\left(\Sigma_i \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_{mi}-y_i-\mu)^2}{\sigma^2}}\right)\right) = +\min_k\left(3\sqrt{2\pi}\sigma + \Sigma_i \frac{1}{2}\frac{(y_{mi}-y_i-\mu)^2}{\sigma^2}\right) =$$

$$+\min_k\left(3\sqrt{2\pi}\sigma + \frac{1}{2}\frac{(0.001-k\bullet 0-\mu)^2}{\sigma^2} + \frac{1}{2}\frac{(2.997-k\bullet 1-\mu)^2}{\sigma^2} + \frac{1}{2}\frac{(-3.011-k\bullet(-1)-\mu)^2}{\sigma^2}\right) =$$

Minimizzazione di –log(L)

$$\min_k F(.) = \left(3\sqrt{2\pi}\sigma + \frac{1}{2}\frac{(0.001 - k\cdot 0 - \mu)^2}{\sigma^2} + \frac{1}{2}\frac{(2.997 - k\cdot 1 - \mu)^2}{\sigma^2} + \frac{1}{2}\frac{(-3.011 - k\cdot(-1) - \mu)^2}{\sigma^2}\right)$$

$$\frac{\partial F(.)}{dk} = 0$$

$$\frac{\partial\left(3\sqrt{2\pi}\sigma + \frac{1}{2}\frac{(0.001 - k\bullet 0 - \mu)^2}{\sigma^2} + \frac{1}{2}\frac{(2.997 - k\bullet 1 - \mu)^2}{\sigma^2} + \frac{1}{2}\frac{(-3.011 - k\bullet(-1) - \mu)^2}{\sigma^2}\right)}{dk} = 0$$

$$\frac{(0.001 - k\bullet 0 - \mu)\cdot 0}{\sigma^2} + \frac{(2.997 - k\bullet 1 - \mu)(-1)}{\sigma^2} + \frac{(-3.011 - k\bullet(-1) - \mu)(+1)}{\sigma^2} = 0$$

$$\frac{1}{\sigma^2}(-2.997 + k - 3.011 + k) + \frac{\mu}{\sigma^2} - \frac{\mu}{\sigma^2} = 0$$

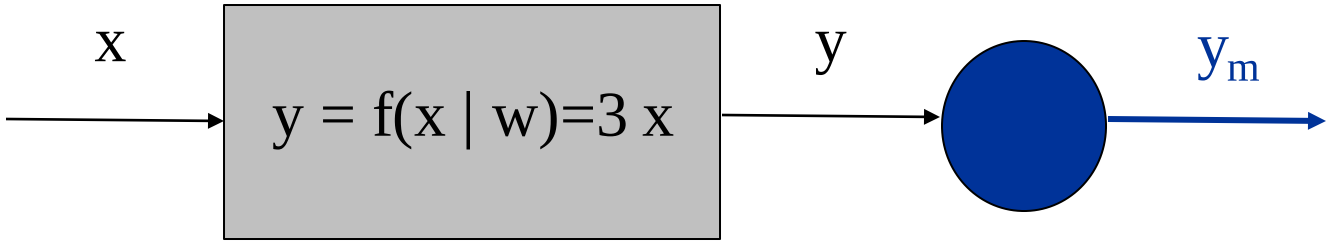
$$\frac{1}{\sigma^2}(-2.997 + k - 3.011 + k) = 0 \implies -6.008 + 2k = 0$$

$$k = 3.004$$

μ

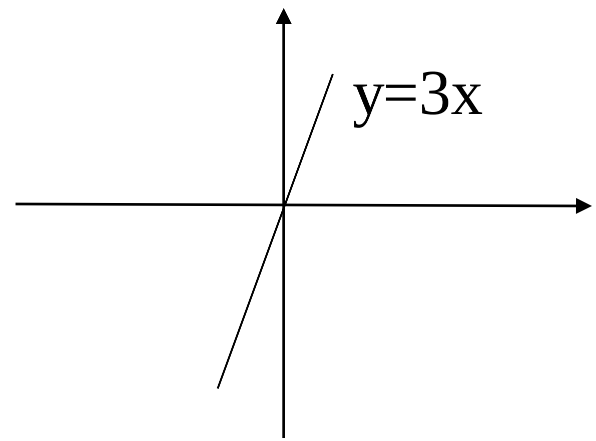
Eliminazione della polarizzazione introdotta dal rumore. Qui rumore a media nulla, $\mu = 0$

Modello lineare



$x = 0, \quad y = 0, \quad y_m = 5.0001$
$x = +1, y = +3, y_m = 5.997$
$x = -1, \; y = -3, \; y_m = -0.011$

Rumore con $\mu = 3$ e stessa varianza



$$p(y_1, y_2, \ldots, y_N) = p(y_1) \cdot p(y_2) \cdot \ldots p(y_N) = L(y_1, y_2, \ldots, y_N \mid k, \mu, \sigma)$$

Unica incognita è k

Minimizzazione di –log(L)

$$\min_k F(.) = \left(3\sqrt{2\pi}\sigma + \frac{1}{2}\frac{(3.001 - k\cdot 0 - \mu)^2}{\sigma^2} + \frac{1}{2}\frac{(5.997 - k\cdot 1 - \mu)^2}{\sigma^2} + \frac{1}{2}\frac{(-0.011 - k\cdot(-1) - \mu)^2}{\sigma^2} \right)$$

$$\frac{\partial F(.)}{dk} = 0$$

$$\frac{\partial \left(3\sqrt{2\pi}\sigma + \frac{1}{2}\frac{(3.001 - k\bullet 0 - \mu)^2}{\sigma^2} + \frac{1}{2}\frac{(5.997 - k\bullet 1 - \mu)^2}{\sigma^2} + \frac{1}{2}\frac{(-0.011 - k\bullet(-1) - \mu)^2}{\sigma^2} \right)}{dk} = 0$$

$$\frac{(3.001 - k\bullet 0 - \mu)\cdot 0}{\sigma^2} + \frac{(5.997 - k\bullet 1 - \mu)(-1)}{\sigma^2} + \frac{(-0.011 - k\bullet(-1) - \mu)(+1)}{\sigma^2} = 0$$

$$\frac{1}{\sigma^2}(-5.997 + k - 0.011 + k) + \frac{\mu}{\sigma^2} - \frac{\mu}{\sigma^2} = 0$$

$$\frac{1}{\sigma^2}(-5.997 + k - 0.011 + k) = 0 \quad ==> \quad -6.008 + 2k = 0$$

$$k = 3.004$$

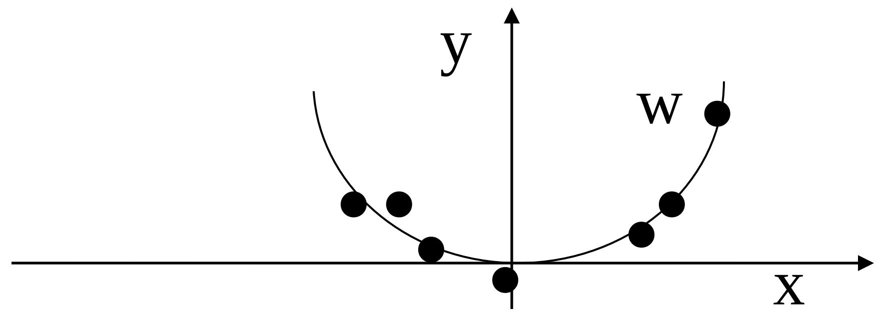
μ

Eliminazione della polarizzazione introdotta dal rumore. Qui rumore a media non nulla, μ = 3. Perche' stesso risultato? (analizzate il residuo).

Esempio di Modello - II

$y=f(x;\ w)$

$y = a\ x^2$



Identificazione: determino i parametri **w** che fittano i punti campionati. La funzione z=f(x) varie forma con il parametro **a** in questo caso. A partire da insiemi di **P**{x,z} determino **w**.

Predizione: Utilizzo il modello (a noto) per predire l'uscita, y_i, in funzione dell'ingresso x_i.

Nota: la funzione f è non lineare in x, ma il modello è lineare in a.

In generale, problemi multi input e multi output: **x, y, w** vettori.

Overview

Densità di probabilità

Stimatori

Modelli

Stima alla massima verosimiglianza di parametri di un modello