Computazione per l'interazione naturale: Regressione lineare probabilistica

Corso di Interazione uomo-macchina II

Prof. Giuseppe Boccignone

Dipartimento di Informatica
Università di Milano

boccignone@di.unimi.it
http://boccignone.di.unimi.it/IUM2_2014.html

Regressione lineare //modelli probabilistici: stima di ML

- Assumiamo rumore gaussiano additivo $t = f(x; \mathbf{w}) + \epsilon$

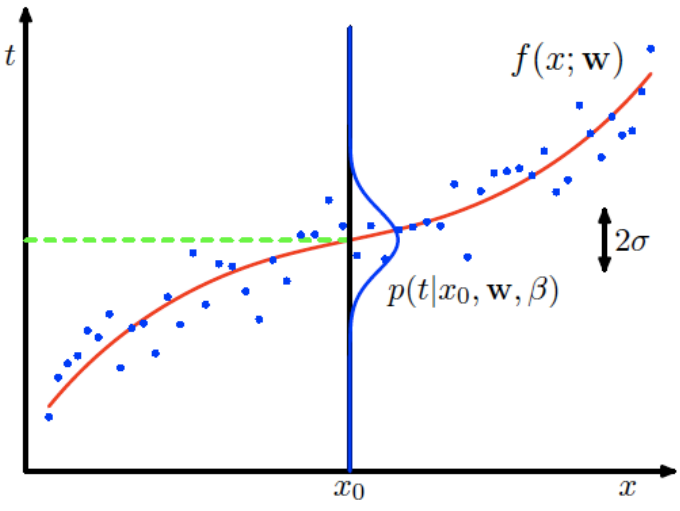
$$\epsilon \sim \mathcal{N}(0, \sigma)$$

precisione $\beta = 1/\sigma^2$

- Allora

$$t|x \sim \mathcal{N}(f(x; \mathbf{w}), \sigma)$$

$$p(t|x) = \mathcal{N}(f(x; \mathbf{w}), \sigma)$$



Regressione lineare //modelli probabilistici: stima di ML

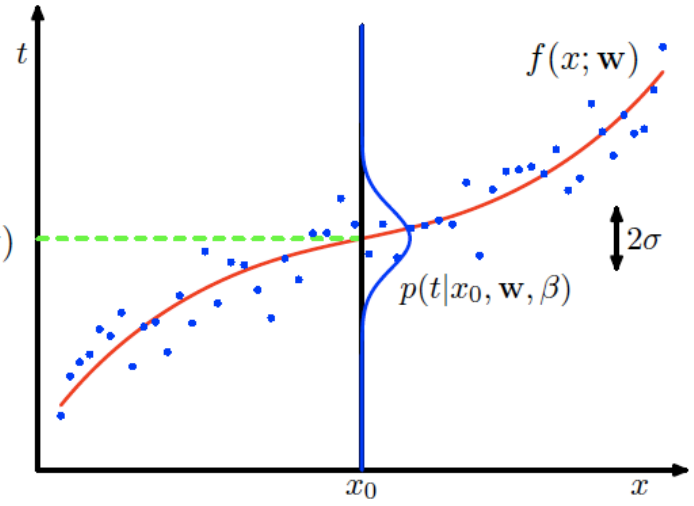
- Abbiamo N osservazioni nel training set

$$(x_1, t_1), \cdots, (x_N, t_N) = (\mathbf{x}, \mathbf{t})$$

- Considerando la probabilità congiunta dei target

$$p(t_1, t_2, \cdots, t_N | x_1, x_2, \cdots, x_N, \mathbf{w}) = p(\mathbf{t}|\mathbf{x}, \mathbf{w})$$

- Vogliamo stimare i parametri w



Introduzione alla Statistica Bayesiana //metodologia generale

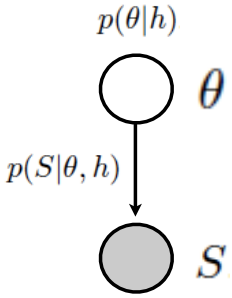
- Obiettivo: predire un dato x sulla base di n osservazioni S = {x1, . . . , xn}
- Tre step:
  - Specifica del modello h con parametri di tuning θ per la generazione dei dati in S
  - Inferenza (learning dei parametri)

$$p(\theta|S, h) = \frac{p(S|\theta, h)p(\theta|h)}{p(S|h)} \propto p(S|\theta, h)p(\theta|h)$$

  - Predizione

$$p(x|S) = \int p(x, \theta|S)d\theta = \int p(x|\theta)p(\theta|S)d\theta$$

Modello o ipotesi h



Statistica Bayesiana //stime puntuali

- Rinunciando ad un approccio completamente Bayesiano, si possono ottenere stime puntuali dei parametri (ovvero i parametri diventano “numeri” e non VA)
  - Stima Maximum A Posteriori (MAP)

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|S)$$

  - Stima di massima verosimiglianza (Maximum Likelihood, ML)

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\, p(S|\theta)$$

$$l(\theta) = \underset{\theta}{\operatorname{argmax}}\, \log p(S|\theta)$$

Esempio: stima di massima verosimiglianza //caso Gaussiano

- Insieme di campioni x1, . . . , xn da distribuzione Gaussiana di parametri ignoti (identicamente distribuiti)
- Campioni estratti indipendentemente:

$$p(S|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

L’ipotesi i.i.d

- allora

$$\hat{\theta} = L(\theta) = \underset{\theta}{\operatorname{argmax}}\, p(S|\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{k=1}^{n} p(x_k|\theta)$$

- oppure usando la log-verosimiglianza

$$l(\theta) = \underset{\theta}{\operatorname{argmax}}\, \log p(S|\theta) = \underset{\theta}{\operatorname{argmax}}\, \log \sum_{k=1}^{n} p(x_k|\theta)$$

$$\Rightarrow \quad \nabla_\theta \log p(S|\theta) = 0$$

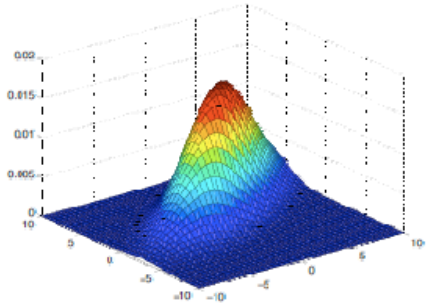
Esempio: stima di massima verosimiglianza //caso Gaussiano

- Un vettore aleatorio $\mathbf{x} \in \mathbb{R}^n$

$$\mathbf{x} \sim Normal(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

forma quadratica di **x**



$$(x_1, t_1), \cdots, (x_N, t_N) = (\mathbf{x}, \mathbf{t})$$

- per il singolo campione

$$\log p(\mathbf{x}_i|\mu) = -\frac{1}{2}\log\left((2\pi)^d|\boldsymbol{\Sigma}|\right) - \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$$

Esempio: stima di massima verosimiglianza // $\boldsymbol{\mu}$ della Gaussiana

- Se $\mathbf{x} \sim Normal(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con covarianza nota e media da stimare
- per il singolo campione

$$\log p(\mathbf{x}_i|\mu) = -\frac{1}{2}\log\left((2\pi)^d|\boldsymbol{\Sigma}|\right) - \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$$

- derivando il secondo termine rispetto a $\boldsymbol{\mu}$ (il primo è costante rispetto a teta)

$$\nabla_\theta \log p(\mathbf{x}_i|\mu) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$$

$$\frac{\partial}{\partial \mathbf{s}}(\mathbf{x}-\mathbf{s})^T\mathbf{W}(\mathbf{x}-\mathbf{s}) = -2\mathbf{W}(\mathbf{x}-\mathbf{s})$$

- per tutti i campioni

$$\nabla_\theta \log p(S|\mu) = \nabla_\theta \sum_{i=1}^{n} \log p(x_i|\mu) = \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$$

- ponendo uguale a 0

$$\sum_{i=1}^{n}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}) = 0 \quad \Rightarrow \quad \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

media empirica

Esempio: stima di massima verosimiglianza // Parametri della Gaussiana

- Se $\mathbf{x} \sim Normal(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con covarianza e media da stimare
- per il singolo campione

$$\log p(\mathbf{x}_i|\mu) = -\frac{1}{2}\log\left((2\pi)^d|\boldsymbol{\Sigma}|\right) - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

- derivando il secondo termine rispetto a $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ ripetendo il procedimento di prima

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

media empirica

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

covarianza empirica

Regressione lineare //stima di ML dei parametri

- Training set $(x_1, t_1), \cdots, (x_N, t_N) = (\mathbf{x}, \mathbf{t})$

L'ipotesi i.i.d

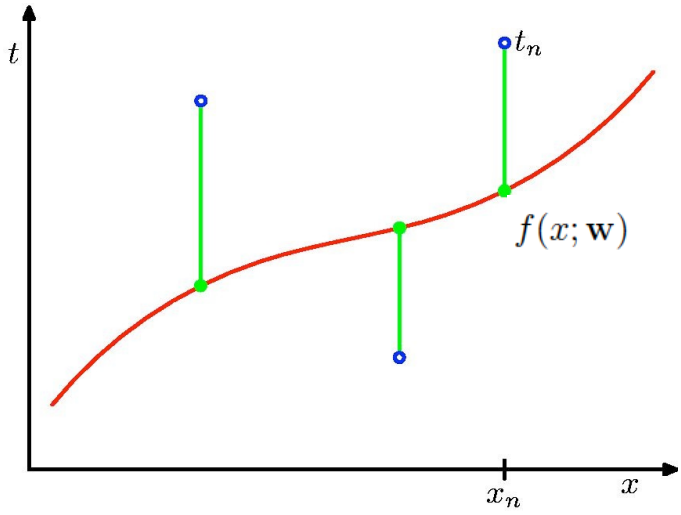
- La likelihood è

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma) = \prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma) = \prod_{n=1}^{N} \mathcal{N}(f(x_n; \mathbf{w}), \sigma)$$

$$\begin{aligned}
\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma) &= \sum_{n=1}^{N} \log p(t_n|x_n, \mathbf{w}, \sigma) \\
&= \sum_{n=1}^{N} \log \mathcal{N}(f(x_n; \mathbf{w}), \sigma) \\
&= \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}|t_n - f(x_n; \mathbf{w})|^2\right) \\
&= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}|t_n - f(x_n; \mathbf{w})|^2
\end{aligned}$$

funzione di costo

Regressione lineare //modelli probabilistici: stima di ML



Funzione di loss (errore) quadratica

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}|t_n - f(x_n;\mathbf{w})|^2$$

Regressione lineare //modelli probabilistici: stima di ML

- Abbiamo che

$$\max \quad \log p(\mathbf{t}|\mathbf{x},\mathbf{w},\sigma) \quad \Longleftrightarrow \quad \min E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}|t_n - f(x_n;\mathbf{w})|^2$$

punto stazionario

$$\frac{\partial \log L}{\partial \mathbf{w}} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n(t_n - \mathbf{x}_n^{\mathsf{T}}\mathbf{w})$$

$$= \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n t_n - \mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\mathbf{w} = \mathbf{0}.$$

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^{\mathsf{T}}\\ \mathbf{x}_2^{\mathsf{T}}\\ \vdots\\ \mathbf{x}_N^{\mathsf{T}}\end{bmatrix} = \begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_N\end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix}t_1\\ t_2\\ \vdots\\ t_N\end{bmatrix}$$

$$\sum_{n=1}^{N}\mathbf{x}_n t_n \longrightarrow \mathbf{X}^{\mathsf{T}}\mathbf{t}$$

$$\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\mathbf{w} \longrightarrow \mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w}$$

$$\frac{\partial \log L}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^{\mathsf{T}}\mathbf{t} - \mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w}) = \mathbf{0}.$$

Regressione lineare //modelli probabilistici: stima di ML

- Abbiamo che

$$\max \quad \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma) \quad \Longleftrightarrow \quad \min E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} |t_n - f(x_n; \mathbf{w})|^2$$

$$\sum_{n=1}^{N}(t_n - \mathbf{x}^T\widehat{\mathbf{w}})^2$$

$$\Rightarrow \quad \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) = 0$$

punto stazionario

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$$

negativa
il punto stazionario è un max

$$\Rightarrow \quad \widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}.$$

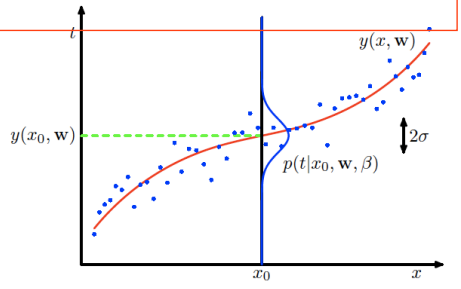
soluzione di ML
=
equazioni normali
per i minimi quadrati

Regressione lineare //modelli probabilistici: stima di ML

- Relazione tra parametri del modello e parametri stimati:

$$\begin{aligned}
E_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma)}\left[\widehat{\mathbf{w}}\right] &= \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma) d\mathbf{t} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \int \mathbf{t} p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma) d\mathbf{t} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma)}\left[\mathbf{t}\right] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w} \\
&= \mathbf{w}
\end{aligned}$$

la curva deterministica



- ML è stimatore unbiased!

Regressione lineare
//modelli probabilistici: stima di ML

- Analogamente possiamo anche fornire una stima ML della varianza del rumore, assumendo $\mathbf{w} = \widehat{\mathbf{w}}$

$$\frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{x}^T\widehat{\mathbf{w}})^2 = 0$$

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^T\widehat{\mathbf{w}})^2 = \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^T(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})$$

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}}) \\ &= \frac{1}{N}\mathbf{t}^\mathsf{T}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\right)\mathbf{t} \\ &= \frac{1}{N}\left(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\hat{\mathbf{t}}\right)\end{aligned}$$

$$\frac{\partial \mathcal{L}^2}{\partial \sigma \partial \sigma} = -\frac{2N^2}{(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})} = -\frac{2N}{\widehat{\sigma^2}}$$

negativa
il punto stazionario è un max

Regressione lineare
//stima di ML dei parametri di generiche funzioni di x

$$f(x_n; \mathbf{w}) = \mathbf{w}^T\phi(\mathbf{x}_i)$$

- La likelihood è

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N}\mathcal{N}(t_i|\mathbf{w}^T\phi(\mathbf{x}_i), \beta^{-1})$$

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^{N}\ln\mathcal{N}(t_i|\mathbf{w}^T\phi(\mathbf{x}_i), \beta^{-1}) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})$$

- Dove la funzione di costo o di errore è

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(t_i - \mathbf{w}^T\phi(\mathbf{x}_i)\right)^2$$

Regressione lineare
//stima di ML dei parametri di generiche funzioni di x

$$f(x_n; \mathbf{w}) = \mathbf{w}^T \phi(\mathbf{x}_i)$$

$$\max \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) \Longleftrightarrow \min E_D(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \left(t_i - \mathbf{w}^T \phi(\mathbf{x}_i) \right)^2$$

$$\Rightarrow \quad 0 = \nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^{N} \left(t_i - \mathbf{w}^T \phi(\mathbf{x}_i) \right) \phi(\mathbf{x}_i)^T$$

$$= \sum_{i=1}^{N} t_i \phi(\mathbf{x}_i)^T - \mathbf{w}^T \left(\sum_{i=1}^{N} \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^T \right)$$

$$\Rightarrow \quad \mathbf{w}_{ML} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t}$$

equazioni normali
per i minimi quadrati

Regressione lineare
//modelli probabilistici: stima di ML

- Abbiamo generalizzato il risultato già ottenuto per la retta

$$\mathbf{w}_{ML} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t}$$

equazioni normali
per i minimi quadrati

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

design
matrix

Regressione lineare //modelli probabilistici: stima di ML

- Possiamo computare il Minimo di ED(w) anche in modo non analitico: metodo di discesa del gradiente (procedura iterativa)
  - Inizializzazione

$$\mathbf{w}^{(0)} = (w_0^{(0)}, w_1^{(0)}, \ldots, w_D^{(0)}) \longrightarrow E_D(\mathbf{w}^{(0)}) = \frac{1}{2}\sum_{i=1}^{N}(t_i - (\mathbf{w}^{(0)})^T\phi(\mathbf{x}_i))^2$$

  - while (~ condizioneTerminazione)

$$w_j^{(i)} := w_j^{(i-1)} - \alpha \frac{\partial E_D(\mathbf{w})}{\partial w_j}\bigg|_{\mathbf{w}=\mathbf{w}^{(i-1)}}$$

    i = i+1;

Regressione lineare //modelli probabilistici: stima di ML

- Possiamo computare il Minimo di ED(w) anche in modo non analitico: metodo di discesa del gradiente (procedura iterativa)
  - Inizializzazione

$$\mathbf{w}^{(0)} = (w_0^{(0)}, w_1^{(0)}, \ldots, w_D^{(0)}) \longrightarrow E_D(\mathbf{w}^{(0)}) = \frac{1}{2}\sum_{i=1}^{N}(t_i - (\mathbf{w}^{(0)})^T\phi(\mathbf{x}_i))^2$$

  - while (~ condizioneTerminazione)

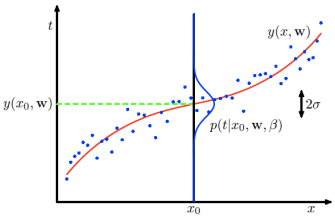
$$\mathbf{w}^{(i)} := \mathbf{w}^{(i-1)} - \alpha(t_i - \mathbf{w}^{(i-1)}\phi(\mathbf{x}_i))\phi(\mathbf{x}_i)$$

    i = i+1;

Regressione lineare //predizione e incertezza

•Possiamo effettuare la predizione sui dati nuovi (test set)

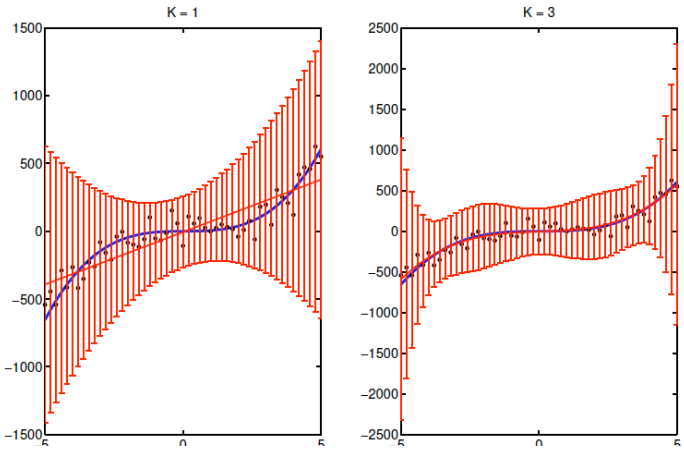
$$t_{new} = \widehat{\mathbf{w}}^T \mathbf{x}_{new}$$



•Qual è l'incertezza sulle stime del nostro modello?

•Usiamo la covarianza...

$$\widehat{t}_{new} \pm \sigma^2_{new}$$



Regressione lineare //predizione e incertezza

•Possiamo effettuare la predizione sui dati nuovi (test set)

$$t_{new} = \widehat{\mathbf{w}}^T \mathbf{x}_{new}$$

•Qual è l'incertezza sulle stime del nostro modello?

•Usiamo la covarianza...

$$\widehat{t}_{new} \pm \sigma^2_{new}$$

$$\begin{aligned}
\sigma^2_{new} &= var\{t_{new}\} = E_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma)}\left[t^2_{new}\right] - \left(E_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma)}\left[t_{new}\right]\right)^2 \\
&= \sigma^2 \mathbf{x}^T_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new} \\
&= \mathbf{x}^T_{new} cov\{\widehat{\mathbf{w}}\}\mathbf{x}_{new}
\end{aligned}$$

l'incertezza è legata alla covarianza dei parametri stimati !!

Regressione lineare //modelli probabilistici: stima di ML

- Qual è l'incertezza sulle stime del nostro modello?
- E' legata alla covarianza dei parametri stimati...

$$\text{cov}\{\hat{\mathbf{w}}\} = E\{(\hat{\mathbf{w}} - E\{\hat{\mathbf{w}}\})(\hat{\mathbf{w}} - E\{\hat{\mathbf{w}}\})^{\mathsf{T}}\} = E\{\hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathsf{T}}\} - E\{\hat{\mathbf{w}}\}E\{\hat{\mathbf{w}}^{\mathsf{T}}\}$$

- Poichè ML è stimatore unbiased: $E\{\hat{\mathbf{w}}\} = \mathbf{w}$

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} \qquad \hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathsf{T}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}\mathbf{t}^{\mathsf{T}}\mathbf{X}\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$$

$$E\{\hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathsf{T}}\} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}E\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\}\mathbf{X}\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$$

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$$

$$\begin{aligned} E\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\} &= E\{(\mathbf{X}\mathbf{w} + \boldsymbol{\epsilon})(\mathbf{X}\mathbf{w} + \boldsymbol{\epsilon})^{\mathsf{T}}\} \\ &= E\{\mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + 2\boldsymbol{\epsilon}\mathbf{w}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\} \\ &= \mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + 2E\{\boldsymbol{\epsilon}\}\mathbf{w}^{\mathsf{T}}\mathbf{X} + E\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\} \\ &= \mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + \sigma^2\mathbf{I} \end{aligned}$$

rumore a media 0

Regressione lineare //modelli probabilistici: stima di ML

- Qual è l'incertezza sulle stime del nostro modello?
- E' legata alla covarianza dei parametri stimati...

$$\text{cov}\{\hat{\mathbf{w}}\} = E\{(\hat{\mathbf{w}} - E\{\hat{\mathbf{w}}\})(\hat{\mathbf{w}} - E\{\hat{\mathbf{w}}\})^{\mathsf{T}}\} = E\{\hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathsf{T}}\} - E\{\hat{\mathbf{w}}\}E\{\hat{\mathbf{w}}^{\mathsf{T}}\}$$

- Si ottiene

$$\begin{aligned} E\{\hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathsf{T}}\} &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}E\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\}\mathbf{X}\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} \\ &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + \sigma^2\mathbf{I})\mathbf{X}\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} \\ &= \mathbf{w}\mathbf{w}^{\mathsf{T}} + \sigma^2\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} \end{aligned}$$

$$\text{cov}\{\hat{\mathbf{w}}\} = E\{\hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathsf{T}}\} - E\{\hat{\mathbf{w}}\}E\{\hat{\mathbf{w}}^{\mathsf{T}}\} = \mathbf{w}\mathbf{w}^{\mathsf{T}} + \sigma^2\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} - \mathbf{w}\mathbf{w}^{\mathsf{T}}$$

$$= \sigma^2\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$$

$$\text{cov}\{\hat{\mathbf{w}}\} = E\{\hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathsf{T}}\} - E\{\hat{\mathbf{w}}\}E\{\hat{\mathbf{w}}^{\mathsf{T}}\} = -\left(\frac{\partial^2\mathcal{L}}{\partial\mathbf{w}\partial\mathbf{w}^{\mathsf{T}}}\right)^{-1}$$

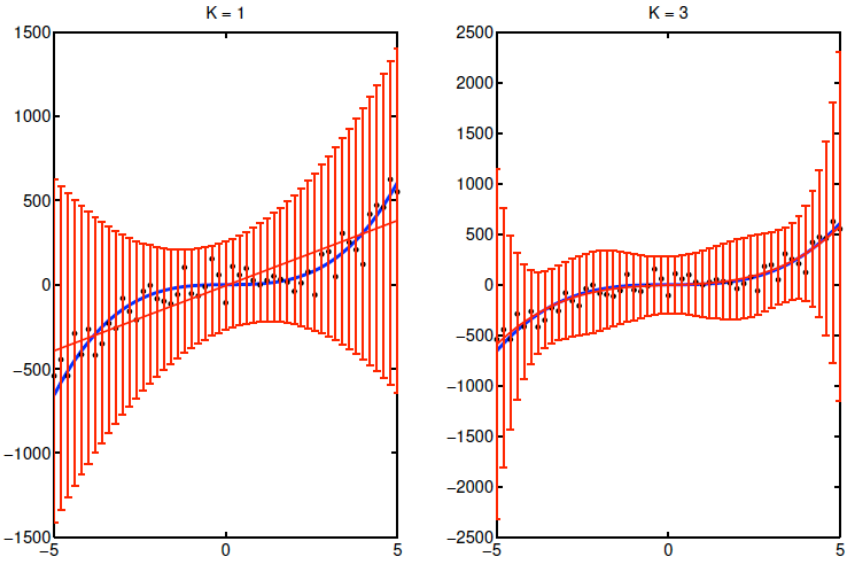
piccola curvatura = elevata varianza = parametri poco significativi

Regressione lineare //modelli probabilistici: stima di ML

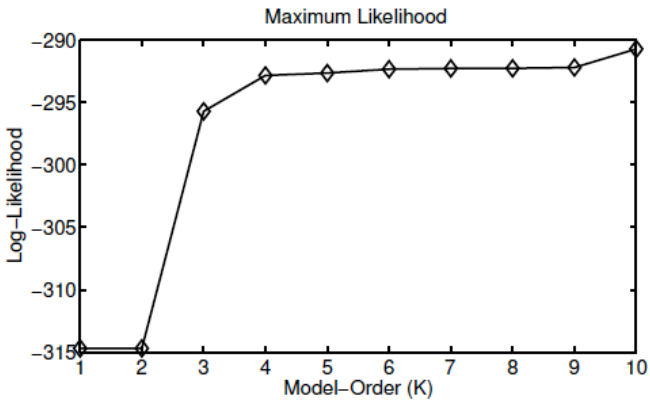
- Qual è l'incertezza sulle stime del nostro modello?

$$\widehat{t}_{new} \pm \sigma^2_{new}$$

$$\begin{aligned}\widehat{t}_{new} &= \mathbf{x}_{new}^{\mathsf{T}} \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} \\ \sigma^2_{new} &= \widehat{\sigma}^2 \mathbf{x}_{new}^{\mathsf{T}} \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{x}_{new} \\ \widehat{\sigma}^2 &= \frac{1}{N}\left(\mathbf{t}^{\mathsf{T}}\mathbf{t} - \mathbf{t}^{\mathsf{T}}\hat{\mathbf{t}}\right)\end{aligned}$$



Regressione lineare //andamento della funzione di log-likelihood



$$\begin{aligned}\log \mathcal{L} &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \widehat{\sigma^2} - \frac{1}{2\widehat{\sigma^2}} N\widehat{\sigma^2} \\ &= -\frac{N}{2}(1 + 2\log 2\pi) - \frac{N}{2}\log \widehat{\sigma^2}\end{aligned}$$

```
Range = 10;
Max_Model_Order = 10;
noise_var = 100;

L=[];
x = [-Range/2:0.2:Range/2]';
N=size(x,1);

f = 5*x.^3  - x.^2 + x;
f_n = f + noise_var*randn(size(x));

[i,j]=sort(x); X=x.^0;

for k=1:Max_Model_Order
    X=[X x.^k];
    w_hat = inv(X'*X)*X'*f_n;
    f_hat = X*w_hat;
    sigma_hat = mean((f_n - f_hat).^2);
    sigma = sigma_hat*diag(X*inv(X'*X)*X');
    L  = [L; -N*log(sqrt(sigma_hat)) - 0.5*N*(1 + log(2*pi))];
    plot(i,f(j),'b');
    hold on
    plot(i,f_n(j),'.k','MarkerSize',15)
    errorbar(i,f_hat(j),sigma(j),'-r.')
    hold off
    pause(1)
end

figure
plot(1:Max_Model_Order,L,'dr--');
```