Scuola universitaria professionale della Svizzera italiana
Bachelor di Ingegneria Informatica

SUPSI

Machine Learning

Lezione 4 - Regressione Lineare: Approccio Generativo

Loris Cannelli, Ricercatore, IDSIA
loris.cannelli@supsi.ch

IDSIA-SUPSI, Galleria 1, Manno

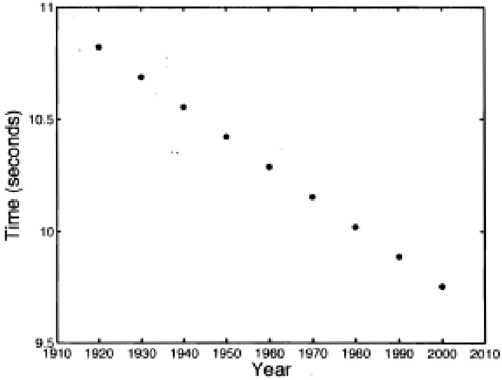
Capitolo 2 del libro
A First Course in Machine Learning

L'approccio generativo

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

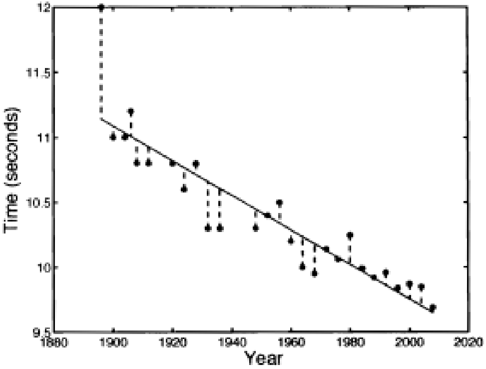
L'approccio generativo

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



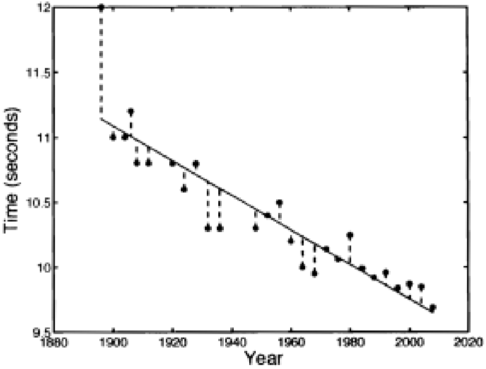
L'approccio generativo

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



L'approccio generativo

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



Come *generare* i punti presenti in figura a partire dal nostro modello lineare?

L'approccio generativo

- Come generare i punti presenti in figura a partire dal nostro modello lineare?

L'approccio generativo

- Come generare i punti presenti in figura a partire dal nostro modello lineare?
- Dovremmo traslare i punti generati da $\mathbf{w}^T\mathbf{x}$

L'approccio generativo

- Come generare i punti presenti in figura a partire dal nostro modello lineare?
- Dovremmo traslare i punti generati da $\mathbf{w}^T\mathbf{x}$
- Possiamo semplicemente aggiungere uno shift

  $$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon$$

  dove ϵ è uno scalare, positivo o negativo

L'approccio generativo

- Come generare i punti presenti in figura a partire dal nostro modello lineare?
- Dovremmo traslare i punti generati da $\mathbf{w}^T\mathbf{x}$
- Possiamo semplicemente aggiungere uno shift

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon$$

  dove ϵ è uno scalare, positivo o negativo

- Potremmo anche modificare il modello lineare con uno *scaling* del tipo $f(\mathbf{x};\mathbf{w}) = \epsilon\mathbf{w}^T\mathbf{x}$, ma quello additivo è più semplice da studiare per iniziare (un modello moltiplicativo di questo tipo descrive in maniera efficace per esempio come degrada la qualità dei pixel in un'immagine)

ϵ

$$f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x} + \epsilon$$

ϵ

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon$$

- Come scegliamo questo shift ϵ che vogliamo sommare al nostro modello lineare?

ϵ

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon$$

- Come scegliamo questo shift ϵ che vogliamo sommare al nostro modello lineare?

- Dato che questo shift deve predirre un comportamento non noto a priori, un modo intuitivo di procedere e presupporre che sia una variabile aleatoria

ϵ

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon$$

- Come scegliamo questo shift ϵ che vogliamo sommare al nostro modello lineare?
- Dato che questo shift deve predirre un comportamento non noto a priori, un modo intuitivo di procedere e presupporre che sia una variabile aleatoria
- Quello che dobbiamo fare, quindi, è definire le caratteristiche di questa variabile aleatoria

Variabile aleatoria Gaussiana

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon$$

Variabile aleatoria Gaussiana

$$f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x} + \epsilon$$

- L'assunzione tipica è che ϵ sia una viariabile aleatoria con distribuzione Gaussiana

Variabile aleatoria Gaussiana

$$f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x} + \epsilon$$

- L'assunzione tipica è che ϵ sia una viariabile aleatoria con distribuzione Gaussiana

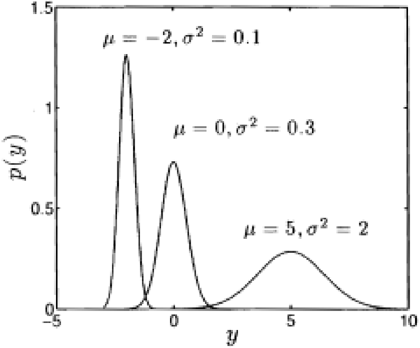
$$p(y|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

Variabile aleatoria Gaussiana

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon$$

- L'assunzione tipica è che ϵ sia una viariabile aleatoria con distribuzione Gaussiana

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



Variabile aleatoria Gaussiana

$$f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}_n + \epsilon_n \quad \forall\, n = 1, \ldots, N$$

Variabile aleatoria Gaussiana

$$f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}_n + \epsilon_n \quad \forall\, n = 1, \ldots, N$$

- Come abbiamo detto, ogni ϵ_n è una variabile aleatoria con distribuzione $p(\epsilon_n)$ Gaussiana

Variabile aleatoria Gaussiana

$$f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}_n + \epsilon_n \quad \forall\, n = 1, \ldots, N$$

- Come abbiamo detto, ogni ϵ_n è una variabile aleatoria con distribuzione $p(\epsilon_n)$ Gaussiana
- Come parametri per la Guassiana, scegliamo per ora valori standard: $\mu = 0, \sigma = 1$

Variabile aleatoria Gaussiana

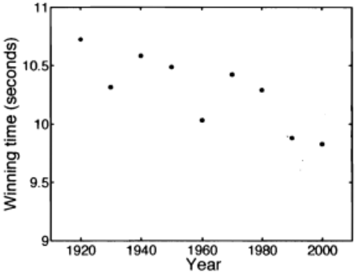
$$f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}_n + \epsilon_n \quad \forall\, n = 1, \ldots, N$$

- Come abbiamo detto, ogni ϵ_n è una variabile aleatoria con distribuzione $p(\epsilon_n)$ Gaussiana
- Come parametri per la Guassiana, scegliamo per ora valori standard: $\mu = 0, \sigma = 1$
- $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ dovrebbero essere in qualche modo correlate tra di loro?

Variabile aleatoria Gaussiana

$$f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}_n + \epsilon_n \quad \forall\, n = 1, \ldots, N$$

- Come abbiamo detto, ogni ϵ_n è una variabile aleatoria con distribuzione $p(\epsilon_n)$ Gaussiana
- Come parametri per la Guassiana, scegliamo per ora valori standard: $\mu = 0, \sigma = 1$
- $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ dovrebbero essere in qualche modo correlate tra di loro?
- Per semplicità assumiamo di no! Ha senso considerare un modello dove gli shift $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ sono indipendenti l'uno dall'altro



Trovare il $\hat{w}$ ottimo

- $t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$

Trovare il $\hat{w}$ ottimo

- $t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- Per le proprietò delle Guassiane: $t_n \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$

Trovare il $\hat{w}$ ottimo

- $t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- Per le proprietò delle Guassiane: $t_n \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$

Esempio

Trovare il $\hat{w}$ ottimo

- $t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- Per le proprietò delle Guassiane: $t_n \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$

Esempio

- $\sigma^2 = 0.05; \mathbf{w} = \begin{bmatrix} 36.416 \\ -0.0133 \end{bmatrix}; \mathbf{x}_n = \begin{bmatrix} 1 \\ 1980 \end{bmatrix}$

Trovare il $\hat{w}$ ottimo

- $t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- Per le proprietò delle Guassiane: $t_n \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$

Esempio

- $\sigma^2 = 0.05; \mathbf{w} = \begin{bmatrix} 36.416 \\ -0.0133 \end{bmatrix}; \mathbf{x}_n = \begin{bmatrix} 1 \\ 1980 \end{bmatrix}$
- $\mu = 36.416 - 0.0133 * 1980 = 10.02$

Trovare il $\hat{w}$ ottimo

- $t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- Per le proprietò delle Guassiane: $t_n \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$

Esempio

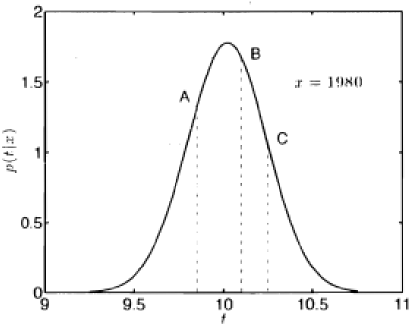
- $\sigma^2 = 0.05; \mathbf{w} = \begin{bmatrix} 36.416 \\ -0.0133 \end{bmatrix}; \mathbf{x}_n = \begin{bmatrix} 1 \\ 1980 \end{bmatrix}$
- $\mu = 36.416 - 0.0133 * 1980 = 10.02$
- Tempo di vittoria reale $t_{1980} = 10.25$

Trovare il $\hat{w}$ ottimo

- $t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- Per le proprietò delle Guassiane: $t_n \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$

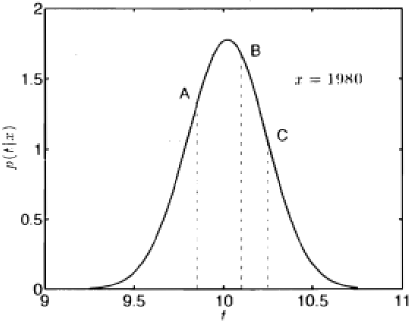
Esempio

- $\sigma^2 = 0.05; \mathbf{w} = \begin{bmatrix} 36.416 \\ -0.0133 \end{bmatrix}; \mathbf{x}_n = \begin{bmatrix} 1 \\ 1980 \end{bmatrix}$
- $\mu = 36.416 - 0.0133 * 1980 = 10.02$
- Tempo di vittoria reale $t_{1980} = 10.25$



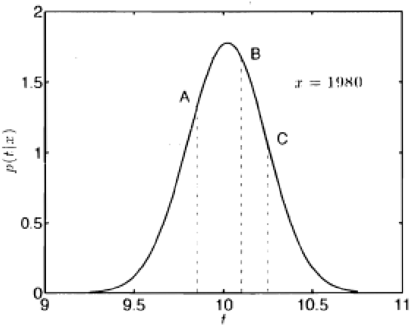
Trovare il $\hat{\mathbf{w}}$ ottimo

$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$



Trovare il $\hat{\mathbf{w}}$ ottimo

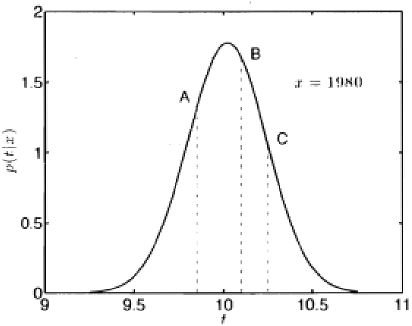
$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$



- Massimizzare la *likelihood* è uno dei concetti chiave del Machine Learning!

Trovare il $\hat{\mathbf{w}}$ ottimo

$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

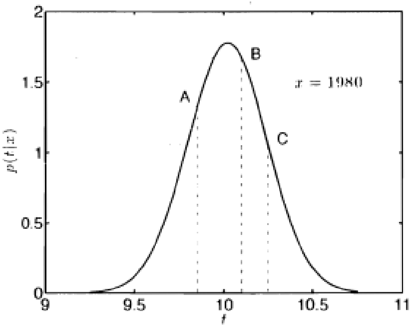


- Massimizzare la *likelihood* è uno dei concetti chiave del Machine Learning!
- Sfruttando le proprietà della Gaussiana e l'indipendenza degli shift, si dimostra analiticamente che il vettore $\hat{\mathbf{w}}$ ottimo è:

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{t}$$

Trovare il $\hat{\mathbf{w}}$ ottimo

$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$



- Massimizzare la *likelihood* è uno dei concetti chiave del Machine Learning!
- Sfruttando le proprietà della Gaussiana e l'indipendenza degli shift, si dimostra analiticamente che il vettore $\hat{\mathbf{w}}$ ottimo è:

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{t}$$

$\Rightarrow$ lo stesso ottenuto nelle lezioni precedenti!

Quanto è valido il nostro modello?

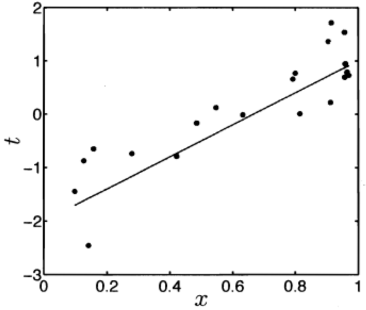
- Prendiamo un vettore $\mathbf{w} = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$ e 20 elementi $(x_1, \ldots, x_{20})$ uniformemente distribuiti tra 0 e 1

Quanto è valido il nostro modello?

- Prendiamo un vettore $\mathbf{w} = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$ e 20 elementi $(x_1, \ldots, x_{20})$ uniformemente distribuiti tra 0 e 1
- Otteniamo i corrispettivi $t_n = w_0 + w_1 x_n + \epsilon_n$, con $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ e $\sigma^2 = 0.25$

Quanto è valido il nostro modello?

- Prendiamo un vettore $\mathbf{w} = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$ e 20 elementi $(x_1, \ldots, x_{20})$ uniformemente distribuiti tra 0 e 1
- Otteniamo i corrispettivi $t_n = w_0 + w_1 x_n + \epsilon_n$, con $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ e $\sigma^2 = 0.25$



Quanto è valido il nostro modello?

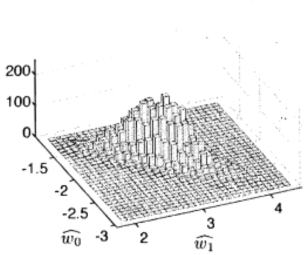
- Generiamo 10000 dataset secondo il modello descritto

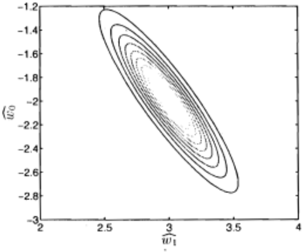
Quanto è valido il nostro modello?

- Generiamo 10000 dataset secondo il modello descritto
- Per ogni dataset facciamo regressione lineare per stimare $\hat{\mathbf{w}}$

Quanto è valido il nostro modello?

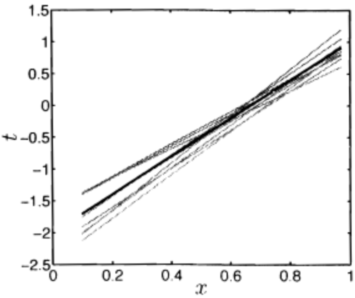
- Generiamo 10000 dataset secondo il modello descritto
- Per ogni dataset facciamo regressione lineare per stimare $\hat{\mathbf{w}}$





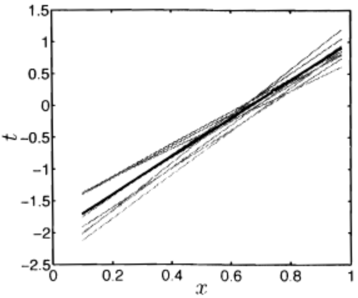
Quanto è valido il nostro modello?

- Generiamo 10000 dataset secondo il modello descritto
- Per ogni dataset facciamo regressione lineare per stimare $\hat{\mathbf{w}}$



Quanto è valido il nostro modello?

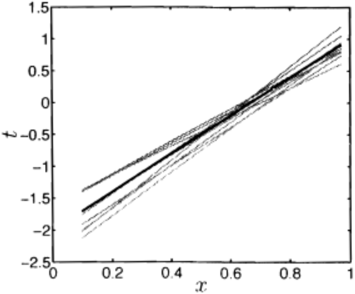
- Generiamo 10000 dataset secondo il modello descritto
- Per ogni dataset facciamo regressione lineare per stimare $\hat{\mathbf{w}}$



- Come è possibile vedere, non siamo troppo lontani dal vero modello. Né troppo in alto, né troppo in basso

Quanto è valido il nostro modello?

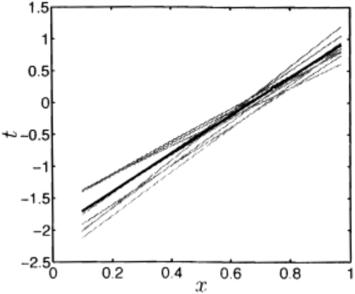
- Generiamo 10000 dataset secondo il modello descritto
- Per ogni dataset facciamo regressione lineare per stimare $\hat{\mathbf{w}}$



- Come è possibile vedere, non siamo troppo lontani dal vero modello. Né troppo in alto, né troppo in basso
- $\Rightarrow$ Si può dimostrare che il nostro regressore lineare è unbiased!

Quanto è valido il nostro modello?

- Generiamo 10000 dataset secondo il modello descritto
- Per ogni dataset facciamo regressione lineare per stimare $\hat{\mathbf{w}}$



- Come è possibile vedere, non siamo troppo lontani dal vero modello. Né troppo in alto, né troppo in basso
- $\Rightarrow$ Si può dimostrare che il nostro regressore lineare è unbiased!
- $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}$. Unbiased: ripetendo l'esperimento molte volte, la media dei risultati sarà sempre più vicina al valore vero

Quanto è valido il nostro modello?

- Di quanta variabilità soffre il nostro regressore lineare?

Quanto è valido il nostro modello?

- Di quanta variabilità soffre il nostro regressore lineare?
- Queste sono informazioni contenute nella matrice di covarianza:

$$\mathrm{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T \mathbf{X} \right)^{-1}$$

Quanto è valido il nostro modello?

- Di quanta variabilità soffre il nostro regressore lineare?
- Queste sono informazioni contenute nella matrice di covarianza:

$$\mathrm{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}$$

- La matrice di covarianza è una matrice quadrata con un numero di righe/colonne uguale alla dimensione di $\mathbf{w}$

Quanto è valido il nostro modello?

- Di quanta variabilità soffre il nostro regressore lineare?
- Queste sono informazioni contenute nella matrice di covarianza:

$$\text{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}$$

- La matrice di covarianza è una matrice quadrata con un numero di righe/colonne uguale alla dimensione di $\mathbf{w}$
- Gli elementi sulla diagonale indicano di quanta variabilità soffre la rispettiva componente di $\hat{\mathbf{w}}$

Quanto è valido il nostro modello?

- Di quanta variabilità soffre il nostro regressore lineare?
- Queste sono informazioni contenute nella matrice di covarianza:

$$\mathrm{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}$$

- La matrice di covarianza è una matrice quadrata con un numero di righe/colonne uguale alla dimensione di $\mathbf{w}$
- Gli elementi sulla diagonale indicano di quanta variabilità soffre la rispettiva componente di $\hat{\mathbf{w}}$
- Gli altri elementi della matrice danno informazioni di correlazione tra diverse componenti

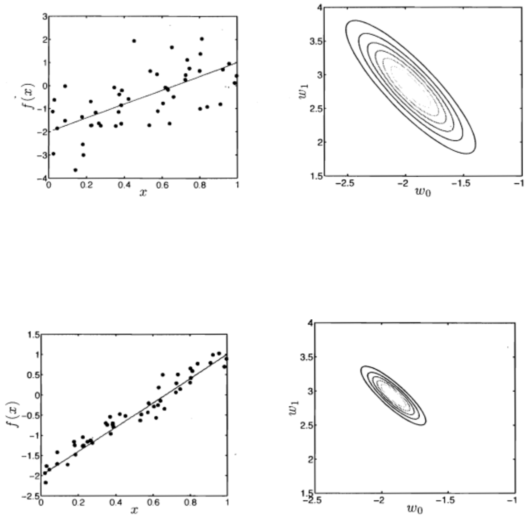
Quanto è valido il nostro modello?

- Di quanta variabilità soffre il nostro regressore lineare?
- Queste sono informazioni contenute nella matrice di covarianza:

$$\mathrm{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}$$

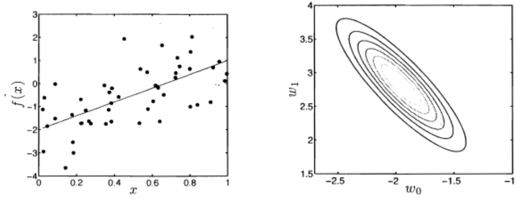
- La matrice di covarianza è una matrice quadrata con un numero di righe/colonne uguale alla dimensione di $\mathbf{w}$
- Gli elementi sulla diagonale indicano di quanta variabilità soffre la rispettiva componente di $\hat{\mathbf{w}}$
- Gli altri elementi della matrice danno informazioni di correlazione tra diverse componenti
  - Un valore vicino a 0 significa che due elementi sono indipendenti
  - Un valore positivo significa che se un valore aumenta, allora anche l'altro deve aumentare per non peggiorare la variabilità del modello
  - Un valore negativo significa che se un valore aumenta, allora l'altro deve diminuire per non peggiorare la variabilità del modello

Curvatura

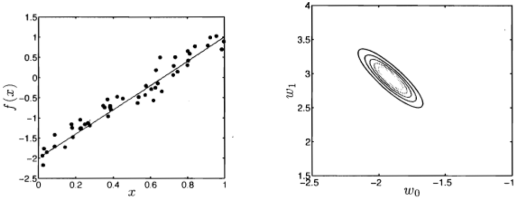


3
2
1
0
−1
−2
−3
−4
f(x)
0
0.2
0.4
0.6
0.8
1
x
4
3.5
3
2.5
2
1.5
w1
−2.5
−2
−1.5
−1
w0
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
f(x)
0
0.2
0.4
0.6
0.8
1
x
4
3.5
3
2.5
2
1.5
w1
−2.5
−2
−1.5
−1
w0

Curvatura

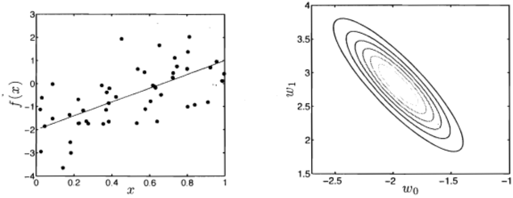


Alta variabilità→Debole curvatura



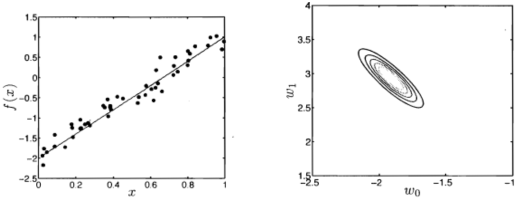
Bassa variabilità→Forte curvatura

Curvatura



Alta variabilità→Debole curvatura

$$\text{cov}[\hat{\mathbf{w}}] = \begin{bmatrix} 0.0784 & -0.12 \\ -0.12 & 0.2466 \end{bmatrix}$$



Bassa variabilità→Forte curvatura

$$\text{cov}[\hat{\mathbf{w}}] = \begin{bmatrix} 0.0031 & -0.0048 \\ -0.0048 & 0.0099 \end{bmatrix}$$

Variabilità σ

$$f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Variabilità σ

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$\mathrm{cov}[\hat{\mathbf{w}}] = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

Variabilità σ

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$\text{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

- Abbiamo visto che il parametro $\hat{\mathbf{w}}$ ottimo è $\hat{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{t}$

Variabilità σ

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$\mathrm{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

- Abbiamo visto che il parametro $\hat{\mathbf{w}}$ ottimo è $\hat{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{t}$
- Qual è la variabilità σ^2 ottima?

Variabilità σ

$$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$\text{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

- Abbiamo visto che il parametro $\hat{\mathbf{w}}$ ottimo è $\hat{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{t}$
- Qual è la variabilità σ^2 ottima?
- Massimizzando la likelihood si ottiene $\hat{\sigma} = \frac{1}{N}\left(\mathbf{t}^\mathbf{T}\mathbf{t} - \mathbf{t}^\mathbf{T}\mathbf{X}\hat{\mathbf{w}}\right)$

Variabilità σ

$$\mathrm{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

Variabilità σ

$$\text{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}$$

- Se riprendiamo i dati dal nostro esempio con 10000 dataset e $\sigma^2 = 0.25$

Variabilità σ

$$\text{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

- Se riprendiamo i dati dal nostro esempio con 10000 dataset e $\sigma^2 = 0.25$
- Si ottiene: $\text{cov}[\hat{\mathbf{w}}] = \begin{bmatrix} 0.0638 & -0.0821 \\ -0.0821 & 0.1317 \end{bmatrix}$

Variabilità σ

$$\text{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

- Se riprendiamo i dati dal nostro esempio con 10000 dataset e $\sigma^2 = 0.25$
- Si ottiene: $\text{cov}[\hat{\mathbf{w}}] = \begin{bmatrix} 0.0638 & -0.0821 \\ -0.0821 & 0.1317 \end{bmatrix}$
- Se invece calcoliamo $\hat{\sigma}^2$ a partire dai dataset otteniamo $\hat{\sigma}^2 = 0.2080$ e $\text{cov}[\hat{\mathbf{w}}] = \begin{bmatrix} 0.0530 & -0.0683 \\ -0.0683 & 0.1095 \end{bmatrix}$ (valori in generale più bassi)

Variabilità σ

$$\text{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

- Se riprendiamo i dati dal nostro esempio con 10000 dataset e $\sigma^2 = 0.25$
- Si ottiene: $\text{cov}[\hat{\mathbf{w}}] = \begin{bmatrix} 0.0638 & -0.0821 \\ -0.0821 & 0.1317 \end{bmatrix}$
- Se invece calcoliamo $\hat{\sigma}^2$ a partire dai dataset otteniamo $\hat{\sigma}^2 = 0.2080$ e $\text{cov}[\hat{\mathbf{w}}] = \begin{bmatrix} 0.0530 & -0.0683 \\ -0.0683 & 0.1095 \end{bmatrix}$ (valori in generale più bassi)
- Quando si ha a disposizione un grande dataset composto da S elementi, si può stimare empiricamente la matrice di covarianza:

$$\widehat{\text{cov}}[\hat{\mathbf{w}}] \triangleq \frac{1}{S}\sum_{s=1}^{S} (\hat{w}_s - \hat{\mu})(\hat{w}_s - \hat{\mu})^T$$

$$\hat{\mu} \triangleq \frac{1}{S}\sum_{s=1}^{S} \hat{\mathbf{w}}_s$$

Variabilità σ

$$\mathrm{cov}[\hat{\mathbf{w}}] = \sigma^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}$$

- Se riprendiamo i dati dal nostro esempio con 10000 dataset e $\sigma^2 = 0.25$
- Si ottiene: $\mathrm{cov}[\hat{\mathbf{w}}] = \begin{bmatrix} 0.0638 & -0.0821 \\ -0.0821 & 0.1317 \end{bmatrix}$
- Se invece calcoliamo $\hat{\sigma}^2$ a partire dai dataset otteniamo $\hat{\sigma}^2 = 0.2080$ e $\mathrm{cov}[\hat{\mathbf{w}}] = \begin{bmatrix} 0.0530 & -0.0683 \\ -0.0683 & 0.1095 \end{bmatrix}$ (valori in generale più bassi)
- Quando si ha a disposizione un grande dataset composto da S elementi, si può stimare empiricamente la matrice di covarianza:

$$\widehat{\mathrm{cov}}[\hat{\mathbf{w}}] \triangleq \frac{1}{S} \sum_{s=1}^{S} \left(\hat{w}_s - \hat{\mu}\right) \left(\hat{w}_s - \hat{\mu}\right)^T$$

$$\hat{\mu} \triangleq \frac{1}{S} \sum_{s=1}^{S} \hat{\mathbf{w}}_s$$

- In questo modo otteniamo $\widehat{\mathrm{cov}[\hat{\mathbf{w}}]} = \begin{bmatrix} 0.0627 & -0.0809 \\ -0.0809 & 0.1301 \end{bmatrix}$ (valori più simili al vero, ma serve S elevato)

Variabilità σ

- $\hat{\sigma}^2$ è *unbiased*?

Variabilità σ

- $\hat{\sigma}^2$ è *unbiased*?
- no!

Variabilità σ

- $\hat{\sigma}^2$ è *unbiased*?
- no!
- Infatti:
$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \left(1 - \frac{D}{N}\right) \neq \sigma$$

Variabilità σ

- $\hat{\sigma}^2$ è *unbiased*?
- no!
- Infatti:
$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \left(1 - \frac{D}{N}\right) \neq \sigma$$
- D è la dimensionalità del vettore di dati $\mathbf{x}_n$ e N è il numero di dati che abbiamo nel nostro dataset per fare Machine Learning

Variabilità σ

- $\hat{\sigma}^2$ è *unbiased*?
- no!
- Infatti:
$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \left(1 - \frac{D}{N}\right) \neq \sigma$$
- D è la dimensionalità del vettore di dati $\mathbf{x}_n$ e N è il numero di dati che abbiamo nel nostro dataset per fare Machine Learning
- La formula dimostra che la stima $\hat{\sigma}^2$ è sempre minore del valore vero σ^2

Variabilità σ

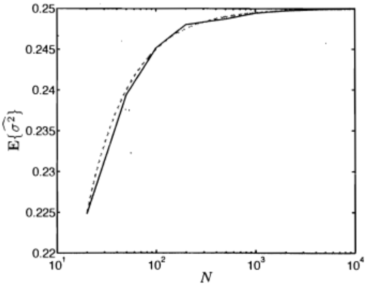
- $\hat{\sigma}^2$ è *unbiased*?
- no!
- Infatti:

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \left(1 - \frac{D}{N}\right) \neq \sigma$$

- D è la dimensionalità del vettore di dati $\mathbf{x}_n$ e N è il numero di dati che abbiamo nel nostro dataset per fare Machine Learning
- La formula dimostra che la stima $\hat{\sigma}^2$ è sempre minore del valore vero σ^2
- La formula dimostra anche che più N è grande -più dati abbiamo- più bassa sarà la variabilità nella stima (come è facile pensare anche intuitivamente)

Variabilità σ

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \left(1 - \frac{D}{N}\right) \neq \sigma$$



Predizioni

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

Predizioni

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

- Quanto è affidabile questa predizione?

Predizioni

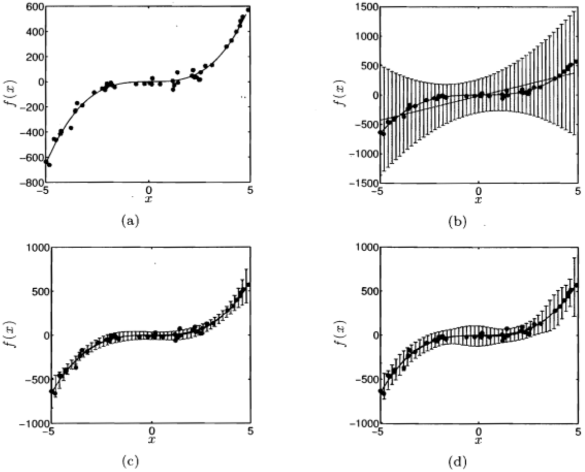
$$t_{\text{new}} = \hat{\mathbf{w}}^T \mathbf{x}_{\text{new}}$$

- Quanto è affidabile questa predizione?
- E'possibile valutarne analiticamente la variabilità:

$$\sigma^2_{\text{new}} = \sigma^2 \mathbf{x}^T_{\text{new}} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{x}_{\text{new}}$$

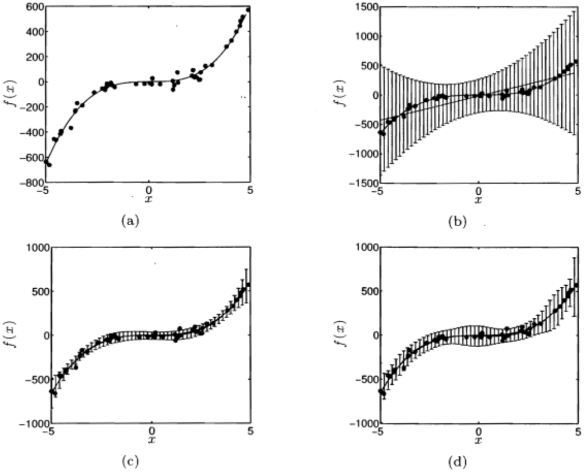
Predizioni

(a) $f(x) = 5x^3 - x^2 + x$
(b) lineare
(c) cubico (polinomio di grado 3) - vero modello
(d) polinomio di grado 6



Predizioni

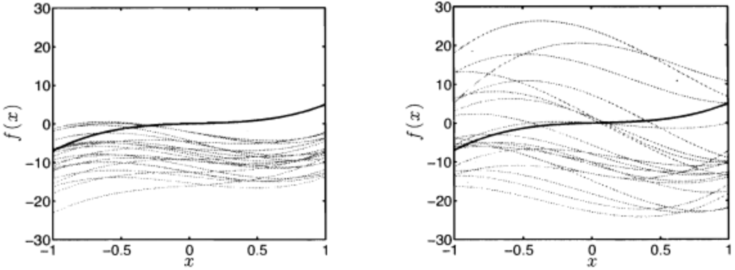
(a) $f(x) = 5x^3 - x^2 + x$
(b) lineare
(c) cubico (polinomio di grado 3) - vero modello
(d) polinomio di grado 6



Predizioni

$$f(x) = 5x^3 - x^2 + x$$

(a) variabilità polinomio di grado 3
(b) variabilità polinomio di grado 6



(a) (b)