Scuola universitaria professionale della Svizzera italiana
Bachelor di Ingegneria Informatica

SUPSI

Machine Learning

Lezione 2 - Regressione Lineare: approfondimento

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

IDSIA-SUPSI, Galleria 1, Manno

Capitolo 1 del libro
A First Course in Machine Learning – Rogers Girolami

Complichiamo un po' le cose

E se volessimo considerare più variabili? Es:

$$t = f(x, s_1, \ldots, s_8; w_0, \ldots, w_9) = w_0 + w_1 x + w_2 s_1 + \ldots + w_9 s_8$$

Complichiamo un po' le cose

E se volessimo considerare più variabili? Es:

$$t = f(x, s_1, \ldots, s_8; w_0, \ldots, w_9) = w_0 + w_1 x + w_2 s_1 + \ldots + w_9 s_8$$

Potremmo fare la stessa analisi vista nella lezione precedente per ottenere $\hat{w}_0, \ldots, \hat{w}_9$, ma la matematica diventerebbe troppo complicata

Complichiamo un po' le cose

E se volessimo considerare più variabili? Es:

$$t = f(x, s_1, \ldots, s_8; w_0, \ldots, w_9) = w_0 + w_1 x + w_2 s_1 + \ldots + w_9 s_8$$

Potremmo fare la stessa analisi vista nella lezione precedente per ottenere $\hat{w}_0, \ldots, \hat{w}_9$, ma la matematica diventerebbe troppo complicata

$\Rightarrow$ Notazione Matriciale

Regressione lineare: notazione matriciale

Ripetiamo gli step visti nella lezione scorsa, utilizzando la notazione matriciale

Regressione lineare: notazione matriciale

Ripetiamo gli step visti nella lezione scorsa, utilizzando la notazione matriciale

Definendo:

$$\mathbf{w} \triangleq \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \qquad \mathbf{x}_n \triangleq \begin{bmatrix} 1 \\ x_n \end{bmatrix}$$

Regressione lineare: notazione matriciale

Ripetiamo gli step visti nella lezione scorsa, utilizzando la notazione matriciale

Definendo:

$$\mathbf{w} \triangleq \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \qquad \mathbf{x}_n \triangleq \begin{bmatrix} 1 \\ x_n \end{bmatrix}$$

Possiamo scrivere:

$$f(x_n; w_0, w_1) = \mathbf{w}^T \mathbf{x}_n = w_0 + w_1 x_n$$

Regressione lineare: notazione matriciale

Ripetiamo gli step visti nella lezione scorsa, utilizzando la notazione matriciale

Definendo:

$$\mathbf{w} \triangleq \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \qquad \mathbf{x}_n \triangleq \begin{bmatrix} 1 \\ x_n \end{bmatrix}$$

Possiamo scrivere:

$$f(x_n; w_0, w_1) = \mathbf{w}^T \mathbf{x}_n = w_0 + w_1 x_n$$

e

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

Regressione lineare: notazione matriciale

Se, inoltre, definiamo:

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \qquad \mathbf{t} \triangleq \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$$

Regressione lineare: notazione matriciale

Se, inoltre, definiamo:

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \qquad \mathbf{t} \triangleq \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$$

Possiamo ottenere una notazione ancora più compatta e comoda:

$$\begin{aligned} \mathcal{L} &= \frac{1}{N} (\mathbf{t} - \mathbf{X}\mathbf{w})^T (\mathbf{t} - \mathbf{X}\mathbf{w}) \\ &= \|\mathbf{t} - \mathbf{X}\mathbf{w}\|_2^2 \end{aligned}$$

Regressione lineare: notazione matriciale

Se, inoltre, definiamo:

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \qquad \mathbf{t} \triangleq \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix}$$

Possiamo ottenere una notazione ancora più compatta e comoda:

$$\begin{aligned} \mathcal{L} &= \frac{1}{N} (\mathbf{t} - \mathbf{X}\mathbf{w})^T (\mathbf{t} - \mathbf{X}\mathbf{w}) \\ &= \|\mathbf{t} - \mathbf{X}\mathbf{w}\|_2^2 \end{aligned}$$

Calcoliamo ora il parametro ottimale $\hat{\mathbf{w}}$ sfruttando questa nuova notazione

Regressione lineare: notazione matriciale

$f(\mathbf{w})$	$\frac{\partial f}{\partial \mathbf{w}}$
$\mathbf{w}^T\mathbf{x}$	$\mathbf{x}$
$\mathbf{x}^T\mathbf{w}$	$\mathbf{x}$
$\mathbf{w}^T\mathbf{w}$	$2\mathbf{w}$
$\mathbf{w}^T\mathbf{C}\mathbf{w}$	$2\mathbf{C}\mathbf{w}$

Regressione lineare: notazione matriciale

$f(\mathbf{w})$	$\frac{\partial f}{\partial \mathbf{w}}$
$\mathbf{w}^{\mathsf{T}}\mathbf{x}$	$\mathbf{x}$
$\mathbf{x}^{\mathsf{T}}\mathbf{w}$	$\mathbf{x}$
$\mathbf{w}^{\mathsf{T}}\mathbf{w}$	$2\mathbf{w}$
$\mathbf{w}^{\mathsf{T}}\mathbf{C}\mathbf{w}$	$2\mathbf{C}\mathbf{w}$

Grazie alla tabella, possiamo facilmente ottenere la derivata di $\mathcal{L}$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{2}{N}\mathbf{X}^T\mathbf{X}\mathbf{w} - \frac{2}{N}\mathbf{X}^T\mathbf{t} = 0$$

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{t}$$

Regressione lineare: notazione matriciale

Matrice Inversa: $\mathbf{A}^{-1}$ è *inversa* di $\mathbf{A}$ se e solo se: $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$

Regressione lineare: notazione matriciale

Matrice Inversa: $\mathbf{A}^{-1}$ è *inversa* di $\mathbf{A}$ se e solo se: $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{t}$$
$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$
$$\mathbf{I}\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

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

Regressione lineare: notazione matriciale

Matrice Inversa: $\mathbf{A}^{-1}$ è *inversa* di $\mathbf{A}$ se e solo se: $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{t}$$
$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$
$$\mathbf{I}\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

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

Se ci viene chiesto di fare una predizione su nuovi dati $\mathbf{x}_{\text{new}}$, basta calcolare:

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

Notazione matriciale: esempio benchmark

n	x_n	t_n
1	1	4.8
2	3	11.3
3	5	17.2

Notazione matriciale: esempio benchmark

n	x_n	t_n
1	1	4.8
2	3	11.3
3	5	17.2

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{bmatrix}, \qquad \mathbf{t} = \begin{bmatrix} 4.8 \\ 11.3 \\ 17.2 \end{bmatrix}$$

Notazione matriciale: esempio benchmark

n	x_n	t_n
1	1	4.8
2	3	11.3
3	5	17.2

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{bmatrix}, \qquad \mathbf{t} = \begin{bmatrix} 4.8 \\ 11.3 \\ 17.2 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 3 & 9 \\ 9 & 35 \end{bmatrix}$$

Notazione matriciale: esempio benchmark

n	x_n	t_n
1	1	4.8
2	3	11.3
3	5	17.2

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{bmatrix}, \qquad \mathbf{t} = \begin{bmatrix} 4.8 \\ 11.3 \\ 17.2 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 3 & 9 \\ 9 & 35 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{24} \begin{bmatrix} 35 & -9 \\ -9 & 3 \end{bmatrix}$$

Notazione matriciale: esempio benchmark

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{24}\begin{bmatrix} 35 & -9 \\ -9 & 3 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{bmatrix} = \frac{1}{24}\begin{bmatrix} 26 & 8 & -10 \\ -6 & 0 & 6 \end{bmatrix}$$

Notazione matriciale: esempio benchmark

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{24}\begin{bmatrix}35 & -9\\ -9 & 3\end{bmatrix} \times \begin{bmatrix}1 & 1 & 1\\ 1 & 3 & 5\end{bmatrix} = \frac{1}{24}\begin{bmatrix}26 & 8 & -10\\ -6 & 0 & 6\end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} = \frac{1}{24}\begin{bmatrix}26 & 8 & -10\\ -6 & 0 & 6\end{bmatrix} \times \begin{bmatrix}4.8\\ 11.3\\ 17.2\end{bmatrix} = \begin{bmatrix}1.8\\ 3.1\end{bmatrix}$$

Notazione matriciale: esempio benchmark

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{24}\begin{bmatrix}35 & -9\\ -9 & 3\end{bmatrix} \times \begin{bmatrix}1 & 1 & 1\\ 1 & 3 & 5\end{bmatrix} = \frac{1}{24}\begin{bmatrix}26 & 8 & -10\\ -6 & 0 & 6\end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} = \frac{1}{24}\begin{bmatrix}26 & 8 & -10\\ -6 & 0 & 6\end{bmatrix} \times \begin{bmatrix}4.8\\ 11.3\\ 17.2\end{bmatrix} = \begin{bmatrix}1.8\\ 3.1\end{bmatrix}$$

$$f(x; w_0, w_1) = 1.8 + 3.1x$$

Regressione lineare: matrice dei dati

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

Regressione lineare: matrice dei dati

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

La matrice dei dati $\mathbf{X}$ l'abbiamo definita come:

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \dots & x_{NK} \end{bmatrix}$$

Regressione lineare: matrice dei dati

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

La matrice dei dati $\mathbf{X}$ l'abbiamo definita come:

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \dots & x_{NK} \end{bmatrix}$$

- Se $N = K + 1$ la matrice è quadrata e il regressore tocca tutti i punti (interpolazione)

Regressione lineare: matrice dei dati

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

La matrice dei dati $\mathbf{X}$ l'abbiamo definita come:

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \dots & x_{NK} \end{bmatrix}$$

- Se $N = K + 1$ la matrice è quadrata e il regressore tocca tutti i punti (interpolazione)
- Se $N > K + 1$ abbiamo più dati che incognite e il regressore approssima i punti

Regressione lineare: matrice dei dati

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

La matrice dei dati $\mathbf{X}$ l'abbiamo definita come:

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \dots & x_{NK} \end{bmatrix}$$

- Se $N = K + 1$ la matrice è quadrata e il regressore tocca tutti i punti (interpolazione)
- Se $N > K + 1$ abbiamo più dati che incognite e il regressore approssima i punti
- $\mathbf{X}$ potrebbe essere non-invertibile o *mal condizionata* (cioè numericamente difficile da invertire)$\Rightarrow$ Singular Value Decomposition (SVD) e pseudoinversa

Regressione lineare: modelli di ordine superiore

Cosa possiamo fare se pensiamo che i nostri dati non abbiano un andamento lineare?

Regressione lineare: modelli di ordine superiore

Cosa possiamo fare se pensiamo che i nostri dati non abbiano un andamento lineare?

Manteniamo il nostro modello

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

lineare nei parametri $\mathbf{w}$ (così da non complicare l'analisi e i calcoli), ma nonlineare nei dati $\mathbf{x}$

Regressione lineare: modelli di ordine superiore

Se, ad esempio, pensiamo che i nostri dati abbiano un andamento quadratico, possiamo aggiungere questa assunzione nel seguente modo.

Regressione lineare: modelli di ordine superiore

Se, ad esempio, pensiamo che i nostri dati abbiano un andamento quadratico, possiamo aggiungere questa assunzione nel seguente modo.

Aggiungiamo ad ogni dato x_n che abbiamo dei termini in più che descrivono l'assunzione che noi facciamo su di essi. Ad esempio, nel caso quadratico:

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_n \end{bmatrix} \Rightarrow \mathbf{x}_n = \begin{bmatrix} 1 \\ x_n \\ x_n^2 \end{bmatrix}$$

Regressione lineare: modelli di ordine superiore

Se, ad esempio, pensiamo che i nostri dati abbiano un andamento quadratico, possiamo aggiungere questa assunzione nel seguente modo.

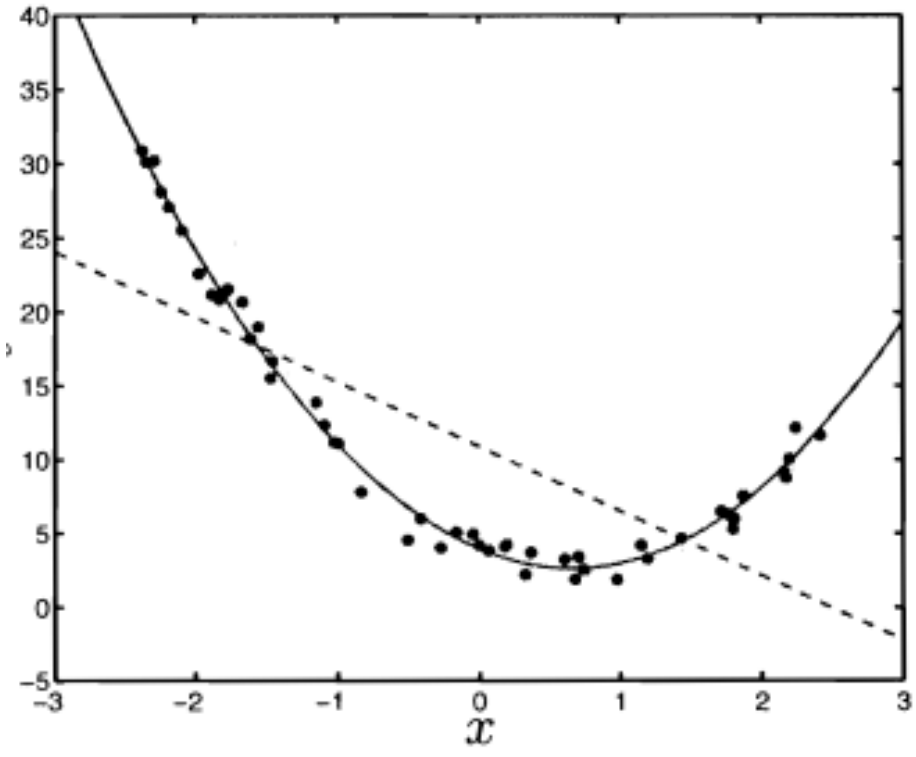
Aggiungiamo ad ogni dato x_n che abbiamo dei termini in più che descrivono l'assunzione che noi facciamo su di essi. Ad esempio, nel caso quadratico:

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_n \end{bmatrix} \Rightarrow \mathbf{x}_n = \begin{bmatrix} 1 \\ x_n \\ x_n^2 \end{bmatrix}$$

Di conseguenza, la matrice dei dati $\mathbf{X}$ diventa automaticamente:

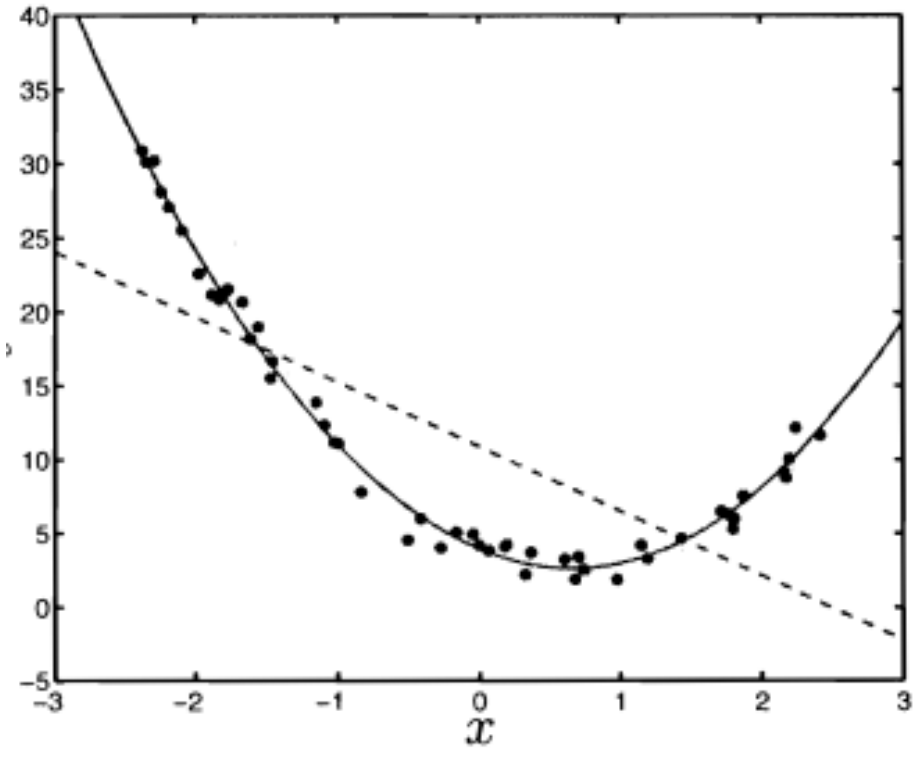
$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{bmatrix}$$

Regressione lineare: modelli di ordine superiore



Regressione lineare con dati quadratici e con dati lineari

Regressione lineare: modelli di ordine superiore



Regressione lineare con dati quadratici e con dati lineari

Regressione lineare: modelli di ordine superiore

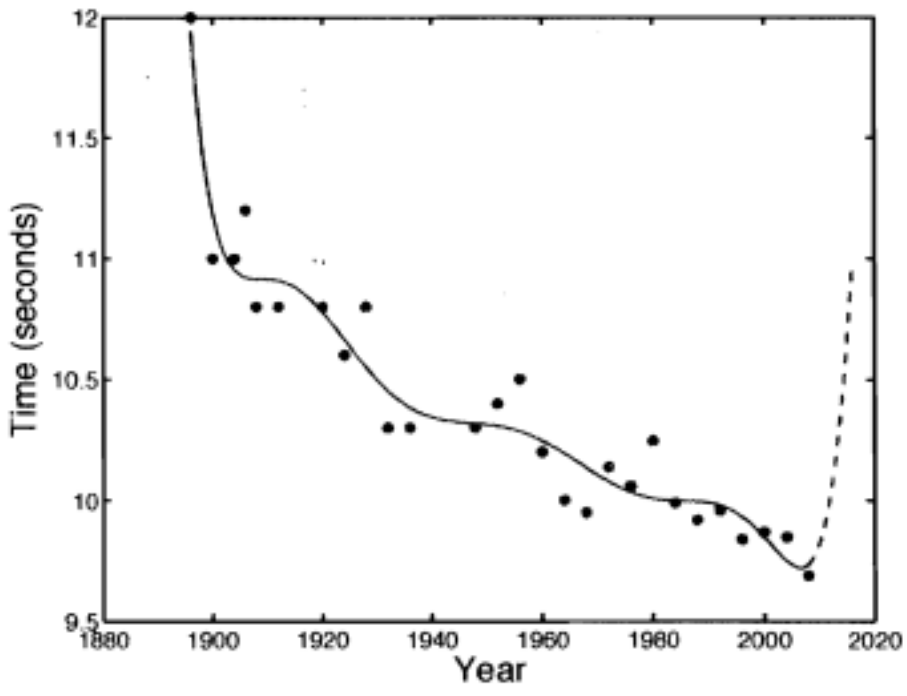
Ovviamente, non siamo limitati a trasformare i dati al secondo ordine, ma possiamo salire a polinomi ordini superiori:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^K \end{bmatrix}$$

Regressione lineare: modelli di ordine superiore

Ovviamente, non siamo limitati a trasformare i dati al secondo ordine, ma possiamo salire a polinomi ordini superiori:

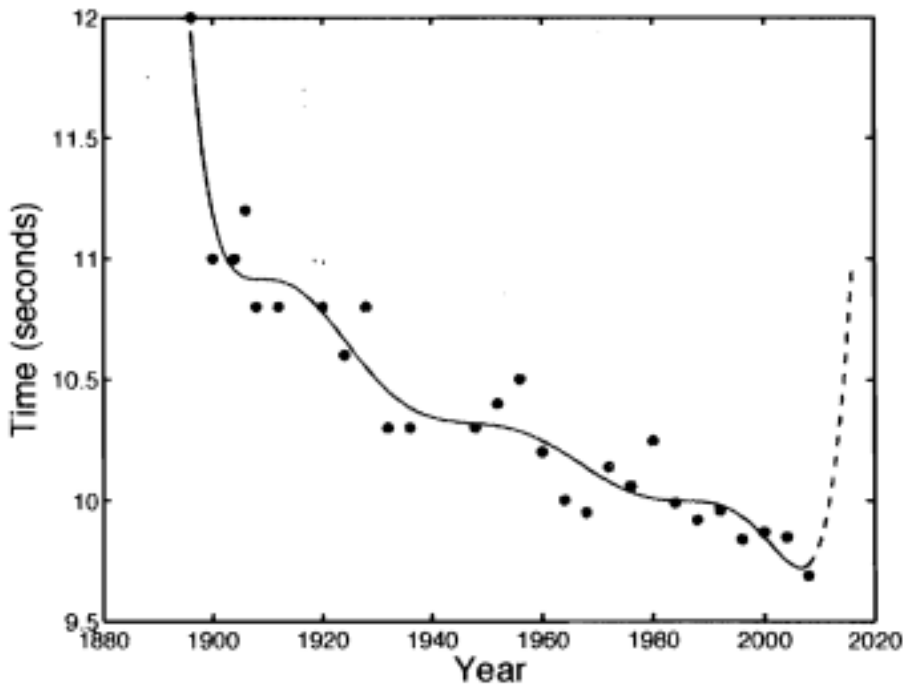
$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^K \end{bmatrix}$$



Regressione lineare: modelli di ordine superiore

Ovviamente, non siamo limitati a trasformare i dati al secondo ordine, ma possiamo salire a polinomi ordini superiori:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^K \end{bmatrix}$$



$$\mathcal{L}^1 = 1.358, \qquad \mathcal{L}^8 = 0.459$$

Regressione lineare: modelli di ordine superiore

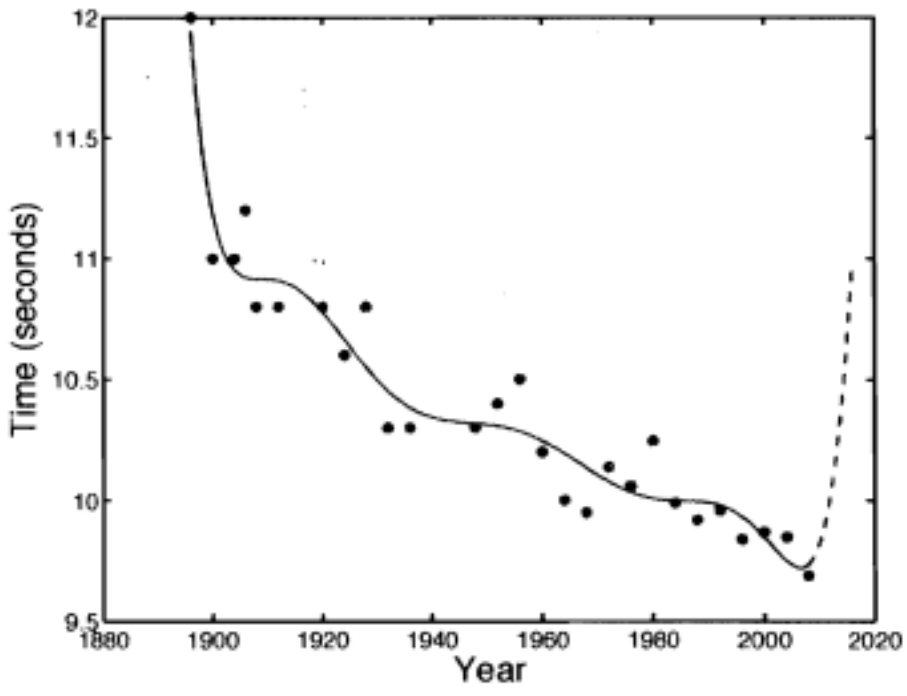
Ovviamente, non siamo limitati a trasformare i dati al secondo ordine, ma possiamo salire a polinomi ordini superiori:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^K \end{bmatrix}$$

Regressione lineare: modelli di ordine superiore

Ovviamente, non siamo limitati a trasformare i dati al secondo ordine, ma possiamo salire a polinomi ordini superiori:

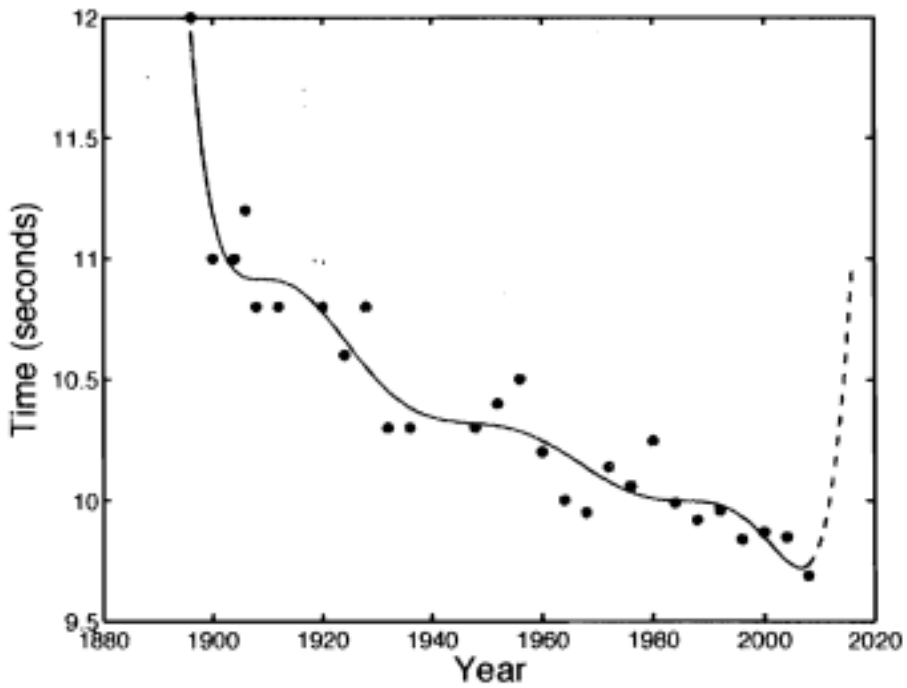
$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^K \end{bmatrix}$$



Regressione lineare: modelli di ordine superiore

Ovviamente, non siamo limitati a trasformare i dati al secondo ordine, ma possiamo salire a polinomi ordini superiori:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^K \end{bmatrix}$$



$$\mathcal{L}^1 = 1.358, \qquad \mathcal{L}^8 = 0.459$$

Regressione lineare: modelli di ordine superiore

Infine, così come abbiamo trasformato i dati utilizzando polinomi, possiamo applicare trasformazioni di altro tipo:

$$\mathbf{X} = \begin{bmatrix} 1 & h_1(x_1) & h_2(x_1) & \dots & h_K(x_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & h_1(x_N) & h_2(x_N) & \dots & h_K(x_N) \end{bmatrix}$$

Regressione lineare: modelli di ordine superiore

Infine, così come abbiamo trasformato i dati utilizzando polinomi, possiamo applicare trasformazioni di altro tipo:

$$\mathbf{X} = \begin{bmatrix} 1 & h_1(x_1) & h_2(x_1) & \dots & h_K(x_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & h_1(x_N) & h_2(x_N) & \dots & h_K(x_N) \end{bmatrix}$$

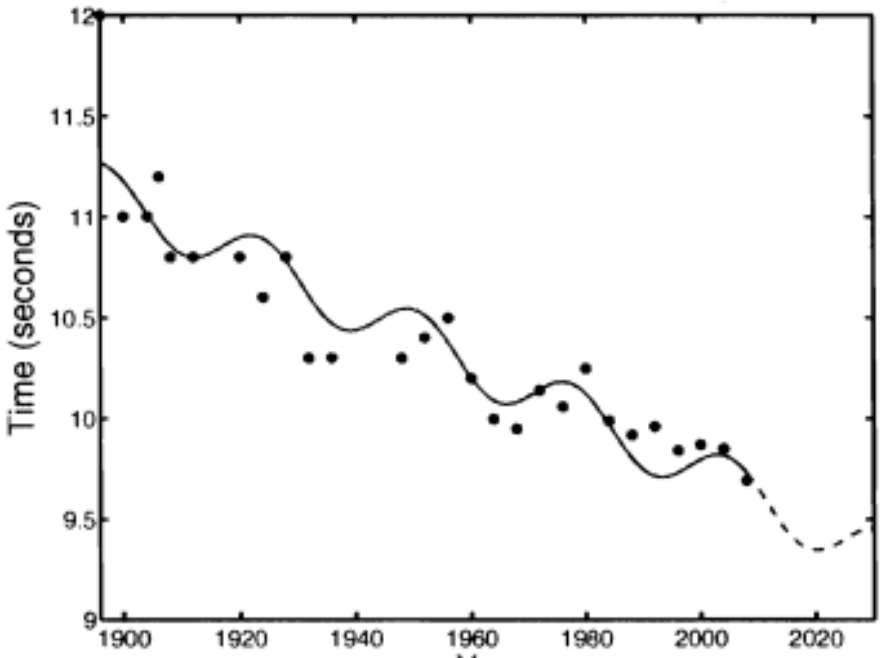
$$h_1(x) = x; h_2(x) = \sin\left(\frac{x-2660}{4.3}\right)$$
$$f(x; \mathbf{w}) = w_0 + w_1 x + w_2 \sin\left(\frac{x-2660}{4.3}\right)$$

Regressione lineare: modelli di ordine superiore

Infine, così come abbiamo trasformato i dati utilizzando polinomi, possiamo applicare trasformazioni di altro tipo:

$$\mathbf{X} = \begin{bmatrix} 1 & h_1(x_1) & h_2(x_1) & \dots & h_K(x_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & h_1(x_N) & h_2(x_N) & \dots & h_K(x_N) \end{bmatrix}$$

$$h_1(x) = x; h_2(x) = \sin\left(\tfrac{x-2660}{4.3}\right)$$
$$f(x; \mathbf{w}) = w_0 + w_1 x + w_2 \sin\left(\tfrac{x-2660}{4.3}\right)$$



- $w_0 = 36.610$
- $w_1 = -0.013$
- $w_2 = -0.133$
- $\mathcal{L} = 1.1037$

Regressione lineare: predizioni e over-fitting

Abbiamo costruito il regressore lineare per poter fare predizioni a partire dai dati a nostra disposizione

Regressione lineare: predizioni e over-fitting

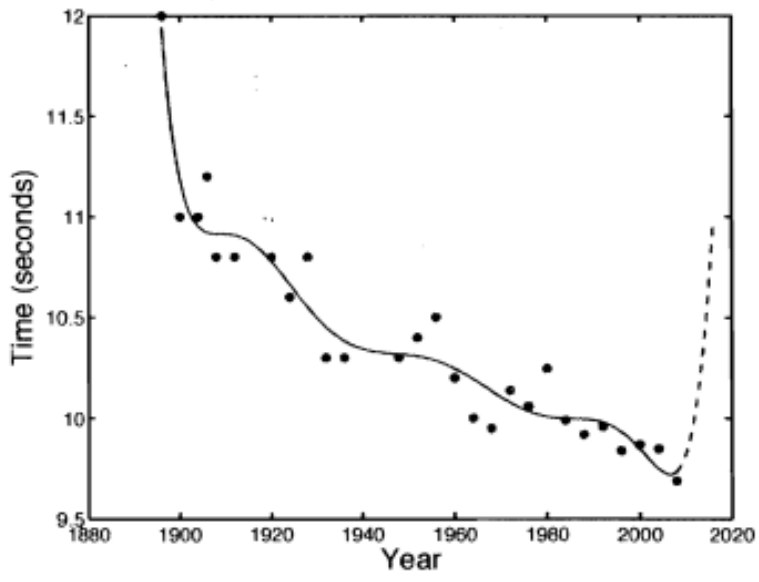
Abbiamo costruito il regressore lineare per poter fare predizioni a partire dai dati a nostra disposizione

Ora dobbiamo misurare la qualità del nostro regressore nel fare queste predizioni
Esempio:

Regressione lineare: predizioni e over-fitting

Abbiamo costruito il regressore lineare per poter fare predizioni a partire dai dati a nostra disposizione

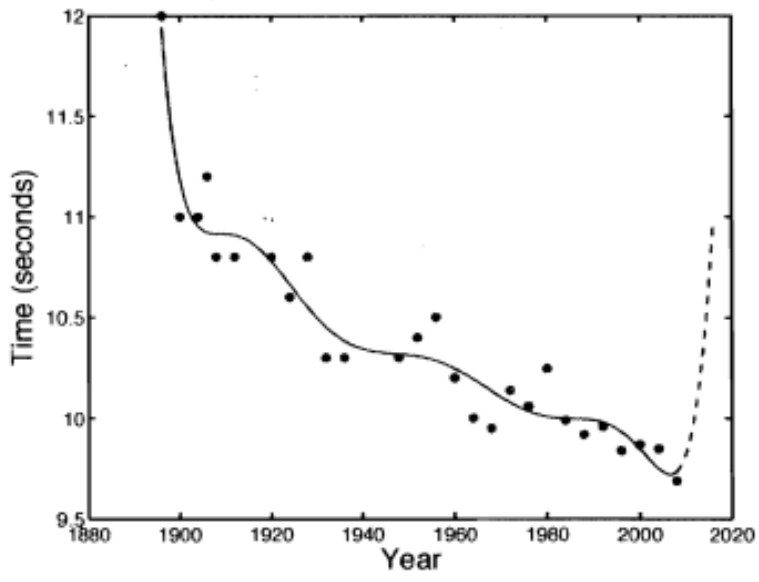
Ora dobbiamo misurare la qualità del nostro regressore nel fare queste predizioni
Esempio:



Regressione lineare: predizioni e over-fitting

Abbiamo costruito il regressore lineare per poter fare predizioni a partire dai dati a nostra disposizione

Ora dobbiamo misurare la qualità del nostro regressore nel fare queste predizioni
Esempio:

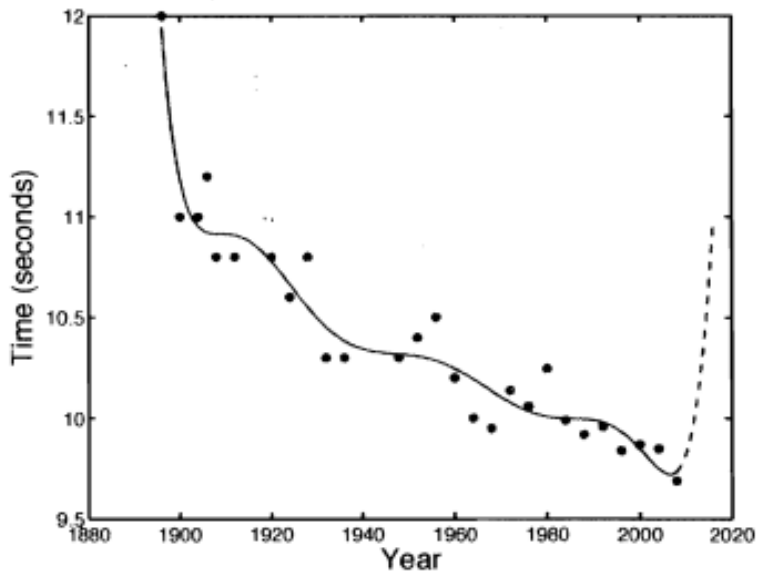


Il polinomio di grado 8 ottiene un valore $\mathcal{L}$ più basso del polinomio di grado 1 nel descrivere i dati di partenza

Regressione lineare: predizioni e over-fitting

Abbiamo costruito il regressore lineare per poter fare predizioni a partire dai dati a nostra disposizione

Ora dobbiamo misurare la qualità del nostro regressore nel fare queste predizioni
Esempio:



Il polinomio di grado 8 ottiene un valore $\mathcal{L}$ più basso del polinomio di grado 1 nel descrivere i dati di partenza

Tuttavia sembra non essere in grado di generalizzare bene rispetto ai dati che non conosce

Regressione lineare: predizioni e over-fitting

Training Set: l'insieme dei dati noti/iniziali. I quali vengono usati per calcolare i parametri del regressore

Regressione lineare: predizioni e over-fitting

Training Set: l'insieme dei dati noti/iniziali. I quali vengono usati per calcolare i parametri del regressore

Validation Set: l'insieme di dati su cui si testa il regressore, per analizzare quanto sono corrette le sue predizioni

Regressione lineare: predizioni e over-fitting

Training Set: l'insieme dei dati noti/iniziali. I quali vengono usati per calcolare i parametri del regressore

Validation Set: l'insieme di dati su cui si testa il regressore, per analizzare quanto sono corrette le sue predizioni

Over-fitting: si parla di over-fitting quando un regressore finisce per fare cattive predizioni, nel tentativo di funzionare al meglio possibile sul training Set

Regressione lineare: predizioni e over-fitting

Nota Bene: In generale più un regressore è di grado elevato, più sarà soggetto ad over-fitting

Regressione lineare: predizioni e over-fitting

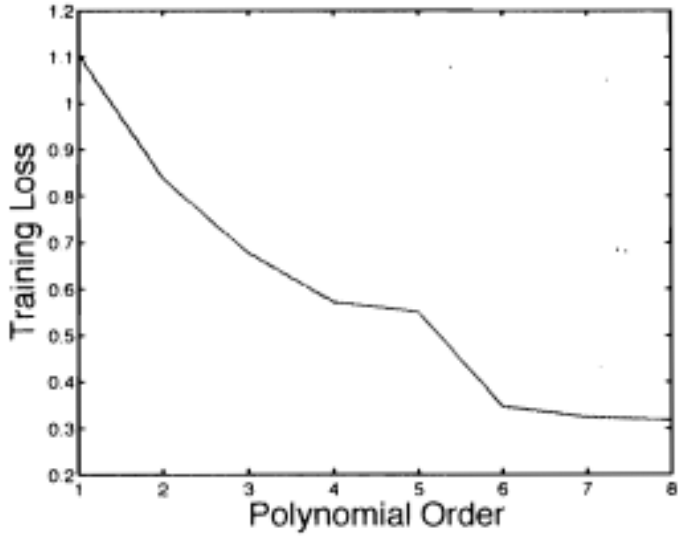
Nota Bene: In generale più un regressore è di grado elevato, più sarà soggetto ad over-fitting

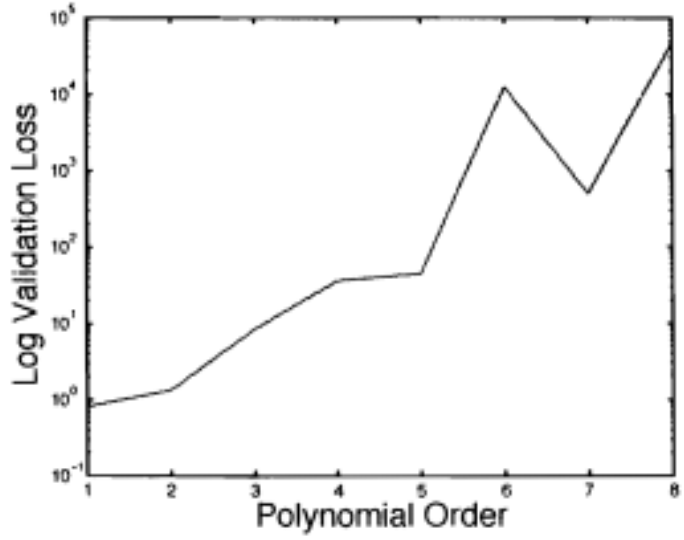
$\Rightarrow$Trade-off tra minimizzare l'errore sul training set e over-fitting

Regressione lineare: predizioni e over-fitting

Nota Bene: In generale più un regressore è di grado elevato, più sarà soggetto ad over-fitting

⇒Trade-off tra minimizzare l'errore sul training set e over-fitting



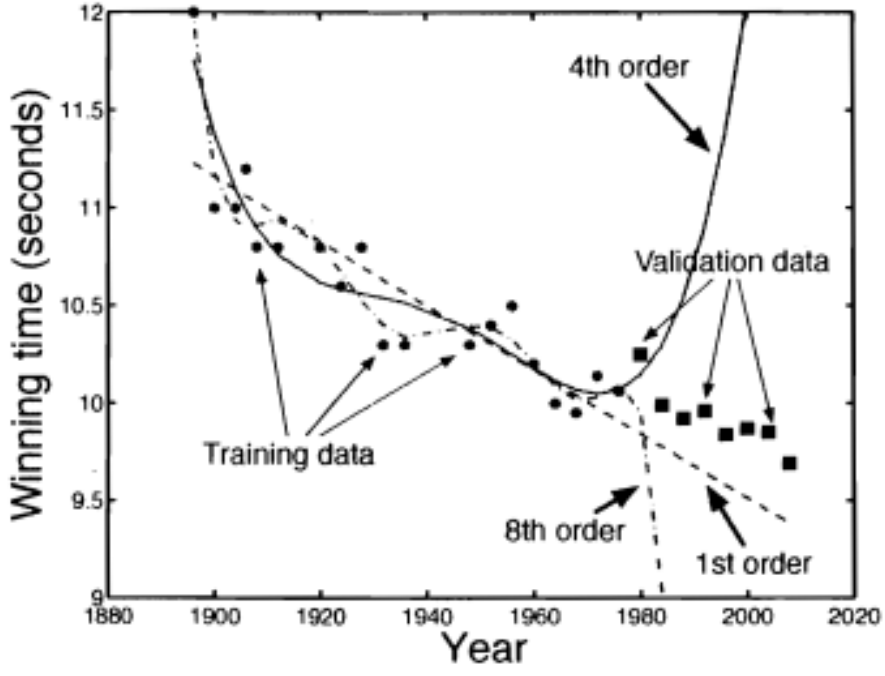


Il validation set è ottenuto rimuovendo dal training set i dati dal 1980 in poi

Regressione lineare: predizioni e over-fitting

Nota Bene: In generale più un regressore è di grado elevato, più sarà soggetto ad over-fitting

⇒Trade-off tra minimizzare l'errore sul training set e over-fitting

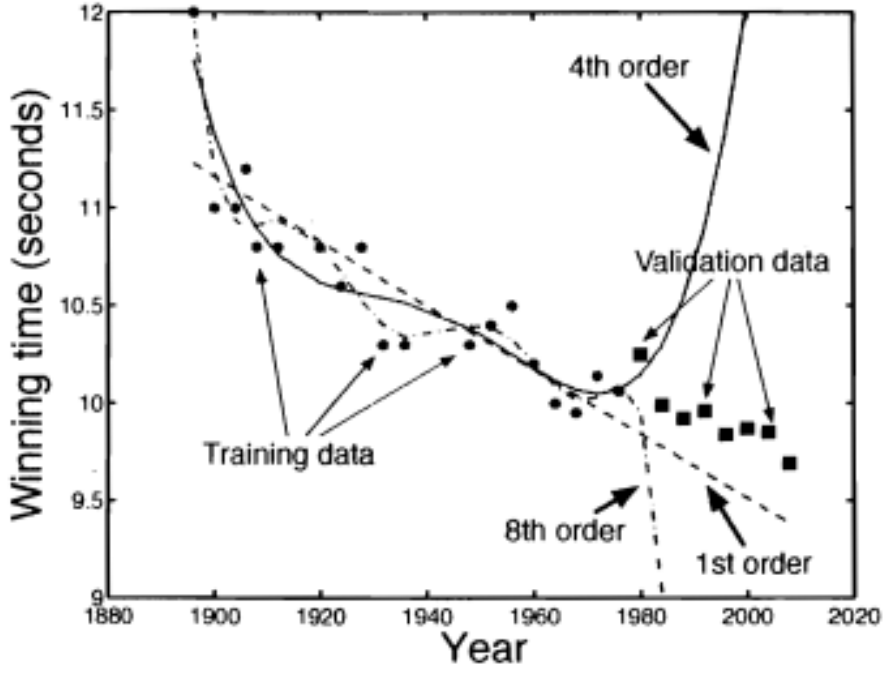


Il validation set è ottenuto rimuovendo dal training set i dati dal 1980 in poi

Regressione lineare: predizioni e over-fitting

Nota Bene: In generale più un regressore è di grado elevato, più sarà soggetto ad over-fitting

⇒Trade-off tra minimizzare l'errore sul training set e over-fitting



Il validation set è ottenuto rimuovendo dal training set i dati dal 1980 in poi

⇒ sembra che il modello migliore sia il regressore di grado 1

Cross-Validation

Spesso abbiamo pochi dati a disposizione: qual è il modo migliore di dividerli in training set e validation set?

Cross-Validation

Spesso abbiamo pochi dati a disposizione: qual è il modo migliore di dividerli in training set e validation set?

Cross-validation:

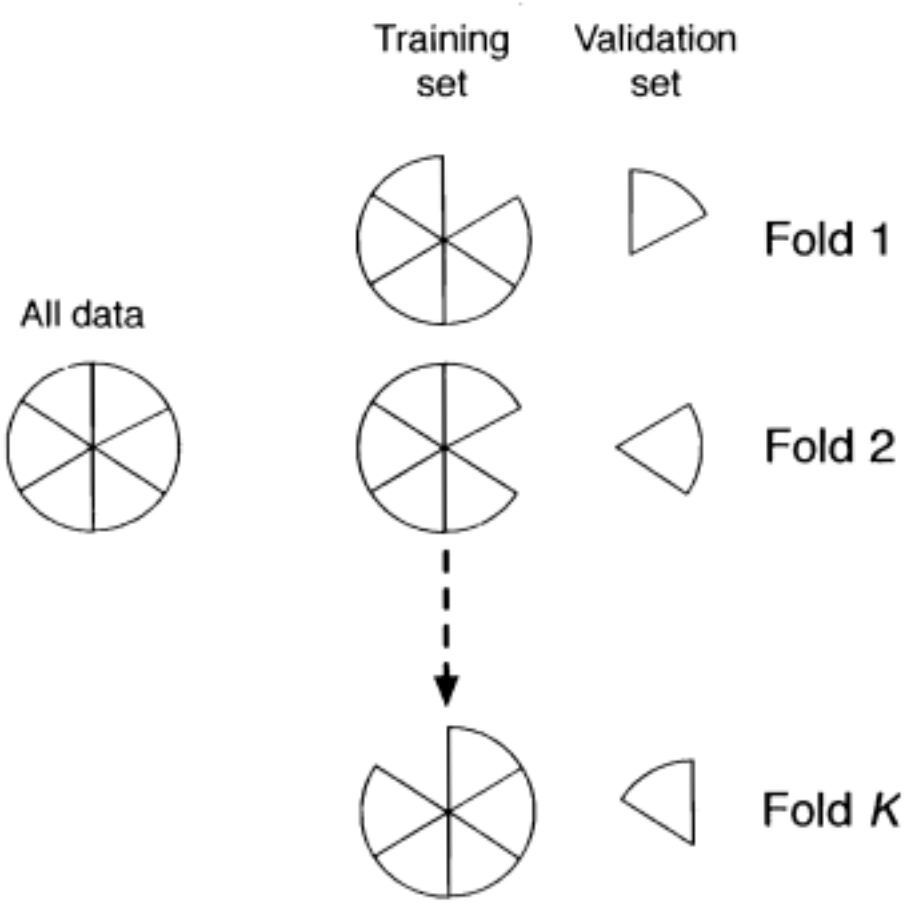
1. si scelgono casualmente alcuni dati che formano il training set ed altri che formano il validation set
2. si costruisce il regressore sul training set selezionato e lo si testa sul validation set, calcolando $\mathcal{L}$
3. si itera il procedimento in 1) e 2), scegliendo dati diversi per il validation set
4. dopo aver svolto un numero fissato di iterazioni, ogni volta con validation set diversi, si calcola la media delle $\mathcal{L}$ ottenute
5. questa media è il valore che indica la qualità del nostro regressore (più è bassa e meglio si sta comportando il regressore scelto)

Cross-Validation

$K-$fold Cross-Validation: i dati vengono divisi in K sottoinsieme di uguale dimensione. Poi ogni sottoinsieme viene usato come validation set per un regressore costruito utilizzando gli altri $K-1$ sottoinsiemi come training set

Cross-Validation

$K-$fold Cross-Validation: i dati vengono divisi in K sottoinsieme di uguale dimensione. Poi ogni sottoinsieme viene usato come validation set per un regressore costruito utilizzando gli altri $K-1$ sottoinsiemi come training set

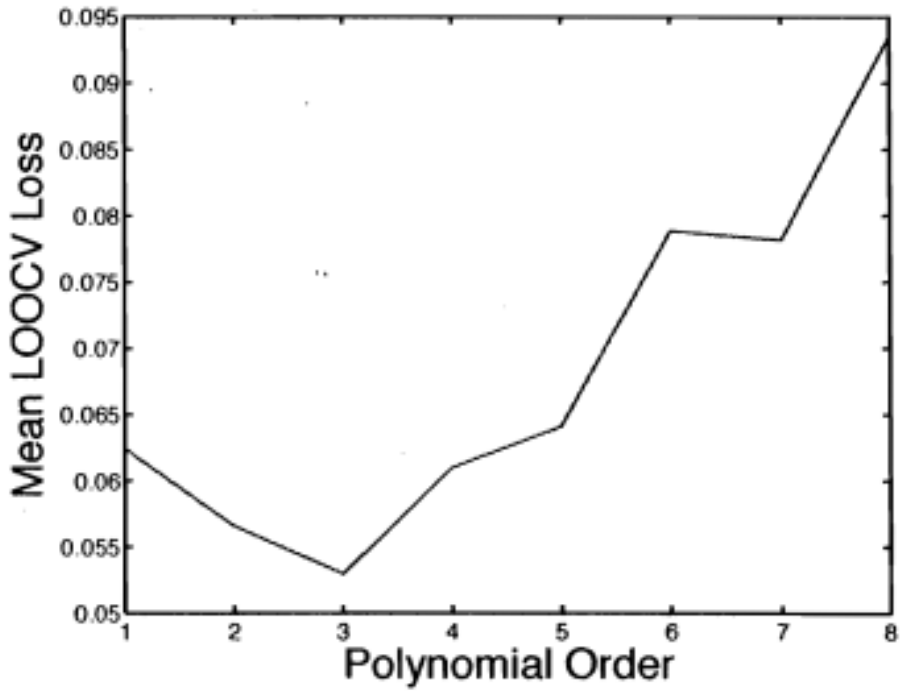


Cross-Validation

Leave-One-Out Cross-Validation (LOOCV): tecnica di K-fold Cross-Validation secondo la quale, avendo a disposizione N dati, si usa $K = N$.

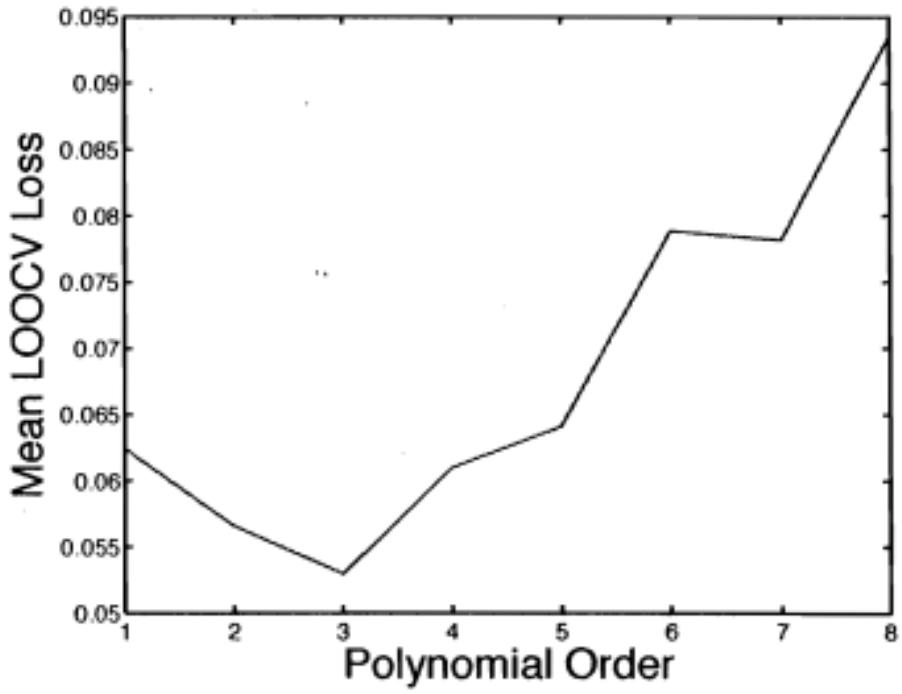
Cross-Validation

Leave-One-Out Cross-Validation (LOOCV): tecnica di K-fold Cross-Validation secondo la quale, avendo a disposizione N dati, si usa $K = N$.



Cross-Validation

Leave-One-Out Cross-Validation (LOOCV): tecnica di K-fold Cross-Validation secondo la quale, avendo a disposizione N dati, si usa $K = N$.



$\Rightarrow$ sembra che il modello migliore sia il regressore di grado 3

Cross-Validation

·Esempio: generiamo artificialmente 50 dati provenienti da un polinomio di terzo grado, poi li corrompiamo con rumore. Questo sarà il nostro dataset

Cross-Validation

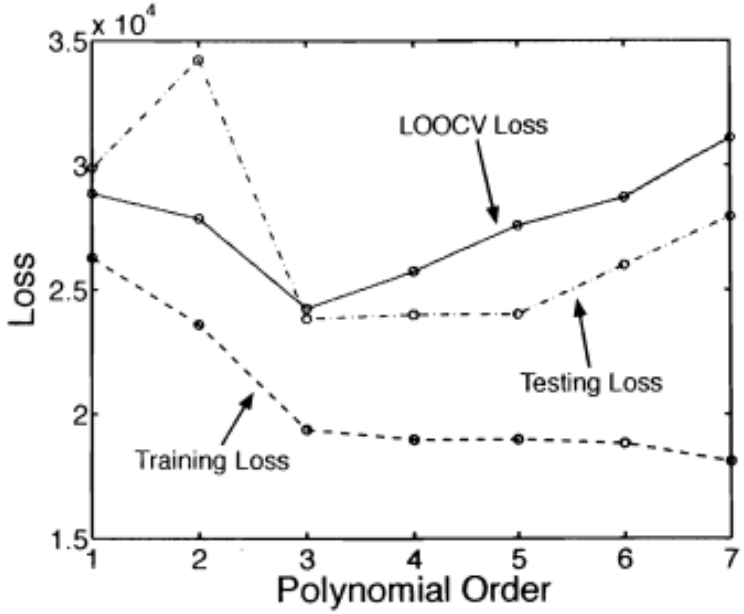
·Esempio: generiamo artificialmente 50 dati provenienti da un polinomio di terzo grado, poi li corrompiamo con rumore. Questo sarà il nostro dataset
·generiamo poi altri 1000 dati, senza corromperli con rumore

Cross-Validation

·Esempio: generiamo artificialmente 50 dati provenienti da un polinomio di terzo grado, poi li corrompiamo con rumore. Questo sarà il nostro dataset
·generiamo poi altri 1000 dati, senza corromperli con rumore
·Costruiamo dei regressori e vediamo secondo il metodo LOOCV e secondo il metodo che sfrutta i 1000 dati in più quale è il grado ottimale per un regressore lineare

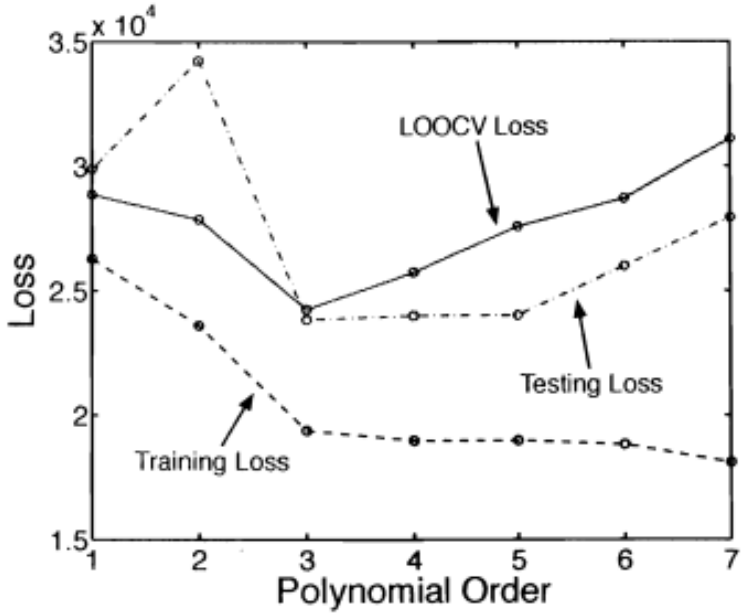
Cross-Validation

·Esempio: generiamo artificialmente 50 dati provenienti da un polinomio di terzo grado, poi li corrompiamo con rumore. Questo sarà il nostro dataset
·generiamo poi altri 1000 dati, senza corromperli con rumore
·Costruiamo dei regressori e vediamo secondo il metodo LOOCV e secondo il metodo che sfrutta i 1000 dati in più quale è il grado ottimale per un regressore lineare



Cross-Validation

·Esempio: generiamo artificialmente 50 dati provenienti da un polinomio di terzo grado, poi li corrompiamo con rumore. Questo sarà il nostro dataset
·generiamo poi altri 1000 dati, senza corromperli con rumore
·Costruiamo dei regressori e vediamo secondo il metodo LOOCV e secondo il metodo che sfrutta i 1000 dati in più quale è il grado ottimale per un regressore lineare



Questo semplice esperimento mostra come l'approccio LOOCV produca un risultato corretto⇒ Buono a sapersi! Dato che difficilmente per esperimenti reali avremo a disposizione un numero elevati di dati (1000, ad esempio) come validation set

Regolarizzazione

- Abbiamo visto che il nostro modello può essere scritto:

Regolarizzazione

- Abbiamo visto che il nostro modello può essere scritto:

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

Regolarizzazione

- Abbiamo visto che il nostro modello può essere scritto:

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

dove:

Regolarizzazione

- Abbiamo visto che il nostro modello può essere scritto:

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

dove:

- $\mathbf{x}$ è il vettore dei dati, ai quali possiamo aver applicato trasformazioni nonlineari per ottenere predizioni migliori
- $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{bmatrix}$ è il vettore dei parametri

Regolarizzazione

- Abbiamo visto che il nostro modello può essere scritto:

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

dove:

  - $\mathbf{x}$ è il vettore dei dati, ai quali possiamo aver applicato trasformazioni nonlineari per ottenere predizioni migliori
  - $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{bmatrix}$ è il vettore dei parametri

- Se, per assurdo, consideriamo $\mathbf{w} = \mathbf{0}$ il nostro modello $f(\mathbf{x}; \mathbf{w})$ darebbe sempre come predizione il risultato 0

Regolarizzazione

- Abbiamo visto che il nostro modello può essere scritto:

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

dove:

  - $\mathbf{x}$ è il vettore dei dati, ai quali possiamo aver applicato trasformazioni nonlineari per ottenere predizioni migliori
  - $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{bmatrix}$ è il vettore dei parametri

- Se, per assurdo, consideriamo $\mathbf{w} = \mathbf{0}$ il nostro modello $f(\mathbf{x}; \mathbf{w})$ darebbe sempre come predizione il risultato 0

- Questo significa che più il valore assoluto degli elementi di $\mathbf{w}$ è elevato, più il nostro regressore è complesso

Regolarizzazione

- Per controllare la complessità del nostro modello bisogna quindi impedire che il valore assoluto degli elementi di $\mathbf{w}$ diventi troppo elevato

Regolarizzazione

- Per controllare la complessità del nostro modello bisogna quindi impedire che il valore assoluto degli elementi di $\mathbf{w}$ diventi troppo elevato

- Un possibile approccio, quindi, è quello inserire il termine $\sum_{i=0}^{K} w_i = \mathbf{w}^T \mathbf{w}$ nella funzione costo $\mathcal{L}$ come regolarizzatore (N.B.: avremmo potuto scegliere $\sum_{i=0}^{K} |w_i|$ come regolarizzatore, ma avrebbe reso l'analisi matematica più complicata)

Regolarizzazione

- Per controllare la complessità del nostro modello bisogna quindi impedire che il valore assoluto degli elementi di $\mathbf{w}$ diventi troppo elevato

- Un possibile approccio, quindi, è quello inserire il termine $\sum_{i=0}^{K} w_i = \mathbf{w}^T\mathbf{w}$ nella funzione costo $\mathcal{L}$ come regolarizzatore (N.B.: avremmo potuto scegliere $\sum_{i=0}^{K} |w_i|$ come regolarizzatore, ma avrebbe reso l'analisi matematica più complicata)

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda\mathbf{w}^T\mathbf{w}$$

Regolarizzazione

- Per controllare la complessità del nostro modello bisogna quindi impedire che il valore assoluto degli elementi di $\mathbf{w}$ diventi troppo elevato

- Un possibile approccio, quindi, è quello inserire il termine $\sum_{i=0}^{K} w_i = \mathbf{w}^T\mathbf{w}$ nella funzione costo $\mathcal{L}$ come regolarizzatore (N.B.: avremmo potuto scegliere $\sum_{i=0}^{K} |w_i|$ come regolarizzatore, ma avrebbe reso l'analisi matematica più complicata)

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda \mathbf{w}^T\mathbf{w}$$

- $\lambda > 0$ è un parametro che possiamo modificare a seconda di quanto importanza vogliamo dare alla regolarizzazione

Regolarizzazione

- Per controllare la complessità del nostro modello bisogna quindi impedire che il valore assoluto degli elementi di $\mathbf{w}$ diventi troppo elevato

- Un possibile approccio, quindi, è quello inserire il termine $\sum_{i=0}^{K} w_i = \mathbf{w}^T\mathbf{w}$ nella funzione costo $\mathcal{L}$ come regolarizzatore (N.B.: avremmo potuto scegliere $\sum_{i=0}^{K} |w_i|$ come regolarizzatore, ma avrebbe reso l'analisi matematica più complicata)

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda \mathbf{w}^T\mathbf{w}$$

- $\lambda > 0$ è un parametro che possiamo modificare a seconda di quanto importanza vogliamo dare alla regolarizzazione

- Più scegliamo λ elevato, più importanza stiamo dando alla regolarizzazione e più richiediamo al nostro regressore di essere semplice

Regolarizzazione

La nuova funzione costo che vogliamo studiare è:

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w}$$

Regolarizzazione

La nuova funzione costo che vogliamo studiare è:

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w}$$

Per minimizzarla e trovare i parametri ottimali $\hat{\mathbf{w}}$ calcoliamo, come sempre la derivata:

Regolarizzazione

La nuova funzione costo che vogliamo studiare è:

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w}$$

Per minimizzarla e trovare i parametri ottimali $\hat{\mathbf{w}}$ calcoliamo, come sempre la derivata:

$$\frac{\partial \mathcal{L}'}{\partial \mathbf{w}} = \frac{2}{N} \mathbf{X}^T \mathbf{X} \mathbf{w} - \frac{2}{N} \mathbf{X}^T \mathbf{t} + 2\lambda \mathbf{w}$$

Regolarizzazione

La nuova funzione costo che vogliamo studiare è:

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w}$$

Per minimizzarla e trovare i parametri ottimali $\hat{\mathbf{w}}$ calcoliamo, come sempre la derivata:

$$\frac{\partial \mathcal{L}'}{\partial \mathbf{w}} = \frac{2}{N} \mathbf{X}^T \mathbf{X} \mathbf{w} - \frac{2}{N} \mathbf{X}^T \mathbf{t} + 2\lambda \mathbf{w}$$

Uguagliando la derivata a 0 otteniamo:

Regolarizzazione

La nuova funzione costo che vogliamo studiare è:

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w}$$

Per minimizzarla e trovare i parametri ottimali $\hat{\mathbf{w}}$ calcoliamo, come sempre la derivata:

$$\frac{\partial \mathcal{L}'}{\partial \mathbf{w}} = \frac{2}{N} \mathbf{X}^T \mathbf{X} \mathbf{w} - \frac{2}{N} \mathbf{X}^T \mathbf{t} + 2\lambda \mathbf{w}$$

Uguagliando la derivata a 0 otteniamo:

$$\frac{2}{N} \mathbf{X}^T \mathbf{X} \mathbf{w} - \frac{2}{N} \mathbf{X}^T \mathbf{t} + 2\lambda \mathbf{w} = 0$$

$$\left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right) \mathbf{w} = \mathbf{X}^T \mathbf{t}$$

$$\left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right)^{-1} \left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right) \mathbf{w} = \left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{t}$$

$$\mathbf{I} \mathbf{w} = \left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{t}$$

Regolarizzazione

La nuova funzione costo che vogliamo studiare è:

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w}$$

Per minimizzarla e trovare i parametri ottimali $\hat{\mathbf{w}}$ calcoliamo, come sempre la derivata:

$$\frac{\partial \mathcal{L}'}{\partial \mathbf{w}} = \frac{2}{N} \mathbf{X}^T \mathbf{X} \mathbf{w} - \frac{2}{N} \mathbf{X}^T \mathbf{t} + 2\lambda \mathbf{w}$$

Uguagliando la derivata a 0 otteniamo:

$$\frac{2}{N} \mathbf{X}^T \mathbf{X} \mathbf{w} - \frac{2}{N} \mathbf{X}^T \mathbf{t} + 2\lambda \mathbf{w} = 0$$

$$\left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right) \mathbf{w} = \mathbf{X}^T \mathbf{t}$$

$$\left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right)^{-1} \left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right) \mathbf{w} = \left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{t}$$

$$\mathbf{I} \mathbf{w} = \left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{t}$$

$$\Rightarrow \hat{\mathbf{w}} = \left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{t}$$

Regolarizzazione

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

Regolarizzazione

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

- Ovviamente, scegliendo $\lambda = 0$ otteniamo lo stesso risultato che abbiamo già derivato in precedenza, per il regressore senza regolarizzazione

Regolarizzazione

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

- Ovviamente, scegliendo $\lambda = 0$ otteniamo lo stesso risultato che abbiamo già derivato in precedenza, per il regressore senza regolarizzazione
- Come già detto, scegliendo valori di λ elevati si obbliga il nostro regressore ad essere meno complicato. Attenzione: bastano leggere variazioni di λ per ottenere regressori molto diversi

Regolarizzazione

Effetto della regolarizzazione su un regressore lineare polinomiale di grado 5