Scuola universitaria professionale della Svizzera italiana
Bachelor di Ingegneria Informatica

SUPSI

Machine Learning

Lezione 3 - Regolarizzazione

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

IDSIA-SUPSI, Galleria 1, Manno

Capitolo 2 del libro
A First Course in Machine Learning

Coefficiente di Determinazione - R^2

- E' un valore tra 0 e 1 che ci aiuta a capire la qualità del nostro regressore

Coefficiente di Determinazione - R^2

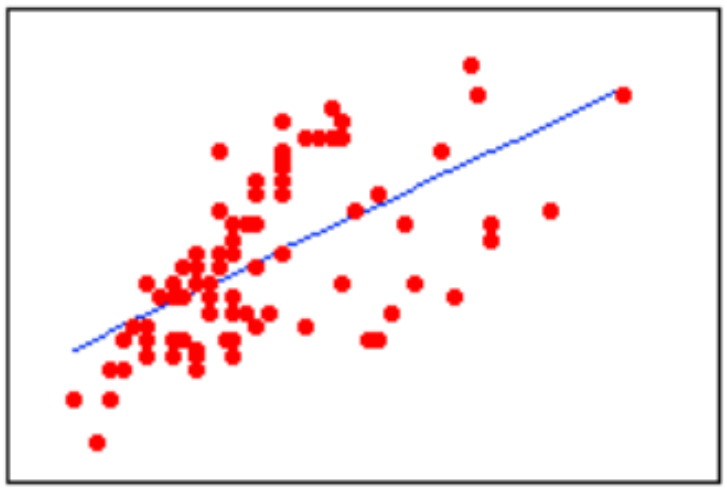
- E' un valore tra 0 e 1 che ci aiuta a capire la qualità del nostro regressore
- Misura la frazione della varianza dei dati osservati che è osservabile grazie ai dati di input

Coefficiente di Determinazione - R^2

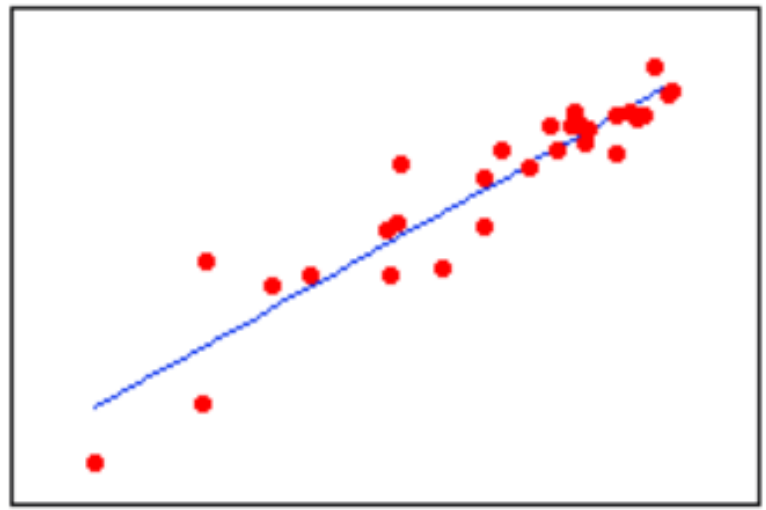
- E' un valore tra 0 e 1 che ci aiuta a capire la qualità del nostro regressore
- Misura la frazione della varianza dei dati osservati che è osservabile grazie ai dati di input
- 0 significa che il regressore spiega nulla della variabilità presente nelle osservazioni
- 1 significa che il regressore spiega tutta la variabilità presente nelle osservazioni

Coefficiente di Determinazione - R^2

- E' un valore tra 0 e 1 che ci aiuta a capire la qualità del nostro regressore
- Misura la frazione della varianza dei dati osservati che è osservabile grazie ai dati di input
- 0 significa che il regressore spiega nulla della variabilità presente nelle osservazioni
- 1 significa che il regressore spiega tutta la variabilità presente nelle osservazioni

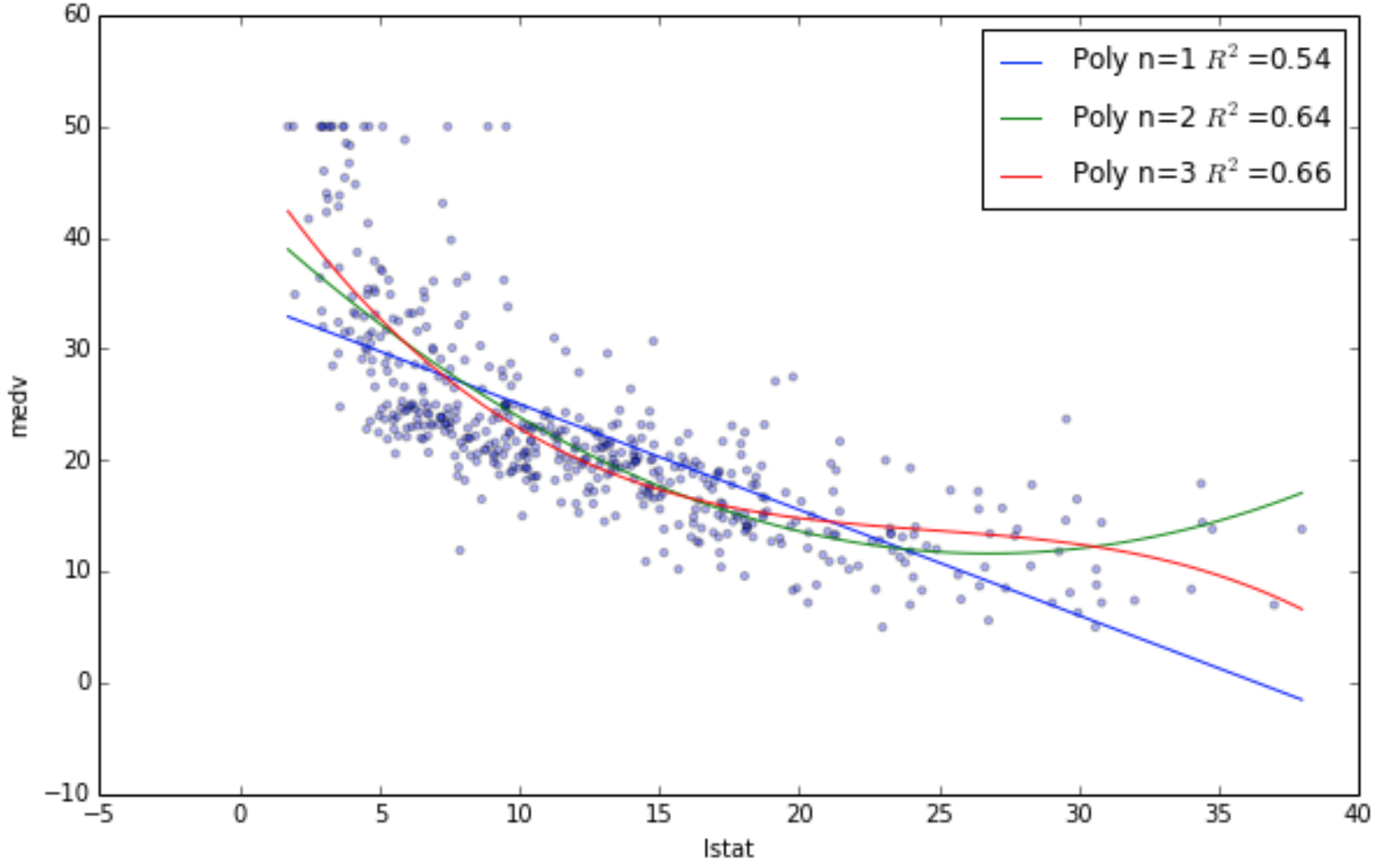


$R^2 = 38\%$

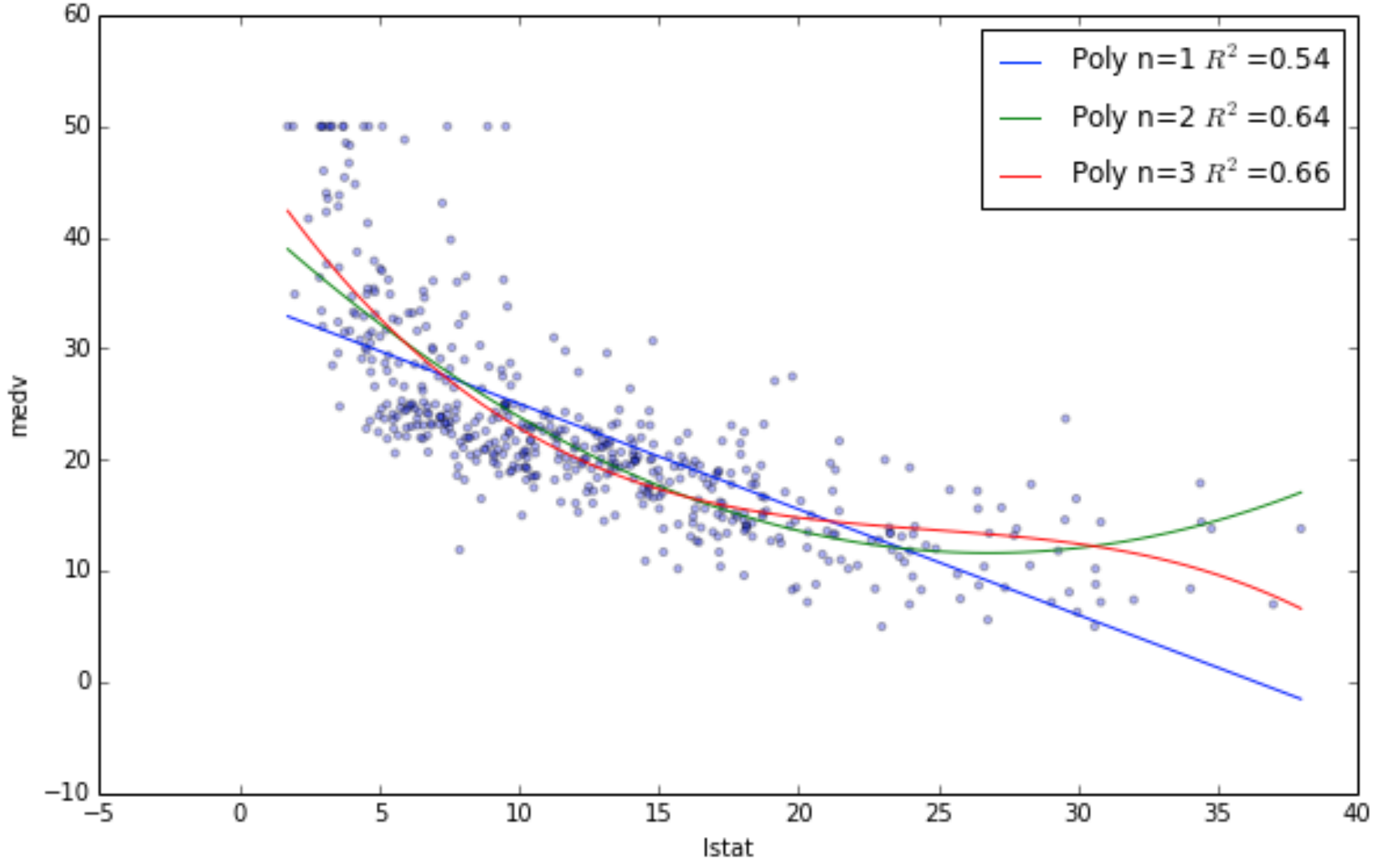


$R^2 = 87.4\%$

Coefficiente di Determinazione - R^2



Coefficiente di Determinazione - R^2



$\Rightarrow$Tentare di aumentare forzatamente R^2 porta all'over-fitting

Coefficiente di Determinazione - R^2

Immaginiamo di avere un set di N input scalari x_n con le relative osservazioni t_n. Costruiamo un regressore lineare che predice per ogni input il valore $f(x_n; w) = w_0 + w_1 x_n$

Coefficiente di Determinazione - R^2

Immaginiamo di avere un set di N input scalari x_n con le relative osservazioni t_n. Costruiamo un regressore lineare che predice per ogni input il valore $f(x_n; w) = w_0 + w_1 x_n$

$$\bar{t} \triangleq \frac{1}{N} \sum_{n=1}^{N} t_n$$

Coefficiente di Determinazione - R^2

Immaginiamo di avere un set di N input scalari x_n con le relative osservazioni t_n. Costruiamo un regressore lineare che predice per ogni input il valore $f(x_n; w) = w_0 + w_1 x_n$

$$\bar{t} \triangleq \frac{1}{N} \sum_{n=1}^{N} t_n$$

Residual Sum of Squares: $SS_{\text{res}} \triangleq \sum_{n=1}^{N} (t_n - f(x_n; w))^2$

Coefficiente di Determinazione - R^2

Immaginiamo di avere un set di N input scalari x_n con le relative osservazioni t_n. Costruiamo un regressore lineare che predice per ogni input il valore $f(x_n; w) = w_0 + w_1 x_n$

$$\bar{t} \triangleq \frac{1}{N} \sum_{n=1}^{N} t_n$$

Residual Sum of Squares: $SS_{\text{res}} \triangleq \sum_{n=1}^{N} (t_n - f(x_n; w))^2$

Total Sum of Squares: $SS_{\text{tot}} \triangleq \sum_{n=1}^{N} (t_n - \bar{t})^2$

Coefficiente di Determinazione - R^2

Immaginiamo di avere un set di N input scalari x_n con le relative osservazioni t_n. Costruiamo un regressore lineare che predice per ogni input il valore $f(x_n; w) = w_0 + w_1 x_n$

$$\bar{t} \triangleq \frac{1}{N} \sum_{n=1}^{N} t_n$$

Residual Sum of Squares: $SS_{\text{res}} \triangleq \sum_{n=1}^{N} (t_n - f(x_n; w))^2$

Total Sum of Squares: $SS_{\text{tot}} \triangleq \sum_{n=1}^{N} (t_n - \bar{t})^2$

$$R^2 \triangleq 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Coefficiente di Determinazione - R^2

Adjusted R^2

Immaginiamo di avere un set di N input vettoriali $\mathbf{x}_n \in \mathbb{R}^K$ con le relative osservazioni t_n. Costruiamo un regressore lineare che predice per ogni input il valore $f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}$

Coefficiente di Determinazione - R^2

Adjusted R^2

Immaginiamo di avere un set di N input vettoriali $\mathbf{x}_n \in \mathbb{R}^K$ con le relative osservazioni t_n. Costruiamo un regressore lineare che predice per ogni input il valore $f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}$

Adjusted R^2: $\bar{R}^2 \triangleq 1 - (1 - R^2) \frac{N-1}{N-K-1}$

Coefficiente di Determinazione - R^2

Adjusted R^2

Immaginiamo di avere un set di N input vettoriali $\mathbf{x}_n \in \mathbb{R}^K$ con le relative osservazioni t_n. Costruiamo un regressore lineare che predice per ogni input il valore $f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}$

Adjusted R^2: $\bar{R}^2 \triangleq 1 - (1 - R^2) \frac{N-1}{N-K-1}$

Adjusted R^2 può essere negativo ed è sempre minore o uguale di R^2

Regolarizzazione

Regolarizzazione - Regressione Ridge

- Abbiamo visto che il nostro modello può essere scritto:

Regolarizzazione - Regressione Ridge

- Abbiamo visto che il nostro modello può essere scritto:

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

Regolarizzazione - Regressione Ridge

- Abbiamo visto che il nostro modello può essere scritto:

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

dove:

Regolarizzazione - Regressione Ridge

- 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 - Regressione Ridge

- 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 - Regressione Ridge

- 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 - Regressione Ridge

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

Regolarizzazione - Regressione Ridge

- 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 - Regressione Ridge

- 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)

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda\mathbf{w}^T\mathbf{w} = \mathcal{L} + \lambda\|\mathbf{w}\|_2^2$$

Regolarizzazione - Regressione Ridge

- 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)

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda\mathbf{w}^T\mathbf{w} = \mathcal{L} + \lambda\|\mathbf{w}\|_2^2$$

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

Regolarizzazione - Regressione Ridge

- 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)

$$\mathcal{L}' \triangleq \mathcal{L} + \lambda\mathbf{w}^T\mathbf{w} = \mathcal{L} + \lambda\|\mathbf{w}\|_2^2$$

- $\lambda > 0$ 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 - Regressione Ridge

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

Regolarizzazione - Regressione Ridge

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

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

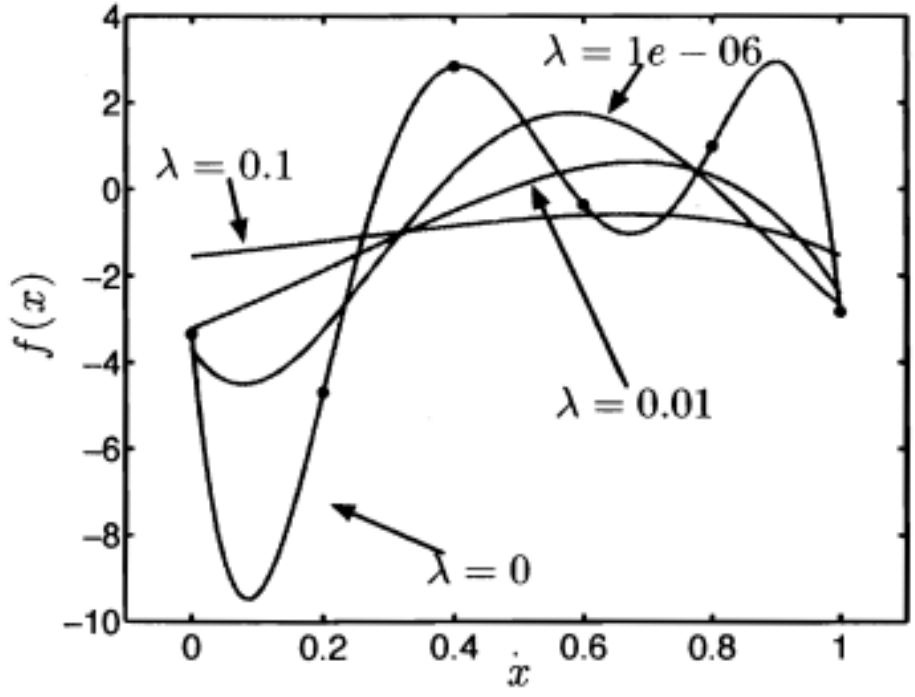
Regolarizzazione - Regressione Ridge

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

- 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 - Regressione Ridge

Effetto della regolarizzazione ridge su un regressore lineare polinomiale di grado 5



Regolarizzazione Generalizzata

- Abbiamo capito che per avere un modello meno complicato e per controllare l'over-fitting dobbiamo *regolarizzare* il vettore dei parametri $\mathbf{w}$

Regolarizzazione Generalizzata

- Abbiamo capito che per avere un modello meno complicato e per controllare l'over-fitting dobbiamo *regolarizzare* il vettore dei parametri $\mathbf{w}$

$$\mathcal{L}' = \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w} = \mathcal{L} + \lambda \sum_{n=1}^{K} w_n^2$$

Regolarizzazione Generalizzata

- Abbiamo capito che per avere un modello meno complicato e per controllare l'over-fitting dobbiamo *regolarizzare* il vettore dei parametri $\mathbf{w}$

$$\mathcal{L}' = \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w} = \mathcal{L} + \lambda \sum_{n=1}^{K} w_n^2$$

- Possiamo generalizzare l'approccio proposto in questo modo:

Regolarizzazione Generalizzata

- Abbiamo capito che per avere un modello meno complicato e per controllare l'over-fitting dobbiamo *regolarizzare* il vettore dei parametri $\mathbf{w}$

$$\mathcal{L}' = \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w} = \mathcal{L} + \lambda \sum_{n=1}^{K} w_n^2$$

- Possiamo generalizzare l'approccio proposto in questo modo:

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n|^q$$

Regolarizzazione Generalizzata

- Abbiamo capito che per avere un modello meno complicato e per controllare l'over-fitting dobbiamo *regolarizzare* il vettore dei parametri $\mathbf{w}$

$$\mathcal{L}' = \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w} = \mathcal{L} + \lambda \sum_{n=1}^{K} w_n^2$$

- Possiamo generalizzare l'approccio proposto in questo modo:

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n|^q$$

- Con $q = 2$ ritorniamo al caso di Regressione Ridge

Regolarizzazione Generalizzata

- Abbiamo capito che per avere un modello meno complicato e per controllare l'over-fitting dobbiamo *regolarizzare* il vettore dei parametri $\mathbf{w}$

$$\mathcal{L}' = \mathcal{L} + \lambda \mathbf{w}^T \mathbf{w} = \mathcal{L} + \lambda \sum_{n=1}^{K} w_n^2$$

- Possiamo generalizzare l'approccio proposto in questo modo:

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n|^q$$

- Con $q = 2$ ritorniamo al caso di Regressione Ridge
- Con $q = 1$ otteniamo il metodo LASSO

Regolarizzazione LASSO

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n| = \mathcal{L} + \lambda \|\mathbf{w}\|_1$$

Regolarizzazione LASSO

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n| = \mathcal{L} + \lambda \|\mathbf{w}\|_1$$

- A differenza della regolarizzazione Ridge che impedisce agli elementi di **w** di diventare troppo grandi,la regolarizzazione LASSO rende alcuni elementi di **w** esattamente 0

Regolarizzazione LASSO

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n| = \mathcal{L} + \lambda \|\mathbf{w}\|_1$$

- A differenza della regolarizzazione Ridge che impedisce agli elementi di **w** di diventare troppo grandi,la regolarizzazione LASSO rende alcuni elementi di **w** esattamente 0
- La regolarizzazione LASSO permette quindi di fare feature selection e classificazione $\Rightarrow$ sopravvivono (cioè, rimangono diversi da 0) solo gli elementi di w associati alle informazioni più significative del nostro regressore

Regolarizzazione LASSO

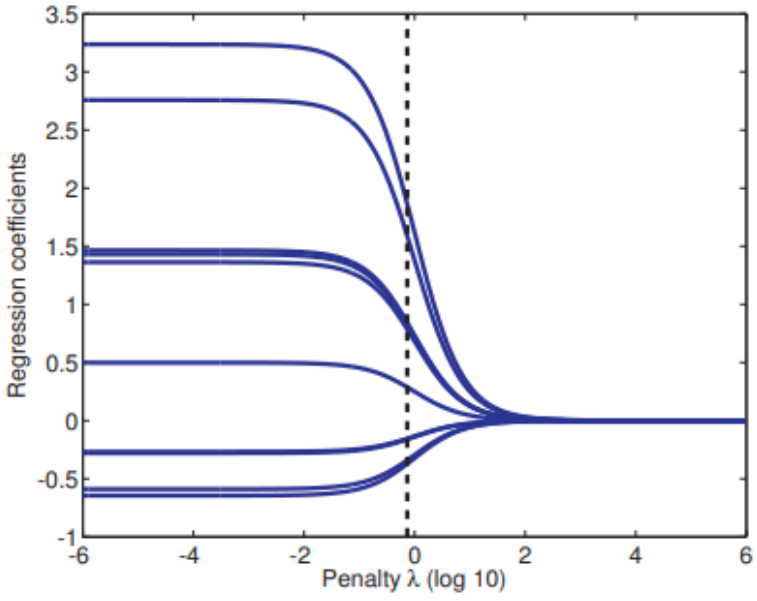
- Dati generati in maniera sintetica, con $\hat{\mathbf{w}} = [3; 2; 1; 0; 0; 0; 0; 0; 0; 0]$

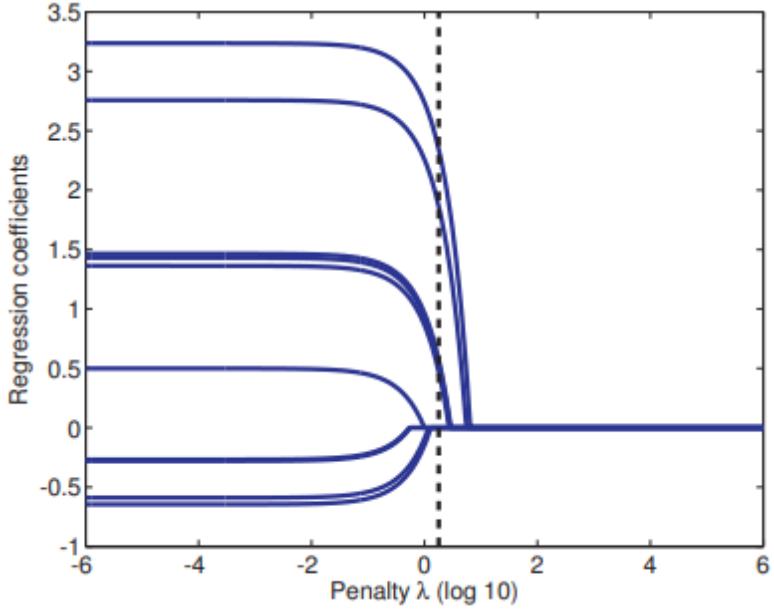
Regolarizzazione LASSO

- Dati generati in maniera sintetica, con $\hat{\mathbf{w}} = [3; 2; 1; 0; 0; 0; 0; 0; 0; 0]$
- Il vettore dei parametri ottimale ha solo 3 componenti significative

Regolarizzazione LASSO

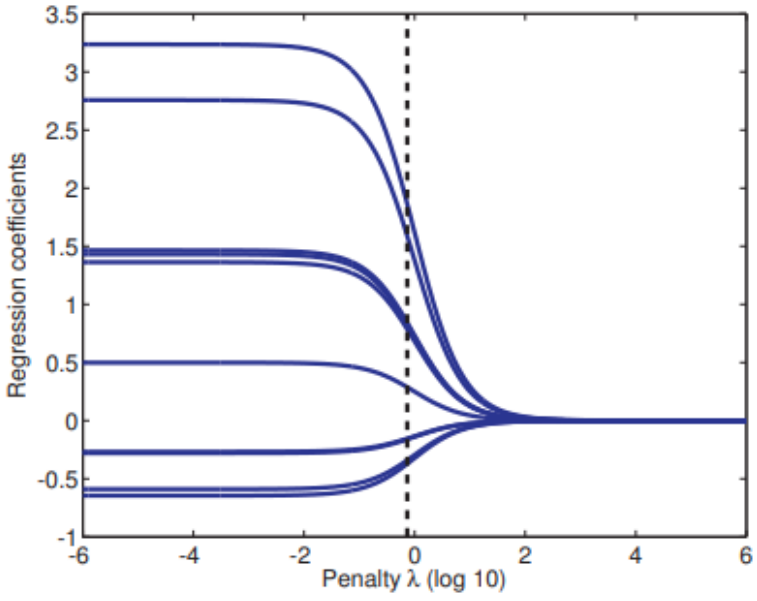
- Dati generati in maniera sintetica, con $\hat{\mathbf{w}} = [3; 2; 1; 0; 0; 0; 0; 0; 0; 0]$
- Il vettore dei parametri ottimale ha solo 3 componenti significative

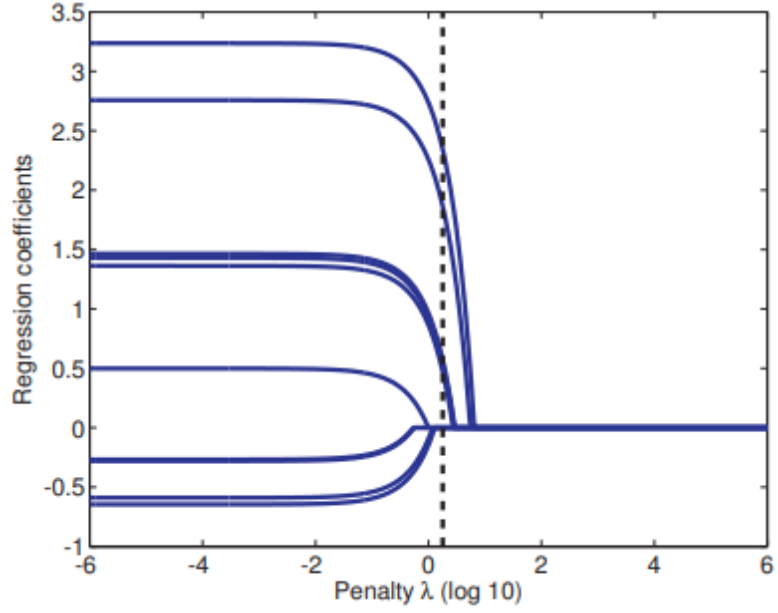




Regolarizzazione LASSO

- Dati generati in maniera sintetica, con $\hat{\mathbf{w}} = [3; 2; 1; 0; 0; 0; 0; 0; 0; 0]$
- Il vettore dei parametri ottimale ha solo 3 componenti significative

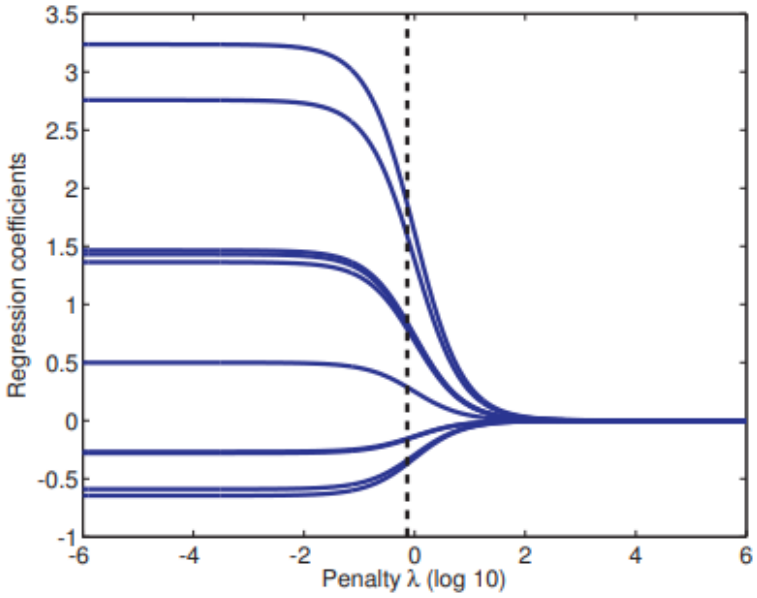


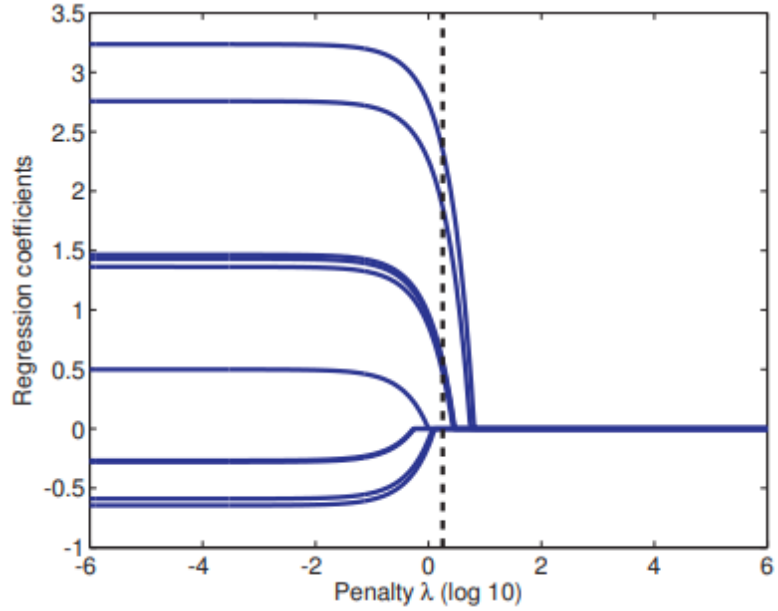


- La regressione ridge non riesce ad ottenere un vettore $\mathbf{w}$ sparso

Regolarizzazione LASSO

- Dati generati in maniera sintetica, con $\hat{\mathbf{w}} = [3; 2; 1; 0; 0; 0; 0; 0; 0; 0]$
- Il vettore dei parametri ottimale ha solo 3 componenti significative





- La regressione ridge non riesce ad ottenere un vettore $\mathbf{w}$ sparso
- La regressione LASSO combina insieme i) il controllare il valore degli elementi in $\mathbf{w}$ e ii) la capacità di operare *variable selection*

LASSO - interpretazione geometrica

- Abbiamo visto che in maniera generica una regressione lineare generalizzata consiste nel minimizzare la funzione costo:

LASSO - interpretazione geometrica

- Abbiamo visto che in maniera generica una regressione lineare generalizzata consiste nel minimizzare la funzione costo:

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n|^q$$

LASSO - interpretazione geometrica

- Abbiamo visto che in maniera generica una regressione lineare generalizzata consiste nel minimizzare la funzione costo:

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n|^q$$

- E' facile dimostrare che minimizzare $\mathcal{L}'$ è equivalente a minimizzare $\mathcal{L}$ rispettando il vincolo: $\sum_{n=1}^{K} |w_n|^q \leq \bar{\lambda}$

LASSO - interpretazione geometrica

- Abbiamo visto che in maniera generica una regressione lineare generalizzata consiste nel minimizzare la funzione costo:

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n|^q$$

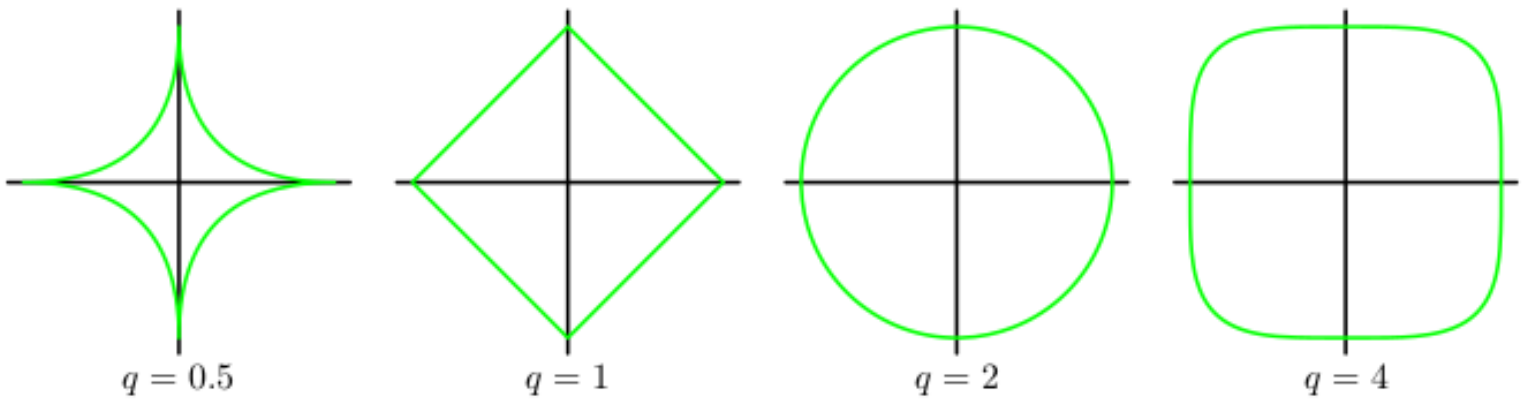
- E' facile dimostrare che minimizzare $\mathcal{L}'$ è equivalente a minimizzare $\mathcal{L}$ rispettando il vincolo: $\sum_{n=1}^{K} |w_n|^q \leq \bar{\lambda}$
  - $\mathcal{L}$ ricordiamo essere una funzione quadratica/parabolica
  - $\bar{\lambda}$ è un valore ottenibile a partire da λ

LASSO - interpretazione geometrica

- Abbiamo visto che in maniera generica una regressione lineare generalizzata consiste nel minimizzare la funzione costo:

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n|^q$$

- E' facile dimostrare che minimizzare $\mathcal{L}'$ è equivalente a minimizzare $\mathcal{L}$ rispettando il vincolo: $\sum_{n=1}^{K} |w_n|^q \leq \bar{\lambda}$
  - $\mathcal{L}$ ricordiamo essere una funzione quadratica/parabolica
  - $\bar{\lambda}$ è un valore ottenibile a partire da λ



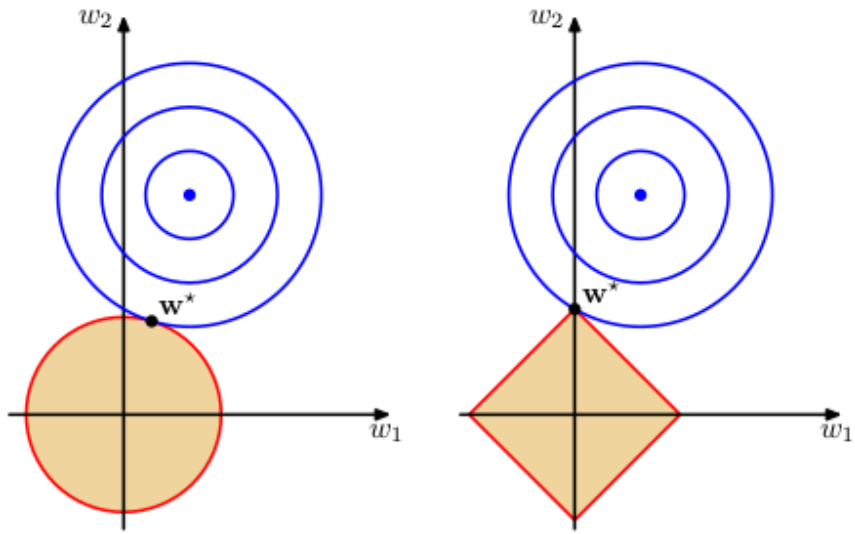
Forma del vincolo imposto dalla regolarizzazione con $\bar{\lambda} = 1$

LASSO - interpretazione geometrica

- Abbiamo visto che in maniera generica una regressione lineare generalizzata consiste nel minimizzare la funzione costo:

$$\mathcal{L}' = \mathcal{L} + \lambda \sum_{n=1}^{K} |w_n|^q$$

- E' facile dimostrare che minimizzare $\mathcal{L}'$ è equivalente a minimizzare $\mathcal{L}$ rispettando il vincolo: $\sum_{n=1}^{K} |w_n|^q \leq \bar{\lambda}$
  - $\mathcal{L}$ ricordiamo essere una funzione quadratica/parabolica
  - $\bar{\lambda}$ è un valore ottenibile a partire da λ



La regolarizzazione LASSO trova una soluzione sparsa con $\hat{w}_1 = 0$

Regolarizzazione Elastic Net

- Nel corso del tempo sono state proposte ed utilizzate decine di regolarizzazioni diverse

Regolarizzazione Elastic Net

- Nel corso del tempo sono state proposte ed utilizzate decine di regolarizzazioni diverse
- Il LASSO tende a comportarsi male quando molte variabili sono correlate tra di loro

Regolarizzazione Elastic Net

- Nel corso del tempo sono state proposte ed utilizzate decine di regolarizzazioni diverse
- Il LASSO tende a comportarsi male quando molte variabili sono correlate tra di loro
- La regolarizzazione Elastic Net prova ad unire i vantaggi di Ridge e LASSO:

Regolarizzazione Elastic Net

- Nel corso del tempo sono state proposte ed utilizzate decine di regolarizzazioni diverse
- Il LASSO tende a comportarsi male quando molte variabili sono correlate tra di loro
- La regolarizzazione Elastic Net prova ad unire i vantaggi di Ridge e LASSO:

$$\begin{aligned}\mathcal{L}' &= \mathcal{L} + \lambda_1 \sum_{n=1}^{K} |w_n| + \lambda_2 \sum_{n=1}^{N} w_n^2 \\ &= \mathcal{L} + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2\end{aligned}$$

Algoritmi

- LASSO contiene la norma l_1 (cioè $\|\cdot\|_1$), quindi è nondifferenziabile

Algoritmi

- LASSO contiene la norma l_1 (cioè $\|\cdot\|_1$), quindi è nondifferenziabile
- Elastic Net sembra rendere la funzione costo $\mathcal{L}'$ piuttosto complicata (contiene sempre la norma l_1)

Algoritmi

- LASSO contiene la norma l_1 (cioè $\|\cdot\|_1$), quindi è nondifferenziabile
- Elastic Net sembra rendere la funzione costo $\mathcal{L}'$ piuttosto complicata (contiene sempre la norma l_1)
- Nel caso semplice di regolarizzazione ridge, il vettore di parametri ottimale si ottiene come: $\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X} + N\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{t} \Rightarrow$ Richiede una quantità enorme di tempo calcolare la matrice inversa quando $\mathbf{X}$ è di grandi dimensioni (Big Data)

Algoritmi

- LASSO contiene la norma l_1 (cioè $\|\cdot\|_1$), quindi è nondifferenziabile
- Elastic Net sembra rendere la funzione costo $\mathcal{L}'$ piuttosto complicata (contiene sempre la norma l_1)
- Nel caso semplice di regolarizzazione ridge, il vettore di parametri ottimale si ottiene come: $\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X} + N\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{t} \Rightarrow$ Richiede una quantità enorme di tempo calcolare la matrice inversa quando $\mathbf{X}$ è di grandi dimensioni (Big Data)

Come risolvere questi problemi?

Algoritmi

- LASSO contiene la norma l_1 (cioè $\|\cdot\|_1$), quindi è nondifferenziabile
- Elastic Net sembra rendere la funzione costo $\mathcal{L}'$ piuttosto complicata (contiene sempre la norma l_1)
- Nel caso semplice di regolarizzazione ridge, il vettore di parametri ottimale si ottiene come: $\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X} + N\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{t} \Rightarrow$ Richiede una quantità enorme di tempo calcolare la matrice inversa quando $\mathbf{X}$ è di grandi dimensioni (Big Data)

Come risolvere questi problemi?

$\Rightarrow$ Algoritmi Iterativi

Algoritmi

- LASSO contiene la norma l_1 (cioè $\|\cdot\|_1$), quindi è nondifferenziabile
- Elastic Net sembra rendere la funzione costo $\mathcal{L}'$ piuttosto complicata (contiene sempre la norma l_1)
- Nel caso semplice di regolarizzazione ridge, il vettore di parametri ottimale si ottiene come: $\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X} + N\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{t} \Rightarrow$ Richiede una quantità enorme di tempo calcolare la matrice inversa quando $\mathbf{X}$ è di grandi dimensioni (Big Data)

Come risolvere questi problemi?

$\Rightarrow$ Algoritmi Iterativi

Ne esistono centinaia:

Algoritmi

- LASSO contiene la norma l_1 (cioè $\|\cdot\|_1$), quindi è nondifferenziabile
- Elastic Net sembra rendere la funzione costo $\mathcal{L}'$ piuttosto complicata (contiene sempre la norma l_1)
- Nel caso semplice di regolarizzazione ridge, il vettore di parametri ottimale si ottiene come: $\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X} + N\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{t} \Rightarrow$ Richiede una quantità enorme di tempo calcolare la matrice inversa quando $\mathbf{X}$ è di grandi dimensioni (Big Data)

Come risolvere questi problemi?
$\Rightarrow$ Algoritmi Iterativi

Ne esistono centinaia:

- Gradient Descent
- Proximal Gradient
- Coordinate Descent
- Neural Networks
- Netwon Methods
- ...

Metodo di Discesa del Gradiente (Gradient Descent)

- Caso semplice - funzione costo senza regolarizzazione:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2 = \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{t}\|_2^2$$

Metodo di Discesa del Gradiente (Gradient Descent)

- Caso semplice - funzione costo senza regolarizzazione:

  $\mathcal{L}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} (t_n - f(\mathbf{x}_n; \mathbf{w}))^2 = \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{t}\|_2^2$

- Scegli un punto di partenza random: $\mathbf{w}^0 \in \mathbb{R}^K$

Metodo di Discesa del Gradiente (Gradient Descent)

- Caso semplice - funzione costo senza regolarizzazione:

  $$\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2 = \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{t}\|_2^2$$

- Scegli un punto di partenza random: $\mathbf{w}^0 \in \mathbb{R}^K$
- Scegli un passo (stespize o learning rate): $\alpha \in (0; 1]$

Metodo di Discesa del Gradiente (Gradient Descent)

- Caso semplice - funzione costo senza regolarizzazione:
  $\mathcal{L}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} (t_n - f(\mathbf{x}_n; \mathbf{w}))^2 = \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{t}\|_2^2$
- Scegli un punto di partenza random: $\mathbf{w}^0 \in \mathbb{R}^K$
- Scegli un passo (stespize o learning rate): $\alpha \in (0; 1]$
- Inizializza un contatore per le iterazioni svolte: $k = 1$

Metodo di Discesa del Gradiente (Gradient Descent)

- Caso semplice - funzione costo senza regolarizzazione:
  $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2 = \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{t}\|_2^2$
- Scegli un punto di partenza random: $\mathbf{w}^0 \in \mathbb{R}^K$
- Scegli un passo (stespize o learning rate): $\alpha \in (0; 1]$
- Inizializza un contatore per le iterazioni svolte: $k = 1$

Il metodo di Discesa del Gradiente aggiorna il vettore dei parametri secondo la seguente regola finché non raggiunge il valore ottimale o fino a quando un numero di iterazioni prefissato viene raggiunto:

Metodo di Discesa del Gradiente (Gradient Descent)

- Caso semplice - funzione costo senza regolarizzazione:
  $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2 = \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{t}\|_2^2$
- Scegli un punto di partenza random: $\mathbf{w}^0 \in \mathbb{R}^K$
- Scegli un passo (stespize o learning rate): $\alpha \in (0; 1]$
- Inizializza un contatore per le iterazioni svolte: $k = 1$

Il metodo di Discesa del Gradiente aggiorna il vettore dei parametri secondo la seguente regola finché non raggiunge il valore ottimale o fino a quando un numero di iterazioni prefissato viene raggiunto:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha\nabla\mathcal{L}(\mathbf{w}^k)$$

Metodo di Discesa del Gradiente (Gradient Descent)

- Caso semplice - funzione costo senza regolarizzazione:
  $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2 = \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{t}\|_2^2$
- Scegli un punto di partenza random: $\mathbf{w}^0 \in \mathbb{R}^K$
- Scegli un passo (stespize o learning rate): $\alpha \in (0; 1]$
- Inizializza un contatore per le iterazioni svolte: $k = 1$

Il metodo di Discesa del Gradiente aggiorna il vettore dei parametri secondo la seguente regola finché non raggiunge il valore ottimale o fino a quando un numero di iterazioni prefissato viene raggiunto:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha\nabla\mathcal{L}(\mathbf{w}^k)$$

$$\nabla\mathcal{L}(\mathbf{w}^k) = \frac{2}{N}\left(\mathbf{X}^T\mathbf{X}\mathbf{w}^k - \mathbf{X}^T\mathbf{t}\right)$$

Metodo di Discesa del Gradiente (Gradient Descent)

- Caso semplice - funzione costo senza regolarizzazione:
  $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2 = \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{t}\|_2^2$
- Scegli un punto di partenza random: $\mathbf{w}^0 \in \mathbb{R}^K$
- Scegli un passo (stespize o learning rate): $\alpha \in (0; 1]$
- Inizializza un contatore per le iterazioni svolte: $k = 1$

Il metodo di Discesa del Gradiente aggiorna il vettore dei parametri secondo la seguente regola finché non raggiunge il valore ottimale o fino a quando un numero di iterazioni prefissato viene raggiunto:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha\nabla\mathcal{L}(\mathbf{w}^k)$$

$$\nabla\mathcal{L}(\mathbf{w}^k) = \frac{2}{N}\left(\mathbf{X}^T\mathbf{X}\mathbf{w}^k - \mathbf{X}^T\mathbf{t}\right)$$

$\mathcal{L}$ è una funzione convessa, quindi questo metodo converge a un valore ottimo $\hat{\mathbf{w}}$. La velocità di convergenza va come $\frac{1}{k}$ (il che significa che se si vuole ottenere un'accuratezza di 0.1, servono circa 1/0.1=10 iterazioni)