Introduzione al Machine Learning: Rischio atteso, rischio empirico, generalizzazione

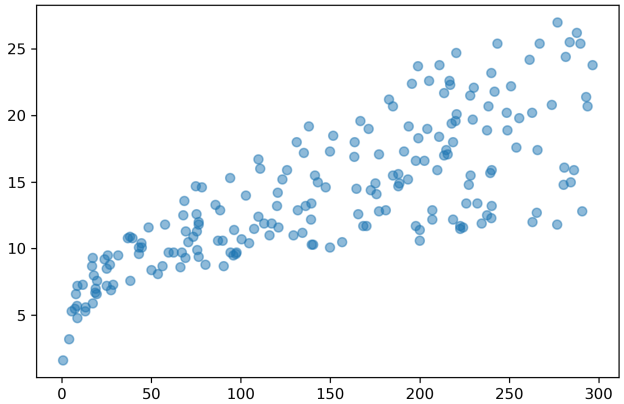
Vincenzo Bonifaci

Esempio: Ritorno da investimenti pubblicitari

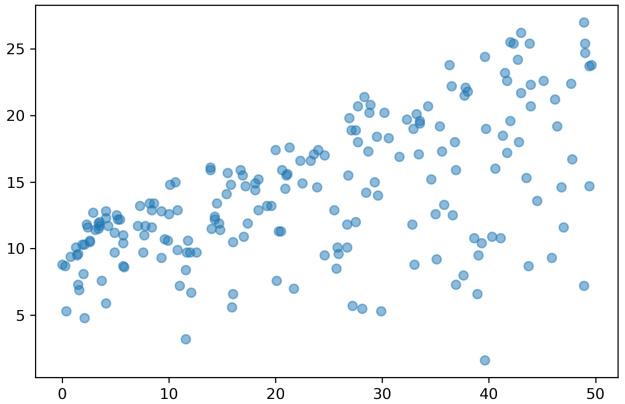
Input: investimenti pubblicitari via TV, radio e giornali in un mercato (in migliaia di Euro)
Output: unità di prodotto vendute in quel mercato (in migliaia)

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9
...	...	...	...	...

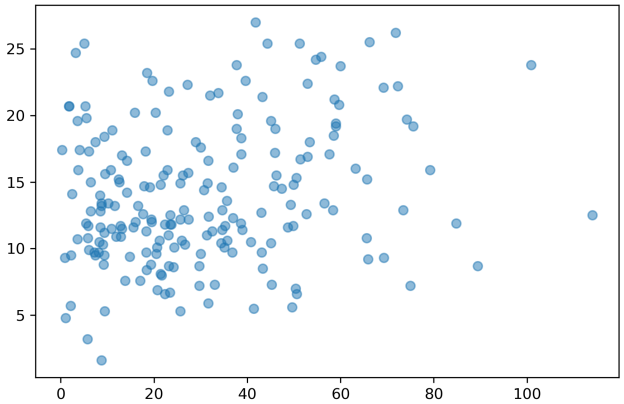
Esempio: Ritorno da investimenti pubblicitari



`sales` vs. `TV`



`sales` vs. `radio`



`sales` vs. `newspaper`

Problemi di predizione: input e output

- Spazio degli input $\mathcal{X}$
  Es.: insieme degli investimenti $\langle$ tv, radio, giornali $\rangle$ $(\mathbb{R}^3_+)$
- Spazio degli output $\mathcal{Y}$
  Es.: insieme delle possibili quantità di prodotto vendute $(\mathbb{R})$

Osservati un certo numero di esempi (x, y), vogliamo trovare una *regola di predizione* (o *ipotesi*)

$$h : \mathcal{X} \to \mathcal{Y}$$

che ricostruisca in maniera "accurata" la relazione ingresso-uscita

Nei problemi di *regressione* l'output è quantitativo (*numerico*)

Nei problemi di *classificazione* l'output è qualitativo (*categorico*)

Funzione di costo [Loss function]

Come quantifichiamo l'accuratezza di una regola di predizione $h : \mathcal{X} \to \mathcal{Y}$ su un particolare esempio?

Una *funzione di costo* è una funzione ℓ che prende una regola di predizione h ed un esempio $(x, y) \in \mathcal{X} \times \mathcal{Y}$, e restituisce un numero nonnegativo

$$\ell(h, (x, y)) \in \mathbb{R}_+$$

Una funzione di costo misura la discrepanza tra l'etichetta predetta ($\hat{y} = h(x)$) e l'etichetta osservata (y)

Esempi di funzioni di costo

- *Quadrato dell'errore*:

$$\ell(h,(x,y)) \stackrel{\text{def}}{=} (h(x)-y)^2$$

  Tipica dei problemi di regressione

- *Funzione costo 0-1*:

$$\ell(h,(x,y)) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } h(x)=y \\ 1 & \text{se } h(x)\neq y \end{cases}$$

  Tipica dei problemi di classificazione

Rischio atteso

Come quantifichiamo l'accuratezza di una regola di predizione $h : \mathcal{X} \to \mathcal{Y}$ in generale?

Assunzione fondamentale

Gli esempi (x, y) sono generati in modo indipendente da una distribuzione di probabilità (ignota) $\mathcal{D}$ sull'insieme $\mathcal{X} \times \mathcal{Y}$

Il *rischio atteso* di una regola di predizione h è

$$\mathrm{RA}(h) \stackrel{\text{def}}{=} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h,(x,y))\right]$$

A parole: il rischio atteso di h è il valore atteso della funzione di costo su h quando gli esempi sono generati dalla distribuzione $\mathcal{D}$

Quantifica il costo medio degli errori di predizione

Il problema del machine learning supervisionato

Problema del machine learning supervisionato

Fissata una distribuzione (ignota) $\mathcal{D}$ su $\mathcal{X} \times \mathcal{Y}$, cerca una regola di predizione che minimizzi il rischio atteso:

$$\underset{h:\mathcal{X}\to\mathcal{Y}}{\text{minimize}}\, \mathrm{RA}(h)$$

Il rischio atteso dipende dalla distribuzione ignota $\mathcal{D}$...

Come minimizzarlo, visto che non conosciamo $\mathcal{D}$?!

Rischio empirico

Non conosciamo $\mathcal{D}$ ma abbiamo degli *esempi* dalla distribuzione $\mathcal{D}$

Il *rischio empirico* di h sugli esempi $S = \{ (x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)}) \}$ è

$$\mathrm{RE}_S(h) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} \ell(h, (x^{(i)}, y^{(i)}))$$

Possiamo usare il rischio empirico come *surrogato* del rischio atteso: se il numero di esempi m è sufficientemente grande, è lecito sperare che i due valori siano vicini

Il principio ERM

Empirical Risk Minimization (ERM)

Dato un insieme di esempi S (generati da $\mathcal{D}$), cerca una regola di predizione che minimizzi il rischio empirico su S:

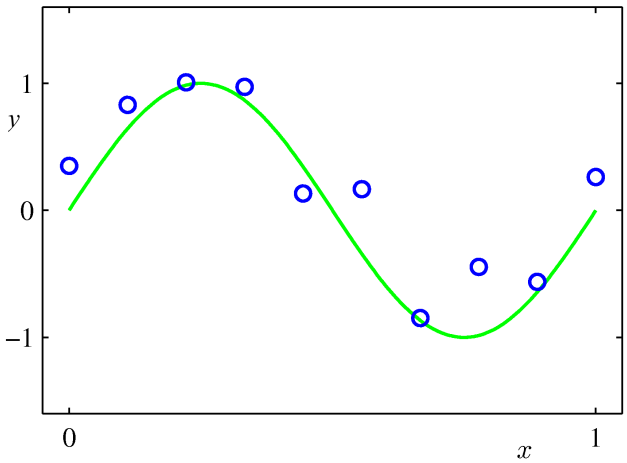
$$\underset{h}{\text{minimize}}\, \mathrm{RE}_S(h) \left(\equiv \underset{h}{\text{minimize}}\, \frac{1}{m} \sum_{i=1}^{m} \ell(h, (x^{(i)}, y^{(i)})) \right)$$

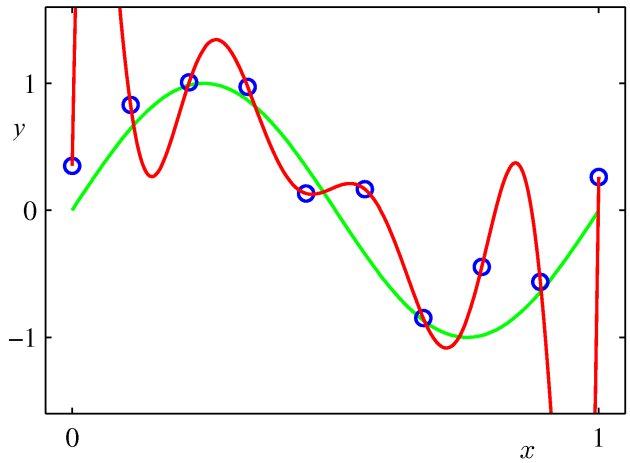
L'insieme S di esempi osservati dal learner è detto *training set*

Il problema del learning supervisionato diventa così un problema di ottimizzazione (nello spazio delle regole)

Il sovradattamento (overfitting)

Sebbene l'ERM sia un principio intuitivo, esso può completamente fallire senza le dovute cautele!





In questo esempio, la regola scelta (la funzione rossa) è *sovradattata* ai dati (*overfitting*):

"Spiega" perfettamente le osservazioni, ma non è un buon modello della distribuzione da cui i dati sono generati (funzione verde + rumore)

Il suo rischio empirico è nullo, ma il suo rischio atteso è alto

ERM con una classe di ipotesi ristretta

Un approccio per ovviare al problema dell'overfitting consiste nel limitare l'insieme delle possibili regole di predizione (ipotesi) h

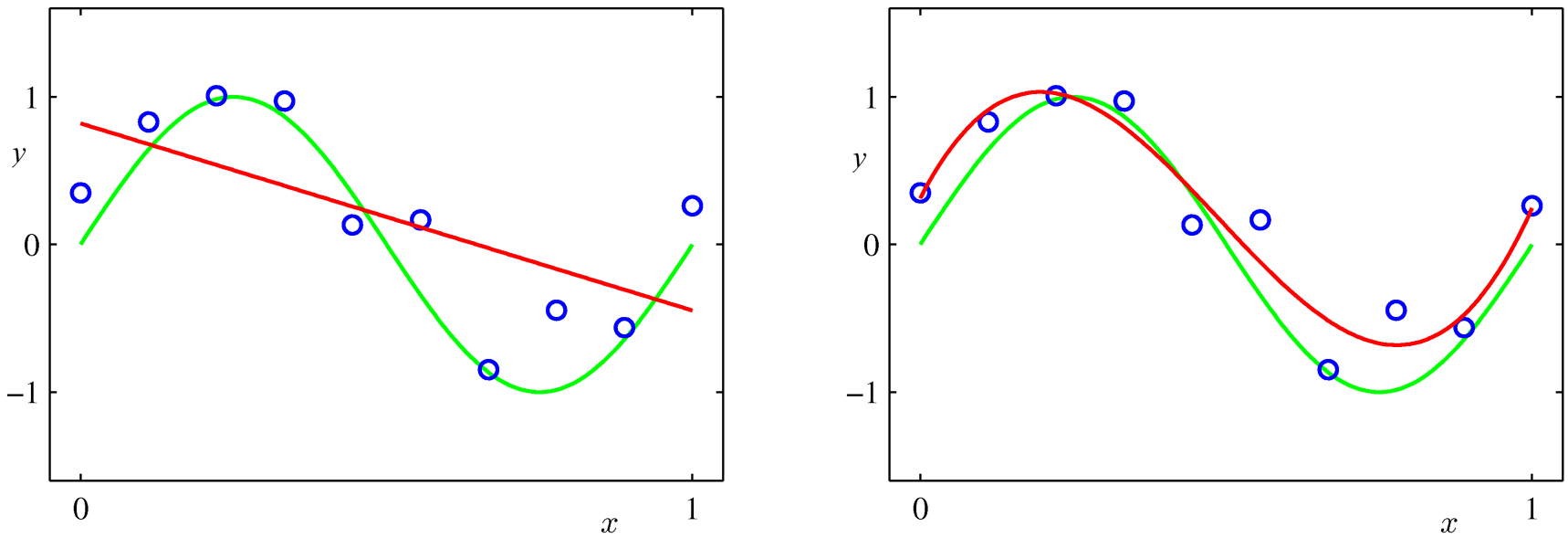
Anziché considerare la classe di tutte le funzioni da $\mathcal{X}$ a $\mathcal{Y}$, consideriamo solo una sua sottoclasse $\mathcal{H}$ (*insieme delle ipotesi*)

Applichiamo il principio ERM restringendoci alle ipotesi in $\mathcal{H}$:

$$\underset{h \in \mathcal{H}}{\text{minimize}}\, \mathrm{RE}_S(h)$$

- La classe $\mathcal{H}$ può incorporare la conoscenza pregressa del problema considerato, limitando la *complessità* delle ipotesi
- La classe $\mathcal{H}$ introduce un *pregiudizio (bias) induttivo*: tutte le regole non in $\mathcal{H}$ sono scartate a priori

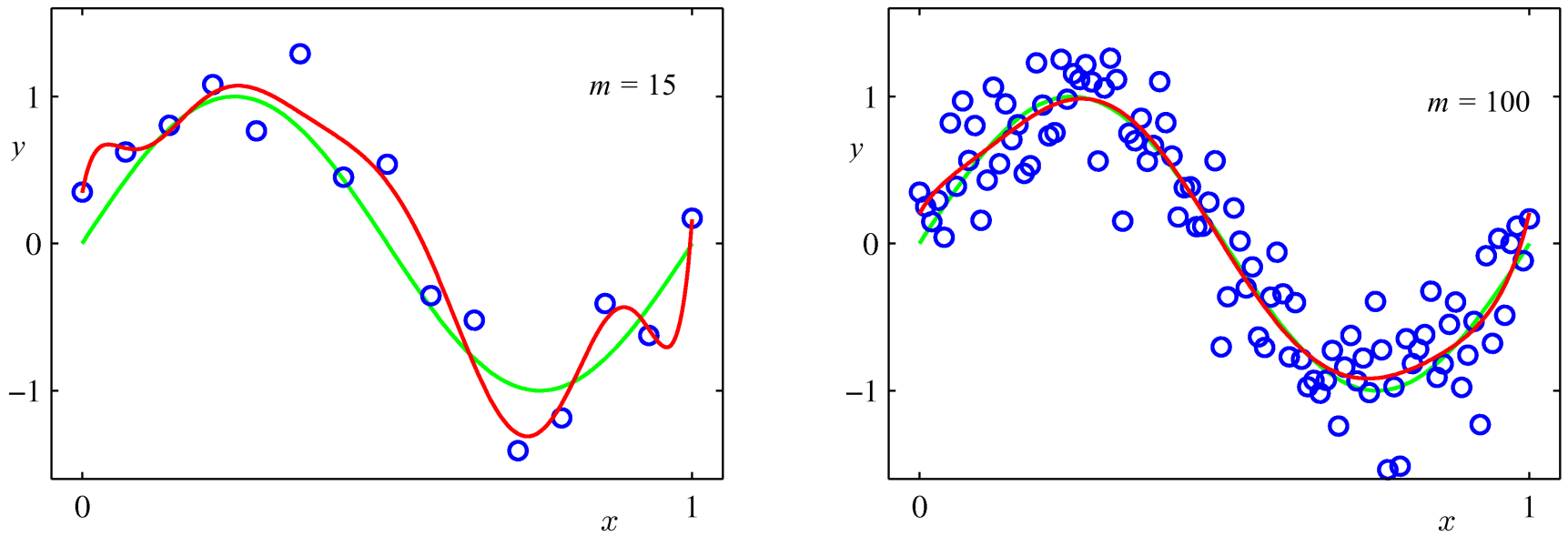
Compromesso bias-varianza



Fitting di un polinomio di grado 1 (sinistra) e di grado 3 (destra)

- Modelli più semplici hanno più bias: possono esibire *underfitting*

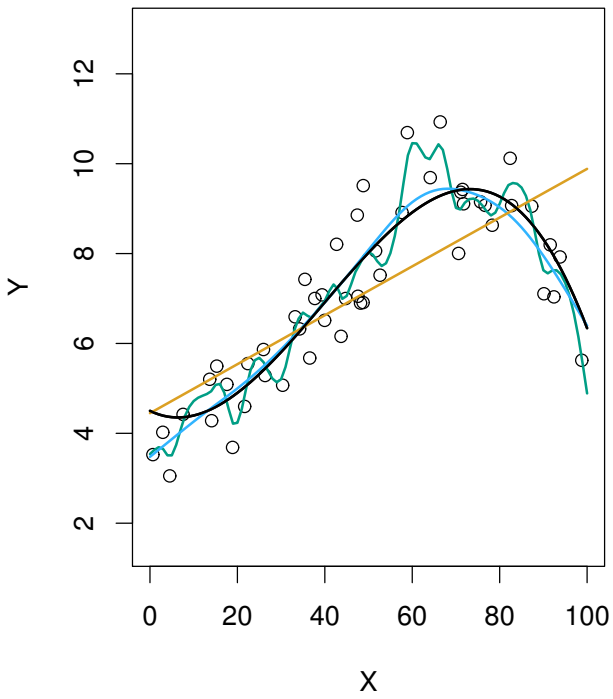
Compromesso bias-varianza

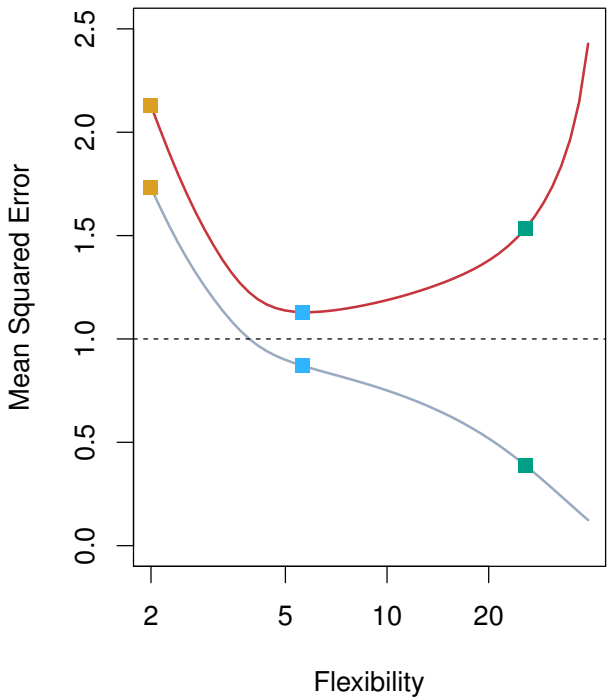


Fitting di un polinomio di grado 9 con 15 esempi (sinistra) e con 100 esempi (destra)

- Modelli più complessi hanno più varianza: richiedono più esempi per evitare *overfitting*

Compromesso bias-varianza





- Sinistra: I dati sono generati sommando la curva nera con un termine di rumore
  Le altre curve rappresentano regressioni polinomiali di grado 1, 5, e 23
- Destra: La curva grigia rappresenta il rischio empirico in funzione del grado
  La curva rossa rappresenta il rischio atteso in funzione del grado

Regressione lineare

Nella *regressione lineare*, l'insieme delle ipotesi è l'insieme $\mathcal{H}_{lin}$ delle funzioni lineari (affini) da $\mathcal{X} \equiv \mathbb{R}^d$ a $\mathcal{Y} \equiv \mathbb{R}$:

$$h \in \mathcal{H}_{lin} \Leftrightarrow h(x) = w_0 + w_1 x_1 + \ldots + w_d x_d \qquad (w_0, \ldots, w_d \in \mathbb{R})$$

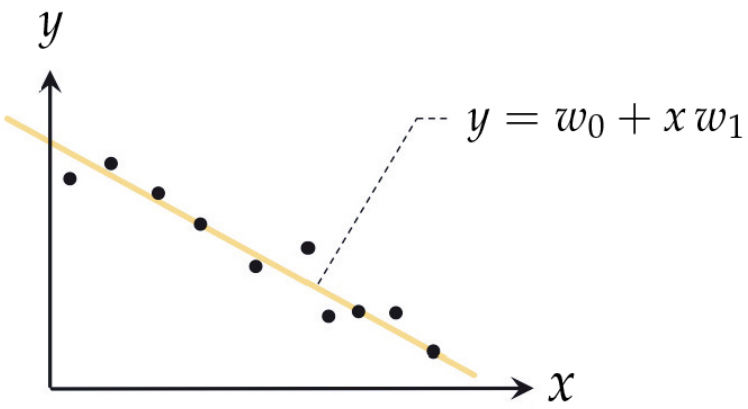
Useremo spesso la convenzione $x_0 \stackrel{\text{def}}{=} 1$, così da poter scrivere $h(x) = w^\top x$

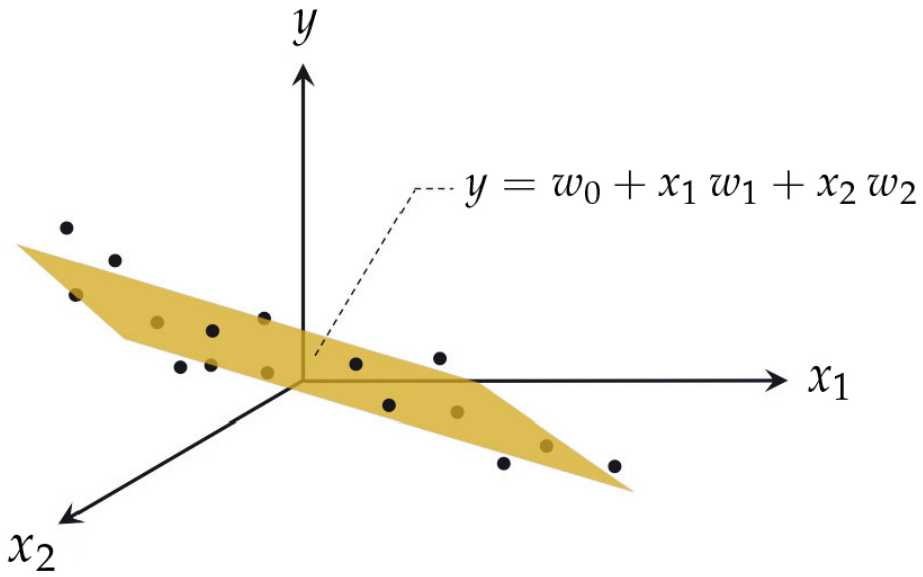
- w_0 è l'*intercetta* (valore previsto dal modello quando x è nullo)
- w_k è il *coefficiente* che esprime la dipendenza di $h(x)$ dalla k-esima componente di x

Una funzione di costo comunemente utilizzata è quella quadratica:

$$\ell(h, (x, y)) = (h(x) - y)^2$$

Regressione lineare



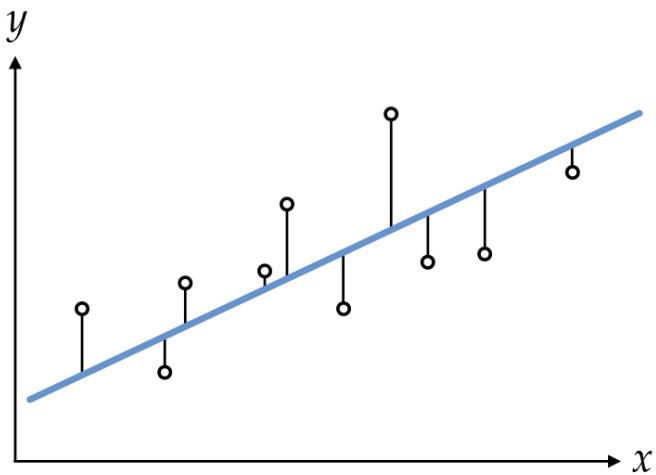


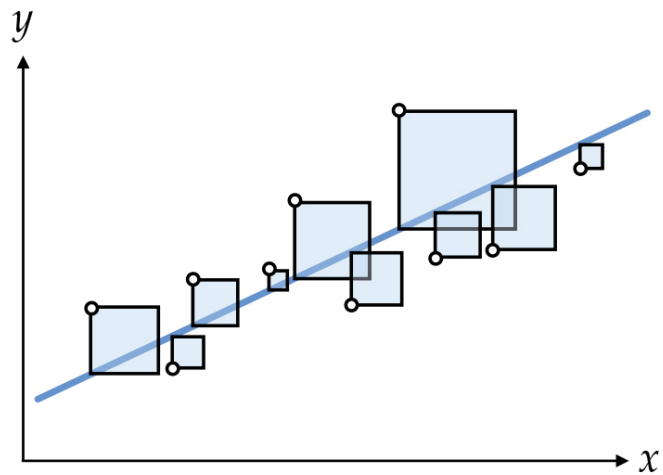
ERM per la regressione lineare

Nella regressione lineare con costo quadratico, il rischio empirico è dato dall'*errore quadratico medio* [*mean squared error*]:

Mean Squared Error (MSE)

$$\mathrm{RE}_S(h) = \frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2 = \frac{1}{m}\|Xw - y\|^2$$





ERM per la regressione lineare

Il minimizzatore del rischio empirico è calcolabile a partire dai dati attraverso una formula esplicita:

$$w^* = \left(\sum_{i=1}^{m} x^{(i)} x^{(i)\top} \right)^{-1} \left(\sum_{i=1}^{m} y^{(i)} x^{(i)} \right) = (X^\top X)^{-1} X^\top y$$

Infatti (si dimostra) deve soddisfare le cosiddette equazioni normali:

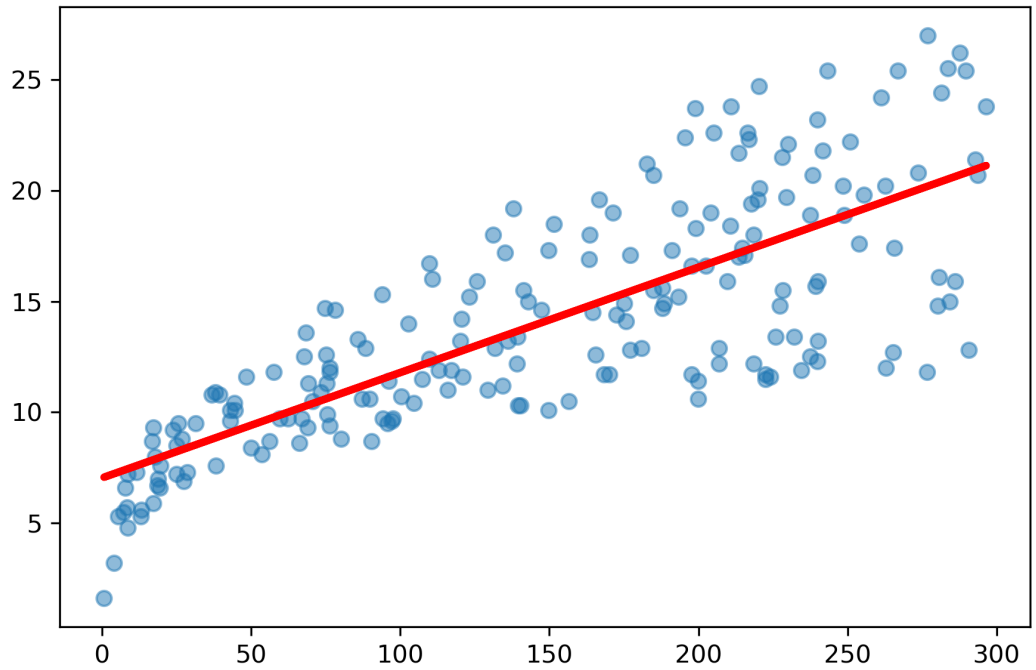
Equazioni normali

Se w^* minimizza l'errore quadratico medio, allora

$$X^\top X w^* = X^\top y$$

Nella pratica, w^* è calcolato con metodi numerici di fattorizzazione (Singular Value Decomposition – SVD), più stabili rispetto alle equazioni normali e che non richiedono l'esistenza dell'inversa

Esempio: regressione di `sales` su `TV`



$$\texttt{sales} \approx w_0 + w_1 \cdot \texttt{TV}$$

- Intercetta $w_0 = 7.03 \Rightarrow$ 7030 unità di prodotto vendute senza investimenti
- Coefficiente $w_1 = 0.047 \Rightarrow$ 47 unità di prodotto in più ogni 1000\$ di pubblicità in TV

Come valutare la qualità del modello?

Qualità del fit (rischio empirico)
$\neq$
qualità del modello (rischio atteso)

Possiamo stimare il rischio atteso di un'ipotesi h utilizzando un altro insieme di esempi di test T (*test set*)

Se gli esempi in T provengono dalla distribuzione (ignota) $\mathcal{D}$, allora con sufficienti esempi, il rischio empirico su T sarà una buona stima del rischio atteso:

$$\mathrm{RE}_T(h) \approx \mathrm{RA}(h) \text{ se } T \text{ è grande}$$

Training set e test set

In pratica, avremo un solo insieme di dati a disposizione

Separiamo a caso i dati di esempio a nostra disposizione in due insiemi:

Training set e test set

- Il *training set* S è usato per trovare l'ipotesi h col miglior fit:

$$\underset{h \in \mathcal{H}}{\text{minimize}}\, \mathrm{RE}_S(h)$$

- Il *test set* T è usato per stimare il rischio atteso di h:

$$\mathrm{RA}(h) \approx \mathrm{RE}_T(h)$$

- La separazione tra S e T è necessaria affinché gli esempi usati per stimare $\mathrm{RA}(h)$ siano indipendenti da h
- La separazione deve essere casuale, affinché S e T seguano la stessa distribuzione
- Mai usare gli esempi di training per testare il modello!

Riepilogo

In ogni metodo di apprendimento che segue il principio ERM:

1. Si assume una classe di ipotesi $\mathcal{H}$
2. Si assume una funzione di costo ℓ
3. Dato un training set di m esempi, attraverso un algoritmo di ottimizzazione si sceglie $h \in \mathcal{H}$ in modo da minimizzare

$$\sum_{i=1}^{m} \ell(h, (x^{(i)}, y^{(i)}))$$

4. Il rischio atteso di h viene stimato attraverso un test set
5. L'ipotesi h viene utilizzata per le predizioni successive:
   - Per ogni nuovo input x', la predizione è $h(x')$