Introduzione a modelli e metodi di regressione

Vincenzo Bonifaci

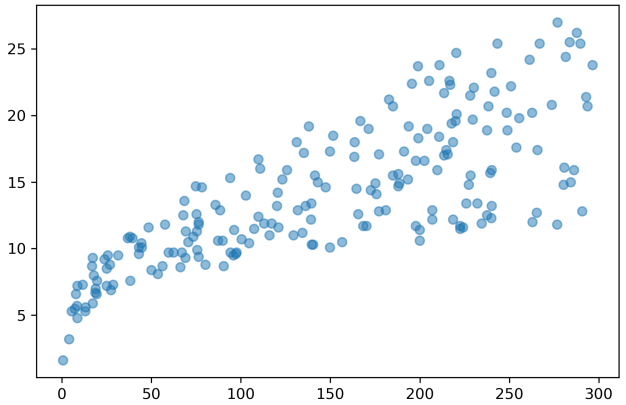
IN550 – Machine Learning

Esempio: Ritorno da investimenti pubblicitari

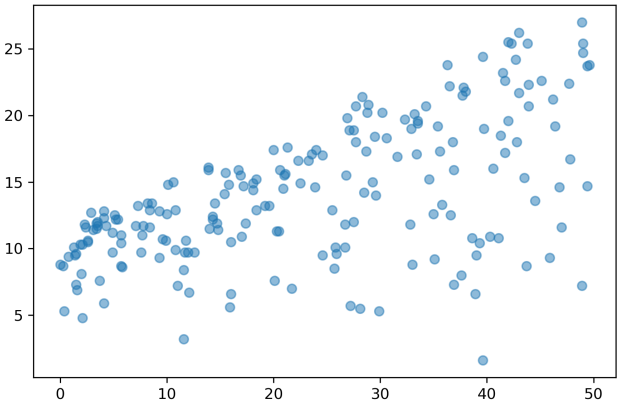
Input: investimenti pubblicitari via TV, radio e giornali in un mercato (*input*, *predittori*, *feature*, *variabili indipendenti*)
Output: unità di prodotto vendute in quel mercato (*output*, *responso*, *etichetta*, *variabile dipendente*)

	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
...	...	...	...	...

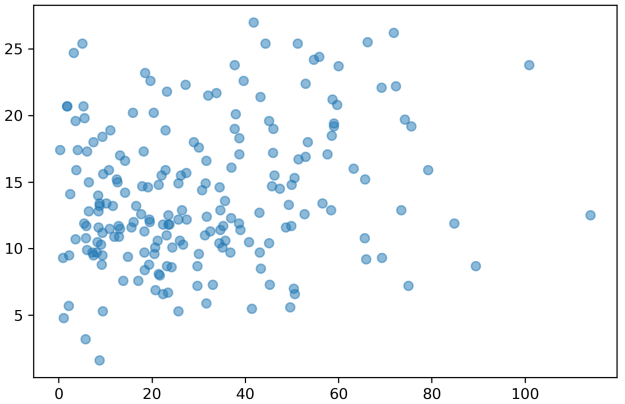
Esempio: Ritorno da investimenti pubblicitari



`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*)

Funzioni di costo [loss functions]

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 reale nonnegativo

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

Esempi di funzioni di costo

- *Quadrato dell'errore*:

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

- *Funzione costo 0-1*:

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

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 $\mathcal{X} \times \mathcal{Y}$

$\mathcal{D}$ è ignota poiché è proprio la relazione ingresso-uscita che l'algoritmo cerca di apprendere!

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

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

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

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:

$$\begin{aligned} &\underset{h}{\text{minimize}}\, L_{\mathcal{D}}(h) \\ \equiv &\underset{h}{\text{minimize}}\, \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h,(x,y))\right] \end{aligned}$$

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

Come possiamo 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

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

Possiamo usare il rischio empirico come *surrogato* del rischio atteso: se il numero di esempi m è sufficientemente grande, si può 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:

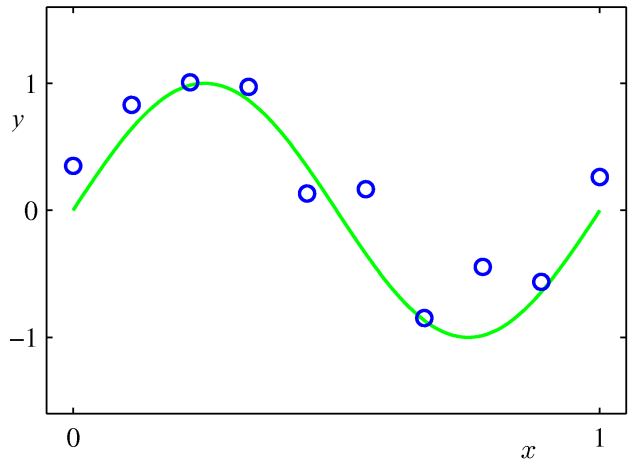
$$\underset{h}{\text{minimize}}\, L_S(h) \left(\equiv \underset{h}{\text{minimize}}\, \frac{1}{m} \sum_{i=1}^{m} \ell(h, (x^{(i)}, y^{(i)})) \right)$$

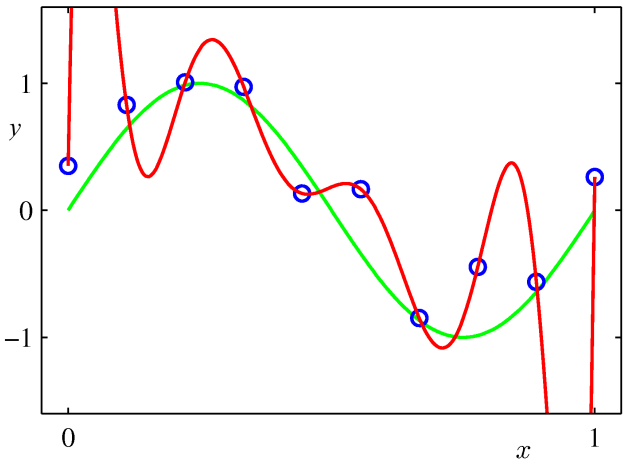
L'insieme S di esempi osservati dal learner è detto *training set*

Applicando l'ERM, il problema del learning supervisionato è rimpiazzato da un problema di ottimizzazione (nello spazio delle regole h)

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

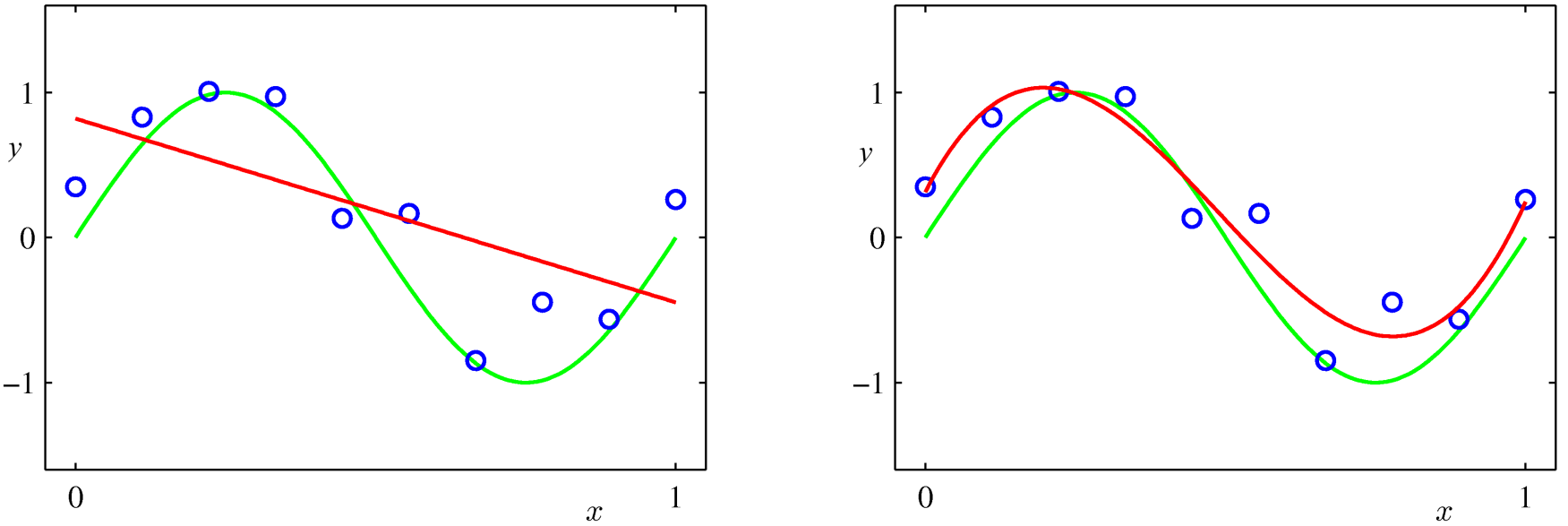
Anziché considerare la classe $\mathcal{Y}^{\mathcal{X}}$ di tutte le funzioni da $\mathcal{X}$ a $\mathcal{Y}$, consideriamo solo una sua sottoclasse $\mathcal{H}$ (*insieme delle ipotesi*)

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

$$\underset{h \in \mathcal{H}}{\text{minimize}}\, L_S(h)$$

- 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

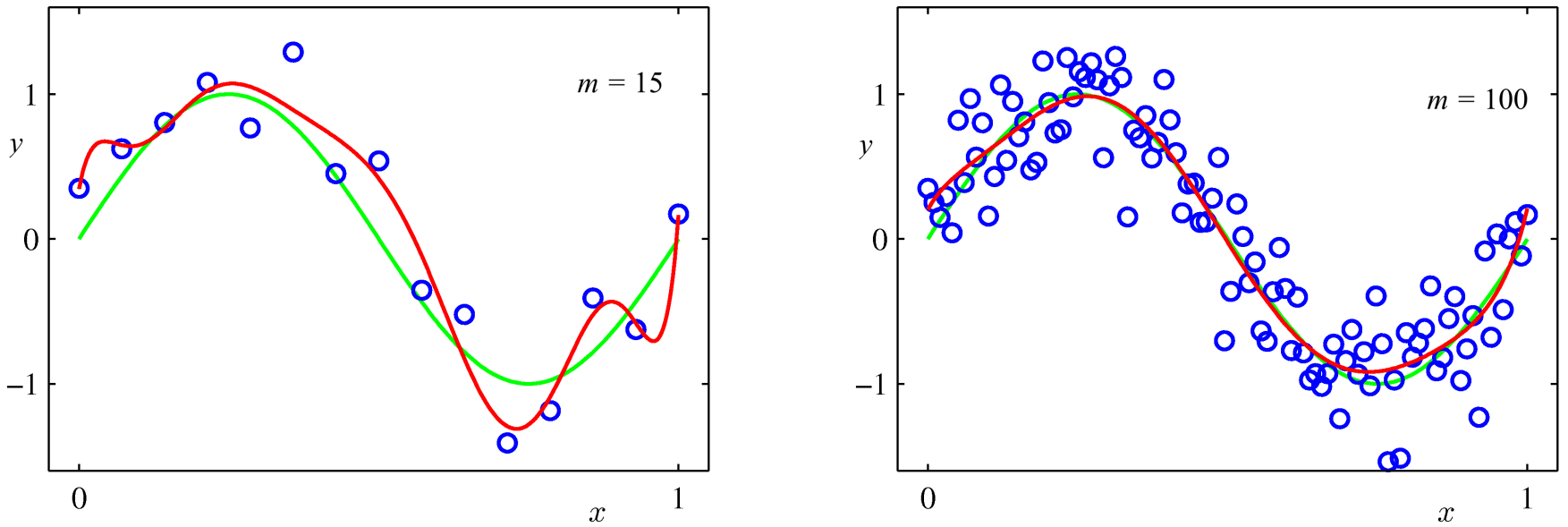
Compromesso bias-varianza



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

- Modelli più semplici hanno più bias (possono esibire *underfitting*)

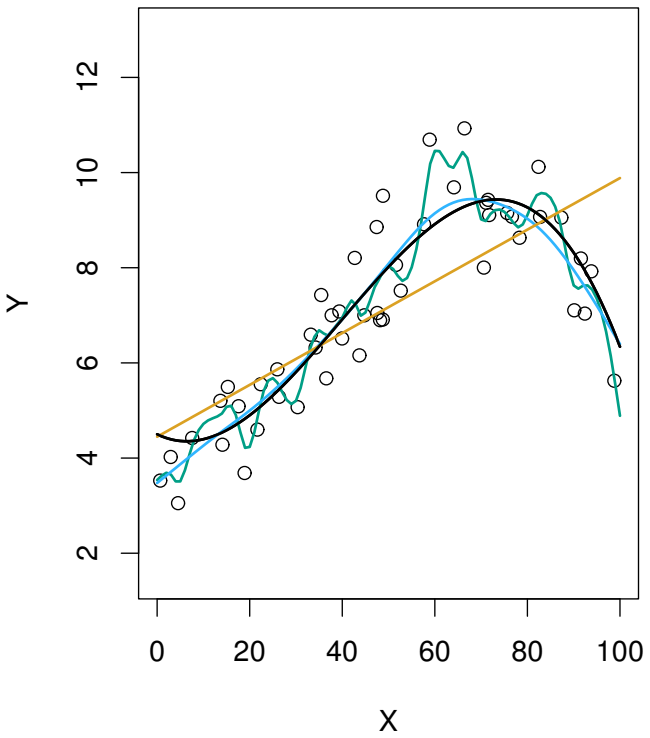
Compromesso bias-varianza

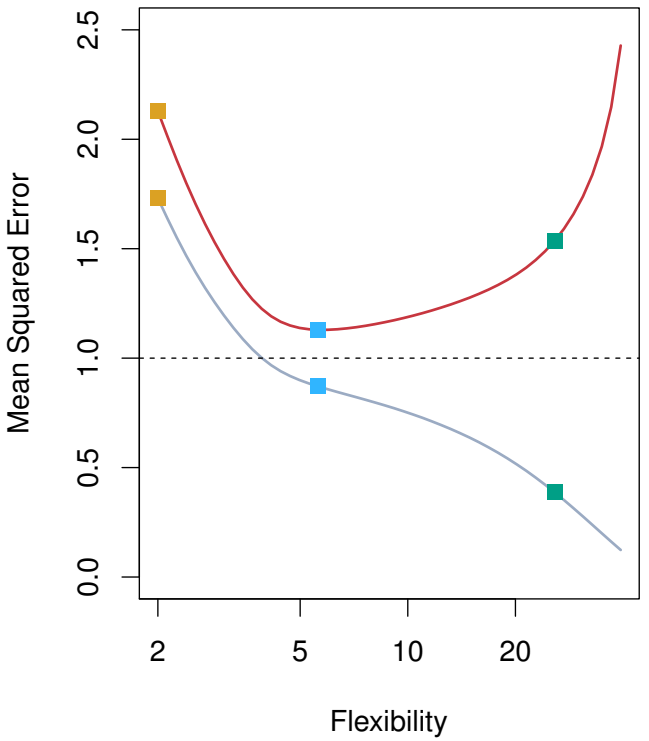


Fitting di un polinomio di grado 9 con 15 esempi (sinistra) e con 100 esempi (destra)

- Modelli più complessi hanno più varianza (richiedono più esempi)

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
  La curva rossa rappresenta il rischio atteso

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 $\mathcal{Y} \equiv \mathbb{R}$:

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

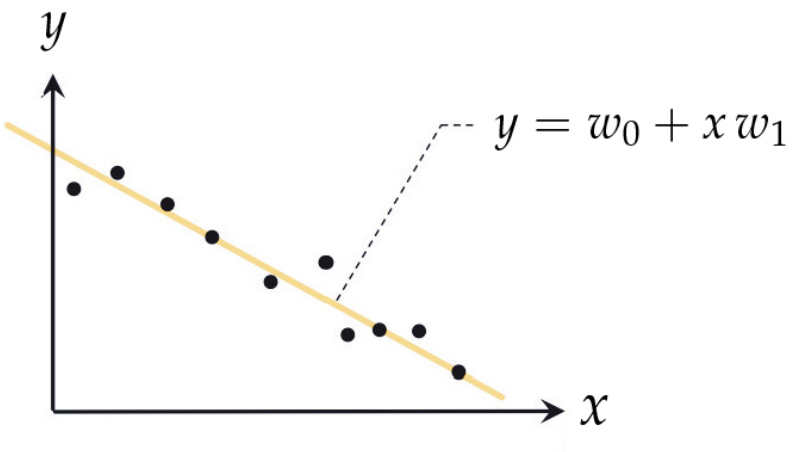
Useremo spesso la convenzione $x_0 \stackrel{\text{def}}{=} 1$, così da poter scrivere $h(x) = w^\top x$

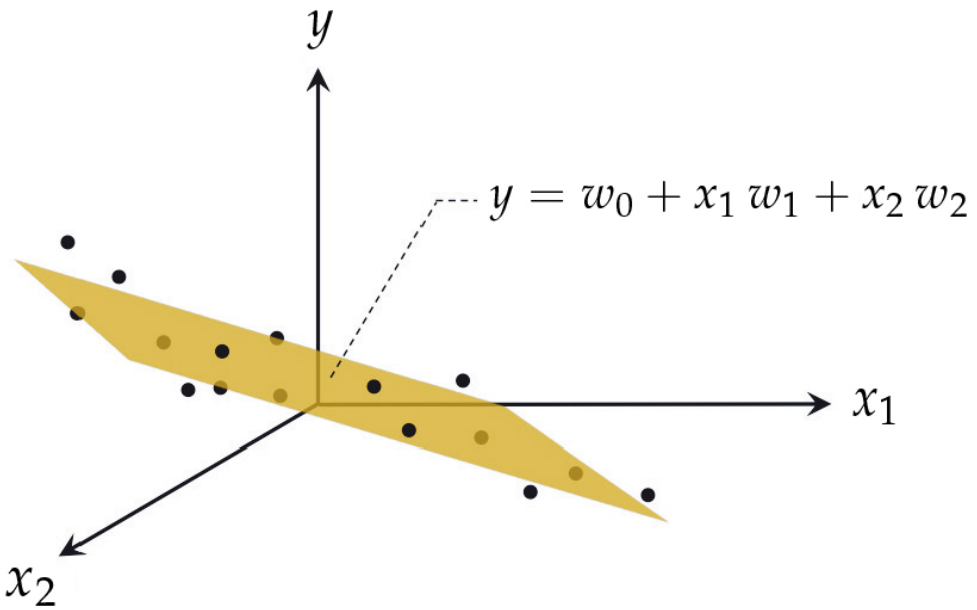
- w_0 è l'*intercetta* (valore previsto dal modello quando x è nullo)
- w_k è il *coefficiente* che esprime la dipendenza di $h(x)$ dalla k-esima componente di x

Una funzione di costo comunemente utilizzata è quella quadratica:

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

Regressione lineare



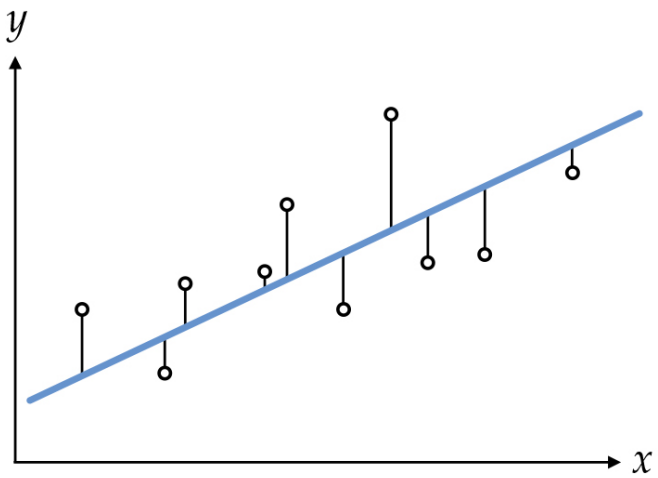


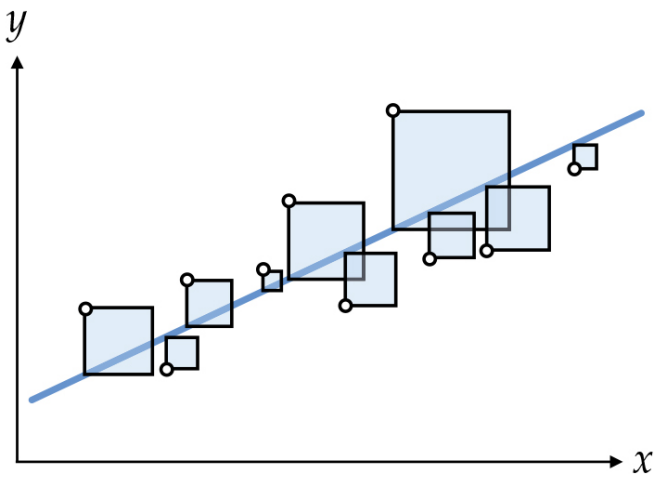
ERM per la regressione lineare

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

Mean Squared Error (MSE)

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





ERM per la regressione lineare

Il minimizzatore del rischio empirico qui è esprimibile in forma chiusa:

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

in quanto (come vedremo) deve soddisfare le cosiddette equazioni normali:

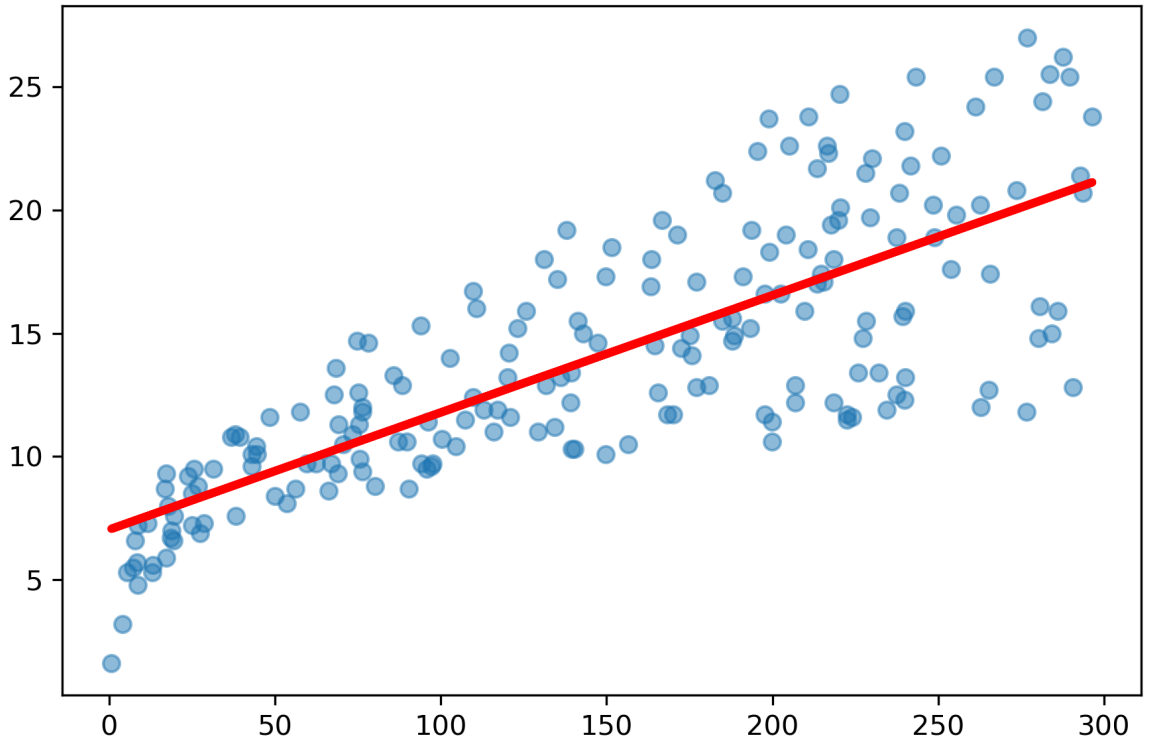
Equazioni normali

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

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

Nella pratica, w^* è calcolata con metodi numerici di fattorizzazione (Singular Value Decomposition – SVD), più stabili rispetto alla formula 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 fit?

Si può usare il rischio empirico (in questo caso: errore quadratico medio)

Nella regressione lineare, si può anche usare la statistica R^2:

Coefficiente R^2

$$R^2 \stackrel{\text{def}}{=} 1 - \frac{\bar{L}_S}{L_S^{(0)}},$$

- $\bar{L}_S = \min_{h \in \mathcal{H}} L_S(h)$ è l'errore quadratico medio della migliore ipotesi lineare
- $L_S^{(0)}$ è l'errore quadratico medio dell'ipotesi costante $h_0(x) = \frac{1}{m} \sum_i y_i$

Come valutare la qualità del fit?

R^2 è un valore tra 0 e 1 che rappresenta la proporzione di variabilità dei dati di output y "spiegata" dal modello

- R^2 coincide con il quadrato del coefficiente di correlazione tra il responso osservato (y) e quello previsto dal modello ($h(x)$)
- Rispetto al rischio empirico, R^2 ha il vantaggio di essere normalizzato
- R^2 è specifico per la funzione costo quadratica

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 insieme di esempi di test T (*test set*)

Gli esempi in T provengono ancora dalla distribuzione (ignota) $\mathcal{D}$

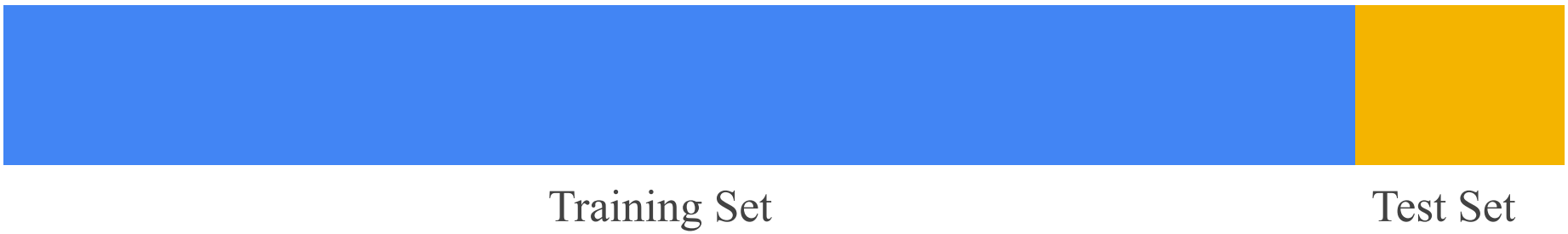
Con sufficienti esempi, il rischio empirico su T sarà una buona stima del rischio atteso:

$$L_T(h) \approx L_{\mathcal{D}}(h)$$

Training set e test set

Nella 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:

$$\underset{h \in \mathcal{H}}{\text{minimize}}\, L_S(h)$$

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

$$L_T(h) \approx L_{\mathcal{D}}(h)$$

  - La separazione è necessaria affinché gli esempi usati per stimare $L_{\mathcal{D}}(h)$ siano indipendenti da h
  - La separazione deve essere casuale, affinché S e T seguano la stessa distribuzione
  - Mai usare gli esempi di test per fare il training del modello!

L'ipotesi Bayesiana

L'*ipotesi Bayesiana* h^* è quella che tra tutte minimizza il rischio atteso:

$$h^* = \operatorname*{argmin}_{h \in \mathcal{Y}^{\mathcal{X}}} L_{\mathcal{D}}(h)$$

Poiché

$$L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell] = \mathbb{E}_x[\mathbb{E}_{y|x}[\ell|x]]$$

abbiamo che per ogni x, $h^*(x)$ minimizza $\mathbb{E}_{y|x}[\ell|x]$

Infatti se così non fosse, potremmo ridefinire $h^*(x)$, diminuendo $L_{\mathcal{D}}(h^*)$

$L_{\mathcal{D}}(h^*)$ è detto *rischio bayesiano*

Ipotesi Bayesiana: esempio

Nel caso della regressione con costo quadratico, $\ell = (h(x) - y)^2$, quindi:

$$\begin{aligned}\mathbb{E}_{y|x}[(h-y)^2|x] &= \mathbb{E}[h^2 - 2hy + y^2|x] \\ &= h^2 - 2h\,\mathbb{E}[y|x] + \mathbb{E}[y^2|x] \\ &= h^2 - 2h\,\mathbb{E}[y|x] + (\mathbb{E}[y|x])^2 + \mathrm{Var}[y|x] \\ &= (h - \mathbb{E}[y|x])^2 + \mathrm{Var}[y|x]\end{aligned}$$

Il secondo termine non dipende da h
Il primo termine è minimizzato se $h(x) = \mathbb{E}[y|x]$

Ipotesi Bayesiana per la regressione con costo quadratico

$$h^*(x) = \mathbb{E}[y|x]$$

Decomposizione bias-varianza nella regressione

Teorema (Decomposizione bias-varianza del rischio atteso)

Se $\ell(h, (x, y)) = (h(x) - y)^2$, vale la decomposizione

$$L_{\mathcal{D}}(h) = (\text{bias}_h)^2 + \text{varianza}_h + \text{rischio bayesiano}$$

dove:

- $\text{bias}_h \stackrel{\text{def}}{=} \mathbb{E}[h(x) - h^*(x)]$
- $\text{varianza}_h \stackrel{\text{def}}{=} \text{Var}[h(x)] = \mathbb{E}[(h(x) - \mathbb{E}[h(x)])^2]$
- $\text{rischio bayesiano} \stackrel{\text{def}}{=} L_{\mathcal{D}}(h^*) = \mathbb{E}[(h^*(x) - y)^2]$
- $h^*(x) = \mathbb{E}[y|x]$

Decomposizione del rischio atteso in generale

Teorema (Decomposizione stima-approssimazione del rischio atteso)

Sia $\mathcal{H}$ una qualunque classe di ipotesi.
Se $h \in \mathcal{H}$, $\bar{h} = \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$, $h^* = \operatorname{argmin}_{h \in \mathcal{Y}^{\mathcal{X}}} L_{\mathcal{D}}(h)$, allora

$$L_{\mathcal{D}}(h) = (L_{\mathcal{D}}(h) - L_{\mathcal{D}}(\bar{h})) + (L_{\mathcal{D}}(\bar{h}) - L_{\mathcal{D}}(h^*)) + L_{\mathcal{D}}(h^*)$$

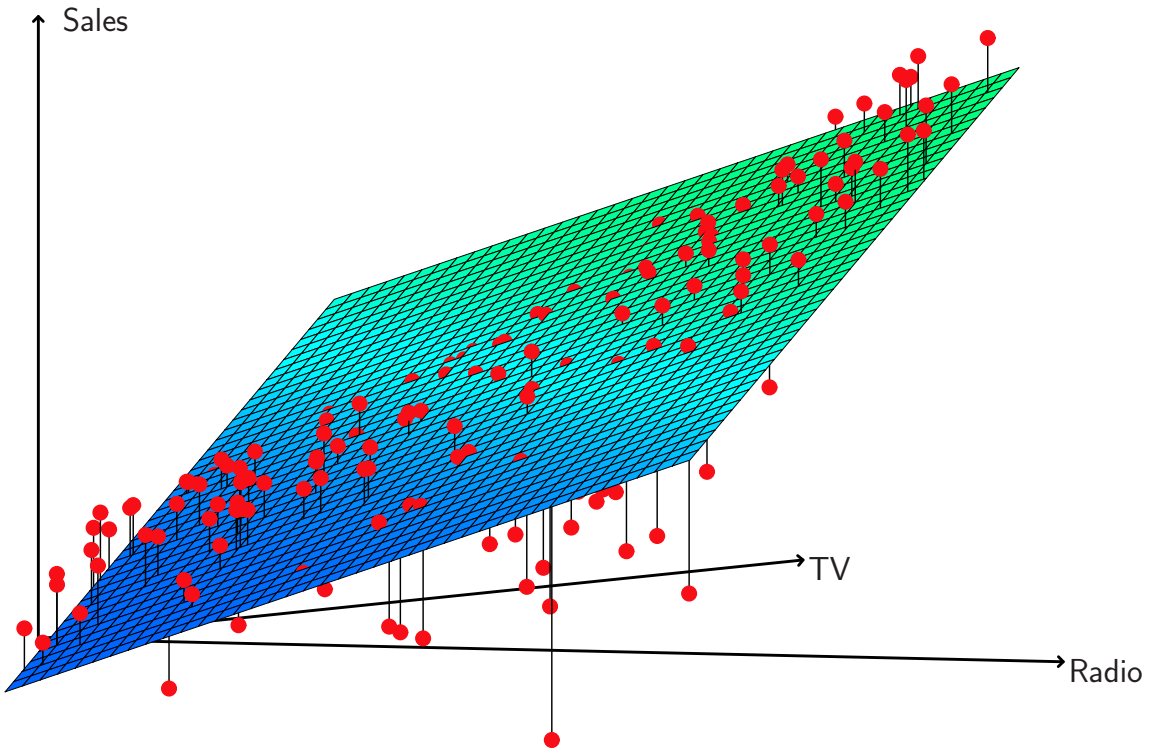
dove:

- $L_{\mathcal{D}}(h) - L_{\mathcal{D}}(\bar{h}) \geq 0$ (*errore di stima*)
- $L_{\mathcal{D}}(\bar{h}) - L_{\mathcal{D}}(h^*) \geq 0$ (*errore di approssimazione*)
- $L_{\mathcal{D}}(h^*) \geq 0$ (*rischio bayesiano*)

NB. La decomposizione bias-varianza non è un caso particolare della decomposizione stima-approssimazione, anche se qualitativamente simile.

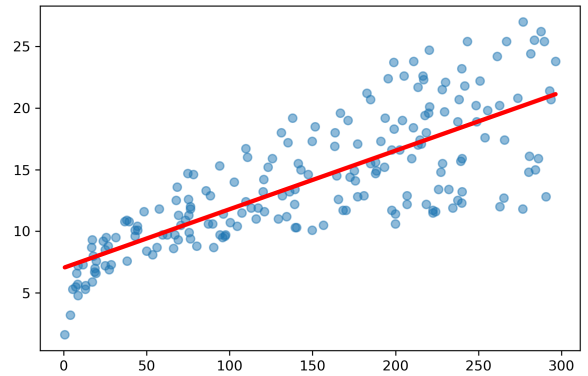
Regressione lineare multipla

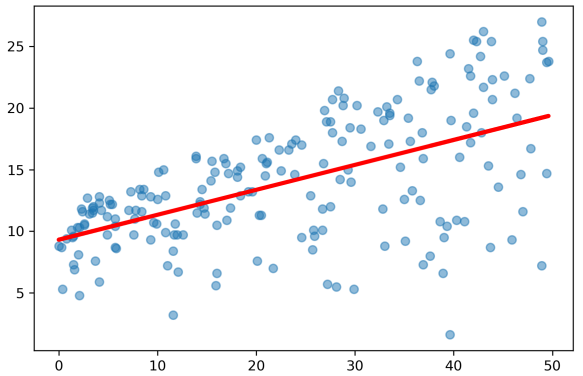
Come dipendono le vendite dagli investimenti in `TV` e `radio`?

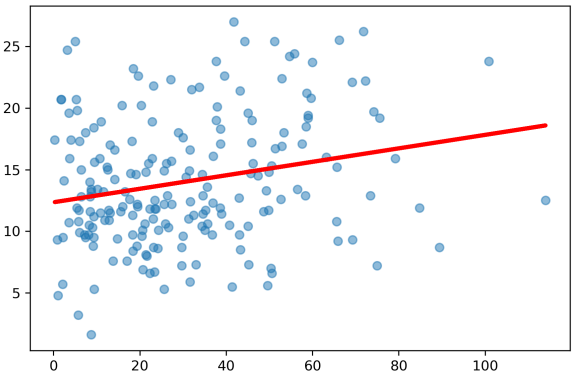


$$\texttt{sales} \approx w_0 + w_1 \cdot \texttt{TV} + w_2 \cdot \texttt{radio}$$

Regressione lineare semplice vs. multipla







Variabili utilizzate	R^2_{train}	MSE_{train}	MSE_{test}
`TV`	58.8%	10.6	10.2
`radio`	35.6%	16.6	24.2
`newspaper`	6.4%	24.1	32.1
`TV, radio, newspaper`	90.7%	2.4	4.4
`TV, radio`	90.7%	2.4	4.4

Il problema di individuare le variabili più rilevanti è detto *feature selection*

Variabili qualitative

Finora abbiamo assunto che tutti gli input siano quantitativi

Come trattare input di tipo *qualitativo*?

Es.: Se vogliamo stimare il reddito di un dipendente, potremmo avere a disposizione un dato sul sesso del dipendente (maschio/femmina)

Possiamo definire la variabile

$$x_{\text{sesso}}^{(i)} = \begin{cases} 1 & \text{se il dipendente } i\text{-esimo è femmina} \\ 0 & \text{se il dipendente } i\text{-esimo è maschio} \end{cases}$$

Il coefficiente relativo a questa variabile indicherà la dipendenza del reddito dal sesso (differenza media di reddito tra dipendenti femmine e maschi)

One-Hot Encoding

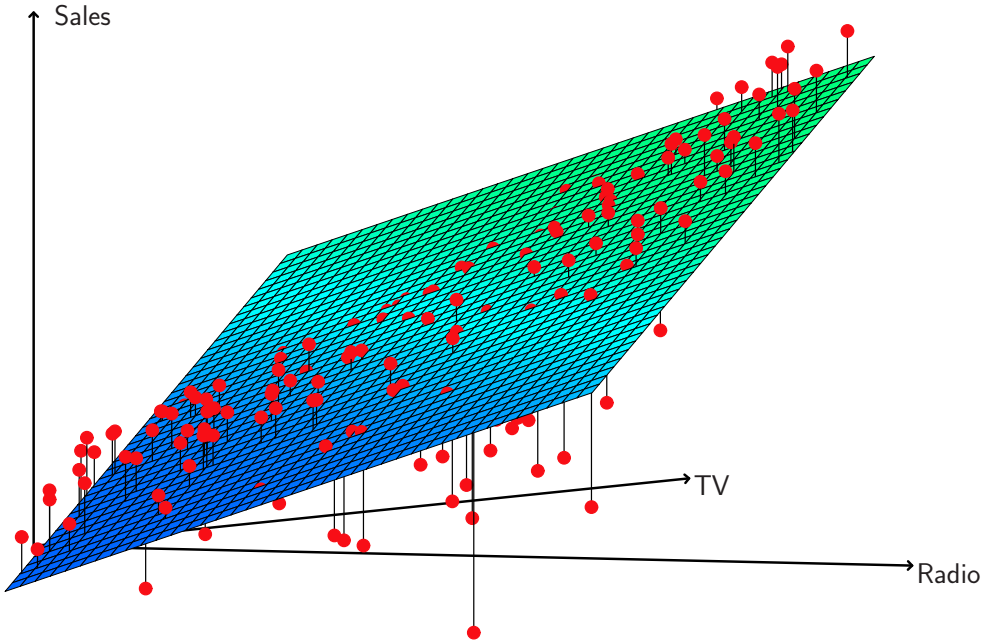
Se i valori possibili sono $K > 2$, non è corretto rappresentarli con una singola variabile, ma possiamo creare K variabili binarie

Esempio: `dieta` $\in$ {vegetariana, vegana, onnivora}

$$
\begin{aligned}
(\ldots 1\ 0\ 0 \ldots) &\text{ vegetariana} \\
(\ldots 0\ 1\ 0 \ldots) &\text{ vegana} \\
(\ldots 0\ 0\ 1 \ldots) &\text{ onnivora}
\end{aligned}
$$

Questo schema è detto *one-hot encoding*

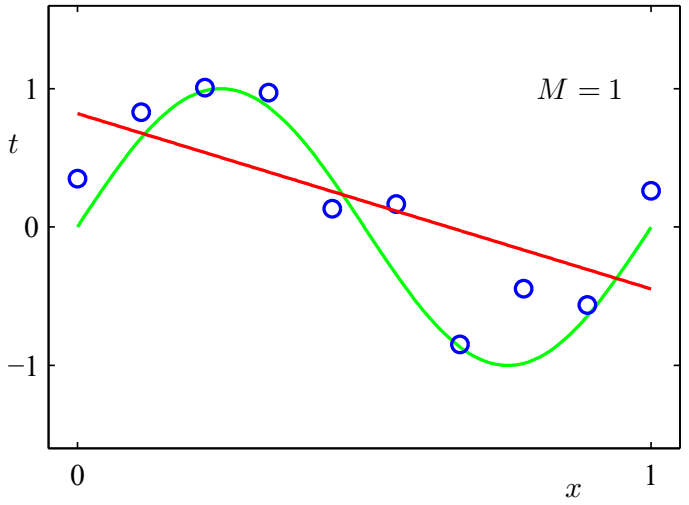
Modellare interazioni tra le variabili (*feature crossing*)



C'è una sinergia tra gli investimenti in `TV` e `radio`?
Proviamo a includere una *variabile sintetica*: `TV` $\times$ `radio`

Variabili utilizzate	R^2_{train}	MSE_{train}	MSE_{test}
`TV, radio, TV` $\times$ `radio`	97.3%	0.7	1.6

Regressione polinomiale (unidimensionale)



Per alcuni problemi, sembrano preferibili regole di predizione non-lineari

La classe dei *regressori polinomiali* di grado n è

$$\mathcal{H}^n_{poly} = \{x \mapsto h(x)\}$$

dove h è un polinomio di grado n: $h(x) = w_0 + w_1 x + \ldots + w_n x^n$

Regressione polinomiale (unidimensionale)

Definiamo la funzione $\phi : \mathbb{R} \to \mathbb{R}^{n+1}$

$$\phi(x) = (1, x, x^2, \ldots, x^n)$$

$$h(x) = w_0 + w_1 x + \ldots + w_n x^n = w^\top \phi(x)$$

è ora una funzione lineare di w e dell'input "espanso" $\phi(x)$

Quindi il vettore w può essere determinato con una regressione lineare, usando gli input espansi $\phi(x)$

Regressione lineare generalizzata

In effetti possiamo usare un qualunque vettore di nuove feature $\phi : \mathbb{R}^d \to \mathbb{R}^{n+1}$ definite a partire dall'input x, per esempio

$$\phi(x) = (1,\, x_2^3,\, \sin x_1,\, \sqrt{|x_3 - x_4|})$$

con ipotesi della forma

$$h(x) = w_0\phi_0 + w_1\phi_1 + \ldots w_n\phi_n = w^\top \phi$$

Il punto cruciale è che l'ipotesi è ancora lineare rispetto al parametro w (ovviamente non lo è più rispetto all'input x)

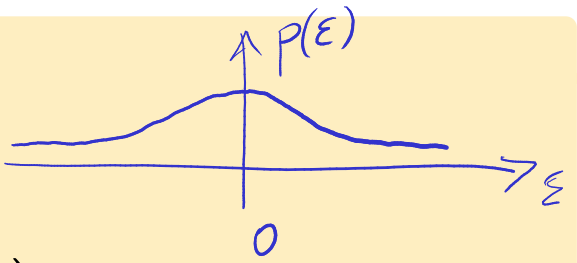
Regressione lineare: interpretazione probabilistica

La funzione costo quadratica può essere giustificata su basi probabilistiche

Consideriamo la seguente assunzione sul processo che genera i dati:

$$y^{(i)} = w^\top x^{(i)} + \epsilon^{(i)}, \qquad i = 1, \ldots, m$$

dove ogni $\epsilon^{(i)}$ è un termine di rumore con distribuzione gaussiana con media nulla e varianza σ^2:



$$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$

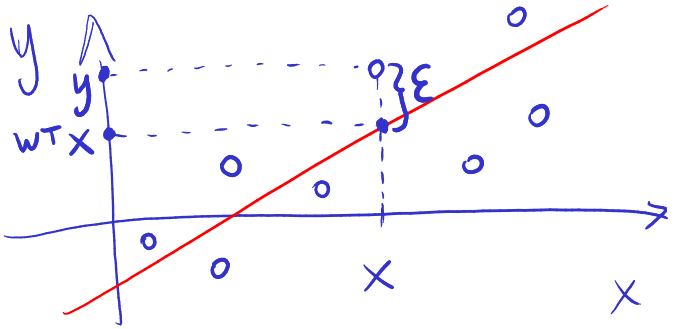
$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

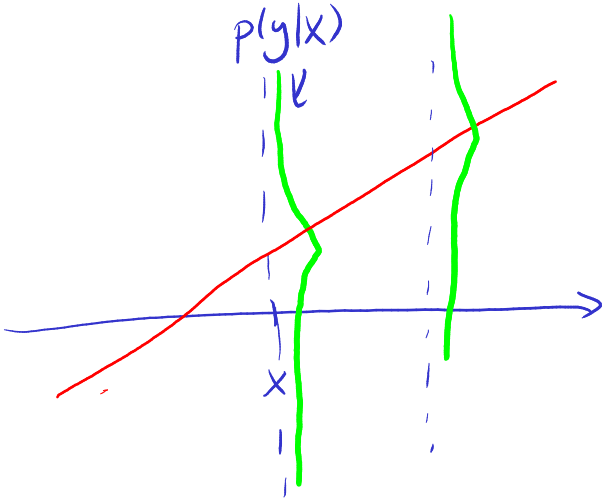
Regressione lineare: interpretazione probabilistica

Poiché $\epsilon^{(i)} = y^{(i)} - w^\top x^{(i)}$, ne consegue

$$p(y^{(i)}|x^{(i)}; w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - w^\top x^{(i)})^2}{2\sigma^2}\right)$$

In altri termini, $(y^{(i)}|x^{(i)}; w) \sim \mathcal{N}(w^\top x^{(i)}, \sigma^2)$





Regressione lineare: interpretazione probabilistica

La *verosimiglianza* [*likelihood*] del parametro w è

Funzione di verosimiglianza condizionata

$$\mathcal{L}_{\text{condizionata}}(w) \stackrel{\text{def}}{=} p(y^{(1)}, \ldots, y^{(m)} | x^{(1)}, \ldots x^{(m)}; w)$$

(probabilità di quegli output fissati gli input quando il parametro è w)

Maximum Likelihood Estimation (MLE)

Una metodologia consolidata per la stima del parametro consiste nel selezionare il parametro che massimizza la funzione verosimiglianza:

$$\underset{w}{\text{maximize}}\, \mathcal{L}(w)$$

Regressione lineare: interpretazione probabilistica

Nel caso della regressione lineare,

$$\mathcal{L}(w) = p(y^{(1)}, \ldots, y^{(m)} | x^{(1)}, \ldots x^{(m)}; w)$$

gli esempi sono indipendenti

$$= \prod_{i=1}^{m} p(y^{(i)} | x^{(i)}; w)$$

assunzione probabilistica

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - w^\top x^{(i)})^2}{2\sigma^2}\right)$$

la massimizzazione è equivalente

$$\log \mathcal{L}(w) = m \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{m} \frac{1}{2\sigma^2}(y^{(i)} - w^\top x^{(i)})^2$$

che è massimizzata quando w minimizza $\sum_{i=1}^{m}(y^{(i)} - w^\top x^{(i)})^2$

Regressione lineare: interpretazione probabilistica

In altre parole, nella regressione lineare,

- il principio ERM con costo quadratico, e
- il principio MLE con assunzione di rumore gaussiano

selezionano lo stesso parametro w^* (e quindi la stessa ipotesi h_{w^*})

Altre assunzioni sul rumore possono portare ad altre funzioni costo

Senza il principio ERM: Regressione non parametrica

Gli approcci visti finora sono *parametrici*: le ipotesi sono rappresentabili con un numero prefissato di parametri (per es. $w_0, w_1, \ldots, w_d$), scelti secondo il principio ERM

Nei metodi *non parametrici* le ipotesi non sono rappresentabili con un numero prefissato di parametri

- Sono più flessibili (minore bias)
- Richiedono più esempi (maggiore varianza)

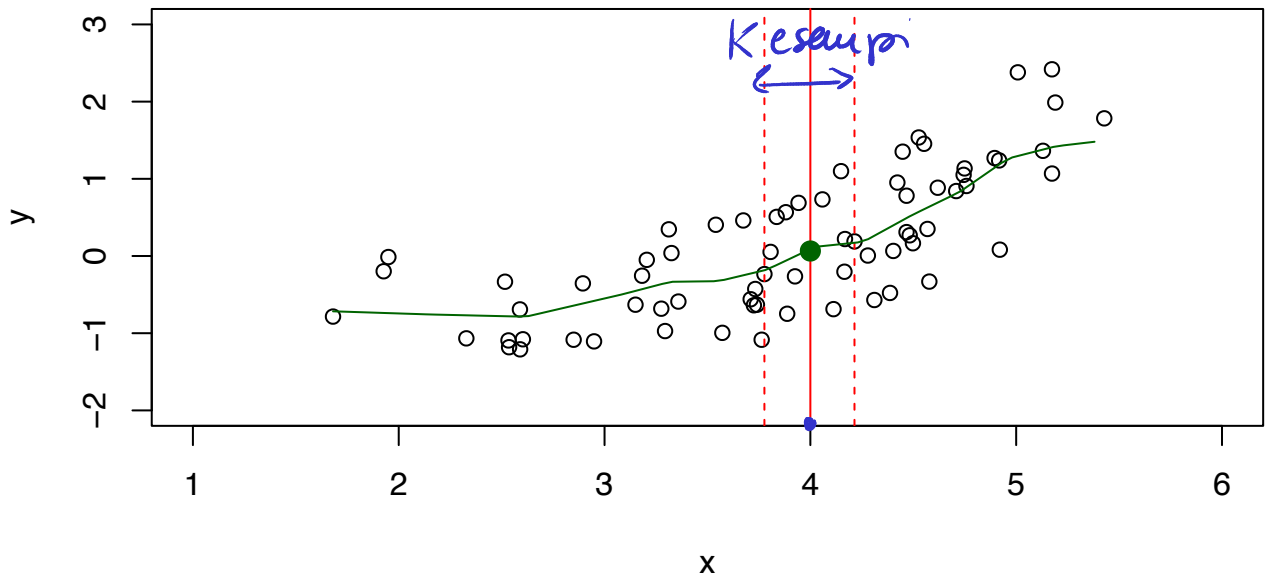
In generale, non si conformano al principio ERM ma si appoggiano direttamente alle osservazioni
(*instance-based learning* o *memory-based learning*)

Regressione K-Nearest Neighbor (K-NN)

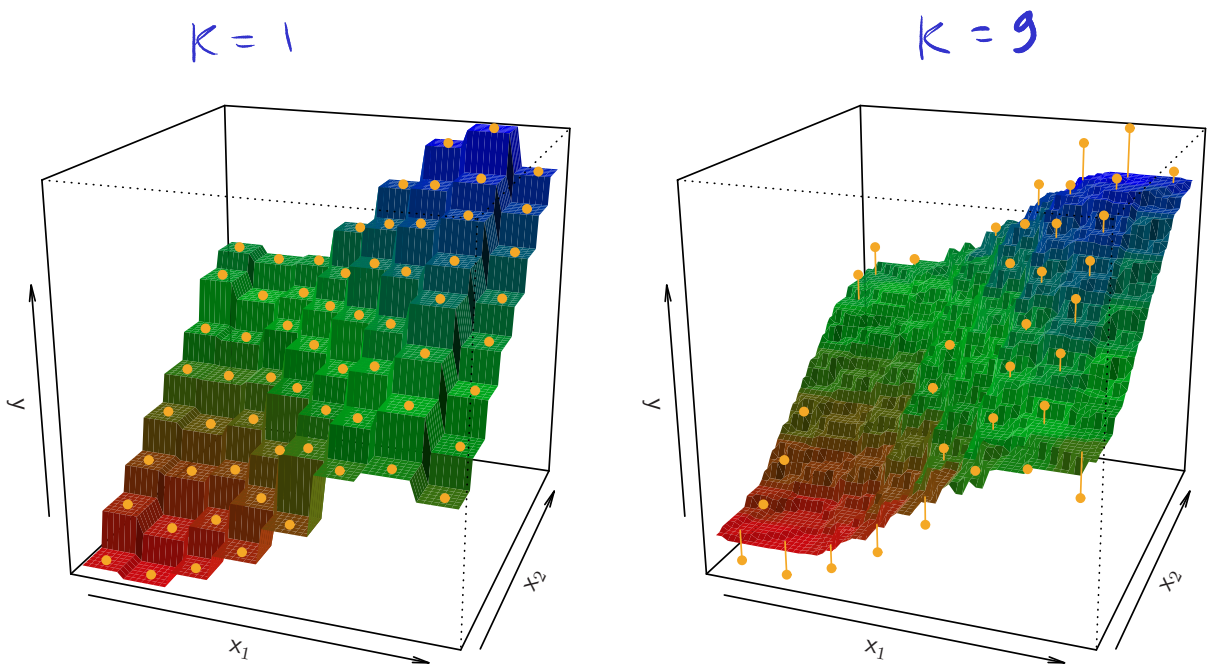
Regressione K-Nearest Neighbor (K-NN)

Sia $K \geq 1$ e sia $x \in \mathbb{R}^{d+1}$ il punto di cui si vuole stimare il responso $h(x)$

1. Identifica i K esempi $x^{(1)}, \ldots, x^{(K)}$ più vicini ad x (in termini di distanza euclidea, o altra funzione distanza)
2. Restituisci la media del responso su quegli esempi: $h(x) = \frac{1}{K} \sum_{i=1}^{K} y^{(i)}$



Regressione K-NN: Esempio

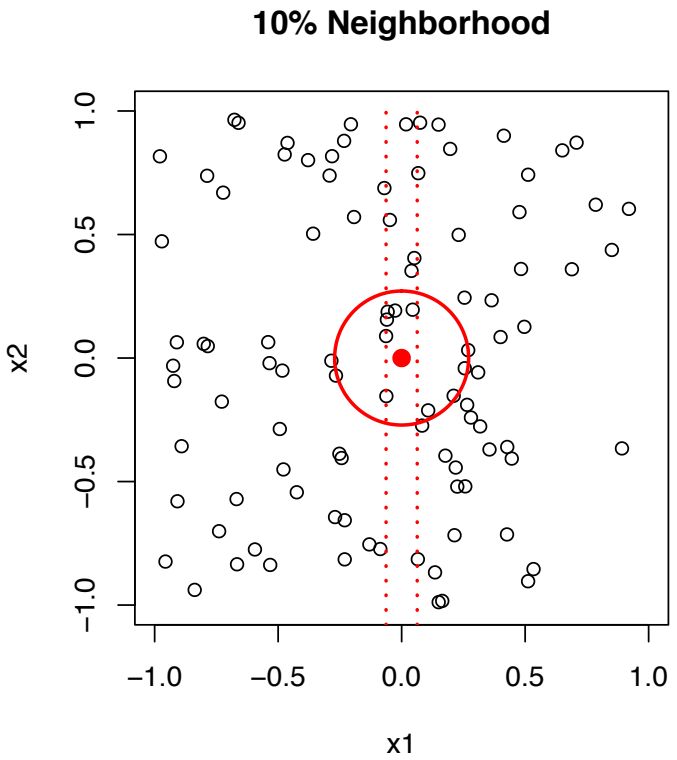


Regressione K-NN su un dataset bidimensionale di 64 osservazioni (punti arancioni)
Sinistra: $K = 1$, destra: $K = 9$

Regressione K-NN: Considerazioni

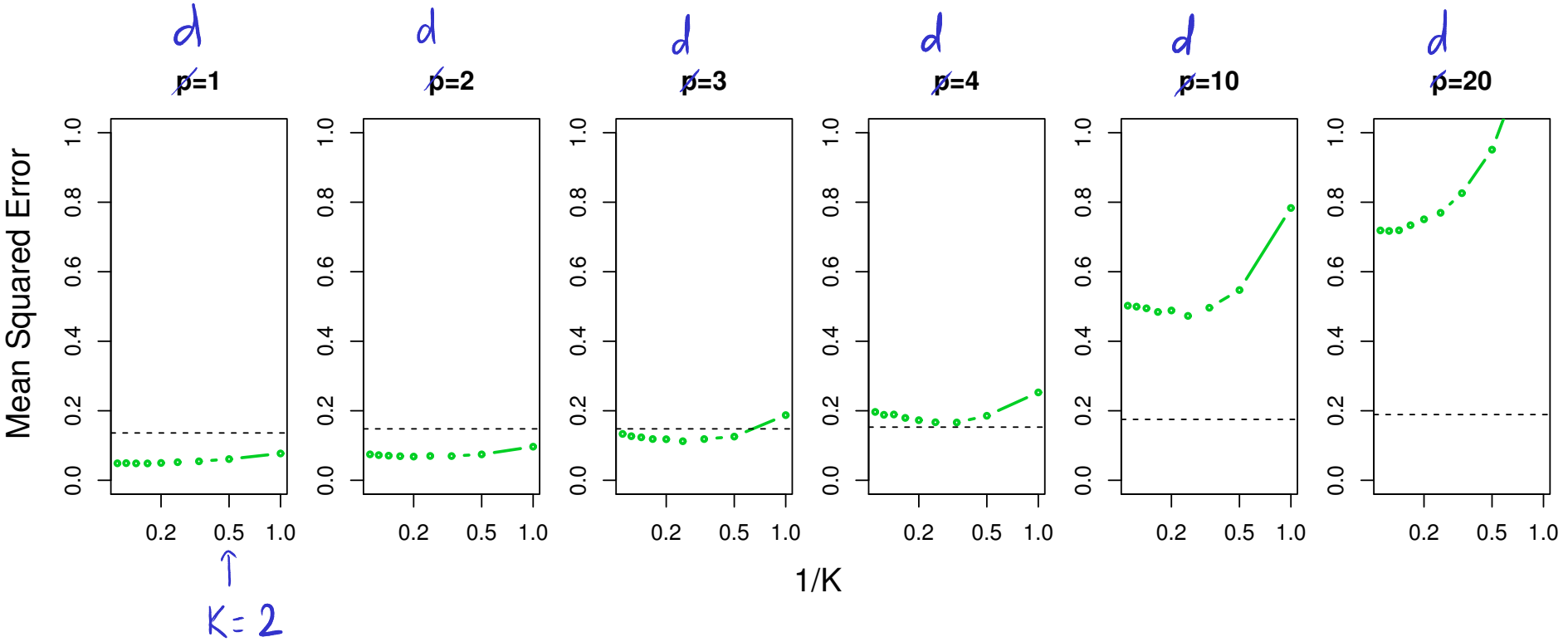
- Il metodo NN richiede accesso a tutti gli esempi ogni volta che effettua una predizione
- Tende ad essere efficace per d piccolo (ad esempio, $d \leq 4$) e m relativamente grande
- Può dare risultati scarsi per d grande: in molte dimensioni, i K punti più vicini possono essere molto lontani

Valutare
$\|x - x'\|$
richiede tempo
$O(d)$
$(x, x' \in \mathbb{R}^{d+1})$



Valuta
$\|x - x^{(1)}\|$
$\|x - x^{(2)}\|$
$\vdots$
$\|x - x^{(m)}\|$
↓
Totale $O(m \cdot d)$ operazioni

Regressione K-NN vs. regressione lineare



MSE di test per una regressione lineare (linea tratteggiata nera) vs. quello di una regressione K-NN (curva verde) per una distribuzione non-lineare in 1 variabile e indipendente dalle altre $p - 1$ variabili

Tipologie di regressione viste finora

Nome	Forma delle ipotesi $h(x)$	Funzione costo $\ell(h, (x, y))$
Regressione lineare (semplice)	$w_0 + w_1 x$	$(h(x) - y)^2$
Regressione lineare (multipla)	$w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$	$(h(x) - y)^2$
Regressione lineare generalizzata	$w_0 + w_1 \phi_1(x) + \ldots + w_n \phi_n(x)$	$(h(x) - y)^2$
Regressione K-NN	nessuna (non segue il principio ERM, a meno che $K = 1$)†	$(h(x) - y)^2$

†**Esercizio (non banale)**. Definire una classe di ipotesi $\mathcal{H}$ tale che la soluzione del problema ERM su $\mathcal{H}$ sia la stessa trovata dall'algoritmo di regressione 1-NN.

Regressione con altre funzioni di costo

Come trattare funzioni di costo diverse da quella quadratica?

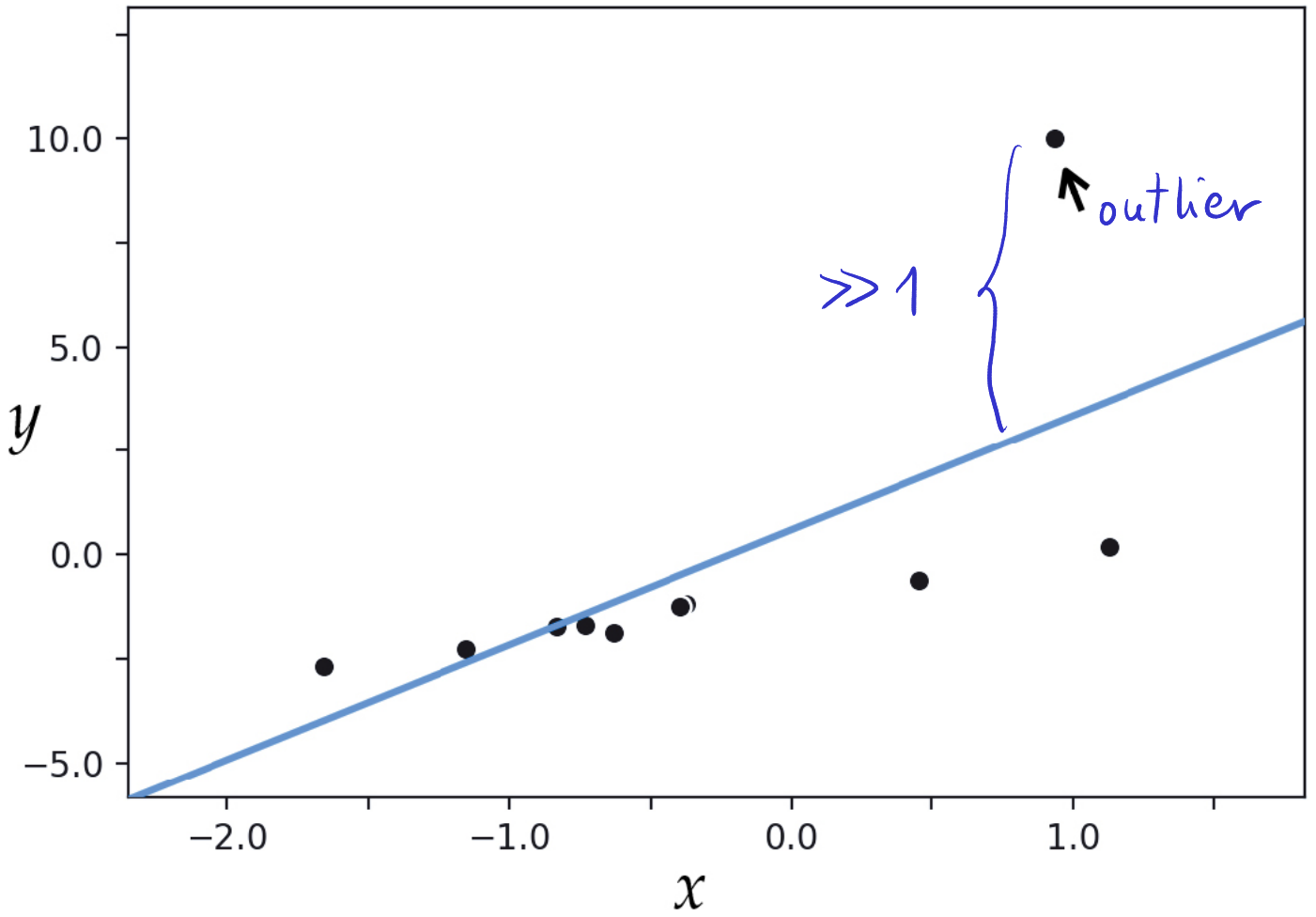
Per esempio, nella regressione *Least Absolute Deviation* (*LAD*),

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

Per una vasta classe di funzioni di costo (convesse e/o differenziabili) esiste una metodologia generale di ottimizzazione: la discesa del gradiente

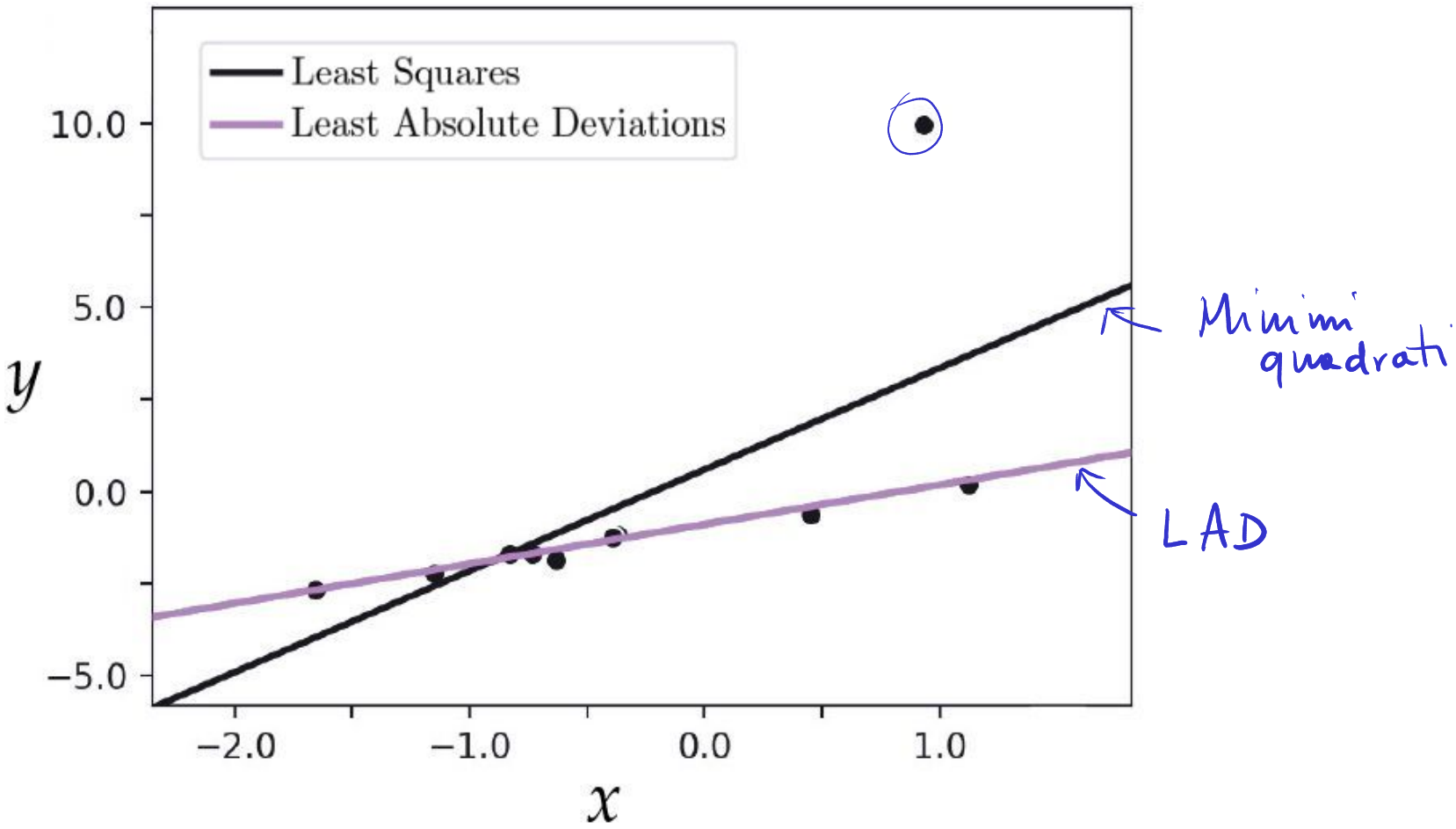
Rischio empirico: $\frac{1}{m} \sum_{i=1}^{m} |h(x^{(i)}) - y^{(i)}|$

Influenza di un esempio "anomalo" (*outlier*)



Il costo quadratico assegna grande importanza agli errori grandi ($\gg 1$)

Outlier: Metodo dei minimi quadrati vs. LAD



- Il costo LAD è più robusto rispetto agli outlier
- Svantaggio: l'ipotesi ottima LAD non è esprimibile in forma chiusa

Influenza di esempi duplicati nella regressione lineare

Supponiamo che l'esempio $(x^{(i)}, y^{(i)})$ compaia β_i volte

Il rischio empirico (in questo caso, l'MSE) diventa

$$L_S(h) = \frac{1}{\beta_1 + \ldots + \beta_m} \sum_{i=1}^{m} \beta_i (h(x^{(i)}) - y^{(i)})^2$$

$$m' = \beta_1 + \ldots + \beta_m$$

Esempi duplicati e regressione lineare pesata

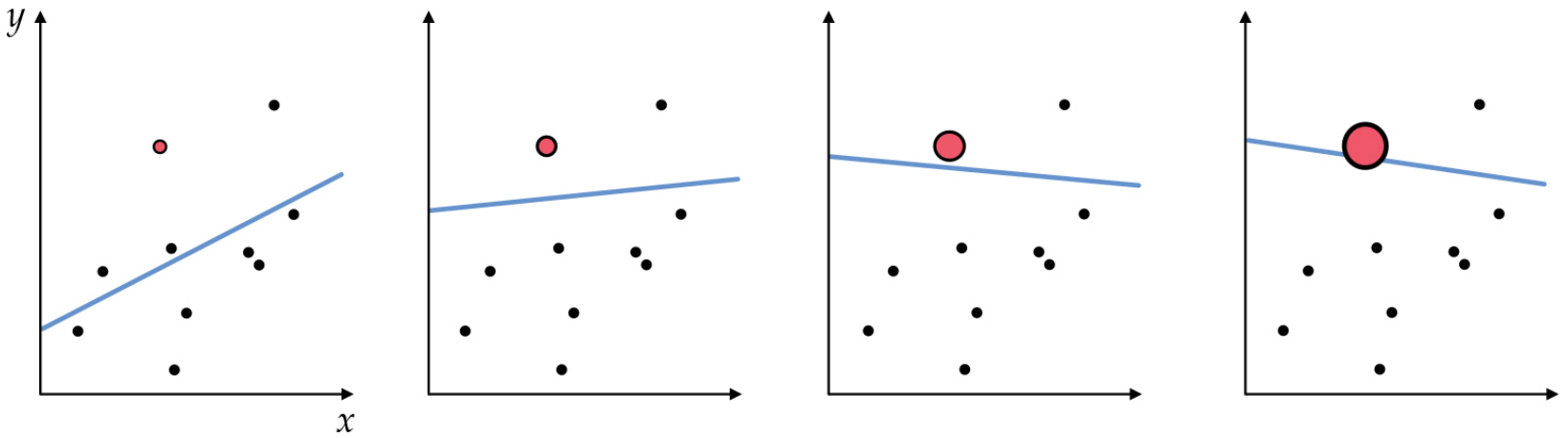
La minimizzazione del rischio empirico

$$L_S(h) = \frac{1}{\beta_1 + \ldots + \beta_m} \sum_{i=1}^{m} \beta_i (h(x^{(i)}) - y^{(i)})^2$$

$$= \frac{1}{\beta_1 + \ldots + \beta_m} \sum_{i=1}^{m} (h(\sqrt{\beta_i} x^{(i)}) - \sqrt{\beta_i} y^{(i)})^2$$

può essere interpretata come una *regressione lineare pesata*:

- L'esempio $(x^{(i)}, y^{(i)})$ è scalato di un fattore $\sqrt{\beta_i}$



Regressione multi-output

Finora abbiamo supposto output unidimensionali: $y^{(i)} \in \mathbb{R}^{1\times 1}$ ← $c = 1$

$h(x^{(i)}) = x^{(i)\top} w$ $= w^\top x^{(i)}$

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

Come gestire più variabili di output? Es. $y^{(i)} \in \mathbb{R}^{1\times c}$

Il vettore di parametri w diventa una matrice $W \in \mathbb{R}^{(d+1)\times c}$

$h(x^{(i)}) = x^{(i)\top} W$

$\ell(h, (x^{(i)}, y^{(i)})) = \left\| h(x^{(i)}) - y^{(i)}) \right\|^2$

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{c} (h(x^{(i)})_k - y_k^{(i)})^2$$

$$= \sum_{k=1}^{c} \left[\frac{1}{m} \sum_{i=1}^{m} (x^{(i)\top} w^{(k)} - y_k^{(i)})^2 \right]$$

Ciascuna colonna $w^{(k)}$ di W può essere ottimizzata indipendentemente

Il risultato è equivalente a c regressioni con output unidimensionale