Introduzione al Machine Learning: Modelli e metodi di regressione

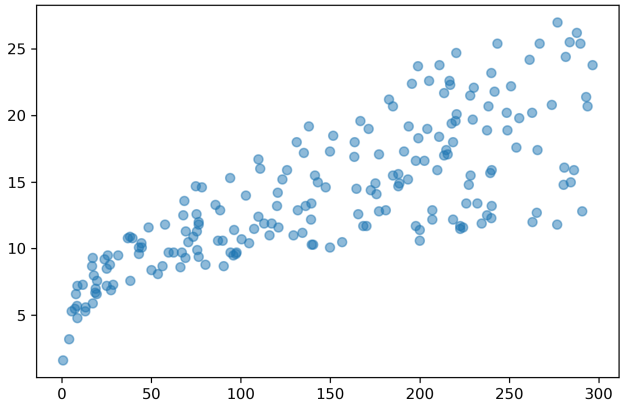
Vincenzo Bonifaci

Esempio: Ritorno da investimenti pubblicitari

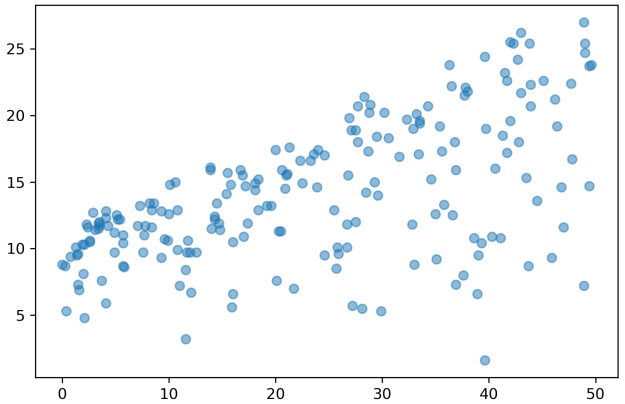
Input: investimenti pubblicitari via TV, radio e giornali in un mercato (in migliaia di Euro)
Output: unità di prodotto vendute in quel mercato (in migliaia)

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

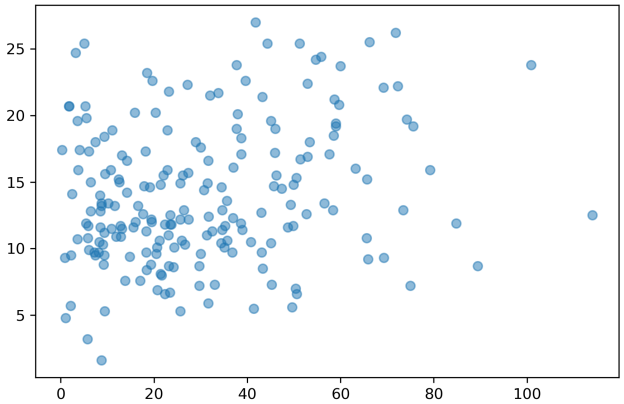
Esempio: Ritorno da investimenti pubblicitari



`sales` vs. `TV`



`sales` vs. `radio`



`sales` vs. `newspaper`

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

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

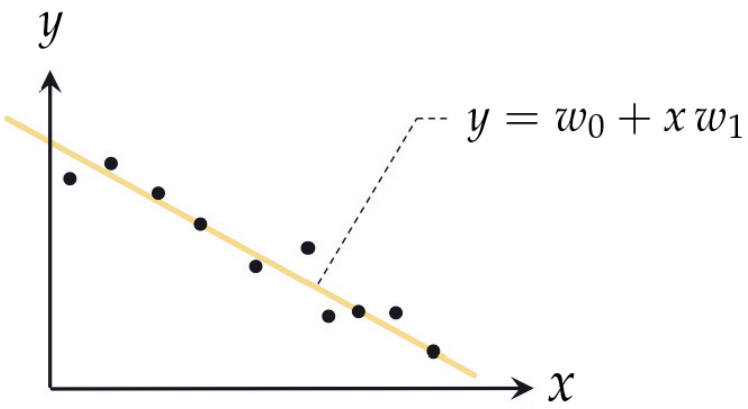
Useremo spesso la convenzione $x_0 \stackrel{\text{def}}{=} 1$, così da poter scrivere $h(x) = w^\top x$

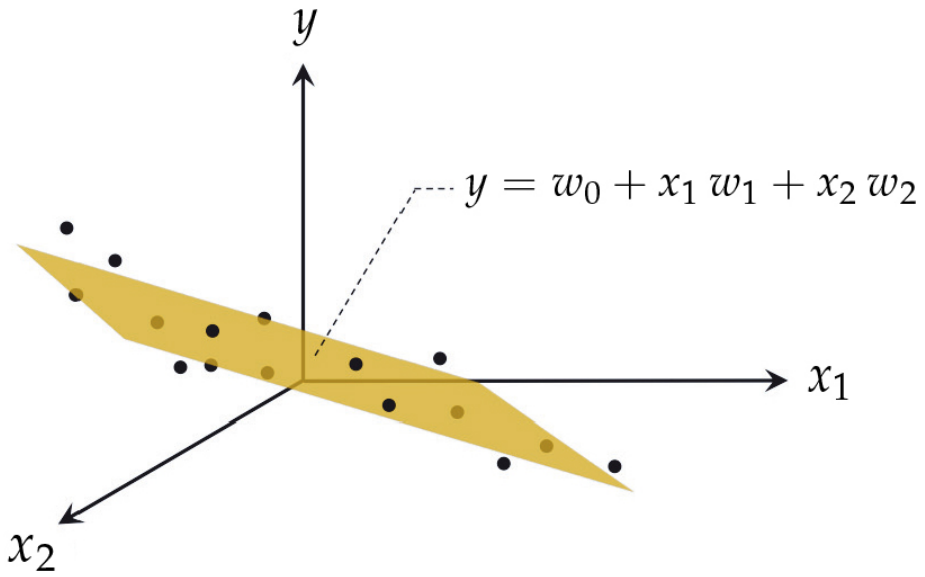
- w_0 è l'*intercetta* (valore previsto dal modello quando x è nullo)
- w_k è il *coefficiente* che esprime la dipendenza di $h(x)$ dalla k-esima componente di x

Una funzione di costo comunemente utilizzata è quella quadratica:

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

Regressione lineare



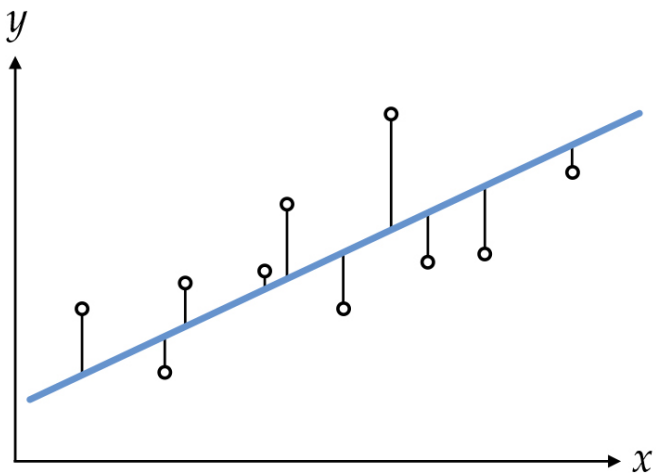


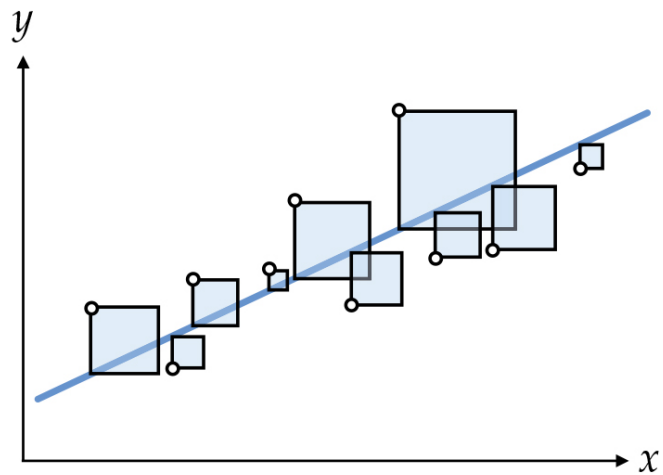
ERM per la regressione lineare

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

Minimizzare l'errore quadratico medio significa trovare il vettore $w \in \mathbb{R}^{d+1}$ che minimizza:

$$\|Xw - y\|^2$$

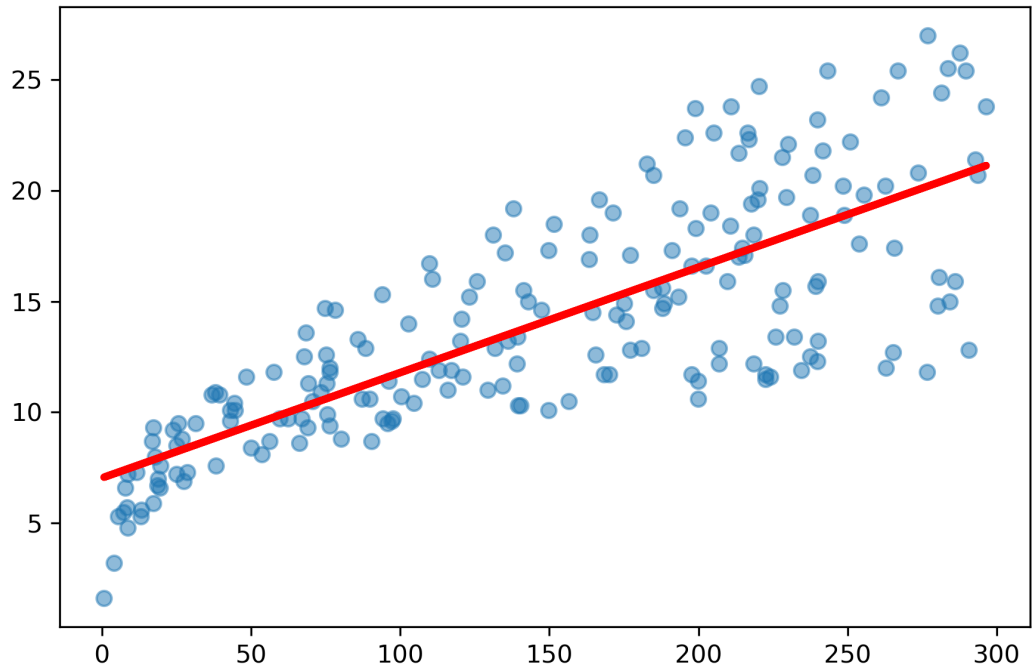
Equazioni normali

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

$$X^\top X w^* = X^\top y, \text{ quindi } w^* = (X^\top X)^{-1} y$$

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 fit?

In generale, si usa il rischio empirico (in questo caso: l'errore quadratico medio)

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

Coefficiente R^2

$$R^2 \stackrel{\text{def}}{=} 1 - \frac{\mathrm{RE}_*}{\mathrm{RE}_0},$$

- RE_* è l'errore quadratico medio della migliore ipotesi lineare $h(x) = w_0 + w_1 x_1 + \ldots + w_d x_d$ calcolata sul campione
- RE_0 è l'errore quadratico medio della migliore ipotesi costante $h(x) = w_0$, data da $h_0(x) = \frac{1}{m}\sum_i y_i$

Come valutare la qualità del fit?

Se calcolato sul campione, R^2 è un valore tra 0 e 1 ed è 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
- Se gli errori quadratici sono calcolati rispetto ad un campione diverso da quello usato per costruire la regressione, si parla di R^2 fuori campione
- Se R^2 è calcolato fuori campione, può essere negativo

Come valutare la qualità del modello?

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

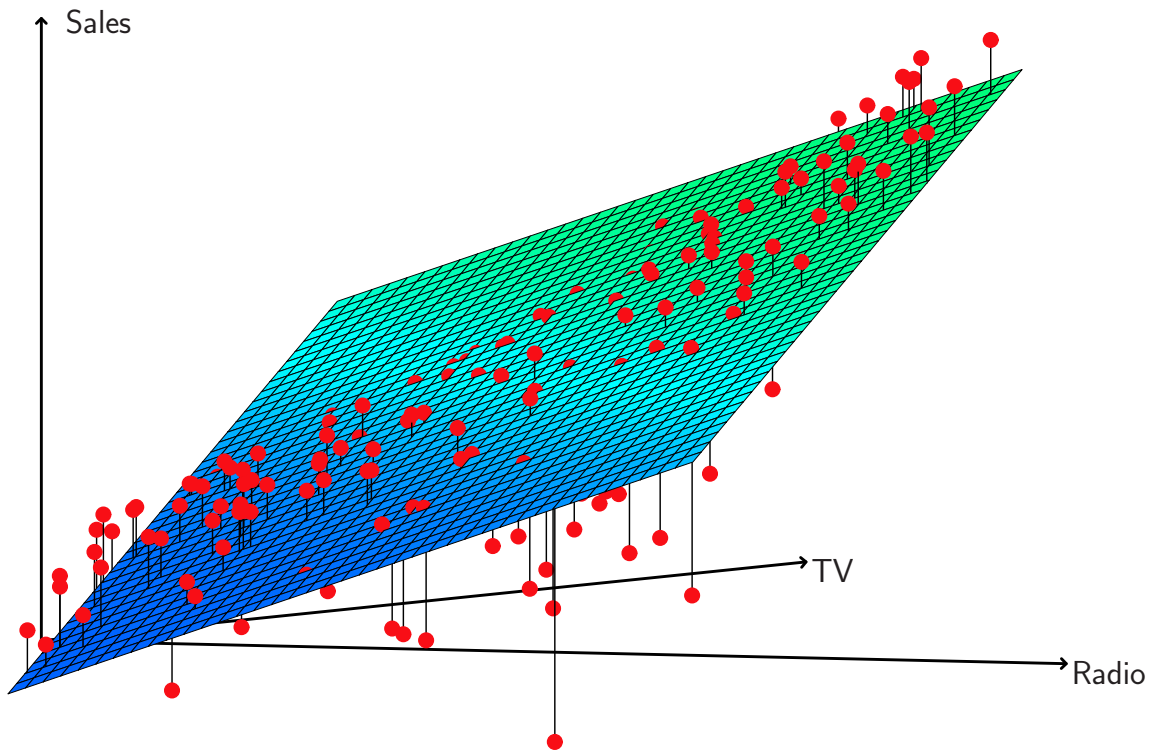
Possiamo stimare il rischio atteso di un'ipotesi h utilizzando un insieme di esempi di test T (*test set*)

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:

$$\mathrm{RE}_T(h) \approx \mathrm{RA}(h) \quad \text{per } T \text{ grande}$$

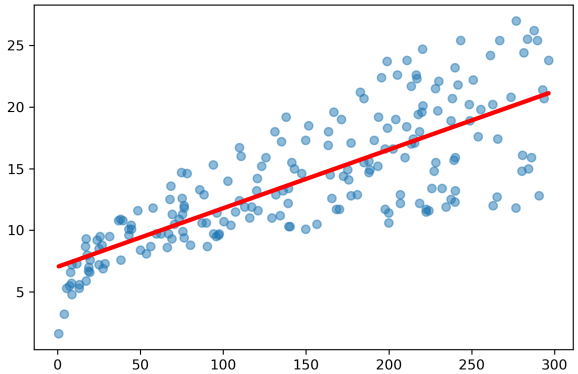
Regressione lineare multipla

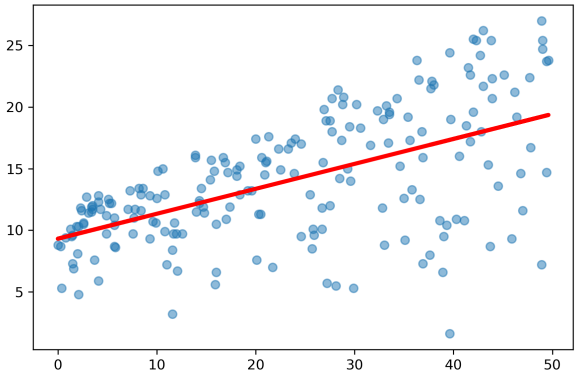
Come dipendono le vendite dagli investimenti in `TV` e `radio`?

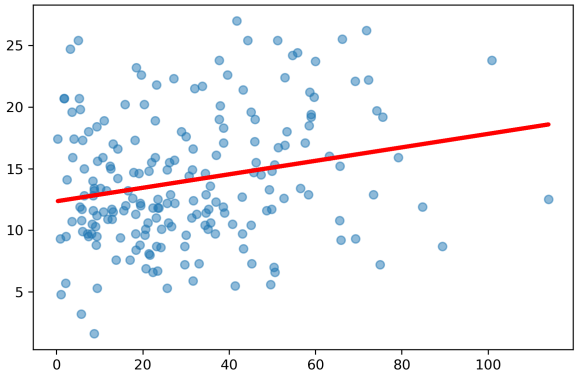


$$\texttt{sales} \approx w_0 + w_1 \cdot \texttt{TV} + w_2 \cdot \texttt{radio}$$

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 numerici

Come trattare input di tipo *categorici*?

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

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

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

One-hot encoding

Se le categorie possibili sono $K > 2$, non è corretto rappresentare il dato con una sola variabile, ma possiamo creare K variabili binarie

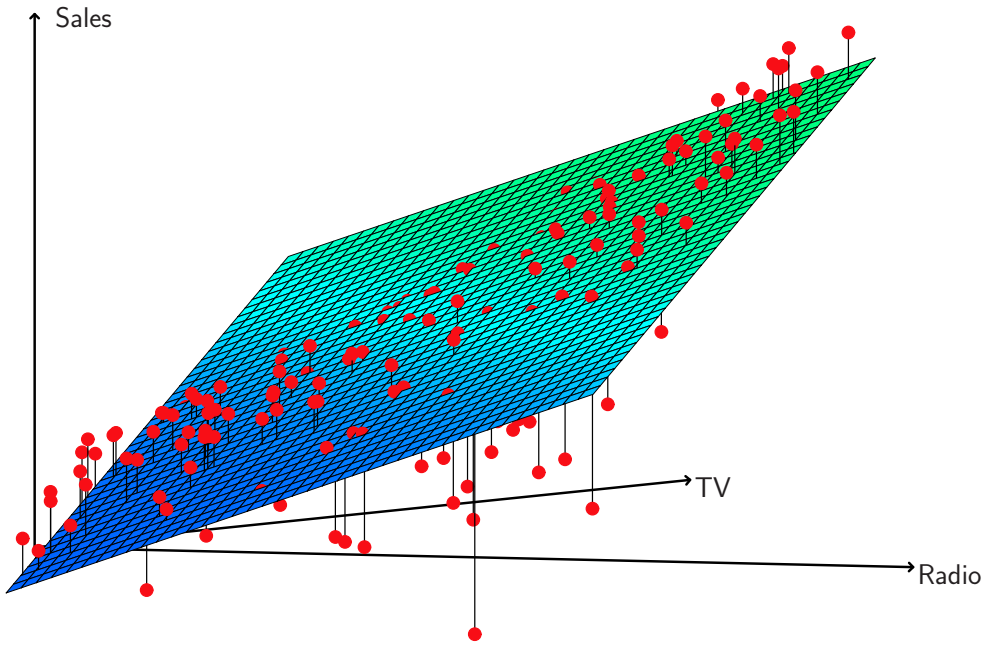
Esempio: `dieta` $\in$ {vegetariana, vegana, onnivora}

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

Questo schema è detto *one-hot encoding*

Modellare interazioni tra le variabili (*feature crossing*)

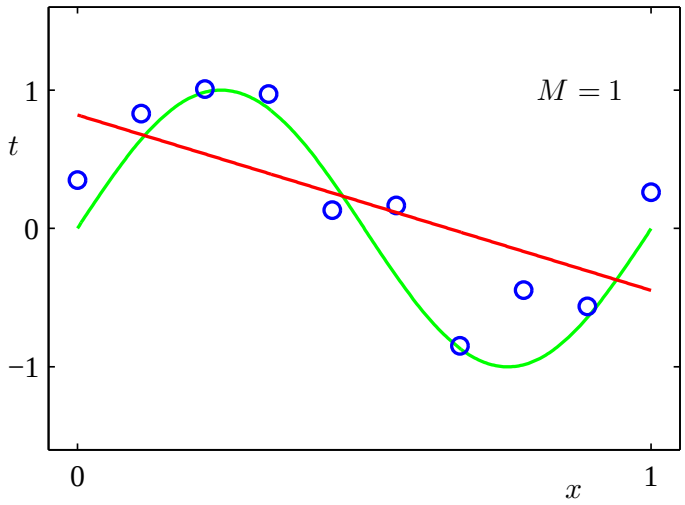
C'è una sinergia tra gli investimenti in `TV` e `radio`?



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 utili 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

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

L'ipotesi è ancora lineare rispetto al vettore dei parametri w (anche se non lo è più rispetto all'input x)

Possiamo ottimizzare w allo stesso modo, usando ϕ al posto di x

Senza il principio ERM: Regressione non parametrica

Gli approcci visti finora sono *parametrici*: le ipotesi sono rappresentabili con un numero prefissato di parametri (ad 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

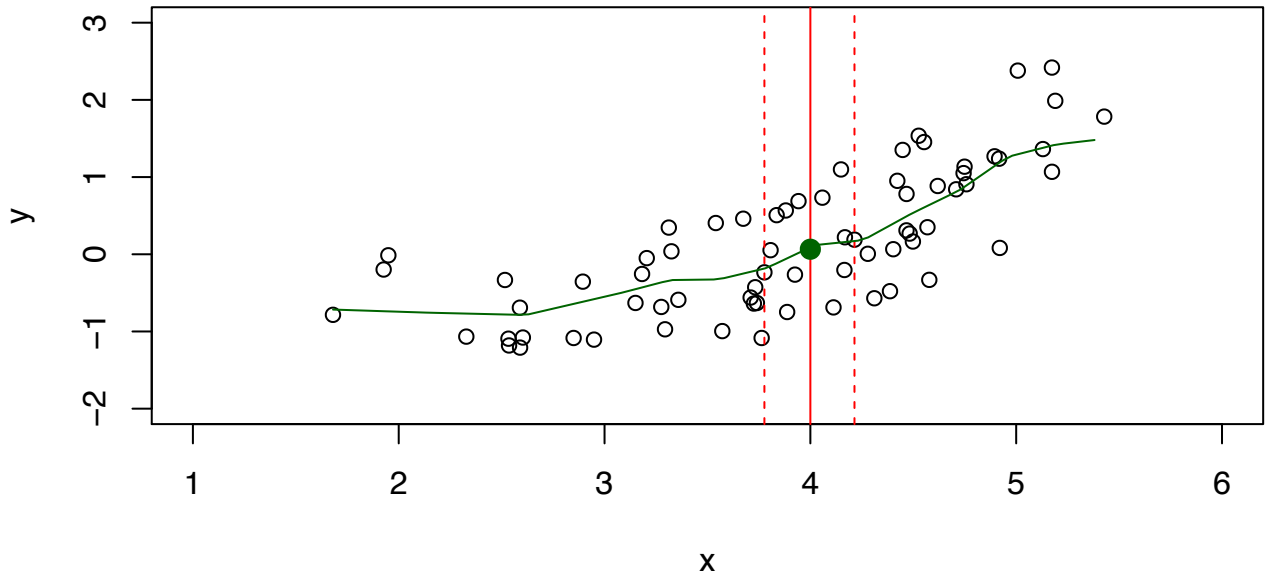
- Sono più flessibili (minore bias)
- Richiedono più esempi (maggiore varianza)

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

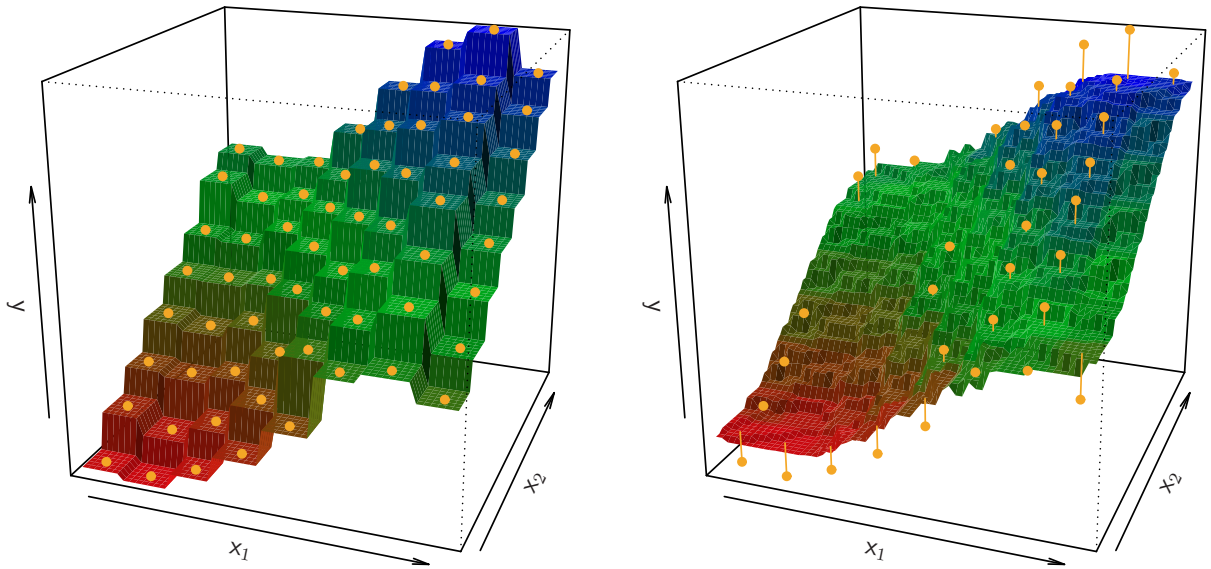
Regressione K-Nearest Neighbor (K-NN)

Sia $K \geq 1$ e sia x 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 quei K esempi: $h(x) = \frac{1}{K} \sum_{i=1}^{K} y^{(i)}$



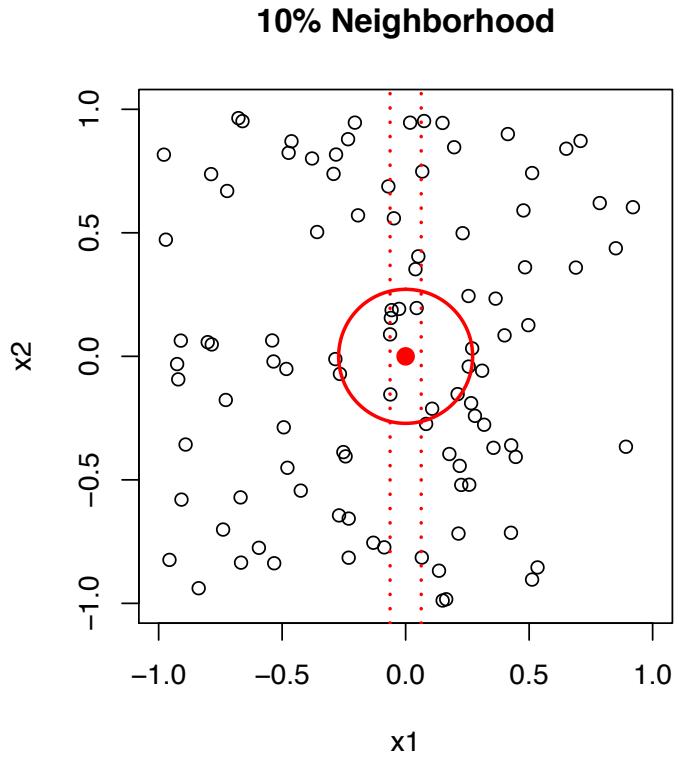
Regressione K-NN: Esempio



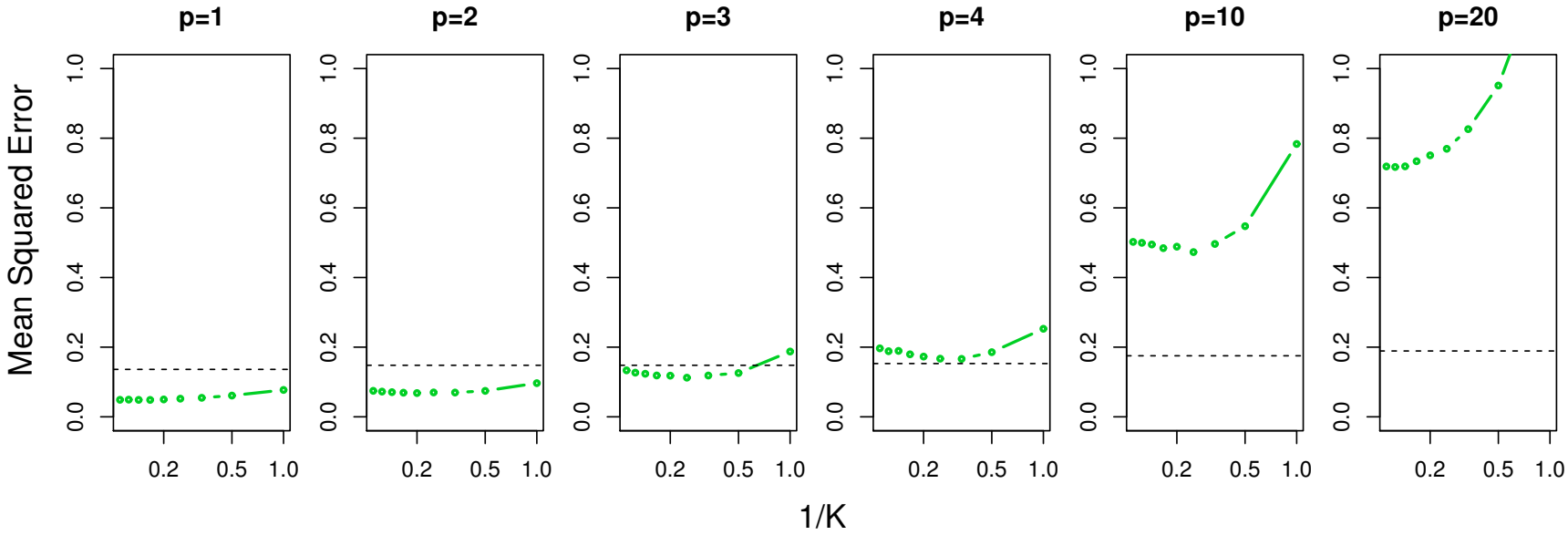
Regressione K-NN su un dataset bidimensionale di 64 osservazioni (punti arancioni)
Sinistra: $K = 1$, destra: $K = 9$

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 relativamente lontani



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	$(h(x) - y)^2$

Regressione con altre funzioni di costo

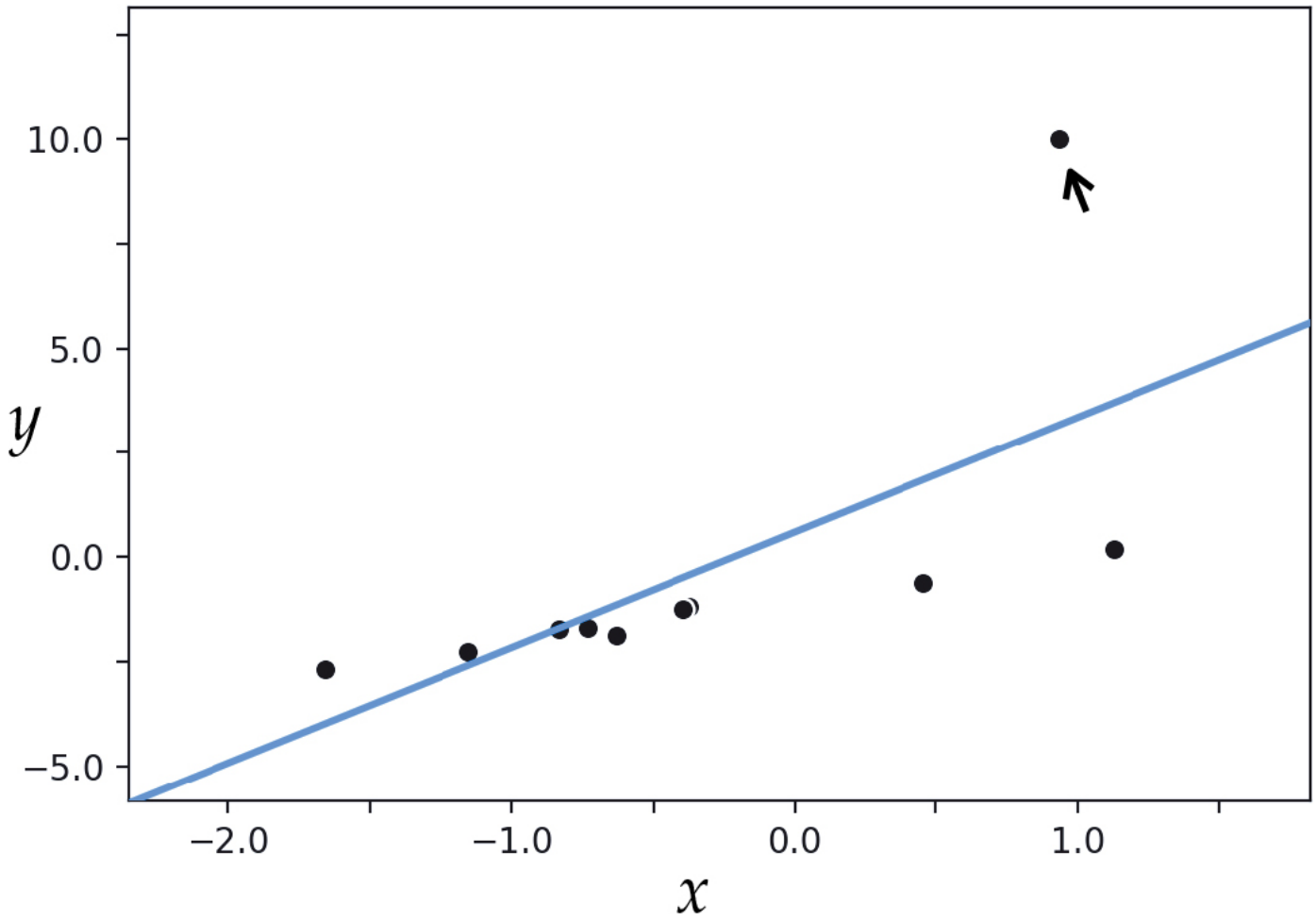
Come trattare funzioni di costo diverse da quella quadratica?

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

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

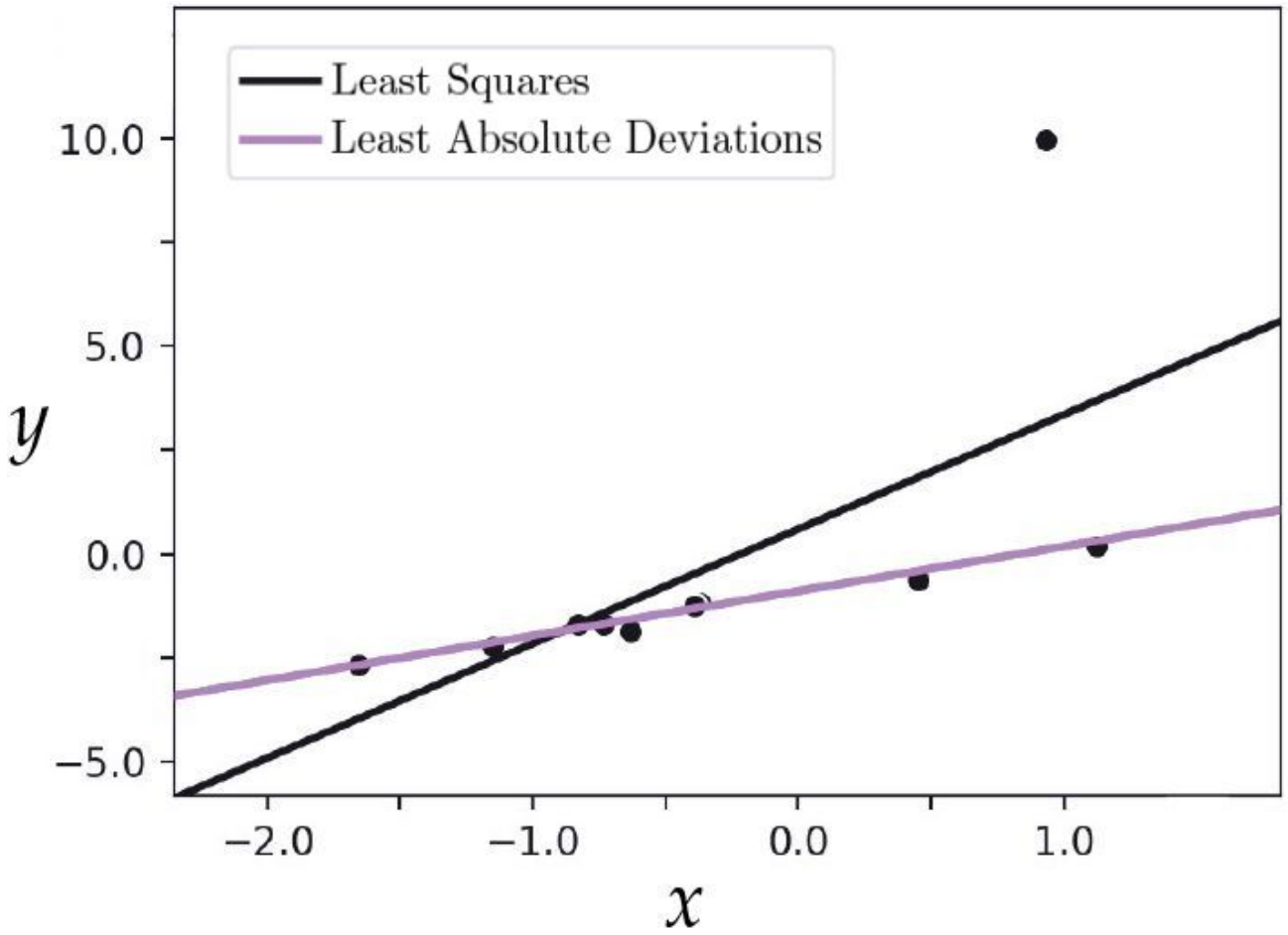
Per una vasta classe di funzioni di costo (convesse e/o differenziabili) esiste una metodologia generale di ottimizzazione: la discesa del gradiente

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



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

Outlier: Metodo dei minimi quadrati vs. LAD



- Il costo LAD è più robusto rispetto agli outlier
- 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 (con costo quadratico) diventa

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

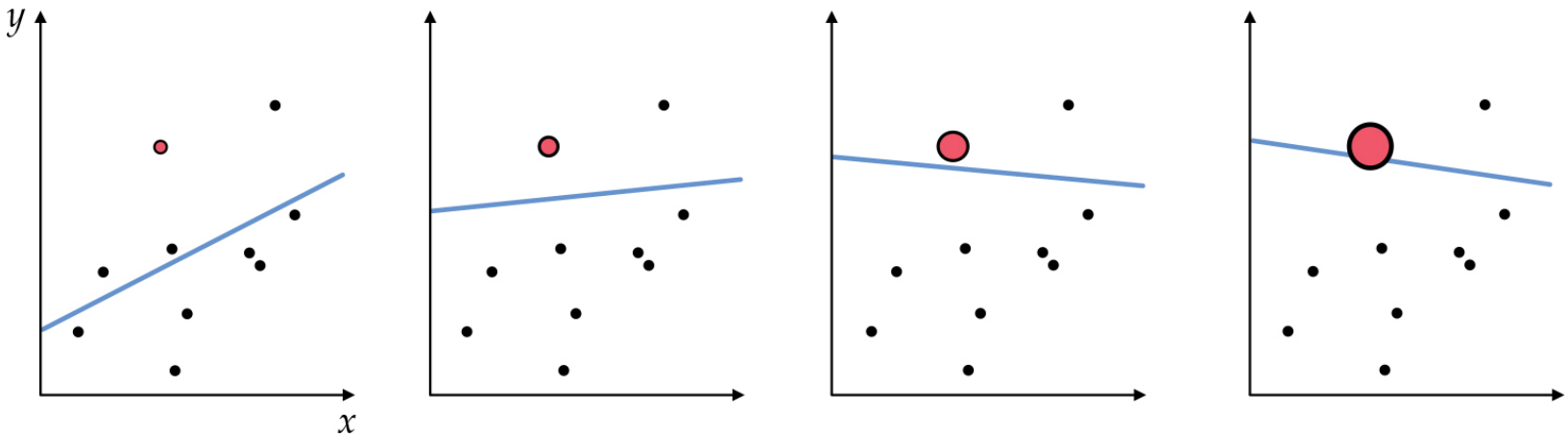
Esempi duplicati e regressione lineare pesata

La minimizzazione del rischio empirico

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

può essere interpretata come una *regressione lineare pesata*

L'esempio $(x^{(i)}, y^{(i)})$ è pesato con un fattore β_i



Regressione multi-output

Abbiamo supposto un singolo output: $y^{(i)} \in \mathbb{R}^{1\times 1}$

Come gestire più variabili di output? Es. $y^{(i)} \in \mathbb{R}^{1\times c}$
Il vettore di parametri w diventa una matrice $W \in \mathbb{R}^{(d+1)\times c}$

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

Problema equivalente a c regressioni con output singolo