Università di Genova | MaLGa

Metodi predittivi

Introduzione alla Data Science

Nicoletta Noceti

Oggi

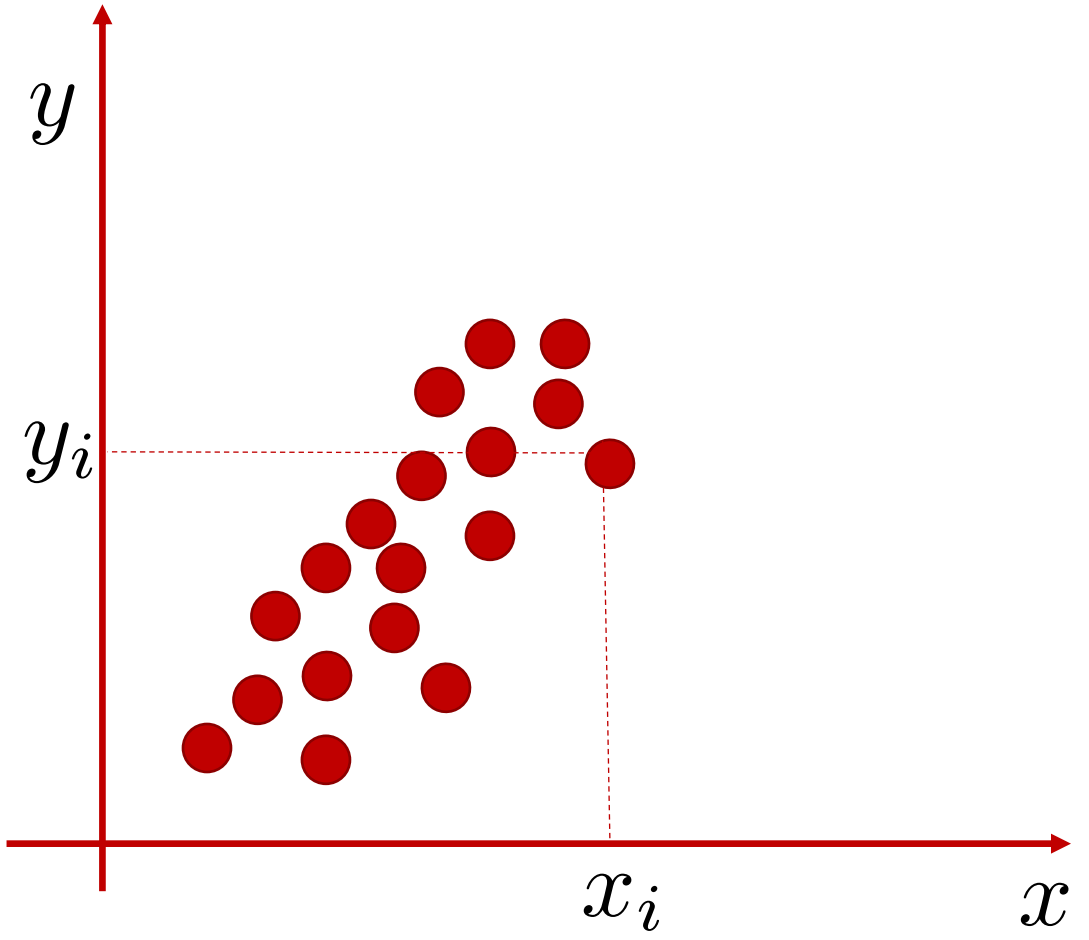
- Introduciamo oggi modelli con capacità predittive che imparano da esempi
- Cosa sono gli esempi? Sono i dati!

UniGe | MaLGa

Machine Learning

Esempio: regressione

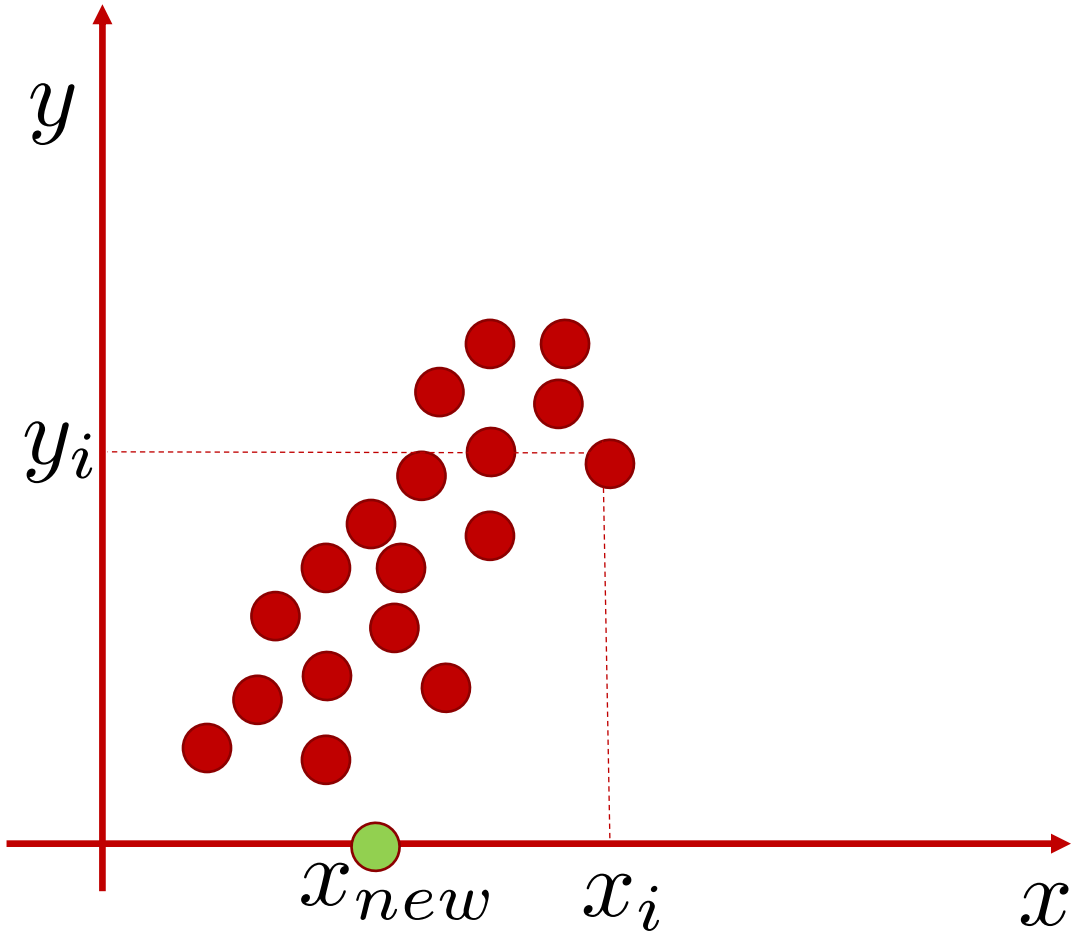
$$Y \subseteq \mathbb{R}$$



$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: regressione

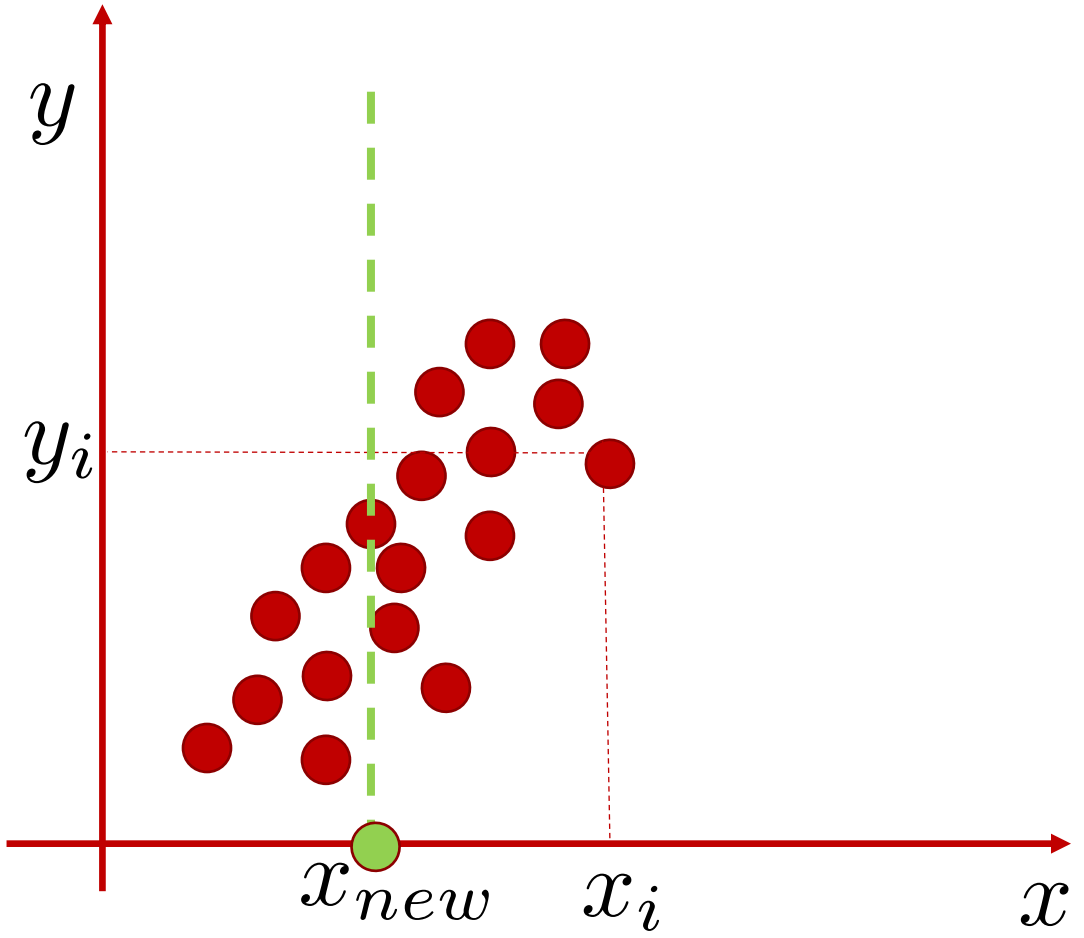
$$Y \subseteq \mathbb{R}$$



$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: regressione

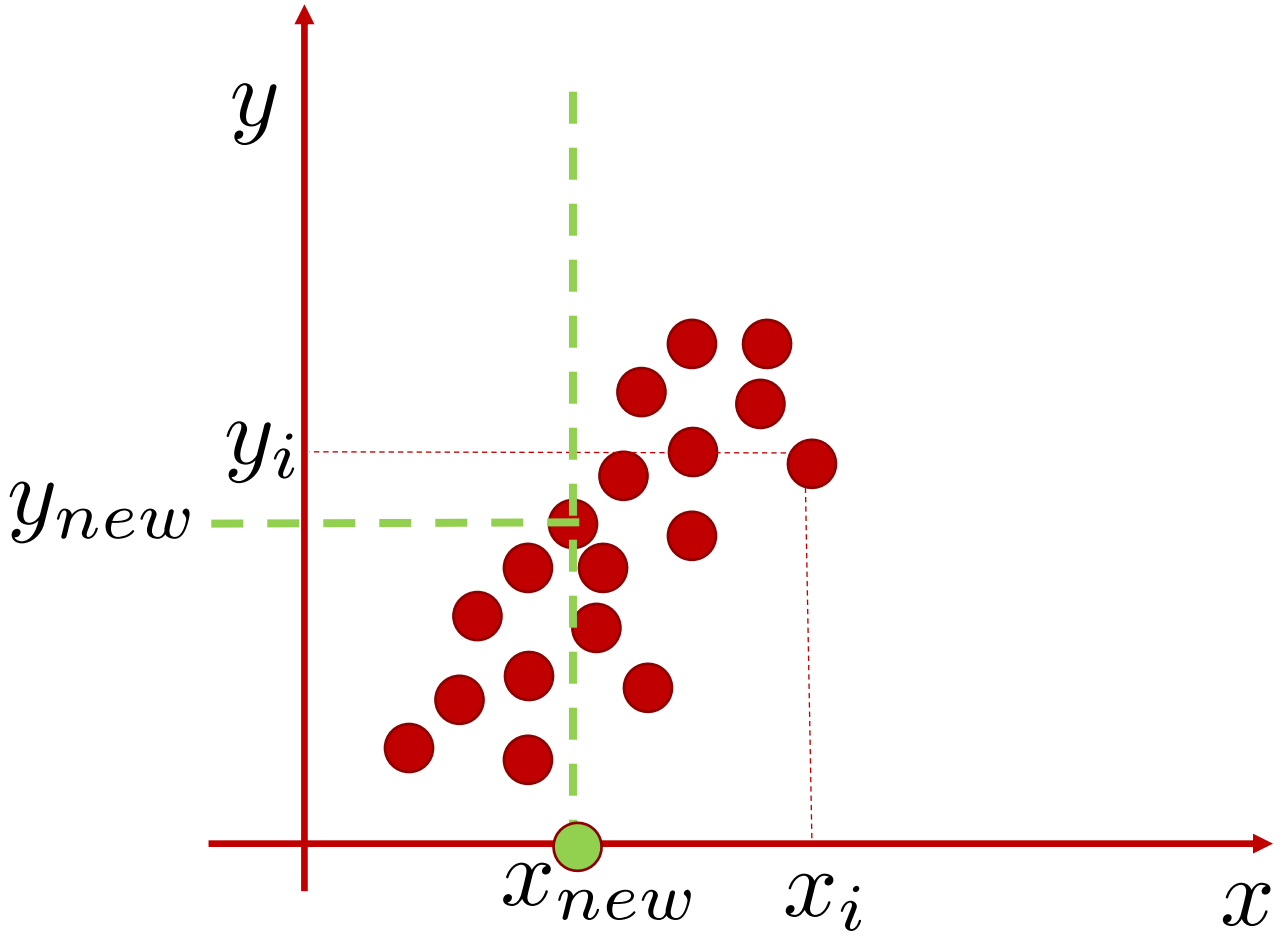
$$Y \subseteq \mathbb{R}$$



$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: regressione

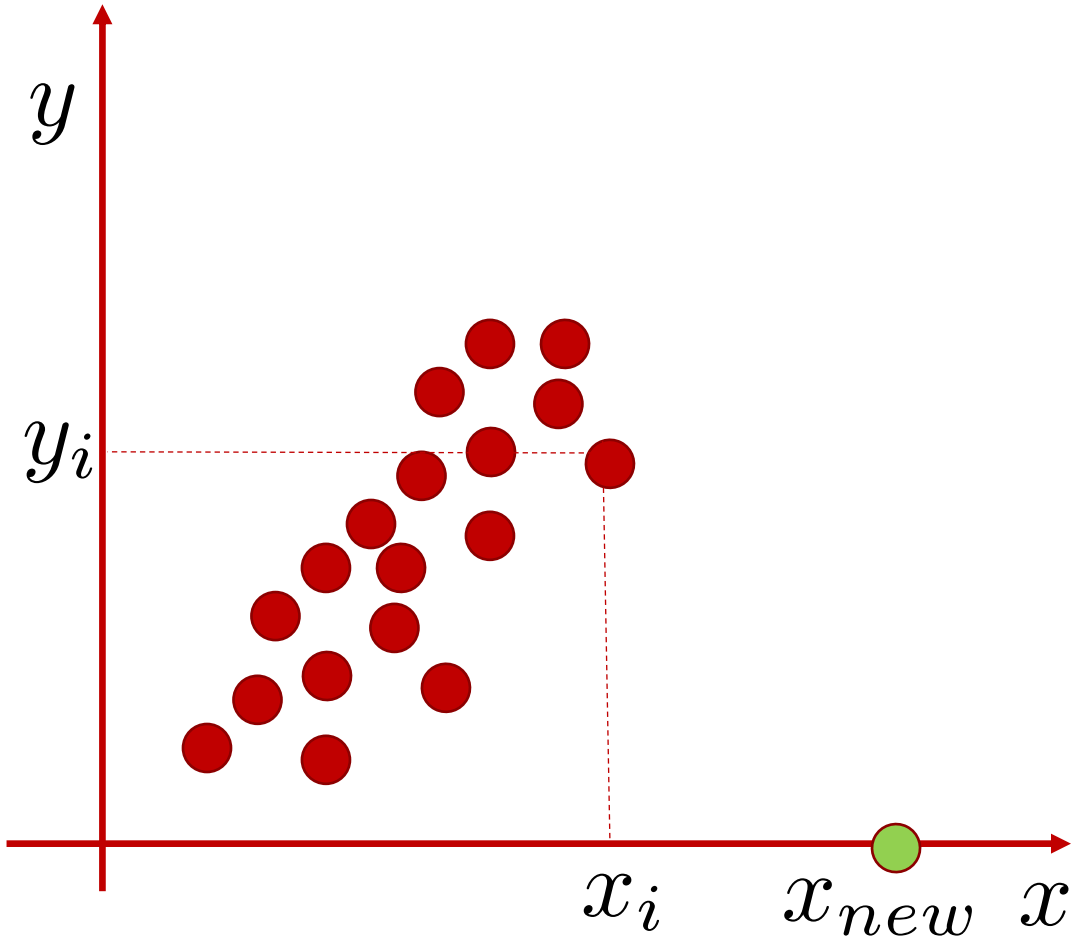
$$Y \subseteq \mathbb{R}$$



$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: regressione

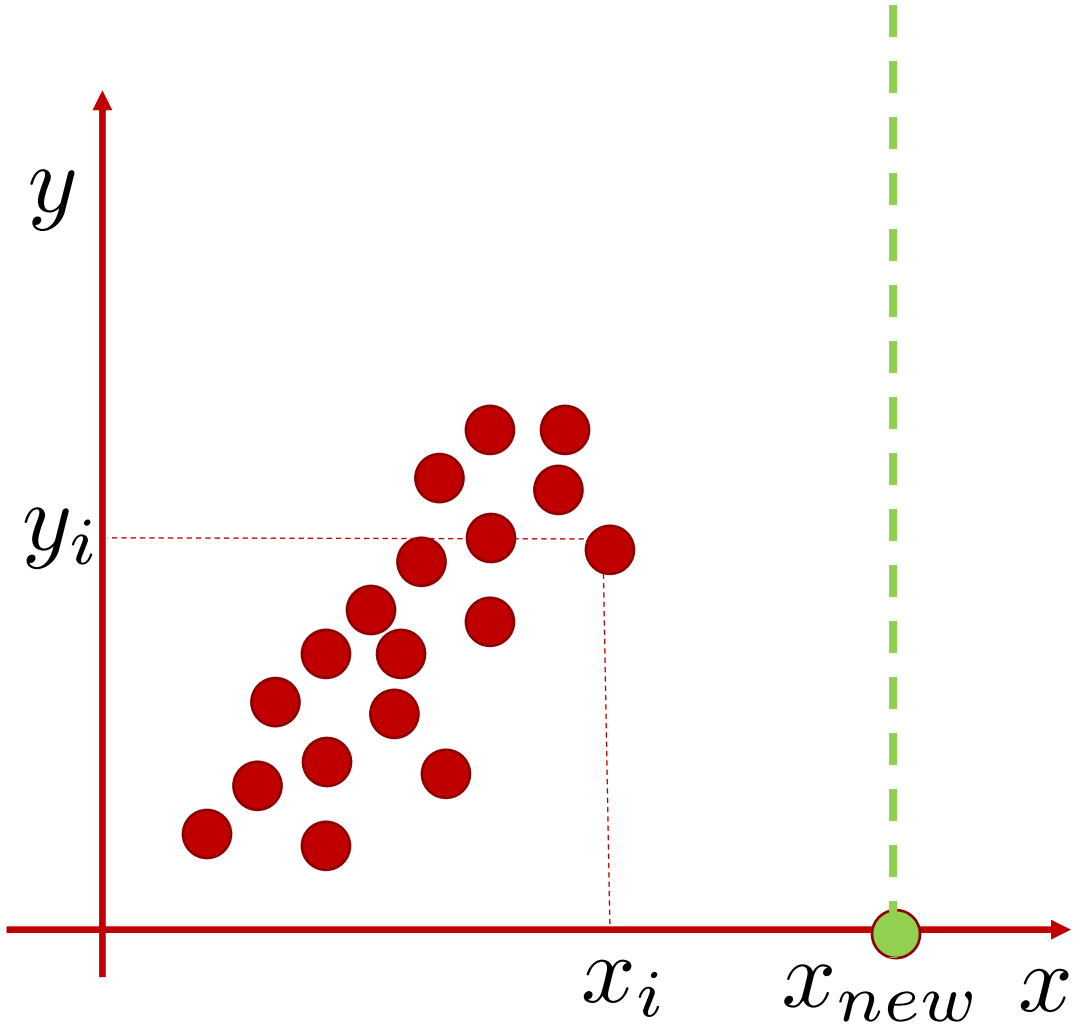
$$Y \subseteq \mathbb{R}$$



$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: regressione

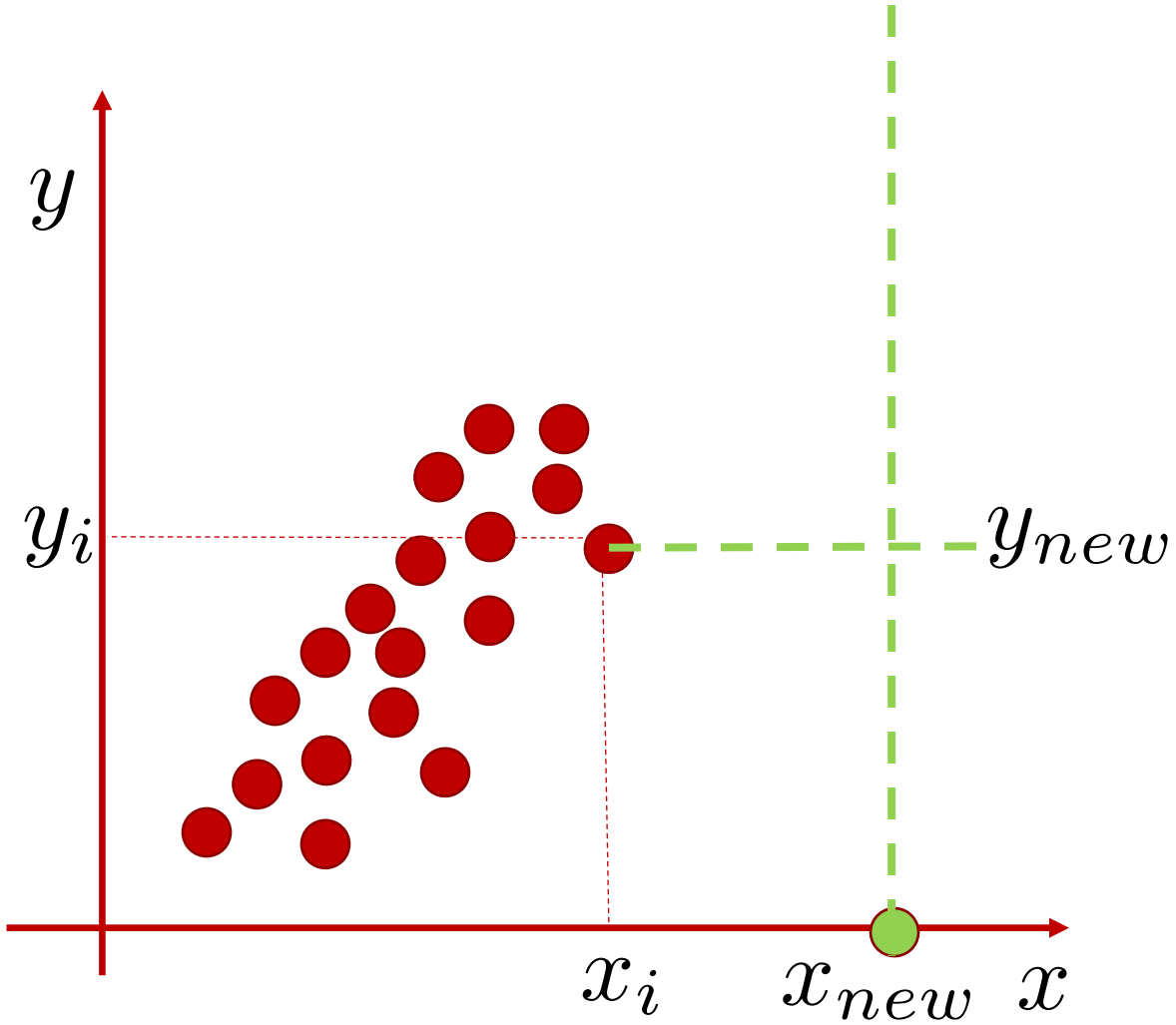
$$Y \subseteq \mathbb{R}$$



$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: regressione

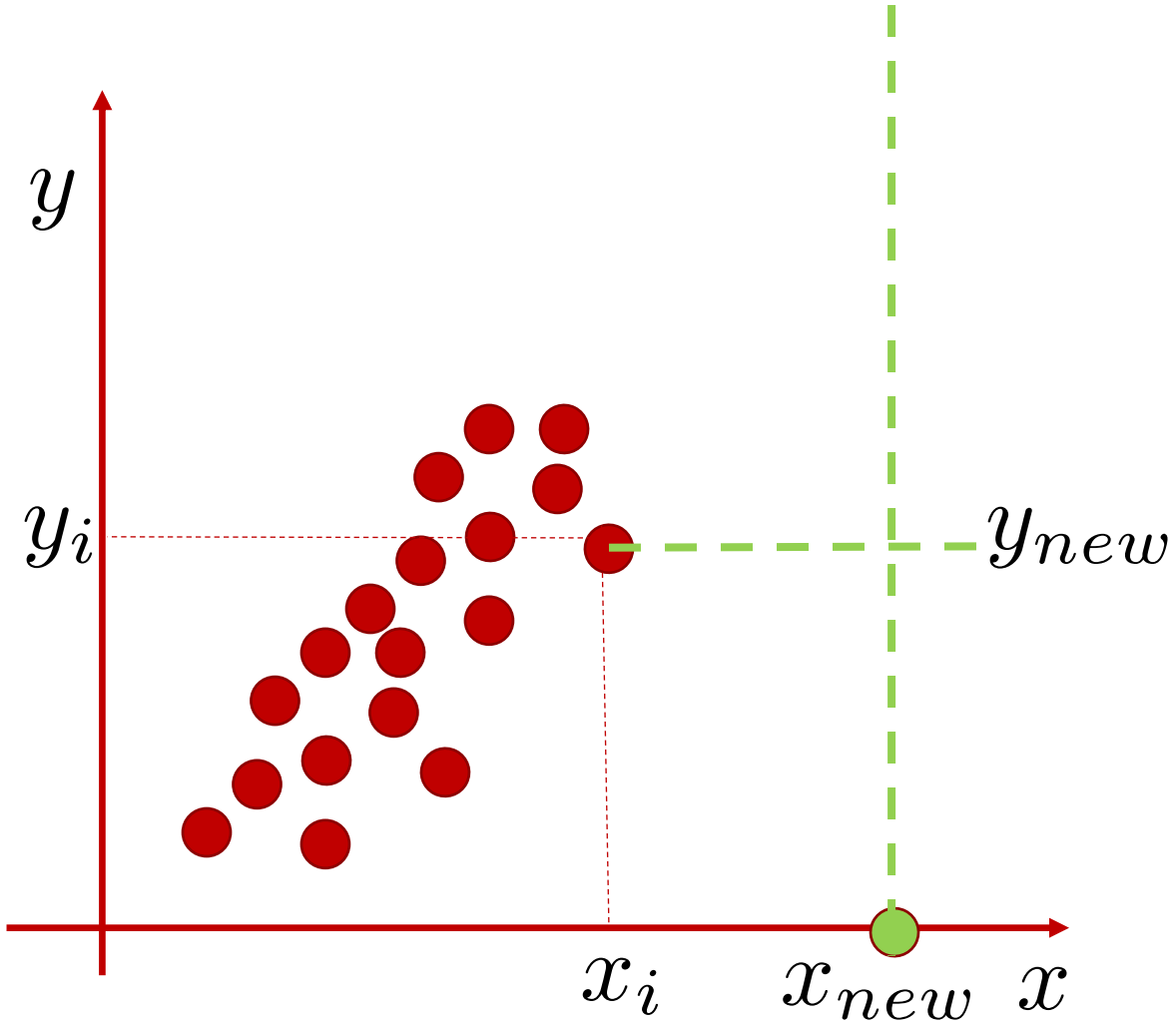
$$Y \subseteq \mathbb{R}$$



$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: regressione

$$Y \subseteq \mathbb{R}$$

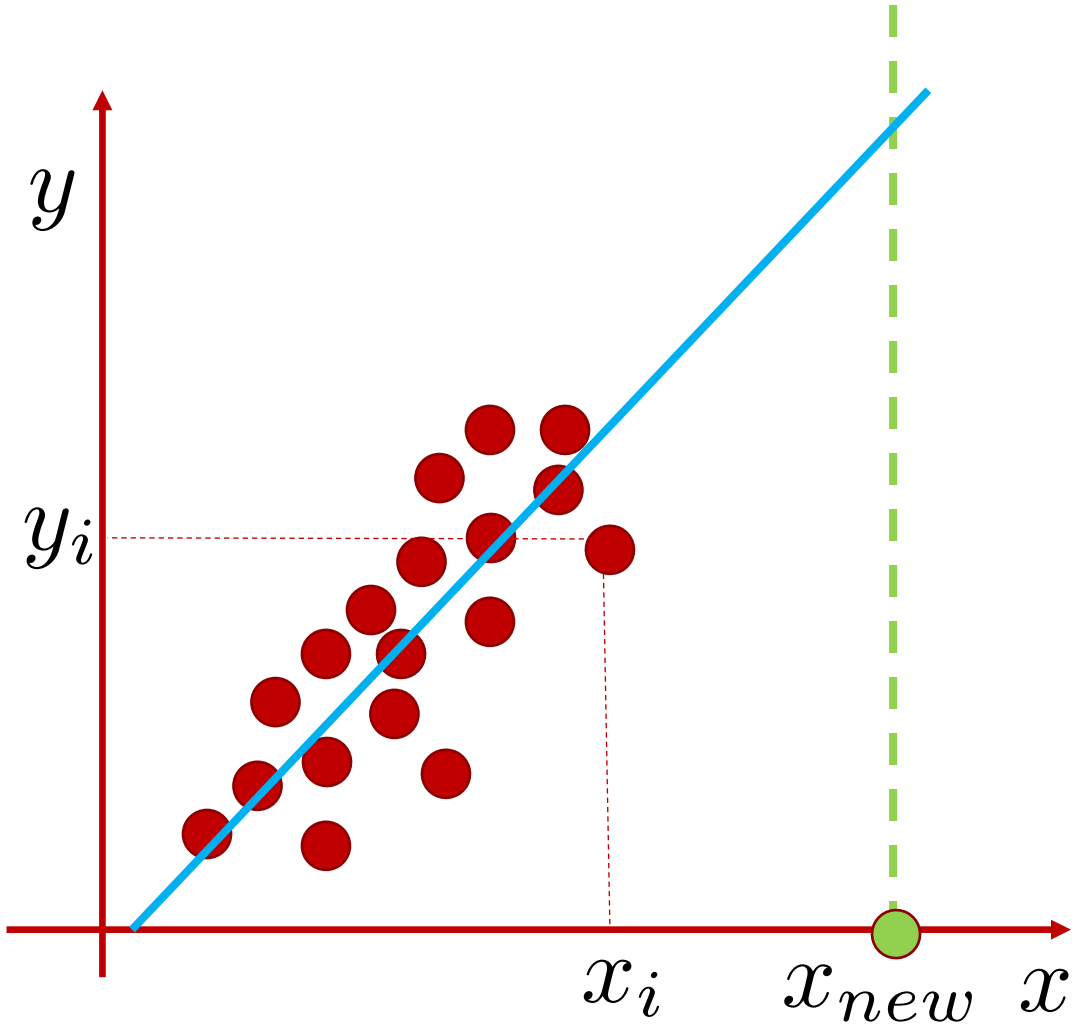


E' una predizione "ragionevole"?

$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: regressione

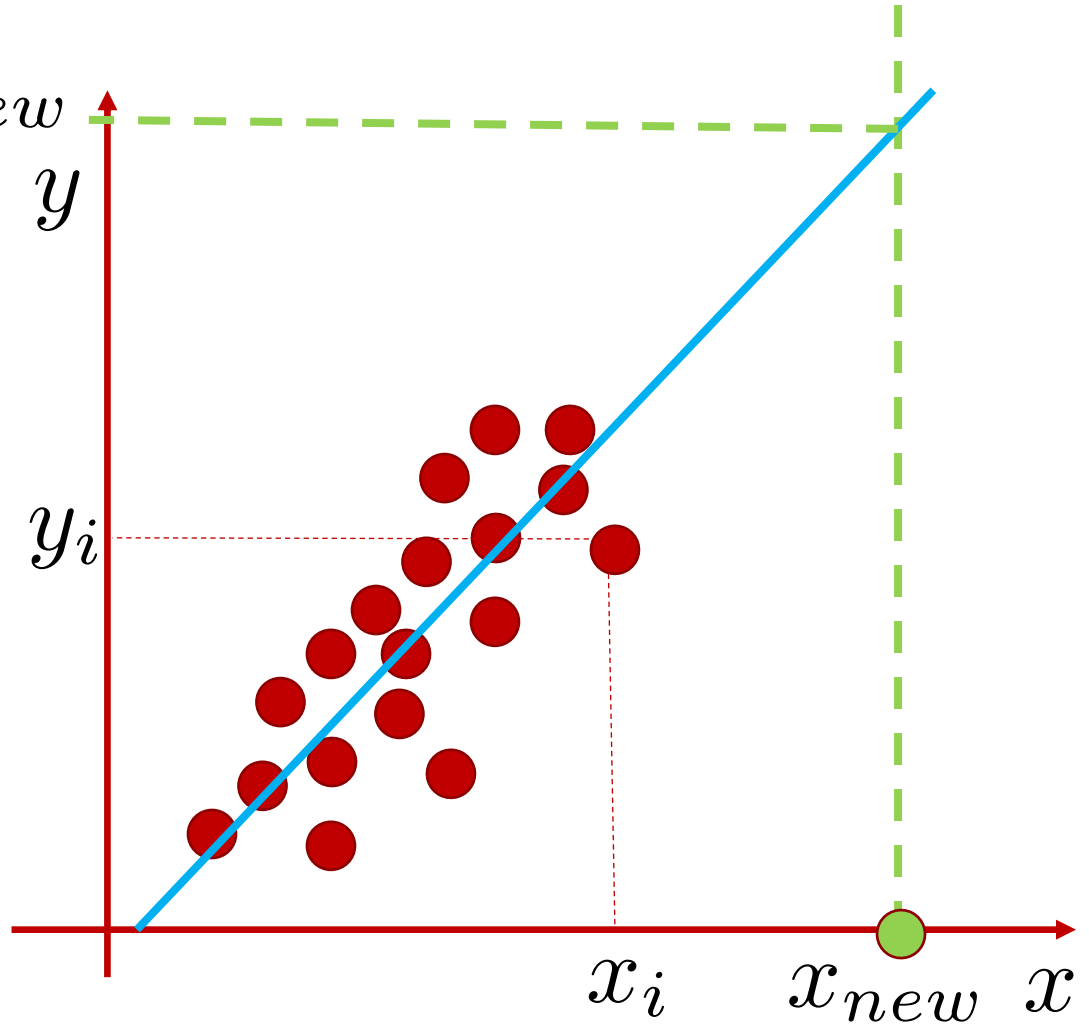
$$Y \subseteq \mathbb{R}$$



$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: regressione

$$Y \subseteq \mathbb{R}$$



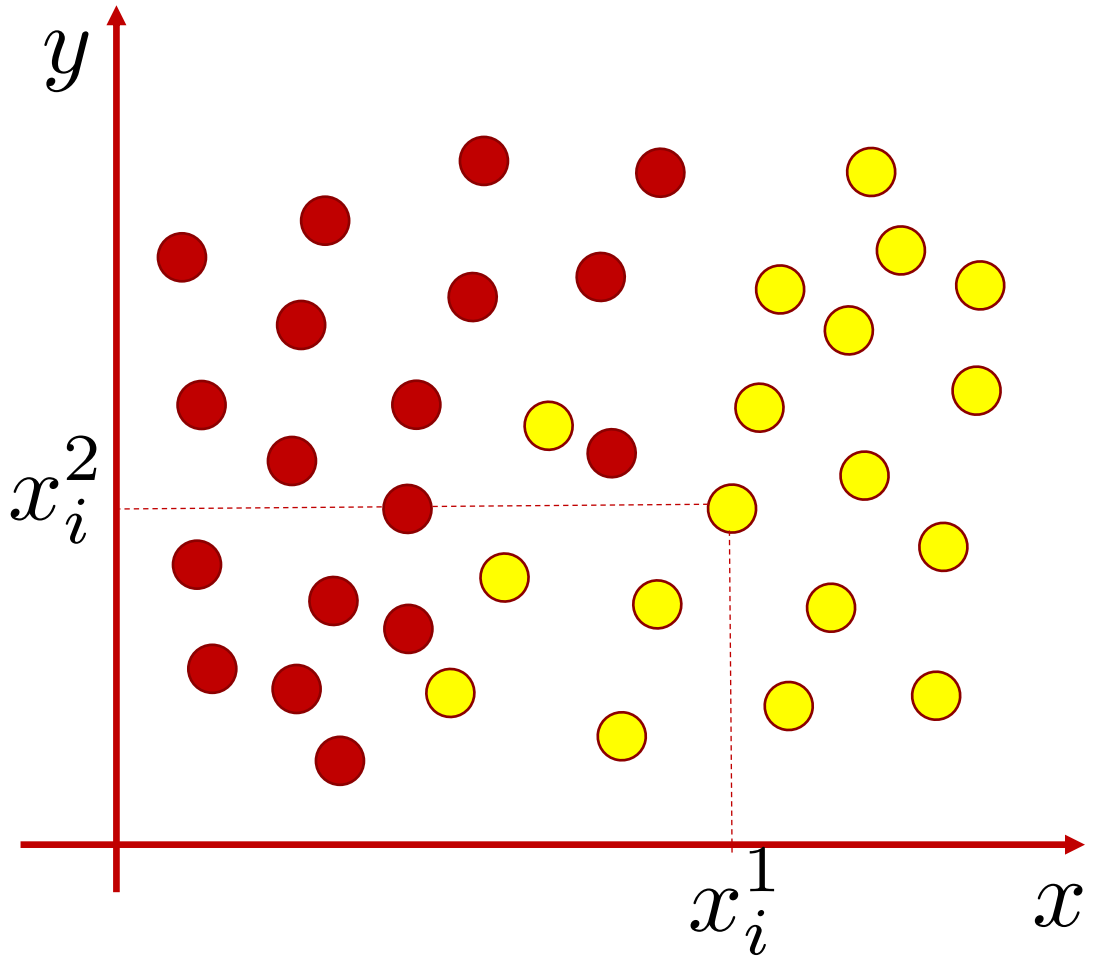
$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: classificazione (binaria)

$$Y = \{-1, 1\}$$

$$X \subseteq \mathbb{R}^2$$

$$x_i = [x_i^1, x_i^2]$$



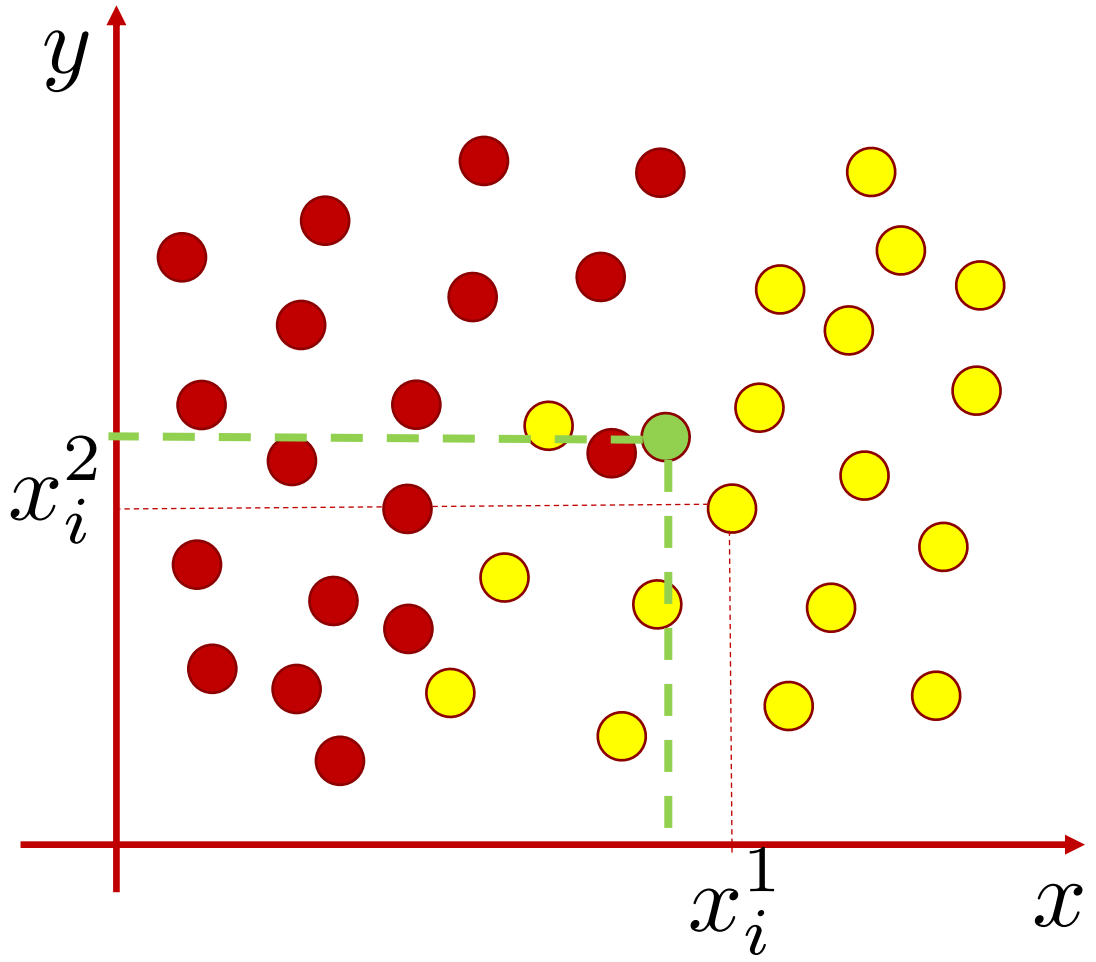
$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: classificazione (binaria)

$$Y = \{-1, 1\}$$

$$X \subseteq \mathbb{R}^2$$

$$x_i = [x_i^1, x_i^2]$$



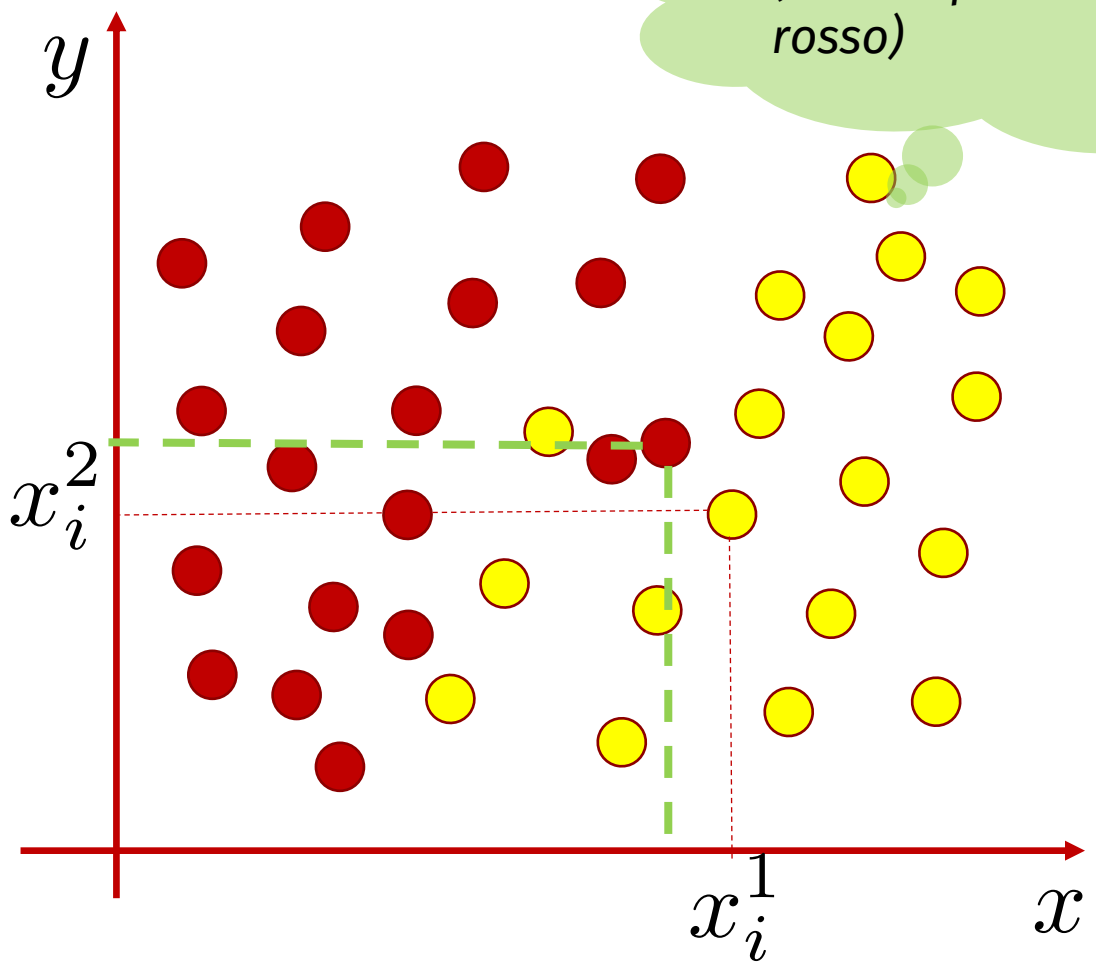
$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: classificazione (binaria)

$$Y = \{-1, 1\}$$

$$X \subseteq \mathbb{R}^2$$

$$x_i = [x_i^1, x_i^2]$$



La classe del nuovo punto è -1 (ossia il punto dovrebbe essere rosso)

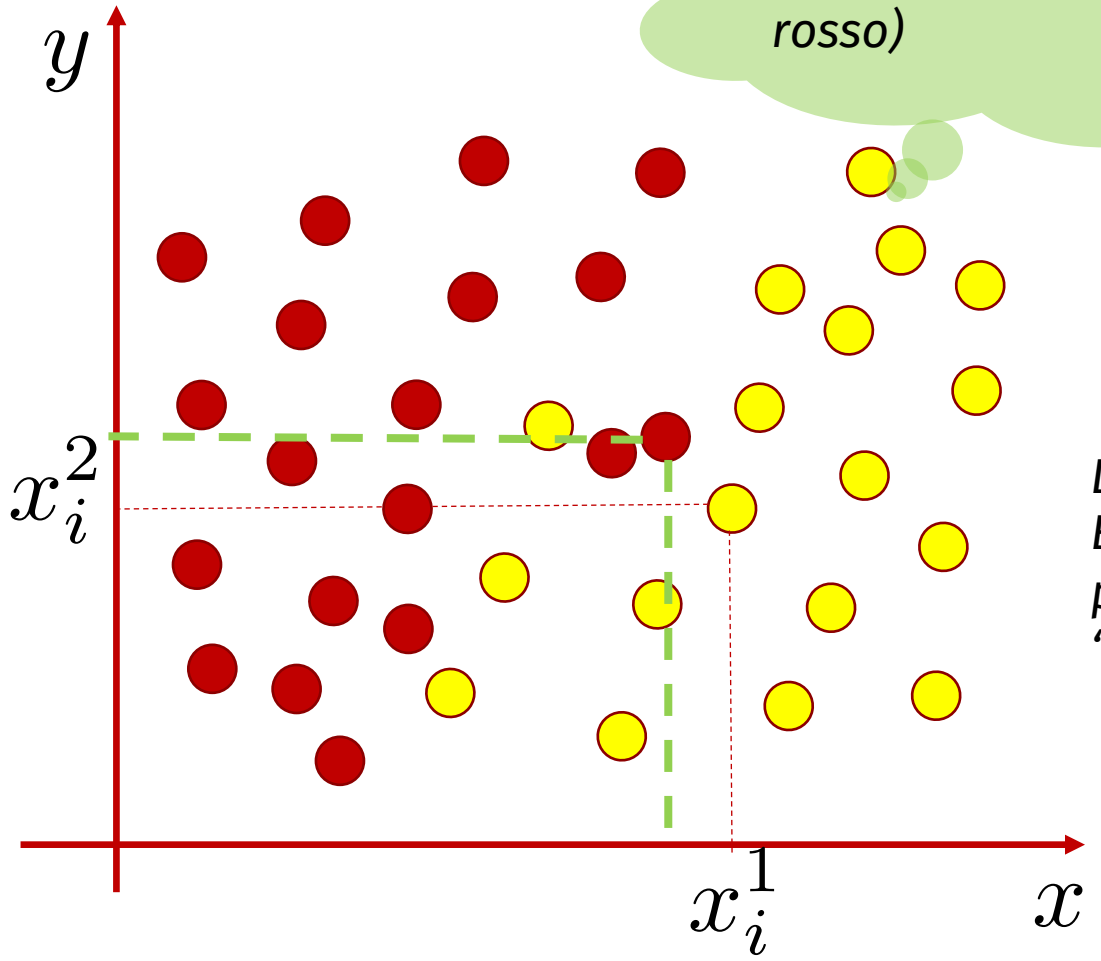
$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: classificazione (binaria)

$$Y = \{-1, 1\}$$

$$X \subseteq \mathbb{R}^2$$

$$x_i = [x_i^1, x_i^2]$$



La classe del nuovo punto è -1 (ossia il punto dovrebbe essere rosso)

Di nuovo… E' una predizione "ragionevole"?

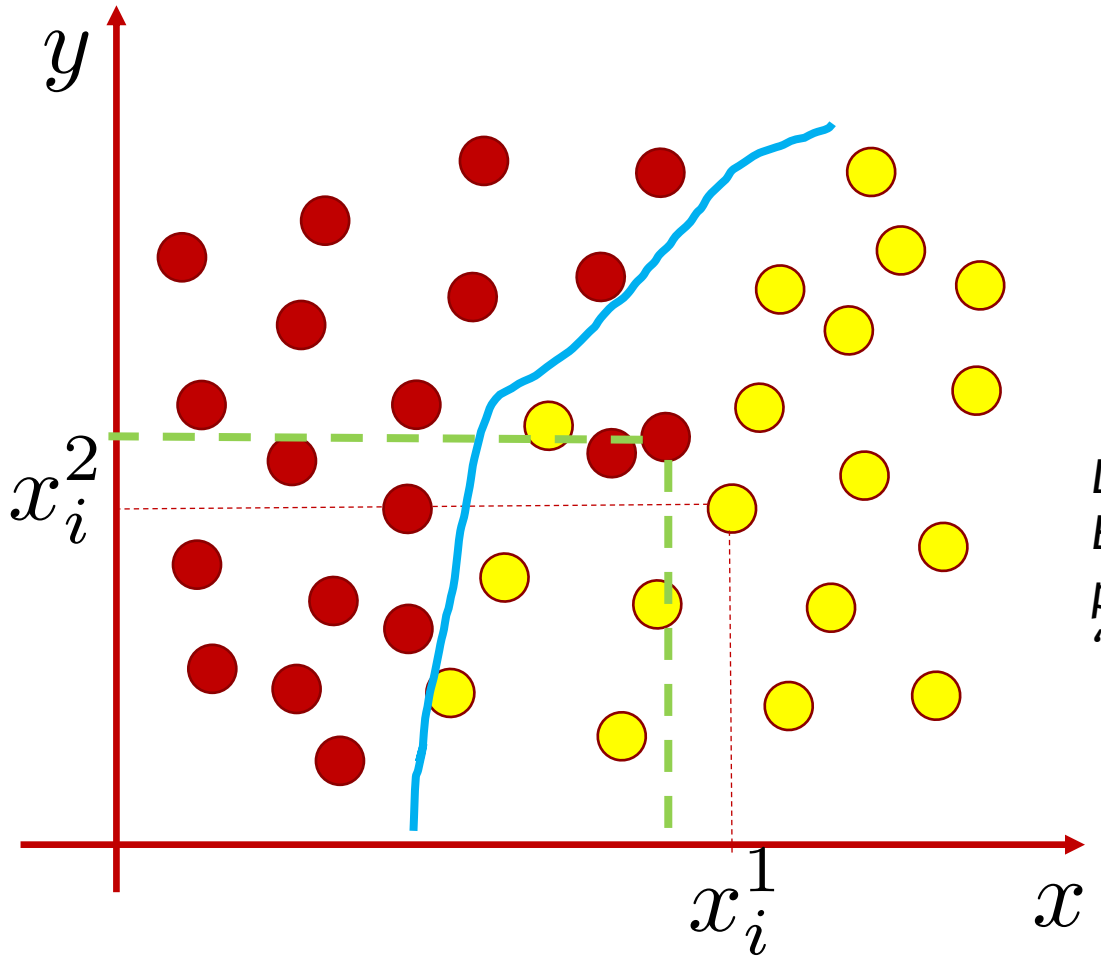
$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: classificazione (binaria)

$$Y = \{-1, 1\}$$

$$X \subseteq \mathbb{R}^2$$

$$x_i = [x_i^1, x_i^2]$$



Di nuovo…
E' una predizione "ragionevole"?

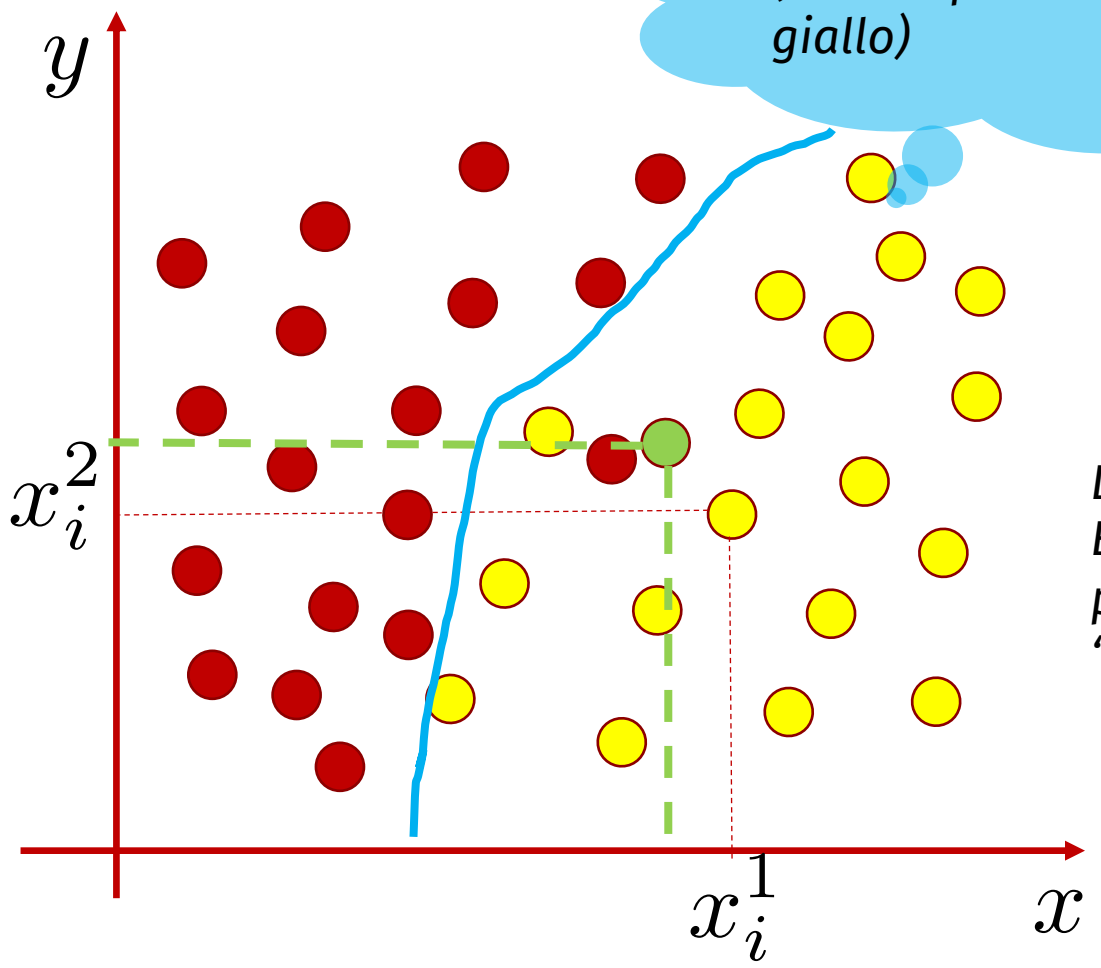
$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Esempio: classificazione (binaria)

$$Y = \{-1, 1\}$$

$$X \subseteq \mathbb{R}^2$$

$$x_i = [x_i^1, x_i^2]$$



La classe del nuovo punto è 1 (ossia il punto dovrebbe essere giallo)

Di nuovo… E' una predizione "ragionevole"?

$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Una definizione

Con il termine Machine Learning ci riferiamo ad una classe di metodi in grado di imparare da esempi invece di essere esplicitamente programmati a fare qualcosa

Due macro-tipologie

- Machine Learning supervisionato (simula l'imparare con un insegnante)
- Machine Learning non supervisionato (simula l'imparare senza un insegnante)

Oggi parliamo di Machine Learning supervisionato

Definiamo il problema

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

è il training set, ossia l'insieme di dati da cui il metodo imparerà

Ogni coppia rappresenta un possibile (input-output) del nostro problema

$$x_k \in \mathbb{R}^d \qquad y_k \in ?$$ *DIPENDE! Ci torniamo tra un attimo*

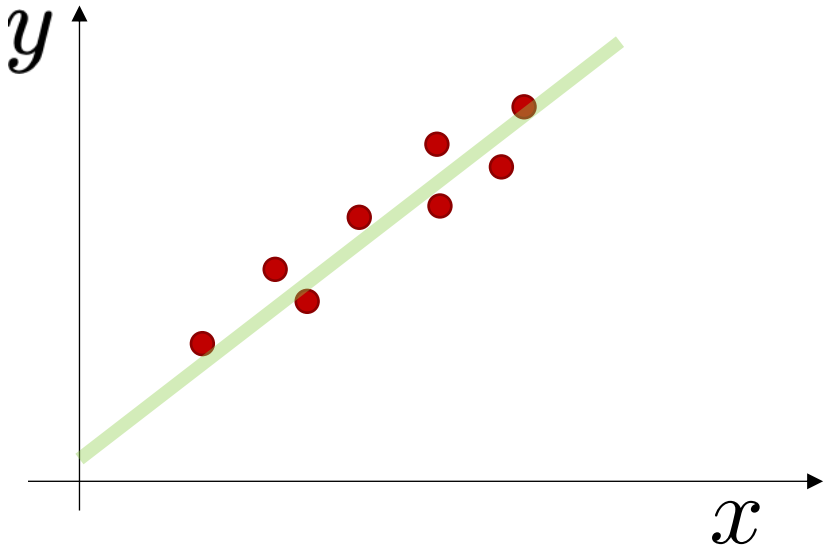
Lo scopo del Machine Learning supervisionato è stimare una funzione f tale che

$$f(x_k) = y_k \quad \text{per ogni} \quad (x_k, y_k) \in S$$

... ma questo non basta

Chi è y? Dipende dal problema...

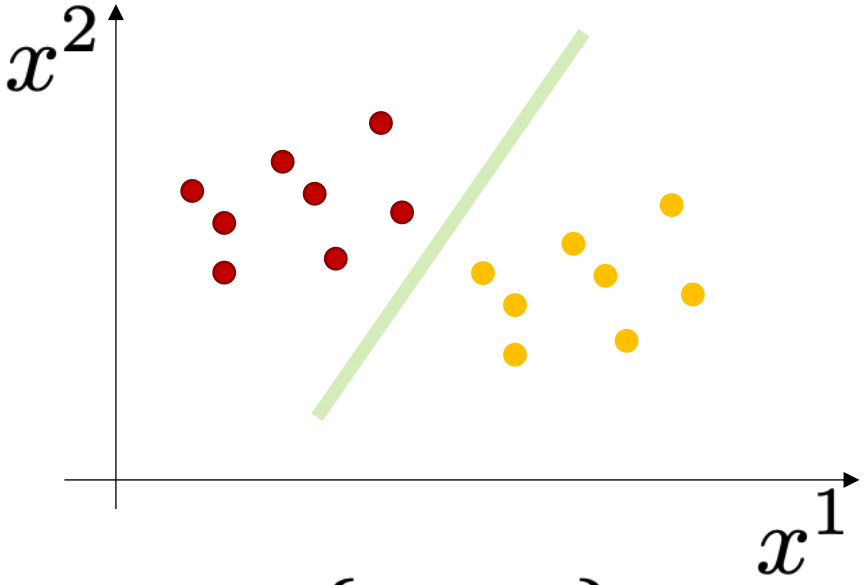
Regressione



$$y_k \in \mathbb{R}$$

$$f : \mathbb{R}^d \to \mathbb{R}$$

Classificazione



$$y_k \in \{-1, 1\}$$

$$f : \mathbb{R}^d \to \{-1, 1\}$$

Definiamo meglio il problema

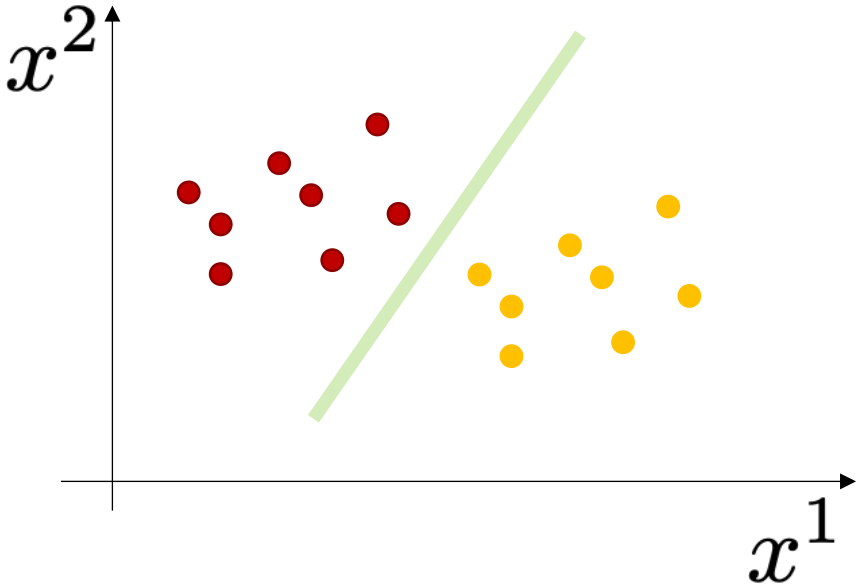
La funzione che cerchiamo deve avere due importanti proprietà:

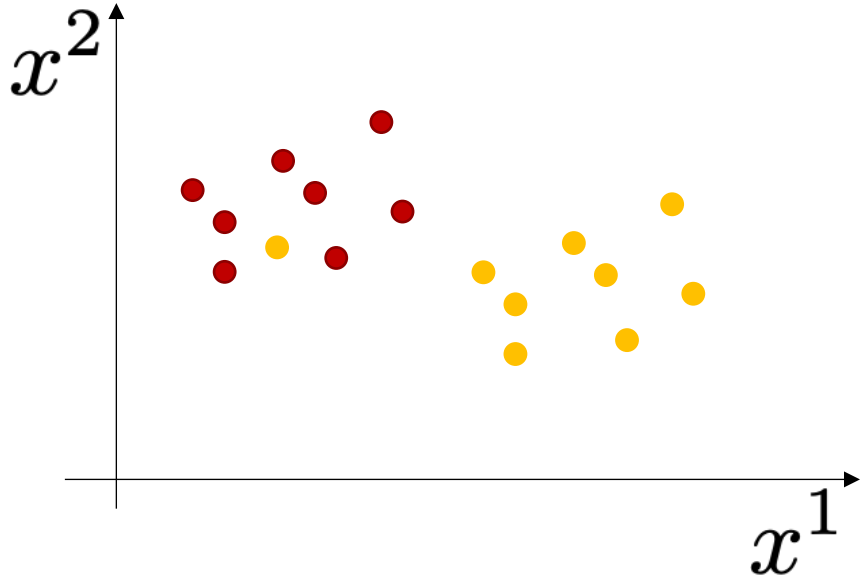
- Capacità di rappresentare i dati di training (fitting)
- La capacità di generalizzare ai dati che avremo a disposizione in futuro (stability/generalization)

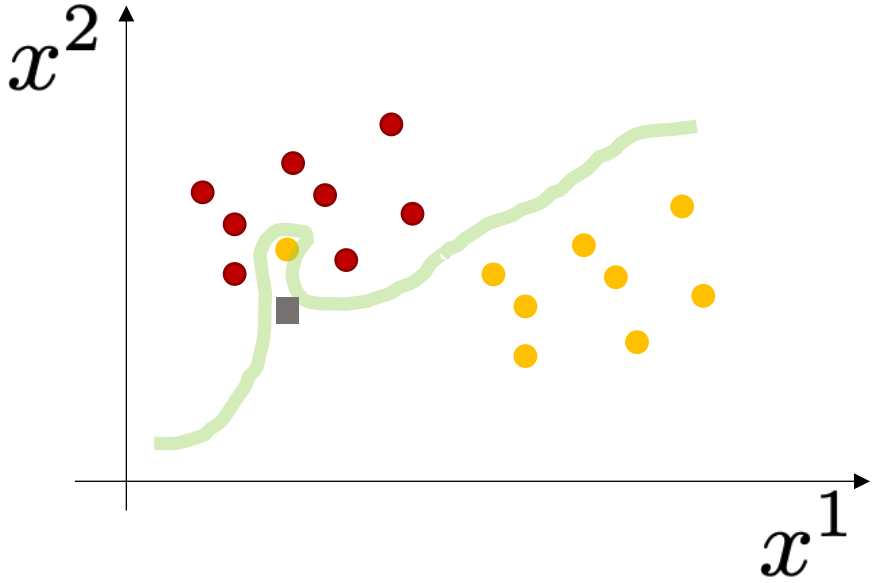
Ogni algoritmo di Machine Learning ambisce a trovare delle soluzioni che siano un compromesso tra queste due proprietà

È tale compromesso che ci garantisce la giusta capacità predittiva!

Cerchiamo di capirlo con un esempio

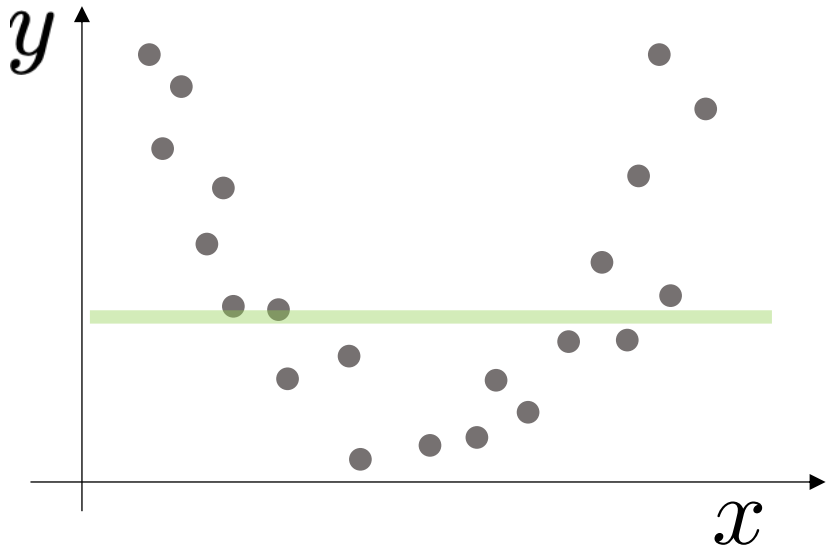


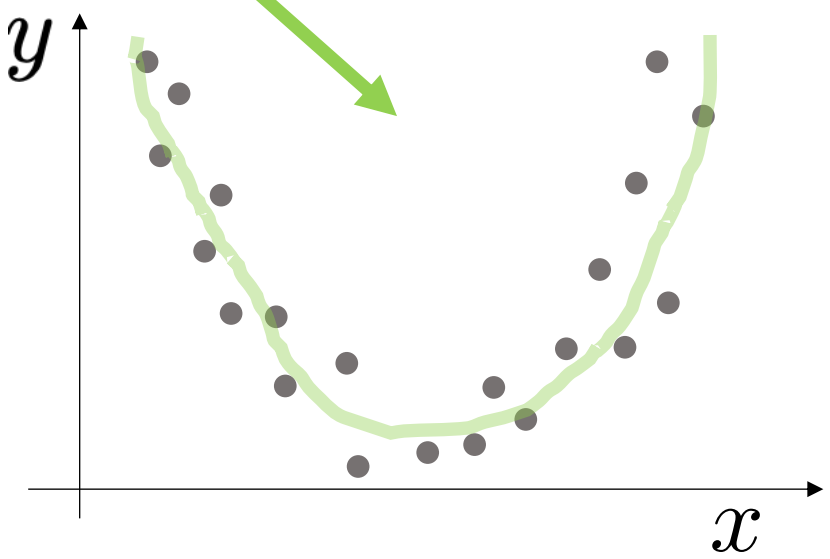


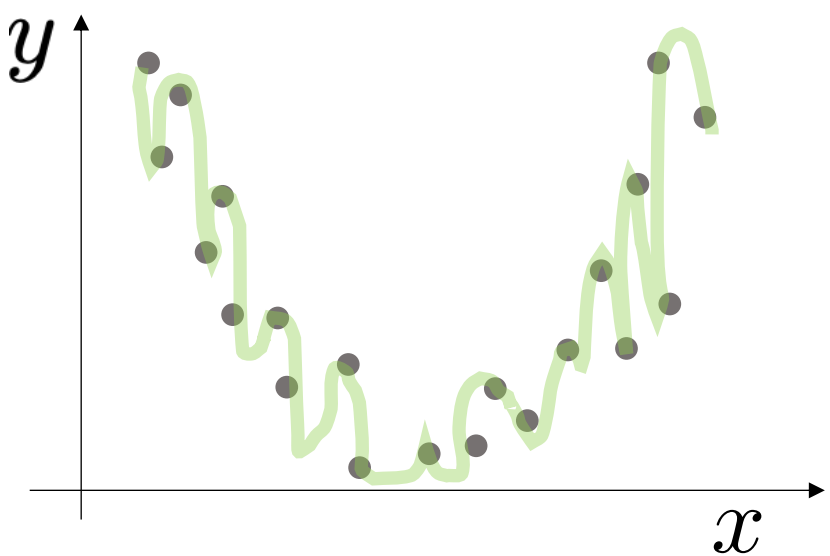


Il nuovo punto verrebbe classificato nel modo sbagliato perché la funzione che abbiamo stimato crede troppo ai dati di training e non generalizza bene a nuovi punti

Il compromesso







Troppo stabile!
→ Non rappresenta i dati: se i dati cambiano, la funzione non cambia

Troppo complicata!
→ Rappresenta perfettamente i dati
→ Non generalizza: se i dati cambiano anche di poco, la funzione cambia profondamente

Allenare un modello di ML supervisionato

Come raggiungere il compromesso?

- Allenare un modello di Machine Learning significa stimare la funzione (di regressione o classificazione) che può dipendere da parametri e iper-parametri
- Vedremo che per allenare un modello di Machine Learning supervisionato ci serviranno 3 insiemi:
  - Insieme di Training → Serve per stimare la funzione
  - Insieme di Validation → Serve per verificare che la funzione stimata con si comporti bene anche su nuovi dati
  - Insieme di Test → Usato SOLO ALLA FINE, serve per verificare la qualità della funzione finale

Allenare un modello di ML supervisionato

Uno schema generale

Supponiamo di voler allenare un modello che prevede di considerare funzioni f_w^θ che dipendono da parametri w e un iperparametro θ

- Per ogni possibile valore θ_i che θ può assumere:
  - Stimiamo la funzione $f_w^{\theta_i}$ (ossia determiniamo i valori di w) sul training set
  - Verifichiamo il comportamento di $f_w^{\theta_i}$ sul validation set (usando errore o accuratezza)
- Tra le funzioni stimate, selezioniamo quella che si comporta meglio sul validation set
- Verifichiamo il comportamento di tale funzione sul test set

UniGe | MaLGa

Regressione lineare

Definizione

- Cerchiamo una relazione lineare tra input e output

$$x = (x^1, x^2, \ldots, x^d) \in \mathbb{R}^d \qquad y \in \mathbb{R}$$

$$y = f(x) = \alpha^0 + \alpha^1 x^1 + \alpha^2 x^2 + \cdots + \alpha^d x^d$$

- Come facciamo a trovare la funzione «migliore»?

Quantificare la bontà di una funzione di regressione

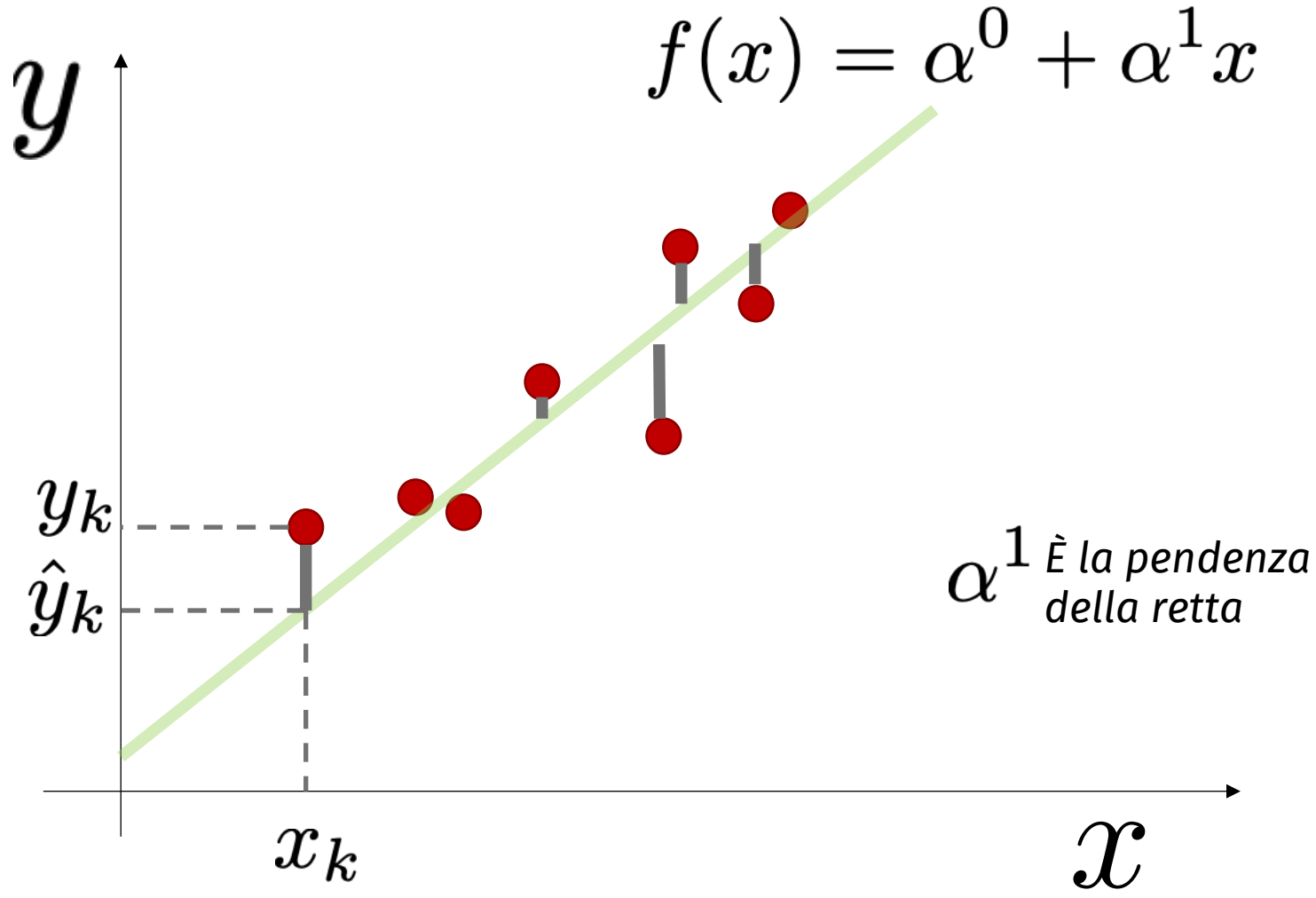
- Immaginiamo di dover valutare la bontà di una funzione stimata $\hat{f}$
- Tale funzione ci fornirà una predizione su ogni input del training set

$$\hat{f}(x_k) = \hat{y}_k$$

- A questo punto possiamo valutare quanto la predizione si discosta dalla verità (ecco dove usiamo la supervisione, ossia l'insegnante!) calcolando i residui:

$$R = \sum_{k=1}^{n} (\hat{y}_k - y_k)^2$$

Cosa sono i residui?



α^0 *È l'intercetta*

α^1 *È la pendenza della retta*

Training del modello

- Il processo di identificare, tra tante soluzioni possibili, la soluzione migliore a partire dai dati a disposizione prende il nome di fase di training di un metodi di Machine Learning
- Ricordate che l'insieme S viene chiamato training set
- In cosa consiste il training del metodo di regressione lineare? Dobbiamo trovare i parametri della f migliore, ossia i coefficienti… come?

Training del metodo di regressione lineare

$$\alpha^* = \arg \min_{\alpha \in \mathbb{R}^{d+1}} \sum_{k=1}^{n} (\hat{y}_k - y_k)^2$$

Calcoliamo il gradiente della funzione rispetto ai coefficienti e poniamo il gradiente = 0 per trovare i coefficienti che minimizzano il funzionale

Valutare la qualità della funzione stimata

- La metrica che viene di solito utilizzata in regressione è l'errore quadratico medio

$$MSE = \frac{1}{N}\sum_{k=1}^{N}(\hat{y}_k - y_k)^2$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\hat{y}_k - y_k)^2}$$

UniGe | MaLGa

Regressione Logistica

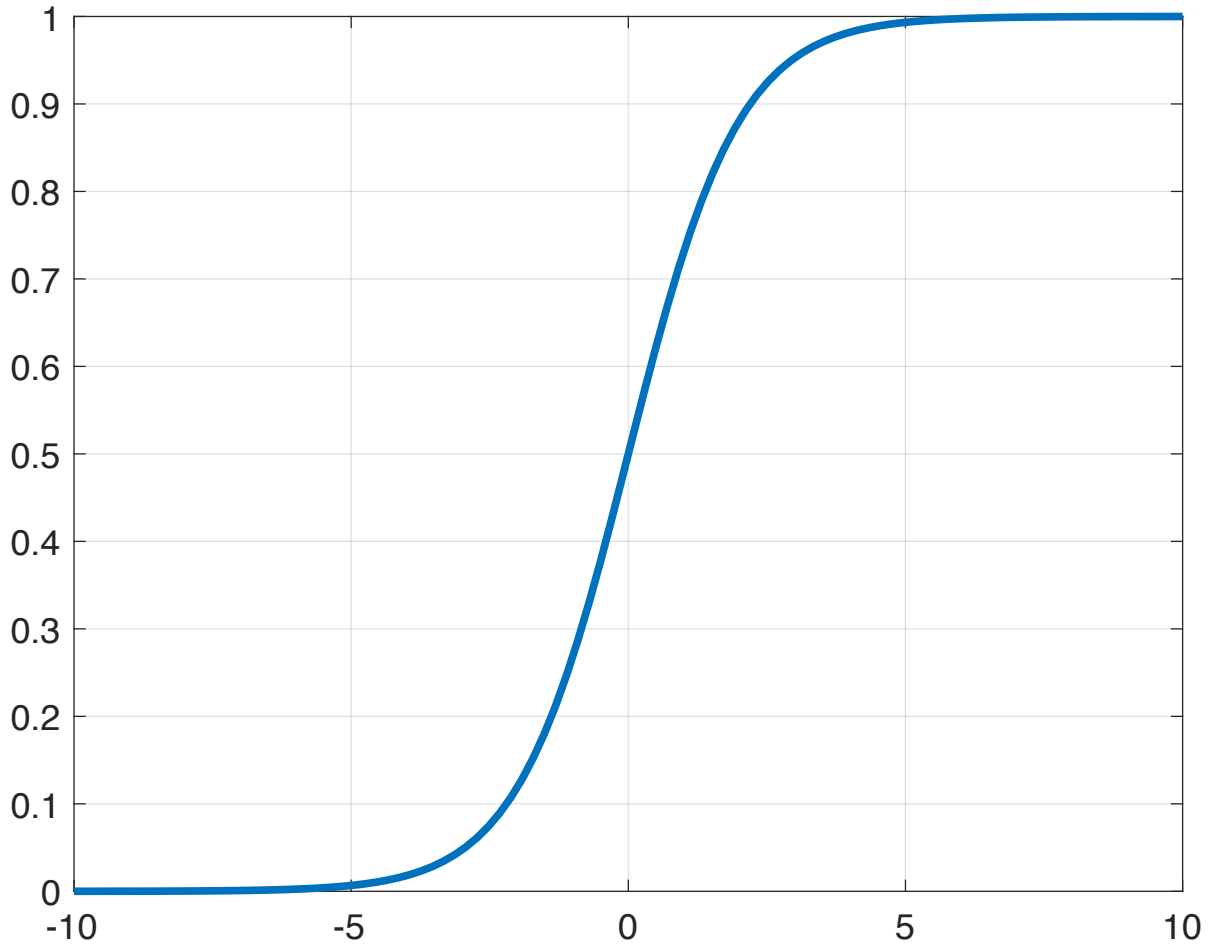
Definizione e formule

- Si tratta di una generalizzazione del modello di regressione lineare ai problemi di classificazione
- Con la regressione logistica prevediamo la probabilità che il dato appartenga ad una certa classe

$$Pr(y = 1|x) = \frac{e^{\alpha^0 + \alpha^1 x}}{1 + e^{\alpha^0 + \alpha^1 x}} = \frac{e^{\hat{f}(x)}}{1 + e^{\hat{f}(x)}}$$

La funzione che ottieniamo

La y può assumere valori in [0,1]



La x può assumere qualunque valore

Come facciamo ad ottenere in output l'etichetta della classe, ossia 1 o -1?

Come ci si arriva? Il concetto di odds

L'odds di un evento è il rapporto tra la sua probabilità p e la probabilità che non accada, cioè 1 – p (evento complementare)

Facciamo un esempio: immaginiamo che delle 3000 persone che entrano in un negozio 1000 acquistino effettivamente qualcosa

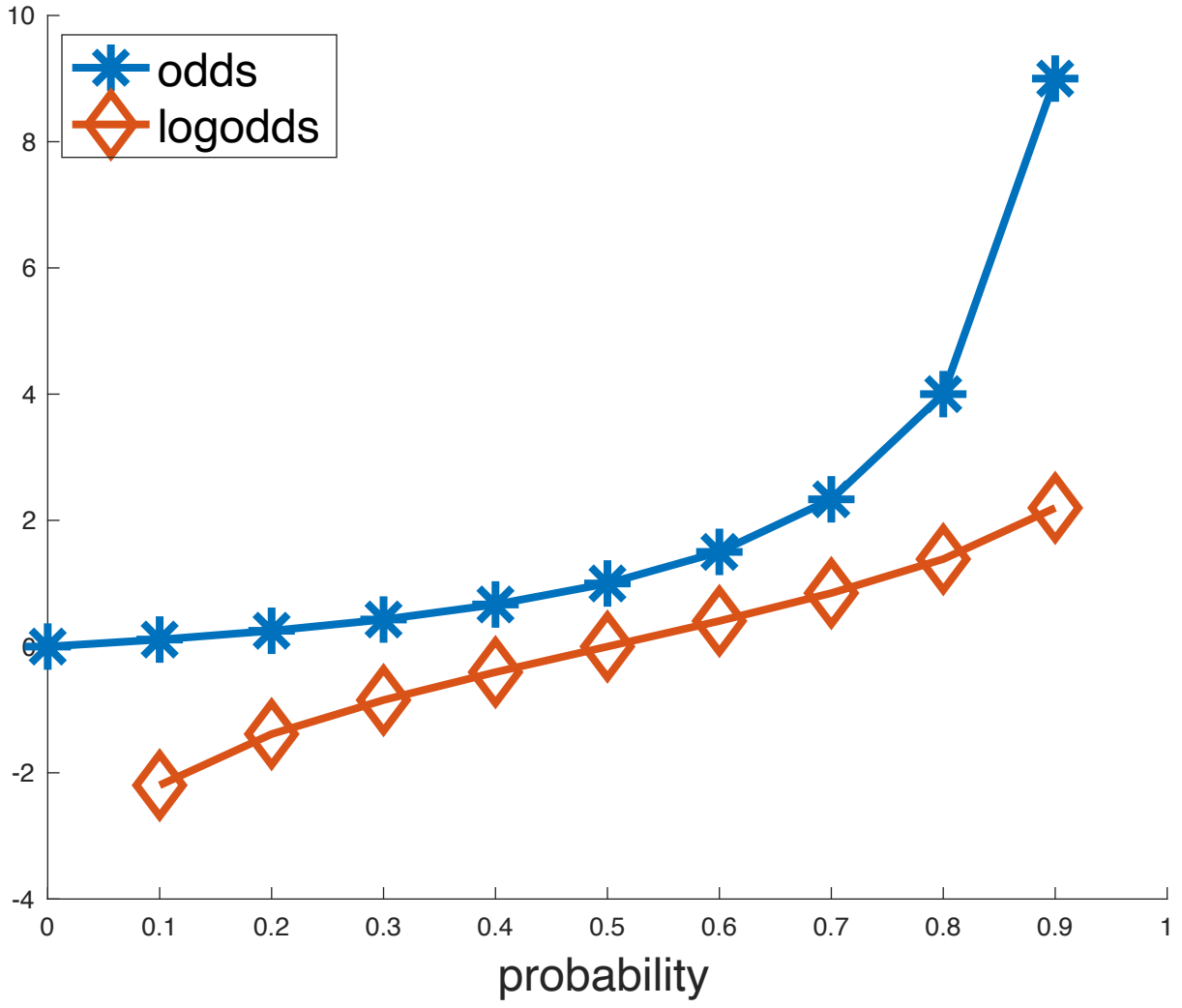
- La probabilità che si verifichi l'evento "persona acquista qualcosa" è $\frac{1}{3}$
- La probabilità dell'evento complementare è $1 - \frac{1}{3} = \frac{2}{3}$
- L'odds dell'evento "persona acquista qualcosa" è dunque

$$\frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{3}\frac{3}{2} = \frac{1}{2}$$

Studiamo le proprietà dell'odds

probability	odds	logodds
0.1	0.111111	-2.197225
0.2	0.250000	-1.386294
0.3	0.428571	-0.847298
0.4	0.666667	-0.405465
0.5	1.000000	0.000000
0.6	1.500000	0.405465
0.7	2.333333	0.847298
0.8	4.000000	1.386294
0.9	9.000000	2.197225

; double click to hide



Deriviamo la regressione logistica

Vogliamo stimare il logaritmo dell'odds usando un modello di regressione lineare

$$log_e(\frac{p}{1-p}) = \alpha^0 + \alpha^1 x$$

$$\frac{p}{1-p} = e^{\alpha^0 + \alpha^1 x}$$

$$p = e^{\alpha^0 + \alpha^1 x}(1-p) \qquad p = \frac{e^{\alpha^0 + \alpha^1 x}}{1 + e^{\alpha^0 + \alpha^1 x}}$$

UniGe | MaLGa

Classificazione Bayesiana naïve

Ricordate il teorema di Bayes?

- Supponiamo che H sia l'ipotesi relativa a determinati dati e che D siano i dati disponibili
- Possiamo usare il teorema di Bayes per ottenere la probabilità che la nostra ipotesi sia corretta sulla base dei dati disponibili

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}$$

- Qual è la probabilità che la mia ipotesi sia vera sulla base dei dati che ho?

Ricordate il teorema di Bayes?

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}$$

- P(H) è la probabilità dell'ipotesi (probabilità a priori)
- P(H|D) è la probabilità dell'ipotesi dopo aver osservato i dati (probabilità a posteriori)
- P(D|H) è la probabilità dei dati sotto l'ipotesi data (probabilità)
- P(D) è la probabilità dei dati sotto qualsiasi ipotesi (costante di normalizzazione)

Bayes e classificazione

- Possiamo usare il teorema di Bayes per classificare un dato come appartenente o no ad una certa classe calcolando

$$P(Classe|x) = \frac{P(x|Classe)P(Classe)}{P(x)}$$

Esempio

In questa tabella consideriamo la variabile Play

“in funzione” del meteo

Proviamo ad applicare l’algoritmo

di Naïve Bayes partendo dal costruire

una tabella delle frequenze

Weather	Play
Rainy	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Sunny	No
Rainy	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Sunny	No
Sunny	Yes
Rainy	No
Overcast	Yes
Overcast	Yes

Tabella delle frequenze

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

Dalle probabilità alla classificazione

Weather	No	Yes	
Overcast	0	5	5/14= 0.35
Rainy	2	2	4/14=0.29
Sunny	2	3	5/14=0.35
All	4/14=0.29	10/14=0.71	

P(Yes|Sunny)= P(Sunny|Yes)*P(Yes)/P(Sunny)

P(Sunny|Yes)= 3/10= 0.3

P(Sunny)= 0.35

P(Yes)=0.71

So P(Yes|Sunny) = 0.3*0.71/0.35= **0.60**

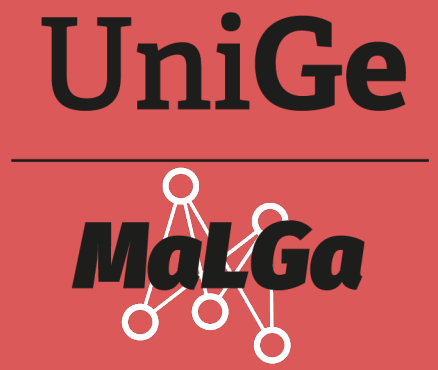
P(No|Sunny)= P(Sunny|No)*P(No)/P(Sunny)

P(Sunny|NO)= 2/4=0.5

P(No)= 0.29

P(Sunny)= 0.35

So P(No|Sunny)= 0.5*0.29/0.35 = **0.41**


UniGe
MaLGa