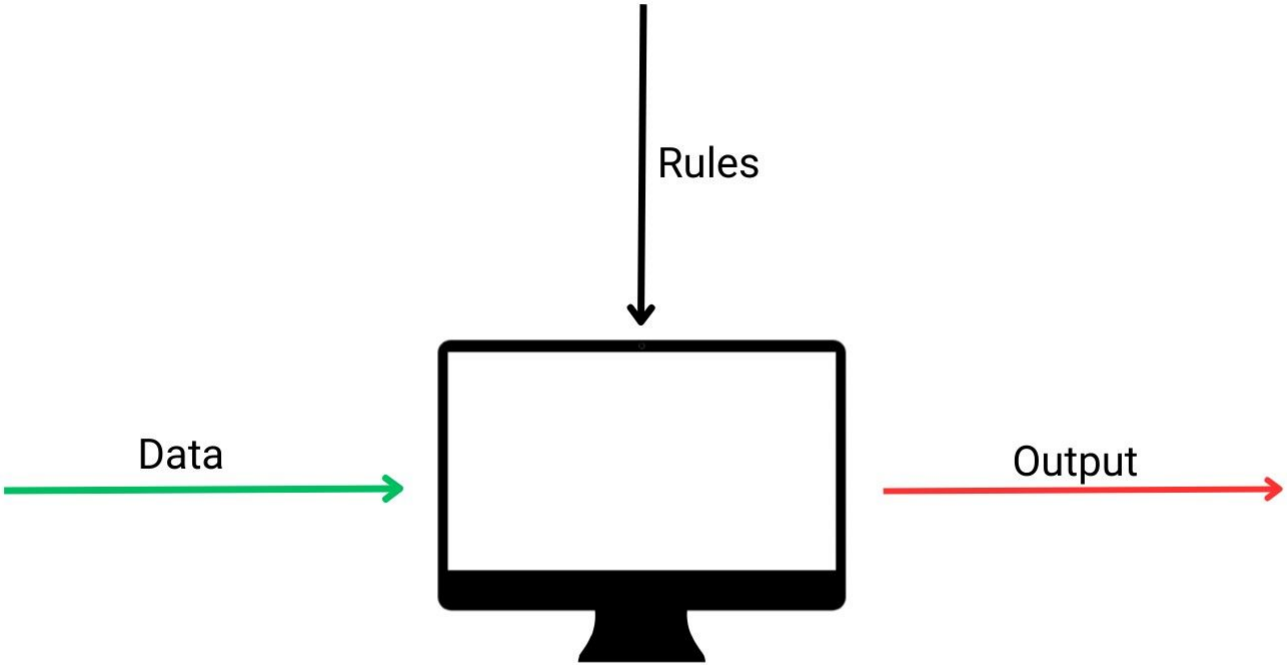
ANTICIPARE LA CRESCITA CON LE NUOVE COMPETENZE SUI BIG DATA – EDIZIONE 2

Operazione Rif. PA 2019-11596/RER "Anticipare la crescita con le nuove competenze sui Big Data", approvata dalla Regione Emilia-Romagna con DGR n° 789 del 20 maggio 2019 e co-finanziata dal Fondo Sociale Europeo PO 2014-2020

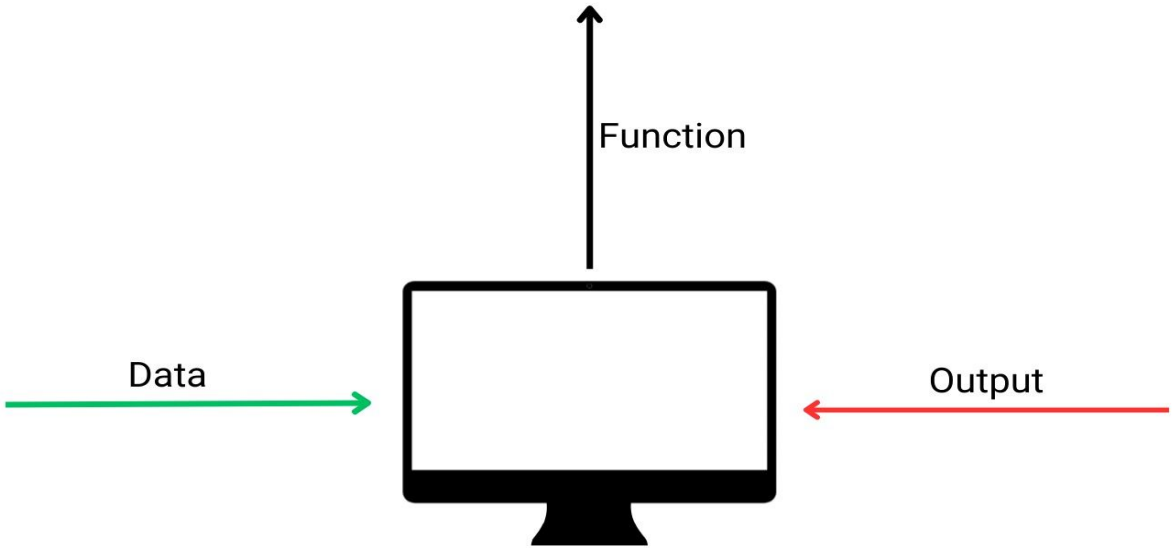
Programma della lezione

- Panoramica del Machine Learning
- 2 modelli di machine learning
  - KNN
  - SVM

Programmazione tradizionale



Machine Learning



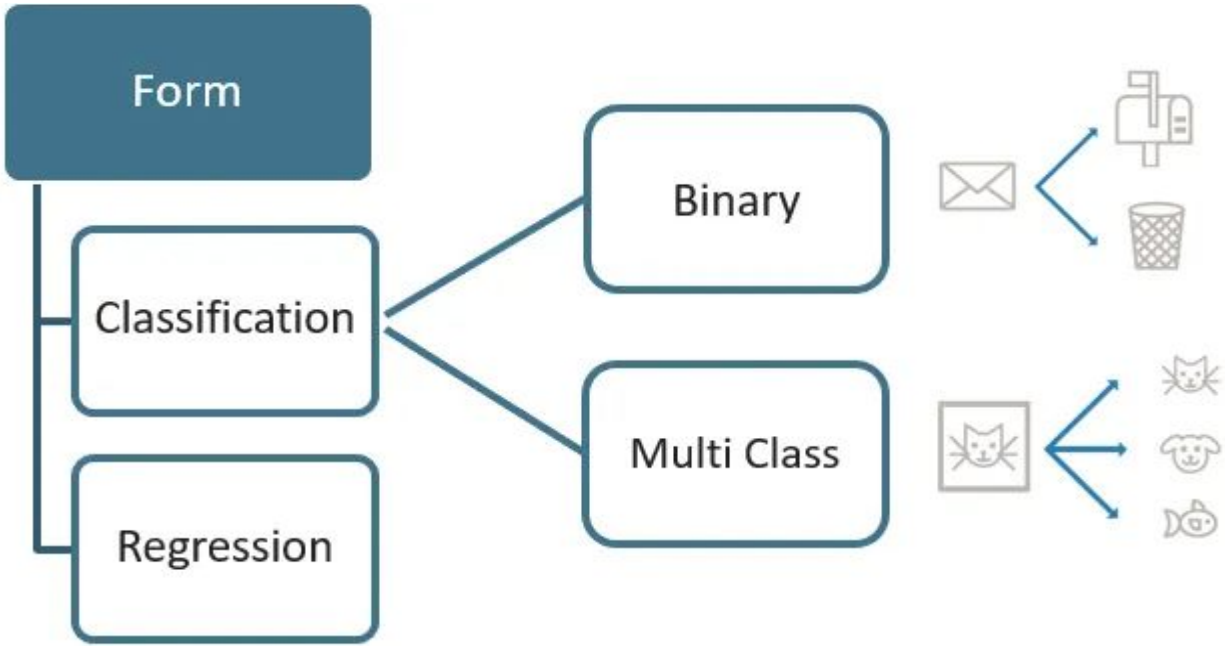
Machine Learning

$$y = f(x)$$

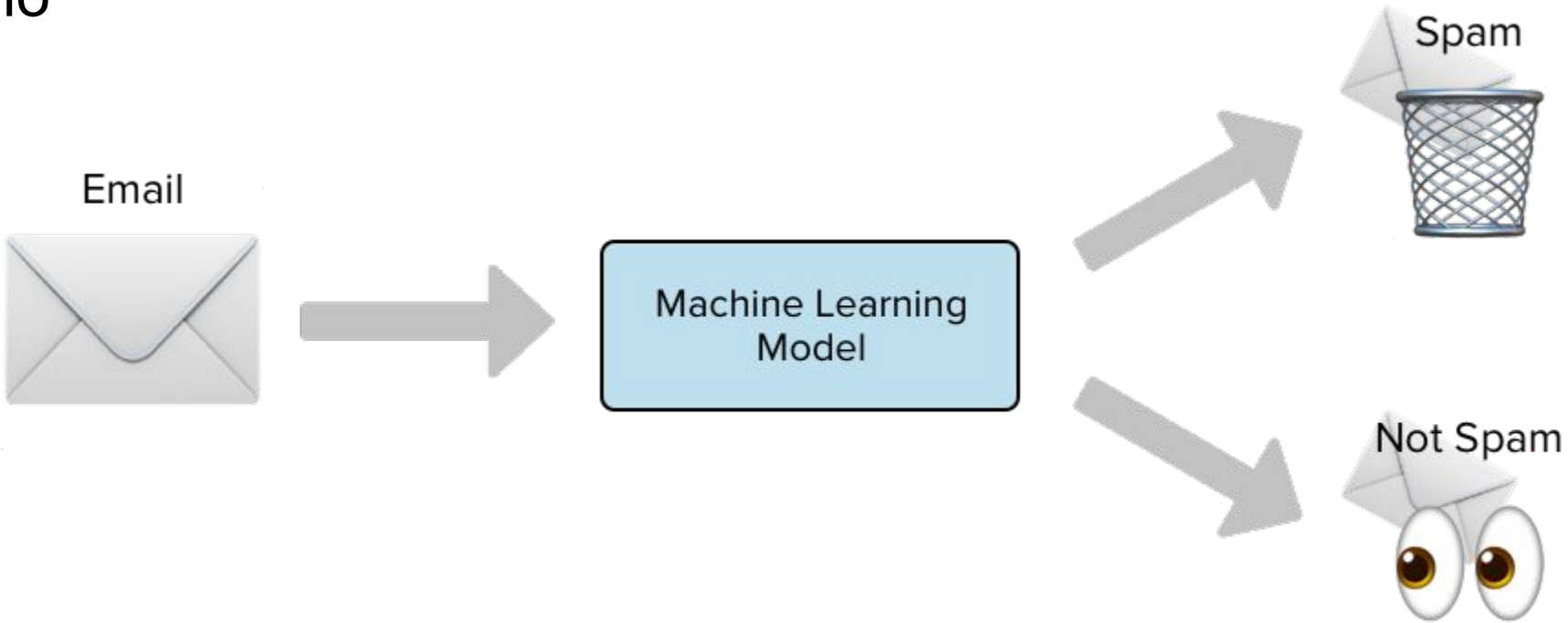
Machine Learning

$$y = f(x) + \varepsilon$$

Esistono due tipi di Supervised Learning

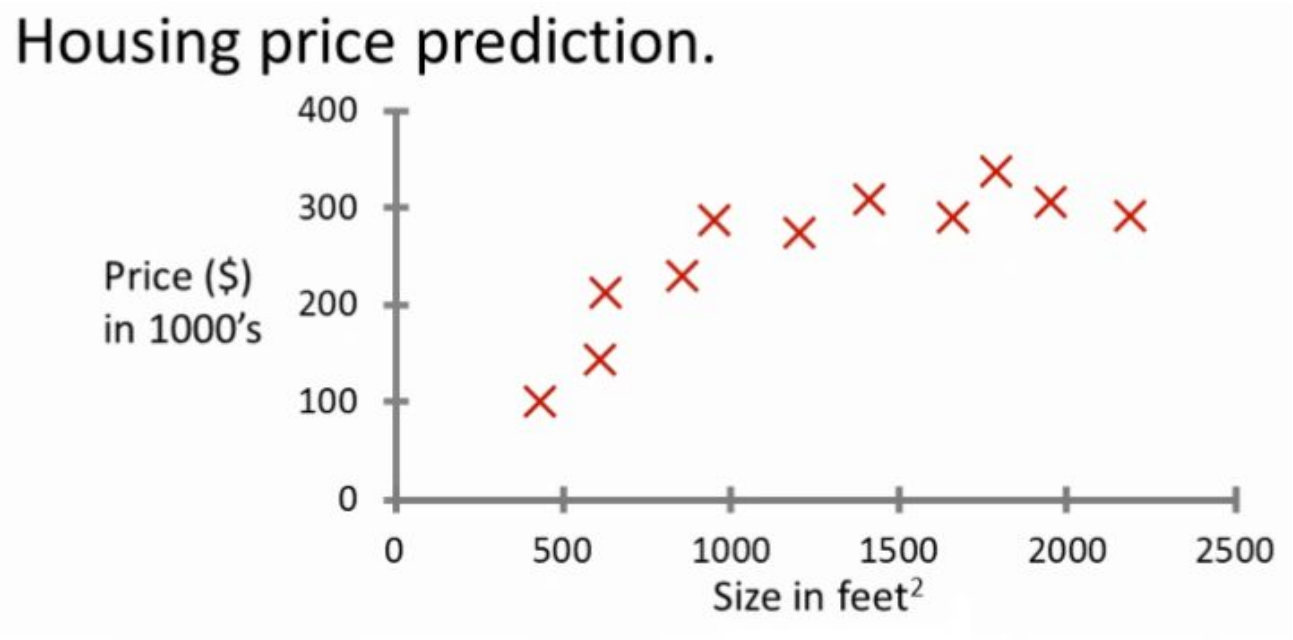


Esempio: vogliamo prevedere se un’email è uno spam o no



È un problema di regressione o classificazione?

Esempio: vogliamo prevedere il prezzo di una casa, data la sua dimensione

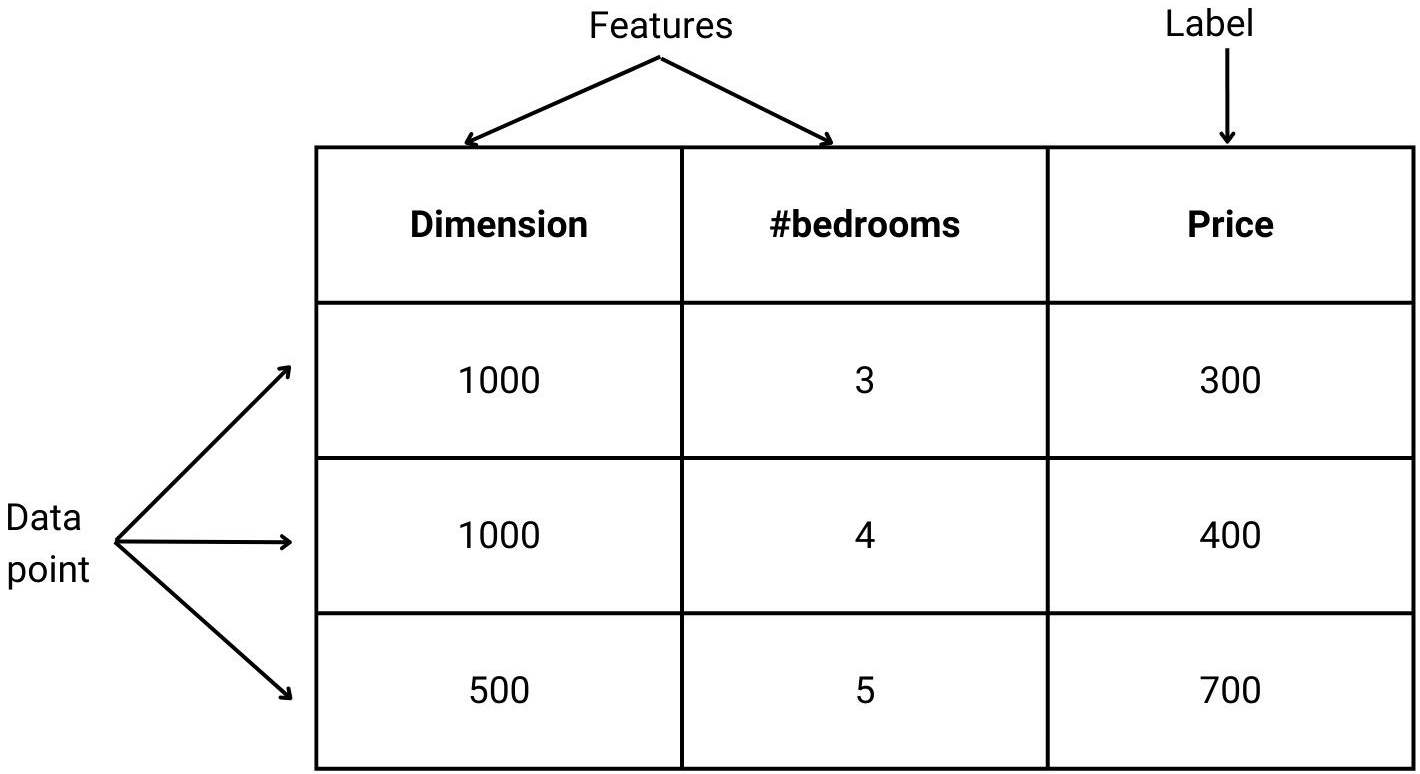


È un problema di regressione o classificazione?

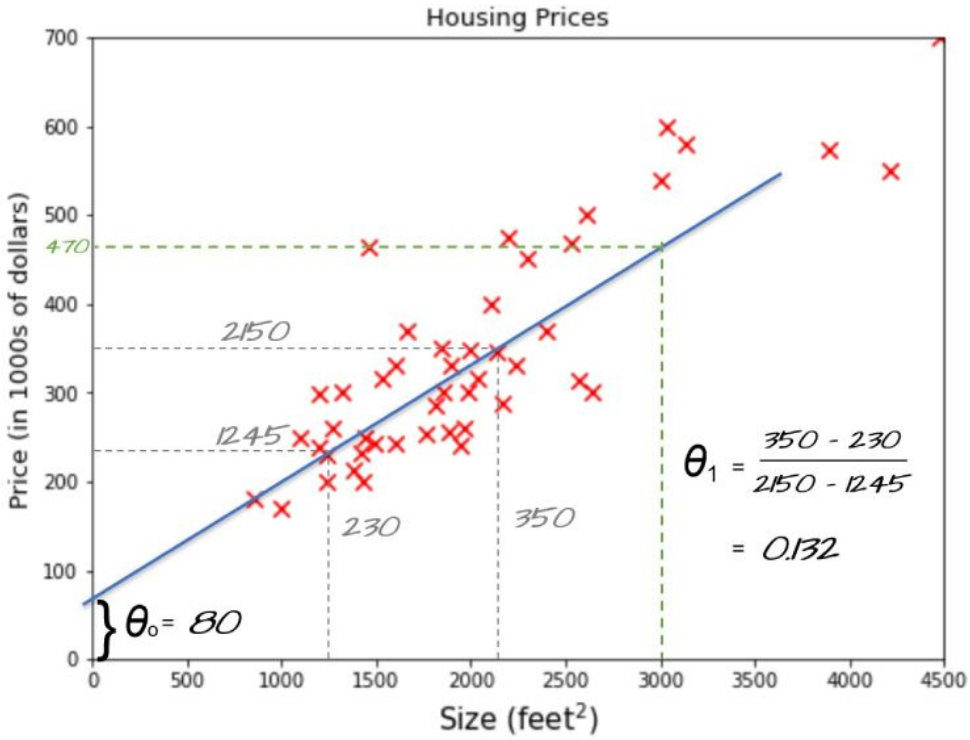
Alcune notazioni:

- Il valore che vogliamo prevedere (prezzo) si chiama **label/target**
- I valori che uso come input (dimensione) si chiamano **features**
- Ogni campione/record si chiama **data point**

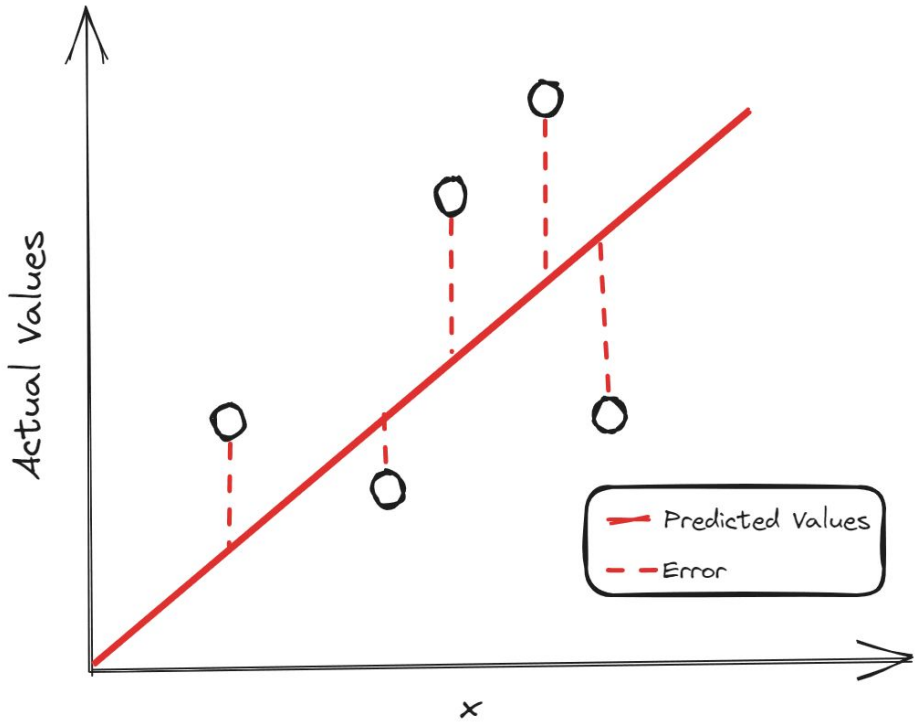
Dimension	#bedrooms	Price
1000	3	300
1000	4	400
500	5	700


Features
Label
Dimension
#bedrooms
Price
1000
3
300
1000
4
400
500
5
700
Data point

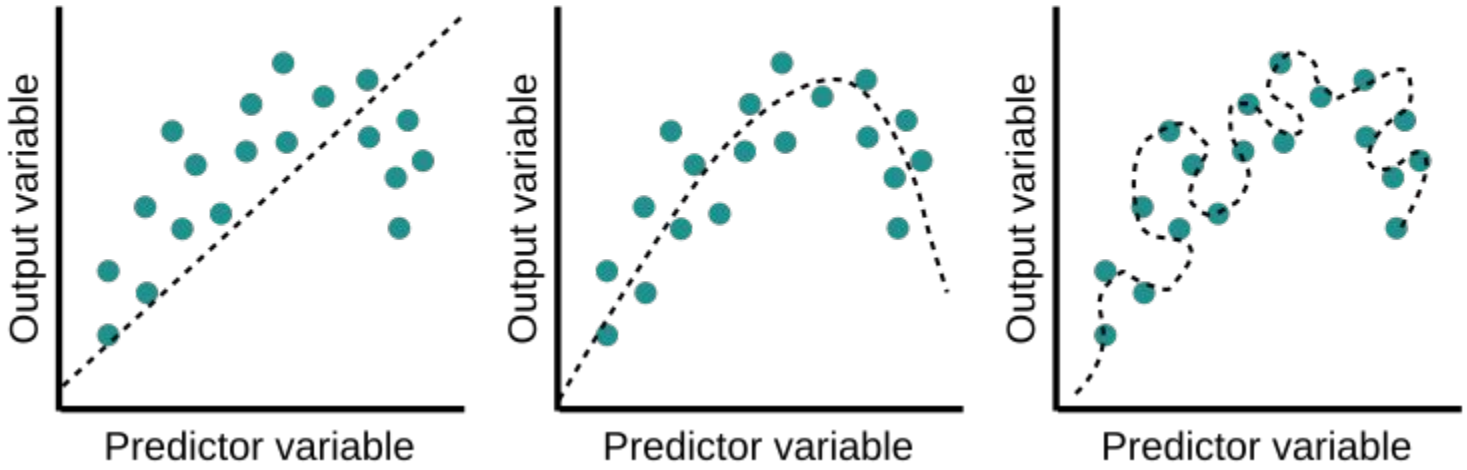
Esempio: funzione di fitting



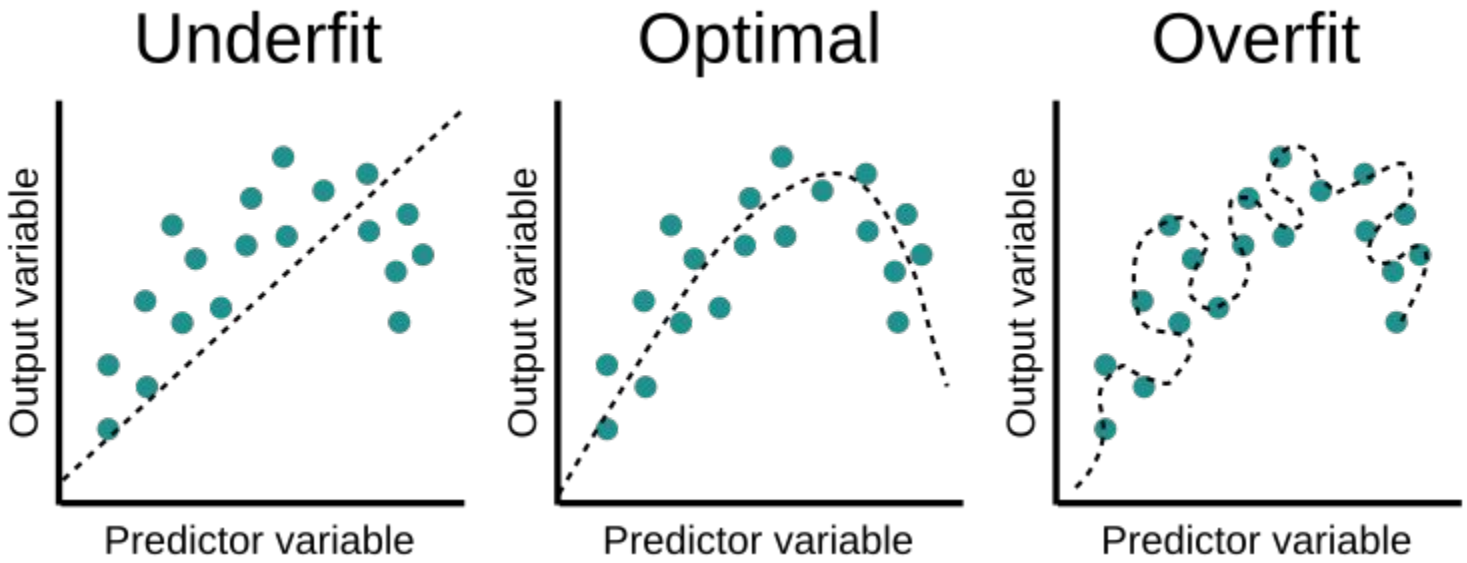
Come valutiamo il nostro modello? Funzione di costo



Quali di questi modelli ha il costo migliore?



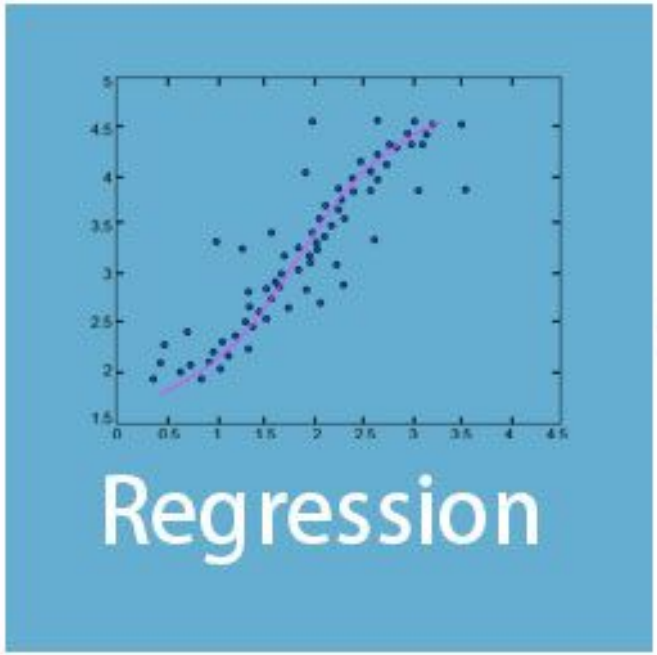
Quali di questi modelli ha il costo migliore?



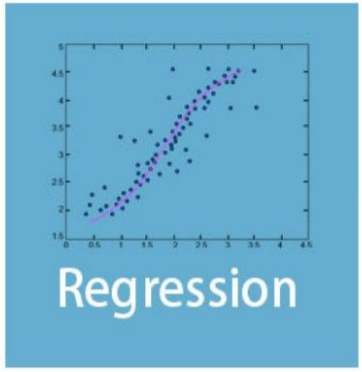
Soluzione: dividiamo i nosti dati in un “training set” e un “test set”

A cosa ci servono il training e il test set?

- **Training set** serve per addestrare il modello
- **Test set** si usa per valutare per il modello addestrato
  - per valutare la performance del modello su dati che non ha mai visto
  - Per valutare la sua abilità di generalizzare

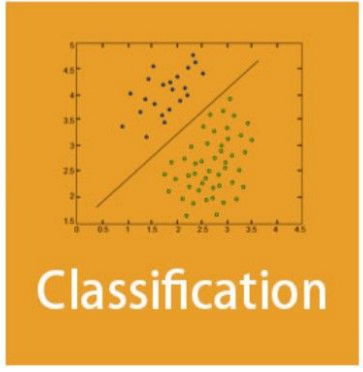

Regression
VS
Classification

Supervised Learning



KNN

SVM

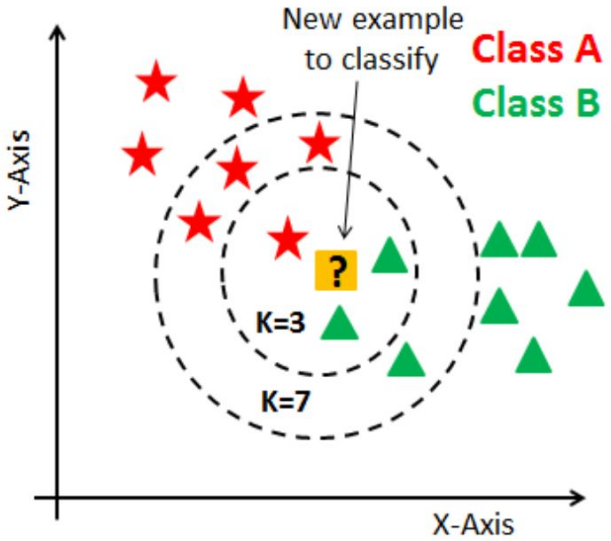


KNN

SVM

K Nearest Neighbors

- È un modello statistico utilizzato sia per la classificazione che per la regressione
- K si riferisce al numero di vicini più prossimi alla nuova osservazione, di cui si vuole fare la previsione



KNN - Fasi

1. Preprocessing: Normalizzazione
2. Scegliere il valore di k
3. Definire la metrica di distanza
4. Calcolare la distanza da tutti i punti nel training set
5. Ordinare le distanze e selezionare i K punti più vicini
6. Assegna
   a. Classe più frequente tra i k punti più vicini alla nuova osservazione
   b. Media (o mediana) dei valori più vicini alla nuova osservazione

Distance functions

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

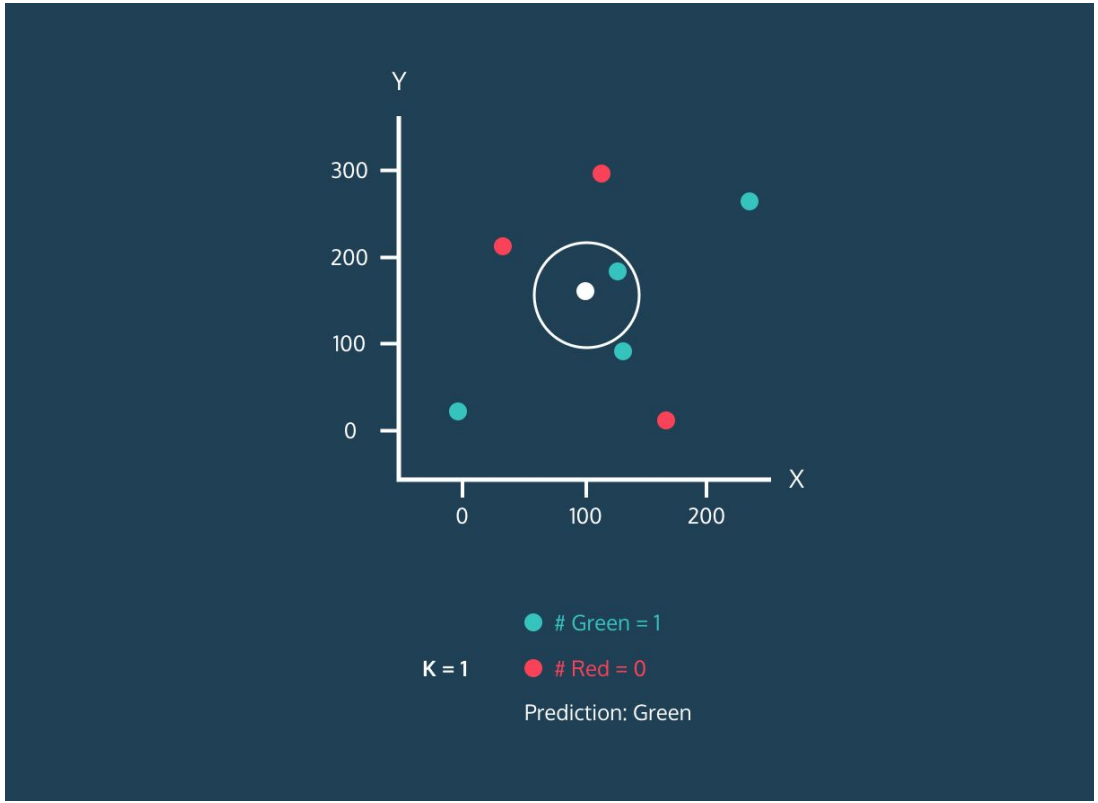
Manhattan

$$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski

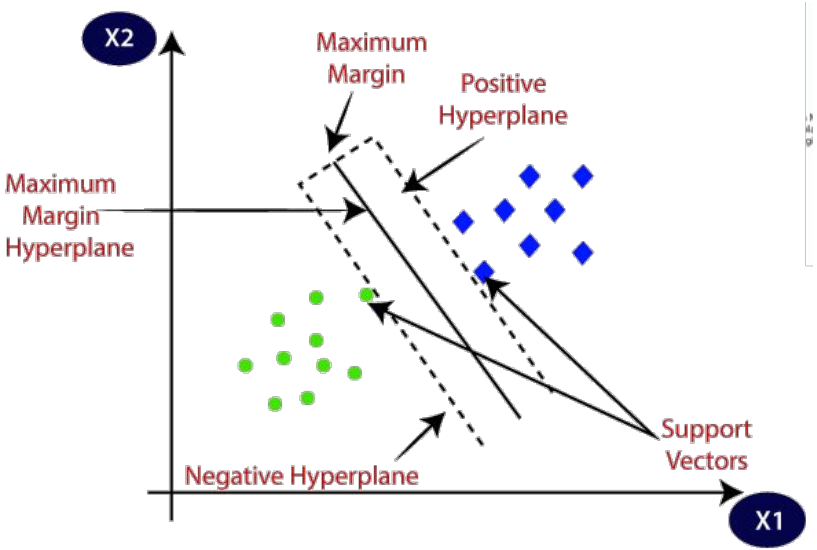
$$\left(\sum_{i=1}^{k}\left(|x_i - y_i|\right)^q\right)^{1/q}$$

K Nearest Neighbors



Support Vector Machines

- È un modello supervisionato utilizzato sia per la classificazione che per la regressione
- Caratteristiche:
  - SVM cerca un iperpiano che massimizzi il margine tra le due classi
  - Kernel-trick: Support Vector Machine utilizza una tecnica chiamata **kernel trick** per trasformare i dati non linearmente separabili in uno spazio in cui possono essere linearmente separabili

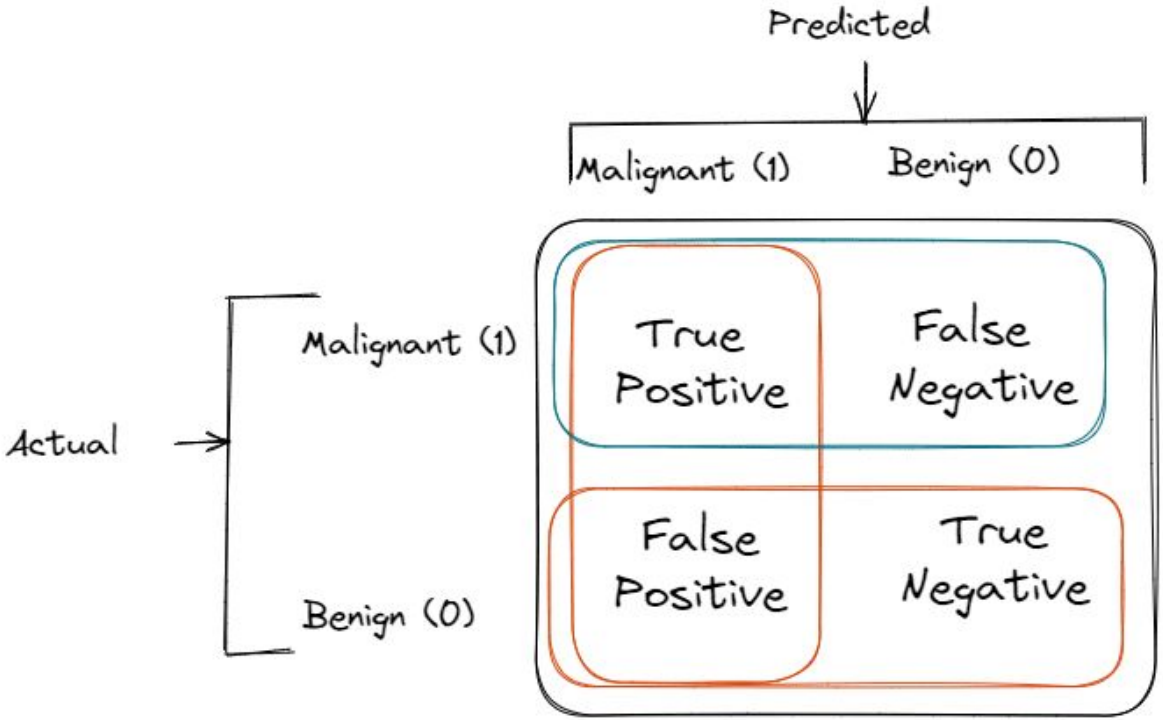


Support Vector Machines

Come valuto le performance del mio modello

- Classificazione
  - Accuratezza
  - Precision
  - Recall
  - F1-Score
- Regressione
  - MAE
  - MSE
  - RMSE

Confusion Matrix



Accuracy

$$\text{Accuracy} = (\text{True Positives} + \text{False Negatives}) / \text{Total Cases}$$

- Si usa quando il dataset è bilanciato
  - Esempio: la classe ‘tumore benigno’ deve avere lo stesso numero di osservazioni della classe ‘tumore maligno’
- Ci poniamo la domanda: Il nostro modello ha fatto le previsioni corrette per entrambe le classi?

Precision

$$\text{Precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives})$$

- Si usa quando
  - il dataset è **sbilanciato**
  - Si vuole minimizzare il numero di falsi positivi
- Falso positivo significa che il nostro modello dice che il paziente ha un tumore maligno, ma in realtà non ha un tumore maligno

Recall

$$\text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives})$$

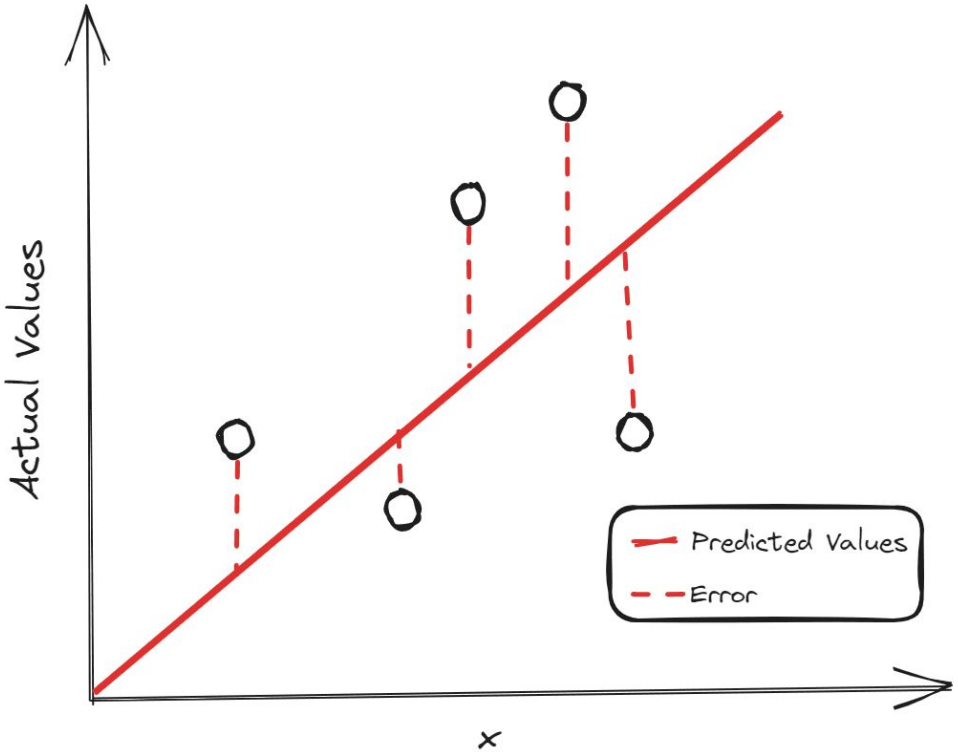
- Si usa quando
  - il dataset è **sbilanciato**
  - Si vuole minimizzare il numero di falsi negativi
- Falso negativo significa che il paziente ha un tumore maligno, ma il modello non lo ha identificato

F1-score

$$F1\text{-}Score = 2 \times (Precision \times Recall) \ / \ (Precision + Recall)$$

- Si usa quando
  - il dataset è **sbilanciato**
  - Si vuole una misura che riassume Precision e Recall
-

Regressione



MAE

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

- Differenza in valore assoluto tra i valori delle previsioni e i valori attuali della variabile target

MSE

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Differenza al quadrato tra i valori delle previsioni e i valori attuali della variabile target
  - Gli errori grandi peseranno di più

RMSE

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

- Radice quadrata del MSE
  - Gli errori grandi peseranno di più
  -

Risorse utili

- Statistical Learning di Trevor Hastie
- Statquest