

# Machine learning

## Fasi: algoritmo ML supervisionato

- 1) Suddivisione dei dati in train e test
- 2) Scelta del modello che dipende da una serie di parametri
- 3) Identificazione dei parametri durante la fase di training (imparando dai dati) utilizzando i dati target.
- 4) Applicazione del modello con i parametri stimati ai dati di test per verificare, tramite misure di errore sempre utilizzando i dati target, l'efficacia del modello ottenuto.
- 5) Applicazione del modello a dati non appartenenti al data set. (GENERALIZZAZIONE)

Train 70 / 80 %

Test 30 / 20 %

- Sono scelti casualmente

Train set → è il sottoinsieme di dati utilizzato per addestrare il modello ML.

Test set → campione di dati utilizzato per fornire una valutazione imparziale dell'addestramento del modello finale al set di dati di addestramento

Oversetting → si verifica quando un modello ML apprende troppo bene i dettagli presenti nei dati di addestramento, adattandosi in modo eccessivo a questi dati.

Di conseguenza, il modello ottiene prestazioni eccellenti sui dati di training, ma fallisce nel generalizzare su dati nuovi o di test.

Underfitting → si verifica quando un modello troppo semplice o non riesce a cogliere la complessità e le relazioni sottostanti nei dati.

In questo caso il modello mostra scarse prestazioni sia sui dati di addestramento sia su quelli di test.

Class Imbalance → accade quando una delle due classi è significativamente meno rappresentata rispetto all'altra.

Accuratezza → Percentuale Predizioni corrette  $\frac{TP+TN}{FN+TN+FP+TP}$

## Data Science e Classificazione

### Tipi di Variabili

- ↳ Numeriche o Quantitative: Variabili rappresentanti numeri.
  - Accettano un ordinamento naturale
  - Su di esse possono essere eseguite operazioni.  
es. Temperatura
- ↳ Categroiche o Qualitative: Variabili rappresentanti un insieme finito di etichette  
es. Genere

		True class		Measures
		Positive	Negative	
Predicted class	Positive	True positive $TP$	False positive $FP$	Positive predictive value (PPV) $\frac{TP}{TP+FP}$
	Negative	False negative $FN$	True negative $TN$	Negative predictive value (NPV) $\frac{TN}{FN+TN}$
Measures	Sensitivity	Specificity	Accuracy	$\frac{TP+TN}{TP+FN+FP+TN}$

Target o output → insieme delle variabili che vogliamo prevedere

Variabili input → variabili contenenti le informazioni che possiamo sfruttare per prevedere le variabili di output

Sia le variabili di input  $x$  che di output  $y$  siano dei vettori.

$$x = (x_1, x_2, \dots, x_p) \quad y = (y_1, y_2, \dots, y_m)$$

→ Rappresenta

Una Colonna

differente del dataset.

- **Problema Univariato** →  $x$  è una variabile scalare  
↳ La variabile contiene un solo valore
- **Problema Multivariato** → se  $x = (x_1, x_2, \dots, x_p)$  → vettoriale

## Categorie di Problemi di ML

**Apprendimento Supervisionato** → sono presenti sia la variabile di input  $x$  sia quella di output  $y$

**Apprendimento Non-Supervisionato** → la variabile di output  $y$  non è presente nel dataset.  
↳ Vogliacuo cercare delle strutture dette cluster tra gli input

**Apprendimento Semi-Supervisionato** → quando le variabili di output  $y$  sono presenti soltanto in una frazione ridotta delle osservazioni.

Noi supponiamo sempre problemi di tipo Supervisionato

### Tra i problemi di tipo Supervisionato:

- **Regressione** → quando la variabile di output  $y$  è di tipo numerico.
- **Classificazione** → quando la variabile di output  $y$  è di tipo categorico.

Regressione Multivariata → es. Sia possibile prevedere la temperatura dato il mese.

Per ciascuna osservazione,  $x \in \mathbb{R}^p$ ,  $y \in \mathbb{R}$   
input (features)      output (target)

- La relazione tra  $x$  e  $y$  è modellizzata da una funzione  $h: \mathbb{R}^p \rightarrow \mathbb{R}$  detta predittore

**Predittore** → funzione che associa ad ogni variabile  $x$  la corrispondente  $y$

$h_{\theta}$     ↳ Il predittore dipenderà da un'insieme di parametri ( $\theta$ ). La relazione tra il predittore e i suoi parametri definiscono l'algoritmo

↳ Obiettivo:  $h_{\theta}(x_i) \approx y_i \quad \forall i=1, \dots, N$

↳ fissando una misura di errore → funzione di loss  $l(y, \hat{y})$

La funzione di loss → misura l'errore per ogni singolo addestramento

$l(h_{\theta}(x_i), y_i)$     ↳ Quanto sbaglia il modello  $h_{\theta}$  su questa singola coppia input-output

## Empiric Risk minimization (ERM)

Minimizzazione del rischio Empirico  $\rightarrow$  E' una strategia o un principio di apprendimento.

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(h_{\theta}(x_i), y_i)$$

$\hookrightarrow$  e' il processo di trovare i parametri del modello  $\theta$  che funzionano meglio sul dataset selezionato.

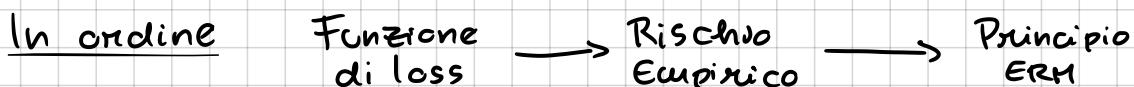
$\hookrightarrow$  Addestramento (training)

### Funzione di rischio Empirico

$\hookrightarrow$  E' una funzione macroscopica. E' la media delle funzioni di loss su tutto il dataset di addestramento.

$$R(\theta) = \frac{1}{N} \sum_{i=1}^N l(h_{\theta}(x_i), y_i) \rightarrow$$
 e' la funzione obiettivo che il principio ERM ci dice di minimizzare

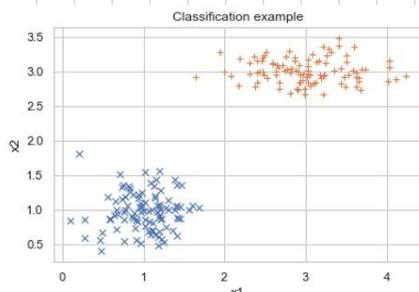
- Rappresenta l'errore "empirico" ovvero sui dati osservati del nostro modello.



## Problemi di Classificazione

- Problema in cui la variabile di output  $y$  e' di tipo Categorica.  $\rightarrow$  Numero finito di valori detti classi  $C_k$

Diagramma Bidimensionale dei valori  $x = \{x_1, x_2\}$  contenuti in  $S$



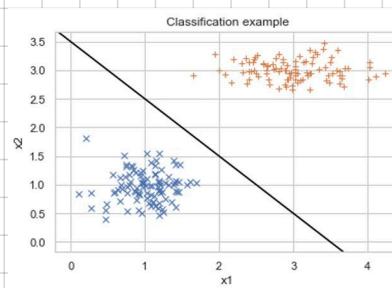
Come faccio ad implementare un algoritmo che, sfruttando i dati in  $S$ , impara a classificare dei nuovi punti come '+' o 'x', conoscendo  $x_1$  e  $x_2$ ?

$\hookrightarrow$  retta che separa le due classi

Predittore lineare  $\rightarrow$  Predittore che separa le classi attraverso una retta.

$$h_{\theta}(x_1, x_2) = \begin{cases} + & \text{se } ax_1 + bx_2 + c \geq 0 \\ - & \text{se } ax_1 + bx_2 + c < 0 \end{cases}$$

Retta Separatrice  $\rightarrow$  retta di equazione  
 $\square \quad ax_1 + bx_2 + c = 0$  che separa perfettamente le due classi



Dataset linearemente Separabile  $\rightarrow$  Esiste almeno un retta separatrice  $\sqcap$ .

## Regressione Logica

↳ Classificatore lineare si basa su una sigmoide  $\sigma(x) = \frac{1}{1+e^{-x}}$

### Predittore di una Regressione Logistica

$$h_{\theta}(x_1, x_2) = \frac{1}{1+e^{-(ax_1+bx_2+c)}} \quad \text{risultato } [0,1]$$

Probabilità che  $h_{\theta}$  attribuisce al punto  $(x_1, x_2)$  di appartenere alla classe 1.

$$\begin{cases} 1 & \text{se } h_{\theta}(x_1, x_2) > 0.5 \\ 0 & \text{se } h_{\theta}(x_1, x_2) \leq 0.5 \end{cases}$$

## Classificatore a Massimo Margine (MMC)

↳ Se  $S$  è linearmente separabile, allora esistono infinite rette separatrici

↳ Quale scegliamo?

- Quella che massimizza la distanza tra le due classi

$$\Pi = \max_{\Pi} \min_{i=1,2} d(\Pi, C_i)$$

- Definita la retta separatrice

- Classificatore a Massimo Margine (MMC) → individua la migliore distanza fra le due classi

$$h_{\theta}(x_1, x_2) = \begin{cases} + & \text{se } ax_1+bx_2+c > 0 \\ - & \text{se } ax_1+bx_2+c \leq 0 \end{cases}$$

→ I dati non sono mai linearmente separabili, la maggior parte delle volte le classi si sovrappongono

## Support Vector Classifier (SVC)

- Non esiste nessuna retta separatrice tra le due classi. → Aggiungo il parametro costo

Parametro costo → Parametro che controlla il numero massimo di input che possono oltre passare la linea di separazione

$$\hookrightarrow C > 0$$

Nuova retta  $\Pi(C) \rightarrow$  nuovo predittore  $h_{\theta}(x_1, x_2) \rightarrow$  Support Vector Classifier (SVC)

## Caso p-dimensionale

Iperpiano  $\rightarrow$  In  $\mathbb{R}^p$ , un iperpiano è il luogo dei punti che rispettano l'equazione

Equazione Lineare  $a_0 + \sum_{i=1}^p a_i x_i = a_0 + a_1 x_1 + \dots + a_p x_p = 0$

## Iperpiano Separatore

Dato un problema di classificazione di C classi su un dataset S, un iperpiano  $\Pi$  di equazione  $a_0 + \sum_{i=1}^p a_i x_i = 0$  che separa perfettamente le due classi si detto iperpiano separatore.

Linearmente Separabile  $\rightarrow$  se esiste almeno un iperpiano separatore  $\Pi$

Conseguenze  $\rightarrow$  Aggiungiamo MHC e SVC per il caso p-dimensionale

MHC  
p-dimensionale  $f_\theta(x) = \begin{cases} + & \text{se } a_0 + \sum_{i=1}^p a_i x_i \geq 0 \\ - & \text{se } a_0 + \sum_{i=1}^p a_i x_i < 0 \end{cases}$

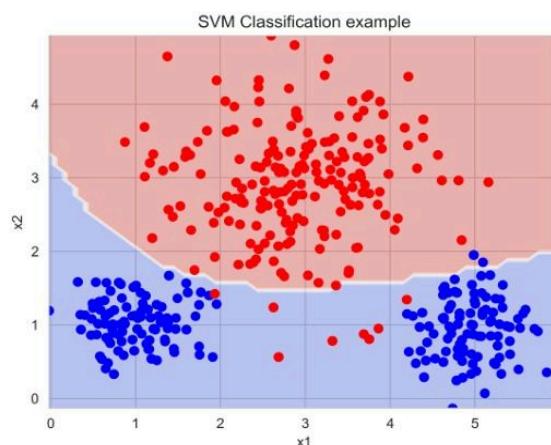
Iperpiano a massimo margine  $\rightarrow a_0 + \sum_{i=1}^p a_i x_i = 0$

## Dati non linearmente separabili

Fino ad adesso abbiamo sempre visto classi separate da una retta, incluso MHC e SVC.

Ci sono casi in cui esistono delle curve "semplici" in grado di separare perfettamente le due classi.  
 $\downarrow$   
POLINOMIO

Esempio di dati separabili da un polinomio.



## Support Vector Machines (SVM)

- Per migliorare l'algoritmo di classificazione.
- Fissiamo una funzione non lineare  $\phi: \mathbb{R}^p \rightarrow \mathbb{R}^d$  t.c.  $\phi(S)$  sia linearmente separabile. dataset ottenuto trasformando  $S$  con  $\phi$

### Tipi di Kernel Function $\phi(x)$

- linear  $\phi(x) = x \rightarrow \text{SVC}$
- poly  $\phi(x) = (1+x)^d \rightarrow \text{Kernel polinomiale}$   
 $\rightarrow d = \text{grado del polinomio (degree)}$
- rbf  $\phi(x) = \exp\left(-\frac{x^2}{\gamma}\right) \rightarrow \text{Kernel radiale (esponenziale)}$   
 $\rightarrow \gamma = (\text{gauca}) \text{ variazione della distribuzione}$
- Sigmoid:  $\phi(x) = \frac{1}{1 + \exp(-x)} \rightarrow \text{Kernel iperbolico}$

→ Fissata la Kernel function  $\phi(x)$

↳ lineare separabilità di  $\phi(S) \rightarrow$  addestriamo un SVC sul dataset trasformato

iperpiano  $\Pi(c)$   
di equazione

$$\alpha_0 + \sum_{i=1}^p \alpha_i \phi(x_i) = 0$$

- Con Costo ( $C > 0$ )

$$h_\theta = \begin{cases} + \text{ se } \alpha_0 + \sum_{i=1}^p \alpha_i \phi(x_i) \geq 0 \\ - \text{ se } \alpha_0 + \sum_{i=1}^p \alpha_i \phi(x_i) < 0 \end{cases}$$

La curva di separazione non è più una retta in  $(\mathbb{R}^p)$ , ma una curva la cui forma dipende dalla funzione del Kernel  $\phi$ .



## Regressione Lineare

- Rappresenta una classe dei problemi supervisionati
- Variabile target di tipo numerico
- Studia la relazione fra due o più variabili, che sono legate in modo NON DETERMINISTICO, per fare inferenze sul modello.

Inferenze → usare i dati che hai a disposizione (campione) per trovare conclusioni generali sulla relazione vera che esiste nella popolazione da cui proviene il campione.

Si usano relazioni fra due o più variabili in modo da potere avere informazioni su una di esse conoscendo i valori dell'altra

• Nel caso lineare supponiamo una relazione lineare fra le variabili  $x$  e  $y$ .

$x \rightarrow$  variabile non casuale  
 $\hookrightarrow$  es. tempo di studio

$$Y = \beta_0 + \beta_1 x$$

$y \rightarrow$  variabile casuale  
 $\hookrightarrow$  es. voto all'esame

Relazione  $\rightarrow$  sarebbe deterministica ma è probabilistica

$\hookrightarrow$  Probabilistica  $\rightarrow$  necessaria quando abbiamo  $n$  variabili e quindi le variabili non hanno una relazione deterministica

Obiettivo  $\rightarrow$  Prendere un valore futuro di  $y$  per un particolare valore di  $x$

$x \rightarrow$  variabile indipendente

$y \rightarrow$  variabile dipendente

In corrispondenza di  $n$  variabili indipendenti  $(x_1, x_2, \dots, x_n)$  si fanno  $n$  valori  $(y_1, y_2, \dots, y_n)$  di variabili indipendenti che sono legati: dalla relazione:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i=1, \dots, n$$

$\varepsilon_i \rightarrow$  Errore Casuale

$\hookrightarrow$  di quanto differisce rispetto al modello esatto

$\hookrightarrow$  distribuzione normale con media 0 e deviazione standard  $\sigma$ . norm(0,  $\sigma^2$ )

### Modello di regressione lineare Semplice

Esistono parametri  $\beta_0, \beta_1, \sigma^2$  t.c. il valore fissato dalla variabile indipendente  $x$ , la variabile dipendente è una variabile aleatoria legata ad  $x$  dal modello:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Come determinare, fra le infinite rette del piano, una buona retta?

$\varepsilon_i \rightarrow$  distribuzione Normale  $\Rightarrow Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$  ha una distribuzione normale  $\sigma = \sigma$

Riprendo la funzione MLE per trovare i parametri.

Funzione di verosimiglianza  $\rightarrow$  rappresenta la probabilità di osservare i dati campionari dati quei parametri

$\hookrightarrow$  Per trovare i parametri  $\beta_0$  e  $\beta_1$  risolvendo la funzione

## Calcoli

TROVO  $\hat{\beta}_0$  e  $\hat{\beta}_1$   
con MLE

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^n f_{Y_i}(x_i) \\ &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp \left\{ -\frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\} \end{aligned}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n Y_i \right) / n}{\sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 / n}$$

$\bar{Y}$  e  $\bar{x}$  sono la media dei valori  $y_i$  e  $x_i$ .

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

altro metodo

Principio dei quadrati minimi (trovo  $\hat{\beta}_0$  e  $\hat{\beta}_1$ )

↪ Detto residuo  $i$ -esimo ( $E_i$ ) la differenza verticale fra l'osservazione  $i$ -esima e la retta di regressione lineare  $(\beta_0 + \beta_1 x_i)$

$$E_i = Y_i - (\beta_0 + \beta_1 x_i)$$

V Funzione somma dei quadrati dei residui

$$f(\beta_0, \beta_1) = \sum_{i=1}^n E_i^2$$

$\hat{\beta}_0$  e  $\hat{\beta}_1$  si ottengono minimizzando  $f(\beta_0, \beta_1)$

Significato dei parametri della retta

$$Y_i = \beta_0 + \beta_1 x_i$$

Es.  $Y(x)$  rappresenta il modello di crescita di un bambino in cm in funzione dei mesi

$$Y(x) = 50 + 0.753 x \quad 0.753 \rightarrow x \text{ mese aumenta di } 0.753$$

I parametri  $\hat{\beta}_0$  e  $\hat{\beta}_1$  → sono variabili aleatorie e dipendono dal campione Considerato

$Y_i$  → variabile aleatoria → perché il processo che genera dati non è deterministico.

Inferenza sui Parametri

Test di  $H_0$  sul parametro più importante  $\hat{\beta}_1$ .

La retta di regressione è parallela all'asse  $x$  oppure no → significa che il valore della variabile aleatoria  $Y$  non cambia al variare di  $x$

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

## Valori Predetti

↳ In samplo → i valori della variabile indipendente in cui si fa la predizione sono nell'insieme dei dati a disposizione. Posso confrontare le predizioni con i valori osservati nelle stesse ascisse.

Out of sample → i valori della variabile indipendente NON sono nell'insieme dei dati a disposizione. Non ho quindi nessun termine di confronto per i valori che vengono predetti

## Coefficiente $R^2$

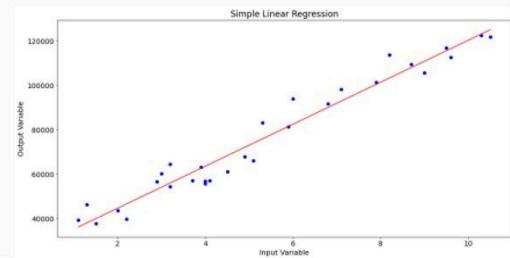
↳ Coefficiente semplice di determinazione → Singolo numero che mi dà indicazioni sulla bontà del modello regressione lineare semplice rispetto al campione di dati a disposizione

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

$y_i$  → valori osservati

$\hat{y}_i$  → valori stimati

$\bar{y}$  →  $\frac{1}{n} \sum_{i=1}^n y_i$  → media dei valori osservati



• Intervallo  $0 \leq R^2 \leq 1$

↳ vicino a 1 → modello di regressione lineare buono

↳ vicino a 0 → modello non rappresentativo del campione → MSE ↓

## Coefficiente semplice di correlazione

↳  $r_c$

$$\rightarrow |r_c| = \sqrt{R^2}$$

MSE → mean square error

↳ altra misura per giudicare la qualità di un stimatore  
↳ indica la discrepanza quadratica media fra i valori osservati dei dati ed i valori dei dati stimati

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad MSE \in [0, +\infty]$$

Bacca

errore  
medio  
quadratico  
nelle stime

⚠ Il coefficiente  $R^2$  da solo non ha significato se non c'è effettivamente una dipendenza lineare fra i dati.

## Analisi dei residui

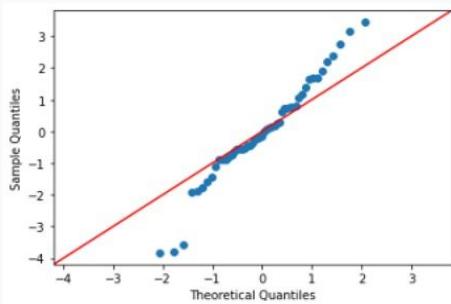
Residui → le differenze fra i valori osservati  $y_i$  e i valori predetti  $\hat{y}_i$  relativi alla stessa ascissa  $x_i$

$$\text{residui} = y_i - \hat{y}_i$$

→ Sui residui viene fatta l'ipotesi di normalità → Nel modello di regressione lineare, che i residui abbiano distribuzione normale con media 0 e ds  $\sigma$

QQ-pot  $\rightarrow$  Se le due distribuzioni sono simili, i punti devono stare molto vicini alla retta.

Test di ipotesi sulla normalità  $\rightarrow$  test di Shapiro-Wilk



$H_0$ : residui normali

$H_a$ : residui non normali

Si calcola il p-value

$p\text{-value} > 0.05 \rightarrow$  i residui seguono una distribuzione normale

$\rightarrow$  Non rifiutiamo  $H_0$

$p\text{-value} \leq 0.05 \rightarrow$  i residui non seguono una distribuzione normale

$\rightarrow$  Rifiutiamo  $H_0$