

Introduzione al Machine Learning: Tipologie di apprendimento

Vincenzo Bonifaci



- 1 Tipologie di apprendimento; demo
- 2 Rischio atteso ed empirico; libreria NumPy
- 3 Metodi di ottimizzazione; es. NumPy
- 4 Regressione lineare e varianti; es. NumPy/SciKit-Learn
- 5 Classificazione Nearest Neighbor; es. NumPy
- 6 Regressione logistica; es. NumPy
- 7 Modelli generativi; es. NumPy
- 8 Validazione e test

Assignment caricati su Teams dopo ogni lezione
(scadenza ultima: 18 settembre 2022)

Definizione di Machine Learning

Arthur Samuel, 1959

L'*apprendimento automatico* (o *machine learning*) è il campo di studi volto a fornire ai calcolatori l'abilità di apprendere [un compito] senza essere stati esplicitamente programmati [per quel compito].

Tom Mitchell, 1997

Un algoritmo *apprende* dall'esperienza E rispetto ad una classe di compiti T ed una misura di performance P se la sua performance sui compiti in T , così come misurata da P , migliora con l'esperienza E .

Machine learning vs. problemi computazionali classici

Esempio di problema computazionale “classico”

Input: un numero intero n

Output: la sua scomposizione in fattori primi

Relazione di input-output specificata in modo formale, matematico

Esempio di problema di machine learning

Input: foto di un animale

Output: nome dell'animale

Relazione di input-output specificata tramite **esempi**
(ingresso, uscita)

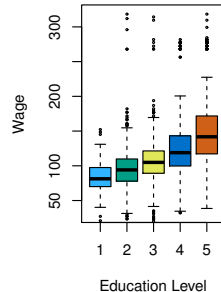
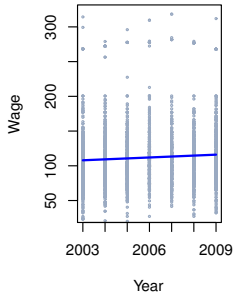
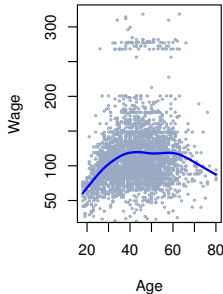
Problemi di apprendimento automatico: esempi

- Determinare la relazione tra salario e titolo di studio
- Identificare messaggi email indesiderati (*spam*)
- Identificare le cifre di un codice di avviamento postale scritto a mano
- Identificare transazioni bancarie fraudolente
- Raggruppare articoli di giornale in base all'argomento
- Raggruppare colture cellulari in base alla tipologia di cancro

Predizione del salario

Input: età, anno di calendario, e titolo di studio di un lavoratore

Output: salario del lavoratore



Sondaggio della popolazione maschile della regione centroatlantica degli USA

Identificazione di email spam

Input: testo di un messaggio email

Output: spam o email

Variabili di input: frequenze relative delle parole e segni di interpunzione più comuni in questi messaggi email

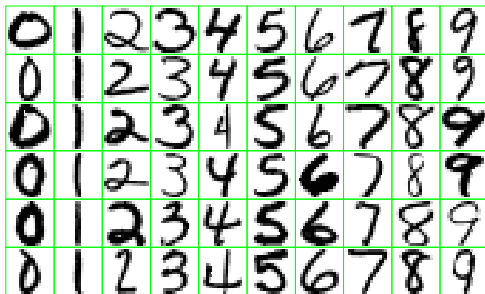
	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Percentuale di occorrenze di ciascuna parola nella classe spam e nella classe email

Riconoscimento di cifre scritte a mano: dataset MNIST

Input: immagine 28×28 pixel a scala di grigi

Output: una cifra da 0 a 9



Variabili di input: $28 \times 28 = 784$ interi tra 0 (nero) e 255 (bianco):
i primi 28 interi descrivono la luminosità dei pixel della prima riga,
i secondi 28 quella dei pixel della seconda riga, ecc.

Identificazione di transazioni bancarie fraudolente

Input: dettagli di una transazione su carta di credito (luogo, tipo di beneficiario, importo, POS/ATM, PIN/chip/striscia, ...)

Output: **probabilità** che la transazione sia fraudolenta



Raggruppamento di articoli di giornale

Input: testi di articoli di giornale

Output: raggruppamento degli articoli per argomento

The screenshot shows the Google News interface. At the top, there's a search bar with the text "Cerca argomenti, località e fonti". Below the search bar, there's a horizontal menu with "Notizie principali" selected. To the left of the main content area, there's a vertical sidebar with various filters: "Per te", "Stai seguendo", "Ricerche salvate", "COVID-19", "Italia", "Dal mondo", "Notizie locali", "Affari", "Scienza e tecnologia", and "Intrattenimento". The main content area displays a list of news articles. The first article is titled "Von der Leyen: 'Tutti in Ue devono avere salari minimi. Con Conte vertice sulla sanità in Italia'" and is from "Il Fatto Quotidiano". It has a sub-header "Ue, Von der Leyen traccia le linee per la ripartenza europea: priorità a sanità, clima e digitale" and is from "la Repubblica". The second article is titled "Il discorso di Ursula von der Leyen: «Il 37% del Recovery Fund per il clima»" and is from "Corriere della Sera". The third article is titled "Von der Leyen: 'E' il momento per l'Europa per allontanarsi da questa fragilità'" and is from "La Stampa". The fourth article is titled "Von der Leyen: 'Organizzeremo con Conte un vertice in Italia sulla sanità' | 'Tutti devono avere un salario minimo'" and is from "TGCOM". To the right of the main content area, there's a link "Altri contenuti di Notizie".

Google News

Cerca argomenti, località e fonti

Notizie principali

Per te

Stai seguendo

Ricerche salvate

COVID-19

Italia

Dal mondo

Notizie locali

Affari

Scienza e tecnologia

Intrattenimento

Notizie

Altri contenuti di Notizie

Notizie sul COVID-19. visualizza le ultime notizie relative al coronavirus (COVID-19)

Von der Leyen: "Tutti in Ue devono avere salari minimi. Con Conte vertice sulla sanità in Italia"

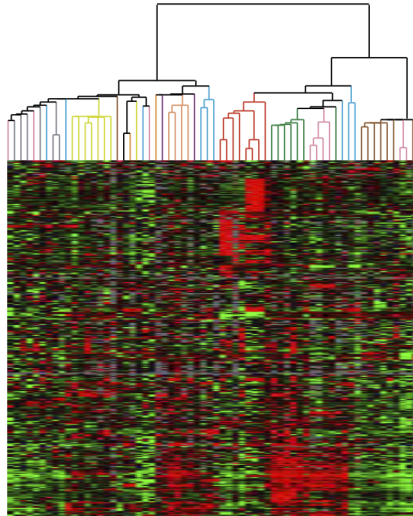
Il Fatto Quotidiano · 1 ora fa

- Ue, Von der Leyen traccia le linee per la ripartenza europea: priorità a sanità, clima e digitale
la Repubblica · 37 minuti fa
- Il discorso di Ursula von der Leyen: «Il 37% del Recovery Fund per il clima»
Corriere della Sera · 1 ora fa
- Von der Leyen: "E' il momento per l'Europa per allontanarsi da questa fragilità"
La Stampa · 4 ore fa
- Von der Leyen: "Organizzeremo con Conte un vertice in Italia sulla sanità" | "Tutti devono avere un salario minimo"
TGCOM · 2 ore fa

Raggruppamento di tipologie di cancro

Input: misure di espressione genica di colture cellulari

Output: raggruppamento delle colture per tipologia di cancro



Tipologie di problemi di apprendimento

- Apprendimento supervisionato
(problemi di predizione)
 - Classificazione
 - Regressione
 - Stima di probabilità
- Apprendimento non supervisionato
(apprendimento della rappresentazione)
 - Clustering
 - Riduzione della dimensionalità
- Altri tipi di apprendimento
(es. apprendimento per rinforzo)

Problemi di predizione: input e output

- Spazio degli input \mathcal{X}
Es.: immagini RGB 64×64 rappresentanti animali
- Spazio degli output \mathcal{Y}
Es.: i nomi di 100 specie animali

Osservati un certo numero di esempi $(x, y) \in \mathcal{X} \times \mathcal{Y}$, cerchiamo una *regola di predizione* (o *ipotesi*, o *modello*)

$$h : \mathcal{X} \rightarrow \mathcal{Y}$$

Diverse categorie di problemi a seconda del tipo di valori di output: qualitativi (*categorici*), quantitativi (*numerici*), probabilità, ecc.

Classificazione binaria

Es.: Identificazione dello spam

$\mathcal{X} = \{ \text{messaggi email} \}$

$\mathcal{Y} = \{ \text{spam, email} \}$

Classificazione multiclasse

Es.: Riconoscimento di cifre scritte a mano

$\mathcal{X} = \{ \text{immagini } 28 \times 28 \text{ a scala di grigi} \}$

$\mathcal{Y} = \{ 0, 1, 2, \dots, 9 \}$

Regressione

Es.: Predizione del salario in base ad età, anno di interesse, titolo di studio

$$\mathcal{X} = [0, 120] \times [1900, 2100] \times \{elementari, medie, diploma, laurea, dottorato\}$$
$$\mathcal{Y} = [0, \infty)$$

Es.: Stime di una compagnia assicurativa

Qual è l'aspettativa di vita di questa persona?

$$\mathcal{Y} = [0, 120]$$

Quali variabili predittrici (spazio \mathcal{X}) potremmo usare nel secondo caso?

Regressione

Es.: Predizione del salario in base ad età, anno di interesse, titolo di studio

$$\mathcal{X} = [0, 120] \times [1900, 2100] \times \{elementari, medie, diploma, laurea, dottorato\}$$
$$\mathcal{Y} = [0, \infty)$$

Es.: Stime di una compagnia assicurativa

Qual è l'aspettativa di vita di questa persona?

$$\mathcal{Y} = [0, 120]$$

Quali variabili predittive (spazio \mathcal{X}) potremmo usare nel secondo caso?

età, sesso, fumatore/non fumatore, pressione sanguigna, livello di colesterolo...

Stima di probabilità

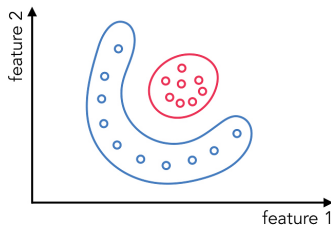
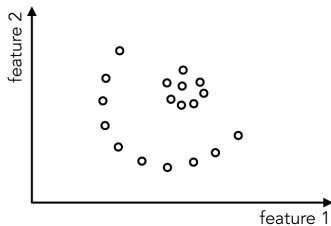
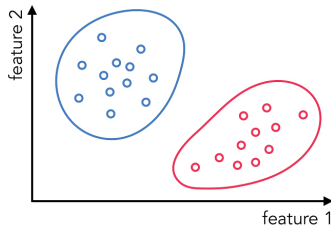
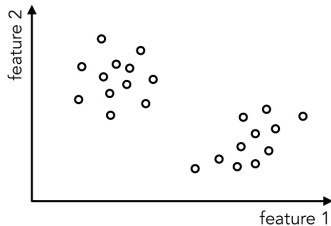
$\mathcal{Y} = [0, 1]$ rappresenta possibili valori di **probabilità**

Es.: Transazioni via carta di credito

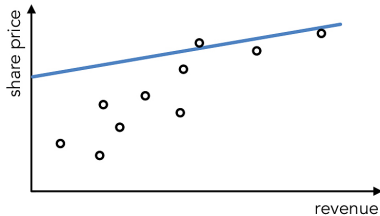
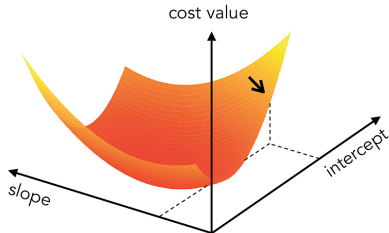
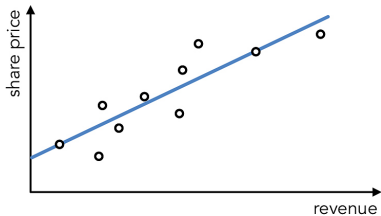
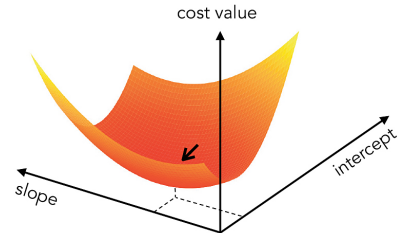
- x = dettagli della transazione
- y = probabilità che la transazione sia fraudolenta

Apprendimento non supervisionato: Clustering

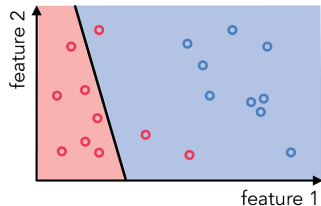
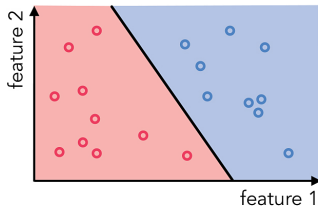
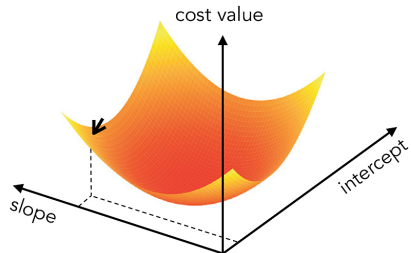
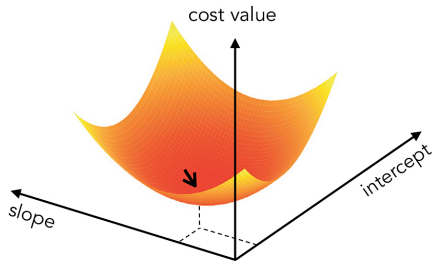
Clustering (raggruppamento): nessuna variabile di uscita!



Il ruolo dell'ottimizzazione matematica



Il ruolo dell'ottimizzazione matematica



Terminologia e notazione

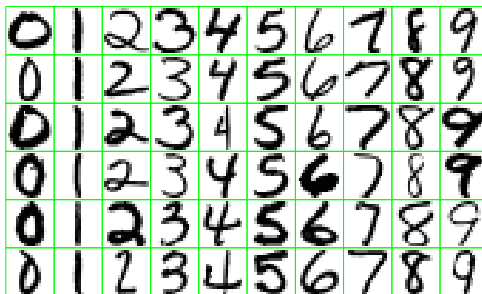
Termini	Sinonimi	Notazione
esempio	osservazione, punto dati	$(x, y), (x^{(i)}, y^{(i)})$
variabile di input	ingresso, predittore, feature, variabile indipendente	x_k
variabile di output	uscita, responso, etichetta, variabile dipendente	y
dati di apprendimento	campione statistico	S
modello	ipotesi, regola di predizione	h

$$\text{matrice dei dati: } \left[\begin{array}{ccc|c} x_1^{(1)} & \dots & x_d^{(1)} & y^{(1)} \\ x_1^{(2)} & \dots & x_d^{(2)} & y^{(2)} \\ & \dots & & \\ x_1^{(m)} & \dots & x_d^{(m)} & y^{(m)} \end{array} \right] = [X \ y]$$

Esempio: il dataset MNIST

60,000 immagini di esempio ($m = 60000$)

Ogni immagine è un vettore di 784 interi ($d = 784$)



Variabili di input: $28 \times 28 = 784$ interi tra 0 (nero) e 255 (bianco):
i primi 28 interi descrivono la luminosità dei pixel della prima riga,
i secondi 28 quella dei pixel della seconda riga, ecc.

Esempio: il dataset MNIST

$$x^{(1)} = [0, 0, 34, 31, 69, \dots, 0, 0]$$

$$y^{(1)} = 9$$

...

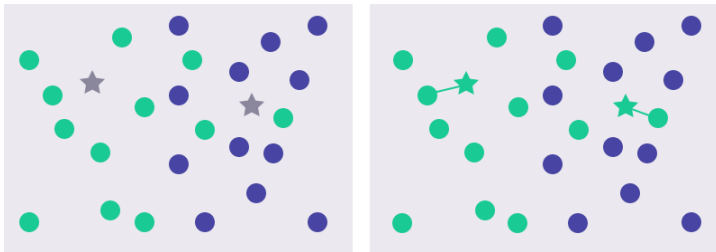
$$x^{(60000)} = [97, 25, 120, 101, 97, \dots, 255, 200]$$

$$y^{(60000)} = 5$$

La *distanza euclidea* tra il vettore x e il vettore x' è

$$\|x - x'\| = \sqrt{\sum_{j=1}^d (x_j - x'_j)^2}.$$

Un semplice algoritmo di predizione: Nearest Neighbor



Per classificare un nuovo vettore x , cerca il vettore nel dataset più vicino ad x e restituiscine l'etichetta:

- Trova l'indice $i \in \{1, 2, \dots, m\}$ che minimizza $\|x - x^{(i)}\|$
- Restituisci $y^{(i)}$