

IN550 Machine Learning

Tipologie di apprendimento

Vincenzo Bonifaci

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 .

Metodi deduttivi vs. metodi induttivi

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 apprendimento

Input: foto di un animale

Output: nome dell'animale

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

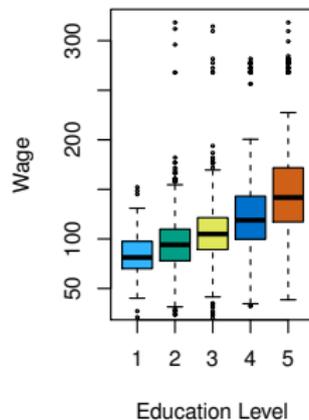
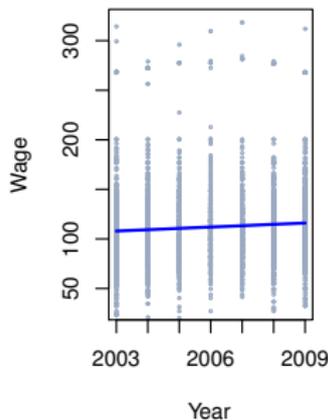
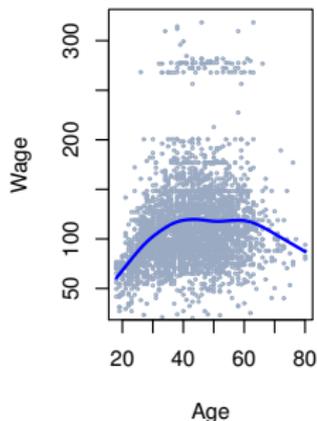
Problemi di apprendimento automatico: esempi

- 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



Dati da un 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

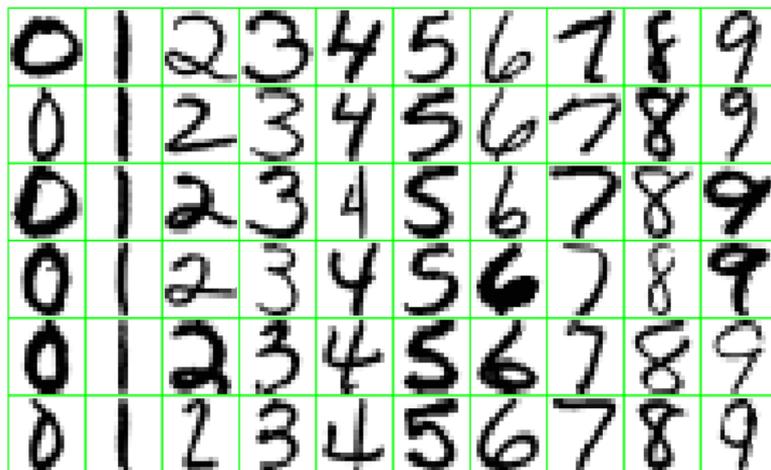
	<code>george</code>	<code>you</code>	<code>hp</code>	<code>free</code>	<code>!</code>	<code>edu</code>	<code>remove</code>
<code>spam</code>	0.00	2.26	0.02	0.52	0.51	0.01	0.28
<code>email</code>	1.27	1.27	0.90	0.07	0.11	0.29	0.01

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

Riconoscimento di cifre scritte a mano

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 e 255:

i primi 28 interi descrivono la luminosità dei pixel della prima riga,

i secondi 28 descrivono la luminosità 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 left is the Google News logo. To its right is a search bar with the placeholder text "Cerca argomenti, località e fonti". Below the search bar is a horizontal menu with several categories: "Notizie principali" (highlighted in blue), "Per te", "Stai seguendo", "Ricerche salvate", "COVID-19", "Italia", "Dal mondo", "Notizie locali", "Affari", "Scienza e tecnologia", and "Intrattenimento". To the right of the search bar, the word "Notizie" is displayed, followed by a link "Altri contenuti di Notizie". Below this is a blue button that says "Notizie sul COVID-19. visualizza le ultime notizie relative al coronavirus (COVID-19)" with a right-pointing arrow. The main content area features a news article with the headline "Von der Leyen: 'Tutti in Ue devono avere salari minimi. Con Conte vertice sulla sanità in Italia'". The article is attributed to "Il Fatto Quotidiano" and is dated "1 ora fa". To the right of the headline is a small image of Ursula von der Leyen. Below the headline is a list of three bullet points, each with a sub-headline and a source/age:

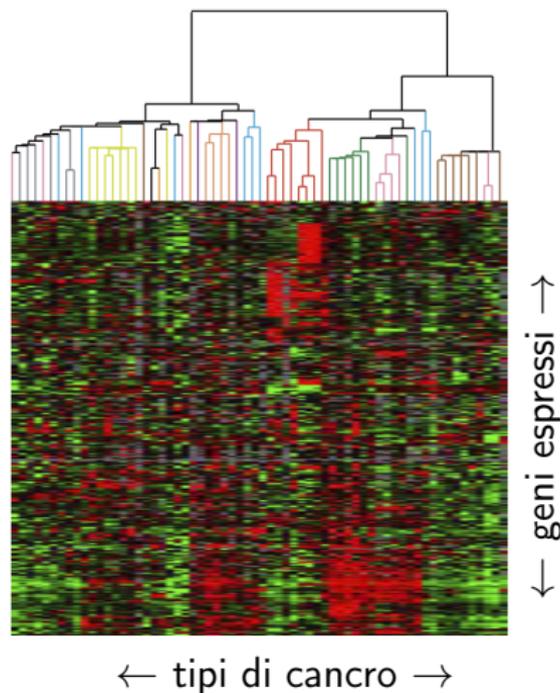
- **Ue, Von der Leyen traccia le linee per la ripartenza europea: priorità a sanità, clima e digitale**
la Repubblica - 37 minuti fa
- **Il discorso di Ursula von der Leyen: «Il 37% del Recovery Fund per il clima»**
Corriere della Sera - 1 ora fa
- **Von der Leyen: "E' il momento per l'Europa per allontanarsi 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"**

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
- Apprendimento non supervisionato
(apprendimento della rappresentazione)
 - Clustering
 - Riduzione della dimensionalità
- Apprendimento per rinforzo
(basato su azioni e ricompense)

Problemi di predizione: input e output

- Spazio degli ingressi \mathcal{X}
Per es., immagini RGB 32×32 rappresentanti animali
- Spazio delle etichette \mathcal{Y}_0
Per es., i nomi di 100 animali
- Spazio delle uscite \mathcal{Y} (in genere $\mathcal{Y} \supseteq \mathcal{Y}_0$, ma non sempre)

Osservati un certo numero di esempi $(x, y) \in \mathcal{X} \times \mathcal{Y}_0$, 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:
(1) qualitativi (*categorici*), (2) quantitativi (*numerici*), (3) probabilità...

Classificazione binaria

Es.: Identificazione dello spam

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

$$\mathcal{Y}_0 = \{ \text{spam, email} \}$$

Classificazione multiclasse

Es.: Riconoscimento di cifre scritte a mano

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

$$\mathcal{Y}_0 = \{ 0, 1, 2, \dots, 9 \}$$

Regressione

Es.: Predizione del salario in base ad età e titolo di studio

$$\mathcal{X} = [0, 120] \times \{elementari, medie, diploma, laurea, dottorato\}$$

$$\mathcal{Y}_0 = [0, \infty)$$

Es.: Stime di una compagnia assicurativa

Qual è l'aspettativa di vita di questa persona?

$$\mathcal{Y}_0 = [0, 120]$$

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

Regressione

Es.: Predizione del salario in base ad età e titolo di studio

$$\mathcal{X} = [0, 120] \times \{elementari, medie, diploma, laurea, dottorato\}$$

$$\mathcal{Y}_0 = [0, \infty)$$

Es.: Stime di una compagnia assicurativa

Qual è l'aspettativa di vita di questa persona?

$$\mathcal{Y}_0 = [0, 120]$$

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

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

Classificazione probabilistica

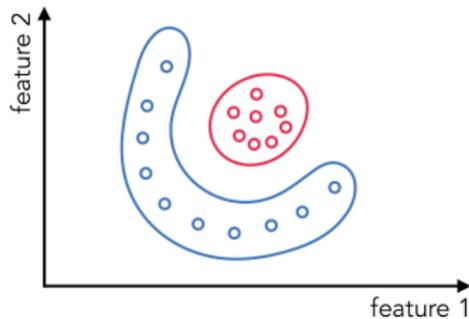
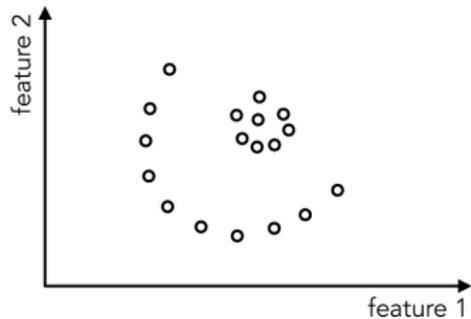
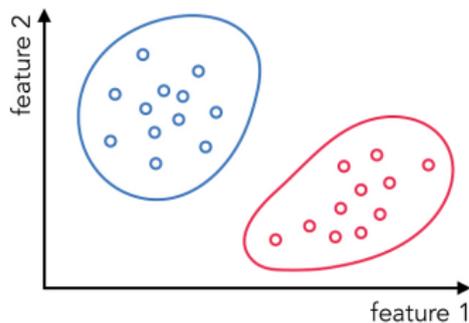
$$\mathcal{Y}_0 = \{0, 1\}$$

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

Es.: Transazioni via carta di credito

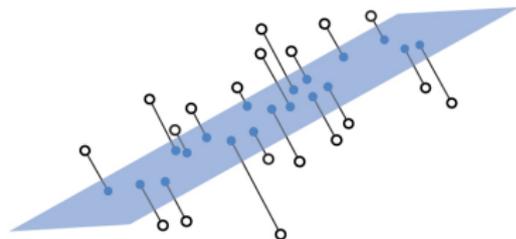
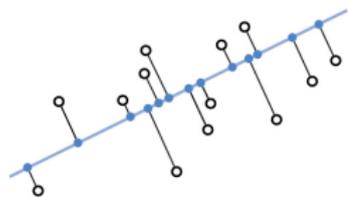
- x = dettagli della transazione
- y = probabilità che la transazione sia fraudolenta

Apprendimento non supervisionato: Clustering



Apprendimento non supervisionato:

Riduzione della dimensionalità



Apprendimento per rinforzo

Non basato su **esempi** ma solo su *azioni* e *ricompense*

Modella l'interazione tra un *agente* e un *ambiente*:

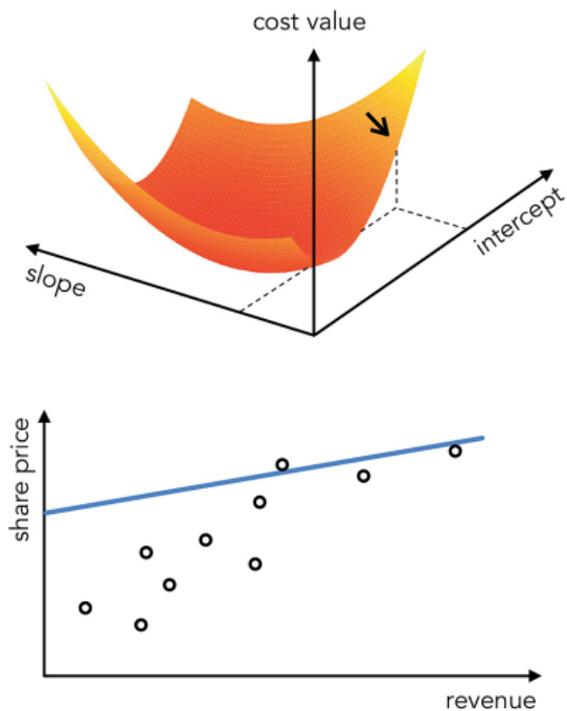
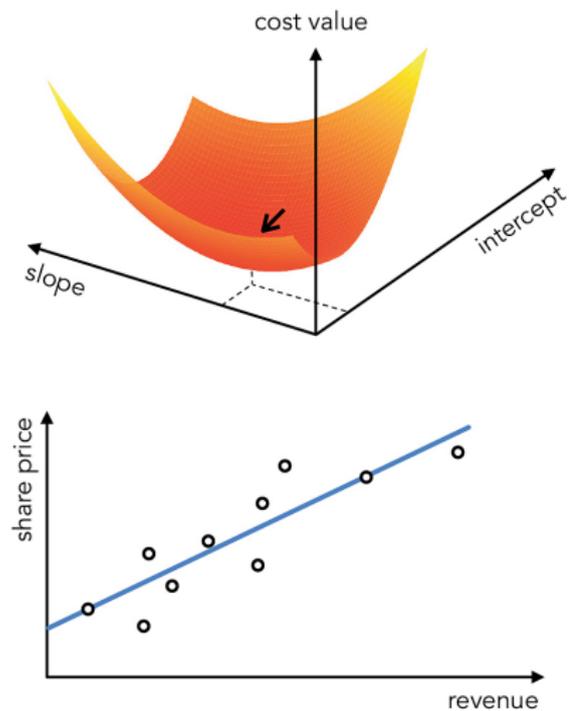
- 1 L'agente sceglie un' *azione* in base allo *stato* dell'ambiente
- 2 L'ambiente risponde con una *ricompensa* e altera il proprio stato
- 3 L'agente sceglie una nuova azione, e così via

Esempi: compravendite finanziarie, robotica, giochi. . .

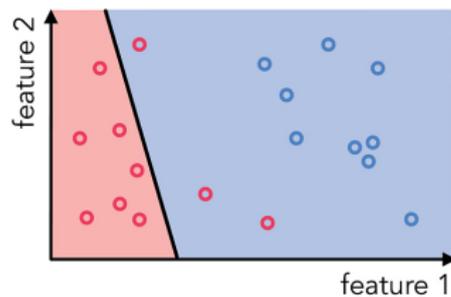
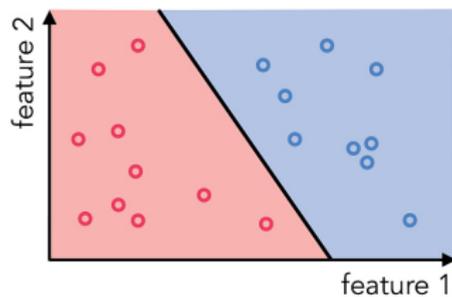
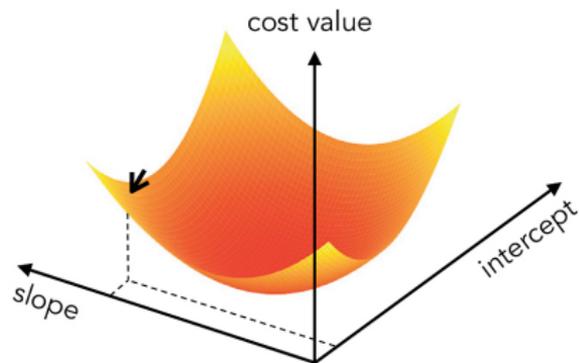
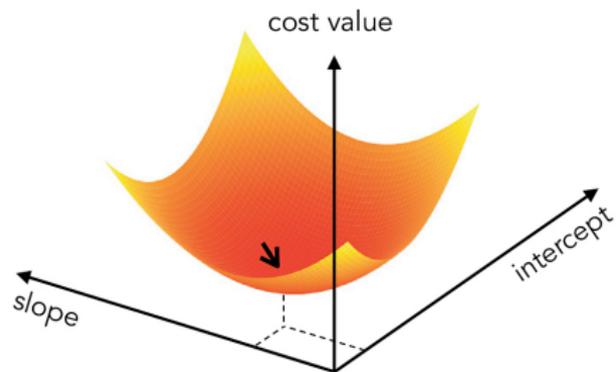
Tipicamente più difficile dell'apprendimento supervisionato:

- Assenza di esempi espliciti
È un orso? No. anziché È un orso? No, è un cane.
- I dati sono intrinsecamente dinamici (dipendono dalle azioni)

Il ruolo dell'ottimizzazione



Il ruolo dell'ottimizzazione



Terminologia e notazione

Termine	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
ipotesi	modello, regola di predizione	h

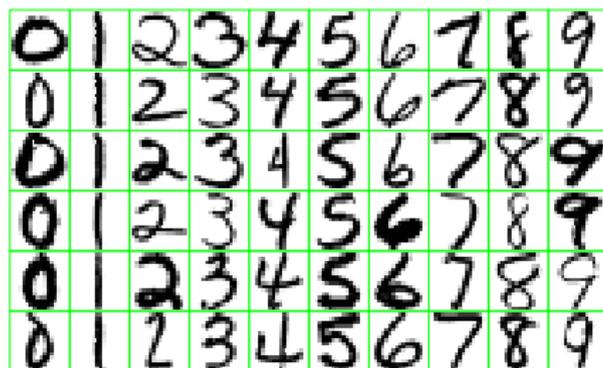
matrice dei dati:

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

Esempio: il dataset MNIST

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

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

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

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

...

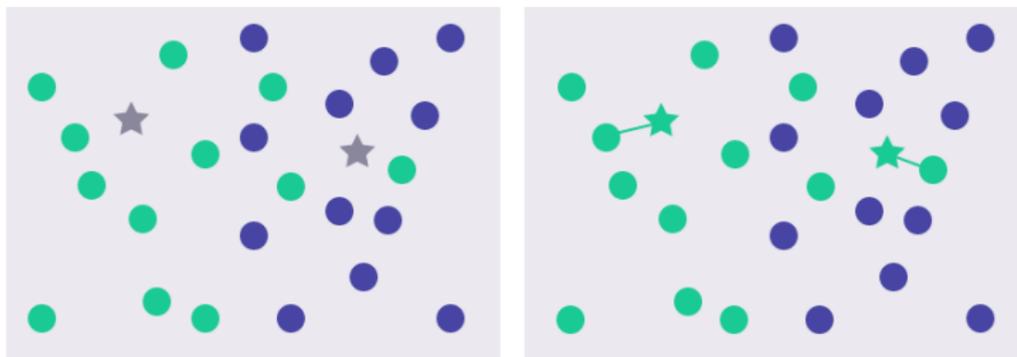
$$x^{(60000)} = [97, 25, 120, 101, 97, \dots, 255, 200]$$

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

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 più vicino ad x nel dataset, e restituiscine l'etichetta:

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