La Rilevanza nell'Apprendimento Statistico

Vivere è imparare

Konrad Lorenz

Matteo Marsili *The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy*

Nonostante lo straordinario sviluppo dell'intelligenza artificiale, la nostra comprensione teorica dell'apprendimento statistico, che ne è alla base, è ancora insufficiente. Offrirò qui una prospettiva basata sul concetto di rilevanza. La rilevanza misura la quantità di informazione sul processo generativo contenuto in un campione di dati o nella rappresentazione interna di un sistema che impara. Spiegherò perché le proprietà di sistemi che imparano differiscono marcatamente da quelle tipiche di sistemi fisici. In particolare, la criticalità emerge come una caratteristica generica di dati massimamente informativi o di sistemi efficienti per l'apprendimento. L'esposizione che segue è discorsiva. Si rimanda agli articoli originali per una trattazione più estesa.

La rivoluzione digitale e l'avvento dell'era dei *Big Data*, hanno portato l'interesse della comunità scientifica sul problema dell'apprendimento statistico automatico, o *machine learning*. Quest'ultimo è alla base degli sviluppi più sorprendenti in Intelligenza Artificiale, poiché la capacità di apprendere è una componente fondamentale di ogni comportamento intelligente. A dispetto delle sue spettacolari applicazioni, dalle traduzioni automatiche alla sconfitta del campione di Go da parte di *AlphaGo*, il nostro grado di comprensione di questi algoritmi è probabilmente paragonabile a quanto si sapeva del motore a vapore nei primi decenni della rivoluzione industriale. Gli ingredienti necessari per realizzare motori potenti (pistoni, pulegge, vapore, ecc.) erano noti ben prima che fosse raggiunta una comprensione soddisfacente del fenomeno. Questa comprensione giunse grazie al lavoro di Carnot, su macchine ideali, che rivelò l'esistenza di una funzione di stato — l'entropia — che gioca un ruolo fondamentale nei processi termodinamici. Allo stesso modo, gli ingredienti per l'apprendimento statistico (ad esempio *algorithms, architectures, and structured data* secondo [1]) sono noti, ma il nostro livello di comprensione è tuttora insoddisfacente.

Infatti, anche se sistemi esperti come quelli di traduzione automatica hanno prestazioni compa-

rabili a quelle dell'intelligenza umana, i dati su cui sono allenati sono generati dall'Uomo. Inoltre, queste prestazioni sono limitate a compiti ben specifici e richiedono costi energetici ordini di grandezza maggiori di quelli impiegati dal cervello umano. Infine, l'inferenza statistica negli organismi viventi deve essere efficiente in regimi in cui i dati sono estremamente scarsi. Le falene che riescono a inferire la posizione della femmina più efficientemente, dalle deboli tracce di feromoni che questa rilascia nell'aria, trasmetteranno più probabilmente i propri geni. Dunque, comprendere l'apprendimento statistico dal punto di vista teorico non è solo importante per le applicazioni di intelligenza artificiale, ma anche per comprendere i sistemi viventi.

Gli strumenti matematici a nostra disposizione per affrontare questi problemi sono gli stessi su cui si basa la statistica classica. Le ultime decadi sono state testimoni di una notevole convergenza tra discipline diverse, dalla teoria dell'informazione, dei codici e della complessità, alla probabilità, la statistica e la meccanica statistica. Per affrontare il problema dell'apprendimento in regimi rilevanti per l'intelligenza artificiale, i concetti della statistica classica devono essere declinati in nuovi ambiti, in cui i dati e i modelli hanno una dimensionalità comparabile se non maggiore al numero di campioni disponibili, ed in cui non abbiamo idea di quale sia il vero modello generativo.

Campioni massimamente informativi: Rilevanza e criticalità

Quanta informazione contiene un campione statistico sul processo generativo sottostante? Nella statistica classica, quando il modello generativo appartiene ad una famiglia parametrica di distribuzioni, questa domanda trova una risposta nel concetto di statistiche sufficienti. Queste sono le osservabili che possiamo calcolare dai dati, e che permettono di stimare i parametri del modello. Tutto il resto dell'informazione contenuta nel campione è rumore non informativo. Possiamo quantificare quanto impariamo, calcolando l'informazione mutua tra il campione ed i parametri del modello. E troviamo che *i)* il numero di bit che impariamo cresce solo con il logaritmo del numero N di dati[1], una frazione risibile dell'informazione contenuta nei dati, che corrisponde allo spazio necessario a immagazzinare i dati sul nostro computer (che cresce con N). Inoltre, *ii)* impariamo quel numero di bit sia che il modello stimato sia quello giusto o meno, o che i dati siano informativi o totalmente aleatori. Per avere garanzie che quei bit che acquisiamo svelino davvero informazioni sul processo generativo, dobbiamo comparare modelli diversi in modo Bayesiano. Tuttavia, il numero dei modelli che dovremmo considerare cresce più che esponenzialmente con la dimensionalità dei dati. Ciò rende questo approccio impraticabile (ma si veda [2]).

In assenza di informazioni *a priori* su quale sia il modello generativo, la massima di Socrate, secondo cui l'unica cosa che possiamo sapere è "saper di non sapere", ci viene in soccorso. Infatti, mentre è difficile identificare una struttura precisa nei dati, la mancanza assoluta di informazione ha una sua matematizzazione precisa nel principio di massima entropia. Questo permette di quantificare esattamente la parte non informativa di un campione in quei sotto-campioni di osservazioni che si presentano lo stesso numero di volte. Il resto dell'informazione, che corrisponde all'entropia di Shannon della distribuzione delle frequenze, è un limite superiore della quantità di informazione che il campione contiene sul modello generativo. Questa entropia, che chiameremo **rilevanza** in quel che segue, è una grandezza fondamentale che ci permette di stimare quanto il nostro campione possa essere informativo. Nel linguaggio comune, il termine rilevanza è spesso inteso in senso relativo. Qui viene inteso in senso assoluto, o meglio relativo al processo generativo dei dati, che ci è ignoto. Per spiegare in modo colorito, ciò che è rilevante per dei neuroni di un'area del cervello o per proteine coinvolte nella stessa funzione biologica, è ciò di cui i neuroni o le proteine parlerebbero al bar, davanti ad un caffè.

Questa caratterizzazione matematica ci permette di identificare campioni di dati massimamente informativi in quelli che hanno massima

[1] Il motivo è che l'inferenza ci permette di ridurre l'incertezza su ogni parametro del modello di un fattore $1/\sqrt{N}$, ed il numero di bit necessari per rappresentare un numero a quella precisione cresce come $\frac{1}{2}\log N$.

rilevanza. Questo problema è ben posto, una volta che consideriamo campioni con lo stesso livello di **risoluzione**. La risoluzione è la quantità di bit necessaria ad immagazzinare un dato del campione sul nostro computer, ed è misurata dall'entropia di Shannon dei dati. La risoluzione può essere scelta dall'osservatore[2]. Ma una volta determinato il livello di dettaglio, la rilevanza è una proprietà intrinseca dei dati, che ci svela informazioni sulla natura del processo generativo.

La figura 1 riporta un esempio della dipendenza della rilevanza in funzione della risoluzione, per rappresentazioni del comportamento temporale di $N = 4000$ azioni del New York Stock Exchange (NYSE) estratte con tre diversi algoritmi di *data clustering*. Si rimanda a [3] per ulteriori dettagli. In breve, gli algoritmi classificano le azioni in *clusters* s, la cui grandezza k_s corrisponde al numero di azioni appartenenti al cluster s. La risoluzione e la rilevanza sono date da

$$\hat{H}[s] = -\sum_s \hat{p}_s \log \hat{p}_s, \quad (1)$$

$$\hat{H}[k] = -\sum_k \hat{p}_k \log \hat{p}_k \quad (2)$$

rispettivamente, dove $\hat{p}_s = k_s/N$ e $\hat{p}_k$ sono le probabilità empiriche che un'azione presa a caso appartenga al *cluster* s, e che tale cluster contenga $k_s = k$ azioni, rispettivamente. Al variare del numero di *clusters* considerati, $\hat{H}[s]$ e $\hat{H}[k]$ tracciano una curva sul piano. Tale curva fornisce una misura di quanto la rappresentazione che diversi algoritmi di *data clustering* estraggono dai dati sono informative sul processo generativo. Se esaminiamo i tre diversi algoritmi riportati nella figura (si veda [3]), scopriamo che l'algoritmo che realizza valori più elevati di $H[k]$ è basato su un modello più aderente alla dinamica delle azioni finanziarie rispetto agli altri, e si avvicina più degli altri alla "vera" classificazione, che è quella data dalla *Security and Exchange Commission* del *New York Stock Exchange*.

In Fig. 1 distinguiamo due regimi. Uno in cui $\hat{H}[k]$ cresce con la risoluzione $\hat{H}[s]$. Questo è il regime in cui ogni stato s è campionato un numero sufficiente di volte e possiamo aspettarci che la distribuzione empirica fornisca una approssimazione ragionevole della vera distribuzione $p(s)$. Ciò non vale per la parte destra, in cui $\hat{H}[k]$ decresce con $\hat{H}[s]$. Questo è il regime di cui parlavamo sopra, in cui la statistica classica non è applicabile e in cui operano l'intelligenza artificiale e quella naturale.

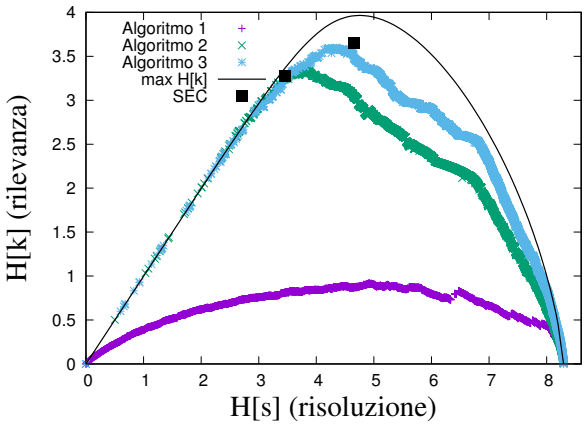

Figura 1: *Andamento di $H[k]$ in funzione di $H[s]$ per tre diversi algoritmi di data clustering, applicati a dati finanziari (si veda [3] per dettagli). Il massimo teorico è riportato dalla curva continua. I due punti ■ corrispondono alla classificazione delle azioni della Security and Exchange Commission (SEC).*

La stessa figura riporta anche la curva corrispondente al valore massimo di $\hat{H}[k]$ per un dato valore di $\hat{H}[s]$. La caratteristica dei campioni che realizzano tale massimo è quella di esibire criticalità statistica, cioè

$$\hat{p}_k \sim k^{-2-\mu}. \quad (3)$$

L'esponente $-\mu$ corrisponde, approssimativamente alla pendenza della curva $\hat{H}[k]$ in funzione di $\hat{H}[s]$. Ciò significa che quando comprimiamo i dati[3], riducendo $\hat{H}[s]$ di un bit, la rilevanza $\hat{H}[k]$ aumenta di μ bit. Possiamo quindi distinguere ulteriormente due sotto-regimi. Quando $\mu > 1$ l'ulteriore compressione rivela una quantità di informazione $\Delta\hat{H}[k]$ maggiore di quella $(-\Delta H[s])$ eliminata. Quando invece $\mu < 1$ la compressione avviene a spese di informazione potenzialmente rilevante. Il punto $\mu = 1$ caratterizza dunque quei campioni massimamente informativi con un livello ottimale di compres-

[2] Ad esempio, possiamo descrivere lo stato di una molecola a diversi livelli di dettaglio, dalla particella puntuale alla descrizione meccanico quantistica in termini di numeri quantici.

[3] Che corrisponde a ridurre il numero dei clusters, nell'esempio di Fig.1.

sione. In questo punto, la statistica di frequenze Eq. (3) corrisponde alla celebre legge di Zipf [4]. Alla luce di quanto detto, l'osservazione della legge di Zipf nella distribuzione delle parole nel linguaggio, o nella frequenza di antigeni nel sistema immunitario – per citare solo due esempi dell'ubiquità di tale legge – può essere letta come conseguenza che sia il linguaggio che il sistema immunitario sono rappresentazioni efficienti per la comunicazione o la difesa dai patogeni, rispettivamente. Più in generale, la prospettiva qui delineata fornisce una chiara risposta alla domanda sul perché la biologia sembri operare in condizioni di criticalità [5]: i sistemi biologici non sono ottimizzati ad un punto critico. Operano su rappresentazioni massimamente informative.

Questa caratterizzazione delle rappresentazioni massimamente informative apre la strada a metodi di inferenza da dati alto-dimensionali *model free*: una direzione di ricerca molto promettente. Ad esempio, [6] mostra come il concetto di rilevanza sia in grado di identificare quei neuroni altamente informativi per la navigazione, unicamente sulla base della loro attività neurale.

Se la criticalià è l'impronta di campioni massimamente informativi, la dovremmo ritrovare nella teoria dei codici, che si occupa appunto di comprimere dati generati da una sorgente in modo efficiente. Questo è stato infatti verificato per codici che realizzano una lunghezza di descrizione minima (*minimum description length*) in [7]. Per questi codici è possibile dimostrare esplicitamente che si tratta di sistemi critici, posizionati esattamente al punto critico di una transizione di fase [7].

La struttura delle macchine ideali per l'apprendimento

Mentre la criticalità è un eccezione per i sistemi fisici, che si verifica solo per valori particolari dei parametri esterni (e.g. temperatura e pressione), la criticalità statistica è la norma per sistemi evoluti al fine di estrarre rappresentazioni efficienti da dati complessi. Per capirlo, dobbiamo aprire la scatola di questi sistemi, e guardare alla distribuzione dei livelli energetici ed alle Hamiltoniane che li caratterizzano. Ci rendiamo allora conto che il problema dell'apprendimento è di natura duale rispetto a quello della meccanica statistica. In entrambi i casi siamo interessati alla statistica degli stati s interni di un sistema in contatto con un sistema $\vec{x}$ esterno. Nella meccanica statistica, s è lo stato di un sottosistema in contatto termico con l'ambiente $\vec{x}$. L'Hamiltoniana $\mathcal{H}$ è data, ed il principio di massima entropia indica nella distribuzione di Gibbs-Boltzmann

$$\arg \max_{p(s):\langle \mathcal{H} \rangle = E} H[s] = \frac{1}{Z} e^{-\mathcal{H}(s)/T}, \qquad (4)$$

il risultato, dove $H[s] = -\sum_s p(s) \log p(s)$ è l'entropia di Shannon. Eq. (4) rende esplicito che l'unica informazione che il sistema fisico ricorda del suo ambiente $\vec{x}$ è l'energia media $\langle \mathcal{H} \rangle$, o equivalentemente la temperatura T, che deve coincidere con quella dell'ambiente[4]. Come conseguenza, la distribuzione dell'energia $\mathcal{H}$ del sistema è concentrata in un piccolo intervallo dell'ordine di $1/\sqrt{n}$ attorno al valor medio $\langle \mathcal{H} \rangle = E$, dove $n \gg 1$ è il numero di gradi di libertà[5]. Questo risultato prende il nome di **Proprietà Asintotica dell'Equipartizione** (PAE) [9] in probabilità. Questa afferma che quasi sicuramente, tutti gli stati osservati hanno la stessa probabilità $p(s) \sim e^{-nH[s]}$ e che, come conseguenza, il numero di tali stati tipici è $\sim e^{nH[s]}$.

In un sistema ottimizzato per apprendere e classificare gli stati $\vec{x}$ dell'ambiente, l'Hamiltoniana assume il naturale significato di costo informativo $\mathcal{H}(s) = -\log p(s)$. Per riuscire a ricordare il massimo numero di dettagli del suo ambiente $\vec{x}$, che in questo caso sono i dati che il sistema vuole imparare, un tale sistema deve lottare contro la PAE, modellando l'Hamiltoniana $\mathcal{H}(s)$ in modo tale che dati $\vec{x}$ e $\vec{x}'$ con una diversa struttura, corrispondano a stati interni s e s' con energie significativamente diverse $\mathcal{H}(s) \neq \mathcal{H}(s')$, come dimostrato in [10]. Ne consegue che i sistemi che estraggono rappresentazioni efficienti dai dati debbano avere una distribuzione di livelli d'energia più larga possibile. Usando ancora l'entropia

[4] Si osservi che $\langle \mathcal{H} \rangle$ è la statistica sufficiente della distribuzione Eq. (4). L'unica variabile necessaria a determinare il parametro coniugato T.

[5] Tale intervallo si allarga solo quando il sistema è al punto critico di una transizione di fase, cioè quando la distinzione tra due diverse fasi termodinamiche scompare. Ad esempio, come mostrato in [8], per il modello di Ising in campo medio, le fluttuazioni dell'energia sono dell'ordine di $n^{-3/4}$ alla temperatura critica.

come misura quantitativa di informazione, ciò porta al **principio di massima rilevanza**

$$\mathcal{H}^*(s) = \arg \max_{\mathcal{H}(s):\ \langle \mathcal{H} \rangle = H[s]} H[E], \qquad (5)$$

dove

$$H[E] = -\sum_E p(E) \log p(E), \qquad (6)$$

e $p(E)$ è la probabilità che uno stato tipico della macchina abbia energia $\mathcal{H}(s) = E$. La rilevanza $H[E]$ misura la quantità di informazione che la rappresentazione s mantiene sul processo generativo dei dati $\vec{x}$, alla risoluzione $H[s]$. A sua volta, $H[s]$ è il costo informativo medio della rappresentazione. Le soluzioni del problema (5) sono caratterizzate da una densità di stati esponenziale $W(E) = \sum_s \delta\left(\mathcal{H}(s) - E\right) \simeq W_0 e^{\mu E}$ o, equivalentemente, da distribuzioni esponenziali dell'energia

$$p(E) = W(E) e^{-E} \simeq e^{(\mu - 1)E}. \qquad (7)$$

Questo è l'analogo della criticalità statistica che caratterizza campioni massimamente informativi[6] (Eq. 3). Il parametro μ gioca ancora lo stesso ruolo. La rilevanza cresce con la risoluzione $H[s]$ quando $\mu \leq 1$. Per $\mu = 1$ la rilevanza $H[E]$ raggiunge il suo massimo, corrispondente ad una distribuzione $p(E)$ approssimativamente uniforme (Eq. 7). Per valori maggiori di μ, $H[E]$ decresce con la risoluzione. Nei termini della differenza

$$H[s|E] = H[s] - H[E] = \langle \log W(E) \rangle, \qquad (8)$$

che fornisce una misura del rumore non-informativo della rappresentazione, è facile comprendere che l'interrelazione tra risoluzione e rilevanza è della stessa natura di quello discusso sopra per un campione: La diminuzione di un bit in $H[s]$ riduce $H[s|E]$ di $\mu - 1$ bit. Per $\mu > 1$ la compressione elimina solo dettagli non informativi della rappresentazione e rivela dettagli rilevanti, mentre per $\mu < 1$ una frazione $1 - \mu$ dell'informazione è persa.

Il confronto tra le equazioni (4) e (5) rende esplicita la relazione duale dei due problemi. Tale dualità si manifesta nel fatto che l'entropia $H[s|E] = \langle \log W(E) \rangle$ è una funzione concava dell'energia $H[s] = \langle E \rangle$ in meccanica statistica, mentre è convessa nell'apprendimento ottimale, cioè per le soluzioni dell'Eq. (5). Inoltre la criticalità si verifica solo in condizioni speciali in meccanica statistica, mentre essa è una caratteristica tipica di macchine allenate nell'apprendimento di strutture complesse di dati. è curioso osservare come "essere critici nell'apprendimento" non sia solo importante nel senso comune del termine, ma anche nell'accezione tecnica che il termine assume in meccanica statistica.

La rilevanza $H[E]$ fornisce anche un limite inferiore alla quantità di informazione che la rappresentazione interna di una macchina che impara estrae sulle caratteristiche nascoste (*hidden features*) dei dati [8]. Le *hidden features* appaiono come le statistiche sufficienti della rappresentazione interna che la macchina genera nell'apprendimento. Massimizzare $H[E]$ implica dunque estrarre il massimo di informazioni possibile sulle *hidden features*.

Il principio di massima rilevanza trova riscontro nel comportamento di macchine reali (*Restricted Boltzmann Machines, Deep Belief Networks, etc.*), come dimostrato in [8, 11]. Dove invece fallisce totalmente è nel caso di macchine di apprendimento Gaussiane, in cui la rilevanza resta costante nel processo di apprendimento, indipendentemente dalla struttura dei dati. Un fallimento illuminante, che rivela la differenza sostanziale tra apprendimento e stima di parametri [8]: non esistono *hidden features* in un modello Gaussiano, dal momento che le statistiche sufficienti sono ben visibili *a priori*. Tale modello impara solo medie e covarianze, ed è cieco rispetto ad ogni struttura dei dati che vada oltre ai primi due momenti.

Questo è un ulteriore indicazione della sostanziale differenza tra la statistica classica, che opera in regimi in cui la complessità dei modelli è limitata dai dati disponibili, e l'intelligenza artificiale dove l'apprendimento si avvale di macchine sovra-parametrizzate, che possono adattare la loro Hamiltoniana per modellare strutture complesse di dati.

[6] Infatti le frequenza k_s/N dello stato s in un campione approssima la probabiliità $e^{-\mathcal{H}(s)}$ di quello stato. Il cambio di variabili $\mathcal{H}(s) \simeq -\log(k_s/N)$ trasforma l'Eq. (7) nell'Eq. (3).

La meccanica statistica di macchine ideali di apprendimento

L'apprendimento genera una distribuzione $p(s, \vec{x})$ i cui massimi descrivono coppie $(s, \vec{x})$ di dati tipici $\vec{x}$ e rappresentazioni interne s corrispondenti. Le proprietà generiche del problema

$$\max_{s,\vec{x}} \log p(s, \vec{x}) = \max_{s,\vec{x}} \left[\mathcal{H}(s) + \mathcal{V}(\vec{x}|s)\right], \qquad (9)$$

possono essere studiate grazie alla teoria delle statistiche degli estremi, a seconda della distribuzione dei livelli di energia del sistema $\mathcal{H}(s)$ e dell'Hamiltoniana di interazione con l'ambiente $\mathcal{V}(\vec{x}|s)$. Tale studio è stato realizzato in Ref. [12], assumendo per entrambi una distribuzione esponenziale *stretched* di parametro γ,

$$P\{\mathcal{H}(s) \geq E\} = e^{-(E/\Delta)^\gamma}, \qquad (10)$$
$$P\{\mathcal{V}(\vec{x}|s) \geq V\} = e^{-V^\gamma}. \qquad (11)$$

Ciò permette di paragonare le proprietà di sistemi fisici, come il celebre *Random Energy Model* (REM) [13] ($\gamma = 2$) a quelle di macchine di apprendimento ideali ($\gamma = 1$). Il parametro Δ in Eq. (10) regola la scala di energia dei livelli della rappresentazione interna $\Delta \sim \mathcal{H}(s)$. Per ogni valore di γ, si osserva una transizione di fase, al variare di Δ, tra uno stato in cui la distribuzione $p(s)$ è concentrata su un numero finito di stati ($\Delta > \Delta_c$), ed uno stato "disordinato" in cui invece $p(s)$ si estende su un numero esponenziale di stati ($\Delta < \Delta_c$). Nell'apprendimento, la fase ordinata ($\Delta > \Delta_c$) corrisponde a rappresentazioni compresse dei dati.

Tale transizione è continua per $\gamma > 1$ (sistemi fisici) e discontinua per $\gamma < 1$. Inoltre, per $\gamma > 1$ la soglia critica Δ_c decresce all'aumentare della dimensionalità relativa del sistema s, rispetto all'ambiente $\vec{x}$, che per sistemi fisici corrisponde al bagno termico. Ciò rende possibile una descrizione statistica di un sistema, indipendentemente dalla conoscenza di tutte le variabili che possono influenzarlo. Ad esempio, se la struttura di una proteina dipendesse da ogni dettaglio delle interazioni con altre molecole presenti nella cellula, non sarebbe possibile inferirne la struttura dalla sequenza di amino-acidi. Tale predizione statistica risulta invece possibile [14]. Per $\gamma < 1$ invece, più fattori influenzano lo stato interno s del sistema, e meno tale stato è predicibile. Inoltre, non è possibile ricostruire l'Hamiltoniana di sistemi con $\gamma < 1$ attraverso un processo di misurazione sperimentale. Ciò significa che i sistemi con $\gamma < 1$ non sono "apprendibili". Ciò conferisce un ruolo speciale alle macchine ideali di apprendimento ($\gamma = 1$) come quelle che separano i sistemi che possiamo conoscere da quelli inaccessibili. Inoltre, solo per $\gamma = 1$ si verifica la situazione speciale in cui Δ_c è indipendente dalla dimensionalità di $\vec{x}$. Per una macchina di apprendimento, in cui $\vec{x}$ corrisponde ai dati ed s alla rappresentazione interna, questa proprietà è fondamentale. Ad esempio, desideriamo macchine che classifichino le immagini nello stesso modo, indipendentemente dalla risoluzione in pixels delle immagini stesse. Solo sistemi con una distribuzione esponenziale di stati ($\gamma = 1$) hanno questa proprietà.

Conclusioni

Le applicazioni di intelligenza artificiale pongono delle domande fondamentali sull'apprendimento e sulla natura delle macchine che estraggono rappresentazioni efficienti dai dati. La fisica statistica ha fornito importanti contributi in questa direzione. Queste pagine suggeriscono che, per sviluppare un quadro concettuale aderente alla realtà di sistemi efficienti di apprendimento, sia necessario comprendere le peculiari differenze tra questi sistemi e i sistemi fisici. Lo sviluppo di macchine di apprendimento più efficienti o di metodi per estrarre variabili rilevanti da dati alto-dimensionali, basate su questo quadro concettuale, sono promettenti direzioni di ricerca.

Più in generale, come osservato da Paul Davies [15], non abbiamo criteri quantitativi per distinguere materia inanimata da forme di vita. Ciò è cruciale, ad esempio, nell'astrobiologia per capire se la vita si è sviluppata anche altrove nell'Universo. Se l'apprendimento è una caratteristica distintiva della vita, il concetto di rilevanza può contribuire concretamente a colmare questa lacuna.

[1] L. Zdeborovà: *Understanding deep learning is also a job for physicists*, Nat. Phys. 16, (2020) 602.

[2] C. de Mulatier, P. P. Mazza, M. Marsili: *Statistical Inference of Minimally Complex Models*, arXiv:2008.00520 [cs.AI] (2020).

[3] M. Marsili, I. Mastromatteo, Y. Roudi: *On sampling and modeling complex systems* Journal of Statistical Mechanics: Theory and Experiment (2013) P09003

[4] G. K. Zipf: *Human Behavior and the Principle of Least Effort* Addison-Wesley, Cambridge, Mass. (1949).

[5] T. Mora, W. Bialek: *Are biological systems poised at criticality?*, Journal of Statistical Physics, 144 (2011) 268.

[6] R. J. Cubero, M. Marsili, Y. Roudi: *Multiscale relevance and informative encoding in neuronal spike trains* Journal of computational neuroscience, 48 (2020) 85.

[7] R. J. Cubero, M. Marsili, Y Roudi: *Minimum description length codes are critical* Entropy 20 (2018) 755

[8] O. Duranthon, M. Marsili, R. Xie: *Maximal Relevance and Optimal Learning Machines* arXiv:1909.12792 (2019)

[9] Thomas M. Cover, Joy A. Thomas: *Elements of information theory* John Wiley & Sons, New York (1999).

[10] R. J. Cubero, J. Jo, M. Marsili, Y Roudi, J. Song: *Statistical criticality arises in most informative representations* Journal of Statistical Mechanics: Theory and Experiment (2019) 063402.

[11] J. Song, M. Marsili, Jungyo Jo: *Resolution and relevance trade-offs in deep learning* Journal of Statistical Mechanics: Theory and Experiment (2018) 123406.

[12] M. Marsili *The peculiar statistical mechanics of optimal learning machines* Journal of Statistical Mechanics: Theory and Experiment (2019) 103401.

[13] B. Derrida, *Random-energy model: An exactly solvable model of disordered systems* Physical Review B 24 (1981) 2613.

[14] F. Morcos, et al. *Direct-coupling analysis of residue coevolution captures native contacts across many protein families* Proceedings of the National Academy of Sciences 108 (2011) E1293

[15] P. Davies, *Does new physics lurk inside living matter?*, Physics Today 73 (2020) 34.

Matteo Marsili: è *Senior Research Scientist* presso l'*Abdus Salam International Centre for Theoretical Physics*. Si occupa di applicazioni interdisciplinari della meccanica statistica.