Un’analisi preliminare della metrologia dell’intelligenza artificiale
Luca Mari, Francesco Bertolotti

Tutto_Misure, 2, 2024

[2.5.24]
In numeri precedenti di Tutto_Misure ci siamo già occupati in questa rubrica di intelligenza artificiale (“In dialogo con un agente artificiale a proposito di qualche argomento di metrologia”, https://issuu.com/tutto_misure/docs/tm.2-2023, e “Una breve introduzione ai sistemi di intelligenza artificiale nella prospettiva della metrologia”, https://issuu.com/tutto_misure/docs/tm.3-2023). Riprendiamo qui l’argomento, proseguendo nell’esplorazione di questioni all’intersezione tra metrologia e intelligenza artificiale (IA), e mantenendo a proposito di quest’ultima una prospettiva ampia, dunque in riferimento di prima tutto al caso generale dei sistemi di machine learning (ML), di cui i sistemi generativi (GenAI) e poi quelli conversazionali (chatbot) sono specializzazioni.
Di fronte a questioni complesse – e il nostro tema è certamente tale (ma nel seguito cercheremo di renderlo comprensibile anche a chi non è già esperto di ML) – è spesso appropriato operare analiticamente, identificando classi di sotto-problemi da trattare almeno in parte separatamente l’uno dall’altro. Per il nostro argomento, una distinzione basilare è tra
– *IA per la metrologia* (“misurare con la IA”)
e
– *metrologia della IA* (“misurare la IA”).
Proponiamo qui qualche considerazione preliminare, e tentativamente fondativa, sul secondo punto, a partire dalla constatazione che quanto più i sistemi di IA saranno diffusi tanto più sarà importante saper valutare in modo affidabile e socialmente condiviso la loro qualità.
In una sua recente risoluzione legislativa, il cosiddetto “AI Act”, lo stesso Parlamento europeo si è pronunciato esplicitamente al proposito: “*in cooperazione con i portatori di interessi e le organizzazioni pertinenti, quali le autorità di metrologia e di analisi comparativa, la Commissione dovrebbe incoraggiare, se del caso, lo sviluppo di parametri di riferimento e metodologie di misurazione per i sistemi di IA. A tal fine, la Commissione dovrebbe prendere atto dei partner internazionali che operano nel settore della metrologia, collaborando con essi, e dei pertinenti indicatori di misurazione relativi all’IA.*” (Regolamento sull’intelligenza artificiale - Risoluzione legislativa del Parlamento europeo del 13 marzo 2024, punto 74, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_IT.html).

Anche dato il contesto per cui stiamo scrivendo, è opportuno chiedersi, prima di tutto, se i sistemi di IA possano essere propriamente oggetto di misurazione, e dunque se l’obiettivo stesso di applicare gli insegnamenti della metrologia all’IA sia fondato. In quanto sistemi software con una struttura specificata formalmente, i sistemi di IA sono entità informazionali, non empiriche: non dovremmo concluderne che le loro proprietà si possono valutare, tipicamente calcolandole in qualche modo, ma non misurare? Per fare un esempio non equivoco, il numero di divisori di un certo numero intero è una sua proprietà quantitativa che sappiamo valutare in modo oggettivo e intersoggettivo, e nondimeno non diremmo che stiamo facendo una misurazione quando stiamo valutando tale proprietà.
D’accordo: la misurazione è un processo che progettiamo e realizziamo, non un processo naturale, e dunque non ha molto senso cercare il “significato vero” del termine “misurazione”. Ma è chiaro che non ci sono benefici nel perdere di specificità semantica, trattando “valutazione” e “misurazione” come sinonimi quando i loro significati sono stati invece abitualmente distinti. Per fare un altro esempio, le funzioni matematiche si valutano: perché dovremmo sostenere che si misurano?
Nel mondo del software la distinzione tra valutazione e misurazione non è però mantenuta in modo così chiaro, e infatti per esempio è abituale leggere l’espressione “misurare il numero di linee di codice di un programma”, una proprietà per cui potrebbe esserci una qualche incertezza di definizione (si contano le linee di commento? ecc.) ma che evidentemente non è empirica. Proprio relativamente alla distinzione empirico -

informazionale, i sistemi software sono entità interessanti: è informazionale il loro codice sorgente, che infatti non ha una definita collocazione spazio-temporale (e dunque in questo senso non è giustificato considerare che il numero di linee di codice sia una proprietà misurabile), ma nel momento in cui il codice viene eseguito, dunque da un particolare sistema hardware in particolari condizioni ambientali, se ne ottiene un sistema con un comportamento empirico. In più, il comportamento di un sistema di ML deriva dalla combinazione di fattori così complessi (la struttura del sistema, la procedura e i dati di addestramento, il sistema hardware sottostante, l'interazione con l'utente) da poter essere propriamente caratterizzato come un fenomeno empirico. Pertanto, sì, ha senso cercare di applicare quella che in precedenti articoli di questa rubrica abbiamo chiamato la "cultura metrologica" (per esempio in "Il ruolo sociale della cultura metrologica: qualche ipotesi", https://issuu.com/tutto_misure/docs/tm.1-2022) per caratterizzare il comportamento dei sistemi di ML, con l'obiettivo di acquisire su di esso informazioni sufficientemente oggettive e intersoggettive.

Una seconda premessa. Un sistema di ML può essere un'entità complessa, con un comportamento complesso, la cui qualità può essere complessa da misurare, anche perché si tratta di una proprietà multidimensionale. Contribuiscono infatti a essa molteplici fattori: la robustezza del sistema, la quantità di energia che l'hardware sottostante richiede per funzionare, la rapidità con cui risponde alle richieste che gli vengono inviate, la ripetibilità delle risposte che fornisce, e così via. Dunque, *ceteris paribus*, è di migliore qualità un sistema più robusto, che richiede meno energia per funzionare, e così via. Per ognuno di questi fattori si pone di principio una questione di misurabilità. Nell'analisi che segue ci limitiamo a considerare un concetto complessivo di qualità, nella logica secondo cui la qualità di un sistema è la sua capacità di realizzare ciò per cui è stato progettato: come possiamo misurare la *quality as fitness for purpose* dei sistemi di AI, dunque?

Il comportamento di un sistema di ML è il risultato non solo della costruzione / programmazione della sua struttura ma anche dell'addestramento a cui è stato sottoposto. Pur con tutte le ovvie differenze del caso, c'è perciò un senso nello studiare una risposta a questa domanda in analogia a quello che accade nella relazione tra docenti e studenti: i docenti valutano (misurano?) la qualità del "comportamento cognitivo" degli studenti in funzione di quello che si aspettano da loro; possiamo comportarci analogamente, appunto, a proposito di quei peculiari "studenti artificiali" che sono i sistemi di ML?
Seguendo ancora una volta un metodo analitico, distinguiamo tre categorie generali di sistemi di ML, ognuna delle quali con proprie caratteristiche e condizioni a proposito della misurabilità della qualità del comportamento dei sistemi stessi.

La prima categoria contiene i sistemi di ML progettati per calcolare previsioni nella forma di classificazioni o regressioni, quando la variabile obiettivo è categorica o numerica rispettivamente (esempi: k-nearest neighbors (KNN), regressione logistica, alberi delle decisioni, ma anche già reti neurali). Esempi di applicazioni in questa categoria sono i sistemi di riconoscimento di caratteri scritti a mano, i sistemi di raccomandazione, i sistemi di analisi delle opinioni (*sentiment analysis*).
Prendendo in esame il caso più semplice, in cui si vuole prevedere come classificare degli individui in una di due classi, occorre disporre di dati "di addestramento" (*training set*), che includono per ogni individuo sia valori per una o più variabili in funzione delle quali classificare (*features*) sia il valore (classe A / classe B) della variabile obiettivo della classificazione (*target*). L'addestramento consiste allora nell'adattare il sistema in modo da renderlo capace di assegnare un valore alla variabile obiettivo di nuovi individui, in questo modo appunto classificandoli: si tratta dunque di una forma di taratura del sistema.
La misurazione della qualità del comportamento di questi sistemi è basata su strategie e tecniche ben note, dato il fatto che durante l'addestramento è possibile prevedere il valore della variabile obiettivo e confrontarlo con il dato corrispondente, che funge dunque da "valore vero" del misurando. Si può così valutare a priori per esempio l'accuratezza del comportamento del sistema, calcolata come il rapporto tra il numero di individui classificati correttamente e il numero di individui presi in esame nella classificazione, e

poi sintetizzata in indici come il numero di veri positivi, falsi positivi, ecc., eventualmente raccolti in una cosiddetta matrice di confusione.
Rimangono comunque di principio due problemi: uno, i possibili bias, nel caso in cui il training set non sia sufficientemente rappresentativo dell'intera popolazione, una situazione che può essere interpretata come una scelta non corretta dei campioni di misura; e due, la possibilità che la variabile obiettivo non sia stabile nella sua relazione con i classificatori (in statistica si direbbe che non si è "in condizioni IID", cioè con variabili indipendenti e distribuite identicamente), cosa che renderebbe obsoleti i risultati dell'addestramento e quindi inaffidabili i risultati della classificazione. In questo la metrologia potrebbe offrire qualche utile insegnamento, in particolare a proposito della caratterizzazione e della gestione dell'incertezza di misura.
La misurazione della qualità del comportamento di questa prima categoria di sistemi di ML è analoga, nella relazione tra docenti e studenti, alla valutazione dei test a scelta multipla. Certo, si tratta di strumenti con limiti significativi – a rischio di bias ("studiare per l'esame"), sottocampionamento, ecc. – ma se la qualità di un test può essere garantita, valutare le competenze degli studenti mediante quel test è un processo non problematico.

Nel contesto del machine learning oggi sono sempre più importanti i sistemi di IA generativa: è la seconda categoria della nostra analisi, considerando per ora sistemi senza capacità di mantenere il contesto (*context-free*) e quindi in grado solo di gestire un singolo scambio domanda-risposta. Esempi di applicazioni in questa categoria sono i sistemi che producono sommari di testi o immagini a partire da testi. L'esempio canonico di questa seconda categoria, quello da cui tra l'altro la GenAI si è sviluppata, sono però i sistemi di traduzione automatica, in cui data la richiesta di un testo in una certa lingua il risultato atteso è un testo corrispondente tradotto in un'altra lingua.
Valutare la qualità del comportamento di un sistema di questo genere è un problema ovviamente più complesso del precedente, ma anche in questo caso si può assumere l'ipotesi che per ogni richiesta esista una "risposta esatta", dunque nell'esempio una traduzione che sarebbe considerata corretta per ogni testo in esame. Un insieme di coppie (testo da tradurre, testo tradotto) è perciò utilizzabile per l'addestramento di un sistema di ML, in questo caso tipicamente una rete neurale, e in particolare una rete ricorrente (RNN, https://en.wikipedia.org/wiki/Recurrent_neural_network) fino a qualche anno fa e oggi usualmente un transformer (https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)), cioè una complessa funzione parametrica il cui addestramento ha lo scopo di assegnare valori appropriati ai parametri della funzione stessa.
Il confronto fra il risultato previsto dal funzionamento della rete e il risultato atteso, dunque corrispondente al confronto fra valore misurato e valore vero, è qui evidentemente più complesso che nel caso della categoria precedente, e richiede delle procedure specifiche. Sempre a proposito di traduzioni, un esempio è *bilingual evaluation understudy* (BLEU, https://en.wikipedia.org/wiki/BLEU), un algoritmo che valuta la similarità tra la traduzione prodotta da un sistema di ML (il valore misurato) e la traduzione di un traduttore umano professionista (il valore vero).
La misurazione della qualità del comportamento di questa seconda categoria di sistemi di ML è analoga, nella relazione tra docenti e studenti, alla valutazione della qualità di saggi, riassunti, traduzioni, …, un processo decisamente più complesso della valutazione di un test a scelta multipla, e per il quale stabilire criteri sufficientemente oggettivi e intersoggettivi non è così ovvio, ma che è già stato studiato e al cui proposito la psicometria ha già sviluppato strumenti di supporto, come le *construct maps*.

La terza categoria di sistemi di ML specializza la seconda, eliminando la condizione di indipendenza dal contesto grazie alla capacità di gestire conversazioni, così che ogni risposta può dipendere non solo dall'ultima richiesta posta, ma dall'intero contenuto della conversazione fino a quel momento, possibilmente inclusi i documenti che sono stati fatti leggere. Sistemi con questa capacità sono i chatbot, da qualche mese diventati simbolo della GenAI, come ChatGPT, Gemini, Copilot, Claude, Perplexity, Llama, Mistral, e così via.

Nell'interazione con un chatbot, la successione delle richieste non può essere completamente standardizzata, dato che ogni nuova richiesta in una conversazione potrebbe dipendere dalle risposte che il chatbot ha fornito in precedenza, una condizione con qualche analogia con l'effetto fisico di isteresi ma che evidentemente può produrre effetti ben più complessi di quest'ultimo. Ciò rende la definizione stessa di cosa sia il misurando (la citata "qualità del comportamento") una questione elusiva: contribuiscono, e come, la correttezza semantica, la specificità, la profondità, l'organizzazione dei contenuti, …?
La misurazione della qualità del comportamento di questa terza categoria di sistemi di ML è analoga, nella relazione tra docenti e studenti, alla valutazione della qualità di esami orali: in questo caso, l'obiettivo di valutare in modo sufficientemente oggettivi e intersoggettivi è davvero arduo. Non per nulla, i benchmark con cui sono valutati i chatbot, e più in generale i "modelli di linguaggio" (due esempi: AI2 Reasoning Challenge (ARC, https://arxiv.org/abs/1803.05457) e HellaSwag (https://arxiv.org/abs/1905.07830); per modelli "aperti" si veda in particolare la Open LLM Leaderboard di Hugging Face, https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard) assumono tipicamente un singolo scambio richiesta-risposta, evitando così a priori le complessità di contesto che lo sviluppo di una conversazione genera. Questa limitazione amplifica poi il rischio di addestramento "per adattamento ai test", una forma di *overfitting* notoriamente sintetizzata nella cosiddetta "legge di Goodhart: "quando un criterio di valutazione diventa un obiettivo, cessa di essere un buon criterio di valutazione". È perciò che può diventare sensato rinunciare a una procedura di misurazione analitica, e operare "a scatola chiusa" limitandosi a raccogliere statistiche sulle preferenze degli utenti nel confronto tra coppie di modelli, mantenuti anonimi per evitare condizionamenti, come fa LMSYS Chatbot Arena Leaderboard (https://chat.lmsys.org/?leaderboard) impiegando il sistema di valutazione Elo (https://it.wikipedia.org/wiki/Elo_(scacchi)).

Se tutto ciò non fosse già abbastanza complesso, ricordiamo infine che, in logica di "qualità totale", la *quality as fitness for purpose* non è ancora l'obiettivo ultimo, dato che lo scopo (*purpose*) stesso potrebbe diventare oggetto di valutazione, cosa che nel caso dei chatbot ha a che vedere con i criteri e i contenuti del loro addestramento. I chatbot, "macchine da conversazione", sono infatti inevitabilmente ideologici nella loro interazione, e come decidere se una certa ideologia è appropriata o no è evidentemente una questione interamente extra-metrologica.
Questi sistemi ci mettono di fronte a tante sfide, dunque…1610