Riflessioni sulla caratterizzazione metrologica dei sistemi di Machine Learning per la classificazione
Luca Mari, Dario Petri

Tutto_Misure, 3, 2025

[20.8.25]
La crescente diffusione dei sistemi di intelligenza artificiale nella nostra società rende sempre più importante *valutare in modo appropriato la qualità dell'informazione che essi producono*, e a questo scopo quella che abbiamo chiamato *cultura metrologica* (per esempio in "Il ruolo sociale della cultura metrologica: qualche ipotesi", pubblicato nel numero 1-2022 di Tutto_Misure) può avere un ruolo significativo. L'intelligenza artificiale è però una disciplina assai ampia (come abbiamo mostrato nella serie "Un'introduzione all'Intelligenza Artificiale Generativa", e in particolare nella prima parte, pubblicata nel numero 4-2024 di Tutto_Misure), che oggi si sviluppa a partire dalle numerose realizzazioni dei sistemi di Machine Learning, entità a comportamento appreso invece che programmato.
Proponiamo qui qualche appunto in vista di una caratterizzazione metrologica sistematica dei sistemi di Machine Learning, e lo facciamo cominciando dal caso strutturalmente più semplice dei sistemi finalizzati a *classificare* l'informazione che viene fornita loro, un tema che ci riconduce alla metrologia delle proprietà classificatorie (*nominal properties*, in inglese; ne abbiamo discusso in "Verso un'incertezza di classificazione - La cultura metrologica nella valutazione delle proprietà classificatorie", pubblicato nel numero 2-2019 di Tutto_Misure).

Qualche premessa è opportuna, per chiarire il contesto della nostra analisi.
Una classificazione è un processo formalizzato mediante una funzione, tipicamente non iniettiva, $f : X \to Y$, in cui X è l'insieme delle entità da classificare e Y è l'insieme mediante cui si classifica, e su cui si assume non sia definita una struttura algebrica (o se definita non sia utilizzata per definire la funzione: altrimenti il processo così formalizzato sarebbe più che solo una classificazione, per esempio un ordinamento o una regressione).
La funzione f è usualmente intesa come la formalizzazione sia della proprietà degli oggetti, elementi di X, per cui si classifica, sia del processo di classificazione; perciò si può scrivere per esempio $colore(oggetto) = rosso$ per intendere che quel certo oggetto ha un certo colore e che quel colore è stato classificato come rosso.
A volte Y è presentato come un insieme di "etichette" (*label*, in inglese: un termine usato frequentemente nel Machine Learning), o di "nomi" (in riferimento al termine "*nominal scale*" usato da Stevens nella sua classificazione dei tipi di scale), ma le cose non stanno così; classificare oggetti colorati etichettandoli con "rosso" o "red" non è sufficiente a produrre classificazioni diverse. La notazione non è perciò $colore(oggetto) = \text{rosso}$, ma appunto $colore(oggetto) = rosso$, che è effettivamente un'equazione: l'oggetto è di colore rosso, e non (solo) il nome (italiano) del colore dell'oggetto è "rosso".
Gli elementi di Y non sono nemmeno le classi in cui si classifica: una classe è un insieme, e Y non è un insieme di insiemi: è invece un insieme di "proprietà rappresentative", una per ogni classe della classificazione, e designate con simboli o altre entità linguistiche (nell'esempio, rosso è un colore e non una classe di oggetti, usualmente designato in italiano con "rosso" e in inglese con "red"). Ciò mostra che la funzione f associa oggetti dotati una certa proprietà a proprietà rappresentative, e che l'equazione corrispondente dovrebbe essere scritta più correttamente

$$f(x) = y \in Y$$

per esempio

$$colore(oggetto) = rosso \in \{bianco, nero, rosso, blu, altro\_colore\}$$

come per altro accade anche per le valutazioni in scala ordinale (per esempio per riportare il gradimento di un certo oggetto in scala Likert occorre indicare appunto la scala: $gradimento(oggetto) = 3 \in \{1,2,3,4\}$ è ovviamente diverso da $gradimento(oggetto) = 3 \in \{1,2,\ldots,9,10\}$). Con ciò, per mantenere esplicito il

riferimento lessicale al fatto che ci stiamo occupando di classificazioni, nel seguito e quando la cosa non genera ambiguità, chiameremo "classi" gli elementi di Y.

L'assenza di una struttura algebrica su Y implica che le (proprietà rappresentative, e perciò le) classi possibili, debbano essere elencate esplicitamente, e quindi siano in numero finito. D'altra parte, Y può essere anche complesso, come quando la classificazione è

– *multiclasse* (come nel caso in cui è dato un insieme Z di classi "di base" e le classi in Y sono sottoinsiemi di Z, e quindi $Y = 2^Z$: se per esempio Z è un insieme di colori, Y contiene tutte le combinazioni di tali colori) o
– *per classi strutturate* (di nuovo, è dato un insieme Z di classi "di base" e le classi in Y sono strutture di Z, per esempio successioni di elementi di Z: è questo, per esempio, il caso delle traduzioni, che possono essere interpretate come classificazioni in cui Z contiene le parole della lingua di arrivo e Y le frasi di quella stessa lingua).

A proposito di X, nei casi semplici è un insieme di oggetti che hanno una proprietà classificatoria, come il colore o la forma geometrica, e questo rende la classificazione analoga a una misurazione:

– in una misurazione, $f : X \to Y$ corrisponde a una funzione da un insieme di oggetti che hanno una grandezza a un insieme di valori della grandezza, per esempio da un insieme di oggetti con lunghezza a un insieme di valori in metri;
– in una classificazione, $f : X \to Y$ corrisponde a una funzione da un insieme di oggetti che hanno una proprietà classificatoria a un insieme di classi della classificazione, per esempio da un insieme di oggetti colorati a un insieme predefinito di colori.

Ma anche X può essere complesso, e può avere una propria struttura, come nel caso di un'immagine il cui contenuto è da classificare o di un testo da tradurre. E in questi casi è poi da chiarire se l'oggetto della classificazione sia l'entità di partenza (l'immagine, il testo) o l'entità ottenuta da una sua codifica (l'elenco delle caratteristiche – *feature*, in inglese – riconosciute nell'immagine o la successione dei vettori di embedding che codificano numericamente le caratteristiche semantiche delle parti identificate nel testo).

Un'ulteriore complessità può poi manifestarsi nella funzione di classificazione, che – come accenneremo sotto – potrebbe associare ogni elemento di X non a un singolo elemento di Y, ma a una distribuzione di probabilità su Y, e quindi la classificazione potrebbe essere realizzata in due fasi, in cui, data la distribuzione prodotta nella prima fase, la seconda fase consiste nella scelta di una classe, tipicamente la classe che compare più frequentemente nella distribuzione, cioè la sua moda.

Anche il caso strutturalmente più semplice dei sistemi di Machine Learning, appunto quello dei sistemi di classificazione, può dunque presentare elementi di significativa complessità dal punto di vista metrologico, a partire dalla questione della definizione dell'oggetto per cui si classifica, ciò che nel caso delle grandezze corrisponde alla proprietà che si intende misurare, cioè il misurando. A volte è chiaro che l'oggetto per cui si classifica è una proprietà (per esempio la forma dei caratteri alfanumerici quando si fa *optical character recognition*, OCR) e perciò la classificazione è interpretabile come una valutazione di tale proprietà, cioè come l'assegnazione a tale proprietà di un valore, che in questo caso è la proprietà rappresentativa della classe a cui la proprietà è riconosciuta appartenere (per esempio la forma di un certo carattere alla classe della lettera "a"). Ma altre volte, come nel menzionato caso delle traduzioni, che la classificazione sia una forma di valutazione (se si sta traducendo dall'italiano all'inglese, l'articolo "the" è il valore dell'articolo "il"?) è meno ovvio.

Con ciò, cominciamo la nostra analisi prendendo in considerazione esempi in cui le classificazioni sono interpretabili in modo non controverso come valutazioni di proprietà (classificatorie, ovviamente).

Analogamente a quanto accade per i sistemi di misura e la misurazione, anche a proposito di sistemi di classificazione e del processo di classificazione è importante distinguere tra *valutazioni della qualità del comportamento dei sistemi* e *valutazioni della qualità dei risultati di tale comportamento*.

Ovviamente, la qualità del comportamento di un sistema influenza – perché statisticamente limita – la qualità dei risultati che se ne possono ottenere. Inoltre, la qualità dei risultati può essere valutata senza preoccuparsi,

almeno di principio, del processo con cui sono stati ottenuti, dunque "a scatola chiusa", ossia a prescindere dal fatto che il sistema sia di Machine Learning o di altro genere, mentre per valutare la qualità del comportamento di un sistema è generalmente utile osservare il comportamento stesso, dunque "a scatola aperta".
Con questa precisazione, proseguiamo queste *riflessioni* con qualche considerazione su come può essere valutata la qualità dei risultati di un processo di classificazione.

Conoscere la quantità di informazione ottenuta con una classificazione è chiaramente importante quando il risultato ottenuto è utilizzato per supportare una decisione. A tal fine, analogamente a quanto avviene per la misurazione, è appropriato formalizzare il risultato in termini probabilistici. Dato che le classi nell'insieme Y sono in numero finito, si può descrivere in modo completo l'informazione fornita da una classificazione con una funzione di massa di probabilità (*probability mass function*, PMF, in inglese) $P_i = \{(y_j, p(y_j|x_i))\}$, che stabilisce la probabilità $p(y_j|x_i)$ che l'oggetto x_i appartenga alla classe y_j in Y.
Per esempio, il risultato di un processo di OCR per un certo carattere x_i potrebbe essere $P_i = \{("a", 0.75), (d, 0.20), (o, 0.05), (\text{ogni altro carattere}, 0.00)\}$.
Tale PMF è ottenibile mediante un processo di taratura del sistema di classificazione, che negli aspetti essenziali coincide con quello di uno strumento di misura. Ma mentre si assume tipicamente che uno strumento di misura abbia una funzione di trasduzione / taratura con una forma analitica semplice, e quindi siano sufficienti pochi punti di taratura (per esempio due nel caso di una funzione lineare), l'assenza di struttura algebrica su Y impone che la taratura sia eseguita per ogni classe y_j in Y, per far sì che il sistema sia in grado di classificare correttamente ogni oggetto x_i candidato. È dunque chiaro che nel caso di un sistema di Machine Learning la taratura si realizza nell'addestramento (*training*, in inglese) del sistema.
Il fatto che, per ogni oggetto x_i da classificare, un sistema di classificazione tarato produca non una singola classe, ma una distribuzione di probabilità è l'indicazione del riconoscimento della presenza di incertezza strumentale nel processo, che, analogamente a quanto raccomandato dalla *Guida all'Espressione dell'Incertezza nella Misurazione* (la GUM) può avere varie cause / componenti, valutabili con metodi statistici ("di categoria A") o di altro genere ("di categoria B"). Anche nel caso delle classificazioni, è comunque in generale da considerare la presenza di un'incertezza di definizione della proprietà da classificare (che nel caso di un processo di OCR potrebbe manifestarsi nella presenza di caratteri di forma inerentemente ambigua), che opera come un limite inferiore all'incertezza di classificazione ottenibile (se anche a un lettore umano non è chiaro se un certo carattere sia una a o una d, un comportamento certo da parte di un classificatore automatico sarebbe sospetto…).

Stabilire e poi riportare come risultato di una classificazione un'intera distribuzione di probabilità potrebbe essere comunque complesso, e spesso non necessario, essendo tipicamente sufficiente considerare solo la parte più rilevante dell'informazione associata alla distribuzione, per esempio riportando solo la classe moda con la relativa probabilità di classificazione corretta (ossia $("a", 0.75)$ nell'esempio precedente), oppure un sottoinsieme di classi con la relativa probabilità complessiva (per esempio, sempre con riferimento all'esempio sopra, $\{"a","d"\}$ con probabilità $0.75 + 0.20 = 0.95$). Questa seconda modalità di esprimere il risultato di una classificazione è analoga a operare con intervalli e probabilità di copertura nella misurazione, e può essere particolarmente utile quando l'informazione ottenuta ha lo scopo di supportare una decisione.
Come nel caso dell'incertezza di misura, può essere appropriato riassumere l'informazione associata alla distribuzione di probabilità relativa alla classificazione dell'oggetto x_i mediante un singolo indice di incertezza come, in particolare, l'entropia di Shannon, definita come $H_i = -\sum_j p(y_j|x_i) \log_2 p(y_j|x_i)$. Il valore massimo di H_i, pari a $\log_2 n$, con n pari al numero di classi in Y, si ottiene nel caso di massima incertezza, quando la distribuzione è uniforme e quindi la classificazione non ha prodotto alcuna

informazione, mentre il valore minimo, 0, si ottiene nel caso di minima incertezza, quando la proprietà da classificare è assegnata con certezza a una singola classe e quindi la classificazione ha prodotto l'informazione più specifica possibile dato l'insieme Y scelto per la classificazione (si noti che l'entropia è calcolabile anche per distribuzioni di probabilità relative a proprietà quantitative: una ragione per cui per queste si tende a sintetizzare l'informazione mediante l'incertezza tipo – cioè la deviazione standard – è che questa è un indice relativo alla media della distribuzione, e quindi esplicitamente al valore misurato, mentre l'entropia non dipende dalla moda della distribuzione).

Concludiamo con un cenno a proposito della valutazione della qualità del comportamento dei sistemi di classificazione. Analogamente a quanto accade per la caratterizzazione metrologica del comportamento di un sistema di misura, in questo caso si suppone di conoscere le classi in cui gli oggetti dovrebbero essere classificati, dunque i "valori veri" delle proprietà, corrispondenti alle "etichette" nei dataset utilizzati per l'addestramento dei sistemi di Machine Learning. Questo rende possibile calcolare prima di tutto le matrici di confusione – che nel caso di classificazioni binarie corrisponde a contare il numero di classificazioni corrette (veri positivi e veri negativi) e non corrette (falsi positivi e falsi negativi) – e da queste, indici come l'accuratezza (il rapporto tra numero di classificazioni corrette e di classificazioni complessive), la precisione (il rapporto tra numero di veri positivi e di positivi), e così via.
Tali indici forniscono però solo un'informazione sulla qualità "media" dei risultati forniti da un classificatore, e non specificamente sulla qualità del risultato fornito nel caso di una singola classificazione, perché ovviamente in tal caso la "classe vera" non è nota: analogamente a quanto accade nella misurazione, per valutare la qualità dei risultati che si ottengono dall'uso di questi sistemi si ricorre non all'accuratezza, ma all'incertezza.