Machine learning

A CURA DI LINO POLO

Idea di fondo

Il machine learning permette di creare un software che impara come fa un bambino, cioè tramite l'induzione di principi generali a partire dall'osservazione dei dati e con la capacità di ottenere nuova conoscenza a partire dall'informazione a disposizione. Un bambino nasce senza conoscere niente, però sa come studiare per apprendere, quindi va a scuola da un maestro, studia per imparare, si crea un'esperienza, svolge un esame, ottiene un voto per valutare che cosa ha imparato, se il voto è basso ripete l'anno scolastico. Nello stesso modo, una soluzione di machine learning può studiare, fare esperienza, scoprire le regolarità statistiche che si celano nei dati. In tal modo, consegue l'importante possibilità di generalizzare, cioè funzionare con la stessa efficacia anche con dati non considerati in precedenza.

Un'altra analogia con l'agricoltura può far comprendere il ruolo delle parti coinvolte. Gli algoritmi di apprendimento sono i semi, i dati sono il terreno, i modelli che apprendono sono le piante adulte. Lo sviluppatore di machine learning è come l'agricoltore: semina, irriga e concima il terreno, controlla lo stato di salute del raccolto, ma per il resto non interferisce.

Meccanismo di funzionamento

Esistono varie definizioni di machine learning, traducibile come apprendimento automatico. Quelle meno formali indicano il machine learning come la scienza per far apprendere i dati al computer senza che sia stato esplicitamente programmato con delle regole. Un altro modo indica il machine learning come un mapping (corrispondenza) tra valori di input e output, che può essere modellato con una funzione non lineare. Data la non linearità, conviene impiegare un approccio di machine learning basato su funzioni y(x;w) che contengono parametri w modificabili secondo i dati x. Il programmatore non deve specificare tutti i parametri w che caratterizzano il dato compito, perché questi vengono trovati tramite il modello.

Meccanismo di funzionamento

Una definizione formale viene espressa nel modo seguente. Dato un insieme di addestramento di N esempi definiti come coppie di dati input e output associato (x1,y1), (x2,y2), …, (xN,yN) dove yi è stato generato da una funzione non nota y=f(x), trovare una funzione h che approssimi la funzione sconosciuta f. La funzione h è l'ipotesi, il machine learning cerca la funzione h che meglio spiega i dati in uscita e che meglio opera su dati nuovi non usati nell'addestramento. Si tratta, perciò, di apprendimento induttivo basato esclusivamente sull'osservazione dei dati esistenti.

Vantaggi e svantaggi

Prima di decidere l'uso degli strumenti offerti dal machine learning è bene avere chiaro cosa c'è di buono e cosa c'è di cattivo, per sapere a cosa si va incontro ed evitare problemi.

I vantaggi riguardano gli aspetti seguenti:

ridotto intervento umano;

possibilità di miglioramento continuo;

facile identificazione di trend e pattern;

può gestire dati con molte dimensioni e varia tipologia;

può risolvere molte applicazioni;

ampia letteratura tecnica e software;

disponibilità di vari modelli da adeguare alle necessità.

Contesti ideali

le regole e le equazioni sono troppo complesse da scrivere, per esempio nel caso del riconoscimento facciale;

le regole per descrivere un'attività cambiano continuamente, per esempio nell'analisi dei virus e dei malware nel software;

bisogna adattarsi alla continua evoluzione della natura dei dati, per esempio nella descrizione dell'andamento dei titoli alla Borsa di Milano;

Gli svantaggi riguardano gli aspetti seguenti:

il machine learning è un problema matematico mal posto, cioè un problema che non ammette una soluzione unica, per cui bisogna fare diverse prove sperando di trovare qualche scelta a proprio favore;

può non essere facile comprendere perché è stata data una specifica risposta (machine learning explainability), un aspetto necessario per correggere errori e aiutare una persona che deve decidere;

Contesti ideali

nessun metodo è il migliore per casi sia generali sia specifici, bisogna provare molti algoritmi e alla fine selezionare quello che offre le migliori prestazioni;

mancanza di alta precisione nel dato ottenuto, difficilmente si ottengono valori netti come 100% oppure 1, bensì valori del tipo 99% o 0,9 e similari;

richiede tempo e risorse di calcolo;

bisogna dare un'adeguata interpretazione degli input e fare un'attenta interpretazione del risultato;

servono molti dati con elevata qualità;

è difficile trovare un compromesso tra velocità, accuratezza, complessità;

è difficile identificare le fonti di rumore, dovute agli errori presenti nei dati, o agli eventi casuali che non si possono facilmente prevedere;

Differenza rispetto ad altri approcci

Programmazione esplicita

Un programmatore scrive esplicitamente un programma considerando tutte le possibili diverse condizioni in cui possono trovarsi i dati. Il machine learning può ricavare le regole per calcolare il risultato osservando gli esempi di input e corrispondente output, senza nessun intervento umano; le regole apprese vengono ulteriormente utilizzate per migliorare il processo decisionale quando vengono immessi nuovi dati di input.

Ciò che viene realizzato dal programmatore è adeguato solo per lo specifico problema; un approccio con il machine learning può essere adattato a diversi problemi grazie alla sua flessibilità.

Nell'approccio tradizionale si prendono il dato in input e un algoritmo che indica come elaborarlo, si crea il codice con un linguaggio di programmazione, viene eseguito il codice con l'input e si ottiene il risultato in output.

Programmazione esplicita

Con il machine learning si procede in questo modo:

1. si prendono coppie di dati come input e output associato, oppure solo input, un modello matematico che descrive un modo di ragionare dell'AI, per esempio il neural network, un algoritmo per realizzare la computazione con il modello scelto, che indica come calcolare il neurone, la rete, back propagation ecc.;

2. con un linguaggio di programmazione e un ambiente di sviluppo si crea il codice necessario per realizzare concretamente il modello e farlo apprendere, per esempio si calcolano le variabili contenenti i pesi tra i neuroni secondo i dati;

3. si ottiene il codice pronto per essere messo in esercizio, composto dalle istruzioni da eseguire e dal valore delle variabili ottimizzate secondo il problema e il modello scelto, in cui mettere un nuovo input e calcolare un output adeguato.

Data mining

Il termine data mining è basato sull'analogia con le operazioni dei minatori che scavano all'interno delle miniere grandi quantità di materiale di poco valore per trovare l'oro. Nel data mining, l'oro è l'informazione, precedentemente sconosciuta o indiscernibile; le operazioni di scavo sono le tecniche di esplorazione dei dati, tra cui quelle fornite dal machine learning, molto utili per analizzare i dati ed estrarre informazioni utili.

Qualche differenza si può notare pensando al machine learning che utilizza algoritmi per migliorare l'esperienza con un determinato compito, mentre il data mining si concentra sull'analisi dei dati per scoprire modelli o proprietà mai visti prima e applica una gamma più ampia di algoritmi. Il machine learning si concentra sullo studio e sulla riproduzione di conoscenze specificamente note, mentre il data mining è esplorativo e cerca conoscenze sconosciute.

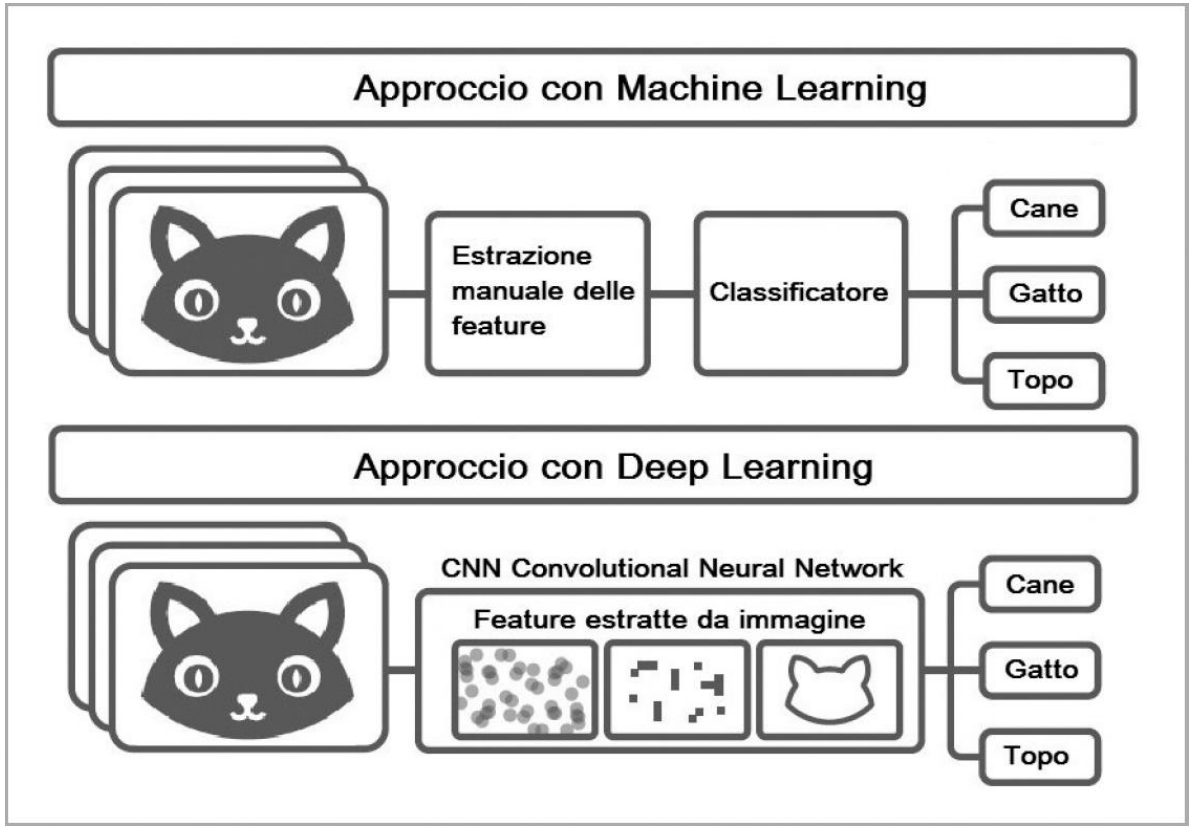
Data science

Data science è un termine generico che include data analytics, data mining, machine learning e altre competenze specifiche per fare analisi di dati relativa al proprio contesto di lavoro. Data science include diversi aspetti della gestione dei dati, come l'acquisizione dei dati da una o più fonti, la pulizia dei dati, la preparazione dei dati, la creazione di nuovi punti dati basati su dati esistenti, l'esecuzione dell'analisi dei dati. Comprende anche l'utilizzo di tecniche di machine learning sui dati per dedurre le conoscenze con cui creare un algoritmo che esegue un'attività su dati non ancora considerati.

Deep learning

Il deep learning, costituisce un sottoinsieme molto particolare. Viene usato il concetto di apprendimento, cambia sostanzialmente il modo con cui vengono estratte le informazioni per descrivere cosa bisogna apprendere. Per esempio, si può considerare l'immagine di un gatto e le possibilità espresse nella Figura. Il machine learning deve estrarre feature riguardanti il colore, la forma più o meno rotonda della testa, il contorno formato dall'oggetto, la posizione e la forma degli occhi ecc. riguardanti caratteristiche scelte dall'analista, come input in un classificatore da scegliere in maniera apposita per ottenere la risposta. Al contrario, il deep learning permette di estrarre una vasta serie di feature molto semplici in maniera automatica direttamente sull'immagine, da elaborare nei successivi livelli neurali fino a ottenere la risposta.

Deep learning



Categorie di soluzioni

Il machine learning ormai è in grado di essere applicato a un'enorme quantità di contesti, tanto

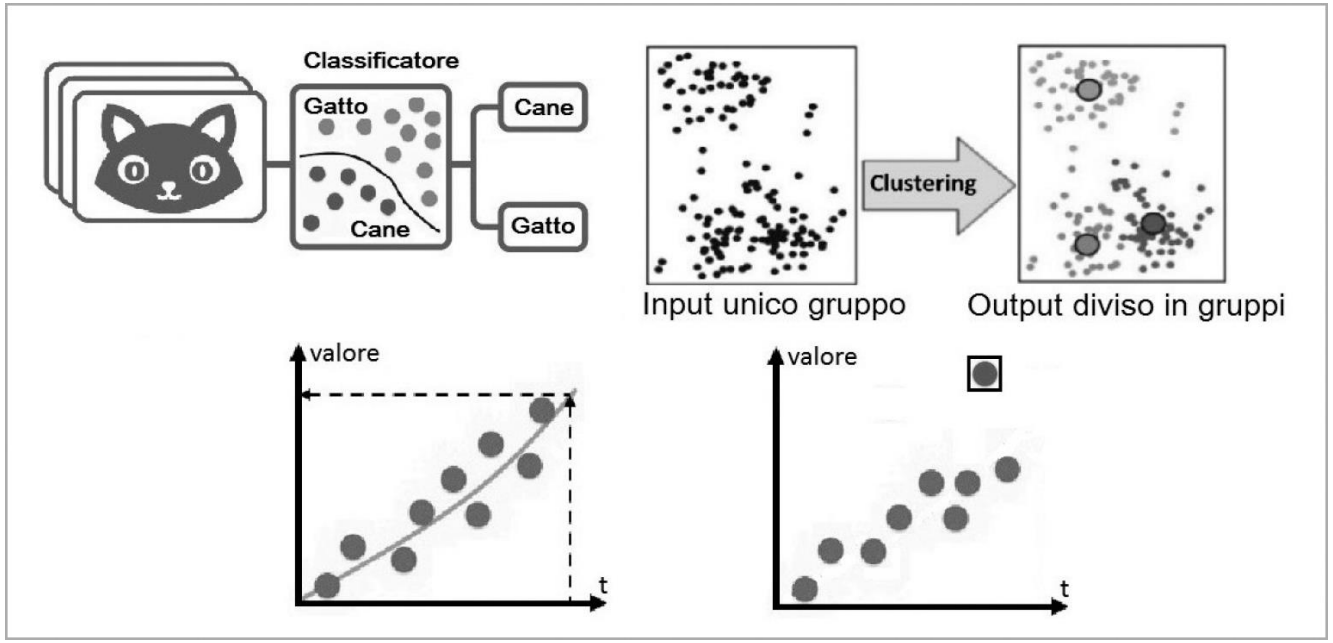
che è difficile riepilogarli in poche pagine. Per avere qualche idea, si possono cercare sul motore di ricerca le parole "review machine learning", "state of art machine learning", "best machine learning", "current applications of machine learning" seguite dalle parole che descrivono il problema da risolvere.

Le principali soluzioni rientrano in queste grandi categorie, in base al tipo di output desiderato.

Classificazione

Gli input sono divisi in due o più classi e il sistema deve produrre un modello in grado di assegnare a un input una o più classi tra quelle disponibili, per esempio la classificazione di immagini secondo gli oggetti contenuti. La variabile di output desiderata è categorica, qualitativa nominale (per esempio, il nome di ogni classe da considerare) o ordinale (per esempio, 1, 2 ecc. per indicare prima classe, seconda classe ecc.). Il classificatore indica il sistema incaricato di eseguire la classificazione. La classificazione è binaria se bisogna distinguere tra due classi, è multi-classe se bisogna distinguere tra molte classi. Per esempio, nella classificazione binaria delle email un servizio per trovare lo spam ha due output come (è spam, non è spam), oppure (classe1, classe 2) e (positivo, negativo).

Classificazione



Clustering

Un insieme di dati viene raggruppato in cluster (gruppi) in modo che i dati nello stesso gruppo abbiano caratteristiche simili e dati in gruppi distinti abbiano caratteristiche molto diverse. È simile alla classificazione, però i gruppi non sono noti a priori e bisogna calcolarli trovando pattern nascosti o strutture intrinseche nei dati, per esempio la suddivisione in gruppi dei clienti secondo il loro comportamento. La variabile output desiderata è quantitativa, dovendo esprimere un numero che indica appartenenza del dato a un gruppo, oppure una misura di distanza per dire quanto il dato è lontano dal centro che rappresenta il gruppo.

Regressione

L’output ha un dominio continuo e non discreto, come accade nel dominio dei numeri reali, quindi è una variabile quantitativa. Per esempio: la predizione di un dato per stimare cosa accadrà in futuro in base a una serie temporale che descrive il passato, stimare il risultato di un sensore elettronico, stimare il numero mancante in una serie di dati.

Trovare anomalie

Con questo approccio si trova un dato diverso da tutti gli altri, corrispondente all'indicare qualcosa da cancellare, un rumore o un dato da evidenziare, per esempio una frode in pagamento da bloccare oppure l'immagine di un biscotto rotto da buttare. Si risolvono questi problemi facendoli rientrare nell'ambito di una classificazione tra due classi corrispondenti ad ammesso e non ammesso, ma se occorre una precisione maggiore è meglio studiare algoritmi specifici.

Forme di apprendimento

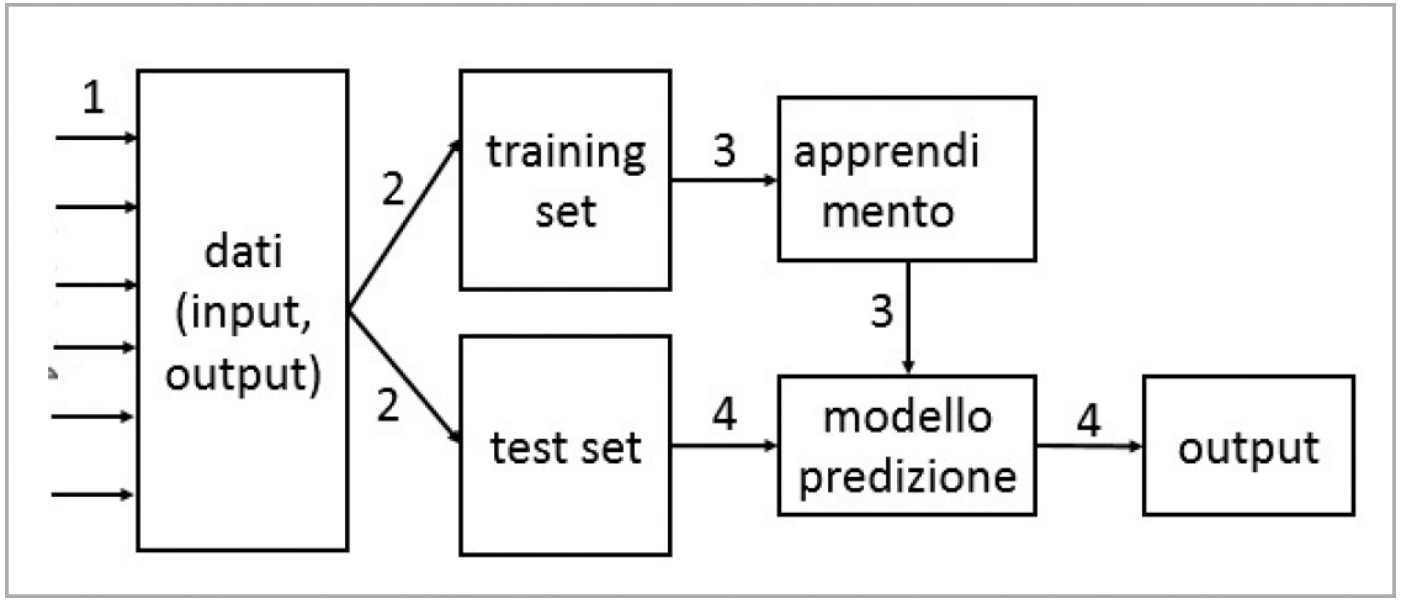
Le modalità con cui il machine learning permette agli algoritmi di fare **learning** (apprendimento, a volte indicato come **addestramento**) con i dati sono classificate in cinque categorie, caratterizzate dal tipo di feedback su cui si basa il sistema di apprendimento. Si tratta di un aspetto fondamentale nella progettazione, infatti viene subito specificato quando bisogna descrivere la soluzione a un problema.

Supervised learning

Nel **supervised learning** (apprendimento supervisionato) vengono presentati al modello scelto gli esempi formati dagli input e relativi output desiderati, per esempio un'immagine e un'etichetta (label) per descrivere l'oggetto contenuto, con lo scopo di apprendere una regola generale in grado di mappare gli input negli output, come se ci fosse un maestro che supervisiona lo studente mentre impara dalle sue lezioni.

Le coppie di input e output vengono divise in due gruppi distinti: training set formato dall'80% dei dati con cui addestrare l'algoritmo, test set formato dal restante 20% per valutare la prestazione, come un esame finale in cui viene dato un voto su quanto si è bravi. Bisogna stare attenti nell'inserire i dati in questi due insiemi: tutte le casistiche possibili devono essere coperte in entrambi gli insiemi, non devono esserci sbilanciamenti con molti dati riguardanti una classe e pochi altri riguardanti le restanti classi. Spesso l'assegnazione dei dati a questi due insiemi viene svolta con un'estrazione casuale, per non subire influenze indotte dall'esperto umano, più o meno consapevolmente. Permette di affrontare i problemi di classificazione e regressione descritti nei paragrafi precedenti.

Supervised learning



Durante il processo di training possono verificarsi due complicazioni. Se il modello scelto è troppo semplice le prestazioni sono scadenti, l'apprendimento è troppo scarso perché non viene compresa la realtà che ha generato i dati usati. In tal caso, si parla di **underfitting** (**sotto-apprendimento**).

Se il modello scelto è troppo complesso rischia di comprendere una versione distorta della realtà che ha generato i dati usati, perché apprende in maniera molto buona soltanto i dati usati e risponde molto male per i dati non ancora considerati. Si tratta di una mancanza di elasticità; il modello non è in grado di generalizzare per rispondere efficacemente a dati diversi, come quando a scuola la maestra mette un voto basso allo studente che ha imparato a memoria senza averci capito niente. In tal caso, si parla di **overfitting** (sovra-apprendimento).

Come si suol dire, la virtù sta nel mezzo; un modello può capire la realtà tanto da poterla descrivere e predire se viene creato con il giusto compromesso tra semplicità e complessità. Si può accettare un errore maggiore nel training set in cambio di un errore minore nel test set. In tal caso, si parla di well fitted (ben addestrato).

Unsupervised learning

Con **unsupervised learning** (apprendimento non supervisionato, indicato anche come **addestramento non supervisionato**) vengono forniti al modello scelto solo gli esempi formati dagli input, senza alcun output atteso, con lo scopo di fargli apprendere in autonomia una qualche struttura nei dati d'ingresso, come se lo studente potesse studiare da solo senza nessun maestro che lo guidi.

I risultati possono essere influenzati dalle decisioni su quali dati esporre all'algoritmo e in quale ordine.

Permette di gestire i problemi di clustering trovando i gruppi di dati in base alle caratteristiche simili; è utile anche per fare association learning, cioè trovare regole associative, come quelle in cui si determina che se una persona compra il prodotto A allora probabilmente preferisce comprare anche il prodotto B e si collegano i prodotti A e B.

Semi-supervised

Si possono combinare i due approcci precedenti con una prima fase supervised sui dati aventi input e output associato, e una successiva fase unsupervised su dati di cui non si conosce l'output associato. Gli input e gli output forniti forniscono il modello generale che si può estrapolare e applicare ai dati rimanenti.

Reinforcement

Con il reinforcement learning (apprendimento con rinforzo) si interagisce con un ambiente dinamico in cui raggiungere un certo obiettivo; a mano a mano che si esplora il dominio del problema vengono forniti dei feedback in termini di ricompense o punizioni secondo il comportamento eseguito, in modo da indirizzare verso la soluzione migliore.

Transfer

Il transfer learning (apprendimento con trasferimento) impara ad affrontare un certo problema generico, poi si prende la conoscenza creata per usarla nell'affrontare un altro problema simile o più specifico. Il grosso vantaggio consiste nel non rifare di nuovo il learning, nel risparmiare tempo, nell'avere una libreria di soluzioni pronte per essere adattate all'esigenza.