Capitolo 1

Introduzione al machine learning

Stando ai racconti di fantascienza, l'invenzione dell'intelligenza artificiale dovrebbe condurre inevitabilmente a guerre apocalittiche fra le macchine e i loro creatori. Le storie partono tutte dalla realtà di oggi: i computer imparano le regole dei giochi più semplici, come il *tris* (*tic-tac-toe*) e poi ad automatizzare dei compiti di routine. Ma il tempo passa, e alle macchine viene affidato il compito di controllare il traffico e le comunicazioni, poi i droni e infine i missili… L'evoluzione delle macchine entra in una tragica spirale quando i computer divengono senzienti e imparano a decidere e a imparare autonomamente. Non avendo quindi più bisogno di programmatori umani, non trovano nulla di meglio da fare che eliminare il genere umano.

Fortunatamente, almeno fino al momento in cui scrivo queste pagine, le macchine hanno ancora bisogno dei nostri input.

Anche se le impressioni relative al machine learning possano essere dipinte con questi colori sinistri, gli algoritmi impiegati oggi sono troppo specifici per pensare che possano essere generalizzati al punto di sviluppare un'autocoscienza. L'obiettivo del machine learning, oggi, non è quello di creare un cervello artificiale, ma piuttosto quello di assisterci nel nostro tentativo di comprendere la grande massa di dati oggi disponibili, e trarne informazioni utili.

Mettendo da parte le idee errate sul futuro, entro la fine di questo capitolo, dovreste avere una conoscenza un po' più chiara di che cos'è il machine learning. Introdurrò anche i concetti fondamentali

In questo capitolo

che definiscono e distinguono i più comuni approcci al machine learning. In particolare, esamineremo i seguenti argomenti.

- Le origini, le applicazioni e le trappole del machine learning.
- In quale modo i computer trasformano i dati in conoscenza e azione.
- I passi da svolgere per applicare un algoritmo di machine learning ai dati.

Il campo del machine learning fornisce un insieme di algoritmi che trasformano i dati in conoscenza pronta all'uso. Vedremo come è facile impiegare R per iniziare ad applicare il machine learning ai problemi del mondo reale.

Le origini del machine learning

Fin dalla nascita, siamo letteralmente sommersi di dati. I nostri sensori fisici, gli occhi, le orecchie, il naso, la lingua e i nervi, si trovano costantemente a essere raggiunti da dati grezzi che il nostro cervello traduce in immagini, suoni, odori, gusti e senso del tatto. Grazie alle parole, siamo anche in grado di condividere queste esperienze con altri.
Dall'avvento della lingua scritta, le osservazioni umane hanno potuto essere registrate. I cacciatori monitoravano i movimenti delle prede; i primi astronomi registravano l'allineamento dei pianeti e delle stelle; e i governanti registravano i pagamenti delle tasse, le nascite e le morti. Al giorno d'oggi, tali osservazioni, e molte altre ancora, sono sempre più automatizzate e registrate in modo sistematico all'interno di sempre più ricchi database. L'invenzione dei sensori elettronici ha ulteriormente contribuito a un'esplosione in termini di volume e ricchezza dei dati registrati. Sensori specializzati, come videocamere, microfoni, nasi chimici, lingue elettroniche e sensori di pressione imitano la capacità umana di vedere, udire, odorare, gustare e percepire sulla pelle. Questi sensori elaborano i dati in modo molto differente da quanto farebbe un essere umano. A differenza di quanto accade per l'attenzione limitata e soggettiva di un essere umano, un sensore elettronico non si ferma mai e non ha emozioni a controllare le sue percezioni.

> **NOTA**
>
> Sebbene i sensori non siano "inquinati" dalla nostra soggettività, non necessariamente presentano un'unica, definitiva rappresentazione della realtà. Alcuni hanno un errore di misurazione intrinseco, dovuto a limiti dell'hardware. Altri hanno dei limiti percettivi. Una fotografia in bianco e nero fornisce una rappresentazione differente di un soggetto, rispetto a una fotografia a colori. Analogamente, un microscopio fornisce una rappresentazione della realtà molto differente da un telescopio.

E così, fra database e sensori, molti aspetti della nostra vita vengono registrati. Governi, aziende e singole persone stanno registrando e studiando una grande massa di dati, da quelli più monumentali a quelli più banali. Sensori meteorologici registrano dati di temperatura e pressione; videocamere di sorveglianza osservano senza sosta marciapiedi e gallerie; e ogni genere di comportamento elettronico viene monitorato: transazioni, comunicazioni, relazioni sui social media e molto altro ancora.
Questo vero diluvio di dati ha condotto alcuni ad affermare che siamo entrati nell'era dei *big data*, ma quel nome forse non è del tutto corretto. Gli esseri umani sono sempre stati circondati da grandi quantità di dati. Ciò che rende unica la nostra era è che abbiamo

enormi quantità di dati registrati, molti dei quali accessibili direttamente tramite computer. *Dataset* sempre più grandi e interessanti sono sempre più accessibili e a portata di mano, tramite una semplice ricerca web. Questa grande ricchezza di informazioni ha la potenzialità di informare le nostre azioni, offrendoci un modo sistematico per acquisire il senso dei dati che abbiamo a disposizione.

Il campo di studi coinvolto nello sviluppo di algoritmi computerizzati in grado di trasformare i dati in azioni intelligenti si chiama *machine learning*: l'apprendimento automatico. Questo campo trae la sua origine in un ambiente in cui i dati disponibili, i metodi statistici e la potenza di calcolo crescono ed evolvono simultaneamente. La crescita nel volume dei data necessitava di maggiore potenza di calcolo, che a sua volta ha dato origine allo sviluppo di metodi statistici atti ad analizzare questi grossi dataset. Ciò ha creato un ciclo di avanzamenti tecnologici che ha consentito la raccolta di dati sempre più numerosi e interessanti, e arriviamo così all'oggi, in cui abbiamo a disposizione innumerevoli flussi di dati su, praticamente, ogni argomento immaginabile.

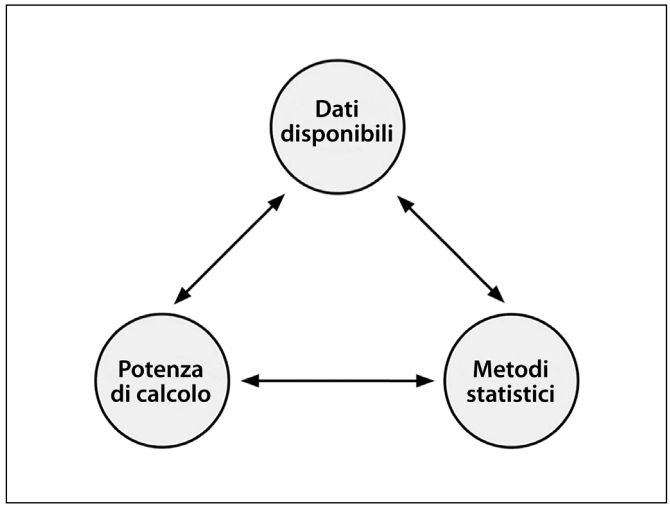


Figura 1.1 Il ciclo di avanzamento che ha reso possibile il machine learning.

Un parente stretto del machine learning, il *data mining*, si occupa della generazione di nuove conoscenze da grandi database. Il data mining prevede una ricerca sistematica di elementi di intelligenza utile. Sebbene vi siano alcuni disaccordi sull'entità della sovrapposizione fra machine learning e data mining, un possibile elemento di separazione consiste nel fatto che il machine learning si occupa di insegnare ai computer a utilizzare i dati per risolvere un problema, mentre il data mining si occupa di insegnare ai computer a identificare dei pattern, degli schemi che gli esseri umani utilizzeranno poi per risolvere un problema.

Praticamente tutte le attività di data mining prevedono l'impiego del machine learning, ma non tutte le attività di machine learning richiedono il data mining. Per esempio, potreste applicare il machine learning per eseguire il data mining di dati sul traffico automobilistico, alla ricerca di pattern correlati ai tassi di incidenti. D'altra parte, se il computer sta imparando a guidare un'auto, si tratterà di un compito di puro machine learning, senza data mining.

NOTA

Il termine "data mining" viene talvolta utilizzato anche in senso denigratorio per descrivere la pratica ingannevole di selezionare appositamente solo quei dati che operano a supporto di una teoria.

Usi e abusi del machine learning

Molti hanno sentito parlare di Deep Blue, un computer giocatore di scacchi che nel 1997 fu il primo a vincere una partita contro un campione del mondo. Un altro famoso computer, Watson, sconfisse due avversari umani in un gioco a quiz televisivo, Jeopardy, nel 2011. Sulla base di questi stupefacenti risultati, alcuni hanno ipotizzato che l'intelligenza dei computer sostituirà gli operatori umani nelle professioni che riguardano la tecnologia dell'informazione, così come le macchine hanno sostituito i lavoratori nei campi e nelle catene di montaggio.
La verità è che anche se le macchine hanno raggiunto risultati così impressionanti, sono tuttora relativamente limitate nella loro capacità di comprendere a fondo un problema. Esse sono pura potenza di calcolo intellettiva, ma senza direzione. Un computer può anche essere più abile di un essere umano di trovare subdoli schemi all'interno di grandi database, ma avrà comunque bisogno di un essere umano per motivare la sua analisi e trasformarne i risultati in un'azione significativa.

RIFERIMENTO

Senza sottovalutare del tutto i risultati conseguiti da Deep Blue e Watson, è importante notare che nessuno dei due aveva nemmeno lontanamente l'intelligenza di un bambino di cinque anni. Per ulteriori informazioni su perché "confrontare le intelligenze è una faccenda 'scivolosa'" consiglio la lettura dell'articolo di "Popular Science" *FYI: Which Computer Is Smarter, Watson Or Deep Blue?*, di Will Grunewald: `https://www.popsci.com/science/article/2012-12/fyi-which-computer-smarter-watson-or-deep-blue`.

Le macchine non sono brave a porre domande e nemmeno sanno quali domande porre. Sono molto più brave a rispondere alle domande, sempre che la domanda sia formulata in un modo che il computer sia in grado di comprendere. Gli attuali algoritmi di machine learning fanno per l'essere umano quello che un segugio fa per il suo addestratore: il senso dell'olfatto del cane può anche essere molto più sviluppato di quello del suo padrone, ma senza un'accurata indicazione, il segugio potrebbe inseguire la sua coda.
Per comprendere meglio le applicazioni del machine learning nel mondo reale, considereremo ora alcuni casi in cui è stato impiegato con successo, alcune aree in cui ci sono spazi di miglioramento e alcune situazioni in cui potrebbe fare più danni che altro.

I successi del machine learning

Il machine learning può avere particolare successo quando estende, piuttosto che sostituire, le conoscenze specializzate di un esperto sul campo. Può funzionare con i medici, in prima linea nel tentativo di debellare il cancro; assiste gli ingegneri e i programmatori nei loro tentativi di creare abitazioni e automobili più intelligenti; e aiuta gli scienziati sociali a costruire conoscenze sul funzionamento delle società.

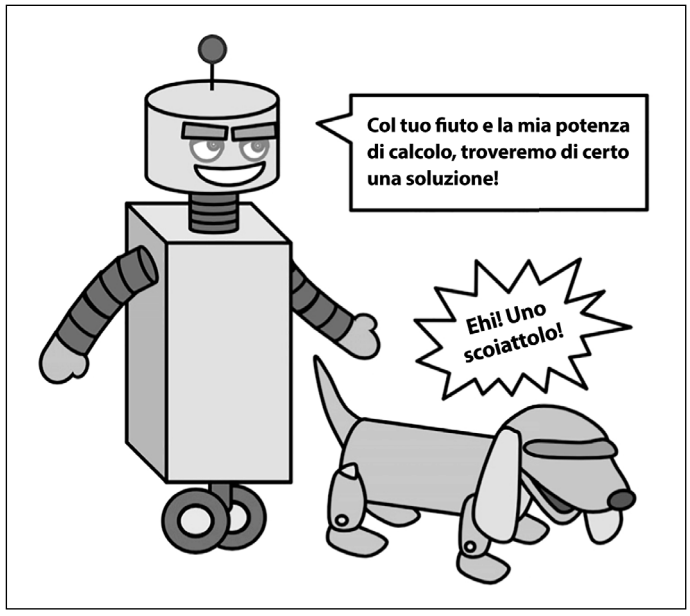


Figura 1.2 Gli algoritmi di machine learning sono strumenti potenti, ma richiedono di procedere in una direzione precisa.

Con questi obiettivi, il machine learning viene impiegato da innumerevoli aziende, laboratori, ospedali ed enti governativi. Ogni attività che generi o aggreghi i dati in questo modo impiega almeno un algoritmo di machine learning.
Anche se è impossibile elencare ogni singolo caso d'uso del machine learning, un'occhiata alle più recenti storie di successo identifica alcuni esempi particolarmente evidenti.

- Identificazione dei messaggi di spam nella posta elettronica.
- Segmentazione del comportamento dei clienti per predisporre messaggi pubblicitari mirati.
- Previsioni meteorologiche e su cambiamenti climatici a lungo termine.
- Riduzione delle transazioni fraudolente tramite carta di credito.
- Stime fattuali dei danni economici a seguito di trombe d'aria e altri disastri naturali.
- Predizione dei risultati delle elezioni.
- Sviluppo di algoritmi per il pilotaggio automatico di droni e automobili.
- Ottimizzazione dell'uso dell'energia nelle abitazioni e negli uffici.
- Proiezione delle aree in cui è più probabile si verifichino attività criminali.
- Scoperta di sequenze genetiche correlate a determinate malattie.

Entro la fine di questo libro, conoscerete gli algoritmi di machine learning comunemente impiegati per insegnare ai computer a svolgere questi compiti. Per il momento, dico solo che indipendentemente dal contesto, il processo di machine learning è lo stesso. Indipendentemente dal compito, un algoritmo prende dei dati e identifica al loro interno dei pattern, degli "schemi", i quali costituiranno la base di ulteriori azioni.

I limiti del machine learning

Sebbene il machine learning sia ampiamente utilizzato e offra incredibili potenzialità, è importante comprenderne i limiti. Il machine learning, attualmente, emula un sottoinsieme relativamente limitato delle capacità del cervello umano. Offre una scarsa flessibilità in termini di estrapolazione al di fuori di rigidi parametri e non conosce il buonsenso. Detto questo, occorre applicare un'estrema cautela nel considerare esattamente che cosa ha imparato un algoritmo, prima di lasciarlo libero di operare nel mondo reale.
Senza una vita di esperienze passate su cui contare, i computer sono limitati anche nella loro capacità di creare semplici inferenze sui successivi passi logici da intraprendere. Prendete, per esempio, i banner pubblicitari presenti in molti siti web. Questi vengono proposti all'utente in base a schemi appresi mediante tecniche di data mining sulla cronologia di navigazione di milioni di utenti. Sulla base di questi dati, qualcuno che navighi in siti web dove si vendono scarpe è interessato ad acquistare scarpe e pertanto dovrà vedere pubblicità di scarpe. Il problema è che questo diviene un ciclo senza fine, in cui, anche dopo l'acquisto delle scarpe, verranno presentate nuove pubblicità di scarpe, invece di pubblicità di lacci e lucidi.
Molti conoscono bene i limiti del machine learning in termini di capacità di comprendere o tradurre una lingua, o di riconoscere il parlato e la scrittura a mano. Forse il primo esempio di questo tipo di problema risale a un episodio del 1994 di "The Simpsons", che mostrava una parodia del tablet Apple Newton. In quel tempo, il Newton era riconosciuto come un ottimo sistema di riconoscimento della scrittura a mano. Sfortunatamente per Apple, occasionalmente falliva il suo scopo in modo esilarante. L'episodio illustra il problema attraverso una sequenza in cui la frase "Beat up Martin" (Picchia Martin) è stata interpretata dal Newton come "Eat up Martha" (Mangia Martha).

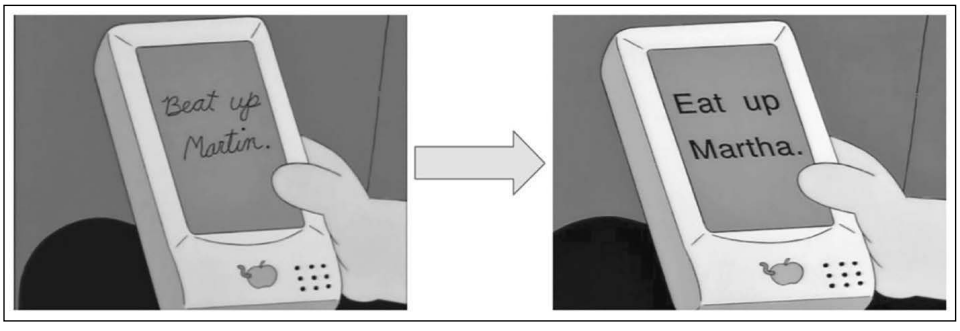


Figura 1.3 Due esilaranti frame tratti da Lisa on Ice, "The Simpsons", 20th Century Fox, 1994.

L'elaborazione automatica del linguaggio umano è abbastanza migliorata dai tempi dell'Apple Newton, tanto che Google, Apple e Microsoft sono piuttosto sicure della loro capacità di offrire servizi ad attivazione vocale di assistenza, come Google Assistant, Siri e Cortana. Tuttavia, questi servizi piuttosto spesso faticano a rispondere a domande relativamente semplici. Inoltre, i servizi di traduzione online talvolta interpretano erroneamente frasi che anche un bambino piccolo comprenderebbe immediatamente, e la funzionalità di predizione del testo presente in molti dispositivi ha portato allo sviluppo di numerosi siti umoristici di "errori di autocorrezione" che illustrano la capacità dei computer di comprendere semplici parole ma sbagliando clamorosamente il contesto.

Alcuni di questi errori certamente sono prevedibili. La lingua umana è complicata, dotata com'è di più livelli di testo e sottintesi, e perfino gli esseri umani talvolta sbagliano il contesto. Nonostante il fatto che il machine learning sta migliorando rapidamente nell'elaborazione del linguaggio umano, le sue continue carenze illustrano un fatto importante: che la qualità del machine learning dipende dalla qualità dei dati da cui apprende. Se il contesto non è esplicito nei dati di input, allora proprio come un essere umano, il computer farà del suo meglio per tentare di indovinare il significato della frase in base ai limiti delle sue esperienze passate.

L'etica del machine learning

Fondamentalmente, il machine learning è uno strumento che ci assiste nel cercare di trarre un significato dalla complessità dei dati del mondo. Come qualsiasi altro strumento, può essere impiegato in modo buono o cattivo. I casi in cui il machine learning sbaglia di più è quando è applicato in modo così vago, o così insensato, da trattare gli esseri umani come cavie da laboratorio, automi o consumatori decerebrati. Un processo apparentemente innocuo può condurre a conseguenze imprevedibili, quando vengono automatizzate da un computer privo di emozioni. Per questo motivo, coloro che impiegano il machine learning o il data mining sarebbero negligenti se non considerassero le implicazioni etiche del loro lavoro.
A causa della relativa "infanzia" della disciplina del machine learning e della velocità con la quale sta progredendo, gli aspetti legali e sociali relativi al suo operato sono un po' vaghi e costantemente in divenire. Occorre applicare una certa cautela nel raccogliere o analizzare i dati con lo scopo di evitare di infrangere le leggi, di violare i termini del servizio o gli accordi sull'utilizzo dei dati o di abusare della fiducia o di violare la privacy dei clienti o del pubblico.

> **NOTA**
>
> Il motto aziendale informale di Google, un'azienda che forse più di ogni altra raccoglie dati sugli individui, era: "Don't be evil" (Non essere cattivo). Un messaggio piuttosto chiaro, ma anche insufficiente. Un approccio migliore potrebbe consistere nel seguire il giuramento di Ippocrate, un principio medico che stabilisce che *Primum non nocere*: "Per prima cosa, non nuocere".

I rivenditori impiegano comunemente il machine learning per scopi pubblicitari, per promozioni mirate, per la gestione degli inventari o per la disposizione degli articoli sugli scaffali. Molti supermercati hanno dotato le loro casse di apparecchi che stampano coupon per promozioni basate sulla cronologia di acquisti del cliente. In cambio di una manciata di dati personali, il cliente riceve degli sconti sugli specifici prodotti che desidera acquistare. A prima vista, questo può apparire relativamente innocuo, ma considerate che cosa accade spingendo troppo avanti questa pratica.
Un racconto, probabilmente apocrifo, riguarda un grosso rivenditore degli Stati Uniti, che ha impiegato il machine learning per identificare le future madri per l'invio di appositi coupon. Il rivenditore sperava che se queste future madri avessero ricevuto sconti interessanti, sarebbero poi diventate clienti fedeli, che poi avrebbero acquistato grandi quantità di pannolini, accessori per bambini e giocattoli.

Armato dei suoi metodi di machine learning, il rivenditore ha identificato gli articoli presenti nella cronologia degli acquisti che fossero in grado di far prevedere, con un elevato grado di sicurezza, non solo che una donna era in attesa, ma anche la data approssimativa della futura nascita.
Dopo che il rivenditore ha impiegato questi dati per una mailing list promozionale, un tizio, arrabbiato, ha risposto chiedendo spiegazioni: come mai sua figlia aveva ricevuto dei coupon per articoli di maternità? Era furioso, perché il rivenditore sembrava incoraggiare gravidanze da teenager! La storia prosegue, con la catena di negozi che offre le proprie scuse, seguite poi dalle scuse del padre, che, dopo aver parlato con sua figlia ha scoperto che, in effetti, era in dolce attesa!
Che la storia sia del tutto vera o inventata, la morale è che occorre applicare del sano buon senso prima di applicare indiscriminatamente i risultati di un'analisi effettuata facendo uso del machine learning. Questo, in particolare, quando il problema coinvolge informazioni riservate, come dati sulla salute. Con un po' di cautela in più, il rivenditore avrebbe potuto prevedere questa situazione e avrebbe potuto applicare una maggiore discrezione nel decidere come rivelare gli schemi che la sua analisi di machine learning aveva individuato.

RIFERIMENTO

Per ulteriori informazioni sul modo in cui i rivenditori impiegano il machine learning per identificare le gravidanze, consultate l'articolo del "New York Times Magazine" *How Companies Learn Your Secrets*, di Charles Duhigg, 2012: `https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html`.

Man mano che gli algoritmi di machine learning vengono applicati sempre più, scopriamo che i computer possono apprendere alcuni comportamenti sconsiderati delle società umane. Purtroppo, tutto ciò comprende il fatto di perpetuare le discriminazioni di "razza" o genere e il rafforzamento degli stereotipi negativi. Per esempio, i ricercatori hanno scoperto che il servizio di pubblicità online di Google presenta con maggiore probabilità offerte di lavoro ad alto reddito agli uomini che alle donne ed è più propenso a mostrare messaggi sulla valutazione del background criminale alle persone di colore.
Per dimostrare che questi tipi di passi falsi non si limitano alla Silicon Valley, un servizio chatbot per Twitter sviluppato da Microsoft è stato prontamente messo offline dopo che aveva iniziato a diffondere propaganda filonazista e antifemminista. Spesso, degli algoritmi che a prima vista sembrano "*content neutral*" iniziano ben presto a riflettere le opinioni della maggioranza o le ideologie dominanti. Un algoritmo creato da Beauty.AI per riflettere un canone oggettivo di bellezza umana ha dato origine a controversie, in quanto favoriva quasi esclusivamente le opinioni dei "bianchi". Immaginate le conseguenze se un concetto di questo tipo fosse stato applicato a un software di riconoscimento facciale per la prevenzione di attività criminali!

RIFERIMENTO

Per ulteriori informazioni sulle conseguenze del machine learning nel mondo reale e i problemi di discriminazione, consultate l'articolo del "New York Times" *When Algorithms Discriminate*, di Claire Cain Miller, 2015: `https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html`.

Per limitare la capacità degli algoritmi di discriminare in modo illegale, alcune giurisdizioni adottano apposite leggi che prevengono l'utilizzo di dati di razza, etnia, religione o di altro tipo per scopi commerciali. Tuttavia, il fatto di escludere questi dati da un progetto potrebbe non essere sufficiente, perché gli algoritmi di machine learning possono comunque imparare inavvertitamente a discriminare. Se un determinato segmento della popolazione tende a vivere in una determinata regione, acquista un determinato prodotto o comunque si comporta in un modo che lo identifica univocamente come un gruppo, gli algoritmi di machine learning possono inferire le informazioni protette a partire da altri fattori. In tali casi, potreste dover de-identificare completamente queste persone, escludendo ogni possibile informazione identificativa, e questo al di là delle altre informazioni già protette.

A parte le conseguenze legali, un utilizzo inappropriato dei dati può essere dannoso a prescindere da tutto. I clienti possono non sentirsi a loro agio o addirittura spaventarsi vedendo resi pubblici certi aspetti della loro vita che considerano privati. Negli ultimi anni, un certo numero di applicazioni web di alto profilo ha sperimentato un esodo di massa di utenti che si sono sentiti sfruttati, quando i termini di servizio dell'applicazione sono stati cambiati o i loro dati sono stati utilizzati per finalità che andavano oltre quello che gli utenti avevano concordato originariamente. Il fatto che le aspettative sulla privacy differiscano per contesto, età e cultura complica ulteriormente la decisione su quale possa essere un "uso appropriato" dei dati personali. Sarebbe saggio considerare le implicazioni culturali del proprio lavoro prima di intraprendere un nuovo progetto, oltre a considerare le sempre più restrittive leggi promulgate dall'Unione Europea: il *General Data Protection Regulation* (GDPR) e le inevitabili politiche che ne discendono.

NOTA

Il fatto che *possiate* utilizzare dei dati per un determinato fine non sempre implica che poi dovreste *effettivamente* utilizzare tali dati.

Infine, è importante notare che mentre gli algoritmi di machine learning divengono sempre più importanti nella nostra vita quotidiana, crescono anche gli incentivi a un loro utilizzo e sfruttamento da parte di soggetti pericolosi. Talvolta, gli hacker "sportivi" vogliono semplicemente mettere a soqquadro gli algoritmi per farsi quattro risate e per avere un attimo di notorietà, come nel caso del "Google bombing" il metodo *crowd-source* per ingannare gli algoritmi di Google e spingerli a valutare in modo particolarmente elevato la pagina desiderata.

Ma altre volte, gli effetti sono ben più gravi. Un esempio recente riguarda l'ondata di cosiddette *fake-news* e di influenze elettorali indebite, propagate tramite la manipolazione degli algoritmi pubblicitari ed elettorali che prendevano di mira le persone in base alla loro personalità online. Per evitare di fornire questo tipo di controllo a estranei, nella creazione di sistemi di machine learning, è fondamentale considerare come questi possano essere influenzati da un determinato individuo o gruppo di potere.

RIFERIMENTO

Lo studioso di *social media* danah boyd (scritto in lettere minuscole) ha presentato un discorso alla *Strata Data Conference 2017* a New York che trattava l'importanza del rafforzamento degli algoritmi di machine learning contro gli attacchi dall'esterno. Per un riassunto, consultate: `https://points.datasociety.net/your-datais-being-manipulated-a7e31a83577b`.

Le conseguenze degli attacchi fraudolenti agli algoritmi di machine learning possono addirittura essere fatali. I ricercatori hanno dimostrato che creando un *adversarial attack* che distorce in modo subdolo un segnale stradale con graffiti attentamente configurati, un hacker può indurre un veicolo a guida autonoma a interpretare erroneamente un segnale di Stop, causando potenzialmente un incidente. Ma anche in assenza di intenti malvagi, i bug e gli errori umani hanno già condotto a incidenti fatali, che hanno coinvolto veicoli a guida autonoma Uber e Tesla. Dati questi esempi, è di primaria importanza ed è un obiettivo etico fondamentale che gli addetti al machine learning si preoccupino del fatto che i loro algoritmi siano ben utilizzati e non abusati nel mondo reale.

Come apprendono le macchine

Una definizione formale del machine learning, attribuita allo scienziato dei computer Tom M. Mitchell, afferma che una macchina apprende ogni volta che è in grado di utilizzare la propria esperienza in modo da migliorare le proprie prestazioni su esperienze simili, in futuro. Sebbene questa definizione sia intuitiva, essa ignora completamente il processo mediante il quale l'esperienza può essere tradotta in un'azione futura e, come sappiamo, fra il "dire" (o anche l'"imparare") e il "fare" c'è di mezzo il mare!
Mentre il cervello umano è naturalmente in grado di apprendere fin dalla nascita, le condizioni tali per cui i computer possano apprendere devono essere rese esplicite. Per questo motivo, sebbene esso non sia strettamente necessario per comprendere le basi teoriche dell'apprendimento, questo concetto di fondo ci aiuta a comprendere, distinguere e implementare gli algoritmi di machine learning.

> **NOTA**
>
> Nel confrontare il machine learning con l'apprendimento umano, potreste trovarvi a esaminare la vostra mente sotto una luce differente.

Indipendentemente dal fatto che il soggetto che apprende sia un essere umano o una macchina, il processo di base dell'apprendimento è simile e può essere suddiviso in quattro elementi correlati.

- *Archivio dati*: utilizza l'osservazione, la memoria e i ricordi per fornire una base fattuale per i successivi ragionamenti.
- *Astrazione*: prevede la traduzione dei dati memorizzati in rappresentazioni e concetti più ampi.
- *Generalizzazione*: impiega i dati astratti per creare conoscenza e inferenze che governano l'azione in nuovi contesti.
- *Valutazione*: fornisce un meccanismo di feedback per misurare l'utilità della conoscenza appresa e informare i potenziali miglioramenti.

Sebbene qui il processo di apprendimento sia stato concettualizzato sotto forma di quattro componenti distinti, lo scopo di questa suddivisione è meramente illustrativo. In realtà, l'intero processo di apprendimento è interconnesso in modo inestricabile. Negli esseri umani, questo processo è del tutto inconscio.

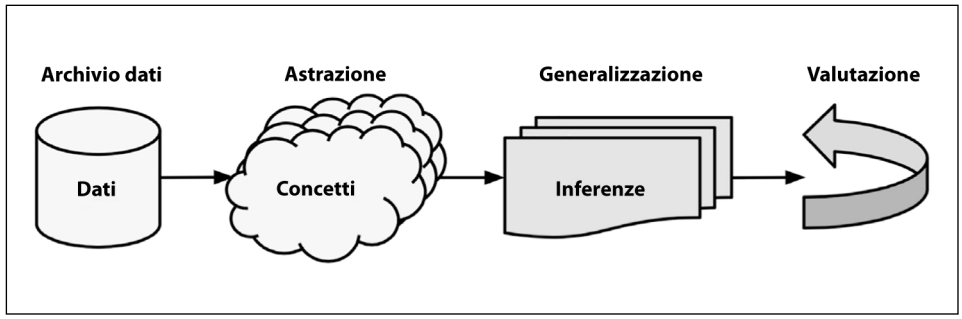


Figura 1.4 Il processo di apprendimento.

Raccogliamo, deduciamo, eseguiamo deduzioni e intuiamo ciò che accade entro i confini del nostro sguardo, e poiché questo processo è nascosto, ogni differenza da persona a persona è attribuita a un vago concetto di soggettività. Al contrario, i computer rendono questi processi espliciti, e poiché l'intero processo è trasparente, la conoscenza appresa può essere esaminata, trasferita, utilizzata per azioni future e trattata come una "scienza" dei dati: *data science*.

Il termine *data science* suggerisce il fatto che esista una relazione fra i dati, la macchina e coloro che guidano il processo di apprendimento. L'utilizzo sempre crescente del termine nelle offerte di lavoro e nei programmi di studi riflette la sua importanza come un campo di studi legato alla teoria statistica e computazionale, così come pure all'infrastruttura tecnologica che rende possibile il machine learning e le sue applicazioni. Il campo spesso chiede ai suoi addetti di essere ottimi raccontatori di storie, equilibrando l'audacia nell'uso dei dati con i limiti di ciò che può essere inferito e previsto a partire dai dati. Per essere ottimi "scienziati dei dati", pertanto, occorre un'ottima conoscenza del funzionamento degli algoritmi di apprendimento.

Archivio dati

Ogni apprendimento parte dai dati. Gli esseri umani e anche i computer utilizzano un archivio dati quale base per ogni ragionamento avanzato. In un essere umano, esso consiste di un cervello il quale impiega segnali elettrochimici che si muovono in una rete di cellule per conservare ed elaborare le osservazioni per poi richiamarle in un futuro a breve e lungo termine. I computer offrono funzionalità simili di memorizzazione a breve e lungo termine impiegando però hard disk, memorie flash e memoria RAM, sistemi che operano in combinazione con un'unità centrale di elaborazione, la CPU.

Può sembrare ovvio, ma la capacità di memorizzare e poi recuperare i dati, da sola è insufficiente per permettere l'apprendimento. I dati memorizzati sono semplicemente tanti 0 e 1 su un disco. Si tratta di una raccolta di dati memorizzati, senza alcun significato, in assenza di un contesto più ampio. Senza un livello di comprensione superiore, la conoscenza è puro ricordo, limitata a quanto è stato osservato e senza nessuna implicazione per il futuro.

Per comprendere meglio le sfumature di questa idea, può essere utile riflettere sull'ultima volta che vi siete trovati a studiare per un test difficile, magari un esame universitario o una certificazione per la carriera professionale. Amereste avere una memoria *eidetica*

(fotografica)? In tal caso, purtroppo occorre dire che essa non vi sarebbe di grande aiuto. Anche se riusciste a memorizzare perfettamente il materiale, questo apprendimento meccanico non offrirebbe particolari vantaggi senza conoscere esattamente le domande e le risposte somministrate in sede di esame. Altrimenti, sareste costretti a memorizzare le risposte di ogni domanda concepibile, su un argomento che potenzialmente potrebbe dare origine a un numero infinito di domande. Ovviamente, questa è una strategia insostenibile.
Al contrario, un approccio migliore consisterebbe nel dedicare del tempo a uno studio selettivo, memorizzando un insieme relativamente limitato di idee rappresentative, sviluppando però una reale comprensione delle relazioni fra le idee e applicandole poi a circostanze impreviste. In questo modo, riuscirete a identificare gli schemi generali più importanti, invece di ridurvi a memorizzare ogni singolo dettaglio, sfumatura e potenziale applicazione.

Astrazione

Il processo di *astrazione* funziona assegnando un significato più ampio ai dati memorizzati: i dati grezzi finiscono per rappresentare un'idea o un concetto più ampio, astratto. Questo tipo di connessione, supponiamo fra un oggetto e la sua rappresentazione, è esemplificata dal famoso dipinto di René Magritte *La Trahison des images* (Figura 1.5).

Figura 1.5 "Questa non è una pipa." Fonte: http://collections.lacma.org/node/239578.

Il dipinto rappresenta una pipa con la didascalia *Ceci n'est pas une pipe* ("Questa non è una pipa"). Ciò che Magritte sta dicendo è che la rappresentazione di una pipa non è una pipa: il dipinto non si può caricare di tabacco e fumare. Ma, nonostante il fatto che la pipa non sia reale, chiunque osservi il dipinto lo riconoscerà come una pipa. Questo suggerisce che gli osservatori sono certamente in grado di connettere l'immagine della pipa all'idea di una pipa, a un ricordo di un oggetto pipa che possono aver tenuto in mano. Le connessioni astratte come questa sono alla base della *rappresentazione della conoscenza*, la formazione delle strutture logiche che assistono nel trasformare delle pure informazioni sensoriali in conoscenze significative.
Nel corso di un processo artificiale di rappresentazione della conoscenza, il computer riassume in un *modello* i dati grezzi memorizzati, come il modello è una descrizione esplicita degli schemi presenti nei dati. Proprio come la pipa di Magritte, la rappresentazione del modello trae vita dai dati grezzi. Esso rappresenta un'idea più grande della somma delle sue parti.

Vi sono molti tipi differenti di modelli. Forse già ne conoscete alcuni. Ecco alcuni esempi:

- equazioni matematiche;
- diagrammi relazionali, come alberi e grafi;
- regole logiche `if/else`;
- raggruppamenti di dati noti come cluster.

La scelta del modello in genere non viene lasciata alla macchina. Al contrario, il compito di apprendimento e il tipo dei dati disponibili informano la scelta del modello. Più avanti in questo stesso capitolo, discuteremo più in dettaglio i metodi utilizzabili per scegliere il tipo di modello appropriato.

Il processo di configurazione di un modello a un dataset è detto *training*, addestramento. Dopo che il modello è stato addestrato, i dati saranno stati trasformati in una forma astratta che condensa l'informazione originale.

NOTA

Potreste chiedervi perché questo passo sia chiamato "training", addestramento, anziché "learning", apprendimento. Innanzitutto, notate che il processo di apprendimento non termina con un'astrazione dei dati: chi apprende deve ancora generalizzare e valutare il suo addestramento. In secondo luogo, la parola "training" (e anche la parola "addestramento) connota meglio il fatto che l'essere umano insegnante addestra la macchina studente a comprendere i dati in un modo ben preciso.

È importante notare che un modello appreso non fornisce nuovi dati, eppure produce nuova conoscenza. Come può accadere? La risposta è che imponendo una struttura preconcetta ai dati sottostanti otteniamo una conoscenza precedentemente invisibile. Presuppone l'esistenza di un nuovo concetto che descrive un modo in cui relazionare fra loro elementi presenti nei dati.

Prendete, per esempio, la scoperta della gravità. Adattando delle equazioni ai dati osservati, Sir Isaac Newton ha inferito il concetto di gravità, ma la forza che noi chiamiamo gravità è sempre stata presente. Semplicemente non è stata riconosciuta fino a quando Newton non l'ha espressa come un concetto astratto che mette in relazione alcuni dati con altri dati: in particolare, diventando il termine g in un modello che spiega le osservazioni relative alla caduta degli oggetti (Figura 1.6).

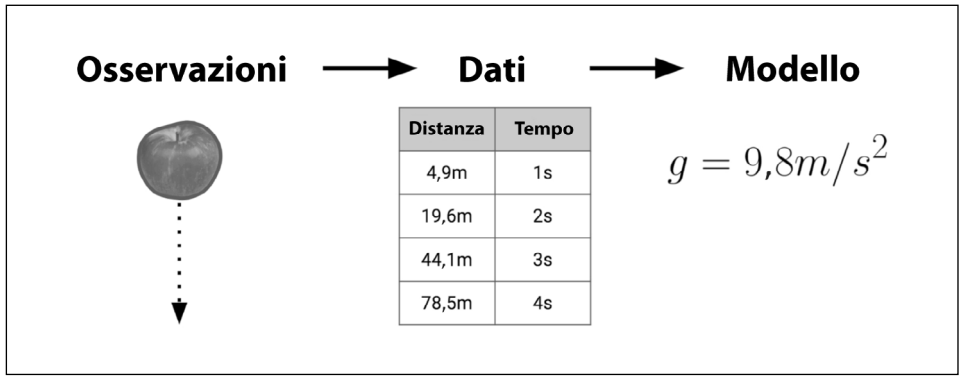


Figura 1.6 I modelli sono astrazioni che spiegano i dati osservati.

La maggior parte dei modelli non porterà allo sviluppo di teorie che scuoteranno il mondo scientifico nei secoli a venire. Tuttavia, la vostra astrazione potrebbe portare alla scoperta di importanti e precedentemente invisibili, schemi e relazioni fra i dati. Un modello addestrato su dati dei genomi potrebbe trovare determinati geni che, quando combinati fra loro, sono responsabili dell'insorgere del diabete, le banche potrebbero scoprire un tipo apparentemente innocuo di transazione che, sistematicamente, compare appena prima di un'attività fraudolenta o gli psicologi potrebbero identificare una combinazione di tratti caratteriali che indicano un nuovo disturbo. Questi schemi nascosti sono sempre stati presenti ma presentando le informazioni in un nuovo formato, possono concettualizzare una nuova idea.

Generalizzazione

Il passo successivo nel processo di apprendimento consiste nell'utilizzare la conoscenza astratta per un'azione futura. Tuttavia, fra gli innumerevoli schemi nascosti che possono essere identificati durante il processo di astrazione e le miriadi di modi per modellare tali schemi, alcuni schemi potrebbero rivelarsi più utili di altri. A meno che la produzione di astrazioni non venga limitata a un insieme utile, potremmo ritrovarci là dove siamo partiti, con una grande quantità di informazioni ma nessuna conoscenza utilizzabile.
Formalmente, il termine *generalizzazione* è definito come il processo di trasformazione della conoscenza astratta in una forma che possa essere utilizzata per un'azione futura, applicata a compiti simili ma non identici a quelli esaminati in precedenza. Essa opera come una ricerca sull'intero set di modelli (ovvero le teorie o inferenze) che *può* essere definito a partire dai dati durante l'addestramento.
Se potete immaginare un ipotetico insieme contenente ogni possibile modo in cui i dati possono essere astratti, la generalizzazione prevede la riduzione di questo insieme in un insieme più piccolo e gestibile di scoperte importanti.
Nella generalizzazione, il sistema di apprendimento è incaricato di ridurre gli schemi che ha scoperto ai soli rilevanti per i suoi compiti futuri. Normalmente, non è possibile ridurre il numero di schemi, i "pattern", esaminandoli uno per uno e valutandoli in base alla loro futura utilità. Al contrario, generalmente gli algoritmi di machine learning impiegano dei sistemi che riducono più rapidamente lo spazio di ricerca. Per questo scopo l'algoritmo impiegherà dei metodi *euristici*, che sono supposizioni informate riguardanti dove è possibile trovare le inferenze più utili.

> **NOTA**
>
> L'euristica utilizza approssimazioni e altre regole empiriche, il che significa che essa non ha alcuna garanzia di trovare il modello migliore per i dati. Tuttavia, senza adottare queste tecniche, sarebbe impossibile trovare informazioni utili in un dataset di grandi dimensioni.

L'euristica viene comunemente utilizzata dagli esseri umani per generalizzare l'esperienza nei confronti di nuovi scenari. Se vi è mai capitato di utilizzare l'istinto per prendere una decisione rapida prima di valutare appieno le circostanze, avete fatto ricorso esattamente alla vostra euristica mentale.
L'incredibile capacità umana di prendere decisioni rapidamente, spesso non conta su una "logica di calcolo", ma piuttosto su "euristiche emozionali". Talvolta, ciò può condurre a conclusioni illogiche. Per esempio, molte persone esprimono i propri timori nel compiere

un viaggio in aereo, piuttosto che in automobile, ma statisticamente le automobili sono più pericolose degli aerei. Questo comportamento può essere spiegato con l'euristica della disponibilità, che è la tendenza delle persone a stimare la probabilità di un evento con la facilità con cui è possibile richiamare alla mente degli esempi. Gli incidenti aerei hanno una grande risonanza nei media. Trattandosi di eventi traumatici, non è facile dimenticarsene, mentre gli incidenti automobilistici solo raramente raggiungono i notiziari. Ma la follia dell'euristica inapplicata non si limita agli esseri umani. Anche l'euristica impiegata dagli algoritmi di machine learning talvolta porta a conclusioni errate. Se l'algoritmo sbaglia in modo *sistematico*, si dice che ha un *bias*, parola difficilmente traducibile in italiano, ma che reca in sé i significati di "scarto", inteso come "differenza" e "pregiudizio", volendo attribuire all'algoritmo caratteristiche più umane. Questo implica che l'algoritmo sbagli in un modo coerente o prevedibile.
Per esempio, supponete che un algoritmo di machine learning abbia imparato a identificare dei volti trovando due cerchi, che rappresentano gli occhi, posizionati sopra una linea orizzontale, che rappresenta la bocca. L'algoritmo potrebbe incontrare problemi, essere soggetto a un *bias negativo*, se esposto a volti non conformi al suo modello. Volti con occhiali, presi di profilo, angolati o con certe tonalità della pelle potrebbero non essere interpretati come volti dall'algoritmo. Analogamente, potrebbe benissimo essere soggetto a un *bias positivo* nei confronti di certe tonalità di pelle, forme di volti o altre caratteristiche che meglio si adattano alla sua conoscenza del mondo (Figura 1.7).

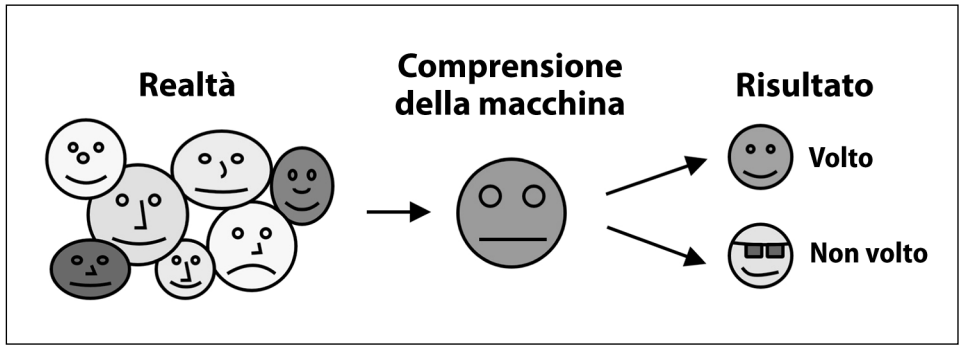


Figura 1.7 Il processo di generalizzazione dell'esperienza può condurre a un bias.

Se consideriamo la traduzione "pregiudizio", la parola "bias" convoglia soprattutto connotazioni piuttosto negative. Si considera una qualità, per una testata giornalistica il fatto di essere esente da pregiudizi, perché fornirà le notizie in modo obiettivo, slegato dall'emotività. Tuttavia, considerate per un momento la possibilità che un certo bias possa essere utile. Senza un po' di arbitrarietà, potrebbe essere un po' difficile decidere fra più scelte, in competizione fra loro, ognuna delle quali ha specifici punti di forza e di debolezza. In effetti, vari studi nel campo della psicologia suggeriscono che coloro che sono nati con un danno in quella porzione del cervello responsabile delle emozioni sono in realtà inaffidabili in termini di capacità di prendere decisioni e potrebbero dedicare ore a elucubrare su decisioni anche banali, come quale maglietta indossare o che cosa mangiare a pranzo. Paradossalmente, il bias da un lato ci rende "ciechi" nei confronti di alcune informazioni, ma dall'altro ci consente di utilizzare le informazioni rimanenti

per intraprendere un'azione. Questo è il modo in cui gli algoritmi di machine learning scelgono uno fra gli innumerevoli modi di comprendere un insieme di dati.

Valutazione

Il bias, dunque, è un male necessario associato ai processi di astrazione e generalizzazione che sono inerenti a ogni compito di apprendimento. Con lo scopo di guidare l'azione nei confronti di un numero infinito di possibilità, ogni apprendimento deve avere un bias. Di conseguenza, ogni strategia di apprendimento ha dei punti deboli; non esiste un unico algoritmo di apprendimento superiore a tutti gli altri. Pertanto, il passo finale nel processo di apprendimento consiste nel valutare il suo successo e misurare le prestazioni dell'apprendimento nonostante i suoi bias. Le informazioni acquisite nella fase di valutazione possono così essere utilizzate per informare un ulteriore, eventuale, addestramento.

RIFERIMENTO

Quando avrete applicato con successo una tecnica di machine learning, sarete tentati di applicarla a ogni compito. È importante resistere a questa tentazione, poiché nessun approccio al machine learning si può dire *migliore* in qualsiasi circostanza. Questo fatto è descritto dal cosiddetto teorema *No Free Lunch* (non esiste un pasto gratis), introdotto da David Wolpert nel 1996. Per ulteriori informazioni: `http://www.no-free-lunch.org`.

In generale, la valutazione si verifica dopo che un modello è stato addestrato su un dataset di training iniziale. Poi, il modello viene valutato su un dataset di test distinto, con lo scopo di giudicare con quale efficacia la sua caratterizzazione dei dati di addestramento si generalizza su nuovi dati, casi differenti e mai esaminati prima. Vale la pena di notare che è estremamente raro che un modello sia perfettamente generalizzabile a ogni caso possibile: gli errori sono quasi sempre inevitabili.
In parte, i modelli falliscono nella loro capacità di generalizzare perfettamente a causa del problema del rumore, un termine che descrive delle variazioni inspiegabili nei dati. La presenza di dati rumorosi è causata da eventi apparentemente casuali, come i seguenti.

- Errori di misurazione dovuti a sensori non precisi, che talvolta sommano o sottraggono una piccola quantità alle loro letture.
- Problemi legati agli individui, come quando una persona sottoposta a un sondaggio fornisce una risposta a casaccio con il solo scopo di terminare più rapidamente.
- Problemi di qualità dei dati, ovvero dati mancanti, nulli, troncati, scritti in modo errato o danneggiati.
- Fenomeni così complessi o così poco noti che hanno un impatto apparentemente casuale sui dati.

Il tentativo di modellare il rumore è alla base di un problema chiamato *overfitting*; poiché la maggior parte dei dati rumorosi ha per definizione una natura inspiegabile, il tentativo di spiegare il rumore produrrà modelli non generalizzabili appieno a nuovi casi. I tentativi di spiegare il rumore producono inoltre modelli più complessi, che mancano il bersaglio (Figura 1.8) rappresentato dagli schemi effettivamente presenti nei dati che stiamo tentando di identificare, inseguendo i dati rumorosi.

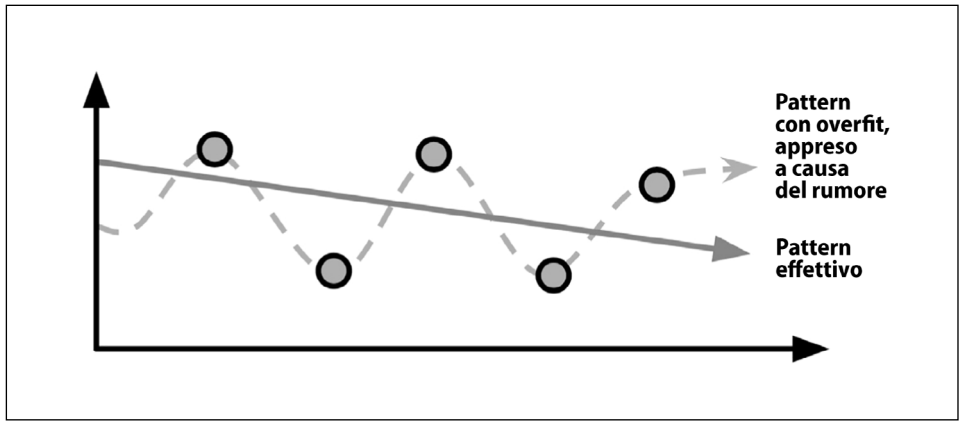


Figura 1.8 La modellazione del rumore produce generalmente modelli più complessi, che mancano il bersaglio degli schemi nascosti.

Un modello che si comporta relativamente bene durante l'addestramento ma relativamente male durante la valutazione si dice che è andato in *overfitting* sul dataset di addestramento, poiché non generalizza bene sul dataset di test. Si è troppo adattato alle specifiche caratteristiche del dataset di addestramento, compresi gli errori (il rumore) presenti in esso. In termini pratici, questo significa che ha identificato nei dati uno schema che non è utile per guidare le azioni future; il processo di generalizzazione è fallito. Le soluzioni del problema dell'overfitting sono specifiche del particolare approccio al machine learning. Per il momento, il concetto che è importante memorizzare è che occorre tenere conto di questo problema. La qualità con la quale i metodi sono in grado di gestire dei dati rumorosi evitando l'overfitting è un elemento distintivo dei vari metodi.

Il machine learning in pratica

Finora, ci siamo concentrati sul funzionamento teorico del machine learning. Per applicare il processo di apprendimento a compiti legati al mondo reale, impiegheremo un processo costituito da cinque passi. Indipendentemente dal compito, ogni algoritmo di machine learning può essere applicato seguendo questi passi.

1. *Raccolta dei dati*: il passo di raccolta dei dati prevede l'acquisizione del materiale di apprendimento che l'algoritmo impiegherà per generare conoscenza pronta all'uso. Nella maggior parte dei casi, i dati avranno bisogno di essere combinati in un'unica origine, come un file di testo, un foglio di lavoro o un database.
2. *Esplorazione e preparazione dei dati*: la qualità di ogni progetto di machine learning si basa in larga misura sulla qualità dei suoi dati di input. Pertanto, è importante conoscere bene i dati e le loro sfumature nel corso di una fase di esplorazione dei dati. È poi necessario ulteriore lavoro per preparare i dati per il processo di apprendimento. Questo prevede la correzione o pulizia dei dati erronei, per eliminare i dati non utili, e la ricodifica dei dati, per renderli conformi agli input dal sistema di apprendimento.
3. *Addestramento del modello*: nel momento in cui i dati sono preparati per la fase di analisi, probabilmente vi sarete fatti un'idea di ciò che potrete apprendere dai dati.

Lo specifico compito di machine learning scelto informerà la scelta di un algoritmo appropriato, e l'algoritmo rappresenterà i dati sotto forma di un modello.

4. *Valutazione del modello*: ogni modello di machine learning fornisce una soluzione con bias al problema di apprendimento, il che significa che è importante valutare la qualità dell'apprendimento dell'algoritmo in base alla sua esperienza. A seconda del tipo di modello impiegato, potreste essere in grado di valutare l'accuratezza del modello impiegando un dataset di test o potreste aver bisogno di sviluppare misure prestazionali specifiche per l'applicazione.
5. *Miglioramento del modello*: se è necessario ottenere prestazioni migliori, diviene necessario utilizzare strategie più avanzate per elevare le prestazioni del modello. Talvolta può essere necessario scegliere proprio un altro tipo di modello. Potreste dover affiancare ai vostri dati ulteriori dati o svolgere altro lavoro preparatorio, come previsto nel passo 2 di questo processo.

Dopo aver completato questi passi, se il modello si comporta bene, può essere applicato al compito previsto. A seconda della situazione, potreste utilizzare il vostro modello per fornire punteggi per predizioni (possibilmente in tempo reale); per proiezioni su dati finanziari; per generare conoscenze utili per scopi di marketing o di ricerca; per automatizzare dei compiti, come la consegna della corrispondenza o la gestione dei voli. I successi e fallimenti del modello sviluppato possono anche fornire ulteriori dati per addestrare la successiva generazione del sistema di apprendimento.

Tipi di dati di input

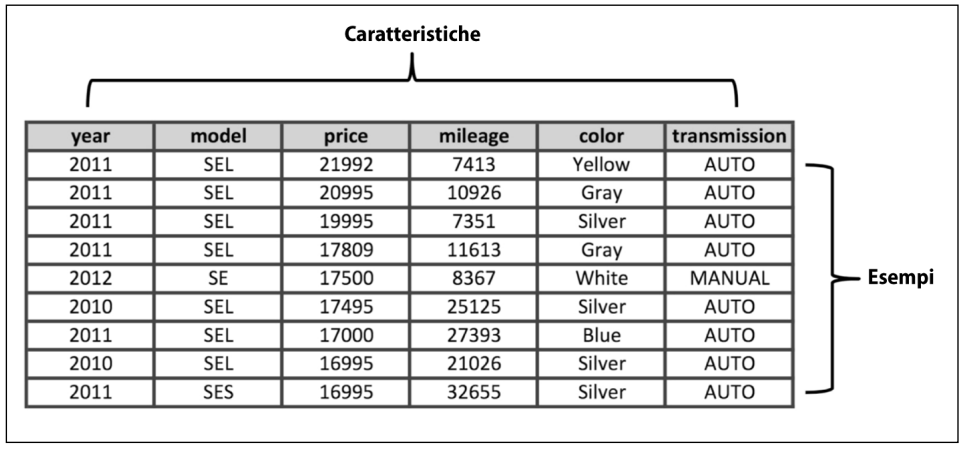
La pratica del machine learning prevede la necessità di far corrispondere alle caratteristiche dei dati di input gli inevitabili bias degli algoritmi di apprendimento disponibili. Pertanto, prima di applicare il machine learning ai problemi del mondo reale, è importante comprendere la terminologia utilizzata per distinguere i dataset di input.
Il termine *unità di osservazione* viene utilizzato per descrivere la più piccola entità dotata di proprietà misurate che siano di interesse per uno studio. Di solito, l'unità di osservazione assume la forma di persone, oggetti o cose, transazioni, punti temporali, regioni geografiche o misurazioni. Talvolta, le unità di osservazione vengono combinate in singole unità, come gli anni-uomo, per denotare i casi in cui la stessa persona viene monitorata nel corso di più anni, e ogni anno-uomo contiene i dati di una persona per un anno.

> **NOTA**
>
> L'unità di osservazione è correlata, ma non è la stessa cosa di, all'unità di analisi, che è la più piccola unità dalla quale può essere tratta un'inferenza. Sebbene spesso esse coincidano, le unità osservate e analizzate non sempre sono la stessa cosa. Per esempio, i dati osservati dalle persone (le unità di osservazione) possono essere utilizzati per analizzare le tendenze nei vari paesi (le unità di analisi).

I dataset che conservano le unità di osservazione e le loro proprietà possono essere descritti come raccolte dei seguenti elementi.

- *Esempi*: istanze delle unità di osservazione per le quali sono state registrate le proprietà.
- *Caratteristiche*: proprietà o attributi registrati degli esempi e che possono essere utili per l'apprendimento.

È più facile comprendere le caratteristiche e gli esempi a partire da una situazione reale. Per esempio, per costruire un algoritmo di apprendimento in grado di identificare i messaggi di posta elettronica spam, le unità di osservazione sarebbero i messaggi di posta elettronica, gli esempi sarebbero i singoli messaggi e le caratteristiche sarebbero le parole presenti nei messaggi.
Per un algoritmo per l'analisi del cancro, le unità di osservazione sarebbero i pazienti, gli esempi comprenderebbero un campione casuale di pazienti e le caratteristiche potrebbero essere i marcatori genomici delle cellule sottoposte a biopsia, oltre alle caratteristiche dei pazienti, come il peso, l'altezza o la pressione sanguigna.
Le persone e le macchine differiscono quanto a complessità dei dati di input che sono abituate a manipolare. Gli esseri umani sono abilissimi a sfruttare dati non strutturati, come testi, immagini e suoni. Inoltre sono flessibili e in grado di gestire casi in cui alcune osservazioni sono ricche di caratteristiche, mentre altre ne hanno molto poche.
Al contrario, i computer generalmente richiedono l'uso di dati strutturati, il che significa che ciascun esempio del fenomeno deve avere le stesse caratteristiche, e queste caratteristiche devono essere organizzate in una forma che il computer sia in grado di comprendere. Per utilizzare appieno la "forza bruta" della macchina, i dataset non strutturati normalmente richiedono una trasformazione dei dati di input in una forma strutturata. Il foglio di lavoro rappresentato nella Figura 1.9 mostra dei dati che sono stati raccolti sotto forma di *matrice*. Nei dati una matrice, ogni riga del foglio di lavoro è un esempio e ogni colonna è una caratteristica. Qui, le righe indicano gli esempi di automobili in vendita, mentre le colonne registrano le caratteristiche di ciascuna automobile, come il prezzo, il chilometraggio, il colore e il tipo di trasmissione. I dati in formato matriciale sono di gran lunga quelli più utilizzati nel campo del machine learning. Come vedrete nei prossimi capitoli, quando in alcune applicazioni si trovano altre forme di dati, vengono comunque trasformate in una matrice prima di passare al machine learning.



year	model	price	mileage	color	transmission
2011	SEL	21992	7413	Yellow	AUTO
2011	SEL	20995	10926	Gray	AUTO
2011	SEL	19995	7351	Silver	AUTO
2011	SEL	17809	11613	Gray	AUTO
2012	SE	17500	8367	White	MANUAL
2010	SEL	17495	25125	Silver	AUTO
2011	SEL	17000	27393	Blue	AUTO
2010	SEL	16995	21026	Silver	AUTO
2011	SES	16995	32655	Silver	AUTO

Figura 1.9 Un semplice dataset sotto forma di matrice che descrive le automobili in vendita.

Le caratteristiche di un dataset possono assumere varie forme. Se una caratteristica è misurabile numericamente, si dice che è una caratteristica *numerica*. Alternativamente, se una caratteristica comprende un insieme di categorie, la caratteristica è detta *categorica* o *nominale*. Uno speciale tipo di variabile categorica è detta *ordinale*, per designare una

variabile nominale le cui categorie cadono in un elenco ordinato. Un esempio di una variabile ordinale è rappresentato dalle taglie degli abiti, per esempio *small*, *medium* e *large*; un altro potrebbe essere il grado di soddisfazione del cliente su una scala da "molto insoddisfatto" a "piuttosto soddisfatto" a "molto soddisfatto". Per ogni dataset, riflettere su che cosa rappresentano le caratteristiche, il loro tipo e le loro unità, aiuterà a determinare un algoritmo di machine learning appropriato per il compito di apprendimento in questione.

Tipi di algoritmi di machine learning

Gli algoritmi di machine learning si suddividono in categorie in base al loro scopo. Conoscere le categorie di algoritmi di apprendimento è un primo passo essenziale nella direzione di un utilizzo dei dati per intraprendere l'azione desiderata.
Un *modello predittivo* viene utilizzato per quei compiti che richiedono, come implica il nome, la predizione di un valore impiegando altri valori presenti nel dataset. L'algoritmo di apprendimento tenta di scoprire e modellare la relazione esistente fra la caratteristica *target* (la caratteristica da predire) e le altre caratteristiche.
Nonostante la parola "predizione" venga comunemente impiegata per implicare una qualche forma di previsione, non necessariamente i modelli predittivi prevedono eventi futuri. Per esempio, un modello predittivo potrebbe essere utilizzato per predire eventi passati, come la data del concepimento di un bambino sulla base degli attuali livelli ormonali della madre. I modelli predittivi possono essere utilizzati anche in tempo reale, per controllare i semafori nell'ora di punta.
Ora, poiché i modelli predittivi ricevono precise istruzioni su ciò di cui hanno bisogno per apprendere e su come è previsto che essi apprendano, il processo di addestramento di un modello predittivo è detto *apprendimento con supervisione*. La supervisione non prevede però un coinvolgimento da parte di un essere umano, ma piuttosto il fatto che i valori target forniscono un modo mediante il quale il sistema di apprendimento può conoscere la qualità del suo apprendimento del compito che gli è stato sottoposto. In termini più formali, dato un insieme di dati, un algoritmo di apprendimento con supervisione tenta di ottimizzare una funzione (il modello) per trovare la combinazione di valori delle caratteristiche che producono l'output che rappresenta il nostro target.
Il compito di machine learning con supervisione frequentemente utilizzato che consiste nel predire a quale categoria appartiene un esempio è chiamato *classificazione*. È facile immaginare i possibili usi di un classificatore. Per esempio, potreste predire se:

- un messaggio di posta elettronica è spam;
- una persona ha un tumore;
- una squadra di calcio vincerà o perderà;
- un richiedente non ottempererà a un prestito.

Nella classificazione, la caratteristica target da predire è una caratteristica categorica nota come *classe*, che si suddivide in categorie chiamate *livelli*. Una classe può avere due o più livelli, e i livelli possono essere o non essere ordinali. La classificazione è così ampiamente utilizzata nel machine learning che esistono vari tipi di algoritmi di classificazione, i cui punti di forza e punti deboli li rendono adatti a diversi tipi di dati di input. Ne vedremo alcuni esempi più avanti in questo stesso capitolo e nel corso di questo libro.
I sistemi di apprendimento con supervisione possono essere utilizzati anche per predire dei dati numerici, come il reddito, i valori di laboratorio, i voti nei test o il numero di

elementi. Per predire tali valori numerici, una forma molto comune di *predizione numerica* prevede l'adattamento di modelli di regressione lineare ai dati di input. Sebbene la regressione non sia l'unico metodo di predizione numerica, è di gran lunga quello più ampiamente utilizzato. I metodi a regressione sono ampiamente utilizzati per le previsioni, in quanto essi quantificano in termini esatti l'associazione fra gli input e il target, comprendendo sia l'entità sia l'incertezza della relazione.

NOTA

Poiché è facile convertire dei numeri in categorie (per esempio, chi ha da 13 a 19 anni rientra nella categoria dei teenager) e delle categorie in numeri (per esempio, assegnando 1 a tutti i maschi e 0 a tutte le femmine), il confine fra modelli di classificazione e i modelli di predizione numerica non è necessariamente netto.

Un *modello descrittivo* viene utilizzato per compiti che si avvantaggerebbero di una conoscenza acquisita riepilogando i dati in modi nuovi e interessanti. Al contrario dei modelli predittivi, che predicono un target cui siamo interessati, in un modello descrittivo, non vi è una caratteristica che sia più importante di ogni altra. In realtà, poiché non vi è alcun target da apprendere, il processo di addestramento di un modello descrittivo è chiamato *apprendimento senza supervisione*. Sebbene possa essere più difficile immaginare applicazioni per i modelli descrittivi (dopotutto, quanto può essere "buono" un sistema di apprendimento che non deve apprendere nulla di specifico) essi sono impiegati abbastanza regolarmente per il data mining.
Per esempio, il compito di modellazione descrittiva chiamato *pattern discovery* (individuazione di schemi) consente di identificare associazioni utili presenti all'interno dei dati. Il *pattern discovery* è l'obiettivo dell'*analisi del carrello della spesa*, che viene applicata ai dati d'acquisto dai rivenditori. Qui, i rivenditori sperano di identificare quegli articoli che vengono frequentemente acquistati insieme, in modo che le informazioni apprese possano essere impiegate per raffinare le tattiche di marketing. Per esempio, se un rivenditore scopre che un certo costume da bagno viene frequentemente acquistato insieme alla crema solare, potrebbe riposizionare i due articoli in modo più ravvicinato, oppure prevedere una promozione per l'acquisto di entrambi i prodotti insieme.

NOTA

Originariamente impiegato solo in contesti di vendita al dettaglio, il metodo *pattern discovery* è ora utilizzato in modi piuttosto innovativi. Per esempio, può essere impiegato per individuare degli schemi nei comportamenti fraudolenti o difetti genetici o per identificare le aree di attività criminale.

Il compito della modellazione descrittiva di dividere un dataset in gruppi omogenei è chiamato *clustering*. Talvolta il clustering viene impiegato per l'*analisi della segmentazione*, che identifica dei gruppi di individui che hanno comportamenti o informazioni demografiche simili, per indirizzare loro campagne pubblicitarie basate proprio su queste caratteristiche che essi condividono. Con questo approccio, la macchina identifica i cluster, ma è comunque richiesto un intervento umano per interpretarli. Per esempio, dati cinque cluster di clienti di un negozio, il team di marketing avrà bisogno di comprendere le differenze fra questi gruppi, con lo scopo di creare una promozione adatta a ogni specifico gruppo. Nonostante questo intervento umano, l'impegno sarà comunque inferiore rispetto a creare un'offerta specifica per ogni singolo cliente.

Infine, una classe di algoritmi di machine learning chiamata *meta-learner* non è legata a uno specifico compito di apprendimento, ma si concentra piuttosto sull'imparare a imparare in modo più efficace. Un algoritmo di meta-learning impiega i risultati degli apprendimenti passati per informare i nuovi apprendimenti.
Questo prevede l'uso di algoritmi di apprendimento che imparano a collaborare in team chiamati *ensemble*, così come pure algoritmi che evolvono nel corso del tempo, in un processo chiamato *reinforcement learning* (apprendimento a rafforzamento). Il meta-learning può essere utile per problemi molto compressi o quando le prestazioni di un algoritmo predittivo devono essere le più accurate possibili.
Alcune delle scoperte più interessanti, oggi, nell'ambito del machine learning riguardano proprio il campo del meta-learning. Per esempio, l'*adversarial learning* prevede l'apprendimento basato sui punti deboli di un modello, con lo scopo di migliorare le sue prestazioni future o renderlo più resistente agli attacchi. Esiste anche un forte investimento in termini di ricerca e attività di sviluppo per creare ensemble sempre più grandi e veloci, in grado di modellare vasti dataset impiegando computer ad alte prestazioni o ambienti di cloud-computing.

Tipi di dati di input e relativi algoritmi

La seguente tabella elenca i vari tipi di algoritmi di machine learning trattati in questo libro. Sebbene si tratti solo di una frazione dell'intero insieme di algoritmi di machine learning disponibili, questi metodi rappresenteranno una base sufficiente per comprendere altri metodi che vi capiterà di incontrare in futuro.

Tabella 1.1 Tipi di algoritmi di machine learning trattati nel libro.

Modello	Compito di apprendimento	Capitolo
Algoritmi di apprendimento con supervisione		
K-Nearest Neighbors	Classificazione	Capitolo 3
Naive Bayes	Classificazione	Capitolo 4
Alberi decisionali	Classificazione	Capitolo 5
Sistemi a regole di classificazione	Classificazione	Capitolo 5
Regressione lineare	Predizione numerica	Capitolo 6
Alberi di regressione	Predizione numerica	Capitolo 6
Alberi di modelli	Predizione numerica	Capitolo 6
Reti neurali	Doppio uso	Capitolo 7
Macchine a vettori di supporto	Doppio uso	Capitolo 7
Algoritmi di apprendimento senza supervisione		
Regole associative	Rilevamento pattern (schemi)	Capitolo 8
k-means clustering	Clustering	Capitolo 9
Algoritmi di meta-learning		
Bagging	Doppio uso	Capitolo 11
Boosting	Doppio uso	Capitolo 11
Foreste casuali	Doppio uso	Capitolo 11

Per iniziare ad applicare il machine learning a un progetto reale, avrete bisogno di determinare a quale dei quattro compiti di apprendimento appartiene il vostro progetto: classificazione, predizione numerica, rilevamento pattern o clustering? Il compito guiderà quindi la scelta dell'algoritmo. Per esempio, se state ricercando uno schema e il compito è di rilevamento pattern, probabilmente impiegherete le regole associative. Analogamente, per un problema di clustering probabilmente impiegherete l'algoritmo k-means e per un compito di predizione numerica utilizzerete l'analisi a regressione o degli alberi di regressione.

Per la classificazione, è più difficile trovare una corrispondenza diretta fra un problema di apprendimento e un classificatore appropriato. In questi casi, è utile considerare le varie distinzioni esistenti fra gli algoritmi, distinzioni che risalteranno all'occhio solo studiando approfonditamente ciascuno dei classificatori. Per esempio, nell'ambito dei problemi di classificazione, gli alberi decisionali producono modelli di facile comprensione, mentre i modelli a reti neurali sono notoriamente difficili da interpretare. Se state progettando un modello di valutazione del credito, questa potrebbe essere una distinzione importante, poiché alcune leggi obbligano a informare il richiedente dei motivi per i quali gli è stato rifiutato il prestito. Anche se la rete neurale è più efficace nel predire problemi nel rimborso dei prestiti, se le sue predizioni non possono essere spiegate, si rivela inutilizzabile per questa applicazione.

Per assistere nella scelta dell'algoritmo, in ogni capitolo elencherò i punti di forza e i punti deboli di ciascun algoritmo di apprendimento. Sebbene talvolta troverete che queste caratteristiche escluderebbero determinati modelli, in molti casi la scelta dell'algoritmo sarà arbitraria. Quando è così, siete sostanzialmente liberi di scegliere l'algoritmo con il quale vi trovate più a vostro agio. Altre volte, quando l'obiettivo primario è l'accuratezza della predizione, potreste dover mettere alla prova più modelli e scegliere poi quello che sembra più adatto, oppure impiegare un algoritmo di meta-learning che combini più sistemi di apprendimento differenti, così da sfruttare i punti di forza di ognuno di essi.

Machine Learning con R

Molti degli algoritmi richiesti per svolgere attività di machine learning non sono inclusi nell'ambito dell'installazione base di R. Tuttavia, tali algoritmi sono disponibili grazie a un'ampia community di esperti, che ha condiviso gratuitamente il proprio lavoro. Questi algoritmi dovranno essere installati manualmente su R. Grazie alla qualifica di R, che è un software gratuito e open-source, non vi è alcuna quota da pagare per questa funzionalità. Una raccolta di funzioni R che può essere condivisa fra gli utenti è chiamata *package*. Esistono package gratuiti per ognuno degli algoritmi di machine learning trattati in questo libro. In realtà, questo libro tratta solo una piccola porzione di tutti i package di machine learning di R.

Se siete interessati a conoscere l'ampiezza dei package R, potete osservarne un elenco su *Comprehensive R Archive Network* (CRAN), una raccolta di siti web e ftp distribuiti in tutto il mondo e che forniscono le versioni più aggiornate del software e dei package R. Se avete ottenuto il software R tramite un download, probabilmente avete usato CRAN. Il sito web CRAN è disponibile all'URL `http://cran.r-project.org/index.html`.

NOTA

Se non avete ancora R, il sito web CRAN fornisce istruzioni sull'installazione e informazioni su come cercare aiuto in caso di problemi.

Il link *Packages* situato nel menu a sinistra vi porterà a una pagina nella quale potete sfogliare i package in ordine alfabetico o ordinati per data di pubblicazione. Al momento [della traduzione in italiano di queste pagine, NdT], esiste un totale di 15.675 package, oltre il doppio rispetto al momento in cui è stata pubblicata la seconda edizione di questo libro e oltre il triplo rispetto a quando è stata pubblicata la prima edizione [originale inglese, NdT]. È evidente che la community di R sta fiorendo e questa tendenza non sembra dar segni di rallentamento!
Il link *Task Views*, sempre sul lato sinistro della pagina web di CRAN, fornisce un elenco selezionato dei package suddivisi per argomento. La vista dedicata al machine learning, che elenca i package trattati in questo libro (e molti altri ancora), è disponibile all'URL: `https://cran.r-project.org/index.html`.

Installazione dei package R

Nonostante il ricco insieme di add-on disponibili per R, il formato dei package ne rende l'installazione e l'utilizzo davvero molto semplici. Per illustrare l'utilizzo dei package, installeremo e caricheremo il package RWeka sviluppato da Kurt Hornik, Christian Buchta e Achim Zeileis (per ulteriori informazioni, consultate *Open-Source Machine Learning: R Meets Weka*, in "Computational Statistics" Vol. 24, pp 225-232). Il package RWeka fornisce una raccolta di funzioni che offrono a R l'accesso agli algoritmi di machine learning tramite il package Java Weka, di Ian H. Witten ed Eibe Frank. Per ulteriori informazioni su Weka, consultate: `http://www.cs.waikato.ac.nz/~ml/weka`.

NOTA

Per poter utilizzare il package RWeka, avrete bisogno di installare anche Java, se non è già presente sul vostro sistema (molti computer sono dotati di Java preinstallato). Java è un insieme di strumenti di programmazione, disponibile gratuitamente, che permette l'utilizzo di applicazioni multipiattaforma come Weka. Per ulteriori informazioni e per scaricare Java per il vostro sistema, visitate il sito `http://www.java.com`.

Il modo più diretto per installare un package prevede l'uso della funzione `install.packages()`. Per installare il package RWeka, digitate al prompt dei comandi di R:

```
> install.packages("RWeka")
```

R si connetterà alla pagina di CRAN e scaricherà il package nel formato corretto per il vostro sistema operativo. Alcuni package, come RWeka, richiedono la presenza di altri package per poter essere utilizzati. Queste sono chiamate *dipendenze*. Per default, l'installer scaricherà e installerà automaticamente tutte le dipendenze.

NOTA

La prima volta che installerete un package, R potrebbe chiedervi di scegliere un sito *mirror* di CRAN. In questo caso, scegliete il mirror che vi sembra più vicino al vostro luogo di residenza. Questo, generalmente, garantisce un download più rapido.

Le opzioni di installazione di default sono appropriate per la maggior parte dei sistemi. Tuttavia, in alcuni casi, potreste voler installare un package in un'altra posizione. Per esempio, se non avete i privilegi root o di amministrazione sul vostro sistema, potreste aver bisogno di specificare un percorso di installazione alternativo. Questo può essere ottenuto tramite l'opzione `lib` come segue:

```
> install.packages("RWeka", lib = "/percorso/della/libreria")
```

La funzione di installazione fornisce anche altre opzioni per l'installazione da un file locale, dal codice sorgente o per l'impiego di versioni sperimentali. Potete scoprire il funzionamento di queste opzioni tramite il sistema di Help, impiegando il seguente comando:

```
> ?install.packages
```

Più in generale, l'operatore *punto interrogativo* può essere utilizzato per ottenere un aiuto per ogni funzione R. Basta semplicemente digitare un `?` prima del nome della funzione.

Caricamento e download dei package R

Per risparmiare memoria, R non carica per default ogni package installato.
Al contrario, i package vengono caricati dagli utenti con la funzione `library()` ogni volta che è necessario.

NOTA

Il nome di questa funzione conduce alcuni a confondere i due termini "libreria" e "package". Tuttavia, per la precisione, una libreria fa riferimento solo alla posizione in cui vengono installati i package e non a un package.

Per caricare il package RWeka precedentemente installato, potete digitare il seguente comando:

```
> library(RWeka)
```

Oltre a RWeka, esistono molti altri package R, che ci troveremo a impiegare nei prossimi capitoli. Le istruzioni di installazione verranno fornite nel momento in cui avremo bisogno di tali package.
Per scaricare un package R, si utilizza la funzione `detach()`. Per esempio, per scaricare il package RWeka di cui abbiamo parlato in queste pagine, basta impiegare il seguente comando:

```
> detach("package:RWeka", unload = TRUE)
```

Questo comando ha l'effetto di liberare tutte le risorse impiegate dal package.

Installazione di RStudio

Prima di iniziare a lavorare con R, è assolutamente consigliabile installare anche l'applicazione desktop open source RStudio (Figura 1.10). RStudio è un'ulteriore interfaccia con R, che comprende funzionalità che rendono più agevole, più comodo e più interattivo l'utilizzo del codice R. È disponibile gratuitamente all'URL: `https://www.rstudio.com`.

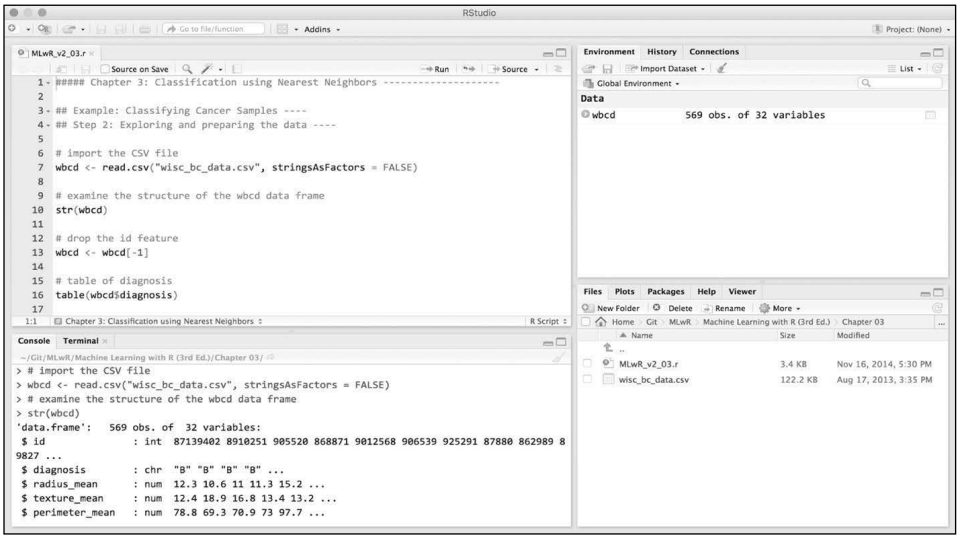


Figura 1.10 L'ambiente desktop RStudio facilità l'utilizzo di R.

L'interfaccia di RStudio comprende un editor per il codice, una console a riga di comando per R, un browser per i file e un browser per gli oggetti R. Gli elementi sintattici del codice R vengono automaticamente colorati e l'output, i grafici e le immagini prodotte da R vengono visualizzate direttamente all'interno dell'ambiente, il che aiuta molto a seguire le istruzioni e i programmi più lunghi e complessi. Altre caratteristiche più avanzate migliorano la gestione del progetto e dei package; l'integrazione con strumenti di controllo del codice sorgente e della versione, come Git e Subversion; la gestione della connessione ai database; la compilazione dell'output di R in formato HTML, PDF o Microsoft Word.

RStudio è uno dei principali motivi per cui R è una delle scelte preferite dei *data scientist* oggi. Ingloba tutta la potenza della programmazione in R, la sua ampia libreria dedicata al machine learning e i package statistici in un'unica interfaccia di facile uso e facile installazione. RStudio non è solo un ambiente ideale per imparare a usare R, ma è anche in grado di estendersi insieme al suo utilizzatore, mano a mano che ne scoprirà le funzionalità più avanzate.

Riepilogo

Il machine learning trae la sua origine all'intersezione fra statistica, scienza dei database e scienza dei computer. Si tratta di uno strumento potente, in grado di trovare conoscenze pratiche all'interno di grandi quantità di dati. Tuttavia, come abbiamo visto in questo capitolo, occorre applicare una certa cautela, con lo scopo di evitare gli abusi dell'applicazione del machine learning nel mondo reale.

Sostanzialmente, il processo di apprendimento prevede l'astrazione dei dati in una rappresentazione strutturata e la generalizzazione della struttura a determinare un'azione che può essere valutata in termini di utilità. In termini pratici, un sistema di apprendimento automatico impiega dei dati contenenti esempi e caratteristiche del concetto da apprendere, e poi raggruppa questi dati a formare un modello, che viene poi impiegato per scopi predittivi o descrittivi. Questi scopi possono essere raggruppati in compiti che comprendono attività di classificazione, predizione numerica, rilevamento di schemi (pattern) e clustering. Fra i tanti possibili metodi, gli algoritmi di machine learning vengono scelti sulla base dei dati di input e del compito di apprendimento da svolgere.

R fornisce il supporto per il machine learning sotto forma di package sviluppati dalla community. Questi strumenti sono disponibili al download gratuitamente, ma devono essere installati prima dell'uso. Ciascun capitolo di questo libro introdurrà tali package ogni volta che saranno necessari.

Nel prossimo capitolo, introdurremo meglio i principali comandi di R impiegati per gestire e preparare i dati al machine learning. Potreste anche sentirvi tentati di saltare questo passo e di passare direttamente alle applicazioni, ma una regola empirica suggerisce che l'80 percento o più del tempo dedicato a un tipico progetto di machine learning è dedicato al passo di preparazione dei dati, il cosiddetto *data wrangling*. Di conseguenza, il vostro investimento iniziale sullo scoprire come "fare le cose" si ripagherà abbondantemente più avanti.