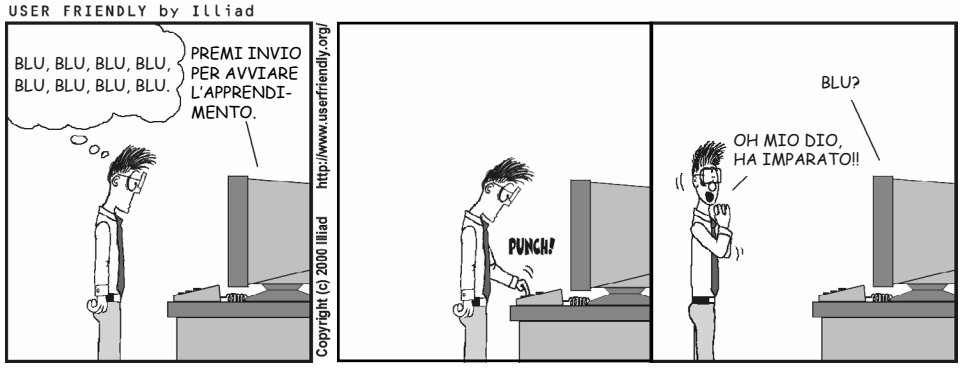
Capitolo 1

Il machine learning: buon senso applicato da un computer



In questo capitolo

- **Sono felice di unirmi a voi in questo viaggio**
- **Occorre essere esperti di matematica e programmazione per studiare il machine learning?**
- **Cos'è quindi il machine learning?**
- **Convincere le macchine a prendere decisioni con i dati: la sequenza ricordare-formulare-prevedere**
- **Riepilogo**

Sono felice di unirmi a voi in questo viaggio

Benvenuti! Sono davvero felice di unirmi a voi in questo vostro viaggio nella comprensione del machine learning. In generale, il machine learning è un processo mediante il quale un computer risolve problemi e prende decisioni più o meno allo stesso modo di un essere umano.
Questo libro veicola un messaggio importante: che il machine learning è un argomento facile. Non è necessario avere solide basi matematiche e competenze di programmazione per seguire queste pagine. Avrete bisogno di un po' di matematica di base, certo, ma i requisiti principali sono il buon senso, un buon intuito visivo e il desiderio di apprendere e applicare questi metodi a tutto ciò che vi appassiona e in cui desiderate migliorare il mondo. Mi sono divertito moltissimo a scrivere questo libro, perché amo approfondire la mia comprensione di questo argomento e spero che anche voi vi divertirete a leggerlo e ad approfondire il machine learning!

Il machine learning è ovunque

Questa affermazione sembra essere ogni giorno più vera. Fatico a immaginare un singolo aspetto della vita che non possa essere migliorato in un modo o nell'altro dal machine learning. Il machine learning può intervenire in qualsiasi lavoro che richieda ripetizioni, analisi dei dati e raccolta di conclusioni. Negli ultimi anni, il machine learning ha registrato un enorme sviluppo grazie ai progressi nella potenza di calcolo e all'ubiquità della raccolta dei dati. Provo a citare solo alcune applicazioni del machine learning: sistemi di raccomandazione, riconoscimento di immagini, elaborazione di testi, auto a guida autonoma, riconoscimento dello spam, diagnosi mediche… ma l'elenco è sterminato. Forse avete un obiettivo o un campo in cui vorreste avere un impatto (o forse lo state già avendo). Molto probabilmente il machine learning può certamente essere applicato a questo campo: forse è proprio questo che vi ha portato a leggere questo libro. Scopriamolo insieme!

Occorre essere esperti di matematica e programmazione per studiare il machine learning?

No. Lo studio del machine learning richiede solo immaginazione, creatività e una mente visiva. Il machine learning si occupa dell'individuazione di modelli che emergono dal mondo e dell'utilizzo di tali modelli per fare previsioni sul futuro. Se vi piace trovare modelli e individuare correlazioni, allora potete utilizzare il machine learning. Se vi dicessi che ho smesso di fumare e che sto mangiando più verdure e facendo più attività fisica, che cosa predireste che accadrà alla mia salute, tra un anno? Che forse migliorerà. Se vi dicessi che sono passato dall'indossare maglioni rossi a maglioni verdi, che cosa predireste che accadrà alla mia salute tra un anno? Che forse non cambierà molto (potrebbe, ma non in base alle informazioni che vi ho dato). Individuare queste correlazioni e questi modelli è ciò di cui si occupa il machine learning. L'unica differenza è che nel machine learning impieghiamo formule e numeri per definire questi schemi, per fare in modo che i computer li possano elaborare.

Per eseguire il machine learning sono necessarie alcune conoscenze di matematica e programmazione, ma non è necessario essere esperti. Se però *siete* esperti in uno di questi campi, o in entrambi, troverete sicuramente che le vostre competenze saranno premiate. Ma se non lo siete, potete comunque imparare a usare il machine learning e apprendere i concetti matematici e di programmazione man mano che procedete. In questo libro introduco tutti i concetti matematici di cui avremo bisogno nel momento in cui ne avremo bisogno. Quanto al programmare, la quantità di codice di machine learning che scriverete dipende da voi. I lavori nel campo del machine learning spaziano da chi passa tutto il giorno a programmare a chi non programma affatto. Molti pacchetti, API e strumenti ci aiutano a eseguire il machine learning con una minima attività di programmazione. Ogni giorno, il machine learning diventa sempre più disponibile per tutti nel mondo e sono felice che voi siate saliti su questo carro!

Formule e codice sono divertenti, se li considerate un linguaggio

Nella maggior parte dei libri sul machine learning, gli algoritmi vengono spiegati matematicamente utilizzando formule, derivate e così via. Sebbene queste descrizioni precise dei metodi funzionino bene nella pratica, una formula isolata può creare più confusione che chiarezza. Tuttavia, come in una partitura musicale, una formula può nascondere dietro la confusione una bella melodia. Consideriamo per esempio questa formula: $\sum_{i=1}^{4} i$. A prima vista sembra brutta, ma rappresenta una somma molto semplice, cioè 1 + 2 + 3 + 4. E che dire di $\sum_{i=1}^{n} w_i$? Questa è semplicemente la somma di molti (n) numeri. Ma quando penso alla somma di molti numeri, preferisco immaginare qualcosa come 3 + 2 + 4 + 27, piuttosto che $\sum_{i=1}^{n} w_i$. Ogni volta che vedo una formula, devo subito immaginarla con un piccolo esempio, e allora l'immagine si fa più chiara nella mia mente. Quando vedo qualcosa come $P(A \mid B)$, che cosa mi viene in mente? Questa è una probabilità condizionata, quindi penso a una frase del tipo "La probabilità che si verifichi un evento A dato il verificarsi di un altro evento B". Per esempio, se A rappresenta "pioggia oggi" e B rappresenta "vivere nella foresta amazzonica", la formula $P(A \mid B) = 0.8$ significa semplicemente "La probabilità che *oggi piova*, dal momento che *viviamo nella foresta amazzonica*, è dell'80%."

Se amate le formule, tranquilli: questo libro le contiene ancora. Ma appariranno solo dopo l'esempio che le illustra.

Lo stesso accade per il codice. Se osserviamo il codice, può sembrarci complicato e potremmo avere difficoltà a immaginare che qualcuno possa inserire tutte queste cose in testa. Tuttavia, il codice non è altro che una sequenza di passaggi e normalmente ciascuno di questi passaggi è semplice. In questo libro scriveremo del codice, ma sarà suddiviso in semplici passaggi e ogni passaggio sarà spiegato attentamente con esempi o illustrazioni. Nel corso dei primi capitoli programmeremo i modelli da zero, per capire come funzionano. Nei capitoli successivi, tuttavia, i modelli diventano più complicati. Per questi utilizzeremo pacchetti come Scikit-Learn, Turi Create o Keras, che hanno già implementato la maggior parte degli algoritmi di machine learning con grande chiarezza e potenza.

Cos'è quindi il machine learning?

Per definire il machine learning, definiamo innanzitutto un termine più generale: l'intelligenza artificiale.

Che cos'è l'intelligenza artificiale?

L'*intelligenza artificiale* (IA) è un termine generale, che definiamo come segue.

> *Intelligenza artificiale*: l'insieme di tutti i compiti in cui un computer può prendere decisioni.

In molti casi, un computer prende queste decisioni imitando il modo in cui un essere umano prende le decisioni. In altri casi, possono imitare processi evolutivi, processi genetici o processi fisici. Ma in generale, ogni volta che vediamo un computer risolvere un problema da solo, che si tratti di guidare un'auto, trovare un percorso tra due punti, diagnosticare un paziente o consigliare un film, stiamo guardando l'intelligenza artificiale.

Che cos'è il machine learning?

Il machine learning è simile all'intelligenza artificiale e spesso le loro definizioni sono confuse. *Il machine learning* (ML) è una parte dell'intelligenza artificiale e lo definiamo come segue.

> *Machine learning*: l'insieme di tutte le attività in cui un computer può prendere decisioni *sulla base di dati*.

Che cosa significa questo? Permettetemi di illustrarlo con il diagramma nella Figura 1.1.

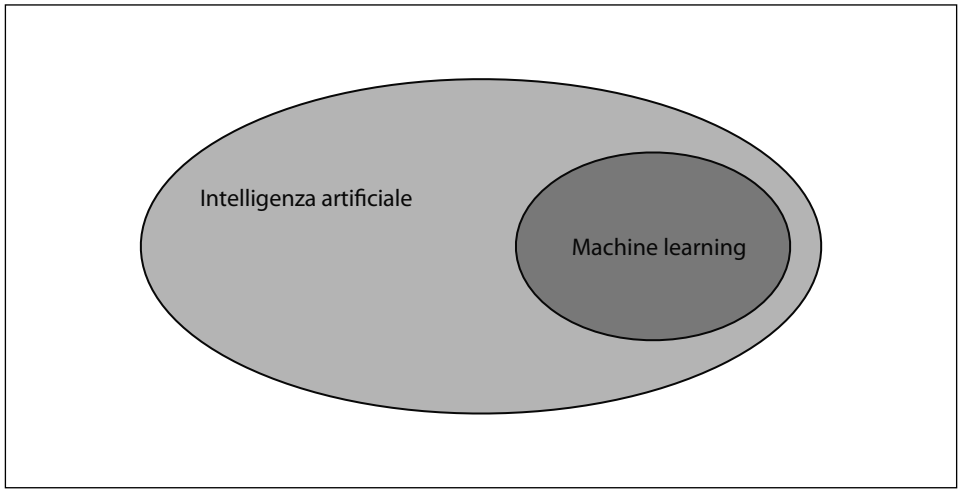


Figura 1.1 Il machine learning è una parte dell'intelligenza artificiale.

Torniamo a osservare in che modo gli esseri umani prendono le decisioni. In termini generali, prendiamo decisioni nei due modi seguenti:

- utilizzando la logica e il ragionamento;
- utilizzando la nostra esperienza.

Per esempio, immaginate di dover decidere quale auto acquistare. Potete osservare attentamente le caratteristiche dell'auto, come il prezzo, il consumo di carburante e la spaziosità, e cercare di trovare la combinazione più adatta al nostro budget. Questo significa usare la logica e il ragionamento. Se invece chiediamo a tutti i nostri amici quali auto possiedono, che cosa apprezzano o non apprezzano del loro mezzo, poi creiamo un elenco di informazioni e lo utilizziamo per decidere, allora stiamo usando l'esperienza (in questo caso, le esperienze dei nostri amici).

Il machine learning rappresenta il secondo metodo: prendere decisioni utilizzando la nostra esperienza. Nel mondo dei computer, il termine usato per parlare di *esperienza* è *dati*. Pertanto, nel machine learning, i computer prendono decisioni sulla base di dati. E così, ogni volta che utilizziamo un computer per risolvere un problema o per prendere una decisione utilizzando solo dati, stiamo utilizzando il machine learning. Colloquialmente, potremmo descrivere il machine learning nel modo seguente: "Il machine learning applica il buon senso, e ad applicarlo è un computer".

Il fatto di passare dalla risoluzione dei problemi utilizzando qualsiasi mezzo necessario al fare la stessa cosa utilizzando solo dati può sembrare un piccolo passo per un computer, ma è stato un passo enorme per l'umanità (Figura 1.2). Un tempo, se volevamo far eseguire un compito a un computer, dovevamo scrivere un programma, vale a dire un intero insieme di istruzioni che il computer avrebbe dovuto seguire. Questo processo è utile per compiti semplici, ma alcuni compiti sono troppo complicati per questo schema di comportamento. Consideriamo per esempio il compito di identificare se un'immagine contiene una mela. Se iniziassimo a scrivere un programma per sviluppare questo compito, scopriremmo subito che è difficile.

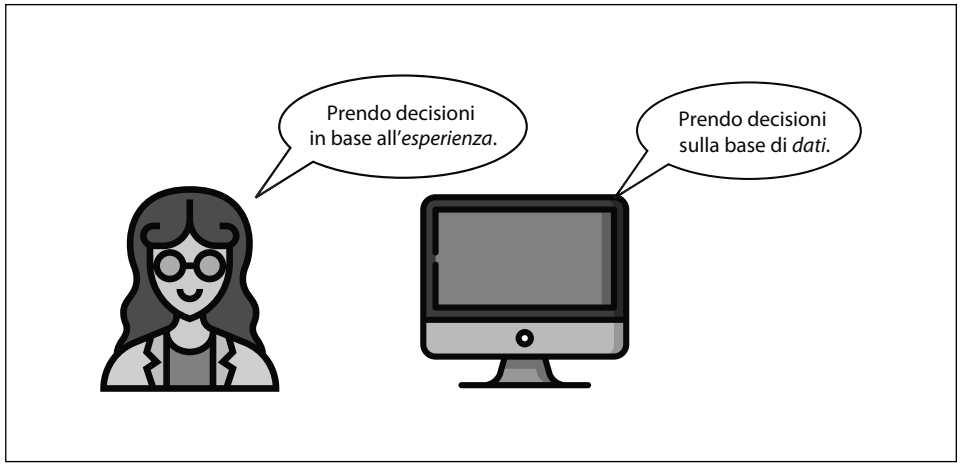


Figura 1.2 Il machine learning comprende tutte le attività in cui i computer prendono decisioni sulla base di dati. Allo stesso modo in cui gli esseri umani prendono decisioni basate su esperienze precedenti, i computer possono prendere decisioni basate su dati precedenti.

Facciamo un passo indietro e poniamoci la seguente domanda. Come abbiamo imparato, come esseri umani, che aspetto ha una mela? Il modo in cui abbiamo imparato la maggior parte delle parole non è stato grazie a qualcuno che ci ha spiegato il loro significato; le abbiamo imparate ripetendole. Abbiamo visto molti oggetti nel corso della nostra infanzia, e gli adulti ci hanno detto qual era il loro nome. Per sapere che cosa fosse una mela, abbiamo visto molte mele nel corso degli anni, e intanto sentivamo la parola *mela*, finché un giorno abbiamo capito che cos'era una mela. Nel machine learning, questo è ciò che facciamo fare al computer. Mostriamo al computer tante immagini e gli diciamo quali contengono una mela (che costituisce i dati). Ripetiamo questo processo finché il computer non rileva i modelli e gli attributi corretti che costituiscono una mela. Alla fine del processo, quando diamo al computer una nuova immagine, può utilizzare questi modelli per determinare se l'immagine contiene una mela. Naturalmente, dobbiamo ancora programmare il computer in modo che rilevi questo schema. Per farlo abbiamo a disposizione diverse tecniche, che impareremo nel corso del libro.

Che cos'è il deep learning?

Così come il machine learning fa parte dell'intelligenza artificiale, il deep learning fa parte del machine learning. Nel paragrafo precedente, abbiamo scoperto che utilizziamo varie tecniche per fare in modo che il computer impari dai dati. Una di queste tecniche ha funzionato straordinariamente bene, quindi ha un proprio campo di studio, chiamato *deep learning* (DL), che definiamo come segue e come vediamo nella Figura 1.3.

> *Deep learning*: il campo del machine learning che utilizza determinati oggetti chiamati *reti neurali*.

Che cosa sono le reti neurali? Ne parleremo nel Capitolo 10. Il deep learning è probabilmente il tipo di machine learning più utilizzato, perché funziona davvero bene. Se stiamo usando una qualsiasi applicazione all'avanguardia, che esegue il riconoscimento di immagini, che genera testo, che gioca a Go o se conduciamo un'auto a guida autonoma, molto probabilmente stiamo usando il deep learning, in un modo o nell'altro.

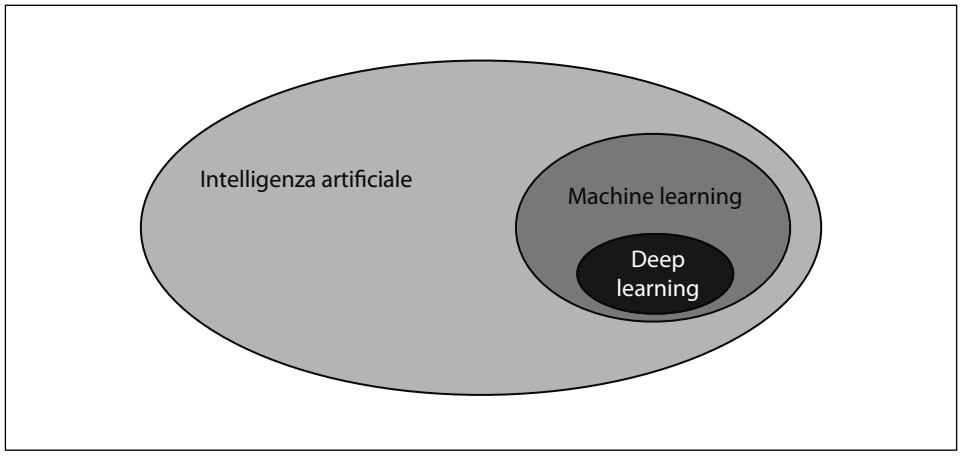


Figura 1.3 Il deep learning è una parte del machine learning.

In altre parole, il deep learning fa parte del machine learning, che a sua volta fa parte dell'intelligenza artificiale. Se questo libro riguardasse i trasporti, l'intelligenza artificiale sarebbe costituita dai veicoli, il machine learning dalle automobili e il deep learning dalle Ferrari.

Convincere le macchine a prendere decisioni con i dati: la sequenza ricordare-formulare-prevedere

Nel paragrafo precedente abbiamo parlato del fatto che il machine learning è costituito da un insieme di tecniche che utilizziamo per far sì che il computer prenda decisioni sulla base di dati. In questo paragrafo scopriremo che cosa significa prendere decisioni sulla base di dati e come funzionano alcune di queste tecniche. Per farlo, analizziamo nuovamente il processo utilizzato dagli esseri umani per prendere decisioni sulla base dell'esperienza. Questo è ciò che viene chiamato *sequenza ricordare-formulare-prevedere*, rappresentata nella Figura 1.4. L'obiettivo del machine learning è quello di insegnare ai computer a pensare allo stesso modo, con la stessa sequenza.

Come pensano gli esseri umani?

Quando noi, come esseri umani, dobbiamo prendere una decisione in base alla nostra esperienza, normalmente utilizziamo questa sequenza.

1. *Ricordiamo* situazioni passate che erano simili.
2. *Formuliamo* una regola generale.
3. Usiamo questa regola per *prevedere* che cosa può accadere in futuro.

Per esempio, se la domanda è "Pioverà, oggi?", il nostro processo per formulare un'ipotesi può essere il seguente.

1. *Ricordiamo* che la settimana scorsa ha piovuto la maggior parte delle volte.
2. *Formuliamo* l'idea che, in questo luogo, piove la maggior parte delle volte.
3. *Prevediamo* che oggi pioverà.

Potremmo avere ragione o torto, ma stiamo cercando di fare la previsione più accurata possibile sulla base delle informazioni in nostro possesso.

Termini tipici del machine learning: modelli e algoritmi

Prima di approfondire l'argomento con ulteriori esempi che illustrano le tecniche utilizzate nel machine learning, definiamo alcuni termini utili che utilizzeremo in questo libro. Sappiamo che nel machine learning facciamo in modo che il computer impari a risolvere un problema utilizzando i dati. Il modo in cui il computer risolve il problema consiste nell'utilizzare i dati per costruire un *modello*. Che cos'è un modello? Definiamo un modello come segue.

> *Modello*: un insieme di regole che rappresentano i dati e possono essere utilizzate per fare previsioni.

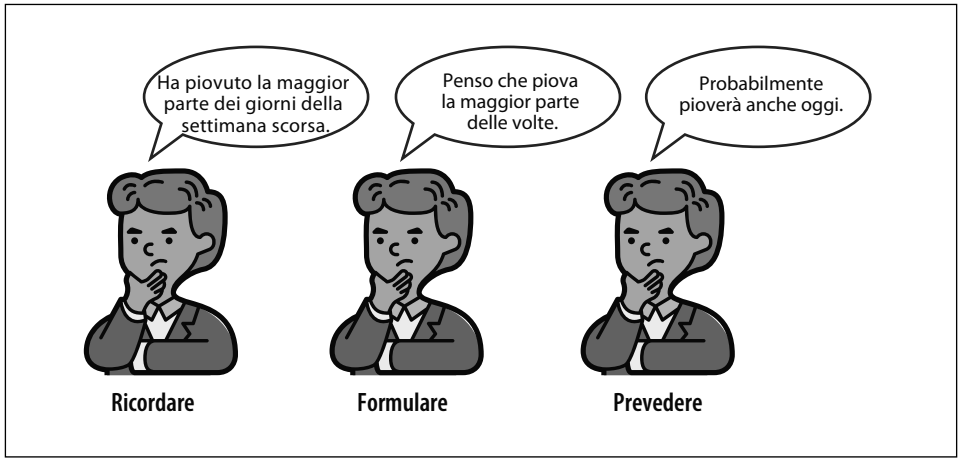


Figura 1.4 La sequenza ricordare-formulare-prevedere è la principale che utilizziamo in questo libro. Si compone di tre passaggi: (1) ricordiamo i dati precedenti, poi (2) formuliamo una regola generale, infine (3) usiamo quella regola per fare previsioni sul futuro.

Possiamo considerare un modello come una rappresentazione della realtà utilizzando un insieme di regole che imitano il più fedelmente possibile i dati disponibili. Nell'esempio della pioggia del paragrafo precedente, il modello era la nostra rappresentazione della realtà, ovvero un mondo in cui piove per la maggior parte del tempo. Questo è un mondo semplice con una regola: *piove la maggior parte del tempo*. Questa rappresentazione può essere accurata o meno, ma secondo i dati è la rappresentazione più accurata della realtà che possiamo formulare. Successivamente utilizziamo questa regola per fare previsioni su dati ancora ignoti.

Un *algoritmo* è il processo che abbiamo utilizzato per costruire il modello. In questo esempio, il processo è semplice: abbiamo osservato quanti giorni ha piovuto e ci siamo resi conto che erano la maggioranza. Naturalmente, gli algoritmi di machine learning possono diventare molto più complicati di così, ma alla fine sono sempre composti da una serie di passaggi. Ecco la nostra definizione di algoritmo.

> *Algoritmo*: una procedura, o una serie di passaggi, utilizzata per risolvere un problema o eseguire un calcolo. In questo libro, l'obiettivo di un algoritmo è quello di costruire un modello.

In breve, un modello è ciò che utilizziamo per fare previsioni e un algoritmo è ciò che utilizziamo per costruire il modello. Queste due definizioni sono facili da confondere e spesso vengono usate l'una al posto dell'altra, ma per mantenerle distinte, esaminiamo alcuni esempi.

Esempi di modelli utilizzati dagli esseri umani

In questo paragrafo ci concentriamo su un'applicazione comune del machine learning: il rilevamento dello spam. Negli esempi seguenti, rileveremo le e-mail di spam e quelle non spam. Le e-mail non spam vengono anche chiamate *ham* (prosciutto, ovvero "ciccia").

Spam è il termine comunemente utilizzato per la posta fastidiosa o indesiderata, come le "catene di sant'Antonio", le promozioni e così via. Il termine deriva da uno sketch del 1972 dei Monty Python in cui ogni piatto del menu di un ristorante conteneva come ingrediente lo Spam (un blocco di carne macinata in scatola non… "particolarmente ricercato"). Tra gli sviluppatori di software, il termine *ham* viene così utilizzato per tutte le e-mail che non sono di spam.

Esempio 1: un fastidioso amico e la sua posta elettronica

In questo esempio, al nostro amico Bob piace inviarci e-mail. Molte delle sue e-mail sono di spam, sotto forma di catene di sant'Antonio. Stiamo iniziando a spazientirci. È sabato e abbiamo appena ricevuto la notifica di un'e-mail da Bob. Possiamo indovinare se questa e-mail è spam o ham senza guardarla?

Per capirlo, utilizziamo il metodo ricordare-formulare-prevedere. Per prima cosa *ricordiamo*, per esempio, le ultime dieci e-mail che abbiamo ricevuto da Bob. Questi sono i dati, e ci ricordiamo che sei di esse erano spam e le altre quattro erano ham. Da queste informazioni possiamo *formulare* il seguente modello.

- *Modello 1*: sei e-mail su dieci di Bob sono di spam.

Questa regola sarà il nostro modello. Attenzione, non è necessario che questa regola sia vera. Potrebbe essere del tutto sbagliata. Ma sulla base dei dati, questo è il meglio che possiamo ottenere, quindi accetteremo la possibilità di essere in errore. Più avanti in questo libro impareremo a valutare i modelli e a migliorarli quando necessario.

Ora che abbiamo la nostra regola, possiamo usarla per *prevedere* se l'e-mail è spam. Se sei e-mail su dieci di Bob sono di spam, allora possiamo supporre che questa nuova e-mail abbia il 60% di probabilità di essere spam e il 40% di probabilità di essere ham. A giudicare da questa regola, è abbastanza sicuro pensare che l'e-mail sia di spam. Pertanto, prevediamo che l'e-mail sia di spam (Figura 1.5).

Ancora una volta: la nostra previsione potrebbe essere sbagliata. Potremmo aprire l'e-mail e renderci conto che è "buona". Ma abbiamo fatto la previsione *al meglio delle nostre conoscenze*. Questo è ciò che fa il machine learning.

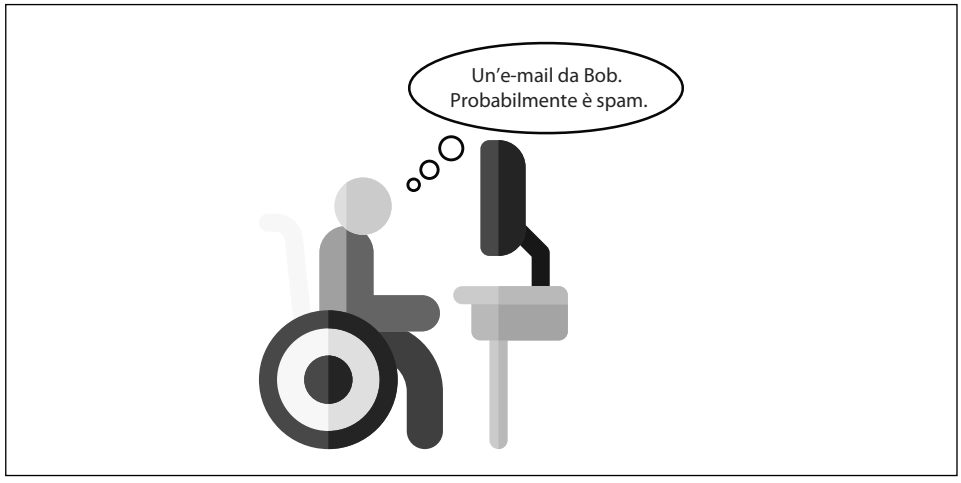


Figura 1.5 Un modello di machine learning molto semplice.

Potreste pensare se sia possibile fare di meglio. Sembra che giudichiamo ogni e-mail di Bob allo stesso modo, ma potrebbero esserci anche altre informazioni che possono aiutarci a distinguere le e-mail di spam da quelle ham. Proviamo ad analizzare un po' di più le e-mail. Per esempio, vediamo quando Bob ha inviato le e-mail, per vedere se emerge uno schema.

Esempio 2: un fastidioso amico che ogni tanto ci invia e-mail

Osserviamo più attentamente le e-mail che Bob ci ha inviato nel mese precedente. Per la precisione, vediamo in che giorno le ha inviate. Ecco le e-mail con le date e l'informazione spam/ham.

- Lunedì: ham.
- Martedì: ham.
- Sabato: spam.
- Domenica: spam.
- Domenica: spam.
- Mercoledì: ham.
- Venerdì: ham.
- Sabato: spam.
- Martedì: ham.
- Giovedì: ham.

Ora le cose cambiano. Riuscite a vedere uno schema? Sembra che ogni e-mail inviata da Bob durante la settimana sia "buona", mentre ogni e-mail inviata durante il fine settimana sia spam. Questo ha senso: forse durante la settimana ci invia e-mail di lavoro, mentre durante il fine settimana ha tempo per inviare spam e decide di sfogarsi liberamente. Quindi, possiamo *formulare* una regola, o modello, più intelligente, come segue.

- *Modello 2*: ogni e-mail che Bob ci invia durante la settimana è ham, mentre quelle che invia durante il fine settimana sono di spam.

Ora vediamo che giorno è oggi. Se è domenica e abbiamo appena ricevuto un'e-mail da Bob, possiamo *prevedere* con grande sicurezza che l'e-mail che ci ha inviato sia spam (Figura 1.6). Facciamo questa previsione e, senza guardare, mandiamo l'e-mail nella spazzatura e andiamo avanti con la nostra giornata.

Esempio 3: le cose si stanno complicando!

Ora, diciamo che per un po' procediamo con questa regola, ma un giorno incontriamo Bob per strada e ci chiede: "Perché non sei venuto alla mia festa di compleanno?". Non abbiamo idea di cosa stia parlando. A quanto pare, domenica scorsa ci ha inviato un invito alla sua festa di compleanno e ce la siamo persa. Perché ce la siamo persa? Perché l'ha inviata nel fine settimana e abbiamo pensato che fosse spam. Sembra che abbiamo bisogno di costruire un modello migliore. Torniamo a esaminare le e-mail di Bob: questo è il nostro passaggio *ricordare*.

Vediamo se riusciamo a trovare uno schema.

- 1 KB: ham.
- 2 KB: ham.

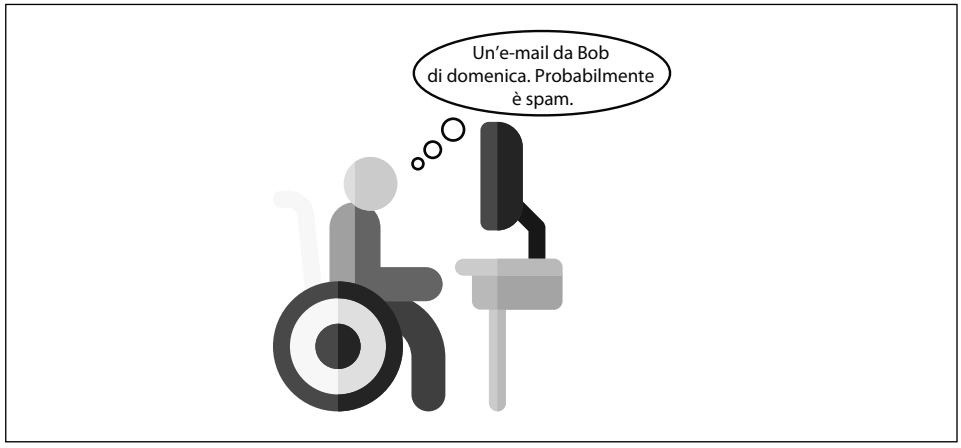


Figura 1.6 Un modello di machine learning leggermente più complesso.

- 16 KB: spam.
- 20 KB: spam.
- 18 KB: spam.
- 3 KB: ham.
- 5 KB: ham.
- 25 KB: spam.
- 1 KB: ham.
- 3 KB: ham.

Che cosa notiamo? Sembra che le e-mail di grandi dimensioni tendano a essere spam, mentre quelle più piccole tendono a essere ham. Questo è logico, perché le e-mail di spam spesso contengono allegati di grandi dimensioni.
Possiamo quindi *formulare* la seguente regola.

- *Modello 3*: qualsiasi e-mail di dimensione pari o superiore a 10 KB è spam, mentre qualsiasi e-mail di dimensione inferiore a 10 KB è ham.

Ora che abbiamo formulato la nostra regola, possiamo fare una *previsione.* Guardiamo l'e-mail che abbiamo ricevuto oggi da Bob e la dimensione è 19 KB. Quindi concludiamo che si tratta di spam (Figura 1.7).
Siamo arrivati alla fine della storia? Neanche lontanamente.
Ma prima di proseguire, tenete presente che per fare le nostre previsioni abbiamo utilizzato il giorno della settimana e la dimensione dell'e-mail. Questi sono esempi di *caratteristiche*. Quello della caratteristica è uno dei concetti più importanti di questo libro.

> *Caratteristica* (*Feature*, *in inglese*): qualsiasi proprietà o qualità dei dati che il modello può utilizzare per fare previsioni.

Potete immaginare che ci siano molte altre caratteristiche che potrebbero indicare se un'e-mail è spam o ham. Ve ne vengono in mente altre? Nei prossimi paragrafi esamineremo alcune altre caratteristiche.

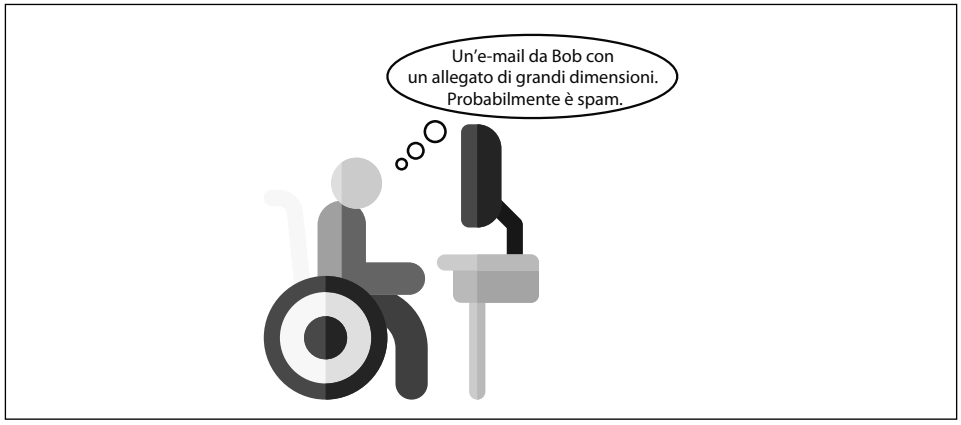


Figura 1.7 Un altro modello di machine learning, leggermente più complesso.

Esempio 4: andiamo oltre

I nostri due classificatori erano buoni, perché hanno escluso le e-mail di grandi dimensioni e quelle inviate nei fine settimana. Ognuno di essi utilizza esattamente una di queste due caratteristiche. E se volessimo una regola che consideri entrambe le caratteristiche? Regole come le seguenti potrebbero funzionare.

- *Modello 4*: se un'e-mail è più grande di 10 KB o viene inviata nel fine settimana, va classificata come spam. Altrimenti va classificata come ham.
- *Modello 5*: se l'e-mail viene inviata durante la settimana, deve essere più grande di 15 KB per essere classificata come spam. Se viene inviata durante il fine settimana, deve essere più grande di 5 KB per essere classificata come spam. Altrimenti va classificato come ham.

Ma possiamo diventare ancora più complicati.

- *Modello 6*: consideriamo il numero del giorno, dove lunedì è 0, martedì è 1, mercoledì è 2, giovedì è 3, venerdì è 4, sabato è 5 e domenica è 6. Se aggiungiamo il numero del giorno e la dimensione dell'e-mail (in KB) e il risultato è pari o superiore a 12, l'e-mail va classificata come spam (Figura 1.8). Altrimenti va classificato come ham.

Figura 1.8 Un modello di machine learning ancora più complesso.

Tutti questi sono modelli validi. Possiamo continuare a creare altri modelli aggiungendo ulteriori livelli di complessità o esaminando ancora più caratteristiche. Ora la domanda è: qual è il modello migliore? È qui che iniziamo ad aver bisogno dell'aiuto di un computer.

Esempi di modelli utilizzati dalle macchine

L'obiettivo è quello di far sì che il computer pensi nel modo in cui pensiamo noi, ovvero utilizzi la sequenza ricordare-formulare-prevedere. In pratica, ecco che cosa fa il computer in ciascuno dei passaggi.

1. *Ricordare*: studia un'enorme tabella di dati.
2. *Formulare*: crea dei modelli esaminando molte regole e formule e controlla quale modello approssima meglio i dati.
3. *Prevedere*: utilizza il modello per fare previsioni sui dati futuri.

Questo processo non è molto diverso da quello che abbiamo fatto nel paragrafo precedente. Il grande progresso sta nel fatto che il computer può costruire i modelli rapidamente, esaminando molte formule e combinazioni di regole finché non ne trova una che approssima particolarmente bene i dati disponibili. Per esempio, possiamo creare un classificatore di spam con caratteristiche come il mittente, la data e l'ora del giorno, il numero di parole, il numero di errori di ortografia e la presenza di determinate parole come *acquista* o *vinci*. Un modello potrebbe facilmente somigliare alla seguente affermazione logica.

- *Modello 7:*
  - Se l'e-mail contiene due o più errori di ortografia, va classificata come spam.
  - Se contiene un allegato di dimensioni maggiori di 10 KB, va classificata come spam.
  - Se il mittente non è presente nella nostra lista dei contatti, va classificata come spam.
  - Se contiene le parole *acquista e vinci*, va classificata come spam.
  - Altrimenti va classificata come ham.

Potrebbe anche somigliare alla seguente formula.

- *Modello 8*: se (dimensione) + 10 (numero errori di ortografia) – (numero occorrenze parola "compra") + 4 (numero presenze parola "acquista") > 10, allora classifica il messaggio come spam (Figura 1.9). Altrimenti classificalo come ham.

Ora la domanda è: "Qual è la regola migliore?". La risposta rapida è: "Quella che approssima meglio i dati", sebbene la risposta vera sia "Quella che si generalizza meglio sui nuovi dati". Alla fine, potremmo ritrovarci con una regola complicata, ma che il computer può formulare e utilizzare per fare rapidamente previsioni. La nostra prossima domanda è: "Come possiamo costruire il modello migliore?". Questo è esattamente l'argomento di questo libro.

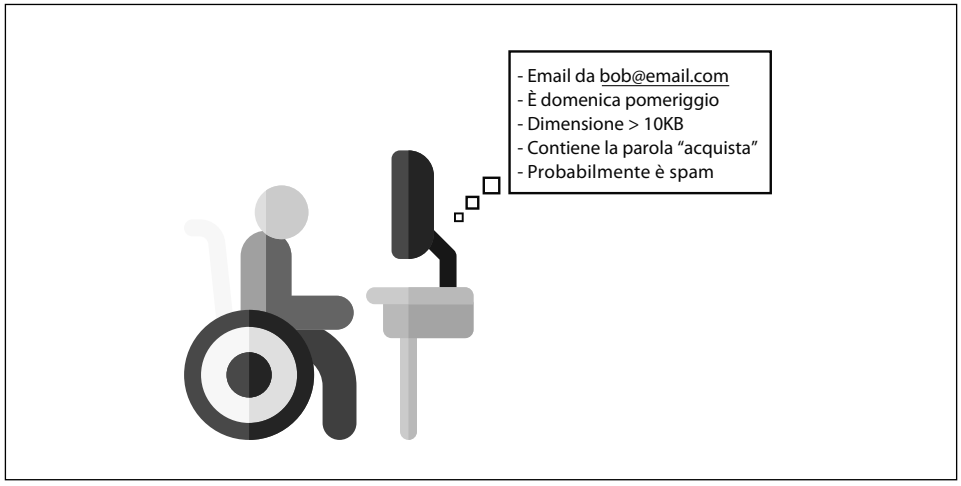


Figura 1.9 Un modello di machine learning molto più complesso, trovato da un computer.

Riepilogo

- Il machine learning è facile. Chiunque può apprenderlo e utilizzarlo, indipendentemente dal proprio background culturale. Tutto ciò che serve è avere la voglia di imparare e avere grandi idee da realizzare.
- Il machine learning è estremamente utile ed è utilizzato nella maggior parte delle discipline. Dalla scienza alla tecnologia, dai problemi sociali alla medicina, il machine learning sta avendo un impatto sulla vita quotidiana e continuerà a farlo.
- Il machine learning è buon senso, ma applicato da un computer. Imita gli schemi di pensiero degli esseri umani per prendere decisioni in modo rapido e accurato.
- Proprio come gli esseri umani prendono decisioni in base all'esperienza, i computer possono prendere decisioni sulla base di dati precedenti. Questo è esattamente ciò che fa il machine learning.

Il machine learning utilizza la sequenza ricordare-formulare-prevedere, come segue.

1. *Ricordare*: considera i dati precedenti.
2. *Formulare*: costruisci un modello, o una regola, sulla base di questi dati.
3. *Prevedere*: utilizza il modello per fare previsioni sui dati futuri.