Capitolo 1

Essere uno scienziato dei dati

In questo capitolo

- **Che cosa si intende per scienza dei dati?**
- **Il diagramma di Venn e la scienza dei dati**
- **Ancora un po' di terminologia**
- **Scienza dei dati: casi di studio**

Qualunque sia il campo professionale nel quale prestate la vostra opera, IT, moda, cibo o finanza, non vi è alcun dubbio che i dati influenzano in qualche modo la vostra vita e il vostro lavoro. Certamente, nel corso di questa settimana, avete avuto o, almeno, avete ascoltato una conversazione relativa ai dati. I media si occupano sempre più spesso di notizie relative a fughe di dati (*data leaks*), cybercrimine e al fatto che i dati possono fornire molte informazioni sulla nostra vita. Perché proprio ora? Cos'è che rende la nostra epoca così proficua per le aziende che si occupano di dati?

Nel Diciannovesimo secolo, il mondo era nel pieno dell'*era industriale*. Il genere umano stava esplorando la sua collocazione nel mondo della produzione industriale fianco a fianco con gigantesche invenzioni meccaniche. I capitani d'industria, come Henry Ford, individuarono grandi opportunità di mercato grazie all'impiego di queste macchine e sono stati in grado di ottenere profitti che, in precedenza, erano inimmaginabili. Naturalmente l'era industriale ebbe i suoi pro e i suoi contro. Da un lato la produzione di massa metteva sempre più beni nelle mani di sempre più consumatori, ma dall'altro, quella è l'epoca in cui è iniziata la nostra battaglia contro l'inquinamento. Giunto il Ventesimo secolo, eravamo ormai piuttosto abili nella creazione di grandi macchine; l'obiettivo diventò così quello di renderle sempre più piccole e veloci. L'era industriale era finita ed era stata sostituita da quella che chiamiamo *era dell'informazione*. Abbiamo iniziato a usare le macchine per raccogliere

e conservare informazioni (dati) su noi stessi e il nostro ambiente con lo scopo di comprendere meglio il nostro universo.
All'inizio degli anni Quaranta, macchine come l'ENIAC (considerato uno dei primi, se non "il" primo, computer) erano in grado di calcolare equazioni matematiche e di eseguire modelli e simulazioni come mai prima di allora.

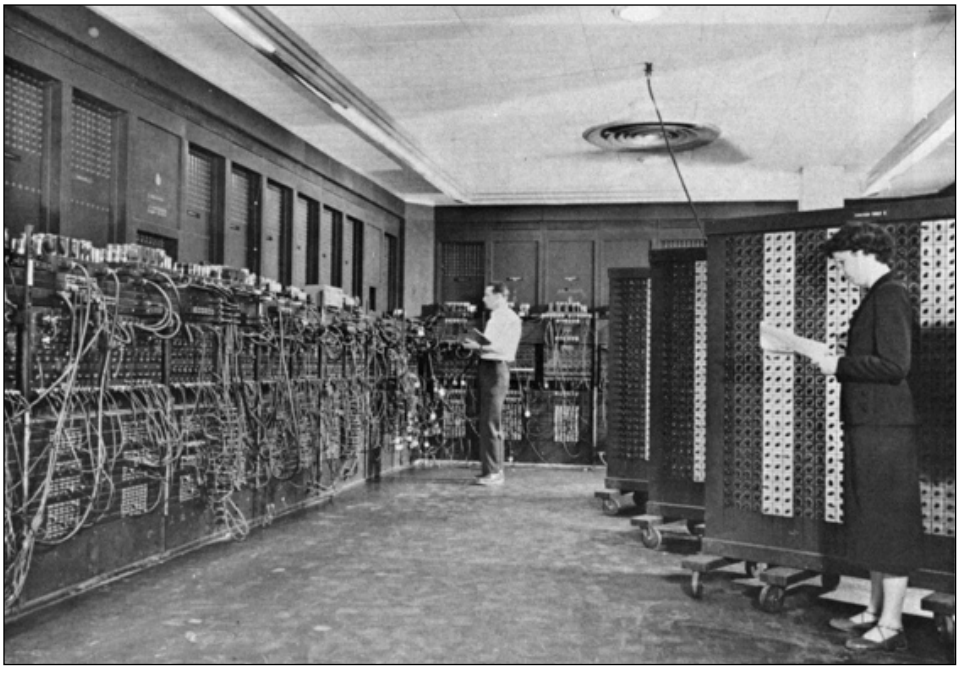


Figura 1.1 Il computer ENIAC (http://ftp.arl.mil/ftp/historic-computers/).

Alla fine abbiamo trovato un buon assistente di laboratorio che era in grado di macinare numeri meglio di noi! Come per l'era industriale, l'era dell'informazione ha avuto aspetti positivi e altri negativi. Fra le cose buone annoveriamo straordinari oggetti tecnologici, come gli smartphone e la televisione. Fra le cose cattive, niente di così terribile come l'inquinamento, ma comunque un bel problema per il Ventunesimo secolo: troppi dati.
È così: l'era dell'informazione, nella sua continua ricerca di generare dati, ha fatto letteralmente esplodere la produzione di dati elettronici. Secondo le stime, nel 2011 abbiamo creato circa 1 800 miliardi di GB di dati (dedicate un attimo a riflettere sull'immensa quantità di cui stiamo parlando). L'anno successivo, nel 2012, siamo saliti a 2 800 miliardi di GB di dati! Questa cifra è destinata a esplodere ulteriormente, e le stime indicano che entro il 2020, in un solo anno produrremo addirittura 40 000 GB di dati. Ognuno di noi contribuisce a comporre questa cifra a ogni tweet, a ogni post su Facebook, ogni volta che salviamo un curriculum scritto con Microsoft Word o che inviamo alla mamma una foto tramite un'app.
Non solo creiamo dati a un livello senza precedenti, ma li consumiamo anche a un tasso sempre più accelerato. Nel solo 2013, un utente medio di uno smartphone usava circa 1 GB di dati al mese. Oggi, tale cifra si stima abbia superato di gran lunga i 2 GB al mese.

Non cerchiamo solo quel banalissimo test della personalità: quello che cerchiamo è la conoscenza. Di tutti quei dati che ci sono "là fuori", qualcosa di utile ci deve essere, per me! Di conseguenza, in pieno ventunesimo secolo, ci rimane un problema. Abbiamo un'enorme quantità di dati e ne creiamo sempre di più. Abbiamo costruito macchine sempre più piccole in grado di raccogliere dati "24/7" e ora il nostro compito consiste nel capirne il senso. Questo significa essere nell'*era dei dati*. È un'era in cui usiamo macchine che i nostri antenati del Diciannovesimo secolo non potevano nemmeno sognare e i dati da noi creati nel Ventesimo secolo per creare informazioni e fonti di conoscenza delle quali può beneficiare ogni essere umano. Gli Stati Uniti hanno creato un apposito ruolo di governo per il capo dei data scientist. Le società che si occupano di tecnologie, come Reddit, che finora non si erano dotate di un esperto di scienza dei dati nel proprio staff, ora li stanno cercando ovunque. Il vantaggio è piuttosto ovvio: usare i dati per effettuare previsioni e simulazioni accurate ci permette di conoscere il funzionamento del mondo come mai prima d'ora.
Sembra una gran cosa, ma… come si fa?
Questo capitolo esplorerà la terminologia e il vocabolario impiegati dagli esperti di scienza dei dati. Introdurremo le parole chiave e le frasi che sono essenziali per parlare di scienza dei dati. Vedremo anche perché usare la scienza dei dati e i tre domini da cui deriva, prima di iniziare a esaminare il codice in Python, il linguaggio prevalentemente usato in questo libro:

- terminologia di base della scienza dei dati;
- i tre domini della scienza dei dati;
- elementi di base della sintassi di Python.

Che cosa si intende per scienza dei dati?

Prima di procedere, esaminiamo alcune definizioni di base che utilizzeremo molto spesso nel corso di questo libro. Uno degli aspetti più sorprendenti (in senso positivo e negativo) di una disciplina così giovane è che queste definizioni possono differire anche molto a seconda del libro, dell'articolo o del documento in cui si trovano.

Terminologia di base

Le definizioni seguenti sono sufficientemente generalizzate da poter essere impiegate nelle conversazioni quotidiane e sul luogo di lavoro in base agli scopi di questo libro, che è un'introduzione ai principi della scienza dei dati.
Iniziamo definendo che cosa si intende per dati. Può sembrare un po' sciocca come prima definizione, ma è anche la più fondamentale. Ogni volta che usiamo la parola "dati", facciamo riferimento a una collezione di informazioni in un formato *organizzato* o *non organizzato*.

- *Dati organizzati*: si tratta di dati ordinati in una struttura a righe e colonne, dove ogni riga rappresenta un'unica *osservazione* e le colonne rappresentano le *caratteristiche* di tale osservazione.

- *Dati non organizzati*: questo è il tipo di dati in formato libero, normalmente testo o audio grezzo o segnali che devono essere analizzati meglio per poter essere organizzati. Ogni volta che aprite Excel (o un altro programma per la creazione di fogli elettronici), vi viene presentata una struttura a righe e colonne in attesa di dati organizzati. Questi programmi non gestiscono molto bene i dati non organizzati. Nella maggior parte dei casi, abbiamo a che fare con dati organizzati dai quali è più facile trarre conoscenze, ma questo non ci impedisce di esaminare anche il puro testo e i metodi di elaborazione dei dati che si presentano in modo non organizzato.

La scienza dei dati è l'arte e la scienza che consiste nel trarre conoscenza dai dati.
Una piccola definizione per un argomento così grande, ma è giusto che sia così! La scienza dei dati si occupa di talmente tante cose che per elencarle sarebbero necessarie pagine e pagine (ci ho provato, ma poi tali pagine mi sono state cancellate).
La scienza dei dati consiste nel prendere i dati, utilizzarli per acquisire conoscenza e poi usare tale conoscenza per:

- prendere decisioni;
- prevedere il futuro;
- comprendere il passato e il presente;
- creare nuovi servizi o prodotti.

Questo libro tratta i metodi della scienza dei dati, esaminando come si elaborano i dati, come si acquisiscono conoscenze e come si usano queste conoscenze per prendere decisioni informate ed effettuare previsioni.
La scienza dei dati consiste nell'usare i dati per acquisire nuove conoscenze che altrimenti rimarrebbero nascoste nei dati.
Per esempio, immaginate di essere seduti a tavola insieme ad altre tre persone. Uno di voi deve prendere una decisione sulla base di determinati dati. Le opinioni da considerare sono quattro. La scienza dei dati permette di mettere sul tavolo anche una quinta, una sesta e perfino una settima opinione.
Questo è il motivo per cui la scienza dei dati non può sostituire la mente umana, ma affiancarsi ad essa, stare al suo fianco. La scienza dei dati non dovrebbe essere considerata come una soluzione totale alla ricerca di informazioni dai dati; non è altro che un'opinione, un'opinione molto informata, ma pur sempre un'opinione. Diciamo che… merita un posto a quella tavola…

Perché scienza dei dati?

In questa era dei dati, è chiaro che ci troviamo di fronte a un surplus di dati. Ma perché mai dovremmo aver bisogno di un vocabolario interamente nuovo? Che cosa c'è di sbagliato nelle nostre precedenti forme di analisi? Innanzitutto, il semplice volume di dati rende praticamente impossibile per un essere umano analizzarli in un tempo ragionevole. I dati possono essere raccolti in varie forme e da fonti differenti, e spesso si trovano in una forma molto poco organizzata.
I dati possono essere mancanti, incompleti, o semplicemente errati. Spesso, le scale di misurazione sono molto differenti, al punto che è difficile confrontarle. Immaginate di dover mettere in relazione dei dati relativi ai prezzi delle auto usate. Una caratteristica di un'auto può essere l'anno di produzione e un'altra potrebbe essere la quantità di

chilometri percorsi. Dopo aver "ripulito" i dati (argomento che occuperà gran parte di questo libro), le relazioni fra i dati risulteranno più ovvie, e la conoscenza che in precedenza si trovava sepolta sotto milioni di righe di dati finalmente si svela. Uno dei principali obiettivi della scienza dei dati consiste nel creare tecniche e procedure esplicite per scoprire e applicare queste relazioni nei dati.
In precedenza, abbiamo esaminato la scienza dei dati in base a una prospettiva più storica, ma dedichiamo un attimo a discutere qual è il suo ruolo nel mondo di oggi, attraverso un semplicissimo esempio.

Esempio: Sigma Tecnologies

Ben Runkle, CEO di Sigma Tecnologies, sta cercando di risolvere un grosso problema. La società continua a perdere clienti affezionati. Ben non sa perché se ne vanno, ma deve fare qualcosa, e in fretta. È convinto che per ridurre il tasso di abbandono, deve sviluppare nuovi prodotti e funzionalità, e consolidare le attuali tecnologie. Per procedere in sicurezza, convoca il capo dei suoi data scientist, Dr. Jessie Hughan. Tuttavia, questi non è convinto che i nuovi prodotti e servizi, da soli, bastino a salvare la società. Piuttosto, sta pensando alle trascrizioni dei contatti con il servizio clienti. Mostra a Runkle le trascrizioni più recenti e vi trova qualcosa di sorprendente:

- "Non so come si esporta questa cosa; me lo può dire?"
- "Dov'è il pulsante per creare una nuova lista?"
- "Sa per caso dov'è il cursore?"
- "Se oggi non riesco a capire come fare, è un bel problema…"

È chiaro che i clienti hanno dei problemi con l'attuale interfaccia utente, e non si lamentano tanto della carenza di funzionalità. Runkle e Hughan hanno pertanto organizzato un'importante revisione dell'interfaccia utente e le vendite sono andate a gonfie vele.
Naturalmente, l'*elaborazione scientifica* richiesta da quest'ultimo esempio è stata minima, ma aiuta a chiarire le cose. Tendiamo a considerare le persone come Runkle, come a dei dirigenti. Il tipico CEO "sanguigno" vuole prendere tutte le decisioni rapidamente e provare diverse soluzioni fino a trovarne una funzionante. L'approccio di Haghun è molto più analitico. Vuole risolvere il problema almeno quanto lo vuole Runkle, ma per trovare le risposte si rivolge ai dati prodotti dagli utenti più che all'istinto. La scienza dei dati si occupa dell'applicazione delle competenze del ragionamento analitico, per usarle come farebbe un dirigente.
Entrambe queste mentalità hanno la loro collocazione nelle aziende; tuttavia, le idee della scienza dei dati sono governate dal modo di procedere scelto da Hagun: usare come fonte di informazioni i dati generati dalla società, piuttosto che scegliere semplicemente una soluzione e provare a vedere se funziona.

Il diagramma di Venn e la scienza dei dati

Molti pensano ancora che gli aspetti matematici e di programmazione della scienza dei dati siano all'altezza solo di un laureato in matematica o di un genio. Niente di più falso. La conoscenza della scienza dei dati inizia da tre aree molto semplici.

- Matematica e statistica: l'uso di equazioni e formule per svolgere l'analisi.
- Programmazione: la capacità di usare istruzioni per ottenere risultati tramite un computer.
- Conoscenza del dominio: la conoscenza del dominio del problema (medicina, finanza, scienze sociali e così via).

Il seguente diagramma di Venn offre una rappresentazione grafica dell'intersezione fra queste tre aree della scienza dei dati.

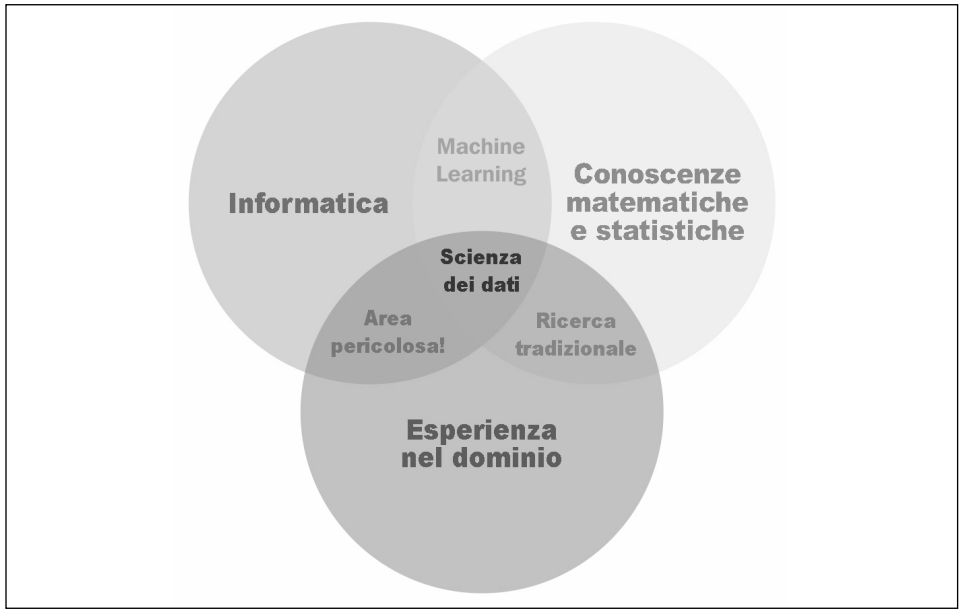


Figura 1.2 Il diagramma di Venn della scienza dei dati.

Coloro che sono più dotati di conoscenze informatiche potranno sviluppare e poi programmare complessi algoritmi usando un linguaggio di programmazione. Eventuali *competenze in matematica e statistica* vi consentiranno di individuare e valutare algoritmi e di adattare una procedura alla vostra specifica situazione. L'esperienza nel *settore dell'attività* (esperienza nel dominio) vi consentirà di applicare concetti e risultati in un modo più appropriato ed efficace.

Il fatto di avere almeno due di queste tre qualità può sicuramente rendervi competenti, ma lascia anche aperto qualche problema. Supponiamo che siate abilissimi programmatori e che abbiate anche una competenza formale nell'ambito finanziario. Potreste creare un sistema automatizzato che vi consenta di vendere o acquistare azioni al posto vostro, ma senza le competenze matematiche per valutare gli algoritmi e, pertanto, per evitare di perdere denaro sul lungo peridodo. È solo potendo contare su competenze in programmazione, matematica e conoscenza del dominio che potete sfruttare con profitto la scienza dei dati.

Quella che potrebbe forse sorprendervi è la *conoscenza del dominio*. Si tratta semplicemente dell'insieme di competenze relative all'area di lavoro. Se un analista finanziario dovesse

iniziare ad analizzare dati relativi agli attacchi cardiaci, avrà quasi certamente bisogno dell'aiuto di un cardiologo per capire il senso di tutti quei numeri.
La scienza dei dati è proprio l'intersezione delle tre aree menzionate in precedenza. Per trarre conoscenze dai dati, dobbiamo essere in grado di utilizzare la programmazione per accedere ai dati, di conoscere la matematica su cui si basano i modelli che deriviamo, e soprattutto, comprendere la collocazione delle nostre analisi nel dominio in cui ci troviamo a operare. Questo comprende la presentazione dei dati. Se stiamo creando un modello per prevedere gli attacchi cardiaci nei pazienti, è meglio creare un PDF di informazioni o un'app dove poter inserire dei numeri e ottenere una rapida previsione? Tutte queste decisioni devono essere prese dal data scientist.

NOTA

Notate anche che l'intersezione fra matematica e programmazione è il machine learning. Questo libro esaminerà il machine learning più in dettaglio più avanti, ma è importante notare che senza un'esplicita capacità di generalizzare i modelli o i risultati applicandoli a un dominio, gli algoritmi di machine learning rimangono quello che sono: algoritmi come altri, presenti sul vostro computer. Potreste avere a disposizione il migliore algoritmo per predire il cancro. Potreste essere in grado di predire il cancro con un'accuratezza del 99 percento sulla base dei dati dei pazienti di cancro, ma se non capite come applicare questo modello nella pratica, in modo che medici e infermieri possano impiegarlo con facilità, il vostro modello potrebbe essere del tutto inutile.

Il libro parla ampiamente di elementi matematici e di programmazione. La conoscenza del dominio deriva sia dall'uso pratico della scienza dei dati sia dalla lettura di esempi di analisi svolte da altri.

Matematica

Molte persone smettono di ascoltare il proprio interlocutore non appena sentono pronunciare la parola "matematica". Poi faranno strane smorfie nel tentativo di celare il loro radicale disprezzo per l'argomento. Questo libro vi guiderà attraverso i concetti matematici impiegati nella scienza dei dati, in particolare la statistica e il calcolo delle probabilità. Utilizzeremo questi sotto-domini della matematica per creare quelli che vengono chiamati modelli.
Un *modello dei dati* fa riferimento a una relazione organizzata e formale fra gli elementi dei dati, che normalmente hanno lo scopo di simulare un fenomeno fisico.
Sostanzialmente, useremo la matematica per formalizzare le relazioni esistenti fra alcune variabili. Come ex matematico puro e attuale insegnante di matematica, so quanto ciò possa essere difficile. Quindi farò del mio meglio per spiegare tutto con la massima chiarezza. Fra le tre aree che compongono la scienza dei dati, la matematica è quella che ci consente di passare da un dominio all'altro. Conoscere la teoria ci consentirà di applicare un modello realizzato per il mondo della moda a un modello adatto all'ambito finanziario. La matematica trattata in questo libro va dalla più semplice algebra alla modellazione probabilistica e statistica avanzata. Non saltate questi capitoli, anche se già conoscete questi argomenti o, al contrario, li temete. Ogni concetto matematico che introduco, sarà accompagnato da spiegazioni, esempi e dalle motivazioni che mi hanno spinto a parlarne. La matematica presentata in questo libro è essenziale per chi vuole operare sulla scienza dei dati.

Esempio: modelli riproduttori/reclute

In biologia, usiamo, insieme a molti altri, un modello detto riproduttori/reclute (*spawners/recruits*) per valutare la salute biologica di una specie. Si tratta di una semplice relazione fra il numero di unità parentali sane di una specie e il numero di nuove unità nel gruppo di animali. In un dataset pubblico del numero di riproduttori e reclute nella popolazione dei salmoni, la relazione fra le due popolazioni è rappresentata dal seguente grafico. Possiamo vedere che esiste assolutamente un qualche tipo di relazione positiva (se cresce uno, cresce anche l'altro). Ma come possiamo formalizzare questa relazione? Per esempio, se conoscessimo il numero di riproduttori in una popolazione, potremmo prevedere il numero di reclute prodotte da tale gruppo, e viceversa?
Sostanzialmente, i modelli ci consentono di specificare una variabile per ottenere l'altra. Considerate il seguente esempio.

$$Recruits = 0.5 * Spawners + 60$$

In questo esempio, supponiamo di sapere che un gruppo di salmoni abbia *1.15* (migliaia) di riproduttori. Allora, avremmo il seguente:

$$Recruits = 0.5 * 1.15 + 60$$

$$Recruits = 60.575$$

Questo risultato (in migliaia di unità) può essere molto utile per stimare come sta cambiando la salute di una popolazione. Se possiamo creare questi modelli, possiamo osservare in modo visuale come la relazione fra le due variabili possa cambiare.
Vi sono molti tipi di modelli per i dati, inclusi i modelli probabilistici e statistici. Entrambi sono sottoinsiemi di un paradigma più ampio, chiamato *machine learning*, apprendimento automatico. L'idea che sta alla base di questi tre argomenti è il fatto che usiamo i dati per ottenere il miglior modello possibile. Non contiamo più sull'istinto, ma sui dati.
Lo scopo di questo esempio consiste nel dimostrare come possiamo definire le relazioni fra gli elementi dei dati usando equazioni matematiche. Il fatto che qui siano stati usati dati relativi alla salute della popolazione di salmoni era del tutto irrilevante! In tutto questo libro, esamineremo relazioni riguardanti investimenti di marketing, dati sul sentiment, recensioni di ristoranti e molto altro ancora. Il motivo principale di tale varietà è che vorrei che voi (i lettori) vi sentiste esposti a quanti più domini possibile.
La matematica e la programmazione sono come degli strumenti che consentono ai data scientist di analizzare e applicare le proprie competenze praticamente ovunque.

Programmazione

Siamo onesti. Probabilmente siete convinti che i computer siano più interessanti della matematica. Ok, vi capisco. Nei notiziari è difficile trovare notizie riguardanti le nuove frontiere della matematica mentre abbondano quelle sulle nuove tecnologie.

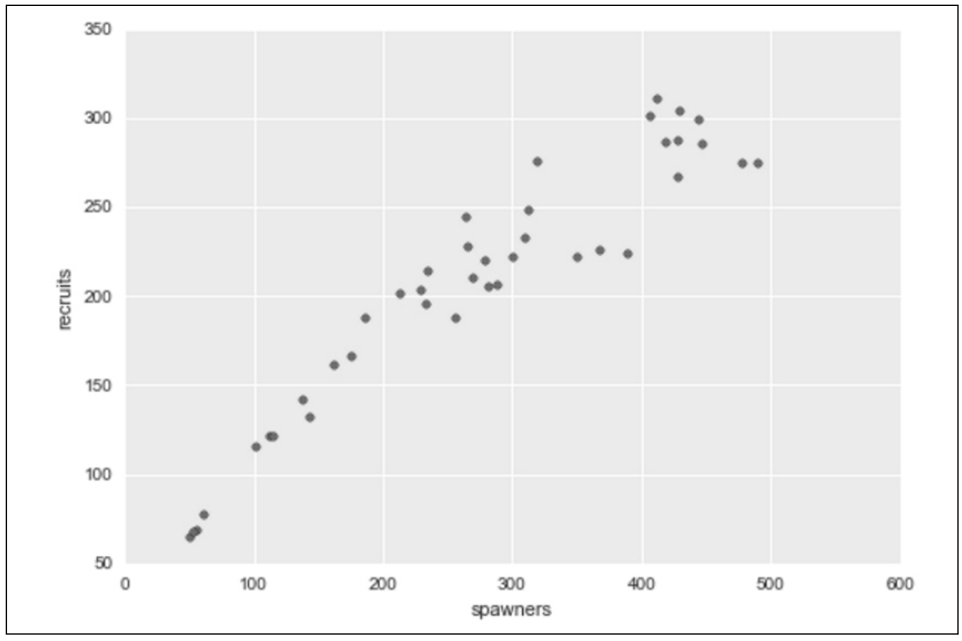


Figura 1.3 Rappresentazione del modello riproduttori-reclute.

Non vi fiondereste davanti alla TV per sentir parlare di una recentissima teoria sui numeri primi, ma magari per un servizio sul nuovo smartphone e di come sia eccezionale per scattare foto favolose ai gattini di casa. I linguaggi di programmazione sono il mezzo con cui comunichiamo con la macchina e le chiediamo di svolgere un determinato lavoro. Un computer può parlare più linguaggi, così come un libro può essere scritto in più lingue; analogamente, la scienza dei dati può essere svolta in più linguaggi. Python, *Julia* e *R* sono solo alcuni dei tanti linguaggi disponibili. Questo libro però si concentrerà esclusivamente sull'uso di Python.

Perché Python?

Utilizzeremo Python per svariate ragioni.

- Python è un linguaggio estremamente semplice da leggere e scrivere, anche per chi non ha mai programmato, il che renderà più comprensibili i prossimi esempi sia durante la lettura, sia successivamente, dopo che avrete finito di leggere questo libro.
- Si tratta di uno dei linguaggi più largamente utilizzati, in ambiente produttivo e accademico (e uno dei linguaggi a diffusione più rapida, come è evidente a tutti).
- La comunità online del linguaggio è vasta e amichevole. Ciò significa che una rapida ricerca su Google fornirà molti risultati relativi a persone che hanno affrontato e risolto una situazione analoga (a volte esattamente la stessa).
- Python offre moduli pronti all'uso di scienza dei dati, utilizzabili da chi è alle prime armi come dal veterano dell'argomento.

Quest'ultima è probabilmente la ragione principale per cui ci concentreremo su Python. Questi moduli precompilati sono non solo potenti, ma anche facili da usare. Entro la fine dei primi capitoli, vi ritroverete a usarli con grande facilità. Alcuni di questi moduli sono:

- `pandas`
- `sci-kit learn`
- `seaborn`
- `numpy/scipy`
- `requests` (per estrarre dati dal Web)
- `BeautifulSoup` (per il parsing Web-HTML)

Le basi di Python

Prima di procedere, è importante formalizzare molti dei requisiti in termini di competenze di programmazione in Python.
In Python, le *variabili* rappresentano oggetti. Inizialmente ci concentreremo solo alcuni semplici tipi di oggetti.

- `int` (un numero intero).
  - Esempi: `3`, `6`, `99`, `-34`, `34`, `11111111`.
- `float` (un numero decimale). Attenzione: come separatore si usa il punto decimale.
  - Esempi: `3.14159`, `2.71`, `-0.34567`.
- `boolean` (`True` o `False`).
  - L'affermazione "Domenica cade nel fine settimana" è vera, `True`.
  - L'affermazione "Martedì cade nel fine settimana" è falsa, `False`.
  - L'affermazione "pi è esattamente il rapporto fra la circonferenza e il diametro di un cerchio" è vera, `True`.
- `string` (testo o parole composte da caratteri).
  - `"Mi piacciono gli hamburger"`.
  - `"Matt è un grande"`.
  - Un tweet è una stringa.
- `list` (una collezione di oggetti).
  - Esempio: `[1, 5.4, True, "mela"]`.

Dobbiamo inoltre conoscere l'uso degli operatori logici. Per questi operatori, tenete sempre in considerazione il tipo di dati `boolean`. Ogni operatore restituirà un risultato `True` o `False`. Osserviamo i seguenti esempi.

- `==` restituisce il valore `True` se le espressioni da entrambi i lati sono uguali; altrimenti restituisce il valore `False`.
  - `3 + 4 == 7` (restituisce `True`)
  - `3 - 2 == 7` (restituisce `False`)
- `<` (minore di).
  - `3 < 5` (`True`)
  - `5 < 3` (`False`)

- <= (minore o uguale a).
  - 3 <= 3 (True)
  - 5 <= 3 (False)
- > (maggiore di).
  - 3 > 5 (False)
  - 5 > 3 (True)
- >= (maggiore o uguale a).
  - 3 >= 3 (True)
  - 5 >= 3 (False)

Programmando in Python, si usa il segno di cancelletto (#) per creare un "commento", che non verrà elaborato come codice ma ha il solo scopo di comunicare con il lettore. Tutto ciò che si trova a destra del segno # è un commento del codice eseguito.

Un semplice esempio in Python

In Python, usiamo spazi e tabulazioni per indicare le operazioni che appartengono ad altre righe di codice.

NOTA

Notate l'uso dell'istruzione `if`: quando l'istruzione dopo l'`if` è `True`, viene eseguita la parte rientrata, come si può vedere dal seguente codice:

```
X = 5.8
Y = 9.5

X + Y == 15.3          # Questo è vero, True

X - Y == 15.3          # Questo è falso, False
if x + y == 15.3:      # Se l'espressione è vera:
    print "True!"      # stampa qualcosa
```

L'istruzione `print "True!"` appartiene alla riga `if x + y == 15.3:` che la precede, e infatti è rientrata sotto di essa. Questo significa che l'istruzione `print` verrà eseguita se e solo se `x + y` è uguale a `15.3`.

Notate che la seguente variabile `list`, `my_list`, può contenere più tipi di oggetti. Questa ha un `int`, un `float`, un `boolean` e una `string` (in questo ordine):

```
my_list = [1, 5.7, True, "mele"]

len(my_list) == 4    # quattro oggetti nella lista

my_list[0] == 1      # il primo oggetto

my_list[1] == 5.7    # il secondo oggetto
```

Da notare nel codice precedente.

- Il comando `len` restituisce la lunghezza della lista (ovvero 4).

- In Python gli indici partono da `0`: la maggior parte dei linguaggi di programmazione inizia a contare da zero e non da uno. Quindi se avete bisogno del primo elemento, il suo indice sarà `0` e se cercate il novantacinquesimo, il suo indice sarà `94`.

Esempio: parsing di un tweet

Ecco ora dell'altro codice Python. In questo esempio, occorre analizzare dei tweet relativi alle quotazioni delle azioni (uno dei più importanti casi di studio di questo libro cercherà di prevedere i movimenti di mercato sulla base del sentiment che circola nei social media).

```
tweet = "RT @j_o_n_dnger: $TWTR now top holding for Andor, unseating $AAPL"

words_in_tweet = tweet.split(' ')           # lista delle parole del tweet

for word in words_in_tweet:                 # per ogni parola della lista
    if "$" in word:                         # se la parola ha un "cashtag"
        print "THIS TWEET IS ABOUT", word   # avverti l'utente
```

Occorre, riga per riga, dire alcune cose su questo frammento di codice.

- Definiamo una variabile che conterrà del testo (in Python si parla di stringa). In questo esempio, il tweet in questione è `"rt @robdv: $twtr now top holding for Andor, unseating $AAPL"`.
- La variabile `words_in_tweet` trasforma il tweet in token (ne separa le parole). Se doveste stampare questa variabile, il risultato sarebbe il seguente:

```
['RT',
'@robdv:',
'$TWTR',
'now',
'top',
'holding',
'for',
'Andor,',
'unseating',
'$AAPL']
```

- Eseguiamo un'iterazione su questa lista di parole. Questo è il significato del ciclo `for`. Significa che esaminiamo le singole parole, una per una.
- Ecco un'altra istruzione `if`. Per ogni parola di questo tweet, se la parola contiene il carattere `$` (il modo in cui si fa riferimento alle sigle delle azioni in Twitter).
- Se l'istruzione `if` precedente è vera (ovvero, se il tweet contiene un cashtag), stampalo e mostralo all'utente.

L'output di questo codice sarà il seguente:

```
THIS TWEET IS ABOUT $TWTR
THIS TWEET IS ABOUT $AAPL
```

Otteniamo questo output poiché queste sono le uniche parole del tweet che hanno un cashtag. Ogni volta che verrà usato Python in questo libro, cercheremo di essere il più possibile espliciti sul significato di ciascuna riga di codice.

Conoscenza del dominio

Come abbiamo visto in precedenza, questa categoria si concentra principalmente sulla conoscenza dell'argomento trattato. Per esempio, se siete un analista finanziario che lavora sui dati del mercato delle azioni, avrete già una grande conoscenza del dominio. Se siete un giornalista che cerca informazioni sui tassi di adozione nel mondo, vorrete consultare un esperto di questo campo. Questo libro cercherà di mostrare degli esempi tratti da vari domini di problemi, fra cui la medicina, il marketing, la finanza e perfino gli avvistamenti di UFO!

Questo significa che se non siete medici, non potete lavorare su dati medici? Naturalmente no! I migliori esperti di scienza dei dati possono applicare le loro competenze a qualsiasi area, anche se non la conoscono approfonditamente. Gli esperti di scienza dei dati sono in grado di adattarsi al campo di lavoro e contribuire in modo appropriato una volta terminata la loro analisi.

Un elemento fondamentale della conoscenza del dominio è la presentazione. A seconda della vostra audience, può essere molto importante presentare i risultati delle vostre ricerche. L'intero valore dei vostri risultati dipende dall'efficacia della vostra comunicazione. Potete prevedere il movimento del mercato con un'accuratezza del 99.99 percento, ma se il vostro programma è difficile da far funzionare, i vostri risultati saranno del tutto inutilizzati. Analogamente, anche se la vostra capacità di comunicare i risultati è inappropriata per il campo studiato, i vostri risultati rimarranno inutilizzati.

Ancora un po' di terminologia

Sembra il momento opportuno per definire un altro po' di termini. A questo punto del libro, sarete probabilmente ansiosi di sentir parlare di scienza dei dati e di conoscere nuove parole e frasi. Ecco la terminologia di base che incontrerete.

- *Machine learning*: fa riferimento al fatto di concedere ai computer la capacità di apprendere dai dati senza "regole" esplicite fornite da un programmatore. Abbiamo già parlato in precedenza, sempre in questo capitolo, del concetto di machine learning in termini di una somma di competenze di programmazione e matematiche. Qui stiamo cercando di formalizzare questa definizione. Il machine learning combina la potenza dei computer con gli algoritmi di apprendimento intelligenti per automatizzare l'individuazione di relazioni nei dati e per creare potenti modelli per i dati. Già che parliamo di modelli per i dati, ci occuperemo dei seguenti due tipi di modelli.
- *Modello probabilistico*: fa riferimento all'impiego del calcolo delle probabilità per trovare una relazione fra elementi che includa un certo grado di casualità.
- *Modello statistico*: fa riferimento al fatto di sfruttare i teoremi tratti dalla statistica per formalizzare le relazioni esistenti fra gli elementi dei dati in una (normalmente) semplice formula matematica.

Sebbene i modelli statistici e probabilistici possano essere applicati ai computer e, pertanto, possano essere considerati machine learning, manterremo separate queste definizioni, in quanto gli algoritmi di machine learning, in genere, cercano di individuare le relazioni in altri modi.

Parleremo di modelli statistici e probabilistici nei prossimi capitoli.

- L'*analisi esplorativa dei dati* (EDA – *Exploratory Data Analysis*) fa riferimento alla preparazione dei dati per standardizzare i risultati e acquisire conoscenze in modo rapido. L'EDA si occupa di visualizzazione e preparazione dei dati. Si tratta dell'attività che trasforma i dati grezzi in dati organizzati e che si occupa di correggere quei punti dei dati che risultano essere mancanti o errati. Durante lo svolgimento dell'analisi EDA, creeremo molti tipi di grafici e poi useremo tali grafici per identificare le caratteristiche e le relazioni chiave da sfruttare poi nei nostri modelli per i dati.
- Il *data mining* è il processo di ricerca delle relazioni esistenti fra gli elementi dei dati. Il data mining è quella parte della scienza dei dati nella quale cerchiamo di trovare relazioni fra le variabili (pensate al modello riproduttori-reclute).
- Ho cercato a fatica di non usare, finora, il termine *big data*. Questo perché sono convinto che questo termine sia molto abusato. E poi perché questa definizione varia da persona a persona: "grandi dati". I *big data* sono quei dati che sono troppo estesi per poter essere elaborati da un'unica macchina (se vi si è bloccato il portatile, forse è perché ha sofferto di una forma di *big data*).

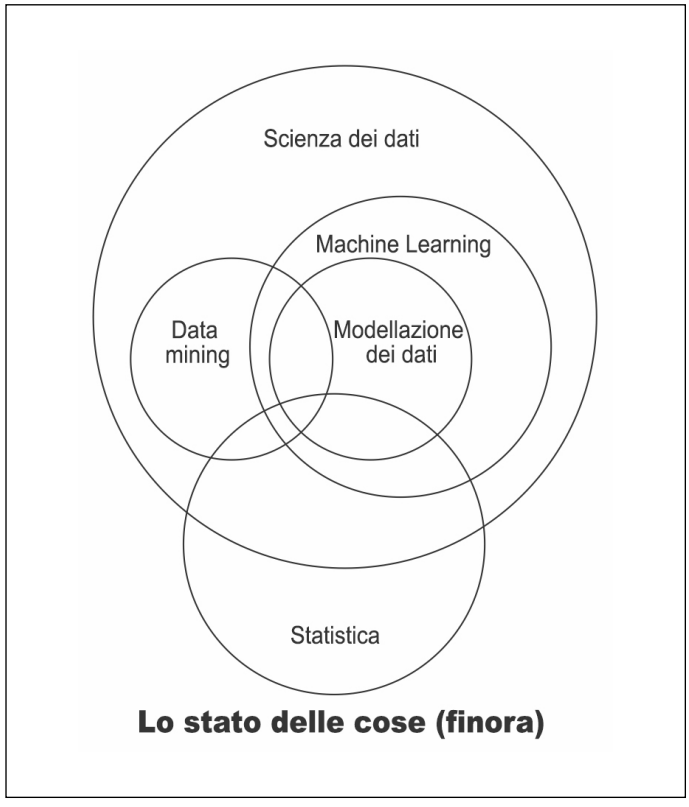


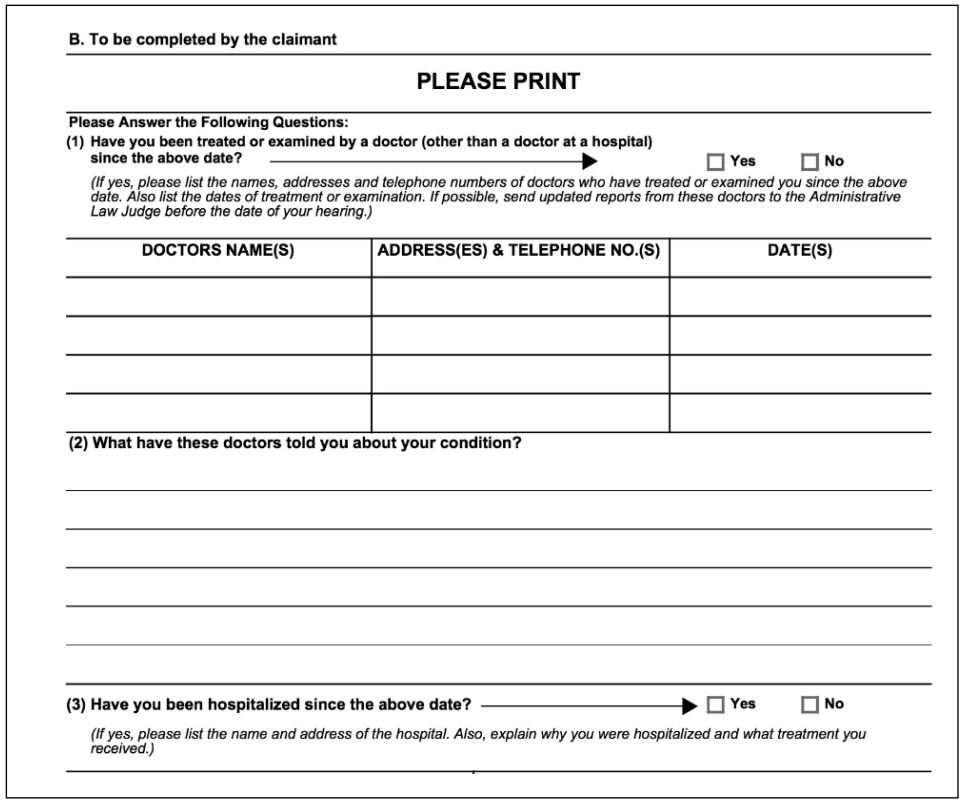
Figura 1.4 Lo stato della scienza dei dati (finora).
Questo schema è incompleto e ha solo scopi rappresentativi.

Scienza dei dati: casi di studio

La combinazione di matematica, programmazione e conoscenza del dominio è ciò che rende così potente la scienza dei dati. Spesso è difficile, per una persona sola, eccellere in tutte e tre queste aree. Questo perché in genere le aziende assumono interi team di esperti di scienza dei dati, invece di un'unica persona. Esaminiamo alcuni ottimi esempi di scienza dei dati in azione, e i loro risultati.

Caso di studio: automatizzare il flusso dei documenti

La modulistica statale è davvero ostica, sia per l'impiegato che deve leggerla, sia per la persona che deve compilarla [NdT: l'esempio è in inglese, ma sono sicuro che il lettore converrà sul fatto che la modulistica italiana presenti lo stesso problema]. Alcuni moduli richiedono anche un paio d'anni prima di essere recepiti; non è un po' assurdo? Vediamo un esempio.



B. To be completed by the claimant

PLEASE PRINT

Please Answer the Following Questions:

(1) Have you been treated or examined by a doctor (other than a doctor at a hospital) since the above date? ☐ Yes ☐ No

(If yes, please list the names, addresses and telephone numbers of doctors who have treated or examined you since the above date. Also list the dates of treatment or examination. If possible, send updated reports from these doctors to the Administrative Law Judge before the date of your hearing.)

DOCTORS NAME(S)	ADDRESS(ES) & TELEPHONE NO.(S)	DATE(S)

(2) What have these doctors told you about your condition?

(3) Have you been hospitalized since the above date? ☐ Yes ☐ No

(If yes, please list the name and address of the hospital. Also, explain why you were hospitalized and what treatment you received.)

Figura 1.5 Semplice modulo sanitario (in inglese).

C'è di peggio… Tutto sommato si tratta solo di testo. Compilare questo campo, poi quello, poi quell'altro e così via. Potete immaginare quanto invece possa essere difficile

per un impiegato leggere questi moduli tutto il giorno, modulo dopo modulo dopo modulo. Ci deve essere un modo migliore!
In effetti c'è. Elder Research Inc. è in grado di leggere questi dati non organizzati e di automatizzare il 20 percento di tutti i moduli nell'ambito della disabilità. Questo significa che un computer è in grado di leggere il 20 percento di questi moduli scritti a mano e dare la propria opinione sull'approvazione.
Non solo: la società esterna incaricata di valutare le approvazioni dei moduli, ha dato ai moduli approvati automaticamente un voto più alto rispetto a quelli approvati da una persona. Pertanto, non solo il computer ha gestito il 20 percento del carico, ma, in media, si è comportato meglio di un essere umano.

Licenziare tutti?

Prima che finisca per ricevere tonnellate di e-mail di persone infuriate, convinte che la scienza dei dati porterà alla fine del lavoro svolto da persone in carne e ossa, tenete in considerazione il fatto che il computer è stato in grado di gestire solo il 20 percento del carico. Questo significa che probabilmente si è comportato malissimo sul restante 80 percento dei moduli! Questo perché il computer è stato probabilmente efficacissimo nei *moduli più semplici*. I moduli che un essere umano ha elaborato in qualche minuto, avrebbero richiesto solo pochi secondi al computer. Ma questi minuti si sommano e, poco a poco, ogni essere umano si trova a risparmiare un'ora al giorno!
I moduli che un essere umano trova semplici da interpretare, sono probabilmente altrettanto semplici per il computer. È proprio quando il testo del modulo è molto conciso o quando l'autore ha deviato abbondantemente dalla grammatica corretta che il computer inizia ad avere problemi. Questo modello è notevole, perché consente agli esseri umani di dedicare più tempo ai moduli di difficile interpretazione, sui quali possono concentrarsi con maggiore attenzione senza essere distratti dalla mole di documenti da valutare.

NOTA

Notate che ho usato la parola *modello*. Come ricorderete, un modello è una relazione fra elementi. In questo caso, la relazione è fra le parole scritte e l'approvazione della richiesta.

Caso di studio: spese di marketing

Un dataset mostra la relazione esistente fra il denaro speso per attività di marketing in TV, radio e riviste. L'obiettivo è quello di analizzare la relazione esistente fra i tre media e scoprire come essi influenzano la vendita di un prodotto. I dati disponibili sono in una struttura a righe e colonne. Ogni riga rappresenta una regione commerciale e le colonne ci dicono quanto denaro è stato speso su ogni medium e il profitto ottenuto in tale regione.

NOTA

Normalmente, gli esperti di scienza dei dati domandano quali sono state le unità e la scala impiegate. In questo caso, le voci TV, Radio e Riviste sono misurate in "migliaia di dollari" e le vendite in "migliaia di pezzi venduti". Questo significa che nella prima regione sono stati spesi 230 100 dollari in pubblicità alla TV, 37 800 dollari in pubblicità alla radio e 69 200 dollari in pubblicità sulle riviste. In quella stessa regione sono stati venduti 22 100 pezzi.

Tabella 1.1 Budget pubblicitario.

	TV	Radio	Riviste	Vendite
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

Per esempio, nella terza regione, abbiamo speso 17 200 dollari il pubblicità televisiva e abbiamo venduto 9 300 articoli.
Se tracciamo il grafico di ogni variabile rispetto alle vendite, otteniamo i seguenti risultati:

```
import seaborn as sns
sns.pairplot(dati, x_vars=['TV','Radio','Newspaper'], y_vars='Sales')
```

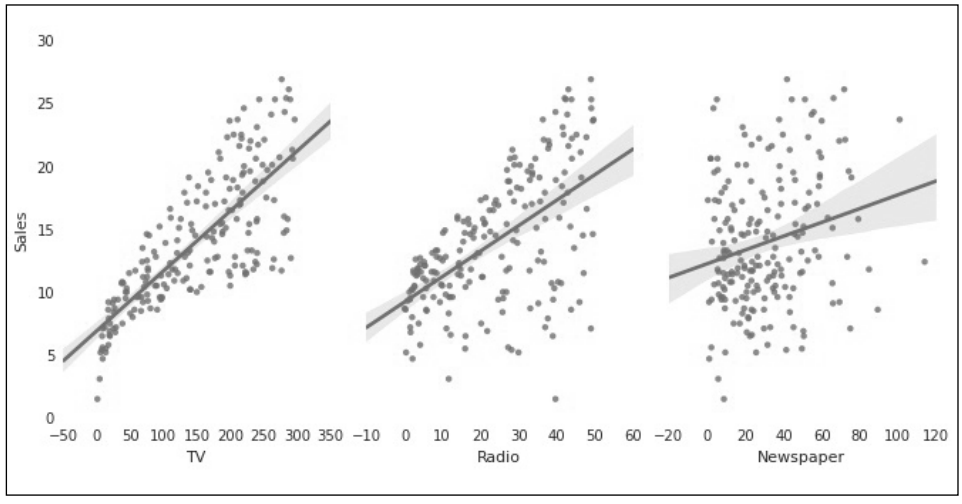


Figura 1.6 Grafici delle spese pubblicitarie.

Notate come nessuna di queste variabili produca davvero una linea netta e, pertanto, da sole potrebbero non essere sufficienti per prevedere le vendite. I dati relativi alla TV sembrano indicare una relazione, ma comunque non si tratta di un gran risultato. In questo caso, per modellare le vendite dobbiamo trovare un modello più complesso rispetto a quello che abbiamo usato nel modello riproduttori-reclute e combinare tutte e tre le variabili.
Questo tipo di problema è molto comune nella scienza dei dati. In questo esempio, stiamo cercando di identificare le caratteristiche chiave associate alle vendite di un prodotto. Se riusciamo a isolare queste caratteristiche chiave, allora possiamo sfruttare queste relazioni e adattare il nostro modello di spesa pubblicitaria nelle varie aree con la speranza di incrementare le vendite.

Caso di studio: i contenuti di un'offerta di lavoro

Cercate un lavoro nell'ambito della scienza dei dati? Ottimo, lasciate che vi aiuti. In questo Caso di studio, ho tratto dal Web mille offerte di lavoro di società che assumono attivamente esperti di scienza dei dati (nel gennaio 2016). L'obiettivo consiste nell'individuare alcune delle principali parole chiave usate nelle offerte di lavoro.

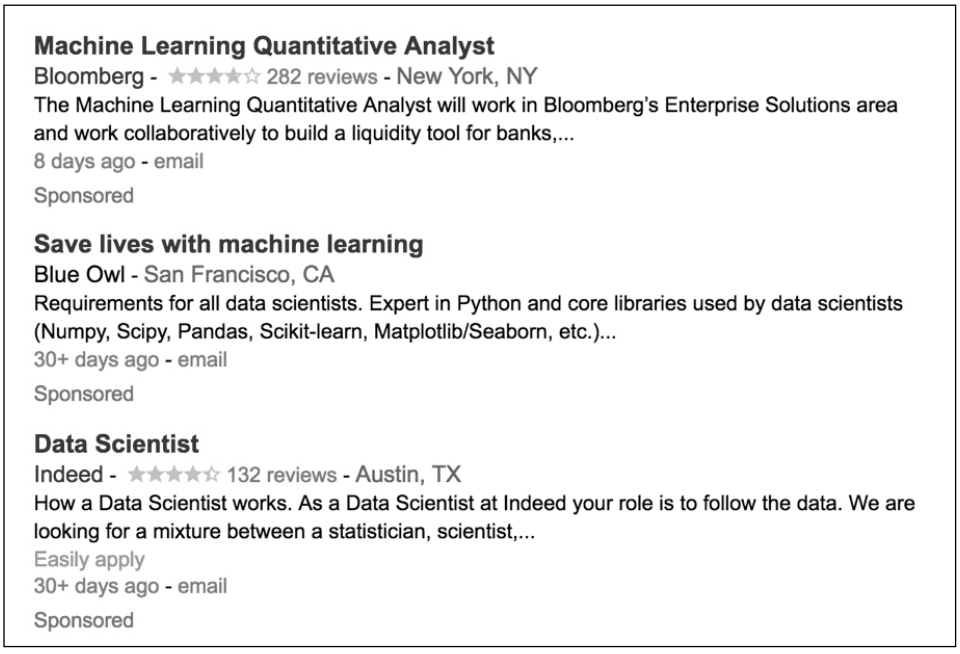


Figura 1.7 Un esempio di ricerche di lavoro nell'ambito della scienza dei dati. (Notate che il secondo annuncio chiede competenze nelle librerie core di Python; ne parleremo più avanti.)

```
import requests
# estrazione di dati dal Web

from BeautifulSoup import BeautifulSoup
# usato per il parsing HTML

from sklearn.feature_extraction.text import CountVectorizer
# usato per contare il numero di parole e frasi (utilizzeremo molto questo modulo)
```

Le prime due importazioni servono per scaricare i dati dal sito web (`Indeed.com`); la terza importazione conta semplicemente il numero di volte che si presenta una parola o frase.

```
texts = []
# contiene le offerte di lavoro in questa lista

for index in range(0,1000,10):    # le prime 100 pagine di indeed
    page = 'indeed.com/jobs?q=data+scientist&start='+str(index)
```

```
    # identifica l'url dell'elenco di offerte

    web_result = requests.get(page).text
    # usa le richieste per visitare l'url

    soup BeautifulSoup(web_result)
    # parsing dell'html della pagina risultante

    for listing in soup.findAll('span', {'class':'summary'}:
    # per ogni offerta nella pagina

        texts.append(listing.text)
        # aggiungi il testo dell'offerta alla nosta lista
```

Prima di perdervi, tutto quello che fa questo ciclo è attraversare cento pagine di offerte di lavoro e, per ogni pagina, estrarre ogni offerta di lavoro. La variabile più importante qui è `texts`, che è una lista di oltre mille offerte di lavoro:

```
type(texts) # == lista

vect = CountVectorizer(ngram_range=(1,2), stop_words='english')
# Conta le frasi di una o due parole

matrix = vect.fit_transform(texts)
# impara il vocabolario dei testi

print len(vect.get_feature_names()) # quante caratteristiche esistono
# In questo caso ci sono 11293 frasi totali di una o due parole
```

Qui ho omesso del codice, che però si trova nel repository GitHub di questo libro. I risultati sono i seguenti (rappresentati come la "frase", seguita dal numero di occorrenze):

```
experience 320
machine 306
learning 305
machine learning 294
techniques 266
statistical 215
team 197
analytics 173
business 167
statistics 159
algorithms 152
datamining 149
software 144
applied 141
programming 132
understanding 127
world 127
research 125
datascience 123
```

```
methods 122
join 122
quantitative 122
group 121
real 120
large 120
```

Notiamo alcuni aspetti degni di nota.

- Le espressioni "machine learning" ed "experience" sono in cima alla lista. L'esperienza viene con la pratica e con questo libro potrete iniziare a capire cos'è il machine learning.
- Queste parole sono tallonate dalle parole che richiedono competenze statistiche, le quali implicano conoscenze matematiche e teoriche.
- La parola "team" è molto alta, perché dovrete lavorare in un team di esperti di scienza dei dati; non è un lavoro da "lupi solitari".
- Le parole tipiche dell'informatica, come "algorithms" e "programming" sono decisamente prevalenti.
- Le parole "tecniques", "understanding" e "methods" implicano un approccio più teorico, un'ambivalenza di ogni dominio.
- La parola "business" implica uno specifico dominio del problema.

Vi sono molte cose interessanti da notare in questo caso di studio, ma il fatto più eclatante è la presenza di molte parole e frasi chiave che rimarcano il ruolo della scienza dei dati. Non si tratta solo di matematica, programmazione o conoscenza del dominio; è davvero la combinazione di queste tre idee (riunite in un'unica persona o diffuse su un team) a rendere possibile (e così potente) la scienza dei dati.

Riepilogo

All'inizio di questo capitolo, ho posto una semplice domanda, che cosa si intende per scienza dei dati? Ecco la risposta. La scienza dei dati è divertente e si occupa di giochi e modellazione. Ci deve necessariamente essere un prezzo da pagare nella nostra ricerca di macchine e algoritmi sempre più potenti. Mentre cerchiamo metodi innovativi per scoprire le tendenze nei dati, vi è una bestia che si annida nell'ombra. Non sto parlando della curva di apprendimento della matematica o della programmazione, né sto facendo riferimento all'eccesso di dati. L'era industriale ci ha lasciato nel bel mezzo di una vera e propria battaglia contro l'inquinamento. La successiva era dell'informazione ci ha lasciato un'enorme mole di dati. Quali pericoli, dunque, reca in sé l'era dei dati?
L'era dei dati può condurci verso qualcosa di molto più sinistro – la disumanizzazione dell'individuo, attraverso masse di dati.
Sempre più persone si gettano a capofitto nel campo della scienza dei dati, molti dei quali senza una particolare esperienza teorica nel campo della matematica o dell'informatica; questo può anche meravigliare. Un qualsiasi esperto di scienza dei dati ha accesso ai dati di milioni di profili, tweet, recensioni online e molto altro ancora.
Tuttavia, entrando nel mondo della scienza dei dati senza adeguate conoscenze teoriche e competenze di programmazione e senza alcun rispetto per il dominio sul quale si

lavora, si corre il rischio di semplificare eccessivamente il fenomeno che si sta cercando di modellare.
Per esempio, supponiamo che vogliate automatizzare la rete di vendita realizzando un semplice programma che cerchi in LinkedIn alcune specifiche parole chiave nel profilo di una persona.

```
keywords = ["Saas", "Sales", "Enterprise"]
```

Ottimo, ora potete scorrere rapidamente LinkedIn per trovare le persone corrispondenti ai vostri criteri. Ma cosa accade se tale persona ha specificato "Software as a Service" invece di "Saas" o ha scritto in modo errato la parola "enterprise"? Come potrà, il vostro modello, gestire queste potenziali buone corrispondenze? Non possono essere trascurati solo perché un cosiddetto "esperto" di scienza dei dati ha commesso un banalissimo errore di eccesso di generalizzazione.
Il programmatore ha deciso di semplificare la ricerca di un'altra figura, cercando tre semplici parole chiave e in tal modo ha lasciato perdere una grande parte di opportunità.
Nel prossimo capitolo, esploreremo i vari tipi di dati che esistono al mondo, dal testo scritto a mano a file ben strutturati in righe e colonne. Esamineremo le operazioni matematiche che possono essere svolte sui vari tipi di dati e dedurremo conoscenze sulla base della forma in cui giungono i dati.