Capitolo 1

Che cosa si intende per scienza dei dati?

È impossibile misurare l'imponderabile.
Kozma Prutkov, autore russo

Sono sicuro che abbiate già almeno un'idea di cosa si intenda con *scienza dei dati*, ma vale la pena di iniziare da qui! La scienza dei dati è la disciplina che consiste nell'estrarre conoscenze dai dati. È un mix di informatica (per le strutture di dati, gli algoritmi, la rappresentazione, la capacità di elaborare grandi quantità di dati e programmazione), statistica (per le regressioni e l'inferenza) e conoscenza del dominio (perché dobbiamo sapere quali domande porre e come interpretare i risultati).
La scienza dei dati, tradizionalmente, si occupa di vari argomenti dissimili, alcuni dei quali potreste forse già conoscere, mentre altri magari li incontrerete per la prima volta in queste pagine.

- *Database*, che forniscono le funzionalità di memorizzazione e integrazione delle informazioni. Troverete informazioni sui database relazionali e sui sistemi di archiviazione dei documenti nel Capitolo 4, *Utilizzare i database*.
- *Analisi del testo* ed *elaborazione del linguaggio naturale*, che ci consentono di eseguire "calcoli sulle parole" trasformando un testo qualitativo in variabili quantitative. Siete interessati agli strumenti per l'analisi del sentiment? Ne parleremo nell'Unità 16, *Elaborazione di testi in linguaggio naturale*.
- *Analisi numerica dei dati* e *data mining*, alla ricerca di schemi coerenti e di relazioni fra le variabili.

In questo capitolo

Questi sono gli argomenti del Capitolo 5, *Usare i dati numerici tabulari,* e del Capitolo 6, *Manipolare serie di dati e frame*.

- *Analisi delle reti complesse*, un argomento tutt'altro che complesso. Riguarda le reti complesse, ovvero collezioni di entità arbitrarie interconnesse. Il Capitolo 7, *Utilizzo dei dati delle reti*, semplifica l'analisi delle reti complesse.
- *Rappresentazione dei dati*, che non deve essere solo gradevole, ma anche utile, in particolare quando occorre convincere gli sponsor a proseguire nell'impegno. Se un'immagine vale più di mille parole, allora il Capitolo 8, *Rappresentazione grafica*, vale da solo il resto del libro.
- *Machine learning* (comprendendo il clustering, gli alberi decisionali, la classificazione e le reti neurali), che tenta di far "pensare" i computer, in modo che possano effettuare previsioni sulla base dei dati del campione. Il Capitolo 10, *Machine Learning*, spiega come fare.
- *Elaborazione delle serie temporali* e, più in generale, *elaborazione dei segnali digitali*, strumenti indispensabili per gli analisti del mercato di borsa, per gli economisti e per i ricercatori nei domini audio e video.
- *Analisi di grandi quantità di dati*; normalmente si fa riferimento all'analisi di una massa di dati non strutturati (testo, audio, video) superiore a 1 TB, prodotti e catturati ad una frequenza elevata. Si tratta di un argomento davvero troppo "grande" per poter essere trattato in questo libro.

Indipendentemente dal tipo di analisi, la scienza dei dati è prima di tutto scienza e solo in seconda battuta "stregoneria". Pertanto, è un processo che segue una sequenza di base piuttosto rigorosa, che inizia con l'acquisizione dei dati e termina con un report sui risultati.

In questo capitolo, esamineremo i processi di base della scienza dei dati: i passi di un tipico studio di analisi dei dati, dove acquisire i dati e la struttura tipica del report di un progetto.

Unità 1
La tipica sequenza di analisi dei dati

I passi di un tipico studio di analisi dei dati, in genere sono coerenti con la sequenza prevista per qualsiasi altro ambito scientifico.
Una scoperta, in scienza dei dati, inizia con una domanda cui trovare una risposta e con il tipo di analisi da applicare. Il tipo di analisi più semplice è quello *descrittivo*, in cui il dataset viene descritto facendo riferimento alle sue misure aggregate, spesso in una forma visuale. Indipendentemente da come si procederà, bisogna almeno descrivere i dati! Durante l'analisi *esplorativa* dei dati, si cerca di trovare nuove relazioni fra le variabili disponibili. Se abbiamo a disposizione solo pochi dati campione e vogliamo usarli per descrivere una popolazione di maggiori dimensioni, l'analisi *inferenziale* statistica è lo strumento perfetto. Un analista *predittivo* impara dal passato per prevedere il futuro. L'analisi *causale* identifica le variabili che influenzano le altre. Infine, l'analisi dei dati *meccanicistica* esplora esattamente il modo in cui una variabile influenza un'altra variabile.
Tuttavia, la qualità della vostra analisi dipende dalla qualità dei dati utilizzati. Qual è l'aspetto di un dataset ideale? Quali dati conterrebbero la risposta alla vostra domanda in un mondo ideale? A proposito, il dataset ideale potrebbe non esistere affatto o essere difficile, se non impossibile, da ottenere. A seconda delle situazioni, magari un dataset più piccolo o meno ricco di caratteristiche potrebbe funzionare altrettanto bene.
Fortunatamente, trarre dati grezzi dal Web o da un database non è poi così difficile, e Python offre moltissimi strumenti che assistono nel download e nella decifrazione dei dati. Ne riparleremo nell'Unità 2, *Sequenza di acquisizione dei dati*.
In questo mondo imperfetto, i dati perfetti non esistono. Nei dati vi possono essere valori mancanti, valori anomali e altri elementi "non standard". Fra gli esempi di dati "sporchi" vi possono essere date di nascita nel futuro, valori negativi di età e peso, indirizzi email inutilizzabili (`noreply@`). Una volta ottenuti i dati grezzi, il passo successivo è usare degli strumenti di "pulizia" dei dati e le vostre conoscenze statistiche per regolarizzare il dataset. Una volta che nei file abbiamo a disposizione i dati "puliti", potete utilizzarli per svolgere analisi descrittive ed esplorative. L'output di questo passo spesso include dei grafici a dispersione (Unità 44), istogrammi e riepiloghi statistici (Unità 46). Questi strumenti vi daranno un'idea del significato dei dati, un'intuizione che è indispensabile per le successive ricerche, in particolare se il dataset ha molte dimensioni.
E ora siete solo a un passo dall'iniziare a prevedere il futuro. I vostri strumenti operativi sono i modelli dei dati che, se adeguatamente addestrati, sono in grado di apprendere dal passato per prevedere il futuro. Senza mai dimenticare di misurare la qualità dei modelli realizzati e l'accuratezza delle loro previsioni!
A questo punto dovete togliervi dal capo il cappellino dell'esperto di statistica e di programmatore e indossare quello di esperto del dominio. Avete a disposizione determinati risultati, ma quale significato hanno nel dominio? In altre parole, tali risultati interessano a qualcuno e sono in grado di cambiare le cose? Immaginate di essere un revisore, incaricato di valutare il vostro stesso lavoro. Dove avete ragione, dove avete sbagliato e dove potreste ottenere risultati migliori o differenti, se aveste un'altra *chance*? Dovreste usare altri tipi di dati, svolgere altri tipi di analisi, porre domande differenti o costruire

un modello differente? Qualcuno vi porrà queste domande, ed è meglio che siate voi i primi. Iniziate quindi a ricercare le risposte quando siete ancora profondamente immersi nel contesto.

Infine, e non meno importante, dovete produrre un report che spieghi come e perché avete elaborato i dati, quali modelli avete utilizzato e quali conclusioni e previsioni sono possibili. Parleremo di struttura dei report verso la fine di questo capitolo, nell'Unità 3, *Struttura dei report*.

Dato che il suo scopo è quello di fungere da supporto per la scelta delle aree appropriate della scienza dei dati basandosi sul linguaggio Python, il *focus* di questo libro è rivolto principalmente ai primi, meno formalizzati e più creativi passi della tipica sequenza di analisi dei dati: ottenere, ripulire, organizzare e dimensionare i dati. La modellazione dei dati, inclusa la modellazione dei dati a fini predittivi, è appena accennata (sarebbe sbagliato trascurare del tutto la modellazione, perché è lì che si può individuare la "vera magia" dell'analisi dei dati!). In generale, l'interpretazione, la verifica e la rappresentazione dei risultati sono molto specifiche del dominio e pertanto appartengono a testi più specializzati.

Unità 2
Sequenza di acquisizione dei dati

L'acquisizione dei dati riguarda tutto ciò che è necessario per ottenere gli artifatti che contengono i dati di input provenienti dalle varie fonti, per estrarre i dati da tali artifatti e per convertirli in rappresentazioni adatte per le successive elaborazioni, come illustrato nella seguente figura.

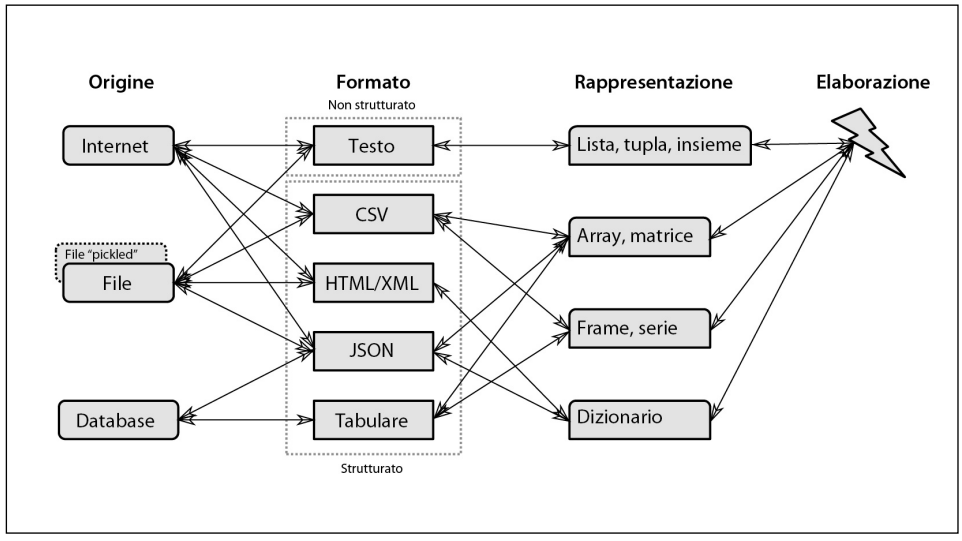

Figura 1.1

Le tre principali fonti di dati sono Internet (in particolare, il World Wide Web), i database e i file locali (magari precedentemente scaricati a mano o usando ulteriore software). Alcuni dei file locali possono essere stati prodotti da altri programmi Python e contenere dati serializzati o "pickled" (vedere l'Unità 12, *Pickling dei dati*).
I formati di dati presenti negli artifatti possono essere i più disparati. Nei prossimi capitoli, vedremo come gestire e manipolare i formati più comuni.

- Testo standard non strutturato in linguaggio naturale (l'inglese, l'italiano, il cinese ecc.).
- Dati strutturati, fra cui:
  - dati tabulari in file CSV (*Comma Separated Values*);
  - dati tabulari tratti da database;
  - dati a tag in formato HTML (*HyperText Markup Language*) o, in generale, in XML (*eXtensible Markup Language*);
  - dati a tag in formato JSON (*JavaScript Object Notation*).

A seconda della struttura originaria dei dati estratti e dello scopo e della natura delle successive elaborazioni, i dati usati negli esempi di questo libro vengono rappresentati

usando le strutture di dati native di Python (liste e dizionari) o strutture di dati avanzate che supportano operazioni specializzate (array `numpy` e frame `pandas`).
Ho cercato di fare in modo che la sequenza di elaborazione dei dati (ottenimento, pulizia e trasformazione di dati grezzi; analisi descrittiva ed esplorativa dei dati; modellazione dei dati e previsione) fosse completamente automatizzata. Per questo motivo, ho evitato di usare strumenti grafici interattivi, i quali solo raramente sono realizzati per operare in modalità batch, e solo raramente registrano una cronologia delle operazioni. Per favorire la modularità, la riusabilità e la recuperabilità, ho suddiviso questa lunga sequenza in sequenze più piccole, salvando i risultati intermedi in file in formato pickle (Unità 12) o JSON (Unità 15).
L'automazione della sequenza porta, naturalmente, ad avere codice riproducibile: un insieme di script Python che chiunque può eseguire per convertire i dati grezzi originari nei risultati finali descritti nel report, teoricamente senza alcuna ulteriore interazione umana. Altri ricercatori potranno usare il codice riproducibile per convalidare i vostri modelli e i vostri risultati e per applicare ai loro problemi il processo che avete sviluppato.

Unità 3
Struttura dei report

Il report è ciò che noi, esperti di scienza dei dati, forniamo al committente, al cliente. In genere un report include i seguenti elementi.

- Sintesi (una breve descrizione del progetto).
- Introduzione.
- Metodi usati per l'acquisizione e l'elaborazione dei dati.
- Risultati ottenuti (escludendo i risultati intermedi e insignificanti, da specificare, semmai, nell'Appendice).
- Conclusioni.
- Appendice.

Oltre ai risultati e ai contenuti grafici non essenziali, l'Appendice deve contenere tutto il codice riproducibile usato per elaborare i dati: script ben commentati che possono essere eseguiti senza alcun parametro e interazione da parte dell'utente.
L'ultima ma non meno importante parte che deve essere fornita è rappresentata dai dati grezzi: tutti i file di dati necessari per eseguire il codice in modo riproducibile, a meno che tali file siano stati forniti dal committente e che non siano stati modificati in alcun modo. Un file `README` in genere spiega la provenienza dei dati e il formato di ogni file di dati allegato.
Considerate questa struttura come una raccomandazione, non come un dogma. Il vostro committente e il vostro buon senso potranno suggerirvi le implementazioni alternative.

Esercitazioni

In questo capitolo introduttivo, abbiamo esaminato i processi che stanno alla base della scienza dei dati: i passi tipici di uno studio di analisi dei dati, dove ottenere i dati e i vari formati dei dati e, infine, la tipica struttura del report di un progetto. La parte rimanente del libro introduce le caratteristiche di Python che sono essenziali per applicare le basi della scienza dei dati, così come i moduli Python che offrono un supporto algoritmico e statistico a un progetto di scienza dei dati di moderata complessità.
Prima di procedere, svolgiamo un semplice progetto per provare a usare Python. Per tradizione, i programmatori sono soliti presentare un nuovo linguaggio di programmazione scrivendo un programma che produce in output le parole "Hello, World!". E non vi è alcun buon motivo per infrangere questa tradizione.

Hello, World! (☆)

Scrivete un programma che produca in output le parole "Hello, World!" (senza gli apici) sulla riga di comando Python.