Presentazione dell'edizione Italiana

La data science è una moderna disciplina fondata su principi, tecniche ed algoritmi di area informatica e matematica che ha come obiettivo l'estrazione di conoscenza e valore dai dati con metodo scientifico. La conoscenza ricavata dai dati è alla base di previsioni e decisioni aziendali e governative, di scoperte in numerosi settori scientifici [1] [2], di studi di fenomeni sociali, ma anche delle applicazioni e sistemi di intelligenza artificiale[3] che stanno rivoluzionando la società e l'economia[4].

Il successo internazionale di questo volume, in ambito universitario e professionale, si deve principalmente alla straordinaria capacità di introdurre la disciplina da zero, seguendo un approccio innovativo ed efficace basato sul coding. Infatti ogni concetto è sviscerato ed accompagnato dall'applicazione pratica in Python, un linguaggio di programmazione che è diventato lo standard de facto in questo ampio settore disciplinare che va dall'analisi dei dati all'intelligenza artificiale con metodi ed algoritmi di machine learning.

Il libro offre un ampio trattamento dei metodi, delle tecniche e degli algoritmi di data science, dai fondamenti della disciplina fino agli algoritmi più popolari di machine learning supervisionato e non supervisionato. Due capitoli sono interamente dedicati a reti neurali e deep learning che sono alla base delle tecnologie di intelligenza artificiale di maggior successo nella computer vision[5] e nel natural language processing[6] [7]. Inoltre un capitolo è dedicato ai sistemi di recommendation che, grazie alla massa crescente dei dati disponibili, sono diventati

[1] L. E. Juarez-Orozco, T. Maaniitty, et al. Refining the long-term prognostic value of hybrid PET/CT through machine learning. European Heart Journal - Cardiovascular Imaging, Volume 20, Issue Supplement_3, June 2019

[2] S. Mostafa Mousavi, William L. Ellsworth, Weiqiang Zhu, Lindsay Y. Chuang, Gregory C. Beroza. Earthquake transformer an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nature Communication 11, 3952 (2020)

[3] David Silver, Julian Schrittwieser, Karen Simonyan, et. al. Mastering the game of Go without human knowledge. Nature 550, 354–359 (2017)

[4] Davenport, T., Guha, A., Grewal, D. et al. How artificial intelligence will change the future of marketing. J. of the Acad. Mark. Sci. 48, 24–42 (2020)

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. CVPR 2016. 770-778. June 2016. Las Vegas, NV, USA

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 NAACL, Volume 1. 4171–4186. June 2019. Minneapolis, Minnesota, USA

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is All You Need. NIPS 2017: 5998-6008. December 2017. Long Beach, CA, USA

ingredienti fondamentali e trasversali ad un'ampia varietà di domini applicativi per prevedere, ad esempio, le propensioni di acquisto di prodotti e servizi di ogni singolo cliente, ma anche le notizie, i post, le nuove connessioni social di suo maggior interesse ed in generale le informazioni più rilevanti rispetto a metriche di similarità con altri utenti.

Queste capacità di profilazione delle persone sono così efficaci che già dal 2013 studi scientifici[8] hanno evidenziato come circa 70 like facebook di un utente siano sufficienti per determinarne automaticamente ed accuratamente informazioni altamente sensibili, tra cui orientamento sessuale, etnia, opinioni religiose e politiche, tratti della personalità, intelligenza, felicità, età, sesso, se fa uso di stupefacenti e se i genitori sono separati. Ciò evidenzia che questi strumenti sollevano anche problemi etici che il libro affronta in un capitolo dove evidenzia i rischi che possono derivare dalle azioni di analisi dei dati, rischi di cui ogni data scientist dovrebbe avere piena consapevolezza.

Il volume non richiede particolari competenze pregresse e contiene anche un capitolo introduttivo a Python, per questo può essere impiegato sia come riferimento in un primo corso universitario di data science, sia a supporto di insegnamenti di data mining e di machine learning, ma anche dei corsi di informatica degli ultimi tre anni del liceo scientifico, dell'istituto tecnico industriale ad indirizzo informatica e dell'istituto tecnico commerciale ad indirizzo ragionieri programmatori.

I concetti fondamentali di algebra lineare, statistica e probabilità alla base della data science sono introdotti in capitoli dedicati e spiegati con esempi facilmente comprensibili, al punto che il libro offre un percorso d'ingresso anche a chi intende avvicinarsi a questa disciplina per la prima volta e senza avere una particolare preparazione in matematica.

Secondo un recente sondaggio internazionale di Gartner group[9] [10] la data science è diventata una priorità strategica negli investimenti di numerose aziende e l'offerta di lavoro per data scientist è almeno un ordine di grandezza superiore al numero di esperti che forma l'accademia. I rapidi avanzamenti, indotti dagli investimenti e dalla ricerca scientifica del settore, avranno un impatto epocale che, secondo gli esperti mondiali, sarà superiore a quello della rivoluzione industriale del settecento e della capillare diffusione dell'energia elettrica nel secolo scorso.

Gianluca Moro, PhD
Docente e ricercatore del Dipartimento di Informatica - Scienza e Ingegneria
Università di Bologna, Campus di Cesena
Via dell'Università 50, 47522 Cesena (FC)
email: nome.cognome@unibo.it
https://www.unibo.it/sitoweb/gianluca.moro

[8] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. PNAS April 9, 2013 110 (15) 5802-5805

[9] https://gtnr.it/3gGZ6W6

[10] https://gtnr.it/3oRK6aX

Prefazione alla seconda edizione

Sono straordinariamente orgoglioso della prima edizione di questo libro, Si è dimostrato davvero il libro che volevo fosse. Parecchi anni di sviluppi nella scienza dei dati, di progressi nell'ecosistema Python e di crescita personale come sviluppatore e formatore, però, hanno *cambiato* come penso debba presentarsi un primo libro di data science.

Nella vita non ci sono quasi mai le seconde occasioni; per i libri, però, esistono le seconde edizioni.

Così, ho riscritto tutto il codice e gli esempi utilizzando Python 3.6 (e molte delle nuove caratteristiche che vi sono state introdotte, come le annotazioni di tipo). Ho messo maggiormente l'accento, in tutto il libro, sulla "pulizia" nella scrittura del codice. Ho sostituito alcuni degli esempi "giocattolo" della prima edizione con altri, più realistici, utilizzando dataset "reali". Ho aggiunto nuovo materiale su temi come l'apprendimento profondo, la statistica e l'elaborazione del linguaggio naturale, così da essere più coerente con quello su cui è probabile uno scienziato dei dati oggi debba lavorare. (Ho anche eliminato del materiale che sembrava meno pertinente.) Ho ripassato con attenzione tutto il libro, sistemando errori, riscrivendo spiegazioni che mi sembravano meno chiare di quel che avrebbero dovuto essere e dando una rinfrescata a qualche battuta.

La prima edizione era un buon libro, questa edizione è ancora meglio. Buona lettura!

Joel Grus
Seattle, WA
2019

Convenzioni utilizzate in questo libro

In questo libro sono utilizzate le convenzioni tipografiche seguenti:

Corsivo Indica nuovi termini, URL, indirizzi di posta elettronica, nomi di file e suffissi di file.

`Monospaziato` È utilizzato per i listati dei programmi e, all'interno dei testo, per evidenziare elementi dei programmi, come nomi di variabili o di funzioni, database, tipi di dati, variabili d'ambiente, enunciati e parole chiave.

`Monospaziato grassetto` Indica comandi o altro testo che deve essere scritto da parte dell'utente, esattamente come si presenta.

`Monospaziato corsivo` Indica testo che deve essere sostituito da valori forniti dall'utente o da valori determinati dal contesto.

Quest'icona indica un suggerimento.

Quest'icona indica una nota di carattere generale.

Quest'icona indica un avvertimento o una nota cautelativa.

Uso degli esempi di codice

Materiale aggiuntivo (esempi di codice, esercizi ecc.) può essere scaricato da *https://github.com/joelgrus/data-science-from-scratch*.

Questo libro vuole essere di aiuto per il vostro lavoro. In generale, se con il libro è offerto codice di esempio, potete usarlo nei vostri programmi e nella vostra documentazione. Non è necessario contattarci per chiedere l'autorizzazione, a meno che riproduciate una parte significativa del codice. Per esempio, per scrivere un programma che usa vari frammenti di codice estratti dal libro non è richiesta autorizzazione. Per vendere o distribuire un CD-ROM di esempi tratti da libri O'Reilly invece è richiesta un'autorizzazione. Rispondere a una domanda citando questo libro e citando codice di esempio non richiede autorizzazione. È necessaria un'autorizzazione invece per poter incorporare una quantità significativa di codice d'esempio tratto da questo libro nella documentazione di un vostro prodotto.

L'attribuzione è gradita ma non richiesta. Un'attribuzione solitamente comprende titolo, autore, editore e ISBN. Per esempio: "*Data Science from Scratch*, Second Edition, by Joel Grus (O'Reilly). Copyright 2019 Joel Grus, 978-1-492-04113-9."

Se pensate che l'uso che intendete fare di esempi di codice ecceda i limiti del fair use o delle autorizzazioni indicate sopra, contattateci all'indirizzo *permissions@oreilly.com*.

O'Reilly Online Learning

Per quasi quarant'anni, *O'Reilly Media* ha fornito formazione, conoscenze e insight su temi tecnologici e di business per aiutare le aziende ad avere successo.

La nostra rete peculiare di esperti e innovatori condivide la propria conoscenza e le proprie competenze attraverso libri, articoli, convegni e la nostra piattaforma di formazione online. La piattaforma di formazione online della O'Reilly dà accesso on-demand a corsi di formazione, percorsi di apprendimento approfondito, ambienti di codifica interattiva e un'ampia raccolta di testi e video fornito da O'Reilly e da oltre 200 altri editori. Per maggiori informazioni, visitate il sito *http://oreilly.com*.

Come contattarci

Indirizzate commenti e domande relativi a questo libro all'editore:

- O'Reilly Media, Inc.
- 1005 Gravenstein Highway North
- Sebastopol, CA 95472
- 800-998-9938 (dagli Stati Uniti o dal Canada)
- 707-829-0515 (internazionali o locali)
- 707-829-0104 (fax)

Esiste una pagina web per questo libro, dove si possono trovare errata corrige, esempi ed eventuali ulteriori informazioni. La pagina si trova all'indirizzo *http://bit.ly/data-science-from-scratch-2e*.

Per commenti o domande tecniche sul libro, inviate un'email a *bookquestions@oreilly.com*.

Per maggiori informazioni sui nostri libri, i corsi, i convegni e le notizie, consultate il nostro sito web: *http://www.oreilly.com*.

Cercateci su Facebook: *http://facebook.com/oreilly*

Seguiteci su Twitter: *http://twitter.com/oreillymedia*

Vedeteci su YouTube: *http://www.youtube.com/oreillymedia*

Ringraziamenti

Innanzitutto, voglio ringraziare Mike Loukides per aver accettato la mia proposta per questo libro (e per aver insistito che lo riducessi a dimensioni ragionevoli). Gli sarebbe stato facile dire, "Chi è questo tizio che continua a spedirmi capitoli di prova, e come faccio a liberarmene?". Sono molto grato che non l'abbia fatto. Voglio ringraziare anche i miei redattori, Michele Cronin e Marie Beaugureau,

per avermi guidato lungo il processo della pubblicazione e per aver portato il libro a una condizione molto migliore di quella a cui avrei saputo condurlo io da solo.

Non avrei mai potuto scrivere questo libro se non avessi studiato la data science e probabilmente non avrei mai imparato la data science se non fosse stato per l'influenza di Dave Hsu, Igor Tatarinov, John Rauser e del resto della banda del Farecast. (È passato così tanto tempo che allora nemmeno veniva chiamata data science!) Anche tutti a Coursera e DataTau meritano ampio credito.

Sono grato anche ai "lettori in beta" e ai revisori. Jay Fundling ha trovato una marea di errori e mi ha evidenziato molte spiegazioni poco chiare; il libro ora è molto migliore (e molto più corretto) grazie a lui. Debashis Ghosh ha fatto un grande lavoro verificando tutte le mie statistiche. Andrew Musselman ha suggerito di abbassare i toni della polemica "le persone che preferiscono R a Python sono moralmente riprovevoli" e penso che alla fine sia stato un buon consiglio. Anche Trey Causey, Ryan Matthew Balfanz, Loris Mularoni, Núria Pujol, Rob Jefferson, Mary Pat Campbell, Zach Geary, Denise Mauldin, Jimmy O'Donnell e Wendy Grus hanno fornito feedback preziosi. Grazie a tutte le persone che hanno letto la prima edizione e hanno contribuito a migliorare questo libro. Gli errori eventualmente rimasti sono ovviamente responsabilità mia.

Devo molto alla comunità #datascience di Twitter, per avermi esposto a una grande quantità di nuovi concetti, per avermi fatto conoscere moltissime ottime persone e per avermi fatto sentire tanto inadeguato da dovermi mettere a scrivere un libro per compensare. Un grazie speciale a Trey Causey (ancora), per avermi ricordato (senza volerlo) di includere un capitolo sull'algebra lineare, e a Sean J. Taylor per avere evidenziato (senza volerlo) un paio di enormi lacune nel capitolo su "Lavorare con i dati".

Soprattutto, un grazie enorme a Ganga e Madeline. L'unica cosa più difficile dello scrivere un libro è vivere con qualcuno che sta scrivendo un libro e non ce l'avrei mai fatta senza il loro sostegno.

Prefazione alla prima edizione

Data science

Quello del "data scientist" è stato definito "the sexiest job of the 21st century", probabilmente da qualcuno che non ha mai visitato una caserma dei pompieri. La data science, comunque, è un campo molto alla moda e in continua crescita, e non c'è bisogno di darsi molto da fare per trovare qualche analista che pronostica affannosamente che entro i prossimi dieci anni avremo bisogno di miliardi e miliardi di scienziati dei dati più di quelli attualmente esistenti.

Ma che cos'è la data science o "scienza dei dati"? In fin dei conti, non possiamo produrre scienziati dei dati se non sappiamo che cosa sia la data science. In base a un diagramma di Venn abbastanza famoso nel settore, la data science si trova all'intersezione di:

- capacità di hacking,
- conoscenze di matematica e statistica,
- ampie competenze settoriali.

Inizialmente avrei voluto scrivere un libro che parlasse di tutti i tre ambiti, ma mi sono rapidamente reso conto che per trattare approfonditamente "ampie competenze settoriali" sarebbero servite decine di migliaia di pagine. A quel punto, ho deciso di concentrarmi sui primi due. Il mio obiettivo è aiutarvi a sviluppare le capacità di hacking di cui avrete bisogno per iniziare a fare data science, e aiutarvi a sentirvi a vostro agio con la matematica e la statistica che sono al centro della scienza dei dati.

È un obiettivo piuttosto ambizioso per un libro. Il modo migliore per apprendere le capacità di hacking (di codifica, di programmazione) è scrivere codice. Leggendo questo libro, vi farete una buona idea del mio modo di manipolare le cose, che non sarà necessariamente il modo migliore per voi. Avrete una buona idea di alcune degli strumenti che uso io, che non saranno necessariamente gli strumenti migliori per voi. Avrete una buona comprensione del modo in cui affronto i problemi dei dati, che non sarà necessariamente il modo migliore di affrontare i problemi dei dati per voi. L'intento (e la speranza) è che i miei esempi vi motivino a provare a fare a modo vostro. Tutto il codice e i dati del libro sono disponibili su GitHub in modo che possiate iniziare senza fatica.

Analogamente, il modo migliore per imparare la matematica è fare matematica. Questo decisamente non è un libro di matematica e, nella maggior parte dei casi, qui non "faremo matematica", ma non si può davvero fare data science senza *qualche* conoscenza di probabilità, statistica e algebra lineare. Questo significa che, quando sarà opportuno, ci tufferemo in equazioni matematiche, intuizioni matematiche, assiomi matematici e versioni "da fumetto" di grandi idee matematiche. Spero che non abbiate timore di tuffarvi con me.

Nel complesso, spero anche di darvi l'idea che giocare con i dati è divertente, perché, beh, giocare con i dati è divertente! (In particolare, se lo si paragona con certe alternative, come compilare la denuncia dei redditi o estrarre carbone in miniera.)

Da zero

Esiste un numero grandissimo di librerie, framework, moduli e strumenti per la data science che implementano in modo efficiente gli algoritmi e le tecniche più frequenti (e anche quelli meno frequenti). Se diventerete scienziati dei dati, dovrete fare la conoscenza approfondita di NumPy, scikit-learn, pandas e un ampio assortimento di altre librerie. Sono ottime per fare data science, ma sono anche un buon modo per iniziare a fare data science senza capirne realmente nulla.

In questo libro, affronteremo la data science da zero. Questo significa che costruiremo strumenti e implementeremo algoritmi manualmente, per capirli meglio. Ho riflettuto molto su come creare implementazioni ed esempi che fossero chiari, ben commentati e leggibili. Nella maggior parte dei casi, gli strumenti che costruiremo saranno illuminanti ma privi di utilità pratica: funzioneranno bene su piccoli dataset "giocattolo", ma andranno a rotoli se applicati alla scala del Web.

In tutto il libro, indicherò librerie che potete usare per applicare queste tecniche a dataset di dimensioni maggiori, ma qui non le useremo.

Si discute molto (ed è una buona cosa) su quale sia il linguaggio migliore per imparare la data science. Molti sono convinti che sia R, il linguaggio di programmazione statistica. (Per noi, sbagliano.) Qualcuno suggerisce Java o Scala. Secondo me, la scelta ovvia è Python.

Python ha molte caratteristiche che lo rendono particolarmente adatto per apprendere (e per fare) data science:

- è gratuito;
- scrivere codice in Python è relativamente semplice (e lo è anche comprenderlo);
- ha moltissime librerie utili per la scienza dei dati.

Sono restio a dire che Python sia il linguaggio di programmazione che preferisco. Esistono altri linguaggi che trovo più piacevoli, meglio progettati o in cui è

semplicemente più divertente scrivere codice. Tuttavia, praticamente tutte le volte in cui inizio un nuovo progetto di data science finisco per usare Python. Tutte le volte che devo realizzare rapidamente un prototipo che funzioni, finisco per usare Python. E tutte le volte che voglio dimostrare concetti di data science in modo chiaro e facilmente comprensibile finisco per usare Python. Quindi, questo libro usa Python.

L'obiettivo del libro però non è quello di insegnare Python (anche se quasi sicuramente leggendolo lo imparerete un po'). Ci sarà un intero capitolo che offre un corso accelerato in cui verranno evidenziate le caratteristiche più importanti per i nostri scopi, ma se non sapete nulla di programmazione in Python (o di programmazione in generale), forse è meglio che affrontiate, a complemento di questo libro qualche tipo di tutorial di "Python per principianti".

Il resto della nostra introduzione alla data science seguirà la stessa impostazione: entreremo nei dettagli quando farlo sarà fondamentale o illuminante, mentre in altre occasioni lascerò che risolviate i dettagli per conto vostro (o che andiate a cercarveli su Wikipedia).

Nel corso degli anni, ho formato numerosi scienziati dei dati. Non tutti sono diventati stelle di prima grandezza destinate a cambiare il mondo, ma alla fine erano tutti scienziati dei dati migliori di quanto non fossero all'inizio. E mi sono convinto che chiunque abbia un po' di attitudine per la matematica e un po' di competenze di programmazione possiede le materie prime per fare data science. Tutto quello che gli serve è una mente curiosa, la disponibilità a lavorare sodo, e questo libro. Perciò, ecco il libro.