Introduzione all'apprendimento statistico

con applicazioni in R

Gareth James • Daniela Witten • Trevor Hastie
Robert Tibshirani

Introduzione all'apprendimento statistico

con applicazioni in R

Edizione italiana a cura di
Silvia Salini
Sabrina Gaito
Patrizia Boracchi
Federico Ambrogi
Giancarlo Manzi
ed Elia Biganzoli

PICCIN

First published in English under the title
An Introduction to Statistical Learning; with Applications in R
by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, edition: 1

Gareth James
Department of Data Sciences and Operations
University of Southern California
Los Angeles, CA, USA

Daniela Witten
Department of Biostatistics
University of Washington
Seattle, WA, USA

Trevor Hastie
Department of Statistics
Stanford University
Stanford, CA, USA

Robert Tibshirani
Department of Statistics
Stanford University
Stanford, CA, USA

ISBN 978-88-299-3094-4

Stampato in Italia

© 2021, by Piccin Nuova Libraria S.p.A., Padova
www.piccin.it

Ai nostri genitori:

Alison e Michael James

Chiara Nappi e Edward Witten

Valerie e Patrick Hastie

Vera e Sami Tibshirani

e alle nostre famiglie:

Michael, Daniel, e Catherine

Tessa, Theo, e Ari

Samantha, Timothy, e Lynda

Charlie, Ryan, Julie, e Cheryl

Presentazione dell'edizione italiana

Quando all'inizio del 2019 l'Editore Dr. Nicola Piccin ci offrì la traduzione del libro *An Introduction to Statistical Learning* di James, Witten, Hastie, Tibshirani la nostra prima reazione fu l'onore per la proposta di occuparci di un'opera così popolare, amata da ricercatori, studenti e appassionati di *Data Science*. Il testo non richiede conoscenze di calcolo troppo avanzate e allo stesso tempo offre efficaci esempi tutoriali in R. Subito dopo arrivarono i dubbi sull'opportunità di tradurre in italiano questa opera in un mondo sempre più legato all'uso dell'inglese come lingua unica per la scienza.

Ci aiutò l'atteggiamento dell'Editore Piccin, evidentemente meno preoccupato di avere un "blockbuster" di vendite nel già difficile mercato dell'editoria universitaria, ma sinceramente animato da un'operazione di sviluppo culturale/educativo nella tradizione della sua casa editrice.

Ci fa piacere infatti ricordare la vicinanza storica della Piccin allo sviluppo della Statistica e della Biometria in Italia.

Effettivamente e fortunatamente, studenti, docenti e ricercatori sviluppano la propria esperienza continuando l'uso dell'italiano insieme a un gergo che mutua termini vari dall'originale inglese. Abbiamo quindi cercato di mantenere lo stile corrente nell'uso della terminologia e i Lettori vorranno perdonare eventuali disomogeneità, fornendo i propri consigli per le possibili revisioni.

Non nascondiamo le difficoltà del lavoro di traduzione svolto nel corso dell'anno 2020. Con l'arrivo della pandemia Covid-19, il gruppo di noi traduttori si è infatti ritrovato unito sul fronte della ricerca nella *Data Science* applicata all'Epidemiologia per combattere il virus e tutti impegnati nella didattica online o *blended* nelle aule con gli studenti presenti a turni alterni.

Ci auguriamo allora che l'opera tradotta possa essere apprezzata e utile per lo sviluppo di tante nuove carriere di giovani appassionati nella Scienza dei Dati.

Oltre a ringraziare ancora il Dr. Nicola Piccin per questa sfida che abbiamo voluto raccogliere, estendiamo il ringraziamento allo staff editoriale Piccin e in particolare alla Dr.ssa Giulia Barollo, responsabile redazionale dell'opera che ci ha seguito con assidua pazienza e squisita cortesia. Un ultimo doveroso e affettuoso ringraziamento va alle nostre famiglie che ci hanno sostenuto anche per questa parte del lavoro in un momento così difficile per tutti.

Milano, novembre 2020

Silvia Salini
Sabrina Gaito
Patrizia Boracchi
Federico Ambrogi
Giancarlo Manzi
ed Elia Biganzoli

Data Science Research Center –
Università degli Studi di Milano La Statale

Prefazione

L'apprendimento statistico (*statistical learning*) si riferisce a una serie di strumenti per modellizzare e comprendere set di dati complessi. Si tratta di un'area sviluppata di recente nella statistica e si fonde con sviluppi paralleli nell'informatica e, in particolare, nel *machine learning*. Il campo comprende molti metodi come il *lasso* e la regressione sparsa, la classificazione e gli alberi di regressione, il *boosting* e le *support vector machines*.

Con l'esplosione dei problemi legati ai "Big Data", l'apprendimento statistico è diventato un campo di forte interesse in molte aree scientifiche, nonché nel marketing, finanza e altre discipline aziendali. Le persone con capacità relative allo *statistical learning* sono molto richieste.

Uno dei primi libri in quest'area - *The Elements of Statistical Learning* (ESL) (Hastie, Tibshirani e Friedman) - è stato pubblicato nel 2001, con una seconda edizione nel 2009. L'ESL è diventato un testo popolare non solo in statistica ma anche in campi correlati. Uno dei motivi della popolarità di ESL è il suo stile relativamente accessibile. Ma ESL è destinato a persone con una formazione avanzata nelle scienze matematiche. *An Introduction to Statistical Learning* (ISL) è nato dalla percepita necessità di un trattamento più ampio e meno tecnico di questi argomenti. In questo nuovo libro, trattiamo molti degli stessi argomenti di ESL, ma ci concentriamo maggiormente sulle applicazioni dei metodi e meno sui dettagli matematici. Abbiamo creato esercitazioni che illustrano come implementare ciascuno dei metodi di apprendimento statistico utilizzando il diffuso pacchetto software statistico R. Queste esercitazioni forniscono al lettore una preziosa esperienza pratica.

Questo libro è adatto a studenti universitari avanzati o studenti di corsi magistrali in statistica o campi quantitativi correlati o per persone di altre discipline che desiderano utilizzare strumenti di apprendimento statistico per analizzare i propri dati. Può essere usato come libro di testo per un corso che copre uno o due semestri.

Vorremmo ringraziare diversi lettori per i preziosi commenti sulle bozze preliminari di questo libro: Pallavi Basu, Alexandra Chouldechova, Patrick Danaher, Will Fithian, Luella Fu, Sam Gross, Max Grazier G'Sell, Courtney Paulson, Xinghao Qiao, Elisa Sheng, Noah Simon, Kean Ming Tan, e Xin Lu Tan.

It's tough to make predictions, especially about the future.

-Yogi Berra

Los Angeles, USA — Gareth James
Seattle, USA — Daniela Witten
Palo Alto, USA — Trevor Hastie
Palo Alto, USA — Robert Tibshirani

Traduttori

Federico Ambrogi
Professore Associato
Dipartimento di Scienze Cliniche
e di Comunità
Università degli Studi di Milano
La Statale

Elia Biganzoli
Professore Ordinario
Dipartimento di Scienze Cliniche
e di Comunità
Università degli Studi di Milano
La Statale

Patrizia Boracchi
Professore Associato
Dipartimento di Scienze Cliniche
e di Comunità
Università degli Studi di Milano
La Statale

Sabrina Tiziana Gaito
Professore Associato
Dipartimento di Informatica
"Giovanni degli Antoni"
Università degli Studi di Milano
La Statale

Giancarlo Manzi
Professore Associato
Dipartimento di Economia,
Management e Metodi Quantitativi
Università degli Studi di Milano
La Statale

Silvia Salini
Professore Associato
Dipartimento di Economia,
Management e Metodi Quantitativi
Università degli Studi di Milano
La Statale

Indice generale