

# **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  
Copyright © Springer Science+Business Media, LLC, part of Springer Nature, 2013 \*  
This edition has been translated and published under licence from  
Springer Science+Business Media, LLC, part of Springer Nature.  
Springer Science+Business Media, LLC, part of Springer Nature takes no responsibility  
and shall not be made liable for the accuracy of the translation.

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

Opera coperta dal diritto d'autore – tutti i diritti sono riservati.  
Questo testo contiene materiale, testi ed immagini, coperto da copyright e non può essere copiato, riprodotto, distribuito, trasferito, noleggiato, licenziato o trasmesso in pubblico, venduto, prestato a terzi, in tutto o in parte, o utilizzato in alcun altro modo o altrimenti diffuso, se non previa espressa autorizzazione dell'editore. Qualsiasi distribuzione o fruizione non autorizzata del presente testo, così come l'alterazione delle informazioni elettroniche, costituisce una violazione dei diritti dell'editore e dell'autore e sarà sanzionata civilmente e penalmente secondo quanto previsto dalla L. 633/1941 e ss.mm.

ISBN 978-88-299-3094-4

Stampato in Italia

---

© 2021, by Piccin Nuova Libreria S.p.A., Padova  
[www.piccin.it](http://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  
Seattle, USA  
Palo Alto, USA  
Palo Alto, USA

Gareth James  
Daniela Witten  
Trevor Hastie  
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

<b>Presentazione dell'edizione italiana</b>	<b>vii</b>
<b>Prefazione</b>	<b>ix</b>
<b>Traduttori</b>	<b>xi</b>
<b>1 Introduzione</b>	<b>1</b>
<b>2 L'apprendimento statistico</b>	<b>15</b>
2.1 Cos'è l'apprendimento statistico? . . . . .	15
2.1.1 Perché stimare $f^P$ . . . . .	17
2.1.2 Come stimare $f^P$ . . . . .	21
2.1.3 Il trade-off tra l'accuratezza della previsione e l'interpretabilità del modello. . . . .	24
2.1.4 Apprendimento supervisionato e non supervisionato. . . . .	26
2.1.5 Problemi di regressione e di classificazione a confronto. . . . .	28
2.2 La valutazione dell'accuratezza di un modello . . . . .	29
2.2.1 La misurazione della qualità dell'adattamento . . . . .	29
2.2.2 Il trade-off tra distorsione (bias) e varianza. . . . .	33
2.2.3 Il contesto della classificazione . . . . .	37
2.3 Laboratorio: introduzione a R. . . . .	42
2.3.1 Comandi di base . . . . .	42
2.3.2 Grafici . . . . .	45
2.3.3 Indicizzazione dei dati . . . . .	47
2.3.4 Caricamento dei dati. . . . .	48
2.3.5 Riepiloghi grafici e numerici aggiuntivi . . . . .	50
2.4 Esercizi . . . . .	52

<b>3</b>	<b>Regressione lineare</b>	<b>59</b>
3.1	Regressione lineare semplice . . . . .	61
3.1.1	Stima dei coefficienti . . . . .	61
3.1.2	Valutare l'accuratezza della stima dei coefficienti . . . . .	63
3.1.3	Valutare l'accuratezza del modello . . . . .	68
3.2	Regressione lineare multipla . . . . .	71
3.2.1	Stima dei coefficienti di regressione . . . . .	72
3.2.2	Alcune importanti questioni . . . . .	75
3.3	Altre considerazioni sul modello di regressione . . . . .	82
3.3.1	Predittori qualitativi . . . . .	82
3.3.2	Estensioni del modello lineare . . . . .	86
3.3.3	Potenziati problemi . . . . .	92
3.4	Il piano di marketing . . . . .	102
3.5	Confronto della regressione lineare con il $K$ -nearest neighbors . . . . .	104
3.6	Laboratorio: regressione lineare . . . . .	109
3.6.1	Librerie . . . . .	109
3.6.2	Regressione lineare semplice . . . . .	110
3.6.3	Regressione lineare multipla . . . . .	113
3.6.4	Termini di interazione . . . . .	115
3.6.5	Trasformazioni non lineari dei predittori . . . . .	116
3.6.6	Predittori qualitativi . . . . .	118
3.6.7	Scrivere funzioni . . . . .	120
3.7	Esercizi . . . . .	120
<b>4</b>	<b>Classificazione</b>	<b>129</b>
4.1	Panoramica dei problemi di classificazione . . . . .	130
4.2	Perché non la regressione lineare? . . . . .	131
4.3	Regressione logistica . . . . .	132
4.3.1	Il modello logistico . . . . .	133
4.3.2	Stima dei coefficienti di regressione . . . . .	135
4.3.3	Effettuare predizioni . . . . .	136
4.3.4	Regressione logistica multipla . . . . .	137
4.3.5	Regressione logistica per più di 2 classi di risposta . . . . .	139
4.4	Analisi discriminante lineare . . . . .	140
4.4.1	Uso del teorema di Bayes nella classificazione . . . . .	140
4.4.2	Analisi discriminante lineare per $p = 1$ . . . . .	141
4.4.3	Analisi discriminante lineare con $p > 1$ . . . . .	144
4.4.4	Analisi discriminante quadratica . . . . .	151
4.5	Confronto tra i metodi di classificazione . . . . .	153
4.6	Laboratorio: regressione logistica, LDA, QDA, e KNN . . . . .	156
4.6.1	Il data set Stock Market . . . . .	156
4.6.2	Regressione logistica . . . . .	158
4.6.3	Analisi discriminante lineare . . . . .	163
4.6.4	Analisi discriminante quadratica . . . . .	165
4.6.5	$K$ -nearest neighbors . . . . .	166
4.6.6	Un'applicazione al data set Caravan Insurance . . . . .	167
4.7	Esercizi . . . . .	171

<b>5</b>	<b>Metodi di resampling</b>	<b>177</b>
5.1	Cross-validation . . . . .	178
5.1.1	Validation Set. . . . .	178
5.1.2	Leave-one-out cross-validation . . . . .	180
5.1.3	$K$ -fold cross-validation. . . . .	183
5.1.4	Bias-variance trade-off per $k$ -fold cross-validation. . . . .	185
5.1.5	La cross-validation nei problemi di classificazione. . . . .	186
5.2	Il bootstrap . . . . .	188
5.3	Laboratorio: la cross-validation e il bootstrap . . . . .	192
5.3.1	L'approccio del set di validazione. . . . .	192
5.3.2	La cross-validation di tipo "leave-one-out". . . . .	193
5.3.3	La cross validation di tipo " $k$ -fold". . . . .	195
5.3.4	Il bootstrap . . . . .	195
5.4	Esercizi . . . . .	199
<b>6</b>	<b>Scelta del modello lineare e regolarizzazione</b>	<b>205</b>
6.1	Scelta di un sottoinsieme di variabili . . . . .	207
6.1.1	Scelta del miglior sottoinsieme . . . . .	207
6.1.2	Scelta del modello iterativa (stepwise). . . . .	209
6.1.3	La scelta del modello ottimo . . . . .	212
6.2	Metodi di shrinkage . . . . .	216
6.2.1	Regressione ridge. . . . .	217
6.2.2	Il lasso . . . . .	221
6.2.3	La scelta del parametro di tuning . . . . .	229
6.3	Metodi di riduzione della dimensionalità . . . . .	230
6.3.1	Regressione delle componenti principali . . . . .	232
6.3.2	Minimi quadrati parziali . . . . .	239
6.4	Considerazioni con alta dimensionalità . . . . .	240
6.4.1	Dati ad alta dimensionalità. . . . .	240
6.4.2	Cosa non funziona in alta dimensionalità? . . . . .	241
6.4.3	Regressione in problemi ad alta dimensionalità . . . . .	243
6.4.4	Interpretazione dei risultati con alta dimensionalità . . . . .	245
6.5	Laboratorio 1: metodi di scelta di un sottoinsieme di variabili. . . . .	246
6.5.1	Scelta del miglior sottoinsieme . . . . .	246
6.5.2	Scelta iterativa in avanti e all'indietro . . . . .	249
6.5.3	Scelta del modello con dati di validazione e cross-validazione . . . . .	250
6.6	Laboratorio 2: regressione ridge e il lasso. . . . .	253
6.6.1	Regressione ridge. . . . .	254
6.6.2	Il lasso . . . . .	258
6.7	Laboratorio 3: regressione PCR e PLS . . . . .	259
6.7.1	Regressione delle componenti principali . . . . .	259
6.7.2	Minimi quadrati parziali . . . . .	261
6.8	Esercizi . . . . .	262
<b>7</b>	<b>Andare oltre la linearità</b>	<b>269</b>
7.1	Regressione polinomiale . . . . .	270
7.2	Funzioni a gradino . . . . .	272

7.3	Funzioni base . . . . .	274
7.4	Spline di regressione . . . . .	274
7.4.1	Polinomiale a tratti . . . . .	275
7.4.2	Vincoli e spline . . . . .	275
7.4.3	La rappresentazione delle basi della spline . . . . .	277
7.4.4	Scegliere il numero e la posizione dei nodi . . . . .	278
7.4.5	Confronto con la regressione polinomiale . . . . .	280
7.5	Spline smussanti . . . . .	281
7.5.1	Una panoramica delle spline smussanti . . . . .	281
7.5.2	Scegliere il parametro di smussamento $\lambda$ . . . . .	282
7.6	Regressione locale . . . . .	284
7.7	Modelli additivi generalizzati . . . . .	286
7.7.1	GAM per problemi di regressione . . . . .	287
7.7.2	GAM per problemi di classificazione . . . . .	290
7.8	Laboratorio: modellamento non-lineare . . . . .	291
7.8.1	Regressione polinomiale e funzioni a gradino . . . . .	292
7.8.2	Spline . . . . .	297
7.8.3	GAM . . . . .	299
7.9	Esercizi . . . . .	302
<b>8</b>	<b>Metodi ad albero</b>	<b>309</b>
8.1	Le basi degli alberi decisionali . . . . .	309
8.1.1	Alberi di regressione . . . . .	310
8.1.2	Alberi di classificazione . . . . .	317
8.1.3	Alberi versus modelli lineari . . . . .	320
8.1.4	Vantaggi e svantaggi degli alberi. . . . .	321
8.2	Bagging, foreste random, boosting . . . . .	322
8.2.1	Bagging . . . . .	322
8.2.2	Foreste random. . . . .	325
8.2.3	Boosting . . . . .	327
8.3	Laboratorio: alberi decisionali . . . . .	329
8.3.1	Stima di alberi di classificazione. . . . .	329
8.3.2	Stima di alberi di regressione . . . . .	334
8.3.3	Bagging e foreste random . . . . .	335
8.3.4	Boosting . . . . .	337
8.4	Esercizi . . . . .	338
<b>9</b>	<b>Support vector machines (SVM)</b>	<b>343</b>
9.1	Classificatore a margine massimo . . . . .	344
9.1.1	Cos'è un iperpiano? . . . . .	344
9.1.2	Classificazione mediante un iperpiano separante. . . . .	345
9.1.3	Il classificatore a margine massimo . . . . .	347
9.1.4	Costruzione del classificatore a margine massimo . . . . .	348
9.1.5	Il caso non separabile . . . . .	349
9.2	Classificatori a vettori di supporto . . . . .	350
9.2.1	Panoramica del classificatore support vector. . . . .	350
9.2.2	Dettagli del classificatore support vector . . . . .	351



9.3	Support vector machines (SVM)	355
9.3.1	Classificazione con confini di decisione non-lineari	355
9.3.2	La support vector macchine	356
9.3.3	Un'applicazione a dati di malattia cardiaca	360
9.4	SVM con più di due classi	361
9.4.1	Classificazione uno contro uno	361
9.4.2	Classificazione uno contro tutti	362
9.5	Relazione con la regressione logistica	362
9.6	Laboratorio: support vector machines (SVM)	365
9.6.1	Classificatore support vector	365
9.6.2	Support vector machines (SVM)	369
9.6.3	Curve ROC	371
9.6.4	SVM con classi multiple	372
9.6.5	Applicazione a dati di espressione genica	373
9.7	Esercizi	374
<b>10</b>	<b>Apprendimento non-supervisionato</b>	<b>379</b>
10.1	La sfida dell'apprendimento non supervisionato	379
10.2	Analisi delle componenti principali	380
10.2.1	Cosa sono le componenti principali?	381
10.2.2	Un'altra interpretazione delle componenti principali	385
10.2.3	Altro sulla PCA	386
10.2.4	Altri usi per le componenti principali	390
10.3	Metodi di clustering	391
10.3.1	$K$ -means clustering	392
10.3.2	Clustering gerarchico	395
10.3.3	Questioni pratiche nel clustering	405
10.4	Laboratorio 1: Principal Components Analysis	407
10.5	Laboratorio 2: clustering	410
10.5.1	$K$ -means clustering	410
10.5.2	Clustering gerarchico	412
10.6	Laboratorio 3: NCI60 Data Example	413
10.6.1	PCA del data set NCI60	414
10.6.2	Clustering delle osservazioni del data set NCI60	417
10.7	Esercizi	420
	<b>Indice analitico</b>	<b>425</b>

