Trevor Hastie
Robert Tibshirani
Jerome Friedman

Gli elementi dell'apprendimento statistico

Data mining, inferenza e previsione

Edizione italiana sulla seconda in lingua inglese a cura di

Elia Biganzoli
Patrizia Boracchi
Sabrina Gaito
Federica Nicolussi
Silvia Salini
Federico Ambrogi
Giancarlo Manzi
Giuseppe Marano
Matteo Zignani

Con il coordinamento di Elia Biganzoli
e la collaborazione di Giacomo Biganzoli

PICCIN

First published in English under the title
The Elements of Statistical Learning; Data Mining, Inference, and Prediction, Second Edition
by Trevor Hastie, Robert Tibshirani and Jerome Friedman, edition: 2

Trevor Hastie
Stanford University
Dept. of Statistics
Stanford CA 94305
USA
hastie@stanford.edu

Robert Tibshirani
Stanford University
Dept. of Statistics
Stanford CA 94305
USA
tibs@stanford.edu

Jerome Friedman
Stanford University
Dept. of Statistics
Stanford CA 94305
USA
jhf@stanford.edu

ISBN 978-88-299-3272-6

www.piccin.it

Ai nostri genitori:

Valerie e Patrick Hastie

Vera e Sami Tibshirani

Florence e Harry Friedman

e alle nostre famiglie:

Samantha, Timothy e Lynda

Charlie, Ryan, Julie e Cheryl

Melanie, Dora, Monika e Ildiko

Presentazione dell'Edizione Italiana

Quando l'Editore Dr. Nicola Piccin ci ha proposto la traduzione italiana del testo originale *The Elements of Statistical Learning* di Hastie, Tibshirani e Friedman, in continuità con il nostro precedente lavoro di traduzione del testo successivo di James, Witten, Hastie e Tibshirani *An Introduction to Statistical Learning*, abbiamo accolto la sfida con una profonda consapevolezza della sua importanza scientifica, ma anche del grande impegno che essa avrebbe richiesto. Si tratta infatti di un testo fondamentale nella moderna Statistica e nell'apprendimento automatico, ampiamente riconosciuto per la sua profondità teorica e per l'impatto che ha avuto, e continua ad avere, nella formazione di intere generazioni di studiosi, ricercatori e data scientist.

Rispetto alla precedente opera di James, Witten, Hastie e Tibshirani, che abbiamo avuto l'onore di tradurre nel 2020, questo volume presenta un livello tecnico e formale significativamente più elevato. La struttura matematica più densa, il lessico talvolta gergale e la presenza di riferimenti avanzati alla teoria statistica e computazionale hanno reso il lavoro di traduzione ancora più impegnativo. In particolare, ci siamo confrontati con la necessità di tradurre termini e concetti specialistici mantenendo al tempo stesso la precisione scientifica e la coerenza con l'uso corrente nella comunità italiana.

Come nel progetto precedente, abbiamo condiviso con l'Editore PICCIN la convinzione che offrire un'opera così rilevante in lingua italiana non fosse un semplice esercizio editoriale, ma un contributo culturale ed educativo volto a stimolare il dialogo tra colleghi, docenti e studenti. In un contesto in cui l'inglese domina come lingua veicolare della scienza, crediamo ancora che una traduzione ben ponderata possa rafforzare la comprensione, favorire l'inclusione formativa e promuovere una terminologia il più possibile condivisa all'interno della nostra comunità accademica.

Abbiamo cercato di mantenere uno stile rigoroso ma accessibile, rispettando le scelte lessicali consolidate e introducendo con cautela nuove proposte terminologiche, sempre con l'obiettivo di rendere il testo utile allo studio, all'insegnamento e alla pratica scientifica. Ci scusiamo in anticipo per eventuali disomogeneità residue e siamo grati a chi vorrà segnalarci suggerimenti per future revisioni.

Se il precedente lavoro di traduzione era stato condotto durante il periodo della pandemia da Covid-19, quando i lockdown avevano paradossalmente reso possibile dedicare tempo alla riflessione e al lavoro condiviso a distanza, in questo secondo caso il maggiore impegno richiesto dal testo si è dovuto confrontare con tempi più limitati e con l'intensificarsi delle attività accademiche e professionali in presenza.

Ringraziamo nuovamente il Dr. Nicola Piccin per averci offerto questa nuova opportunità, che abbiamo vissuto come un'estensione naturale del precedente impegno. Un ringraziamento speciale va alla Dr.ssa Cecilia Allegri per la sua costante disponibilità e attenzione nella gestione e nella lavorazione redazionale insieme al redattore Dr. Stefano Girardi e a tutto lo staff editoriale PICCIN. Infine, il ringraziamento va alle nostre famiglie, che ci hanno sostenuto anche in questo nuovo lavoro, nato in un periodo complesso ma stimolante per l'evoluzione della scienza dei dati.

Milano, Luglio 2025
Elia Biganzoli con Patrizia Boracchi,
Sabrina Gaito, Federica Nicolussi,
Silvia Salini, Federico Ambrogi,
Giancarlo Manzi, Giuseppe Marano,
Matteo Zignani e Giacomo Biganzoli

Prefazione alla Seconda Edizione

Crediamo in Dio, tutti gli altri devono portare dati.

–William Edwards Deming (1900-1993)[1]

Siamo stati gratificati dalla popolarità della prima edizione di quest'opera. Ciò, insieme al ritmo serrato della ricerca nel campo dell'apprendimento statistico, ci ha motivato ad aggiornare il nostro libro con la seconda edizione.

Abbiamo aggiunto quattro nuovi capitoli e aggiornato alcuni dei capitoli esistenti. Poiché molti lettori hanno familiarità con l'impaginazione della prima edizione, abbiamo cercato di modificarla il meno possibile. Ecco un riepilogo delle principali modifiche:

[1] Sul Web, questa citazione è stata ampiamente attribuita sia a Deming che a Robert W. Hayden; tuttavia il Professor Hayden ci ha detto che non può rivendicare alcun merito per questa citazione, e ironicamente non siamo riusciti a trovare alcun "dato" che confermi che Deming abbia effettivamente affermato ciò.

Capitolo	Cosa c'è di nuovo
1. Introduzione	
2. Panoramica sull'apprendimento supervisionato	
3. Metodi lineari per la regressione	Algoritmo LAR e generalizzazioni del lasso
4. Metodi lineari di classificazione	Percorso lasso per la regressione logistica
5. Espansione delle funzioni di base e regolarizzazione	Ulteriori illustrazioni di RKHS
6. Metodi kernel	
7. Valutazione e selezione dei modelli	Punti di forza e insidie del cross-convalida
8. Inferenza e media dei modelli	
9. Modelli additivi, alberi e metodi affini	
10. Boosting e alberi additivi	Nuovo esempio dall'ecologia; parte del materiale suddiviso nel Capitolo 16.
11. Reti neurali	Le reti neurali bayesiane e la sfida NIPS 2003
12. Support vector machine e discriminanti flessibili	Algoritmo del percorso per il classificatore SVM
13. Metodi prototipali e vicino più prossimo	
14. Apprendimento non supervisionato	Clustering spettrale, kernel PCA, PCA sparsa, analisi archetipica della fattorizzazione di matrici non negative, riduzione della dimensione non lineare, algoritmo di page ranking di Google, un approccio diretto all'ICA
15. Random forest	Nuovo
16. Apprendimento d'insieme	Nuovo
17. Modelli grafici non orientati	Nuovo
18. Problemi in grandi dimensioni	Nuovo

Alcune ulteriori note:

- La nostra prima edizione era ostile ai lettori daltonici; in particolare, tendevamo a favorire i contrasti rosso/verde particolarmente fastidiosi. In questa edizione abbiamo cambiato in larga misura la tavolozza dei colori, sostituendo quanto sopra con un contrasto arancione/blu.
- Abbiamo cambiato il titolo del Capitolo 6 per evitare confusione con il metodo kernel di machine learning che viene discusso nel contesto delle macchine a vettori di supporto (Cap. 12) e più in generale nei Capitoli 5 e 14.
- Nella prima edizione, la discussione sulla stima del tasso di errore nel Capitolo 7 era approssimativa, poiché non distinguevamo chiaramente le nozioni di tasso di errore condizionato (condizionato al training set) e di tasso incondizionato. Abbiamo risolto questo problema nella nuova edizione.
- I Capitoli 15 e 16 seguono naturalmente il Capitolo 10, e probabilmente è meglio leggerli in quest'ordine.
- Nel Capitolo 17 non abbiamo affrontato una trattazione completa dei modelli grafici e abbiamo discusso solo i modelli non orientati e alcuni nuovi metodi

per la loro stima. A causa della mancanza di spazio, abbiamo specificamente omesso la copertura dei modelli grafici diretti.

- Il Capitolo 18 esplora il problema "$p >> N$", che è l'apprendimento in spazi caratteristici ad alta dimensionalità. Questi problemi sorgono in molte aree, compresi gli studi genomici e proteomici e la classificazione dei documenti.

Ringraziamo i tanti lettori che hanno individuato gli errori (troppo numerosi) nella prima edizione. Ci scusiamo per questi e abbiamo fatto del nostro meglio per evitarne in questa nuova edizione. Ringraziamo Mark Segal, Bala Rajaratnam e Larry Wasserman per i commenti su alcuni dei nuovi capitoli, e molti studenti laureati e post-dottorato di Stanford che hanno offerto commenti, in particolare Mohammed AlQuraishi, John Boik, Holger Hoefling, Arian Maleki, Donal McMahon, Saharon Rosset, Babak Shababa, Daniela Witten, Ji Zhu e Hui Zou. Ringraziamo John Kimmel per la sua pazienza nel guidarci attraverso questa nuova edizione. RT dedica questa edizione alla memoria di Anna McPhee.

Trevor Hastie
Robert Tibshirani
Jerome Friedman

Stanford, California
Agosto 2008

Prefazione alla Prima Edizione

Stiamo annegando nelle informazioni e siamo affamati di conoscenza.

–Rutherford D. Roger

Il campo della Statistica è costantemente messo alla prova dai problemi che la scienza e l'industria portano al suo ingresso. All'inizio, questi problemi derivavano spesso da esperimenti agricoli e industriali ed erano di portata relativamente piccola. Con l'avvento dei computer e dell'era dell'informazione, i problemi statistici sono esplosi sia in termini di dimensioni che di complessità. Le sfide nei settori dell'archiviazione, dell'organizzazione e della ricerca dei dati hanno portato al nuovo campo del "data mining"; problemi statistici e computazionali in biologia e medicina hanno creato la "bioinformatica". In molti campi vengono generate grandi quantità di dati e il compito dello statistico è quello di dare un senso a tutto ciò: estrarre modelli e tendenze importanti e capire "cosa dicono i dati". Lo chiamiamo *imparare dai dati* (o *inferenza induttiva*, N.d.C.).

Le sfide legate all'apprendimento dai dati hanno portato a una rivoluzione nello stato delle scienze statistiche. Dato che la computazione gioca un ruolo così importante, non sorprende che gran parte di questo nuovo sviluppo sia stato realizzato da ricercatori in altri campi, come l'informatica e l'ingegneria.

I problemi di apprendimento che prendiamo in considerazione possono essere approssimativamente classificati come *supervisionato* o *senza supervisione*. Nell'apprendimento supervisionato, l'obiettivo è prevedere il valore di una misura di risultato sulla base di una serie di misure di input; nell'apprendimento non supervisionato non esiste una misura dei risultati e l'obiettivo è descrivere le associazioni e i modelli tra una serie di misure di input.

Questo libro è il nostro tentativo di riunire molte delle nuove idee importanti nell'apprendimento e di spiegarle in un quadro statistico. Sebbene siano necessari alcuni dettagli matematici, sottolineiamo i metodi e le loro basi concettuali piuttosto che le loro proprietà teoriche. Di conseguenza, speriamo che questo libro piaccia non solo agli statistici, ma anche ai ricercatori e ai professionisti in un'ampia varietà di campi.

Proprio come abbiamo imparato molto dai ricercatori esterni al campo della statistica, il nostro punto di vista statistico può aiutare gli altri a comprendere meglio diversi aspetti dell'apprendimento:

> *Non esiste una vera interpretazione di nulla; l'interpretazione è un veicolo al servizio della comprensione umana. Il valore dell'interpretazione sta nel consentire agli altri di pensare fruttuosamente a un'idea.*
>
> –Andreas Buja

Desideriamo riconoscere il contributo di molte persone alla concezione e al completamento di questo libro. David Andrews, Leo Breiman, Andreas Buja, John Chambers, Bradley Efron, Geoffrey Hinton, Werner Stuetzle e John Tukey hanno fortemente influenzato le nostre carriere. Balasubramanian Narasimhan ci ha dato consigli e aiuto su molti problemi computazionali e ha mantenuto un eccellente ambiente informatico. Shin-Ho Bang ha aiutato nella produzione di numerose figure. Lee Wilkinson ha fornito preziosi consigli sulla produzione del colore. Ilana Belitskaya, Eva Cantoni, Maya Gupta, Michael Jordan, Shanti Gopatam, Radford Neal, Jorge Picazo, Bogdan Popescu, Olivier Renaud, Saharon Rosset, John Storey, Ji Zhu, Mu Zhu, due revisori e molti studenti hanno letto parti del manoscritto e proposto suggerimenti utili. John Kimmel è stato di supporto, paziente e disponibile in ogni fase; MaryAnn Brickner e Frank Ganz hanno guidato uno straordinario team di produzione in Springer. Trevor Hastie desidera ringraziare il dipartimento di statistica dell'Università di Cape Town per la sua ospitalità durante le fasi finali di lavorazione di questo libro. Ringraziamo con gratitudine NSF e NIH per il loro sostegno a questo lavoro. Infine, vorremmo ringraziare le nostre famiglie e i nostri genitori per il loro affetto e sostegno.

Trevor Hastie
Robert Tibshirani
Jerome Friedman

Stanford, California
Maggio 2001

> *I silenziosi statistici hanno cambiato il nostro mondo; non scoprendo nuovi fatti o sviluppi tecnici, ma cambiando il modo in cui ragioniamo, sperimentiamo e formiamo le nostre opinioni…*
>
> –Ian Hacking

Curatori dell'Edizione Italiana

Elia Biganzoli
Professore Ordinario di Statistica medica
Dipartimento di Scienze Biomediche e Cliniche
Università degli Studi di Milano La Statale

Patrizia Boracchi
Professoressa Associata di Statistica medica
Dipartimento di Scienze Biomediche e Cliniche
Università degli Studi di Milano La Statale

Sabrina Gaito
Professoressa Ordinaria di Informatica
Dipartimento di Informatica "Giovanni degli Antoni"
Università degli Studi di Milano La Statale

Federica Nicolussi
Professoressa Associata di Statistica
MOX – Dipartimento di Matematica
Politecnico di Milano

Silvia Salini
Professoressa Ordinaria di Statistica
Dipartimento di Economia, Management e Metodi Quantitativi
Università degli Studi di Milano La Statale

Federico Ambrogi
Professore Ordinario di Statistica medica
Dipartimento di Scienze Cliniche e di Comunità
Università degli Studi di Milano La Statale

Giancarlo Manzi
Professore Associato di Statistica
Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza
Sapienza Università di Roma

Giuseppe Marano
Biostatistico Senior
Dipartimento di Scienze Biomediche e Cliniche
Università degli Studi di Milano La Statale

Matteo Zignani
Professore Associato di Informatica
Dipartimento di Informatica "Giovanni degli Antoni"
Università degli Studi di Milano La Statale

Con la collaborazione di
Giacomo Biganzoli
Università degli Studi di Milano La Statale

Indice generale

1 Introduzione 1

2 Panoramica sull'apprendimento supervisionato. 9

2.1 Introduzione 9
2.2 Tipi di variabili e terminologia 9
2.3 Due semplici approcci alla previsione: minimi quadrati e vicini più prossimi . . 11
2.3.1 Modelli lineari e minimi quadrati 11
2.3.2 Metodi del vicino più prossimo 14
2.3.3 Dai minimi quadrati ai vicini più prossimi 15
2.4 Teoria delle decisioni statistiche 18
2.5 Metodi locali in alte dimensioni 22
2.6 Modelli statistici, apprendimento supervisionato e approssimazione di funzioni 27
2.6.1 Un modello statistico per la distribuzione congiunta $\Pr(X, Y)$ 28
2.6.2 Apprendimento supervisionato 28
2.6.3 Approssimazione di funzioni 29
2.7 Modelli di regressione strutturata 31
2.7.1 Difficoltà del problema 31
2.8 Classi di stimatori ristretti 33
2.8.1 Penalità di smussamento e metodi bayesiani 33
2.8.2 Metodi del kernel e regressione locale 33
2.8.3 Funzioni di base e metodi del dizionario 34
2.9 Selezione del modello e compromesso bias-varianza 35
Note bibliografiche 37
Esercizi 38

3 Metodi lineari per la regressione 41
3.1 Introduzione 41
3.2 Modelli di regressione lineare e minimi quadrati 41
3.2.1 Esempio: cancro alla prostata 47
3.2.2 Il teorema di Gauss-Markov 49
3.2.3 Regressione multipla da regressione semplice univariata 50
3.2.4 Output multipli 53
3.3 Selezione di un sottoinsieme 54
3.3.1 Selezione del miglior sottoinsieme 55
3.3.2 La selezione del modello con le procedure forward (in avanti), backward (all'indietro) e stepwise (per passi) 55
3.3.3 Regressione in avanti per fasi 58
3.3.4. Esempio dei dati sul cancro alla prostata (continuazione) 58
3.4 I metodi di riduzione (shrinkage) 59
3.4.1 La regressione ridge 59
3.4.2 Il metodo lasso 65
3.4.3 Discussione: selezione di sottoinsiemi, regressione ridge e lasso . . . 66
3.4.4 Regressione ad angolo minimo 69
3.5 Metodi che usano le direzioni derivate dagli input 75
3.5.1 Regressione sulle componenti principali 75
3.5.2 Minimi quadrati parziali 76
3.6 Discussione: un confronto tra i metodi di selezione e riduzione 78
3.7 Riduzione e selezione per output multipli 80
3.8 Approfondimento sul lasso e sugli algoritmi correlati di percorso 82
3.8.1 Regressione incrementale in avanti per fasi (incremental forward stagewise regression) 82
3.8.2 Algoritmo per il percorso lineare a tratti 84
3.8.3 Il selettore di Dantzig 85
3.8.4 Lasso raggruppato 86
3.8.5 Ulteriori proprietà del lasso 86
3.8.6 Ottimizzazione del percorso delle coordinate 88
3.9 Considerazioni computazionali 89
Note bibliografiche 89
Esercizi 90

4 Metodi lineari di classificazione 95
4.1 Introduzione 95
4.2 Regressione lineare di una matrice indicatrice 97
4.3 Analisi discriminante lineare 100
4.3.1 Analisi discriminante quadratica 105
4.3.2 Calcoli per LDA 107
4.3.3 Analisi discriminante lineare a ranghi ridotti 107
4.4 Regressione logistica 112
4.4.1 Stimare i modelli di regressione logistica 113
4.4.2 Esempio: cardiopatia in Sudafrica 115
4.4.3 Approssimazioni quadratiche e inferenza 117
4.4.4 Regressione logistica regolarizzata L_1 118

4.4.5 Regressione logistica o LDA? 119
4.5 Iperpiani di separazione 121
4.5.1 Algoritmo di apprendimento del percettrone di Rosenblatt 123
4.5.2 Iperpiani di separazione ottimali 124
Note bibliografiche 127
Esercizi 127

5 Espansione delle funzioni di base e regolarizzazione 131
5.1 Introduzione 131
5.2 Polinomi a tratti e spline 133
5.2.1 Spline cubiche naturali 136
5.2.2 Esempio: cardiopatia in Sudafrica (continuazione) 137
5.2.3 Esempio: riconoscimento del fonema 139
5.3 Filtraggio ed estrazione delle variabili 141
5.4 Smoothing spline 142
5.4.1 Gradi di libertà e matrici smoother 144
5.5 Selezione automatica dei parametri di smussamento 147
5.5.1 Fissare i gradi di libertà 148
5.5.2 Tradeoff bias-varianza 148
5.6 Regressione logistica non parametrica 151
5.7 Spline multidimensionali 152
5.8 Regolarizzazione e riproduzione degli spazi di Hilbert con kernel 157
5.8.1 Spazi di funzioni generati da kernel 158
5.8.2 Esempi di RKHS 160
5.9 Smussamento con le ondine 164
5.9.1 Basi delle ondine e la trasformata a ondine 167
5.9.2 Filtraggio adattivo delle ondine 168
Note bibliografiche 170
Esercizi 171
Appendice: calcoli per le spline 175
B-spline 175
Calcoli per le smoothing spline 177

6 Metodi kernel 179
6.1 Smussatori kernel unidimensionali 180
6.1.1 Regressione lineare locale 182
6.1.2 Regressione polinomiale locale 185
6.2 Selezione della larghezza del kernel 186
6.3 Regressione locale in $\mathbb{R}^p$ 188
6.4 Modelli di regressione locale strutturata in $\mathbb{R}^p$ 189
6.4.1 Kernel strutturati 191
6.4.2 Funzioni di regressione strutturata 191
6.5 Verosimiglianza locale e altri modelli 193
6.6 Stima e classificazione della densità del kernel 196
6.6.1 Stima della densità del kernel 196
6.6.2 Classificazione della densità del kernel 198
6.6.3 Il classificatore di Naive Bayes 198

6.7 Funzioni a base radiale e nuclei . . . 200
6.8 Modelli di misture per la stima e classificazione della densità . . . 202
6.9 Considerazioni computazionali . . . 204
Note bibliografiche . . . 204
Esercizi . . . 204

7 Valutazione e selezione dei modelli . . . 207
7.1 Introduzione . . . 207
7.2 Bias, varianza e complessità del modello . . . 207
7.3 Decomposizione bias-varianza . . . 211
7.3.1 Esempio: tradeoff bias-varianza . . . 214
7.4 Ottimismo dell'errore di training . . . 216
7.5 Stime dell'errore di previsione nel campione . . . 218
7.6 Numero effettivo di parametri . . . 220
7.7 Approccio bayesiano e criterio BIC . . . 221
7.8 Minima lunghezza di descrizione . . . 223
7.9 Dimensione di Vapnik-Chervonenkis . . . 225
7.9.1 Esempio (continuazione) . . . 227
7.10 Cross-validazione . . . 229
7.10.1 Cross-validazione a *K* strati . . . 229
7.10.2 Il modo giusto e il modo sbagliato di fare la cross-validazione . . . 233
7.10.3. La cross-validazione funziona davvero? . . . 235
7.11 Metodi bootstrap . . . 237
7.11.1 Esempio (continuazione) . . . 240
7.12 Errore di test condizionale o atteso? . . . 241
Note bibliografiche . . . 242
Esercizi . . . 245

8 Inferenza e media dei modelli . . . 249
8.1 Introduzione . . . 249
8.2 Bootstrap e metodi di massima verosimiglianza . . . 249
8.2.1 Esempio di smussamento . . . 249
8.2.2 Inferenza tramite massima verosimiglianza . . . 252
8.2.3 Bootstrap e massima verosimiglianza . . . 255
8.3 Metodi bayesiani . . . 255
8.4 Relazione tra bootstrap e inferenza bayesiana . . . 258
8.5 Algoritmi EM . . . 260
8.5.1 Modello di mistura con due componenti . . . 260
8.5.2 L'algoritmo EM in generale . . . 264
8.5.3 EM come procedura di massimizzazione-massimizzazione . . . 266
8.6 MCMC per campionare dalla distribuzione a posteriori . . . 267
8.7 Bagging . . . 270
8.7.1 Esempio: alberi con dati simulati . . . 271
8.8 Model averaging e stacking . . . 276
8.9 Ricerca stocastica: bumping . . . 278
Note bibliografiche . . . 280
Esercizi . . . 280

9 Modelli additivi, alberi e metodi affini 283
9.1 Modelli additivi generalizzati 283
9.1.1 Fitting di modelli additivi 285
9.1.2 Esempio: regressione logistica additiva 287
9.1.3 Riepilogo 292
9.2 Metodi basati sugli alberi 292
9.2.1 Contesto 292
9.2.2 Alberi di regressione 294
9.2.3 Alberi di classificazione 296
9.2.4 Altre problematiche 297
9.2.5 Esempio dello spam (continuazione) 300
9.3 PRIM: bump hunting 304
9.3.1 Esempio dello spam (continuazione) 307
9.4 MARS: Multivariate adaptive regression splines 308
9.4.1 Esempio spam (continuazione) 312
9.4.2 Esempio (dati simulati) 312
9.4.3 Altre problematiche 313
9.5 Misture gerarchiche di esperti 315
9.6 Dati mancanti 317
9.7 Considerazioni computazionali 319
Note bibliografiche 320
Esercizi 320

10 Boosting e alberi additivi 323
10.1 Metodi di boosting 323
10.1.1 Schematizzazione di questo capitolo 326
10.2 Adattamento di un modello additivo con il boosting 327
10.3 Modellazione additiva progressiva 328
10.4 Perdita esponenziale e AdaBoost 329
10.5 Perché la funzione di perdita esponenziale? 330
10.6 Funzioni di perdita e robustezza 332
10.7 Procedure "pronte all'uso" per il data mining 336
10.8 Esempio: dati spam 338
10.9 Alberi con boosting 341
10.10 Ottimizzazione numerica tramite gradient boosting 343
10.10.1 Steepest descent 343
10.10.2 Gradient boosting 344
10.10.3 Implementazioni del gradient boosting 345
10.11 Alberi della giusta dimensione per il boosting 346
10.12 Regolarizzazione 349
10.12.1 Shrinkage 349
10.12.2 Sottocampionamento 350
10.13 Interpretazione 352
10.13.1 Importanza relativa delle variabili predittive 352
10.13.2 Grafici di dipendenza parziale 354
10.14 Esempi 355
10.14.1 California housing 356

10.14.2 New Zealand Fish 359
10.14.3 Dati demografici 364
Note bibliografiche 366
Esercizi 368

11 Reti neurali 373
11.1 Introduzione 373
11.2 Projection pursuit regression 373
11.3 Reti neurali 376
11.4 Addestramento di una rete neurale 379
11.5 Problemi nell'addestramento delle reti neurali 381
11.5.1 Inizializzazione 381
11.5.2 Overfitting 382
11.5.3 Scaling dell'input 382
11.5.4 Numero di unità nascoste e di livelli 384
11.5.5 Minimi molteplici 384
11.6 Esempio: dati simulati 385
11.7 Esempio: codici ZIP 386
11.8 Discussione 392
11.9 Reti neurali bayesiane e competizione NIPS 2003 393
11.9.1 Bayes, boosting e bagging 394
11.9.2 Confronto delle performance 396
11.10 Considerazioni computazionali 398
Note bibliografiche 398
Esercizi 399

12 Support vector machine e discriminanti flessibili 401
12.1 Introduzione 401
12.2 Classificatore support vector 401
12.2.1 Derivazione del classificatore support vector 404
12.2.2 Esempio di mixture (continuazione) 405
12.3 Support vector machine e kernel 407
12.3.1 Addestrare una SVM per la classificazione 407
12.3.2 SVM come un metodo di penalizzazione 410
12.3.3 Stima di funzione e reproducing kernel 412
12.3.4 SVM e il problema della "maledizione della dimensionalità" 413
12.3.5 Un path algorithm per un classificatore SVM 416
12.3.6 SVM per la regressione 418
12.3.7 Regressione e kernel 420
12.3.8 Discussione 421
12.4 Generalizzazione dell'analisi discriminante lineare 422
12.5 Analisi discriminante flessibile 423
12.5.1 Calcolo di FDA 427
12.6 Analisi discriminante con penalità 429
12.7 Analisi discriminante con misture 431
12.7.1 Esempio: forme d'onda 434

Note bibliografiche 437
Esercizi 438

13 Metodi prototipali e vicino più prossimo 441
13.1 Introduzione 441
13.2 Metodi prototipali 441
13.2.1 Clustering *K*-means 442
13.2.2 Apprendimento a quantizzazione vettoriale 444
13.2.3 Misture gaussiane 445
13.3 Classificatori *k*-vicini più prossimi 445
13.3.1 Esempio: uno studio comparativo 450
13.3.2 Esempio: il nearest-neighbors di ordine *k* e la classificazione di immagini 452
13.3.3 Metriche invarianti e distanza di tangente 453
13.4 Metodi nearest-neighbors adattivi 457
13.4.1 Esempio 460
13.4.2 Riduzione dimensionale globale per il metodo nearest-neighbors 461
13.5 Considerazioni computazionali 462
Note bibliografiche 463
Esercizi 463

14 Apprendimento non supervisionato 467
14.1 Introduzione 467
14.2 Regole di associazione 469
14.2.1 Market basket analysis (analisi del paniere di mercato) 470
14.2.2 L'algoritmo Apriori 471
14.2.3 Esempio: market basket analysis 474
14.2.4 Addestramento non supervisionato come addestramento supervisionato 476
14.2.5 Regole di associazione generalizzate 479
14.2.6 Scelta del metodo di addestramento supervisionato 480
14.2.7 Esempio: market basket analysis (continuazione) 481
14.3 Analisi dei cluster 483
14.3.1 Matrici di prossimità 484
14.3.2 Dissomiglianze basate sugli attributi 485
14.3.3 Dissimilarità degli oggetti 486
14.3.4 Algoritmi di clustering 488
14.3.5 Algoritmi combinatori 488
14.3.6 *K*-means 490
14.3.7 Misture gaussiane come clustering *K*-means soft 492
14.3.8 Esempio: dati dei microarray di tumori umani 493
14.3.9 Quantizzazione vettoriale 495
14.3.10 *K*-medoids 496
14.3.11 Problemi pratici 498
14.3.12 Clustering gerarchico 500
14.4 Mappe autorganizzanti 508

14.5 Componenti principali, curve e superfici 514
14.5.1 Componenti principali 514
14.5.3 Curve principali e superfici 521
14.5.3 Clustering spettrale 523
14.5.4 Componenti principali basate sul kernel 526
14.5.5 Componenti principali sparse 529
14.6 Fattorizzazione matriciale non negativa 531
14.6.1 Analisi archetipale 533
14.7 Analisi delle componenti indipendenti e ricerca esplorativa della proiezione 535
14.7.1 Variabili latenti e analisi fattoriale 536
14.7.2 Analisi delle componenti indipendenti 538
14.7.3 Individuazione esplorativa della proiezione 543
14.7.4 Un approccio diretto per l'ICA 545
14.8 Scaling multidimensionale 548
14.9 Riduzione di dimensione non lineare e scaling multidimensionale locale 550
14.10 Algoritmo Google PageRank 554
Note bibliografiche 555
Esercizi 557

15 Random forest 565
15.1 Introduzione 565
15.2 Definizione delle random forest 565
15.3 Dettagli sulle random forest 570
15.3.1 Campioni out-of-bag 570
15.3.2 Misure di importanza delle variabili 571
15.3.3 Grafici di prossimità 571
15.3.4 Random forest e overfitting 573
15.4 Analisi delle random forest 575
15.4.1 Varianza ed effetto di de-correlazione 575
15.4.2 Distorsione (bias) 578
15.4.3 Classificatore adattivo dei vicini più prossimi 578
Note bibliografiche 579
Esercizi 580

16 Apprendimento d'insieme 583
16.1 Introduzione 583
16.2 Boosting e percorsi di regolarizzazione 585
16.2.1 Regressione penalizzata 585
16.2.2 Principio della "scommessa sulla scarsità" 588
16.2.3 Percorsi di regolarizzazione, overfitting e margini 591
16.3 Apprendimento d'insieme 594
16.3.1 Apprendimento di un buon insieme 595
16.3.2 Regole d'insieme 599
Note bibliografiche 601
Esercizi 602

17 Modelli grafici non orientati 603
17.1 Introduzione 603
17.2 Grafi markoviani e loro proprietà 604
17.3 Modelli grafici non orientati per variabili continue 608
17.3.1 Stima dei parametri quando la struttura del grafo è nota 609
17.3.2 Stima della struttura del grafo 613
17.4 Modelli grafici non orientati per variabili discrete 616
17.4.1 Stima dei parametri quando la struttura del grafo è nota 617
17.4.2 Nodi non osservati 619
17.4.3 Stima della struttura del grafo 620
17.4.4 Macchine di Boltzmann ristrette 620
Note bibliografiche 622
Esercizi 623

18 Problemi in grandi dimensioni: $p >> N$ 627
18.1 Quando p è molto più grande di N 627
18.2 Analisi discriminante lineare diagonale e centroidi contratti più prossimi . . 629
18.3 Classificatori lineari con regolarizzazione quadratica 632
18.3.1 Analisi discriminante regolarizzata 634
18.3.2 Regressione logistica con regolarizzazione quadratica 635
18.3.3 Classificatore di tipo support vector machine 635
18.3.4 Selezione delle variabili 636
18.3.5 Scorciatoie computazionali quando $p >> N$ 637
18.4 Classificatori lineari con regolarizzazione L_1 638
18.4.1 Applicazione del lasso alla spettroscopia della massa proteica . . . 642
18.4.2 Lasso fuso per dati funzionali 644
18.5 Classificazione quando le variabili non sono disponibili 646
18.5.1 Esempio: kernel di stringhe e classificazione delle proteine 646
18.5.2 Classificazione e altri modelli che usano kernel di prodotti interni e distanze a coppie 647
18.5.3 Esempio: classificazione di abstract di articoli scientifici 650
18.6 Regressione per grandi dimensioni: componenti principali supervisionate . . . 652
18.6.1 Collegamento ai modelli per variabili latenti 656
18.6.2 Relazione con il metodo dei minimi quadrati parziali 657
18.6.3 Pre-condizionamento per la selezione delle variabili 659
18.7 Valutazione delle variabili e problema dei test multipli 661
18.7.1 Tasso di falsi positivi 664
18.7.2 Punti di taglio asimmetrici e procedura SAM 667
18.7.3 Interpretazione bayesiana del FDR 669
Note bibliografiche 670
Esercizi 671

Bibliografia 675

Indice analitico 697