Prefazione

I dati acquistano speciale valore se opportunamente analizzati e se i risultati vengono interpretati, comunicati in modo chiaro e distribuiti efficientemente ai beneficiari. Solo in tal modo i dati possono essere usati per prendere decisioni con cognizione di causa. La *data science* è l'elemento chiave di questa catena. È una moderna disciplina fondata su principi, tecniche e algoritmi di area statistica, informatica e matematica che ha come obiettivo l'estrazione di informazione, conoscenza e valore dai dati con metodo scientifico. L'informazione si manifesta in relazioni di vario genere, talvolta inattese – i *pattern* – e la conoscenza si concretizza in modelli e nuove ipotesi che caratterizzano il processo generatore dei dati, pur semplificandolo in generale.

Oggigiorno la *data science* è diventata una priorità strategica negli investimenti di numerose aziende e l'offerta di lavoro in posizioni da *data scientist* è almeno un ordine di grandezza superiore a quella di tutte le altre discipline. I rapidi avanzamenti indotti dalla ricerca scientifica e dagli investimenti del settore avranno un impatto epocale che secondo alcune previsioni sarà superiore a quello della rivoluzione industriale e della capillare diffusione dell'energia elettrica del secolo scorso.

Questo libro intende offrire un ampio trattamento dei metodi e delle tecniche degli algoritmi di data science, dai fondamenti della disciplina fino agli algoritmi non supervisionati per la classificazione robusta per prevedere, ad esempio, le propensioni di acquisto di prodotti e servizi di ogni singolo cliente e in generale le informazioni più rilevanti rispetto a metriche di similarità con altri utenti. Queste capacità di profilazione della clientela sono così efficaci che in alcuni casi, sono sufficienti circa 70 'like' Facebook/Instagram di un utente per determinare accuratamente e automaticamente informazioni altamente sensibili sul suo comportamento.

Per questo motivo, la *data science* è divenuta anche oggetto di riflessioni etico-filosofiche e interventi legislativi per regolarne l'applicazione, specie quando l'obiettivo è "imparare" dai dati per dotare la macchina di metodi di ragionamento e comportamento adattivo, nel senso che viene tradizionalmente attribuito all'intelligenza artificiale. Va detto che questo libro non si inoltrerà in questi aspetti, pur fornendo basi essenziali per comprenderne i meccanismi, il senso e la portata.

Questo testo introduce la disciplina della *data science* dalle basi, seguendo l'approccio innovativo del *coding*. Ogni concetto, infatti, è sviscerato ed accompagnato dall'applicazione pratica in MATLAB, uno dei linguaggi di programmazione più diffusi al mondo in questo settore disciplinare, specie tra fisici, ingegneri ed economisti. Lo abbiamo scelto per il nostro libro perché il

suo ambiente di sviluppo classico, che accoglie codice leggibile ed aperto (ossia senza ricorso a pratiche astruse volte ad accelerarne l'esecuzione), garantisce chiarezza didattica ed eccellenti prestazioni nel calcolo numerico e nel processamento di strutture dati vettorizzabili (come le matrici), che danno grande efficienza alle simulazioni e visualizzazioni grafiche 2D/3D.

Il volume non richiede particolari competenze pregresse in quanto assume la conoscenza solo dei concetti di base della statistica descrittiva e inferenziale a livello d'un corso istituzionale universitario. Il libro contiene anche una parte introduttiva relativa a MATLAB, per renderlo utilizzabile da chi non ha esperienza pregressa di programmazione. Ogni metodo è quindi presentato con un esercizio codificato in MATLAB e la relativa soluzione, per facilitare la comprensione di ogni singolo passo della procedura. Vengono poi illustrate applicazioni a problemi reali che mostrano l'utilità operativa dei diversi metodi descritti. Pensiamo che, con queste caratteristiche, il volume sia utile anche agli analisti in posizioni decisionali - dirigenti aziendali, esperti di marketing, operatori finanziari, data manager, gestori di database - con conoscenza di base degli indici statistici e volontà ad applicare con adeguata consapevolezza metodologica le tecniche di analisi dei dati a supporto delle scelte aziendali.

Si discute spesso su quale sia il linguaggio migliore da affiancare allo studio della *data science*: palesemente la bilancia oggi pende verso R e Python, essenzialmente per il fatto di essere *open source* e quindi gratuiti. Abbiamo detto che trasparenza, chiarezza didattica e velocità del codice sono alla base della scelta MATLAB per questo libro, ma ci sono altri agromenti sui quali invitiamo il lettore a riflettere, specie dopo aver sperimentato gli esercizi proposti nel libro.

- La scrittura di codice MATLAB è particolarmente naturale e semplice se confrontata ad R e Python e quindi adatta ai neofiti della disciplina.
- Le sue numerosissime librerie sono state sviluppate da esperti in svariati settori applicativi, che vanno dall'ingegneria aerospaziale all'automotive, dall'analisi delle serie storiche agli algoritmi di analisi testuali.
- L'ambiente di sviluppo offre buona integrazione con gli altri linguaggi di programmazione. Oltre a poter eseguire funzioni Python o R da MATLAB, è possibile tradurre in automatico codice MATLAB in linguaggio C/C++ per aumentare la velocità dell'esecuzione e per facilitarne il *porting* su processori *embedded* utilizzati nella produzione industriale (per esempio Raspberry Pi).
- MATLAB offre un'ambiente per la creazione di interfacce grafiche avanzate e per l'inserimento di controlli come barre di scorrimento, slider, ca-

selle a discesa semplice ed intuitivo. Questo rende lo strumento prezioso a fini didattici.

- MATLAB è integrato in GitHub – un circuito molto usato per lo sviluppo collettivo di progetti software, dotato di servizi web sofisticati per la revisione e condivisione del codice sorgente – e quindi non richiede particolari conoscenze sui servizi di *versionamento del software*.

- Va infine considerato che il codice distribuito da MATLAB offre le garanzie richieste dall'uso in settori critici come quello *embedded*. In altre parole, un data scientist che sviluppa in MATLAB non si deve preoccupare se a sua insaputa sta invocando "software coming with ABSOLUTELY NO WARRANTY".

A questa lista aggiungiamo un elemento personale, essendo noi artefici di un progetto software denominato FSDA – *Flexible Statistics and Data Analysis* – che estende l'utilizzo di MATLAB con centinaia di funzioni che vanno dall'analisi robusta dei dati multidimensionali ad una serie di strumenti ed interfacce grafiche che rendono agevole e automatico l'utilizzo di tecniche avanzate di *data science*. Il toolbox si installa agendo con un semplice 'click' sul pulsante Add-Ons|Get Add-Ons nella scheda Home di MATLAB. Siamo convinti che il lavoro del data scientist possa beneficiare enormemente dall'uso di FSDA, a meno di fingere che i dati raccolti nelle applicazioni reali siano così come ce li immagiamo in un mondo platonico: senza errori, anomalie e aderenti al modello che si siamo prefigurati.

La struttura del libro è la seguente. Nel primo capitolo viene presentato l'ambiente MATLAB senza dare nulla per scontato. Si illustrano i diversi tipi di dati MATLAB (character, string, array, table, struct), le tecniche per gestirli e trattarli nei dati reali e i concetti di base della programmazione (costrutti `if else end` e cicli `for`).

Nel capitolo "Algebra lineare di base" si richiamano concetti basilari di algebra lineare, che costituiscono il presupposto necessario per la data science.

Nel capitolo "La matrice dei dati e le analisi univariate" vengono riepilogati i diversi indici statistici per l'analisi univariata, ponendo attenzione sulla loro implementazione. Viene anche illustrata l'implementazione delle distribuzioni di frequenze, la costruzione degli intervalli di confidenza e la presentazione grafica dei risultati per l'intero campione o per sottogruppi di unità.

Nel capitolo "Variabili casuali: densità e distribuzioni" vengono illustrate le principali distribuzioni teoriche che potrebbero aver generato la distribuzione empirica osservata. Per ogni distribuzione si illustra come calcolare in

MATLAB la funzione di densità, la funzione di ripartizione, i quantili e come generare numeri casuali da questa distribuzione.

Il capitolo "I trattamenti preliminari dei dati" è dedicato al "data preprocessing", che è una fase molto importante dell'analisi poiché consente di eliminare incongruenze ed errori nelle informazioni rilevate che potrebbero inficiare i risultati. Questo lavoro di "pulizia dei dati" (*data cleaning*), deve sempre precedere le elaborazioni statistiche e la costruzione dei modelli successivi. Particolare attenzione in questo testo viene data agli algoritmi di statistica robusta che consentono di resistere alla presenza di valori anomali e/o errori di rilevazione.

Nel capitolo "La relazione tra le variabili quantitative: correlazione e cograduazione" vengono illustrati i metodi per analizzare e testare le relazioni tra variabili quantitative.

I temi trattati nei capitoli successivi sono i seguenti.

1. L'associazione, che studia le relazioni tra le variabili qualitative nominali oppure ordinali.

2. Le rappresentazioni grafiche univariate, bivariate e multidimensionali, che danno una prima idea delle strutture (pattern) presenti nei dati.

3. Gli strumenti avanzati di algebra lineare. La loro conoscenza è necessaria per comprendere le tecniche multivariate di *data science* . L'obiettivo di questo capitolo non è quello di fornire un trattamento completo di tali tecniche e né tanto meno quello di fornire una serie di dimostrazioni matematiche; intendiamo semplicemente gettare le basi per capire i dettagli delle tecniche avanzate dei capitoli successivi.

4. Le distanze e gli indici di similarità, che mettono in luce le differenze e le analogie tra le unità statistiche esaminate e costituisco la premessa necessaria per la segmentazione comportamentale e/o in generale per la profilazione dei dati (*cluster analysis*) e/o per comprendere le tecniche avanzate di riduzione delle dimensioni.

5. L'analisi delle componenti principali, che consente di ridurre la dimensionalità del problema con riferimento alle variabili, così da focalizzare l'attenzione sugli aspetti più importanti. Diversamente da altri testi di *data science* che trattano questi aspetti complessi come "black box" – ossia, si limitano a fornire i comandi per invocare determinate procedure e produrre un certo output a fronte di un certo input – e diversamente dai testi classici di statistica che sono prevalentemente concentrati sugli aspetti matematici, lasciati senza riscontro pratico, nel nostro testo

vengono sviscerate tutte le istruzioni, riga per riga, necessare a replicare ogni passaggio intermedio: riteniamo che questo metodo di studio sia indispensabile per arrivare a "possedere" completamente i dettagli della tecnica statistica. Per coloro che invece già conoscono i dettagli delle diverse tecniche, vengono fornite GUI che consentono di interagire in maniera semplice ed interattiva con l'output automaticamente prodotto.

6. La *cluster analysis*, che consente di individuare gruppi omogenei di unità, trovando applicazioni salienti nella segmentazione del mercato. In questo capitolo si illustra anche come combinare le tecniche precedenti di riduzione delle dimensioni con quelle di clustering.

7. L'analisi delle corrispondenze che consente di visualizzare le relazioni tra le modalità di fenomeni qualitativi presentati in tabelle di contingenza.

Ogni progetto di successo è sempre accompagnato da alcuni presupposti e da un sogno. Il presupposto di questo libro è che gli ingredienti per fare *data science* si limitano a un minimo di strumenti matematici e un pò di programmazione: tutto quello che serve per metterli a frutto è una mente curiosa e la disponibilità a impegnarsi. Il sogno è quello di fornire alle persone curiose gli strumenti idonei a far piacere la materia, illustrare gli elementi di base della programmazione statistica e impostare una carriera da *data scientist* senza troppo sforzo. È consuetudine chiudere con dovuti ringraziamenti.

- Nel nostro caso, sentiamo di rivolgerci con speciale gratitudine a Francesca Perino, Stefano Olivieri e Giovanna Galliano, per aver riconosciuto in noi grande passione per la programmazione statistica e aver incoraggiato e supportato l'uso avanzato di MATLAB dall'interno della loro azienda, Mathworks Italia, coinvolgendo anche colleghi come Paola Vallauri, Fabrizia Grande e Alessio Conte che hanno avuto un ruolo importante nel facilitare l'uso e la diffusione a fini didattici di MATLAB all'Università di Parma.

  Grazie a tutti loro, il nostro gruppo ora beneficia anche di supporto tecnico da parte di specialisti ed esperti sviluppatori della casa madre dell'azienda, a Natick (Boston, US), troppo numerosi per essere menzionati in questa sede. Siamo stupiti di aver trovato in una azianda una tale curiosità per il nostro lavoro (che ha finalità non commerciali), tenacia nella ricerca di soluzioni adatte al nostro caso, creatività delle stesse, e apertura mentale nel cogliere dal nostro lavoro spunti per il miglioramento del loro: caratteristiche queste che evidentemente accomunano i ricercatori in *data science* indipendentemente dalla loro collocazione professionale.

- Vogliamo sinceramente ringraziare Valentin Todorov, ben noto nella comunità R, per aver valutato il nostro progetto FSDA senza preconcetti ed averci aiutato a creare un ponte tra R e MATLAB col solo fine di dimostrare i benefici che quest'ultimo ambiente di sviluppo può offrire a statistici e data scientist. Valentin sa passare con maestria da un linguaggio all'altro (e a molti altri), cogliendo opportunità e proponendo miglioramenti ed obiettivi di indubbio valore strategico per il nostro progetto.

- Questo libro può essere anche letto ad aggiornamento e integrazione (specie in chiave software) di libri precedenti concepiti a supporto della didattica. In particolare, il nostro testo eredita alcuni dataset e problemi dal libro di Sergio Zani e Andrea Cerioli (2007), *Analisi dei dati e data mining per le decisioni aziendali, Giuffrè Editore, Milano.*

  In particolare, va riconosciuto a Sergio Zani il merito di aver fondato un gruppo di data scientist presso l'Università di Parma evidentemente aperto a collaborazioni in ambienti internazionali, come dimostrato dai coautori di questo testo, e ad Andrea Cerioli l'impegno volto al miglioramento di tecniche chiave nella *data science* come il "model-based clustering" robusto e all'esplorazione di temi che sembrano guadagnare spazio in *data science* come l'applicazione della legge di Benford alla validazione dei dati statistici.

Marco Riani
Aldo Corbellini
Fabrizio Laurini
Gianluca Morelli
Domenico Perrotta
Francesca Torti