Prefazione

Data Science e MATLAB

I dati acquistano un valore speciale quando vengono analizzati correttamente e i risultati ottenuti sono interpretati con chiarezza e comunicati in modo efficace ai destinatari. Solo così i dati possono essere utilizzati per prendere decisioni informate e strategiche. La scienza dei dati – a cui ci riferiremo d'ora in poi con il termine inglese, *data science* – rappresenta l'elemento chiave di questa catena. È una disciplina moderna che si basa sui principi, sulle tecniche e sugli algoritmi derivanti da campi quali la statistica, l'informatica e la matematica. Il suo obiettivo fondamentale è l'estrazione di informazione, conoscenza e valore dai dati attraverso un approccio scientifico rigoroso. L'informazione si manifesta sotto forma di relazioni di vario genere, talvolta inattese – i cosiddetti *pattern* – e la conoscenza si concretizza in modelli e nuove ipotesi che descrivono il processo generatore dei dati, pur semplificandolo in modo generale.

Oggigiorno, la *data science* è diventata una priorità strategica negli investimenti di molte aziende, con un'offerta di lavoro per posizioni da *data scientist* che supera di almeno un ordine di grandezza quella di tutte le altre discipline. I rapidi avanzamenti derivanti dalla ricerca scientifica e dagli investimenti settoriali stanno avendo un impatto epocale che, secondo alcune previsioni, supererà quello della rivoluzione industriale e della diffusione capillare dell'energia elettrica nel secolo scorso. Un esempio di queste innovazioni è l'emergere prepotente dei modelli linguistici di grandi dimensioni (LLM, Large Language Models) e di tecnologie come GPT. Nonostante queste straordinarie novità, è fondamentale che i ricercatori mantengano una solida conoscenza degli elementi di base della *data science* da cui questi strumenti avanzati e rivoluzionari sono nati e si sono sviluppati. Senza queste basi, non sarà possibile comprenderne le proprietà e, soprattutto, gestire i risultati che essi forniscono quando chiamati a risolvere problemi in campo matematico-statistico e per l'analisi di dati complessi.

Questo libro intende offrire un ampio trattamento dei metodi e delle tecniche degli algoritmi di *data science*, partendo dai fondamenti della disciplina fino ad arrivare agli algoritmi non supervisionati per la classificazione robusta. Questi metodi possono essere utilizzati, ad esempio, per prevedere le propensioni di acquisto di prodotti e servizi di ogni singolo cliente e, più in generale, per estrarre informazioni rilevanti basate su metriche di similarità con altri utenti. Le capacità di profilazione della clientela sono così avanzate che, in alcuni casi, bastano circa 70 'like' su Facebook o Instagram di un uten-

te per determinare automaticamente e con precisione informazioni altamente sensibili sul suo comportamento.

Per questo motivo, la *data science* è divenuta anche oggetto di riflessioni etico-filosofiche e interventi legislativi volti a regolarne l'applicazione, soprattutto quando l'obiettivo è "imparare" dai dati per dotare le macchine di metodi di ragionamento e comportamenti adattivi, come tradizionalmente attribuito all'intelligenza artificiale. Va sottolineato che questo libro non esplorerà in dettaglio questi aspetti, ma insistiamo ancora sul fatto che fornirà le basi essenziali per comprenderne i meccanismi, il significato e la portata.

Questo testo introduce la disciplina della *data science* partendo dalle basi, seguendo l'approccio innovativo del *coding*. Ogni concetto viene dettagliato e affiancato da applicazioni pratiche in MATLAB, uno dei linguaggi di programmazione più diffusi al mondo in questo settore, soprattutto tra fisici, ingegneri ed economisti. Abbiamo scelto MATLAB per il nostro libro perché il suo ambiente di sviluppo classico, che favorisce codice leggibile e aperto (senza ricorrere a pratiche complesse per l'accelerazione dell'esecuzione), assicura chiarezza didattica ed eccellenti prestazioni nel calcolo numerico. Inoltre, offre un'ottima gestione delle strutture dati vettorizzabili, come le matrici, che migliorano notevolmente l'efficienza delle simulazioni e delle visualizzazioni grafiche 2D/3D. Lo abbiamo inoltre scelto per il suo supporto all'approccio "Open Science", che promuove la trasparenza e la collaborazione, elementi chiave per la crescita della ricerca scientifica e la diffusione dei risultati.

Il volume non richiede particolari competenze pregresse, presupponendo soltanto la conoscenza dei concetti fondamentali della statistica descrittiva e inferenziale a livello di un corso istituzionale universitario. Il libro include anche una sezione introduttiva su MATLAB, rendendolo accessibile anche a chi non ha esperienza pregressa di programmazione. Ogni metodo è presentato con un esercizio programmato in MATLAB e la relativa soluzione, per facilitare la comprensione di ogni singolo passo della procedura. Inoltre, vengono presentate applicazioni a problemi reali che dimostrano l'utilità pratica dei diversi metodi descritti. Crediamo che, grazie a queste caratteristiche, il volume possa essere un valido strumento anche per analisti in posizioni decisionali – dirigenti aziendali, esperti di marketing, operatori finanziari, data manager, gestori di database – che abbiano una conoscenza di base degli indici statistici e la volontà di applicare con consapevolezza metodologica le tecniche di analisi dei dati a supporto delle decisioni aziendali.

Si discute spesso su quale sia il linguaggio migliore da affiancare allo studio della Data Science: attualmente, la bilancia tende verso R e Python, principalmente perché sono gratuiti e "open source". Tuttavia, è importante distinguere

tra il concetto di "open source", che si riferisce alla disponibilità del codice sorgente in forma accessibile e modificabile, e la gratuità. Anche MATLAB offre trasparenza e accessibilità se il codice sviluppato è aperto e comprensibile, piuttosto che compilato e nascosto. MATLAB presenta anche una serie di altri vantaggi che meritano attenzione e riflessione, specialmente dopo aver sperimentato gli esercizi proposti in questo libro.

- La scrittura di codice in MATLAB è particolarmente intuitiva e semplice, soprattutto se confrontata con R e Python, rendendolo ideale per i principianti della disciplina. Il codice prodotto è altamente leggibile e ben strutturato, facilitando la sua manutenzione e l'aggiornamento nel tempo.

- Le librerie di MATLAB sono state sviluppate da esperti di diversi settori applicativi, che spaziano dall'ingegneria aerospaziale all'automotive e dall'analisi delle serie storiche agli algoritmi di analisi testuale, offrendo una vasta gamma di strumenti affidabili.

- L'ambiente di sviluppo di MATLAB è ben integrato con altri linguaggi di programmazione. Oltre alla possibilità di eseguire funzioni Python o R, è possibile convertire codice MATLAB in C/C++ automaticamente, il che migliora la velocità e facilita il *porting* su processori *embedded*, ampiamente utilizzati nella produzione industriale, come quelli con architettura ARM, incluso il Raspberry Pi.

- MATLAB offre un ambiente semplice ed intuitivo per creare interfacce grafiche avanzate e aggiungere controlli come barre di scorrimento, slider e menù a discesa, rendendosi particolarmente utile per scopi didattici.

- È inoltre ben noto che MATLAB eccelle nella visualizzazione dei dati, offrendo strumenti avanzati per la creazione di grafici complessi e la manipolazione interattiva dei dati. Questo rende la verifica e l'interpretazione dei risultati più immediata e intuitiva.

- MATLAB è integrato con GitHub, una piattaforma ampiamente utilizzata per lo sviluppo collaborativo di software, con servizî web avanzati per la revisione e la condivisione del codice sorgente, eliminando la necessità di conoscenze approfondite sui servizi di *versionamento del software*.

- Vale la pena notare che il codice fornito da MATLAB offre le garanzie richieste per l'uso in settori critici come l'embedded. Ciò significa che un

data scientist che sviluppa in MATLAB non deve preoccuparsi di invocare software con improbabili garanzie, come indicato dall'avviso "software coming with ABSOLUTELY NO WARRANTY", che incontriamo spesso nei pacchetti R. In altre parole, MATLAB offre un robusto sistema di supporto ufficiale, comprese documentazioni dettagliate e forum dedicati. Questo può essere particolarmente utile anche per i principianti, e non solo per chi si trova di fronte a problemi complessi che richiedono un supporto tecnico competente.

A questa lista aggiungiamo un elemento personale: siamo gli sviluppatori di un progetto software chiamato FSDA (*Flexible Statistics and Data Analysis*), che estende le funzionalità di MATLAB con centinaia di funzioni che spaziano dall'analisi robusta dei dati multidimensionali a una serie di strumenti e interfacce grafiche che semplificano l'uso di tecniche avanzate di Data Science. Il toolbox può essere installato con un semplice clic sul pulsante Add-Ons|Get Add-Ons nella scheda HOME di MATLAB. Siamo convinti che il lavoro del data scientist possa trarre grande beneficio dall'uso di FSDA, a meno di pretendere che i dati raccolti nelle applicazioni reali siano privi di errori, anomalie e perfettamente conformi al modello ideale che immaginiamo.

Struttura e contenuti del libro

La struttura del libro è essenzialmente divisa in due parti. La prima parte copre i concetti di base e gli strumenti essenziali per l'utilizzo di MATLAB nella *data science*. La seconda parte è costituita da capitoli più specialistici, che offrono un'analisi più approfondita e tecnica dei vari strumenti e dei metodi avanzati nella disciplina.

I capitoli seguenti fanno parte della prima sezione logica, dedicata ai fondamenti della *data science* con MATLAB.

- Il primo introduce l'ambiente MATLAB, senza dare nulla per scontato. Vengono illustrati i diversi tipi di dati in MATLAB (character, string, array, table, struct e dictionary) e le tecniche per gestirli e utilizzarli nei dati reali, oltre ai concetti di base della programmazione, come i costrutti `if else end` e i cicli `for`.

- Nel capitolo "Algebra lineare", vengono richiamati i concetti fondamentali di algebra lineare, che costituiscono il fondamento necessario per la Data Science.

- Nel capitolo "Analisi esplorative dei dati e tabelle pivot", si riepilogano i diversi indici statistici per l'analisi univariata, con particolare attenzione alla loro implementazione. Si illustrano anche le distribuzioni di frequenze, la costruzione degli intervalli di confidenza e la presentazione grafica dei risultati per l'intero campione o per sottogruppi di unità, con un focus specifico sulla costruzione delle tabelle pivot.

- Il capitolo "Importazione dei dati dal mondo web (in tempo reale)" è dedicato a mostrare come caricare in MATLAB dati provenienti dai principali provider mondiali di dati e software del mercato economico-finanziario. È possibile importare dati da GitHub (o da qualsiasi altro sistema di controllo della versione per la gestione del codice sorgente nello sviluppo di software), dalle principali banche dati finanziarie (es. Bloomberg, LSEG), economiche (FRED), dai social media (es. 𝕏) e dagli istituti di statistica (es. ISTAT) in tempo reale con una sola riga di codice. Viene discussa anche l'importazione di dataset di grandi dimensioni e le opzioni avanzate per determinare il formato delle variabili. Poiché i dati finanziari ed economici spesso si presentano sotto forma di serie storiche, questo capitolo introduce anche le `timetable` e gli strumenti per la gestione delle date e per il cambiamento della periodicità della serie.

- Nel capitolo "Variabili casuali: densità e distribuzioni", vengono esplorate le principali distribuzioni teoriche che potrebbero spiegare la distribuzione empirica osservata. Per ciascuna distribuzione, si spiega come calcolare in MATLAB la funzione di densità, la funzione di ripartizione, i quantili e come generare numeri casuali da tale distribuzione.

- Il capitolo "I trattamenti preliminari dei dati" si concentra sul "data preprocessing", una fase cruciale dell'analisi che consente di correggere incongruenze ed errori nelle informazioni raccolte, evitando che questi influenzino negativamente i risultati. Questo processo di pulizia dei dati (*data cleaning*) deve sempre precedere le elaborazioni statistiche e la costruzione dei modelli successivi. In particolare, il testo dedica un'attenzione particolare agli algoritmi di statistica robusta, che permettono di resistere alla presenza di valori anomali e/o errori di rilevazione.

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

I temi trattati nei capitoli successivi, che coprono gli approfondimenti avanzati, sono i seguenti:

- Rappresentazioni grafiche. Si esplorano le rappresentazioni grafiche univariate, bivariate e multidimensionali, offrendo una prima idea delle strutture (pattern) presenti nei dati.
- Associazione. Si studiano le relazioni tra le variabili qualitative nominali e ordinali, fornendo strumenti per valutare le connessioni tra tali variabili.
- Distanze e indici di similarità. Il capitolo mette in evidenza le differenze e le analogie tra le unità statistiche esaminate, fornendo la base necessaria per la segmentazione comportamentale e la profilazione dei dati (*cluster analysis*) oltre alla comprensione delle tecniche avanzate di riduzione delle dimensioni.
- Analisi delle componenti principali. L'analisi delle componenti principali permette di ridurre la dimensionalità del problema rispetto alle variabili, concentrandosi sugli aspetti più importanti. A differenza di altri testi di *data science*, che trattano questi argomenti complessi come "black box" limitandosi a fornire i comandi per ottenere output specifici, o di testi classici di statistica prevalentemente concentrati sugli aspetti matematici senza riscontro pratico, il nostro testo esamina tutte le istruzioni necessarie per replicare ogni passaggio intermedio, riga per riga. Questo metodo è fondamentale per "possedere" completamente i dettagli delle tecniche statistiche. Per coloro già familiari con i dettagli delle tecniche, sono fornite GUI per un'interazione semplice e interattiva con l'output prodotto.
- Analisi delle corrispondenze. Il capitolo consente di visualizzare le relazioni tra le modalità di fenomeni qualitativi presentati in tabelle di contingenza.
- Cluster analysis. Viene illustrato come individuare gruppi omogenei di unità, un processo con applicazioni rilevanti nella segmentazione del mercato. Inoltre, il capitolo descrive come combinare le tecniche di riduzione delle dimensioni con quelle di clustering.
- Analisi delle serie storiche. Il capitolo spiega come modellare, in modo parametrico e/o non parametrico, le serie temporali di natura economica e finanziaria osservate a varie frequenze. Particolare attenzione è dedicata all'analisi di serie storiche reali, dalla serie del riscaldamento globale

a quella delle vendite al dettaglio, fino all'analisi della serie storica del FTSEMIB della Borsa Valori di Milano. Il capitolo copre dagli strumenti di base agli argomenti avanzati di analisi delle serie storiche.

Alla fine di ogni capitolo, abbiamo aggiunto una serie di esercizi di riepilogo che speriamo possano aiutare a comprendere meglio i concetti esposti e a esplorare aspetti aggiuntivi. Le soluzioni degli esercizi, assieme a materiale extra, sono disponibili online nello spazio della casa editrice, accessibile tramite le credenziali associate al libro. Le pagine online dedicate al libro sono organizzate secondo l'indice dei capitoli e sono interattive, permettendo ai lettori di scaricare il materiale integrativo e rispondere a domande per verificare la comprensione dei concetti. Abbiamo incluso nel sito alcune parti delle prime due edizioni del libro per rendere la lettura del testo principale più fluida.

La terza edizione introduce molte novità, a partire dalla compatibilità con MATLAB 2025a e successive versioni. Una delle innovazioni principali è l'integrazione di un "copilot", che utilizza strumenti avanzati di intelligenza artificiale addestrati sulla documentazione specifica di MATLAB per supportare la scrittura automatica del codice. Questi strumenti offrono suggerimenti e assistenza contestuale, beneficiando di un addestramento mirato anziché di un modello LLM (Large Language Model) generalista.

La versione R2025a di MATLAB segna una rivoluzione significativa anche nell'interfaccia e nell'ergonomia del software, con l'introduzione di un nuovo desktop web-based, gestione multitab delle figure, miglioramenti nelle prestazioni e un editor avanzato (supportato, come dicevamo sopra, dall'intelligenza artificiale). Queste trasformazioni sostanziali hanno reso indispensabile la pubblicazione della terza edizione del libro, poiché era essenziale allineare il contenuto con le ultime innovazioni del software, garantendo così ai lettori un'esperienza aggiornata e pertinente.

Ringraziamenti

Ogni progetto di successo nasce da alcuni presupposti e da un sogno. Il presupposto di questo libro è che gli ingredienti per fare *data science* consistono in un minimo di strumenti matematici e un po' di programmazione: tutto ciò che serve è una mente curiosa e la volontà di imparare. Il sogno è quello di fornire alle persone curiose gli strumenti necessari per apprezzare la materia, comprendere gli elementi di base della programmazione statistica e avviare una carriera da *data scientist* senza troppe difficoltà. È consuetudine chiudere con i dovuti ringraziamenti:

- Innanzitutto, ringraziamo tutti gli studenti che hanno segnalato refusi nelle prime due edizioni del libro. La terza edizione beneficia dei preziosi feedback ricevuti. Un ringraziamento particolare va a Angela Borrello, Federico Baio, Giacomo Boschi, Eleonora Sula, Georgiana Flotta e Marika Palme. Tutte le segnalazioni sono consultabili alla pagina `https://github.com/UniprJRC/DSconMATLAB/issues`.

- Esprimiamo speciale gratitudine a Paolo Panarese, Francesca Perino, Stefano Olivieri e Giovanna Galliano per aver riconosciuto la nostra passione per la programmazione statistica e incoraggiato l'uso avanzato di MATLAB all'interno di MathWorks Italia, coinvolgendo anche colleghi come Paola Vallauri, Fabrizia Grande e Alessio Conte, che hanno avuto un ruolo importante nel facilitare l'uso didattico di MATLAB all'Università di Parma.

  Grazie a loro, il nostro gruppo ora beneficia del supporto tecnico di specialisti ed esperti sviluppatori della sede centrale a Natick (Boston, US), troppo numerosi per essere menzionati tutti. Ringraziamo in particolare Rob Purser, Jos Martin e Andy Campbell, che con i loro contributi e dei rispettivi team hanno significativamente incrementato l'efficacia e la produttività del nostro lavoro e del progetto FSDA. Siamo colpiti dalla curiosità aziendale verso il nostro lavoro, dalla tenacia nella ricerca di soluzioni adatte, dalla creatività e apertura mentale nel recepire spunti utili, caratteristiche comuni tra i ricercatori in *data science* indipendentemente dalla loro posizione professionale.

Marco Riani
Aldo Corbellini
Luigi Grossi
Fabrizio Laurini
Gianluca Morelli
Domenico Perrotta
Tommaso Proietti
Francesca Torti