Data Science and Predictive Analytics

Ivo D. Dinov

Data Science and Predictive Analytics

Biomedical and Health Applications using R

Springer

Ivo D. Dinov
University of Michigan–Ann Arbor
Ann Arbor, Michigan, USA

Additional material to this book can be downloaded from http://extras.springer.com.

ISBN 978-3-319-72346-4 ISBN 978-3-319-72347-1 (eBook)
https://doi.org/10.1007/978-3-319-72347-1

Library of Congress Control Number: 2018930887

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

. . . dedicated to my lovely and encouraging wife, Magdalena, my witty and persuasive kids, Anna-Sophia and Radina, my very insightful brother, Konstantin, and my nurturing parents, Yordanka and Dimitar . . .

Foreword

Instructors, formal and informal learners, working professionals, and readers looking to enhance, update, or refresh their interactive data skills and methodological developments may selectively choose sections, chapters, and examples they want to cover in more depth. Everyone who expects to gain new knowledge or acquire computational abilities should review the overall textbook organization before they decide what to cover, how deeply, and in what order. The organization of the chapters in this book reflects an order that may appeal to many, albeit not all, readers.

Chapter 1 (Motivation) presents (1) the DSPA mission and objectives, (2) several driving biomedical challenges including Alzheimer's disease, Parkinson's disease, drug and substance use, and amyotrophic lateral sclerosis, (3) provides demonstrations of brain visualization, neurodegeneration, and genomics computing, (4) identifies the six defining characteristics of big (biomedical and healthcare) data, (5) explains the concepts of *data science* and *predictive analytics*, and (6) sets the DSPA expectations.

Chapter 2 (Foundations of *R*) justifies the use of the statistical programming language *R* and (1) presents the fundamental programming principles; (2) illustrates basic examples of data transformation, generation, ingestion, and export; (3) shows the main mathematical operators; and (4) presents basic data and probability distribution summaries and visualization.

In **Chap. 3 (Managing Data in R)**, we present additional *R* programming details about (1) loading, manipulating, visualizing, and saving *R* Data Structures; (2) present sample-based statistics measuring central tendency and dispersion; (3) explore different types of variables; (4) illustrate scrapping data from public websites; and (5) show examples of cohort-rebalancing.

A detailed discussion of **Visualization** is presented in **Chap. 4** where we (1) show graphical techniques for exposing composition, comparison, and relationships in multivariate data; and (2) present 1D, 2D, 3D, and 4D distributions along with surface plots.

The foundations of **Linear Algebra and Matrix Computing** are shown in **Chap. 5**. We (1) show how to create, interpret, process, and manipulate

second-order tensors (matrices); (2) illustrate variety of matrix operations and their interpretations; (3) demonstrate linear modeling and solutions of matrix equations; and (4) discuss the eigen-spectra of matrices.

Chapter 6 (Dimensionality Reduction) starts with a simple example reducing 2D data to 1D signal. We also discuss (1) matrix rotations, (2) principal component analysis (PCA), (3) singular value decomposition (SVD), (4) independent component analysis (ICA), and (5) factor analysis (FA).

The discussion of machine learning model-based and model-free techniques commences in **Chap. 7 (Lazy Learning – Classification Using Nearest Neighbors)**. In the scope of the k-nearest neighbor algorithm, we present (1) the general concept of divide-and-conquer for splitting the data into training and validation sets, (2) evaluation of model performance, and (3) improving prediction results.

Chapter 8 (Probabilistic Learning: Classification Using Naive Bayes) presents the naive Bayes and linear discriminant analysis classification algorithms, identifies the assumptions of each method, presents the Laplace estimator, and demonstrates step by step the complete protocol for training, testing, validating, and improving the classification results.

Chapter 9 (Decision Tree Divide and Conquer Classification) focuses on decision trees and (1) presents various classification metrics (e.g., entropy, misclassification error, Gini index), (2) illustrates the use of the C5.0 decision tree algorithm, and (3) shows strategies for pruning decision trees.

The use of linear prediction models is highlighted in **Chap. 10 (Forecasting Numeric Data Using Regression Models)**. Here, we present (1) the fundamentals of multivariate linear modeling, (2) contrast regression trees vs. model trees, and (3) present several complete end-to-end predictive analytics examples.

Chapter 11 (Black Box Machine-Learning Methods: Neural Networks and Support Vector Machines) lays out the foundation of Neural Networks as silicon analogues to biological neurons. We discuss (1) the effects of network layers and topology on the resulting classification, (2) present support vector machines (SVM), and (3) demonstrate classification methods for optical character recognition (OCR), iris flowers clustering, Google trends and the stock market prediction, and quantifying quality of life in chronic disease.

Apriori Association Rules Learning is presented in **Chap. 12** where we discuss (1) the foundation of association rules and the Apriori algorithm, (2) support and confidence measures, and (3) present several examples based on grocery shopping and head and neck cancer treatment.

Chapter 13 (k-Means Clustering) presents (1) the basics of machine learning clustering tasks, (2) silhouette plots, (3) strategies for model tuning and improvement, (4) hierarchical clustering, and (5) Gaussian mixture modeling.

General protocols for measuring the performance of different types of classification methods are presented in **Chap. 14 (Model Performance Assessment)**. We discuss (1) evaluation strategies for binary, categorical, and continuous outcomes; (2) confusion matrices quantifying classification and prediction accuracy; (3) visualization of algorithm performance and ROC curves; and (4) introduce the foundations of internal statistical validation.

Chapter 15 (Improving Model Performance) demonstrates (1) strategies for manual and automated model tuning, (2) improving model performance with meta-learning, and (3) ensemble methods based on bagging, boosting, random forest, and adaptive boosting.

Chapter 16 (Specialized Machine Learning Topics) presents some technical details that may be useful for some computational scientists and engineers. There, we discuss (1) data format conversion; (2) SQL data queries; (3) reading and writing XML, JSON, XLSX, and other data formats; (4) visualization of network bioinformatics data; (4) data streaming and on-the-fly stream classification and clustering; (5) optimization and improvement of computational performance; and (6) parallel computing.

The classical approaches for feature selection are presented in **Chap. 17 (Variable/Feature Selection)** where we discuss (1) filtering, wrapper, and embedded techniques, and (2) show the entire protocols from data collection and preparation to model training, testing, evaluation and comparison using recursive feature elimination.

In **Chap. 18 (Regularized Linear Modeling and Controlled Variable Selection)**, we extend the mathematical foundation we presented in Chap. 5 to include *fidelity* and *regularization* terms in the objective function used for model-based inference. Specifically, we discuss (1) computational protocols for handling complex high-dimensional data, (2) model estimation by controlling the false-positive rate of selection of critical features, and (3) derivations of effective forecasting models.

Chapter 19 (BigBig Longitudinal Data Analysis) is focused on interrogating time-varying observations. We illustrate (1) time series analysis, e.g., ARIMA modeling, (2) structural equation modeling (SEM) with latent variables, (3) longitudinal data analysis using linear mixed models, and (4) the generalized estimating equations (GEE) modeling.

Expanding upon the term-frequency and inverse document frequency techniques we saw in Chap. 8, Chap. 20 **(Natural Language Processing/Text Mining)** provides more details about (1) handling unstructured text documents, (2) term frequency (TF) and inverse document frequency (IDF), and (3) the cosine similarity measure.

Chapter 21 (Prediction and Internal Statistical Cross Validation) provides a broader and deeper discussion of method validation, which started in Chap. 14. Here, we present (1) general prediction and forecasting methods, (2) demonstrate internal statistical n-fold cross-validation, and (3) comparison strategies for multiple prediction models.

Chapter 22 (Function Optimization) presents technical details about minimizing objective functions, which are present virtually in any data science oriented inference or evidence-based translational study. Here, we explain (1) constrained and unconstrained cost function optimization, (2) Lagrange multipliers, (3) linear and quadratic programming, (4) general nonlinear optimization, and (5) data denoising.

The last chapter of this textbook is **Chap. 23 (Deep Learning)**. It covers (1) perceptron activation functions, (2) relations between artificial and biological

neurons and networks, (3) neural nets for computing exclusive OR (XOR) and negative AND (NAND) operators, (3) classification of handwritten digits, and (4) classification of natural images.

We compiled a few dozens of biomedical and healthcare case-studies that are used to demonstrate the presented DSPA concepts, apply the methods, and validate the software tools. For example, **Chap. 1** includes high-level driving biomedical challenges including dementia and other neurodegenerative diseases, substance use, neuroimaging, and forensic genetics. **Chapter 3** includes a traumatic brain injury (TBI) case-study, **Chap. 10** described a heart attacks case-study, and **Chap. 11** uses a quality of life in chronic disease data to demonstrate optical character recognition that can be applied to automatic reading of handwritten physician notes. **Chapter 18** presents a predictive analytics Parkinson's disease study using neuroimaging-genetics data. **Chapter 20** illustrates the applications of natural language processing to extract quantitative biomarkers from unstructured text, which can be used to study hospital admissions, medical claims, or patient satisfaction. **Chapter 23** shows examples of predicting clinical outcomes for amyotrophic lateral sclerosis and irritable bowel syndrome cohorts, as well as quantitative and qualitative classification of biological images and volumes. Indeed, these represent just a few examples, and the readers are encouraged to try the same methods, protocols and analytics on other research-derived, clinically acquired, aggregated, secondary-use, or simulated datasets.

The online appendices (http://DSPA.predictive.space) are continuously expanded to provide more details, additional content, and expand the DSPA methods and applications scope. Throughout this textbook, there are cross-references to appropriate chapters, sections, datasets, web services, and live demonstrations (Live Demos). The sequential arrangement of the chapters provides a suggested reading order; however, alternative sorting and pathways covering parts of the materials are also provided. Of course, readers and instructors may further choose their own coverage paths based on specific intellectual interests and project needs.

Preface

Genesis

Since the turn of the twenty-first century, the evidence overwhelming reveals that the rate of increase for the amount of data we collect doubles each 12–14 months (Kryder's law). The growth momentum of the volume and complexity of digital information we gather far outpaces the corresponding increase of computational power, which doubles each 18 months (Moore's law). There is a substantial imbalance between the increase of data inflow and the corresponding computational infrastructure intended to process that data. This calls into question our ability to extract valuable information and actionable knowledge from the mountains of digital information we collect. Nowadays, it is very common for researchers to work with petabytes (PB) of data, $1PB = 10^{15}$ bytes, which may include nonhomologous records that demand unconventional analytics. For comparison, the Milky Way Galaxy has approximately 2×10^{11} stars. If each star represents a byte, then one petabyte of data correspond to 5,000 Milky Way Galaxies.

This data storage-computing asymmetry leads to an explosion of innovative data science methods and disruptive computational technologies that show promise to provide effective (semi-intelligent) decision support systems. Designing, understanding and validating such new techniques require deep within-discipline basic science knowledge, transdisciplinary team-based scientific collaboration, open-scientific endeavors, and a blend of exploratory and confirmatory scientific discovery. There is a pressing demand to bridge the widening gaps between the needs and skills of practicing data scientists, advanced techniques introduced by theoreticians, algorithms invented by computational scientists, models constructed by biosocial investigators, network products and Internet of Things (IoT) services engineered by software architects.

Purpose

The purpose of this book is to provide a sufficient methodological foundation for a number of modern data science techniques along with hands-on demonstration of implementation protocols, pragmatic mechanics of protocol execution, and interpretation of the results of these methods applied on concrete case-studies. Successfully completing the Data Science and Predictive Analytics (DSPA) training materials (http://predictive.space) will equip readers to (1) understand the computational foundations of Big Data Science; (2) build critical inferential thinking; (3) lend a tool chest of *R* libraries for managing and interrogating raw, derived, observed, experimental, and simulated big healthcare datasets; and (4) furnish practical skills for handling complex datasets.

Limitations/Prerequisites

Prior to diving into DSPA, the readers are strongly encouraged to review the prerequisites and complete the self-assessment pretest. Sufficient remediation materials are provided or referenced throughout. The DSPA materials may be used for variety of graduate level courses with durations of 10–30 weeks, with 3–4 instructional credit hours per week. Instructors can refactor and present the materials in alternative orders. The DSPA chapters in this book are organized sequentially. However, the content can be tailored to fit the audience's needs. Learning data science and predictive analytics is not a linear process – many alternative pathways

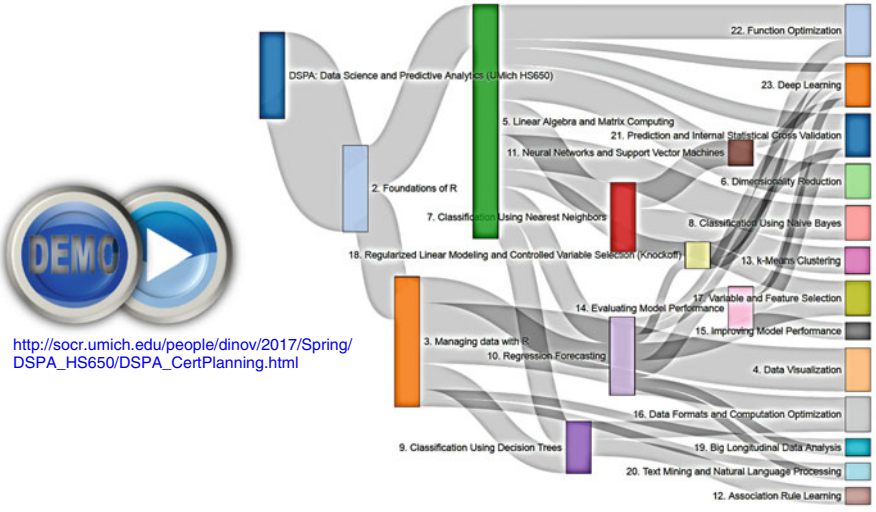

Fig. 1 DSPA topics flowchart

can be completed to gain complementary competencies. We developed an interactive and dynamic flowchart (http://socr.umich.edu/people/dinov/courses/DSPA_Book_FlowChart.html) that highlights several tracks illustrating reasonable pathways starting with Foundations of *R* and ending with specific competency topics. The content of this book may also be used for self-paced learning or as a refresher for working professionals, as well as for formal and informal data science training, including massive open online courses (MOOCs). The DSPA materials are designed to build specific data science skills and predictive analytic competencies, as described by the Michigan Institute for Data Science (MIDAS).

Scope of the Book

Throughout this book, we use a constructive definition of "Big Data" derived by examining the common characteristics of many dozens of biomedical and healthcare case-studies, involving complex datasets that required special handling, advanced processing, contemporary analytics, interactive visualization tools, and translational interpretation. These six characteristics of "Big Data" are defined in the Motivation Chapter as size, heterogeneity and complexity, representation incongruency, incompleteness, multiscale format, and multisource origins. All supporting electronic materials, including datasets, assessment problems, code, software tools, videos, and appendices, are available online at http://DSPA.predictive.space.

This textbook presents a balanced view of the mathematical formulation, computational implementation, and health applications of modern techniques for managing, processing, and interrogating big data. The intentional focus on human health applications is demonstrated by a diverse range of biomedical and healthcare case-studies. However, the same techniques could be applied in other domains, e.g., climate and environmental sciences, biosocial sciences, high-energy physics, astronomy, etc., that deal with complex data possessing the above characteristics. Another specific feature of this book is that it solely utilizes the statistical computing language *R*, rather than any other scripting, user-interface based, or software programming alternatives. The choice for *R* is justified in the Foundations Chapter.

All techniques presented here aim to obtain data-driven and evidence-based scientific inference. This process starts with collecting or retrieving an appropriate dataset, and identifying sources of data that need to be harmonized and aggregated into a joint computable data object. Next, the data are typically split into training and testing components. Model-based or model-free methods are fit, estimated, or learned on the training component and then validated on the complementary testing data. Different types of expected outcomes and results from this process include prediction, prognostication, or forecasting of specific clinical traits (computable phenotypes), clustering, or classification that labels units, subjects, or cases in the data. The final steps include algorithm fine-tuning, assessment, comparison, and statistical validation.

Acknowledgements

The work presented in this textbook relies on deep basic science, as well as holistic interdisciplinary connections developed by scholars, teams of scientists, and transdisciplinary collaborations. Ideas, datasets, software, algorithms, and methods introduced by the wider scientific community were utilized throughout the DSPA resources. Specifically, methodological and algorithmic contributions from the fields of computer vision, statistical learning, mathematical optimization, scientific inference, biomedical computing, and informatics drove the concept presentations, data-driven demonstrations, and case-study reports. The enormous contributions from the entire *R* statistical computing community were critical for developing these resources. We encourage community contributions to expand the techniques, bolster their scope and applications, enhance the collection of case-studies, optimize the algorithms, and widen the applications to other data-intense disciplines or complex scientific challenges.

The author is profoundly indebted to all of his direct mentors and advisors for nurturing my curiosity, inspiring my studies, guiding the course of my career, and providing constructive and critical feedback throughout. Among these scholars are Gencho Skordev (Sofia University); Kenneth Kuttler (Michigan Tech University); De Witt L. Sumners and Fred Huffer (Florida State University); Jan de Leeuw, Nicolas Christou, and Michael Mega (UCLA); Arthur Toga (USC); and Brian Athey, Patricia Hurn, Kathleen Potempa, Janet Larson, and Gilbert Omenn (University of Michigan).

Many other colleagues, students, researchers, and fellows have shared their expertise, creativity, valuable time, and critical assessment for generating, validating, and enhancing these open-science resources. Among these are Christopher Aakre, Simeone Marino, Jiachen Xu, Ming Tang, Nina Zhou, Chao Gao, Alexandr Kalinin, Syed Husain, Brady Zhu, Farshid Sepehrband, Lu Zhao, Sam Hobel, Hanbo Sun, Tuo Wang, and many others. Many colleagues from the Statistics Online Computational Resource (SOCR), the Big Data for Discovery Science (BDDS) Center, and the Michigan Institute for Data Science (MIDAS) provided encouragement and valuable suggestions.

The development of the DSPA materials was partially supported by the US National Science Foundation (grants 1734853, 1636840, 1416953, 0716055, and 1023115), US National Institutes of Health (grants P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30AG053760), and the Elsie Andresen Fiske Research Fund.

Ann Arbor, MI, USA Ivo D. Dinov

DSPA Application and Use Disclaimer

The Data Science and Predictive Analytics (DSPA) resources are designed to help scientists, trainees, students, and professionals learn the foundation of data science, practical applications, and pragmatics of dealing with concrete datasets, and to experiment in a sandbox of specific case-studies. Neither the author nor the publisher have control over, or make any representation or warranties, expressed or implied, regarding the use of these resources by researchers, users, patients, or their healthcare provider(s), or the use or interpretation of any information stored on, derived, computed, suggested by, or received through any of the DSPA materials, code, scripts, or applications. All users are solely responsible for deriving, interpreting, and communicating any information to (and receiving feedback from) the user's representatives or healthcare provider(s).

Users, their proxies, or representatives (e.g., clinicians) are solely responsible for reviewing and evaluating the accuracy, relevance, and meaning of any information stored on, derived by, generated by, or received through the application of any of the DSPA software, protocols, or techniques. The author and the publisher cannot and do not guarantee said accuracy. *The DSPA resources, their applications, and any information stored on, generated by, or received through them are not intended to be a substitute for professional or expert advice, diagnosis, or treatment. Always seek the advice of a physician or other qualified professional with any questions regarding any real case-study (e.g., medical diagnosis, conditions, prediction, and prognostication). Never disregard professional advice or delay seeking it because of something read or learned through the use of the DSPA material or any information stored on, generated by, or received through the SOCR resources*.

All readers and users acknowledge that the DSPA copyright owners or licensors, in their sole discretion, may from time to time make modifications to the DSPA resources. Such modifications may require corresponding changes to be made in the code, protocols, learning modules, activities, case-studies, and other DSPA materials. Neither the author, publisher, nor licensors shall have any obligation to furnish any maintenance or support services with respect to the DSPA resources.

The DSPA resources are intended for educational purposes only. They are not intended to offer or replace any professional advice nor provide expert opinion. Please speak to qualified professional service providers if you have any specific concerns, case-studies, or questions.

Biomedical, Biosocial, Environmental, and Health Disclaimer

All DSPA information, materials, software, and examples are provided for general education purposes only. Persons using the DSPA data, models, tools, or services for any medical, social, healthcare, or environmental purposes should not rely on accuracy, precision, or significance of the DSPA reported results. While the DSPA resources may be updated periodically, users should independently check against other sources, latest advances, and most accurate peer-reviewed information.

Please consult appropriate professional providers prior to making any lifestyle changes or any actions that may impact those around you, your community, or various real, social, and virtual environments. Qualified and appropriate professionals represent the single best source of information regarding any Biomedical, Biosocial, Environmental, and Health decisions. None of these resources have either explicit or implicit indication of FDA approval!

Any and all liability arising directly or indirectly from the use of the DSPA resources is hereby disclaimed. The DSPA resources are provided "as is" and without any warranty expressed or implied. All direct, indirect, special, incidental, consequential, or punitive damages arising from any use of the DSPA resources or materials contained herein are disclaimed and excluded.

Notations

The following common notations are used throughout this textbook.

Notation	Description
 http://www.socr.umich.edu/people/dinov/courses/DSPA_Topics.html	**A link to an interactive live web demonstration.** Some of these Live Demos require modern Java and JavaScript enabled browsers and Internet access.

`require(ggplot2)`
`# Comments Loading required package: ggplot2`
`Data_R_SAS_SPSS_Pubs <-`
 	
`read.csv('https://umich.edu/data', header=T)`
`df <- data.frame(Data_R_SAS_SPSS_Pubs)`
`# convert to long format`
`df <- melt(df , id.vars = 'Year', variable.name = 'Software')`
`ggplot(data=df, aes(x=Year, y=value, color=Software, group = Software)) +`
`geom_line() + geom`
`## 3 1 a`
…
`## 20 3 c`
`data_long`
`## CaseID Gender Feature Measurement`
`## 1 1 M Age 5.0`
`## 2 2 F Age 6.0`
	R fragments of code, reported results in the output shell, or comments. The complete library of all code presented in the textbook is available in electronic format on the DSPA site.
	Note that:
	"**#**" is used for comments,
	"**##**" indicates *R* textual output,
	the R code is **color-coded** to identify different types of comments, instructions, commands and parameters,
	Output like "**## … ##**" suggests that some of the *R* output is deleted or compressed to save space, and
	indenting is used to visually determine the scope of a method, command, or an expression

$\leftarrow$ *or* $\rightarrow$	In an asymptotic or limiting sense, tending to, convergence, or approaching a value or a limit.
$\ll$ *or* $\gg$	Left hand size is substantially smaller or larger than the right hand side.
$\sim$	Depending on the context, model definition, similar to, approximately equal to, or equivalent (in probability distribution sense).
`package::function`	A standard reference notation to functions members of specific *R* packages.
Case-studies	https://umich.instructure.com/courses/38100/files/folder/Case_Studies
Electronic Materials	http://DSPA.predictive.space
Also see the **Glossary** and the **Index**, located in the end of the book.	

Fig. 2 Common DSPA notations

Contents