Preface

For eons we humans have sought out *rules* or *patterns* that accurately describe how important systems in the world around us work, whether these systems be agricultural, biological, physical, financial, etc. We do this because such rules allow us to understand a system better, accurately predict its future behavior and ultimately, control it. However, the process of finding the ”right” rule that seems to govern a given system has historically been no easy task. For most of our history *data* (glimpses of a given system at work) has been an extremely scarce commodity. Moreover, our ability to *compute*, to try out various rules to see which most accurately represents a phenomenon, has been limited to what we could accomplish by hand. Both of these factors naturally limited the range of phenomena scientific pioneers of the past could investigate and inevitably forced them to use philosophical and/or visual approaches to rule-finding. Today, however, we live in a world awash in data, and have colossal computing power at our fingertips. Because of this, we lucky descendants of the great pioneers can tackle a much wider array of problems and take a much more empirical approach to rule-finding than our forbears could. Machine learning, the topic of this textbook, is a term used to describe a broad (and growing) collection of pattern-finding algorithms designed to properly identify system rules empirically and by leveraging our access to potentially enormous amounts of data and computing power.

In the past decade the user base of machine learning has grown dramatically. From a relatively small circle in computer science, engineering, and mathematics departments the users of machine learning now include students and researchers from every corner of the academic universe, as well as members of industry, data scientists, entrepreneurs, and machine learning enthusiasts. This textbook is the result of a complete tearing down of the standard curriculum of machine learning into its most fundamental components, and a curated re-assembly of those pieces (painstakingly polished and organized) that we feel will most benefit this broadening audience of learners. It contains fresh and intuitive yet rigorous descriptions of the most fundamental concepts necessary to conduct research, build products, and tinker.

Book Overview

The second edition of this text is a complete revision of our first endeavor, with virtually every chapter of the original rewritten from the ground up and eight new chapters of material added, doubling the size of the first edition. Topics from the first edition, from expositions on gradient descent to those on One-versus-All classification and Principal Component Analysis have been reworked and polished. A swath of new topics have been added throughout the text, from derivative-free optimization to weighted supervised learning, feature selection, nonlinear feature engineering, boosting-based cross-validation, and more.

While heftier in size, the intent of our original attempt has remained unchanged: to explain machine learning, from first principles to practical implementation, in the simplest possible terms. A big-picture breakdown of the second edition text follows below.

Part I: Mathematical Optimization (Chapters 2–4)

Mathematical optimization is the workhorse of machine learning, powering not only the tuning of individual machine learning models (introduced in Part II) but also the framework by which we determine appropriate models themselves via cross-validation (discussed in Part III of the text).

In this first part of the text we provide a complete introduction to mathematical optimization, from basic zero-order (derivative-free) methods detailed in Chapter 2 to fundamental and advanced first-order and second-order methods in Chapters 3 and 4, respectively. More specifically this part of the text contains complete descriptions of local optimization, *random search* methodologies, *gradient descent*, and *Newton's method*.

Part II: Linear Learning (Chapters 5–9)

In this part of the text we describe the fundamental components of cost function based machine learning, with an emphasis on linear models.

This includes a complete description of *supervised learning* in Chapters 5–7 including linear regression, two-class, and multi-class classification. In each of these chapters we describe a range of perspectives and popular design choices made when building supervised learners.

In Chapter 8 we similarly describe *unsupervised learning*, and Chapter 9 contains an introduction to fundamental *feature engineering* practices including popular *histogram* features as well as various input normalization schemes, and *feature selection* paradigms.

Part III: Nonlinear Learning (Chapters 10–14)

In the final part of the text we extend the fundamental paradigms introduced in Part II to the general nonlinear setting.

We do this carefully beginning with a basic introduction to nonlinear supervised and unsupervised learning in Chapter 10, where we introduce the motivation, common terminology, and notation of nonlinear learning used throughout the remainder of the text.

In Chapter 11 we discuss how to *automate* the selection of appropriate nonlinear models, beginning with an introduction to *universal approximation*. This naturally leads to detailed descriptions of *cross-validation*, as well as *boosting*, *regularization*, *ensembling*, and *K-folds* cross-validation.

With these fundamental ideas in-hand, in Chapters 12–14 we then dedicate an individual chapter to each of the three popular universal approximators used in machine learning: *fixed-shape kernels*, *neural networks*, and *trees*, where we discuss the strengths, weaknesses, technical eccentricities, and usages of each popular universal approximator.

To get the most out of this part of the book we strongly recommend that Chapter 11 and the fundamental ideas therein are studied and understood before moving on to Chapters 12–14.

Part IV: Appendices

This shorter set of appendix chapters provides a complete treatment on advanced optimization techniques, as well as a thorough introduction to a range of subjects that the readers will need to understand in order to make full use of the text.

Appendix A continues our discussion from Chapters 3 and 4, and describes *advanced first- and second-order optimization techniques*. This includes a discussion of popular extensions of gradient descent, including *mini-batch optimization*, *momentum acceleration*, *gradient normalization*, and the result of combining these enhancements in various ways (producing e.g., the RMSProp and Adam first order algorithms) – and Newton's method – including *regularization* schemes and *Hessian-free* methods.

Appendix B contains a tour of *computational calculus* including an introduction to the derivative/gradient, higher-order derivatives, the Hessian matrix, numerical differentiation, forward and backward (backpropagation) automatic differentiation, and Taylor series approximations.

Appendix C provides a suitable background in *linear and matrix algebra*, including vector/matrix arithmetic, the notions of spanning sets and orthogonality, as well as eigenvalues and eigenvectors.

Readers: How To Use This Book

This textbook was written with first-time learners of the subject in mind, as well as for more knowledgeable readers who yearn for a more intuitive and serviceable treatment than what is currently available today. To make full use of the text one needs only a basic understanding of vector algebra (mathematical functions, vector arithmetic, etc.) and computer programming (for example, basic proficiency with a dynamically typed language like Python). We provide complete introductory treatments of other prerequisite topics including linear algebra, vector calculus, and automatic differentiation in the appendices of the text. Example ″roadmaps,″ shown in Figures 0.1–0.4, provide suggested paths for navigating the text based on a variety of learning outcomes and university courses (ranging from a course on the essentials of machine learning to special topics – as described further under ″Instructors: How to use this Book″ below).

We believe that *intuitive leaps precede intellectual ones*, and to this end defer the use of probabilistic and statistical views of machine learning in favor of a fresh and consistent geometric perspective throughout the text. We believe that this perspective not only permits a more intuitive understanding of individual concepts in the text, but also that it helps establish revealing connections between ideas often regarded as fundamentally distinct (e.g., the logistic regression and Support Vector Machine classifiers, kernels and fully connected neural networks, etc.). We also highly emphasize the importance of *mathematical optimization* in our treatment of machine learning. As detailed in the ″Book Overview″ section above, optimization is the workhorse of machine learning and is fundamental at many levels – from the tuning of individual models to the general selection of appropriate nonlinearities via cross-validation. Because of this a strong understanding of mathematical optimization is requisite if one wishes to deeply understand machine learning, and if one wishes to be able to implement fundamental algorithms.

To this end, we place significant emphasis on the design and implementation of algorithms throughout the text with implementations of fundamental algorithms given in Python. These fundamental examples can then be used as building blocks for the reader to help complete the text's programming exercises, allowing them to ″get their hands dirty″ and ″learn by doing,″ practicing the concepts introduced in the body of the text. While in principle any programming language can be used to complete the text's coding exercises, we highly recommend using Python for its ease of use and large support community. We also recommend using the open-source Python libraries NumPy, autograd, and matplotlib, as well as the Jupyter notebook editor to make implementing and testing code easier. A complete set of installation instructions, datasets, as well as starter notebooks for many exercises can be found at

https://github.com/jermwatt/machine_learning_refined

Instructors: How To Use This Book

Chapter slides associated with this textbook, datasets, along with a large array of instructional interactive `Python` widgets illustrating various concepts throughout the text, can be found on the github repository accompanying this textbook at

`https://github.com/jermwatt/machine_learning_refined`

This site also contains instructions for installing `Python` as well as a number of other free packages that students will find useful in completing the text's exercises.

This book has been used as a basis for a number of machine learning courses at Northwestern University, ranging from introductory courses suitable for undergraduate students to more advanced courses on special topics focusing on optimization and deep learning for graduate students. With its treatment of foundations, applications, and algorithms this text can be used as a primary resource or in fundamental component for courses such as the following.

Machine learning essentials treatment: an introduction to the essentials of machine learning is ideal for undergraduate students, especially those in quarter-based programs and universities where a deep dive into the entirety of the book is not feasible due to time constraints. Topics for such a course can include: gradient descent, logistic regression, Support Vector Machines, One-versus-All and multi-class logistic regression, Principal Component Analysis, K-means clustering, the essentials of feature engineering and selection, cross-validation, regularization, ensembling, bagging, kernel methods, fully connected neural networks, and trees. A recommended roadmap for such a course – including recommended chapters, sections, and corresponding topics – is shown in Figure 0.1.

Machine learning full treatment: a standard machine learning course based on this text expands on the essentials course outlined above both in terms of breadth and depth. In addition to the topics mentioned in the essentials course, instructors may choose to cover Newton's method, Least Absolute Deviations, multi-output regression, weighted regression, the Perceptron, the Categorical Cross Entropy cost, weighted two-class and multi-class classification, online learning, recommender systems, matrix factorization techniques, boosting-based feature selection, universal approximation, gradient boosting, random forests, as well as a more in-depth treatment of fully connected neural networks involving topics such as batch normalization and early-stopping-based regularization. A recommended roadmap for such a course – including recommended chapters, sections, and corresponding topics – is illustrated in Figure 0.2.

Mathematical optimization for machine learning and deep learning: such a course entails a comprehensive description of zero-, first-, and second-order optimization techniques from Part I of the text (as well as Appendix A) including: coordinate descent, gradient descent, Newton's method, quasi-Newton methods, stochastic optimization, momentum acceleration, fixed and adaptive steplength rules, as well as advanced normalized gradient descent schemes (e.g., Adam and RMSProp). These can be followed by an in-depth description of the feature engineering processes (especially standard normalization and PCA-sphering) that speed up (particularly first-order) optimization algorithms. All students in general, and those taking an optimization for machine learning course in particular, should appreciate the fundamental role optimization plays in identifying the ″right″ nonlinearity via the processes of boosting and regulariziation based cross-validation, the principles of which are covered in Chapter 11. Select topics from Chapter 13 and Appendix B – including backpropagation, batch normalization, and forward/backward mode of automatic differentiation – can also be covered. A recommended roadmap for such a course – including recommended chapters, sections, and corresponding topics – is given in Figure 0.3.

Introductory portion of a course on deep learning: such a course is best suitable for students who have had prior exposure to fundamental machine learning concepts, and can begin with a discussion of appropriate first order optimization techniques, with an emphasis on stochastic and mini-batch optimization, momentum acceleration, and normalized gradient schemes such as Adam and RMSProp. Depending on the audience, a brief review of fundamental elements of machine learning may be needed using selected portions of Part II of the text. A complete discussion of fully connected networks, including a discussion of backpropagation and forward/backward mode of automatic differentiation, as well as special topics like batch normalization and early-stopping-based cross-validation, can then be made using Chapters 11, 13, and Appendices A and B of the text. A recommended roadmap for such a course – including recommended chapters, sections, and corresponding topics – is shown in Figure 0.4. Additional recommended resources on topics to complete a standard course on deep learning – like convolutional and recurrent networks – can be found by visiting the text's github repository.

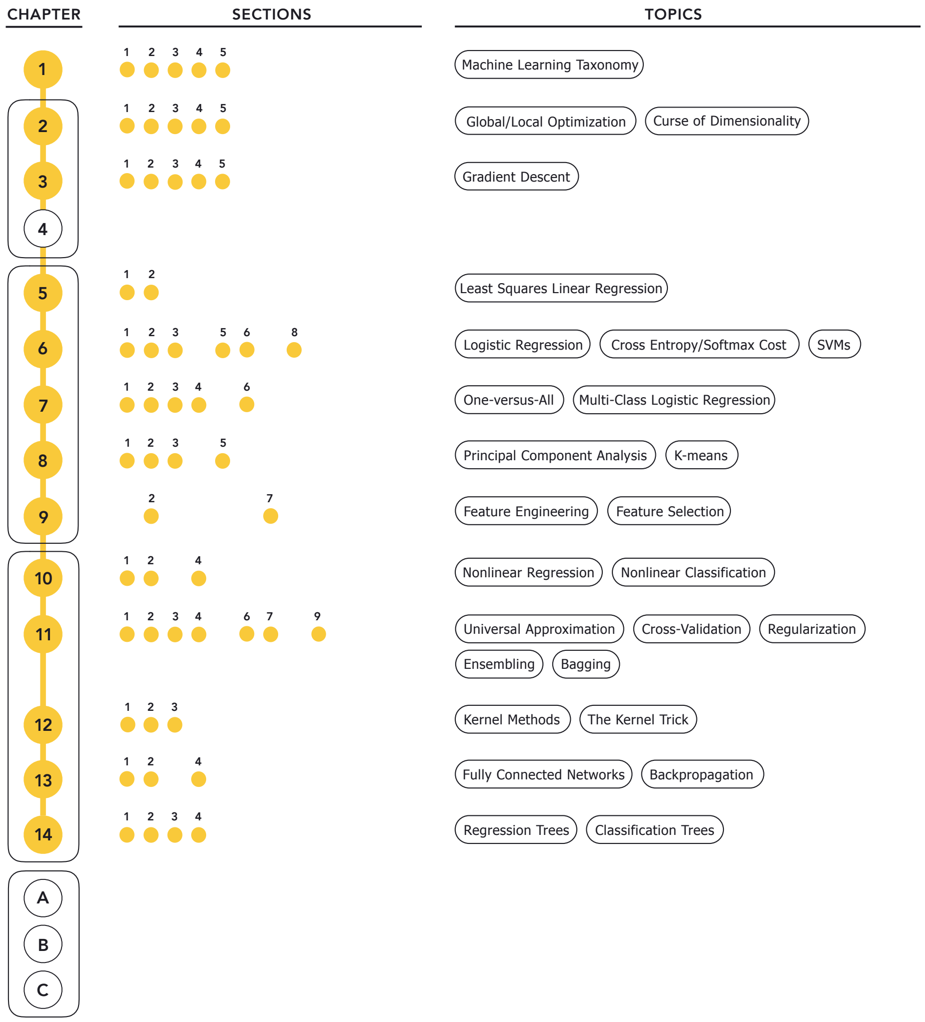

Figure 0.1 Recommended study roadmap for a course on the essentials of machine learning, including requisite chapters (left column), sections (middle column), and corresponding topics (right column). This essentials plan is suitable for time-constrained courses (in quarter-based programs and universities) or self-study, or where machine learning is not the sole focus but a key component of some broader course of study. Note that chapters are grouped together visually based on text layout detailed under ”Book Overview” in the Preface. See the section titled ”Instructors: How To Use This Book” in the Preface for further details.

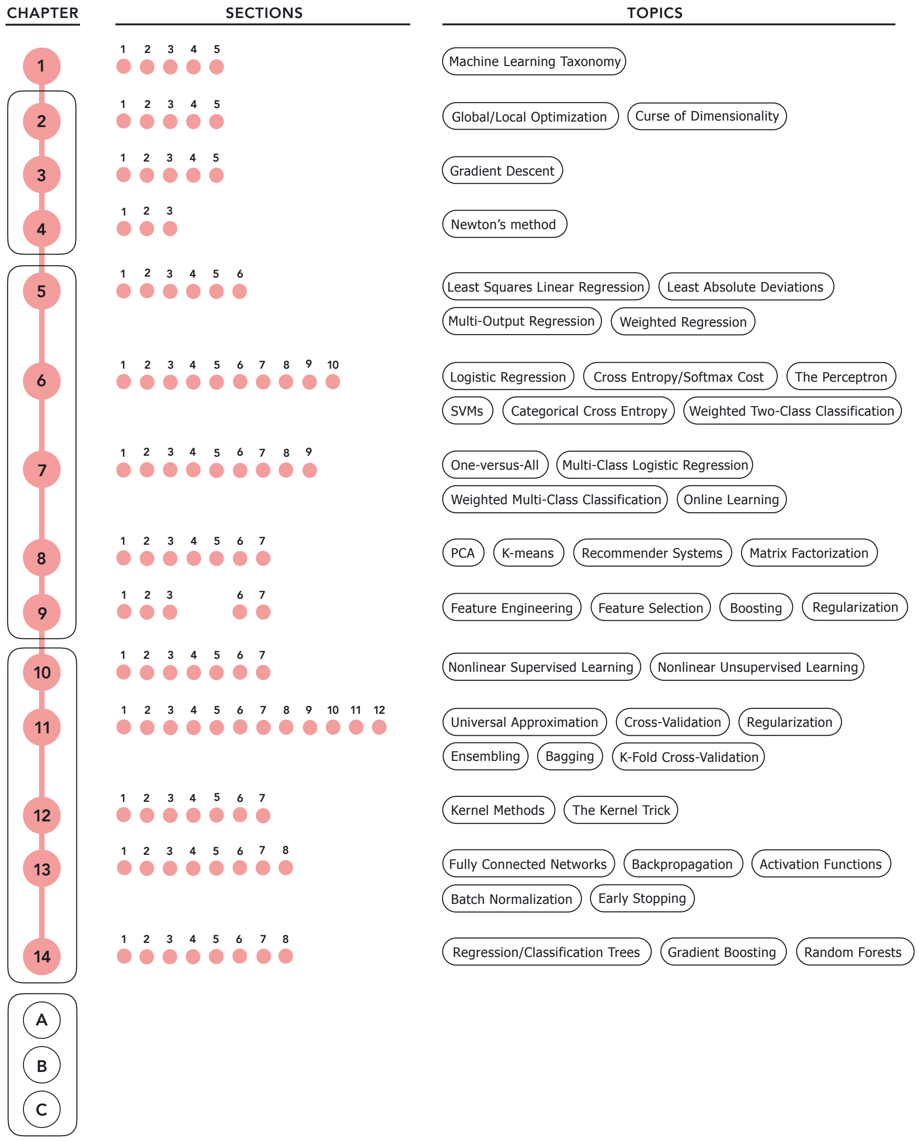

Figure 0.2 Recommended study roadmap for a full treatment of standard machine learning subjects, including chapters, sections, as well as corresponding topics to cover. This plan entails a more in-depth coverage of machine learning topics compared to the essentials roadmap given in Figure 0.1, and is best suited for senior undergraduate/early graduate students in semester-based programs and passionate independent readers. See the section titled ”Instructors: How To Use This Book” in the Preface for further details.

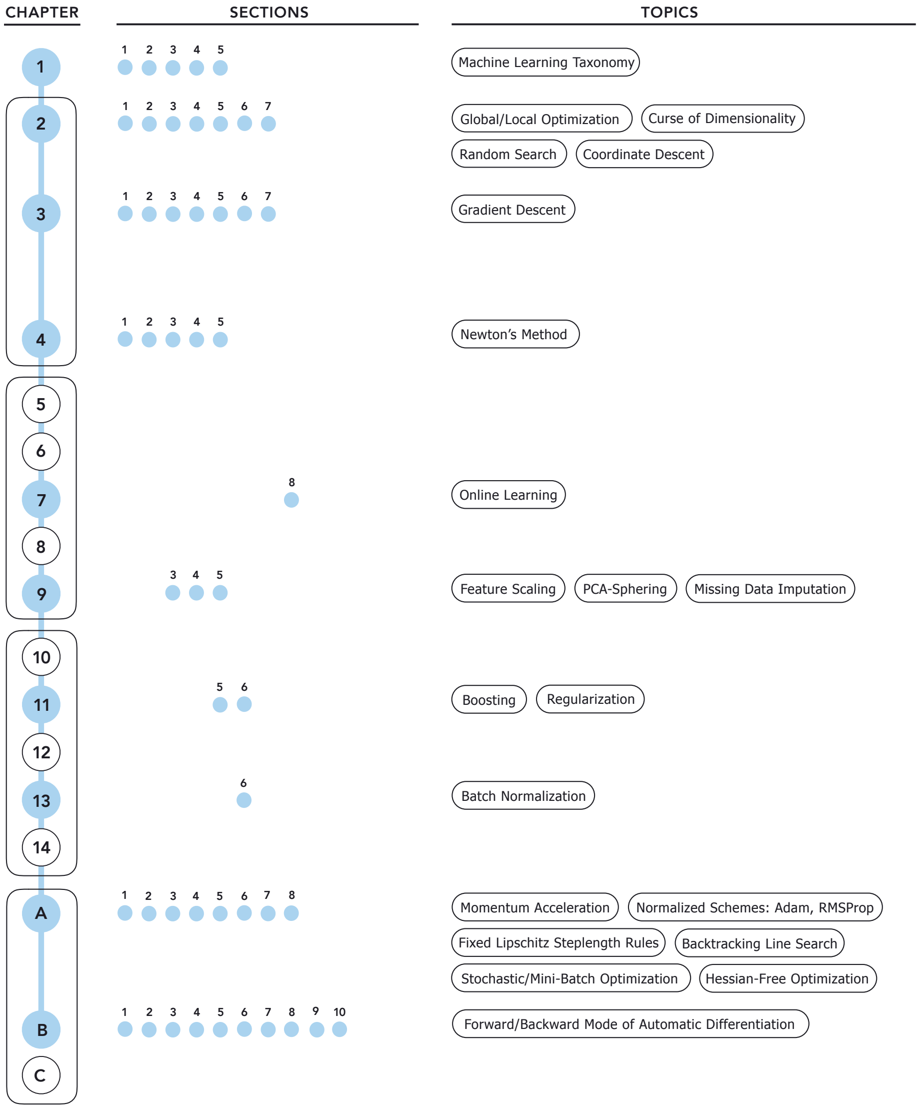

Figure 0.3 Recommended study roadmap for a course on mathematical optimization for machine learning and deep learning, including chapters, sections, as well as topics to cover. See the section titled ”Instructors: How To Use This Book” in the Preface for further details.

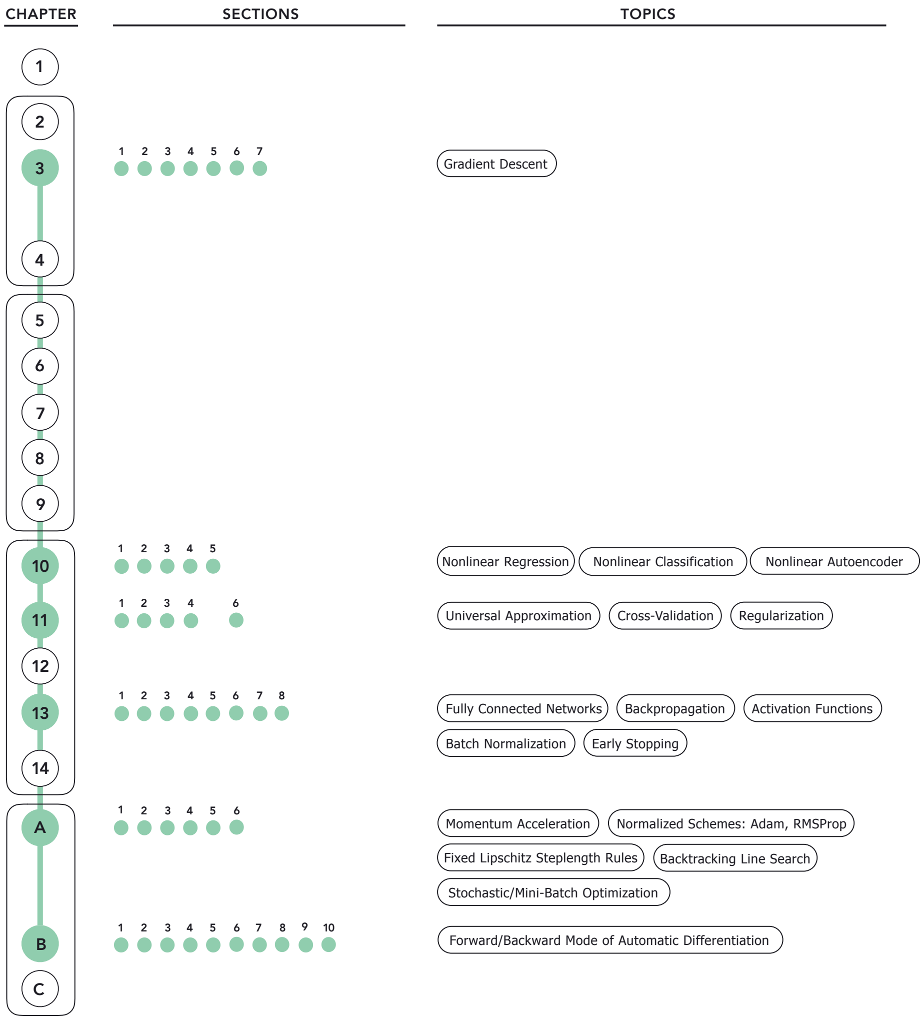

Figure 0.4 Recommended study roadmap for an introductory portion of a course on deep learning, including chapters, sections, as well as topics to cover. See the section titled ”Instructors: How To Use This Book” in the Preface for further details.

Acknowledgements

This text could not have been written in anything close to its current form without the exceedingly gracious assistance of the editorial team at Cambridge University Press including Julie Lancashire, Charlie Howell, Julia Ford, and Lisa Pinto. A very special thanks as well goes to our line editor Beverley Lawrence.

We also owe an enormous debt to the work of the genius-angels in the `Python` open-source community, particularly authors and contributors of `NumPy`, `Jupyter`, and `matplotlib`. We are especially grateful to the authors and contributors of `autograd` including Dougal Maclaurin, David Duvenaud, Matt Johnson, and Jamie Townsend, as `autograd` allowed us to experiment and iterate on a host of new ideas included in the second edition of this text that greatly improved it as well as, we hope, the learning experience for its readers.

We are also very grateful for the many students over the years that provided insightful feedback on the content of this text, with special thanks to Bowen Tian who provided copious amounts of insightful feedback on early drafts of the work.

Finally, a big thanks to Mark McNess Rosengren and the entire Standing Passengers crew for helping us stay caffeinated during the writing of this text.