Preface

Data science has been recognized as a science since 2001, roughly. Its origin lies in technological advances that are generating nearly inconceivable volumes of data. The rate at which new data are being produced is not likely to slow for some time. As a society, we have realized that these data provide opportunities to learn about the systems and processes generating the data. But data in its original form is of relatively little value. Paradoxically, the more of it that there is, the less the value. It has to be reduced to extract value from it. Extracting information from data is the subject of data science.

Becoming a successful practitioner of data science is a real challenge. The knowledge base incorporates demanding topics from statistics, computer science, and mathematics. On top of that, domain-specific knowledge, if not critical, is very helpful. Preparing students in these three or four areas is necessary. But at some point, the subject areas need to be brought together as a coherent package in what we consider to be a course in *data science*. A student that lacks a course that actually teaches data science is not well prepared to practice data science. This book serves as a backbone for a course that brings together the main subject areas.

We've paid attention to the needs of employers with respect to entry-level data scientists—and what they say is lacking from the skills of these new data scientists. What is most lacking are programming abilities. From the educators' point of view, we want to teach principles and theory—the stuff that's needed by students to learn on their own. We're not going to be able to teach them everything they need in their careers, or even in the short term. But teaching principles and foundations is the best preparation for independent learning. Fortuitously, there is a subject that encompasses both principles and programming—algorithms. Therefore, this book has been written about the algorithms of data science.

Algorithms for Data Science focuses on the principles of data reduction and core algorithms for analyzing the data of data science. Understanding the fundamentals is crucial to be able to adapt existing algorithms and create new algorithms. The text provides many opportunities for the reader to develop and improve their programming skills. Every algorithm discussed at length is accompanied by a tutorial that guides the reader through implementation of the algorithm in either `Python` or `R`. The algorithm is then applied to a real-world data set. Using real data allows us to talk about domain-specific problems. Regrettably, our self-imposed coding edict eliminates some important predictive analytic algorithms because of their complexity.

We have two audiences in mind. One audience is practitioners of data science and the allied areas of statistics, mathematics, and computer science. This audience would read the book if they have an interest in improving their analytical skills, perhaps with the objective of working as a data scientist. The second audience are upper-division undergraduate and graduate students in data science, business analytics, mathematics, statistics, and computer science. This audience would be engaged in a course on data analytics or self-study.

Depending on the sophistication of the audience, the book may be used for a one- or two-semester course on data analytics. If used for a one-semester course, the instructor has several options regarding the course content. All options begin with Chaps. 1 and 2 so that the concepts of data reduction and data dictionaries are firmly established.

1. If the instructional emphasis is on computation, then Chaps. 3 and 4 on methods for massively large data and distributed computing would be covered. Chapter 12 works with streaming data, and so this chapter is a nice choice to close the course. Chapter 7 on healthcare analytics is optional and might be covered as time allows. The tutorials of Chap. 7 involve relatively large and challenging data sets. These data sets provide the student and instructor with many opportunities for interesting projects.
2. A course oriented toward general analytical methods might pass over Chaps. 3 and 4 in favor of data visualization (Chap. 5) and linear regression (Chap. 6). The course could close with Chap. 9 on k-nearest neighbor prediction functions and Chap. 11 on forecasting.
3. A course oriented toward predictive analytics would focus on Chaps. 9 and 10 on k-nearest neighbor and naïve Bayes prediction functions. The course would close with Chaps. 11 and 12 on forecasting and streaming data.

Acknowledgments

We thank Brett Kassner, Jason Kolberg, and Greg St. George for reviewing chapters and Guy Shepard for help solving hardware problems and unraveling network mysteries. Many thanks to Alex Philp for anticipating the future and breaking trail. We thank Leonid Kalachev and Peter Golubstov for many interesting conversations and insights.

Missoula, MT, USA Brian Steele
Missoula, MT, USA John Chandler
Missoula, MT, USA Swarna Reddy

http://www.springer.com/978-3-319-45795-6

Algorithms for Data Science

Steele, B.; Chandler, J.; Reddy, S.

2016, XXIII, 430 p. 48 illus., 30 illus. in color., Hardcover
ISBN: 978-3-319-45795-6