Machine Learning

HANDS-ON FOR DEVELOPERS AND TECHNICAL PROFESSIONALS

JASON BELL

WILEY

Machine Learning

Machine Learning

Hands-On for Developers and Technical Professionals

Second Edition

Jason Bell

WILEY

Published simultaneously in Canada

ISBN: 978-1-119-64214-5
ISBN: 978-1-119-64225-1 (ebk)
ISBN: 978-1-119-64219-0 (ebk)

Manufactured in the United States of America

Library of Congress Control Number: 2019956691

To all the developers who just wanted to get the code working without reading all the math stuff first.

About the Author

Jason Bell has worked in software development for more than 30 years. Currently he focuses on large-volume data solutions and helping retail and finance customers gain insight from data with machine learning. He is also an active committee member for several international technology conferences.

About the Technical Editor

Jacob Andresen works as a senior software developer based in Copenhagen, Denmark. He has been working as a software developer and consultant in information retrieval systems and web applications since 2002.

Acknowledgments

"Never again!" I think those were my final words after completing the first edition of this book. Five years later, and here we are again. When the call comes, you immediately think, "Well, it can't be hard, can it?"

To the Team

Jim Minatel, Devon Lewis, Janet Wehner, Pete Gaughan, and the rest of the team at Wiley, thank you for giving your blessing to this second edition and putting your faith in me to revise an awful lot of content. Apologies for the spelling mistakes and those *colour/color* occurrences. Many thanks to Jacob Andresen for giving a technical overview on the content of the book. His enthusiasm for the project was wonderful.

Most Excellent Friends and Collaborators

Dearest friends and acquaintances, thank you: Jennifer Michael, Marie Bentall, Tim Brundle, Stephen Houston, Garrett Murphy, Clare Conway, Tom Spinks, Matt Johnston, Alan Edwards, Colin Mitchell, Simon Hewitt, Mary McKenna, Alan Thorburn, Colin McHale, Dan Lyons, Victoria McCallum, Andrew Bolster, Eoin McFadden, Catherine Muldoon, Amanda Paver, Ben Lorica, Alastair Croll, Mark Madsen, Ellen Friedman, Ted Dunning, Sophia DeMartini, Bruce Durling, Francine Bennett, Michelle Varron, Elise Huard, Antony Woods, John Stephenson, McCraigMcCraig of the Clan McCraig, everyone on the Clojurians Slack Channel, the Strata Data community, Carla Gaggini, Kiki Schirr, Wendy

Devolder, Brian O'Neill, Anthony O'Connor, Tom Gray, Deepa Mann-Kler, Alan Hook, Michelle Douglas, Pete Harwood, Jen Samuel, and Colin Masters. There are loads I've forgotten, I know. I'm sorry.

And Finally

To my wife, Wendy, and my daughter, Clarissa, for absolutely everything and encouraging me to do these projects to the best of my nerdy ability. I couldn't have done it without you both.

To the rest of my family, Maggie, Fern, Andrew, Kerry, Ian and Margaret, William and Sylvia, thank you for all the support and kind words. William, if I need any more help, I'll call you.

The Bios That Never Made It. . .

"He has the boots and jacket that were the envy of many men."

"A dab hand at late-night YouTube videos of 80s pop stars."

"Jason Bell learned to play bass guitar on Saturday afternoons while pretending to work in a music shop."

Thanks to everyone who reads this book. I hope it's helpful in your journey. It's an honor and privilege that you chose to read it. Now I believe it's time for a cup of tea.

Contents

Introduction

Well, times have changed since writing the first edition of this book. Between 2014 and now there is more emphasis on data and what it can do for us but also how that power can be used against us. Hardware has gotten better, processing has gotten much faster, and the ability to classify, predict, and decide based on our data is extraordinary. At the same time, we've become much more aware of the risks of how data is used, the biases that can happen, and that a lot of black-box models don't always get things right.

Still, it's an exciting time to be involved. We still create more data than we can sensibly process. New ideas involving machine learning are being presented daily. The appetite for learning has grown rapidly, too.

Data mining and machine learning have been around a number of years already. When you look closely, the machine learning algorithms that are being applied aren't any different from what they were years ago; what is new is how they are applied at scale. When you look at the number of organizations that are creating the data, it's really, in my opinion, a minority. Google, Facebook, Twitter, Netflix, and a small handful of others are the ones getting the majority of mentions in the headlines with a mixture of algorithmic learning and tools that enable them to scale. So, the real question you should ask is, "How does all this apply to the rest of us?"

Data with large scale, near-instant processing, has come to the fore. The emphasis has moved from batch systems like Hadoop to more streaming-based systems like Kafka. I admit there will be times in this book when I look at the Big Data side of machine learning—it's a subject I can't ignore—but it's only a small factor in the overall picture of how to get insight from the available data. It is important to remember that I am talking about tools, and the key is figuring out which tools are right for the job you are trying to complete.

Aims of This Book

This book is about machine learning and not about Big Data. It's about the various techniques used to gain insight from your data. By the end of the book, you will have seen how various methods of machine learning work, and you will also have had some practical explanations on how the code is put together, leaving you with a good idea of how you could apply the right machine learning techniques to your own problems.

There's no right or wrong way to use this book. You can start at the beginning and work your way through, or you can just dip in and out of the parts you need to know at the time you need to know them.

"Hands-On" Means Hands-On

Many books on the subject of machine learning that I've read in the past have been very heavy on theory. That's not a bad thing. If you're looking for in-depth theory with really complex-looking equations, I applaud your rigor. Me? I'm more hands-on with my approach to learning and to projects. My philosophy is quite simple.

- Start with a question in mind.
- Find the theory I need to learn.
- Find lots of examples I can learn from.
- Put them to work in my own projects.

As a software developer, I like to see lots of examples. As a teacher, I like to get as much hands-on development time as possible but also get the message across to students as simply as possible. There's something about fingers on keys, coding away on your IDE, and getting things to work that's rather appealing, and it's something that I want to convey in the book.

Everyone has his or her own learning styles. I believe this book covers the most common methods, so everybody will benefit.

"What About the Math?"

Like arguing that your favorite football team is better than another or trying to figure out whether Jimmy Page is a better guitarist than Jeff Beck (I prefer Beck), there are some things that will be debated forever and a day. One such debate is how much math you need to know before you can start doing machine learning.