

O'REILLY®

Compliments of



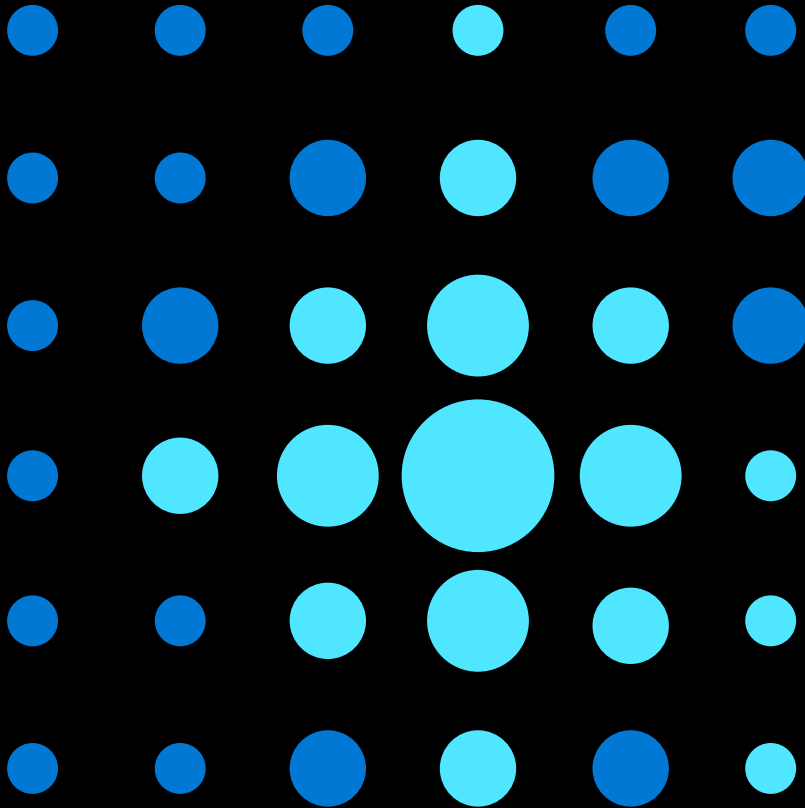
Microsoft
Azure



Thoughtful Machine Learning with Python

A TEST-DRIVEN APPROACH

Matthew Kirk



Build machine learning models easily and quickly

Overcome the complexity of building, training, and deploying machine learning models. Accelerate your path to production, scale on demand, and gain insights from cloud to edge.

Find out more about using Azure Machine Learning service with your favorite open-source tools and frameworks.

[Learn more >](#)

Thoughtful Machine Learning with Python

A Test-Driven Approach

Matthew Kirk

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Thoughtful Machine Learning with Python

by Matthew Kirk

Copyright © 2017 Matthew Kirk. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (<http://oreilly.com/safari>). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editors: Mike Loukides and Shannon Cutt

Production Editor: Nicholas Adams

Copyeditor: James Fraleigh

Proofreader: Charles Roumeliotis

Indexer: Wendy Catalano

Interior Designer: David Futato

Cover Designer: Randy Comer

Illustrator: Rebecca Demarest

January 2017: First Edition

Revision History for the First Edition

2017-01-10: First Release

2017-10-20: Second Release

See <http://oreilly.com/catalog/errata.csp?isbn=9781491924136> for release details.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Thoughtful Machine Learning with Python*, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

This work is part of a collaboration between O'Reilly and Microsoft. See our [statement of editorial independence](#).

978-1-491-08304-7

[LSI]

Table of Contents

Foreword.....	ix
Preface.....	xi
1. Probably Approximately Correct Software.....	1
Writing Software Right	2
SOLID	2
Testing or TDD	4
Refactoring	5
Writing the Right Software	6
Writing the Right Software with Machine Learning	7
What Exactly Is Machine Learning?	7
The High Interest Credit Card Debt of Machine Learning	8
SOLID Applied to Machine Learning	9
Machine Learning Code Is Complex but Not Impossible	12
TDD: Scientific Method 2.0	12
Refactoring Our Way to Knowledge	13
The Plan for the Book	13
2. A Quick Introduction to Machine Learning.....	15
What Is Machine Learning?	15
Supervised Learning	15
Unsupervised Learning	16
Reinforcement Learning	17
What Can Machine Learning Accomplish?	17
Mathematical Notation Used Throughout the Book	18
Conclusion	19
3. K-Nearest Neighbors.....	21
How Do You Determine Whether You Want to Buy a House?	21

How Valuable Is That House?	22
Hedonic Regression	22
What Is a Neighborhood?	23
K-Nearest Neighbors	24
Mr. K's Nearest Neighborhood	25
Distances	25
Triangle Inequality	25
Geometrical Distance	26
Computational Distances	27
Statistical Distances	29
Curse of Dimensionality	31
How Do We Pick K?	32
Guessing K	32
Heuristics for Picking K	33
Valuing Houses in Seattle	35
About the Data	36
General Strategy	36
Coding and Testing Design	36
KNN Regressor Construction	37
KNN Testing	39
Conclusion	42
4. Naive Bayesian Classification.....	43
Using Bayes' Theorem to Find Fraudulent Orders	43
Conditional Probabilities	44
Probability Symbols	44
Inverse Conditional Probability (aka Bayes' Theorem)	46
Naive Bayesian Classifier	47
The Chain Rule	47
Naiveté in Bayesian Reasoning	47
Pseudocount	49
Spam Filter	50
Setup Notes	50
Coding and Testing Design	50
Data Source	51
EmailObject	51
Tokenization and Context	55
SpamTrainer	57
Error Minimization Through Cross-Validation	64
Conclusion	67

5. Decision Trees and Random Forests.....	69
The Nuances of Mushrooms	70
Classifying Mushrooms Using a Folk Theorem	71
Finding an Optimal Switch Point	72
Information Gain	73
GINI Impurity	74
Variance Reduction	75
Pruning Trees	75
Ensemble Learning	76
Writing a Mushroom Classifier	78
Conclusion	86
6. Hidden Markov Models.....	87
Tracking User Behavior Using State Machines	87
Emissions/Observations of Underlying States	89
Simplification Through the Markov Assumption	91
Using Markov Chains Instead of a Finite State Machine	91
Hidden Markov Model	92
Evaluation: Forward-Backward Algorithm	92
Mathematical Representation of the Forward-Backward Algorithm	92
Using User Behavior	93
The Decoding Problem Through the Viterbi Algorithm	96
The Learning Problem	97
Part-of-Speech Tagging with the Brown Corpus	97
Setup Notes	98
Coding and Testing Design	98
The Seam of Our Part-of-Speech Tagger: CorpusParser	99
Writing the Part-of-Speech Tagger	101
Cross-Validating to Get Confidence in the Model	107
How to Make This Model Better	109
Conclusion	109
7. Support Vector Machines.....	111
Customer Happiness as a Function of What They Say	112
Sentiment Classification Using SVMs	112
The Theory Behind SVMs	113
Decision Boundary	114
Maximizing Boundaries	115
Kernel Trick: Feature Transformation	115
Optimizing with Slack	118
Sentiment Analyzer	118
Setup Notes	118

Coding and Testing Design	119
SVM Testing Strategies	120
Corpus Class	120
CorpusSet Class	123
Model Validation and the Sentiment Classifier	126
Aggregating Sentiment	130
Exponentially Weighted Moving Average	130
Mapping Sentiment to Bottom Line	131
Conclusion	132
8. Neural Networks.....	133
What Is a Neural Network?	134
History of Neural Nets	134
Boolean Logic	134
Perceptrons	135
How to Construct Feed-Forward Neural Nets	135
Input Layer	136
Hidden Layers	138
Neurons	139
Activation Functions	140
Output Layer	145
Training Algorithms	145
The Delta Rule	146
Back Propagation	146
QuickProp	147
RProp	147
Building Neural Networks	149
How Many Hidden Layers?	149
How Many Neurons for Each Layer?	150
Tolerance for Error and Max Epochs	150
Using a Neural Network to Classify a Language	151
Setup Notes	151
Coding and Testing Design	151
The Data	152
Writing the Seam Test for Language	152
Cross-Validating Our Way to a Network Class	155
Tuning the Neural Network	158
Precision and Recall for Neural Networks	159
Wrap-Up of Example	159
Conclusion	159

9. Clustering.....	161
Studying Data Without Any Bias	161
User Cohorts	162
Testing Cluster Mappings	164
Fitness of a Cluster	164
Silhouette Coefficient	164
Comparing Results to Ground Truth	165
K-Means Clustering	165
The K-Means Algorithm	165
Downside of K-Means Clustering	167
EM Clustering	167
Algorithm	168
The Impossibility Theorem	169
Example: Categorizing Music	170
Setup Notes	170
Gathering the Data	170
Coding Design	171
Analyzing the Data with K-Means	172
EM Clustering Our Data	173
The Results from the EM Jazz Clustering	178
Conclusion	180
10. Improving Models and Data Extraction.....	181
Debate Club	181
Picking Better Data	182
Feature Selection	182
Exhaustive Search	184
Random Feature Selection	186
A Better Feature Selection Algorithm	186
Minimum Redundancy Maximum Relevance Feature Selection	187
Feature Transformation and Matrix Factorization	189
Principal Component Analysis	189
Independent Component Analysis	190
Ensemble Learning	192
Bagging	193
Boosting	193
Conclusion	195
11. Putting It Together: Conclusion.....	197
Machine Learning Algorithms Revisited	197
How to Use This Information to Solve Problems	199
What's Next for You?	199

Index..... 201

Foreword

Machine learning is not an entirely new subject, but it has gained more popularity in recent years as organizations accelerate development of AI solutions.

Author Matthew Kirk takes readers through the basics of machine learning, with topics such as neural networks, K-Nearest Neighbors (KNNs), clustering, and other algorithms; applying test-driven development (TDD); exploring techniques for improving ML models; and more. This practical guide features code examples with Python's NumPy, Pandas, Scikit-Learn, and SciPy data science libraries. Kirk brings these learnings full circle, with references to real-world examples and engaging, hands-on exercises.

While this book is not intended to be an exhaustive introduction to machine learning, it is designed to help the readers learn the fundamentals, understand the various machine learning algorithms and their applications, and develop a framework to build machine learning solutions.

Microsoft designed Azure Machine Learning service to provide a platform to build, train, and deploy machine learning models easily from cloud to edge. We hope you enjoy the book and consider Azure Machine Learning to accelerate your path to developing high-quality models and AI solutions.

— *Bharat Sandhu*
Director, Azure AI Platform
Microsoft

Preface

I wrote the first edition of *Thoughtful Machine Learning* out of frustration over my coworkers' lack of discipline. Back in 2009 I was working on lots of machine learning projects and found that as soon as we introduced support vector machines, neural nets, or anything else, all of a sudden common coding practice just went out the window.

Thoughtful Machine Learning was my response. At the time I was writing 100% of my code in Ruby and wrote this book for that language. Well, as you can imagine, that was a tough challenge, and I'm excited to present a new edition of this book rewritten for Python. I have gone through most of the chapters, changed the examples, and made it much more up to date and useful for people who will write machine learning code. I hope you enjoy it.

As I stated in the first edition, my door is always open. If you want to talk to me for any reason, feel free to drop me a line at matt@matthewkirk.com. And if you ever make it to Seattle, I would love to meet you over coffee.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic

Indicates new terms, URLs, email addresses, filenames, and file extensions.

`Constant width`

Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

Constant width bold

Shows commands or other text that should be typed literally by the user.

Constant width italic

Shows text that should be replaced with user-supplied values or by values determined by context.



This element signifies a general note.

Using Code Examples

Supplemental material (code examples, exercises, etc.) is available for download at <http://github.com/thoughtfulml/examples-in-python>.

This book is here to help you get your job done. In general, if example code is offered with this book, you may use it in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: “*Thoughtful Machine Learning with Python* by Matthew Kirk (O'Reilly). Copyright 2017 Matthew Kirk, 978-1-491-92413-6.”

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at permissions@oreilly.com.

O'Reilly Safari



Safari (formerly Safari Books Online) is a membership-based training and reference platform for enterprise, government, educators, and individuals.

Members have access to thousands of books, training videos, Learning Paths, interactive tutorials, and curated playlists from over 250 publishers, including O'Reilly Media, Harvard Business Review, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Adobe, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe

Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, and Course Technology, among others.

For more information, please visit <http://oreilly.com/safari>.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at <http://bit.ly/thoughtful-machine-learning-with-python>.

To comment or ask technical questions about this book, send email to bookquestions@oreilly.com.

For more information about our books, courses, conferences, and news, see our website at <http://www.oreilly.com>.

Find us on Facebook: <http://facebook.com/oreilly>

Follow us on Twitter: <http://twitter.com/oreilymedia>

Watch us on YouTube: <http://www.youtube.com/oreilymedia>

Acknowledgments

I've waited over a year to finish this book. My diagnosis of testicular cancer and the sudden death of my dad forced me take a step back and reflect before I could come to grips with writing again. Even though it took longer than I estimated, I'm quite pleased with the result.

I am grateful for the support I received in writing this book: everybody who helped me at O'Reilly and with writing the book. Shannon Cutt, my editor, who was a rock and consistently uplifting. Liz Rush, the sole technical reviewer who was able to make it through the process with me. Stephen Elston, who gave helpful feedback. Mike Loukides, for humoring my idea and letting it grow into two published books. Alexey Porotnikov who helped me extensively with the Python coding examples.

I also want to give special thanks to Alexey Porotnikov (<https://github.com/alpo>) who painstakingly helped me convert all these examples from Ruby to Python and also from Python 2 to Python 3. Seriously, thank you!

I'm grateful for friends, most especially Curtis Fanta. We've known each other since we were five. Thank you for always making time for me (and never being deterred by my busy schedule).

To my family. For my nieces Zoe and Darby, for their curiosity and awe. To my brother Jake, for entertaining me with new music and movies. To my mom Carol, for letting me discover the answers, and advising me to take physics (even though I never have). You all mean so much to me.

To the Le family, for treating me like one of their own. Thanks to Liliana for the Lego dates, and Sayone and Alyssa for being bright spirits in my life. For Martin and Han for their continual support and love. To Thanh (Dad) and Kim (Mom) for feeding me more food than I probably should have, and for giving me multimeters and books on opamps. Thanks for being a part of my life.

To my grandma, who kept asking when she was going to see the cover. You're always pushing me to achieve, be it through Boy Scouts or owning a business. Thank you for always being there.

To Sophia, my wife. A year ago, we were in a hospital room while I was pumped full of painkillers...and we survived. You've been the most constant pillar of my adult life. Whenever I take on a big hairy audacious goal (like writing a book), you always put your needs aside and make sure I'm well taken care of. You mean the world to me.

Last, to my dad. I miss your visits and our camping trips to the woods. I wish you were here to share this with me, but I cherish the time we did have together. This book is for you.