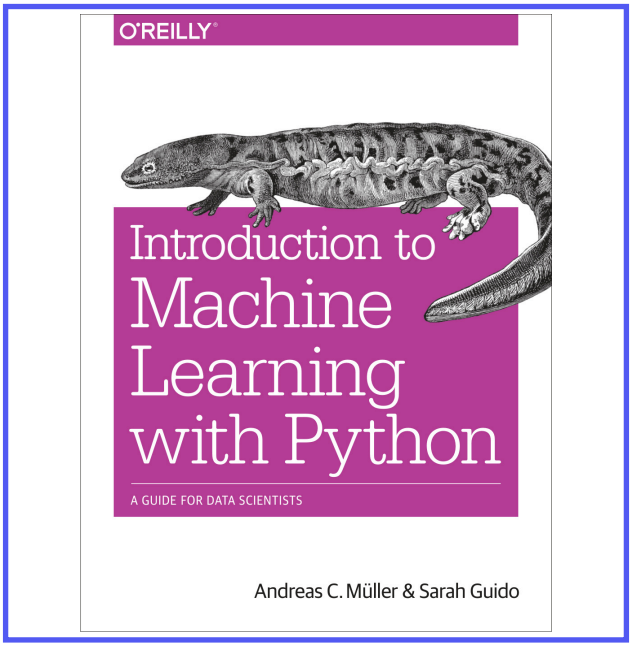


ebookbig.com

or scan the QR code to access the digital study portal.

2. Supervised Learning. 25
Classification and Regression 25
Generalization, Overfitting, and Underfitting 26
Relation of Model Complexity to Dataset Size 29
Supervised Machine Learning Algorithms 29
Some Sample Datasets 30
k-Nearest Neighbors 35
Linear Models 45
Naive Bayes Classifiers 68
Decision Trees 70
Ensembles of Decision Trees 83
Kernelized Support Vector Machines 92
Neural Networks (Deep Learning) 104
Uncertainty Estimates from Classifiers 119
The Decision Function 120
Predicting Probabilities 122
Uncertainty in Multiclass Classification 124
Summary and Outlook 127

3. Unsupervised Learning and Preprocessing. 131
Types of Unsupervised Learning 131
Challenges in Unsupervised Learning 132
Preprocessing and Scaling 132
Different Kinds of Preprocessing 133
Applying Data Transformations 134
Scaling Training and Test Data the Same Way 136
The Effect of Preprocessing on Supervised Learning 138
Dimensionality Reduction, Feature Extraction, and Manifold Learning 140
Principal Component Analysis (PCA) 140
Non-Negative Matrix Factorization (NMF) 156
Manifold Learning with t-SNE 163
Clustering 168
k-Means Clustering 168
Agglomerative Clustering 182
DBSCAN 187
Comparing and Evaluating Clustering Algorithms 191
Summary of Clustering Methods 207
Summary and Outlook 208

4. Representing Data and Engineering Features. 211
Categorical Variables 212
One-Hot-Encoding (Dummy Variables) 213

Numbers Can Encode Categoricals 218
Binning, Discretization, Linear Models, and Trees 220
Interactions and Polynomials 224
Univariate Nonlinear Transformations 232
Automatic Feature Selection 236
Univariate Statistics 236
Model-Based Feature Selection 238
Iterative Feature Selection 240
Utilizing Expert Knowledge 242
Summary and Outlook 250

5. Model Evaluation and Improvement. 251
Cross-Validation 252
Cross-Validation in scikit-learn 253
Benefits of Cross-Validation 254
Stratified k-Fold Cross-Validation and Other Strategies 254
Grid Search 260
Simple Grid Search 261
The Danger of Overfitting the Parameters and the Validation Set 261
Grid Search with Cross-Validation 263
Evaluation Metrics and Scoring 275
Keep the End Goal in Mind 275
Metrics for Binary Classification 276
Metrics for Multiclass Classification 296
Regression Metrics 299
Using Evaluation Metrics in Model Selection 300
Summary and Outlook 302

6. Algorithm Chains and Pipelines. 305
Parameter Selection with Preprocessing 306
Building Pipelines 308
Using Pipelines in Grid Searches 309
The General Pipeline Interface 312
Convenient Pipeline Creation with make_pipeline 313
Accessing Step Attributes 314
Accessing Attributes in a Grid-Searched Pipeline 315
Grid-Searching Preprocessing Steps and Model Parameters 317
Grid-Searching Which Model To Use 319
Summary and Outlook 320

7. Working with Text Data. 323
Types of Data Represented as Strings 323

Example Application: Sentiment Analysis of Movie Reviews 325
Representing Text Data as a Bag of Words 327
Applying Bag-of-Words to a Toy Dataset 329
Bag-of-Words for Movie Reviews 330
Stopwords 334
Rescaling the Data with tf–idf 336
Investigating Model Coefficients 338
Bag-of-Words with More Than One Word (n-Grams) 339
Advanced Tokenization, Stemming, and Lemmatization 344
Topic Modeling and Document Clustering 347
Latent Dirichlet Allocation 348
Summary and Outlook 355

8. Wrapping Up. 357
Approaching a Machine Learning Problem 357
Humans in the Loop 358
From Prototype to Production 359
Testing Production Systems 359
Building Your Own Estimator 360
Where to Go from Here 361
Theory 361
Other Machine Learning Frameworks and Packages 362
Ranking, Recommender Systems, and Other Kinds of Learning 363
Probabilistic Modeling, Inference, and Probabilistic Programming 363
Neural Networks 364
Scaling to Larger Datasets 364
Honing Your Skills 365
Conclusion 366

Index. 367

Preface

Machine learning is an integral part of many commercial applications and research projects today, in areas ranging from medical diagnosis and treatment to finding your friends on social networks. Many people think that machine learning can only be applied by large companies with extensive research teams. In this book, we want to show you how easy it can be to build machine learning solutions yourself, and how to best go about it. With the knowledge in this book, you can build your own system for finding out how people feel on Twitter, or making predictions about global warming. The applications of machine learning are endless and, with the amount of data available today, mostly limited by your imagination.

Who Should Read This Book

This book is for current and aspiring machine learning practitioners looking to implement solutions to real-world machine learning problems. This is an introductory book requiring no previous knowledge of machine learning or artificial intelligence (AI). We focus on using Python and the `scikit-learn` library, and work through all the steps to create a successful machine learning application. The methods we introduce will be helpful for scientists and researchers, as well as data scientists working on commercial applications. You will get the most out of the book if you are somewhat familiar with Python and the NumPy and `matplotlib` libraries.

We made a conscious effort not to focus too much on the math, but rather on the practical aspects of using machine learning algorithms. As mathematics (probability theory, in particular) is the foundation upon which machine learning is built, we won't go into the analysis of the algorithms in great detail. If you are interested in the mathematics of machine learning algorithms, we recommend the book *The Elements of Statistical Learning* (Springer) by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, which is available for free at the authors' website. We will also not describe how to write machine learning algorithms from scratch, and will instead focus on

how to use the large array of models already implemented in `scikit-learn` and other libraries.

Why We Wrote This Book

There are many books on machine learning and AI. However, all of them are meant for graduate students or PhD students in computer science, and they're full of advanced mathematics. This is in stark contrast with how machine learning is being used, as a commodity tool in research and commercial applications. Today, applying machine learning does not require a PhD. However, there are few resources out there that fully cover all the important aspects of implementing machine learning in practice, without requiring you to take advanced math courses. We hope this book will help people who want to apply machine learning without reading up on years' worth of calculus, linear algebra, and probability theory.

Navigating This Book

This book is organized roughly as follows:

- Chapter 1 introduces the fundamental concepts of machine learning and its applications, and describes the setup we will be using throughout the book.
- Chapters 2 and 3 describe the actual machine learning algorithms that are most widely used in practice, and discuss their advantages and shortcomings.
- Chapter 4 discusses the importance of how we represent data that is processed by machine learning, and what aspects of the data to pay attention to.
- Chapter 5 covers advanced methods for model evaluation and parameter tuning, with a particular focus on cross-validation and grid search.
- Chapter 6 explains the concept of pipelines for chaining models and encapsulating your workflow.
- Chapter 7 shows how to apply the methods described in earlier chapters to text data, and introduces some text-specific processing techniques.
- Chapter 8 offers a high-level overview, and includes references to more advanced topics.

While Chapters 2 and 3 provide the actual algorithms, understanding all of these algorithms might not be necessary for a beginner. If you need to build a machine learning system ASAP, we suggest starting with Chapter 1 and the opening sections of Chapter 2, which introduce all the core concepts. You can then skip to "Summary and Outlook" on page 127 in Chapter 2, which includes a list of all the supervised models that we cover. Choose the model that best fits your needs and flip back to read the

section devoted to it for details. Then you can use the techniques in Chapter 5 to evaluate and tune your model.

Online Resources

While studying this book, definitely refer to the `scikit-learn` website for more in-depth documentation of the classes and functions, and many examples. There is also a video course created by Andreas Müller, "Advanced Machine Learning with scikit-learn," that supplements this book. You can find it at *http://bit.ly/advanced_machine_learning_scikit-learn*.

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. Also used for commands and module and package names.

`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 tip or suggestion.



This element signifies a general note.



This icon indicates a warning or caution.

Using Code Examples

Supplemental material (code examples, IPython notebooks, etc.) is available for download at *https://github.com/amueller/introduction_to_ml_with_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: "*An Introduction to Machine Learning with Python* by Andreas C. Müller and Sarah Guido (O'Reilly). Copyright 2017 Sarah Guido and Andreas Müller, 978-1-449-36941-5."

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*.

Safari® Books Online



Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of plans and pricing for enterprise, government, education, and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que,

Peachpit Press, 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, Course Technology, and hundreds more. For more information about Safari Books Online, please visit us online.

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/intro-machine-learning-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/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

Acknowledgments

From Andreas

Without the help and support of a large group of people, this book would never have existed.

I would like to thank the editors, Meghan Blanchette, Brian MacDonald, and in particular Dawn Schanafelt, for helping Sarah and me make this book a reality.

I want to thank my reviewers, Thomas Caswell, Olivier Grisel, Stefan van der Walt, and John Myles White, who took the time to read the early versions of this book and provided me with invaluable feedback—in addition to being some of the cornerstones of the scientific open source ecosystem.

I am forever thankful for the welcoming open source scientific Python community, especially the contributors to `scikit-learn`. Without the support and help from this community, in particular from Gael Varoquaux, Alex Gramfort, and Olivier Grisel, I would never have become a core contributor to `scikit-learn` or learned to understand this package as well as I do now. My thanks also go out to all the other contributors who donate their time to improve and maintain this package.

I'm also thankful for the discussions with many of my colleagues and peers that helped me understand the challenges of machine learning and gave me ideas for structuring a textbook. Among the people I talk to about machine learning, I specifically want to thank Brian McFee, Daniela Huttenkoppen, Joel Nothman, Gilles Louppe, Hugo Bowne-Anderson, Sven Kreis, Alice Zheng, Kyunghyun Cho, Pablo Baberas, and Dan Cervone.

My thanks also go out to Rachel Rakov, who was an eager beta tester and proofreader of an early version of this book, and helped me shape it in many ways.

On the personal side, I want to thank my parents, Harald and Margot, and my sister, Miriam, for their continuing support and encouragement. I also want to thank the many people in my life whose love and friendship gave me the energy and support to undertake such a challenging task.

From Sarah

I would like to thank Meg Blanchette, without whose help and guidance this project would not have even existed. Thanks to Celia La and Brian Carlson for reading in the early days. Thanks to the O'Reilly folks for their endless patience. And finally, thanks to DTS, for your everlasting and endless support.