
Introduction to Machine Learning with Python

A Guide for Data Scientists

Andreas C. Müller and Sarah Guido

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Introduction to Machine Learning with Python

by Andreas C. Müller and Sarah Guido

Copyright © 2017 Sarah Guido, Andreas Müller. 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://safaribooksonline.com>). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Dawn Schanafelt

Production Editor: Kristen Brown

Copyeditor: Rachel Head

Proofreader: Jasmine Kwityn

Indexer: Judy McConville

Interior Designer: David Futato

Cover Designer: Karen Montgomery

Illustrator: Rebecca Demarest

October 2016: First Edition

Revision History for the First Edition

2016-09-22: First Release

See <http://oreilly.com/catalog/errata.csp?isbn=9781449369415> for release details.

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

While the publisher and the authors have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the authors 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.

978-1-449-36941-5

[LSI]

Table of Contents

Preface.....	vii
1. Introduction.....	1
Why Machine Learning?	1
Problems Machine Learning Can Solve	2
Knowing Your Task and Knowing Your Data	4
Why Python?	5
scikit-learn	5
Installing scikit-learn	6
Essential Libraries and Tools	7
Jupyter Notebook	7
NumPy	7
SciPy	8
matplotlib	9
pandas	10
mglearn	11
Python 2 Versus Python 3	12
Versions Used in this Book	12
A First Application: Classifying Iris Species	13
Meet the Data	14
Measuring Success: Training and Testing Data	17
First Things First: Look at Your Data	19
Building Your First Model: k-Nearest Neighbors	20
Making Predictions	22
Evaluating the Model	22
Summary and Outlook	23

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 <code>make_pipeline</code>	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