

O'REILLY®

Third
Edition

Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow

Concepts, Tools, and Techniques
to Build Intelligent Systems

powered by



Aurélien Géron

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

Through a recent series of breakthroughs, deep learning has boosted the entire field of machine learning. Now, even programmers who know close to nothing about this technology can use simple, efficient tools to implement programs capable of learning from data. This bestselling book uses concrete examples, minimal theory, and production-ready Python frameworks (Scikit-Learn, Keras, and TensorFlow) to help you gain an intuitive understanding of the concepts and tools for building intelligent systems.

With this updated third edition, author Aurélien Géron explores a range of techniques, starting with simple linear regression and progressing to deep neural networks. Numerous code examples and exercises throughout the book help you apply what you've learned. Programming experience is all you need to get started.

- Use Scikit-Learn to track an example ML project end to end
- Explore several models, including support vector machines, decision trees, random forests, and ensemble methods
- Exploit unsupervised learning techniques such as dimensionality reduction, clustering, and anomaly detection
- Dive into neural net architectures, including convolutional nets, recurrent nets, generative adversarial networks, autoencoders, diffusion models, and transformers
- Use TensorFlow and Keras to build and train neural nets for computer vision, natural language processing, generative models, and deep reinforcement learning

Aurélien Géron is a machine learning consultant. A former Googler, he led YouTube's video classification team from 2013 to 2016. He was also a founder and CTO of Wifirst from 2002 to 2012, a leading wireless ISP in France, and a founder and CTO of Polyconseil in 2001, a telecom consulting firm.

"An exceptional resource to study machine learning. You will find clear-minded, intuitive explanations, and a wealth of practical tips."

—François Chollet
Author of Keras, author of
Deep Learning with Python

"This book is a great introduction to the theory and practice of solving problems with neural networks; I recommend it to anyone interested in learning about practical ML."

—Pete Warden
Mobile Lead for TensorFlow

DATA SCIENCE / MACHINE LEARNING

US \$79.99

CAN \$99.99

ISBN: 978-1-098-12597-4



THIRD EDITION

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

*Concepts, Tools, and Techniques to
Build Intelligent Systems*

Aurélien Géron

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

by Aurélien Géron

Copyright © 2023 Aurélien Géron. 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 (<https://oreilly.com>). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Acquisitions Editor: Nicole Butterfield

Development Editors: Nicole Taché and
Michele Cronin

Production Editor: Beth Kelly

Copyeditor: Kim Cofer

Proofreader: Rachel Head

Indexer: Potomac Indexing, LLC

Interior Designer: David Futato

Cover Designer: Karen Montgomery

Illustrator: Kate Dullea

March 2017: First Edition

September 2019: Second Edition

October 2022: Third Edition

Revision History for the Third Edition

2022-10-03: First Release

See <https://oreilly.com/catalog/errata.csp?isbn=9781492032649> for release details.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc.

The views expressed in this work are those of the author, and do not represent the publisher's views. 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.

978-1-098-12597-4

[LSI]

Table of Contents

Preface.....	xv
--------------	----

Part I. The Fundamentals of Machine Learning

1. The Machine Learning Landscape.....	3
What Is Machine Learning?	4
Why Use Machine Learning?	5
Examples of Applications	8
Types of Machine Learning Systems	9
Training Supervision	10
Batch Versus Online Learning	17
Instance-Based Versus Model-Based Learning	21
Main Challenges of Machine Learning	27
Insufficient Quantity of Training Data	27
Nonrepresentative Training Data	28
Poor-Quality Data	30
Irrelevant Features	30
Overfitting the Training Data	30
Underfitting the Training Data	33
Stepping Back	33
Testing and Validating	34
Hyperparameter Tuning and Model Selection	34
Data Mismatch	35
Exercises	37

2. End-to-End Machine Learning Project.....	39
Working with Real Data	39
Look at the Big Picture	41
Frame the Problem	41
Select a Performance Measure	43
Check the Assumptions	46
Get the Data	46
Running the Code Examples Using Google Colab	46
Saving Your Code Changes and Your Data	48
The Power and Danger of Interactivity	49
Book Code Versus Notebook Code	50
Download the Data	50
Take a Quick Look at the Data Structure	51
Create a Test Set	55
Explore and Visualize the Data to Gain Insights	60
Visualizing Geographical Data	61
Look for Correlations	63
Experiment with Attribute Combinations	66
Prepare the Data for Machine Learning Algorithms	67
Clean the Data	68
Handling Text and Categorical Attributes	71
Feature Scaling and Transformation	75
Custom Transformers	79
Transformation Pipelines	83
Select and Train a Model	88
Train and Evaluate on the Training Set	88
Better Evaluation Using Cross-Validation	89
Fine-Tune Your Model	91
Grid Search	91
Randomized Search	93
Ensemble Methods	95
Analyzing the Best Models and Their Errors	95
Evaluate Your System on the Test Set	96
Launch, Monitor, and Maintain Your System	97
Try It Out!	100
Exercises	101
 3. Classification.....	 103
MNIST	103
Training a Binary Classifier	106
Performance Measures	107

Measuring Accuracy Using Cross-Validation	107
Confusion Matrices	108
Precision and Recall	110
The Precision/Recall Trade-off	111
The ROC Curve	115
Multiclass Classification	119
Error Analysis	122
Multilabel Classification	125
Multioutput Classification	127
Exercises	129
4. Training Models.....	131
Linear Regression	132
The Normal Equation	134
Computational Complexity	137
Gradient Descent	138
Batch Gradient Descent	142
Stochastic Gradient Descent	145
Mini-Batch Gradient Descent	148
Polynomial Regression	149
Learning Curves	151
Regularized Linear Models	155
Ridge Regression	156
Lasso Regression	158
Elastic Net Regression	161
Early Stopping	162
Logistic Regression	164
Estimating Probabilities	164
Training and Cost Function	165
Decision Boundaries	167
Softmax Regression	170
Exercises	173
5. Support Vector Machines.....	175
Linear SVM Classification	175
Soft Margin Classification	176
Nonlinear SVM Classification	178
Polynomial Kernel	180
Similarity Features	181
Gaussian RBF Kernel	181
SVM Classes and Computational Complexity	183

SVM Regression	184
Under the Hood of Linear SVM Classifiers	186
The Dual Problem	189
Kernelized SVMs	190
Exercises	193
6. Decision Trees.....	195
Training and Visualizing a Decision Tree	195
Making Predictions	197
Estimating Class Probabilities	199
The CART Training Algorithm	199
Computational Complexity	200
Gini Impurity or Entropy?	201
Regularization Hyperparameters	201
Regression	204
Sensitivity to Axis Orientation	206
Decision Trees Have a High Variance	207
Exercises	208
7. Ensemble Learning and Random Forests.....	211
Voting Classifiers	212
Bagging and Pasting	215
Bagging and Pasting in Scikit-Learn	217
Out-of-Bag Evaluation	218
Random Patches and Random Subspaces	219
Random Forests	220
Extra-Trees	220
Feature Importance	221
Boosting	222
AdaBoost	222
Gradient Boosting	226
Histogram-Based Gradient Boosting	230
Stacking	232
Exercises	235
8. Dimensionality Reduction.....	237
The Curse of Dimensionality	238
Main Approaches for Dimensionality Reduction	239
Projection	239
Manifold Learning	241
PCA	243

Preserving the Variance	243
Principal Components	244
Projecting Down to d Dimensions	245
Using Scikit-Learn	246
Explained Variance Ratio	246
Choosing the Right Number of Dimensions	247
PCA for Compression	249
Randomized PCA	250
Incremental PCA	250
Random Projection	252
LLE	254
Other Dimensionality Reduction Techniques	256
Exercises	257
9. Unsupervised Learning Techniques.....	259
Clustering Algorithms: k-means and DBSCAN	260
k-means	263
Limits of k-means	272
Using Clustering for Image Segmentation	273
Using Clustering for Semi-Supervised Learning	275
DBSCAN	279
Other Clustering Algorithms	282
Gaussian Mixtures	283
Using Gaussian Mixtures for Anomaly Detection	288
Selecting the Number of Clusters	289
Bayesian Gaussian Mixture Models	292
Other Algorithms for Anomaly and Novelty Detection	293
Exercises	294

Part II. Neural Networks and Deep Learning

10. Introduction to Artificial Neural Networks with Keras.....	299
From Biological to Artificial Neurons	300
Biological Neurons	301
Logical Computations with Neurons	303
The Perceptron	304
The Multilayer Perceptron and Backpropagation	309
Regression MLPs	313
Classification MLPs	315
Implementing MLPs with Keras	317

Building an Image Classifier Using the Sequential API	318
Building a Regression MLP Using the Sequential API	328
Building Complex Models Using the Functional API	329
Using the Subclassing API to Build Dynamic Models	336
Saving and Restoring a Model	337
Using Callbacks	338
Using TensorBoard for Visualization	340
Fine-Tuning Neural Network Hyperparameters	344
Number of Hidden Layers	349
Number of Neurons per Hidden Layer	350
Learning Rate, Batch Size, and Other Hyperparameters	351
Exercises	353
11. Training Deep Neural Networks.....	357
The Vanishing/Exploding Gradients Problems	358
Glorot and He Initialization	359
Better Activation Functions	361
Batch Normalization	367
Gradient Clipping	372
Reusing Pretrained Layers	373
Transfer Learning with Keras	375
Unsupervised Pretraining	377
Pretraining on an Auxiliary Task	378
Faster Optimizers	379
Momentum	379
Nesterov Accelerated Gradient	381
AdaGrad	382
RMSProp	383
Adam	384
AdaMax	385
Nadam	386
AdamW	386
Learning Rate Scheduling	388
Avoiding Overfitting Through Regularization	392
ℓ_1 and ℓ_2 Regularization	393
Dropout	394
Monte Carlo (MC) Dropout	397
Max-Norm Regularization	399
Summary and Practical Guidelines	400
Exercises	402

12. Custom Models and Training with TensorFlow.....	403
A Quick Tour of TensorFlow	403
Using TensorFlow like NumPy	407
Tensors and Operations	407
Tensors and NumPy	409
Type Conversions	409
Variables	410
Other Data Structures	410
Customizing Models and Training Algorithms	412
Custom Loss Functions	412
Saving and Loading Models That Contain Custom Components	413
Custom Activation Functions, Initializers, Regularizers, and Constraints	415
Custom Metrics	416
Custom Layers	419
Custom Models	422
Losses and Metrics Based on Model Internals	424
Computing Gradients Using Autodiff	426
Custom Training Loops	430
TensorFlow Functions and Graphs	433
AutoGraph and Tracing	435
TF Function Rules	437
Exercises	438
 13. Loading and Preprocessing Data with TensorFlow.....	 441
The tf.data API	442
Chaining Transformations	443
Shuffling the Data	445
Interleaving Lines from Multiple Files	446
Preprocessing the Data	448
Putting Everything Together	449
Prefetching	450
Using the Dataset with Keras	452
The TFRecord Format	453
Compressed TFRecord Files	454
A Brief Introduction to Protocol Buffers	454
TensorFlow Protobufs	456
Loading and Parsing Examples	457
Handling Lists of Lists Using the SequenceExample Protobuf	459
Keras Preprocessing Layers	459
The Normalization Layer	460
The Discretization Layer	463

The CategoryEncoding Layer	463
The StringLookup Layer	465
The Hashing Layer	466
Encoding Categorical Features Using Embeddings	466
Text Preprocessing	471
Using Pretrained Language Model Components	473
Image Preprocessing Layers	474
The TensorFlow Datasets Project	475
Exercises	477
14. Deep Computer Vision Using Convolutional Neural Networks.	479
The Architecture of the Visual Cortex	480
Convolutional Layers	481
Filters	484
Stacking Multiple Feature Maps	485
Implementing Convolutional Layers with Keras	487
Memory Requirements	490
Pooling Layers	491
Implementing Pooling Layers with Keras	493
CNN Architectures	495
LeNet-5	498
AlexNet	499
GoogLeNet	502
VGGNet	505
ResNet	505
Xception	509
SENet	510
Other Noteworthy Architectures	512
Choosing the Right CNN Architecture	514
Implementing a ResNet-34 CNN Using Keras	515
Using Pretrained Models from Keras	516
Pretrained Models for Transfer Learning	518
Classification and Localization	521
Object Detection	523
Fully Convolutional Networks	525
You Only Look Once	527
Object Tracking	530
Semantic Segmentation	531
Exercises	535

15. Processing Sequences Using RNNs and CNNs.	537
Recurrent Neurons and Layers	538
Memory Cells	540
Input and Output Sequences	541
Training RNNs	542
Forecasting a Time Series	543
The ARMA Model Family	549
Preparing the Data for Machine Learning Models	552
Forecasting Using a Linear Model	555
Forecasting Using a Simple RNN	556
Forecasting Using a Deep RNN	557
Forecasting Multivariate Time Series	559
Forecasting Several Time Steps Ahead	560
Forecasting Using a Sequence-to-Sequence Model	562
Handling Long Sequences	565
Fighting the Unstable Gradients Problem	565
Tackling the Short-Term Memory Problem	568
Exercises	576
 16. Natural Language Processing with RNNs and Attention.	 577
Generating Shakespearean Text Using a Character RNN	578
Creating the Training Dataset	579
Building and Training the Char-RNN Model	581
Generating Fake Shakespearean Text	582
Stateful RNN	584
Sentiment Analysis	587
Masking	590
Reusing Pretrained Embeddings and Language Models	593
An Encoder–Decoder Network for Neural Machine Translation	595
Bidirectional RNNs	601
Beam Search	603
Attention Mechanisms	604
Attention Is All You Need: The Original Transformer Architecture	609
An Avalanche of Transformer Models	620
Vision Transformers	624
Hugging Face’s Transformers Library	629
Exercises	633
 17. Autoencoders, GANs, and Diffusion Models.	 635
Efficient Data Representations	637
Performing PCA with an Undercomplete Linear Autoencoder	639

Stacked Autoencoders	640
Implementing a Stacked Autoencoder Using Keras	641
Visualizing the Reconstructions	642
Visualizing the Fashion MNIST Dataset	643
Unsupervised Pretraining Using Stacked Autoencoders	644
Tying Weights	645
Training One Autoencoder at a Time	646
Convolutional Autoencoders	648
Denoising Autoencoders	649
Sparse Autoencoders	651
Variational Autoencoders	654
Generating Fashion MNIST Images	658
Generative Adversarial Networks	659
The Difficulties of Training GANs	663
Deep Convolutional GANs	665
Progressive Growing of GANs	668
StyleGANs	671
Diffusion Models	673
Exercises	681
18. Reinforcement Learning.....	683
Learning to Optimize Rewards	684
Policy Search	685
Introduction to OpenAI Gym	687
Neural Network Policies	691
Evaluating Actions: The Credit Assignment Problem	693
Policy Gradients	694
Markov Decision Processes	699
Temporal Difference Learning	703
Q-Learning	704
Exploration Policies	706
Approximate Q-Learning and Deep Q-Learning	707
Implementing Deep Q-Learning	708
Deep Q-Learning Variants	713
Fixed Q-value Targets	713
Double DQN	714
Prioritized Experience Replay	714
Dueling DQN	715
Overview of Some Popular RL Algorithms	716
Exercises	720

19. Training and Deploying TensorFlow Models at Scale.....	721
Serving a TensorFlow Model	722
Using TensorFlow Serving	722
Creating a Prediction Service on Vertex AI	732
Running Batch Prediction Jobs on Vertex AI	739
Deploying a Model to a Mobile or Embedded Device	741
Running a Model in a Web Page	744
Using GPUs to Speed Up Computations	746
Getting Your Own GPU	747
Managing the GPU RAM	749
Placing Operations and Variables on Devices	752
Parallel Execution Across Multiple Devices	753
Training Models Across Multiple Devices	756
Model Parallelism	756
Data Parallelism	759
Training at Scale Using the Distribution Strategies API	765
Training a Model on a TensorFlow Cluster	766
Running Large Training Jobs on Vertex AI	770
Hyperparameter Tuning on Vertex AI	772
Exercises	776
Thank You!	777
 A. Machine Learning Project Checklist.....	 779
 B. Autodiff.....	 785
 C. Special Data Structures.....	 793
 D. TensorFlow Graphs.....	 801
 Index.....	 811