

O'REILLY

Hands-On Machine Learning with Scikit-Learn & TensorFlow

CONCEPTS, TOOLS, AND TECHNIQUES
TO BUILD INTELLIGENT SYSTEMS



powered by



Aurélien Géron

Hands-On Machine Learning with Scikit-Learn and TensorFlow

Concepts, Tools, and Techniques to Build Intelligent Systems

Part I. The Fundamentals of Machine Learning

1. The Machine Learning Landscape

What is Machine Learning?

Supervised/Unsupervised Learning

Aurélien Géron

Overfitting	23
Underfitting	25
Stepping Back	26
Testing and Validation	29
Exercises	31

2. End-to-End Machine Learning Project

Working with Real Data	33
Look at the Big Picture	34
Frame the Problem	35
Select a Performance Measure	36

Hands-On Machine Learning with Scikit-Learn and TensorFlow

by Aurélien Géron

Copyright © 2017 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 (<http://oreilly.com/safari>). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Nicole Tache

Indexer: Wendy Catalano

Production Editor: Nicholas Adams

Interior Designer: David Futato

Copyeditor: Rachel Monaghan

Cover Designer: Randy Comer

Proofreader: Charles Roumeliotis

Illustrator: Rebecca Demarest

March 2017: First Edition

Revision History for the First Edition

2017-03-10: First Release

2017-06-09: Second Release

See <http://oreilly.com/catalog/errata.csp?isbn=9781491962299> for release details.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, 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.

978-1-491-96229-9

[LSI]

Preface

The Machine Learning Tsunami

In 2006, Geoffrey Hinton et al. published a paper¹ showing how to train a deep neural network capable of recognizing handwritten digits with state-of-the-art precision (>98%). They branded this technique “Deep Learning.” Training a deep neural net was widely considered impossible at the time,² and most researchers had abandoned the idea since the 1990s. This paper revived the interest of the scientific community and before long many new papers demonstrated that Deep Learning was not only possible, but capable of mind-blowing achievements that no other Machine Learning (ML) technique could hope to match (with the help of tremendous computing power and great amounts of data). This enthusiasm soon extended to many other areas of Machine Learning.

Fast-forward 10 years and Machine Learning has conquered the industry: it is now at the heart of much of the magic in today’s high-tech products, ranking your web search results, powering your smartphone’s speech recognition, and recommending videos, beating the world champion at the game of Go. Before you know it, it will be driving your car.

Machine Learning in Your Projects

So naturally you are excited about Machine Learning and you would love to join the party!

Perhaps you would like to give your homemade robot a brain of its own? Make it recognize faces? Or learn to walk around?

1 Available on Hinton’s home page at <http://www.cs.toronto.edu/~hinton/>.

2 Despite the fact that Yann Lecun’s deep convolutional neural networks had worked well for image recognition since the 1990s, although they were not as general purpose.

Or maybe your company has tons of data (user logs, financial data, production data, machine sensor data, hotline stats, HR reports, etc.), and more than likely you could unearth some hidden gems if you just knew where to look; for example:

- Segment customers and find the best marketing strategy for each group
- Recommend products for each client based on what similar clients bought
- Detect which transactions are likely to be fraudulent
- Predict next year's revenue
- And more (<https://www.kaggle.com/wiki/DataScienceUseCases>)

Whatever the reason, you have decided to learn Machine Learning and implement it in your projects. Great idea!

Objective and Approach

This book assumes that you know close to nothing about Machine Learning. Its goal is to give you the concepts, the intuitions, and the tools you need to actually implement programs capable of *learning from data*.

We will cover a large number of techniques, from the simplest and most commonly used (such as linear regression) to some of the Deep Learning techniques that regularly win competitions.

Rather than implementing our own toy versions of each algorithm, we will be using actual production-ready Python frameworks:

- Scikit-Learn (<http://scikit-learn.org/>) is very easy to use, yet it implements many Machine Learning algorithms efficiently, so it makes for a great entry point to learn Machine Learning.
- TensorFlow (<http://tensorflow.org/>) is a more complex library for distributed numerical computation using data flow graphs. It makes it possible to train and run very large neural networks efficiently by distributing the computations across potentially thousands of multi-GPU servers. TensorFlow was created at Google and supports many of their large-scale Machine Learning applications. It was open-sourced in November 2015.

The book favors a hands-on approach, growing an intuitive understanding of Machine Learning through concrete working examples and just a little bit of theory. While you can read this book without picking up your laptop, we highly recommend you experiment with the code examples available online as Jupyter notebooks at <https://github.com/ageron/handson-ml>.

Prerequisites

This book assumes that you have some Python programming experience and that you are familiar with Python's main scientific libraries, in particular NumPy (<http://numpy.org/>), Pandas (<http://pandas.pydata.org/>), and Matplotlib (<http://matplotlib.org/>).

Also, if you care about what's under the hood you should have a reasonable understanding of college-level math as well (calculus, linear algebra, probabilities, and statistics).

If you don't know Python yet, <http://learnpython.org/> is a great place to start. The official tutorial on python.org (<https://docs.python.org/3/tutorial/>) is also quite good.

If you have never used Jupyter, Chapter 2 will guide you through installation and the basics: it is a great tool to have in your toolbox.

If you are not familiar with Python's scientific libraries, the provided Jupyter notebooks include a few tutorials. There is also a quick math tutorial for linear algebra.

Roadmap

This book is organized in two parts. Part I, *The Fundamentals of Machine Learning*, covers the following topics:

- What is Machine Learning? What problems does it try to solve? What are the main categories and fundamental concepts of Machine Learning systems?
- The main steps in a typical Machine Learning project.
- Learning by fitting a model to data.
- Optimizing a cost function.
- Handling, cleaning, and preparing data.
- Selecting and engineering features.
- Selecting a model and tuning hyperparameters using cross-validation.
- The main challenges of Machine Learning, in particular underfitting and overfitting (the bias/variance tradeoff).
- Reducing the dimensionality of the training data to fight the curse of dimensionality.
- The most common learning algorithms: Linear and Polynomial Regression, Logistic Regression, k-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forests, and Ensemble methods.

Part II, *Neural Networks and Deep Learning*, covers the following topics:

- What are neural nets? What are they good for?
- Building and training neural nets using TensorFlow.
- The most important neural net architectures: feedforward neural nets, convolutional nets, recurrent nets, long short-term memory (LSTM) nets, and autoencoders.
- Techniques for training deep neural nets.
- Scaling neural networks for huge datasets.
- Reinforcement learning.

The first part is based mostly on Scikit-Learn while the second part uses TensorFlow.



Don't jump into deep waters too hastily: while Deep Learning is no doubt one of the most exciting areas in Machine Learning, you should master the fundamentals first. Moreover, most problems can be solved quite well using simpler techniques such as Random Forests and Ensemble methods (discussed in Part I). Deep Learning is best suited for complex problems such as image recognition, speech recognition, or natural language processing, provided you have enough data, computing power, and patience.

Other Resources

Many resources are available to learn about Machine Learning. Andrew Ng's ML course on Coursera (<https://www.coursera.org/learn/machine-learning/>) and Geoffrey Hinton's course on neural networks and Deep Learning (<https://www.coursera.org/course/neuralnets>) are amazing, although they both require a significant time investment (think months).

There are also many interesting websites about Machine Learning, including of course Scikit-Learn's exceptional User Guide (http://scikit-learn.org/stable/user_guide.html). You may also enjoy Dataquest (<https://www.dataquest.io/>), which provides very nice interactive tutorials, and ML blogs such as those listed on Quora (<http://goo.gl/GwtU3A>). Finally, the Deep Learning website (<http://deeplearning.net/>) has a good list of resources to learn more.

Of course there are also many other introductory books about Machine Learning, in particular:

- Joel Grus, *Data Science from Scratch* (O'Reilly). This book presents the fundamentals of Machine Learning, and implements some of the main algorithms in pure Python (from scratch, as the name suggests).

- Stephen Marsland, *Machine Learning: An Algorithmic Perspective* (Chapman and Hall). This book is a great introduction to Machine Learning, covering a wide range of topics in depth, with code examples in Python (also from scratch, but using NumPy).
- Sebastian Raschka, *Python Machine Learning* (Packt Publishing). Also a great introduction to Machine Learning, this book leverages Python open source libraries (Pylearn 2 and Theano).
- Yaser S. Abu-Mostafa, Malik Magdon-Ismael, and Hsuan-Tien Lin, *Learning from Data* (AMLBook). A rather theoretical approach to ML, this book provides deep insights, in particular on the bias/variance tradeoff (see Chapter 4).
- Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach, 3rd Edition* (Pearson). This is a great (and huge) book covering an incredible amount of topics, including Machine Learning. It helps put ML into perspective.

Finally, a great way to learn is to join ML competition websites such as Kaggle.com this will allow you to practice your skills on real-world problems, with help and insights from some of the best ML professionals out there.

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 tip or suggestion.



This element signifies a general note.



This element indicates a warning or caution.

Using Code Examples


Supplemental material (code examples, exercises, etc.) is available for download at <https://github.com/ageron/handson-ml>.

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: “*Hands-On Machine Learning with Scikit-Learn and TensorFlow* by Aurélien Géron (O'Reilly). Copyright 2017 Aurélien Géron, 978-1-491-96229-9.”

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**® 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/hands-on-machine-learning-with-scikit-learn-and-tensorflow>.

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

I would like to thank my Google colleagues, in particular the YouTube video classification team, for teaching me so much about Machine Learning. I could never have started this project without them. Special thanks to my personal ML gurus: Clément Courbet, Julien Dubois, Mathias Kende, Daniel Kitachewsky, James Pack, Alexander Pak, Anosh Raj, Vitor Sessak, Wiktor Tomczak, Ingrid von Glehn, Rich Washington, and everyone at YouTube Paris.

I am incredibly grateful to all the amazing people who took time out of their busy lives to review my book in so much detail. Thanks to Pete Warden for answering all my TensorFlow questions, reviewing Part II, providing many interesting insights, and of course for being part of the core TensorFlow team. You should definitely check out

his blog (<https://petewarden.com/>)! Many thanks to Lukas Biewald for his very thorough review of Part II: he left no stone unturned, tested all the code (and caught a few errors), made many great suggestions, and his enthusiasm was contagious. You should check out his blog (<https://lukasbiewald.com/>) and his cool robots (<https://goo.gl/Eu5u28>)! Thanks to Justin Francis, who also reviewed Part II very thoroughly, catching errors and providing great insights, in particular in Chapter 16. Check out his posts (<https://goo.gl/28ve8z>) on TensorFlow!

Huge thanks as well to David Andrzejewski, who reviewed Part I and provided incredibly useful feedback, identifying unclear sections and suggesting how to improve them. Check out his website (<http://www.david-andrzejewski.com/>)! Thanks to Grégoire Mesnil, who reviewed Part II and contributed very interesting practical advice on training neural networks. Thanks as well to Eddy Hung, Salim Sémaoune, Karim Matrah, Ingrid von Glehn, Iain Smears, and Vincent Guilbeau for reviewing Part I and making many useful suggestions. And I also wish to thank my father-in-law, Michel Tessier, former mathematics teacher and now a great translator of Anton Chekhov, for helping me iron out some of the mathematics and notations in this book and reviewing the linear algebra Jupyter notebook.

And of course, a gigantic “thank you” to my dear brother Sylvain, who reviewed every single chapter, tested every line of code, provided feedback on virtually every section, and encouraged me from the first line to the last. Love you, bro!

Many thanks as well to O’Reilly’s fantastic staff, in particular Nicole Tache, who gave me insightful feedback, always cheerful, encouraging, and helpful. Thanks as well to Marie Beaugureau, Ben Lorica, Mike Loukides, and Laurel Ruma for believing in this project and helping me define its scope. Thanks to Matt Hacker and all of the Atlas team for answering all my technical questions regarding formatting, asciidoc, and LaTeX, and thanks to Rachel Monaghan, Nick Adams, and all of the production team for their final review and their hundreds of corrections.

Last but not least, I am infinitely grateful to my beloved wife, Emmanuelle, and to our three wonderful kids, Alexandre, Rémi, and Gabrielle, for encouraging me to work hard on this book, asking many questions (who said you can’t teach neural networks to a seven-year-old?), and even bringing me cookies and coffee. What more can one dream of?

Table of Contents

40	Exercises	40
41	Check the Assumptions	41
42	Download the Data	42
43	Take a Quick Look at the Data Structure	43
44	Create a Test Set	44
45	Discover and Visualize the Data to Gain Insights	45
46	Visualizing Geographical Data	46
47	Looking for Correlations	47
48	Experimenting with Attribute Combinations	48
49	Prepare the Data for Machine Learning Algorithms	49
50	Data Cleaning	50
51	Handling Text and Categorical Attributes	51
52	Custom Transformers	52
53	Regularized Linear Models	53
54	Logistic Regression	54
55	Linear Models	55
56	Transformation Pipelines	56
57	Feature Selection	57
58	Training and Evaluating	58
59	Better Evaluation Using Cross-Validation	59
60	Five-Tuple Your Model	60
61	Grid Search	61
62	Randomized Search	62
63	Resample Methods	63
64	Analyze the Best Models and Their	64
65	Evaluate Your System on the Test Set	65
66	Launch, Monitor, and Maintain	66
67	Try It Out	67
68	Exercises	68
69	Classification	69
70	MNIST	70
71	Training a Binary Classifier	71
72	Performance Measures	72
73	Measuring Accuracy Using Cross-Validation	73
74	Confusion Matrix	74
75	Precision and Recall	75
76	Precision/Recall Tradeoff	76
77	The ROC Curve	77
78	Multiclass Classification	78
79	Error Analysis	79
80	Multilabel Classification	80
81	Multiclass Output Classification	81
82	Check the Assumptions	82
83	Download the Data	83
84	Take a Quick Look at the Data Structure	84
85	Create a Test Set	85
86	Discover and Visualize the Data to Gain Insights	86
87	Visualizing Geographical Data	87
88	Looking for Correlations	88
89	Experimenting with Attribute Combinations	89
90	Prepare the Data for Machine Learning Algorithms	90
91	Data Cleaning	91
92	Handling Text and Categorical Attributes	92
93	Custom Transformers	93
94	Regularized Linear Models	94
95	Logistic Regression	95
96	Linear Models	96
97	Transformation Pipelines	97
98	Feature Selection	98
99	Training and Evaluating	99
100	Better Evaluation Using Cross-Validation	100
101	Five-Tuple Your Model	101
102	Grid Search	102
103	Randomized Search	103
104	Resample Methods	104
105	Analyze the Best Models and Their	105
106	Evaluate Your System on the Test Set	106
107	Launch, Monitor, and Maintain	107
108	Try It Out	108
109	Exercises	109
110	Classification	110
111	MNIST	111
112	Training a Binary Classifier	112
113	Performance Measures	113
114	Measuring Accuracy Using Cross-Validation	114
115	Confusion Matrix	115
116	Precision and Recall	116
117	Precision/Recall Tradeoff	117
118	The ROC Curve	118
119	Multiclass Classification	119
120	Error Analysis	120
121	Multilabel Classification	121
122	Multiclass Output Classification	122
123	Check the Assumptions	123
124	Download the Data	124
125	Take a Quick Look at the Data Structure	125
126	Create a Test Set	126
127	Discover and Visualize the Data to Gain Insights	127
128	Visualizing Geographical Data	128
129	Looking for Correlations	129
130	Experimenting with Attribute Combinations	130
131	Prepare the Data for Machine Learning Algorithms	131
132	Data Cleaning	132
133	Handling Text and Categorical Attributes	133
134	Custom Transformers	134
135	Regularized Linear Models	135
136	Logistic Regression	136
137	Linear Models	137
138	Transformation Pipelines	138
139	Feature Selection	139
140	Training and Evaluating	140
141	Better Evaluation Using Cross-Validation	141
142	Five-Tuple Your Model	142
143	Grid Search	143
144	Randomized Search	144
145	Resample Methods	145
146	Analyze the Best Models and Their	146
147	Evaluate Your System on the Test Set	147
148	Launch, Monitor, and Maintain	148
149	Try It Out	149
150	Exercises	150
151	Classification	151
152	MNIST	152
153	Training a Binary Classifier	153
154	Performance Measures	154
155	Measuring Accuracy Using Cross-Validation	155
156	Confusion Matrix	156
157	Precision and Recall	157
158	Precision/Recall Tradeoff	158
159	The ROC Curve	159
160	Multiclass Classification	160
161	Error Analysis	161
162	Multilabel Classification	162
163	Multiclass Output Classification	163
164	Check the Assumptions	164
165	Download the Data	165
166	Take a Quick Look at the Data Structure	166
167	Create a Test Set	167
168	Discover and Visualize the Data to Gain Insights	168
169	Visualizing Geographical Data	169
170	Looking for Correlations	170
171	Experimenting with Attribute Combinations	171
172	Prepare the Data for Machine Learning Algorithms	172
173	Data Cleaning	173
174	Handling Text and Categorical Attributes	174
175	Custom Transformers	175
176	Regularized Linear Models	176
177	Logistic Regression	177
178	Linear Models	178
179	Transformation Pipelines	179
180	Feature Selection	180
181	Training and Evaluating	181
182	Better Evaluation Using Cross-Validation	182
183	Five-Tuple Your Model	183
184	Grid Search	184
185	Randomized Search	185
186	Resample Methods	186
187	Analyze the Best Models and Their	187
188	Evaluate Your System on the Test Set	188
189	Launch, Monitor, and Maintain	189
190	Try It Out	190
191	Exercises	191
192	Classification	192
193	MNIST	193
194	Training a Binary Classifier	194
195	Performance Measures	195
196	Measuring Accuracy Using Cross-Validation	196
197	Confusion Matrix	197
198	Precision and Recall	198
199	Precision/Recall Tradeoff	199
200	The ROC Curve	200
201	Multiclass Classification	201
202	Error Analysis	202
203	Multilabel Classification	203
204	Multiclass Output Classification	204
205	Check the Assumptions	205
206	Download the Data	206
207	Take a Quick Look at the Data Structure	207
208	Create a Test Set	208
209	Discover and Visualize the Data to Gain Insights	209
210	Visualizing Geographical Data	210
211	Looking for Correlations	211
212	Experimenting with Attribute Combinations	212
213	Prepare the Data for Machine Learning Algorithms	213
214	Data Cleaning	214
215	Handling Text and Categorical Attributes	215
216	Custom Transformers	216
217	Regularized Linear Models	217
218	Logistic Regression	218
219	Linear Models	219
220	Transformation Pipelines	220
221	Feature Selection	221
222	Training and Evaluating	222
223	Better Evaluation Using Cross-Validation	223
224	Five-Tuple Your Model	224
225	Grid Search	225
226	Randomized Search	226
227	Resample Methods	227
228	Analyze the Best Models and Their	228
229	Evaluate Your System on the Test Set	229
230	Launch, Monitor, and Maintain	230
231	Try It Out	231
232	Exercises	232
233	Classification	233
234	MNIST	234
235	Training a Binary Classifier	235
236	Performance Measures	236
237	Measuring Accuracy Using Cross-Validation	237
238	Confusion Matrix	238
239	Precision and Recall	239
240	Precision/Recall Tradeoff	240
241	The ROC Curve	241
242	Multiclass Classification	242
243	Error Analysis	243
244	Multilabel Classification	244
245	Multiclass Output Classification	245
246	Check the Assumptions	246
247	Download the Data	247
248	Take a Quick Look at the Data Structure	248
249	Create a Test Set	249
250	Discover and Visualize the Data to Gain Insights	250
251	Visualizing Geographical Data	251
252	Looking for Correlations	252
253	Experimenting with Attribute Combinations	253
254	Prepare the Data for Machine Learning Algorithms	254
255	Data Cleaning	255
256	Handling Text and Categorical Attributes	256
257	Custom Transformers	257
258	Regularized Linear Models	258
259	Logistic Regression	259
260	Linear Models	260
261	Transformation Pipelines	261
262	Feature Selection	262
263	Training and Evaluating	263
264	Better Evaluation Using Cross-Validation	264
265	Five-Tuple Your Model	265
266	Grid Search	266
267	Randomized Search	267
268	Resample Methods	268
269	Analyze the Best Models and Their	269
270	Evaluate Your System on the Test Set	270
271	Launch, Monitor, and Maintain	271
272	Try It Out	272
273	Exercises	273
274	Classification	274
275	MNIST	275
276	Training a Binary Classifier	276
277	Performance Measures	277
278	Measuring Accuracy Using Cross-Validation	278
279	Confusion Matrix	279
280	Precision and Recall	280
281	Precision/Recall Tradeoff	281
282	The ROC Curve	282
283	Multiclass Classification	283
284	Error Analysis	284
285	Multilabel Classification	285
286	Multiclass Output Classification	286
287	Check the Assumptions	287
288	Download the Data	288
289	Take a Quick Look at the Data Structure	289
290	Create a Test Set	290
291	Discover and Visualize the Data to Gain Insights	291
292	Visualizing Geographical Data	292
293	Looking for Correlations	293
294	Experimenting with Attribute Combinations	294
295	Prepare the Data for Machine Learning Algorithms	295
296	Data Cleaning	296
297	Handling Text and Categorical Attributes	297
298	Custom Transformers	298
299	Regularized Linear Models	299
300	Logistic Regression	300
301	Linear Models	301
302	Transformation Pipelines	302
303	Feature Selection	303
304	Training and Evaluating	304
305	Better Evaluation Using Cross-Validation	305
306	Five-Tuple Your Model	306
307	Grid Search	307
308	Randomized Search	308
309	Resample Methods	309
310	Analyze the Best Models and Their	310
311	Evaluate Your System on the Test Set	311
312	Launch, Monitor, and Maintain	312
313	Try It Out	313
314	Exercises	314
315	Classification	315
316	MNIST	316
317	Training a Binary Classifier	317
318	Performance Measures	318
319	Measuring Accuracy Using Cross-Validation	319
320	Confusion Matrix	320
321	Precision and Recall	321
322	Precision/Recall Tradeoff	322
323	The ROC Curve	323
324	Multiclass Classification	324
325	Error Analysis	325
326	Multilabel Classification	326
327	Multiclass Output Classification	327
328	Check the Assumptions	328
329	Download the Data	329
330	Take a Quick Look at the Data Structure	330
331	Create a Test Set	331
332	Discover and Visualize the Data to Gain Insights	332
333	Visualizing Geographical Data	333
334	Looking for Correlations	334
335	Experimenting with Attribute Combinations	335
336	Prepare the Data for Machine Learning Algorithms	336
337	Data Cleaning	337
338	Handling Text and Categorical Attributes	338
339	Custom Transformers	339
340	Regularized Linear Models	340
341	Logistic Regression	341
342	Linear Models	342
343	Transformation Pipelines	343
344	Feature Selection	344
345	Training and Evaluating	345
346	Better Evaluation Using Cross-Validation	346
347	Five-Tuple Your Model	347
348	Grid Search	348
349	Randomized Search	349
350	Resample Methods	350
351	Analyze the Best Models and Their	351
352	Evaluate Your System on the Test Set	352
353	Launch, Monitor, and Maintain	353
354	Try It Out	354
355	Exercises	355
356	Classification	356
357	MNIST	357
358	Training a Binary Classifier	358
359	Performance Measures	359
360	Measuring Accuracy Using Cross-Validation	360
361	Confusion Matrix	361
362	Precision and Recall	362
363	Precision/Recall Tradeoff	363
364	The ROC Curve	364
365	Multiclass Classification	365
366	Error Analysis	366
367	Multilabel Classification	367
368	Multiclass Output Classification	368
369	Check the Assumptions	369
370	Download the Data	370
371	Take a Quick Look at the Data Structure	371
372	Create a Test Set	372
373	Discover and Visualize the Data to Gain Insights	373
374	Visualizing Geographical Data	374
375	Looking for Correlations	375
376	Experimenting with Attribute Combinations	376
377	Prepare the Data for Machine Learning Algorithms	377
378	Data Cleaning	378
379	Handling Text and Categorical Attributes	379
380	Custom Transformers	380
381	Regularized Linear Models	381
382	Logistic Regression	382
383	Linear Models	383
384	Transformation Pipelines	384
385	Feature Selection	385
386	Training and Evaluating	386
387	Better Evaluation Using Cross-Validation	387
388	Five-Tuple Your Model	388
389	Grid Search	389
390	Randomized Search	390
391	Resample Methods	391
392	Analyze the Best Models and Their	392
393	Evaluate Your System on the Test Set	393
394	Launch, Monitor, and Maintain	394
395	Try It Out	395
396	Exercises	396
397	Classification	397
398	MNIST	398
399	Training a Binary Classifier	399
400	Performance Measures	400
401	Measuring Accuracy Using Cross-Validation	401
402	Confusion Matrix	402
403	Precision and Recall	403
404	Precision/Recall Tradeoff	404
405	The ROC Curve	405
406	Multiclass Classification	406
407	Error Analysis	407
408	Multilabel Classification	408
409	Multiclass Output Classification	409
410	Check the Assumptions	410
411	Download the Data	411
412	Take a Quick Look at the Data Structure	412
413	Create a Test Set	413
414	Discover and Visualize the Data to Gain Insights	414
415	Visualizing Geographical Data	415
416	Looking for Correlations	416
417	Experimenting with Attribute Combinations	417
418	Prepare the Data for Machine Learning Algorithms	418
419	Data Cleaning	419
420	Handling Text and Categorical Attributes	420
421	Custom Transformers	421
422	Regularized Linear Models	422
423	Logistic Regression	423
424	Linear Models	424
425	Transformation Pipelines	425
426	Feature Selection	426
427	Training and Evaluating	427
428	Better Evaluation Using Cross-Validation	428
429	Five-Tuple Your Model	429
430	Grid Search	430
431	Randomized Search	431
432	Resample Methods	432
433	Analyze the Best Models and Their	433
434	Evaluate Your System on the Test Set	434
435	Launch, Monitor, and Maintain	435
436	Try It Out	436
437	Exercises	437
438	Classification	438
439	MNIST	439
440	Training a Binary Classifier	440
441	Performance Measures	441
442	Measuring Accuracy Using Cross-Validation	442
443	Confusion Matrix	443
444	Precision and Recall	444
445	Precision/Recall Tradeoff	445
446	The ROC Curve	446
447	Multiclass Classification	447
448	Error Analysis	448
449	Multilabel Classification	449
450	Multiclass Output Classification	450
451	Check the Assumptions	451
452	Download the Data	452
453	Take a Quick Look at the Data Structure	453
454	Create a Test Set	454
455	Discover and Visualize the Data to Gain Insights	455
456	Visualizing Geographical Data	456
457	Looking for Correlations	457
458	Experimenting with Attribute Combinations	458
459	Prepare the Data for Machine Learning Algorithms	459
460	Data Cleaning	460
461	Handling Text and Categorical Attributes	461
462	Custom Transformers	462
463	Regularized Linear Models	463
464	Logistic Regression	464
465	Linear Models	465
466	Transformation Pipelines	466
467	Feature Selection	467
468	Training and Evaluating	468
469		

Check the Assumptions	40
Get the Data	40
Create the Workspace	40
Download the Data	43
Take a Quick Look at the Data Structure	45
Create a Test Set	49
Discover and Visualize the Data to Gain Insights	53
Visualizing Geographical Data	53
Looking for Correlations	56
Experimenting with Attribute Combinations	59
Prepare the Data for Machine Learning Algorithms	60
Data Cleaning	61
Handling Text and Categorical Attributes	63
Custom Transformers	65
Feature Scaling	66
Transformation Pipelines	66
Select and Train a Model	68
Training and Evaluating on the Training Set	69
Better Evaluation Using Cross-Validation	70
Fine-Tune Your Model	72
Grid Search	72
Randomized Search	75
Ensemble Methods	75
Analyze the Best Models and Their Errors	75
Evaluate Your System on the Test Set	76
Launch, Monitor, and Maintain Your System	77
Try It Out!	77
Exercises	78
3. Classification.....	79
MNIST	79
Training a Binary Classifier	82
Performance Measures	82
Measuring Accuracy Using Cross-Validation	83
Confusion Matrix	84
Precision and Recall	86
Precision/Recall Tradeoff	87
The ROC Curve	91
Multiclass Classification	93
Error Analysis	96
Multilabel Classification	100
Multioutput Classification	101

Exercises	102
4. Training Models	105
Linear Regression	106
The Normal Equation	108
Computational Complexity	110
Gradient Descent	111
Batch Gradient Descent	114
Stochastic Gradient Descent	117
Mini-batch Gradient Descent	119
Polynomial Regression	121
Learning Curves	123
Regularized Linear Models	127
Ridge Regression	127
Lasso Regression	130
Elastic Net	132
Early Stopping	133
Logistic Regression	134
Estimating Probabilities	134
Training and Cost Function	135
Decision Boundaries	136
Softmax Regression	139
Exercises	142
5. Support Vector Machines	145
Linear SVM Classification	145
Soft Margin Classification	146
Nonlinear SVM Classification	149
Polynomial Kernel	150
Adding Similarity Features	151
Gaussian RBF Kernel	152
Computational Complexity	153
SVM Regression	154
Under the Hood	156
Decision Function and Predictions	156
Training Objective	157
Quadratic Programming	159
The Dual Problem	160
Kernelized SVM	161
Online SVMs	164
Exercises	165

6. Decision Trees	167
Training and Visualizing a Decision Tree	167
Making Predictions	169
Estimating Class Probabilities	171
The CART Training Algorithm	171
Computational Complexity	172
Gini Impurity or Entropy?	172
Regularization Hyperparameters	173
Regression	175
Instability	177
Exercises	178
7. Ensemble Learning and Random Forests	181
Voting Classifiers	181
Bagging and Pasting	185
Bagging and Pasting in Scikit-Learn	186
Out-of-Bag Evaluation	187
Random Patches and Random Subspaces	188
Random Forests	189
Extra-Trees	190
Feature Importance	190
Boosting	191
AdaBoost	192
Gradient Boosting	195
Stacking	200
Exercises	202
8. Dimensionality Reduction	205
The Curse of Dimensionality	206
Main Approaches for Dimensionality Reduction	207
Projection	207
Manifold Learning	210
PCA	211
Preserving the Variance	211
Principal Components	212
Projecting Down to d Dimensions	213
Using Scikit-Learn	214
Explained Variance Ratio	214
Choosing the Right Number of Dimensions	215
PCA for Compression	216
Incremental PCA	217
Randomized PCA	218

Kernel PCA	218
Selecting a Kernel and Tuning Hyperparameters	219
LLE	221
Other Dimensionality Reduction Techniques	223
Exercises	224

Part II. Neural Networks and Deep Learning

9. Up and Running with TensorFlow	229
Installation	232
Creating Your First Graph and Running It in a Session	232
Managing Graphs	234
Lifecycle of a Node Value	235
Linear Regression with TensorFlow	235
Implementing Gradient Descent	237
Manually Computing the Gradients	237
Using autodiff	238
Using an Optimizer	239
Feeding Data to the Training Algorithm	239
Saving and Restoring Models	241
Visualizing the Graph and Training Curves Using TensorBoard	242
Name Scopes	245
Modularity	246
Sharing Variables	248
Exercises	251
10. Introduction to Artificial Neural Networks	253
From Biological to Artificial Neurons	254
Biological Neurons	255
Logical Computations with Neurons	256
The Perceptron	257
Multi-Layer Perceptron and Backpropagation	261
Training an MLP with TensorFlow's High-Level API	264
Training a DNN Using Plain TensorFlow	265
Construction Phase	265
Execution Phase	269
Using the Neural Network	269
Fine-Tuning Neural Network Hyperparameters	270
Number of Hidden Layers	270
Number of Neurons per Hidden Layer	271
Activation Functions	272

Exercises	272
11. Training Deep Neural Nets	275
Vanishing/Exploding Gradients Problems	275
Xavier and He Initialization	277
Nonsaturating Activation Functions	279
Batch Normalization	282
Gradient Clipping	286
Reusing Pretrained Layers	287
Reusing a TensorFlow Model	287
Reusing Models from Other Frameworks	289
Freezing the Lower Layers	290
Caching the Frozen Layers	291
Tweaking, Dropping, or Replacing the Upper Layers	292
Model Zoos	292
Unsupervised Pretraining	293
Pretraining on an Auxiliary Task	294
Faster Optimizers	295
Momentum Optimization	295
Nesterov Accelerated Gradient	297
AdaGrad	298
RMSProp	300
Adam Optimization	300
Learning Rate Scheduling	303
Avoiding Overfitting Through Regularization	305
Early Stopping	305
ℓ_1 and ℓ_2 Regularization	305
Dropout	307
Max-Norm Regularization	309
Data Augmentation	311
Practical Guidelines	312
Exercises	313
12. Distributing TensorFlow Across Devices and Servers	315
Multiple Devices on a Single Machine	316
Installation	316
Managing the GPU RAM	319
Placing Operations on Devices	320
Parallel Execution	323
Control Dependencies	325
Multiple Devices Across Multiple Servers	325
Opening a Session	327