

Daniel Whitenack

Machine Learning With Go

Implement Regression, Classification, Clustering, Time-series Models, Neural Networks, and More using the Go Programming Language



Packt>

Machine Learning With Go

The mission of this book is to turn readers into productive, innovative data analysts who leverage Go to build robust and valuable applications. To this end, the book clearly introduces the technical aspects of building predictive models in Go, but it also helps the reader understand how machine learning workflows are applied in real-world scenarios.

Machine Learning With Go shows readers how to be productive in machine learning while also producing applications that maintain a high level of integrity. It also gives readers patterns to overcome challenges that are often encountered when trying to integrate machine learning in an engineering organization.

The readers will begin by gaining a solid understanding of how to gather, organize, and parse real-work data from a variety of sources. Readers will then develop a solid statistical toolkit that will allow them to be able to quickly gain intuition about the content of a dataset. Finally, the readers will gain hands-on experience implementing essential machine learning techniques (regression, classification, clustering, and so on) with relevant Go packages.

Finally, the reader will have a solid machine learning mindset and a powerful Go toolkit of techniques, packages, and examples implementations.

Things you will learn:

- Learn about data gathering, organization, parsing, and cleaning
- Explore matrices, linear algebra, statistics and probability
- See how to evaluate and validate models
- Look at regression, classification, and clustering
- Learn about neural networks and deep learning
- Utilize time series models and anomaly detection
- Get to grips with techniques for deploying and distributing analyses and models
- Optimize machine learning workflow techniques

Packt
www.packtpub.com

\$ 49.99 US
£ 41.99 UK

Prices do not include local sales
Tax or VAT where applicable



Machine Learning With Go

Daniel Whitenack



Machine Learning With Go

Implement Regression, Classification, Clustering, Time-series Models, Neural Networks, and More using the Go Programming Language

Daniel Whitenack

Packt

BIRMINGHAM - MUMBAI

Machine Learning With Go

Copyright © 2017 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: September 2017

Production reference: 1210917

Published by Packt Publishing Ltd.

Livery Place

35 Livery Street

Birmingham

B3 2PB, UK.

ISBN 978-1-78588-210-4

www.packtpub.com

Credits

Author

Daniel Whitenack

Copy Editor

Tasneem Fatehi

Reviewers

Niclas Jern

Richard Townsend

Project Coordinator

Manthan Patel

Commissioning Editor

Veena Pagare

Proofreader

Safis Editing

Acquisition Editor

Varsha Shetty

Indexer

Tejal Daruwale Soni

Content Development Editor

Snehal Kolte

Graphics

Tania Dutta

Technical Editor

Sagar Sawant

Production Coordinator

Deepika Naik

About the Author

Daniel Whitenack (@dwhitena), PhD, is a trained data scientist working with Pachyderm (@pachydermIO). Daniel develops innovative, distributed data pipelines that include predictive models, data visualizations, statistical analyses, and more. He has spoken at conferences around the world (GopherCon, JuliaCon, PyCon, ODSC, Spark Summit, and more), teaches data science/engineering at Purdue University (@LifeAtPurdue), and, with Ardan Labs (@ardanlabs), maintains the Go kernel for Jupyter, and is actively helping to organize contributions to various open source data science projects.

I would like to thank my wife for her boundless patience and support while writing this book.

I would also like to acknowledge the many wonderful gophers and data scientists that have mentored me, collaborated with me, and encouraged me. These include Bill Kennedy, Brendan Tracey, Sebastien Binet, Alex Sanchez, the whole team at Pachyderm (including Joey Zwicker and Joe Doliner), Chris Tava, Mat Ryer, David Hernandez, Xuanyi Chew, the team at Minio (including Anand Babu Periasamy and Garima Kapoor), and many more!

Soli Deo Gloria

About the Reviewers

Niclas Jern has been using Go to solve interesting problems at scale since Go 1.0. He graduated from Abo Akademi University with an MSc in computer engineering, majoring in software engineering.

He enjoys using Go and other programming languages to tackle problems, especially in the fields of data processing and machine learning. His hobbies include long walks, lifting heavy metal objects at the gym, the occasional rant at <http://www.njern.co>, and spending quality time with his wife and daughter.

Niclas currently works at Walkbase, a company he founded together with some of his classmates from Abo Akademi university, where he leads the engineering team, which is tackling the data processing problems that come with revolutionizing retail analytics.

Richard Townsend became the top contributor to GoLearn in 2014 (and hence is responsible for a lot of its odd behavior) while studying for his undergraduate degree at Warwick University. Since then, he's worked for a top UK technology company on everything from embedded systems to Android operating system frameworks, and currently spends his time optimizing web browsers. He still spends a significant amount of time on sentiment analysis (co-authoring two papers on it) and other natural language processing tasks like part of speech tagging often using the latest deep learning technologies.

www.PacktPub.com

For support files and downloads related to your book, please visit www.PacktPub.com. Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details. At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.



<https://www.packtpub.com/mapt> Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at <https://www.amazon.com/dp/1785882104>.

If you'd like to join our team of regular reviewers, you can email us at customerreviews@packtpub.com. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

Table of Contents

Preface	1
<hr/>	
Chapter 1: Gathering and Organizing Data	9
<hr/>	
Handling data - Gopher style	10
Best practices for gathering and organizing data with Go	12
CSV files	13
Reading in CSV data from a file	14
Handling unexpected fields	15
Handling unexpected types	16
Manipulating CSV data with data frames	19
JSON	21
Parsing JSON	21
JSON output	24
SQL-like databases	25
Connecting to an SQL database	25
Querying the database	26
Modifying the database	27
Caching	28
Caching data in memory	29
Caching data locally on disk	29
Data versioning	31
Pachyderm jargon	32
Deploying/installing Pachyderm	32
Creating data repositories for data versioning	33
Putting data into data repositories	34
Getting data out of versioned data repositories	34
References	34
Summary	35
<hr/>	
Chapter 2: Matrices, Probability, and Statistics	37
<hr/>	
Matrices and vectors	37
Vectors	38
Vector operations	38
Matrices	40
Matrix operations	42
Statistics	43

Distributions	44
Statistical measures	45
Measures of central tendency	45
Measures of spread or dispersion	47
Visualizing distributions	50
Histograms	50
Box plots	54
Probability	57
Random variables	57
Probability measures	58
Independent and conditional probability	58
Hypothesis testing	59
Test statistics	60
Calculating p-values	60
References	62
Summary	63
Chapter 3: Evaluation and Validation	65
<hr/>	
Evaluation	65
Continuous metrics	66
Categorical metrics	70
Individual evaluation metrics for categorical variables	70
Confusion matrices, AUC, and ROC	75
Validation	78
Training and test sets	79
Holdout set	83
Cross validation	84
References	86
Summary	87
Chapter 4: Regression	89
<hr/>	
Understanding regression model jargon	89
Linear regression	90
Overview of linear regression	91
Linear regression assumptions and pitfalls	93
Linear regression example	94
Profiling the data	94
Choosing our independent variable	98
Creating our training and test sets	101
Training our model	103
Evaluating the trained model	104
Multiple linear regression	108
Nonlinear and other types of regression	113

References	118
Summary	118
Chapter 5: Classification	119
<hr/>	
Understanding classification model jargon	120
Logistic regression	120
Overview of logistic regression	121
Logistic regression assumptions and pitfalls	125
Logistic regression example	125
Cleaning and profiling the data	126
Creating our training and test sets	131
Training and testing the logistic regression model	133
k-nearest neighbors	139
Overview of kNN	139
kNN assumptions and pitfalls	141
kNN example	142
Decision trees and random forests	144
Overview of decision trees and random forests	144
Decision tree and random forest assumptions and pitfalls	145
Decision tree example	146
Random forest example	147
Naive bayes	148
Overview of naive bayes and its big assumption	148
Naive bayes example	148
References	150
Summary	151
Chapter 6: Clustering	153
<hr/>	
Understanding clustering model jargon	154
Measuring Distance or Similarity	154
Evaluating clustering techniques	156
Internal clustering evaluation	156
External clustering evaluation	161
k-means clustering	162
Overview of k-means clustering	162
k-means assumptions and pitfalls	165
k-means clustering example	166
Profiling the data	166
Generating clusters with k-means	169
Evaluating the generated clusters	171
Other clustering techniques	174

References	175
Summary	175
Chapter 7: Time Series and Anomaly Detection	177
Representing time series data in Go	178
Understanding time series jargon	181
Statistics related to time series	182
Autocorrelation	182
Partial autocorrelation	187
Auto-regressive models for forecasting	190
Auto-regressive model overview	190
Auto-regressive model assumptions and pitfalls	191
Auto-regressive model example	192
Transforming to a stationary series	192
Analyzing the ACF and choosing an AR order	195
Fitting and evaluating an AR(2) model	196
Auto-regressive moving averages and other time series models	202
Anomaly detection	203
References	205
Summary	205
Chapter 8: Neural Networks and Deep Learning	207
Understanding neural net jargon	208
Building a simple neural network	209
Nodes in the network	210
Network architecture	212
Why do we expect this architecture to work?	213
Training our neural network	214
Utilizing the simple neural network	221
Training the neural network on real data	222
Evaluating the neural network	224
Introducing deep learning	226
What is a deep learning model?	227
Deep learning with Go	228
Setting up TensorFlow for use with Go	230
Retrieving and calling a pretrained TensorFlow model	230
Object detection using TensorFlow from Go	232
References	236
Summary	236
Chapter 9: Deploying and Distributing Analyses and Models	237
Running models reliably on remote machines	238

A brief introduction to Docker and Docker jargon	238
Docker-izing a machine learning application	240
Docker-izing the model training and export	240
Docker-izing model predictions	245
Testing the Docker images locally	250
Running the Docker images on remote machines	252
Building a scalable and reproducible machine learning pipeline	253
Setting up a Pachyderm and Kubernetes cluster	254
Building a Pachyderm machine learning pipeline	256
Creating and filling the input repositories	257
Creating and running the processing stages	261
Updating pipelines and examining provenance	266
Scaling pipeline stages	268
References	270
Summary	271
Appendix: Algorithms/Techniques Related to Machine Learning	273
Gradient descent	273
Entropy, information gain, and related methods	276
Backpropagation	278
Index	283
