

Cognitive Systems Monographs

Volume 35

Series editors

Rüdiger Dillmann, University of Karlsruhe, Karlsruhe, Germany
e-mail: ruediger.dillmann@kit.edu

Yoshihiko Nakamura, Tokyo University, Tokyo, Japan
e-mail: nakamura@ynl.t.u-tokyo.ac.jp

Stefan Schaal, University of Southern California, Los Angeles, USA
e-mail: sschaal@usc.edu

David Vernon, University of Skövde, Skövde, Sweden
e-mail: david@vernon.eu

About this Series

The Cognitive Systems Monographs (COSMOS) publish new developments and advances in the fields of cognitive systems research, rapidly and informally but with a high quality. The intent is to bridge cognitive brain science and biology with engineering disciplines. It covers all the technical contents, applications, and multidisciplinary aspects of cognitive systems, such as Bionics, System Analysis, System Modelling, System Design, Human Motion, Understanding, Human Activity Understanding, Man-Machine Interaction, Smart and Cognitive Environments, Human and Computer Vision, Neuroinformatics, Humanoids, Biologically motivated systems and artefacts Autonomous Systems, Linguistics, Sports Engineering, Computational Intelligence, Biosignal Processing, or Cognitive Materials as well as the methodologies behind them. Within the scope of the series are monographs, lecture notes, selected contributions from specialized conferences and workshops.

Advisory Board

Heinrich H. Bülthoff, MPI for Biological Cybernetics, Tübingen, Germany

Masayuki Inaba, The University of Tokyo, Japan

J.A. Scott Kelso, Florida Atlantic University, Boca Raton, FL, USA

Oussama Khatib, Stanford University, CA, USA

Yasuo Kuniyoshi, The University of Tokyo, Japan

Hiroshi G. Okuno, Kyoto University, Japan

Helge Ritter, University of Bielefeld, Germany

Giulio Sandini, University of Genova, Italy

Bruno Siciliano, University of Naples, Italy

Mark Steedman, University of Edinburgh, Scotland

Atsuo Takanishi, Waseda University, Tokyo, Japan

More information about this series at <http://www.springer.com/series/8354>

Mark Hoogendoorn · Burkhardt Funk

Machine Learning for the Quantified Self

On the Art of Learning from Sensory Data

 Springer

Mark Hoogendoorn
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam
The Netherlands

Burkhardt Funk
Institut für Wirtschaftsinformatik
Leuphana Universität Lüneburg
Lüneburg, Niedersachsen
Germany

ISSN 1867-4925

ISSN 1867-4933 (electronic)

Cognitive Systems Monographs

ISBN 978-3-319-66307-4

ISBN 978-3-319-66308-1 (eBook)

<https://doi.org/10.1007/978-3-319-66308-1>

Library of Congress Control Number: 2017949497

© Springer International Publishing AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*Live as if you were to die tomorrow.
Learn as if you were to live forever.*

Mahatma Gandhi

Foreword

Sensors are all around us, and increasingly on us. We carry smartphones and watches, which have the potential to gather enormous quantities of data. These data are often noisy, interrupted, and increasingly high dimensional. A challenge in data science is how to put this veritable fire hose of noisy data to use and extract useful summaries and predictions.

In this timely monograph, Mark Hoogendoorn and Burkhardt Funk face up to the challenge. Their choice of material shows good mastery of the various subfields of machine learning, which they bring to bear on these data. They cover a wide array of techniques for supervised and unsupervised learning, both for cross-sectional and time series data. Ending each chapter with a useful set of thinking and computing problems adds a helpful touch. I am sure this book will be welcomed by a broad audience, and I hope it is a big success.

June 2017

Trevor Hastie
Stanford University, Stanford, CA, USA

Preface

Self-tracking has become part of a modern lifestyle; wearables and smartphones support self-tracking in an easy fashion and change our behavior such as in the health sphere. The amount of data generated by these devices is so overwhelming that it is difficult to get useful insight from it. Luckily, in the domain of artificial intelligence, techniques exist that can help out here: machine learning approaches are well suited to assist and enable one to analyze this type of data. While there are ample books that explain machine learning techniques, self-tracking data comes with its own difficulties that require dedicated techniques such as learning over time and across users. In this book, we will explain the complete loop to effectively use self-tracking data for machine learning; from cleaning the data, the identification of features, finding clusters in the data, algorithms to create predictions of values for the present and future, to learning how to provide feedback to users based on their tracking data. All concepts we explain are drawn from state-of-the-art scientific literature. To illustrate all approaches, we use a case study of a rich self-tracking dataset obtained from the crowdsignals platform. While the book is focused on the self-tracking data, the techniques explained are more widely applicable to sensory data in general, making it useful for a wider audience.

Who should read this book? The book is intended for students, scholars, and practitioners with an interest in analyzing sensory data and user-generated content to build their own algorithms and applications. We will explain the basics of the suitable algorithms, and the underlying mathematics will be explained as far as it is beneficial for the application of the methods. The focus of the book is on the application side. We provide implementation in both Python and R of nearly all algorithms we explain throughout the book and make the code available for all the case studies we present in the book as well.

Additional material is available on the website of the book (ml4qs.org):

- Code examples are available in Python and R
- Datasets used in the book and additional sources to be explored by readers
- Up-to-date list of scientific papers and text books related to the book's theme

We have been researchers in this field for over ten years and would like to thank everybody who formed the body of knowledge that has become the basis for this book. First of all, we would like to thank the people at crowdsignals.io for providing us with the dataset that is used throughout the book, Evan Welbourne in particular. Furthermore, we want to thank the colleagues who contributed to the book: Dennis Becker, Ward van Breda, Vincent Bremer, Gusz Eiben, Eoin Grau, Evert Haasdijk, Ali el Hassouni, Floris den Hengst, and Bart Kamphorst. We also want to thank all the graduate students that participated in the Machine Learning for the Quantified Self course at the Vrije Universiteit Amsterdam in June 2017 and provided feedback on a preliminary version of the book that was used as reader during the course. Mark would like to thank (in the order of appearance in his academic career) Maria Gini, Catholijn Jonker, Jan Treur, Gusz Eiben, and Peter Szolovits for being such great sources of inspiration.

And of course, the writing of this book would not have been possible without our loving family and friends. Mark would specifically like to thank his parents for their continuous support and his friends for helping him in getting the proper relaxation in the busy book-writing period. Burkhardt is very grateful to his family, especially his wife Karen Funk and his two daughters, for allowing him to often work late and to spend almost half a year at the University of Virginia and Stanford University during his sabbatical.

Amsterdam, The Netherlands
Lüneburg, Germany
August 2017

Mark Hoogendoorn
Burkhardt Funk

Contents

1	Introduction	1
1.1	The Quantified Self	2
1.2	The Goal of this Book	4
1.3	Basic Terminology	5
1.3.1	Data Terminology	5
1.3.2	Machine Learning Terminology	7
1.4	Basic Mathematical Notation	8
1.5	Overview of the Book	10
 Part I Sensory Data and Features		
2	Basics of Sensory Data	15
2.1	Crowdsignals Dataset	15
2.2	Converting the Raw Data to an Aggregated Data Format	17
2.3	Exploring the Dataset	19
2.4	Machine Learning Tasks	23
2.5	Exercises	24
2.5.1	Pen and Paper	24
2.5.2	Coding	24
3	Handling Noise and Missing Values in Sensory Data	25
3.1	Detecting Outliers	27
3.1.1	Distribution-Based Models	28
3.1.2	Distance-Based Models	30
3.2	Imputation of Missing Values	34
3.3	A Combined Approach: The Kalman Filter	35
3.4	Transformation	37
3.4.1	Lowpass Filter	38
3.4.2	Principal Component Analysis	38

- 3.5 Case Study 42
 - 3.5.1 Outlier Detection 43
 - 3.5.2 Missing Value Imputation 45
 - 3.5.3 Kalman Filter 46
 - 3.5.4 Data Transformation 47
- 3.6 Exercises. 49
 - 3.6.1 Pen and Paper 49
 - 3.6.2 Coding. 50
- 4 Feature Engineering Based on Sensory Data 51**
 - 4.1 Time Domain 51
 - 4.1.1 Numerical Data 52
 - 4.1.2 Categorical Data 54
 - 4.1.3 Mixed Data 56
 - 4.2 Frequency Domain 58
 - 4.2.1 Fourier Transformations 58
 - 4.2.2 Features in Frequency Domain 60
 - 4.3 Features for Unstructured Data 62
 - 4.3.1 Pre-processing Text Data. 62
 - 4.3.2 Bag of Words 63
 - 4.3.3 TF-IDF 63
 - 4.3.4 Topic Modeling. 64
 - 4.4 Case Study 65
 - 4.4.1 Time Domain 66
 - 4.4.2 Frequency Domain 67
 - 4.4.3 New Dataset 68
 - 4.5 Exercises. 69
 - 4.5.1 Pen and Paper 69
 - 4.5.2 Coding. 70
- Part II Learning Based on Sensory Data**
- 5 Clustering 73**
 - 5.1 Learning Setup 73
 - 5.2 Distance Metrics 74
 - 5.2.1 Individual Data Points Distance Metrics 74
 - 5.2.2 Person Level Distance Metrics 77
 - 5.3 Non-hierarchical Clustering 82
 - 5.4 Hierarchical Clustering 84
 - 5.4.1 Agglomerative Clustering 84
 - 5.4.2 Divisive Clustering 87
 - 5.5 Subspace Clustering 88
 - 5.6 Datastream Clustering. 91
 - 5.7 Performance Evaluation 93

- 5.8 Case Study 94
 - 5.8.1 Non-hierarchical Clustering 94
 - 5.8.2 Hierarchical Clustering 98
- 5.9 Exercises. 98
 - 5.9.1 Pen and Paper 98
 - 5.9.2 Coding. 100
- 6 Mathematical Foundations for Supervised Learning 101**
 - 6.1 Learning Process and Elements 101
 - 6.1.1 Unknown Target Function. 102
 - 6.1.2 Observed Data. 104
 - 6.1.3 Error Measure 105
 - 6.1.4 Hypothesis Set and the Learning Machine. 107
 - 6.1.5 Model Selection and Evaluation 111
 - 6.2 Learning Theory 114
 - 6.2.1 PAC Learnability. 114
 - 6.2.2 VC-Dimension and VC-Bound 116
 - 6.2.3 Implications. 118
 - 6.3 Exercises. 120
 - 6.3.1 Pen and Paper 120
 - 6.3.2 Coding. 121
- 7 Predictive Modeling without Notion of Time. 123**
 - 7.1 Learning Setup 123
 - 7.2 Feedforward Neural Networks 125
 - 7.2.1 Perceptron 125
 - 7.2.2 Multi-layer Perceptron 128
 - 7.2.3 Convolutional Neural Networks. 129
 - 7.3 Support Vector Machines 131
 - 7.4 K-Nearest Neighbor 134
 - 7.5 Decision Trees 135
 - 7.6 Naive Bayes 139
 - 7.7 Ensembles. 140
 - 7.7.1 Bagging 141
 - 7.7.2 Boosting 141
 - 7.8 Predictive Modeling for Data Streams 144
 - 7.9 Practical Considerations 145
 - 7.9.1 Feature Selection 145
 - 7.9.2 Regularization 147
 - 7.10 Case Study 148
 - 7.10.1 Classification: Predicting the Activity Label 149
 - 7.10.2 Regression: Predicting the Heart Rate 157

7.11	Exercises.	163
7.11.1	Pen and Paper	163
7.11.2	Coding.	164
8	Predictive Modeling with Notion of Time	167
8.1	Learning Setup	167
8.2	Time Series Analysis	168
8.2.1	Basic Concepts	169
8.2.2	Filtering and Smoothing	170
8.2.3	Autoregressive Integrated Moving Average Model—ARIMA	173
8.2.4	Estimating and Forecasting Time Series Models	176
8.2.5	Example Application	177
8.3	Neural Networks.	181
8.3.1	Recurrent Neural Networks	182
8.3.2	Echo State Networks	184
8.4	Dynamical Systems Models	186
8.4.1	Example Based on Bruce’s Data	186
8.4.2	Parameter Optimization	188
8.5	Case Study	195
8.5.1	Tuning Parameters.	195
8.5.2	Results.	197
8.6	Exercises.	201
8.6.1	Pen and Paper	201
8.6.2	Coding.	201
9	Reinforcement Learning to Provide Feedback and Support	203
9.1	Basic Setting.	203
9.2	One-Step SARSA Temporal Difference Learning	208
9.3	Q-Learning	210
9.4	SARSA(λ) and Q(λ).	211
9.5	Approximate Solutions	212
9.6	Discretizing the State Space	212
9.7	Exercises.	213
9.7.1	Pen and Paper	213
9.7.2	Coding.	214
 Part III Discussion		
10	Discussion	217
10.1	Learning Full Circle	217
10.2	Heterogeneity	218
10.3	Effective Data Collection and Reuse.	219
10.4	Data Processing and Storage.	219

10.5	Better Predictive Modeling and Clustering	220
10.6	Validation	221
References	223
Index	229