An Introduction to Data Analysis using Aggregation Functions in R

Simon James

An Introduction to Data Analysis using Aggregation Functions in R

Springer

Simon James
School of Information Technology
Deakin University
Burwood, VIC, Australia

ISBN 978-3-319-46761-0 ISBN 978-3-319-46762-7 (eBook)
DOI 10.1007/978-3-319-46762-7

Library of Congress Control Number: 2016953188

© Springer International Publishing AG 2016
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.

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

Preface

This introduction to the area of aggregation and data analysis is intended for computer science or business students who are interested in tools for summarizing and interpreting data, but who do not necessarily have a formal mathematics background. It was motivated by our need for such a text in postgraduate data analytics subjects at my university; however it could also be of interest to undergraduate data science students. By reading through the chapters and completing the questions, you will be introduced to some of the issues that can arise when we try to make sense of data. It is hoped that you will gain an appreciation for the usefulness of the results established in the field of data aggregation and their wide applicability. I personally find the study of aggregation functions to be very enjoyable and fulfilling, as it allows for a nice mix of theoretically interesting results and practical applications. The rise of data analytics and the growing need for companies to learn more from their data also make the need for intelligent data aggregation techniques more indispensable than ever.

The overall aim of this book is to allow future data analysts to become aware of aggregation functions theory and methods in an accessible way, focusing on a fundamental understanding of the data and summarization tools that complements any study in statistical or machine learning techniques. To this end, included in each of the chapters is an `R` tutorial giving an introduction to commands and techniques relevant to the topics studied. These tutorials assume no programming background and will give you some exposure to one of the most widely adopted (and freely available) computing languages used in data analysis.

For a more in-depth overview of aggregation functions, there are some great existing monographs, including the following to name a few.

- Beliakov, G., Bustince, H. and Calvo, T.: A Practical Guide to Averaging Functions. Springer, Berlin, New York (2015)
- Beliakov, G., Pradera, A. and Calvo, T.: Aggregation Functions: A Guide for Practitioners. Springer, Heidelberg (2007)
- Gagolewski, M.: Data Fusion. Theory, Methods and Applications. Institute of Computer Science, Polish Academy of Sciences (2015)

- Grabisch, M., Marichal, J.-L., Mesiar, R. and Pap, E.: Aggregation Functions. Cambridge University press, Cambridge, (2009)
- Torra, V. and Narukawa, Y.: Modeling Decisions. Information Fusion and Aggregation Operators. Springer, Berlin, Heidelberg (2007)

Whilst focusing on established results, I hope that the following pages will offer a fresh perspective on the recent trends in aggregation research. In this respect I am indebted to a number of colleagues and researchers who lead this field, many of whom I have been fortunate enough to meet or collaborate with over the span of my short career. In particular, I would like to acknowledge Humberto Bustince, Radko Mesiar, and Michal Baczynski, who have all made me feel very welcome when visiting their institutions and participating in conferences overseas. Special mention should be made regarding the works of Michel Grabisch and Ronald R. Yager on fuzzy integrals and OWA operators respectively, which have been a huge influence in my understanding of these areas (and so I hope I do them justice in the short introductions provided). I am grateful to my friend and ecology research expert, Dale Nimmo, who has helped identify a number of interesting applications aggregation theory. A huge thank-you to Marek Gagolewski for our recent discussions and for his extensive comments on drafts of this text; I am looking forward to our future research endeavors! Finally (on the academic side), I would like to thank Gleb Beliakov for continuing to be a mentor to my research. Certainly the majority of these chapters have been influenced by what you have taught me and our work together.

Lastly I would like to acknowledge the support of family (in particular my Mum, who proofread some chapters of this book), friends (especially Rachel, who has been a great support while I've been working on this book in addition to her usual honest and constructive comments), and my work colleagues (a very busy 2016 was made much more bearable by having people like Lauren, Elicia, Guy, Tim, Lei, Crystal, Menuri, Gang, Shui, Vicky, Sutharshan, and Michelle to talk to).

Melbourne, VIC, Australia Simon James
August 2016

Contents

1 Aggregating Data with Averaging Functions 1

1.1 The Problem in Data: Average Returns 2

1.2 Background Concepts 4

1.2.1 What is a Function? 4

1.2.2 Multivariate Functions and Vectors 7

1.3 The Arithmetic Mean 10

1.3.1 Definition 10

1.3.2 Properties 12

1.4 The Median 14

1.5 The Geometric and Harmonic Means 16

1.5.1 The Geometric Mean 16

1.5.2 More Properties 18

1.5.3 The Harmonic Mean 19

1.6 Averaging Aggregation Functions 21

1.7 Uses of Aggregation Functions 23

1.7.1 Sporting Statistics 24

1.7.2 Simple Forecasting 24

1.7.3 Indices of Diversity and Equity 25

1.8 Summary of Formulas 25

1.9 Practice Questions 26

1.10 R Tutorial 27

1.10.1 Entering Commands in the Console 27

1.10.2 Basic Mathematical Operations 27

1.10.3 Assignment of Variables 28

1.10.4 Assigning a Vector to a Variable 29

1.10.5 Basic Operations on Vectors 29

1.10.6 Existing Functions 30

1.10.7 Creating New Functions 32

1.11 Practice Questions Using R 33

References 34

2 Transforming Data 37
2.1 The Problem in Data: Multicriteria Evaluation 38
2.1.1 Which is Better: Higher or Lower? 39
2.1.2 Consistent Scales 39
2.1.3 Differences in Distribution 39
2.2 Background Concepts 41
2.2.1 Arrays and Matrices (**X**) 41
2.2.2 Matrix/Array Entries (x_{ij}) 41
2.2.3 Matrix/Array Rows and Columns ($\mathbf{x}_i, \mathbf{x}_j$) 42
2.3 Negations and Utility Transformations 42
2.4 Scaling, Standardization and Normalization 46
2.5 Log and Polynomial Transformations 50
2.6 Piecewise-Linear Transformations 52
2.7 Functions Built from Transformations 56
2.8 Power Means 57
2.9 Quasi-Arithmetic Means 60
2.10 Summary of Formulas 61
2.11 Practice Questions 62
2.12 R Tutorial 63
2.12.1 Replacing Values 63
2.12.2 Arrays and Matrices 64
2.12.3 Reading a Table 65
2.12.4 Transforming Variables 67
2.12.5 Rank-Based Scores 67
2.12.6 Using `if()` for Cases 68
2.12.7 Plotting in Two Variables 69
2.12.8 Defining Power Means 71
2.13 Practice Questions Using R 72
References 72

3 Weighted Averaging 75
3.1 The Problem in Data: Group Decision Making 76
3.2 Background Concepts 78
3.2.1 Regression Parameters 78
3.3 Weighting Vectors 79
3.3.1 Interpreting Weights 81
3.3.2 Example: Welfare Functions 81
3.4 Weighted Power Means 82
3.5 Weighted Medians 84
3.6 Examples of Other Weighting Conventions 85
3.6.1 Simpson's Dominance Index 85
3.6.2 Entropy 86
3.7 The Borda Count 87
3.8 Summary of Formulas 88
3.9 Practice Questions 89

3.10 R Tutorial 91
3.10.1 Weighted Arithmetic Means 91
3.10.2 Weighted Power Means 91
3.10.3 Default Values for Functions 92
3.10.4 Weighted Median 92
3.10.5 Borda Counts 93
3.11 Practice Questions Using R 95
References 95

4 Averaging with Interaction 97
4.1 The Problem in Data: Supplementary Analytics 98
4.2 Background Concepts 99
4.2.1 Trimmed and Winsorized Mean 99
4.3 Ordered Weighted Averaging 101
4.3.1 Definition 101
4.3.2 Properties 102
4.3.3 Special Cases 102
4.3.4 Orness 103
4.3.5 Defining Weighting Vectors 104
4.4 The Choquet Integral 108
4.4.1 Fuzzy Measures 108
4.4.2 Definition 110
4.4.3 Special Cases 111
4.4.4 Calculation 111
4.4.5 Examples 113
4.4.6 Properties 115
4.4.7 The Shapley Value 116
4.5 Summary of Formulas 117
4.6 Practice Questions 118
4.7 R Tutorial 120
4.7.1 OWA Operator 121
4.7.2 Choquet Integral 122
4.7.3 Orness 123
4.7.4 Shapley Values 125
4.8 Practice Questions Using R 125
References 126

5 Fitting Aggregation Functions to Empirical Data 129
5.1 The Problem in Data: Recommender Systems 130
5.2 Background Concepts 134
5.2.1 Optimization and Linear Constraints 135
5.2.2 Evaluating and Interpreting a Model's Accuracy 139
5.2.3 Flexibility and Overfitting 142
5.3 Learning Weights for Aggregation Functions 143
5.3.1 Fitting a Weighted Power (or Quasi-Arithmetic) Mean 143
5.3.2 Fitting an OWA Function 144
5.3.3 Fitting the Choquet Integral 145

5.4 Using Aggregation Models for Analysis and Prediction 146
5.4.1 Comparing Different Averaging Functions 149
5.4.2 Making Inferences About the Importance of Each Variable 155
5.4.3 Make Inferences About Tendency Toward Lower or Higher Inputs 156
5.4.4 Predicting the Outputs for Unknown/New Data 156
5.5 Reliability 157
5.5.1 Do Our y_i Values Represent the ‘Ground Truth’? 157
5.5.2 Are We Evaluating an Approach or the Model Itself? 158
5.6 Conclusions 159
5.7 Summary of Formulas 160
5.8 Practice Questions Using R 161
References 162

6 Solutions 163
Aggregating Data with Averaging Functions: Solutions 163
1.9 Practice Questions 163
1.10 Practice Questions Using R 165
Transforming Data: Solutions 168
2.11 Practice Questions 168
2.13 Practice Questions Using R 172
Weighted Aggregation: Solutions 176
3.9 Practice Questions 176
3.11 Practice Questions Using R 180
Averaging with Interaction: Solutions....................................... 182
4.6 Practice Questions 182
4.7 Practice Questions Using R 190
Fitting Aggregation Functions to Empirical Data: Solutions............... 192
5.6 Practice Questions Using R 192

Index... 197