Machine Learning for Text

Charu C. Aggarwal

Machine Learning for Text

Second Edition

Springer

Charu C. Aggarwal
Mohegan Lake
NY, USA

A Solution Manual to this book can be downloaded from https://link.springer.com/book/10.1007/978-3-030-96623-2

ISBN 978-3-030-96622-5 ISBN 978-3-030-96623-2 (eBook)
https://doi.org/10.1007/978-3-030-96623-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my wife Lata, my daughter Sayani,
and my late parents Dr. Prem Sarup and Mrs. Pushplata Aggarwal.

Preface

> "If it is true that there is always more than one way of construing a text, it is not true that all interpretations are equal."– Paul Ricoeur

The rich area of text analytics draws ideas from information retrieval, machine learning, and natural language processing. Each of these areas is an active and vibrant field in its own right, and numerous books have been written in each of these different areas. As a result, many of these books have covered some aspects of text analytics, but they have not covered all the areas that a book on learning from text is expected to cover.

At this point, a need exists for a focussed book on machine learning from text. This book is a first attempt to integrate all the complexities in the areas of machine learning, information retrieval, and natural language processing in a holistic way, in order to create a coherent and integrated book in the area. Therefore, the chapters are divided into three categories:

1. *Fundamental algorithms and models:* Many fundamental applications in text analytics, such as matrix factorization, clustering, and classification, have uses in domains beyond text. Nevertheless, these methods need to be tailored to the specialized characteristics of text. Chapters 1 through 8 will discuss core analytical methods in the context of machine learning from text.

2. *Information retrieval and ranking:* Many aspects of information retrieval and ranking are closely related to text analytics. For example, ranking SVMs and link-based ranking are often used for learning from text. Chapter 9 will provide an overview of information retrieval methods from the point of view of text mining.

3. *Sequence- and natural language-centric models:* Although multidimensional representations can be used for basic applications in text analytics, the true richness of the text representation can be leveraged by treating text as sequences. Chapters 10 through 16 will discuss these advanced topics like sequence embedding, deep learning, transformers, pre-trained language models, information extraction, knowledge graphs, summarization, question-answering, opinion mining, text segmentation, and event extraction.

Because of the diversity of topics covered in this book, some careful decisions have been made on the scope of coverage. A complicating factor is that many machine learning techniques depend on the use of basic natural language processing and information retrieval methodologies. This is particularly true of the sequence-centric approaches discussed in Chapters 10

through 16 that are more closely related to natural language processing. Examples of analytical methods that rely on natural language processing include information extraction, event extraction, opinion mining, and text summarization, which frequently leverage basic natural language processing tools like linguistic parsing or part-of-speech tagging. Needless to say, natural language processing is a full fledged field in its own right (with excellent books dedicated to it). Therefore, a question arises on how much discussion should be provided on techniques that lie on the interface of natural language processing and text mining without deviating from the primary scope of this book. Our general principle in making these choices has been to focus on *mining* and *machine learning* aspects. If a specific natural language or information retrieval method (e.g., part-of-speech tagging) is not *directly* about text analytics, we have illustrated how to *use* such techniques (as black-boxes) rather than discussing the internal algorithmic details of these methods. Basic techniques like part-of-speech tagging have matured in algorithmic development, and have been commoditized to the extent that many open-source tools are available with little difference in relative performance. Therefore, we only provide working definitions of such concepts in the book, and the primary focus will be on their utility as off-the-shelf tools in mining-centric settings. The book provides pointers to the relevant books and open-source software in each chapter in order to enable additional help to the student and practitioner.

The book is written for graduate students, researchers, and practitioners. The exposition has been simplified to a large extent, so that a graduate student with a reasonable understanding of linear algebra and probability theory can understand the book easily. Numerous exercises are available along with a solution manual to aid in classroom teaching.

Throughout this book, a vector or a multidimensional data point is annotated with a bar, such as $\overline{X}$ or $\overline{y}$. A vector or multidimensional point may be denoted by either small letters or capital letters, as long as it has a bar. Vector dot products are denoted by centered dots, such as $\overline{X} \cdot \overline{Y}$. A matrix is denoted in capital letters without a bar, such as R. Throughout the book, the $n \times d$ document-term matrix is denoted by D, with n documents and d dimensions. The individual documents in D are therefore represented as d-dimensional row vectors, which are the bag-of-words representations. On the other hand, vectors with one component for each data point are usually n-dimensional column vectors. An example is the n-dimensional column vector $\overline{y}$ of class variables of n data points.

What Is New in the Second Edition

The second edition of the book emphasizes deep learning and natural language processing. Chapter 10 on deep learning has been significantly enhanced with discussions on different types of neural networks as well as language models like ELMo. Chapter 11 is entirely new, and it discusses transformers and pre-trained language models. Deep learning methods have been added to the chapter on text summarization with a special focus on abstractive summarization (Chapter 12). The information extraction chapter has now been updated to an integrated chapter on information extraction and knowledge graphs. The addition of knowledge graphs also lays the ground for a completely new chapter on question-answering (Chapter 14). Deep learning methods for sentiment analysis are also introduced in the book.

Mohegan Lake, NY, USA Charu C. Aggarwal

The original version of the book has been revised. A correction to this book can be found at https://doi.org/10.1007/978-3-030-96623-2_17

Acknowledgments

Acknowledgements for the First Edition

I would like to thank my family including my wife, daughter, and my parents for their love and support. I would also like to thank my manager Nagui Halim for his support during the writing of this book.

This book has benefitted from significant feedback and several collaborations that i have had with numerous colleagues over the years. I would like to thank Quoc Le, Chih-Jen Lin, Chandan Reddy, Saket Sathe, Shai Shalev-Shwartz, Jiliang Tang, Suhang Wang, and ChengXiang Zhai for their feedback on various portions of this book and for answering specific queries on technical matters. I would particularly like to thank Saket Sathe for commenting on several portions, and also for providing some sample output from a neural network to use in the book. For their collaborations, I would like to thank Tarek F. Abdelzaher, Jing Gao, Quanquan Gu, Manish Gupta, Jiawei Han, Alexander Hinneburg, Thomas Huang, Nan Li, Huan Liu, Ruoming Jin, Daniel Keim, Arijit Khan, Latifur Khan, Mohammad M. Masud, Jian Pei, Magda Procopiuc, Guojun Qi, Chandan Reddy, Saket Sathe, Jaideep Srivastava, Karthik Subbian, Yizhou Sun, Jiliang Tang, Min-Hsuan Tsai, Haixun Wang, Jianyong Wang, Min Wang, Suhang Wang, Joel Wolf, Xifeng Yan, Mohammed Zaki, ChengXiang Zhai, and Peixiang Zhao. I would particularly like to thank Professor ChengXiang Zhai for my earlier collaborations with him in text mining. I would also like to thank my advisor James B. Orlin for his guidance during my early years as a researcher.

Finally, I would like to thank Lata Aggarwal for helping me with some of the figures created using PowerPoint graphics in this book.

Acknowledgements for the Second Edition

I would like to thank Roy Lee, Chandan Reddy, and Jiliang Tang for their feedback on the second edition of the book.

Contents

Author Biography

Charu C. Aggarwal is a Distinguished Research Staff Member (DRSM) at the IBM T. J. Watson Research Center in Yorktown Heights, New York. He completed his undergraduate degree in Computer Science from the Indian Institute of Technology at Kanpur in 1993 and his Ph.D. from the Massachusetts Institute of Technology in 1996.

He has worked extensively in the field of data mining. He has published more than 400 papers in refereed conferences and journals and authored over 80 patents. He is the author or editor of 20 books, including textbooks on data mining, recommender systems, and outlier analysis. Because of the commercial value of his patents, he has thrice been designated a Master Inventor at IBM. He is a recipient of an IBM Corporate Award (2003) for his work on bio-terrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, and a recipient of two IBM Outstanding Technical Achievement Awards (2009, 2015) for his work on data streams/high-dimensional data. He received the EDBT 2014 Test of Time Award for his work on condensation-based privacy-preserving data mining. He is a recipient of the IEEE ICDM Research Contributions Award (2015) and ACM SIGKDD Innovation Award, which are the two most prestigious awards for influential research contributions in the field of data mining. He is also a recipient of the W. Wallace McDowell Award, which is the highest award given solely by the IEEE Computer Society across the field of Computer Science.

He has served as the general co-chair of the IEEE Big Data Conference (2014) and as the program co-chair of the ACM CIKM Conference (2015), the IEEE ICDM Conference (2015), and the ACM KDD Conference (2016). He served as an associate editor of the IEEE Transactions on Knowledge and Data Engineering from 2004 to 2008. He is an associate editor of the IEEE Transactions on Big Data, an action editor of the Data Mining and Knowledge Discovery Journal, and an associate editor of the Knowledge and Information Systems Journal. He has served or currently serves as the editor-in-chief of the ACM Transactions on Knowledge Discovery from Data as well as the ACM SIGKDD Explorations. He is also an editor-in-chief of ACM Books. He serves on the advisory board of the Lecture Notes on Social Networks, a publication by Springer. He has served as the vice-president of the SIAM Activity Group on Data Mining and is a member of the SIAM industry committee. He is a fellow of the SIAM, ACM, and the IEEE, for "contributions to knowledge discovery and data mining algorithms."