

Preface

A search for the word “zettabyte” will return a page that predicts that we will enter the zettabyte age around 2015. To store this amount of data on DVDs would require over 215 billion disks. The search itself (on September 20, 2011) returned 775,000 results, many relevant, many irrelevant, and many duplicates. The challenge is to elicit knowledge from all this data.

Scientific endeavors are constantly generating more and more data. As new generations of instruments are created, one of the characteristics is typically more sensitive results. In turn, this typically means more data is generated. New techniques made available by new instrumentation, techniques, and understanding allows us to consider approaches such as genome-wide association studies (GWAS) that were outside of our ability to consider just a few years ago. Again, the challenge is to elicit knowledge from all this data.

But as we continue to generate this ever-increasing amount of data, we would also like to know what relationships and patterns exist between the data. This, in essence, is the goal of data mining: find the patterns within the data. This is what this book is about. Is there some quantity X that is related to some other quantity Y that isn't obvious to us? If so, what could those relationships tell us? Is there something novel, something new, that these patterns tell us? Can it advance our knowledge?

There is no obvious end in sight to the increasing generation of data. To the contrary, as tools, techniques, and instrumentation continue to become smaller, cheaper, and thus, more available, it is likely that the opposite will be the case and data will continue to be generated in ever-increasing volumes. It is for this reason that automated approaches to processing data, understanding data, and finding these patterns will be even more important.

This leads to the major challenge of a book like this: what to include and what to leave out. We felt it important to cover as much of the theory of data mining as possible, including statistical, analytical, visualization, and machine learning techniques. Much exciting work is being done under the umbrella of machine learning, and much of it is seeing fruition within the data mining discipline itself. To say that

this covers a wide area – and a multitude of sins – is an understatement. To those readers who question why we included a particular topic, and to those who question why we omitted some other topic, we can only apologize for this. In writing a book aimed at introducing data mining techniques to the life sciences, describing a broad range of techniques is a necessity to allow the researcher to select the most appropriate tools for his/her investigation. Many of the techniques we discuss are not necessarily in widespread use, but we believe can be valuable to the researcher.

Many people over many years triggered our enthusiasm for developing this text and thanks to them that we have created this book. Particular thanks goes to Bruce Lucarelli and Viswanath Balasubramanian for their contributions and insights on various parts of the text.

Cincinnati, OH, USA

Rob Sullivan



<http://www.springer.com/978-1-58829-942-0>

Introduction to Data Mining for the Life Sciences

Sullivan, R.

2012, XVIII, 638 p., Hardcover

ISBN: 978-1-58829-942-0

A product of Humana Press