

Preface

The data mining literature contains many excellent titles that address the needs of users with a variety of interests ranging from decision making to pattern investigation in biological data. However, these books do not deal with the mathematical tools that are currently needed by data mining researchers and doctoral students and we felt that it is timely to produce a new version of our book that integrates the mathematics of data mining with its applications. We emphasize that this book is about mathematical tools for data mining and *not* about data mining itself; despite this, many substantial applications of mathematical concepts in data mining are included. The book is intended as a reference for the working data miner.

We present several areas of mathematics that, in our opinion are vital for data mining: *set theory*, including partially ordered sets and combinatorics; *linear algebra*, with its many applications in linear algorithms; *topology* that is used in understanding and structuring data, and *graph theory* that provides a powerful tool for constructing data models.

Our set theory chapter begins with a study of functions and relations. Applications of these fundamental concepts to such issues as equivalences and partitions are discussed. We have also included a précis of universal algebra that covers the needs of subsequent chapters.

Partially ordered sets are important on their own and serve in the study of certain algebraic structures, namely lattices, and Boolean algebras. This is continued with a combinatorics chapter that includes such topics as the inclusion–exclusion principle, combinatorics of partitions, counting problems related to collections of sets, and the Vapnik–Chervonenkis dimension of collections of sets.

An introduction to topology and measure theory is followed by a study of the topology of metric spaces, and of various types of generalizations and specializations of the notion of metric. The dimension theory of metric spaces is essential for recent preoccupations of data mining researchers with the applications of fractal theory to data mining.

A variety of applications in data mining are discussed, such as the notion of entropy, presented in a new algebraic framework related to partitions rather than random distributions, level-wise algorithms that generalize the Apriori technique, and generalized measures and their use in the study of frequent item sets.

Linear algebra is present in this new edition with three chapters that treat linear spaces, norms and inner products, and spectral theory. The inclusion of these

chapters allowed us to expand our treatment of graph theory and include many new applications.

A final chapter is dedicated to clustering that includes basic types of clustering algorithms, techniques for evaluating cluster quality, and spectral clustering.

The text of this second edition, which appears 7 years after the publication of the first edition, was reorganized, corrected, and substantially amplified. Each chapter ends with suggestions for further reading. Over 700 exercises and supplements are included; they form an integral part of the material. Some of the exercises are in reality supplemental material. For these, we include solutions. The mathematics required for making the best use of our book is a typical three-semester sequence in calculus.

Boston, January 2014
Villeneuve d'Ascq

Dan A. Simovici
Chabane Djeraba

Mathematical Tools for Data Mining
Set Theory, Partial Orders, Combinatorics

Simovici, D.; Djeraba, C.

2014, XI, 831 p. 93 illus., Hardcover

ISBN: 978-1-4471-6406-7