Mathematics of Data Science
A Computational Approach to Clustering and Classification
Daniela Calvetti and Erkki Somersalo

This self-contained textbook provides a solid mathematical basis for understanding popular data science algorithms for clustering and classification and shows that an in-depth understanding of the mathematics powering these algorithms gives insight into the underlying data. It presents a step-by-step derivation of these algorithms, outlining their implementation from scratch in a computationally sound way. The book proposes different ways of visualizing high-dimensional data to unveil hidden internal structures and includes graphical explanations and computed examples using publicly available data sets.

2020 • x + 189 pages • Softcover • 978-1-611976-36-6
List $64.00 • SIAM Member $44.80 • DI01

Nonnegative Matrix Factorization
Nicolas Gillis

Nonnegative matrix factorization (NMF) in its modern form has become a standard tool in the analysis of high-dimensional data sets. This book provides a comprehensive and up-to-date account of the most important aspects of the NMF problem and is the first book to cover in detail the theoretical aspects of NMF, including geometric interpretation, nonnegative rank, complexity, and uniqueness. It explains why understanding these theoretical insights is key to using this computational tool effectively and meaningfully. Nonnegative Matrix Factorization is accessible to a wide audience and is ideal for anyone interested in the workings of NMF. It discusses new results on identifiability and complexity and the separable NMF and contains MATLAB codes for readers to run numerical examples.

2020 • xxvi + 350 pages • Softcover • 978-1-611976-40-3
List $87.00 • SIAM Member $60.90 • DI02

Mining Imperfect Data
With Examples in R and Python
Second Edition
Ronald K. Pearson

It has been estimated that as much as 80% of the total effort in a typical data analysis project is taken up with data preparation, including reconciling and merging data from different sources, identifying and interpreting various data anomalies, and selecting and implementing appropriate treatment strategies for the anomalies that are found. This book focuses on the identification and treatment of data anomalies, including examples that highlight different types of anomalies, their potential consequences if left undetected and untreated, and options for dealing with them. The book also emphasizes the range of open-source tools available for identifying and treating data anomalies, mostly in R but also with several examples in Python.

2020 • xxvi + 481 pages • Softcover • 978-1-611976-26-7
List $94.00 • SIAM Member $65.80 • MN04

Data Clustering
Theory, Algorithms, and Applications
Second Edition
Guojun Gan, Chaoqun Ma, and Jianhong Wu

Data clustering, also known as cluster analysis, is an unsupervised process that divides a set of objects into homogeneous groups. Development in the area has exploded, especially in clustering algorithms for big data and open-source software for cluster analysis. This second edition reflects these new developments, covers the basics of data clustering, includes a list of popular clustering algorithms, and provides program code that helps users implement clustering algorithms.

2020 • xxiv + 406 pages • Softcover • 978-1-611976-32-8
List $94.00 • SIAM Member $65.80 • MN05