
Reviewed by Andrew Binstock

March 18, 2020

Although big data was trendy five years ago, today organizations are trying to make sense of the flood of data they now have. Because data mining proved insufficient, the largely ignored field of AI has finally come into its own. And out of AI's many subdisciplines, machine learning has emerged as the principal means for organizations to gain a true understanding of their data.

If machine learning has caught your interest, you'll find many videos and tutorials on the web. Most are of limited value, either because they are too theoretical or because they focus on only one tool in the lengthy machine-learning toolchain. The scarcity of useful offerings makes the present book such a satisfying find. It's a hands-on overview of machine-learning, written for the Java developer.

The author, who has more than a decade of experience working with the vast data troves generated from retail sales, starts with a full explanation of what machine learning is and what it can deliver.

He then moves into the prerequisites: finding the data and preparing it for use by the tools that are presented across the remaining 12 chapters. After explaining the sources of usable data (public datasets, APIs, web scraping), the cleaning of data for use, and the classification of data, he provides an overview of the statistics you typically will use in analyzing the data.

After explaining the statistical methods, such as linear regression, he launches into decision trees and presents them using the Weka toolkit. That toolkit is used again in the next chapter on clustering (there, referring to the machine-learning technique of identifying items with common traits, rather than to hardware architecture). As the author does in every chapter, he starts with an approachable explanation of what the technique is, why it's important, and what tools can be used to implement it. These presentations are remarkably clear. Due to the well-thought-out explanations, many times, I came to understand items that had previously been mere concepts. These presentations lead immediately to implementation code using a small dataset. That code is usually Java, and in some places the equivalent operations are shown in Clojure as well. One chapter, though, has substantial new material shown in Clojure only.

These chapters are followed by coverage of association rules learning and support vector machines (SVMs), both of which are demonstrated using Weka. Then comes a very lengthy presentation on neural nets and implementations using both Weka and Deeplearning4J (commonly referred to as DL4J).
The rest of the book (which comprises some 360 pages plus appendices) provides demonstrations of the machine-learning concepts as applied to specific types of data. The first data type is text, which uses Apache Tika and Google’s word2vec. The author then moves on to images, which serve as a way to introduce convolutional neural networks.

Finally, there come two chapters on the big data engines commonly associated with machine learning: Kafka and Spark. The chapter on Kafka spans more than 60 pages and includes an illuminating explanation of the technology, followed by the steps to set up a single-node cluster and, later, by the steps to set up larger clusters. Not only does the author show the code for the sample project, but he carefully steps you through the planning necessary for the project and the decisions that were made in this particular implementation. Unfortunately for Java developers, a large section is implemented in Compojure, a Clojure-based web framework.

The chapter on Spark is shorter and frequently relies on comparisons with the preceding coverage of Kafka.

A final chapter, which really should be viewed as an appendix, is an introduction to the R language, which is often used in machine learning. This concise coverage is intended as an aid to the reader who might be working with colleagues who use R, so the basic explanation of R offered here facilitates communication.

Overall, this is about as fine and complete an introduction to machine learning as is presently available for Java developers. It’s an excellent work that I can recommend, but not entirely without reservations. There were some irksome aspects. The first and most irritating is the poor presentation of code. In some chapters, keywords are bolded but in others they are not. In some, statics are italicized but in others they are not. In all chapters, the code frequently wraps to the beginning of the next line. I don’t understand such carelessness, especially in light of how trivially easy it is to format code properly.

There are other lapses, which are frustrating things to find in any professional book. For example, one chapter includes a full-page listing for which the entire explanation consists of a single sentence. Elsewhere, the author refers to things incorrectly and in other places he is inconsistent with his explanations. All books have some annoyances, but this is a second edition, so it’s reasonable to expect that all these failings would have been removed.

It’s too bad that there is this sloppiness in an otherwise excellent book. While I still believe this is the best introduction presently available for Java developers, with more care it could be the definitive treatment.

---

**Andrew Binstock**

Andrew Binstock (@platypusguy) was formerly the editor in chief of *Java Magazine*. Previously, he was the editor of *Dr. Dobbs Journal*. He co-founded the company behind the open-source iText PDF library, which was acquired in 2015. His book on algorithm implementation in C went through 16 printings before joining the long tail. Previously, he was the editor in chief of *UNIX Review* and, earlier, the founding editor of the *C Gazette*. He lives in Silicon Valley with his wife. When not coding or editing, he studies piano.