Probability and Statistics for Machine Learning: A Textbook

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, New York

November 8, 2023

To my wife Lata, my daughter Sayani,
and all my mathematics teachers

About the Book

This book teaches probability and statistics from the machine learning perspective. The chapters of this book belong to three categories:

1. *The basics of probability and statistics:* These chapters focus on the basics of probability and statistics, and cover the key principles of these topics. Chapter 1 provides an overview of the area of probability and statistics and its relationship to machine learning. The fundamentals of probability and statistics are covered in Chapters 2 through 5.

2. *From probability to machine learning:* Many machine learning applications are addressed using probabilistic models, whose parameters are then learned in a data-driven manner. Chapter 6 through 9 explore how different models from probability and statistics are applied to machine learning. Perhaps the most important tool that bridges the gap from data to probability is maximum-likelihood estimation, which is a foundational concept from the perspective of machine learning.

3. *Advanced topics:* Chapter 10 is devoted to discrete-state Markov processes. It explores the application of probability and statistics to a temporal and sequential setting, although the applications extended to more complex settings such as graphical data. Chapter 11 covers a number of useful concepts in extreme-value analysis.

The style of writing promotes the learning of probability and statistics simultaneously with a probabilistic perspective on the modeling of machine learning applications. The book contains over 200 worked examples in order to elucidate key concepts. Exercises are included both within the text of the chapters and at the end of the chapters. The book is written for a broad audience, including graduate students, researchers, and practitioners.

About the Author

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. in Operations Research from the Massachusetts Institute of Technology in 1996.

He has published more than 400 papers in refereed conferences and journals, and has applied for or been granted more than 80 patents. He is author or editor of 20 books, including textbooks on linear algebra, machine learning (for text), neural networks, recommender systems, and outlier analysis. Because of the commercial value of his patents, he has thrice been designated a Master Inventor at IBM. He has received several internal and external awards, including the EDBT Test-of-Time Award (2014), the ACM SIGKDD Innovation Award (2019), and the IEEE ICDM Research Contributions Award (2015). He is also a recipient of the W. Wallace McDowell Award, the highest award given solely by the IEEE Computer Society across the field of computer science. He has served as an editor-in-chief of the ACM SIGKDD Explorations and is currently serving as an editor-in-chief of the ACM Transactions on Knowledge Discovery from Data as well as ACM Books. He is a fellow of the SIAM, ACM, and the IEEE, for "contributions to knowledge discovery and data mining algorithms."

Contents

Preface

"Lies, damned lies, and statistics." — Mark Twain

Most of machine learning is directly or indirectly related to probability and statistics. After all, machine learning is all about making predictions based on data, which inevitably leads to statistical methods. These statistical methods are often couched as *models*, which use *probabilities* to quantify the likelihoods of events. Therefore, having a strong background in probability and statistics is critical.

The familiarity required with probability and statistics often goes well beyond what is taught in undergraduate curricula. As a result, this presents a challenge to beginners in the field. In many cases, the type of techniques required from probability and statistics are specific to machine learning, which is not covered by generic courses on probability and statistics. This book therefore develops a treatment of probability and statistics from the specific perspective of machine learning.

This book teaches probability and statistics with a specific focus on machine learning applications. As a natural consequence of this approach many key concepts in machine learning are covered in detail. Therefore, it is possible to learn a significant amount of machine learning during the process of learning probability and statistics in this book. The chapters of this book are organized as follows:

1. *The basics of probability and statistics:* These chapters focus on the basics of probability and statistics, and cover the key principles of these topics. Chapter 1 provides an overview of the area of probability and statistics and its relationship to machine learning. The fundamentals of probability and statistics are covered in Chapters 2 through 5.

2. *From probability to machine learning:* Many machine learning applications are addressed using probabilistic models, whose parameters are then learned in a data-driven manner. Chapter 6 through 9 explore how different models from probability and statistics are applied to machine learning. Perhaps the most important tool that bridges the gap from data to probability is maximum-likelihood estimation, which is a foundational concept from the perspective of machine learning.

3. *Advanced topics:* Chapter 10 is devoted to discrete-state Markov processes. It explores the application of probability and statistics to a temporal and sequential setting, although the applications extended to more complex settings such as graphical data. Chapter 11 covers a number of useful concepts in extreme-value analysis.

More than 200 worked examples are provided in the book in order to elucidate different concepts. Furthermore, the book contains unsolved exercises both within and at the end of chapters. The worked examples should be solved without looking at the solution as one reads the chapter. This will lead to slower progress but a better understanding. In-chapter exercises are often similar to worked examples — hints for solving the more difficult of these exercises are often given immediately after the exercise in order to help the reader along. Exercises at the end of the chapter are intended to be solved as refreshers after completing the chapter. An instructor solution manual is available containing solutions to end-of-chapter exercises. There are a total of about 600 (solved, in-chapter, and end-of-chapter) exercises in the book. Therefore, the book provides ample opportunity for practice.

Prerequisites for the Book

The main challenge with writing such a book is that it is grounded in a solid understanding of basic mathematics. A knowledge of calculus (at the high-school level) is absolutely essential for understanding the book. There are some concepts that also require a notation-level understanding of multivariate and vector calculus but these concepts are largely self-explanatory in nature (and only require an understanding of basic definitions). A basic understanding of vectors is needed, although a detailed understanding of linear algebra is not assumed. The only concept that is used repeatedly in the book is the concept of eigenvectors and principal component analysis, which is described from first principles in Chapter 2. Except for a single (clearly demarcated) section in Chapter 10, advanced concepts in linear algebra are not needed for understanding the material. The book makes an effort to point out sections of the book that can be skipped over without loss in continuity (by clearly demarcating them). These sections are somewhat advanced and used only occasionally in machine learning. The corresponding sections have been marked by an asterisk (*) in the section header.

Notations

Throughout this book, a vector or a multidimensional data point is annotated with a vector right arrow, such as $\vec{X}$ or $\vec{y}$. 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 $\vec{x} \cdot \vec{y}$. A matrix is denoted in capital letters without a vector symbol, such as R. Random variables are also denoted by capital letters, and the difference between a random variable and a matrix is usually obvious from the underlying context. Samples of random variables are denoted by small letters. Throughout the book, the $n \times d$ matrix corresponding to the entire training data set is denoted by D, with n data points and d dimensions. The individual data points in D are therefore d-dimensional row vectors, and are often denoted by $\vec{x}_1 \ldots \vec{x}_n$. Note that these vectors use small letters rather than capital letters, because they are assumed to be samples from some underlying data distribution. Vectors with one component for each data point (observation) are usually n-dimensional column vectors. An example is the n-dimensional column vector $\vec{y}$ of class variables of n data points. An observed value y_i is distinguished from a predicted value $\hat{y}_i$ by a circumflex at the top of the variable.