

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

1	Probability and Statistics: An Introduction	1
1.1	Introduction	1
1.1.1	The Interplay Between Probability, Statistics, and Machine Learning	2
1.2	Representing Data	2
1.2.1	Numeric Multidimensional Data	3
1.2.2	Categorical and Mixed Attribute Data	4
1.3	Summarizing and Visualizing Data	6
1.4	The Basics of Probability and Probability Distributions	8
1.4.1	Populations versus Samples	12
1.4.2	Modeling Populations with Samples	13
1.4.3	Handling Dependence in Data Samples	15
1.5	Hypothesis Testing	15
1.6	Basic Problems in Machine Learning	16
1.6.1	Clustering	16
1.6.2	Classification and Regression Modeling	17
1.6.3	Outlier Detection	20
1.7	Summary	21
1.8	Further Reading	22
1.9	Exercises	22
2	Summarizing and Visualizing Data	23
2.1	Introduction	23
2.2	Summarizing Data	24
2.2.1	Univariate Summarization	24
2.2.2	Multivariate Summarization	31
2.3	Data Visualization	44
2.3.1	Univariate Visualization	44
2.3.2	Multivariate Visualization	50
2.4	Applications to Data Preprocessing	55
2.4.1	Univariate Preprocessing Methods	55
2.4.2	Whitening: A Multivariate Preprocessing Method	57
2.5	Summary	59
2.6	Further Reading	60

2.7	Exercises	60
3	Probability Basics and Random Variables	63
3.1	Introduction	63
3.2	Sample Spaces and Events	64
3.3	The Counting Approach to Probabilities	71
3.4	Set-Wise View of Events	72
3.5	Conditional Probabilities and Independence	75
3.6	The Bayes Rule	77
3.6.1	The Observability Perspective: Posteriors versus Likelihoods	80
3.7	The Basics of Probability Distributions	81
3.7.1	Closed-Form View of Probability Distributions	82
3.7.2	Continuous Distributions	84
3.7.3	Multivariate Probability Distributions	87
3.8	Distribution Independence and Conditionals	90
3.8.1	Independence of Distributions	90
3.8.2	Conditional Distributions	91
3.8.3	Example: A Simple 1-Dimensional Knowledge-Based Bayes Classifier	94
3.9	Summarizing Distributions	95
3.9.1	Expectation and Variance	96
3.9.2	Distribution Covariance	102
3.9.3	Useful Multivariate Properties Under Independence	104
3.10	Compound Distributions	106
3.10.1	Total Probability Rule in Continuous Hypothesis Spaces	107
3.10.2	Bayes Rule in Continuous Hypothesis Spaces	109
3.11	Functions of Random Variables (*)	112
3.11.1	Distribution of the Function of a Single Random Variable	112
3.11.2	Distribution of the Sum of Random Variables	115
3.11.3	Geometric Derivation of Distributions of Functions	117
3.12	Summary	119
3.13	Further Reading	120
3.14	Exercises	120
4	Probability Distributions	125
4.1	Introduction	125
4.2	The Uniform Distribution	126
4.3	The Bernoulli Distribution	129
4.4	The Categorical Distribution	130
4.5	The Geometric Distribution	133
4.6	The Binomial Distribution	136
4.7	The Multinomial Distribution	139
4.8	The Exponential Distribution	143
4.9	The Poisson Distribution	147
4.10	The Normal Distribution	150
4.10.1	Multivariate Normal Distribution: Independent Attributes	157
4.10.2	Multivariate Normal Distribution: Dependent Attributes	159
4.11	The Student's t -Distribution	163
4.12	The χ^2 -Distribution	168
4.13	Mixture Distributions: The Realistic View	173

4.13.1	Why Mixtures are Ubiquitous: A Motivating Example	174
4.13.2	The Basic Generative Process of a Mixture Model	175
4.13.3	Some Useful Results for Prediction	176
4.13.4	The Conditional Independence Assumption	177
4.14	Moments of Random Variables (*)	178
4.14.1	Central and Standardized Moments	178
4.14.2	Moment Generating Functions	180
4.15	Summary	183
4.16	Further Reading	183
4.17	Exercises	183
5	Hypothesis Testing and Confidence Intervals	187
5.1	Introduction	187
5.2	The Central Limit Theorem	189
5.3	Sampling Distribution and Standard Error	190
5.4	The Basics of Hypothesis Testing	192
5.4.1	Confidence Intervals	196
5.4.2	When Population Standard Deviations Are Not Available	199
5.4.3	The One-Tailed Hypothesis Test	202
5.5	Hypothesis Tests For Differences in Means	204
5.5.1	Unequal Variance t -Test	204
5.5.1.1	Tightening the Degrees of Freedom	208
5.5.2	Equal Variance t -Test	208
5.5.3	Paired t -Test	210
5.6	χ^2 -Hypothesis Tests	213
5.6.1	Standard Deviation Hypothesis Test	213
5.6.2	χ^2 -Goodness-of-Fit Test	216
5.6.3	Independence Tests	218
5.7	Analysis of Variance (ANOVA)	220
5.8	Machine Learning Applications of Hypothesis Testing	223
5.8.1	Evaluating the Performance of a Single Classifier	226
5.8.2	Comparing Two Classifiers	227
5.8.3	χ^2 -Statistic for Feature Selection in Text	228
5.8.4	Fisher Discriminant Index for Feature Selection	229
5.8.5	Fisher Discriminant Index for Classification (*)	231
5.9	Summary	235
5.10	Further Reading	236
5.11	Exercises	236
6	Reconstructing Probability Distributions from Data	239
6.1	Introduction	239
6.2	Maximum Likelihood Estimation	241
6.2.1	Comparing Likelihoods with Posteriors	246
6.3	Reconstructing Common Distributions from Data	246
6.3.1	The Uniform Distribution	246
6.3.2	The Bernoulli Distribution	247
6.3.3	The Geometric Distribution	248
6.3.4	The Binomial Distribution	249
6.3.5	The Multinomial Distribution	251

6.3.6	The Exponential Distribution	252
6.3.7	The Poisson Distribution	253
6.3.8	The Normal Distribution	254
6.3.9	Multivariate Distributions with Dimension Independence	256
6.3.10	Gaussian Distribution with Dimension Dependence	257
6.4	Mixture of Distributions: The EM Algorithm	259
6.5	Kernel Density Estimation	265
6.6	Reducing Reconstruction Variance	268
6.6.1	Variance in Maximum Likelihood Estimation	269
6.6.2	Prior Beliefs with Maximum A Posteriori (MAP) Estimation	272
6.6.3	Kernel Density Estimation: Role of Bandwidth	277
6.7	The Bias-Variance Trade-Off	279
6.8	Popular Distributions Used as Conjugate Priors (*)	283
6.8.1	Gamma Distribution	284
6.8.2	Beta Distribution	286
6.8.3	Dirichlet Distribution	288
6.9	Summary	292
6.10	Further Reading	292
6.11	Exercises	293
7	Regression	295
7.1	Introduction	295
7.2	The Basics of Regression	296
7.2.1	Interpreting the Coefficients	297
7.2.2	Feature Engineering Trick for Dropping Bias	297
7.2.3	Regression: A Central Problem in Statistics and Linear Algebra	299
7.3	Two Perspectives on Linear Regression	300
7.3.1	The Linear Algebra Perspective	300
7.3.2	The Probabilistic Perspective	302
7.4	Solutions to Linear Regression	306
7.4.1	Closed-Form Solution to Squared-Loss Regression	306
7.4.2	The Case of One Non-Trivial Predictor Variable	310
7.4.3	Solution with Gradient Descent for Squared Loss	313
7.4.4	Gradient Descent For L_1 -Loss Regression	316
7.5	Handling Categorical Predictors	316
7.6	Overfitting and Regularization	318
7.6.1	Closed-Form Solution for Regularized Formulation	321
7.6.2	Solution Based on Gradient Descent	322
7.6.3	LASSO Regularization	323
7.7	A Probabilistic View of Regularization	323
7.8	Evaluating Linear Regression	326
7.8.1	Evaluating In-Sample Properties of Regression	326
7.8.2	Out-of-Sample Evaluation	330
7.9	Nonlinear Regression	331
7.9.1	Interpretable Feature Engineering	332
7.9.2	Explicit Feature Engineering with Similarity Matrices	335
7.9.3	Implicit Feature Engineering with Similarity Matrices	338
7.10	Summary	341
7.11	Further Reading	341

7.12 Exercises	341
8 Classification: A Probabilistic View	343
8.1 Introduction	343
8.2 Generative Probabilistic Models	344
8.2.1 Continuous Numeric Data: The Gaussian Distribution	347
8.2.2 Binary Data: The Bernoulli Distribution	351
8.2.3 Sparse Numeric Data: The Multinomial Distribution	355
8.2.4 Plate Diagrams for Generative Processes	358
8.3 Loss-Based Formulations: A Probabilistic View	360
8.3.1 Least-Squares Classification	362
8.3.1.1 The Probabilistic Interpretation and Its Problems	365
8.3.1.2 Practical Issues with Least Squares Classification	365
8.3.2 Logistic Regression	366
8.3.2.1 Maximum Likelihood Estimation for Logistic Regression	367
8.3.2.2 Gradient Descent and Stochastic Gradient Descent	370
8.3.2.3 Interpreting Updates in Terms of Error Probabilities	370
8.3.3 Multinomial Logistic Regression	371
8.3.3.1 The Probabilistic Model	372
8.3.3.2 Maximum Likelihood Estimation	372
8.3.3.3 Probabilistic Interpretation of Gradient Descent Updates	374
8.4 Beyond Classification: Ordered Logit Model	375
8.4.1 Maximum Likelihood Estimation for Ordered Logit	376
8.5 Summary	377
8.6 Further Reading	378
8.7 Exercises	378
9 Unsupervised Learning: A Probabilistic View	381
9.1 Introduction	381
9.2 Mixture Models for Clustering	382
9.2.1 Continuous Numeric Data: The Gaussian Distribution	386
9.2.2 Binary Data: The Bernoulli Distribution	390
9.2.3 Sparse Numeric Data: The Multinomial Distribution	392
9.3 Matrix Factorization	395
9.3.1 The Squared Loss Model	396
9.3.1.1 Probabilistic Interpretation of Squared Loss	397
9.3.1.2 Regularization	399
9.3.1.3 Application to Incomplete Data: Recommender Systems	400
9.3.2 Probabilistic Latent Semantic Analysis	401
9.3.2.1 Example of PLSA	405
9.3.2.2 Alternative Plate Diagram for PLSA	406
9.3.3 Logistic Matrix Factorization	407
9.4 Outlier Detection	409
9.4.1 The Mahalanobis Method: A Probabilistic View of Whitening	410
9.4.2 Mixture Models in Outlier Detection	414
9.4.3 Matrix Factorization for Outlier detection	415
9.5 Summary	418
9.6 Further Reading	418
9.7 Exercises	418

10 Discrete State Markov Processes	421
10.1 Introduction	421
10.2 Markov Chains	423
10.2.1 Steady-State Behavior of Markov Chains	426
10.2.2 Transient Behavior of Markov Chains	428
10.2.3 Periodic Markov Chains	432
10.2.4 Ergodicity	434
10.2.5 Different Cases of Ergodicity and Non-Ergodicity	437
10.2.6 Properties and Applications of Non-Ergodic Markov Chains	438
10.2.7 Probabilities of Absorbing Outcomes	444
10.2.8 The View from Matrix Algebra (*)	447
10.3 Machine Learning Applications of Markov Chains	449
10.3.1 PageRank	449
10.3.2 Application to Vertex Classification	453
10.4 Markov Chains to Generative Models	457
10.5 Hidden Markov Models	457
10.5.1 Formal Definition and Techniques for HMMs	461
10.5.2 Evaluation: Computing the Fit Probability for Observed Sequence	462
10.5.3 Explanation: Determining the Most Likely State Sequence for Observed Sequence	463
10.5.4 Training: Baum-Welch Algorithm	463
10.6 Applications of Hidden Markov Models	465
10.6.1 Mixture of HMMs for Clustering	465
10.6.2 Outlier Detection	467
10.6.3 Classification	467
10.7 Summary	468
10.8 Further Reading	468
10.9 Exercises	469
11 Probabilistic Inequalities and Extreme Value Analysis	471
11.1 Introduction	471
11.2 Jensen's Inequality	472
11.3 Markov and Chebychev Inequalities	476
11.4 Approximations for Sums of Random Variables	480
11.4.1 The Chernoff Bound	482
11.4.2 The Normal Approximation to the Binomial Distribution	486
11.4.3 The Poisson Approximation to the Binomial Distribution	489
11.4.4 The Hoeffding Inequality	490
11.5 Comparing Tail Inequalities with Approximation Estimates	494
11.6 Summary	496
11.7 Further Reading	497
11.8 Exercises	497

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 $\vec{x}_1 \dots \vec{x}_n$. 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.