Machine Learning for Engineers

This self-contained introduction to machine learning, designed from the start with engineers in mind, will equip students and researchers with everything they need to start applying machine learning principles and algorithms to real-world engineering problems. With a consistent emphasis on the connections between estimation, detection, information theory, and optimization, it includes: an accessible overview of the relationships between machine learning and signal processing, providing a solid foundation for further study; clear explanations of the differences between state-of-the-art machine learning techniques and conventional model-driven methods, equipping students and researchers with the tools necessary to make informed technique choices; demonstration of the links between information-theoretical concepts and their practical engineering relevance; and reproducible examples using MATLAB®, enabling hands-on experimentation. Assuming only a basic understanding of probability and linear algebra, and accompanied by lecture slides and solutions for instructors, this is the ideal introduction to machine learning for engineering students of all disciplines

Osvaldo Simeone is Professor of Information Engineering at King's College London, where he directs the King's Communications, Learning & Information Processing (KCLIP) lab. He is a Fellow of the IET and of the IEEE.

Machine Learning for Engineers

Osvaldo Simeone
King's College London

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781316512821
DOI: 10.1017/9781009072205

First published 2022

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Simeone, Osvaldo, author.
Title: Machine learning for engineers : principles and algorithms through signal processing and information theory / Osvaldo Simeone, King's College London.
Description: First edition. | Cambridge ; New York, NY : Cambridge University Press, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2022010244 | ISBN 9781316512821 (hardback)
Subjects: LCSH: Engineering–Data processing. | Machine learning. |
BISAC: TECHNOLOGY & ENGINEERING / Signals & Signal Processing
Classification: LCC TA345 .S5724 2022 | DDC 620.00285–dc23/eng/20220504
LC record available at https://lccn.loc.gov/2022010244

ISBN 978-1-316-51282-1 Hardback

Additional resources for this publication at www.cambridge.org/simeone

Contents

Part I Introduction and Background

Part II Fundamental Concepts and Algorithms

Part III Advanced Tools and Algorithms

Part V Epilogue

Preface

Overview

This book provides a self-contained introduction to the field of machine learning through the lens of signal processing and information theory for an audience with an engineering background. This preface explains why the book was written and what you will find in it.

Background and Motivation

Advances in machine learning and artificial intelligence (AI) have made available new tools that are revolutionizing science, engineering, and society at large. Modern machine learning techniques build on conceptual and mathematical ideas from stochastic optimization, linear algebra, signal processing, Bayesian inference, as well as information theory and statistical learning theory. Students and researchers working in different fields of engineering are now expected to have a general grasp of machine learning principles and algorithms, and to be able to assess the relative relevance of available design solutions spanning the space between model- and data-based methodologies. This book is written with this audience in mind.

In approaching the field of machine learning, students of signal processing and information theory may at first be ill at ease in reconciling the similarity between the techniques used in machine learning – least squares, gradient descent, maximum likelihood – with differences in terminology and emphasis (and hype?). Seasoned signal processing and information-theory researchers may in turn find the resurgence of machine learning somewhat puzzling ("didn't we write off that technique three decades ago?"), while still being awed by the scale of current applications and by the efficiency of state-of-the-art methods. They may also pride themselves on seeing many of the ideas originating in their communities underpin machine learning solutions that have wide societal and economic repercussions.

Existing books on the subject of machine learning come in different flavors: Some are compilations of algorithms mostly intended for computer scientists; and others focus on specific aspects, such as optimization, Bayesian reasoning, or theoretical principles. Books that have been used for many years as references, while still relevant, appear to be partly outdated and superseded by more recent research papers.

In this context, what seems to be missing is a textbook aimed at engineering students and researchers that can be used for self-study, as well as for undergraduate and graduate courses alongside modules on statistical signal processing, information theory, and optimization. An ideal text should provide a principled introduction to machine learning that highlights connections with estimation, detection, information theory, and optimization, while offering a concise but extensive coverage of state-of-the-art topics and simple, reproducible examples. Filling this gap in the bookshelves of engineering libraries is the ambition of this book.

As I end this first section without mentioning "deep learning", some readers may start to worry. Indeed, deep learning is seen in this book "merely" as a particularly effective way to define and train a specific class of models. I have attempted to place it in the context of related methods, and I have often preferred to focus on simpler models that are better suited to illustrate the underlying ideas. While running counter to the general trend on the computer science side of machine learning, the focus on simple, reproducible examples is intended as a means to strip away aspects of scale to reveal concepts, intuition, and key techniques.

Intended Audience

This book is intended for a general audience of students, engineers, and researchers with a background in probability and signal processing. (I also hope that it will be of some interest to students and researchers working in information theory, who may not be aware of some of the connections explored here.)

To offer a self-contained introduction to these intended readers, the text introduces supervised and unsupervised learning in a systematic fashion – including the necessary background on linear algebra, probability, and optimization – taking the reader from basic tools to state-of-the-art methods within a unified, coherent presentation. Information-theoretic concepts and metrics are used throughout the text to serve as training criteria and performance measures. Later chapters explore topics that are subject to intense research at the time of writing, in the hope that they can serve as launching pads for original research and new contributions.

Why This Book?

I am often asked by colleagues and students with a background in engineering to suggest "the best place to start" to get into the field of machine learning. I typically respond with a list of books: For a general, but slightly outdated introduction, read this book; for a detailed survey of methods based on Bayesian signal processing, check this other reference; to learn about optimization for machine learning, I found this text useful; for theoretical aspects, here is another text; and, for more recent developments, a list of papers is attached. Unfortunately, the size and number of these references may be intimidating.

I started writing this book in the form of lectures notes for a course taught at the New Jersey Institute of Technology (NJIT). My motivation was, and remains, that of distilling the array of references mentioned above into a single text that offers a balanced presentation connecting to other courses on statistical signal processing, information theory, and optimization. This initial effort led to the monograph [1], which contains some of the material covered here.

Years later, across the ocean, now with a family and forced to work from home, I have had a chance to revise the original text [1] for a module on machine learning I am currently teaching at King's College London (KCL). The current book was born as a result of this activity. It is meant to serve as a full textbook, including background material; expanded discussions on principles and algorithms, such as automatic differentiation, contrastive learning, and energy-based models; completely new chapters on optimization, variational inference and learning, information-theoretic learning, Bayesian learning, transfer learning, meta-learning, and federated learning, as well as on further topics for research; new examples and figures; and end-of-chapter problems.

The number of machine learning books is quite extensive and growing. There are classical books, such as the texts by Bishop [2] and Hastie, Tibshirani, and Friedman [3], which provide general introductions in the form of an extensive review of techniques, not including more recent developments. In this book, I have taken a different approach, striving for unification within the framework of information theory and probabilistic models, with the aim of also covering more advanced material. The text by Murphy [4] focuses on probabilistic models, and can serve as a complementary reference to fill in the details of specific techniques that are not detailed here. Theoretical books, such as the excellent reference by Shalev-Shwartz and Ben-David [5], may not be the most natural choice for researchers interested in general principles, intuition, and applications, rather than detailed theoretical derivations. The text by Watt, Borhani, and Katsaggelos [6] takes an optimization perspective, and can also be used as a complement for this book. The book by Theodoridis [7] is an extensive reference on Bayesian learning. For readers interested in programming for deep learning, a useful reference is the book by Zhang, Lipton, Li, and Smola [8]. There are also various books that focus on applications, but mostly in the form of edited contributions (e.g., [9]).

Using This Book

This book can be used for self-study, as a reference for researchers seeking as an entry point in the field of machine learning, as well as a textbook. Courses on machine learning are primarily taught in computer science departments, but they are also increasingly part of the portfolio of engineering departments. This text can be adopted for a senior undergraduate course or for a more advanced graduate course on machine learning within an engineering curriculum.

An undergraduate course, like the one I currently teach at KCL, can cover Part I and Part II by discussing Chapter 1 to Chapter 6 in full, while including a selection of topics from Chapter 7 (see the Organization section below). A more advanced course could be taught by selecting topics from Chapter 7, as well as Part III and Part IV.

A note of warning for the more formally minded readers: While I have tried to be precise in the use of notation and in the formulation of statements, I have purposely avoided a formal – theorem, proof, remark – writing style, attempting instead to follow a more “narrative” and intuitive approach. I have endeavored to include only results that are easy to interpret and prove, and the text includes short proofs for the main theoretical results. I have also excluded any explicit mention of measure-theoretic aspects, and some results are provided without detailing all the underlying technical assumptions.

Throughout the text, I have made an effort to detail simple, reproducible examples that can be programmed in MATLAB® or Python without making use of specialized libraries in a relatively short amount of time. In particular, all examples in the book have been programmed in MATLAB® (with one exception for Fig. 13.17). Complex, large-scale experiments can be easily found in the literature and in books such as [8].

End-of-chapter problems are provided for Chapter 2, Part II, and Part III. (For the topics covered in Part IV, the reader is invited to reproduce the experiments provided in the text and to search the literature for open problems and more complex experiments.) The problems offer a mix of analytical and programming tasks, and they can be solved using the material and tools covered in the text. Solutions for instructors can be found at the book's website. Questions requiring programming are indicated with an asterisk (*).

When teaching, I find it useful to discuss some of end-of-chapter problems as part of tutorial sessions that encompass both theoretical and programming questions. I have also found it easier to use MATLAB®, particularly with students having little expertise in programming.

A list of recommended resources is included at the end of each chapter. The main purpose of these notes is to offer pointers for further reading. I have not attempted to assign credit for the techniques presented in each chapter to the original authors, preferring instead to refer to textbooks and review papers.

In writing, editing, and re-editing the text, I was reminded of the famous spider (*Cyclosa octotuberculata* if you wish to look it up) that weaves information about the positions of previously observed preys into its web. Like the spider, we as engineers use mathematics as a form of extended cognition, outsourcing mental load to the environment. In making various passes through the text, I have picked up my previous trails on the web, discovering new ideas and correcting old ones. I realize that the process would be never-ending, weaving and unweaving, and I offer my apologies to the reader for any mistakes or omissions in the current text. A list of errors will be maintained at the book's website.

Organization

The text is divided into four main parts. The first part includes a general introduction to the field of machine learning and to the role of engineers in it, as well as a background chapter used to set the notation and review the necessary tools from probability and linear algebra. The second part introduces the basics of supervised and unsupervised learning, including algorithmic principles and theory. The third part covers more advanced material, encompassing statistical learning theory, exponential family of distributions, approximate Bayesian inference, information maximization and estimation, and Bayesian learning. Finally, the fourth part moves beyond conventional centralized learning to address topics such as transfer learning, meta-learning, and federated learning. Following is a detailed description of the chapters.

Part I: Introduction and Background
Chapter 1. When and How to Use Machine Learning. This chapter motivates the use of machine learning – a data-driven inductive bias-based design methodology – as opposed to the conventional model-driven domain knowledge-based design engineers are more accustomed to; it provides general guidelines for the use of machine learning; and it introduces the landscape of machine learning frameworks (supervised, unsupervised, and reinforcement learning).
Chapter 2. Background. Chapter 2 reviews the necessary key concepts and tools from probability and linear algebra.

Part II: Fundamental Concepts and Algorithms
Chapter 3. Inference, or Model-Driven Prediction. As a benchmark for machine learning, this chapter reviews optimal Bayesian inference, which refers to the ideal scenario in which the statistical model underlying data generation is known. The chapter focuses on detection and estimation under both hard and soft predictors. It provides an introduction to key information-theoretic metrics as performance measures adopted for optimal inference, namely cross entropy, Kullback–Liebler (KL) divergence, entropy, mutual information, and free energy.

Chapter 4. Supervised Learning: Getting Started. Chapter 4 covers the principles of supervised learning in the frequentist framework by introducing key definitions, such as inductive bias, model class, loss function, population loss, generalization, approximation and estimation errors, and overfitting and underfitting. It distinguishes between the problems of training hard and soft predictors, demonstrating the connections between the two formulations.

Chapter 5. Optimization for Machine Learning. The previous chapter considers only problems for which closed-form, or simple numerical, solutions to the learning problem are available. Chapter 5 presents background on optimization that is needed to move beyond the simple settings studied in Chapter 4, by detailing gradient descent, stochastic gradient descent, backpropagation, as well as their properties.

Chapter 6. Supervised Learning: Beyond Least Squares. This chapter builds on the techniques introduced in Chapters 4 and 5 to cover linear models and neural networks, as well as generative models, for binary and multi-class classification. The chapter also introduces mixture models and non-parametric techniques, as well as extensions to regression.

Chapter 7. Unsupervised Learning. Chapter 7 introduces unsupervised learning tasks within a unified framework based on latent-variable models, distinguishing directed discriminative and generative models, autoencoders, and undirected models. Specific algorithms such as K-means clustering, expectation maximization (EM), energy model-based training, and contrastive representation learning, are detailed. An underlying theme of the chapter is the use of supervised learning as a subroutine to solve unsupervised learning problems.

Part III: Advanced Tools and Algorithms

Chapter 8. Statistical Learning Theory. The previous chapters have left open an important question: How many samples are needed to obtain desirable generalization performance? This chapter addresses this question by reviewing the basics of statistical learning theory with an emphasis on probably approximately correct (PAC) learning theory. Limitations of the theory and open problems are also discussed.

Chapter 9. Exponential Family of Distributions. Another limitation of the previous chapters is their reliance on Gaussian and Bernoulli distributions. This chapter provides a general framework to instantiate probabilistic models based on the exponential family of distributions (which includes as special cases Gaussian and Bernoulli distributions). Generalized linear models (GLMs) are also introduced as conditional extensions of exponential-family models.

Chapter 10. Variational Inference and Variational Expectation Maximization. As discussed in Chapter 7, Bayesian inference is a key subroutine in enabling learning for models with latent variables, including most unsupervised learning techniques and mixture models for supervised learning. Exact Bayesian inference becomes quickly intractable for model sizes of practical interest, and, in order to scale up such methods, one needs to develop approximate Bayesian inference strategies. This chapter elaborates on such techniques by focusing on variational inference (VI) and variational EM (VEM). As an application of VEM, the chapter describes variational autoencoders (VAE).

Chapter 11. Information-Theoretic Inference and Learning. All the inference and learning problems studied in the previous chapters can be interpreted as optimizing specific information-theoretic metrics – most notably the free energy in the general framework of VI and VEM. This chapter explores this topic in more detail by covering both likelihood-based and likelihood-free learning problems. For likelihood-based models, problem formulations, such as maximum entropy, are introduced as generalized forms of VI and VEM. For likelihood-free models, two-sample estimators of information-theoretic metrics – most generally f-divergences – are described and leveraged to define generative adversarial networks (GANs).

Chapter 12. Bayesian Learning. This chapter provides an introduction to Bayesian learning, including motivation, examples, and main techniques. State-of-the-art parametric methods, such as stochastic gradient Langevin dynamics (SGLD), are covered, along with non-parametric methods such as Gaussian processes (GPs).

Part IV: Beyond Centralized Single-Task Learning
Chapter 13. Transfer Learning, Multi-task Learning, Continual Learning, and Meta-learning. As discussed in Part II and Part III, machine learning typically assumes that the statistical conditions during training match those during testing, and that training is carried out separately for each learning problem. This chapter introduces formulations of learning problems that move beyond this standard setting, including transfer learning, multi-task learning, continual learning, and meta-learning. Among the specific algorithms covered by the chapter are likelihood- and regularization-based continual learning, as well as model agnostic meta-learning (MAML). Bayesian perspectives on these formulations are also presented.

Chapter 14. Federated Learning. Previous chapters assume the standard computing architecture of centralized data processing. A scenario that is of increasing interest involves separate learners, each with its own local data set. Training algorithms that operate in a decentralized way without the exchange of local data sets are often labeled as "federated". This chapter provides an introduction to federated learning, presenting basic algorithms such as federated averaging (FedAvg) and generalizations thereof, covering privacy aspects via differential privacy (DP), and Bayesian solutions based on federated VI.

Part V: Epilogue
Chapter 15. Beyond This Book. This chapter provides a brief look at topics not covered in the main text, including probabilistic graphical models, causality, quantum machine learning, machine unlearning, and general AI.

Bibliography

[1] O. Simeone, "A brief introduction to machine learning for engineers," *Foundations and Trends in Signal Processing*, vol. 12, no. 3–4, pp. 200–431, 2018.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[4] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[5] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[6] J. Watt, R. Borhani, and A. Katsaggelos, *Machine Learning Refined: Foundations, Algorithms, and Applications*. Cambridge University Press, 2020.

[7] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, 2015.

[8] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. https://d2l.ai, 2020.

[9] M. R. Rodrigues and Y. C. Eldar, *Information-Theoretic Methods in Data Science*. Cambridge University Press, 2021.

Acknowledgements

I have read somewhere that writing is all about compressing time – the time, that is, spent by the writer in researching, selecting and organizing, and finally committing to page. A reader may need only five minutes to read a paragraph that the writer spent hours laboring on. Compression – writing – is successful if is nearly lossless, allowing the reader to reproduce the ideas in the mind of the writer with limited effort. Calibrating the level of compression has been an ongoing concern in writing this text, and I have been helped in making choices by readers of the original monograph [1] and by students at NJIT and KCL who have attended my courses on machine learning. A special mention goes to Kfir Cohen, who has patiently sifted through hundreds of slides for typos and imperfections (and prepared Fig. 7.2).

I have learned many of the topics covered in this book alongside colleagues and students: variational inference with Dr. Hyeryung Jang and Dr. Nicolas Skatchkovsky; Bayesian learning with Dr. Rahif Kassab; meta-learning with Dr. Sangwoo Park and Dr. Sharu Theresa Jose; differential privacy with Dr. Dongzhu Liu; information-theoretic learning with Dr. Jingjing Zhang; and the list goes on. Professor Bipin Rajendran has motivated me to look into probabilistic machine learning, and I am grateful for his illuminating explanations; Professor Petar Popovski has been inspirational as ever throughout the long period of gestation of this book and has provided very useful and detailed comments; Dr. Jakob Hoydis has provided useful initial feedback; and Dr. Onur Sahin has offered much needed reminders not to forget to use cases. I am grateful to all my colleagues at NJIT and at KCL, as well as my collaborators around the world, for making my work life so fulfilling. Thank you also to Dr. Hari Chittoor for spotting a number of typos, to Dr. Yusha Liu for preparing Fig. 12.5, and to Dr. Sangwoo Park for preparing Fig. 13.17.

I would like to gratefully acknowledge support by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 725732).

At Cambridge University Press, I have been lucky to work with Helen Shannon, Elizabeth Horne, and Jane Adams. Special thanks also to Julie Lancashire for believing in this project from the beginning and to Khaled Makhshoush for graciously agreeing to design the cover of this book. Finally, I am grateful to John King for his excellent and thoughtful work on editing the text.

This work is dedicated to my parents, to Lisa, Noah, Lena, and to my extended family across two continents.

Notation

General Conventions

- Random variables or random vectors – both abbreviated as rvs – are represented using Roman typeface, while their values and realizations are indicated by the corresponding standard font. For instance, the equality $\mathrm{x} = x$ indicates that rv x takes value x.
- An exception is made for rvs denoted by Greek letters, for which the same symbol is used for both rvs and realizations, with their use made clear by the context.
- All vectors are taken to be in column form.
- Matrices are indicated using uppercase fonts, with Roman typeface used for random matrices.
- Calligraphic fonts are used for sets.
- The distribution of an rv x, which may be either a probability mass function (pmf) for a discrete rvs or a probability density function (pdf) for continuous rvs, is denoted as $p(x)$. To specify a given numerical value x for rv x, we will also write $p(\mathrm{x} = x)$.
- For the Kullback–Liebler (KL) divergence, which involves two distributions $p(x)$ and $q(x)$, we use both notations $\mathrm{KL}(p||q)$ and $\mathrm{KL}(p(x)||q(x))$. The latter notation is particularly useful for conditional distributions $p(x|y)$ and $q(x|y)$, in which case the KL divergence $\mathrm{KL}(p(x|y)||q(x|y))$ is evaluated for a fixed value y. The same approach is applied to the other information-theoretic metrics listed below.
- In numerical calculations, we will often only keep the first two decimal digits by rounding up the second digit at each intermediate step.

Notations

General

$\mathbb{R}$	set of real numbers
$\mathbb{R}^N$	set of $N \times 1$ vectors of real numbers
$\mathbb{R}^+$	set of non-negative real numbers
$(a, b]$	interval between two real numbers a and b with $a < b$, excluding a and including b
(a, b)	interval between two real numbers a and b with $a < b$, excluding a and b
$[a, b]$	interval between two real numbers a and b with $a < b$, including a and b
$\{1, \ldots, K\}$	set including all integer numbers from 1 to K
$\{(a_k)_{k=1}^K\}$ or $\{a_k\}_{k=1}^K$	set $\{a_1, \ldots, a_K\}$
$x_{\mathcal{S}}$	set of elements x_k indexed by the integers $k \in \mathcal{S}$
$\mathcal{S}^N$	set of all $N \times 1$ vectors with entries taking values in set $\mathcal{S}$
$\lvert\mathcal{S}\rvert$	cardinality of a set $\mathcal{S}$
$\propto$	proportional to
$\exists$	there exists
$\sum_x$	sum over all values of variable x

Functions

$f(\cdot)$	a function
$f(x)$	output of function $f(x)$ for input x, or the function itself
$\nabla f(x)$	gradient vector of function $f(x)$
$\nabla^2 f(x)$	Hessian matrix of function $f(x)$
$\log(\cdot)$	natural logarithm
$\log_2(\cdot)$	logarithm in base 2
$\lvert \cdot \rvert$	absolute value (magnitude)
$\mathbb{1}(\cdot)$	indicator function: it equals 1 if the argument is true and 0 otherwise
$\delta(\cdot)$	Dirac delta function or Kronecker delta function ($\delta(x-x') = \mathbb{1}(x=x')$)
$\sigma(\cdot)$	sigmoid function: $\sigma(x) = 1/(1+\exp(-x))$
$\lceil \cdot \rceil$	ceiling function (smallest larger integer)
$\mathrm{step}(\cdot)$	step function: $\mathrm{step}(x) = 1$ if $x > 0$, and $\mathrm{step}(x) = 0$ if $x < 0$
$\min_x f(x)$	minimization problem for function $f(\cdot)$ over x
$\arg\min_x f(x)$	an element x in the set of minimizers of function $f(x)$
$\mathcal{O}(f(x))$	a function of the form $af(x) + b$ for some constants a and b

Linear Algebra

$[x]_i$	ith element of vector x
$[A]_{ij}$ or $[A]_{i,j}$	(i,j)th element of matrix A
$[A]_{r:}$ or $[A]_{:c}$	rth row and cth column of matrix A, respectively
x^T and X^T	transpose of vector x and matrix X
$\lvert\lvert a\rvert\rvert^2 = \sum_{i=1}^N a_i^2$	quadratic, or ℓ_2, norm of a vector $a = [a_1, \ldots, a_N]^T$
$\lvert\lvert a\rvert\rvert_1 = \sum_{i=1}^N \lvert a_i\rvert$	ℓ_1 norm
$\lvert\lvert a\rvert\rvert_0$	ℓ_0 pseudo-norm, which returns the number of non-zero entries of vector a
I_L	$L \times L$ identity matrix
I	identity matrix when dimension is clear from context
0_L	$L \times L$ all-zero matrix or $L \times 1$ all-zero vector, as clear from the context
1_L	$L \times 1$ all-one vector
$\det(\cdot)$	determinant of a square matrix
$\mathrm{tr}(\cdot)$	trace of a square matrix
$\mathrm{diag}(\cdot)$	column vector of elements on the main diagonal of the argument matrix
$\mathrm{Diag}(\cdot)$	square diagonal matrix with main diagonal given by the argument vector
$\odot$	element-wise product

Probability

$\mathrm{x} \sim p(x)$	rv x is distributed according to distribution $p(x)$
$\mathrm{x}_n \underset{\text{i.i.d.}}{\sim} p(x)$	rvs $\mathrm{x}_n \sim p(x)$ are independent and identically distributed (i.i.d.)
$p(x\vert y)$ or $p(x\vert\mathrm{y} = y)$	conditional distribution of x given the observation of rv $\mathrm{y} = y$
$(\mathrm{x}\vert\mathrm{y} = y) \sim p(x\vert y)$	rv x is drawn according to the conditional distribution $p(x\vert\mathrm{y} = y)$
$\mathrm{E}_{\mathrm{x}\sim p(x)}[\cdot]$	expectation of the argument over the distribution of the rv $\mathrm{x} \sim p(x)$
$\mathrm{E}_{\mathrm{x}\sim p(x\vert y)}[\cdot]$	conditional expectation of the argument over the distribution $p(x\vert y)$

$\Pr_{\mathrm{x}\sim p(x)}[\cdot]$	probability of the argument over the distribution of the rv $\mathrm{x} \sim p(x)$
$\Pr[\cdot]$	probability of the argument when the distribution is clear from the context
$\mathrm{Bern}(x\|q)$	Bernoulli pmf with parameter $q \in [0,1]$
$\mathrm{Cat}(x\|q)$	categorical pmf with parameter vector q
$\mathcal{N}(x\|\mu,\Sigma)$	multivariate Gaussian pdf with mean vector μ and covariance matrix Σ
$\mathrm{Beta}(z\|a,b)$	beta distribution with parameters a and b
$\mathcal{U}(x\|a,b)$	uniform pdf in the interval $[a,b]$
$\mathrm{Var}(p(x))$ or $\mathrm{Var}(p)$	variance of rv $\mathrm{x} \sim p(x)$
$\mathrm{Var}(\mathrm{x})$	variance of rv $\mathrm{x} \sim p(x)$ when the distribution $p(x)$ is clear from the context

Information-Theoretic Metrics

$\mathrm{H}(p(x))$ or $\mathrm{H}(p)$	entropy of rv $\mathrm{x} \sim p(x)$
$\mathrm{H}(\mathrm{x})$	entropy of rv $\mathrm{x} \sim p(x)$ when the distribution $p(x)$ is clear from the context
$\mathrm{H}(\mathrm{x}\|\mathrm{y})$	conditional entropy of rv x given rv y
$\mathrm{I}(\mathrm{x};\mathrm{y})$	mutual information between rvs x and y
$\mathrm{H}(p\|\|q)$	cross entropy between distributions p and q
$\mathrm{KL}(p\|\|q)$	KL divergence between distributions p and q
$\mathrm{F}(p\|\|\tilde{p})$	free energy for distribution p and unnormalized distribution $\tilde{p}$
$\mathrm{JS}(p\|\|q)$	Jensen–Shannon divergence between distributions p and q
$\mathrm{D}_f(p\|\|q)$	f-divergence between distributions p and q
$\mathrm{IPM}(p\|\|q)$	integral probability metric between distributions p and q

Learning-Related Quantities

$\mathcal{D}$	training set
$L_p(\theta)$	population loss as a function of the model parameter θ
$L_{\mathcal{D}}(\theta)$	training loss as a function of the model parameter θ
$\theta_{\mathcal{D}}^{ERM}$	model parameter obtained via empirical risk minimization (ERM)
$\theta_{\mathcal{D}}^{ML}$	model parameter obtained via maximum likelihood (ML)
$\theta_{\mathcal{D}}^{MAP}$	model parameter obtained via maximum a posteriori (MAP)
$\mathrm{ExpFam}(x\|\eta)$	distribution in the exponential family with natural parameter vector η
$\mathrm{ExpFam}(x\|\mu)$	distribution in the exponential family with mean parameter vector μ

Acronyms

ADF	Assumed Density Filtering
AI	Artificial Intelligence
BFL	Bayesian Federated Learning
BN	Bayesian Network
CDF	Cumulative Distribution Function
CDL	Contrastive Density Learning
CRL	Contrastive Representation Learning
DAG	Directed Acyclic Graph
DP	Differential Privacy
DV	Donsker–Varadhan
ELBO	Evidence Lower BOund
EM	Expectation Maximization
EP	Expectation Propagation
ERM	Empirical Risk Minimization
EWC	Elastic Weight Consolidation
FIM	Fisher Information Matrix
FL	Federated Learning
FVI	Federated Variational Inference
GAN	Generative Adversarial Network
GD	Gradient Descent
GGN	Generalized Gauss–Newton (GGN)
GLM	Generalized Linear Model
GOFAI	Good Old Fashioned AI
GP	Gaussian Process
GVEM	Generalized Variational Expectation Maximization
GVI	Generalized Variational Inference
i.i.d.	independent identically distributed
IPM	Integral Probability Metric
ITM	Information-Theoretic Measure
JS	Jensen–Shannon
KDE	Kernel Density Estimation
KL	Kullback–Leibler
K-NN	K-Nearest Neighbors
LDR	Log-Distribution Ratio
LLR	Log-Likelihood Ratio
LS	Least Squares
MAML	Model Agnostic Meta-Learning
MAP	Maximum A Posteriori
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo

MDL	Minimum Description Length
ML	Maximum Likelihood
MMD	Maximum Mean Discrepancy
MRF	Markov Random Field
OMP	Orthogonal Matching Pursuit
PAC	Probably Approximately Correct
PCA	Principal Component Analysis
pdf	probability density function
PGM	Probabilistic Graphical Model
pmf	probability mass function
QDA	Quadratic Discriminant Analysis
RBM	Restricted Boltzmann Machine
RKHS	Reproducing Kernel Hilbert Space
rv	random variable, or random vector
SGD	Stochastic Gradient Descent
SGLD	Stochastic Gradient Langevin Dynamics
SG-MCMC	Stochastic Gradient Markov Chain Monte Carlo
SVGD	Stein Variational Gradient Descent
TV	Total Variation
VAE	Variational AutoEncoder
VC	Vapnik–Chervonenkis
VCL	Variational Continual Learning
VEM	Variational Expectation Maximization
VI	Variational Inference