MACHINE LEARNING EVALUATION

As machine learning gains widespread adoption and integration in a variety of applications, including safety and mission-critical systems, the need for robust evaluation methods grows more urgent. This book compiles scattered information on the topic from research papers and blogs to provide a centralized resource that is accessible to students, practitioners, and researchers across the sciences. The book examines meaningful metrics for diverse types of learning paradigms and applications, unbiased estimation methods, rigorous statistical analysis, fair training sets, and meaningful explainability, all of which are essential for building robust and reliable machine learning products. In addition to standard classification, the book discusses unsupervised learning, regression, image segmentation, and anomaly detection. It also covers topics such as industry-strength evaluation, fairness, and responsible AI. Implementations using Python and scikit-learn are available on the book's website.

NATHALIE JAPKOWICZ is Professor and Chair of the Department of Computer Science at American University, Washington, DC. She previously taught at the University of Ottawa. Her current research focuses on lifelong anomaly detection and hate speech detection. She has also researched one-class learning and the class imbalance problem extensively. She has received numerous awards, including Test of Time and Distinguished Service awards.

ZOIS BOUKOUVALAS is Assistant Professor in the Department of Mathematics and Statistics at American University, Washington, DC. His research focuses on the development of interpretable multimodal machine learning algorithms, and he has been the lead principal investigator of several research grants. Through his research and teaching activities, he is creating environments that encourage and support the success of underrepresented students for entry into machine learning careers.

MACHINE LEARNING EVALUATION

Towards Reliable and Responsible AI

NATHALIE JAPKOWICZ
American University, Washington, DC

ZOIS BOUKOUVALAS
American University, Washington, DC

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, 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 Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781316518861

DOI: 10.1017/9781009003872

When citing this work, please include a reference to the DOI 10.1017/9781009003872

First published 2025

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

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-316-51886-1 Hardback

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will
remain, accurate or appropriate.

This book is dedicated to my mother, Suzanne. Maman, thank you for your encouragement and unshakable trust in my abilities. Although Norrin's books about international history and politics are more approachable than mine, I believe that you will, nonetheless, be happy to see it in print!

— Nathalie

To my beloved wife and colleague, Maria Barouti, and my dear parents, Panagiotis Boukouvalas and Niki Iliopoulou. Each of you has played a significant role in shaping my journey, and I dedicate this book to you as a token of my love and appreciation!

— Zois

Contents

Part I Preliminary Considerations

Part II Evaluation for Classification

Part III Evaluation for Other Settings

Cambridge University Press & Assessment
978-1-316-51886-1 — Machine Learning Evaluation: Towards Reliable and Responsible AI
Nathalie Japkowicz , Zois Boukouvalas
Frontmatter
More Information

Appendices

Abbreviations

10-fold CV	10-fold cross-validation
5 × 2 CV test	5 × 2 cross-validation test
5-fold CV	5-fold cross-validation
AI	artificial intelligence
AIC	Akaike information criterion
ANOVA	analysis of variance
AUC	area under the ROC curve
AUROC	area under the ROC curve
BERT	bidirectional encoder representations from transformers
BIC	Bayesian information criterion
BLEU	bilingual evaluation understudy
CD	critical difference
CDF	cumulative distribution function
CI	confidence interval
CNN	convolutional neural network
CV	cross-validation
DBI	Davis–Bouldin index
DBSCAN	density-based spatial clustering of applications with noise
DEA	data envelopment analysis
DET	detection error tradeoff
EM	excess mass criterion
EM algorithm	expectation maximization algorithm
EMR	exact match ratio
ERM	empirical risk minimization
FN	false negative
FP	false positive
FPR	false positive rate
FWER	family wise error rate

GAN	generative adversarial networks
HDI	high density interval
HL	Hamming loss
HS	Hamming score
i.i.d.	independently and identically distributed
ICA	independent component analysis
k-fold CV	k-fold cross-validation
LC	loss comparison
LDA	linear discriminant analysis
LIME	local interpretable model-agnostic explanation
LOO	leave-one-out
MAE	mean absolute error
MAP	maximum a posteriori
MCMC	Markov chain Monte Carlo
MDS	multi-Dimensional scaling
MSE	mean squared error
MV	mass volume criterion
NEC	normalized expected cost
NHST	null hypothesis significance testing
NLP	natural language processing
NMF	nonnegative matrix factorization
NPV	negative predictive value
PCA	principal component analysis
PDD	partial dependence plots
PDF	probability density function
Pdf	probability distribution function
PPCA	probabilistic principal component analysis
PPV	positive predictive value
PU	positive-unlabeled
RBF	radial basis function
RMSE	root mean squared error
ROC	receiver operating characteristic
ROPE	region of practical equivalence
ROUGE	recall-oriented understudy for gisting evaluation
SCAR	selected completely at random
SCM	set covering machine
SHAP	Shapley additive explanations
SIM	simple and intuitive measure
SOM	self-organizing maps
SRM	structural risk minimization

SSE	sum of squares error
SST	sum of squares totals
SVM	support vector machine
TN	true negative
TP	true positive
TPR	true positive rate
t-SNE	t-distributed stochastic neighbor embedding
VAE	variational autoencoders
XAI	explainable AI
XGBoost	extreme gradient boosting
XML	explainable machine learning

Preface

This book was born a few months after the beginning of the pandemic, at 9:45pm on June 2, 2020, to be exact, when Zois responded positively by email to Nathalie's invitation to participate in its creation. Nathalie had believed, for several years by then, that her 2011 book on evaluation, coauthored with Mohak Shah, was not capturing enough of the new developments in machine learning, and she was interested in reviving the project. Mohak's industry career, characterized by many moves and increasing sets of responsibilities, had made it impossible for Mohak to commit to the project. Nathalie's move from the University of Ottawa to American University, along with a new role as department chair, had made it difficult for Nathalie to envision working on a project as demanding as a new book without a partner. Nevertheless, with the popularization of machine learning through advances in deep learning and the creation of the data science discipline, it became clear that writing a comprehensive book on machine learning evaluation was becoming an urgent matter. Zois' positive email reply was the long-awaited catalyst! Being forced to stay home because of the pandemic reinforced our decision, and armed with a contract from Cambridge University Press, we set out to work. It took time – a lot longer than expected – but, despite new courses to teach and new administrative tasks to attend to, we made it happen and are pleased with the result!

Fear might have been an important motivator for this book. As more and more products embedding machine learning components are hitting the market, and as their applications are increasing in impact, we believe that it is urgent to inform society of ways to make these products safe and reliable. We view it as crucial for everyone involved in the design, application, and release of machine learning products to understand how to conduct machine learning evaluation, and maybe more importantly, to be aware of its uncertainty and the risks it comes with. If products like self-driving cars and automated hiring systems are to be deployed, everyone involved in the process, even the consumer, should know about the risks associated with their use.

The book plays a number of important roles. Foremost, it is meant to educate machine learning practitioners about the need for a thorough evaluation and the fact that evaluation is intimately linked to the application at hand. There are many evaluation tools appropriate for different applications, and each focuses on the specific goals the application is trying to achieve. Is the goal to diagnose patients as accurately as possible, like in a medical screening device? Is it to balance resources and needs in the fairest possible way, like in a public health program? Is it to minimize financial risk to make a business viable? Whatever the purpose of the evaluation, a particular set of techniques will be most appropriate to handle it, and the central purpose of the book is to expose the reader to these different tools so that, when the time comes, he or she will be able to select those most appropriate for the purpose of the application.

The second goal of the book is to sensitize the reader to the fact that neither machine learning nor its evaluation should be approached with certainty. Machine learning can produce unexpected results and evaluation, while able to roughly assess the worth of a machine learning-based system, can miss some very important limitations of that system. Nevertheless, neither the learning algorithms nor their evaluation are the results of random calculations, and by considering the fascinating mechanisms used by both learning algorithms and evaluation mechanisms, we hope to give readers the tools necessary to either build sufficient trust in an algorithm or know when to reject one as not being reliable enough for deployment.

On the practical side, the third goal of the book is to provide tools that make the implementation of the evaluation process easier. Some of these tools already exist, and in such cases, we point the reader to them and illustrate their uses. Others are homemade and, likewise, shared at https://github.com/zoisboukouvalas/MachineLearningEvaluation_TowardsReliableResponsibleAI.

Finally, the fourth role filled by our book concerns the future. This is the beginning of machine learning deployment. The evaluation of machine learning systems is bound to evolve and improve as more and more approaches get deployed. For the time being, the certainty with which evaluation tools are assessed depends on statistical principles rather than customer reactions to deployed products. We hope that with our book as a basis, future machine learning evaluators will be ready to move to the next level of evaluation to integrate other considerations such as safety, ethics, and legality.

Acknowledgments

There are many people who made the writing of this book possible and to whom we are immensely grateful. First and foremost, we would like to thank Mohak Shah for generously allowing us to reuse some of the content he provided in the 2011 book. Mohak also took time out of his busy schedule to give us advice on the organization of various chapters and of the overall book. His discerning eye and swift analytical skills led to an improved framing of our discussion and to the correction or clarification of certain passages. We are very honored to have had him be the first reader of the complete draft. We would also like to thank all our colleagues and students for their support during this undertaking and the valuable discussions we held with them. They include Michael Baron, Colin Bellinger, Paula Branco, Nicolas Cloutier, Roberto Corizzo, Evan Crothers, Kamil Faber, Kushankur Ghosh, William Klement, Bartek Krawczyk, Zhen Liu, Caitlin Moroney, Sunday Okechukwu, Sabrina Ripsman, Myles Russell, Liam Spoletini, Herna Viktor, and Bei Xiao.

Roberto Corizzo and Colin Bellinger helped us think through the best evaluation practices for the many problems we collaborated on. Both, with their deep understanding of the field, attention to detail, and creativity, guided us in refining our thoughts on specific aspects of evaluation. Evan Crothers helped us truly understand the ethical issues that creep into most methodological decisions as well as the societal dangers of the methods we deploy. Liam Spoletini and Sunday Okechukwu are owed a huge thank you for all the effort they put into the creation of the GitHub site that accompanies the book. Without it, the book would not be as useful. Nicolas Cloutier, a high-school intern, impressed us greatly and deserves a big thank you as well for unearthing the Cochran Q test for use in the multi-classifier single-domain case, a situation that had eluded us for many years! We also thank Herna Viktor, Roberto Corizzo, Kamil Faber, and Sabrina Ripsman for suggesting the addition of material that enhanced the quality of the book.

Many thanks to our editor at Cambridge University Press, Lauren Cowles, for her encouragements throughout the project and her excellent advice. We would also like to thank Arman Chowdhury, our editorial assistant, Clare Dennison, our senior content manager, and Jasintha Jacob Srinivasan, our project manager. We are extremely grateful, as well, to Eleanor Bolton, our copyeditor, for her careful reading of our manuscript.

Nathalie would also like to thank her family for putting up with her throughout the writing of the book. This often translated into last-minute meals hastily put together (and usually cooked in the microwave), less-than-ideal orderliness in the house, and little attention to homework or other matters. I will try to amend for my failings during the writing of the book (unless I just dive into another project and keep this up for another few years!). Thank you to my husband, Norrin, for his constant encouragement and advice; juggling the chairing of our respective departments at the same time and throughout the pandemic, along with managing our academic, family, and other pursuits has been an interesting proposition! I couldn't have done it without his love and support! Thank you to Shira, now a fellow scientist and great sounding board, and to Dafna for patiently enduring my obsession with computer work; as everybody knows, you are a great kid! Thank you as well to my mother, Suzanne, in Paris, who is always excited to hear about my projects. I am sorry that my father, Michel, who passed away 15 years ago, could not enjoy this moment. I am also sorry that my father-in-law, Michael, who passed away during the writing of the book, is not with us today as we would have enjoyed discussing its content together. Thank you to my mother-in-law, Toba, for her constant care and support.

Zois would like to express his heartfelt gratitude to his wife, Maria Barouti, who is not only a dedicated researcher but also a passionate machine learning educator. Throughout numerous hours of discussion and collaboration, she has generously shared her expertise, insights, and enthusiasm for the field of machine learning. Her guidance has been invaluable in helping me understand what is important in machine learning and how it should be presented to better benefit newcomers to this field. Maria, your visionary outlook for academia's future has given it a profound sense of relevance and purpose. Your commitment to educating others and empowering them with the knowledge that will shape the future of machine learning advancements leaves me in awe. Lastly, I would like to extend my deepest thanks to my beloved parents, Panagiotis and Niki, who have been my unwavering guiding stars, illuminating my path through every challenge and triumph. Your boundless love, selfless sacrifices, and unwavering encouragement have played a pivotal role in shaping the person I am today. I am forever indebted to the values and principles you instilled in me, forming the very foundation of my character.