

1

Introduction

Recent advances in deep learning have spearheaded impressive automated capabilities in visual and natural language processing and resulted in machine learning becoming a very popular discipline. While a decade or so ago only a handful of individuals – well read in matters of artificial intelligence (AI) and machine learning – perceived the inherent contributions that machine learning could bring to their field of expertise, today many people believe that machine learning will solve all of their problems: for better or for worse, machine learning has become ubiquitous. In such a climate, evaluating machine learning’s performance is of paramount importance, especially when large-scale deployments of tools based on its foundations are under consideration.

Performance evaluation in machine learning and, more generally, artificial intelligence has always been an important aspect of the field. Back in 1950, Turing famously designed a test he called the “imitation game,” now known as the “Turing test,” to evaluate the “intelligence” of the machine. In truth, any computer system must be subjected to testing and evaluation, and the software engineering discipline, born in the mid to late 1980s, created strict guidelines on how to do so. The issue with AI and machine learning, however, is that their testing and evaluation are not easy feats. While in traditional computer systems the range of outcomes is narrow and often quantifiable, in AI and machine learning, it is neither. The goal here is to get the machine to behave “intelligently.” Yet, what is intelligence? First, not every human being behaves the same way, and second, humans must function in unpredictable settings. These and many other considerations are what make the evaluation of AI and machine learning systems ill-defined. In fact, with the release of ChatGPT in 2022, the debate is on, once again, as it probably was when ELIZA, one of the first computer programs able to converse in English with people, was released in the 1960s. Though the book does not go into the more philosophical side of the problem, one thing is clear: in order to mimic humans adequately, the computer program needs to be able to *generalize* to unknown situations, and, as a result, it has to include nondeterministic “behavior.” It is this lack of determinism, or the system’s *inductive* behavior, that makes testing in the context of machine learning so complex, and this is one aspect of the problem into which our book takes a deep dive.

Another aspect relates to the fact that machine learning is now mature enough to be embedded in products designed to hit the market. Its evaluation is thus leaving the realm of philosophical questioning and becoming a practical matter. This shift from an inconse-

quential practice necessary to publish research papers, but not robust enough to truly assess the practical value of the resulting product, is a crucial one. Another essential function of this book is to bridge the gap between these two evaluation paradigms. It presents many different evaluation methods with the goal of covering the various situations that may arise when deploying a product: Is the product competitive? Robust? Safe? Fair? What is its expected benefit–cost ratio? While we cannot envision all the situations that could arise, the book gives the readers many tools that will allow them to determine what evaluation practice will serve their specific situation best. This is particularly transparent in the case of evaluation metrics discussed in Chapters 4, 5, 8, and 9, where different applications (e.g., medical, information retrieval, security) and learning paradigms (e.g., supervised, anomaly detection, time-series analysis, unsupervised) call for different types of measurements with overlaps occasionally occurring.

1.1 Motivation for This Book

The deep learning revolution is what motivated us most to write this book. Until deep learning started bearing fruit a few years ago, the technology was not mature enough to be considered for widespread use in sensitive applications such as self-driving cars, automatic screening of job applicants, and so on. As a result, we reasoned, the evaluation methods previously considered may not be robust enough for the types of applications machine learning is now encountering. This is not to say that machine learning tools were not evaluated and used in practical applications before, but instead, it is to emphasize that their uses are becoming more widespread now, and their domains of application more and more sensitive. This, we feel, requires further scrutiny. Moreover, while machine learning evaluation was a topic of conversation in the research community and within a circle of practitioners tightly linked to that research community, at this point, the need to understand how to evaluate machine learning algorithms adequately has spread to a much larger audience. It encompasses many scientific circles beyond the traditional machine learning community (e.g., chemists, biologists, physicists, environmental scientists, medical researchers), the business community with its various sectors, and even social studies and the arts.

The purpose of this book is to present a concise, yet complete, intuitive, yet formal, presentation of machine learning evaluation. The book has a predecessor, cowritten by one of the current coauthors, but although the new book builds upon the old one, it departs radically from it by increasing its coverage, updating its suggested methodologies, and, generally speaking, proposing a more robust approach. Not insignificantly, the new book also expands the reach of the discussion to the broader community.

An additional motivation for this book, which also ties together the reasons previously mentioned, is that although the situation has improved, evaluation of machine learning algorithms is often seen as an annoying and non-rewarding task. After all, creating new algorithms is a lot more exciting than testing them! More often than not, researchers or practitioners feel that they *need* to perform the task of evaluation to satisfy crusty conference or journal reviewers or bosses, but that the task is well below their skill levels. The view in this book, actually, is that, first, the evaluation of machine learning algorithms is

a fascinating field of study in and of itself, and, second, the result of this evaluation has become of extreme practical importance now that the world embraces machine learning and embeds it in its products.

Indeed, the field has, at last, matured to the point where the safety of the technology must be considered in ways similar to the way in which the safety of other technological or medical advances have had to be assessed for many years. While such evaluation might have been seen as overkill, and perhaps rightly so, in the earlier years of machine learning, now that self-driving cars use the technology as well as medical diagnostic or hiring systems, the issue cannot be ignored any longer, and approaches similar to phased medical trials or other industrial-strength evaluation need to be put in place. In fact, the time might have come for teams of evaluators, independent from the developers, to test AI products. Perhaps, even, the equivalent of the CDC or FDA for AI needs to be put in place to approve or prevent AI-based products from reaching the market in order to protect the consumer from physical or psychological harm and ensure the fairness of the product.¹ The purpose of this book is to lay out the various tools available to conduct a rigorous performance evaluation for a variety of machine learning paradigms including those used in supervised learning, unsupervised learning, image processing applications, large language models, and so on. The tools presented should also help evaluate different aspects of practical uses that may differ from one area of application to the next. To make the task of evaluation easier, the book refers the reader to existing evaluation tools, or provides new ones, where they are lacking, on what is currently the most prevalent machine learning development platform, *scikit-learn*.

There are a variety of reasons why we believe that the time has come to write this book. As mentioned previously, in the 13 years since the publication of its predecessor, there has been an explosion of people using machine learning tools, and trying to make sense of how to evaluate their performance. This new audience is not as homogenous as it was 13 years ago since, in addition to computer scientists, it includes many statisticians, data scientists, and practitioners of various disciplines ranging from the core sciences to the social sciences, including the medical sciences, education, journalism, and even arts disciplines such as literature, visual arts, and music.

In addition to the explosion and diversity of new users, there has also been a data explosion, bringing into focus different types of data (e.g., images, text), much larger amounts (sometimes in the order of terabytes), and a true desire to leverage the data toward robust industrial products. This last change makes the issue of evaluation essential and raises questions that were not considered with the same urgency in the past, such as privacy, bias, and explainability matters, as well as domain and task-dependent considerations. These, along with the model's correctness, need to be properly assessed before a product can be deployed or commercialized. By the same token, new machine learning tasks have emerged or, at least, become more prevalent. This includes computer vision tasks such as image

¹ A practical advantage of this suggestion, by the way, would be to free developers from having to perform as rigorous an evaluation of their ideas since these could be pitted professionally at a later stage. More to the point, however, such a division of labor would lead to a less biased evaluation process since the evaluators would have no stake in the products they are testing.

segmentation, unsupervised learning, and data stream analysis. Each of these new tasks requires new analysis tools.

In summary, the different types of data and their amounts, the new tasks that have emerged, as well as the stricter evaluation imperatives that are – or should be – in effect, may, in certain cases, require the use of evaluation tools different from those discussed previously, and ready access to these tools is imperative.

1.2 Contents and Organization of the Book

Machine learning evaluation typically refers to the evaluation of classification. The scope of this book extends beyond the task of classification, although classification remains a predominant aspect of machine learning and is covered thoroughly.

The presentation of the book is designed to appeal to both machine learning researchers and practitioners. Specifically, we provide both an informal discussion and access to simple-to-use tools with clear guidelines on when to use them (when possible), as well as more formal, theoretical explanations to support some of these practical considerations. To reflect the current trend in the field, all of the code provided is written in Python and uses the scikit-learn package (though some of the tools exist, in a previous version, in R).

The book is organized into four parts. Part I reviews essential statistical and machine learning concepts that are needed to provide context for the remainder of the book. This includes random variables, distributions, confidence intervals, and hypothesis testing on the statistical side; as well as the concepts of loss function, risk, empirical and structural risk minimization, regularization, the bias–variance tradeoff, clustering, dimensionality reduction, latent variable modeling, and generative learning on the machine learning side. In addition, Part I presents the de facto way in which machine learning evaluation is conducted, reviewing, along the way, well-known concepts such as the confusion matrix, micro- and macro- averaging, as well as well-known classification, regression, and clustering metrics; error estimation methods such as the holdout and k -fold cross-validation, and basic statistical tests such as the t -test and the sign test. Many of the evaluation methods covered in this part are well known to most practitioners of machine learning. In addition to reviewing them, we explain why they are, oftentimes, not sufficient. This prompts a discussion of why it is necessary to go beyond the material covered in the first part of the book, thus motivating the need for its further three parts.

Part II discusses machine learning evaluation in the important classification setting. In particular, it discusses the metrics that have been proposed for that setting, paying particular attention to the issues of class imbalances, costs, uncertainty, and calibration; the error-estimation/resampling approaches that have been proposed and their relationship to the bias and variance of the error estimates; and the different approaches to statistical analysis, including null hypothesis statistical testing, confidence intervals, effect size, and power analysis, as well as newer Bayesian analysis approaches that have recently been proposed in the machine learning literature.

Part III then turns to machine learning tasks other than classification. Evaluation methods for many tasks are presented, including those for classical paradigms such as regression anal-

ysis, time-series analysis, outlier detection, and reinforcement learning, and also newer tasks such as positive-unlabeled classification, ordinal classification, multi-labeled classification, image segmentation, text generation, data stream mining, and lifelong learning. A full chapter is then devoted to the important unsupervised learning paradigm, and evaluation methods for tasks including clustering and hierarchical clustering, dimensionality reduction, latent variable models, and generative models are discussed.

Finally, Part IV turns to practical considerations related to evaluation and deployment. First, machine learning evaluation is presented in a software engineering light whose goal is to herald the future of machine learning's use in industrial applications. Topics include data, algorithms, and platform imperfections, online testing, along with a description of current industry practice, and suggestions for improvements. The next chapter turns to the question of how to practice machine learning in a responsible manner. In particular, it dives into the issues of data and algorithmic bias, fairness, explainability, privacy, and security among others, and advocates the need for human-centered machine learning.

The book concludes with a discussion of how the performance evaluation components discussed throughout the book unify into an overall framework for in-laboratory evaluation. This is followed by a discussion of how to move from a laboratory setting to a deployment setting based on the material covered in Part IV of the book. Associated with this deployment, we emphasize the potential social consequences of machine learning technology, together with their causes, and suggest that these potential social effects should be considered as part of the evaluation framework.

The book comes accompanied by the Github site https://github.com/zoisboukouvalas/MachineLearningEvaluation_TowardsReliableResponsibleAI, which was written by American University graduate students Liam Spoletini and Sunday Okechukwu. The site provides Python code that illustrates how to implement the different evaluation methods discussed in the book. This, therefore, provides a quick way for machine learning designers and practitioners to apply evaluation techniques to their applications.