Synthesis Lectures on
Human Language Technologies

SYNTHESIS
COLLECTION OF TECHNOLOGY

Stefan Riezler · Michael Hagmann

Validity, Reliability, and Significance

Empirical Methods for NLP and Data Science

Second Edition

Springer

Synthesis Lectures on Human Language Technologies

The series publishes topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

Stefan Riezler · Michael Hagmann

Validity, Reliability, and Significance

Empirical Methods for NLP and Data Science

Second Edition

Springer

Stefan Riezler
Department of Computational Linguistics
and Interdisciplinary Center for Scientific
Computing
Heidelberg University
Heidelberg, Germany

Michael Hagmann
Department of Computational Linguistics
Heidelberg University
Heidelberg, Germany

ISSN 1947-4040 ISSN 1947-4059 (electronic)
Synthesis Lectures on Human Language Technologies
ISBN 978-3-031-57064-3 ISBN 978-3-031-57065-0 (eBook)
https://doi.org/10.1007/978-3-031-57065-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

To Sabine and Janna and Ida.

Preface to the Second Edition

The last few years have seen a paradigm shift in Natural Language Processing (NLP), caused by the rise of Large Language Models (LLMs). Due to the scalability of LLMs and their ability to encode vast amounts of knowledge, it suffices to prompt a pre-trained LLM with a few task-specific demonstrations (a.k.a. in-context examples) in order to tap into rich representations of concepts that are useful for a multitude of NLP tasks. However, common practices such as in-context learning with API-served black-box LLMs raise the question whether the methods to analyze validity, reliability, and significance of machine learning results that were introduced in the first edition of this book are still applicable in this new setup.

As has been shown in several recent studies, dataset bias and shortcut learning are problems for few-shot prompting of LLMs in a similar way as they are problematic for training or fine-tuning of any other machine learning model (Du et al., 2024; Kung & Peng, 2023; Tang et al., 2023; Zhao et al., 2021). Problems of (in)validity thus transfer to the era of LLMs, however, with the additional complexity of interactions between spurious features in prompts and in training examples (Webson & Pavlick, 2022). Fortunately, validity tests like the GAM-based circularity test are in principle applicable to predictions of any black-box model trained on non-public data, thus most methods for validity testing introduced Chap. 2 can be saved into the new age of machine learning without changes.

The main addition to this second edition is Chap. 5 where we show how to apply the LMEM-based methods for reliability testing, variance component analysis (Chap. 3) and significance testing (Chap. 4) to analyze the *inferential reproducibility* (Goodman et al., 2016) of research results. Such an analysis regards various sources of nondeterminism as inherent and sometimes irreducible conditions of measurement that contribute to variance in performance evaluation in an interesting way. The focus is then on incorporating several such sources of variance, including their interaction with data properties, into an analysis of the significance and reliability of machine learning evaluation. Since the goal of an analysis of inferential reproducibility is to draw inferences beyond particular instances of trained models, it relies on explicit sources of variability in model training. For in-context learning of LLMs, sources of randomness include the number, ordering, and similarity metric of the in-context examples. For training or fine-tuning of LLMs, typical sources of variability are meta-parameter settings or data characteristics. We exemplify inferential reproducibility using an algorithm for regularized fine-tuning of pre-trained LLMs (Aghajanyan et al., 2021) under various algorithm-level and data-level factors of non-determinism. Furthermore, we present the reproducibility study in a way that ensures it is replicable itself, by links to code and data, and by a walk-through of our open-source code.

In addition to introducing a new chapter in this edition, inconsistencies in notation and some typographical errors have been fixed. We refrained from overloading the book with new experimental results and instead refer the reader to the conference publication of Hagmann et al. (2023a) that is related to Chap. 5, and to an extension of Sect. 2.4.3 to more complex examples from medical data science in the conference paper of Hagmann et al. (2023b).

We would like to thank several people who contributed greatly to this second edition. First of all, we are grateful to Graeme Hirst and Susanne Filler for clearing the path to this edition so efficiently from technical and administrative obstacles. Furthermore, we would like to thank Philipp Meier and Shigehiko Schamoni, our co-authors on the respective above-mentioned publications accompanying this book, for allowing us to incorporate parts of our joint work into this edition.

Heidelberg, Germany
February 2024

Stefan Riezler
Michael Hagmann

References

Aghajanyan, A., Shrivastava, A., Gupta, A., Goyal, N., Zettlemoyer, L., & Gupta, S. (2021). Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations* (ICLR). Available from: https://openreview.net/forum?id=OQ08SN70M1V

Du, M., He, F., Zou, N., Tao, D., and Hu, X. (2024). Shortcut learning of large language models in natural language understanding. In Communications of the ACM, 67(1):110−120. Available from: https://doi.org/10.1145/3596490

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Sci Transl Med*, 8(341):1–6. Available from: https://doi.org/10.1126/scitranslmed.aaf5027

Hagmann, M.,Meier, P., and Riezler, S. (2023a). Towards inferential reproducibility of machine learning research. In *The Eleventh International Conference on Learning Representations* (ICLR). Available from: https://openreview.net/forum?id=li4GQCQWkv

Hagmann, M., Schamoni, S., and Riezler, S. (2023b). Validity problems in clinical machine learning by indirect data labeling using consensus definitions. In *Extended Abstract presented at Machine Learning for Health (ML4H) symposium*, New Orleans, United States. Available from: https://doi.org/10.48550/arXiv.2311.03037

Kung, P.-N. & Peng, N. (2023). Do models really learn to follow instructions? An empirical study of instruction tuning. In *Proceedings of the 61st AnnualMeeting of the Association for Computational Linguistics* (ACL), Toronto, Canada. Available from: http://dx.doi.org/10.18653/v1/2023.acl-short.113

Tang, R., Kong, D., Huang, L., & Xue, H. (2023). Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics* (ACL), Toronto, Canada. Available from: http://dx.doi.org/10.18653/v1/2023.findings-acl.284

Webson, A. & Pavlick, E. (2022). Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL:HLT), Seattle, WA, USA. Available from: https://aclanthology.org/2022.naacl-main.167

Zhao, Z.,Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning* (ICML), virtual. Available from: https://proceedings.mlr.press/v139/zhao21c.html

Preface to the First Edition

There is a particular book that accompanied the first author since his days as a doctoral student: Paul R. Cohen's textbook *Empirical Methods for Artificial Intelligence* (Cohen, 1995). The book was introduced to him by Mark Johnson, with the recommendation that it contained essential information for an empirical researcher that is not easily available in a comparably concise form anywhere else. This assessment of Cohen's book is still valid today.

Myriad books on machine learning, deep learning, and artificial intelligence have been published since Cohen's book appeared in 1995. With rare exceptions such as Hardt and Recht (2022), however, questions about data practices, the concepts of validity and reliability, or techniques of exploratory data analysis are not mentioned in contemporary books on machine learning. A discussion of confirmatory techniques for statistical hypothesis testing and their relevance for practical machine learning research is also not integrated in most machine learning textbooks. For these topics, Cohen's exposition of exploratory and confirmatory techniques of empirical science is still the to-go textbook. However, Cohen's book has not been updated since its publication date.

The goal of our book is to extend and update Cohen's book using model-based techniques to address the questions of validity, reliability, and significance in empirical machine learning research. In our book, these techniques are based on interpretable probabilistic models as described in Wood (2017). These models are not necessarily more recent than Cohen's book, but they possess the necessary expressiveness to model experimental data from data annotation and machine learning prediction experiments, and they are associated with proven statistical properties for drawing inferences about the parameters and models. The goal of our book is to provide the reader with an instrument in the form of model-based statistical tests that enable assessing the methodological questions of validity, reliability, and significance. We showcase our techniques on examples from the authors' areas of expertise—NLP and medical data science—and hope that the proposed techniques will also be of use to readers from other areas of machine learning and artificial intelligence.

Heidelberg, Germany
November 2021

Stefan Riezler
Michael Hagmann

References

Cohen, P. R. (1995). *Empirical methods for artificial intelligence*. The MIT Press. Available from: https://mitpress.mit.edu/9780262534178/empirical-methods-for-artificial-intelligence/

Hardt, M. and Recht, B. (2022). Patterns, predictions, and actions: A story about machine learning. Princeton University Press. Available from: https://mlstory.org

Wood, S. N. (2017). Generalized Additive Models. An Introduction with R. Chapman & Hall/CRC, second edition. Available from: https://doi.org/10.1201/9781315370279

Acknowledgments

This book would not have been possible without the help of several people who are actively involved in empirical machine learning research, and who were willing to read early drafts of our book and comment on its relevance to their own work.

Firstly, we would like to thank the students at the departments of computational linguistics and computer science who participated in two iterations of a seminar class on the topics of the book, and who detected and corrected many mistakes in earlier versions of the book. We are indebted to Nathan Berger for proofreading our writing as a native speaker of American English and for comments on the intelligibility of the contents as a Ph.D. student in computer science. We would like to thank Michael Staniek for his critical comments on the coherence of our argumentation and the usefulness of the presented methods for a Ph.D. student in computational linguistics. We are indebted to Mayumi Ohta for testing our R scripts and for contributing to the experimental material that illustrates our statistical tests.

We thank Artem Sokolov for patiently going over several early versions of the book, for a great many discussions on all aspects of our work, and for serving as an endless source of recommendations on related work.

Last, we would like to thank Graeme Hirst and Michael Morgan for giving us the chance to publish this book in the first place and for selecting two excellent reviewers. We thank the anonymous reviewers for providing feedback on all levels of detail of the book and for giving invaluable guidance on how to present our material in a clear and appealing form.

Clearly, various errors and shortcomings remain, and we would be grateful if readers could point them out to us so that they can be corrected.

November 2021

Stefan Riezler
Michael Hagmann

Contents

About the Authors

Stefan Riezler is a full professor in the Department of Computational Linguistics at Heidelberg University, Germany since 2010, and also co-opted in Informatics at the Department of Mathematics and Computer Science. He received his Ph.D. (with distinction) in Computational Linguistics from the University of Tübingen in 1998, conducted post-doctoral work at Brown University in 1999, and spent a decade in industry research (Xerox PARC, Google Research). His research focus is on interactive machine learning for natural language processing problems, especially for the application areas of cross-lingual information retrieval and statistical machine translation. He is engaged as an editorial board member of the main journals of the field—*Computational Linguistics* and *Transactions of the Association for Computational Linguistics*—and is a regular member of the program committee of various natural language processing and machine learning conferences. He has published more than 100 journal and conference papers in these areas. He also conducts interdisciplinary research as a member of the Interdisciplinary Center for Scientific Computing (IWR), for example, on the topic of early prediction of sepsis using machine learning and natural language processing techniques.

Michael Hagmann is a graduate research assistant in the Department of Computational Linguistics at Heidelberg University, Germany, since 2019. He received an M.Sc. in Statistics (with distinction) from the University of Vienna, Austria in 2016, and a Ph.D. in Computational Linguistics from Heidelberg University in 2023. He received an award for the best Master's thesis in Applied Statistics from the Austrian Statistical Society. He has worked as a medical statistician at the medical faculty of Heidelberg University in Mannheim, Germany, and in the section for Medical Statistics at the Medical University of Vienna, Austria. His research focus is on statistical methods for data science and, recently, NLP. He has published more than 50 papers in journals for medical research and mathematical statistics.

Introduction 1

1.1 Empirical Methods in Machine Learning

Machine learning is a research field that has been explored for several decades, and recently has begun to affect many areas of modern life under the reinvigorated label of artificial intelligence. The goal of machine learning can be described as learning a mathematical function to make predictions on unseen test data, based on given training data, without explicit programmed instructions on how to perform the task. The methods employed for learning functional relationships between inputs and outputs heavily build on methods of mathematical optimization (Bottou et al., 2018). While optimization problems are formalized as minimization of empirical risk functions on given training data, the important twist in machine learning is that it aims to optimize prediction performance in expectation, thus enabling generalization to unseen test data. The development and analysis of techniques for generalization is the topic of the dedicated sub-field of statistical learning theory (Bousquet et al., 2004; Vapnik, 1998; von Luxburg & Schölkopf, 2011). Statistical learning theory can be seen as the methodological basis of machine learning, and central concepts of statistical learning theory have been compared to Popper's ideas of falsifiability of a scientific theory (Corfield et al., 2009). In a similar spirit, comparisons of the methodology of machine learning and empirical science have led to direct advertisements of "Machine Learning as Philosophy of Science" (Korb, 2004).

Let us contrast this proposition with the practical workflow of a machine learning researcher conducting empirical research in natural language processing (NLP) and data science. Most empirical research in these areas follows the paradigm of adopting or establishing a set of input representations and output labels that are split into portions for training, development, and testing. The data in these splits are assumed to represent independent samples from an identical distribution (so-called i.i.d. samples). Furthermore, data in the splits are made i.i.d. artificially, e.g., by shuffling data at random between splits (Arjovsky et al., 2019) or by experience replay (Schölkopf, 2022). The i.i.d. assumption is crucial for the

S. Riezler and M. Hagmann, *Validity, Reliability, and Significance*, Synthesis Lectures
on Human Language Technologies, https://doi.org/10.1007/978-3-031-57065-0_1

consistency guarantees from statistical learning theory to apply (Vapnik, 1998; von Luxburg & Schölkopf, 2011). Furthermore, it can be seen an acknowledgment of basic principles of experimental control by a randomized experimental design (Cox & Reid, 2000; Mead et al., 2012). A typical NLP or data science project then starts with optimizing the parameters of a machine learning model on given training data, tuning meta-parameters on development data, and ends with testing the model using a standard automatic evaluation metric on benchmark test data. We call this scheme of a machine learning process the **train-dev-test paradigm** of NLP and data science.[1]

The train-dev-test paradigm allows the researcher to happily focus on improving model performance, with the only limit being the computational budget to train and re-train complex models, such as deep neural networks, under extensive exploration of meta-parameters, but without having to ask any questions about the data themselves, about what the machine learning model learned from them, or how the learning process is influenced by diverse sources of variability. Such questions are typically thought of as extraneous to the machine learning process, and standard statistical learning theory does not provide answers to them. However, as we will show in this book, processes like data annotation or model evaluation that happen before or after machine learning crucially influence the entire machine learning process. The viewpoint advocated in this book is that answers to questions about bias and consistency in data annotation, about representations of raw input data, or about variability of machine learning models with respect to meta-parameters and test data, should be an integral part of the methodology of machine learning. The current discussion of methodological issues in empirical machine learning is at the state of informal guidance by Dos and Don'ts (Bowman & Dahl, 2021; Lones, 2021). The goal of this book is to analyze problems in the train-dev-test paradigm from the viewpoint of the methodology of empirical sciences—a point of view that is independent of and orthogonal to statistical learning theory[2]—and to answer them by concrete statistical techniques.

The methodological questions that will be addressed in this book include the question of **validity**—does a machine learning model predict what it purports to predict? For example, we might want to scrutinize surprisingly good results on hard tasks like natural language understanding, and ask whether successful machine learning models do understand language or instead rely on superficial patterns that are highly, but spuriously, correlated with target classes (Clark et al., 2019). Similarly, observed superior performance in data mining might be due to illegitimate leakage of information correlated with the target (Kaufmann et al., 2011), and exact prediction of the target in medical informatics might be based on using

[1] Clearly, this paradigm is pervasive in machine learning and artificial intelligence in general, for example, in the area of image processing that uses similar methods and exhibits similar problems as the area of natural language processing. We will frequently refer to examples from related areas, but keep our focus on running examples from the areas of NLP and medical data science.

[2] The orthogonality of our methodological point of view to statistical learning theory is shown by the fact that it applies to classical learning theory as well as to more recent approaches (Arjovsky et al., 2019; Kawaguchi et al., 2022; Shen et al., 2021).