From Data to Decisions: Placing Machine Learning Challenges In Context

David J. Stracuzzi **Maximillian G. Chen** **Michael C. Darling** **Matthew G. Peterson**
Sandia National Laboratories
Albuquerque, NM 87123
{djstrac | mgchen | mcdarli | mgpeter}@sandia.gov

1 Background

We argue that machine learning challenges can foster advances in the art and practice of data analysis by designing contests that include the context of decision-making applications. In particular, uncertainty plays an important role in the decision making that accompanies data analysis. Challenge contests advance both the state of the art and the state of practice by requiring participants to apply modeling and prediction techniques to real world data. The ensuing difficulties in dealing with noisy and messy data, particularly data that may not adequately address the question of interest, requires participants to adapt and generalize algorithms in new and exciting ways. However, machine learning challenges in their current form neglect the larger decision-making context in which machine learning algorithms are often used. In the following, we present examples in which we attempt to situate a realistic machine learning problem into a decision-making context.

Consider a smartphone's auto-replacement spell-checker. Spell checkers frequently provide mediocre corrections and completions that can turn simple typos into nonsensical statements, or worse, alter the author's intent. Such issues stem both from limited language modeling resources and from *a neglect of decision-making context*. For a given misspelling and context, the number of possible corrections can be large and difficult to rank. This corresponds to high uncertainty — many similarly plausible solutions — and automatically selecting a replacement amounts to random guessing among alternatives. The available information is not sufficient to indicate one word over another, yet from the author's perspective, one word is not equivalent to another. By selecting a replacement essentially at random, the algorithm neglects its context of improving communication.

2 Decision Making Demands Consideration of Uncertainty

Broadly speaking, machine learning considers problems in which data describe a relationship of interest, such as misspellings to correct spellings, but the relationship is either too complex to code manually or incompletely understood (Mitchell, 1997). Such problems tend to be ill-defined (Hadamard, 1923), meaning that the resulting model parameterizations may not be unique and may be highly sensitive to the input data. When we consider these problems in a decision context, ranging from determining plagiarism for course grading to climate modeling for policy decisions, we raise critical questions that extend beyond classic model selection, parameterization, and evaluation.

- How sure are we that the model's response is correct for the given input?
- What other responses are also plausible?
- What are the sources of variability and what can be done to reduce their impact?

Decisions often carry consequences, so decision making requires that we address questions about the veracity of our data, models, and output. For example, Friedman and Zeckhauser (2012) show that efforts to reduce or eliminate uncertainty in the intelligence community can lead to consequence neglect, or undue focus on the probability of a scenario occurring while ignoring its consequences.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Traditional performance evaluation metrics, such as confusion matrices, ROC curves, and F-scores all use some version of true and false positive and negative rates to quantify a classifier's performance. These methods provide a global measure of a classifier's ability to discriminate among examples of different classes. However, none of these methods quantifies the variability in a classifier's output. Addressing variability requires identifying potential sources of uncertainty and explicitly evaluating their effects.

Example 1: Authorship and Plagiarism One recent KDD Cup competition focused on authorship determination (KDD, 2013), in which the task is to identify authorship in the context of individual authors publishing under multiple variations of their name and multiple authors with similar names. Relevant applications include automated disambiguation systems, which help to organize publications for the purposes of search and author rating. Both cases entail important implications for errors: if an author's work is misassigned, then searches may not turn up pieces of related work, the author may not receive credit for the work, and their rating or reputation may not reflect the full body of their research. Similar issues arise in plagiarism analysis, which has also been the source of several competitions (see PAN, 2015, for example). Simple cases of copy-and-paste aside, the assignment of origination can have substantial impact on both reputation and financial success.

Critical questions in both cases are not limited to *what's the performance of the analysis system on training or validation data?* To make a justifiable decision, we want to know how well the validated model speaks to the specific text in question. A decision maker, such as a course instructor in the plagiarism case, needs to know what the analysis determined *and* the degree of uncertainty in the result. Is it a clear case of plagiarism? Or a borderline case? Would varying word choice in a few places make a difference? Designers of machine learning systems need to include the analyses required to answer such questions in the system's design, which represents a substantial departure from the current state of practice in many cases. Inclusion of the matched passage mitigates concerns about plagiarism detection, but the larger point that machine learning algorithms can and should provide more information about their analyses still holds.

Example 2: Malicious URL Classification The cyber realm is particularly vulnerable to issues of nuance. For example, in URL classification, the goal is to determine whether the target web site is malicious given only the link text and structure (Darling et al., 2015). Malicious link content evolves over time as adversaries attempt to invade new systems, so a detection system that was 99% accurate yesterday may be less reliable today. The general task is therefore to alert users to potential threats, but the expectation of concept drift means that the system needs to recognize conditions under which threat detection errors are more likely. In practice, the machine learning system needs to evaluate either the degree to which a new example resembles the original training data, or the uncertainty in the resulting classification. Large differences from the training data or high uncertainty would indicate an unreliable classification, though not necessarily an incorrect one. The user may then need to take some action, but this is preferable to exposing a computer or network to exploitation. As with the previous examples, the performance criteria depends on both the model's ability to evaluate examples and the model's ability to determine the quality of its own output.

3 Conclusion

Machine learning algorithms are one part of a larger pipeline, and future challenge problems should include the post-analysis decision making. In practice, this means that evaluation metrics must relate to *application goals* instead of focusing on traditional machine learning performance criteria. In the spell-checking example, this may reward not making a correction if the the system cannot disambiguate among alternatives. For authorship and cyber security, this may mean providing uncertainty information above and beyond yes/no classifications or even point-estimate probabilities. The criteria by which machine learning research is judged often differs from the criteria by which applications are judged. As a simple illustration, compare the differences between the review criteria for the research-focused AAAI (2018) conference and the application-centric Innovative Applications of AI (IAAI-18, 2018) conference. Machine learning competitions are well-positioned to establish more nuanced and application-driven evaluation measures in the community.

Acknowledgments

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

AAAI-18 (2018). AAAI-18 call for papers. `http://www.aaai.org/Conferences/AAAI/2018/aaai18call.php`. Accessed: 2017-10-03.

Darling, M., Heileman, G., Gressel, G., Ashok, A., and Poornachandran, P. (2015). A lexical approach for classifying malicious URLs. In *Proceedings of the 2015 International Conference on High Performance Computing and Simulation*, pages 195–202, Amsterdam, Netherlands. IEEE Press.

Friedman, J. A. and Zeckhauser, R. (2012). Assessing uncertainty in intelligence. *Intelligence and National Security*, 27(6):824–847.

Hadamard, J. (1923). *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Yale University Press, New Haven, CT.

IAAI-18 (2018). IAAI-18 call for papers. `http://www.aaai.org/Conferences/IAAI/2018/iaai18call.php`. Accessed: 2017-10-03.

KDD (2013). Determine whether an author has written a given paper. `http://www.kdd.org/kdd-cup/view/kdd-cup-2013-track-1`. Accessed: 2017-09-27.

Mitchell, T. M. (1997). *Machine Learning*. WCB/McGraw-Hill, Boston, MA.

PAN (2015). Plagiarism detection / text reuse detection. `http://pan.webis.de/clef15/pan15-web/plagiarism-detection.html`. Accessed: 2017-09-27.