Towards international standards for evaluating machine learning

Frank Rudzicz[1,2,3,4] and **P Alison Paprica**[4,5] and **Marta Janczarski**[6]
[1] Li Ka Shing Knowledge Institute, St Michael's Hospital
[2] Department of Computer Science, University of Toronto
[3] Surgical Safety Technologies Inc.
[4] Vector Institute for Artificial Intelligence
[5] Institute for Health Policy, Management and Evaluation, University of Toronto [6] Standards Council of Canada

Abstract

Various international efforts to standardize artificial intelligence have begun, and many of these efforts involve issues related to privacy, trustworthiness, safety, and public well-being, which are topics that don't necessarily have international consensus, and may not for the foreseeable future. Meanwhile, the pursuit of achieving state-of-the-art accuracy in machine learning has resulted in a somewhat *ad hoc* application of empirical methodology that may limit the correctness of the computation of those accuracies, resulting in unpredictable applicability of those models. Trusting the objective quantitative performance of our systems is itself a safety concern and should inform the earliest standards towards safety in AI.

Introduction

Implementing international standards is a primary method to ensure the safety of a process or product. Outlining a set of specifications enables consumers and producers to both abide by the requirements set out to reduce risk and harm. There are two principal avenues to employ these safety standards. The first is to develop a *conformity assessment scheme* against a set standard; this provides a certification that can clearly and quickly determine whether a product or process meets agreed-upon requirements. The process of certification often includes both testing the product or process, and analyzing the documentation throughout the design and creation process. The second avenue is *incorporation by reference of standards in regulation*. Regional and federal regulations incorporate national and international standards to ensure uniformity and a level of product requirements that are expected. Prominent examples include building codes, and the Consumer Product Safety Act[1]. For example, the International Standards Organization (ISO) committee on consumer policy (COPOLCO) provides a potential platform to investigate the consumer impact of AI, in particular through aspects of safety (ISO/IEC 2014).

Developing foundational machine learning (ML) has, to some extent, been *ad hoc*, susceptible to trends, and relatively undirected. There is no inherent moral or ethical problem with this approach except for the *potential* to consequently evaluate our systems in a cursory or incorrect way. Without careful consideration to empirical methodology, the truth and generalizability of our own statements may be suspect. Without any conceptual barriers to overcome, success in innocuous 'toy' tasks, such as distinguishing images of cats from images of dogs (Elson et al. 2007), can quickly transfer to other tasks with superficially similar data types, such as distinguishing images of malignant skin lesions from benign ones (Esteva et al. 2017). That is, ML allows for modelling procedures to be easily transferred across data sets without necessarily considering possible covariates hidden in those data sets, nor the potential consequences of false positives or false negatives. Indeed, not accounting for skin colour in those images may aversely affect underrepresented populations in the data (Adamson and Smith 2018; Lashbrook 2018). The relative ease with which modern ML can be implemented may reveal unintended biases of this type or, more concerning, biases that we cannot understand.

The dawn of AI standards

In addition to various regional efforts, two international organizations are leading the global standardization effort in AI. The IEEE has launched the Global Initiative on Ethics of Autonomous and Intelligent Systems to address some of the societal concerns that are emerging with AI. These include areas such as data governance and privacy, algorithmic bias, transparency, ethically driven robots and autonomous systems, failsafe design and wellbeing metrics. Also, the ISO has recently created a new technical subcommittee (SC) in the area of artificial intelligence, ISO/JTC 1 SC 42, whose scope covers foundational standards as well as issues related to safety and trustworthiness. At their first plenary in April 2018, the subcommittee created the following study groups:

Computational approaches and characteristics This concerns different technologies (e.g., ML algorithms, reasoning) used by AI systems including their properties and characteristics. This will include specialized AI systems (e.g., NLP or computer vision) to understand and identify their underlying computational approaches, architectures, and characteristics, and industry practices, processes and methods for the application of AI systems.

Trustworthiness This concerns approaches to establish trust in AI systems, e.g., through transparency, verifia-

[1] https://laws-lois.justice.gc.ca/eng/acts/C-1.68/

bility, explainability, controllability. Engineering pitfalls, typical threats and risks, their mitigation techniques, and approaches to robustness, accuracy, privacy, and safety will also be investigated.

Use cases and applications This focuses on application domains for AI (e.g., social networks and embedded systems) and the different context of their use (e.g., health care, smart homes, and autonomous cars).

Work items related to the specification of performance of ML models, and comparison between models, are relevant to each of these study groups.

Standards for evaluating ML models

Machine learning research is driven, to a large extent, by the search for mechanisms that achieve greater accuracy than competing approaches. Unfortunately, there have been several methodological limitations or misapplications that have hindered the comparison of models, or made such comparisons forfeit. Indeed, references are routinely made in the literature to 'state-of-the-art' performance, sometimes involving minuscule differences on small data sets, and in relatively esoteric tasks. Broad acceptance of such empirical procedures, or their assumed generalizability, makes the supposed direct comparison between approaches tenuous at best, and suspicious at worse.

When comparing the performance of two or more algorithms, the following aspects must be carefully controlled and reported:

Implementation For example, if an algorithm can be accelerated (e.g., by using GPU processing) in such a way that can affect outcomes (e.g., if a there is a stopping condition in time), then this must be made explicit.

Hyper-parameters If the hyper-parameters of a ML model are optimized, the hyper-parameters of the comparative ML models should also be optimized, except when hyper-parameters themselves are being compared.

Preprocessing Preprocessing steps will not unjustly favour one model over another. For example, if a classifier requires 'stop words' (e.g., prepositions) to be retained in an NLP task, those words should not be removed. Moreover, preprocessing should be consistent across all data. For example, if outliers, incomplete data, or noise are removed, it must be done uniformly.

Training and testing data When several machine learning models are being compared, the data used to train those models or, separately, evaluate those models, should be identical. These data should be ecologically valid, statistically indistinct, or otherwise similar to data expected to be observed in deployment.

Representative data The data should be as free of sampling bias as possible. That is, the distribution of classes in the data should be identical to their distribution in the real world, to the extent possible. There may be special considerations to this point. To some extent, models trained on historical data may encapsulate biases from the past that the developers wish removed, such as demographic biases towards recidivism or gender biases in word embeddings (Bolukbasi et al. 2016); if the system is meant to enable decision support prospectively, techniques to mitigate bias should be taken.

Appropriate baselines Any classifier of interest should be compared against at least one representative, appropriate baseline. Trivial baselines should not be considered. A trivial baseline, for example, always predicts the majority class and, in general, is not the result of a machine learning process.

Appropriate measures There is a tendency to report accuracy or area under the precision-recall (or other operator characteristic) curves in nominal classification; however, this is not always correct. For example, systems that predict cause-of-death according to international standards for disease coding should not merely report accuracy, but should include the cause-specific mortality fraction (CSMF), which is the fraction of in-hospital deaths for a given cause normalized over all causes (Murray et al. 2007). CSMF accuracy is therefore a measure of predictive quality at the population level, which quantifies how closely the estimated CSMF values approximate the truth. In fact, when it is possible to compute its associated coefficients, the chance-corrected version of CSMF accuracy should be used instead (Flaxman et al. 2015). Clearly, the measure used, itself, can be highly context-dependent and result in very different outcomes for different samples.

Limiting information leakage It is necessary to partition data between training sets and test sets in such a way so that no latent information exists across sets, other than directly obtained from observation variables. This can occur when latent information is highly correlated to labels, annotation, or other supervised information.

For example, a system may be designed to classify between people with and without neurodegeneration from audio (Fraser, Meltzer, and Rudzicz 2015) and have multiple data points recorded from each human subject in the data. Some acoustic features, such as vocal jitter or phonation rates, may be used to identify pathology cross-sectionally, but they can also be used to identify the speaker themselves. Since each speaker is associated with a label for the outcome, even if individual samples are partitioned across training and test sets, it would be inappropriate if an individual speaker is represented in both sets. This is because any model could learn the identity of a speaker from the training data, and apply the known label to test data, tainting the results. Leave-one-out cross-validation is one mitigation strategy.

Limiting channel effects A channel effect occurs when a classifier may learn characteristics of the *manner* in which data were recorded, in addition to the *nature* of the data themselves. For example, a hospital-based system may be designed to classify among patient data. However, if all or most patients with complex cancers seek treatment in urban centres, then a classifier may learn to associate those cancers with certain regions.

Channel effects can be caused by the mechanism used

to obtain the data, any preprocessing that occurred on one or more proper subsets of the data, the identity of the individual or individuals obtaining the data, or environmental changes in which data were recorded, for example. If these effects cannot be controlled, they must be accounted for as covariates during statistical significance testing. Additionally, strategies have been developed to explicitly factor out channel effects, as with i-Vectors through expectation-maximization, probabilistic linear discriminant analysis, and factor analysis (Verma and Das 2015).

Furthermore, appropriate statistical tests of significance must be undertaken, when possible, in order to establish whether there is any *meaningful* difference between approaches. A difference of 0.5%, for example, on a single test set is not necessarily conclusive with regards to the models compared. Naturally, tests of significance can also be misused (i.e., so-called 'p-hacking'), so effort must be taken to choose appropriate tests. For example, if a test has an assumed distribution (as in standard t-tests), then the validity of that assumption in the data should also be evaluated (e.g., through a Lilliefors or Kolmogorov-Smirnov test). If multiple comparisons are made (e.g., through multiple hyperparameterizations), then this must be accounted for also, e.g., through a Bonferroni test. Alternatively, standardized computations of effect sizes (e.g., Cohen's d) can mitigate against the risks of p-hacking. Finally, where possible, all relevant covariates must be accounted for in the model; as mentioned above, this includes all aspects in the channel, including confounding variables in the data themselves, that could effect the outcomes, as well as the interactions between those variables.

The assessment of nominal classification or continuous regression has been implicit in the discussion above, but the same principles may be taken in evaluating reinforcement learning, for example.

Recommendations

Ensuring the safety and maintenance of AI systems will be the subject of various standardization efforts, including explainable models, unintended biases (including cultural, social, historical, or sampling biases), human-machine interaction, and scalable oversight. However, the clear-eyed refinement of the actual evaluation methodologies will be crucial to many of these challenges. When so many independent researchers and organizations (across academic, commercial, or governmental sectors) are actively and competitively engaged to achieve state-of-the-art performance, it is essential to be able to objectively and quantitatively establish that performance correctly, consistently, and with expected minimal levels of reporting, otherwise any claims should *not* be trusted.

Endorsement versus enforcement

As international standards are optional, a core concern is whether these standards will be enforceable by governments, given a variety of attitudes towards artificial intelligence by the governments represented on standardization bodies. Perhaps more concerning is whether standards in AI will even be acknowledged or endorsed by developers and practitioners, especially in academia. Agreeing upon a minimal set of standards for evaluation, across sectors, may require a broad cultural change within the ML community. Indeed, Dror et al. (2018) showed through a meta-analysis that, while the ML and natural language processing communities are driven by experimental results, statistical significance testing is ignored or misused *most* of the time. Therefore, as machine learning becomes an increasingly applied science, empirical methods should be emphasized early in relevant University and educational programs, including Computer Science.

Naturally, innovation should continue to be encouraged. In fact, lowering certain regulatory barriers may promote certain safe uses of ML in the service of the public well-being, including healthcare. The point is not that risk should be averted – there is substantial evidence that undertaking *certain kinds* of risk can be beneficial *when those risks are understood*. The challenge we face is that we do not truly understand the risks of AI and ML – we barely understand how to assess those systems in the first place, which should be among our first priorities.

With some exceptions, such as cases where human rights are at risk, we do not propose that standards in AI should inhibit our scientific exploration in any way. Nor do we propose limits to the capabilities of AI systems – those will largely be dependent on national or regional laws or regulations. Rather, the international standardization community has the opportunity to place certain expectations as to how we, the ML community, evaluate our own work. Trusting the objective quantitative performance of our systems is itself a safety concern.

Final comments

This paper represents the beginning of a long, multi-year process in surveying challenges or shortcomings in the evaluation of performance of machine learning, especially as it relates to the trustworthiness of that performance. These are not a complete set of requirements, nor are the recommendations fully expressed. Training software, rather than explicitly programming it, poses unique challenges and imposes a specific development and test life cycle suitable that is distinct from traditional software development and, crucially, regulation and standardization. Accurately controlling for the behaviours of deployed machine learning, through a quantitative evaluation of its performance, is crucial.

Acknowledgements

Rudzicz is the Canadian Chair of the mirror committee to ISO/JTC 1 SC 42 on Artificial Intelligence.

References

Adamson, A. S., and Smith, A. 2018. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology* 154(11):1247–1248.

Bolukbasi, T.; Chang, K.-w.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to Computer Programmer as Woman

is to Homemaker? Debiasing Word Embeddings. In *NIPS*, 1–9.

Dror, R.; Baumer, G.; Shlomov, S.; and Reichart, R. 2018. The Hitchhiker' s Guide to Testing Statistical Significance in Natural Language Processing. In *56th Annual Meeting ofthe Association for Computational Linguistics*, 1–10.

Elson, J.; Douceur, J. R.; Howell, J.; and Saul, J. 2007. Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization. *Proceedings of 14th ACM Conference on Computer and Communications Security (CCS)* 366–374.

Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118.

Flaxman, A. D.; Serina, P. T.; Hernandez, B.; Murray, C. J. L.; Riley, I.; and Lopez, A. D. 2015. Measuring causes of death in populations: a new metric that corrects cause-specific mortality fractions for chance. *Population health metrics* 13(1):28.

Fraser, K. C.; Meltzer, J. A.; and Rudzicz, F. 2015. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease* 49(2):407–422.

ISO/IEC. 2014. Safety aspects – guidelines for their inclusion in standards. Standard, International Organization for Standardization, Geneva, CH.

Lashbrook, A. 2018. AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind. *The Atlantic*.

Murray, C. J. L.; Lopez, A. D.; Barofsky, J. T.; Bryson-Cahn, C.; and Lozano, R. 2007. Estimating Population Cause-Specific Mortality Fractions from in-Hospital Mortality: Validation of a New Method. *PLoS Medicine* 4(11):e326.

Verma, P., and Das, P. K. 2015. i-Vectors in speech processing applications: a survey. *International Journal of Speech Technology* 18(4):529–546.