Improving AI robustness with automatic hidden stratification detection

Mathieu Carrière
Inria

Maxime Duval
Kili Technology

The adoption of AI algorithms in the industry is now widespread, yet, applications of classical machine learning models often face several difficulties, one of them being the lack of relevant data to feed supervised algorithms. Industrial applications are often specific or specialized and the datasets you can find in the open-source world do not cover entirely all use cases. The goal of Kili Technology is to empower companies and data scientists in their daily machine learning challenges by reducing the workload associated to the annotation of unsupervised datasets. We have a first class application allowing users to label images, text, speech or videos according to their AI task (classification, object detection, entity recognition, ...).

Machine learning models perform well on data that "look" or "behave" similarly to data that they were trained on. In practice, though, it is unfeasible to collect enough training data to account for all potential test time scenarios. Current ML systems may fail when encountering out-of-distribution data. There is also a fundamental limitation to how we collect data and train models. Models may learn the "wrong" things, such as spurious correlations and dependencies on confounding variables that hold for most, but not all, of the data.

Christopher Ré, at Stanford, leads HazyResearch, a lab working on "the next generation of machine learning systems". One of the area they are working on is subgroup robustness or hidden stratification. With standard classification, we assign a single label for each sample in our dataset, and train a model to correctly predict those labels. However, several distinct data subsets or "subgroups" might exist among datapoints that all share the same label, and these labels may only coarsely describe the meaningful variation within the population (see examples here for medical [7] or imaging [6]). A new method, GEORGE [10], was developed on this topic to automatically identify hidden stratification. Using those subgroups, one can then boost model performance [9]. Alternatively, approaches based on Topological Data Analysis (TDA) have also been developed by the DataShape team at Inria. TDA [1, 3] is a field of data science that aims at discovering and encoding complex, topological patterns hidden in the data. As such, it also has been used for identifying hidden groups and structures in various data sets, that standard data science methods were not able to detect [5, 8]. More formally, recent advances in the TDA community have permitted to define stratifications of data sets w.r.t. model confidence [2], as well as topology-based criterion for detecting outliers and out-of-distribution data [4].

As a research intern, you goal will be to develop and improve automatic discovery of these hidden strata, and more generally, finding all possible failure modes of a model by analyzing datasets and models systematically in conjunction. This includes identifying and isolating data points that may be considered outliers, estimating performance on unlabeled data that is streaming to a deployed model, and generating rich summaries of how the data distribution may be shifting over time. Future directions for slice discovery will continue to improve our understanding of how to find slices that are interpretable, task-relevant, error-prone and susceptible to distribution shift. The internship will be co-supervised by Mathieu Carrière, DataShape (Inria) and Maxime Duval (Kili Technology). The internship will be hosted in Kili's offices, located 40 rue du Colisée, 75008 Paris, in a great work environment.

Contact : mathieu.carriere@inria.fr

References

[1] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[2] Mathieu Carrière and Bertrand Michel. Statistical analysis of Mapper for stochastic and multivariate filters. In *CoRR*. arXiv:1912.10742, 2019.

[3] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Society, 2010.

[4] Théo Lacombe, Yuichi Ike, Mathieu Carrière, Frédéric Chazal, Marc Glisse, and Yuhei Umeda. Topological Uncertainty: monitoring trained neural networks through persistence of activation graphs. In *30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*. International Joint Conferences on Artificial Intelligence Organization, 2021.

[5] Monica Nicolau, Arnold Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7265–7270, 2011.

[6] Luke Oakden-Rayner. Improving medical ai safety by addressing hidden stratification, Oct 2019.

[7] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

[8] Abbas Rizvi, Pablo Cámara, Elena Kandror, Thomas Roberts, Ira Schieren, Tom Maniatis, and Raúl Rabadán. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35:551–560, 2017.

[9] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[10] Nimit S Sohoni, Jared A Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *arXiv preprint arXiv:2011.12945*, 2020.