Simulating Real-World Imperfections: Assessing Machine Learning Model Robustness Against Label Noise and Distribution Shifts

David Jackson [1], Paul Groth [1], Hazar Harmouch [1]

[1] University of Amsterdam, Amsterdam, The Netherlands
{d.i.jackson, p.groth, h.harmouch}@uva.nl

In real-world applications, binary classification models frequently encounter obstacles such as label noise, distribution shifts, and particular data attributes, all of which can adversely impact both model performance and fairness. This work focuses exclusively on tabular data, conducting a thorough examination into the robustness of state-of-the-art machine learning models against label noise and distribution shifts. We simulate various forms of noise and assess the response of multiple models to these adversities. Our study introduces six types of noise, encompassing random label noise, conditional feature noise, class-conditional noise, and temporal/contextual noise, designed to mimic real-world inaccuracies without altering feature values. We also investigate complex noise interactions, such as combined conditional and class-conditional noise, to model intricate error patterns.

We study the impact of noise rates that vary from mild to severe on model accuracy and fairness by systematically changing noise rates to simulate real-world data of varying qualities. Our analysis includes an extensive array of models, including but not limited to CatBoost, LightGBM, XGBoost, TabTransformer, FT Transformer, MLP, ResNet, NODE, and SAINT, as well as fairness-oriented methods like DRO and Group DRO. Using both in-distribution and out-of-distribution data, we leverage the 'TableShift' benchmark [1], which provides 15 datasets specifically crafted for such evaluations. These datasets represent a wide variety of real-world problems, from diverse domains, each with unique data characteristics.

Our approach provides a holistic framework to simulate noise and evaluate model robustness. It acts as an important benchmark that allows researchers to compare various machine learning algorithms and evaluate the efficacy of error detection and data cleaning techniques. The present work, therefore, provides important knowledge on the behavior of models when conditions of information are noisy or change dynamically, along with several valuable insights for the development of more robust and equitable machine learning systems.

References

[1] Gardner, J., Popovic, Z., & Schmidt, L. (1983). Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems*, 36.