

A replication study of transformer-based TabPFN for assessing the applicability of neural-network based solutions in tabular classification.

Kartikey Chauhan - 501259284

Overview

Deep Learning has revolutionized the field of AI and led to remarkable achievements in applications involving image and text data. In particular, large transformer-based models trained on massive corpora, are disrupting machine learning in many areas. However, when it comes to tabular (a.k.a. structured) data, traditional machine learning methods, such as gradient-boosted decision trees have shown superior performance over deep learning. Recently, TabPFN [3] proposes a radical change to how tabular classification is done, introducing a pre-trained Transformer that is able to perform classification without training. This project aims to replicate and do an empirical review of TabPFN’s claims, while potentially exploring avenues for further scaling and modifications based on latest advancements in the Transformer space.

Objectives

- Replication of the transformer model creation and training.
- Comparative analysis of its classification and runtime performance against traditional ML methods (*LogReg*, *KNN*), AutoML systems (*Auto-sklearn 2.0*, *Autogluon*), and boosting machines (*XGBoost*, *CatBoost*, *LightGBM*).
- Analysis of potential model scaling and extension via hyperparameter optimization, extended training, or alternative architectures. Currently, TabPFN’s scale is limited (maximum of 1000 training data points, 100 numerical features without missing values, 10 classes).

Datasets

- The comparative analysis will be based on popular datasets from [OpenML.org](https://openml.org), unseen by the model’s priors, and extending the evaluation scope to more than the 18 datasets used in the paper.

Evaluation

- The **Receiver Operating Characteristic Area Under the Curve** (ROC AUC) and **Mean Time** to train and predict the classes are chosen as key metrics. The ROC AUC is a comprehensive metric that is particularly useful for imbalanced datasets, where accuracy can be misleading. Additionally, the ROC AUC is not affected by changes in the class distribution. Meanwhile, mean time to train and predict assesses the model’s computational efficiency, the crucial advantage of a pretrained network.

References

- [1] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Frank Hutter, Michel Lang, Rafael G. Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. Openml benchmarking suites. *arXiv:1708.03731v2 [stat.ML]*, 2019.
- [2] Benjamin Feuer, Chinmay Hegde, and Niv Cohen. Scaling tabpfn: Sketching and feature selection for tabular prior-data fitted networks. <https://arxiv.org/abs/2311.10609>, 2023.
- [3] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. <https://arxiv.org/abs/2207.01848>, 2023.
- [4] Andreas Müller, Carlo Curino, and Raghu Ramakrishnan. Mothenet: A foundational hypernetwork for tabular classification. <https://arxiv.org/abs/2312.08598>, 2023.
- [5] Guri Zabërgja, Arlind Kadra, and Josif Grabocka. Tabular data: Is attention all you need? <https://arxiv.org/abs/2402.03970>, 2024.