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

Kartikey Chauhan[1]

[1]Data Science & Analytics, Toronto Metropolitan University

Abstract

We perform a replication study of TabPFN, a transformer-based model for tabular data classification. We evaluate the model on a set of benchmark datasets and compare its performance against traditional machine learning methods and state-of-the-art AutoML systems. Our results show that TabPFN outperforms the baselines on most datasets, demonstrating the potential of neural-network based solutions in tabular classification tasks. We also provide an empirical review of TabPFN's claims and discuss potential avenues for further scaling and modifications based on the latest advancements in the Transformer space.

Keywords: TabPFN, Transformers, Tabular Data, Classification, Machine Learning

1. Introduction

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 hollmann2023tabpfn [1] 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.

2. Background on TabPFN

2.1 TabPFN Architecture

TabPFN proposed a two-stage approach for tabular data classification:

- First, meta-learn to approximate Bayesian inference using synthetic datasets.
- Second, use the labeled samples *in context* to classify unlabelled samples.

TabPFN is trained on a large number of synthetic datasets generated from a carefully designed prior distribution. This prior incorporates principles from causal reasoning and a preference for simple structures. Here are some key details about the prior used in TabPFN:

[1] https://github.com/automl/TabPFN

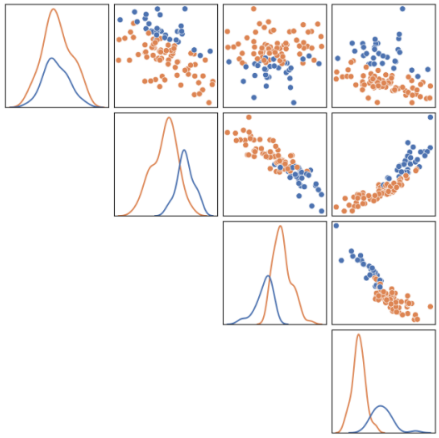

Figure 1: Visualization of Prior samples.

1. **Structural Causal Models (SCMs)**: The prior entails a space of structural causal models, which are used to generate the synthetic datasets. This allows TabPFN to learn the causal relationships and structures present in tabular data.

2. **Preference for Simplicity**: The prior has a preference for simpler causal structures over more complex ones. This bias towards simplicity helps TabPFN avoid overfitting and generalize better to new datasets.

3. **Bayesian Neural Networks (BNNs)**: In addition to SCMs, the prior also includes Bayesian neural networks as a possible data generating mechanism. This allows TabPFN to capture non-linear relationships in the data.

4. **Varying Dataset Characteristics**: The synthetic datasets are generated with varying numbers of features (up to 100), classes (up to 10), and sample sizes (up to 1024). This exposure to diverse dataset characteristics during training helps TabPFN generalize to a wide range of tabular problems.

5. **Hyperparameter Sampling**: The prior also includes sampling of various hyperparameters, such as MLP weight dropout, input feature scaling, and whether to sample the target variable from the last MLP layer or not.

By training on this carefully designed prior distribution, TabPFN learns to approximate Bayesian inference on tabular data tasks, allowing it to quickly adapt to new datasets through in-context learning without additional training or hyperparameter tuning.

```
TransformerModel(
  (transformer_encoder): TransformerEncoderDiffInit(
    (layers): ModuleList(
      (0-11): 12 x TransformerEncoderLayer(
        (self_attn): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=512, out_features=512, bias=True
        )
        (linear1): Linear(in_features=512, out_features=1024, bias=True)
        (dropout): Dropout(p=0.0, inplace=False)
        (linear2): Linear(in_features=1024, out_features=512, bias=True)
        (norm1): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
        (norm2): LayerNorm((512,), eps=1e-05, elementwise_affine=True)(dropout1): Dropout(p=0.0, i
        (dropout2): Dropout(p=0.0, inplace=False)
```

```
        )
      )
    )
    (encoder): Linear(in_features=100, out_features=512, bias=True)
    (y_encoder): Linear(in_features=1, out_features=512, bias=True)
    (decoder): Sequential(
      (0): Linear(in_features=512, out_features=1024, bias=True)
      (1): GELU(approximate='none')
      (2): Linear(in_features=1024, out_features=10, bias=True)
    )
    (criterion): CrossEntropyLoss()
)
```

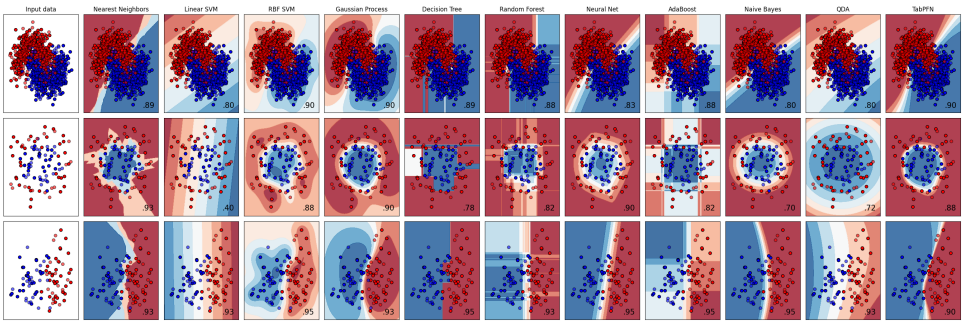

Figure 2: Decision boundaries on toy datasets generated with scikit-learn

2.2 How to create sections and subsections

Simply use the section and subsection commands, as in this example document! With Overleaf, all the formatting and numbering is handled automatically according to the template you've chosen. If you're using the Visual Editor, you can also create new sections and subsections via the buttons in the editor toolbar.

2.3 This is an example for second level head - subsection head

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

2.3.1 This is an example for third level head - subsubsection head

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

This is an example for fourth level head - paragraph head
Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

3. Replication and Evaluation

3.1 Baselines

We start with evaluation of the existing model provided the authors of TabPFN. We compare against five standard ML methods and two state-of-the-art AutoML systems for tabular data.

3.1.1 Datasets

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

This is an example for fourth level head - paragraph head
Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

4. Conclusions

Some conclusions here.

A. Some Notation

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

A.1 Appendix subsection title here

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.