What exactly has TabPFN learned to do?

Calvin McCarter

More results here!

Background on TabPFN

TabPFN[1] is a prior-data fitted network (PFN) that was pretrained to perform in-context learning on fresh tabular classification problems.

- Meta-learned Transformer model (26M parameters)
- Trained using synthetically-generating data: structural causal models are randomly generated, then training datasets are sampled from each causal model.
- On fresh classification tasks, no training (i.e. weight updating) is needed; instead, training data is given as context.

[1] TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. N. Hollmann, S. Muller, K. Eggensperger, F. Hutter. ICLR 2023.

Questions

- Was TabPFN overfit to tabular benchmarks?
- Has TabPFN learned fundamental principles of statistical learning?

Our approach

Understand what TabPFN-the-model-artifact does:

Observe TabPFN's test outputs as a function of its test inputs, for a variety of training sets:

- Simple settings, so we can plot the learned functions
- Out-of-domain settings, so we can evaluate generalization

TabPFN, with training data, is just a function:

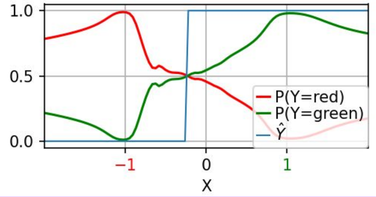
$$x_{\text{test}} \xrightarrow{f_{\theta_{\text{TabPFN}}}(X_{\text{train}}, y_{\text{train}})} \hat{y}_{\text{test}}$$

Simple synthetic settings

1d binary classification

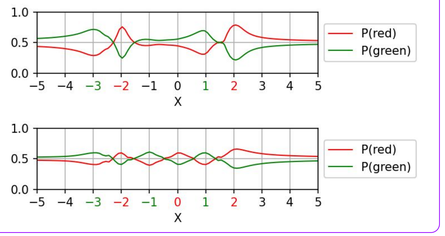
➔ expects right-skewed data

➔ has epistemic uncertainty outside the range of its training data



➔ cannot perform periodic interpolation

➔ cannot perform periodic extrapolation

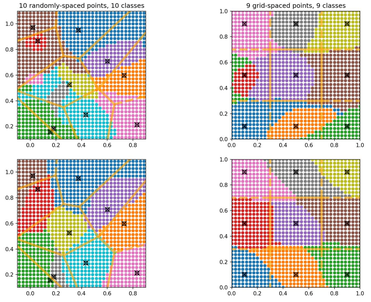


2d multi-class classification

Has TabPFN learned the principle of nearest-neighbor classification? Not without ensembling!

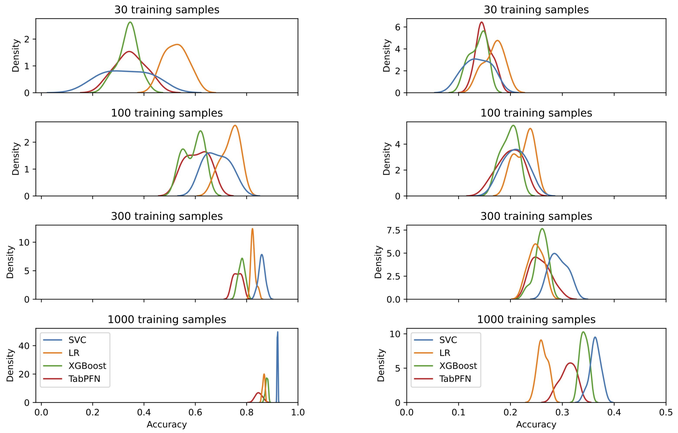
No ensembling:

32 ensembles:



Out-of-domain settings

Compare support vector (SVC), logistic regression (LR), XGBoost, and TabPFN on problems in computer vision and genomics.



MNIST CIFAR-10

Conclusions

TabPFN:

- does not appear overfit to tabular benchmarks.
- appears to *sometimes* implement fundamental statistical learning principles.

Pretrained tabular ML models should be evaluated like LLMs:

- Theoretical guarantees and benchmark results are not enough!
- Evaluate specific model artifacts, not only methods.