Tabular DL

Парфенова Анна
Руководитель: Игнашин Игорь

1 апреля 2025 г.

Введение и предпосылки

- Градиентный бустинг по-прежнему часто превосходит DL на табличных данных (но есть еще трансформеры)
- Недооценённый аспект: эмбеддинги числовых признаков
- Цель — изучить, как более выразительные эмбеддинги могут улучшить DL

О проекте TabM

- Репозиторий: `http://github.com/yandex-research/tabm`
- Основные цели:
  - Разработка SOTA-архитектур для табличных данных
  - Изучение альтернатив классическим GBDT-моделям
  - Исследование способов представления признаков, особенно числовых

Основная идея

- Вместо прямой подачи скалярных признаков — перевод их в векторы (эмбеддинги)
- Эти эмбеддинги затем обрабатываются стандартным backbone (MLP, Transformer и т.д.)
- Эмбеддинги можно строить различными способами: от простых линейных до периодических

Разные подходы к эмбеддингу

- **Piecewise Linear Encoding (PLE)** — биннинг + линейная интерполяция
- **Периодические функции** — $\sin(2\pi cx), \cos(2\pi cx)$, параметры c обучаемые
- **Классические блоки DL** — линейные слои, ReLU и т.п.

Комбинации моделей

- Backbone + Embedding:
  - Backbones: MLP, ResNet, Transformer
  - Embeddings: PLE, Periodic, Linear/ReLU, AutoDis
- Пример: `MLP-PLR`, `ResNet-T-LR`, `Transformer-PLR`

	GE ↑	CH ↑	CA ↓	HO ↓	AD ↑	OT ↑	HI ↑	FB ↓	SA ↑	CO ↑	MI ↓	Avg. Rank
CatBoost	0.692	0.861	0.430	3.093	0.873	0.825	0.727	5.226	0.924	0.967	**0.741**	3.6 ± 2.9
XGBoost	0.683	0.859	0.434	3.152	**0.875**	0.827	0.726	5.338	0.919	0.969	0.742	4.6 ± 2.7
MLP	0.665	0.856	0.486	3.109	0.856	0.822	0.727	5.616	0.913	0.968	0.746	8.5 ± 2.6
MLP-LR	0.679	0.861	0.463	3.012	0.859	0.826	0.731	5.477	0.924	0.972	0.744	5.5 ± 2.7
MLP-Q-LR	0.682	0.859	0.433	3.080	0.867	0.818	0.724	**5.144**	0.924	0.974	0.745	5.1 ± 1.9
MLP-T-LR	0.673	0.861	0.435	3.099	0.870	0.821	0.727	5.409	0.924	0.973	0.746	5.1 ± 1.7
MLP-PLR	**0.700**	0.858	0.453	**2.975**	0.874	**0.830**	**0.734**	5.388	**0.924**	0.975	0.743	3.0 ± 2.4
ResNet	0.690	0.861	0.483	3.081	0.856	0.821	0.734	5.482	0.918	0.968	0.745	6.7 ± 3.3
ResNet-LR	0.672	0.862	0.450	2.992	0.859	0.822	0.733	5.415	0.923	0.971	0.743	5.6 ± 2.7
ResNet-Q-LR	0.674	0.859	0.427	3.066	0.868	0.815	0.729	5.309	0.923	0.976	0.746	4.7 ± 2.0
ResNet-T-LR	0.683	0.862	**0.425**	3.030	0.872	0.822	0.731	5.471	0.923	0.975	0.744	4.1 ± 1.9
ResNet-PLR	0.691	0.861	0.443	3.040	**0.874**	0.825	0.734	5.400	0.924	0.975	0.743	3.2 ± 1.3
Transformer-L	0.668	0.861	0.455	3.188	0.860	0.824	0.727	5.434	0.924	0.973	0.743	5.9 ± 2.2
Transformer-LR	0.666	0.861	0.446	3.193	0.861	0.824	0.733	5.430	0.924	0.973	0.743	5.2 ± 2.2
Transformer-Q-LR	0.690	0.857	**0.425**	3.143	0.868	0.818	0.726	5.471	**0.924**	0.975	0.744	4.4 ± 2.2
Transformer-T-LR	0.686	0.862	**0.423**	3.149	0.871	0.823	0.733	5.515	0.924	**0.976**	0.744	3.7 ± 2.2
Transformer-PLR	0.686	**0.864**	0.449	3.091	0.873	0.823	0.734	5.581	**0.924**	0.975	0.743	3.9 ± 2.5

Какая работа была проделана

- Изучена статья On Embeddings for Numerical Features in Tabular Deep Learning от команды Яндекса
- Запущены уже существующие эксперименты проекта TabM
- Предприняты попытки сделать модель вида MLP-PLR и проверить ее эффективность

Будущая работа

- Хотим улучшить бенчмарки за счет более обучаемых периодических эмбеддингов
- Отследить насколько хорошо они обучаются (или насколько плохо)
- Попробовать сделать их более обучаемыми за счет какого-то кастомного оптимайзера
- Регуляризация на веса:
  - L_2 на параметры частот
  - **Total Variation Loss**: $TV(w) = \sum_i |w_i - w_{i+1}|$