# Дистилляция табличных нейронных сетей

Аушев Ислам

18 декабря 2023 г

ШКОЛА АНАЛИЗА ДАННЫХ

# Постановка задачи

На данный момент нейронные сети показывают SOTA результаты для многих доменов данных.

Исключением являются табличные данные, для которых чаще всего используют GBDT подходы.

Однако, в последнее время нейронные сети показывают хорошие результаты и для таблиц

Yandex Research : https://arxiv.org/pdf/2106.11959.pdf

| | CA ↓ | AD ↑ | HE ↑ | JA ↑ | HI ↑ | AL ↑ | EP ↑ | YE ↓ | CO ↑ | YA ↓ | MI ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Default hyperparameters | | | | | | | | | | | |
| XGBoost | 0.462 | **0.874** | 0.348 | 0.711 | 0.717 | 0.924 | 0.8799 | 9.192 | 0.964 | 0.761 | 0.751 |
| CatBoost | **0.428** | 0.873 | 0.386 | 0.724 | 0.728 | 0.948 | 0.8893 | 8.885 | 0.910 | 0.749 | 0.744 |
| FT-Transformer | 0.454 | 0.860 | **0.395** | **0.734** | **0.731** | **0.966** | **0.8969** | **8.727** | **0.973** | **0.747** | **0.742** |

# Постановка задачи

С ростом качества нейронных сетей на табличных данных увеличился и размер моделей, что сказалось на их производительности.

Скорость инференса моделей:

| | CA | HE | JA | HI | AL | YE |
|---|---|---|---|---|---|---|
| MLP | 0.009 | 0.034 | 0.053 | 0.046 | 0.081 | 0.402 |
| AutoInt | 0.071 | 0.277 | 1.273 | 0.217 | 2.221 | 9.680 |
| ResNet | 0.026 | 0.295 | 0.257 | 0.137 | 0.667 | 1.033 |
| Node | 1.078 | 3.353 | 8.075 | 5.663 | 6.433 | 26.141 |
| FT-transformer | 0.123 | 1.307 | 2.929 | 0.898 | 6.433 | 3.736 |
| CatBoost | 0.016 | 0.248 | 0.127 | 0.064 | 5.862 | 0.184 |

Одним из способов увеличения производительности модели является дистилляция.

# Постановка задачи

Дистилляция - способ обучения нейронных сетей на основе знаний, полученных из уже ранее обученной модели

Размер модели ученика берут меньше, чем у учителя.
Distillation Loss:

$$-\sum_i y_i \log(p_i) + \alpha \cdot D_{KL}(q||p)$$

Кросс-энтропия ученика + KL-дивергенция между учителем и учеником.

DistillBERT : https://arxiv.org/pdf/1910.01108.pdf

| Model | Score | CoLA | MNLI | MRPC | QNLI | QQP | RTE | SST-2 | STS-B | WNLI |
|---|---|---|---|---|---|---|---|---|---|---|
| ELMo | 68.7 | 44.1 | 68.6 | 76.6 | 71.1 | 86.2 | 53.4 | 91.5 | 70.4 | 56.3 |
| BERT-base | 79.5 | 56.3 | 86.7 | 88.6 | 91.8 | 89.6 | 69.3 | 92.7 | 89.0 | 53.5 |
| DistilBERT | 77.0 | 51.3 | 82.2 | 87.5 | 89.2 | 88.5 | 59.9 | 91.3 | 86.9 | 56.3 |

# Постановка задачи

Во многих случаях Catboost отрабатывает быстрее, чем табличные нейронные сети.

Можно ли улучшить качество GBDT, имея на руках обученную нейронную сеть?

# Цель задачи

Основная цель : исследование возможностей дистилляции нейронных сетей на GBDT

# Идея

Пусть $X$ - трейн-датасет, $y$ - таргет, N - нейросеть-учитель. Обучаем $N$ на данных $(X, y)$, получаем новые метки $y'$

$$y' = N(X)$$

Обучаем GBDT на данных $(X, y')$

# Идея

Мотивация такого метода :

1. Данный метод не использует нейросеть-учитель во время инференса
2. Наиболее общий подход, который не учитывает специфику конкретной выборки данных
3. Интуитивно напоминает метод дистилляции между нейронными сетями

# Ход решения

Для проведения экспериментов были взяты 6 датасетов : California Housing(CA), Helena(HE), Jannis(JA), Higgs(HI), Aloi(AL), Year(YE).

В качестве учителя рассматривались 5 моделей : MLP, AutoInt, Node, ResNet, FT-Transformer.

В качестве GBDT модели CatBoost

# Ход решения

Выбор моделей объясняется разнообразием архитектур

MLP - Fully Connected Layers

ResNet - Fully Connected Layers + Residual Connections

Node - Differentiable Decision Trees

AutoInt - Attention

FT-Transformer - Attention + Tricks

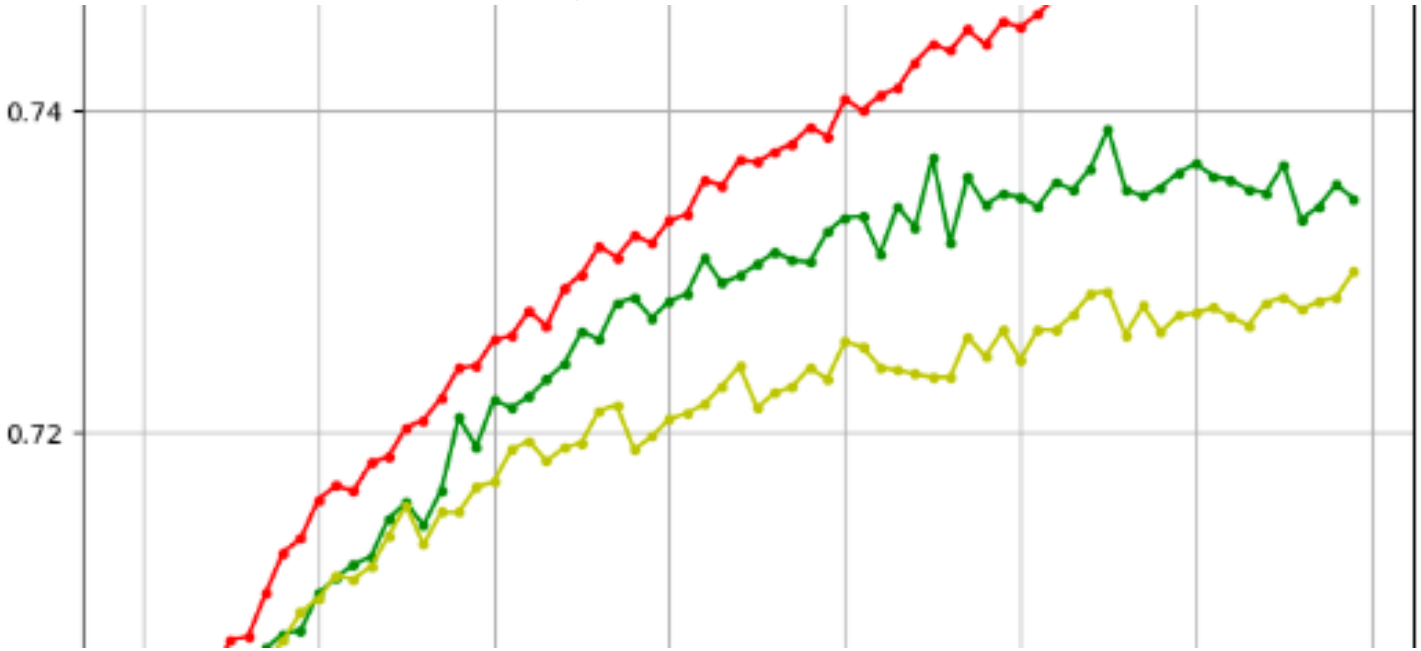
# Ход решения

Для лучшего воспроизведения результатов, были использованы гиперпараметры и методы предобработки данных, описанные в статье Yandex Research : https://arxiv.org/pdf/2106.11959.pdf

# Ход решения

Табличные нейросети учатся нестабильно, метрика скачет от эпохи к эпохе, что осложняет дальнейшую дистилляцию.

Метрика на шаге обучения:




Используем early-stopping по валидации.

# Ход решения

Как оказалось, дистилляция очень сильно зависит от обучения модели, модели имеющие близкие метрики могут давать разные результаты для бустинга.

Поэтому обучаем много моделей, а затем собираем из них ансамбль, агрегируя результаты каждой модели. Далее учим CatBoost на таргетах ансамбля.

Размер ансамблей варьируется от 3 до 5

# Результаты

Качество моделей:

| | CA | HE | JA | HI | AL | YE |
|---|---|---|---|---|---|---|
| MLP | $0.506 \pm 2e{-3}$ | $0.384 \pm 2e{-3}$ | $0.717 \pm 5e{-4}$ | $0.721 \pm 1e{-3}$ | $0.949 \pm 1e{-3}$ | $8.861 \pm 6e{-3}$ |
| AutoInt | $0.482 \pm 3e{-3}$ | $0.370 \pm 2e{-3}$ | $0.712 \pm 6e{-3}$ | $0.721 \pm 6e{-3}$ | $0.940 \pm 6e{-3}$ | $8.963 \pm 2e{-2}$ |
| ResNet | $0.496 \pm 3e{-3}$ | $0.395 \pm 1e{-3}$ | $0.723 \pm 7e{-4}$ | $0.721 \pm 1e{-3}$ | $0.961 \pm 2e{-4}$ | $8.810 \pm 4e{-3}$ |
| Node | $0.474 \pm 7e{-4}$ | $0.357 \pm 1e{-3}$ | $0.724 \pm 2e{-3}$ | $0.725 \pm 1e{-3}$ | — | $8.792 \pm 4e{-3}$ |
| FT-Transformer | $0.465 \pm 2e{-3}$ | $0.390 \pm 1e{-3}$ | $0.732 \pm 1e{-3}$ | $0.729 \pm 1e{-3}$ | $0.951 \pm 2e{-3}$ | $8.852 \pm 4e{-4}$ |
| CatBoost | 0.451 | 0.382 | 0.724 | 0.725 | 0.945 | 8.886 |

Качество дистилляции:

| | CA | HE | JA | HI | AL | YE |
|---|---|---|---|---|---|---|
| MLP | $0.485 \pm 1e{-2}$ | $0.377 \pm 2e{-3}$ | $0.716 \pm 2e{-3}$ | $0.720 \pm 8e{-4}$ | $0.948 \pm 2e{-3}$ | $8.905 \pm 6e{-3}$ |
| AutoInt | $0.471 \pm 3e{-3}$ | $0.367 \pm 3e{-4}$ | $0.710 \pm 3e{-3}$ | $0.722 \pm 1e{-3}$ | $0.949 \pm 3e{-4}$ | $9.007 \pm 3e{-3}$ |
| ResNet | $0.484 \pm 2e{-3}$ | $0.383 \pm 9e{-4}$ | $0.717 \pm 1e{-3}$ | $0.721 \pm 2e{-3}$ | $0.952 \pm 3e{-3}$ | $8.868 \pm 5e{-3}$ |
| Node | $0.484 \pm 3e{-3}$ | $0.355 \pm 8e{-4}$ | $0.724 \pm 3e{-4}$ | $0.723 \pm 1e{-3}$ | — | $8.860 \pm 1e{-3}$ |
| FT-Transformer | $0.469 \pm 7e{-3}$ | $0.380 \pm 2e{-3}$ | $0.725 \pm 2e{-3}$ | $0.725 \pm 1e{-3}$ | $0.947 \pm 4e{-4}$ | $8.921 \pm 6e{-3}$ |

В каждой строке результат дистилляции модели в бустинг.

Красным цветом указан прирост метрики у CatBoost

Качество дистилляции ансамблей:

| | CA | HE | JA | HI | AL | YE |
|---|---|---|---|---|---|---|
| MLP | 0.479 | 0.377 | 0.716 | 0.721 | 0.947 | 8.900 |
| AutoInt | 0.468 | 0.371 | 0.713 | 0.723 | 0.951 | 8.978 |
| ResNet | 0.484 | 0.384 | 0.719 | 0.722 | 0.952 | 8.863 |
| Node | 0.484 | 0.355 | 0.724 | 0.725 | — | 8.857 |
| FT-Transformer | 0.466 | 0.383 | 0.728 | 0.726 | 0.947 | 8.917 |
| CatBoost | 0.451 | 0.382 | 0.724 | 0.725 | 0.945 | 8.886 |

Дистилляция дает прирост качества, дистилляция с ансамблей показывает лучший результат.
Результат дистилляции прямо пропорционален качеству модели.

# Результаты

Отметим некоторые наблюдения из результатов :

1. Если качество исходной модели хуже, чем качество бустинга, то дистилляция не дает выигрыша
2. Если качество исходной модели лучше, чем качество бустинга, то почти всегда дистилляция дает прирост метрики
3. Лучше дистиллировать ансамбли, а не отдельные модели

# Перспективы

1. В данной работе был рассмотрен самый примитивный способ дистилляции, однако, даже при таком подходе качество бустинга можно улучшить. Возможно изучение других способов.
2. В датасетах, используемых в данном проекте, отсутствуют категории. В дальнейшем планируется провести эксперименты на данных с категориальными признаками.
3. Из результатов работы нельзя точно сказать, какая архитектура наиболее удобна для бустинга в качестве учителя, это можно изучать дальше.