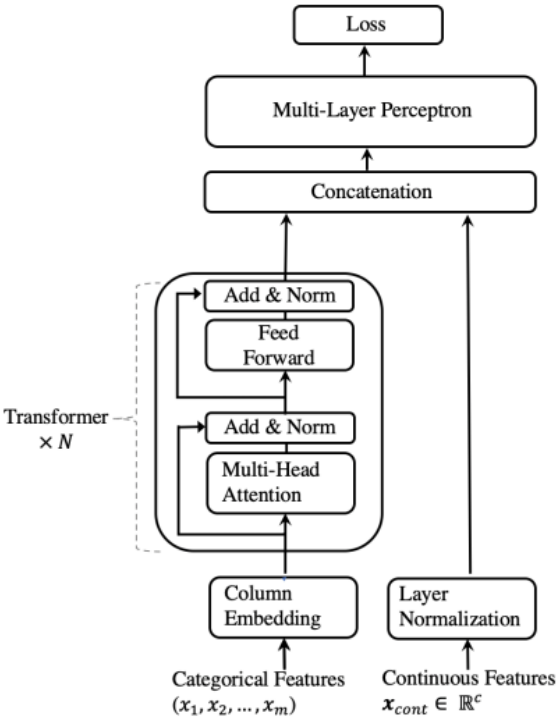
Хабибуллин Тимур
24 мая 2022 г.

Устранение пропусков в табличных данных

TabTransformer - Sota в Deep Learning для работы с табличными данными, который позволяет решать задачи классификации/регрессии.



Основной принцип работы:

1. Создаются контекстуальные эмбеддниги для *категориальных* признаков (для каждого столбца обучается дополнительный эмбеддинг для значения NaN)
2. Они конкатенируются вместе со значениями непрерывных признаков
3. Большой полученный вектор переддается на вход MLP, который позволяет решать поставленную задачу
4. Функция потерь зависит от постановки задачи - классификация (cross entropy loss) или регрессия (MSE)

Хочется на основе этой модели как-то научиться устранять пропуски в данных, а именно - пропуски в категориальных признаках.

Первая идея:

1. Обучить TabTransformer на части данных без пропусков для основной задачи (классификация/регрессия), модель научится делать контекстуальные эмбеддинги для всех значений категориальных фичей.

2. После этого в оставшейся части данных с пропусками для каждой фичи $x_i = \text{NaN}$ пробовать подобрать лучшее значение $x_i = a^*$: перебрать разные значения пропущенной фичи $x_i = a$ (фичи категориальные, поэтому их ограниченное количество) и посмотреть, где получается лучшее сопоставление эмбеддинга w_i контексту (т.е. сравнить с представлениями остальных категориальных фичей w_j после инфренса), которое можно выразить через поиск минимума NLLL:

$$L(a) = -\frac{1}{m}\sum_{j \neq i} \log P(x_i = a | x_j)$$

$$P(x_i = a | x_j) = \frac{\exp(\langle w_i, w_j \rangle)}{\sum\limits_{j \neq i} \exp(\langle w_i, w_j \rangle}$$

$$a^* = \arg\min_a L(a)$$

Вторая идея:

1. Сделать предобучение с помощью Mask Language Model (MLM) - убрать лейблы для решения основной задачи (классификация/регрессия), взять часть данных без пропусков, на каждом шаге делать маски и научиться предсказывать, что спрятано за масками. Это можно сделать как в BERT - эмбеддниги, которые получились для маскок, просто пропускаются через FFNN с выходом, длина которого есть количество всех уникальных значений фичей среди всех столбцов (типо словарик), берется argmax Softmax(...) и тд

2. С помощью этой предобученной модели в режиме MLM предсказать значения в $x_i = \text{NaN}$

Эту предобученную модель можно потом доучить на датасете без пропусков для решения основной задачи.

Датасеты без пропусков интересно попробовать скормить для решения основной задачи с помощью CatBoost и TabTransformer, а также доучить предобученный TabTransformer. Результаты этих моделей можно сравнить результатами работы CatBoost и TabTransformer на исходных датасетах с пропусками

Ожидается, что после устранения пропусков категориальных фичей метрики (в статье используется AUC) у обеих моделей должны стать лучше.