Улучшение нейросетевых методов предсказания функции распределения для применения в задачах глубокого машинного обучения

Мишустина Маргарита Владимировна

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: mishustina.rita@yandex.ru

Научный руководитель — *Китов Виктор Владимирович*

В задачах традиционного и глубокого машинного обучения часто необходимо исследовать распределения признаков. Полезным инструментом для работы с распределением является функция распределения. С помощью функции распределении признака можно, например, бинаризовать значения фактора по квантилям, а затем заменить эти значения на обучаемые квантили ([1]). Кроме того, значения многомерной функции распределения могут быть использованы для поиска выбросов (то есть таких объектов, которые существенно отличаются от остальных): объект, чья функция распределения будет выходить за границы определенных по валидации квантилей (например, q_5 и q_{95}) – считаются выбросами.

В работе улучшается существующий непараметрический нейросетевой метод оценки распределений с помощью изменения способа сэмплирования целевой переменной. Разработанный метод позволяет добиться уменьшения значения функции потерь на синтетических данных, сгенерированных из различных распределений (например, нормальное распределение, многомерное нормальное распределение, распределение Коши и тд). Обновленная схема работа выглядит так:

1. $t = 1$, $x_1 \leq x_2 \leq \ldots \leq x_N$, $x_i \in \mathbb{R}^1$, $\forall i \in 1, \ldots, N$ – берем точки из минибатча,
2. $u_1 \leq u_2 \leq \ldots \leq u_N \sim U[0,1]$ – генерируем таргеты
3. $$J = \sum_{n=1}^{N} \Big[H(x_n, w) - u_n \Big]^2 +$$

$$\underbrace{+\lambda\sum_{k=1}^{K}\mathbb{I}(H(y_k,w)>H(y_k+\Delta,w))\Big[H(y_k,w)-H(y_k+\Delta,w)\Big]^2}_{\text{штраф за монотонность}},$$

– оптимизируем функцию потерь где $y_1, y_2, \ldots, y_K$ — любой набор точек из обучающего набора данных, $\Delta > 0$ – гиперпараметр, λ – коэффициент регуляризации.

4. $t = t + 1$

В силу естественного ограничения на монотонность функции распределения, в исходном алгоритме используется дополнительная регуляризация, отвечающая за то, чтобы результирующая нейронная сеть представляла собой неубывающую функцию. В работе предлагается способ избавиться от этой регуляризации путем наложения ограничения на архитектуру и веса модели. В предложенной модели используются линейные слои и регуляризация (композиция данных слоев является монотонной при условии неотрицательности весов линейных слоев), при этом неотрицательность весов контролируется, например, таким образом: $\hat{W} = \max(W, 0)$. Данная модификация позволяет несколько ускорить абсолютное время сходимости метода.

Кроме того, новый метод обобщается на многомерный случай. Пример работы обобщенного метода для смеси нормальных распределений представлен в разделе «иллюстрации». Схема обобщенного метода выглядит следующим образом:

1. $t = 1$, $x_1, x_2, \ldots, x_N$, $x_i \in \mathbb{R}^D$, $\forall i \in 1, \ldots, N$ – берем точки из мини-батча (D – размерность пространства),

2. $$J = \sum_{n=1}^{N}\Big[H(x_n,w)-t_n\Big]^2+$$

$$+\lambda\sum_{d=1}^{D}\sum_{k=1}^{K}\mathbb{I}(H(y_{k1},\ldots,y_{kd}+\Delta,\ldots,y_{kD}),w)>$$

$$>H((y_{k1},\ldots,y_{kd},\ldots,y_{kD}),w))\cdot$$

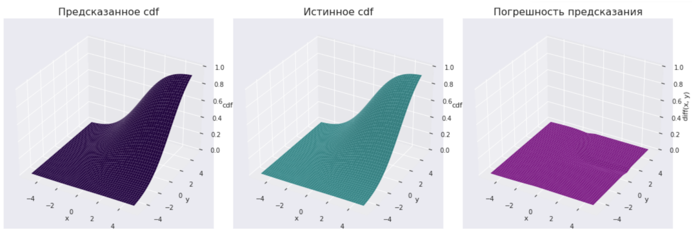
$$\cdot\Big[H((y_{k1},\ldots,y_{kd}+\Delta,\ldots,y_{kD}),w)-H((y_{k1},\ldots,y_{kd},\ldots,y_{kD}),w)\Big]^2,$$

– оптимизируем обобщенную функцию потерь, где $y_1, y_2, \ldots, y_K$ — любой набор точек из обучающего набора данных, $\Delta > 0$, λ – коэффициент регуляризации.

3. $t = t + 1$.

Преимущество нового метода перед традиционным статистическим методом показывается на только на синтетических данных, упомянутых ранее, но также и в практических задачах машинного обучения - обнаружении выбросов и кодировании вещественных признаков через эмбеддинги.

Иллюстрации



Пример работы разработанного обобщения для смеси нормальных распределений.

Литература

1. Gorishniy Y., Rubachev I., Babenko A. On embeddings for numerical features in tabular deep learning //arXiv preprint arXiv:2203.05556. – 2022.
2. Nonparametric Density Estimation for High-Dimensional Data - Algorithms and Applications (https://arxiv.org/pdf/1904.00176.pdf)