ICIEA 2022

16 - 19 Dec 22
Chengdu, China

ICIEA22-000426
Investigation of Rounding Algorithms Combining Data Distribution Characteristics

Mengyuan ZHU* , Qian ZHANG, Yunwei ZHANG, Tao SHEN, Baochang ZHANG
China Information Technology Designing & Consulting Institute Co., Ltd.

Abstract

The default ratings for the scenarios like user ratings or recommendation systems are usually expressed as integers. However, the prediction results of regression models based on machine learning or deep learning are usually floating-point decimals. The bias between integer ratings and floating-point decimals expanded the mean absolute error (MAE). In this paper, we first compared the results among three conventional rounding methods, i.e., rounding up, rounding down, and rounding off. The results show that conventional rounding method does not consider the information of original data distribution. Therefore, we established two novel rounding algorithms combining the real-world data distribution, which aim to reduce the MAE of the LightGBM regression outputs. First, an adjacent rounding algorithm combining the data distribution of adjacent labels was proposed. By this means, the predicted value should be the one with the higher distribution frequency between the two integer labels. Then, we extended the rounding algorithm to the top-n label values with higher distribution frequency than others. Moreover, a global optimal rounding algorithm is proposed by taking the distance information between each predicted value and the first n labels as another influencing factor. The proposed method has demonstrated potential applications in recommender systems, user ratings and other scenarios.

Introduction

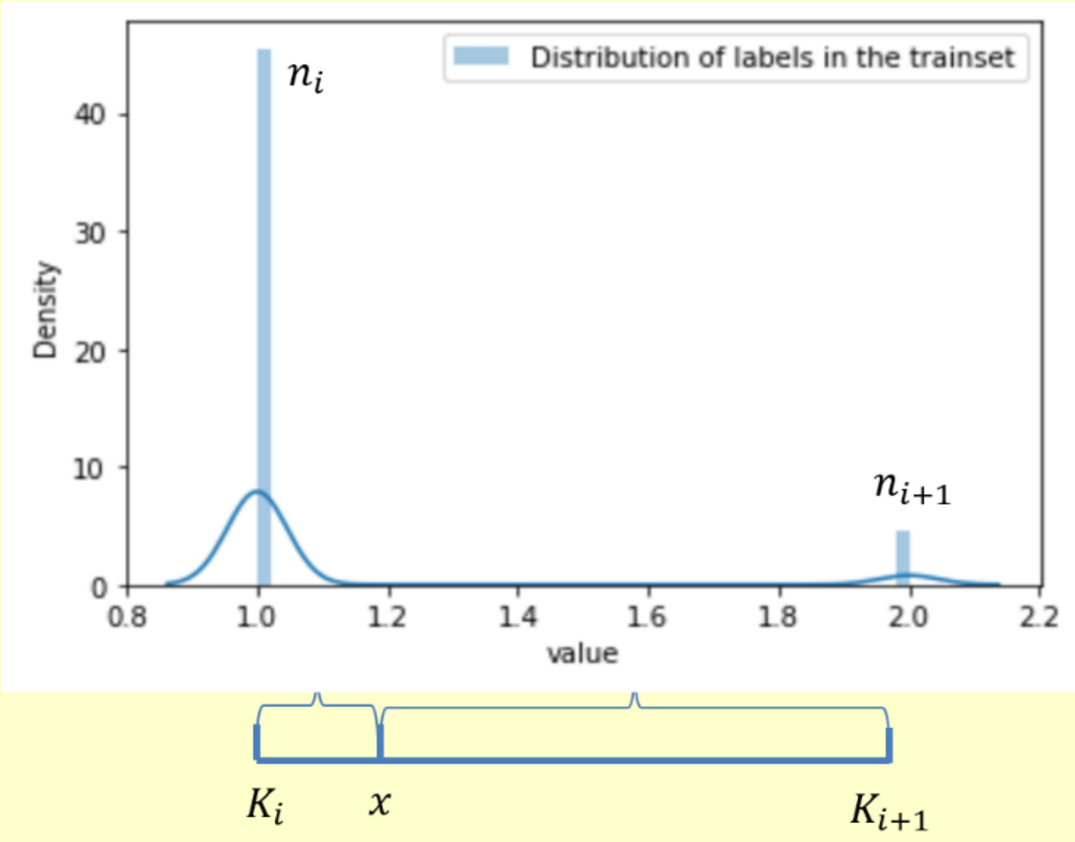
The predicted results of the regression models based on machine learning or deep learning should be floating-point decimals. However, the default ratings are usually expressed as integers for user ratings or recommendation systems. Therefore, the output of the regression model needs to be rounded to eliminate the bias between integer ratings and floating point decimals. However, conventional methods do not take into account the data distribution characteristics of the ratings in the train set and therefore cannot largely decrease the original MAE as expected.

Methods

We proposes two novel rounding algorithms considering the characteristics of the rating data distribution.

A. Adjacent rounding algorithm

We associate the threshold value with the data distribution of user ratings in the train set, setting the threshold according to the proportion of the distribution of adjacent ratings.
Let the result predicted by the regression model be x, which is a floating-point decimal between two adjacent integer labels K_i and K_{i+1}, The frequencies of the labels K_i and K_{i+1} in the train set are n_i and n_{i+1}, respectively, $i \in [0, 9]$. As shown in Fig. 4, the probability of a label's appearance can represent the frequency that the predicted score belongs to this label.



Adjacent rounding algorithm

$$F(x) = \begin{cases} 1, x < 1 \\ K_i, K_i \le x < K_i + \frac{n_i}{n_i + n_{i+1}}, \\ K_{i+1}, \frac{n_i}{n_i + n_{i+1}} \le x \le K_{i+1} \\ 10, x > 10 \end{cases}$$

B. Global Optimal Rounding Algorithm

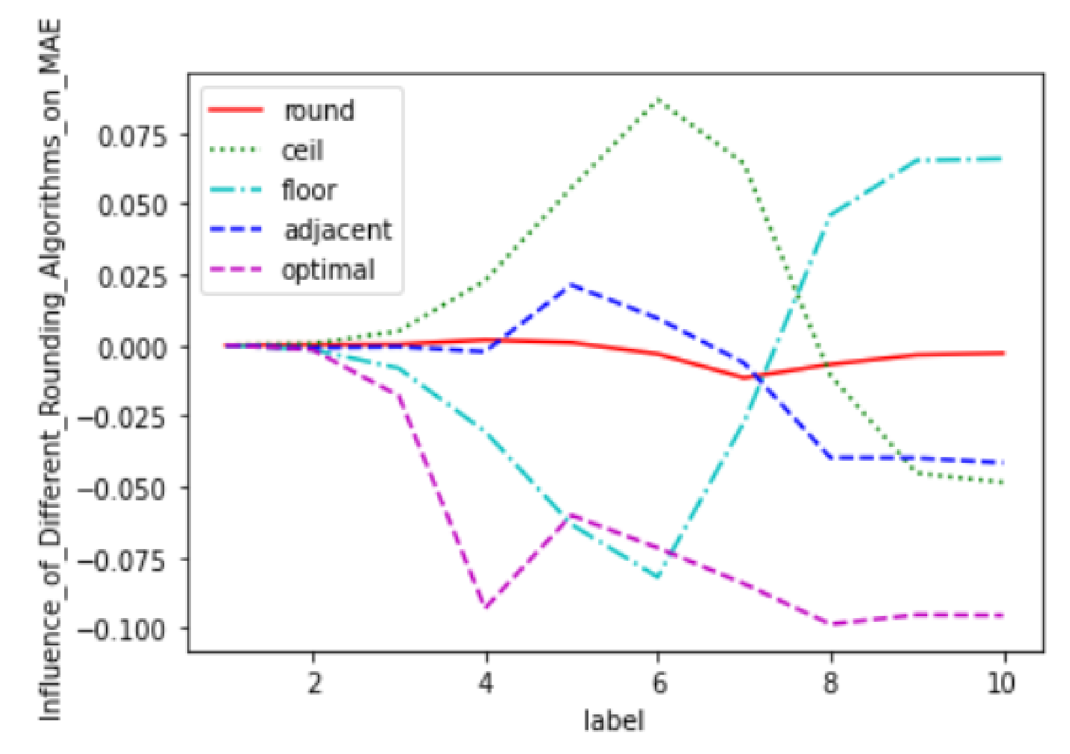
the top *N* labels with the highest frequency of appearance can be selected as the rounding object of the predicted value. The global optimal rounding function $F(x)$ can be written as:

$$F(x) = \begin{cases} K_i, if\ x = K_i \\ argmax\left(\frac{f(K_i)}{|x - K_i|^p}\right), if\ x \ne K_i \end{cases}, i = 0,1 \ldots 9$$

Equation calculates the distance between it and the top n-th user rating label values K_i with the highest frequency, and the respective occurrence probabilities $f(K_i)$ of the top n-th label values K_i in the train set.

Results

The proposed adjacent rounding algorithm and the global optimal rounding algorithm were applied to process the user rating data, respectively.



Influence of adjacent rounding and global optimal rounding on MAE

As can be seen, compared with the conventional rounding methods, the two rounding algorithms proposed considering the data distribution characteristics can reduce the MAE of the result predicted by the regression model, thereby improving the prediction accuracy. The global optimal rounding algorithm can further reduce MAE of prediction result and truth value and improve the accuracy of the prediction results, because it considers multiple factors, such as the data distribution and the distance between each predicted value and the user rating label.

Summary

This paper established two novel rounding algorithms combining the real-world data distribution, which aim to reduce the MAE of the LightGBM regression outputs. Compared with the conventional rounding methods, the proposed methods successfully considers various factors affecting the rounding propensity, including the distance of the predicted value from the adjacent integer label, the frequency distribution of each integer label, the top N-th user rating categories. Therefore, each predicted value has a greater probability of orienting to its appropriate integer. The proposed rounding algorithm takes the data distribution into account and thus improves the prediction accuracy of the regression model.