CatBoost
Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

29 Jul 2022

1. Introduction

Category boosting (CatBoost) is a supervised machine learning algorithm that is used for both classification and regression. Just like many other boosting algortihms, it uses a sequence of decision trees with reduced loss compared to previous trees. It has a fast learning process and gives very high accuracy when the dataset has a lot of categorical features.

2. Features

Quantization

In CatBoost, data is split into buckets for improving speed. This is done by a technique called quantization. Categorical labels are converted to numerical labels before quantization. There are many different quantization modes available in CatBoost like

- Median
- Uniform
- UniformAndQuantiles
- MaxLogSum
- MinEntropy
- GreedyLogSum

Tree Generation

The decision trees are generated by a greedy algorithm. The splitting of leaves happens based on the buckets obtained from quantization. At each level, many possible decisions are enumerated and a number of penalty functions are calculated for each of them. The decision with the least penalty is selected for splitting. After each tree is made, the classification objects are randomly permuted and the structure of the next tree is decided by a metric, which is calculated sequentially.

Symmetric Trees

Unlike other boosting algorithms, CatBoost builds symmetric trees. In a symmetric tree, all decision nodes in the same level split using the same condition. This prevents overfitting and makes the model faster.

3. Questions

i) What is the biggest advantage of CatBoost?
ii) What is a drawback of CatBoost?
iii) List some parameters used in CatBoost.
iv) Explain the median mode of quantization.
v) Rank CatBoost, XGBoost and LightGBM in terms of speed.

4. Answers

i) CatBoost outperforms other boosting algorithms when the input dataset has many categorical features. It handles categorical and textual features automatically.
ii) CatBoost does not deal with sparse datasets very well.
iii) Some commonly used parameters in CatBoost are quantization mode, loss function, learning rate and the maximum number of trees.
iv) In the median mode of quantization, every bucket has approximately the same number of objects.
v) LightGBM > CatBoost > XGBoost