LightGBM
Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

27 Jul 2022

1. Introduction

Light gradient boosting machine (LightGBM) is a supervised machine learning algorithm that is used for both classification and regression. As the name suggests, it is a lightweight boosting algorithm that provides high accuracy with low computational power. It is very efficient and is much faster than other boosting algorithms.

2. Features

Leaf-wise Trees

LightGBM builds its decision trees leaf-wise as opposed to level-wise, which is done by most other boosting algorithms including XGBoost. It means that instead of building the tree level-by-level, a particular leaf is selected and the tree is split at that leaf converting it into a decision node. Typically, the leaf with the maximum delta loss is chosen for splitting. Leaf-wise generation of trees improves the accuracy of the model.

Histogram-based Algorithms

The tree generation itself is done using histogram-based algorithms. The data is split into a number of groups known as bins. These bins are used to build the decision trees instead of the individual data values. The use of histogram-based algorithms makes the model faster and reduces memory requirements. This is because the number of bins is usually much smaller than the number of data samples.

Gradient-Based One-Side Sampling

Gradient-based one-side sampling (GOSS) is a sampling technique. Not every data in the training set is used to generate the trees. The gradients are first sorted. Samples with high gradients are all chosen because high gradients imply high error, so it is important to consider all of them for training. Samples with low gradients however, are not as important and are randomly sampled to obtain a few of them. These sampled data points along with the ones with high gradients are together used to train the decision trees. Sampling improves accuracy as it prevents overfitting.

Exclusive Feature Bundling

Exclusive feature bundling (EFB) is a feature engineering technique. Often, data contains exclusive features viz. features that rarely take non-zero values simultaneously. EFB condenses these exclusive features into something known as a bundle. This greatly reduces the number of features and improves the accuracy of the model.

3. Questions

i) List some hyperparameters involved in LightGBM.
ii) What makes LightGBM faster than XGBoost?
iii) How does LightGBM deal with sparse feature spaces?
iv) Give an example where you finid exclusive features.
v) What is a naive way to bundle the above exclusive features?

4. Answers

i) Some important hyperparameters involved in LightGBM are
   - maximum number of leaves in the tree
   - minimum number of data points in one leaf
   - maximum depth of the tree

ii) The histogram-based algorithms that LightGBM uses are the main reason why LightGBM is usually faster than XGBoost, which uses pre-sort-based algorithms.
iii) Exclusive feature bundling is used to deal with sparse feature spaces. Features that are sparse can be condensed into bundles.
iv) In multi-class classification, one-hot encoding is often used to convert cardinal categorical labels into numerical labels. For example,

	Category A	Category B	Category C
Feature 1	1	0	0
Feature 2	0	1	0
Feature 3	0	0	1

v) A naive way to bundle the above exclusive features would be

	Category A	Category B	Category C
Bundled Feature	1	2	3