MACHINE LEARNING - REGRESSION

Boosting Algorithm

- An ensemble modeling technique

Prepared by Mariappan

18 Dec, 2024

Presented to Hope AI, Coimbatore

Boosting?

Overview

- Boosting is a machine learning algorithm for primarily reducing bias, and also variance in supervised learning. and a family of machine learning algorithms that convert weak learners to strong ones.

- Improved Performance in Model.
- Ability to Handle Complex Data/Complicated Data Pattern.
- Robustness to Noise - Effectively reducing the impact of noisy samples in final prediction.
- Flexibility -Allowing for customization and adaptation to various problem domains.

Proponents

AdaBoost

Adaptive Boosting
Useful for classification

Gradient Boosting

Fitting new models to the residual errors of prior models.
XGBoost and LightGBM are popular.
Applicable for both regression and classification

How Boosting Algorithm works?

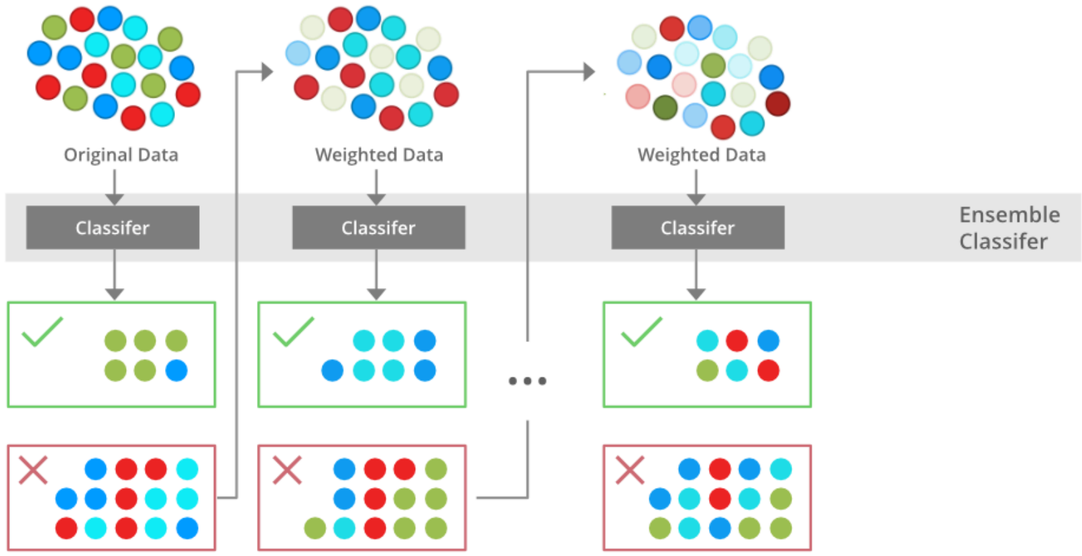
‘Boosting’ refers to a family of algorithms which converts weak learner to strong learners.

Step 1: Analyse and draw decision stupms (Algorithms).

Step 2: False prediction higher weightage.

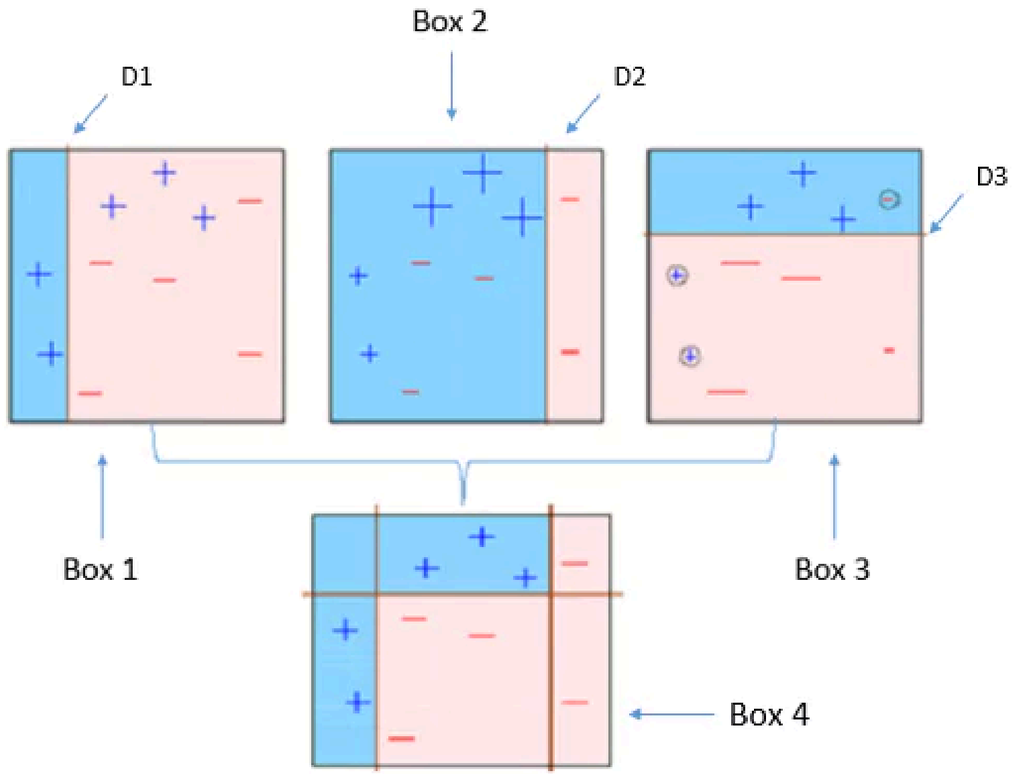
Step 3: Repeat Step 2 untill right prediction.

Boosting basically tries to reduce the bias error which arises when models are not able to identify relevant trends in the data. This happens by evaluating the difference between the predicted value and the actual value.



AdaBoost

Best out-of-the-box classifier



What it is ?

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

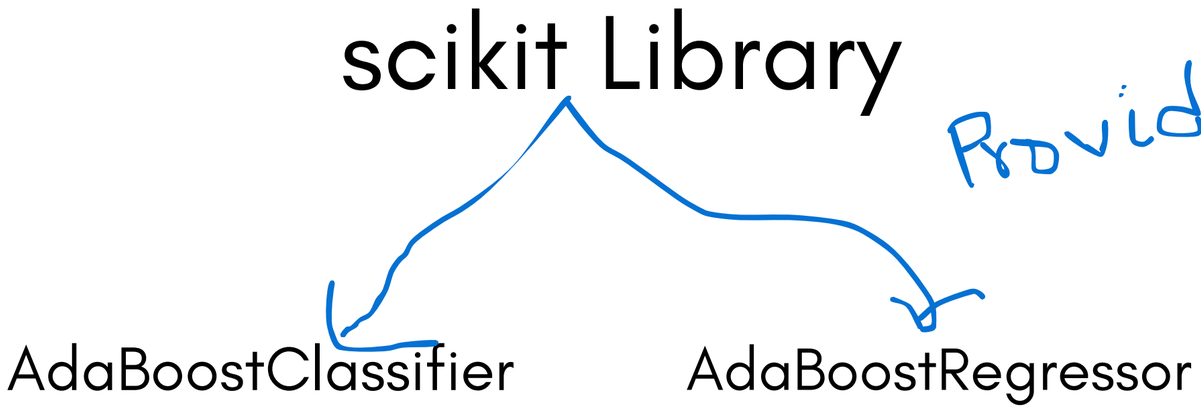
Relevance

AdaBoost is an ensemble learning method (also known as “meta-learning”) which was initially created to increase the efficiency of binary classifiers. AdaBoost uses an iterative approach to learn from the mistakes of weak classifiers, and turn them into strong ones.

Pseudocode of AdaBoost

1. Initially set uniform example weights.
2. for Each base learner do:
3. Train base learner with a weighted sample.
4. Test base learner on all data.
5. Set learner weight with a weighted error.
6. Set example weights based on ensemble predictions.
7. end for

It is used in Facial Recognition systems.



Pros:

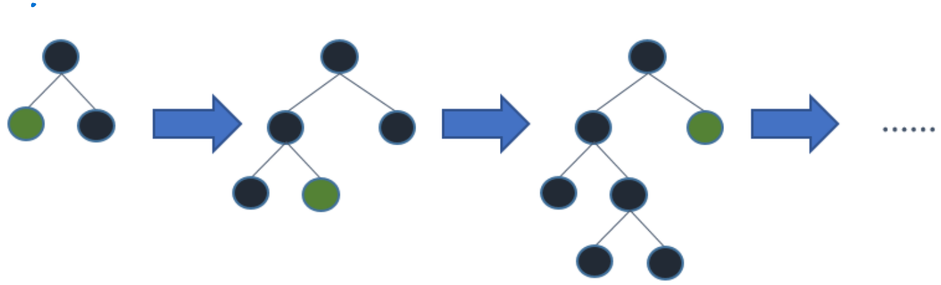
1. It is easier to use with less need for tweaking parameters
2. AdaBoost is not prone to overfitting
3. To improve the accuracy of your weak classifiers hence making it flexible
4. Found use cases in text and image classification as well.

Cons:

1. Boosting technique learns progressively, it is important to ensure that you have quality data.
2. AdaBoost is extremely sensitive to Noisy data and outliers. It is recommended to eliminate them.
3. AdaBoost has also been proven to be slower than XGBoost.

LightGBM- Light Gradient Boosting Machine

LightGBM uses histogram-based algorithms



Leaf-wise tree growth

What it is ?

LightGBM is an ensemble learning framework developed by Microsoft, a gradient boosting method, which constructs a strong learner by sequentially adding weak learners in a gradient descent manner. It optimizes memory usage and training time with techniques like Gradient-based One-Side Sampling (GOSS).

Relevance

Light GBM uses leaf wise splitting over depth-wise splitting which enables it to converge much faster but also leads to overfitting. Light GBM is almost 7 times faster than XGBOOST and is a much better approach when dealing with large datasets

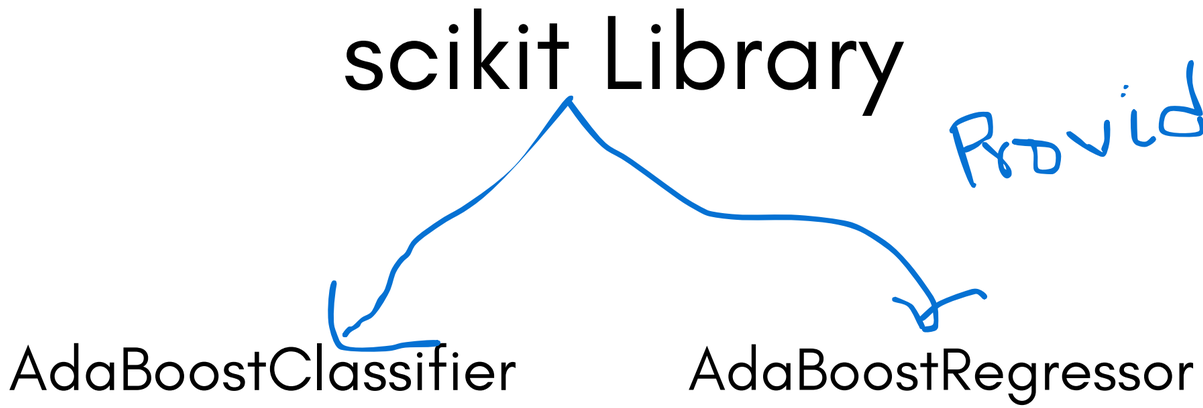
Characterstics

1. It splits the tree leaf wise with the best fit
2. It grows leaf-wise
3. It is surprisingly very fast, hence the word 'Light.

Tuning Parameters

1. For best fit: num_leaves,min_data_in_leaf &max_depth
2. For faster speed:bagging_fraction,feature_fraction & max_bin.
3. For better accuracy: num_leaves, max_bin.

It is used in finance, healthcare, marketing.



Pros:

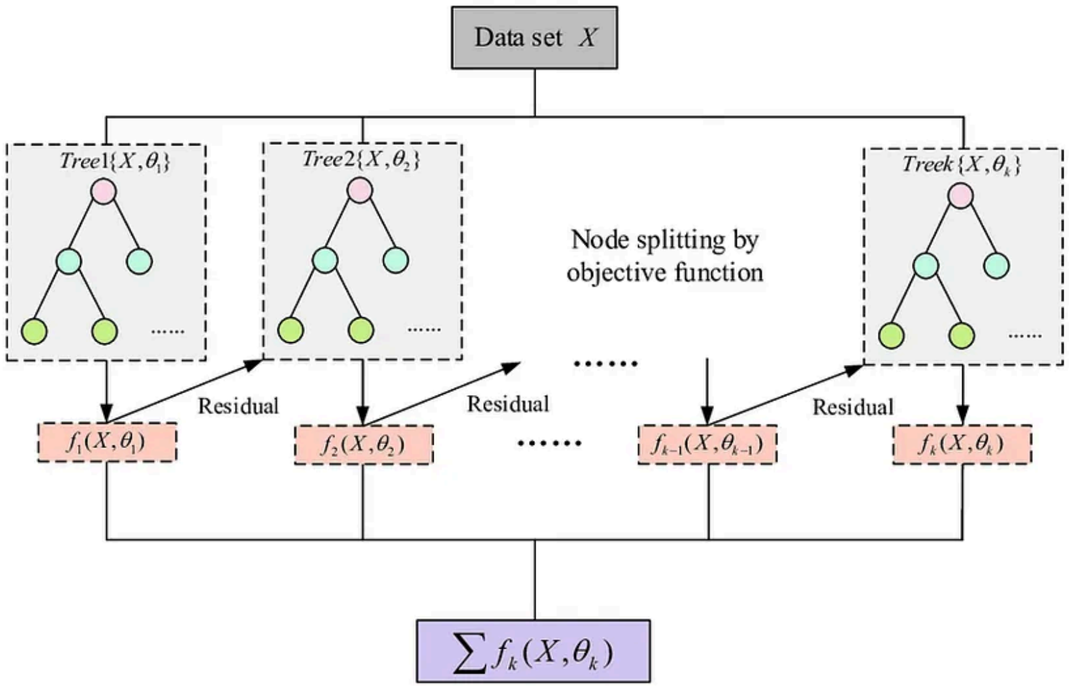
1. Faster Speed and Higher Accuracy.
2. Better Accuracy.
3. Support for Parallel and Distributed GPU Learning.
4. Capability to Handle Large-Scale Data (bigdata).

Cons:

1. Overfitting- LightGBM is its sensitivity to hyperparameters. selecting the optimal values can be challenging and may require extensive experimentation.
2. Compatibility with Datasets.
3. AdaBoost has also been proven to be slower than XGBoost.

XGBoost -Extreme Gradient Boosting

Sequential ensemble learning



What it is ?

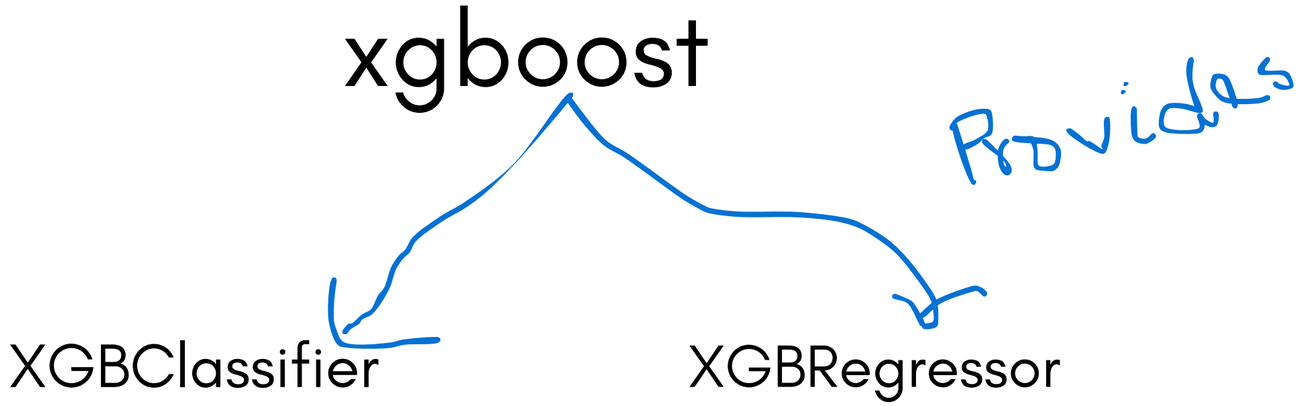
It is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

How it works?

XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed. With XGBoost, trees are built in parallel, instead of sequentially like GBDT

Usage

It is used in sales prediction,,Malware classification, Kaggle, Learning to rank



Pros:

1. It's efficient handling of missing values.
2. Built-in support for parallel processing.
3. XGBoost has a wide range of hyperparameters to optimize performance.
4. It suitable for large datasets.

Cons:

1. Computational Complexity in resource-constrained systems.
2. XGBoost will be prone to overfitting small datasets.
3. Finding the optimal set of parameters can be time-consuming and requires expertise.
4. XGBoost will be memory-intensive while handing large datasets.
5. Handling Image recognition and Natural Language Processing tasks.
6. Handling Unstructured Data.