XGBoost

XGBoost is an open-source software library that implements optimized distributed gradient boosting machine learning algorithms under the Gradient Boosting framework.

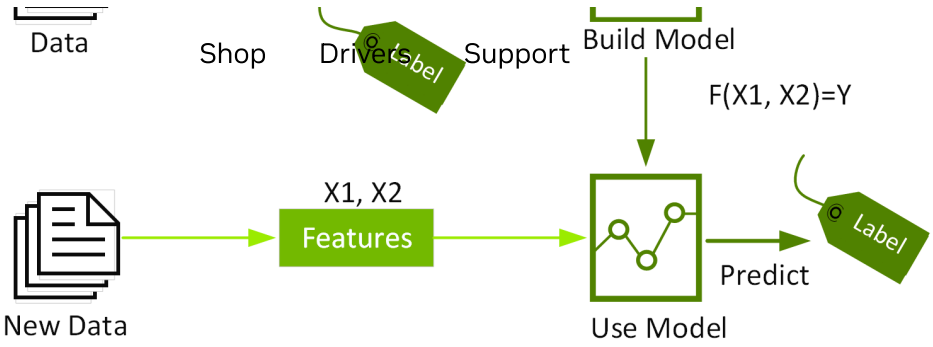
What is XGBoost?

XGBoost, which stands for Extreme Gradient Boosting, 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.

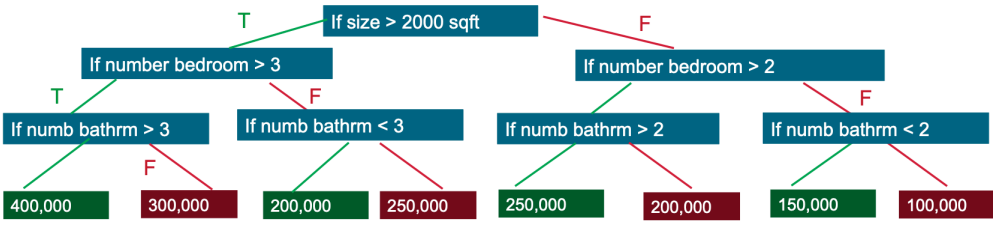
It's vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting.

Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset's features.

NVIDIA

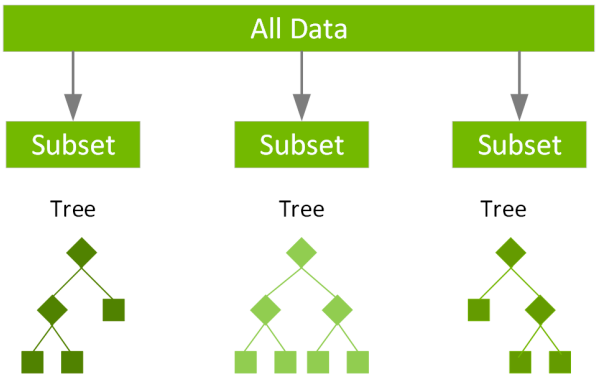


Decision trees create a model that predicts the label by evaluating a tree of if-then-else true/false feature questions, and estimating the minimum number of questions needed to assess the probability of making a correct decision. Decision trees can be used for classification to predict a category, or regression to predict a continuous numeric value. In the simple example below, a decision tree is used to estimate a house price (the label) based on the size and number of bedrooms (the features).



A Gradient Boosting Decision Trees (GBDT) is a decision tree ensemble learning algorithm similar to random forest, for classification and regression. Ensemble learning algorithms combine multiple machine learning algorithms to obtain a better model.

Both random forest and GBDT build a model consisting of multiple decision trees. The difference is in how the trees are built and combined.



The term “gradient boosting” comes from the idea of “boosting” or improving a single weak model by combining it with a number of other weak models in order to generate a collectively strong model. Gradient boosting is an extension of boosting where the process of additively generating weak models is formalized as a gradient descent algorithm over an objective function. Gradient boosting sets targeted outcomes for the next model in an effort to minimize errors. Targeted outcomes for each case are based on the gradient of the error (hence the name gradient boosting) with respect to the prediction.

GBDTs iteratively train an ensemble of shallow decision trees, with each iteration using the error residuals of the previous model to fit the next model. The final prediction is a weighted sum of all of the tree predictions. Random forest “bagging” minimizes the variance and overfitting, while GBDT “boosting” minimizes the bias and underfitting.

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. It follows a level-wise strategy, scanning across gradient values and using these partial sums to evaluate the quality of splits at every possible split in the training set.

Why XGBoost?

XGBoost gained significant favor in the last few years as a result of helping individuals and teams win virtually every Kaggle structured data competition. In these competitions, companies and researchers post data after which statisticians and data miners compete to produce the best models for predicting and describing the data.

Initially both Python and R implementations of XGBoost were built. Owing to its popularity, today XGBoost has package implementations for Java, Scala, Julia, Perl, and other languages. These implementations have opened the XGBoost library to even more developers and improved its appeal throughout the Kaggle community.

XGBoost has been integrated with a wide variety of other tools and packages such as scikit-learn for Python enthusiasts and caret for R users. In addition, XGBoost is

XGBoost Benefits and Attributes

The list of benefits and attributes of XGBoost is extensive, and includes the following:

- A large and growing list of data scientists globally that are actively contributing to XGBoost open source development
- Usage on a wide range of applications, including solving problems in regression, classification, ranking, and user-defined prediction challenges
- A library that's highly portable and currently runs on OS X, Windows, and Linux platforms
- Cloud integration that supports AWS, Azure, Yarn clusters, and other ecosystems
- Active production use in multiple organizations across various vertical market areas
- A library that was built from the ground up to be efficient, flexible, and portable

XGBoost and Data Scientists

It's noteworthy for data scientists that XGBoost and XGBoost machine learning models have the premier combination of prediction performance and processing time compared with other algorithms. This has been borne out by various benchmarking studies and further explains its appeal to data scientists.

How XGBoost Runs Better with GPUs

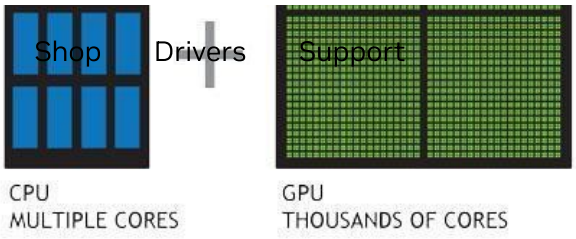
CPU-powered machine learning tasks with XGBoost can literally take hours to run. That's because creating highly accurate, state-of-the-art prediction results involves the creation of thousands of decision trees and the testing of large numbers of parameter combinations. Graphics processing units, or GPUs, with their massively parallel architecture consisting of thousands of small efficient cores, can launch thousands of parallel threads simultaneously to supercharge compute-intensive tasks.

NVIDIA

US

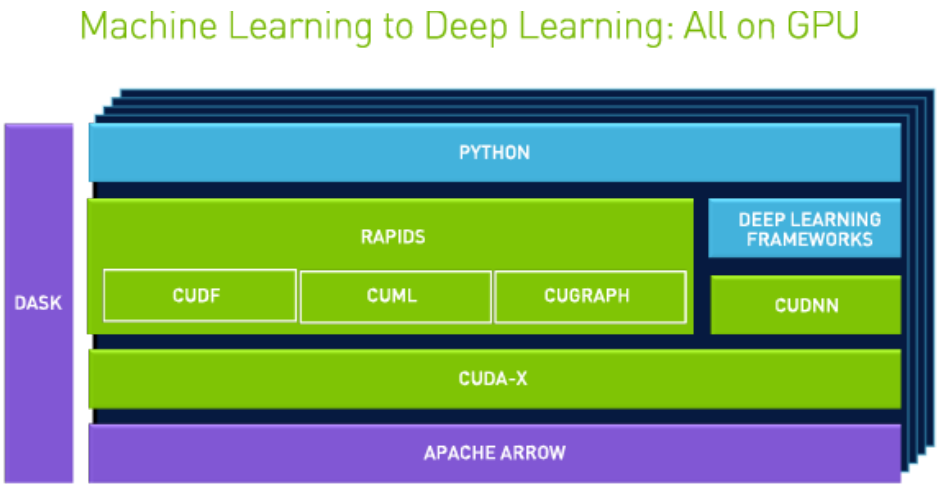
NVIDIA

US



NVIDIA developed NVIDIA RAPIDS™—an open-source data analytics and machine learning acceleration platform—or executing end-to-end data science training pipelines completely in GPUs. It relies on NVIDIA CUDA® primitives for low-level compute optimization, but exposes that GPU parallelism and high memory bandwidth through user-friendly Python interfaces.

Focusing on common data preparation tasks for analytics and data science, RAPIDS offers a familiar DataFrame API that integrates with scikit-learn and a variety of machine learning algorithms without paying typical serialization costs. This allows acceleration for end-to-end pipelines—from data prep to machine learning to deep learning. RAPIDS also includes support for multi-node, multi-GPU deployments, enabling vastly accelerated processing and training on much larger dataset sizes.



XGBoost + RAPIDS

The RAPIDS team works closely with the Distributed Machine Learning Common (DMLC) XGBoost organization, and XGBoost now includes seamless, drop-in GPU acceleration. This significantly speeds up model training and improves accuracy for better predictions.

NVIDIA US

with XGBoost, so you can share a single, high-speed memory pool.

GPU-Accelerated XGBoost

The GPU-accelerated XGBoost algorithm makes use of fast parallel prefix sum operations to scan through all possible splits, as well as parallel radix sorting to repartition data. It builds a decision tree for a given boosting iteration, one level at a time, processing the entire dataset concurrently on the GPU.

GPU-Accelerated, End-to-End Data Pipelines with Spark + XGBoost

NVIDIA understands that machine learning at scale delivers powerful predictive capabilities for data scientists and developers and, ultimately, to end users. But this at-scale learning depends upon overcoming key challenges to both on-premises and cloud infrastructure, like speeding up pre-processing of massive data volumes and then accelerating compute-intensive model training.

NVIDIA's initial release of spark-xgboost enabled training and inferencing of XGBoost machine learning models across Apache Spark nodes. This has helped make it a leading mechanism for enterprise-class distributed machine learning.

GPU-Accelerated Spark XGBoost speeds up pre-processing of massive volumes of data, allows larger data sizes in GPU memory, and improves XGBoost training and tuning time.

Next Steps

- Find out more about:
  - Machine Learning
  - RAPIDS
  - Apache Spark
- Check out our free ebook all about Spark 3
- Learn more about the RAPIDS Accelerator for Apache Spark

About Us

Company Overview

Investors

Venture Capital (NVentures)

NVIDIA Foundation

Research

Corporate Sustainability

Technologies

Careers

Newsroom

Company Blog

Technical Blog

Webinars

Stay Informed

Events Calendar

GTC AI Conference

NVIDIA On-Demand

Popular Links

Developers

Partners

Executive Insights

Startups and VCs

NVIDIA Connect for ISVs

Documentation

Technical Training

Professional Services for Data Science