KDnuggets

Subscribe to KDnuggets News | | Contact

search KDnuggets Search

- SOFTWARE
- News/Blog
- Top stories
- Opinions
- Tutorials
- JOBS
- Companies
- Courses
- Datasets
- EDUCATION
- Certificates
- Meetings
- Webinars

KDnuggets Home » News » 2017 » Oct » Tutorials, Overviews » XGBoost, a Top Machine Learning Method on Kaggle, Explained ( 17:n38 )

# XGBoost, a Top Machine Learning Method on Kaggle, Explained

Oct 2017 Silver KDnuggets Blog

<= Previous post
Next post =>

http likes 232
Like 114 Share 114 Tweet Share 105

Tags: Algorithms, Data Science, Explained, Kaggle, Machine Learning

Looking to boost your machine learning competitions score? Here’s a brief summary and introduction to a powerful and popular tool among Kagglers, XGBoost.

---

**By Ilan Reinstein, KDnuggets.**

eXtreme Gradient Boosting
custom tree building algorithm
XGBoost
by Tianqi Chen
Used for:
• classification
• regression
• ranking
with custom loss functions
Interfaces for Python and R, can be executed on YARN

*What is XGBoost?*

XGBoost has become a widely used and really popular tool among Kaggle competitors and Data Scientists in industry, as it has been battle tested for production on large-scale problems. It is a highly flexible and versatile tool that can work through most regression, classification and ranking problems as well as user-built objective functions. As an open-source software, it is easily accessible and it may be used through different platforms and interfaces. The amazing portability and compatibility of the system permits its usage on all three Windows, Linux and OS X. It also supports training on distributed cloud platforms like AWS, Azure, GCE among others and it is easily connected to large-scale cloud dataflow systems such as Flink and Spark. Although it was built and initially used in the Command Line Interface (CLI) by its creator (Tianqi Chen), it can also be loaded and used in various languages and interfaces such as Python, C++, R, Julia, Scala and Java.

Its name stands for **eXtreme Gradient Boosting**, it was developed by Tianqi Chen and now is part of a wider collection of open-source libraries developed by the Distributed Machine Learning Community (DMLC). XGBoost is a scalable and accurate implementation of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed. Specifically, it was engineered to exploit every bit of memory and hardware resources for tree boosting algorithms.

The implementation of XGBoost offers several advanced features for model tuning, computing environments and algorithm enhancement. It is capable of performing the three main forms of gradient boosting (Gradient Boosting (GB), Stochastic GB and Regularized GB) and it is robust enough to support fine tuning and addition of regularization parameters. According to Tianqi Chen, the latter is what makes it superior and different to other libraries.

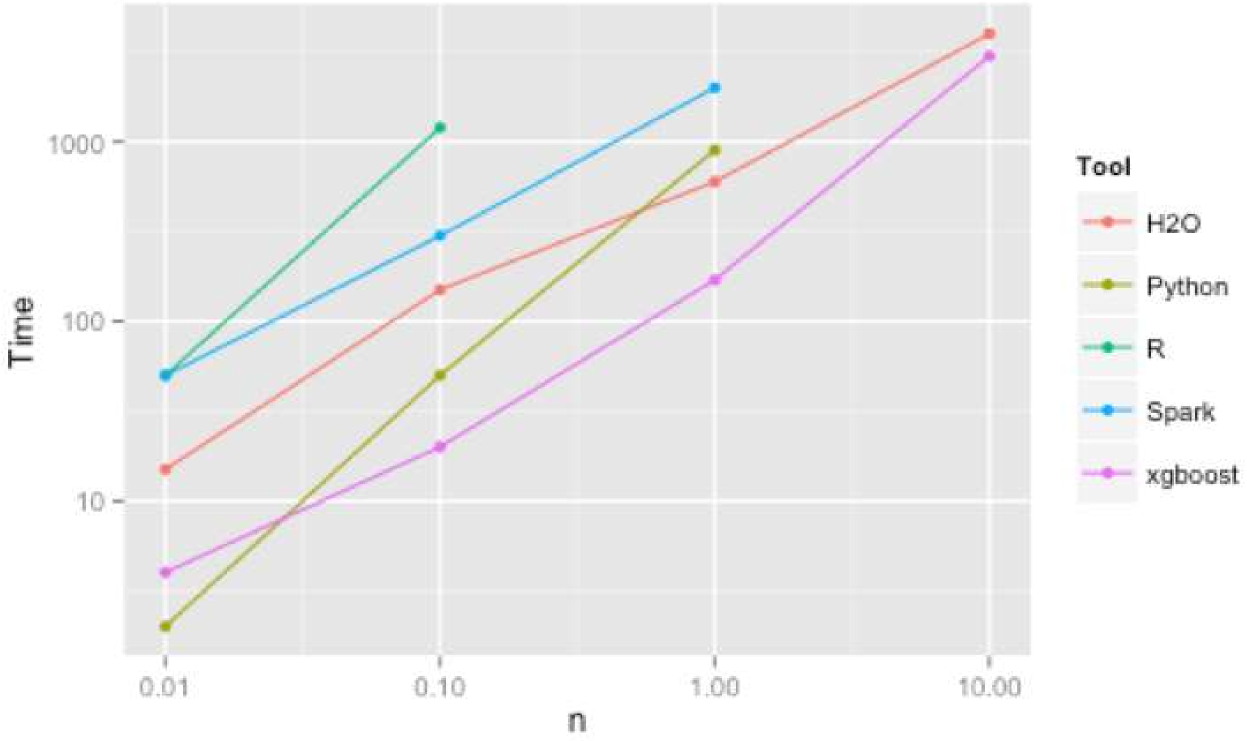
> “…xgboost used a more regularized model formalization to control over-fitting, which gives it better performance.”- Tianqi Chen on Quora

System-wise, the library’s portability and flexibility allow the use of a wide variety of computing environments like parallelization for tree construction across several CPU cores; distributed computing for large models; Out-of-Core computing; and Cache Optimization to improve hardware usage and efficiency.

The algorithm was developed to efficiently reduce computing time and allocate an optimal usage of memory resources. Important features of implementation include handling of missing values (Sparse Aware), Block Structure to support parallelization in tree construction and the ability to fit and boost on new data added to a trained model (Continued Training).

*Why use XGBoost?*

As we already mentioned, the key features of this library rely on *model performance and execution speed*. A well-structured clear benchmark done by Szilard Pafka, shows how XGBoost outperforms several other well-known implementations of gradient tree boosting.



This comparison in Figure 1 helps us grasp the power of the tool and see how well balanced its benefits are, i.e., it does not seem to sacrifice speed over accuracy or vice versa. It starts to become clear why more Kagglers are using it every day, it is a semi-perfect equilibrium of both performance and time-efficiency.

*How does it work?*

Before moving on to the details of the algorithm, let's set some basic definitions to make our life easier and get an intuitive and complete understanding of this popular tool.

First, let's clarify the concept of boosting. This is an ensemble method that seeks to create a strong classifier (model) based on "weak" classifiers. In this context, weak and strong refer to a measure of how correlated are the learners to the actual target variable. By adding models on top of each other iteratively, the errors of the previous model are corrected by the next predictor, until the training data is accurately predicted or reproduced by the model. If you want to dig into boosting a bit more, check out information about a popular implementation called AdaBoost (Adaptive Boosting) here.

Now, gradient boosting also comprises an ensemble method that sequentially adds predictors and corrects previous models. However, instead of assigning different weights to the classifiers after every iteration, this method fits the new model to new residuals of the previous prediction and then minimizes the loss when adding the latest prediction. So, in the end, you are updating your model using gradient descent and hence the name, gradient boosting. This is supported for both regression and classification problems. XGBoost specifically, implements this algorithm for decision tree boosting with an additional custom regularization term in the objective function.

*Getting started with XGBoost*

You may download and install XGBoost regardless of which interface you are using. To learn more on how to use on each specific platform please follow the instructions on this link. You will also find official documentation and tutorials here.

For further information on the source code and examples, you may visit this DMLC repository on Github.

For more information on boosting and gradient boosting the following resources might be helpful:

- The official published paper by Tianqi Chen is available for download from Arxiv
- The official documentation XGBoost page
- Here is a great presentation that summarizes the math in a very intuitive way
- Some Wikipedia Articles give a good general idea of the history and the math behind the algorithms:

Thanks to Jason Brownlee for the inspiration of this post, more resources on Boosting and XGBoost are available on his post.

**Related:**

- Lessons Learned From Benchmarking Fast Machine Learning Algorithms
- A Simple XGBoost Tutorial Using the Iris Dataset
- XGBoost: Implementing the Winningest Kaggle Algorithm in Spark and Flink

**<= Previous post**
**Next post =>**

## Top Stories Past 30 Days

**Most Popular**

1. I wasn't getting hired as a Data Scientist. So I sought data on who is.
2. 10 Great Python Resources for Aspiring Data Scientists
3. Which Data Science Skills are core and which are hot/emerging ones?
4. Advice on building a machine learning career and reading research papers by Prof. Andrew Ng
5. Python Libraries for Interpretable Machine Learning
6. Object-oriented programming for data scientists: Build your ML estimator
7. TensorFlow vs PyTorch vs Keras for NLP

**Most Shared**

1. Python Libraries for Interpretable Machine Learning
2. 10 Great Python Resources for Aspiring Data Scientists
3. I wasn't getting hired as a Data Scientist. So I sought data on who is.
4. TensorFlow vs PyTorch vs Keras for NLP
5. Which Data Science Skills are core and which are hot/emerging ones?
6. The 5 Graph Algorithms That Data Scientists Should Know
7. My journey path from a Software Engineer to BI Specialist to a Data Scientist

## Latest News

- Overcoming Deep Learning Stumbling Blocks
- The Last SQL Guide for Data Analysis You'll Ever ...
- Research Guide for Neural Architecture Search
- 6 Must See Deep Learning Experts at ODSC West 2019 R...
- 5 Fundamental AI Principles
- Recreating Imagination: DeepMind Builds Neural Networks...

Open for Innovation
KNIME
KNIME Fall Summit 2019
November 5 - 8 in Austin, TX
Use promo code KDNUGGETS for 10% off tickets

**Top Stories Last Week**

## Most Popular

1. The Future of Analytics and Data Science

2. **Which Data Science Skills are core and which are hot/emerging ones?**
3. **6 bits of advice for Data Scientists**
4. **Help Your Career Survive DataGeddon**
5. **Automatic Version Control for Data Scientists**
6. **10 Great Python Resources for Aspiring Data Scientists**
7. **5 Famous Deep Learning Courses/Schools of 2019**

## Most Shared

1. **5 Famous Deep Learning Courses/Schools of 2019**
2. **12 Deep Learning Researchers and Leaders**
3. **Natural Language in Python using spaCy: An Introduction**
4. **A Single Function to Streamline Image Classification with Keras**
5. **What is Hierarchical Clustering?**
6. **The Future of Analytics and Data Science**
7. **6 bits of advice for Data Scientists**

KDnuggets Home » News » 2017 » Oct » Tutorials, Overviews » XGBoost, a Top Machine Learning Method on Kaggle, Explained ( 17:n38 )

**Subscribe to KDnuggets News**

X