
xgboost

Release 0.80

xgboost developers

Sep 28, 2018

Contents

1	Contents	3
1.1	Installation Guide	3
1.2	Get Started with XGBoost	10
1.3	XGBoost Tutorials	12
1.4	Frequently Asked Questions	30
1.5	XGBoost GPU Support	32
1.6	XGBoost Parameters	34
1.7	XGBoost Python Package	42
1.8	XGBoost R Package	64
1.9	XGBoost JVM Package	81
1.10	XGBoost.jl	96
1.11	XGBoost Command Line version	96
1.12	Contribute to XGBoost	96
	Python Module Index	101

XGBoost is an optimized distributed gradient boosting library designed to be highly **efficient**, **flexible** and **portable**. It implements machine learning algorithms under the **Gradient Boosting** framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.

1.1 Installation Guide

Note: Pre-built binary wheel for Python

If you are planning to use Python, consider installing XGBoost from a pre-built binary wheel, available from Python Package Index (PyPI). You may download and install it by running

```
# Ensure that you are downloading one of the following:
# * xgboost-{version}-py2.py3-none-manylinux1_x86_64.whl
# * xgboost-{version}-py2.py3-none-win_amd64.whl
pip3 install xgboost
```

- The binary wheel will support GPU algorithms (*gpu_exact*, *gpu_hist*) on machines with NVIDIA GPUs. **However, it will not support multi-GPU training; only single GPU will be used.** To enable multi-GPU training, download and install the binary wheel from [this page](#).
- Currently, we provide binary wheels for 64-bit Linux and Windows.

1.1.1 Building XGBoost from source

This page gives instructions on how to build and install XGBoost from scratch on various systems. It consists of two steps:

1. First build the shared library from the C++ codes (*libxgboost.so* for Linux/OSX and *xgboost.dll* for Windows). (For R-package installation, please directly refer to [R Package Installation](#).)
2. Then install the language packages (e.g. Python Package).

Note: Use of Git submodules

XGBoost uses Git submodules to manage dependencies. So when you clone the repo, remember to specify `--recursive` option:

```
git clone --recursive https://github.com/dmlc/xgboost
```

For windows users who use github tools, you can open the git shell and type the following command:

```
git submodule init
git submodule update
```

Please refer to *Trouble Shooting* section first if you have any problem during installation. If the instructions do not work for you, please feel free to ask questions at [the user forum](#).

Contents

- *Building the Shared Library*
 - *Building on Ubuntu/Debian*
 - *Building on OSX*
 - *Building on Windows*
 - *Building with GPU support*
 - *Customized Building*
- *Python Package Installation*
- *R Package Installation*
- *Trouble Shooting*

1.1.2 Building the Shared Library

Our goal is to build the shared library:

- On Linux/OSX the target library is `libxgboost.so`
- On Windows the target library is `xgboost.dll`

The minimal building requirement is

- A recent C++ compiler supporting C++11 (g++-4.8 or higher)

We can edit `make/config.mk` to change the compile options, and then build by `make`. If everything goes well, we can go to the specific language installation section.

Building on Ubuntu/Debian

On Ubuntu, one builds XGBoost by running

```
git clone --recursive https://github.com/dmlc/xgboost
cd xgboost; make -j4
```


Building on OSX

Install with pip: simple method

First, make sure you obtained `gcc-5` (newer version does not work with this method yet). Note: installation of `gcc` can take a while (~ 30 minutes).

```
brew install gcc@5
```

Then install XGBoost with `pip`:

```
pip3 install xgboost
```

You might need to run the command with `sudo` if you run into permission errors.

Build from the source code - advanced method

First, obtain `gcc-7` with homebrew (<https://brew.sh/>) if you want multi-threaded version. Clang is okay if multi-threading is not required. Note: installation of `gcc` can take a while (~ 30 minutes).

```
brew install gcc@7
```

Now, clone the repository:

```
git clone --recursive https://github.com/dmlc/xgboost
cd xgboost; cp make/config.mk ./config.mk
```

Open `config.mk` and uncomment these two lines:

```
export CC = gcc
export CXX = g++
```

and replace these two lines as follows: (specify the GCC version)

```
export CC = gcc-7
export CXX = g++-7
```

Now, you may build XGBoost using the following command:

```
make -j4
```

You may now continue to [Python Package Installation](#).

Building on Windows

You need to first clone the XGBoost repo with `--recursive` option, to clone the submodules. We recommend you use [Git for Windows](#), as it comes with a standard Bash shell. This will highly ease the installation process.

```
git submodule init
git submodule update
```

XGBoost support compilation with Microsoft Visual Studio and MinGW.

Compile XGBoost using MinGW

After installing [Git for Windows](#), you should have a shortcut named `Git Bash`. You should run all subsequent steps in `Git Bash`.

In MinGW, `make` command comes with the name `mingw32-make`. You can add the following line into the `.bashrc` file:

```
alias make='mingw32-make'
```

(On 64-bit Windows, you should get [MinGW64](#) instead.) Make sure that the path to MinGW is in the system `PATH`.

To build with MinGW, type:

```
cp make/mingw64.mk config.mk; make -j4
```

Compile XGBoost with Microsoft Visual Studio

To build with Visual Studio, we will need CMake. Make sure to install a recent version of CMake. Then run the following from the root of the XGBoost directory:

```
mkdir build
cd build
cmake .. -G"Visual Studio 12 2013 Win64"
```

This specifies an out of source build using the MSVC 12 64 bit generator. Open the `.sln` file in the build directory and build with Visual Studio. To use the Python module you can copy `xgboost.dll` into `python-package/xgboost`.

After the build process successfully ends, you will find a `xgboost.dll` library file inside `./lib/` folder, copy this file to the the API package folder like `python-package/xgboost` if you are using Python API.

Unofficial windows binaries and instructions on how to use them are hosted on [Guido Tapia's blog](#).

Building with GPU support

XGBoost can be built with GPU support for both Linux and Windows using CMake. GPU support works with the Python package as well as the CLI version. See [Installing R package with GPU support](#) for special instructions for R.

An up-to-date version of the CUDA toolkit is required.

From the command line on Linux starting from the XGBoost directory:

```
mkdir build
cd build
cmake .. -DUSE_CUDA=ON
make -j
```

Note: Enabling multi-GPU training

By default, multi-GPU training is disabled and only a single GPU will be used. To enable multi-GPU training, set the option `USE_NCCL=ON`. Multi-GPU training depends on NCCL2, available at [this link](#). Since NCCL2 is only available for Linux machines, **multi-GPU training is available only for Linux**.

```
mkdir build
cd build
cmake .. -DUSE_CUDA=ON -DUSE_NCCL=ON
make -j
```

On Windows, see what options for generators you have for CMake, and choose one with [arch] replaced with Win64:

```
cmake -help
```

Then run CMake as follows:

```
mkdir build
cd build
cmake .. -G"Visual Studio 14 2015 Win64" -DUSE_CUDA=ON
```

Note: Visual Studio 2017 Win64 Generator may not work

Choosing the Visual Studio 2017 generator may cause compilation failure. When it happens, specify the 2015 compiler by adding the `-T` option:

```
make .. -G"Visual Studio 15 2017 Win64" -T v140,cuda=8.0 -DR_LIB=ON -DUSE_CUDA=ON
```

To speed up compilation, the compute version specific to your GPU could be passed to cmake as, e.g., `-DGPU_COMPUTE_VER=50`. The above cmake configuration run will create an `xgboost.sln` solution file in the build directory. Build this solution in release mode as a x64 build, either from Visual studio or from command line:

```
cmake --build . --target xgboost --config Release
```

To speed up compilation, run multiple jobs in parallel by appending option `-- /MP`.

Customized Building

The configuration file `config.mk` modifies several compilation flags: - Whether to enable support for various distributed filesystems such as HDFS and Amazon S3 - Which compiler to use - And some more

To customize, first copy `make/config.mk` to the project root and then modify the copy.

Python Package Installation

The python package is located at `python-package/`. There are several ways to install the package:

1. Install system-wide, which requires root permission:

```
cd python-package; sudo python setup.py install
```

You will however need Python `distutils` module for this to work. It is often part of the core python package or it can be installed using your package manager, e.g. in Debian use

```
sudo apt-get install python-setuptools
```

Note: Re-compiling XGBoost

If you recompiled XGBoost, then you need to reinstall it again to make the new library take effect.

2. Only set the environment variable `PYTHONPATH` to tell python where to find the library. For example, assume we cloned *xgboost* on the home directory `~`. then we can added the following line in `~/.bashrc`. This option is **recommended for developers** who change the code frequently. The changes will be immediately reflected once you pulled the code and rebuild the project (no need to call `setup` again)

```
export PYTHONPATH=~/.xgboost/python-package
```

3. Install only for the current user.

```
cd python-package; python setup.py develop --user
```

4. If you are installing the latest XGBoost version which requires compilation, add MinGW to the system PATH:

```
import os
os.environ['PATH'] = os.environ['PATH'] + ';C:\\Program Files\\mingw-w64\\x86_64-5.3.
↳0-posix-seh-rt_v4-rev0\\mingw64\\bin'
```

R Package Installation

Installing pre-packaged version

You can install *xgboost* from CRAN just like any other R package:

```
install.packages("xgboost")
```

Or you can install it from our weekly updated drat repo:

```
install.packages("drat", repos="https://cran.rstudio.com")
drat::addRepo("dmlc")
install.packages("xgboost", repos="http://dmlc.ml/drat/", type = "source")
```

For OSX users, single threaded version will be installed. To install multi-threaded version, first follow [Building on OSX](#) to get the OpenMP enabled compiler. Then

- Set the `Makevars` file in highest priority for R.

The point is, there are three `Makevars` : `~/.R/Makevars`, `xgboost/R-package/src/Makevars`, and `/usr/local/Cellar/r/3.2.0/R.framework/Resources/etc/Makeconf` (the last one obtained by running `file.path(R.home("etc"), "Makeconf")` in R), and `SHLIB_OPENMP_CXXFLAGS` is not set by default!! After trying, it seems that the first one has highest priority (surprise!).

Then inside R, run

```
install.packages("drat", repos="https://cran.rstudio.com")
drat::addRepo("dmlc")
install.packages("xgboost", repos="http://dmlc.ml/drat/", type = "source")
```

Installing the development version

Make sure you have installed git and a recent C++ compiler supporting C++11 (e.g., g++-4.8 or higher). On Windows, Rtools must be installed, and its bin directory has to be added to PATH during the installation. And see the previous subsection for an OSX tip.

Due to the use of git-submodules, `devtools::install_github` can no longer be used to install the latest version of R package. Thus, one has to run git to check out the code first:

```
git clone --recursive https://github.com/dmlc/xgboost
cd xgboost
git submodule init
git submodule update
cd R-package
R CMD INSTALL .
```

If the last line fails because of the error `R: command not found`, it means that R was not set up to run from command line. In this case, just start R as you would normally do and run the following:

```
setwd('wherever/you/cloned/it/xgboost/R-package/')
install.packages('.', repos = NULL, type="source")
```

The package could also be built and installed with cmake (and Visual C++ 2015 on Windows) using instructions from the next section, but without GPU support (omit the `-DUSE_CUDA=ON` cmake parameter).

If all fails, try *Building the shared library* to see whether a problem is specific to R package or not.

Installing R package with GPU support

The procedure and requirements are similar as in *Building with GPU support*, so make sure to read it first.

On Linux, starting from the XGBoost directory type:

```
mkdir build
cd build
cmake .. -DUSE_CUDA=ON -DR_LIB=ON
make install -j
```

When default target is used, an R package shared library would be built in the build area. The `install` target, in addition, assembles the package files with this shared library under `build/R-package`, and runs `R CMD INSTALL`.

On Windows, cmake with Visual C++ Build Tools (or Visual Studio) has to be used to build an R package with GPU support. Rtools must also be installed (perhaps, some other MinGW distributions with `gendef.exe` and `dlltool.exe` would work, but that was not tested).

```
mkdir build
cd build
cmake .. -G"Visual Studio 14 2015 Win64" -DUSE_CUDA=ON -DR_LIB=ON
cmake --build . --target install --config Release
```

When `--target xgboost` is used, an R package dll would be built under `build/Release`. The `--target install`, in addition, assembles the package files with this dll under `build/R-package`, and runs `R CMD INSTALL`.

If cmake can't find your R during the configuration step, you might provide the location of its executable to cmake like this: `-DLIBR_EXECUTABLE="C:/Program Files/R/R-3.4.1/bin/x64/R.exe"`.

If on Windows you get a “permission denied” error when trying to write to ...Program Files/R/... during the package installation, create a `.Rprofile` file in your personal home directory (if you don’t already have one in there), and add a line to it which specifies the location of your R packages user library, like the following:

```
.libPaths( unique(c("C:/Users/USERNAME/Documents/R/win-library/3.4", .libPaths())) )
```

You might find the exact location by running `.libPaths()` in R GUI or RStudio.

Trouble Shooting

1. Compile failed after `git pull`

Please first update the submodules, clean all and recompile:

```
git submodule update && make clean_all && make -j4
```

2. Compile failed after `config.mk` is modified

Need to clean all first:

```
make clean_all && make -j4
```

3. Makefile: `dmlc-core/make/dmlc.mk`: No such file or directory

We need to recursively clone the submodule:

```
git submodule init
git submodule update
```

Alternatively, do another clone

```
git clone https://github.com/dmlc/xgboost --recursive
```

1.2 Get Started with XGBoost

This is a quick start tutorial showing snippets for you to quickly try out XGBoost on the demo dataset on a binary classification task.

1.2.1 Links to Other Helpful Resources

- See [Installation Guide](#) on how to install XGBoost.
- See [Text Input Format](#) on using text format for specifying training/testing data.
- See [Tutorials](#) for tips and tutorials.
- See [Learning to use XGBoost by Examples](#) for more code examples.

1.2.2 Python

```
import xgboost as xgb
# read in data
dtrain = xgb.DMatrix('demo/data/agaricus.txt.train')
dtest = xgb.DMatrix('demo/data/agaricus.txt.test')
# specify parameters via map
param = {'max_depth':2, 'eta':1, 'silent':1, 'objective':'binary:logistic' }
num_round = 2
bst = xgb.train(param, dtrain, num_round)
# make prediction
preds = bst.predict(dtest)
```

1.2.3 R

```
# load data
data(agaricus.train, package='xgboost')
data(agaricus.test, package='xgboost')
train <- agaricus.train
test <- agaricus.test
# fit model
bst <- xgboost(data = train$data, label = train$label, max.depth = 2, eta = 1,
  ↪nrounds = 2,
                nthread = 2, objective = "binary:logistic")
# predict
pred <- predict(bst, test$data)
```

1.2.4 Julia

```
using XGBoost
# read data
train_X, train_Y = readlibsvm("demo/data/agaricus.txt.train", (6513, 126))
test_X, test_Y = readlibsvm("demo/data/agaricus.txt.test", (1611, 126))
# fit model
num_round = 2
bst = xgboost(train_X, num_round, label=train_Y, eta=1, max_depth=2)
# predict
pred = predict(bst, test_X)
```

1.2.5 Scala

```
import ml.dmlc.xgboost4j.scala.DMatrix
import ml.dmlc.xgboost4j.scala.XGBoost

object XGBoostScalaExample {
  def main(args: Array[String]) {
    // read training data, available at xgboost/demo/data
    val trainData =
      new DMatrix("/path/to/agaricus.txt.train")
    // define parameters
    val paramMap = List(
      "eta" -> 0.1,
      "max_depth" -> 2,
```

(continues on next page)

(continued from previous page)

```

    "objective" -> "binary:logistic").toMap
  // number of iterations
  val round = 2
  // train the model
  val model = XGBoost.train(trainData, paramMap, round)
  // run prediction
  val predTrain = model.predict(trainData)
  // save model to the file.
  model.saveModel("/local/path/to/model")
}
}

```

1.3 XGBoost Tutorials

This section contains official tutorials inside XGBoost package. See [Awesome XGBoost](#) for more resources.

1.3.1 Introduction to Boosted Trees

XGBoost stands for “Extreme Gradient Boosting”, where the term “Gradient Boosting” originates from the paper *Greedy Function Approximation: A Gradient Boosting Machine*, by Friedman. This is a tutorial on gradient boosted trees, and most of the content is based on [these slides](#) by Tianqi Chen, the original author of XGBoost.

The **gradient boosted trees** has been around for a while, and there are a lot of materials on the topic. This tutorial will explain boosted trees in a self-contained and principled way using the elements of supervised learning. We think this explanation is cleaner, more formal, and motivates the model formulation used in XGBoost.

Elements of Supervised Learning

XGBoost is used for supervised learning problems, where we use the training data (with multiple features) x_i to predict a target variable y_i . Before we learn about trees specifically, let us start by reviewing the basic elements in supervised learning.

Model and Parameters

The **model** in supervised learning usually refers to the mathematical structure of by which the prediction y_i is made from the input x_i . A common example is a *linear model*, where the prediction is given as $\hat{y}_i = \sum_j \theta_j x_{ij}$, a linear combination of weighted input features. The prediction value can have different interpretations, depending on the task, i.e., regression or classification. For example, it can be logistic transformed to get the probability of positive class in logistic regression, and it can also be used as a ranking score when we want to rank the outputs.

The **parameters** are the undetermined part that we need to learn from data. In linear regression problems, the parameters are the coefficients θ . Usually we will use θ to denote the parameters (there are many parameters in a model, our definition here is sloppy).

Objective Function: Training Loss + Regularization

With judicious choices for y_i , we may express a variety of tasks, such as regression, classification, and ranking. The task of **training** the model amounts to finding the best parameters θ that best fit the training data x_i and labels y_i . In order to train the model, we need to define the **objective function** to measure how well the model fit the training data.

A salient characteristic of objective functions is that they consist two parts: **training loss** and **regularization term**:

$$\text{obj}(\theta) = L(\theta) + \Omega(\theta)$$

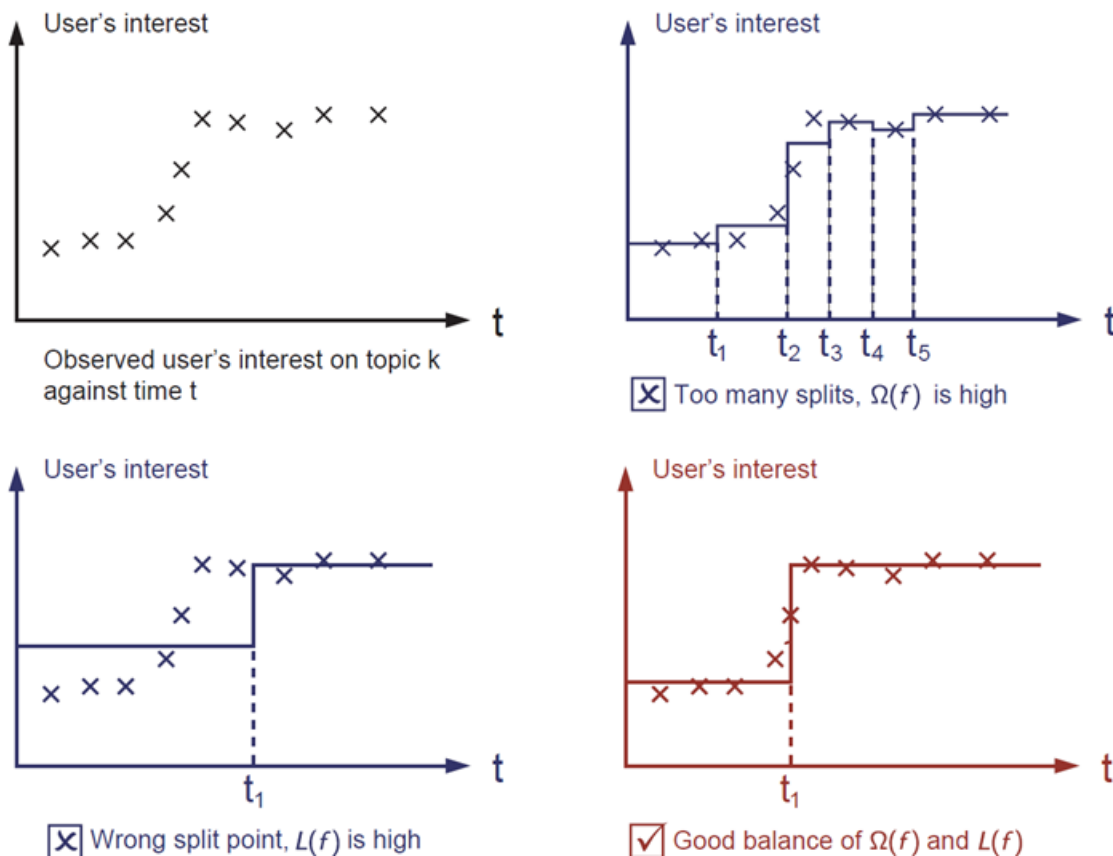
where L is the training loss function, and Ω is the regularization term. The training loss measures how *predictive* our model is with respect to the training data. A common choice of L is the *mean squared error*, which is given by

$$L(\theta) = \sum_i (y_i - \hat{y}_i)^2$$

Another commonly used loss function is logistic loss, to be used for logistic regression:

$$L(\theta) = \sum_i [y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i})]$$

The **regularization term** is what people usually forget to add. The regularization term controls the complexity of the model, which helps us to avoid overfitting. This sounds a bit abstract, so let us consider the following problem in the following picture. You are asked to *fit* visually a step function given the input data points on the upper left corner of the image. Which solution among the three do you think is the best fit?



The correct answer is marked in red. Please consider if this visually seems a reasonable fit to you. The general principle is we want both a *simple* and *predictive* model. The tradeoff between the two is also referred as **bias-variance tradeoff** in machine learning.

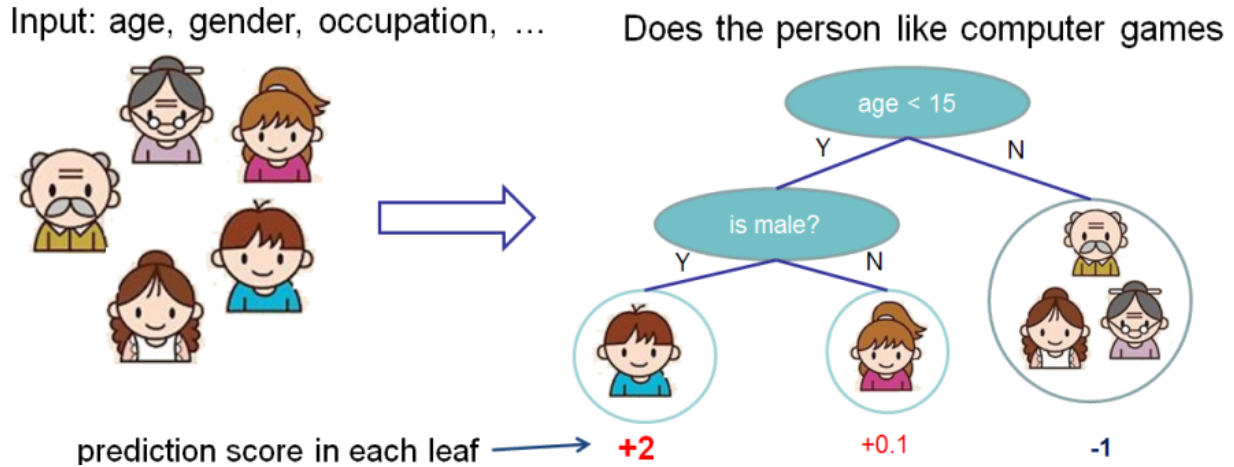
Why introduce the general principle?

The elements introduced above form the basic elements of supervised learning, and they are natural building blocks of machine learning toolkits. For example, you should be able to describe the differences and commonalities between

gradient boosted trees and random forests. Understanding the process in a formalized way also helps us to understand the objective that we are learning and the reason behind the heuristics such as pruning and smoothing.

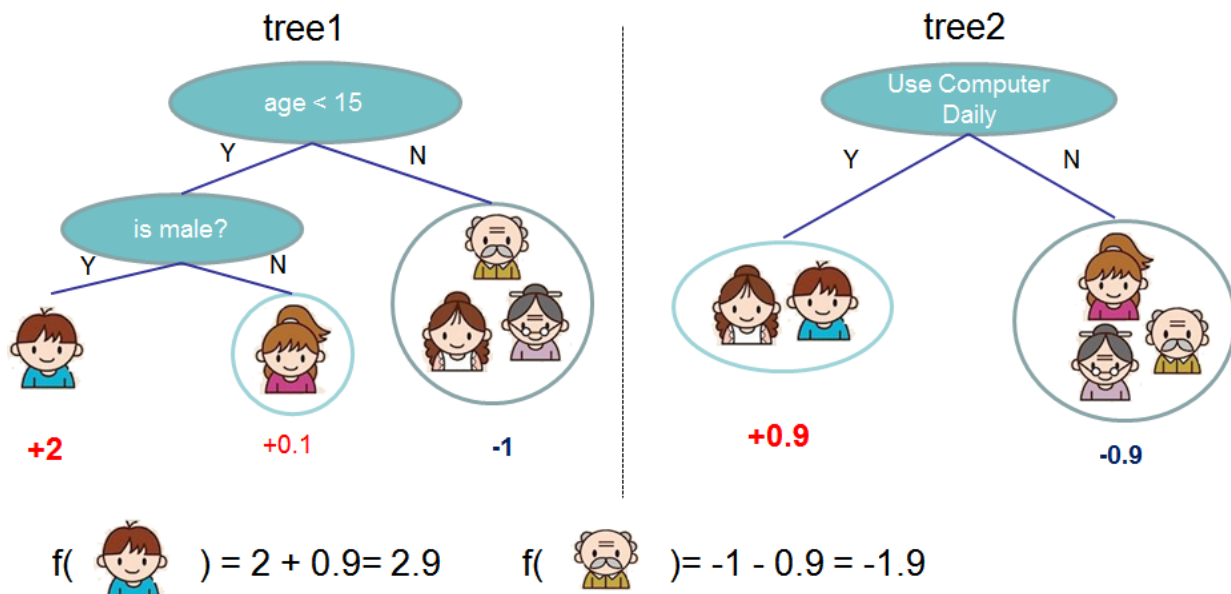
Decision Tree Ensembles

Now that we have introduced the elements of supervised learning, let us get started with real trees. To begin with, let us first learn about the model choice of XGBoost: **decision tree ensembles**. The tree ensemble model consists of a set of classification and regression trees (CART). Here's a simple example of a CART that classifies whether someone will like computer games.



We classify the members of a family into different leaves, and assign them the score on the corresponding leaf. A CART is a bit different from decision trees, in which the leaf only contains decision values. In CART, a real score is associated with each of the leaves, which gives us richer interpretations that go beyond classification. This also allows for a principled, unified approach to optimization, as we will see in a later part of this tutorial.

Usually, a single tree is not strong enough to be used in practice. What is actually used is the ensemble model, which sums the prediction of multiple trees together.



Here is an example of a tree ensemble of two trees. The prediction scores of each individual tree are summed up to

get the final score. If you look at the example, an important fact is that the two trees try to *complement* each other. Mathematically, we can write our model in the form

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F}$$

where K is the number of trees, f is a function in the functional space \mathcal{F} , and \mathcal{F} is the set of all possible CARTs. The objective function to be optimized is given by

$$\text{obj}(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Now here comes a trick question: what is the *model* used in random forests? Tree ensembles! So random forests and boosted trees are really the same models; the difference arises from how we train them. This means that, if you write a predictive service for tree ensembles, you only need to write one and it should work for both random forests and gradient boosted trees. (See [Treelite](#) for an actual example.) One example of why elements of supervised learning rock.

Tree Boosting

Now that we introduced the model, let us turn to training: How should we learn the trees? The answer is, as is always for all supervised learning models: *define an objective function and optimize it!*

Let the following be the objective function (remember it always needs to contain training loss and regularization):

$$\text{obj} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i)$$

Additive Training

The first question we want to ask: what are the **parameters** of trees? You can find that what we need to learn are those functions f_i , each containing the structure of the tree and the leaf scores. Learning tree structure is much harder than traditional optimization problem where you can simply take the gradient. It is intractable to learn all the trees at once. Instead, we use an additive strategy: fix what we have learned, and add one new tree at a time. We write the prediction value at step t as $\hat{y}_i^{(t)}$. Then we have

$$\begin{aligned} \hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ &\dots \\ \hat{y}_i^{(t)} &= \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{aligned}$$

It remains to ask: which tree do we want at each step? A natural thing is to add the one that optimizes our objective.

$$\begin{aligned} \text{obj}^{(t)} &= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i) \\ &= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{constant} \end{aligned}$$

If we consider using mean squared error (MSE) as our loss function, the objective becomes

$$\begin{aligned}\text{obj}^{(t)} &= \sum_{i=1}^n (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \sum_{i=1}^t \Omega(f_i) \\ &= \sum_{i=1}^n [2(\hat{y}_i^{(t-1)} - y_i)f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + \text{constant}\end{aligned}$$

The form of MSE is friendly, with a first order term (usually called the residual) and a quadratic term. For other losses of interest (for example, logistic loss), it is not so easy to get such a nice form. So in the general case, we take the *Taylor expansion of the loss function up to the second order*:

$$\text{obj}^{(t)} = \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + \text{constant}$$

where the g_i and h_i are defined as

$$\begin{aligned}g_i &= \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \\ h_i &= \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})\end{aligned}$$

After we remove all the constants, the specific objective at step t becomes

$$\sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

This becomes our optimization goal for the new tree. One important advantage of this definition is that the value of the objective function only depends on g_i and h_i . This is how XGBoost supports custom loss functions. We can optimize every loss function, including logistic regression and pairwise ranking, using exactly the same solver that takes g_i and h_i as input!

Model Complexity

We have introduced the training step, but wait, there is one important thing, the **regularization term**! We need to define the complexity of the tree $\Omega(f)$. In order to do so, let us first refine the definition of the tree $f(x)$ as

$$f_t(x) = w_{q(x)}, w \in R^T, q : R^d \rightarrow \{1, 2, \dots, T\}.$$

Here w is the vector of scores on leaves, q is a function assigning each data point to the corresponding leaf, and T is the number of leaves. In XGBoost, we define the complexity as

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

Of course, there is more than one way to define the complexity, but this one works well in practice. The regularization is one part most tree packages treat less carefully, or simply ignore. This was because the traditional treatment of tree learning only emphasized improving impurity, while the complexity control was left to heuristics. By defining it formally, we can get a better idea of what we are learning and obtain models that perform well in the wild.

The Structure Score

Here is the magical part of the derivation. After re-formulating the tree model, we can write the objective value with the t -th tree as:

$$\begin{aligned}\text{obj}^{(t)} &\approx \sum_{i=1}^n [g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \\ &= \sum_{j=1}^T [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T\end{aligned}$$

where $I_j = \{i | q(x_i) = j\}$ is the set of indices of data points assigned to the j -th leaf. Notice that in the second line we have changed the index of the summation because all the data points on the same leaf get the same score. We could further compress the expression by defining $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$:






$$\text{obj}^{(t)} = \sum_{j=1}^T [G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2] + \gamma T$$

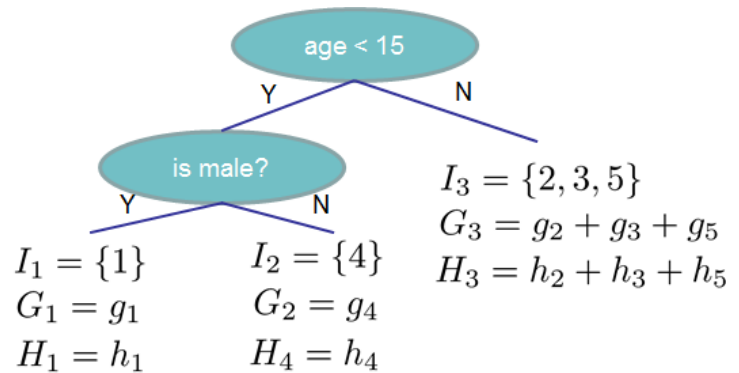
In this equation, w_j are independent with respect to each other, the form $G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2$ is quadratic and the best w_j for a given structure $q(x)$ and the best objective reduction we can get is:

$$\begin{aligned}w_j^* &= -\frac{G_j}{H_j + \lambda} \\ \text{obj}^* &= -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T\end{aligned}$$

The last equation measures *how good* a tree structure $q(x)$ is.

Instance index gradient statistics

1		g_1, h_1
2		g_2, h_2
3		g_3, h_3
4		g_4, h_4
5		g_5, h_5



$$Obj = -\sum_j \frac{G_j^2}{H_j + \lambda} + 3\gamma$$

The smaller the score is, the better the structure is

If all this sounds a bit complicated, let's take a look at the picture, and see how the scores can be calculated. Basically, for a given tree structure, we push the statistics g_i and h_i to the leaves they belong to, sum the statistics together, and use the formula to calculate how good the tree is. This score is like the impurity measure in a decision tree, except that it also takes the model complexity into account.

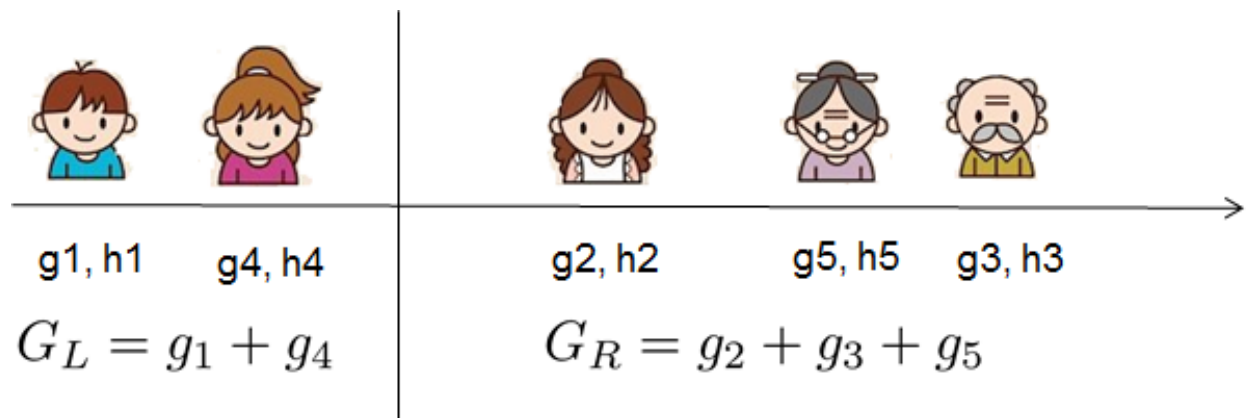
Learn the tree structure

Now that we have a way to measure how good a tree is, ideally we would enumerate all possible trees and pick the best one. In practice this is intractable, so we will try to optimize one level of the tree at a time. Specifically we try to split a leaf into two leaves, and the score it gains is

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

This formula can be decomposed as 1) the score on the new left leaf 2) the score on the new right leaf 3) The score on the original leaf 4) regularization on the additional leaf. We can see an important fact here: if the gain is smaller than γ , we would do better not to add that branch. This is exactly the **pruning** techniques in tree based models! By using the principles of supervised learning, we can naturally come up with the reason these techniques work :)

For real valued data, we usually want to search for an optimal split. To efficiently do so, we place all the instances in sorted order, like the following picture.



A left to right scan is sufficient to calculate the structure score of all possible split solutions, and we can find the best split efficiently.

Final words on XGBoost

Now that you understand what boosted trees are, you may ask, where is the introduction for XGBoost? XGBoost is exactly a tool motivated by the formal principle introduced in this tutorial! More importantly, it is developed with both deep consideration in terms of **systems optimization** and **principles in machine learning**. The goal of this library is to push the extreme of the computation limits of machines to provide a **scalable**, **portable** and **accurate** library. Make sure you try it out, and most importantly, contribute your piece of wisdom (code, examples, tutorials) to the community!

1.3.2 Distributed XGBoost YARN on AWS

This is a step-by-step tutorial on how to setup and run distributed **XGBoost** on an AWS EC2 cluster. Distributed XGBoost runs on various platforms such as MPI, SGE and Hadoop YARN. In this tutorial, we use YARN as an example since this is a widely used solution for distributed computing.

Note: XGBoost with Spark

If you are preprocessing training data with Spark, consider using *XGBoost4J-Spark*.

Prerequisite

We need to get a [AWS key-pair](#) to access the AWS services. Let us assume that we are using a key `mykey` and the corresponding permission file `mypem.pem`.

We also need [AWS credentials](#), which includes an `ACCESS_KEY_ID` and a `SECRET_ACCESS_KEY`.

Finally, we will need a S3 bucket to host the data and the model, `s3://mybucket/`

Setup a Hadoop YARN Cluster

This sections shows how to start a Hadoop YARN cluster from scratch. You can skip this step if you have already have one. We will be using `yarn-ec2` to start the cluster.

We can first clone the `yarn-ec2` script by the following command.

```
git clone https://github.com/tqchen/yarn-ec2
```

To use the script, we must set the environment variables `AWS_ACCESS_KEY_ID` and `AWS_SECRET_ACCESS_KEY` properly. This can be done by adding the following two lines in `~/.bashrc` (replacing the strings with the correct ones)

```
export AWS_ACCESS_KEY_ID=[your access ID]
export AWS_SECRET_ACCESS_KEY=[your secret access key]
```

Now we can launch a master machine of the cluster from EC2:

```
./yarn-ec2 -k mykey -i mypem.pem launch xgboost
```

Wait a few mininutes till the master machine gets up.

After the master machine gets up, we can query the public DNS of the master machine using the following command.

```
./yarn-ec2 -k mykey -i mypem.pem get-master xgboost
```

It will show the public DNS of the master machine like `ec2-xx-xx-xx.us-west-2.compute.amazonaws.com` Now we can open the browser, and type (replace the DNS with the master DNS)

```
ec2-xx-xx-xx.us-west-2.compute.amazonaws.com:8088
```

This will show the job tracker of the YARN cluster. Note that we may have to wait a few minutes before the master finishes bootstrapping and starts the job tracker.

After the master machine gets up, we can freely add more slave machines to the cluster. The following command add `m3.xlarge` instances to the cluster.

```
./yarn-ec2 -k mykey -i mypem.pem -t m3.xlarge -s 2 addslave xgboost
```

We can also choose to add two spot instances

```
./yarn-ec2 -k mykey -i mypem.pem -t m3.xlarge -s 2 addspot xgboost
```

The slave machines will start up, bootstrap and report to the master. You can check if the slave machines are connected by clicking on the Nodes link on the job tracker. Or simply type the following URL (replace DNS ith the master DNS)

```
ec2-xx-xx-xx.us-west-2.compute.amazonaws.com:8088/cluster/nodes
```

One thing we should note is that not all the links in the job tracker work. This is due to that many of them use the private IP of AWS, which can only be accessed by EC2. We can use ssh proxy to access these packages. Now that we have set up a cluster with one master and two slaves, we are ready to run the experiment.

Build XGBoost with S3

We can log into the master machine by the following command.

```
./yarn-ec2 -k mykey -i mypem.pem login xgboost
```

We will be using S3 to host the data and the result model, so the data won't get lost after the cluster shutdown. To do so, we will need to build XGBoost with S3 support. The only thing we need to do is to set `USE_S3` variable to be true. This can be achieved by the following command.

```
git clone --recursive https://github.com/dmlc/xgboost
cd xgboost
cp make/config.mk config.mk
echo "USE_S3=1" >> config.mk
make -j4
```

Now we have built the XGBoost with S3 support. You can also enable HDFS support if you plan to store data on HDFS by turning on `USE_HDFS` option. XGBoost also relies on the environment variable to access S3, so you will need to add the following two lines to `~/ .bashrc` (replacing the strings with the correct ones) on the master machine as well.

```
export AWS_ACCESS_KEY_ID=AKIAIOSFODNN7EXAMPLE
export AWS_SECRET_ACCESS_KEY=wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY
export BUCKET=mybucket
```

Host the Data on S3

In this example, we will copy the example dataset in XGBoost to the S3 bucket as input. In normal usecases, the dataset is usually created from existing distributed processing pipeline. We can use `s3cmd` to copy the data into mybucket (replace `{BUCKET}` with the real bucket name).

```
cd xgboost
s3cmd put demo/data/agaricus.txt.train s3://{BUCKET}/xgb-demo/train/
s3cmd put demo/data/agaricus.txt.test s3://{BUCKET}/xgb-demo/test/
```

Submit the Jobs

Now everything is ready, we can submit the XGBoost distributed job to the YARN cluster. We will use the `dmlc-submit` script to submit the job.

Now we can run the following script in the distributed training folder (replace `{BUCKET}` with the real bucket name)

```
cd xgboost/demo/distributed-training
# Use dmlc-submit to submit the job.
../../dmlc-core/tracker/dmlc-submit --cluster=yarn --num-workers=2 --worker-cores=2\
  ../../xgboost mushroom.aws.conf nthread=2\
  data=s3://{BUCKET}/xgb-demo/train\
  eval[test]=s3://{BUCKET}/xgb-demo/test\
  model_dir=s3://{BUCKET}/xgb-demo/model
```


All the configurations such as `data` and `model_dir` can also be directly written into the configuration file. Note that we only specified the folder path to the file, instead of the file name. XGBoost will read in all the files in that folder as the training and evaluation data.

In this command, we are using two workers, and each worker uses two running threads. XGBoost can benefit from using multiple cores in each worker. A common choice of working cores can range from 4 to 8. The trained model will be saved into the specified model folder. You can browse the model folder.

```
s3cmd ls s3://${BUCKET}/xgb-demo/model/
```

The following is an example output from distributed training.

```
16/02/26 05:41:59 INFO dmlc.Client: jobname=DMLC[nworker=2]:xgboost,username=ubuntu
16/02/26 05:41:59 INFO dmlc.Client: Submitting application application_1456461717456_
↪0015
16/02/26 05:41:59 INFO impl.YarnClientImpl: Submitted application application_
↪1456461717456_0015
2016-02-26 05:42:05,230 INFO @tracker All of 2 nodes getting started
2016-02-26 05:42:14,027 INFO [05:42:14] [0] test-error:0.016139 train-error:0.
↪014433
2016-02-26 05:42:14,186 INFO [05:42:14] [1] test-error:0.000000 train-error:0.
↪001228
2016-02-26 05:42:14,947 INFO @tracker All nodes finishes job
2016-02-26 05:42:14,948 INFO @tracker 9.71754479408 secs between node start and job_
↪finish
Application application_1456461717456_0015 finished with state FINISHED at_
↪1456465335961
```

Analyze the Model

After the model is trained, we can analyse the learnt model and use it for future prediction tasks. XGBoost is a portable framework, meaning the models in all platforms are *exchangeable*. This means we can load the trained model in python/R/Julia and take benefit of data science pipelines in these languages to do model analysis and prediction.

For example, you can use [this IPython notebook](#) to plot feature importance and visualize the learnt model.

Troubleshooting

If you encounter a problem, the best way might be to use the following command to get logs of stdout and stderr of the containers and check what causes the problem.

```
yarn logs -applicationId yourAppId
```

Future Directions

You have learned to use distributed XGBoost on YARN in this tutorial. XGBoost is a portable and scalable framework for gradient boosting. You can check out more examples and resources in the [resources page](#).

The project goal is to make the best scalable machine learning solution available to all platforms. The API is designed to be able to portable, and the same code can also run on other platforms such as MPI and SGE. XGBoost is actively evolving and we are working on even more exciting features such as distributed XGBoost python/R package.

1.3.3 DART booster

XGBoost mostly combines a huge number of regression trees with a small learning rate. In this situation, trees added early are significant and trees added late are unimportant.

Vinayak and Gilad-Bachrach proposed a new method to add dropout techniques from the deep neural net community to boosted trees, and reported better results in some situations.

This is a instruction of new tree booster `dart`.

Original paper

Rashmi Korlakai Vinayak, Ran Gilad-Bachrach. “DART: Dropouts meet Multiple Additive Regression Trees.” [JMLR](#).

Features

- Drop trees in order to solve the over-fitting.
 - Trivial trees (to correct trivial errors) may be prevented.

Because of the randomness introduced in the training, expect the following few differences:

- Training can be slower than `gbtree` because the random dropout prevents usage of the prediction buffer.
- The early stop might not be stable, due to the randomness.

How it works

- In m -th training round, suppose k trees are selected to be dropped.
- Let $D = \sum_{i \in \mathbf{K}} F_i$ be the leaf scores of dropped trees and $F_m = \eta \tilde{F}_m$ be the leaf scores of a new tree.
- The objective function is as follows:

$$\text{Obj} = \sum_{j=1}^n L\left(y_j, \hat{y}_j^{m-1} - D_j + \tilde{F}_m\right) + \Omega\left(\tilde{F}_m\right).$$

- D and F_m are overshooting, so using scale factor

$$\hat{y}_j^m = \sum_{i \notin \mathbf{K}} F_i + a \left(\sum_{i \in \mathbf{K}} F_i + b F_m \right).$$

Parameters

The booster `dart` inherits `gbtree` booster, so it supports all parameters that `gbtree` does, such as `eta`, `gamma`, `max_depth` etc.

Additional parameters are noted below:

- `sample_type`: type of sampling algorithm.
 - `uniform`: (default) dropped trees are selected uniformly.
 - `weighted`: dropped trees are selected in proportion to weight.
- `normalize_type`: type of normalization algorithm.
 - `tree`: (default) New trees have the same weight of each of dropped trees.

$$\begin{aligned}
 a \left(\sum_{i \in \mathbf{K}} F_i + \frac{1}{k} F_m \right) &= a \left(\sum_{i \in \mathbf{K}} F_i + \frac{\eta}{k} \tilde{F}_m \right) \\
 &\sim a \left(1 + \frac{\eta}{k} \right) D \\
 &= a \frac{k + \eta}{k} D = D, \\
 a &= \frac{k}{k + \eta}
 \end{aligned}$$

- forest: New trees have the same weight of sum of dropped trees (forest).

$$\begin{aligned}
 a \left(\sum_{i \in \mathbf{K}} F_i + F_m \right) &= a \left(\sum_{i \in \mathbf{K}} F_i + \eta \tilde{F}_m \right) \\
 &\sim a (1 + \eta) D \\
 &= a (1 + \eta) D = D, \\
 a &= \frac{1}{1 + \eta}.
 \end{aligned}$$

- `rate_drop`: dropout rate.
 - range: [0.0, 1.0]
- `skip_drop`: probability of skipping dropout.
 - If a dropout is skipped, new trees are added in the same manner as `gbtree`.
 - range: [0.0, 1.0]

Sample Script

```

import xgboost as xgb
# read in data
dtrain = xgb.DMatrix('demo/data/agaricus.txt.train')
dtest = xgb.DMatrix('demo/data/agaricus.txt.test')
# specify parameters via map
param = {'booster': 'dart',
         'max_depth': 5, 'learning_rate': 0.1,
         'objective': 'binary:logistic', 'silent': True,
         'sample_type': 'uniform',
         'normalize_type': 'tree',
         'rate_drop': 0.1,
         'skip_drop': 0.5}
num_round = 50
bst = xgb.train(param, dtrain, num_round)
# make prediction
# ntree_limit must not be 0
preds = bst.predict(dtest, ntree_limit=num_round)

```

Note: Specify `ntree_limit` when predicting with test sets

By default, `bst.predict()` will perform dropouts on trees. To obtain correct results on test sets, disable dropouts by specifying a nonzero value for `ntree_limit`.

1.3.4 Monotonic Constraints

It is often the case in a modeling problem or project that the functional form of an acceptable model is constrained in some way. This may happen due to business considerations, or because of the type of scientific question being investigated. In some cases, where there is a very strong prior belief that the true relationship has some quality, constraints can be used to improve the predictive performance of the model.

A common type of constraint in this situation is that certain features bear a **monotonic** relationship to the predicted response:

$$f(x_1, x_2, \dots, x, \dots, x_{n-1}, x_n) \leq f(x_1, x_2, \dots, x', \dots, x_{n-1}, x_n)$$

whenever $x \leq x'$ is an **increasing constraint**; or

$$f(x_1, x_2, \dots, x, \dots, x_{n-1}, x_n) \geq f(x_1, x_2, \dots, x', \dots, x_{n-1}, x_n)$$

whenever $x \leq x'$ is a **decreasing constraint**.

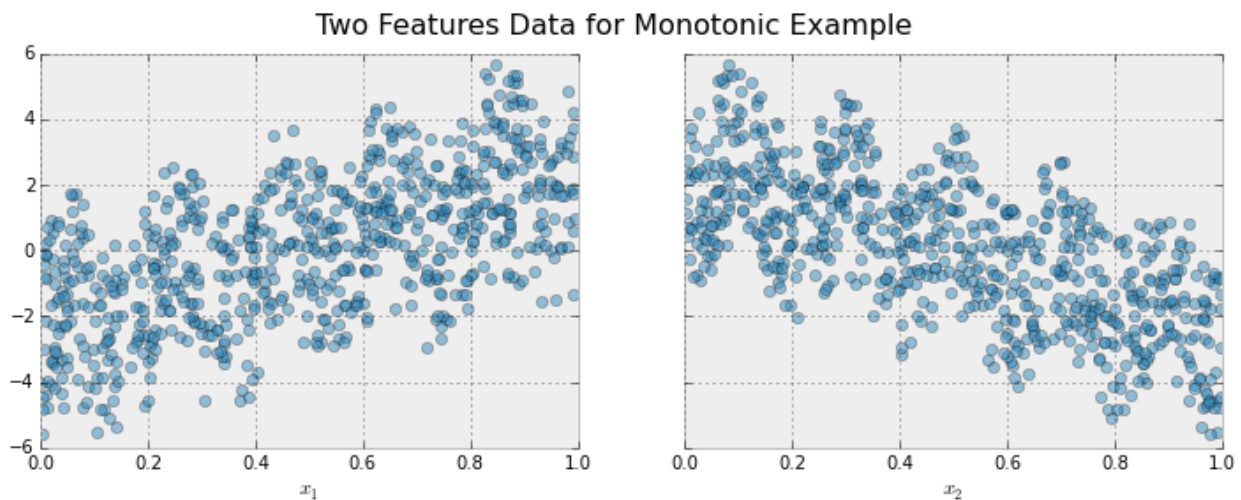
XGBoost has the ability to enforce monotonicity constraints on any features used in a boosted model.

A Simple Example

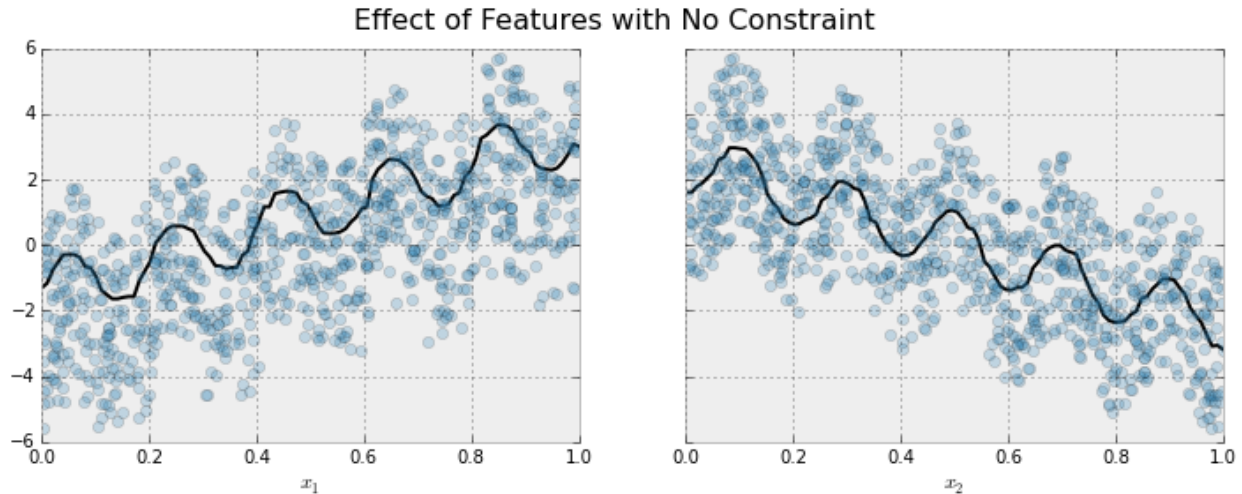
To illustrate, let's create some simulated data with two features and a response according to the following scheme

$$y = 5x_1 + \sin(10\pi x_1) - 5x_2 - \cos(10\pi x_2) + N(0, 0.01), x_1, x_2 \in [0, 1]$$

The response generally increases with respect to the x_1 feature, but a sinusoidal variation has been superimposed, resulting in the true effect being non-monotonic. For the x_2 feature the variation is decreasing with a sinusoidal variation.

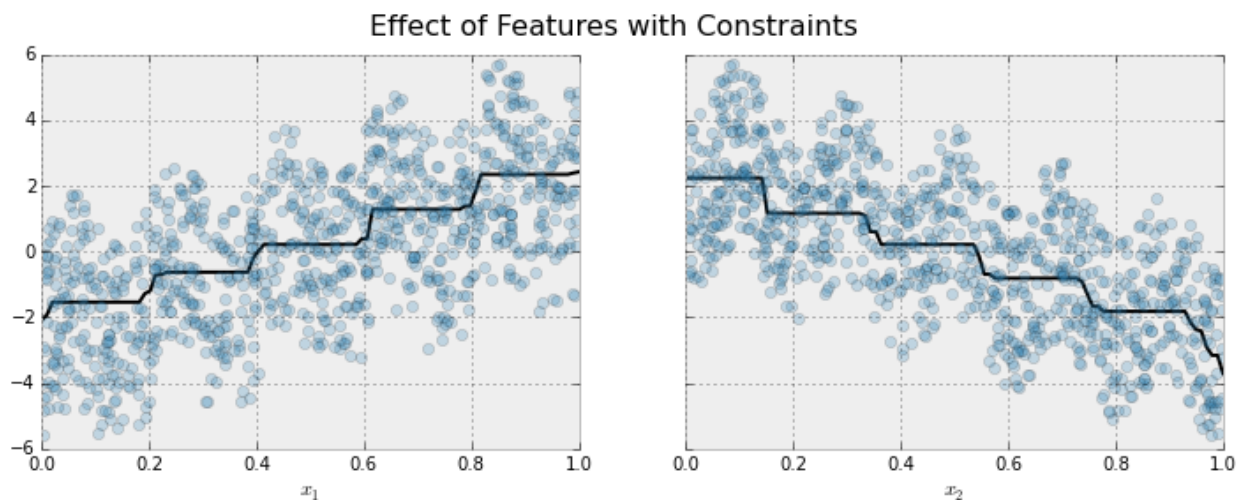


Let's fit a boosted tree model to this data without imposing any monotonic constraints:



The black curve shows the trend inferred from the model for each feature. To make these plots the distinguished feature x_i is fed to the model over a one-dimensional grid of values, while all the other features (in this case only one other feature) are set to their average values. We see that the model does a good job of capturing the general trend with the oscillatory wave superimposed.

Here is the same model, but fit with monotonicity constraints:



We see the effect of the constraint. For each variable the general direction of the trend is still evident, but the oscillatory behaviour no longer remains as it would violate our imposed constraints.

Enforcing Monotonic Constraints in XGBoost

It is very simple to enforce monotonicity constraints in XGBoost. Here we will give an example using Python, but the same general idea generalizes to other platforms.

Suppose the following code fits your model without monotonicity constraints

```
model_no_constraints = xgb.train(params, dtrain,
                                num_boost_round = 1000, evals = evallist,
                                early_stopping_rounds = 10)
```

Then fitting with monotonicity constraints only requires adding a single parameter

```
params_constrained = params.copy()
params_constrained['monotone_constraints'] = "(1,-1)"

model_with_constraints = xgb.train(params_constrained, dtrain,
                                   num_boost_round = 1000, evals = evallist,
                                   early_stopping_rounds = 10)
```

In this example the training data X has two columns, and by using the parameter values $(1, -1)$ we are telling XGBoost to impose an increasing constraint on the first predictor and a decreasing constraint on the second.

Some other examples:

- $(1, 0)$: An increasing constraint on the first predictor and no constraint on the second.
- $(0, -1)$: No constraint on the first predictor and a decreasing constraint on the second.

Choice of tree construction algorithm. To use monotonic constraints, be sure to set the `tree_method` parameter to one of `exact`, `hist`, and `gpu_hist`.

Note for the ‘hist’ tree construction algorithm. If `tree_method` is set to either `hist` or `gpu_hist`, enabling monotonic constraints may produce unnecessarily shallow trees. This is because the `hist` method reduces the number of candidate splits to be considered at each split. Monotonic constraints may wipe out all available split candidates, in which case no split is made. To reduce the effect, you may want to increase the `max_bin` parameter to consider more split candidates.

1.3.5 Text Input Format of DMatrix

Basic Input Format

XGBoost currently supports two text formats for ingesting data: LibSVM and CSV. The rest of this document will describe the LibSVM format. (See [this Wikipedia article](#) for a description of the CSV format.)

For training or predicting, XGBoost takes an instance file with the format as below:

Listing 1: `train.txt`

```
1 101:1.2 102:0.03
0 1:2.1 10001:300 10002:400
0 0:1.3 1:0.3
1 0:0.01 1:0.3
0 0:0.2 1:0.3
```

Each line represent a single instance, and in the first line ‘1’ is the instance label, ‘101’ and ‘102’ are feature indices, ‘1.2’ and ‘0.03’ are feature values. In the binary classification case, ‘1’ is used to indicate positive samples, and ‘0’ is used to indicate negative samples. We also support probability values in $[0,1]$ as label, to indicate the probability of the instance being positive.

Auxiliary Files for Additional Information

Note: all information below is applicable only to single-node version of the package. If you’d like to perform distributed training with multiple nodes, skip to the section *Embedding additional information inside LibSVM file*.

Group Input Format

For [ranking task](#), XGBoost supports the group input format. In ranking task, instances are categorized into *query groups* in real world scenarios. For example, in the learning to rank web pages scenario, the web page instances are grouped by their queries. XGBoost requires an file that indicates the group information. For example, if the instance file is the `train.txt` shown above, the group file should be named `train.txt.group` and be of the following format:

Listing 2: `train.txt.group`

```
2
3
```

This means that, the data set contains 5 instances, and the first two instances are in a group and the other three are in another group. The numbers in the group file are actually indicating the number of instances in each group in the instance file in order. At the time of configuration, you do not have to indicate the path of the group file. If the instance file name is `xxx`, XGBoost will check whether there is a file named `xxx.group` in the same directory.

Instance Weight File

Instances in the training data may be assigned weights to differentiate relative importance among them. For example, if we provide an instance weight file for the `train.txt` file in the example as below:

Listing 3: `train.txt.weight`

```
1
0.5
0.5
1
0.5
```

It means that XGBoost will emphasize more on the first and fourth instance (i.e. the positive instances) while training. The configuration is similar to configuring the group information. If the instance file name is `xxx`, XGBoost will look for a file named `xxx.weight` in the same directory. If the file exists, the instance weights will be extracted and used at the time of training.

Note: Binary buffer format and instance weights

If you choose to save the training data as a binary buffer (using `save_binary()`), keep in mind that the resulting binary buffer file will include the instance weights. To update the weights, use the `set_weight()` function.

Initial Margin File

XGBoost supports providing each instance an initial margin prediction. For example, if we have a initial prediction using logistic regression for `train.txt` file, we can create the following file:

Listing 4: `train.txt.base_margin`

```
-0.4
1.0
3.4
```

XGBoost will take these values as initial margin prediction and boost from that. An important note about `base_margin` is that it should be margin prediction before transformation, so if you are doing logistic loss, you will need to put in value before logistic transformation. If you are using XGBoost predictor, use `pred_margin=1` to output margin values.

Embedding additional information inside LibSVM file

This section is applicable to both single- and multiple-node settings.

Query ID Columns

This is most useful for [ranking task](#), where the instances are grouped into query groups. You may embed query group ID for each instance in the LibSVM file by adding a token of form `qid:xx` in each row:

Listing 5: `train.txt`

```
1 qid:1 101:1.2 102:0.03
0 qid:1 1:2.1 10001:300 10002:400
0 qid:2 0:1.3 1:0.3
1 qid:2 0:0.01 1:0.3
0 qid:3 0:0.2 1:0.3
1 qid:3 3:-0.1 10:-0.3
0 qid:3 6:0.2 10:0.15
```

Keep in mind the following restrictions:

- You are not allowed to specify query ID's for some instances but not for others. Either every row is assigned query ID's or none at all.
- The rows have to be sorted in ascending order by the query IDs. So, for instance, you may not have one row having large query ID than any of the following rows.

Instance weights

You may specify instance weights in the LibSVM file by appending each instance label with the corresponding weight in the form of `[label]:[weight]`, as shown by the following example:

Listing 6: `train.txt`

```
1:1.0 101:1.2 102:0.03
0:0.5 1:2.1 10001:300 10002:400
0:0.5 0:1.3 1:0.3
1:1.0 0:0.01 1:0.3
0:0.5 0:0.2 1:0.3
```

where the negative instances are assigned half weights compared to the positive instances.

1.3.6 Notes on Parameter Tuning

Parameter tuning is a dark art in machine learning, the optimal parameters of a model can depend on many scenarios. So it is impossible to create a comprehensive guide for doing so.

This document tries to provide some guideline for parameters in XGBoost.

Understanding Bias-Variance Tradeoff

If you take a machine learning or statistics course, this is likely to be one of the most important concepts. When we allow the model to get more complicated (e.g. more depth), the model has better ability to fit the training data, resulting in a less biased model. However, such complicated model requires more data to fit.

Most of parameters in XGBoost are about bias variance tradeoff. The best model should trade the model complexity with its predictive power carefully. [Parameters Documentation](#) will tell you whether each parameter will make the model more conservative or not. This can be used to help you turn the knob between complicated model and simple model.

Control Overfitting

When you observe high training accuracy, but low test accuracy, it is likely that you encountered overfitting problem.

There are in general two ways that you can control overfitting in XGBoost:

- The first way is to directly control model complexity.
 - This includes `max_depth`, `min_child_weight` and `gamma`.
- The second way is to add randomness to make training robust to noise.
 - This includes `subsample` and `colsample_bytree`.
 - You can also reduce stepsize `eta`. Remember to increase `num_round` when you do so.

Handle Imbalanced Dataset

For common cases such as ads clickthrough log, the dataset is extremely imbalanced. This can affect the training of XGBoost model, and there are two ways to improve it.

- If you care only about the overall performance metric (AUC) of your prediction
 - Balance the positive and negative weights via `scale_pos_weight`
 - Use AUC for evaluation
- If you care about predicting the right probability
 - In such a case, you cannot re-balance the dataset
 - Set parameter `max_delta_step` to a finite number (say 1) to help convergence

1.3.7 Using XGBoost External Memory Version (beta)

There is no big difference between using external memory version and in-memory version. The only difference is the filename format.

The external memory version takes in the following filename format:

`filename#cacheprefix`

The `filename` is the normal path to libsvm file you want to load in, and `cacheprefix` is a path to a cache file that XGBoost will use for external memory cache.

Note: External memory is not available with GPU algorithms

External memory is not available when `tree_method` is set to `gpu_exact` or `gpu_hist`.

The following code was extracted from `demo/guide-python/external_memory.py`:

```
dtrain = xgb.DMatrix('../data/agaricus.txt.train#dtrain.cache')
```

You can find that there is additional `#dtrain.cache` following the libsvm file, this is the name of cache file. For CLI version, simply add the cache suffix, e.g. `"../data/agaricus.txt.train#dtrain.cache"`.

Performance Note

- the parameter `nthread` should be set to number of **physical** cores
 - Most modern CPUs use hyperthreading, which means a 4 core CPU may carry 8 threads
 - Set `nthread` to be 4 for maximum performance in such case

Distributed Version

The external memory mode naturally works on distributed version, you can simply set path like

```
data = "hdfs://path-to-data/#dtrain.cache"
```

XGBoost will cache the data to the local position. When you run on YARN, the current folder is temporal so that you can directly use `dtrain.cache` to cache to current folder.

Usage Note

- This is a experimental version
- Currently only importing from libsvm format is supported
 - Contribution of ingestion from other common external memory data source is welcomed

1.4 Frequently Asked Questions

This document contains frequently asked questions about XGBoost.

1.4.1 How to tune parameters

See *Parameter Tuning Guide*.

1.4.2 Description on the model

See *Introduction to Boosted Trees*.

1.4.3 I have a big dataset

XGBoost is designed to be memory efficient. Usually it can handle problems as long as the data fit into your memory. (This usually means millions of instances) If you are running out of memory, checkout [external memory version](#) or [distributed version](#) of XGBoost.

1.4.4 Running XGBoost on Platform X (Hadoop/Yarn, Mesos)

The distributed version of XGBoost is designed to be portable to various environment. Distributed XGBoost can be ported to any platform that supports [rabbit](#). You can directly run XGBoost on Yarn. In theory Mesos and other resource allocation engines can be easily supported as well.

1.4.5 Why not implement distributed XGBoost on top of X (Spark, Hadoop)

The first fact we need to know is going distributed does not necessarily solve all the problems. Instead, it creates more problems such as more communication overhead and fault tolerance. The ultimate question will still come back to how to push the limit of each computation node and use less resources to complete the task (thus with less communication and chance of failure).

To achieve these, we decide to reuse the optimizations in the single node XGBoost and build distributed version on top of it. The demand of communication in machine learning is rather simple, in the sense that we can depend on a limited set of API (in our case [rabbit](#)). Such design allows us to reuse most of the code, while being portable to major platforms such as Hadoop/Yarn, MPI, SGE. Most importantly, it pushes the limit of the computation resources we can use.

1.4.6 How can I port the model to my own system

The model and data format of XGBoost is exchangeable, which means the model trained by one language can be loaded in another. This means you can train the model using R, while running prediction using Java or C++, which are more common in production systems. You can also train the model using distributed versions, and load them in from Python to do some interactive analysis.

1.4.7 Do you support LambdaMART

Yes, XGBoost implements LambdaMART. Checkout the objective section in [parameters](#).

1.4.8 How to deal with Missing Value

XGBoost supports missing value by default. In tree algorithms, branch directions for missing values are learned during training. Note that the gblinear booster treats missing values as zeros.

1.4.9 Slightly different result between runs

This could happen, due to non-determinism in floating point summation order and multi-threading. Though the general accuracy will usually remain the same.

1.4.10 Why do I see different results with sparse and dense data?

“Sparse” elements are treated as if they were “missing” by the tree booster, and as zeros by the linear booster. For tree models, it is important to use consistent data formats during training and scoring.

1.5 XGBoost GPU Support

This page contains information about GPU algorithms supported in XGBoost. To install GPU support, checkout the *Installation Guide*.

Note: CUDA 8.0, Compute Capability 3.5 required

The GPU algorithms in XGBoost require a graphics card with compute capability 3.5 or higher, with CUDA toolkits 8.0 or later. (See [this list](#) to look up compute capability of your GPU card.)

1.5.1 CUDA Accelerated Tree Construction Algorithms

Tree construction (training) and prediction can be accelerated with CUDA-capable GPUs.

Usage

Specify the `tree_method` parameter as one of the following algorithms.

Algorithms

tree_method	Description
gpu_exact	The standard XGBoost tree construction algorithm. Performs exact search for splits. Slower and uses considerably more memory than <code>gpu_hist</code> .
gpu_hist	Equivalent to the XGBoost fast histogram algorithm. Much faster and uses considerably less memory. NOTE: Will run very slowly on GPUs older than Pascal architecture.

Supported parameters

parameter	gpu_exact	gpu_hist
subsample		
colsample_bytree		
colsample_bylevel		
max_bin		
gpu_id		
n_gpus		
predictor		
grow_policy		
monotone_constraints		

GPU accelerated prediction is enabled by default for the above mentioned `tree_method` parameters but can be switched to CPU prediction by setting `predictor` to `cpu_predictor`. This could be useful if you want to

conserve GPU memory. Likewise when using CPU algorithms, GPU accelerated prediction can be enabled by setting `predictor` to `gpu_predictor`.

The device ordinal can be selected using the `gpu_id` parameter, which defaults to 0.

Multiple GPUs can be used with the `gpu_hist` tree method using the `n_gpus` parameter, which defaults to 1. If this is set to -1 all available GPUs will be used. If `gpu_id` is specified as non-zero, the gpu device order is $\text{mod}(\text{gpu_id} + i) \% \text{n_visible_devices}$ for $i=0$ to $\text{n_gpus}-1$. As with GPU vs. CPU, multi-GPU will not always be faster than a single GPU due to PCI bus bandwidth that can limit performance.

Note: Enabling multi-GPU training

Default installation may not enable multi-GPU training. To use multiple GPUs, make sure to read [Building with GPU support](#).

The GPU algorithms currently work with CLI, Python and R packages. See [Installation Guide](#) for details.

Listing 7: Python example

```
param['gpu_id'] = 0
param['max_bin'] = 16
param['tree_method'] = 'gpu_hist'
```

Benchmarks

You can run benchmarks on synthetic data for binary classification:

```
python tests/benchmark/benchmark.py
```

Training time on 1,000,000 rows x 50 columns with 500 boosting iterations and 0.25/0.75 test/train split on i7-6700K CPU @ 4.00GHz and Pascal Titan X yields the following results:

tree_method	Time (s)
gpu_hist	13.87
hist	63.55
gpu_exact	161.08
exact	1082.20

See [GPU Accelerated XGBoost](#) and [Updates to the XGBoost GPU algorithms](#) for additional performance benchmarks of the `gpu_exact` and `gpu_hist` tree methods.

1.5.2 References

Mitchell R, Frank E. (2017) Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science* 3:e127 <https://doi.org/10.7717/peerj-cs.127>

Nvidia Parallel Forall: Gradient Boosting, Decision Trees and XGBoost with CUDA

Authors

- Rory Mitchell
- Jonathan C. McKinney

- Shankara Rao Thejaswi Nanditale
- Vinay Deshpande
- ... and the rest of the H2O.ai and NVIDIA team.

Please report bugs to the user forum <https://discuss.xgboost.ai/>.

1.6 XGBoost Parameters

Before running XGBoost, we must set three types of parameters: general parameters, booster parameters and task parameters.

- **General parameters** relate to which booster we are using to do boosting, commonly tree or linear model
- **Booster parameters** depend on which booster you have chosen
- **Learning task parameters** decide on the learning scenario. For example, regression tasks may use different parameters with ranking tasks.
- **Command line parameters** relate to behavior of CLI version of XGBoost.

Note: Parameters in R package

In R-package, you can use `.` (dot) to replace underscore in the parameters, for example, you can use `max.depth` to indicate `max_depth`. The underscore parameters are also valid in R.

- *General Parameters*
 - *Parameters for Tree Booster*
 - *Additional parameters for Dart Booster (`booster=dart`)*
 - *Parameters for Linear Booster (`booster=gblinear`)*
 - *Parameters for Tweedie Regression (`objective=reg:tweedie`)*
- *Learning Task Parameters*
- *Command Line Parameters*

1.6.1 General Parameters

- `booster` [default= `gbtree`]
 - Which booster to use. Can be `gbtree`, `gblinear` or `dart`; `gbtree` and `dart` use tree based models while `gblinear` uses linear functions.
- `silent` [default=0]
 - 0 means printing running messages, 1 means silent mode
- `nthread` [default to maximum number of threads available if not set]
 - Number of parallel threads used to run XGBoost
- `num_pbuffer` [set automatically by XGBoost, no need to be set by user]

- Size of prediction buffer, normally set to number of training instances. The buffers are used to save the prediction results of last boosting step.
- `num_feature` [set automatically by XGBoost, no need to be set by user]
 - Feature dimension used in boosting, set to maximum dimension of the feature

Parameters for Tree Booster

- `eta` [default=0.3, alias: `learning_rate`]
 - Step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features, and `eta` shrinks the feature weights to make the boosting process more conservative.
 - range: [0,1]
- `gamma` [default=0, alias: `min_split_loss`]
 - Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger `gamma` is, the more conservative the algorithm will be.
 - range: [0,∞]
- `max_depth` [default=6]
 - Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. 0 indicates no limit. Note that limit is required when `grow_policy` is set of `depthwise`.
 - range: [0,∞]
- `min_child_weight` [default=1]
 - Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than `min_child_weight`, then the building process will give up further partitioning. In linear regression task, this simply corresponds to minimum number of instances needed to be in each node. The larger `min_child_weight` is, the more conservative the algorithm will be.
 - range: [0,∞]
- `max_delta_step` [default=0]
 - Maximum delta step we allow each leaf output to be. If the value is set to 0, it means there is no constraint. If it is set to a positive value, it can help making the update step more conservative. Usually this parameter is not needed, but it might help in logistic regression when class is extremely imbalanced. Set it to value of 1-10 might help control the update.
 - range: [0,∞]
- `subsample` [default=1]
 - Subsample ratio of the training instances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees. and this will prevent overfitting. Subsampling will occur once in every boosting iteration.
 - range: (0,1]
- `colsample_bytree` [default=1]
 - Subsample ratio of columns when constructing each tree. Subsampling will occur once in every boosting iteration.
 - range: (0,1]

- `colsample_bylevel` [default=1]
 - Subsample ratio of columns for each split, in each level. Subsampling will occur each time a new split is made. This parameter has no effect when `tree_method` is set to `hist`.
 - range: (0,1]
- `lambda` [default=1, alias: `reg_lambda`]
 - L2 regularization term on weights. Increasing this value will make model more conservative.
- `alpha` [default=0, alias: `reg_alpha`]
 - L1 regularization term on weights. Increasing this value will make model more conservative.
- `tree_method` string [default= `auto`]
 - The tree construction algorithm used in XGBoost. See description in the [reference paper](#).
 - Distributed and external memory version only support `tree_method=approx`.
 - Choices: `auto`, `exact`, `approx`, `hist`, `gpu_exact`, `gpu_hist`
 - * `auto`: Use heuristic to choose the fastest method.
 - For small to medium dataset, exact greedy (`exact`) will be used.
 - For very large dataset, approximate algorithm (`approx`) will be chosen.
 - Because old behavior is always use exact greedy in single machine, user will get a message when approximate algorithm is chosen to notify this choice.
 - * `exact`: Exact greedy algorithm.
 - * `approx`: Approximate greedy algorithm using quantile sketch and gradient histogram.
 - * `hist`: Fast histogram optimized approximate greedy algorithm. It uses some performance improvements such as bins caching.
 - * `gpu_exact`: GPU implementation of `exact` algorithm.
 - * `gpu_hist`: GPU implementation of `hist` algorithm.
- `sketch_eps` [default=0.03]
 - Only used for `tree_method=approx`.
 - This roughly translates into $O(1 / \text{sketch_eps})$ number of bins. Compared to directly select number of bins, this comes with theoretical guarantee with sketch accuracy.
 - Usually user does not have to tune this. But consider setting to a lower number for more accurate enumeration of split candidates.
 - range: (0, 1)
- `scale_pos_weight` [default=1]
 - Control the balance of positive and negative weights, useful for unbalanced classes. A typical value to consider: `sum(negative instances) / sum(positive instances)`. See [Parameters Tuning](#) for more discussion. Also, see Higgs Kaggle competition demo for examples: [R](#), [py1](#), [py2](#), [py3](#).
- `updater` [default= `grow_colmaker,prune`]
 - A comma separated string defining the sequence of tree updaters to run, providing a modular way to construct and to modify the trees. This is an advanced parameter that is usually set automatically, depending on some other parameters. However, it could be also set explicitly by a user. The following updater plugins exist:

- * `grow_colmaker`: non-distributed column-based construction of trees.
- * `distcol`: distributed tree construction with column-based data splitting mode.
- * `grow_histmaker`: distributed tree construction with row-based data splitting based on global proposal of histogram counting.
- * `grow_local_histmaker`: based on local histogram counting.
- * `grow_skmaker`: uses the approximate sketching algorithm.
- * `sync`: synchronizes trees in all distributed nodes.
- * `refresh`: refreshes tree's statistics and/or leaf values based on the current data. Note that no random subsampling of data rows is performed.
- * `prune`: prunes the splits where $\text{loss} < \text{min_split_loss}$ (or γ).
- In a distributed setting, the implicit updater sequence value would be adjusted to `grow_histmaker`, `prune`.
- `refresh_leaf` [default=1]
 - This is a parameter of the `refresh` updater plugin. When this flag is 1, tree leafs as well as tree nodes' stats are updated. When it is 0, only node stats are updated.
- `process_type` [default= default]
 - A type of boosting process to run.
 - Choices: `default`, `update`
 - * `default`: The normal boosting process which creates new trees.
 - * `update`: Starts from an existing model and only updates its trees. In each boosting iteration, a tree from the initial model is taken, a specified sequence of updater plugins is run for that tree, and a modified tree is added to the new model. The new model would have either the same or smaller number of trees, depending on the number of boosting iterations performed. Currently, the following built-in updater plugins could be meaningfully used with this process type: `refresh`, `prune`. With `process_type=update`, one cannot use updater plugins that create new trees.
- `grow_policy` [default= depthwise]
 - Controls a way new nodes are added to the tree.
 - Currently supported only if `tree_method` is set to `hist`.
 - Choices: `depthwise`, `lossguide`
 - * `depthwise`: split at nodes closest to the root.
 - * `lossguide`: split at nodes with highest loss change.
- `max_leaves` [default=0]
 - Maximum number of nodes to be added. Only relevant when `grow_policy=lossguide` is set.
- `max_bin`, [default=256]
 - Only used if `tree_method` is set to `hist`.
 - Maximum number of discrete bins to bucket continuous features.
 - Increasing this number improves the optimality of splits at the cost of higher computation time.
- `predictor`, [default='cpu_predictor']
 - The type of predictor algorithm to use. Provides the same results but allows the use of GPU or CPU.

- * `cpu_predictor`: Multicore CPU prediction algorithm.
- * `gpu_predictor`: Prediction using GPU. Default when `tree_method` is `gpu_exact` or `gpu_hist`.

Additional parameters for Dart Booster (`booster=dart`)

Note: Using `predict()` with DART booster

If the booster object is DART type, `predict()` will perform dropouts, i.e. only some of the trees will be evaluated. This will produce incorrect results if data is not the training data. To obtain correct results on test sets, set `ntree_limit` to a nonzero value, e.g.

```
preds = bst.predict(dtest, ntree_limit=num_round)
```

- `sample_type` [default= uniform]
 - Type of sampling algorithm.
 - * `uniform`: dropped trees are selected uniformly.
 - * `weighted`: dropped trees are selected in proportion to weight.
- `normalize_type` [default= tree]
 - Type of normalization algorithm.
 - * `tree`: new trees have the same weight of each of dropped trees.
 - Weight of new trees are $1 / (k + \text{learning_rate})$.
 - Dropped trees are scaled by a factor of $k / (k + \text{learning_rate})$.
 - * `forest`: new trees have the same weight of sum of dropped trees (forest).
 - Weight of new trees are $1 / (1 + \text{learning_rate})$.
 - Dropped trees are scaled by a factor of $1 / (1 + \text{learning_rate})$.
- `rate_drop` [default=0.0]
 - Dropout rate (a fraction of previous trees to drop during the dropout).
 - range: [0.0, 1.0]
- `one_drop` [default=0]
 - When this flag is enabled, at least one tree is always dropped during the dropout (allows Binomial-plus-one or epsilon-dropout from the original DART paper).
- `skip_drop` [default=0.0]
 - Probability of skipping the dropout procedure during a boosting iteration.
 - * If a dropout is skipped, new trees are added in the same manner as `gbtree`.
 - * Note that non-zero `skip_drop` has higher priority than `rate_drop` or `one_drop`.
 - range: [0.0, 1.0]

Parameters for Linear Booster (`booster=gblinear`)

- `lambda` [default=0, alias: `reg_lambda`]
 - L2 regularization term on weights. Increasing this value will make model more conservative. Normalised to number of training examples.
- `alpha` [default=0, alias: `reg_alpha`]
 - L1 regularization term on weights. Increasing this value will make model more conservative. Normalised to number of training examples.
- `updater` [default= `shotgun`]
 - Choice of algorithm to fit linear model
 - * `shotgun`: Parallel coordinate descent algorithm based on shotgun algorithm. Uses ‘hogwild’ parallelism and therefore produces a nondeterministic solution on each run.
 - * `coord_descent`: Ordinary coordinate descent algorithm. Also multithreaded but still produces a deterministic solution.

Parameters for Tweedie Regression (`objective=reg:tweedie`)

- `tweedie_variance_power` [default=1.5]
 - Parameter that controls the variance of the Tweedie distribution $\text{var}(y) \sim E(y)^{\text{tweedie_variance_power}}$
 - range: (1,2)
 - Set closer to 2 to shift towards a gamma distribution
 - Set closer to 1 to shift towards a Poisson distribution.

1.6.2 Learning Task Parameters

Specify the learning task and the corresponding learning objective. The objective options are below:

- `objective` [default=`reg:linear`]
 - `reg:linear`: linear regression
 - `reg:logistic`: logistic regression
 - `binary:logistic`: logistic regression for binary classification, output probability
 - `binary:logitraw`: logistic regression for binary classification, output score before logistic transformation
 - `binary:hinge`: hinge loss for binary classification. This makes predictions of 0 or 1, rather than producing probabilities.
 - `gpu:reg:linear`, `gpu:reg:logistic`, `gpu:binary:logistic`, `gpu:binary:logitraw`: versions of the corresponding objective functions evaluated on the GPU; note that like the GPU histogram algorithm, they can only be used when the entire training session uses the same dataset
 - `count:poisson`: poisson regression for count data, output mean of poisson distribution
 - * `max_delta_step` is set to 0.7 by default in poisson regression (used to safeguard optimization)

- `survival:cox`: Cox regression for right censored survival time data (negative values are considered right censored). Note that predictions are returned on the hazard ratio scale (i.e., as $HR = \exp(\text{marginal_prediction})$ in the proportional hazard function $h(t) = h_0(t) * HR$).
- `multi:softmax`: set XGBoost to do multiclass classification using the softmax objective, you also need to set `num_class`(number of classes)
- `multi:softprob`: same as softmax, but output a vector of `ndata * nclass`, which can be further reshaped to `ndata * nclass` matrix. The result contains predicted probability of each data point belonging to each class.
- `rank:pairwise`: set XGBoost to do ranking task by minimizing the pairwise loss
- `reg:gamma`: gamma regression with log-link. Output is a mean of gamma distribution. It might be useful, e.g., for modeling insurance claims severity, or for any outcome that might be [gamma-distributed](#).
- `reg:tweedie`: Tweedie regression with log-link. It might be useful, e.g., for modeling total loss in insurance, or for any outcome that might be [Tweedie-distributed](#).
- `base_score` [default=0.5]
 - The initial prediction score of all instances, global bias
 - For sufficient number of iterations, changing this value will not have too much effect.
- `eval_metric` [default according to objective]
 - Evaluation metrics for validation data, a default metric will be assigned according to objective (rmse for regression, and error for classification, mean average precision for ranking)
 - User can add multiple evaluation metrics. Python users: remember to pass the metrics in as list of parameters pairs instead of map, so that latter `eval_metric` won't override previous one
 - The choices are listed below:
 - * `rmse`: [root mean square error](#)
 - * `mae`: [mean absolute error](#)
 - * `logloss`: [negative log-likelihood](#)
 - * `error`: Binary classification error rate. It is calculated as $\#(\text{wrong cases}) / \#(\text{all cases})$. For the predictions, the evaluation will regard the instances with prediction value larger than 0.5 as positive instances, and the others as negative instances.
 - * `error@t`: a different than 0.5 binary classification threshold value could be specified by providing a numerical value through 't'.
 - * `merror`: Multiclass classification error rate. It is calculated as $\#(\text{wrong cases}) / \#(\text{all cases})$.
 - * `mlogloss`: [Multiclass logloss](#).
 - * `auc`: [Area under the curve](#)
 - * `ndcg`: [Normalized Discounted Cumulative Gain](#)
 - * `map`: [Mean average precision](#)
 - * `ndcg@n`, `map@n`: 'n' can be assigned as an integer to cut off the top positions in the lists for evaluation.
 - * `ndcg-`, `map-`, `ndcg@n-`, `map@n-`: In XGBoost, NDCG and MAP will evaluate the score of a list without any positive samples as 1. By adding "-" in the evaluation metric XGBoost will evaluate these score as 0 to be consistent under some conditions.

- * poisson-nloglik: negative log-likelihood for Poisson regression
- * gamma-nloglik: negative log-likelihood for gamma regression
- * cox-nloglik: negative partial log-likelihood for Cox proportional hazards regression
- * gamma-deviance: residual deviance for gamma regression
- * tweedie-nloglik: negative log-likelihood for Tweedie regression (at a specified value of the tweedie_variance_power parameter)
- seed [default=0]
 - Random number seed.

1.6.3 Command Line Parameters

The following parameters are only used in the console version of XGBoost

- num_round
 - The number of rounds for boosting
- data
 - The path of training data
- test:data
 - The path of test data to do prediction
- save_period [default=0]
 - The period to save the model. Setting save_period=10 means that for every 10 rounds XGBoost will save the model. Setting it to 0 means not saving any model during the training.
- task [default= train] options: train, pred, eval, dump
 - train: training using data
 - pred: making prediction for test:data
 - eval: for evaluating statistics specified by eval[name]=filename
 - dump: for dump the learned model into text format
- model_in [default=NULL]
 - Path to input model, needed for test, eval, dump tasks. If it is specified in training, XGBoost will continue training from the input model.
- model_out [default=NULL]
 - Path to output model after training finishes. If not specified, XGBoost will output files with such names as 0003.model where 0003 is number of boosting rounds.
- model_dir [default= models/]
 - The output directory of the saved models during training
- fmap
 - Feature map, used for dumping model
- dump_format [default= text] options: text, json
 - Format of model dump file

- `name_dump` [default= `dump.txt`]
 - Name of model dump file
- `name_pred` [default= `pred.txt`]
 - Name of prediction file, used in pred mode
- `pred_margin` [default=0]
 - Predict margin instead of transformed probability

1.7 XGBoost Python Package

This page contains links to all the python related documents on python package. To install the package package, checkout [Installation Guide](#).

1.7.1 Contents

Python Package Introduction

This document gives a basic walkthrough of xgboost python package.

List of other Helpful Links

- [Python walkthrough code collections](#)
- [Python API Reference](#)

Install XGBoost

To install XGBoost, follow instructions in [Installation Guide](#).

To verify your installation, run the following in Python:

```
import xgboost as xgb
```

Data Interface

The XGBoost python module is able to load data from:

- LibSVM text format file
- Comma-separated values (CSV) file
- NumPy 2D array
- SciPy 2D sparse array
- Pandas data frame, and
- XGBoost binary buffer file.

(See [Text Input Format of DMatrix](#) for detailed description of text input format.)

The data is stored in a `DMatrix` object.

- To load a libsvm text file or a XGBoost binary file into `DMatrix`:

```
dtrain = xgb.DMatrix('train.svm.txt')
dtest = xgb.DMatrix('test.svm.buffer')
```

- To load a CSV file into *DMatrix*:

```
# label_column specifies the index of the column containing the true label
dtrain = xgb.DMatrix('train.csv?format=csv&label_column=0')
dtest = xgb.DMatrix('test.csv?format=csv&label_column=0')
```

(Note that XGBoost does not support categorical features; if your data contains categorical features, load it as a NumPy array first and then perform [one-hot encoding](#).)

- To load a NumPy array into *DMatrix*:

```
data = np.random.rand(5, 10) # 5 entities, each contains 10 features
label = np.random.randint(2, size=5) # binary target
dtrain = xgb.DMatrix(data, label=label)
```

- To load a `scipy.sparse` array into *DMatrix*:

```
csr = scipy.sparse.csr_matrix((data, (row, col)))
dtrain = xgb.DMatrix(csr)
```

- To load a Pandas data frame into *DMatrix*:

```
data = pandas.DataFrame(np.arange(12).reshape((4,3)), columns=['a', 'b', 'c'])
label = pandas.DataFrame(np.random.randint(2, size=4))
dtrain = xgb.DMatrix(data, label=label)
```

- Saving *DMatrix* into a XGBoost binary file will make loading faster:

```
dtrain = xgb.DMatrix('train.svm.txt')
dtrain.save_binary('train.buffer')
```

- Missing values can be replaced by a default value in the *DMatrix* constructor:

```
dtrain = xgb.DMatrix(data, label=label, missing=-999.0)
```

- Weights can be set when needed:

```
w = np.random.rand(5, 1)
dtrain = xgb.DMatrix(data, label=label, missing=-999.0, weight=w)
```

Setting Parameters

XGBoost can use either a list of pairs or a dictionary to set *parameters*. For instance:

- Booster parameters

```
param = {'max_depth': 2, 'eta': 1, 'silent': 1, 'objective': 'binary:logistic'}
param['nthread'] = 4
param['eval_metric'] = 'auc'
```

- You can also specify multiple eval metrics:

```
param['eval_metric'] = ['auc', 'ams@0']

# alternatively:
# plst = param.items()
# plst += [('eval_metric', 'ams@0')]
```

- Specify validations set to watch performance

```
evallist = [(dtest, 'eval'), (dtrain, 'train')]
```

Training

Training a model requires a parameter list and data set.

```
num_round = 10
bst = xgb.train(param, dtrain, num_round, evallist)
```

After training, the model can be saved.

```
bst.save_model('0001.model')
```

The model and its feature map can also be dumped to a text file.

```
# dump model
bst.dump_model('dump.raw.txt')
# dump model with feature map
bst.dump_model('dump.raw.txt', 'featmap.txt')
```

A saved model can be loaded as follows:

```
bst = xgb.Booster({'nthread': 4}) # init model
bst.load_model('model.bin') # load data
```

Early Stopping

If you have a validation set, you can use early stopping to find the optimal number of boosting rounds. Early stopping requires at least one set in `evals`. If there's more than one, it will use the last.

```
train(..., evals=evals, early_stopping_rounds=10)
```

The model will train until the validation score stops improving. Validation error needs to decrease at least every `early_stopping_rounds` to continue training.

If early stopping occurs, the model will have three additional fields: `bst.best_score`, `bst.best_iteration` and `bst.best_ntree_limit`. Note that `xgboost.train()` will return a model from the last iteration, not the best one.

This works with both metrics to minimize (RMSE, log loss, etc.) and to maximize (MAP, NDCG, AUC). Note that if you specify more than one evaluation metric the last one in `param['eval_metric']` is used for early stopping.

Prediction

A model that has been trained or loaded can perform predictions on data sets.


```
# 7 entities, each contains 10 features
data = np.random.rand(7, 10)
dtest = xgb.DMatrix(data)
ypred = bst.predict(dtest)
```

If early stopping is enabled during training, you can get predictions from the best iteration with `bst.best_ntree_limit`:

```
ypred = bst.predict(dtest, ntree_limit=bst.best_ntree_limit)
```

Plotting

You can use plotting module to plot importance and output tree.

To plot importance, use `xgboost.plot_importance()`. This function requires `matplotlib` to be installed.

```
xgb.plot_importance(bst)
```

To plot the output tree via `matplotlib`, use `xgboost.plot_tree()`, specifying the ordinal number of the target tree. This function requires `graphviz` and `matplotlib`.

```
xgb.plot_tree(bst, num_trees=2)
```

When you use IPython, you can use the `xgboost.to_graphviz()` function, which converts the target tree to a `graphviz` instance. The `graphviz` instance is automatically rendered in IPython.

```
xgb.to_graphviz(bst, num_trees=2)
```

Python API Reference

This page gives the Python API reference of xgboost, please also refer to Python Package Introduction for more information about python package.

- [Core Data Structure](#)
- [Learning API](#)
- [Scikit-Learn API](#)
- [Plotting API](#)

Core Data Structure

Core XGBoost Library.

class `xgboost.DMatrix`(*data*, *label=None*, *missing=None*, *weight=None*, *silent=False*, *feature_names=None*, *feature_types=None*, *nthread=None*)

Bases: `object`

Data Matrix used in XGBoost.

DMatrix is a internal data structure that used by XGBoost which is optimized for both memory efficiency and training speed. You can construct DMatrix from `numpy.array`s

Parameters

- **data** (*string/numpy array/scipy.sparse/pd.DataFrame/DataTable*) – Data source of DMatrix. When data is string type, it represents the path libsvm format txt file, or binary file that xgboost can read from.
- **label** (*list or numpy 1-D array, optional*) – Label of the training data.
- **missing** (*float, optional*) – Value in the data which needs to be present as a missing value. If None, defaults to np.nan.
- **weight** (*list or numpy 1-D array, optional*) – Weight for each instance.
- **silent** (*boolean, optional*) – Whether print messages during construction
- **feature_names** (*list, optional*) – Set names for features.
- **feature_types** (*list, optional*) – Set types for features.
- **nthread** (*integer, optional*) – Number of threads to use for loading data from numpy array. If -1, uses maximum threads available on the system.

feature_names

Get feature names (column labels).

Returns **feature_names**

Return type *list* or *None*

feature_types

Get feature types (column types).

Returns **feature_types**

Return type *list* or *None*

get_base_margin()

Get the base margin of the DMatrix.

Returns **base_margin**

Return type *float*

get_float_info(field)

Get float property from the DMatrix.

Parameters **field** (*str*) – The field name of the information

Returns **info** – a numpy array of float information of the data

Return type *array*

get_label()

Get the label of the DMatrix.

Returns **label**

Return type *array*

get_uint_info(field)

Get unsigned integer property from the DMatrix.

Parameters **field** (*str*) – The field name of the information

Returns **info** – a numpy array of unsigned integer information of the data

Return type *array*

get_weight()

Get the weight of the DMatrix.

Returns `weight`

Return type `array`

num_col()

Get the number of columns (features) in the DMatrix.

Returns `number of columns`

Return type `int`

num_row()

Get the number of rows in the DMatrix.

Returns `number of rows`

Return type `int`

save_binary(fname, silent=True)

Save DMatrix to an XGBoost buffer.

Parameters

- **fname** (*string*) – Name of the output buffer file.
- **silent** (*bool (optional; default: True)*) – If set, the output is suppressed.

set_base_margin(margin)

Set base margin of booster to start from.

This can be used to specify a prediction value of existing model to be `base_margin`. However, remember margin is needed, instead of transformed prediction e.g. for logistic regression: need to put in value before logistic transformation see also `example/demo.py`

Parameters **margin** (*array like*) – Prediction margin of each datapoint

set_float_info(field, data)

Set float type property into the DMatrix.

Parameters

- **field** (*str*) – The field name of the information
- **data** (*numpy array*) – The array of data to be set

set_float_info_numpy2d(field, data)

Set float type property into the DMatrix for numpy 2d array input

Parameters

- **field** (*str*) – The field name of the information
- **data** (*numpy array*) – The array of data to be set

set_group(group)

Set group size of DMatrix (used for ranking).

Parameters **group** (*array like*) – Group size of each group

set_label(label)

Set label of dmatrix

Parameters `label` (*array like*) – The label information to be set into DMatrix

set_label_np2d (`label`)
Set label of dmatrix

Parameters `label` (*array like*) – The label information to be set into DMatrix from numpy 2D array

set_uint_info (`field`, `data`)
Set uint type property into the DMatrix.

Parameters

- **field** (*str*) – The field name of the information
- **data** (*numpy array*) – The array of data to be set

set_weight (`weight`)
Set weight of each instance.

Parameters `weight` (*array like*) – Weight for each data point

set_weight_np2d (`weight`)

Set weight of each instance for numpy 2D array

Parameters `weight` (*array like*) – Weight for each data point in numpy 2D array

slice (`rindex`)
Slice the DMatrix and return a new DMatrix that only contains *rindex*.

Parameters `rindex` (*list*) – List of indices to be selected.

Returns `res` – A new DMatrix containing only selected indices.

Return type *DMatrix*

class `xgboost.Booster` (`params=None`, `cache=()`, `model_file=None`)
Bases: `object`

A Booster of of XGBoost.

Booster is the model of xgboost, that contains low level routines for training, prediction and evaluation.

Parameters

- **params** (*dict*) – Parameters for boosters.
- **cache** (*list*) – List of cache items.
- **model_file** (*string*) – Path to the model file.

attr (`key`)
Get attribute string from the Booster.

Parameters `key` (*str*) – The key to get attribute from.

Returns `value` – The attribute value of the key, returns None if attribute do not exist.

Return type *str*

attributes ()
Get attributes stored in the Booster as a dictionary.

Returns `result` – Returns an empty dict if there's no attributes.

Return type dictionary of attribute_name: attribute_value pairs of strings.

boost (*dtrain, grad, hess*)

Boost the booster for one iteration, with customized gradient statistics.

Parameters

- **dtrain** (*DMatrix*) – The training DMatrix.
- **grad** (*list*) – The first order of gradient.
- **hess** (*list*) – The second order of gradient.

copy ()

Copy the booster object.

Returns **booster** – a copied booster model

Return type *Booster*

dump_model (*fout, fmap="", with_stats=False*)

Dump model into a text file.

Parameters

- **foout** (*string*) – Output file name.
- **fmap** (*string, optional*) – Name of the file containing feature map names.
- **with_stats** (*bool optional*) – Controls whether the split statistics are output.

eval (*data, name='eval', iteration=0*)

Evaluate the model on mat.

Parameters

- **data** (*DMatrix*) – The dmatrix storing the input.
- **name** (*str, optional*) – The name of the dataset.
- **iteration** (*int, optional*) – The current iteration number.

Returns **result** – Evaluation result string.

Return type *str*

eval_set (*evals, iteration=0, feval=None*)

Evaluate a set of data.

Parameters

- **evals** (*list of tuples (DMatrix, string)*) – List of items to be evaluated.
- **iteration** (*int*) – Current iteration.
- **feval** (*function*) – Custom evaluation function.

Returns **result** – Evaluation result string.

Return type *str*

get_dump (*fmap="", with_stats=False, dump_format='text'*)

Returns the dump the model as a list of strings.

get_fscore (*fmap=""*)

Get feature importance of each feature.

Parameters **fmap** (*str optional*) – The name of feature map file

get_score (*fmap="", importance_type='weight'*)

Get feature importance of each feature. Importance type can be defined as:

- ‘weight’: the number of times a feature is used to split the data across all trees.
- ‘gain’: the average gain across all splits the feature is used in.
- ‘cover’: the average coverage across all splits the feature is used in.
- ‘total_gain’: the total gain across all splits the feature is used in.
- ‘total_cover’: the total coverage across all splits the feature is used in.

Parameters

- **fmap** (*str* (optional)) – The name of feature map file.
- **importance_type** (*str*, default ‘weight’) – One of the importance types defined above.

get_split_value_histogram (*feature*, *fmap*=”, *bins*=None, *as_pandas*=True)

Get split value histogram of a feature

Parameters

- **feature** (*str*) – The name of the feature.
- **fmap** (*str* (optional)) – The name of feature map file.
- **bin** (*int*, default None) – The maximum number of bins. Number of bins equals number of unique split values *n_unique*, if *bins* == None or *bins* > *n_unique*.
- **as_pandas** (*bool*, default True) – Return *pd.DataFrame* when pandas is installed. If False or pandas is not installed, return *numpy ndarray*.

Returns

- *a histogram of used splitting values for the specified feature*
- *either as numpy array or pandas DataFrame.*

load_model (*fname*)

Load the model from a file.

The model is loaded from an XGBoost internal binary format which is universal among the various XGBoost interfaces. Auxiliary attributes of the Python Booster object (such as *feature_names*) will not be loaded. To preserve all attributes, pickle the Booster object.

Parameters **fname** (*string* or a *memory buffer*) – Input file name or memory buffer(see also *save_raw*)

load_rabit_checkpoint ()

Initialize the model by load from rabit checkpoint.

Returns **version** – The version number of the model.

Return type *integer*

predict (*data*, *output_margin*=False, *ntree_limit*=0, *pred_leaf*=False, *pred_contribs*=False, *approx_contribs*=False, *pred_interactions*=False, *validate_features*=True)

Predict with data.

Note: This function is not thread safe.

For each booster object, predict can only be called from one thread. If you want to run prediction using multiple thread, call `bst.copy()` to make copies of model object and then call `predict()`.

Note: Using `predict()` with DART booster

If the booster object is DART type, `predict()` will perform dropouts, i.e. only some of the trees will be evaluated. This will produce incorrect results if `data` is not the training data. To obtain correct results on test sets, set `ntree_limit` to a nonzero value, e.g.

```
preds = bst.predict(dtest, ntree_limit=num_round)
```

Parameters

- **data** (`DMatrix`) – The `dmatrix` storing the input.
- **output_margin** (`bool`) – Whether to output the raw untransformed margin value.
- **ntree_limit** (`int`) – Limit number of trees in the prediction; defaults to 0 (use all trees).
- **pred_leaf** (`bool`) – When this option is on, the output will be a matrix of (nsample, ntrees) with each record indicating the predicted leaf index of each sample in each tree. Note that the leaf index of a tree is unique per tree, so you may find leaf 1 in both tree 1 and tree 0.
- **pred_contribs** (`bool`) – When this is True the output will be a matrix of size (nsample, nfeats + 1) with each record indicating the feature contributions (SHAP values) for that prediction. The sum of all feature contributions is equal to the raw untransformed margin value of the prediction. Note the final column is the bias term.
- **approx_contribs** (`bool`) – Approximate the contributions of each feature
- **pred_interactions** (`bool`) – When this is True the output will be a matrix of size (nsample, nfeats + 1, nfeats + 1) indicating the SHAP interaction values for each pair of features. The sum of each row (or column) of the interaction values equals the corresponding SHAP value (from `pred_contribs`), and the sum of the entire matrix equals the raw untransformed margin value of the prediction. Note the last row and column correspond to the bias term.
- **validate_features** (`bool`) – When this is True, validate that the Booster's and data's `feature_names` are identical. Otherwise, it is assumed that the `feature_names` are the same.

Returns prediction

Return type numpy array

`save_model(fname)`

Save the model to a file.

The model is saved in an XGBoost internal binary format which is universal among the various XGBoost interfaces. Auxiliary attributes of the Python Booster object (such as `feature_names`) will not be saved. To preserve all attributes, pickle the Booster object.

Parameters `fname` (`string`) – Output file name

`save_rabit_checkpoint()`

Save the current booster to rabit checkpoint.

`save_raw()`

Save the model to a in memory buffer representation

Returns

Return type a in memory buffer representation of the model

set_attr (***kwargs*)

Set the attribute of the Booster.

Parameters ***kwargs* – The attributes to set. Setting a value to None deletes an attribute.

set_param (*params, value=None*)

Set parameters into the Booster.

Parameters

- **params** (*dict/list/str*) – list of key,value pairs, dict of key to value or simply str key
- **value** (*optional*) – value of the specified parameter, when params is str key

update (*dtrain, iteration, fobj=None*)

Update for one iteration, with objective function calculated internally.

Parameters

- **dtrain** (*DMatrix*) – Training data.
- **iteration** (*int*) – Current iteration number.
- **fobj** (*function*) – Customized objective function.

Learning API

Training Library containing training routines.

`xgboost.train` (*params, dtrain, num_boost_round=10, evals=(), obj=None, feval=None, maximize=False, early_stopping_rounds=None, evals_result=None, verbose_eval=True, xgb_model=None, callbacks=None, learning_rates=None*)

Train a booster with given parameters.

Parameters

- **params** (*dict*) – Booster params.
- **dtrain** (*DMatrix*) – Data to be trained.
- **num_boost_round** (*int*) – Number of boosting iterations.
- **evals** (*list of pairs (DMatrix, string)*) – List of items to be evaluated during training, this allows user to watch performance on the validation set.
- **obj** (*function*) – Customized objective function.
- **feval** (*function*) – Customized evaluation function.
- **maximize** (*bool*) – Whether to maximize feval.
- **early_stopping_rounds** (*int*) – Activates early stopping. Validation error needs to decrease at least every <early_stopping_rounds> round(s) to continue training. Requires at least one item in evals. If there's more than one, will use the last. Returns the model from the last iteration (not the best one). If early stopping occurs, the model will have three additional fields: `bst.best_score`, `bst.best_iteration` and `bst.best_ntree_limit`. (Use `bst.best_ntree_limit` to get the correct value if `num_parallel_tree` and/or `num_class` appears in the parameters)
- **evals_result** (*dict*) – This dictionary stores the evaluation results of all the items in watchlist.

Example: with a watchlist containing [(dtest,'eval'), (dtrain,'train')] and a parameter containing ('eval_metric': 'logloss'), the **evals_result** returns

```
{'train': {'logloss': ['0.48253', '0.35953']},
 'eval': {'logloss': ['0.480385', '0.357756']}}
```

- **verbose_eval** (*bool or int*) – Requires at least one item in evals. If **verbose_eval** is True then the evaluation metric on the validation set is printed at each boosting stage. If **verbose_eval** is an integer then the evaluation metric on the validation set is printed at every given **verbose_eval** boosting stage. The last boosting stage / the boosting stage found by using **early_stopping_rounds** is also printed. Example: with **verbose_eval**=4 and at least one item in evals, an evaluation metric is printed every 4 boosting stages, instead of every boosting stage.
- **learning_rates** (*list or function (deprecated - use callback API instead)*) – List of learning rate for each boosting round or a customized function that calculates eta in terms of current number of round and the total number of boosting round (e.g. yields learning rate decay)
- **xgb_model** (*file name of stored xgb model or 'Booster' instance*) – Xgb model to be loaded before training (allows training continuation).
- **callbacks** (*list of callback functions*) – List of callback functions that are applied at end of each iteration. It is possible to use predefined callbacks by using xgb.callback module. Example: [xgb.callback.reset_learning_rate(custom_rates)]

Returns booster

Return type a trained booster model

`xgboost.cv(params, dtrain, num_boost_round=10, nfold=3, stratified=False, folds=None, metrics=(), obj=None, feval=None, maximize=False, early_stopping_rounds=None, fpreproc=None, as_pandas=True, verbose_eval=None, show_stdv=True, seed=0, callbacks=None, shuffle=True)`

Cross-validation with given parameters.

Parameters

- **params** (*dict*) – Booster params.
- **dtrain** (*DMatrix*) – Data to be trained.
- **num_boost_round** (*int*) – Number of boosting iterations.
- **nfold** (*int*) – Number of folds in CV.
- **stratified** (*bool*) – Perform stratified sampling.
- **folds** (*a KFold or StratifiedKFold instance or list of fold indices*) – Sklearn KFold or StratifiedKFold object. Alternatively may explicitly pass sample indices for each fold. For *n* folds, **folds** should be a length *n* list of tuples. Each tuple is (*in*, *out*) where *in* is a list of indices to be used as the training samples for the *n*th fold and *out* is a list of indices to be used as the testing samples for the *n*th fold.
- **metrics** (*string or list of strings*) – Evaluation metrics to be watched in CV.
- **obj** (*function*) – Custom objective function.
- **feval** (*function*) – Custom evaluation function.
- **maximize** (*bool*) – Whether to maximize feval.

- **early_stopping_rounds** (*int*) – Activates early stopping. CV error needs to decrease at least every <early_stopping_rounds> round(s) to continue. Last entry in evaluation history is the one from best iteration.
- **fpreproc** (*function*) – Preprocessing function that takes (dtrain, dtest, param) and returns transformed versions of those.
- **as_pandas** (*bool, default True*) – Return pd.DataFrame when pandas is installed. If False or pandas is not installed, return np.ndarray
- **verbose_eval** (*bool, int, or None, default None*) – Whether to display the progress. If None, progress will be displayed when np.ndarray is returned. If True, progress will be displayed at boosting stage. If an integer is given, progress will be displayed at every given *verbose_eval* boosting stage.
- **show_stdv** (*bool, default True*) – Whether to display the standard deviation in progress. Results are not affected, and always contains std.
- **seed** (*int*) – Seed used to generate the folds (passed to numpy.random.seed).
- **callbacks** (*list of callback functions*) – List of callback functions that are applied at end of each iteration. It is possible to use predefined callbacks by using xgb.callback module. Example:

```
[xgb.callback.reset_learning_rate(custom_rates)]
```

- **shuffle** (*bool*) – Shuffle data before creating folds.

Returns evaluation history

Return type `list(string)`

Scikit-Learn API

Scikit-Learn Wrapper interface for XGBoost.

```
class xgboost.XGBRegressor(max_depth=3, learning_rate=0.1, n_estimators=100, silent=True,  
objective='reg:linear', booster='gbtree', n_jobs=1, nthread=None,  
gamma=0, min_child_weight=1, max_delta_step=0, subsam-  
ple=1, colsample_bytree=1, colsample_bylevel=1, reg_alpha=0,  
reg_lambda=1, scale_pos_weight=1, base_score=0.5, ran-  
dom_state=0, seed=None, missing=None, **kwargs)
```

Bases: `xgboost.sklearn.XGBModel`, `object`

Implementation of the scikit-learn API for XGBoost regression.

Parameters

- **max_depth** (*int*) – Maximum tree depth for base learners.
- **learning_rate** (*float*) – Boosting learning rate (xgb’s “eta”)
- **n_estimators** (*int*) – Number of boosted trees to fit.
- **silent** (*boolean*) – Whether to print messages while running boosting.
- **objective** (*string or callable*) – Specify the learning task and the corresponding learning objective or a custom objective function to be used (see note below).
- **booster** (*string*) – Specify which booster to use: gbtree, gblinear or dart.

- **nthread** (*int*) – Number of parallel threads used to run xgboost. (Deprecated, please use `n_jobs`)
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. (replaces `nthread`)
- **gamma** (*float*) – Minimum loss reduction required to make a further partition on a leaf node of the tree.
- **min_child_weight** (*int*) – Minimum sum of instance weight(hessian) needed in a child.
- **max_delta_step** (*int*) – Maximum delta step we allow each tree's weight estimation to be.
- **subsample** (*float*) – Subsample ratio of the training instance.
- **colsample_bytree** (*float*) – Subsample ratio of columns when constructing each tree.
- **colsample_bylevel** (*float*) – Subsample ratio of columns for each split, in each level.
- **reg_alpha** (*float* (*xgb's alpha*)) – L1 regularization term on weights
- **reg_lambda** (*float* (*xgb's lambda*)) – L2 regularization term on weights
- **scale_pos_weight** (*float*) – Balancing of positive and negative weights.
- **base_score** – The initial prediction score of all instances, global bias.
- **seed** (*int*) – Random number seed. (Deprecated, please use `random_state`)
- **random_state** (*int*) – Random number seed. (replaces `seed`)
- **missing** (*float*, *optional*) – Value in the data which needs to be present as a missing value. If None, defaults to `np.nan`.
- ****kwargs** (*dict*, *optional*) – Keyword arguments for XGBoost Booster object. Full documentation of parameters can be found here: <https://github.com/dmlc/xgboost/blob/master/doc/parameter.rst>. Attempting to set a parameter via the constructor args and ****kwargs** dict simultaneously will result in a `TypeError`.

Note: ****kwargs** unsupported by scikit-learn

****kwargs** is unsupported by scikit-learn. We do not guarantee that parameters passed via this argument will interact properly with scikit-learn.

Note: A custom objective function can be provided for the `objective` parameter. In this case, it should have the signature `objective(y_true, y_pred) -> grad, hess`:

y_true: array_like of shape `[n_samples]` The target values

y_pred: array_like of shape `[n_samples]` The predicted values

grad: array_like of shape `[n_samples]` The value of the gradient for each sample point.

hess: array_like of shape `[n_samples]` The value of the second derivative for each sample point

apply (*X*, *n_tree_limit=0*)

Return the predicted leaf every tree for each sample.

Parameters

- **X** (*array_like*, *shape*=[*n_samples*, *n_features*]) – Input features matrix.
- **ntree_limit** (*int*) – Limit number of trees in the prediction; defaults to 0 (use all trees).

Returns X_leaves – For each datapoint *x* in *X* and for each tree, return the index of the leaf *x* ends up in. Leaves are numbered within $[0; 2^{**}(\text{self.max_depth}+1))$, possibly with gaps in the numbering.

Return type *array_like*, *shape*=[*n_samples*, *n_trees*]

evals_result()

Return the evaluation results.

If *eval_set* is passed to the *fit* function, you can call *evals_result()* to get evaluation results for all passed *eval_sets*. When *eval_metric* is also passed to the *fit* function, the *evals_result* will contain the *eval_metrics* passed to the *fit* function

Returns evals_result

Return type dictionary

Example

```
param_dist = {'objective': 'binary:logistic', 'n_estimators': 2}

clf = xgb.XGBModel(**param_dist)

clf.fit(X_train, y_train,
        eval_set=[(X_train, y_train), (X_test, y_test)],
        eval_metric='logloss',
        verbose=True)

evals_result = clf.evals_result()
```

The variable *evals_result* will contain:

```
{ 'validation_0': { 'logloss': [ '0.604835', '0.531479' ] },
  'validation_1': { 'logloss': [ '0.41965', '0.17686' ] } }
```

feature_importances_

Feature importances property

Returns feature_importances_

Return type array of shape [*n_features*]

fit (*X*, *y*, *sample_weight*=None, *eval_set*=None, *eval_metric*=None, *early_stopping_rounds*=None, *verbose*=True, *xgb_model*=None, *sample_weight_eval_set*=None)

Fit the gradient boosting model

Parameters

- **X** (*array_like*) – Feature matrix
- **y** (*array_like*) – Labels
- **sample_weight** (*array_like*) – instance weights
- **eval_set** (*list*, *optional*) – A list of (*X*, *y*) tuple pairs to use as a validation set for early-stopping

- **sample_weight_eval_set** (*list, optional*) – A list of the form [L_1, L_2, ..., L_n], where each L_i is a list of instance weights on the i-th validation set.
- **eval_metric** (*str, callable, optional*) – If a str, should be a built-in evaluation metric to use. See doc/parameter.rst. If callable, a custom evaluation metric. The call signature is func(y_predicted, y_true) where y_true will be a DMatrix object such that you may need to call the get_label method. It must return a str, value pair where the str is a name for the evaluation and value is the value of the evaluation function. This objective is always minimized.
- **early_stopping_rounds** (*int*) – Activates early stopping. Validation error needs to decrease at least every <early_stopping_rounds> round(s) to continue training. Requires at least one item in evals. If there's more than one, will use the last. Returns the model from the last iteration (not the best one). If early stopping occurs, the model will have three additional fields: bst.best_score, bst.best_iteration and bst.best_ntree_limit. (Use bst.best_ntree_limit to get the correct value if num_parallel_tree and/or num_class appears in the parameters)
- **verbose** (*bool*) – If *verbose* and an evaluation set is used, writes the evaluation metric measured on the validation set to stderr.
- **xgb_model** (*str*) – file name of stored xgb model or 'Booster' instance Xgb model to be loaded before training (allows training continuation).

get_booster ()

Get the underlying xgboost Booster of this model.

This will raise an exception when fit was not called

Returns booster

Return type a xgboost booster of underlying model

get_params (*deep=False*)

Get parameters.

get_xgb_params ()

Get xgboost type parameters.

load_model (*fname*)

Load the model from a file.

Parameters fname (*string or a memory buffer*) – Input file name or memory buffer(see also save_raw)

predict (*data, output_margin=False, ntree_limit=None*)

Predict with *data*.

Note: This function is not thread safe.

For each booster object, predict can only be called from one thread. If you want to run prediction using multiple thread, call `xgb.copy()` to make copies of model object and then call `predict()`.

Note: Using `predict()` with DART booster

If the booster object is DART type, `predict()` will perform dropouts, i.e. only some of the trees will be evaluated. This will produce incorrect results if *data* is not the training data. To obtain correct results on test sets, set `ntree_limit` to a nonzero value, e.g.

```
preds = bst.predict(dtest, ntree_limit=num_round)
```

Parameters

- **data** (*DMatrix*) – The dmatrix storing the input.
- **output_margin** (*bool*) – Whether to output the raw untransformed margin value.
- **ntree_limit** (*int*) – Limit number of trees in the prediction; defaults to `best_ntree_limit` if defined (i.e. it has been trained with early stopping), otherwise 0 (use all trees).

Returns prediction

Return type numpy array

save_model (*fname*)

Save the model to a file.

Parameters **fname** (*string*) – Output file name

```
class xgboost.XGBClassifier(max_depth=3, learning_rate=0.1, n_estimators=100, silent=True,  
                           objective='binary:logistic', booster='gbtree', n_jobs=1,  
                           nthread=None, gamma=0, min_child_weight=1, max_delta_step=0,  
                           subsample=1, colsample_bytree=1, colsample_bylevel=1,  
                           reg_alpha=0, reg_lambda=1, scale_pos_weight=1, base_score=0.5,  
                           random_state=0, seed=None, missing=None, **kwargs)
```

Bases: `xgboost.sklearn.XGBModel`, `object`

Implementation of the scikit-learn API for XGBoost classification.

Parameters

- **max_depth** (*int*) – Maximum tree depth for base learners.
- **learning_rate** (*float*) – Boosting learning rate (xgb’s “eta”)
- **n_estimators** (*int*) – Number of boosted trees to fit.
- **silent** (*boolean*) – Whether to print messages while running boosting.
- **objective** (*string or callable*) – Specify the learning task and the corresponding learning objective or a custom objective function to be used (see note below).
- **booster** (*string*) – Specify which booster to use: `gbtree`, `gblinear` or `dart`.
- **nthread** (*int*) – Number of parallel threads used to run xgboost. (Deprecated, please use `n_jobs`)
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. (replaces `nthread`)
- **gamma** (*float*) – Minimum loss reduction required to make a further partition on a leaf node of the tree.
- **min_child_weight** (*int*) – Minimum sum of instance weight(hessian) needed in a child.
- **max_delta_step** (*int*) – Maximum delta step we allow each tree’s weight estimation to be.
- **subsample** (*float*) – Subsample ratio of the training instance.

- **colsample_bytree** (*float*) – Subsample ratio of columns when constructing each tree.
- **colsample_bylevel** (*float*) – Subsample ratio of columns for each split, in each level.
- **reg_alpha** (*float* (*xgb's alpha*)) – L1 regularization term on weights
- **reg_lambda** (*float* (*xgb's lambda*)) – L2 regularization term on weights
- **scale_pos_weight** (*float*) – Balancing of positive and negative weights.
- **base_score** – The initial prediction score of all instances, global bias.
- **seed** (*int*) – Random number seed. (Deprecated, please use `random_state`)
- **random_state** (*int*) – Random number seed. (replaces `seed`)
- **missing** (*float*, *optional*) – Value in the data which needs to be present as a missing value. If None, defaults to `np.nan`.
- ****kwargs** (*dict*, *optional*) – Keyword arguments for XGBoost Booster object. Full documentation of parameters can be found here: <https://github.com/dmlc/xgboost/blob/master/doc/parameter.rst>. Attempting to set a parameter via the constructor args and ****kwargs** dict simultaneously will result in a `TypeError`.

Note: ****kwargs** unsupported by scikit-learn

****kwargs** is unsupported by scikit-learn. We do not guarantee that parameters passed via this argument will interact properly with scikit-learn.

Note: A custom objective function can be provided for the `objective` parameter. In this case, it should have the signature `objective(y_true, y_pred) -> grad, hess`:

y_true: `array_like` of shape `[n_samples]` The target values

y_pred: `array_like` of shape `[n_samples]` The predicted values

grad: `array_like` of shape `[n_samples]` The value of the gradient for each sample point.

hess: `array_like` of shape `[n_samples]` The value of the second derivative for each sample point

apply (*X*, *n_tree_limit=0*)

Return the predicted leaf every tree for each sample.

Parameters

- **X** (*array_like*, *shape=[n_samples, n_features]*) – Input features matrix.
- **n_tree_limit** (*int*) – Limit number of trees in the prediction; defaults to 0 (use all trees).

Returns **X_leaves** – For each datapoint `x` in `X` and for each tree, return the index of the leaf `x` ends up in. Leaves are numbered within `[0; 2** (self.max_depth+1))`, possibly with gaps in the numbering.

Return type `array_like`, *shape=[n_samples, n_trees]*

evals_result ()

Return the evaluation results.

If `eval_set` is passed to the `fit` function, you can call `evals_result()` to get evaluation results for all passed `eval_sets`. When `eval_metric` is also passed to the `fit` function, the `evals_result` will contain the `eval_metrics` passed to the `fit` function

Returns `evals_result`

Return type dictionary

Example

```
param_dist = {'objective': 'binary:logistic', 'n_estimators': 2}

clf = xgb.XGBClassifier(**param_dist)

clf.fit(X_train, y_train,
        eval_set=[(X_train, y_train), (X_test, y_test)],
        eval_metric='logloss',
        verbose=True)

evals_result = clf.evals_result()
```

The variable `evals_result` will contain

```
{'validation_0': {'logloss': ['0.604835', '0.531479']},
 'validation_1': {'logloss': ['0.41965', '0.17686']}}
```

`feature_importances_`

Feature importances property

Returns `feature_importances_`

Return type array of shape `[n_features]`

fit (`X`, `y`, `sample_weight=None`, `eval_set=None`, `eval_metric=None`, `early_stopping_rounds=None`, `verbose=True`, `xgb_model=None`, `sample_weight_eval_set=None`)
Fit gradient boosting classifier

Parameters

- **X** (*array_like*) – Feature matrix
- **y** (*array_like*) – Labels
- **sample_weight** (*array_like*) – Weight for each instance
- **eval_set** (*list, optional*) – A list of (`X`, `y`) pairs to use as a validation set for early-stopping
- **sample_weight_eval_set** (*list, optional*) – A list of the form `[L_1, L_2, ..., L_n]`, where each `L_i` is a list of instance weights on the `i`-th validation set.
- **eval_metric** (*str, callable, optional*) – If a `str`, should be a built-in evaluation metric to use. See `doc/parameter.rst`. If callable, a custom evaluation metric. The call signature is `func(y_predicted, y_true)` where `y_true` will be a `DMatrix` object such that you may need to call the `get_label` method. It must return a `str`, value pair where the `str` is a name for the evaluation and value is the value of the evaluation function. This objective is always minimized.
- **early_stopping_rounds** (*int, optional*) – Activates early stopping. Validation error needs to decrease at least every `<early_stopping_rounds>` round(s) to continue training. Requires at least one item in `evals`. If there's more than one, will use

the last. Returns the model from the last iteration (not the best one). If early stopping occurs, the model will have three additional fields: `bst.best_score`, `bst.best_iteration` and `bst.best_ntree_limit`. (Use `bst.best_ntree_limit` to get the correct value if `num_parallel_tree` and/or `num_class` appears in the parameters)

- **verbose** (*bool*) – If *verbose* and an evaluation set is used, writes the evaluation metric measured on the validation set to stderr.
- **xgb_model** (*str*) – file name of stored xgb model or ‘Booster’ instance Xgb model to be loaded before training (allows training continuation).

get_booster()

Get the underlying xgboost Booster of this model.

This will raise an exception when fit was not called

Returns booster

Return type a xgboost booster of underlying model

get_params (*deep=False*)

Get parameters.

get_xgb_params()

Get xgboost type parameters.

load_model (*fname*)

Load the model from a file.

Parameters **fname** (*string or a memory buffer*) – Input file name or memory buffer(see also `save_raw`)

predict (*data, output_margin=False, ntree_limit=None*)

Predict with *data*.

Note: This function is not thread safe.

For each booster object, `predict` can only be called from one thread. If you want to run prediction using multiple thread, call `xgb.copy()` to make copies of model object and then call `predict()`.

Note: Using `predict()` with DART booster

If the booster object is DART type, `predict()` will perform dropouts, i.e. only some of the trees will be evaluated. This will produce incorrect results if *data* is not the training data. To obtain correct results on test sets, set `ntree_limit` to a nonzero value, e.g.

```
preds = bst.predict(dtest, ntree_limit=num_round)
```

Parameters

- **data** (*DMatrix*) – The dmatrix storing the input.
- **output_margin** (*bool*) – Whether to output the raw untransformed margin value.
- **ntree_limit** (*int*) – Limit number of trees in the prediction; defaults to `best_ntree_limit` if defined (i.e. it has been trained with early stopping), otherwise 0 (use all trees).

Returns prediction

Return type numpy array

predict_proba (*data*, *ntree_limit=None*)

Predict the probability of each *data* example being of a given class.

Note: This function is not thread safe

For each booster object, predict can only be called from one thread. If you want to run prediction using multiple thread, call `xgb.copy()` to make copies of model object and then call predict

Parameters

- **data** (*DMatrix*) – The dmatrix storing the input.
- **ntree_limit** (*int*) – Limit number of trees in the prediction; defaults to `best_ntree_limit` if defined (i.e. it has been trained with early stopping), otherwise 0 (use all trees).

Returns prediction – a numpy array with the probability of each data example being of a given class.

Return type numpy array

save_model (*fname*)

Save the model to a file.

Parameters fname (*string*) – Output file name

Plotting API

Plotting Library.

`xgboost.plot_importance` (*booster*, *ax=None*, *height=0.2*, *xlim=None*, *ylim=None*, *title='Feature importance'*, *xlabel='F score'*, *ylabel='Features'*, *importance_type='weight'*, *max_num_features=None*, *grid=True*, *show_values=True*, ***kwargs*)

Plot importance based on fitted trees.

Parameters

- **booster** (*Booster*, *XGBModel* or *dict*) – Booster or XGBModel instance, or dict taken by `Booster.get_fscore()`
- **ax** (*matplotlib Axes*, *default None*) – Target axes instance. If None, new figure and axes will be created.
- **grid** (*bool*, *Turn the axes grids on or off. Default is True (On)*) –
- **importance_type** (*str*, *default "weight"*) – How the importance is calculated: either “weight”, “gain”, or “cover”
 - “weight” is the number of times a feature appears in a tree
 - “gain” is the average gain of splits which use the feature
 - “cover” is the average coverage of splits which use the feature where coverage is defined as the number of samples affected by the split

- **max_num_features** (*int*, *default None*) – Maximum number of top features displayed on plot. If None, all features will be displayed.
- **height** (*float*, *default 0.2*) – Bar height, passed to `ax.barh()`
- **xlim** (*tuple*, *default None*) – Tuple passed to `axes.xlim()`
- **ylim** (*tuple*, *default None*) – Tuple passed to `axes.ylim()`
- **title** (*str*, *default "Feature importance"*) – Axes title. To disable, pass None.
- **xlabel** (*str*, *default "F score"*) – X axis title label. To disable, pass None.
- **ylabel** (*str*, *default "Features"*) – Y axis title label. To disable, pass None.
- **show_values** (*bool*, *default True*) – Show values on plot. To disable, pass False.
- **kwargs** – Other keywords passed to `ax.barh()`

Returns `ax`

Return type matplotlib Axes

`xgboost.plot_tree` (*booster*, *fmap=""*, *num_trees=0*, *rankdir='UT'*, *ax=None*, ***kwargs*)
Plot specified tree.

Parameters

- **booster** (*Booster*, *XGBModel*) – Booster or XGBModel instance
- **fmap** (*str* (*optional*)) – The name of feature map file
- **num_trees** (*int*, *default 0*) – Specify the ordinal number of target tree
- **rankdir** (*str*, *default "UT"*) – Passed to graphviz via `graph_attr`
- **ax** (*matplotlib Axes*, *default None*) – Target axes instance. If None, new figure and axes will be created.
- **kwargs** – Other keywords passed to `to_graphviz`

Returns `ax`

Return type matplotlib Axes

`xgboost.to_graphviz` (*booster*, *fmap=""*, *num_trees=0*, *rankdir='UT'*, *yes_color='#0000FF'*, *no_color='#FF0000'*, ***kwargs*)

Convert specified tree to graphviz instance. IPython can automatically plot the returned graphviz instance. Otherwise, you should call `.render()` method of the returned graphviz instance.

Parameters

- **booster** (*Booster*, *XGBModel*) – Booster or XGBModel instance
- **fmap** (*str* (*optional*)) – The name of feature map file
- **num_trees** (*int*, *default 0*) – Specify the ordinal number of target tree
- **rankdir** (*str*, *default "UT"*) – Passed to graphviz via `graph_attr`
- **yes_color** (*str*, *default '#0000FF'*) – Edge color when meets the node condition.
- **no_color** (*str*, *default '#FF0000'*) – Edge color when doesn't meet the node condition.
- **kwargs** – Other keywords passed to graphviz `graph_attr`

Returns `ax`

Return type matplotlib Axes

1.8 XGBoost R Package

You have found the XGBoost R Package!

1.8.1 Get Started

- Checkout the [Installation Guide](#) contains instructions to install xgboost, and [Tutorials](#) for examples on how to use XGBoost for various tasks.
- Read the [API documentation](#).
- Please visit [Walk-through Examples](#).

1.8.2 Tutorials

XGBoost R Tutorial

Introduction

Xgboost is short for e**X**treme **G**radient **B**oosting package.

The purpose of this Vignette is to show you how to use **Xgboost** to build a model and make predictions.

It is an efficient and scalable implementation of gradient boosting framework by @friedman2000additive and @friedman2001greedy. Two solvers are included:

- *linear* model ;
- *tree learning* algorithm.

It supports various objective functions, including *regression*, *classification* and *ranking*. The package is made to be extendible, so that users are also allowed to define their own objective functions easily.

It has been [used](#) to win several [Kaggle](#) competitions.

It has several features:

- Speed: it can automatically do parallel computation on *Windows* and *Linux*, with *OpenMP*. It is generally over 10 times faster than the classical `gbm`.
- Input Type: it takes several types of input data:
 - *Dense Matrix*: *R*'s *dense* matrix, i.e. `matrix` ;
 - *Sparse Matrix*: *R*'s *sparse* matrix, i.e. `Matrix::dgCMatrix` ;
 - Data File: local data files ;
 - `xgb.DMatrix`: its own class (recommended).
- Sparsity: it accepts *sparse* input for both *tree booster* and *linear booster*, and is optimized for *sparse* input ;
- Customization: it supports customized objective functions and evaluation functions.

Installation

Github version

For weekly updated version (highly recommended), install from *Github*:

```
install.packages("drat", repos="https://cran.rstudio.com")
drat::addRepo("dmlc")
install.packages("xgboost", repos="http://dmlc.ml/drat/", type = "source")
```

Windows user will need to install *Rtools* first.

CRAN version

The version 0.4-2 is on CRAN, and you can install it by:

```
install.packages("xgboost")
```

Formerly available versions can be obtained from the CRAN [archive](#)

Learning

For the purpose of this tutorial we will load **XGBoost** package.

```
require(xgboost)
```

Dataset presentation

In this example, we are aiming to predict whether a mushroom can be eaten or not (like in many tutorials, example data are the same as you will use on in your every day life :-).

Mushroom data is cited from UCI Machine Learning Repository. @Bache+Lichman:2013.

Dataset loading

We will load the *agaricus* datasets embedded with the package and will link them to variables.

The datasets are already split in:

- *train*: will be used to build the model ;
- *test*: will be used to assess the quality of our model.

Why *split* the dataset in two parts?

In the first part we will build our model. In the second part we will want to test it and assess its quality. Without dividing the dataset we would test the model on the data which the algorithm have already seen.

```
data(agaricus.train, package='xgboost')
data(agaricus.test, package='xgboost')
train <- agaricus.train
test <- agaricus.test
```

In the real world, it would be up to you to make this division between `train` and `test` data. The way to do it is out of the purpose of this article, however `caret` package may [help](#).

Each variable is a `list` containing two things, `label` and `data`:

```
str(train)

## List of 2
## $ data :Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
## .. ..@ i      : int [1:143286] 2 6 8 11 18 20 21 24 28 32 ...
## .. ..@ p      : int [1:127] 0 369 372 3306 5845 6489 6513 8380 8384 10991 ...
## .. ..@ Dim     : int [1:2] 6513 126
## .. ..@ Dimnames:List of 2
## .. .. ..$ : NULL
## .. .. ..$ : chr [1:126] "cap-shape=bell" "cap-shape=conical" "cap-shape=convex"
## .. .. ..$ : chr [1:126] "cap-shape=flat" ...
## .. ..@ x      : num [1:143286] 1 1 1 1 1 1 1 1 1 1 ...
## .. ..@ factors : list()
## $ label: num [1:6513] 1 0 0 1 0 0 0 1 0 0 ...
```

`label` is the outcome of our dataset meaning it is the binary *classification* we will try to predict.

Let's discover the dimensionality of our datasets.

```
dim(train$data)
```

```
## [1] 6513 126
```

```
dim(test$data)
```

```
## [1] 1611 126
```

This dataset is very small to not make the **R** package too heavy, however **XGBoost** is built to manage huge dataset very efficiently.

As seen below, the data are stored in a `dgCMatrix` which is a *sparse* matrix and `label` vector is a numeric vector (`{0,1}`):

```
class(train$data)[1]
```

```
## [1] "dgCMatrix"
```

```
class(train$label)
```

```
## [1] "numeric"
```

Basic Training using XGBoost

This step is the most critical part of the process for the quality of our model.

Basic training

We are using the `train` data. As explained above, both `data` and `label` are stored in a `list`.

In a *sparse* matrix, cells containing 0 are not stored in memory. Therefore, in a dataset mainly made of 0, memory size is reduced. It is very usual to have such dataset.

We will train decision tree model using the following parameters:

- `objective = "binary:logistic"`: we will train a binary classification model ;
- `max.depth = 2`: the trees won't be deep, because our case is very simple ;
- `nthread = 2`: the number of cpu threads we are going to use;
- `nrounds = 2`: there will be two passes on the data, the second one will enhance the model by further reducing the difference between ground truth and prediction.

```
bstSparse <- xgboost(data = train$data, label = train$label, max.depth = 2, eta = 1,
  ↪ nthread = 2, nrounds = 2, objective = "binary:logistic")
```

```
## [0] train-error:0.046522
## [1] train-error:0.022263
```

More complex the relationship between your features and your label is, more passes you need.

Parameter variations

Dense matrix

Alternatively, you can put your dataset in a *dense* matrix, i.e. a basic **R** matrix.

```
bstDense <- xgboost(data = as.matrix(train$data), label = train$label, max.depth = 2,
  ↪ eta = 1, nthread = 2, nrounds = 2, objective = "binary:logistic")
```

```
## [0] train-error:0.046522
## [1] train-error:0.022263
```

xgb.DMatrix

XGBoost offers a way to group them in a `xgb.DMatrix`. You can even add other meta data in it. It will be useful for the most advanced features we will discover later.

```
dtrain <- xgb.DMatrix(data = train$data, label = train$label)
bstDMatrix <- xgboost(data = dtrain, max.depth = 2, eta = 1, nthread = 2, nrounds = 2,
  ↪ objective = "binary:logistic")
```

```
## [0] train-error:0.046522
## [1] train-error:0.022263
```

Verbose option

XGBoost has several features to help you to view how the learning progress internally. The purpose is to help you to set the best parameters, which is the key of your model quality.

One of the simplest way to see the training progress is to set the `verbose` option (see below for more advanced techniques).

```
# verbose = 0, no message
bst <- xgboost(data = dtrain, max.depth = 2, eta = 1, nthread = 2, nrounds = 2,
  ↪objective = "binary:logistic", verbose = 0)
```

```
# verbose = 1, print evaluation metric
bst <- xgboost(data = dtrain, max.depth = 2, eta = 1, nthread = 2, nrounds = 2,
  ↪objective = "binary:logistic", verbose = 1)
```

```
## [0] train-error:0.046522
## [1] train-error:0.022263
```

```
# verbose = 2, also print information about tree
bst <- xgboost(data = dtrain, max.depth = 2, eta = 1, nthread = 2, nrounds = 2,
  ↪objective = "binary:logistic", verbose = 2)
```

```
## [11:41:01] amalgamation/./src/tree/updater_prune.cc:74: tree pruning end, 1 roots,
  ↪ 6 extra nodes, 0 pruned nodes, max_depth=2
## [0] train-error:0.046522
## [11:41:01] amalgamation/./src/tree/updater_prune.cc:74: tree pruning end, 1 roots,
  ↪ 4 extra nodes, 0 pruned nodes, max_depth=2
## [1] train-error:0.022263
```

Basic prediction using XGBoost

Perform the prediction

The purpose of the model we have built is to classify new data. As explained before, we will use the `test` dataset for this step.

```
pred <- predict(bst, test$data)

# size of the prediction vector
print(length(pred))
```

```
## [1] 1611
```

```
# limit display of predictions to the first 10
print(head(pred))
```

```
## [1] 0.28583017 0.92392391 0.28583017 0.28583017 0.05169873 0.92392391
```

These numbers doesn't look like *binary classification* $\{0, 1\}$. We need to perform a simple transformation before being able to use these results.

Transform the regression in a binary classification

The only thing that **XGBoost** does is a *regression*. **XGBoost** is using `label` vector to build its *regression* model.

How can we use a *regression* model to perform a binary classification?

If we think about the meaning of a regression applied to our data, the numbers we get are probabilities that a datum will be classified as 1. Therefore, we will set the rule that if this probability for a specific datum is > 0.5 then the observation is classified as 1 (or 0 otherwise).

```
prediction <- as.numeric(pred > 0.5)
print(head(prediction))
```

```
## [1] 0 1 0 0 0 1
```

Measuring model performance

To measure the model performance, we will compute a simple metric, the *average error*.

```
err <- mean(as.numeric(pred > 0.5) != test$label)
print(paste("test-error=", err))
```

```
## [1] "test-error= 0.0217256362507759"
```

Note that the algorithm has not seen the test data during the model construction.

Steps explanation:

1. `as.numeric(pred > 0.5)` applies our rule that when the probability (\Leftrightarrow regression \Leftrightarrow prediction) is > 0.5 the observation is classified as 1 and 0 otherwise ;
2. `probabilityVectorPreviouslyComputed != test$label` computes the vector of error between true data and computed probabilities ;
3. `mean(vectorOfErrors)` computes the *average error* itself.

The most important thing to remember is that **to do a classification, you just do a regression to the label and then apply a threshold**.

Multiclass classification works in a similar way.

This metric is **0.02** and is pretty low: our yummy mushroom model works well!

Advanced features

Most of the features below have been implemented to help you to improve your model by offering a better understanding of its content.

Dataset preparation

For the following advanced features, we need to put data in `xgb.DMatrix` as explained above.

```
dtrain <- xgb.DMatrix(data = train$data, label=train$label)
dtest <- xgb.DMatrix(data = test$data, label=test$label)
```

Measure learning progress with `xgb.train`

Both `xgboost` (simple) and `xgb.train` (advanced) functions train models.

One of the special feature of `xgb.train` is the capacity to follow the progress of the learning after each round. Because of the way boosting works, there is a time when having too many rounds lead to an overfitting. You can see this feature as a cousin of cross-validation method. The following techniques will help you to avoid overfitting or optimizing the learning time in stopping it as soon as possible.

One way to measure progress in learning of a model is to provide to **XGBoost** a second dataset already classified. Therefore it can learn on the first dataset and test its model on the second one. Some metrics are measured after each round during the learning.

in some way it is similar to what we have done above with the average error. The main difference is that below it was after building the model, and now it is during the construction that we measure errors.

For the purpose of this example, we use `watchlist` parameter. It is a list of `xgb.DMatrix`, each of them tagged with a name.

```
watchlist <- list(train=dtrain, test=dtest)

bst <- xgb.train(data=dtrain, max.depth=2, eta=1, nthread = 2, nrounds=2,
  watchlist=watchlist, objective = "binary:logistic")
```

```
## [0] train-error:0.046522 test-error:0.042831
## [1] train-error:0.022263 test-error:0.021726
```

XGBoost has computed at each round the same average error metric than seen above (we set `nrounds` to 2, that is why we have two lines). Obviously, the `train-error` number is related to the training dataset (the one the algorithm learns from) and the `test-error` number to the test dataset.

Both training and test error related metrics are very similar, and in some way, it makes sense: what we have learned from the training dataset matches the observations from the test dataset.

If with your own dataset you have not such results, you should think about how you divided your dataset in training and test. May be there is something to fix. Again, `caret` package may [help](#).

For a better understanding of the learning progression, you may want to have some specific metric or even use multiple evaluation metrics.

```
bst <- xgb.train(data=dtrain, max.depth=2, eta=1, nthread = 2, nrounds=2,
  watchlist=watchlist, eval.metric = "error", eval.metric = "logloss", objective =
  "binary:logistic")
```

```
## [0] train-error:0.046522 train-logloss:0.233376 test-error:0.042831 test-
  logloss:0.226686
## [1] train-error:0.022263 train-logloss:0.136658 test-error:0.021726 test-
  logloss:0.137874
```

`eval.metric` allows us to monitor two new metrics for each round, `logloss` and `error`.

Linear boosting

Until now, all the learnings we have performed were based on boosting trees. **XGBoost** implements a second algorithm, based on linear boosting. The only difference with previous command is `booster = "gblinear"` parameter (and removing `eta` parameter).

```
bst <- xgb.train(data=dtrain, booster = "gblinear", max.depth=2, nthread = 2,
  nrounds=2, watchlist=watchlist, eval.metric = "error", eval.metric = "logloss",
  objective = "binary:logistic")
```

```
## [0] train-error:0.024720 train-logloss:0.184616 test-error:0.022967 test-
↳logloss:0.184234
## [1] train-error:0.004146 train-logloss:0.069885 test-error:0.003724 test-
↳logloss:0.068081
```

In this specific case, *linear boosting* gets slightly better performance metrics than decision trees based algorithm.

In simple cases, it will happen because there is nothing better than a linear algorithm to catch a linear link. However, decision trees are much better to catch a non linear link between predictors and outcome. Because there is no silver bullet, we advise you to check both algorithms with your own datasets to have an idea of what to use.

Manipulating xgb.DMatrix

Save / Load

Like saving models, xgb.DMatrix object (which groups both dataset and outcome) can also be saved using xgb.DMatrix.save function.

```
xgb.DMatrix.save(dtrain, "dtrain.buffer")
```

```
## [1] TRUE
```

```
# to load it in, simply call xgb.DMatrix
dtrain2 <- xgb.DMatrix("dtrain.buffer")
```

```
## [11:41:01] 6513x126 matrix with 143286 entries loaded from dtrain.buffer
```

```
bst <- xgb.train(data=dtrain2, max.depth=2, eta=1, nthread = 2, nrounds=2,
↳watchlist=watchlist, objective = "binary:logistic")
```

```
## [0] train-error:0.046522 test-error:0.042831
## [1] train-error:0.022263 test-error:0.021726
```

Information extraction

Information can be extracted from xgb.DMatrix using getinfo function. Hereafter we will extract label data.

```
label = getinfo(dtest, "label")
pred <- predict(bst, dtest)
err <- as.numeric(sum(as.integer(pred > 0.5) != label))/length(label)
print(paste("test-error=", err))
```

```
## [1] "test-error= 0.0217256362507759"
```

View feature importance/influence from the learnt model

Feature importance is similar to R gbm package's relative influence (rel.inf).

```
importance_matrix <- xgb.importance(model = bst)
print(importance_matrix)
xgb.plot.importance(importance_matrix = importance_matrix)
```

View the trees from a model

You can dump the tree you learned using `xgb.dump` into a text file.

```
xgb.dump(bst, with.stats = T)
```

```
## [1] "booster[0]"
## [2] "0:[f28<-1.00136e-05] yes=1,no=2,missing=1,gain=4000.53,cover=1628.25"
## [3] "1:[f55<-1.00136e-05] yes=3,no=4,missing=3,gain=1158.21,cover=924.5"
## [4] "3:leaf=1.71218,cover=812"
## [5] "4:leaf=-1.70044,cover=112.5"
## [6] "2:[f108<-1.00136e-05] yes=5,no=6,missing=5,gain=198.174,cover=703.75"
## [7] "5:leaf=-1.94071,cover=690.5"
## [8] "6:leaf=1.85965,cover=13.25"
## [9] "booster[1]"
## [10] "0:[f59<-1.00136e-05] yes=1,no=2,missing=1,gain=832.545,cover=788.852"
## [11] "1:[f28<-1.00136e-05] yes=3,no=4,missing=3,gain=569.725,cover=768.39"
## [12] "3:leaf=0.784718,cover=458.937"
## [13] "4:leaf=-0.96853,cover=309.453"
## [14] "2:leaf=-6.23624,cover=20.4624"
```

You can plot the trees from your model using “`xgb.plot.tree`”

```
xgb.plot.tree(model = bst)
```

if you provide a path to `fname` parameter you can save the trees to your hard drive.

Save and load models

Maybe your dataset is big, and it takes time to train a model on it? Maybe you are not a big fan of losing time in redoing the same task again and again? In these very rare cases, you will want to save your model and load it when required.

Hopefully for you, **XGBoost** implements such functions.

```
# save model to binary local file
xgb.save(bst, "xgboost.model")
```

```
## [1] TRUE
```

`xgb.save` function should return `TRUE` if everything goes well and crashes otherwise.

An interesting test to see how identical our saved model is to the original one would be to compare the two predictions.

```
# load binary model to R
bst2 <- xgb.load("xgboost.model")
pred2 <- predict(bst2, test$data)

# And now the test
print(paste("sum(abs(pred2-pred))=", sum(abs(pred2-pred))))
```

```
## [1] "sum(abs(pred2-pred))= 0"
```

result is 0? We are good!

In some very specific cases, like when you want to pilot **XGBoost** from `caret` package, you will want to save the model as a R binary vector. See below how to do it.

```
# save model to R's raw vector
rawVec <- xgb.save.raw(bst)

# print class
print(class(rawVec))
```

```
## [1] "raw"
```

```
# load binary model to R
bst3 <- xgb.load(rawVec)
pred3 <- predict(bst3, test$data)

# pred2 should be identical to pred
print(paste("sum(abs(pred3-pred))=", sum(abs(pred2-pred))))
```

```
## [1] "sum(abs(pred3-pred))= 0"
```

Again 0? It seems that XGBoost works pretty well!

References

Understand your dataset with XGBoost

Introduction

The purpose of this Vignette is to show you how to use **Xgboost** to discover and understand your own dataset better.

This Vignette is not about predicting anything (see [Xgboost presentation](#)). We will explain how to use **Xgboost** to highlight the *link* between the *features* of your data and the *outcome*.

Package loading:

```
require(xgboost)
require(Matrix)
require(data.table)
if (!require('vcd')) install.packages('vcd')
```

VCD package is used for one of its embedded dataset only.

Preparation of the dataset

Numeric VS categorical variables

Xgboost manages only numeric vectors.

What to do when you have *categorical* data?

A *categorical* variable has a fixed number of different values. For instance, if a variable called *Colour* can have only one of these three values, *red*, *blue* or *green*, then *Colour* is a *categorical* variable.

In **R**, a *categorical* variable is called `factor`.

Type `?factor` in the console for more information.

To answer the question above we will convert *categorical* variables to `numeric` one.

Conversion from categorical to numeric variables

Looking at the raw data

In this Vignette we will see how to transform a *dense* `data.frame` (*dense* = few zeroes in the matrix) with *categorical* variables to a very *sparse* matrix (*sparse* = lots of zero in the matrix) of numeric features.

The method we are going to see is usually called *one-hot encoding*.

The first step is to load *Arthritis* dataset in memory and wrap it with `data.table` package.

```
data(Arthritis)
df <- data.table(Arthritis, keep.rownames = F)
```

`data.table` is 100% compliant with **R** `data.frame` but its syntax is more consistent and its performance for large dataset is *best in class* (`dplyr` from **R** and `Pandas` from **Python** included). Some parts of **Xgboost R** package use `data.table`.

The first thing we want to do is to have a look to the first lines of the `data.table`:

```
head(df)
```

```
##      ID Treatment  Sex Age Improved
## 1:  57   Treated Male  27     Some
## 2:  46   Treated Male  29     None
## 3:  77   Treated Male  30     None
## 4:  17   Treated Male  32   Marked
## 5:  36   Treated Male  46   Marked
## 6:  23   Treated Male  58   Marked
```

Now we will check the format of each column.

```
str(df)
```

```
## Classes 'data.table' and 'data.frame':  84 obs. of  5 variables:
## $ ID      : int  57 46 77 17 36 23 75 39 33 55 ...
## $ Treatment: Factor w/ 2 levels "Placebo","Treated": 2 2 2 2 2 2 2 2 2 2 ...
## $ Sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
## $ Age      : int  27 29 30 32 46 58 59 59 63 63 ...
## $ Improved : Ord.factor w/ 3 levels "None"<"Some"<".": 2 1 1 3 3 1 3 1 1 ...
## - attr(*, ".internal.selfref")=<externalptr>
```

2 columns have factor type, one has ordinal type.

ordinal variable :

- can take a limited number of values (like factor) ;
- these values are ordered (unlike factor). Here these ordered values are: Marked > Some > None

Creation of new features based on old ones

We will add some new *categorical* features to see if it helps.

Grouping per 10 years

For the first feature we create groups of age by rounding the real age.

Note that we transform it to `factor` so the algorithm treat these age groups as independent values.

Therefore, 20 is not closer to 30 than 60. To make it short, the distance between ages is lost in this transformation.

```
head(df[,AgeDiscret := as.factor(round(Age/10,0))])
```

```
##      ID Treatment  Sex Age Improved AgeDiscret
## 1:  57   Treated Male  27   Some           3
## 2:  46   Treated Male  29   None           3
## 3:  77   Treated Male  30   None           3
## 4:  17   Treated Male  32   Marked          3
## 5:  36   Treated Male  46   Marked          5
## 6:  23   Treated Male  58   Marked          6
```

Random split in two groups

Following is an even stronger simplification of the real age with an arbitrary split at 30 years old. I choose this value **based on nothing**. We will see later if simplifying the information based on arbitrary values is a good strategy (you may already have an idea of how well it will work...).

```
head(df[,AgeCat:= as.factor(ifelse(Age > 30, "Old", "Young"))])
```

```
##      ID Treatment  Sex Age Improved AgeDiscret AgeCat
## 1:  57   Treated Male  27   Some           3   Young
## 2:  46   Treated Male  29   None           3   Young
## 3:  77   Treated Male  30   None           3   Young
## 4:  17   Treated Male  32   Marked          3    Old
## 5:  36   Treated Male  46   Marked          5    Old
## 6:  23   Treated Male  58   Marked          6    Old
```

Risks in adding correlated features

These new features are highly correlated to the `Age` feature because they are simple transformations of this feature.

For many machine learning algorithms, using correlated features is not a good idea. It may sometimes make prediction less accurate, and most of the time make interpretation of the model almost impossible. GLM, for instance, assumes that the features are uncorrelated.

Fortunately, decision tree algorithms (including boosted trees) are very robust to these features. Therefore we have nothing to do to manage this situation.

Cleaning data

We remove ID as there is nothing to learn from this feature (it would just add some noise).

```
df[, ID:=NULL]
```

We will list the different values for the column `Treatment`:

```
levels(df[, Treatment])
```

```
## [1] "Placebo" "Treated"
```

One-hot encoding

Next step, we will transform the categorical data to dummy variables. This is the **one-hot encoding** step.

The purpose is to transform each value of each *categorical* feature in a *binary* feature $\{0, 1\}$.

For example, the column `Treatment` will be replaced by two columns, `Placebo`, and `Treated`. Each of them will be *binary*. Therefore, an observation which has the value `Placebo` in column `Treatment` before the transformation will have after the transformation the value 1 in the new column `Placebo` and the value 0 in the new column `Treated`. The column `Treatment` will disappear during the one-hot encoding.

Column `Improved` is excluded because it will be our label column, the one we want to predict.

```
sparse_matrix <- sparse.model.matrix(Improved~.-1, data = df)
head(sparse_matrix)
```

```
## 6 x 10 sparse Matrix of class "dgCMatrix"
##
## 1 . 1 1 27 1 . . . . 1
## 2 . 1 1 29 1 . . . . 1
## 3 . 1 1 30 1 . . . . 1
## 4 . 1 1 32 1 . . . . .
## 5 . 1 1 46 . . 1 . . .
## 6 . 1 1 58 . . . 1 . .
```

Formulae `Improved~.-1` used above means transform all *categorical* features but column `Improved` to binary values. The `-1` is here to remove the first column which is full of 1 (this column is generated by the conversion). For more information, you can type `?sparse.model.matrix` in the console.

Create the output numeric vector (not as a sparse Matrix):

```
output_vector = df[, Improved] == "Marked"
```

1. set Y vector to 0;
2. set Y to 1 for rows where `Improved == Marked` is TRUE;
3. return Y vector.

Build the model

The code below is very usual. For more information, you can look at the documentation of `xgboost` function (or at the vignette [Xgboost presentation](#)).

```
bst <- xgboost(data = sparse_matrix, label = output_vector, max.depth = 4,
              eta = 1, nthread = 2, nrounds = 10, objective = "binary:logistic")
```



```
## [0] train-error:0.202381
## [1] train-error:0.166667
## [2] train-error:0.166667
## [3] train-error:0.166667
## [4] train-error:0.154762
## [5] train-error:0.154762
## [6] train-error:0.154762
## [7] train-error:0.166667
## [8] train-error:0.166667
## [9] train-error:0.166667
```

You can see some `train-error: 0.XXXXX` lines followed by a number. It decreases. Each line shows how well the model explains your data. Lower is better.

A model which fits too well may **overfit** (meaning it copy/paste too much the past, and won't be that good to predict the future).

Here you can see the numbers decrease until line 7 and then increase.

It probably means we are overfitting. To fix that I should reduce the number of rounds to `nrounds = 4`. I will let things like that because I don't really care for the purpose of this example :-)

Feature importance

Measure feature importance

Build the feature importance data.table

In the code below, `sparse_matrix@Dimnames[[2]]` represents the column names of the sparse matrix. These names are the original values of the features (remember, each binary column == one value of one *categorical* feature).

```
importance <- xgb.importance(feature_names = sparse_matrix@Dimnames[[2]], model = bst)
head(importance)
```

```
##           Feature      Gain      Cover Frequency
## 1:           Age 0.622031651 0.67251706 0.67241379
## 2: TreatmentPlacebo 0.285750607 0.11916656 0.10344828
## 3:           SexMale 0.048744054 0.04522027 0.08620690
## 4:      AgeDiscret6 0.016604647 0.04784637 0.05172414
## 5:      AgeDiscret3 0.016373791 0.08028939 0.05172414
## 6:      AgeDiscret4 0.009270558 0.02858801 0.01724138
```

The column `Gain` provide the information we are looking for.

As you can see, features are classified by `Gain`.

`Gain` is the improvement in accuracy brought by a feature to the branches it is on. The idea is that before adding a new split on a feature `X` to the branch there was some wrongly classified elements, after adding the split on this feature, there are two new branches, and each of these branch is more accurate (one branch saying if your observation is on this branch then it should be classified as 1, and the other branch saying the exact opposite).

`Cover` measures the relative quantity of observations concerned by a feature.

`Frequency` is a simpler way to measure the `Gain`. It just counts the number of times a feature is used in all generated trees. You should not use it (unless you know why you want to use it).

Improvement in the interpretability of feature importance data.table

We can go deeper in the analysis of the model. In the `data.table` above, we have discovered which features counts to predict if the illness will go or not. But we don't yet know the role of these features. For instance, one of the question we may want to answer would be: does receiving a placebo treatment helps to recover from the illness?

One simple solution is to count the co-occurrences of a feature and a class of the classification.

For that purpose we will execute the same function as above but using two more parameters, `data` and `label`.

```
importanceRaw <- xgb.importance(feature_names = sparse_matrix@Dimnames[[2]], model = xgb
  bst, data = sparse_matrix, label = output_vector)

# Cleaning for better display
importanceClean <- importanceRaw[, `:=` (Cover=NULL, Frequency=NULL)]

head(importanceClean)
```

##	Feature	Split	Gain	RealCover	RealCover %
## 1:	TreatmentPlacebo	-1.00136e-05	0.28575061	7	0.2500000
## 2:	Age	61.5	0.16374034	12	0.4285714
## 3:	Age	39	0.08705750	8	0.2857143
## 4:	Age	57.5	0.06947553	11	0.3928571
## 5:	SexMale	-1.00136e-05	0.04874405	4	0.1428571
## 6:	Age	53.5	0.04620627	10	0.3571429

In the table above we have removed two not needed columns and select only the first lines.

First thing you notice is the new column `Split`. It is the split applied to the feature on a branch of one of the tree. Each split is present, therefore a feature can appear several times in this table. Here we can see the feature `Age` is used several times with different splits.

How the split is applied to count the co-occurrences? It is always `<`. For instance, in the second line, we measure the number of persons under 61.5 years with the illness gone after the treatment.

The two other new columns are `RealCover` and `RealCover %`. In the first column it measures the number of observations in the dataset where the split is respected and the label marked as 1. The second column is the percentage of the whole population that `RealCover` represents.

Therefore, according to our findings, getting a placebo doesn't seem to help but being younger than 61 years may help (seems logic).

You may wonder how to interpret the `< 1.00001` on the first line. Basically, in a sparse Matrix, there is no 0, therefore, looking for one hot-encoded categorical observations validating the rule `< 1.00001` is like just looking for 1 for this feature.

Plotting the feature importance

All these things are nice, but it would be even better to plot the results.

```
xgb.plot.importance(importance_matrix = importanceRaw)
```

```
## Error in xgb.plot.importance(importance_matrix = importanceRaw): Importance matrix
  is not correct (column names issue)
```

Feature have automatically been divided in 2 clusters: the interesting features... and the others.

Depending of the dataset and the learning parameters you may have more than two clusters. Default value is to limit them to 10, but you can increase this limit. Look at the function documentation for more information.

According to the plot above, the most important features in this dataset to predict if the treatment will work are :

- the Age ;
- having received a placebo or not ;
- the sex is third but already included in the not interesting features group ;
- then we see our generated features (AgeDiscret). We can see that their contribution is very low.

Do these results make sense?

Let's check some **Chi2** between each of these features and the label.

Higher **Chi2** means better correlation.

```
c2 <- chisq.test(df$Age, output_vector)
print(c2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$Age and output_vector
## X-squared = 35.475, df = 35, p-value = 0.4458
```

Pearson correlation between Age and illness disappearing is **35.48**.

```
c2 <- chisq.test(df$AgeDiscret, output_vector)
print(c2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$AgeDiscret and output_vector
## X-squared = 8.2554, df = 5, p-value = 0.1427
```

Our first simplification of Age gives a Pearson correlation is **8.26**.

```
c2 <- chisq.test(df$AgeCat, output_vector)
print(c2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$AgeCat and output_vector
## X-squared = 2.3571, df = 1, p-value = 0.1247
```

The perfectly random split I did between young and old at 30 years old have a low correlation of **2.36**. It's a result we may expect as may be in my mind > 30 years is being old (I am 32 and starting feeling old, this may explain that), but for the illness we are studying, the age to be vulnerable is not the same.

Morality: don't let your *gut* lower the quality of your model.

In *data science* expression, there is the word *science* :-)

Conclusion

As you can see, in general *destroying information by simplifying it won't improve your model*. **Chi2** just demonstrates that.

But in more complex cases, creating a new feature based on existing one which makes link with the outcome more obvious may help the algorithm and improve the model.

The case studied here is not enough complex to show that. Check [Kaggle website](#) for some challenging datasets. However it's almost always worse when you add some arbitrary rules.

Moreover, you can notice that even if we have added some not useful new features highly correlated with other features, the boosting tree algorithm have been able to choose the best one, which in this case is the Age.

Linear models may not be that smart in this scenario.

Special Note: What about Random Forests™?

As you may know, **Random Forests™** algorithm is cousin with boosting and both are part of the **ensemble learning** family.

Both train several decision trees for one dataset. The *main* difference is that in Random Forests™, trees are independent and in boosting, the tree N+1 focus its learning on the loss (\Leftrightarrow what has not been well modeled by the tree N).

This difference have an impact on a corner case in feature importance analysis: the *correlated features*.

Imagine two features perfectly correlated, feature A and feature B. For one specific tree, if the algorithm needs one of them, it will choose randomly (true in both boosting and Random Forests™).

However, in Random Forests™ this random choice will be done for each tree, because each tree is independent from the others. Therefore, approximatively, depending of your parameters, 50% of the trees will choose feature A and the other 50% will choose feature B. So the *importance* of the information contained in A and B (which is the same, because they are perfectly correlated) is diluted in A and B. So you won't easily know this information is important to predict what you want to predict! It is even worse when you have 10 correlated features...

In boosting, when a specific link between feature and outcome have been learned by the algorithm, it will try to not refocus on it (in theory it is what happens, reality is not always that simple). Therefore, all the importance will be on feature A or on feature B (but not both). You will know that one feature have an important role in the link between the observations and the label. It is still up to you to search for the correlated features to the one detected as important if you need to know all of them.

If you want to try Random Forests™ algorithm, you can tweak Xgboost parameters!

Warning: this is still an experimental parameter.

For instance, to compute a model with 1000 trees, with a 0.5 factor on sampling rows and columns:

```
data(agaricus.train, package='xgboost')
data(agaricus.test, package='xgboost')
train <- agaricus.train
test <- agaricus.test

#Random Forest™ - 1000 trees
bst <- xgboost(data = train$data, label = train$label, max.depth = 4, num_parallel_
↪tree = 1000, subsample = 0.5, colsample_bytree = 0.5, nrounds = 1, objective =
↪"binary:logistic")
```

```
## [0] train-error:0.002150
```

```
#Boosting - 3 rounds
bst <- xgboost(data = train$data, label = train$label, max.depth = 4, nrounds = 3,
  ↪objective = "binary:logistic")
```

```
## [0]  train-error:0.006142
## [1]  train-error:0.006756
## [2]  train-error:0.001228
```

Note that the parameter `round` is set to 1.

Random Forests™ is a trademark of Leo Breiman and Adele Cutler and is licensed exclusively to Salford Systems for the commercial release of the software.

1.9 XGBoost JVM Package

You have found the XGBoost JVM Package!

1.9.1 Installation

Installation from source

Building XGBoost4J using Maven requires Maven 3 or newer, Java 7+ and CMake 3.2+ for compiling the JNI bindings.

Before you install XGBoost4J, you need to define environment variable `JAVA_HOME` as your JDK directory to ensure that your compiler can find `jni.h` correctly, since XGBoost4J relies on JNI to implement the interaction between the JVM and native libraries.

After your `JAVA_HOME` is defined correctly, it is as simple as run `mvn package` under `jvm-packages` directory to install XGBoost4J. You can also skip the tests by running `mvn -DskipTests=true package`, if you are sure about the correctness of your local setup.

To publish the artifacts to your local maven repository, run

```
mvn install
```

Or, if you would like to skip tests, run

```
mvn -DskipTests install
```

This command will publish the xgboost binaries, the compiled java classes as well as the java sources to your local repository. Then you can use XGBoost4J in your Java projects by including the following dependency in `pom.xml`:

```
<dependency>
  <groupId>ml.dmlc</groupId>
  <artifactId>xgboost4j</artifactId>
  <version>latest_source_version_num</version>
</dependency>
```

For sbt, please add the repository and dependency in `build.sbt` as following:

```
resolvers += "Local Maven Repository" at "file://" + Path.userHome.absolutePath + "/.m2/
↳ repository"

"ml.dmlc" % "xgboost4j" % "latest_source_version_num"
```

If you want to use XGBoost4J-Spark, replace `xgboost4j` with `xgboost4j-spark`.

Note: XGBoost4J-Spark requires Spark 2.3+

XGBoost4J-Spark now requires Spark 2.3+. Latest versions of XGBoost4J-Spark uses facilities of *org.apache.spark.ml.param.shared* extensively to provide for a tight integration with Spark MLLIB framework, and these facilities are not fully available on earlier versions of Spark.

Installation from maven repo

Access release version

Listing 8: maven

```
<dependency>
  <groupId>ml.dmlc</groupId>
  <artifactId>xgboost4j</artifactId>
  <version>latest_version_num</version>
</dependency>
```

Listing 9: sbt

```
"ml.dmlc" % "xgboost4j" % "latest_version_num"
```

This will checkout the latest stable version from the Maven Central.

For the latest release version number, please check [here](#).

if you want to use XGBoost4J-Spark, replace `xgboost4j` with `xgboost4j-spark`.

Access SNAPSHOT version

You need to add GitHub as repo:

Listing 10: maven

```
<repository>
  <id>GitHub Repo</id>
  <name>GitHub Repo</name>
  <url>https://raw.githubusercontent.com/CodingCat/xgboost/maven-repo/</url>
</repository>
```

Listing 11: sbt

```
resolvers += "GitHub Repo" at "https://raw.githubusercontent.com/CodingCat/xgboost/
↳ maven-repo/"
```

Then add dependency as following:

Listing 12: maven

```
<dependency>
  <groupId>ml.dmlc</groupId>
  <artifactId>xgboost4j</artifactId>
  <version>latest_version_num</version>
</dependency>
```

Listing 13: sbt

```
"ml.dmlc" % "xgboost4j" % "latest_version_num"
```

For the latest release version number, please check [here](#).

Note: Windows not supported by published JARs

The published JARs from the Maven Central and GitHub currently only supports Linux and MacOS. Windows users should consider building XGBoost4J / XGBoost4J-Spark from the source. Alternatively, checkout pre-built JARs from [criteo-forks/xgboost-jars](#).

Enabling OpenMP for Mac OS

If you are on Mac OS and using a compiler that supports OpenMP, you need to go to the file `xgboost/jvm-packages/create_jni.py` and comment out the line

```
CONFIG["USE_OPENMP"] = "OFF"
```

in order to get the benefit of multi-threading.

1.9.2 Contents

Getting Started with XGBoost4J

This tutorial introduces Java API for XGBoost.

Data Interface

Like the XGBoost python module, XGBoost4J uses DMatrix to handle data. LIBSVM txt format file, sparse matrix in CSR/CSC format, and dense matrix are supported.

- The first step is to import DMatrix:

```
import ml.dmlc.xgboost4j.java.DMatrix;
```

- Use DMatrix constructor to load data from a libsvm text format file:

```
DMatrix dmat = new DMatrix("train.svm.txt");
```

- Pass arrays to DMatrix constructor to load from sparse matrix.

Suppose we have a sparse matrix

```
1 0 2 0
4 0 0 3
3 1 2 0
```

We can express the sparse matrix in Compressed Sparse Row (CSR) format:

```
long[] rowHeaders = new long[] {0,2,4,7};
float[] data = new float[] {1f,2f,4f,3f,3f,1f,2f};
int[] colIndex = new int[] {0,2,0,3,0,1,2};
int numColumn = 4;
DMatrix dmat = new DMatrix(rowHeaders, colIndex, data, DMatrix.SparseType.CSR,
    ↪numColumn);
```

... or in Compressed Sparse Column (CSC) format:

```
long[] colHeaders = new long[] {0,3,4,6,7};
float[] data = new float[] {1f,4f,3f,1f,2f,2f,3f};
int[] rowIndex = new int[] {0,1,2,2,0,2,1};
int numRows = 3;
DMatrix dmat = new DMatrix(colHeaders, rowIndex, data, DMatrix.SparseType.CSC,
    ↪numRows);
```

- You may also load your data from a dense matrix. Let's assume we have a matrix of form

```
1    2
3    4
5    6
```

Using row-major layout, we specify the dense matrix as follows:

```
float[] data = new float[] {1f,2f,3f,4f,5f,6f};
int nrow = 3;
int ncol = 2;
float missing = 0.0f;
DMatrix dmat = new DMatrix(data, nrow, ncol, missing);
```

- To set weight:

```
float[] weights = new float[] {1f,2f,1f};
dmat.setWeight(weights);
```

Setting Parameters

To set parameters, parameters are specified as a Map:

```
Map<String, Object> params = new HashMap<String, Object>() {
    {
        put("eta", 1.0);
        put("max_depth", 2);
        put("silent", 1);
        put("objective", "binary:logistic");
        put("eval_metric", "logloss");
    }
};
```


Training Model

With parameters and data, you are able to train a booster model.

- Import Booster and XGBoost:

```
import ml.dmlc.xgboost4j.java.Booster;
import ml.dmlc.xgboost4j.java.XGBoost;
```

- Training

```
DMatrix trainMat = new DMatrix("train.svm.txt");
DMatrix validMat = new DMatrix("valid.svm.txt");
// Specify a watch list to see model accuracy on data sets
Map<String, DMatrix> watches = new HashMap<String, DMatrix>() {
    {
        put("train", trainMat);
        put("test", testMat);
    }
};
int nround = 2;
Booster booster = XGBoost.train(trainMat, params, nround, watches, null, null);
```

- Saving model

After training, you can save model and dump it out.

```
booster.saveModel("model.bin");
```

- Generating model dump with feature map

```
// dump without feature map
String[] model_dump = booster.getModelDump(null, false);
// dump with feature map
String[] model_dump_with_feature_map = booster.getModelDump("featureMap.txt",
↳ false);
```

- Load a model

```
Booster booster = XGBoost.loadModel("model.bin");
```

Prediction

After training and loading a model, you can use it to make prediction for other data. The result will be a two-dimension float array (nsample, nclass); for predictLeaf(), the result would be of shape (nsample, nclass*ntrees).

```
DMatrix dtest = new DMatrix("test.svm.txt");
// predict
float[][] predicts = booster.predict(dtest);
// predict leaf
float[][] leafPredicts = booster.predictLeaf(dtest, 0);
```

XGBoost4J-Spark Tutorial (version 0.8+)

XGBoost4J-Spark is a project aiming to seamlessly integrate XGBoost and Apache Spark by fitting XGBoost to Apache Spark's MLLIB framework. With the integration, user can not only uses the high-performant algorithm implementation of XGBoost, but also leverages the powerful data processing engine of Spark for:

- Feature Engineering: feature extraction, transformation, dimensionality reduction, and selection, etc.
- Pipelines: constructing, evaluating, and tuning ML Pipelines
- Persistence: persist and load machine learning models and even whole Pipelines

This tutorial is to cover the end-to-end process to build a machine learning pipeline with XGBoost4J-Spark. We will discuss

- Using Spark to preprocess data to fit to XGBoost/XGBoost4J-Spark's data interface
- Training a XGBoost model with XGBoost4J-Spark
- Serving XGBoost model (prediction) with Spark
- Building a Machine Learning Pipeline with XGBoost4J-Spark
- Running XGBoost4J-Spark in Production

- *Build an ML Application with XGBoost4J-Spark*
 - *Refer to XGBoost4J-Spark Dependency*
 - *Data Preparation*
 - * *Read Dataset with Spark's Built-In Reader*
 - * *Transform Raw Iris Dataset*
 - *Training*
 - *Prediction*
 - * *Batch Prediction*
 - * *Single instance prediction*
 - *Model Persistence*
 - * *Model and pipeline persistence*
 - * *Interact with Other Bindings of XGBoost*
- *Building a ML Pipeline with XGBoost4J-Spark*
 - *Basic ML Pipeline*
 - *Pipeline with Hyper-parameter Tunning*
- *Run XGBoost4J-Spark in Production*
 - *Parallel/Distributed Training*
 - *Gang Scheduling*
 - *Checkpoint During Training*

Build an ML Application with XGBoost4J-Spark

Refer to XGBoost4J-Spark Dependency

Before we go into the tour of how to use XGBoost4J-Spark, we would bring a brief introduction about how to build a machine learning application with XGBoost4J-Spark. The first thing you need to do is to refer to the dependency in Maven Central.

You can add the following dependency in your `pom.xml`.

```
<dependency>
  <groupId>ml.dmlc</groupId>
  <artifactId>xgboost4j-spark</artifactId>
  <version>latest_version_num</version>
</dependency>
```

For the latest release version number, please check [here](#).

We also publish some functionalities which would be included in the coming release in the form of snapshot version. To access these functionalities, you can add dependency to the snapshot artifacts. We publish snapshot version in github-based repo, so you can add the following repo in `pom.xml`:

```
<repository>
  <id>XGBoost4J-Spark Snapshot Repo</id>
  <name>XGBoost4J-Spark Snapshot Repo</name>
  <url>https://raw.githubusercontent.com/CodingCat/xgboost/maven-repo</url>
</repository>
```

and then refer to the snapshot dependency by adding:

```
<dependency>
  <groupId>ml.dmlc</groupId>
  <artifactId>xgboost4j</artifactId>
  <version>next_version_num-SNAPSHOT</version>
</dependency>
```

Note: XGBoost4J-Spark requires Spark 2.3+

XGBoost4J-Spark now requires Spark 2.3+. Latest versions of XGBoost4J-Spark uses facilities of *org.apache.spark.ml.param.shared* extensively to provide for a tight integration with Spark MLLIB framework, and these facilities are not fully available on earlier versions of Spark.

Data Preparation

As aforementioned, XGBoost4J-Spark seamlessly integrates Spark and XGBoost. The integration enables users to apply various types of transformation over the training/test datasets with the convenient and powerful data processing framework, Spark.

In this section, we use [Iris](#) dataset as an example to showcase how we use Spark to transform raw dataset and make it fit to the data interface of XGBoost.

Iris dataset is shipped in CSV format. Each instance contains 4 features, “sepal length”, “sepal width”, “petal length” and “petal width”. In addition, it contains the “class” column, which is essentially the label with three possible values: “Iris Setosa”, “Iris Versicolour” and “Iris Virginica”.

Read Dataset with Spark's Built-In Reader

The first thing in data transformation is to load the dataset as Spark's structured data abstraction, `DataFrame`.

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.types.{DoubleType, StringType, StructField, StructType}

val spark = SparkSession.builder().getOrCreate()
val schema = new StructType(Array(
  StructField("sepal length", DoubleType, true),
  StructField("sepal width", DoubleType, true),
  StructField("petal length", DoubleType, true),
  StructField("petal width", DoubleType, true),
  StructField("class", StringType, true))
val rawInput = spark.read.schema(schema).csv("input_path")
```

At the first line, we create an instance of `SparkSession` which is the entry of any Spark program working with `DataFrame`. The `schema` variable defines the schema of `DataFrame` wrapping Iris data. With this explicitly set schema, we can define the columns' name as well as their types; otherwise the column name would be the default ones derived by Spark, such as `_col0`, etc. Finally, we can use Spark's built-in csv reader to load Iris csv file as a `DataFrame` named `rawInput`.

Spark also contains many built-in readers for other format. The latest version of Spark supports CSV, JSON, Parquet, and LIBSVM.

Transform Raw Iris Dataset

To make Iris dataset be recognizable to XGBoost, we need to

1. Transform String-typed label, i.e. "class", to Double-typed label.
2. Assemble the feature columns as a vector to fit to the data interface of Spark ML framework.

To convert String-typed label to Double, we can use Spark's built-in feature transformer `StringIndexer`.

```
import org.apache.spark.ml.feature.StringIndexer
val stringIndexer = new StringIndexer().
  setInputCol("class").
  setOutputCol("classIndex").
  fit(rawInput)
val labelTransformed = stringIndexer.transform(rawInput).drop("class")
```

With a newly created `StringIndexer` instance:

1. we set input column, i.e. the column containing String-typed label
2. we set output column, i.e. the column to contain the Double-typed label.
3. Then we `fit` `StringIndexer` with our input `DataFrame` `rawInput`, so that Spark internals can get information like total number of distinct values, etc.

Now we have a `StringIndexer` which is ready to be applied to our input `DataFrame`. To execute the transformation logic of `StringIndexer`, we `transform` the input `DataFrame` `rawInput` and to keep a concise `DataFrame`, we drop the column "class" and only keeps the feature columns and the transformed Double-typed label column (in the last line of the above code snippet).

The `fit` and `transform` are two key operations in MLLIB. Basically, `fit` produces a "transformer", e.g. `StringIndexer`, and each transformer applies `transform` method on `DataFrame` to add new column(s) containing transformed

features/labels or prediction results, etc. To understand more about `fit` and `transform`, You can find more details in [here](#).

Similarly, we can use another transformer, `VectorAssembler`, to assemble feature columns “sepal length”, “sepal width”, “petal length” and “petal width” as a vector.

```
import org.apache.spark.ml.feature.VectorAssembler
val vectorAssembler = new VectorAssembler().
  setInputCols(Array("sepal length", "sepal width", "petal length", "petal width")).
  setOutputCol("features")
val xgbInput = vectorAssembler.transform(labelTransformed).select("features",
  ↪ "classIndex")
```

Now, we have a `DataFrame` containing only two columns, “features” which contains vector-represented “sepal length”, “sepal width”, “petal length” and “petal width” and “classIndex” which has Double-typed labels. A `DataFrame` like this (containing vector-represented features and numeric labels) can be fed to XGBoost4J-Spark’s training engine directly.

Training

XGBoost supports both regression and classification. While we use Iris dataset in this tutorial to show how we use XGBoost/XGBoost4J-Spark to resolve a multi-classes classification problem, the usage in Regression is very similar to classification.

To train a XGBoost model for classification, we need to claim a `XGBoostClassifier` first:

```
import ml.dmlc.xgboost4j.scala.spark.XGBoostClassifier
val xgbParam = Map("eta" -> 0.1f,
  "max_depth" -> 2,
  "objective" -> "multi:softprob",
  "num_class" -> 3,
  "num_round" -> 100,
  "num_workers" -> 2)
val xgbClassifier = new XGBoostClassifier(xgbParam).
  setFeaturesCol("features").
  setLabelCol("classIndex")
```

The available parameters for training a XGBoost model can be found in [here](#). In XGBoost4J-Spark, we support not only the default set of parameters but also the camel-case variant of these parameters to keep consistent with Spark’s MLLIB parameters.

Specifically, each parameter in [this page](#) has its equivalent form in XGBoost4J-Spark with camel case. For example, to set `max_depth` for each tree, you can pass parameter just like what we did in the above code snippet (as `max_depth` wrapped in a `Map`), or you can do it through setters in `XGBoostClassifier`:

```
val xgbClassifier = new XGBoostClassifier().
  setFeaturesCol("features").
  setLabelCol("classIndex")
xgbClassifier.setMaxDepth(2)
```

After we set `XGBoostClassifier` parameters and feature/label column, we can build a transformer, `XGBoostClassificationModel` by fitting `XGBoostClassifier` with the input `DataFrame`. This `fit` operation is essentially the training process and the generated model can then be used in prediction.

```
val xgbClassificationModel = xgbClassifier.fit(xgbInput)
```

Prediction

XGBoost4j-Spark supports two ways for model serving: batch prediction and single instance prediction.

Batch Prediction

When we get a model, either `XGBoostClassificationModel` or `XGBoostRegressionModel`, it takes a `DataFrame`, read the column containing feature vectors, predict for each feature vector, and output a new `DataFrame` with the following columns by default:

- `XGBoostClassificationModel` will output margins (`rawPredictionCol`), probabilities(`probabilityCol`) and the eventual prediction labels (`predictionCol`) for each possible label.
- `XGBoostRegressionModel` will output prediction label(`predictionCol`).

Batch prediction expects the user to pass the testset in the form of a `DataFrame`. `XGBoost4J-Spark` starts a `XGBoost` worker for each partition of `DataFrame` for parallel prediction and generates prediction results for the whole `DataFrame` in a batch.

```
val xgbClassificationModel = xgbClassifier.fit(xgbInput)
val results = xgbClassificationModel.transform(testSet)
```

With the above code snippet, we get a result `DataFrame`, result containing margin, probability for each class and the prediction for each instance

features	classIndex	rawPrediction	probability	prediction
[5.1, 3.5, 1.4, 0.2]	0.0	[3.45569849014282...	[0.99579632282257...	0.0
[4.9, 3.0, 1.4, 0.2]	0.0	[3.45569849014282...	[0.99618089199066...	0.0
[4.7, 3.2, 1.3, 0.2]	0.0	[3.45569849014282...	[0.99643349647521...	0.0
[4.6, 3.1, 1.5, 0.2]	0.0	[3.45569849014282...	[0.99636095762252...	0.0
[5.0, 3.6, 1.4, 0.2]	0.0	[3.45569849014282...	[0.99579632282257...	0.0
[5.4, 3.9, 1.7, 0.4]	0.0	[3.45569849014282...	[0.99428516626358...	0.0
[4.6, 3.4, 1.4, 0.3]	0.0	[3.45569849014282...	[0.99643349647521...	0.0
[5.0, 3.4, 1.5, 0.2]	0.0	[3.45569849014282...	[0.99579632282257...	0.0
[4.4, 2.9, 1.4, 0.2]	0.0	[3.45569849014282...	[0.99618089199066...	0.0
[4.9, 3.1, 1.5, 0.1]	0.0	[3.45569849014282...	[0.99636095762252...	0.0
[5.4, 3.7, 1.5, 0.2]	0.0	[3.45569849014282...	[0.99428516626358...	0.0
[4.8, 3.4, 1.6, 0.2]	0.0	[3.45569849014282...	[0.99643349647521...	0.0
[4.8, 3.0, 1.4, 0.1]	0.0	[3.45569849014282...	[0.99618089199066...	0.0
[4.3, 3.0, 1.1, 0.1]	0.0	[3.45569849014282...	[0.99618089199066...	0.0
[5.8, 4.0, 1.2, 0.2]	0.0	[3.45569849014282...	[0.97809928655624...	0.0
[5.7, 4.4, 1.5, 0.4]	0.0	[3.45569849014282...	[0.97809928655624...	0.0
[5.4, 3.9, 1.3, 0.4]	0.0	[3.45569849014282...	[0.99428516626358...	0.0
[5.1, 3.5, 1.4, 0.3]	0.0	[3.45569849014282...	[0.99579632282257...	0.0
[5.7, 3.8, 1.7, 0.3]	0.0	[3.45569849014282...	[0.97809928655624...	0.0
[5.1, 3.8, 1.5, 0.3]	0.0	[3.45569849014282...	[0.99579632282257...	0.0

Single instance prediction

`XGBoostClassificationModel` or `XGBoostRegressionModel` support make prediction on single instance as well. It accepts a single `Vector` as feature, and output the prediction label.

However, the overhead of single-instance prediction is high due to the internal overhead of `XGBoost`, use it carefully!

```
val features = xgbInput.head().getAs[Vector]("features")
val result = xgbClassificationModel.predict(features)
```

Model Persistence

Model and pipeline persistence

A data scientist produces an ML model and hands it over to an engineering team for deployment in a production environment. Reversely, a trained model may be used by data scientists, for example as a baseline, across the process of data exploration. So it's important to support model persistence to make the models available across usage scenarios and programming languages.

XGBoost4j-Spark supports saving and loading XGBoostClassifier/XGBoostClassificationModel and XGBoostRegressor/XGBoostRegressionModel. It also supports saving and loading a ML pipeline which includes these estimators and models.

We can save the XGBoostClassificationModel to file system:

```
val xgbClassificationModelPath = "/tmp/xgbClassificationModel"
xgbClassificationModel.write.overwrite().save(xgbClassificationModelPath)
```

and then loading the model in another session:

```
import ml.dmlc.xgboost4j.scala.spark.XGBoostClassificationModel

val xgbClassificationModel2 = XGBoostClassificationModel.
  ↳load(xgbClassificationModelPath)
xgbClassificationModel2.transform(xgbInput)
```

With regards to ML pipeline save and load, please refer the next section.

Interact with Other Bindings of XGBoost

After we train a model with XGBoost4j-Spark on massive dataset, sometimes we want to do model serving in single machine or integrate it with other single node libraries for further processing. XGBoost4j-Spark supports export model to local by:

```
val nativeModelPath = "/tmp/nativeModel"
xgbClassificationModel.nativeBooster.saveModel(nativeModelPath)
```

Then we can load this model with single node Python XGBoost:

```
import xgboost as xgb
bst = xgb.Booster({'nthread': 4})
bst.load_model(nativeModelPath)
```

Note: Using HDFS and S3 for exporting the models with nativeBooster.saveModel()

When interacting with other language bindings, XGBoost also supports saving-models-to and loading-models-from file systems other than the local one. You can use HDFS and S3 by prefixing the path with `hdfs://` and `s3://` respectively. However, for this capability, you must do **one** of the following:

1. Build XGBoost4J-Spark with the steps described in [here](#), but turning `USE_HDFS` (or `USE_S3`, etc. in the same place) switch on. With this approach, you can reuse the above code example by replacing “nativeModelPath” with a HDFS path.
 - However, if you build with `USE_HDFS`, etc. you have to ensure that the involved shared object file, e.g. `libhdfs.so`, is put in the `LIBRARY_PATH` of your cluster. To avoid the complicated cluster environment configuration, choose the other option.
2. Use bindings of HDFS, S3, etc. to pass model files around. Here are the steps (taking HDFS as an example):
 - Create a new file with

```
val outputStream = fs.create("hdfs_path")
```

where “fs” is an instance of `org.apache.hadoop.fs.FileSystem` class in Hadoop.

- Pass the returned `OutputStream` in the first step to `nativeBooster.saveModel()`:

```
xgbClassificationModel.nativeBooster.saveModel(outputStream)
```

- Download file in other languages from HDFS and load with the pre-built (without the requirement of `libhdfs.so`) version of XGBoost. (The function “`download_from_hdfs`” is a helper function to be implemented by the user)

```
import xgboost as xgb
bst = xgb.Booster({'nthread': 4})
local_path = download_from_hdfs("hdfs_path")
bst.load_model(local_path)
```

Note: Consistency issue between XGBoost4J-Spark and other bindings

There is a consistency issue between XGBoost4J-Spark and other language bindings of XGBoost.

When users use Spark to load training/test data in LIBSVM format with the following code snippet:

```
spark.read.format("libsvm").load("trainingset_libsvm")
```

Spark assumes that the dataset is using 1-based indexing (feature indices starting with 1). However, when you do prediction with other bindings of XGBoost (e.g. Python API of XGBoost), XGBoost assumes that the dataset is using 0-based indexing (feature indices starting with 0) by default. It creates a pitfall for the users who train model with Spark but predict with the dataset in the same format in other bindings of XGBoost. The solution is to transform the dataset to 0-based indexing before you predict with, for example, Python API, or you append `?indexing_mode=1` to your file path when loading with `DMatrx`. For example in Python:

```
xgb.DMatrix('test.libsvm?indexing_mode=1')
```

Building a ML Pipeline with XGBoost4J-Spark

Basic ML Pipeline

Spark ML pipeline can combine multiple algorithms or functions into a single pipeline. It covers from feature extraction, transformation, selection to model training and prediction. XGBoost4j-Spark makes it feasible to embed XGBoost into such a pipeline seamlessly. The following example shows how to build such a pipeline consisting of Spark `MLlib` feature transformer and `XGBoostClassifier` estimator.

We still use `Iris` dataset and the `rawInput` `DataFrame`. First we need to split the dataset into training and test dataset.

```
val Array(training, test) = rawInput.randomSplit(Array(0.8, 0.2), 123)
```

Then we build the ML pipeline which includes 4 stages:

- Assemble all features into a single vector column.
- From string label to indexed double label.
- Use `XGBoostClassifier` to train classification model.
- Convert indexed double label back to original string label.

We have shown the first three steps in the earlier sections, and the last step is finished with a new transformer `IndexToString`:

```
val labelConverter = new IndexToString()
  .setInputCol("prediction")
  .setOutputCol("realLabel")
  .setLabels(stringIndexer.labels)
```

We need to organize these steps as a Pipeline in Spark ML framework and evaluate the whole pipeline to get a `PipelineModel`:

```
import org.apache.spark.ml.feature._
import org.apache.spark.ml.Pipeline

val pipeline = new Pipeline()
  .setStages(Array(assembler, stringIndexer, booster, labelConverter))
val model = pipeline.fit(training)
```

After we get the `PipelineModel`, we can make prediction on the test dataset and evaluate the model accuracy.

```
import org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator

val prediction = model.transform(test)
val evaluator = new MulticlassClassificationEvaluator()
val accuracy = evaluator.evaluate(prediction)
```

Pipeline with Hyper-parameter Tunning

The most critical operation to maximize the power of XGBoost is to select the optimal parameters for the model. Tuning parameters manually is a tedious and labor-consuming process. With the latest version of XGBoost4J-Spark, we can utilize the Spark model selecting tool to automate this process.

The following example shows the code snippet utilizing `CrossValidation` and `MulticlassClassificationEvaluator` to search the optimal combination of two XGBoost parameters, `max_depth` and `eta`. (See *XGBoost Parameters*.) The model producing the maximum accuracy defined by `MulticlassClassificationEvaluator` is selected and used to generate the prediction for the test set.

```
import org.apache.spark.ml.tuning._
import org.apache.spark.ml.PipelineModel
import ml.dmlc.xgboost4j.scala.spark.XGBoostClassificationModel

val paramGrid = new ParamGridBuilder()
  .addGrid(booster.maxDepth, Array(3, 8))
  .addGrid(booster.eta, Array(0.2, 0.6))
```

(continues on next page)

(continued from previous page)

```
.build()
val cv = new CrossValidator()
    .setEstimator(pipeline)
    .setEvaluator(evaluator)
    .setEstimatorParamMaps(paramGrid)
    .setNumFolds(3)

val cvModel = cv.fit(training)

val bestModel = cvModel.bestModel.asInstanceOf[PipelineModel].stages(2)
    .asInstanceOf[XGBoostClassificationModel]
bestModel.extractParamMap()
```

Run XGBoost4J-Spark in Production

XGBoost4J-Spark is one of the most important steps to bring XGBoost to production environment easier. In this section, we introduce three key features to run XGBoost4J-Spark in production.

Parallel/Distributed Training

The massive size of training dataset is one of the most significant characteristics in production environment. To ensure that training in XGBoost scales with the data size, XGBoost4J-Spark bridges the distributed/parallel processing framework of Spark and the parallel/distributed training mechanism of XGBoost.

In XGBoost4J-Spark, each XGBoost worker is wrapped by a Spark task and the training dataset in Spark's memory space is fed to XGBoost workers in a transparent approach to the user.

In the code snippet where we build XGBoostClassifier, we set parameter `num_workers` (or `numWorkers`). This parameter controls how many parallel workers we want to have when training a XGBoostClassificationModel.

Note: Regarding OpenMP optimization

By default, we allocate a core per each XGBoost worker. Therefore, the OpenMP optimization within each XGBoost worker does not take effect and the parallelization of training is achieved by running multiple workers (i.e. Spark tasks) at the same time.

If you do want OpenMP optimization, you have to

1. set `nthread` to a value larger than 1 when creating XGBoostClassifier/XGBoostRegressor
 2. set `spark.task.cpus` in Spark to the same value as `nthread`
-

Gang Scheduling

XGBoost uses [AllReduce](#) algorithm to synchronize the stats, e.g. histogram values, of each worker during training. Therefore XGBoost4J-Spark requires that all of `nthread * numWorkers` cores should be available before the training runs.

In the production environment where many users share the same cluster, it's hard to guarantee that your XGBoost4J-Spark application can get all requested resources for every run. By default, the communication layer in XGBoost will block the whole application when it requires more resources to be available. This process usually brings unnecessary

resource waste as it keeps the ready resources and try to claim more. Additionally, this usually happens silently and does not bring the attention of users.

XGBoost4J-Spark allows the user to setup a timeout threshold for claiming resources from the cluster. If the application cannot get enough resources within this time period, the application would fail instead of wasting resources for hanging long. To enable this feature, you can set with XGBoostClassifier/XGBoostRegressor:

```
xgbClassifier.setTimeoutRequestWorkers(60000L)
```

or pass in `timeout_request_workers` in `xgbParamMap` when building XGBoostClassifier:

```
val xgbParam = Map("eta" -> 0.1f,
  "max_depth" -> 2,
  "objective" -> "multi:softprob",
  "num_class" -> 3,
  "num_round" -> 100,
  "num_workers" -> 2,
  "timeout_request_workers" -> 60000L)
val xgbClassifier = new XGBoostClassifier(xgbParam) .
  setFeaturesCol("features") .
  setLabelCol("classIndex")
```

If XGBoost4J-Spark cannot get enough resources for running two XGBoost workers, the application would fail. Users can have external mechanism to monitor the status of application and get notified for such case.

Checkpoint During Training

Transient failures are also commonly seen in production environment. To simplify the design of XGBoost, we stop training if any of the distributed workers fail. However, if the training fails after having been through a long time, it would be a great waste of resources.

We support creating checkpoint during training to facilitate more efficient recovery from failure. To enable this feature, you can set how many iterations we build each checkpoint with `setCheckpointInterval` and the location of checkpoints with `setCheckpointPath`:

```
xgbClassifier.setCheckpointInterval(2)
xgbClassifier.setCheckpointPath("/checkpoint_path")
```

An equivalent way is to pass in parameters in XGBoostClassifier's constructor:

```
val xgbParam = Map("eta" -> 0.1f,
  "max_depth" -> 2,
  "objective" -> "multi:softprob",
  "num_class" -> 3,
  "num_round" -> 100,
  "num_workers" -> 2,
  "checkpoint_path" -> "/checkpoints_path",
  "checkpoint_interval" -> 2)
val xgbClassifier = new XGBoostClassifier(xgbParam) .
  setFeaturesCol("features") .
  setLabelCol("classIndex")
```

If the training failed during these 100 rounds, the next run of training would start by reading the latest checkpoint file in `/checkpoints_path` and start from the iteration when the checkpoint was built until to next failure or the specified 100 rounds.

XGBoost4J Java API

XGBoost4J Scala API

XGBoost4J-Spark Scala API

XGBoost4J-Flink Scala API

1.10 XGBoost.jl

See [XGBoost.jl Project page](#).

1.11 XGBoost Command Line version

See [XGBoost Command Line walkthrough](#).

1.12 Contribute to XGBoost

XGBoost has been developed and used by a group of active community members. Everyone is more than welcome to contribute. It is a way to make the project better and more accessible to more users.

- Please add your name to [CONTRIBUTORS.md](#) after your patch has been merged.
- Please also update [NEWS.md](#) to add note on your changes to the API or XGBoost documentation.

Guidelines

- *Submit Pull Request*
- *Git Workflow Howtos*
 - *How to resolve conflict with master*
 - *How to combine multiple commits into one*
 - *What is the consequence of force push*
- *Documents*
- *Testcases*
- *Sanitizers*
- *Examples*
- *Core Library*
- *Python Package*
- *R Package*

1.12.1 Submit Pull Request

- Before submit, please rebase your code on the most recent version of master, you can do it by

```
git remote add upstream https://github.com/dmlc/xgboost
git fetch upstream
git rebase upstream/master
```

- If you have multiple small commits, it might be good to merge them together (use git rebase then squash) into more meaningful groups.
- Send the pull request!
 - Fix the problems reported by automatic checks
 - If you are contributing a new module, consider add a testcase in `tests`.

1.12.2 Git Workflow Howtos

How to resolve conflict with master

- First rebase to most recent master

```
# The first two steps can be skipped after you do it once.
git remote add upstream https://github.com/dmlc/xgboost
git fetch upstream
git rebase upstream/master
```

- The git may show some conflicts it cannot merge, say `conflicted.py`.
 - Manually modify the file to resolve the conflict.
 - After you resolved the conflict, mark it as resolved by

```
git add conflicted.py
```

- Then you can continue rebase by

```
git rebase --continue
```

- Finally push to your fork, you may need to force push here.

```
git push --force
```

How to combine multiple commits into one

Sometimes we want to combine multiple commits, especially when later commits are only fixes to previous ones, to create a PR with set of meaningful commits. You can do it by following steps.

- Before doing so, configure the default editor of git if you haven't done so before.

```
git config core.editor the-editor-you-like
```

- Assume we want to merge last 3 commits, type the following commands

```
git rebase -i HEAD~3
```

- It will pop up an text editor. Set the first commit as `pick`, and change later ones to `squash`.
- After you saved the file, it will pop up another text editor to ask you modify the combined commit message.

- Push the changes to your fork, you need to force push.

```
git push --force
```

What is the consequence of force push

The previous two tips requires force push, this is because we altered the path of the commits. It is fine to force push to your own fork, as long as the commits changed are only yours.

1.12.3 Documents

- Documentation is built using sphinx.
- Each document is written in `reStructuredText`.
- You can build document locally to see the effect.

1.12.4 Testcases

- All the testcases are in `tests`.
- We use python nose for python test cases.

1.12.5 Sanitizers

By default, sanitizers are bundled in GCC and Clang/LLVM. One can enable sanitizers with GCC ≥ 4.8 or LLVM ≥ 3.1 , But some distributions might package sanitizers separately. Here is a list of supported sanitizers with corresponding library names:

- Address sanitizer: libasan
- Leak sanitizer: liblsan
- Thread sanitizer: libtsan

Memory sanitizer is exclusive to LLVM, hence not supported in XGBoost.

How to build XGBoost with sanitizers

One can build XGBoost with sanitizer support by specifying `-DUSE_SANITIZER=ON`. By default, address sanitizer and leak sanitizer are used when you turn the `USE_SANITIZER` flag on. You can always change the default by providing a semicolon separated list of sanitizers to `ENABLED_SANITIZERS`. Note that thread sanitizer is not compatible with the other two sanitizers.

```
cmake -DUSE_SANITIZER=ON -DENABLED_SANITIZERS="address;leak" /path/to/xgboost
```

How to use sanitizers with CUDA support

Runing XGBoost on CUDA with address sanitizer (asan) will raise memory error. To use asan with CUDA correctly, you need to configure asan via `ASAN_OPTIONS` environment variable:

```
ASAN_OPTIONS=protect_shadow_gap=0 ../testxgboost
```

For details, please consult [official documentation](#) for sanitizers.

1.12.6 Examples

- Usecases and examples will be in [demo](#).
- We are super excited to hear about your story, if you have blogposts, tutorials code solutions using XGBoost, please tell us and we will add a link in the example pages.

1.12.7 Core Library

- Follow [Google style for C++](#).
- Use C++11 features such as smart pointers, braced initializers, lambda functions, and `std::thread`.
- We use Doxygen to document all the interface code.
- You can reproduce the linter checks by running `make lint`

1.12.8 Python Package

- Always add docstring to the new functions in `numpydoc` format.
- You can reproduce the linter checks by typing `make lint`

1.12.9 R Package

Code Style

- We follow Google's C++ Style guide for C++ code.
 - This is mainly to be consistent with the rest of the project.
 - Another reason is we will be able to check style automatically with a linter.
- You can check the style of the code by typing the following command at root folder.

```
make rcplint
```

- When needed, you can disable the linter warning of certain line with ``// NOLINT (*)`` comments.
- We use [roxygen](#) for documenting the R package.

Rmarkdown Vignettes

Rmarkdown vignettes are placed in [R-package/vignettes](#). These Rmarkdown files are not compiled. We host the compiled version on [doc/R-package](#).

The following steps are followed to add a new Rmarkdown vignettes:

- Add the original rmarkdown to `R-package/vignettes`.
- Modify `doc/R-package/Makefile` to add the markdown files to be build.

- Clone the [dmlc/web-data](#) repo to folder `doc`.
- Now type the following command on `doc/R-package`:

```
make the-markdown-to-make.md
```

- This will generate the markdown, as well as the figures in `doc/web-data/xgboost/knitr`.
- Modify the `doc/R-package/index.md` to point to the generated markdown.
- Add the generated figure to the `dmlc/web-data` repo.
 - If you already cloned the repo to `doc`, this means `git add`
- Create PR for both the markdown and `dmlc/web-data`.
- You can also build the document locally by typing the following command at the `doc` directory:

```
make html
```

The reason we do this is to avoid exploded repo size due to generated images.

R package versioning

Since version 0.6.4.3, we have adopted a versioning system that uses `x.y.z` (or `core_major.core_minor.cran_release`) format for CRAN releases and an `x.y.z.p` (or `core_major.core_minor.cran_release.patch`) format for development patch versions. This approach is similar to the one described in Yihui Xie's [blog post on R Package Versioning](#), except we need an additional field to accomodate the `x.y` core library version.

Each new CRAN release bumps up the 3rd field, while developments in-between CRAN releases would be marked by an additional 4th field on the top of an existing CRAN release version. Some additional consideration is needed when the core library version changes. E.g., after the core changes from 0.6 to 0.7, the R package development version would become 0.7.0.1, working towards a 0.7.1 CRAN release. The 0.7.0 would not be released to CRAN, unless it would require almost no additional development.

Registering native routines in R

According to [R extension manual](#), it is good practice to register native routines and to disable symbol search. When any changes or additions are made to the C++ interface of the R package, please make corresponding changes in `src/init.c` as well.

X

- `xgboost.core`, [45](#)
- `xgboost.plotting`, [62](#)
- `xgboost.sklearn`, [54](#)
- `xgboost.training`, [52](#)

A

`apply()` (xgboost.XGBClassifier method), 59
`apply()` (xgboost.XGBRegressor method), 55
`attr()` (xgboost.Booster method), 48
`attributes()` (xgboost.Booster method), 48

B

`boost()` (xgboost.Booster method), 48
Booster (class in xgboost), 48

C

`copy()` (xgboost.Booster method), 49
`cv()` (in module xgboost), 53

D

DMatrix (class in xgboost), 45
`dump_model()` (xgboost.Booster method), 49

E

`eval()` (xgboost.Booster method), 49
`eval_set()` (xgboost.Booster method), 49
`evals_result()` (xgboost.XGBClassifier method), 59
`evals_result()` (xgboost.XGBRegressor method), 56

F

`feature_importances_` (xgboost.XGBClassifier attribute), 60
`feature_importances_` (xgboost.XGBRegressor attribute), 56
`feature_names` (xgboost.DMatrix attribute), 46
`feature_types` (xgboost.DMatrix attribute), 46
`fit()` (xgboost.XGBClassifier method), 60
`fit()` (xgboost.XGBRegressor method), 56

G

`get_base_margin()` (xgboost.DMatrix method), 46
`get_booster()` (xgboost.XGBClassifier method), 61
`get_booster()` (xgboost.XGBRegressor method), 57
`get_dump()` (xgboost.Booster method), 49

`get_float_info()` (xgboost.DMatrix method), 46
`get_fscore()` (xgboost.Booster method), 49
`get_label()` (xgboost.DMatrix method), 46
`get_params()` (xgboost.XGBClassifier method), 61
`get_params()` (xgboost.XGBRegressor method), 57
`get_score()` (xgboost.Booster method), 49
`get_split_value_histogram()` (xgboost.Booster method), 50
`get_uint_info()` (xgboost.DMatrix method), 46
`get_weight()` (xgboost.DMatrix method), 46
`get_xgb_params()` (xgboost.XGBClassifier method), 61
`get_xgb_params()` (xgboost.XGBRegressor method), 57

L

`load_model()` (xgboost.Booster method), 50
`load_model()` (xgboost.XGBClassifier method), 61
`load_model()` (xgboost.XGBRegressor method), 57
`load_rabit_checkpoint()` (xgboost.Booster method), 50

N

`num_col()` (xgboost.DMatrix method), 47
`num_row()` (xgboost.DMatrix method), 47

P

`plot_importance()` (in module xgboost), 62
`plot_tree()` (in module xgboost), 63
`predict()` (xgboost.Booster method), 50
`predict()` (xgboost.XGBClassifier method), 61
`predict()` (xgboost.XGBRegressor method), 57
`predict_proba()` (xgboost.XGBClassifier method), 62

S

`save_binary()` (xgboost.DMatrix method), 47
`save_model()` (xgboost.Booster method), 51
`save_model()` (xgboost.XGBClassifier method), 62
`save_model()` (xgboost.XGBRegressor method), 58
`save_rabit_checkpoint()` (xgboost.Booster method), 51
`save_raw()` (xgboost.Booster method), 51
`set_attr()` (xgboost.Booster method), 52

`set_base_margin()` (xgboost.DMatrix method), [47](#)
`set_float_info()` (xgboost.DMatrix method), [47](#)
`set_float_info_np2d()` (xgboost.DMatrix method), [47](#)
`set_group()` (xgboost.DMatrix method), [47](#)
`set_label()` (xgboost.DMatrix method), [47](#)
`set_label_np2d()` (xgboost.DMatrix method), [48](#)
`set_param()` (xgboost.Booster method), [52](#)
`set_uint_info()` (xgboost.DMatrix method), [48](#)
`set_weight()` (xgboost.DMatrix method), [48](#)
`set_weight_np2d()` (xgboost.DMatrix method), [48](#)
`slice()` (xgboost.DMatrix method), [48](#)

T

`to_graphviz()` (in module xgboost), [63](#)
`train()` (in module xgboost), [52](#)

U

`update()` (xgboost.Booster method), [52](#)

X

`XGBClassifier` (class in xgboost), [58](#)
`xgboost.core` (module), [45](#)
`xgboost.plotting` (module), [62](#)
`xgboost.sklearn` (module), [54](#)
`xgboost.training` (module), [52](#)
`XGBRegressor` (class in xgboost), [54](#)