教育部教改計畫
開發課程模組

eXtreme Gradient Boosting Introduction

Professor Chien-Mo James Li 李建模
Graduate Institute of Electronics Engineering
National Taiwan University

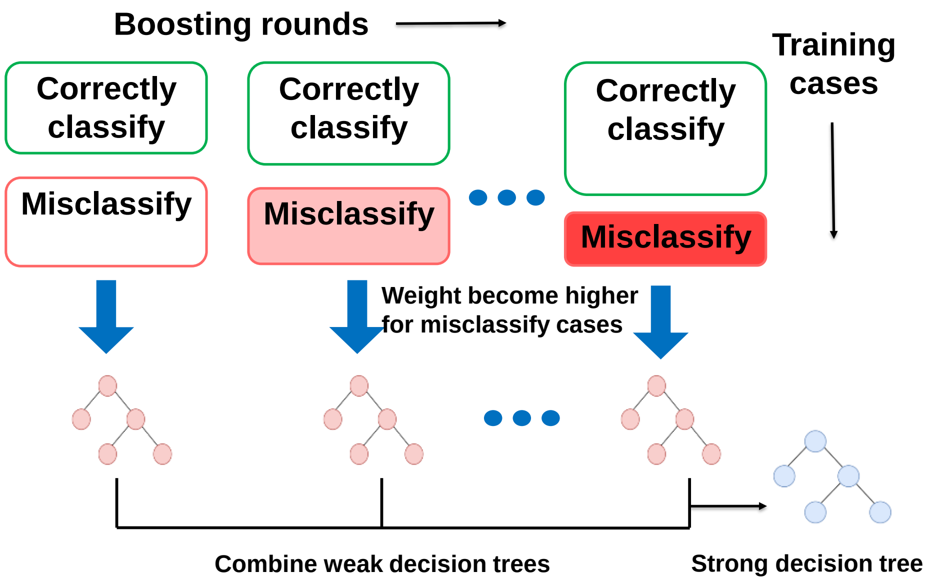
Outline

- **Introduction**
- **Installation**
- **Get Started**
- **Basic Functions**

Boosting

- Often use when training models
  - With decision-tree algorithms
- Train decision tree in a sequence
  - Each decision tree is often weak
- Focus on the misclassified cases
  - Give the incorrect classifications from the first tree a higher weight then input to the next tree
- Combine the weak trees
  - Become a single powerful tree

Boosting (cont.)



Boosting rounds

Training cases

Correctly classify

Misclassify

Correctly classify

Misclassify

Correctly classify

Misclassify

Weight become higher for misclassify cases

Combine weak decision trees

Strong decision tree

Types of Boosting

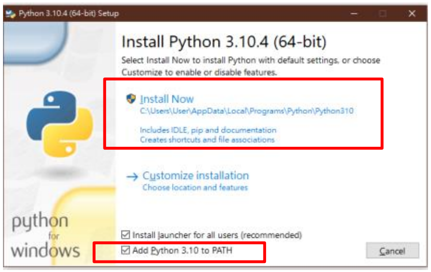
- Adaptive boosting (AdaBoost)
  - Give same weight to each dataset
  - Misclassified cases get higher weight in the next round
  - Stop when the residual error is smaller than threshold
- Gradient boosting (GB)
  - Does not give misclassified cases higher weight
  - Fit a weak learner to the opposite of the gradient of the current fitting error in each iteration
- Extreme gradient boosting (XGBoost)
  - Introduce in next page

Extreme Gradient Boosting (XGBoost)

- XGBoost: A Scalable Tree Boosting System
  - Tianqi Chen, Carlos Guestrin, (2016)
- Using gradient descent
  - Concept similar to GB but different from AdaBoost
- Implements parallel processing
  - 10 times faster than gradient boosting
- Implements regularization to reduce overfitting
- Allows users to define custom optimization objectives and evaluation criteria

Installation

- For Windows and MAC
- Download Python
  - Download link (https://www.python.org/downloads/)
  - Choose the package for your OS
- Install Python
  - Open the exe file
  - Check add Python to PATH
  - Click Install Now



Python 3.10.4 (64-bit) Setup

Install Python 3.10.4 (64-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

Install Now
C:\Users\User\AppData\Local\Programs\Python\Python310

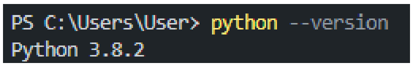
Includes IDLE, pip and documentation
Creates shortcuts and file associations

→ Customize installation
Choose location and features

☑ Install launcher for all users (recommended)
☑ Add Python 3.10 to PATH

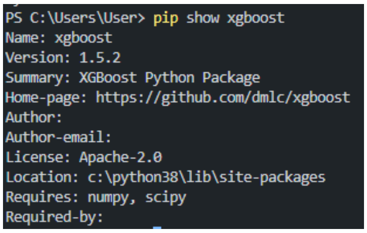
Installation (cont.)

- Check Python version
  - `$python --version`
  - `$python3 --version` for MAC



```
PS C:\Users\User> python --version
Python 3.8.2
```

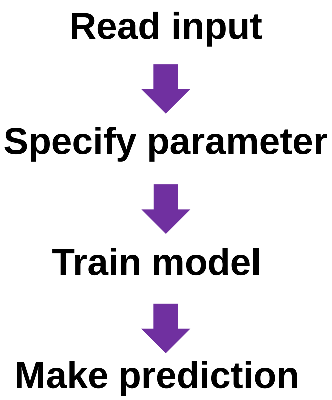
- Install XGBoost
  - `$pip install xgboost`
  - `$pip3 install xgboost` for MAC
- Check package
  - `$pip show xgboost`



```
PS C:\Users\User> pip show xgboost
Name: xgboost
Version: 1.5.2
Summary: XGBoost Python Package
Home-page: https://github.com/dmlc/xgboost
Author:
Author-email:
License: Apache-2.0
Location: c:\python38\lib\site-packages
Requires: numpy, scipy
Required-by:
```

Get Started

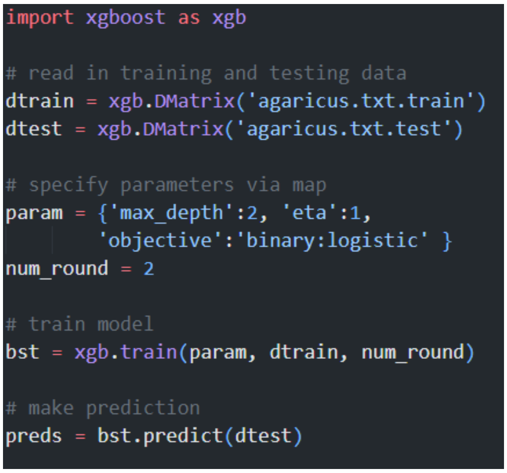
- Train your model
  - Using sample data


Read input

Specify parameter

Train model

Make prediction



```
import xgboost as xgb

# read in training and testing data
dtrain = xgb.DMatrix('agaricus.txt.train')
dtest = xgb.DMatrix('agaricus.txt.test')

# specify parameters via map
param = {'max_depth':2, 'eta':1,
         'objective':'binary:logistic' }
num_round = 2

# train model
bst = xgb.train(param, dtrain, num_round)

# make prediction
preds = bst.predict(dtest)
```

Convert Data to Dmatrix

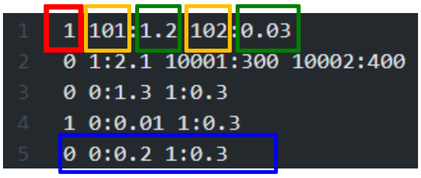
- *dtrain = xgb.DMatrix('agaricus.txt.train')*
  - Input: file path and name
  - Output: data can be used for the model

- *dtrain=xgb.DMatrix(data,label=label,missing=-999.0)*
  - Select specific labels
  - Handle missing data

Input Format

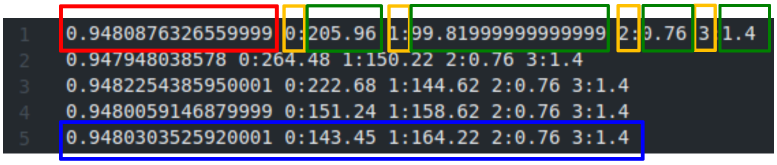
- **Input format example**

Instance label Instance feature: feature value



```
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



```
0.9480876326559999 0:205.96 1:99.81999999999999 2:0.76 3:1.4
0.947948038578 0:264.48 1:150.22 2:0.76 3:1.4
0.9482254385950001 0:222.68 1:144.62 2:0.76 3:1.4
0.9480059146879999 0:151.24 1:158.62 2:0.76 3:1.4
0.9480303525920001 0:143.45 1:164.22 2:0.76 3:1.4
```

Important Training Parameters

- ***booster***: default = gbtree
  - **which booster to use**
- ***nthread***: maximum available threads
  - **number of threads to run XGBoost**
- ***eta***: default = 0.3, range[0, 1]
  - **learning rate**
- ***max_depth***: default = 6, range[0,∞]
  - **Maximum depth of a tree**

Important Training Parameters (cont.)

- ***gamma***: default = 0, range[0,∞]
  - Minimum loss to make a partition on leaf node of tree

- ***subsample***: default = 1, range(0, 1]
  - Subsample ratio of training data

- ***lambda***: default = 1, range[1,∞]
  - L2 regularization term on weights
- ***tree_method***: exact, approx, hist, gpu_hist
  - Tree building method

Important Training Parameters (cont.)

- ***objective***: default = reg:squarederror
  - reg:squarederror: regression with squared loss
  - binary:logistic: logistic regression for binary classification, output probability

- ***eval_metric***: default according to objective
  - rmse: root mean square error
  - mae: mean square error
  - mape: mean absolute percentage error

Train and Predict

- *model = xgb.train(param, dtrain, round)*
  - param: Booster parameter
  - dtrain: Training data
  - round: Number of boosting iterations
  - model: a trained model

- *preds = model.predict(dtest)*
  - model: your trained model
  - dtest: testing data
  - preds: prediction of testing data

Save and Load Model

- *model.save_model(‘name.model')*
  - model: your trained model
  - name: file name of your model

- *model = xgb.Booster()*
  - Function to init model

- *model.load_model(“name.model")*
  - model: the model variable you declare
  - name: file name of your model

Reference Link

- ***https://xgboost.readthedocs.io/en/stable/***
- ***https://zhuanlan.zhihu.com/p/31182879***
- ***https://ithelp.ithome.com.tw/articles/10273094***