Gradient Boosted Regression Trees

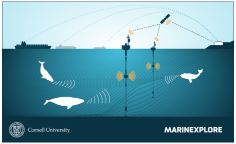
Material: https://github.com/pprett/pydata-gbrt-tutorial

Peter Prettenhofer (@pprett)
DataRobot

Gilles Louppe (@glouppe)
Université de Liège, Belgium

Motivation

GEFCom2012
Wind Forecasting



CARVANA
SELL NO EVIL.

Motivation


NETFLIX
Movies For You
You really liked it...
Heritage Health Prize
Yandex
Web Images Video Mail Translate
Yandex worldwide
Search API
LEARNING TO RANK
March - May 2010
Cornell University
MARINEXPLORE
Expedia
CARVANA
SELL NO EVIL.

Outline

About us

Peter

- @pprett
- Python & ML $\sim$ 6 years
- sklearn dev since 2010

Gilles

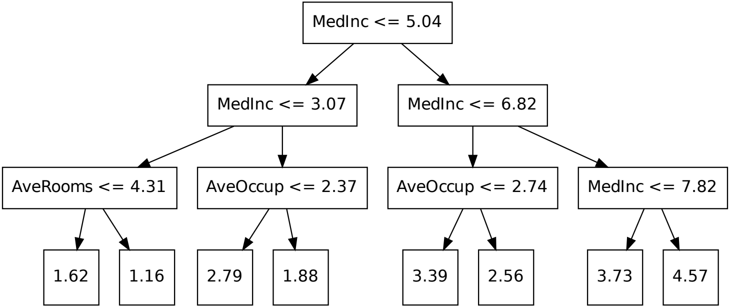
- @glouppe
- PhD student (Liège, Belgium)
- sklearn dev since 2011
  Chief tree hugger

Outline

Machine Learning 101

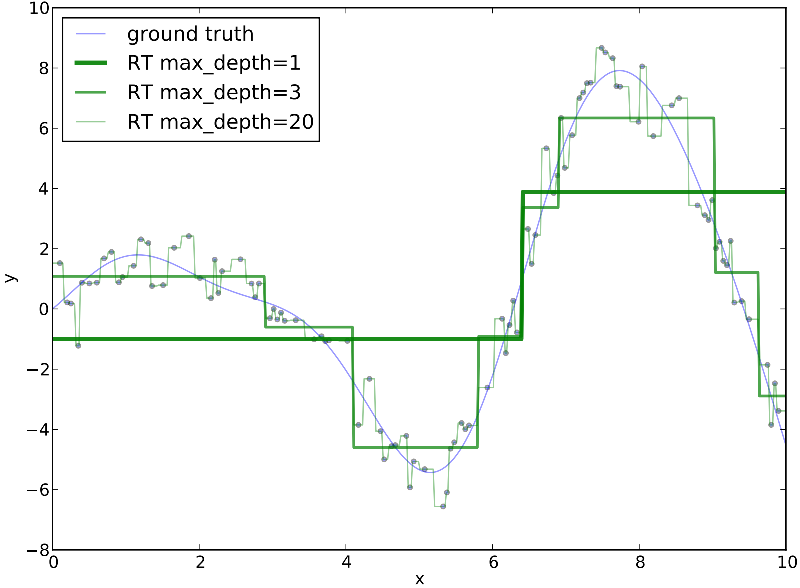
- Data comes as...
  - A set of examples $\{(\mathbf{x}_i, y_i) | 0 \leq i < \text{n\_samples}\}$, with
  - Feature vector $\mathbf{x} \in \mathbb{R}^{\text{n\_features}}$, and
  - Response $y \in \mathbb{R}$ (regression) or $y \in \{-1, 1\}$ (classification)
- Goal is to...
  - Find a function $\hat{y} = f(\mathbf{x})$
  - Such that error $L(y, \hat{y})$ on new (unseen) $\mathbf{x}$ is minimal

Classification and Regression Trees [Breiman et al, 1984]

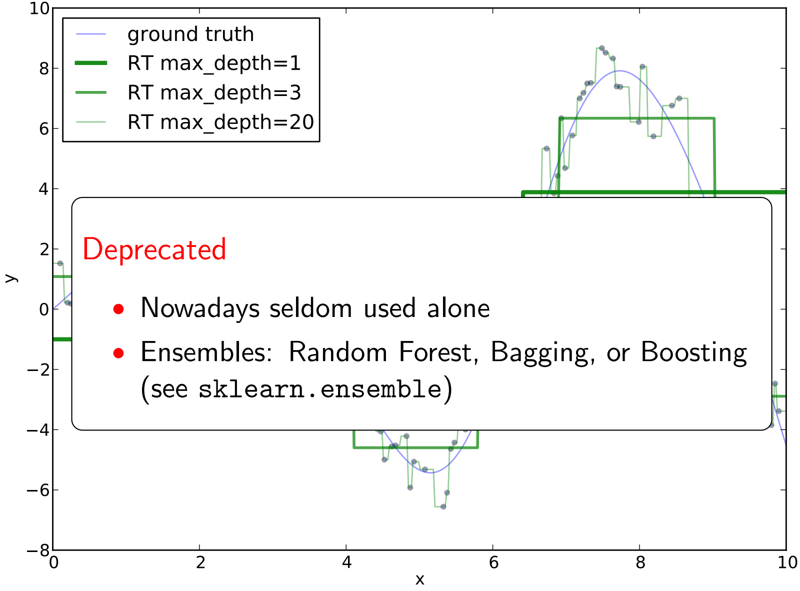


`sklearn.tree.DecisionTreeClassifier|Regressor`

Function approximation with Regression Trees



Function approximation with Regression Trees



Deprecated

- Nowadays seldom used alone
- Ensembles: Random Forest, Bagging, or Boosting (see `sklearn.ensemble`)

Outline

Gradient Boosted Regression Trees

Advantages

- Heterogeneous data (features measured on different scale)
- Supports different loss functions (e.g. huber)
- Automatically detects (non-linear) feature interactions

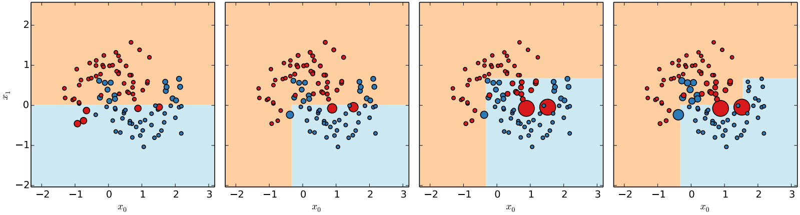
Disadvantages

- Requires careful tuning
- Slow to train (but fast to predict)
- Cannot extrapolate

Boosting

AdaBoost [Y. Freund & R. Schapire, 1995]

- Ensemble: each member is an expert on the errors of its predecessor
- Iteratively re-weights training examples based on errors



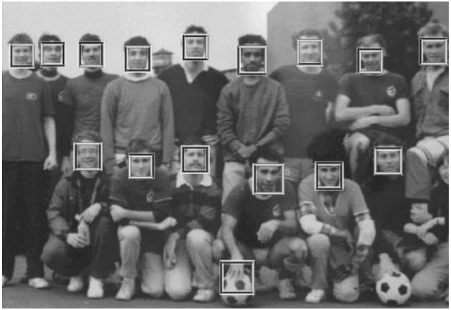
`sklearn.ensemble.AdaBoostClassifier|Regressor`

Boosting

Adal

Huge success

- Viola-Jones Face Detector (2001)



- Freund & Schapire won the Gödel prize 2003

`sklearn.ensemble.AdaBoostClassifier|Regressor`

Gradient Boosting [J. Friedman, 1999]

Statistical view on boosting

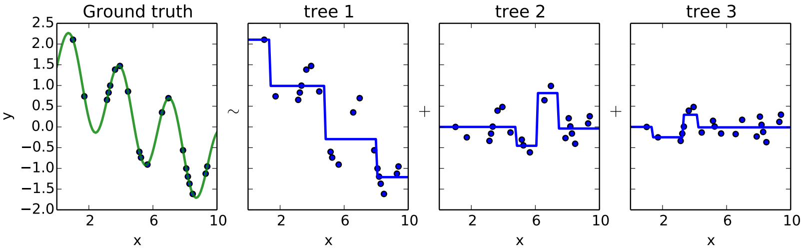
- $\Rightarrow$ Generalization of boosting to arbitrary loss functions

Gradient Boosting [J. Friedman, 1999]

Statistical view on boosting

- $\Rightarrow$ Generalization of boosting to arbitrary loss functions

Residual fitting



`sklearn.ensemble.GradientBoostingClassifier|Regressor`

Functional Gradient Descent

Least Squares Regression

- Squared loss: $L(y_i,\ f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$
- The residual $\sim$ the (negative) gradient $\frac{\partial L(y_i,\, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}$

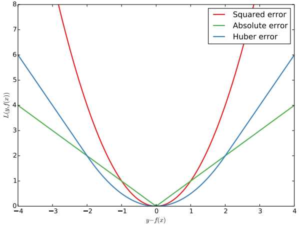
Functional Gradient Descent

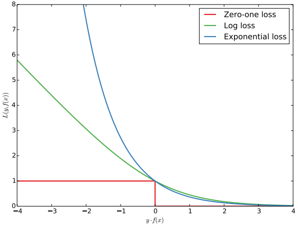
Least Squares Regression

- Squared loss: $L(y_i,\ f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$
- The residual $\sim$ the (negative) gradient $\frac{\partial L(y_i,\, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}$

Steepest Descent

- Regression trees approximate the (negative) gradient
- Each tree is a successive gradient descent step





Outline

Notebook

https://github.com/pprett/pydata-gbrt-tutorial

Outline