Gradient boosting

Vladislav Goncharenko
ML Teamlead, DZEN

girafe ai

MSU, spring 2024

Outline

1. Intuitions
2. Gradient boosting theory
3. Examples
4. Libraries
5. Feature importances
6. Hyperparameter optimization

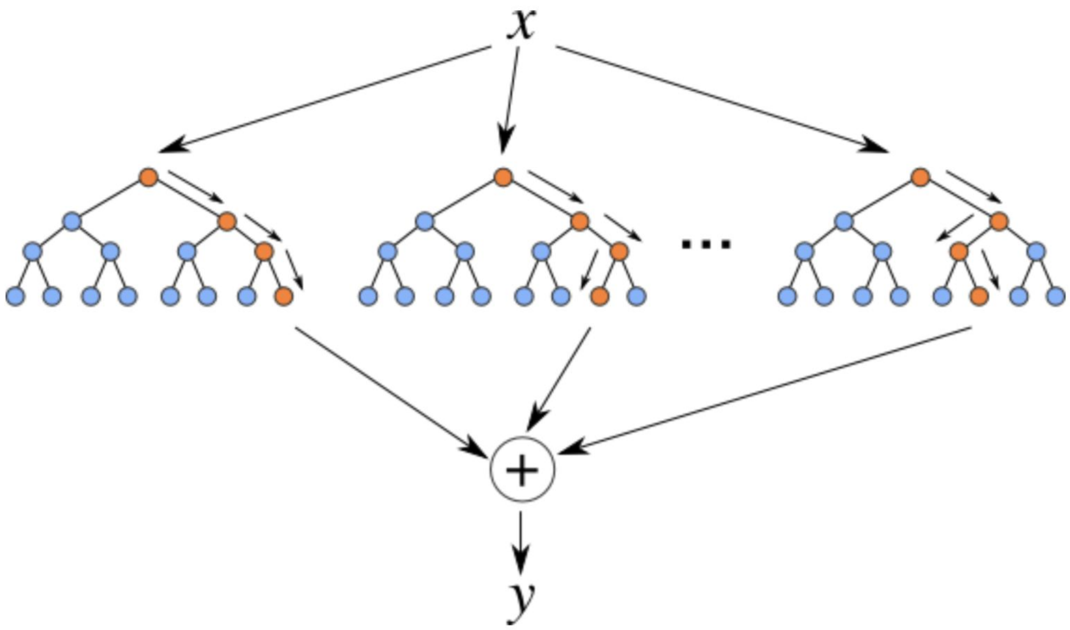
Ensembling recap

girafe
ai

00

Random Forest

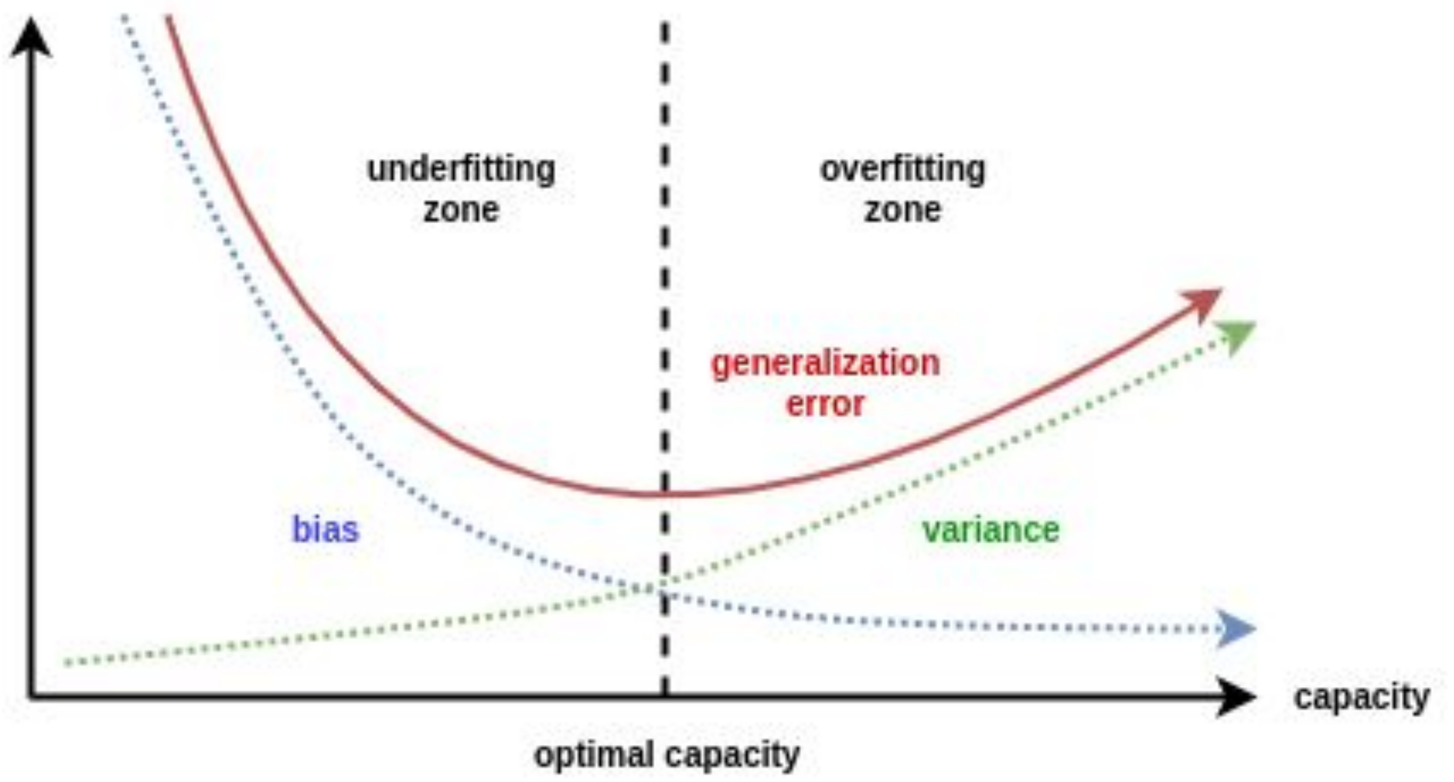
Bagging + RSM = Random Forest



Random Forest

- One of the greatest “universal” models
- There are some modifications: Extremely Randomized Trees, Isolation Forest, etc.

Bias-variance tradeoff



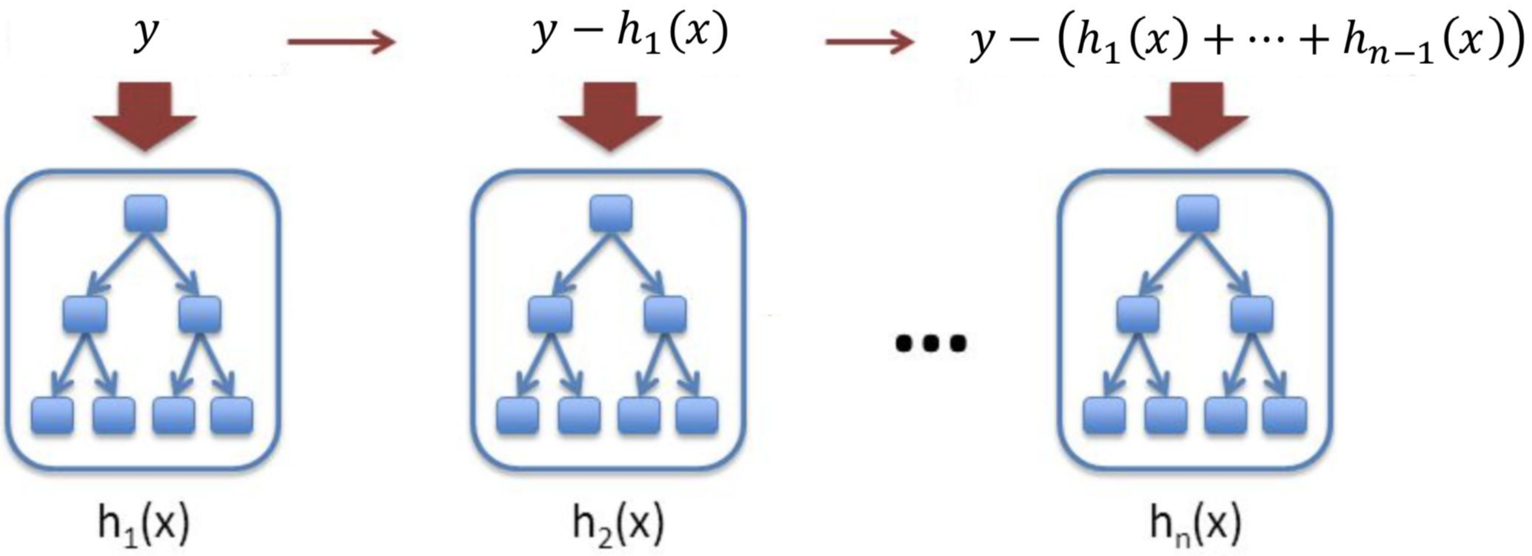
Boosting intuition

girafe
ai

01

Boosting

$$a_n(x) = h_1(x) + \dots + h_n(x)$$

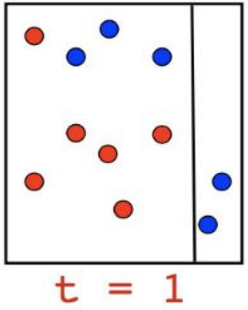


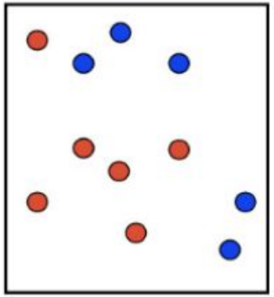
* in case of MSE loss

Boosting: intuition

Binary classification

Use decision stumps

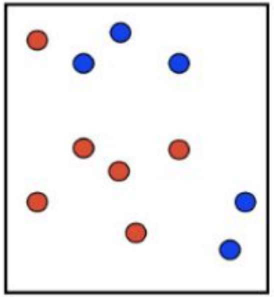


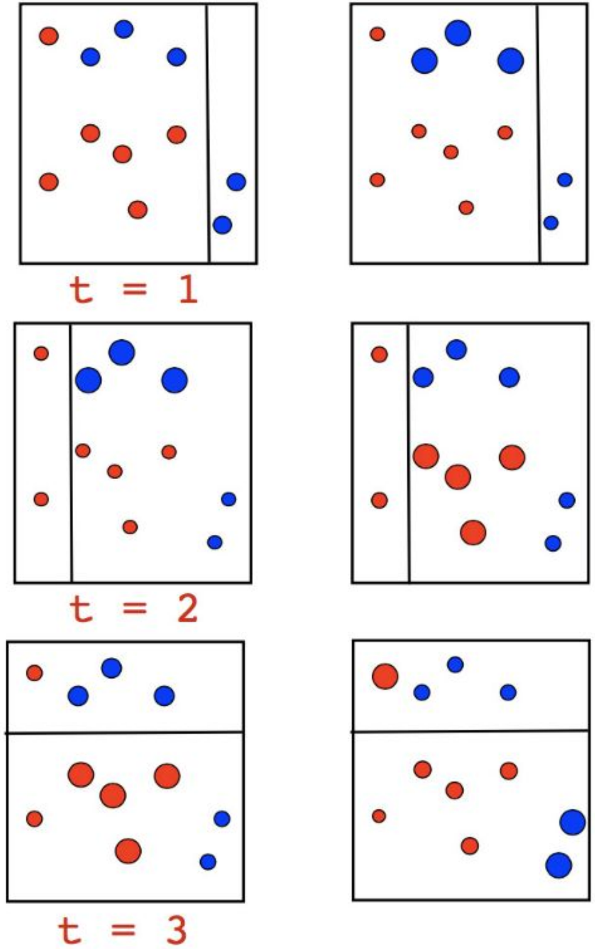


Boosting: intuition

Binary classification

Use decision stumps.

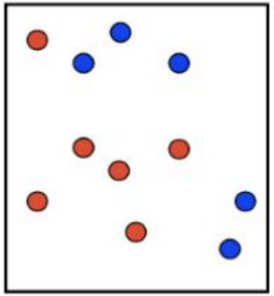


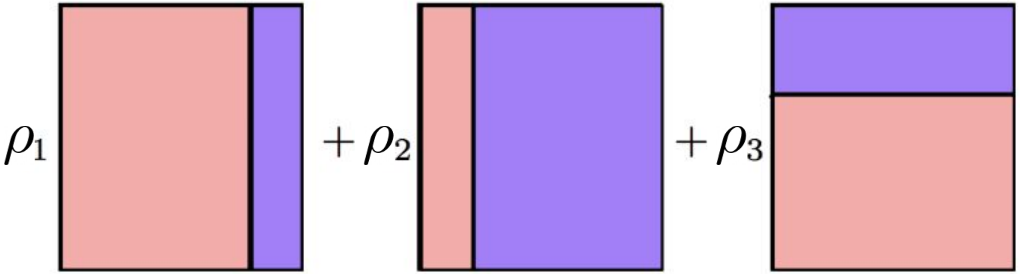


Boosting: intuition

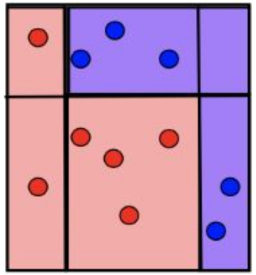
Binary classification

Use decision stumps.





$$\hat{f}_T(x) = \sum_{t=1}^{T} \rho_t h_t(x) \quad =$$



Gradient boosting theory

girafe
ai

02

Gradient boosting: theory

Denote dataset $\{(x_i, y_i)\}_{i=1,\dots,n}$, loss function $L(y, f)$.

Optimal model:

$$\hat{f}(x) = \underset{f(x)}{\arg\min}\ L(y, f(x)) = \underset{f(x)}{\arg\min}\ \mathbb{E}_{x,y}[L(y, f(x))]$$

Let it be from parametric family:

$$\hat{f}(x) = f(x, \hat{\theta}),$$

$$\hat{\theta} = \underset{\theta}{\arg\min}\ \mathbb{E}_{x,y}[L(y, f(x, \theta))]$$

Gradient boosting: theory

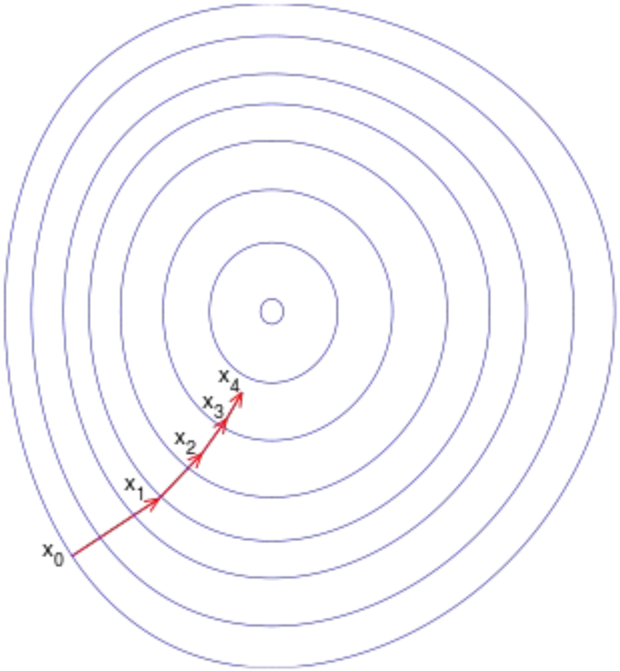
$$\hat{f}(x) = \sum_{i=0}^{t-1} \hat{f}_i(x),$$

$$(\rho_t, \theta_t) = \arg\min_{\rho,\theta} \mathbb{E}_{x,y}[L(y, \hat{f}(x) + \rho \cdot h(x, \theta))],$$

$$\hat{f}_t(x) = \rho_t \cdot h(x, \theta_t)$$

What if we could use gradient descent in space of our models?

Gradient boosting: theory



What if we could use gradient descent in space of our models?

Gradient boosting: theory

$$\hat{f}(x) = \sum_{i=0}^{t-1} \hat{f}_i(x),$$

$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}(x)}, \quad \text{for } i = 1, \ldots, n,$$

$$\theta_t = \arg\min_{\theta} \sum_{i=1}^{n} (r_{it} - h(x_i, \theta))^2,$$

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{n} L(y_i, \hat{f}(x_i) + \rho \cdot h(x_i, \theta_t))$$

Gradient boosting: theory

In linear regression case with MSE loss:

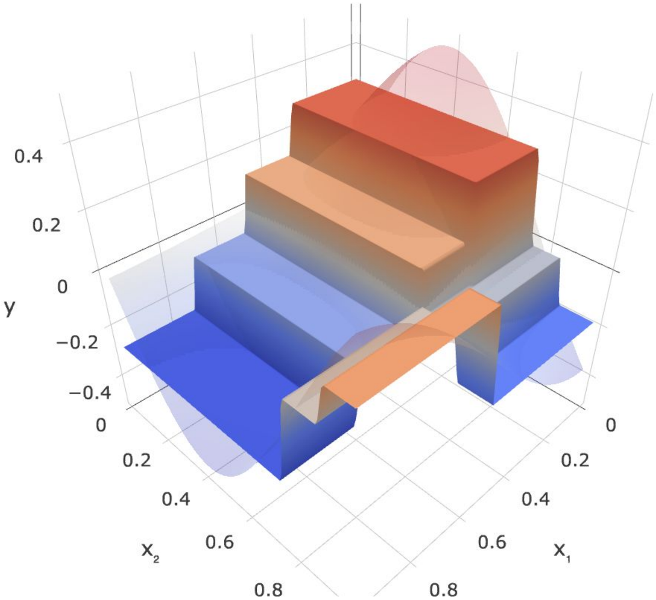
$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}(x)} = -2(\hat{y}_i - y_i) \propto \hat{y}_i - y_i$$

GB examples

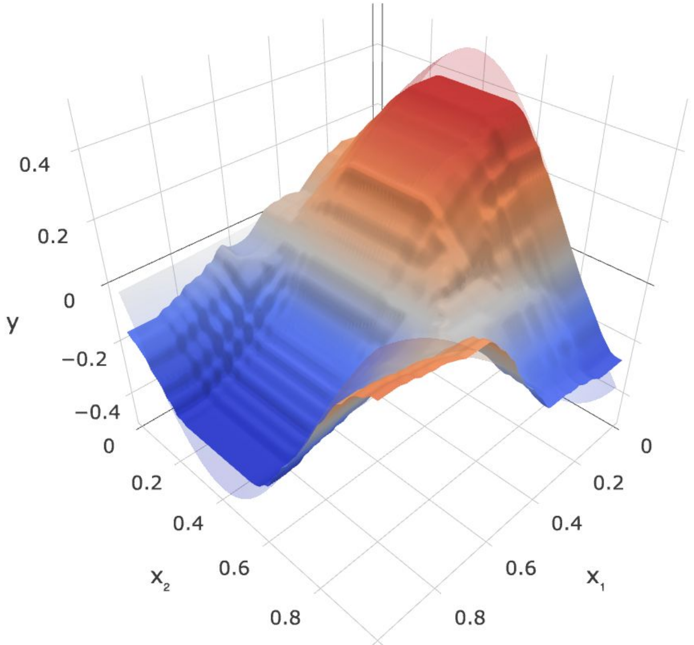
girafe
ai

03

Beautiful demo



One tree



Boosting

http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html

Gradient boosting

What we need:

- Data
- Loss function and its gradient
- Family of algorithms (with constraints if necessary)
- Number of iterations M
- Initial value (GBM by Friedman): constant

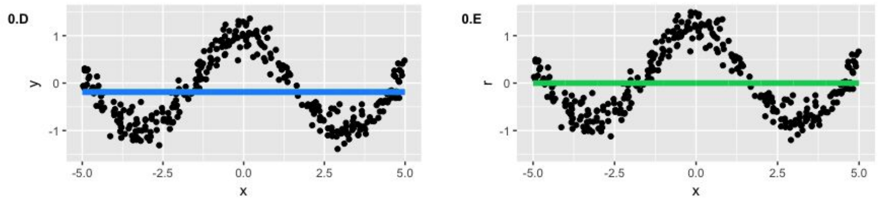
Gradient boosting: example

What we need:

- Data: toy dataset $y = cos(x) + \epsilon, \epsilon \sim \mathcal{N}(0, \frac{1}{5}), x \in [-5, 5]$
- Loss function: MSE
- Family of algorithms: decision trees with depth 2
- Number of iterations M = 3
- Initial value: just mean valu

https://habr.com/ru/company/ods/blog/327250/

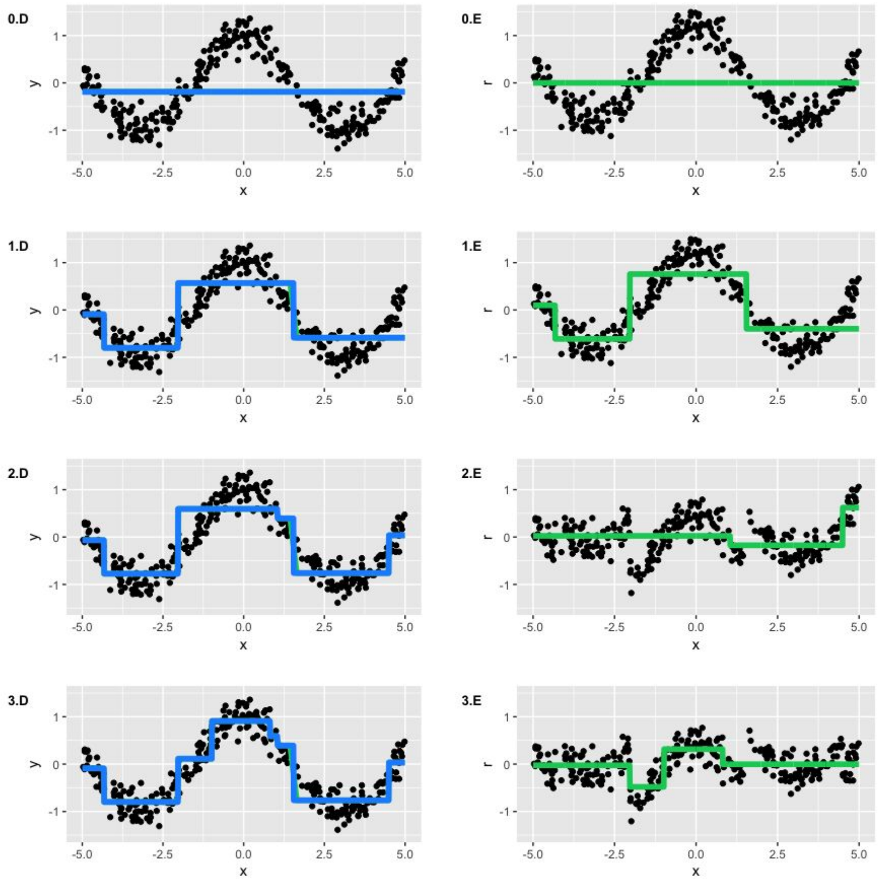
Gradient boosting: example



Left: full ensemble on each step.

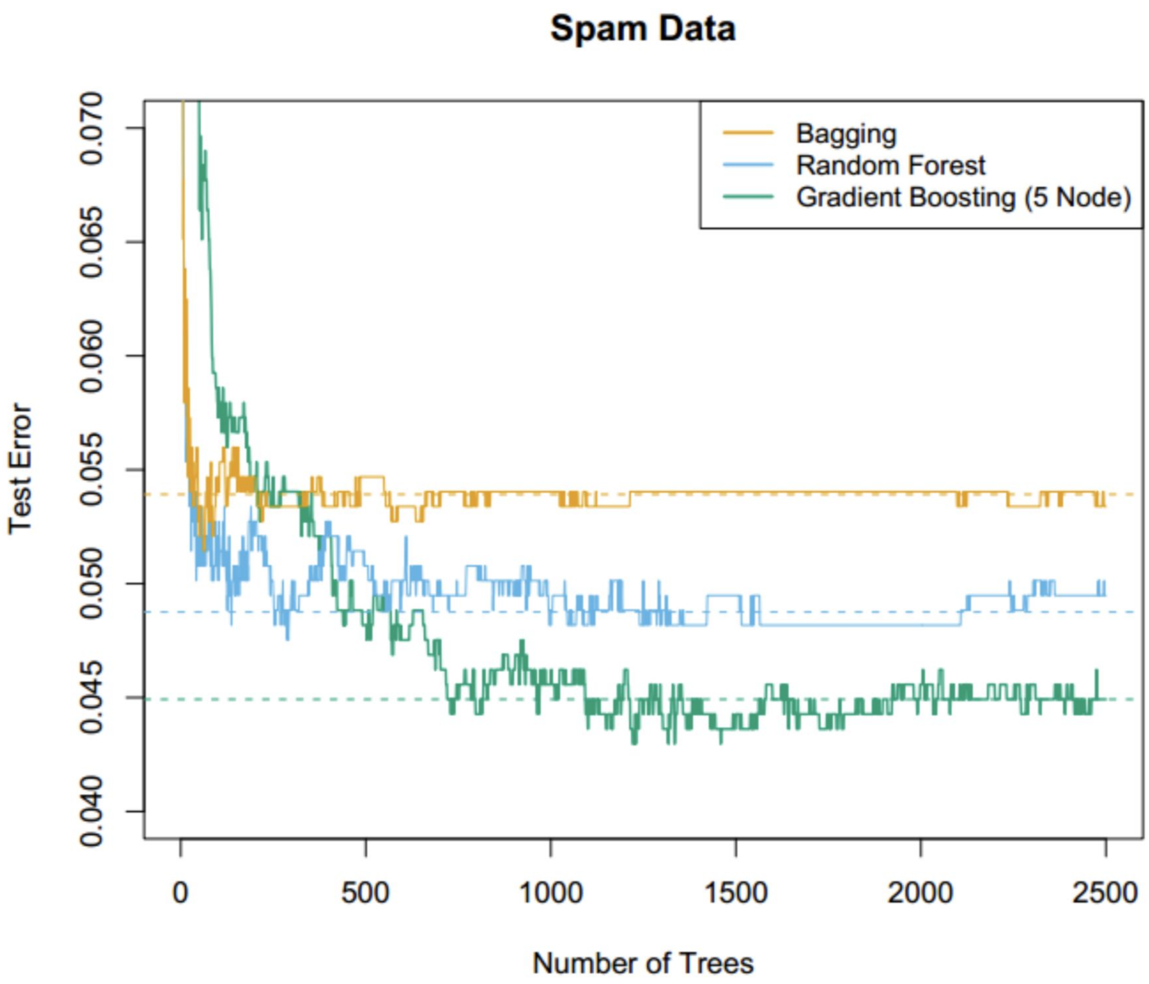
Right: additional tree decisions.

Gradient boosting: example



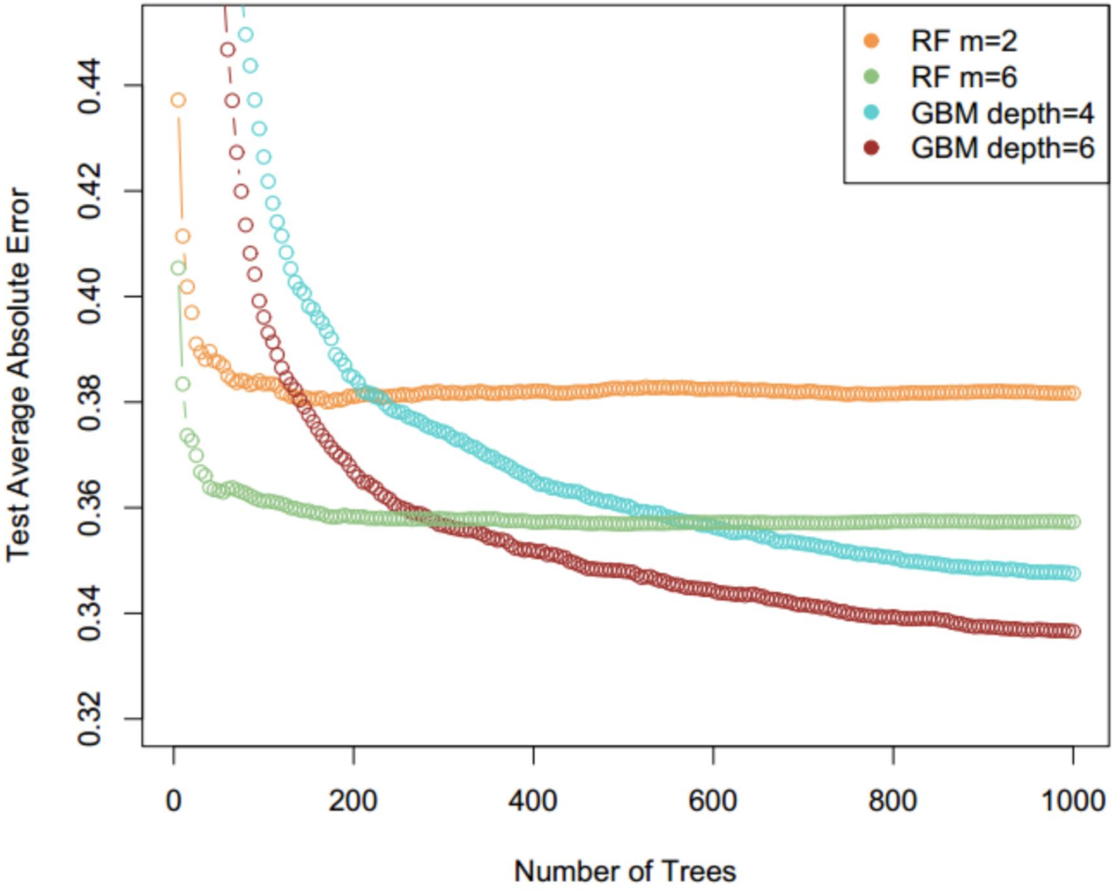
Left: full ensemble on each step.

Right: additional tree decisions

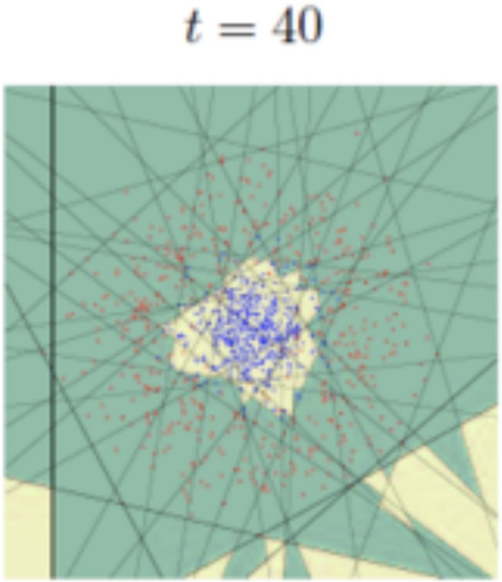


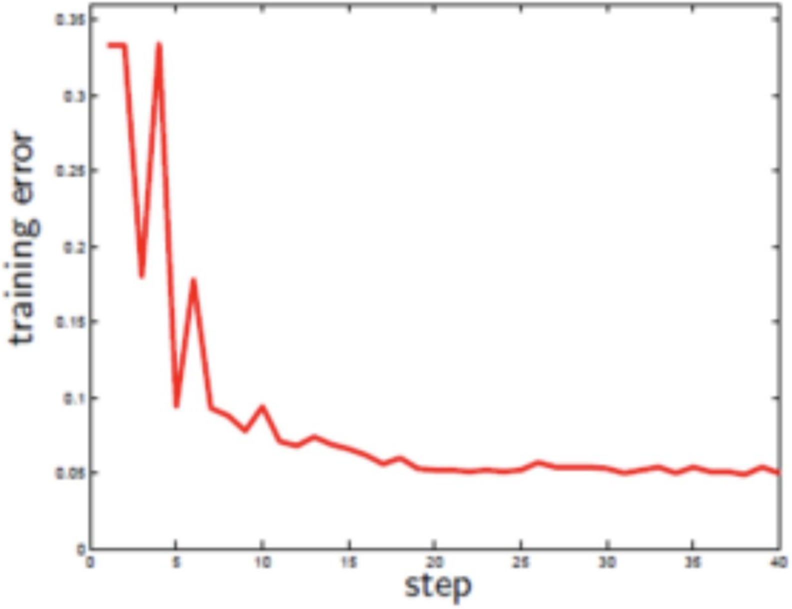
Spam Data
Bagging
Random Forest
Gradient Boosting (5 Node)
Test Error
0.040
0.045
0.050
0.055
0.060
0.065
0.070
0
500
1000
1500
2000
2500
Number of Trees

California Housing Data



Boosting with linear classification methods





Parallelization

Which of the ensembling methods could be parallelized?

- Random Forest: parallel on the forest level (all trees are independent)
- Gradient boosting: parallel on one tree level

Libraries for GB

girafe
ai

04

Main contemporary instruments

1. Catboost by Yandex
2. LightGBM by Microsoft
3. XGboost by community

Definitely not sklearn!

More on boosting

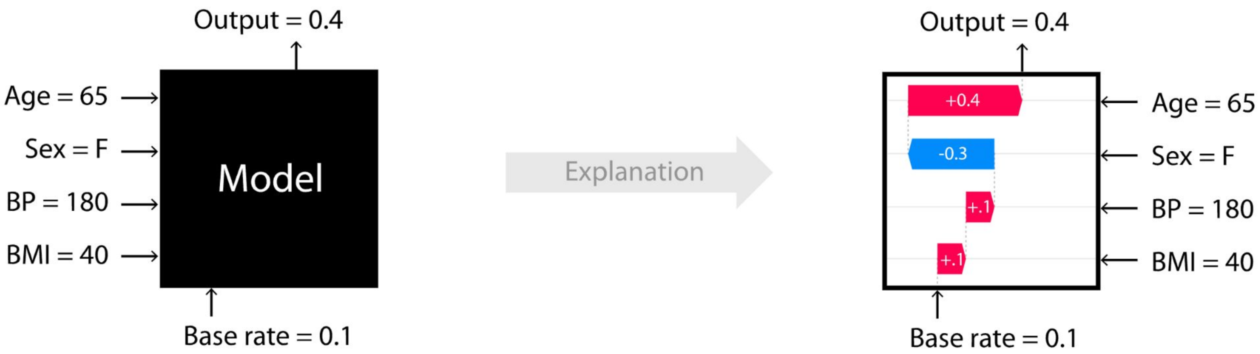
- https://habr.com/ru/companies/ods/articles/645887/
- https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm
- https://towardsdatascience.com/catboost-vs-lightgbm-vs-xgboost-c80f40662924
- https://www.springboard.com/blog/data-science/xgboost-random-forest-catboost-lightgbm/
- https://towardsdatascience.com/performance-comparison-catboost-vs-xgboost-and-catboost-vs-lightgbm-886c1c96db64

Feature importances

girafe
ai

05

Shap values



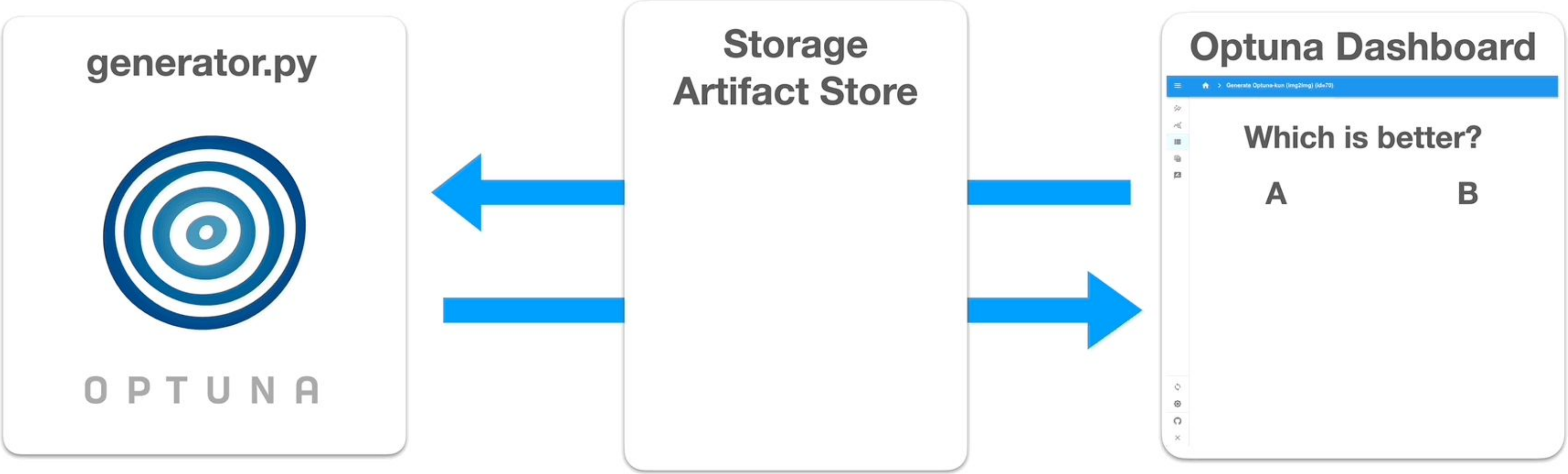
https://github.com/shap/shap

Hyperparameter optimization

girafe
ai

06

Black box or 0 order optimization



https://optuna.org/ and http://hyperopt.github.io/hyperopt/

Revise

1. Intuitions
2. Gradient boosting theory
3. Examples
4. Libraries
5. Feature importances
6. Hyperparameter optimization

Thanks for attention!

Questions?

girafe
ai