

# Композиции моделей: Boosting

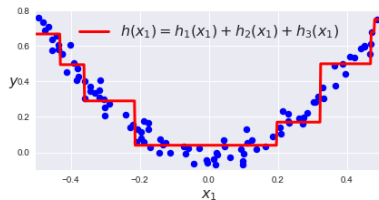
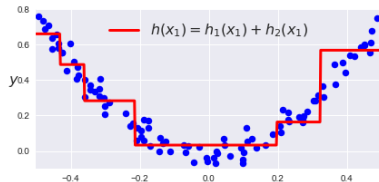
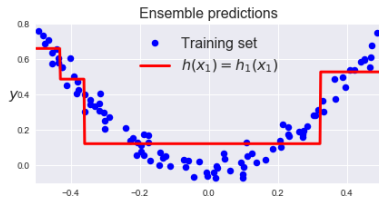
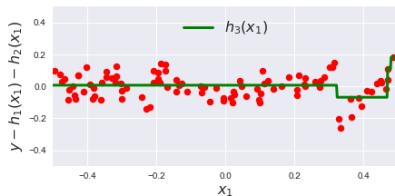
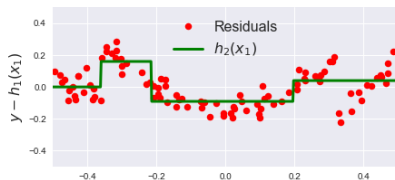
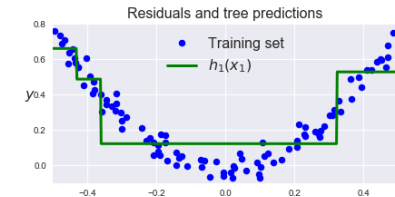
Михаил Андреев

сентябрь 2018

# Boosting : sklearn

- **GradientBoostingClassifier / GradientBoostingRegressor**
- **AdaBoostClassifier / AdaBoostRegressor**

# Boosting : the idea



# Gradient boosting

Fix parameters:  $M, \gamma$ . Initialize:  $F_0(x) = 0$ ;

**for**  $m = 1$  to  $M$  **do**

(1) Compute so-called "pseudo-residuals" (pointwise negative gradient):

$$r_{im} = - \left[ \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \dots, n. ;$$

(2) Approximate the negative gradient by fitting  $h_m(x)$  to pseudo-residuals, i.e. train it using the training set  $\{(x_i, r_{im})\}_{i=1}^n$ ;

(3) Compute multiplier  $\gamma_m$  by solving one-dimensional optimization problem:

$$\gamma_m = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) .;$$

(4)  $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$  ;

**end**

**Result:**  $F_M(x)$

# Gradient boosting

[https://en.wikipedia.org/wiki/Gradient\\_boosting](https://en.wikipedia.org/wiki/Gradient_boosting)

<https://github.com/esokolov/ml-course-hse/blob/master/2017-fall/lecture-notes/lecture09-nsembles.pdf>

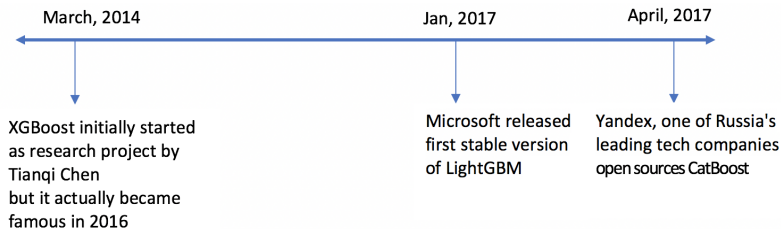
# Boosting: простой пример : подбор параметров

см. `ipy nb`

# Boosting : модификации

- Рандомизация в духе RF: bagging / random subspaces
- Модификации алгоритма построения дерева.

# Boosting: libraries



<https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db>



## Чужой опыт (1)

"Скорость обучения для CatBoost надо задавать больше, чем аналогичный параметр для XGBoost и LightGBM."

<https://kelijah.livejournal.com/216093.html>

"кажется, что CatBoost справляется лучше других алгоритмов градиентного бустинга только с большим количеством деревьев (четырёхзначного порядка). Конечно, сравнивать этот параметр вообще некорректно — Кэтбуст использует другие деревья, ему их нужно больше." <http://nlpx.net/archives/536>

Boosting: попытка показать более сложный пример :  
опыт хождения по граблям

см. ipynb

## Чужой опыт (2)

<https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db>

# Капитан Очевидность на линии

- Похожие, даже одноименные, параметры в разных реализациях имеют разные оптимумы
- Ограничение - не только ЦПУ, но и память.
- версия 0.2 менее оптимизирована, чем версия 0.9.1.1
- know your tools

Спасибо за внимание

