

Машинное обучение

Лекция 05. Ансамбли решающих деревьев

Драль Алексей

<https://www.linkedin.com/in/alexey-dral>

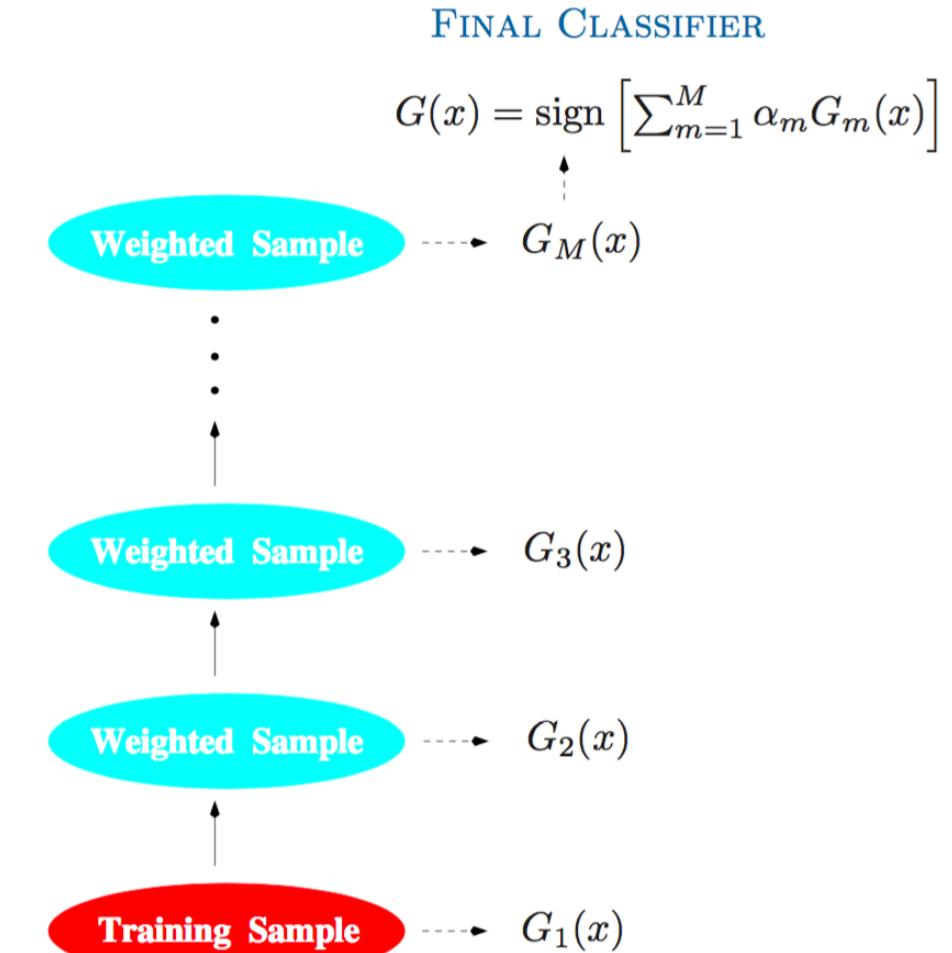
13.03.2018, Москва, ФИВТ МФТИ

План

- Закрепление бустинга: тонкости реализации
- Обобщение до Gradient Tree Boosting / GBDT / GBM / MART
- Эвристики оптимизации и state-of-the-art алгоритмы

AdaBoost.M1 (1997, Freund, Schapire)

- Задача бинарной классификации: $Y \in \{-1, 1\}$
- Def. слабый классификатор (weak classifier)
– ошибка предсказания чуть лучше, чем у случайного угадывания.
- Цель бустинга – последовательное применение слабых классификаторов на измененных версиях данных для получения (сильной) композиции сильных агтов.



Algorithm 10.1 AdaBoost.M1.

1. Initialize the observation weights $w_i = 1/N, i = 1, 2, \dots, N$.
 2. For $m = 1$ to M :
 - (a) Fit a classifier $G_m(x)$ to the training data using weights w_i .
 - (b) Compute
$$\text{err}_m = \frac{\sum_{i=1}^N w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i}.$$
 - (c) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$.
 - (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))], i = 1, 2, \dots, N$.
 3. Output $G(x) = \text{sign} \left[\sum_{m=1}^M \alpha_m G_m(x) \right]$.
-

Algorithm 10.1 AdaBoost.M1.

1. Initialize the observation weights $w_i = 1/N, i = 1, 2, \dots, N$.
 2. For $m = 1$ to M :
 - (a) Fit a classifier $G_m(x)$ to the training data using weights w_i .
 - (b) Compute
$$\text{err}_m = \frac{\sum_{i=1}^N w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i}.$$
 - (c) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$.
 - (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))], i = 1, 2, \dots, N$.
 3. Output $G(x) = \text{sign} \left[\sum_{m=1}^M \alpha_m G_m(x) \right]$.
-



Откуда взялся sign

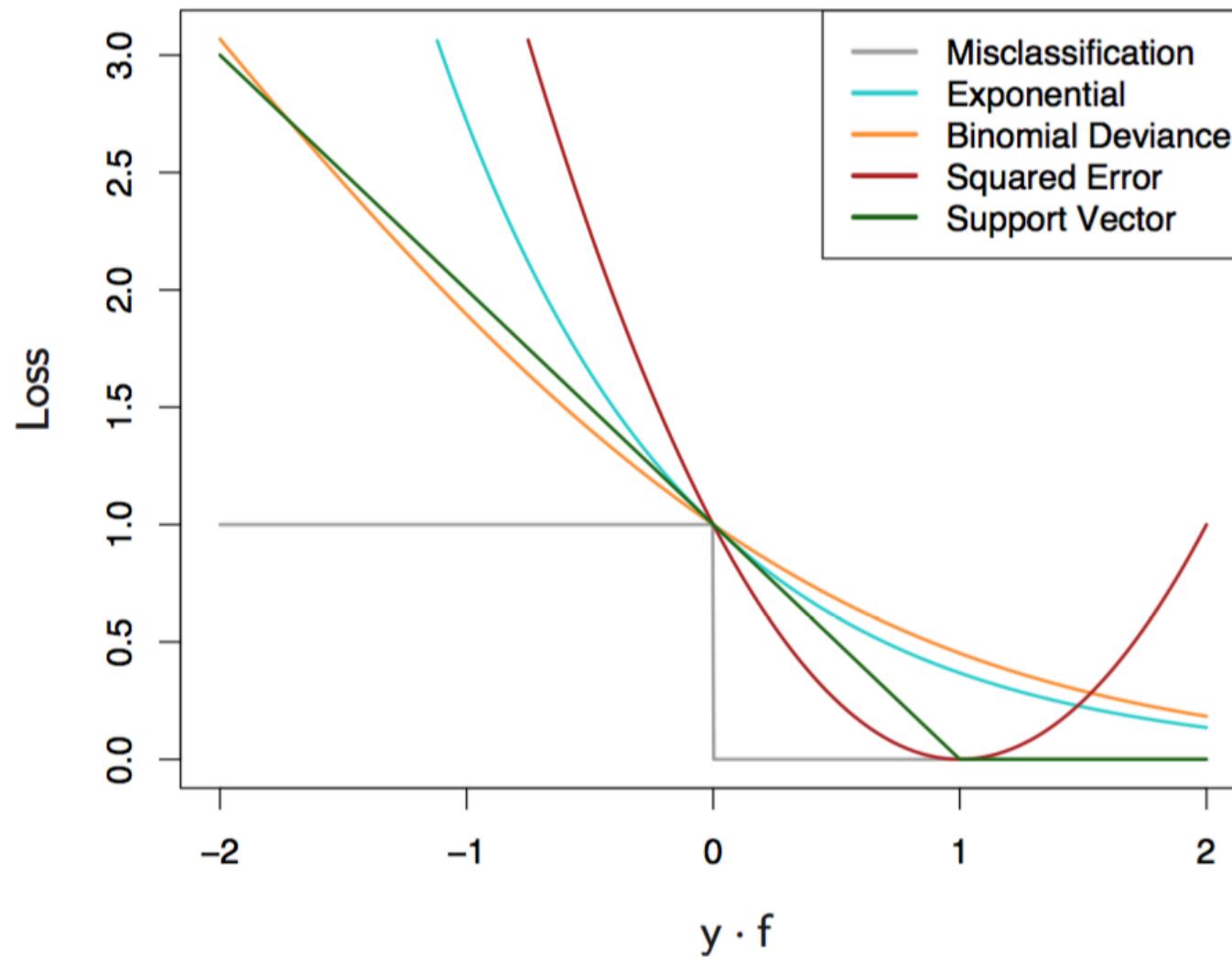
$$f^*(x) = \operatorname{argmin}_{f(x)} E(e^{-Y f(x)}) = \frac{1}{2} \log \left(\frac{\Pr(Y = 1|x)}{\Pr(Y = -1|x)} \right)$$

Откуда взялся sign

$$f^*(x) = \operatorname{argmin}_{f(x)} E(e^{-Y f(x)}) = \frac{1}{2} \log \left(\frac{\Pr(Y = 1|x)}{\Pr(Y = -1|x)} \right)$$

$$f^*(x) = \operatorname{argmin}_{f(x)} E(1 + e^{-2 Y f(x)}) = \frac{1}{2} \log \left(\frac{\Pr(Y = 1|x)}{\Pr(Y = -1|x)} \right)$$

На заметку Data Scientist'у



Algorithm 10.2 *Forward Stagewise Additive Modeling.*

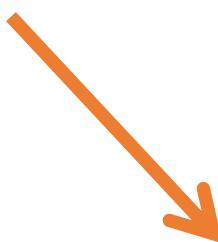
1. Initialize $f_0(x) = 0$.

2. For $m = 1$ to M :

(a) Compute

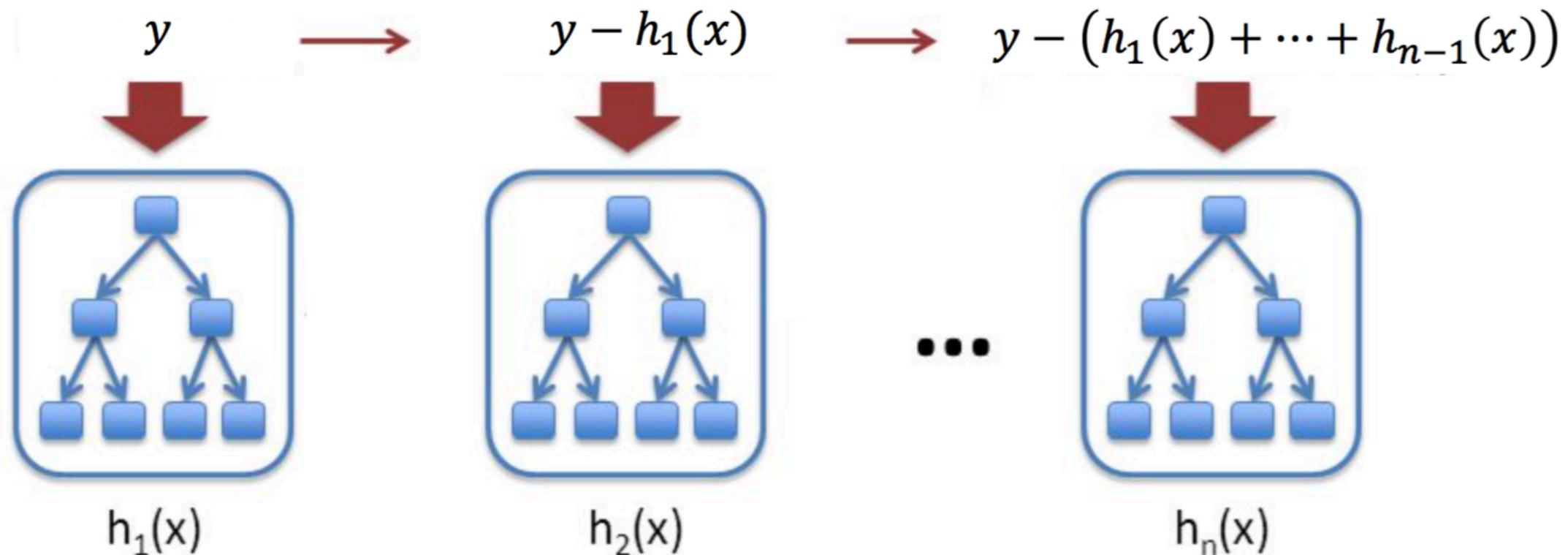
$$(\beta_m, \gamma_m) = \arg \min_{\beta, \gamma} \sum_{i=1}^N L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma)).$$

(b) Set $f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$.



AdaBoost \sim Forward Stagewise Additive Modeling при $L(y, f(x)) = \exp(-y f(x))$

Какая функция потерь?



Более общая формулировка

$$f(x) = \sum_{m=1}^M \beta_m b(x; \gamma_m)$$

$$\min_{\{\beta_m, \gamma_m\}_1^M} \sum_{i=1}^N L \left(y_i, \sum_{m=1}^M \beta_m b(x_i; \gamma_m) \right)$$

Более общая формулировка

$$f(x) = \sum_{m=1}^M \beta_m b(x; \gamma_m)$$

$$\min_{\{\beta_m, \gamma_m\}_1^M} \sum_{i=1}^N L \left(y_i, \sum_{m=1}^M \beta_m b(x_i, \gamma_m) \right)$$

$$\min_{\beta, \gamma} \sum_{i=1}^N L(y_i, \beta b(x_i; \gamma))$$

Algorithm 10.3 *Gradient Tree Boosting Algorithm.*

1. Initialize $f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$.

2. For $m = 1$ to M :

(a) For $i = 1, 2, \dots, N$ compute

$$r_{im} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}.$$

(b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm} , $j = 1, 2, \dots, J_m$.

(c) For $j = 1, 2, \dots, J_m$ compute

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

3. Output $\hat{f}(x) = f_M(x)$.

Как до этого дошли: мысль первая

$$T(x; \Theta) = \sum_{j=1}^J \gamma_j I(x \in R_j)$$

$$\Theta = \{R_j, \gamma_j\}_1^J$$

$$\hat{\Theta} = \arg \min_{\Theta} \sum_{j=1}^J \sum_{x_i \in R_j} L(y_i, \gamma_j)$$

Как до этого дошли: мысль первая

$$T(x; \Theta) = \sum_{j=1}^J \gamma_j I(x \in R_j)$$

$$\Theta = \{R_j, \gamma_j\}_1^J$$

- (1) Зная R_j найти γ_j
- (2) Найти R_j

$$\hat{\Theta} = \arg \min_{\Theta} \sum_{j=1}^J \sum_{x_i \in R_j} L(y_i, \gamma_j) \longrightarrow \tilde{\Theta} = \arg \min_{\Theta} \sum_{i=1}^N \tilde{L}(y_i, T(x_i, \Theta))$$

Обновления в алгоритме 10.2

$$f_M(x) = \sum_{m=1}^M T(x; \Theta_m) \quad \Theta_m = \{R_{jm}, \gamma_{jm}\}_1^{J_m}$$

$$\hat{\Theta}_m = \arg \min_{\Theta_m} \sum_{i=1}^N L(y_i, f_{m-1}(x_i) + T(x_i; \Theta_m)) \quad (*)$$

$$\hat{\gamma}_{jm} = \arg \min_{\gamma_{jm}} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma_{jm}) \quad (**)$$

Мысль вторая – численная оптимизация

$$L(f) = \sum_{i=1}^N L(y_i, f(x_i))$$



$$\hat{\mathbf{f}} = \arg \min_{\mathbf{f}} L(\mathbf{f})$$

$$\mathbf{f} = \{f(x_1), f(x_2), \dots, f(x_N)\}^T$$

$$\mathbf{f}_M = \sum_{m=0}^M \mathbf{h}_m, \quad \mathbf{h}_m \in \mathbb{R}^N$$

Метод наискорейшего спуска

$$\mathbf{h}_m = -\rho_m \mathbf{g}_m \quad \mathbf{g}_m \in \mathbb{R}^N$$

$$g_{im} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x_i)=f_{m-1}(x_i)}$$

$$\rho_m = \arg \min_{\rho} L(\mathbf{f}_{m-1} - \rho \mathbf{g}_m)$$

$$\mathbf{f}_m = \mathbf{f}_{m-1} - \rho_m \mathbf{g}_m$$

Аналогии: что более гибко?

$$\mathbf{h}_m = -\rho_m \mathbf{g}_m \quad \mathbf{g}_m \in \mathbb{R}^N$$

$$g_{im} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x_i)=f_{m-1}(x_i)}$$

$$\rho_m = \arg \min_{\rho} L(\mathbf{f}_{m-1} - \rho \mathbf{g}_m)$$

$$\mathbf{f}_m = \mathbf{f}_{m-1} - \rho_m \mathbf{g}_m$$

$-T(x_i; \Theta_m)$, см. (*)

γ_{jm} , см. (**)

Градиентный бустинг (Gradient Boosting)

$$\tilde{\Theta}_m = \arg \min_{\Theta} \sum_{i=1}^N (-g_{im} - T(x_i; \Theta))^2$$

Получим другие R_{jm} , чем при решении (*), но построение дерева тоже аппроксимация.

Algorithm 10.3 *Gradient Tree Boosting Algorithm.*

1. Initialize $f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$.

2. For $m = 1$ to M :

(a) For $i = 1, 2, \dots, N$ compute

$$r_{im} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}.$$

(b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm} , $j = 1, 2, \dots, J_m$.

(c) For $j = 1, 2, \dots, J_m$ compute

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

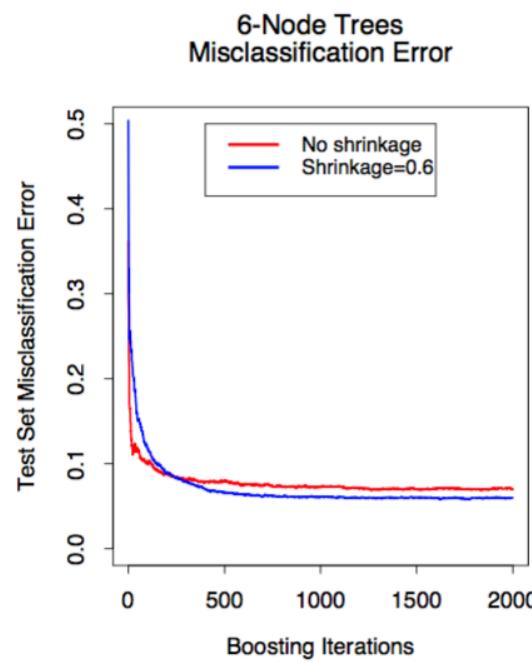
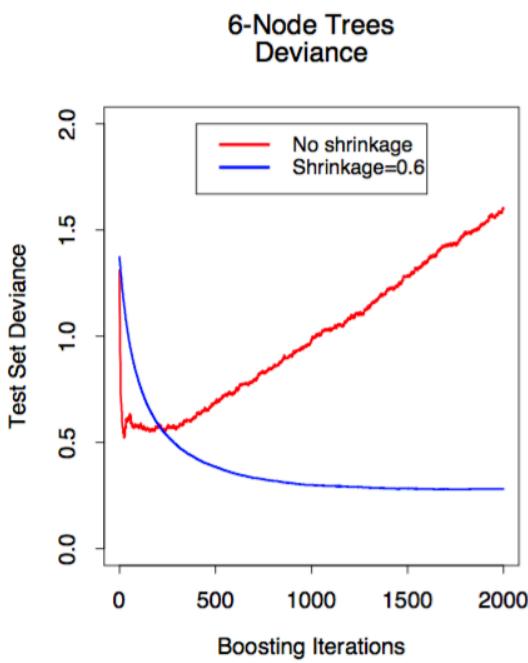
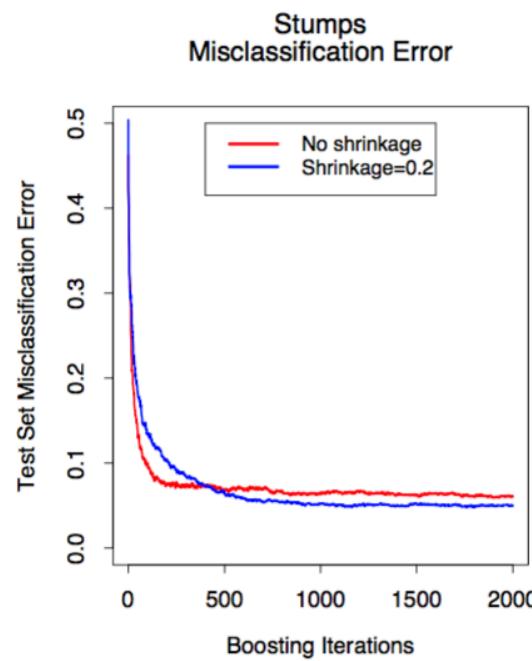
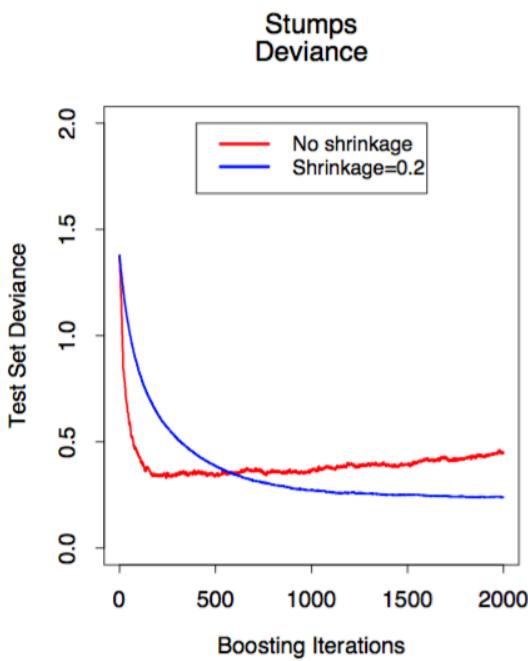
3. Output $\hat{f}(x) = f_M(x)$.

Регуляризация градиентного бустинга

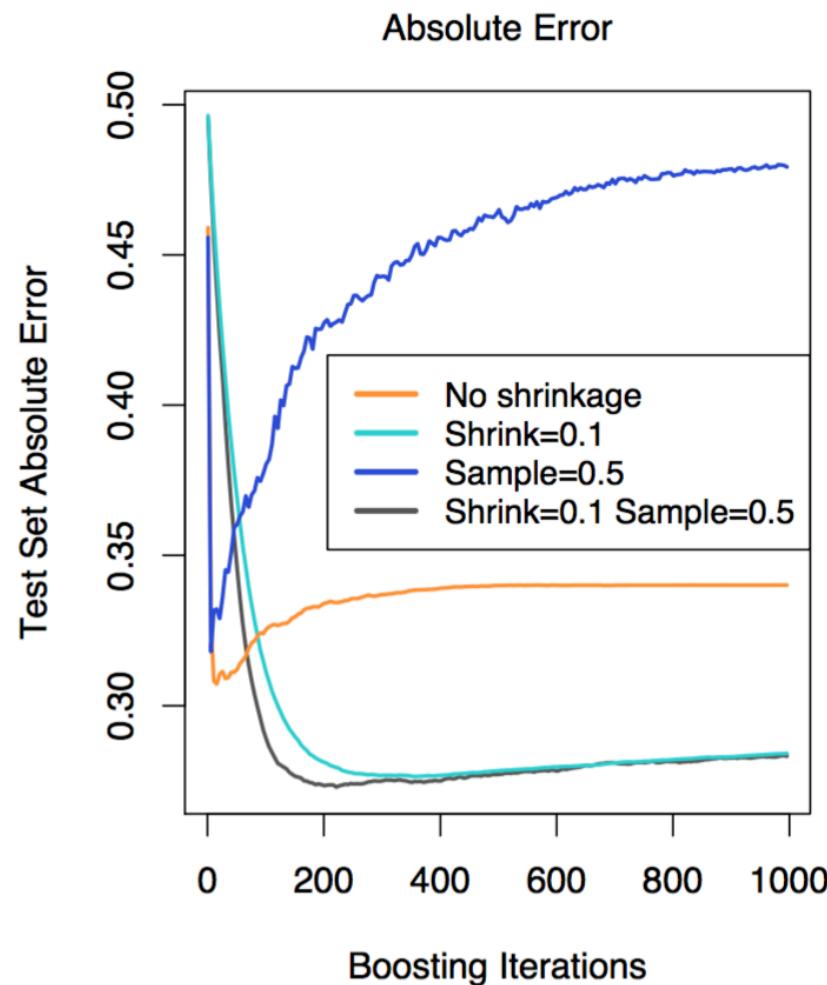
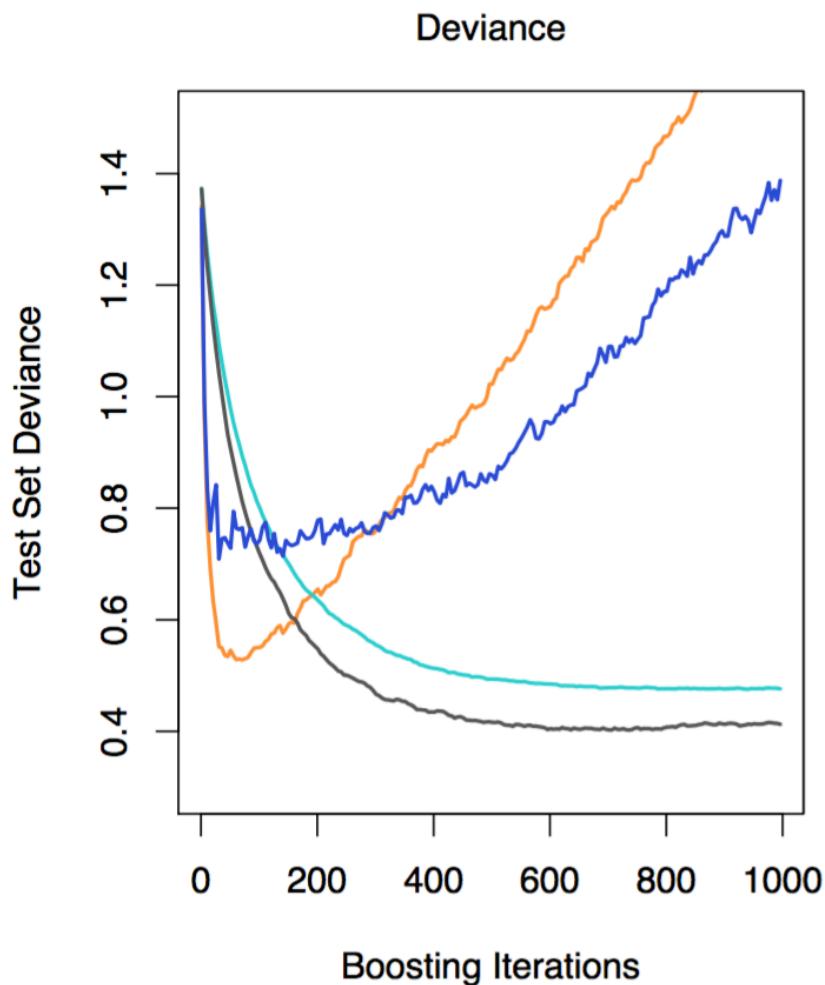
$$(1) \quad f_m(x) = f_{m-1}(x) + \nu \cdot \sum_{j=1}^J \gamma_{jm} I(x \in R_{jm})$$

$0 < \nu < 1$ - learning rate (скорость обучения)

(2) стохастический градиентный спуск (subsampling w/o replacement)



4–Node Trees



Q&A «проверка связи»: какие две
отличительные черты в Gradient Tree Boosting?



Резюме

- Gradient Boosted Model (GBM, Ridgeway - 1999) в отличие от AdaBoost можно сделать робастной к пересекающимся классам (в частности к ошибкам разметки).
- Полезные эвристики:
 - ставить маленький learning rate (но соответственно дольше обучать)
 - использовать subsampling (но аккуратно)
- Численная оптимизация второго порядка – XGBoost (2016), LightGBM (powered by Microsoft, 2017)

Резюме лекций 3-5 (модуль «деревья»)



Decision Tree



Random Forest



Gradient Boosted Model

Обратная связь

Отзывы о прошедших лекциях и семинарах просьба оставлять здесь:

https://ml-mipt.github.io/2018part1_Schedule/

Полезные материалы

- материалы курса ФИВТ МФТИ (включая архивы прошлых лет) -
<https://github.com/ml-mipt/ml-mipt-part1>
- Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. (Chapter 10. Boosting)
- James, G., Witten, D., Hastie, T., Tibshirani, R. An Introduction to Statistical Learning with Applications in R (Chapter 8. Tree-Based Methods)
- материалы курса ФКН ВШЭ -
http://wiki.cs.hse.ru/Машинное_обучение_1
- XGBoost (2016): <https://arxiv.org/abs/1603.02754>; LightGBM (2017) -
<https://github.com/Microsoft/LightGBM>

Задача со звездочкой #1: K-class AdaBoost

Цель: Написать алгоритм (аналогичный 10.1) и доказать его правильность для задачи классификации с К классами (см. задание 10.5 из ESL)

Бонус: 2 балла (первым Зм приславшим)

Оформление: электронный документ (не скан / фото)

Тема письма: “ML-2018, ensemble, K-class AdaBoost, ФИО (группа)”

Высыпать по адресу: aadral@gmail.com