Деревья решений

Кручинин Дмитрий

Научный руководитель: Кураленок И.Е.

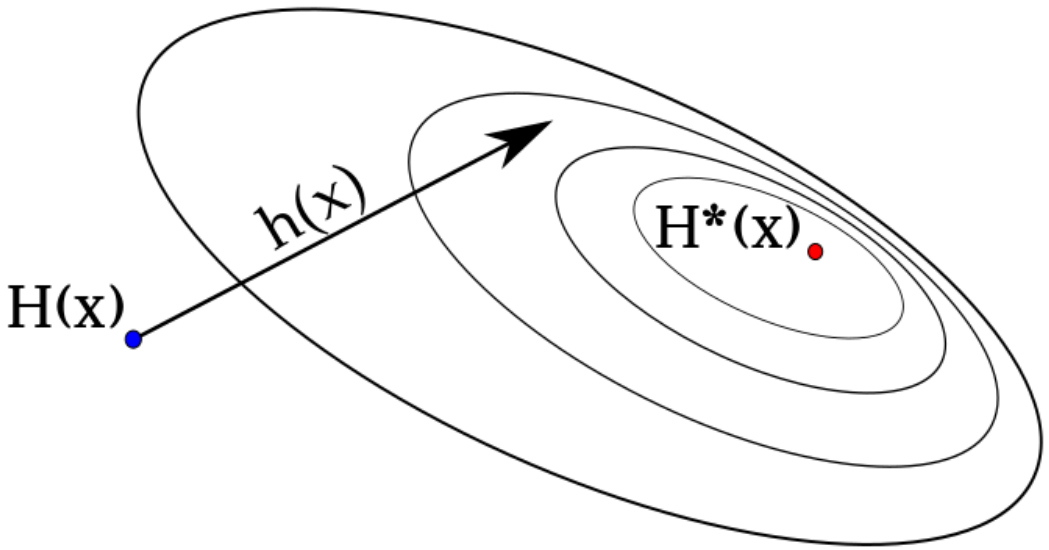
СПбАУ

2 июня 2017 г.

Outline

Градиентный бустинг, обозначения

Датасет: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$.
Решающая функция: $H_T(x) = \sum_{t=1}^T h_t(x)$, где h_t – дерево решений.
Функция потерь: $\mathcal{L}(x, H(x))$.
h_{t+1} строится в соответствии с градиентом $\left.\frac{\partial \mathcal{L}}{\partial H_t}\right|_{H(X)}$.
$H_{t+1}(X) = H_t(X) + \eta h_{t+1}(X)$, $0 < \eta \leq 1$

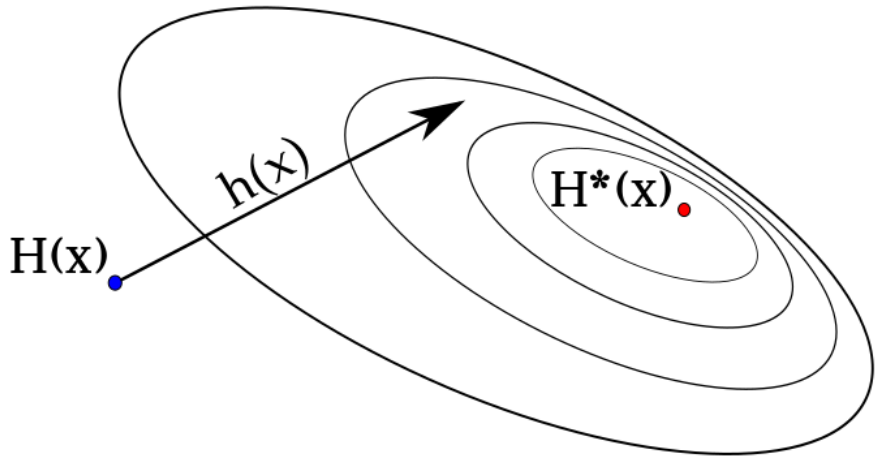


Сходимость и связь со средним

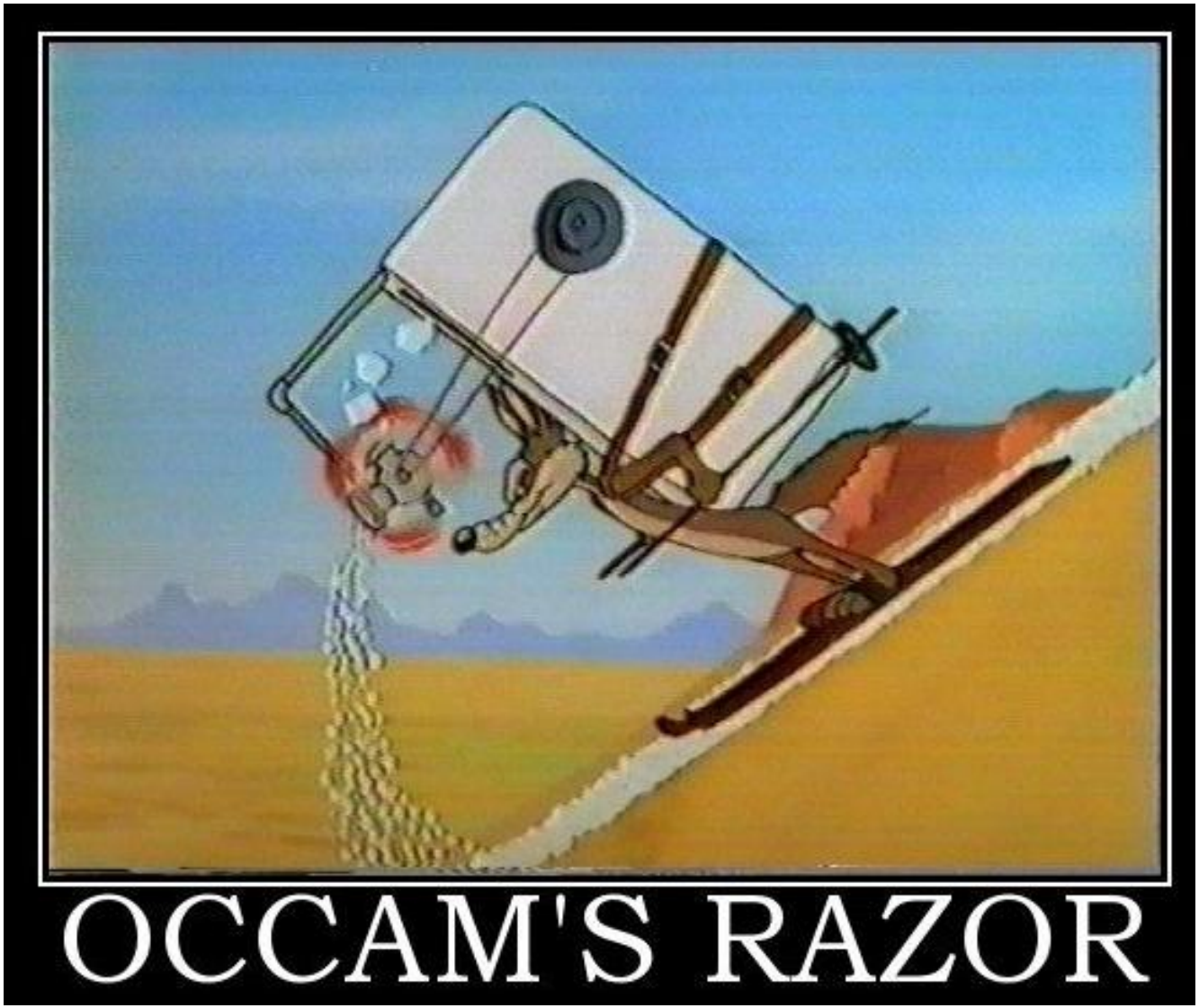
Jerome H. Friedman: "*Decreasing the learning rate increases the best score, but requier more weak learners.*"

- Шаг в бустинге $h_{t+1}(X)$ – это среднее значение (μ).
- Получаем: $p(\mu) \sim \mathcal{N}\left(0, \frac{1}{\tau}\right)$ и $X \sim \mathcal{N}\left(\mu, \frac{1}{\tau}\right)$
- Сопряженная байесовская оценка:

$$\widehat{\mu} = \frac{n\tau\overline{X} + \tau \cdot 0}{n\tau + \tau} = \overline{X}\frac{n}{n+1}$$



Minimum description length



Информационная регуляризация

- Во-первых, обычная регуляризация

$$p(h_{t+1}|X, H_t(X)) = p(X|h_{t+1}, H_t(X)) \cdot p(h_{t+1}|H_t(X)) =$$

$$= p(X|h_{t+1}, H_t(X)) \cdot p(h_{t+1})$$

- Посмотрим на бустинг с точки зрения информации:

$$I\left(H_{t+1}(X)\right) = I\left(H_t(X) + h_{t+1}(X)\right) \leq I(H_t(X)) + I(h_{t+1}(X)) \downarrow$$

- Добавим условие $I(H_{t+1}(X)) \leq I_0$, и из KKT получаем

$$\mathcal{L}(X, H_t(X)) + \Omega(h_{t+1}) + I(H_{t+1}(X))$$

Информационная регуляризация. Как считать $I(h(X))$?

Будем считать $I(h(X))$ будто мы сжимаем $h(X)$ энтропийным кодированием.

$$h(X) = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Для x_i попавших в один лист y_i – одинаковый.

$$I(h(X)) = \sum_{i=1}^{L} l_i \log l_i$$

l_i – количество объектов в i-том листе

Результаты. Датасет Higgs

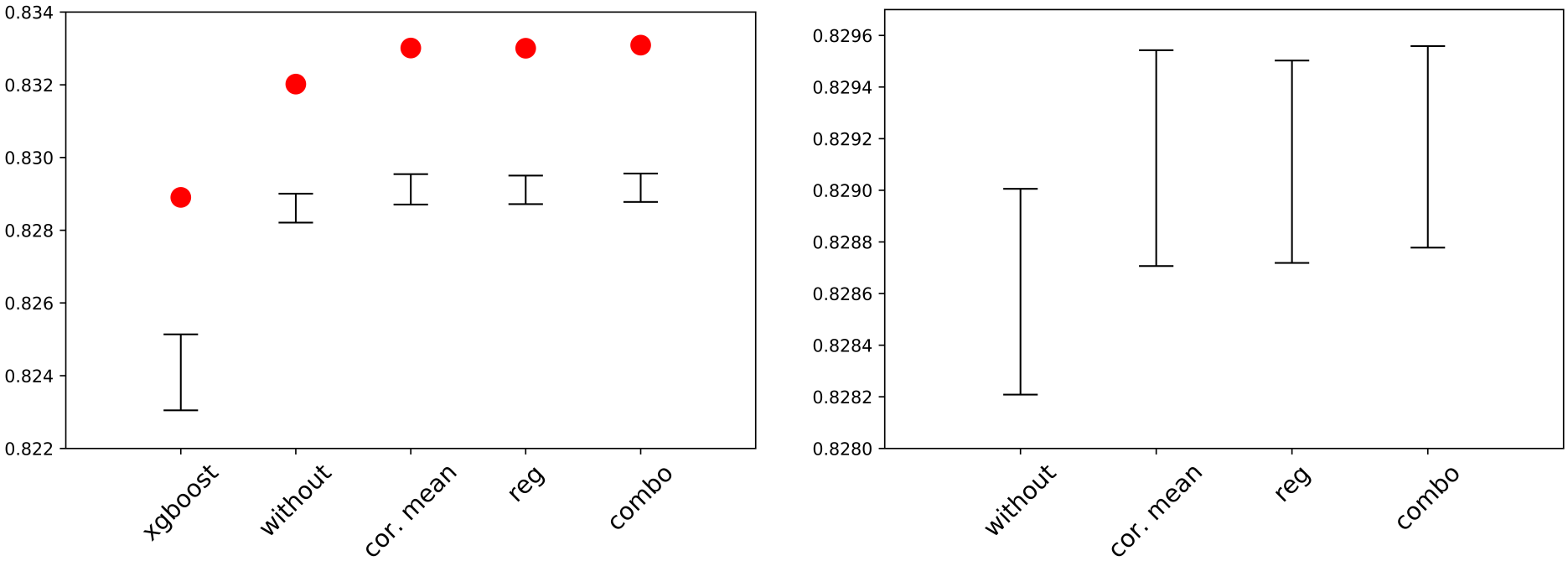


Рис. 1: AUC, размер выборки $\sim$ 1M, кол-во признаков 29

Результаты

Плохие...

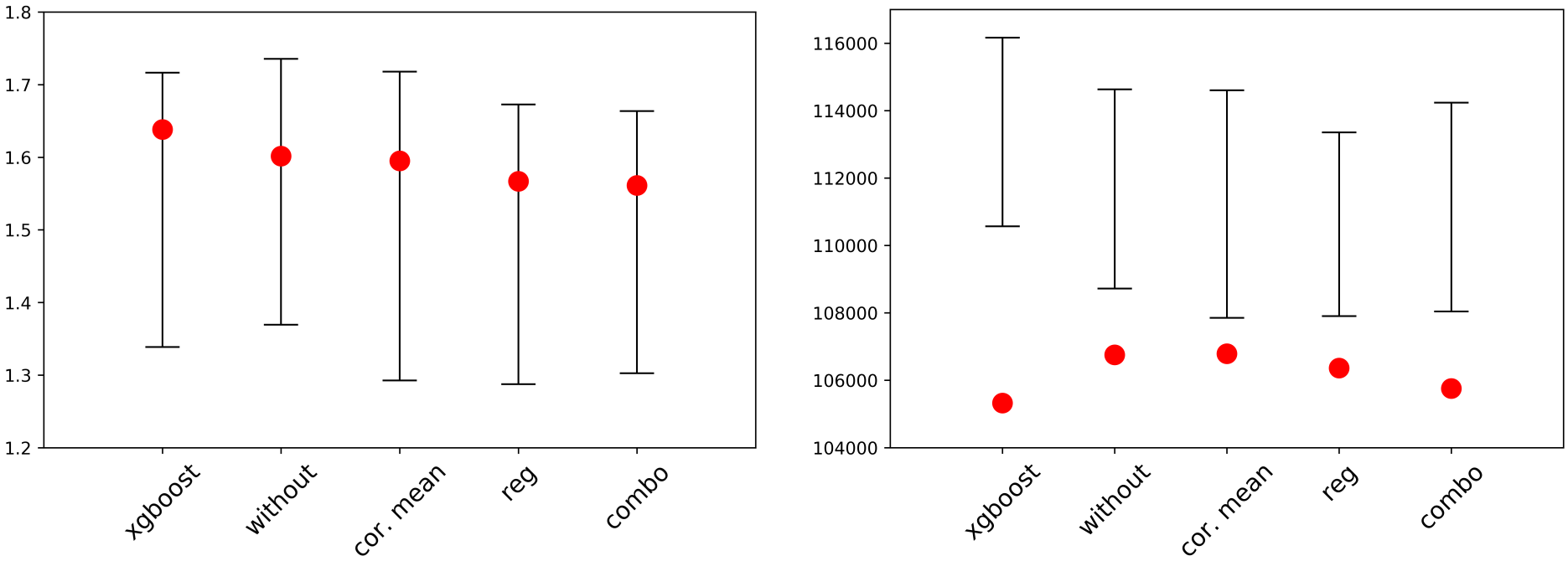


Рис. 2: RMSE, слева датасет CT slices: размер выборки – 53500, 386 признаков
справа – KC house pricing: размер – 21613, 19 признаков

Борьба с random seed. Kendall rank correlation

- Random seed после bootstrap'а?
- "Elen Voorhes: Variations in relevance judgements and the measurement of retrieval effectiveness".
- Возьмем seed как отдельного судью, посчитаем kendall-tau:

0.5533 – ct slice, с порядком: reg → corr. mean → without → xgb.
0.45 – kc house, с тем же порядком. От проблемы до конца не избавились.

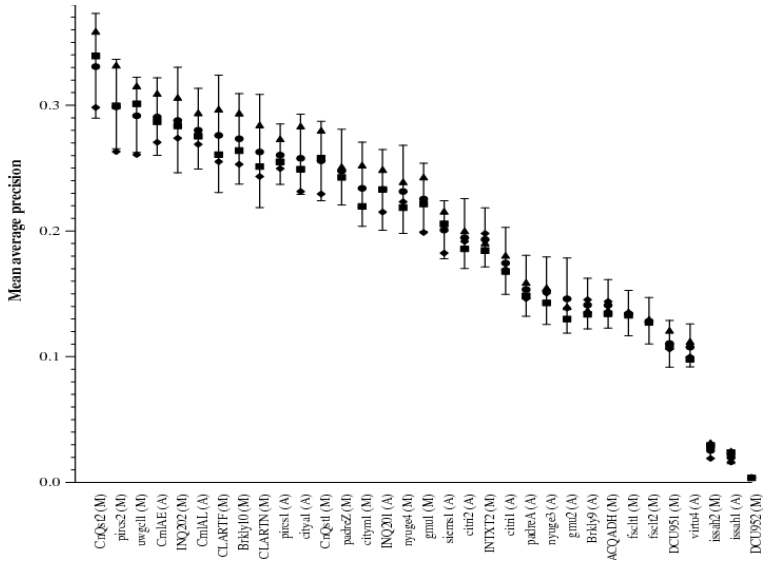


Рис. 3: Confidence intervals for different systems on TREC-4 dataset

Борьба с random seed. Wilcoxon test

	xgboost	**without**	**corr. mean**	**regularization**
xgboost	1.	6.83e-01	1.34e-02	9.77e-17
without	6.83e-01	1.	4.15e-02	4.56e-15
corr. mean	1.34e-02	4.15e-02	1.	1.65e-08
regularization	9.77e-17	4.56e-15	1.65e-08	1.

Таблица 1: Wilcoxon rank sum test. P-values matrix. CT slices.

	xgboost	**without**	**corr. mean**	**regularization**
xgboost	1.	5.28e-08	2.88e-09	2.15e-11
without	5.28e-08	1.	5.30e-02	4.89e-02
corr. mean	2.88e-09	5.30e-02	1.	3.71e-02
regularization	2.15e-11	4.89e-02	3.71e-02	1.

Таблица 2: Wilcoxon rank sum test. P-values matrix. KC house pricing.