Градиентный бустинг смешанных моделей с последовательным усложнением

Шабанов Андрей Александрович, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель — д.ф.-м.н., проф. М. С. Ермаков
Рецензент: ассистент А. Ю. Шлемов

Санкт-Петербург
2014г.

Цели работы

Цели работы

- Построение схем градиентного бустинга с использованием моделей разного типа.
- Анализ достоинств и недостатков предложенных схем по сравнению с стандартными моделями на примере модельных и реальных данных.
- Получение метода оценки характера зависимостей в данных на основе предложенных схем.

План работы

План работы

- Описание GBM
- Описание предложенных алгоритмов.
- Сравнение с классическими схемами
  - на одномерных модельных данных.
  - на многомерных модельных данных.
  - на реальных данных.

Постановка задачи

Входные данные

- $x = (x_1, \ldots, x_p)$, $x_j \in R^N$ - независимые переменные
- Y — выходная переменная
- $\rho(Y, f(x))$ — заданная функция потерь

Требуется построить функцию f:

$$Y_i = f(x_{i_1}, \ldots, x_{i_p}) + \varepsilon, \quad \forall i = \overline{1, N}$$

$$\hat{f}(x) = \underset{f(x)}{\operatorname{argmin}} \, \mathbb{E}_{Y,x} \rho(Y, f(x)).$$

Рассматриваем случай МНК-регрессии:

$$Y \in R, \quad \rho = \frac{1}{2}|Y - f|^2$$

Метод градиентного бустинга

Общий подход бустинговых алгоритмов

$$\hat{f}(x) = \sum_{t=0}^{M} \hat{f}_t(x), \quad \hat{f}_t \leftarrow \hat{f}_{t-1} + \nu_t h(x, \theta_t),$$

$$(\nu_t, \theta_t) = \underset{\nu,\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \rho(Y_i, \hat{f}_{t-1} + \nu h(x_i, \theta)).$$

Метод градиентного бустинга (Friedman, 1999)

- $\hat{f}_0 \equiv \operatorname{argmin}_c \frac{1}{N} \sum_{i=1}^{N} \rho(Y_i, c)$

Для $t = 1, \ldots, M$:

- $\hat{U}_{it} = -\frac{\partial}{\partial f} \rho(Y_i, f)|_{f=\hat{f}_{t-1}(x_i)}, \quad i = 1, ..., N.$
- $\theta_t = \operatorname{argmin}_{\theta,\nu} \sum_{i=1}^{N} [\hat{U}_i - \nu h(x_i, \theta)]^2.$
- $\nu_t = \operatorname{argmin}_{\nu} \sum_{i=1}^{N} \rho(Y_i, \hat{f}_{t-1}(x_i) + \nu h(x_i, \theta_t)).$
- $\hat{f}_t = \hat{f}_{t-1} + \nu_t h(x, \theta_t)$

Метод градиентного бустинга

Градиентный бустинг одномерными моделями в L_2 случае:

$\rho = \frac{1}{2}|Y - f|^2 \Rightarrow \hat{U}_t = Y - \hat{f}_{t-1}.$

- *Начальное приближение* $\hat{f}_0 = \frac{1}{N}\sum Y_i$

Для $t = 1, \ldots, M$*:*

- *Для* $j = 1, \ldots, P$ *независимо строим базовые модели* $h(x_j, \theta_t)$*:*

$$\theta_t = \operatorname{argmin} \sum_{i=1}^{N} (\hat{U}_t - h(x_j, \theta))^2.$$

- *Выбираем лучшую модель:*

$$j_{best} = \operatorname*{argmin}_j \sum_{i=1}^{N} (\hat{U}_t - h(x_j, \theta))^2.$$

- $\hat{f}_t = \hat{f}_{t-1} + h(x_{j_{best}}, \theta_t).$

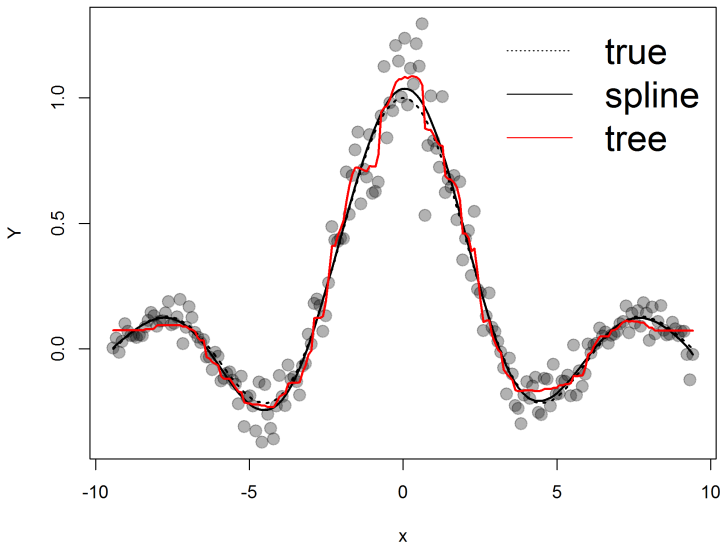
Метод градиентного бустинга

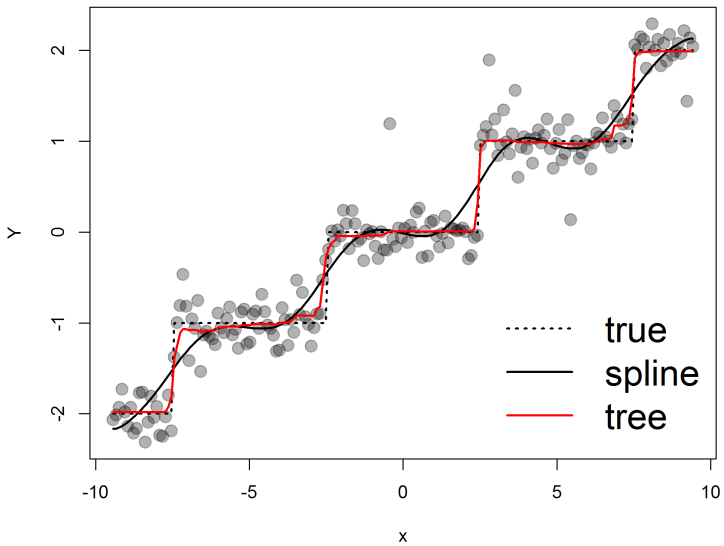
Основные базовые модели

- Линейная модель (Friedman, 2000) $h(\boldsymbol{x}, \theta) = \theta_{j1}\boldsymbol{x}_j + \theta_{j2}$.
- Дерево регрессии (Friedman, 1999) $h_t(\boldsymbol{x}) = \sum_{j=1}^{J} \theta_{jt} I(x \in R_{jt})$.
- P-сплайн (Tutz and Binder, 2006) λ — параметр регуляризации.

$$\hat{f}_j = \underset{f}{\operatorname{argmin}} \sum_{i=1}^{N} (U_i - f(x_{ij}))^2 + \lambda \int (f''(x))^2 dx.$$

Примеры неоптимального выбора типа модели:





Смешанные схемы

Best model choise — *на каждой итерации добавляем модель с наименьшим MSE на out-of-bag множестве*

Increasing complexity — *последовательно добавляем линейные модели (h_1), сплайны (h_2), деревья глубины 1 (h_3), деревья глубины J (h_4).*

Мера "полезности" модели h на итерации t:

$$\Delta MSE_t(h(\cdot)) = \frac{MSE_{t-1} - \frac{1}{N}\sum(\hat{f}_t + h(\theta, x_t) - Y)^2}{MSE_{t-1}}.$$

Алгоритм выбора типа модели

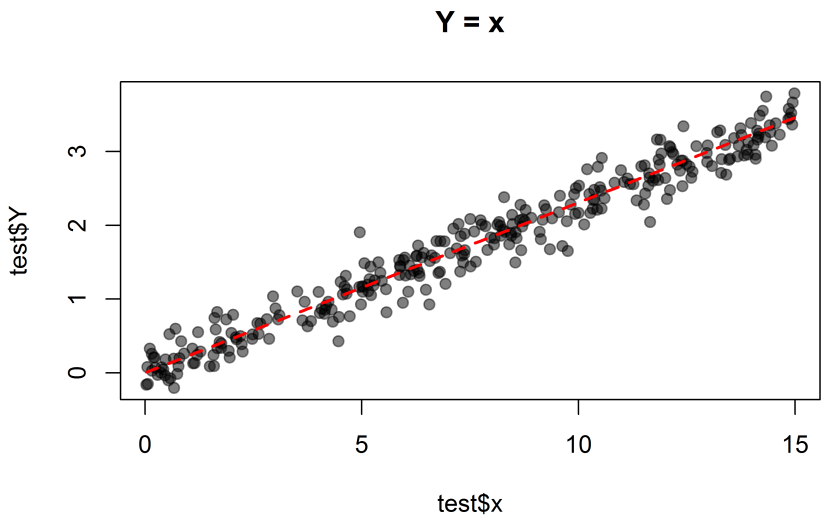
Шаг 1. Добавляем h_1 если $\frac{1}{N}\sum_{i=t-N+1}^{t}(\Delta MSE_t(h_1(\cdot)) > \varepsilon$.

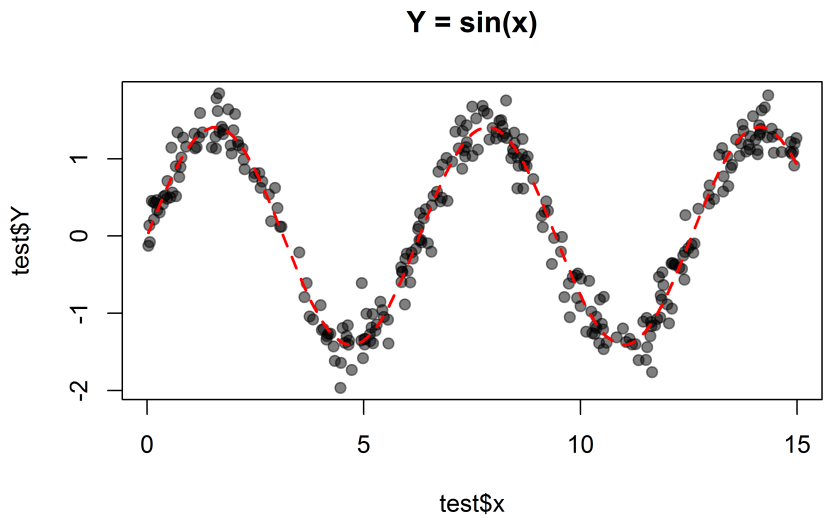
Шаг 2. Добавляем h_2 с вероятностью $\frac{\Delta MSE_t(h_2)}{\Delta MSE_t(h_1)+\Delta MSE_t(h_2)}$, иначе добавляем h_1.

Если число добавлений h_2 за N итераций $> \varepsilon_2 \Rightarrow h_i = h_{i+1}$

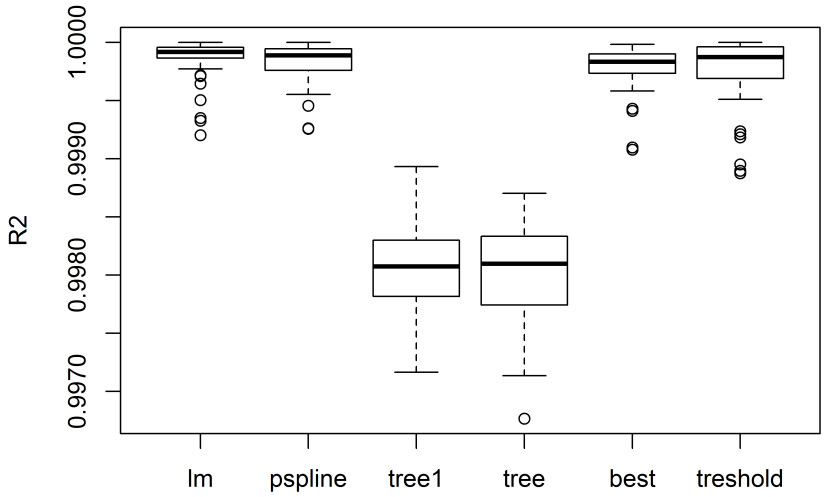
Результаты сравнения на одномерных модельных данных

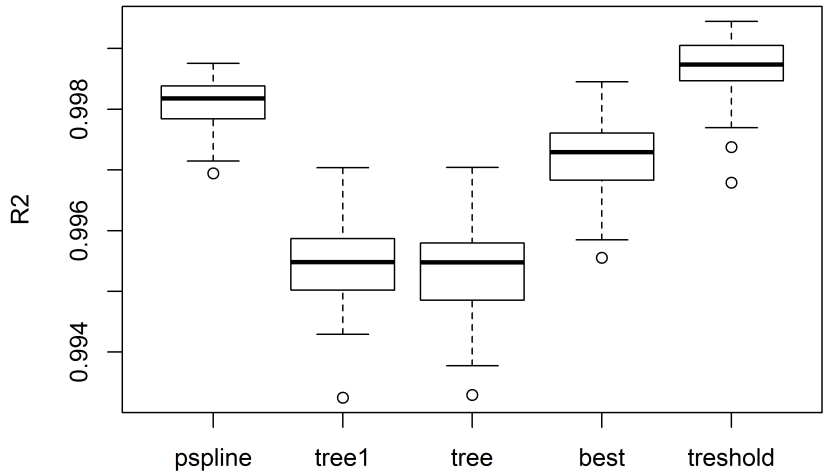
Пример сгенерированных выборок в случае гладких функций:





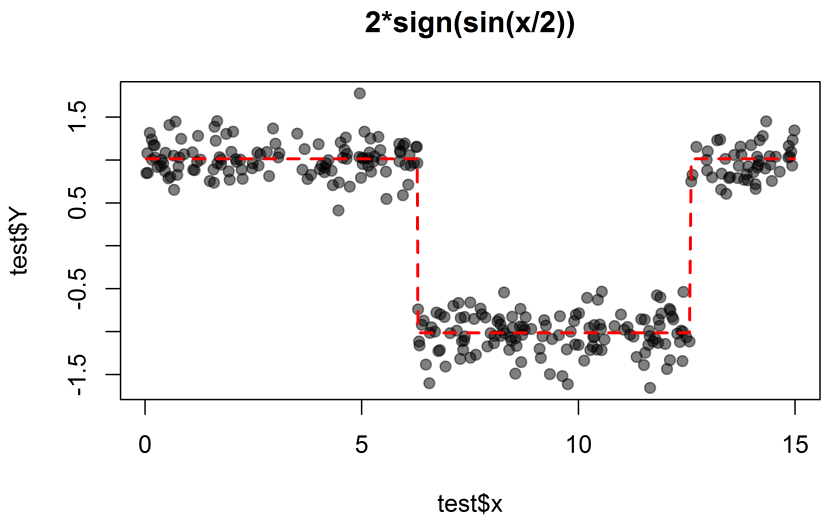
Полученные результаты $R2$ на 100 сгенерированных выборках:

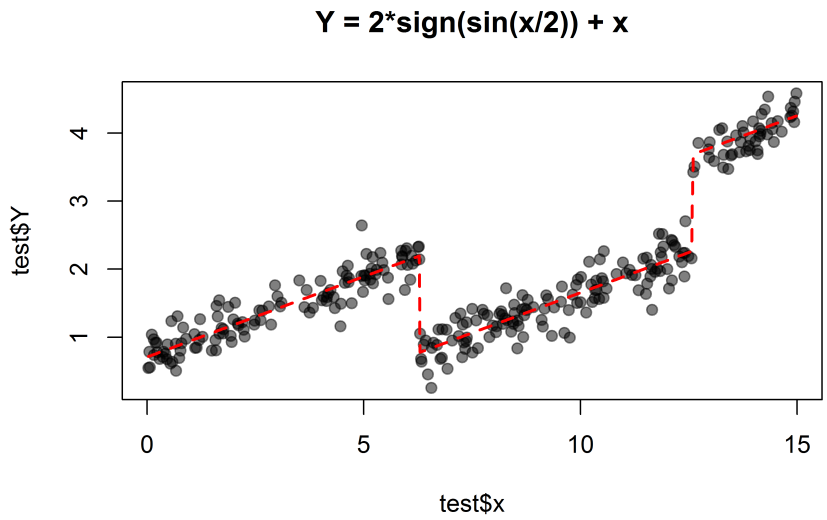




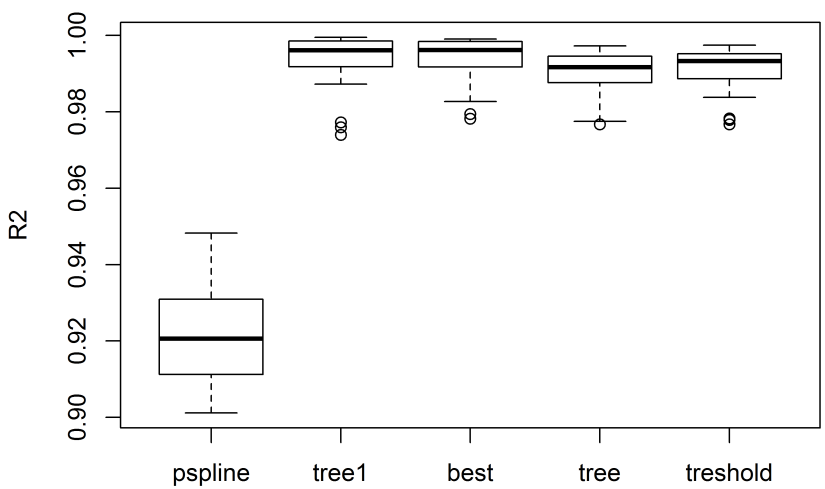
Результаты сравнения на одномерных модельных данных

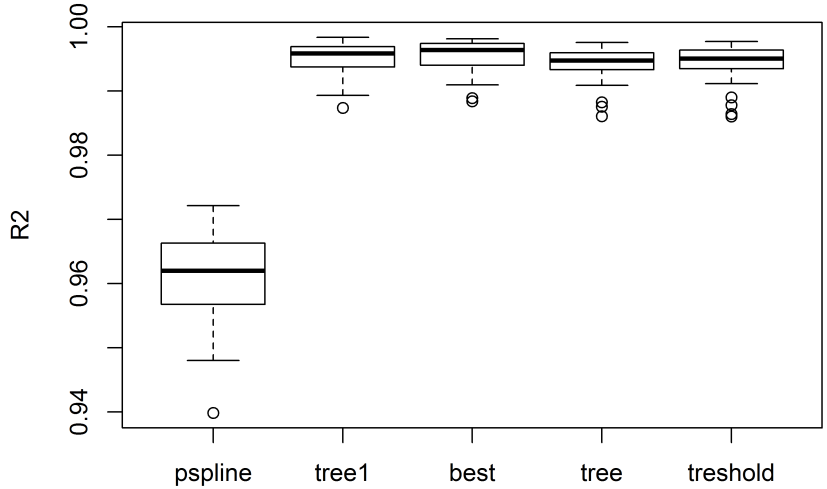
Пример сгенерированных выборок в случае функций с разрывами:





Полученные результаты $R2$ на 100 сгенерированных выборках:





Модельные данные многомерный случай

$$x_i \sim U[0,1], i = 1 \dots 4 \quad x_i \in \{1,2,3,4\}, p = 1/4, i = 5,6,7.$$

Аддитивная зависимость

$$f(x) = 2x_1^2 - 3x_2 - 4x_3 + 5sin(2\pi x_3) + 5I\{(x_6 = 3) \cup (x_6 = 4)\}.$$

Смешанная зависимость

$$f(x) = 2x_1^2 - 3x_2 - 4x_3 + 5sin(2\pi x_3) + 5I\{(x_6 >= 3) \cap (x_7 >= 4)\}.$$

Неаддитивная зависимость

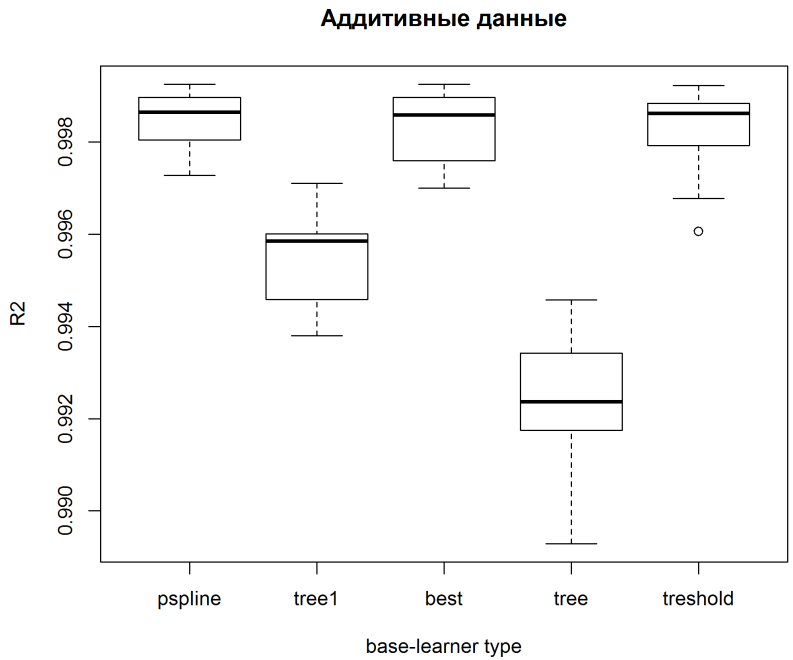
$$f(x) = 2x_1^2 x_2 + I\{x_5 \leq 2\}x_3 sin(2\pi x_3) + 3I\{(x_6 >= 3) \cap (x_7 >= 4)\}.$$

Модельные данные многомерный случай

Рассматриваемые схемы бустинга

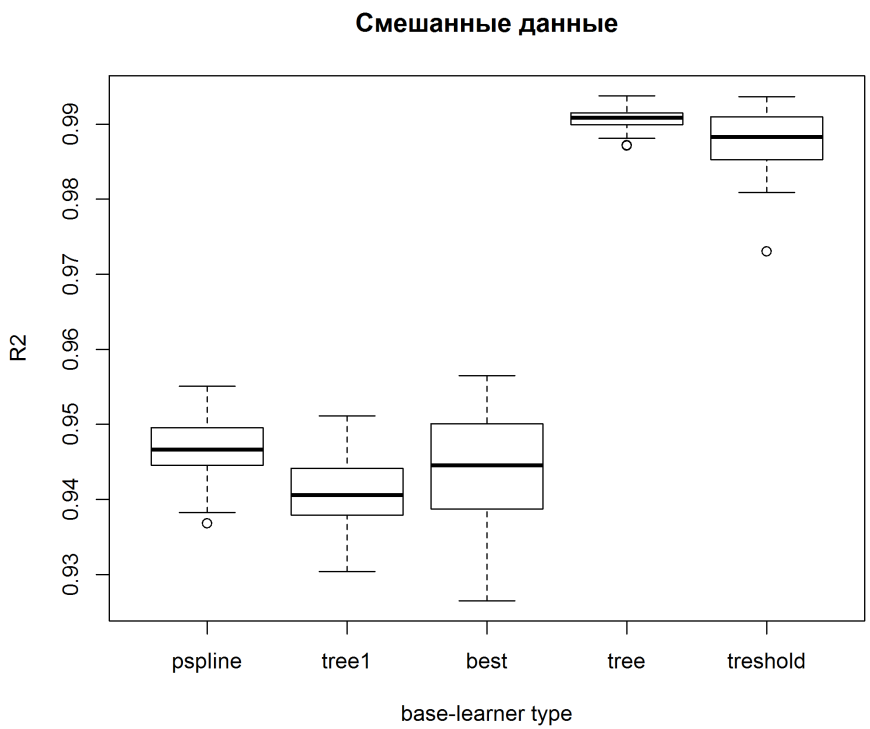
- $pspline$ — Бустинг сплайнами.
- $tree1$ — Бустинг деревьями с глубиной дерева $= 1$.
- $tree$ — Бустинг деревьями глубиной $= 4$.
- $best$ — Аддитивный бустинг лучшей одномерной моделью.
- $treshold$ — Схема с последовательным усложнением.

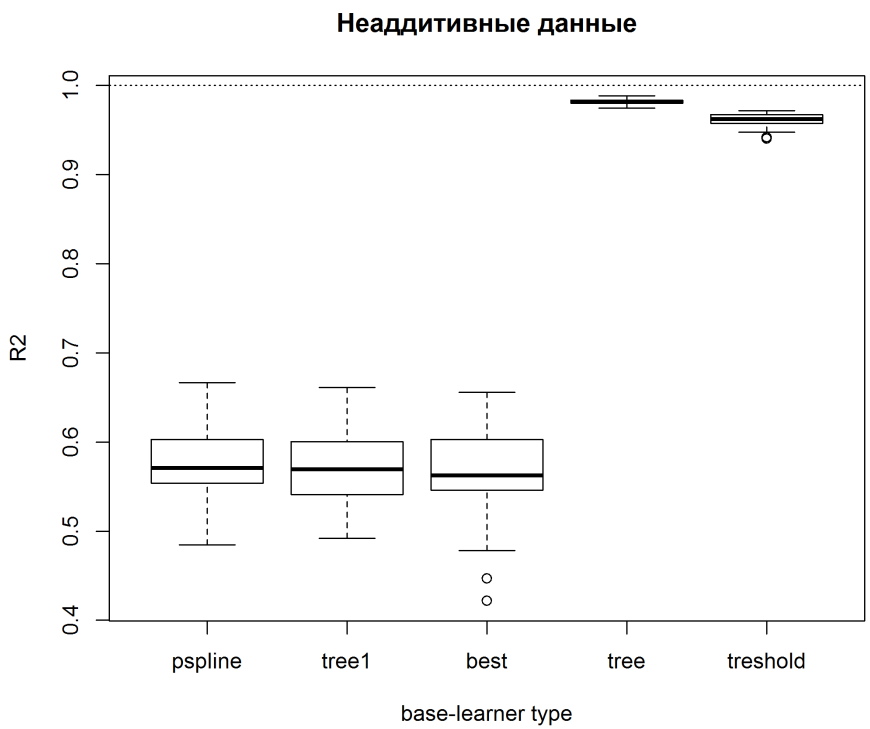
Результат в случае отсутствия эффектов взаимодействия



Модельные данные многомерный случай

Результаты $R2$ на 100 смоделированных выборках в случае наличия эффектов взаимодействия между переменными.
Устойчивость схемы с последовательным усложнением.





Модельные данные выводы

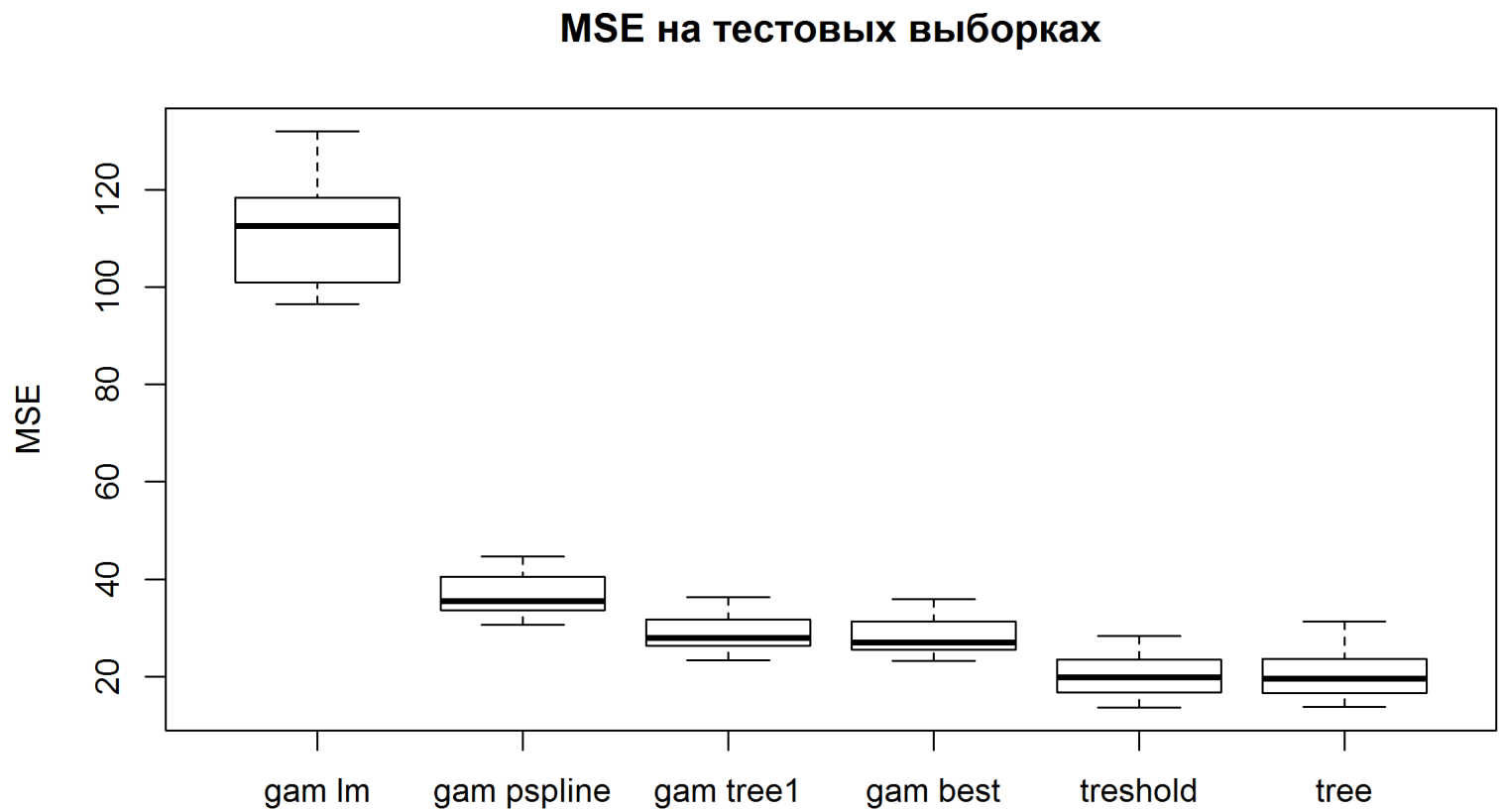
- Отсутствует оптимальная однотипная схема.
- В случае одномерных данных результаты предложенных схем близки и к оптимальным.
- В многомерном случае на аддитивных данных результаты обеих схем близки к оптимальным.
- Добавление эффектов взаимодействия может ухудшать точность прогноза по сравнению с бустингом деревьями решений.

Точность прогноза на реальных данных

Пример использования GBM для прогноза прочности бетона.

http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength
$Y \in R$, 8 зависимых переменных, 1030 индивидов.

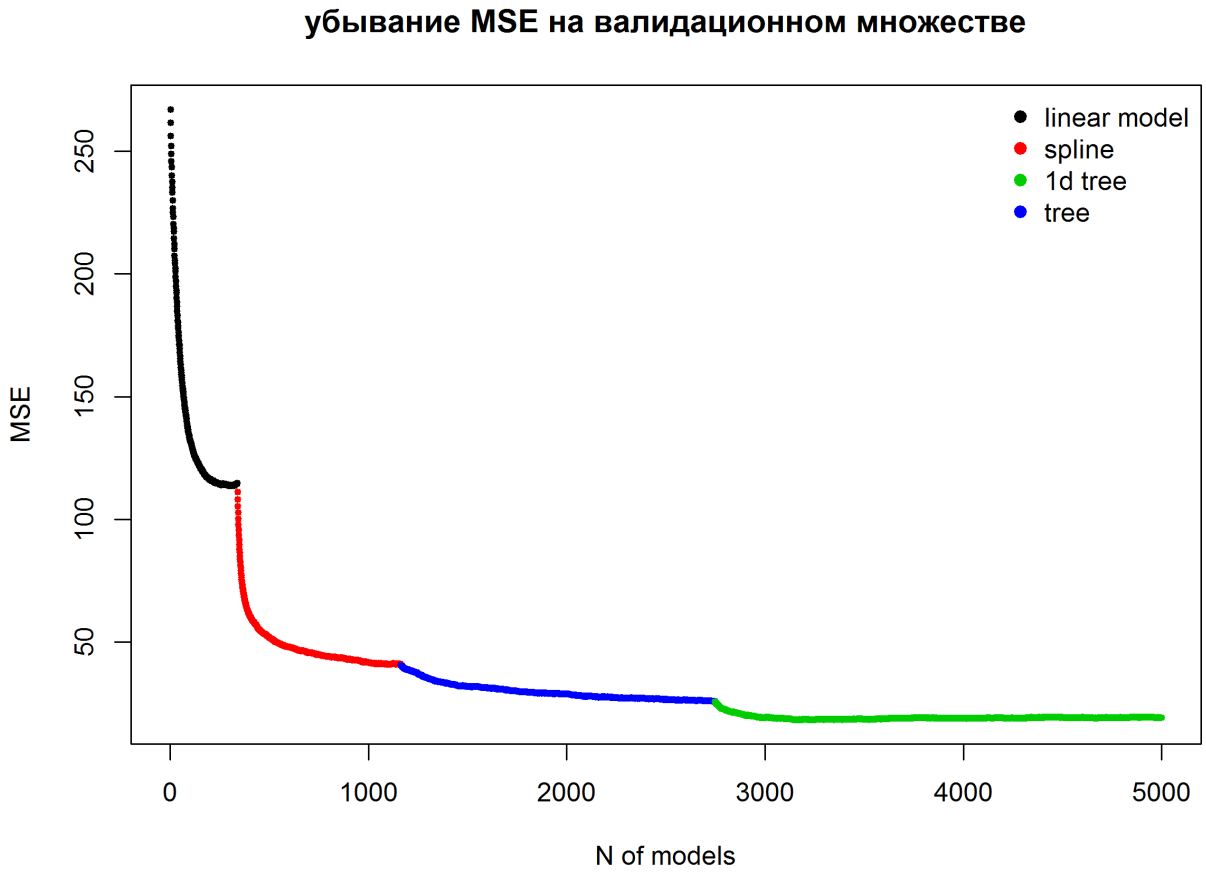
Оценка MSE по 100 бутстрап-выборкам



Оценка характера зависимости данных

Анализ убывания MSE по типам модели позволяет дать наглядную оценку:

- степени нелинейности данных
- наличии/отсутствии эффектов взаимодействия
- обоснованность применения сложных моделей



Вывод

Схема GBM с последовательным усложнением

Преимущества

- Более универсальная схема.
- Точность прогноза близка к оптимальному значению.
- Позволяет оценить характер зависимости данных.

Недостатки

- Ухудшение точности на неаддитивных данных.

Результаты

- Рассмотрены достоинства и недостатки стандартных схем GBM.
- Предложено две схемы выбора типа модели при бустинге.
- Проведено сравнение точности прогноза в зависимости от
  - типа зависимой функции (одномерный случай)
  - наличие отсутствия эффектов взаимодействия (многомерный случай)
- Предложена схема оценки обоснованности применения GBM на деревьях на основе сходимости MSE по типу модели.
- Реализована своя имплементация метода GBM на языке R, позволяющая применять нестандартные методы выбора типа модели.

Спасибо за внимание!