ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УПРАВЛЕНИЯ

Основан в 1919 году

Методы и технологии машинного обучения

Лекция 5: Методы, основанные на деревьях решений

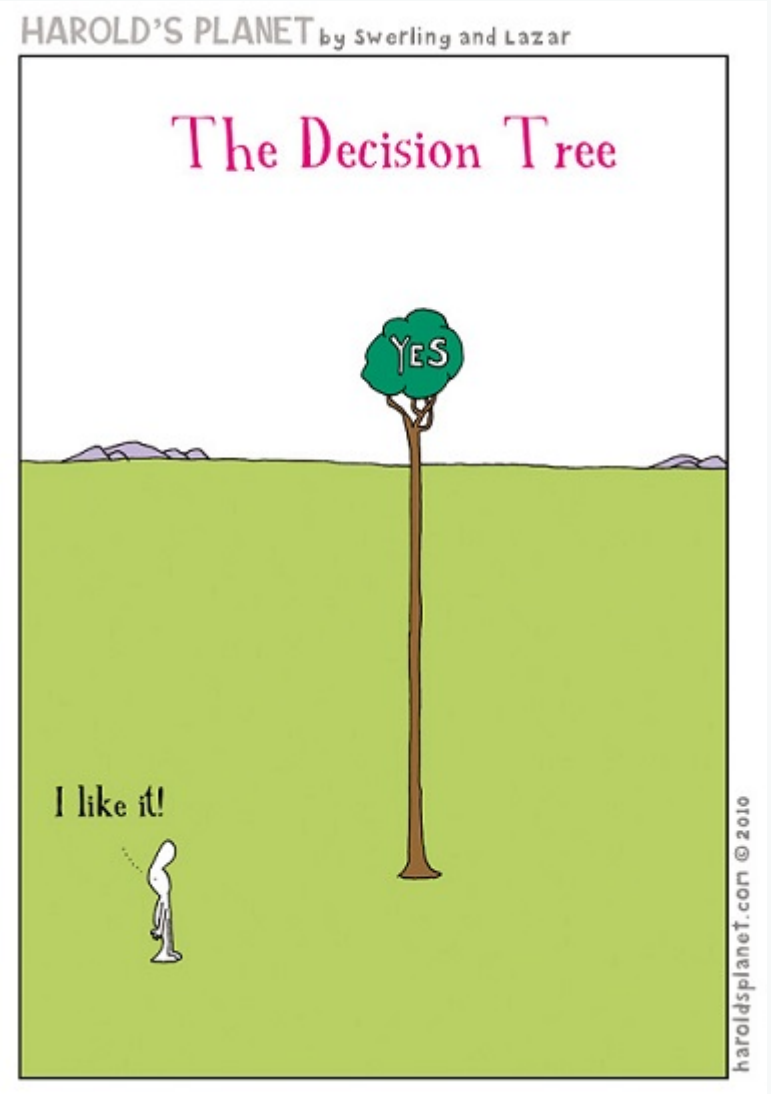
Светлана Андреевна Суязова (Аксюк)
sa_aksyuk@guu.ru

осенний семестр 2021 / 2022 учебного года

План лекции

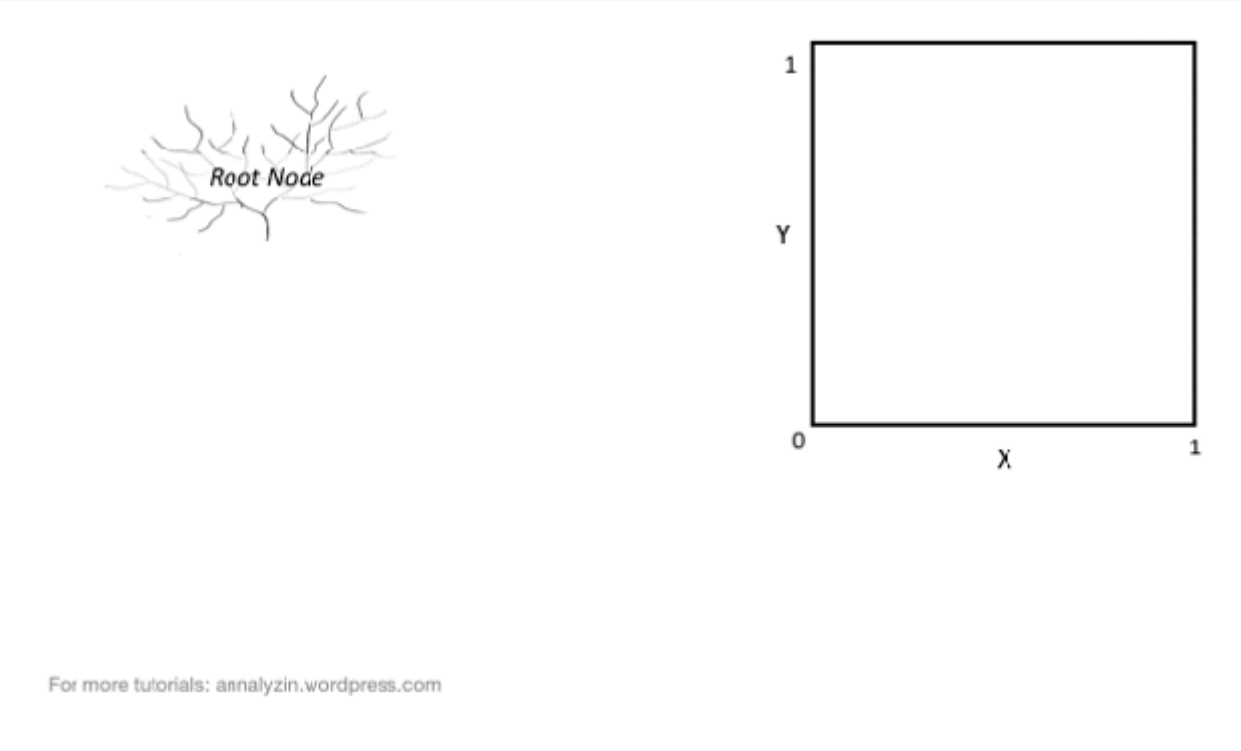
- Деревья решений

- Что такое бутстреп
- Бэггинг, случайный лес, бустинг



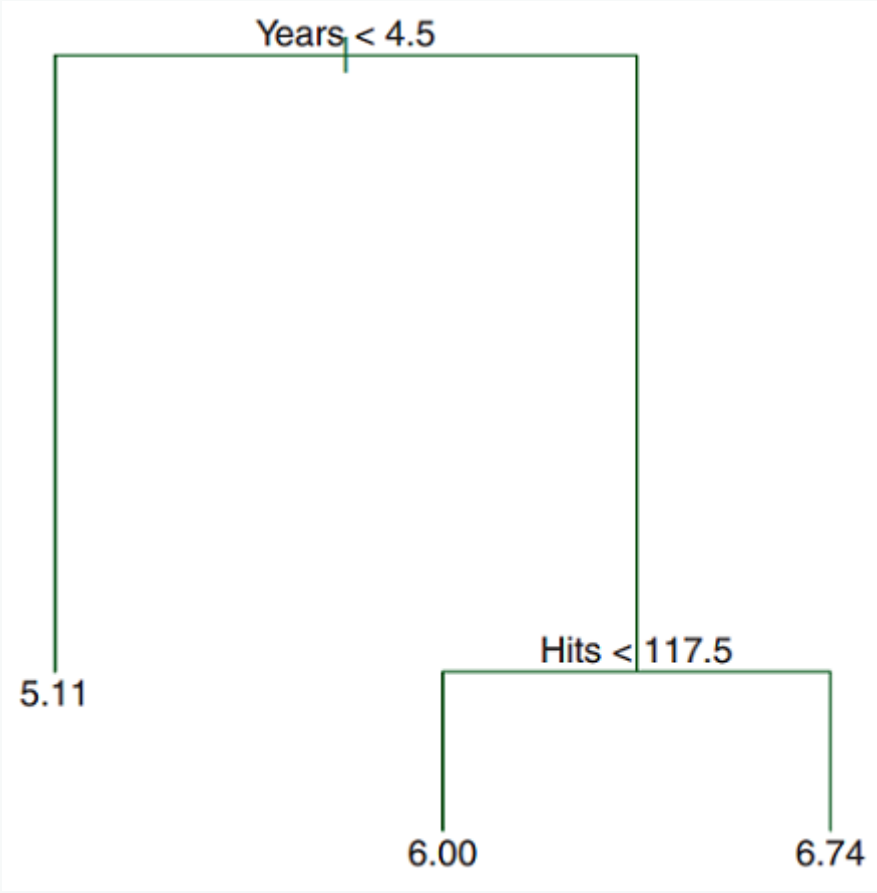
instagram.com/harolds.planet/

Пример на данных `Titanic`.



- Дерево сегментирует пространство X-ов на несколько ограниченных областей
- Деревья применяются в задачах регрессии и классификации
- Предсказание делают по оценке среднего (среднее, медиана, мода) для сегмента, в котором оказалось наблюдение
- Проверка гипотез о характере взаимосвязей между откликом и объясняющими переменными невозможна

Дерево решений в задаче регрессии



Пример на данных `Hitters`, отклик: `log(Salary)`.

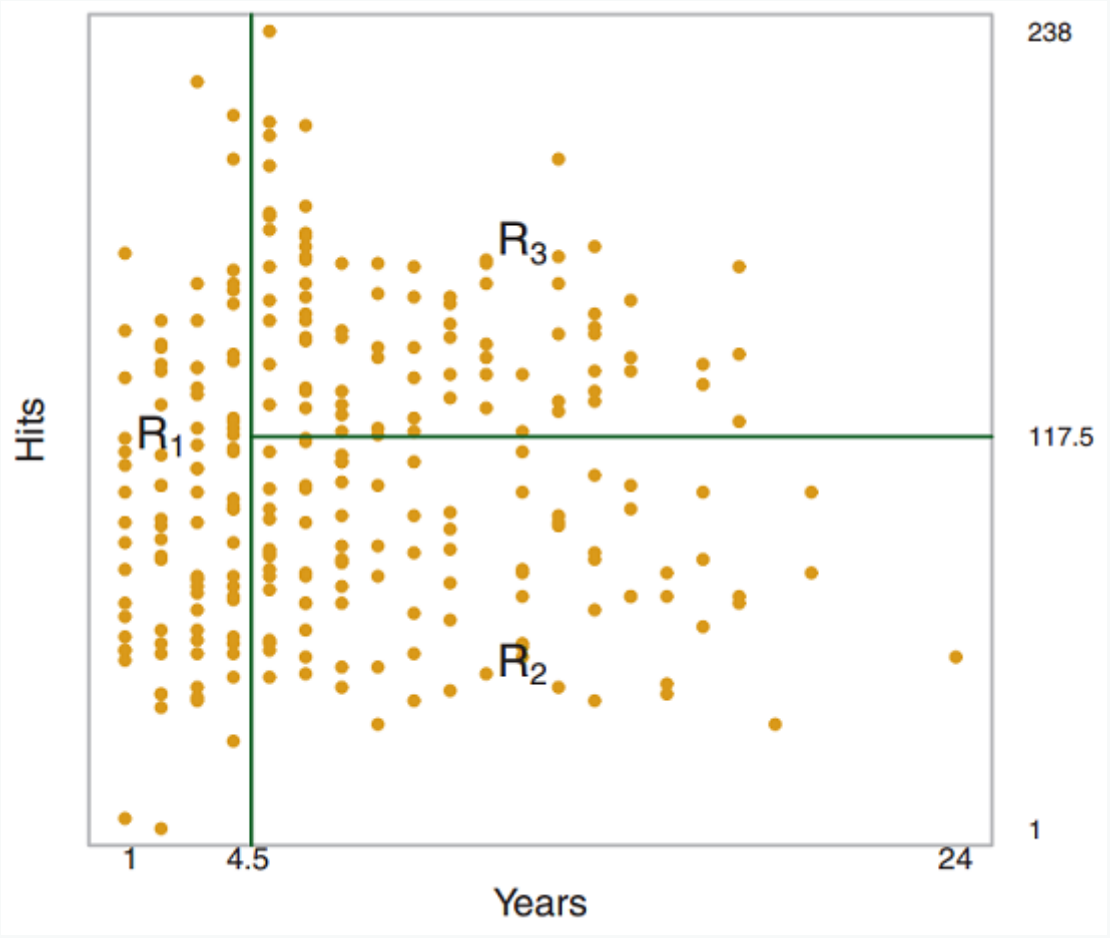
Дерево решений в задаче регрессии

$$
\begin{aligned}
R_1 =& \{X | Years < 4.5\} \\
R_2 =& \{X | Years > 4.5, \\
& \quad Hits < 117.5\} \\
R_3 =& \{X | Years > 4.5, \\
& \quad Hits > 117.5\}
\end{aligned}
$$

Прогнозы:

$$
\begin{aligned}
\hat{y}_{R_1} &= 1000 \cdot \exp^{5.11} \approx 165670\$ \\
\hat{y}_{R_2} &= 1000 \cdot \exp^{6} \approx 403429\$ \\
\hat{y}_{R_3} &= 1000 \cdot \exp^{6.74} \approx 845561\$
\end{aligned}
$$

Конечные узлы (листья):
R_1, R_2, R_3



Интерпретация

```
|--- Years < 4.50
|   |--- log_Salary: 5.11
|--- Years >= 4.50
|   |--- Hits < 117.5
|   |   |--- log_Salary: 6.00
|   |--- Hits >= 117.5
|   |   |--- log_Salary: 6.74
```

- `Years` – главный фактор: `Years` `↑↑` `Salary`
- Для игроков с высоким `Years`. `Hits` ↑↑ `Salary`

Процедура построения

(1) Разбить пространство предикторов на J отдельных непересекающихся многомерных прямоугольников (*контейнеров*) $R_1, R_2, \ldots, R_J$, которые минимизируют RSS:

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2 \to \min$$

(2) Для всех наблюдений, попадающих в область R_J, сделать одинаковое предсказание по среднему отклику у обучающих наблюдений в этой области.

Нисходящий жадный алгоритм
(или *рекурсивное бинарное разбиение*)

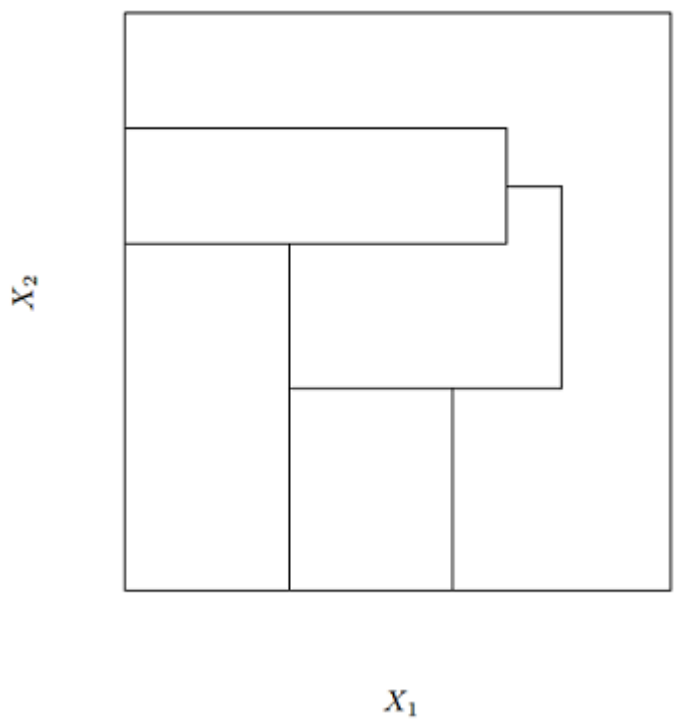
- **нисходящий**: начинается с корневого узла дерева (все наблюдения в одной области)
- **жадный**: на каждом этапе выполняется разбиение, оптимальное для этого этапа, без заглядывания вперёд

Пусть j – номер предиктора, s – точка разрыва:

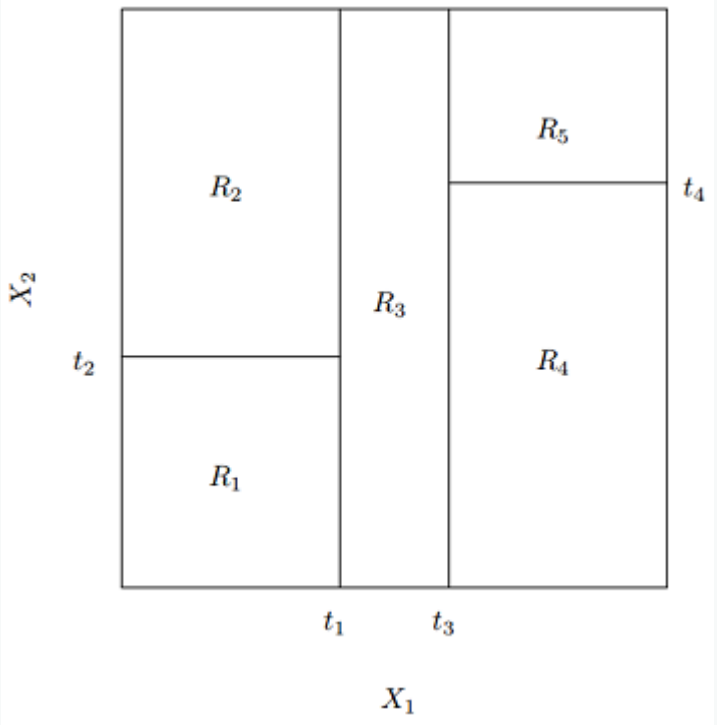
$$\forall j, s: \quad R_1(j,s) = \{X | X_j < s\}, \quad R_2(j,s) = \{X | X_j \geq s\}$$

$$\sum_{i: x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i: x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2 \to \min$$

Пример разбиения

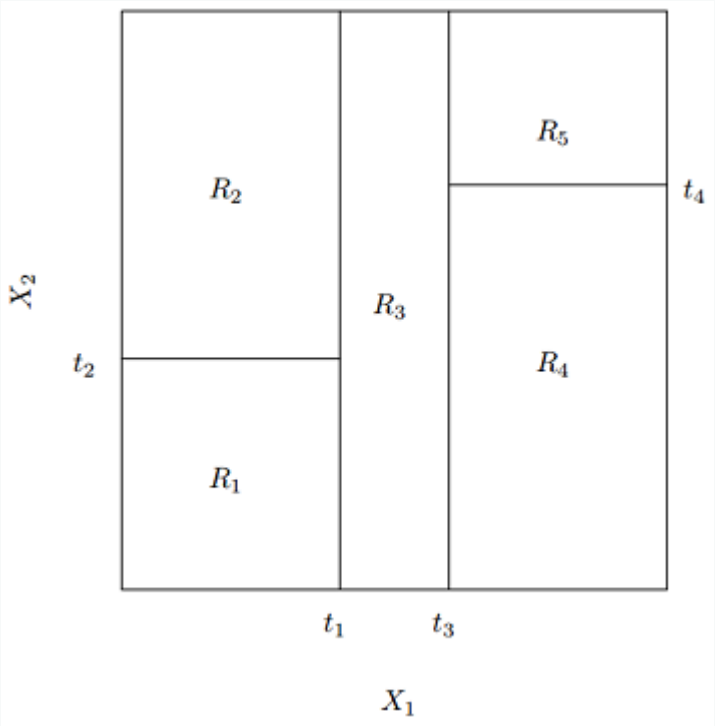


так нисходящим жадным алгоритмом разбить нельзя...

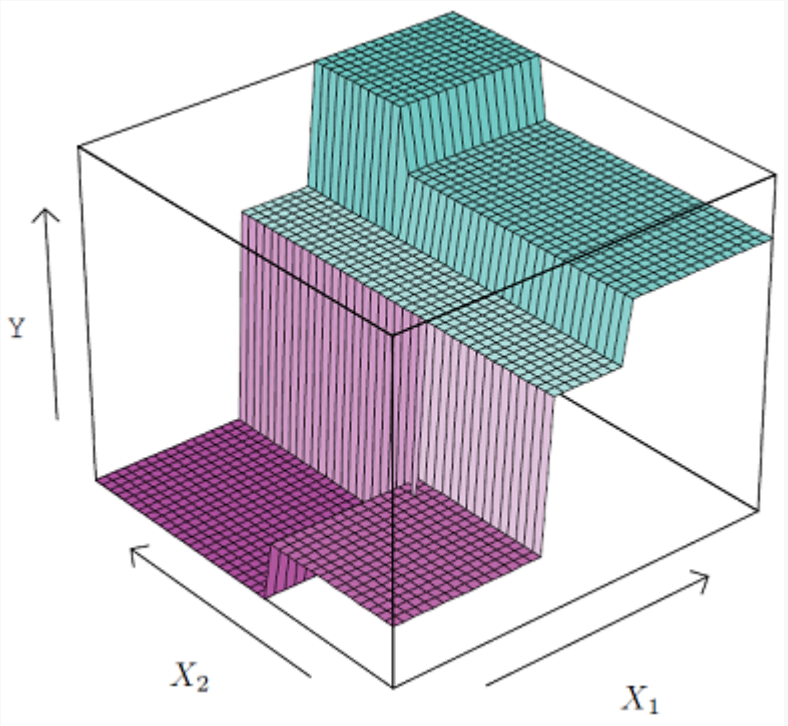


...а так можно

Пример разбиения (задача регрессии)



пример рекурсивного бинарного разбиения



поверхность прогноза

Деревья и линейные модели

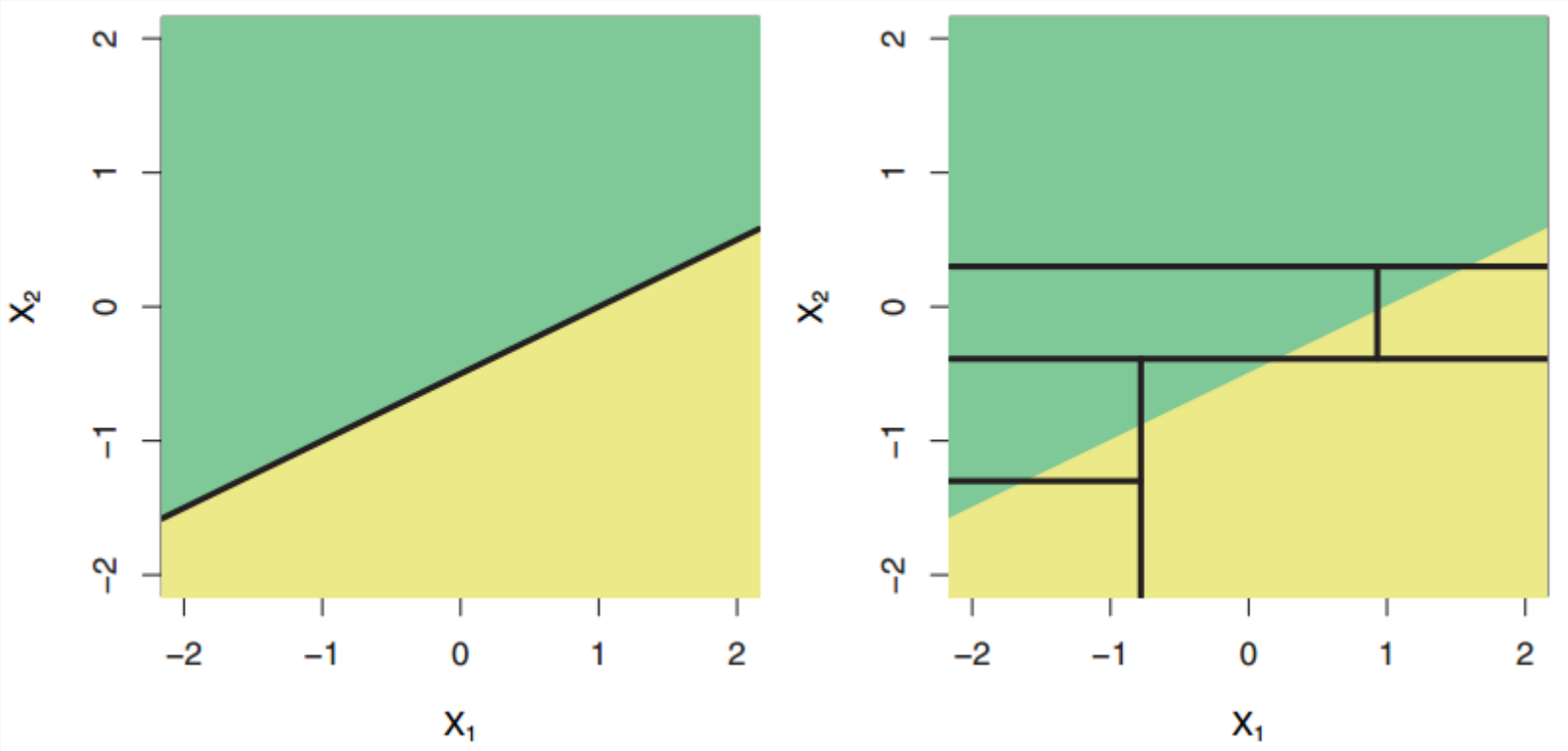
Модель линейной регрессии:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

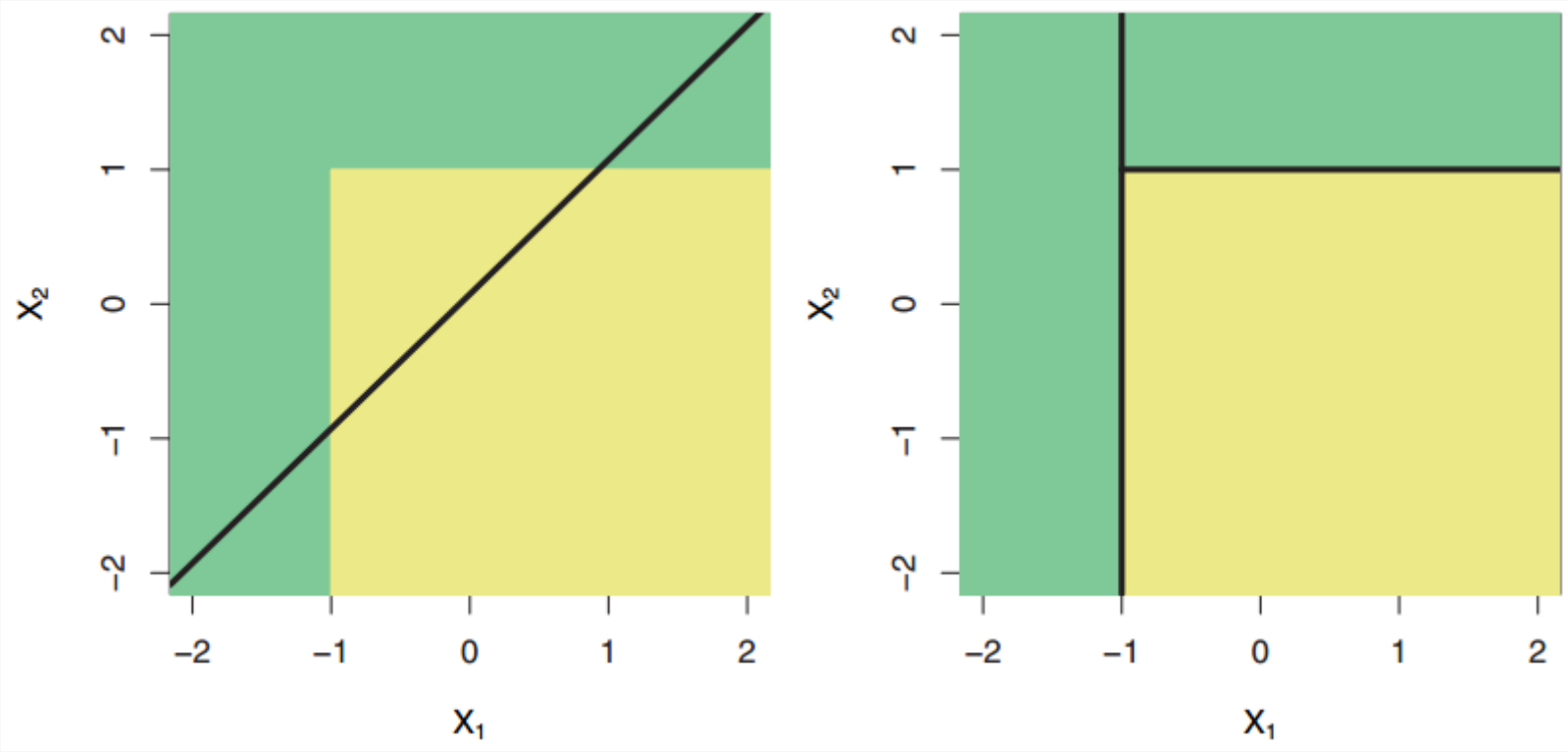
Модель регрессионного дерева:

$$f(X) = \sum_{m=1}^{M} c_m \cdot 1_{(X \in R_m)}$$

где $R_1, \ldots, R_M$ – непересекающиеся области пространства предикторов.



Истинная классифицирующая функция линейна, модель регрессии (слева) работает лучше дерева (справа).



Истинная классифицирующая функция нелинейна, регрессионное дерево (справа) лучше регрессии (слева).

Обрезка ветвей дерева

- устранить переобучение
- снизить число ветвей, чтобы повысить интерпретируемость дерева
- с учётом штрафа на сложность:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha|T| \to \min$$

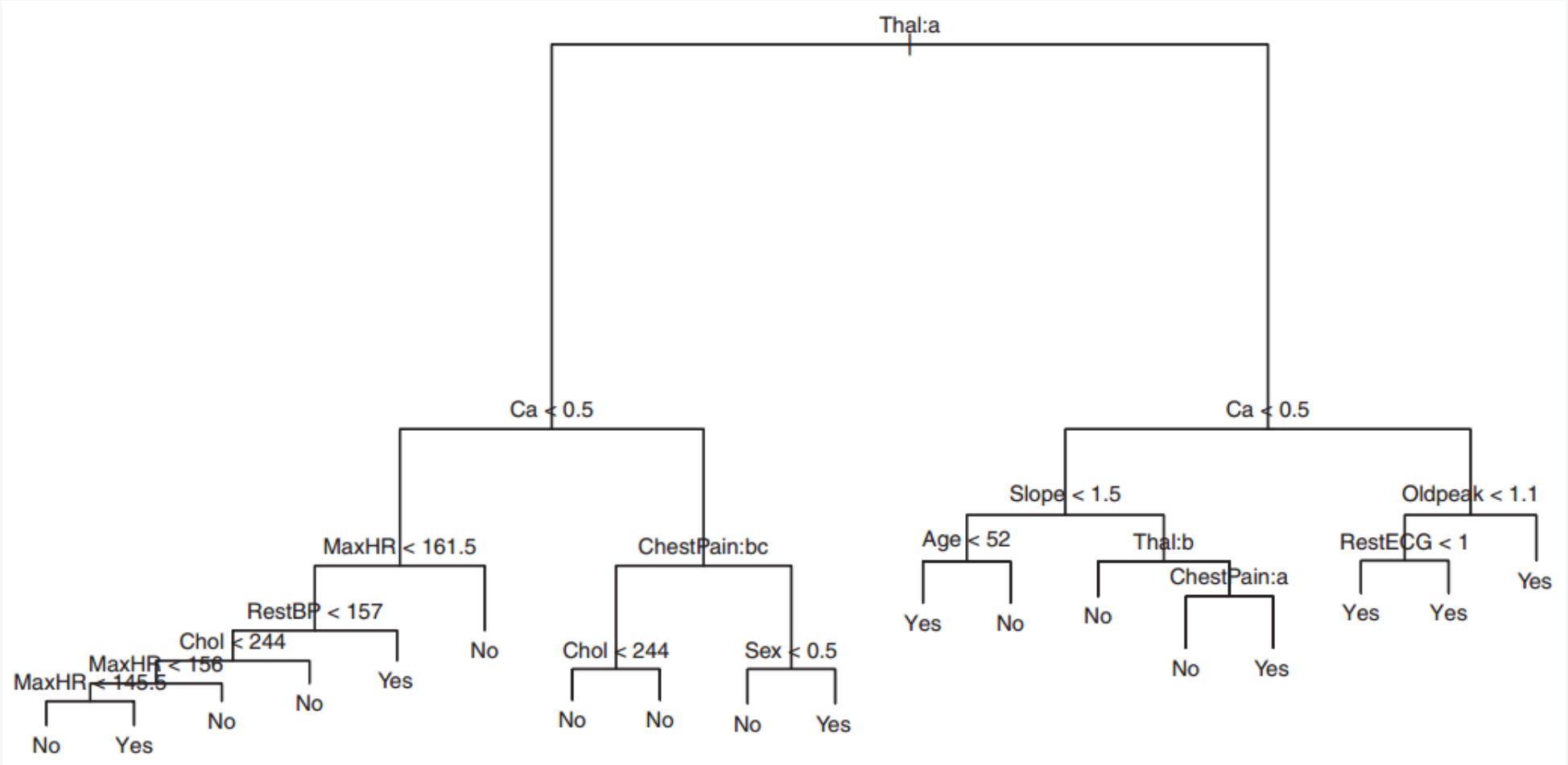
где $|T|$ – число конечных узлов дерева T, R_m – контейнер, соответствующий m-му конечному узлу, $\hat{y}_{R_m}$ – предсказанный отклик в области R_m, α – гиперпараметр

Деревья классификации

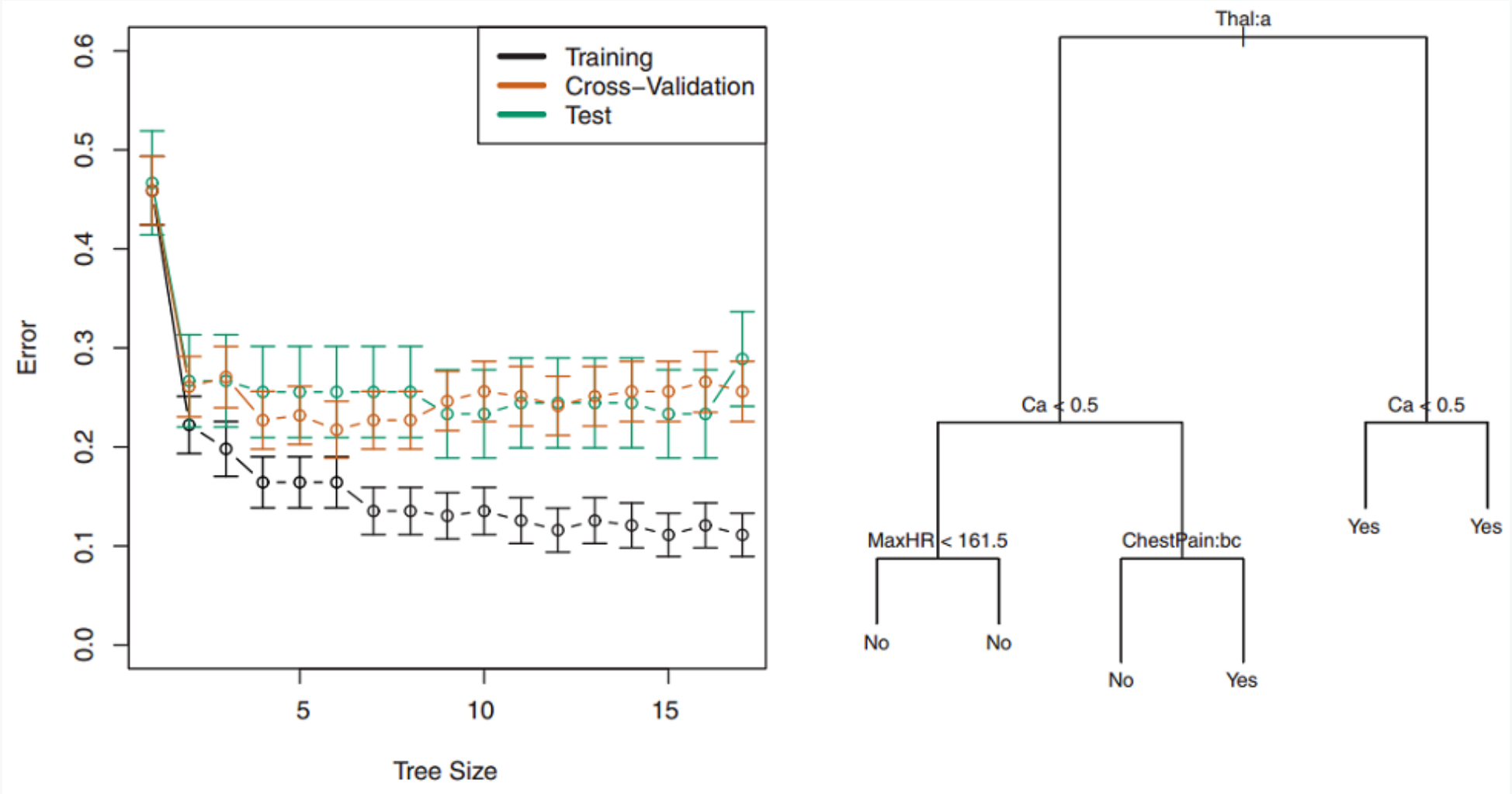
Прогноз по наиболее часто встречающемуся классу.
Оценки точности:

- частота ошибок классификации $E = 1 - \max\limits_k(\hat{p}_{mk})$
- индекс Джинни $G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
- коэффициент перекрёстной энтропии
  $D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$

где $\hat{p}_{mk}$ – доля обучающих наблюдений в m-ой области, принадлежащих классу k.
При низких значениях G, D частоты $\hat{p}_{mk}$ близки к 0 и 1, т.е. узлы "чистые".



Данные `Heart`, необрезанное дерево. Узлы с одинаковым прогнозом различаются частотой верно классифицированных наблюдений.



Данные `Heart`. *Слева:* ошибки в зависимости от числа узлов у обрезанного дерева. *Справа:* дерево с наименьшей ошибкой перекрёстной проверки.

Алгоритмы построения деревьев решений

- ID3 – рекурсивное бинарное разбиение, максимизирует прирост информации (IG) и минимизирует энтропию (H)
- C4.5 – усовершенствованная версия алгоритма ID3
- CART (Classification and Regression Tree) – алгоритм построения деревьев классификации и регрессии, строит бинарные деревья, минимизируя индекс Джини (G)
- CHAID (Chi-square automatic interaction detection) – автоматическое рекурсивное бинарное разбиение на базе критерия Хи-квадрат

Преимущества деревьв решений

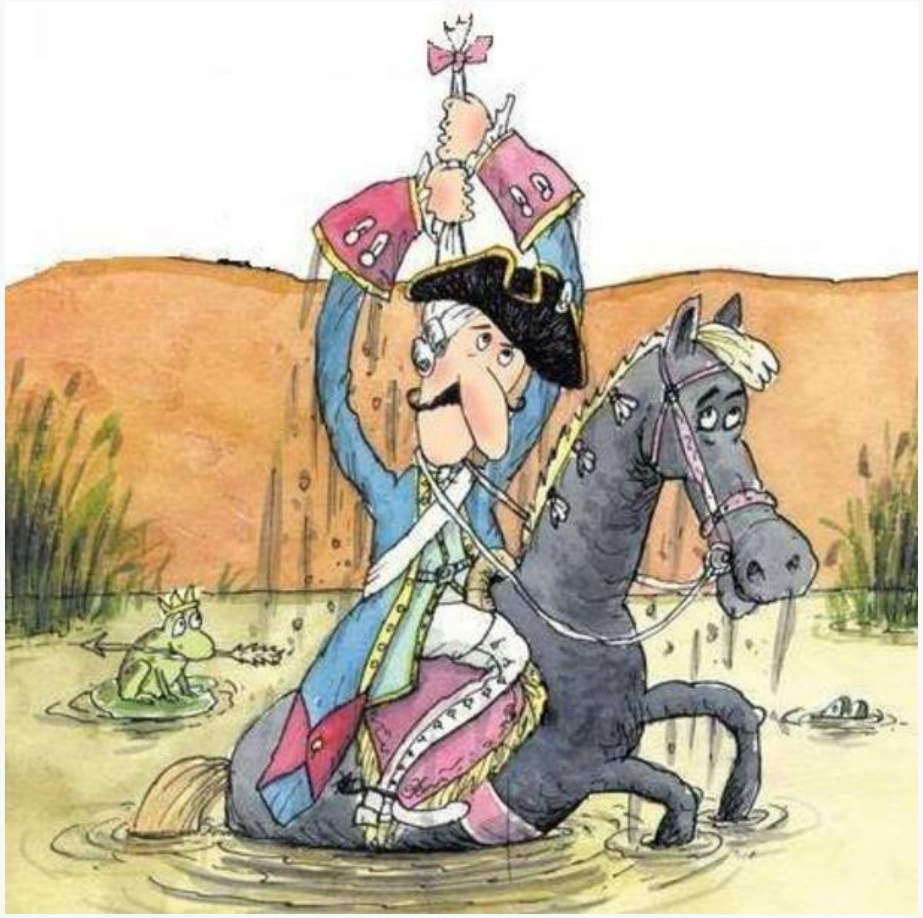
- Принцип работы модели интуитивно понятен
- Деревья решения близки к процессу принятия решений людьми
- Деревья можно представить графически при любой размерности пространства предикторов
- Легко справляются с качественными предикторами, специальные фиктивные переменные на базе категориальных не нужны

Недостатки

- Нестабильность: небольшие изменения входных данных могут сильно повлиять на модель
- Неточность: наилучшее бинарное разбиение в корне дерева не всегда ведёт к точному прогнозу

План лекции

- Деревья решений
- Что такое бутстрап
- Бэггинг, случайный лес, бустинг



Происхождение термина "Bootstrap":

- *1870* – петля на задней части мужского сапога, потянув за которую, можно надеть сапоги
- *1900* – фигура речи "вытягивать себя за петли от сапог" означает выполнение невыполнимого задания
- *1916* – расширение значения идиомы до "совершенствоваться в скрупулезной самостоятельной работе"
- *1953* – последовательность инструкций для загрузки операционной системы компьютера (программа "вытягивает саму себя")

www.thefreedictionary.com; www.etymonline.com

Использование термина "Бутстреп":

- *В программировании:* метод создания компилятора языка программирования, при котором значительная часть кода компилятора создаётся на целевом языке
- *В веб-разработке:* так называется инструмент веб-дизайна, фронт-энд среда разработки, распространяемая по свободной лицензии
- **В статистике:** метод определения статистик вероятностных распределений, основанный на многократной генерации псевдовыборок методом Монте-Карло на основе имеющейся выборки

Пример с инвестированием в два актива

Два финансовых актива обеспечивают доходность X и Y соответственно, X и Y – случайные величины. Долю инвестиций в X обозначим как α, тогда доля инвестиций в Y: $(1 - \alpha)$.

Целдь – минимизировать дисперсию доходности: $\mathrm{Var}(\alpha X + (1 - \alpha)Y) \to \min$

Минимум достигается при:

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

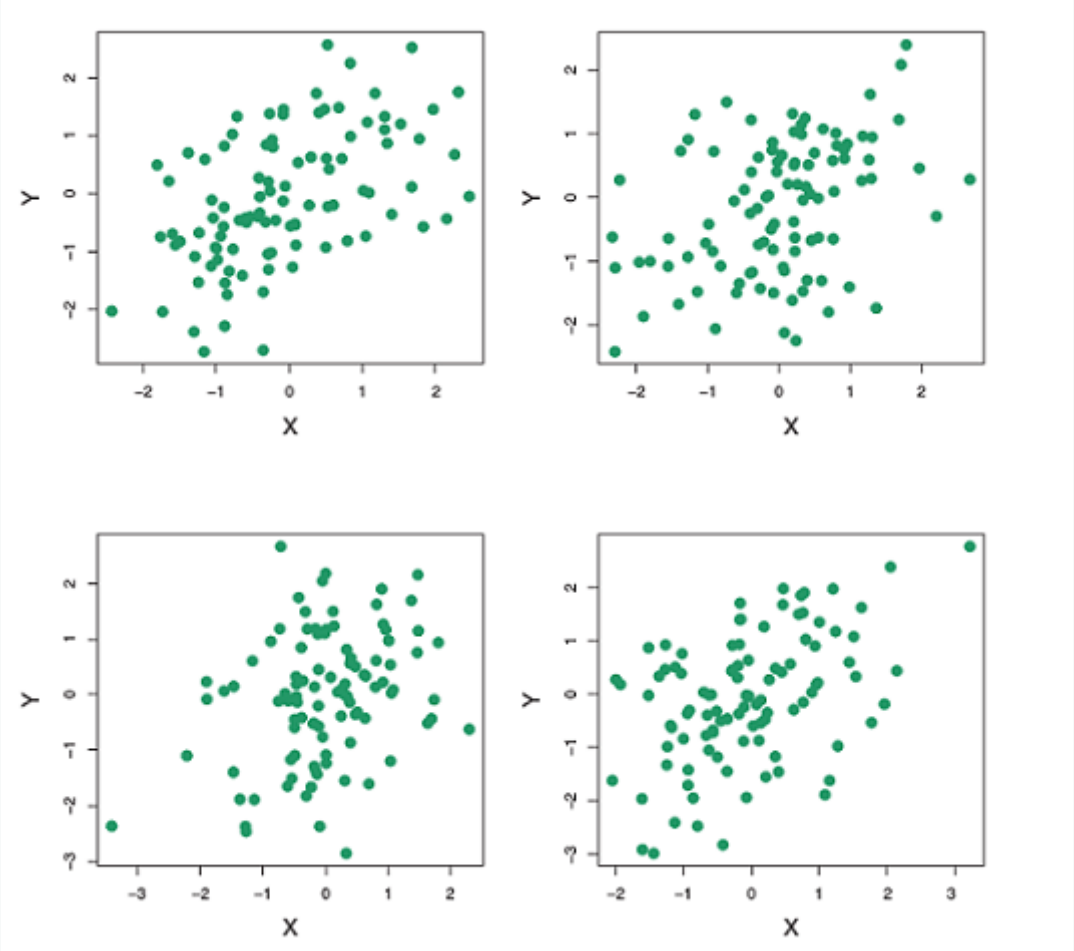
Истинные значения дисперсий и ковариации неизвестны. Мы можем вычислить их оценки: $\hat{\sigma}_X^2 = \hat{\mathrm{Var}}(X)$, $\hat{\sigma}_Y^2 = \hat{\mathrm{Var}}(Y)$, $\hat{\sigma}_{XY} = \hat{\mathrm{Cov}}(X, Y)$.

1. Если данных много, можно взять *много бесповторных выборок* и усреднить оценки
2. Если данных мало, для вычисления оценок можно сделать *много выборок с повторами* из имеющихся данных – **это есть бутстреп**

Пример с инвестированием в два актива

Четыре имитированных выборки X и Y. Истинное значение $\alpha = 0.6$.

Оценки α слева направо, сверху вниз: 0,576, 0,532, 0,657 и 0,651.



Пример с инвестированием в два актива

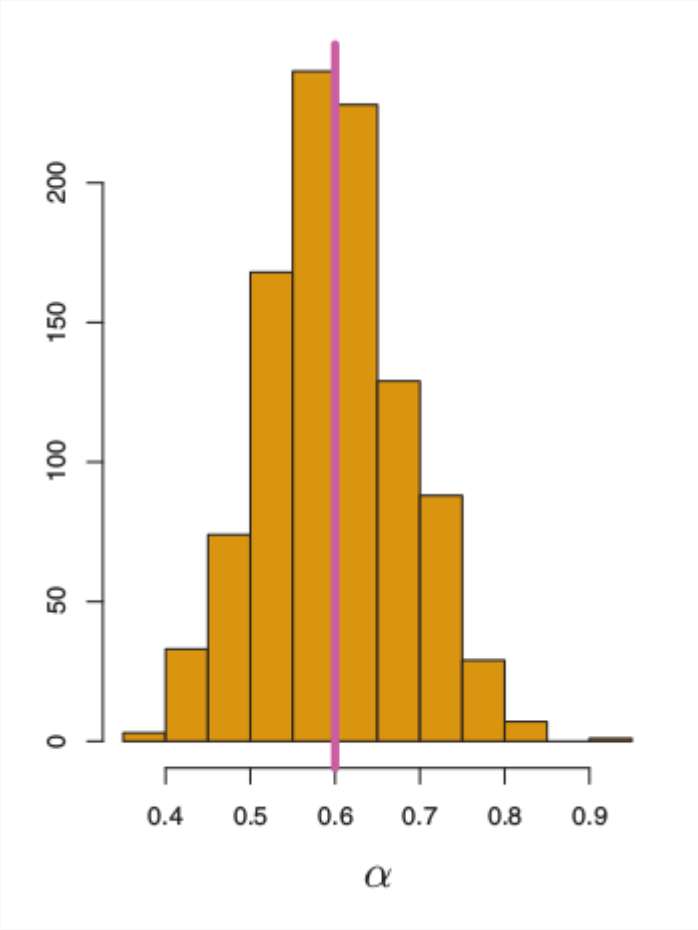
1000 имитированных выборок X и Y. Истинное значение $\alpha = 0.6$.

Средняя оценка α:

$$\bar{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996$$

Стандартная ошибка α:

$$\mathrm{SE}(\hat{\alpha}) = \sqrt{\frac{\sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2}{1000 - 1}} = 0.083$$

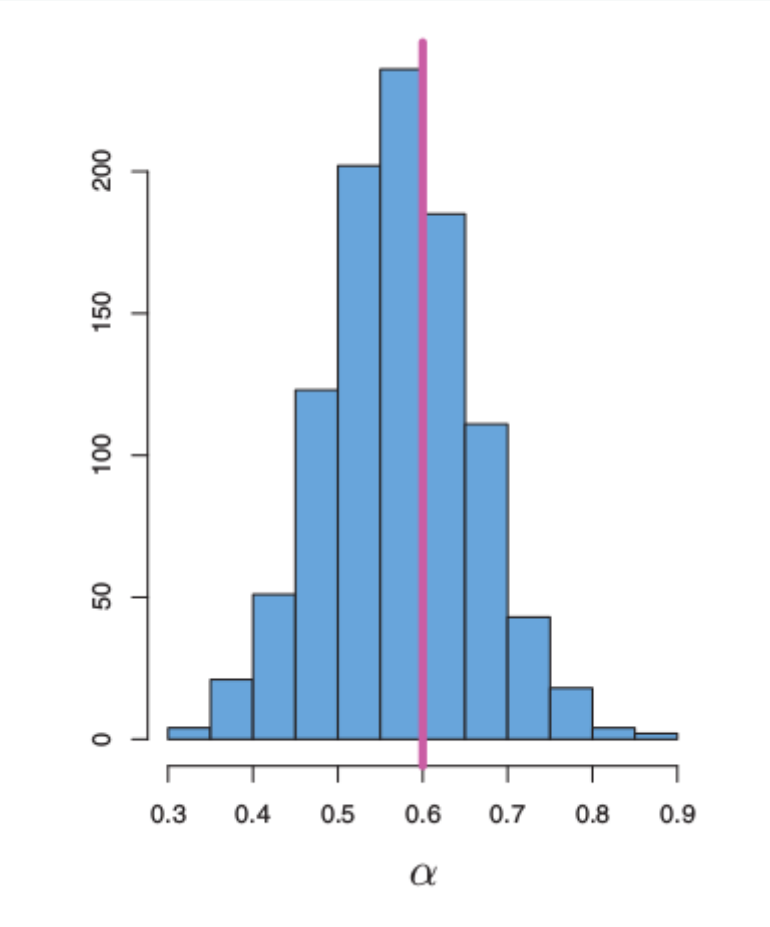


Пример с инвестированием в два актива

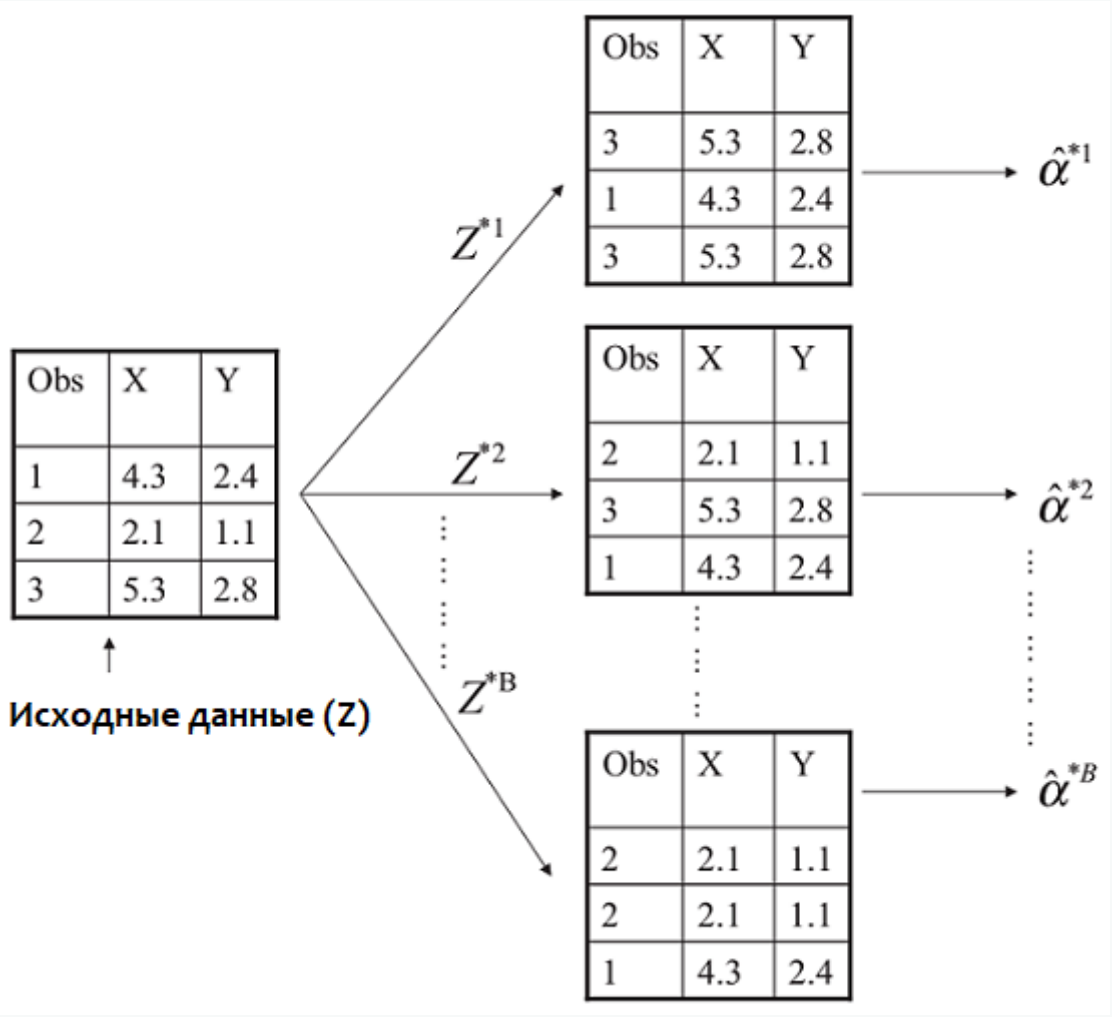
1000 бутстреп выборок из одного набора данных. Истинное значение $\alpha = 0.6$.

Стандартная ошибка α:

$$\mathrm{SE}_B(\hat{\alpha}) = 0.087$$



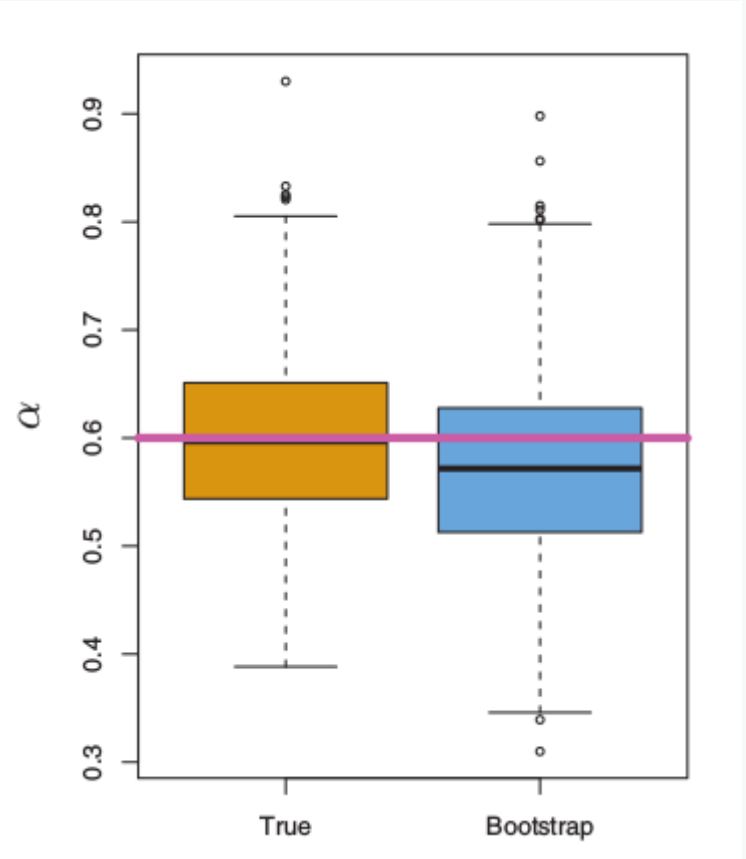
Что происходит с наблюдениями при бутстрепе



Пример с инвестированием в два актива

Множество оценок, сгенерированных на базе генеральной совокупности (*слева*), схоже со множеством оценок, полученных бутстрепом (*справа*).

Бутстреп-оценка может служить для нахождения вариабельности $\hat{\alpha}$



План лекции

- Деревья решений
- Что такое бутстреп

- Бэггинг, случайный лес, бустинг

Бэггинг

Идея: используя **бутстреп**, вырастить много деревьев и усреднить их предсказания.

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

- усреднение n оценок с дисперсией σ^2 даёт оценку с дисперсией σ^2/n
- деревья строятся глубокими и не обрезаются
- B подбираем по оценке ошибки вне выборки
- в задачах классификации вместо усреднения – решение по большинству голосов

Ошибка по оставшимся данным

В бэггинг-модели можно оценить ошибку вне выборки без перекрёстной проверки:

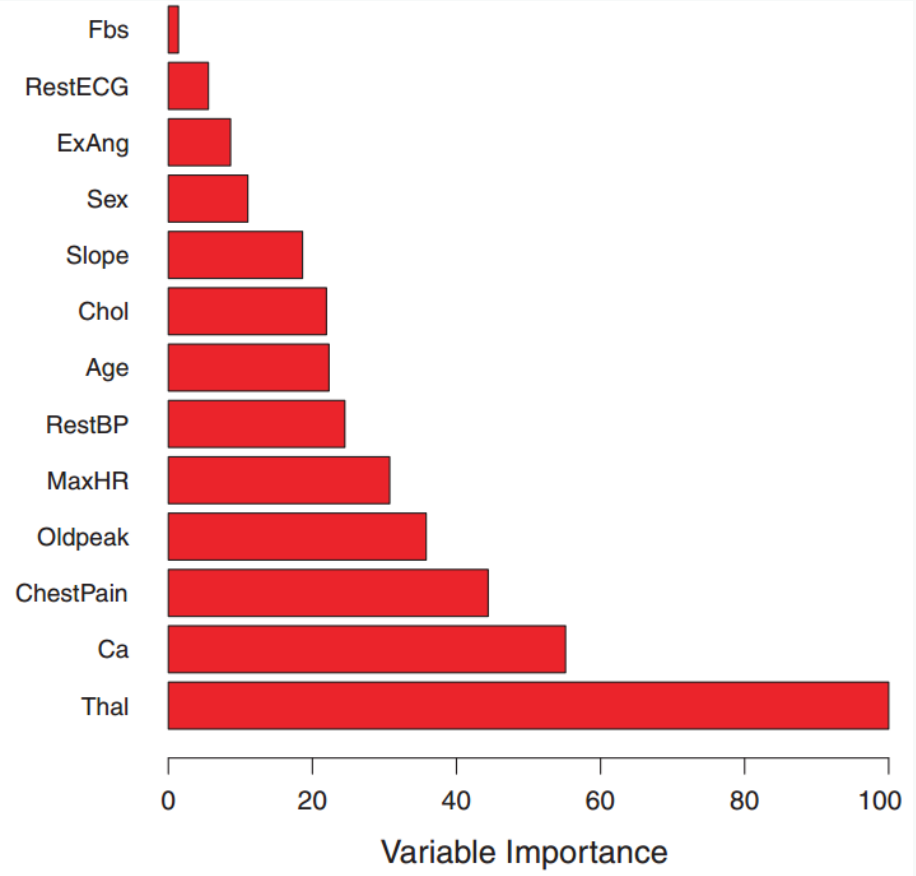
- В среднем каждое дерево "растёт" на $2/3$ наблюдений, $1/3$ – *оставшиеся наблюдения*
- Предсказываем отклик i-го наблюдения с помощью каждого дерева, для которого это наблюдение является оставшимся
- Считаем MSE (регрессия) или частоту ошибок (классификация)

Показатели важности переменных

Бэггинг повышает точность предсказаний, жертвуя интерпретируемостью.

- Считаем, на сколько уменьшается RSS (коэффициент Джинни) при разбиении по предиктору
- Усредняем оценку по всем деревьям
- Большое снижение оценки ошибки прогноза указывает на важный предиктор

На графике: важность предикторов из набора данных `Heart`*.*
По горизонтали: среднее снижение индекса Джинни для каждой переменной относительно макс. значения.



Случайный лес

Идея: усовершенствовать бэггинг, устранив корреляцию между деревьями

- B обучающих бутстреп-выборок
- Каждое дерево строится на m случайно отобранных предикторах из общего количества (p); $m \approx \sqrt{p}$

Метод позволяет снижать влияние очень сильных предикторов и давать шанс остальным.

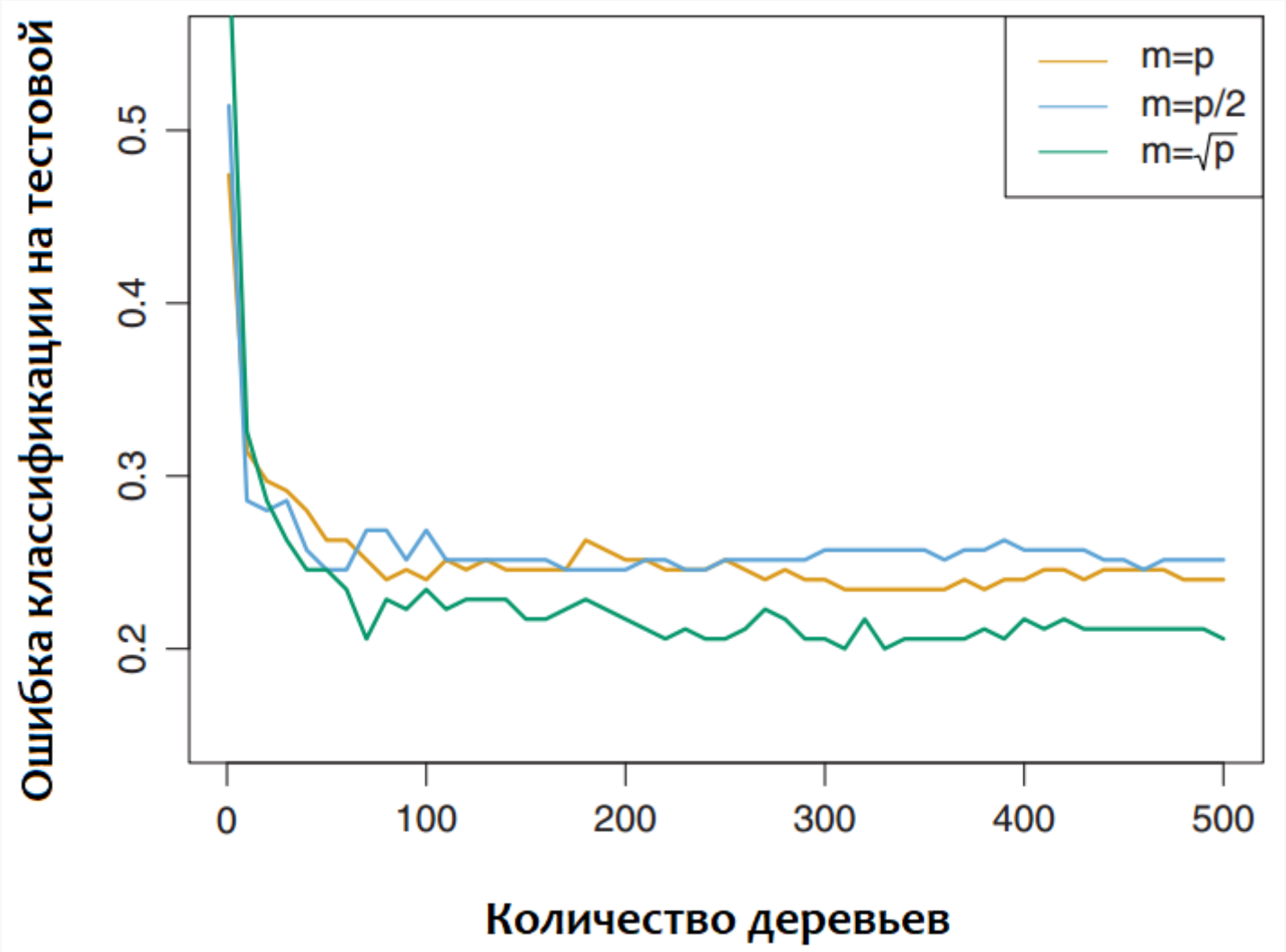
Если $m = p$, получим процедуру бэггинга.

Одно дерево vs бэггинг

Данные по экспрессии 20000 генов в образцах 349 пациентов.

Отклик – тип ткани образца: нормальная ткань или один из 14 видов рака.

Частота ошибки одного дерева: **45.7%**

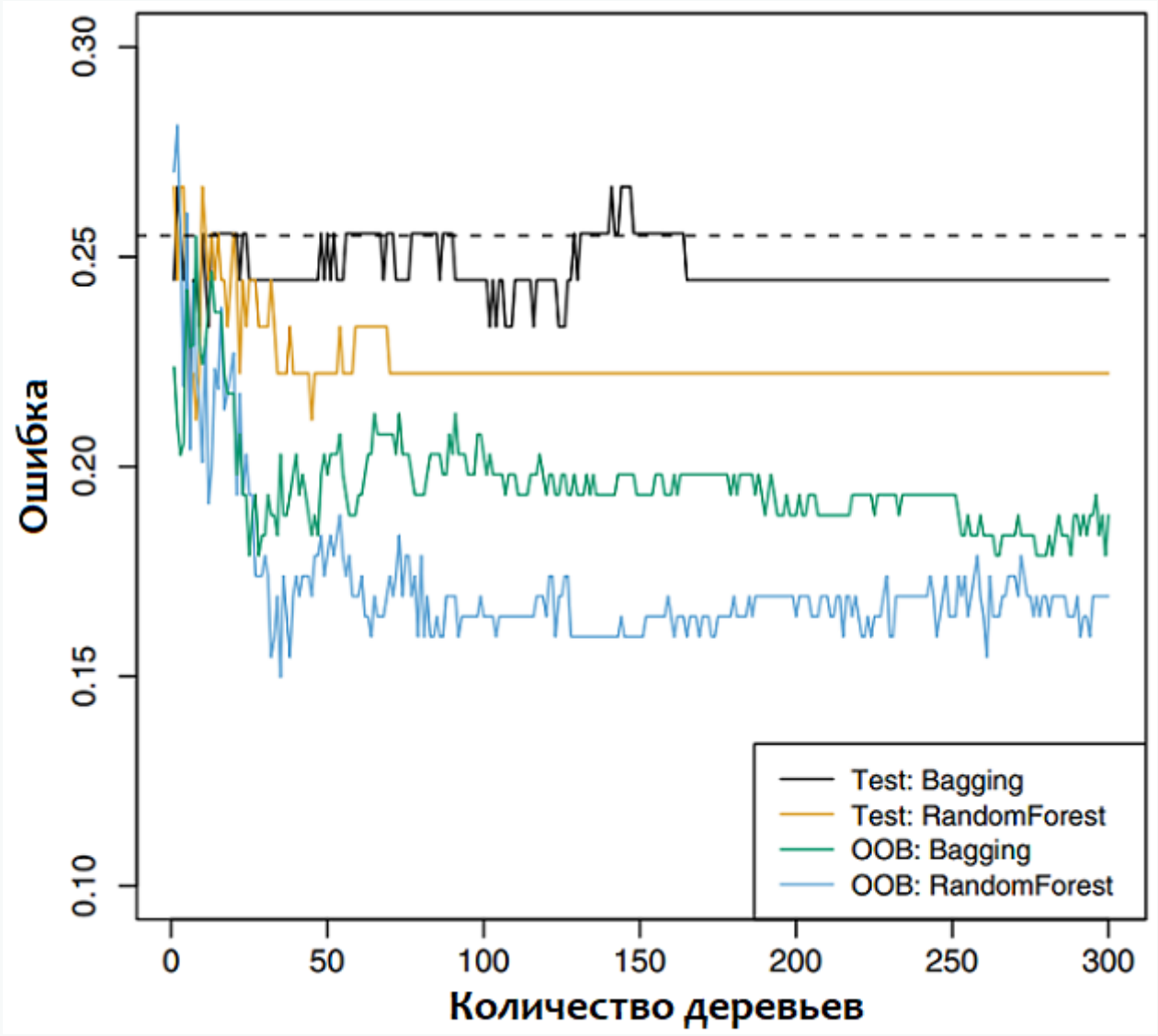


Бэггинг vs случайный лес

Данные `Heart`.

Пунктиром показана ошибка одного дерева на контрольной выборке.

OOB (ошибки по оставшимся данным) ниже ошибок на контрольной выборке.



Бустинг

- вместо бутстрепа применяется метод отбора наблюдений, основанный на результатах построения дерева
- деревья строятся последовательно, следующее – на остатках предыдущего
- деревья неглубокие, с числом узлов $d = 1$, $d = 2$
- при $d = 1$ получаем аналог линейной модели
- модель работает по принципу *медленного обучения*, скорость контролируется гиперпараметром λ

Алгоритм бустинга

(1) Присвоить $\hat{f}(x) = 0$ и $r_i = y_i \forall i$ в обучающей выборке

(2) Для $b = 1, 2, \ldots, B$:

(а) построить дерево $\hat{f}^b$ с d внутренними узлами по обучающим данным (X, r) ;
(б) обновить $\hat{f}$, добавив обрезанную версию нового дерева:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

(в) обновить остатки: $r_i \leftarrow r_i - \lambda \hat{f}^b(x)$

(3) Итоговая модель: $\hat{f}(x) = \sum_{b=1}^{B} \hat{f}^b(x)$

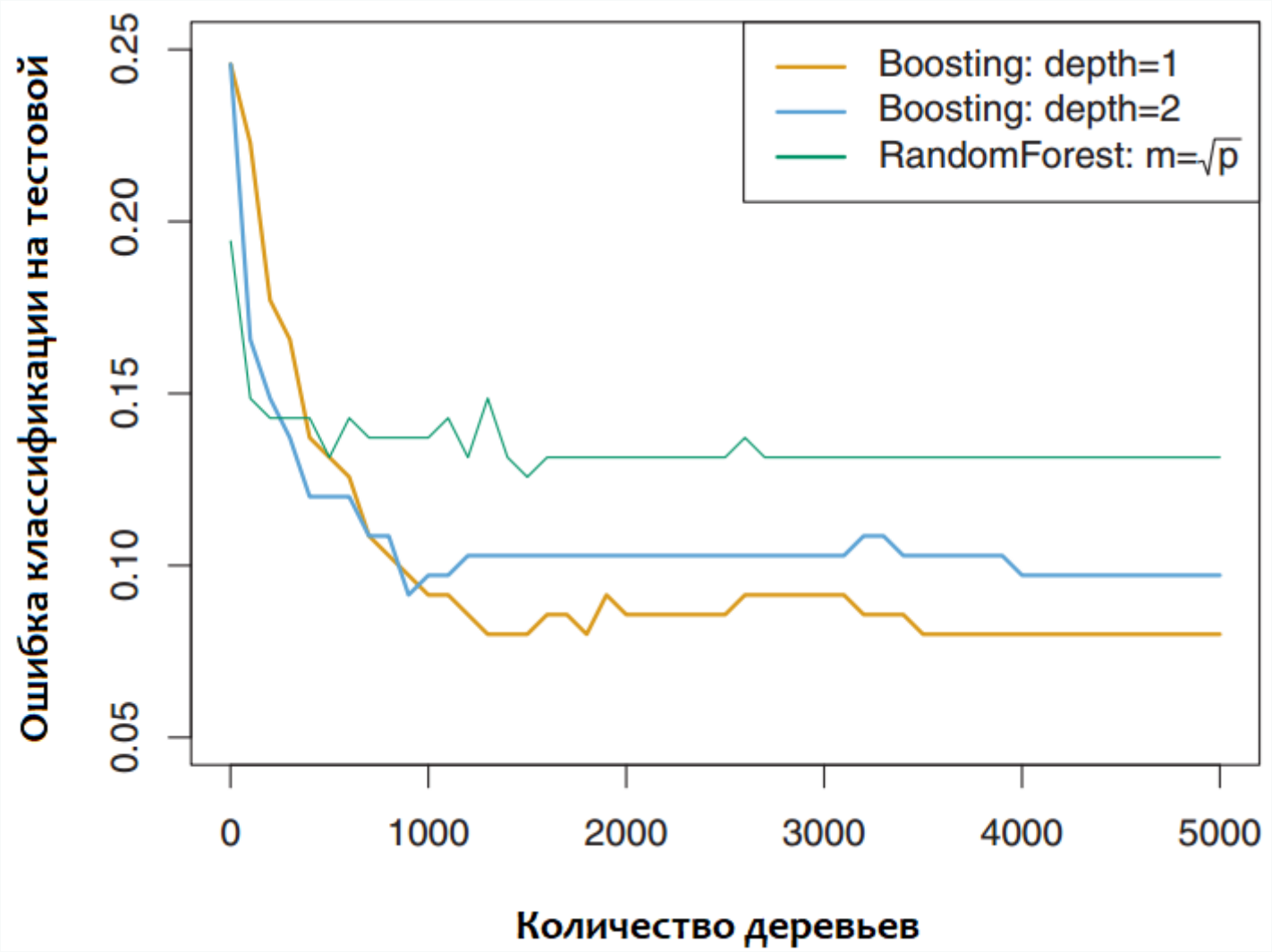
Три гиперпараметра:

- Число деревьев B (подбор перекрёстной проверкой). При больших B возможно переобучение.
- Параметр сжатия λ – скорость обучения (обычно $0.01 \leq \lambda \leq 0.001$). Низкие значения λ соответствуют высоким B.
- Число внутренних узлов деревьев b – глубина взаимодействий между предикторами.

Бустинг vs случайный лес

Данные по экспрессии 20000 генов в образцах 349 пациентов. Отклик – тип ткани образца: нормальная ткань или один из 14 видов рака.

Частота ошибки одного дерева: **24%**



Сравнение методов на основе деревьев решений

Метод	Преимущества	Недостатки
Одиночное дерево	легко визуализировать	неустойчиво к изменению входных данных
Бэггинг	легче оценить ошибку модели (OOB); более устойчив, чем одно дерево	корреляция между деревьями; возможен "перевес" в пользу сильного предиктора
Случайный лес	нечувствительность к несущественным признакам и зашумлённым наборам данных	визуализация деревьев невозможна
Бустинг	медленное обучение, тонкая подстройка по данные	взаимодействие между предикторами ограничено, много параметров для настройки

Источники

1. *Джеймс Г.*, *Уиттон Д.*, *Хасти Т.*, *Тибширани Р.* Введение в статистическое обучение с примерами на языке R. Пер. с англ. С.Э. Мастицкого – М.: ДМК Пресс, **2016** – 450 с.

2. *Бринк Х., Ричардс Дж., Феверолф М.* Машинное обучение. – Спб.: Питер, **2018**. – 336 с.

3. *Анналин Ын, Кеннет Су* Теоретический минимум по Big Data. Всё, что нужно знать о больших данных. – Спб.: Питер, **2019**. – 208 с.

4. *Annalyn Ng* Would you survive a disaster? / kdnuggets.com. URL: https://www.kdnuggets.com/2016/09/decision-trees-disastrous-overview.html

5. Данные `Titanic`, `Hitters`, `Heart` (https://web.stanford.edu/~hastie/ElemStatLearn/data.html).