

Введение в анализ данных

Лекция 9

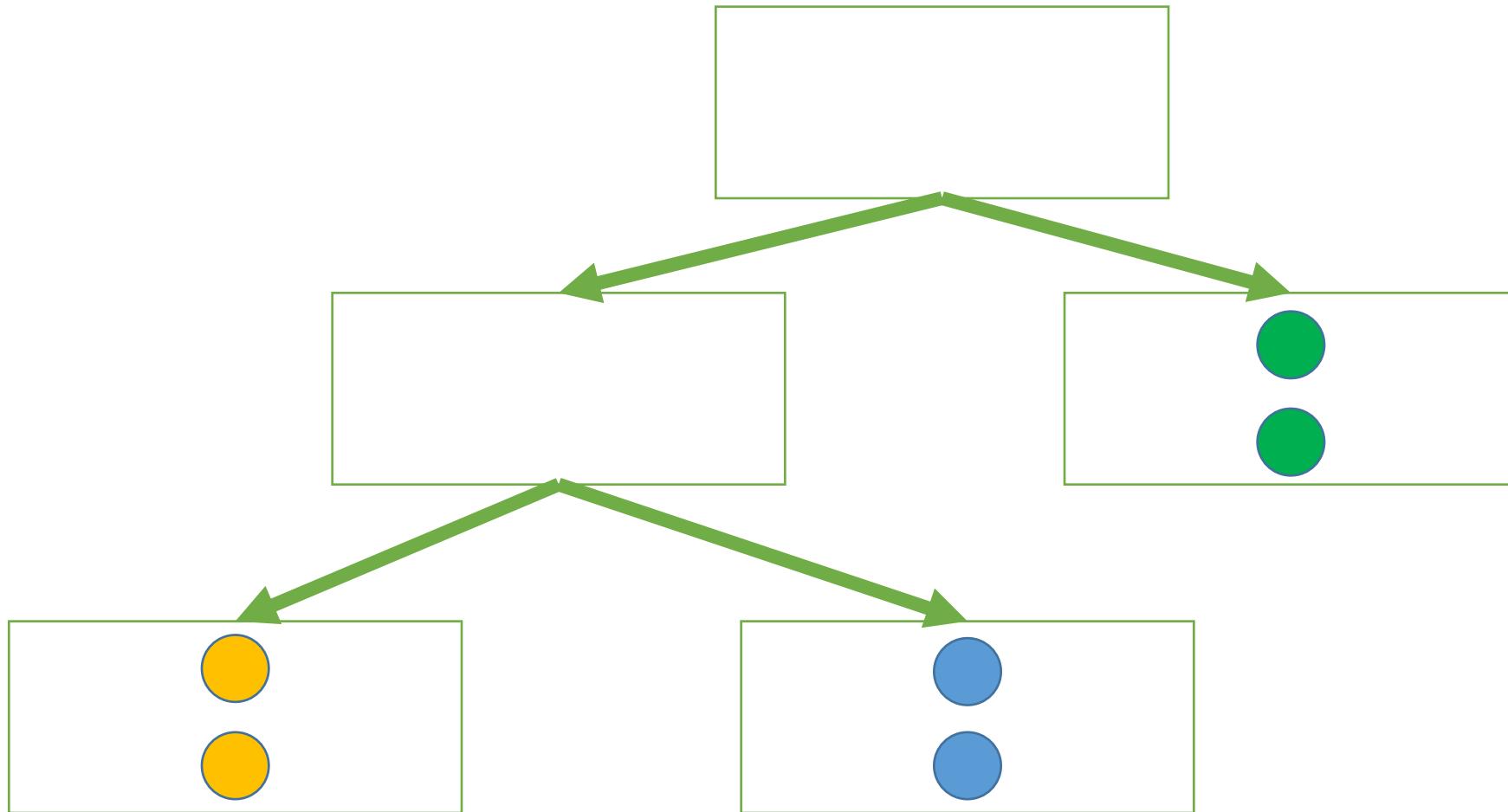
Решающие деревья и случайные леса

Евгений Соколов

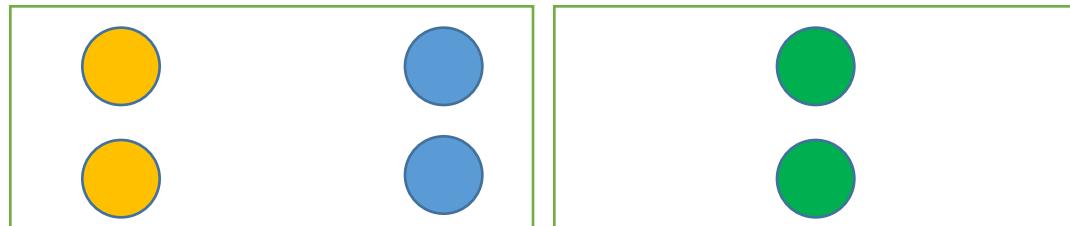
esokolov@hse.ru

НИУ ВШЭ, 2019

Жадное построение



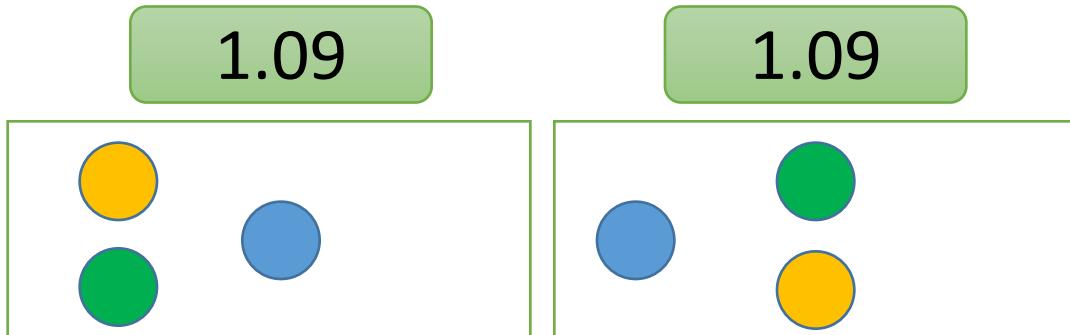
Как сравнить разбиения?



0.693

0

- $(0.5, 0.5, 0)$ и $(0, 0, 1)$
- $H = 0.693 + 0 = 0.693$

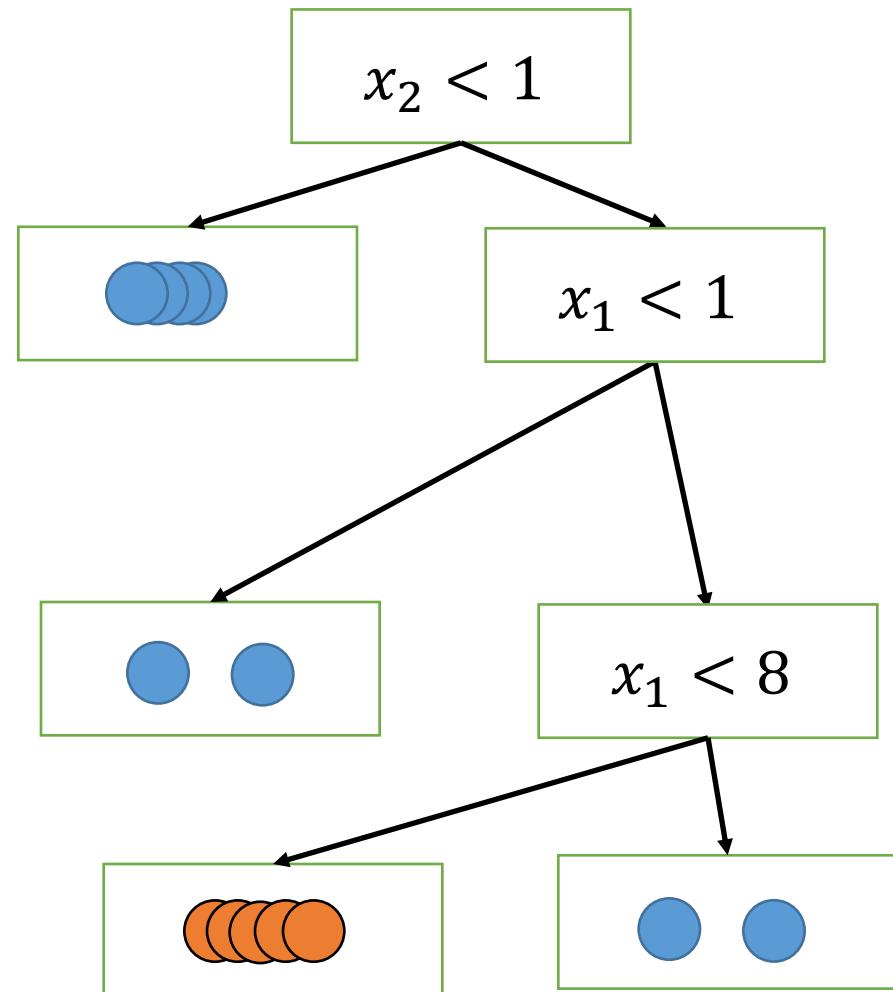
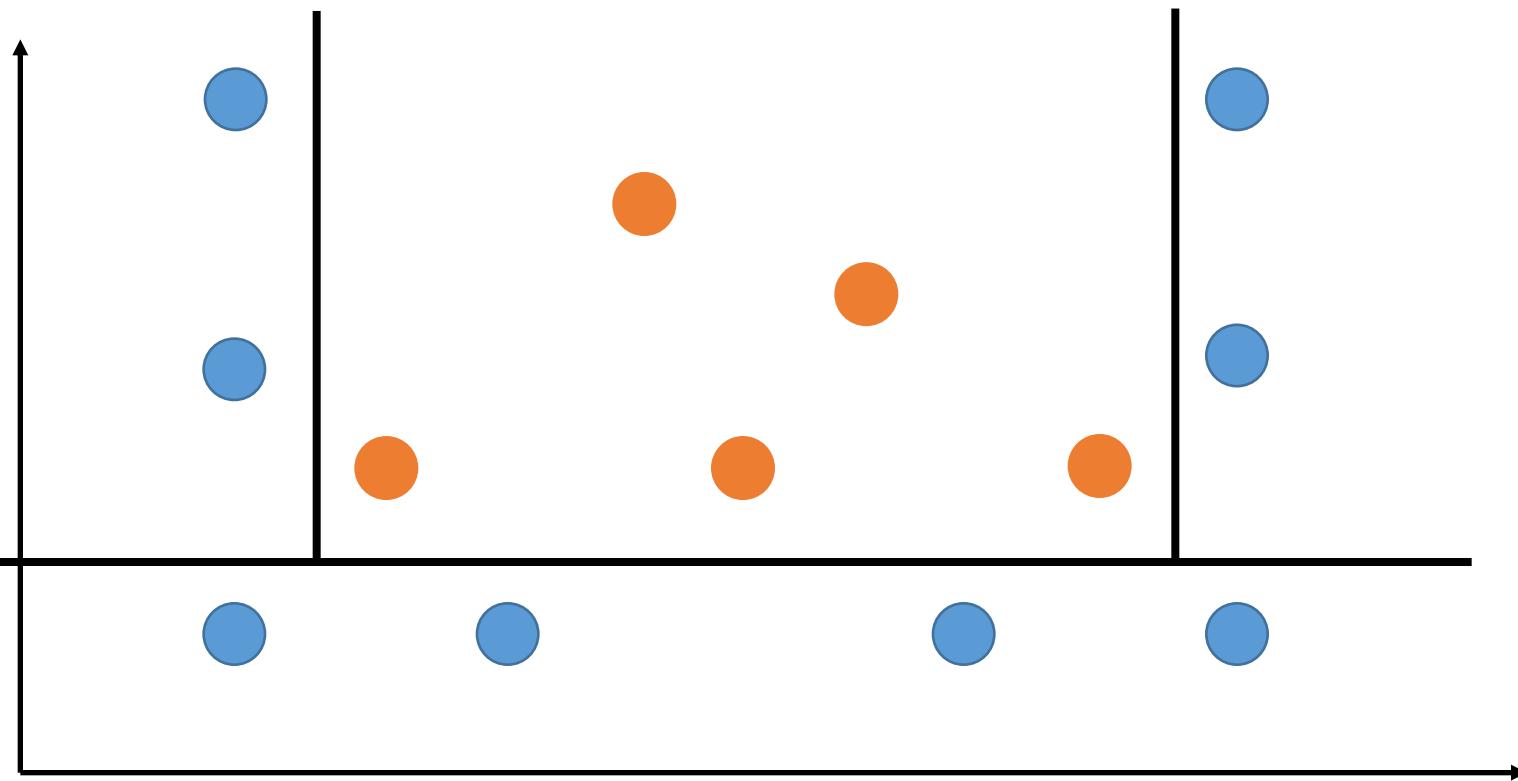


1.09

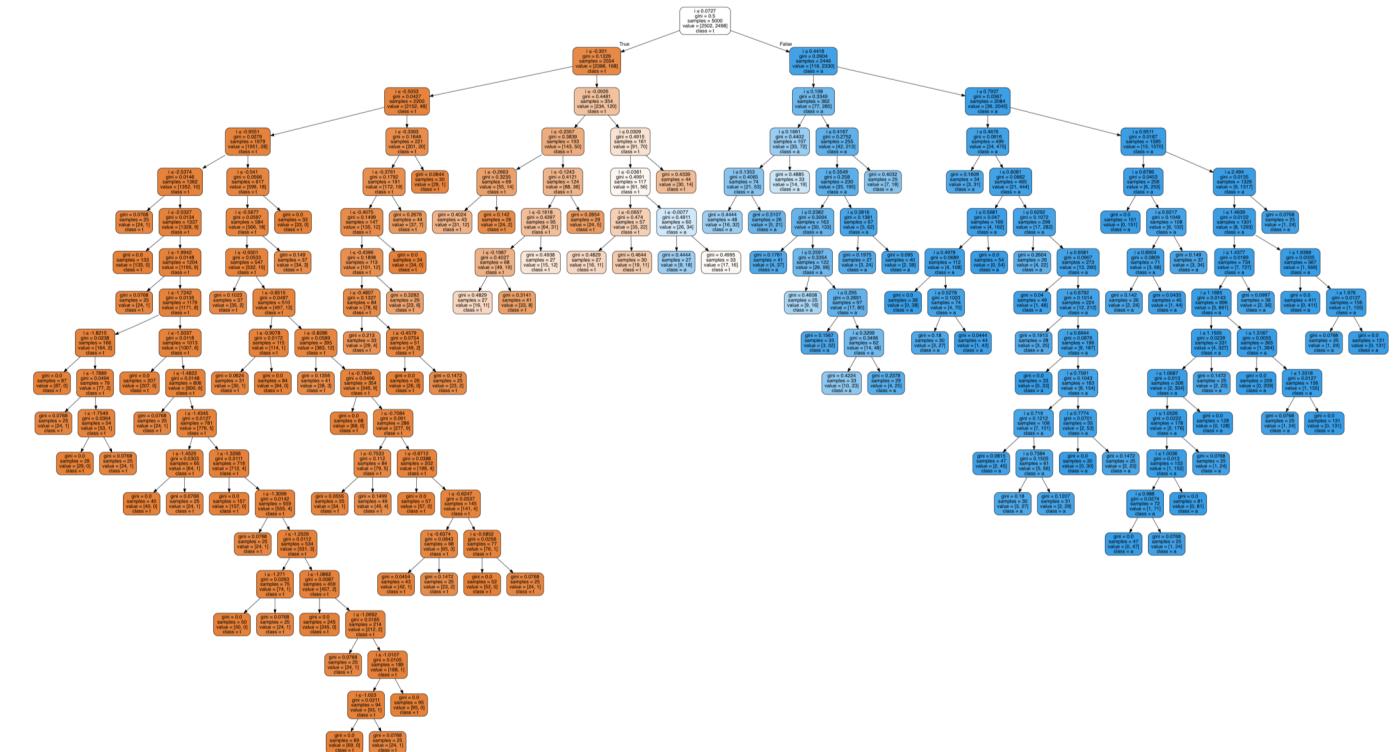
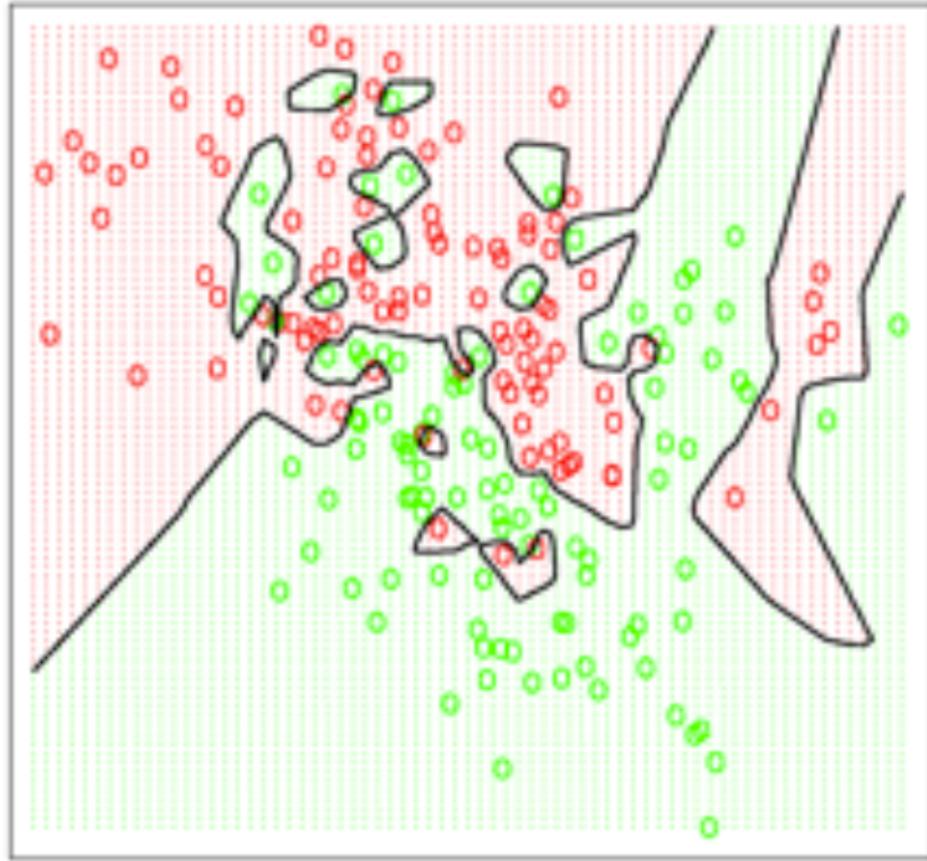
1.09

- $(0.33, 0.33, 0.33)$ и $(0.33, 0.33, 0.33)$
- $H = 1.09 + 1.09 = 2.18$

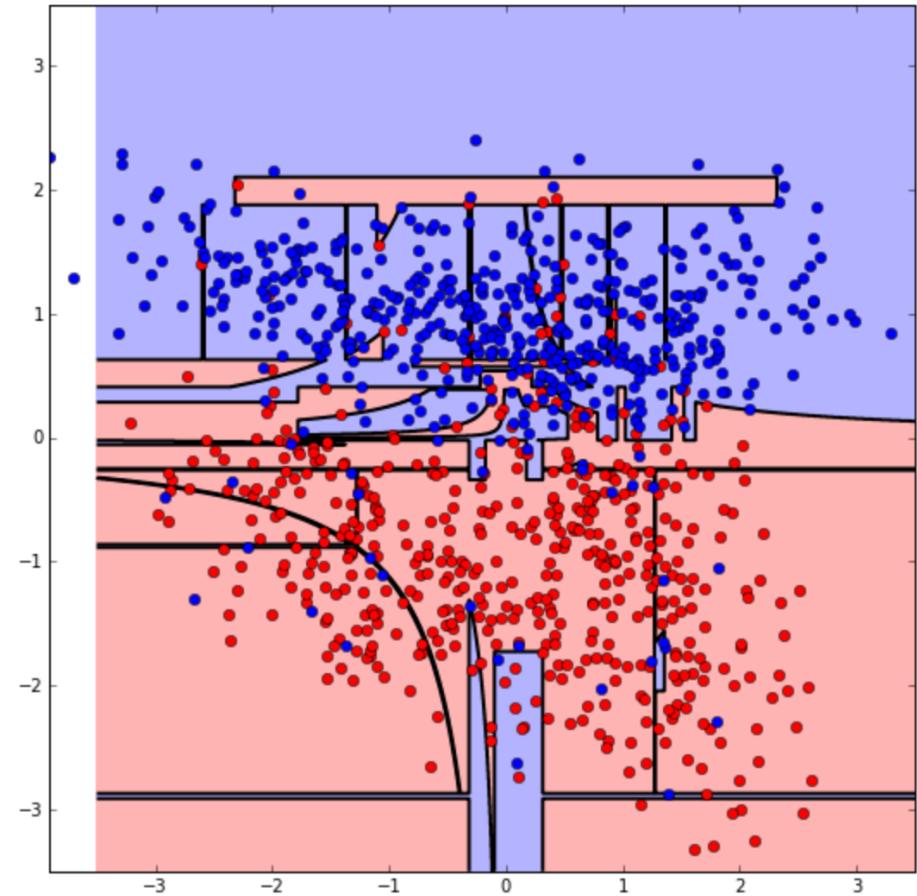
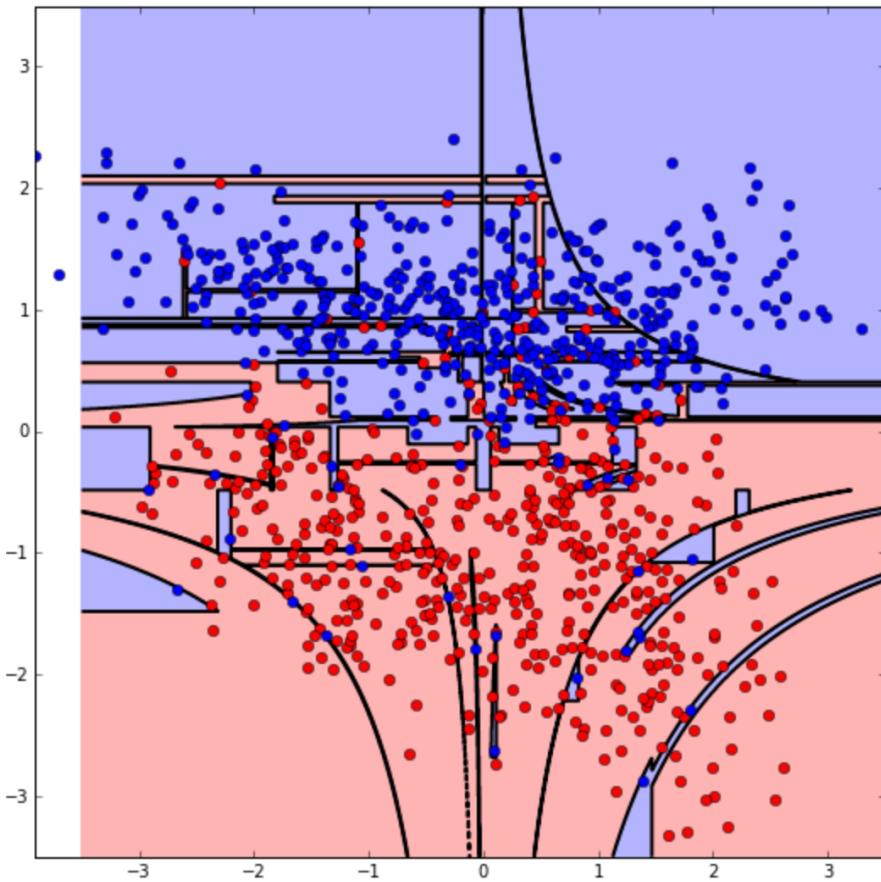
Обучение деревьев



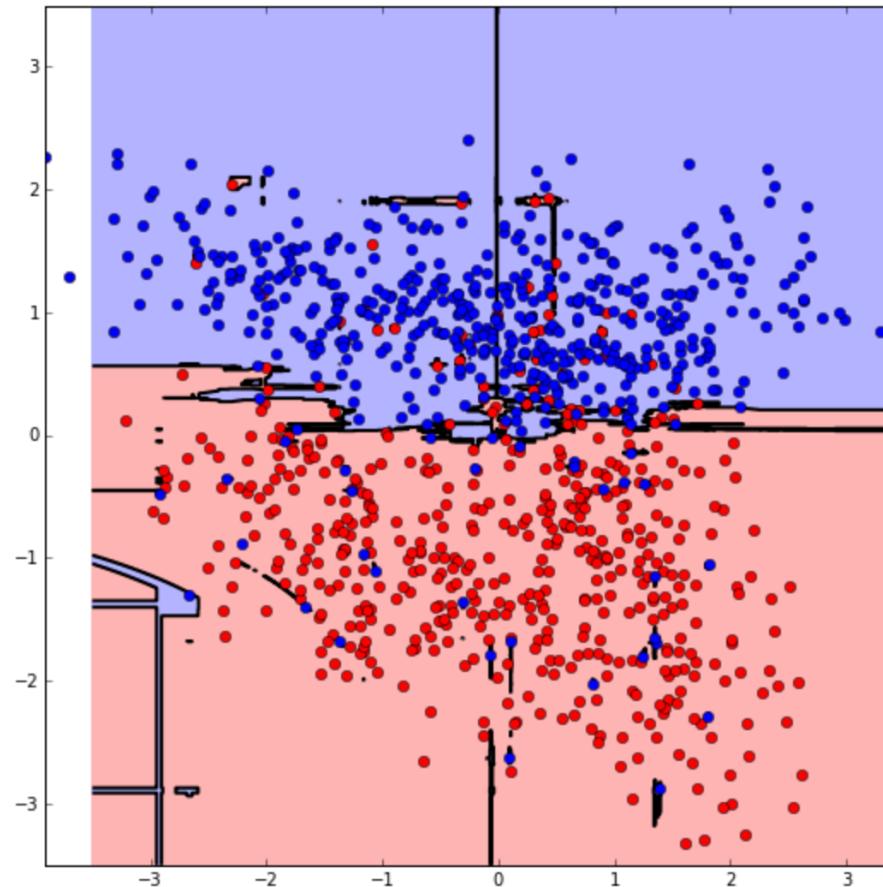
Переобучение деревьев



Неустойчивость деревьев



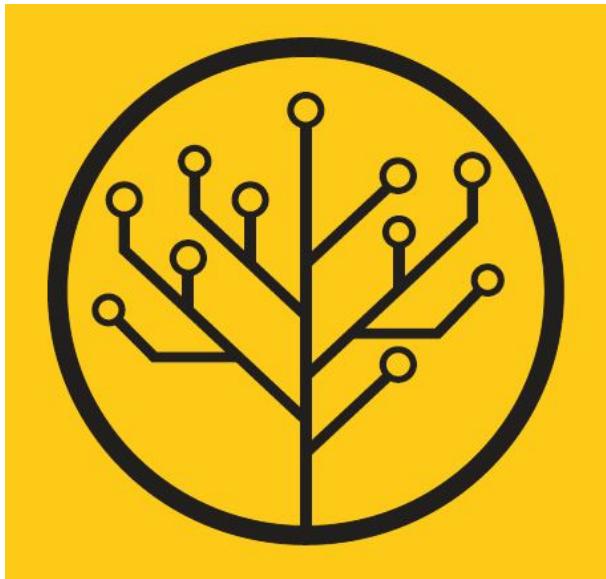
Усреднение деревьев



Композиции алгоритмов

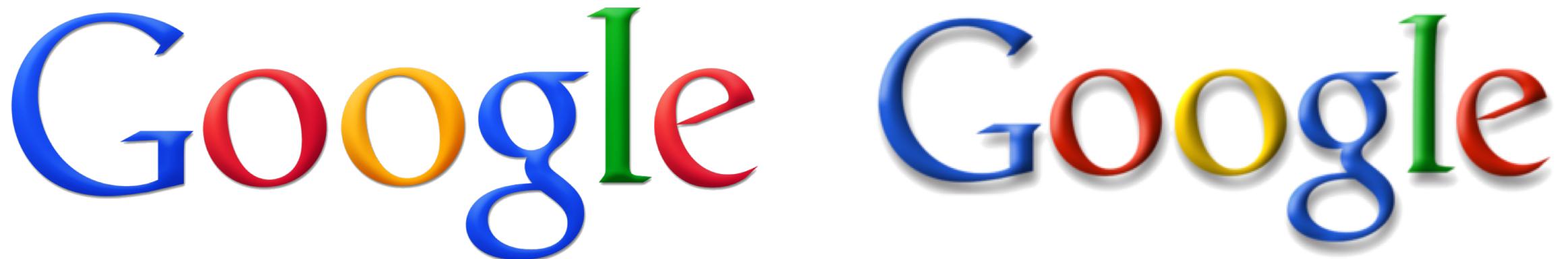
Majority vote

- Как выглядит логотип факультета компьютерных наук?



Majority vote

- Как выглядит логотип факультета компьютерных наук?
- Какой из двух логотипов более старый?



Majority vote

- Как выглядит логотип факультета компьютерных наук?
- Какой из двух логотипов более старый?
- Как выглядит корпус Вышки в Перми?



Majority vote

- Как выглядит логотип факультета компьютерных наук?
- Какой из двух логотипов более старый?
- Как выглядит корпус Вышки в Перми?
- Покоординатный спуск — это метод оптимизации 1-го или 2-го порядка?

Majority vote

- Дано: N базовых алгоритмов $b_1(x), \dots, b_N(x)$
- Каждый хотя бы немного лучше случайного угадывания
- Композиция: класс, за который проголосовало больше всего базовых алгоритмов

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{n=1}^N [b_n(x) = y]$$

Усреднение наблюдений

- Наблюдение: усреднение результатов повышает их точность
- Измерение артериального давления
- Измерение скорости света
- Усреднение соседних пикселей изображения

Усреднение наблюдений

- Сколько лет факультету компьютерных наук?

Усреднение наблюдений

- Сколько лет факультету компьютерных наук?
- Сколько метров в 1 сажени?

Усреднение наблюдений

- Сколько лет факультету компьютерных наук?
- Сколько метров в 1 сажени?
- Сколько лет лектору?

Усреднение наблюдений

- Сколько лет факультету компьютерных наук?
- Сколько метров в 1 сажени?
- Сколько лет лектору?
- Сколько всего стран в мире?

Усреднение наблюдений

- Дано: N базовых алгоритмов $b_1(x), \dots, b_N(x)$
- Каждый хотя бы немного лучше случайного угадывания
- Композиция:

$$a(x) = \frac{1}{N} \sum_{n=1}^N b_n(x)$$

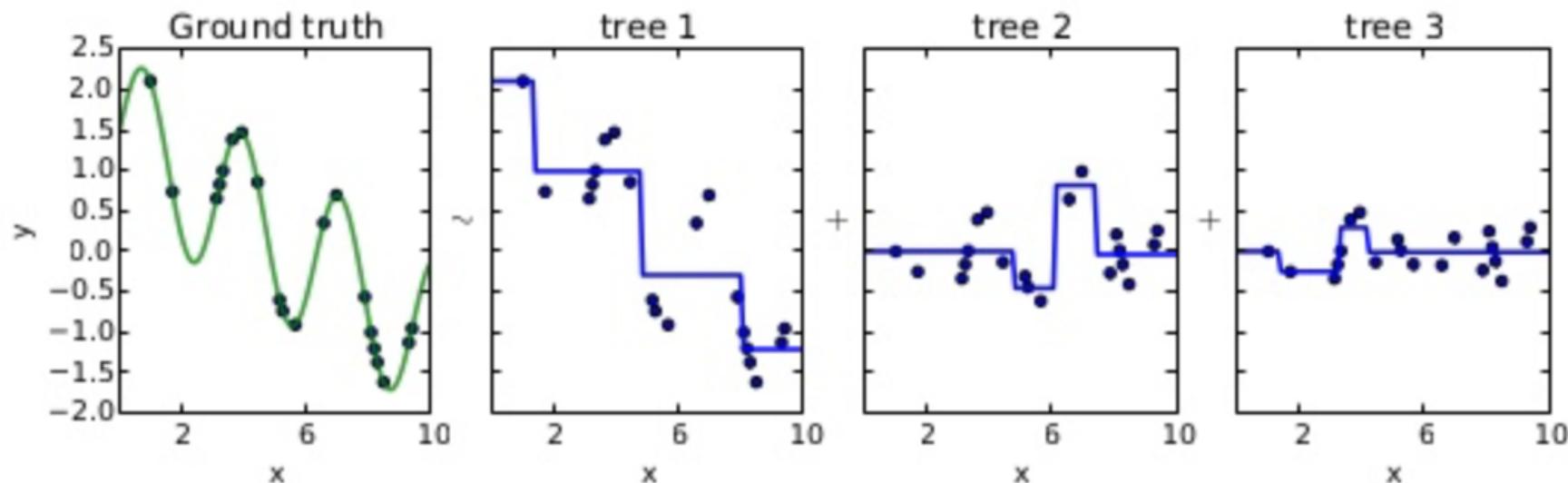
Композиции алгоритмов

- Базовые алгоритмы: $b_1(x), \dots, b_N(x)$
- Композиция: $a(x)$

- Как по одной и той же выборке обучить N различных моделей?

БУСТИНГ

- Каждый следующий алгоритм исправляет ошибки предыдущих
- Яркий пример: градиентный бустинг над решающими деревьями
- В следующий раз



БЭГГИНГ

- Bagging (Bootstrap Aggregation)
- Базовые алгоритмы обучаются независимо
- Каждый обучается на подмножестве данных
- Усреднение ответов или выбор по большинству
- Яркий пример: случайный лес (random forest)

БЭГГИНГ

Идея:

- Обучим много деревьев $b_1(x), \dots, b_N(x)$
- Выберем ответ по большинству:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{n=1}^N [b_n(x) = y]$$

Пример

- Прогнозы деревьев: $-1, -1, 1, -1, 1, -1$

$$a(x) = ?$$

Пример

- Прогнозы деревьев: $-1, -1, 1, -1, 1, -1$

$$a(x) = -1$$

Рандомизация

- Как сделать деревья разными?
- Обучать по подвыборкам!

Рандомизация

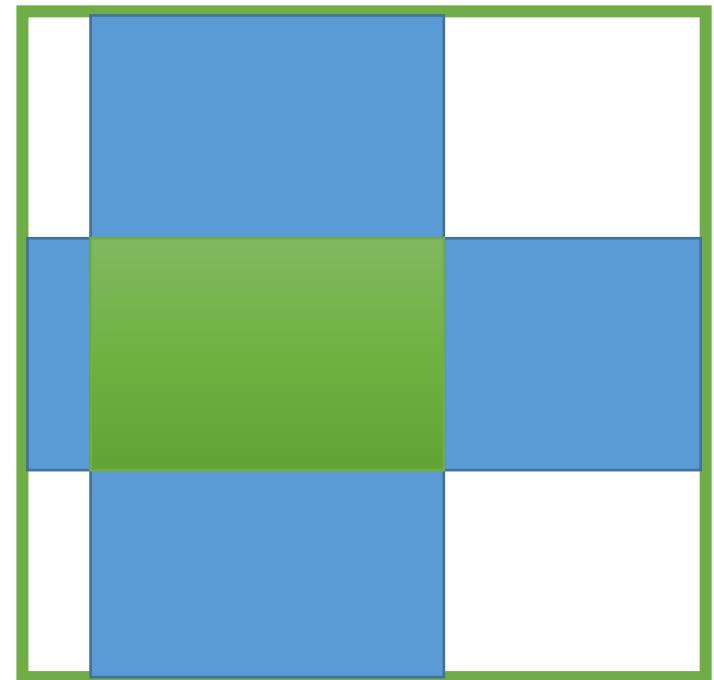
- Популярный подход: бутстрэп
- Выбираем из обучающей выборки ℓ объектов с возвращением
- Пример: $\{x_1, x_2, x_3, x_4\} \rightarrow \{x_1, x_2, x_2, x_4\}$
- Примерно $0.632 * \ell$ различных объектов

Рандомизация

- Другой подход: выбор случайного подмножества объектов
- Гиперпараметр: размер подмножества

Виды рандомизации

- Бэггинг: обучаем на случайной подвыборке
- Метод случайных подпространств: обучаем на случайном подмножестве признаков
- Размер подвыборки/подмножества — гиперпараметр



Рандомизация

- Этого недостаточно
- Как можно рандомизировать сам процесс построения дерева?

Поиск разбиения

- Пусть в вершине m оказалась выборка X_m
- $Q(X_m, j, t)$ — критерий ошибки условия $[x^j \leq t]$
- Ищем лучшие параметры j и t перебором:

$$Q(X_m, j, t) \rightarrow \min_{j,t}$$

Поиск разбиения

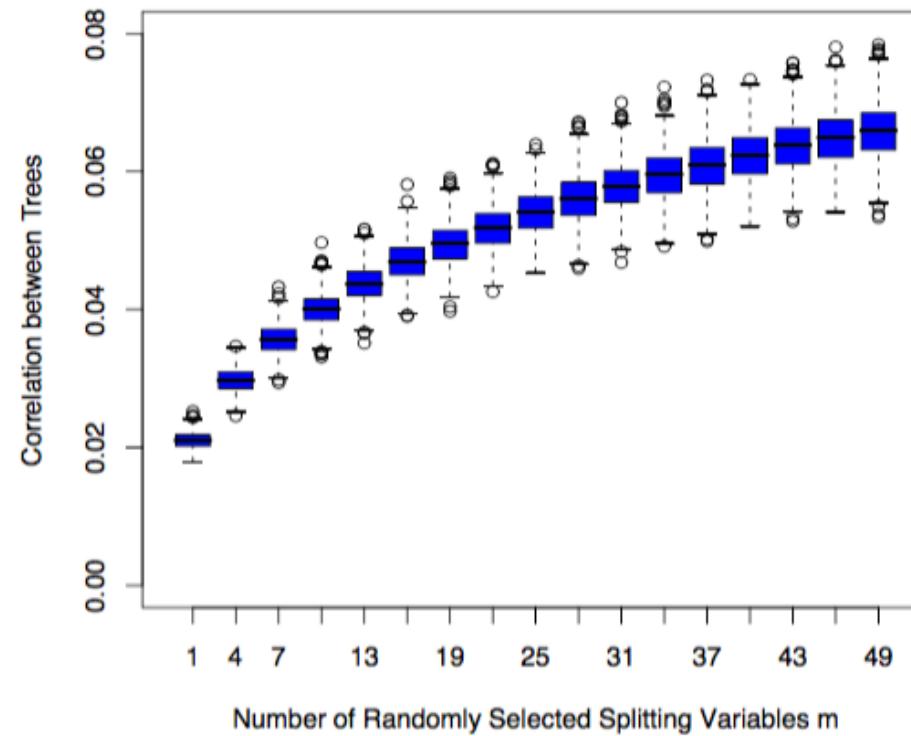
- Пусть в вершине m оказалась выборка X_m
- $Q(X_m, j, t)$ — критерий ошибки условия $[x^j \leq t]$
- Ищем лучшие параметры j и t перебором:

$$Q(X_m, j, t) \rightarrow \min_{j,t}$$

- Случайный лес: выбираем j из случайного подмножества признаков размера q



Корреляция между деревьями



Корреляция между деревьями

Рекомендации для q :

- Регрессия: $q = \frac{d}{3}$
- Классификация: $q = \sqrt{d}$

Случайный лес (Random forest)

1. Для $n = 1, \dots, N$:
2. Сгенерировать выборку \tilde{X} с помощью бутстрата
3. Построить решающее дерево $b_n(x)$ по выборке \tilde{X}
4. Дерево строится, пока в каждом листе не окажется не более n_{min} объектов
5. Оптимальное разбиение ищется среди q случайных признаков

Случайный лес (Random forest)

1. Для $n = 1, \dots, N$:
2. Сгенерировать выборку \tilde{X} с помощью бутстрата
3. Построить решающее дерево $b_n(x)$ по выборке \tilde{X}
4. Дерево строится, пока в каждом листе не окажется не более n_{min} объектов
5. Оптимальное разбиение ищется среди q случайных признаков

Выбираются заново при каждом разбиении!

Случайный лес

- Регрессия:

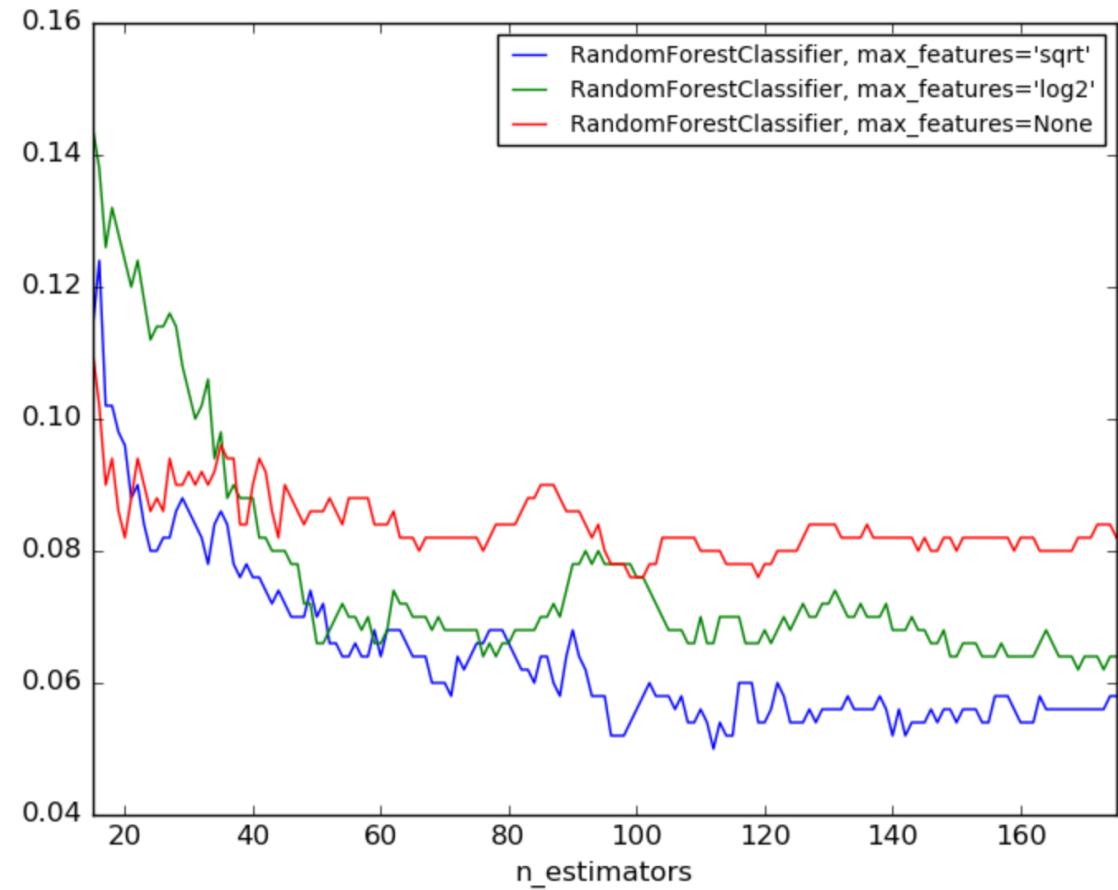
$$a(x) = \frac{1}{N} \sum_{n=1}^N b_n(x)$$

- Классификация:

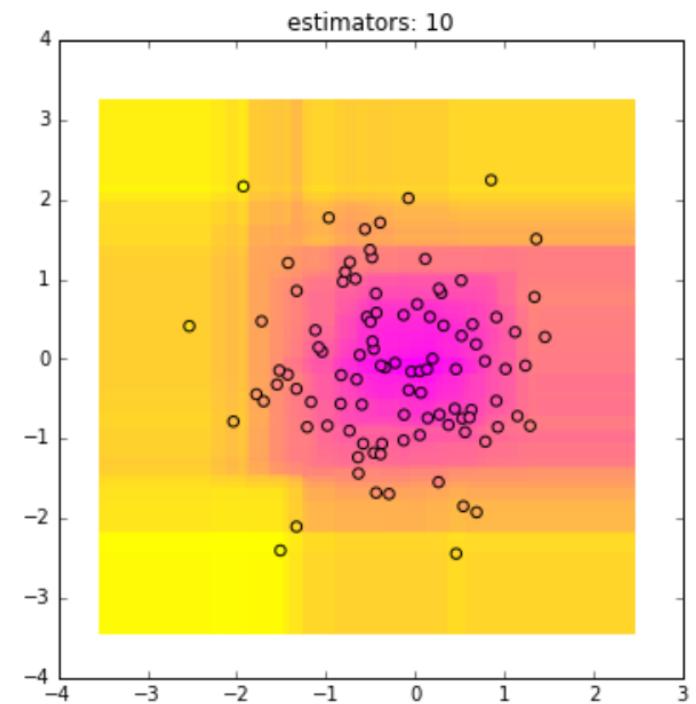
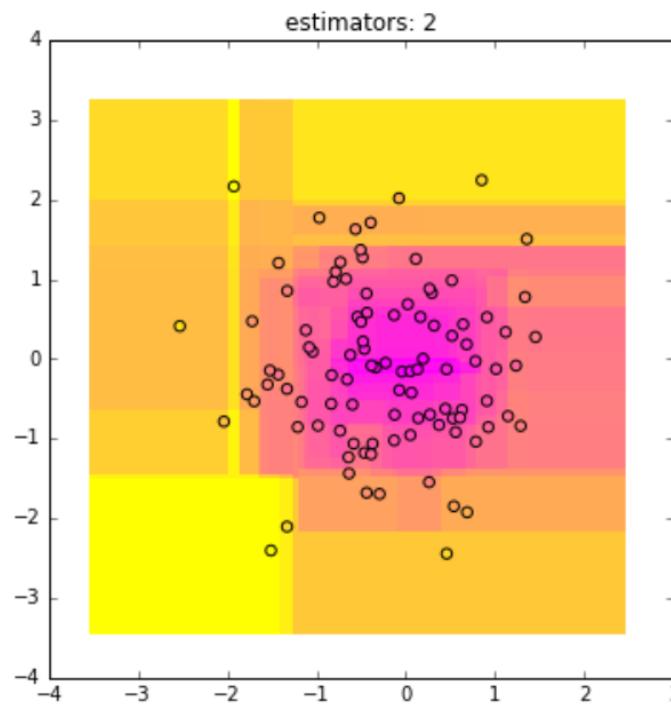
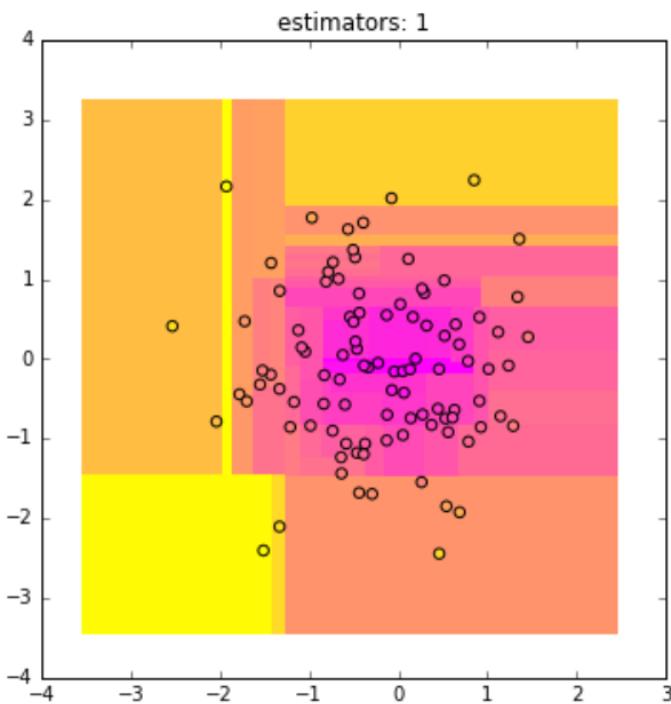
$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{n=1}^N [b_n(x) = y]$$

Ошибка на teste

- Ошибка сначала убывает, а затем остаётся примерно на одном уровне
- Случайный лес не переобучается при росте N



Случайный лес



Out-of-bag

- Каждое дерево обучается примерно на 63% данных
- Остальные объекты — как бы тестовая выборка для этого дерева
- X_n — обучающая выборка для $b_n(x)$
- Можно оценить ошибку на новых данных:

$$Q_{test} = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{1}{\sum_{n=1}^N [x_i \notin X_n]} \sum_{n=1}^N [x_i \notin X_n] b_n(x_i)$$

Важность признаков

Перестановочный метод:

- Проверяем важность j -го признака
- Перемешиваем соответствующий столбец в матрице «объекты-признаки» для тестовой выборки
- Измеряем качество модели
- Если оно слабо изменилось, то признак не очень важный