

Введение в анализ данных

Лекция 9

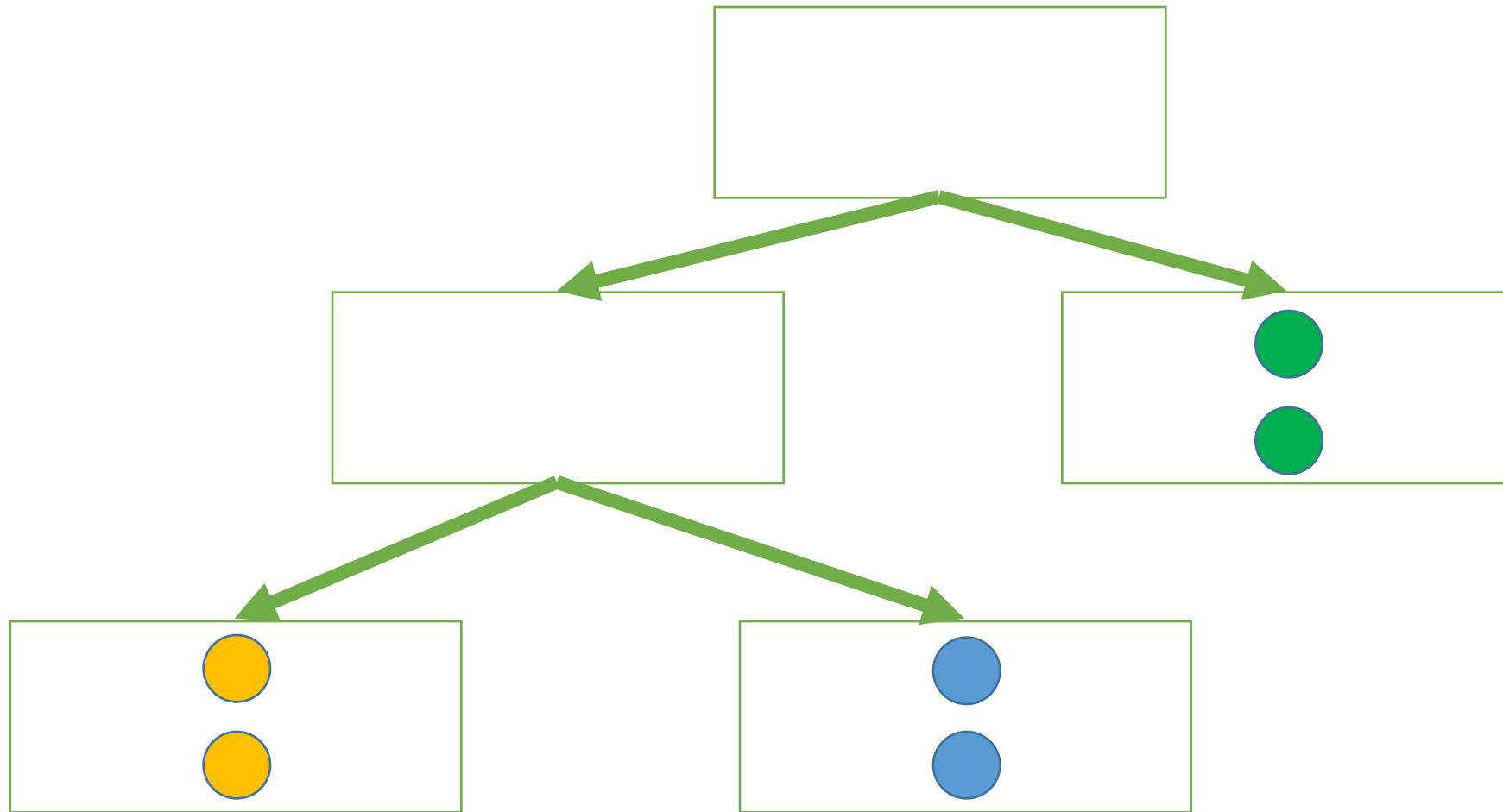
Решающие деревья и случайные леса

Евгений Соколов

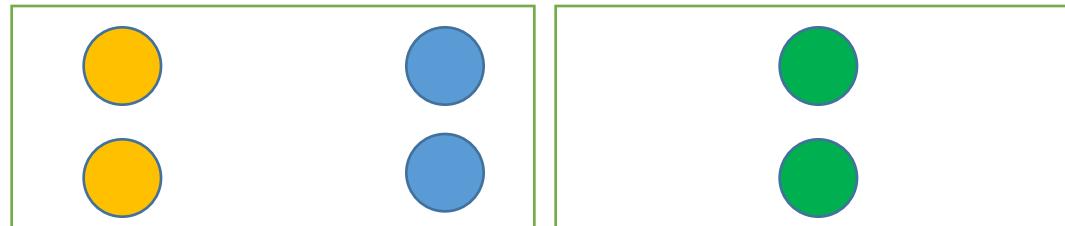
esokolov@hse.ru

НИУ ВШЭ, 2018

Жадное построение



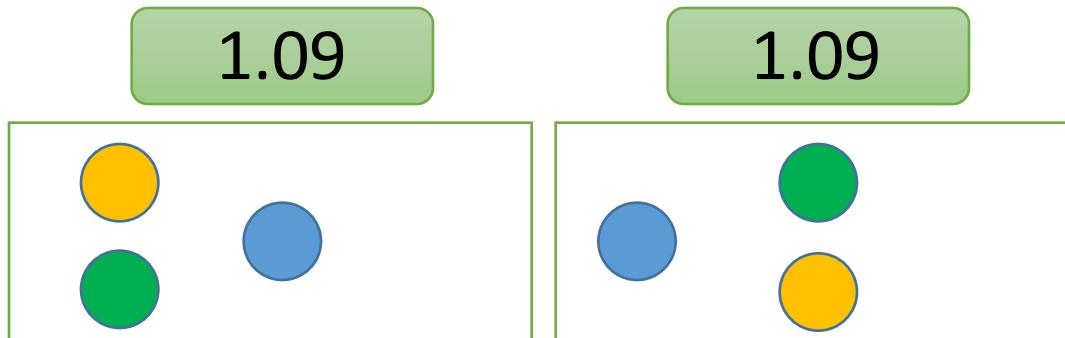
Как сравнить разбиения?



0.693

0

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

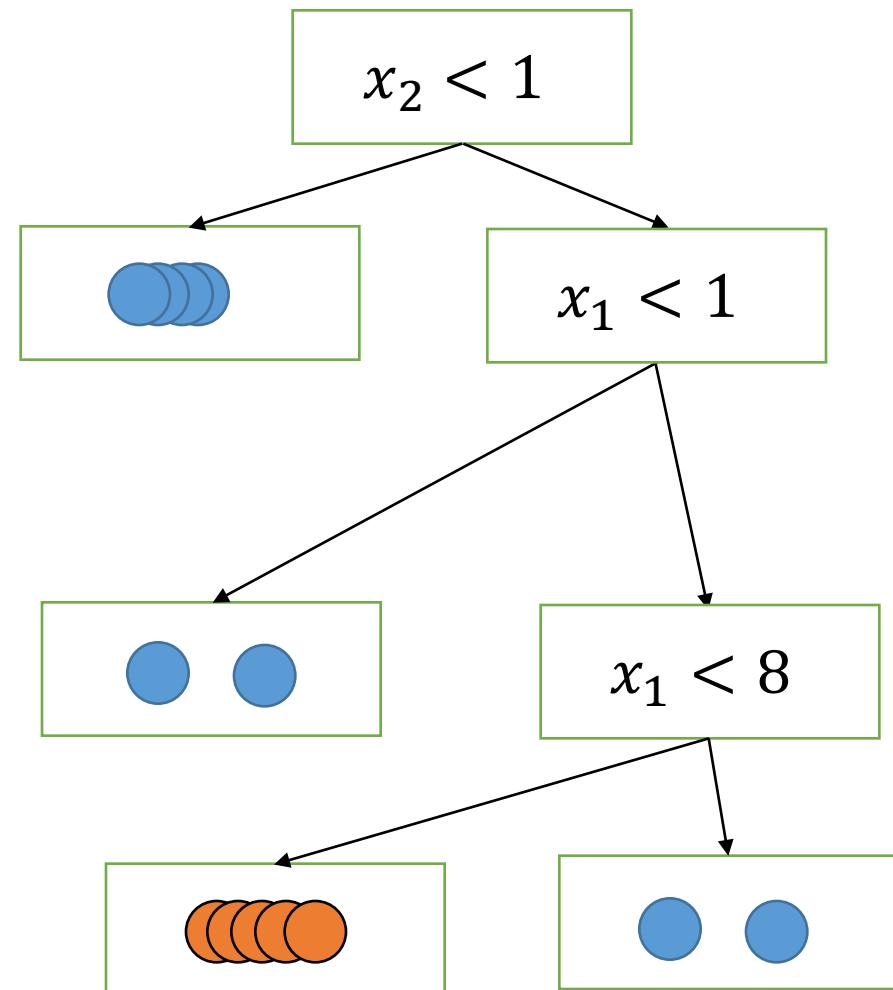
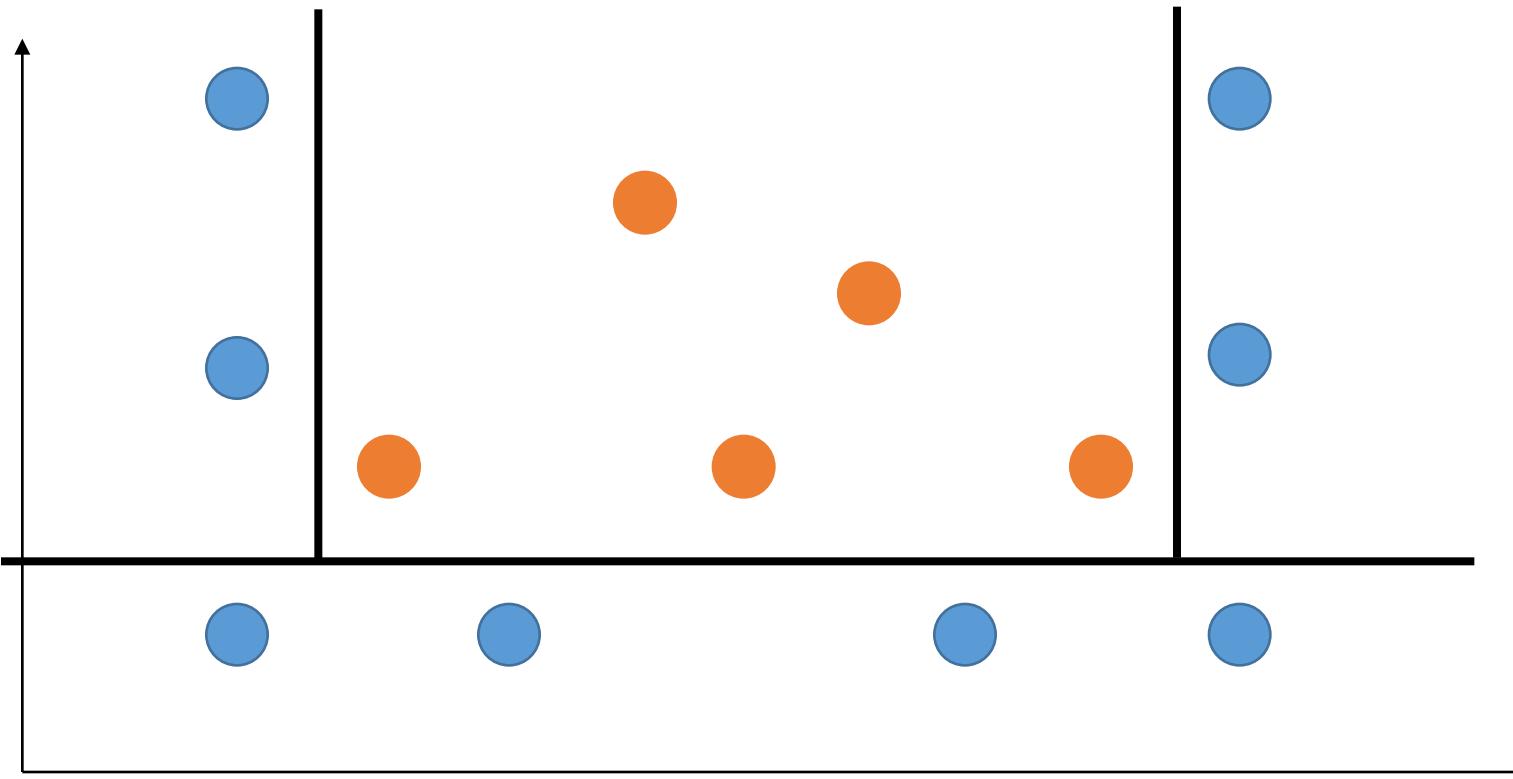


1.09

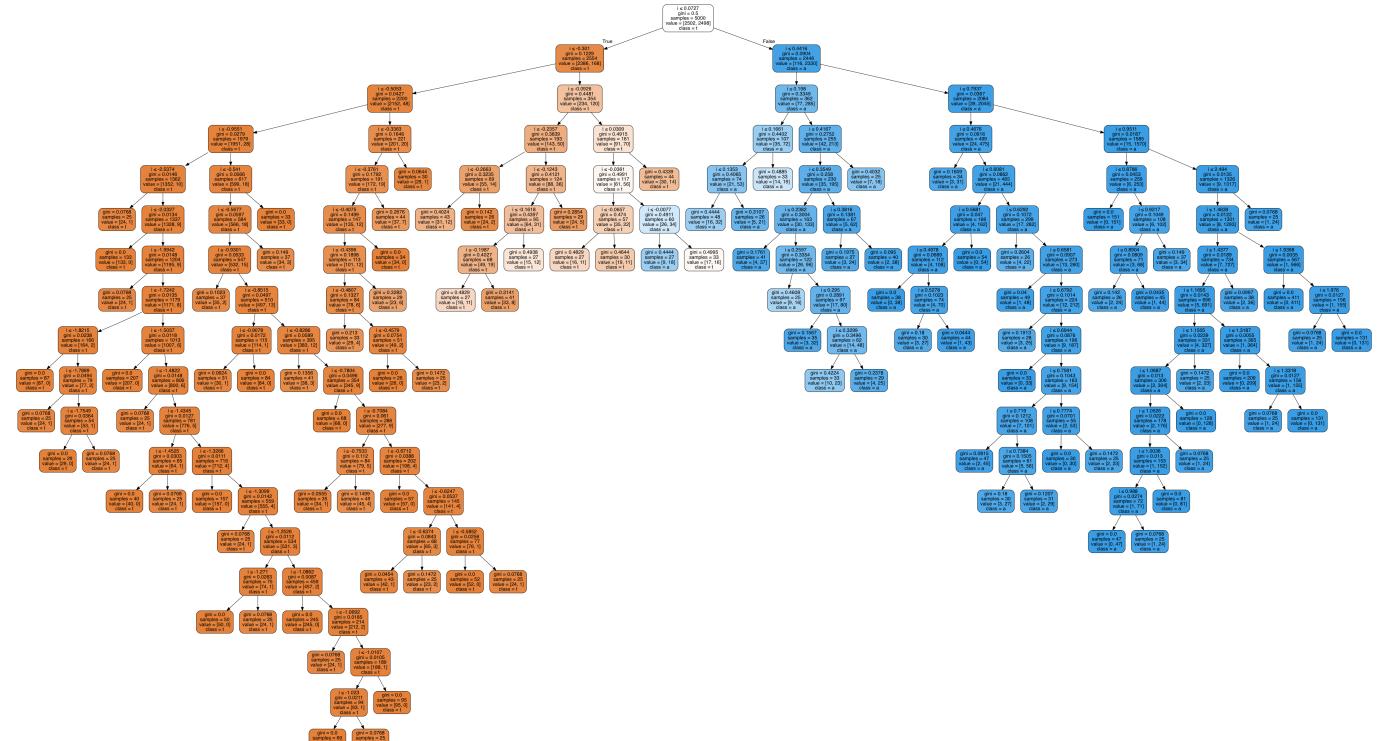
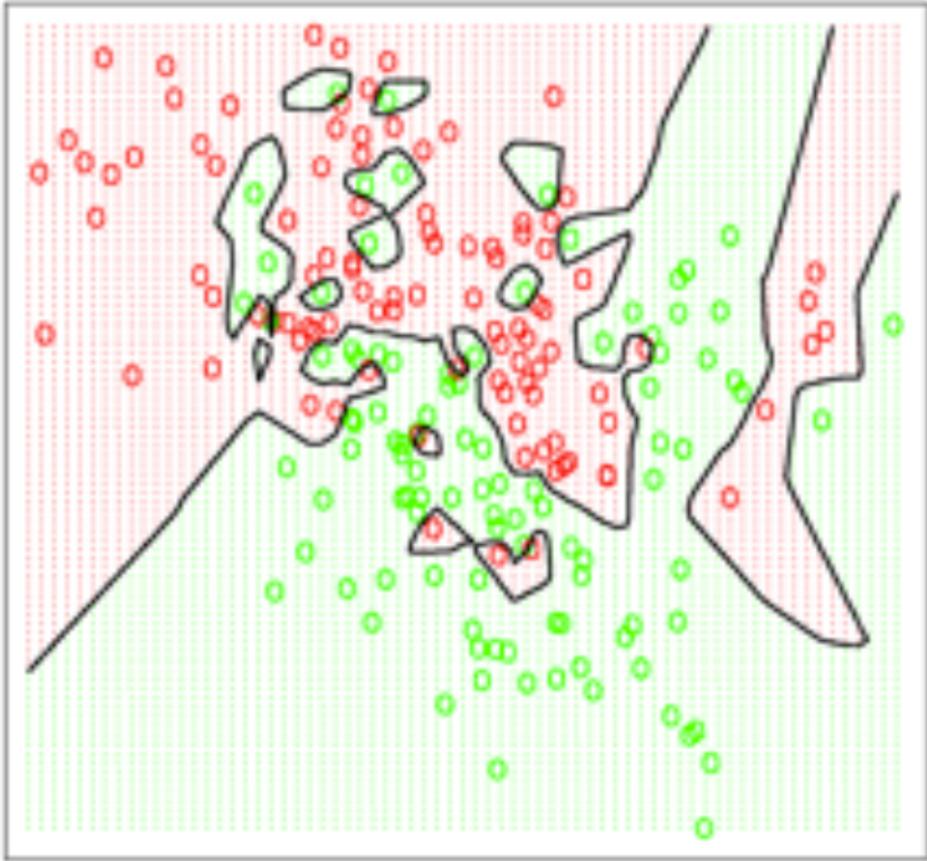
1.09

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

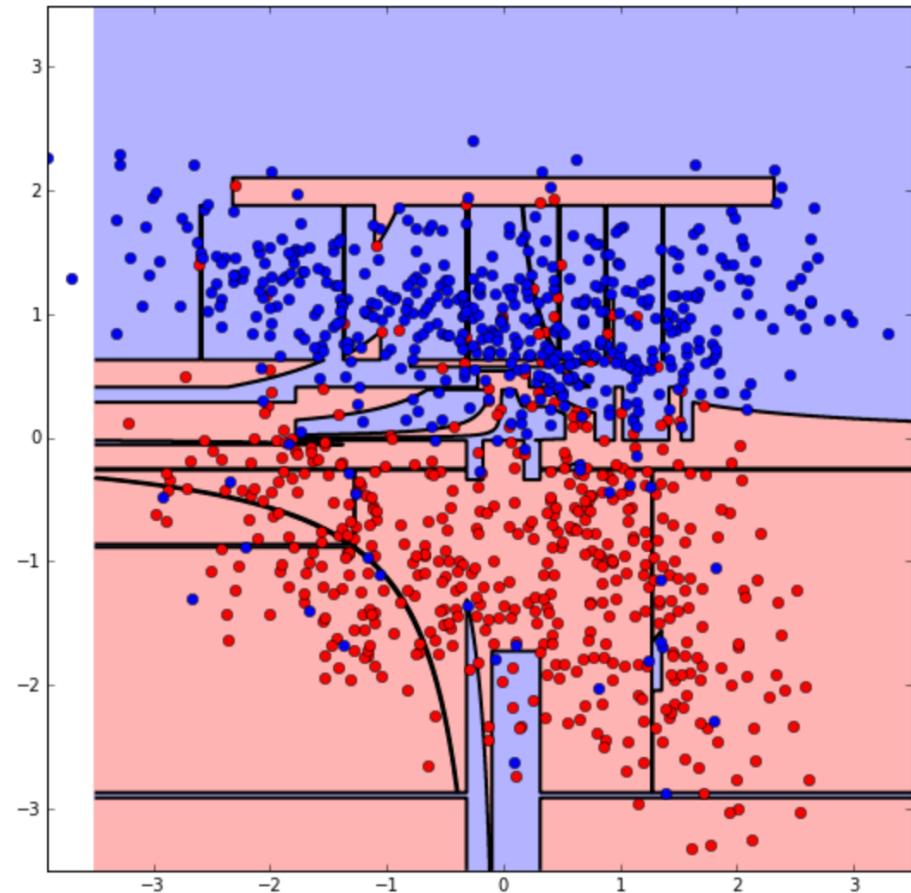
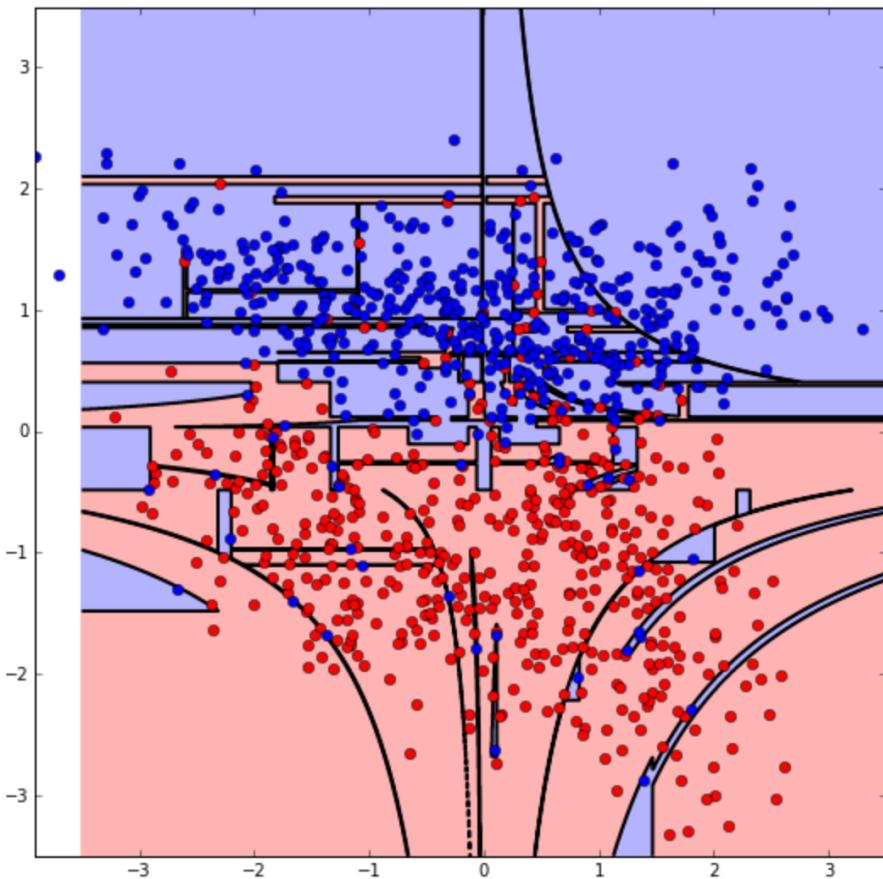
Обучение деревьев



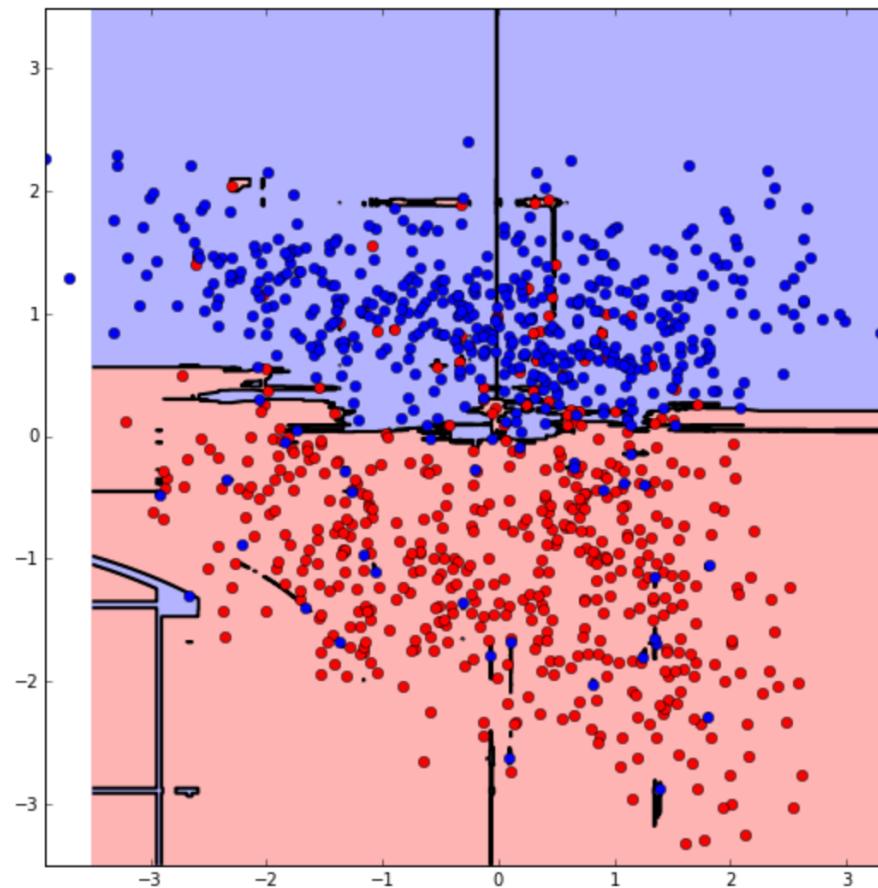
Переобучение деревьев



Неустойчивость деревьев



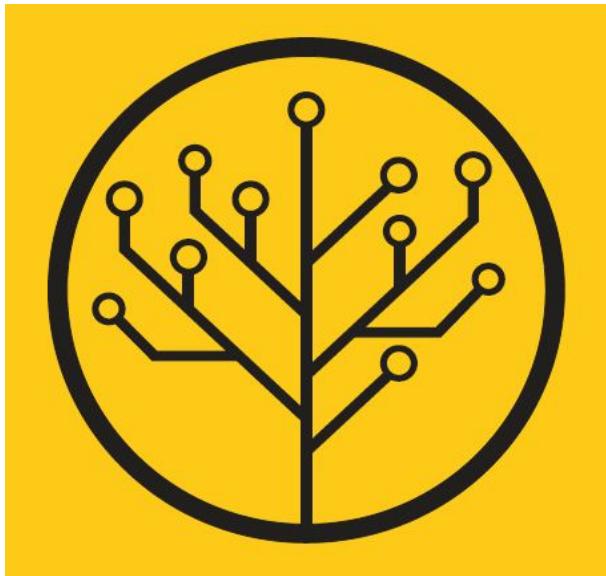
Усреднение деревьев



Композиции алгоритмов

Majority vote

- Как выглядит логотип факультета компьютерных наук?



Majority vote

- Как выглядит логотип факультета компьютерных наук?
- В каком полушарии находится пролив Магеллана?

Majority vote

- Как выглядит логотип факультета компьютерных наук?
- В каком полушарии находится пролив Магеллана?
- Как выглядит выхухоль?



Majority vote

- Как выглядит логотип факультета компьютерных наук?
- В каком полушарии находится пролив Магеллана?
- Как выглядит выхухоль?
- Градиентный 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

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

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

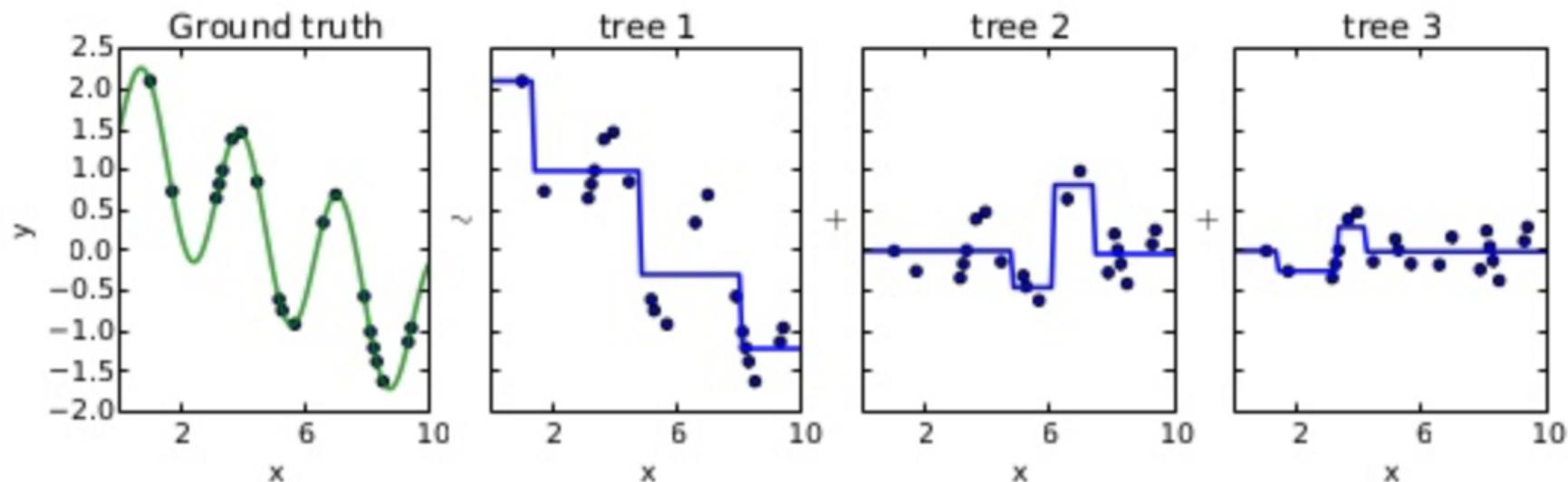
Композиции алгоритмов

- 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 следующем курсе



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

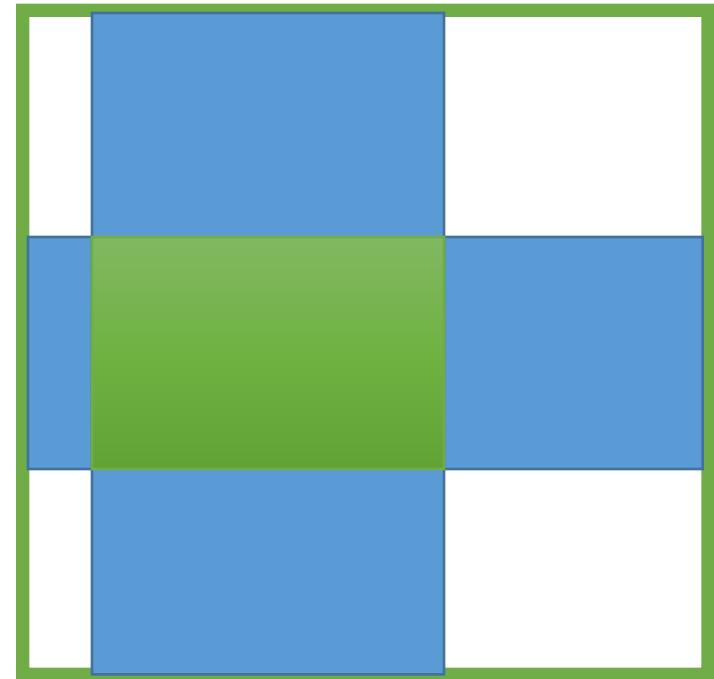
- Популярный подход: бутстрэп
- Выбираем из обучающей выборки ℓ объектов с возвращением
- Пример: $\{x_1, x_2, x_3, x_4\} \rightarrow \{x_1, x_2, x_2, x_4\}$
- Примерно $0.632 * \ell$ различных объектов

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

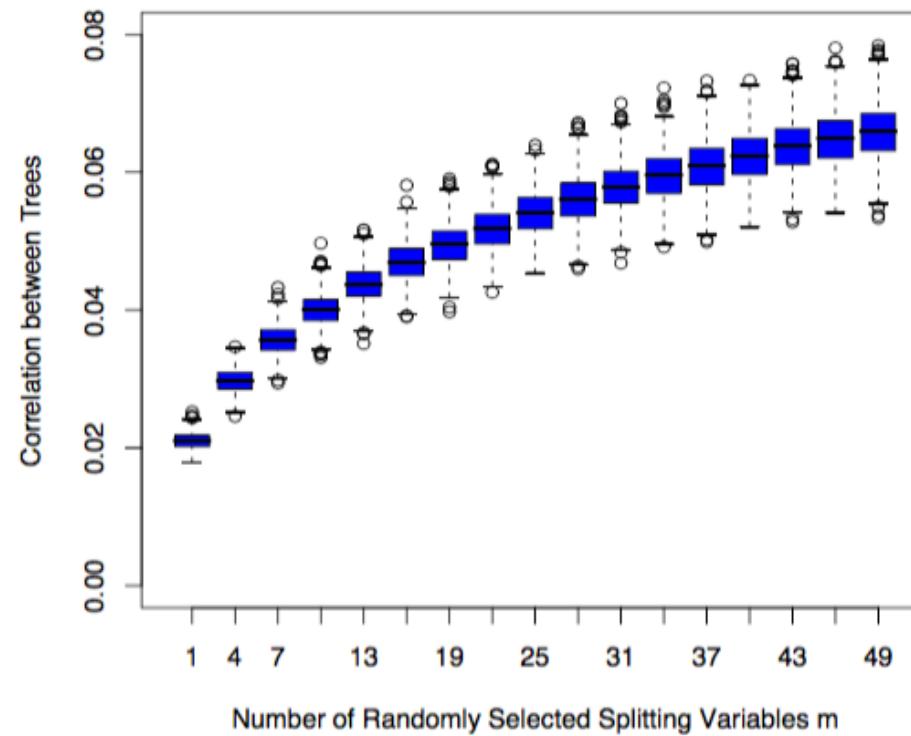
- Пусть в вершине m оказалась выборка X_m
- $Q(X_m, j, t)$ — критерий ошибки условия $[x^j \leq t]$
- Ищем лучшие параметры j и t перебором:

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

- 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:

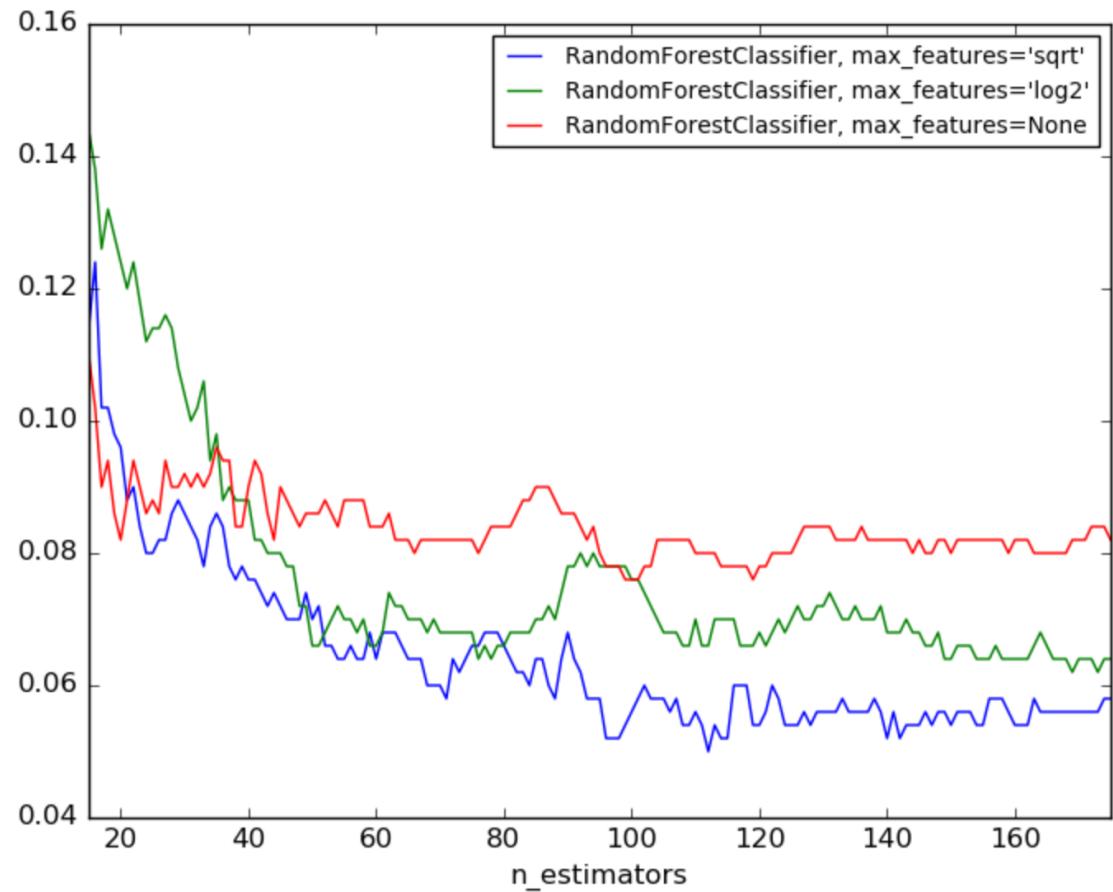
$$a(x) = \frac{1}{N} \sum_{n=1}^N b_n(x)$$

- Классификация:

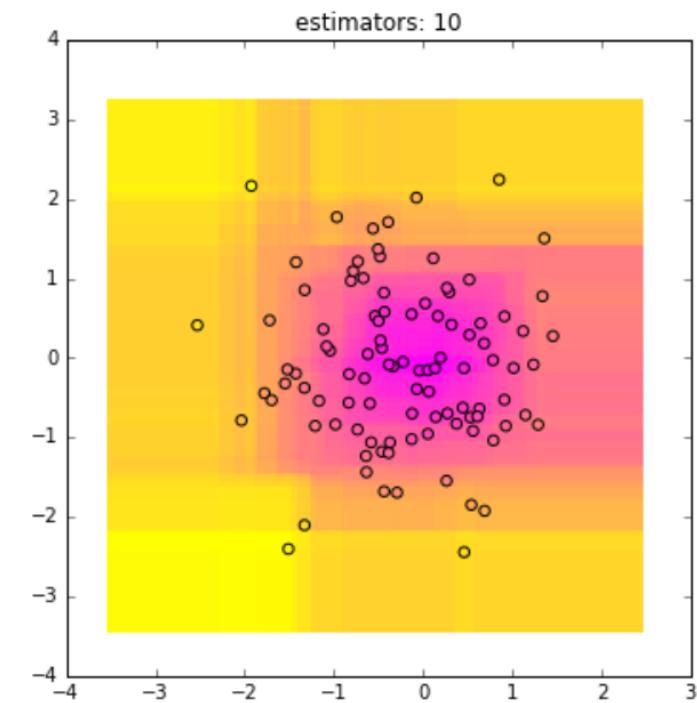
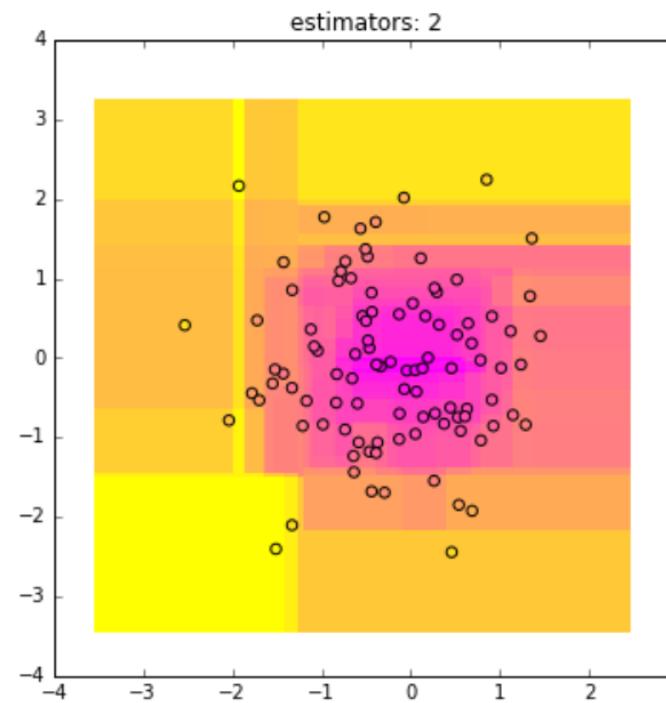
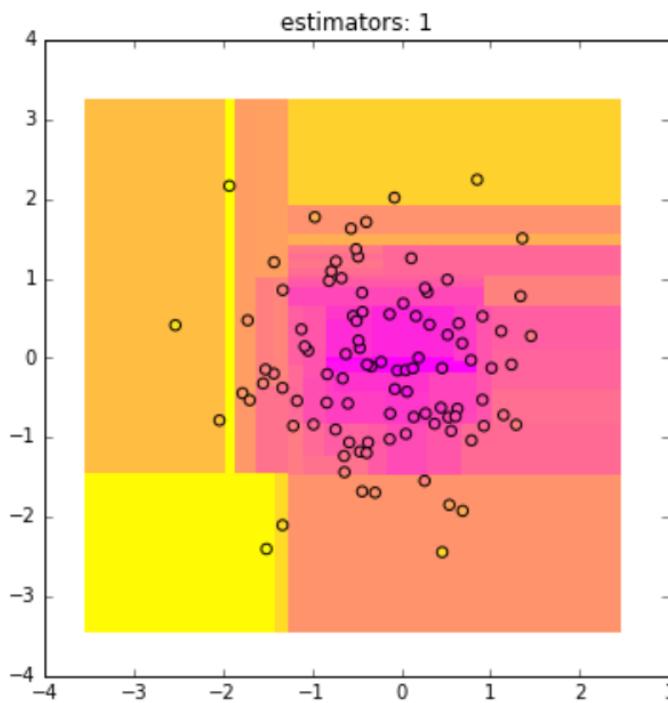
$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{n=1}^N [b_n(x) = y]$$

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

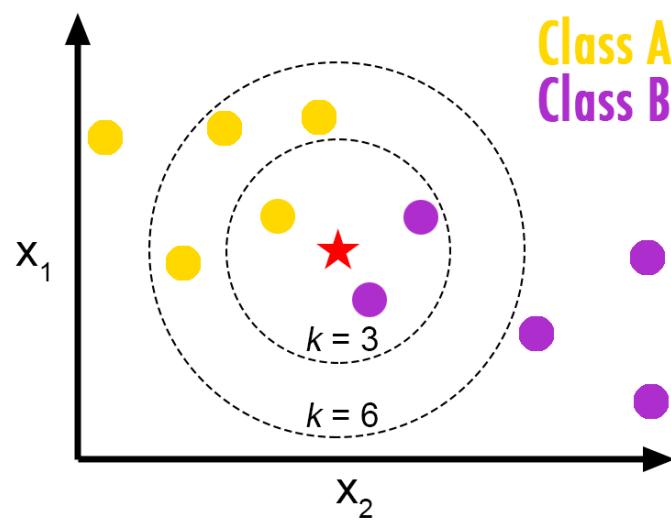
Обучение с учителем
(заключение)

Обучение с учителем

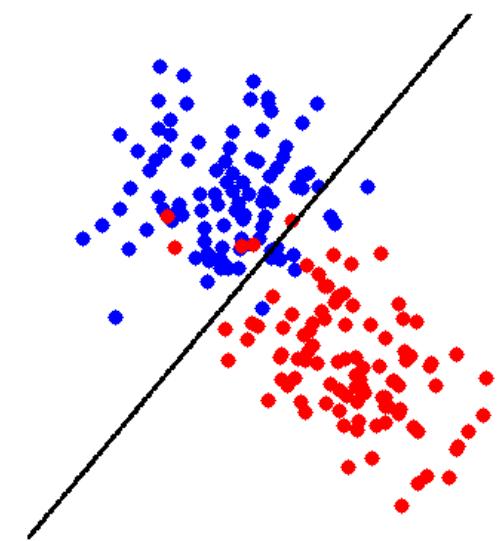
- В нашем курсе: классификация или регрессия
- $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка
- $a(x)$ — алгоритм, модель
- «С учителем» — т.е. на обучающей выборке известны ответы y_i

Обучение с учителем

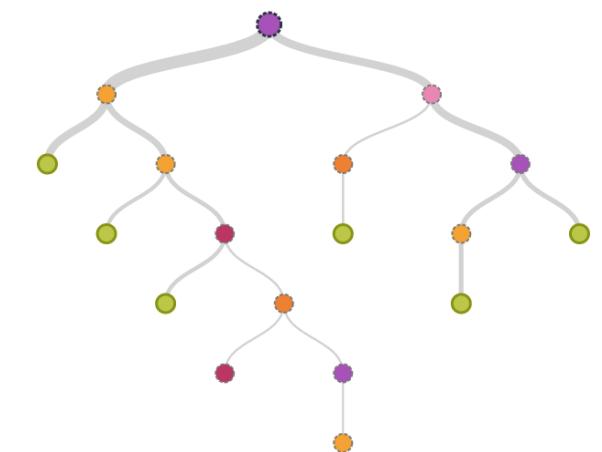
Метод k ближайших
соседей



Линейные модели

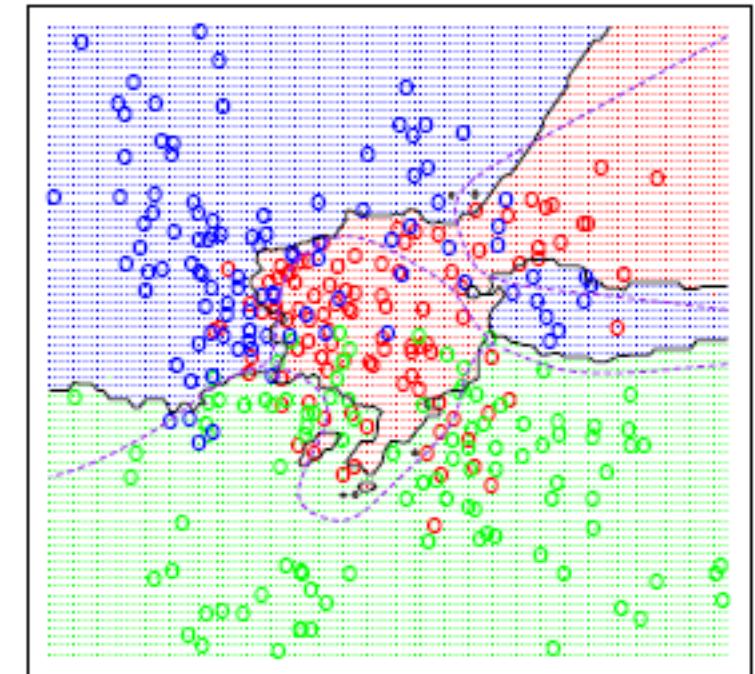


Решающие деревья



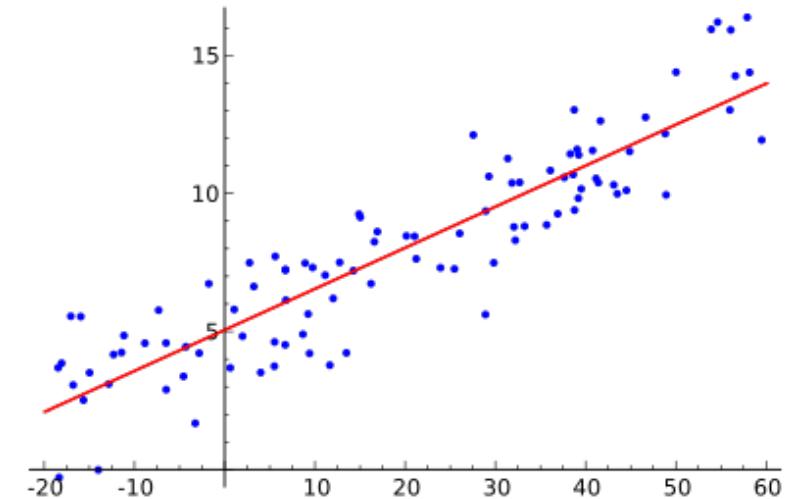
Метод k ближайших соседей

- (+) Очень мало параметров
- (+) Может восстанавливать сложные закономерности
- (-) Нередко показывает плохое качество
- (-) Приличных результатов можно добиться при подборе метрики



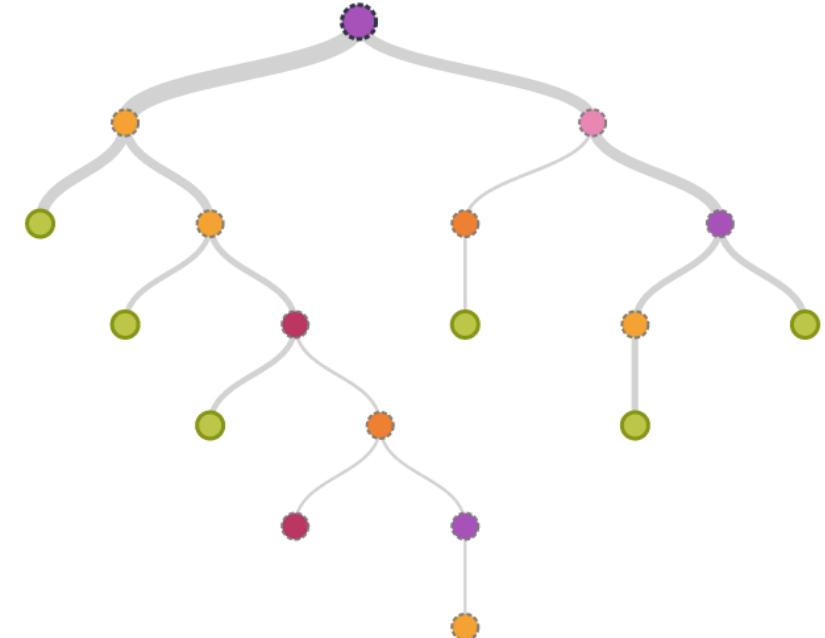
Линейные модели

- (+) Легко контролировать переобучение (регуляризация)
- (+) Быстро обучаются даже на огромных объемах данных
- (+) Хорошо работают при большом числе признаков (например, на категориальных признаках)
- (-) Восстанавливают всего лишь линейные закономерности



Решающие деревья

- (+) Могут дать нулевую ошибку на любой обучающей выборке
- (+) Можно интерпретировать
- (-) Очень легко переобучаются
- (+) Хорошо объединяются в композиции



Обучение с учителем

- Мы изучили основные типы моделей
- Измерять качество в регрессии и классификации тоже научились
- Важные этапы — подготовка данных (откуда их взять?) и разработка признаков
- Пример задачи: автоответ на письма
 - по мотивам статьи «Smart Reply: Automated Response Suggestion for Email», KDD 2016

Автоответ на письма

Сейчас 25% писем-ответов содержат меньше 20 токенов

Требования к автоответу:

- Высокое качество с точки зрения языка и смысла
- Разнообразие
 - Показывать несколько разных вариантов
 - «Yes, I will be there» и «I'll be there»
- Сохранение приватности переписки пользователей

Задачи машинного обучения

Триггеринг:

- Понять, нужен ли для данного письма автоответ
- Письма со сложным вопросом
 - «Where do you want to go today?»
- Письма, на которые ответ не нужен вообще

Задачи машинного обучения

Выбор наиболее подходящих ответов:

- Классификация на K классов
- K — число допустимых ответов

Задачи не про машинное обучение

- Как собрать обучающую выборку для задачи триггеринга?
- Какие автоответы являются допустимыми?
- На каких признаках обучать классификаторы?
- Как добиться разнообразия ответов?

Триггеринг

Данные:

- Положительные примеры — письма, на которые ответили с мобильного устройства
- Отрицательные примеры — письма, не которые не ответили вообще
- 238 миллионов объектов

Допустимые ответы

Как не надо:

- Your the best!
- Thanks hon
- Yup
- Got it thx
- Leave me alone

Допустимые ответы

Как не надо (все предлагаемые ответы — одинаковые по смыслу):

Yes, I'll be there.

Yes, I will be there.

I'll be there.

Yes, I can.

What time?

I'll be there!

I will be there.

Sure, I'll be there.

Yes, I can be there.

Yes!

Допустимые ответы

- Взяли несколько миллионов наиболее частых ответов пользователей
- Кластеризовали их
- Выбрали из каждого кластера пять представителей и проверили допустимость силами асессоров

Разнообразие ответов

- Из каждого кластера выбирается представитель с максимальной оценкой вероятности от классификатора
- Есть смещение в сторону позитивных ответов
- Если топ-3 кандидатов позитивные, то третий заменяется на наиболее вероятный негативный ответ

Задачи не про машинное обучение

- Как собрать обучающую выборку для задачи триггеринга?
- Какие автоответы являются допустимыми?
- На каких признаках обучать классификаторы?
- Как добиться разнообразия ответов?