

Введение в анализ данных

Лекция 11

Решающие деревья и случайные леса

Евгений Соколов

sokolov.evg@gmail.com

НИУ ВШЭ, 2016

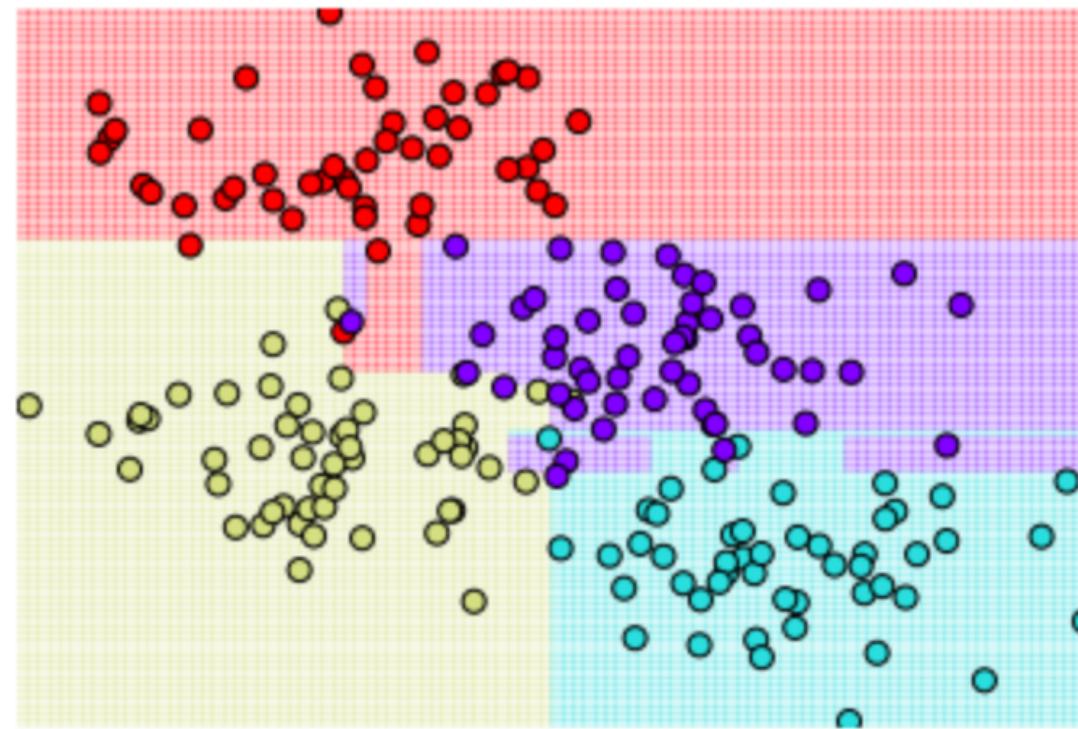


Обучение решающих деревьев

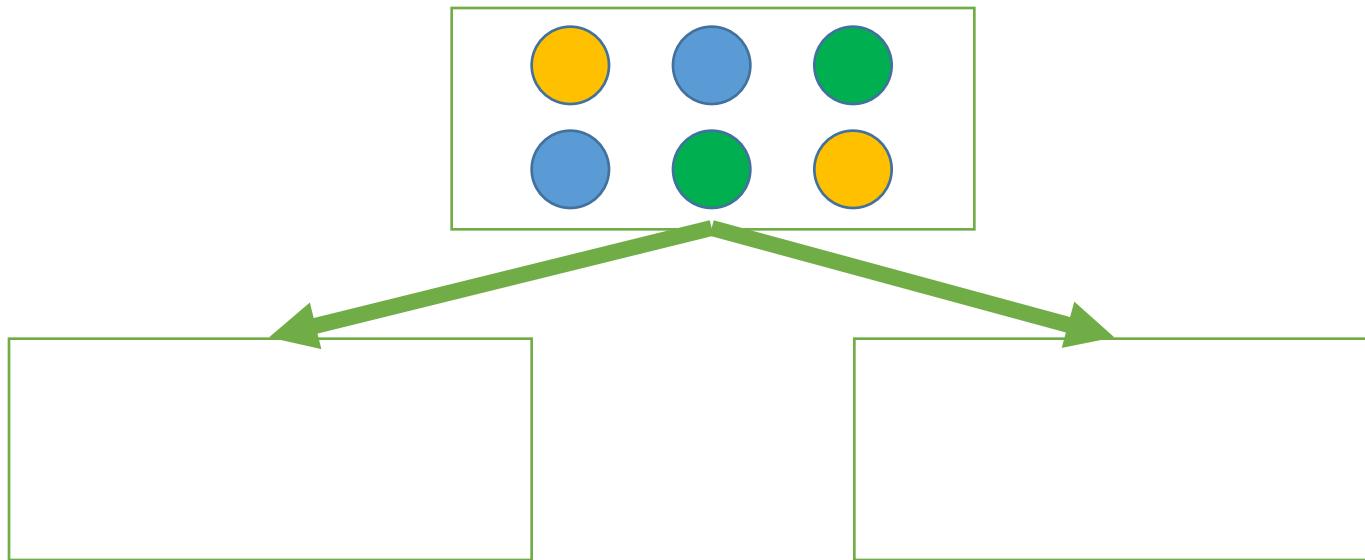
Обучение деревьев

- Жадное построение от корня к листьям
- Разбиение в каждой вершине выбирается так, чтобы сгруппировать похожие объекты
- Условия в вершинах: $[x^j \leq t]$

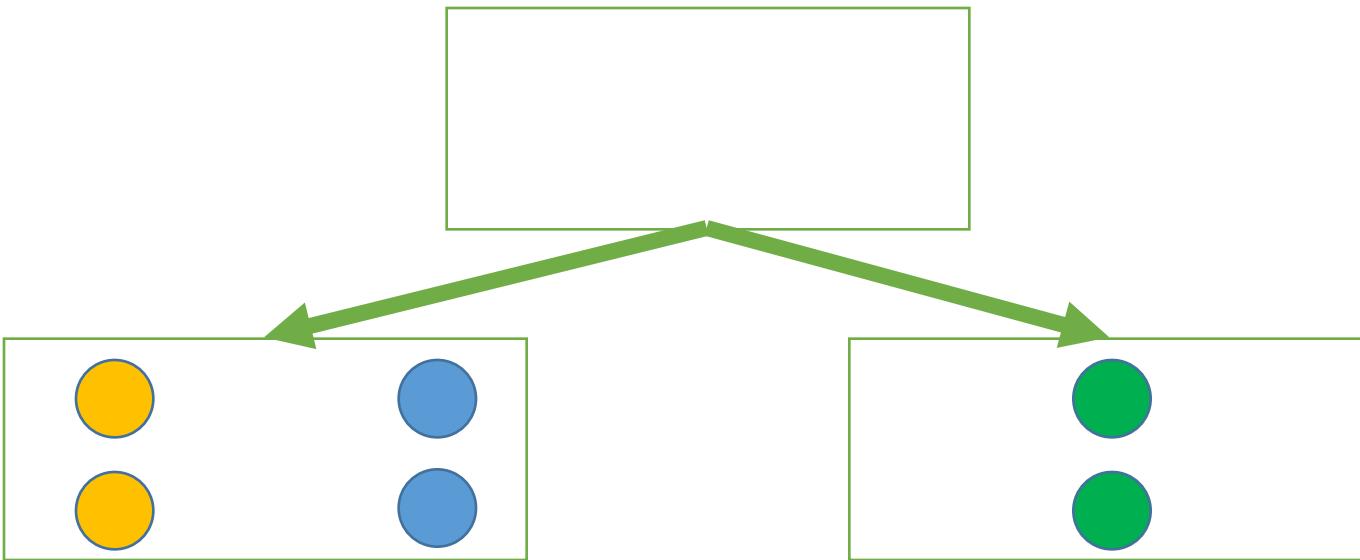
Обучение деревьев



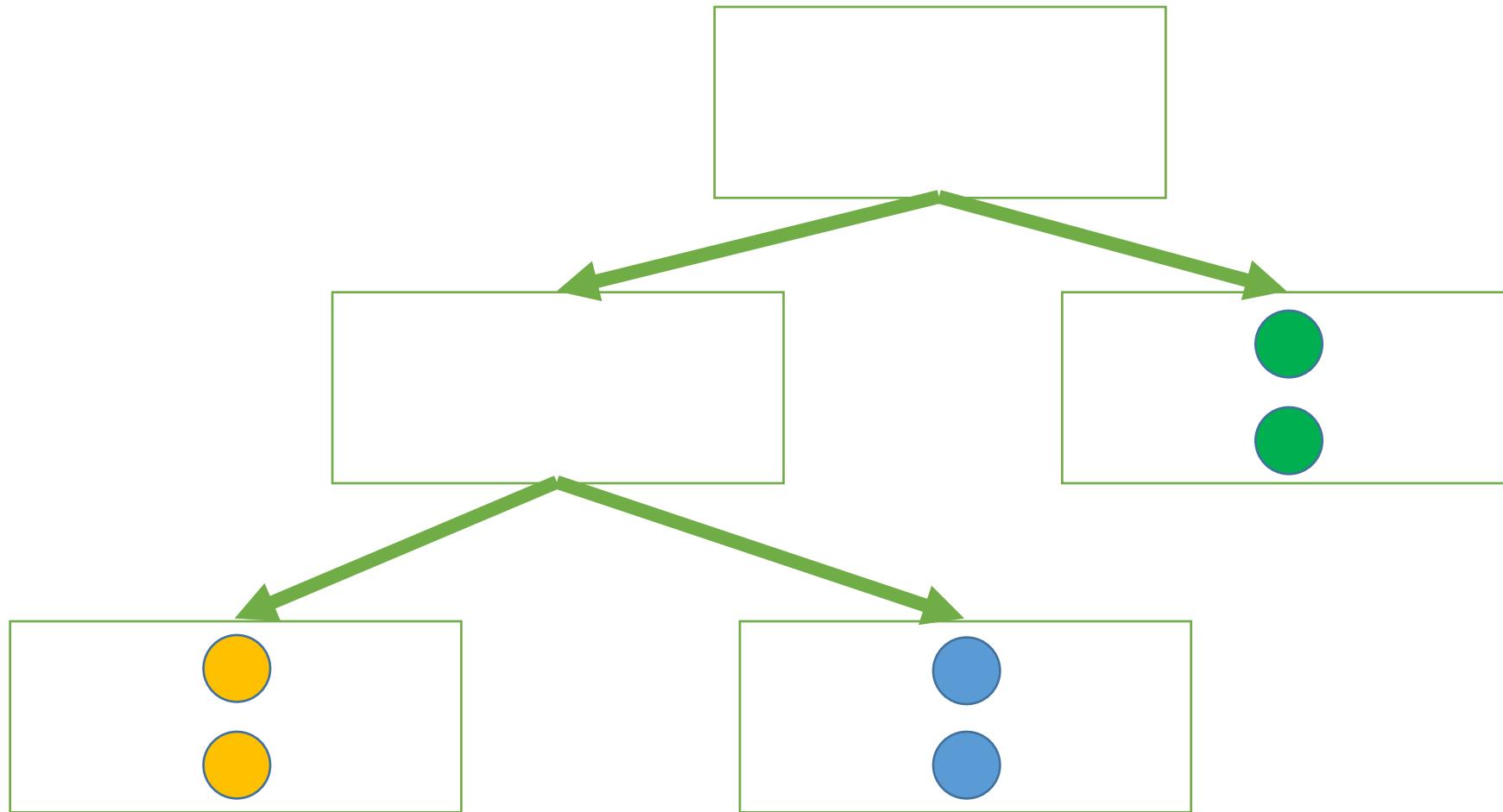
Жадное построение



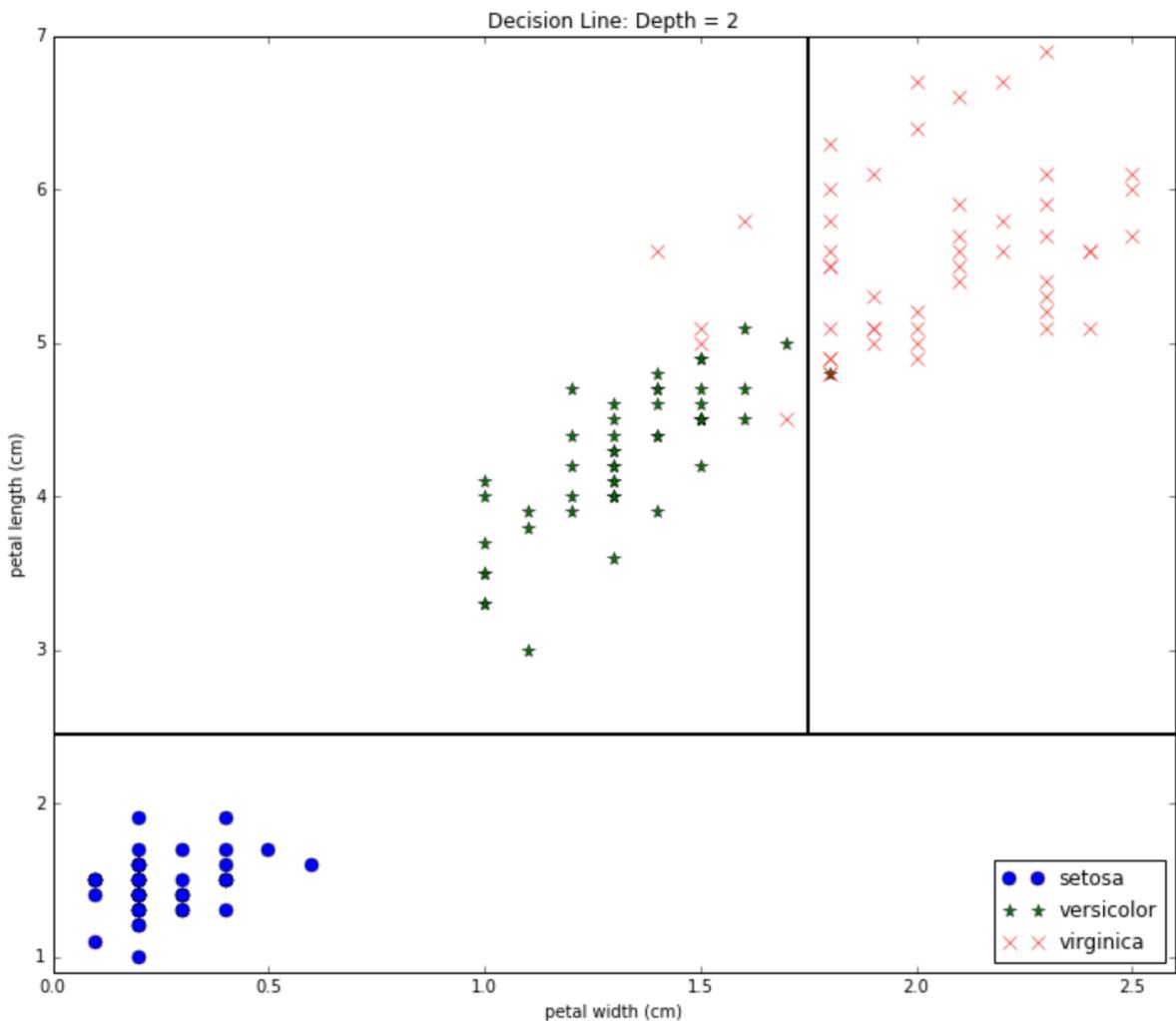
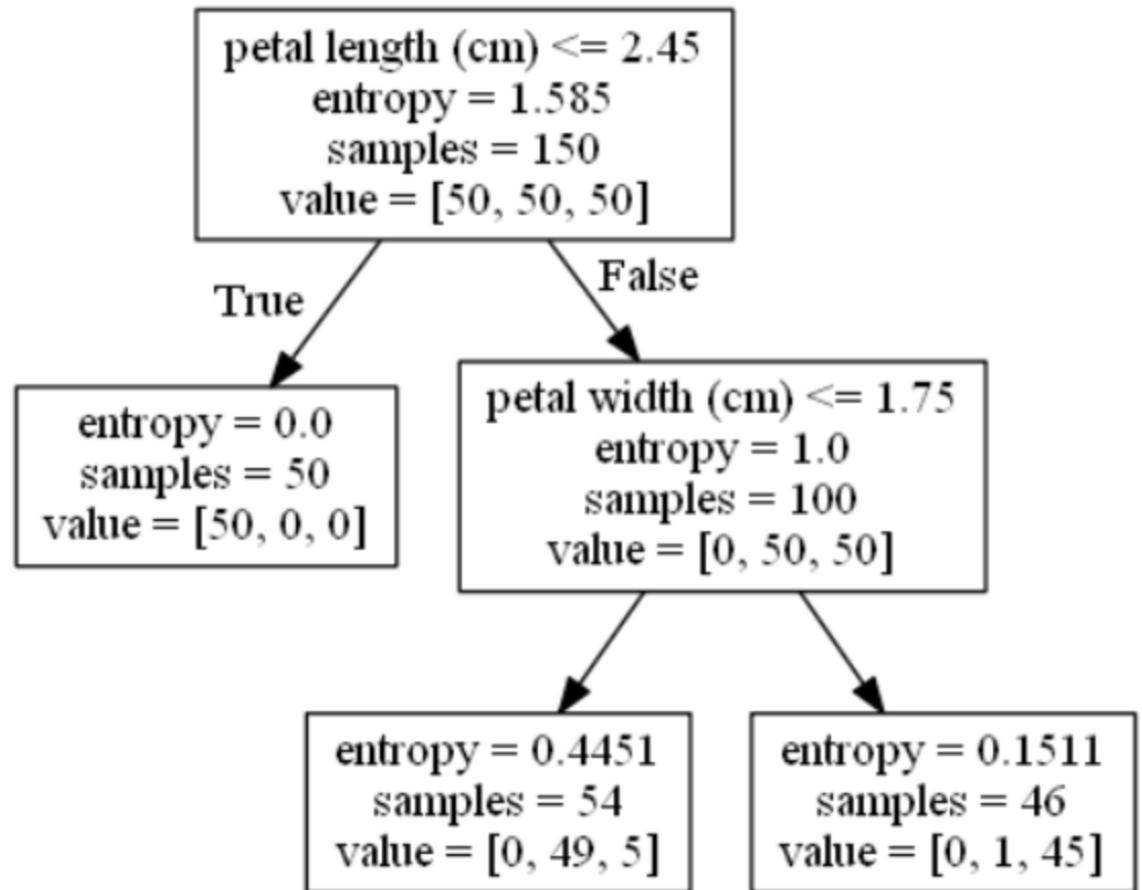
Жадное построение



Жадное построение



Решающее дерево



Жадный алгоритм построения дерева

1. Поместить в корень всю выборку: $X_1 = X$
2. Начать построение с корня: $m = 1$
3. Если выполнен критерий останова для вершины m , то выход
4. Найти лучшее разбиение $[x^j \leq t]$ для вершины m
5. Разбить вершину m на дочерние вершины l и r
6. Повторить шаги 3-6 для дочерних вершин l и r

Поиск разбиения

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

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

Критерий качества

$$Q(X_m, j, t) = H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r)$$

Разброс ответов в левом
листе

Разброс ответов в правом
листе

Критерий информативности

- $H(X)$
- Зависит от ответов на выборке X
- Чем меньше разброс ответов, тем меньше значение $H(X)$

Регрессия

$$\bar{y}(X) = \frac{1}{|X|} \sum_{i \in X} y_i$$

$$H(X) = \frac{1}{|X|} \sum_{i \in X} (y_i - \bar{y}(X))^2$$

Классификация

- Доля объектов класса k в выборке X :

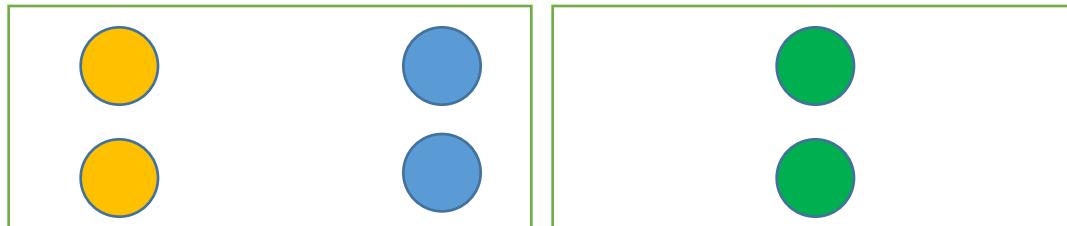
$$p_k = \frac{1}{|X|} \sum_{i \in X} [y_i = k]$$

Энтропийный критерий

$$H(X) = - \sum_{k=1}^K p_k \ln p_k$$

- Считаем, что $0 \ln 0 = 0$
- Если $p_1 = 1, p_2 = \dots = p_K = 0$, то $H(X) = 0$
- Мера отличия распределения классов от вырожденного

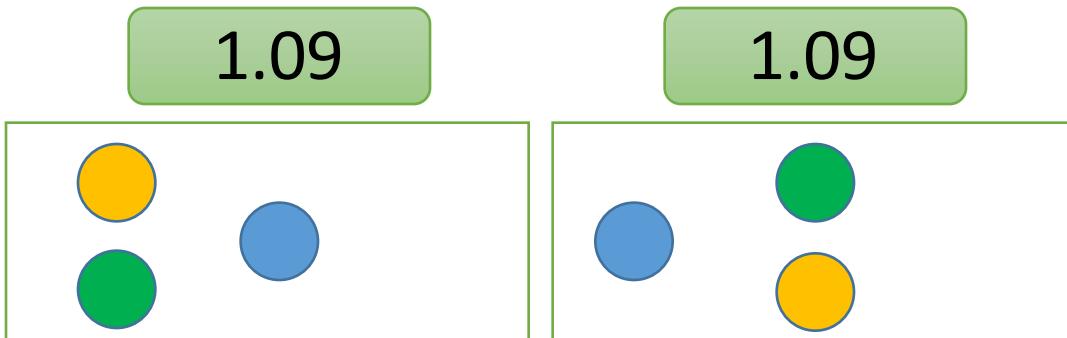
Энтропийный критерий



0.693

0

- $(0.5, 0.5, 0)$ и $(0, 0, 1)$



1.09

1.09

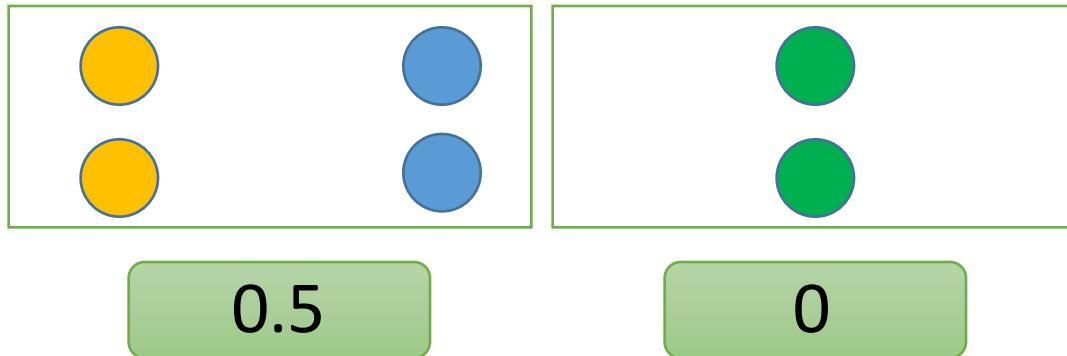
- $(0.33, 0.33, 0.33)$ и $(0.33, 0.33, 0.33)$
- $H = 1.09 + 1.09 = 2.18$

Критерий Джини

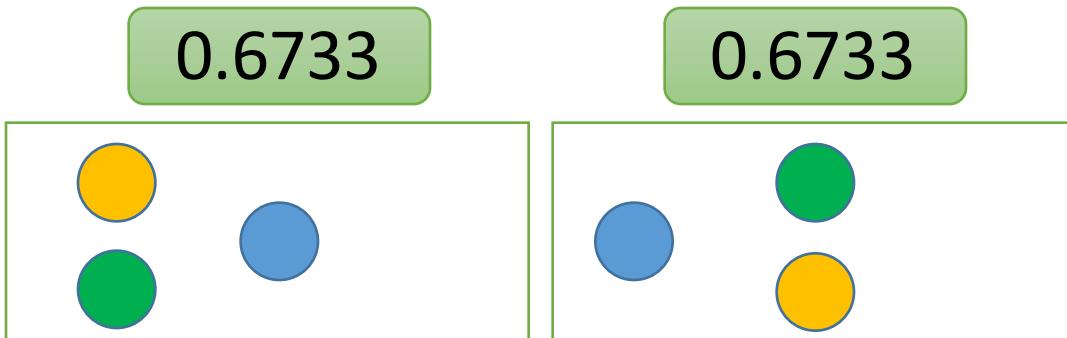
$$H(X) = \sum_{k=1}^K p_k(1 - p_k)$$

- Если $p_1 = 1, p_2 = \dots = p_K = 0$, то $H(X) = 0$
- Вероятность ошибки классификатора, который выдаёт ответы пропорционально p_k

Критерий Джини



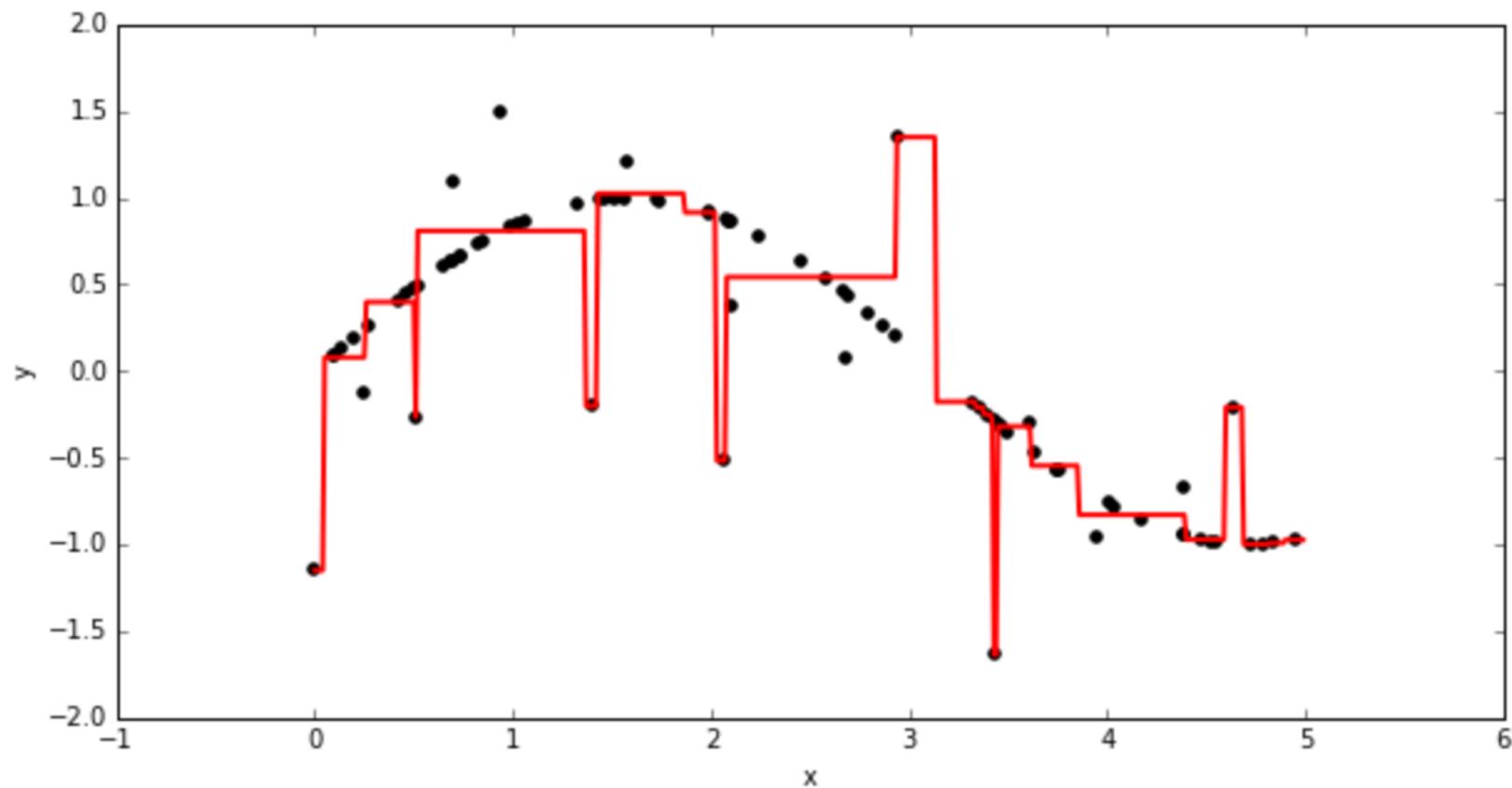
- $(0.5, 0.5, 0)$ и $(0, 0, 1)$



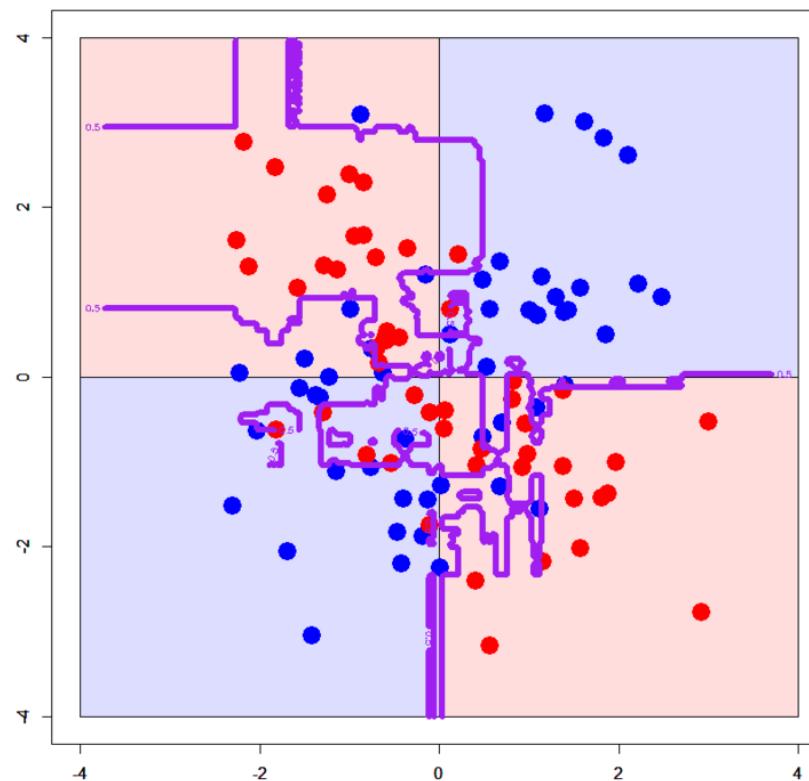
- $(0.33, 0.33, 0.33)$ и $(0.33, 0.33, 0.33)$

Переобучение деревьев и
борьба с ним

Переобучение деревьев



Переобучение деревьев

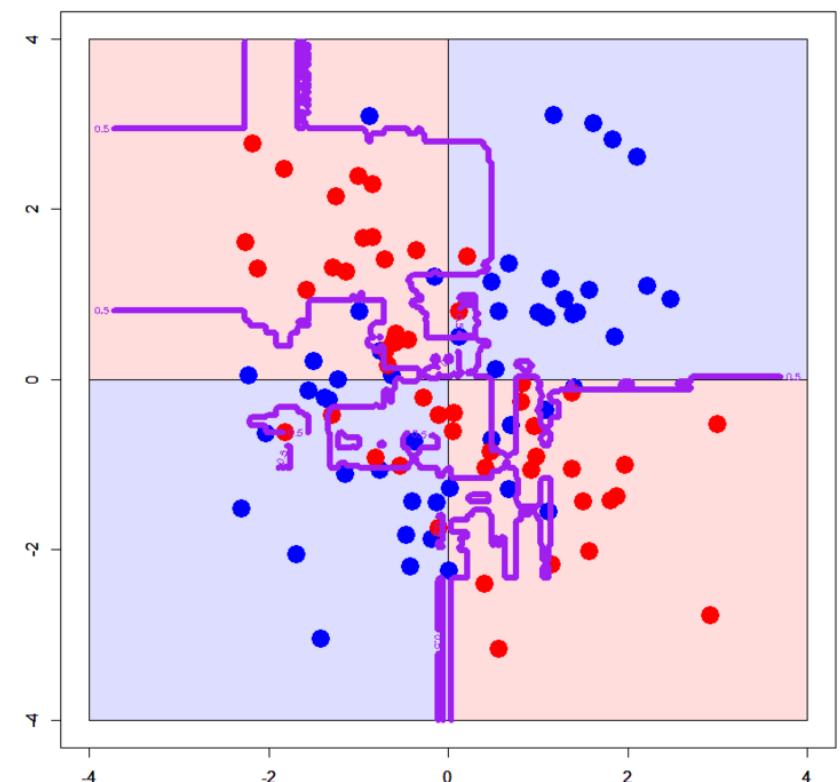


Критерий останова

- Как понять, разбивать вершину или делать листовой?
- Способ борьбы с переобучением

Критерий останова

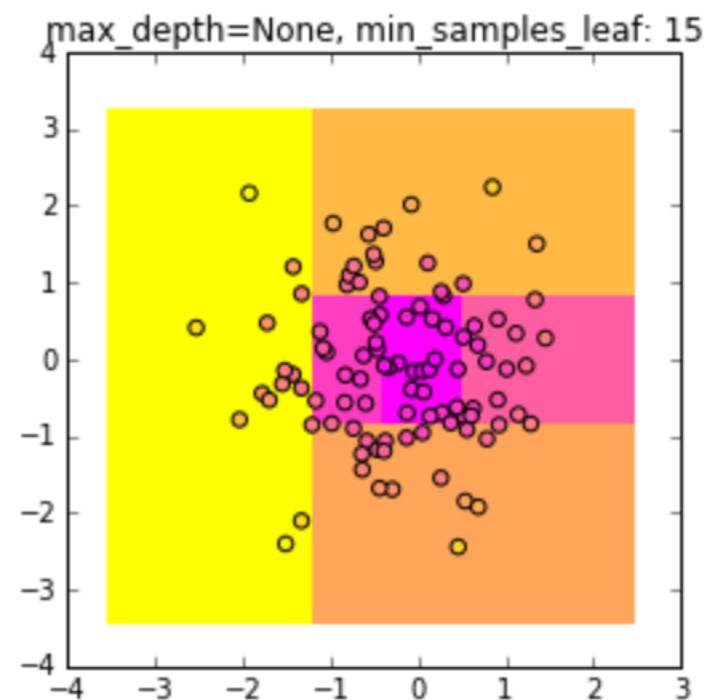
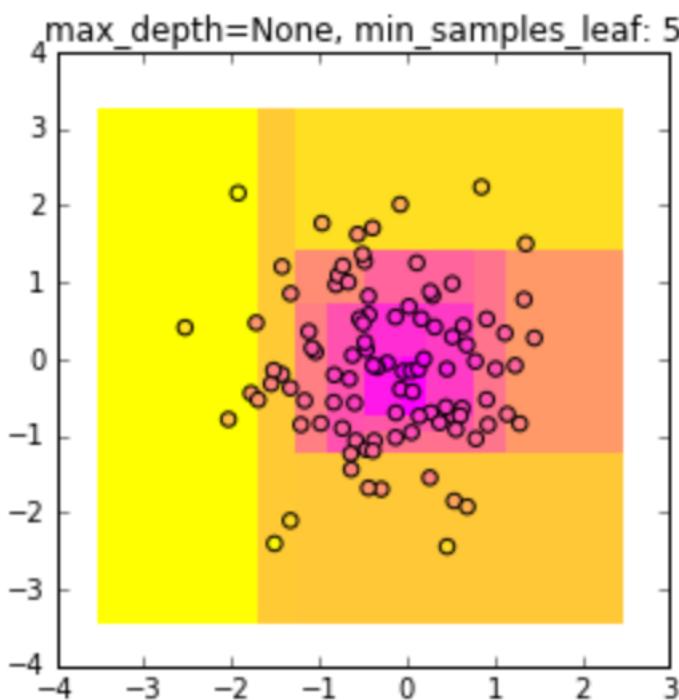
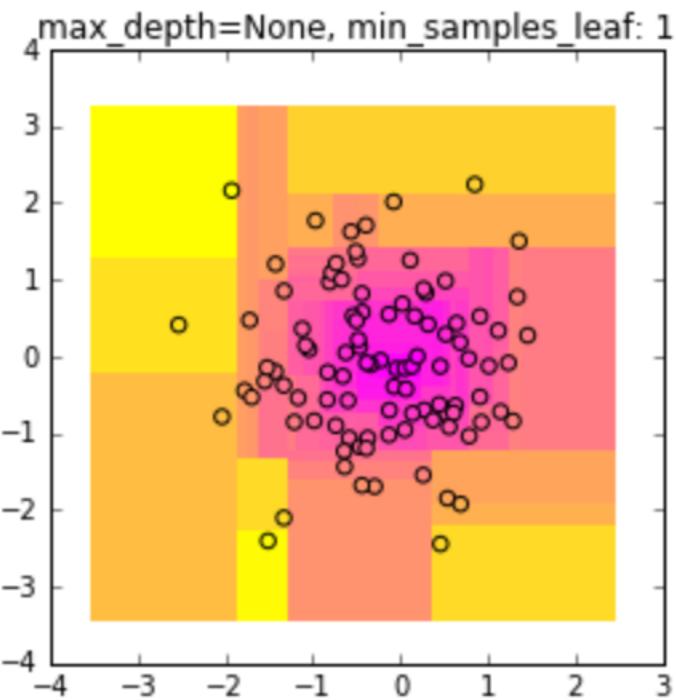
- Все объекты в вершине относятся к одному классу
- Простое условие
- Но приводит к переобучению



Число объектов в листе

- В вершину попало $\leq n$ объектов
- При $n = 1$ получаем максимально переобученные деревья
- n должно быть достаточно, чтобы построить надёжный прогноз
- Рекомендация: $n = 5$

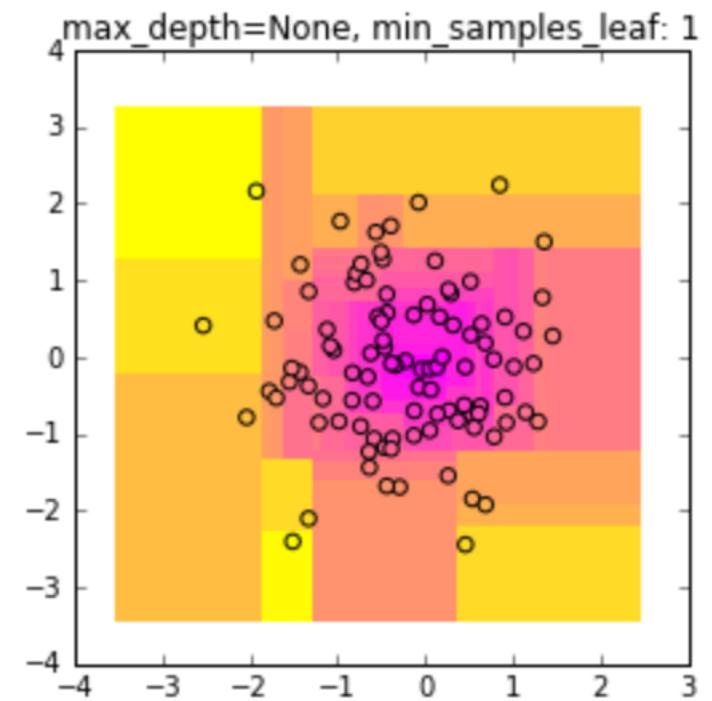
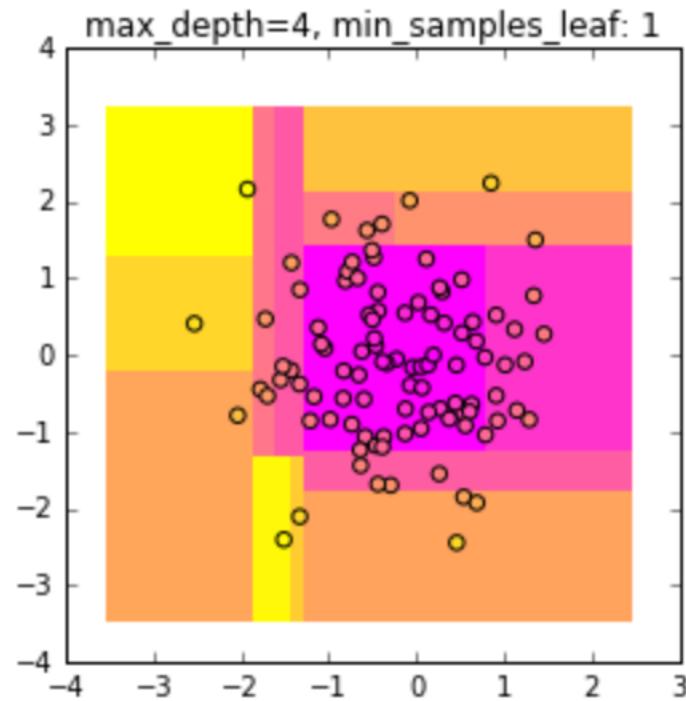
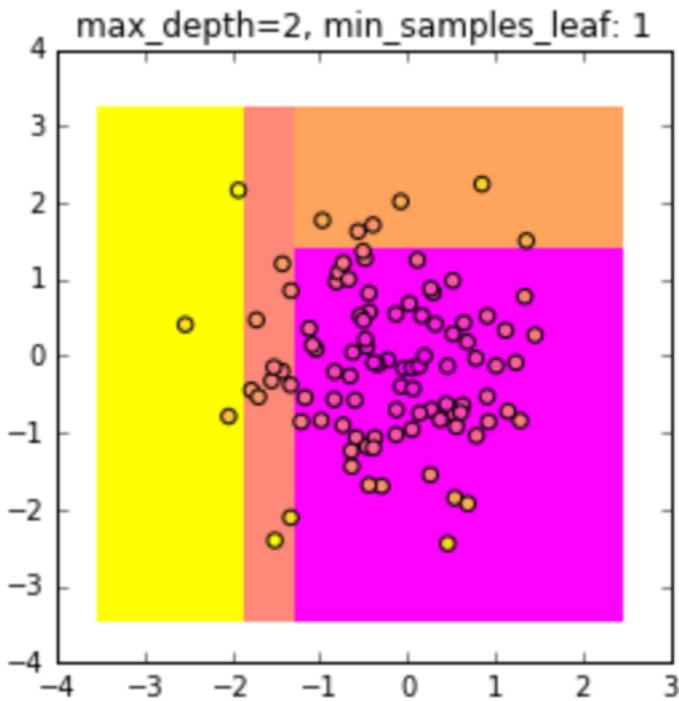
Число объектов в листе



Глубина дерева

- Ограничение на глубину
- Достаточно грубый критерий

Глубина дерева



Стрижка деревьев

- Строим максимально переобученное дерево
- Удаляем листья по некоторому критерию
- Пример: удаляем, пока улучшается ошибка на валидации
- Считается, что работает лучше критериев останова

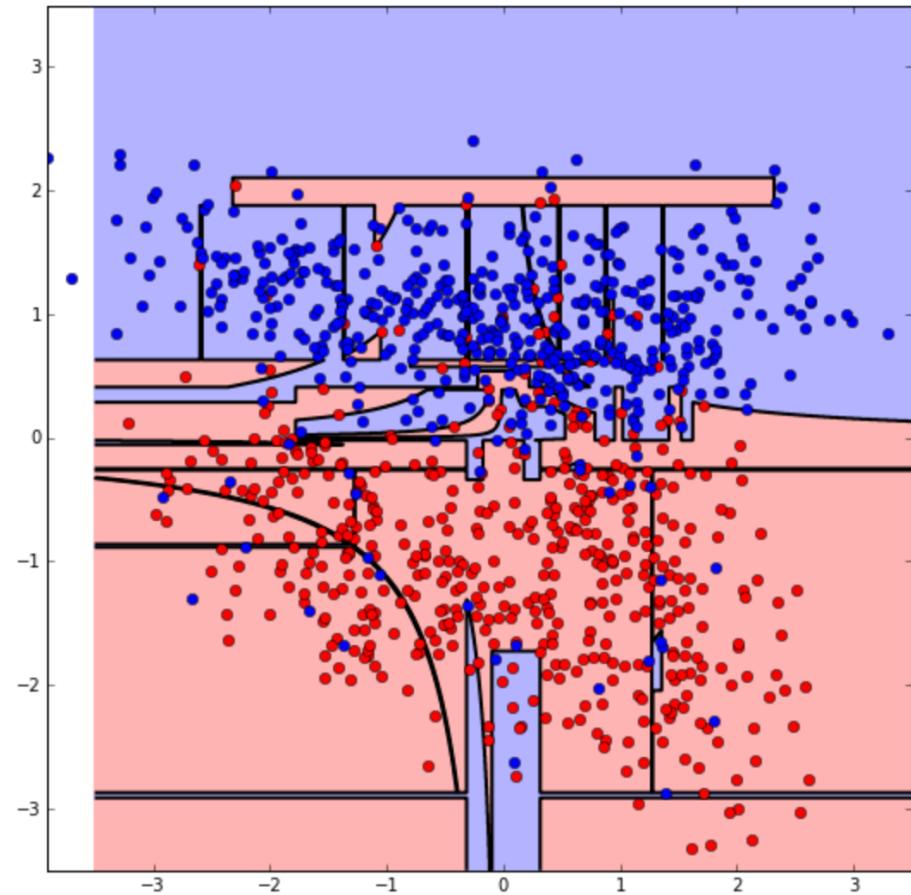
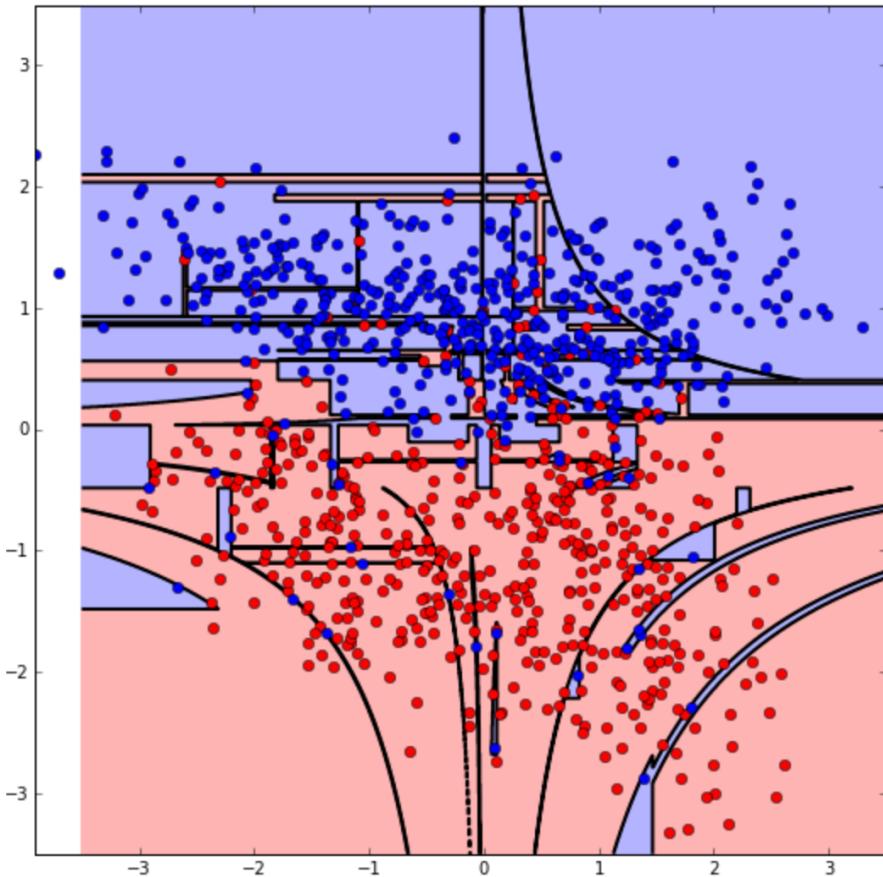
Стрижка деревьев

- Трудоёмкая процедура
- Имеет смысл только при использовании одного дерева
- В композициях деревьев достаточно простых критериев останова

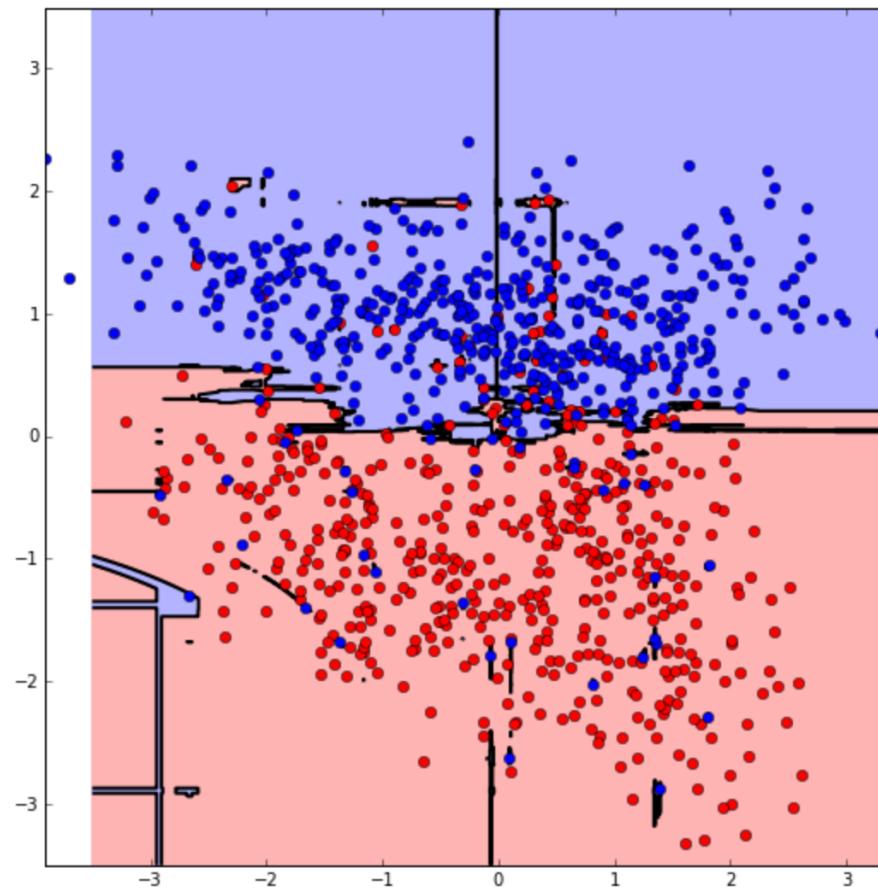
Неустойчивость деревьев

- Структура дерева очень сильно меняется даже при малом изменении выборки
- Пример: обучим два дерева по подвыборкам размером 90% от всего обучения

Неустойчивость деревьев



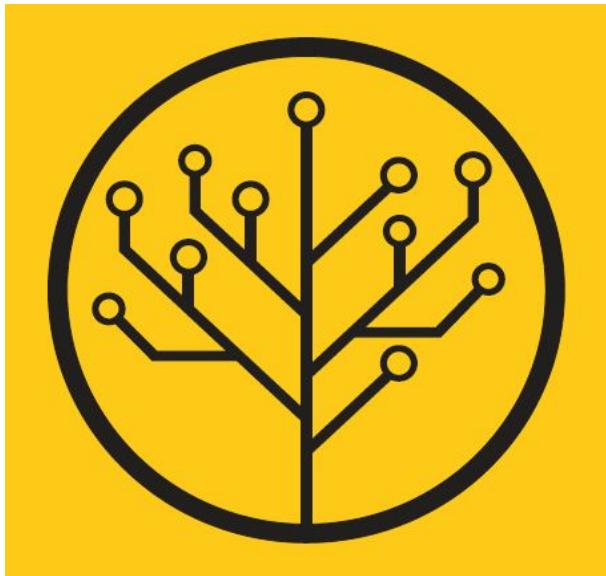
Усреднение деревьев



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 немножко лучше случайного угадывания
- Композиция:

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

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: $a(x)$

- Композиции — одна из наиболее мощных техник в современном машинном обучении

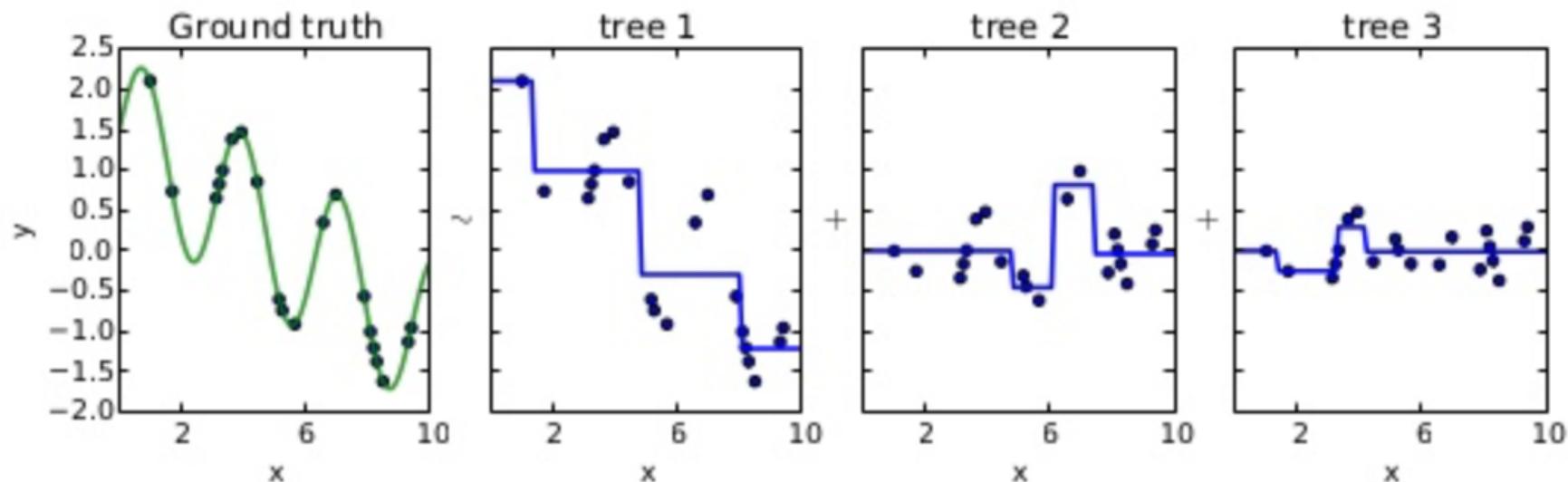
Композиции алгоритмов

- 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: $-1, -1, 1, -1, 1, -1$

$$a(x) = -1$$

Рандомизация

- Как сделать деревья разными?
- Обучать по подвыборкам!

Рандомизация

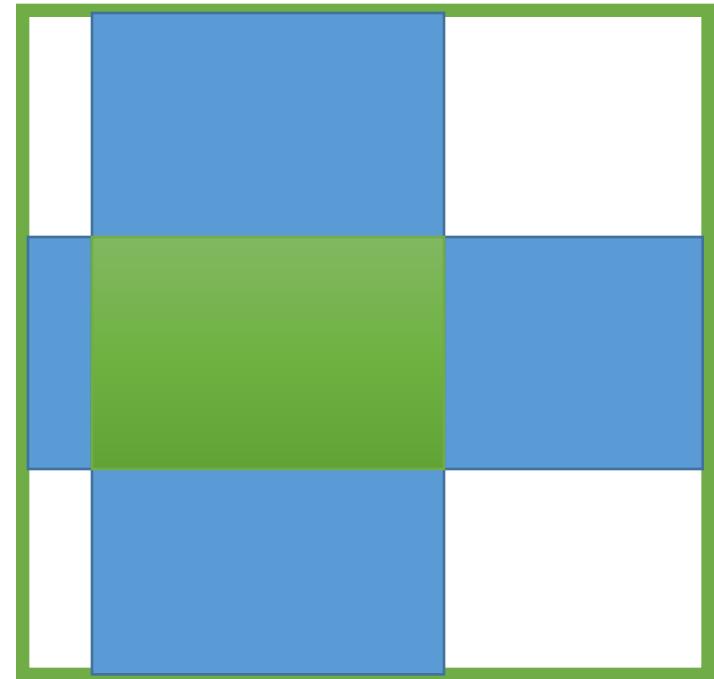
- Популярный подход: бутстрэп
- Выбираем из обучающей выборки ℓ объектов с возвращением
- Пример: $\{x_1, x_2, x_3, x_4\} \rightarrow \{x_1, x_2, x_2, x_4\}$
- Примерно $0.632 * \ell$ различных объектов

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

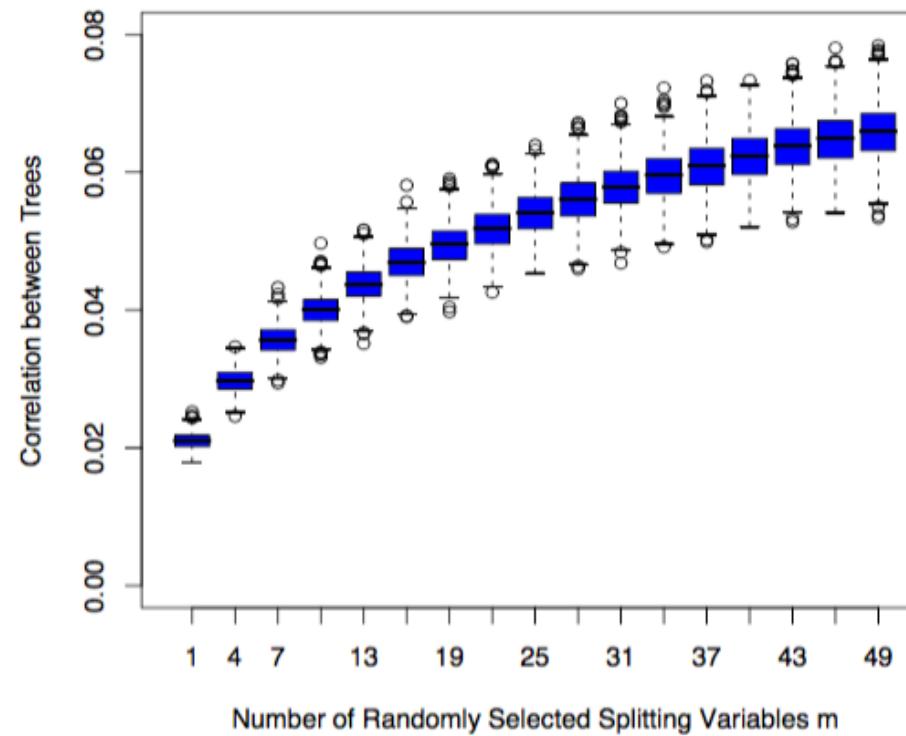
- Пусть в вершине m оказалась выборка X_m
- $Q(X_m, j, t)$ — критерий ошибки условия $[x^j \leq t]$
- Ищем лучшие параметры j и t перебором:

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

- 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

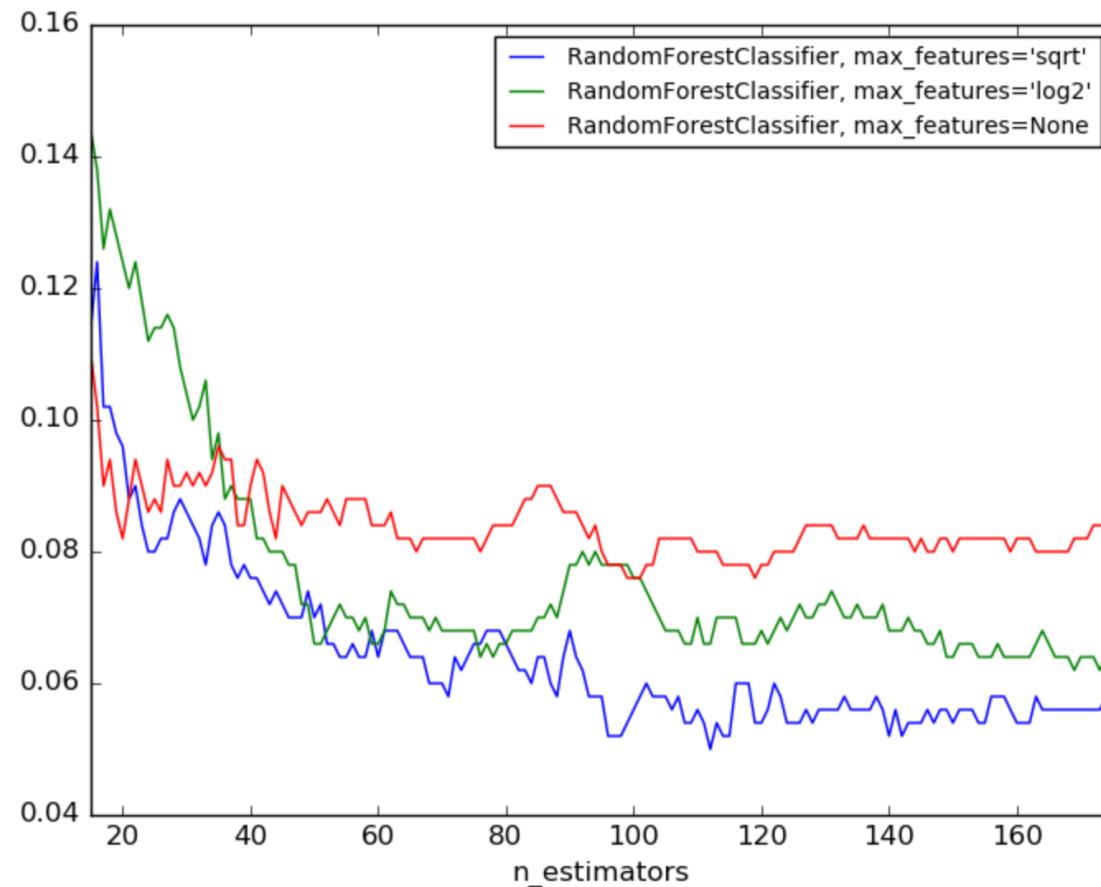
- Регрессия:

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

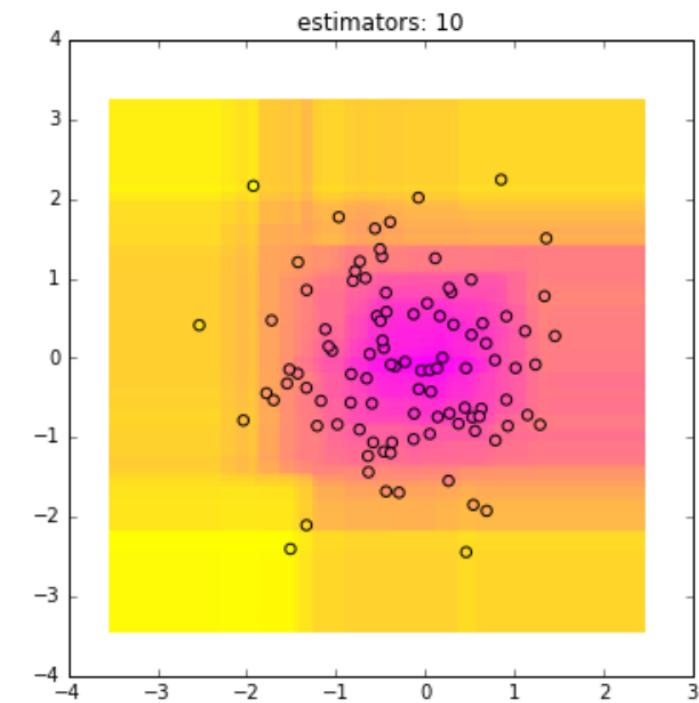
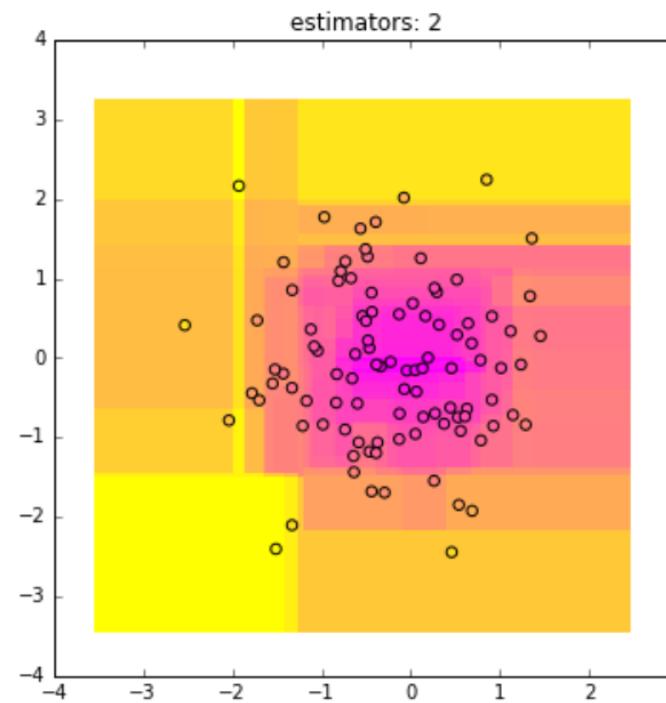
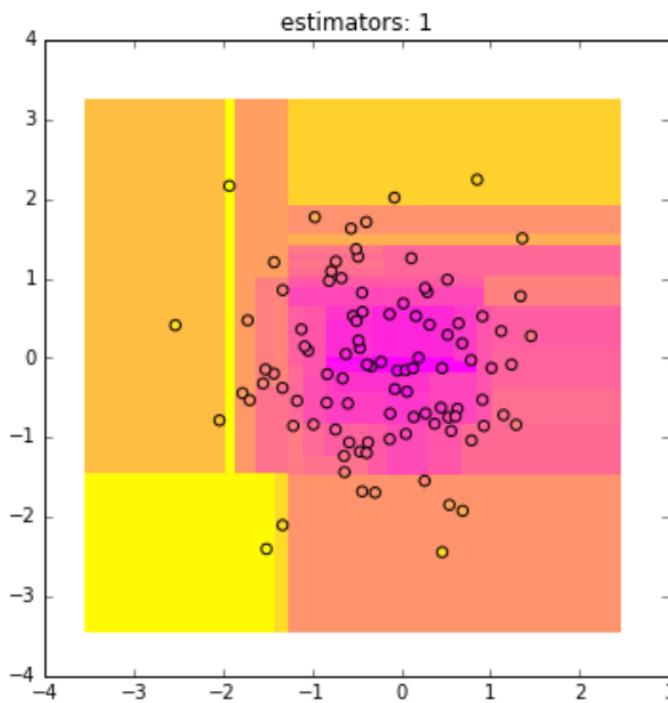
- Классификация:

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

Качество на тесте



Случайный лес



Резюме

- Деревья обучаются жадно
- Разбиения выбираются так, чтобы как можно сильнее уменьшить критерий информативности
- Борьба с переобучением: ограничение глубины или числа объектов в листьях
- Композиции алгоритмов
- Случайные леса

В следующий раз

- Отбор признаков и понижение размерности
- Метод главных компонент