

# Введение в анализ данных

Лекция 8

Решающие деревья

Евгений Соколов

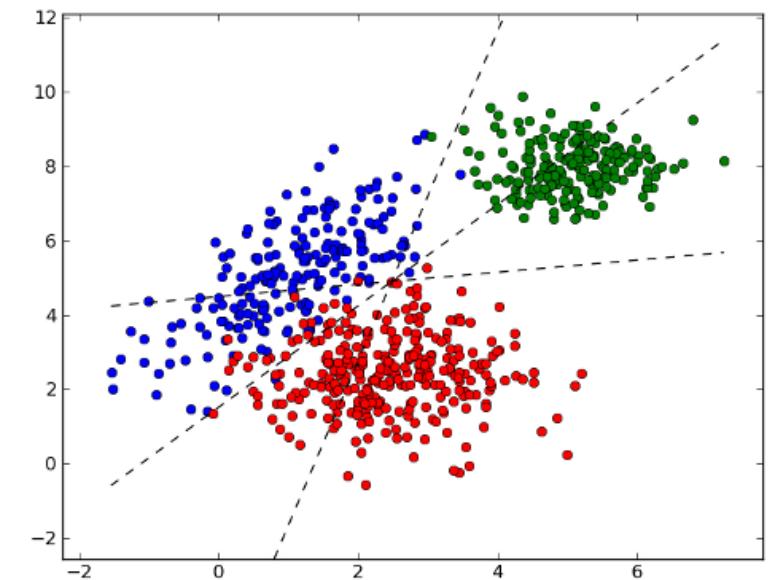
[esokolov@hse.ru](mailto:esokolov@hse.ru)

НИУ ВШЭ, 2018

# Многоклассовые задачи

# Многоклассовая классификация

- $\mathbb{Y} = \{1, 2, \dots, K\}$



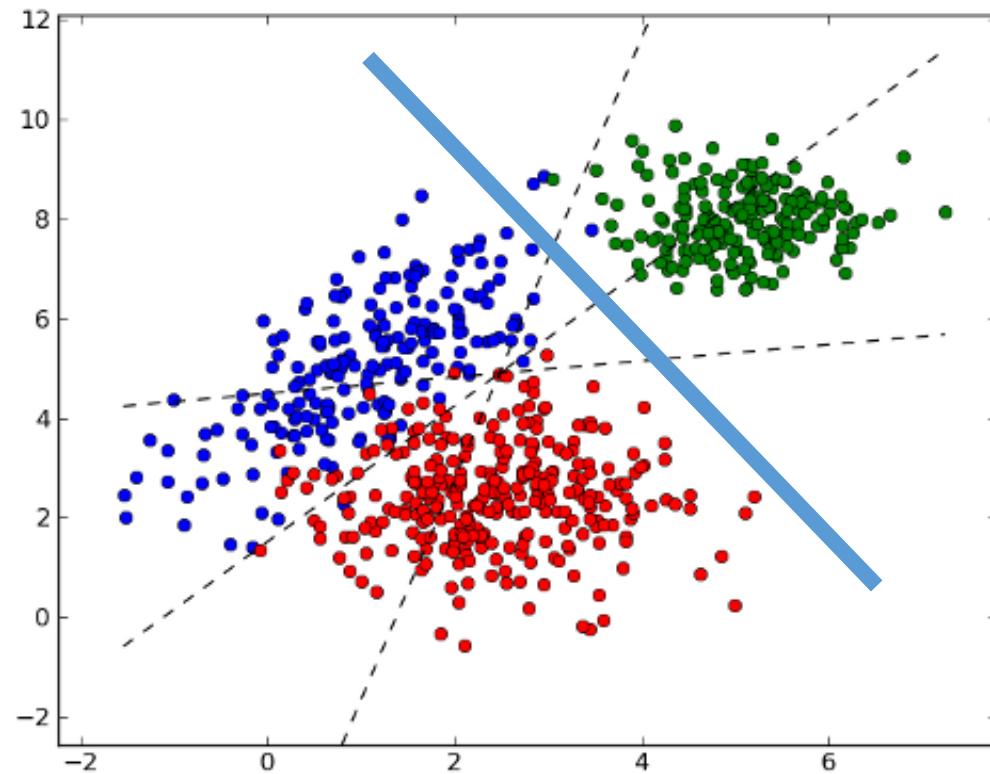
# Бинарная классификация

$$a(x) = \text{sign} \langle w, x \rangle$$

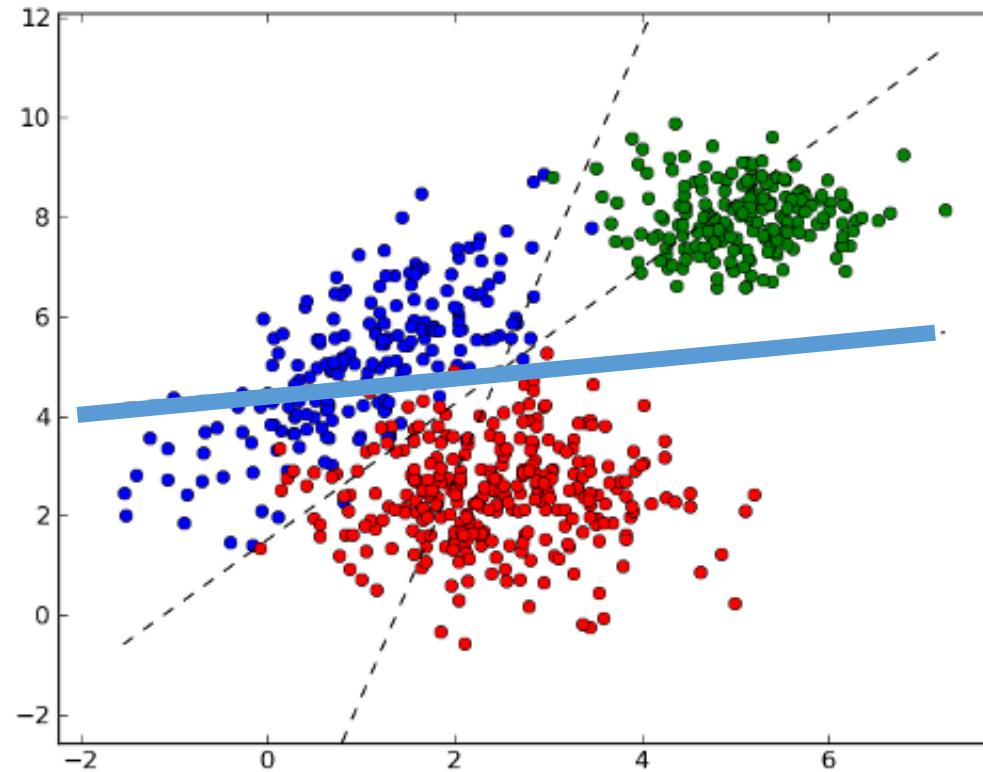
# One-vs-all

- Способ сведения многоклассовой задачи к набору бинарных классификаций
- Обучаем свой классификатор для каждого класса
- Задача: отделение класса от всех остальных

# One-vs-all



# One-vs-all



# One-vs-all

- $K$  задач бинарной классификации

- $k$ -я задача:

- $X = (x_i, [y_i = k])_{i=1}^\ell$

- Классификатор  $a_k(x) = \text{sign} \langle w_k, x \rangle$

- Алгоритм:

$$a(x) = \arg \max_{k \in \{1, \dots, K\}} \langle w_k, x \rangle$$

# Матрица ошибок

	$y = 1$	$y = 2$	...	$y = K$
$a(x) = 1$	$q_{11}$	$q_{12}$	...	$q_{1K}$
$a(x) = 2$	$q_{21}$	$q_{22}$	...	$q_{2K}$
...	...	...	...	...
$a(x) = K$	$q_{K1}$	$q_{K2}$	...	$q_{KK}$

Доля правильных ответов

$$\text{accuracy}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

# Точность и полнота

- Относительно каждого класса
- Можно усреднить точность и полноту по всем классам
- Можно усреднить F-меру

Решающие деревья

# Линейные модели

$$a(x) = w_0 + \sum_{j=1}^d w_j x^j$$

- Веса можно интерпретировать, если признаки масштабированы

# Пример

- Предсказание стоимости квартиры
- Признаки: площадь, этаж, число комнат

$$a(x) = 10 * (\text{площадь}) + 1.1 * (\text{этаж}) + 20 * (\text{число комнат})$$

# Пример

- Зависимость от этажа вряд ли линейная
- Квадратичные признаки:

$$a(x) = 10 * (\text{площадь}) + 1.1 * (\text{этаж}) + 20 * (\text{число комнат}) - 0.2 * (\text{этаж})^2 + 0.5 * (\text{площадь} * \text{число комнат}) + \dots$$

# Пример

- С кубическими признаками будет ещё лучше
- Как интерпретировать признак этаж \* (число комнат)<sup>2</sup>?
- Всего таких признаков 20

# Пример

- Можно бинаризовать признаки:  $[x^j > t]$
- (этаж > 1), (этаж > 2), ..., (этаж > 30)
- Признаков будет на порядки больше
- Легче интерпретировать:  
– 2[этаж > 3][площадь < 40][число комнат < 3]
- Можно использовать  $L_1$ -регуляризацию

# Логические правила

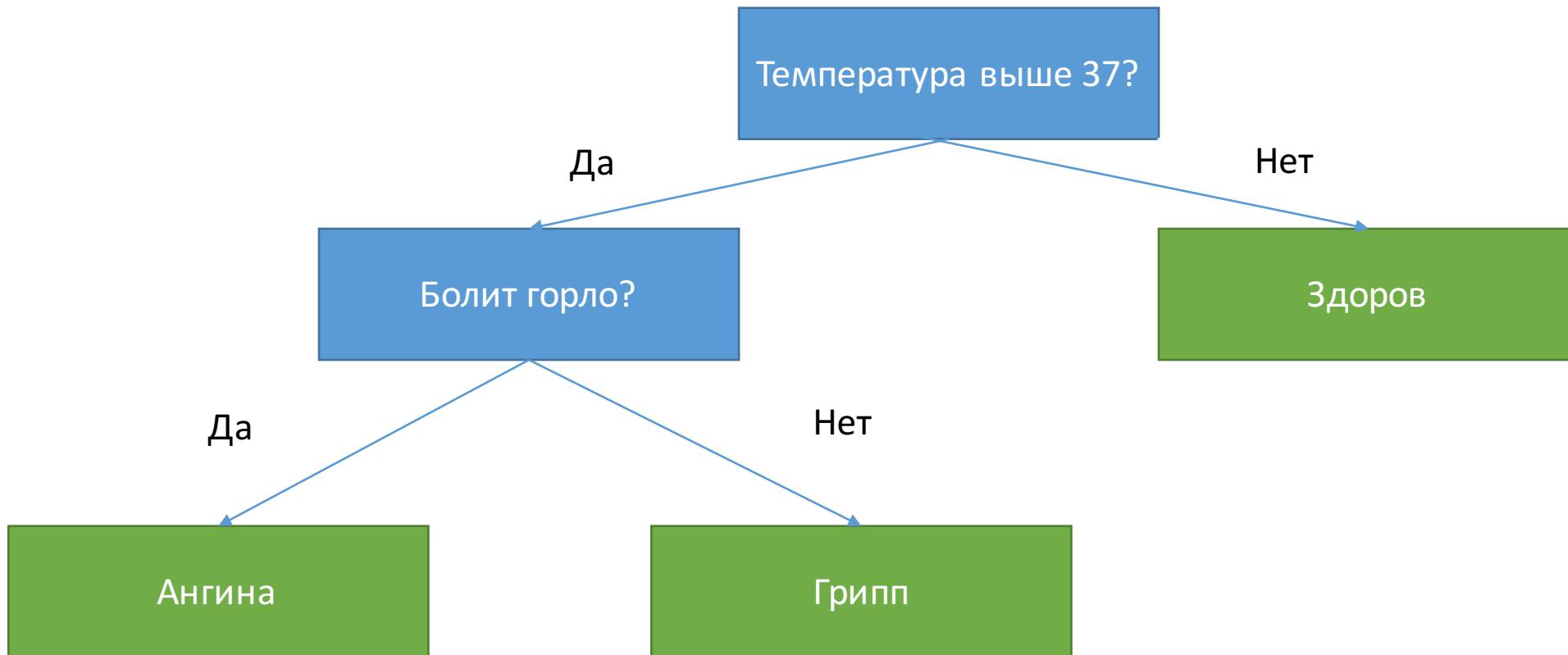
[этаж > 3][площадь < 40][число комнат < 3]

- Легко объяснить заказчику (если  $\leq 5$  условий)
- Позволяют извлекать знания из данных
- Не факт, что оптимальны с точки зрения качества

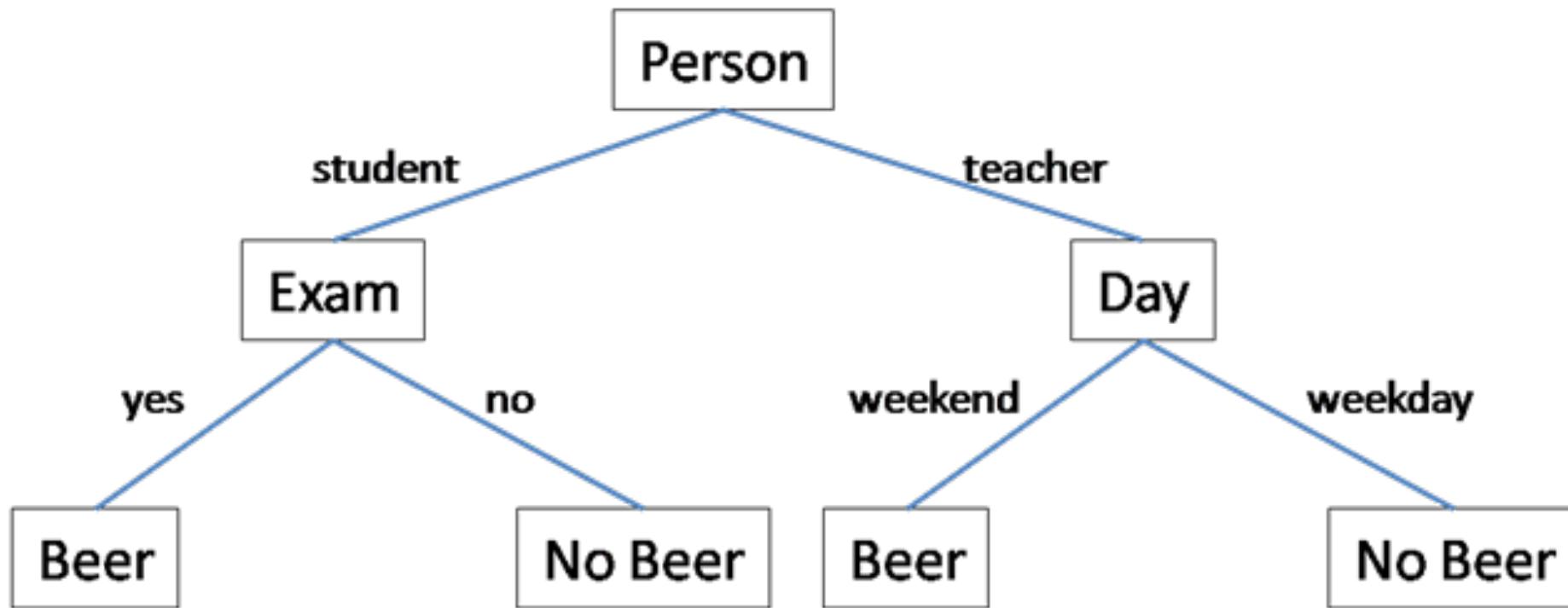
# Логические правила

- Как строить?
- Линейные модели
- Перебор, жадное наращивание
- Решающие деревья

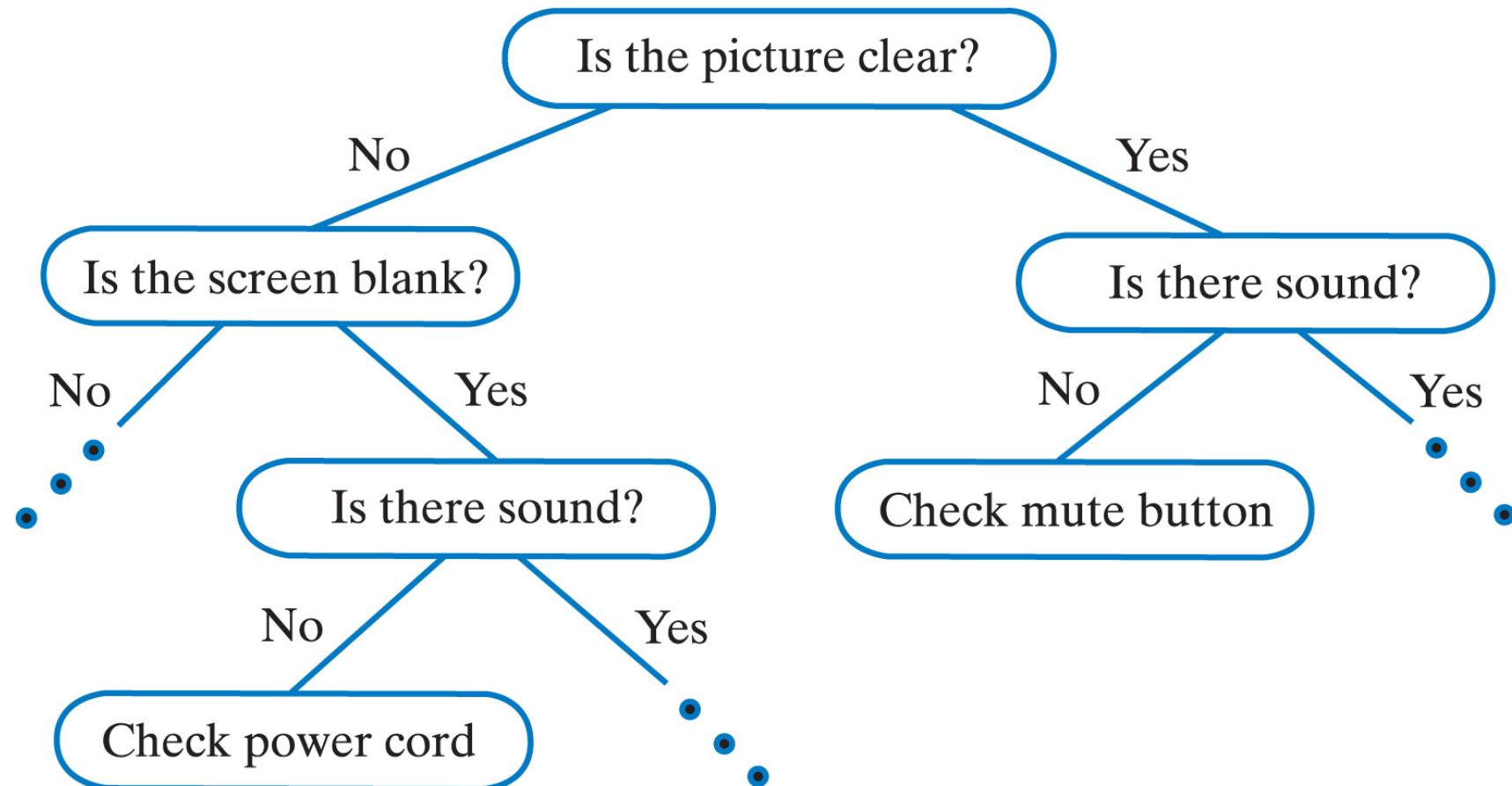
# Медицинская диагностика



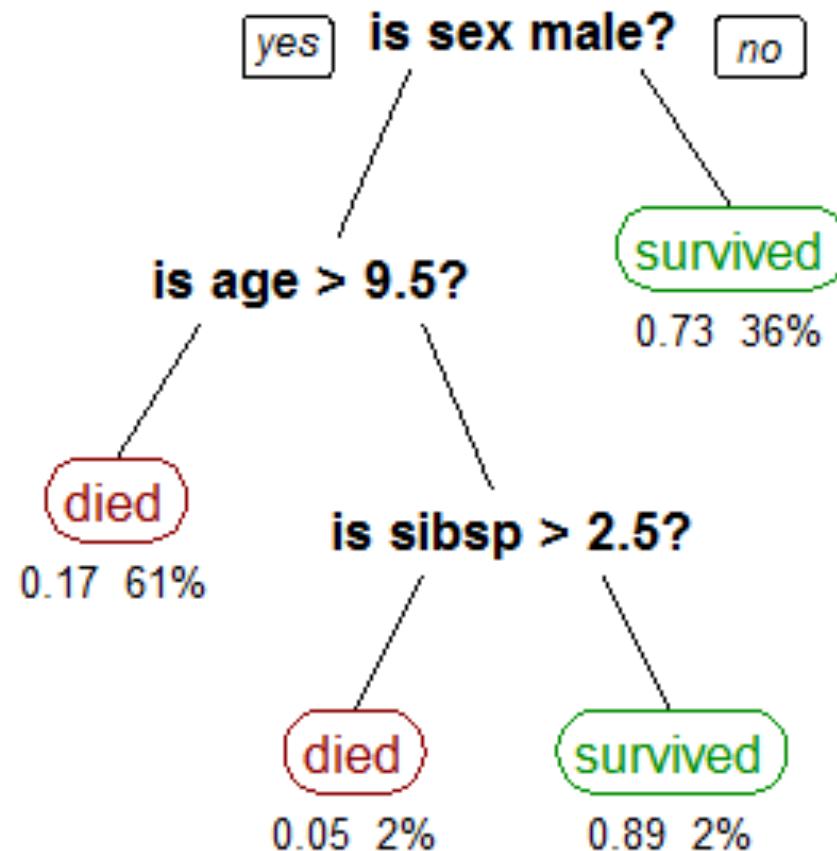
# Приятие решений



# Схема диалога с клиентом

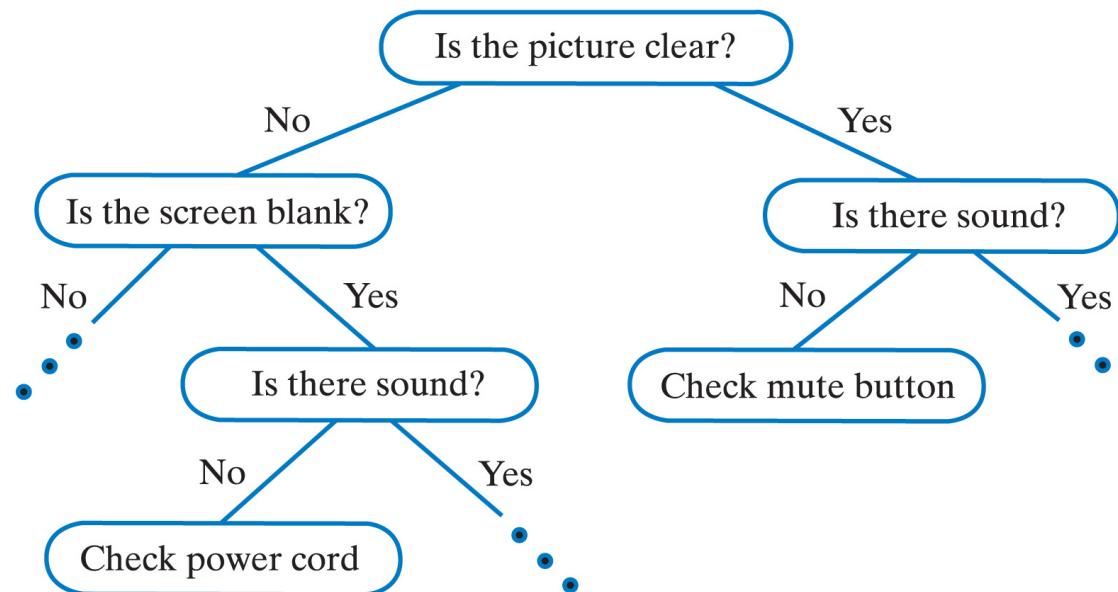


# Пассажиры Титаника



# Решающее дерево

- Бинарное дерево
- В каждой внутренней вершине записано условие
- В каждом листе записан прогноз (решение)



# Условия

- Самые популярные варианты:

$$[x^j \leq t] \quad \text{и} \quad [x^j = t]$$

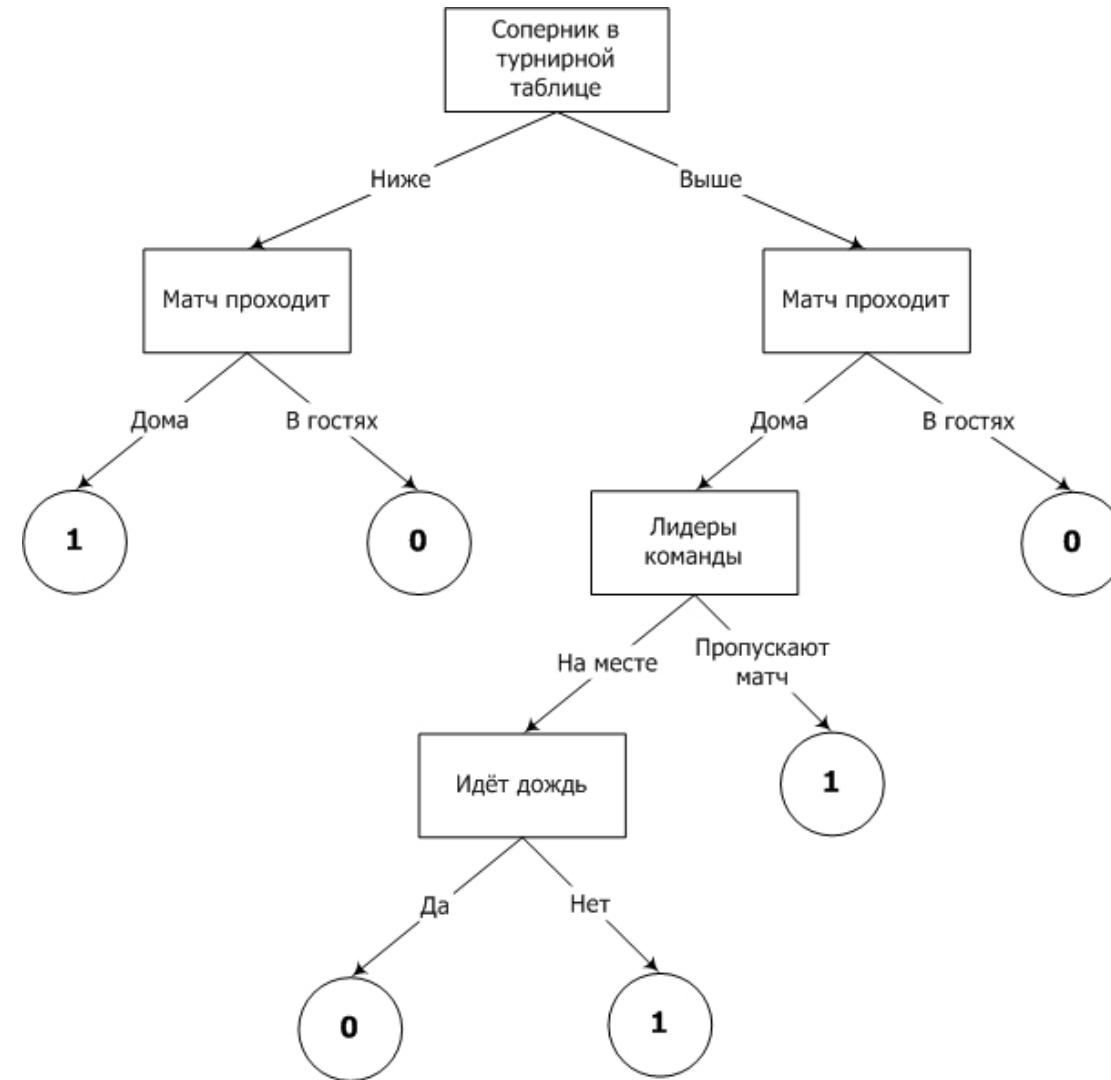
Примеры:

- [этаж = 5]
- [площадь ≤ 30]

# Прогноз в листе

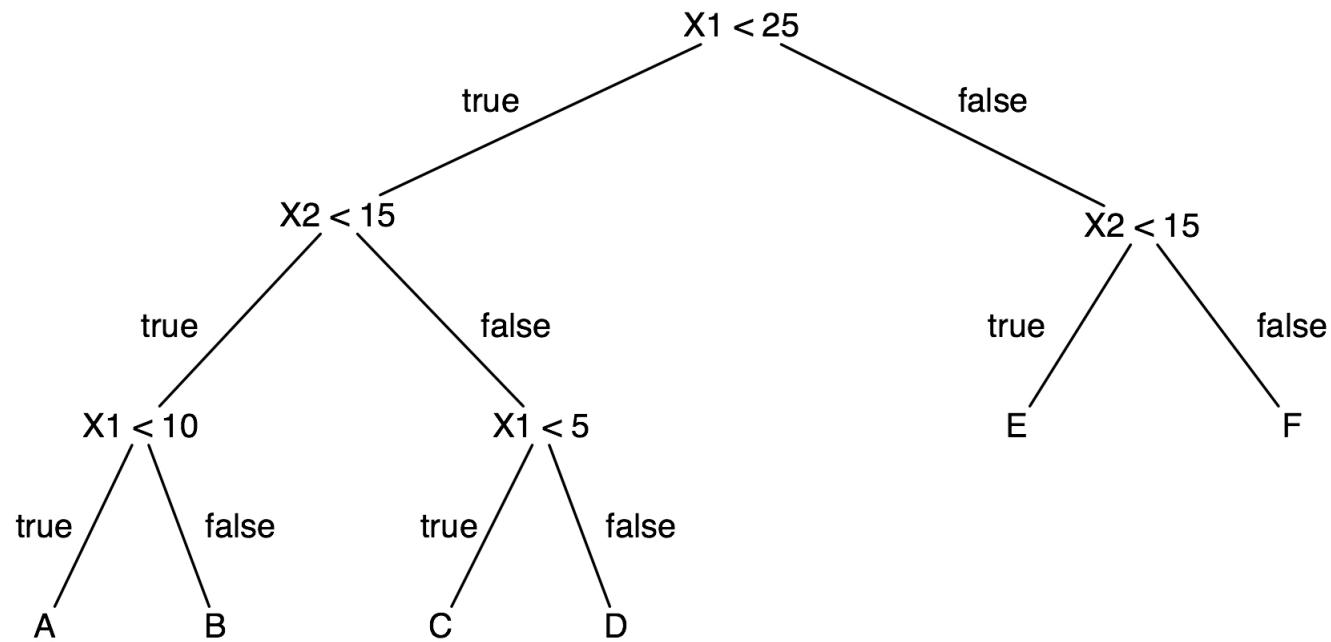
- Регрессия:
  - Вещественное число
- Классификация:
  - Класс
  - Вероятности классов

# Исход футбольного матча

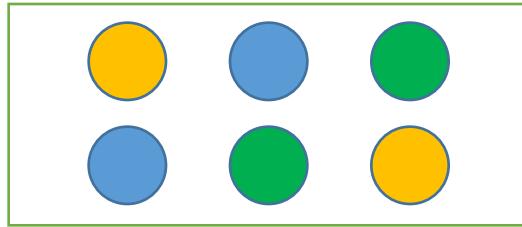


# Жадное построение

- Растим дерево от корня к листьям

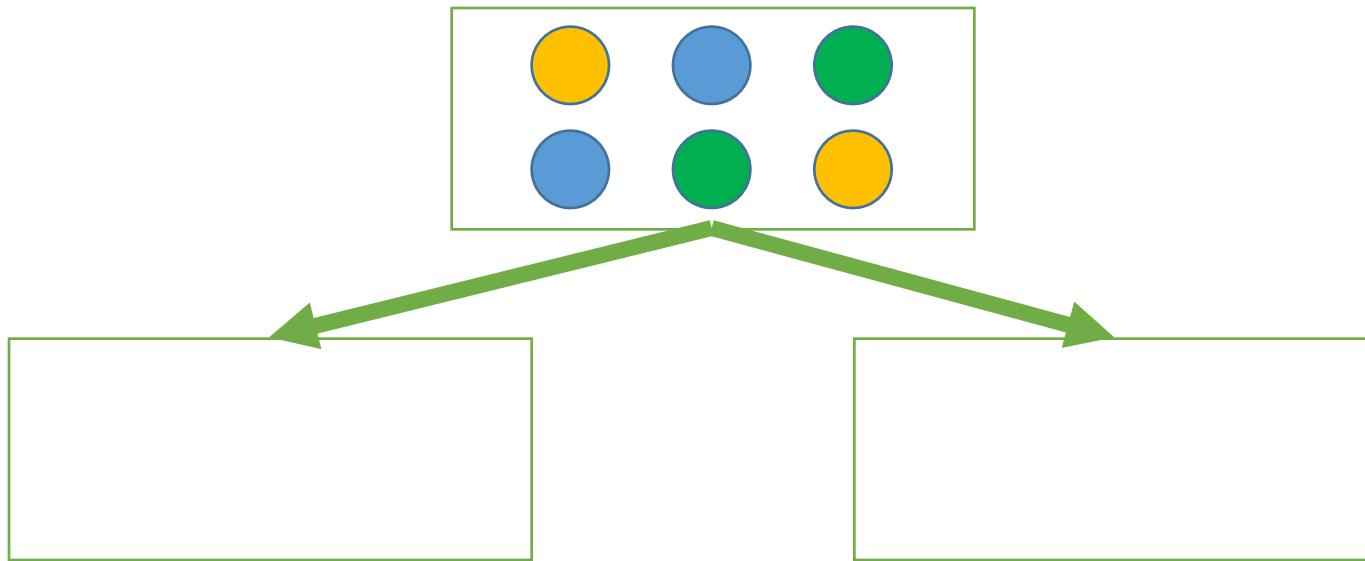


# Жадное построение

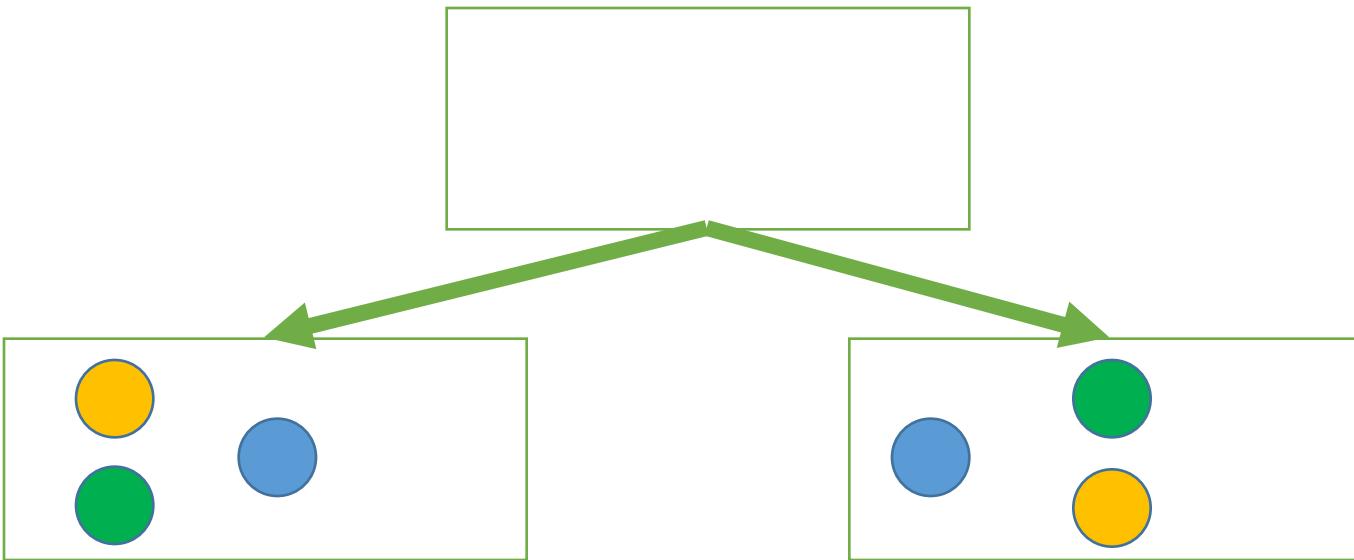


- Как разбить вершину?

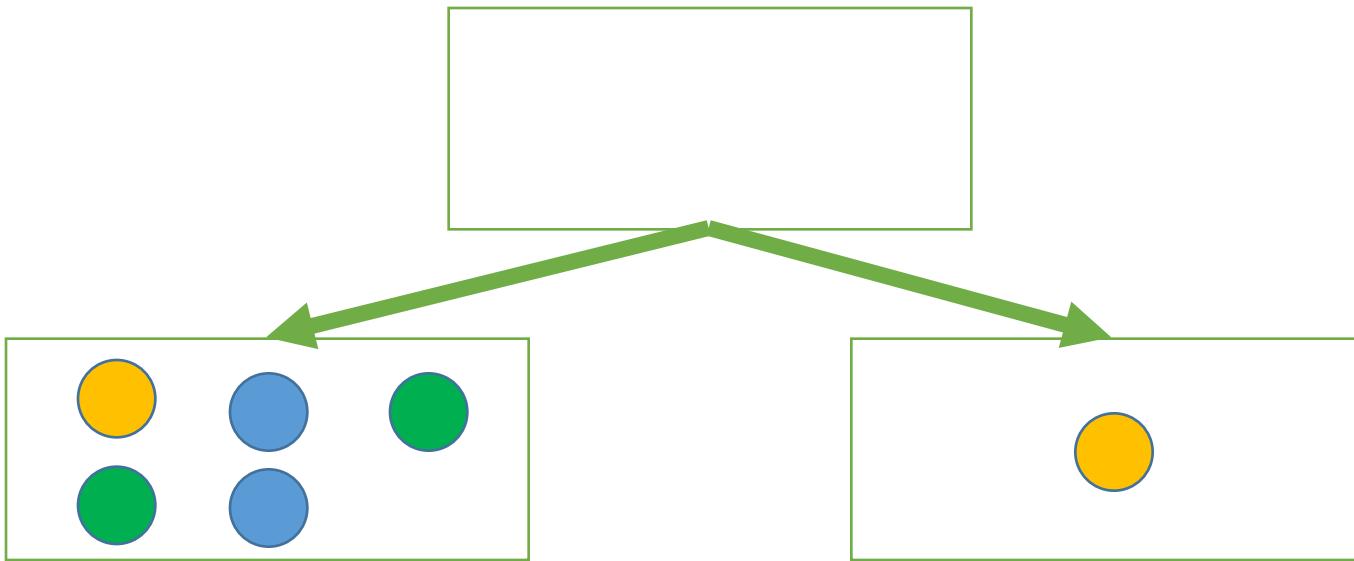
# Жадное построение



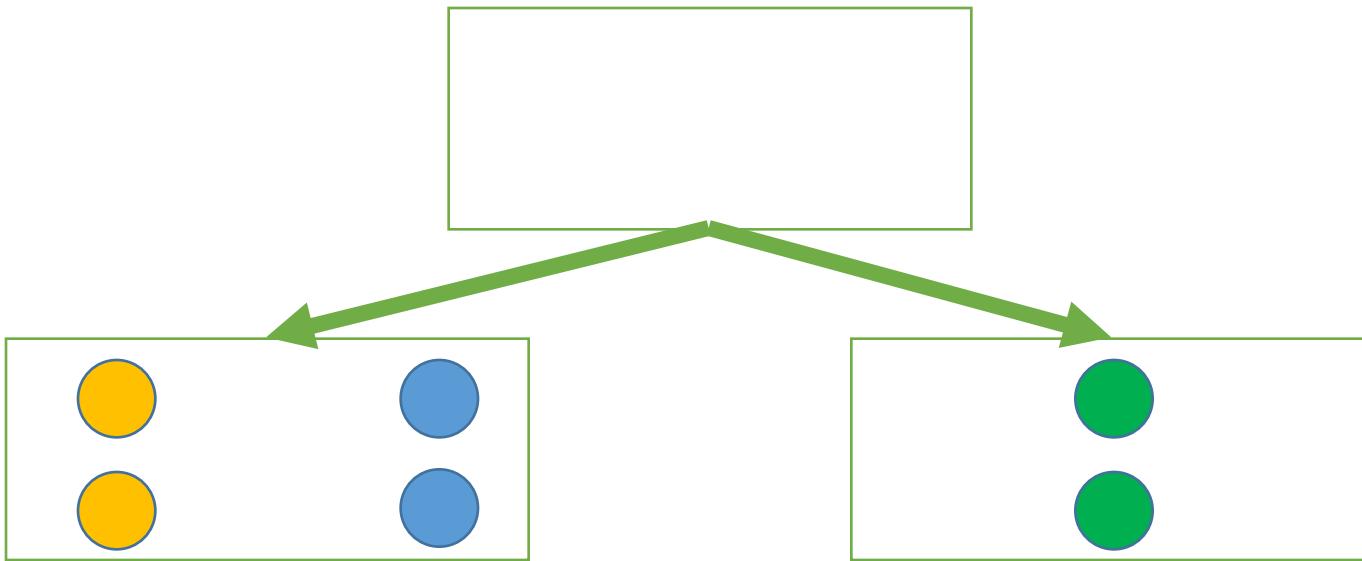
# Жадное построение



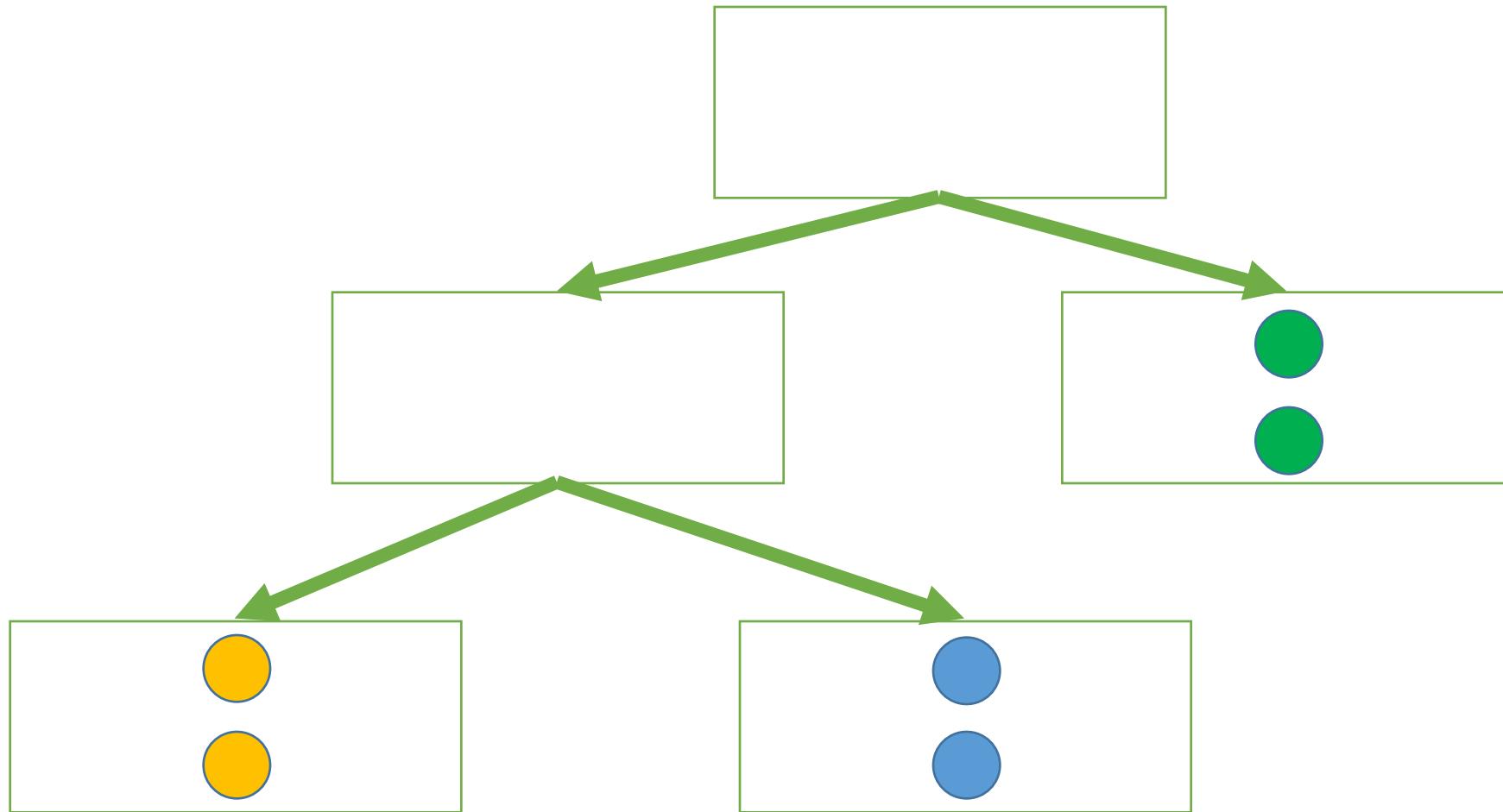
# Жадное построение



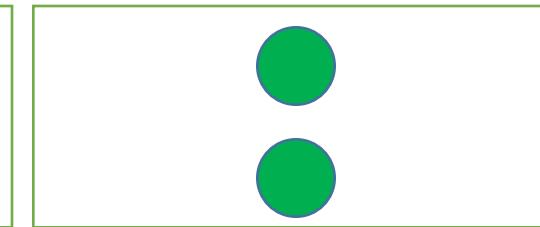
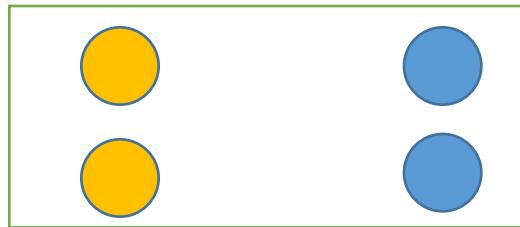
# Жадное построение



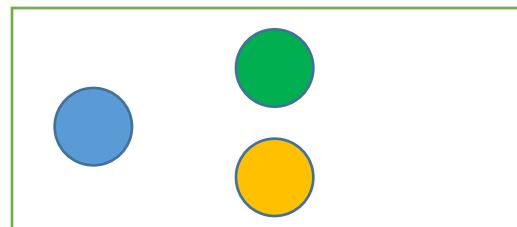
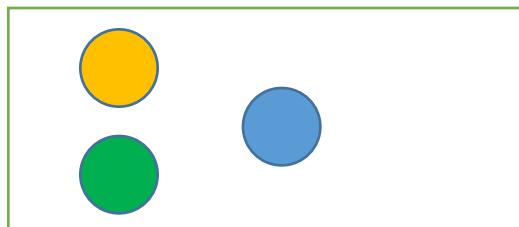
# Жадное построение



# Как сравнить разбиения?

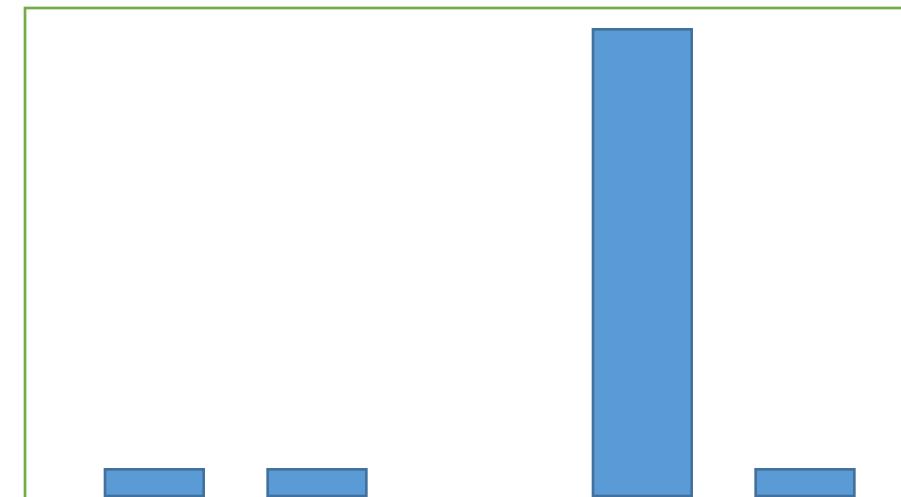
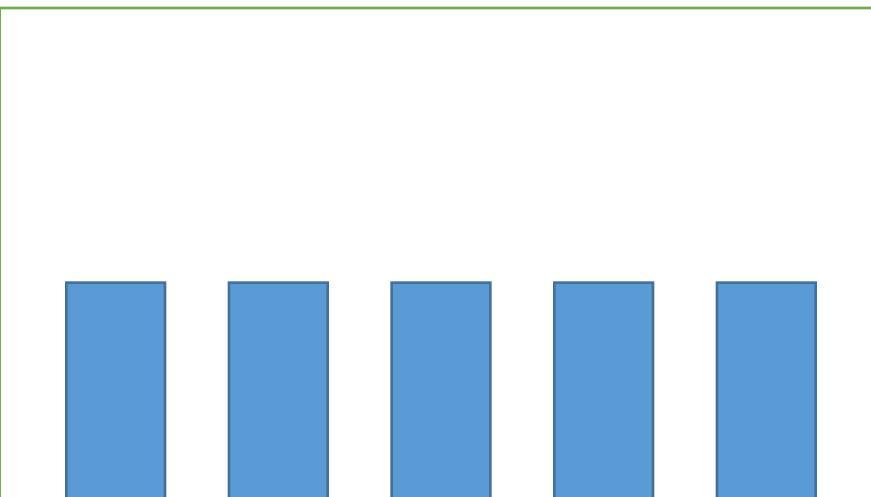


или



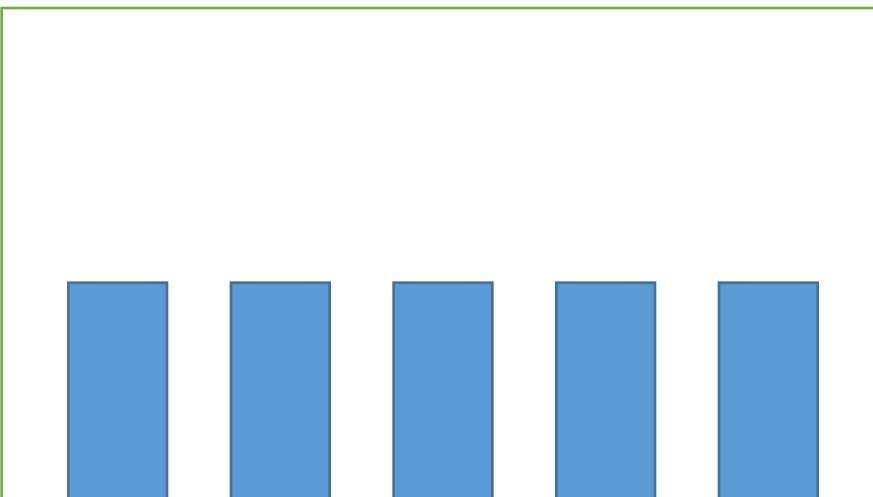
# Энтропия

- Мера неопределённости распределения

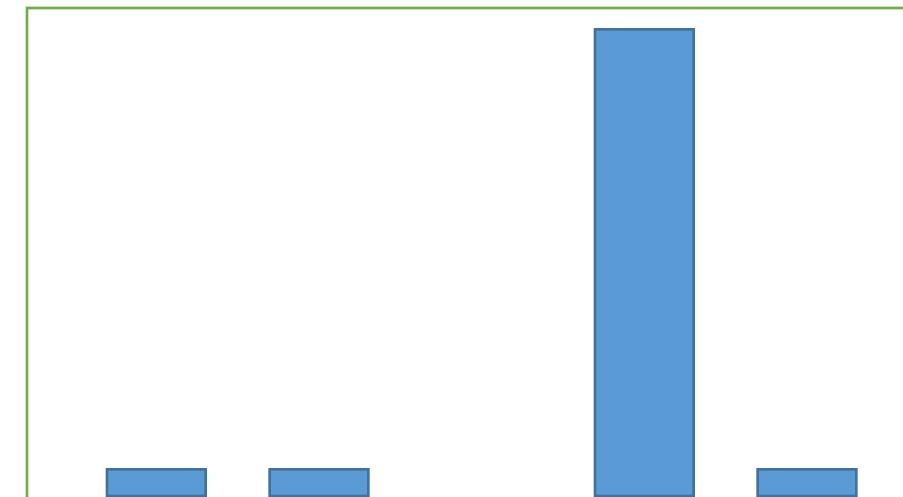


# Энтропия

- Мера неопределённости распределения



Высокая энтропия



Низкая энтропия

# Энтропия

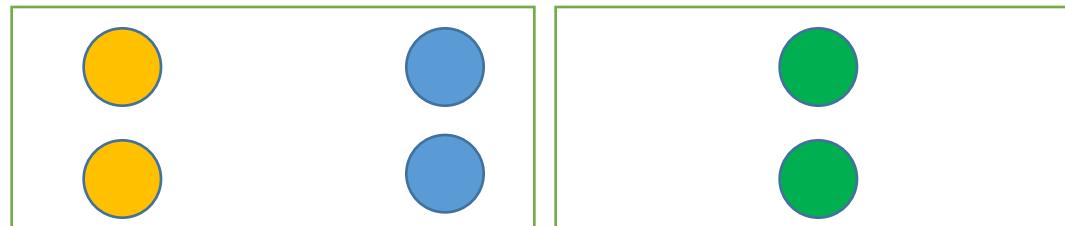
- Дискретное распределение
- Принимает  $n$  значений с вероятностями  $p_1, \dots, p_n$
- Энтропия:

$$H(p_1, \dots, p_n) = - \sum_{i=1}^n p_i \log p_i$$

# Энтропия

- $(0.2, 0.2, 0.2, 0.2, 0.2)$
- $H = 1.60944 \dots$
- $(0.9, 0.05, 0.05, 0, 0)$
- $H = 0.394398 \dots$
- $(0, 0, 0, 1, 0)$
- $H = 0$

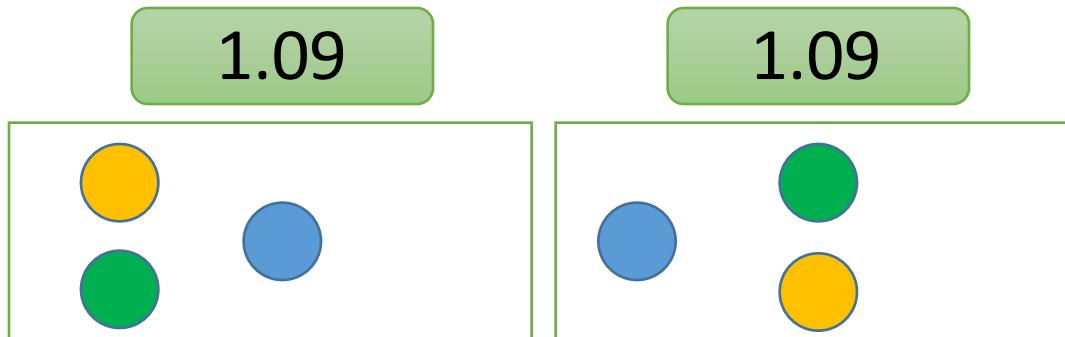
# Как сравнить разбиения?



0.693

0

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

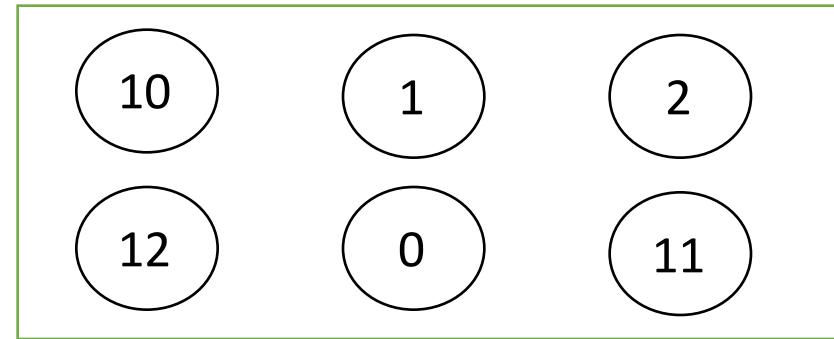


1.09

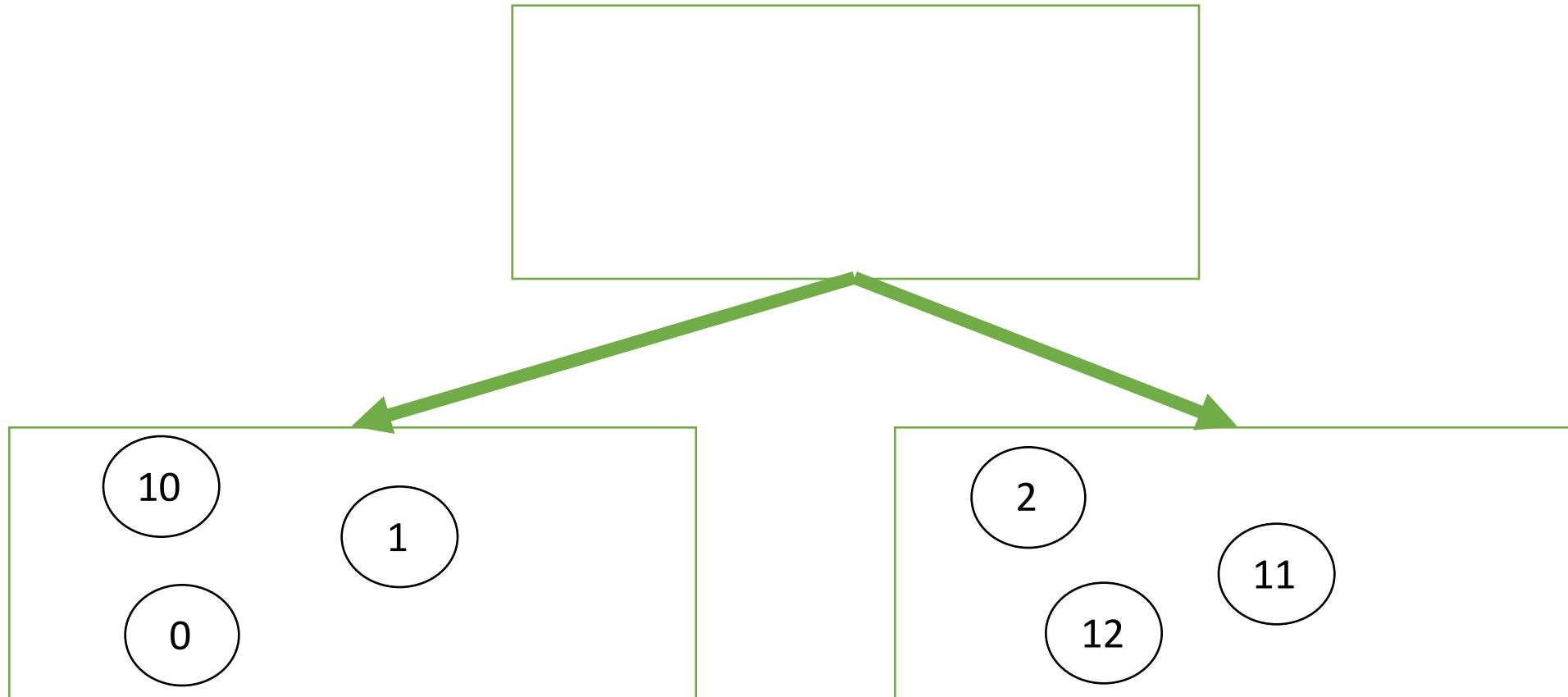
1.09

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

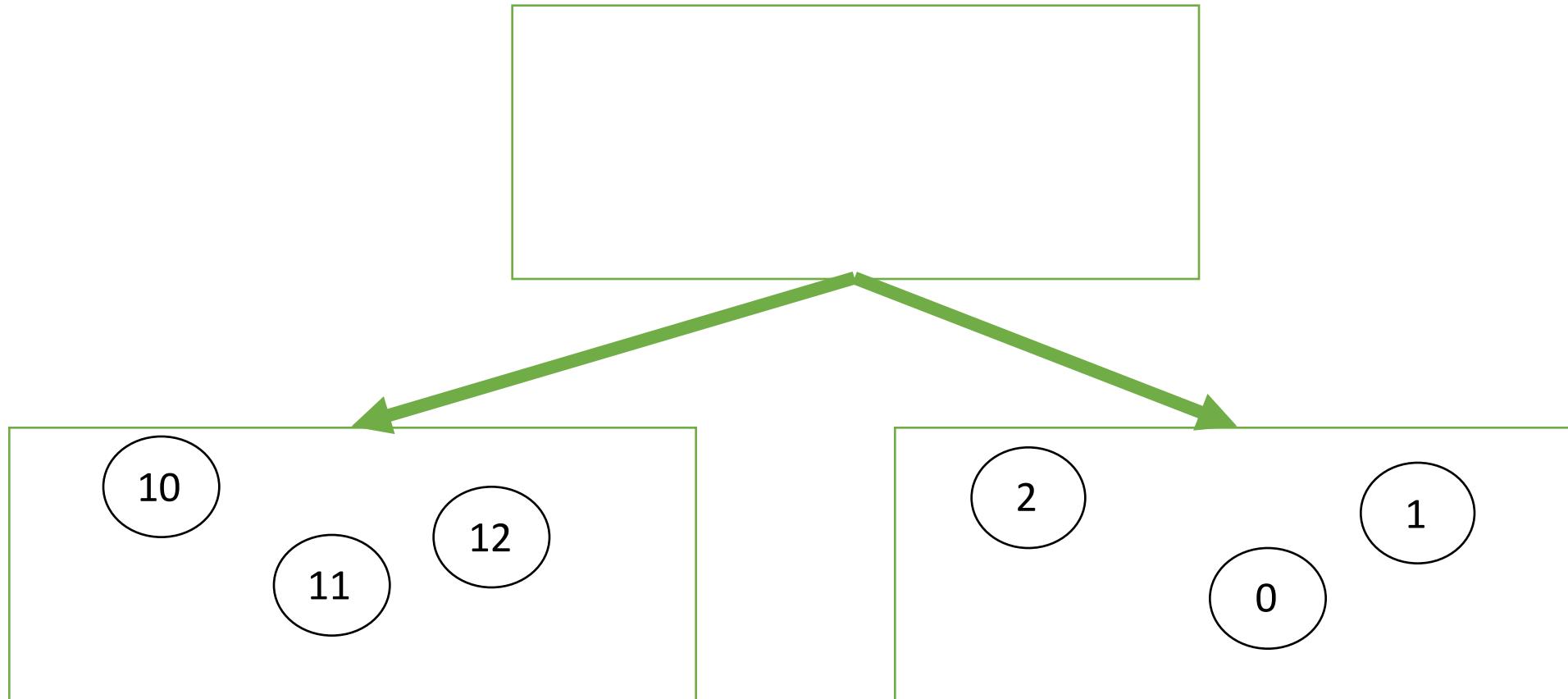
# А для регрессии?



# А для регрессии?



# А для регрессии?



# А для регрессии?

- Выбираем разбиение с наименьшей суммарной дисперсией
- Чем меньше дисперсия, тем меньше неопределённости

# 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:

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

# 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

# Поиск разбиения

- После того, как разбиение найдено:
- Разбиваем  $X_m$  на две части:

$$X_l = \{x \in X_m \mid [x^j \leq t]\}$$

$$X_r = \{x \in X_m \mid [x^j > t]\}$$

- Повторяем процедуру для дочерних вершин

# Критерий останова

- В какой момент прекращать разбиение вершин?
- В вершине один объекты?
- В вершине объекты одного класса?
- Глубина превысила порог?

# Ответ в листе

- Допустим, решили сделать вершину  $m$  листом
- Какой прогноз выбрать?
- Регрессия:

$$a_m = \frac{1}{|X_m|} \sum_{i \in X_m} y_i$$

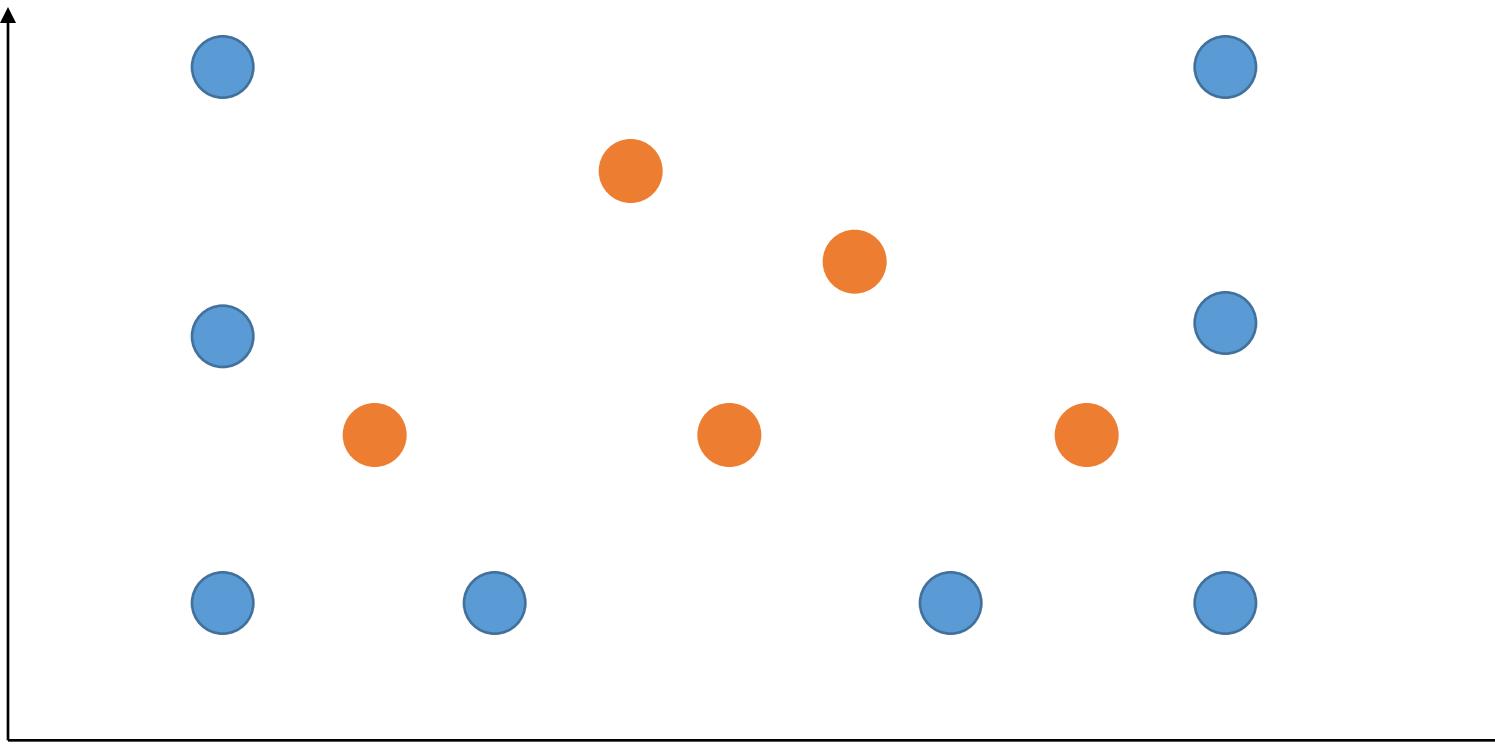
- Классификация:

$$a_m = \arg \max_{y \in \mathbb{Y}} \sum_{i \in X_m} [y_i = y]$$

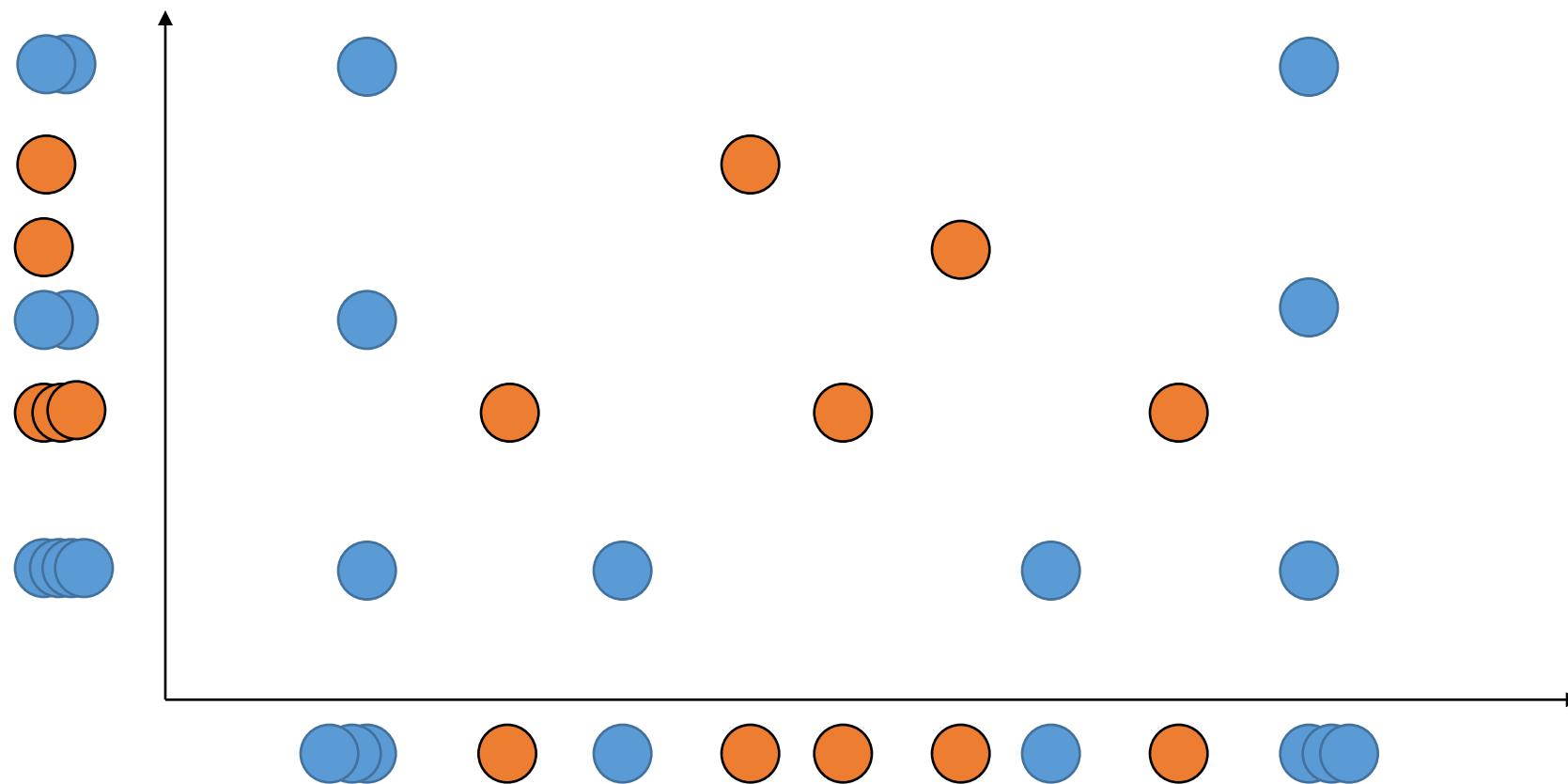
# Жадный алгоритм построения дерева

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  $r$

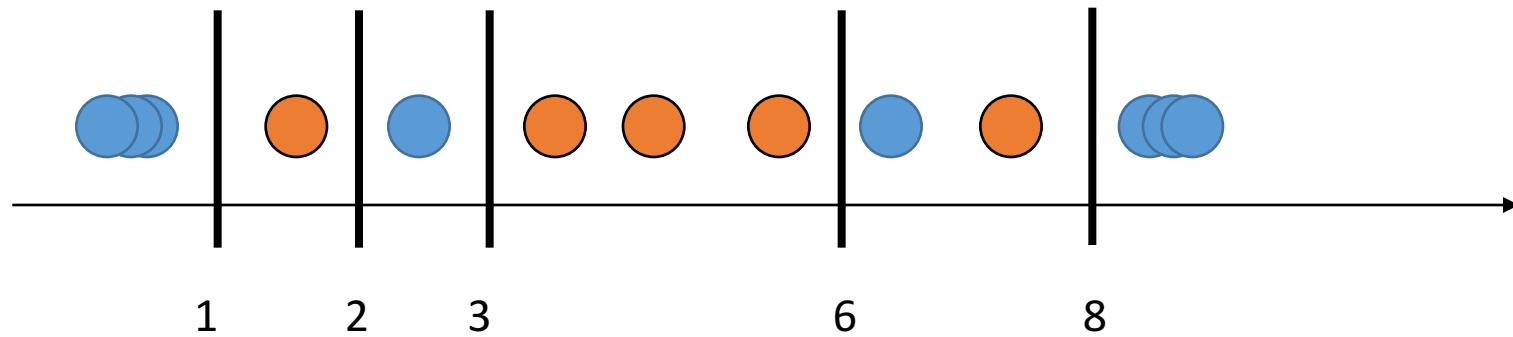
# Обучение деревьев



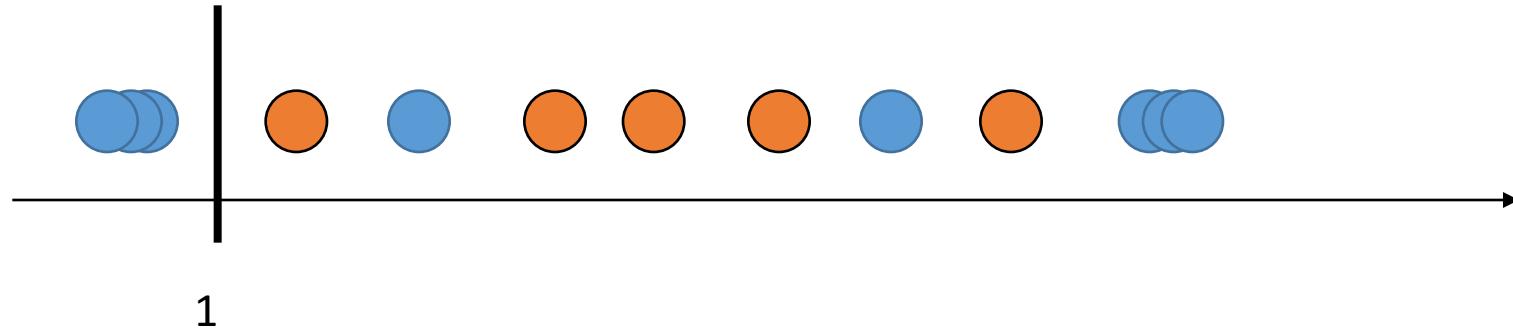
# Признаки



# Разбиения по признаку 1



# Разбиения по признаку 1

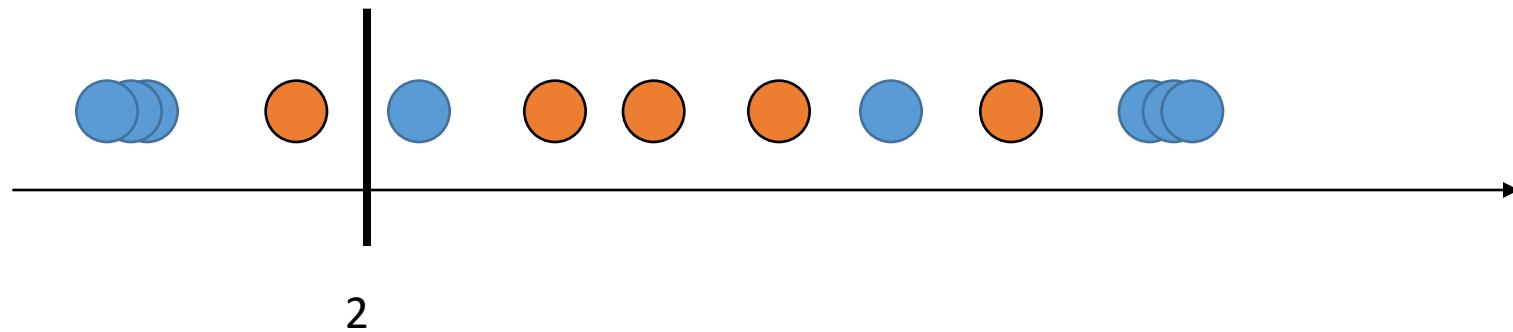


$$(1, 0)$$
$$H(p) = 0$$

$$(1/2, 1/2)$$
$$H(p) = 0.69$$

$$\frac{3}{13} H(p_l) + \frac{10}{13} H(p_r) = 0.53$$

# Разбиения по признаку 1

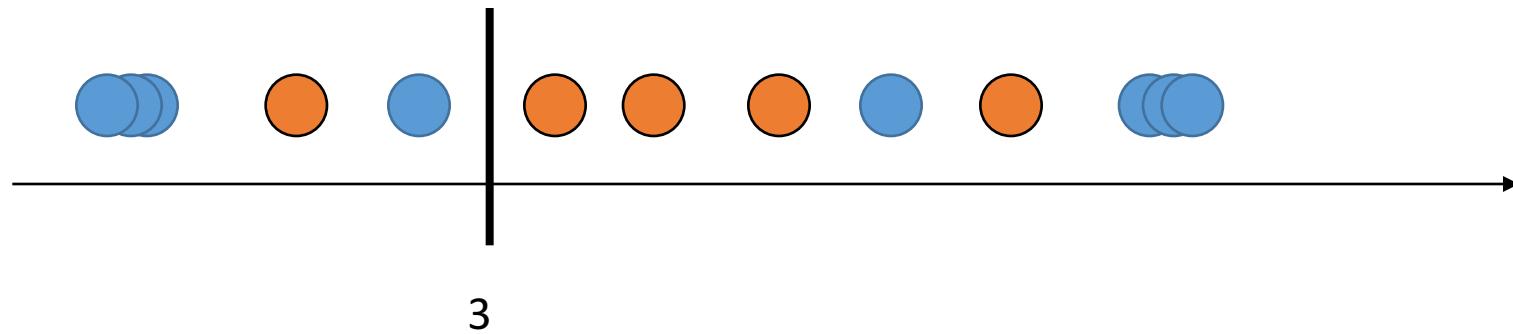


$$(3/4, 1/4)$$
$$H(p) = 0.56$$

$$(5/9, 4/9)$$
$$H(p) = 0.69$$

$$\frac{4}{13}H(p_l) + \frac{9}{13}H(p_r) = 0.65$$

# Разбиения по признаку 1

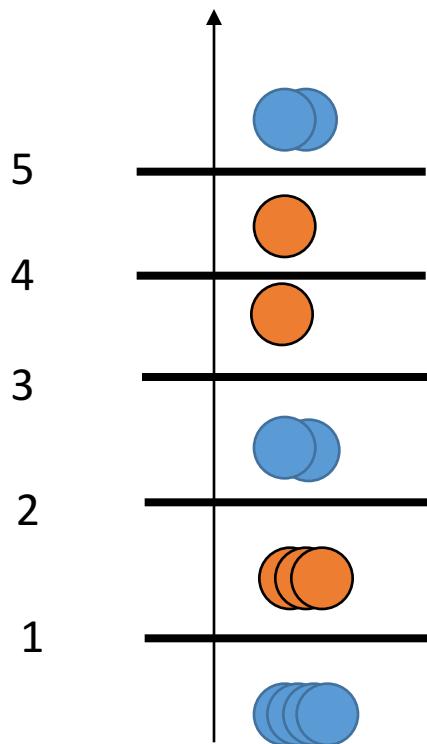


$$(4/5, 1/5)$$
$$H(p) = 0.5$$

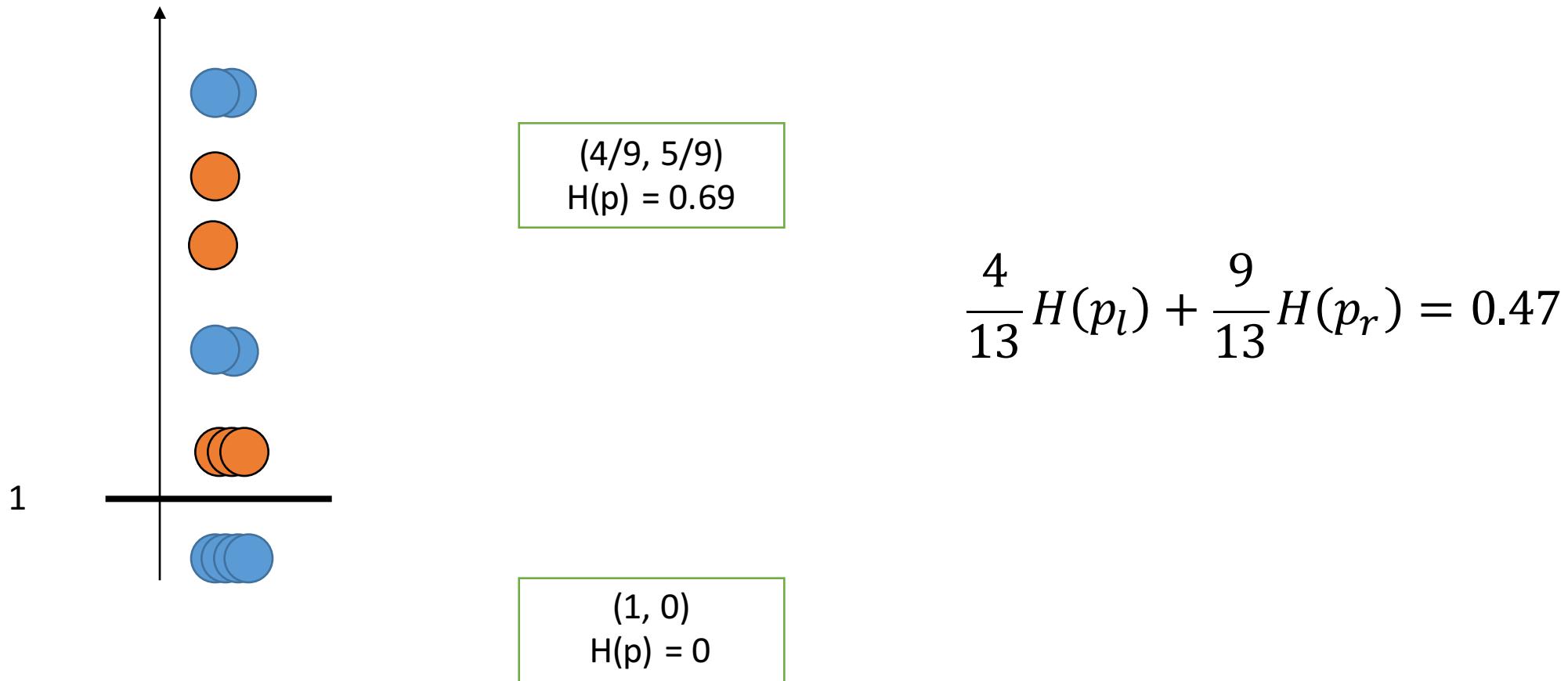
$$(1/2, 1/2)$$
$$H(p) = 0.69$$

$$\frac{5}{13}H(p_l) + \frac{8}{13}H(p_r) = 0.62$$

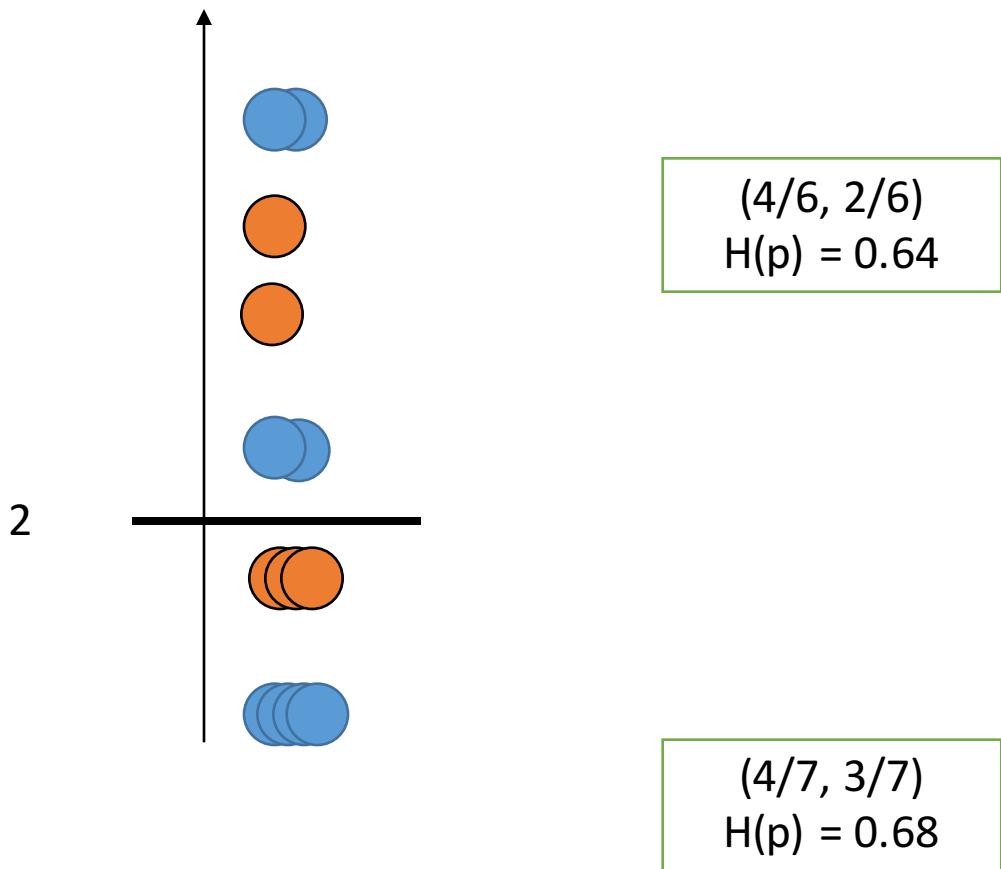
# Разбиения по признаку 2



# Разбиения по признаку 2

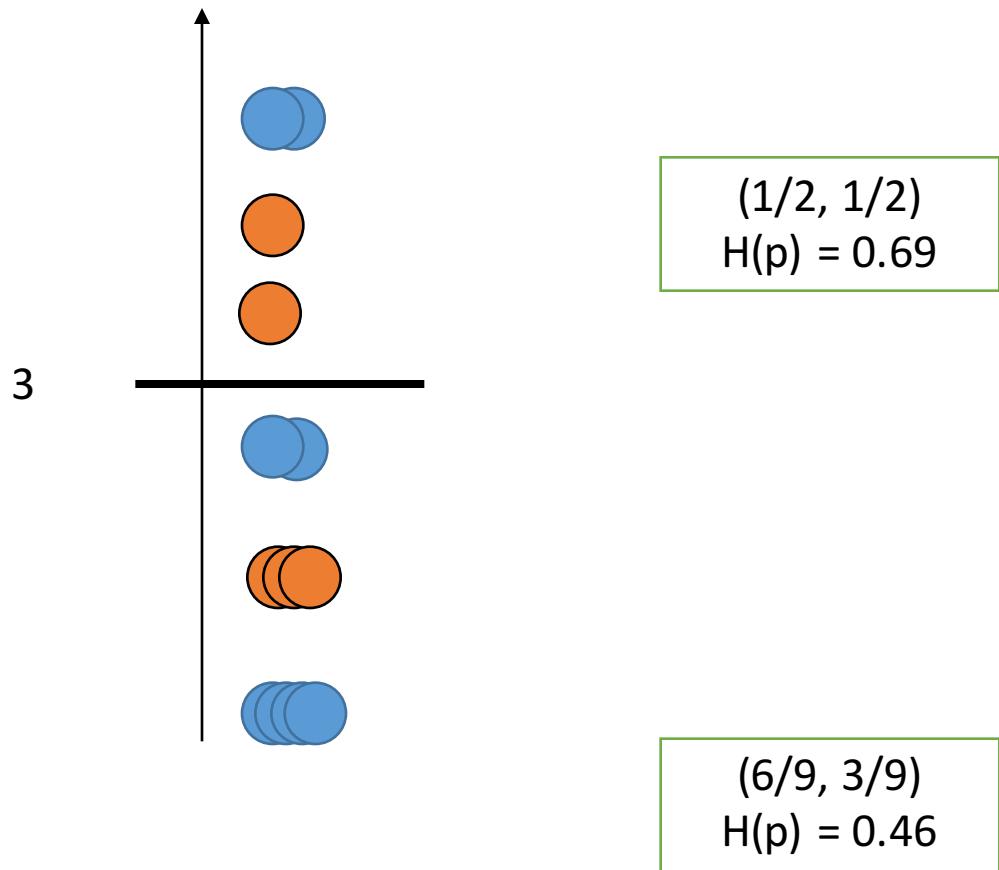


# Разбиения по признаку 2



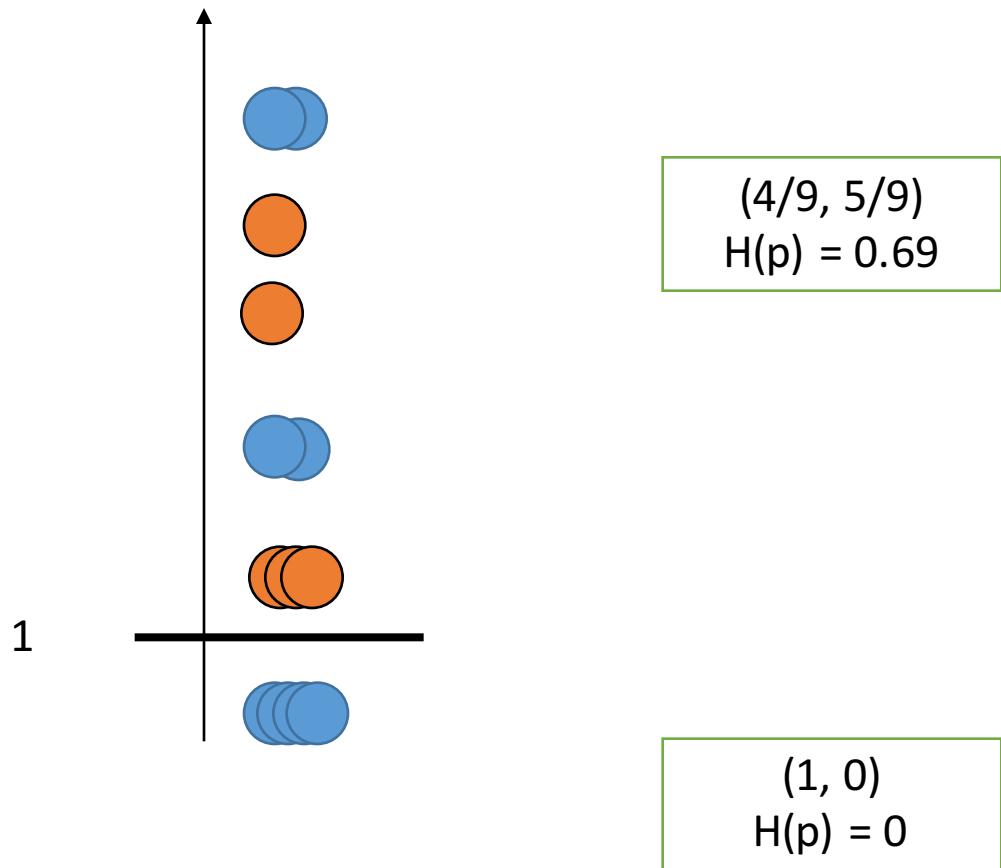
$$\frac{7}{13}H(p_l) + \frac{6}{13}H(p_r) = 0.66$$

# Разбиения по признаку 2



$$\frac{9}{13}H(p_l) + \frac{4}{13}H(p_r) = 0.53$$

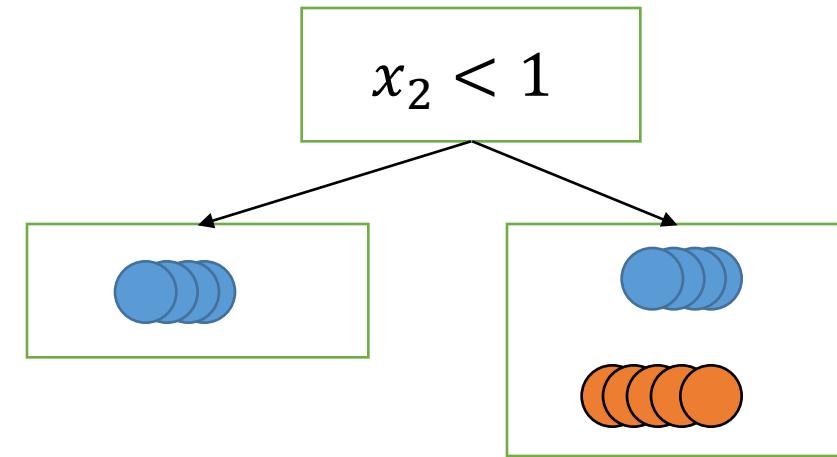
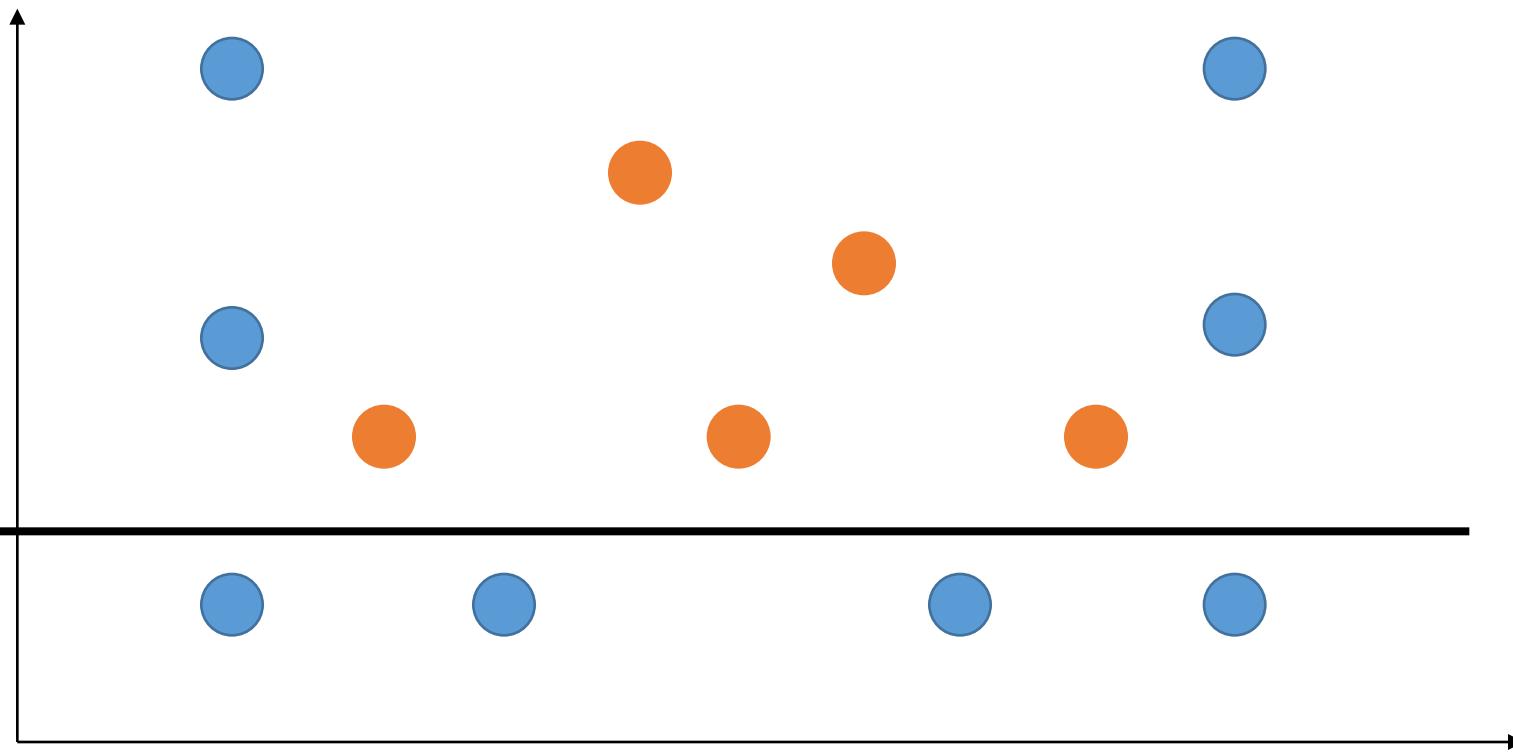
# Разбиения по признаку 2



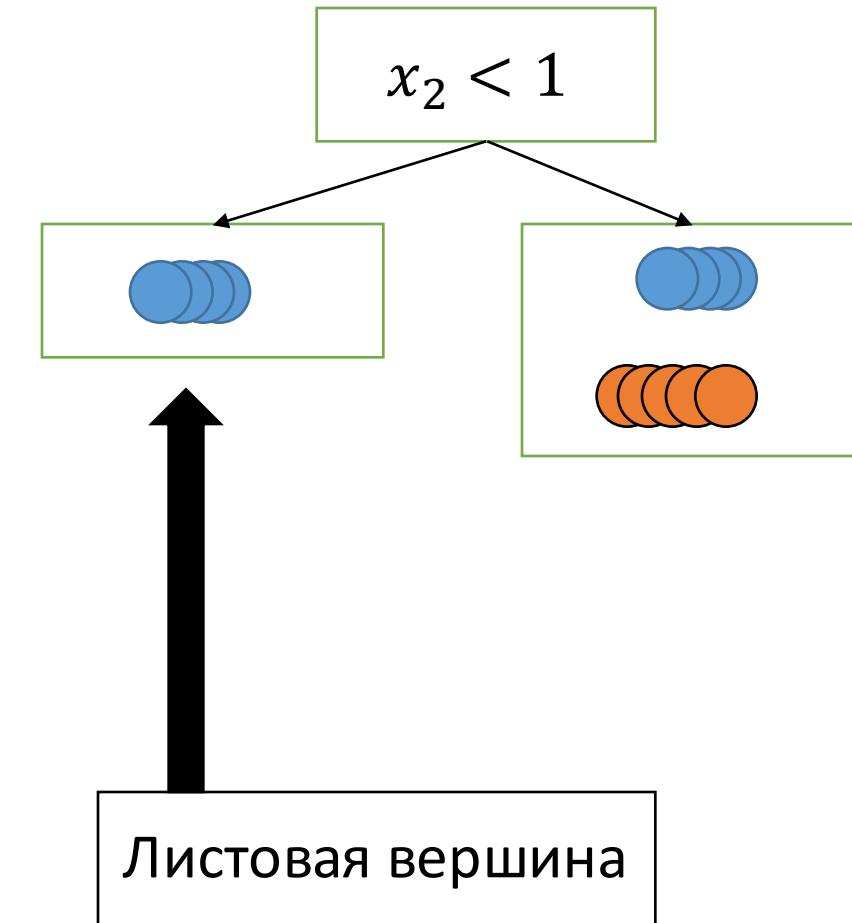
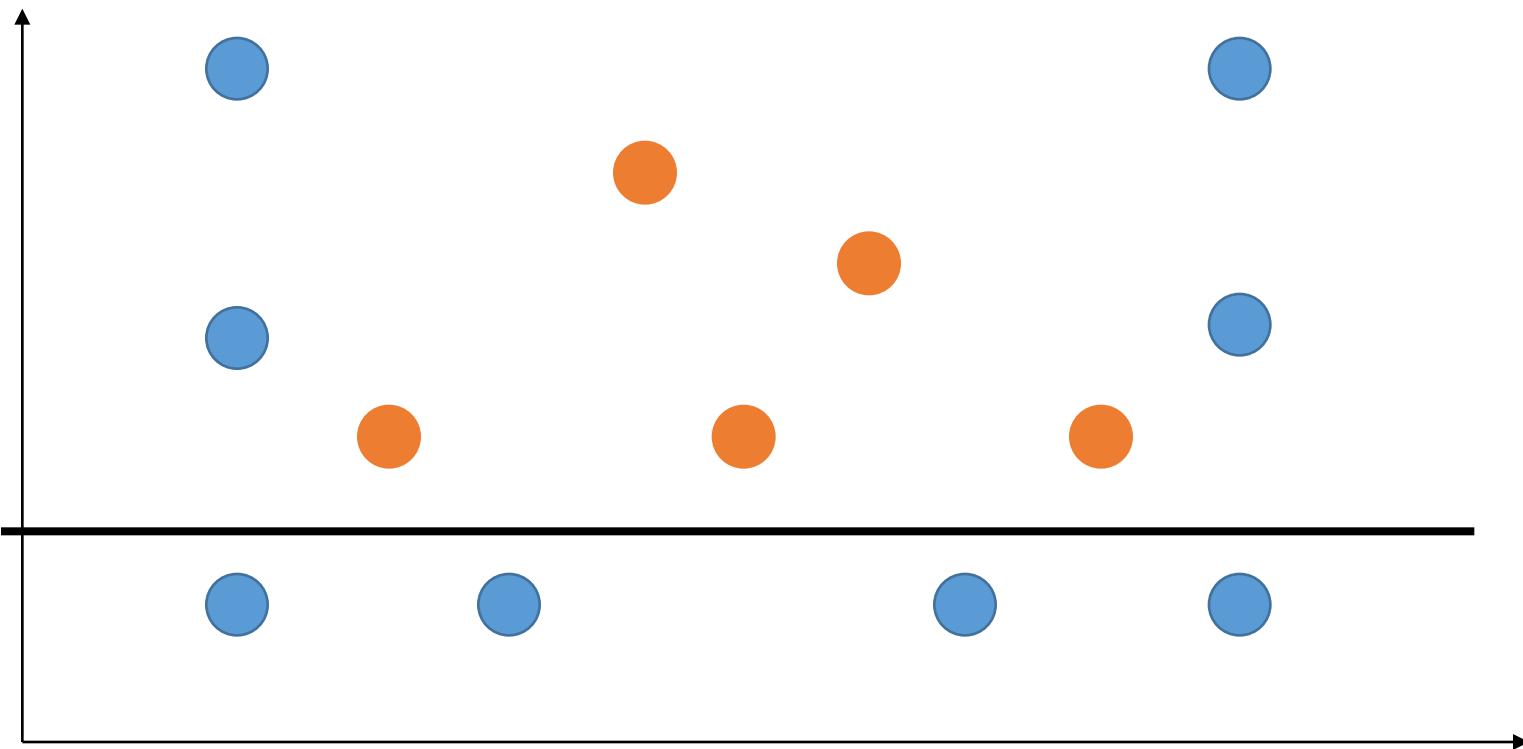
$$\frac{4}{13}H(p_l) + \frac{9}{13}H(p_r) = 0.47$$

Лучшее разбиение!

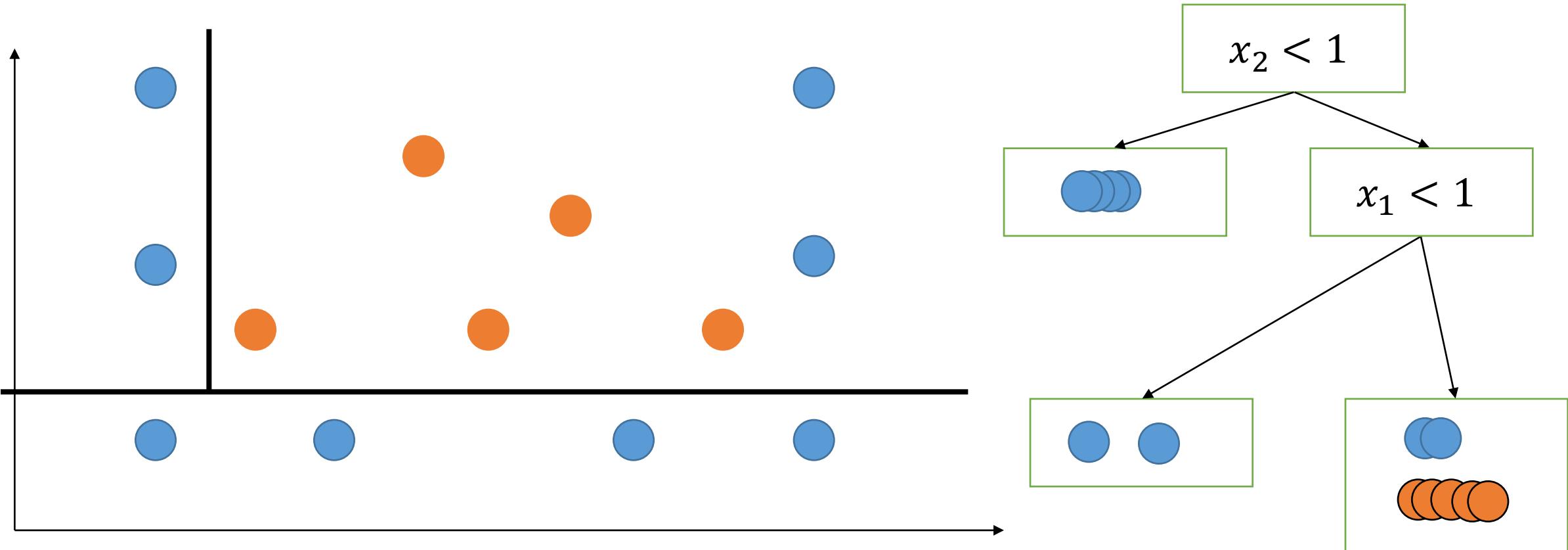
# Обучение деревьев



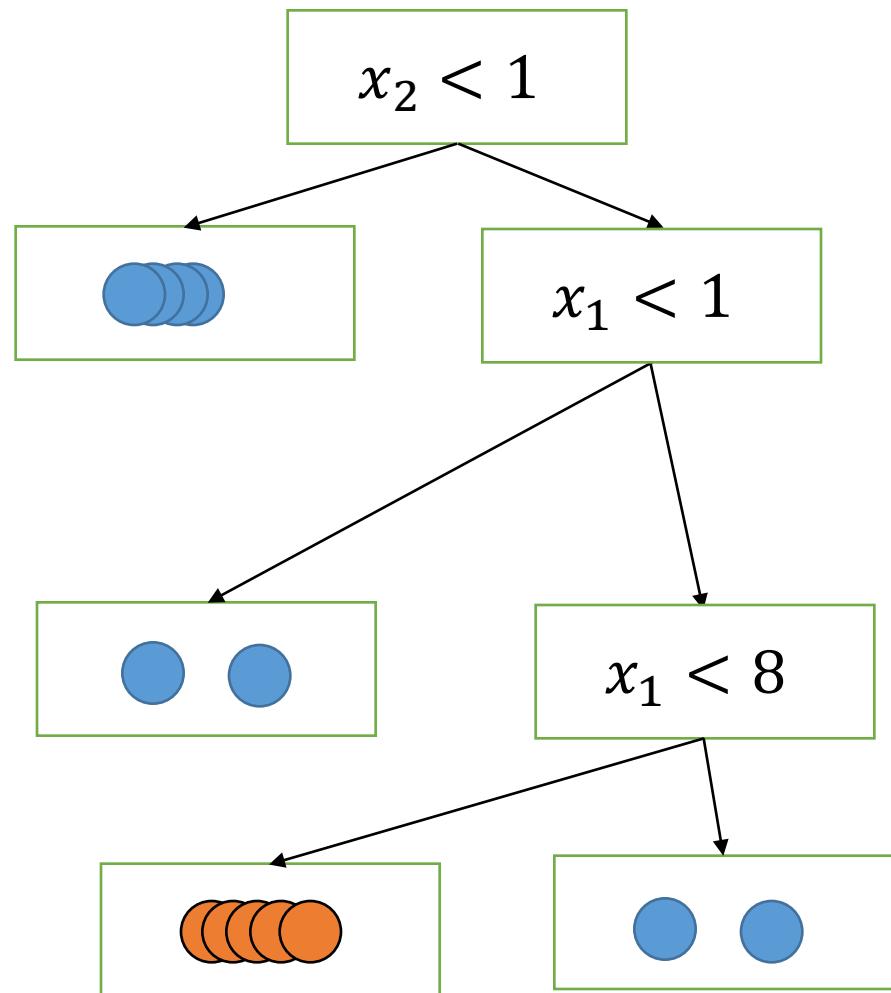
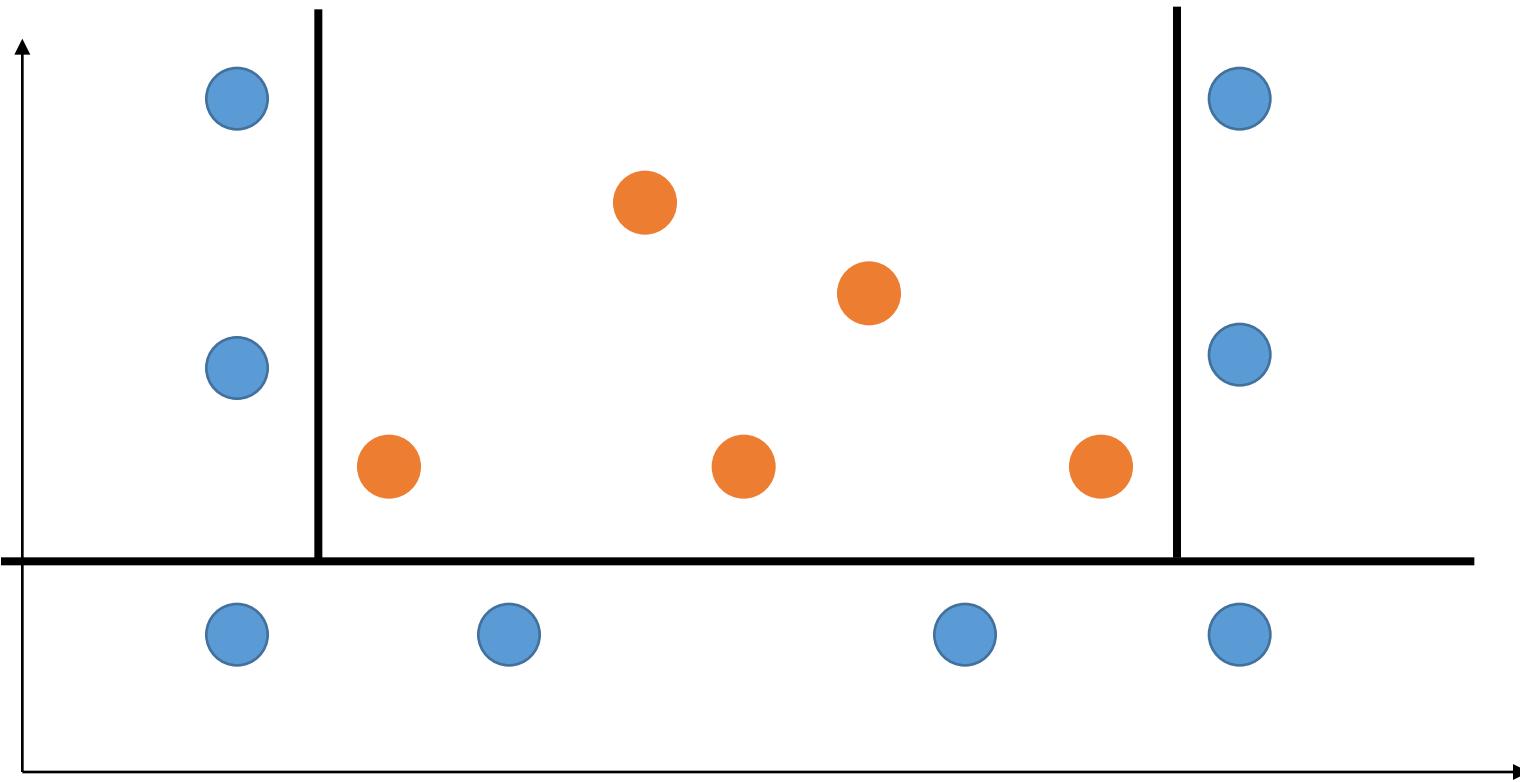
# Обучение деревьев



# Обучение деревьев

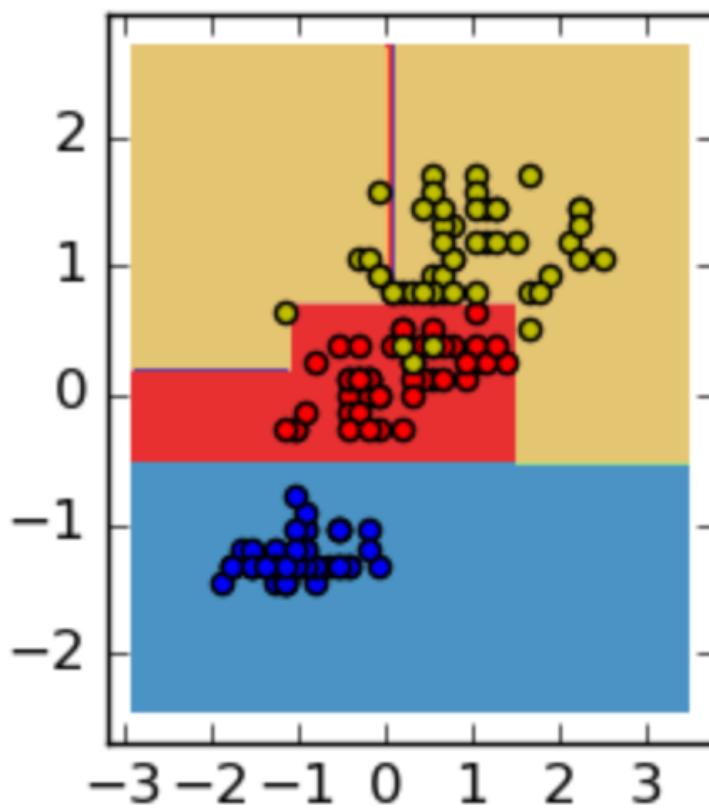


# Обучение деревьев

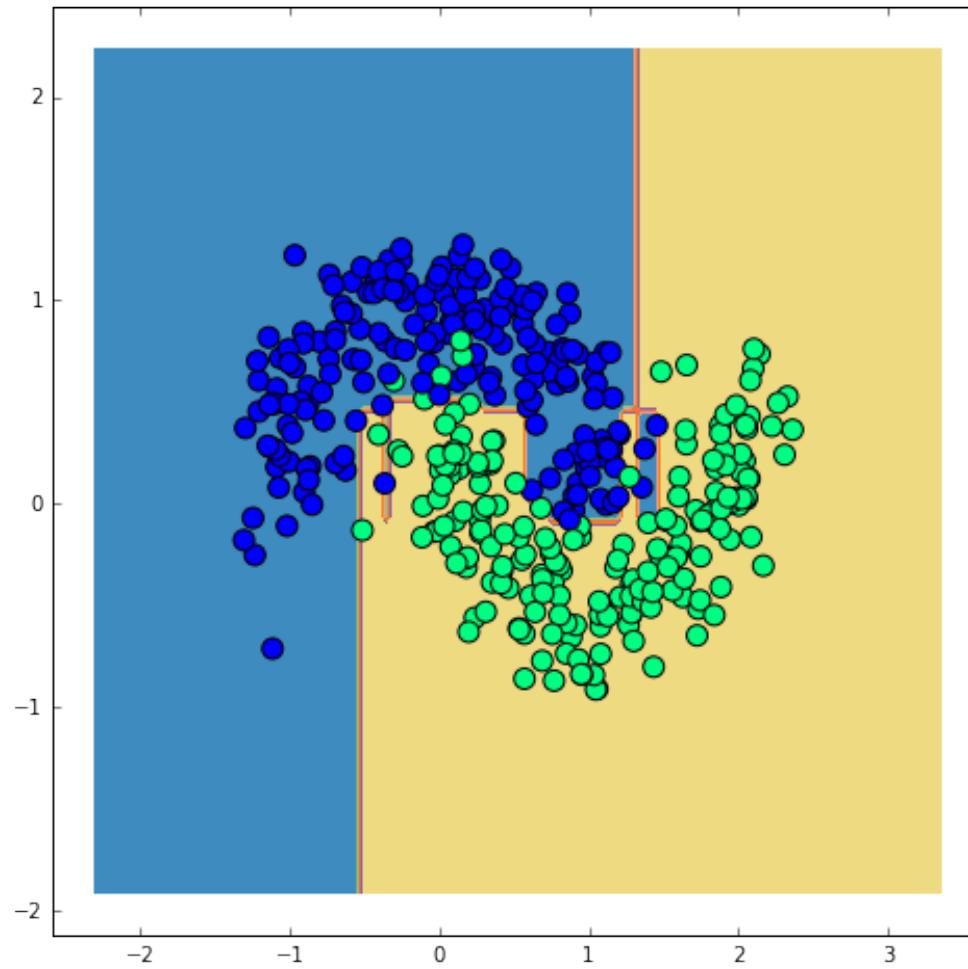


Переобучение деревьев и  
борьба с ним

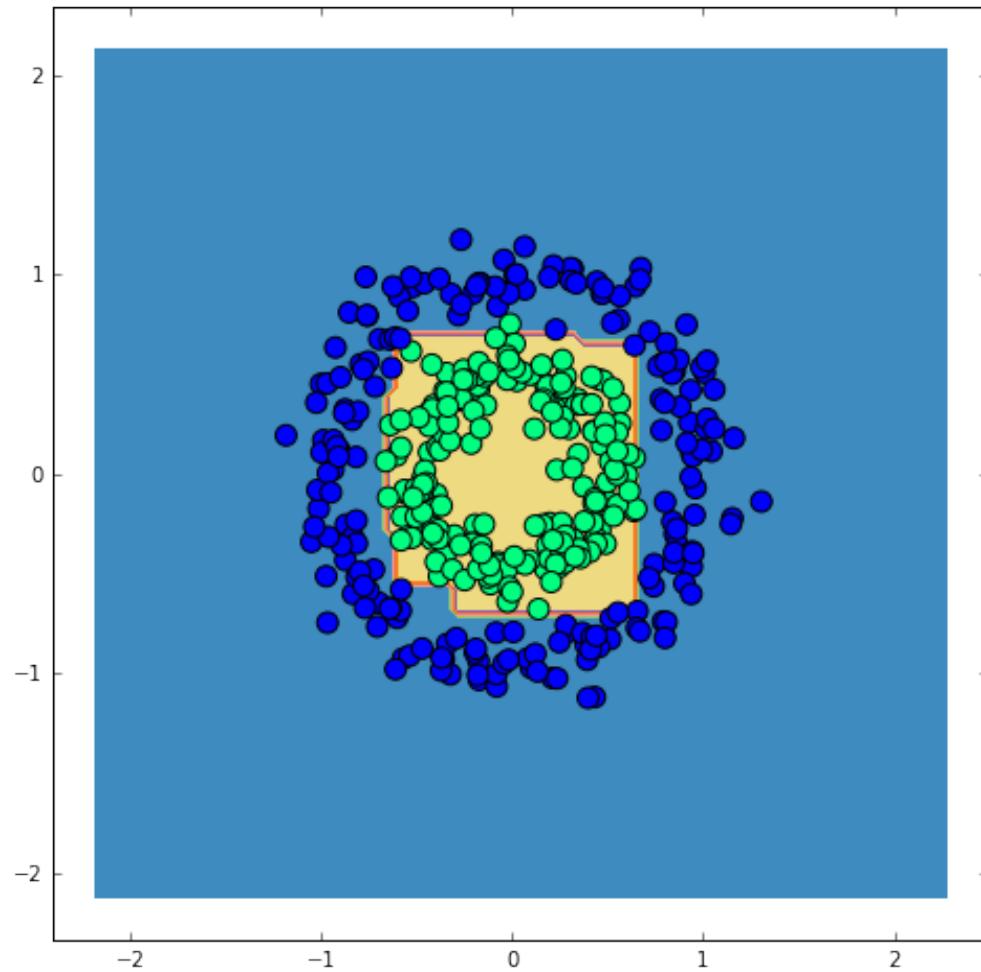
# Классификация



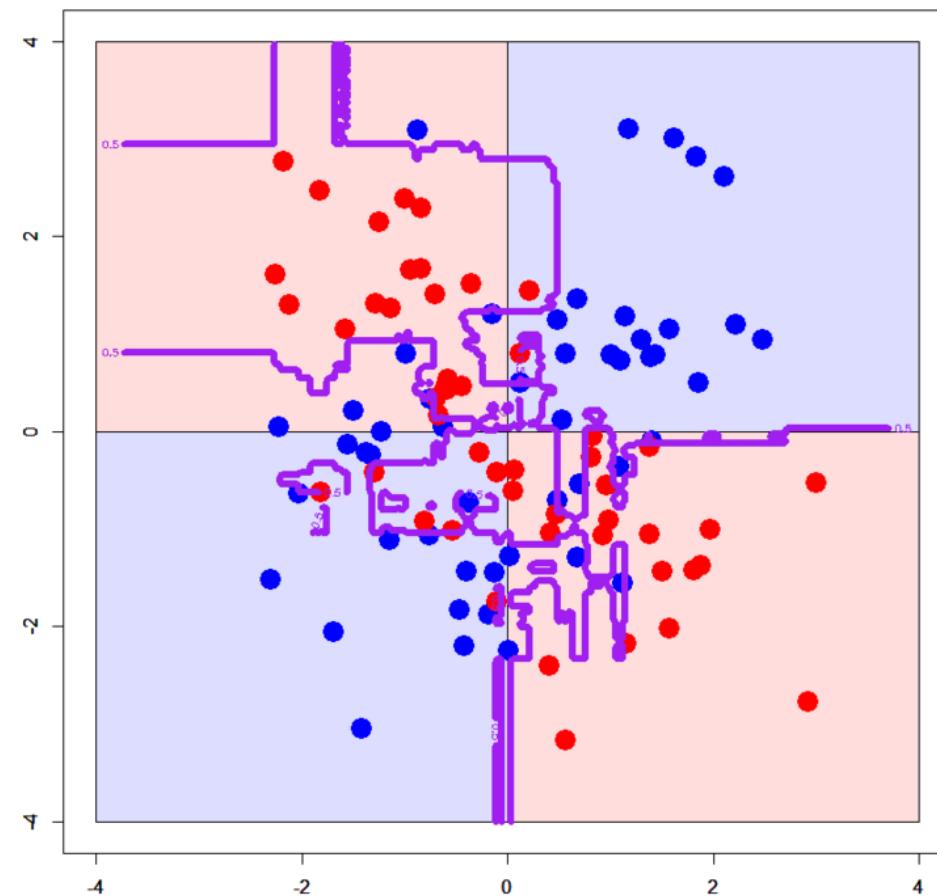
# Классификация



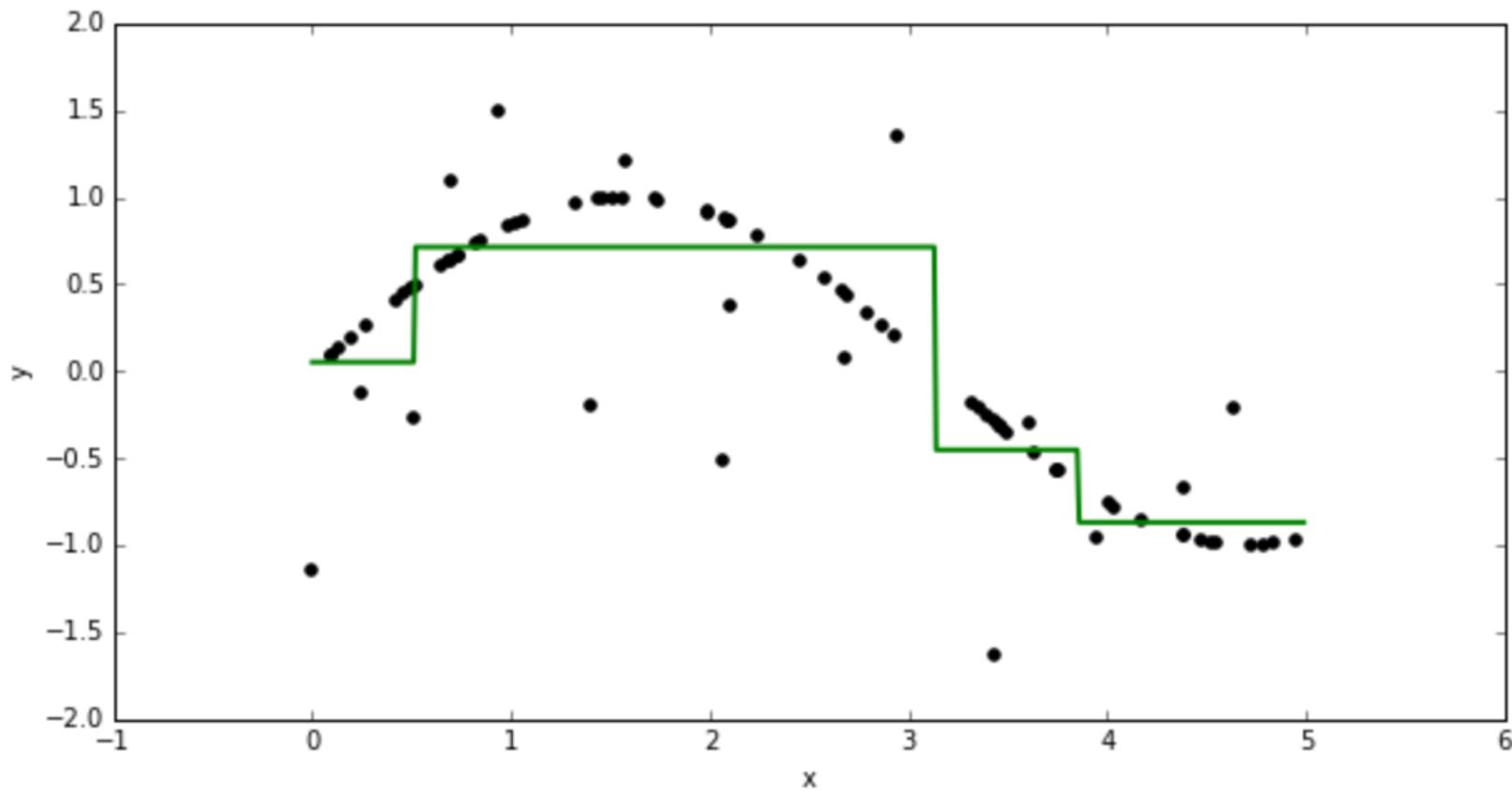
# Классификация



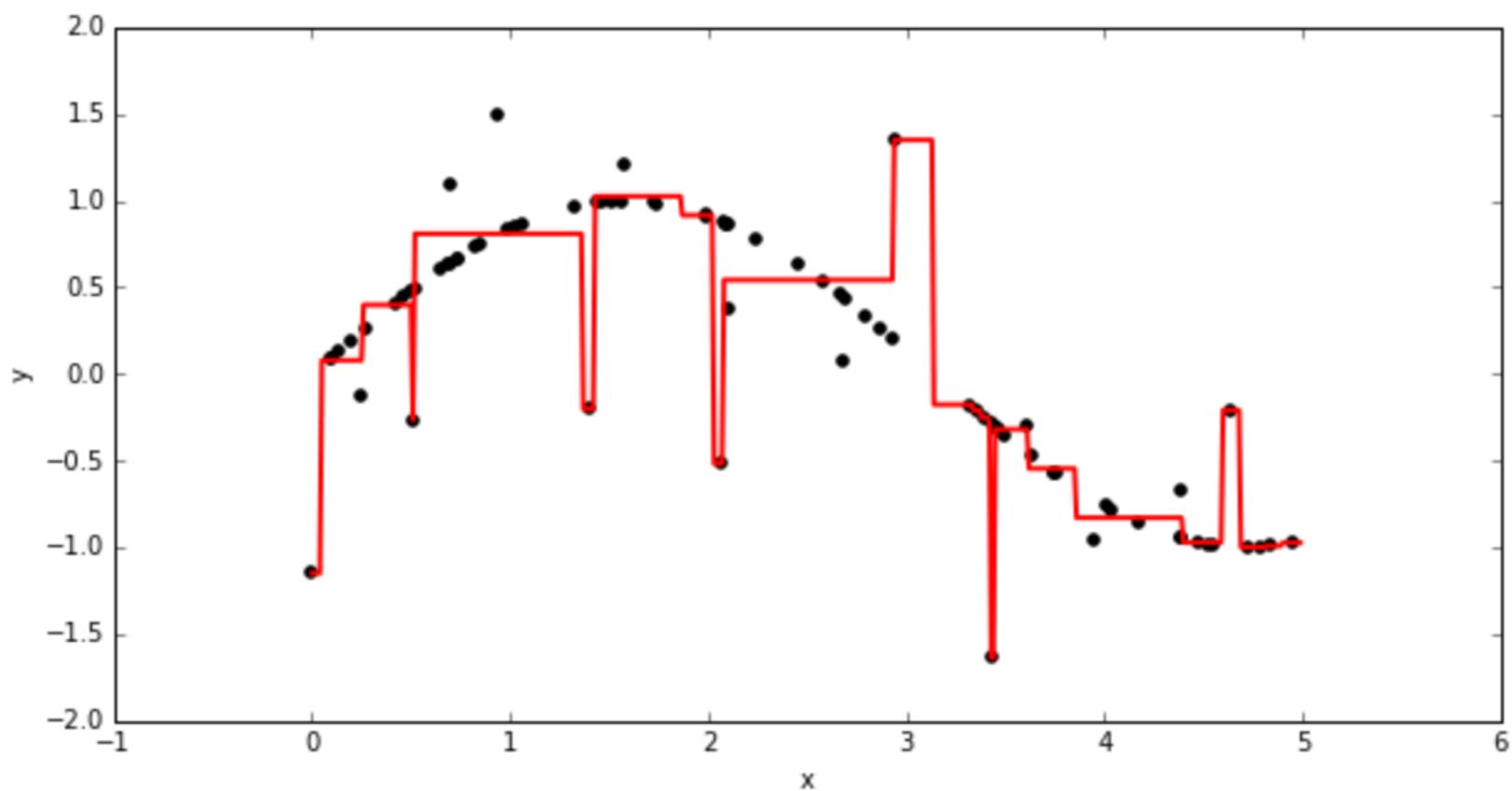
# Классификация



# Регрессия



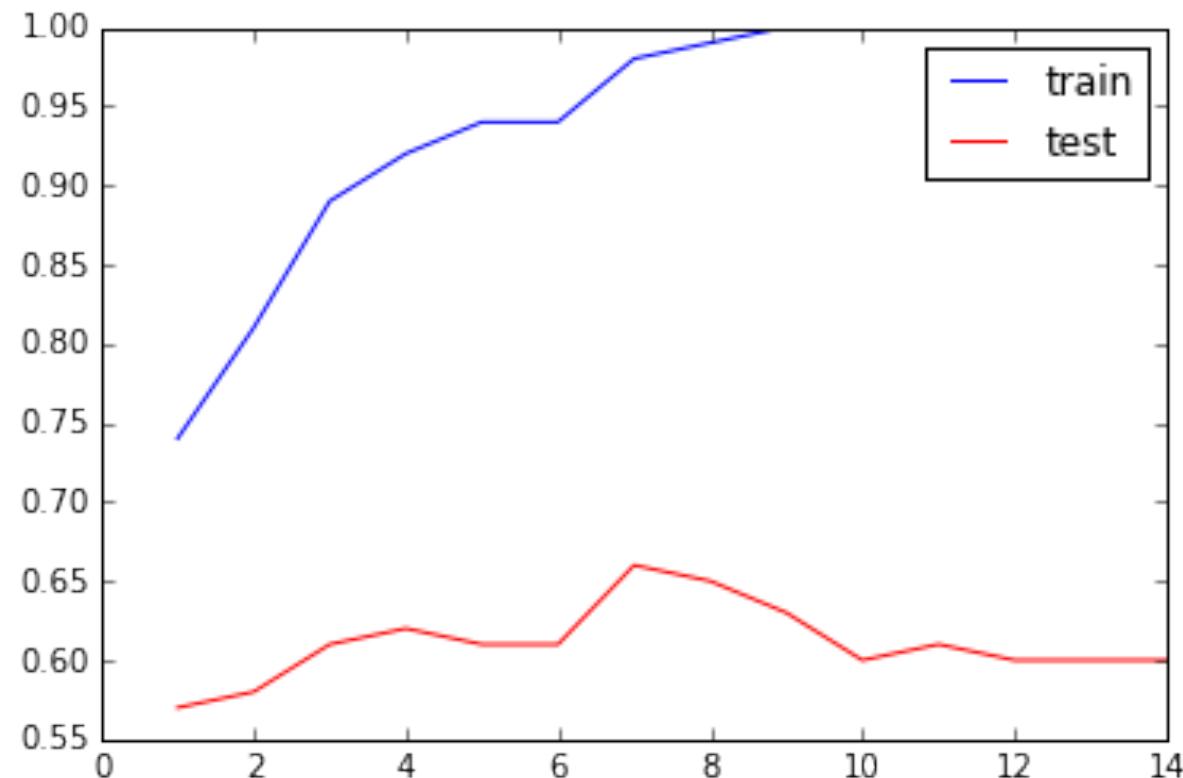
# Регрессия



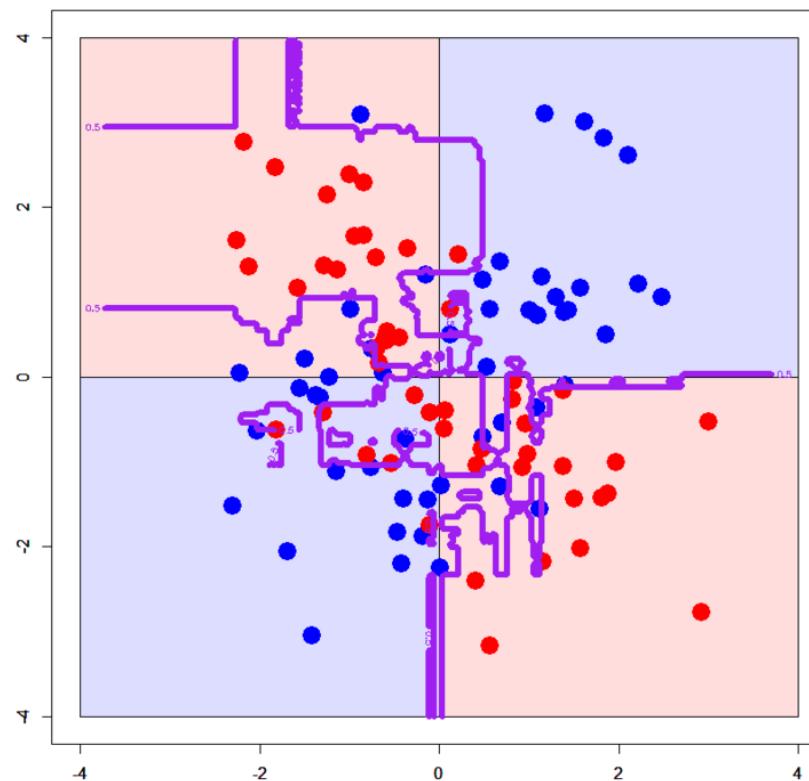
# Решающие деревья

- Восстанавливают сложные закономерности
- Могут построить сколь угодно сложную поверхность
- Чем больше глубина — тем сложнее поверхность
- Склонны к переобучению

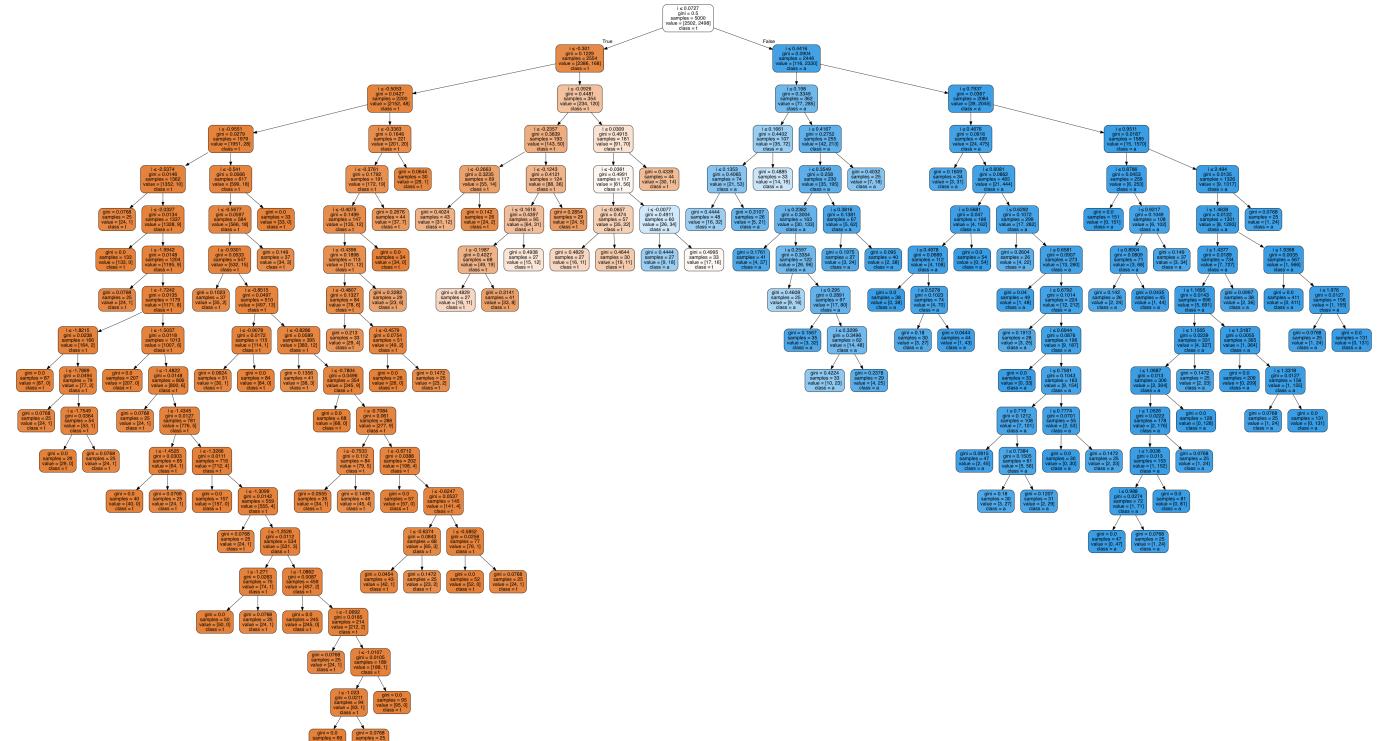
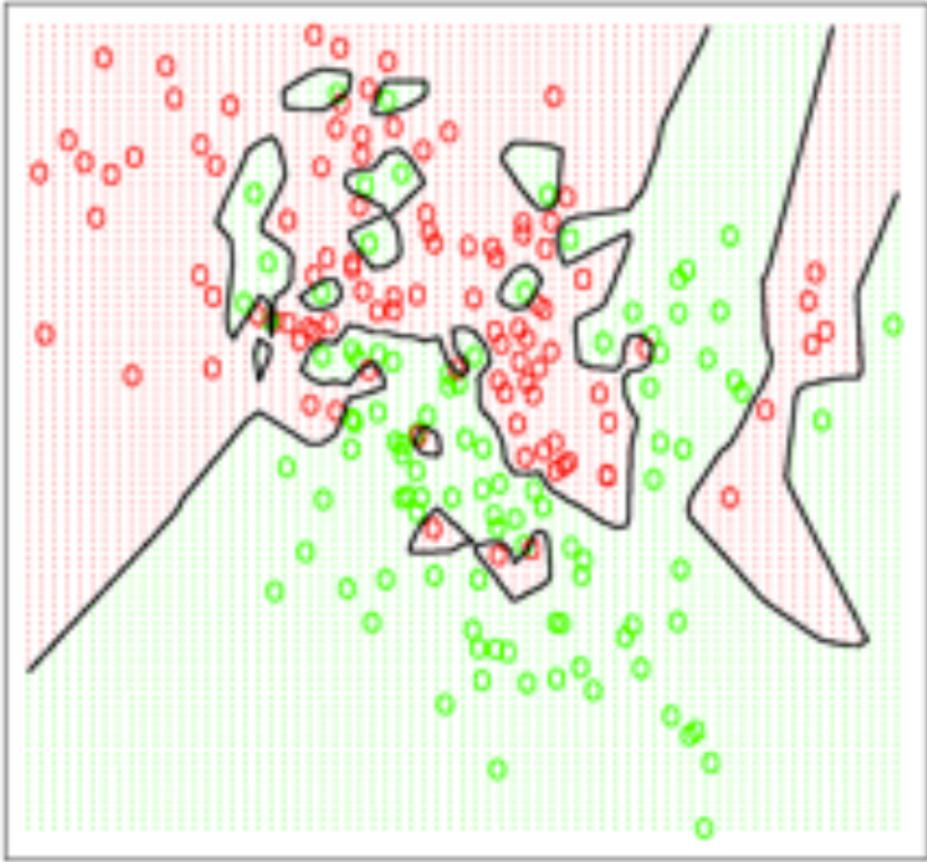
# Глубина деревьев



# Переобучение деревьев



# Переобучение деревьев



# Переобучение деревьев

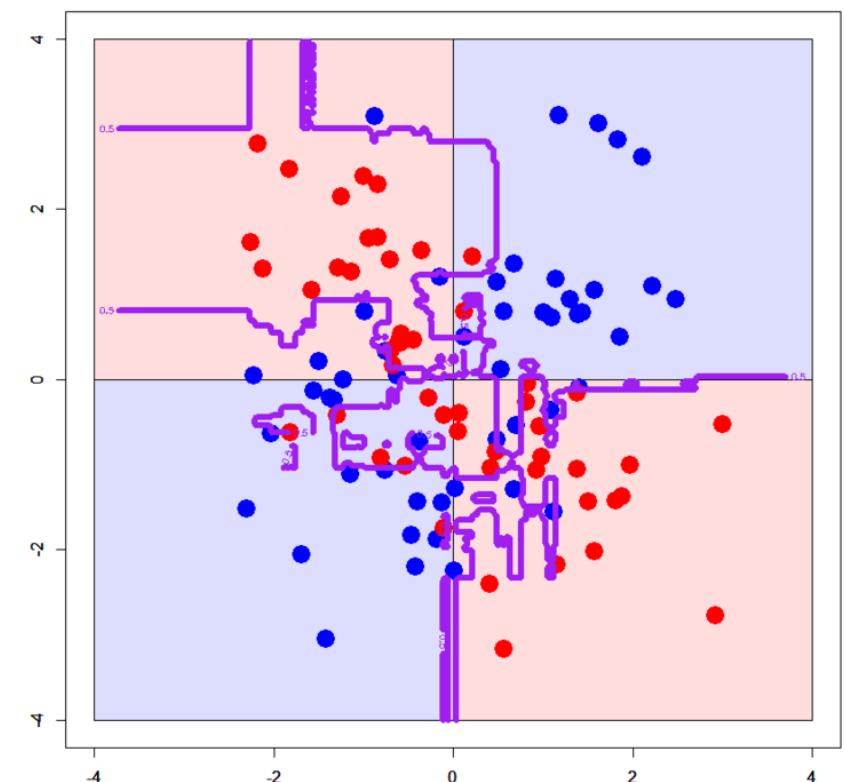
- Дерево может достичь нулевой ошибки на любой выборке
- Как правило, такое дерево окажется переобученным
- Выход — ограничивать глубину или число объектов в листе

# Критерий останова

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

# Критерий останова

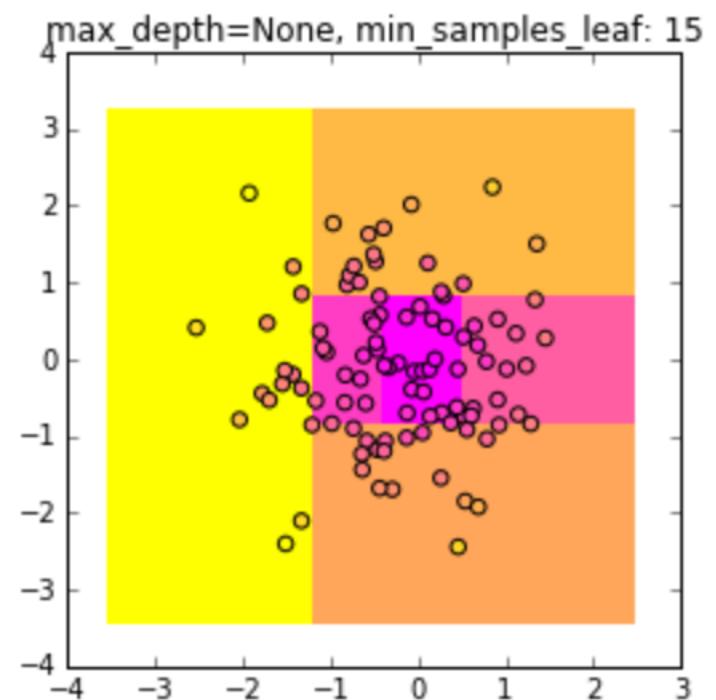
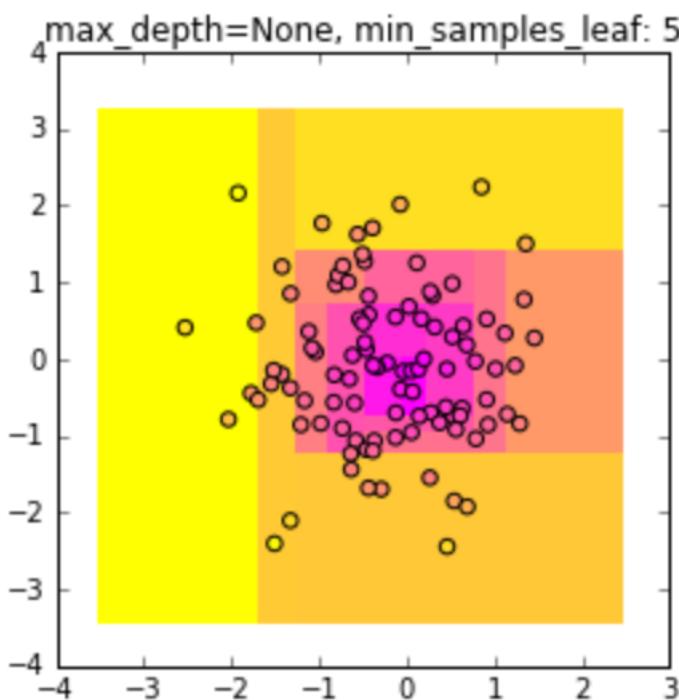
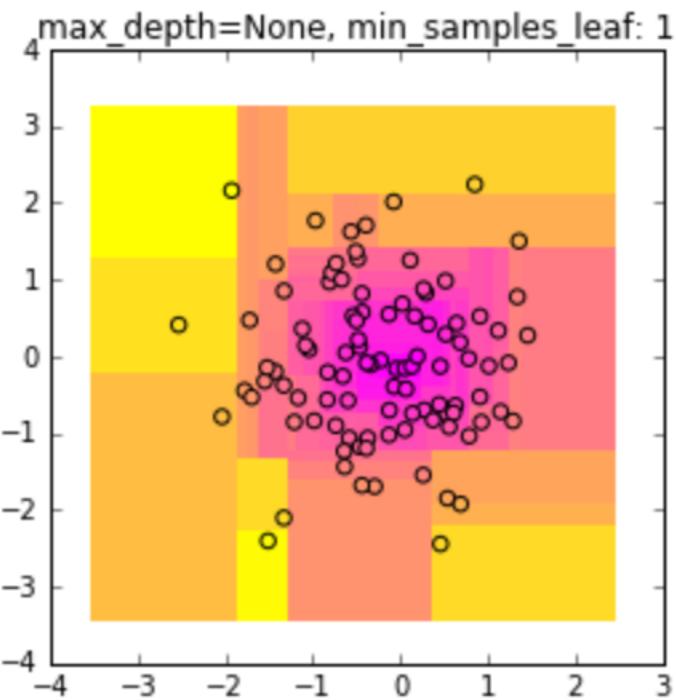
- Все объекты в вершине относятся к одному классу
- Простое условие
- Но приводит к переобучению



# 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:  $n = 5$

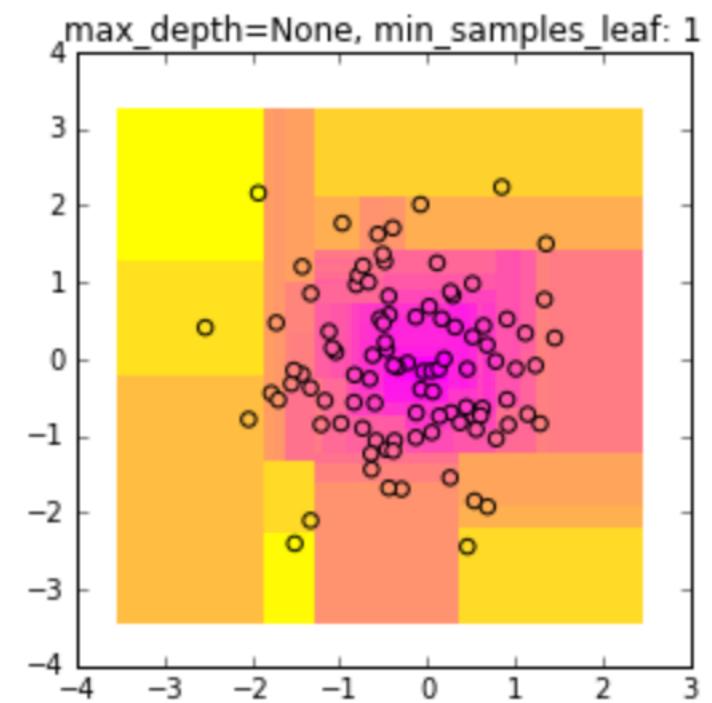
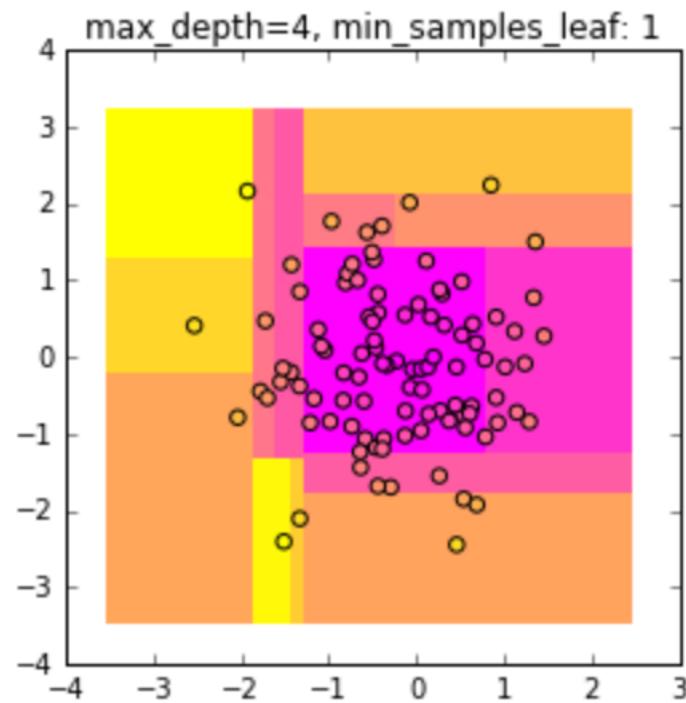
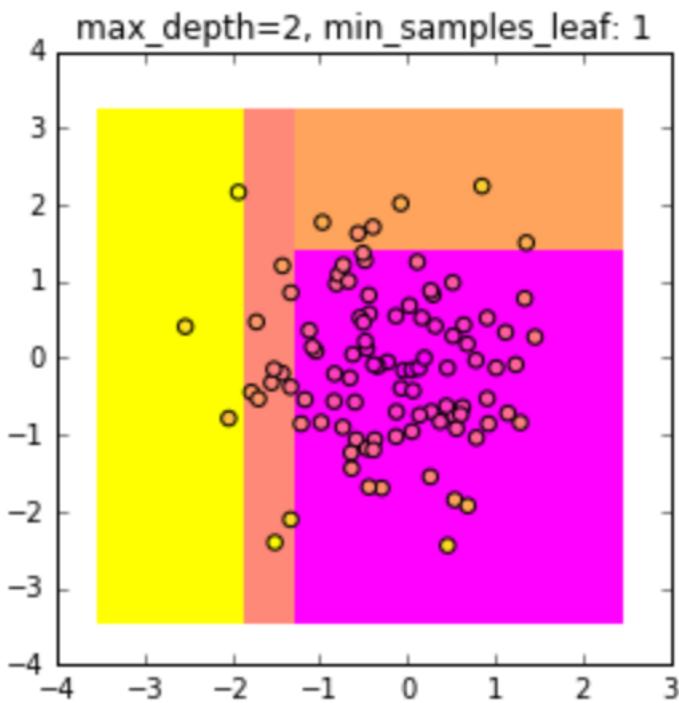
# Число объектов в листе



# Глубина дерева

- Ограничение на глубину
- Достаточно грубый критерий

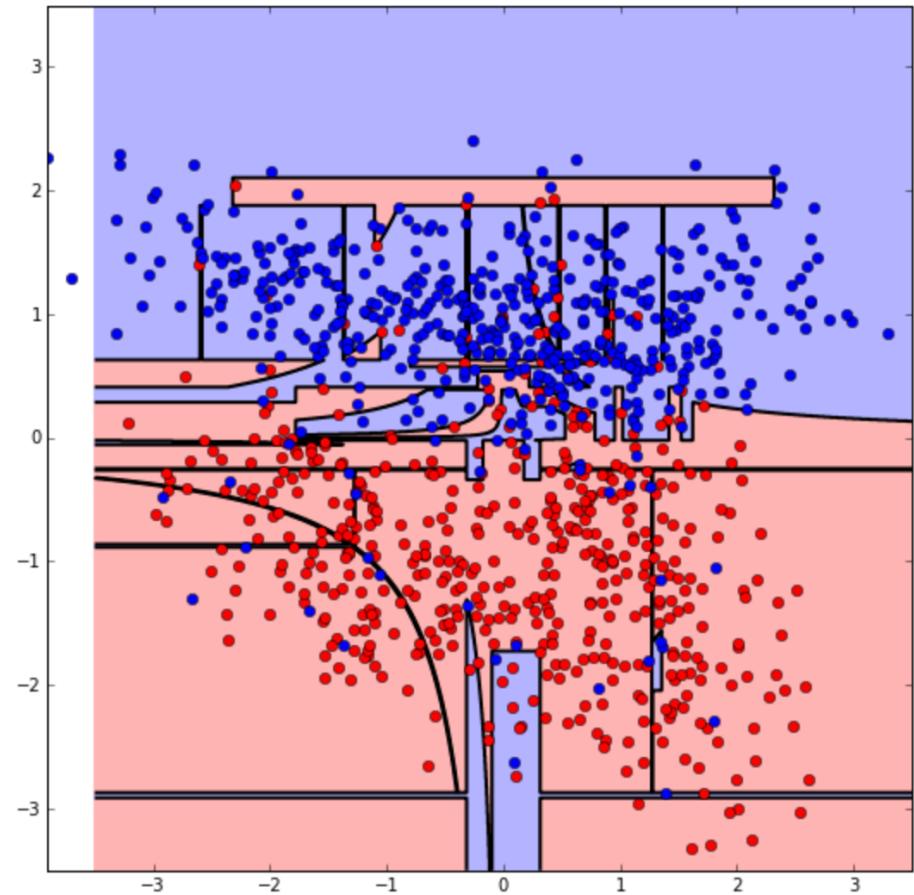
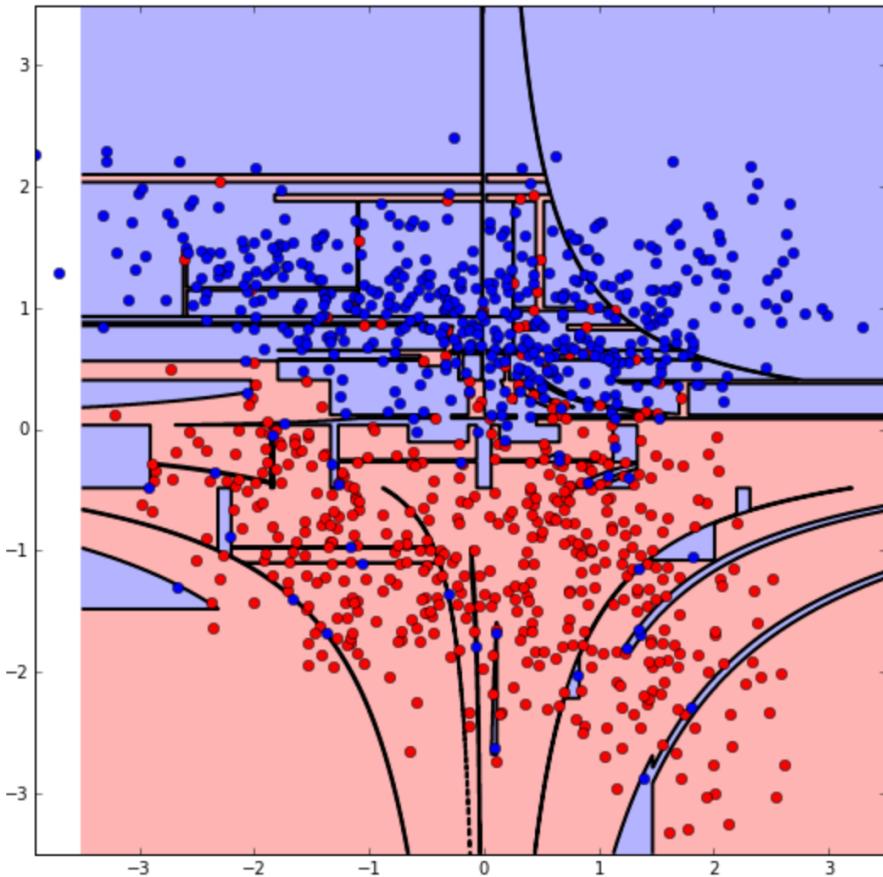
# Глубина дерева



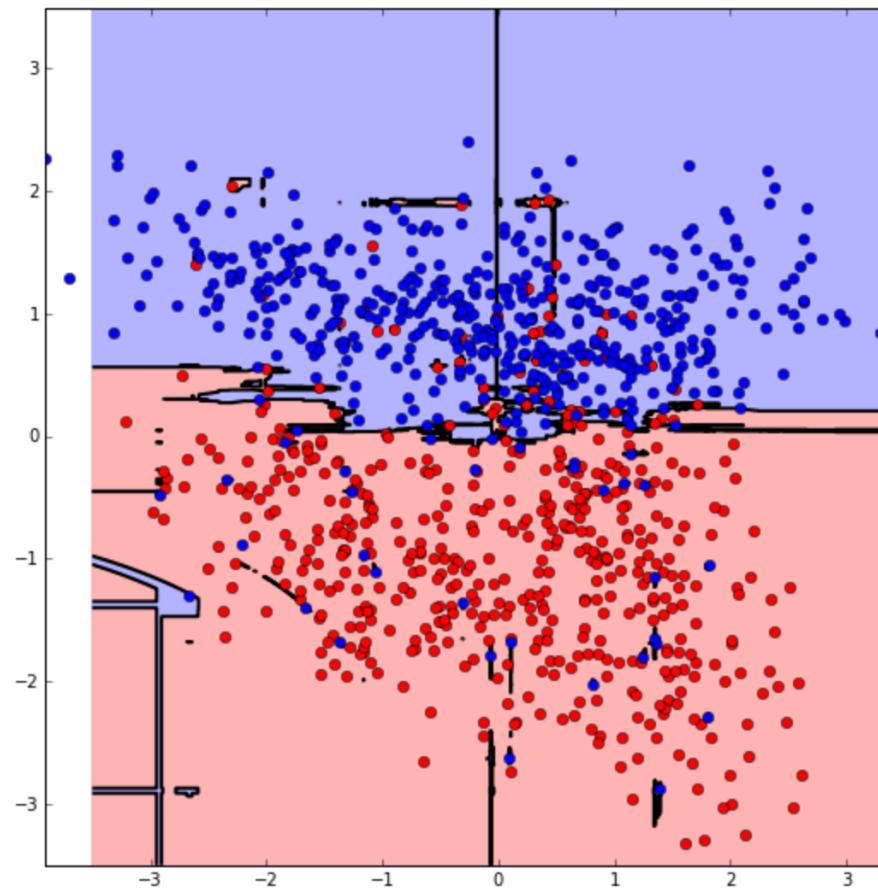
# Неустойчивость деревьев

- 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



# Резюме

- Решающее дерево — очень мощная модель
- Обучение эвристическое
- Много тонкостей с переобучением
- Обычно используется в композициях — на следующей лекции