Введение в анализ данных

Лекция 10

Решающие деревья

Евгений Соколов

sokolov.evg@gmail.com

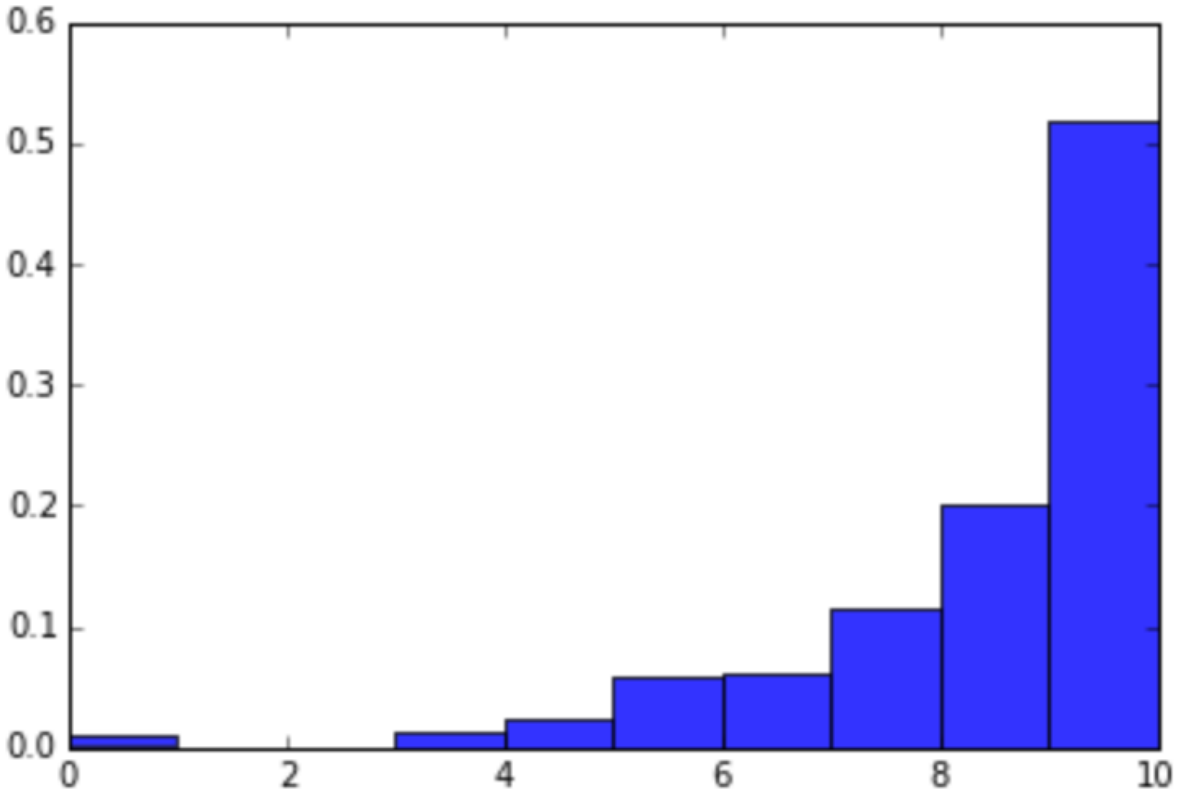
НИУ ВШЭ, 2016

Немного данных

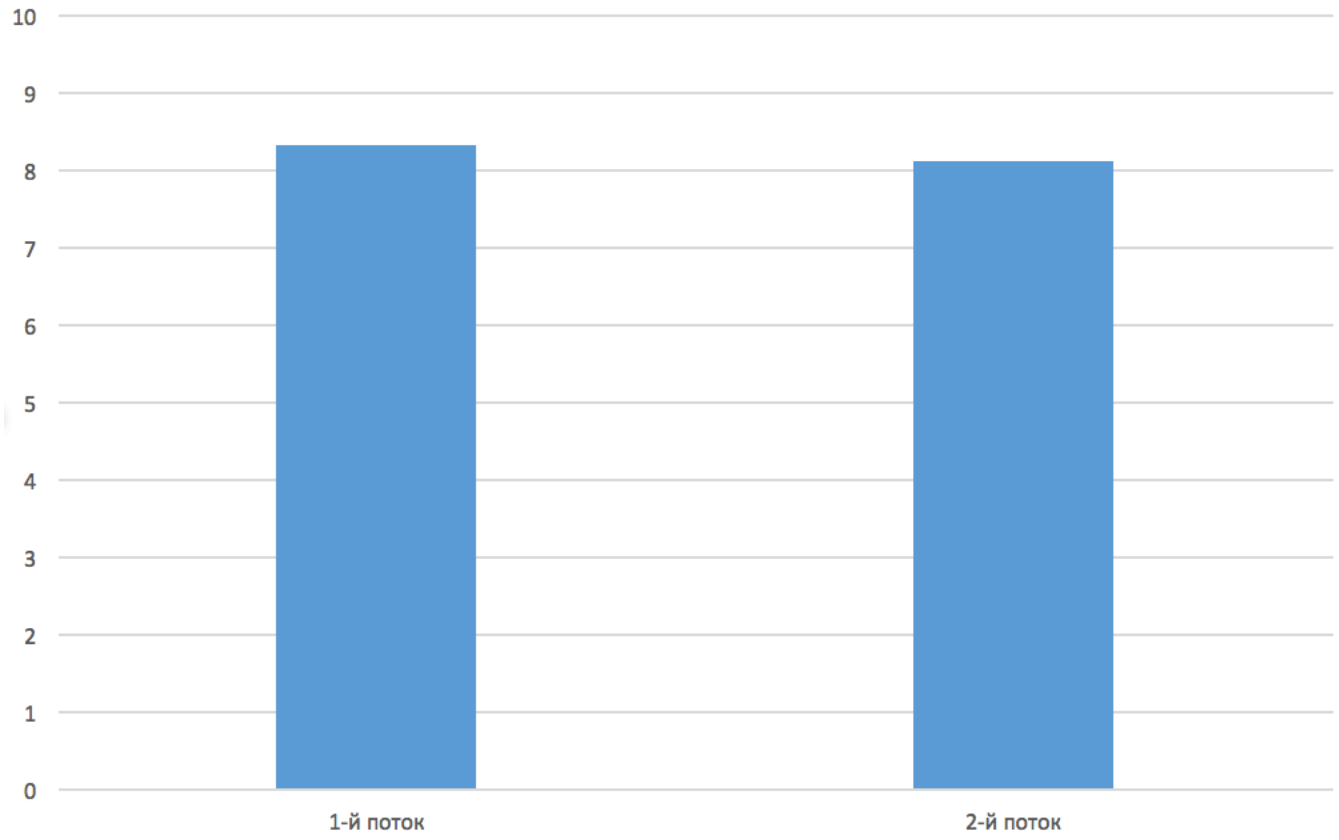
- Коллоквиум сдавали 225 студентов

- 4 группы имеют средний балл выше 9
- Все они — с экономического факультета

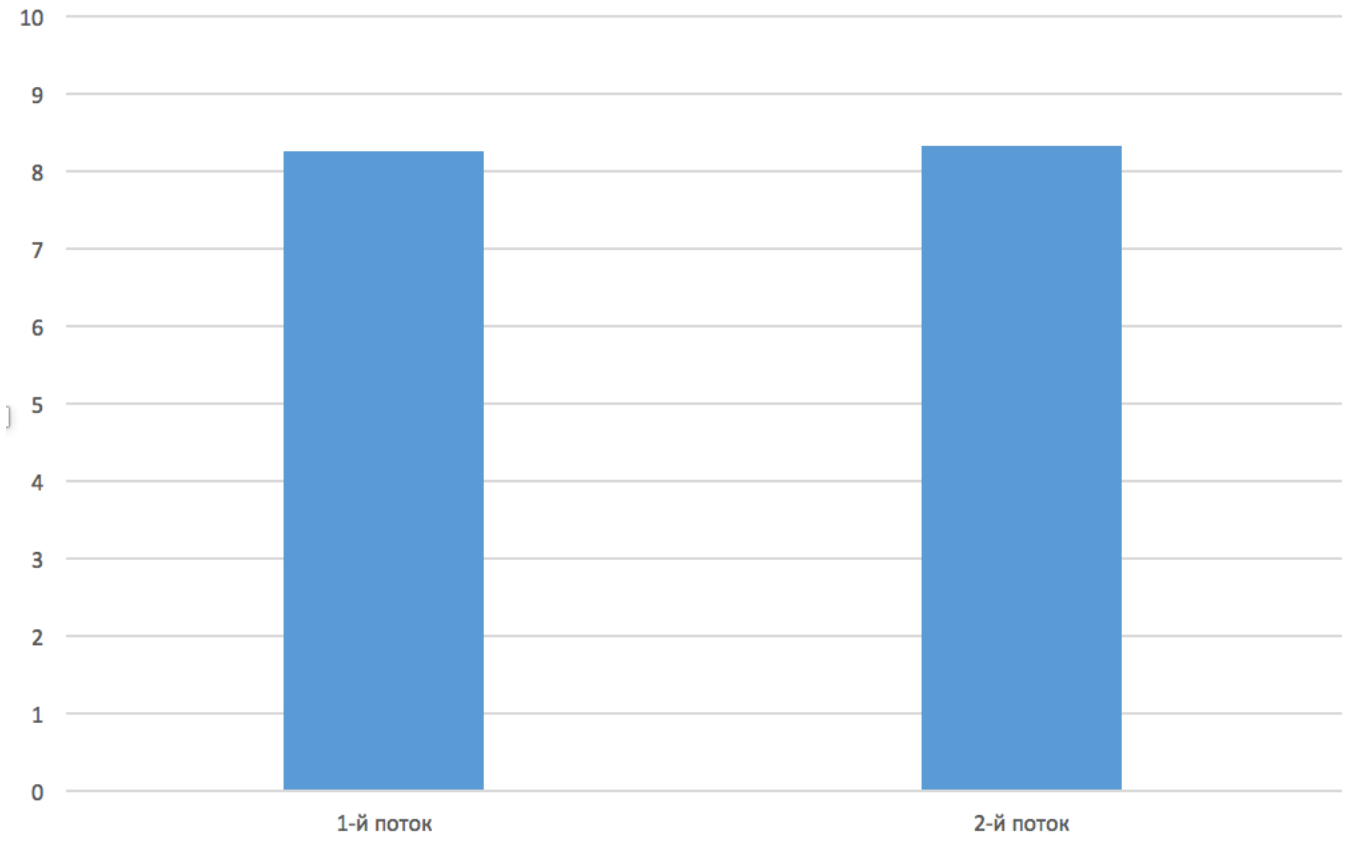
Немного данных



Немного данных



Немного данных



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:

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

Пример

- С кубическими признаками будет ещё лучше
- Как интерпретировать признак $\text{этаж} * (\text{число комнат})^2$?
- Всего таких признаков 20

Пример

- Можно бинаризовать признаки: $[x^j > t]$
- $(\text{этаж} > 1), (\text{этаж} > 2), \ldots, (\text{этаж} > 30)$
- Признаков будет на порядки больше
- Легче интерпретировать:

$$-2[\text{этаж} > 3][\text{площадь} < 40][\text{число комнат} < 3]$$

- Можно использовать L_1-регуляризацию

Логические правила

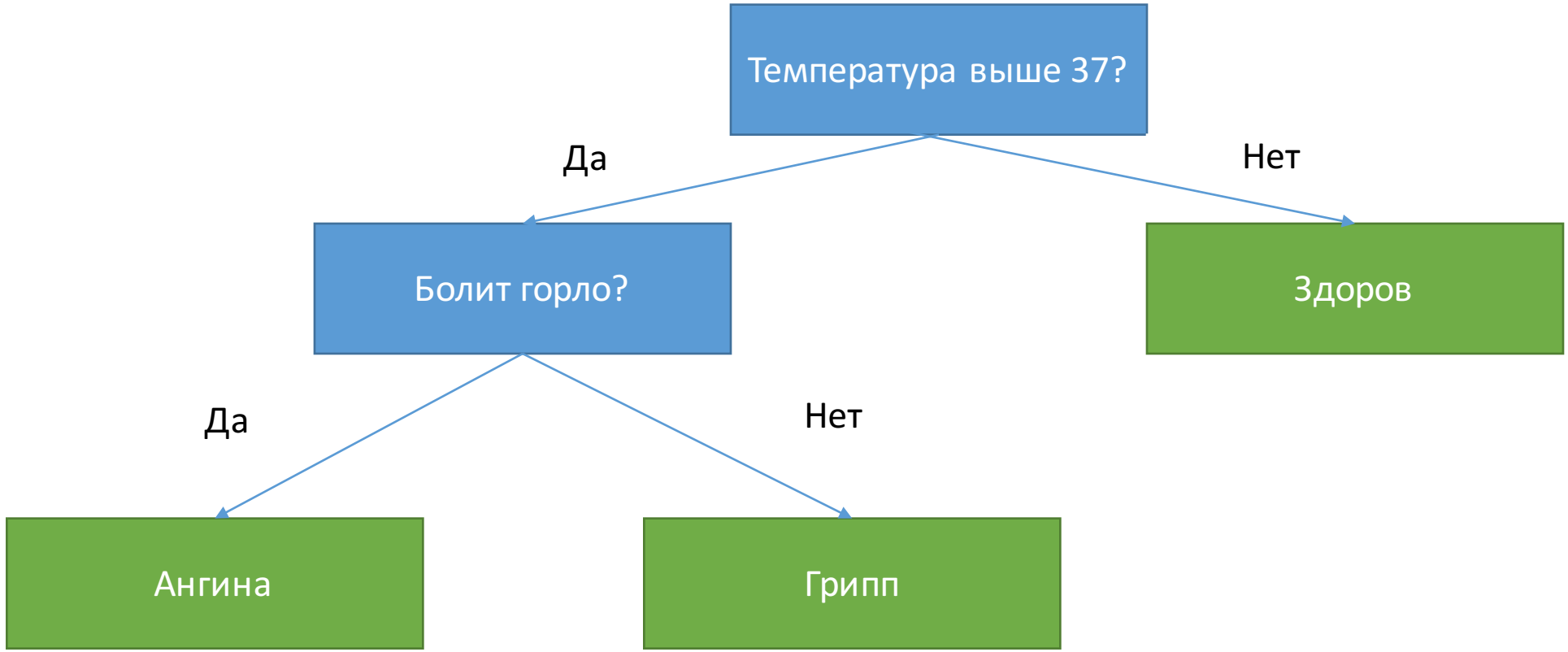
$$[\text{этаж} > 3][\text{площадь} < 40][\text{число комнат} < 3]$$

- Легко объяснить заказчику (если ≤ 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



Принятие решений

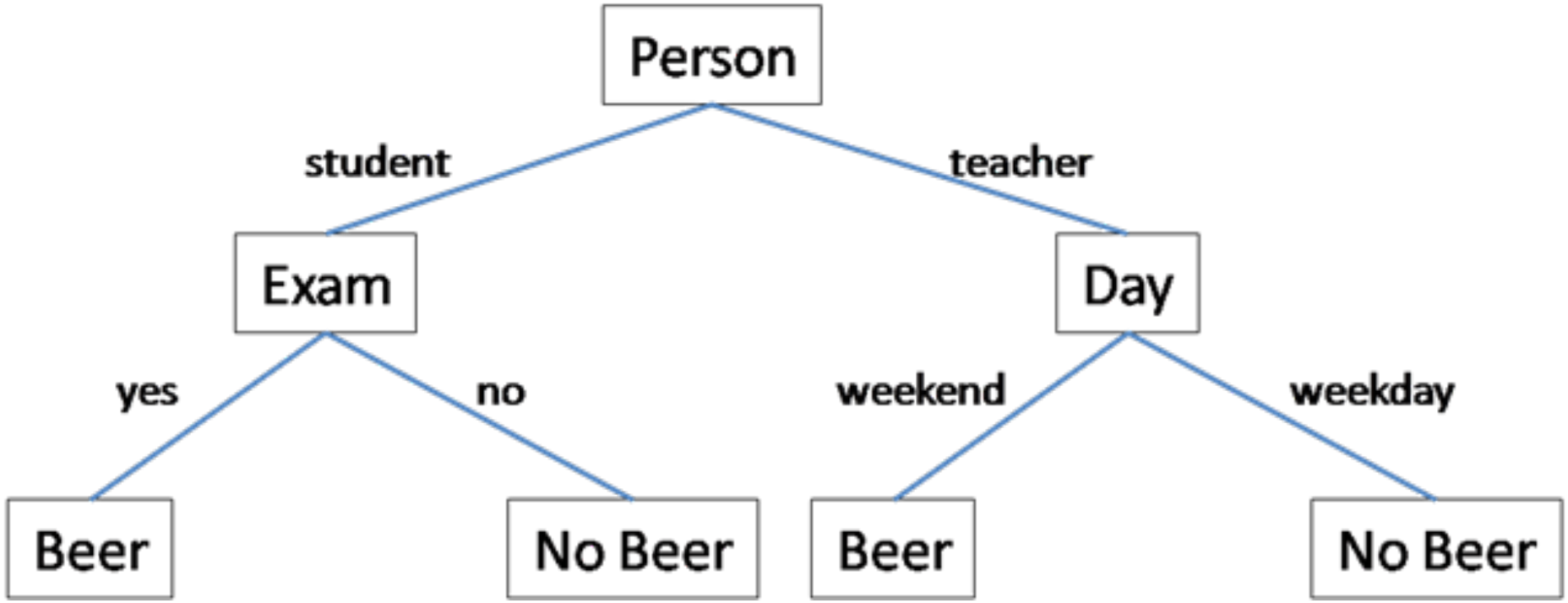
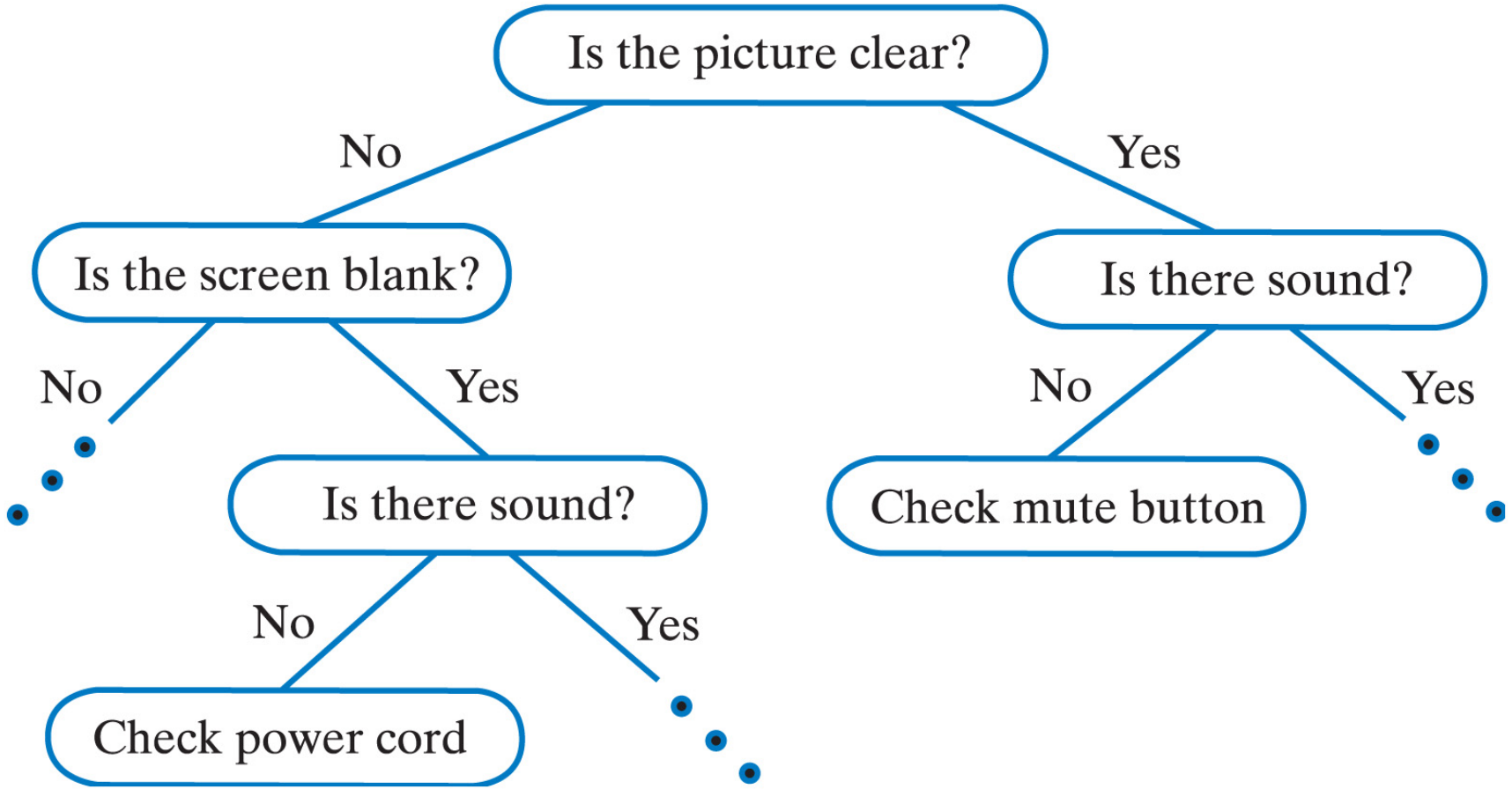
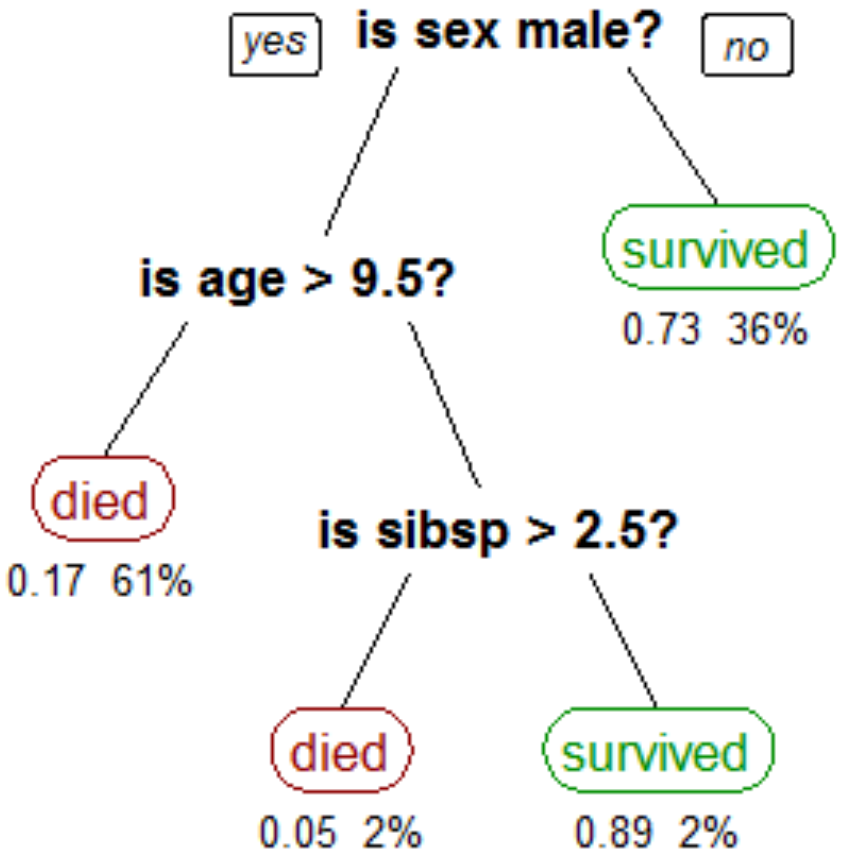


Схема диалога с клиентом

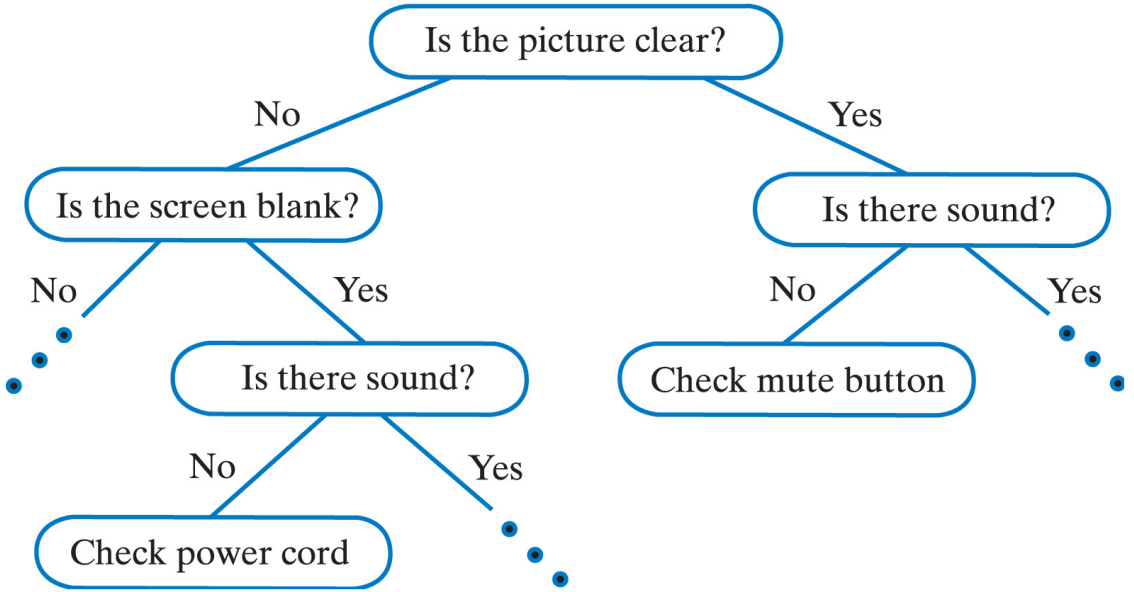


Пассажиры Титаника



Решающее дерево

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



Условия

- Самые популярные варианты:

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

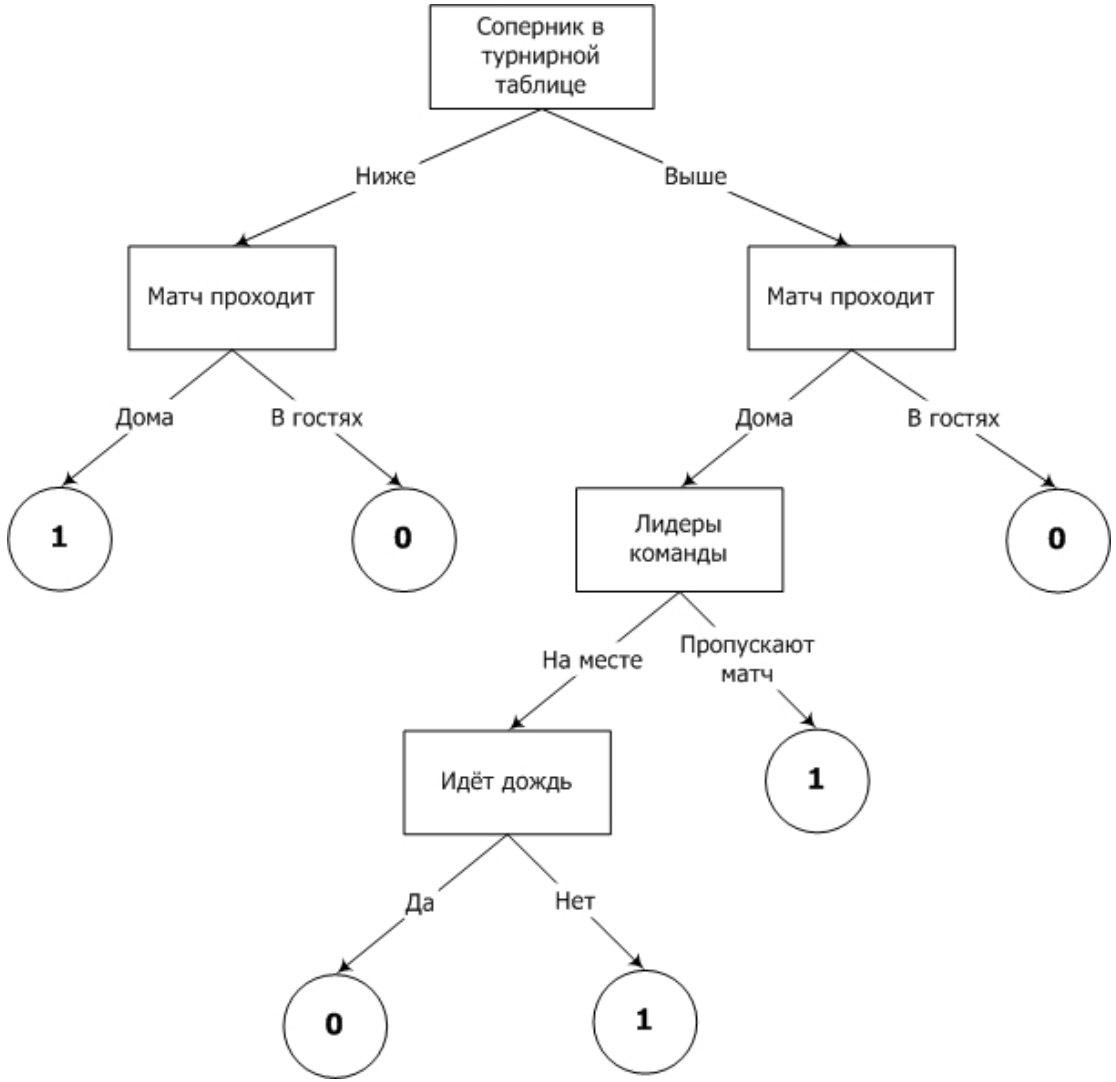
Примеры:

- $[\text{этаж} = 5]$
- $[\text{площадь} \leq 30]$

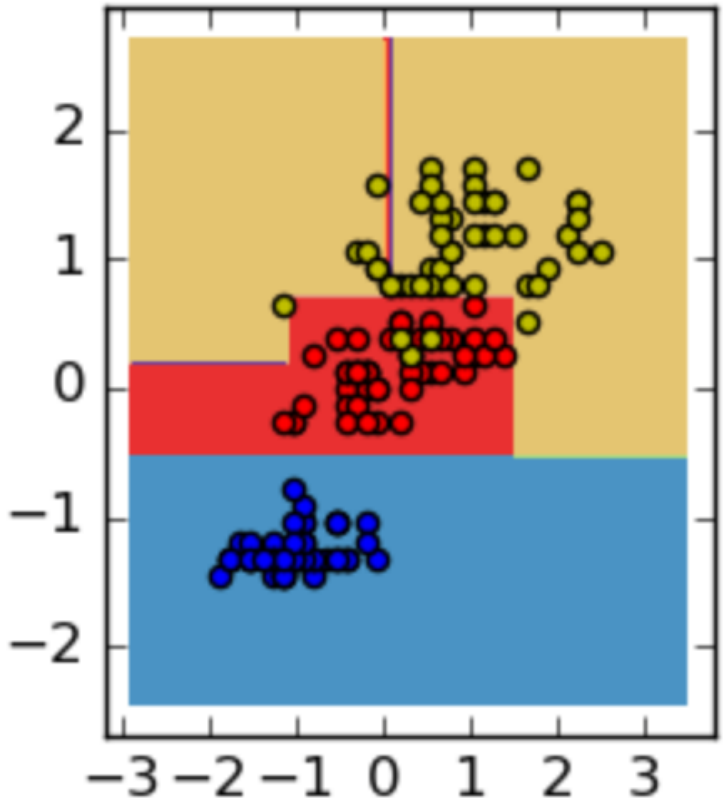
Прогноз в листе

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

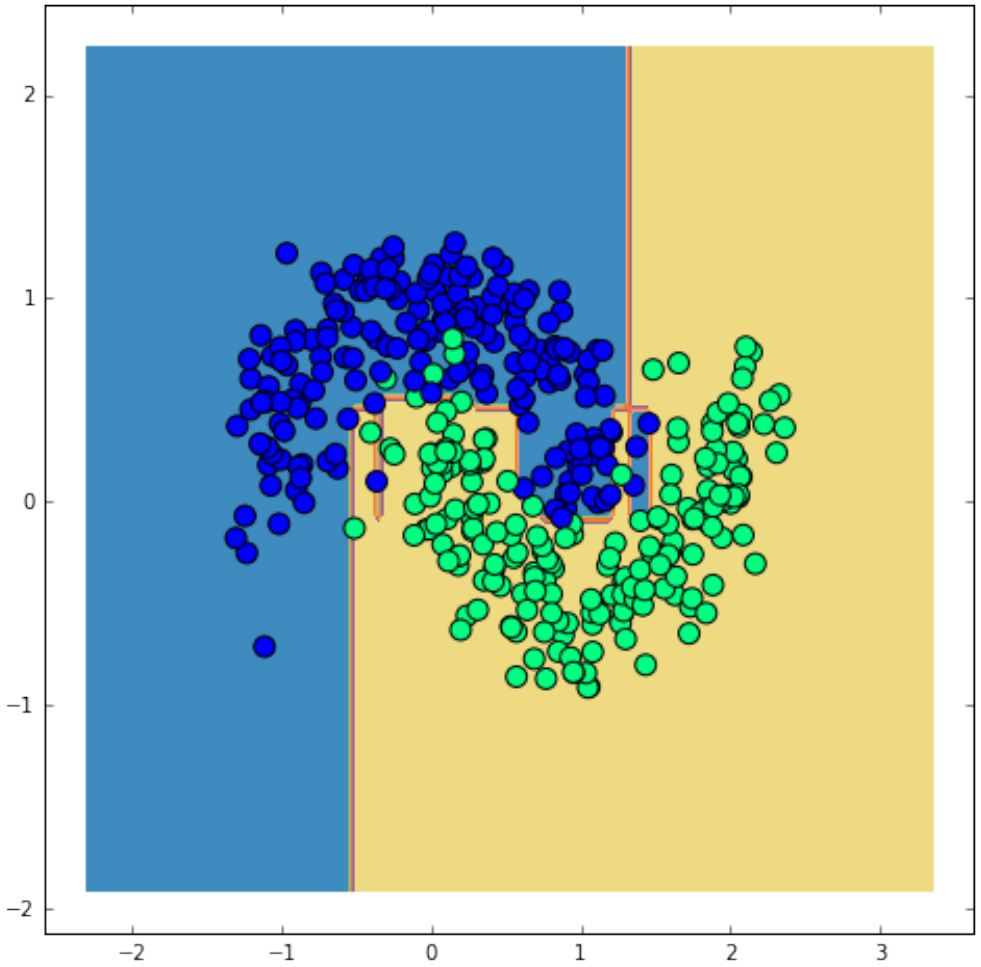
Исход футбольного матча



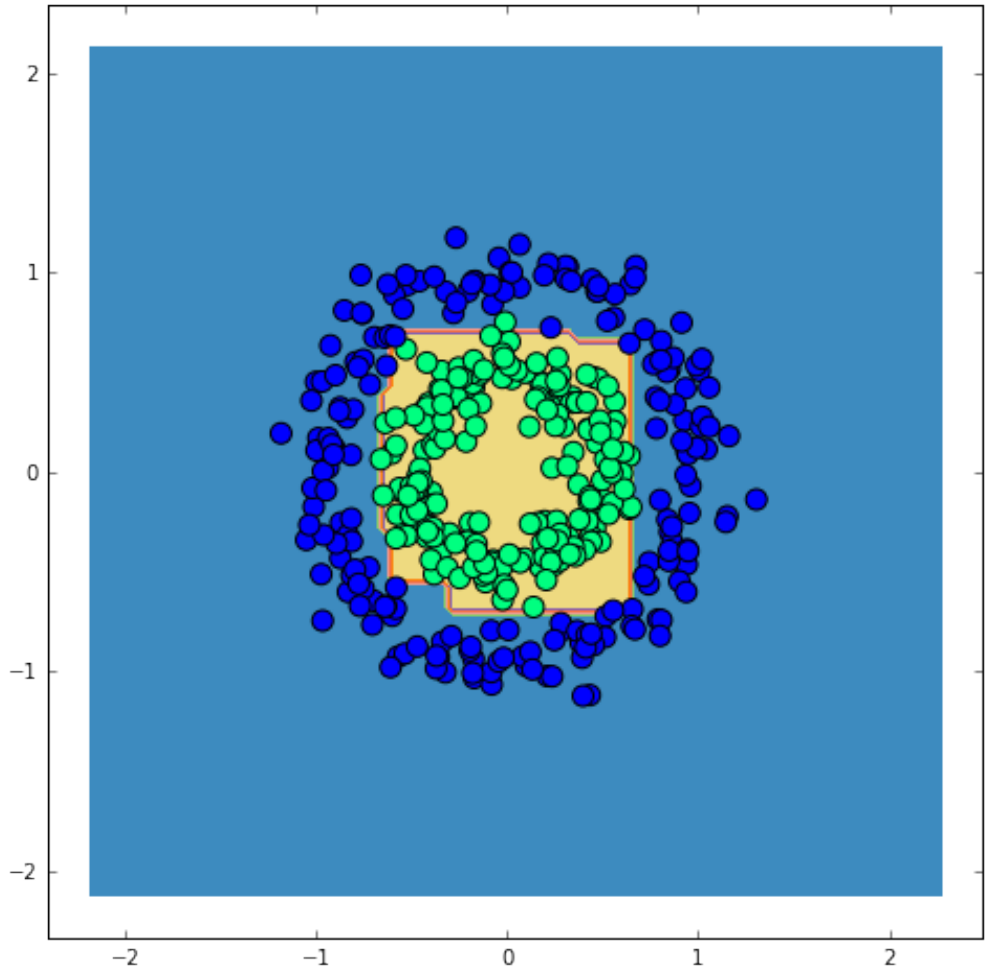
Классификация



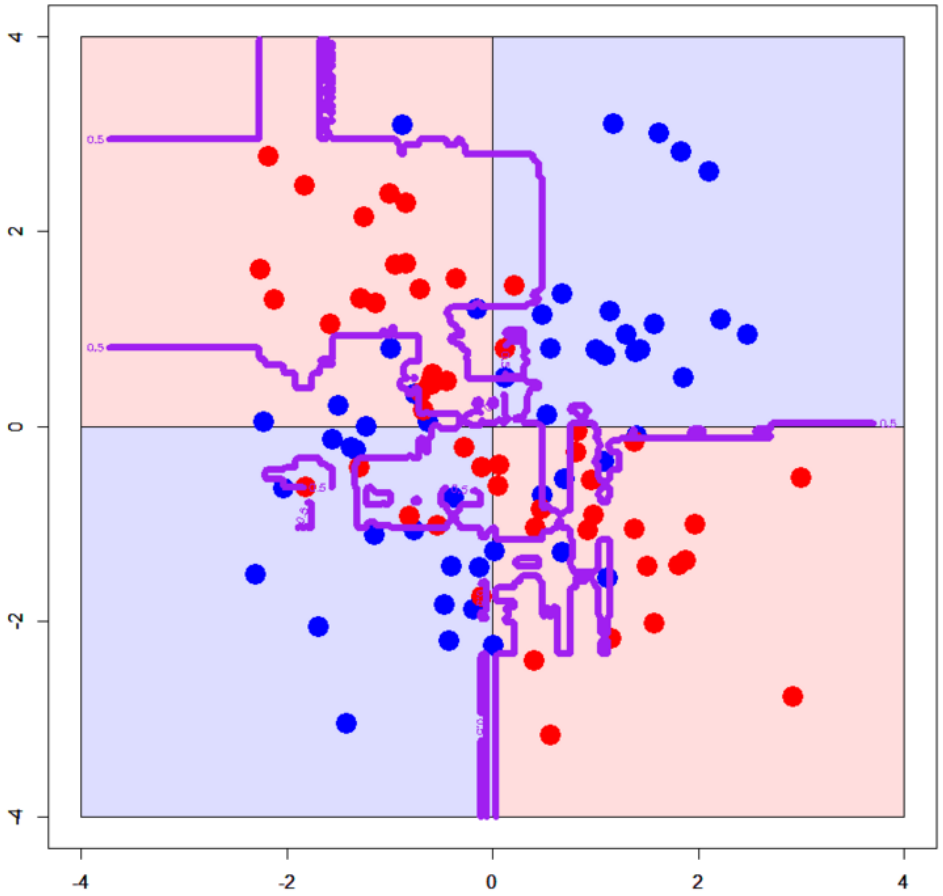
Классификация



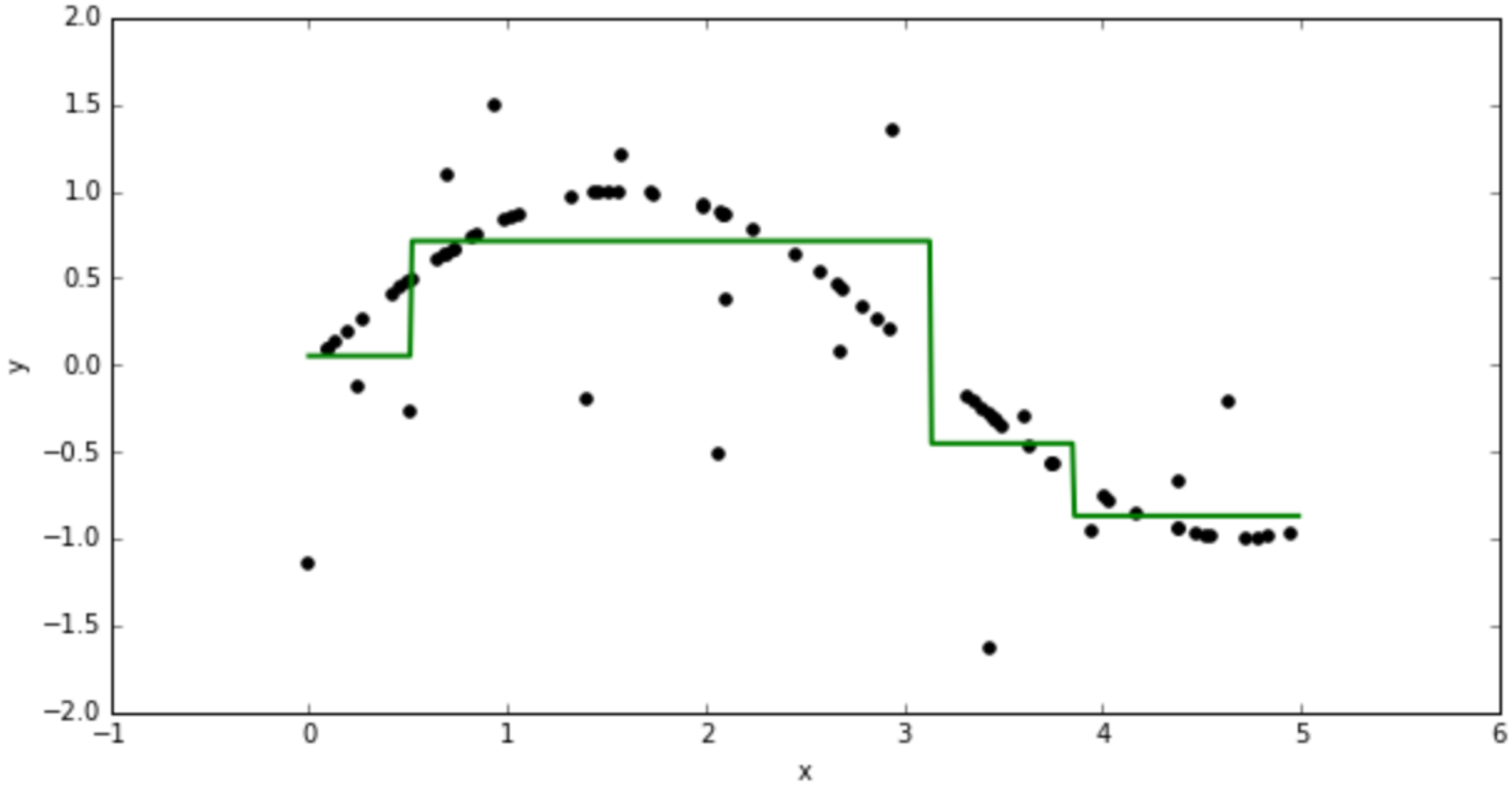
Классификация



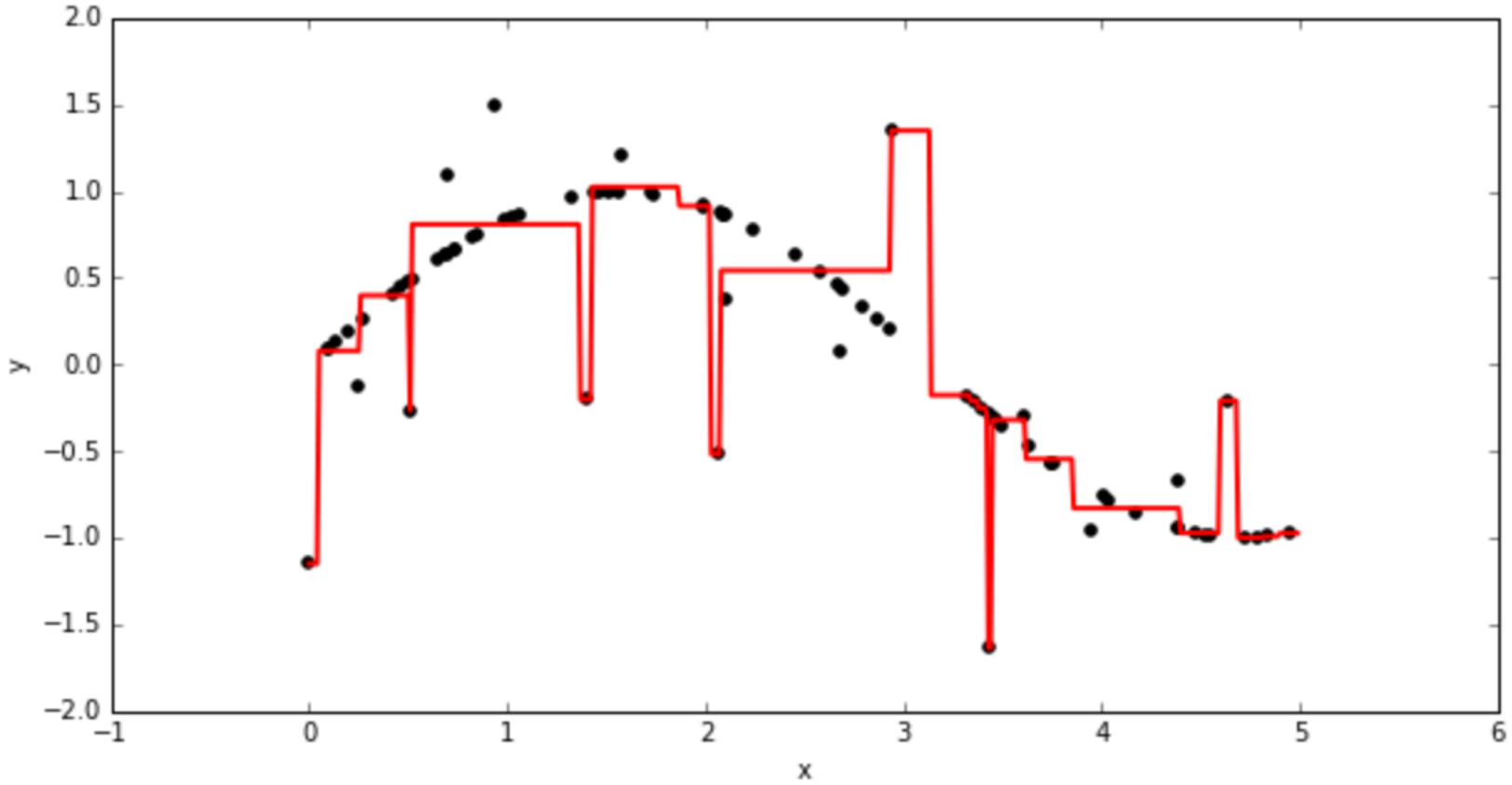
Классификация



Регрессия



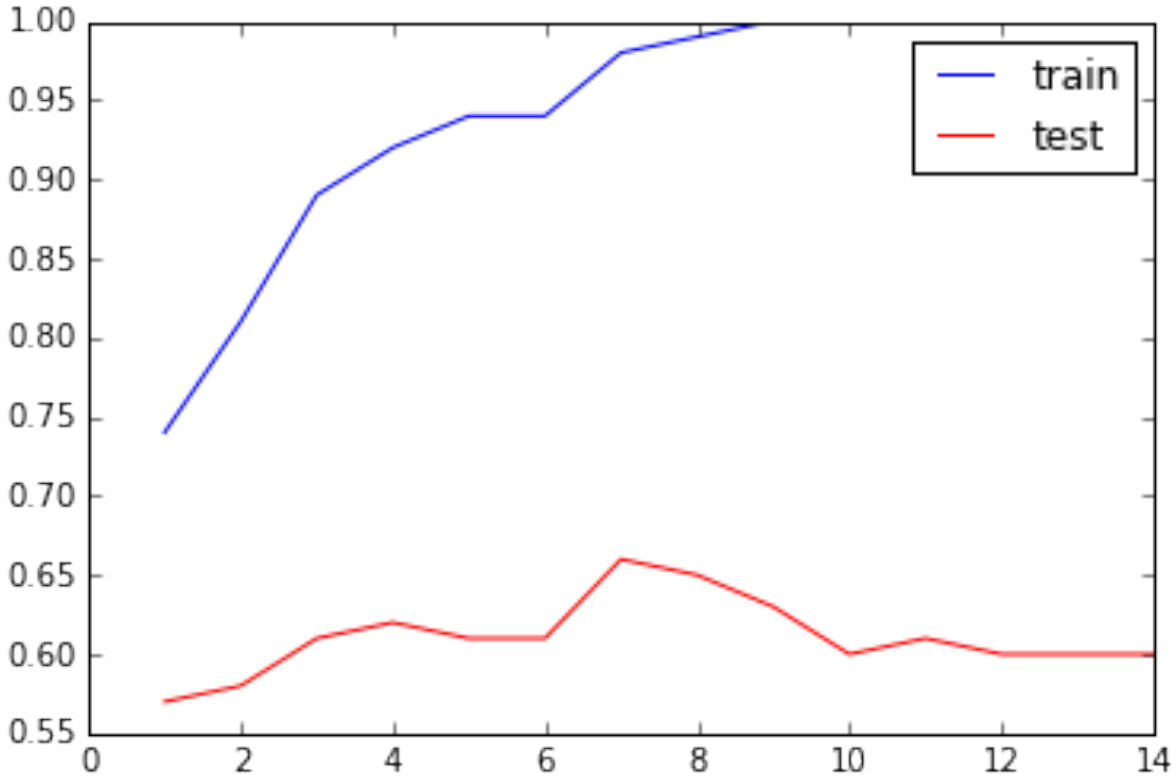
Регрессия



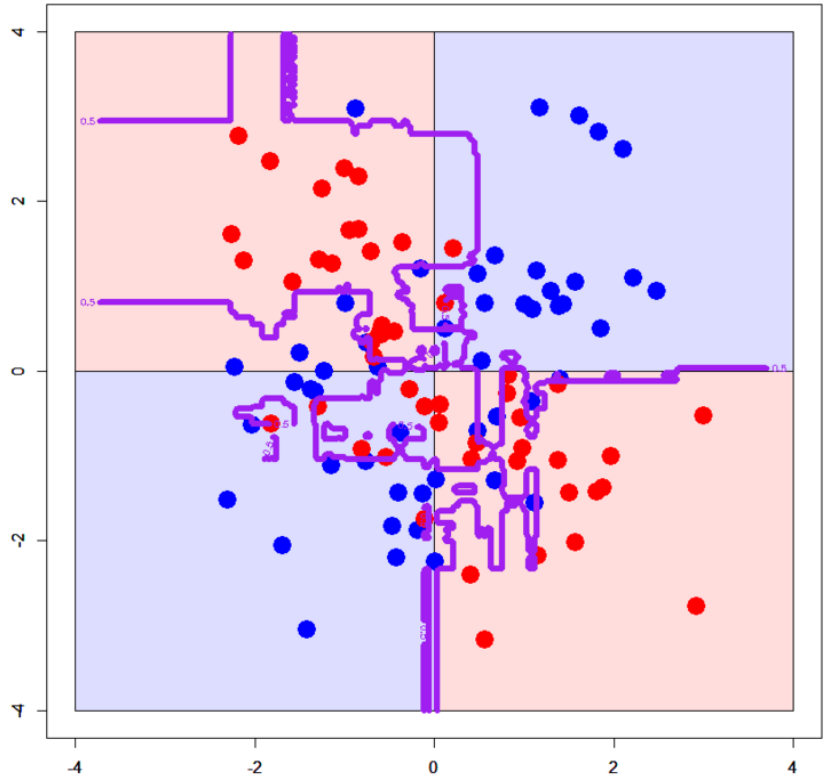
Решающие деревья

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

Глубина деревьев



Переобучение деревьев



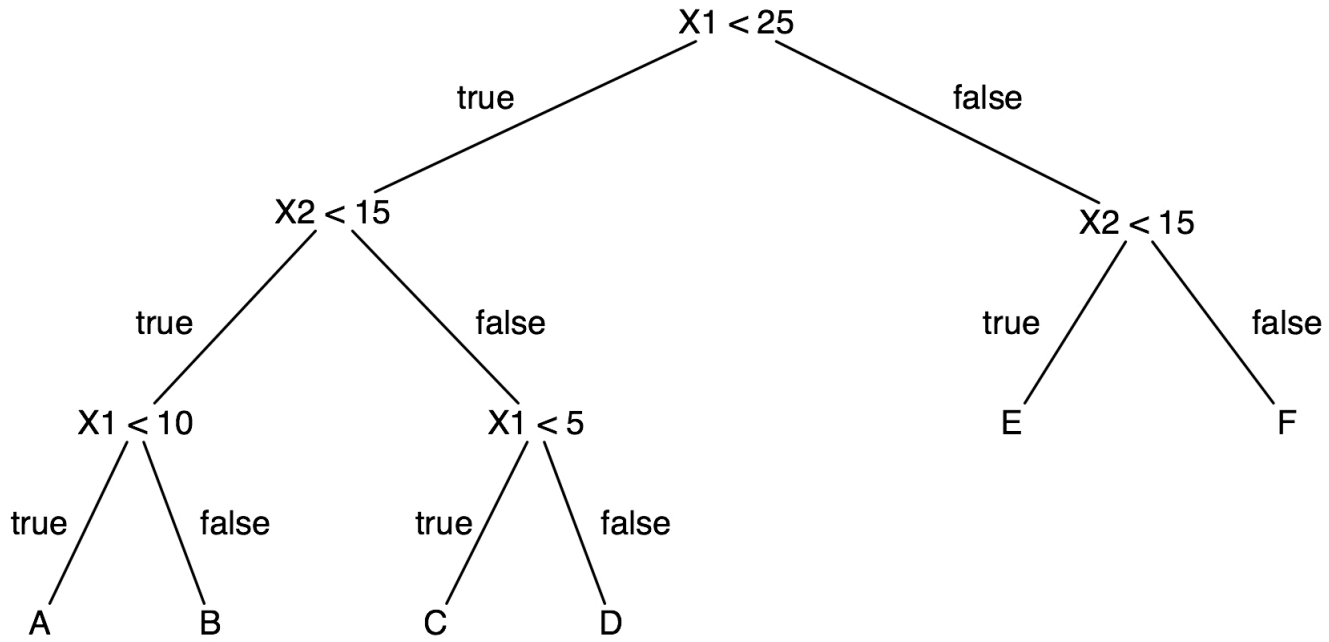
Переобучение деревьев

- Дерево может достичь нулевой ошибки на любой выборке
- Борьба с переобучением: минимальное дерево среди всех с нулевой ошибкой
- NP-полная задача

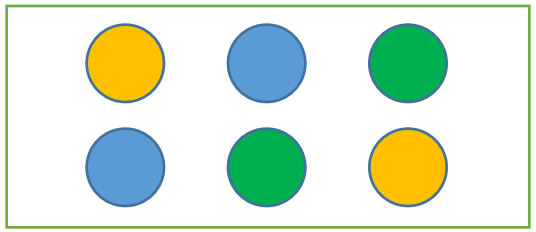
- Выход — жадное построение

Жадное построение

- Растим дерево от корня к листьям

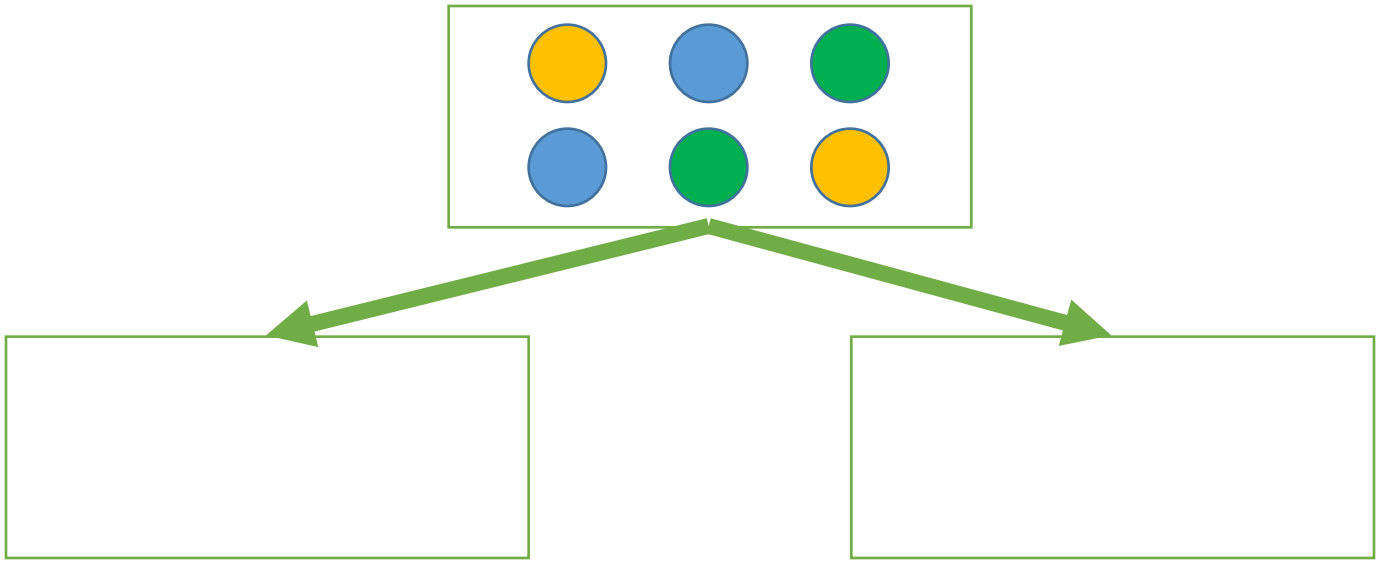


Жадное построение

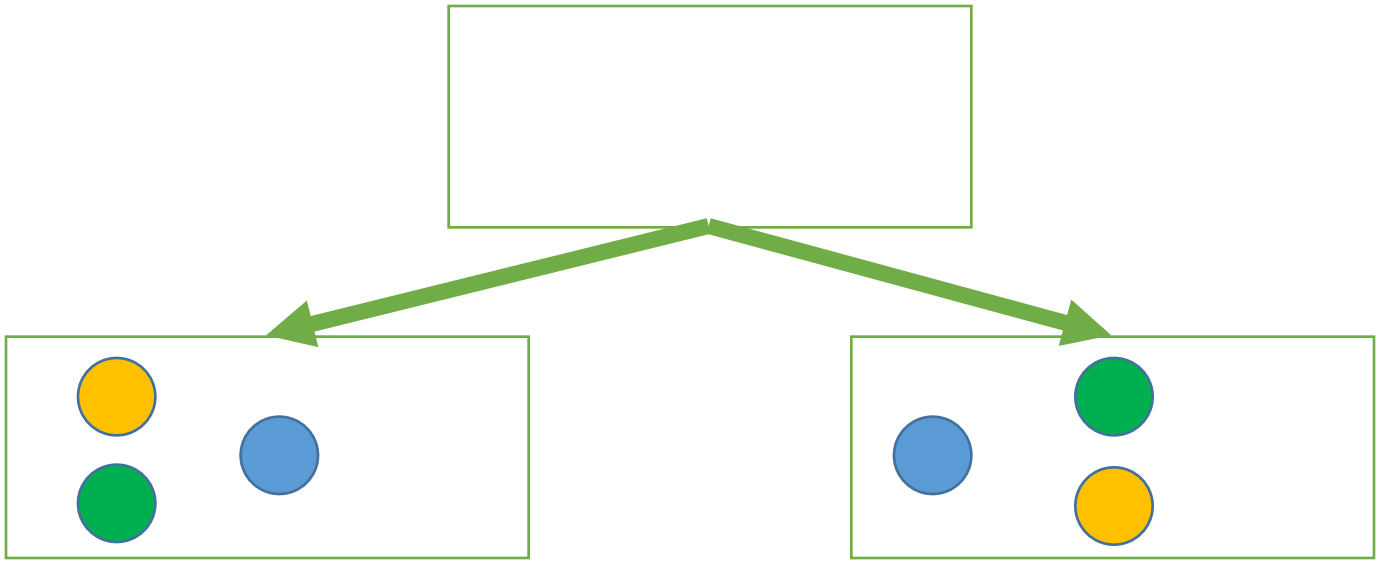


- Как разбить вершину?

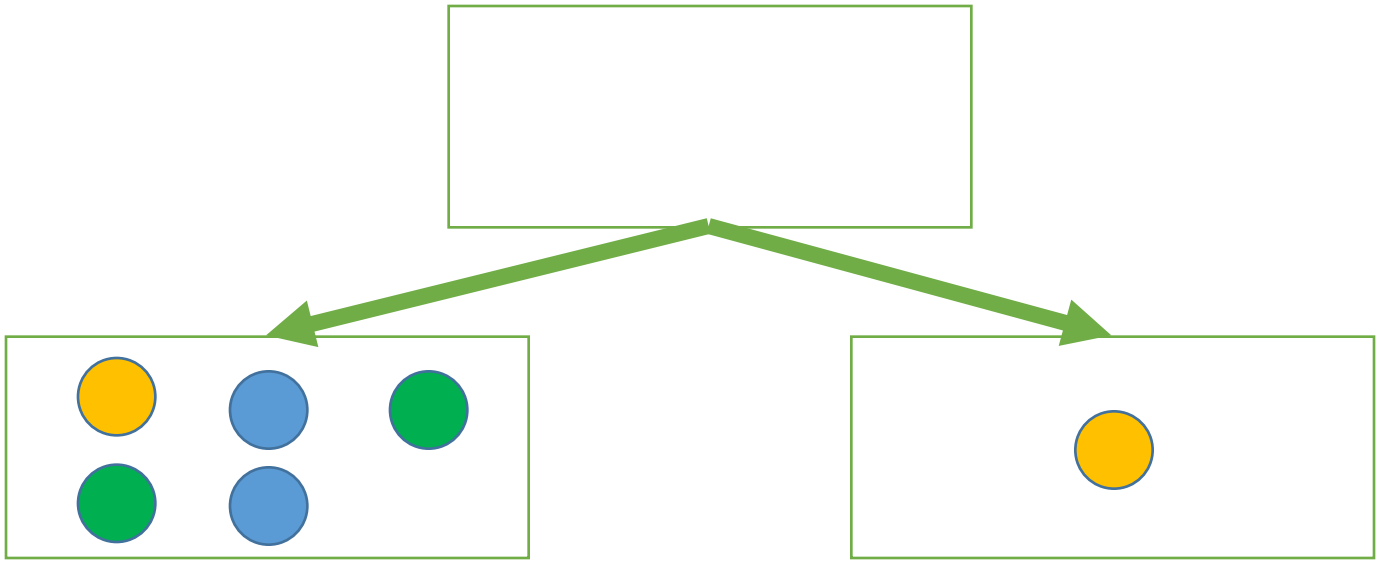
Жадное построение



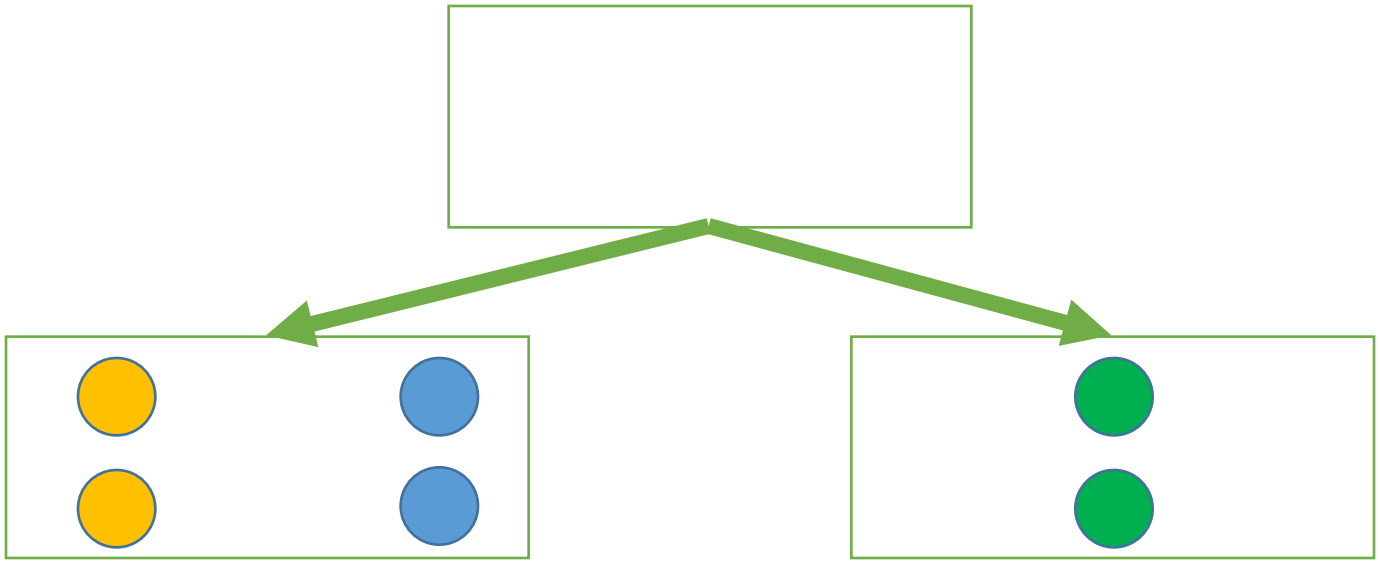
Жадное построение



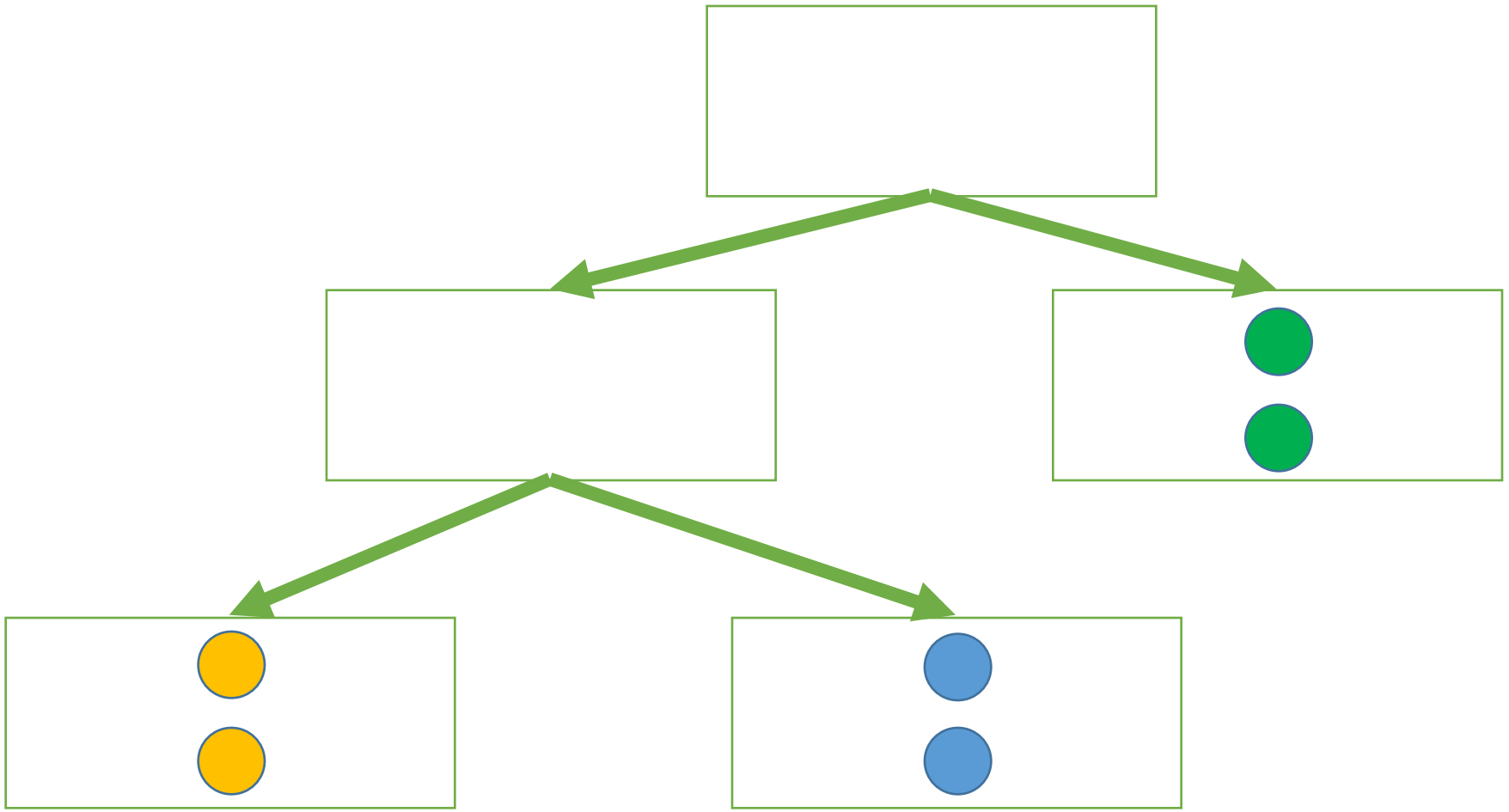
Жадное построение



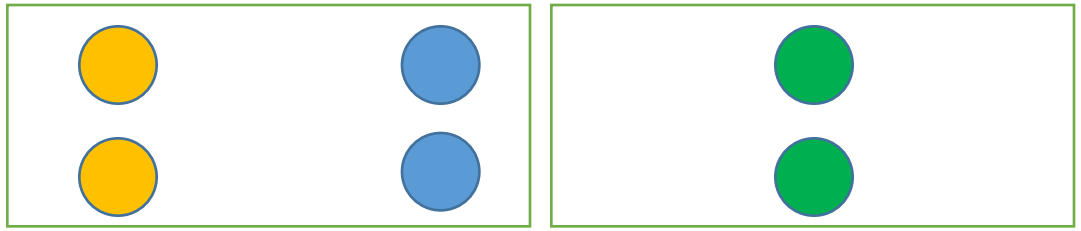
Жадное построение



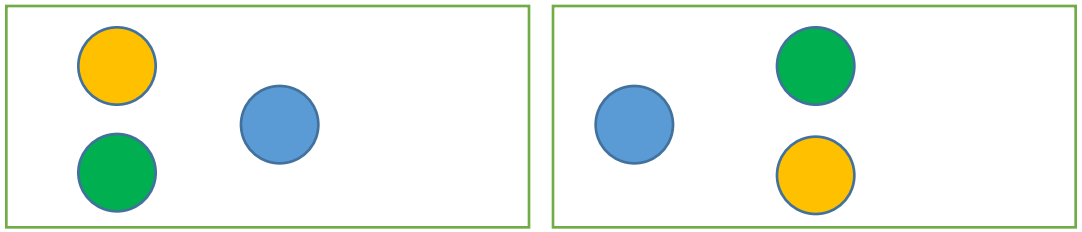
Жадное построение



Как сравнить разбиения?

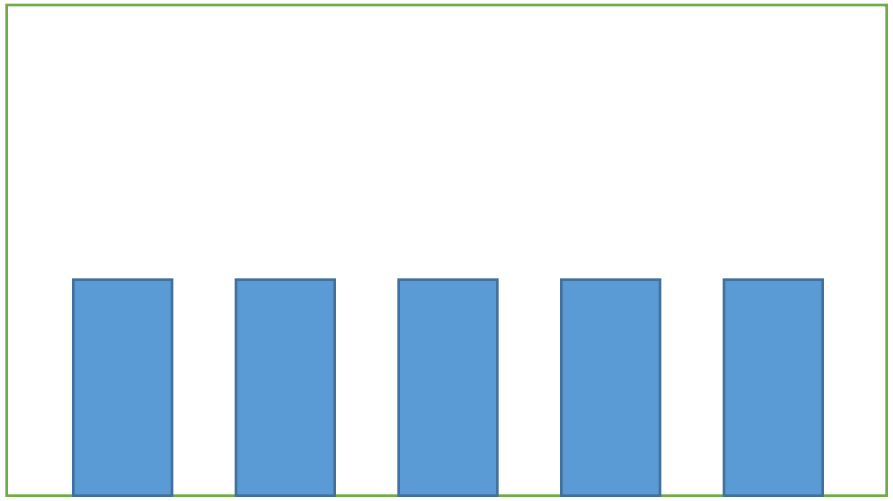


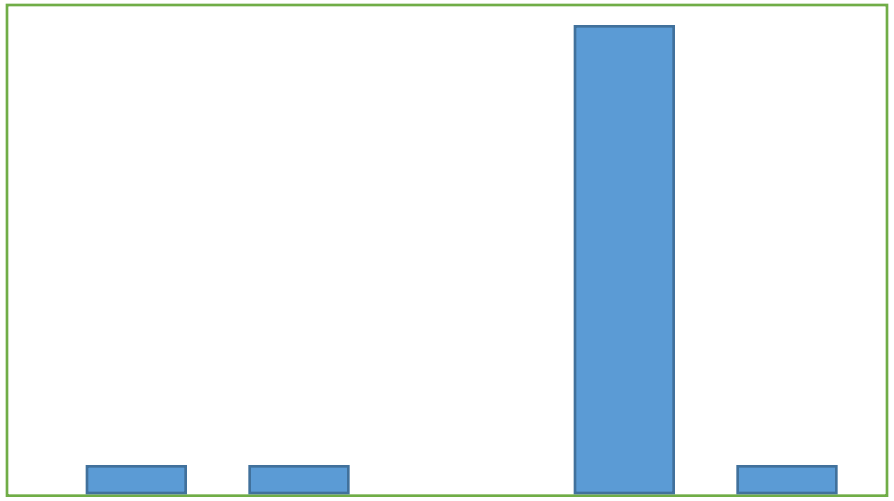
или



Энтропия

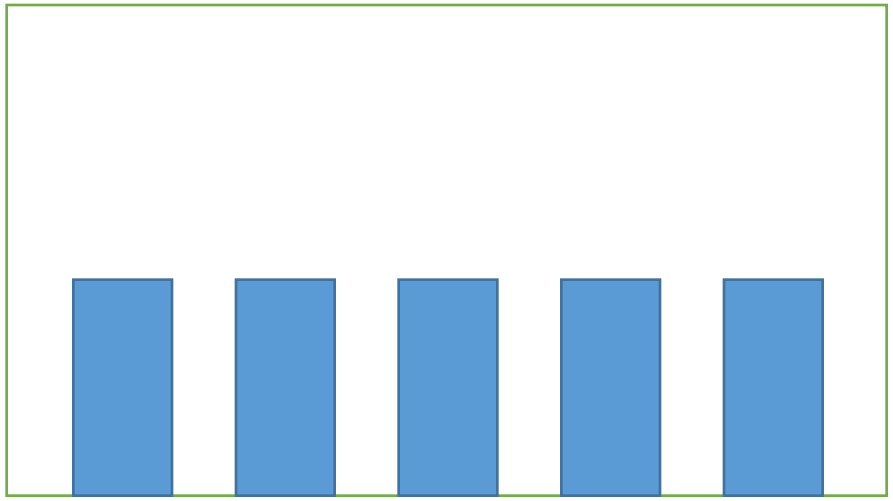
- Мера неопределённости распределения



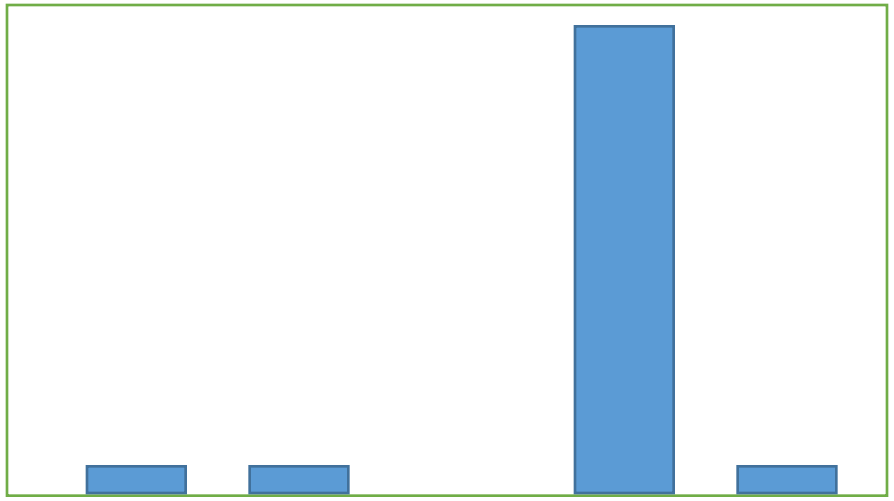


Энтропия

- Мера неопределённости распределения



Высокая энтропия



Низкая энтропия

Энтропия

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

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

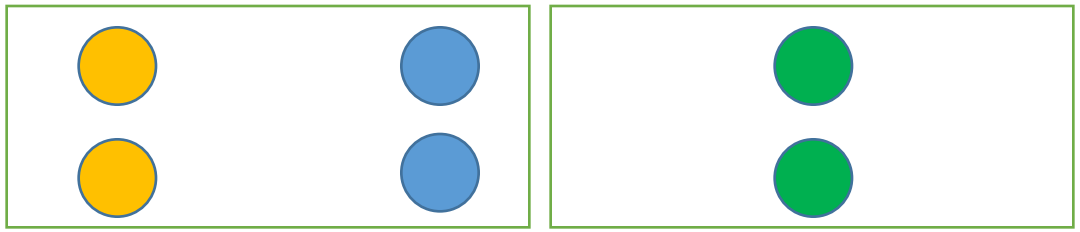
Энтропия

- $(0.2, 0.2, 0.2, 0.2, 0.2)$
- $H = 1.60944 \ldots$

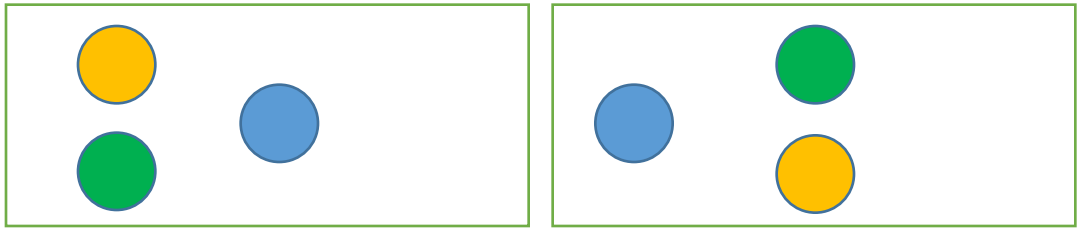
- $(0.9, 0.05, 0.05, 0, 0)$
- $H = 0.394398 \ldots$

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

Как сравнить разбиения?

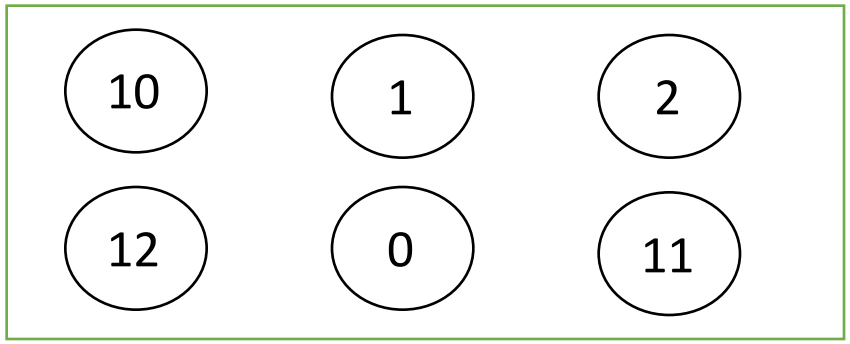


- $(0.5, 0.5, 0)$ и $(0, 0, 1)$
- $H = 0.693 + 0 = 0.693$

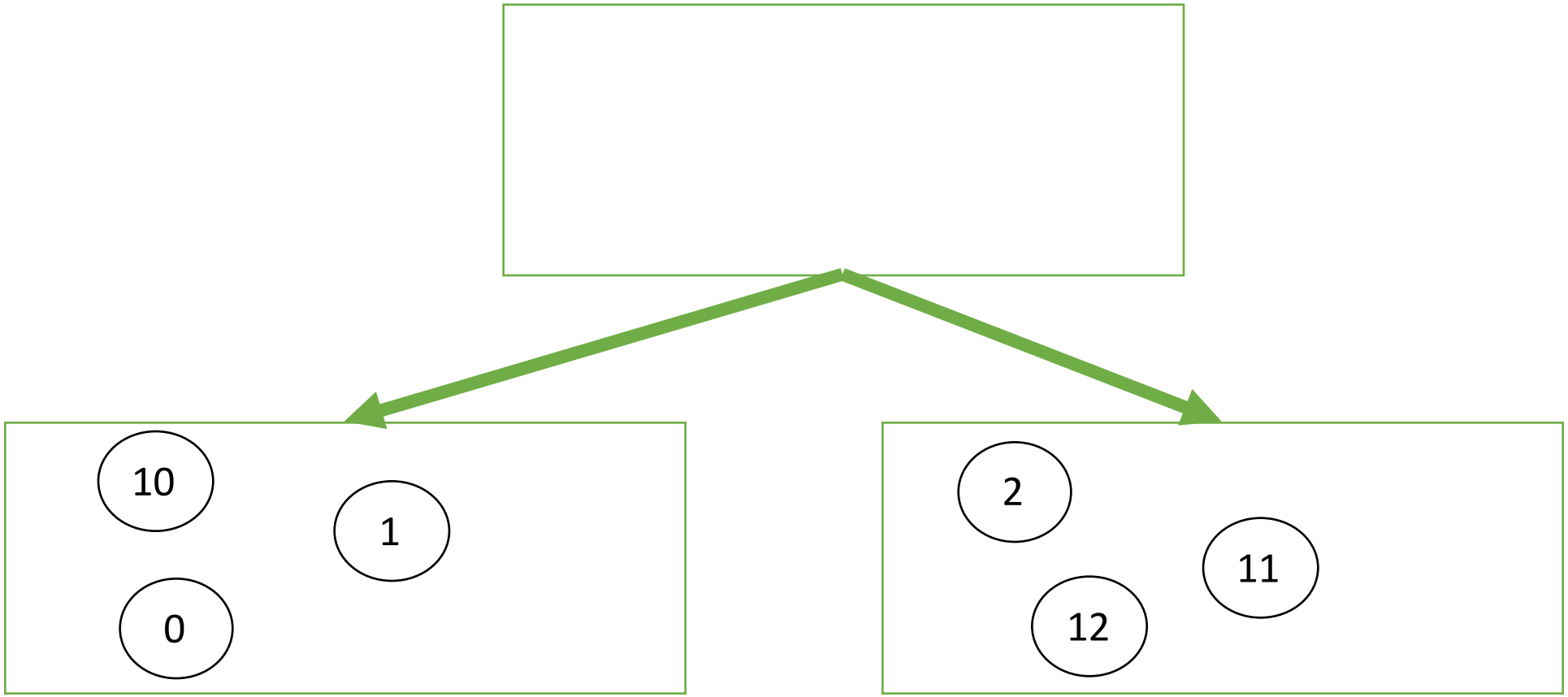


- $(0.33, 0.33, 0.33)$ и $(0.33, 0.33, 0.33)$
- $H = 1.09 + 1.09 = 2.18$

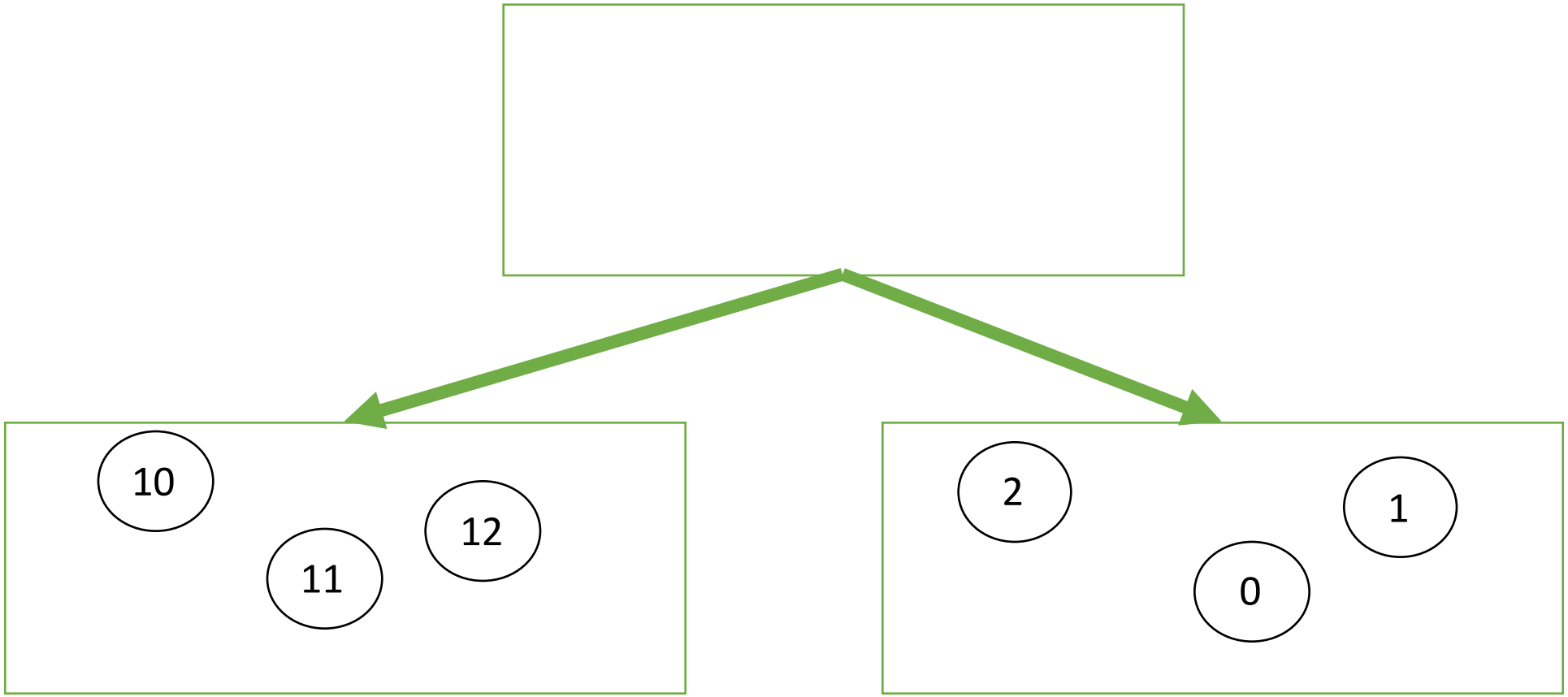
А для регрессии?



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:

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

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 перекращать 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:

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

Ответ в листе

- Допустим, решили сделать вершину m листом
- Какой прогноз выбрать?
- Вероятности классов:

$$a_{mk} = \frac{1}{|X_m|} \sum_{i \in X_m} [y_i = k]$$

Вопросы

- Критерий ошибки разбиения?
- Критерий останова?

Резюме

- Иногда модель нужно интерпретировать
- Решающие деревья легко объяснить
- Решающие деревья легко переобучаются
- Построение деревьев — жадный алгоритм

В следующий раз

- Подробнее про обучение деревьев
- Борьба с переобучением деревьев