

Основы программирования в Python

Лекция 3

Классификация

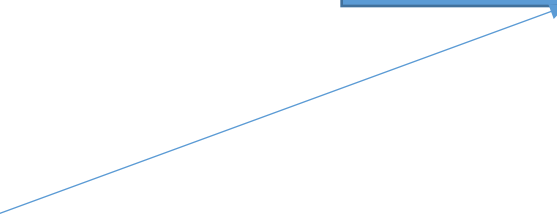
Классификация

- $\mathbb{Y} = \{-1, +1\}$
- -1 — отрицательный класс
- $+1$ — положительный класс
- $a(x)$ должен возвращать одно из двух чисел

Линейная регрессия

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

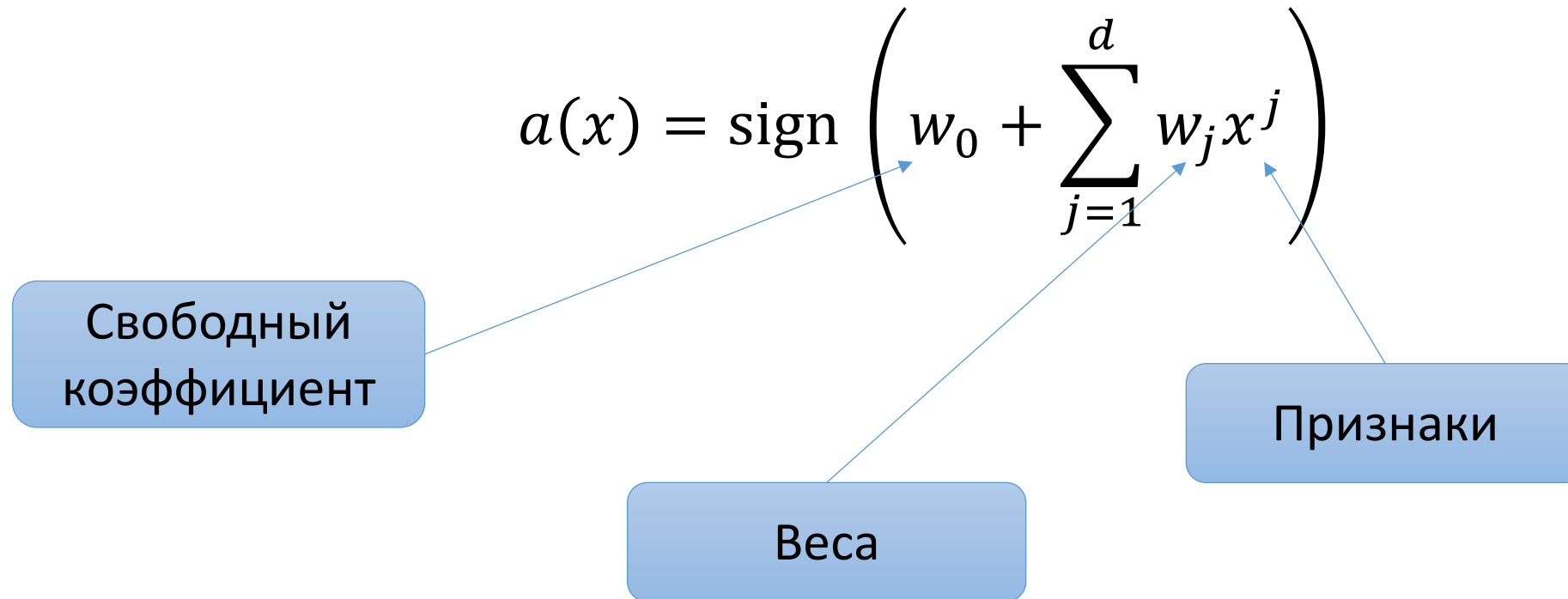
Вещественное
число!



Линейный классификатор

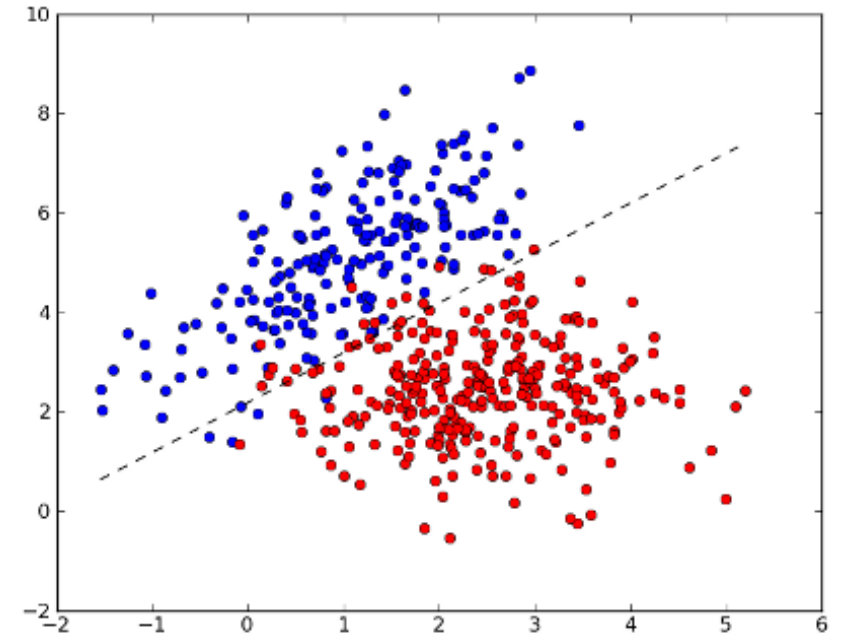
$$a(x) = \text{sign} \left(w_0 + \sum_{j=1}^d w_j x^j \right)$$

Линейный классификатор

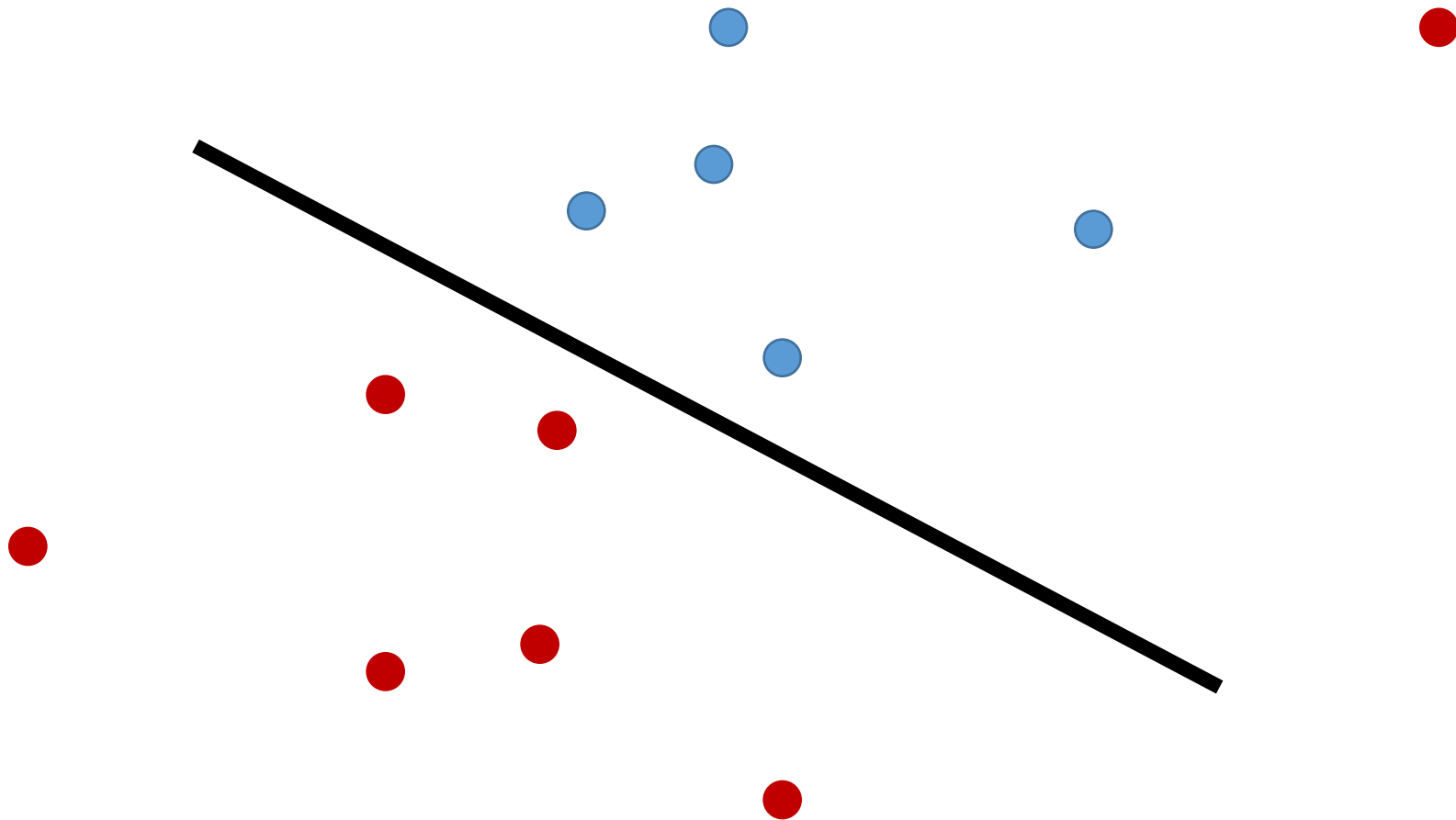


Геометрия линейного классификатора

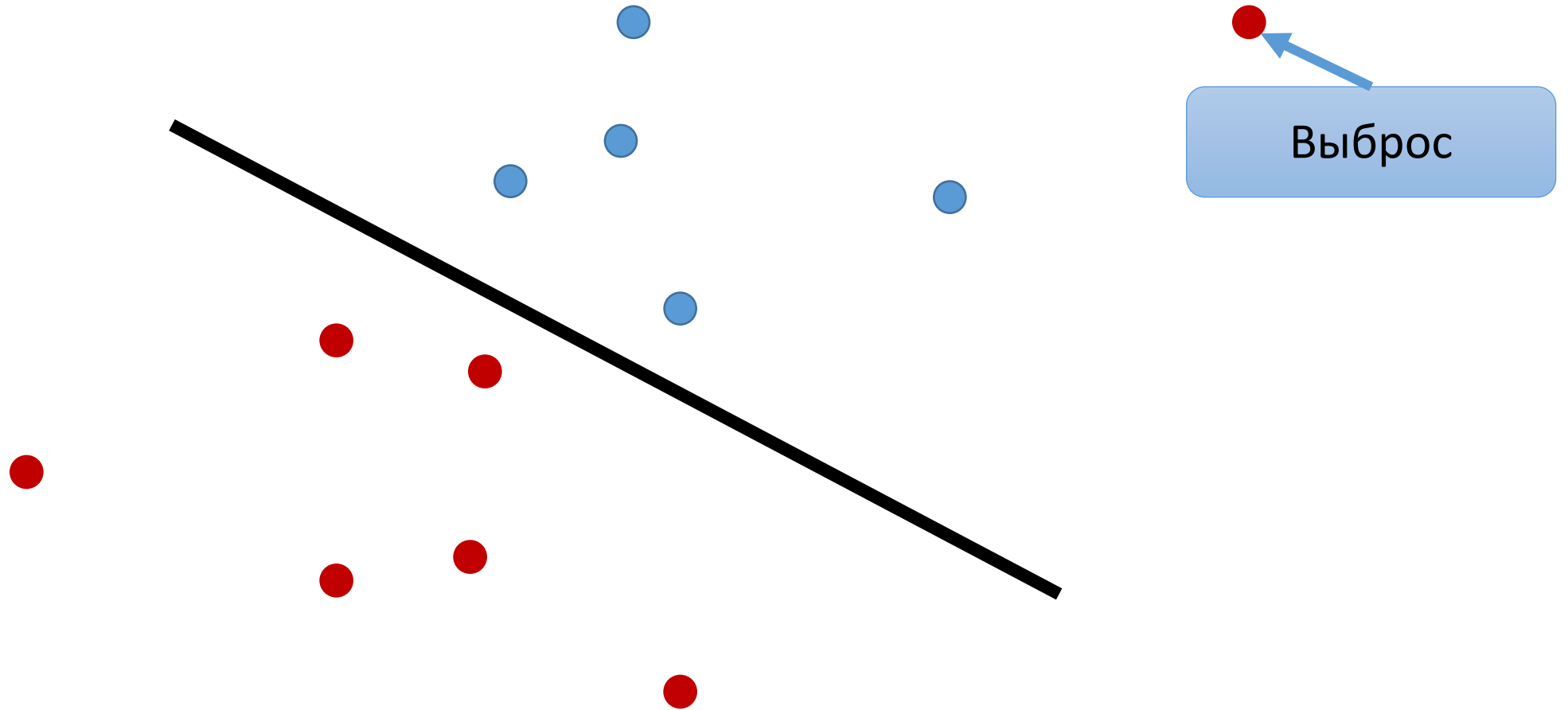
- Линейный классификатор проводит гиперплоскость
- $\langle w, x \rangle < 0$ — объект «слева» от неё
- $\langle w, x \rangle > 0$ — объект «справа» от неё



Геометрия линейного классификатора

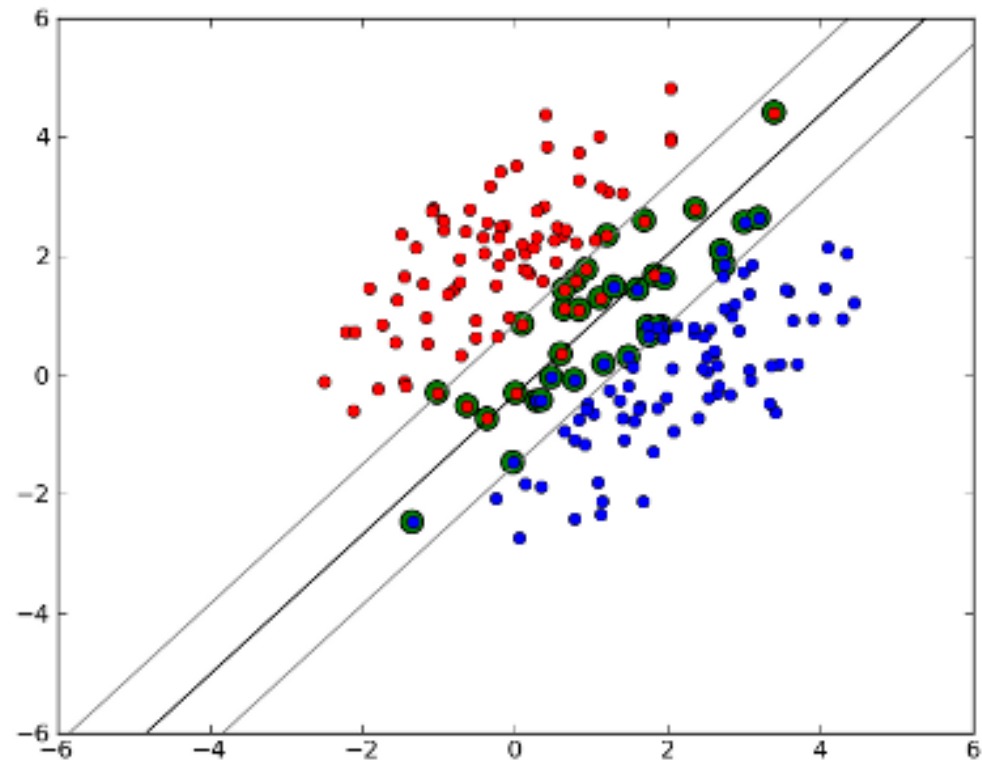


Геометрия линейного классификатора



Отступ

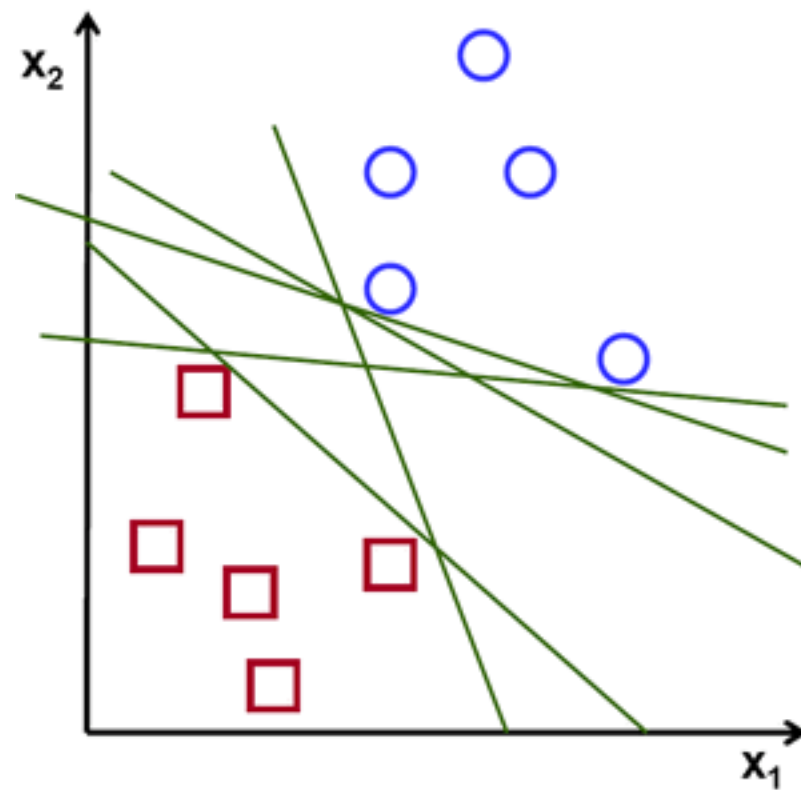
- $M_i = y_i \langle w, x_i \rangle$
- $M_i > 0$ — классификатор дает верный ответ
- $M_i < 0$ — классификатор ошибается
- Чем дальше отступ от нуля, тем больше уверенности



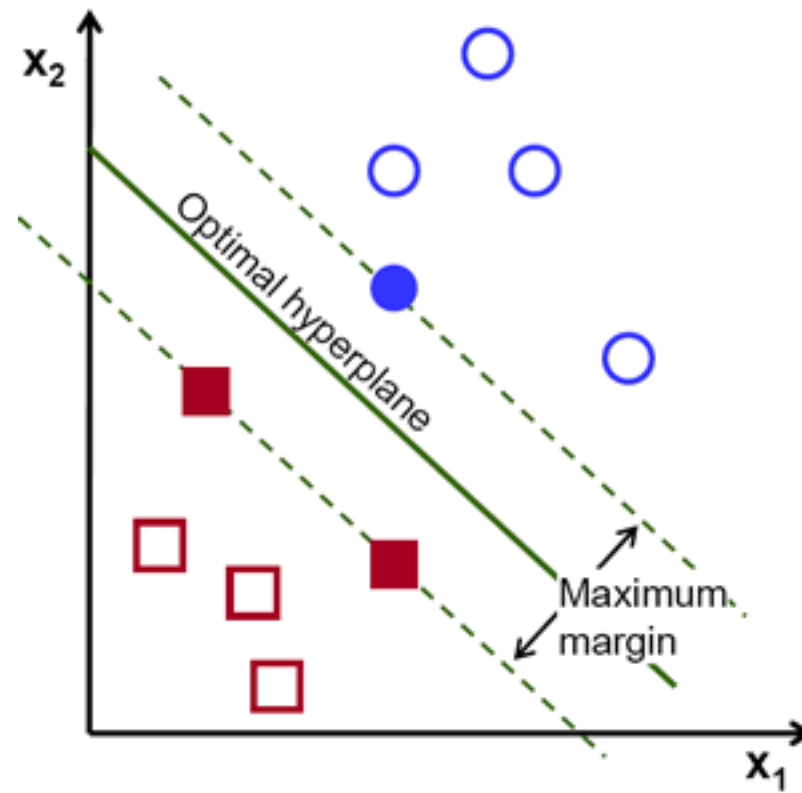
Линейный классификатор

- Линейный классификатор разделяет два класса гиперплоскостью
- Чем больше отступ по модулю, тем дальше объект от гиперплоскости
- Знак отступа говорит о корректности предсказания

SVM



SVM



Качество классификации

| $a(x)$ | y |
|-----------|-----------|
| -1 | -1 |
| +1 | +1 |
| -1 | -1 |
| +1 | -1 |
| +1 | +1 |

- Доля неправильных ответов:

$$\frac{1}{5} = 0.2$$

Качество классификации

- Доля **правильных** ответов:

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

- На английском: **accuracy**

Качество классификации

- Доля **правильных** ответов:

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

- На английском: **accuracy**
- ВАЖНО: не переводите это как «точность»!

Несбалансированные выборки

- Пример:
 - Класс -1: 950 объектов
 - Класс +1: 50 объектов
- $a(x) = -1$
- Доля правильных ответов: 0.95

Несбалансированные выборки

- q_0 — доля объектов самого крупного класса
- Для разумных алгоритмов:

$$\text{accuracy} \in [q_0, 1]$$

- Если получили большой accuracy — посмотрите на баланс классов

Цены ошибок

- Пример: кредитный скоринг
- Модель 1:
 - 80 кредитов вернули
 - 20 кредитов не вернули
- Модель 2:
 - 48 кредитов вернули
 - 2 кредита не вернули
- Кто лучше?

Цены ошибок

- Что хуже?
 - Выдать кредит «плохому» клиенту
 - Не выдать кредит «хорошему» клиенту
- Доля верных ответов не учитывает цены ошибок

Матрица ошибок

| | $y = 1$ | $y = -1$ |
|-------------|---------------------|---------------------|
| $a(x) = 1$ | True Positive (TP) | False Positive (FP) |
| $a(x) = -1$ | False Negative (FN) | True Negative (TN) |

Матрица ошибок

- Модель $a_1(x)$:

| | $y = 1$ | $y = -1$ |
|-------------|---------|----------|
| $a(x) = 1$ | 80 | 20 |
| $a(x) = -1$ | 20 | 80 |

- Модель $a_2(x)$:

| | $y = 1$ | $y = -1$ |
|-------------|---------|----------|
| $a(x) = 1$ | 48 | 2 |
| $a(x) = -1$ | 52 | 98 |

Точность (precision)

- Можно ли доверять классификатору при $a(x) = 1$?

$$\text{precision}(a, X) = \frac{TP}{TP + FP}$$

Точность (precision)

- Модель $a_1(x)$:

| | $y = 1$ | $y = -1$ |
|-------------|---------|----------|
| $a(x) = 1$ | 80 | 20 |
| $a(x) = -1$ | 20 | 80 |

- $\text{precision}(a_1, X) = 0.8$

- Модель $a_2(x)$:

| | $y = 1$ | $y = -1$ |
|-------------|---------|----------|
| $a(x) = 1$ | 48 | 2 |
| $a(x) = -1$ | 52 | 98 |

- $\text{precision}(a_2, X) = 0.96$

Полнота (recall)

- Как много положительных объектов находит классификатор?

$$\text{recall}(a, X) = \frac{TP}{TP + FN}$$

Полнота (recall)

- Модель $a_1(x)$:

| | $y = 1$ | $y = -1$ |
|-------------|---------|----------|
| $a(x) = 1$ | 80 | 20 |
| $a(x) = -1$ | 20 | 80 |

- $\text{recall}(a_1, X) = 0.8$

- Модель $a_2(x)$:

| | $y = 1$ | $y = -1$ |
|-------------|---------|----------|
| $a(x) = 1$ | 48 | 2 |
| $a(x) = -1$ | 52 | 98 |

- $\text{recall}(a_2, X) = 0.48$

Антифрод

- Классификация транзакций на нормальные и мошеннические
- Высокая точность, низкая полнота:
 - Редко блокируем нормальные транзакции
 - Пропускаем много мошеннических
- Низкая точность, высокая полнота:
 - Часто блокируем нормальные транзакции
 - Редко пропускаем мошеннические

Кредитный скоринг

- Неудачных кредитов должно быть не больше 5%
- Ограничение: $\text{precision}(a, X) \geq 0.95$
- Максимизируем полноту

Медицинская диагностика

- Надо найти не менее 80% больных
- Ограничение: $\text{recall}(a, X) \geq 0.8$
- Максимизируем точность

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

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

Пример

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

Логические правила

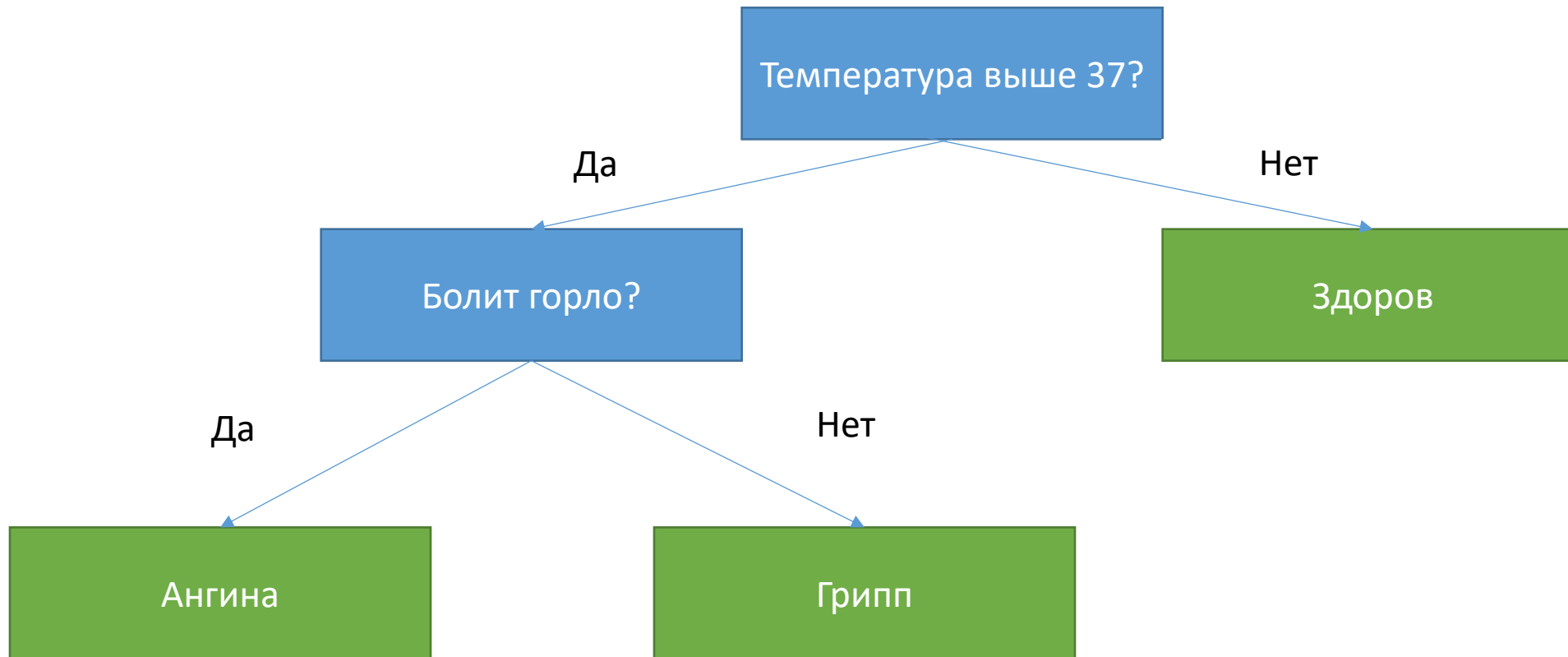
$[\text{этаж} > 3][\text{площадь} < 40][\text{число комнат} < 3]$

- 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
- Решающие деревья

Медицинская диагностика



Принятие решений

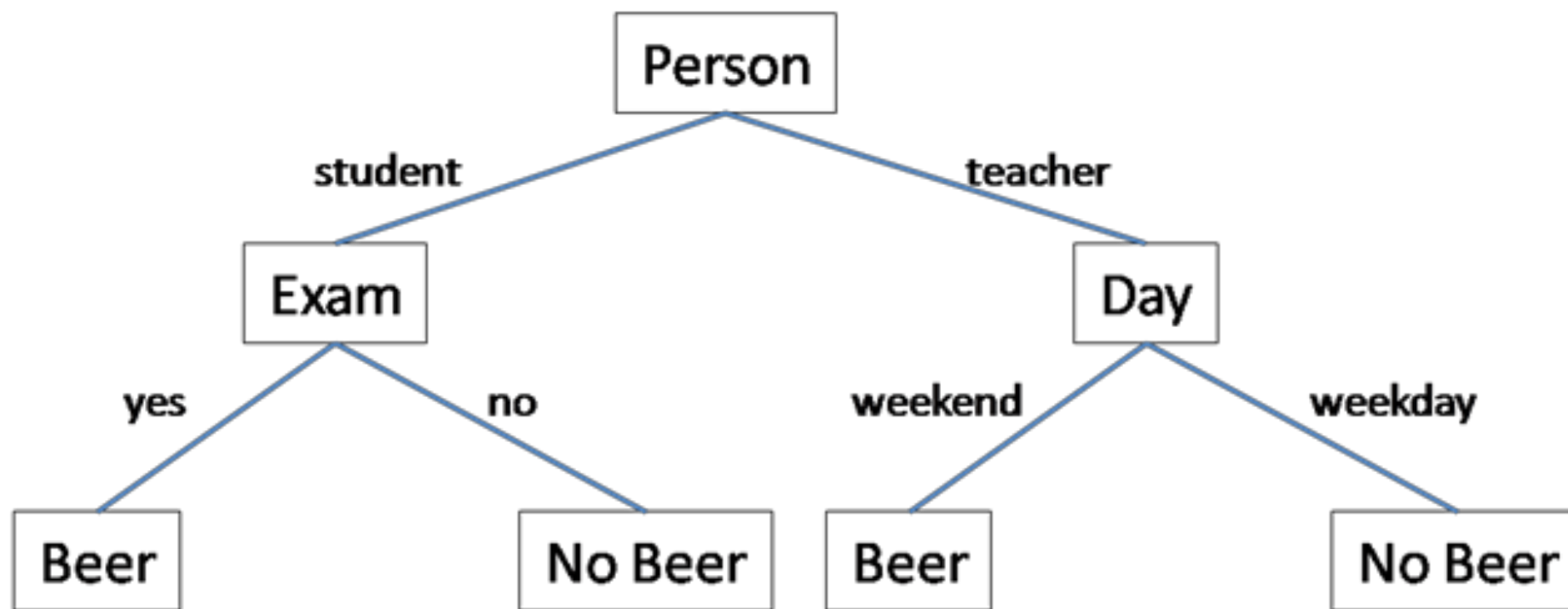
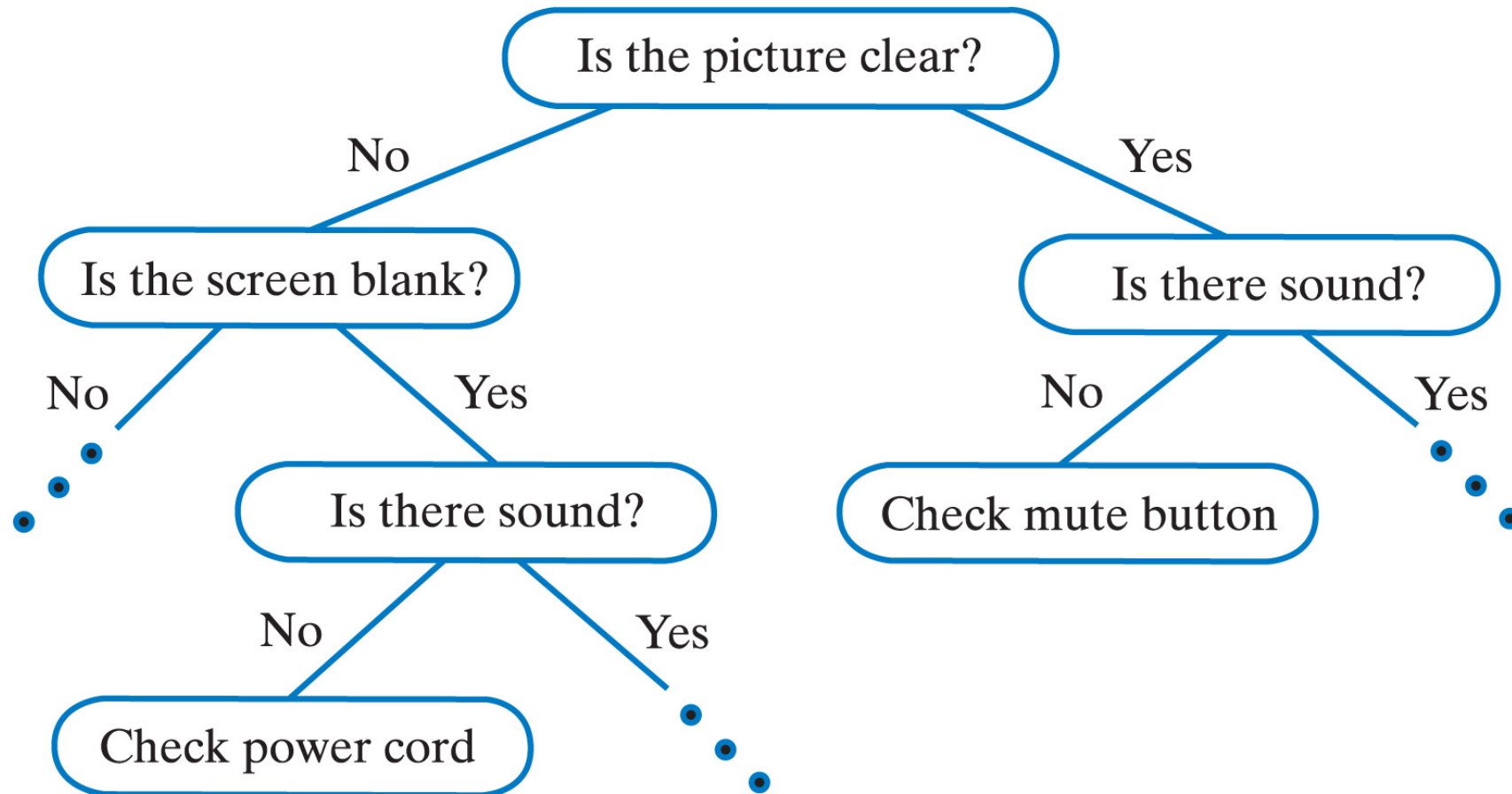
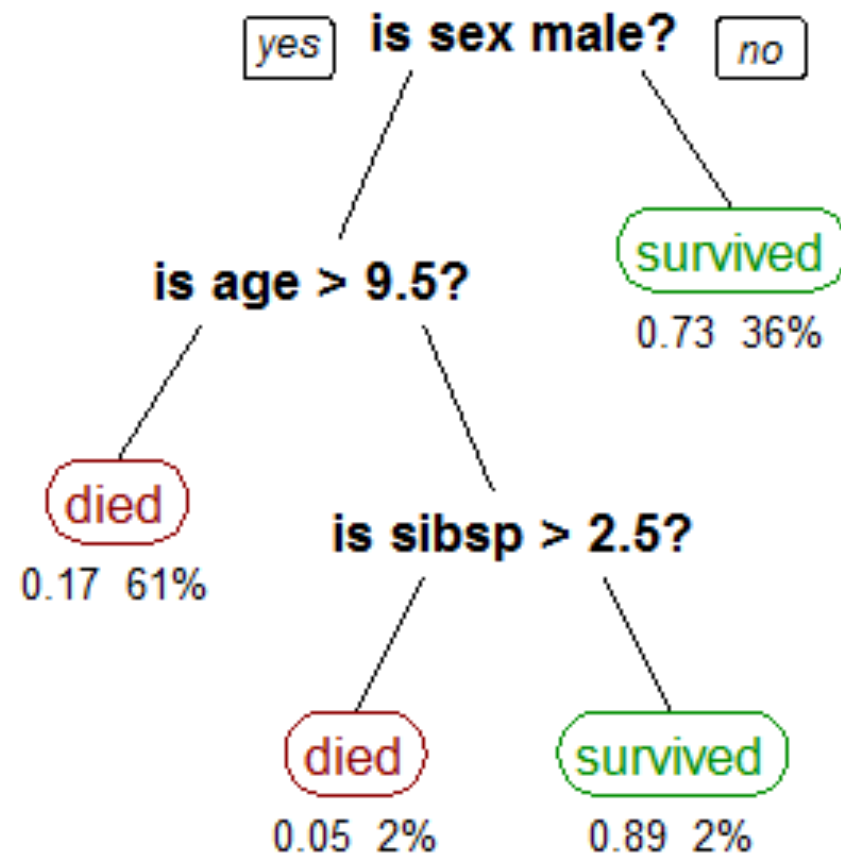


Схема диалога с клиентом

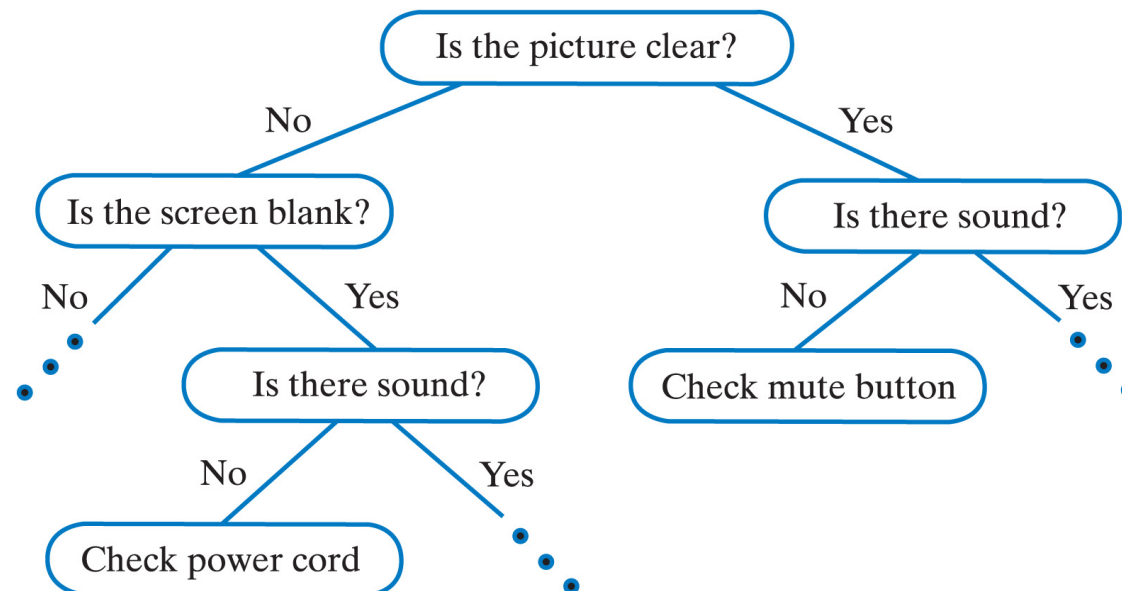


Пассажиры Титаника



Решающее дерево

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



УСЛОВИЯ

- Самые популярные варианты:

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

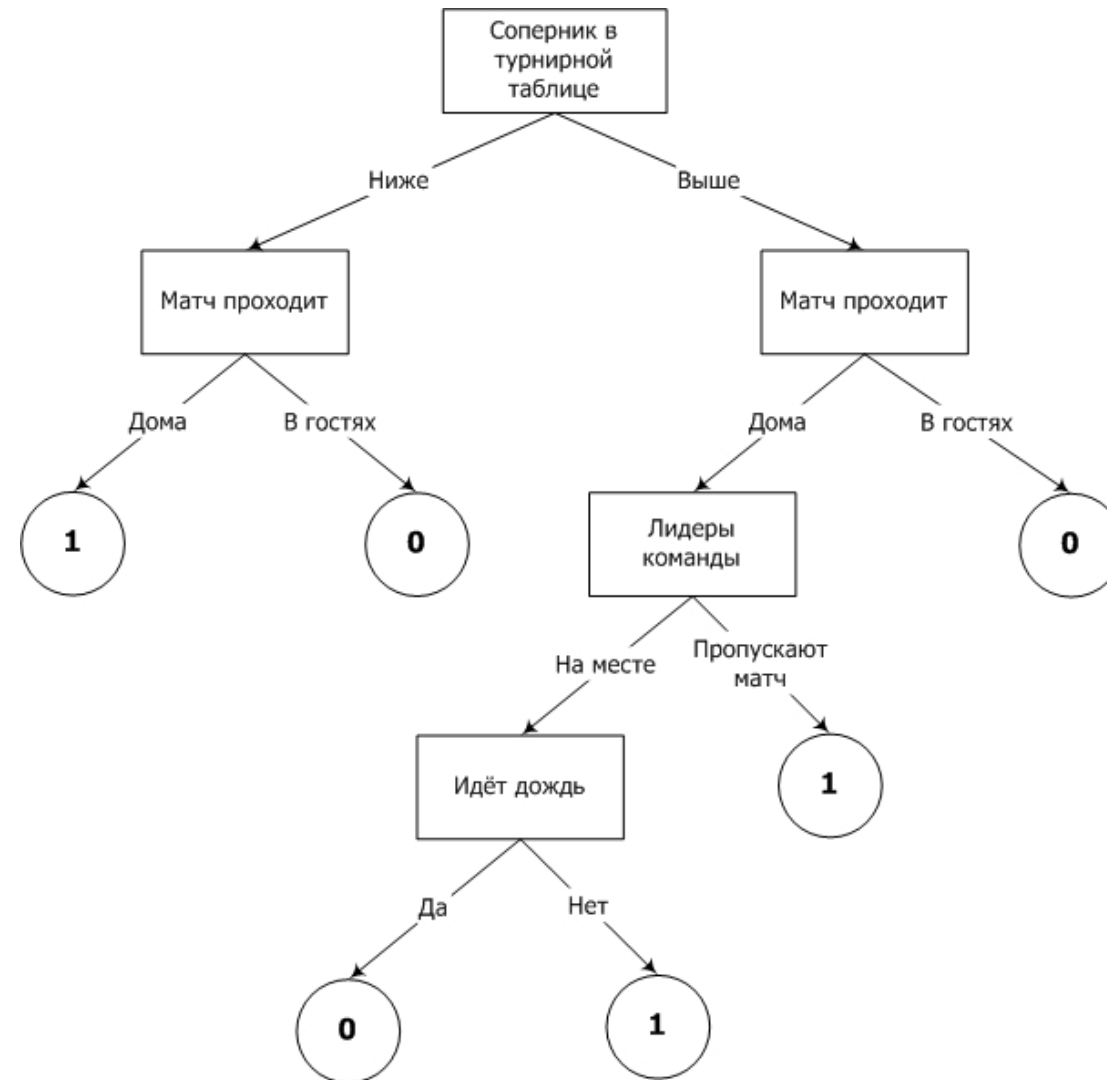
Примеры:

- [этаж = 5]
- [площадь \leq 30]

Прогноз в листе

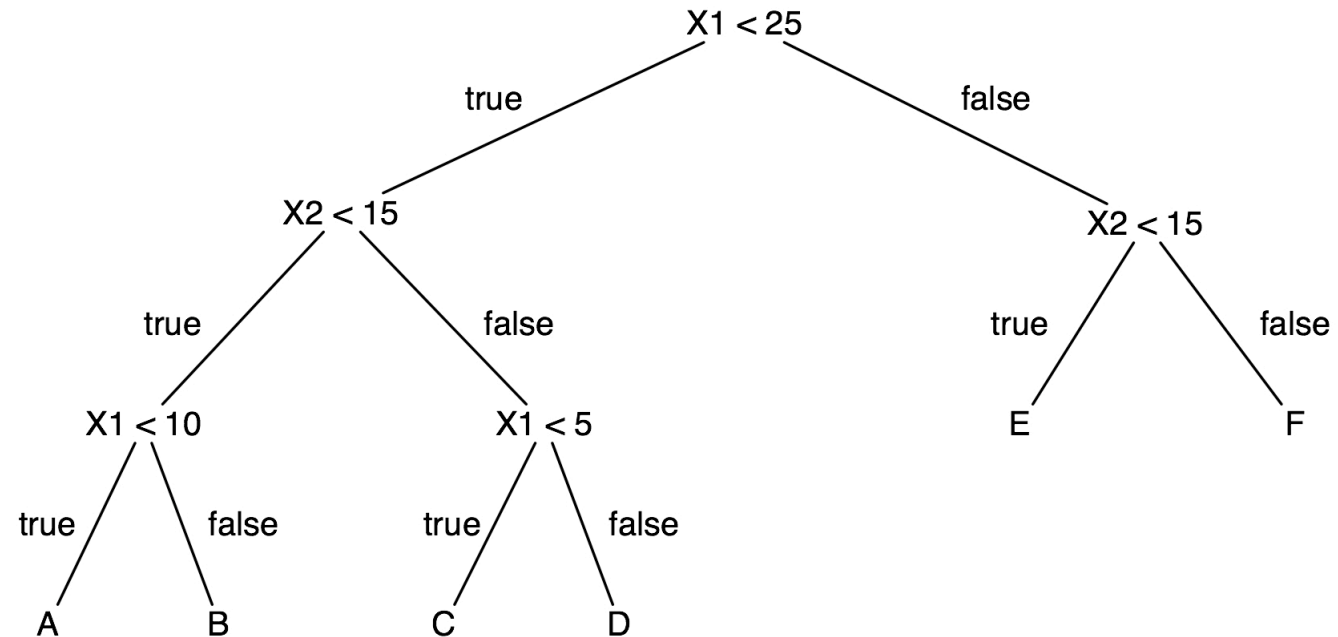
- Регрессия:
 - Вещественное число
- Классификация:
 - Класс
 - Вероятности классов

Исход футбольного матча

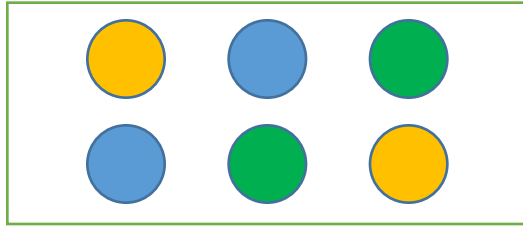


Жадное построение

- Растим дерево от корня к листьям

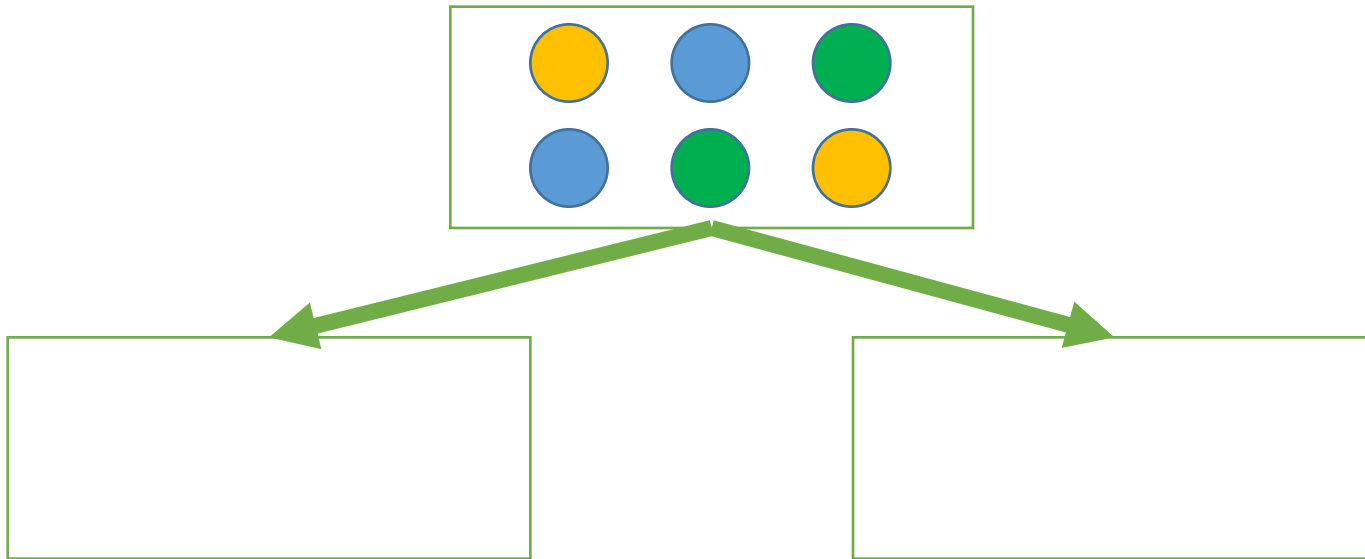


Жадное построение

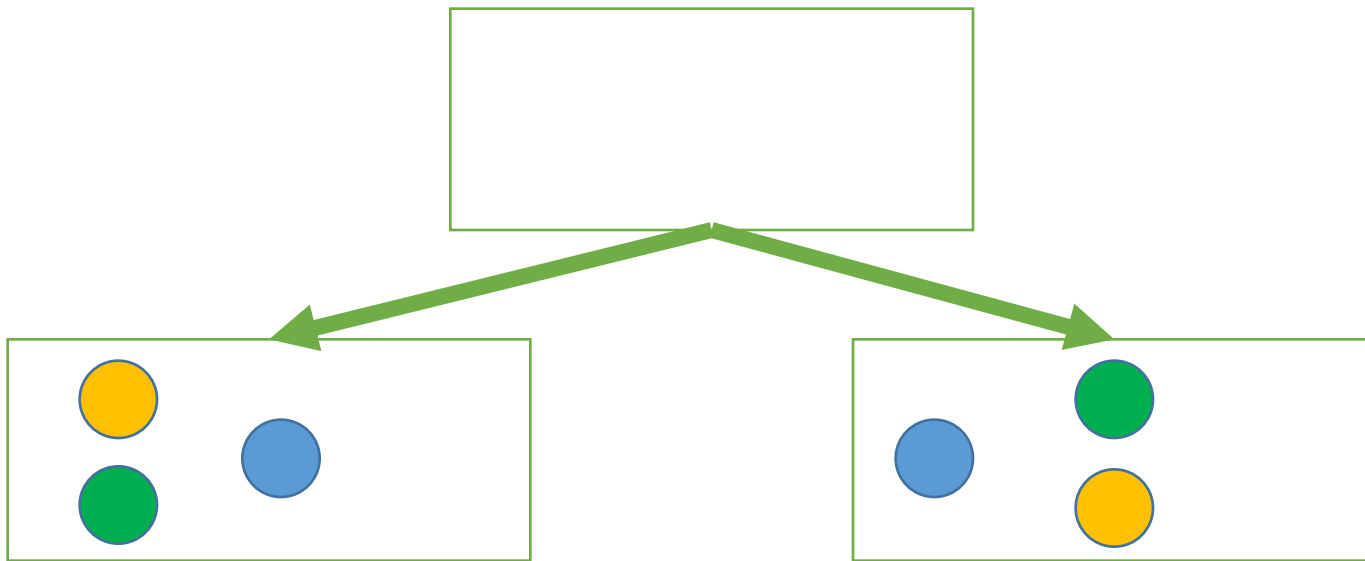


- Как разбить вершину?

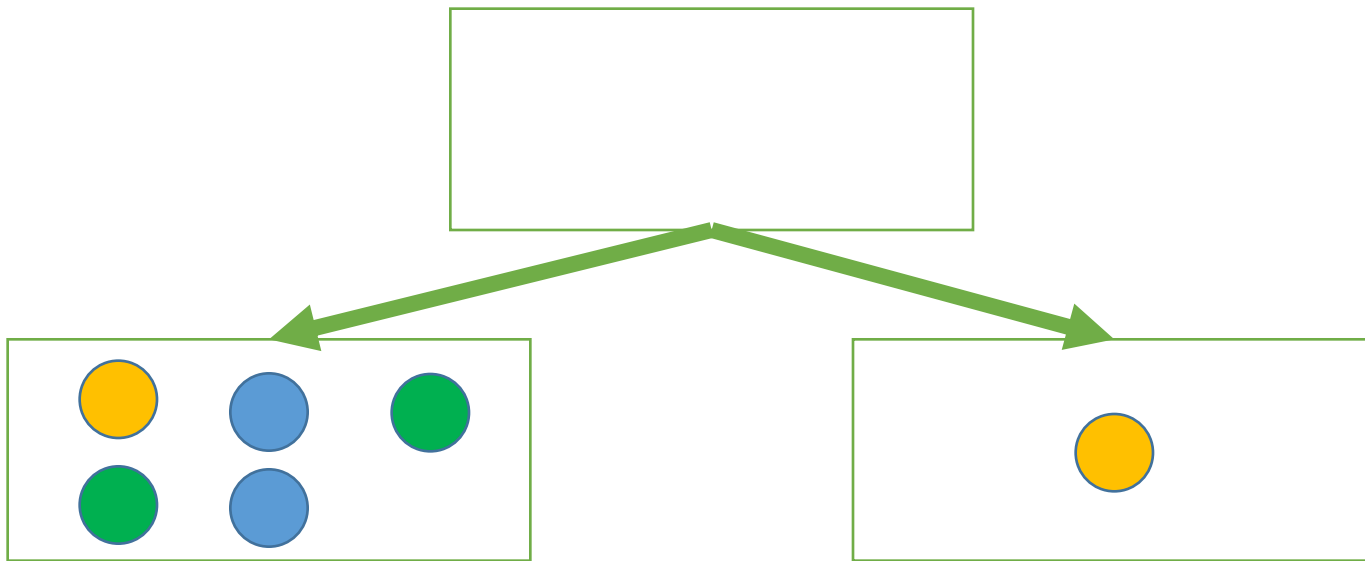
Жадное построение



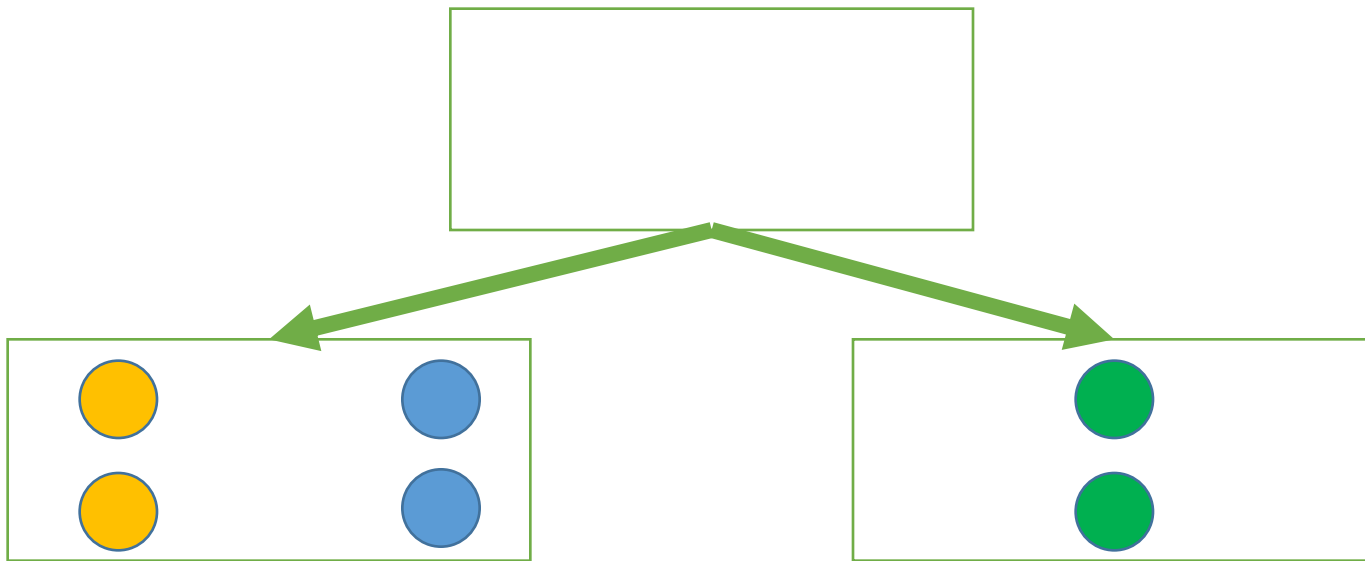
Жадное построение



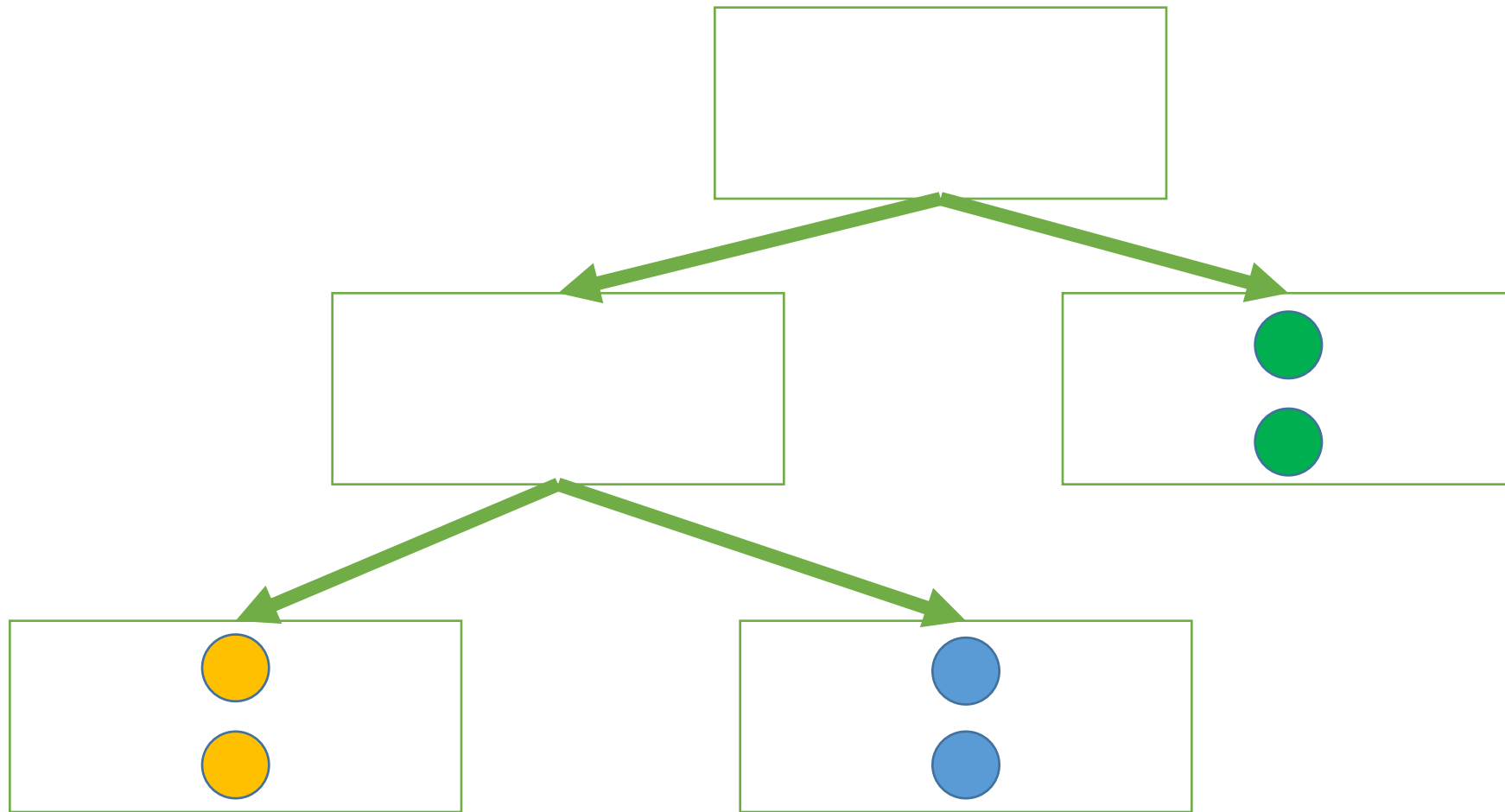
Жадное построение



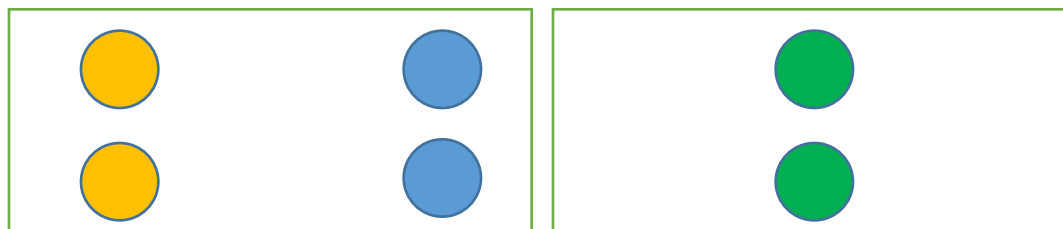
Жадное построение



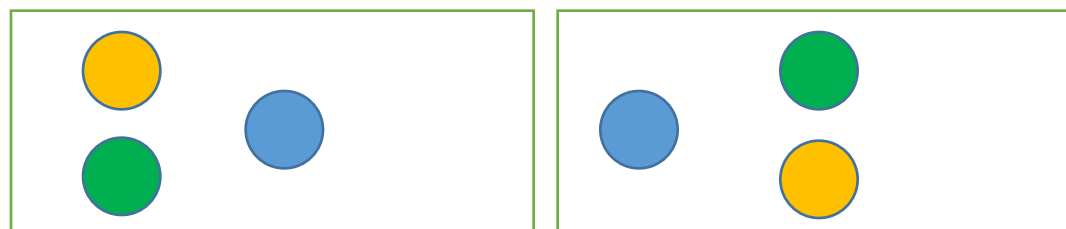
Жадное построение



Как сравнить разбиения?

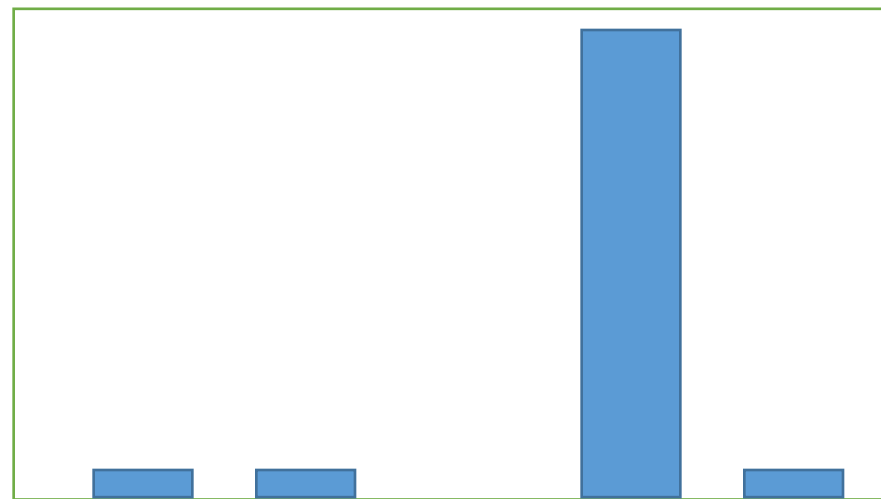
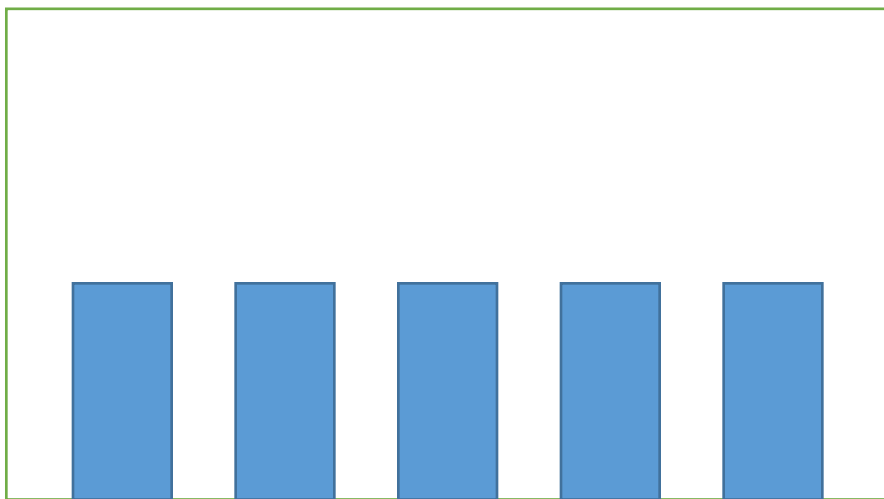


или



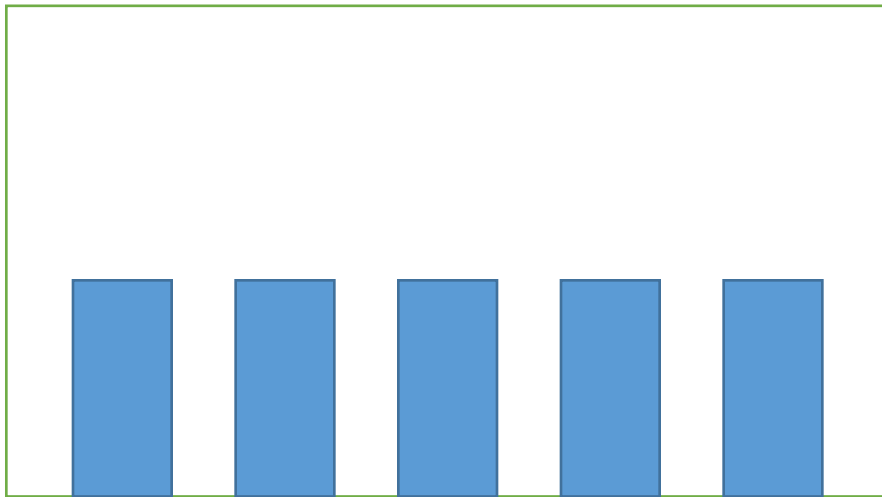
Энтропия

- Мера неопределённости распределения

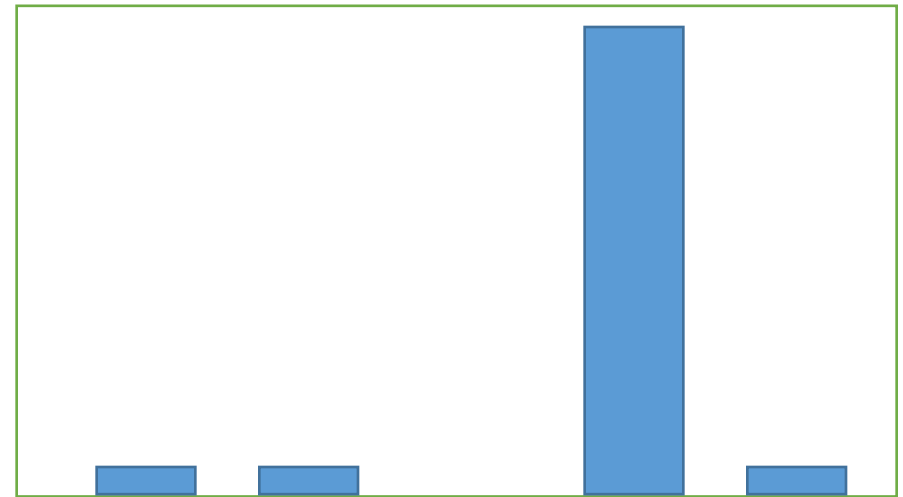


Энтропия

- Мера неопределённости распределения



Высокая энтропия



Низкая энтропия

Энтропия

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

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

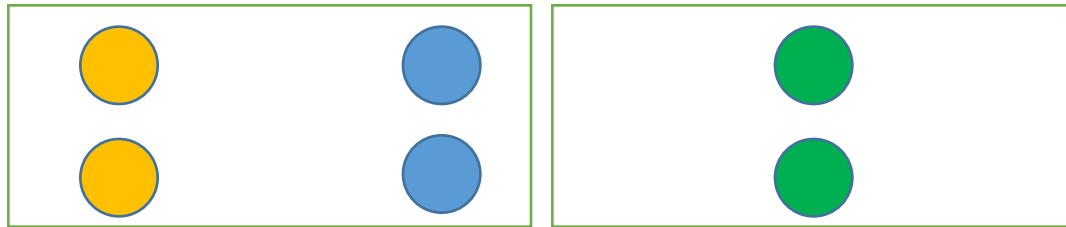
Энтропия

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

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

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

Как сравнить разбиения?



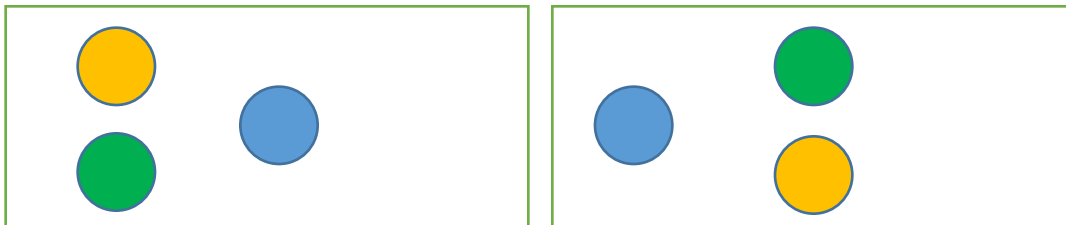
0.693

0

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

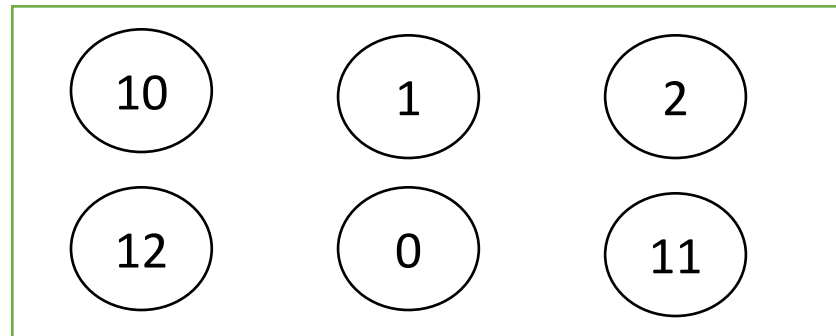
1.09

1.09

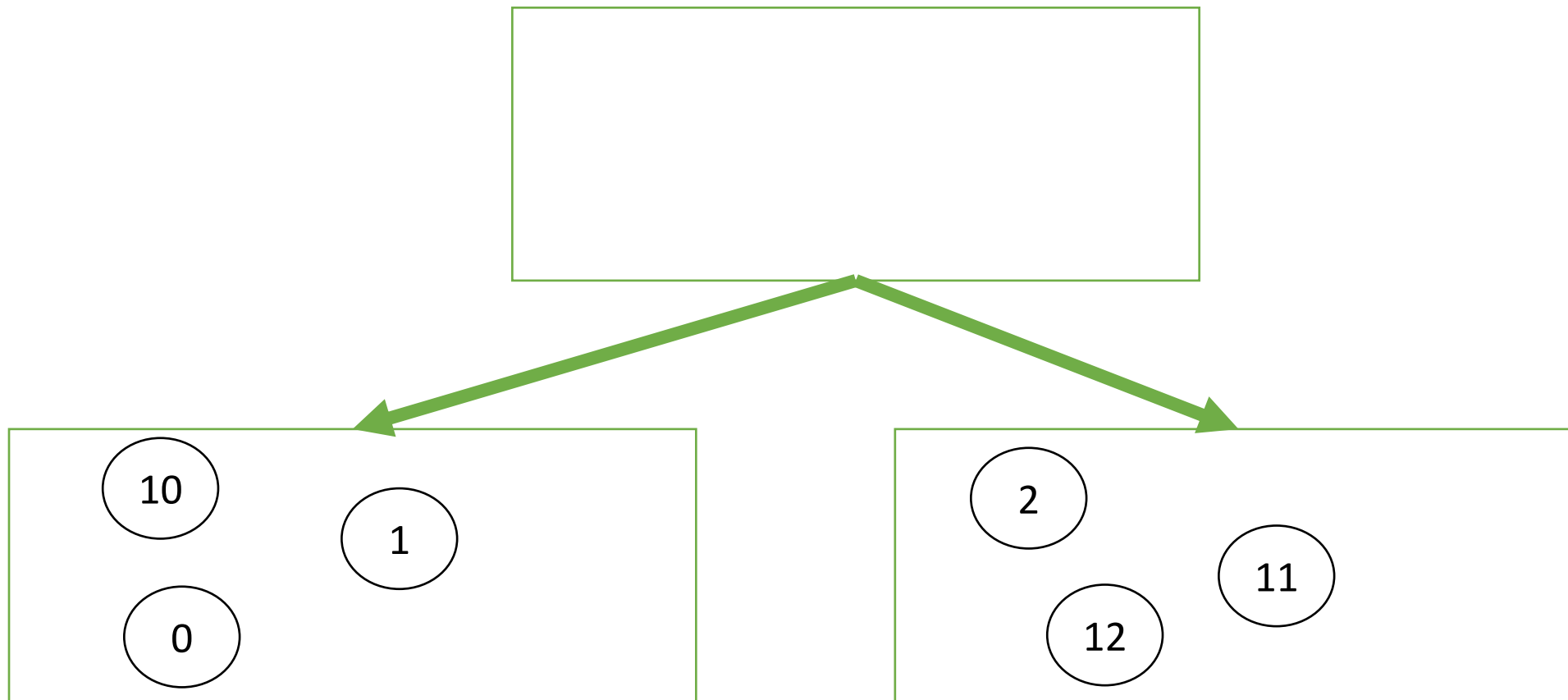


- $(0.33, 0.33, 0.33)$ и $(0.33, 0.33, 0.33)$
- $H = 1.09 + 1.09 = 2.18$

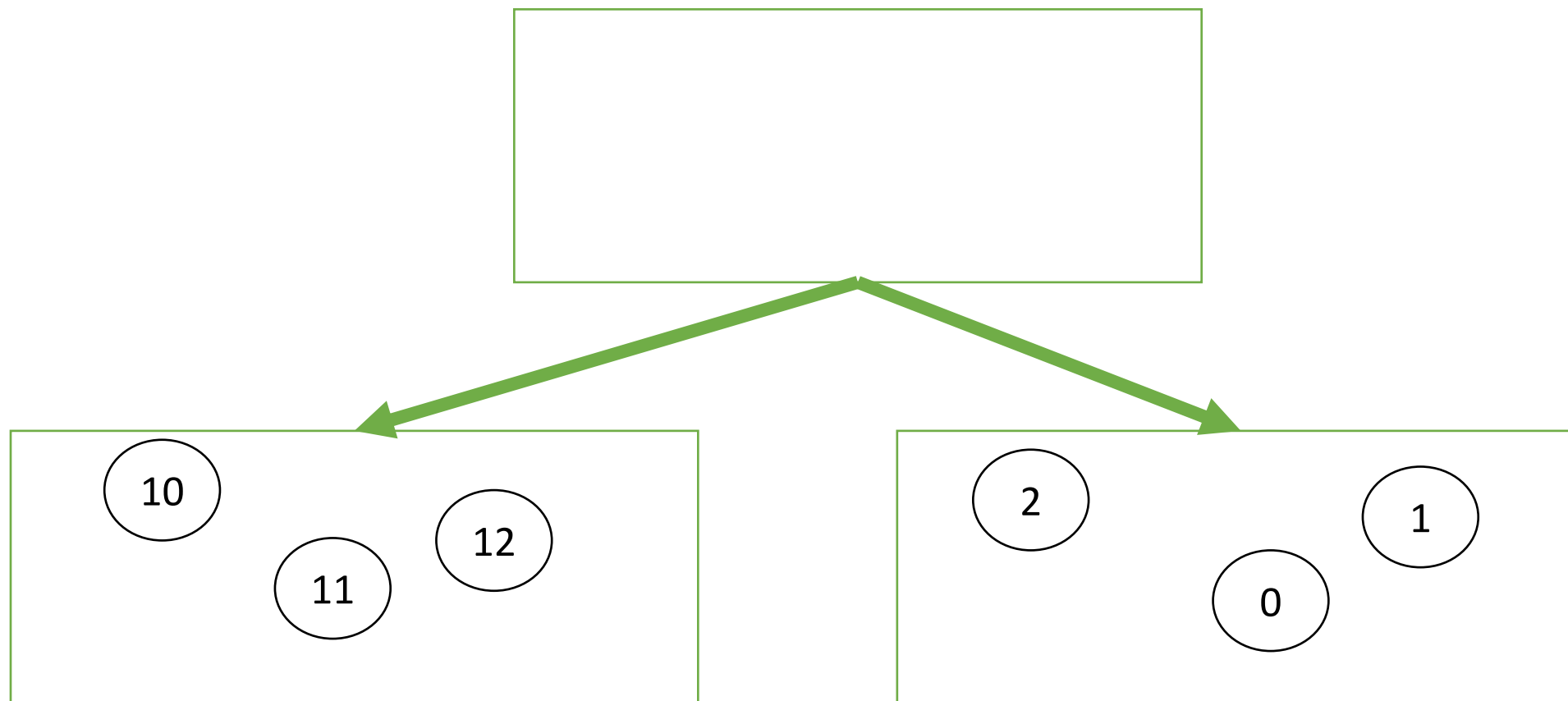
А для регрессии?



А для регрессии?



А для регрессии?



А для регрессии?

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

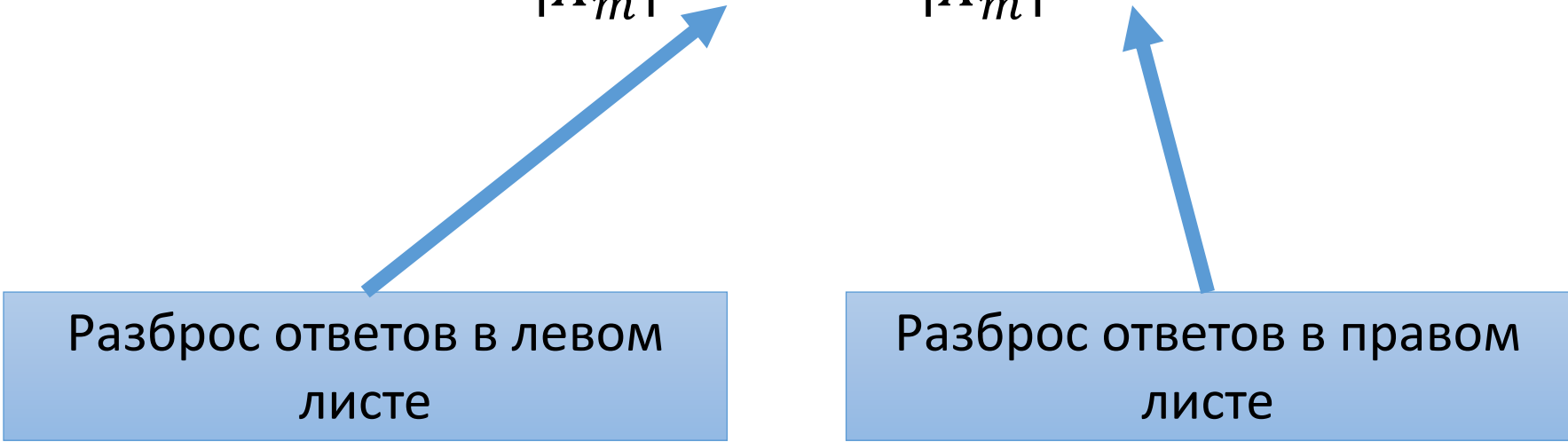
Поиск разбиения

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

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 :

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

Критерий останова

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

Ответ в листе

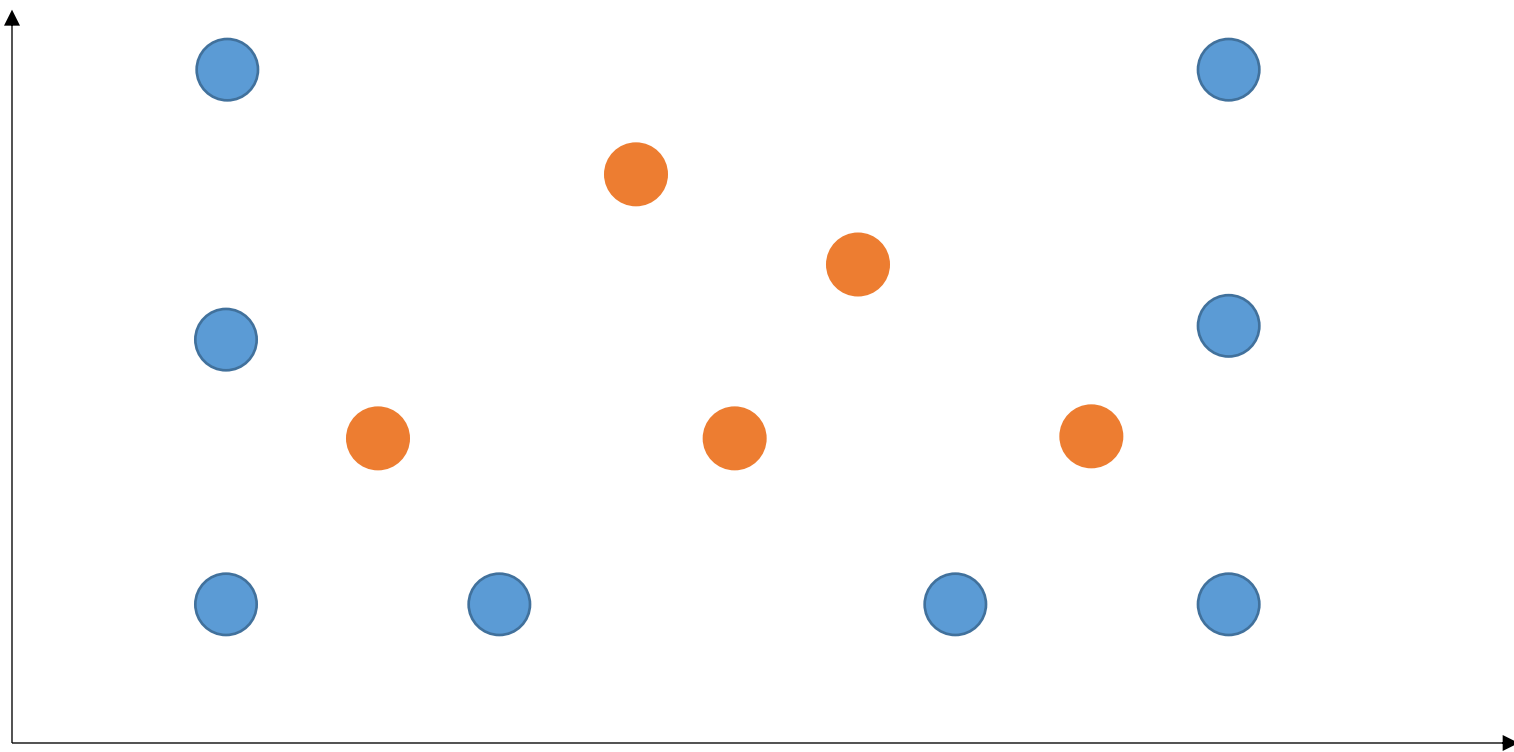
- Допустим, решили сделать вершину t листом
- Какой прогноз выбрать?
- Регрессия:

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

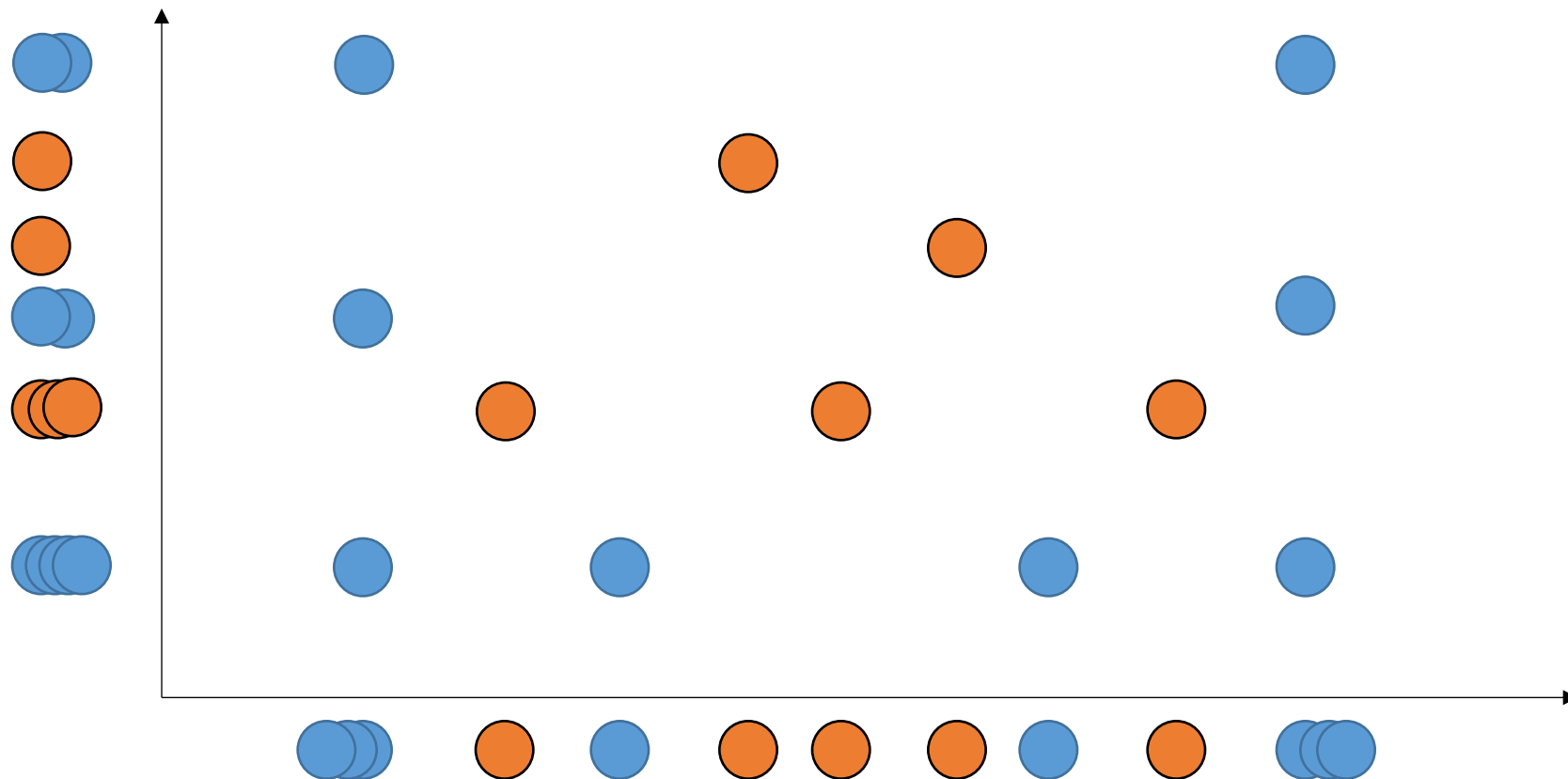
- Классификация:

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

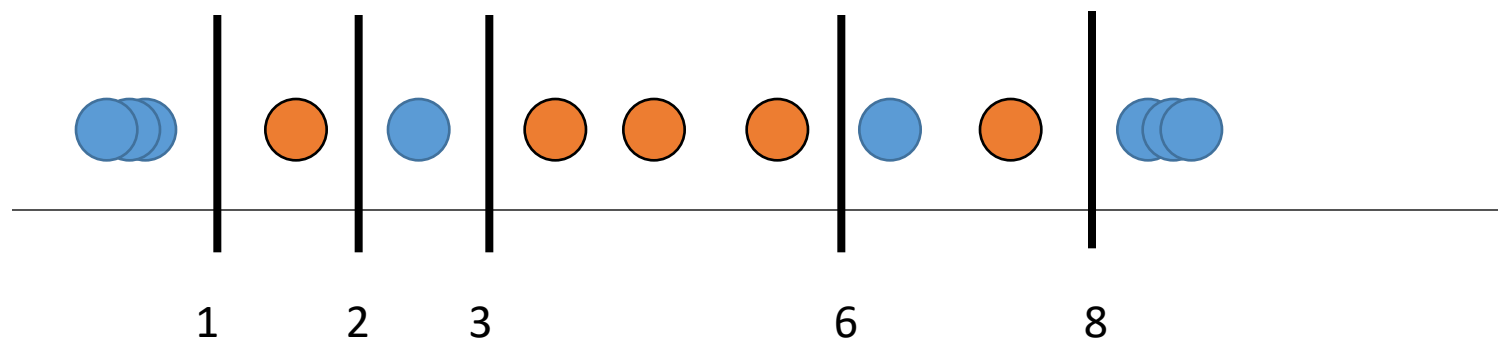
Обучение деревьев



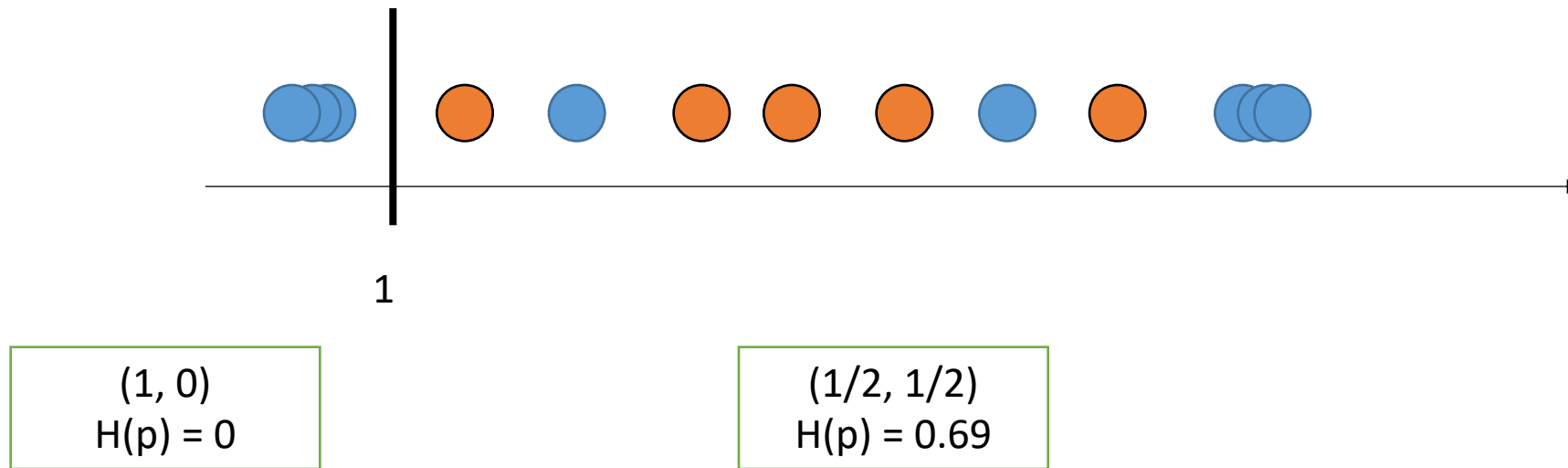
Признаки



Разбиения по признаку 1

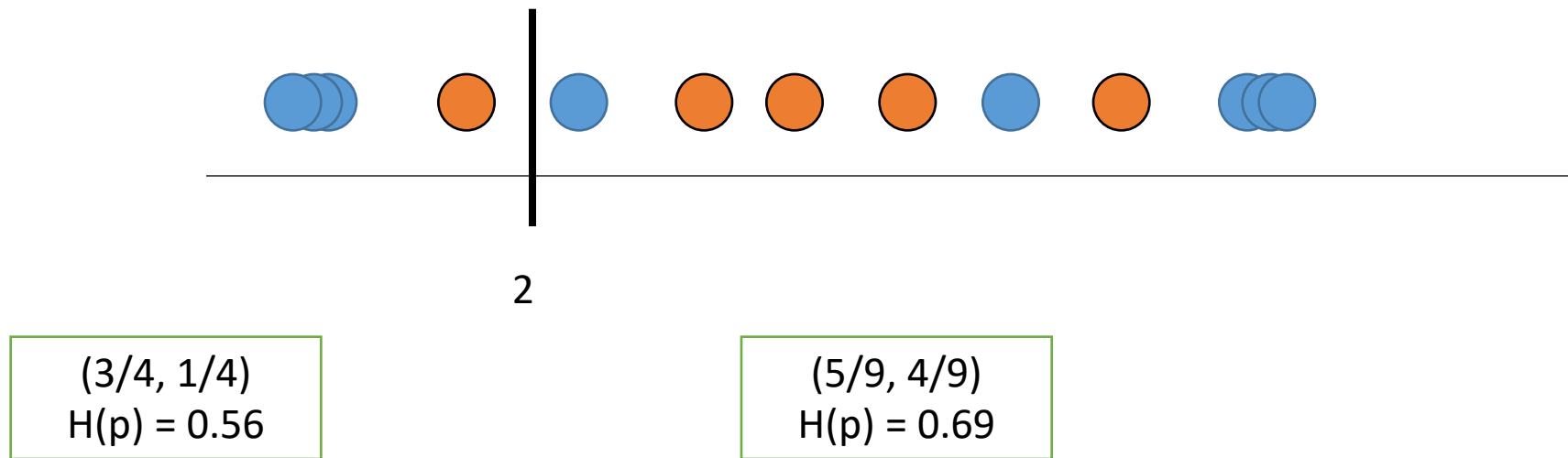


Разбиения по признаку 1



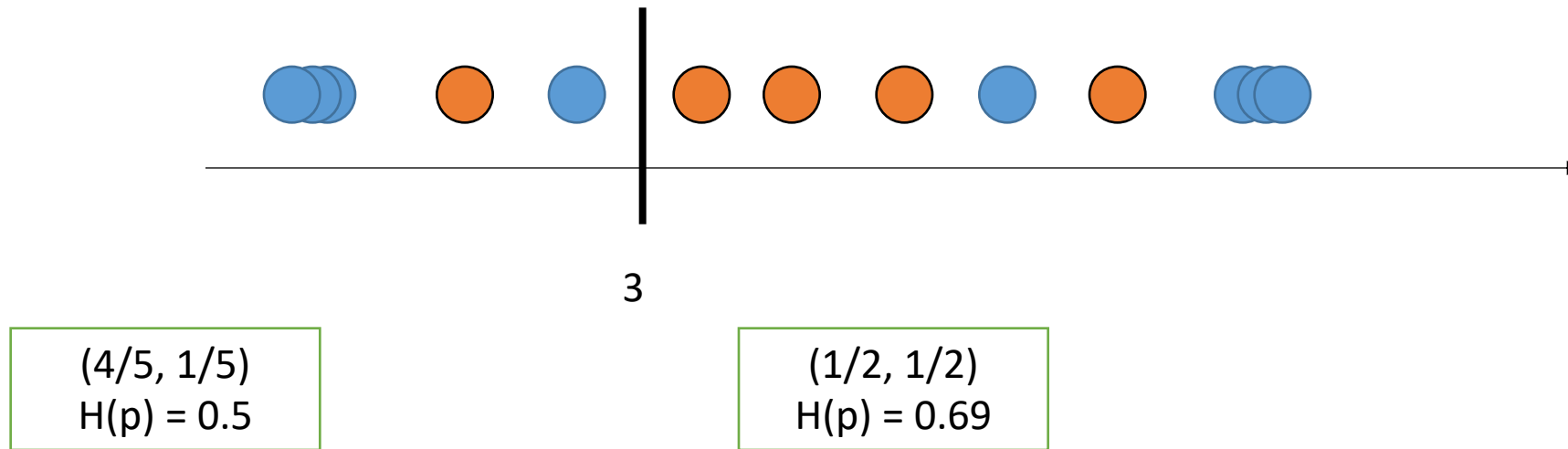
$$\frac{3}{13}H(p_l) + \frac{10}{13}H(p_r) = 0.53$$

Разбиения по признаку 1



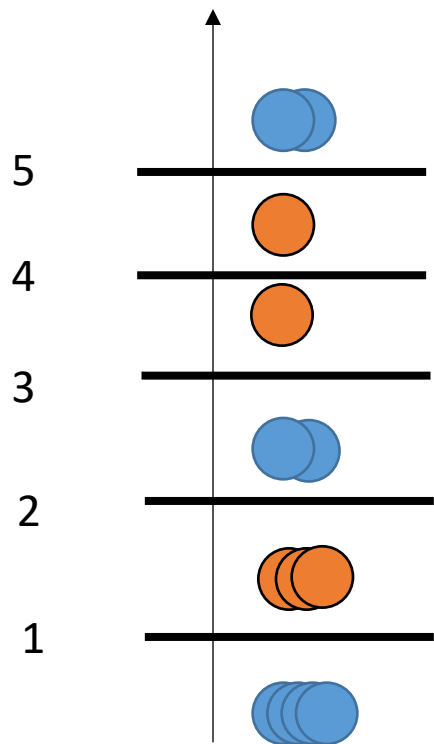
$$\frac{4}{13}H(p_l) + \frac{9}{13}H(p_r) = 0.65$$

Разбиения по признаку 1

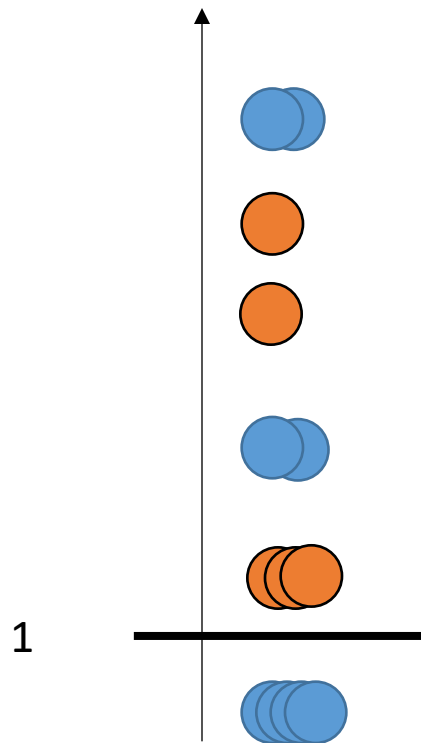


$$\frac{5}{13}H(p_l) + \frac{8}{13}H(p_r) = 0.62$$

Разбиения по признаку 2



Разбиения по признаку 2

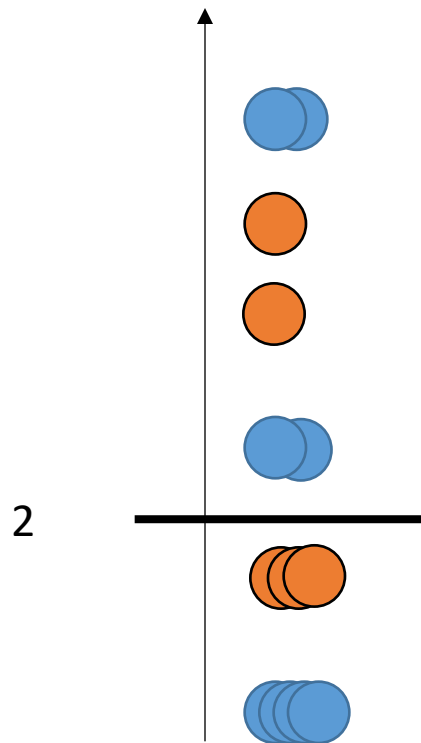


$(4/9, 5/9)$
 $H(p) = 0.69$

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

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

Разбиения по признаку 2

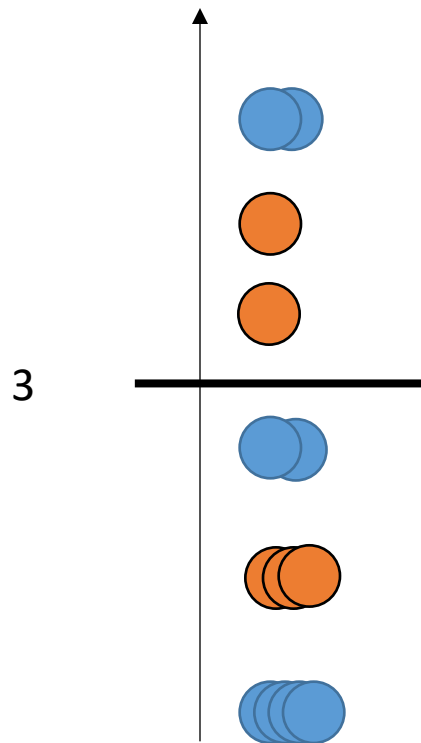


$(4/6, 2/6)$
 $H(p) = 0.64$

$(4/7, 3/7)$
 $H(p) = 0.68$

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

Разбиения по признаку 2

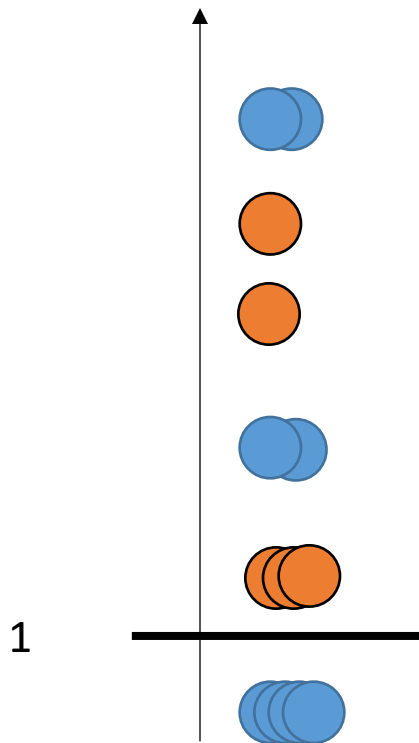


$(1/2, 1/2)$
 $H(p) = 0.69$

$(6/9, 3/9)$
 $H(p) = 0.46$

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

Разбиения по признаку 2



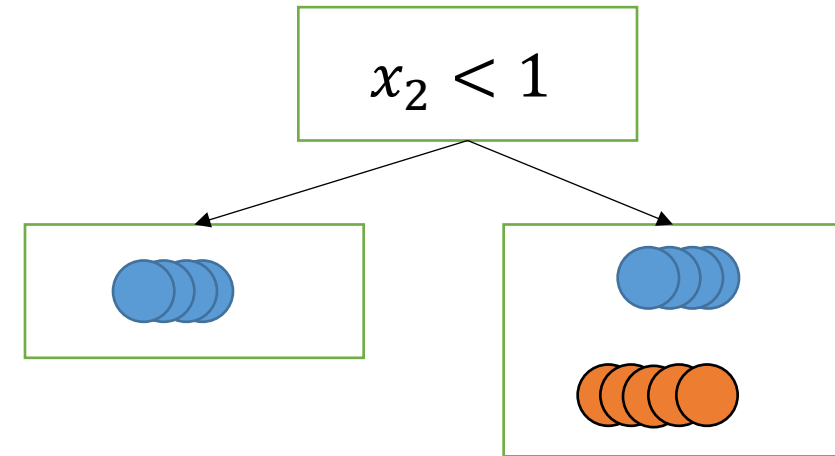
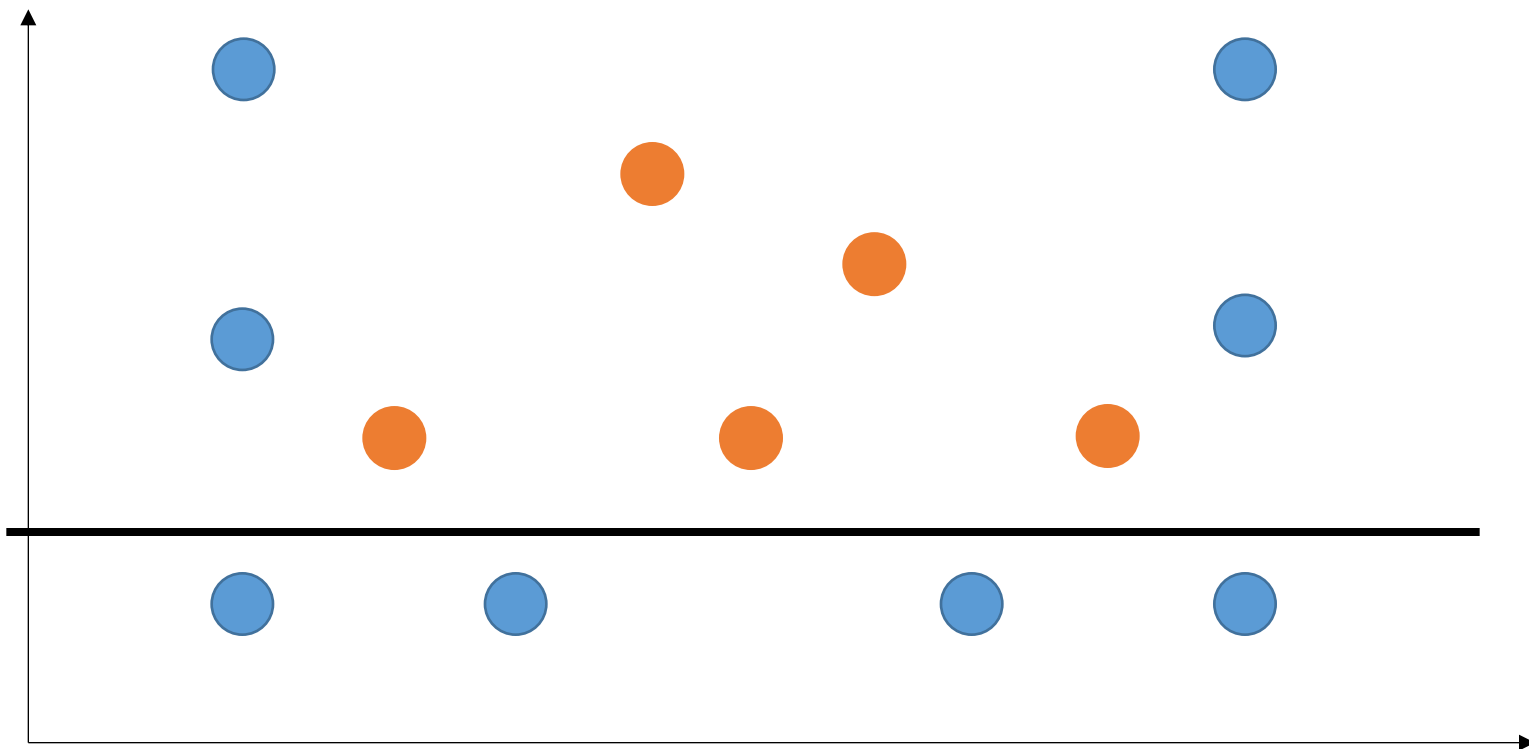
$(4/9, 5/9)$
 $H(p) = 0.69$

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

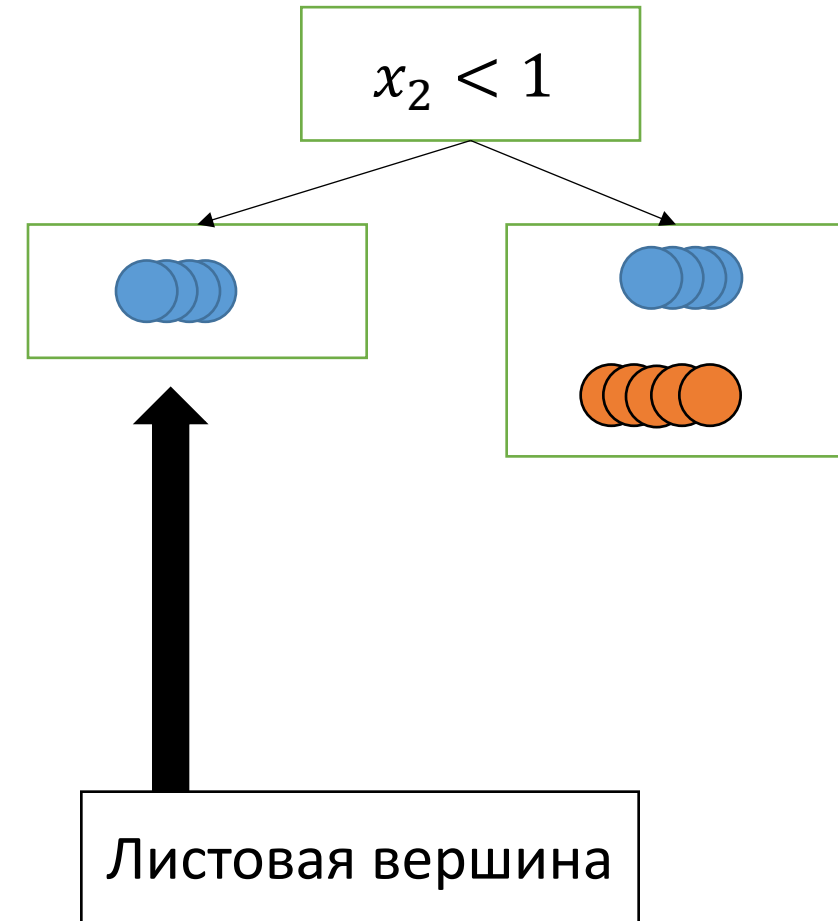
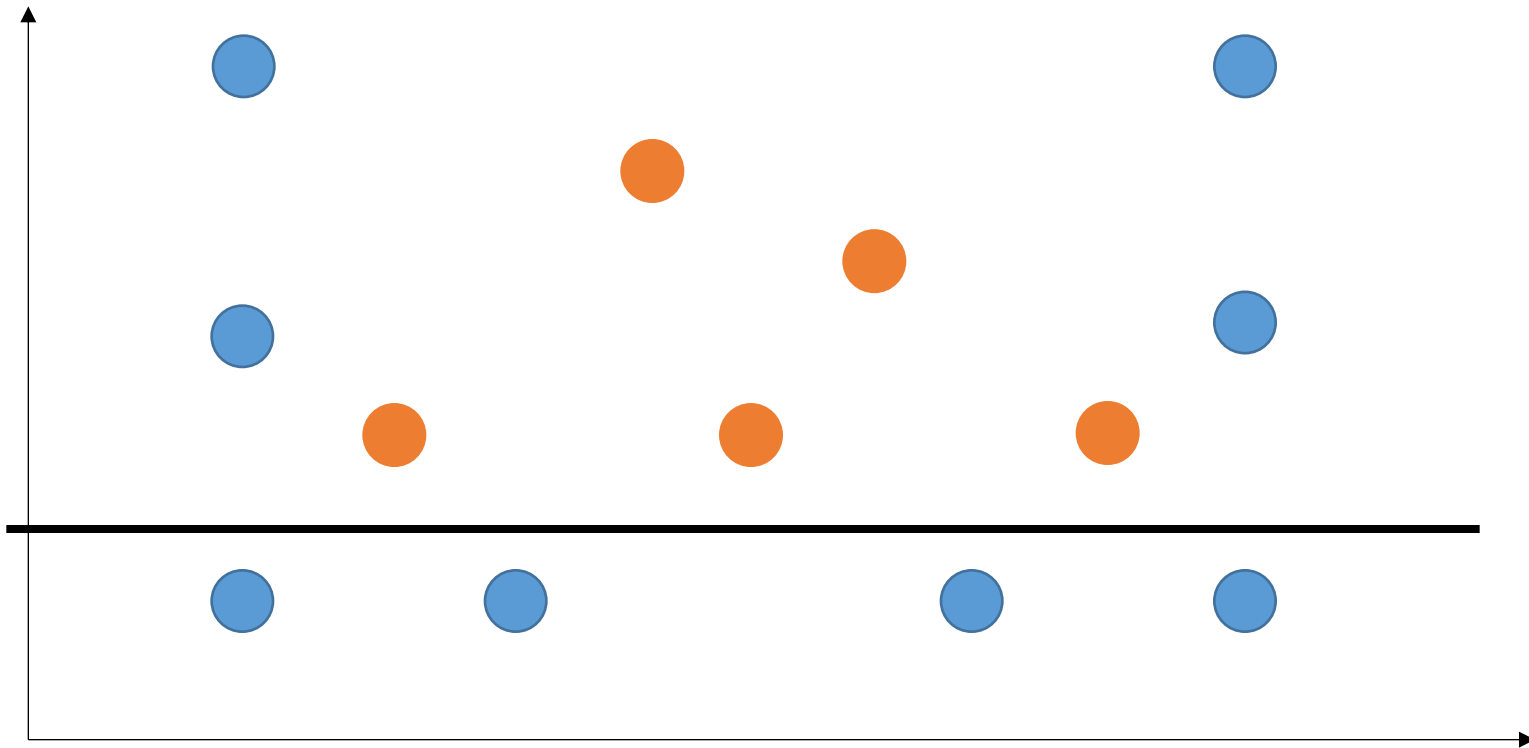
$$\frac{4}{13}H(p_l) + \frac{9}{13}H(p_r) = 0.47$$

Лучшее разбиение!

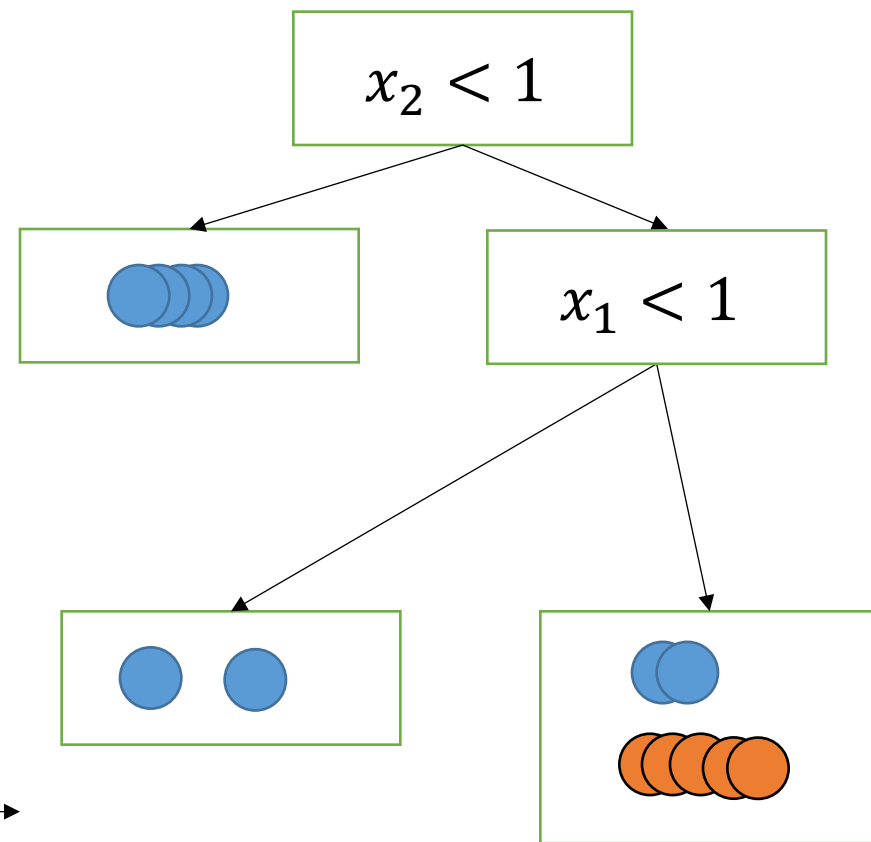
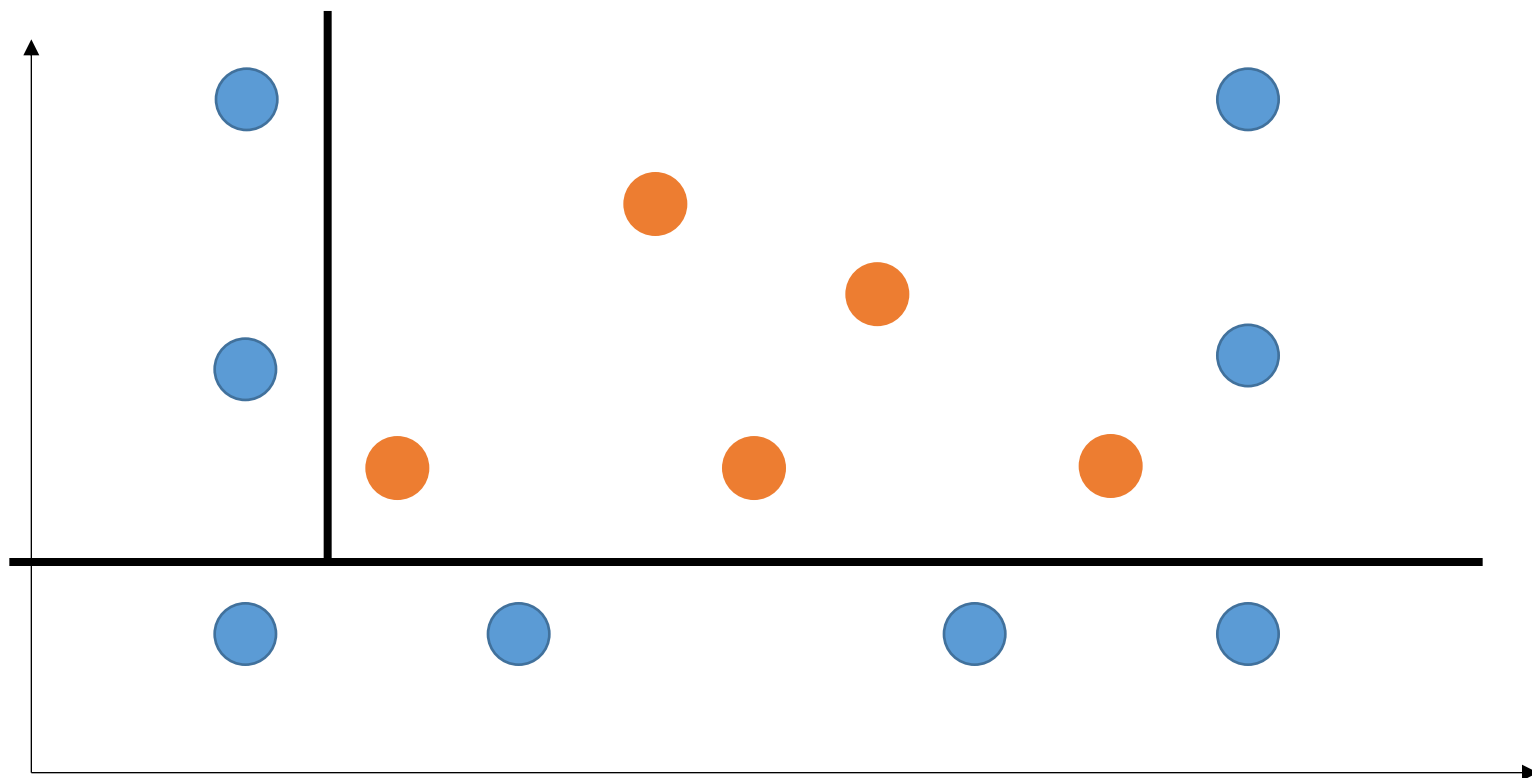
Обучение деревьев



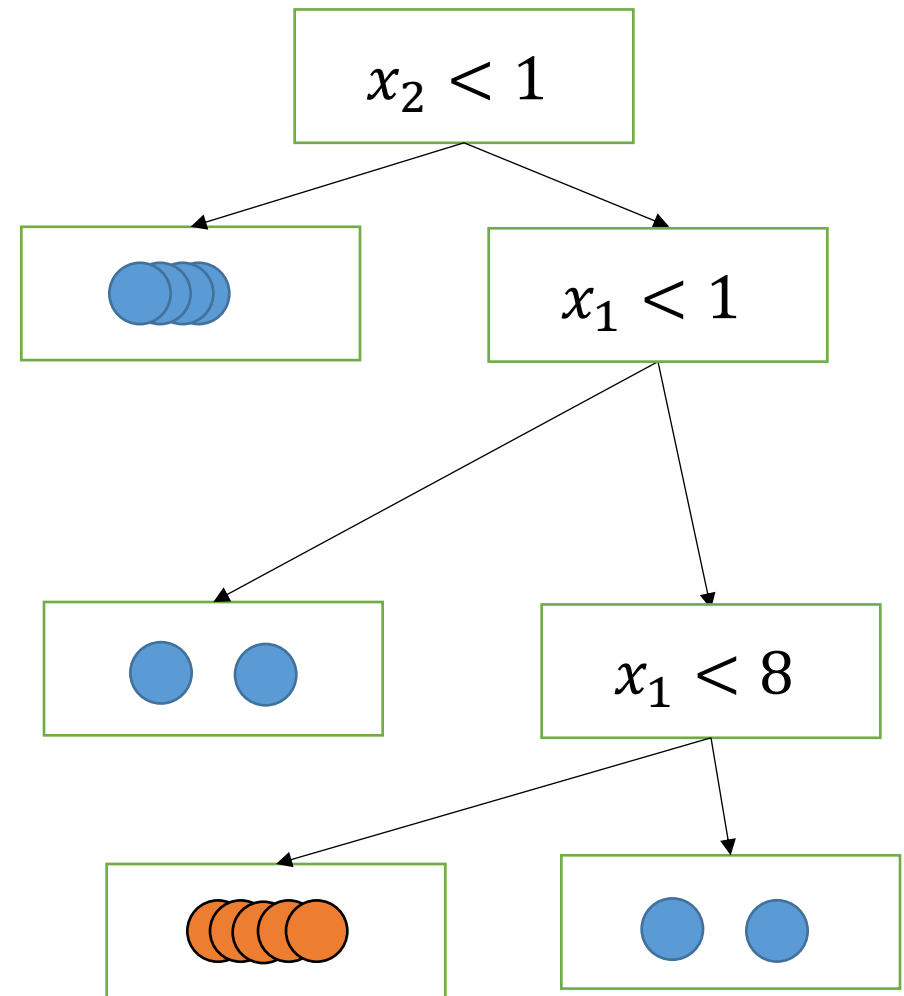
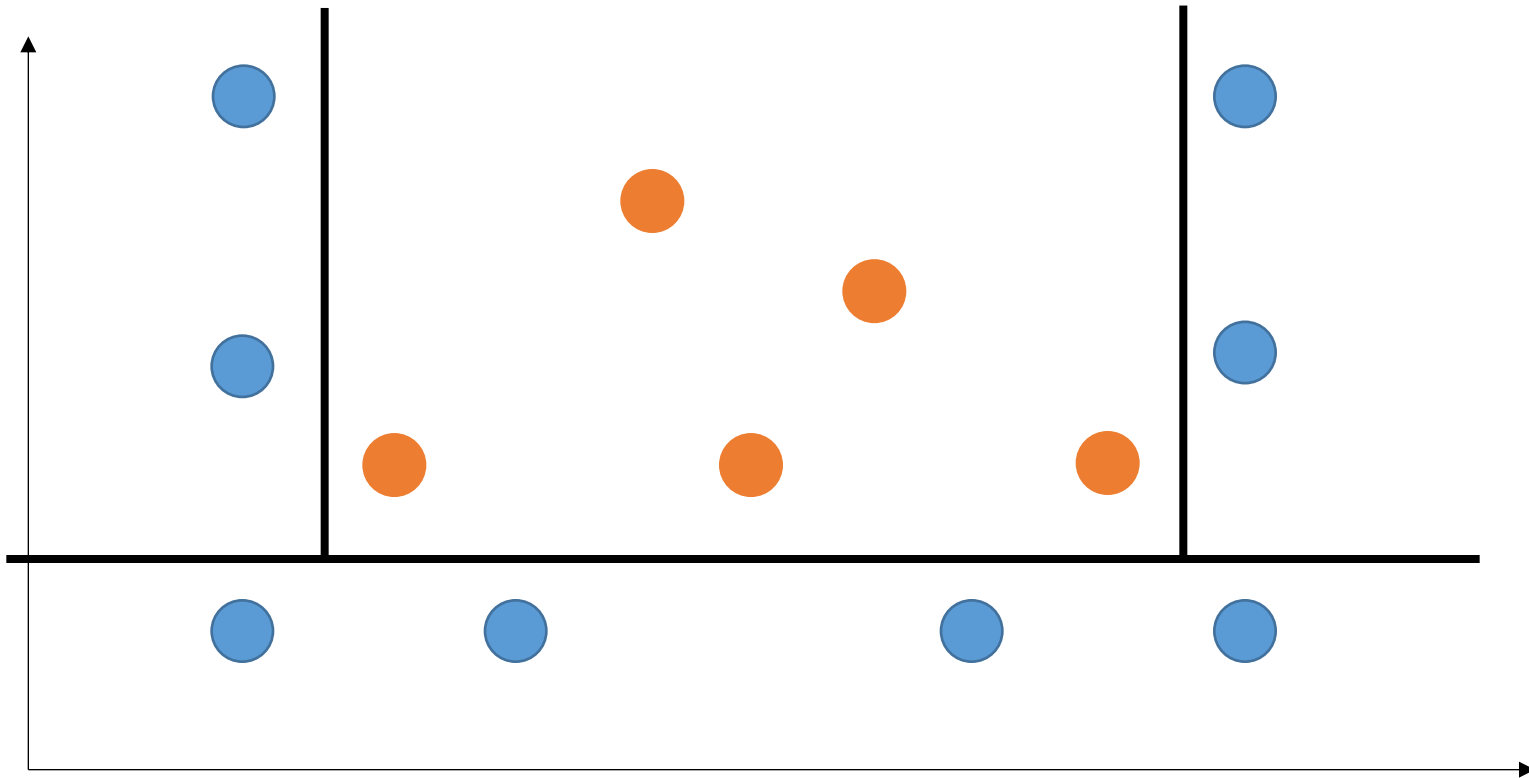
Обучение деревьев



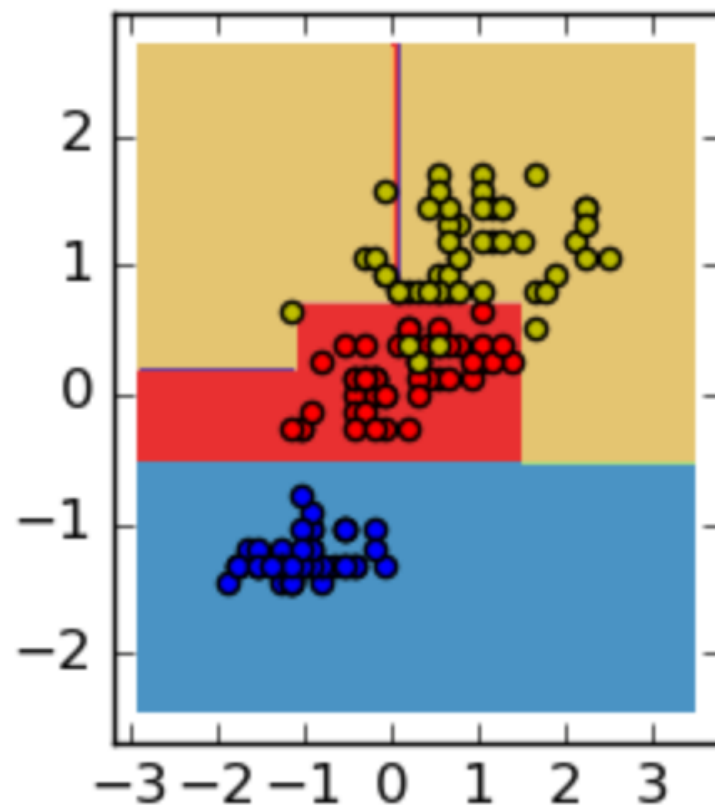
Обучение деревьев



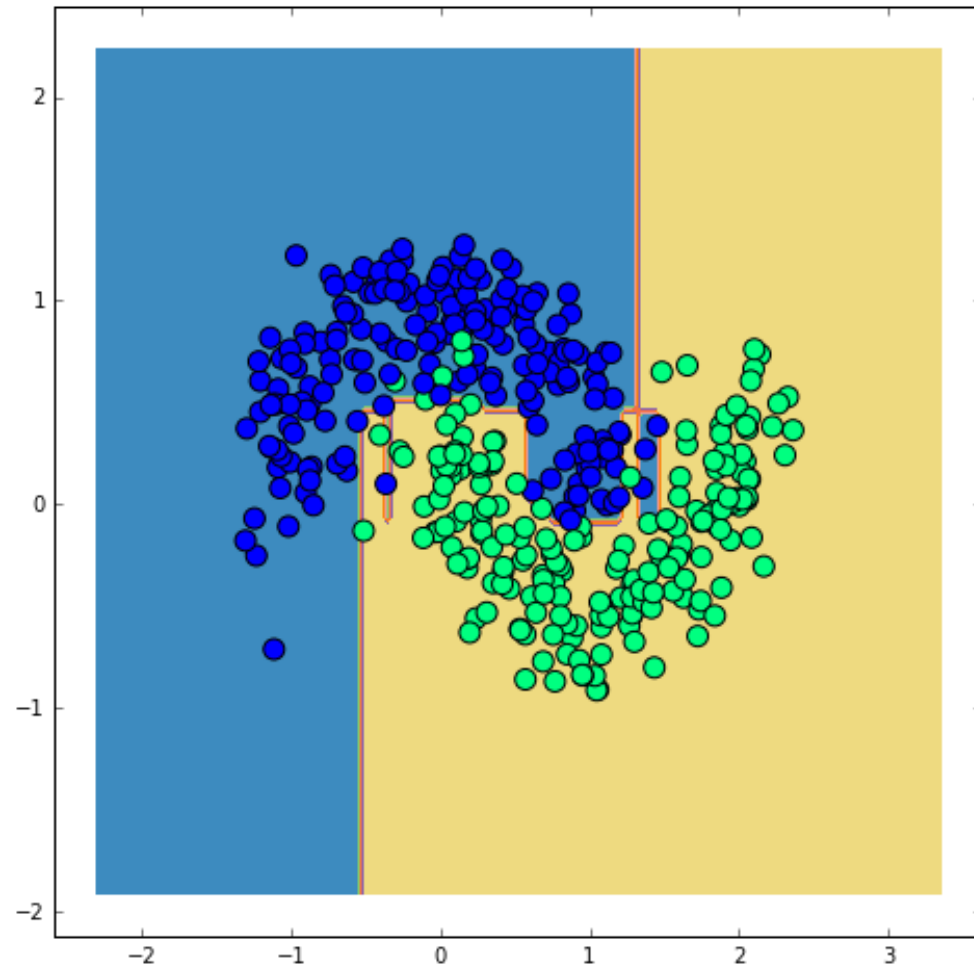
Обучение деревьев



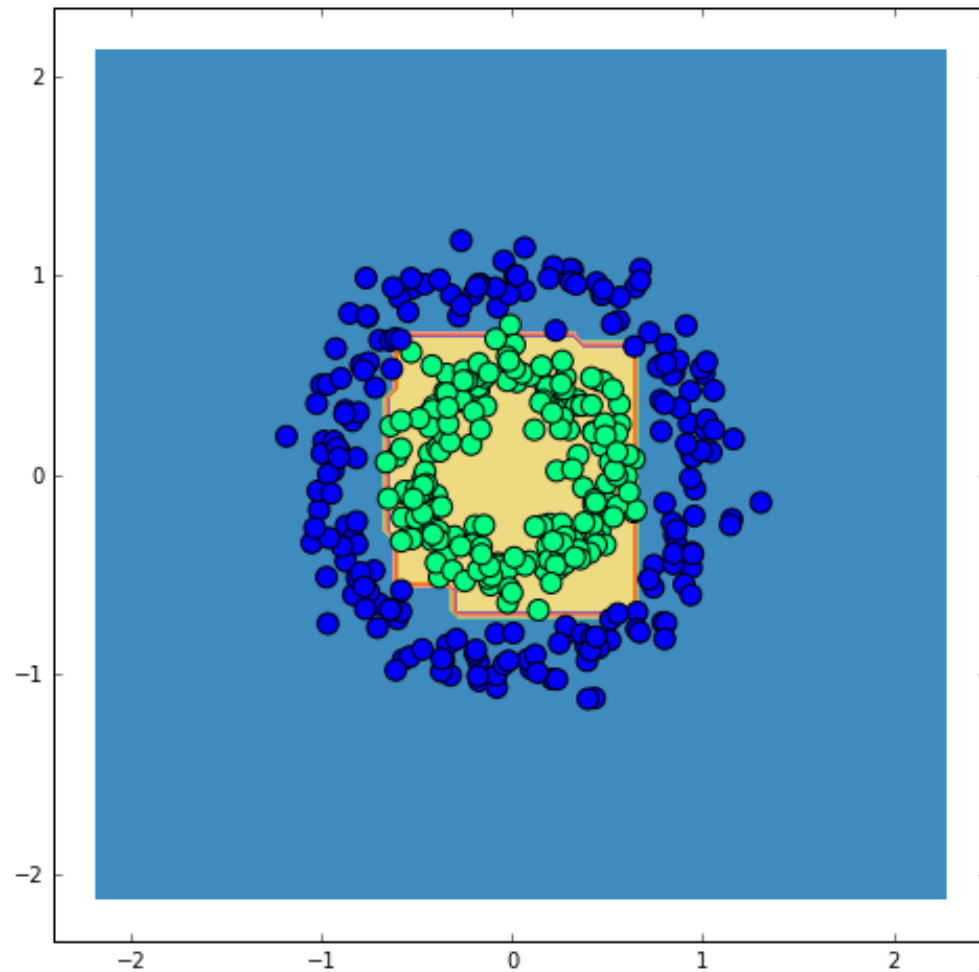
Классификация



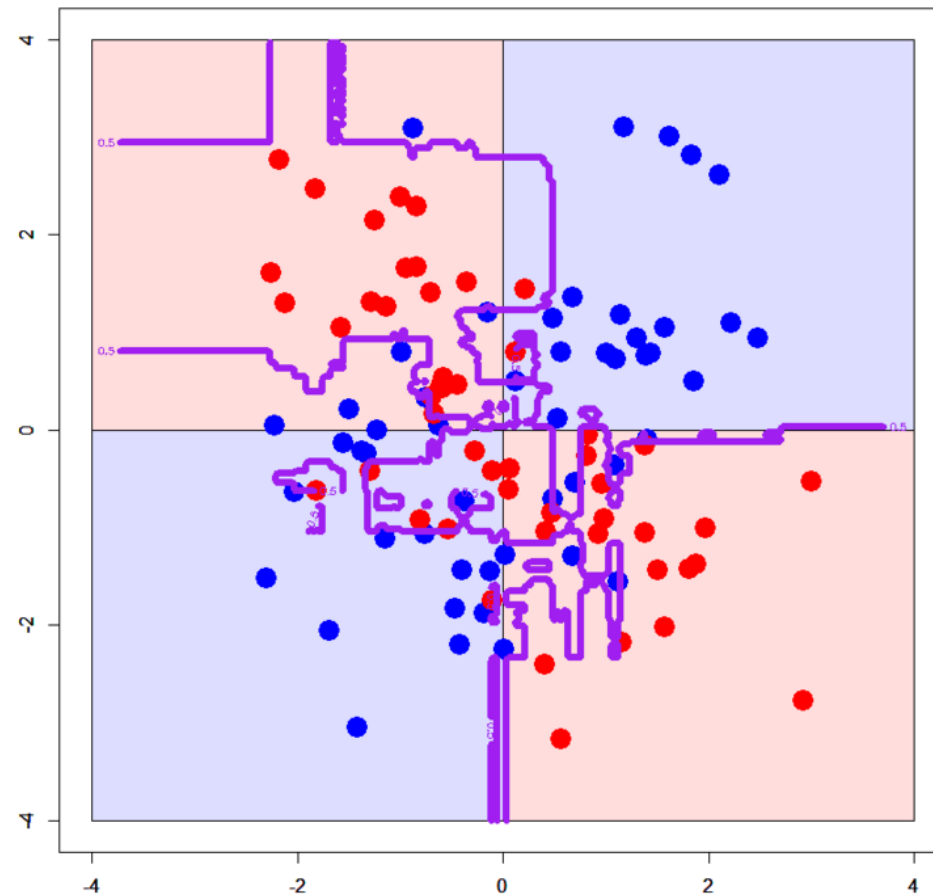
Классификация



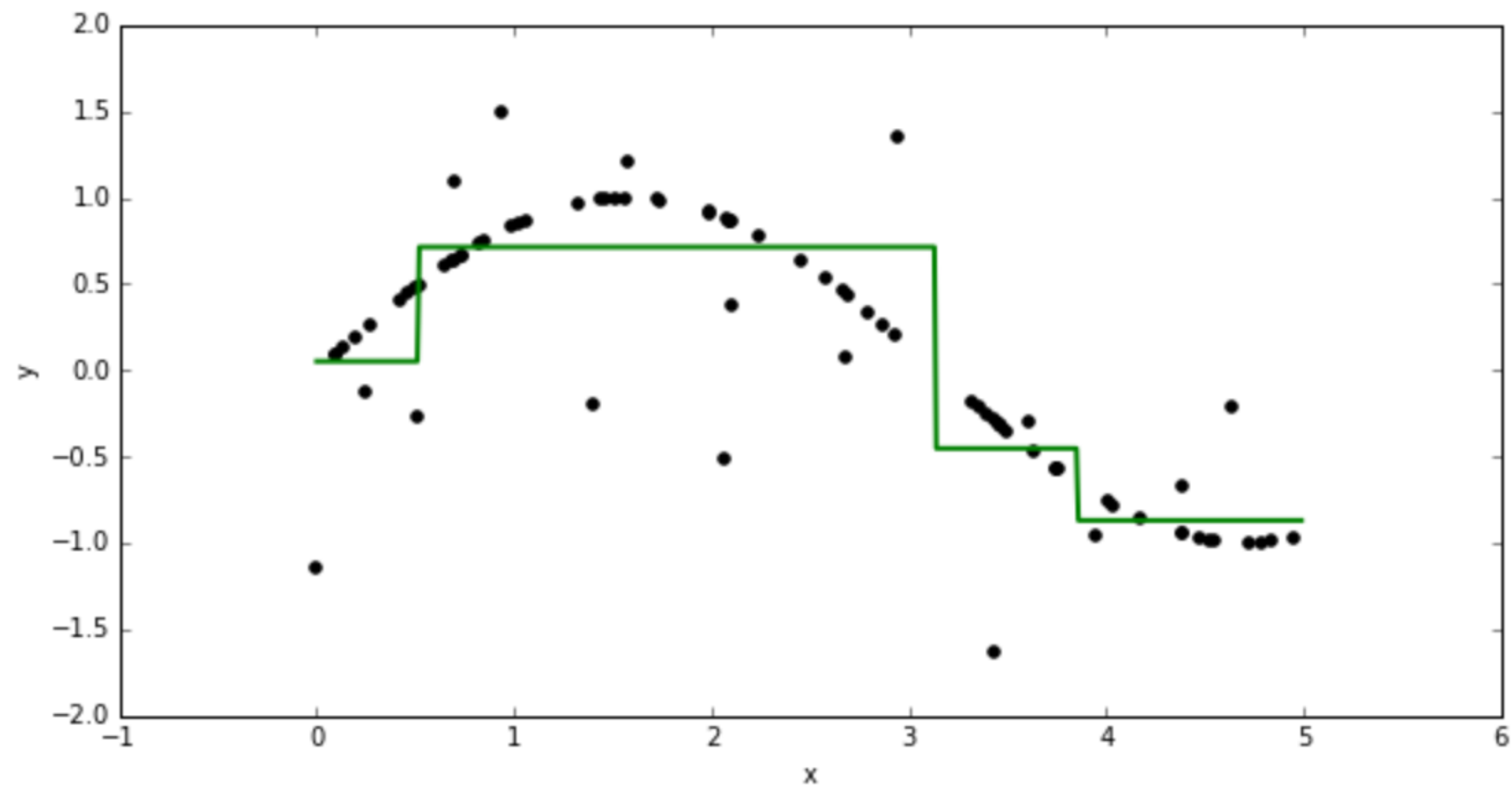
Классификация



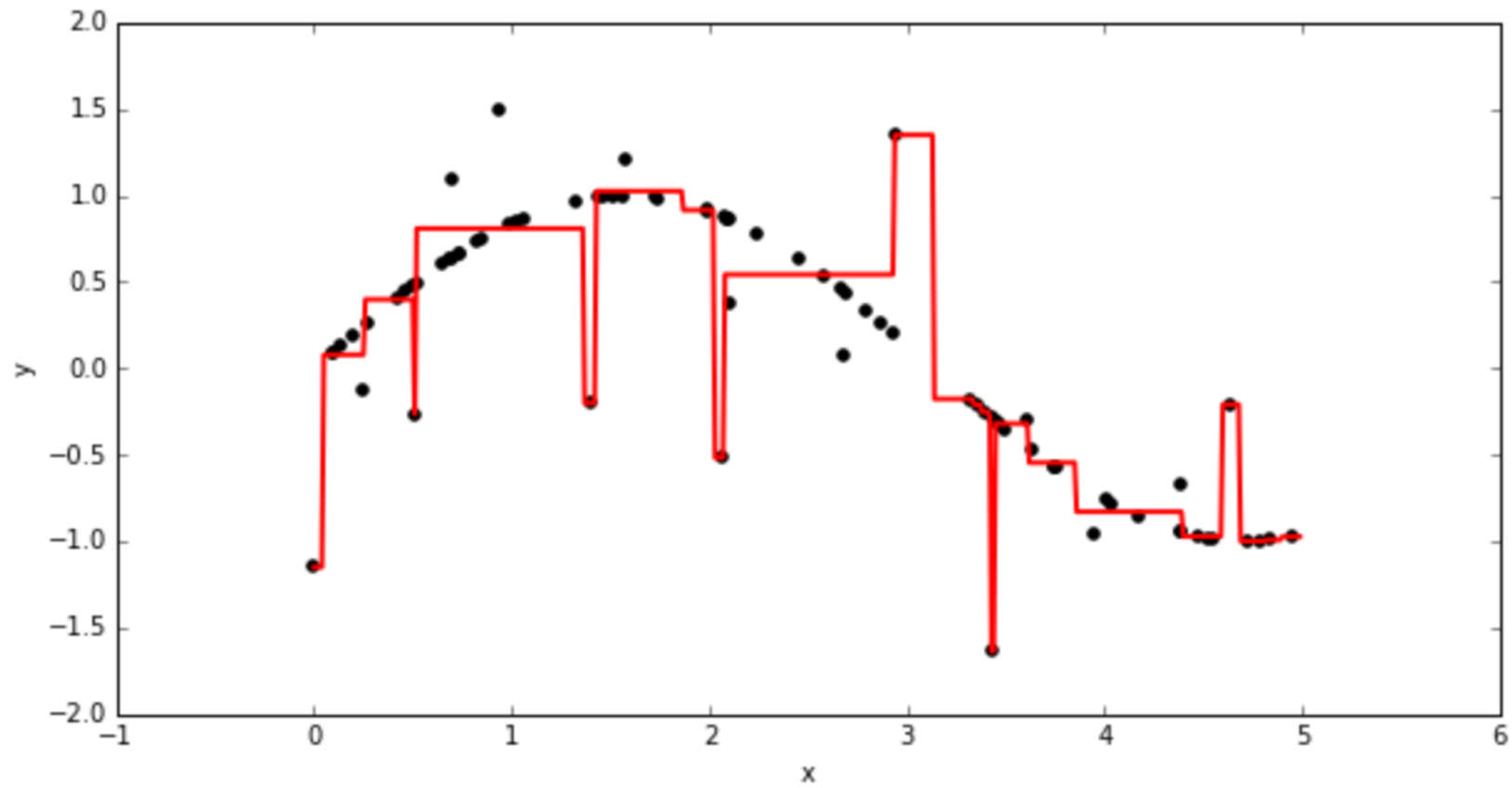
Классификация



Регрессия



Регрессия



Резюме

- Линейные классификаторы разделяют классы гиперплоскостью
- Качество классификации: доля правильных ответов, точность и полнота
- Деревья:
 - Восстанавливают сложные закономерности
 - Могут построить сколь угодно сложную поверхность
 - Чем больше глубина — тем сложнее поверхность
 - Склонны к переобучению