Решающие деревья

Лекция 4

Тест

Повторение

X – множество объектов (их признаковое описание)

Y – множество истинных ответов

$\hat{Y}$ – множество предсказанных ответов. Получаем по формуле:

$$\hat{Y} = f(X),$$

где $f(X)$ – модель машинного обучения

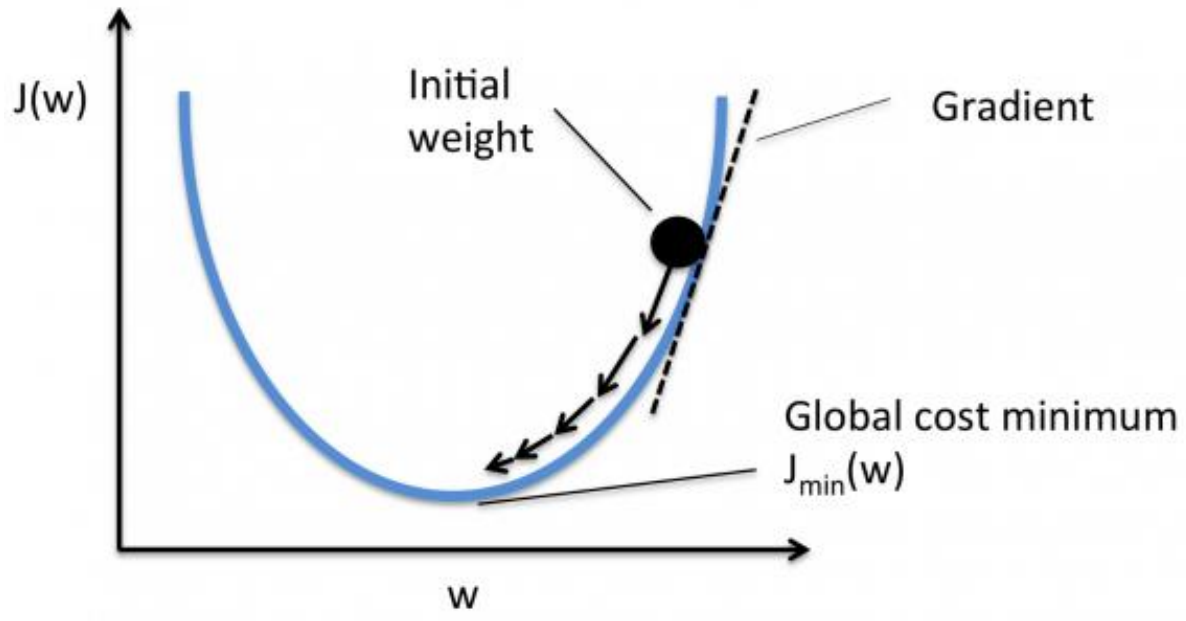
Если модель линейная, то

$$\hat{Y} = X^T W$$

Повторение

Обновление весов модели происходит с учетом **антиградиента функции потерь**.

$$\boldsymbol{W = W - \eta \cdot \nabla_W Q(W)}$$

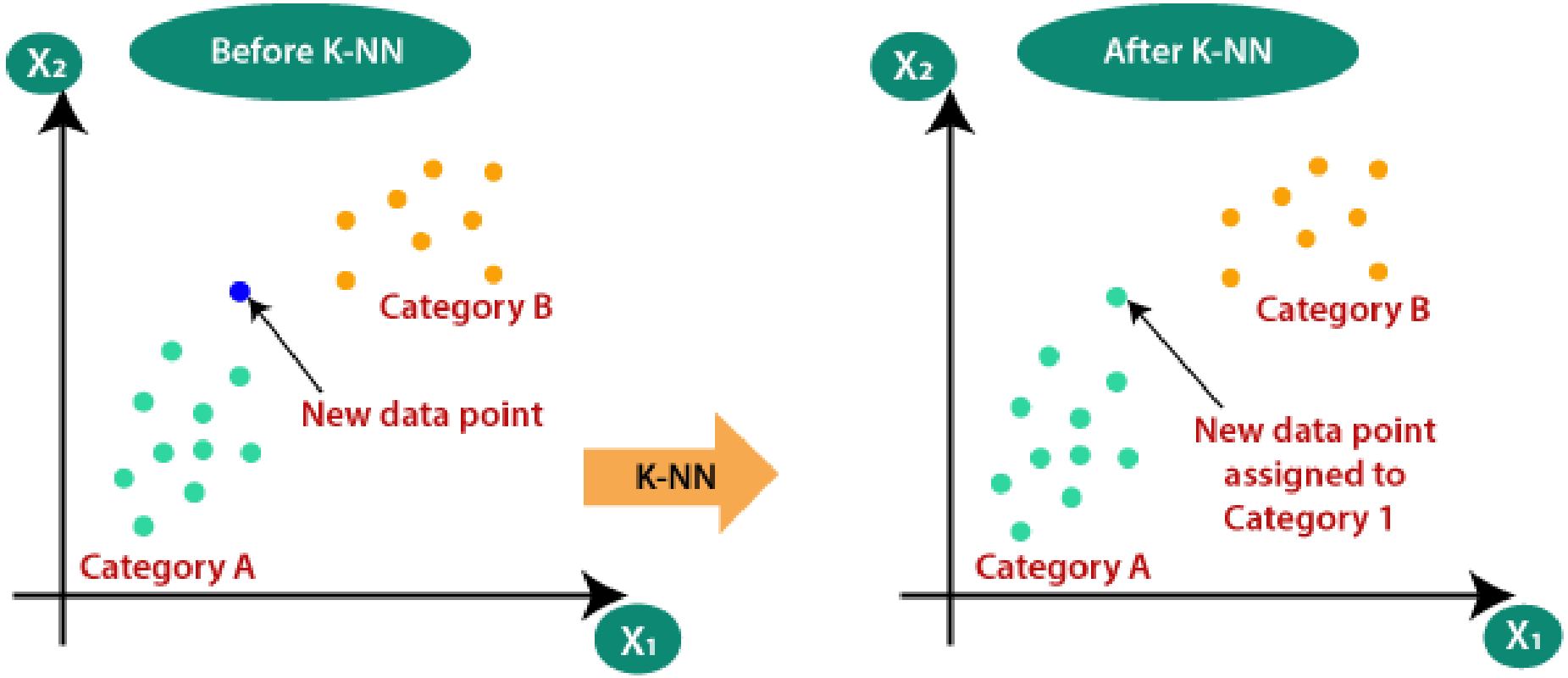


Новый мир без градиента

Есть модели, где градиент не нужен:

- K-ближайших соседей
- Наивный байесовский классификатор (регрессор)
- Решающее дерево

K-ближайших соседей



К-ближайших соседей. Расстояние

Варианты расстояний:

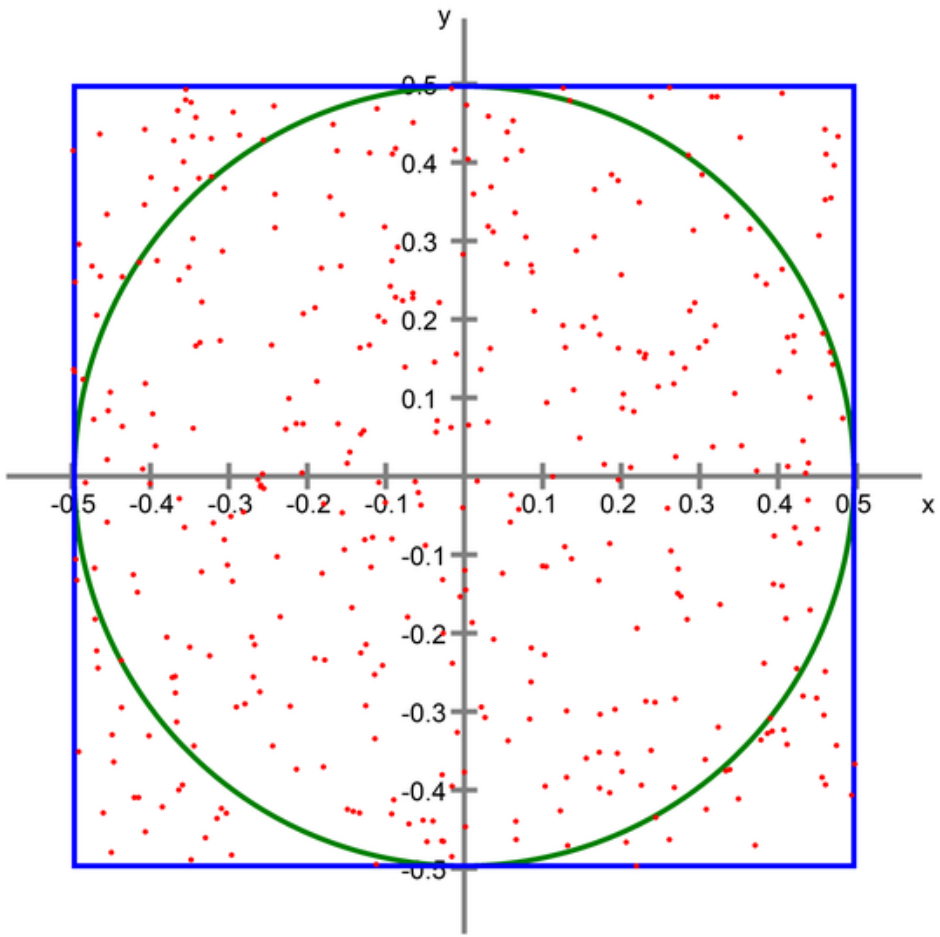
Евклидово (L2): $Q\left(x^k, x^m\right) = \sum_{i=1}^{n} \sqrt{\left(x_i^k - x_i^m\right)^2}$

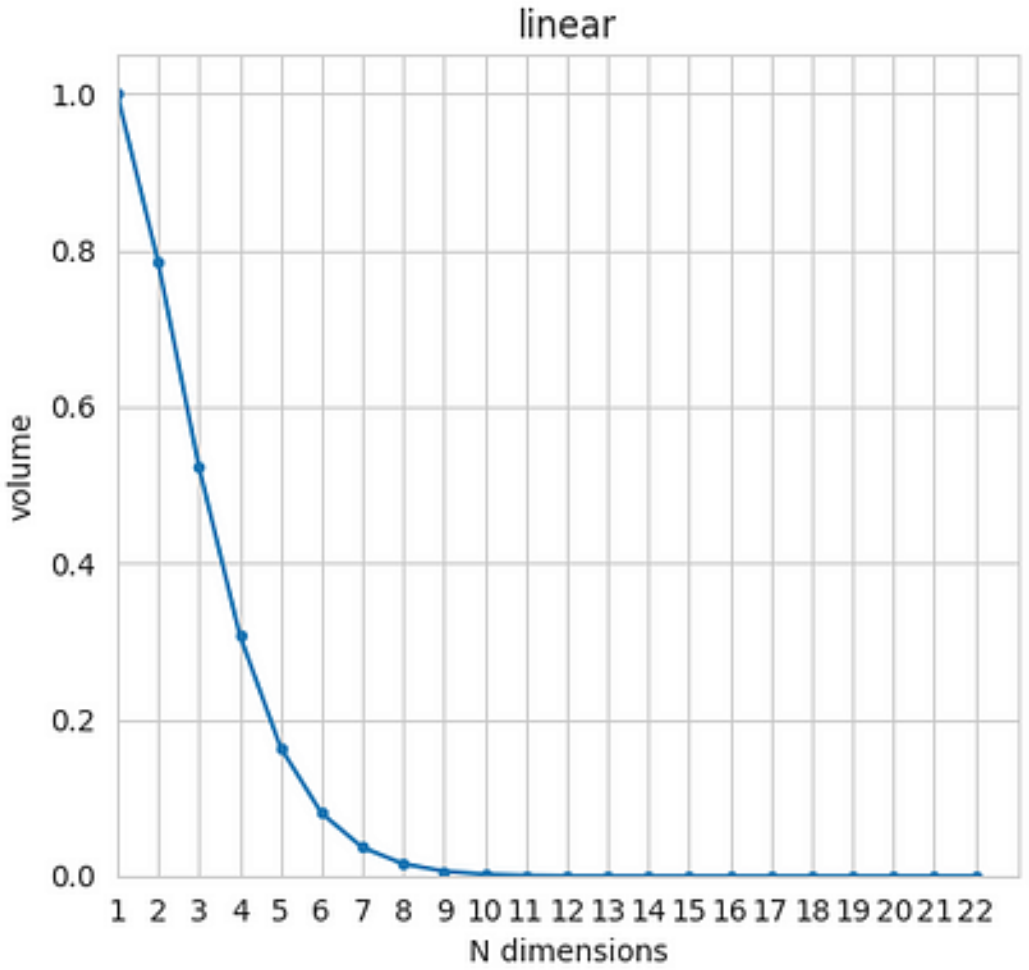
Манхэттенское расстояние (*L1*): $Q\left(x^k, x^m\right) = \sum_{i=1}^{n} |x_i^k - x_i^m|$

Косинусное расстояние: $Q\left(x^k, x^m\right) = \frac{\sum_{i=1}^{n} x_i^k \cdot x_i^m}{\sum_{i=1}^{n} \sqrt{\left(x_i^k\right)^2} \cdot \sum_{i=1}^{n} \sqrt{\left(x_i^m\right)^2}}$

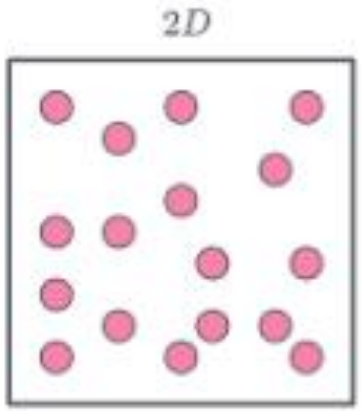
Необходимо масштабировать признаки при использовании KNN.

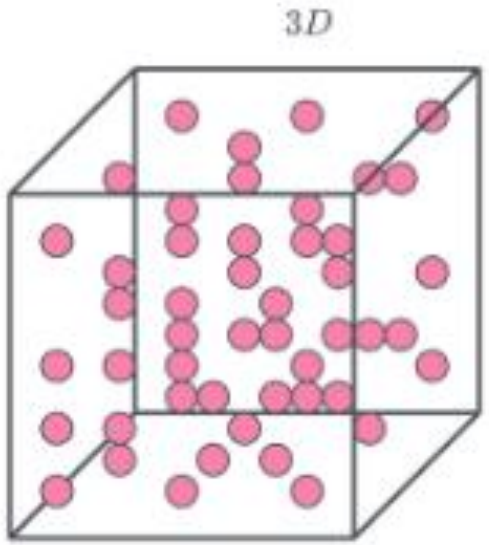
Проклятие размерности

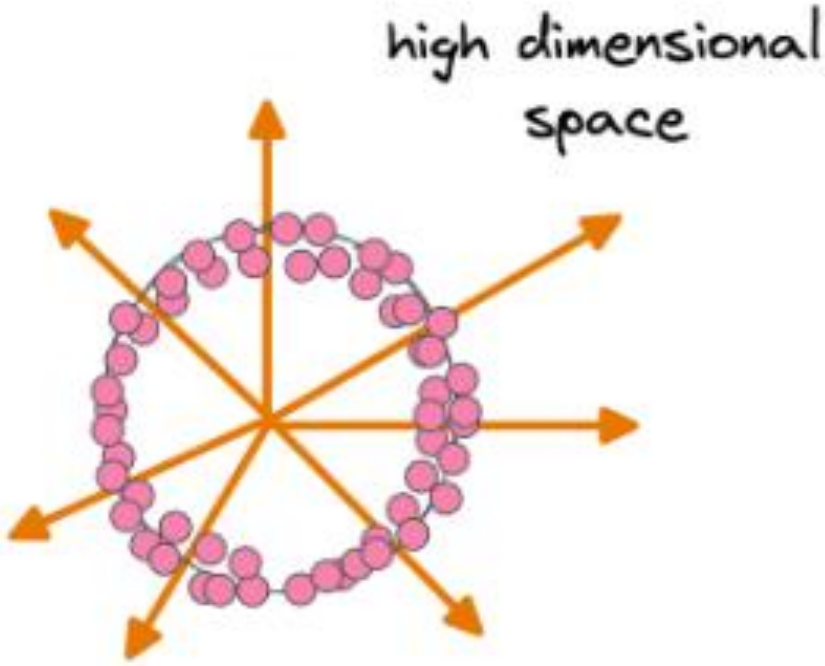




Проклятие размерности







Вопрос на засыпку

В выборке 1000 человек, из них 400 мужчин, 600 женщин.

Целевых событий ($y = 1$) 100 штук, 80 из которых у мужчин.

Какова вероятность целевого событий при условии, что перед нами мужчина?

Наивный байесовский классификатор

Теорема Байеса:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

где $P(A)$ – априорная вероятность события A,

$P(B)$ – априорная вероятность события B,

$P(B|A)$ – условная вероятность наступления события B при наступлении события A (правдоподобие)

Наивный байесовский классификатор

Например:

В выборке 1000 человек, из них 400 мужчин, 600 женщин.

Целевых событий ($y = 1$) 100 штук, 80 из которых у мужчин.

Какова вероятность целевого событий при условии, что перед нами мужчина?

Перепишем в терминах машинного обучения:

$$P(y = 1 | X = \text{«Мужчина»}) = \frac{P(y = 1) \cdot P(X = \text{«Мужчина»} | y = 1)}{P(X = \text{«Мужчина»})}$$

Наивный байесовский классификатор

1000 человек, 400 М, 600 Ж. 100 «1», из них 80 М, 20 Ж.

$$P(y = 1) = \frac{100}{1000} = 0.1$$

$$P(X = \text{«Мужчина»}) = \frac{400}{1000} = 0.4$$

$$P(X = \text{«Мужчина»} | y = 1) = \frac{80}{100} = 0.8$$

Наивный байесовский классификатор

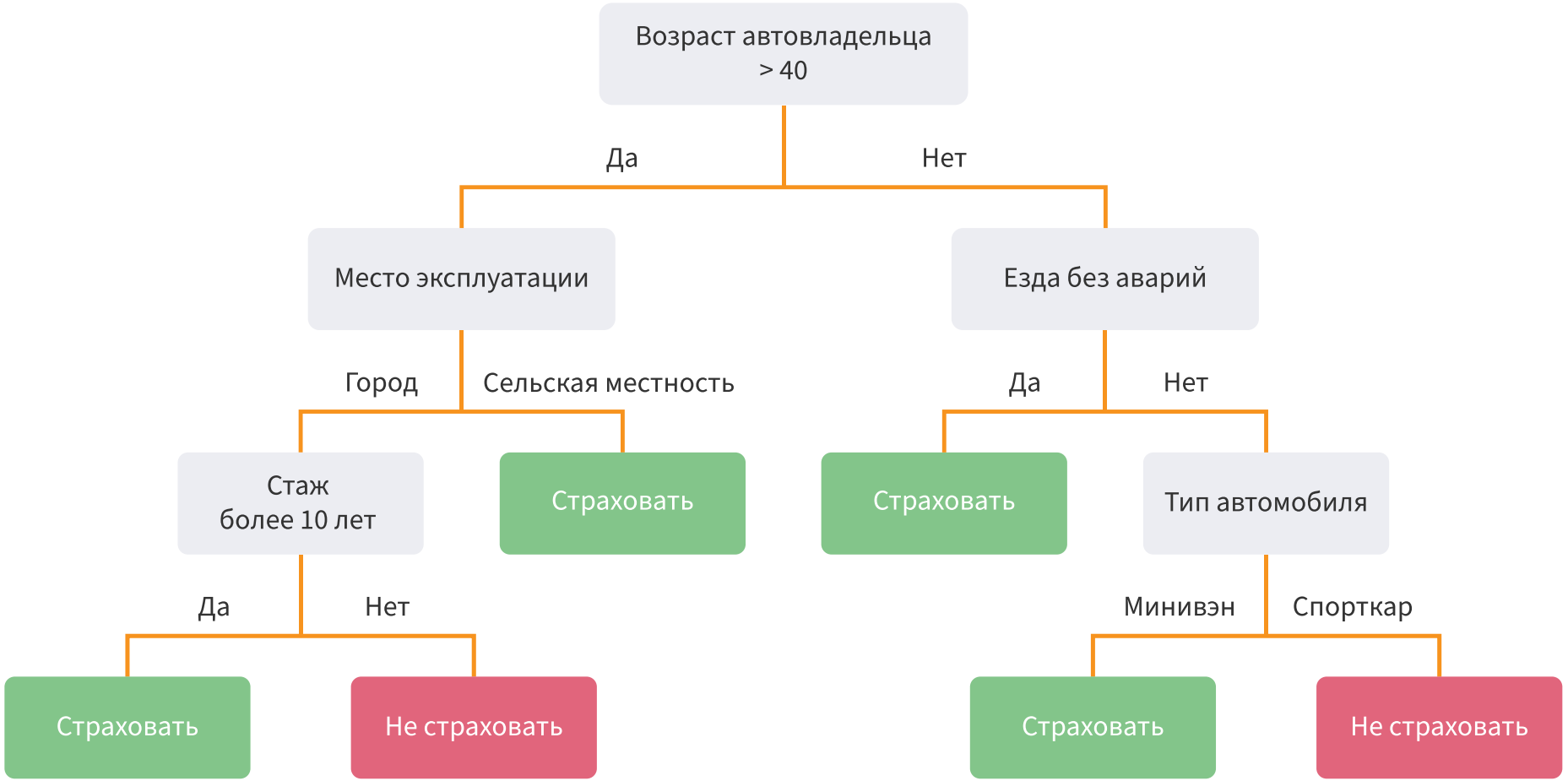
Тогда

$$P(y = 1|X = \text{«Мужчина»}) = \frac{0.1 \cdot 0.8}{0.4} = 0.2 = 20\%$$

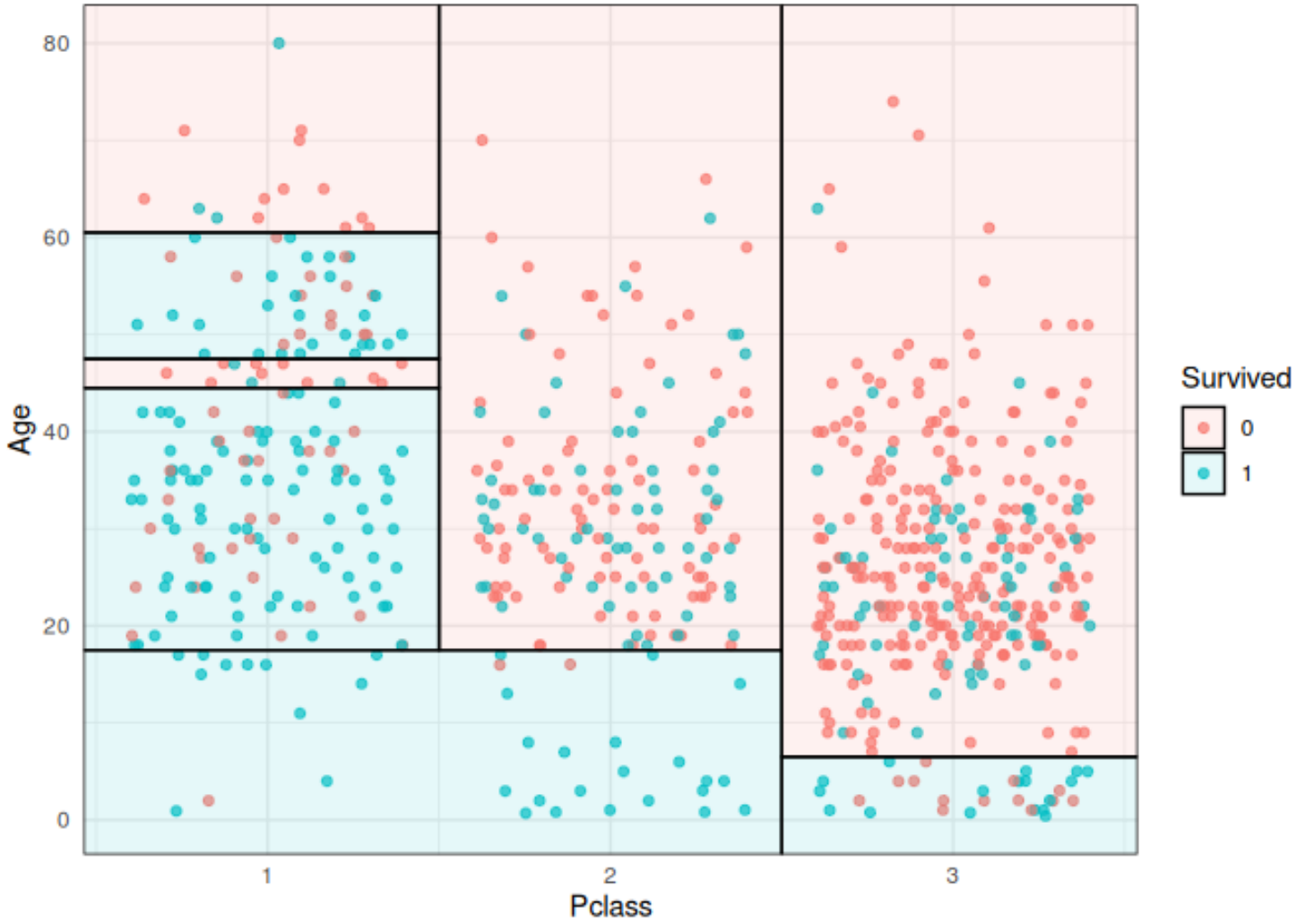
Если признаков несколько, то:

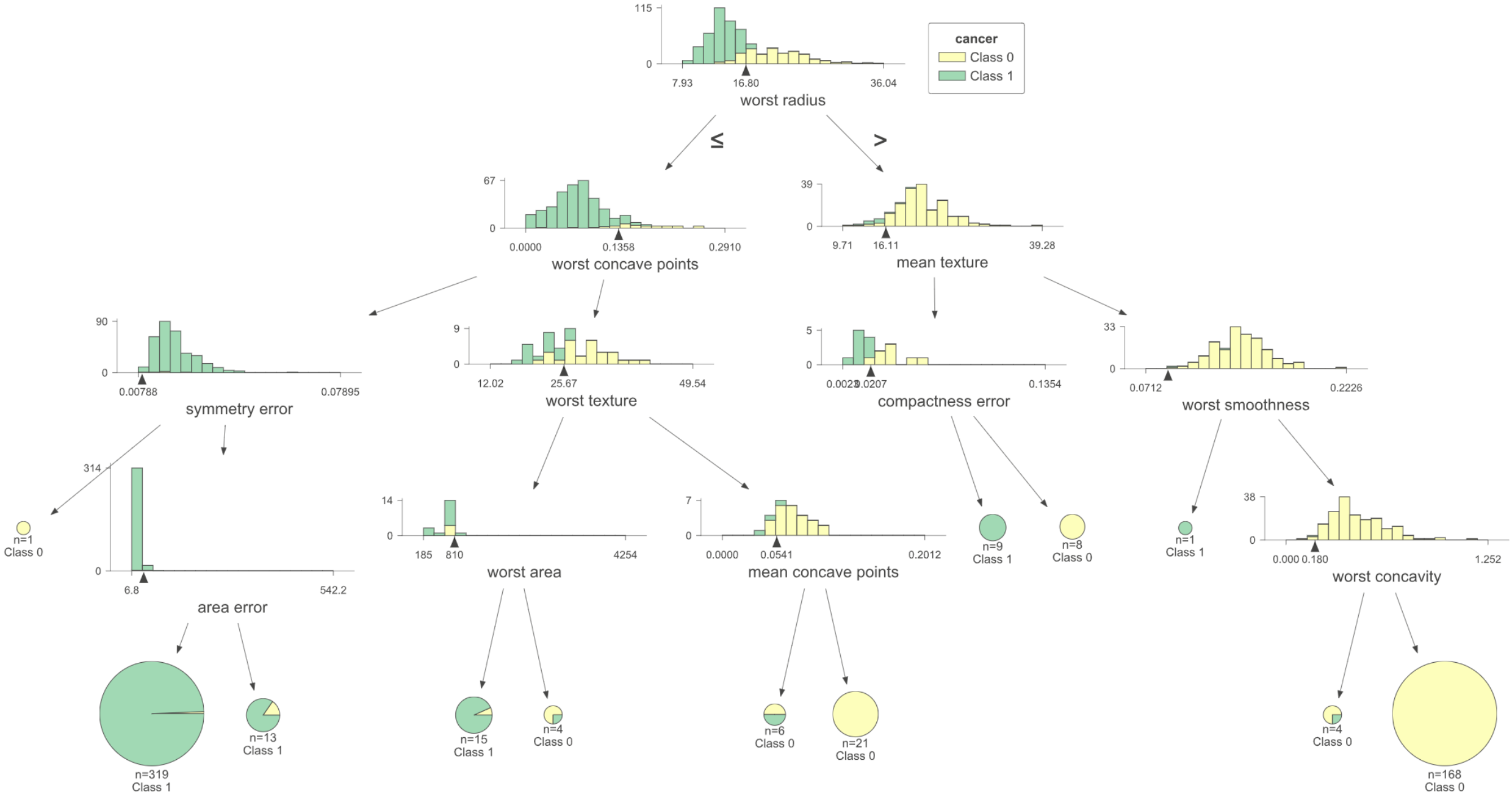
$$P(y = 1|X_1, X_2) = \frac{P(y = 1) \cdot P(X_1|y = 1) \cdot P(X_2|y = 1)}{P(X_1, X_2)}$$

Решающее дерево



Решающее дерево




cancer
Class 0
Class 1
115
0
7.93
16.80
36.04
worst radius
<=
>
67
0
0.0000
0.1358
0.2910
worst concave points
39
0
9.71
16.11
39.28
mean texture
90
0
0.00788
0.07895
symmetry error
9
0
12.02
25.67
49.54
worst texture
5
0
0.0023
0.0207
0.1354
compactness error
33
0
0.0712
0.2226
worst smoothness
314
0
6.8
542.2
area error
14
0
185
810
4254
worst area
7
0
0.0000
0.0541
0.2012
mean concave points
38
0
0.000
0.180
1.252
worst concavity
n=1
Class 0
n=319
Class 1
n=13
Class 1
n=15
Class 1
n=4
Class 0
n=6
Class 0
n=21
Class 0
n=9
Class 1
n=8
Class 0
n=1
Class 1
n=4
Class 0
n=168
Class 0

Решающее дерево. Критерий информативности

Хотим для каждого потенциального разбиения знать, насколько оно хорошо.

Опишем функционал качества $Q(R)$ по формуле:

$$\boldsymbol{Q(R, X_i, t) = H(R_{head}) - \frac{size_{left}}{size_{head}} H(R_{left}) - \frac{size_{right}}{size_{head}} H(R_{right})}$$

где $H(R)$ -- критерий информативности (impurity criterion)

Решающее дерево. Регрессия

Для регрессии введем критерий информативности:

$$H(R) = \frac{1}{N}\sum_{i=0}^{N}(y_i - c)^2 \to min$$

$$H(R) = \frac{1}{N}\sum_{i=0}^{N}\left(y_i - \frac{1}{N}\sum_{i=0}^{N} y_i\right)^2 \to min$$

Решающее дерево. Классификация

1. Оцениваем ошибку классификации

$$H(R) = \frac{1}{N}\sum_{i=0}^{N}[y_i \neq c] \rightarrow min$$

В листе дерево всегда предсказывает наиболее часто встречающийся класс, поэтому:

$$H(R) = \frac{1}{N}\sum_{i=0}^{N}[y_i \neq c] = 1 - p_k$$

Решающее дерево. Классификация

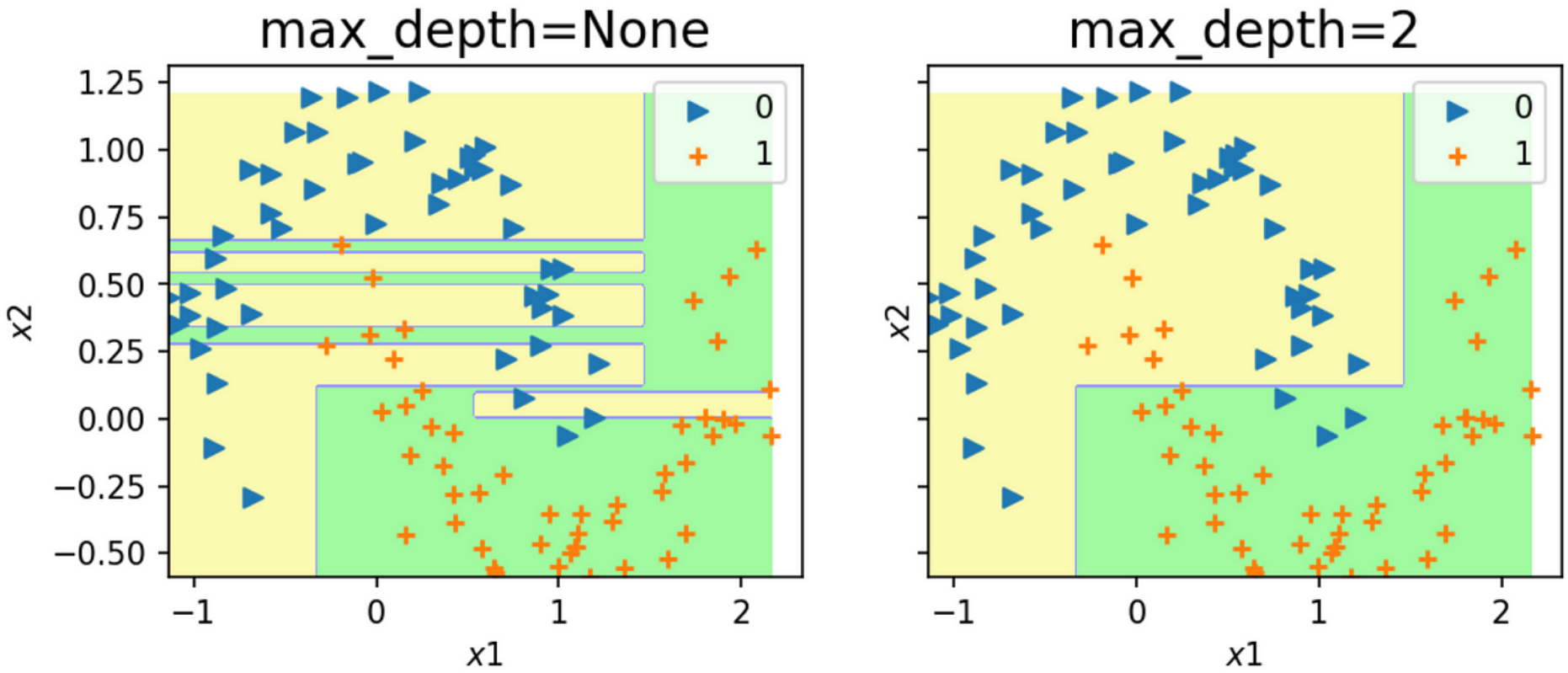
2. Критерий Джини

$$H(R) = -\sum_{k=1}^{K} p_k(1 - p_k) \to min$$

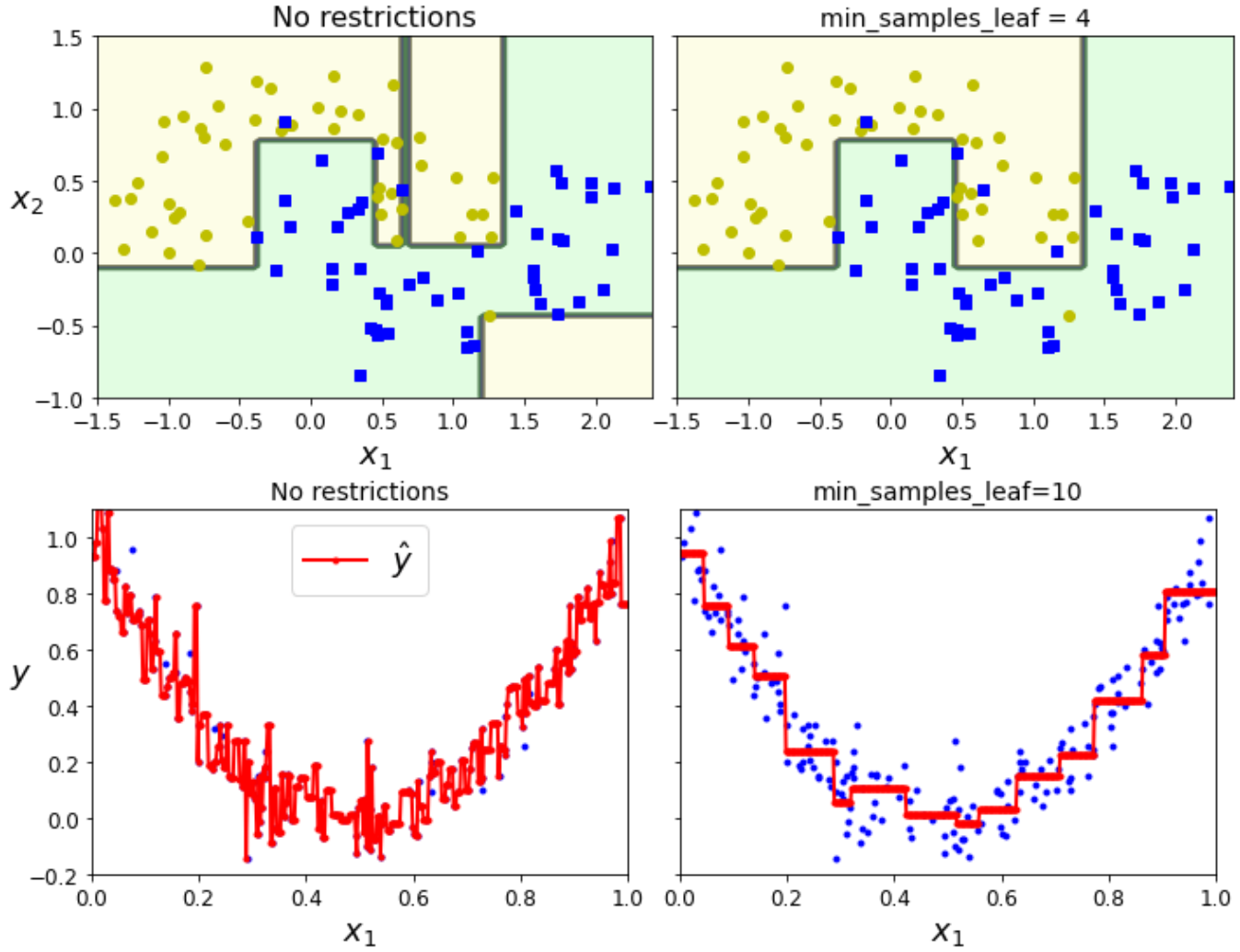
3. Энтропийный критерий

$$H(R) = -\sum_{k=1}^{K} p_k \log p_k \to min$$

Решающее дерево. Глубина



Решающее дерево. Количество наблюдений в листьях



Решающее дерево. Параметры

criterion – критерий разбиения. Для классификации Джини, энтропия или логлосс.

splitter – стратегия разбиение. Best или random.

max_depth – максимальная глубина дерева. От 1 до бесконечности. По умолчанию *None*.

min_samples_split – минимальное количество наблюдений в листе, который будут делить дальше. От 1 до бесконечности. По умолчанию *2*.

min_samples_leaf – минимальное количество наблюдений в дочернем листе. От 1 до бесконечности. По умолчанию *1*.

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Решающие деревья. Обрезка