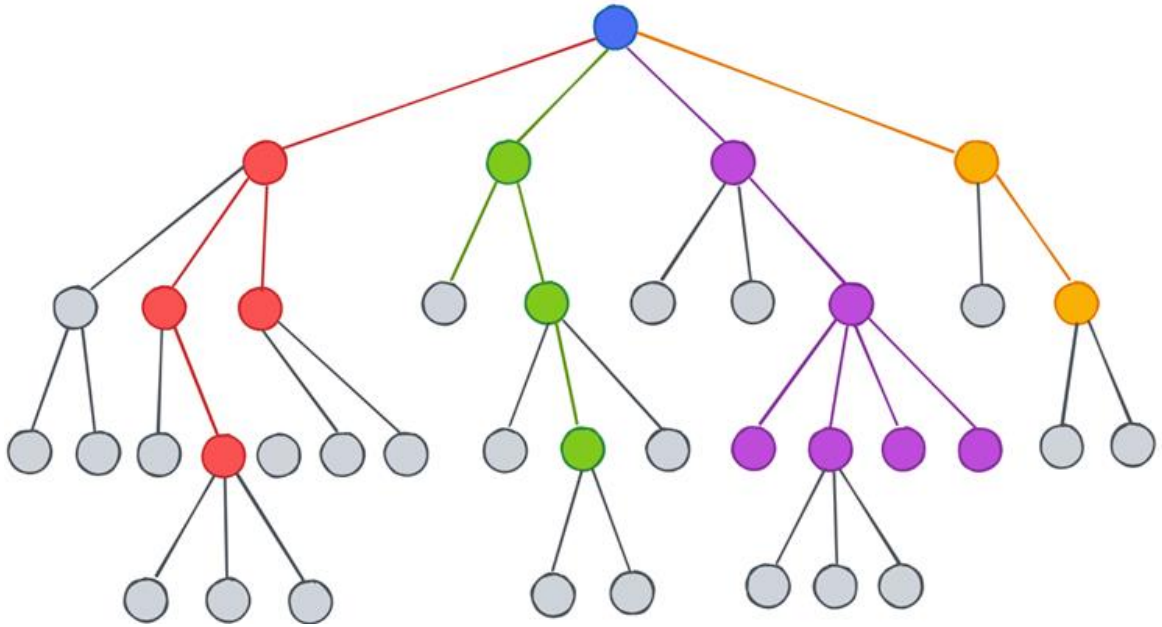
Решающие деревья



Решающие деревья

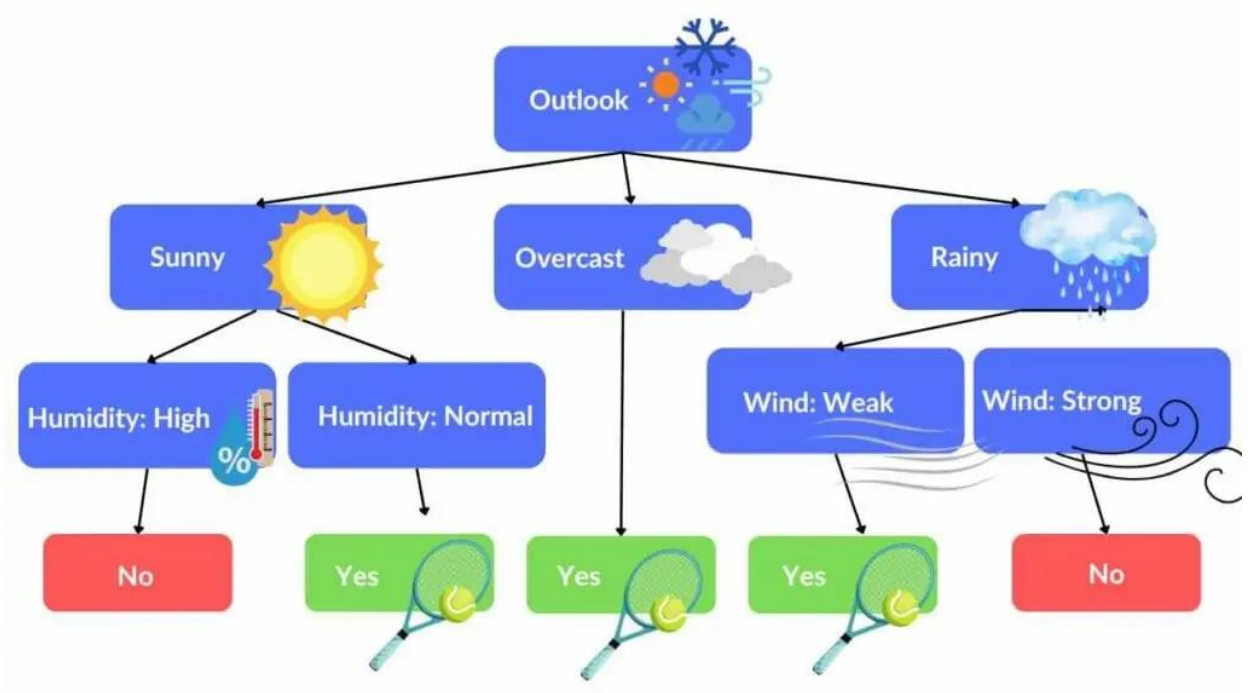
Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Логические правила

- Легко объяснить, как работают
- Находят нелинейные закономерности

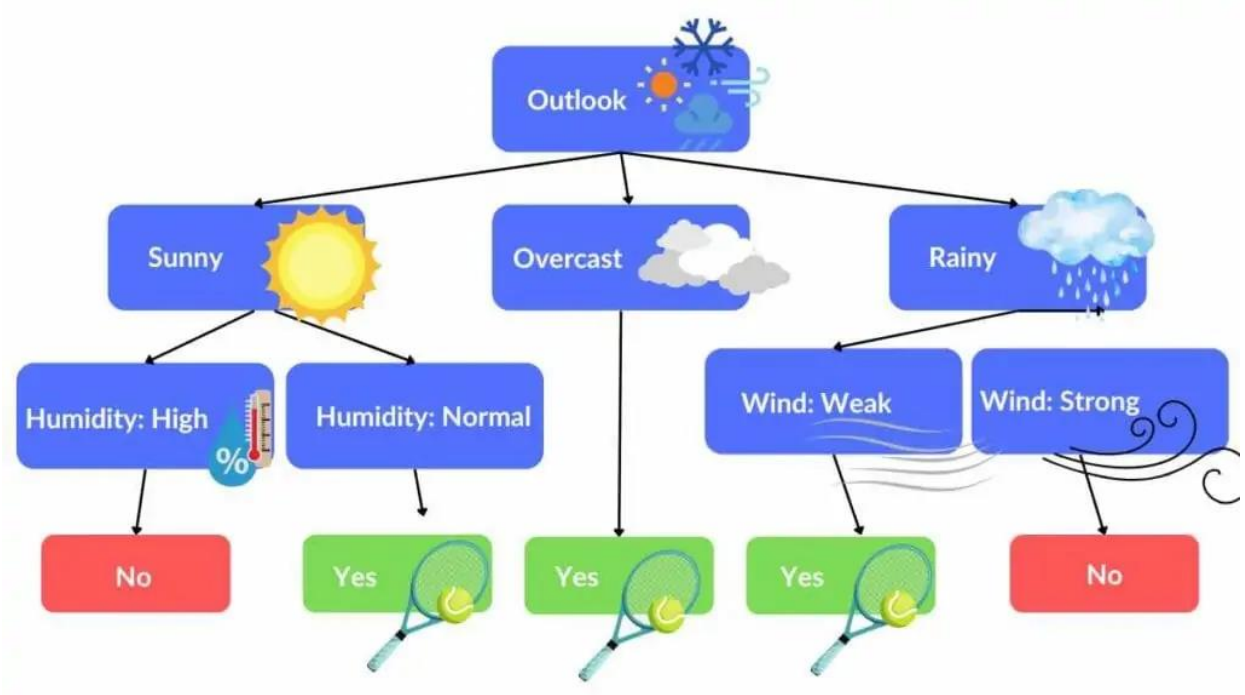
- Нужно как-то искать хорошие логические правила
- Нужно уметь составлять модели из логических правил

Решающие деревья

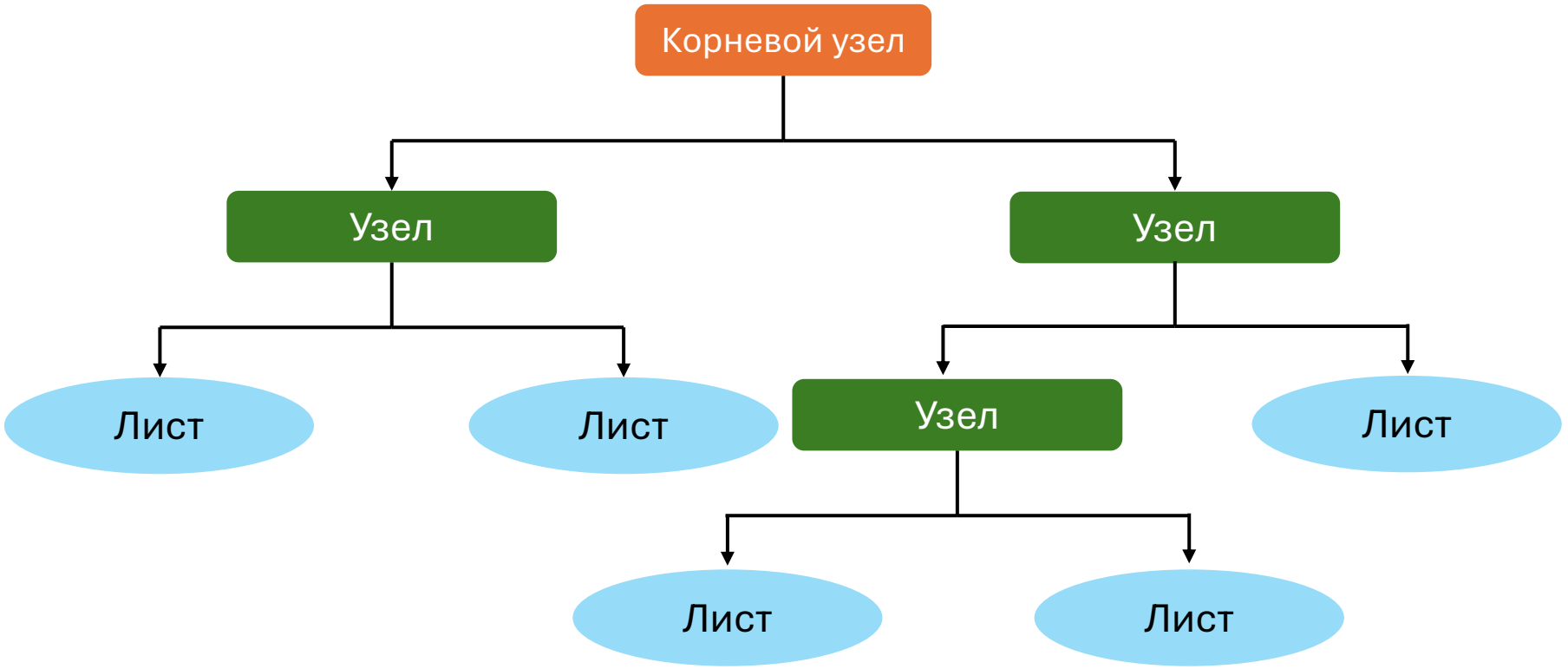


Решающие деревья

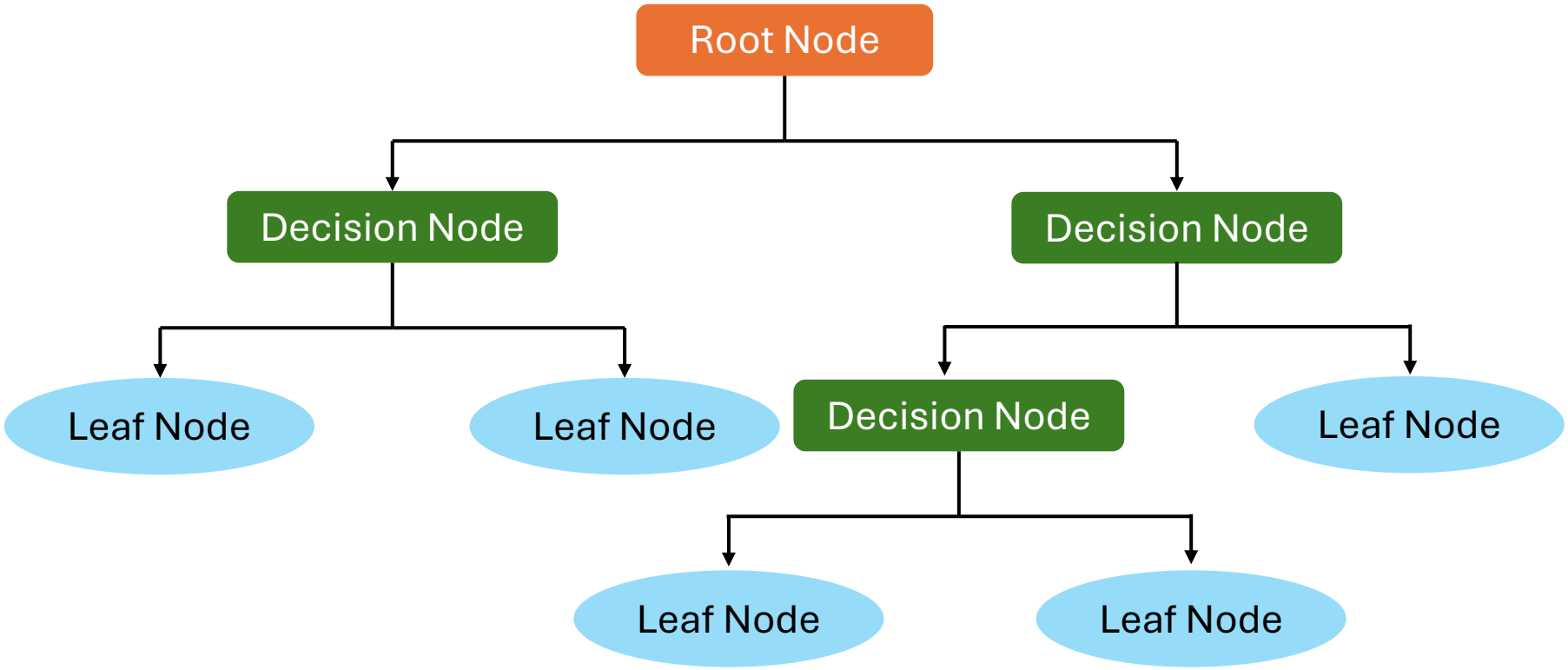
Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No



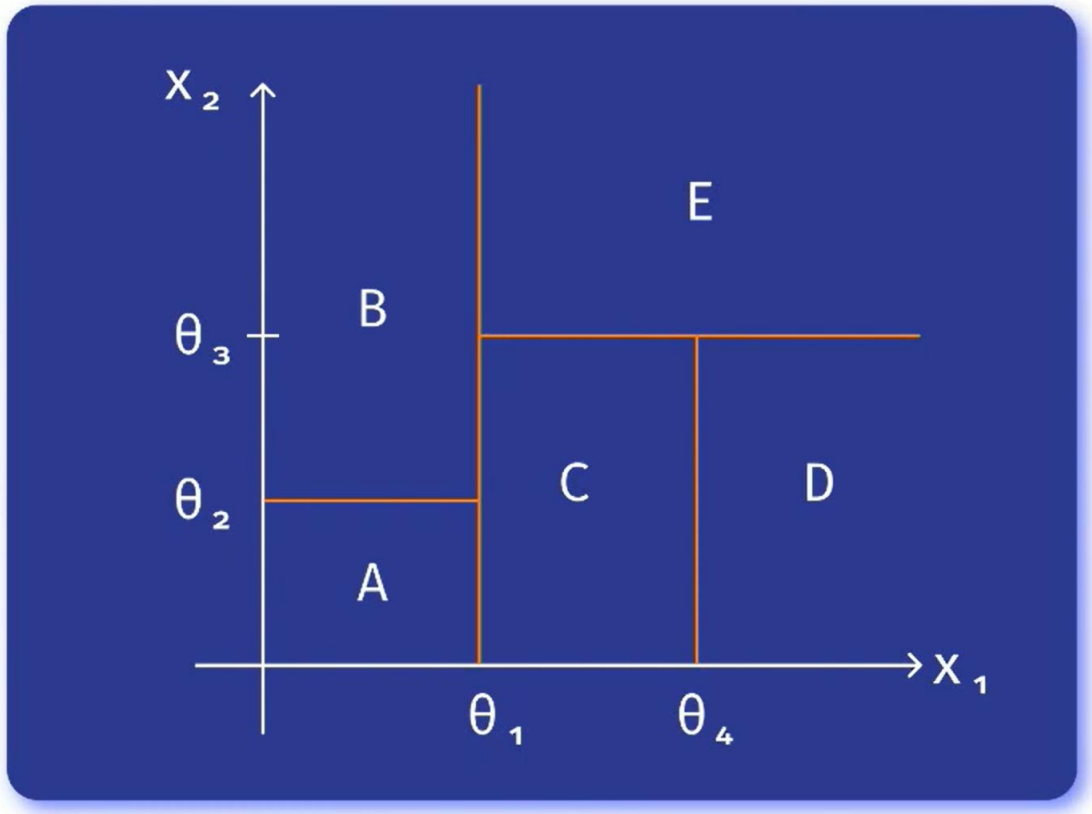
Решающие деревья

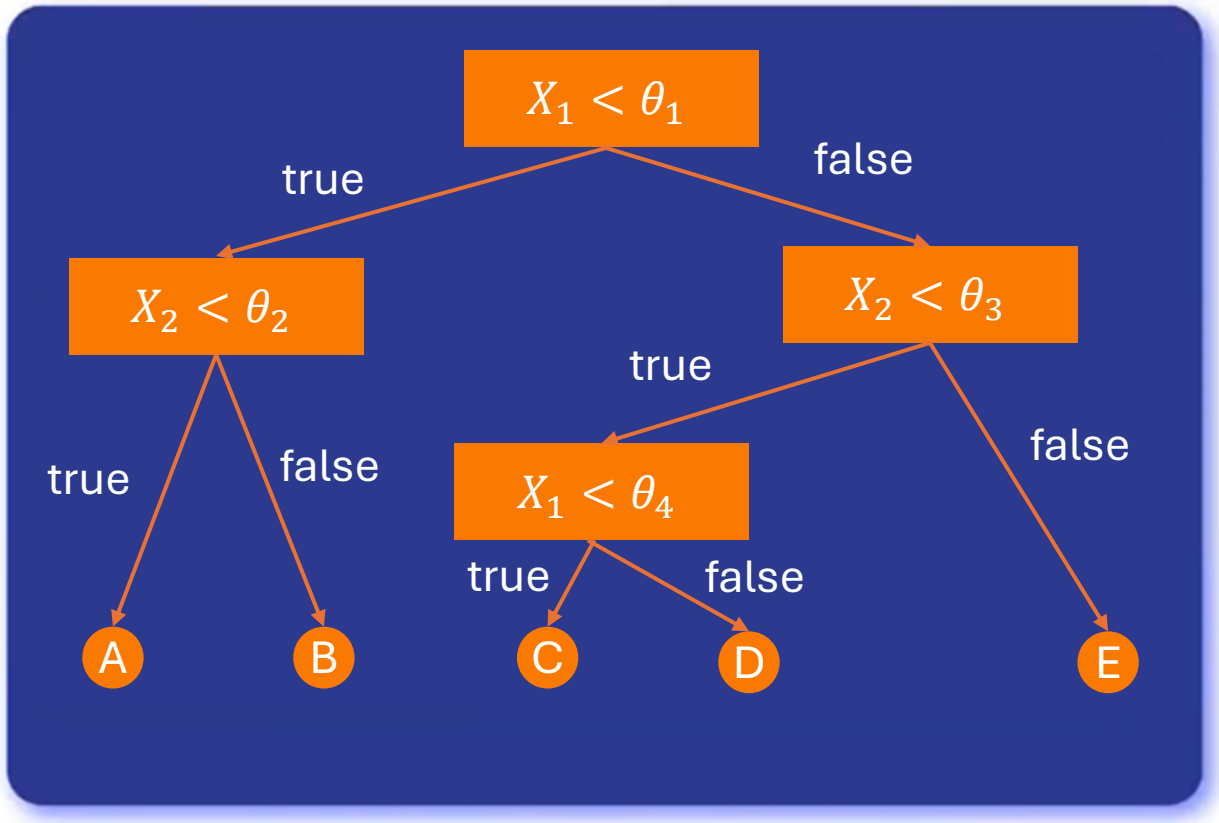


Решающие деревья

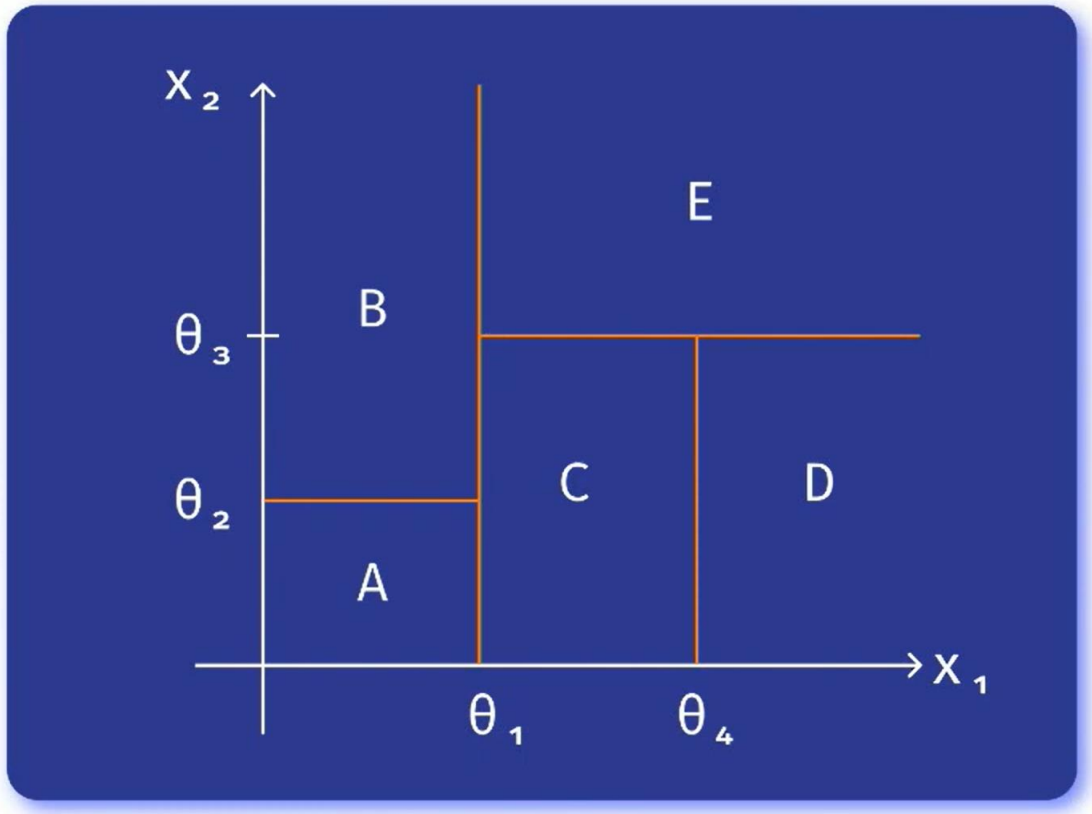


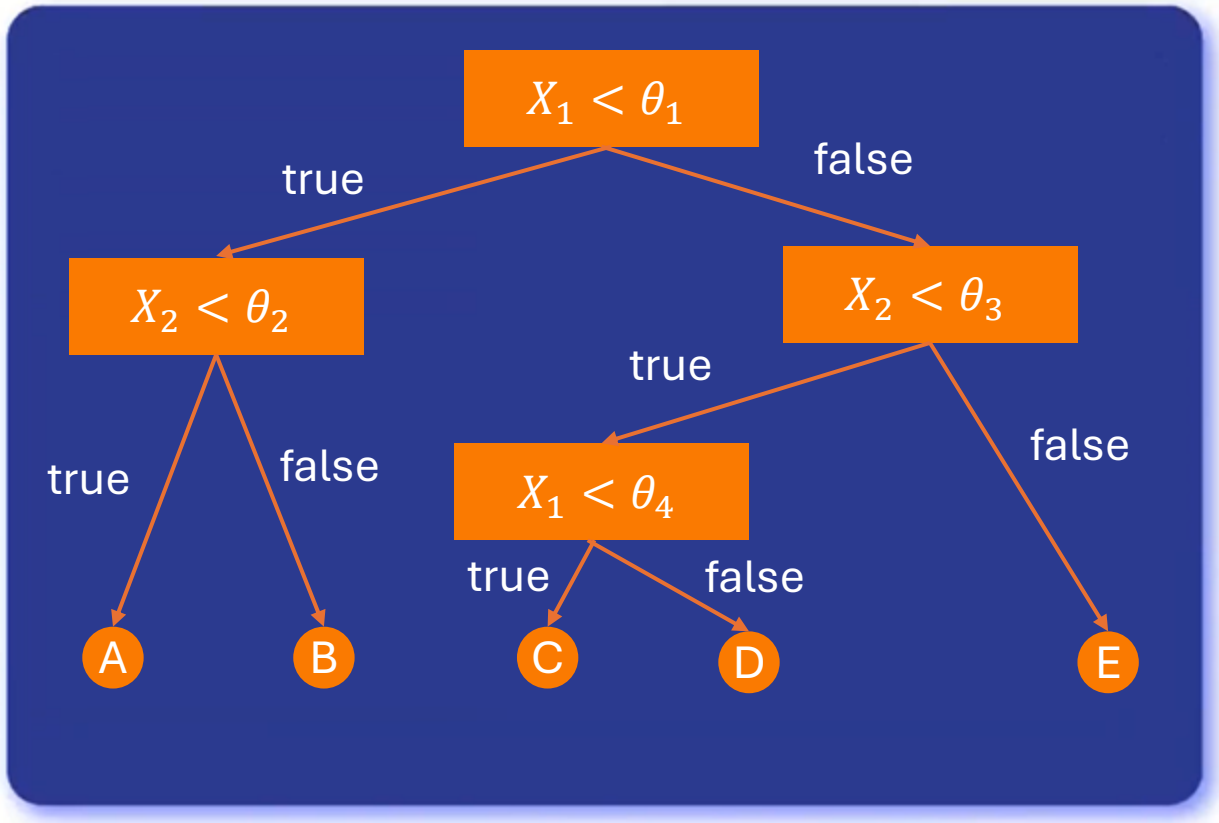
Геометрическая интерпретация





Геометрическая интерпретация

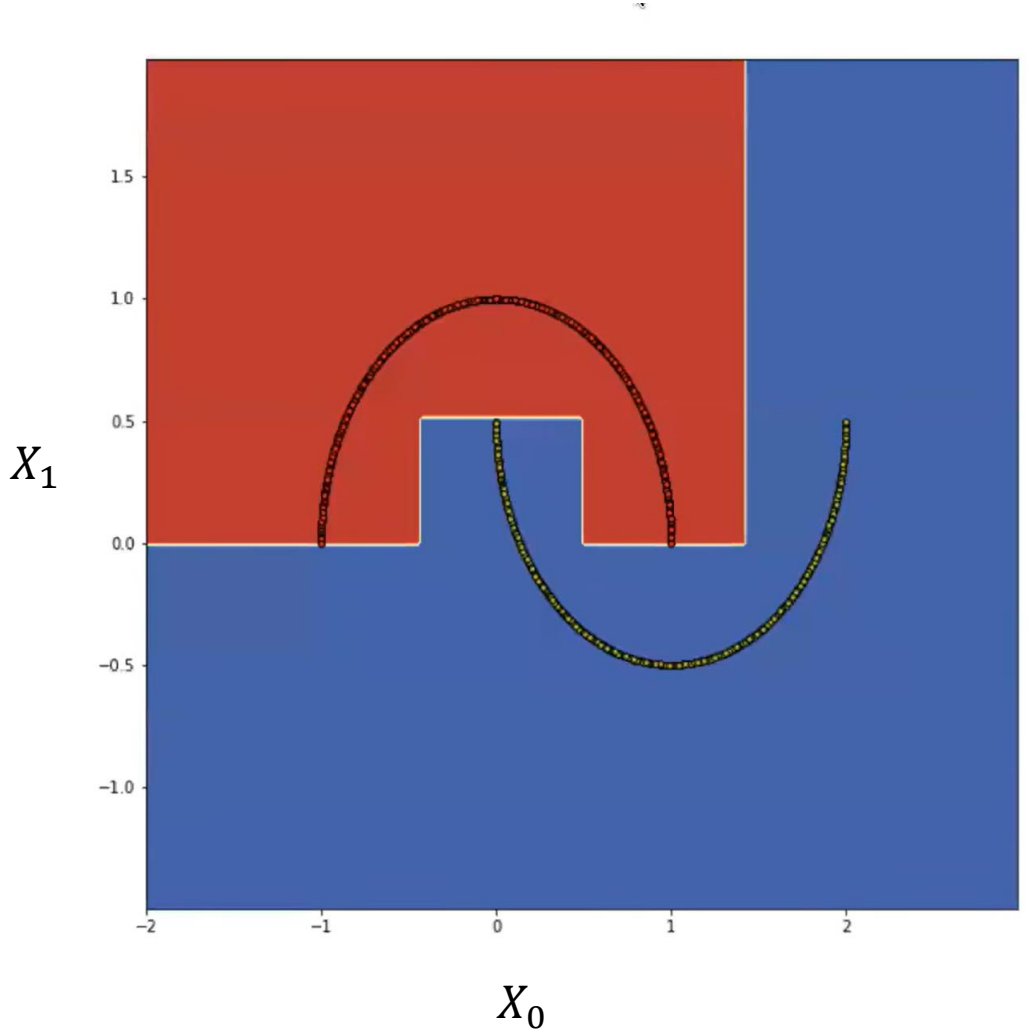


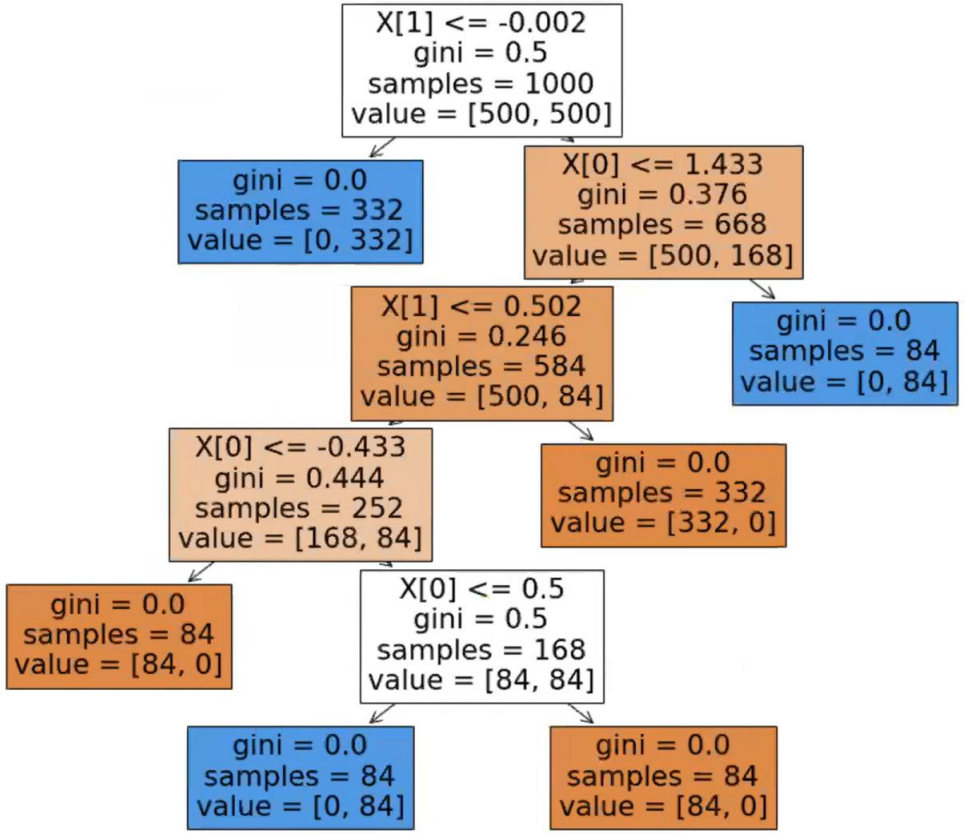


Внутренние вершины: предикаты $[X_j < t]$

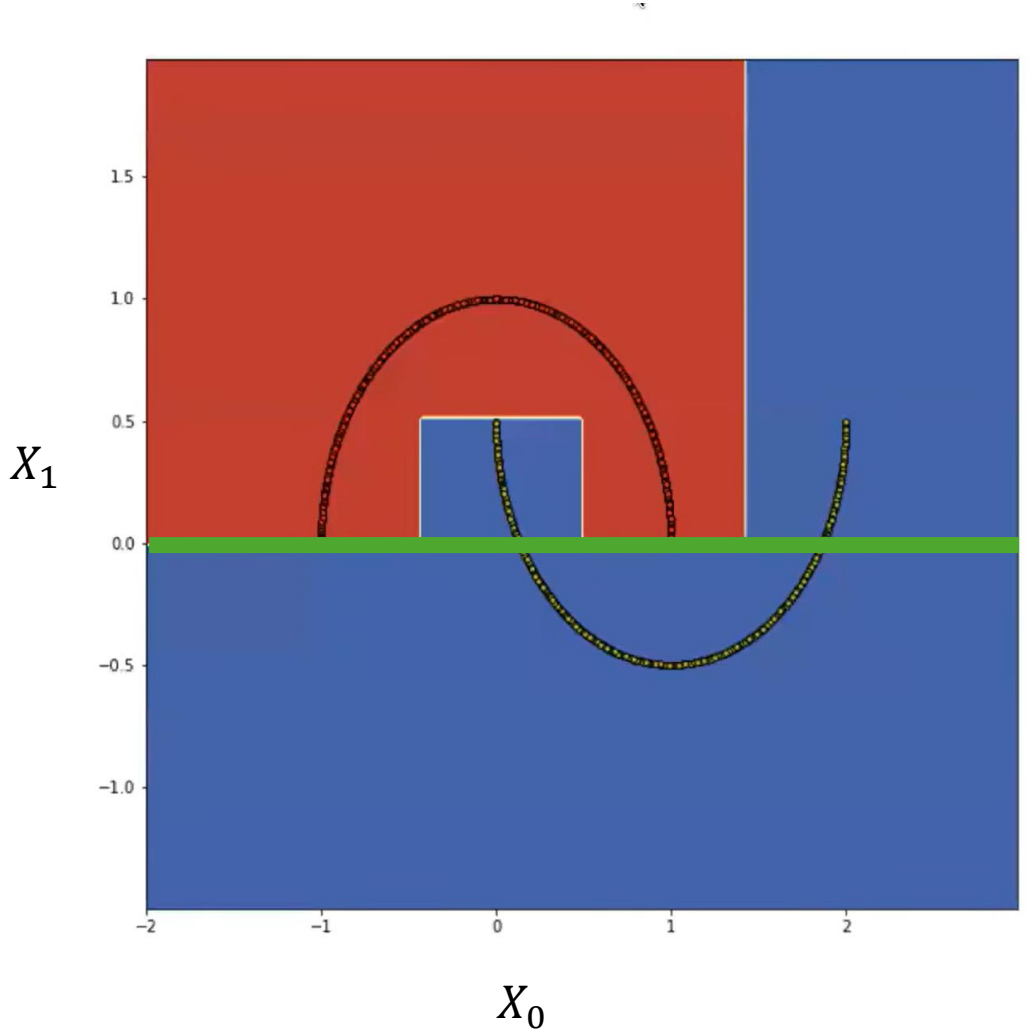
Листья: прогнозы $c \in \mathbb{Y}$

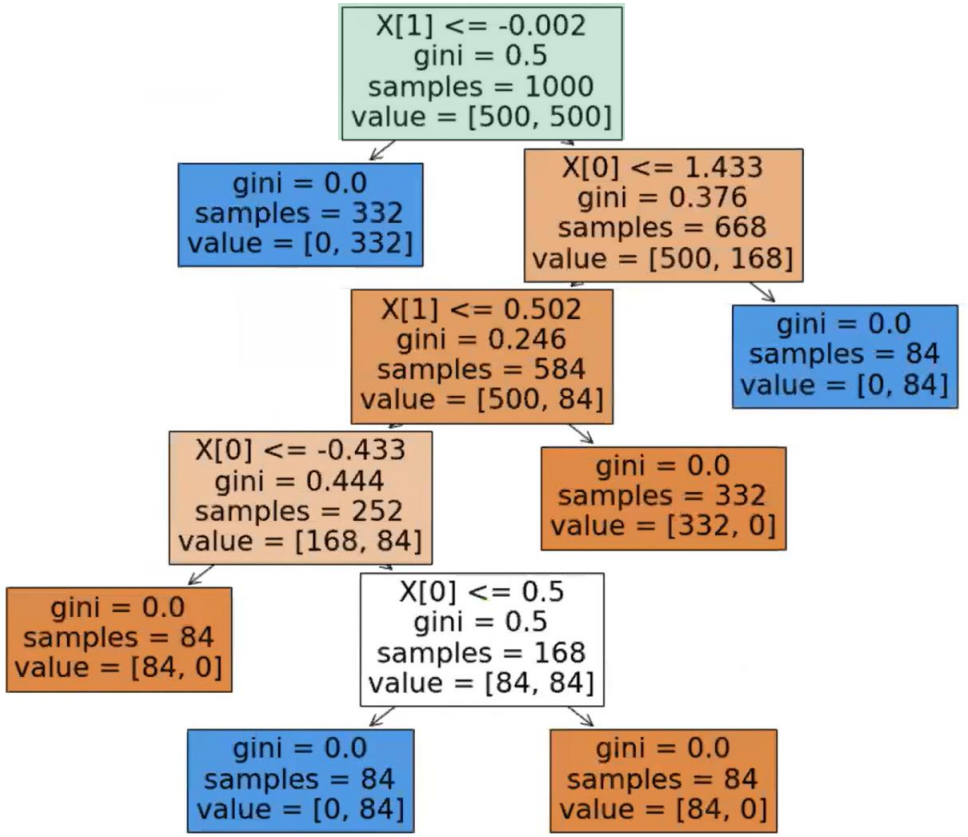
Решающие деревья



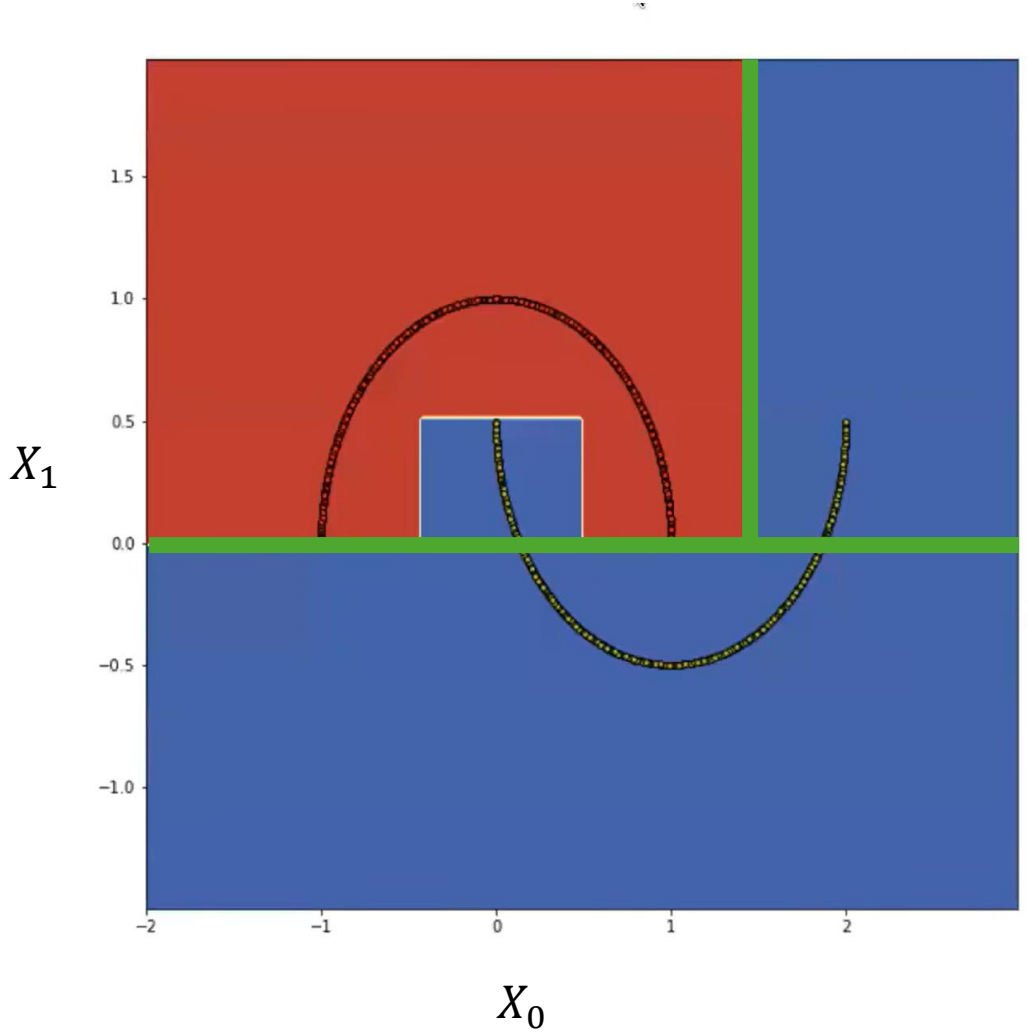


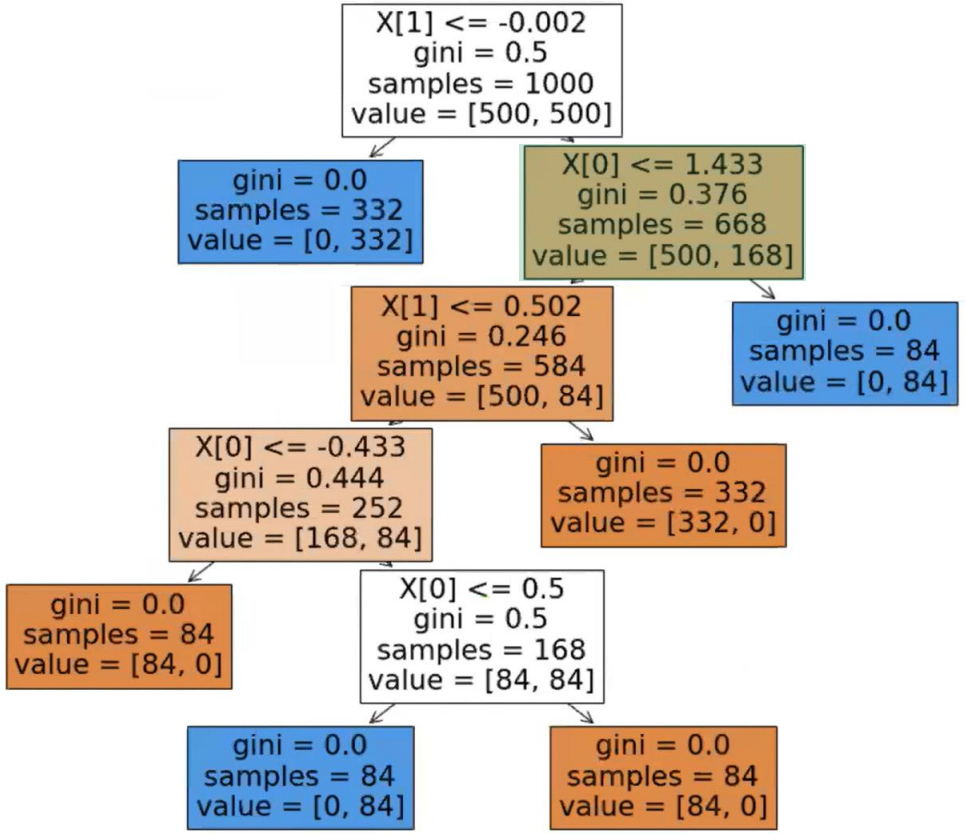
Решающие деревья



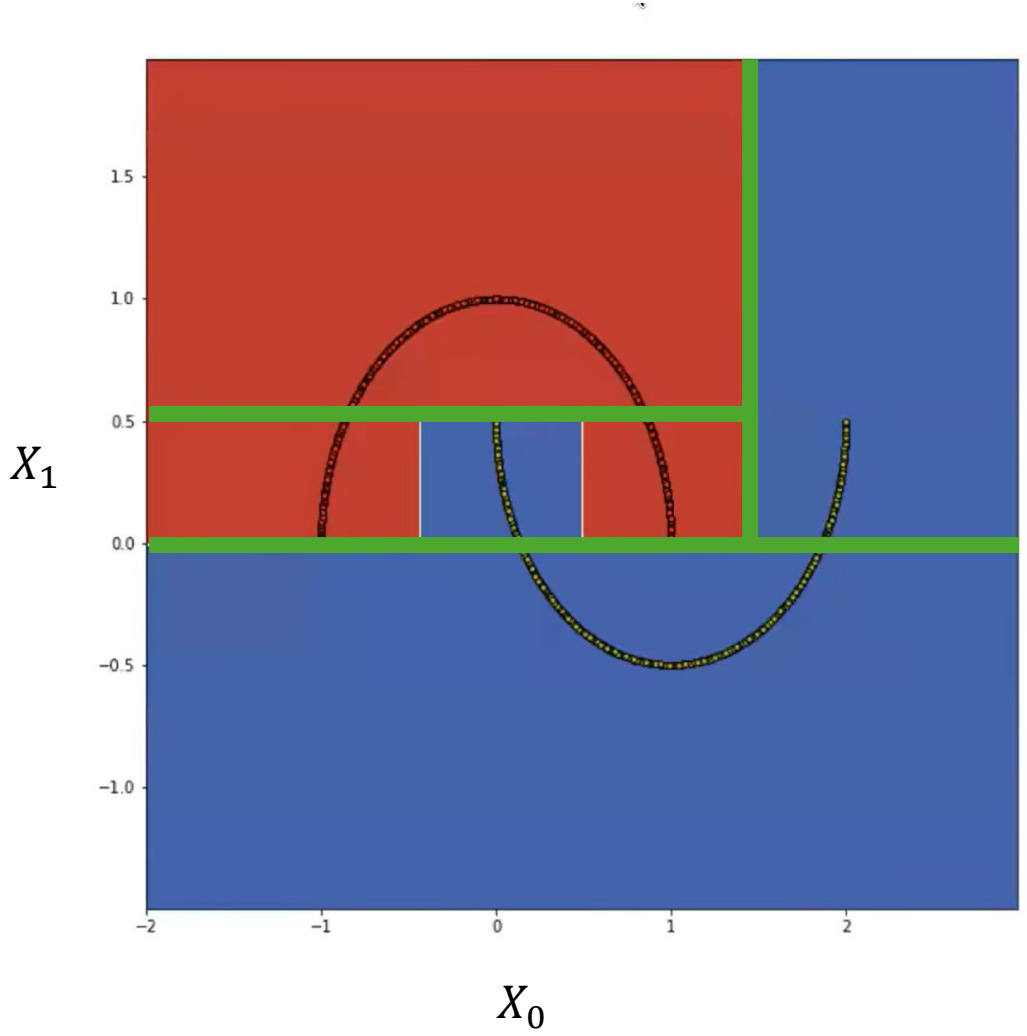


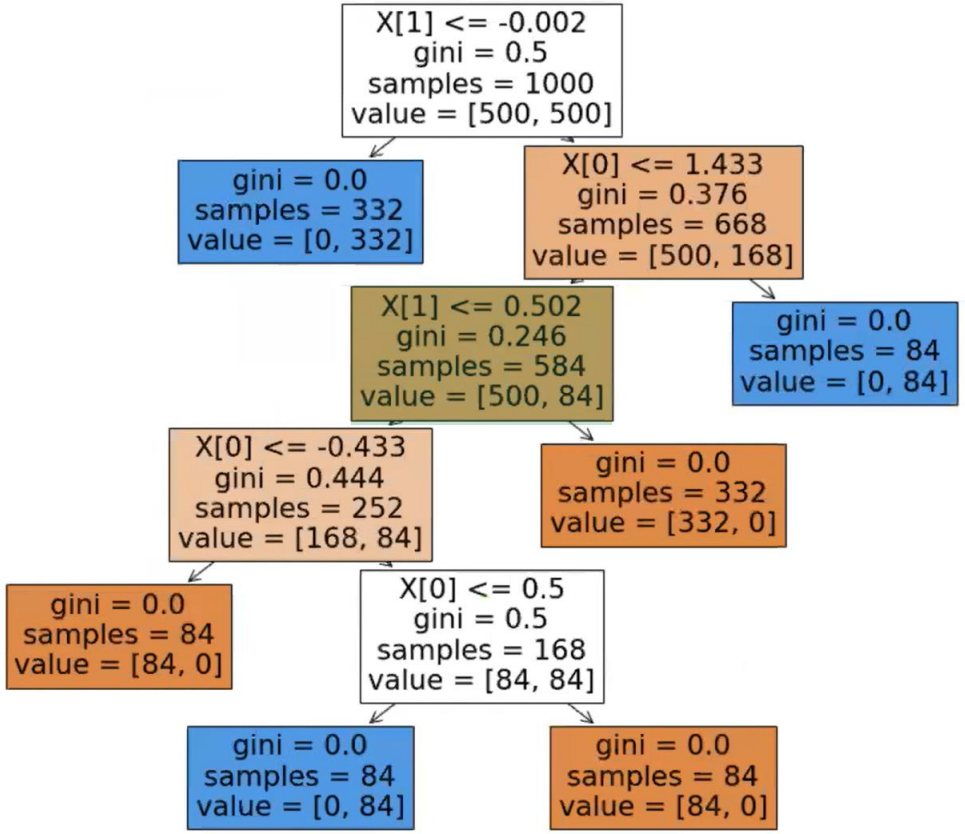
Решающие деревья



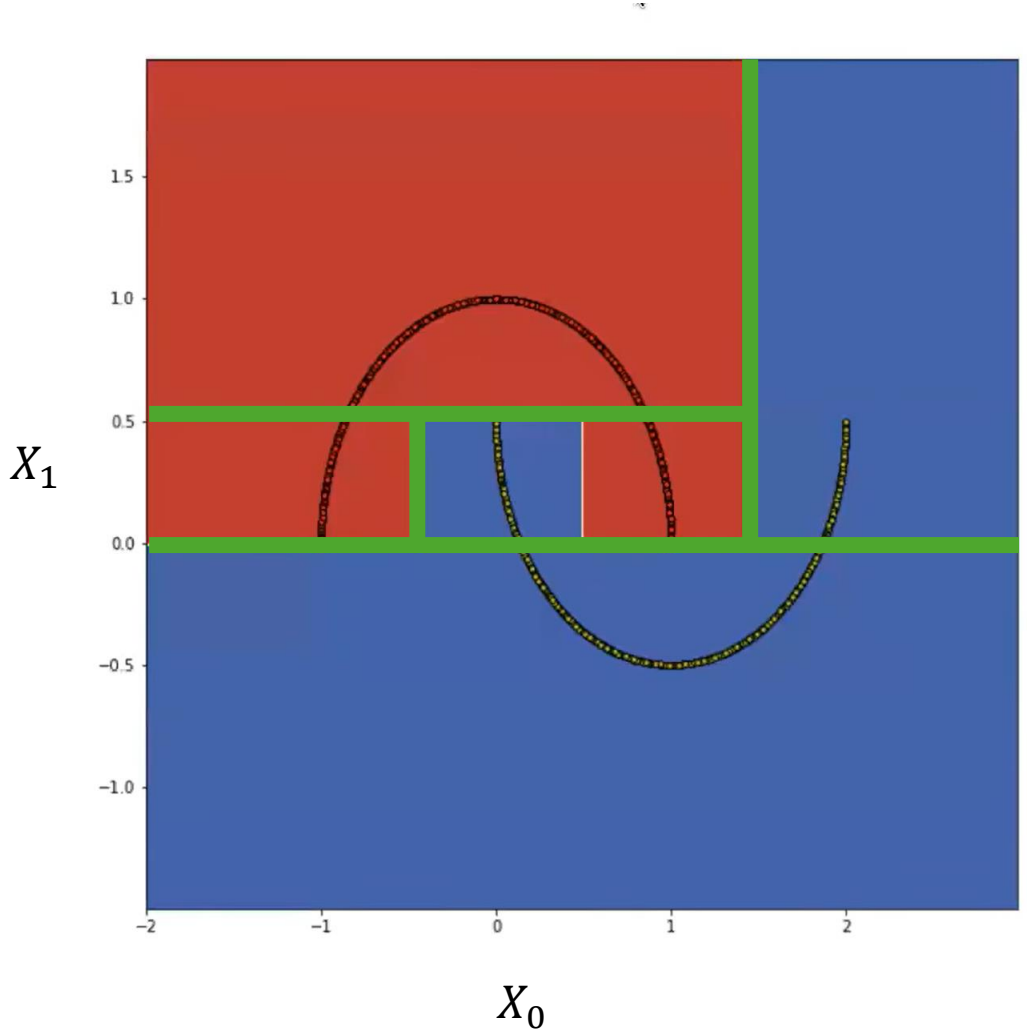


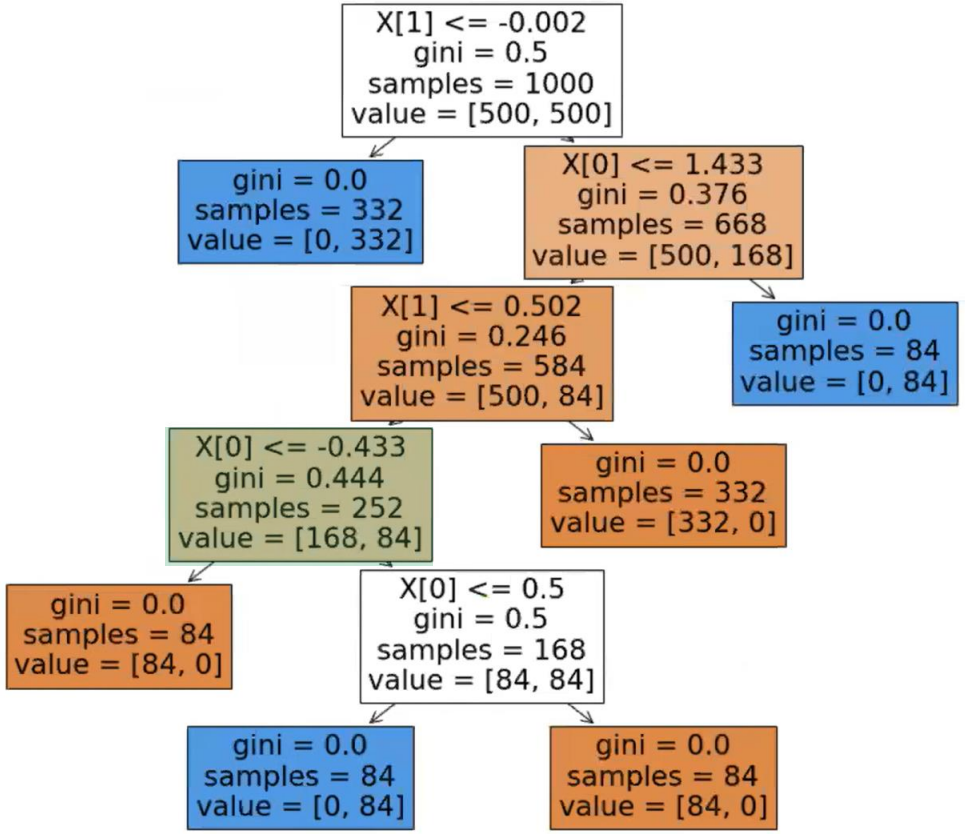
Решающие деревья



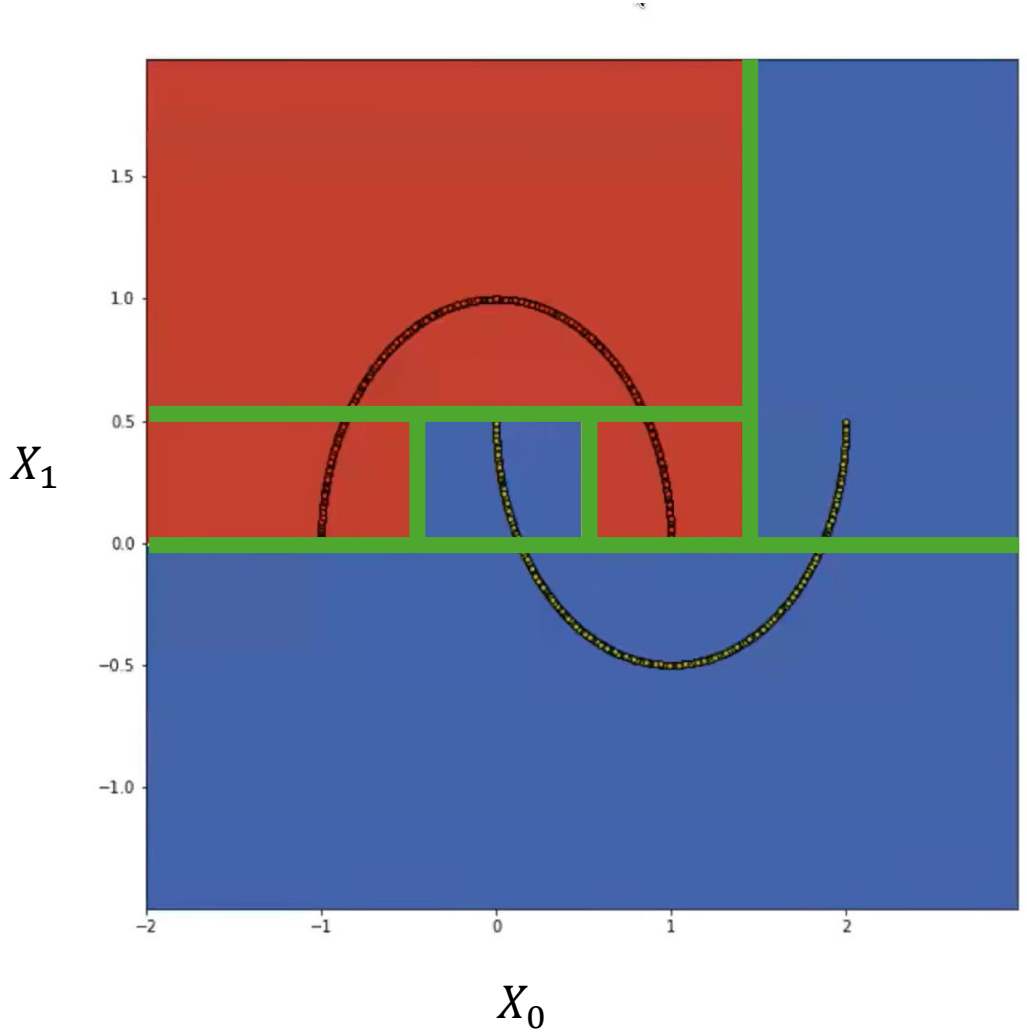


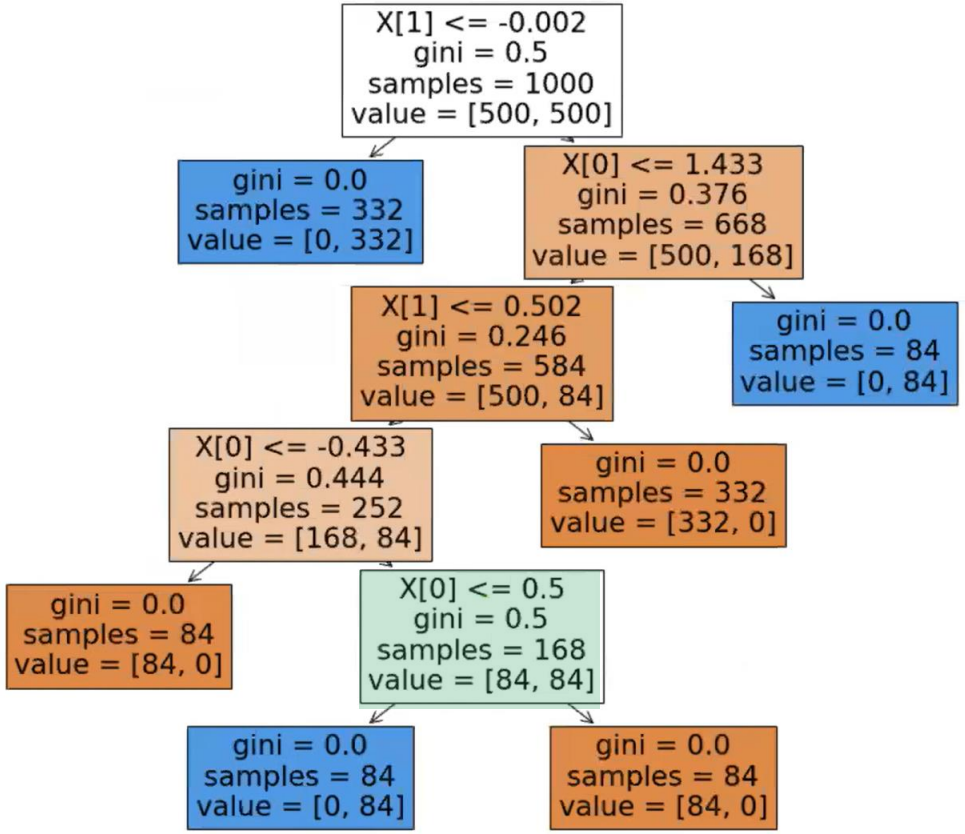
Решающие деревья



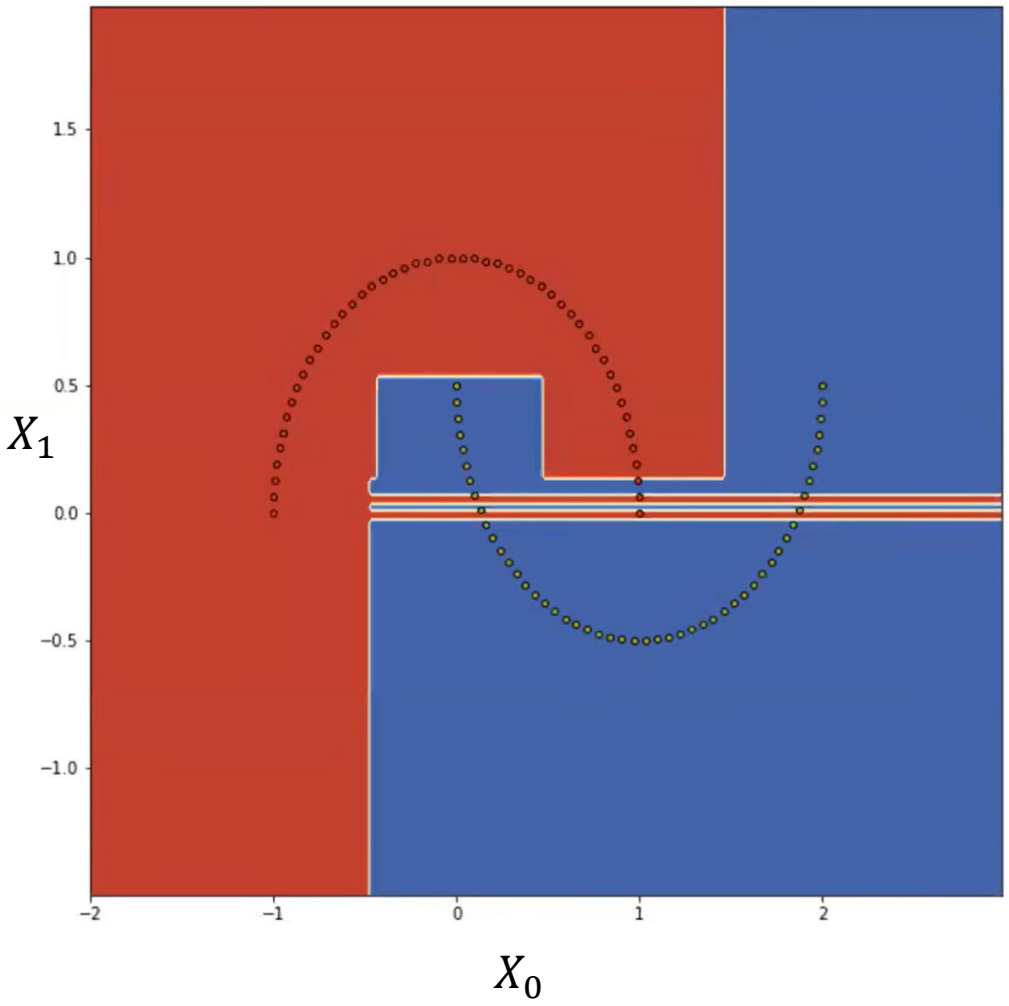


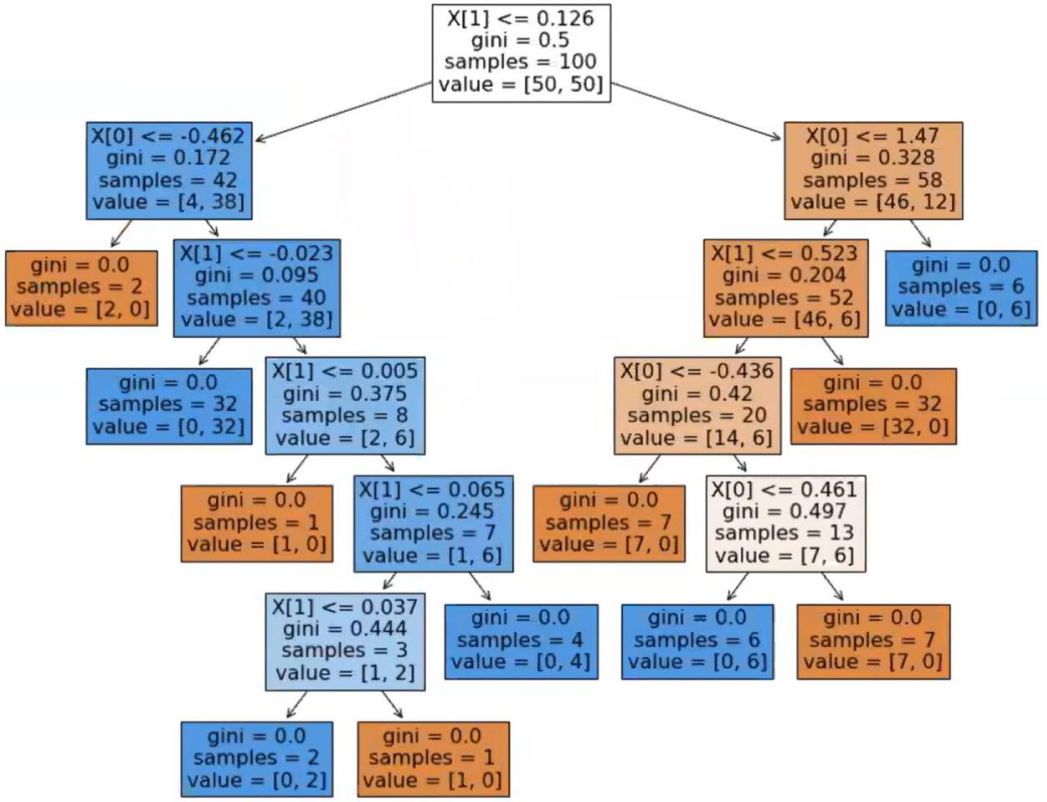
Решающие деревья



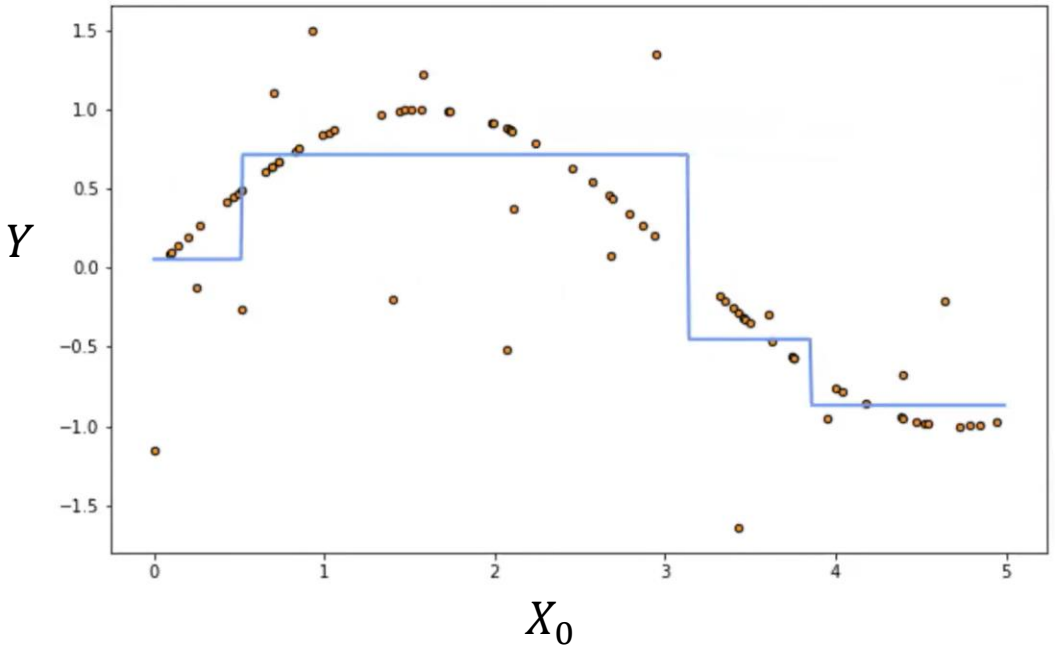


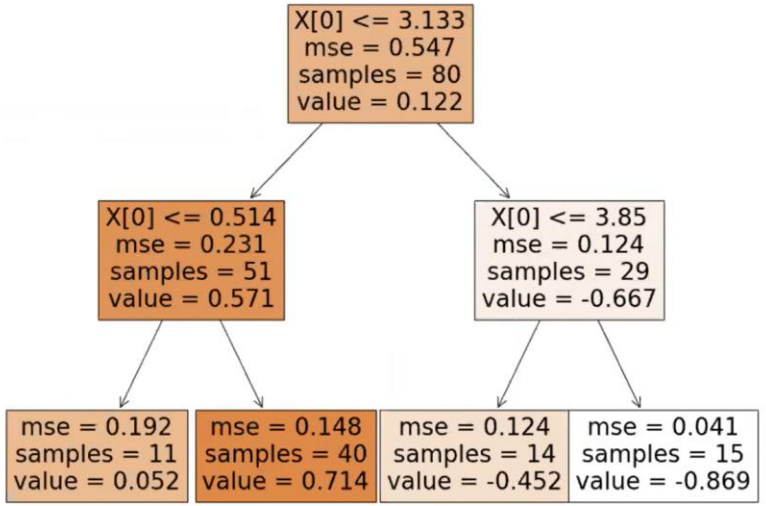
Решающие деревья



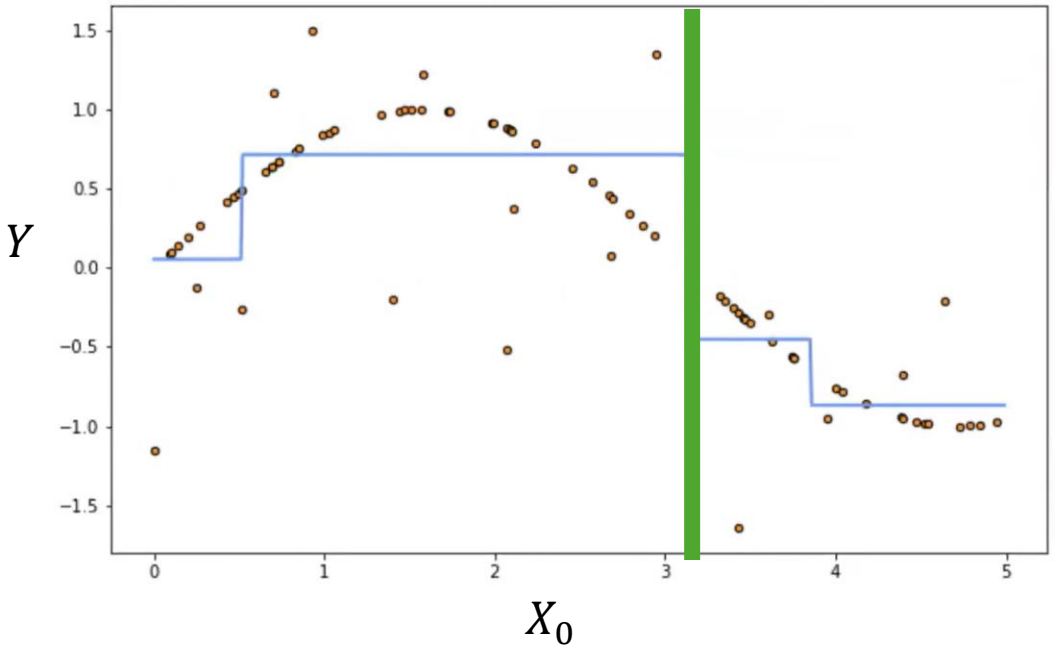


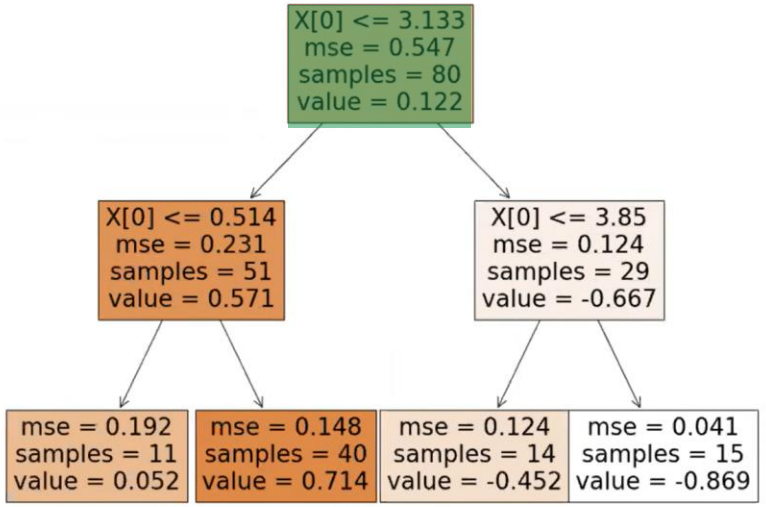
Решающее дерево для регрессии



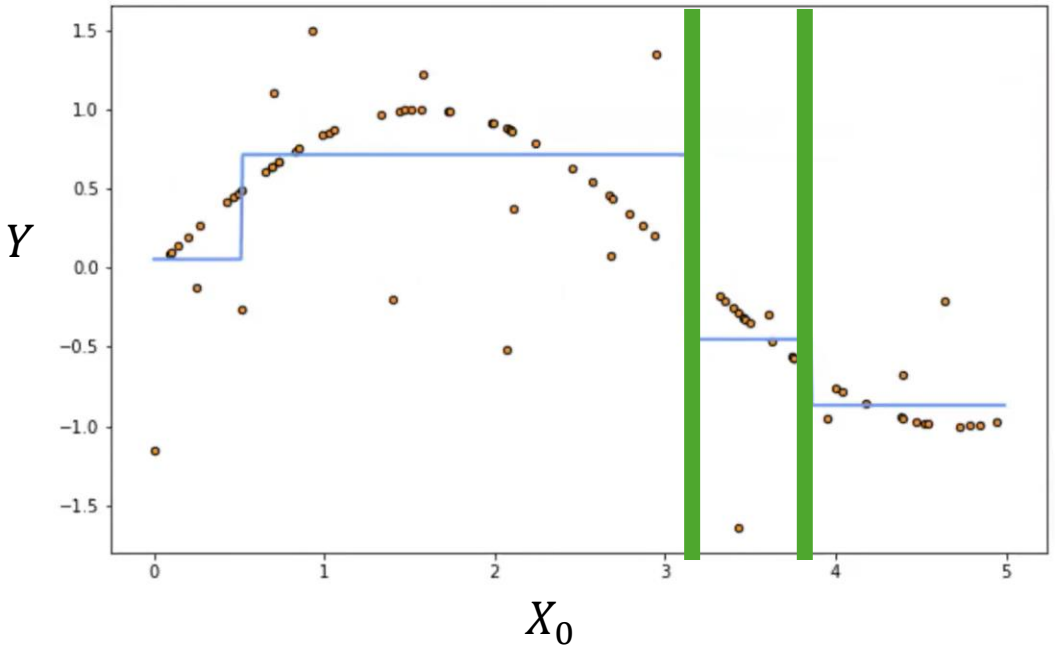


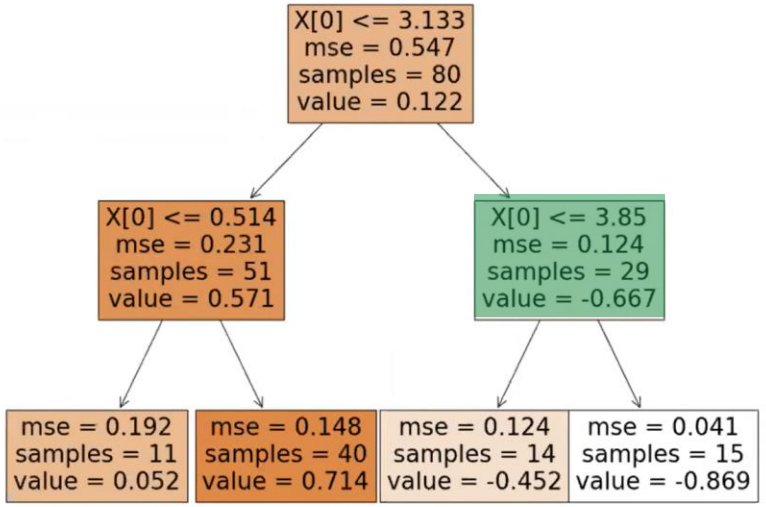
Решающее дерево для регрессии



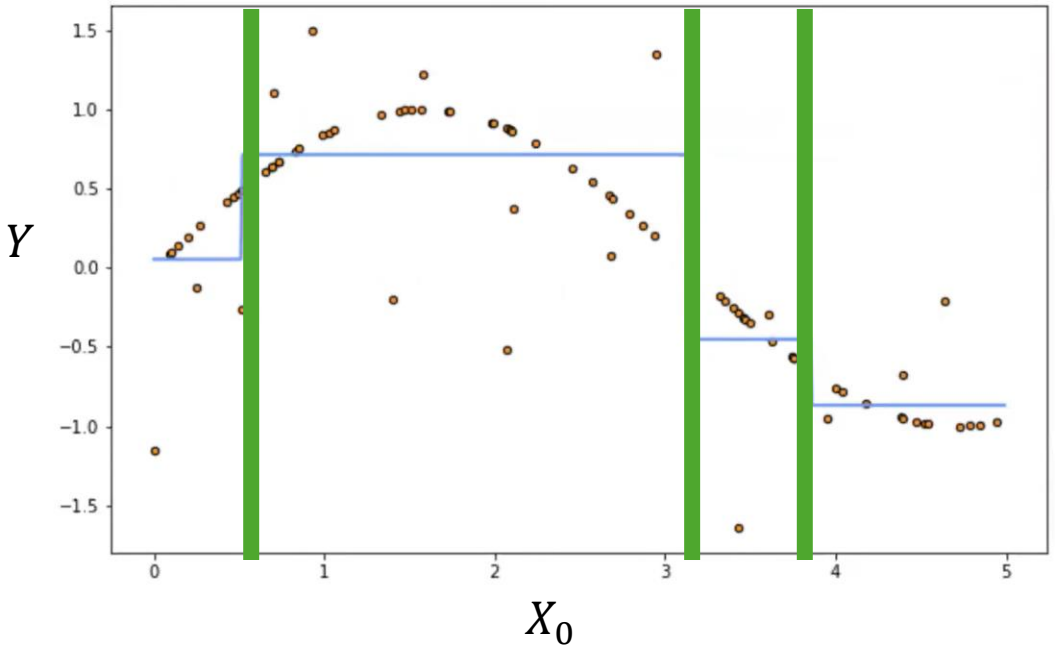


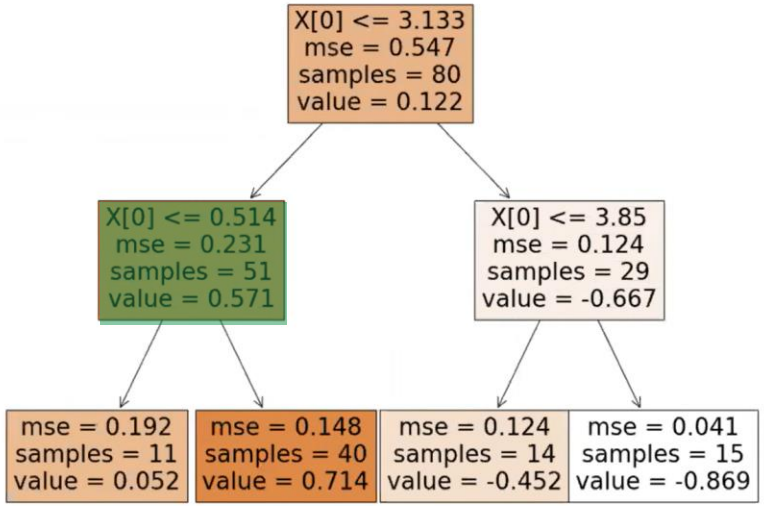
Решающее дерево для регрессии



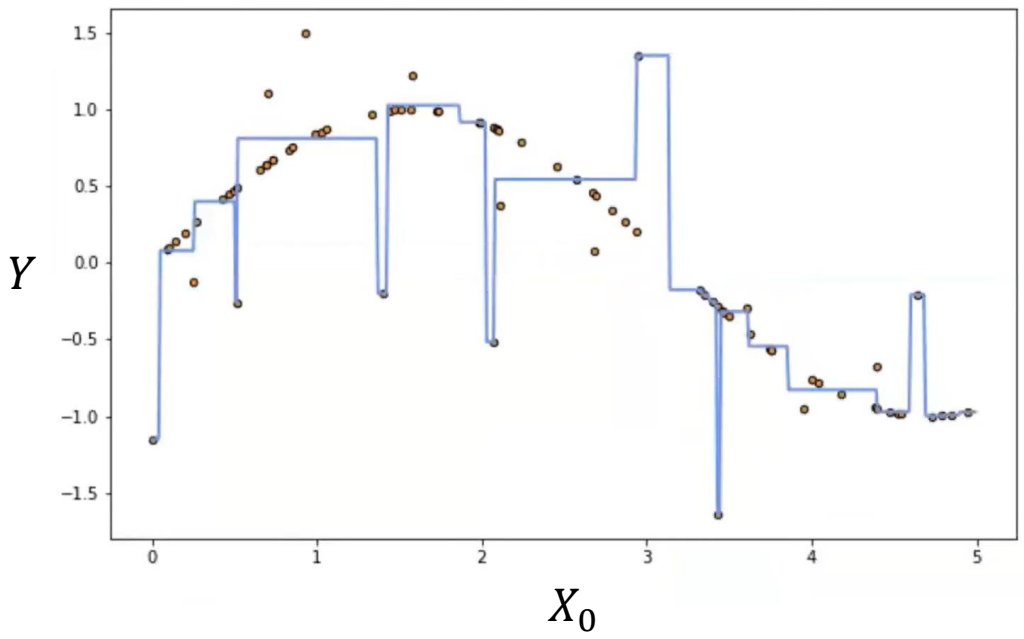


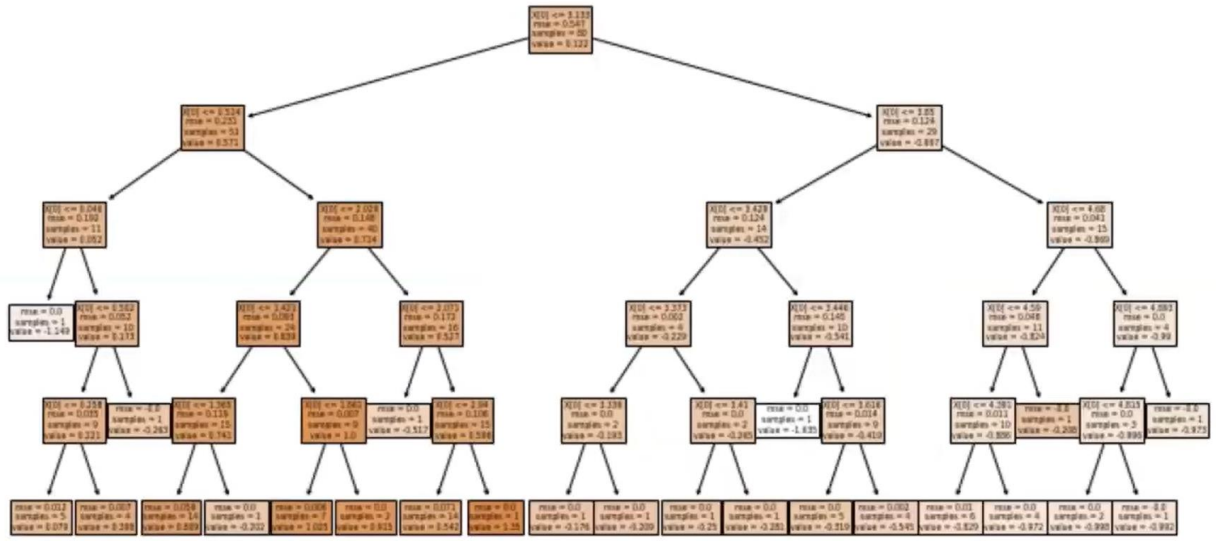
Решающее дерево для регрессии





Решающее дерево для регрессии

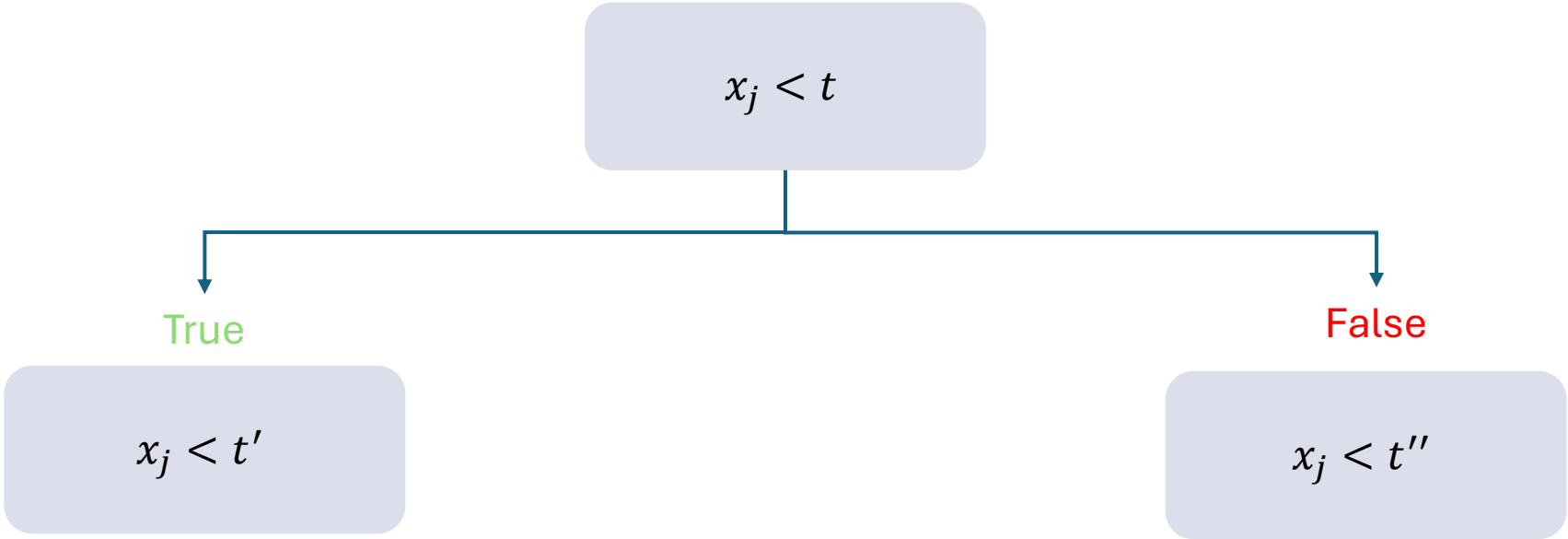




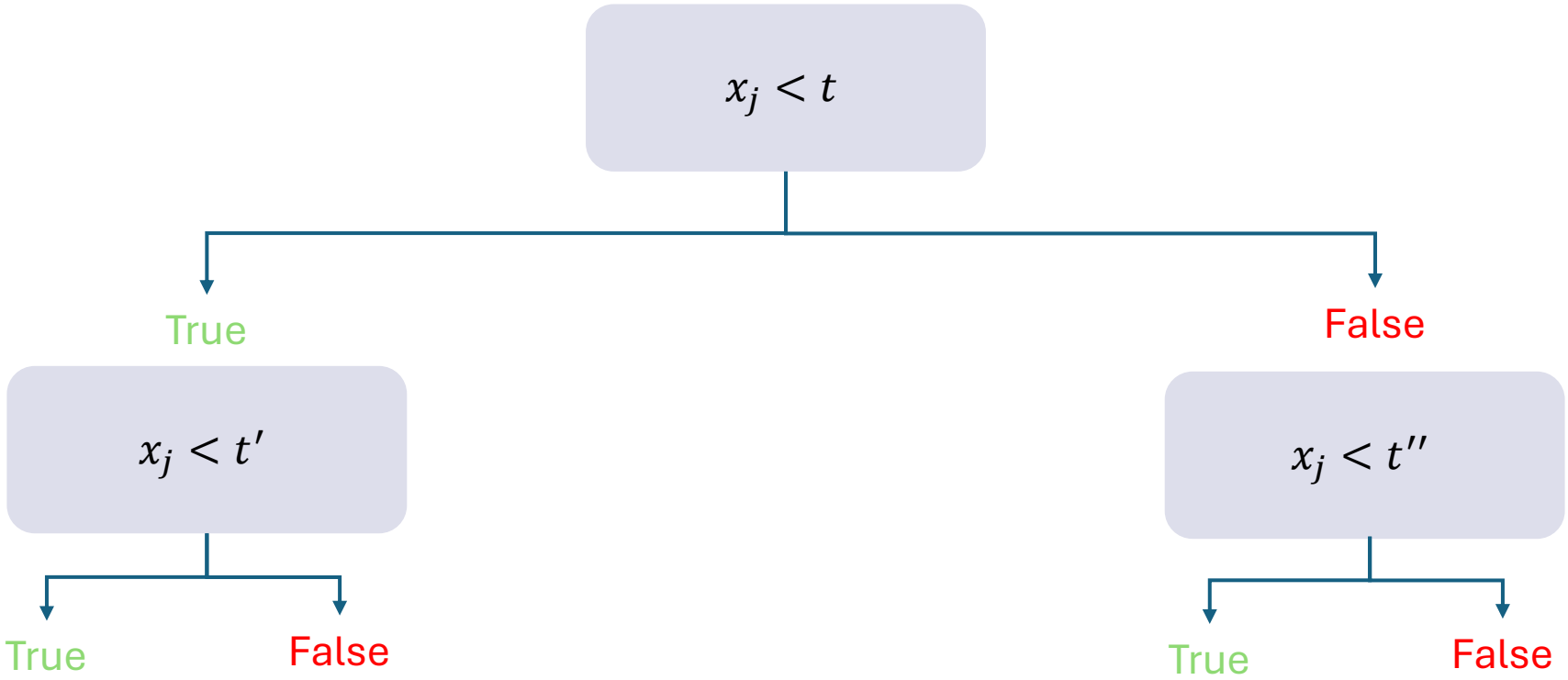
Предикаты

$$x_j < t$$

Предикаты



Предикаты



Прогнозы в листьях

Наш выбор: константные прогнозы $c_v \in \mathbb{Y}$

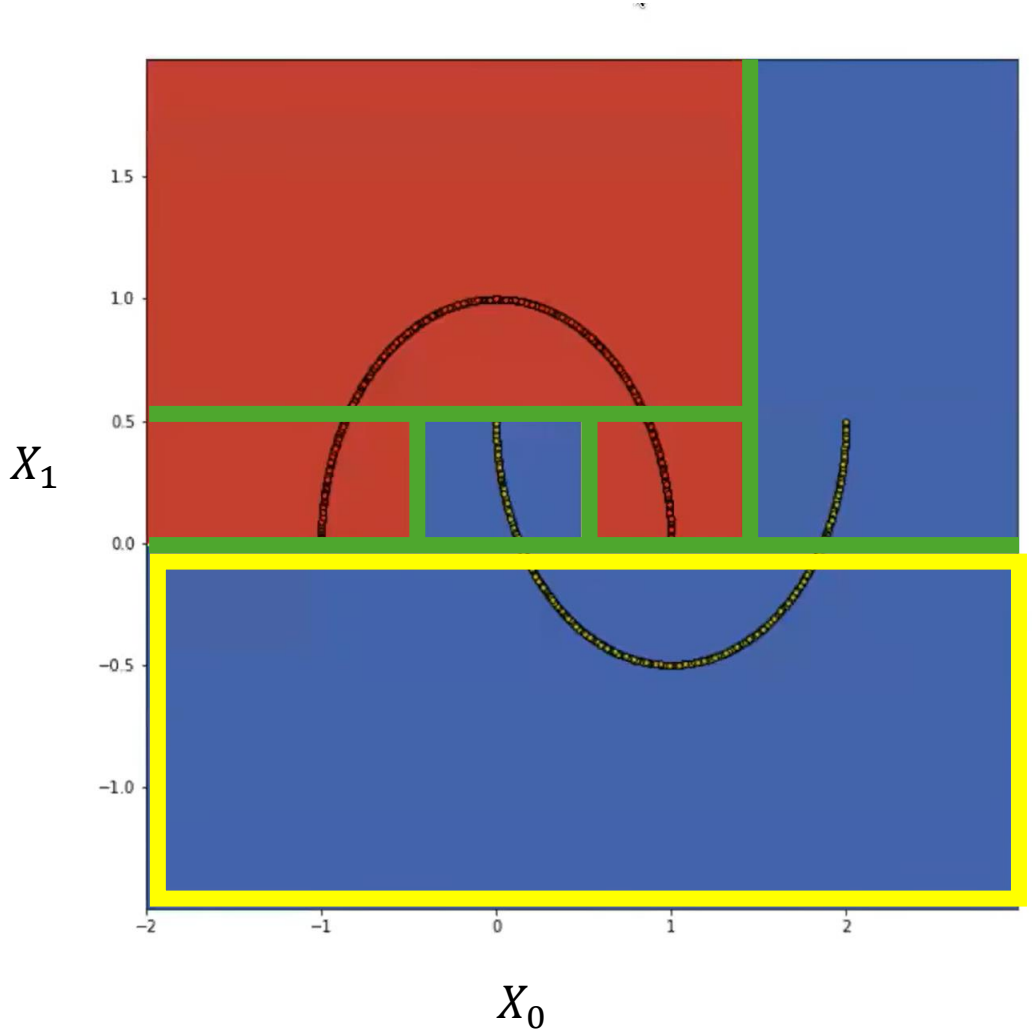
Регрессия:

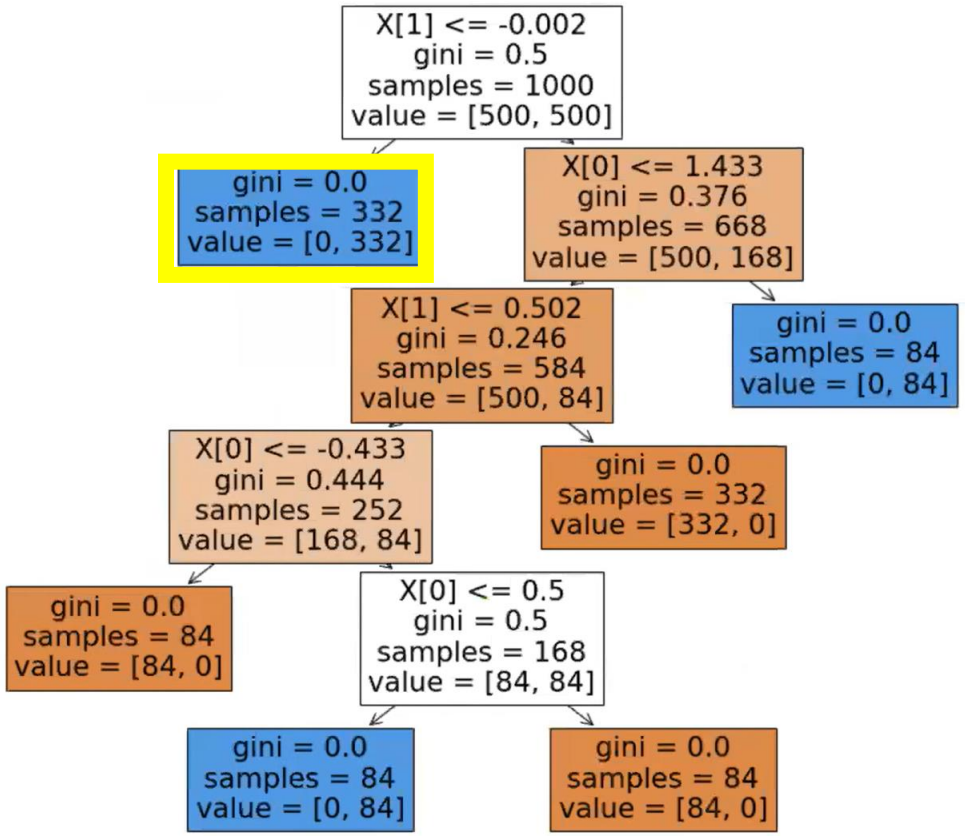
$$c_v = \frac{1}{|R_v|} \sum_{(x_i, y_i) \in R_v} y_1$$

Классификация:

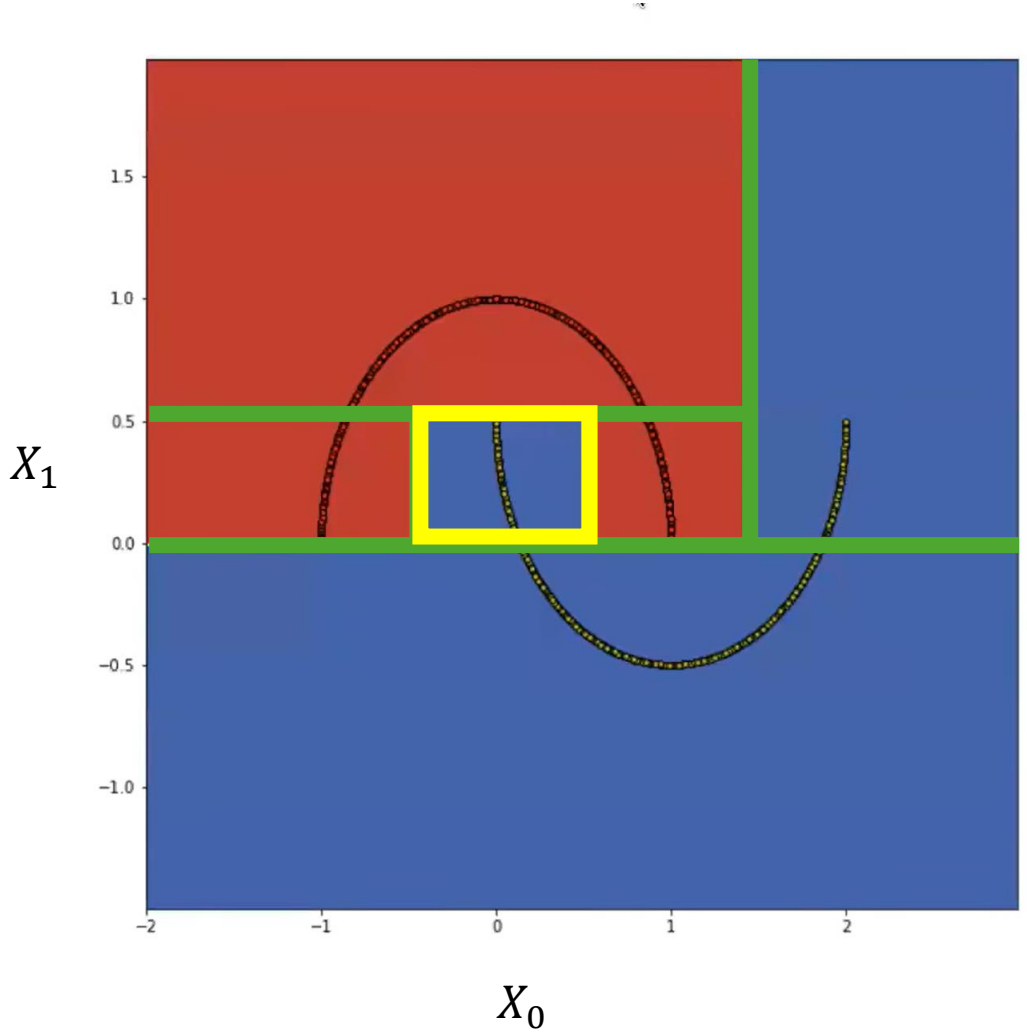
$$c_v = \underset{k \in \mathbb{Y}}{\operatorname{argmax}} \sum_{(x_i, y_i) \in R_v} [y_i = k]$$

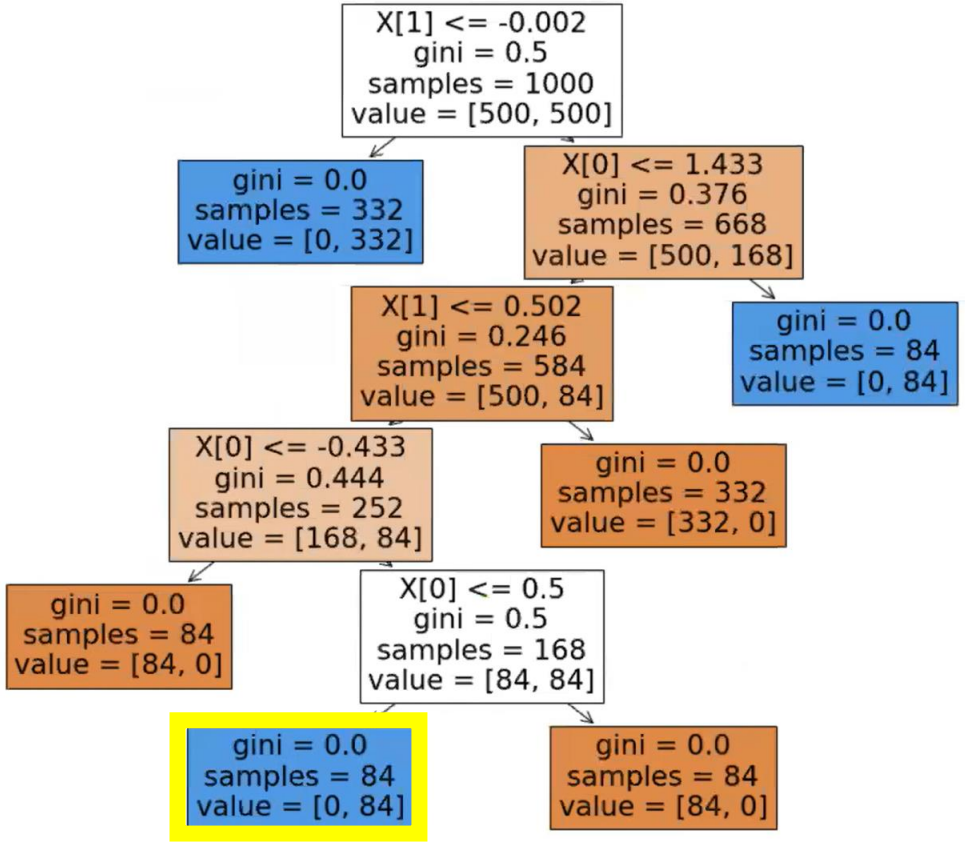
Прогнозы в листьях





Прогнозы в листьях





Формула для дерева

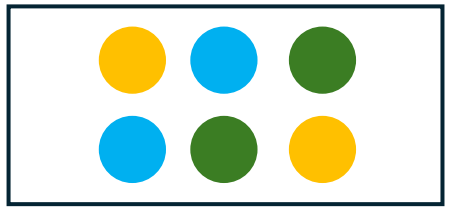
Дерево разбивает признаковое пространство на области $R_1, \dots, R_J$

Каждая область R_j соотвествует листу

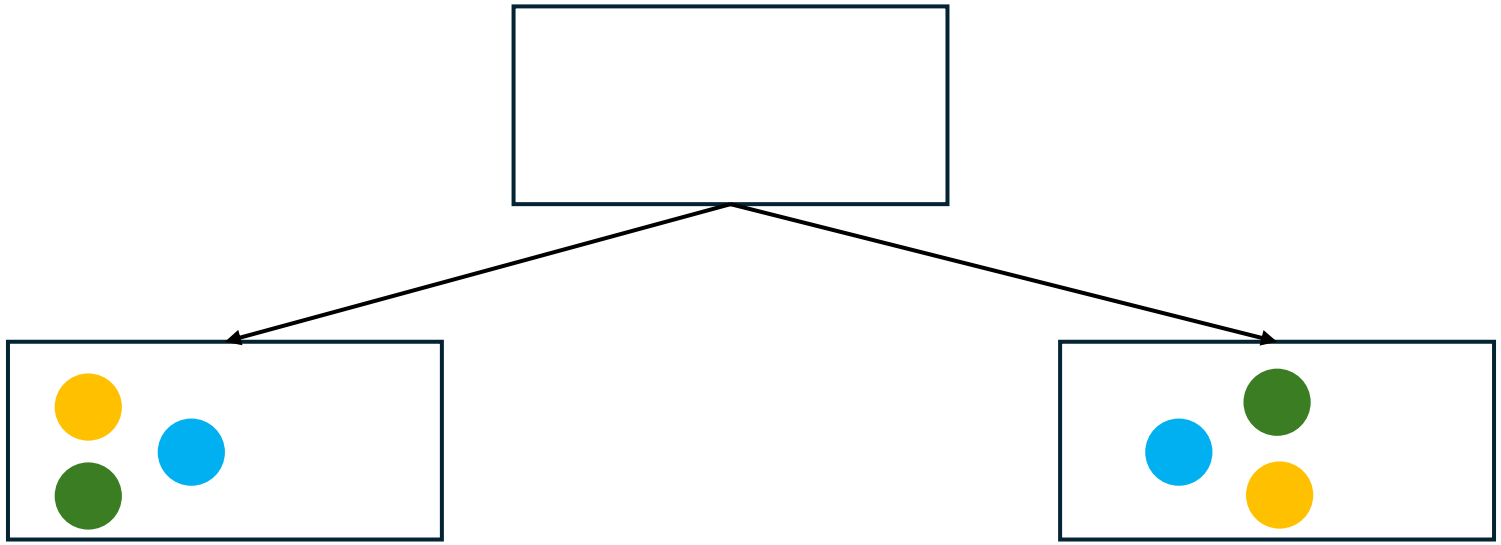
В области R_j прогноз c_j константный

$$a(x) = \sum_{j=1}^{J} c_j [x \in R_j]$$

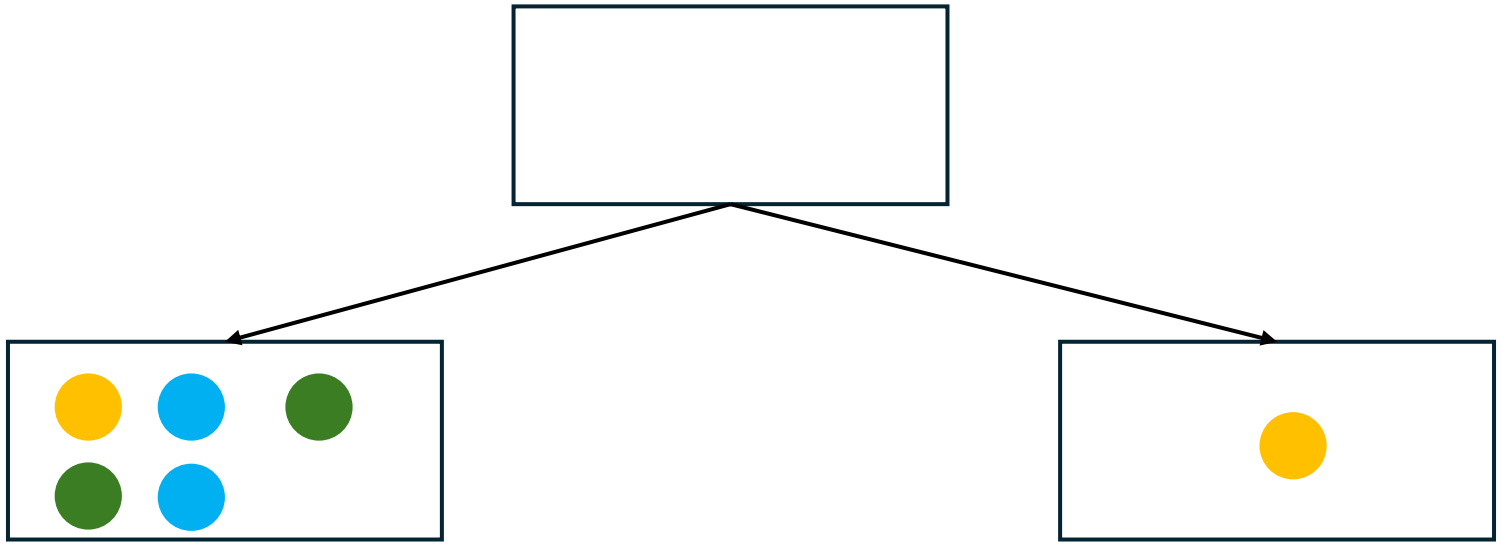
Как разбить вершину?



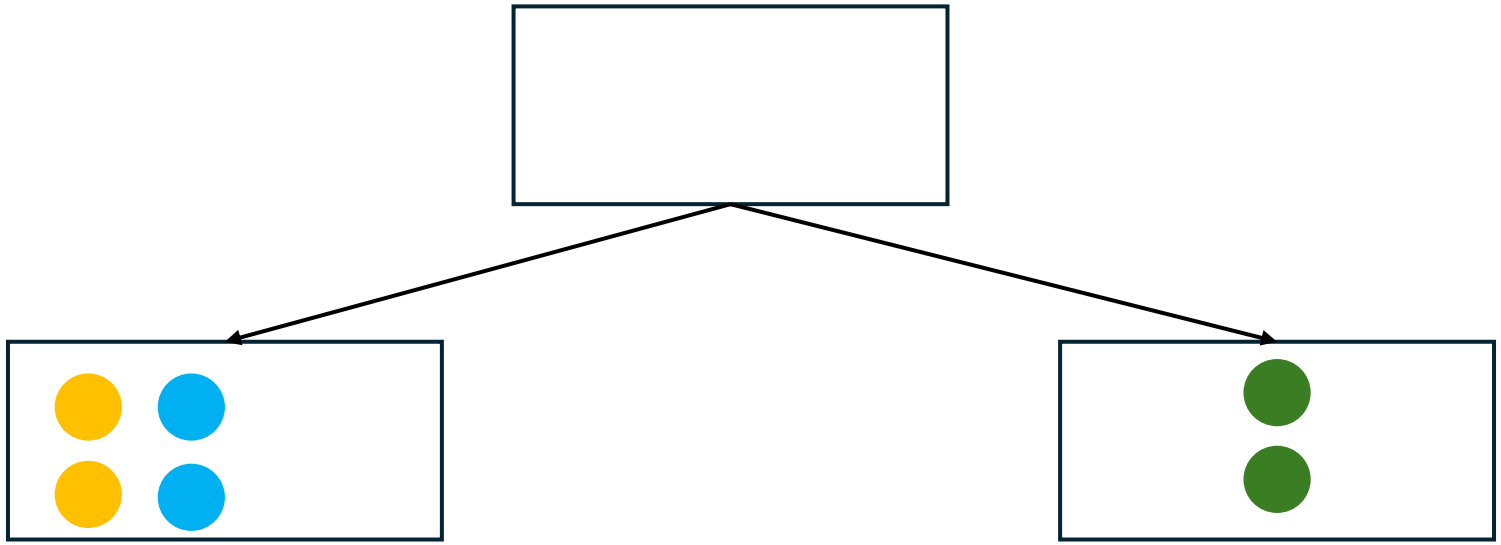
Как разбить вершину?



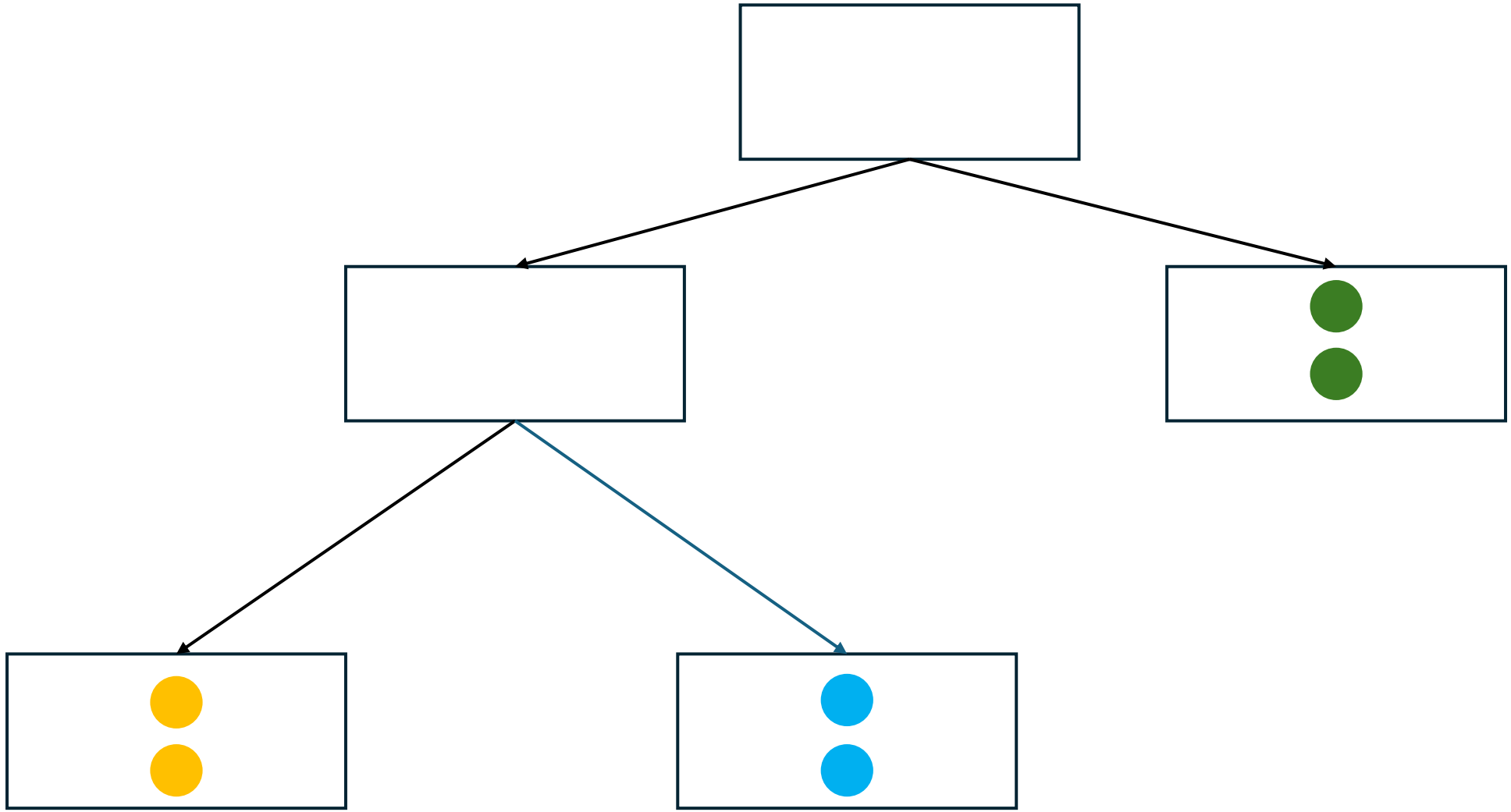
Как разбить вершину?



Как разбить вершину?

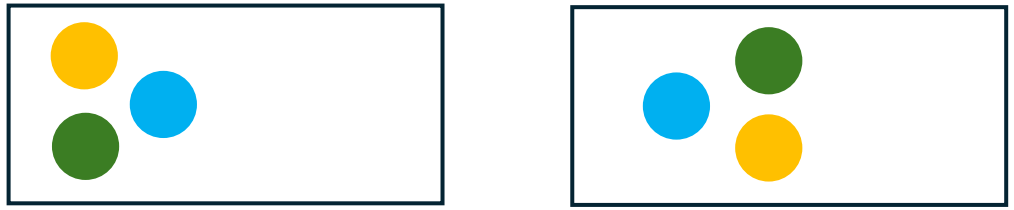


Как разбить вершину?



Как сравнить разбиения?

или



Критерии информативности (impurity criterion)

Возможные функции H(q):

Энтропия:

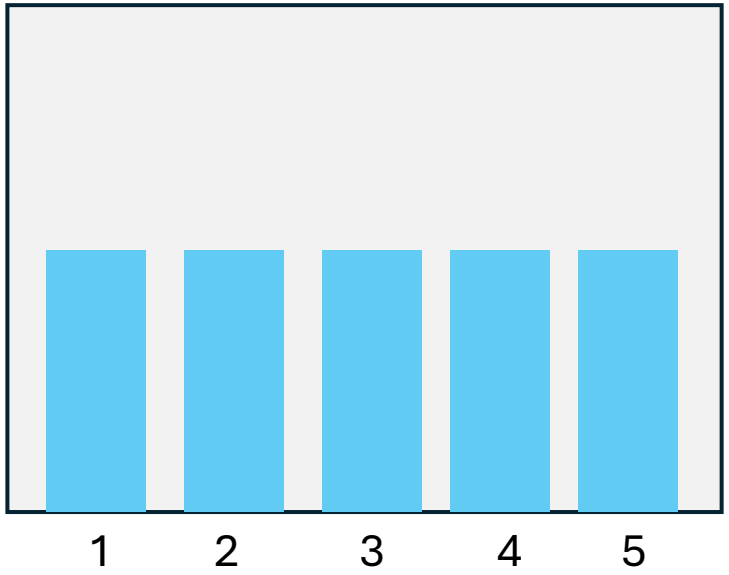
$$H(p_1, \dots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

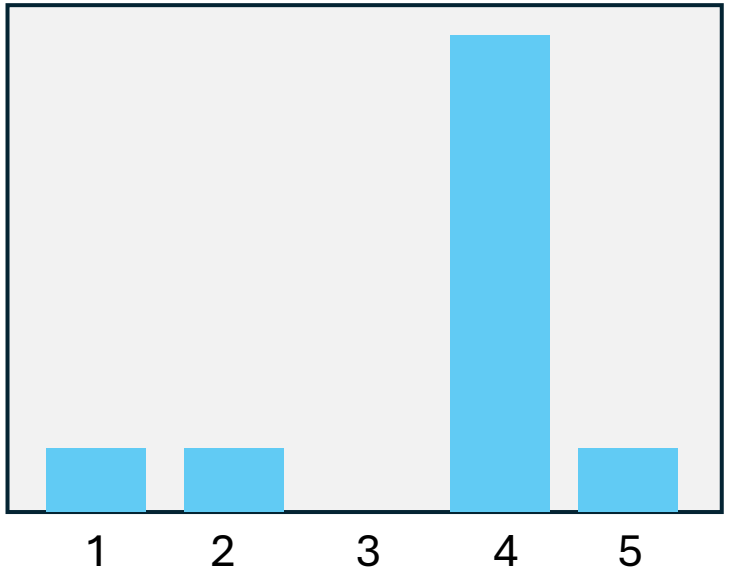
Индекс Джини:

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

Энтропия

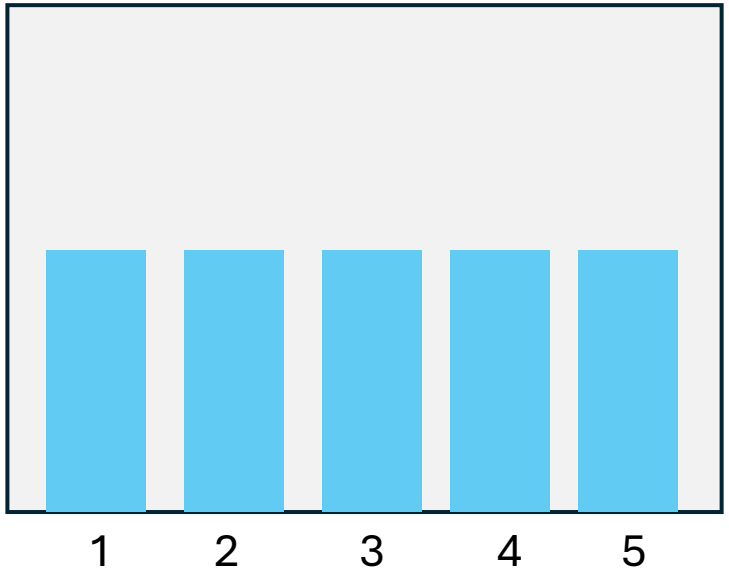
Мера неопределённости распределения



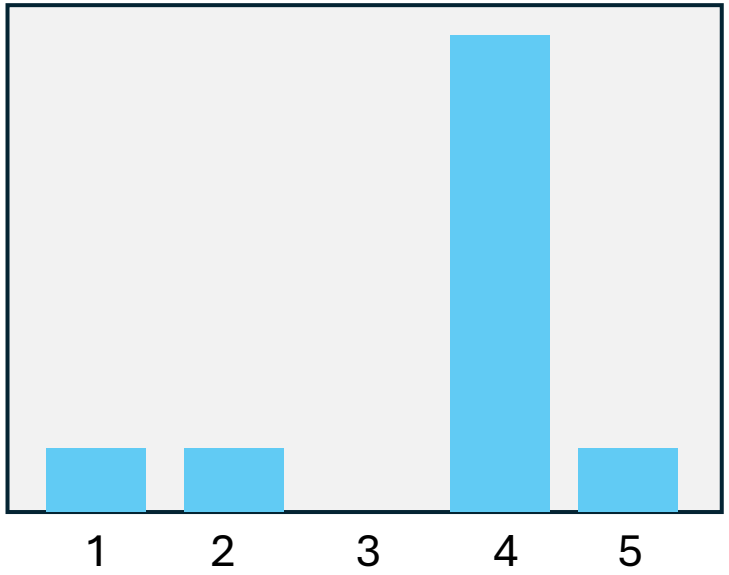


Энтропия

Мера неопределённости распределения



Высокая энтропия



Низкая энтропия

Энтропия

Дискретное распределение

Принимает n значений с вероятностями $p_1, \dots, p_n$

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

Энтропия

Дискретное распределение

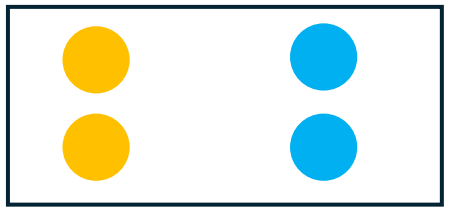
Принимает n значений с вероятностями $p_1, \dots, p_n$

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

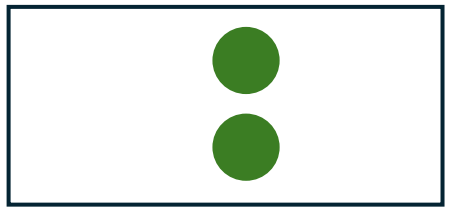
(0.2,0.2,0.2,0.2,0.2)	(0.9,0.05,0.05,0,0)	(0,0,0,1.0)
H=1.60944	H=0.394398	H=0

Как сравнить разбиения?

(0.5,0.5,0)



(0,0,1)



$H(p_L)$= 0.693

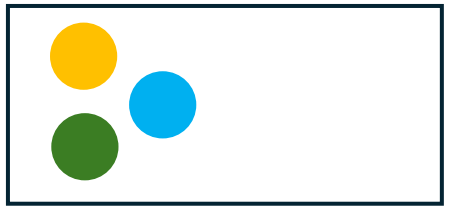
$H(p_R)$= 0

$$H = \frac{L}{Q}H(p_L) + \frac{R}{Q}H(p_R)$$

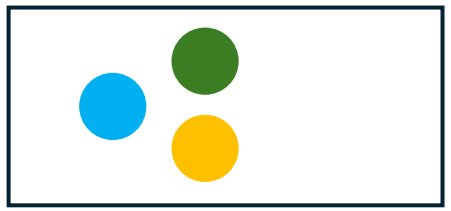
$$H = 0.693 + 0 = 0.693$$

или

(0.3,0.3,0.3)



(0.3,0.3,0.3)



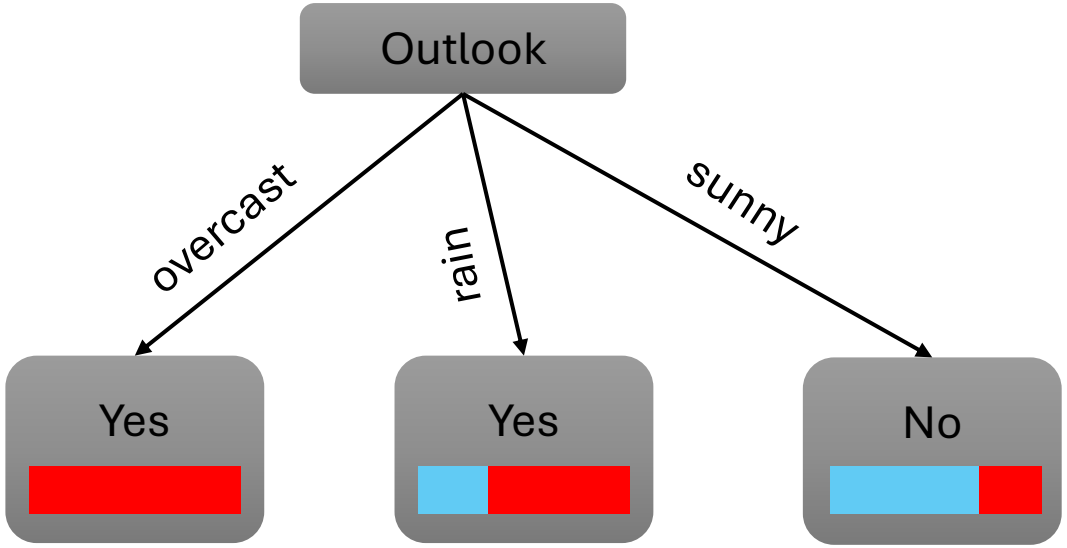
$H(p_L)$= 1.09

$H(p_R)$= 1.09

$$H = 1.09 + 1.09 = 2.18$$

Критерии информативности (impurity criterion)

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

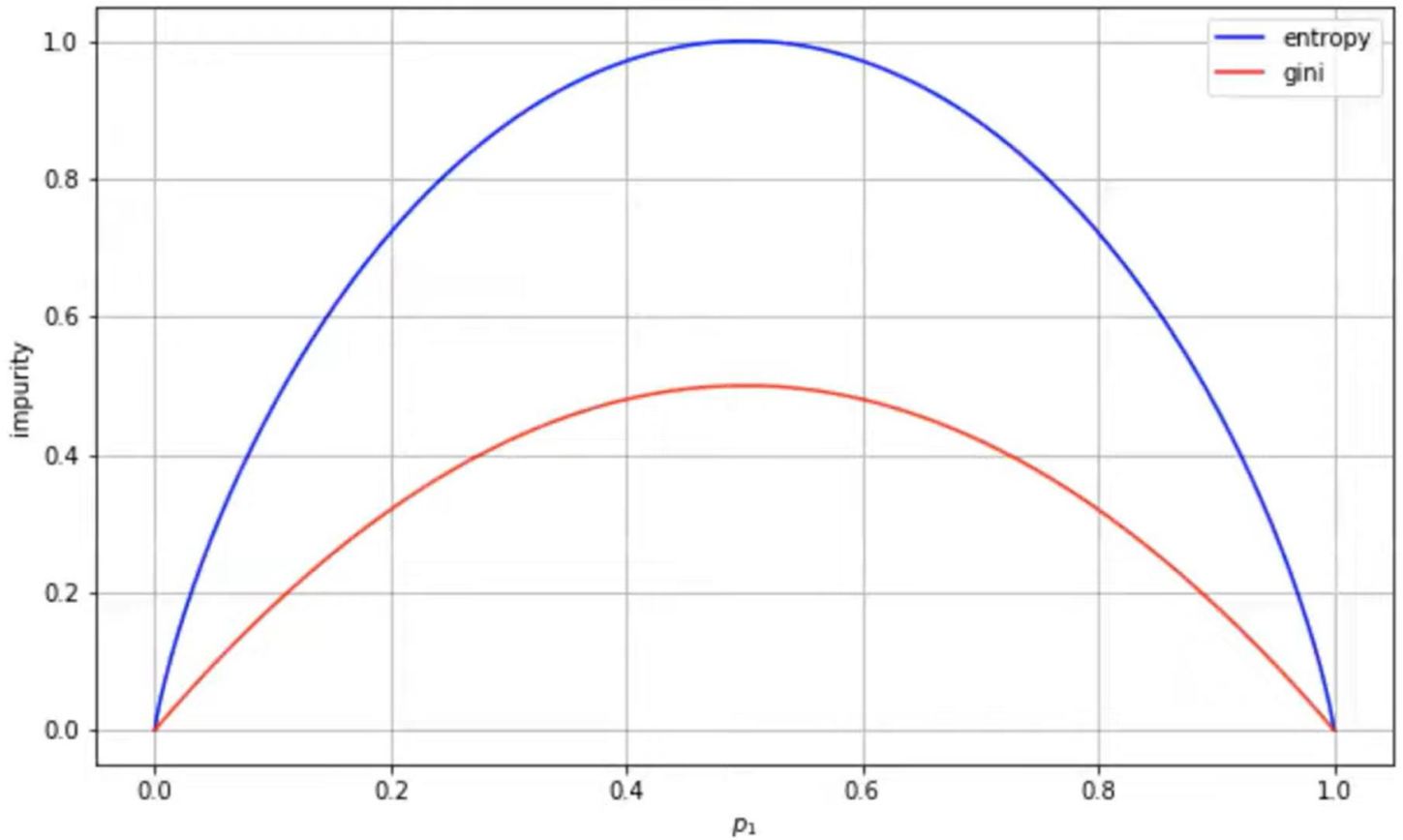


$$H(p_1, \dots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

Характкристика «хаотичности» вершиныы

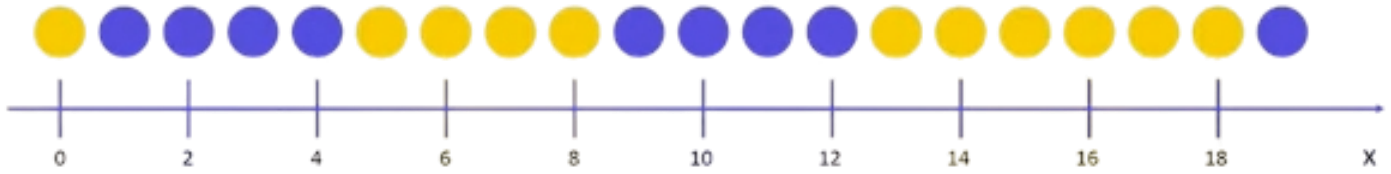
Критерий Джини

$$H(p_1, \ldots, p_n) = \sum_{i=1}^{K} p_i(1 - p_i)$$



Критерии информативности (impurity criterion)

$$Q = 20\,, p_Q = \frac{9}{20}$$



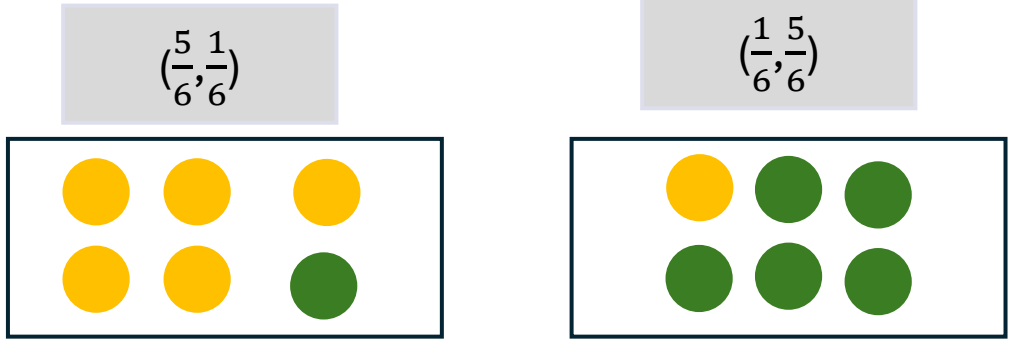
$$L = 13\,, p_L = \frac{8}{13}$$

$$R = 6\,, p_R = \frac{1}{7}$$

$$H(p_L) + H(p_R) \to min$$

Как сравнить разбиения?

$$H = H(p_L) + H(p_R)$$

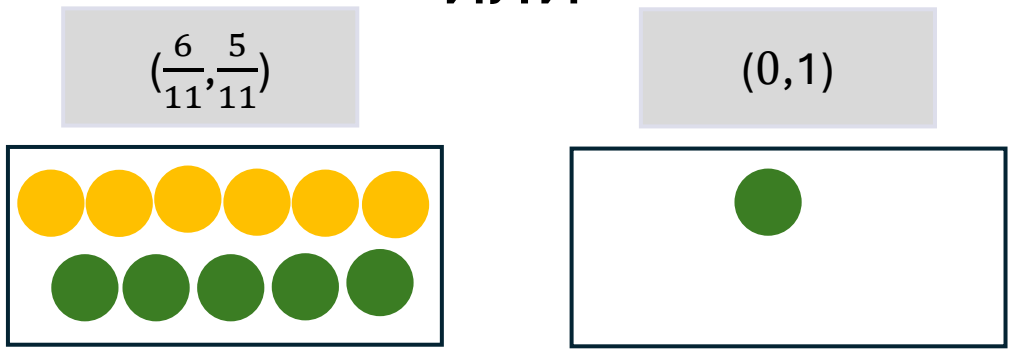


$H(p_L)$= 0.65

$H(p_R)$= 0.65

$$H = 0.65 + 0.65 = 1.3$$

или

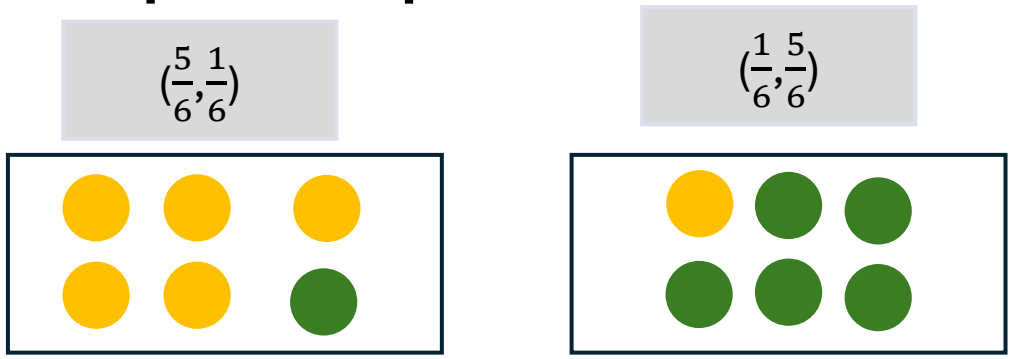


$H(p_L)$= 0.994

$H(p_R)$= 0

$$H = 0.994 + 0 = 0.994$$

Как сравнить разбиения?



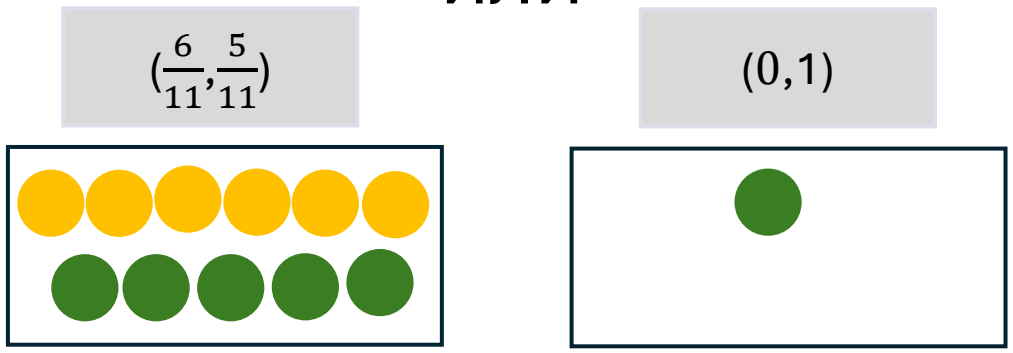
$H(p_L)$= 0.65

$H(p_R)$= 0.65

$$H = \frac{L}{Q}H(p_L) + \frac{R}{Q}H(p_R)$$

$$H = \frac{1}{2} * 0.65 + \frac{1}{2} * 0.65 = 0.65$$

ИЛИ

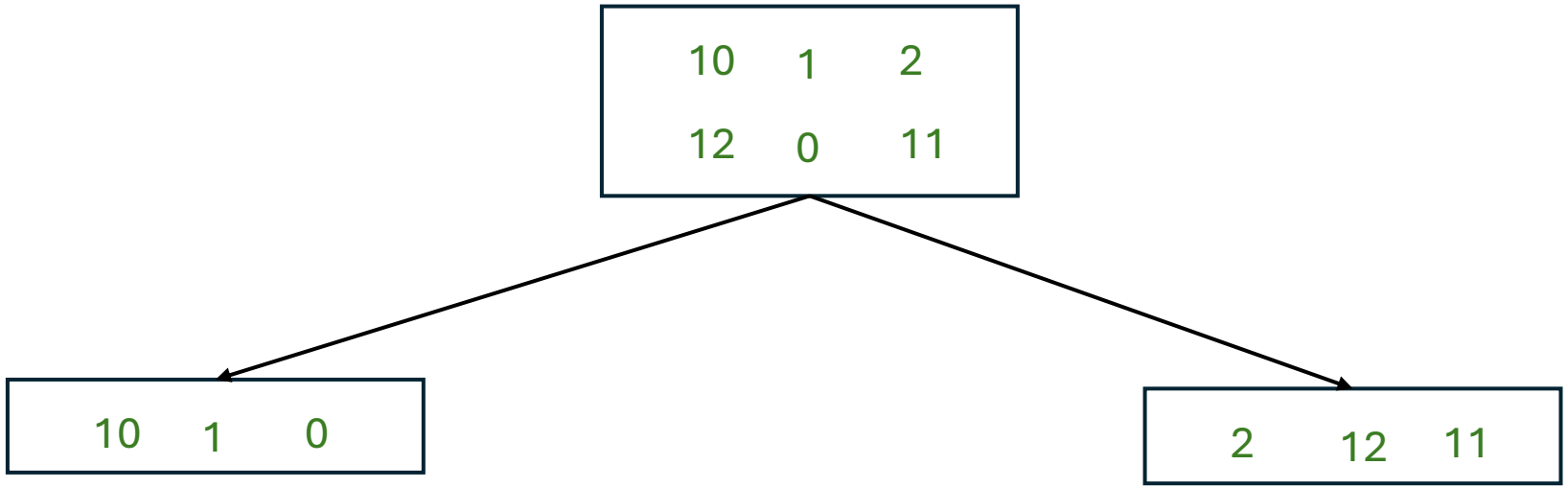


$H(p_L)$= 0.994

$H(p_R)$= 0

$$H = \frac{11}{12} * 0.994 + \frac{1}{12} * 0 = 0.991$$

Для регрессии

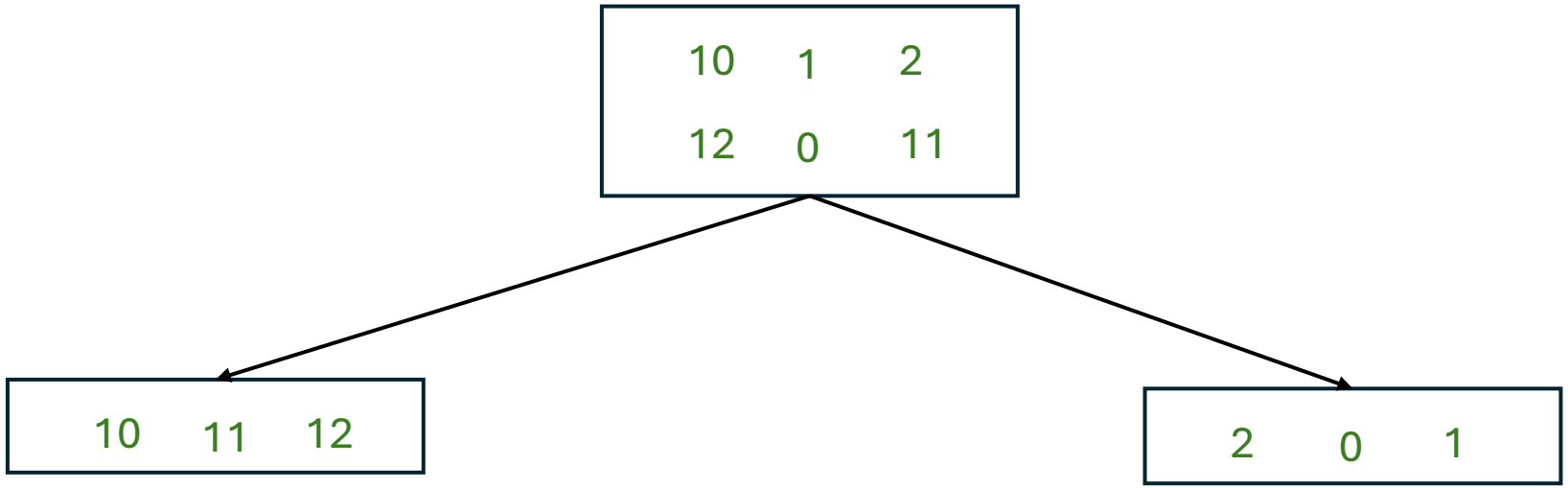


$H(L)$= 20.22

$H(R)$= 20.22

$$\boldsymbol{H = H(L) + H(R) = 40.44}$$

Для регрессии



$H(L)$= 0.66

$H(R)$= 0.66

$$\boldsymbol{H = H(L) + H(R) = 1.33}$$

Для регрессии

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} (y_i - y_R)^2$$

$$y_R = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} y_i$$

Хаотичность вершины можно измерять дисперсией ответов в ней

Жадное построение дерева

Как строить дерево?

Оптимальный вариант: перебрать все возможные деревья

Слишком долго

Жадное построение дерева

Как строить дерево?

- Мы уже умеем выбрать лучший предикат для разбиения вершины
- Будем строить жадно
- Начнём с корня дерева, будем разбивать последовательно, пока не выполнится некоторый критерий остановка

Критерий остановка

- Ограничить глубину
- Ограничить количество листьев
- Задать минимальное число объектов в вершине
- Задать минимальное уменьшение хаотичности при разбиении
- И так далее

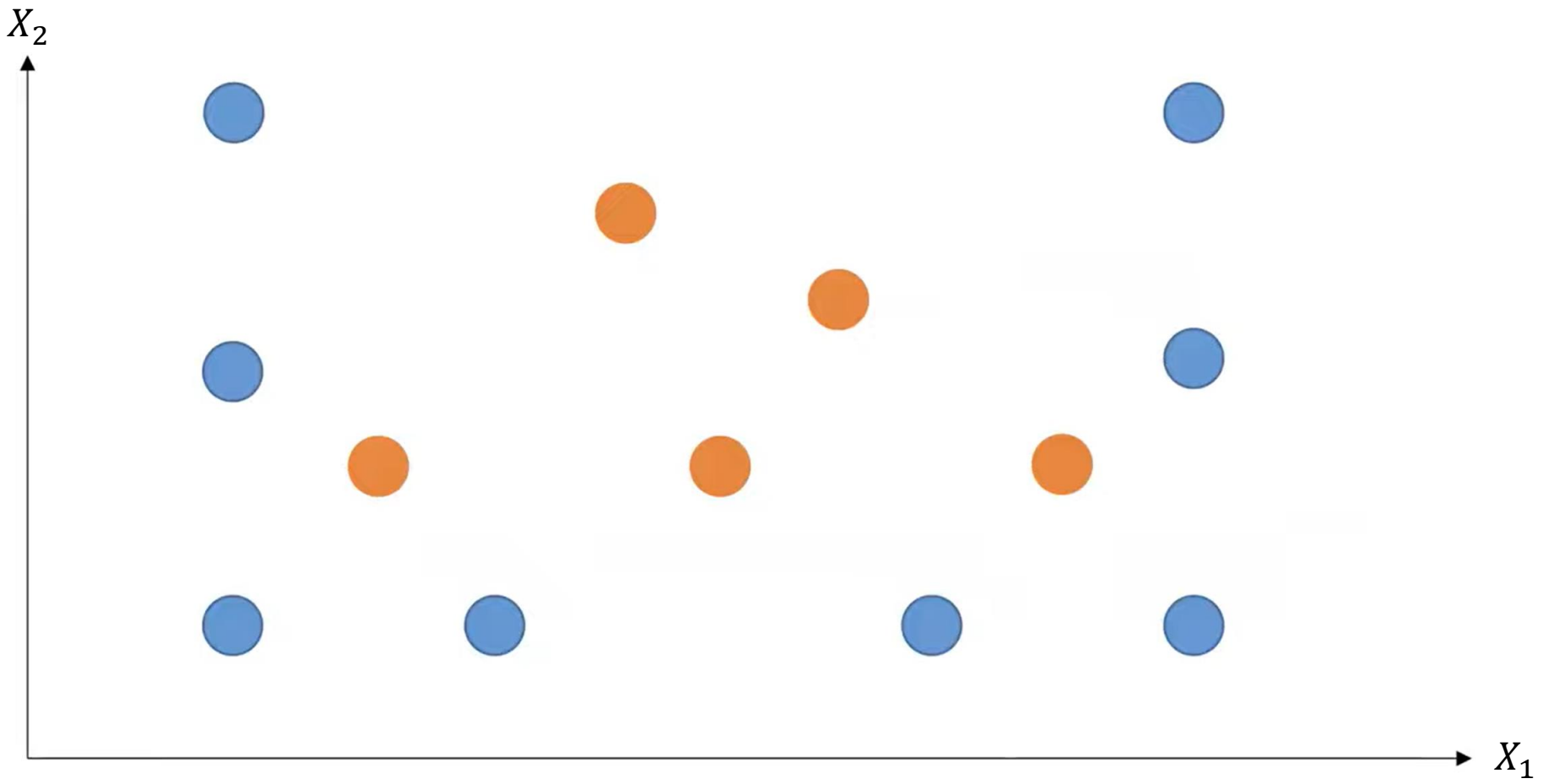
Жадный алгоритм

1. Поместить в корень всю выборку: $R_1 = X$
2. Запустить построение из корня : SplitNode(1, R_1)

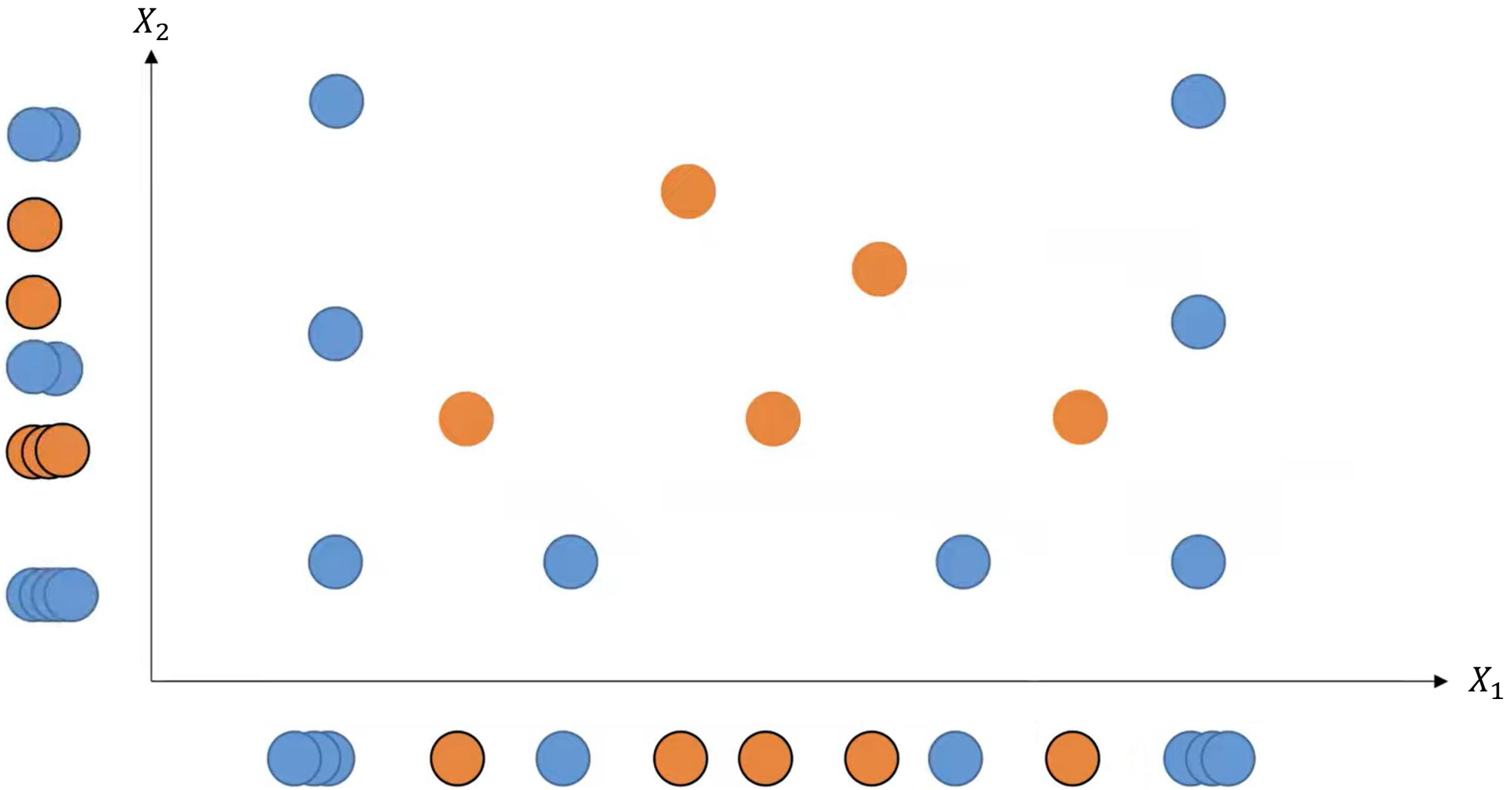
SplitNode(m, R_m)

1. Если выполнен критерий остановка, то выход
2. Ищем лучший предикат: $j, t = \underset{j,t}{\operatorname{argmin}} Q(R_m, j, t)$
3. Разбиваем с его помощью объектов: $R_l = \{(x, y) \in R_m \mid [x_j < t]\}$, $R_r = \{(x, y) \in R_m \mid [x_j \geq t]\}$
4. Повторяем для дочерних вершин: SplitNode(l, R_l), : SplitNode(r, R_r)

Жадный алгоритм

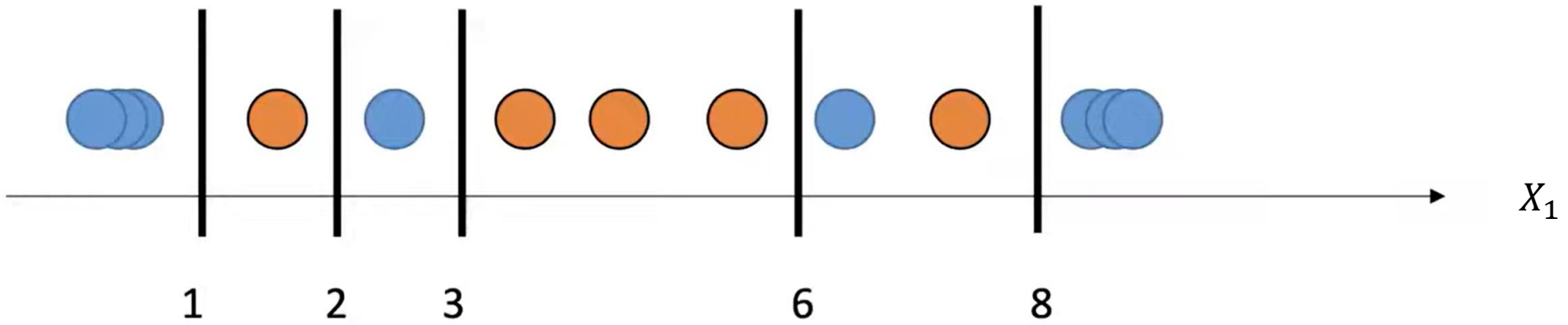


Жадный алгоритм



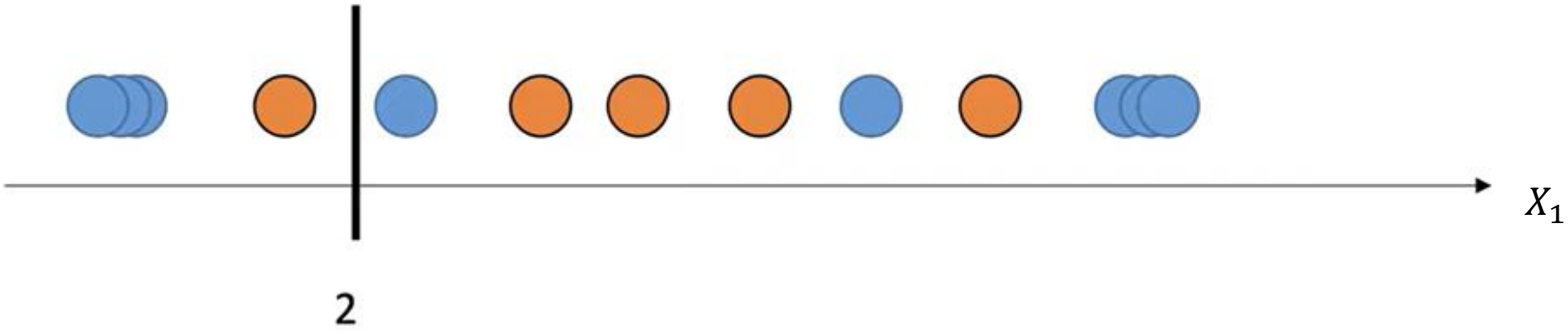
Жадный алгоритм

Разбиения по признаку 1



Жадный алгоритм

Разбиения по признаку 1



$(\frac{3}{4}, \frac{1}{4})$

$H(p_L)$= 0.56

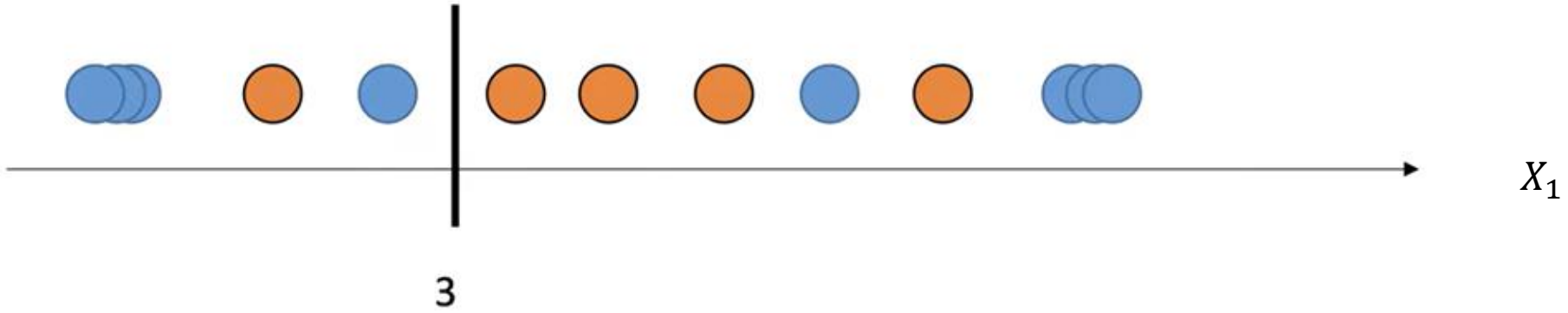
$(\frac{5}{9}, \frac{4}{9})$

$H(p_R)$= 0.69

$$H = \frac{4}{13} H(p_L) + \frac{9}{13} H(p_R) = 0.65$$

Жадный алгоритм

Разбиения по признаку 1



$(\frac{4}{5}, \frac{1}{5})$

$(\frac{1}{2}, \frac{1}{2})$

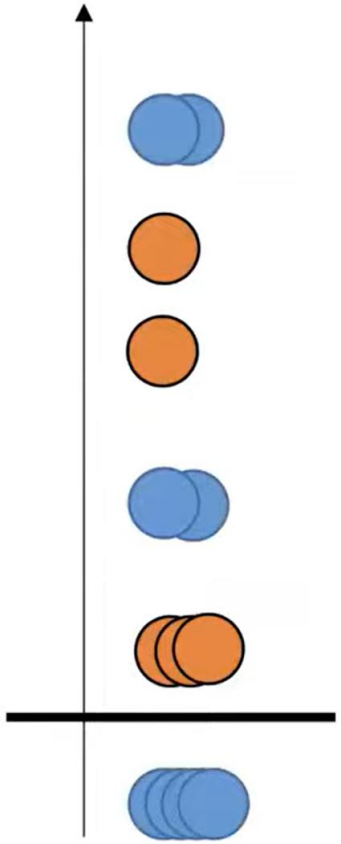
$H(p_L)$= 0.5

$H(p_R)$= 0.69

$$H = \frac{5}{13} H(p_L) + \frac{8}{13} H(p_R) = 0.62$$

Жадный алгоритм

Разбиения по признаку 2



1

X_2

$(\frac{4}{9}, \frac{5}{9})$

$H(p_L)$= 0.69

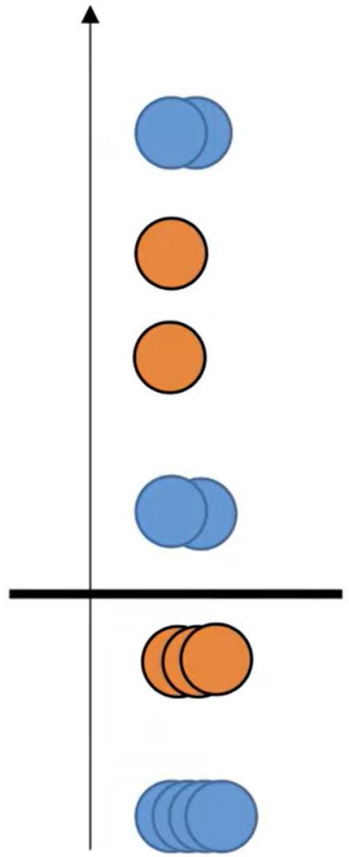
(1,0)

$H(p_R)$= 0

$$H = \frac{4}{13} H(p_L) + \frac{9}{13} H(p_R) = 0.47$$

Жадный алгоритм

Разбиения по признаку 2



2

X_2

$(\frac{4}{6}, \frac{2}{6})$

$H(p_L)$= 0.5

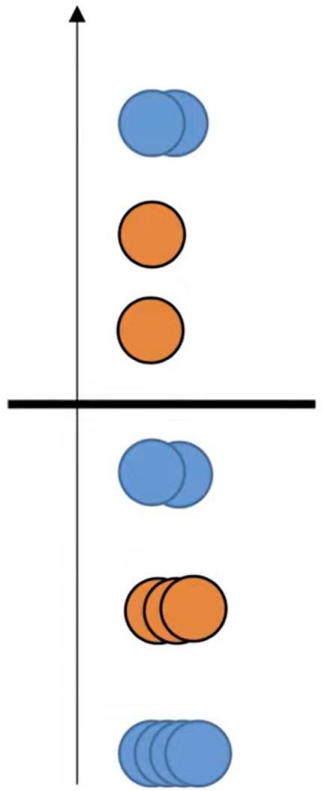
$(\frac{4}{7}, \frac{3}{7})$

$H(p_R)$= 0.69

$$H = \frac{7}{13} H(p_L) + \frac{6}{13} H(p_R) = 0.66$$

Жадный алгоритм

Разбиения по признаку 2



$(\frac{1}{2}, \frac{1}{2})$

$H(p_L)$= 0.69

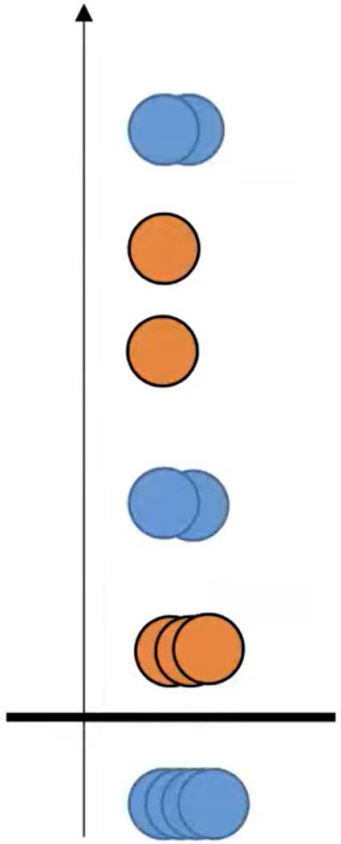
$(\frac{6}{9}, \frac{3}{9})$

$H(p_R)$= 0.46

$$H = \frac{9}{13} H(p_L) + \frac{4}{13} H(p_R) = 0.53$$

Жадный алгоритм

Разбиения по признаку 2



$(\frac{4}{9}, \frac{5}{9})$

$H(p_L)$= 0.69

$(1,0)$

$H(p_R)$= 0

$$H = \frac{4}{13} H(p_L) + \frac{9}{13} H(p_R) = 0.47$$

Лучшее разбиение!

Жадный алгоритм