

Введение в анализ данных

Лекция 10

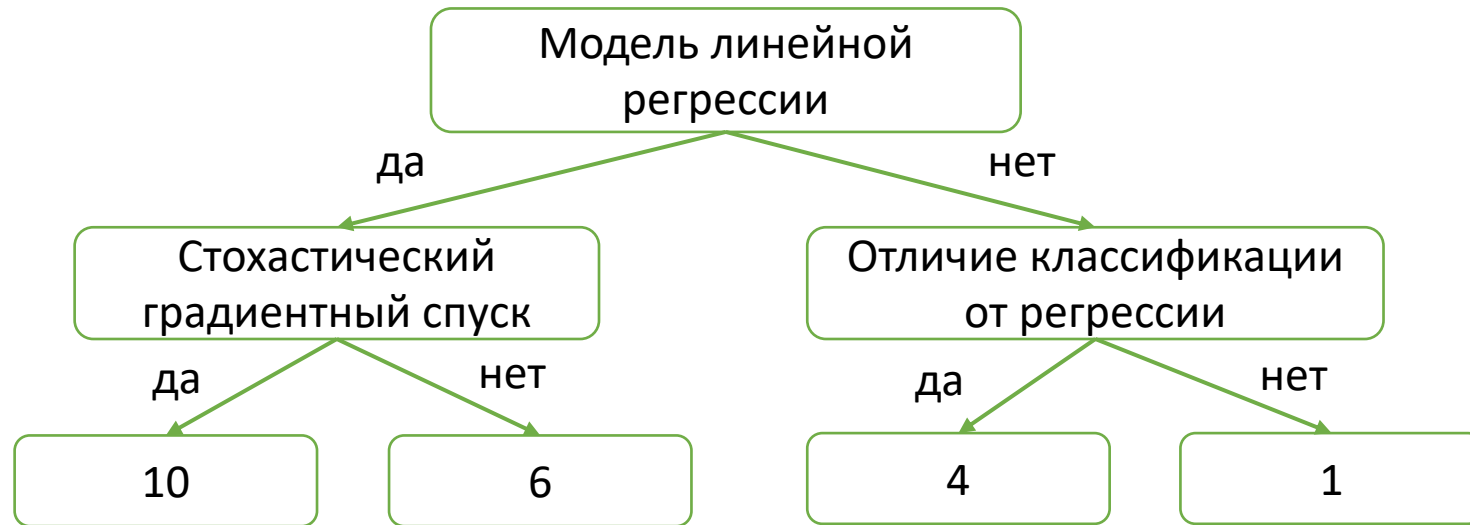
Решающие деревья и композиции моделей

Евгений Соколов

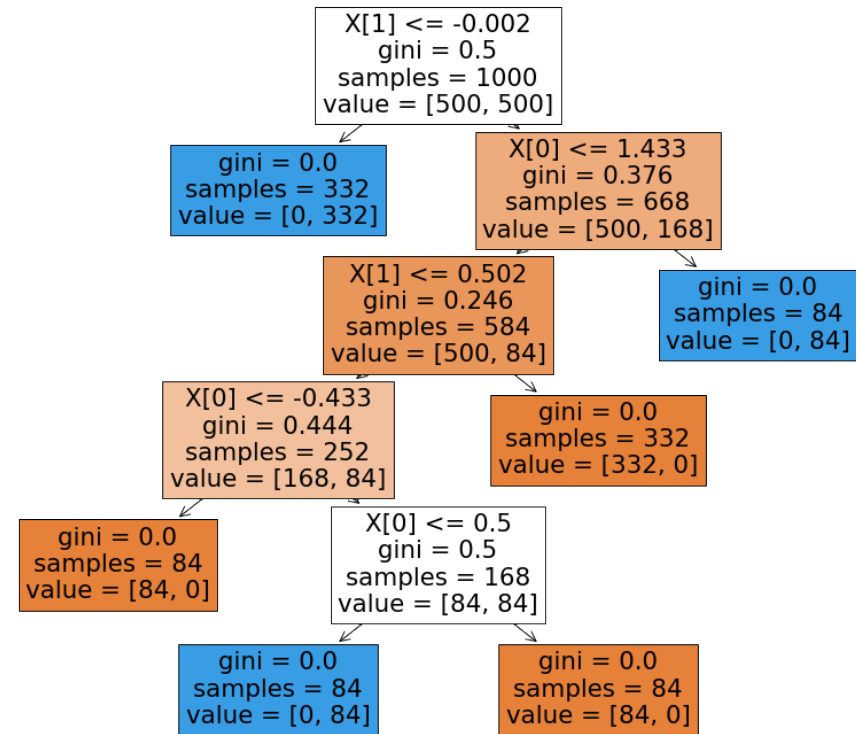
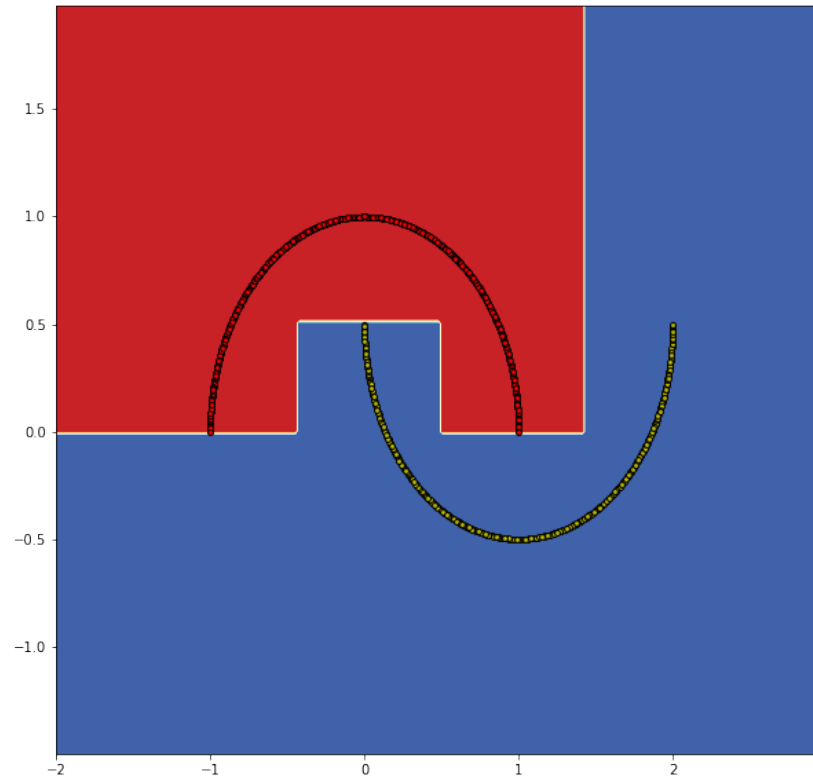
esokolov@hse.ru

НИУ ВШЭ, 2020

Решающее дерево

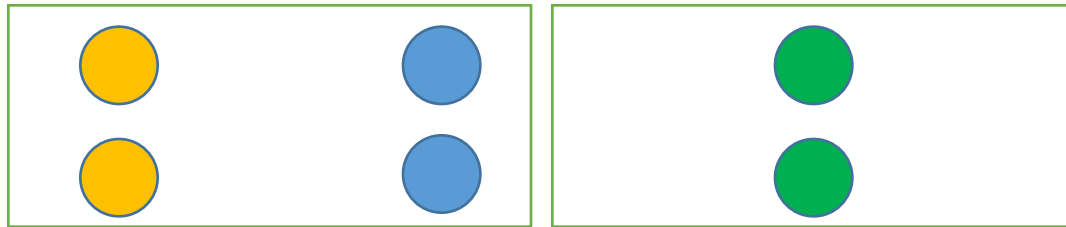


Решающее дерево



Как выбирать предикаты

Как сравнить разбиения?



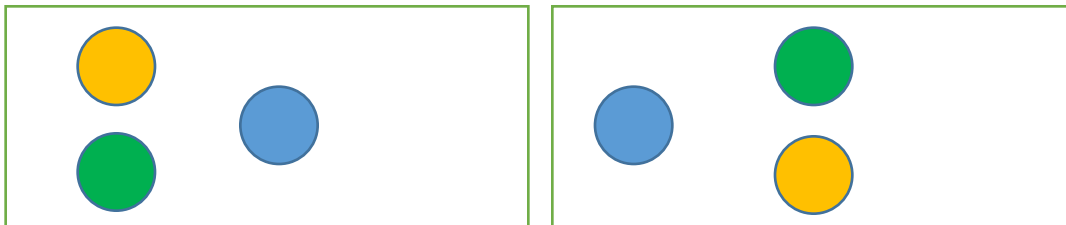
0.693

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

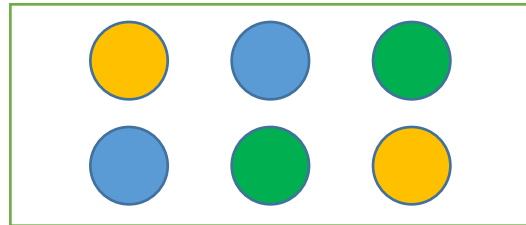
Энтропия

$$H(p_1, \dots, p_K) = - \sum_{i=1}^K p_i \log_2 p_i$$

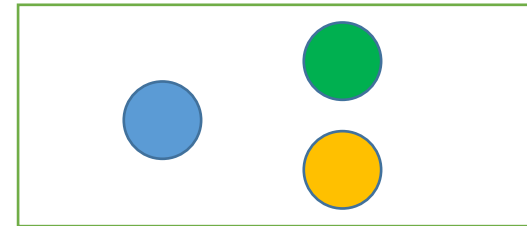
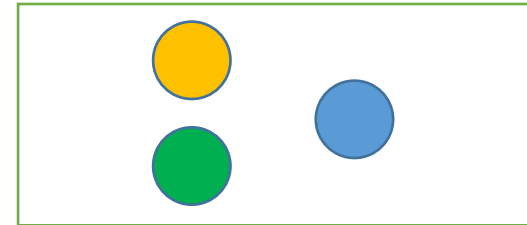
- Характеристика «хаотичности» вершины
- Impurity

Критерий информативности

- Как понять, какой предикат лучше?
- Сравнить хаотичность в исходной вершине и в двух дочерних!

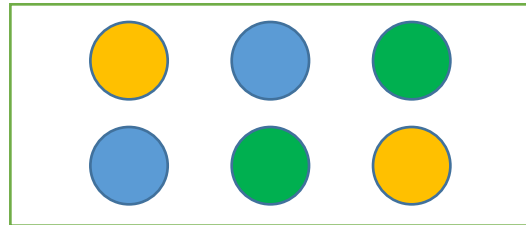


против

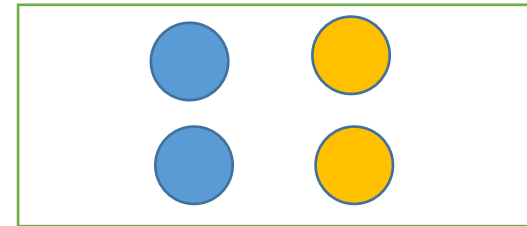
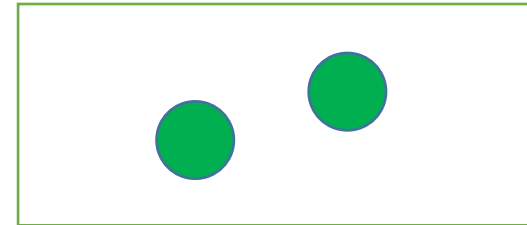


Критерий информативности

- Как понять, какой предикат лучше?
- Сравнить хаотичность в исходной вершине и в двух дочерних!



против



Критерий информативности

- Как понять, какой предикат лучше?
- Сравнить хаотичность в исходной вершине и в двух дочерних!

$$Q(R, j, t) = H(R) - H(R_\ell) - H(R_r) \rightarrow \max_{j, t}$$

Критерий информативности

$$Q(R, j, t) = H(R) - \frac{|R_\ell|}{|R|} H(R_\ell) - \frac{|R_r|}{|R|} H(R_r) \rightarrow \max_{j,t}$$

- Или так:

$$Q(R, j, t) = \frac{|R_\ell|}{|R|} H(R_\ell) + \frac{|R_r|}{|R|} H(R_r) \rightarrow \min_{j,t}$$

Задача регрессии

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

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

- То есть «хаотичность» вершины можно измерять дисперсией ответов в ней

Жадное построение дерева

Как строить дерево?

- Оптимальный вариант: перебрать все возможные деревья, выбрать самое маленькое среди безошибочных
- Слишком долго

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 останова

Критерий останова

- 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: $\text{SplitNode}(1, R_1)$

Жадный алгоритм

- $\text{SplitNode}(m, R_m)$

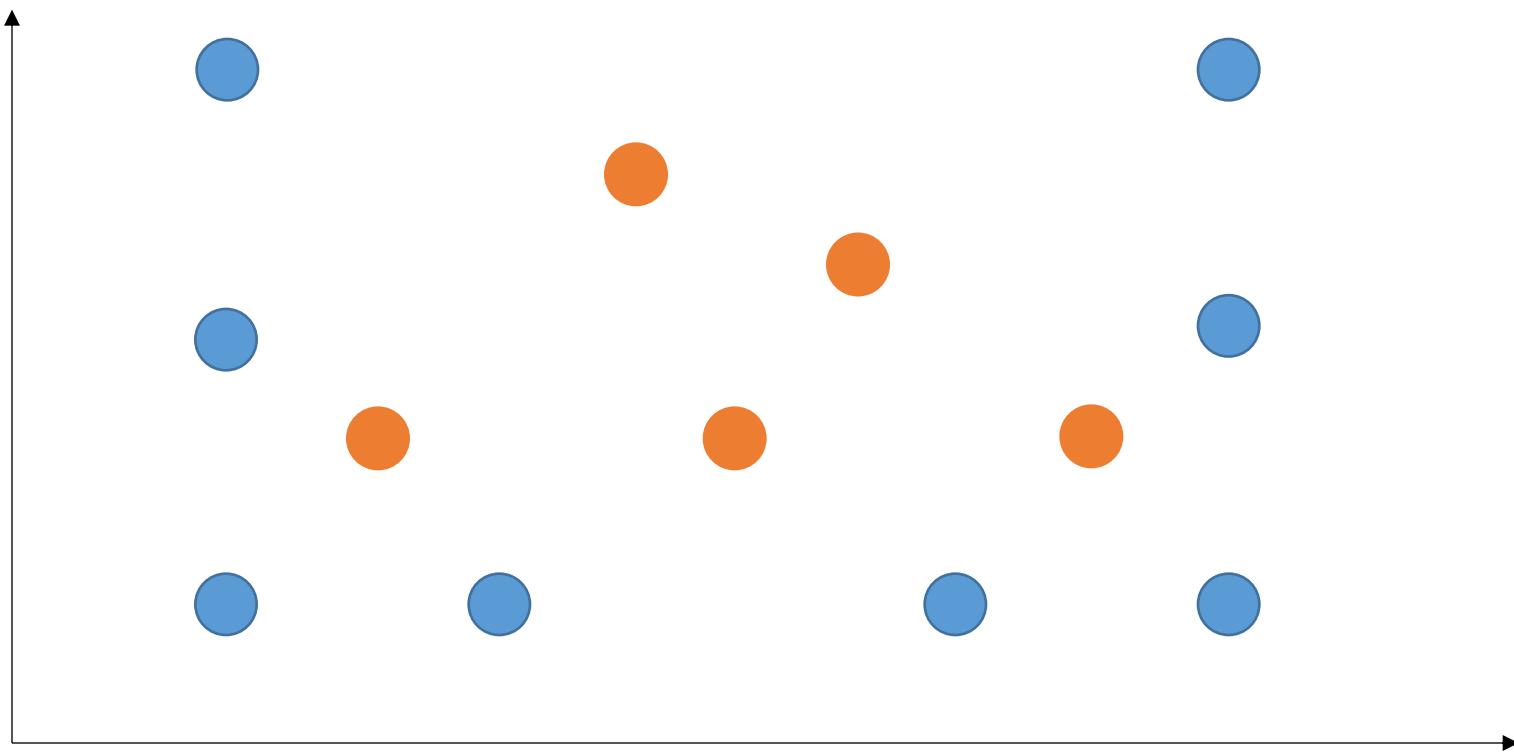
1. Если выполнен критерий останова, то выход

2. Ищем лучший предикат: $j, t = \arg \min_{j, t} Q(R_m, j, t)$

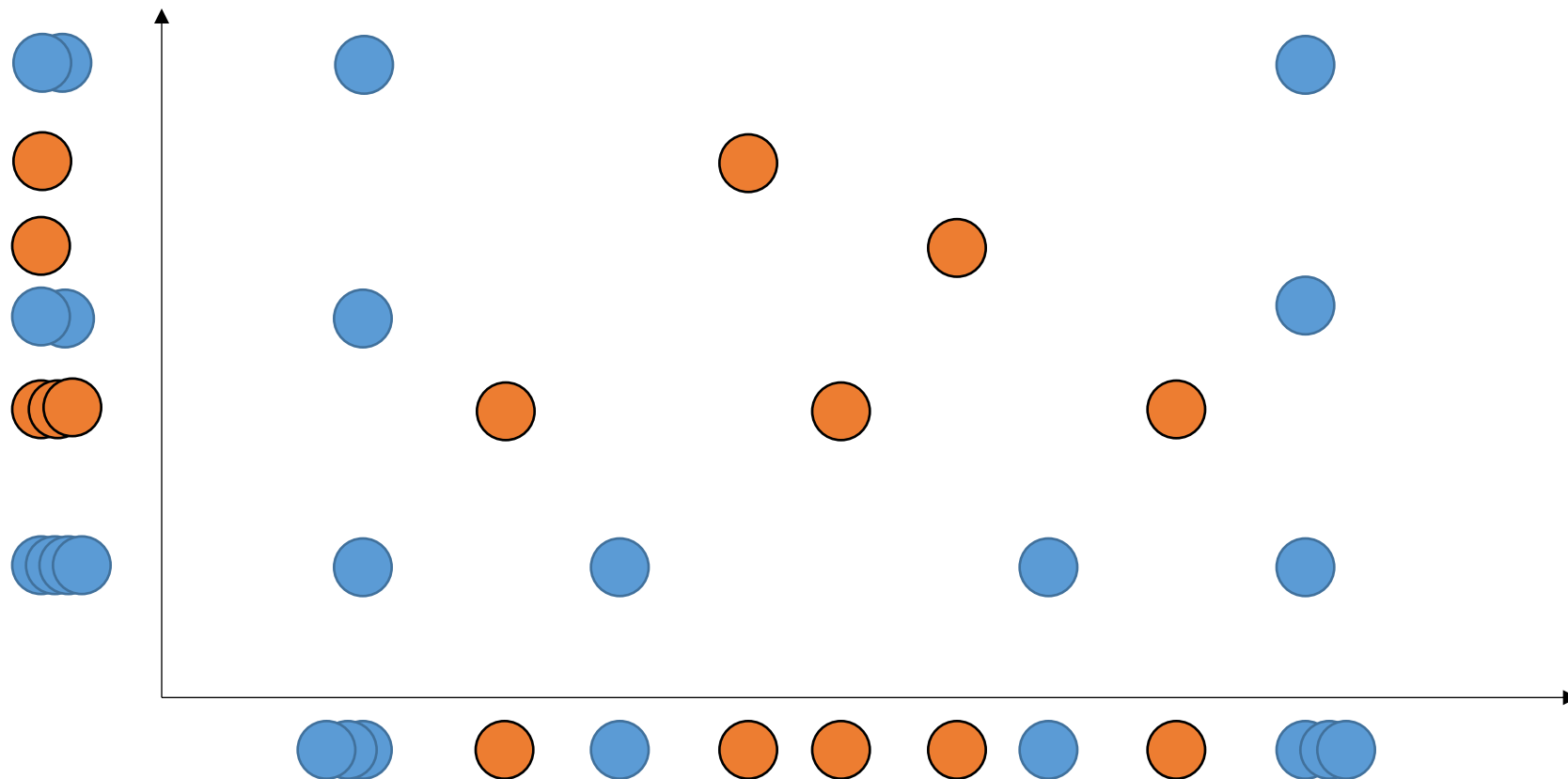
3. Разбиваем с его помощью объекты: $R_\ell = \left\{ \{(x, y) \in R_m \mid [x_j < t] \} \right\},$
 $R_r = \left\{ \{(x, y) \in R_m \mid [x_j \geq t] \} \right\}$

4. Повторяем для дочерних вершин: $\text{SplitNode}(\ell, R_\ell)$ и $\text{SplitNode}(r, R_r)$

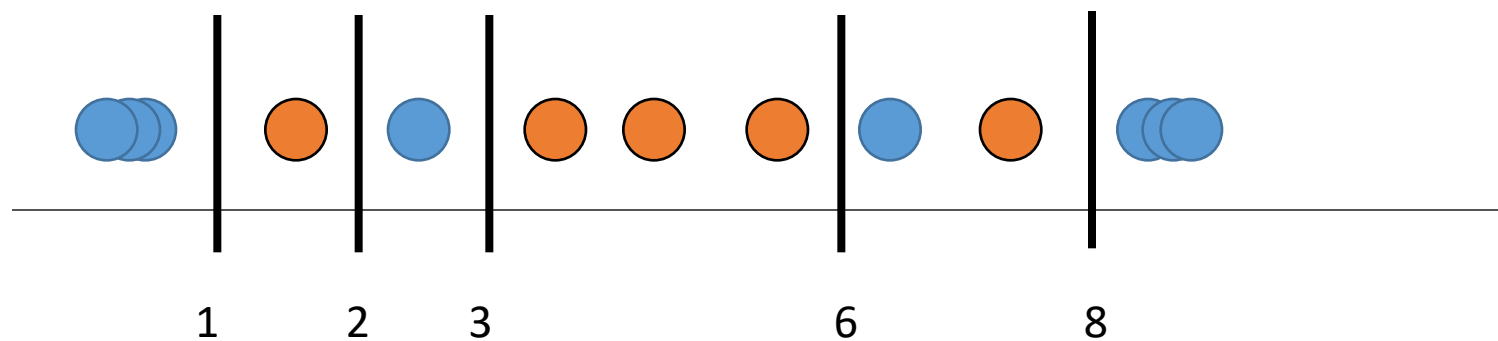
Обучение деревьев



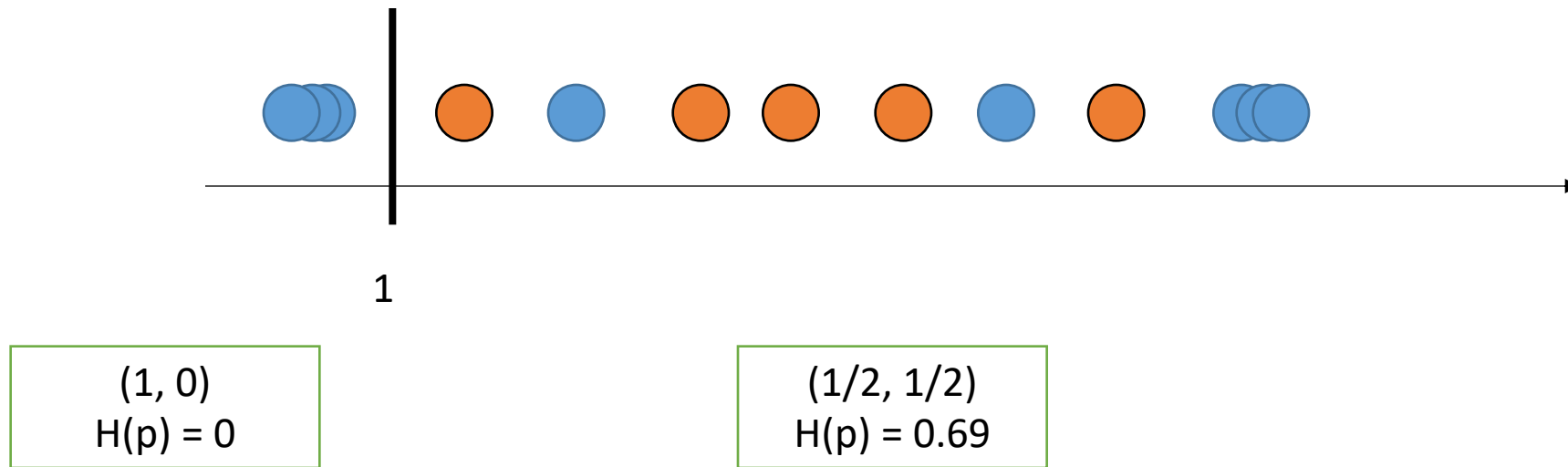
Признаки



Разбиения по признаку 1

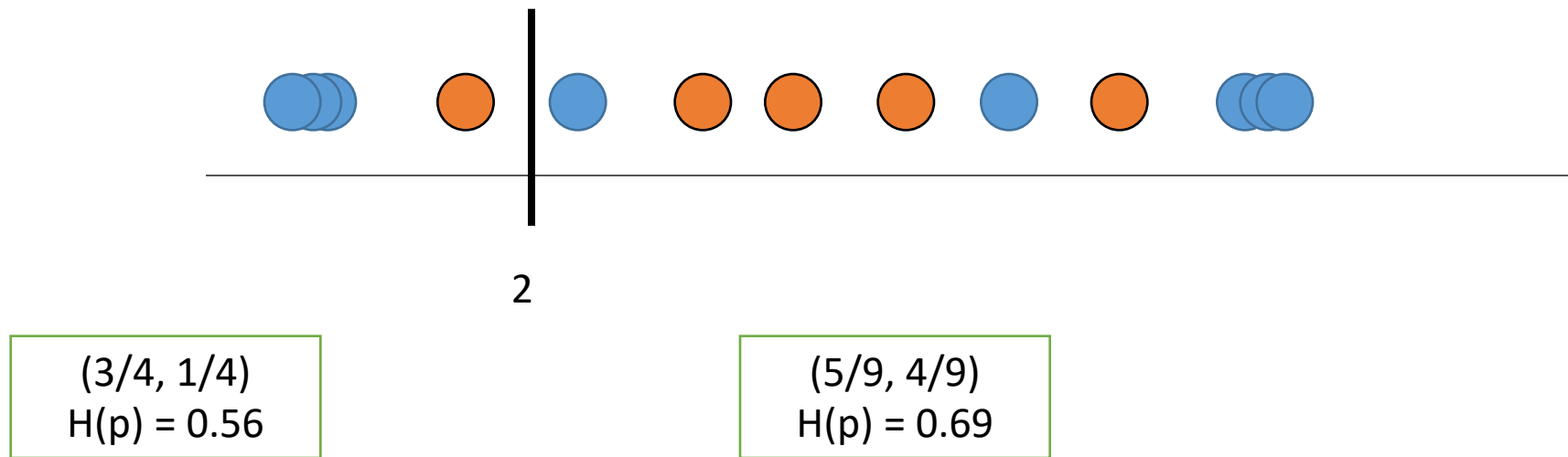


Разбиения по признаку 1



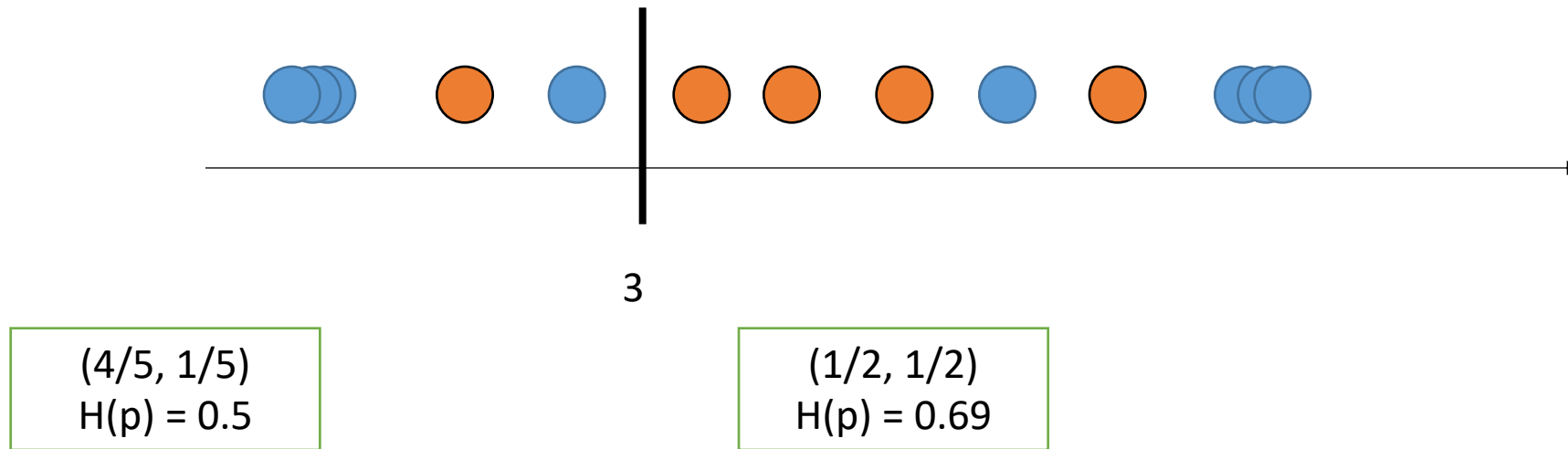
$$\frac{3}{13}H(p_l) + \frac{10}{13}H(p_r) = 0.53$$

Разбиения по признаку 1



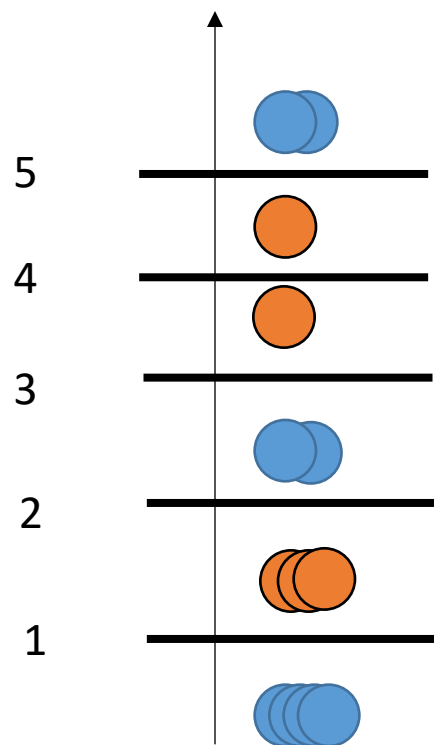
$$\frac{4}{13}H(p_l) + \frac{9}{13}H(p_r) = 0.65$$

Разбиения по признаку 1

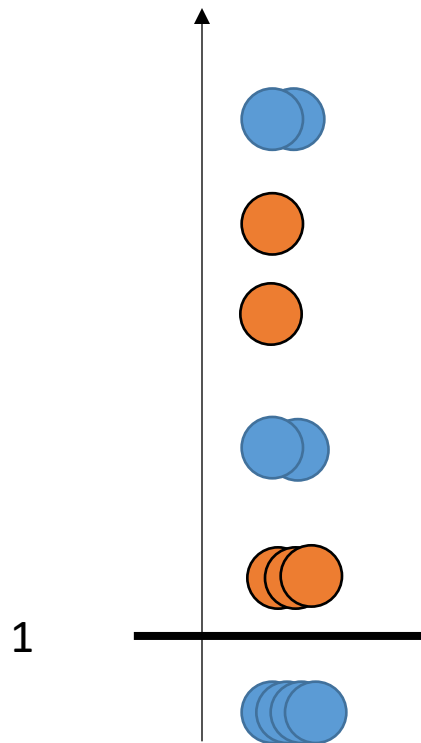


$$\frac{5}{13}H(p_l) + \frac{8}{13}H(p_r) = 0.62$$

Разбиения по признаку 2



Разбиения по признаку 2

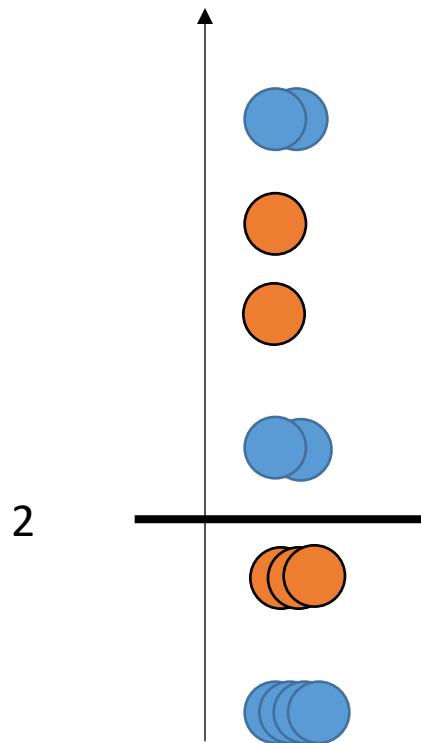


$(4/9, 5/9)$
 $H(p) = 0.69$

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

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

Разбиения по признаку 2

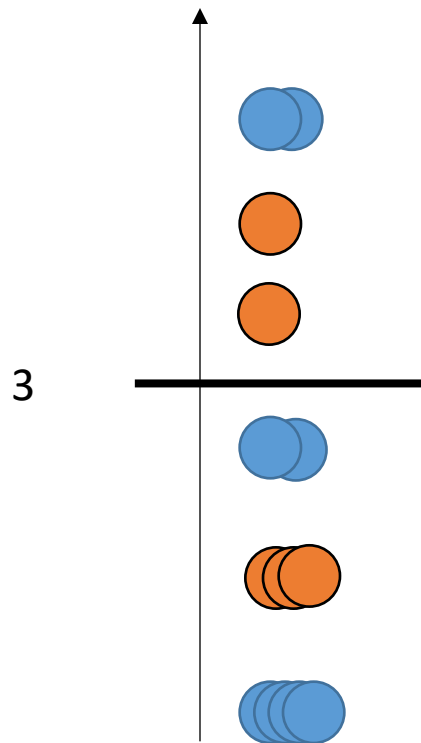


$(4/6, 2/6)$
 $H(p) = 0.64$

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

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

Разбиения по признаку 2

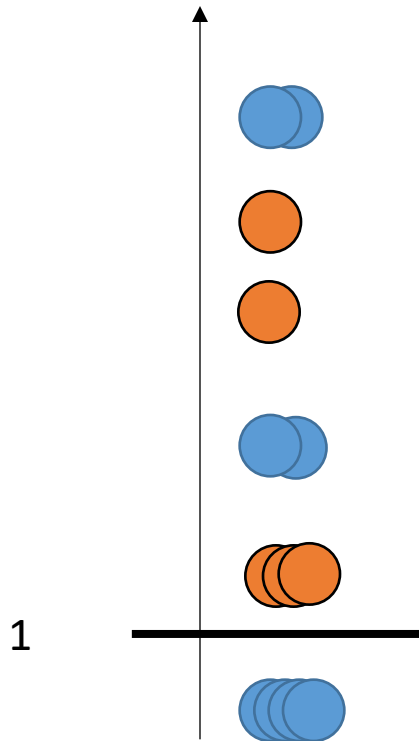


$(1/2, 1/2)$
 $H(p) = 0.69$

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

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

Разбиения по признаку 2



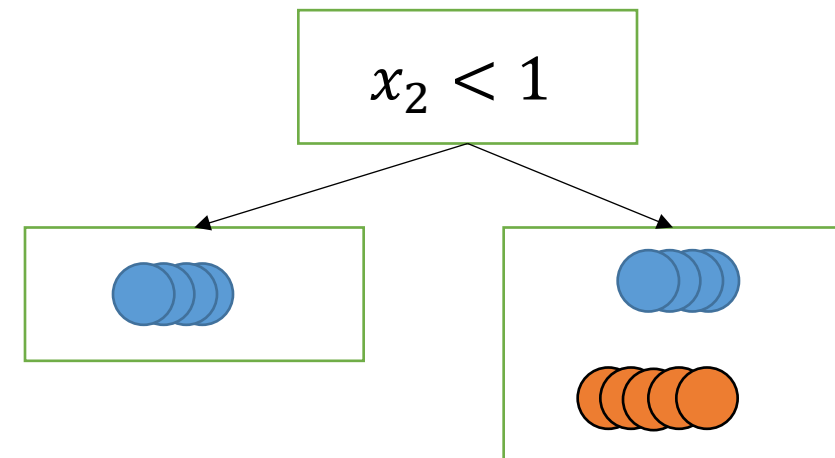
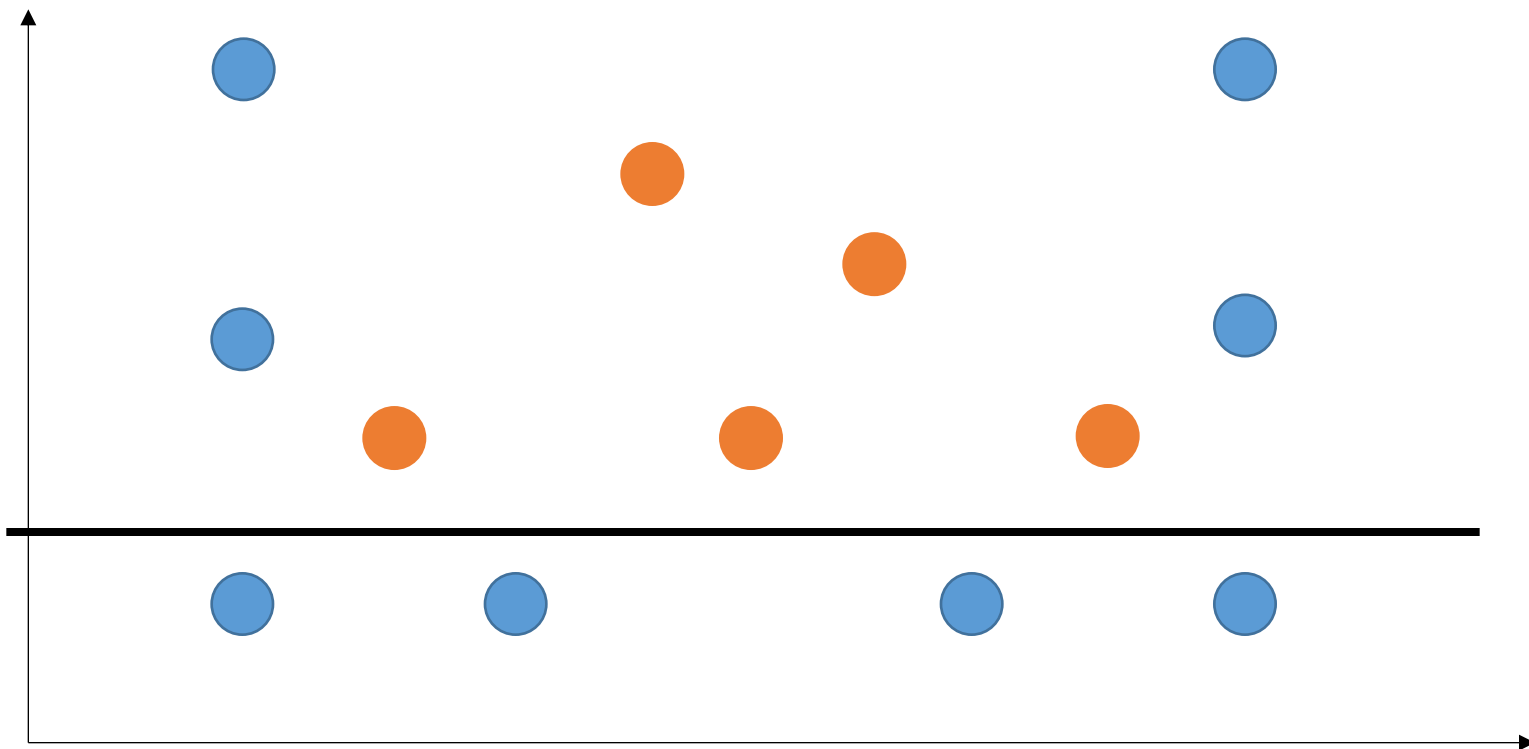
$(4/9, 5/9)$
 $H(p) = 0.69$

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

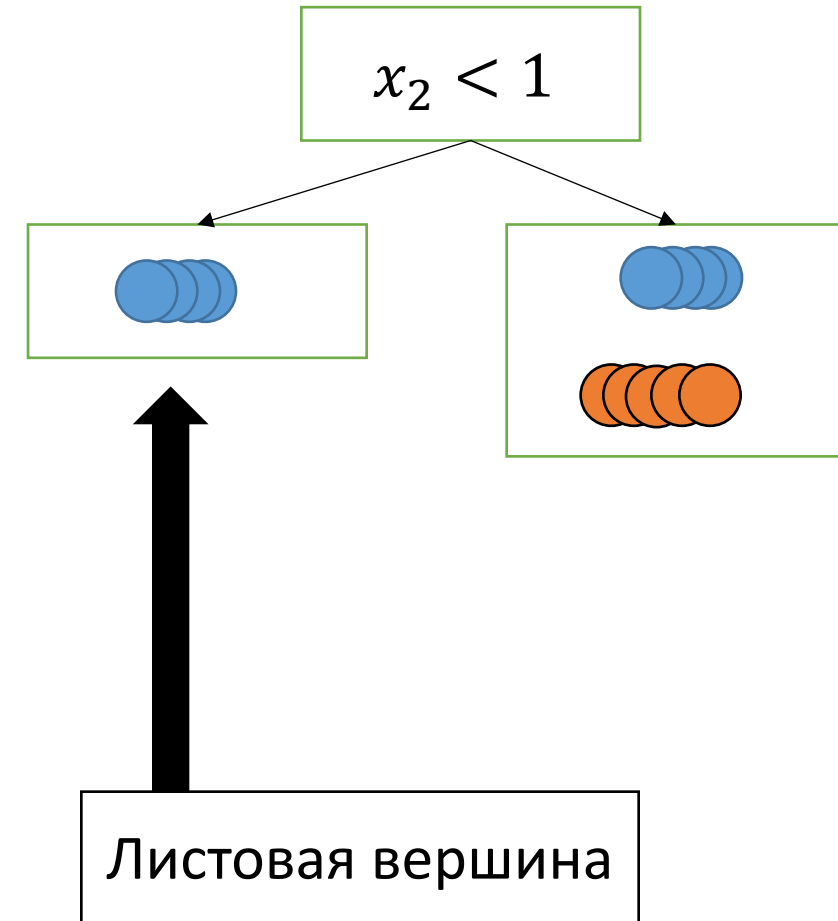
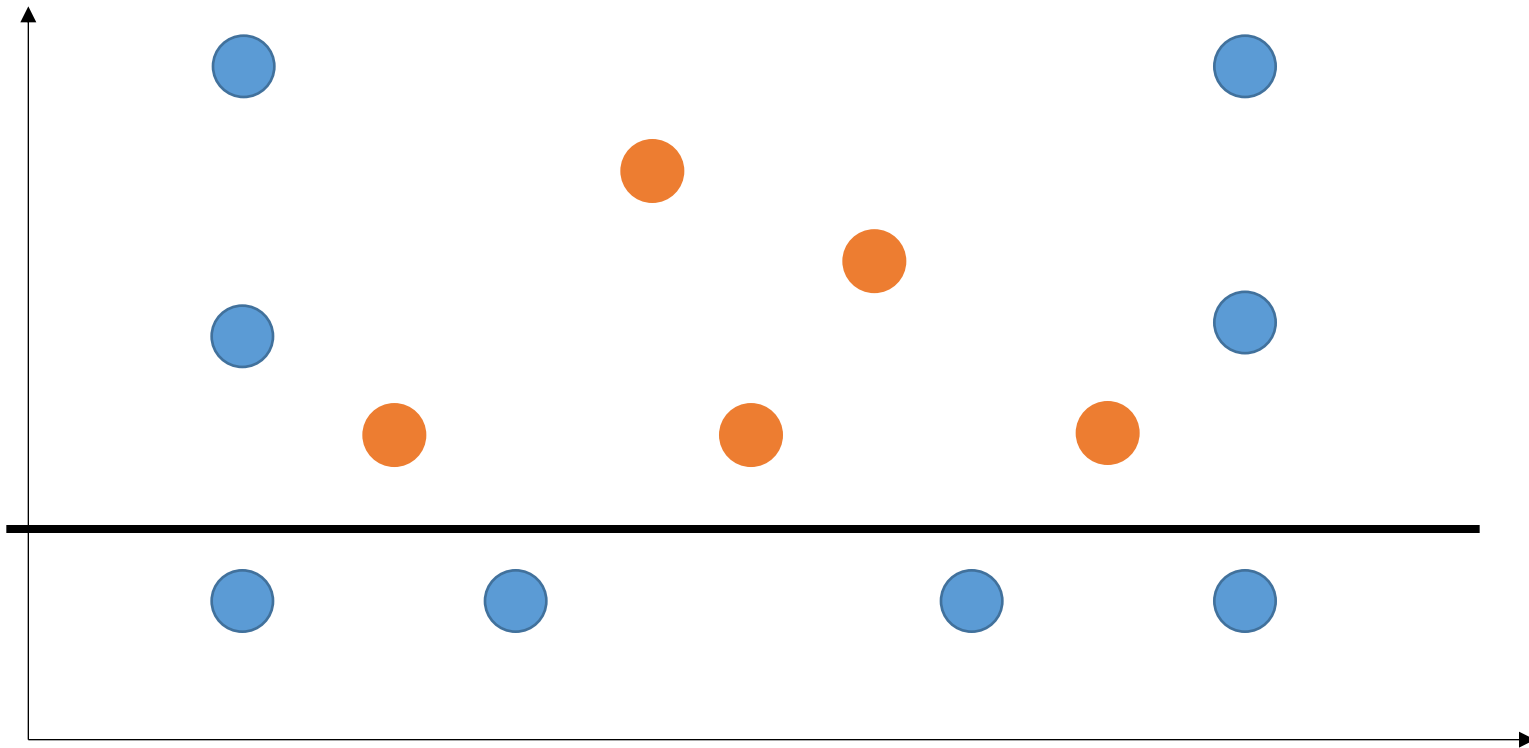
$$\frac{4}{13}H(p_l) + \frac{9}{13}H(p_r) = 0.47$$

Лучшее разбиение!

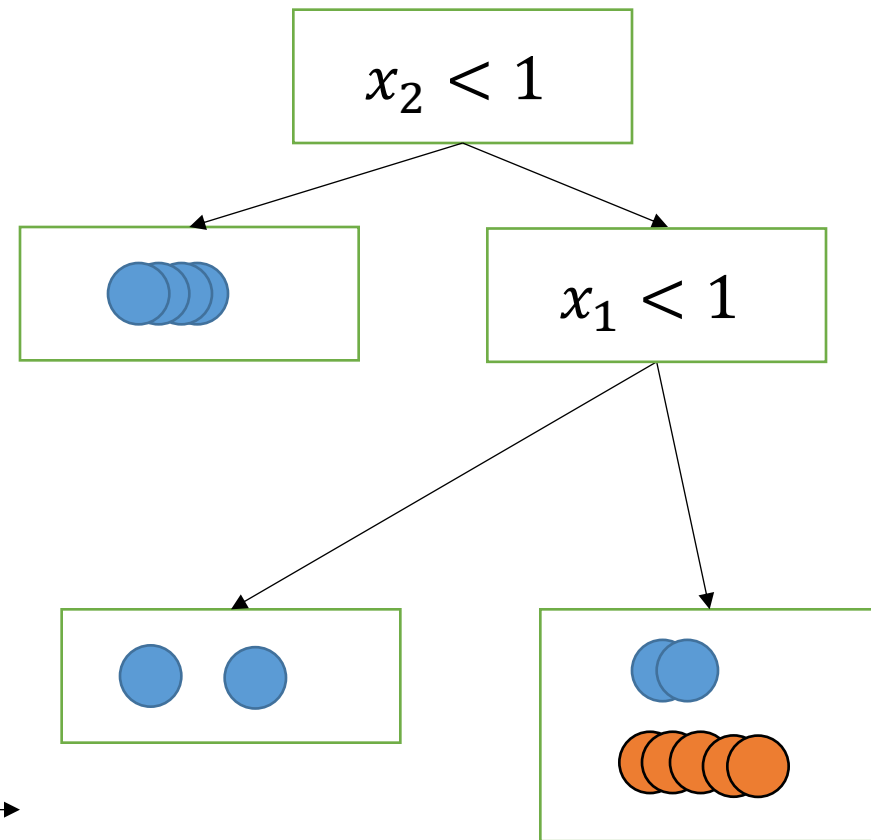
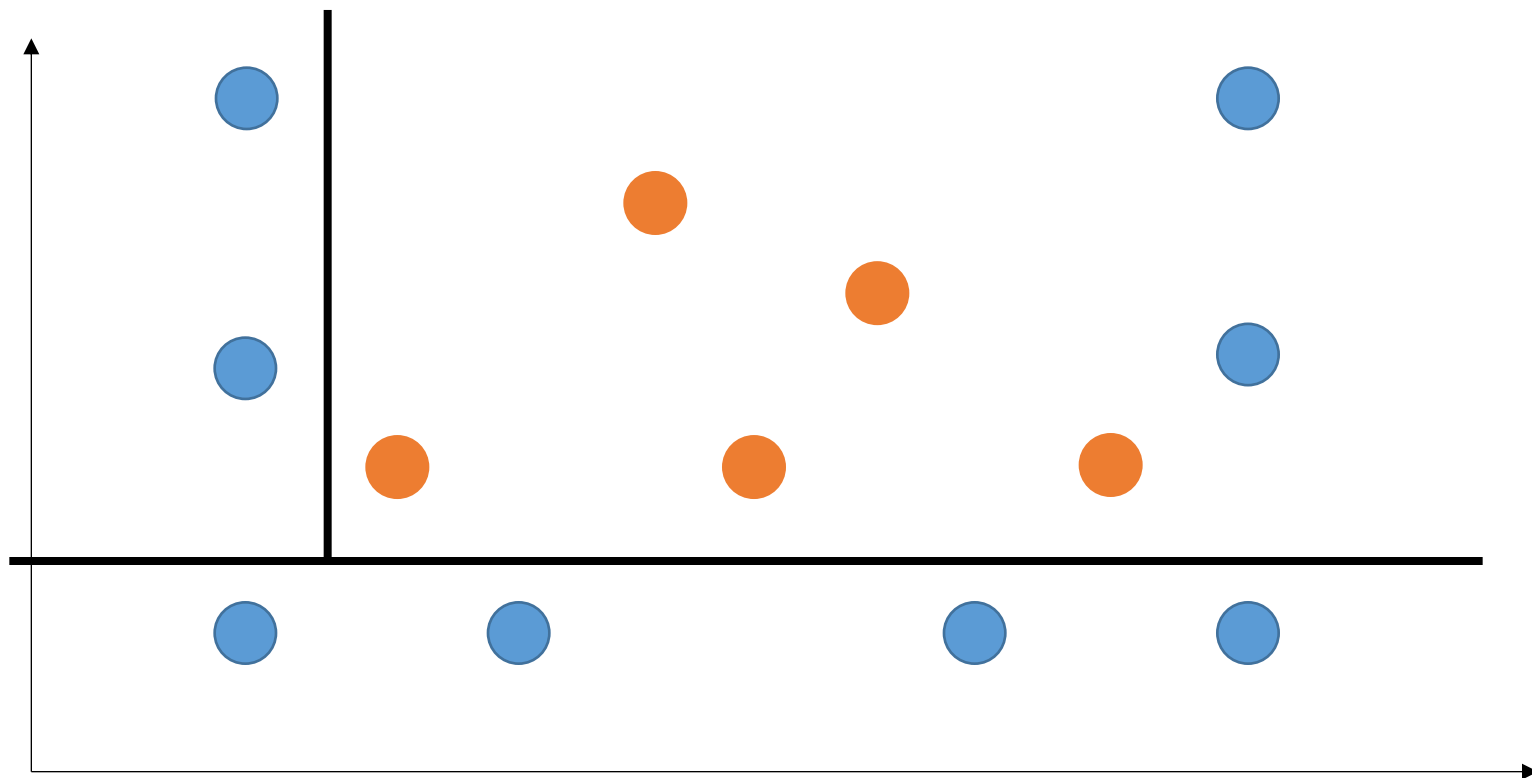
Обучение деревьев



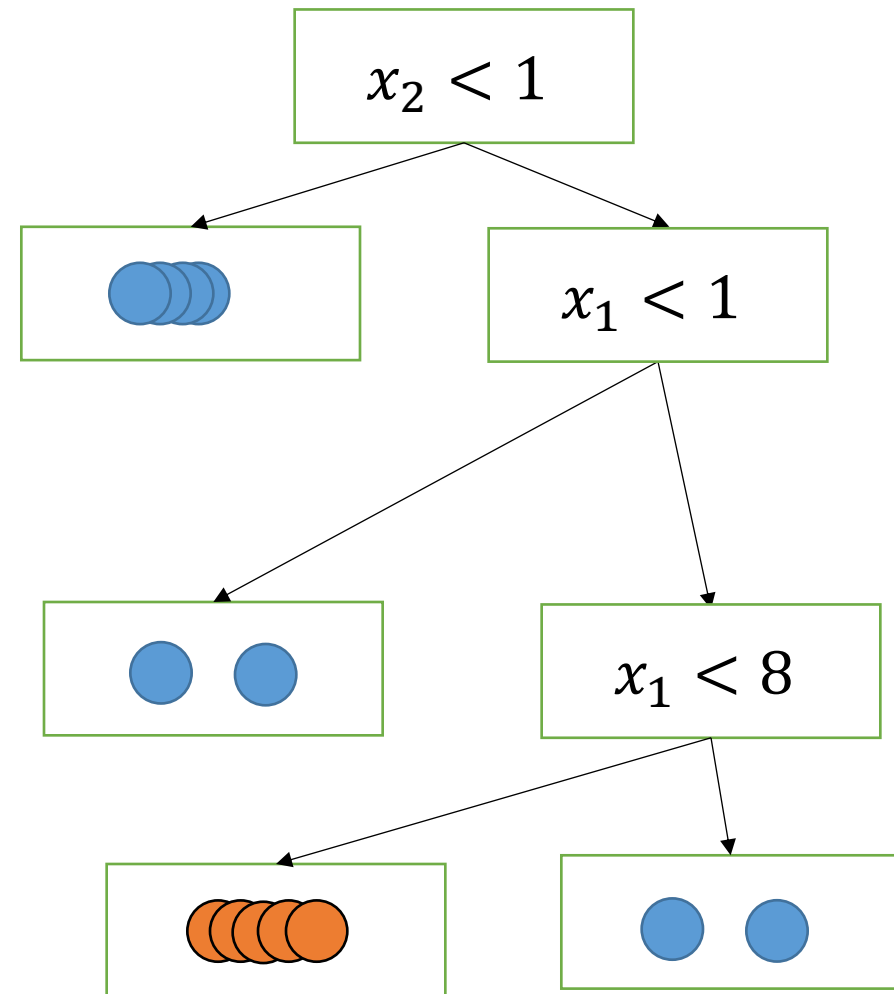
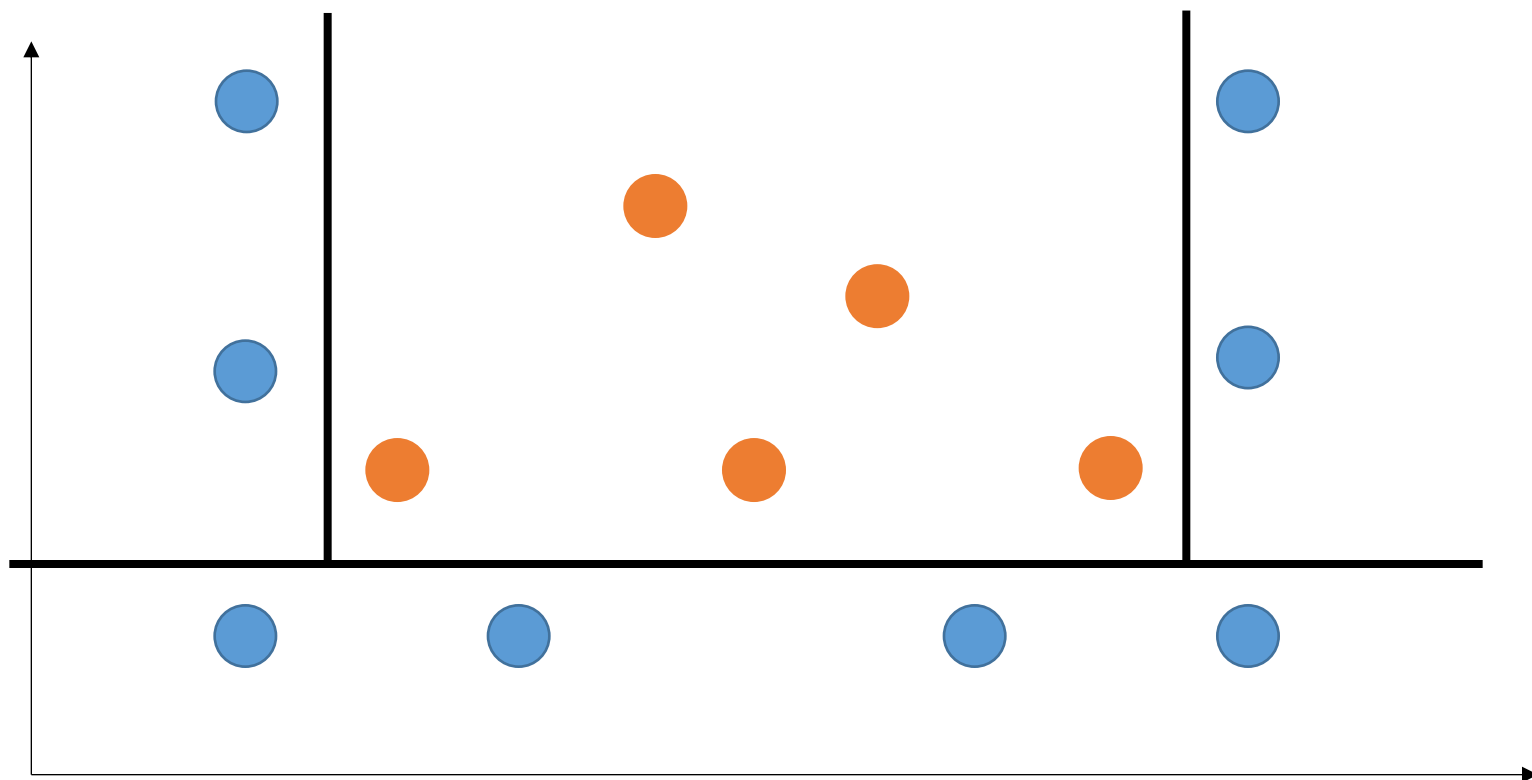
Обучение деревьев



Обучение деревьев



Обучение деревьев



Резюме

- Решающие деревья позволяют строить сложные модели, но есть риск переобучения
- Деревья строятся жадно, на каждом шаге вершина разбивается на две с помощью лучшего из предиктов
- Алгоритм довольно сложный и требует перебора всех предикатов на каждом шаге

Работа с категориальными
признаками

Кодирование категориальных признаков

- Значения признака «район»: $U = \{u_1, \dots, u_m\}$
- Новые признаки вместо x_j : $[x_j = u_1], \dots, [x_j = u_m]$
- One-hot кодирование

Кодирование категориальных признаков

Район		ЦАО	ЮАО	САО
ЦАО		1	0	0
ЮАО		0	1	0
ЦАО		1	0	0
САО		0	0	1
ЮАО		0	1	0

Кодирование категориальных признаков

Район	Цена
ЦАО	10.000.000
ЮАО	4.000.000
ЦАО	9.000.000
САО	7.000.000
ЮАО	5.000.000

Счётчики

- Значения признака x_j : $U_j = \{u_1, \dots, u_m\}$
- Посчитаем все категории в обучающей выборке:

$$\text{count}(j, u_p) = \sum_{i=1}^{\ell} [x_{ij} = u_p]$$

Счётчики

- Значения признака x_j : $U_j = \{u_1, \dots, u_m\}$
- Для регрессии посчитаем суммарный ответ в категории:

$$\text{target}(j, u_p) = \sum_{i=1}^{\ell} [x_{ij} = u_p] y_i$$

Счётчики

- Значения признака x_j : $U_j = \{u_1, \dots, u_m\}$
- Для классификации посчитаем классы в категории:

$$\text{target}_k(j, u_p) = \sum_{i=1}^{\ell} [x_{ij} = u_p] [y_i = k]$$

Счётчики

- Mean-target encoding
- Задача регрессии
- Заменим категориальный признак на числовой:

$$\widetilde{x_{ij}} = \frac{\text{target}(j, x_{ij})}{\text{count}(j, x_{ij})}$$

Счётчики

- Mean-target encoding
- Задача классификации
- Заменим категориальный признак на K числовых:

$$\widetilde{x}_{ij} = \left(\frac{\text{target}_1(j, x_{ij})}{\text{count}(j, x_{ij})}, \dots, \frac{\text{target}_K(j, x_{ij})}{\text{count}(j, x_{ij})} \right)$$

Кодирование категориальных признаков

Район	Счётчик	Цена
ЦАО	9.500.000	10.000.000
ЮАО	4.500.000	4.000.000
ЦАО	9.500.000	9.000.000
САО	7.000.000	7.000.000
ЮАО	4.500.000	5.000.000

Борьба с переобучением в счётчиках

- Проблема в основном с редкими категориями
- Решение 1: добавление шума

Район	Счётчик	Цена
ЦАО	9.130.000	10.000.000
ЮАО	4.023.000	4.000.000
ЦАО	10.124.000	9.000.000
САО	7.942.000	7.000.000
ЮАО	4.728.000	5.000.000

Борьба с переобучением в счётчиках

- Проблема в основном с редкими категориями
- Решение 2: добавление априорных величин в счётчики

$$\widetilde{x_{ij}} = \frac{\text{target}(j, x_{ij}) + a}{\text{count}(j, x_{ij}) + b}$$

Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков

Блок 1

Блок 2

Блок 3

Борьба с переобучением в счётчиках

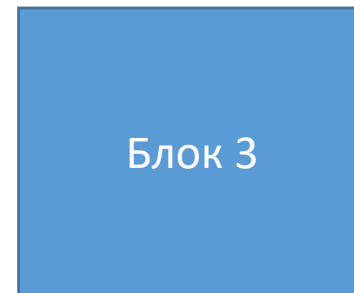
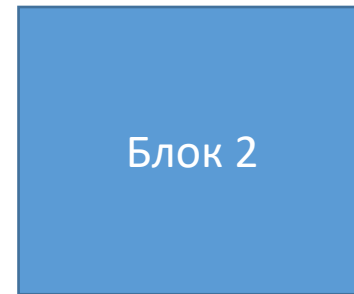
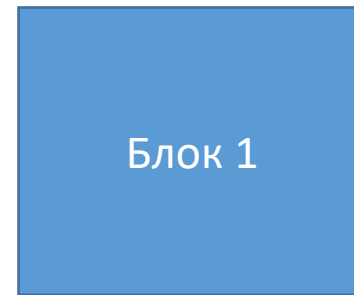
- Решение 3: кросс-валидация счётчиков



Считаем $\text{count}(j, u_p)$ и $\text{target}(j, u_p)$

Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



Считаем $\text{count}(j, u_p)$ и $\text{target}(j, u_p)$

Вычисляем признаки: $\widetilde{x}_{ij} = \frac{\text{target}(j, x_{ij})}{\text{count}(j, x_{ij})}$

Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков

Блок 1

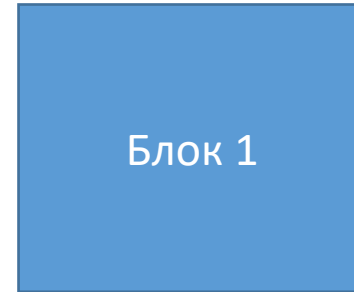
Блок 2

Блок 3

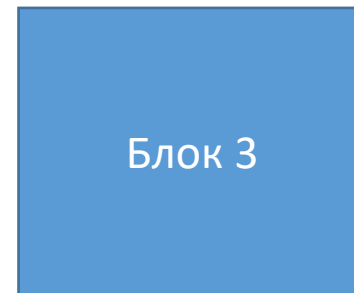
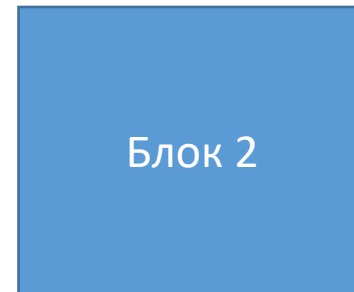
Считаем $\text{count}(j, u_p)$ и $\text{target}(j, u_p)$

Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



Вычисляем признаки: $\widetilde{x}_{ij} = \frac{\text{target}(j, x_{ij})}{\text{count}(j, x_{ij})}$



Считаем $\text{count}(j, u_p)$ и $\text{target}(j, u_p)$

Резюме

- Счётчики позволяют заменить категориальный признак на один числовой
- Могут привести к переобучению
- Можно бороться с ним через добавление шума, априорных значений или кросс-валидацию