

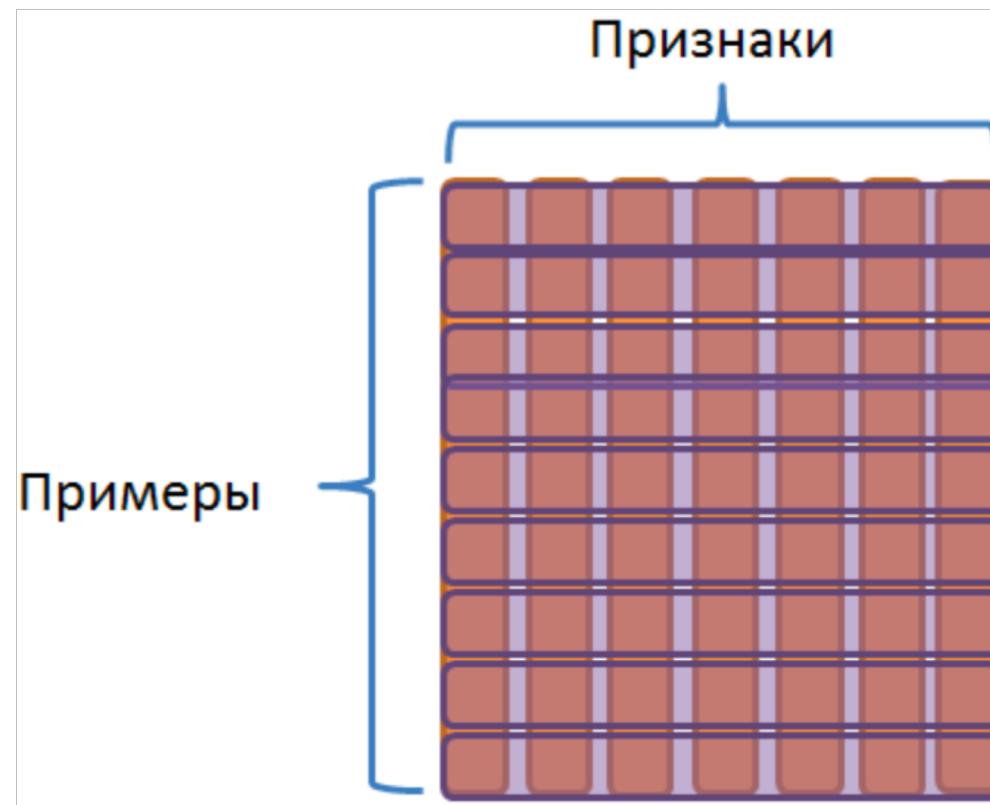


Open Data Science

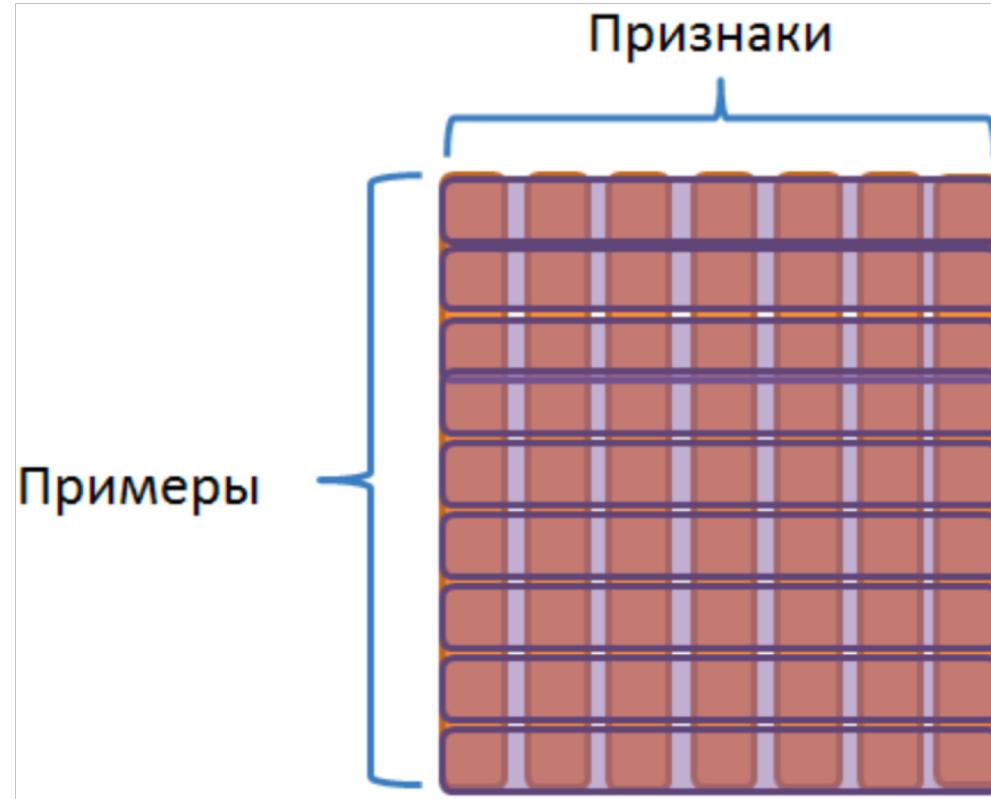


Тема 3.
Задача классификации и деревья решений.

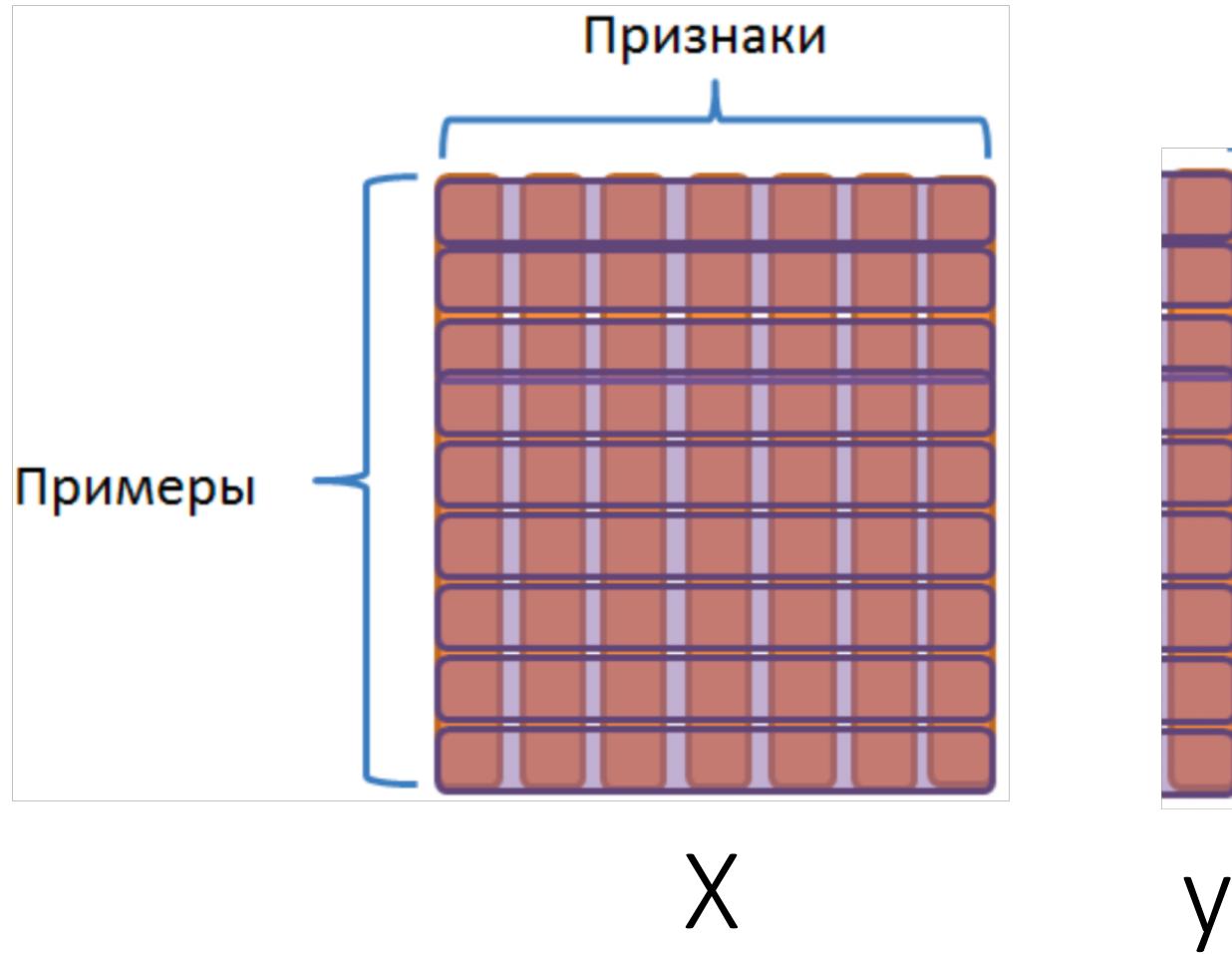
Выборка



Обучение без учителя (Unsupervised)



Обучение с учителем (Supervised)



Классификация



Регрессия



Постановка задачи классификации

Пусть дан набор объектов $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $i \in 1, \dots, N$, полученный из неизвестной закономерности $y = f(\mathbf{x})$. Необходимо построить такую $h(\mathbf{x})$, которая наиболее точно аппроксимирует $f(\mathbf{x})$.

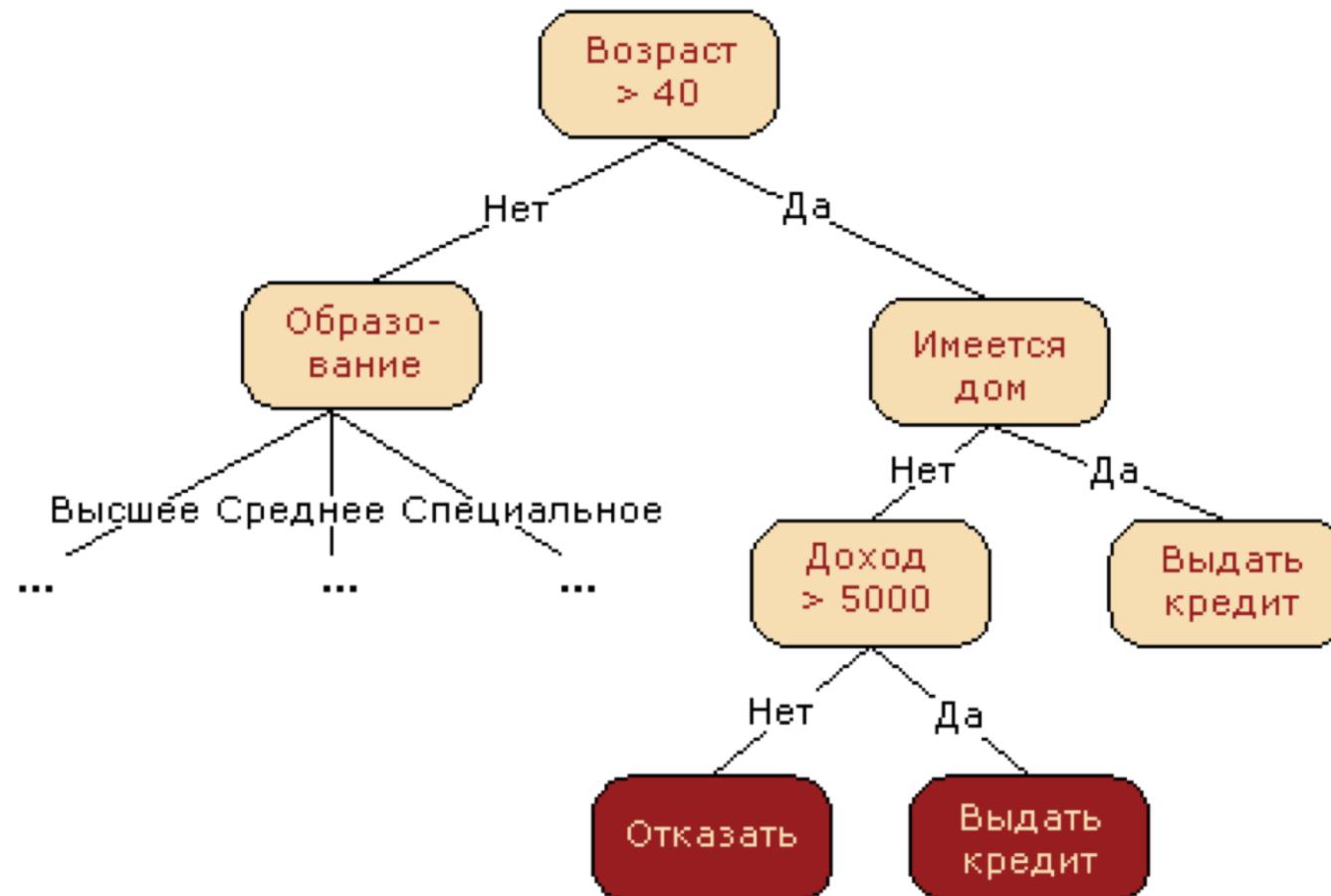
Будем искать неизвестную

$$h(\mathbf{x}) = h(a_1, \dots, a_T)$$

Дерево решений



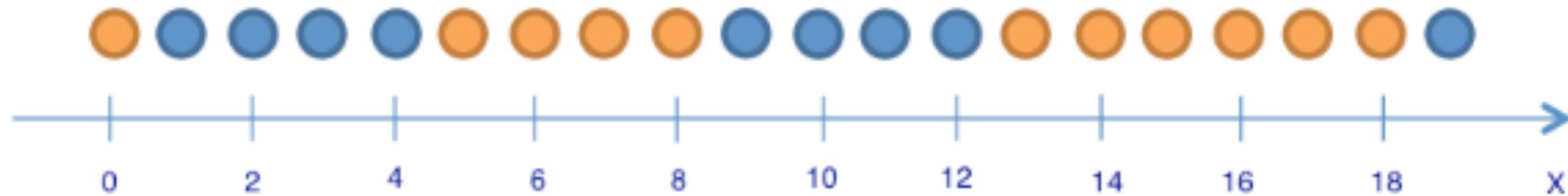
Дерево решений



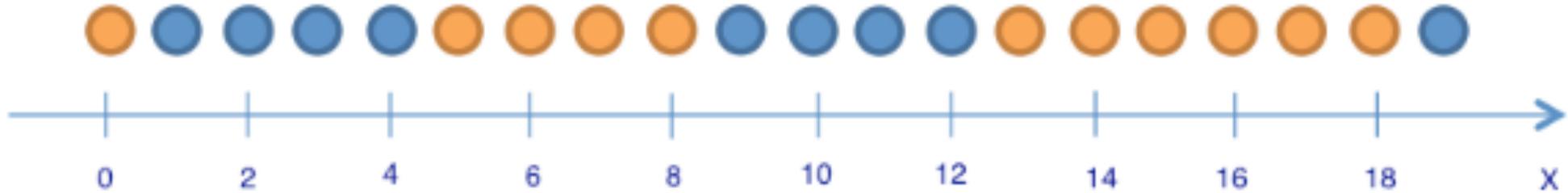
Критерий информативности

- Энтропия Шенона $-\sum_{i=1}^N p_i \log_2(p_i)$
- Индекс Джини $1 - \sum_{i=1}^N p_i^2$

Дерево решений



Дерево решений

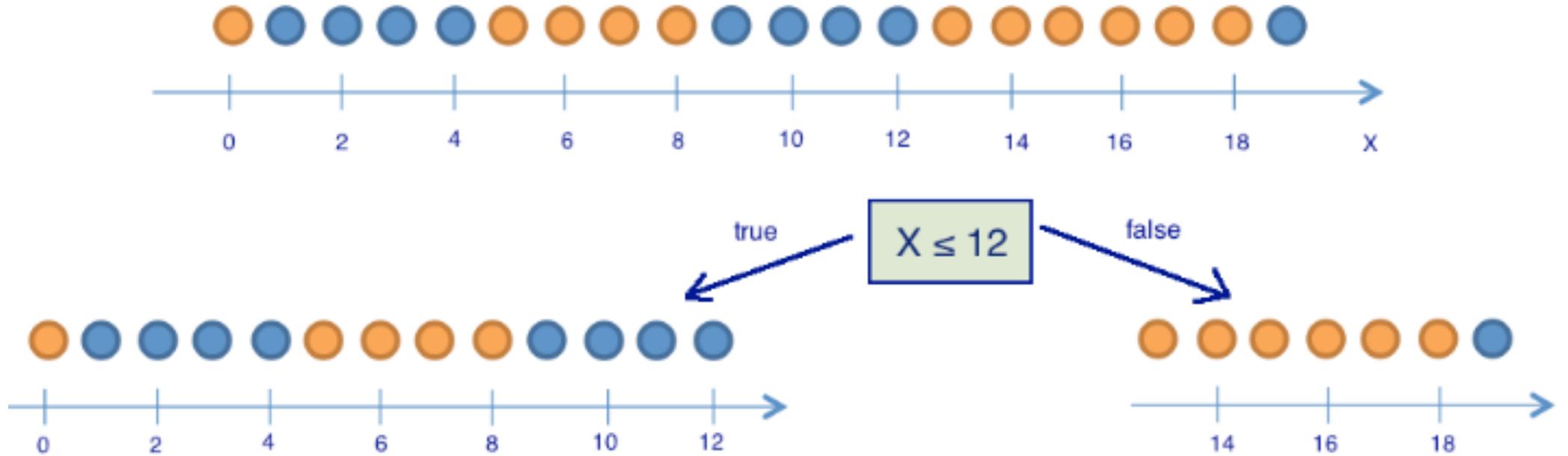


$$\text{Оранжевые: } p_1 = \frac{9}{20}$$

$$\text{Синие: } p_2 = \frac{11}{20}$$

$$\text{Начальная энтропия: } S_0 = -\frac{9}{20} \log_2\left(\frac{9}{20}\right) - \frac{11}{20} \log_2\left(\frac{11}{20}\right) = 0.993$$

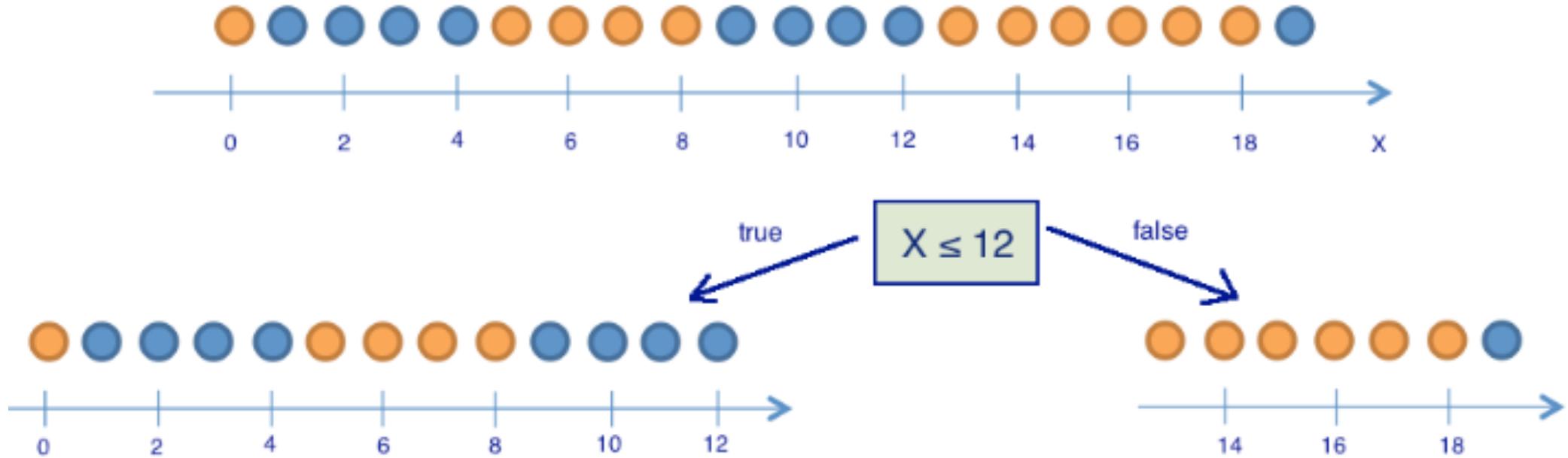
Дерево решений



$$S_1 = -\frac{5}{13} \log_2\left(\frac{5}{13}\right) - \frac{8}{13} \log_2\left(\frac{8}{13}\right) = 0.96$$

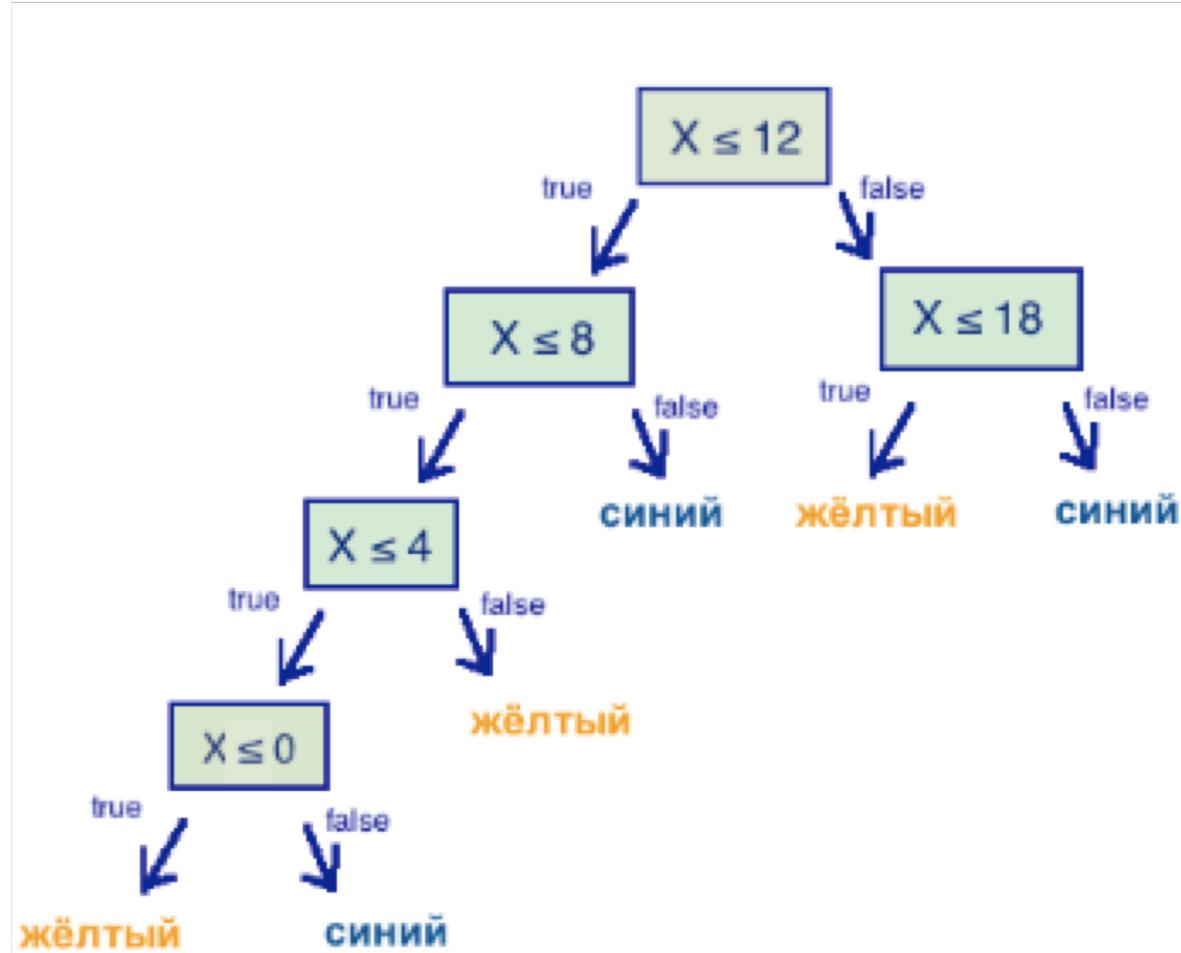
$$S_2 = -\frac{1}{7} \log_2\left(\frac{1}{7}\right) - \frac{6}{7} \log_2\left(\frac{6}{7}\right) = 0.59$$

Дерево решений



$$\text{IG}("X \leq 12") = S_0 - \frac{13}{20}S_1 - \frac{7}{20}S_2 = 0.163$$

Дерево решений



Дерево решений

```
def build(L):
    create node t
    if the stopping criterion is True:
        assign a predictive model to t
    else:
        Find the best binary split L = L_left + L_right
        t.left = build(L_left)
        t.right = build(L_right)
    return t
```

Дерево решений

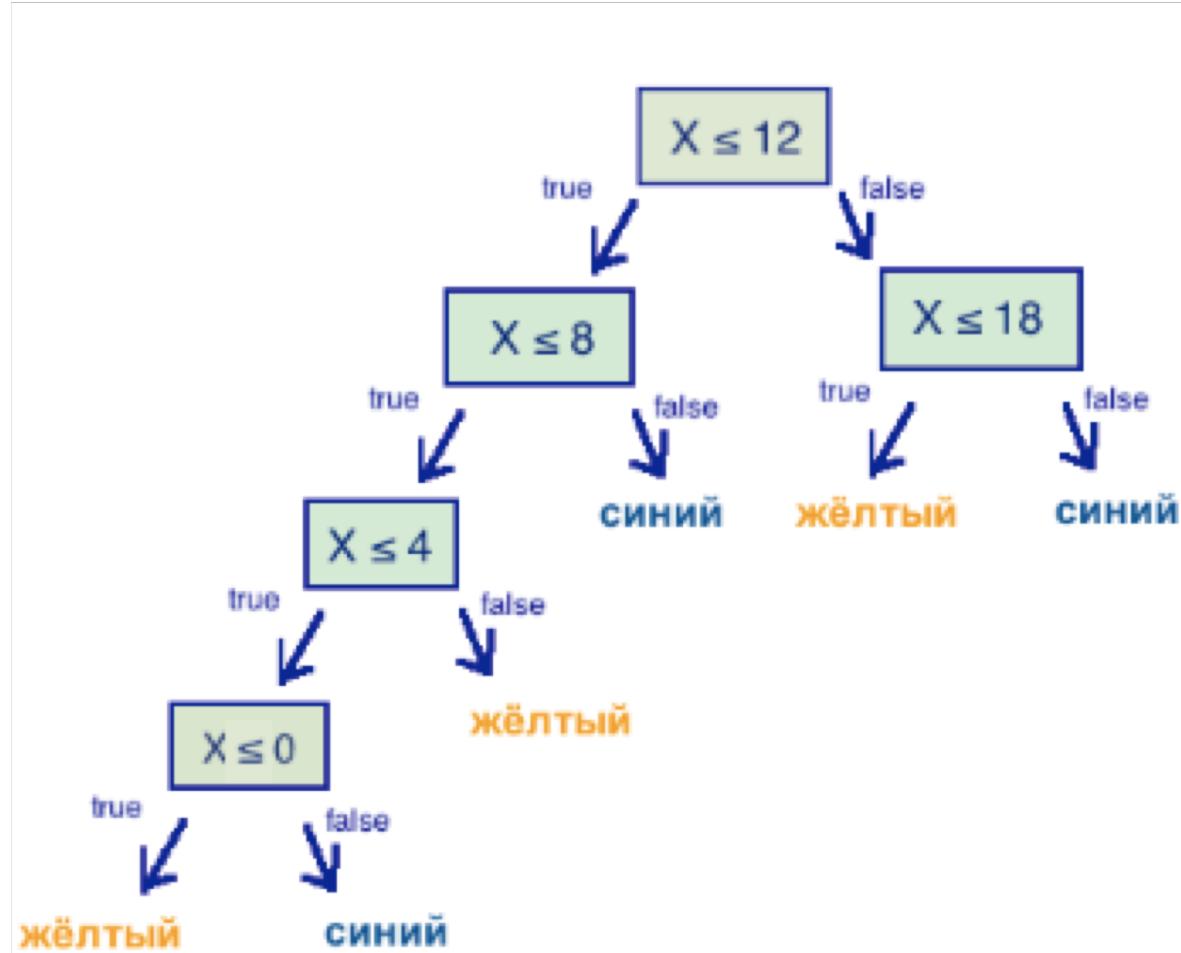
Плюсы

- Интерпретируемость
- Не требует сложной подготовки данных
- Можно оценить модель с помощью статистических тестов

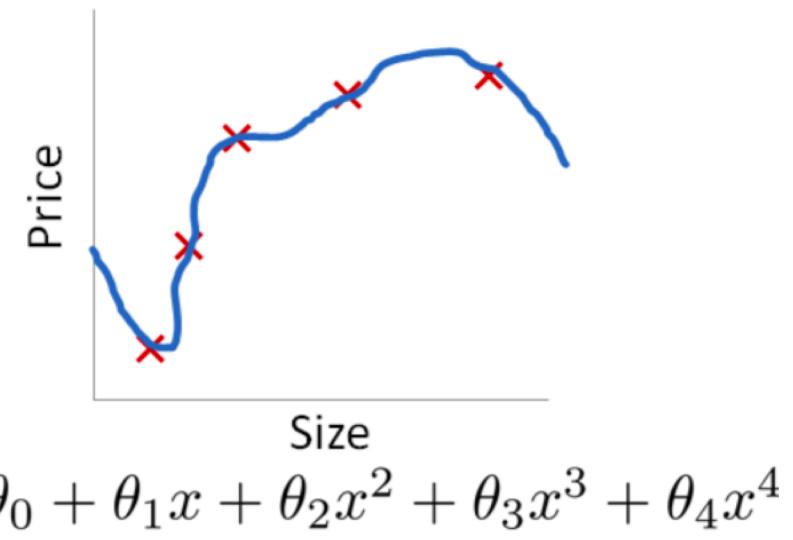
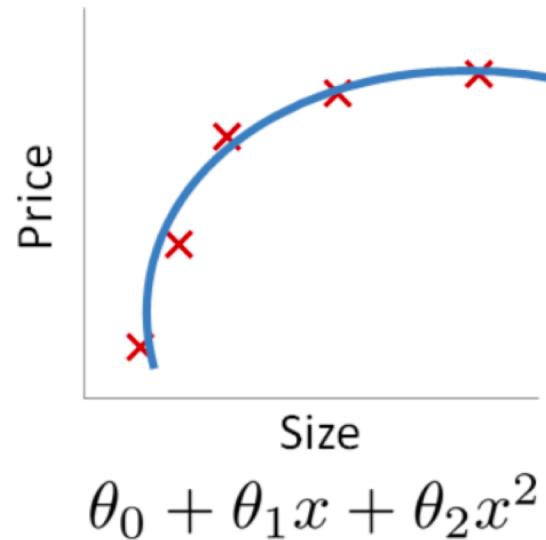
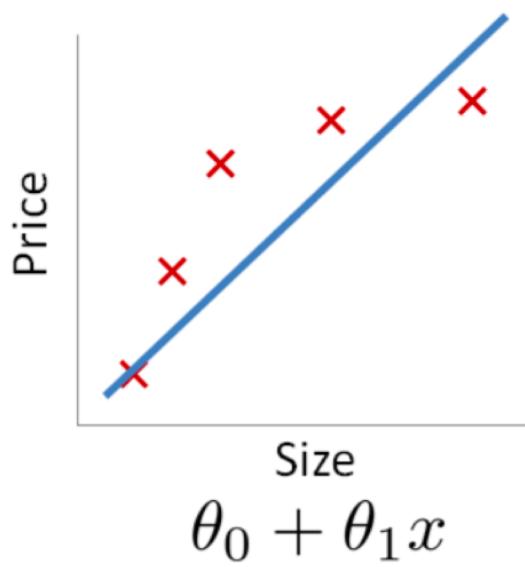
Минусы

- Алгоритм построен на эвристиках
- Слабый алгоритм

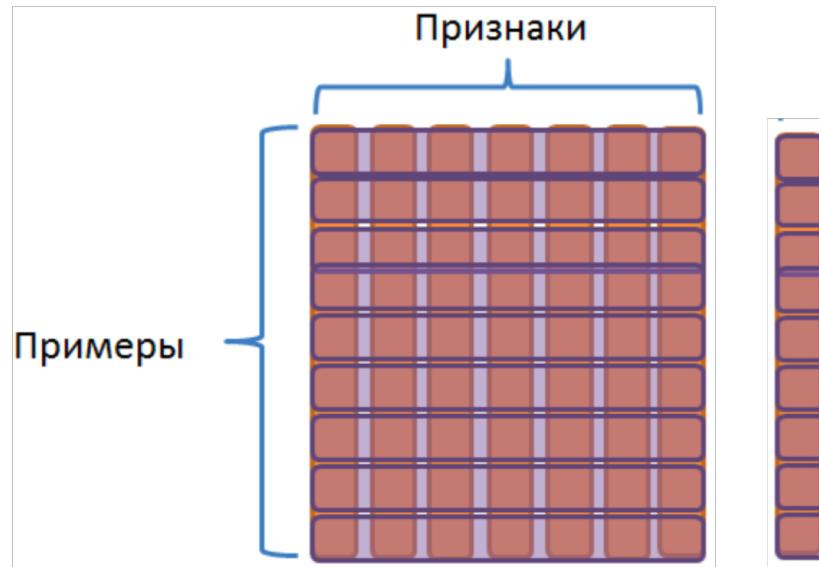
Дерево решений



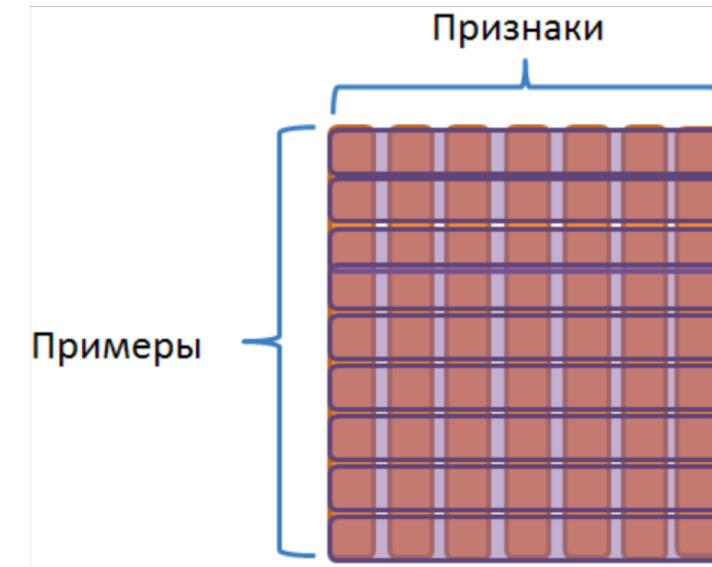
Переобучение



Обучающая и тестовая выборки

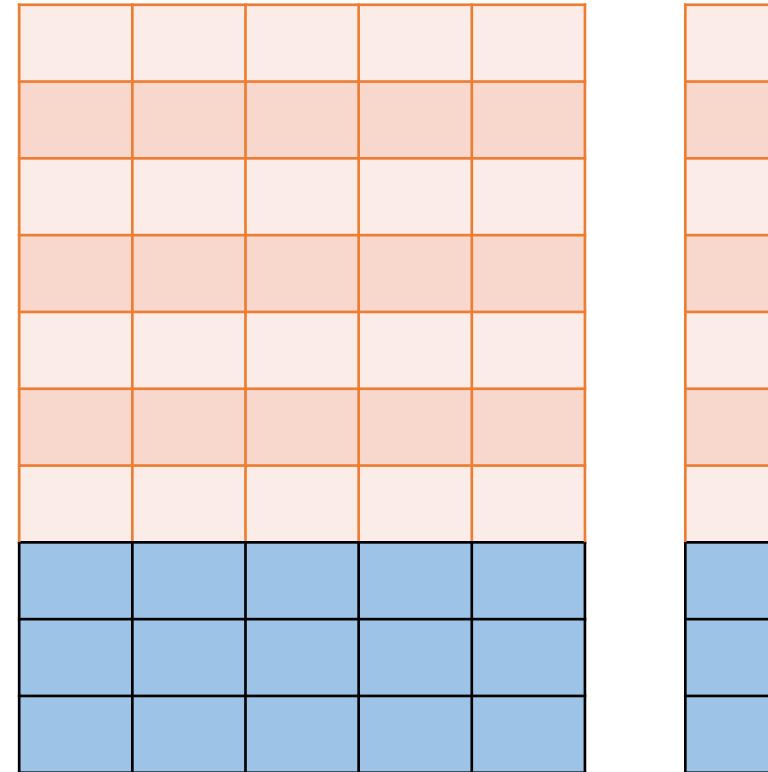


Обучающая выборка



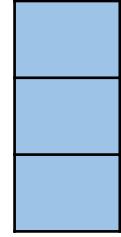
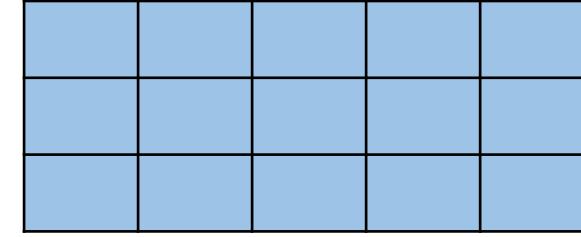
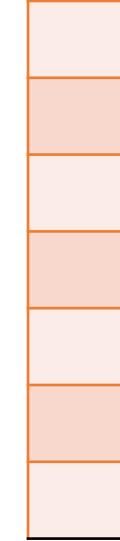
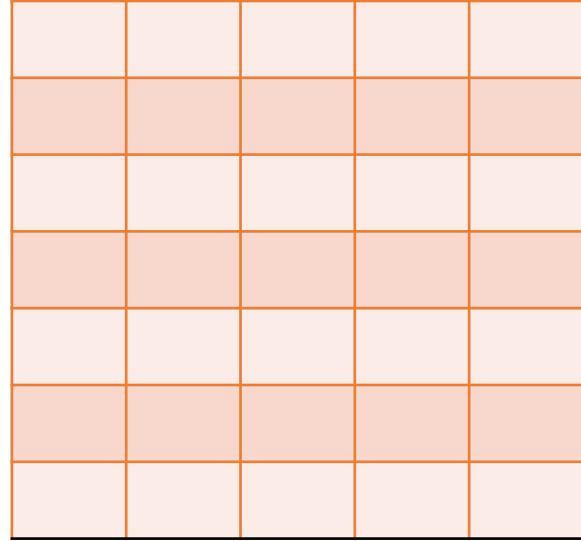
Тестовая выборка

Обучающая и тестовая выборки



Обучающая выборка

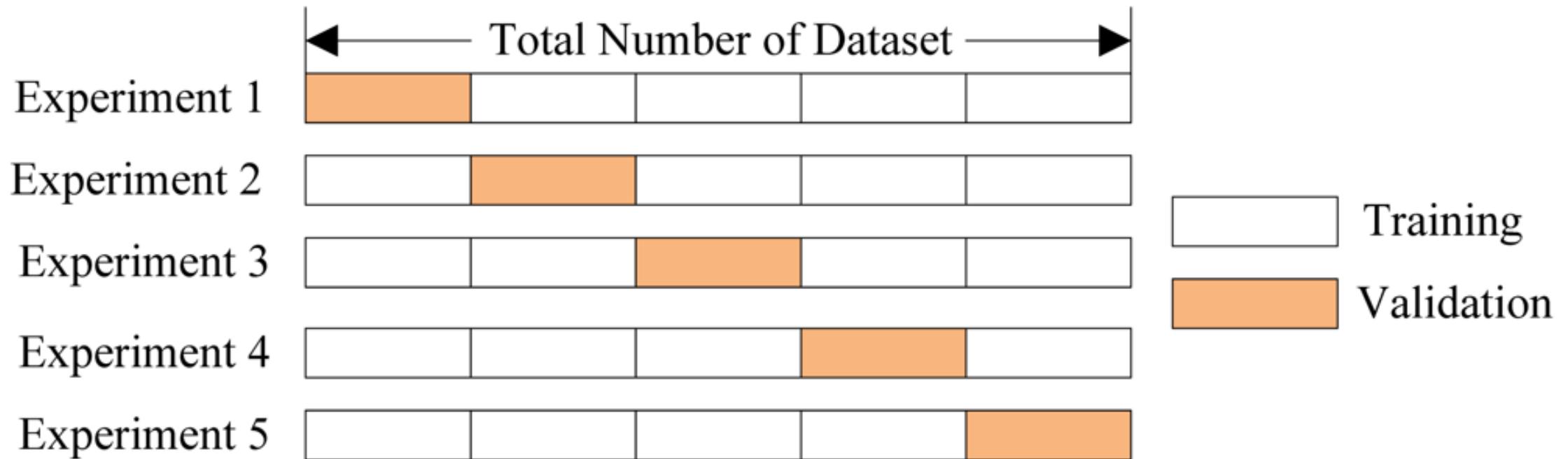
Обучающая и тестовая выборки



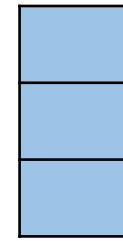
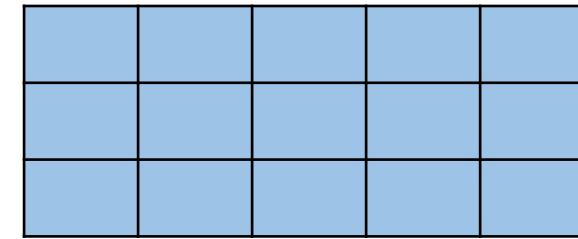
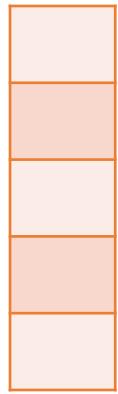
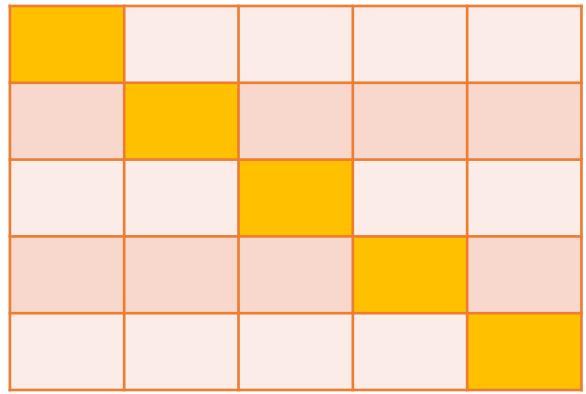
Обучающая выборка
(X_{train} и y_{train})

Тестовая выборка
(X_{test} и y_{test})

Cross-validation



Финальный Pipeline



Кросс-валидация на
обучающей выборке

+

Результат работы на
отложенной выборке

Метрики

Accuracy

Доля правильных ответов алгоритма.

Плюсы

- Интуитивно понятная метрика
- Легко считается

Минусы

- Совершенно бесполезна в задачах с неравными классами

Матрица ошибок

	True	False
Predict True	True Positive (TP)	False Positive (FP)
Predict False	False Negative (FN)	True Negative (TN)

Метрики

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Метрики

Precision

Доля ответов алгоритма, которую алгоритм назвал положительными и которые на самом деле положительные

Recall

Какую долю положительных ответов смог правильно классифицировать наш алгоритм

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Метрики

F-мера

Одна метрика, которая отвечает за улучшение одновременно Precision и Recall.

$$F_{\beta} = (1 + \beta^2) \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

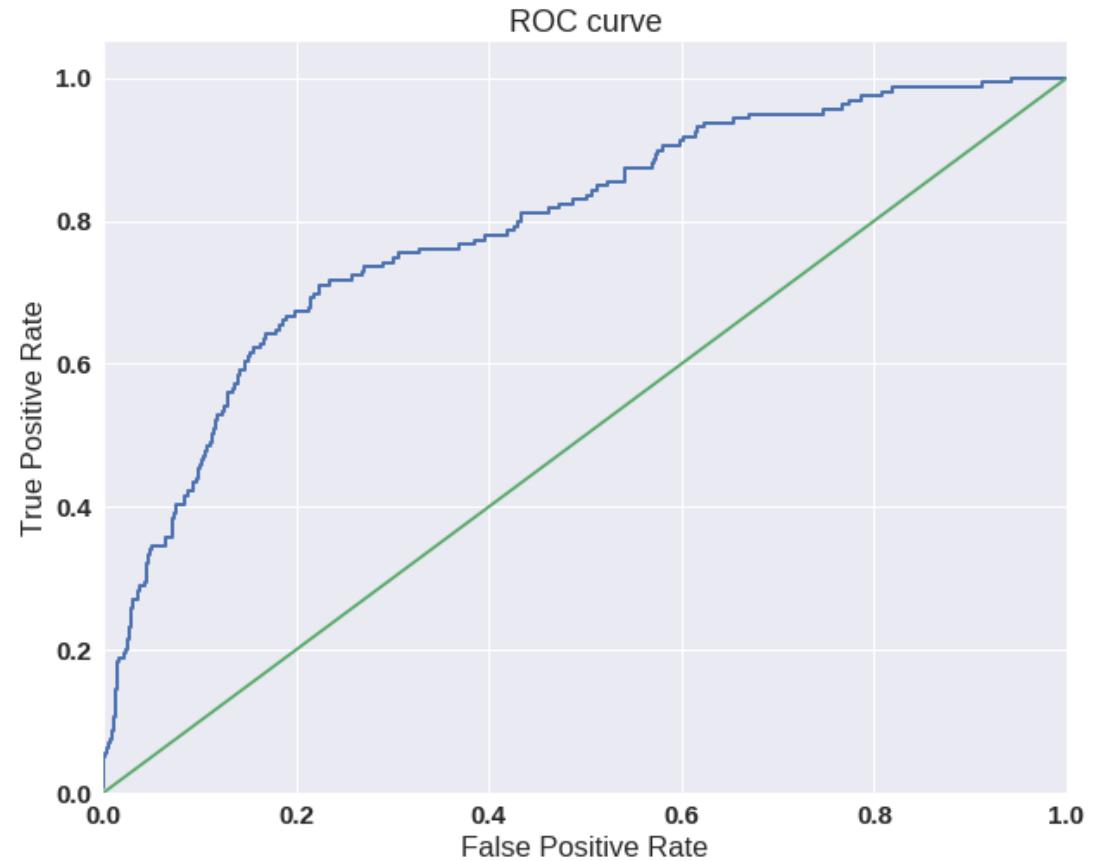
Метрики

ROC-AUC

Площадь под кривой ошибок.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$



Кривая ошибок в координатах (TPR и FPR)

Дерево решений

Плюсы

- Интерпретируемость
- Отлично подходят как базовый алгоритм для ансамблей моделей
- Мало чувствителен к выбросам
- Скорость работы

Минусы

- Очень склонен к переобучению
- Легко изменяет всю структуру дерева от небольшого изменения данных