

Машинное обучение

Лекция 1

Метрики качества, валидация, Деревья решений

Власов Кирилл Вячеславович



2020

Формальная постановка задачи

Дана обучающая выборка (объекты независимы):

$$X_m = \{ (x_1, y_1), \dots, (x_m, y_m) \}$$

Для задачи регрессии - Целевая переменная задана вещественным числом

$$(x_1, y_1) \in \mathbb{R}^m \times \mathbb{Y}, \mathbb{Y} = \mathbb{R}$$

Для задачи классификации - Целевая переменная задана конечным числом меток

$$(x_1, y_1) \in \mathbb{R}^m \times \mathbb{Y}, \mathbb{Y} = \{-1; 1\}$$

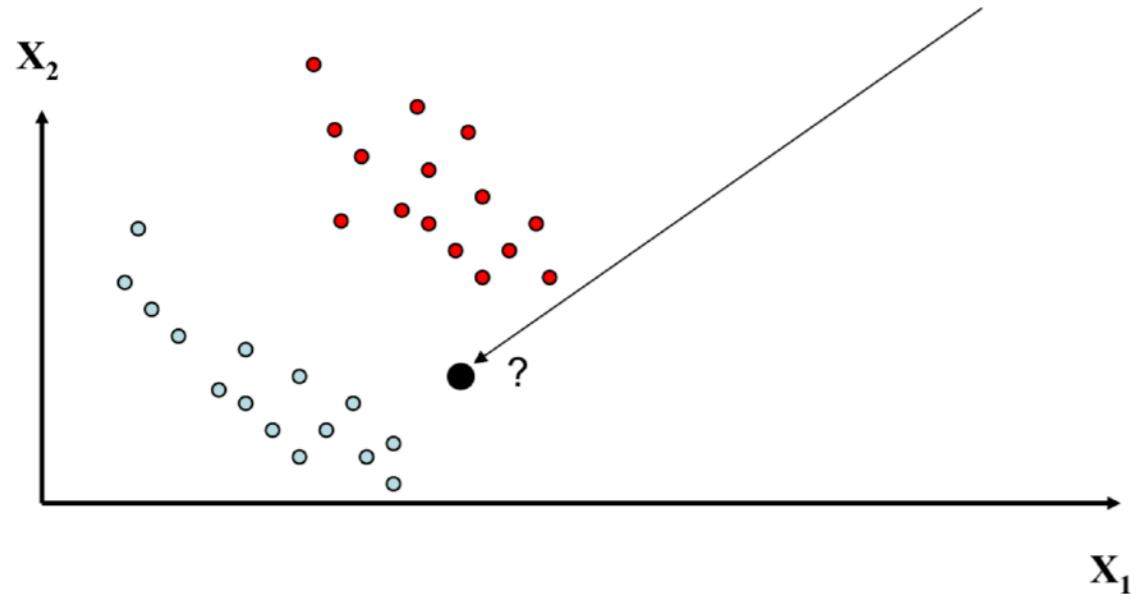
Задать такую функцию $f(x)$ от вектора признаков x , которое выдает ответ для любого возможного наблюдения x

$$f(x): \mathbb{X} \rightarrow \mathbb{Y}$$

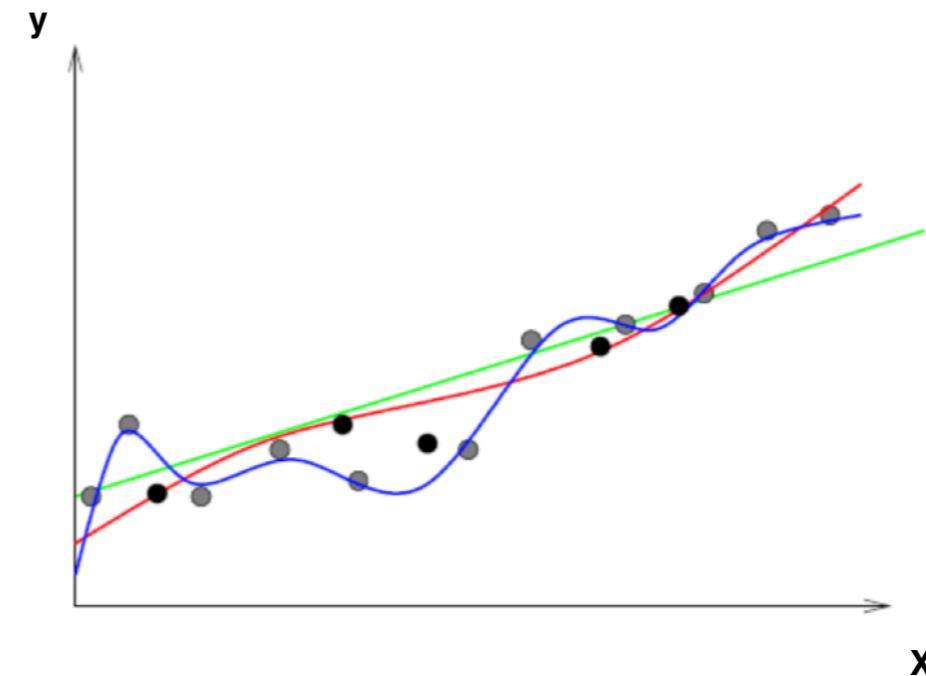
Основная гипотеза МО: Схожим объектам соответствуют схожие объекты

Формальная постановка задачи

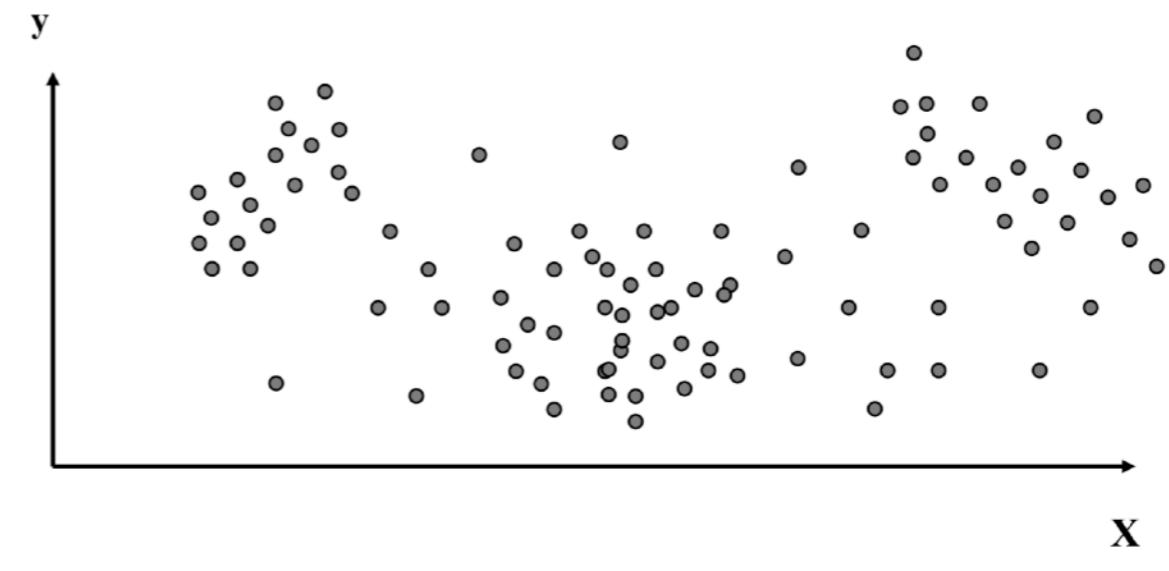
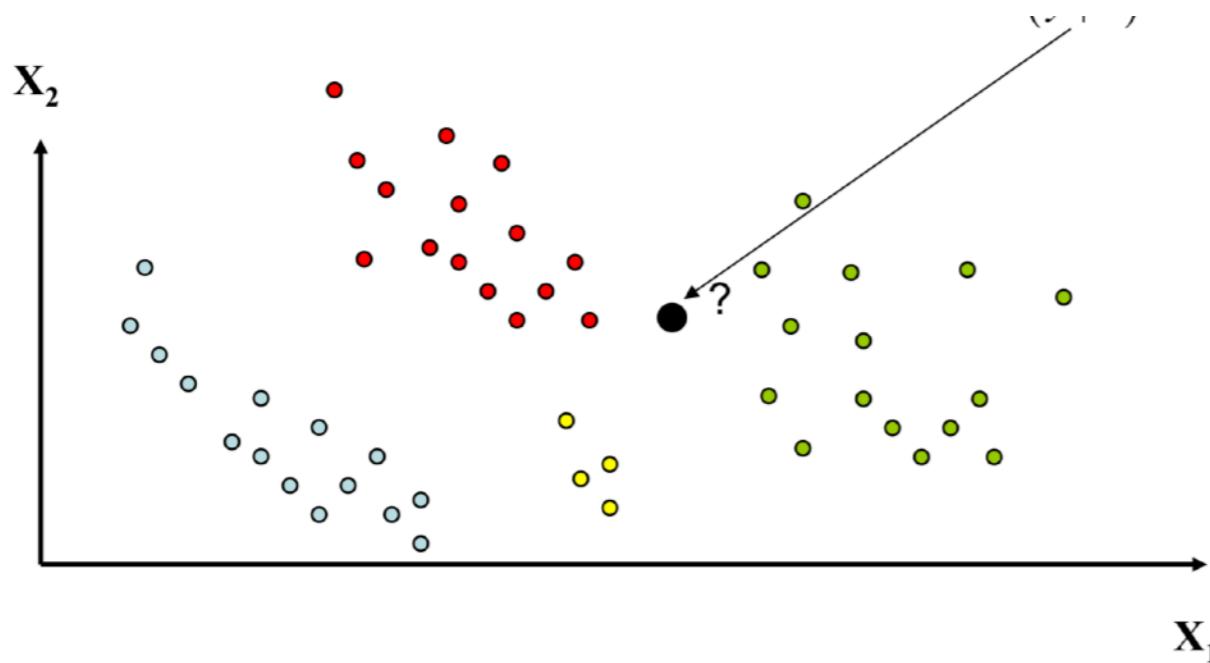
Классификация



Восстановление регрессии

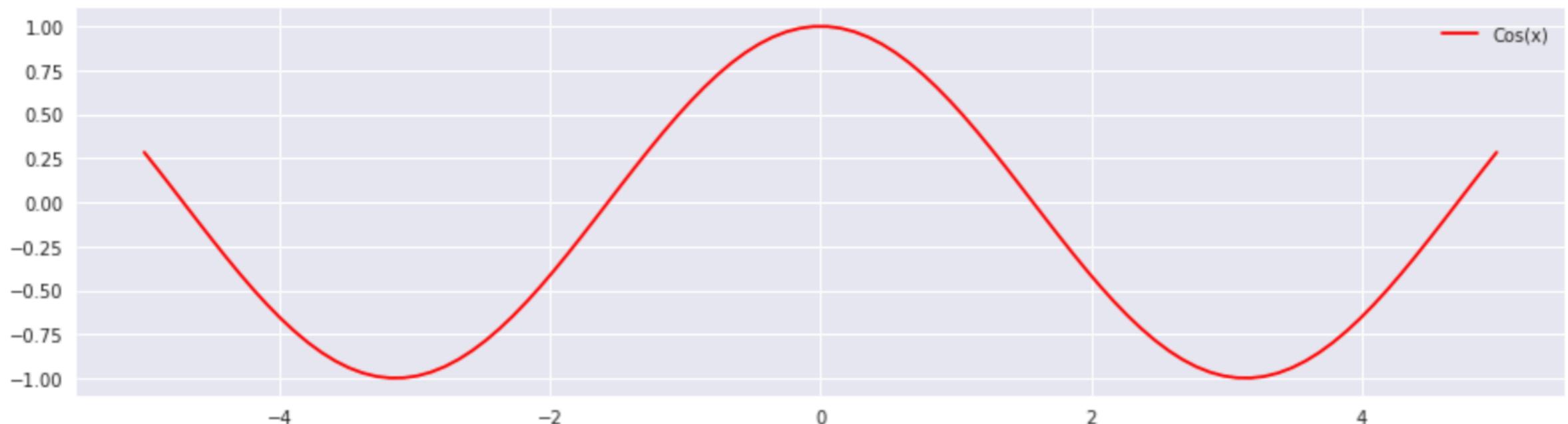


Кластеризация



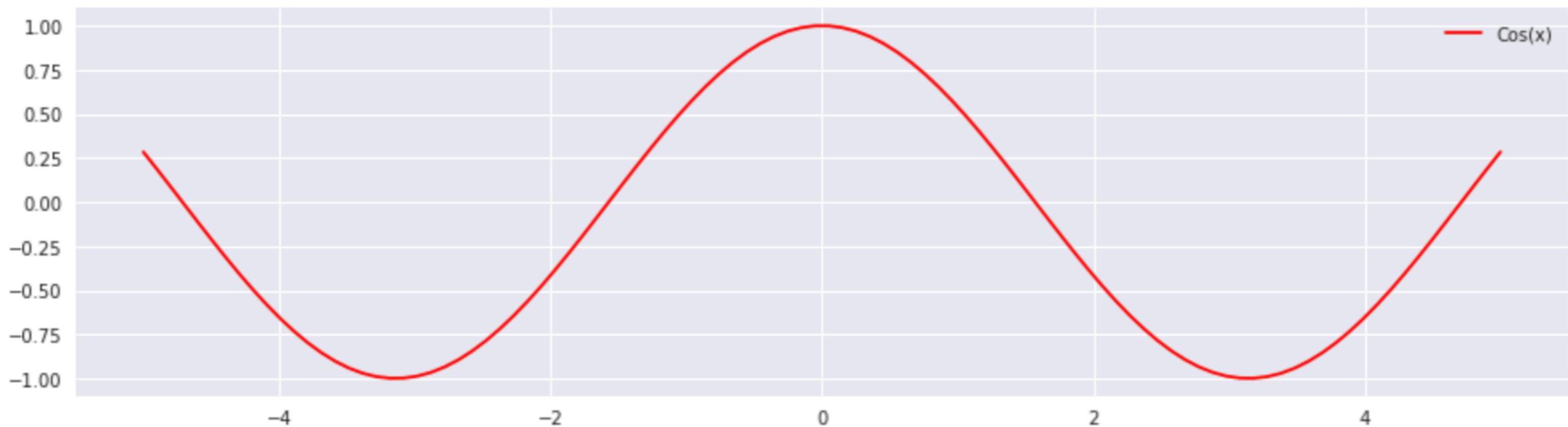
Метрики качества

$$y = \cos(x), x \in [-5, 5]$$

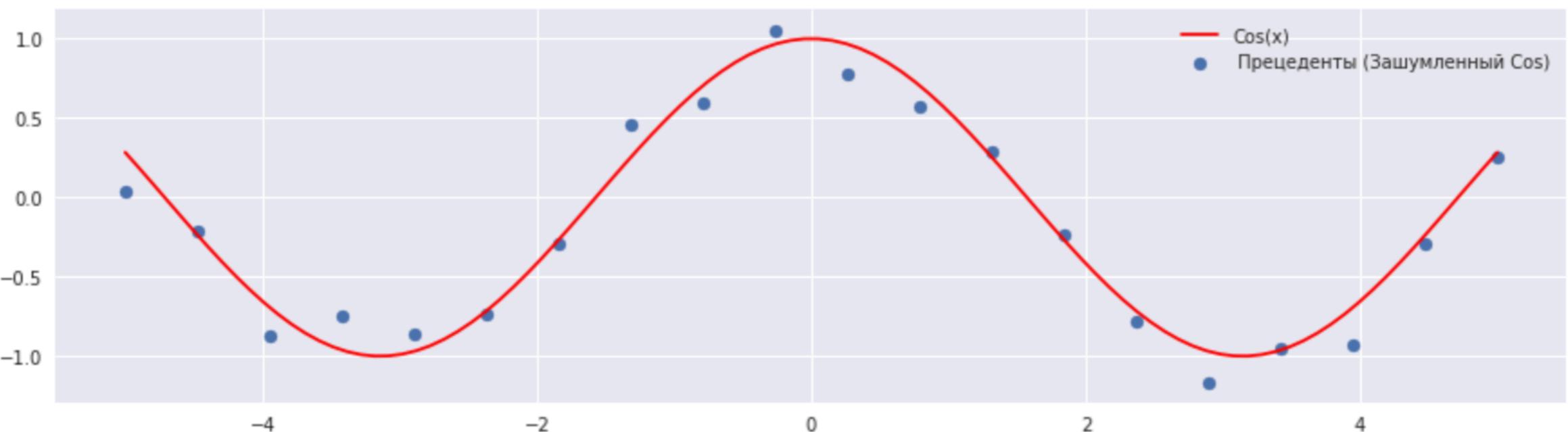


Метрики качества

$$y = \cos(x), x \in [-5, 5]$$

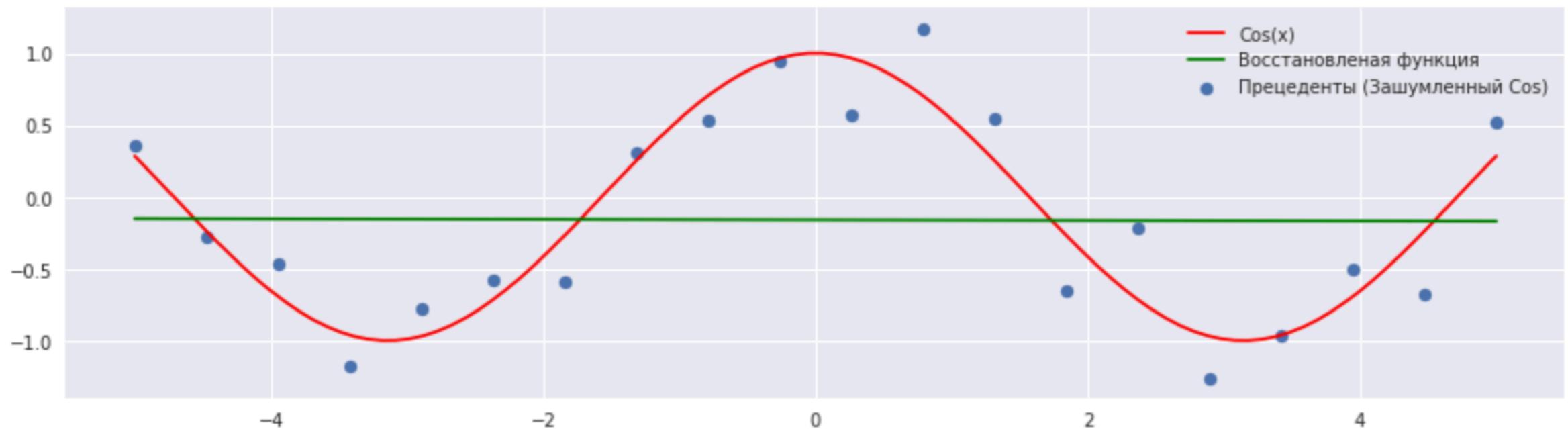


$$y = \cos(x) + \varepsilon, \text{ где } \varepsilon = \mathcal{N}\left(0, \frac{1}{2}\right), x \in [-5, 5]$$



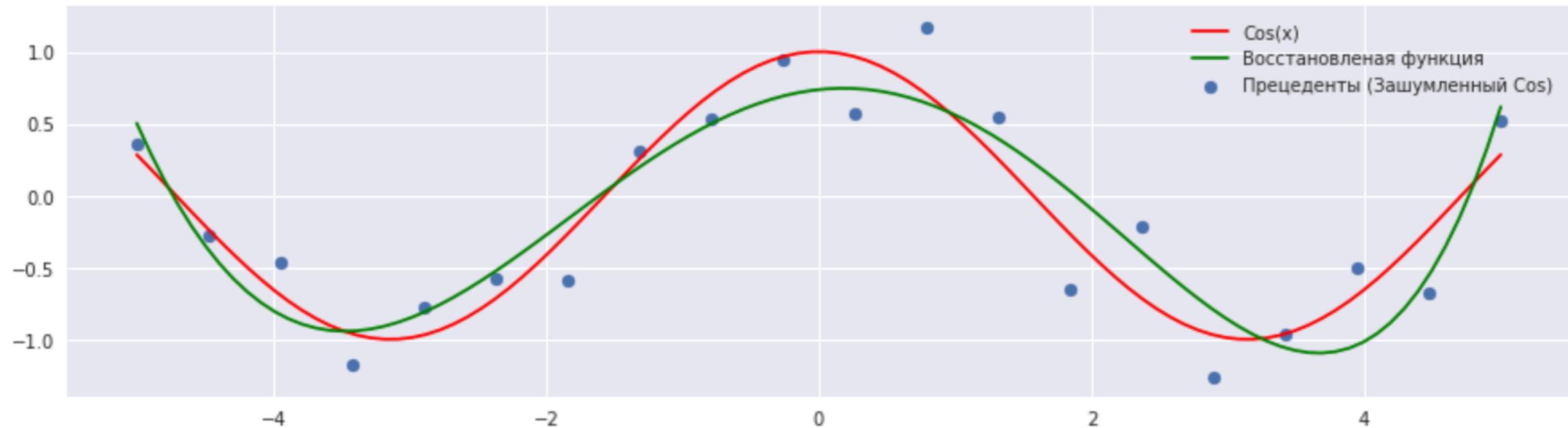
Метрики качества

Восстановим зависимость линейной функцией



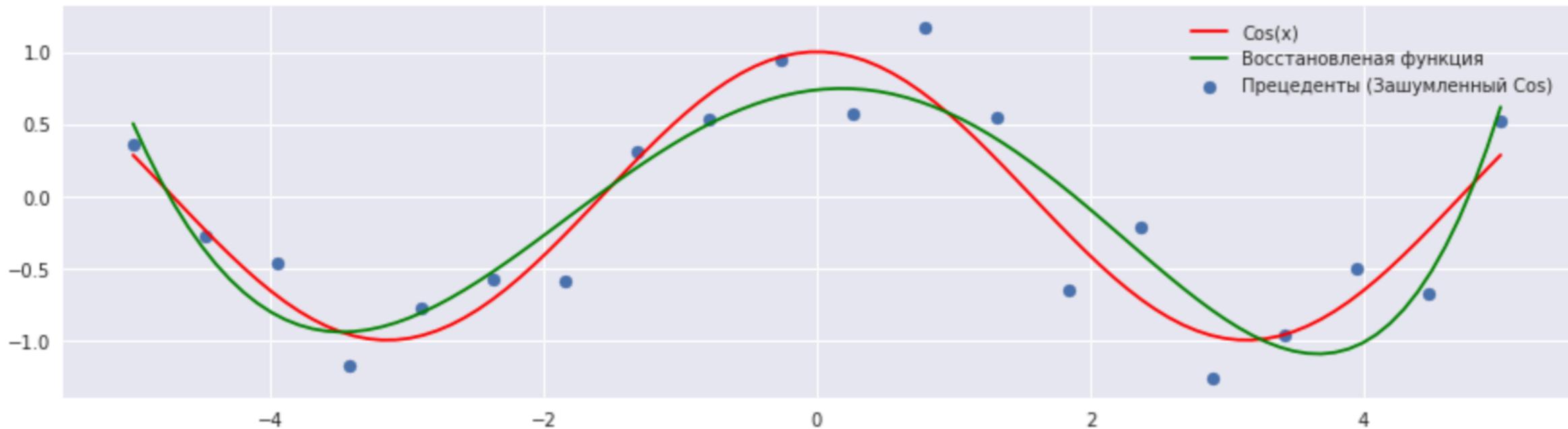
Метрики качества

Восстановим зависимость с помощью полинома 5-ого порядка

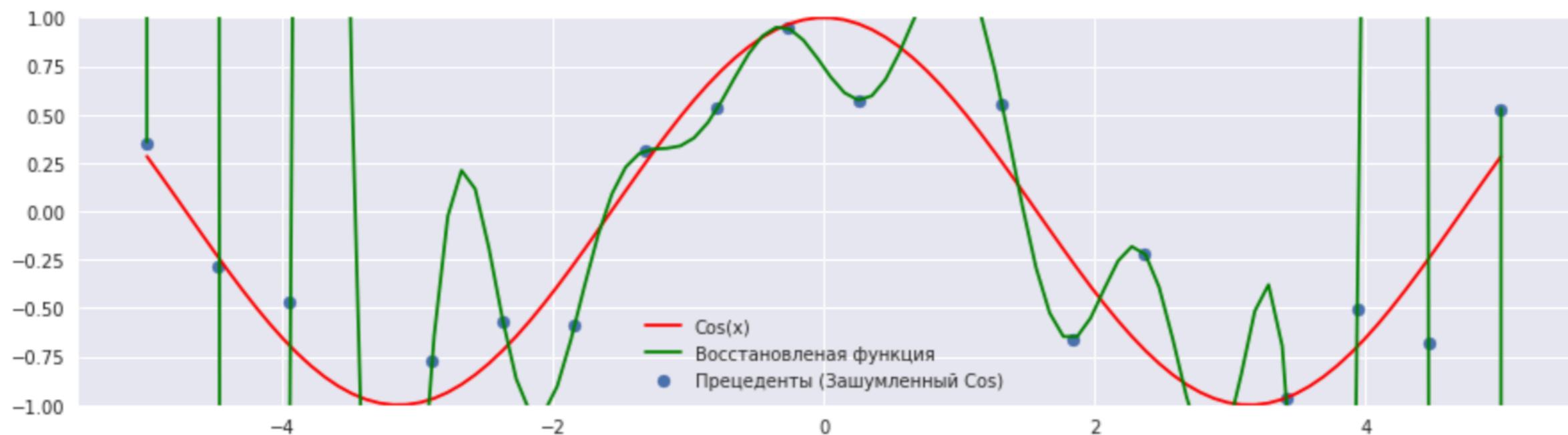


Метрики качества

Восстановим зависимость с помощью полинома 5-ого порядка



Восстановим зависимость с помощью полинома 11-ого порядка



Метрики качества в задачах регрессии

Средняя квадратичная (*Mean Squared Error, MSE*) ошибка:

$$MSE = \frac{1}{l} \sum_{i=1}^l (f(x_i) - y_i)^2$$

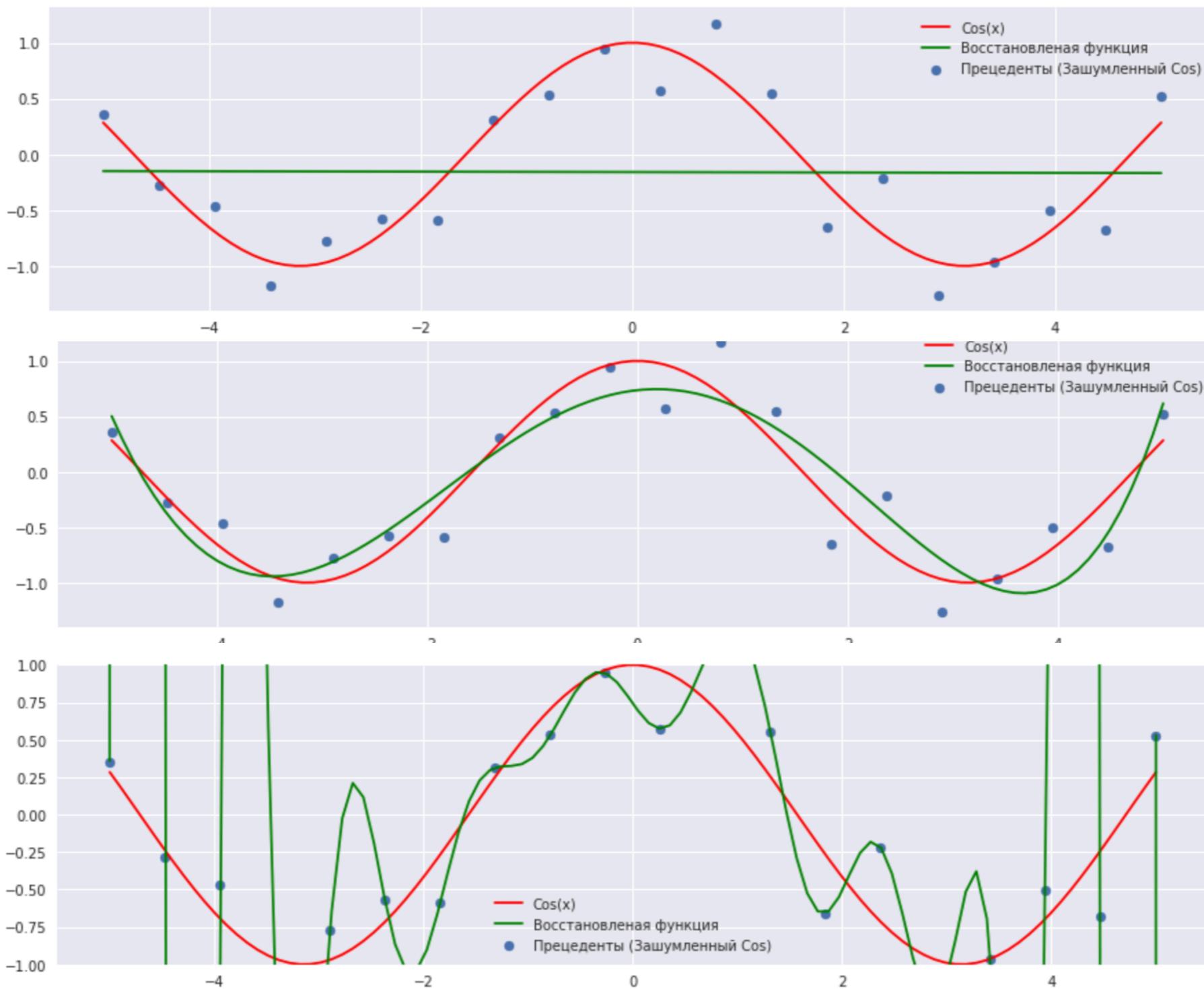
Средняя абсолютная (*Mean Absolute Error, MAE*) ошибка:

$$MAE = \frac{1}{l} \sum_{i=1}^l |f(x_i) - y_i|$$

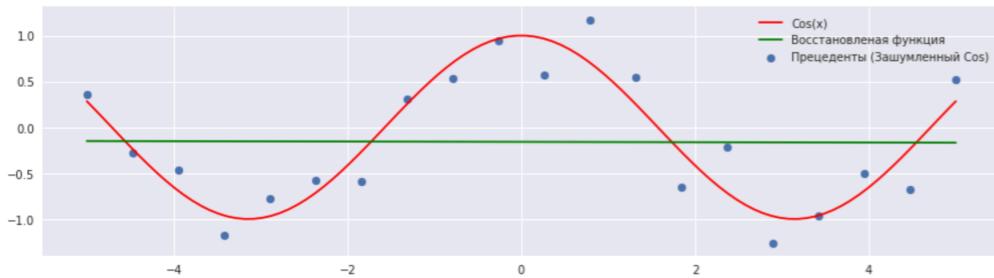
Среднеквадратичный функционал сильнее штрафует за большие отклонения по сравнению со среднеабсолютным, и поэтому более чувствителен к выбросам.

Идеальны для сравнения моделей, но не всегда понятно как их оценивать относительно целевой переменной. Например: MSE - 10 это хорошо, если переменная лежит в пределах интервала (10000, 100000), но плохо, если целевая переменная принимает значения от 0 до 1

Мотивация валидации

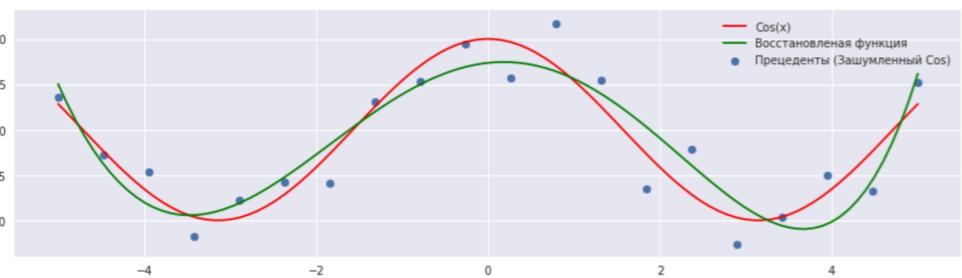


Мотивация валидации

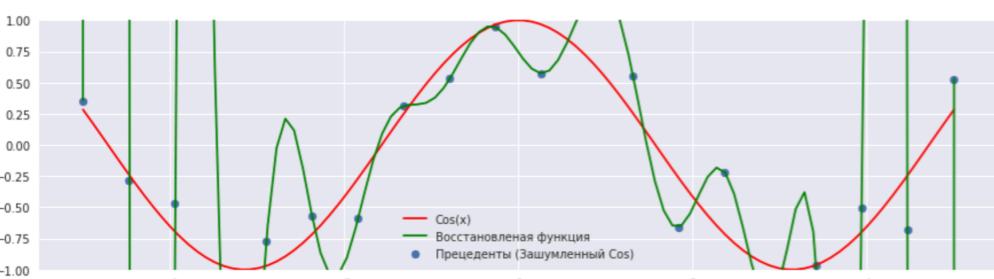


MSE	MAE	R2
-----	-----	----

Линейная модель	0.472	0.586	0.0004
-----------------	-------	-------	--------

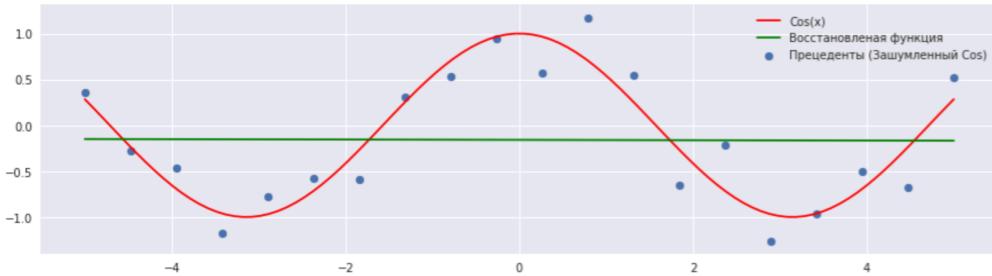


Полином 5-ой степени	0.047	0.179	0.9000
----------------------	-------	-------	--------



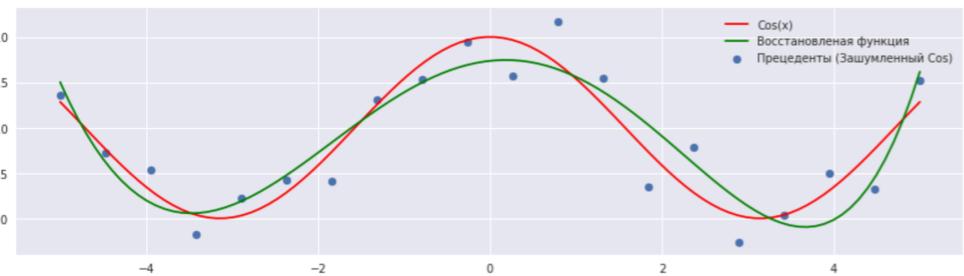
Полином 11-ой степени	0.000	0.000	1.0000
-----------------------	-------	-------	--------

Мотивация валидации



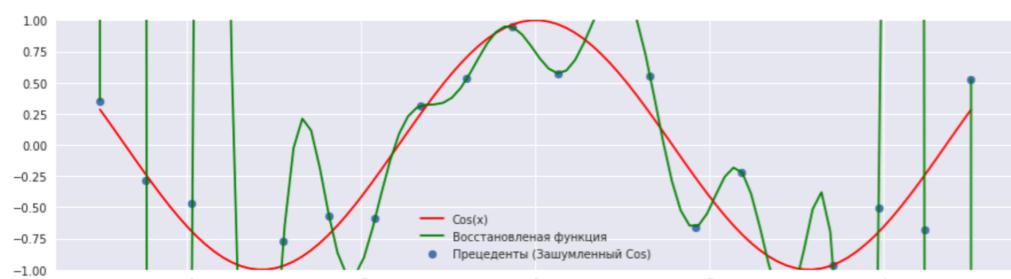
MSE	MAE	R2
-----	-----	----

Линейная модель	0.472	0.586	0.0004
-----------------	-------	-------	--------



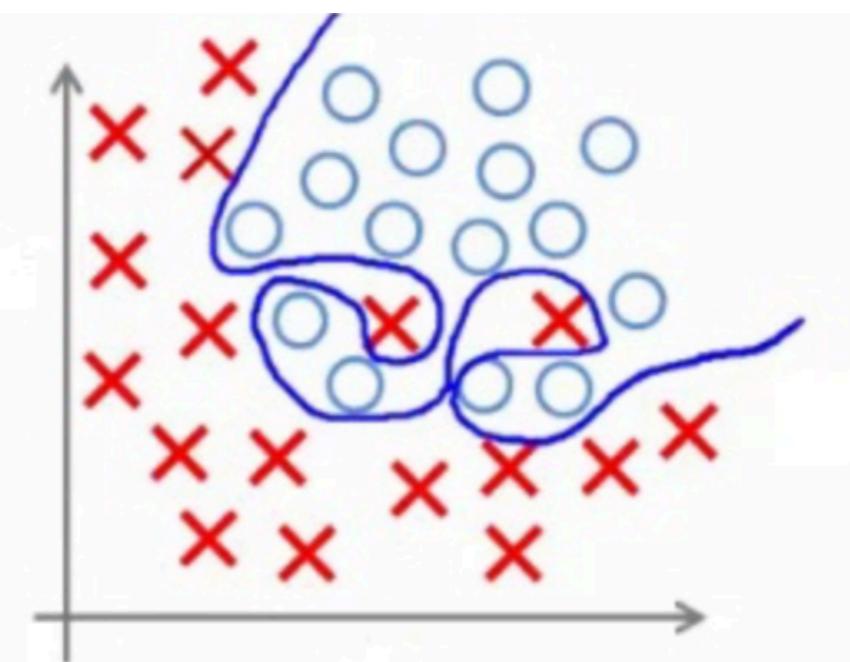
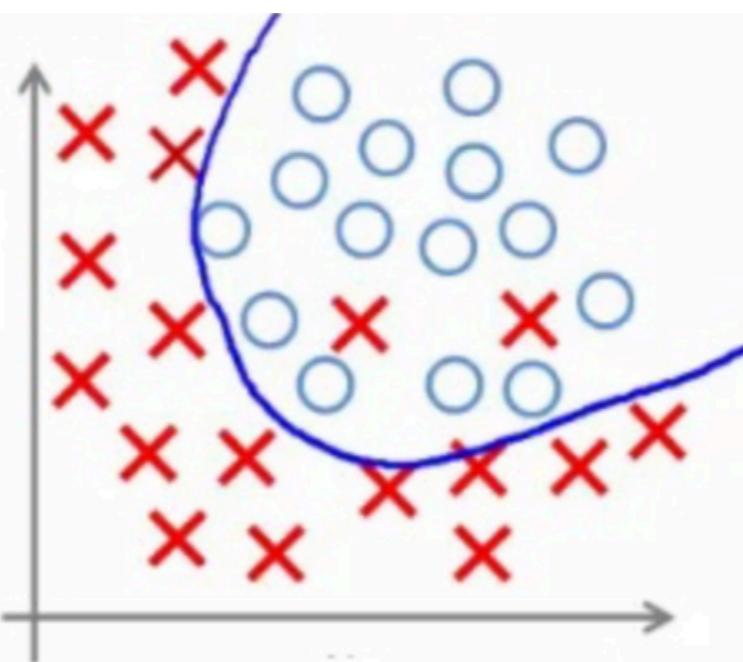
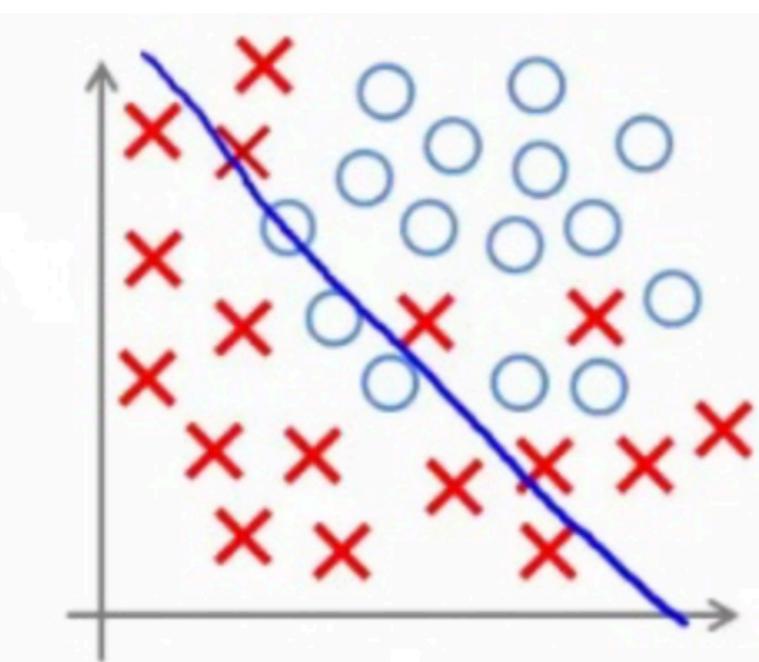
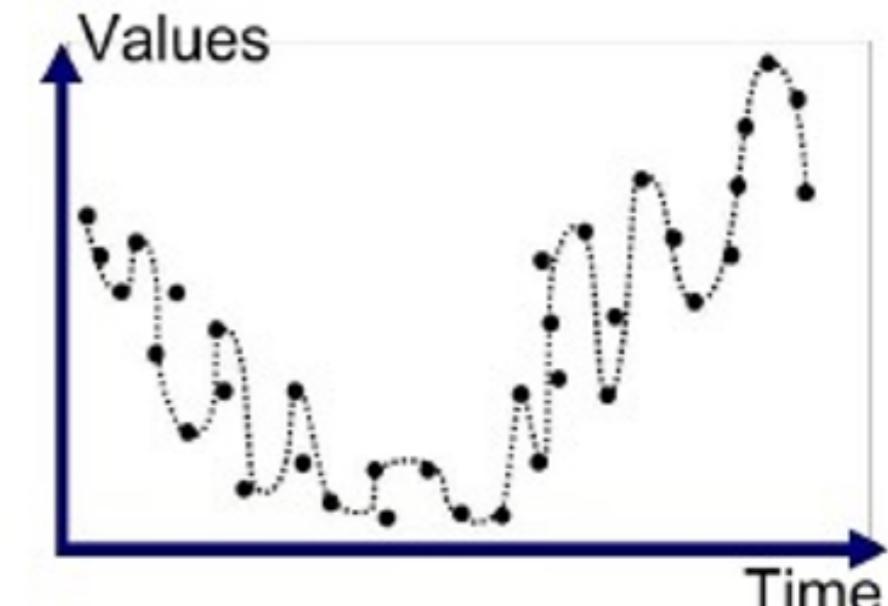
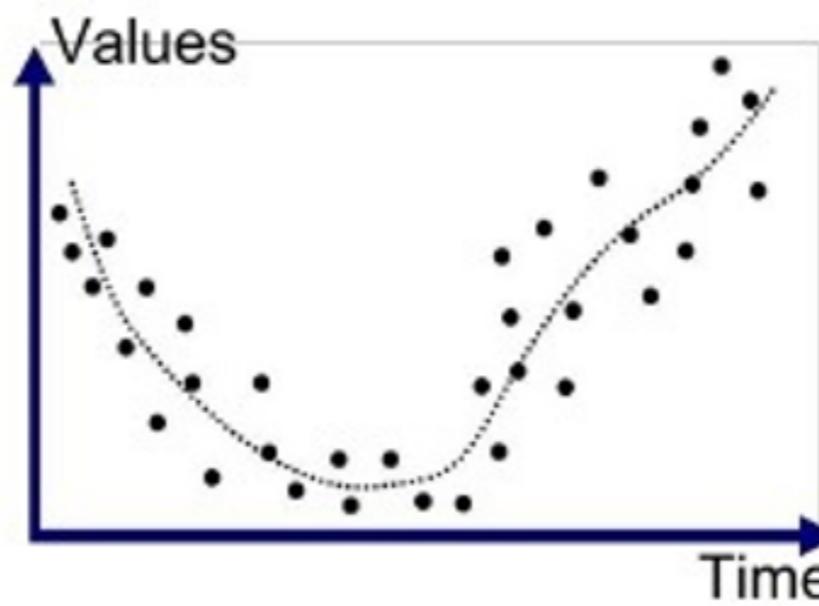
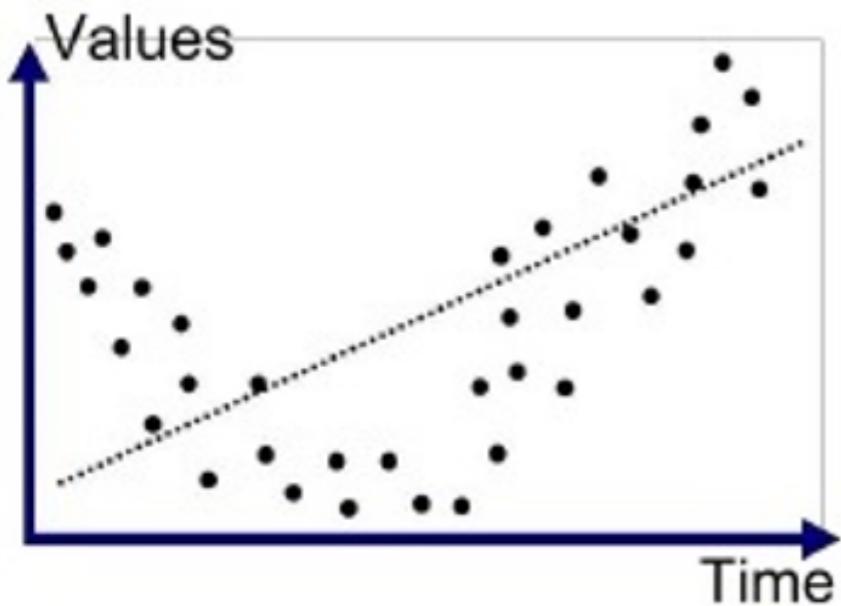
Полином 5-ой степени	0.047	0.179	0.9000
----------------------	-------	-------	--------

Полином 11-ой степени	0.000	0.000	1.0000
-----------------------	-------	-------	--------

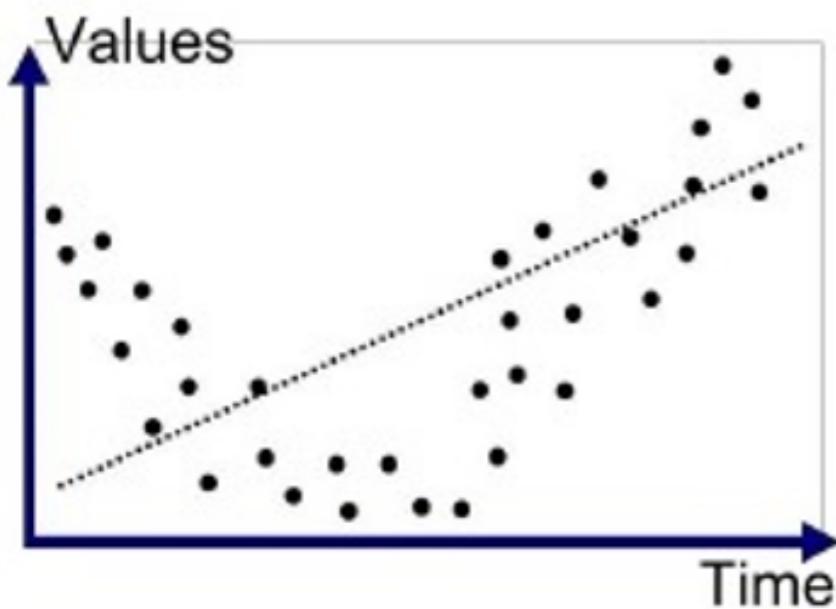


**Обобщающая способность
Полинома 11-ой степени
отсутствует.**

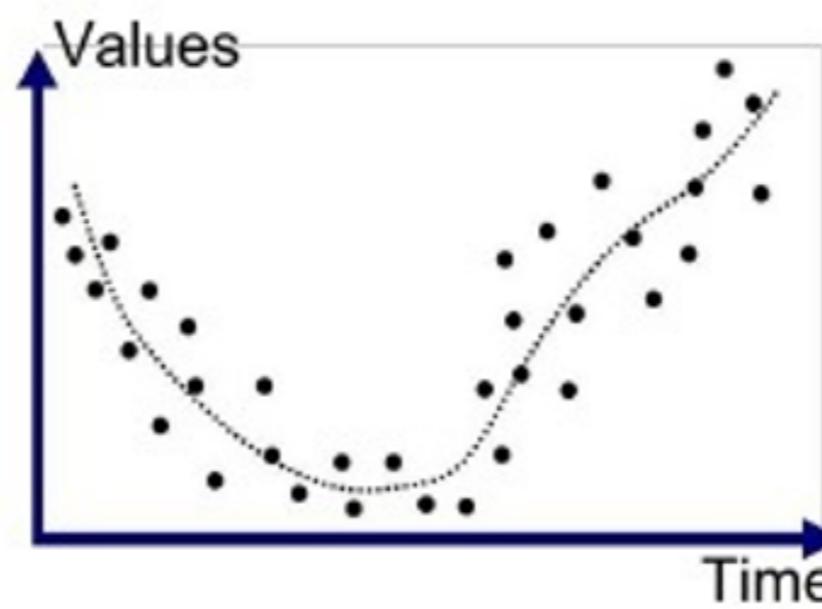
Мотивация валидации



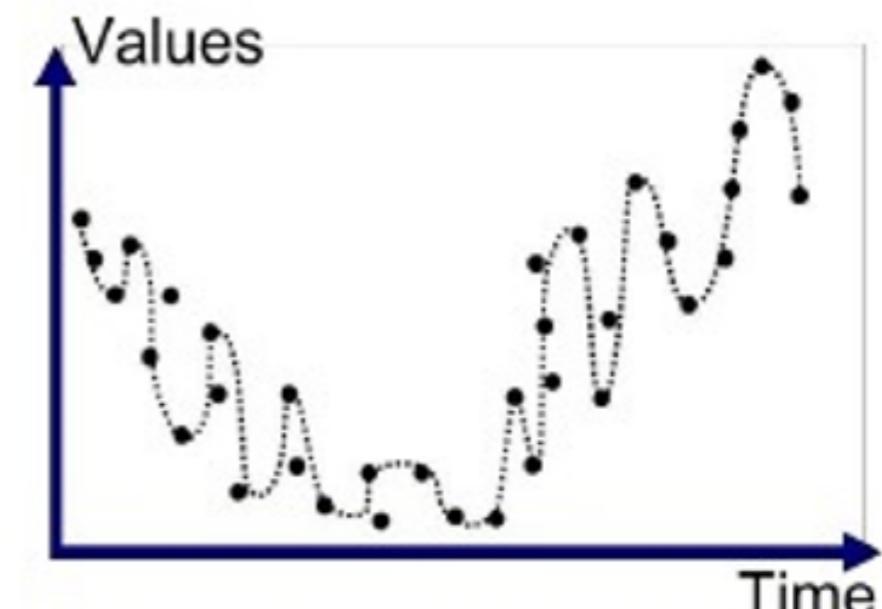
Мотивация валидации



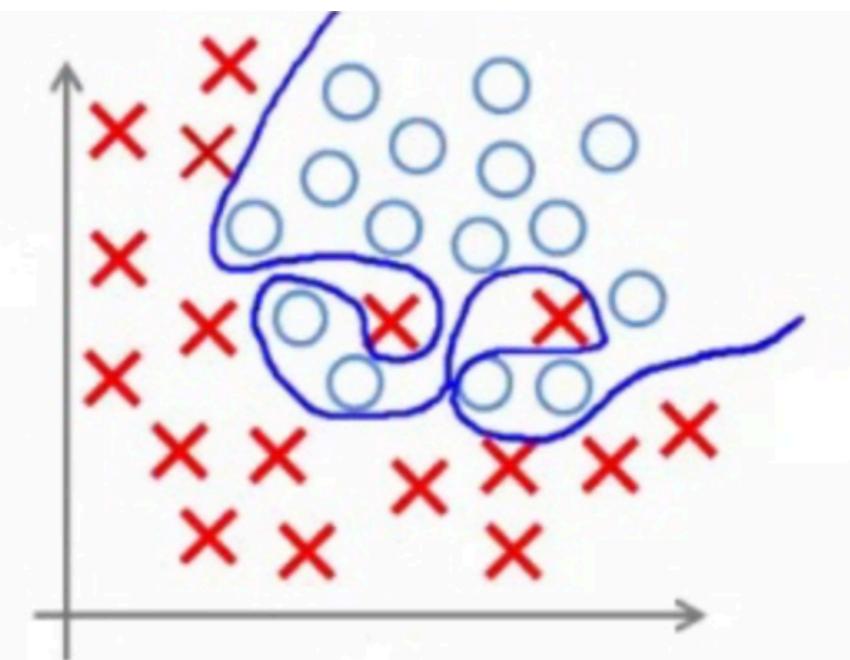
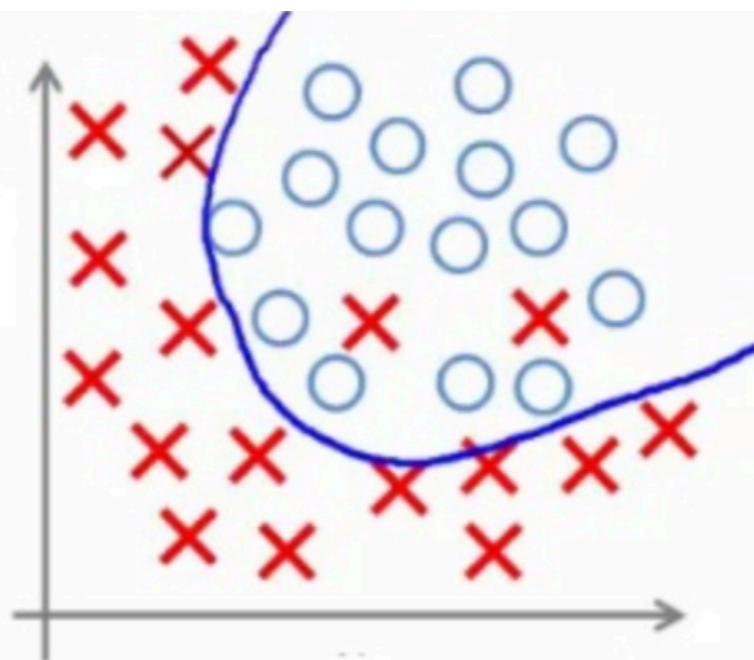
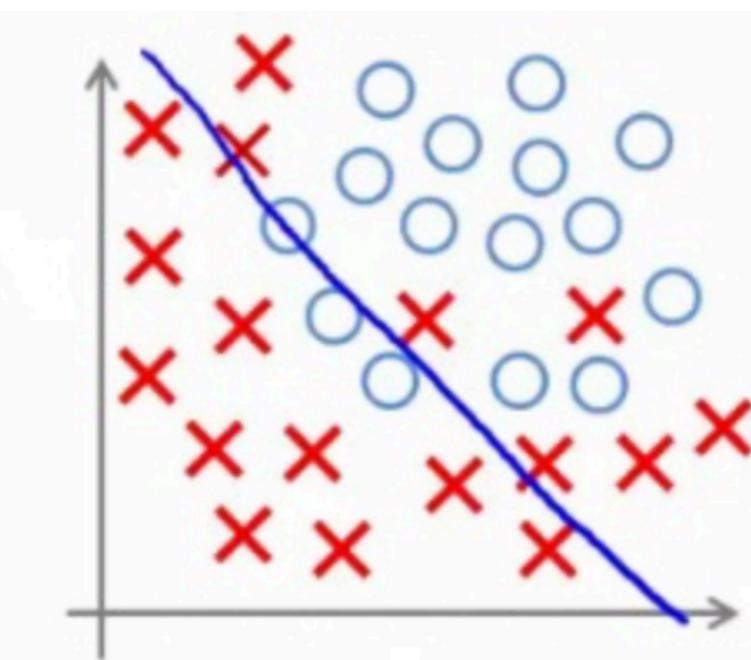
Underfitted



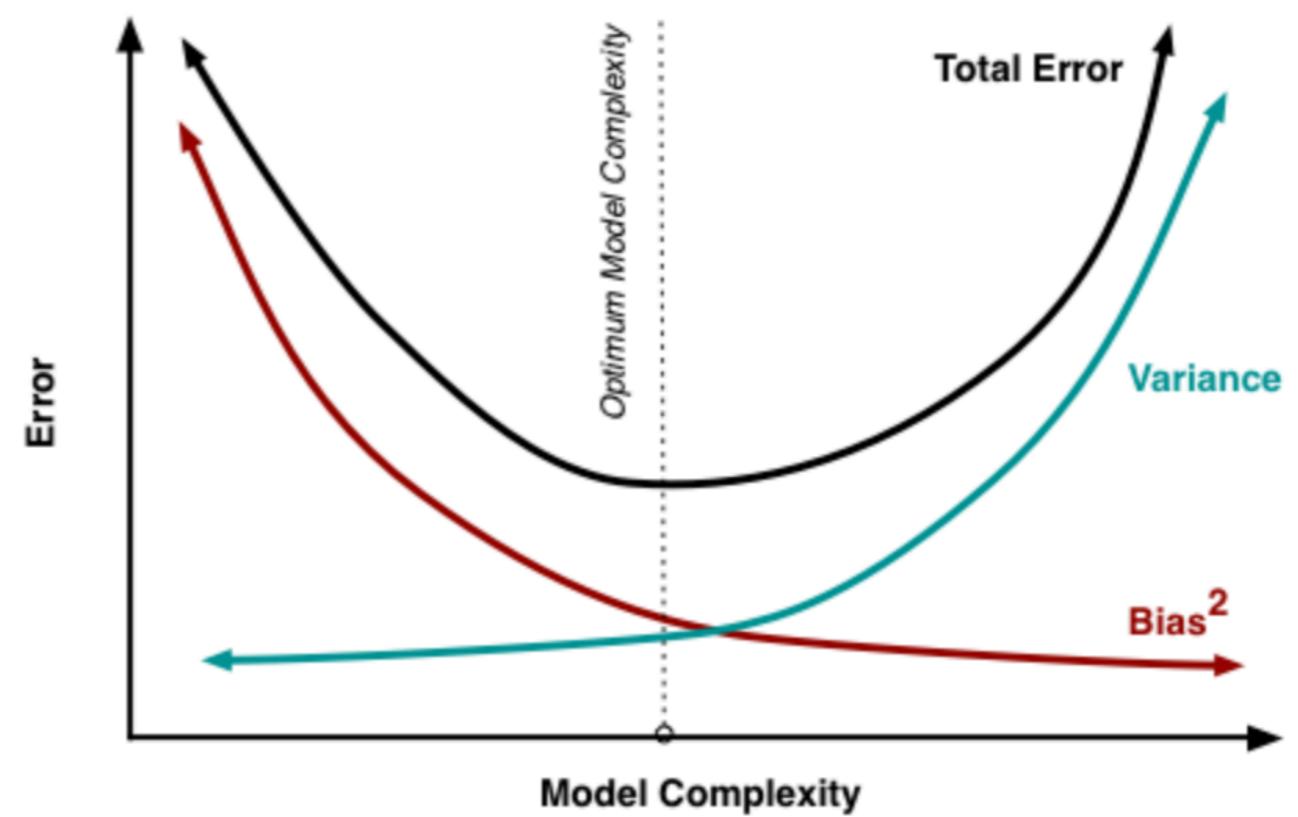
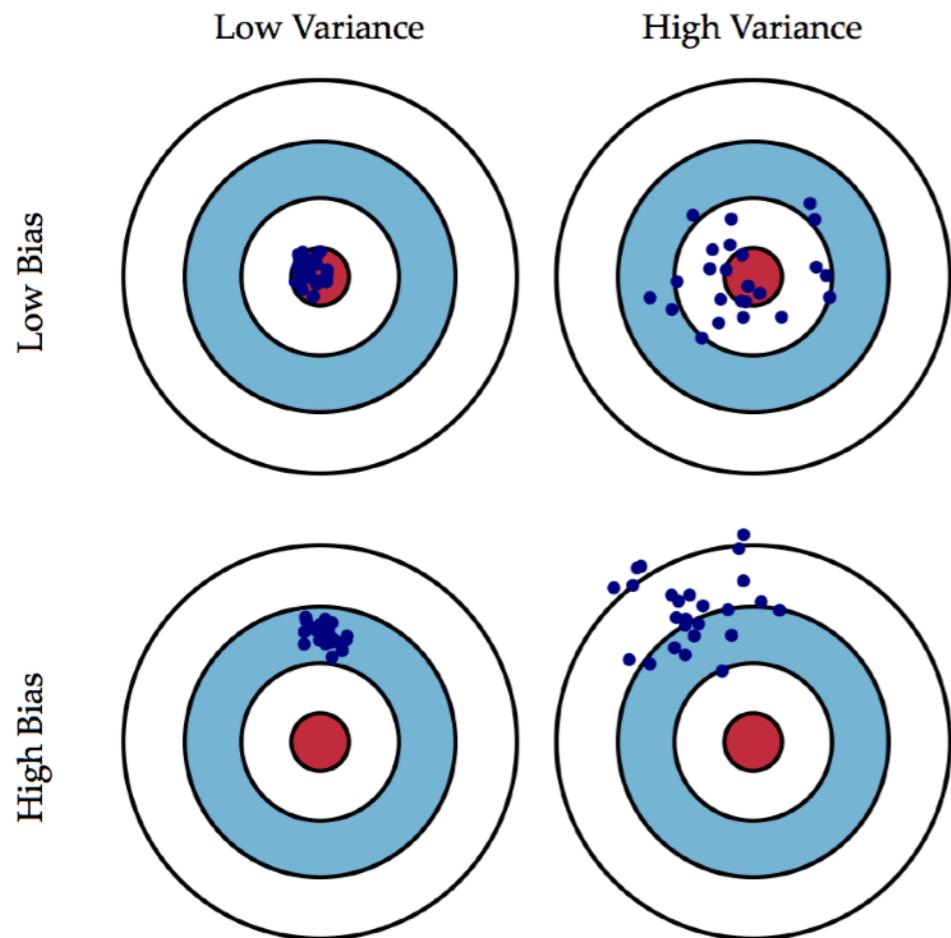
Good Fit/Robust



Overfitted



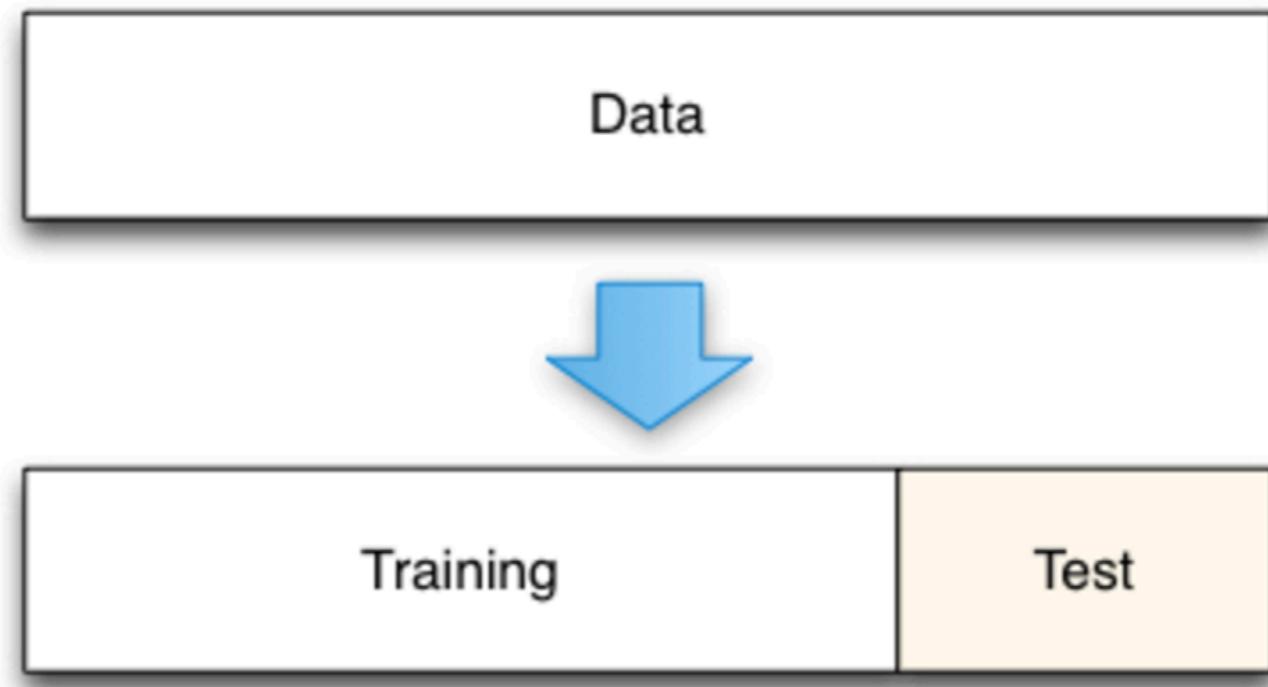
Bias and Variance tradeoff



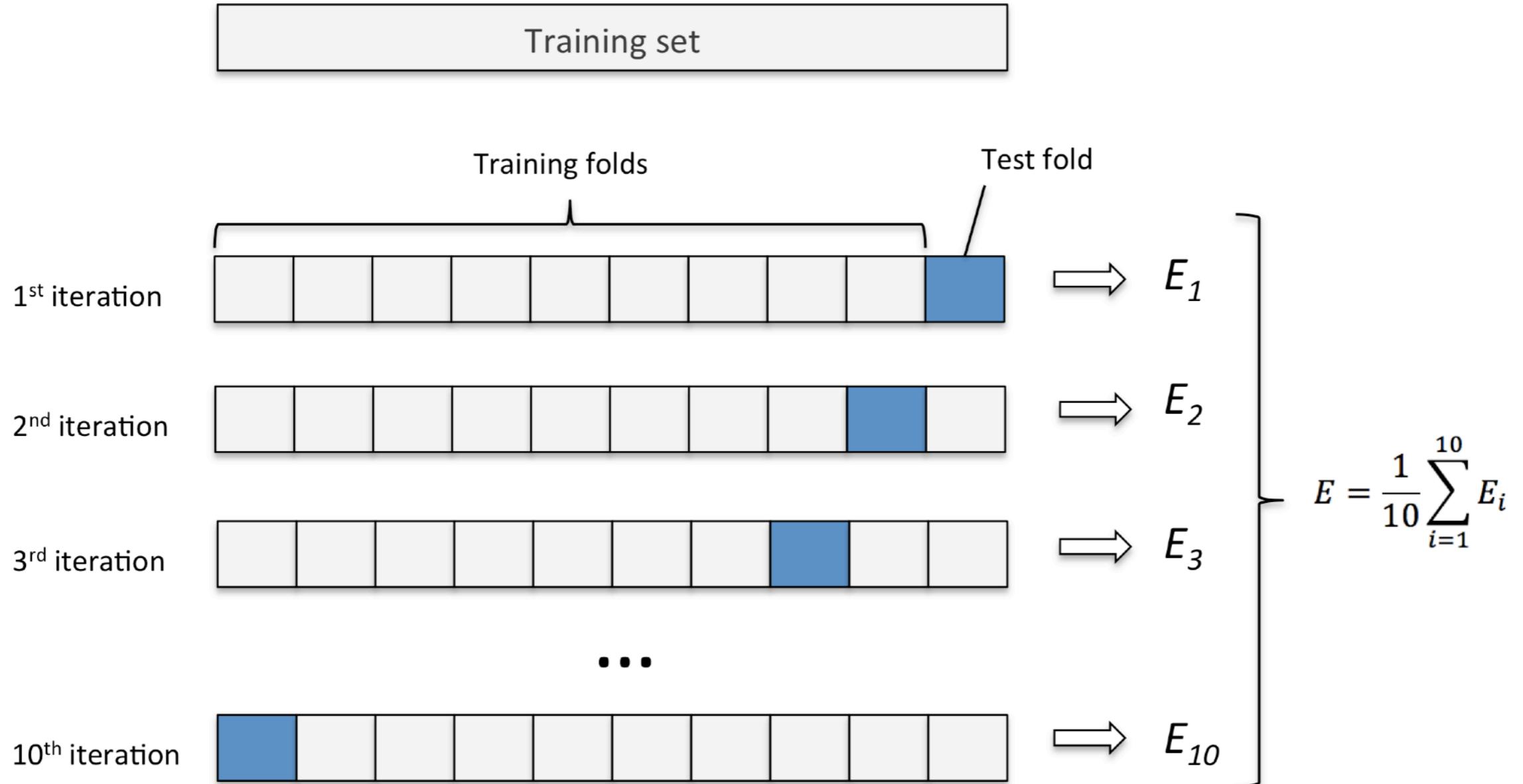
$$Err(x) = E[(Y - \hat{f}(x))^2]$$

$$Err(x) = Bias^2 + Variance + IrreducibleError$$

Стратегии валидации



Стратегии валидации



Деревья принятия решений

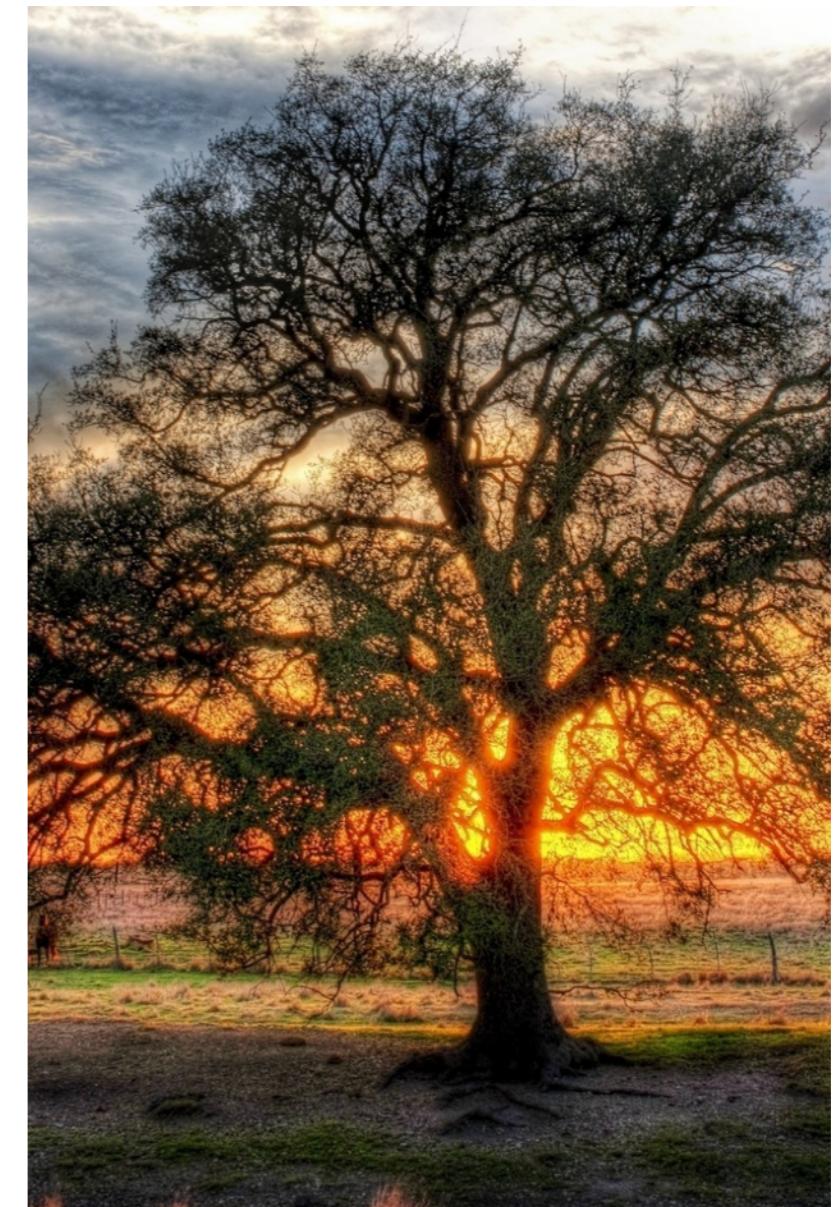
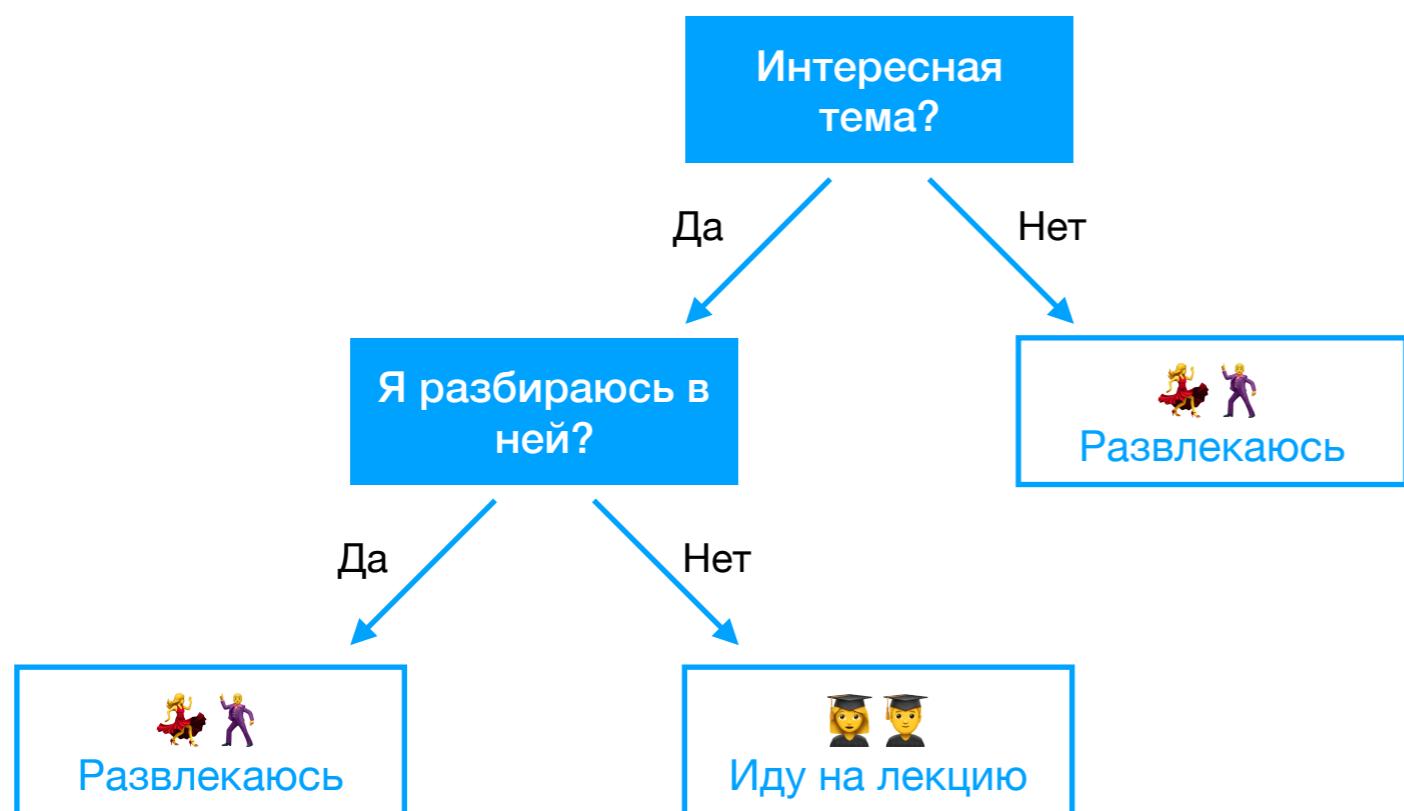
Логический алгоритм
классификации, основанный
на поиске конъюнктивных
закономерностей.



Деревья принятия решений

Логический алгоритм классификации, основанный на поиске конъюнктивных закономерностей.

Пойду ли я на МО сегодня?



Деревья принятия решений

Дерево решений служит обобщением опыта экспертов, средством передачи знаний будущим сотрудникам или моделью бизнес-процесса компании.

Решение о выдаче кредита заемщику принималось на основе некоторых интуитивно (или по опыту) выведенных правил, которые можно представить в виде дерева решений.

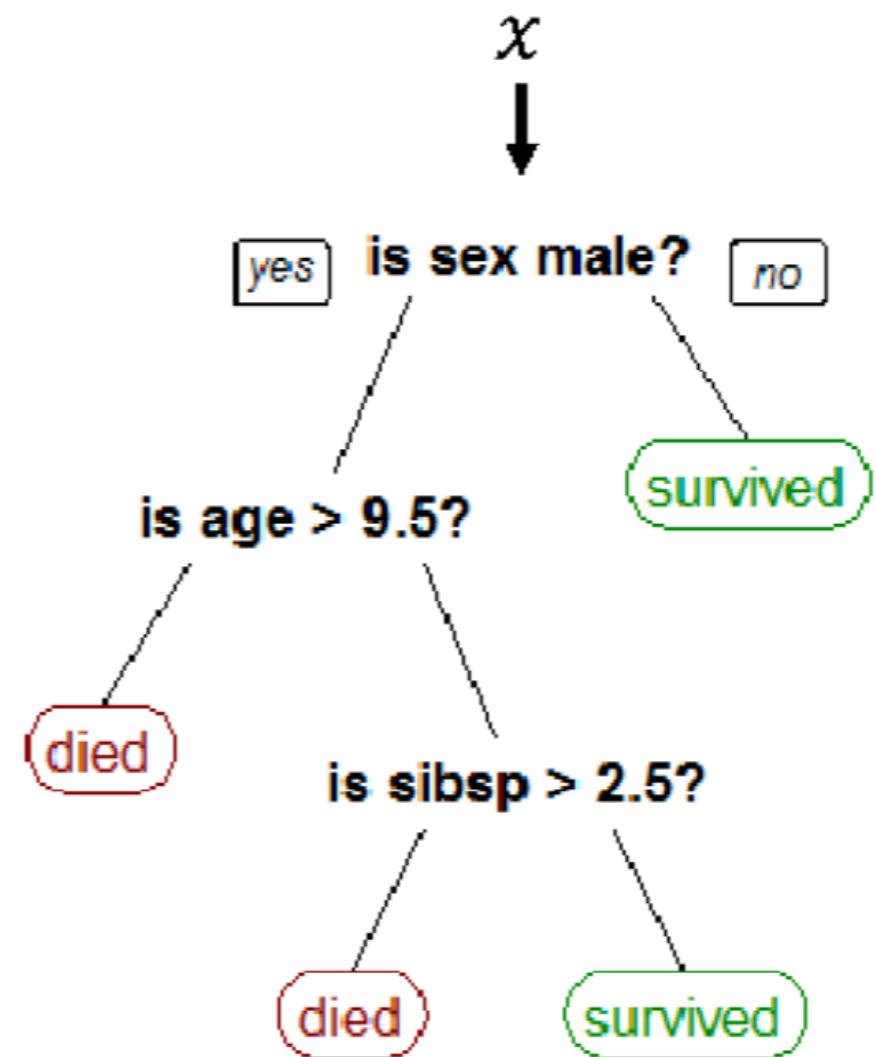
Пример: Кредитный scoring



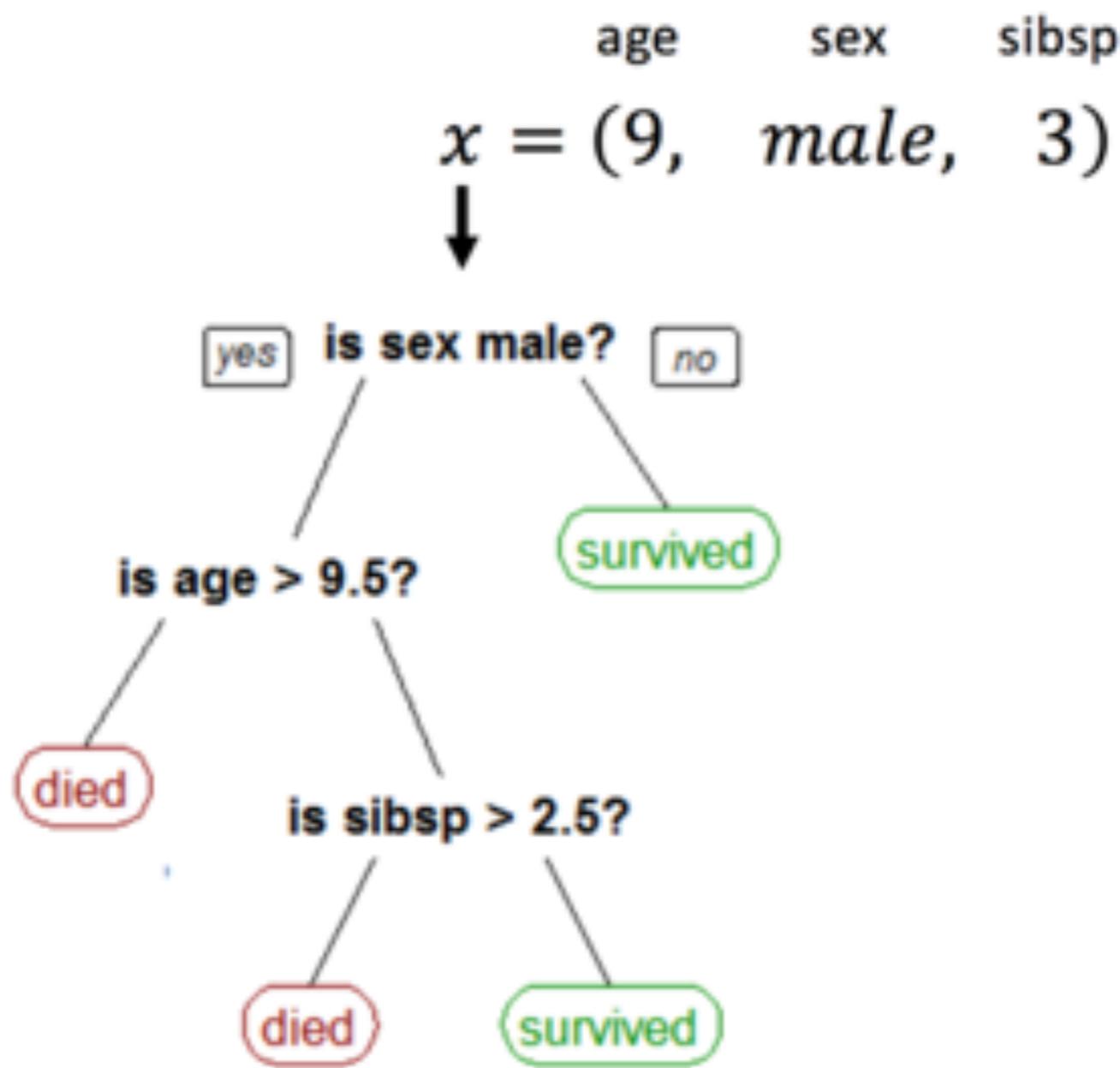
Деревья принятия решений



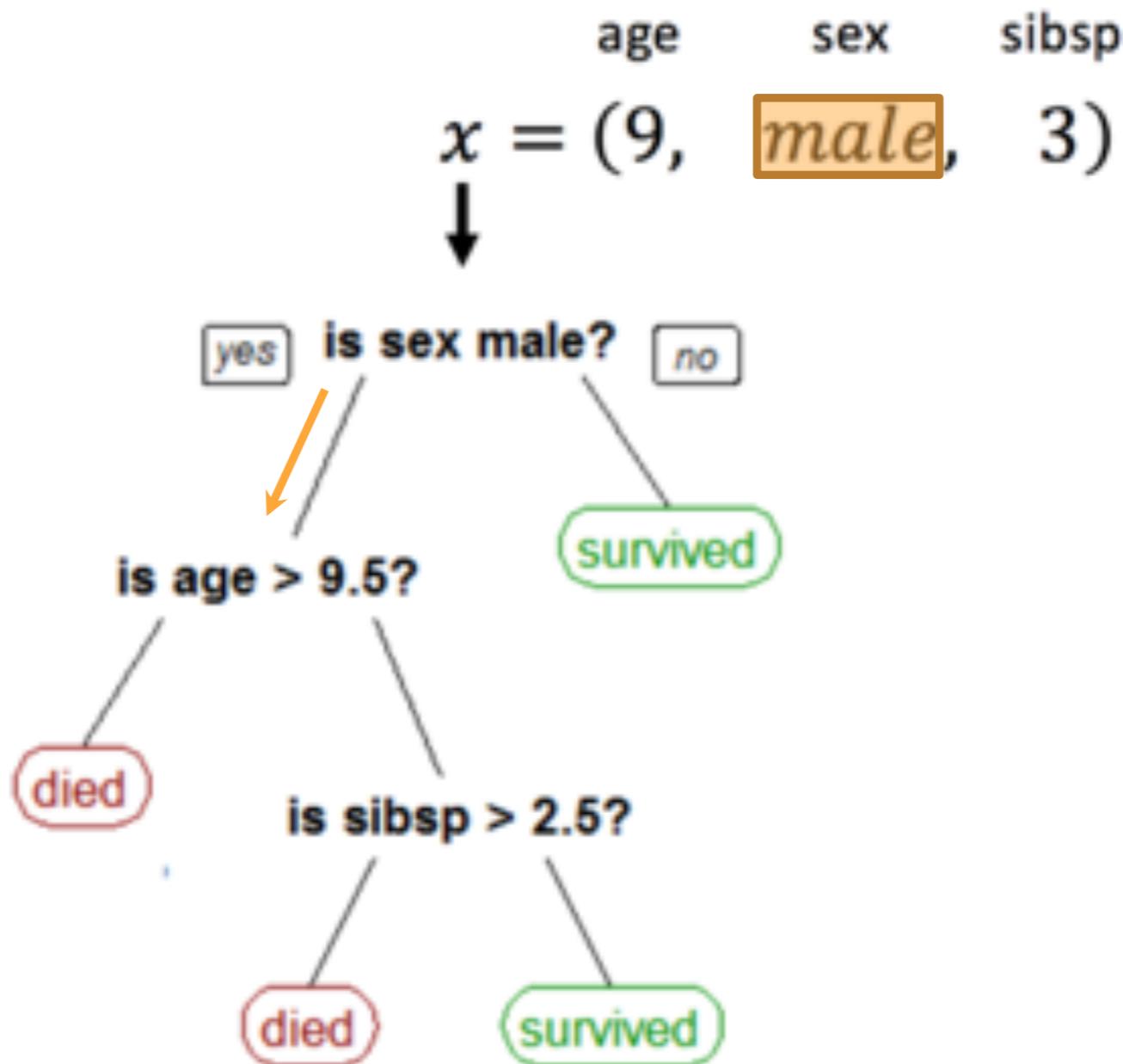
<https://www.kaggle.com/c/titanic>



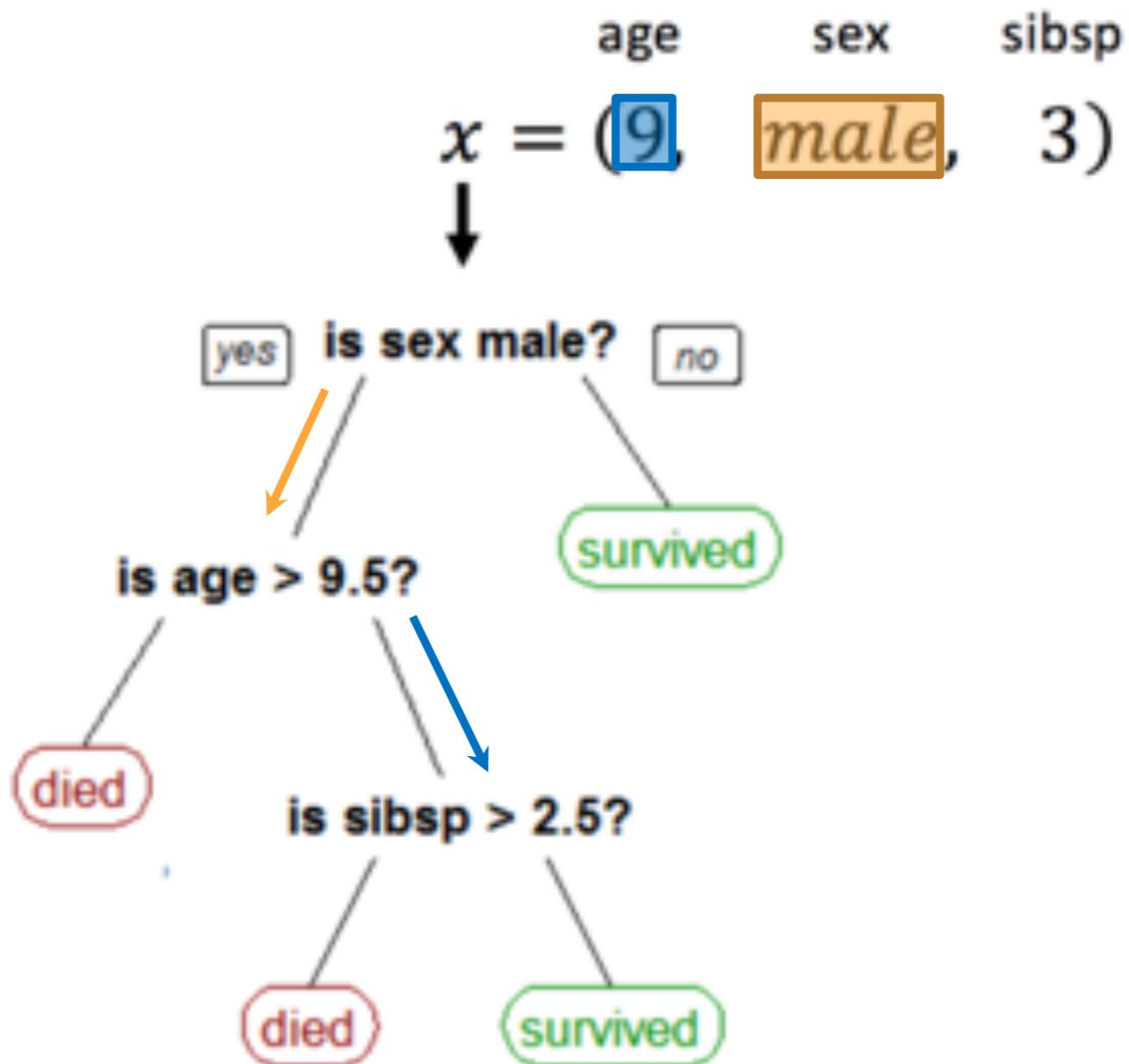
Деревья принятия решений



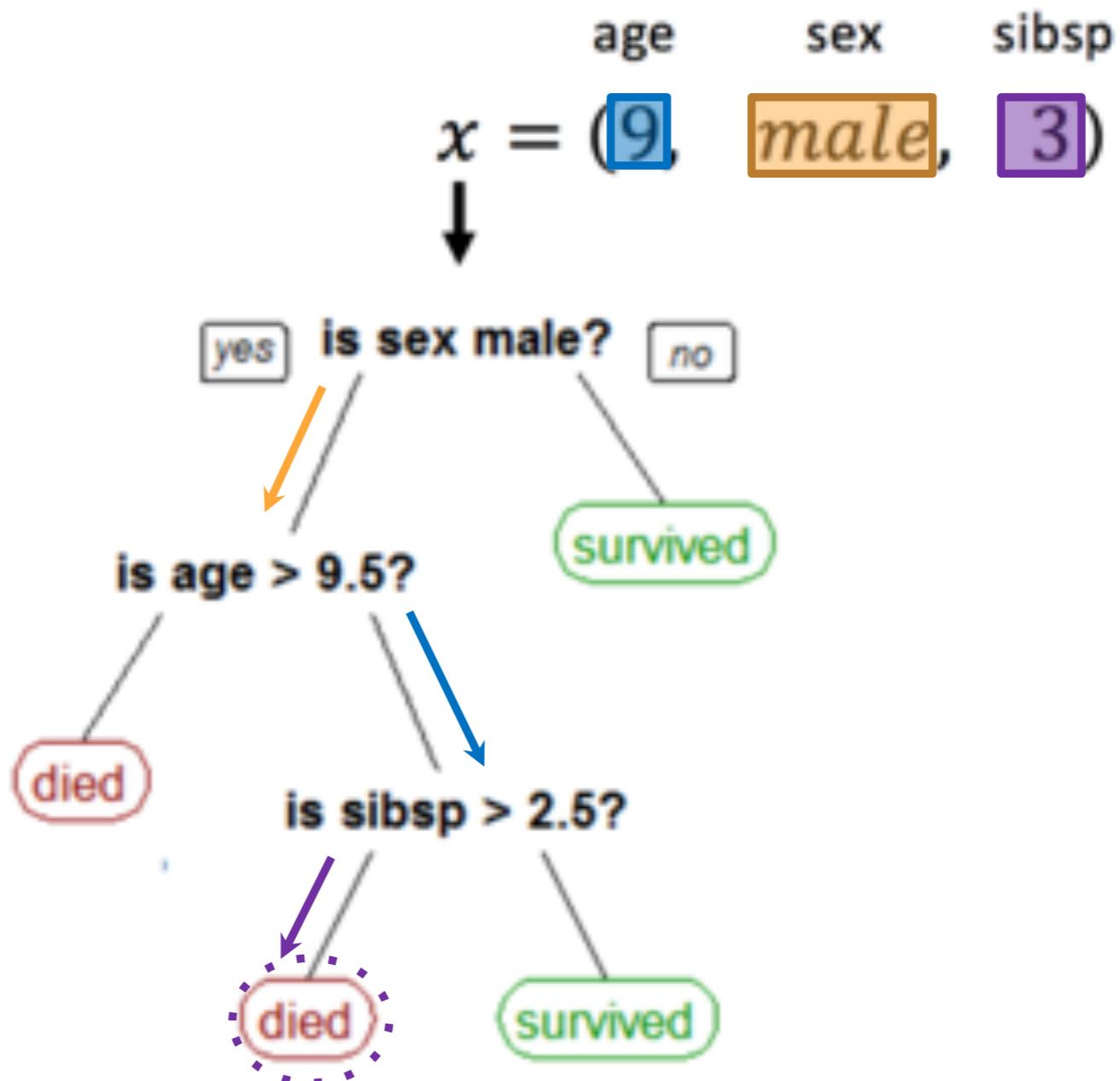
Деревья принятия решений



Деревья принятия решений



Деревья принятия решений



Как вырастить собственное дерево?



Игра «20 вопросов»

Построение дерева принятия решений

Энтропия Шеннона

$$S = - \sum_i^N p_i \log_2(p_i)$$

Энтропия соответствует степени хаоса в системе. Чем выше энтропия, тем менее упорядочена система и наоборот.

Прирост информации

$$IG(Q) = S_0 - \sum_i^q \frac{N_i}{N} S_i$$

Выбирается такое разбиение, при котором прирост информации максимальен

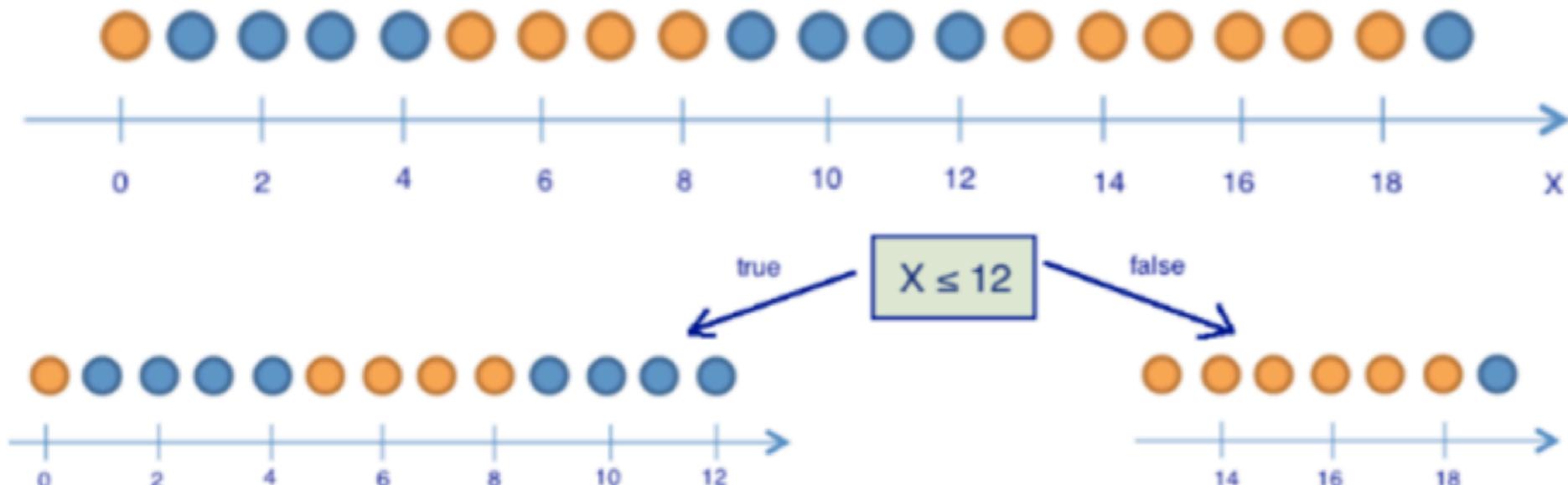
Построение дерева принятия решений



$$S = - \sum_i^N p_i \log_2(p_i)$$

$$IG(Q) = S_0 - \sum_i^q \frac{N_i}{N} S_i$$

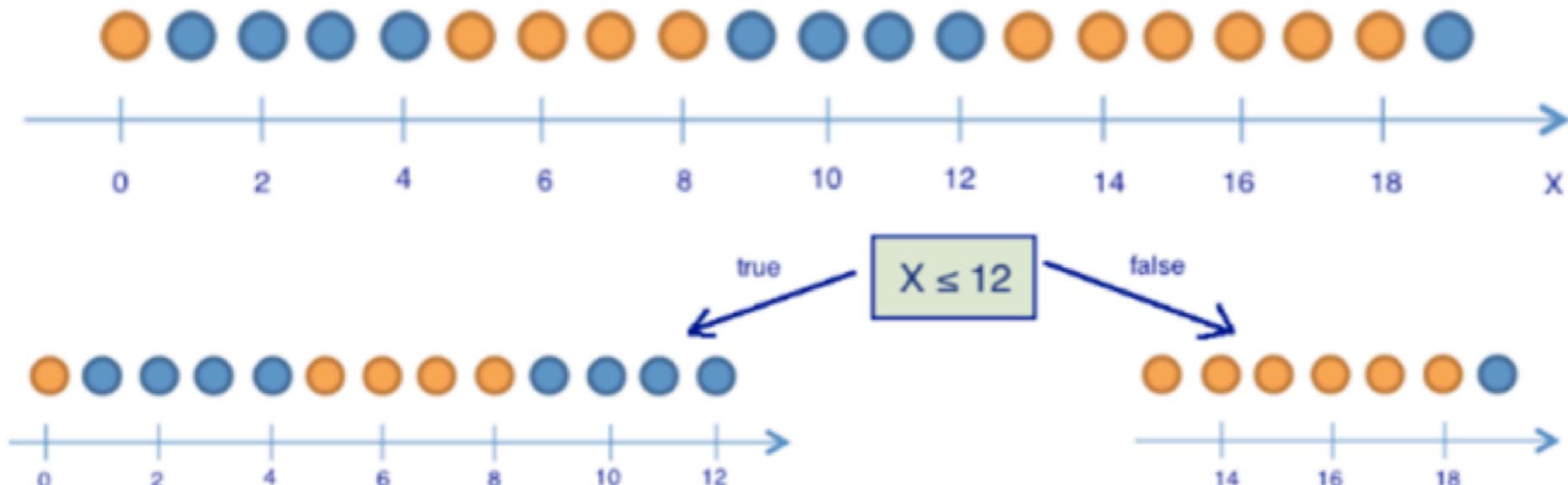
Построение дерева принятия решений



$$S = - \sum_i^N p_i \log_2(p_i)$$

$$IG(Q) = S_0 - \sum_i^q \frac{N_i}{N} S_i$$

Построение дерева принятия решений



$$S_0 = -\frac{9}{20} \log_2 \frac{9}{20} - \frac{11}{20} \log_2 \frac{11}{20} \approx 1$$

$$S_1 = -\frac{5}{13} \log_2 \frac{5}{13} - \frac{8}{13} \log_2 \frac{8}{13} \approx 0,96$$

$$S_2 = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} \approx 0,6$$

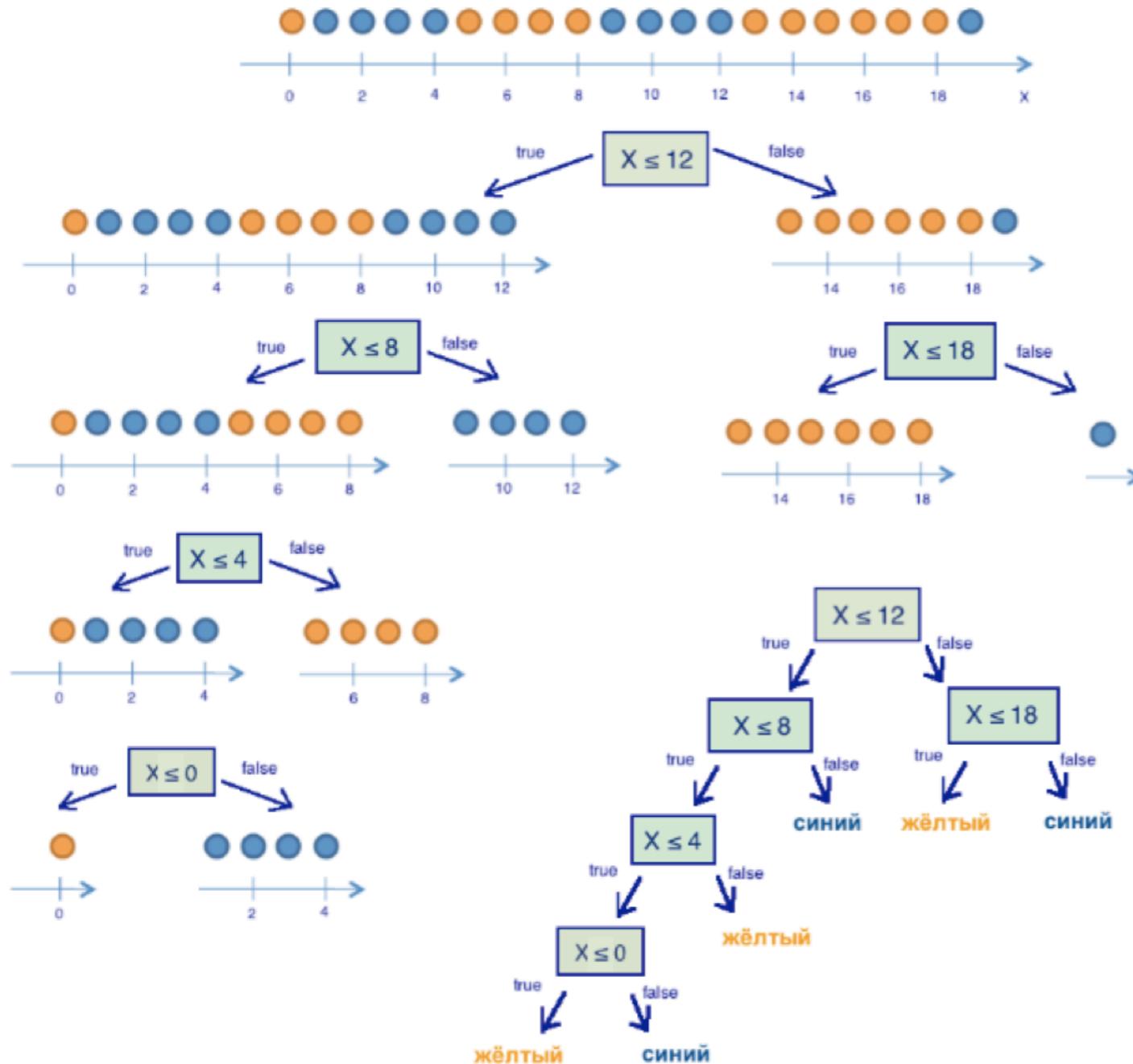
$$IG(x \leq 12) = S_0 - \frac{13}{20} \times S_1 - \frac{7}{20} \times S_2 \approx 0,16$$

разделив шарики на две группы по признаку "координата меньше либо равна 12", мы уже получили более упорядоченную систему, чем в начале.

$$S = - \sum_i^N p_i \log_2(p_i)$$

$$IG(Q) = S_0 - \sum_i^q \frac{N_i}{N} S_i$$

Построение дерева принятия решений



энтропия группы с шариками одного цвета равна 0, что соответствует представлению, что группа шариков одного цвета – упорядоченная.

Построение дерева принятия решений

$$S = - \sum_i^N p_i \log_2(p_i)$$

Энтропийный критерий (Entropy criteria)

$$S = 1 - \sum_{k=1}^n (p_k)^2$$

Неопределенность Джини (Gini impurity)

$$S = 1 - \max_k p_k$$

Ошибка классификации (misclassification error)

почти не используется

Построение дерева принятия решений

Критерии качества как функции от $p+$ (бинарная классификация)

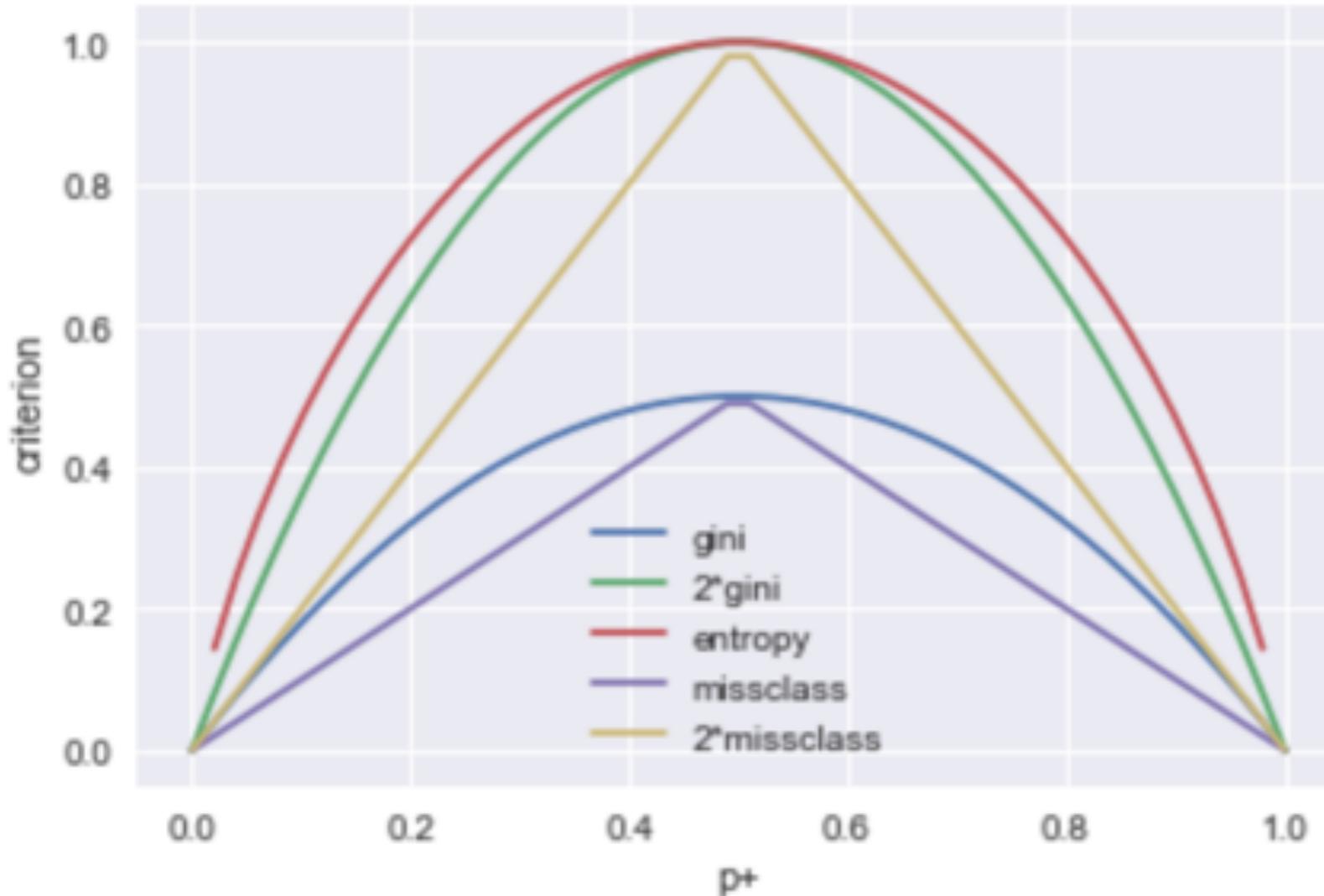


График энтропии очень близок к графику удвоенной неопределенности Джини, поэтому на практике работают почти одинаково.

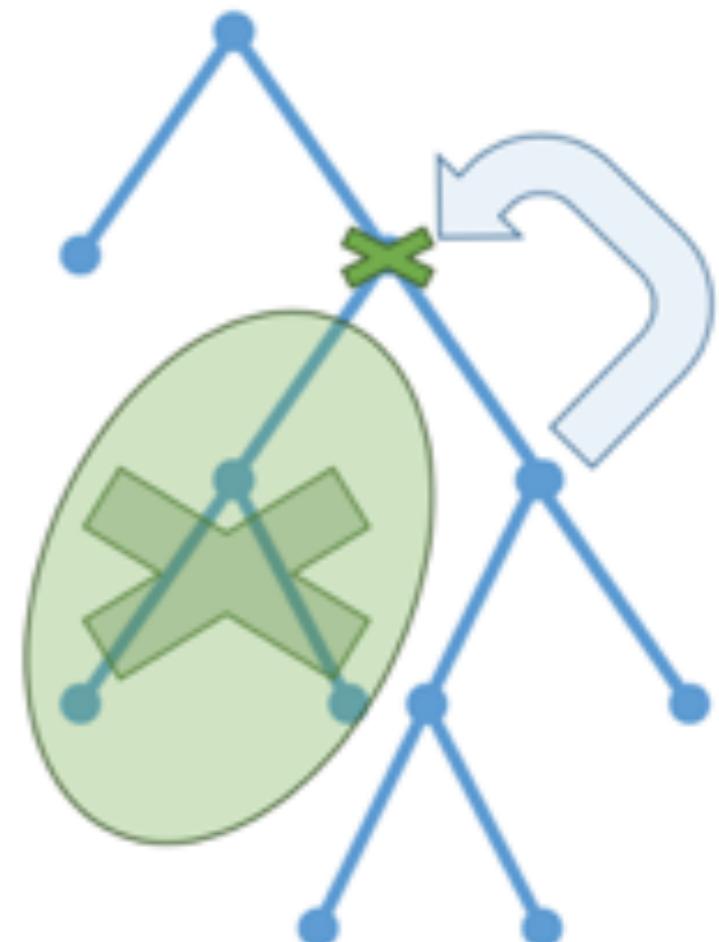
Борьба с переобучением (Pruning)

▶ Pre-pruning (ограничиваем до обучения)

- ▷ Максимальная глубина дерева
- ▷ Минимальное число элементов в узле дерева
- ▷ Минимальное число элементов в для разбиения
- ▷ Минимальный “Information gain”
- ▷ ...

▶ Post-pruning (упрощаем после обучения)

- ▷ Reduced error pruning
- ▷ Cost-complexity pruning
- ▷ ...

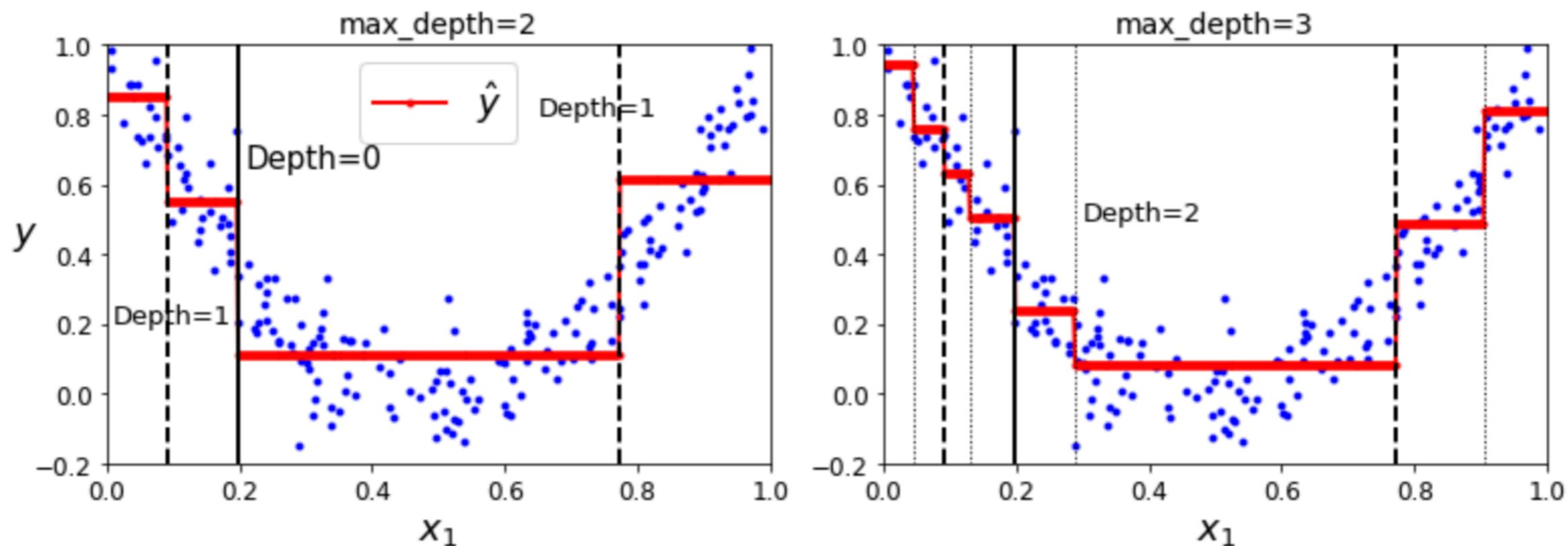


Деревья решений для задачи регрессии



`sklearn.tree.DecisionTreeRegressor`

(`criterion='mse'`, `splitter='best'`, `max_depth=None`, `min_samples_split=2`,
`min_samples_leaf=1`, `min_weight_fraction_leaf=0.0`, `max_features=None`,
`random_state=None`, `max_leaf_nodes=None`, `min_impurity_decrease=0.0`,
`min_impurity_split=None`, `presort=False`)



Какой ответ деревьев в регрессии?

Какая стратегия поведения в листьях регрессионного дерева приводит к меньшему матожиданию ошибки по MSE: отвечать средним значением таргета на объектах обучающей выборки, попавших в лист, или отвечать таргетом для случайного объекта из листа (считая все объекты равновероятными)?

Какой ответ деревьев в регрессии?

Какая стратегия поведения в листьях регрессионного дерева приводит к меньшему матожиданию ошибки по MSE: отвечать средним значением таргета на объектах обучающей выборки, попавших в лист, или отвечать таргетом для случайного объекта из листа (считая все объекты равновероятными)?

- $\hat{y} = \frac{1}{n} \sum_{i=1}^n c_i$

$$\mathbb{E}(y - \frac{1}{n} \sum_{i=1}^n c_i)^2 = \mathbb{E}y^2 + \left(\frac{1}{n} \sum_{i=1}^n c_i \right)^2 - 2 \left(\frac{1}{n} \sum_{i=1}^n c_i \right) \mathbb{E}y$$

- $\hat{y} = X$, где $X \sim U(c)$

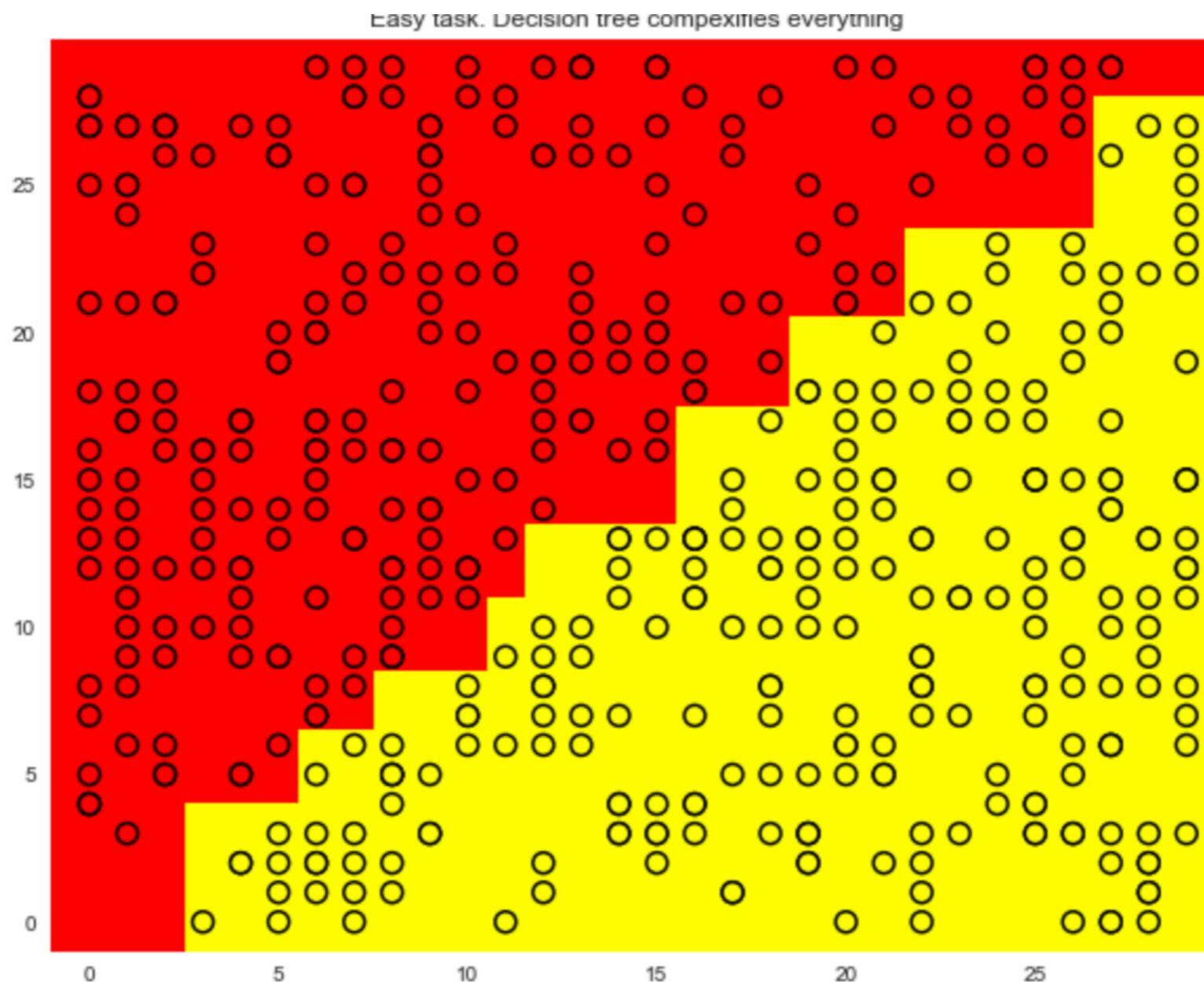
$$\mathbb{E} \frac{1}{n} \sum_{i=1}^n (y - c_i)^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(y - c_i)^2 = \mathbb{E}y^2 + \frac{1}{n} \sum_{i=1}^n c_i^2 - \frac{2}{n} \mathbb{E}y \sum_{i=1}^n c_i$$

Тогда выпишем их разность:

$$\mathbb{E} \frac{1}{n} \sum_{i=1}^n (y - c_i)^2 - \mathbb{E}(y - \bar{c})^2 = \frac{1}{n} \sum_{i=1}^n c_i^2 - \left(\frac{1}{n} \sum_{i=1}^n c_i \right)^2 \geq 0 \text{ (По неравенству Коши-Буняковского)}$$

Получили, что мат. ожидание ошибки для первого поведения меньше, чем для второго.

Сложные случаи для деревьев



Ссылки

1. Статья на habr: Метрики в задачах машинного обучения
2. Семинар из курса Евгения Соколова
3. Открытый курс машинного обучения: Тема 3
4. Семинар Евгения Соколова