Машинное обучение

Власов Кирилл Вячеславович

2018

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

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

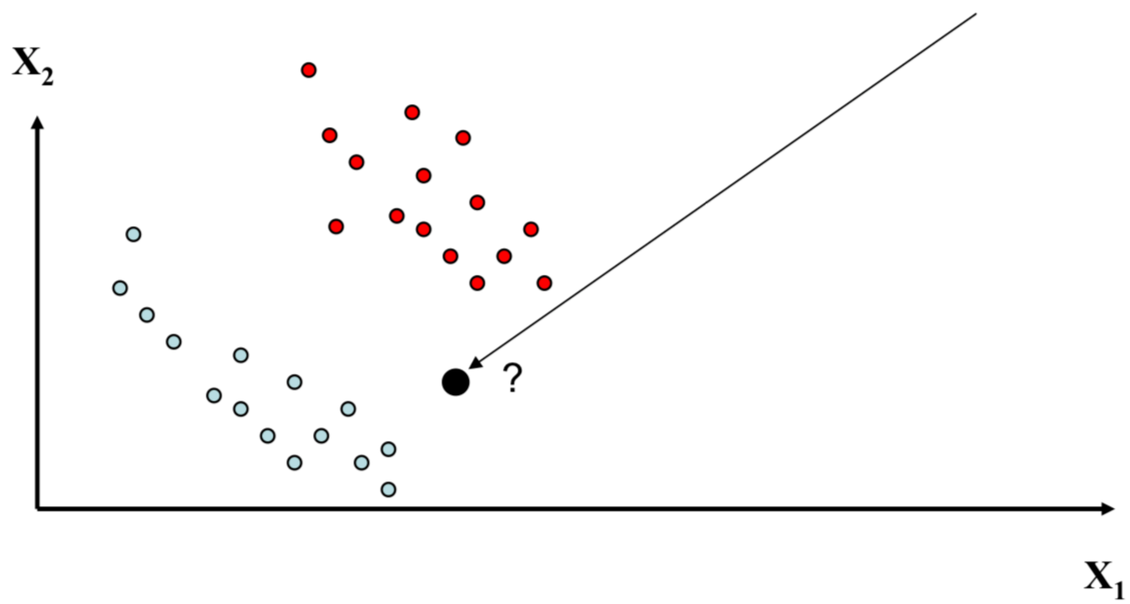
Задать такую функцию f(x) от вектора признаков x, которое выдает ответ для любого возможного наблюдения x

$$f(x)\colon \mathbb{X} \to \mathbb{Y}$$

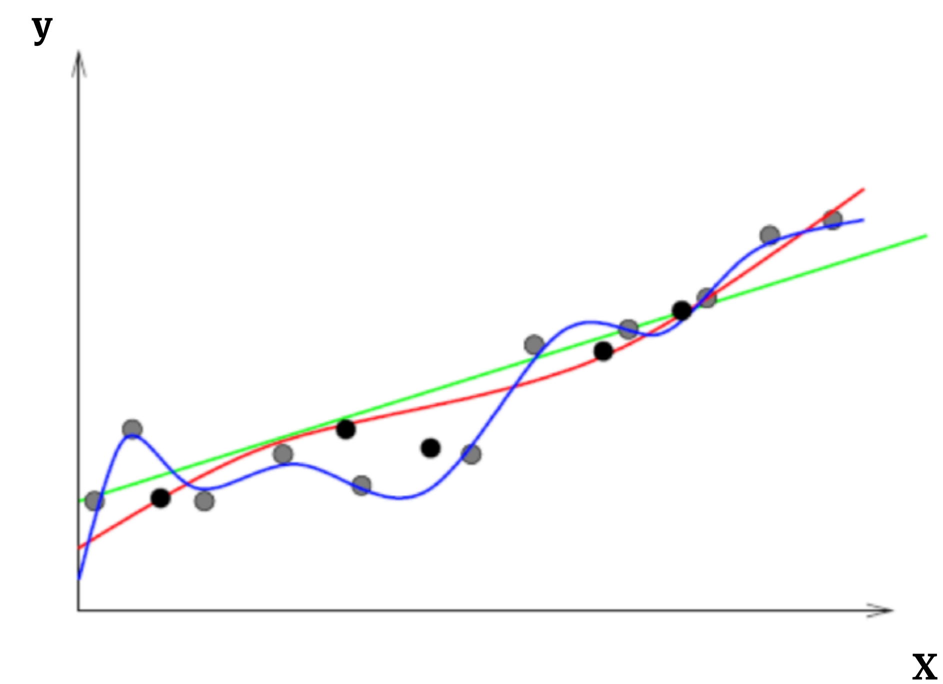
Основная гипотеза МО: Схожим объектам соответствуют схожие объекты

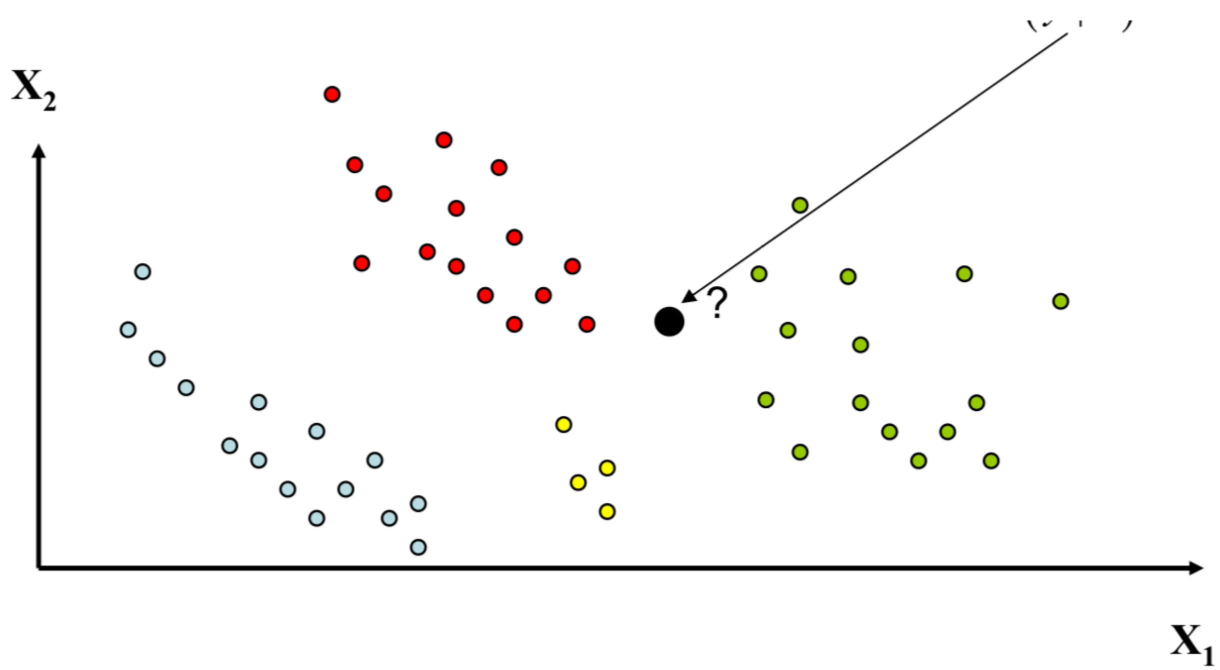
Формальная постановка задачи

Классификация

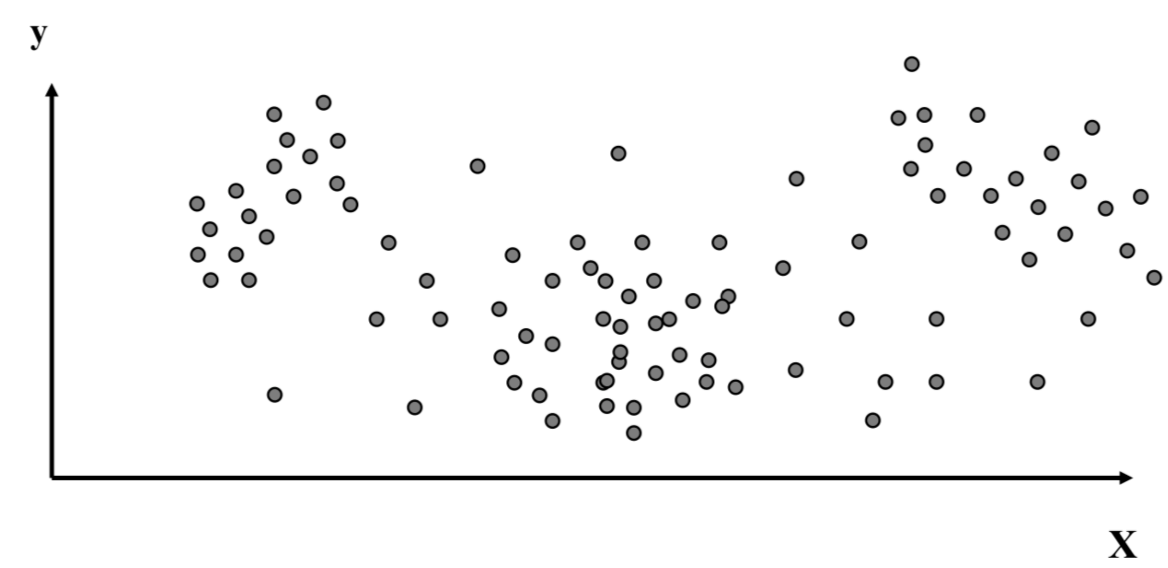


Восстановление регрессии



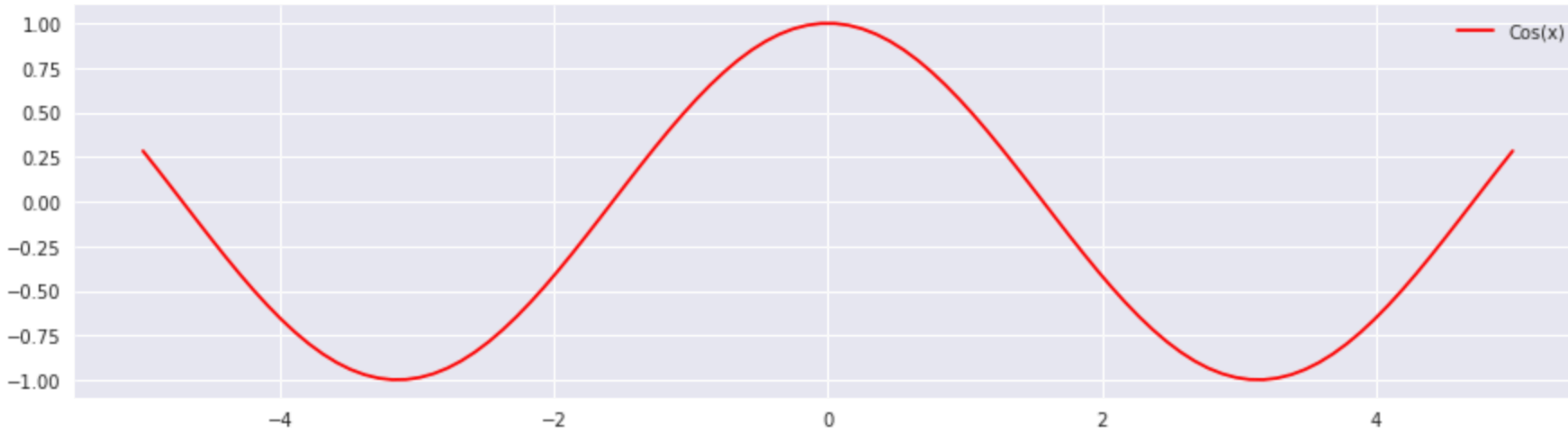


Кластеризация



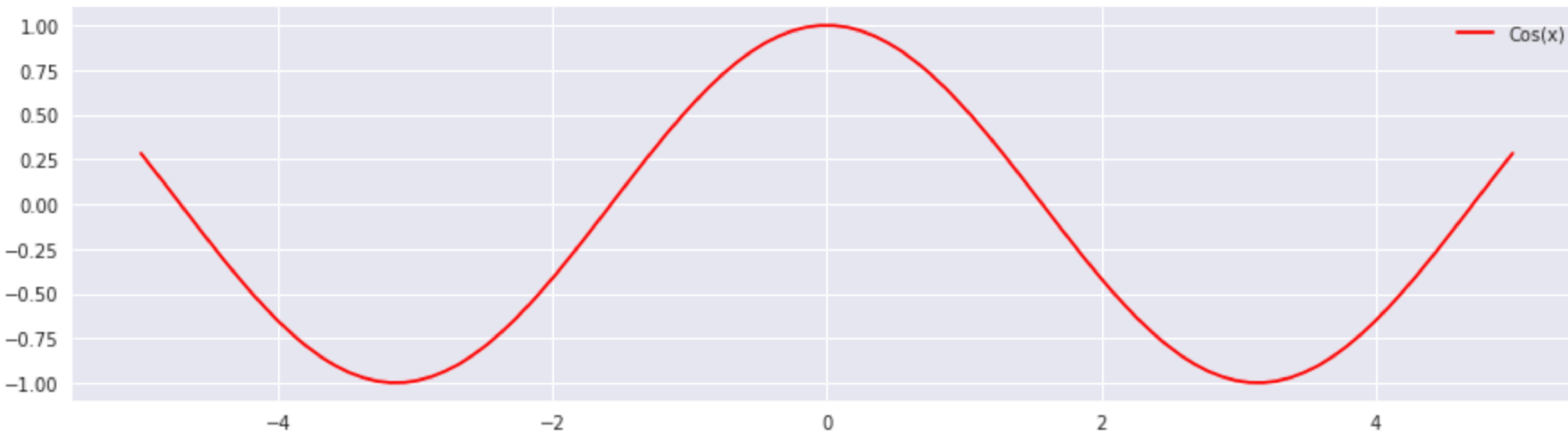
Метрики качества

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

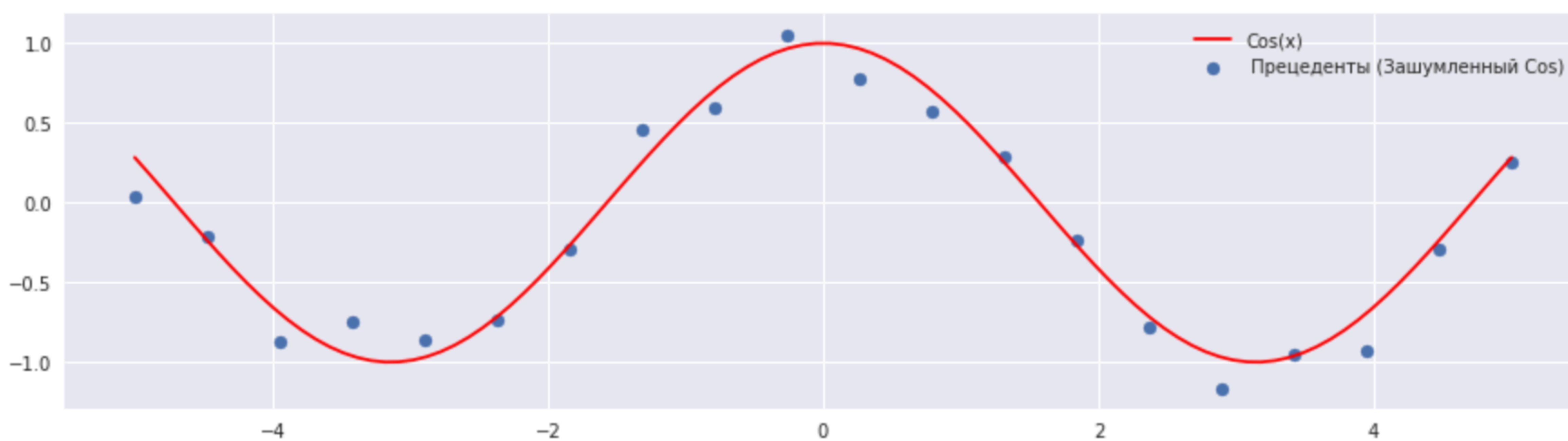


Метрики качества

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

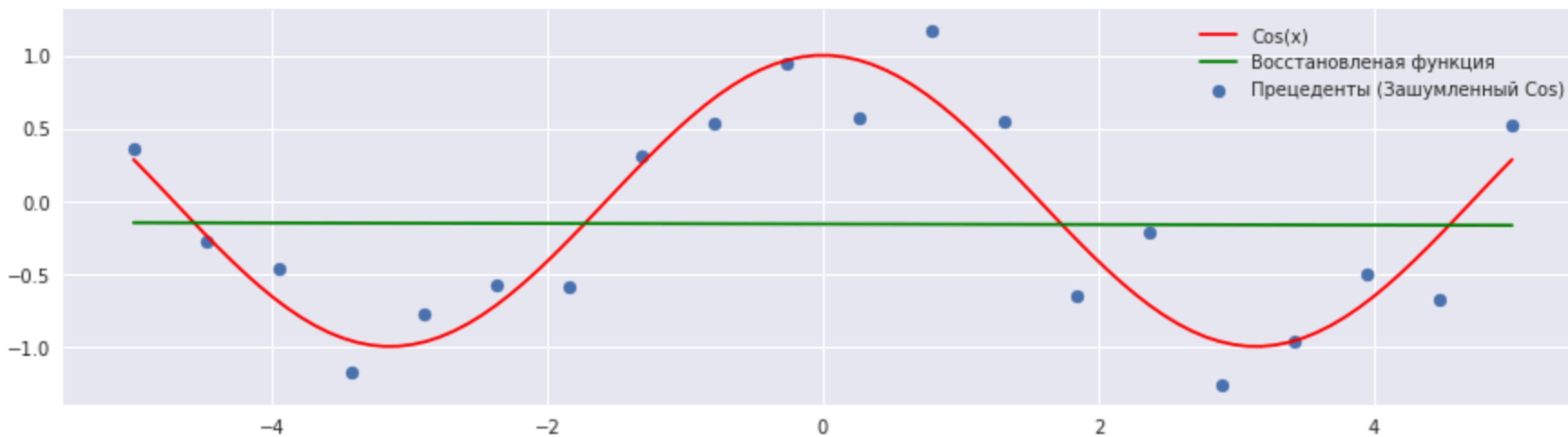


$y = \cos(x) + \varepsilon$, где $\varepsilon = \mathcal{N}\left(0, \frac{1}{2}\right), x \in [-5, 5]$



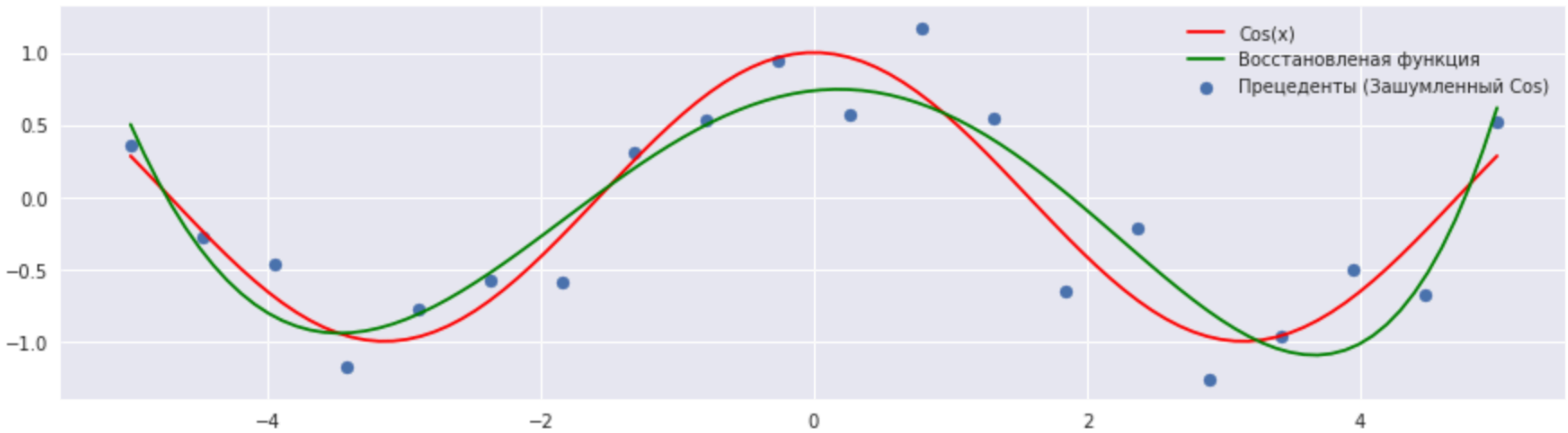
Метрики качества

Восстановим зависимость линейной функцией



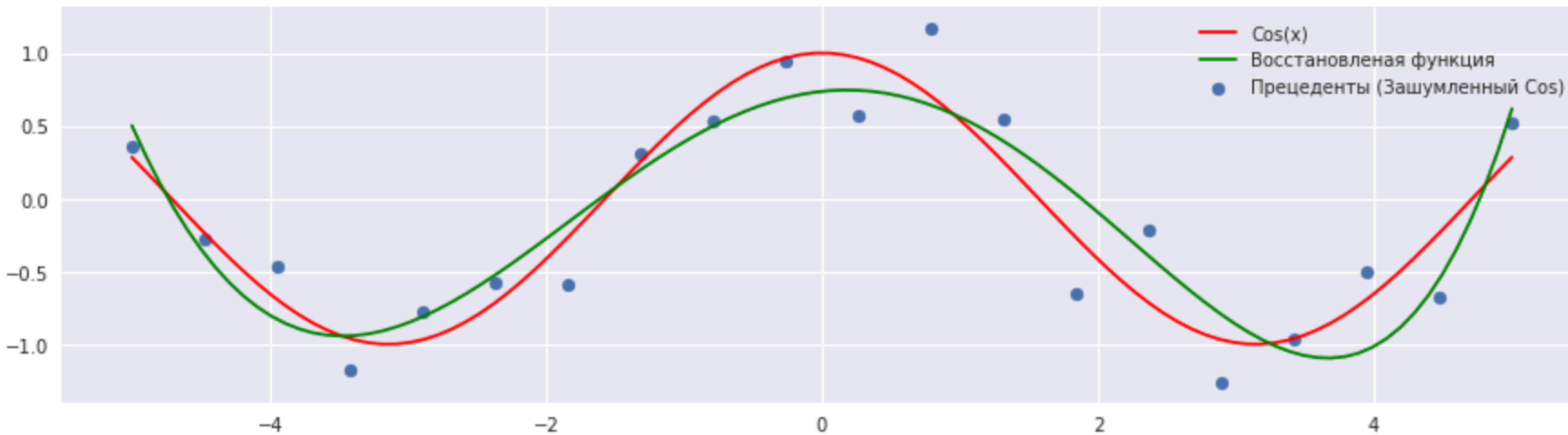
Метрики качества

Восстановим зависимость с помощью полинома 5-ого порядка

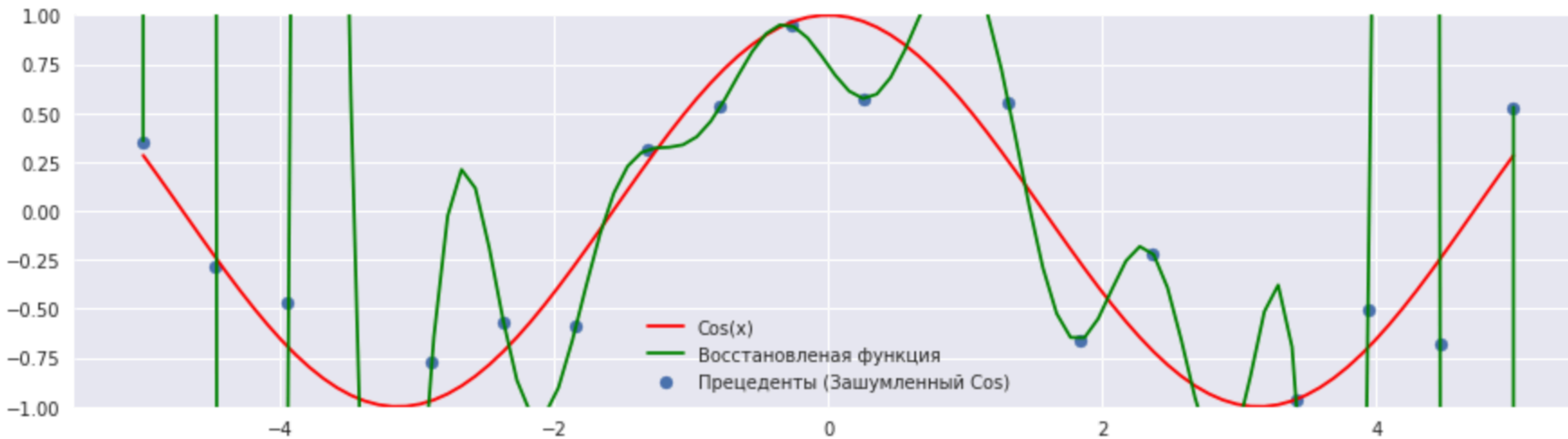


Метрики качества

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:

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

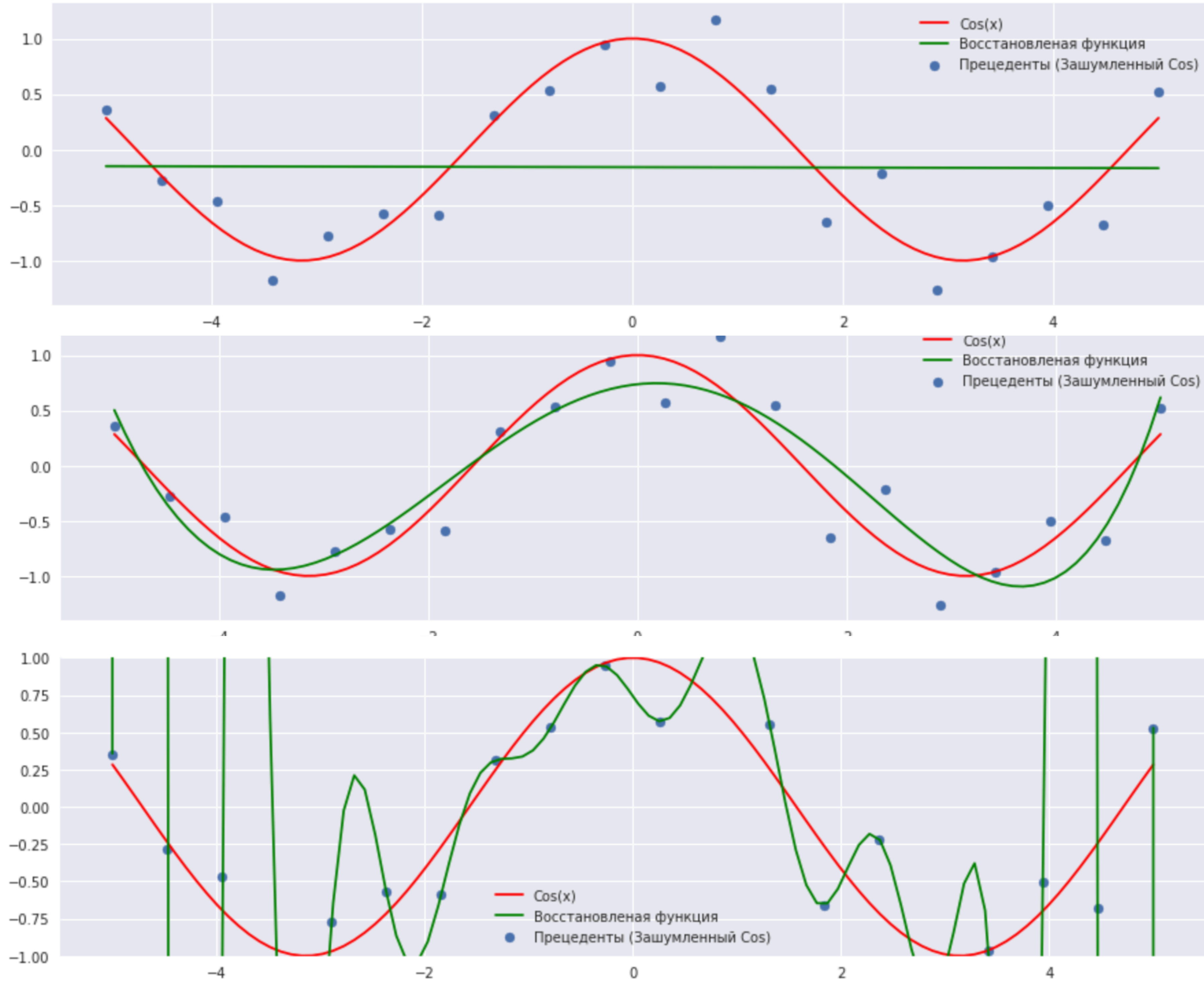
Средняя абсолютная (Mean Absolute Error, MAE) ошибка:

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

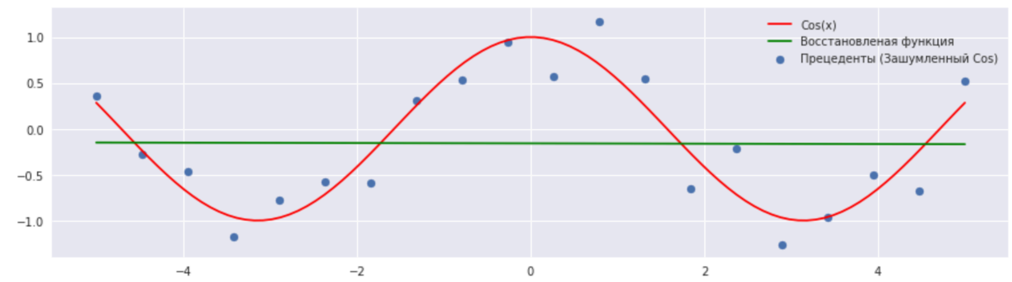
Среднеквадратичный функционал сильнее штрафует за большие отклонения по сравнению со среднеабсолютным, и поэтому более чувствителен к выбросам.

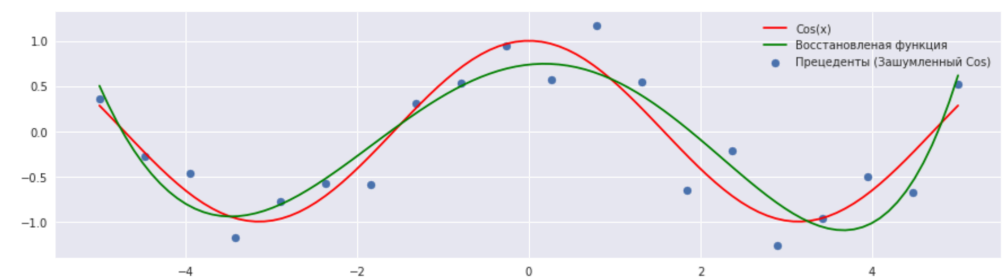
Идеальны для сравнения моделей, но не всегда понятно как их оценивать относительно целевой переменной. Например: MSE - 10 это хорошо, если переменная лежит в пределах интервала (10000, 100000), но плохо, если целевая переменная принимает значения от 0 до 1

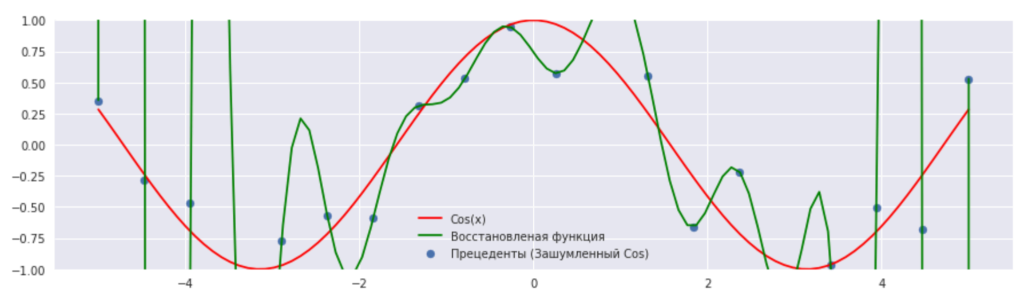
Мотивация валидации



Мотивация валидации

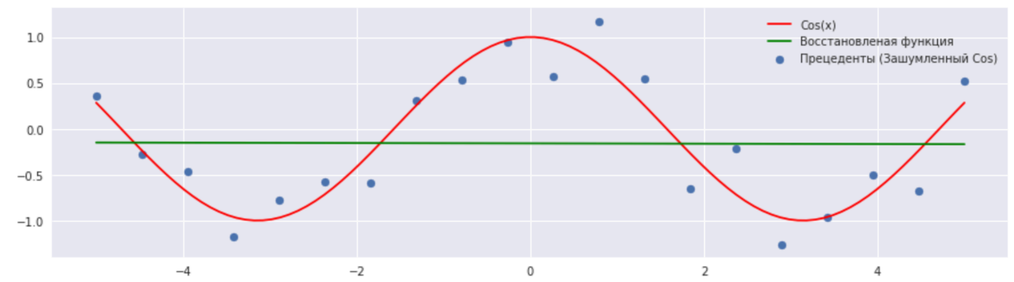


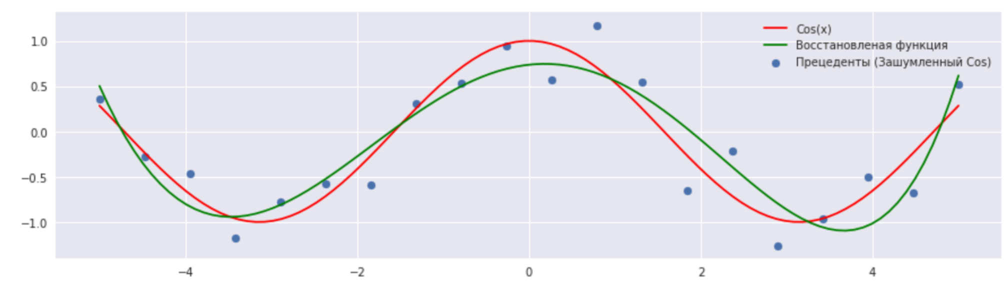


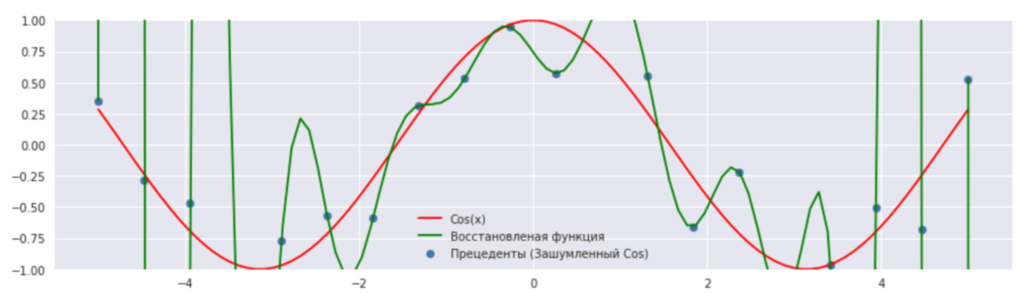


	MSE	MAE	R2
Линейная модель	0.472	0.586	0.0004
Полином 5-ой степени	0.047	0.179	0.9000
Полином 11-ой степени	0.000	0.000	1.0000

Мотивация валидации



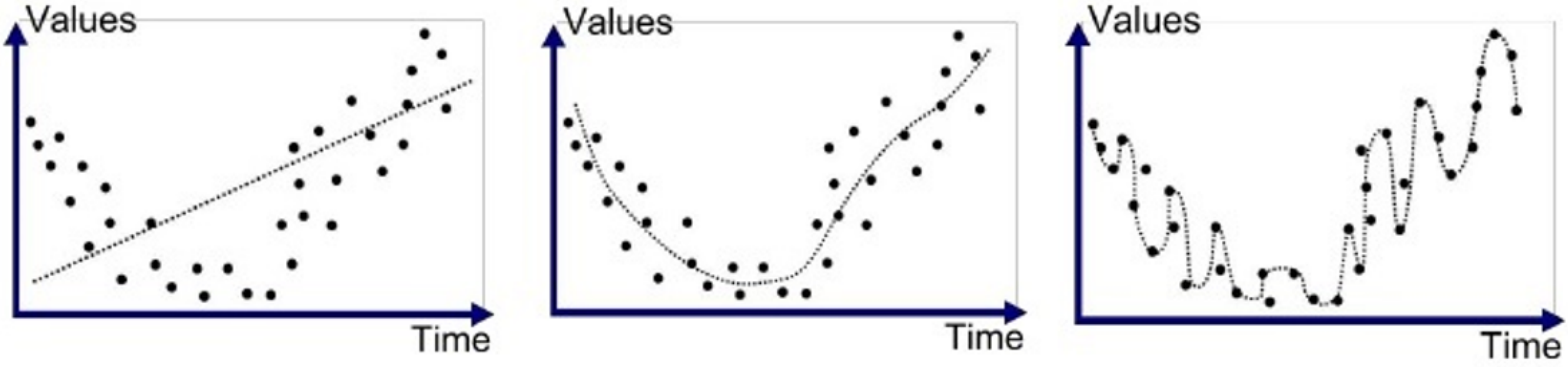


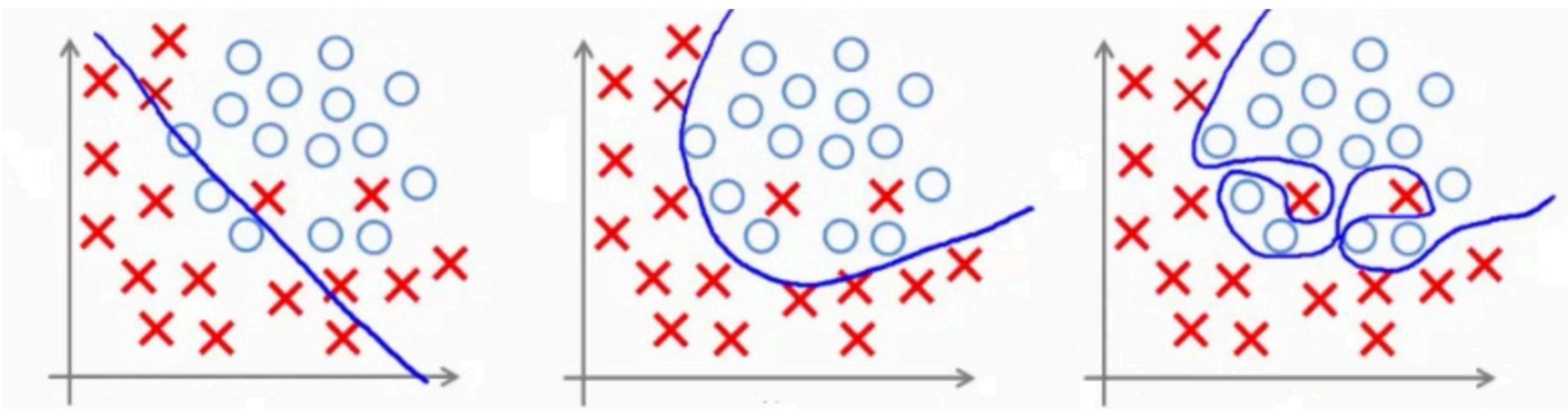


	MSE	MAE	R2
Линейная модель	0.472	0.586	0.0004
Полином 5-ой степени	0.047	0.179	0.9000
Полином 11-ой степени	0.000	0.000	1.0000

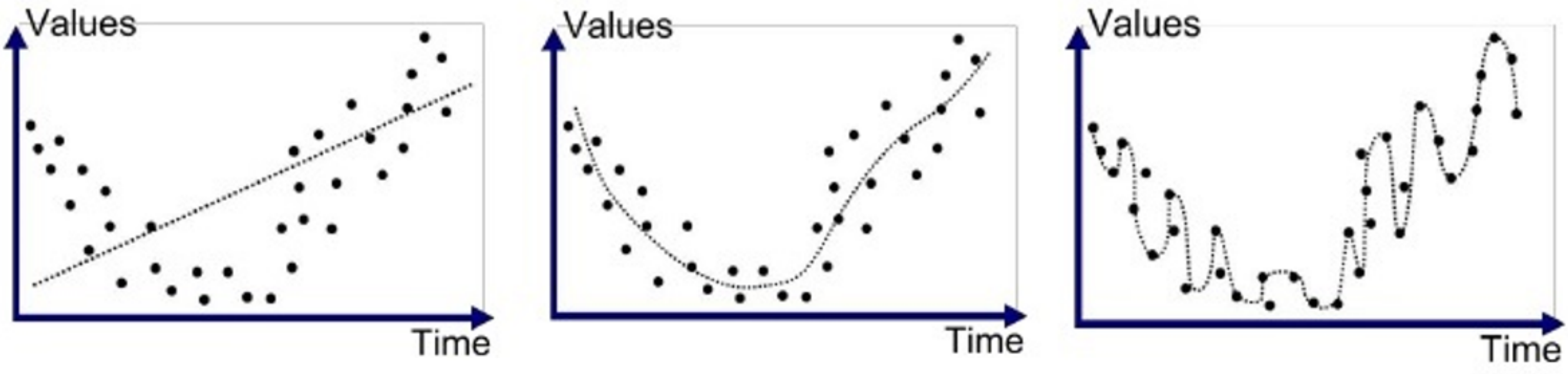
Обобщающая способность Полинома 11-ой степени отсутствует.

Мотивация валидации

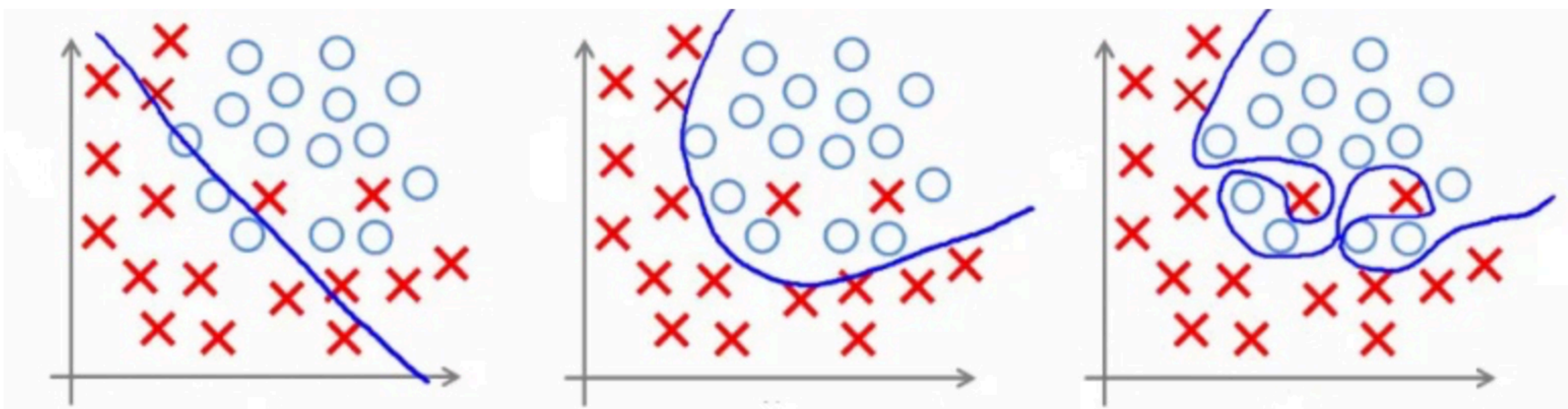




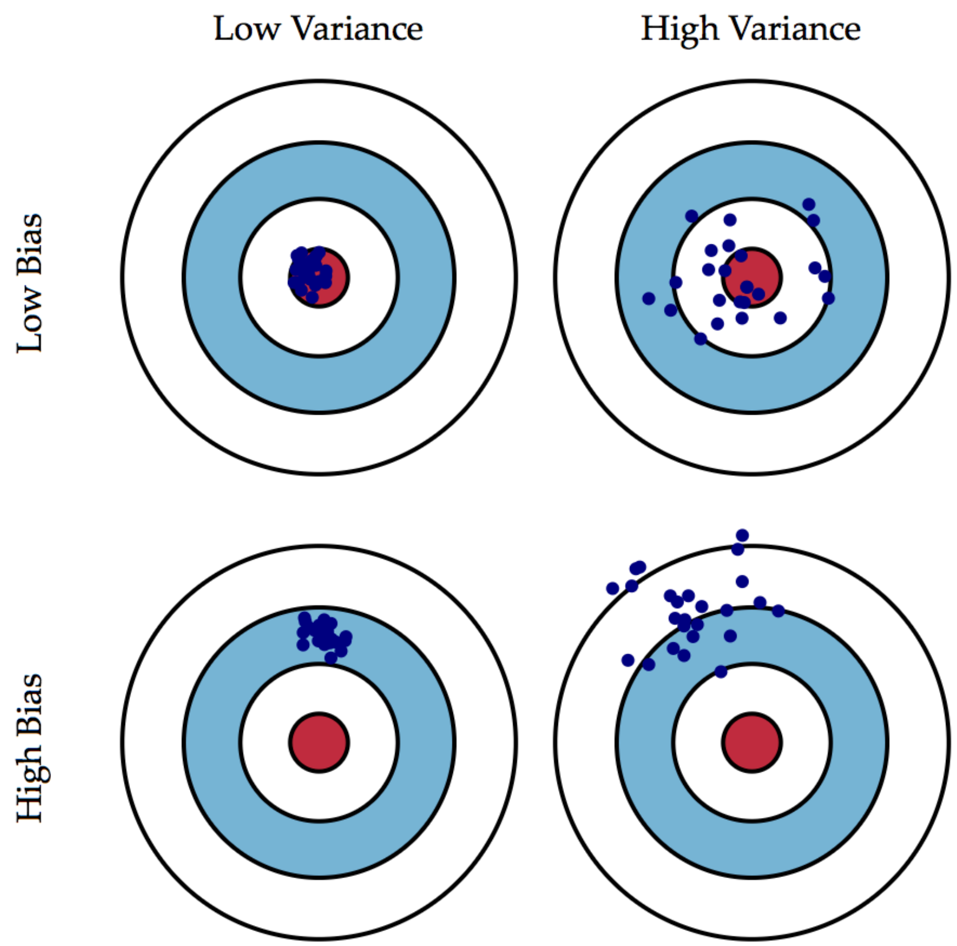
Мотивация валидации

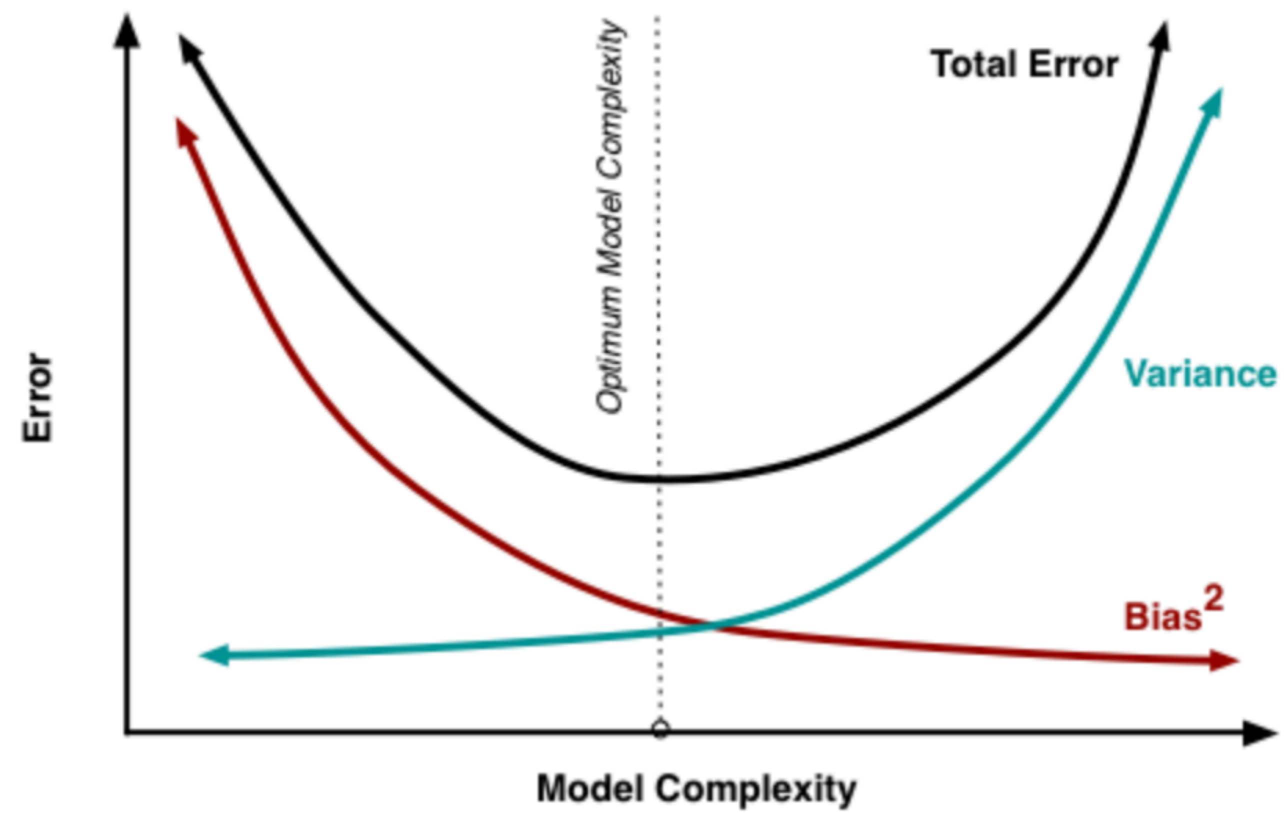


Underfitted Good Fit/Robust Overfitted



Bias and Variance tradeoff



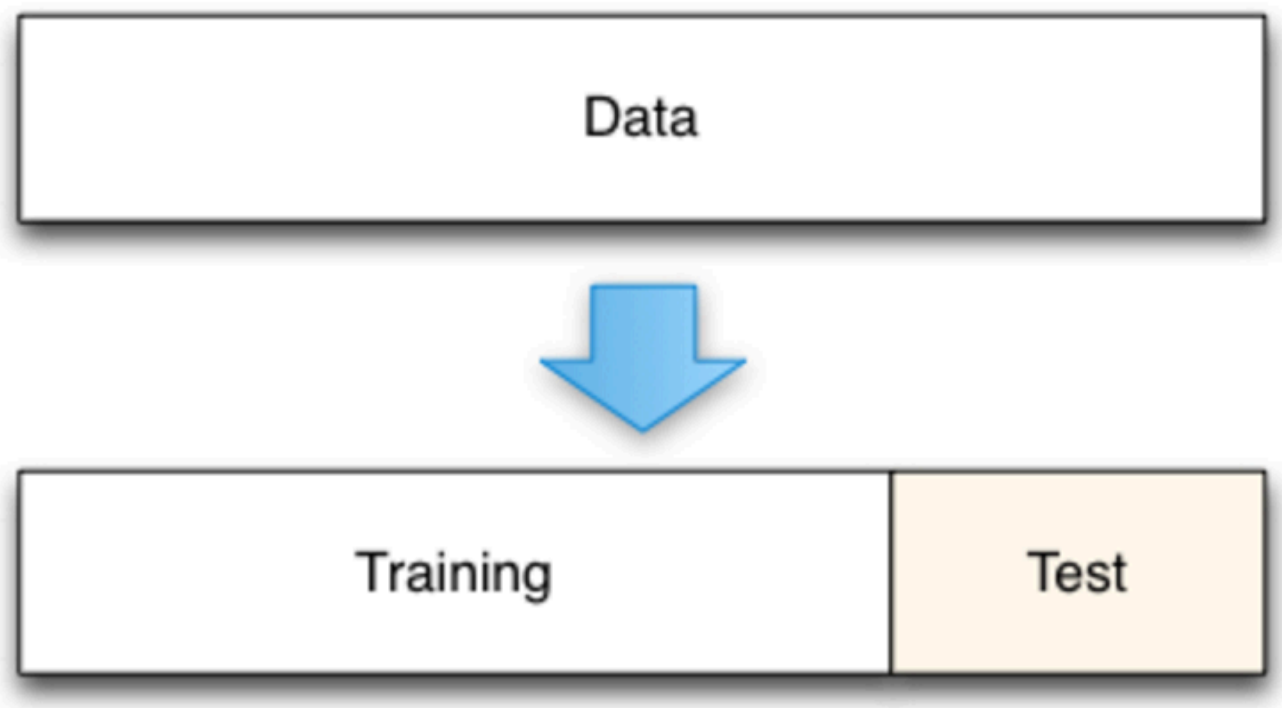


$$Err(x) = E\left[(Y - \hat{f}(x))^2\right]$$

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

http://scott.fortmann-roe.com/docs/BiasVariance.html

Стратегии валидации



Стратегии валидации

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

$$f(x)\colon \mathbb{X} \to \mathbb{Y}$$

Основная гипотеза МО: Схожим объектам соответствуют схожие объекты

Метод k ближайших соседей
(**kNN**, ***k Nearest Neighbours***)

<u>Гипотеза компактности:</u> если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных.

- Вычислить расстояние до каждого из объектов обучающей выборки
- Отобрать k объектов обучающей выборки, расстояние до которых минимально
- Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей

`sklearn.neighbors`.KNeighborsRegressor

(*n_neighbors=5*, *weights='uniform'*, *algorithm='auto'*, *leaf_size=30*, *p=2*, *metric='minkowski'*, *metric_params=None*, *n_jobs=None*, ***kwargs*)

`sklearn.neighbors`.KNeighborsClassifier

(*n_neighbors=5*, *weights='uniform'*, *algorithm='auto'*, *leaf_size=30*, *p=2*, *metric='minkowski'*, *metric_params=None*, *n_jobs=None*, ***kwargs*)

Метод k ближайших соседей
(**kNN**, *k Nearest Neighbours*)

<u>Гипотеза компактности:</u> если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных.

- Вычислить расстояние до каждого из объектов обучающей выборки
- Отобрать k объектов обучающей выборки, расстояние до которых минимально
- Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей

`sklearn.neighbors`**.KNeighborsRegressor**

(*n_neighbors=5*, *weights='uniform'*, *algorithm='auto'*, *leaf_size=30*, *p=2*, *metric='minkowski'*, *metric_params=None*, *n_jobs=None*, ***kwargs*)

`sklearn.neighbors`**.KNeighborsClassifier**

(*n_neighbors=5*, *weights='uniform'*, *algorithm='auto'*, *leaf_size=30*, *p=2*, *metric='minkowski'*, *metric_params=None*, *n_jobs=None*, ***kwargs*)

Выбор числа соседей k

При $k=1$ алгоритм ближайшего соседа неустойчив к шумовым выбросам: он даёт ошибочные классификации не только на самих объектах-выбросах, но и на ближайших к ним объектах других классов.

При $k=l$, наоборот, алгоритм чрезмерно устойчив и вырождается в константу. Таким образом, крайние значения нежелательны.

Метод k ближайших соседей
(**kNN,** *k Nearest Neighbours*)

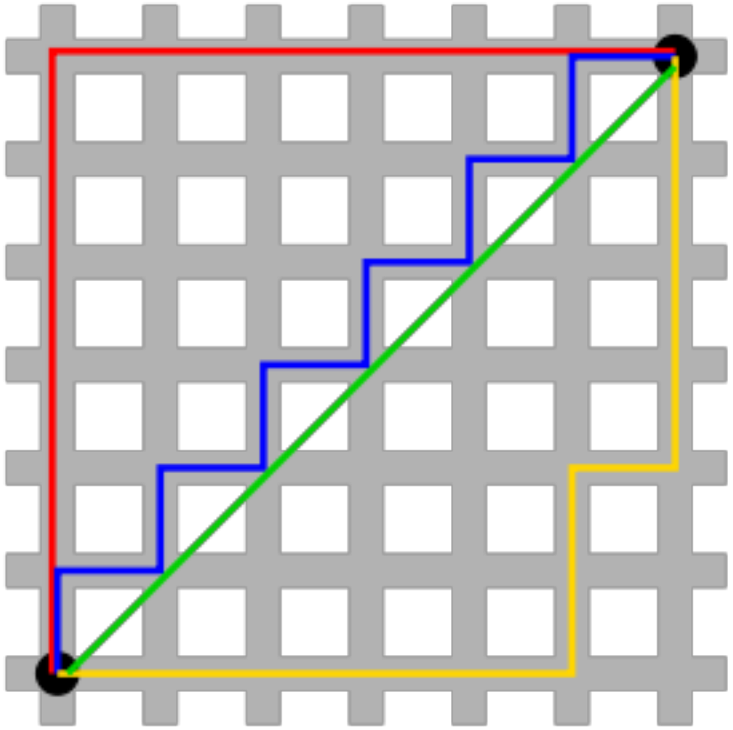
Выбор метрики

Евклидово расстояние ("euclidean")

$$\sqrt{\sum (x-y)^2}$$

Расстояние городских кварталов «манхэттенское расстояние» ("manhattan")

$$\sum |x-y|$$



Расстояние Чебышева "chebyshev"

$$max(x-y)$$

Растояние Минковского "minkowski"

$$\left(\sum |x-y|^p\right)^{\frac{1}{p}}$$

расстояния с параметром p равным 1 (расстояние городских кварталов) или 2 (евклидова метрика).
p = ∞ метрика обращается в расстояние Чебышёва.

Метод k ближайших соседей
(**kNN,** *k Nearest Neighbours*)

Нормирование признаков

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

$$z \in [0,1]$$

Стандартизация признаков

$$z = \frac{x - \mu}{\mu - \sigma}$$

где

μ - Среднее

σ - Стандартное отклонение

Проклятие размерности

В пространстве высокой размерности все объекты примерно одинаково далеки друг от друга; выбор ближайших соседей становится практически произвольным.

Ссылки

1. Статья на habr: Метрики в задачах машинного обучения
2. Семинар из курса Евгения Соколова