



# **Лекция 5: непараметрическая регрессия**

Евгений Борисов

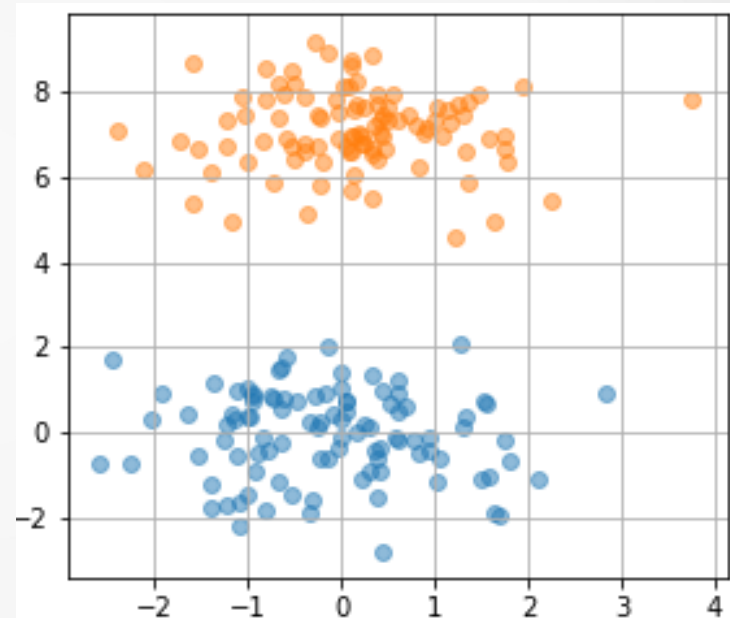
# метрические методы : регрессия

классификация - задача разделения объектов на классы

$X \subset \mathbb{R}^n$  - объекты

$Y \in \{0, 1\}$  - метки классов

$a: X \rightarrow Y$



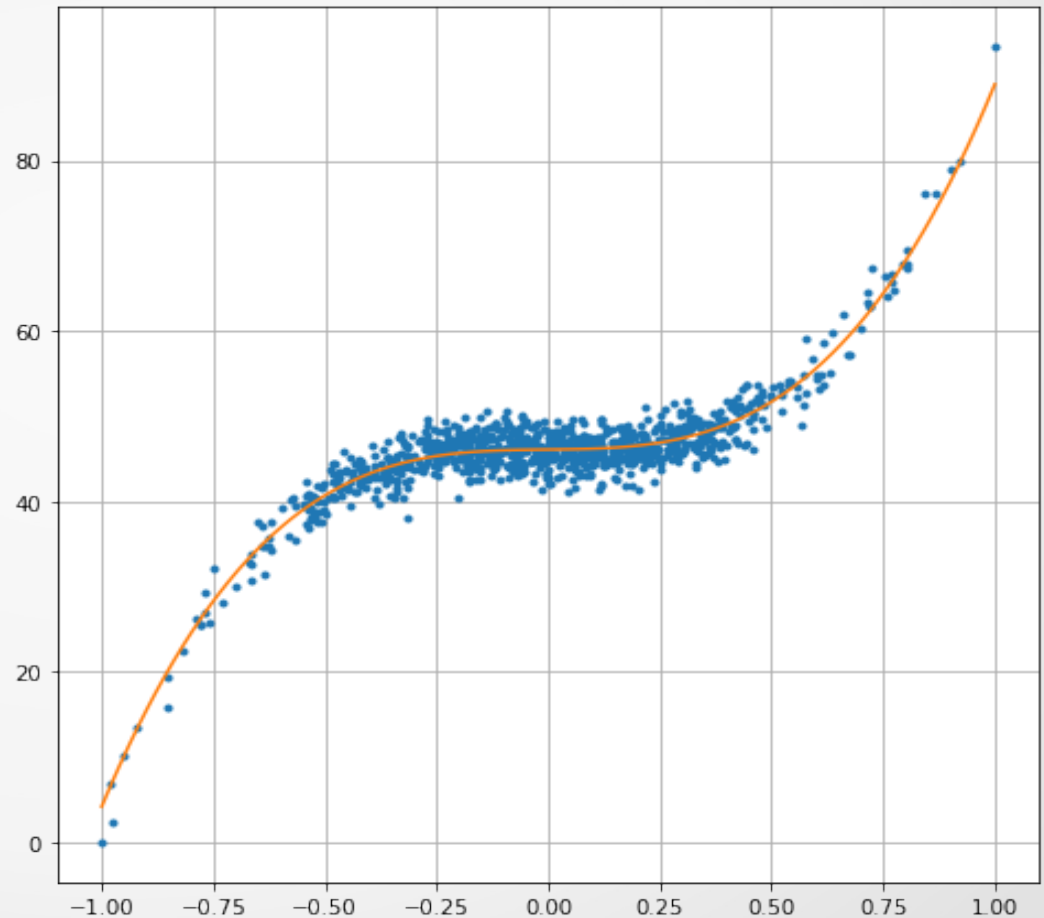
# метрические методы : регрессия

регрессия-задача восстановления зависимости

$X \subset \mathbb{R}^n$  - объекты

$Y \subset \mathbb{R}$  - ответы

$$a: X \rightarrow Y$$



# метрические методы : регрессия

Оценка недвижимости по статистике продаж

цена = **оценка**(  
    район,  
    площадь,  
    этаж,  
    лифт,  
    ремонт,  
    )

# метрические методы : регрессия

$X \subset \mathbb{R}^n$  - объекты

$Y \subset \mathbb{R}$  - ответы

регрессия - задача восстановления зависимости

$$a: X \rightarrow Y$$

# метрические методы : регрессия

регрессия - задача восстановления зависимости

$a: X \rightarrow Y$        $X \subset \mathbb{R}^n$  - объекты       $Y \subset \mathbb{R}$  - ответы

параметрический подход: определяем (допускаем) вид зависимости

$$a = f(x, \theta)$$

... и подбираем параметры решая задачу оптимизации  
(метод наименьших квадратов)

$$Q(\theta, X) = \sum_{i=1}^m (f(x_i, \theta) - y_i)^2 \rightarrow \min_{\theta}$$

# метрические методы : регрессия

регрессия - задача восстановления зависимости

$a: X \rightarrow Y$        $X \subset \mathbb{R}^n$  - объекты       $Y \subset \mathbb{R}$  - ответы

параметрический подход: определяем (допускаем) вид зависимости

$$a = f(x, \theta)$$

... и подбираем параметры решая задачу оптимизации  
(метод наименьших квадратов)

$$Q(\theta, X) = \sum_{i=1}^m (f(x_i, \theta) - y_i)^2 \rightarrow \min_{\theta}$$

недостаток: не вполне понятно как определять вид зависимости

# метрические методы : регрессия

регрессия - задача восстановления зависимости

$a: X \rightarrow Y$        $X \subset \mathbb{R}^n$  - объекты       $Y \subset \mathbb{R}$  - ответы

Непараметрический подход: приближение  $\theta$  в окрестности точки  $u$

$$Q(\theta, X) = \sum_{i=1}^m w_i(u) \cdot (\theta - y_i)^2 \rightarrow \min_{\theta}$$

$w_i(u)$  - вес объекта  $x_i$  относительно  $u$   
убывает при увеличении расстояния



# метрические методы : регрессия

регрессия - задача восстановления зависимости

$a: X \rightarrow Y$        $X \subset \mathbb{R}^n$  - объекты       $Y \subset \mathbb{R}$  - ответы

Непараметрический подход: приближение  $\theta$  в окрестности точки  $u$

$$Q(\theta, X) = \sum_{i=1}^m w_i(u) \cdot (\theta - y_i)^2 \rightarrow \min_{\theta}$$

$w_i(u) = K\left(\frac{\rho(u, x_i)}{h}\right)$  - вес объекта  $x_i$  относительно  $u$   
убывает при увеличении расстояния

$K(r)$  - функция ядра

$h$  - параметр ширина окна сглаживания

# метрические методы : регрессия

регрессия - задача восстановления зависимости

$a: X \rightarrow Y$        $X \subset \mathbb{R}^n$  - объекты       $Y \subset \mathbb{R}$  - ответы

оптимизируем  $Q$  ...

$$Q(\theta, X) = \sum_{i=1}^m w_i(u) \cdot (\theta - y_i)^2 \rightarrow \min_{\theta}$$

$$\frac{\partial Q}{\partial \theta} = 0$$

# метрические методы : регрессия

регрессия - задача восстановления зависимости

$a: X \rightarrow Y$        $X \subset \mathbb{R}^n$  - объекты       $Y \subset \mathbb{R}$  - ответы

оптимизируем  $Q$  ...

$$Q(\theta, X) = \sum_{i=1}^m w_i(u) \cdot (\theta - y_i)^2 \rightarrow \min_{\theta}$$
$$\frac{\partial Q}{\partial \theta} = 0$$

... получаем

формулу Надарая-Ватсона

$$a(u, X) = \frac{\sum_{i=1}^m y_i \cdot w_i(u)}{\sum_{i=1}^m w_i(u)} = \frac{\sum_{i=1}^m y_i \cdot K\left(\frac{\rho(u, x_i)}{h}\right)}{\sum_{i=1}^m K\left(\frac{\rho(u, x_i)}{h}\right)}$$

$w_i(u)$  - вес объекта  $x_i$  относительно  $u$

$K(r)$  - функция ядра

$h$  - параметр ширина окна сглаживания

# метрические методы : регрессия

## метрика - функция расстояния

$$\rho: X \times X \rightarrow [0, \infty)$$

аксиома тождества :  $\rho(x, y) = 0 \Leftrightarrow x = y$

симметрия:  $\rho(x, y) = \rho(y, x)$

неравенство треугольника:  $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$

# метрические методы : регрессия

## метрика - функция расстояния

Евклидова метрика:  $\rho(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$

метрика Минковского:  $\rho(x, y) = \sqrt[n]{\sum_i w_i |x_i - y_i|^n}$

метрика Чебышева:  $\rho(x, y) = \max_i |x_i - y_i|$

косинусная метрика:  $\rho(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$

# метрические методы : регрессия

ядро - неотрицательная интегрируемая функция

$$\int_{-\infty}^{+\infty} K(r) dr = 1$$

$$K(r) = K(-r)$$

прямоугольное

$$\Pi(r) = [|r| \leq 1]$$

треугольное

$$T(r) = (1 - |r|)[|r| \leq 1]$$

квадратичное  
(Епанечникова)

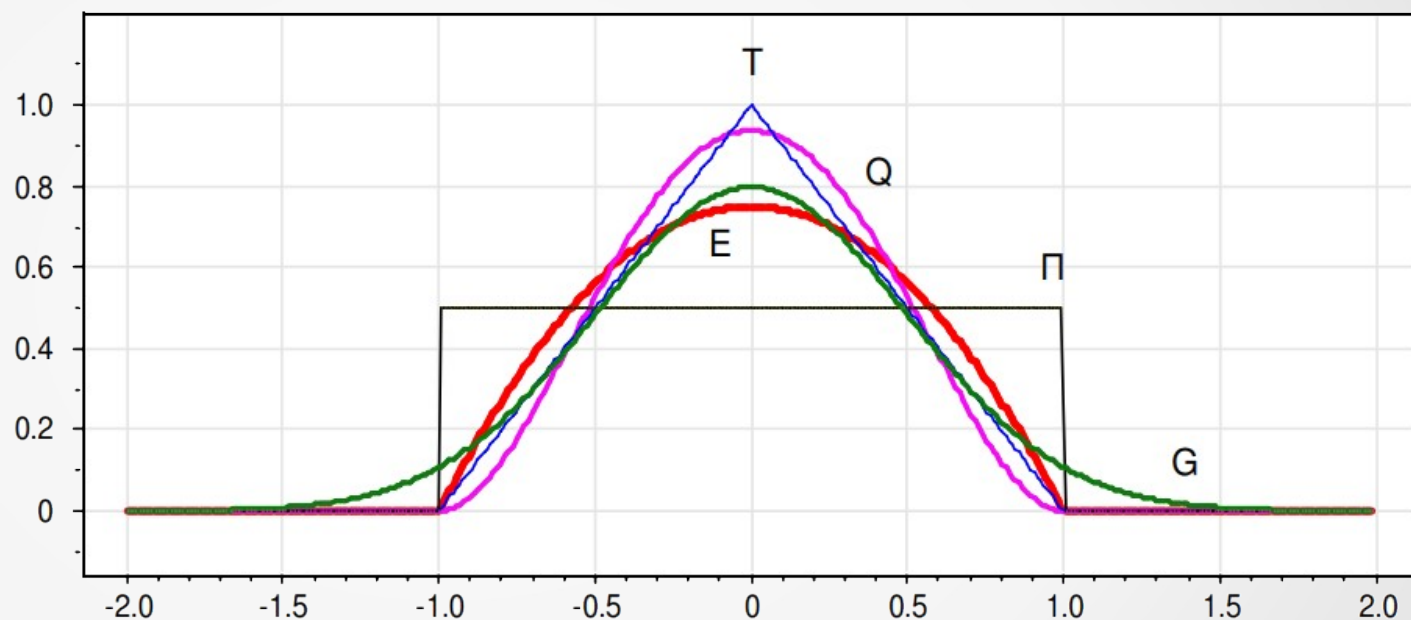
$$E(r) = (1 - r^2)[|r| \leq 1]$$

квартическое

$$Q(r) = (1 - r^2)^2[|r| \leq 1]$$

гауссово

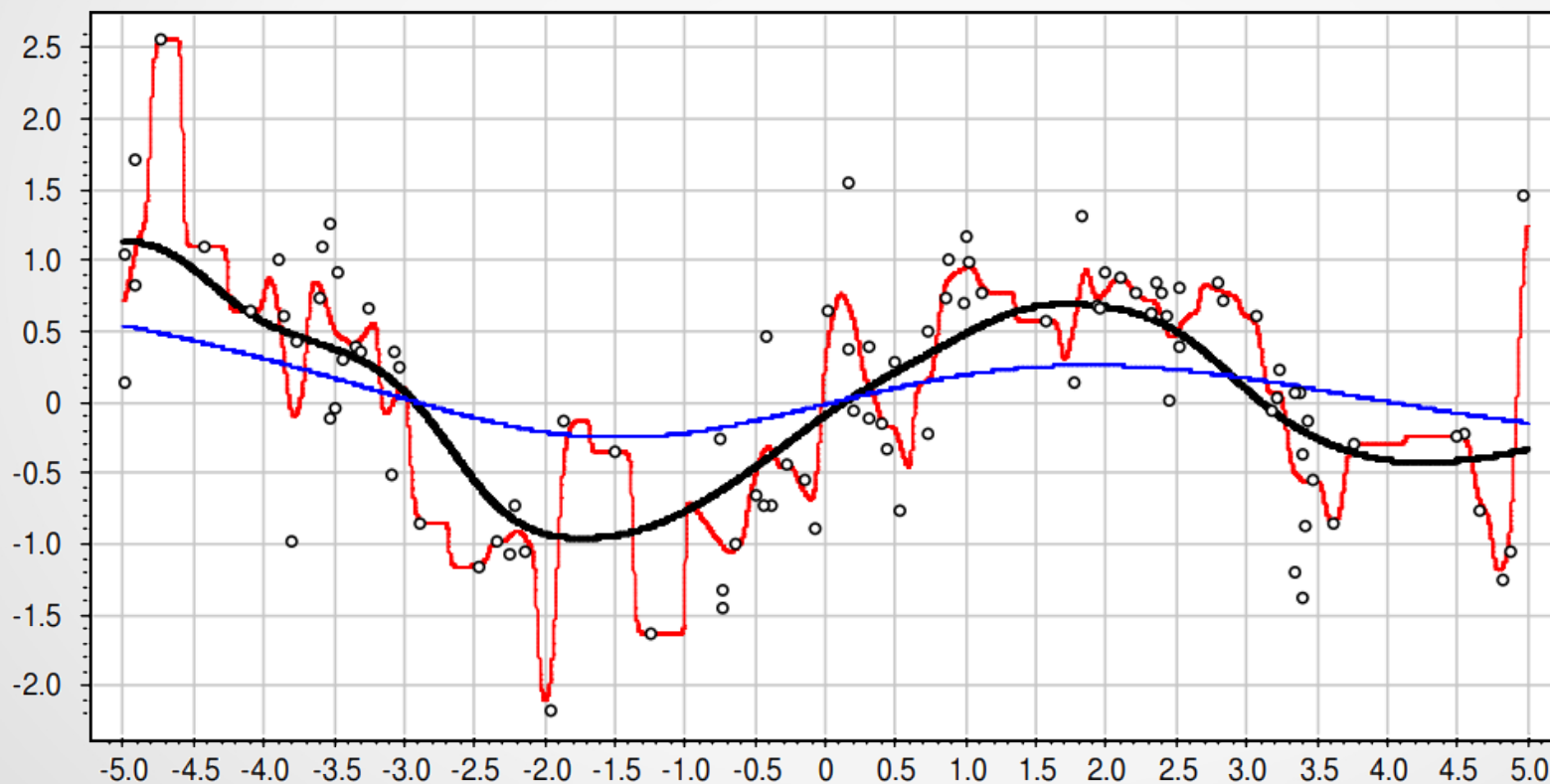
$$G(r) = \exp(-2r^2)$$



# метрические методы : регрессия

## выбор ядра и ширины окна сглаживания

$h \in \{0.1, 1.0, 3.0\}$ , гауссовское ядро  $K(r) = \exp(-2r^2)$



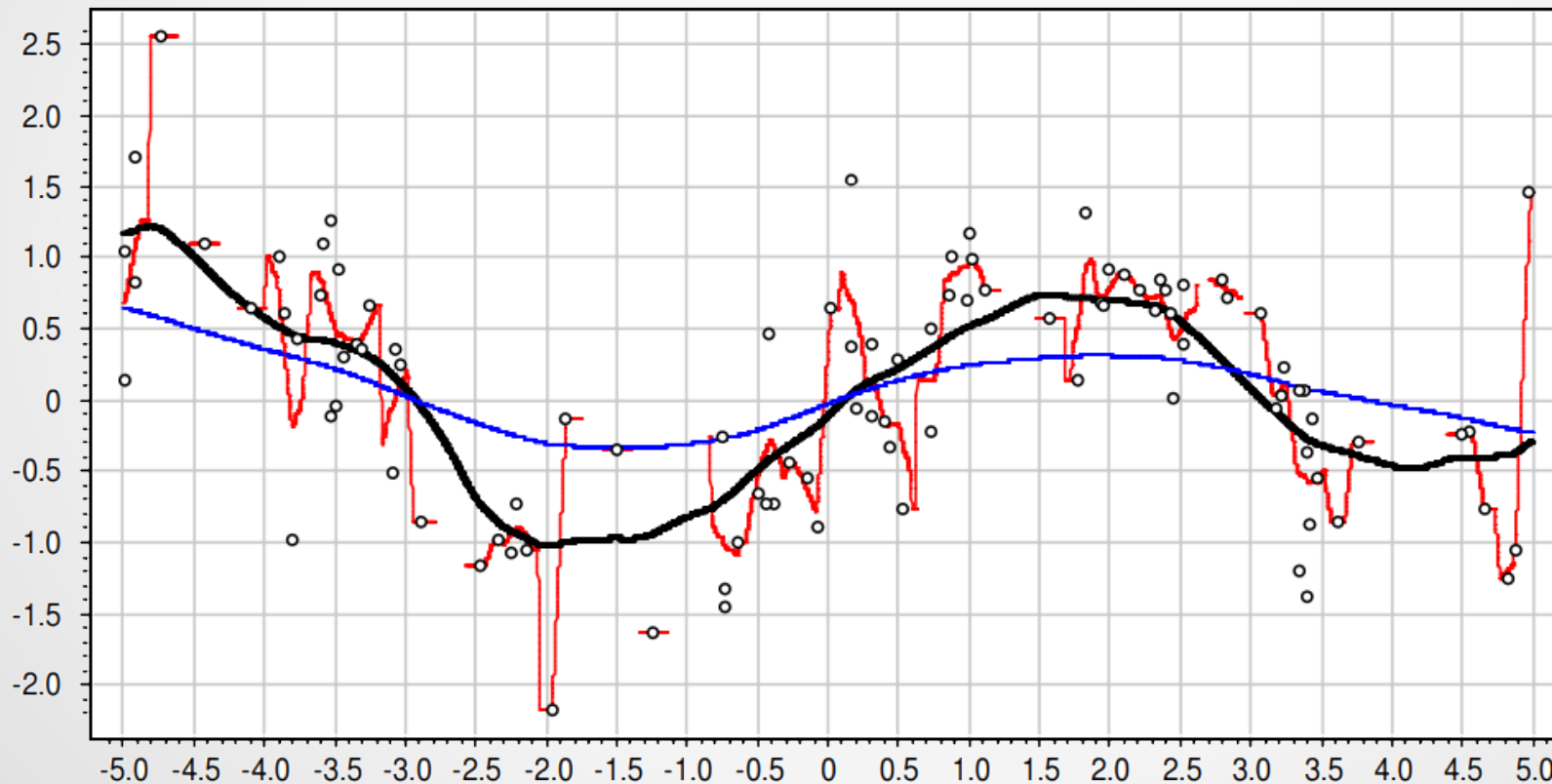
Гауссовское ядро  $\Rightarrow$  гладкая аппроксимация

Ширина окна существенно влияет на точность аппроксимации

# метрические методы : регрессия

## выбор ядра и ширины окна сглаживания

$h \in \{0.1, 1.0, 3.0\}$ , треугольное ядро  $K(r) = (1 - |r|) [ |r| \leq 1 ]$



Треугольное ядро  $\Rightarrow$  кусочно-линейная аппроксимация

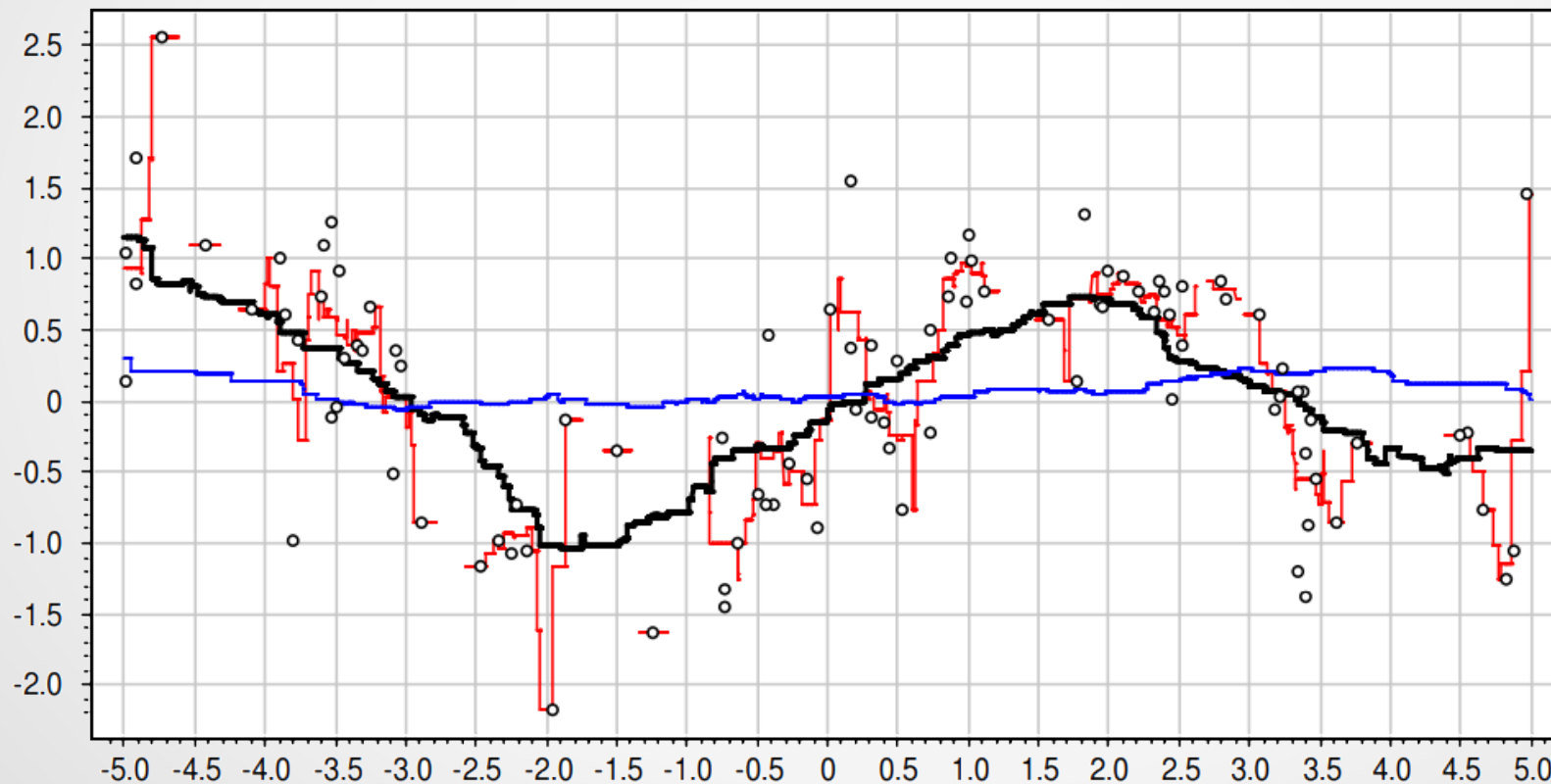
Аппроксимация не определена, если в окне нет точек выборки



# метрические методы : регрессия

## выбор ядра и ширины окна сглаживания

$h \in \{0.1, 1.0, 3.0\}$ , прямоугольное ядро  $K(r) = [ |r| \leq 1 ]$



Прямоугольное ядро  $\Rightarrow$  кусочно-постоянная аппроксимация  
Выбор ядра слабо влияет на точность аппроксимации

# метрические методы : регрессия

ядро  $K(r)$  влияет на гладкость функции  $a(x)$

ширина окна  $h$  влияет на качество аппроксимации  $a(x)$

# метрические методы : регрессия

ядро  $K(r)$  влияет на гладкость функции  $a(x)$

ширина окна  $h$  влияет на качество аппроксимации  $a(x)$

**выбор  $h$  ширины окна сглаживания**

метод скользящего контроля (Leave One Out, LOO)

параметр  $h$  выбираем перебором

проверяем суммарную ошибку на учебном множестве

из учебного набора удаляется текущий (проверяемый) пример

$$LOO(h, X) = \sum_{i=1}^m \left( a(x_i, \{X \setminus x_i\}, h) - y_i \right)^2 \rightarrow \min_h$$

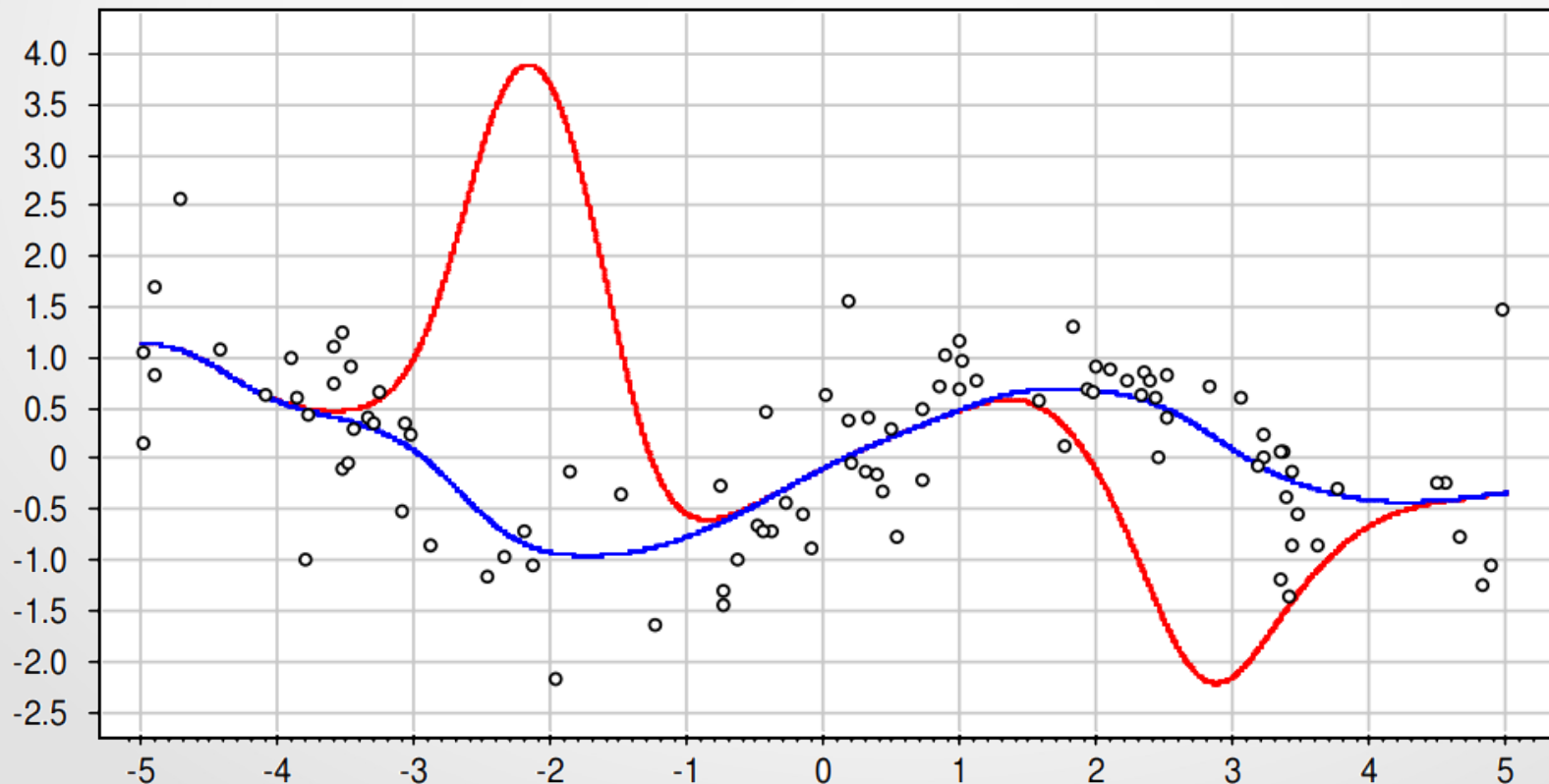
# метрические методы : регрессия

## влияние выбросов

$\ell = 100$ ,  $h = 1.0$ , гауссовское ядро  $K(r) = \exp(-2r^2)$

Две из 100 точек — выбросы с ординатами  $y_i = 40$  и  $-40$

Синяя кривая — выбросов нет



# метрические методы : регрессия

## коррекция влияния выбросов

локально взвешенное сглаживание

$\varepsilon_i = |a(x_i, \{X \setminus x_i\}, h) - y_i|$  - ошибка на учебном примере  $i$

$\gamma_i = \tilde{K}(\varepsilon_i)$  - корректирующий коэффициент,  
чем больше ошибка тем меньше вес,  
(выбираем другое ядро, отличное от  $K$ )

# метрические методы : регрессия

алгоритм LOWESS (locally weighted scatter plot smoothing):  
определяем корректирующие коэффициенты

1. инициализация  $y_i := 1; i = 1, \dots, m$

2. вычисляем оценку скользящего контроля

$$a_i = a(x_i, \{X \setminus x_i\}, h) = \frac{\sum_{j=1, i \neq j}^m y_j y_j \cdot K\left(\frac{\rho(x_i, x_j)}{h}\right)}{\sum_{j=1, i \neq j}^m K\left(\frac{y_j \rho(x_i, x_j)}{h}\right)}; i = 1, \dots, m$$

3. пересчитываем корректирующие коэффициенты

$$y_i := \tilde{K}(|a_i - y_i|); i = 1, \dots, m$$

4. если корректирующие коэффициенты существенно изменились  
то переход на п.2  
иначе конец работы

# метрические методы : регрессия



**Вопросы ?**

# метрические методы : регрессия

git clone [https://github.com/mechanoid5/ml\\_lectorium.git](https://github.com/mechanoid5/ml_lectorium.git)

К.В. Воронцов Метрические методы классификации. - курс  
"Машинное обучение" ШАД Яндекс 2014

К.В. Воронцов Методы восстановления регрессии - курс  
"Машинное обучение" ШАД Яндекс 2014

Борисов Е.С. Модели математической регрессии  
<http://mechanoid.kiev.ua/ml-regression.html>

Формула Надарая-Ватсона  
<http://www.machinelearning.ru/wiki>



# метрические методы : регрессия

## источники данных для экспериментов



sklearn.datasets  
UCI Repository  
kaggle



## задание

выбрать данные в репозитории UCI  
( <https://archive.ics.uci.edu/ml> )

реализовать для них регрессию

реализовать LOO для выбора размера окна сглаживания

реализовать LOWESS для коррекции выбросов