Лекция 9: линейная и нелинейная регрессия

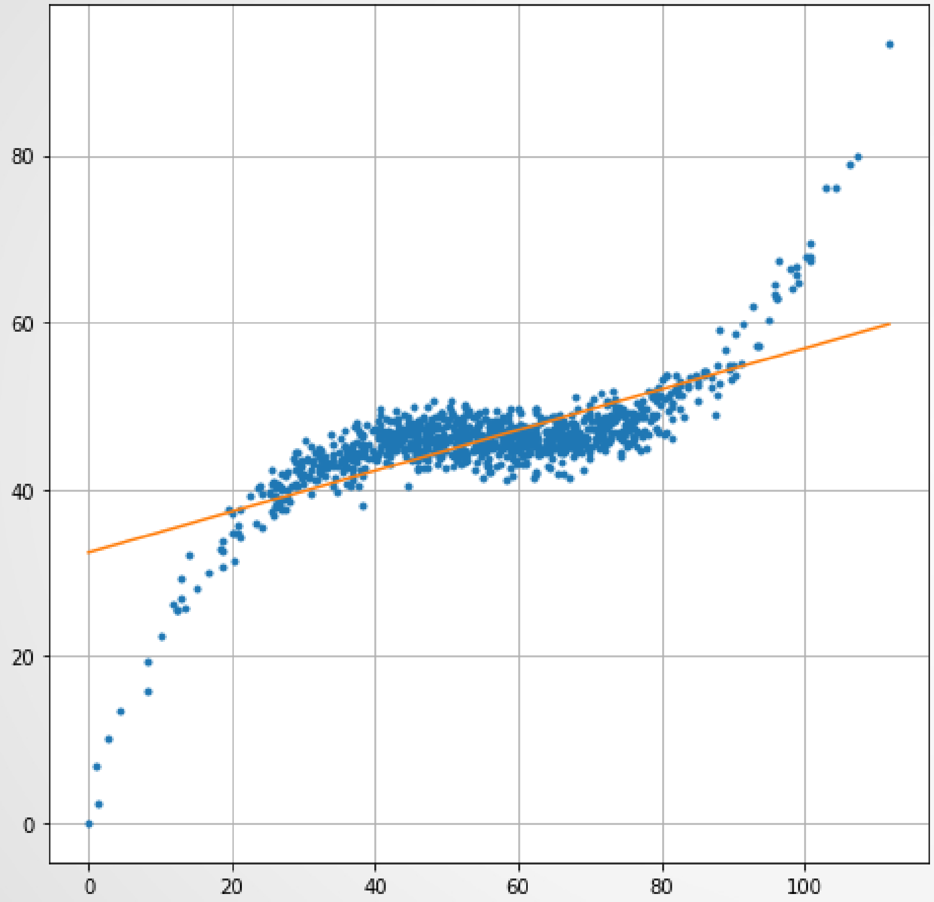
Евгений Борисов

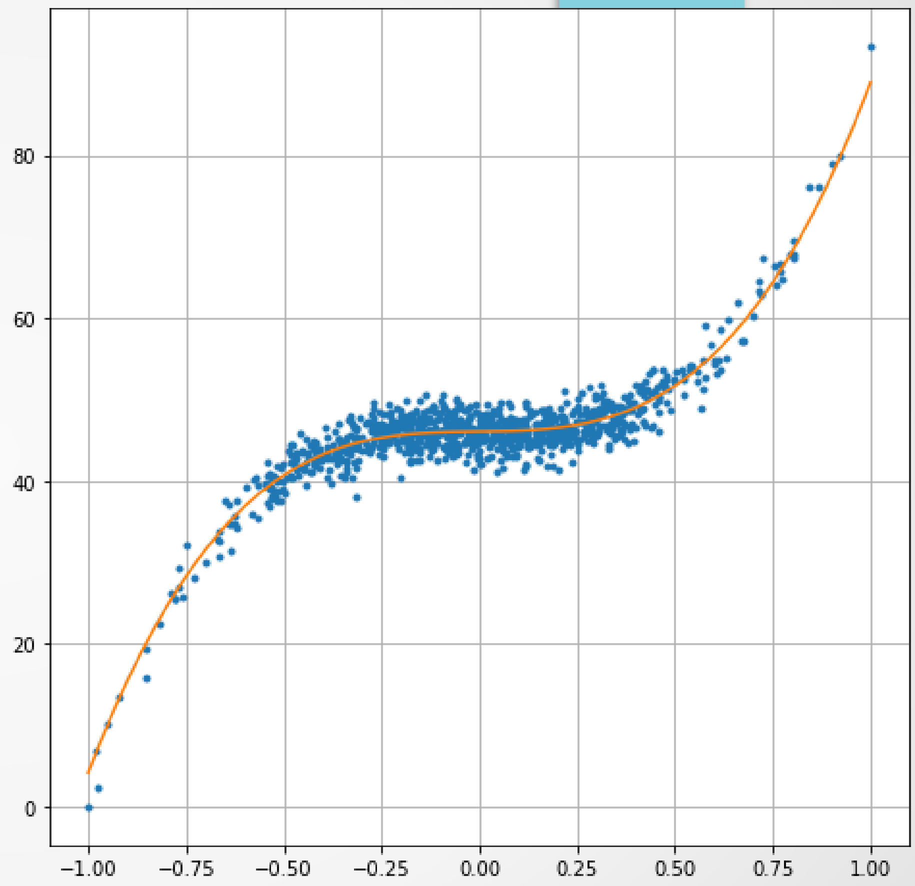
линейная и нелинейная регрессия

Оценка недвижимости по статистике продаж

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

линейная и нелинейная регрессия





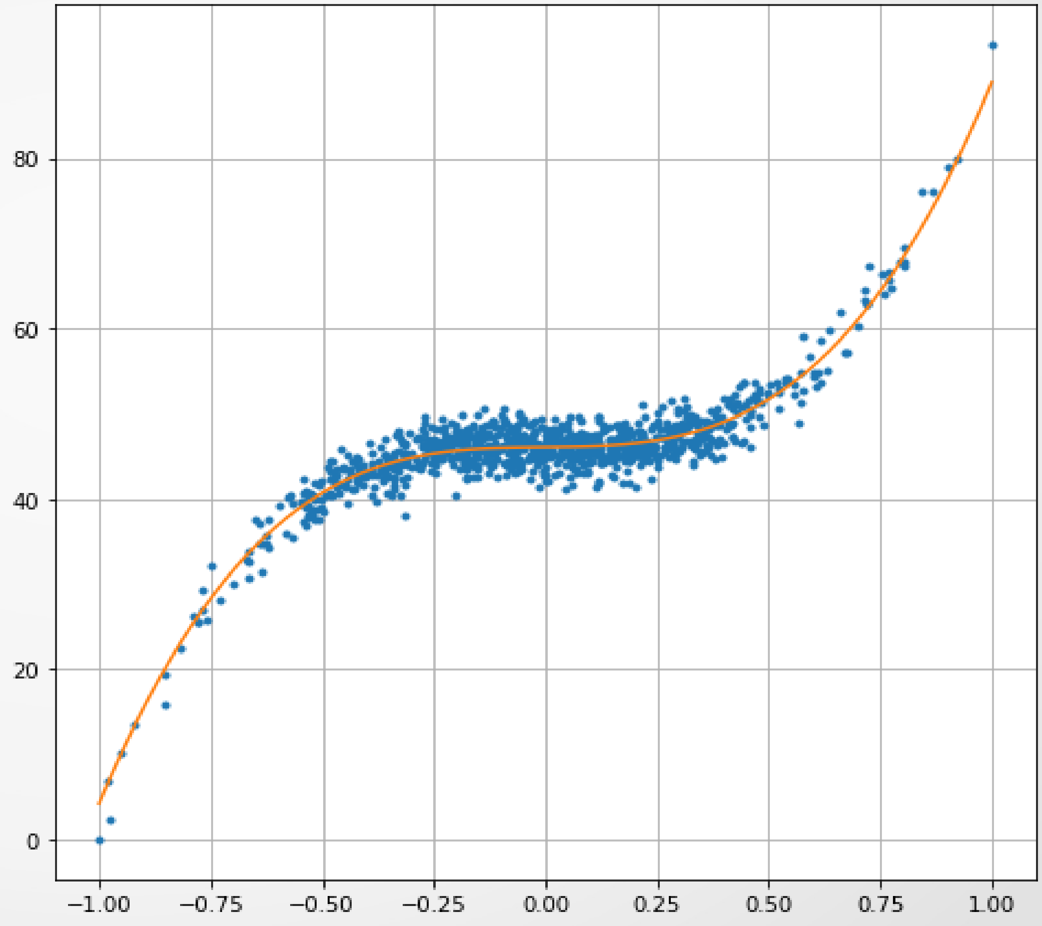
линейная и нелинейная регрессия

регрессия-задача восстановления зависимости

$X \subset \mathbb{R}^n$ - объекты

$Y \subset \mathbb{R}$ - ответы

$$a : X \rightarrow Y$$



линейная и нелинейная регрессия

<u>регрессия</u> - задача восстановления зависимости

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

<u>параметрический подход</u>:определяем (допускаем) вид зависимости

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

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

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

линейная и нелинейная регрессия

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2n} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} & y_m \end{bmatrix}$$

x - вектор-признак
y - ответ (значение функции)
n - размер пространства признаков
m - количество примеров

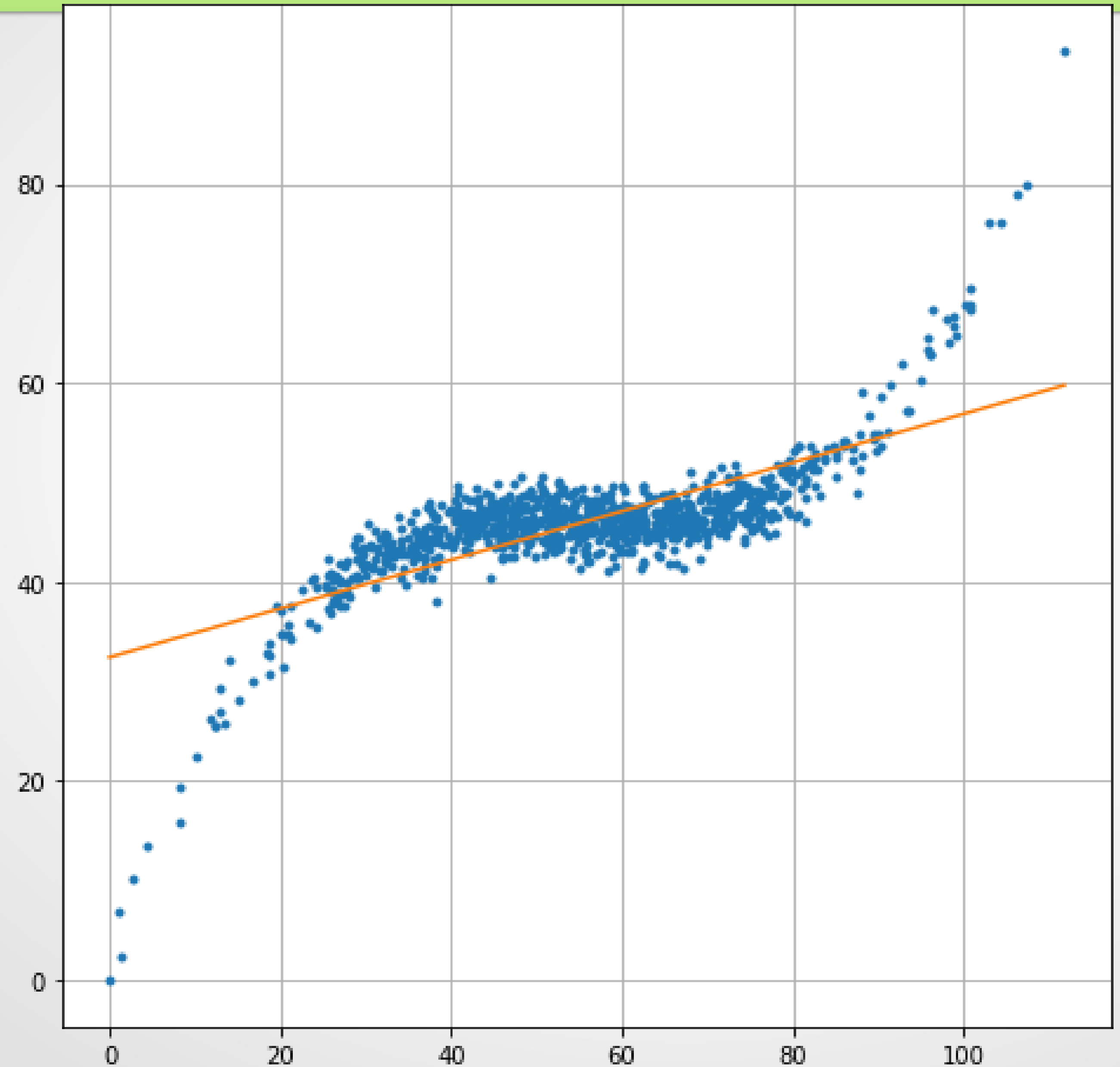
линейная и нелинейная регрессия

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$y = h(x, \theta) = \theta \cdot X_i = \theta_0 + \theta_1 \cdot x_1 + \ldots + \theta_n \cdot x_n$$

x - вектор-признак
y - ответ (значение функции)
n - размер пространства признаков
m - количество примеров
θ - параметры

линейная и нелинейная регрессия



линейная и нелинейная регрессия

$$\theta=\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix}\qquad X=\begin{bmatrix}1 & x_{11} & x_{12} & \cdots & x_{1n}\\ 1 & x_{21} & x_{22} & \cdots & x_{2n}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn}\end{bmatrix}\qquad y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_m\end{bmatrix}$$

$$y=h(x,\theta)=\theta\cdot X_i=\theta_0+\theta_1\cdot x_1+\ldots+\theta_n\cdot x_n$$

$$\theta=?$$

линейная и нелинейная регрессия

$$\theta=\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix}\qquad X=\begin{bmatrix}1 & x_{11} & x_{12} & \cdots & x_{1n}\\ 1 & x_{21} & x_{22} & \cdots & x_{2n}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn}\end{bmatrix}\qquad y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_m\end{bmatrix}$$

$$y=h(x,\theta)=\theta\cdot X_i=\theta_0+\theta_1\cdot x_1+\ldots+\theta_n\cdot x_n$$

<u>метод наименьших квадратов:</u>
минимизация суммы квадратов отклонений
функции от искомых переменных

$$\theta=\left(X^T\cdot X\right)^{-1}\cdot X^T\cdot y$$

линейная и нелинейная регрессия

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$y = h(x, \theta) = \theta \cdot X_i = \theta_0 + \theta_1 \cdot x_1 + \ldots + \theta_n \cdot x_n$$

<u>метод наименьших квадратов:</u>
минимизация суммы квадратов отклонений
функции от искомых переменных

$$\theta = (X^T \cdot X)^{-1} \cdot X^T \cdot y$$

ограничения метода:

не работает для больших наборов данных

X^TX может быть необратимая (вырожденная)
- строки **X** линейно зависимы
- признаков больше чем примеров

линейная и нелинейная регрессия

$E(\theta, x, y) = h(x, \theta) - y$ - функция ошибки

$$J(\theta) = \frac{1}{2 \cdot m} \cdot \sum_{i=1}^{m} \left(h(x_i, \theta) - y_i \right)^2$$

функция потери (loss)
средняя квадратичная ошибка (MSQE)

$\min_{\theta} J(\theta)$ - задача оптимизации

линейная и нелинейная регрессия

$$E(\theta, x, y) = h(x, \theta) - y$$ - функция ошибки

$$J(\theta) = \frac{1}{2 \cdot m} \cdot \sum_{i=1}^{m} \left(h(x_i, \theta) - y_i\right)^2$$

функция потери (loss)
средняя квадратичная ошибка (MSQE)

$$\min_{\theta} J(\theta)$$ - задача оптимизации

метод градиентного спуска:

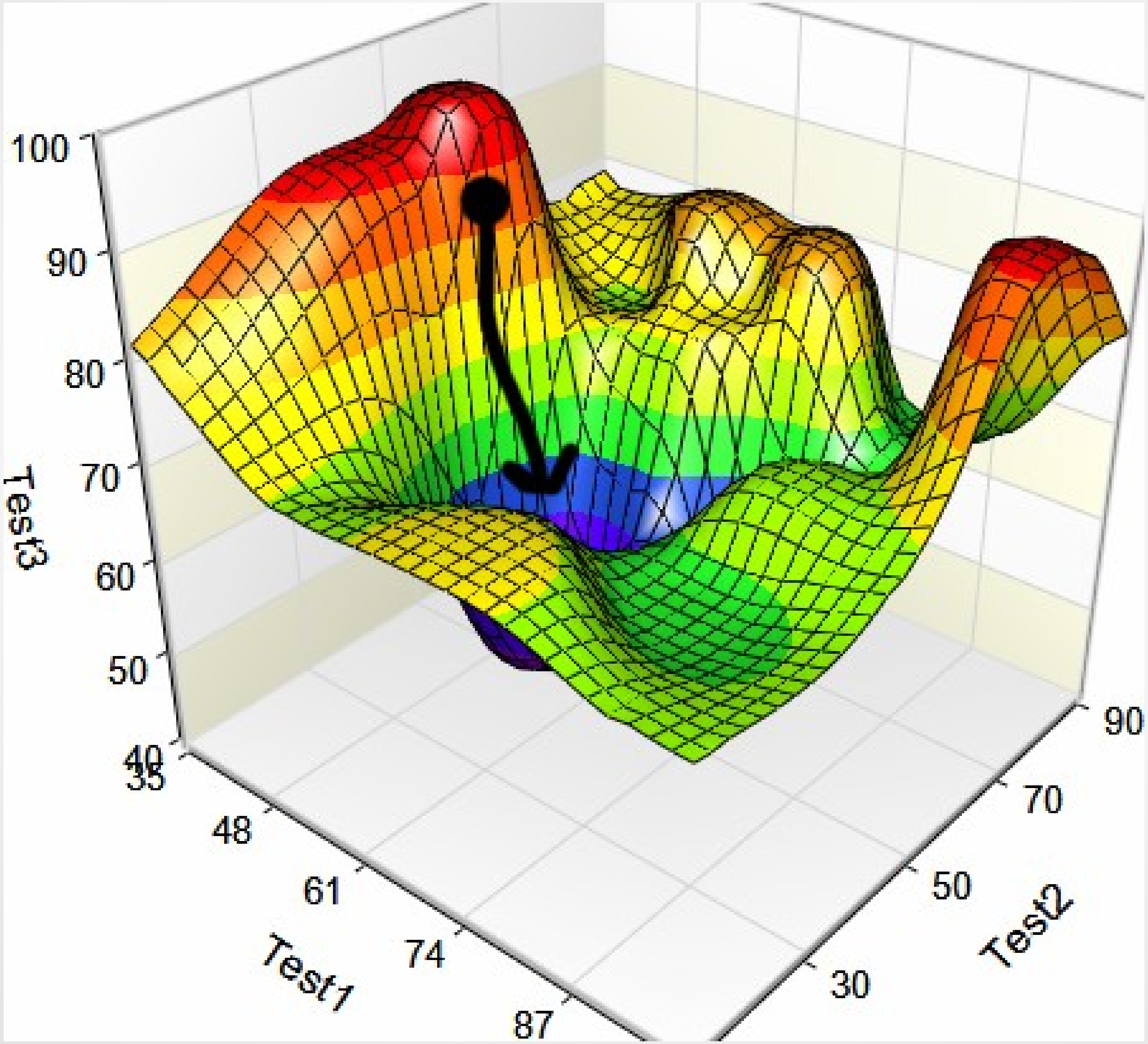
градиент функции - направление наискорейшего возрастания ф-ции

$$\nabla J(\theta) = \left[\frac{\partial J}{\partial \theta_0}, \ldots, \frac{\partial J}{\partial \theta_n}\right]$$

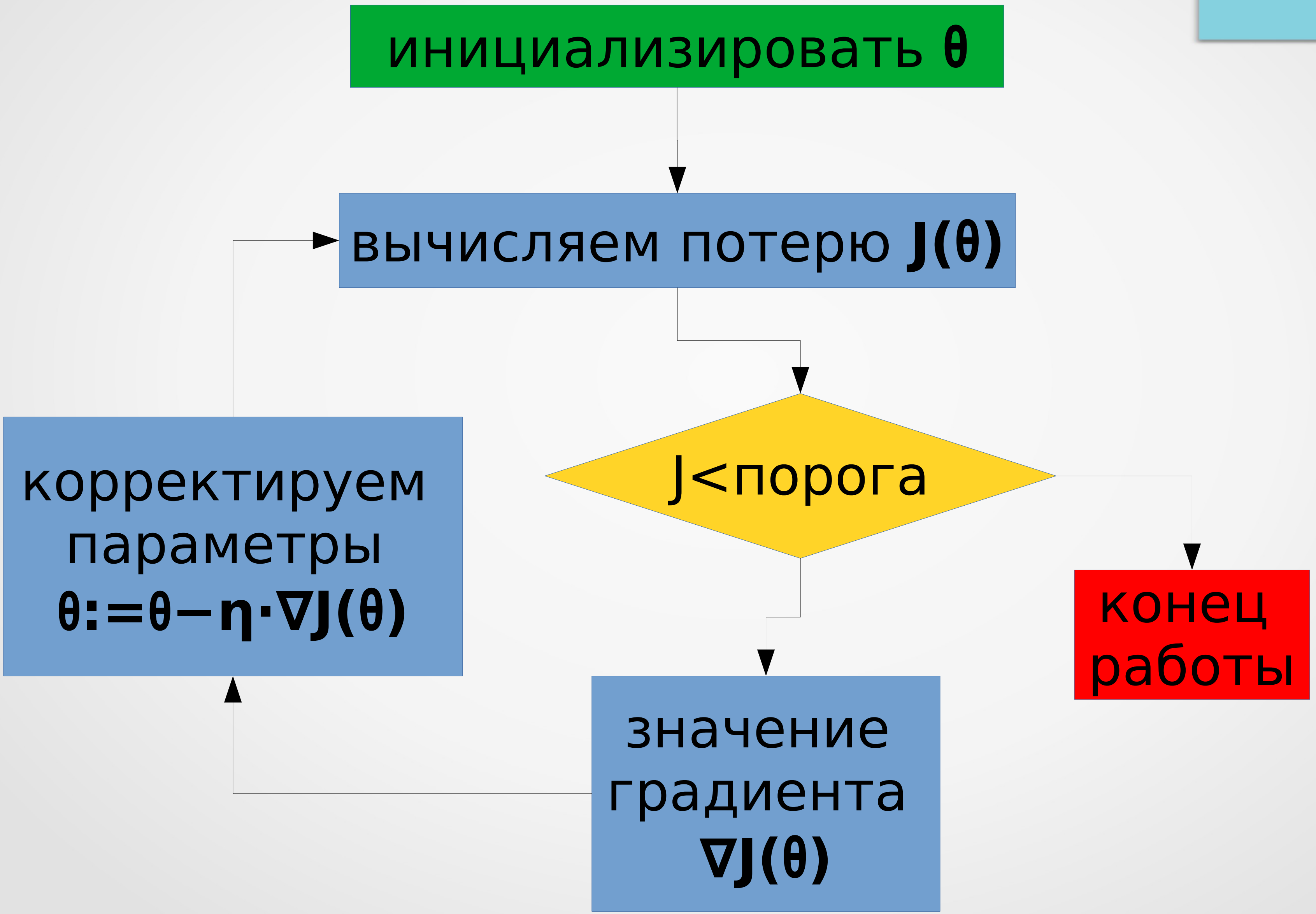
двигаем параметры в противоположную сторону

$$\theta := \theta - \alpha \cdot \nabla J(\theta)$$

линейная и нелинейная регрессия



линейная и нелинейная регрессия



линейная и нелинейная регрессия

функция потери (MSQE)

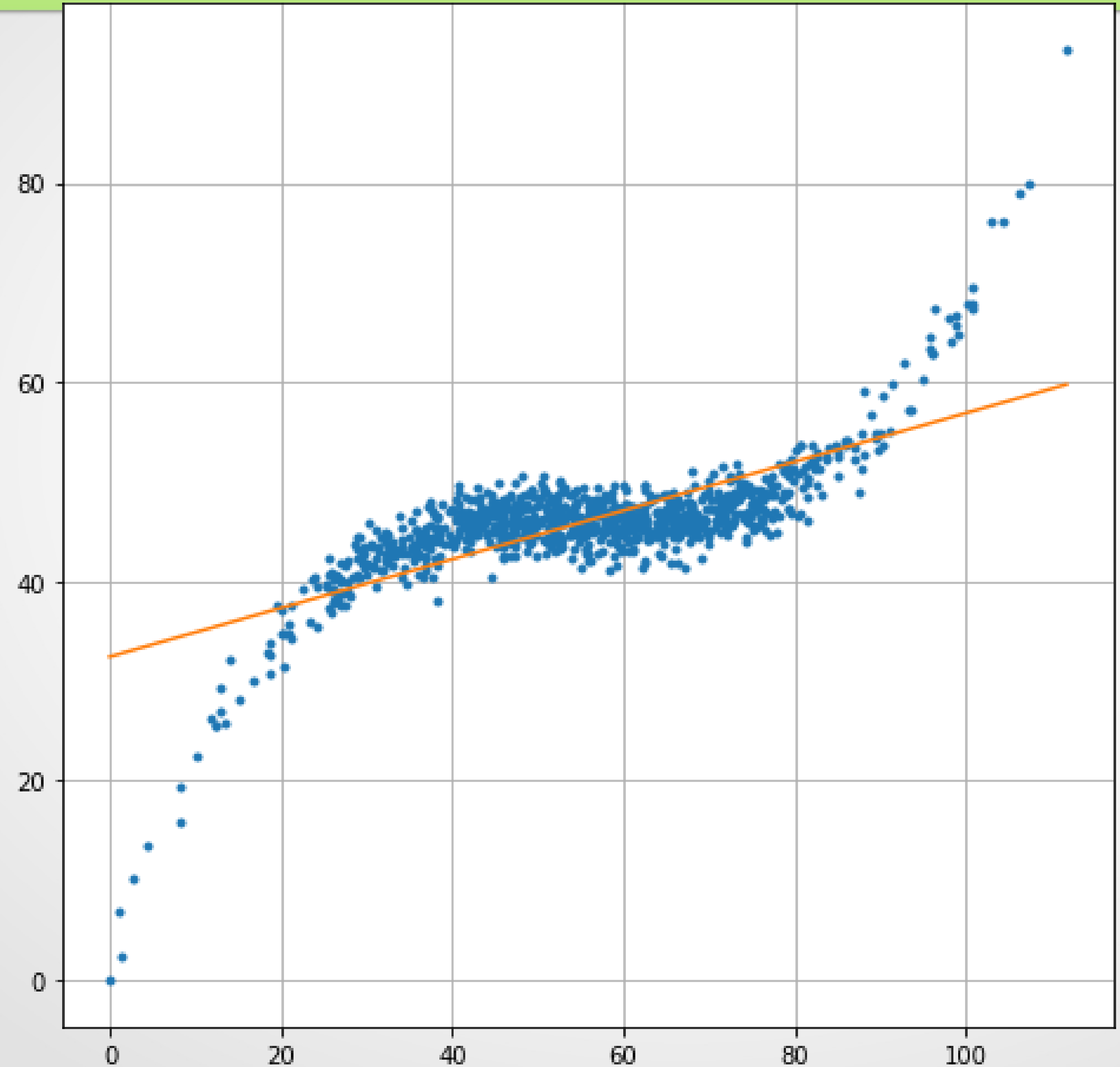
$$J(\theta)=\frac{1}{2\cdot m}\cdot\sum_{i=1}^{m}\left(h(x_i,\theta)-y_i\right)^2$$

градиент и изменение весов

$$\theta_j:=\theta_j-\alpha\cdot\frac{\partial J}{\partial\theta_j}=\theta_j-\alpha\cdot\frac{1}{m}\cdot\sum_{i=1}^{m}\left(h(x_i,\theta)-y_i\right)\cdot x_{ij}$$

$$\theta:=\theta-\alpha\cdot X^T\cdot\left(h(X,\theta)-y\right)$$

линейная и нелинейная регрессия



линейная и нелинейная регрессия

преобразование исходных данных

увеличиваем размерность пространства

строим полином степени **k** на признаках **x**, размера **n**

<u>пример для одномерного пространства (n=1) :</u>

строим полином k=1

линейная: $h_{lin}(\theta, x) = \theta_0 + \theta_1 x$

линейная и нелинейная регрессия

преобразование исходных данных

увеличиваем размерность пространства

строим полином степени **k** на признаках **x**, размера **n**

<u>пример для одномерного пространства (n=1) :</u>

строим полином k=1

линейная: $h_{lin}(\theta, x) = \theta_0 + \theta_1 x$

строим полином k=3

нелинейная: $h_{nlin}(\theta, x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$

линейная и нелинейная регрессия

m примеров, размера **n**, **n+1** параметр θ

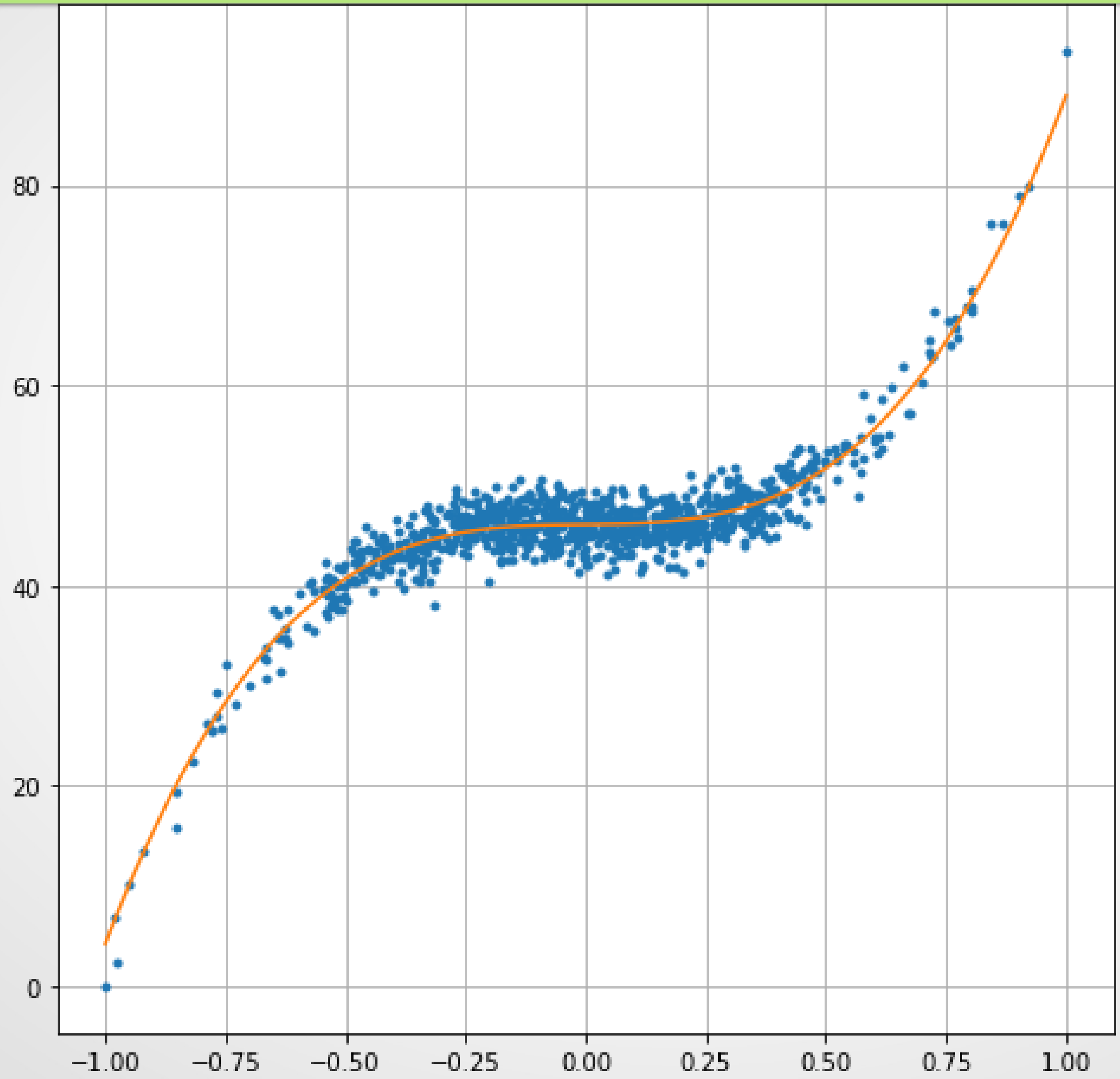
исходные данные $$X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

$$h(x,\theta)=\theta_0+\theta_1\cdot x_1+\theta_2\cdot x_2+\theta_3\cdot x_1\cdot x_2+\theta_4\cdot x_1^2+\theta_5\cdot x_2^2$$

строим полином степени **k** на переменных **x**,

комбинируем столбцы матрицы **X**
число параметров **θ** увеличивается

линейная и нелинейная регрессия



линейная и нелинейная регрессия

Вопросы ?

линейная и нелинейная регрессия

git clone *https://github.com/mechanoid5/ml_lectorium.git*

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

линейная и нелинейная 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