

Дисциплина  
**Основы машинного обучения и нейронные сети**

Лекция 3  
**Линейная регрессия**

# Линейная регрессия: источники

1. Основы машинного обучения (майор ИАД ВШЭ, *Е.А. Соколов*) [Электронный ресурс] / Режим доступа:  
[http://wiki.cs.hse.ru/%D0%9E%D1%81%D0%BD%D0%BE%D0%B2%D1%8B\\_%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE\\_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D1%8F](http://wiki.cs.hse.ru/%D0%9E%D1%81%D0%BD%D0%BE%D0%B2%D1%8B_%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D1%8F),  
свободный (дата обращения 15.10.2023).
2. Официальный канал школы "Deep Learning School" от Физтех-Школы прикладной математики и информатики МФТИ и Лаборатории нейронных систем и глубокого обучения МФТИ [Электронный ресурс] / Режим доступа:  
<https://www.youtube.com/@DeepLearningSchool/about>,  
свободный (дата обращения 15.10.2023).
3. Основы линейной регрессии [Электронный ресурс] / Режим доступа:  
<https://habr.com/ru/articles/514818/>, свободный (дата обращения 15.10.2023).
4. Википедия — свободная энциклопедия [Электронный ресурс] / Режим доступа:  
[https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%B0%D1%8F\\_%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F](https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%B0%D1%8F_%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F), свободный (дата обращения 15.10.2023).
5. *Мэрфи К.П.* Вероятностное машинное обучение. — М.: ДМК Пресс, 2022. — 940 с.: ил.  
*Murphy, Kevin P.* Machine learning: a probabilistic perspective. Cambridge, MIT Press, 2013.

# Обозначения (напоминание)

$\ell$  – размер выборки  $X$  (число прецедентов)

$d$  – длина признакового описания объекта (число признаков объекта)

$(\mathbf{x}_i, y_i)$  -  $i$ -й прецедент выборки  $X^\ell$ ,  $i = 1, 2, \dots, \ell$

$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \dots \\ x_{id} \end{pmatrix}$  -  $i$ -й объект выборки  $X^\ell$ ,  $i = 1, 2, \dots, \ell$

$x_{ij}$  -  $j$ -й признак  $i$ -го объекта выборки  $X^\ell$ ,  $i = 1, 2, \dots, \ell$ ,  $j = 1, 2, \dots, d$

Матрица «объекты–признаки» (feature data) размера  $\ell \times d$  для  $\ell = 9$ ,  $d = 4$ :

$$\mathbf{F} = ||f_j(\mathbf{x}_i)||_{9 \times 4} = \begin{pmatrix} f_1(\mathbf{x}_1) & \dots & f_4(\mathbf{x}_1) \\ \dots & \dots & \dots \\ f_1(\mathbf{x}_9) & \dots & f_4(\mathbf{x}_9) \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{\ell 1} & \dots & x_{\ell d} \end{pmatrix}$$

# Функция, линейная функция: определение

Функция —

отображение, соответствие между двумя множествами, при котором каждому элементу одного множества ( $X$ ) соответствует единственный элемент другого множества ( $Y$ ):  $f: X \rightarrow Y$  или  $f: x \mapsto y$

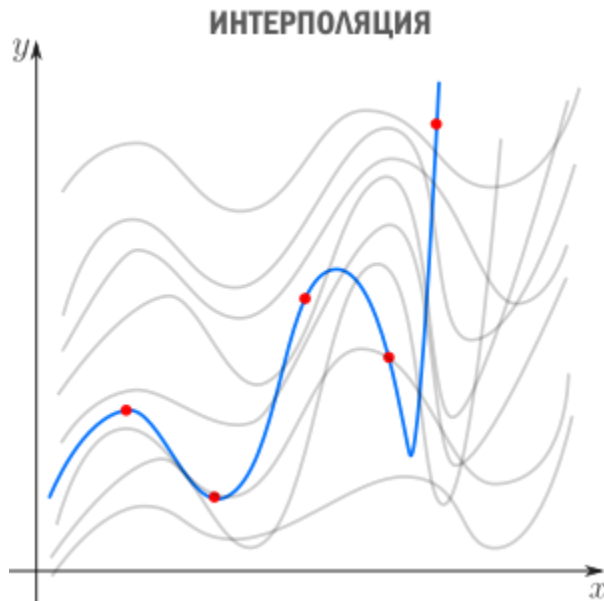
Линейное уравнение —

это алгебраическое уравнение, у которого максимальная степень составляющих его многочленов равна 1:

$$w_1x_1 + w_2x_2 + \dots + w_dx_d + w_0 = 0, \quad (3.2)$$

где  $w_j \in \mathbb{R}$ ,  $j = 1, 2, \dots, d$ .

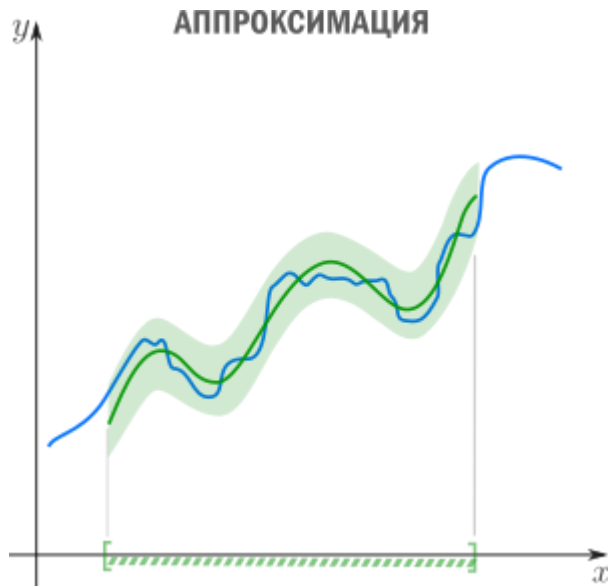
# Интерполяция



Интерполяция — способ выбрать из семейства функций ту, которая проходит через заданные точки. (Интерполяция — прогнозирование поведения функции внутри интервала, экстраполяция — вне интервала) [2].

Примеры: интерполяция полиномами Лагранжа, сплайн-интерполяция, многомерная интерполяция (билинейная, трилинейная, методом ближайшего соседа и т.д).

# Аппроксимация



Аппроксимация — способ выбрать из семейства «простых» функций приближение для «сложной» функции на отрезке, при этом ошибка не должна превышать определенного предела. Используют, когда нужно получить функцию, похожую на данную, но более удобную для вычислений и манипуляций (дифференцирования, интегрирования и т.п). [2]

Примеры: ряд Тейлора на отрезке, аппроксимация ортогональными многочленами, аппроксимация Паде, аппроксимация синуса Бхаскара и т.п.

# Регрессия

Регрессия — способ выбрать из семейства функций ту, которая минимизирует функцию потерь, характеризующую, насколько сильно пробная функция отклоняется от значений в заданных точках. Полученные экспериментально точки содержат ошибку измерений, поэтому разумнее требовать, чтобы функция передавала общую тенденцию, а не точно проходила через все точки. [2]



Регрессия — это «интерполирующая аппроксимация»: мы хотим провести кривую решающей функции как можно ближе к точкам и при этом сохранить ее максимально простой, чтобы уловить общую тенденцию. За баланс между этими противоречивыми желаниями как раз отвечает функция потерь (англ. «loss function» или «cost function»).

# Линейная регрессия: определение

Линейная регрессия (англ. Linear regression) — используемая в статистике и машинном обучении модель зависимости одной (объясняемой, зависимой) переменной  $y$  от другой или нескольких других переменных (признаков, факторов, регрессоров, независимых переменных)  $\mathbf{X}$  с линейной функцией зависимости [3]:

$$a(\mathbf{x}_i) = w_0 + \sum_{j=1}^d w_j x_{ij}, \quad i = 1, 2, \dots, \ell \quad (3.1)$$



# Машинное обучение (вспоминаем лк 1)

Говорят, что компьютерная программа *обучается* на основе опыта  $E$  по отношению к некоторому классу задач  $T$  и меры качества  $P$ , если качество решения задач из  $T$ , измеренное на основе  $P$ , улучшается с приобретением опыта  $E$ .

*Mitchell Tom. Machine Learning. McGraw Hill, 1997.*

# Задача обучения с учителем (обучение по прецедентам)

По обучающей выборке  $X^\ell$  построить *модель* (*алгоритм*)  $a: \mathbb{X} \rightarrow \mathbb{Y}$ , которая приближала бы целевую функцию  $y(\mathbf{x})$ , не только на объектах обучающей выборки, но и на всём множестве  $\mathbb{X}$ .

# Линейная регрессия: постановка задачи

Выборка:  $\mathbf{X} \in \mathbb{R}^{\ell \times d}$ ,  $\mathbf{y} \in \mathbb{R}$

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{\ell 1} & \dots & x_{\ell d} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

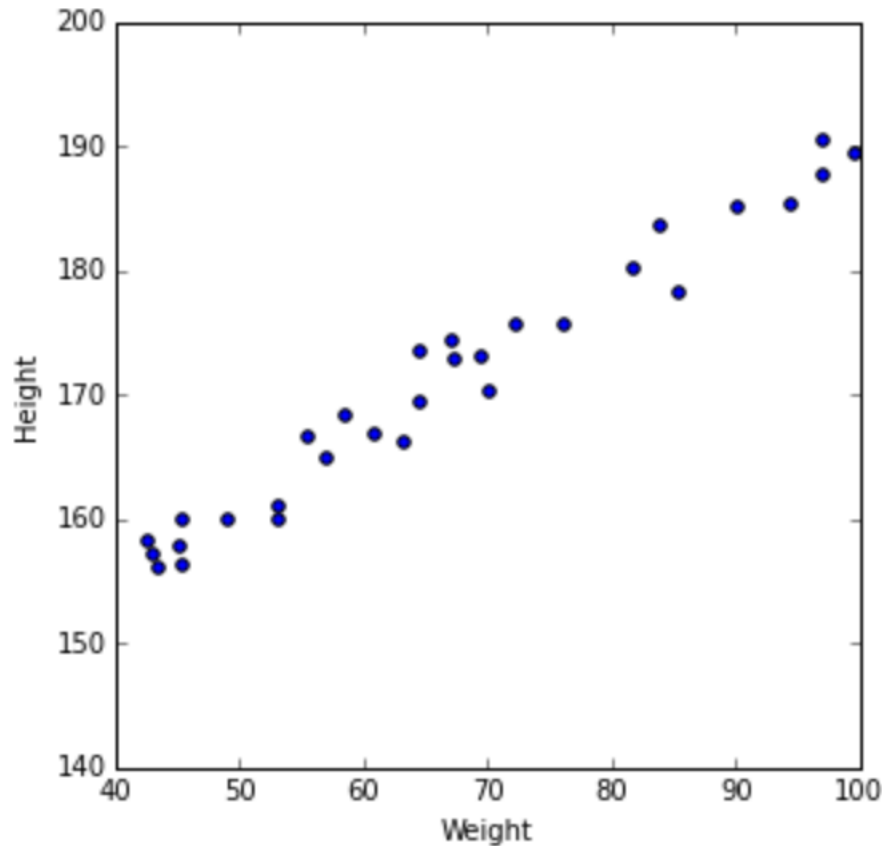
$$a(\mathbf{x}_i) = w_0 + \sum_{j=1}^d w_j x_{ij}, \quad i = 1, 2, \dots, \ell$$

Функция потерь - квадратичная:

$$\mathcal{L}(a(\mathbf{X}), \mathbf{y}) = \sum_{i=1}^{\ell} (a(\mathbf{x}_i) - y_i)^2$$

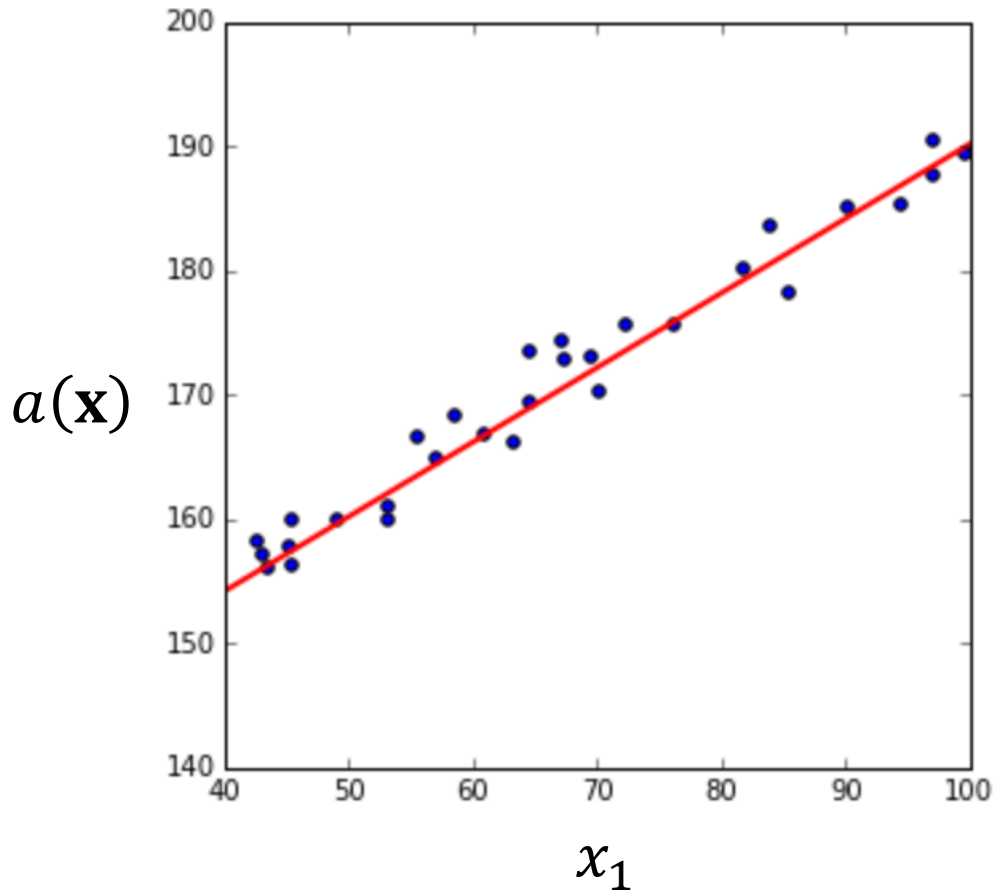
Линейная регрессия + квадратичная функция потерь =  
= метод наименьших квадратов

# Один признак ( $d=1$ ): парная регрессия (1/3)



$a(\mathbf{x})$  - рост человека  
 $x_1$  - вес человека

## Один признак ( $d=1$ ): парная регрессия (2/3)



$a(\mathbf{x})$  - рост человека  
 $x_1$  - вес человека

$$a(\mathbf{x}) = w_0 + w_1 x_1$$

(3.3)

# Один признак ( $d=1$ ): парная регрессия (3/3)

Простейший случай: один признак

Модель:

$$a(\mathbf{x}) = w_0 + w_1 x_1 \quad (3.3)$$

Два параметра:

$w_0$  – смещение по оси  $Oy$  (intercept)

$w_1$  – угловой коэффициент, тангенс угла наклона (slope)

## Два признака ( $d=2$ ) (1/2)

Чуть более сложный случай: два признака

Модель:

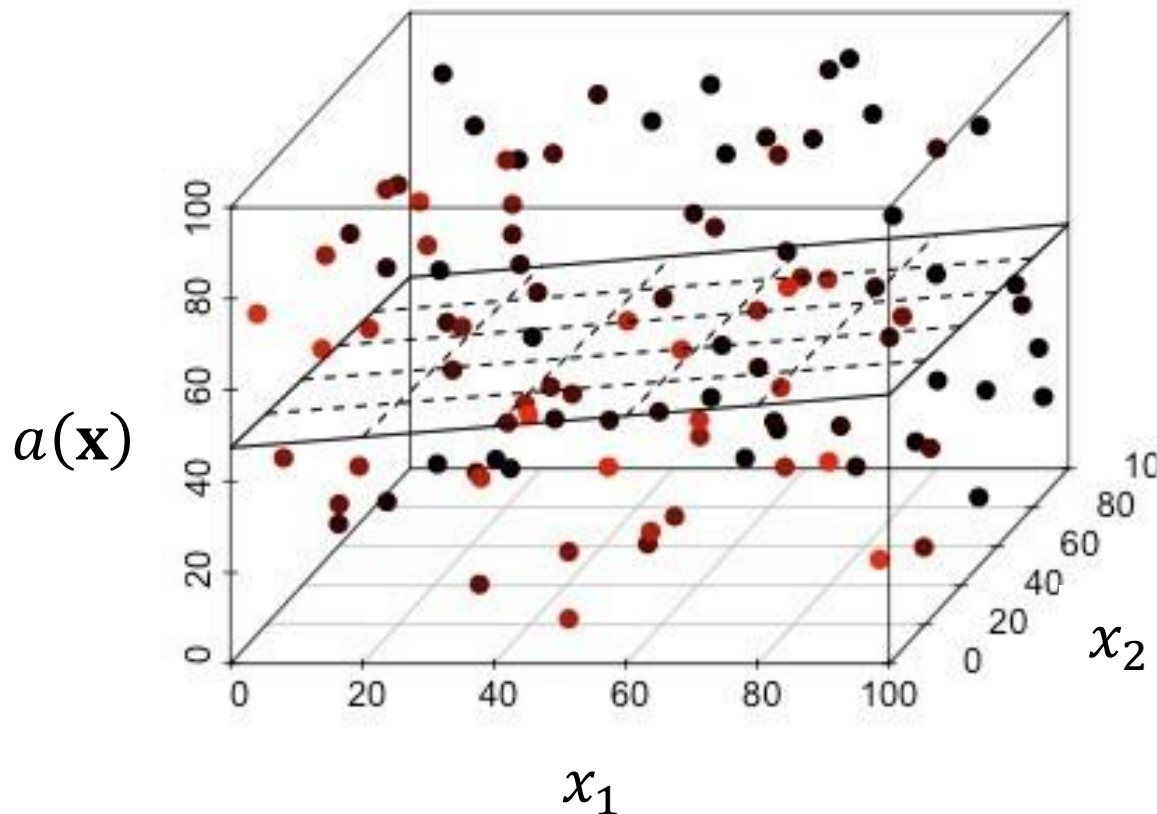
$$a(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 \quad (3.4)$$

Три параметра

## Два признака ( $d=2$ ) (2/2)

$$a(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2$$

(3.4)



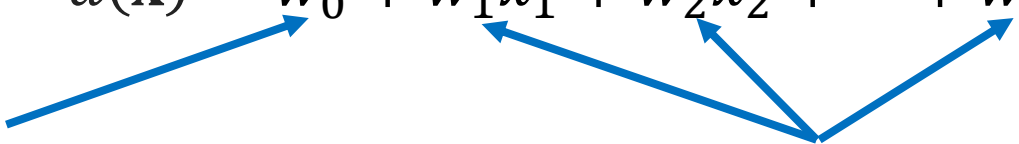


# Много признаков ( $d \geq 1$ ): мультилинейная регрессия (1/3)

Общий случай: много признаков ( $d \geq 1$ ) –  
мультилинейная или многофакторная регрессия

Модель:

$$a(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \dots + w_dx_d \quad (3.1)$$



свободный член (сдвиг, bias, intercept)

веса (параметры, коэффициенты)

Число параметров:  $d + 1$

# Применимость модели линейной регрессии

# Модель линейной регрессии: особенности

$$a(\mathbf{x}, \mathbf{w}) = \langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^T \mathbf{x}$$

- Нет гарантий, что целевая переменная именно так (линейно) зависит от признаков
- Надо формировать (конструировать) признаки так, чтобы каждый признак линейно влиял на целевую переменную.

# Прогнозирование стоимости квартиры (1/5)

- Признаки: площадь, район, расстояние до метро
- Целевая переменная: рыночная стоимость квартиры
- Линейная модель:

$$a(\mathbf{x}) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + \\ + w_3 * (\text{расстояние до метро})$$

## Прогнозирование стоимости квартиры (2/5)

$$a(\mathbf{x}) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + \\ + w_3 * (\text{расстояние до метро})$$

# Прогнозирование стоимости квартиры (3/5)

$$a(\mathbf{x}) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + \\ + w_3 * (\text{расстояние до метро})$$

$x_1$  – численный признак,  
за каждый кв. метр добавляем  $w_1$  рублей к прогнозу

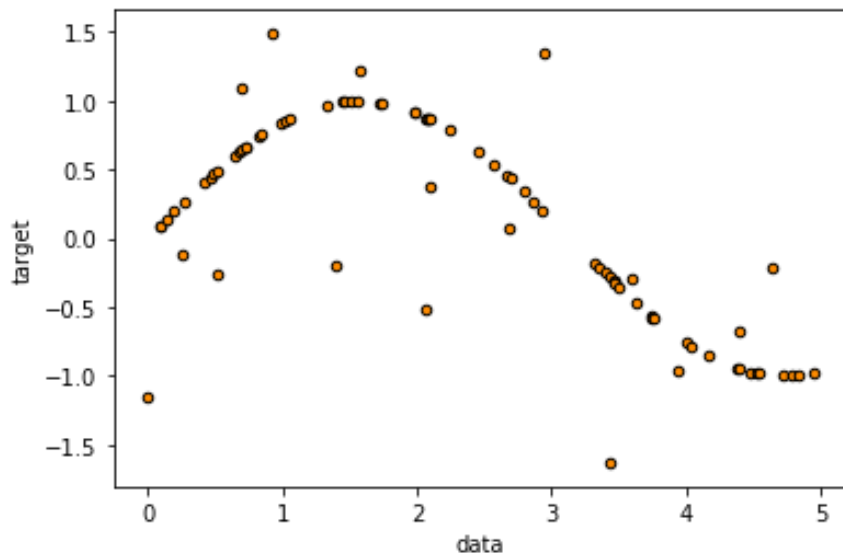
# Прогнозирование стоимости квартиры (4/5)

$$a(\mathbf{x}) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + \\ + w_3 * (\text{расстояние до метро})$$

$x_2$  – категориальный признак,  
нужно перевести в численный

# Прогнозирование стоимости квартиры (5/5)

$$a(\mathbf{x}) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + \\ + w_3 * (\text{расстояние до метро})$$



$x_3$  – численный признак,  
нелинейный



# Кодирование категориальных признаков (1/2)

Значение признака  $x_2$  «район» - из множества  
 $\mathbb{U} = \{u_1, u_2, \dots, u_m\} = \{\text{ЦАО}, \text{ЮАО}, \text{САО}\}$

Новые признаки  $x_{d+k}$  вместо признака  $x_2$  «район»:  
 $[x_{d+k} = u_k], k = 1, 2, \dots, m$

One-hot кодирование (one-hot encoding)

## Кодирование категориальных признаков (2/2)

Район		ЦАО	ЮАО	САО
ЦАО	→	1	0	0
ЮАО		0	1	0
ЦАО		1	0	0
САО		0	0	1
ЮАО		0	1	0

$$\begin{aligned} a(\mathbf{x}) = & w_0 + w_1 \cdot (\text{площадь}) + \\ & + w_2 \cdot (\text{район}) + \\ & + w_3 \cdot (\text{расстояние до метро}) \end{aligned}$$

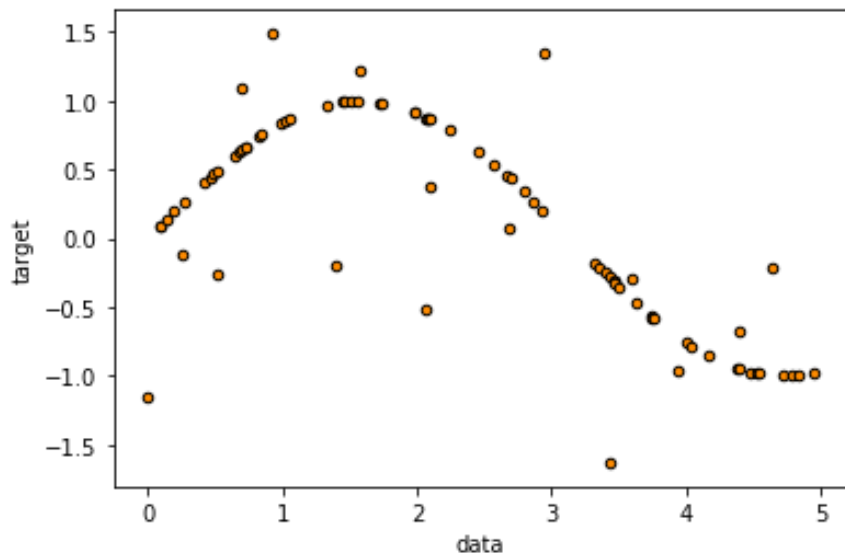
## Кодирование категориальных признаков (2/2)

Район		ЦАО	ЮАО	САО
ЦАО		1	0	0
ЮАО		0	1	0
ЦАО		1	0	0
САО		0	0	1
ЮАО		0	1	0

$$\begin{aligned} a(\mathbf{x}) = & w_0 + w_1 \cdot (\text{площадь}) + \\ & + w_4 \cdot (\text{квартира в ЦАО?}) + \\ & + w_5 \cdot (\text{квартира в ЮАО?}) + \\ & + w_6 \cdot (\text{квартира в САО?}) + \\ & + w_3 \cdot (\text{расстояние до метро}) \end{aligned}$$

# Прогнозирование стоимости квартиры (5/5)

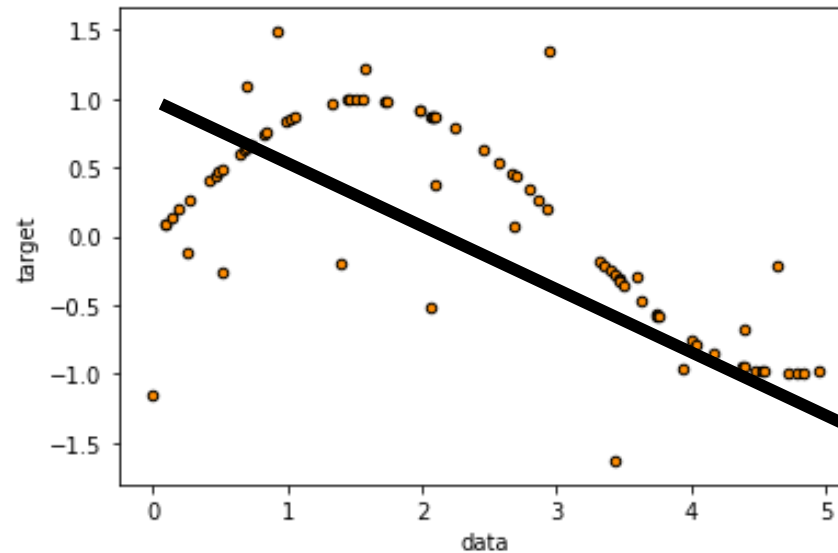
$$a(\mathbf{x}) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + \\ + w_3 * (\text{расстояние до метро})$$



$x_3$  – численный признак,  
нелинейный

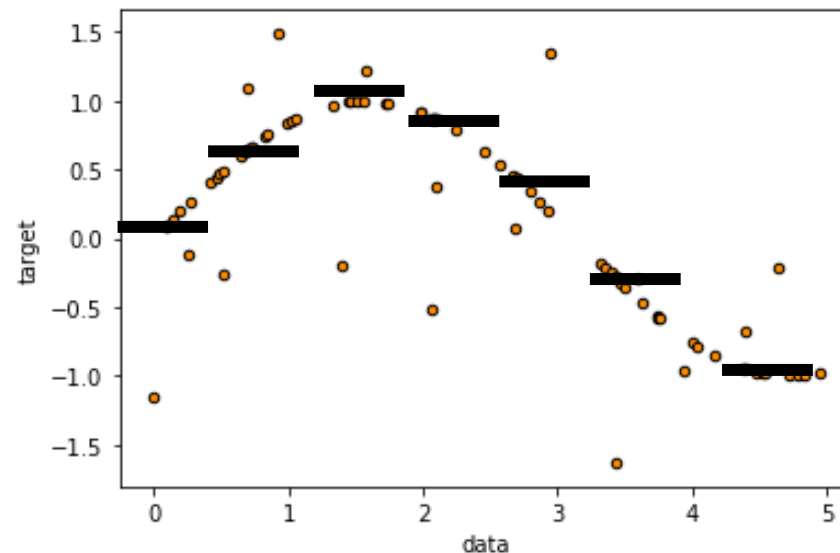
# Работа с сложной зависимостью (1/3)

$$a(\mathbf{x}) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + \\ + w_3 * (\text{расстояние до метро})$$



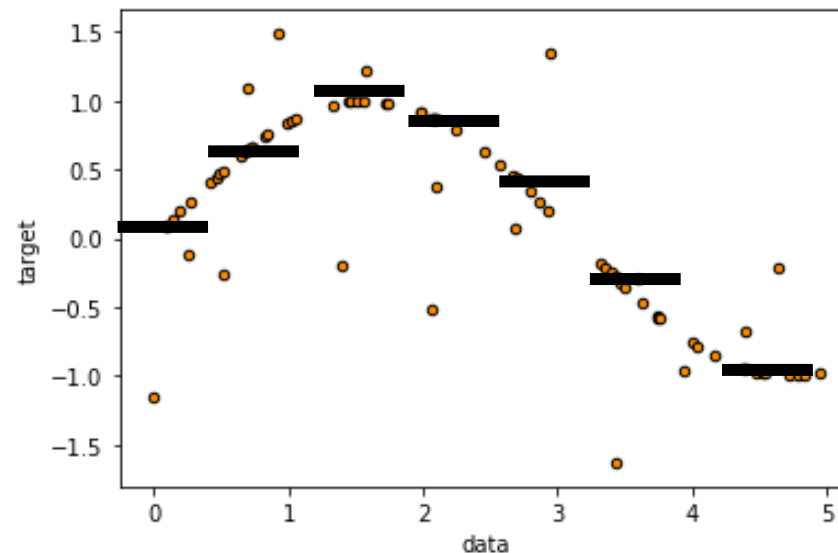
## Работа с сложной зависимостью (2/3)

$$a(\mathbf{x}) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + \\ + w_3 * (\text{расстояние до метро})$$



# Работа с сложной зависимостью (3/3)

$$a(\mathbf{x}) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + \\ + w_3 * [t_0 \leq x_3 < t_1] + \dots + w_{3+n} [t_{n-1} \leq x_3 < t_n]$$



# Линейность по параметрам модели, а не по признакам объекта (1/3)

Переменные в полиномах в линейном уравнении – функции от признаков:

$$\begin{aligned} a(\mathbf{x}, \mathbf{w}) = & w_0 + w_1 * (\text{площадь}) + w_2 * (\text{этаж}) + \\ & + w_3 * (\text{расстояние до метро}) + w_4 * (\text{площадь})^2 + \\ & + w_5 * (\text{этаж})^2 + w_6 * (\text{расстояние до метро})^2 + \\ & + w_7 * (\text{площадь}) * (\text{этаж}) + \dots \end{aligned}$$

Модель  $a(\mathbf{x}, \mathbf{w})$  линейно зависит от параметров  $\mathbf{w}$ , хотя и нелинейно зависит от входных данных  $\mathbf{x}$ .



# Линейность по параметрам модели, а не по признакам объекта (2/3)

$$a(\mathbf{x}, \mathbf{w}) = w_0 + w_1x_1 + w_2x_2 + \dots + w_dx_d = \mathbf{w}^T \mathbf{x}$$

– модель **линейной регрессии**, где  
переменными в многочлене выступают  
признаки объекта

$$a(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \varphi(\mathbf{x})$$

– модель **полиномиальной регрессии**, где  
переменными в многочлене выступают  
предобработанные признаки (конструирование  
признаков)

Например,  $\varphi(x) = (1, x, x^2, \dots, x^D)$

# Линейность по параметрам модели, а не по признакам объекта (3/3)

$$a(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \varphi(\mathbf{x})$$

– модель **полиномиальной регрессии**, где переменными в полиноме выступают предобработанные признаки

**Переменной в полиноме в линейном уравнении** может выступать

- численный признак объекта:  $x_j$
- степень численного признака:  $x_{d+3} = (x_j)^4$
- преобразование численного признака (логарифм, корень и т.д.):  $x_{d+2} = \sqrt{x_j}$
- произведение численных признаков:  $x_{d+4} = x_j \cdot x_k$
- значения из one-hot-encoding (0 и 1)

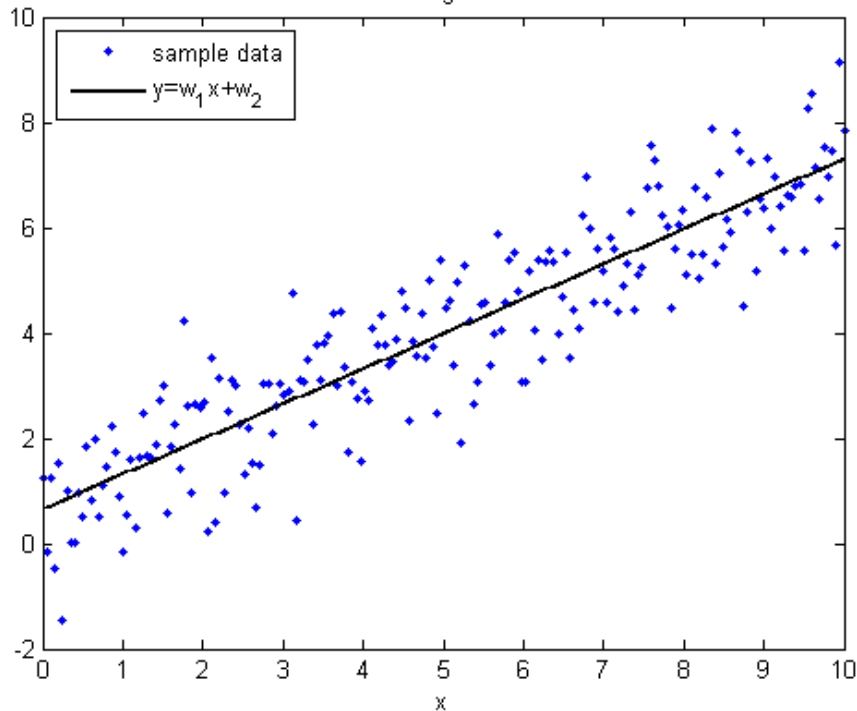
Цель – сделать модель более выразительной, позволять прогнозировать результат не только гиперплоскостью, но и гиперповерхностью.

# Примеры переменных в полиноме в линейном уравнении (1/2)

Переменные - признаки объекта:

$$a(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x}$$

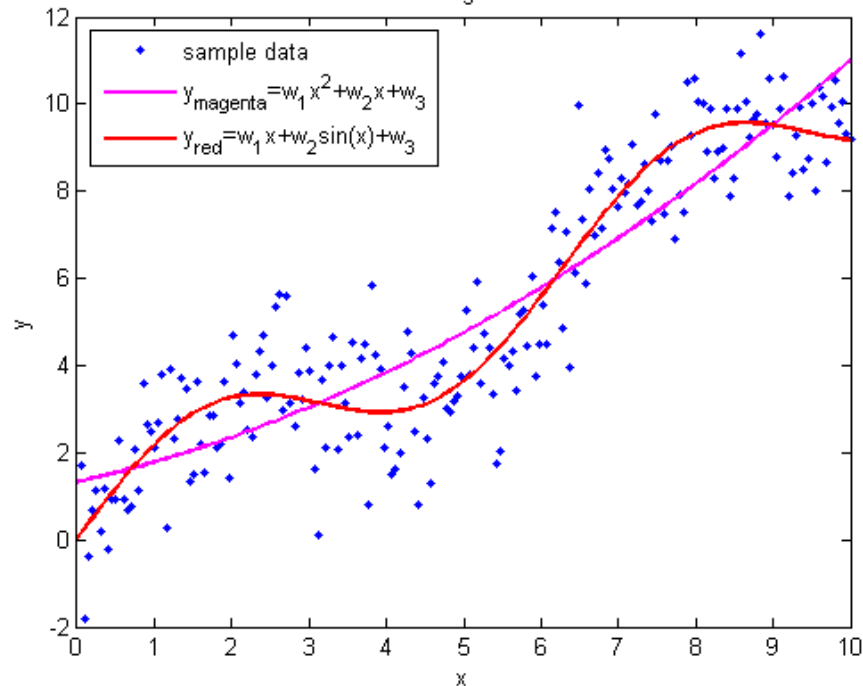
Linear regression



Переменные – функции от признаков:

$$a(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \varphi(\mathbf{x})$$

Linear regression



Соколов Е.А. Материалы курса «Основы машинного обучения», ВШЭ, майнор ИАД (доп. профиль «Интеллектуальный анализ данных»).

07.11.2023

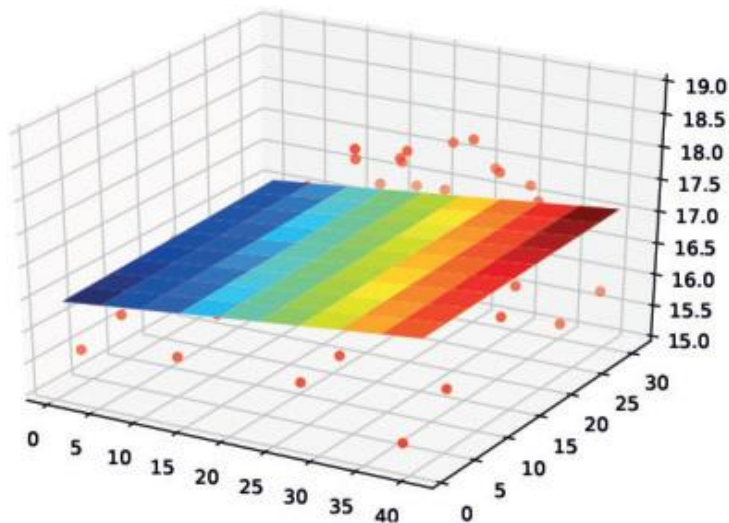
Линейная регрессия

37

# Примеры переменных в полиноме в линейном уравнении (2/2)

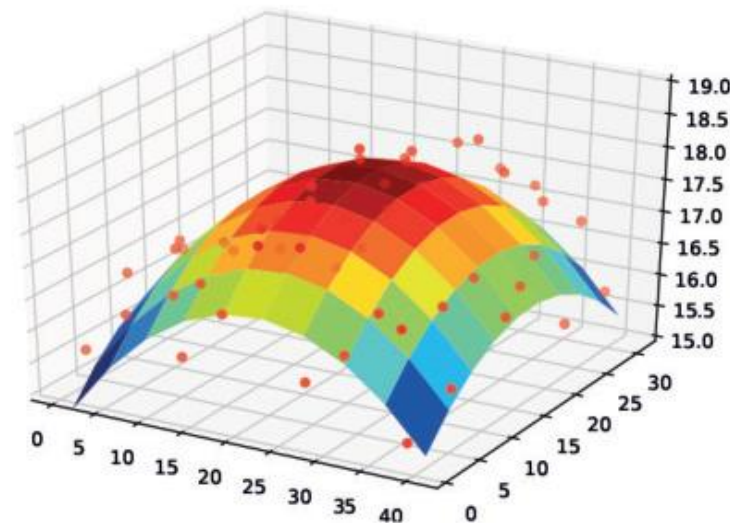
Переменные - признаки объекта:

$$a(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} = \\ = w_0 + w_1 x_1 + w_2 x_2$$



Переменные – функции от признаков:

$$a(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = \\ = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2$$



Прогнозирование температуры воздуха в разных точках лаборатории Intel в Беркли, штат Калифорния (используются с разрешения Ромейна Тибо).

Построено программой по адресу [figures.probml.ai/book1/1.6](https://figures.probml.ai/book1/1.6)

*Мэрфи К.П. Вероятностное машинное обучение. – М.: ДМК Пресс, 2022. – 940 с.: ил.*

# Применимость модели линейной регрессии

Модель линейной регрессии хороша, если данные специально подготовлены:

- признаки предобработаны one-hot кодированием категориальных признаков или бинаризацией численных признаков
- признаки сконструированы для подгонки (поиска весов) модели в виде линейной зависимости

# Умножение матриц: вспоминаем линейную алгебру

$$\mathbf{A} \in \mathbb{R}^{m \times k}, \quad \mathbf{B} \in \mathbb{R}^{k \times n}, \quad \mathbf{C} \in \mathbb{R}^{m \times n}$$

$$\mathbf{A} = (a_{ij})_{\substack{i=1,\dots,m \\ j=1,\dots,k}}; \quad \mathbf{B} = (b_{ij})_{\substack{i=1,\dots,k \\ j=1,\dots,n}}; \quad \mathbf{C} = (c_{ij})_{\substack{i=1,\dots,m \\ j=1,\dots,n}}$$

$$\mathbf{C} = \mathbf{A} \times \mathbf{B}$$

$$c_{ij} = \sum_{s=1}^k a_{is} b_{sj}, \quad i = 1, \dots, m, \quad j = 1, \dots, n$$

# Умножение матриц:

пример (1/5)

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}$$

# Умножение матриц:

пример (2/5)

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & & \\ & & \\ & & \end{pmatrix}$$



# Умножение матриц:

пример (3/5)

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \\ & & \\ & & \end{pmatrix}$$

# Умножение матриц:

пример (4/5)

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 4 \\ & & \\ & & \end{pmatrix}$$

# Умножение матриц:

пример (5/5)

$$\begin{pmatrix} 1 & 2 \\ \boxed{0} & \boxed{1} \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} \boxed{1} & 0 & 0 \\ \boxed{0} & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 4 \\ \boxed{0} & & \end{pmatrix}$$

# Умножение матриц: линейная модель

Матрица «объекты-признаки»:

*объект и его признаки*

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

# Умножение матриц: линейная модель

Матрица «объекты-признаки»:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

*значения признака на всех объектах*

# Линейная модель в матричном виде

$$a(\mathbf{x}_i, \mathbf{w}) = \langle \mathbf{w}, \mathbf{x}_i \rangle = \mathbf{w}^T \mathbf{x}_i, \quad i = 1, 2, \dots, \ell$$

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

$$\mathbf{X}\mathbf{w} = \mathbf{y}$$
$$\begin{pmatrix} \sum_{i=1}^d w_i x_{1i} \\ \sum_{i=1}^d w_i x_{2i} \\ \vdots \\ \sum_{i=1}^d w_i x_{\ell i} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ \cdots \\ \cdots \\ y_\ell \end{pmatrix}$$

# Линейная регрессия: обучение на основе функции потерь

# Задача обучения с учителем (обучение по прецедентам)

По обучающей выборке  $X^\ell$  построить *модель* (*алгоритм*)  $a: \mathbb{X} \rightarrow \mathbb{Y}$ , которая приближала бы целевую функцию  $y(\mathbf{x})$ , не только на объектах обучающей выборки, но и на всём множестве  $\mathbb{X}$ .



# Линейная регрессия: постановка задачи

Выборка:  $\mathbf{X} \in \mathbb{R}^{\ell \times d}$ ,  $\mathbf{y} \in \mathbb{R}^{\ell \times 1}$

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{\ell 1} & \dots & x_{\ell d} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

Модель:  $a(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w}$

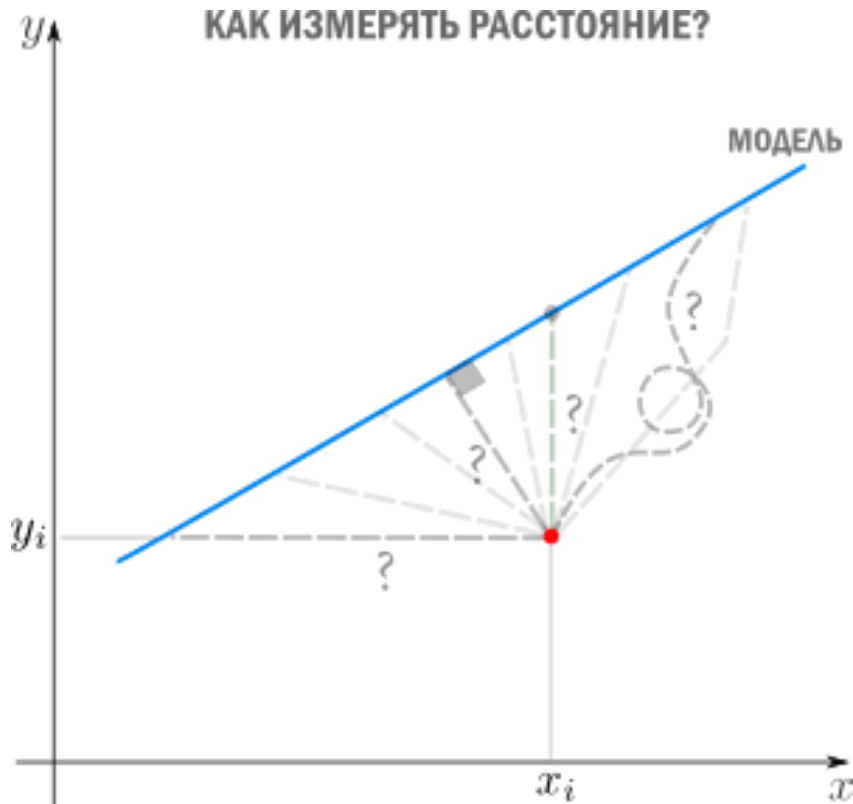
$$a(\mathbf{x}_i) = w_0 + \sum_{j=1}^d w_j x_{ij}, \quad i = 1, 2, \dots, \ell$$

Функция потерь - квадратичная:

$$\mathcal{L}(a(\mathbf{X}), \mathbf{y}) = \sum_{i=1}^{\ell} (a(\mathbf{x}_i) - y_i)^2$$

Линейная регрессия + квадратичная функция потерь =  
= метод наименьших квадратов

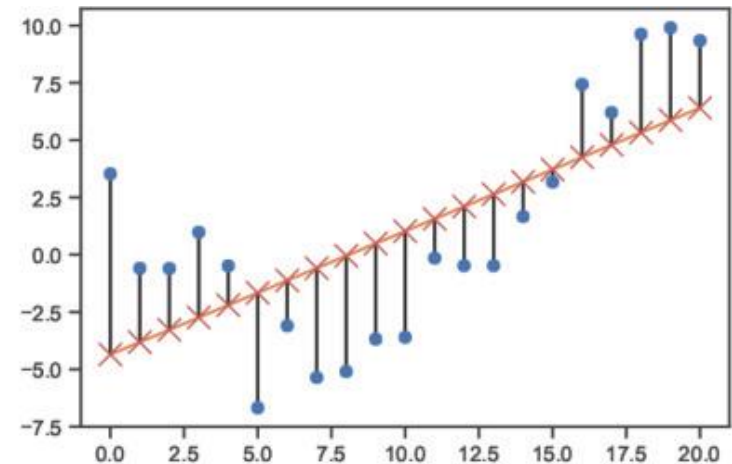
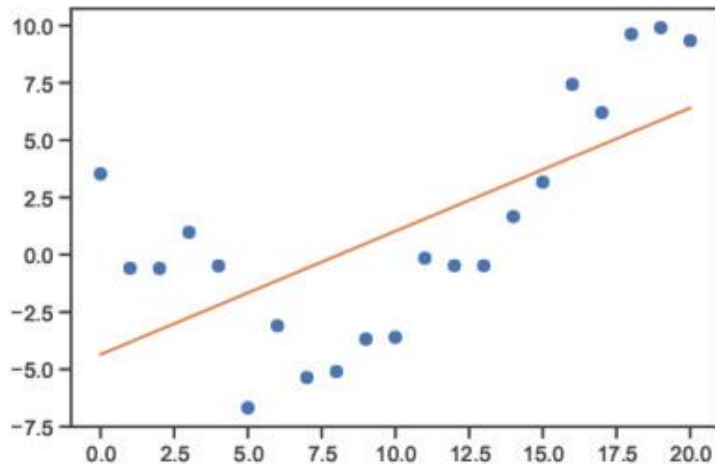
# Метод наименьших квадратов (1/2)



Расстояние от точки до прямой — «геометрическое», длина перпендикуляра.

Расстояние в линейной регрессии — «функциональное», невязка  $(a(x_i) - y_i)$ .

# Метод наименьших квадратов (2/2)



красная линия – модель линейной регрессии  
синие точки – ответы  $y_i$  из выборки  
красные крестики – прогнозы  $a(x_i)$  модели  
вертикальные отрезки - невязки между  $a(x_i)$  и  $y_i$

Цель регрессии: методом наименьших квадратов выбрать линию (т.е.  $w$ ), которая минимизирует сумму квадратов невязок  $\sum_{i=1}^{\ell} (a(x_i) - y_i)^2$ .

Построено программой по адресу [figures.problml.ai/book1/1.5](https://figures.problml.ai/book1/1.5)

*Мэрфи К.П. Вероятностное машинное обучение. – М.: ДМК Пресс, 2022. – 940 с.: ил.*

# Функционал качества

Квадратичная ошибка SSE для линейной регрессии:

$$Q_{SSE}(w_1, \dots, w_d) = \sum_{i=1}^{\ell} (\mathbf{w}_1 x_1 + \dots + \mathbf{w}_d x_d - y_i)^2$$

Среднеквадратичная ошибка MSE для линейной регрессии:

$$Q_{MSE}(w_1, \dots, w_d) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x} \rangle - y_i)^2$$

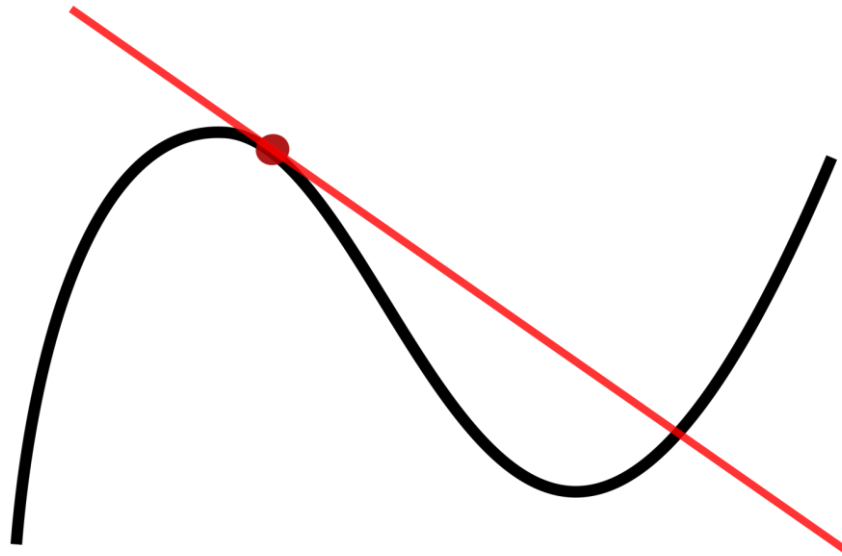
Задача:

$$Q(w_1, \dots, w_d) \rightarrow \min_{\mathbf{w}}$$

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{R}^d} Q(\mathbf{w})$$

# Производная (1/3)

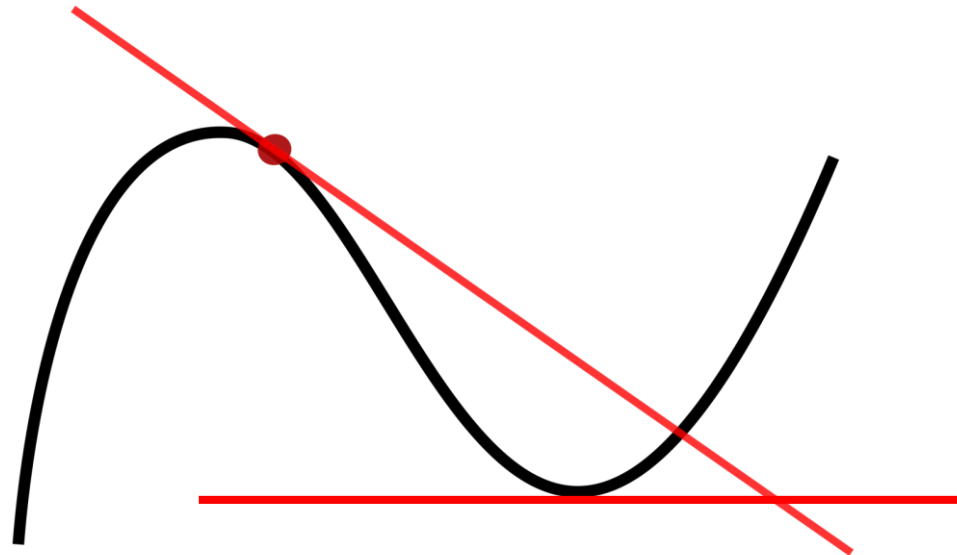
$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$



## Производная (2/3)

Если точка  $x_0$  — экстремум и в ней существует производная, то

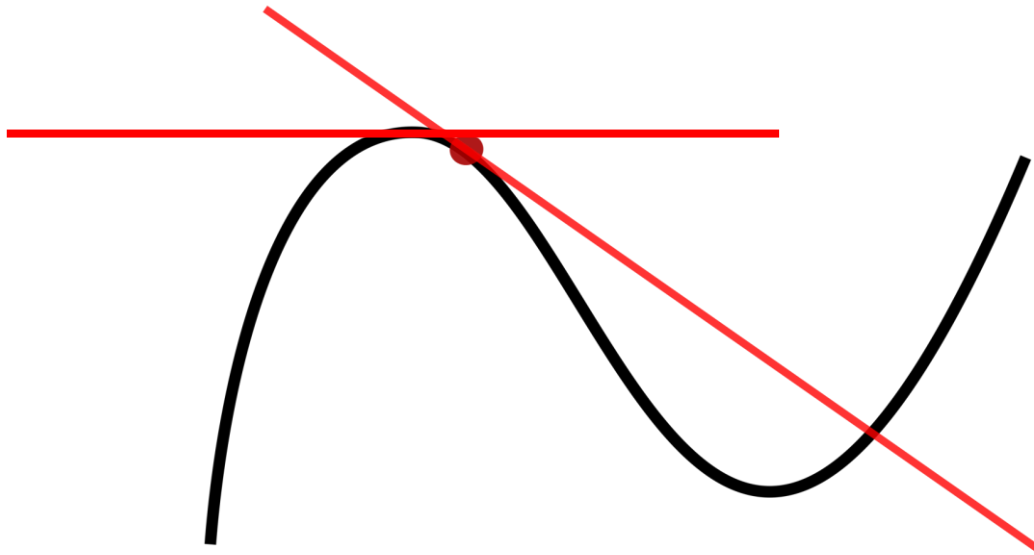
$$f'(x_0) = 0$$



## Производная (3/3)

Если точка  $x_0$  — экстремум и в ней существует производная, то

$$f'(x_0) = 0$$

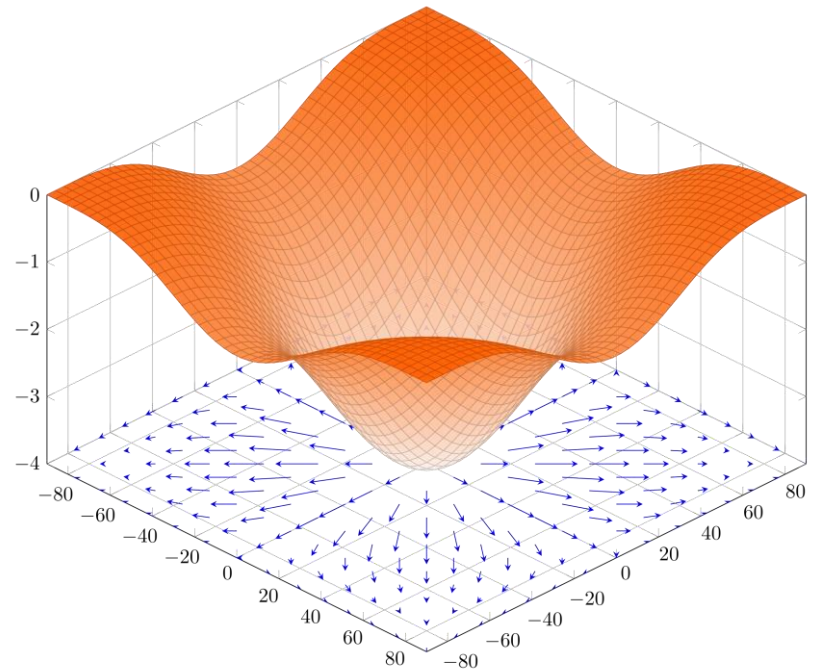


# Градиент

Градиент — вектор, своим направлением указывающий направление наискорейшего роста некоторой скалярной величины.

Компоненты - частные производные

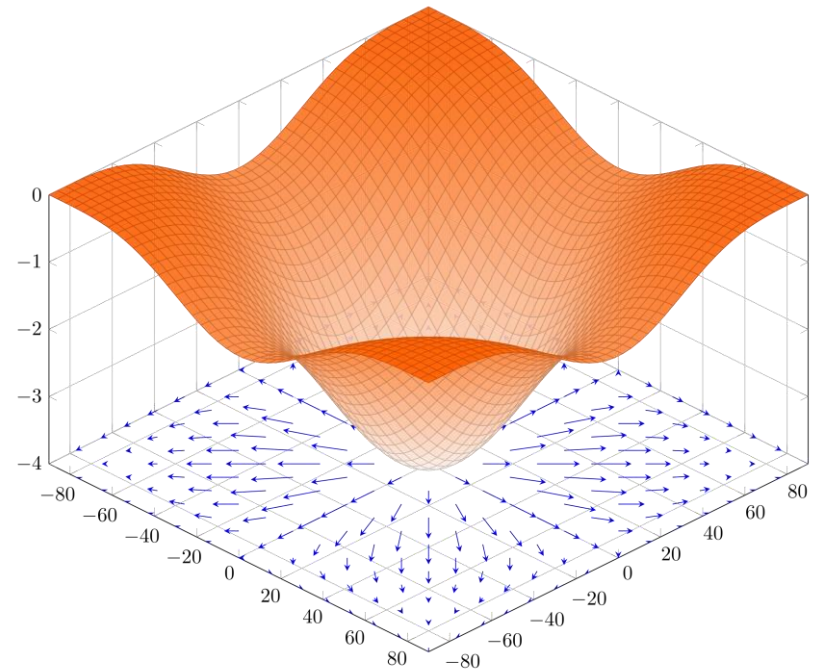
$$\nabla f(x) = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d} \right)$$





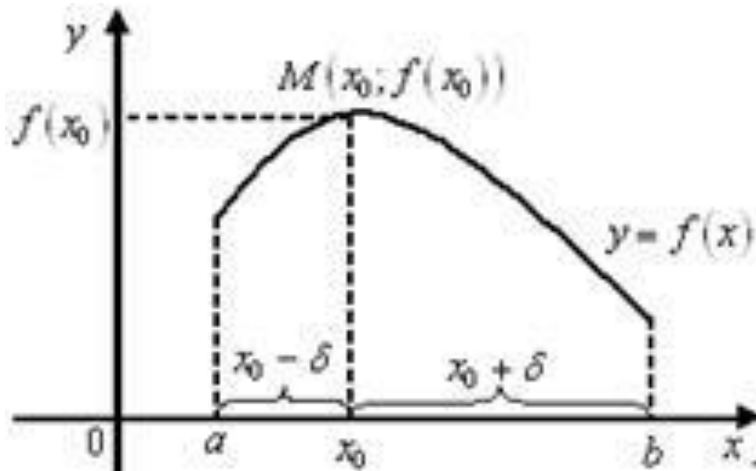
# Важное свойство градиента

- В точке  $x_0$  функция быстрее всего растёт в направлении градиента
- Если градиент равен нулю, то это экстремум



# Определение экстремума

## вспоминаем математический анализ



Теорема 3.1. Для того, чтобы дифференцируемая на  $(a, b)$  функция  $f(x)$  не убывала (не возрастала) на этом интервале, необходимо и достаточно, чтобы  $\nabla f(x_0) \geq 0$  ( $\nabla f(x_0) \leq 0$ )  $\forall x \in (a, b)$ .

Если  $\nabla f(x_0) > 0$ , то  $f(x)$  возрастает.  
Если  $\nabla f(x_0) < 0$ , то  $f(x)$  убывает.

Точка  $x_0$  - точка локального экстремума:

$f(x) - f(x_0) < 0$  – локальный максимум,

$f(x) - f(x_0) > 0$  – локальный минимум.

Наименьшее и наибольшее значения функции  $f(x)$  на  $[a, b]$  - *абсолютные* минимум и максимум или *глобальные* экстремумы функции  $f(x)$ .

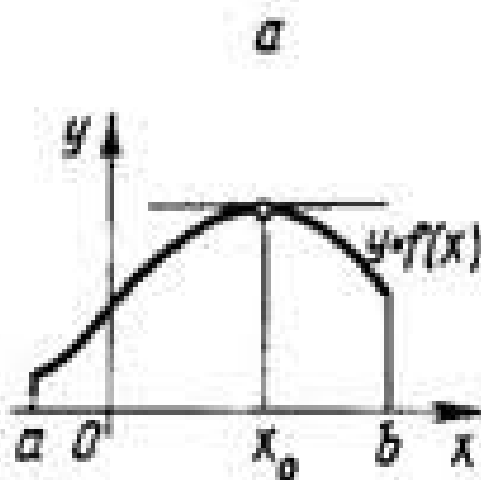
# Необходимое условие существования локального экстремума (1/2) вспоминаем математический анализ

Теорема 3.2. Если в точке  $x_0$  функция  $f(x)$  имеет экстремум, то её производная в ней либо равна нулю  $\nabla f(x_0) = 0$ , либо не существует.

Точка  $x_0$  - *критическая* точка, точка возможного экстремума.

# Необходимое условие существования локального экстремума (2/2)

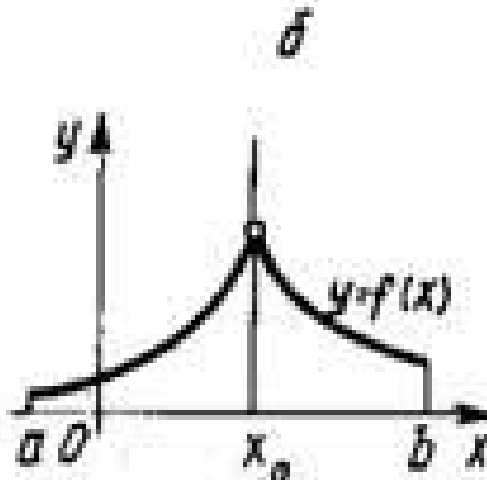
вспоминаем математический анализ



касательная параллельна  
оси абсцисс,

$$\nabla f(x_0) = 0,$$

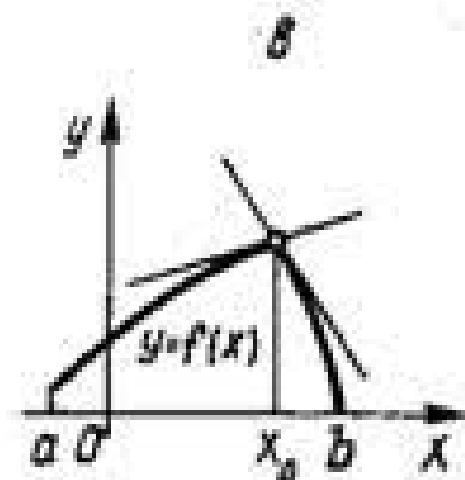
$x_0$  - стационарная точка



касательная  
параллельна  
оси ординат,

$$\nabla f(x_0) = \infty,$$

$x_0$  - точка возврата



существуют не совпадающие  
левая и правая касательные,

$$\nabla f(x_0 - 0) \neq \nabla f(x_0 + 0),$$

$x_0$  - угловая точка

# Достаточные условия (признаки) экстремума (1/3)

вспоминаем математический анализ

Теорема 3.3 (первый достаточный признак существования экстремума функции).

Если производная  $\nabla f(x)$  непрерывной функции  $f(x)$  при переходе через критическую точку  $x_0$  меняет знак с «+» на «—», то  $x_0$  - точка локального максимума, с «-» на «+», то  $x_0$  - точка локального минимума, не меняет знак, то  $x_0$  - не точка локального экстремума.

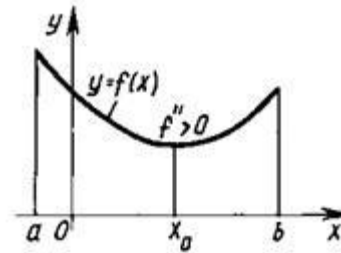
# Достаточные условия (признаки) экстремума (2/3)

вспоминаем математический анализ

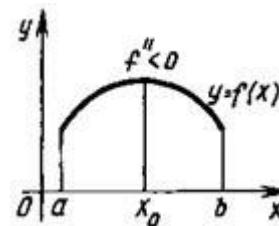
Теорема 3.4 (второй достаточный признак существования экстремума функции).

Стационарная точка  $x_0$  функции  $f(x)$ , дважды дифференцируемой в её  $\delta$ -окрестности  $[x_0 - \delta, x_0 + \delta]$ , является

точкой локального минимума,  
если  $f''(x_0) > 0$ ,



точкой локального максимума,  
если  $f''(x_0) < 0$ .



# Достаточные условия (признаки) экстремума (3/3)

вспоминаем математический анализ

Теорема 3.5 (третий достаточный признак существования экстремума функции).

Пусть функция  $f(x)$   $n$  раз непрерывно дифференцируема в точке  $x_0$  и в этой точке

$$f'(x_0) = f''(x_0) = \dots = f^{(n-1)}(x_0) = 0, \text{ и } f^{(n)}(x_0) \neq 0.$$

Если  $n$  чётное и  $f^{(n)}(x_0) < 0$ , то  $x_0$  - точка локального максимума,  
если  $n$  чётное и  $f^{(n)}(x_0) > 0$ , то  $x_0$  - точка локального минимума,  
если  $n$  нечётное, то  $x_0$  - не точка локального экстремума.

# Линейная регрессия: точное решение (1/2)

$$a(\mathbf{x}_i, \mathbf{w}) = \langle \mathbf{w}, \mathbf{x}_i \rangle = \mathbf{w}^T \mathbf{x}_i, \quad i = 1, 2, \dots, \ell$$

$$\mathbf{X}\mathbf{w} = \mathbf{y}$$

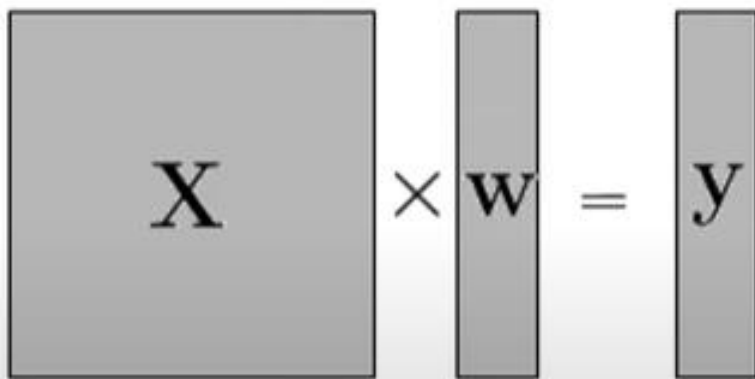
$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_\ell \end{pmatrix}$$

Найти:  $\mathbf{w}$



# Линейная регрессия: точное решение (2/2)

Случай  $\ell = d$ , тогда матрица  $\mathbf{X}$  – квадратная и может иметь обратную. Ищем  $\mathbf{X}^{-1}$ .


$$\mathbf{X} \times \mathbf{w} = \mathbf{y}$$

Система линейных уравнений:

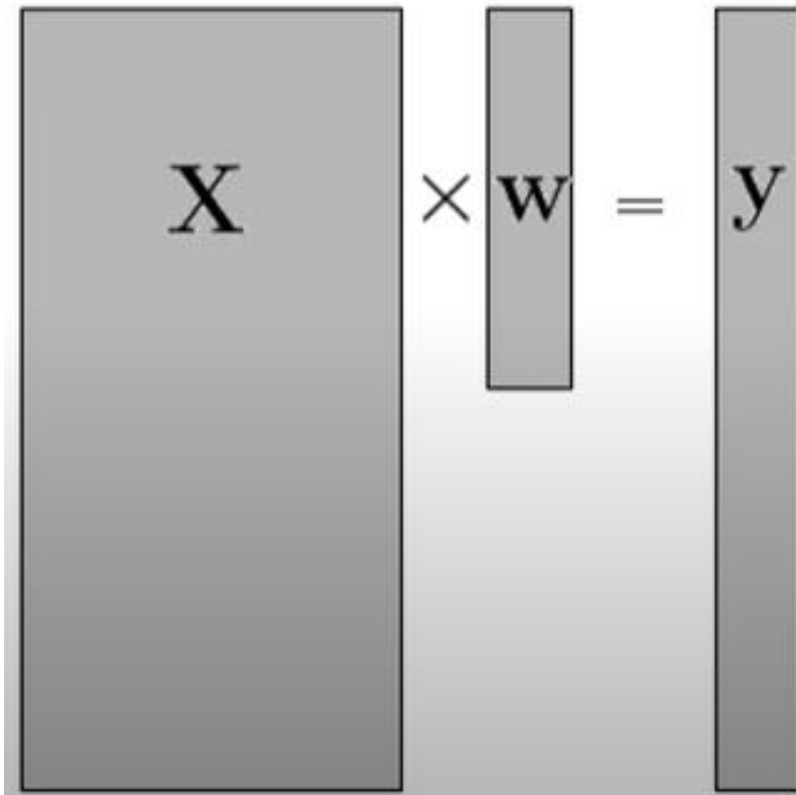
$$\mathbf{X}\mathbf{w} = \mathbf{y}$$

Решение:

$$\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$$

# Линейная регрессия: приближенное решение

Обычно  $\ell \gg d$  (большие данные), тогда имеем переопределенную систему линейных уравнений, матрица  $X$  – прямоугольная,  $\nexists X^{-1}$ . Ищем псевдообратную матрицу  $X^+ = (X^T X)^{-1} X^T$ .


$$X \times w = y$$

Система линейных уравнений:

$$Xw = y$$

Приближенное решение:

$$w = X^+ y = (X^T X)^{-1} X^T y$$

# Линейная регрессия: точное решение через производную (1/2)

Функция потерь

$$\mathcal{L}(a(\mathbf{x}_i, \mathbf{w}), y_i) = \sum_{i=1}^{\ell} (\mathbf{x}_i^T \mathbf{w} - y_i)^2 = (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

Возьмем производную:

$$\frac{\partial}{\partial \mathbf{w}} \mathcal{L}(a(\mathbf{X}, \mathbf{w}), \mathbf{y}) = 2\mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

# Линейная регрессия: точное решение через производную (2/2)

Функция потерь

$$\mathcal{L}(a(\mathbf{x}_i, \mathbf{w}), y_i) = \sum_{i=1}^{\ell} (\mathbf{x}_i^T \mathbf{w} - y_i)^2 = (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

Если столбцы  $\mathbf{X}$  линейно независимы, можно приравнять производную к нулю (иначе  $\nexists (\mathbf{X}^T \mathbf{X})^{-1}$ ):

$$\mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{y}) = 0$$

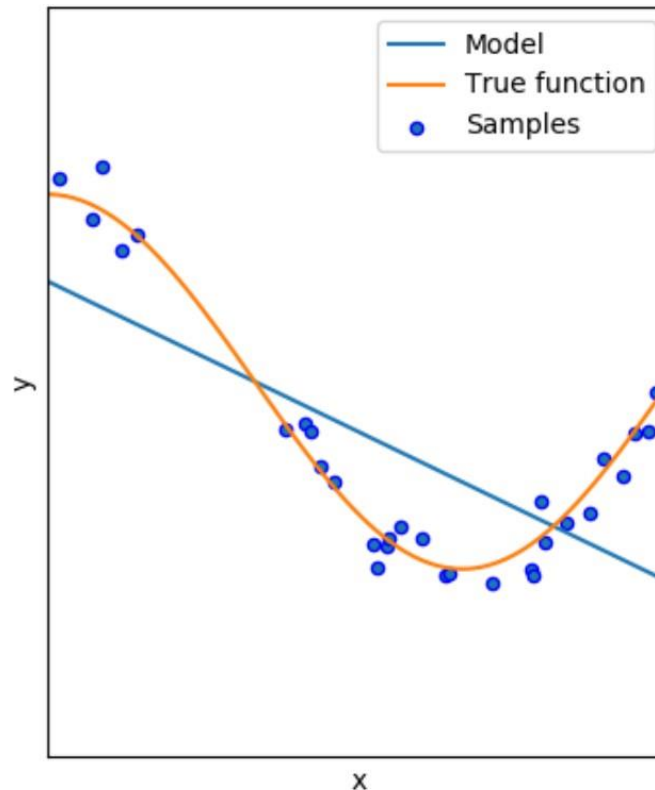
$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# Переобучение и регуляризация линейных моделей

# Нелинейная задача (1/3)

Линейная регрессия 1-го порядка:

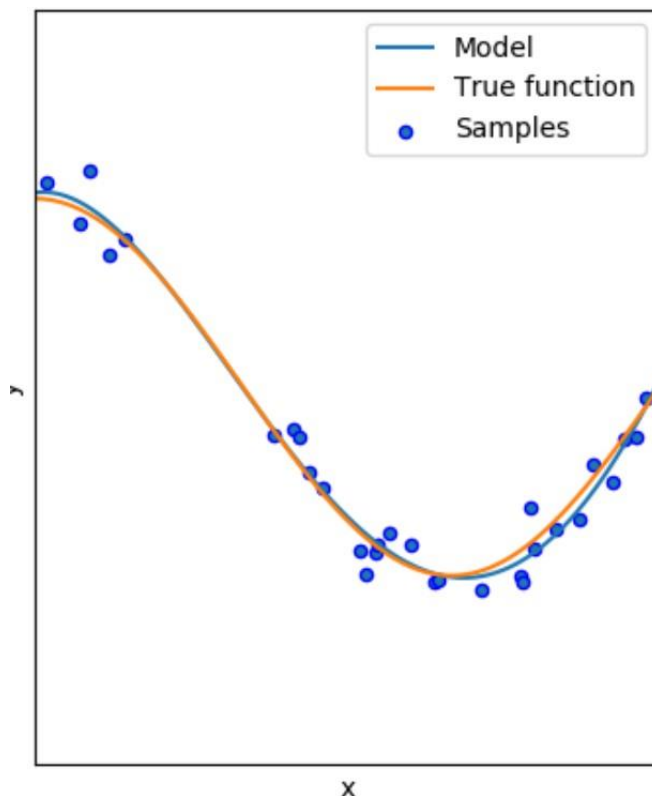
$$a(x) = w_0 + w_1 x$$



## Нелинейная задача (2/3)

Полиномиальная регрессия 4-го порядка:

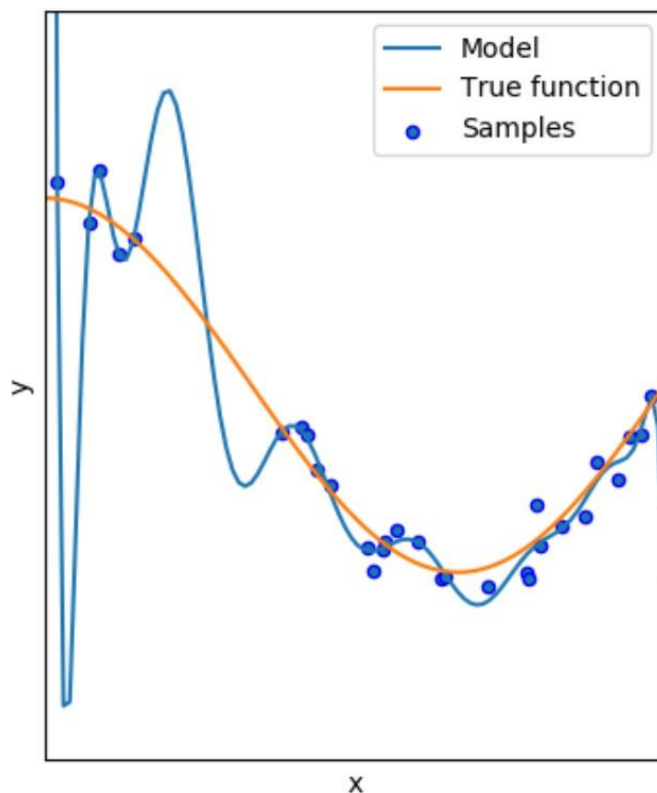
$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4$$



## Нелинейная задача (3/3)

Полиномиальная регрессия 15-го порядка:

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$





## Симптом переобучения (1/2)

Переобученная модель наилучшим образом подстроилась под сложную зависимость в обучающей выборке, но будет плохо работать на новых данных.

Обнаружить переобучение можно на валидационной выборке (validation set).

*Переобучение* – ситуация, когда качество модели на обучающей выборке значительно лучше, чем на валидационной выборке.

## Симптом переобучения (2/2)

Большие коэффициенты в линейной модели — плохо, это симптом переобучения (эмпирическое наблюдение):

$$a(x) = 0.5 + 13458922x - 43983740x^2 + \dots$$

Пример: предсказание роста по весу

$$a(x) = 698x - 41714$$

Изменение веса на 0.01 кг приведет к изменению роста на 7 см - очевидно, неверно.

# Регуляризация (1/4)

*Регуляризация* – множество методов, предназначенных для упрощения структуры модели (чтобы избежать подгонки под обучающую выборку).

Приём: штрафовать за большие веса.

Пример функционала качества:

$$Q_{MSE}(a(\mathbf{X}, \mathbf{w}), \mathbf{y}) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

Регуляризаторы:

$$\|\mathbf{w}\|^2 = \sum_{j=1}^d w_j^2 \text{ (} L_2\text{-норма) или } \|\mathbf{w}\| = \sum_{j=1}^d |w_j| \text{ (} L_1\text{-норма)}$$

## Регуляризация (2/4)

Гребневая регрессия (Ridge regression) — один из методов регуляризации.

Регуляризованный функционал:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}},$$

где  $\lambda \geq 0$  — коэффициент регуляризации (гиперпараметр, подбирать по валидационной выборке),  $\|\mathbf{w}\|^2 = \sum_{i=1}^{\ell} w_i^2$

## Регуляризация (3/4)

Регуляризованный функционал при гребневой регрессии

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$$

Аналитическое решение:

$$\mathbf{w} = (\mathbf{X}^T + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

## Регуляризация (4/4)

LASSO (Least Absolute Shrinkage and Selection Operator) — другой метод регуляризации.

Регуляризованный функционал при LASSO:

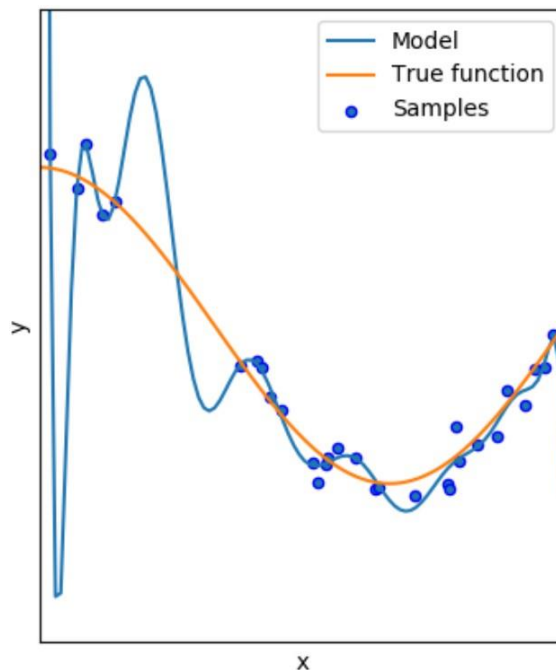
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d |w_j| \rightarrow \min_{\mathbf{w}}$$

При использовании  $L_1$ -нормы некоторые веса зануляются, что приводит к отбору признаков (отсеваются избыточные).

# Эффект гребневой регрессии (1/4)

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$

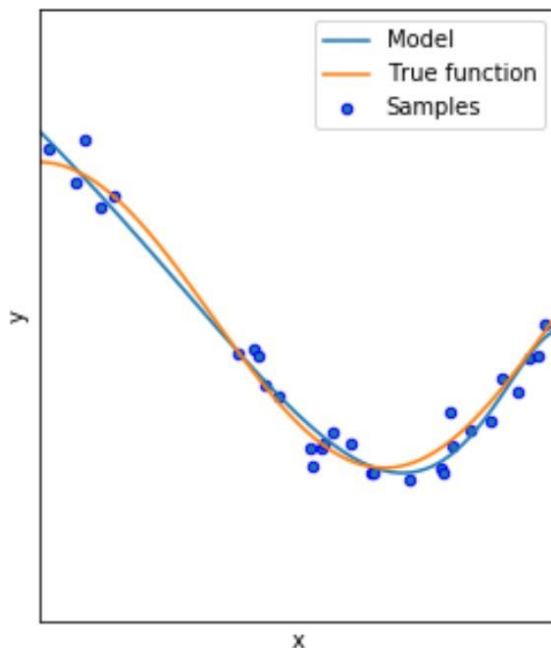
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(\mathbf{x}_i) - y_i)^2 \rightarrow \min_{\mathbf{w}}$$



## Эффект гребневой регрессии (2/4)

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(\mathbf{x}_i) - y_i)^2 + \mathbf{0.01} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$$

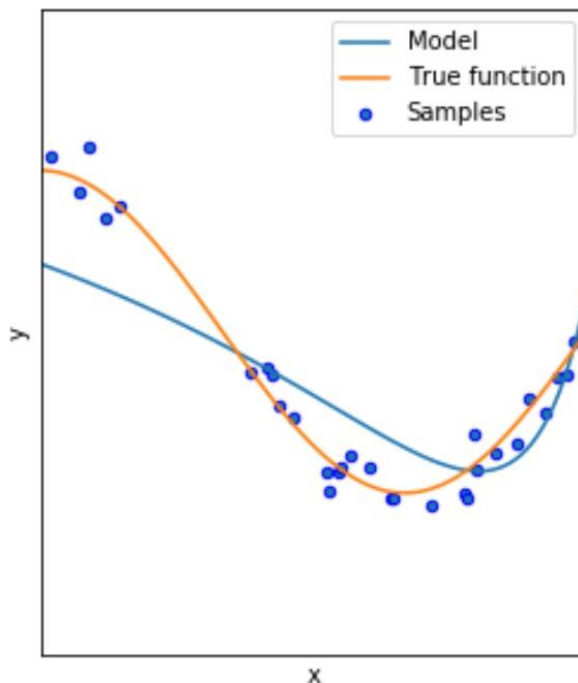




## Эффект гребневой регрессии (3/4)

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$

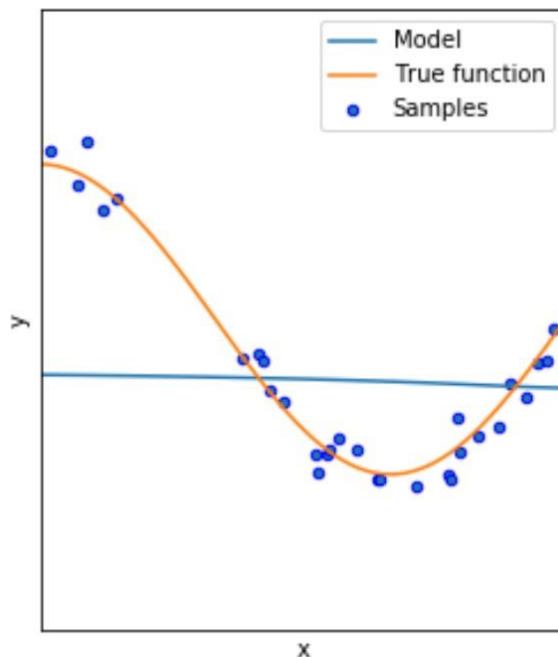
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(\mathbf{x}_i) - y_i)^2 + \mathbf{1} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$$



## Эффект гребневой регрессии (4/4)

$$a(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + \dots + w_{15}x^{15}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(\mathbf{x}_i) - y_i)^2 + \textcolor{red}{100} \|\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$$



# Регуляризаторы

- $\|z\|_2 = \sqrt{\sum_{j=1}^d z_j^2}$  —  $L_2$ -норма
- $\|z\|_1 = \sum_{j=1}^d |z_j|$  —  $L_1$ -норма

# Интерпретация линейных моделей

# Прогнозирование стоимости квартиры

$$\begin{aligned} a(\mathbf{x}, \mathbf{w}) = & 100.000 * (\text{площадь}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

# Прогнозирование стоимости квартиры

$$\begin{aligned} a(\mathbf{x}, \mathbf{w}) = & 100.000 * (\text{площадь}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?

# Прогнозирование стоимости квартиры

$$a(\mathbf{x}, \mathbf{w}) = 100.000 * (\text{площадь в кв. м.}) \\ + 500.000 * (\text{число магазинов рядом}) \\ + 100 * (\text{средний доход жильцов дома})$$

- Чем больше вес, тем важнее признак?

# Прогнозирование стоимости квартиры

$$a(\mathbf{x}, \mathbf{w}) = 10 * (\text{площадь в кв. см.}) \\ + 500.000 * (\text{число магазинов рядом}) \\ + 100 * (\text{средний доход жильцов дома})$$

- Чем больше вес, тем важнее признак?



# Прогнозирование стоимости квартиры

$$a(\mathbf{x}, \mathbf{w}) = 100.000 * (\text{площадь в кв. м.}) \\ + 500.000 * (\text{число магазинов рядом}) \\ + 100 * (\text{средний доход жильцов дома})$$

- Чем больше вес, тем важнее признак?

# Прогнозирование стоимости квартиры

$$a(\mathbf{x}, \mathbf{w}) = 100.000 * (\text{площадь в кв. м.}) \\ + 500.000 * (\text{число магазинов рядом}) \\ + 100 * (\text{средний доход жильцов дома})$$

- Чем больше вес, тем важнее признак?
- Только если признаки масштабированы!

# Масштабирование признаков (1/2)

1. Для  $j$ -го признака,  $j = 1, \dots, d$ , вычисляем среднее значение и стандартное (среднеквадратическое) отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{\ell} \sum_{i=1}^{\ell} x_{ij}$$

$$\sigma_j = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (x_{ij} - \mu_j)^2}$$

## Масштабирование признаков (2/2)

2. Вычитаем из каждого значения признака среднее по выборке и делим на стандартное отклонение по выборке:

$$\hat{x}_{ij} := \frac{x_{ij} - \mu_j}{\sigma_j}, \quad i = 1, \dots, \ell, \quad j = 1, \dots, d$$

# Особенности регуляризации

- Если модель переобучается, то веса используются для запоминания обучающей выборки
- Правильнее масштабировать признаки и регуляризовать модель перед изучением весов

# Pros & Cons линейной регрессии

# Плюсы метода линейной регрессии

- результат легко интерпретируется, более того, по значениям весов наглядно видны наиболее значимые признаки
- модель не переобучается (нетрудно избежать переобучения)
- хорошо работает даже на выборках небольшого объема, когда другие модели не справляются (напр., для нейронных сетей недостаточно данных, а для решающих деревьев наступит переобучение)
- не требует специальных знаний при применении
- простая функция потерь (квадратичная ошибка)

# Минусы метода линейной регрессии

- хорошо описывает только зависимости близкие к линейным

Например, если попробовать приблизить параболу с вершиной в нуле на симметричном отрезке  $[-a, a]$  с помощью линейной регрессии, то получится константа.

- требует предобработки данных, масштабирования