

# Введение в анализ данных

## Лекция 4

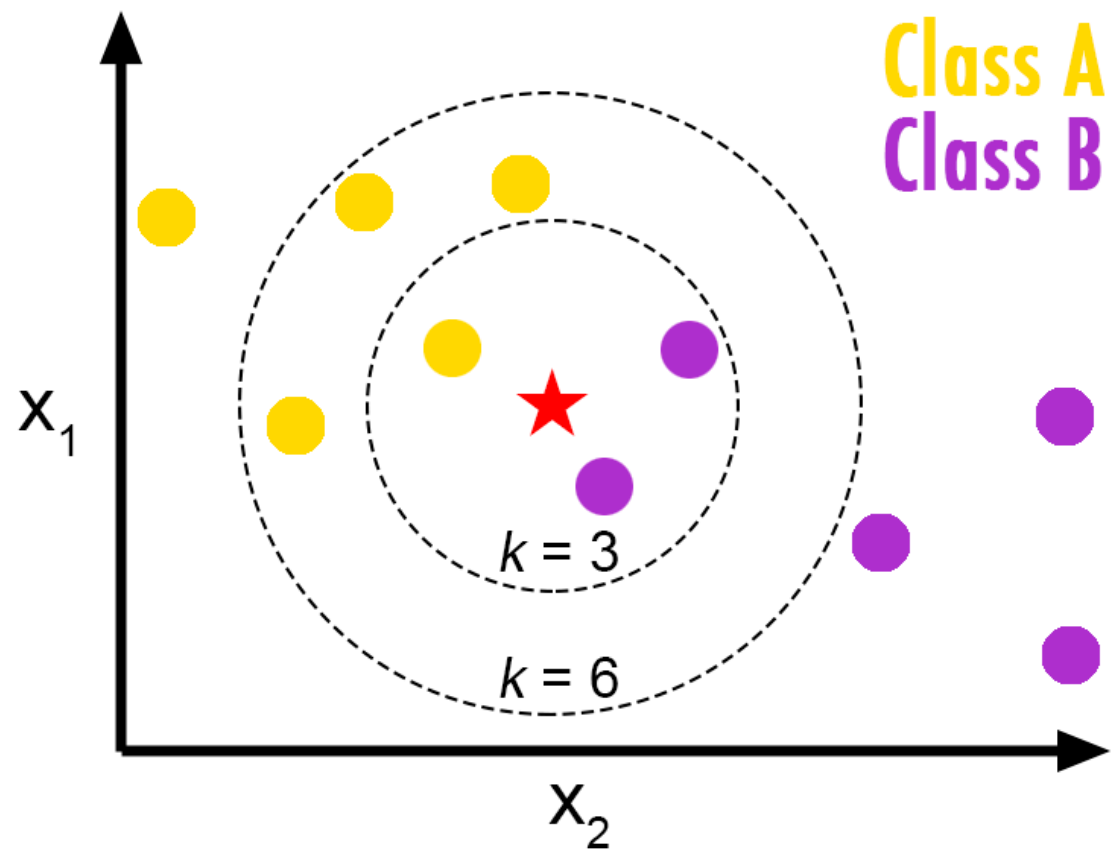
Метрические и линейные методы регрессии

Евгений Соколов

[esokolov@hse.ru](mailto:esokolov@hse.ru)

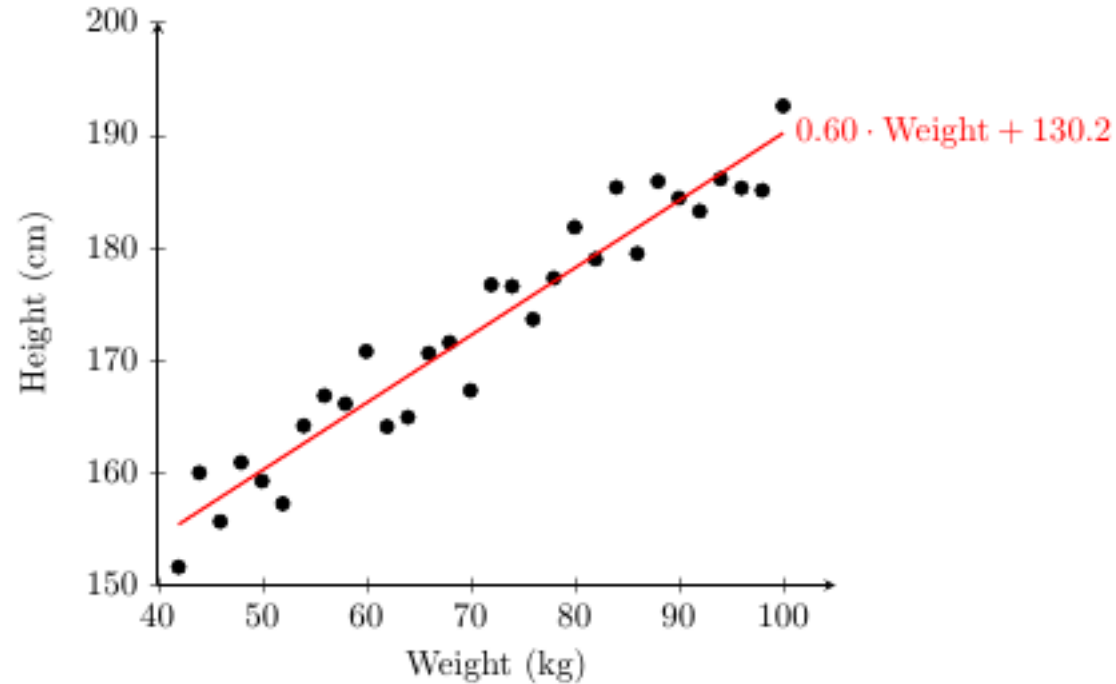
НИУ ВШЭ, 2017

# Метод k ближайших соседей



# Регрессия

- Вещественные ответы:  $\mathbb{Y} = \mathbb{R}$
- (вещественные числа — числа с любой дробной частью)
- Пример: предсказание роста по весу



Среднеквадратичная ошибка

# Функционал ошибки

$a(x)$	$y$	отклонение
11	10	1
9	10	-1
20	10	10
1	10	-9

# Функционал ошибки

- Ошибку надо минимизировать
- Минимизация отклонения ( $a(x) - y$ ) приведёт к провалу

$a(x)$	$y$	отклонение
11	10	1
9	10	-1
20	10	10
1	10	-9

# Функционал ошибки

- Возьмём модуль:  $|a(x) - y|$
- Не имеет производной

$a(x)$	$y$	$ a(x) - y $
11	10	1
9	10	1
20	10	10
1	10	9

# Функционал ошибки

- Возведём в квадрат:  $(a(x) - y)^2$

$a(x)$	$y$	$(a(x) - y)^2$
11	10	1
9	10	1
20	10	100
1	10	81



# Среднеквадратичная ошибка

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

- MSE (Mean Squared Error)

# Среднеквадратичная ошибка

$$Q(w, X) = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2}$$

- RMSE (Root Mean Squared Error)
- В тех же единицах измерения, что и ответы
- Сложные производные из-за корня

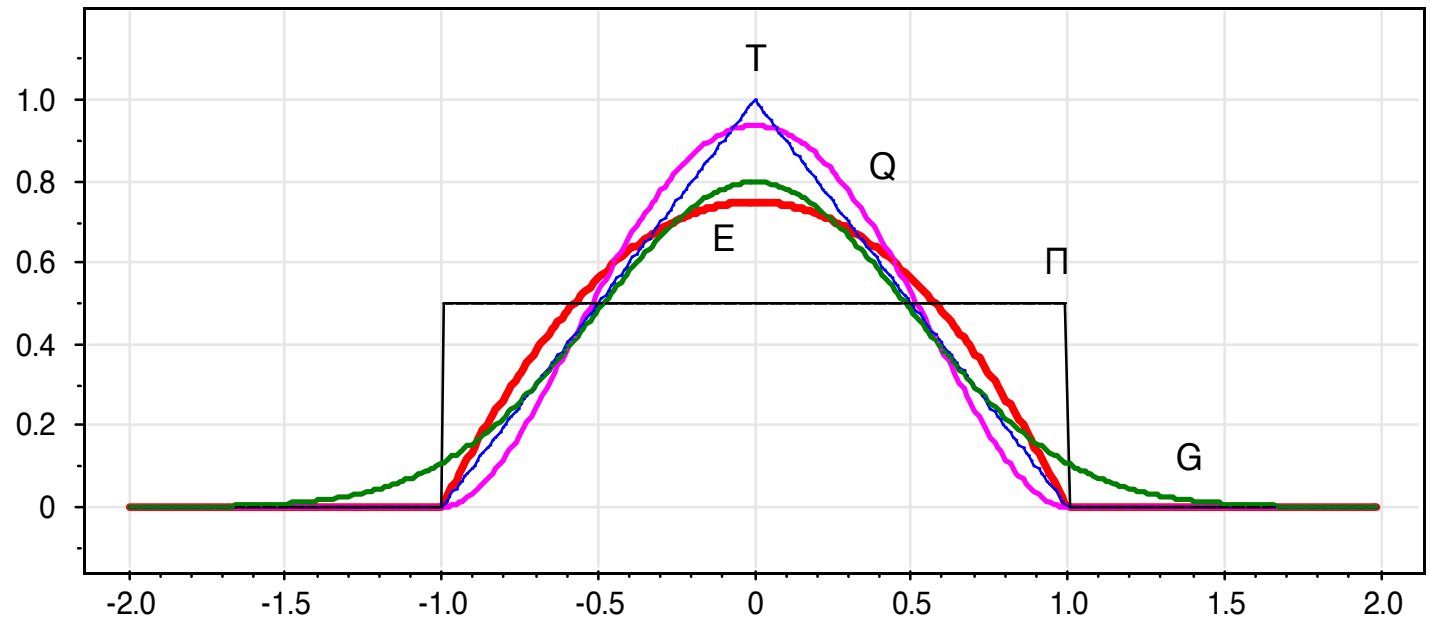
# Метрические методы регрессии

# kNN с весами

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k w_i [y_{(i)} = y]$$

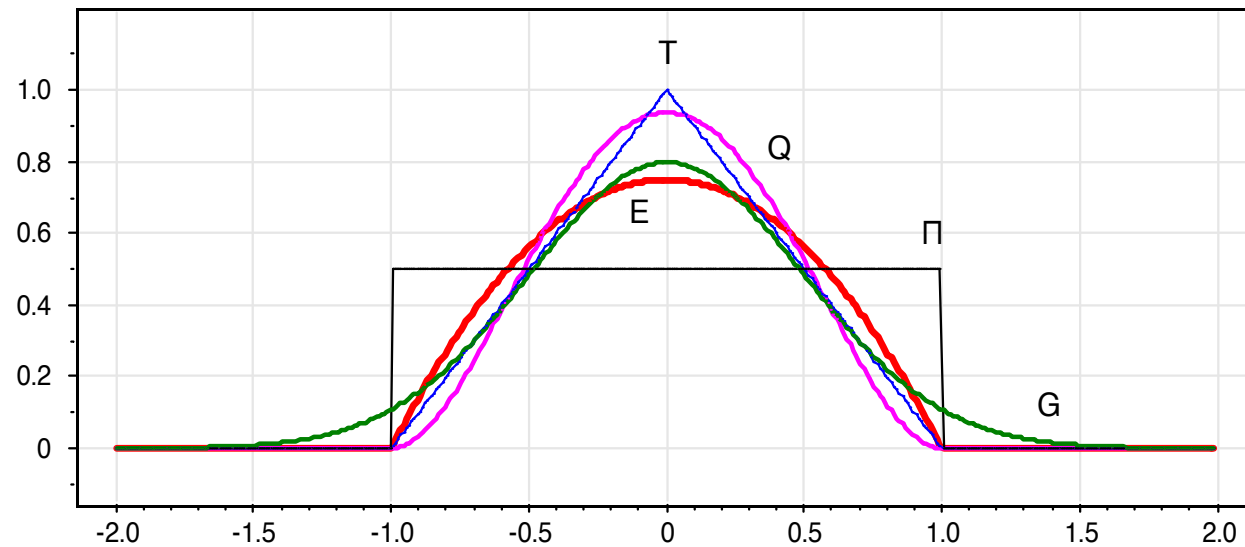
Парзеновское окно:

- $w_i = K \left( \frac{\rho(x, x_{(i)})}{h} \right)$
- $K$  — ядро
- $h$  — ширина окна



# Ядра

- Гауссовское ядро:  $K(z) = (2\pi)^{-0.5} \exp\left(-\frac{1}{2}z\right)$
- И много других



# kNN для регрессии

- Классификация:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k w_i [y_{(i)} = y]$$

- Регрессия:

# kNN для регрессии

- Классификация:

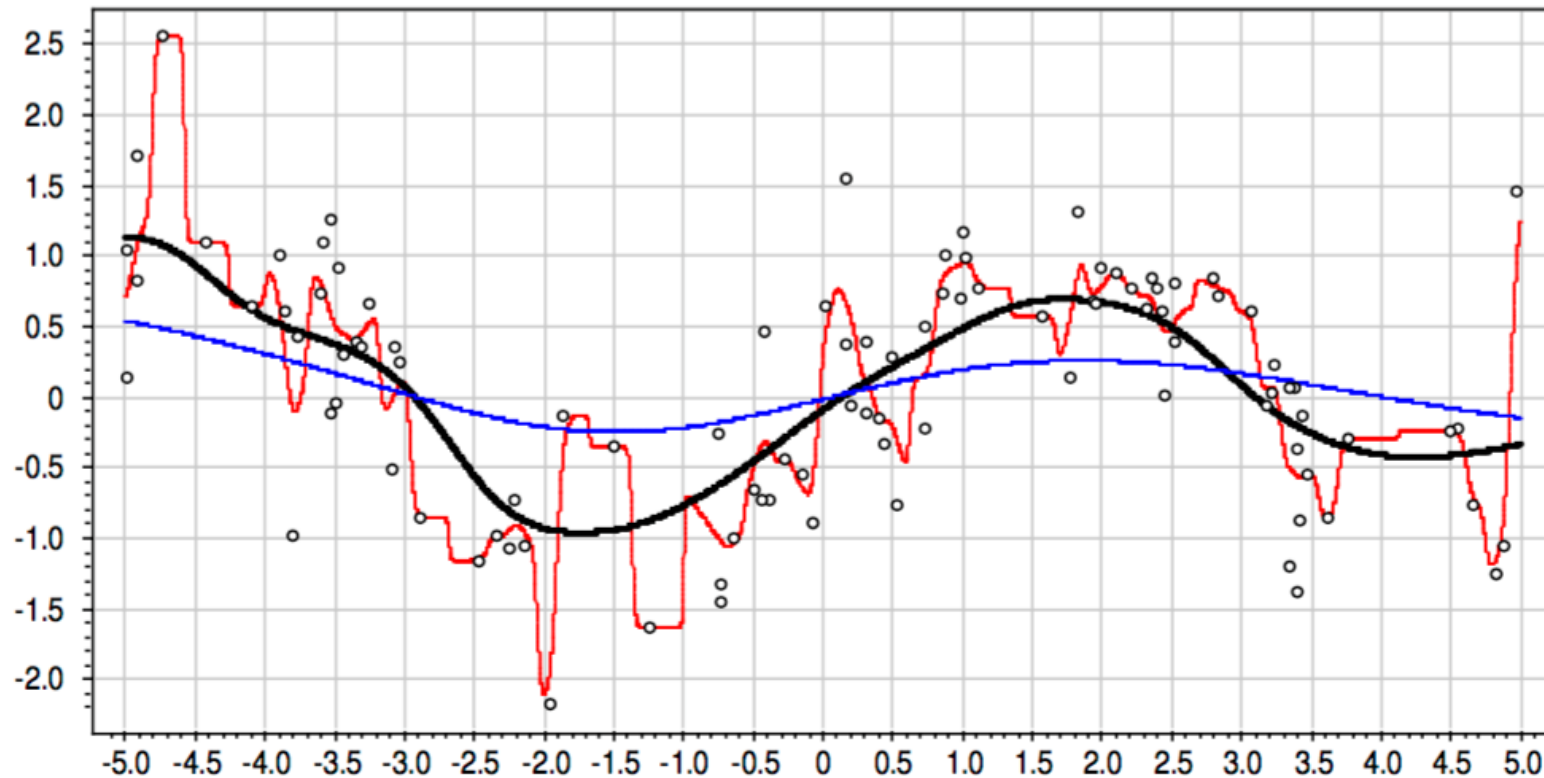
$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k w_i [y_{(i)} = y]$$

- Регрессия:

$$a(x) = \frac{\sum_{i=1}^k w_i y_{(i)}}{\sum_{i=1}^k w_i}$$

# kNN для регрессии

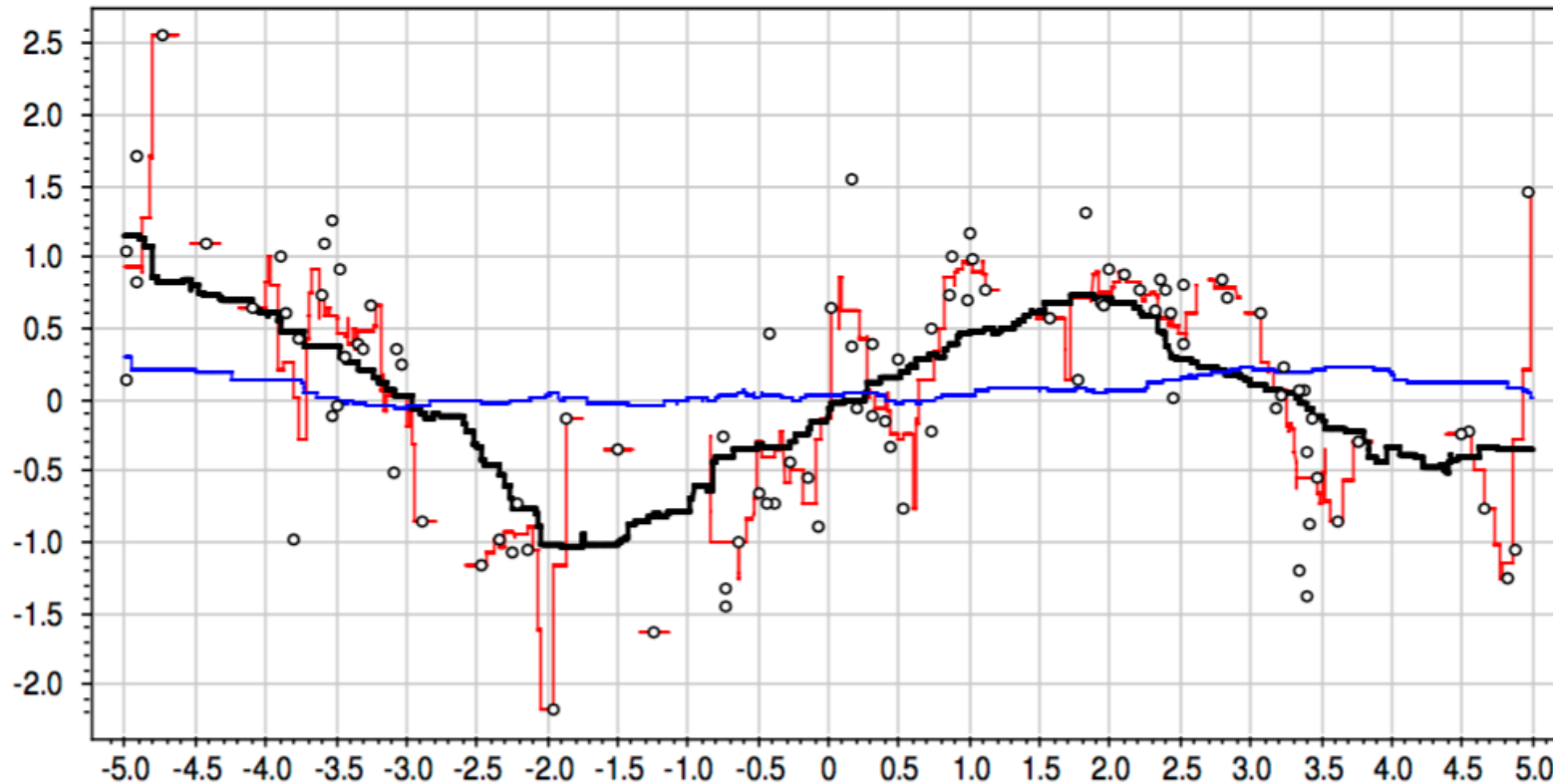
- Гауссовское ядро
- $h \in \{0.1, 1.0, 3.0\}$





# kNN для регрессии

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



Функции расстояния

# Евклидова метрика

$$\rho(x, z) = \sqrt{\sum_{j=1}^d (x_j - z_j)^2}$$

- Более общий вариант — метрика Минковского:

$$\rho(x, z) = \left( \sum_{j=1}^d (x_j - z_j)^p \right)^{1/p}$$

# Чувствительность к масштабу

- Задача: определение пола
- Признаки:
  - Рост
  - Экспрессия гена SRY (от 0 до 1) — у женщин ближе к нулю
- Обучающая выборка:
  - $x_1 = (180, 0.2)$
  - $x_2 = (172, 0.9)$
- Новый объект:  $x = (178, 0.85)$

# Чувствительность к масштабу

- Задача: определение пола
- Признаки:
  - Рост
  - Экспрессия гена SRY (от 0 до 1) — у женщин ближе к нулю
- Обучающая выборка:
  - $x_1 = (180, 0.2)$
  - $x_2 = (172, 0.9)$
- Новый объект:  $x = (178, 0.85)$
- $\rho(x, x_1) = 2.1, \rho(x, x_2) = 5$

# Чувствительность к масштабу

- Если признаки имеют разные масштабы, то будут учитываться лишь самые крупные
- Перед применением kNN выборку необходимо масштабировать!

# Расстояние Джаккарда

- Измеряет расстояния между множествами
- Пример: каждый объект — набор слов или тэгов
- Метрика:

$$\rho(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

# Расстояние Джаккарда

- Пример 1:

- $A = \{\text{комедия, триллер, США}\}$

- $B = \{\text{триллер, ужасы, Великобритания}\}$

- $\rho(A, B) = 1 - \frac{1}{5} = 0.8$

- Пример 2:

- $A = \{\text{комедия, США}\}$

- $B = \{\text{комедия, США}\}$

- $\rho(A, B) = 1 - \frac{2}{2} = 0$

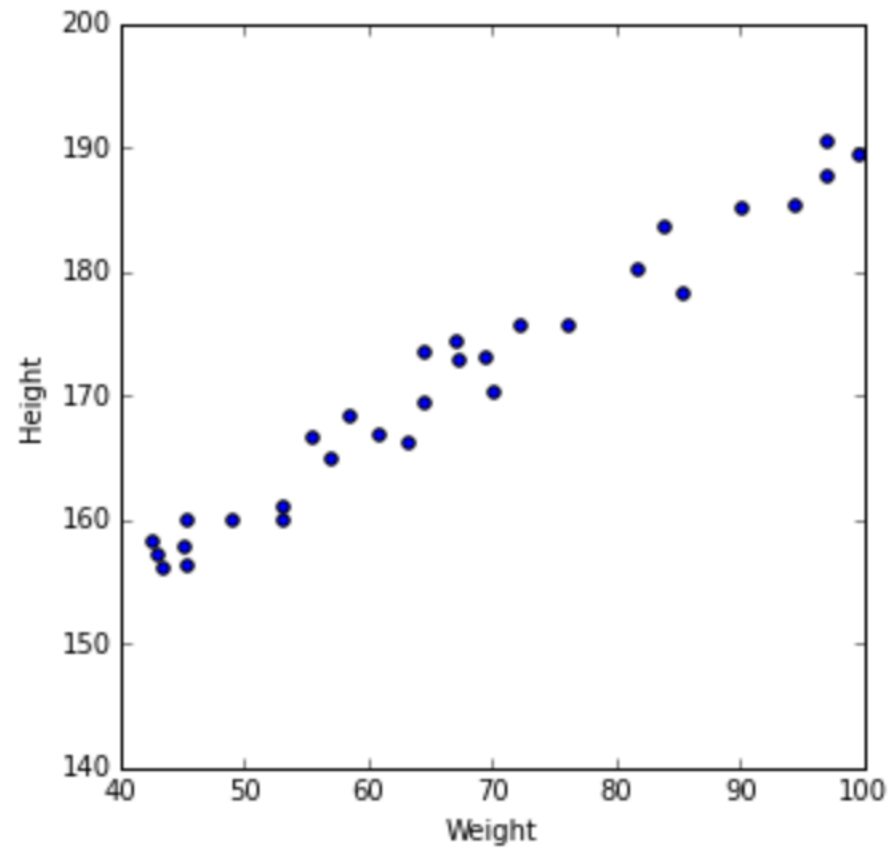


# Резюме по kNN

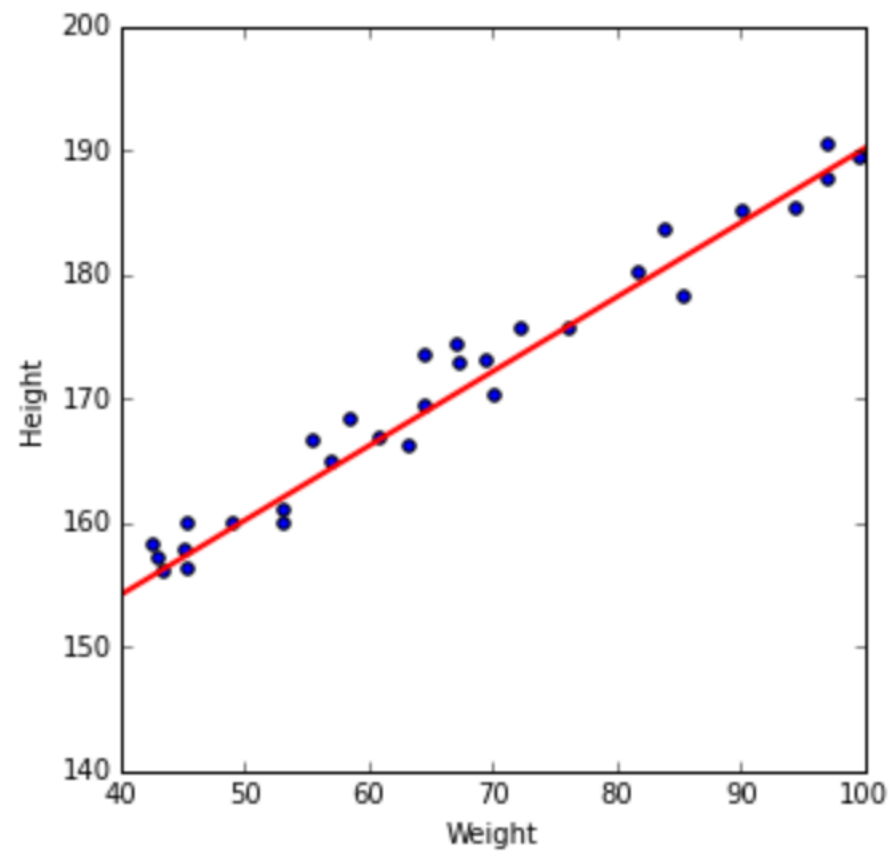
- Метрические методы — одни из самых интуитивных в машинном обучении
- Простая процедура обучения
- Гиперпараметры:
  - функция расстояния
  - число соседей
  - ядро
  - ширина окна

# Линейная регрессия

# Одномерная выборка



# Одномерная выборка



# Парная регрессия

- Простейший случай: один признак
- Модель:  $a(x) = w_1x + w_0$
- Два параметра:  $w_1$  и  $w_0$
- Одна из простейших моделей

# Линейная регрессия

- Взвешенная сумма признаков:


$$a(x) = w_0 + w_1x^1 + \dots + w_dx^d$$

- $x^1, x^2, \dots, x^d$  — значений признаков
- $w_0, w_1, w_2, \dots, w_d$  — параметры
- $w_0$  — смещение

# Линейная регрессия

- Взвешенная сумма признаков:

$$a(x) = w_0 + w_1x^1 + \dots + w_dx^d$$

- $x^1, x^2, \dots, x^d$  — значений признаков
  - $w_0, w_1, w_2, \dots, w_d$  — параметры
  - $w_0$  — смещение
- 

# Единичный признак

$$a(x) = w_0 * 1 + w_1 x^1 + \dots + w_d x^d$$

- $w_0$  — как бы коэффициент при единичном признаке
- Добавим его!

$$\begin{pmatrix} x_{11} & \dots & x_{1d} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{\ell 1} & \dots & x_{\ell d} & 1 \end{pmatrix}$$



# Линейная регрессия

- Везде далее считаем, что среди признаков есть единичный

$$a(x) = w_1 x^1 + \dots + w_d x^d = \langle w, x \rangle$$



Скалярное  
произведение

# Линейная регрессия

- Линейная модель:  $a(x) = w_1x^1 + \dots + w_dx^d = \langle w, x \rangle$
- Обучение:

$$\sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w$$

Функция с  $d$  аргументами

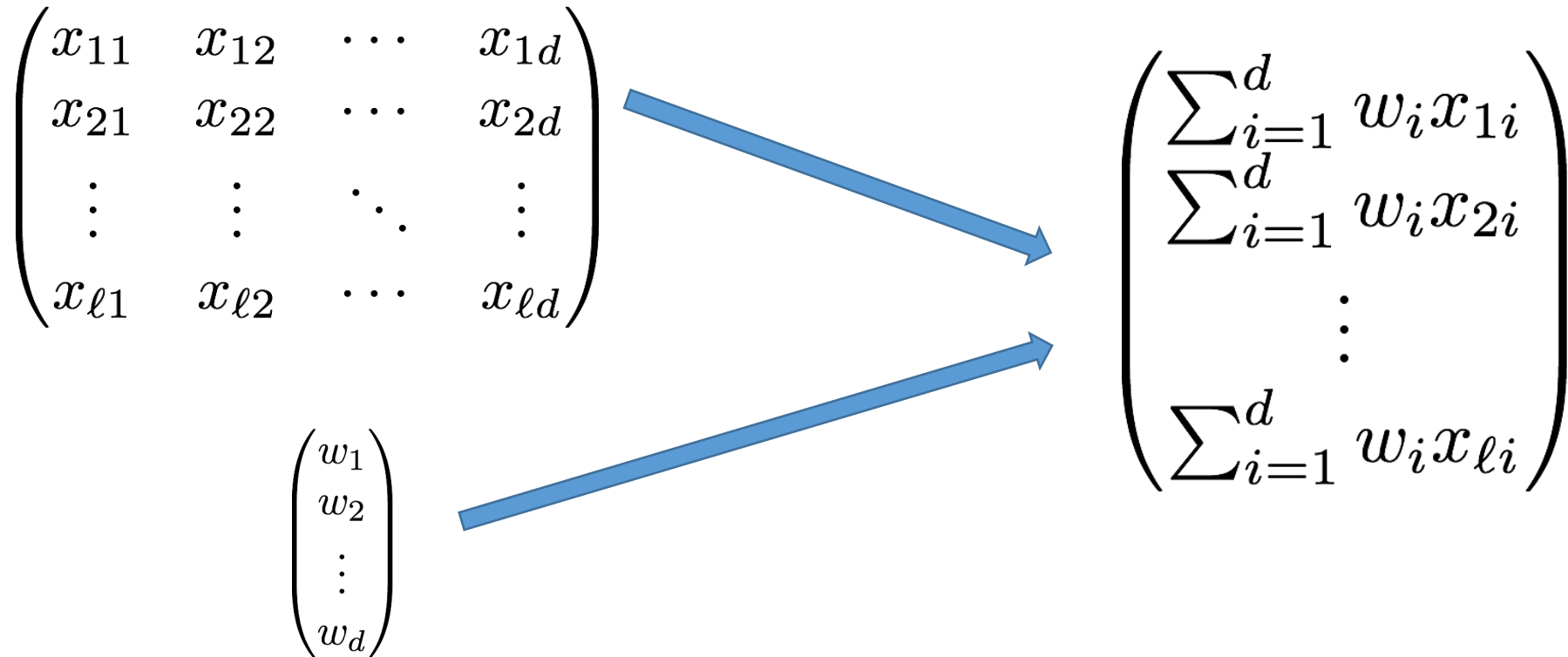
Умножение матриц и MSE

# Векторы и матрицы

- Вектор размера  $d$  — тоже матрица
- Вектор-строка:  $w = (w_1, \dots, w_d) \in \mathbb{R}^{1 \times d}$
- Вектор-столбец:  $w = \begin{pmatrix} w_1 \\ \dots \\ w_d \end{pmatrix} \in \mathbb{R}^{d \times 1}$

# Линейная модель

- $a(x) = w_1x^1 + \dots + w_dx^d$
- Как применить модель к целой выборке?



# Умножение

- Мы еще не вводили умножение матрицы на вектор
- Определим его именно так
- Только для матрицы  $\ell \times d$  и вектора  $d \times 1$

$$Xw = \begin{pmatrix} \sum_{i=1}^d w_i x_{1i} \\ \sum_{i=1}^d w_i x_{2i} \\ \vdots \\ \sum_{i=1}^d w_i x_{\ell i} \end{pmatrix}$$

# Линейные преобразования

- Умножая матрицу  $A \in \mathbb{R}^{m \times n}$  на вектор  $x \in \mathbb{R}^{n \times 1}$ , получаем вектор  $z \in \mathbb{R}^{m \times 1}$
- Матрица задает функцию из  $\mathbb{R}^{n \times 1}$  в  $\mathbb{R}^{m \times 1}$
- Эта функция — линейная:
  - $A(x_1 + x_2) = Ax_1 + Ax_2$
  - $A(\alpha x) = \alpha Ax$
- Любая линейная функция описывается некоторой матрицей

# Линейные преобразования

- Функции можно применять последовательно:  $g(f(x))$
- В том числе линейные:  $A(Bx)$
- Композиция линейных функций — тоже линейная функция:
  - $A(B(x_1 + x_2)) = A(Bx_1) + A(Bx_2)$
  - $A(B(\alpha x)) = \alpha A(Bx)$
- А какая у нее матрица?
- Зададим матричное умножение так, чтобы оно соответствовало композиции линейных преобразований



# Матричное умножение

- Только для матриц  $A \in \mathbb{R}^{m \times k}$  и  $A \in \mathbb{R}^{k \times n}$
- Результат:  $AB = C \in \mathbb{R}^{m \times n}$
- Правило:

$$c_{ij} = \sum_{p=1}^k a_{ip} b_{pj}$$

Пример

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

Пример

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

Пример

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

Пример

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

Пример

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

# Векторный вид MSE

$$Q(w, X) = \frac{1}{\ell} \|Xw - y\|^2$$

- $X$  — матрица объекты-признаки
- $y$  — вектор ответов на обучающей выборке

# Производная и градиент



# Скорость роста

- Численность населения:

1950	1960	1970	1980	1990	2000
2,525,778,669	3,026,002,942	3,691,172,616	4,449,048,798	5,320,816,667	6,127,700,428

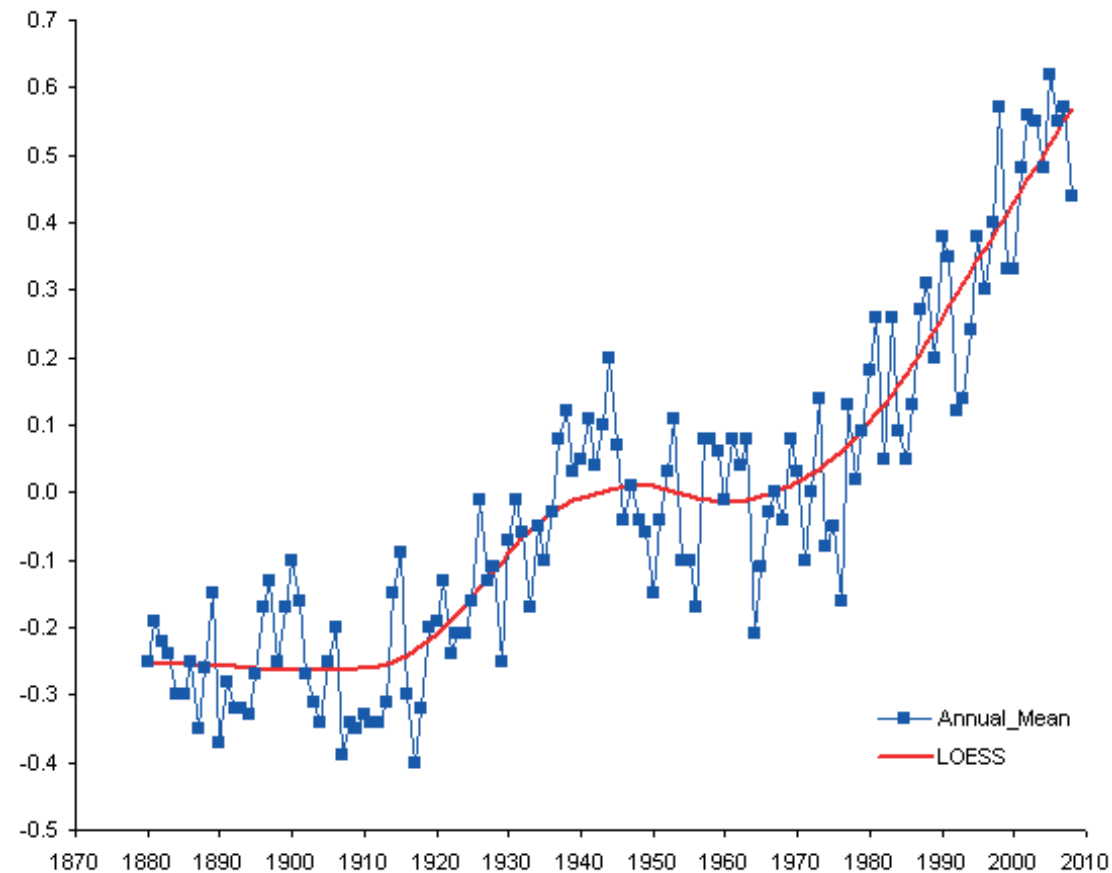
- Скорость роста между 1990 и 2000:

$$\frac{6127700428 - 5320816667}{10} = 80,688,376$$

- Дискретная величина

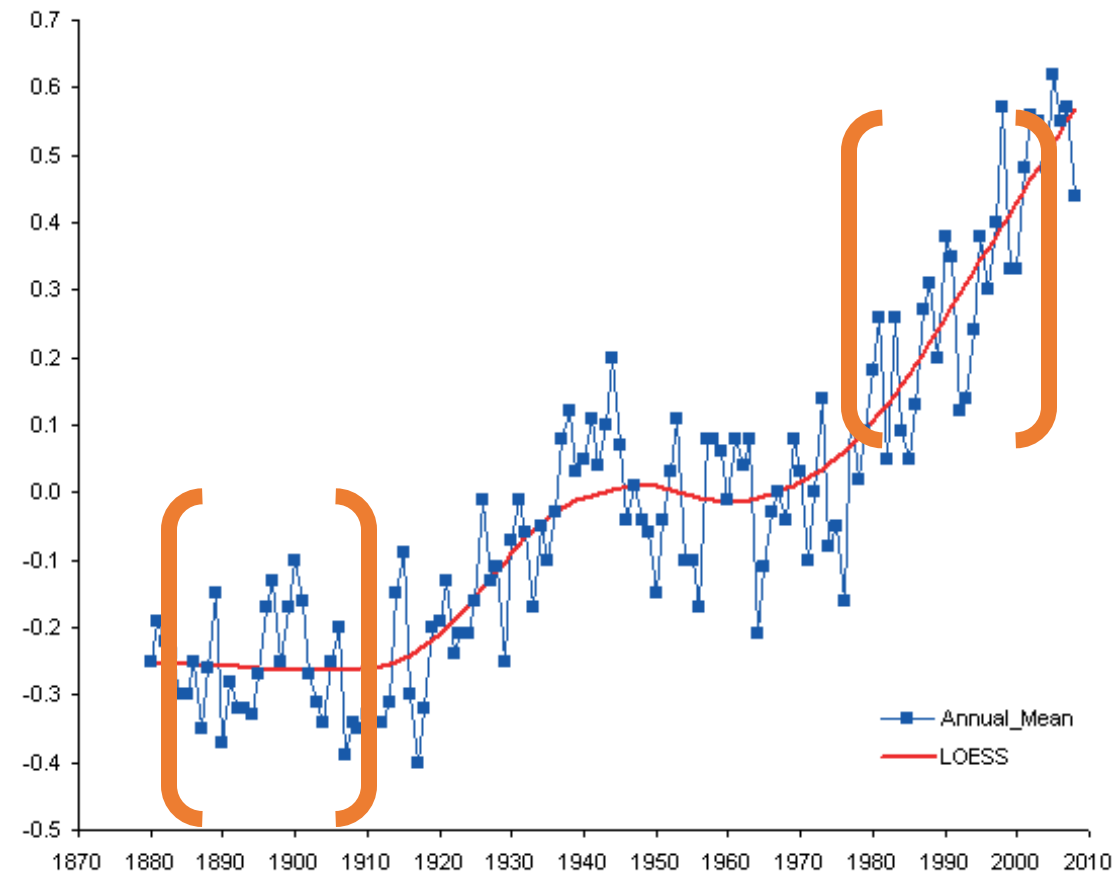
# Скорость роста

- Отклонение температуры от нормы (непрерывная величина):



# Скорость роста

- Отклонение температуры от нормы:



Низкая скорость

Высокая скорость

# Скорость роста

- Можем измерить скорость на интервале  $[x_0, x]$ :

$$\frac{f(x) - f(x_0)}{x - x_0}$$

- Как измерить мгновенную скорость в конкретный момент  $x_0$ ?
- Устремим  $x$  к  $x_0$ !

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

# Скорость роста

- Можем измерить скорость на интервале  $[x_0, x]$ :

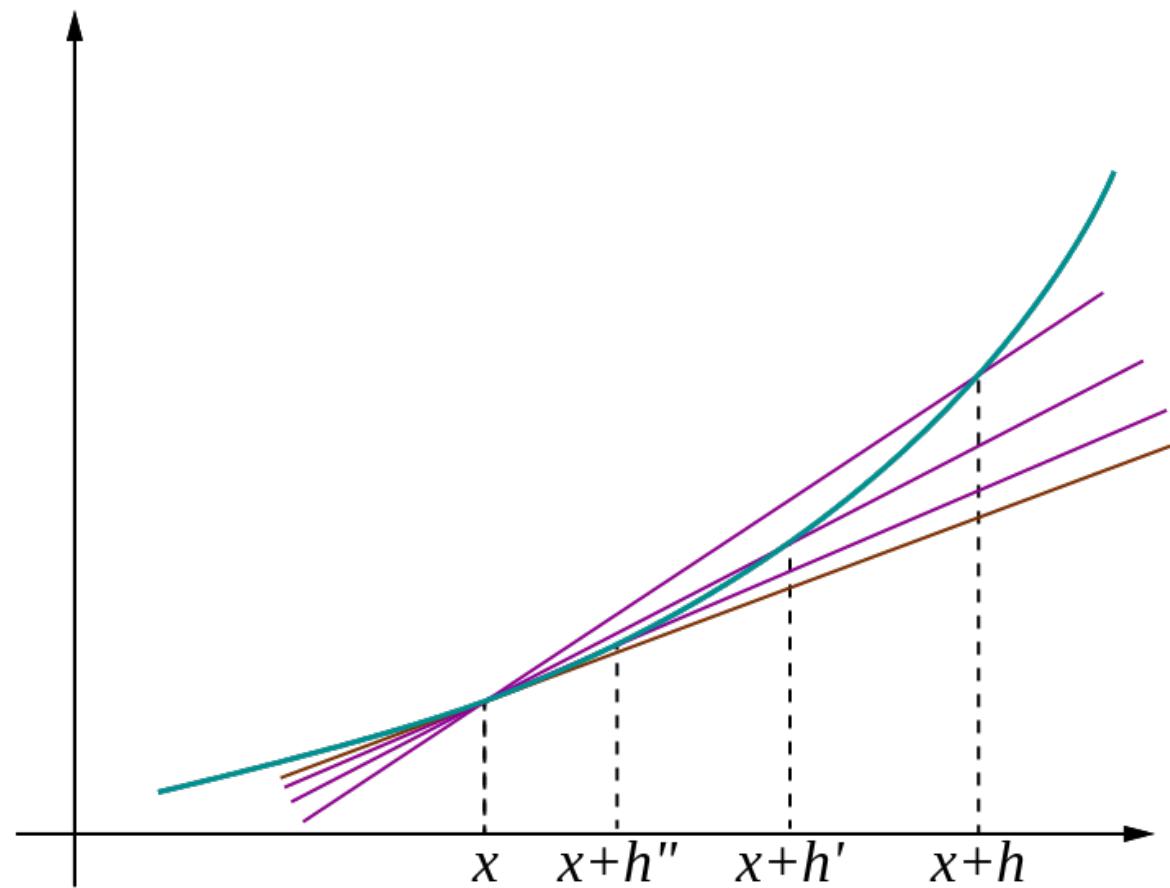
$$\frac{f(x) - f(x_0)}{x - x_0}$$

- Как измерить мгновенную скорость в конкретный момент  $x_0$ ?
- Устремим  $x$  к  $x_0$ !

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

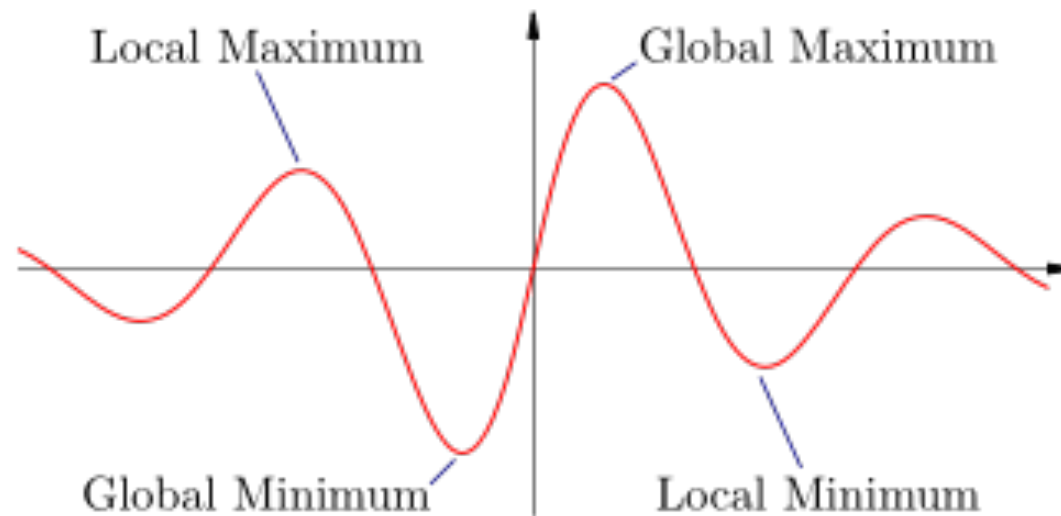
Производная

# Производная



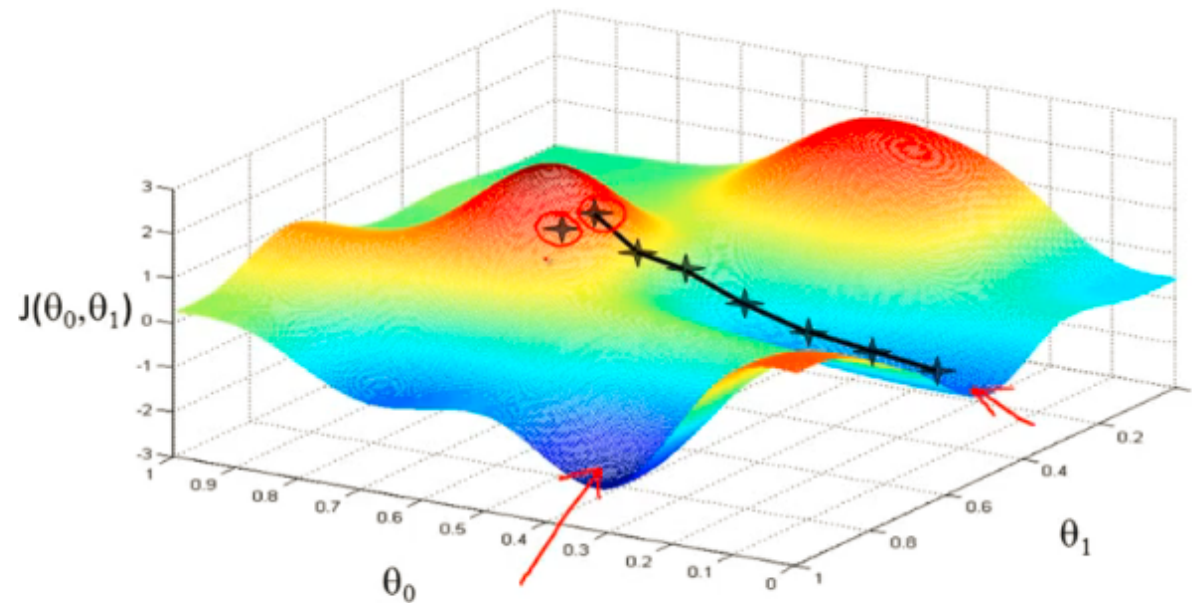
# Экстремумы

- Экстремум — минимум или максимум
- Локальный минимум — меньше всех значений в некоторой окрестности
- Глобальный минимум — меньше всех значений



# Экстремумы

- Локальные минимумы — одна из главных проблем в машинном обучении



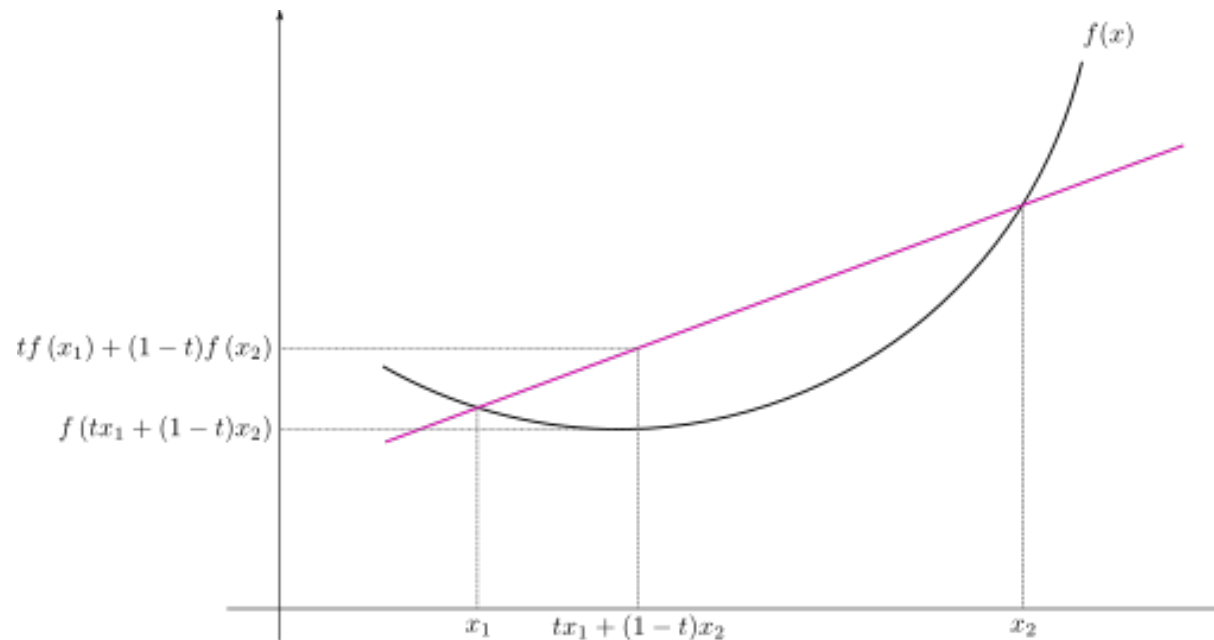


# Условие оптимальности

- Как понять, является ли точка  $x_0$  экстремумом?
- Теорема Ферма: если точка  $x_0$  — экстремум, и в ней существует производная, то  $f'(x_0) = 0$
- Если функция везде имеет производную: решаем  $f'(x) = 0$
- Если с производной проблемы: не повезло
- Даже если производная есть, то что делать с локальными экстремумами?

# Выпуклые функции

- Функция выпуклая, если ее график лежит ниже любого отрезка, соединяющего две точки



# Выпуклые функции

- Функция выпуклая, если во всех точках  $f''(x) \geq 0$
- Важное свойство: любой локальный экстремум выпуклой функции является глобальным
- Решая уравнение  $f'(x) = 0$ , получим глобальные экстремумы
- Вывод: будем стараться выбирать выпуклые функционалы!

# Пример

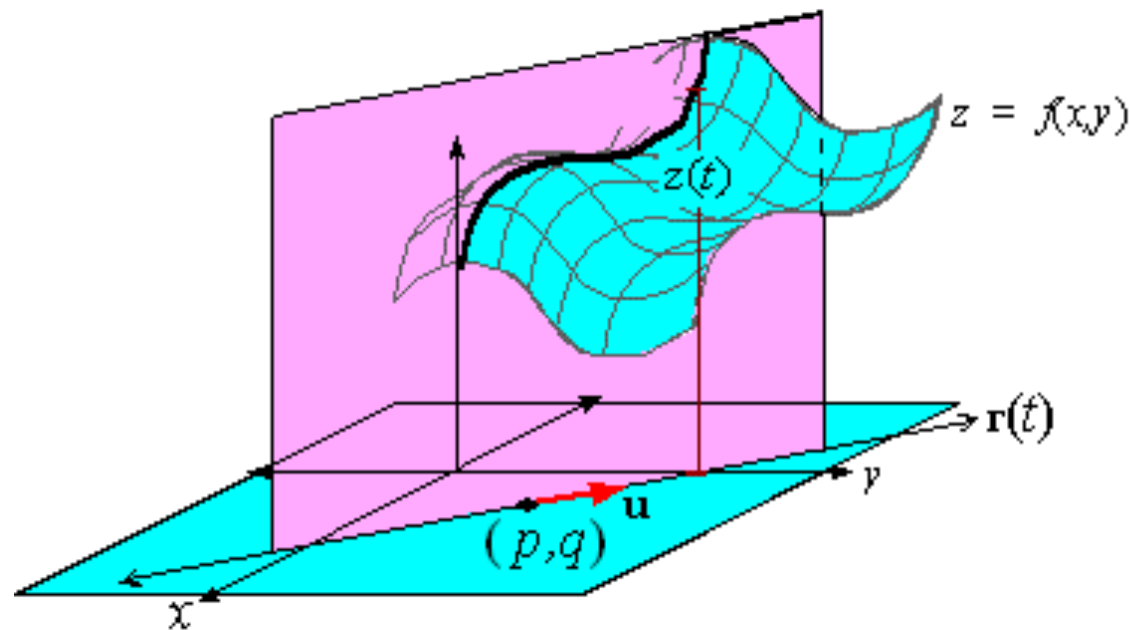
- Функционал качества линейной регрессии:

$$Q(w_1, \dots, w_d) = \sum_{i=1}^{\ell} (w_1 x^1 + \dots + w_d x^d - y_i)^2$$

- Как искать ее минимум?

# Производная по направлению

- С какой скоростью растет функция в конкретном направлении?



# Производная по направлению

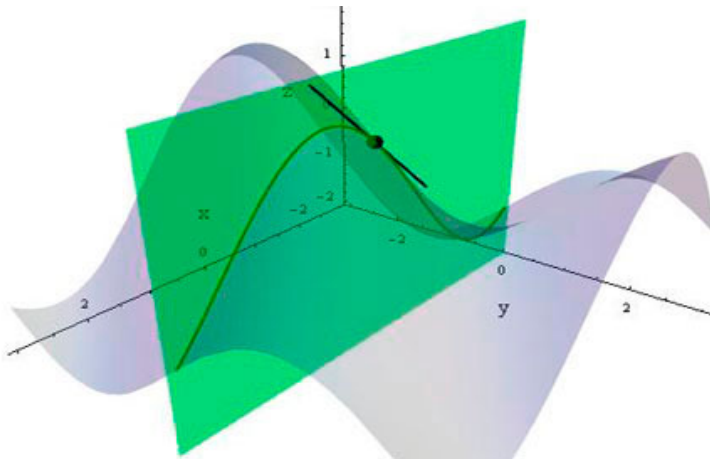
- Направление:  $v$ , причем  $\|v\| = 1$
- Производная:

$$f'_v(x_0) = \lim_{t \rightarrow 0} \frac{f(x_0 + tv) - f(x_0)}{t}$$

# Частные производные

- С какой скоростью функция меняется вдоль переменной  $x_i$ ?
- Частная производная по  $x_i$ :

$$\frac{\partial f}{\partial x_i} = \lim_{t \rightarrow 0} \frac{f(x_1, \dots, x_i + t, \dots, x_d) - f(x_1, \dots, x_i, \dots, x_d)}{t}$$



# Градиент

- Градиент — вектор из частных производных:

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

- У градиента есть очень важное свойство!



# Градиент

- Зафиксируем точку  $x_0$
- В каком направлении функция быстрее всего растёт?

$$f'_v(x_0) \rightarrow \max_v$$

Угол между градиентом и направлением

- Связь производной по направлению и градиента:

$$f'_v(x_0) = \langle \nabla f(x_0), v \rangle = \|\nabla f(x_0)\| * \|v\| * \cos \varphi$$

# Градиент

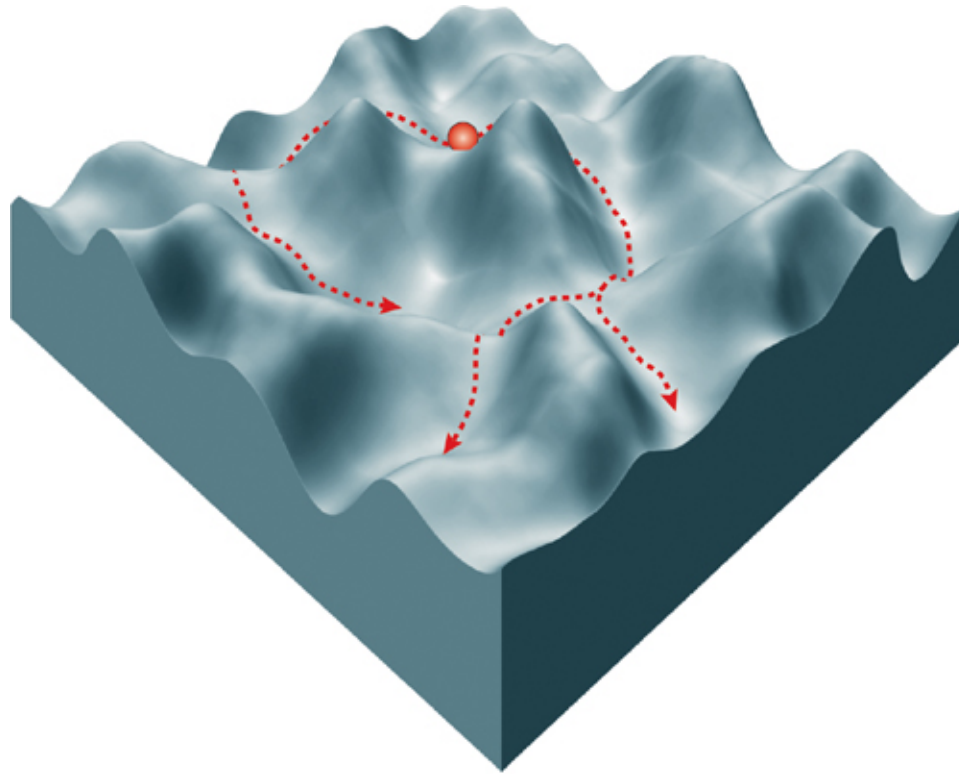
- Произвольная по направлению максимальна, если направление совпадает с градиентом!
- **Градиент — направление наискорейшего роста функции**
- Антиградиент — направление наискорейшего убывания

# Условие оптимальности

- Как понять, является ли точка  $x_0$  экстремумом?
- Обобщение теоремы Ферма: если точка  $x_0$  — экстремум, и в ней существует градиент, то  $\nabla f(x_0) = 0$
- Если функция везде имеет градиент: решаем  $\nabla f(x) = 0$
- Если с градиентом проблемы: не повезло

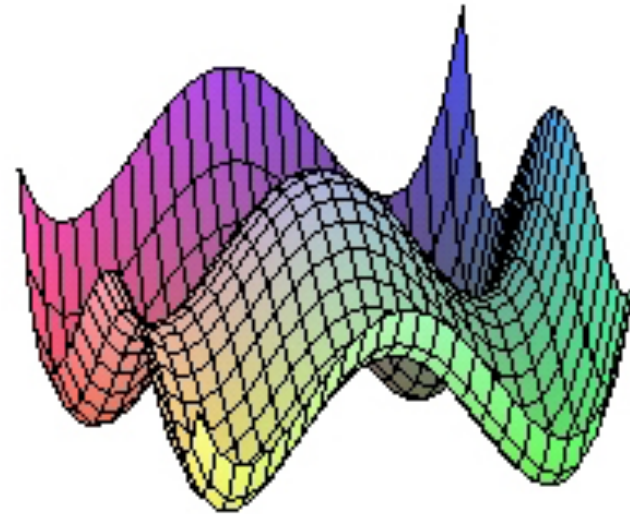
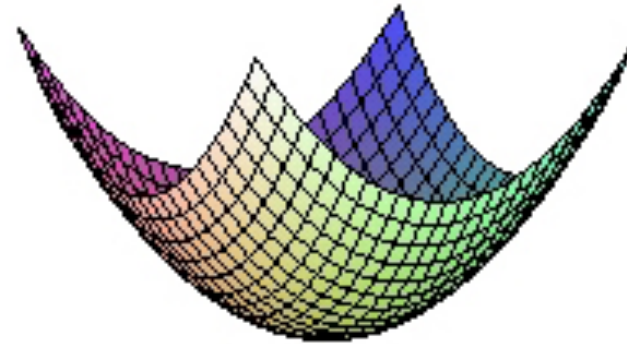
# Экстремумы

- Проблема с локальными экстремумами все еще актуальна



# Выпуклые функции

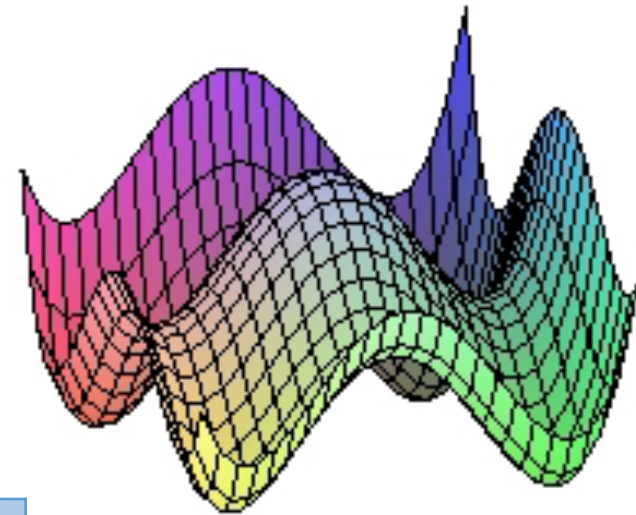
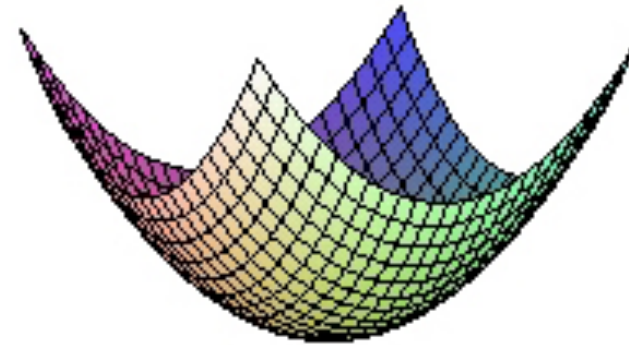
- Функция выпуклая, если ее график лежит ниже отрезка, соединяющего любые две точки



# Выпуклые функции

- Функция выпуклая, если ее график лежит ниже отрезка, соединяющего любые две точки

Выпуклая функция



Невыпуклая функция

# Обучение линейной регрессии

# Задача оптимизации

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w$$

- Градиент существует в любой точке
- Выпуклая функция
- Единственный минимум (не всегда)



# Градиент

$$\nabla Q(w, X) = \left( \frac{\partial Q}{\partial w_1}, \dots, \frac{\partial Q}{\partial w_d} \right)$$

Производные:

$$\frac{\partial Q}{\partial w_j} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i^j (\langle w, x_i \rangle - y_i)$$

# Обучение линейной регрессии

- Векторная запись MSE:

$$Q(w, X) = \frac{1}{\ell} \|Xw - y\|^2$$

- Условие минимума:

$$\nabla Q(w, X) = 0$$

- Что, если попробуем решить эту систему уравнений?

# Обратная матрица

- $A^{-1}$  — обратная к  $A$
- $AA^{-1} = A^{-1}A = I$
- $I$  — единичная матрица
- Только для квадратных матриц
  
- Существует тогда и только тогда, когда  $\det A \neq 0$
- Можно найти с помощью SciPy

# Обучение линейной регрессии

- Условие минимума решается аналитически!

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

- Но обращение матрицы — очень сложная операция
- Градиентный спуск гораздо быстрее

# Резюме

- Линейная регрессия — одна из самых простых моделей в машинном обучении
- Функционал качества: среднеквадратичная ошибка
- Обучение: аналитическая формула или градиентный спуск