Функционал ошибки для классификации

Ошибка классификации

- Доля неправильных ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

- Нотация Айверсона:
  - $[\text{истина}] = 1$
  - $[\text{ложь}] = 0$

Ошибка классификации

$a(x)$	y
-1	-1
$+1$	$+1$
-1	-1
$\mathbf{+1}$	$\mathbf{-1}$
$+1$	$+1$

- Доля неправильных ответов:

$$\frac{1}{5} = 0.2$$

Accuracy

- Доля правильных ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

- На английском: accuracy

Accuracy

- Доля правильных ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

- На английском: accuracy
- ВАЖНО: не переводите это как «точность»!

Оценивание обобщающей способности

Как оценить качество?

- Как алгоритм будет вести себя на новых данных?
- Какая у него будет доля ошибок?
- …или другая метрика качества
- По обучающей выборке нельзя это оценить

Отложенная выборка

- Разбиваем выборку на две части
  - Обучающая выборка
  - Отложенная выборка
- На первой обучаем алгоритм
- На второй измеряем качество

Пропорции разбиения

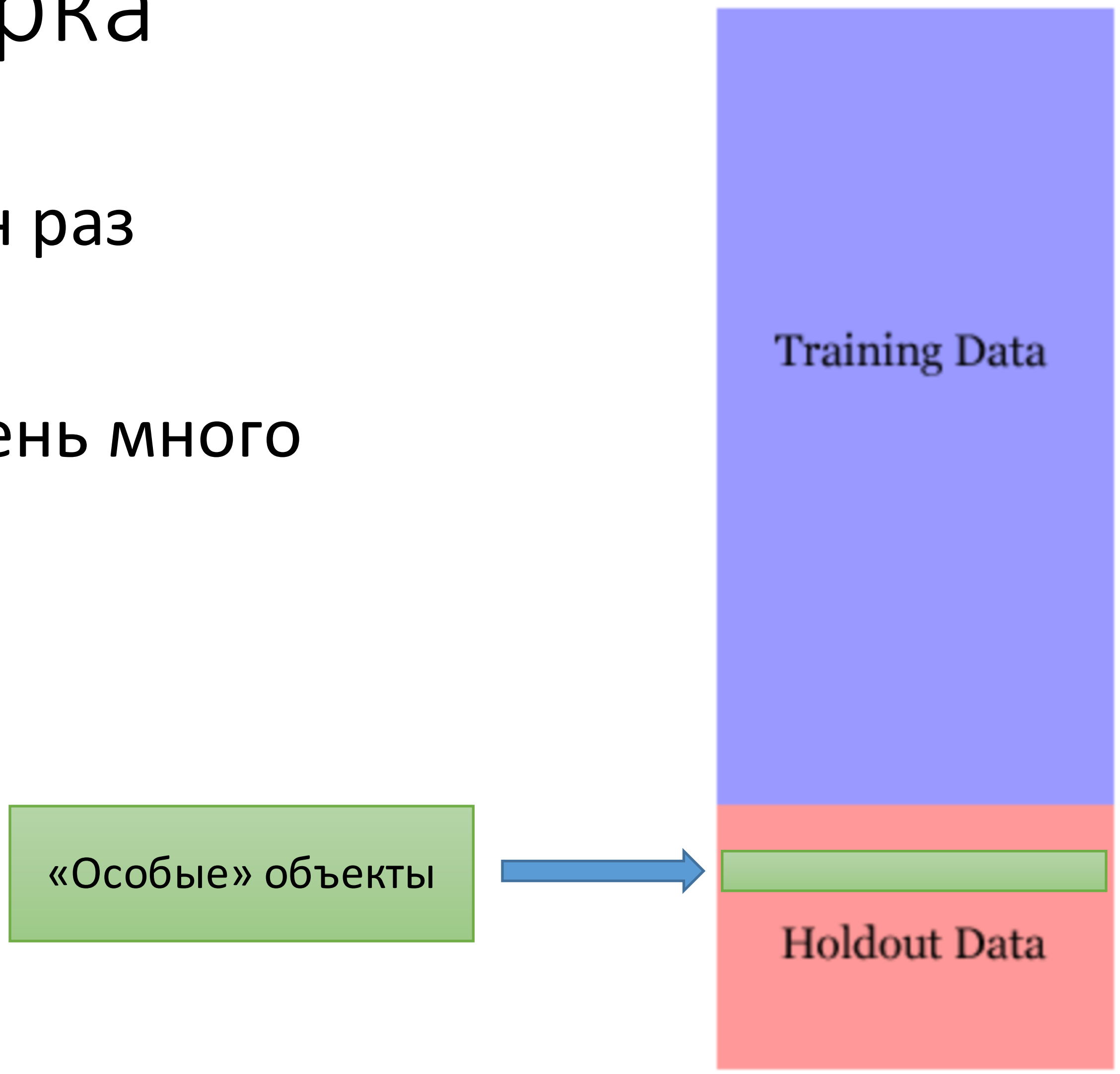
- Маленькая отложенная часть
  - (+) Обучающая выборка репрезентативная
  - (-) Оценка качества ненадежная
- Большая отложенная часть
  - (+) Оценка качества надежная
  - (-) Оценка качества смещенная
- Обычно: 70/30, 80/20, 0.632/0.368

Отложенная выборка

- (+) Обучаем алгоритм один раз
- (-) Зависит от разбиения
- Подходит, если данных очень много

Отложенная выборка

- (+) Обучаем алгоритм один раз
- (-) Зависит от разбиения
- Подходит, если данных очень много



Много отложенных выборок

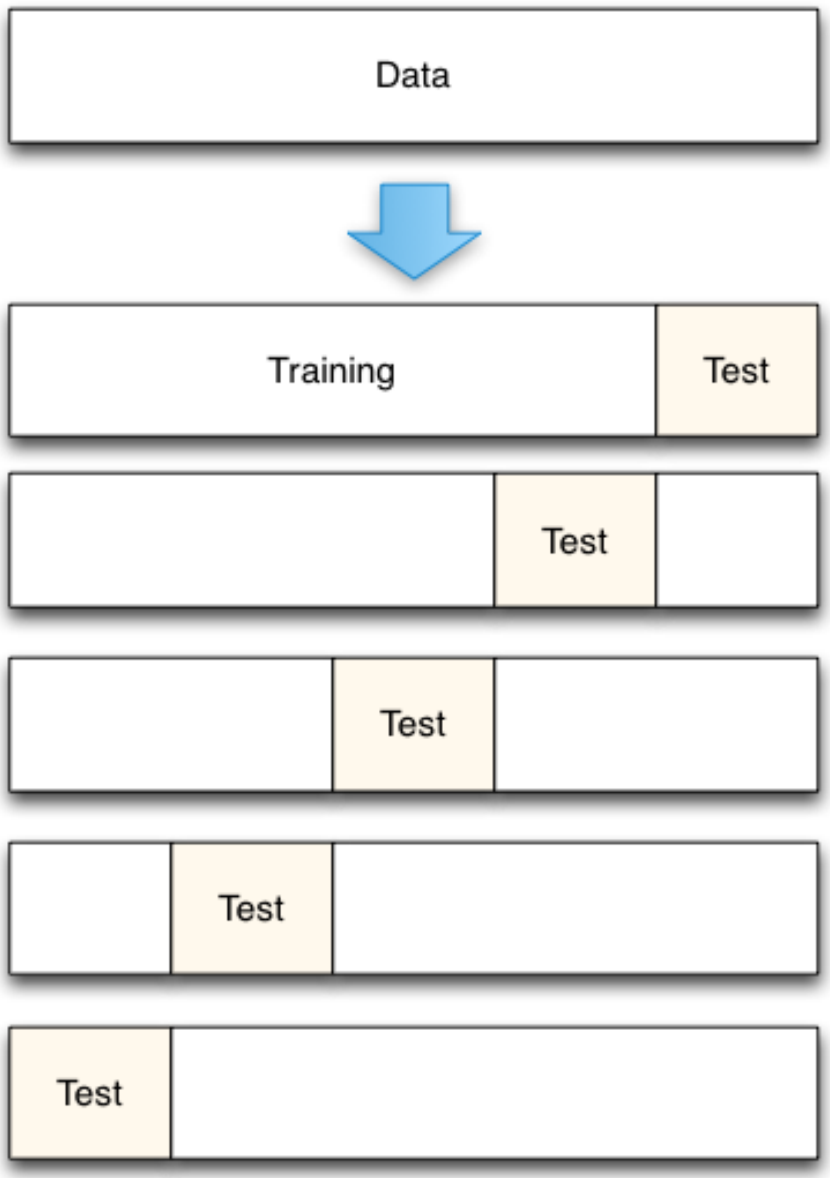
- Улучшение: разбиваем выборку на две части n раз
- Усредняем оценку качества

Много отложенных выборок

- Нет гарантий, что каждый объект побывает в обучении

Кросс-валидация

- Разбиваем выборку на k блоков
- Каждая по очереди выступает как тестовая



Число блоков

- Мало блоков
  - Тестовая выборка всегда большая — (+) надежные оценки
  - Обучение маленькое — (-) смещенные оценки
- Много блоков
  - (-) Ненадежные оценки
  - (+) Несмещенные оценки

Число блоков

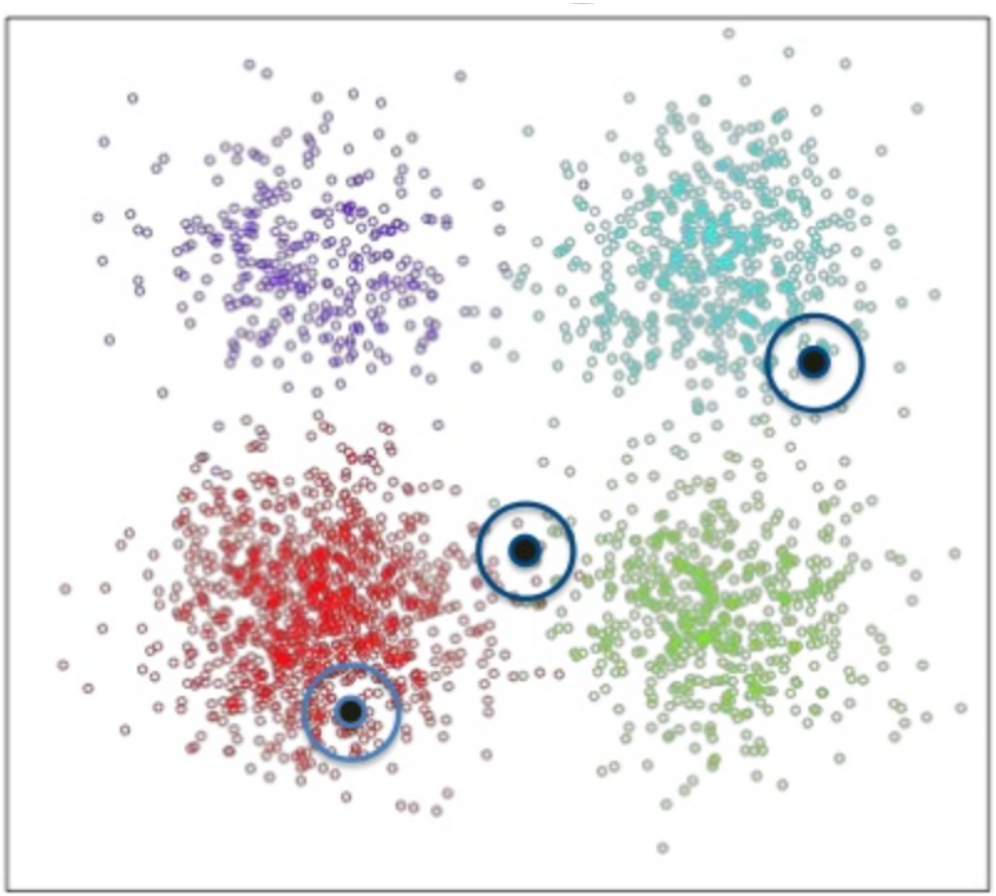
- Обычно: $k = 3, 5, 10$
- Чем больше выборка, тем меньше нужно k
- Чем больше k, тем больше раз надо обучать алгоритм

Совет

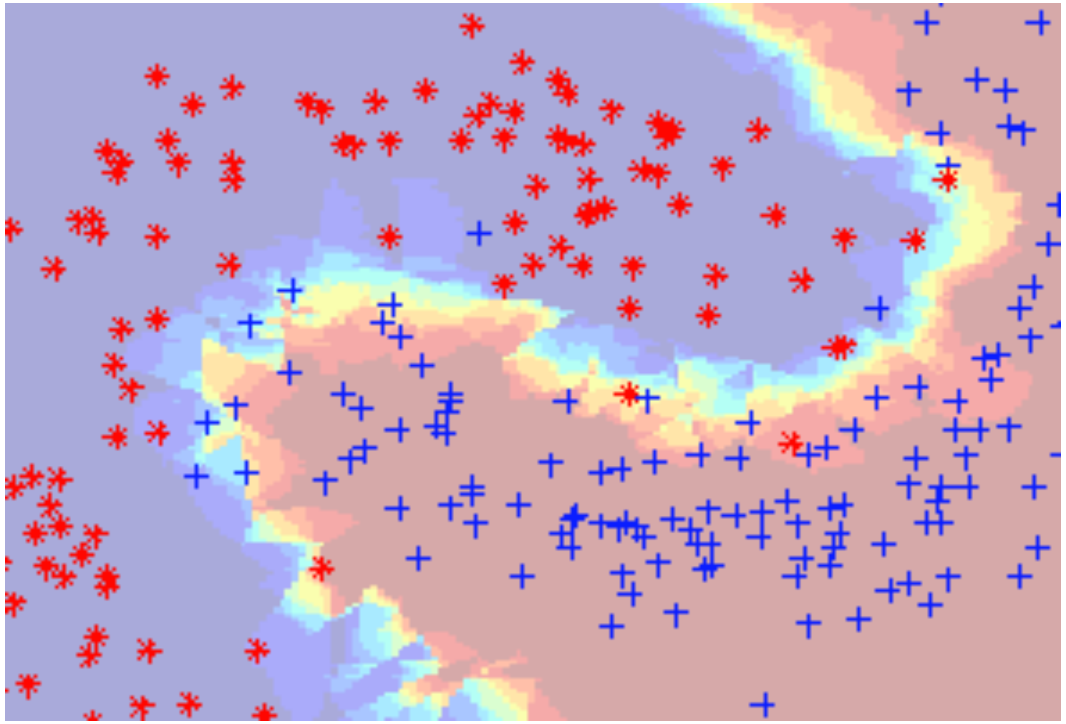
- Перемешивайте выборку!
- Объекты могут быть отсортированы
- При разбиении в обучении могут оказаться только мальчики, в контроле — только девочки

Гипотеза компактности

Гипотеза компактности



Гипотеза компактности



Гипотеза компактности

Гипотеза компактности

- Для классификации: близкие объекты, как правило, лежат в одном классе
- Для регрессии: близким объектам соответствуют близкие ответы

- Что такое «близкие объекты»?

Измерение сходства

- Необходимо ввести расстояние между объектами
- $\rho(x, z)$ — функция расстояния (не обязательно метрика)

- Типичный пример: евклидова метрика

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

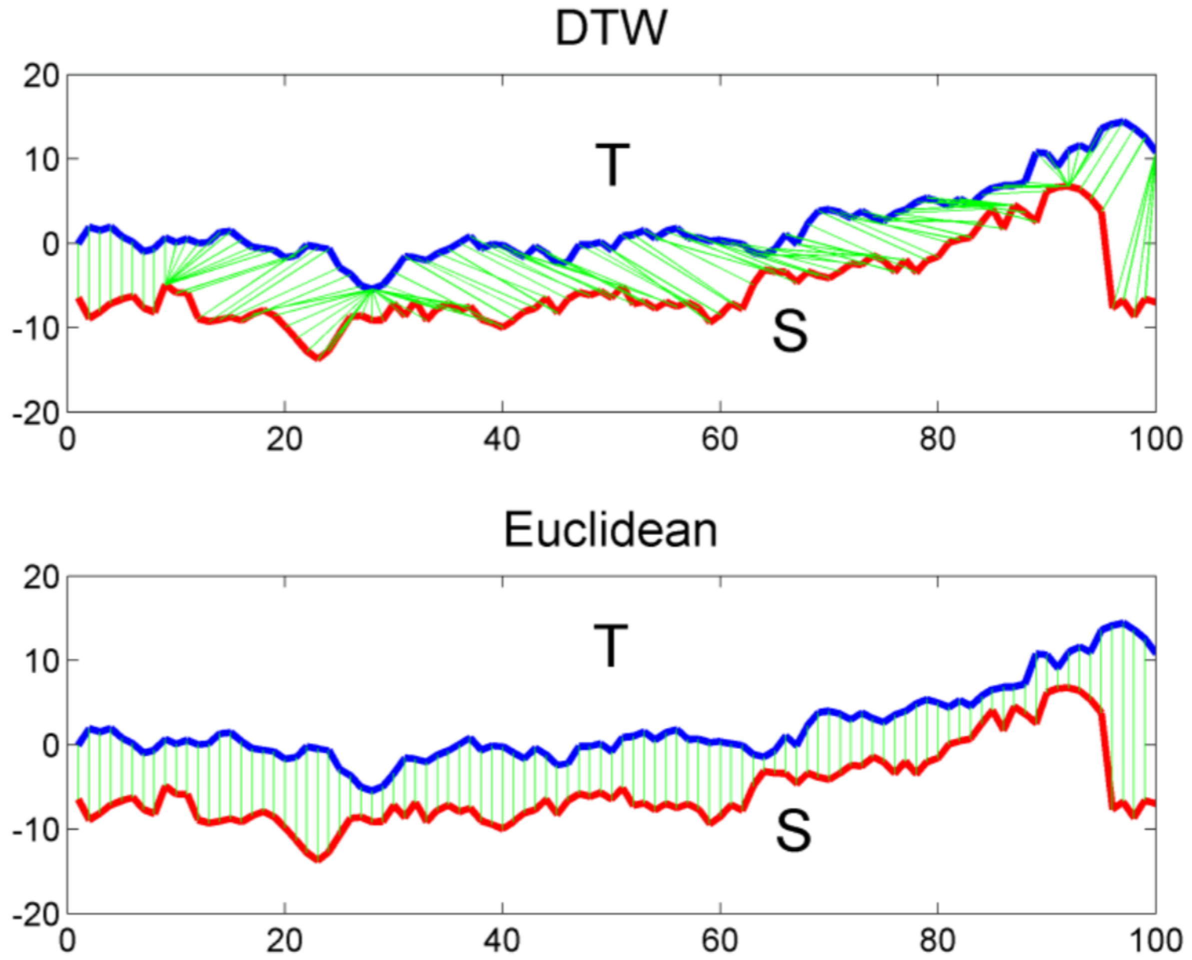
Расстояния на текстах

- Расстояние Левенштейна
- Количество вставок и удалений символов, необходимое для преобразования одной строки в другую

```
CTGGGCTAAAGGTCCCTTAGCC..TTTAGAAAAA.GGGCCATTAGGAAATTGC
CTGGGACTAAA....CCTTAGCCTATTTACAAAAATGGGCCATTAGG...TTGC
```

Расстояния на временных рядах

- Суммарное евклидово расстояние
- Dynamic time warping
- И другие



Метрические методы классификации

Метод k ближайших соседей

- k nearest neighbors (kNN)

- Задача классификации
- Дано: выборка $X = (x_i, y_i)_{i=1}^{\ell}$
- Этап обучения: запоминаем выборку X

Метод k ближайших соседей

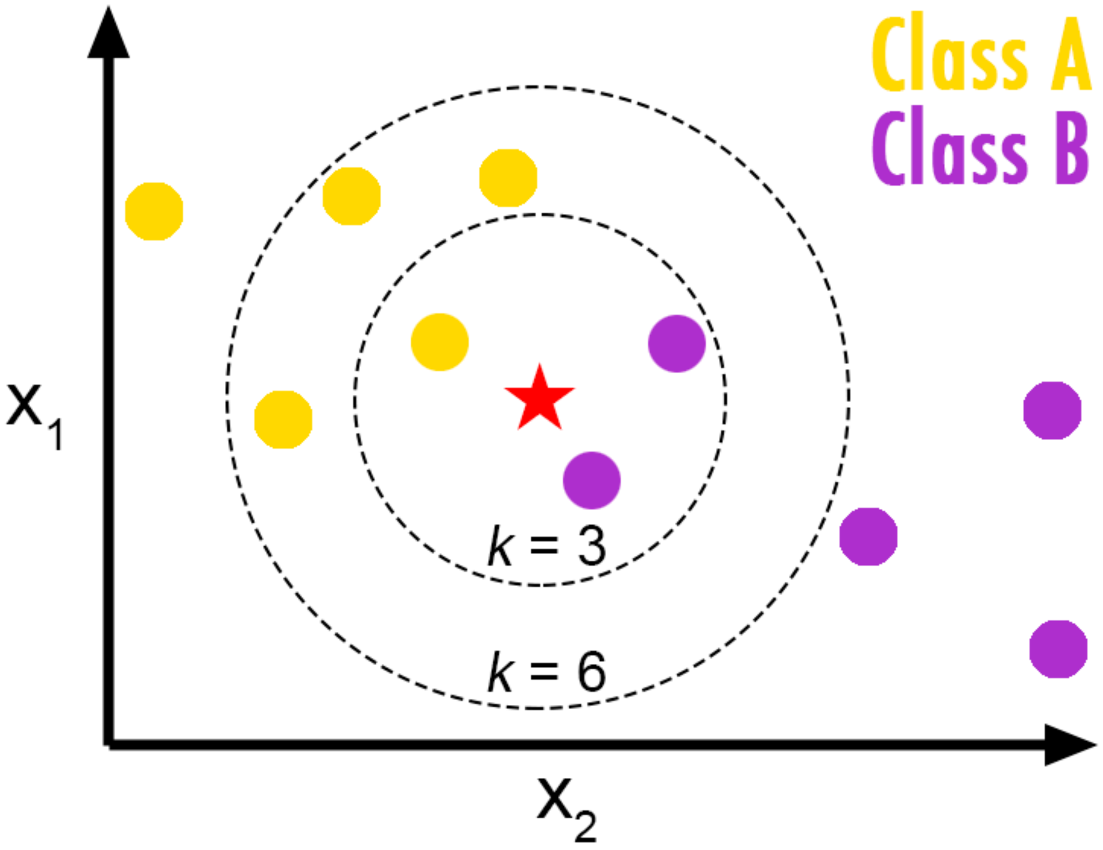
- Новый объект x
- Сортируем объекты обучающей выборки по расстоянию до x:

$$\rho\left(x, x_{(1)}\right) \leq \cdots \leq \rho\left(x, x_{(\ell)}\right)$$

- Выбираем класс, наиболее популярный среди k ближайших соседей:

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

Метод k ближайших соседей

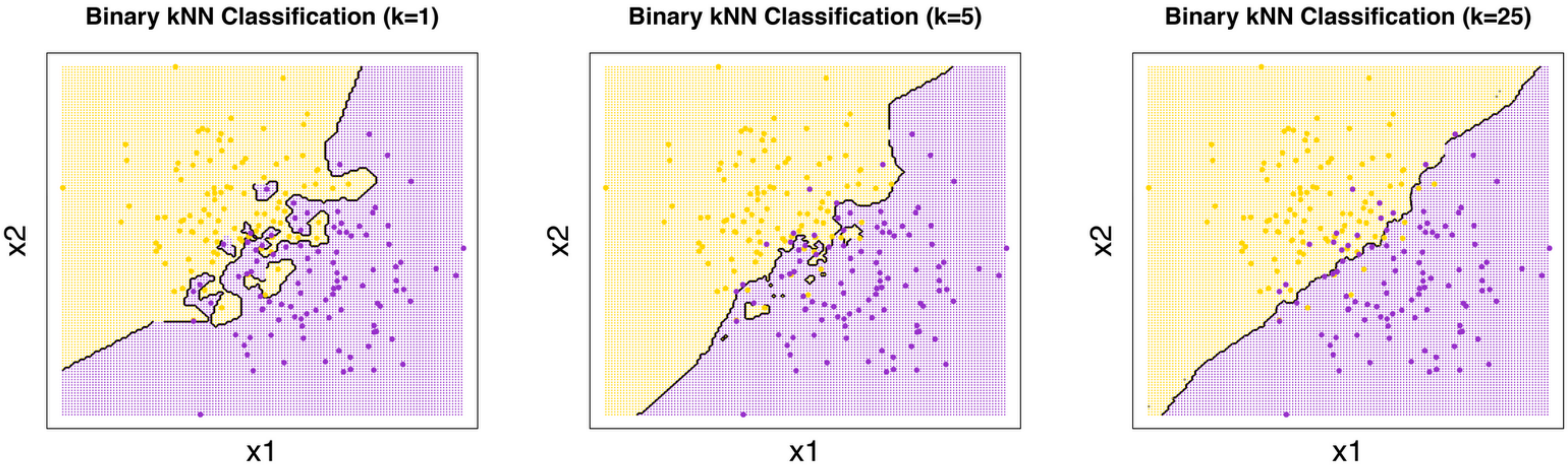


Метод k ближайших соседей

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

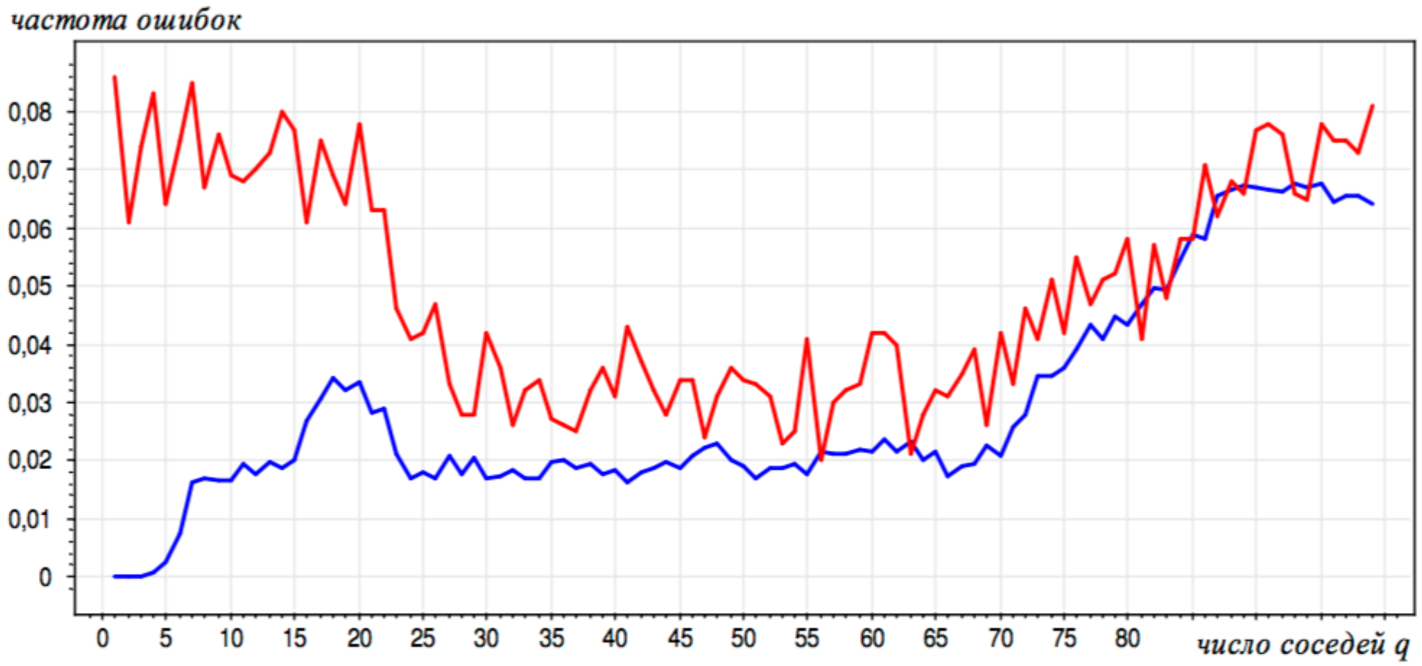
- k — гиперпараметр алгоритма
- Подбирается с помощью holdout-выборки или кросс-валидации
- Чем больше k, тем проще разделяющая поверхность

Выбор числа соседей

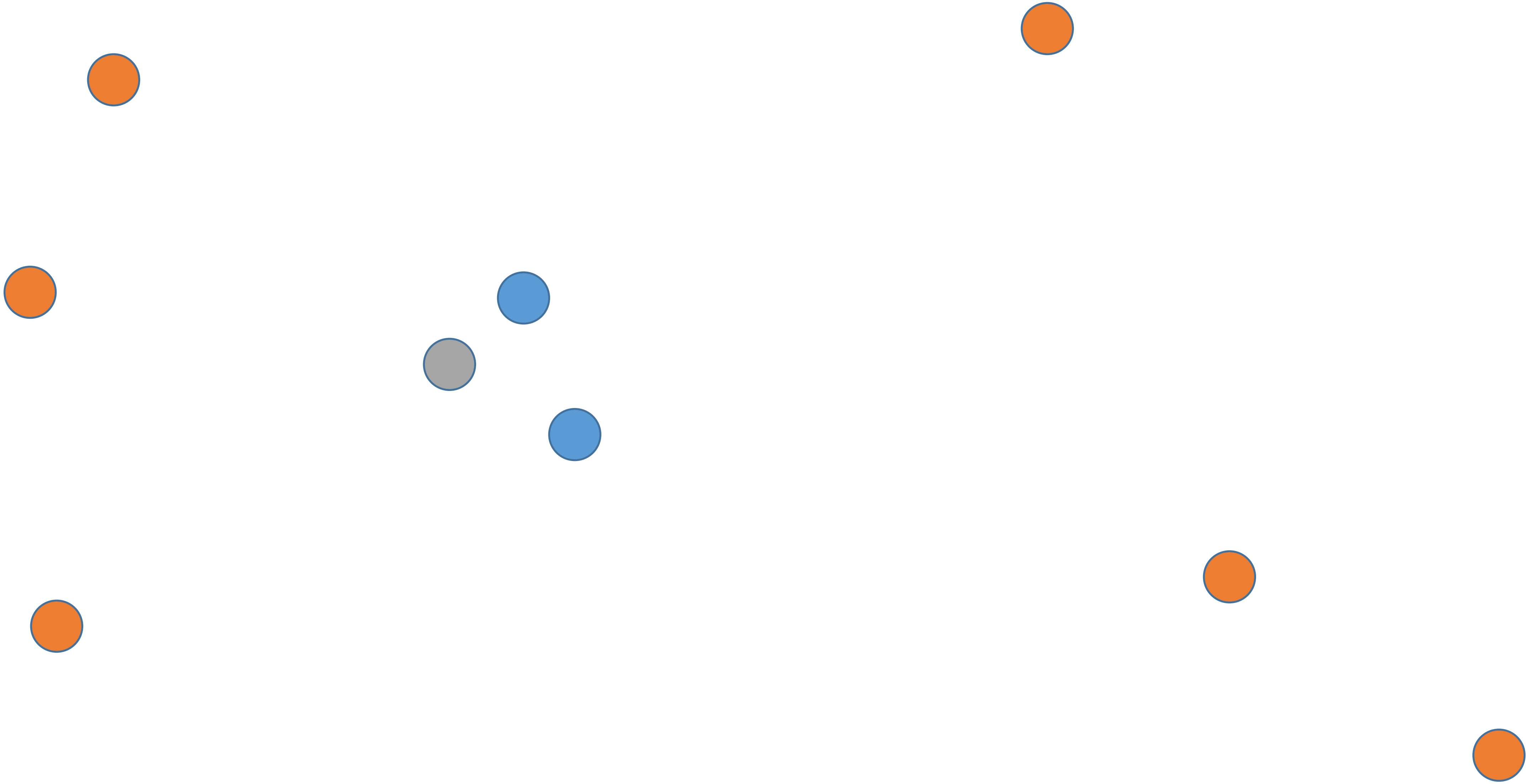


Выбор числа соседей

- Синий — ошибка на обучении
- Красный — ошибка на кросс-валидации



Проблема kNN



Проблема kNN

- Никак не учитываются расстояния до k ближайших соседей
- Более близкие соседи должны быть важнее

kNN с весами

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

Варианты:

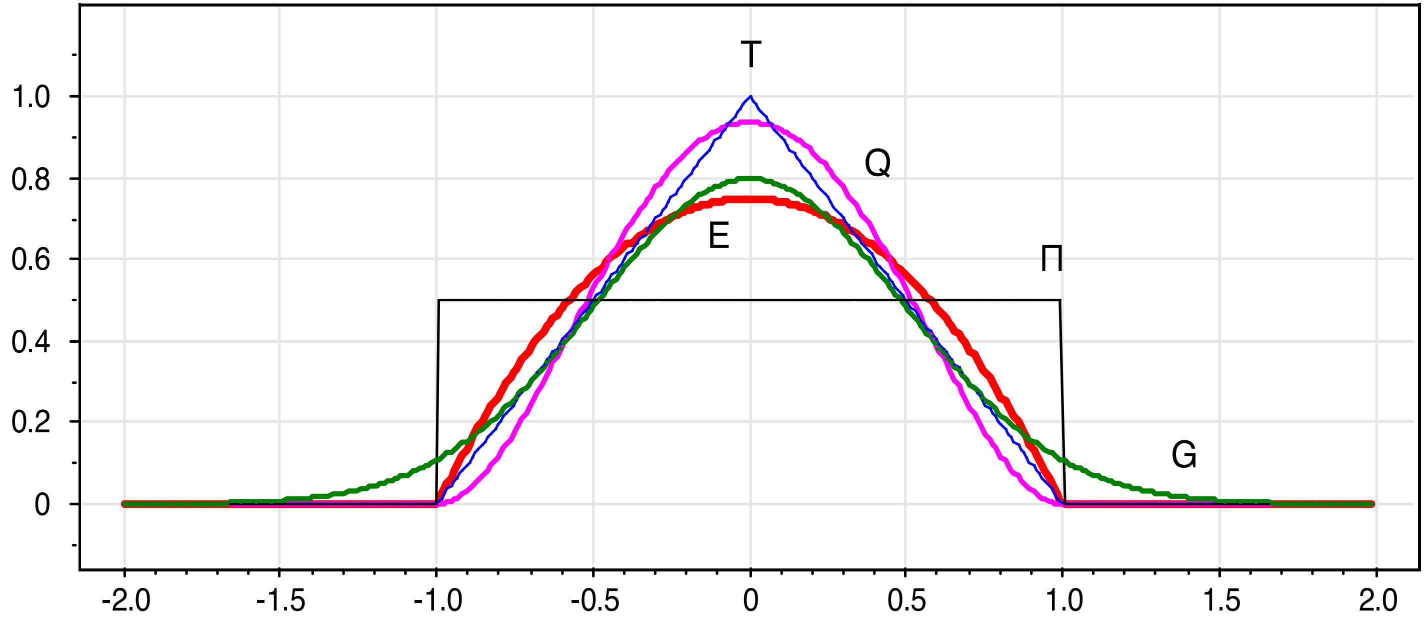
- $w_i = \frac{k+1-i}{k}$
- $w_i = q^i$
- Не учитывают сами расстояния

kNN с весами

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

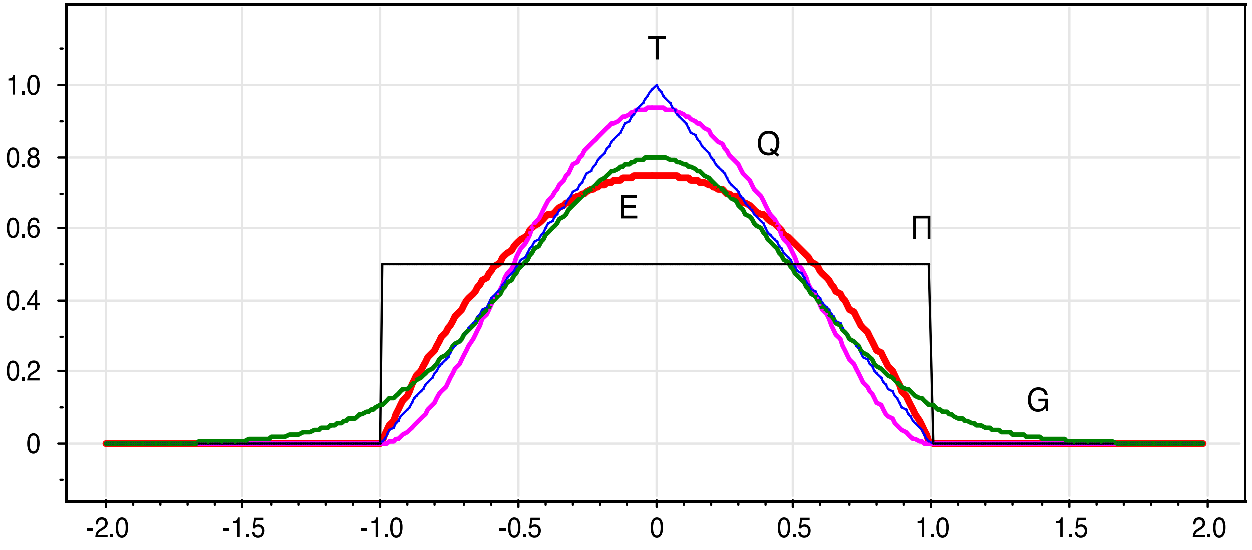
Парзеновское окно:

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

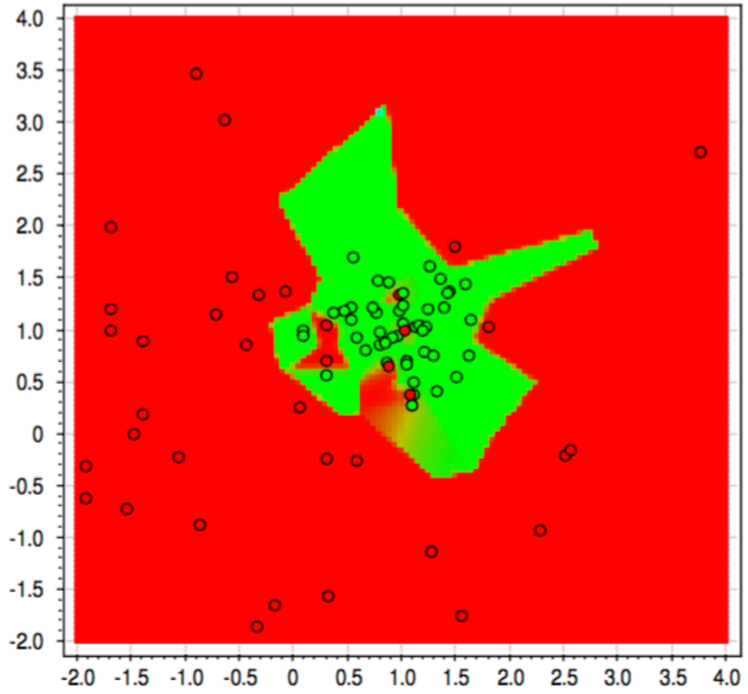


Ядра

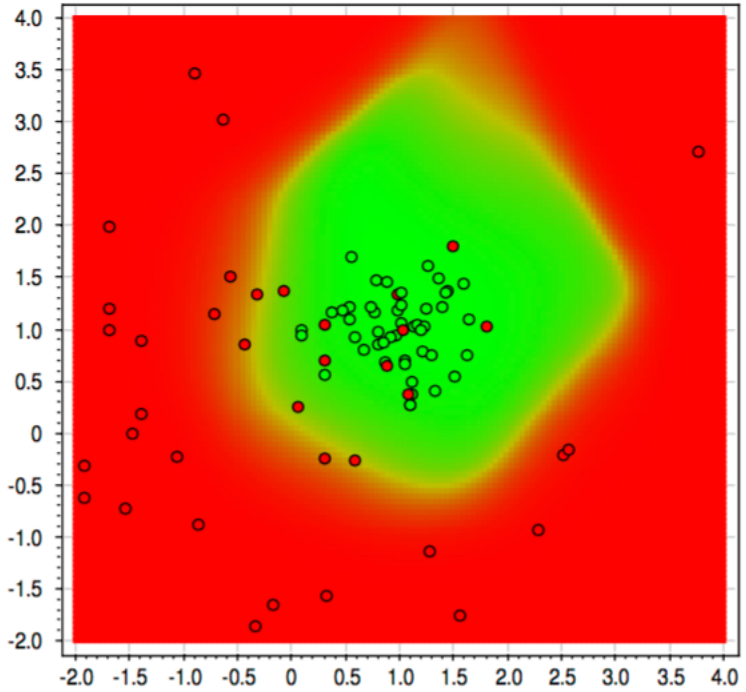
- Гауссовское ядро: $K(z) = (2\pi)^{-0.5} \exp\left(-\frac{1}{2}z\right)$
- И много других



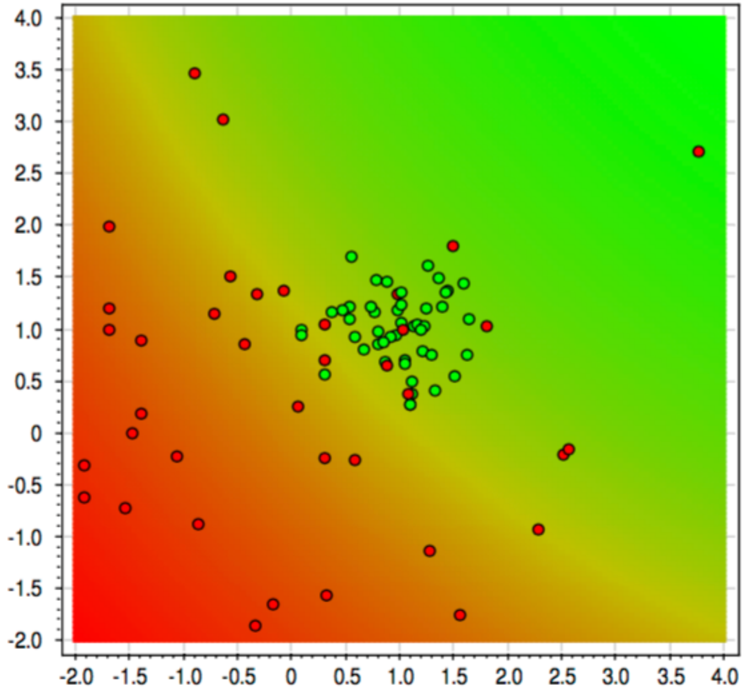
Ядра



$h = 0.05$



$h = 0.5$



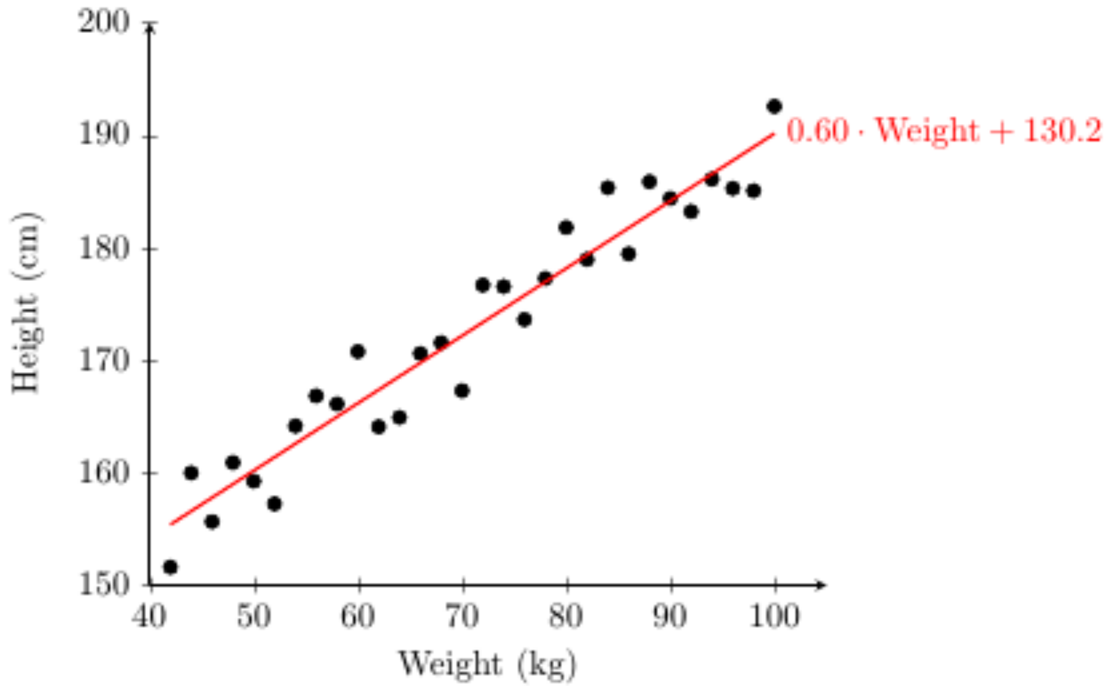
$h = 5$

Особенности kNN

- Обучение как таковое отсутствует — нужно лишь запомнить обучающую выборку
- Для применения модели необходимо вычислить расстояния от нового объекта до всех обучающих объектов
- Применение требует ℓd операций
- Существуют специальные методы для поиска 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

$\boldsymbol{a(x)}$	$\boldsymbol{y}$	**отклонение**
11	10	1
9	10	-1
20	10	10
1	10	-9

Функционал ошибки

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

$\boldsymbol{a(x)}$	$\boldsymbol{y}$	$\boldsymbol{|a(x) - y|}$
11	10	1
9	10	1
20	10	10
1	10	9

Функционал ошибки

- Возведём в квадрат: $(a(x) - y)^2$

$\boldsymbol{a(x)}$	$\boldsymbol{y}$	$\boldsymbol{(a(x) - y)^2}$
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

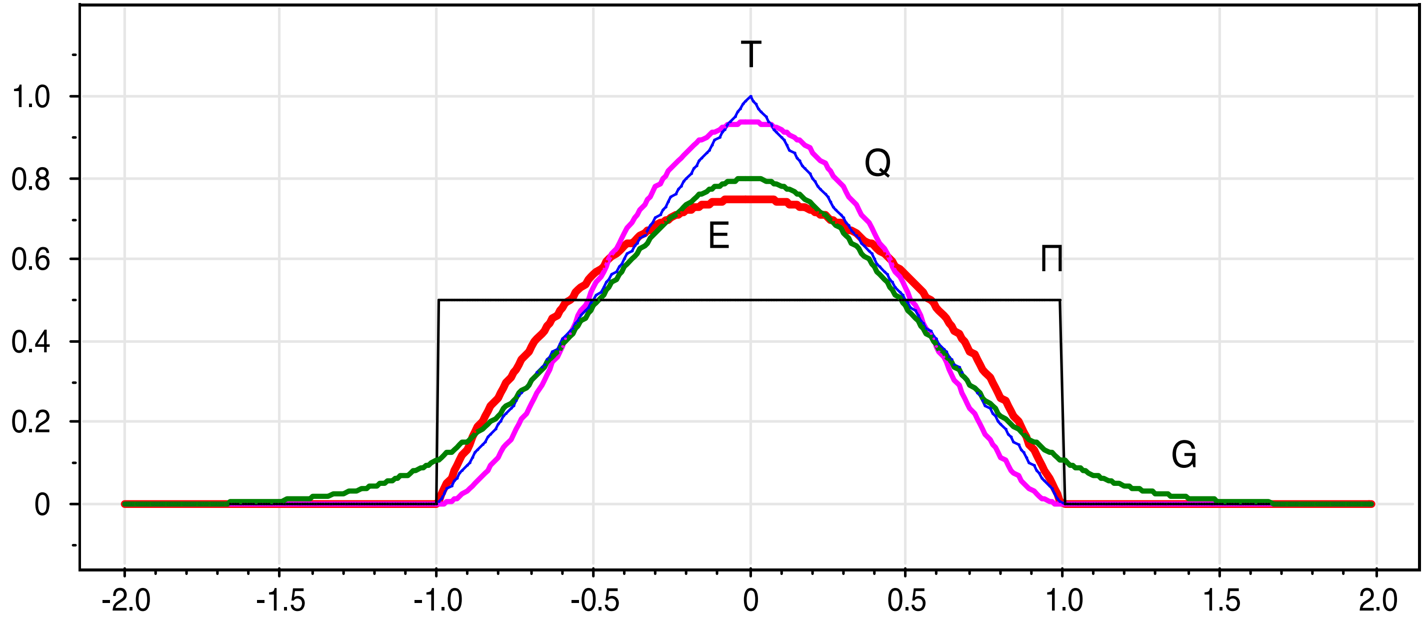
Метрические методы регрессии

kNN с весами

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

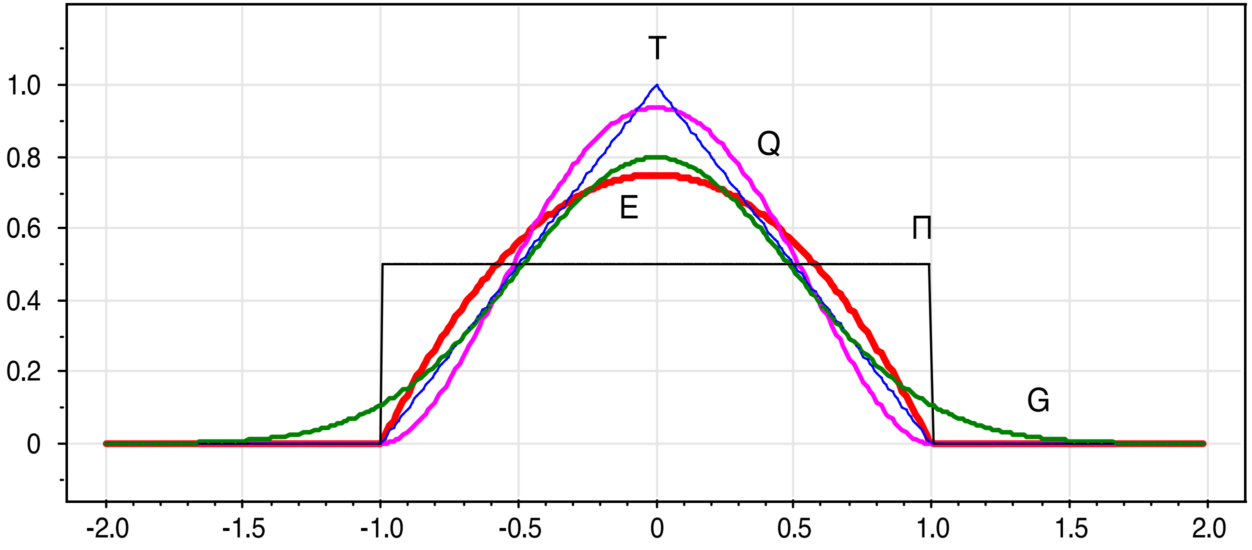
Парзеновское окно:

- $w_i = K\left(\frac{\rho(x, x_{(i)})}{h}\right)$
- 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:

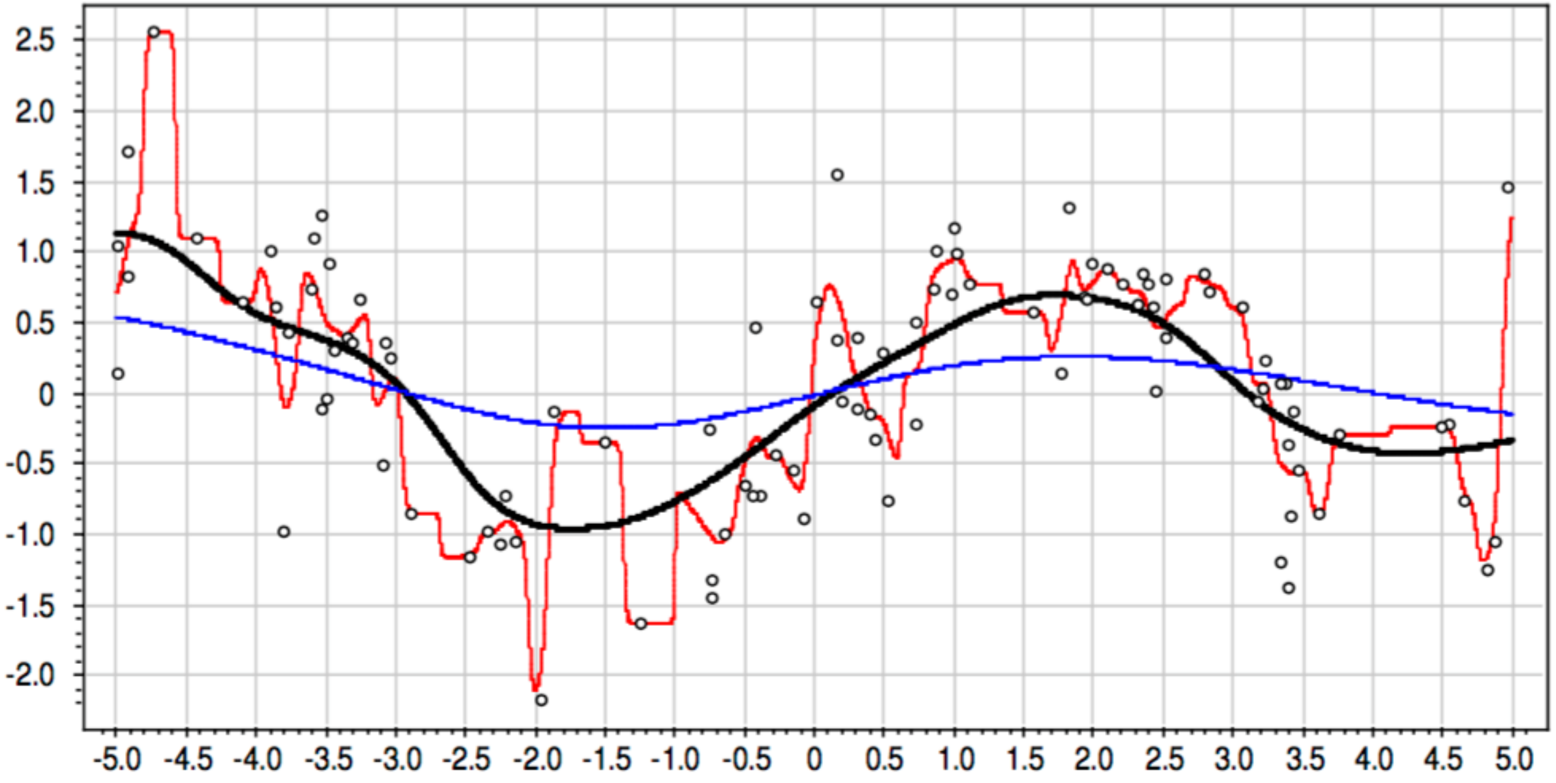
$$a(x) = \arg\max_{y \in \mathbb{Y}} \sum_{i=1}^{k} w_i [y_{(i)} = y]$$

- Регрессия:

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

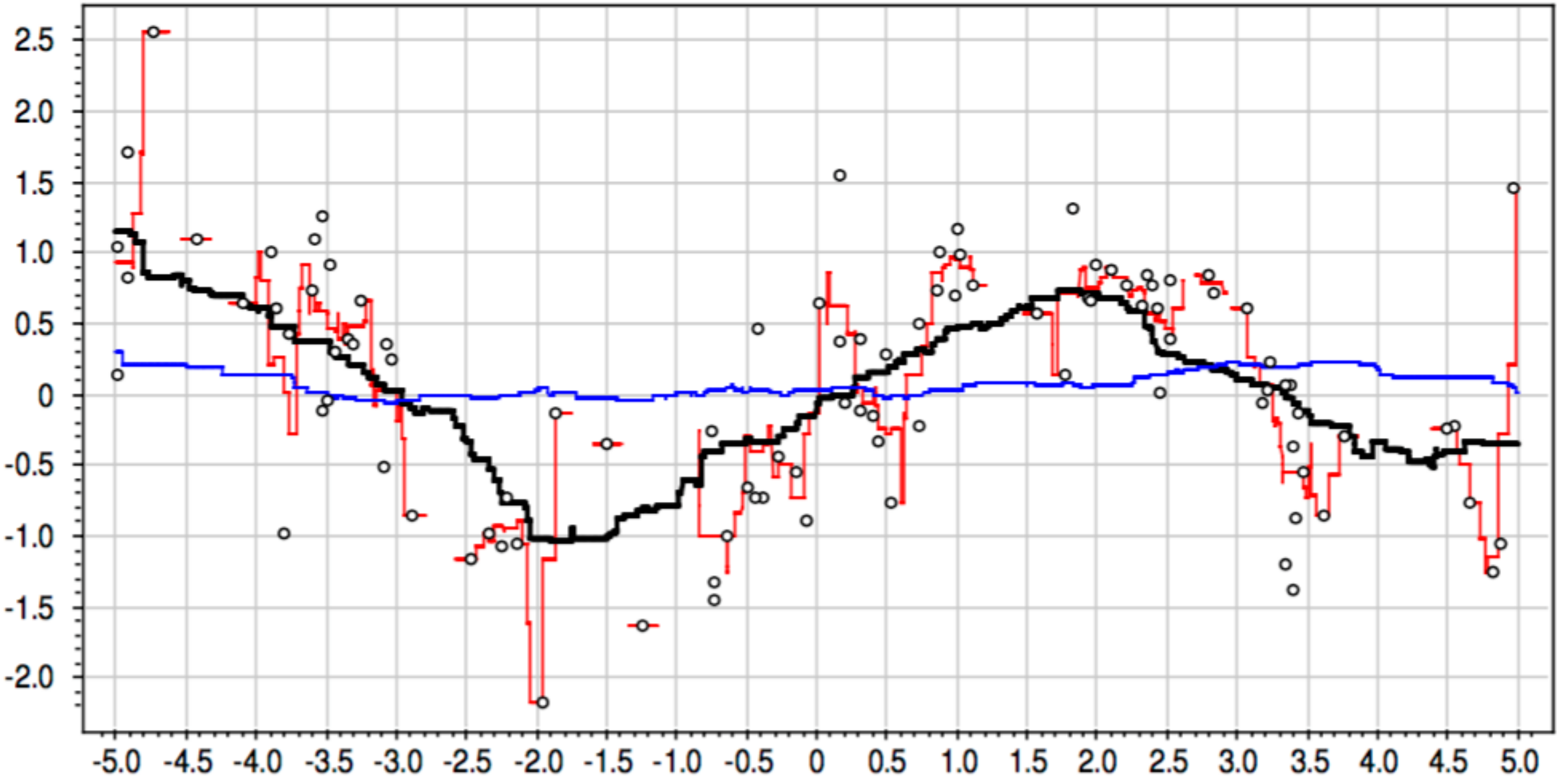
kNN для регрессии

- 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

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

- Более общий вариант — метрика Минковского:

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

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:
  - $A = \{\text{комедия}, \text{триллер}, \text{США}\}$
  - $B = \{\text{триллер}, \text{ужасы}, \text{Великобритания}\}$
- $\rho(A, B) = 1 - \frac{1}{5} = 0.8$

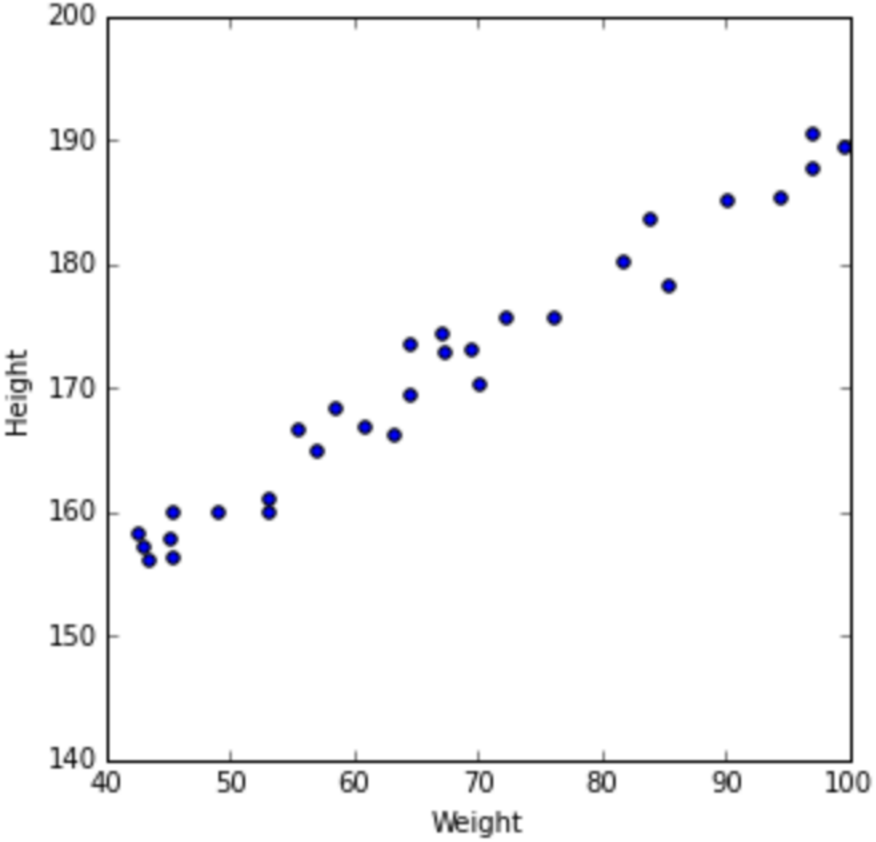
- Пример 2:
  - $A = \{\text{комедия}, \text{США}\}$
  - $B = \{\text{комедия}, \text{США}\}$
- $\rho(A, B) = 1 - \frac{2}{2} = 0$

Резюме по kNN

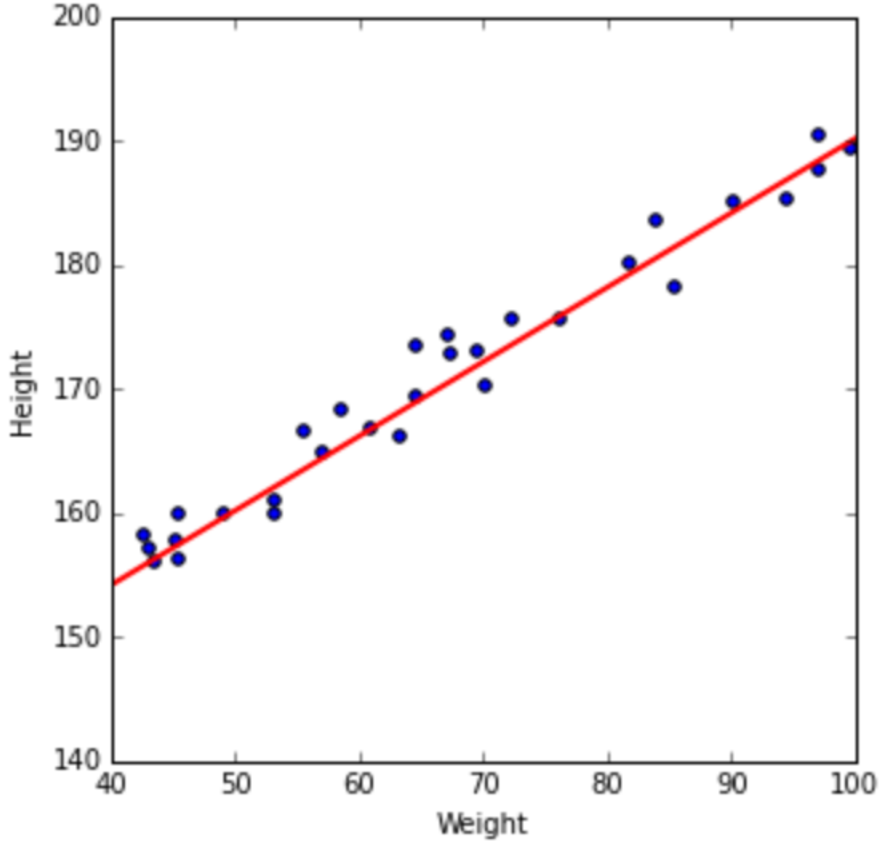
- Метрические методы — одни из самых интуитивных в машинном обучении
- Простая процедура обучения
- Гиперпараметры:
    - функция расстояния
    - число соседей
    - ядро
    - ширина окна

Линейная регрессия

Одномерная выборка



Одномерная выборка



Парная регрессия

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

Линейная регрессия

- Взвешенная сумма признаков:

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

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

Линейная регрессия

- Взвешенная сумма признаков:

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

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

Единичный признак

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

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

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

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

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

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

- $a(x) = w_1 x^1 + \dots + w_d x^d$
- Как применить модель к целой выборке?

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

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

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

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:
  - $A\big(B(x_1 + x_2)\big) = A(Bx_1) + A(Bx_2)$
  - $A\big(B(\alpha x)\big) = \alpha A(Bx)$
- 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

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

Пример

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

Пример

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

Пример

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

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

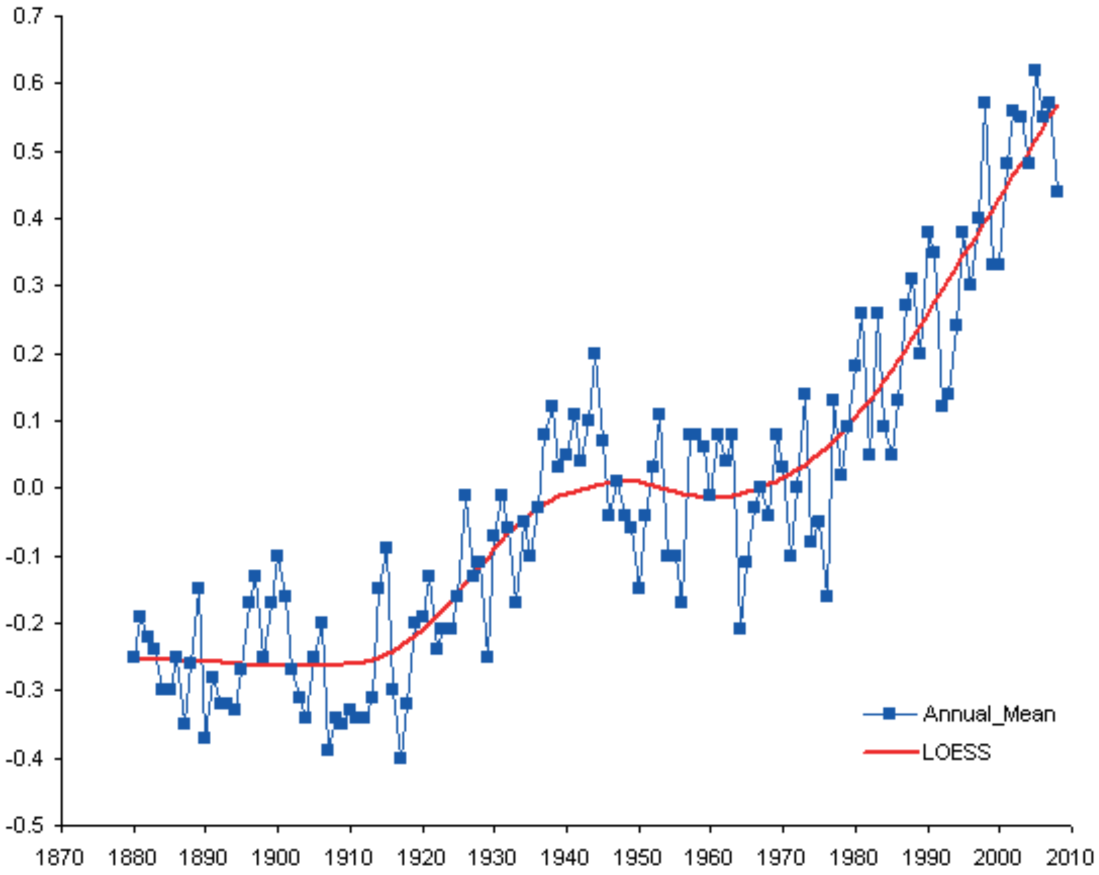
- Скорость роста между 1990 и 2000:

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

- Дискретная величина

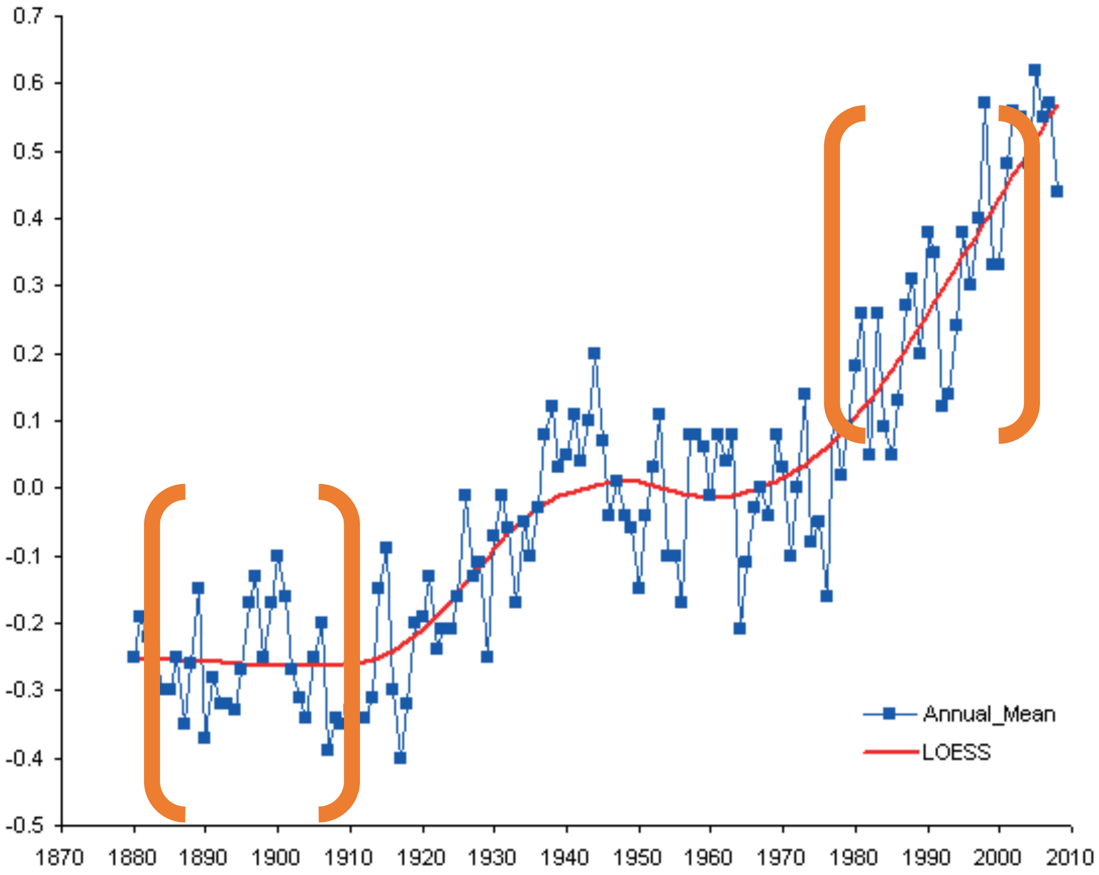
Скорость роста

- Отклонение температуры от нормы (непрерывная величина):



Скорость роста

• Отклонение температуры от нормы:



Высокая скорость

Низкая 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!
$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

Скорость роста

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

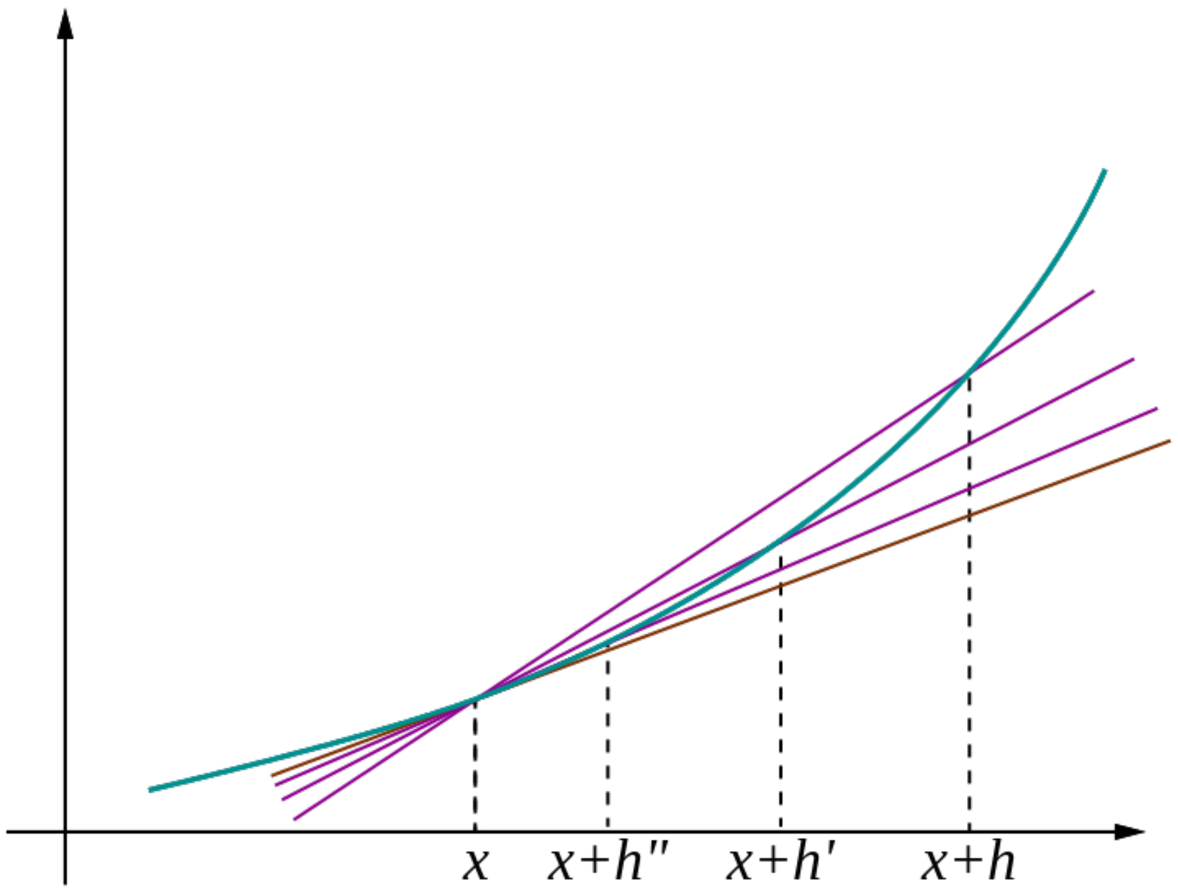
$$\frac{f(x) - f(x_0)}{x - x_0}$$

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

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

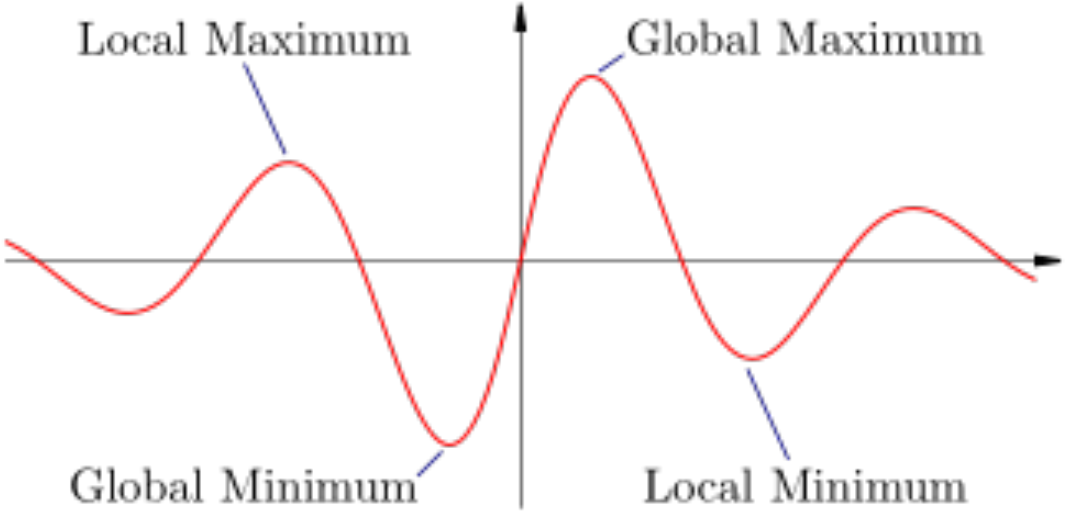
Производная

Производная



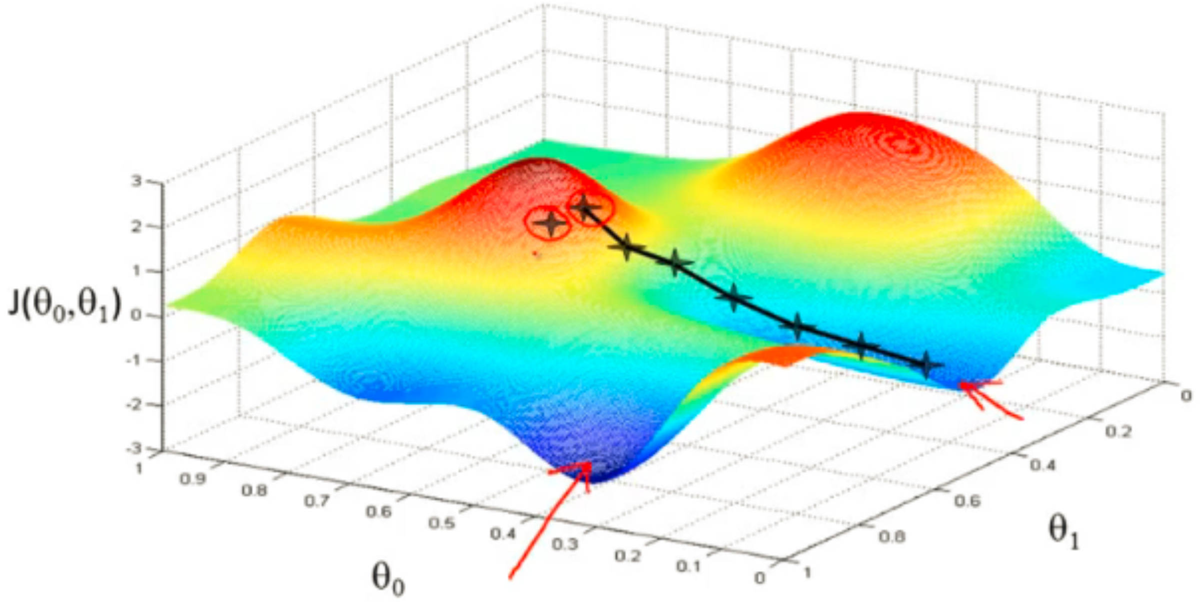
Экстремумы

- 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

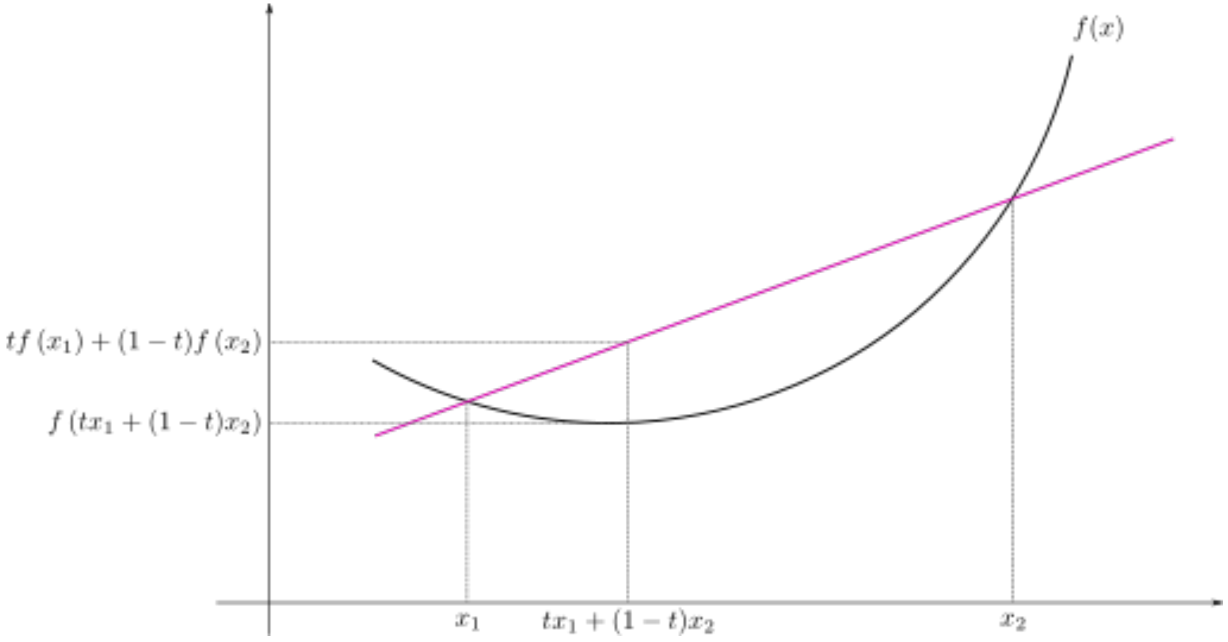
- Как понять, является ли точка x_0 экстремумом?
- Теорема Ферма: если точка x_0 — экстремум, и в ней существует производная, то $f'(x_0) = 0$

- 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

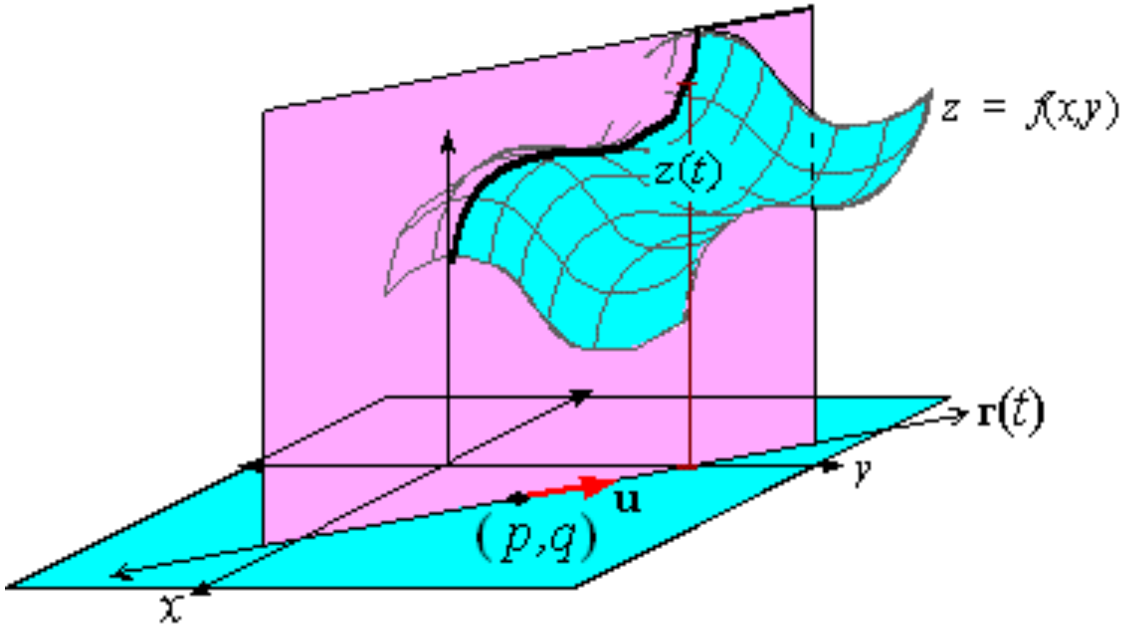
- Функционал качества линейной регрессии:

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

- Как искать ее минимум?

Производная по направлению

- С какой скоростью растет функция в конкретном направлении?



Производная по направлению

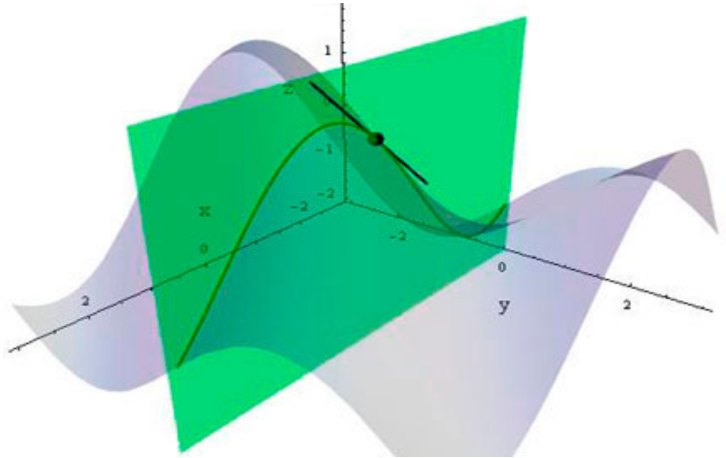
- Направление: v, причем $\|v\| = 1$
- Производная:

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

Частные производные

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

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



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 растет?

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

Угол между градиентом и направлением

- Связь производной по направлению и градиента:

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

Градиент

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

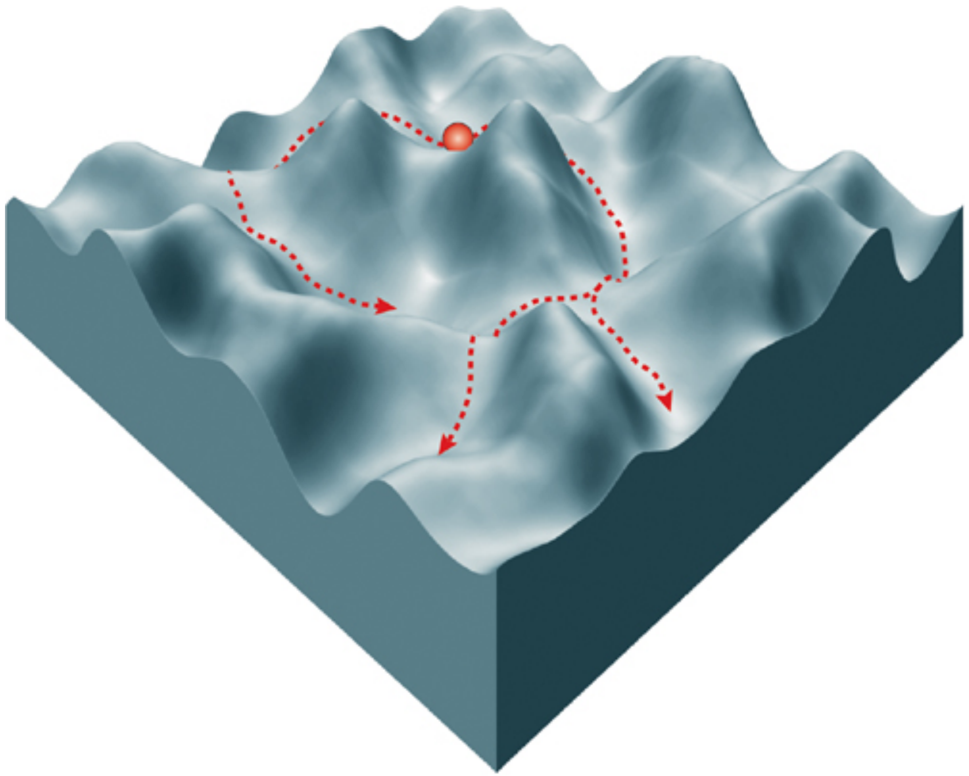
Условие оптимальности

- 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

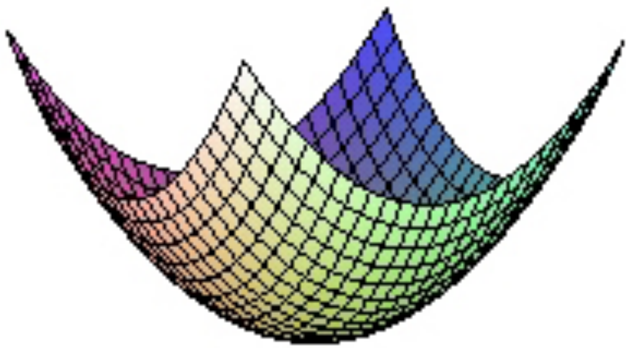
- Проблема с локальными экстремумами все еще актуальна

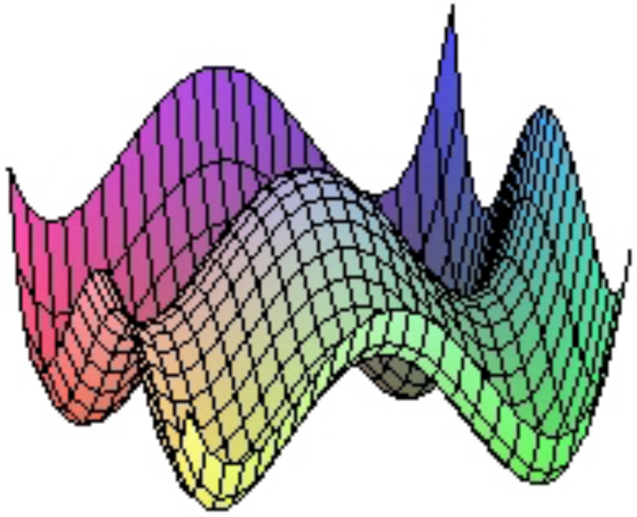


Nature Reviews | Molecular Cell Biology

Выпуклые функции

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

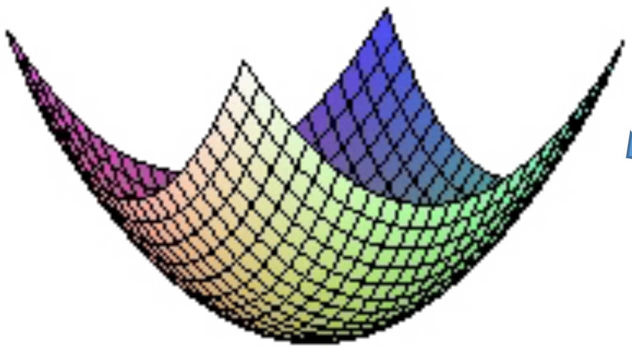


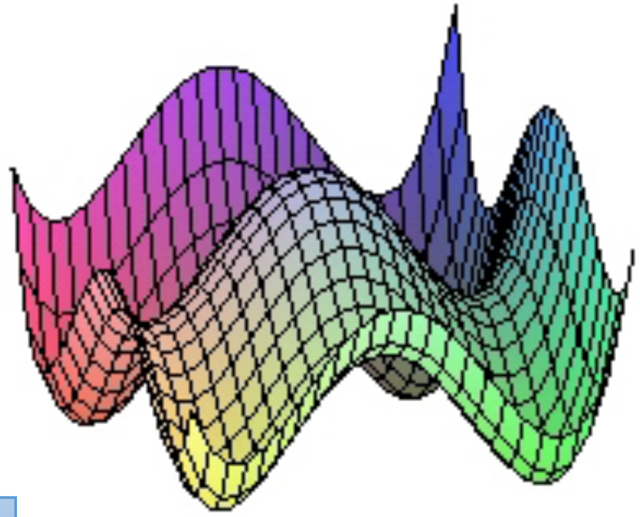


Выпуклые функции

Выпуклая функция

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





Невыпуклая функция

Обучение линейной регрессии

Задача оптимизации

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

- 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

Градиент

- Градиент — вектор из частных производных:

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

- И зачем нам этот вектор?
- У градиента есть очень важное свойство!

Градиент

- Зафиксируем точку x_0
- В каком направлении функция быстрее всего растет?

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

Угол между градиентом и направлением

- Связь производной по направлению и градиента:

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

Градиент

- Производная 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

Методы оптимизации

Поиск минимума

- Функционал качества линейной регрессии:

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

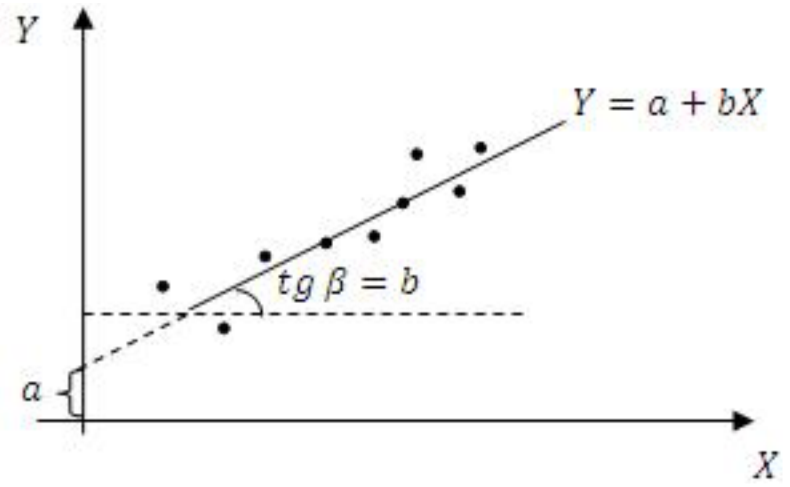
- Как искать минимум?

Поиск минимума

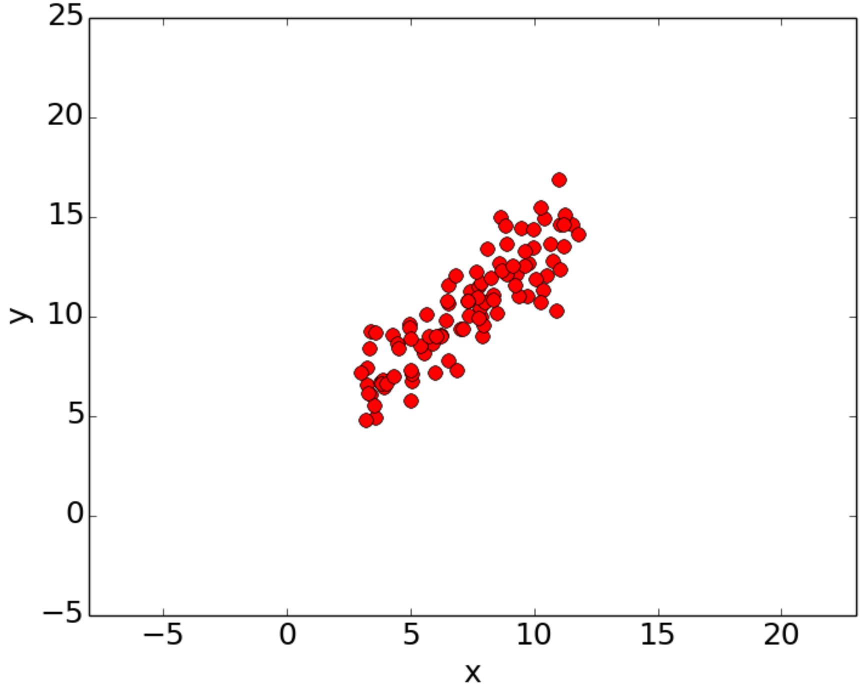
- Можно решать уравнение: $\nabla Q(w) = 0$
- А если уравнение сложное, и аналитически решить нельзя?
- Нужна численная оптимизация

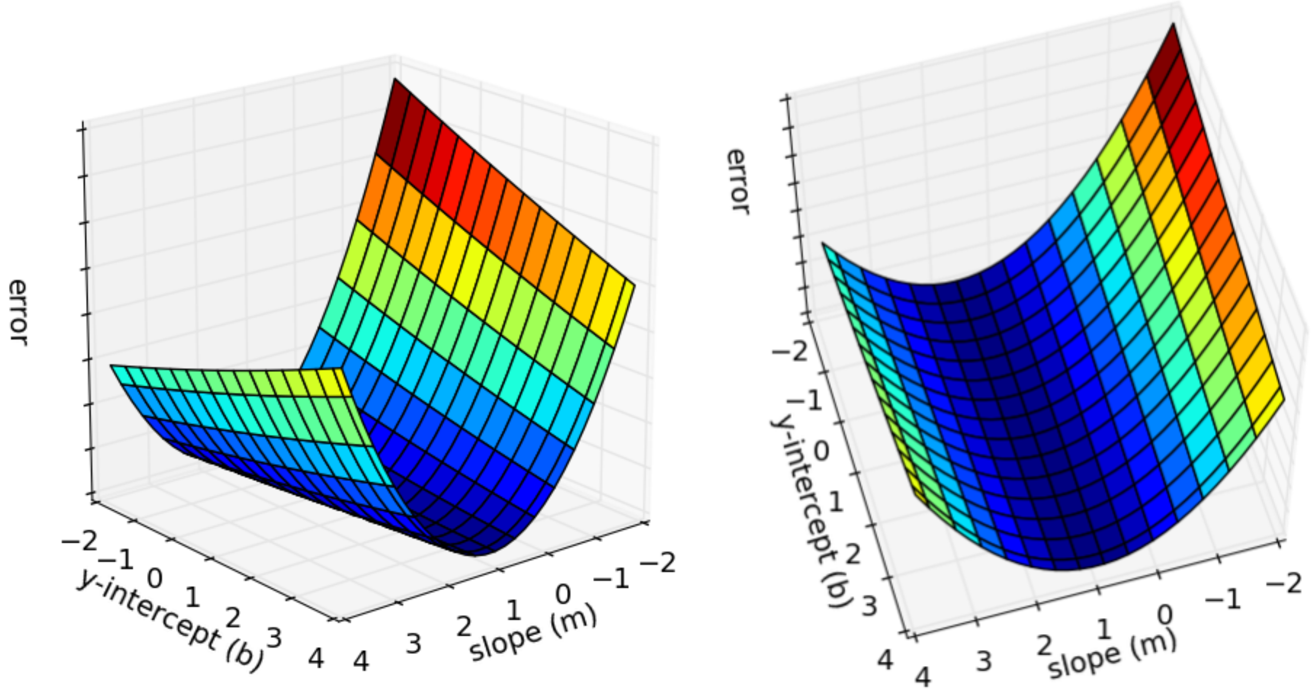
Парная регрессия

- Простейший случай: один признак
- Модель: $a(x) = w_1 x + w_0$
- Два параметра: w_1 и w_0
- Функционал: $Q(w_0, w_1) = \sum_{i=1}^{\ell}(w_1 x_i + w_0 - y_i)^2$



Парная регрессия



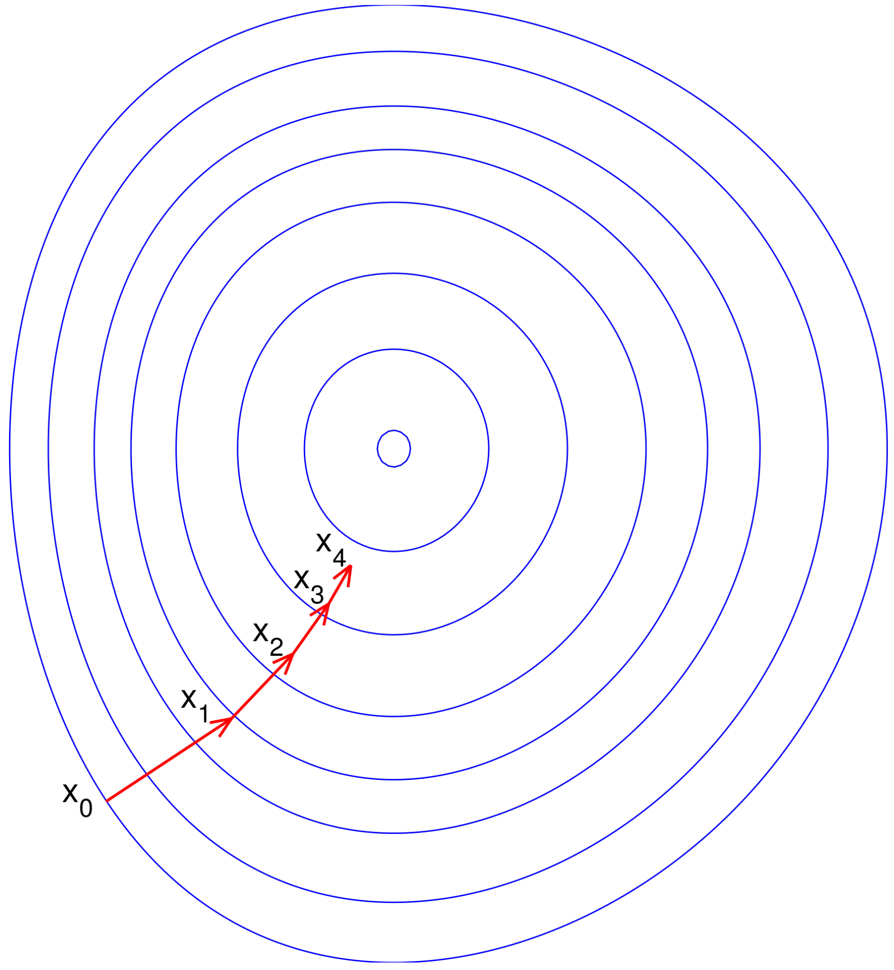


Выборка

Функционал качества

Градиентный спуск

- Допустим, мы выбрали начальное приближение $w^0 = (w_0^0, w_1^0)$
- Как его улучшить?
- Шагнуть в сторону наискорейшего убывания
- То есть в сторону антиградиента!



Градиентный спуск

- Повторять до сходимости:

$$w^t = w^{t-1} - \eta \, \nabla Q(w^{t-1})$$

Новая точка

Размер шага

Градиент в предыдущей точке

Градиентный спуск

- Повторять до сходимости:

$$w^t = w^{t-1} - \eta \, \nabla Q(w^{t-1})$$

- Сходимость: $\|w^t - w^{t-1}\| < \varepsilon$

Градиент для парной регрессии

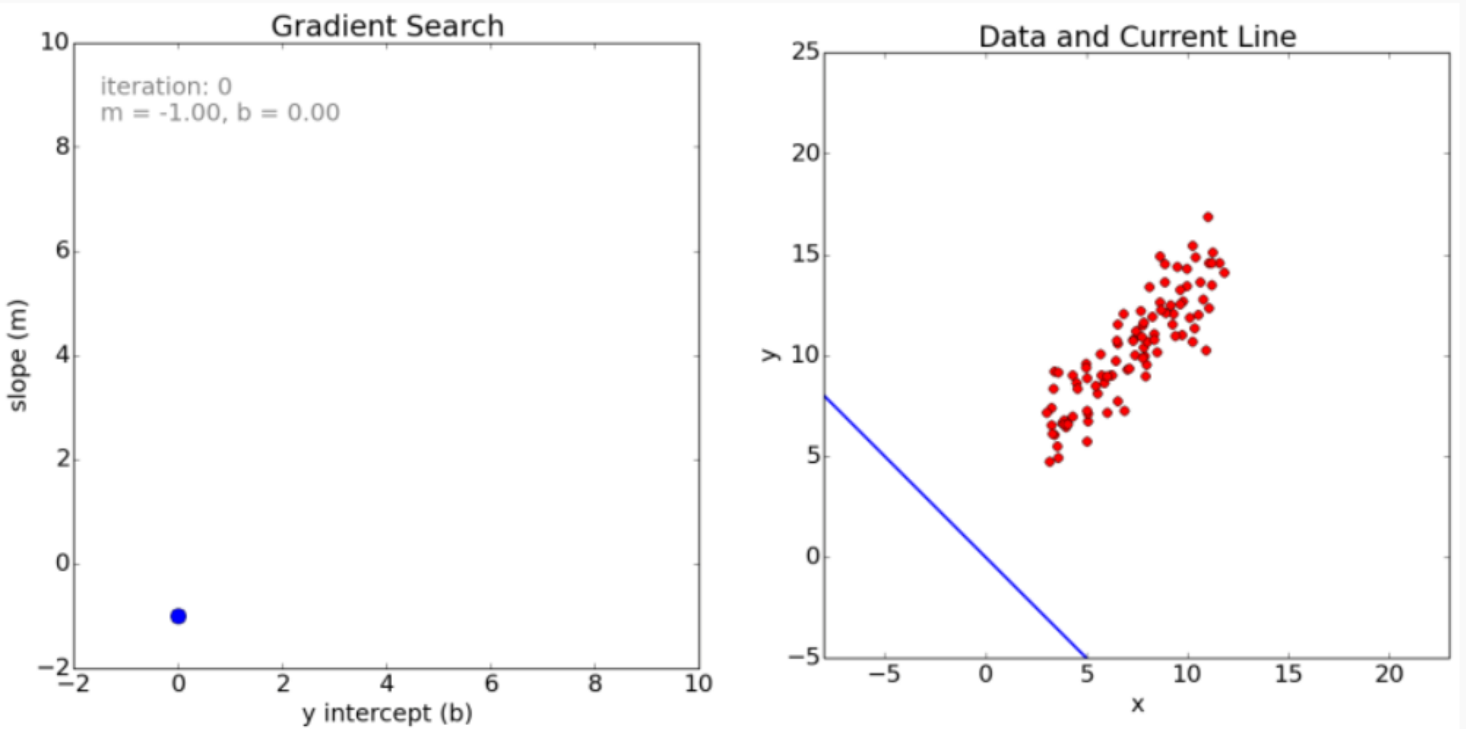
$$Q(w_0, w_1) = \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$$

- Частные производные:

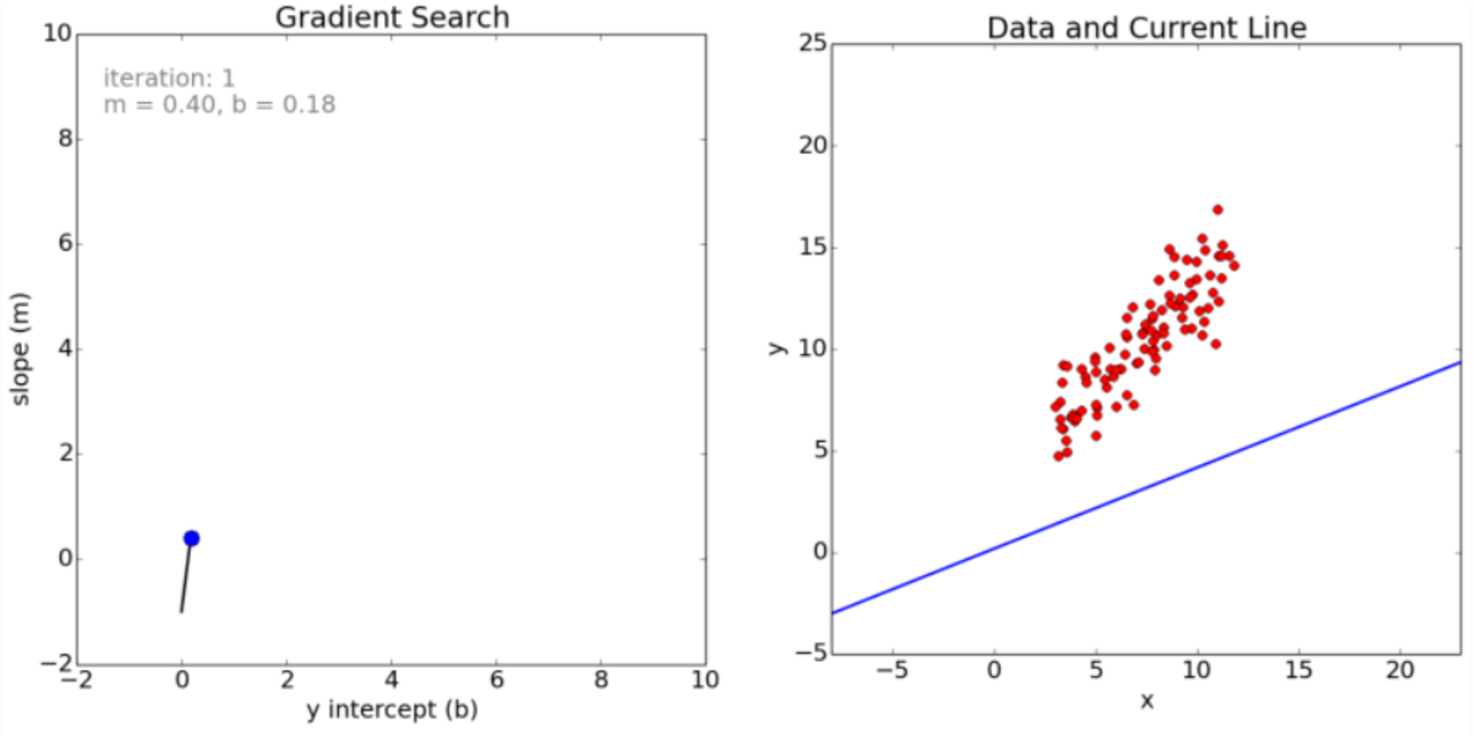
$$\frac{\partial Q}{\partial w_1} = 2 \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i) x_i$$

$$\frac{\partial Q}{\partial w_0} = 2 \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)$$

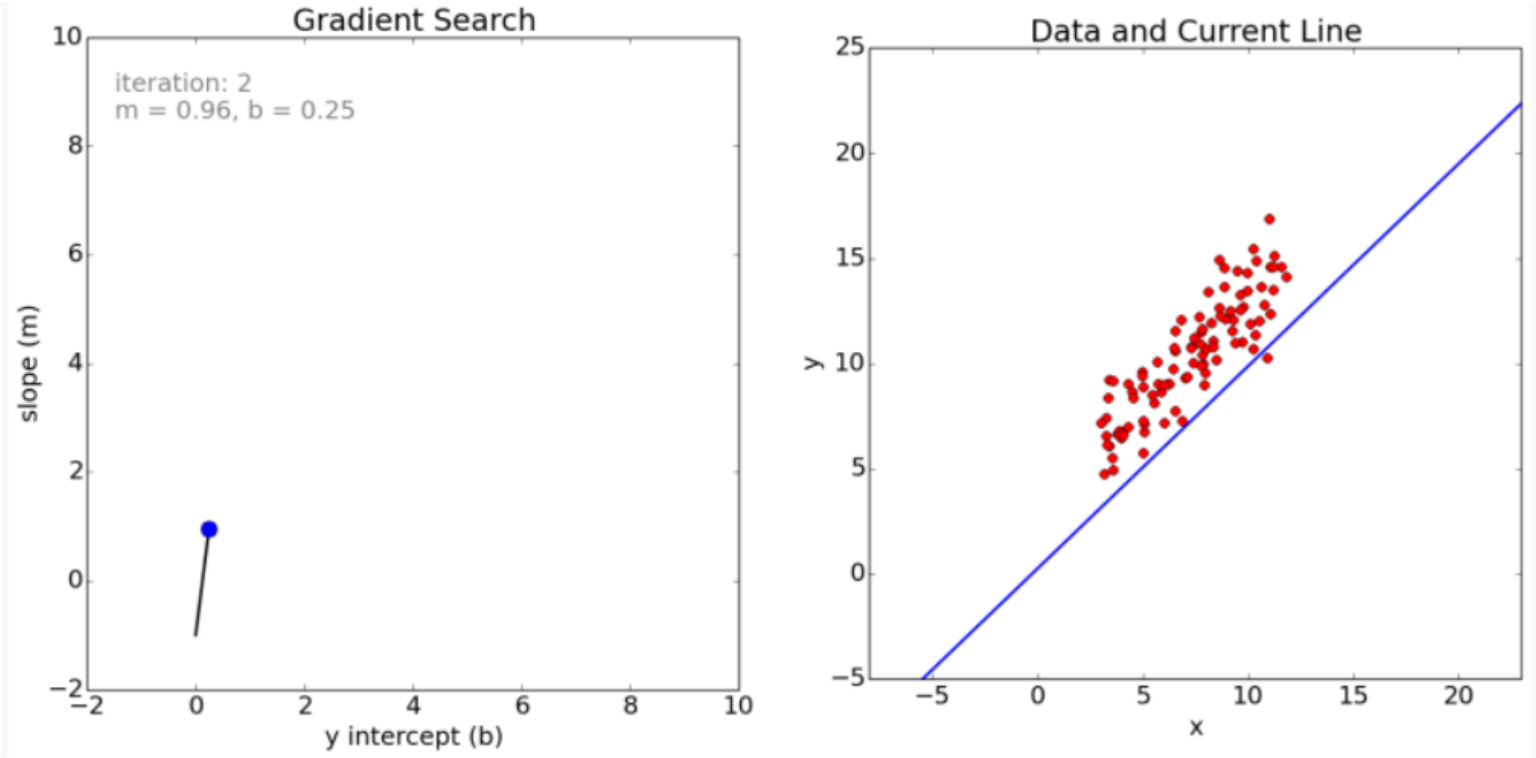
Парная регрессия



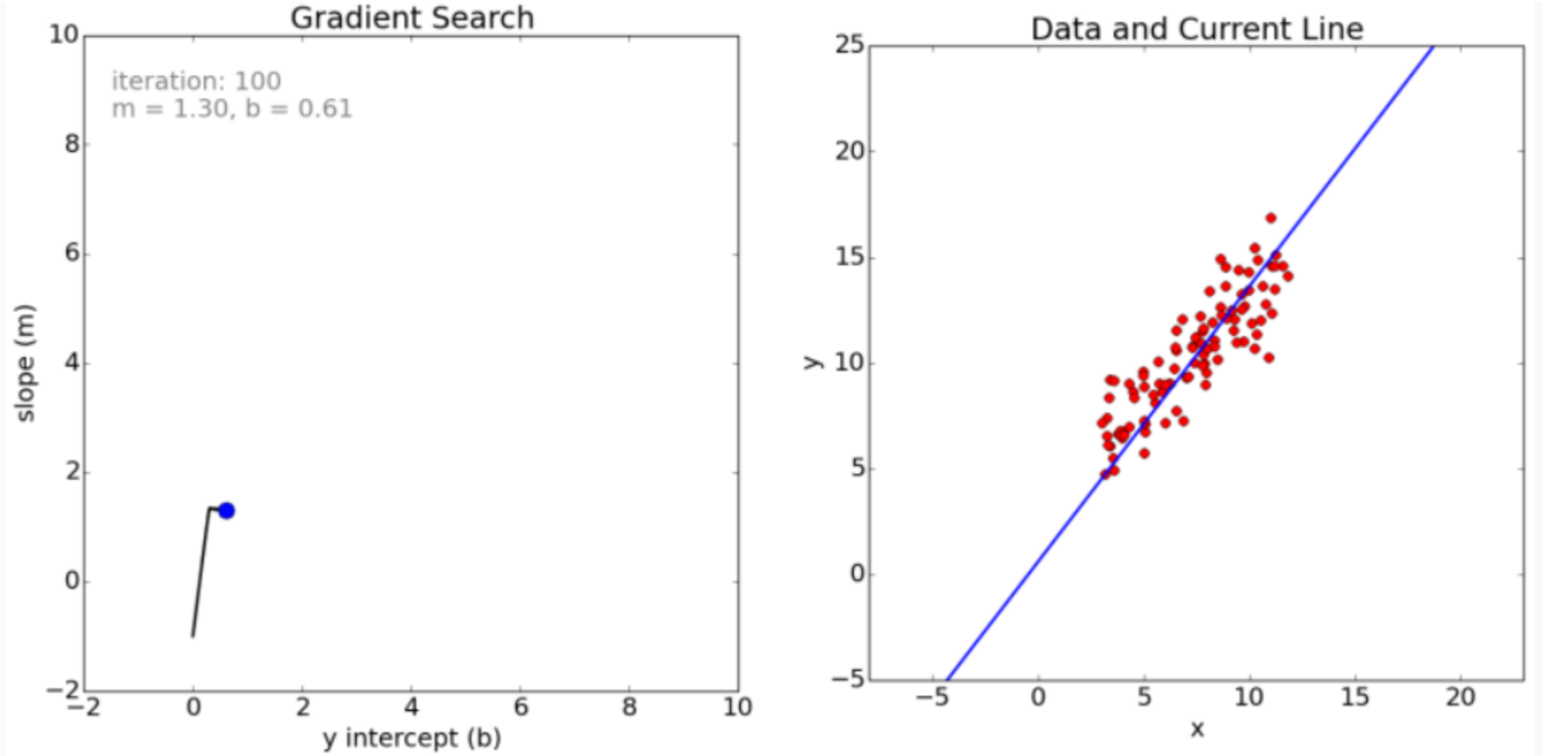
Парная регрессия



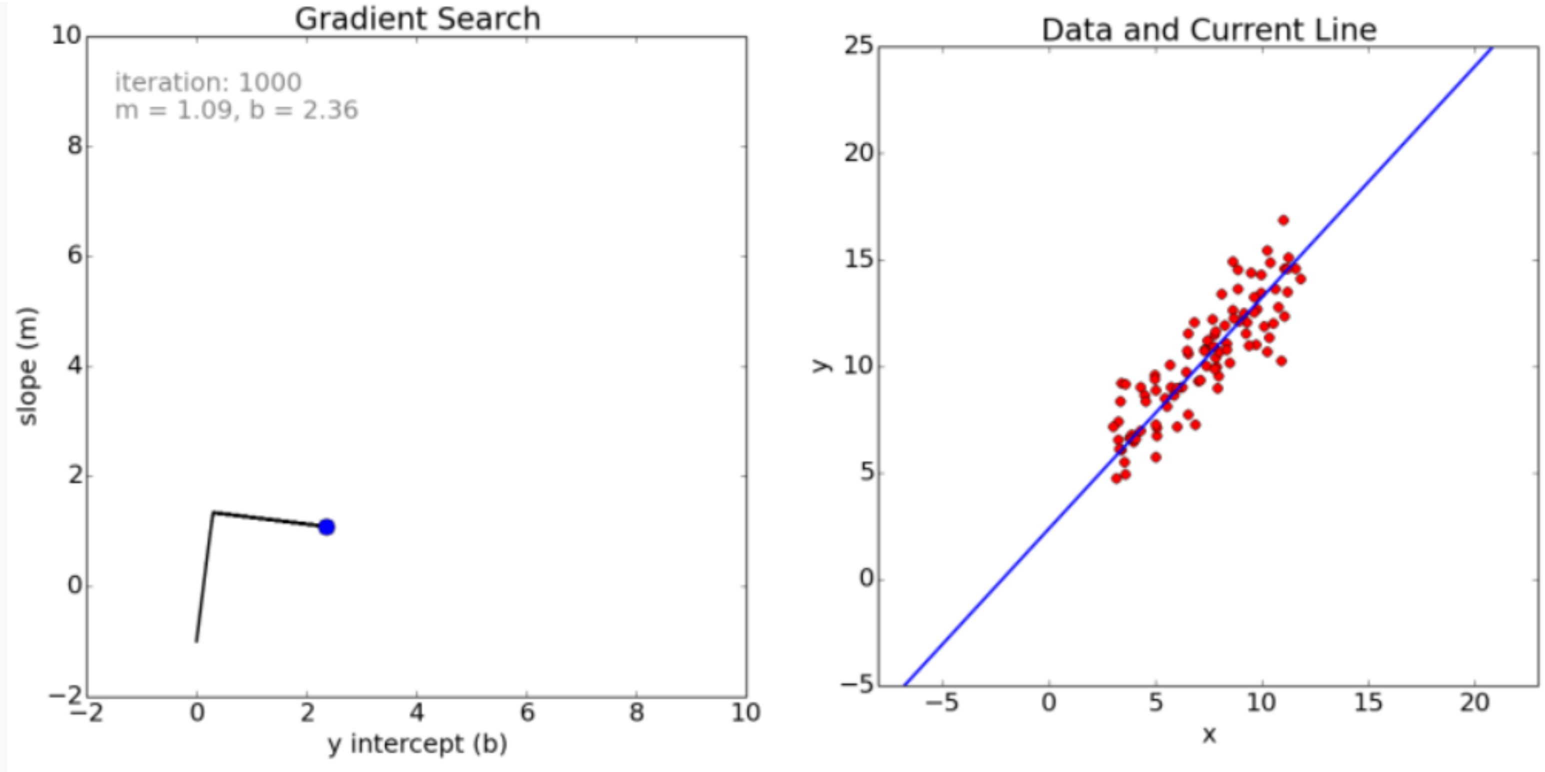
Парная регрессия



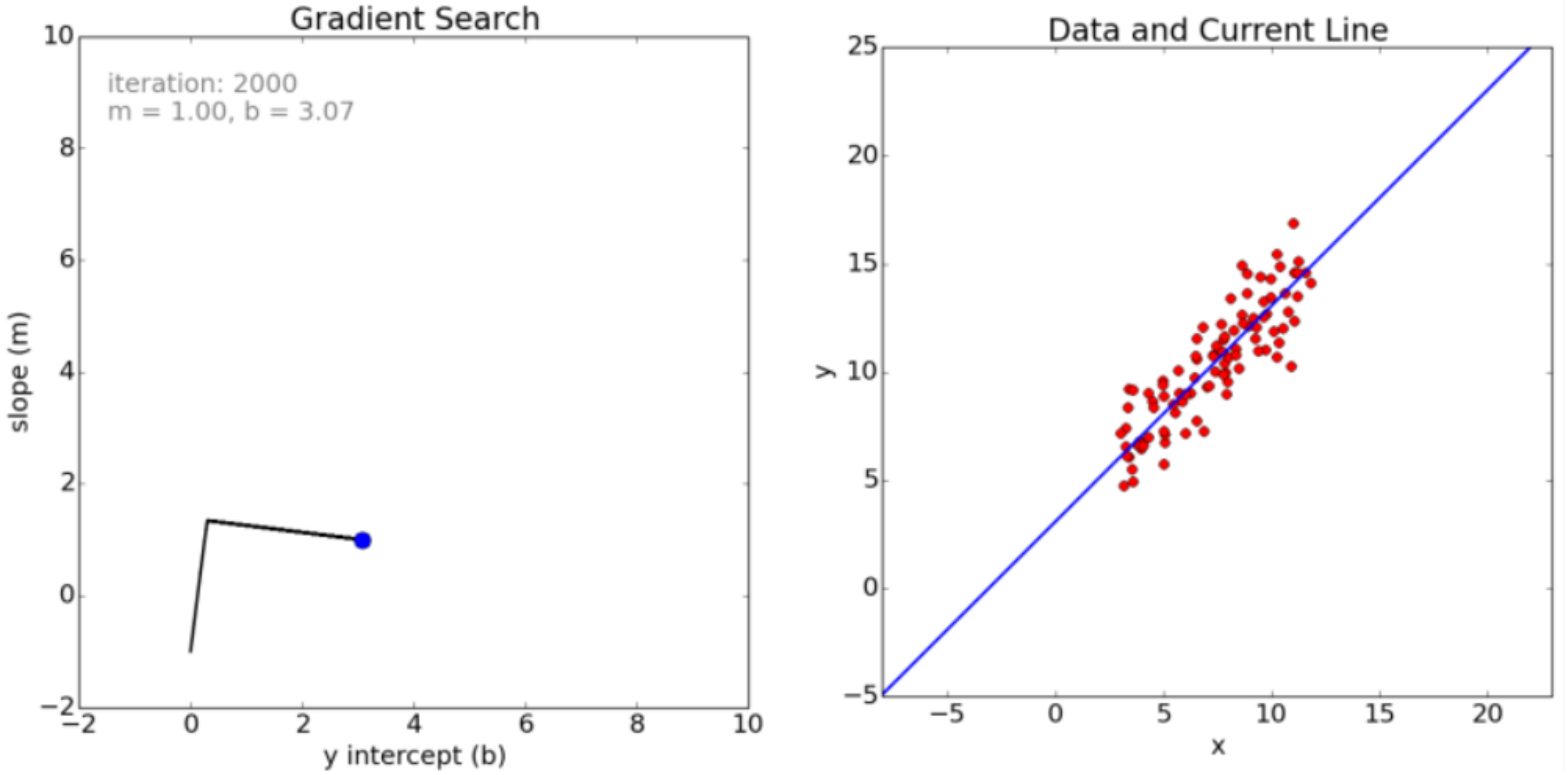
Парная регрессия



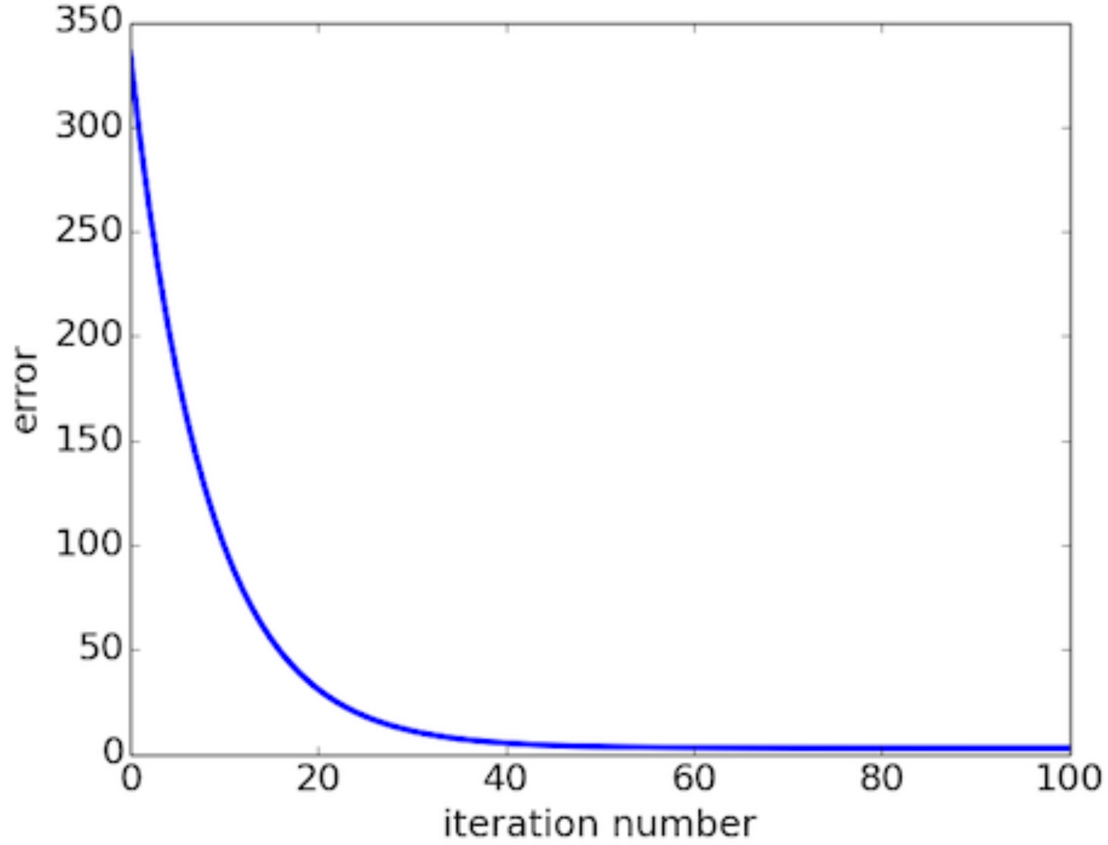
Парная регрессия



Парная регрессия

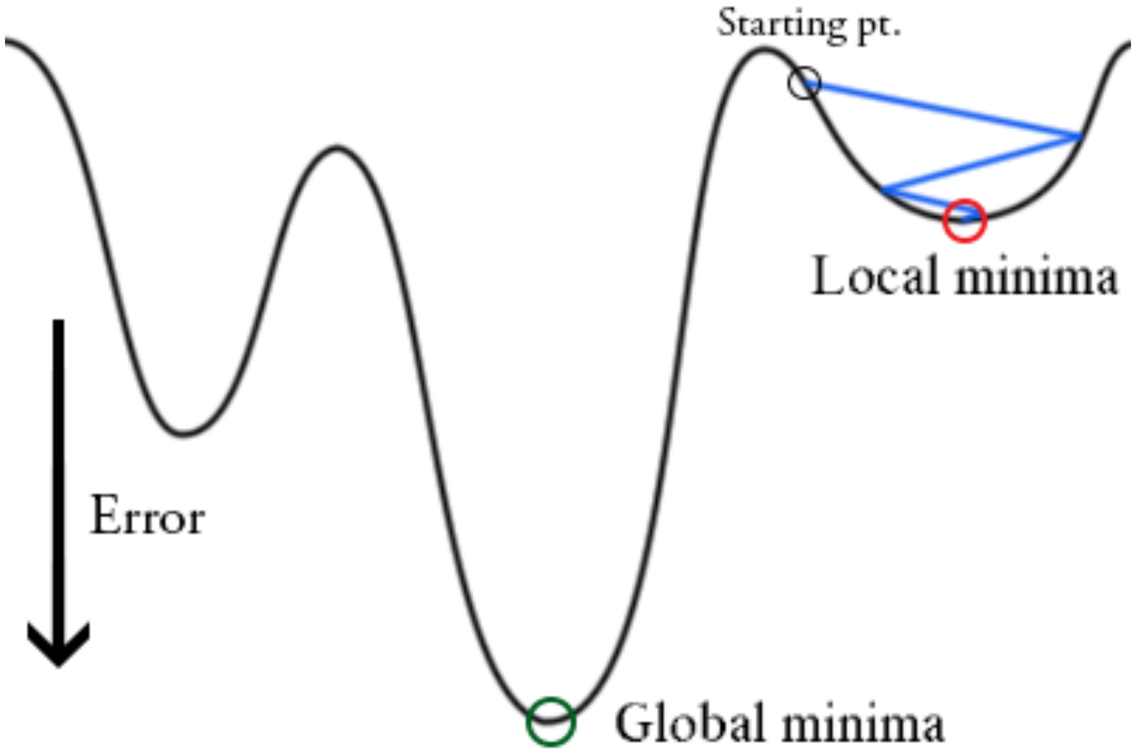


Функционал качества



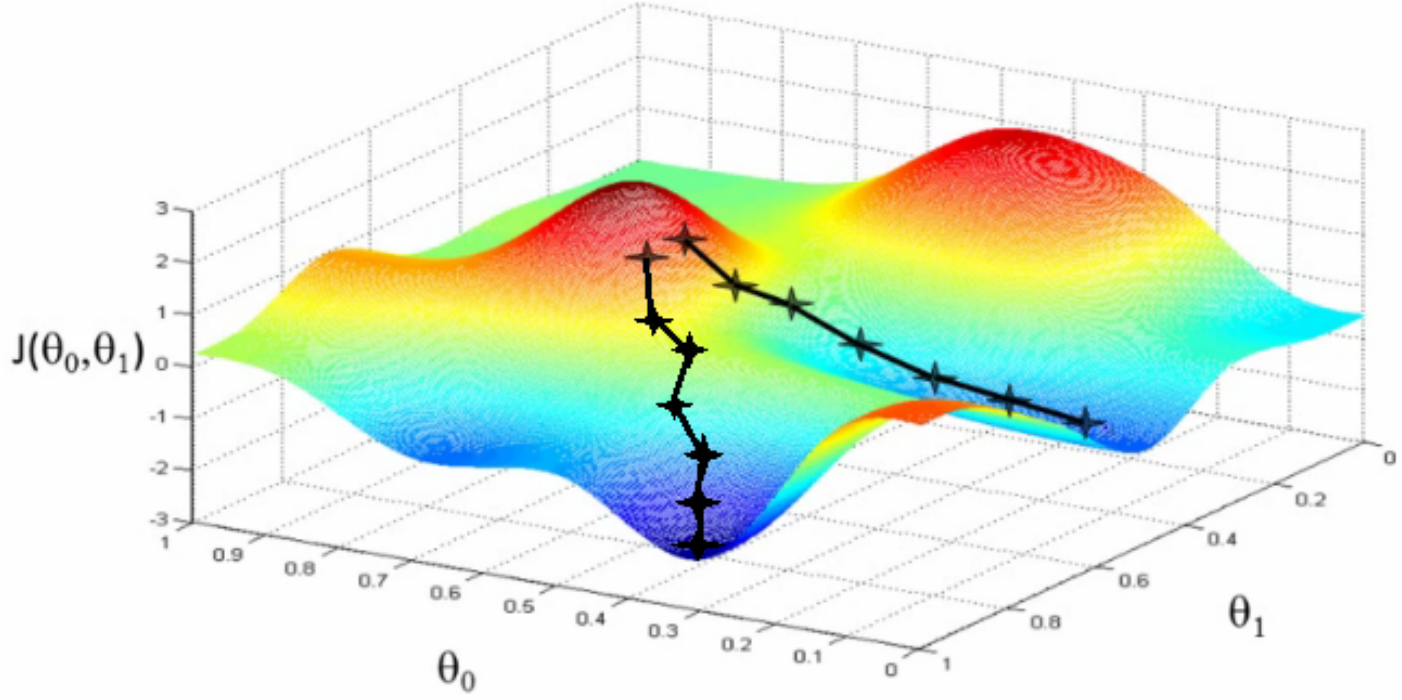
Локальные минимумы

- Градиентный спуск находит только локальные минимумы



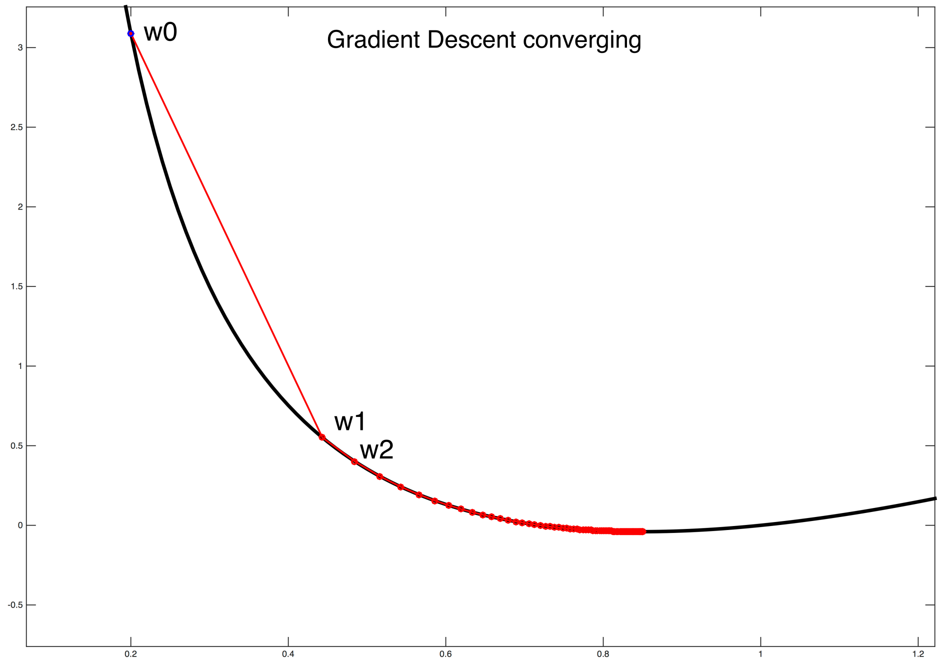
Локальные минимумы

- Результат зависит от начального приближения
- Мультистарт

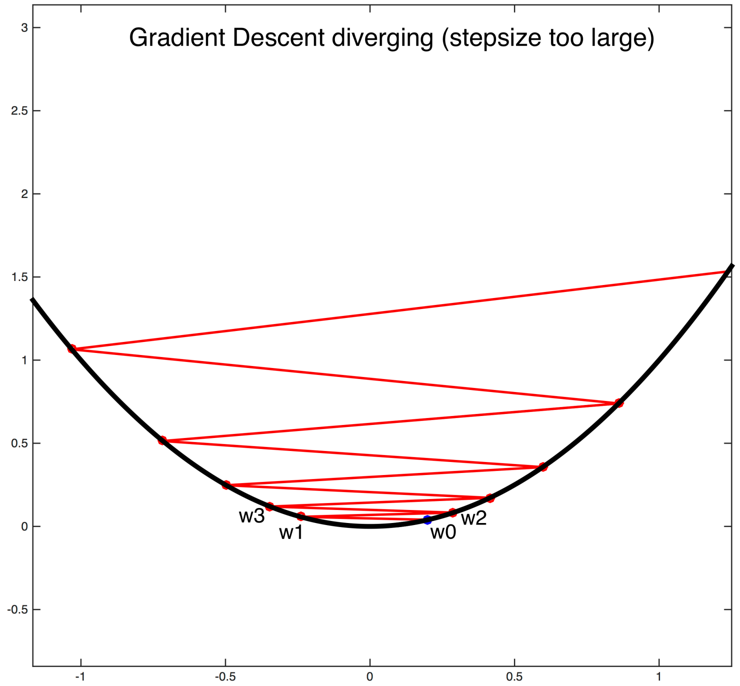


Размер шага

- Выбор размера шага η — искусство



Маленький шаг



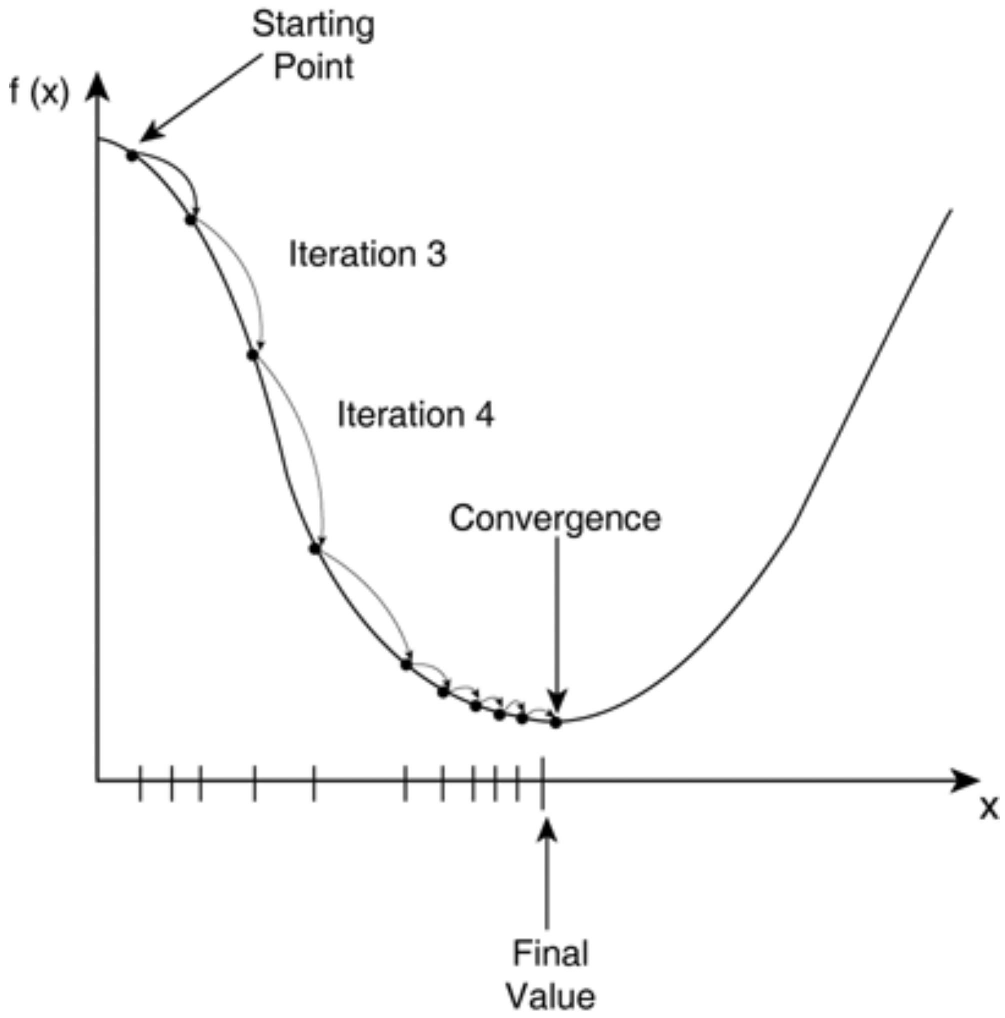
Большой шаг

Размер шага

- Маленький шаг — больше шансов на сходимость, но требуется больше итераций
- Большой шаг — есть риск отсутствия сходимости
- Наискорейший градиентный спуск:
$$\eta_t = \arg\min_{\eta} Q\left(w^{t-1} - \eta\, \nabla Q(w^{t-1})\right)$$
- Нужно делать одномерный поиск на каждой итерации

Размер шага

- Обычно пользуются эвристиками
- Чем ближе к минимуму, тем меньше надо шагать
- Неплохо работает: $\eta_t = \frac{1}{t}$
- Еще лучше: $\eta_t = \lambda \left(\frac{s}{s+t}\right)^p$, где λ, s, p — параметры



Системы линейных уравнений

- $Xw = y$
- Можно решать градиентным спуском следующую задачу:
$$\|Xw - y\|^2 \to \min_w$$
- Функционал выпуклый
- Если решение есть — минимальное значение равно нулю
- Если решения нет — найдем наилучшее приближение

Другие методы оптимизации

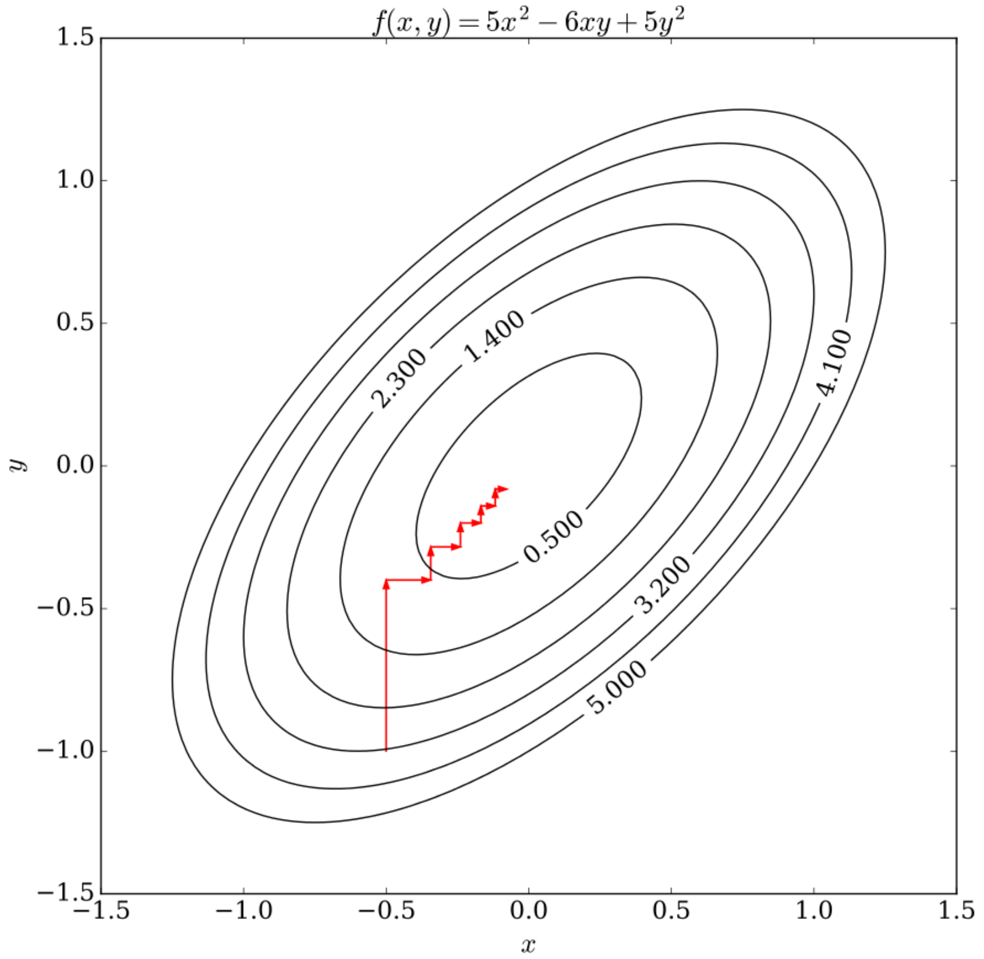
- Методы первого порядка — используют первые производные
  - Градиентный спуск
  - Стохастический градиентный спуск
  - Квазиньютоновские методы, BFGS
  - Stochastic Average Gradient, Nesterov momentum, …
- Методы второго порядка — используют вторые производные
  - Метод Ньютона
- Методы нулевого порядка — без производных
  - Покоординатный спуск
  - Стохастическая оптимизация

Покоординатный спуск

- По очереди меняем каждую координату
- Шаг по каждой координате — случайный, наискорейший, эвристический…

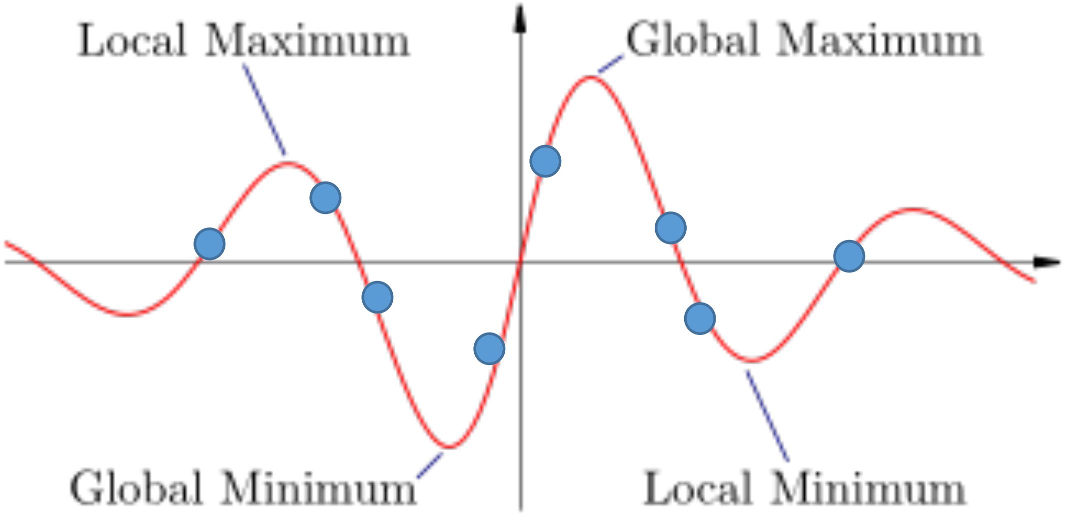
- Быстрые итерации, но может медленно сходиться
- Используется в методе опорных векторов (один из линейных)

Покоординатный спуск



Стохастическая оптимизация

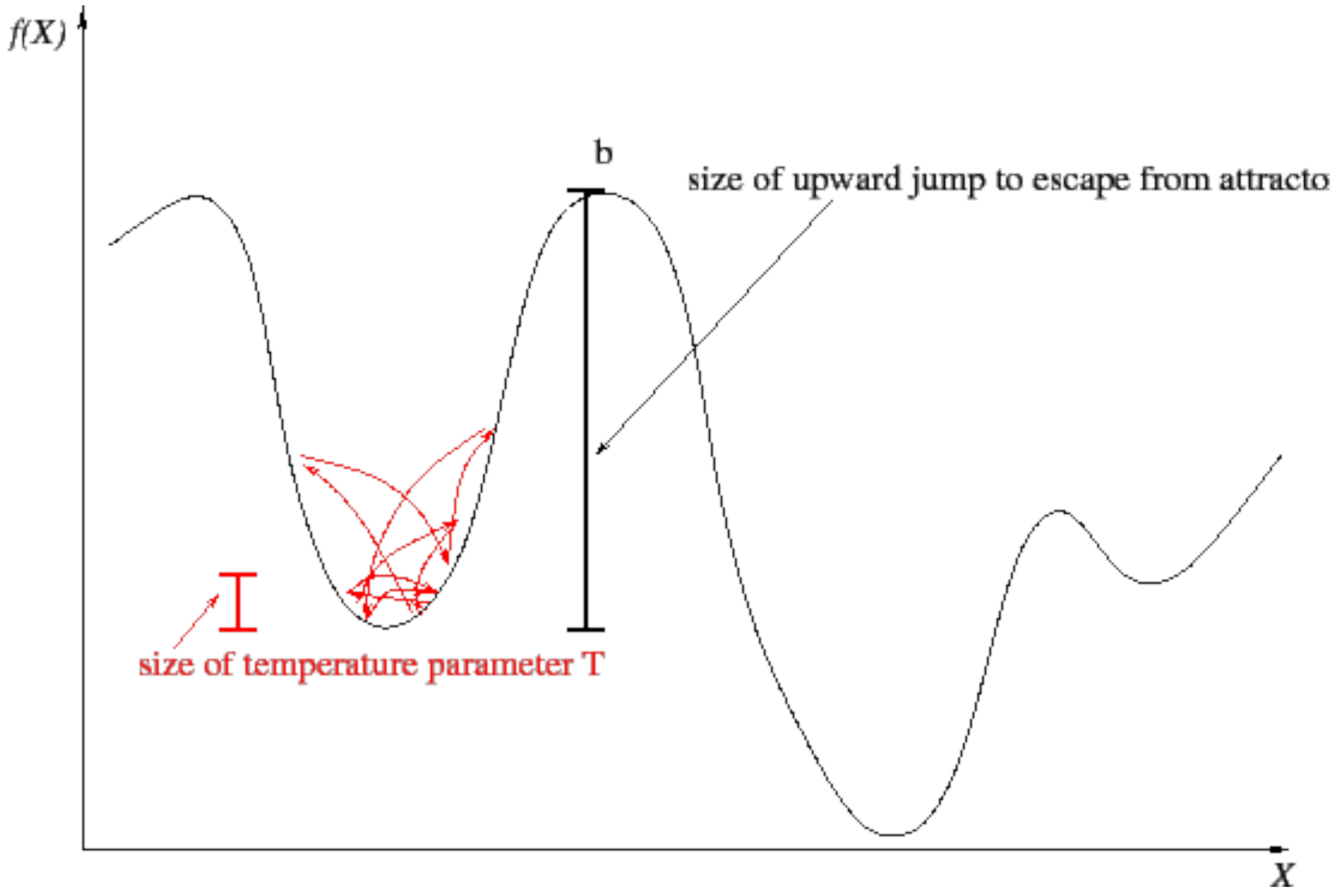
- Простейший алгоритм:
  - Генерируем N раз случайную точку
  - Выбираем ту, на которой значение функционала наименьшее
- Не самый лучший подход
- Нужно более направленной движение



Метод имитации отжига

- w^t — текущее приближение
- Генерируем кандидата w
- Если $Q(w) < Q(w^t)$
  - то переходим: $w^{t+1} = w$
- Если $Q(w) > Q(w^t)$
  - то переходим с вероятностью $\exp\left(-\frac{Q(w)-Q(w^t)}{C_t}\right)$

Метод имитации отжига



Обучение линейной регрессии

Задача оптимизации

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

- Гладкая функция
- Выпуклая функция
- Единственный минимум (не всегда)

Градиентный спуск

- Повторять до сходимости:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

- Сходимость: $\|w^t - w^{t-1}\| < \varepsilon$

Градиент

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

Производные:

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

Нюансы

- Выбор длины шага η — пробуем разные значения
- Выборка должна быть масштабирована
- Признаки не должны коррелировать

Аналитическое решение

- 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?

Аналитическое решение

- Она решается аналитически!

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

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

Мультиколлинеарность

Объекты-признаки

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

- Задача предсказания прибыли магазина в следующем месяце
- Рассмотрим в качестве векторов столбцы матрицы (признаки)

Подозрительные зависимости

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

- Первый и второй признаки: $x_2 = 1000x_1$
- Первый — общий вес товаров в тоннах, второй — в килограммах

Подозрительные зависимости

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

- $x_5 = 0.5x_3 + 0.5x_4$
- Пятый — средняя прибыль за последние два месяца
- Третий и четвертый — прибыль в прошлом и позапрошлом месяце

Линейная зависимость

— один из векторов равен сумме с весами остальных векторов

- Это плохо:
  - Избыточная информация
  - Лишние затраты на хранение данных
  - Вредит некоторым методам машинного обучения

Линейная зависимость

- Пусть дан набор векторов $x_1, \dots, x_n$
- Они линейно зависимы, если
  - существуют такие числа $\beta_1, \dots, \beta_n$,
  - хотя бы одно из которых не равно нулю,
  - что сумма векторов с такими коэффициентами равна нулю

$$\beta_1 x_1 + \cdots + \beta_n x_n = 0$$

Мультиколлинеарность

- Наличие зависимостей между признаками
- Приводит к тому, что решений бесконечное число
- Далеко не все из них имеют хорошую обобщающую способность

Линейная зависимость

- Худший случай — линейно зависимые признаки
- Существуют такие $\alpha = (\alpha_1, \dots, \alpha_d)$, что для любого объекта:

$$\alpha_1 x^1 + \dots + \alpha_d x^d = \langle \alpha, x \rangle = 0$$

Линейная зависимость

- Допустим, мы нашли решение w_*
- Модифицируем: $w_1 = w_* + t\alpha$
- (t — число)
- Ответ нового алгоритма на любом объекте:

$$\langle w_1, x \rangle = \langle w_* + t\alpha, x \rangle = \langle w_*, x \rangle + t\langle \alpha, x \rangle = \langle w_*, x \rangle$$

- w_1 — тоже решение!

Коррелирующие признаки

- Тоже плохо
- Сначала разберёмся с корреляцией

Коэффициент корреляции

$$\rho(\xi, \eta) = \frac{\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)}{\sqrt{\mathbb{D}\xi \mathbb{D}\eta}}$$

Выборочная корреляция:

$$\rho(x, z) = \frac{\sum_{i=1}^{\ell}(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{\ell}(x_i - \bar{x})^2 \sum_{i=1}^{\ell}(z_i - \bar{z})^2}}$$

$$\bar{x} = \frac{1}{\ell}\sum_{j=1}^{\ell} x_j\,; \qquad \bar{z} = \frac{1}{\ell}\sum_{j=1}^{\ell} z_j$$

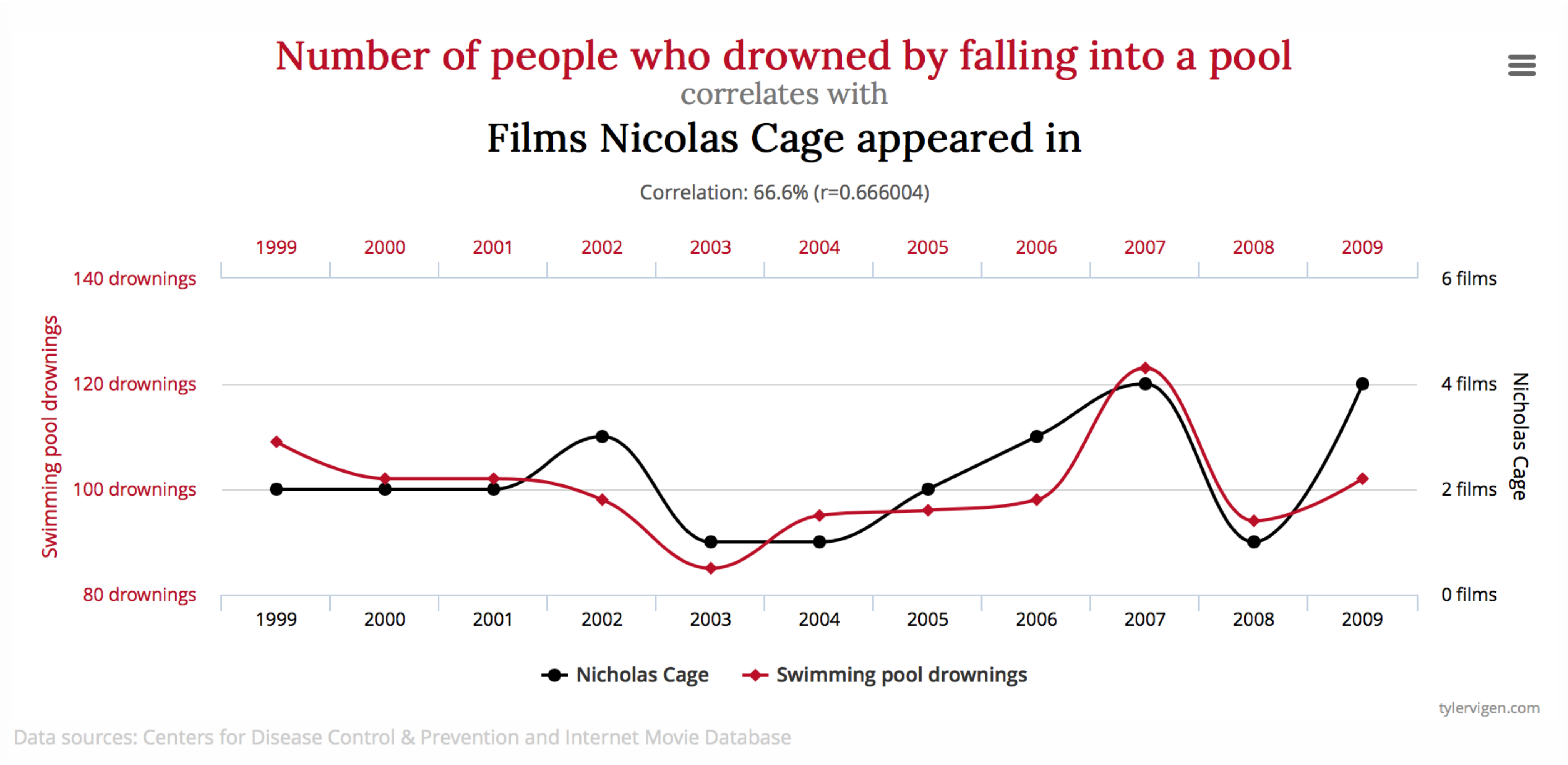
Коэффициент корреляции

$$\rho(x, z) = \frac{\sum_{i=1}^{\ell}(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{\ell}(x_i - \bar{x})^2 \sum_{i=1}^{\ell}(z_i - \bar{z})^2}}$$

- $\rho(x, z) \in [-1, +1]$
- Очень грубо: чем ближе к +1 или -1, тем точнее выполнено уравнение
$$x = az + b$$

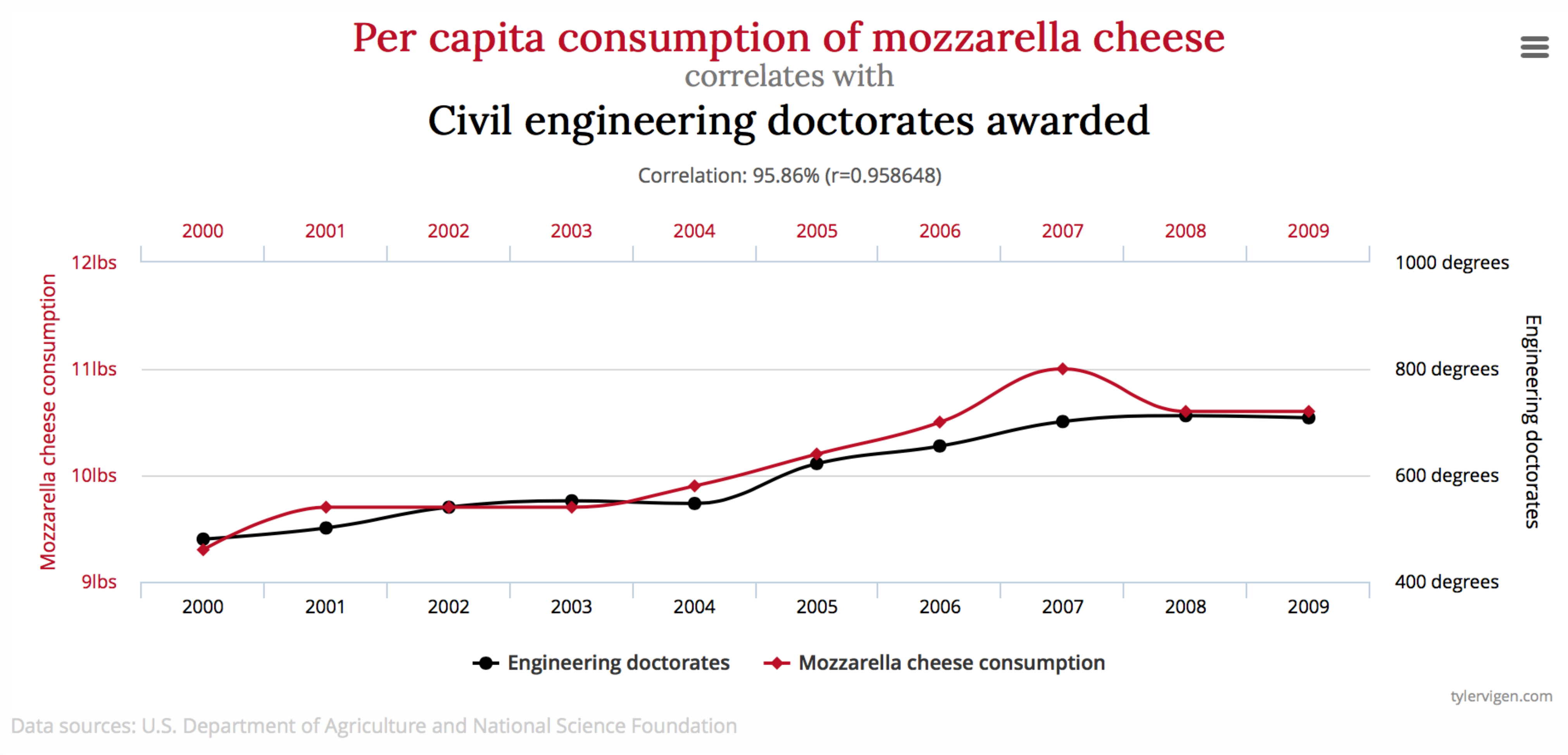
- Мера линейной зависимости

Пример



Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
Correlation: 66.6% (r=0.666004)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com
Data sources: Centers for Disease Control & Prevention and Internet Movie Database

Пример



Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded
Correlation: 95.86% (r=0.958648)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
12lbs
11lbs
10lbs
9lbs
Mozzarella cheese consumption
1000 degrees
800 degrees
600 degrees
400 degrees
Engineering doctorates
Engineering doctorates
Mozzarella cheese consumption
tylervigen.com
Data sources: U.S. Department of Agriculture and National Science Foundation

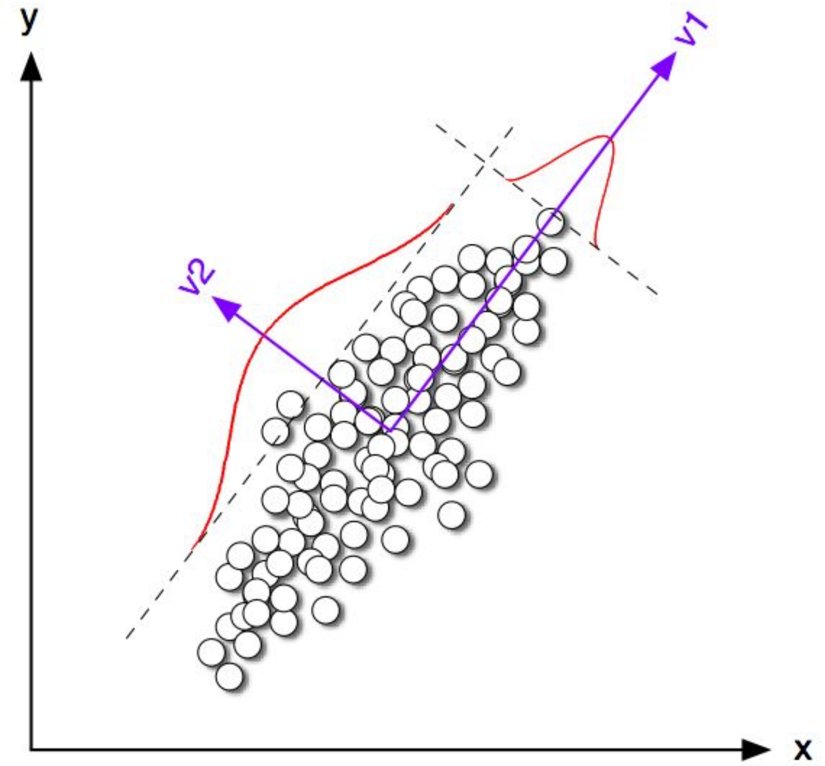
Распространённое заблуждение

- Может показаться, что из корреляции следует причинно-следственная связь
- Это не так!
- Корреляция означает, что события часто происходят вместе
- Но никак не следуют друг из друга

- Больше примеров: http://tylervigen.com/spurious-correlations

Коррелирующие признаки

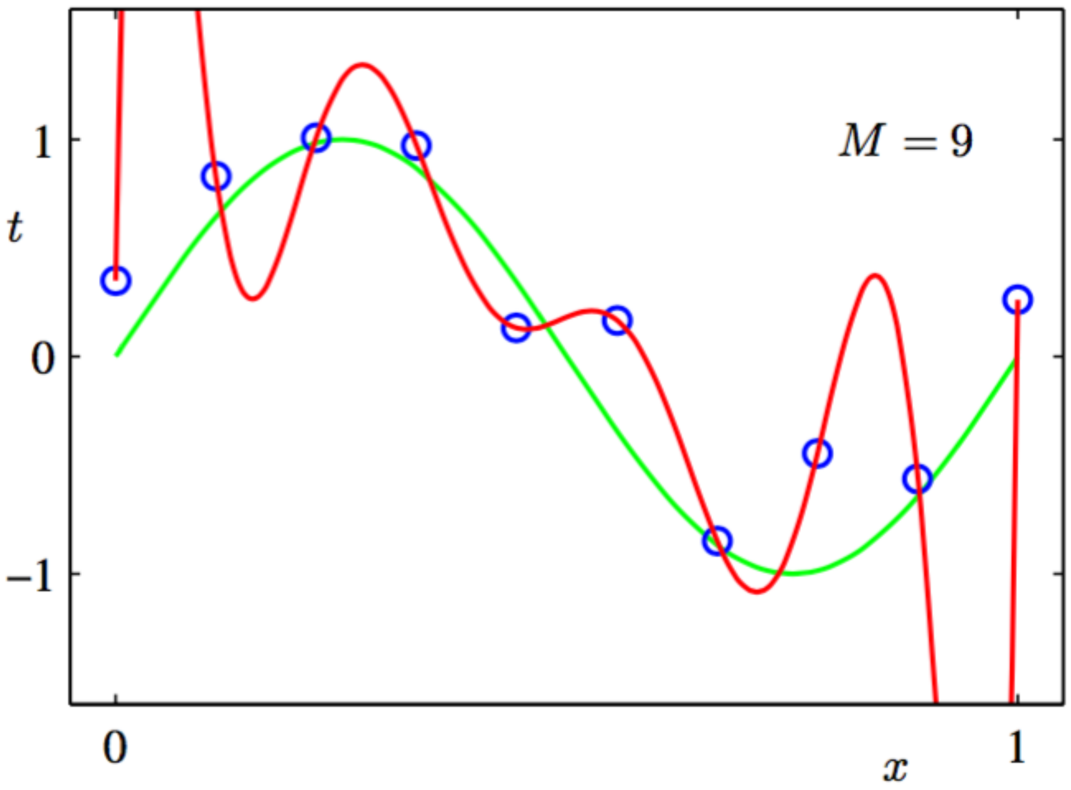
- Плохо, если есть коррелирующие признаки
- Решение: отбор признаков или их декорреляция
- В следующих лекциях



Переобучение и регуляризация

Пример

- Один признак x
- $a(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_9 x^9$



Пример

- Коэффициенты:

$$a(x) = 0.5 + 13458922x - 43983740x^2 + \cdots + 2740x^9$$

- Большие коэффициенты — симптом переобучения
- (эмпирическое наблюдение)

Симптом переобучения

- Большие коэффициенты в линейной модели — это плохо
- Пример: предсказание роста по весу
  - $a(x) = 698x - 41714$
- Изменение веса на 0.01 кг приведет к изменению роста на 7 см
- Не похоже не правильную зависимость

Регуляризация

- Будем штрафовать за большие веса!
- Функционал:

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

- Регуляризатор:

$$\|w\|^2 = \sum_{j=1}^{d} w_j^2$$

Регуляризация

- Регуляризованный функционал ошибки:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(\langle w, x_i\rangle - y_i)^2 + \lambda\|w\|^2 \to \min_w$$

- Всё ещё гладкий и выпуклый

Коэффициент регуляризации

- λ — новый параметр, надо подбирать
- Высокий λ — простые модели
- Низкий λ — риск переобучения
- Нужно балансировать
- Подбор λ — с помощью кросс-валидации

Смысл регуляризации

- Минимизация регуляризованного функционала равносильна решению условной задачи:

$$\begin{cases} \dfrac{1}{\ell}\sum_{i=1}^{\ell}(\langle w, x_i\rangle - y_i)^2 \to \min\limits_{w} \\ \|w\|^2 \le C \end{cases}$$

L_1-регуляризация

- L_1-регуляризатор:

$$\|w\|_1 = \sum_{j=1}^{d} |w_j|$$

- Регуляризованный функционал ошибки:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(\langle w, x_i\rangle - y_i)^2 + \lambda\|w\|_1 \to \min_w$$

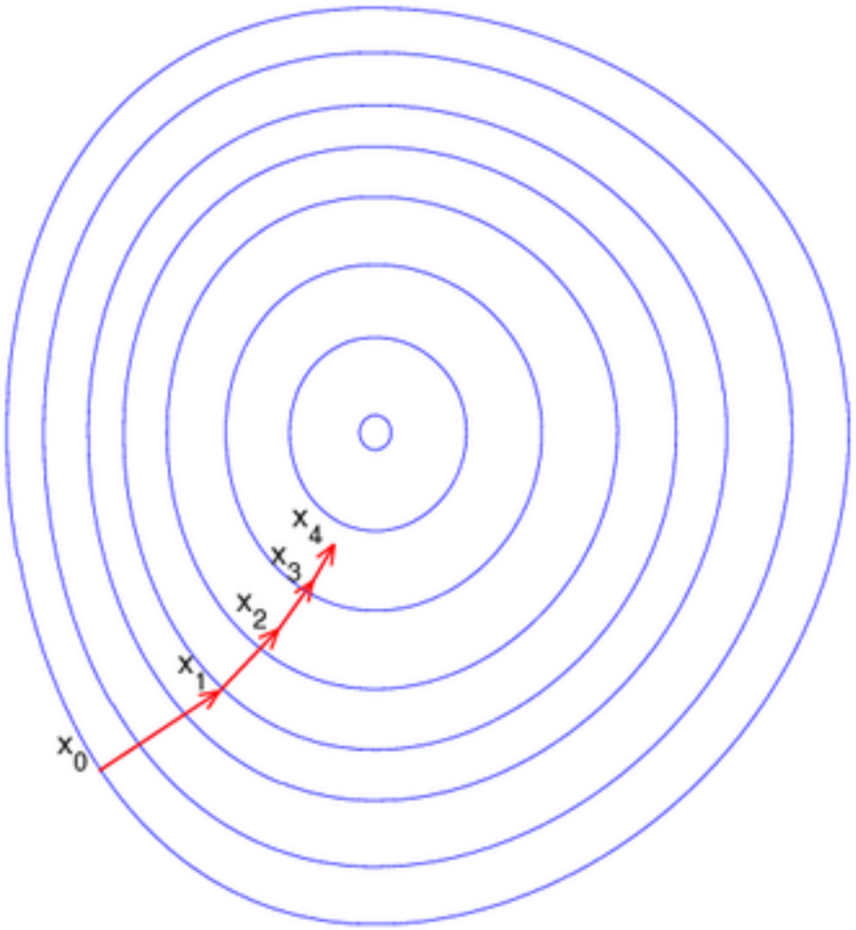
L_1-регуляризация

- Функционал становится негладким
- Сложнее оптимизировать
- Зато производится отбор признаков
- Часть весов в решении будут нулевыми

Масштабирование признаков

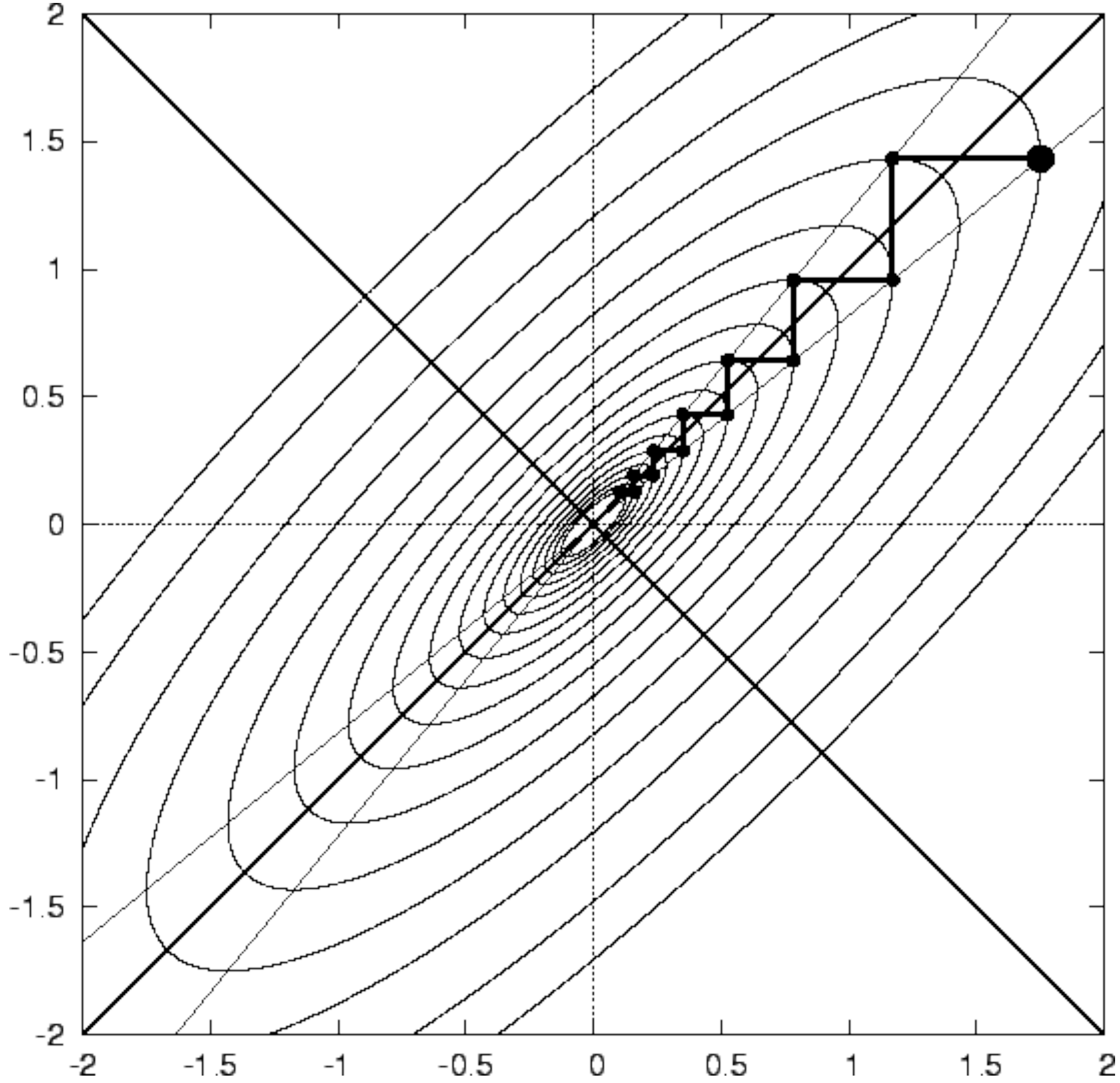
Масштабирование выборки

Хороший случай



Масштабирование выборки

Плохой случай



Масштабирование выборки

- Задача: одобрят ли заявку на грант?
- 1-й признак: сколько успешных заявок было до этого у заявителя
- 2-й признак: год рождения заявителя

- Масштаб: единицы и тысячи

- Все признаки должны иметь одинаковый масштаб

Масштабирование выборки

- Отмасштабируем j-й признак
- Вычисляем среднее и стандартное отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} \left(x_i^j - \mu_j\right)^2}$$

Масштабирование выборки

- Отмасштабируем j-й признак
- Вычтем из каждого значения признака среднее и поделим на стандартное отклонение:

$$x_i^j := \frac{x_i^j - \mu_j}{\sigma_j}$$

Резюме

- Градиентный спуск — универсальный инструмент обучения дифференцируемых моделей
- Линейные зависимости и корреляции в признаках приводят к проблемам при обучении
- Регуляризация — способ борьбы с переобучением
- Масштабирование помогает улучшить сходимость градиентных методов