Лекция 4.
Линейные модели

Введение в машинное обучение

Обозначения

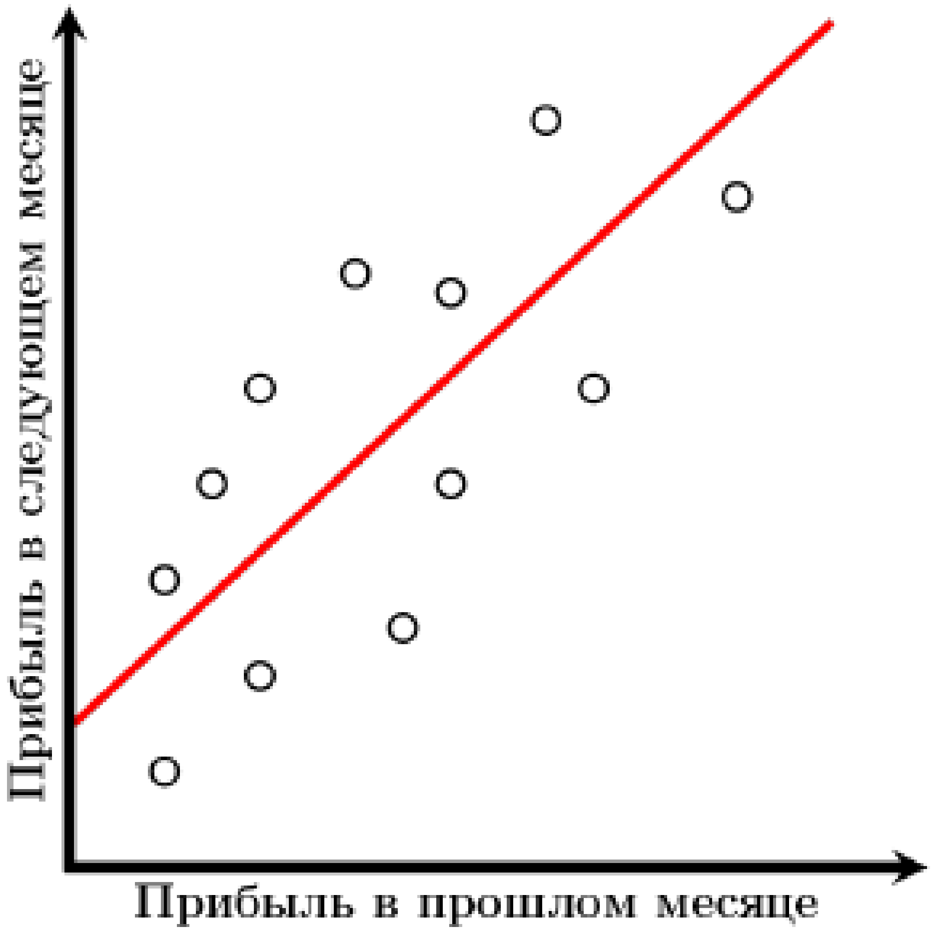
- X – пространство объектов
- Y – пространство ответов
- $x = \left(x^1, \dots, x^d\right)$ – признаковое описание объекта
- $\mathrm{X} = (x_i, y_i)_{i=1}^{l}$ –обучающая выборка
- $a(x)$ – алгоритм, модель
- $Q(a, X)$ – функционал ошибки алгоритма a на выборке X
- Обучение: $a(x) = argmin_{a \in A}\ \mathrm{Q}(\mathrm{a}, \mathrm{X})$
- $Y = \mathbb{R}$

Напоминание

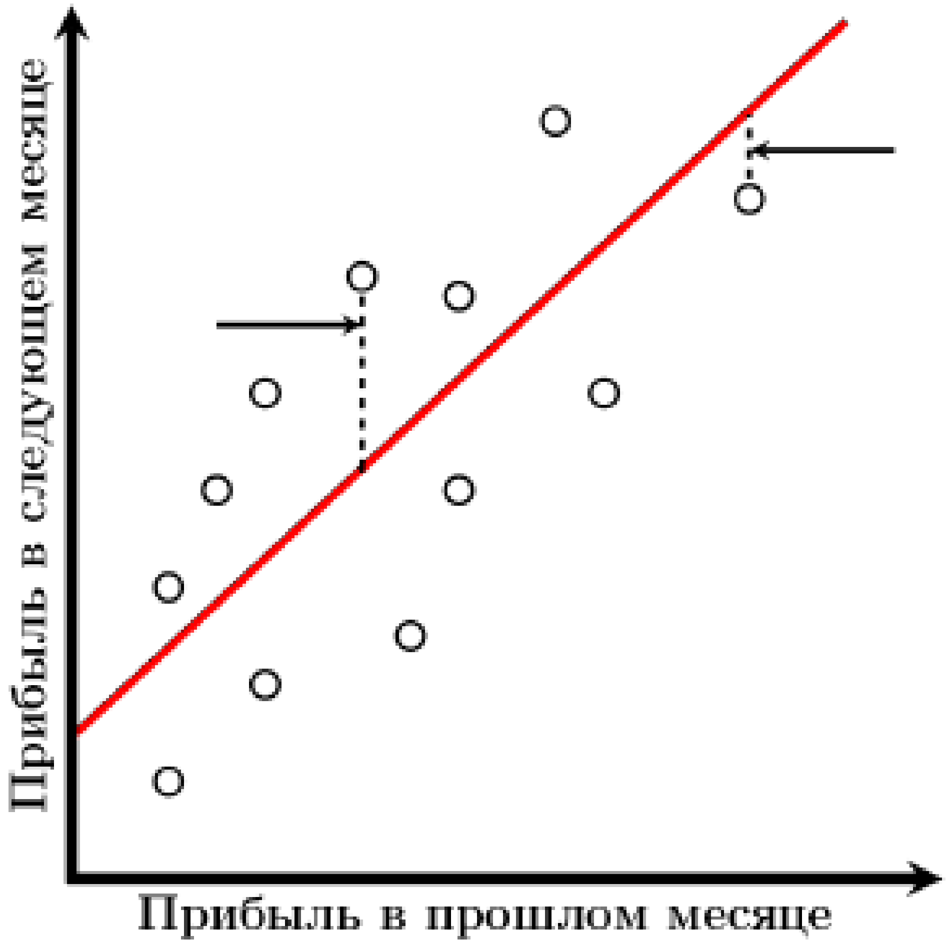
- **Функционал ошибки Q**: способ измерения того, хорошо или плохо работает алгоритм на конкретной выборке
- **Семейство алгоритмов A** : как выглядит множество алгоритмов, из которых выбирается лучший
- **Метод обучения** : как именно выбирается лучший алгоритм из семейства алгоритмов.

Задача регрессии

Задача регрессии



Задача регрессии



Семейство алгоритмов

$$a(x) = \mathbf{w}_0 + \sum_{j=1}^{d} \mathbf{w}_j \, x^j$$

Свободный коэффициент

Веса

Признаки

Семейство алгоритмов

$$a(x) = w_0 + \sum_{j=1}^{d} w_i x_i$$

$$a(x) = \sum_{i=0}^{m} w_i x_i = \langle w, x \rangle$$

$$y = XW + \varepsilon$$

Линейная регрессия

$$y = XW + \varepsilon$$

$$y_i = \sum_{j=0}^{m} w_j X_{ij} + \varepsilon_i$$

- где $y \epsilon R^n -$ целевая переменная
- $w \epsilon R^{m+1} -$ вектор параметров модели(весы)
- $X -$ матрица наблюдений и признаков размерности n строк на $m + 1$ столбцов с полным рангом по столбцам: $rank(X) = m$
- $\varepsilon -$ случайная переменная, соответствующая случайной, непрогнозируемой ошибке модел

Линейная регрессия

На модель накладываются ограничения:

1. Матожидание случайных ошибок равно нулю: $\forall i: E[\varepsilon_i] = 0$
2. Дисперсия случайных ошибок одинакова и конечна:
   $\forall i: Var(\varepsilon_i) = \sigma^2 < \infty$
3. Случайные ошибки не скоррелированы: $\forall i \neq j: Cov(\varepsilon_i, \varepsilon_j) = 0$

Линейная оценка

- Оценка весов называется линейной если :
- $\widehat{w_i} = w_{1i}y_1 + w_{2i}y_2 + \cdots + w_{1n}y_n,$
- Где
  $\forall k\ w_{ki}$ зависит только от наблюдаемых данных X и почти наверняка нелинейно
- Так как решением задачи поиска оптимальной весов будет именно линейная оценка, то и модель называется линейной регрессией
- Один из способов вычислить значения параметров является метод наименьших квадратов (МНК), который минимизирует среднеквадратичную ошибку между реальным значением зависимой переменной и прогнозом, выданной моделью

Метод наименьших квадратов

$$
\begin{aligned}
\mathcal{L}\left(\mathbf{X}, \mathbf{y}, \mathbf{w}\right) &= \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right)^2 \\
&= \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 \\
&= \frac{1}{2n} (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}} \left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)
\end{aligned}
$$

Шпаргалка по матричным производным

$$\frac{\partial}{\partial \mathbf{X}} \mathbf{X}^{\mathrm{T}} \mathbf{A} = \mathbf{A}$$

$$\frac{\partial}{\partial \mathbf{X}} \mathbf{X}^{\mathrm{T}} \mathbf{A} \mathbf{X} = \left(\mathbf{A} + \mathbf{A}^{\mathrm{T}}\right) \mathbf{X}$$

$$\frac{\partial}{\partial \mathbf{A}} \mathbf{X}^{\mathrm{T}} \mathbf{A} \mathbf{y} = \mathbf{X}^{\mathrm{T}} \mathbf{y}$$

$$\frac{\partial}{\partial \mathbf{X}} \mathbf{A}^{-1} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \mathbf{X}} \mathbf{A}^{-1}$$

МНК. Дифференцирование

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \frac{1}{2n} \left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} \right) \\
&= \frac{1}{2n} \left(-2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} \right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \;\; &\Leftrightarrow \;\; \frac{1}{2n} \left(-2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} \right) = 0 \\
&\Leftrightarrow \;\; -\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = 0 \\
&\Leftrightarrow \;\; \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \\
&\Leftrightarrow \;\; \mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}
\end{aligned}
$$

Аналитическое решение

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

- Нужно обращать матрицу d x d – сложность d^3
- Могут возникнуть численные проблемы

Обучение линейной регрессии

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

- d неизвестных
- Есть константный признак
- Выпуклая функция

Матричная запись

- Матрица объекты-признаки

$$X = \begin{array}{cl} \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{\ell 1} & \cdots & x_{\ell d} \end{pmatrix} & \begin{array}{l} \text{Объект} \\ \\ \\ \end{array} \end{array}$$

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

Признак

Вектор ответов

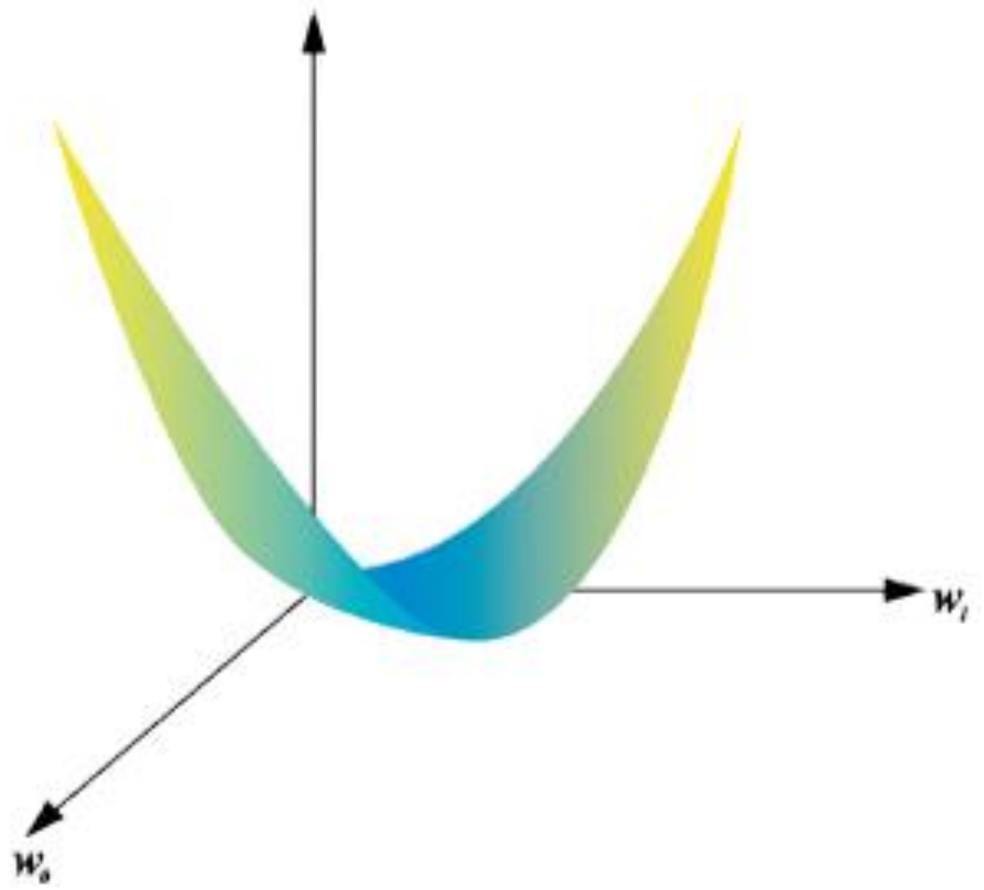
$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_\ell \end{pmatrix}$$

Матричная запись

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

Градиентный спуск

- Функция ошибки гладкая и выпуклая



Градиентный спуск: алгоритм

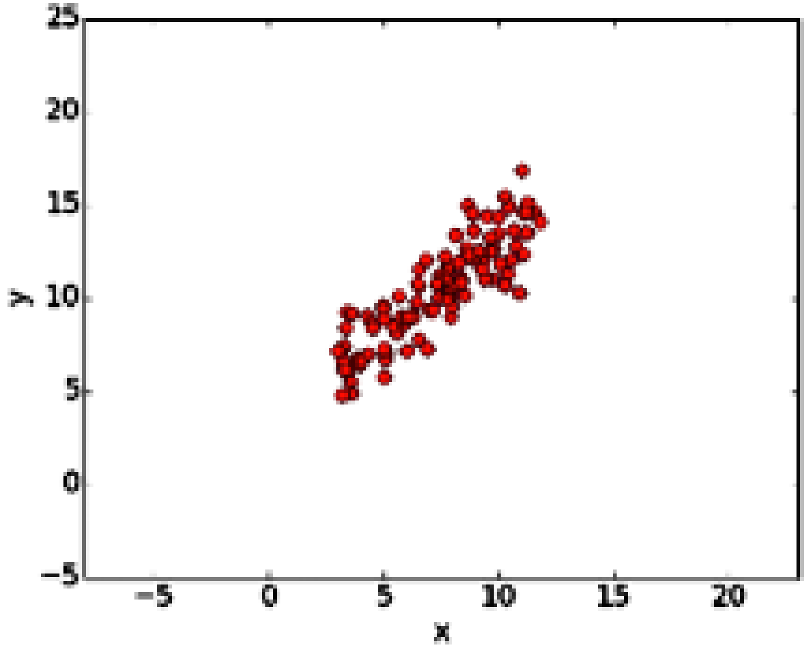
- Инициализация: $w^0 = 0$
- Цикл по t = 1, 2, 3, …:
- $w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, X)$
- Если $\|w^t - w^{t-1}\| < \varepsilon,$ то завершить

Парная регрессия

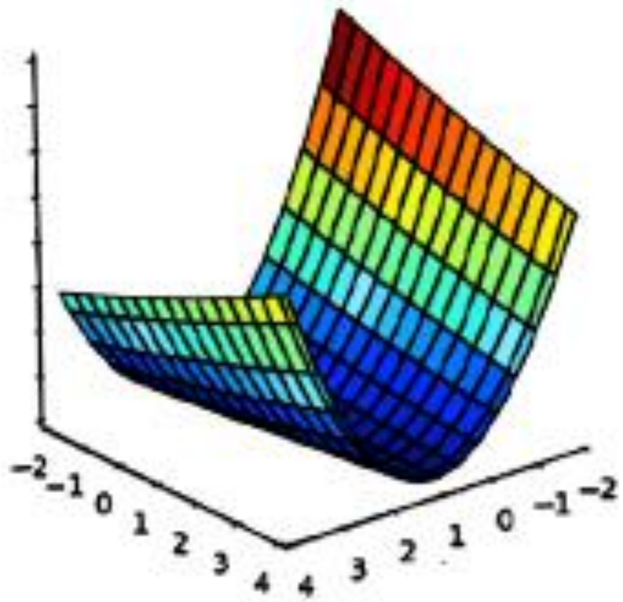
- Простейший случай: один признак
- Модель: $a(x) = w_1 x + w_0$
- Два параметра: w_1 и w_0
- Функционал:

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

Парная регрессия



Выборка



Функционал качества

Градиентный спуск

- Инициализация: $w^0 = 0$
- Цикл по t = 1, 2, 3,…:
- $w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, X)$
- Если $\|w^t - w^{t-1}\| < \varepsilon,$ то завершить

Градиент для парной регрессии

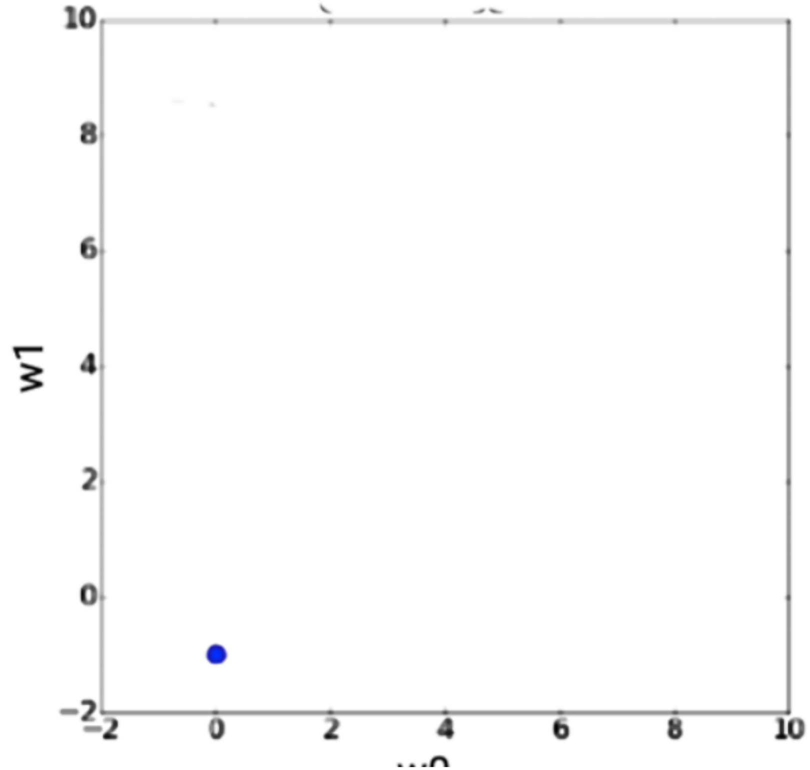
$$Q(w_0, w_1, X) = \frac{1}{l}\sum_{i=1}^{l}(w_1 x_i + w_0 - y_i)^2$$

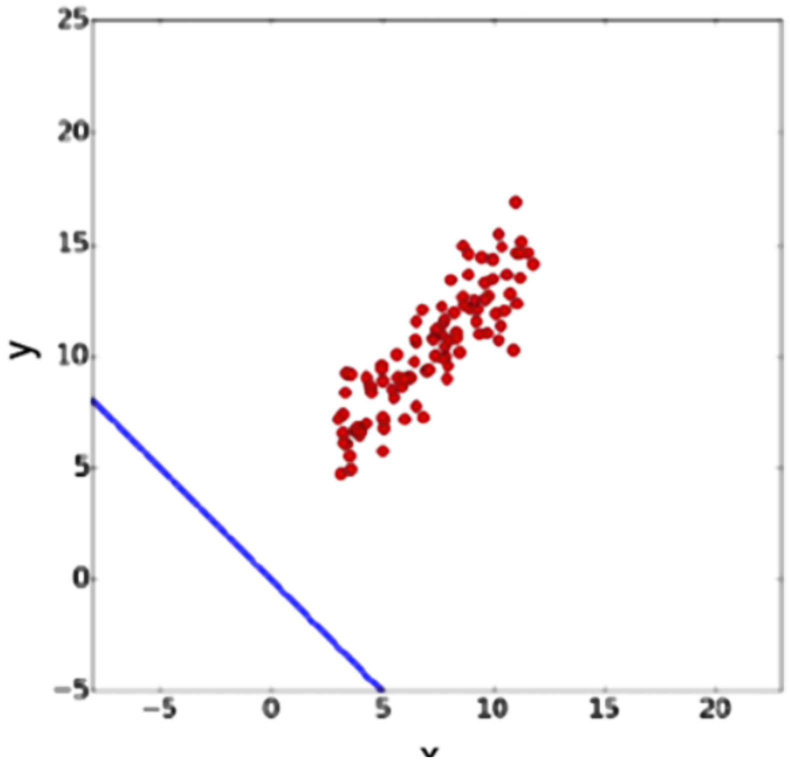
- Частные производные:

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

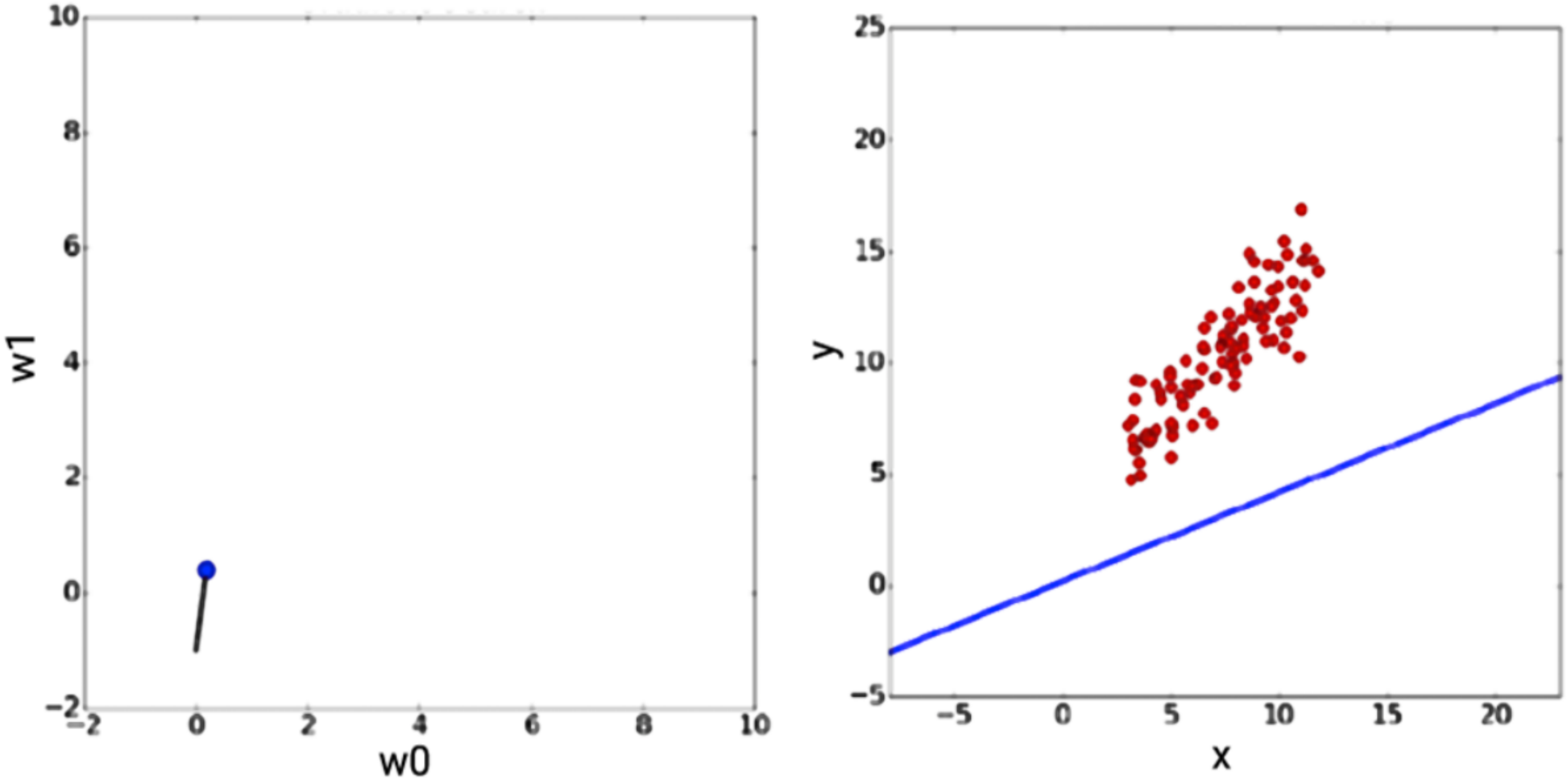
$$\frac{\partial Q}{\partial w_0} = \frac{2}{l}\sum_{i=1}^{l}(w_1 x_i + w_0 - y_i)$$

Парная регрессия – 1 итерация

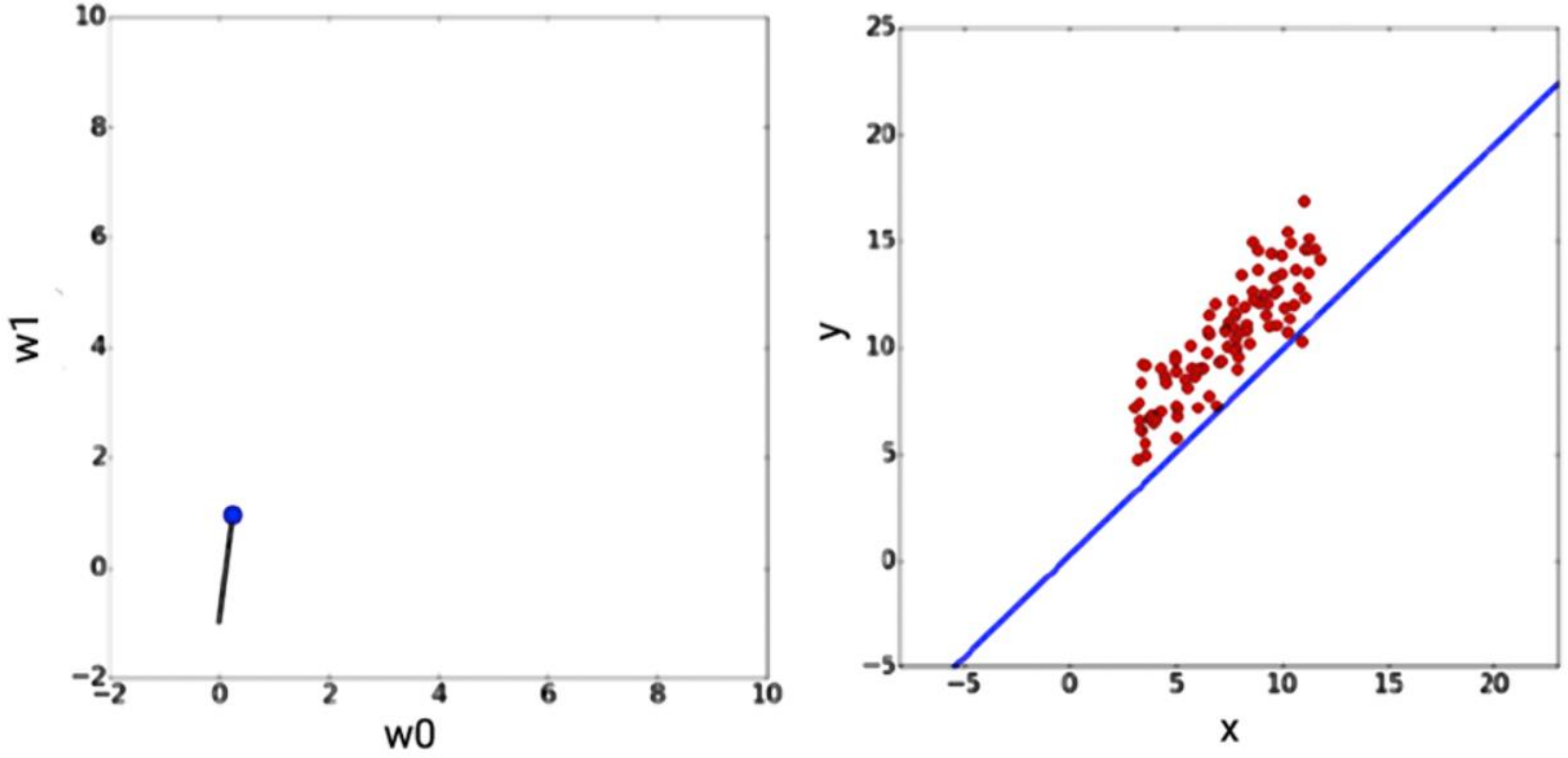




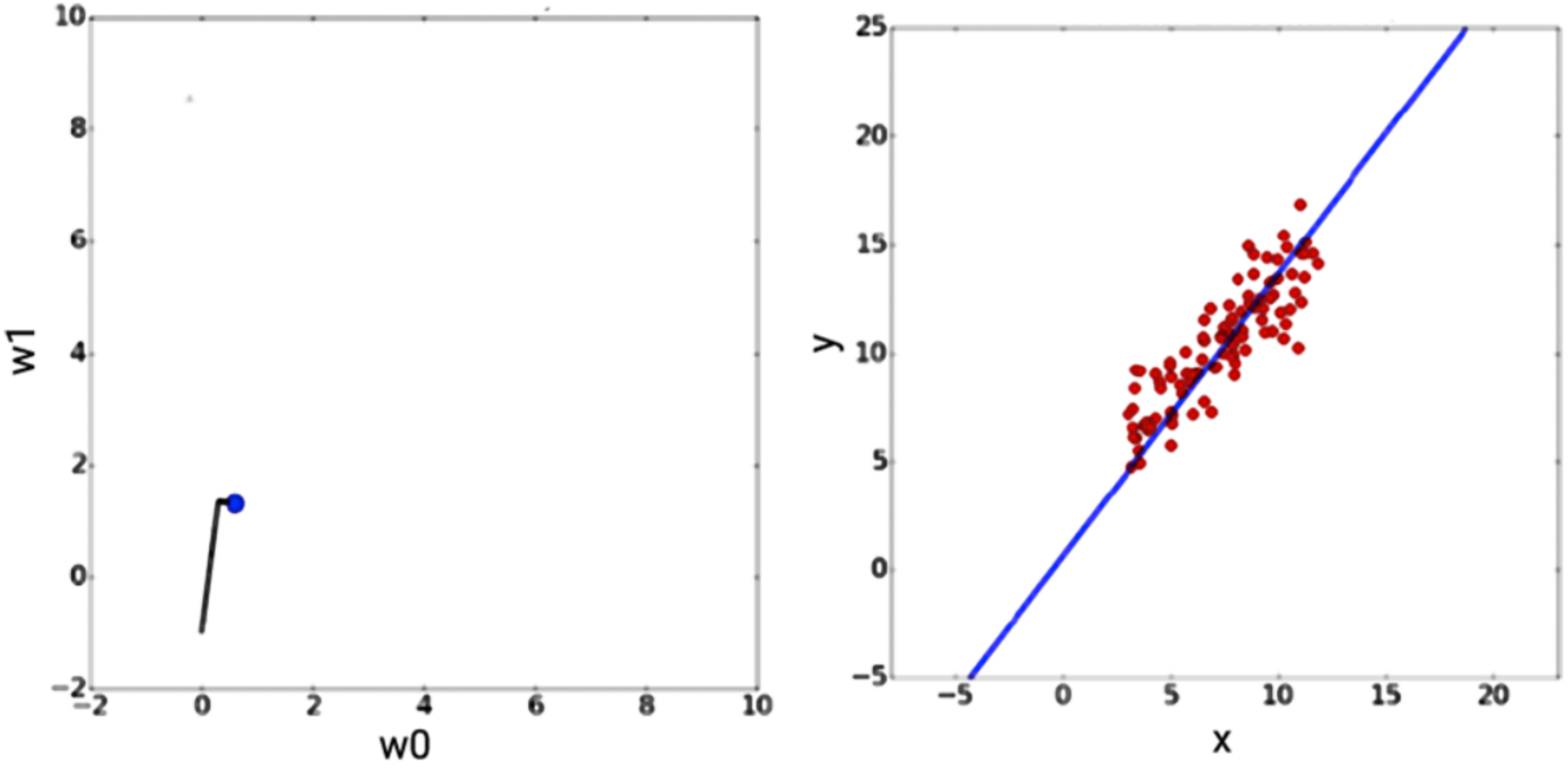
Парная регрессия – 2 итерация



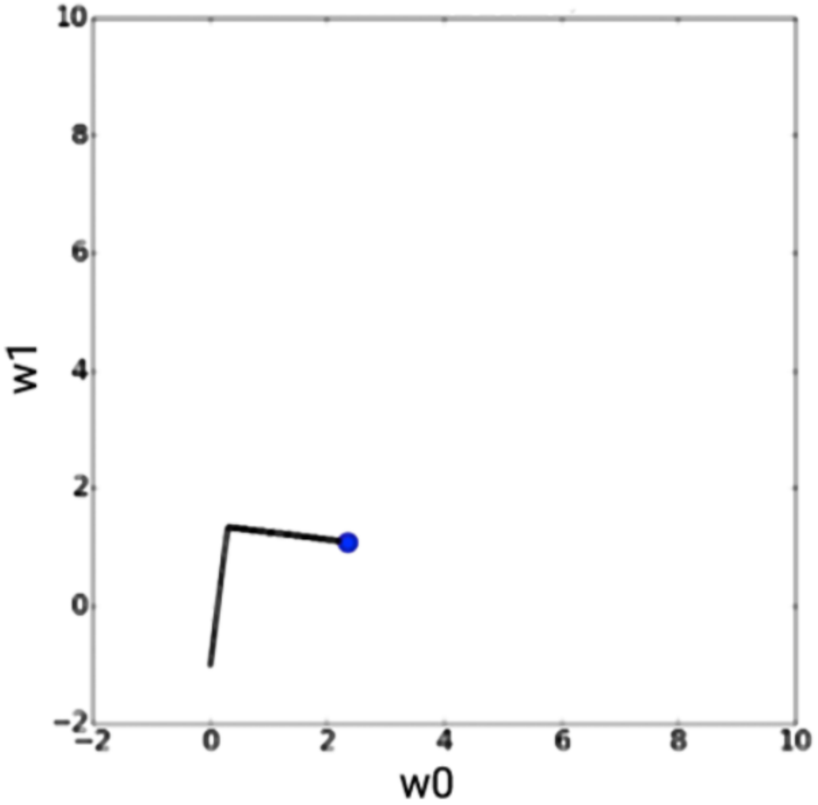
Парная регрессия – 3 итерация

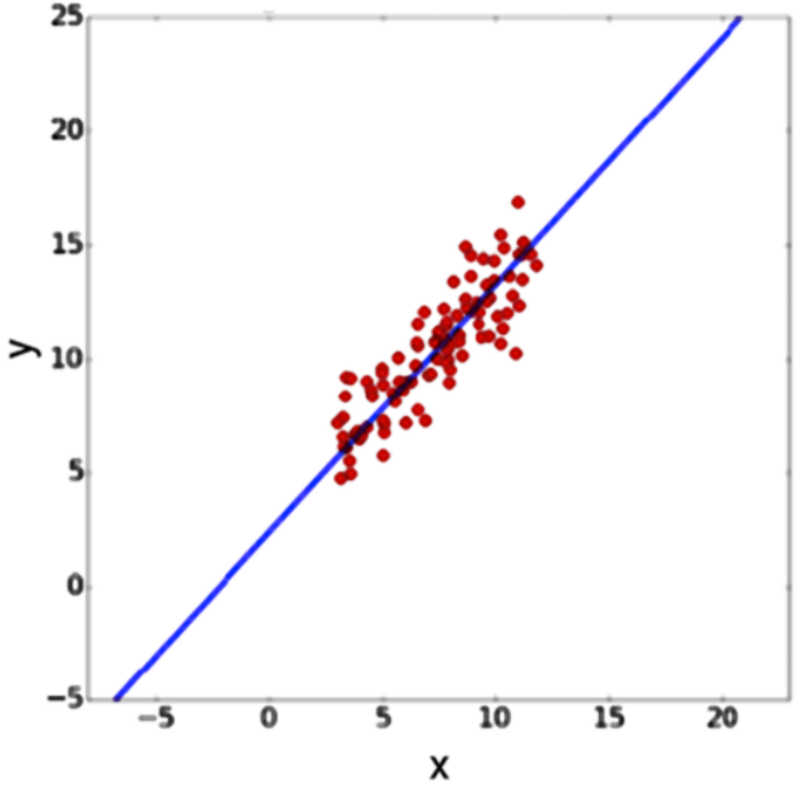


Парная регрессия – 4 – я итерация

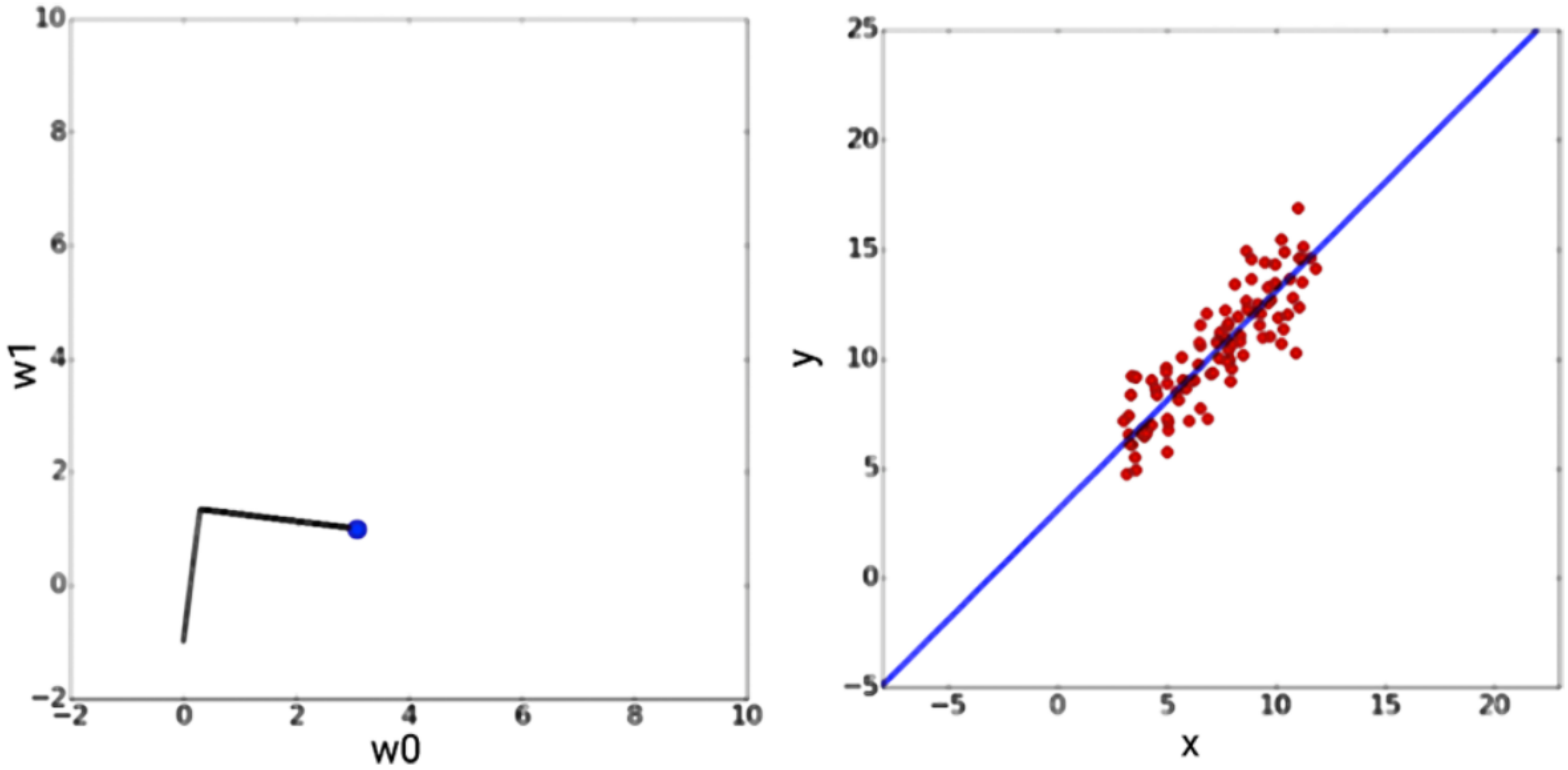


Парная регрессия – 5-я итерация

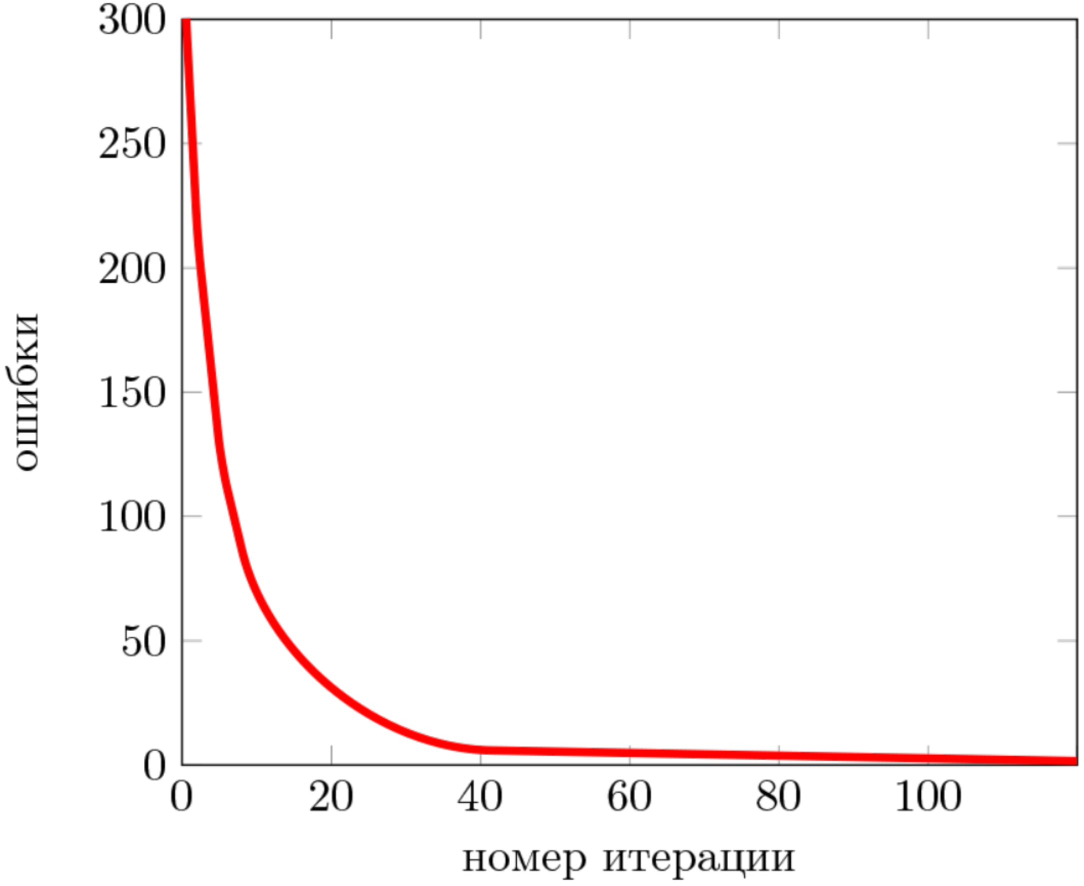




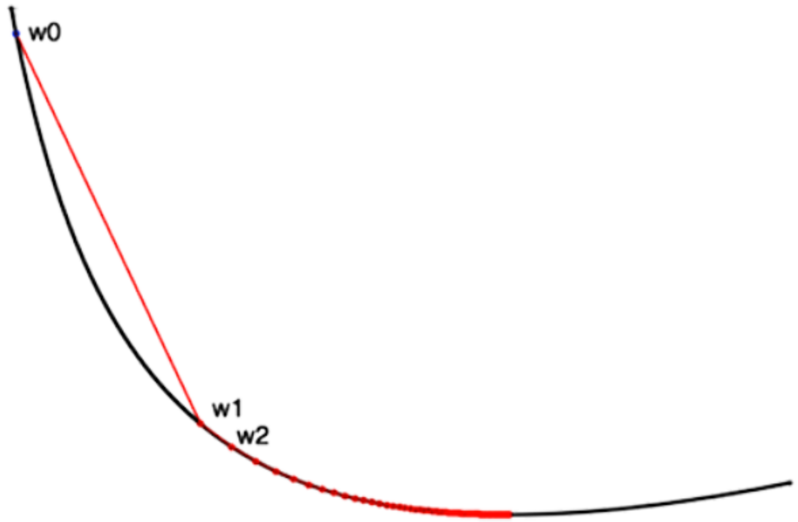
Парная регрессия – 6-я итерация



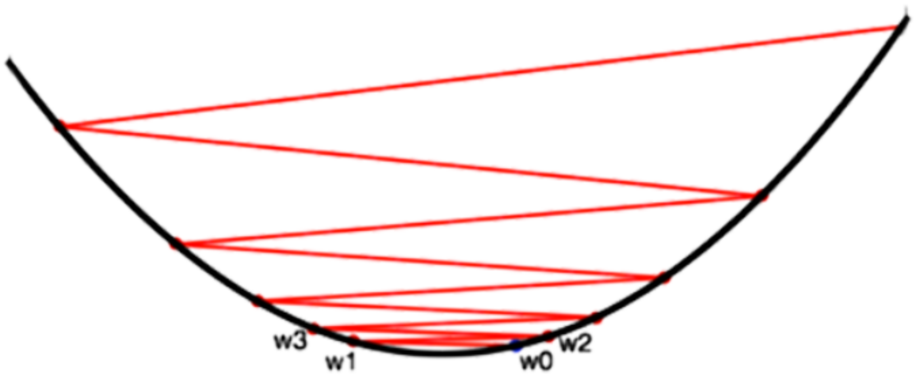
Функционал качества



Размер шага



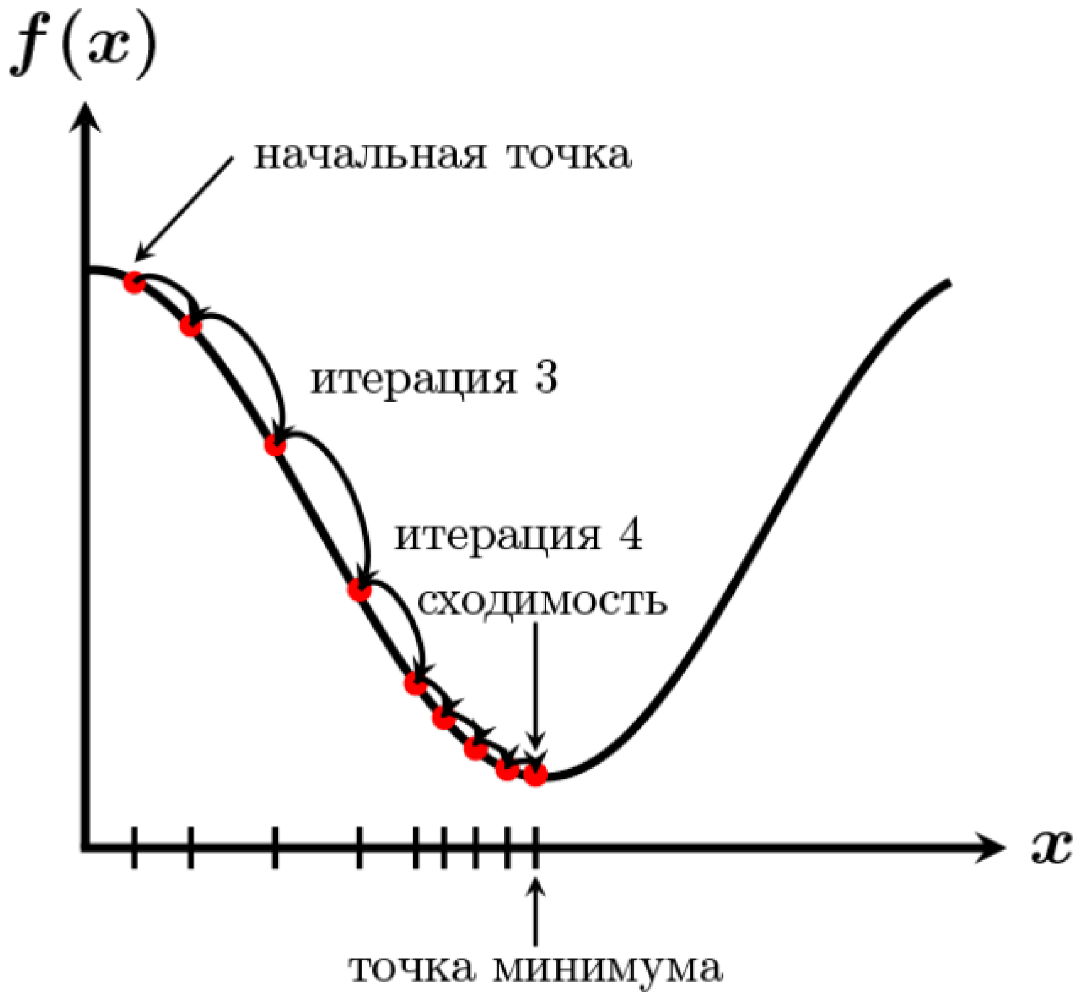
Маленький шаг



Большой шаг

Размер шага

- Размер шага, гиперпараметр, нужно подбирать
- Обычно пользуются эвристиками
- Чем ближе к минимуму, тем меньше надо шагать



Многомерная линейная регрессия

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

- Градиент:

$$\nabla_{\mathrm{w}} Q(\mathrm{w}, X) = \frac{2}{\ell} X^T (X\,\mathrm{w} - y)$$

Стохастический градиентный спуск

- Инициализация: $w^0 = 0$
- Цикл по t = 1, 2, 3, …:
- $w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, X)$
- Если $\|w^t - w^{t-1}\| < \varepsilon, \text{то завершить}$

Градиент функционала

$$\nabla_{\mathbf{w}} Q(\mathbf{w}, X) = \frac{2}{l} X^T (X\,\mathbf{w} - y)$$

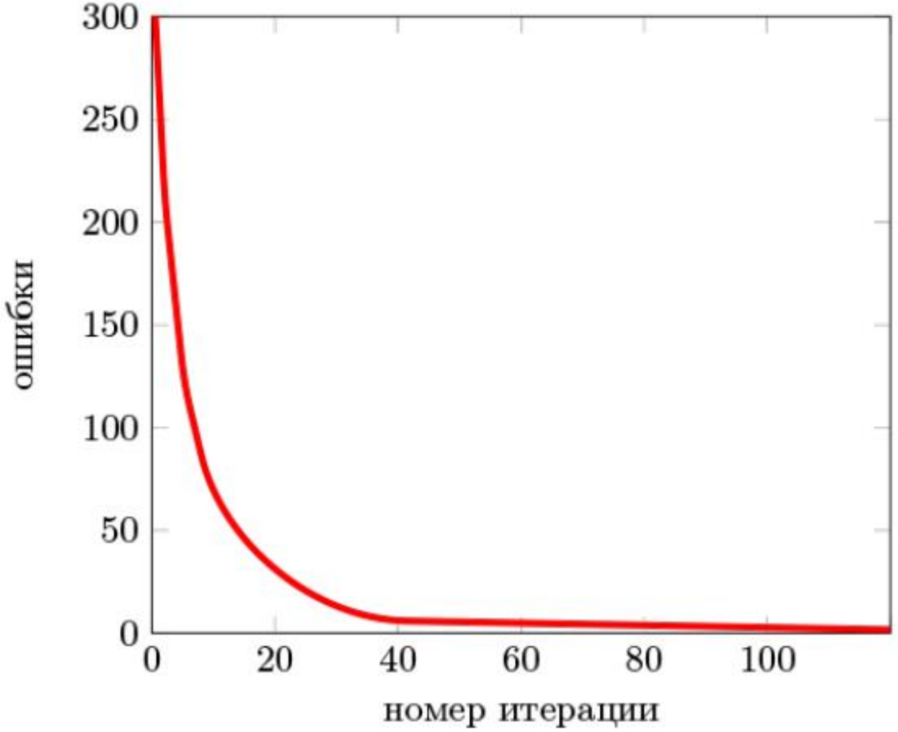
$$\frac{\partial Q}{\partial \mathbf{w}_j} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i^j (\langle \mathbf{w}, x_i \rangle - y_i)$$

Градиентный спуск требует вычисления полного градиента!

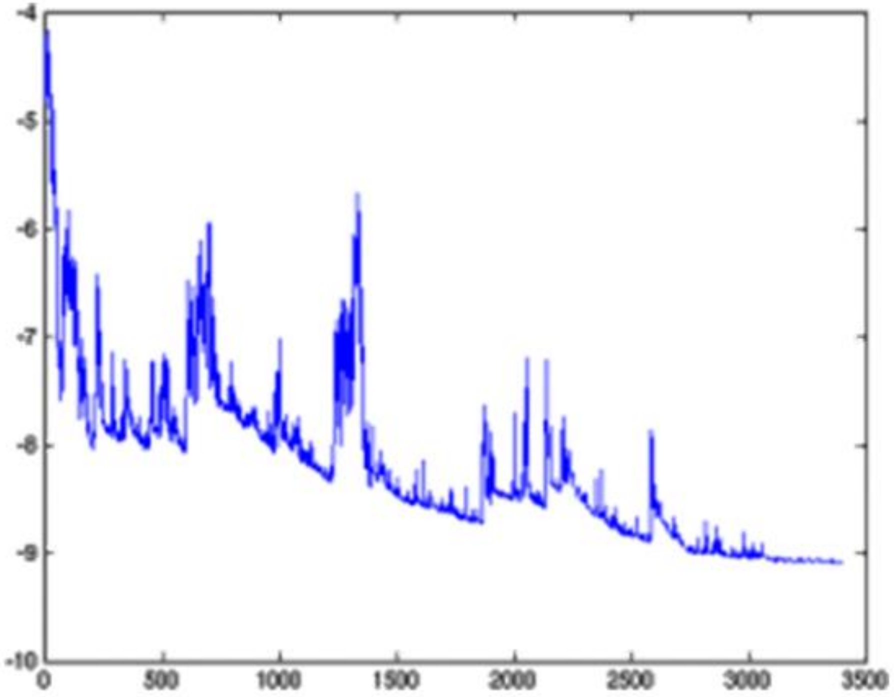
Стохастический градиентный спуск

- Инициализация: $w^0 = 0$
- Цикл по t = 1, 2, 3, …:
- Выбрать случайный объект x_i из X
- $w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, \{x_i\})$
- Если $\|w^t - w^{t-1}\| < \varepsilon, \text{то завершить}$

Стохастический градиентный спуск



Градиентный спуск



Стохастический градиентный спуск

Преимущества SGD

- Быстрее выполняется один шаг
- Не требует хранения выборки в памяти
- Подходит для онлайн обучения

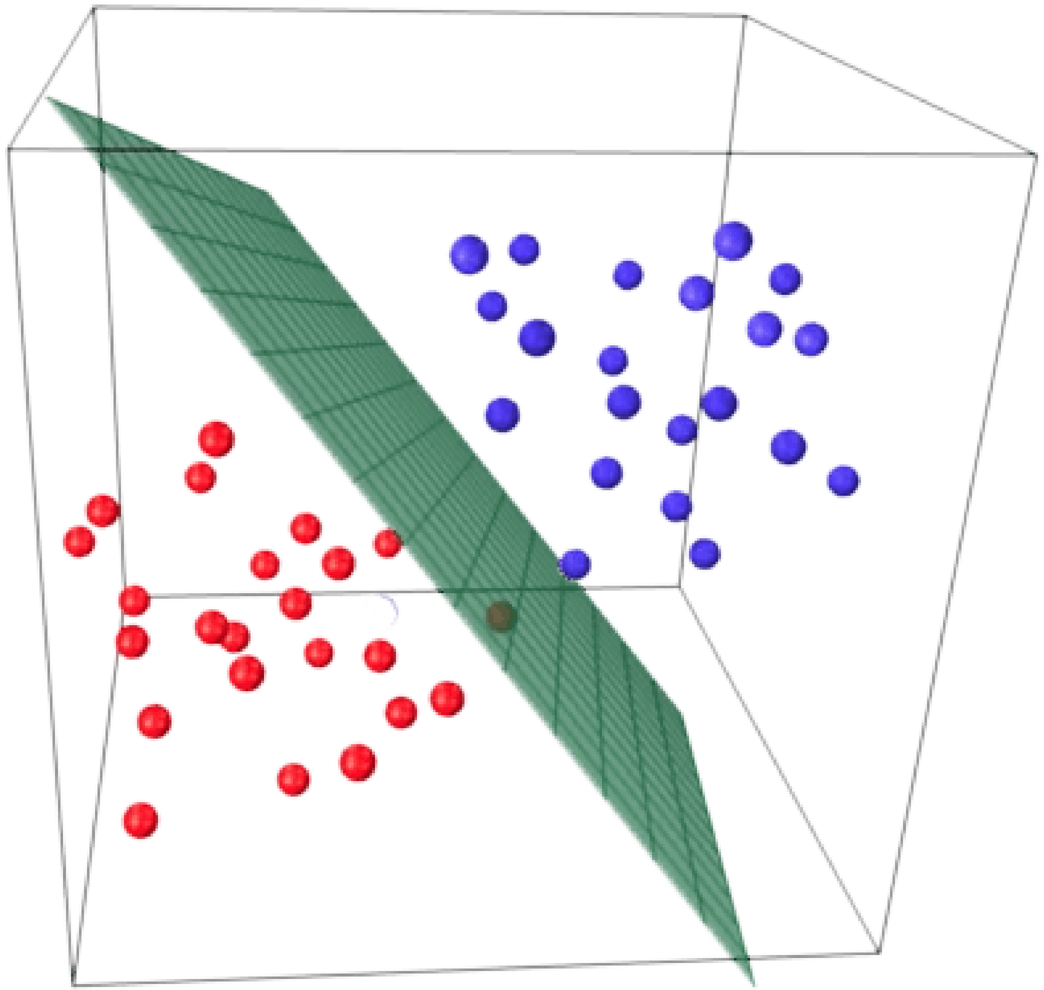
Метод максимального правдоподобия

- Интуитивная оценка вероятности 117/400 = 29 %
- Такая оценка является оценкой максимального правдоподобия
- Разберемся откуда берется данная оценка, вспомни Распределение Бернулли : случайная величина X имеет распределение Бернулли, если она принимает всего два значения (1 и 0 с вероятностями θ и 1 – θ) и имеет следующую функцию распределения вероятности :

$$p(\theta, x) = \theta^x (1-\theta)^{(1-x)}, \ x \in \{0, 1\}$$

Ноутбук Jupyter

Линейный классификатор



Линейный классификатор

$$\boldsymbol{a(x) = \mathbf{sign}\left(\mathbf{w}_0 + \sum_{j=1}^{d} \mathbf{w}_j\, x^j\right)}$$

Свободный коэффициент

Веса

Признаки

Линейный классификатор

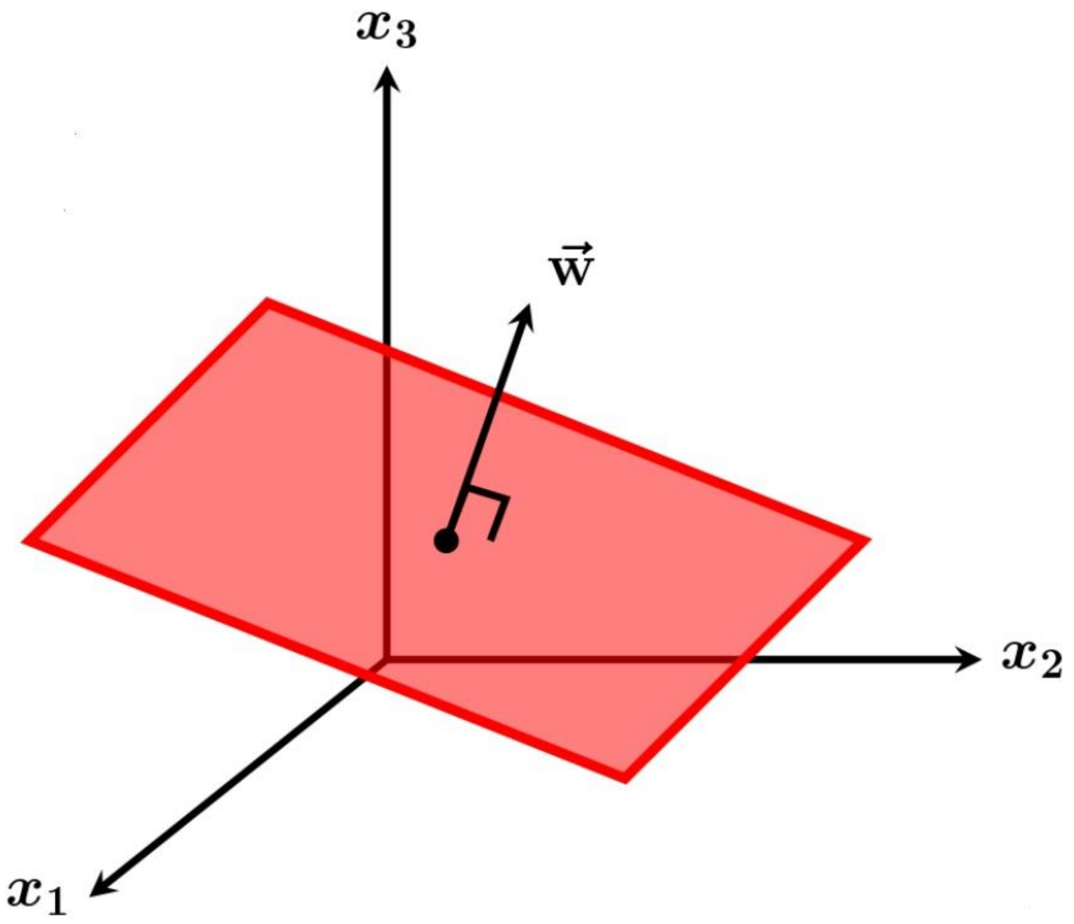
Добавим единичный признак:

$$a(x) = \operatorname{sign} \sum_{j=1}^{d+1} w_j\, x^j = \operatorname{sign}\langle w, x\rangle$$

Уравнение гиперплоскости

$$\boxed{\langle \mathbf{w}, x \rangle = 0}$$

Уравнение гиперплоскости



Расстояние от гиперплоскости

- Чтобы получить расстояние до гиперплоскости, нужно решить почти классическую задачу из курса линейной алгебры: найти расстояние от точки с радиус вектором X_A до плоскости, которая задается уравнением $w^T x = 0$

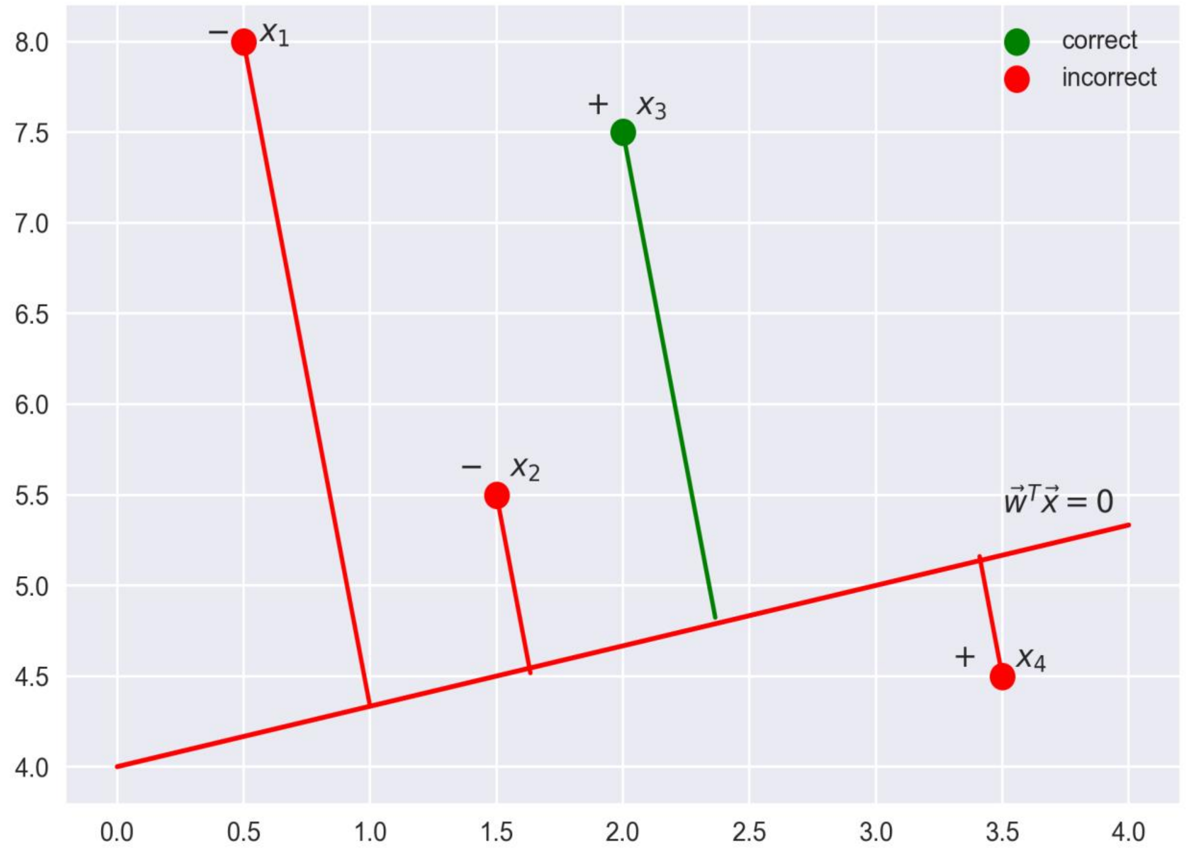
$$\rho(x_A, w^T x = 0) = \frac{w^T x_A}{\|w\|}$$

Когда получим ответ, то поймем, что чем больше по модулю выражение $w^T x_i$, тем дальше точка x_i находится от плоскости $w^T x = 0$

Отступ

$M(x_i) = y_i w^T x_i$ – уверенность модели в классификации объекта x_i

•если отступ большой (по модулю) и положительный, это значит, что метка класса поставлена правильно,На рисунке – x3.
•если отступ большой (по модулю) и отрицательный, значит метка класса поставлена неправильно (скорее всего такой объект – аномалия, например, его метка в обучающей выборке поставлена неправильно). На рисунке – x1.
•если отступ малый (по модулю), то объект находится близко к разделюящей гиперплоскости, а знак отступа определяет, правильно ли объект классифицирован. На рисунке – x2 и x4.



Логистическая регрессия

Логистическая регрессия является частным случаем линейного классификатора, но она обладает способностью оценивать вероятность принадлежности объекта x_i к классу '+'

$$p_+ = \mathrm{P}\left(y_i = 1 \mid \mathbf{x_i}, \mathbf{w}\right)$$

Бизнес предпосылки

Клиент	**Вероятность невозврата**
Mike	0.78
Jack	0.45
Larry	0.13
Kate	0.06
William	0.03
Jessica	0.02

Отказ

p*=0.15

Одобрение

Логистическая регрессия

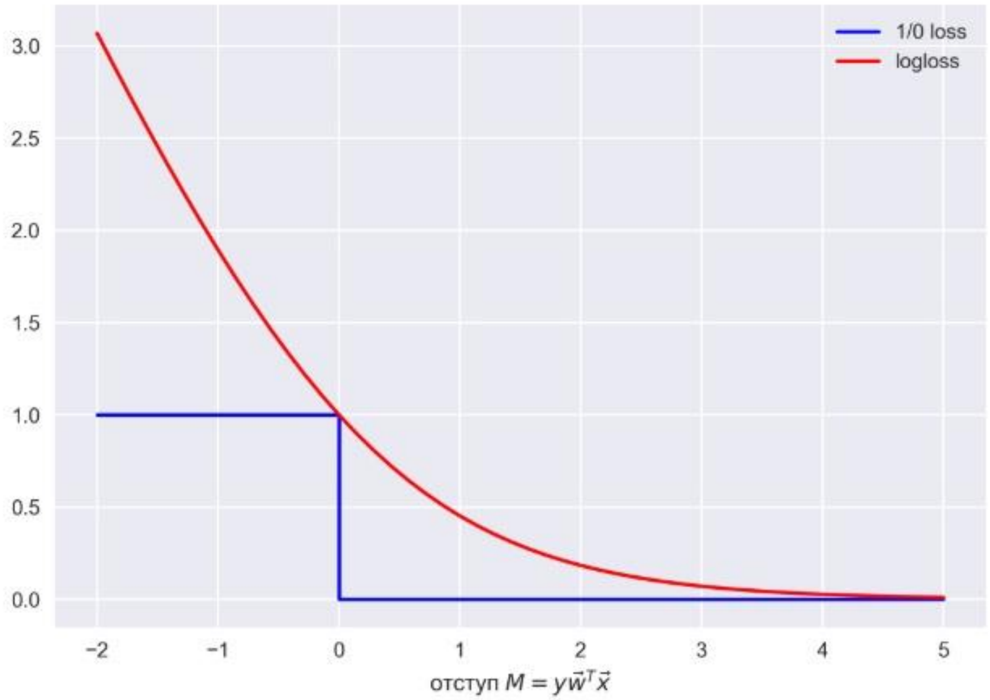
- Мы хотим прогнозировать вероятность $p_+ \in [0, 1]$
- Умеем строить линейный прогноз с помощью МНК: $a(x) = w^T x \in R$
- Каким образом преобразовать полученное значение в вероятность, пределы которы – [0, 1]
- Для этого нужна функция $f : \mathbb{R} \to [0, 1]$
- В логистической регрессии берется функция сигмоид
- $\sigma(z) = \frac{1}{1+e^{-z}}$

Логистическая регрессия

$$\mathcal{L}_{log}(\mathbf{X}, \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{\ell} \log(1 + \exp^{-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_{\mathrm{i}}}).$$

Верхняя оценка

- Посмотрим на новую функцию как функцию от отступа : $L(M) = \log(1 + e^{-M})$
- Нарисуем график 1/0 функций потерь (zero one loss), которая просто штрафует модель на 1 за ошибку на каждом объекте (отступ отрицательный) : $L_{1/0}(M) = [M < 0]$



Верхняя оценка

- Картинка отражает общую идею, что в задаче классификации, не умея напрямую минимизировать число ошибок (по крайней мере, градиентными методами это не сделать – производная 1/0 функциий потерь в нуле обращается в бесконечность), мы минимизируем некоторую ее верхнюю оценку.

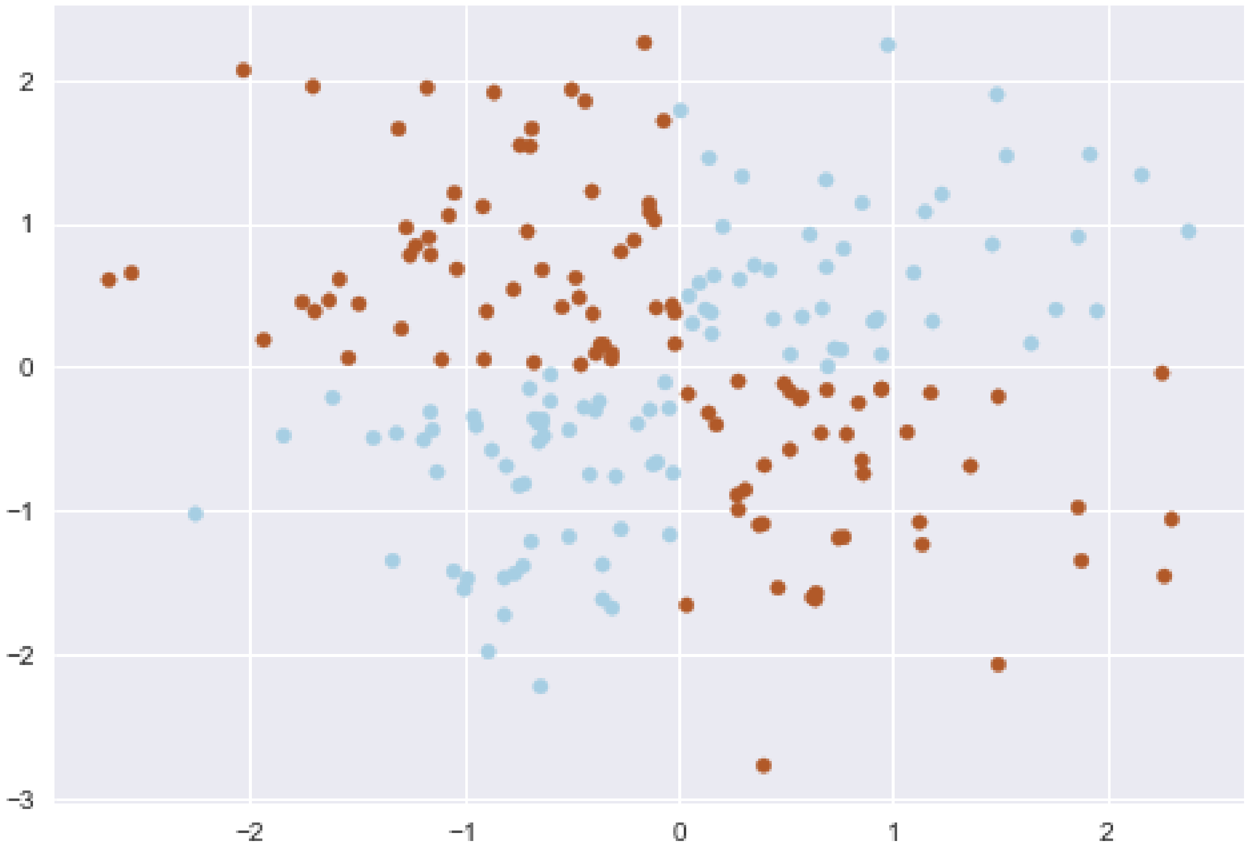
$$\mathcal{L}_{1/0}(\mathbf{X}, \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{\ell}[M(\mathbf{x}_\mathrm{i}) < 0] \leq \sum_{i=1}^{\ell} \log$$
$$(1 + \exp^{-y_i \mathbf{w}^\mathrm{T} \mathbf{x}_\mathrm{i}}) = \mathcal{L}_{\log}(\mathbf{X}, \mathbf{y}, \mathbf{w}),$$

Преимущества

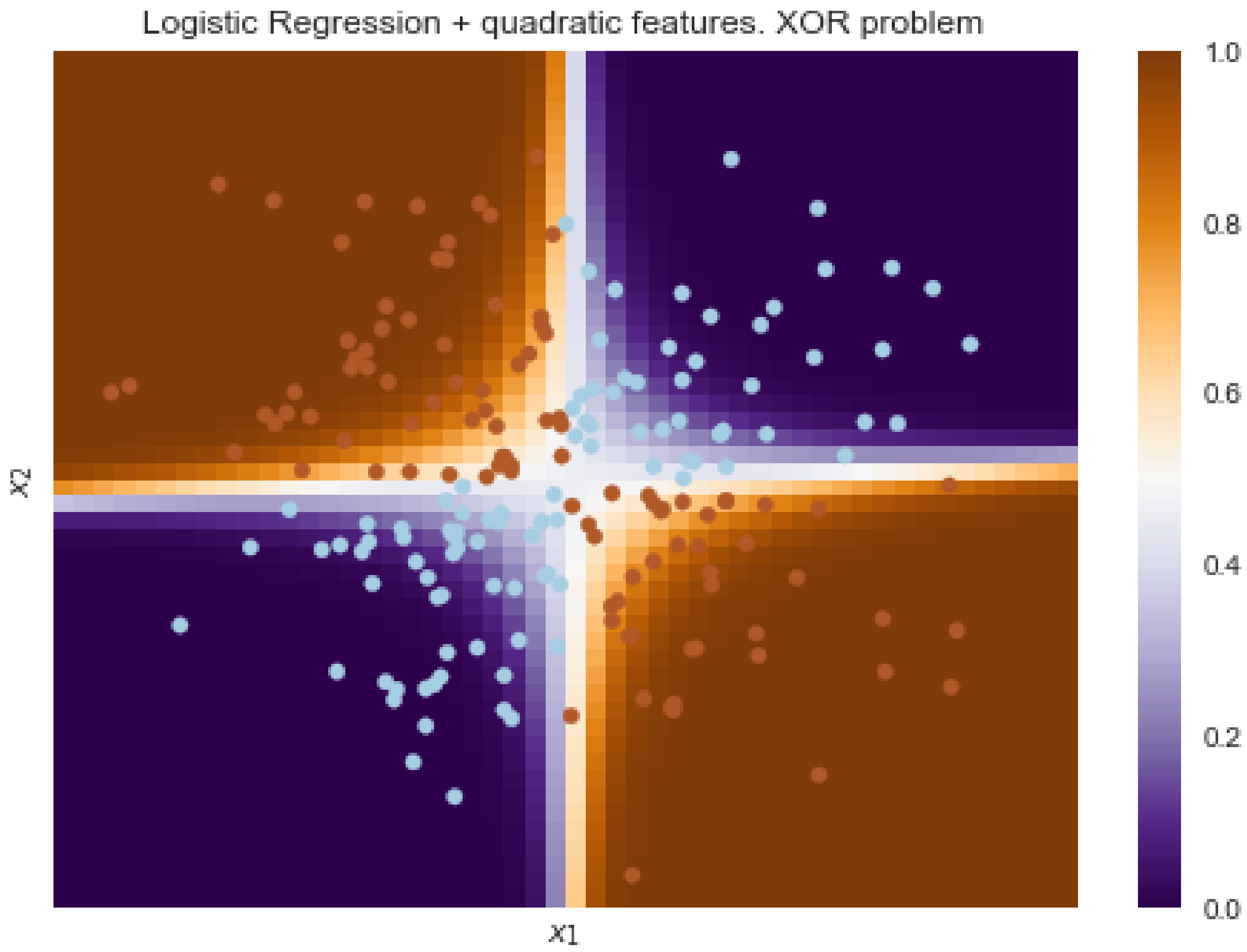
- Линейные модели быстро работают
- Мало параметров
- Хорошо применимы для больших данных

Недостатки: пример XOR проблема

XOR		Input 1	
		0	1
Input 2	0	0	1
	1	1	0



Полиномиальные признаки



Недостатки

- Линейные модели требуют тщательной обработки данных, для достижения хороших результатов
- Линейные модели чувствительны к выбросам и масштабированю
- В идеале, нужно, чтобы выборка соответствовала условиям теоремы Гаусса – Маркова, но это практически никогда не случается.

Спасибо за внимание !!!