ЛИНЕЙНЫЕ МОДЕЛИ В ЗАДАЧАХ РЕГРЕССИИ

ОБУЧЕНИЕ С УЧИТЕЛЕМ

- $\mathbb{X}$ — пространство объектов
- $\mathbb{Y}$ — пространство ответов
- $x = (x^1, \ldots, x^d)$ — признаковое описание
- $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка

- $a(x)$ — алгоритм, модель
- $Q(a, X)$ — функционал ошибки алгоритма a на выборке X
- Обучение: $a(x) = \underset{a \in \mathcal{A}}{\operatorname{argmin}}\, Q(a, X)$

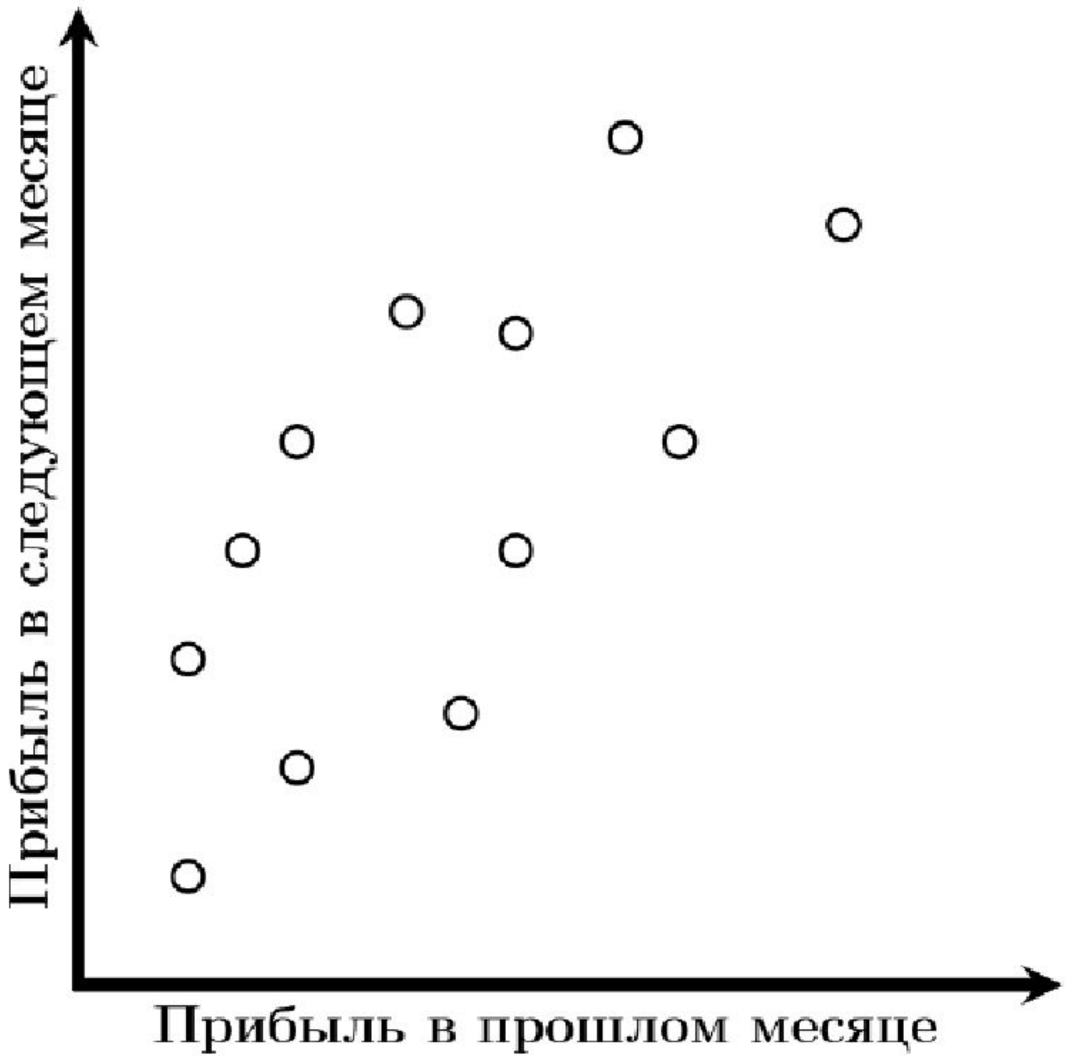
ОБУЧЕНИЕ С УЧИТЕЛЕМ

- Задача регрессии: $\mathbb{Y} = \mathbb{R}$
- Функционал ошибки?
- Семейство алгоритмов?
- Метод обучения?

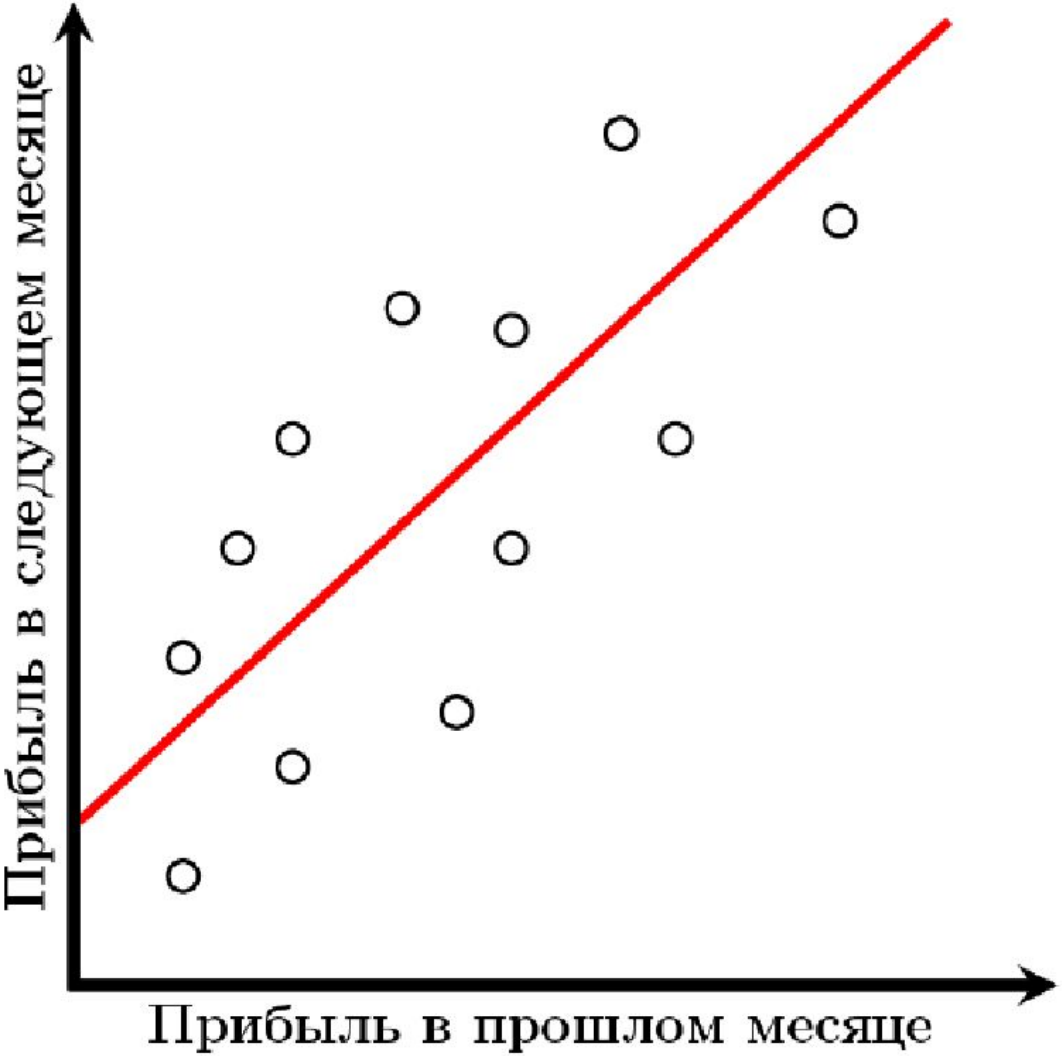
ЛИНЕЙНЫЕ МОДЕЛИ

- Задача регрессии: $\mathbb{Y} = \mathbb{R}$
- Пример: предсказание прибыли магазина

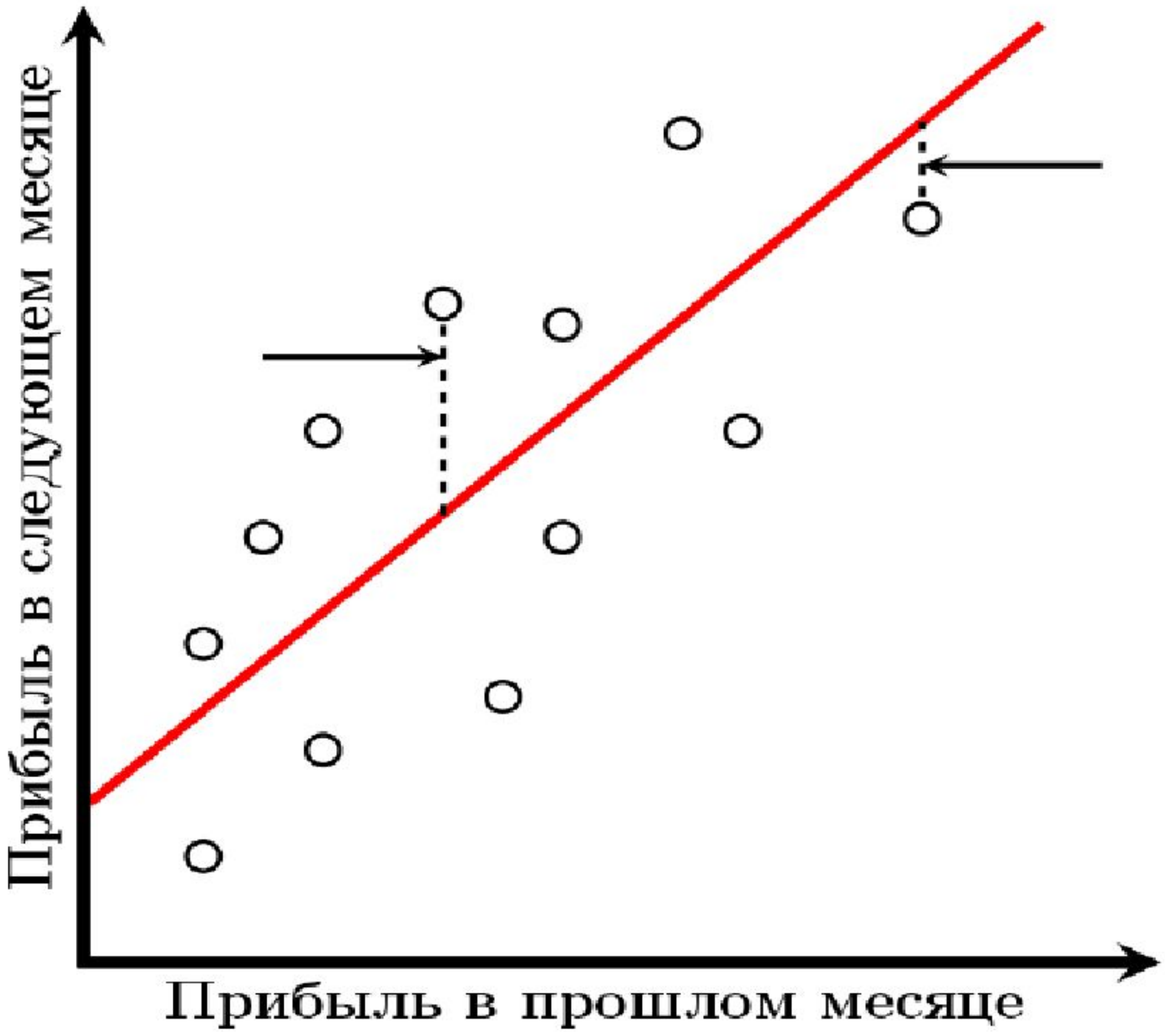
ЛИНЕЙНЫЕ МОДЕЛИ



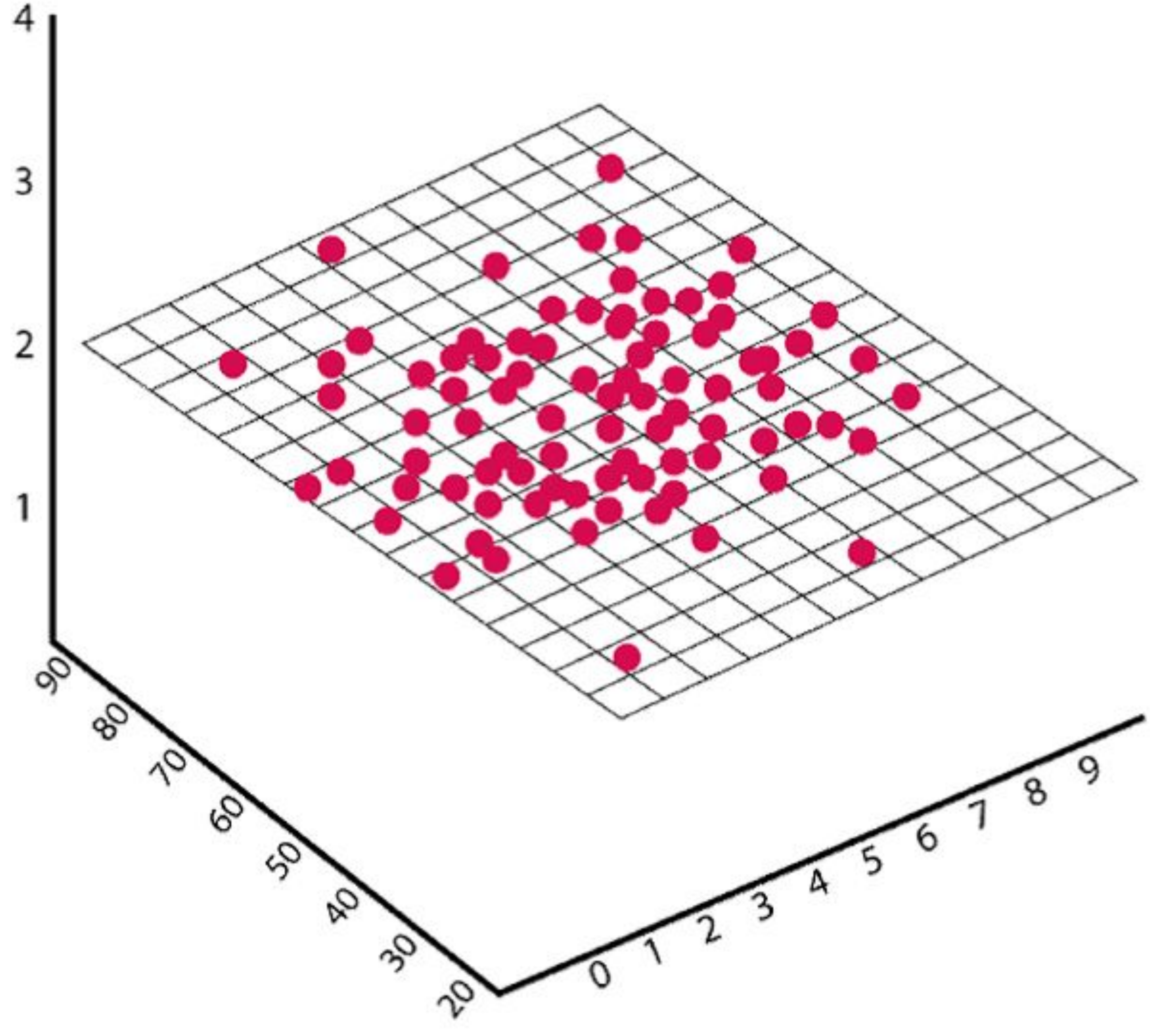
ЛИНЕЙНЫЕ МОДЕЛИ



ЛИНЕЙНЫЕ МОДЕЛИ



ЛИНЕЙНЫЕ МОДЕЛИ



ОБУЧЕНИЕ С УЧИТЕЛЕМ

- Задача регрессии: $\mathbb{Y} = \mathbb{R}$
- **Семейство алгоритмов?**
- Функционал ошибки?
- Метод обучения?

ЛИНЕЙНЫЕ МОДЕЛИ

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

Свободный коэффициент

Веса

Признаки

ЛИНЕЙНЫЕ МОДЕЛИ

- Добавим константный признак

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

Обучение с учителем

- Задача регрессии: $\mathbb{Y} = \mathbb{R}$
- Семейство алгоритмов?
- **Функционал ошибки?**
- Метод обучения?

Среднеквадратичная ошибка

- Отклонение прогноза: $a(x) - y$

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

Среднеквадратичная ошибка

- ~~Отклонение прогноза: $a(x) - y$~~
- Модуль отклонения: $|a(x) - y|$

Среднеквадратичная ошибка

- ~~Отклонение прогноза: $a(x) - y$~~
- ~~Модуль отклонения: $|a(x) - y|$~~
- Квадрат отклонения: $(a(x) - y)^2$

Среднеквадратичная ошибка

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

Среднеквадратичная ошибка

- Для линейной модели:

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

Вещественный вектор

Резюме

- Линейные алгоритмы для регрессии
- Среднеквадратичная ошибка

ОБУЧЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ

ОБУЧЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ

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

- d неизвестных
- Есть константный признак
- Выпуклая функция

Матричная запись

- Матрица «объекты-признаки»:

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

Матричная запись

- Матрица «объекты-признаки»:

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

Признак

Матричная запись

- Вектор ответов:

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

МАТРИЧНАЯ ЗАПИСЬ

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

Аналитическое решение

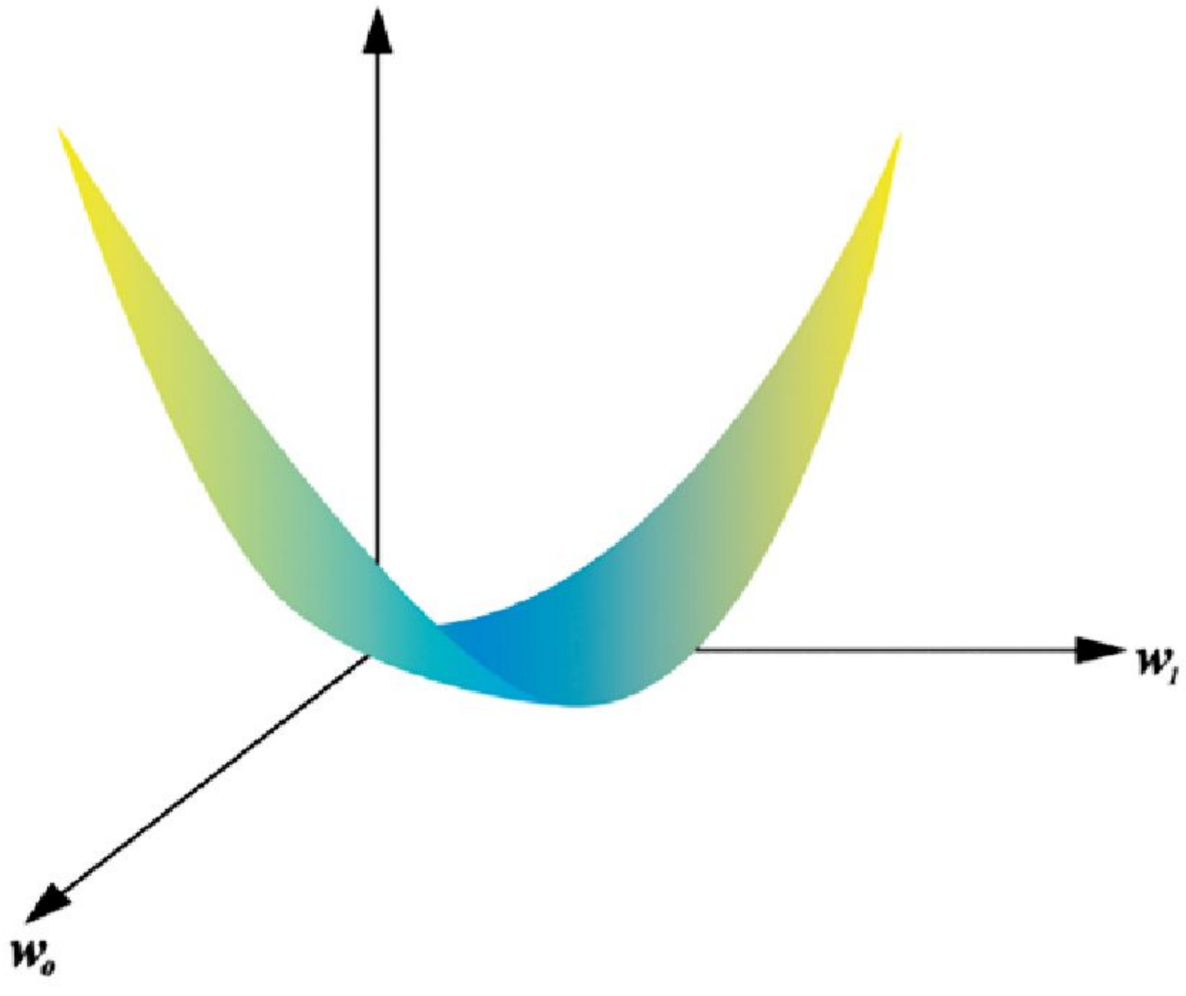
$$\mathbf{w}_* = (X^T X)^{-1} X^T y$$

- Нужно обращать матрицу $d \times d$ — сложность d^3

- Могут возникнуть численные проблемы

ГРАДИЕНТНЫЙ СПУСК

- Функция ошибки гладкая и выпуклая



ГРАДИЕНТНЫЙ СПУСК: НАПОМИНАНИЕ

- Инициализация: $\mathbf{w}^0 = 0$
- Цикл по $t = 1, 2, 3, \dots$:

$$\mathbf{w}^t = \mathbf{w}^{t-1} - \eta_t \nabla Q(\mathbf{w}^{t-1}, X)$$

Если $\| \mathbf{w}^t - \mathbf{w}^{t-1} \| < \varepsilon$, то завершить

Градиентный спуск: напоминание

- $\boxed{\text{Инициализация: } \mathbf{w}^0 = 0}$
- Цикл по $t = 1, 2, 3, \ldots$:

$$\mathbf{w}^t = \mathbf{w}^{t-1} - \eta_t \nabla Q(\mathbf{w}^{t-1}, X)$$

Если $\| \mathbf{w}^t - \mathbf{w}^{t-1} \| < \varepsilon$, то завершить

ГРАДИЕНТНЫЙ СПУСК: НАПОМИНАНИЕ

- Инициализация: $\mathbf{w}^0 = 0$
- Цикл по $t = 1, 2, 3, \ldots$:

$$\boxed{\mathbf{w}^t = \mathbf{w}^{t-1} - \eta_t \nabla Q(\mathbf{w}^{t-1}, X)}$$

Если $\| \mathbf{w}^t - \mathbf{w}^{t-1} \| < \varepsilon$, то завершить

Градиентный спуск: напоминание

- Инициализация: $\mathbf{w}^0 = 0$
- Цикл по $t = 1, 2, 3, \dots$:

$$\mathbf{w}^t = \mathbf{w}^{t-1} - \eta_t \nabla Q(\mathbf{w}^{t-1}, X)$$

Если $\| \mathbf{w}^t - \mathbf{w}^{t-1} \| < \varepsilon$, то завершить

Резюме

- Матричная запись функционала линейной регрессии
- Аналитическое решение
- Градиентный спуск

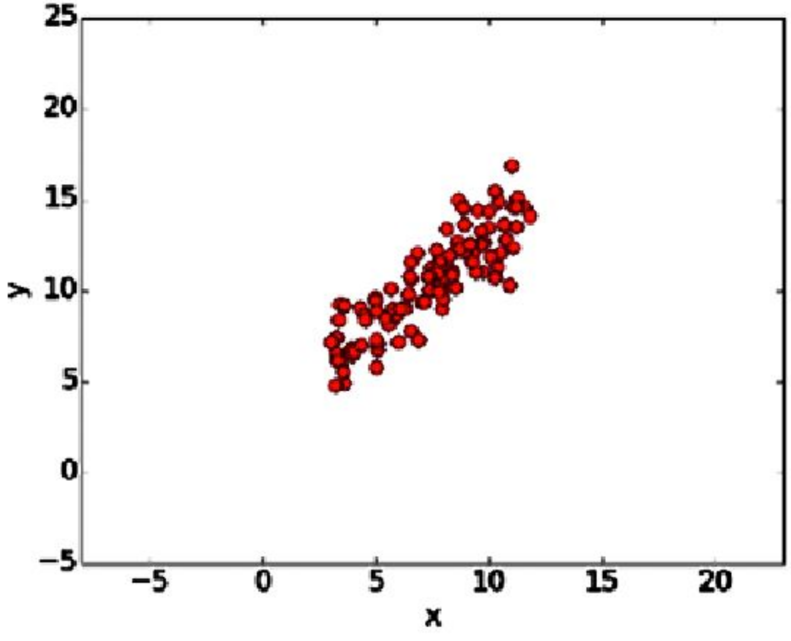
Градиентный спуск для линейной регрессии

ПАРНАЯ РЕГРЕССИЯ

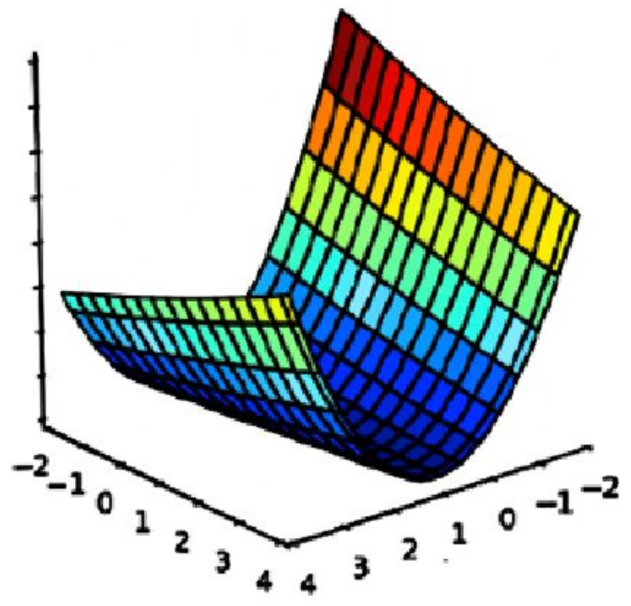
- Простейший случай: один признак
- Модель: $a(x) = \mathrm{w}_1 x + \mathrm{w}_0$
- Два параметра: w_1 и w_0
- Функционал:

$$Q(\mathrm{w}_0, \mathrm{w}_1, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\mathrm{w}_1 x_i + \mathrm{w}_0 - y_i)^2$$

ПАРНАЯ РЕГРЕССИЯ



Выборка



Функционал качества

ГРАДИЕНТНЫЙ СПУСК

- Инициализация: $\mathbf{w}^0 = 0$
- Цикл по $t = 1, 2, 3, \ldots$:

  $$\mathbf{w}^t = \mathbf{w}^{t-1} - \eta_t \nabla Q(\mathbf{w}^{t-1}, X)$$

  Если $\| \mathbf{w}^t - \mathbf{w}^{t-1} \| < \varepsilon$, то завершить

Градиент для парной регрессии

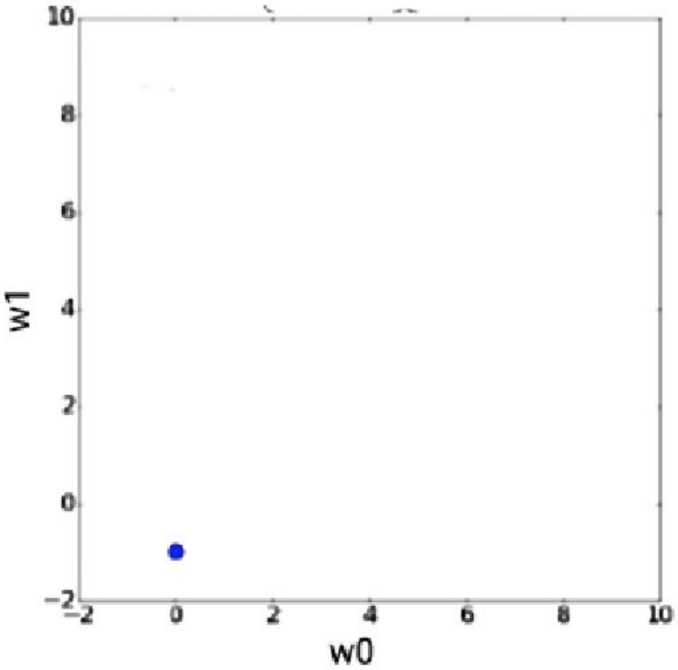
$$Q(\mathbf{w}_0, \mathbf{w}_1, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\mathbf{w}_1 x_i + \mathbf{w}_0 - y_i)^2$$

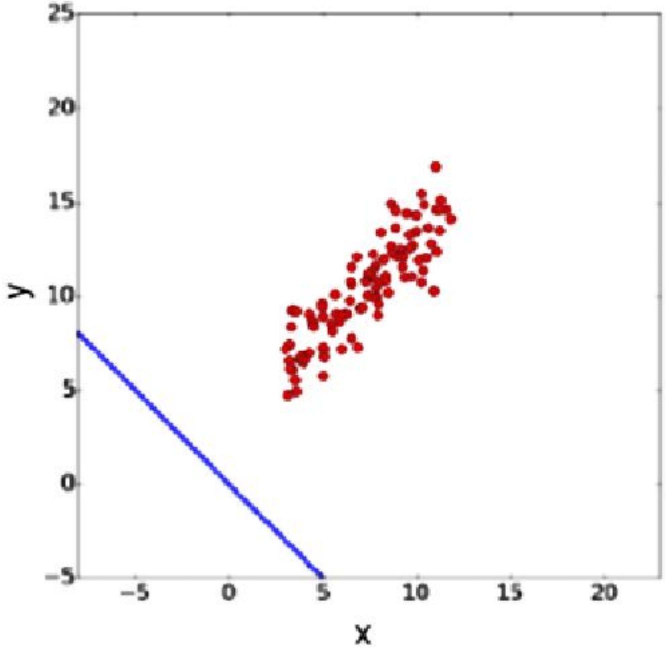
- Частные производные:

$$\frac{\partial Q}{\partial \mathbf{w}_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} (\mathbf{w}_1 x_i + \mathbf{w}_0 - y_i) x_i$$

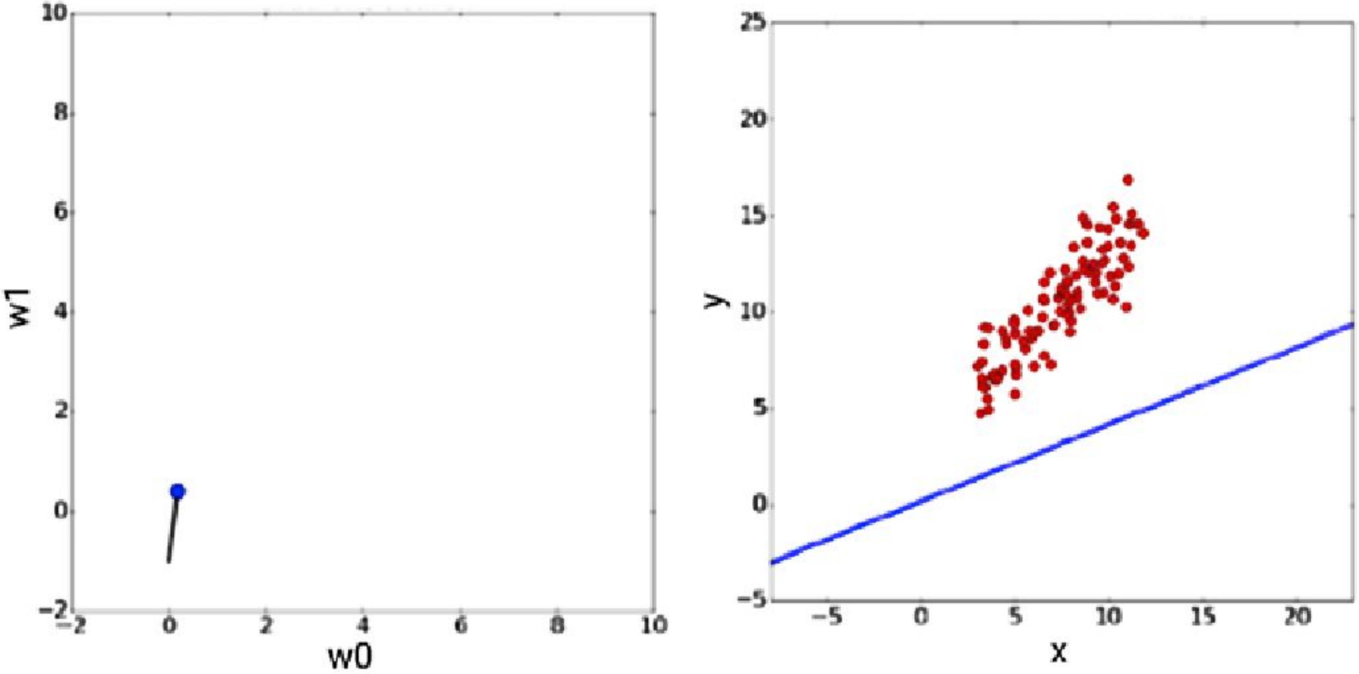
$$\frac{\partial Q}{\partial \mathbf{w}_0} = \frac{2}{\ell} \sum_{i=1}^{\ell} (\mathbf{w}_1 x_i + \mathbf{w}_0 - y_i)$$

ПАРНАЯ РЕГРЕССИЯ

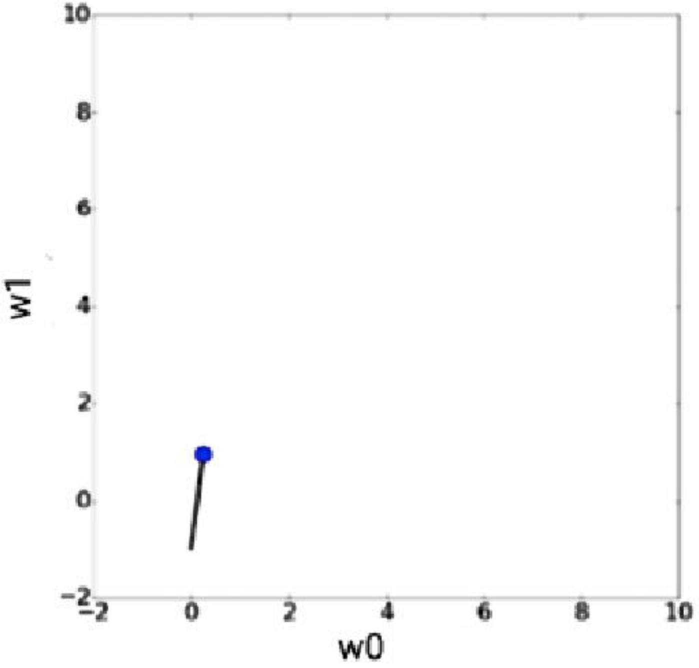


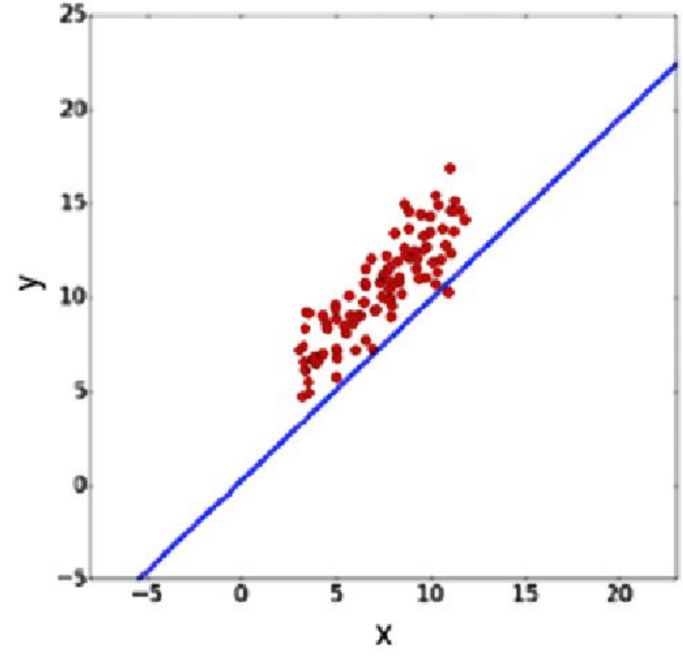


ПАРНАЯ РЕГРЕССИЯ

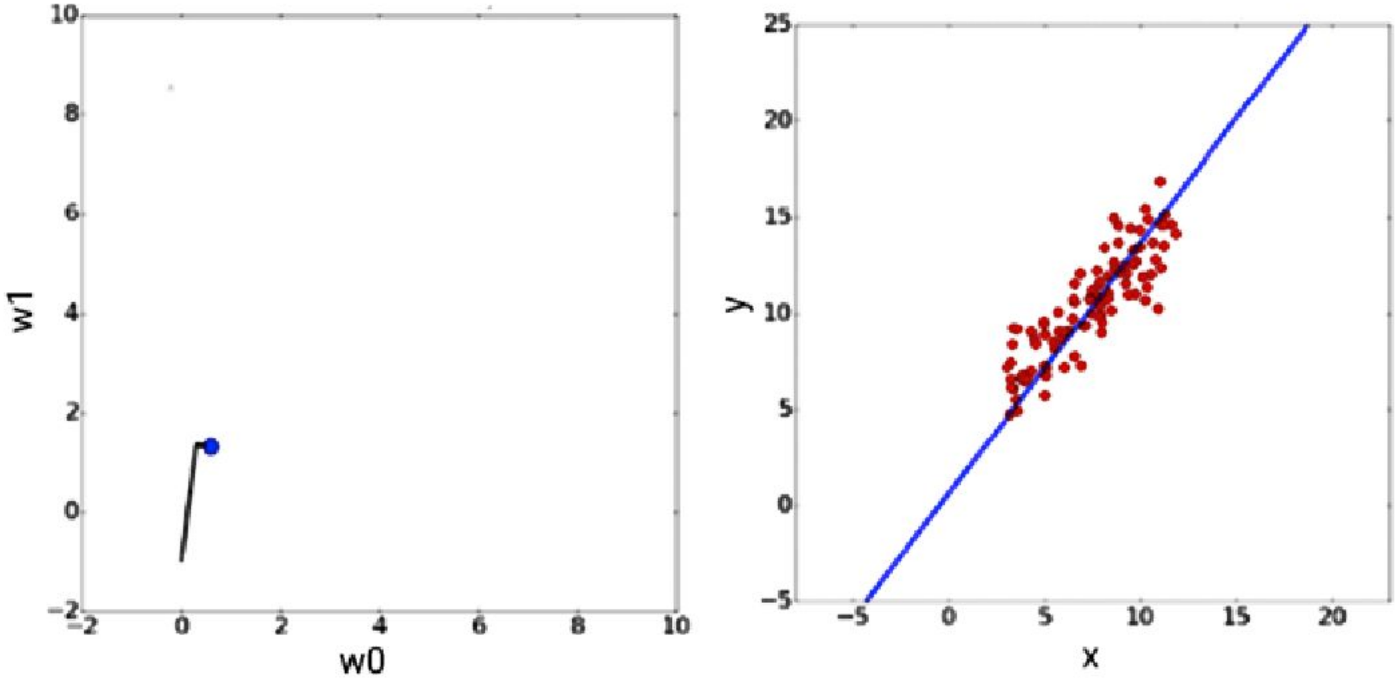


ПАРНАЯ РЕГРЕССИЯ

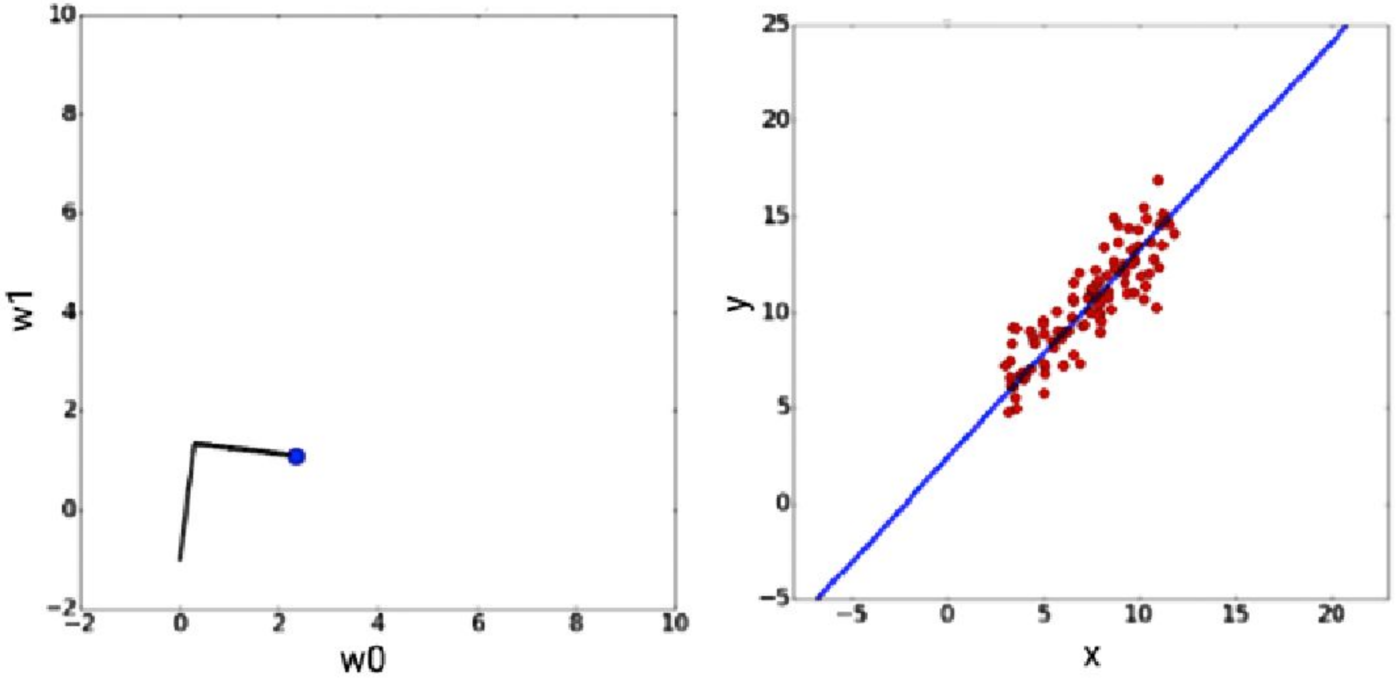




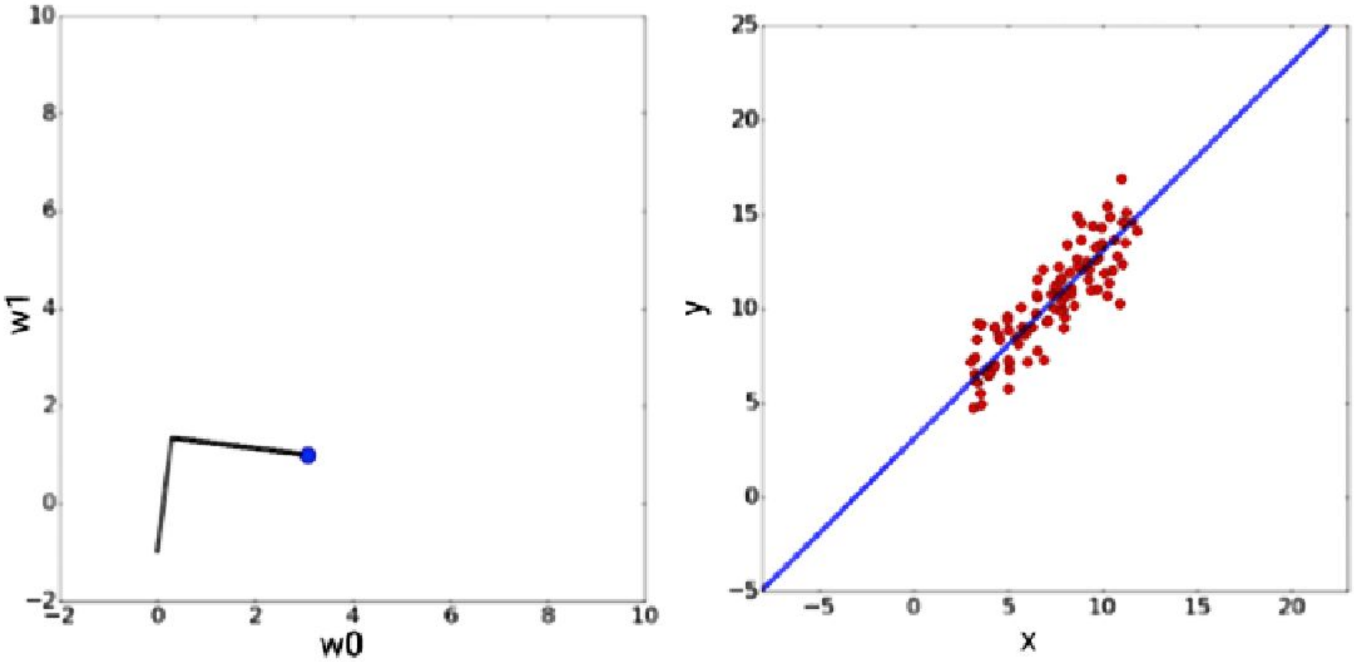
ПАРНАЯ РЕГРЕССИЯ



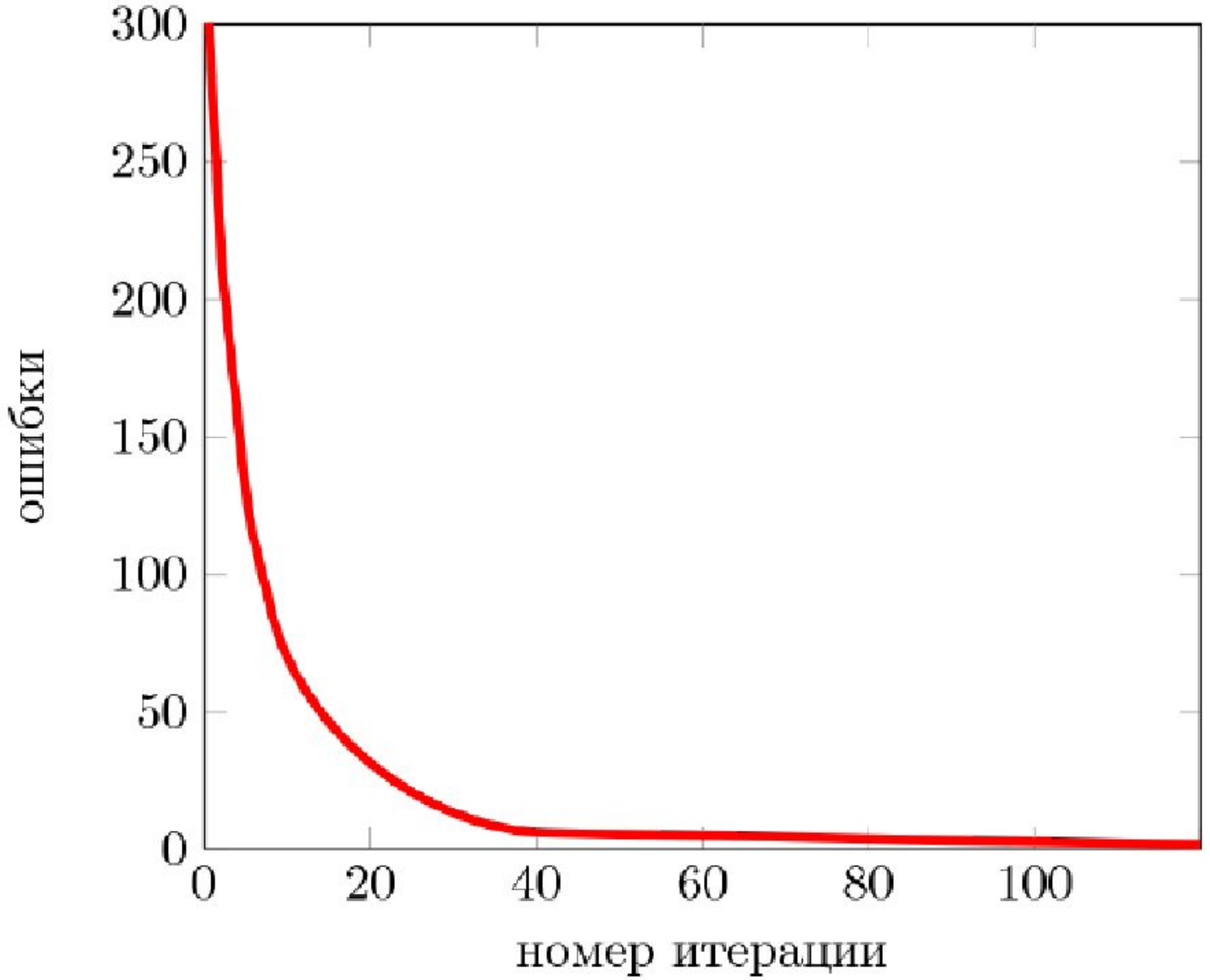
ПАРНАЯ РЕГРЕССИЯ



ПАРНАЯ РЕГРЕССИЯ

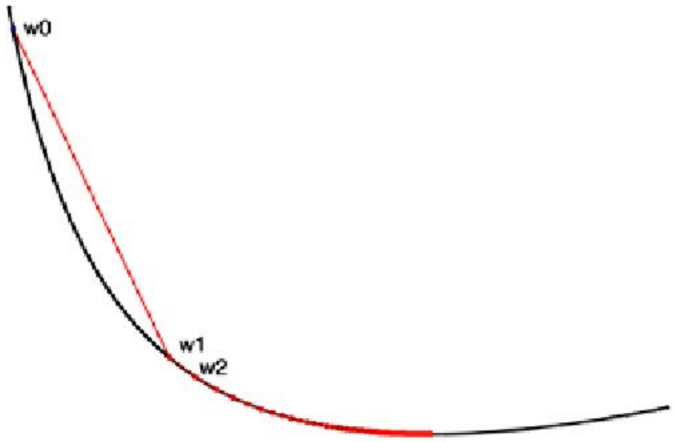


ФУНКЦИОНАЛ КАЧЕСТВА

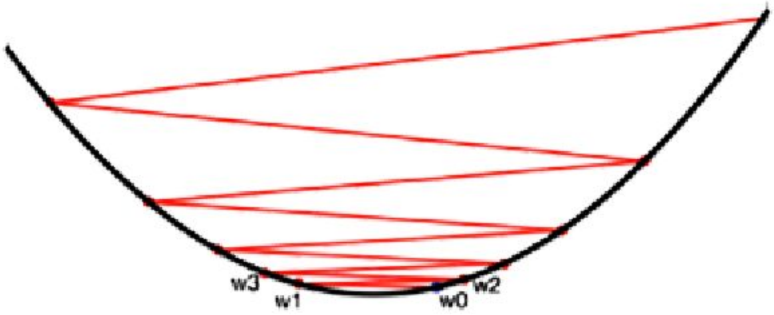


РАЗМЕР ШАГА

- Выбор размера шага η — искусство



Маленький шаг



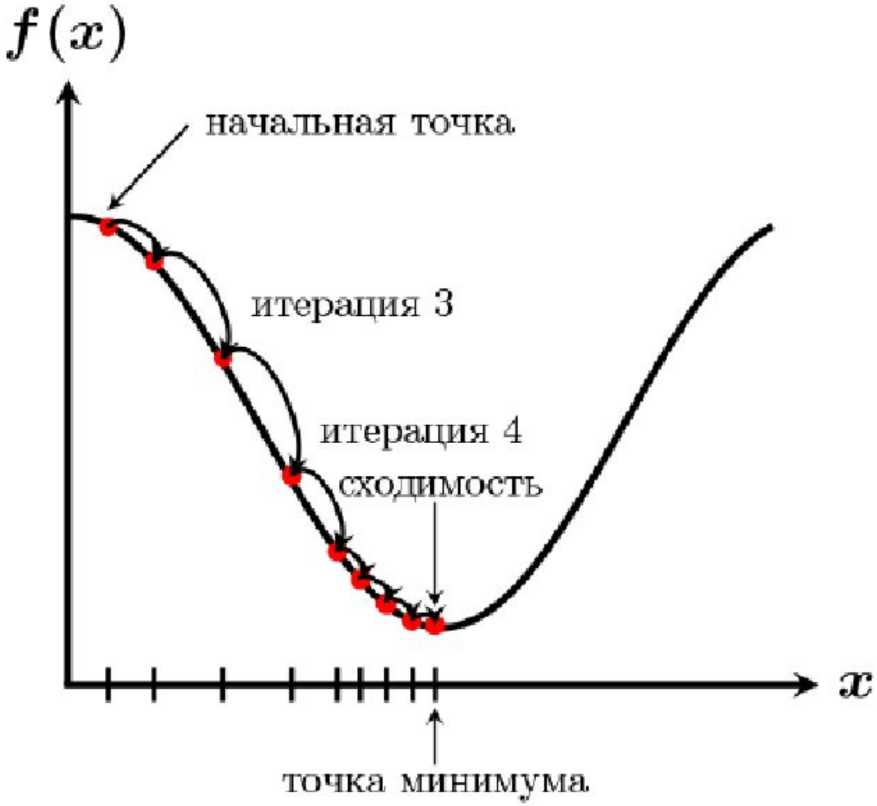
Большой шаг

РАЗМЕР ШАГА

- Неплохо работает: $\eta_t = \frac{k}{t}$
- k — константа, надо подбирать

РАЗМЕР ШАГА

- Обычно пользуются эвристиками
- Чем ближе к минимуму, тем меньше надо шагать



Многомерная линейная регрессия

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

- Градиент:

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

РЕЗЮМЕ

- Градиентный спуск для одномерной и многомерной линейной регрессии
- Важность выбора шага

Стохастический градиентный спуск

Градиентный спуск

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

Градиент функционала

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

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

Градиент функционала

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

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

Суммирование по всей выборке!

ГРАДИЕНТ ФУНКЦИОНАЛА

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

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

Как поменять веса, чтобы улучшить качество на объекте x_i

Градиент функционала

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

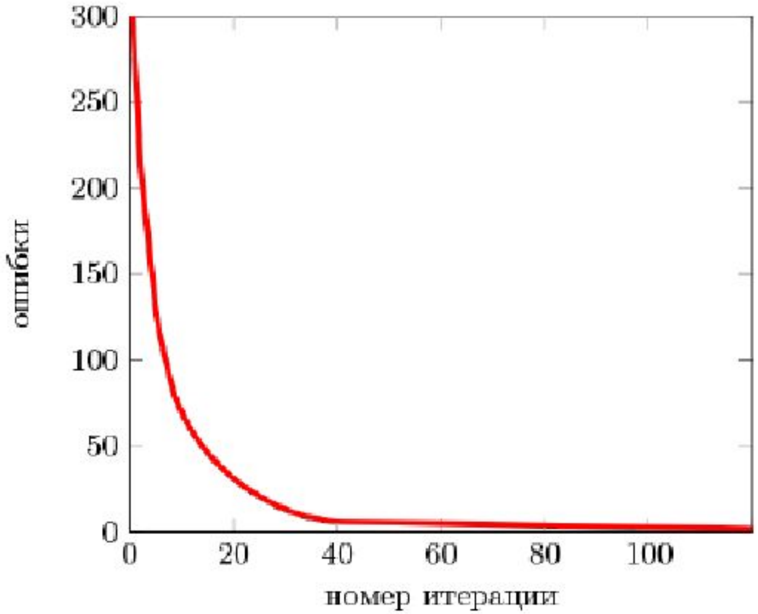
$$\frac{\partial Q}{\partial \mathbf{w}_j} = \boxed{\frac{2}{\ell} \sum_{i=1}^{\ell} x_i^j (\langle \mathbf{w}, x_i \rangle - y_i)}$$

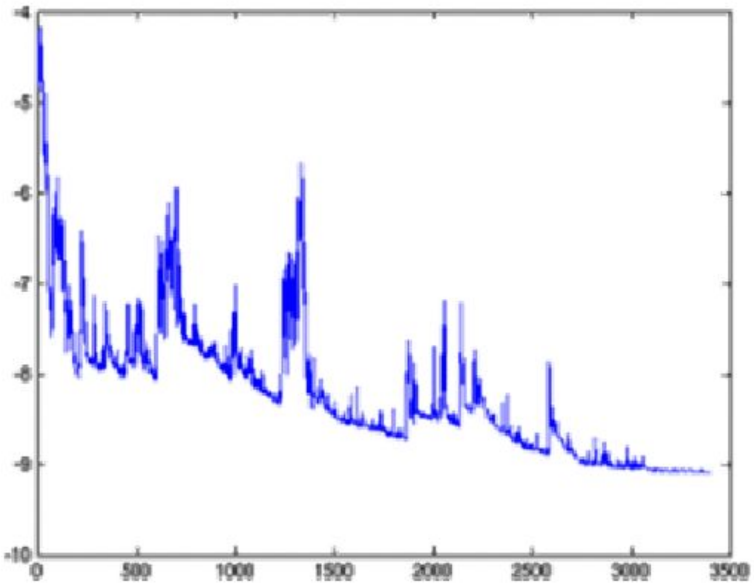
Как поменять веса, чтобы
улучшить качество на всей выборке

Стохастический градиентный спуск

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

Резюме

- Градиентный спуск требует вычисления полного градиента

- Стохастический градиентный спуск использует лишь один объект

- SGD позволяет обучать алгоритм на больших выборках

ЛИНЕЙНАЯ КЛАССИФИКАЦИЯ

ОБУЧЕНИЕ С УЧИТЕЛЕМ

- Задача бинарной классификации: $\mathbb{Y} = \{-1, +1\}$
- Функционал ошибки?
- Семейство алгоритмов?
- Метод обучения?

Линейный классификатор

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

Свободный коэффициент

Веса

Признаки

ЛИНЕЙНЫЙ КЛАССИФИКАТОР

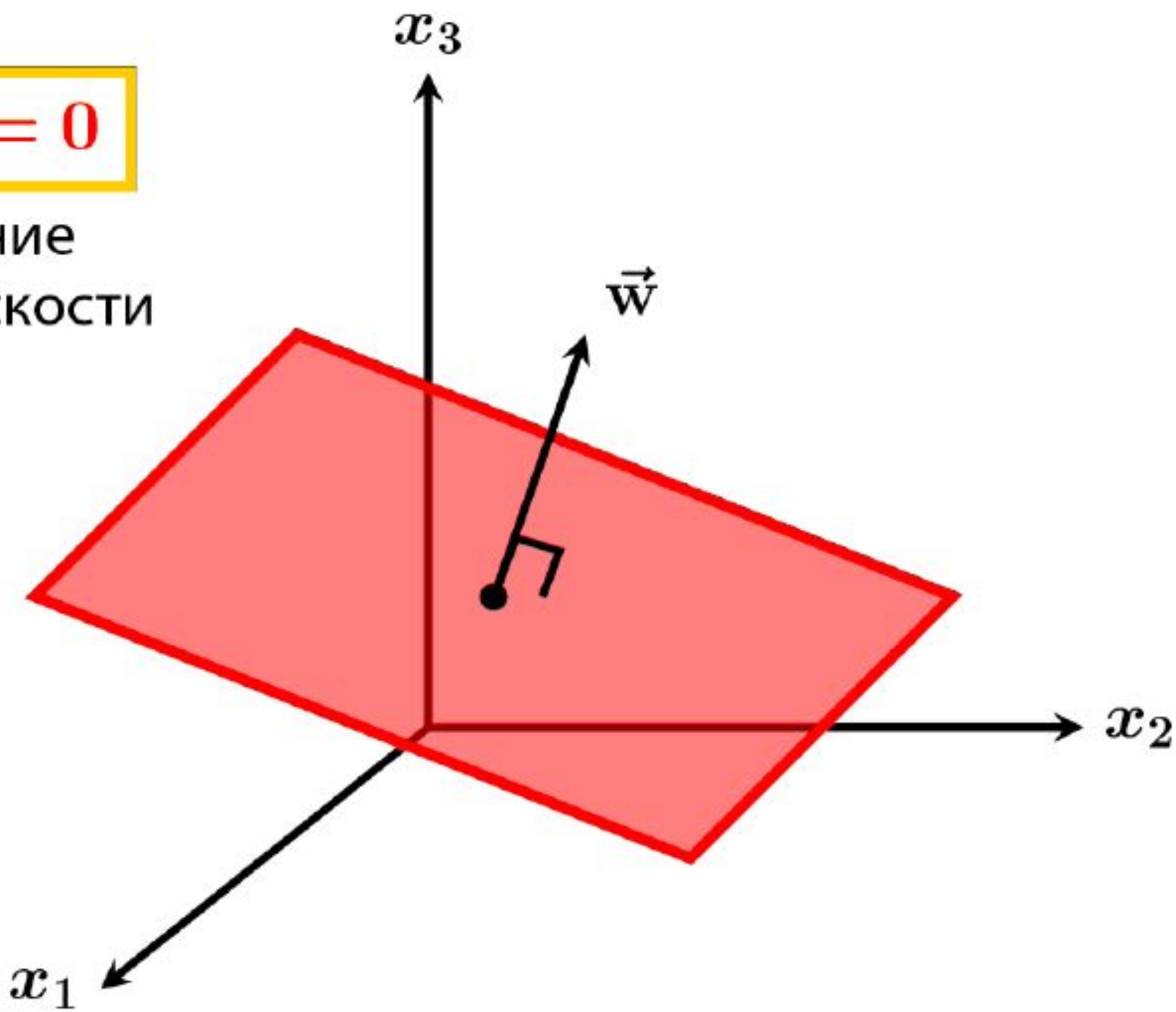
- Добавим единичный признак:

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

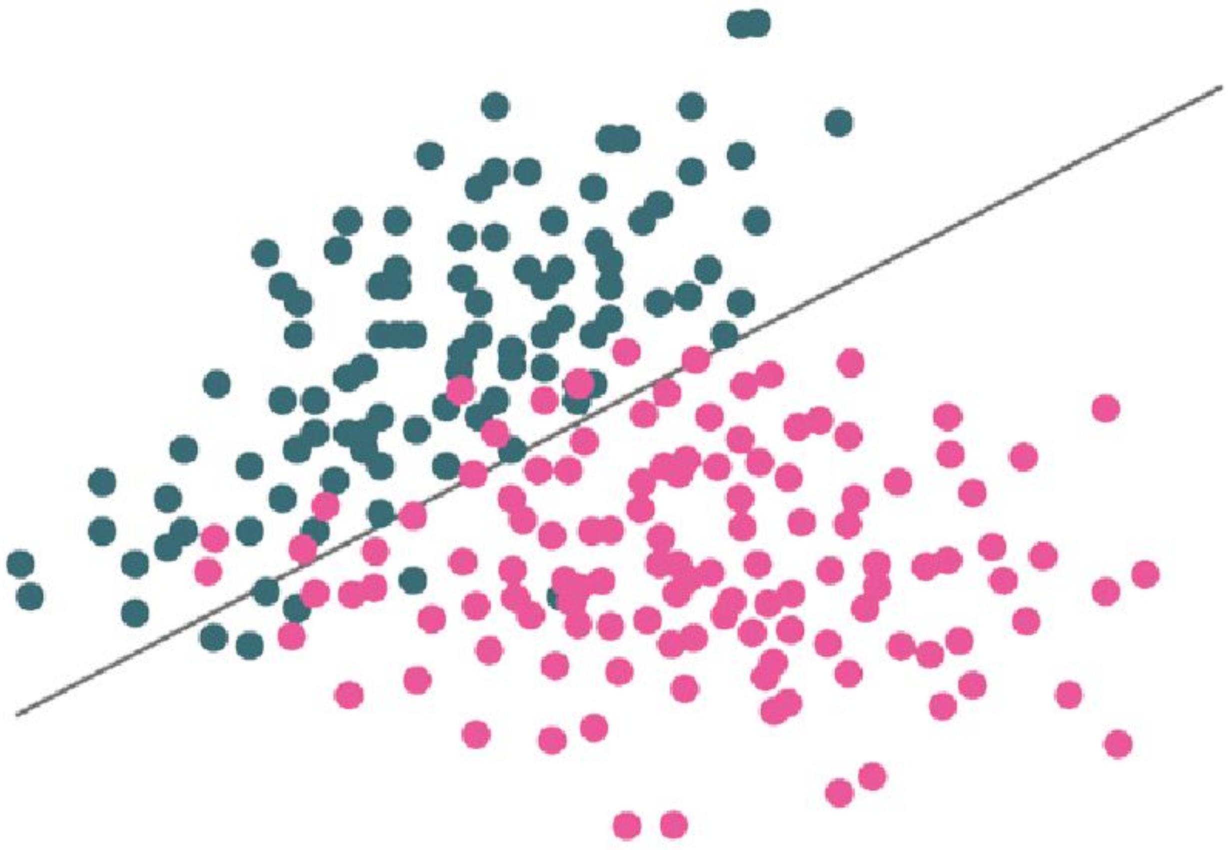
ГЕОМЕТРИЯ ЛИНЕЙНОГО КЛАССИФИКАТОРА

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

Уравнение гиперплоскости

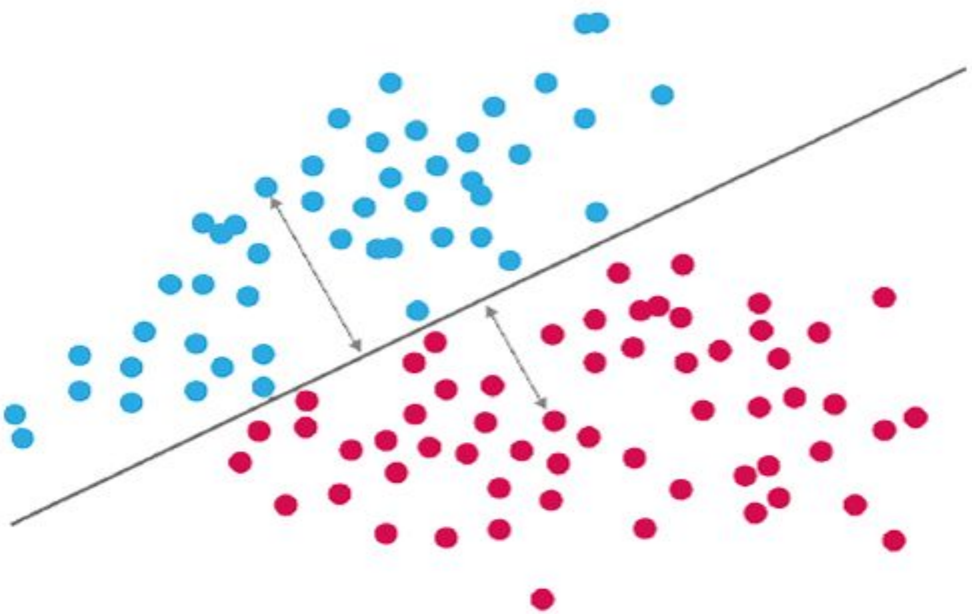


ГЕОМЕТРИЯ ЛИНЕЙНОГО КЛАССИФИКАТОРА



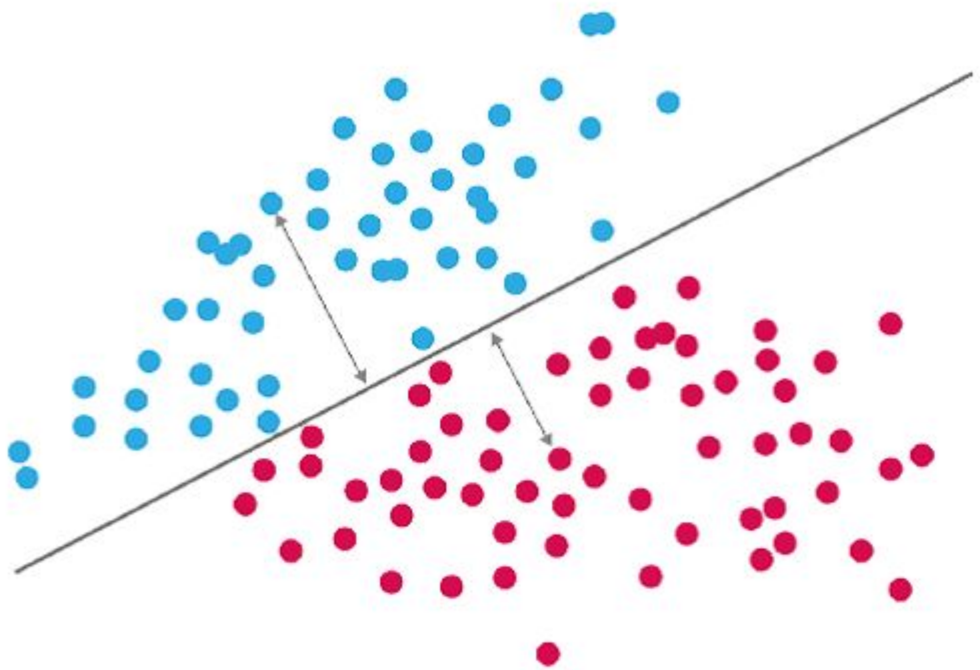
ГЕОМЕТРИЯ ЛИНЕЙНОГО КЛАССИФИКАТОРА

- Расстояние от точки до гиперплоскости $\langle \mathbf{w}, x \rangle = 0$: $$\frac{|\langle \mathbf{w}, x \rangle|}{\| \mathbf{w} \|}$$
- Чем больше $\langle \mathbf{w}, x \rangle$, тем дальше объект от разделяющей гиперплоскости



Отступ

- $M_i = y_i \langle \mathrm{w}, x_i \rangle$
- $M_i > 0$ — классификатор даёт верный ответ
- $M_i < 0$ — классификатор ошибается
- Чем дальше отступ от нуля, тем больше уверенности

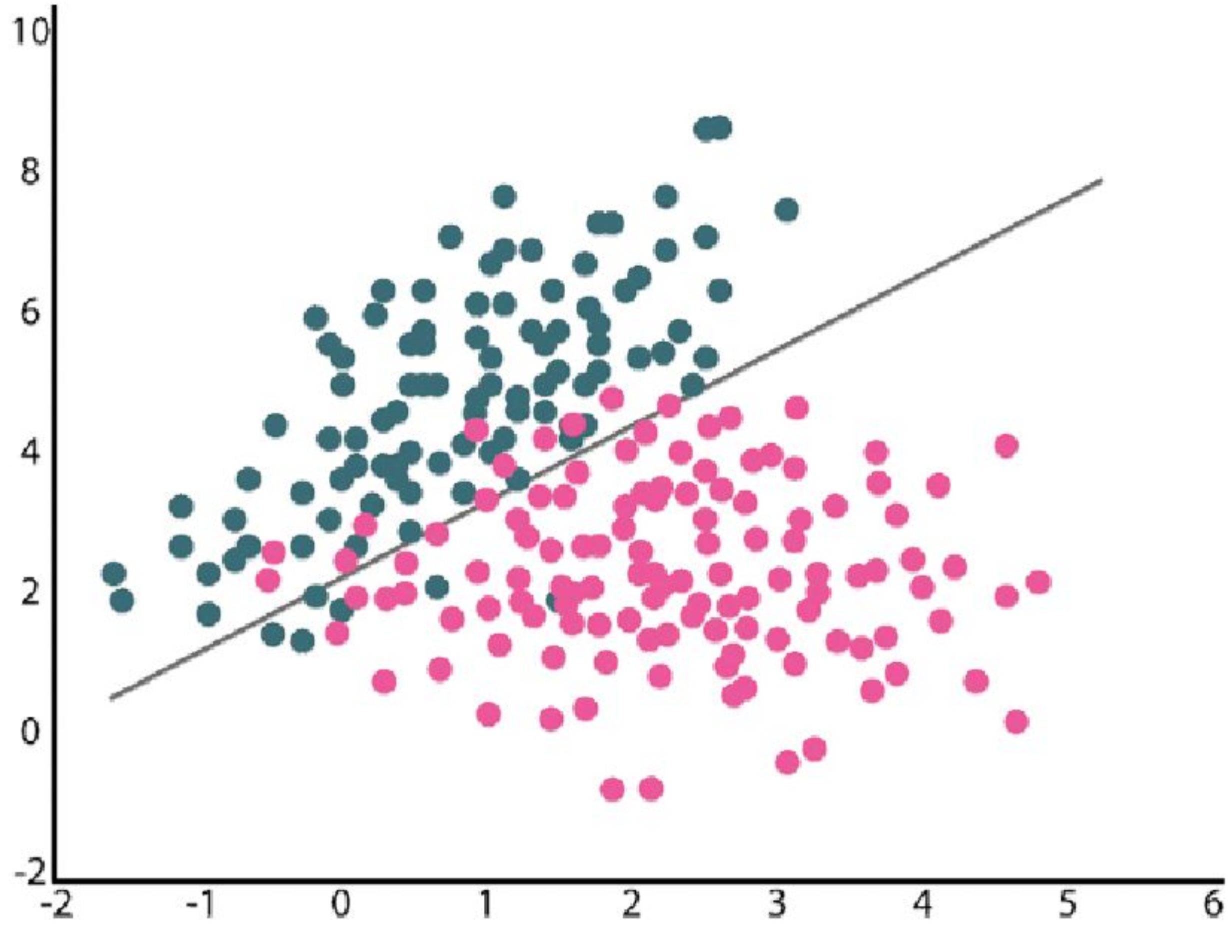


Резюме

- Линейный классификатор разделяет два класса гиперплоскостью
- Чем больше модуль отступа, тем дальше объект от гиперплоскости
- Знак отступа говорит о корректности предсказания

ФУНКЦИИ ПОТЕРЬ В ЗАДАЧАХ КЛАССИФИКАЦИИ

ЛИНЕЙНЫЙ КЛАССИФИКАТОР



ЛИНЕЙНАЯ РЕГРЕССИЯ

- Квадратичное отклонение:

$$L(a, y) = (a - y)^2$$

- Абсолютное отклонение:

$$L(a, y) = |a - y|$$

- ...

ЛИНЕЙНАЯ КЛАССИФИКАЦИЯ

- Доля неправильных ответов:

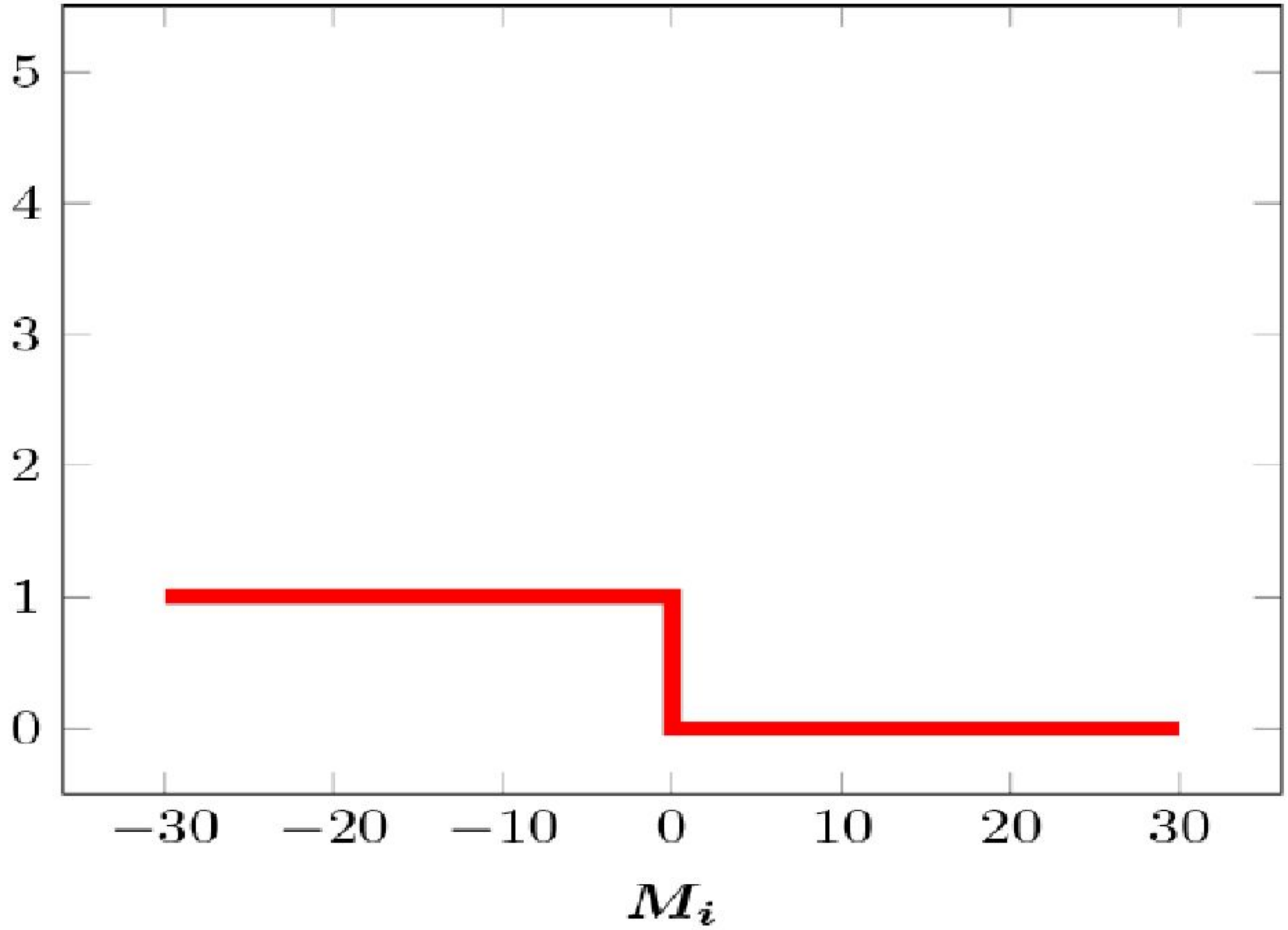
$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

ЛИНЕЙНАЯ КЛАССИФИКАЦИЯ

- Доля неправильных ответов (через отступ):

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\underbrace{y_i \langle \mathrm{w}, x_i \rangle}_{M_i} < 0]$$

ПОРОГОВАЯ ФУНКЦИЯ ПОТЕРЬ



Линейная классификация

- Доля неправильных ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [M_i < 0]$$

- Разрывная функция
- Можно использовать методы негладкой оптимизации
- Но это сложно

ОЦЕНКА ФУНКЦИИ ПОТЕРЬ

- Возьмём любую гладкую оценку пороговой функции:

$$[M < 0] \leq \tilde{L}(M)$$

- Оценим через неё функционал ошибки:

$$Q(a, X) \leq \tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(M_i)$$

Оценка функции потерь

$$Q(a, X) \leq \tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(M_i) \to \min_a$$

Минимизируем верхнюю оценку

Надеемся, что доля ошибок тоже уменьшится

ПРИМЕРЫ ОЦЕНОК

- Логистическая:

$$\tilde{L}(M) = \ln(1 + \exp(-M))$$

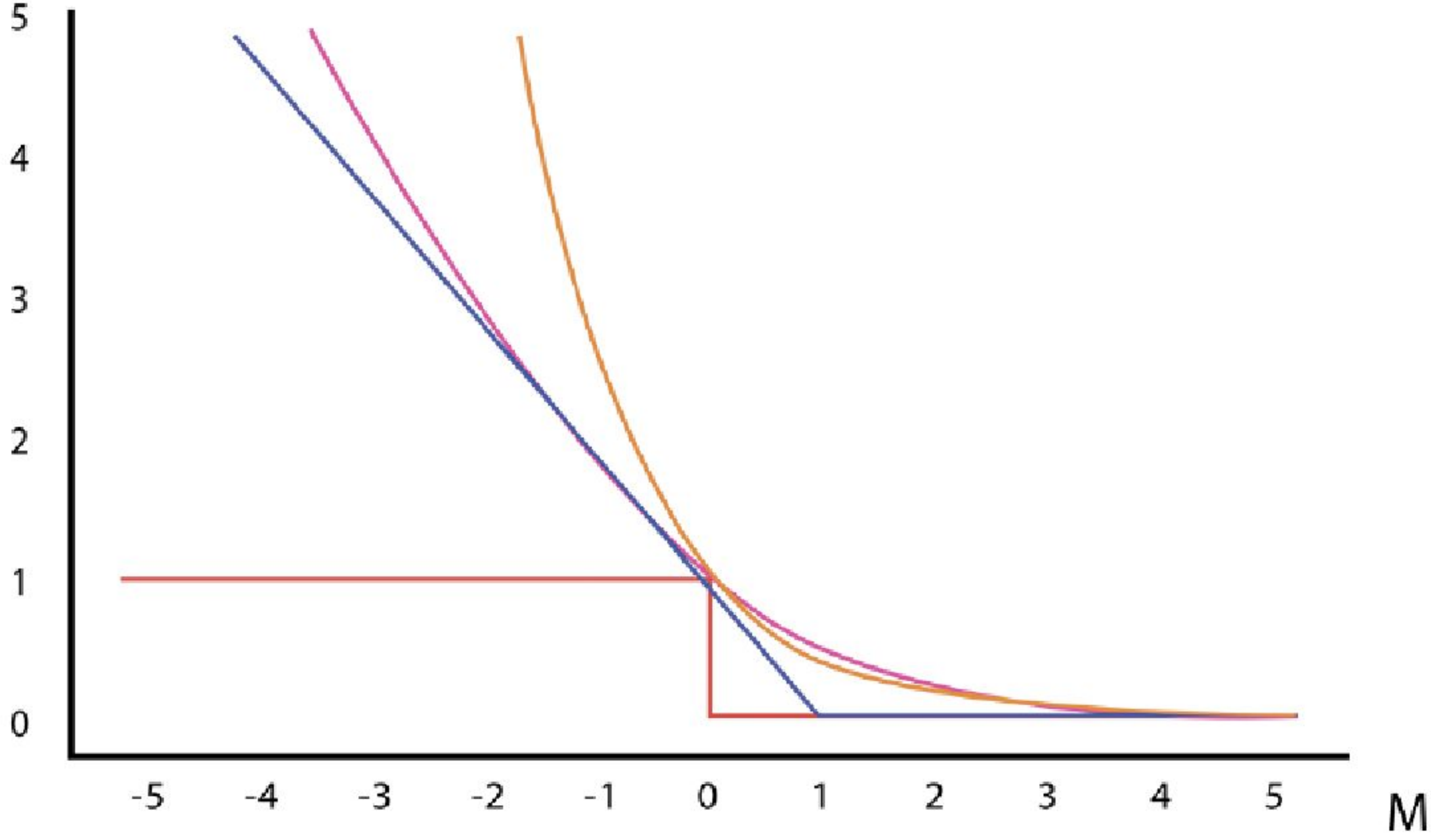
- Экспоненциальная:

$$\tilde{L}(M) = \exp(-M)$$

- Кусочно-линейная:

$$\tilde{L}(M) = \max(0, 1 - M)$$

ПРИМЕРЫ ОЦЕНОК



Логистическая функция потерь

$$\tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \ln\left(1 + \exp\left(-M_i\right)\right)$$

Логистическая функция потерь

$$\tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \ln\left(1 + \exp\left(-y_i \langle \mathrm{w}, x_i \rangle\right)\right)$$

ОБУЧЕНИЕ

- Обучение — с помощью любых методов оптимизации

- Например, стохастический градиентный спуск

РЕЗЮМЕ

- В классификации есть логичный функционал потерь — доля ошибок
- Но он негладкий
- Для гладкости нужно оценить пороговую функцию потерь
- Обучение — градиентные методы