

ЛИНЕЙНЫЕ МОДЕЛИ В ЗАДАЧАХ РЕГРЕССИИ

ОБУЧЕНИЕ С УЧИТЕЛЕМ

- › \mathbb{X} — пространство объектов
 - › \mathbb{Y} — пространство ответов
 - › $x = (x^1, \dots, x^d)$ — 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: $a(x) = \operatorname{argmin}_{a \in \mathcal{A}} Q(a, X)$

ОБУЧЕНИЕ С УЧИТЕЛЕМ

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

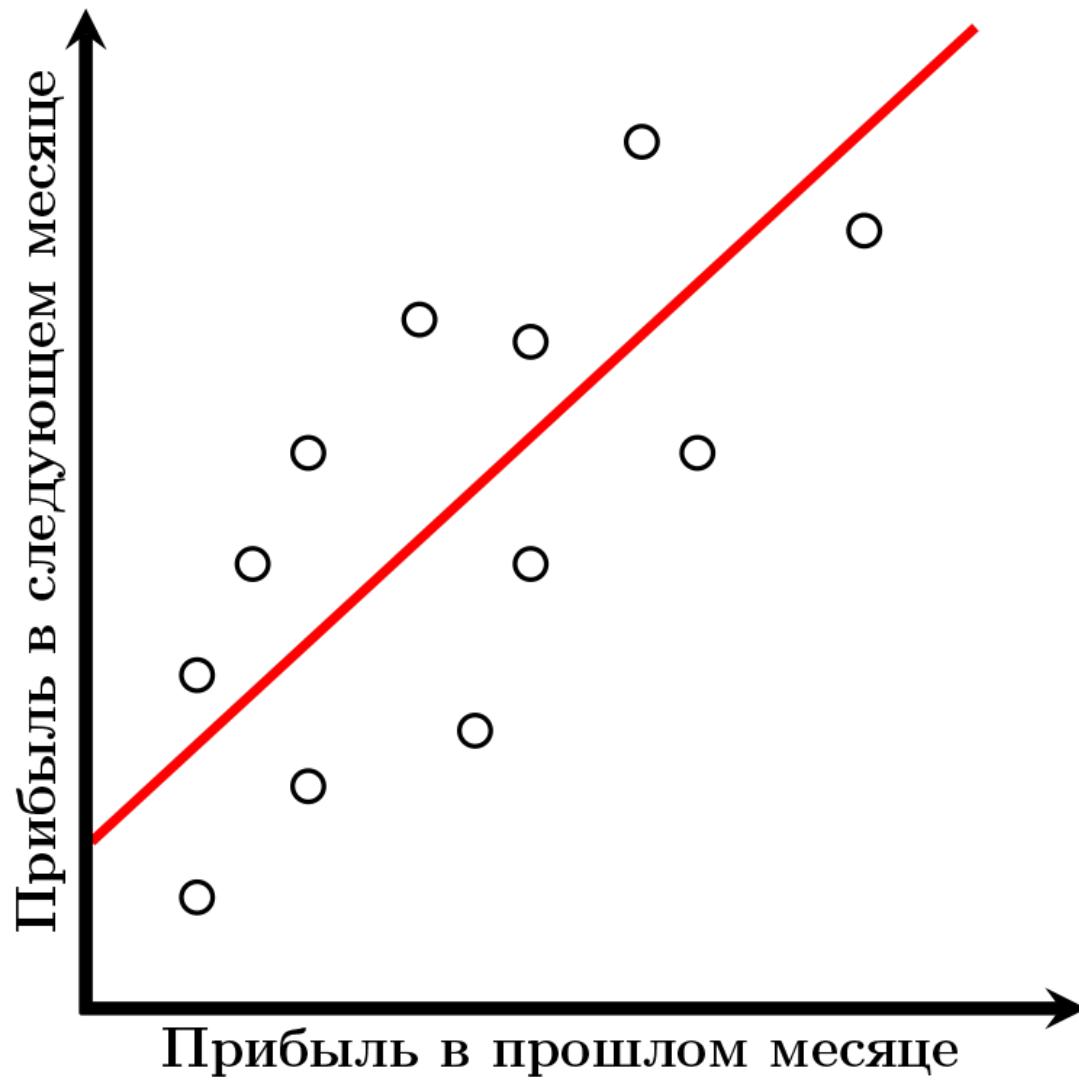
ЛИНЕЙНЫЕ МОДЕЛИ

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

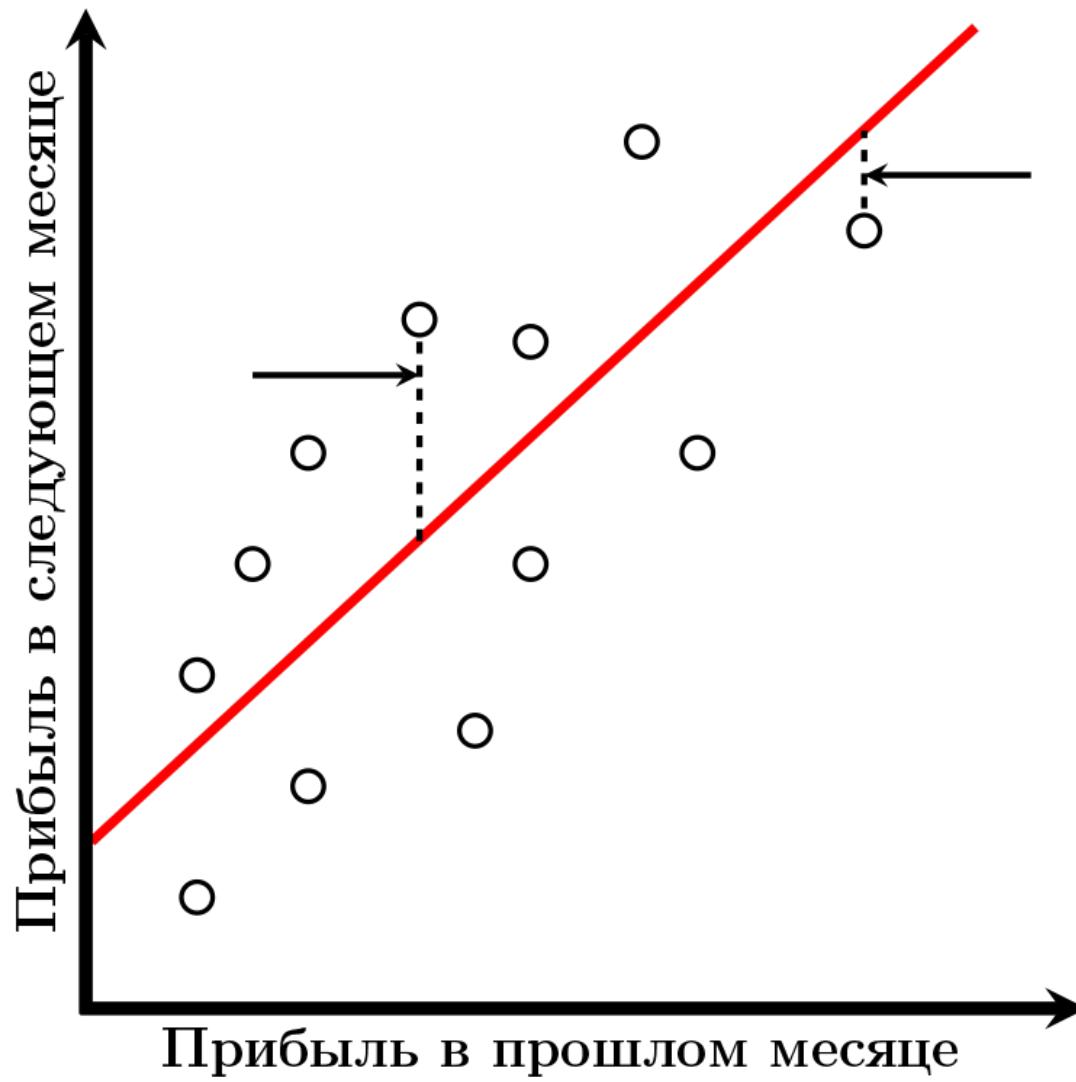
ЛИНЕЙНЫЕ МОДЕЛИ



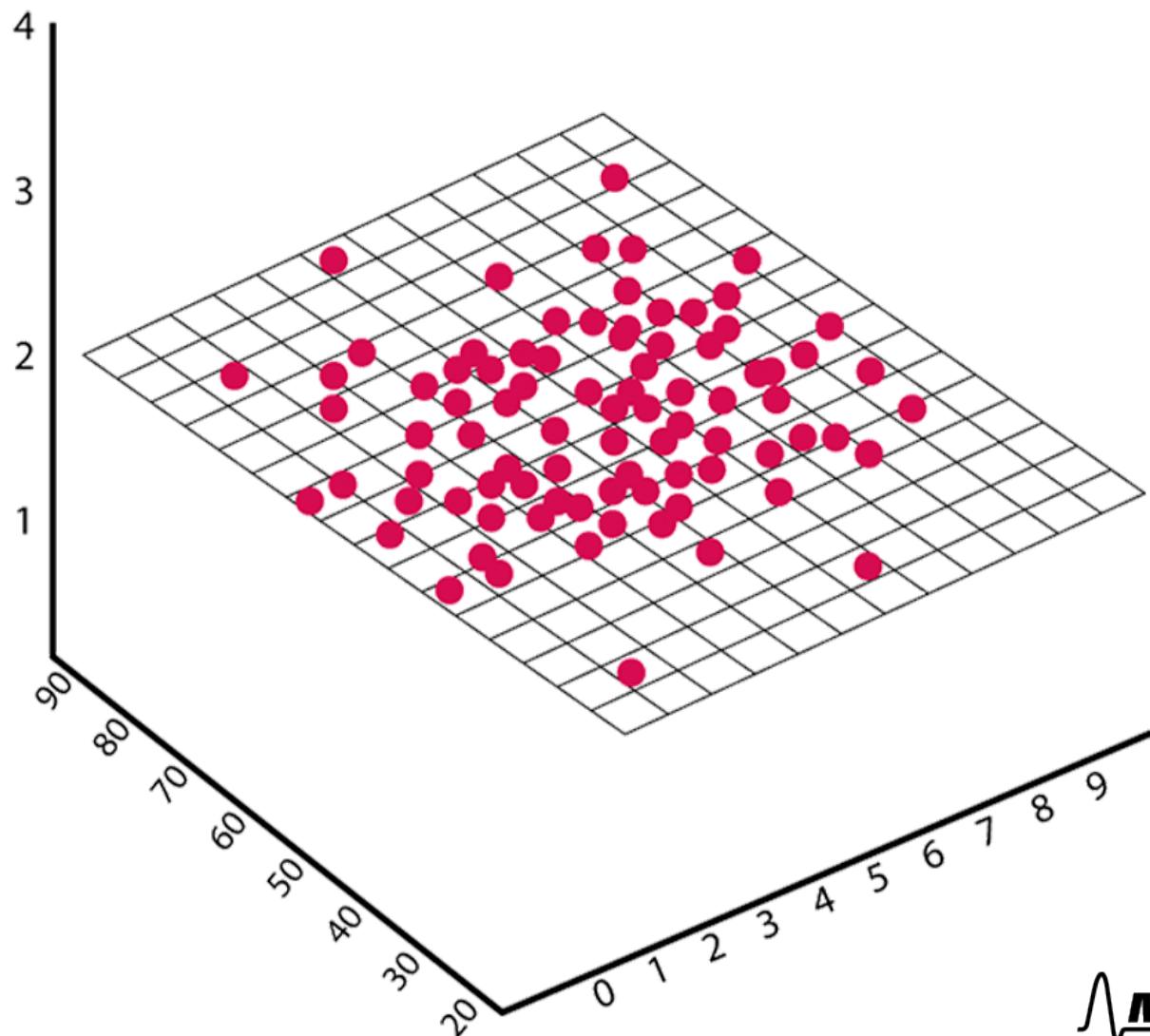
ЛИНЕЙНЫЕ МОДЕЛИ



ЛИНЕЙНЫЕ МОДЕЛИ



ЛИНЕЙНЫЕ МОДЕЛИ



ОБУЧЕНИЕ С УЧИТЕЛЕМ

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

ЛИНЕЙНЫЕ МОДЕЛИ

ЛИНЕЙНЫЕ МОДЕЛИ

› Добавим константный признак

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

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:

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

Вещественный вектор

РЕЗЮМЕ

- › Линейные алгоритмы для регрессии
- › Среднеквадратичная ошибка

ОБУЧЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ

ОБУЧЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ

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

- › d неизвестных
- › Есть константный признак
- › Выпуклая функция

МАТРИЧНАЯ ЗАПИСЬ

» Матрица «объекты-признаки»:

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

Объект

МАТРИЧНАЯ ЗАПИСЬ

» Матрица «объекты-признаки»:

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

Признак

МАТРИЧНАЯ ЗАПИСЬ

› Вектор ответов:

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

МАТРИЧНАЯ ЗАПИСЬ

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

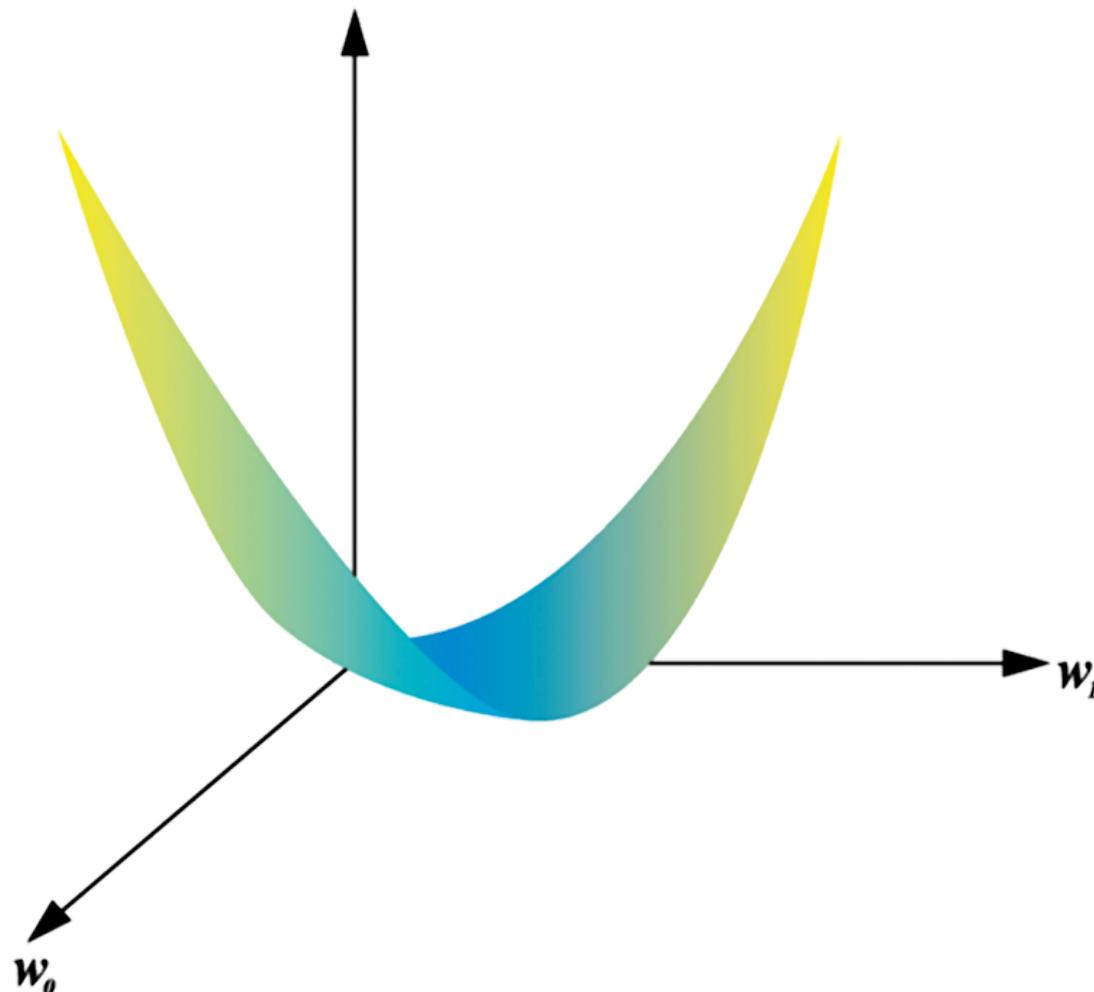
АНАЛИТИЧЕСКОЕ РЕШЕНИЕ

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

- › Нужно обращать матрицу $d \times d$ — сложность d^3
- › Могут возникнуть численные проблемы

ГРАДИЕНТНЫЙ СПУСК

› Функция ошибки гладкая и выпуклая



ГРАДИЕНТНЫЙ СПУСК: НАПОМИНАНИЕ

- › Инициализация: $\mathbf{w}^0 = \mathbf{0}$
- › 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

› Инициализация: $\mathbf{w}^0 = \mathbf{0}$

› 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

- › Инициализация: $\mathbf{w}^0 = \mathbf{0}$
- › 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

- › Инициализация: $\mathbf{w}^0 = \mathbf{0}$
- › 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

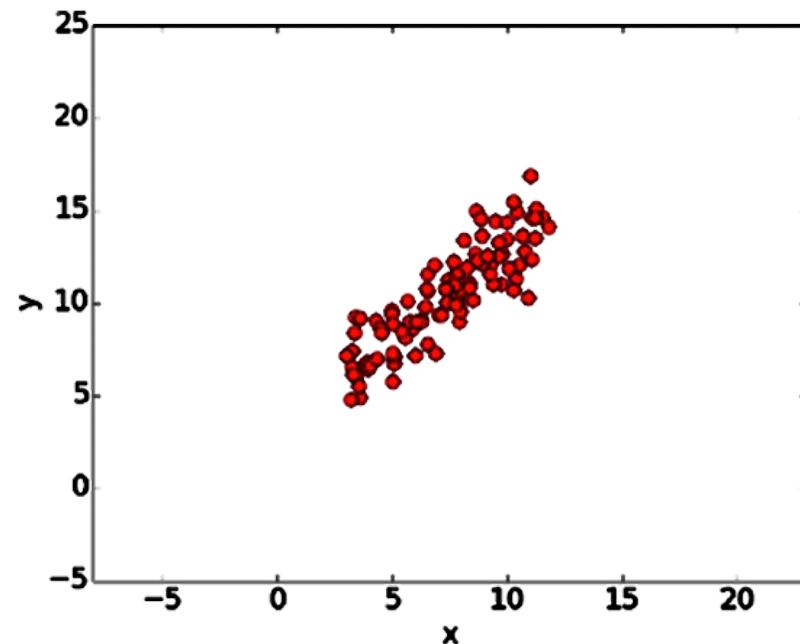
› Модель: $a(x) = w_1 x + w_0$

› Два параметра: w_1 и w_0

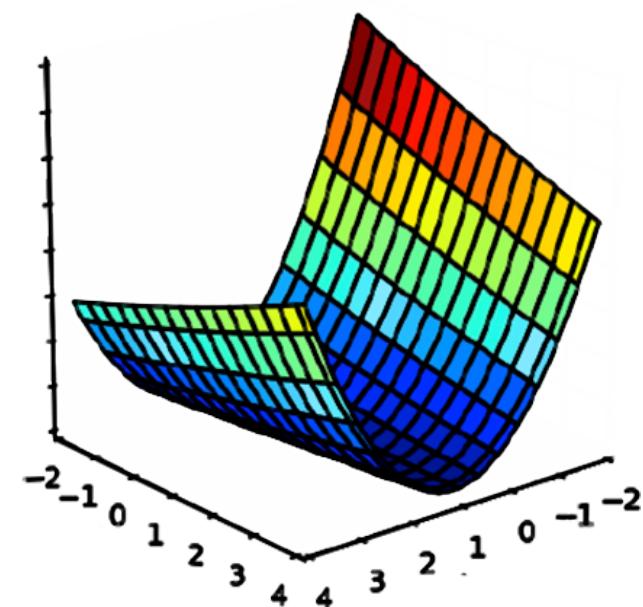
› Функционал:

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

ПАРНАЯ РЕГРЕССИЯ



Выборка



Функционал качества

ГРАДИЕНТНЫЙ СПУСК

- » Инициализация: $\mathbf{w}^0 = \mathbf{0}$
- » 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

ГРАДИЕНТ ДЛЯ ПАРНОЙ РЕГРЕССИИ

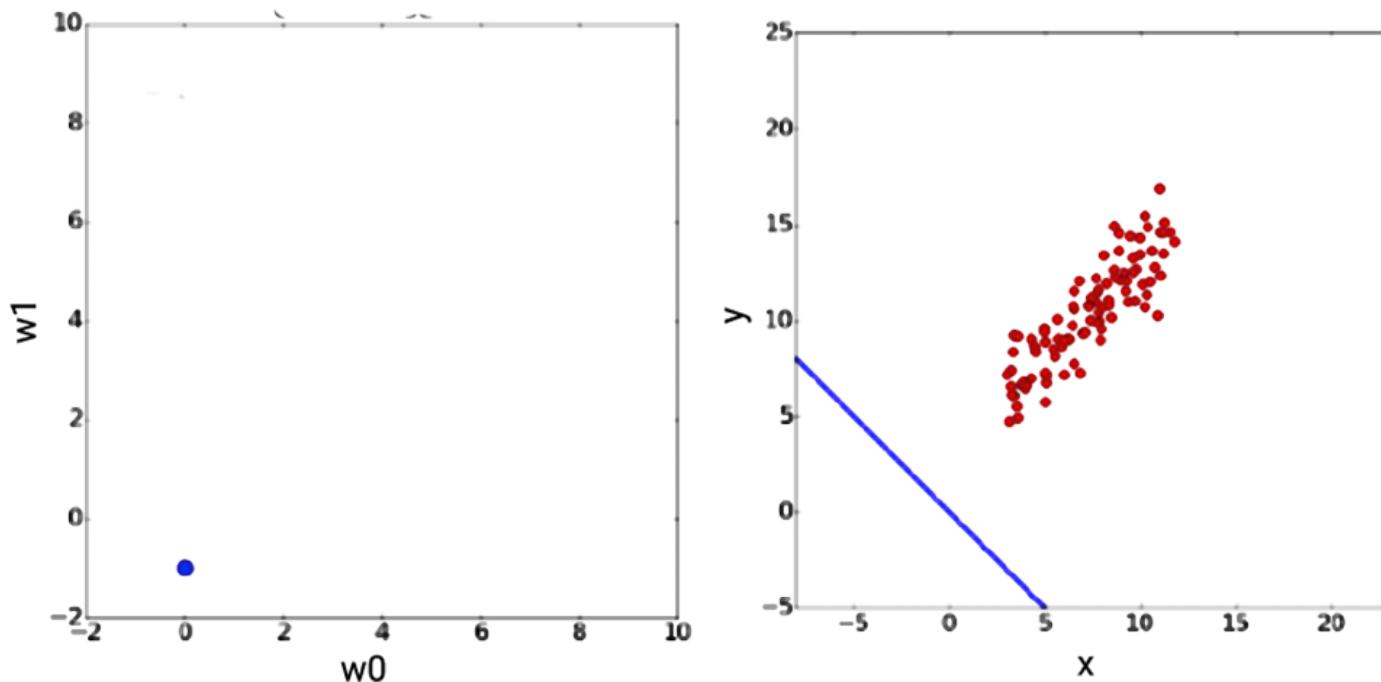
$$Q(w_0, w_1, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$$

› Частные производные:

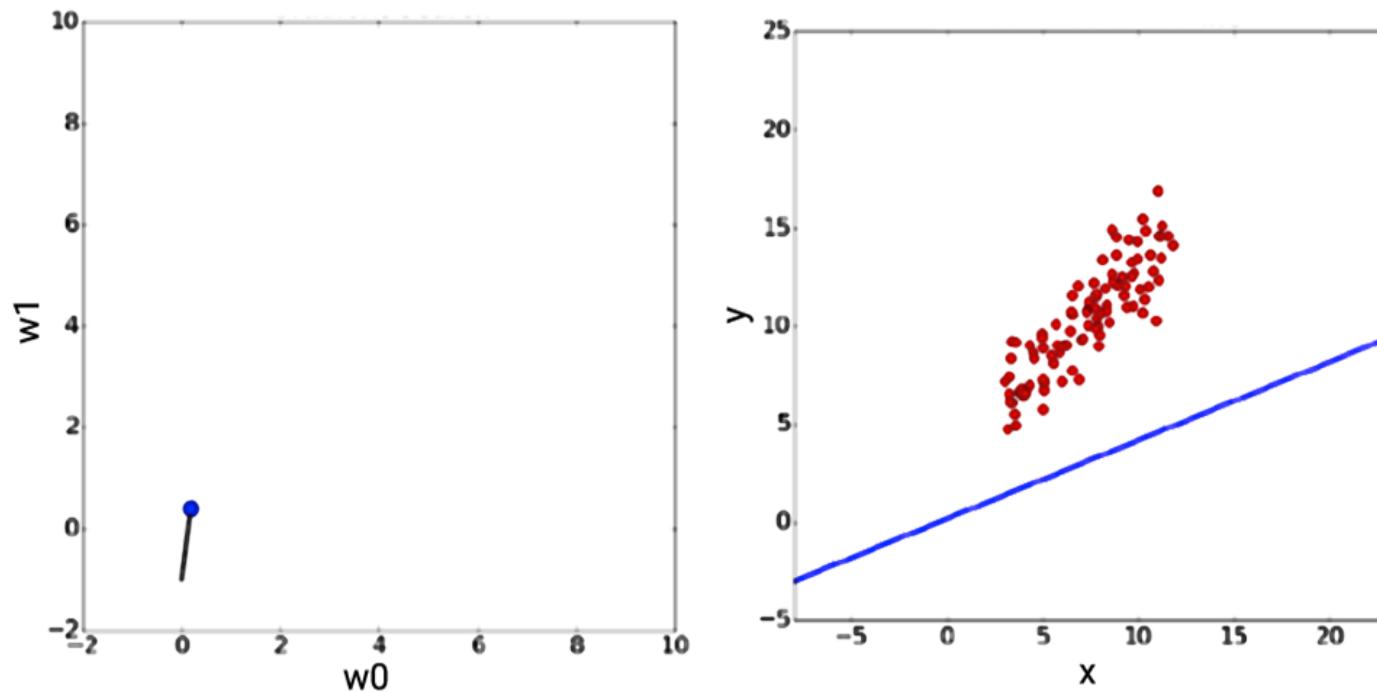
$$\frac{\partial Q}{\partial w_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i) x_i$$

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

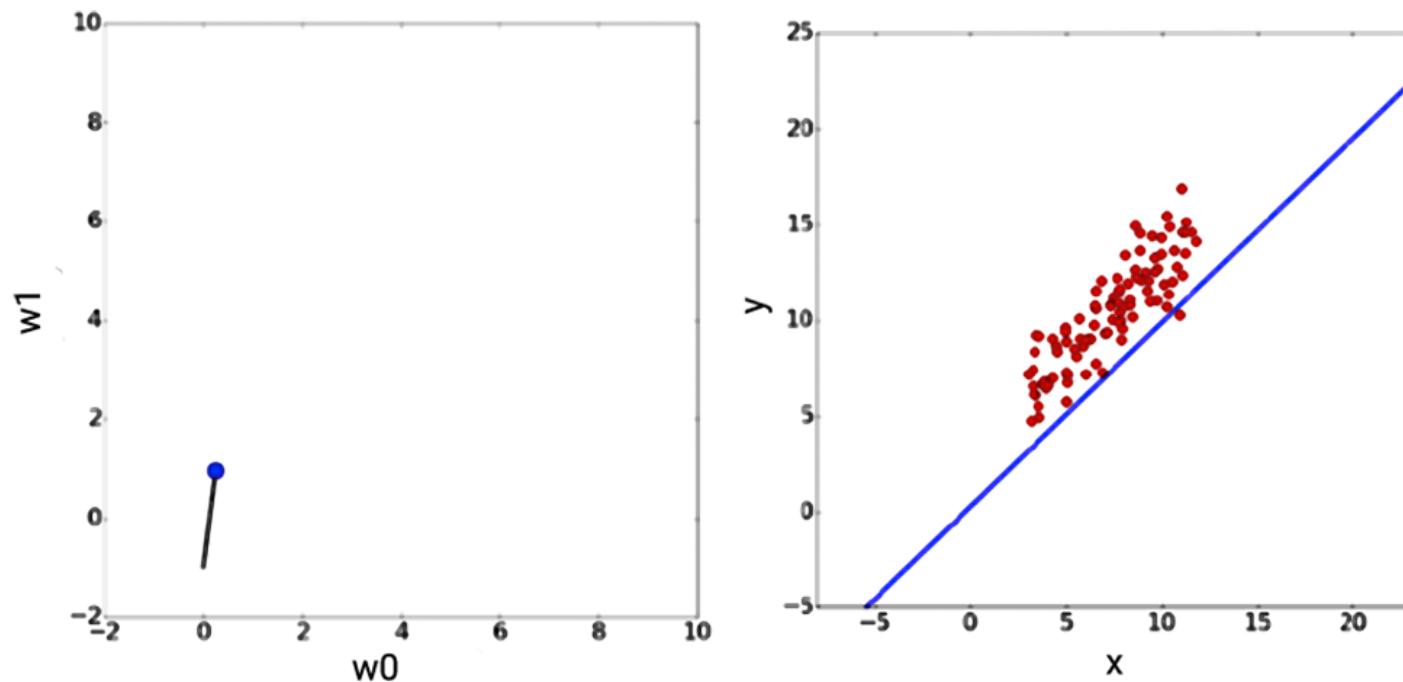
ПАРНАЯ РЕГРЕССИЯ



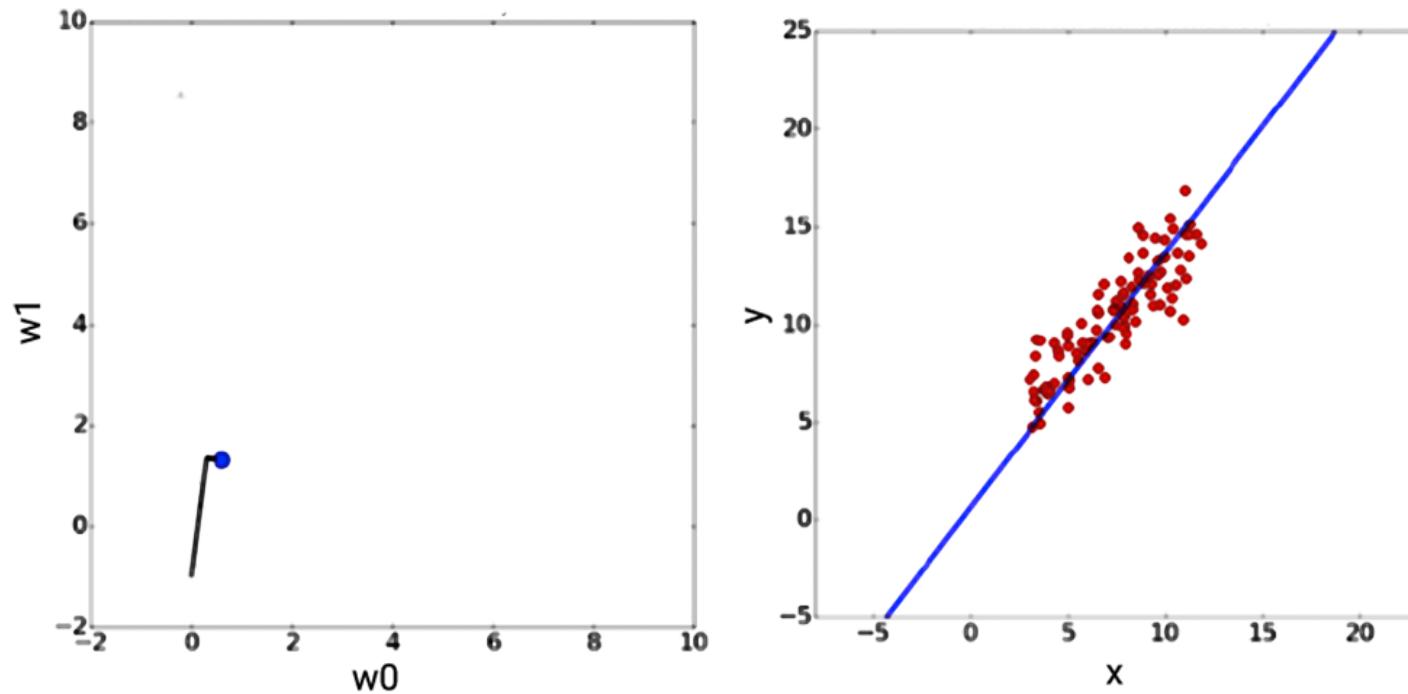
ПАРНАЯ РЕГРЕССИЯ



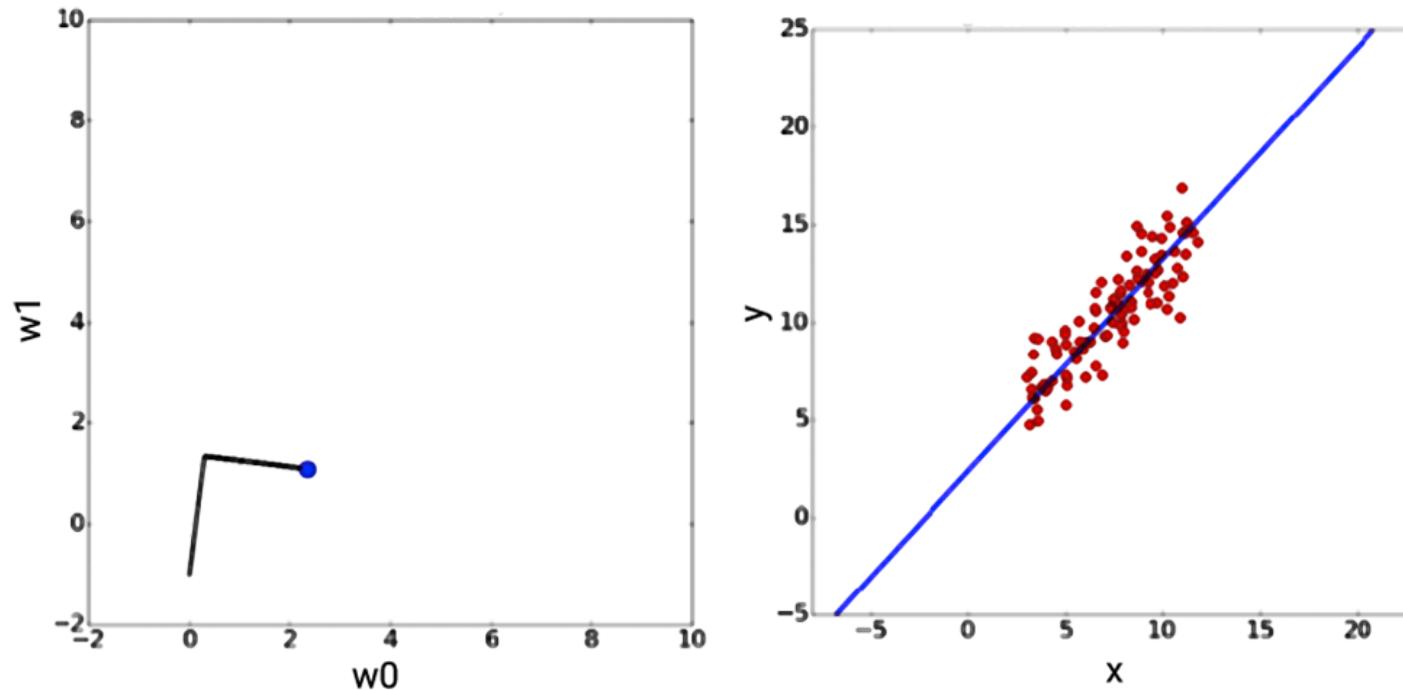
ПАРНАЯ РЕГРЕССИЯ



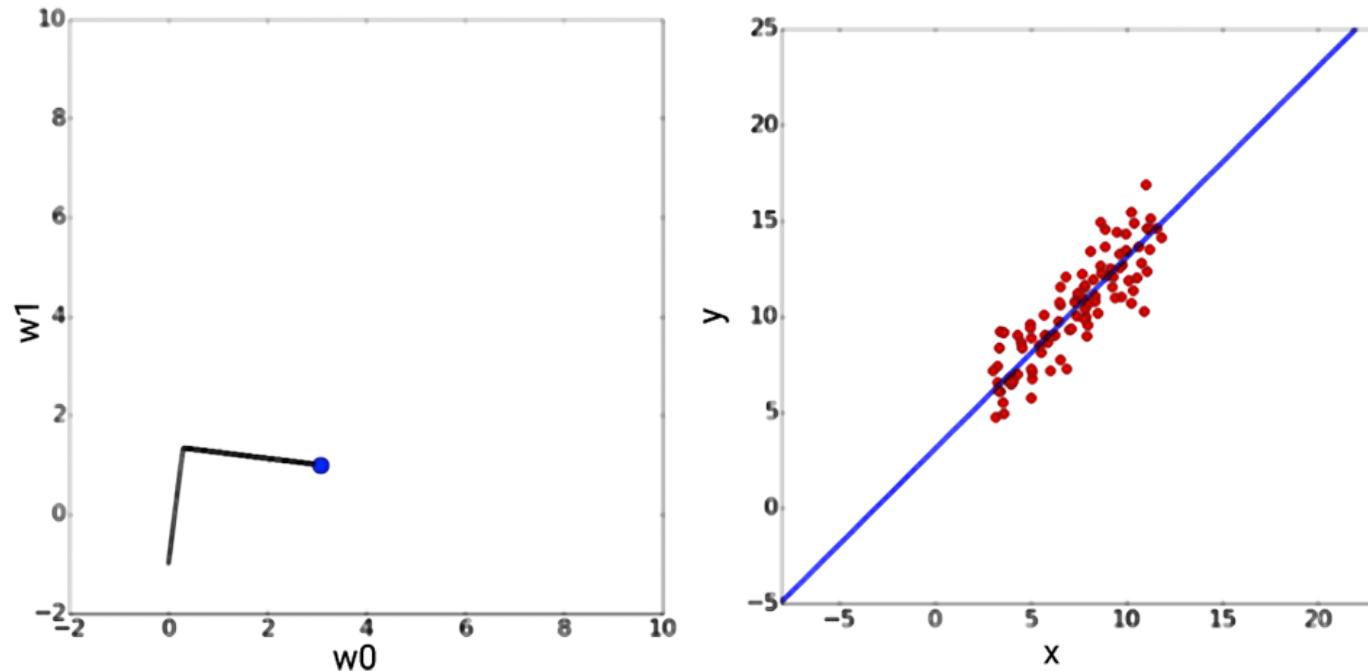
ПАРНАЯ РЕГРЕССИЯ



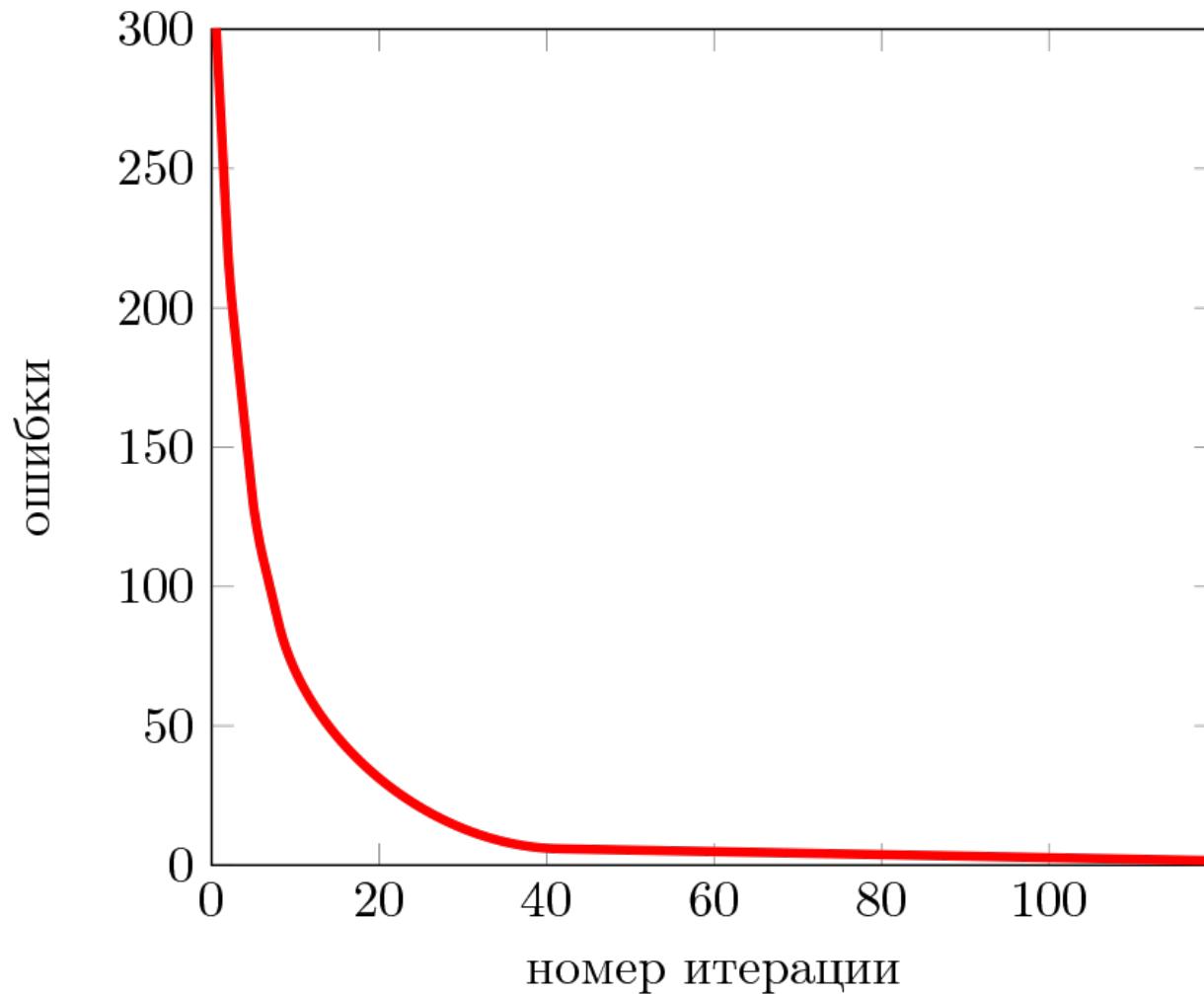
ПАРНАЯ РЕГРЕССИЯ



ПАРНАЯ РЕГРЕССИЯ

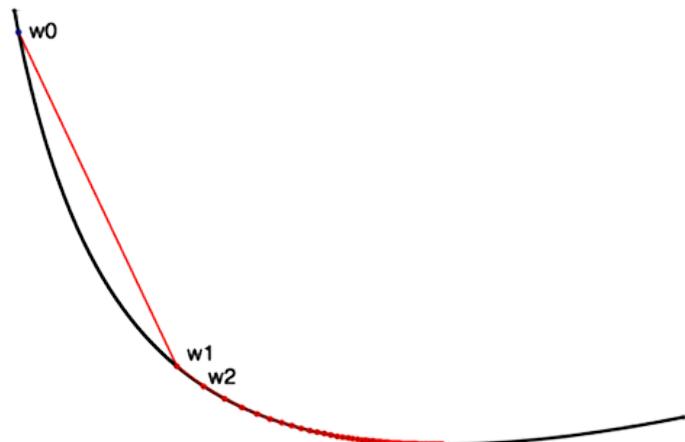


ФУНКЦИОНАЛ КАЧЕСТВА

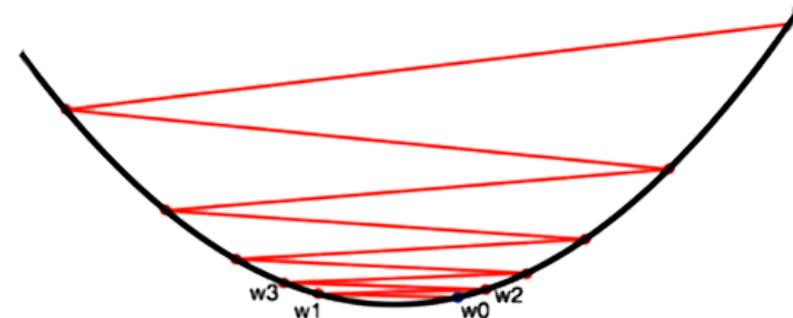


РАЗМЕР ШАГА

› Выбор размера шага η — искусство



Маленький шаг



Большой шаг

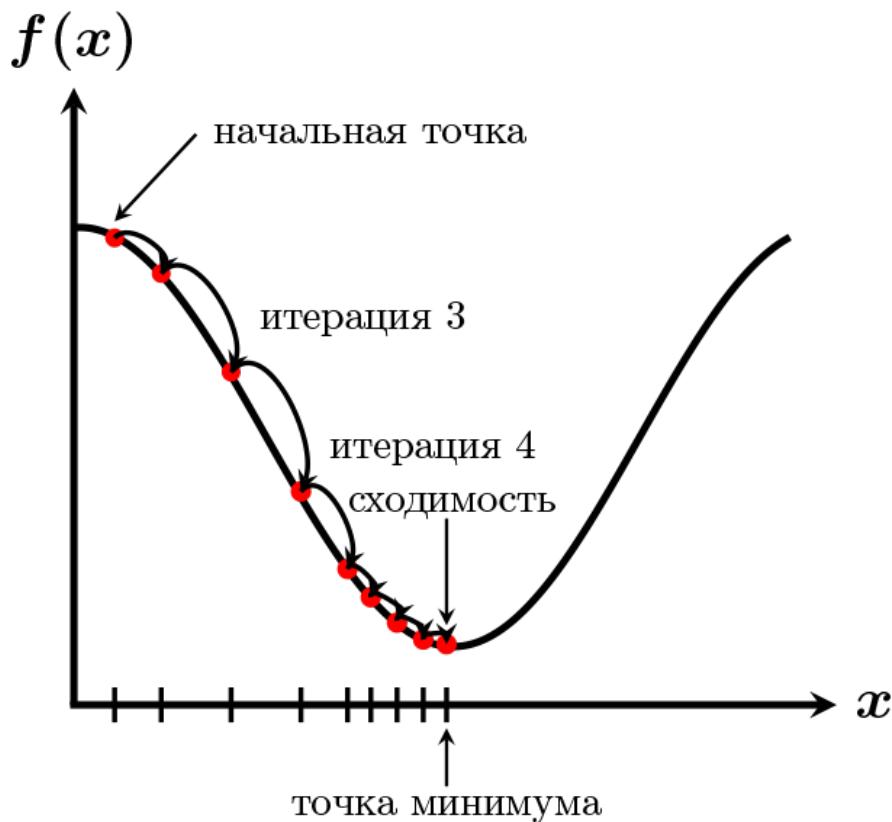
РАЗМЕР ШАГА

- › 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

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

› Градиент:

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

РЕЗЮМЕ

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

СТОХАСТИЧЕСКИЙ ГРАДИЕНТНЫЙ СПУСК

ГРАДИЕНТНЫЙ СПУСК

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

ГРАДИЕНТ ФУНКЦИОНАЛА

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

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

ГРАДИЕНТ ФУНКЦИОНАЛА

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

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

Суммирование по всей выборке!

ГРАДИЕНТ ФУНКЦИОНАЛА

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

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

Как поменять веса, чтобы
улучшить качество на объекте x_i

ГРАДИЕНТ ФУНКЦИОНАЛА

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

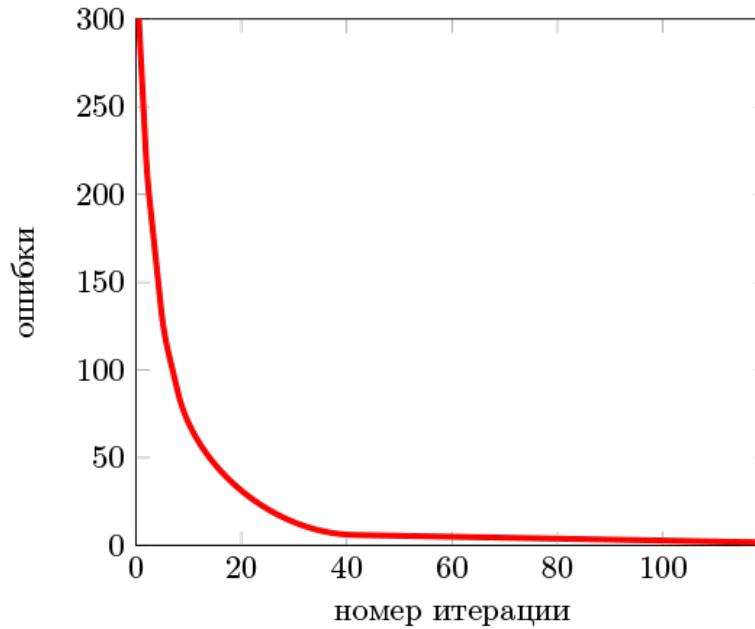
$$\frac{\partial Q}{\partial w_j} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i^j (\langle \mathbf{w}, x_i \rangle - y_i)$$

Как поменять веса, чтобы
улучшить качество на всей выборке

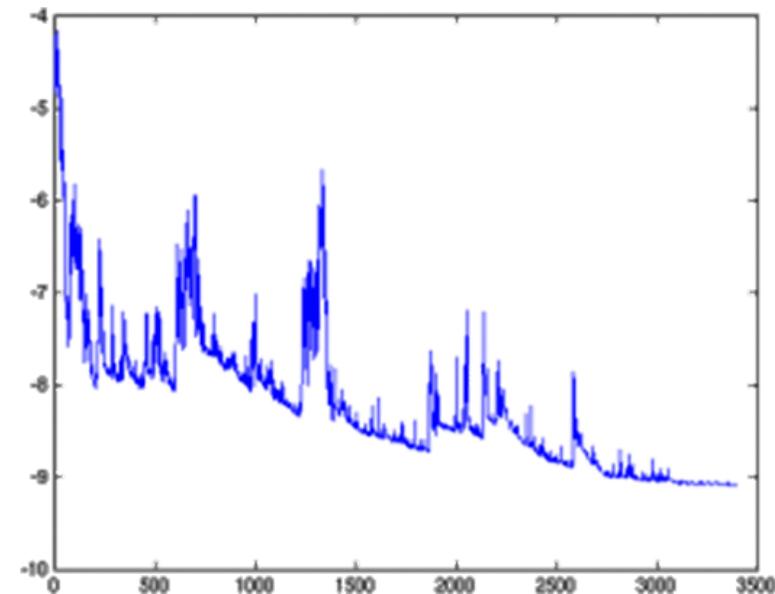
СТОХАСТИЧЕСКИЙ ГРАДИЕНТНЫЙ СПУСК

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

$$a(x) = \text{sign} \left(w_0 + \sum_{j=1}^d w_j x^j \right)$$

Свободный коэффициент

Веса

Признаки

ЛИНЕЙНЫЙ КЛАССИФИКАТОР

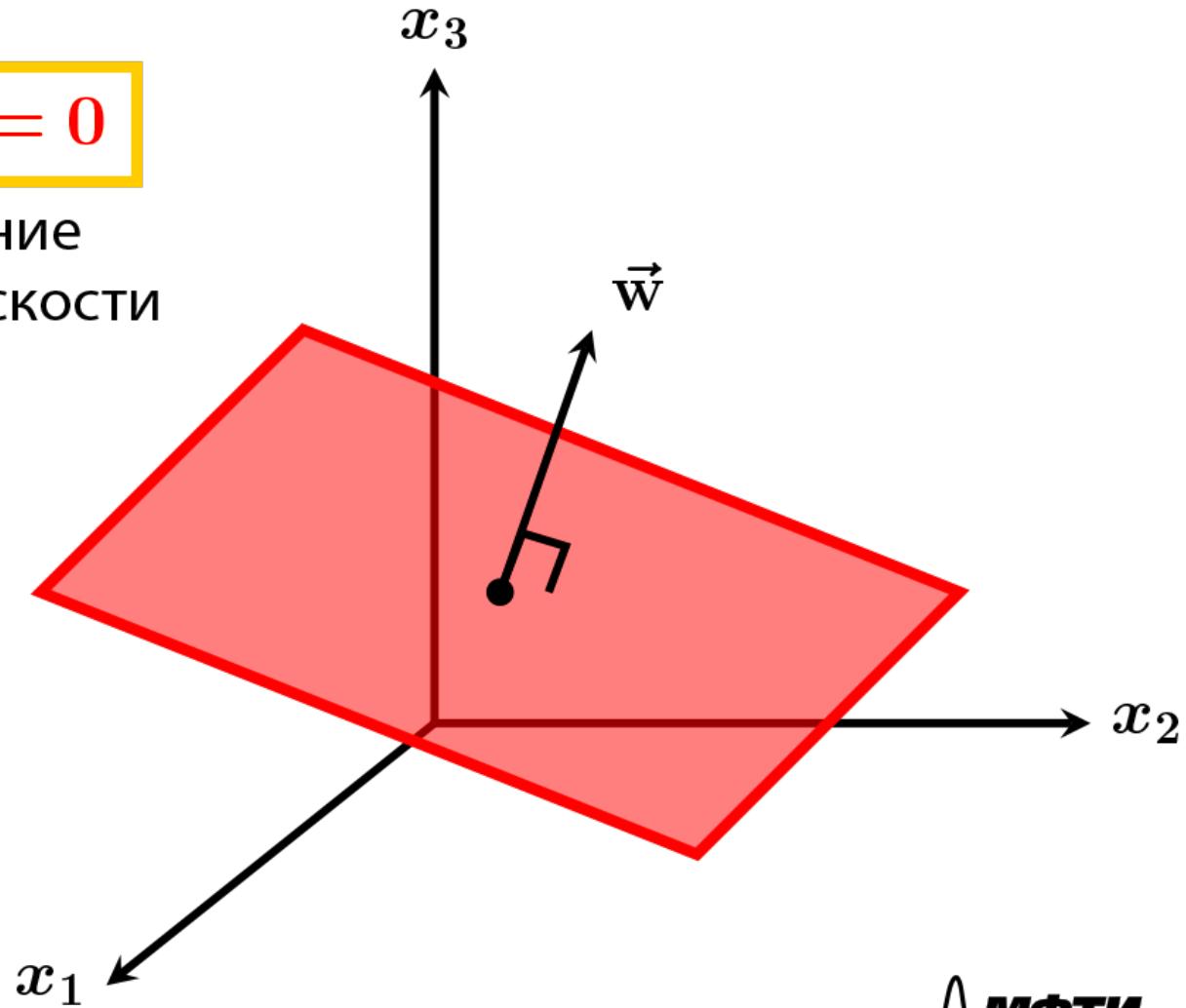
› Добавим единичный признак:

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

ГЕОМЕТРИЯ ЛИНЕЙНОГО КЛАССИФИКАТОРА

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

Уравнение
гиперплоскости



ГЕОМЕТРИЯ ЛИНЕЙНОГО КЛАССИФИКАТОРА

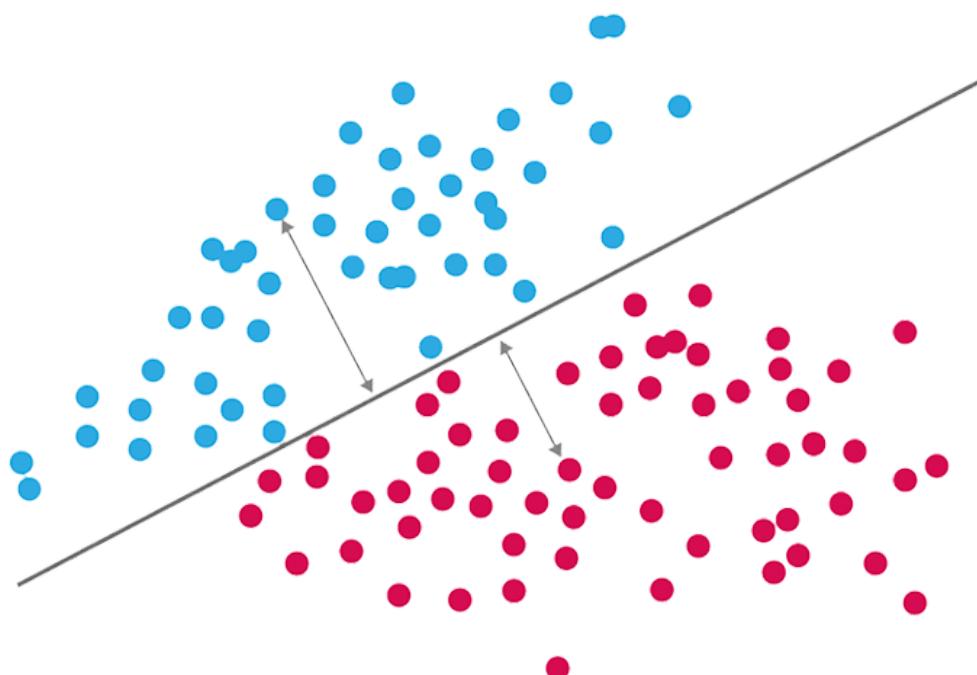


ГЕОМЕТРИЯ ЛИНЕЙНОГО КЛАССИФИКАТОРА

- » Расстояние от точки до гиперплоскости

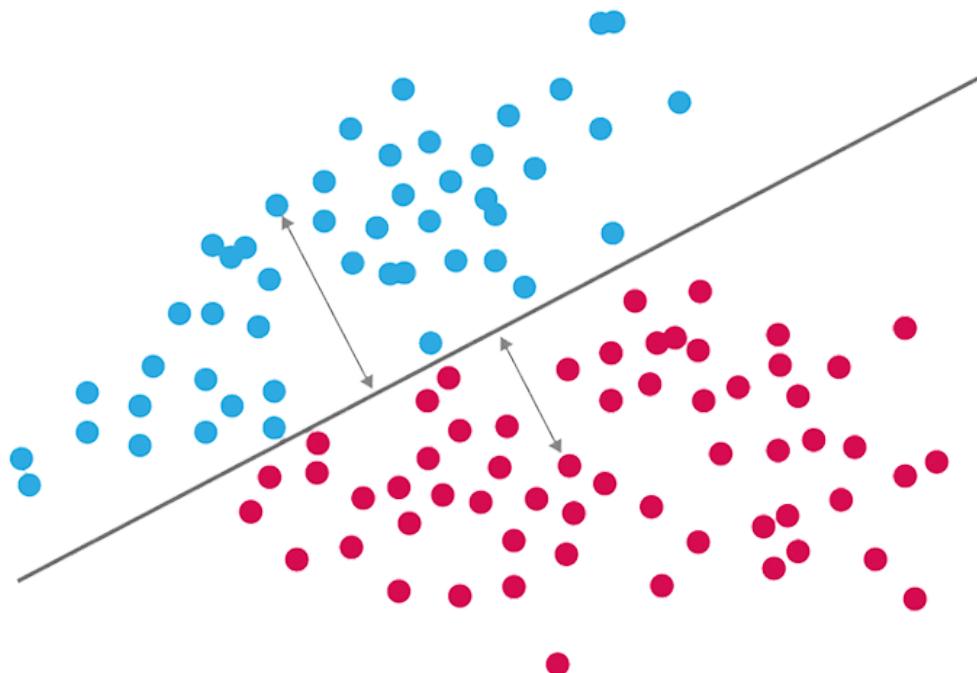
$$\langle \mathbf{w}, \mathbf{x} \rangle = 0: \quad \frac{|\langle \mathbf{w}, \mathbf{x} \rangle|}{\|\mathbf{w}\|}$$

- » Чем больше $\langle \mathbf{w}, \mathbf{x} \rangle$, тем дальше объект от разделяющей гиперплоскости



ОТСТУП

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

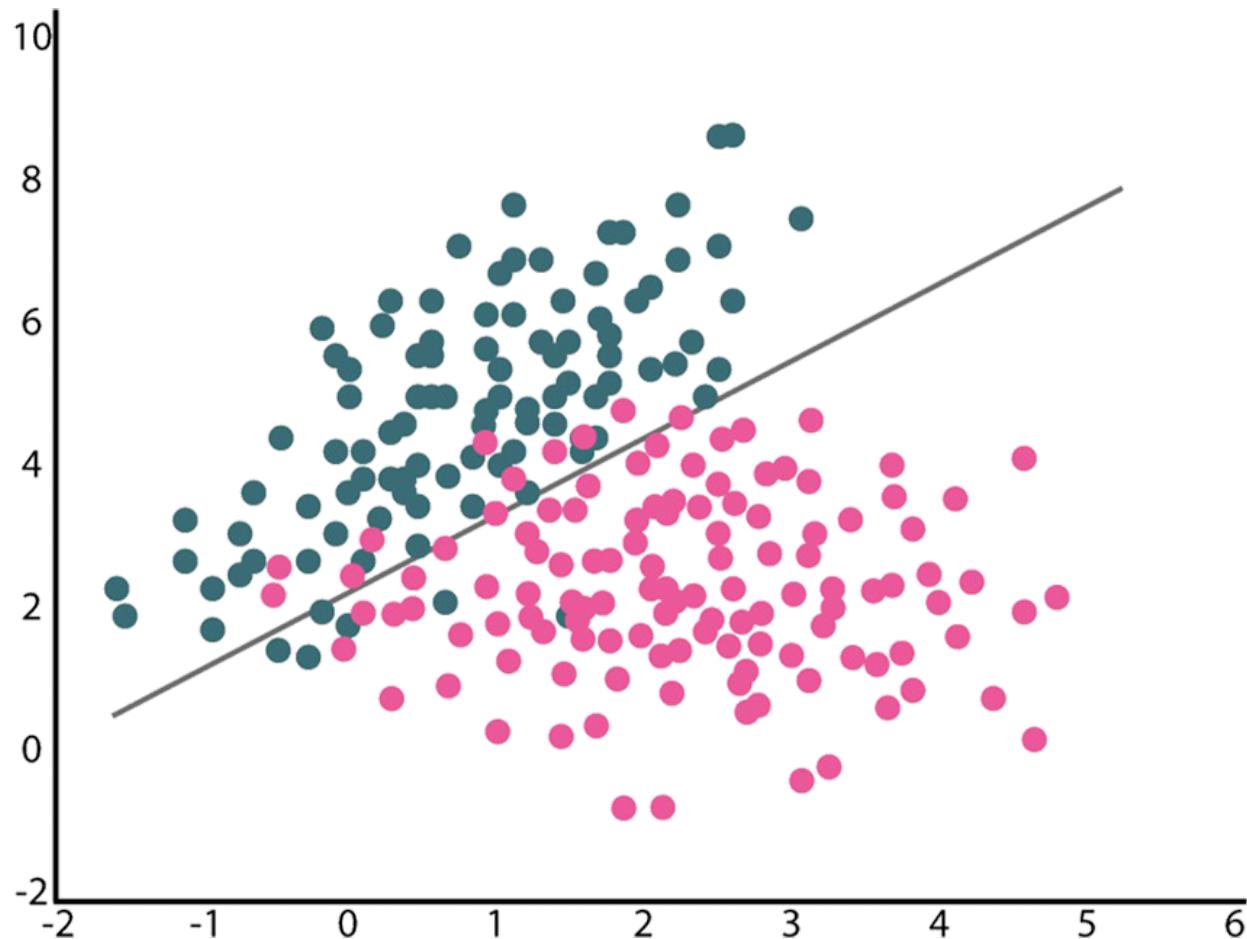


РЕЗЮМЕ

- › 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:

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

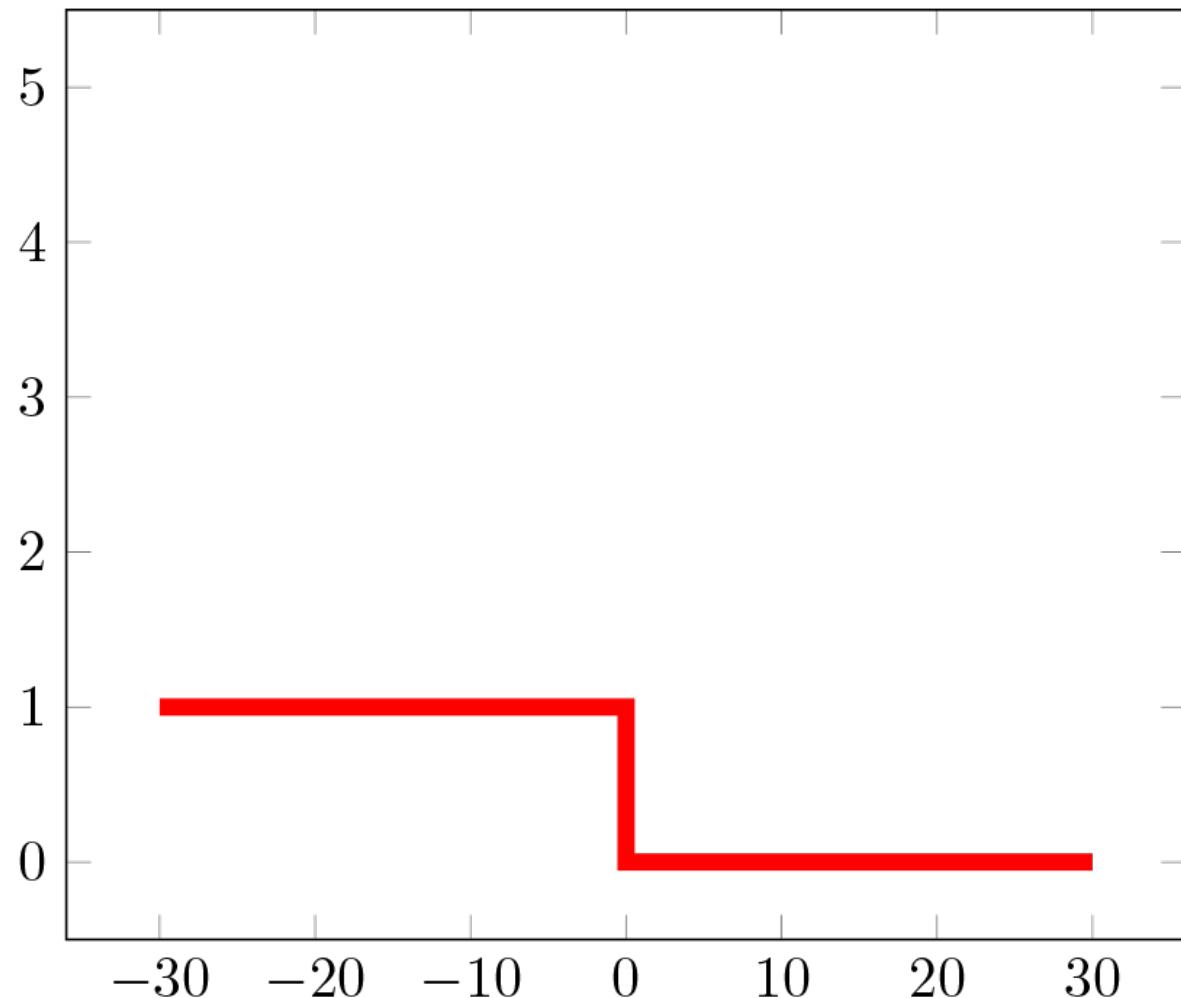
ЛИНЕЙНАЯ КЛАССИФИКАЦИЯ

› Доля неправильных ответов (через отступ):

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [y_i \langle \mathbf{w}, \mathbf{x}_i \rangle < 0]$$

M_i

ПОРОГОВАЯ ФУНКЦИЯ ПОТЕРЬ



M_i

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

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

Минимизируем
верхнюю оценку

Надеемся, что доля
ошибок тоже
уменьшится



ПРИМЕРЫ ОЦЕНОК

› Логистическая:

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

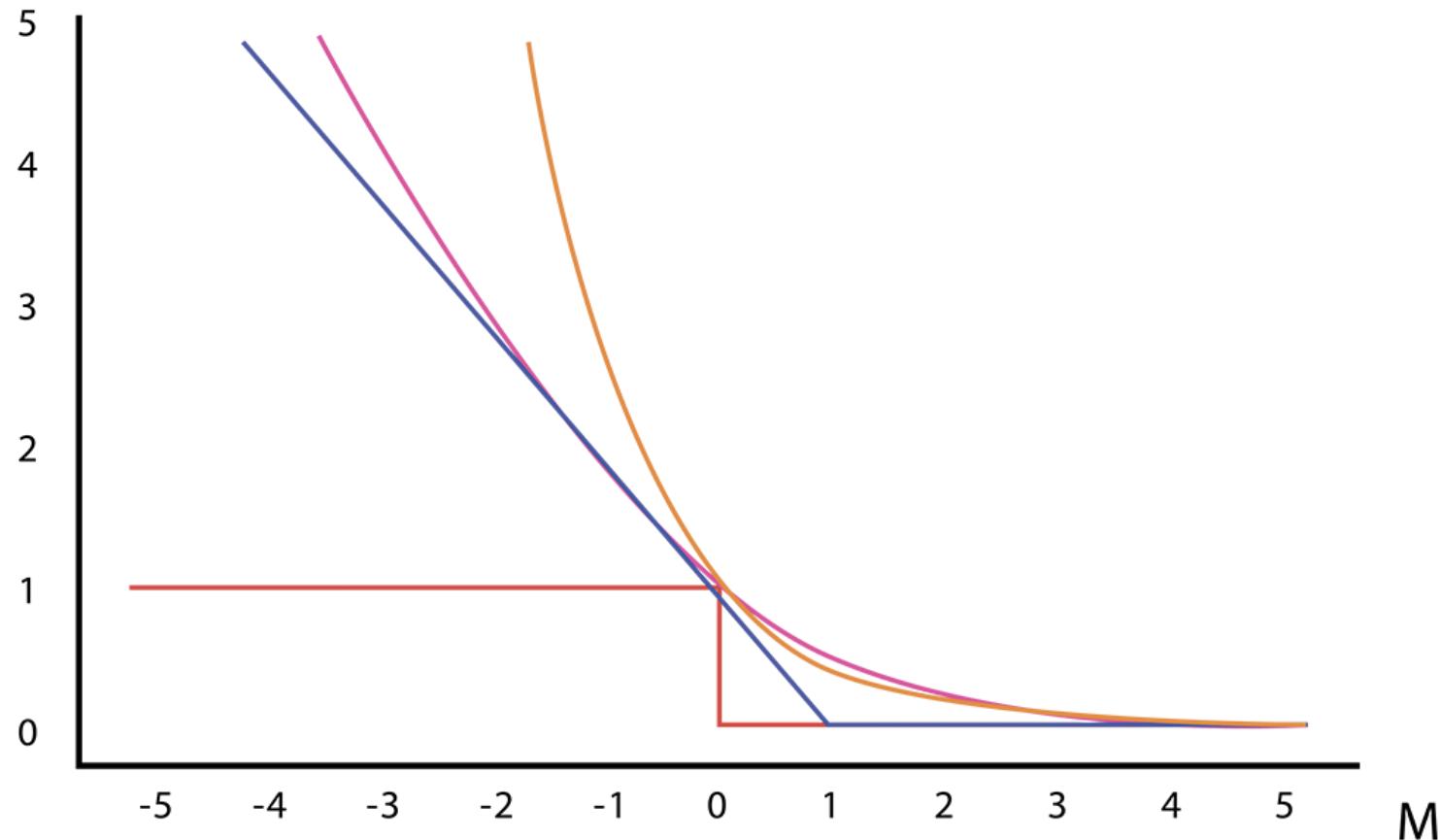
› Экспоненциальная:

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

› Кусочно-линейная:

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

ПРИМЕРЫ ОЦЕНОК



ЛОГИСТИЧЕСКАЯ ФУНКЦИЯ ПОТЕРЬ

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

ЛОГИСТИЧЕСКАЯ ФУНКЦИЯ ПОТЕРЬ

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

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