

Основы машинного обучения

Лекция 7

Функции потерь в регрессии. Линейная классификация.

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2023

Функции потерь в задачах регрессии

Среднеквадратичная ошибка

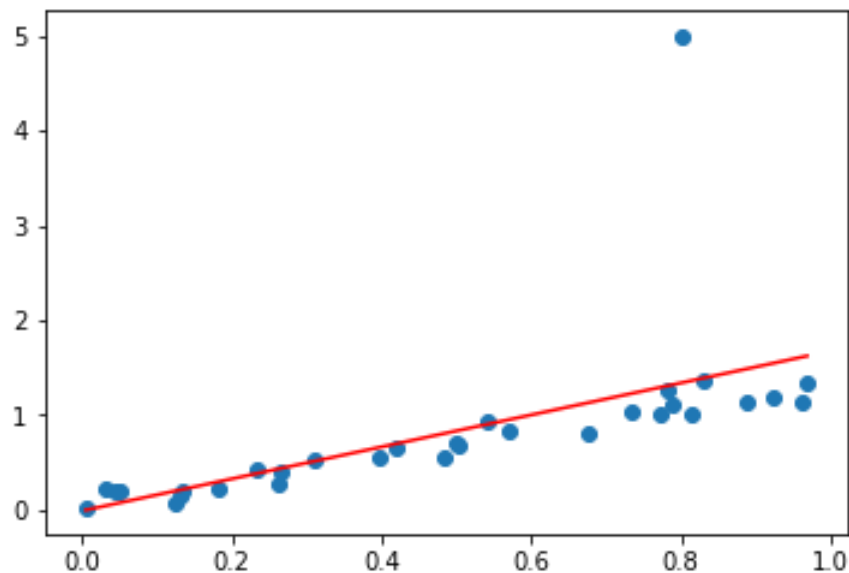
- Частый выбор — квадратичная функция потерь

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

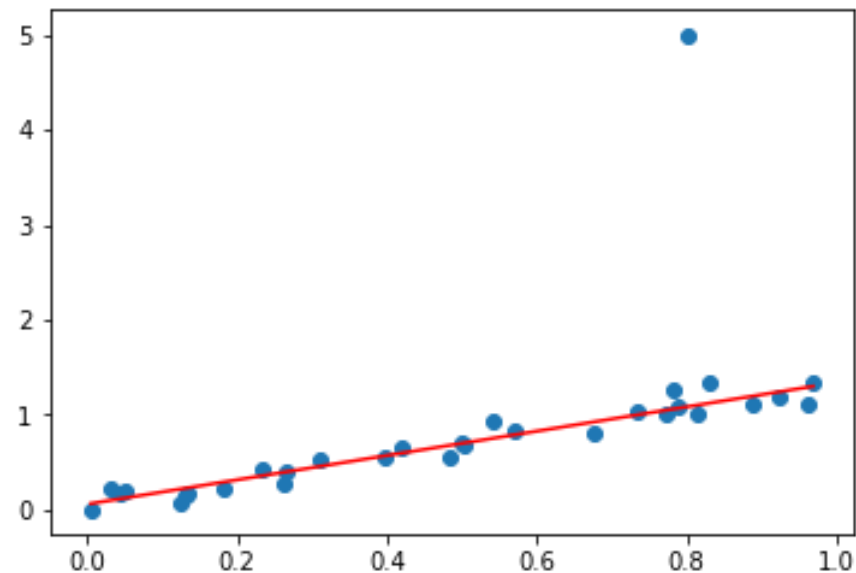
- Функционал ошибки — среднеквадратичная ошибка (mean squared error, MSE)

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

Выбросы



С учётом выброса



Без учёта выброса

Обучение на среднеквадратичную ошибку

Выбросы

$a(x)$	y	$(a(x) - y)^2$
2	1	1
1	2	1
2	3	1
5	4	1
6	5	1
7	100	8649
6	7	1

$$MSE \approx 1236$$

Выбросы

$a(x)$	y	$(a(x) - y)^2$
4	1	9
5	2	9
6	3	9
7	4	9
8	5	9
10	100	8100
10	7	9

$$MSE \approx 1164$$

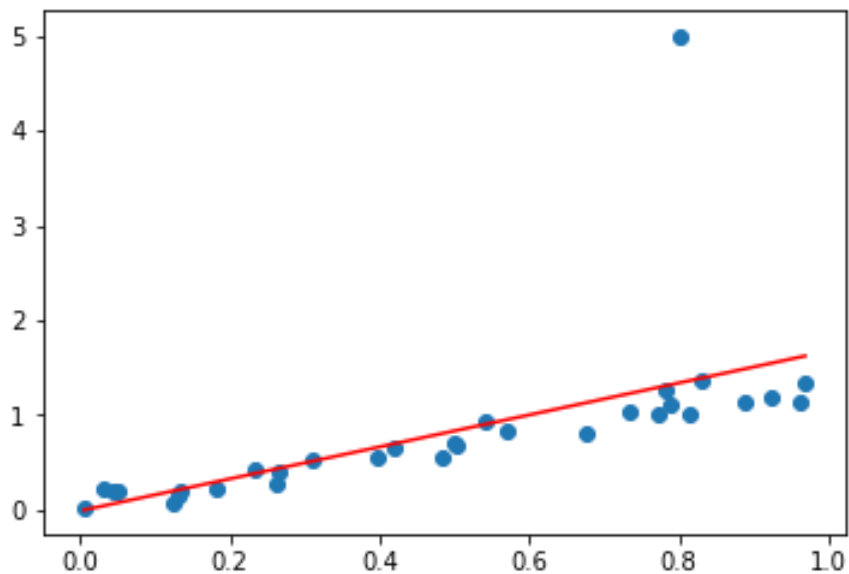
Средняя абсолютная ошибка

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

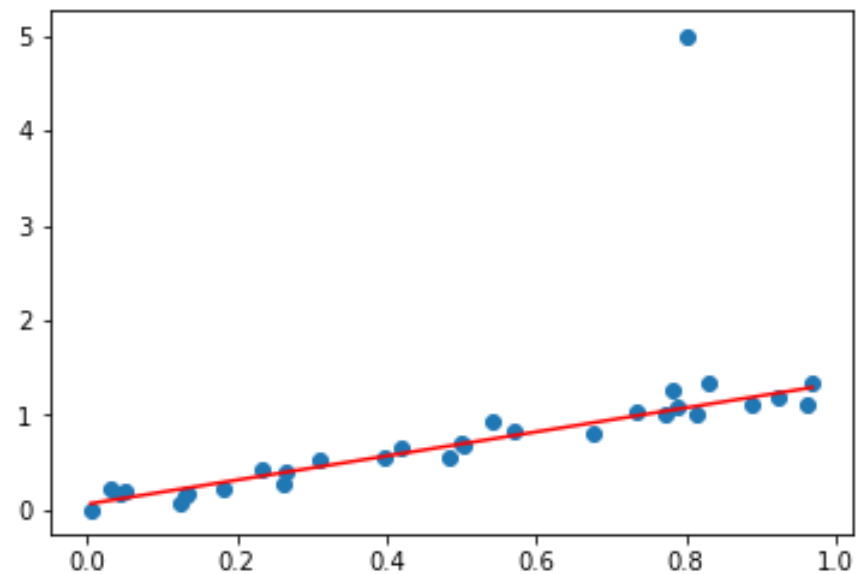
- Функционал ошибки — средняя абсолютная ошибка (mean absolute error, MAE)

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

Выбросы



Обучение на MSE



Обучение на MAE

Выбросы

$a(x)$	y	$ a(x) - y $
2	1	1
1	2	1
2	3	1
5	4	1
6	5	1
7	100	93
6	7	1

$$MAE \approx 14.14$$

Выбросы

$a(x)$	y	$ a(x) - y $
4	1	3
5	2	3
6	3	3
7	4	3
8	5	3
10	100	90
10	7	3

$$MAE \approx 15.43$$

Функция потерь Хубера

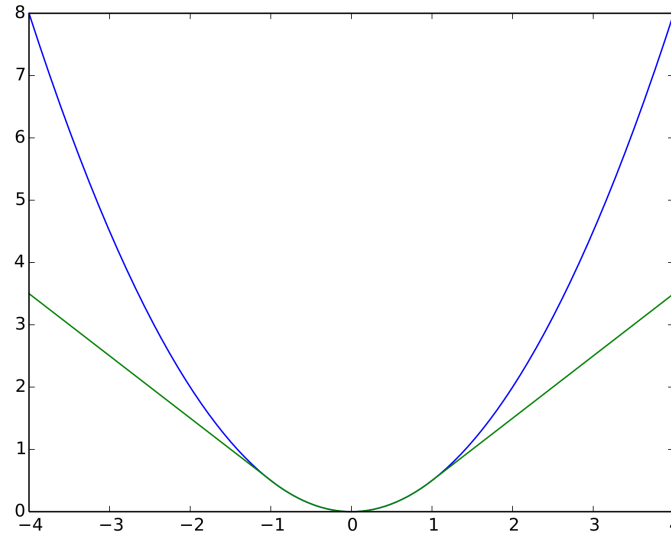
$$L_H(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & |y - a| < \delta \\ \delta \left(|y - a| - \frac{1}{2}\delta \right), & |y - a| \geq \delta \end{cases}$$

- Функционал ошибки:

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

Функция потерь Хубера

$$L_H(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & |y - a| < \delta \\ \delta \left(|y - a| - \frac{1}{2}\delta \right), & |y - a| \geq \delta \end{cases}$$



MAPE

- Mean Absolute Percentage Error (средний модуль относительной ошибки)

$$L(y, a) = \left| \frac{y - a}{y} \right|$$

$$Q(a, X) = \frac{100\%}{\ell} \sum_{i=1}^{\ell} \left| \frac{a(x_i) - y_i}{y_i} \right|$$

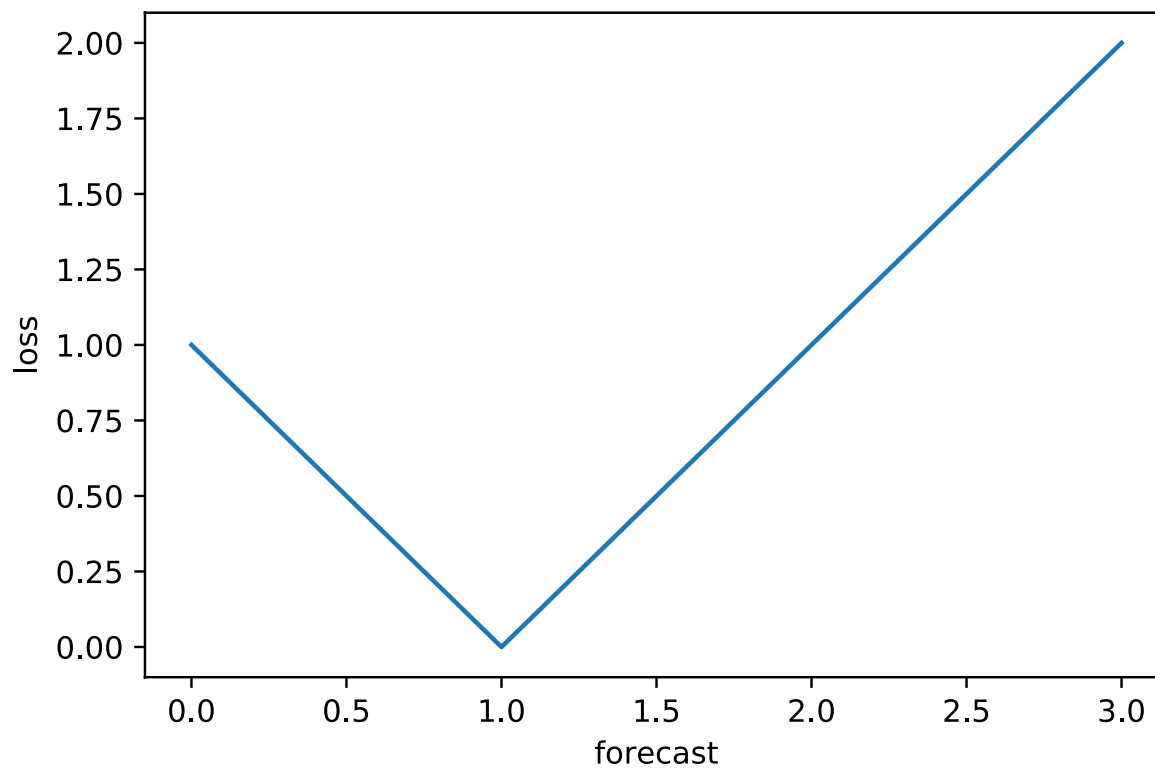
MAPE

$$L(y, a) = \left| \frac{y - a}{y} \right|$$

- Особенности (при $a \geq 0$):
- Недопрогноз штрафует максимум на единицу
- Перепрогноз может быть оштрафован любым числом
- Несимметричная функция потерь (отдаёт предпочтение недопрогнозу)

MAPE

$$L(y, a) = \left| \frac{y - a}{y} \right|$$



SMAPE

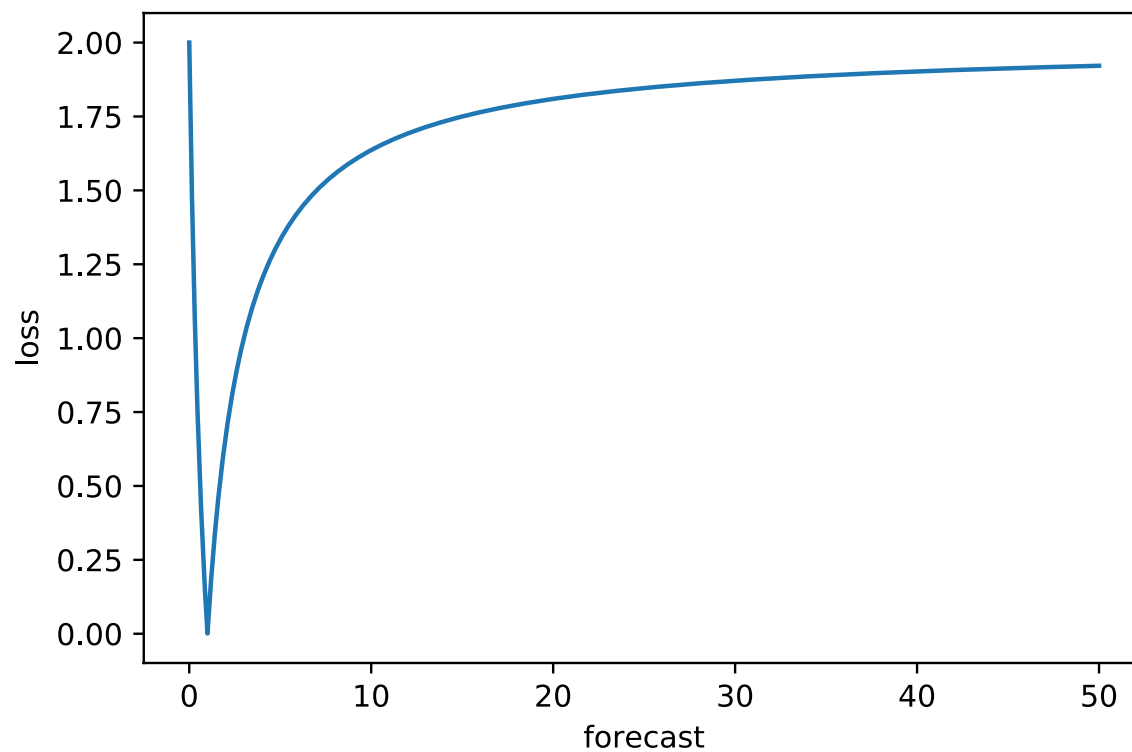
- Symmetric Mean Absolute Percentage Error (симметричный средний модуль относительной ошибки)

$$L(y, a) = \frac{|y - a|}{(|y| + |a|)/2}$$

$$Q(a, X) = \frac{100\%}{\ell} \sum_{i=1}^{\ell} \frac{|y_i - a(x_i)|}{(|y_i| + |a(x_i)|)/2}$$

SMAPE

$$L(y, a) = \frac{|y - a|}{(|y| + |a|)/2}$$



Модель линейной классификации

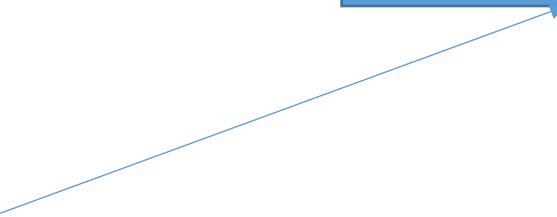
Классификация

- $\mathbb{Y} = \{-1, +1\}$
- -1 — отрицательный класс
- $+1$ — положительный класс
- $a(x)$ должен возвращать одно из двух чисел

Линейная регрессия

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

Вещественное
число!



Линейный классификатор

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

Линейный классификатор

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

Свободный
коэффициент

Веса

Признаки

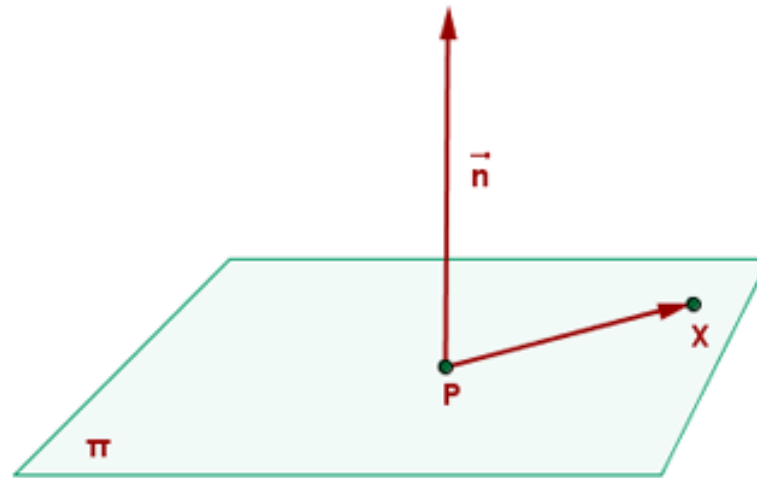
Линейный классификатор

- Будем считать, что есть единичный признак

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

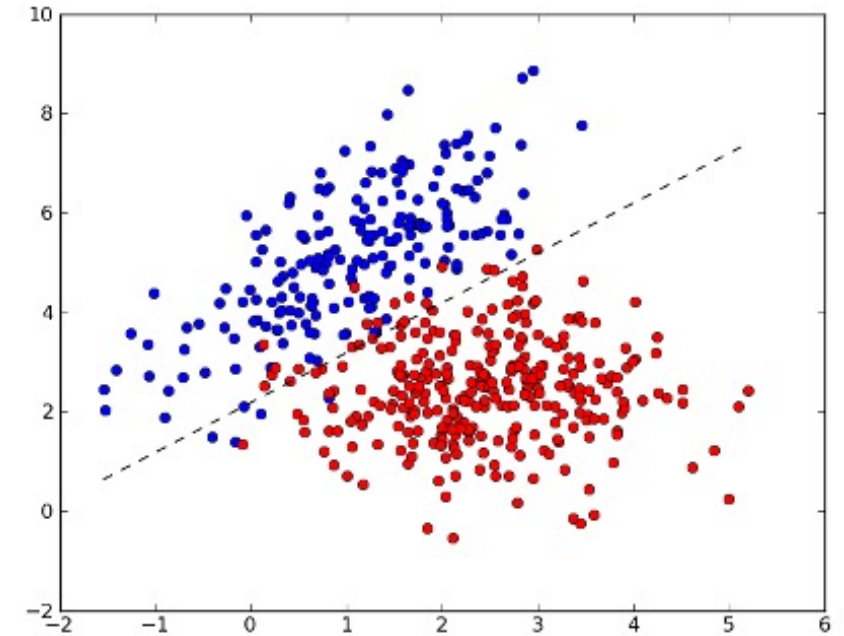
Геометрия линейного классификатора

Уравнение гиперплоскости: $\langle w, x \rangle = 0$



Геометрия линейного классификатора

- Линейный классификатор проводит гиперплоскость
- $\langle w, x \rangle < 0$ — объект «слева» от неё
- $\langle w, x \rangle > 0$ — объект «справа» от неё



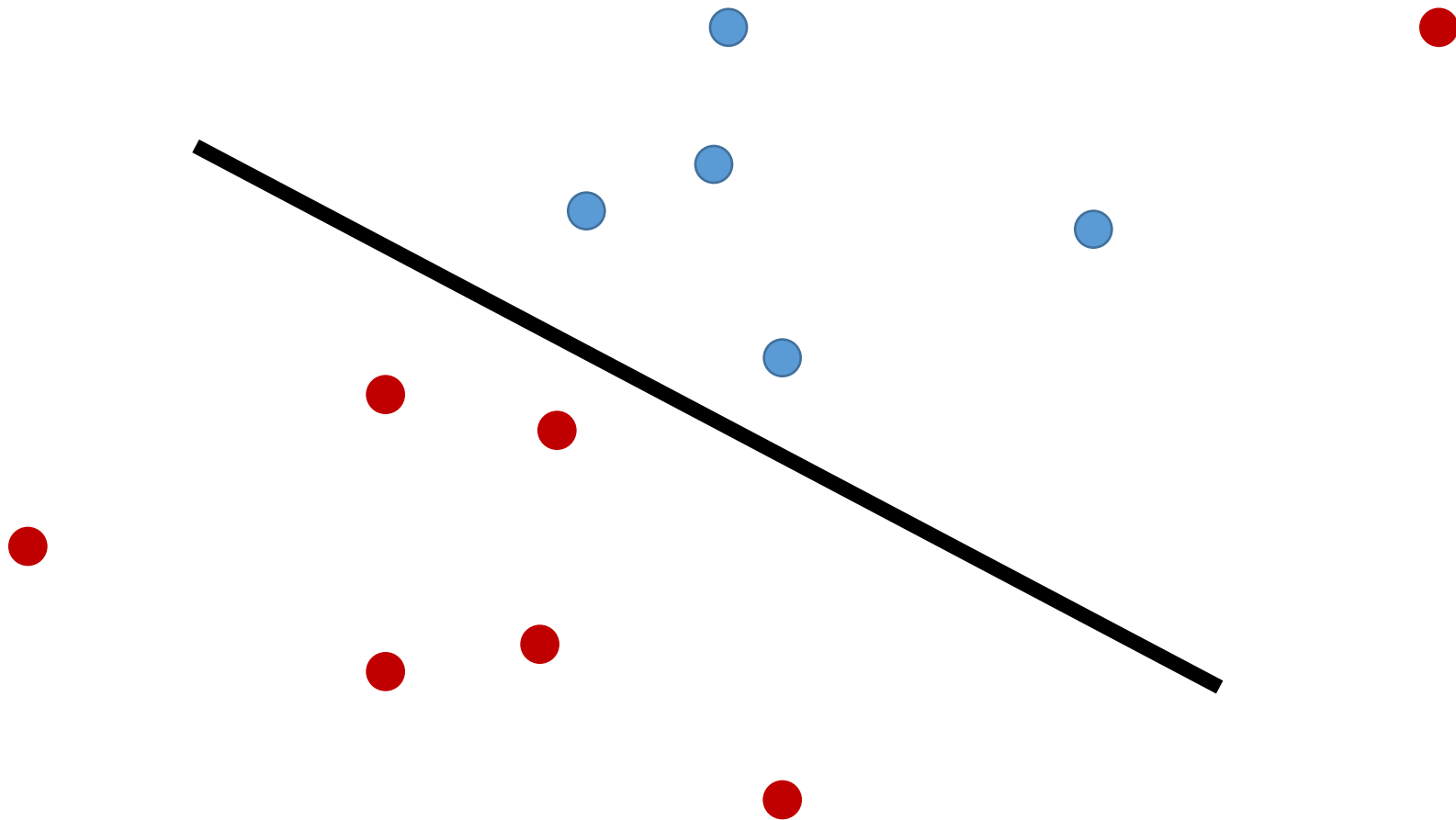
Геометрия линейного классификатора

- Расстояние от точки до гиперплоскости $\langle w, x \rangle = 0$:

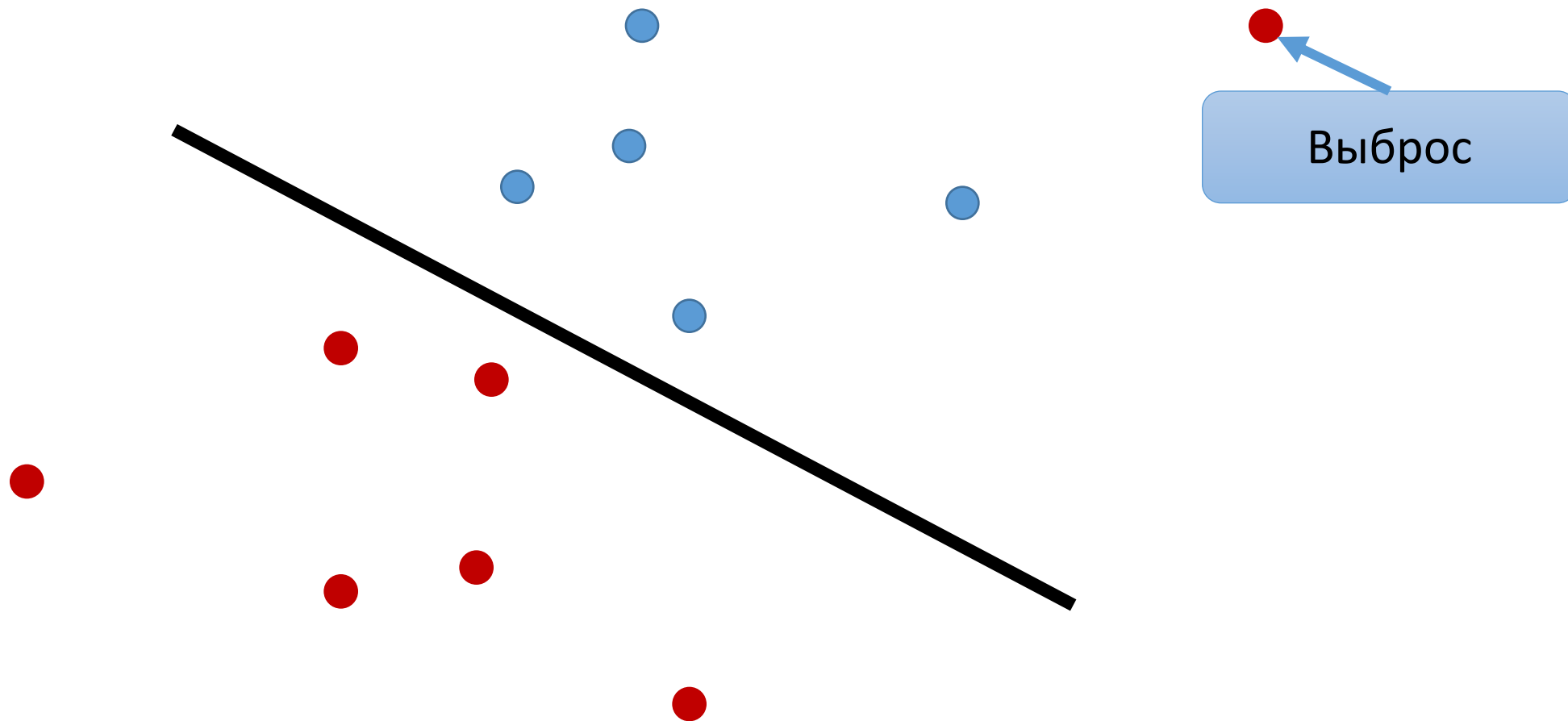
$$\frac{|\langle w, x \rangle|}{\|w\|}$$

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

Геометрия линейного классификатора

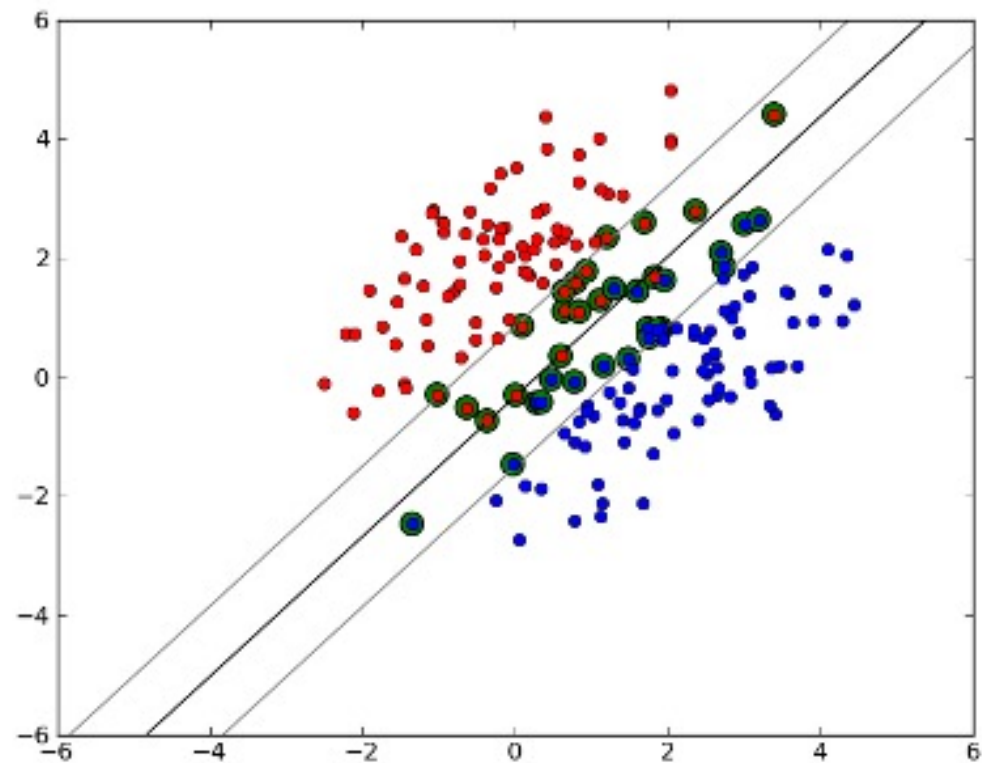


Геометрия линейного классификатора



Отступ

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



Порог

$$a(x) = \text{sign}(\langle w, x \rangle - t)$$

- t — порог классификатора
- Можно подбирать для оптимизации функции потерь, отличной от использованной при обучении

Линейный классификатор

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

Обучение линейных классификаторов

Функция потерь в классификации

- Частый выбор — бинарная функция потерь

$$L(y, a) = [a \neq y]$$

- Функционал ошибки — доля ошибок (error rate)

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

- Нередко измеряют долю верных ответов (accuracy):

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

Доля ошибок для линейного классификатора

- Функционал ошибки:

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\text{sign}(\langle w, x_i \rangle) \neq y_i]$$

- Индикатор — недифференцируемая функция

Отступы для линейного классификатора

- Функционал ошибки:

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\text{sign}(\langle w, x_i \rangle) \neq y_i]$$

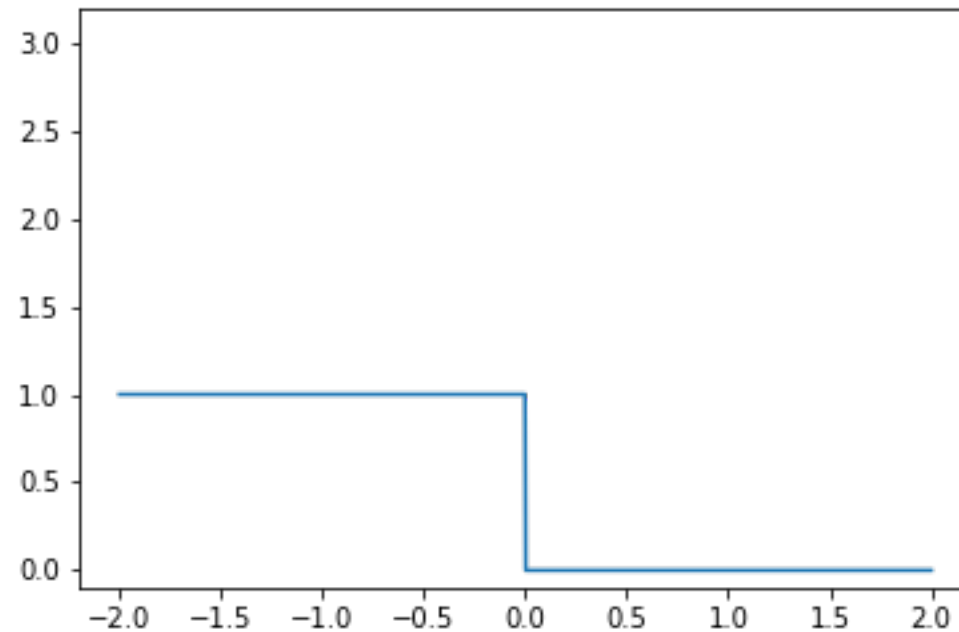
- Альтернативная запись:

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

Отступы для линейного классификатора

$$L(M) = [M < 0]$$

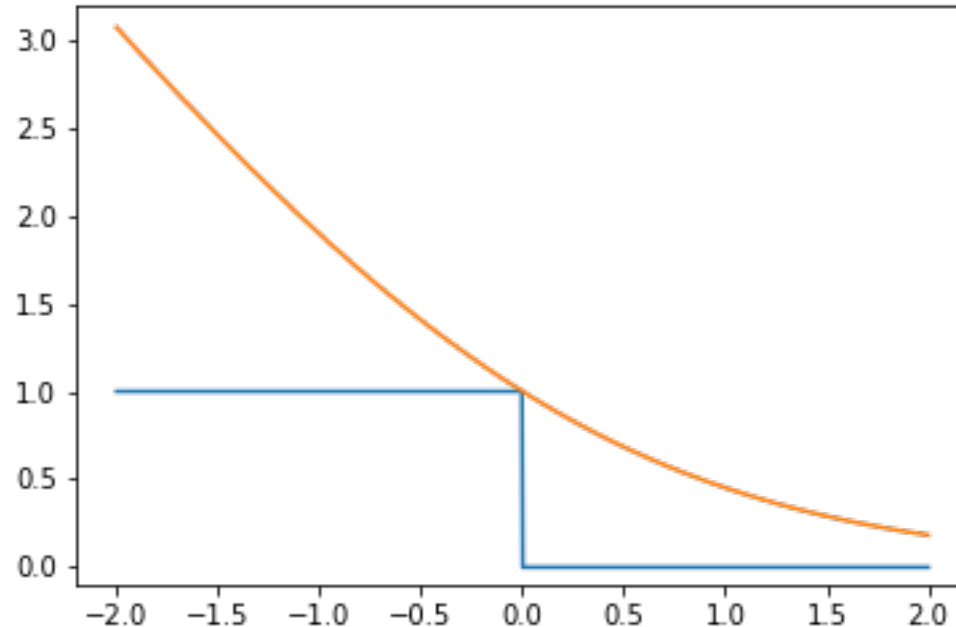
- Нельзя продифференцировать



Верхняя оценка

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

- Оценим сверху дифференцируемой функцией



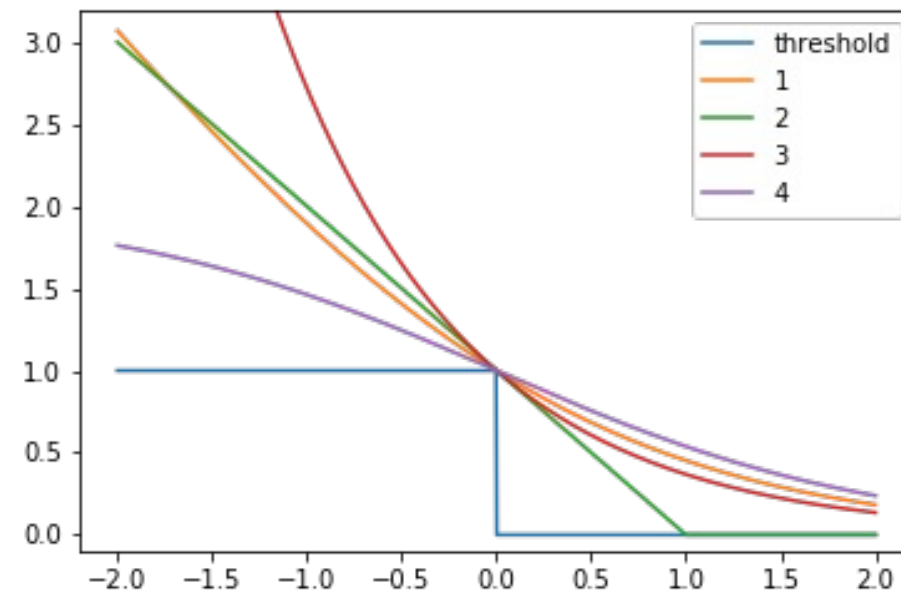
Верхняя оценка

$$0 \leq \frac{1}{\ell} \sum_{i=1}^{\ell} [y_i \langle w, x_i \rangle < 0] \leq \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(y_i \langle w, x_i \rangle) \rightarrow \min_w$$

- Минимизируем верхнюю оценку
- Надеемся, что она прижмёт долю ошибок к нулю

Примеры верхних оценок

1. $\tilde{L}(M) = \log(1 + e^{-M})$ — логистическая
2. $\tilde{L}(M) = \max(0, 1 - M)$ — кусочно-линейная
3. $\tilde{L}(M) = e^{-M}$ — экспоненциальная
4. $\tilde{L}(M) = \frac{2}{1+e^M}$ — сигмоидная



Пример обучения

- Выбираем логистическую функцию потерь:

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

- Вычисляем градиент:

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

Пример обучения

- Делаем градиентный спуск:

$$w^{(t)} = w^{(t-1)} + \eta \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{y_i x_i}{1 + \exp(y_i \langle w, x_i \rangle)}$$

Пример регуляризации

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

- Полностью аналогично линейной регрессии
- Важно не накладывать регуляризацию на свободный коэффициент
- Можно использовать L_1 -регуляризацию