

Машинное обучение

Лекция 6

Линейная классификация

Метрики качества

Сергей Корпачев

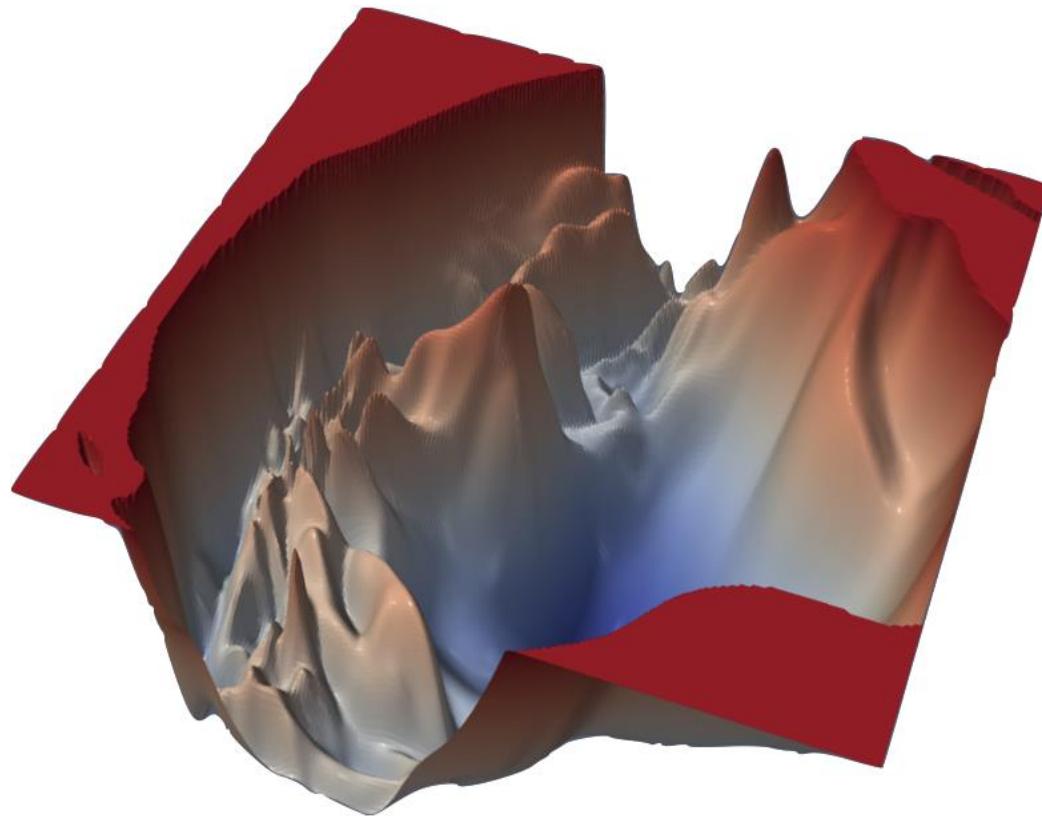
НИУ ВШЭ, 2026

Градиент и его свойства

Локальные минимумы

- Градиентный спуск находит **локальный минимум**
- Мультистарт — запуск градиентного спуска из разных начальных точек
- Может улучшить результат

Локальные минимумы



Длина шага

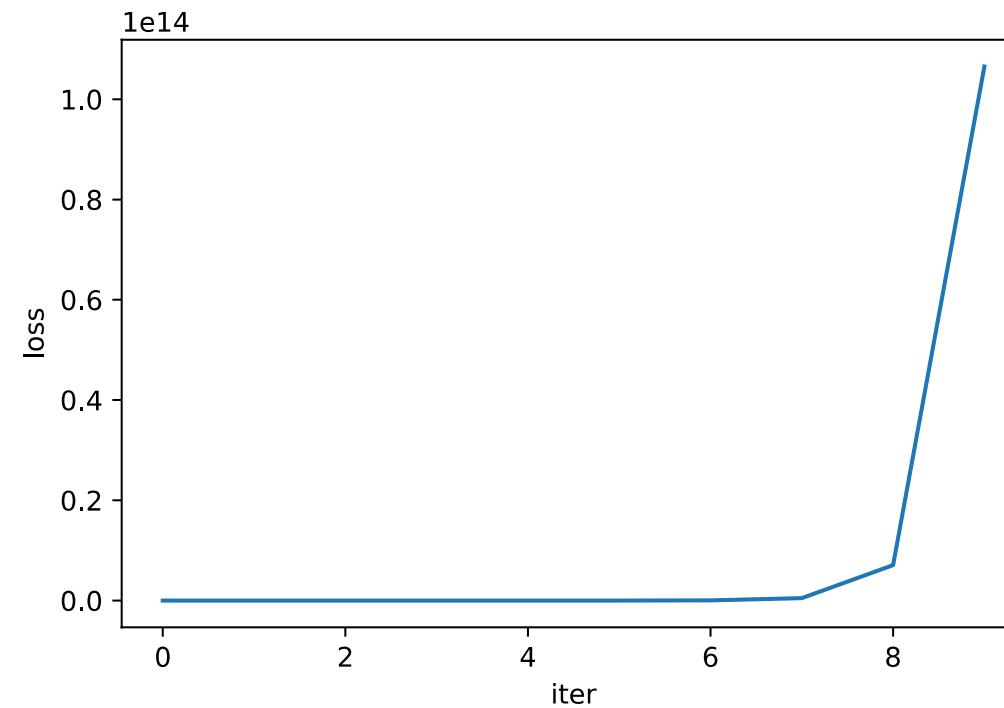
$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

- Позволяет контролировать скорость обучения

Длина шага

Градиент на первом шаге:

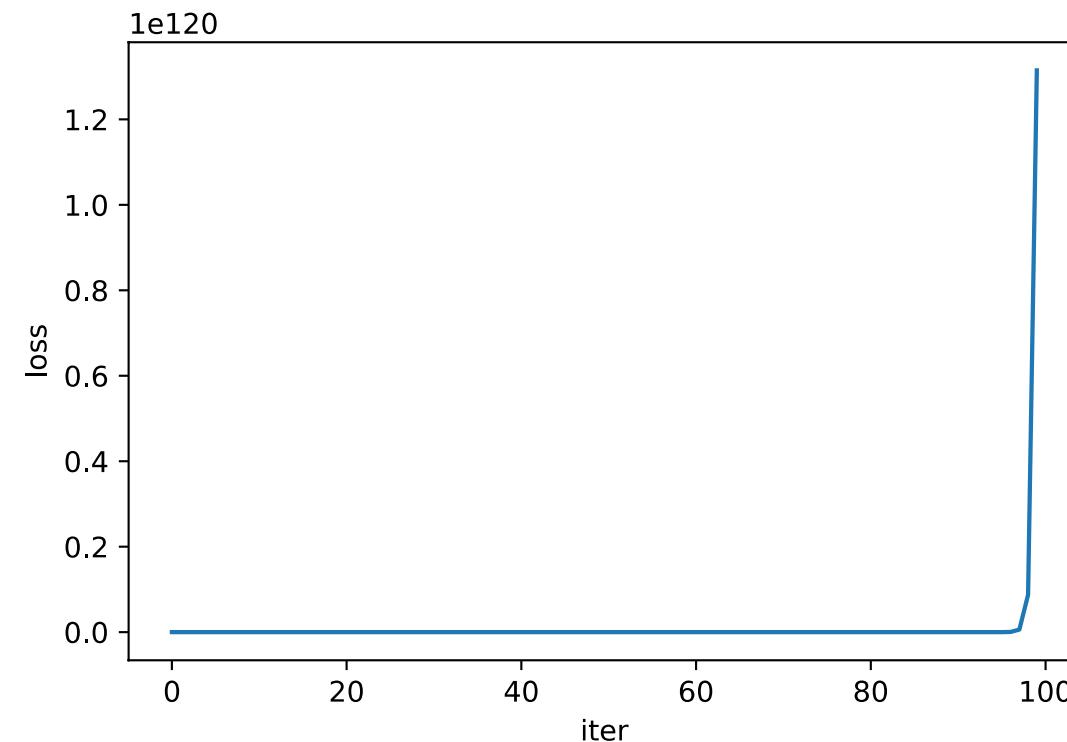
[26.52, 564.80, 682.90, 5097.71, 12110.87]



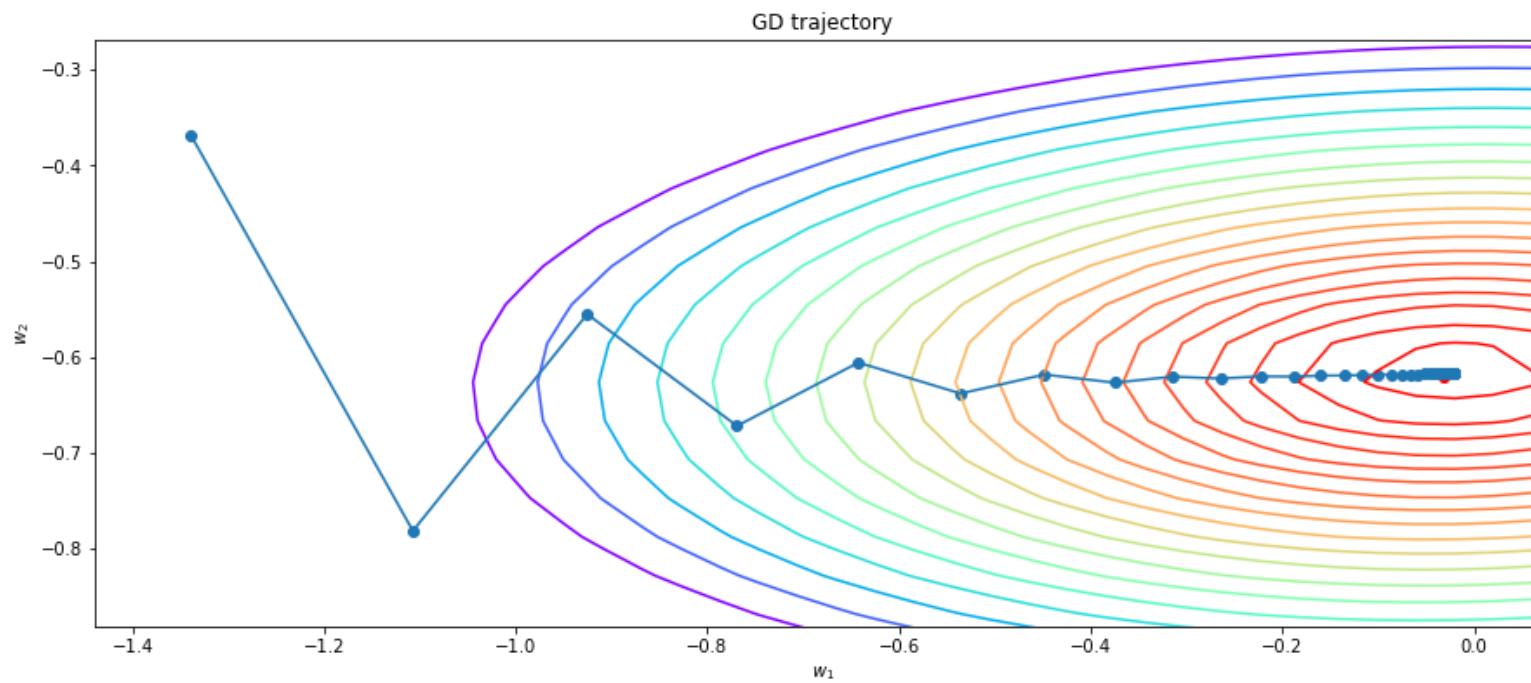
Длина шага

Градиент на первом шаге:

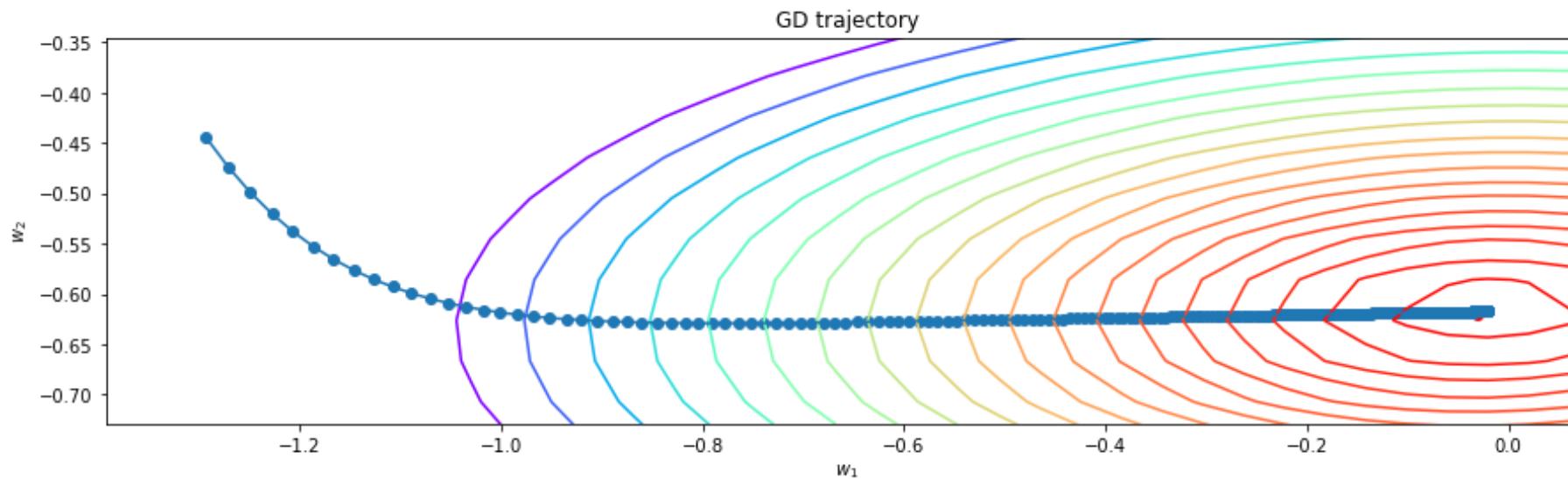
[26.52, 564.80, 682.90, 5097.71, 12110.87]



Длина шага



Длина шага



Длина шага

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

- Позволяет контролировать скорость обучения
- Если сделать длину шага недостаточно маленькой, градиентный спуск может разойтись
- Длина шага — параметр, который нужно подбирать

Переменная длина шага

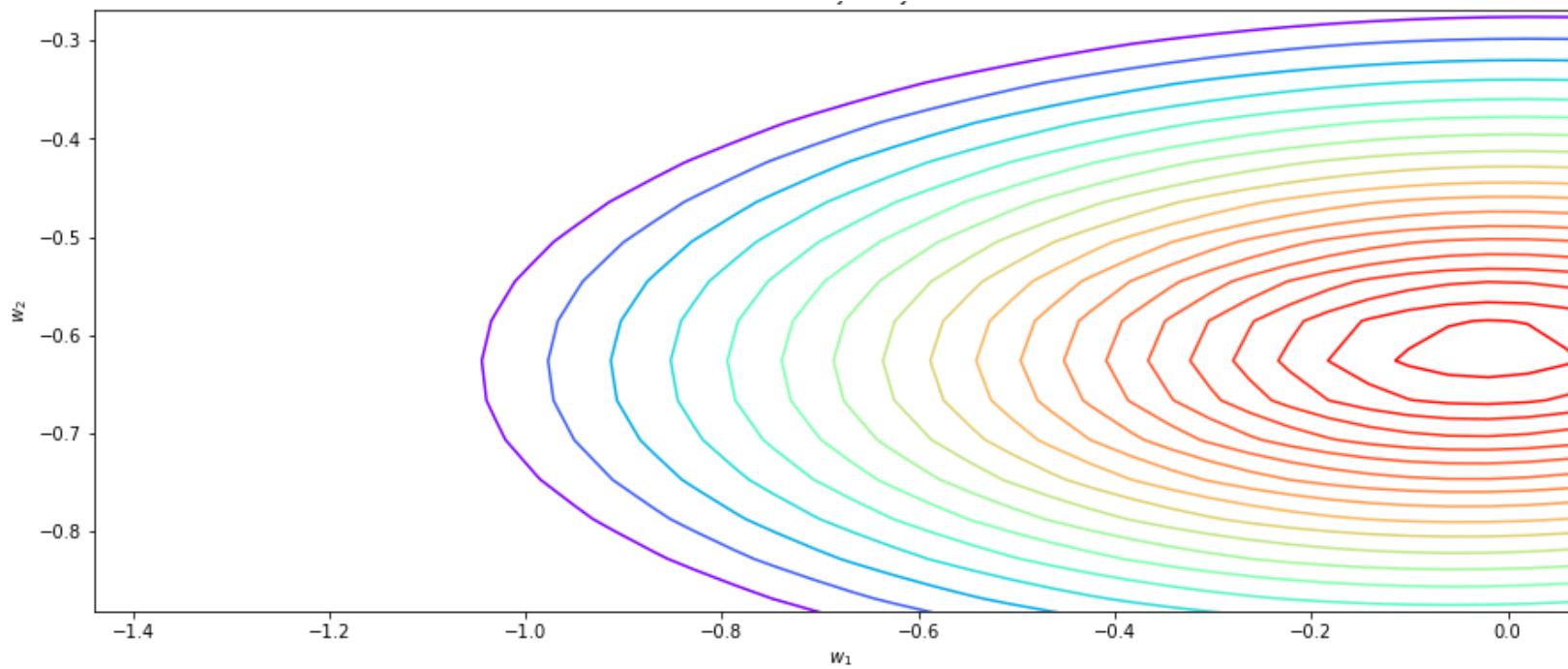
$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1})$$

- Длину шага можно менять в зависимости от шага
- Например: $\eta_t = \frac{1}{t}$
- Ещё вариант: $\eta_t = \lambda \left(\frac{s}{s+t} \right)^p$

Масштабирование признаков

```
[[  0.8194022 -11.97609413 -34.41655678   0.98167246 -34.14405489 ]
 [ -2.83614512  17.19489715   3.29562399  63.8054227  39.70301275 ]
 [  3.10906179  11.26049837   0.51404712  22.64032379 -28.62078735 ]
 [ ...,
 [ -3.61976507  17.63933655  31.65890573  22.5124188 -75.6386039  ]
 [ -1.98472285  3.98588887  29.6135414   -11.11816  33.98746403 ]
 [ -3.34136103 -12.81955782 -19.5542601   12.62435442  50.24876879 ]]
```

Масштабирование признаков



Масштабирование признаков

- Вычтем из каждого значения признака среднее и поделим на стандартное отклонение:

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

Стохастический градиентный спуск

Градиентный спуск

1. Начальное приближение: w^0

2. Повторять:

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

3. Останавливаемся, если

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

Линейная регрессия

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

- $\frac{\partial Q}{\partial w_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_{i1} (\langle w, x \rangle - y_i)$
- ...
- $\frac{\partial Q}{\partial w_d} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_{id} (\langle w, x \rangle - y_i)$
- $\nabla Q(w) = \frac{2}{\ell} X^T (Xw - y)$

Сложности градиентного спуска

- Для вычисления градиента, как правило, надо просуммировать что-то по всем объектам
- И это для одного маленького шага!

Оценка градиента

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

- Градиент:

$$\nabla Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} \nabla L(y_i, a(x_i))$$

- Может, оценить градиент одним слагаемым?

$$\nabla Q(w) \approx \nabla L(y_i, a(x_i))$$

Стохастический градиентный спуск

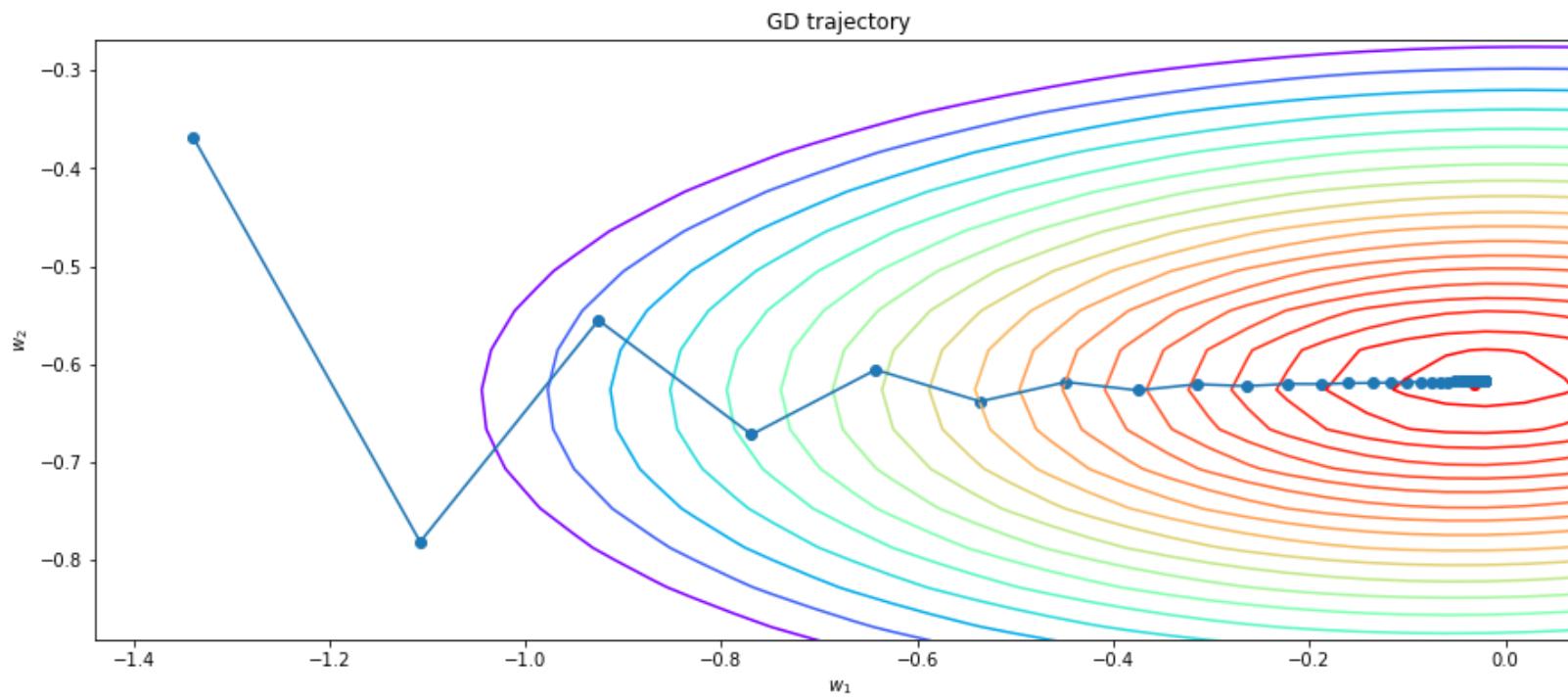
1. Начальное приближение: w^0
2. Повторять, каждый раз выбирая случайный объект i_t :

$$w^t = w^{t-1} - \eta \nabla L(y_{i_t}, a(x_{i_t}))$$

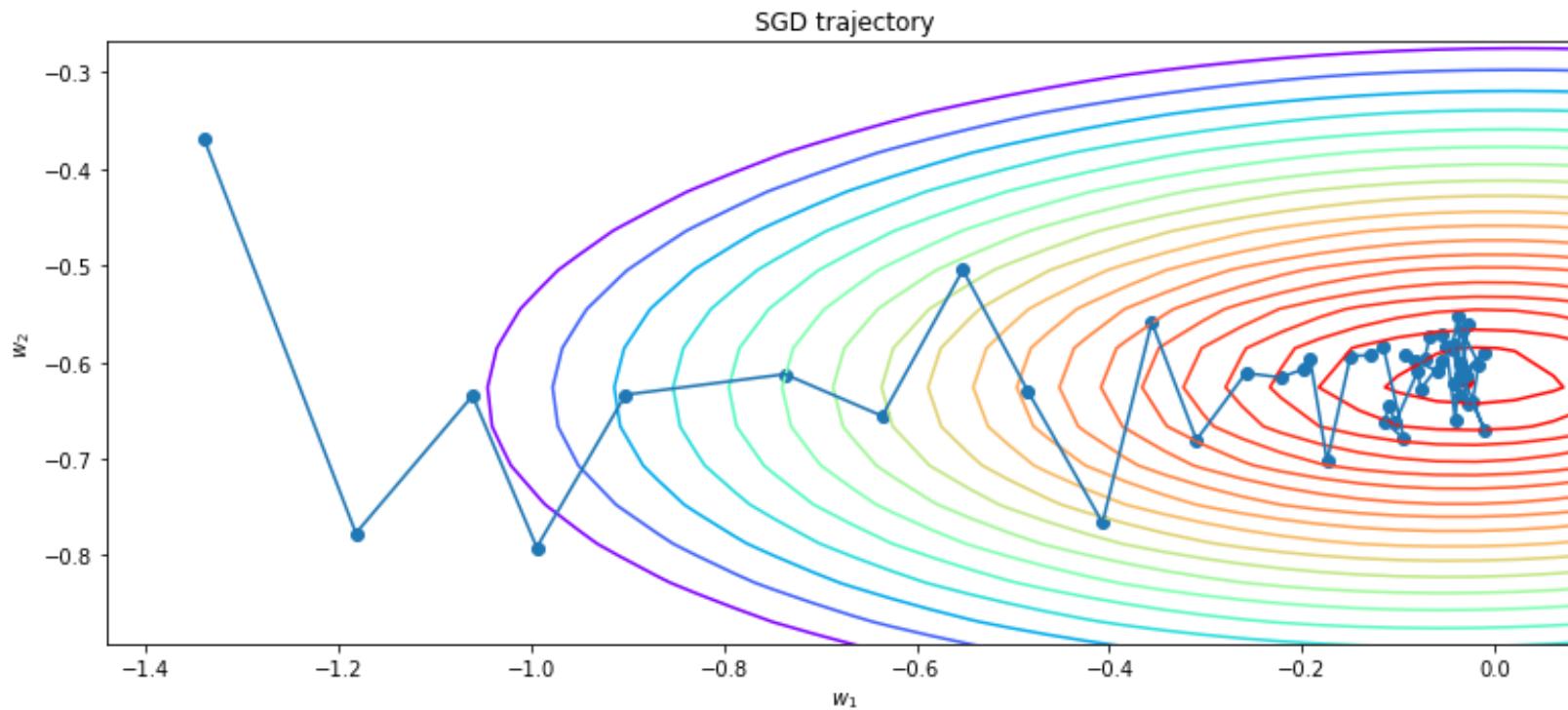
3. Останавливаемся, если

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

Градиентный спуск



Стochastic gradient descent



Стохастический градиентный спуск

1. Начальное приближение: w^0
2. Повторять, каждый раз выбирая случайный объект i_t :

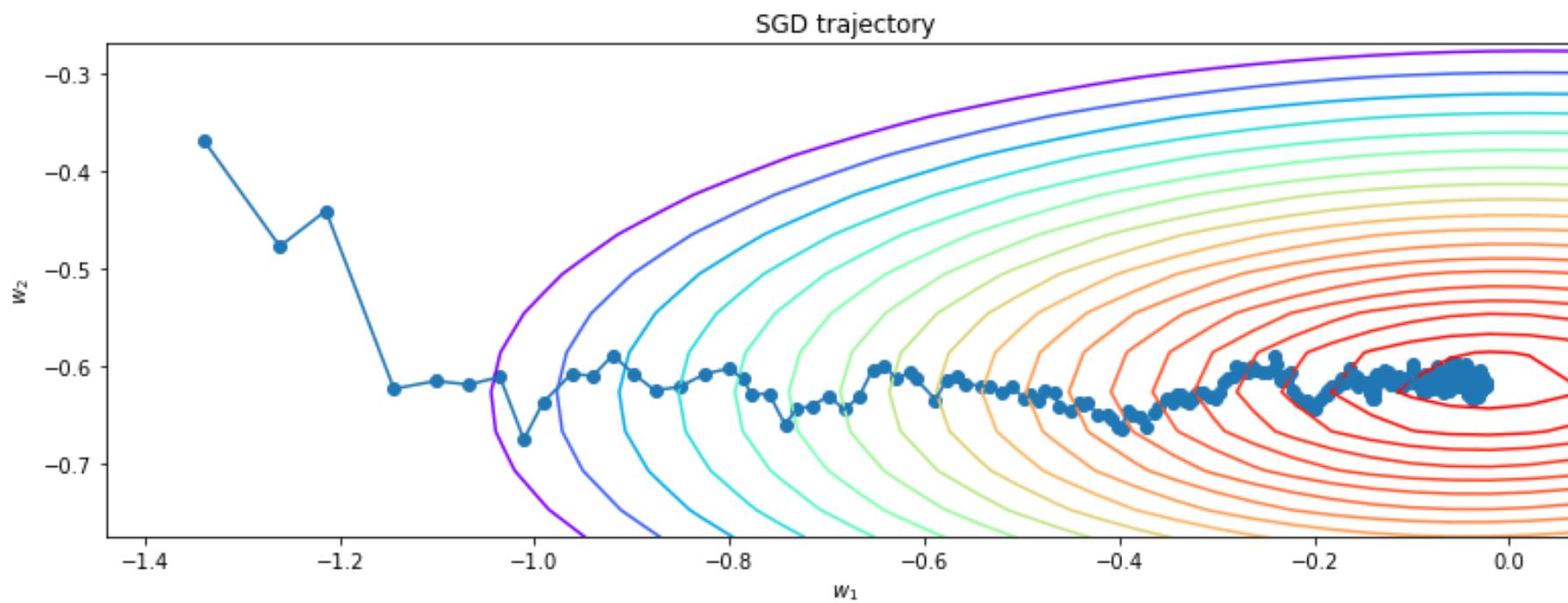
$$w^t = w^{t-1} - \eta_t \nabla L(y_{i_t}, a(x_{i_t}))$$

3. Останавливаемся, если

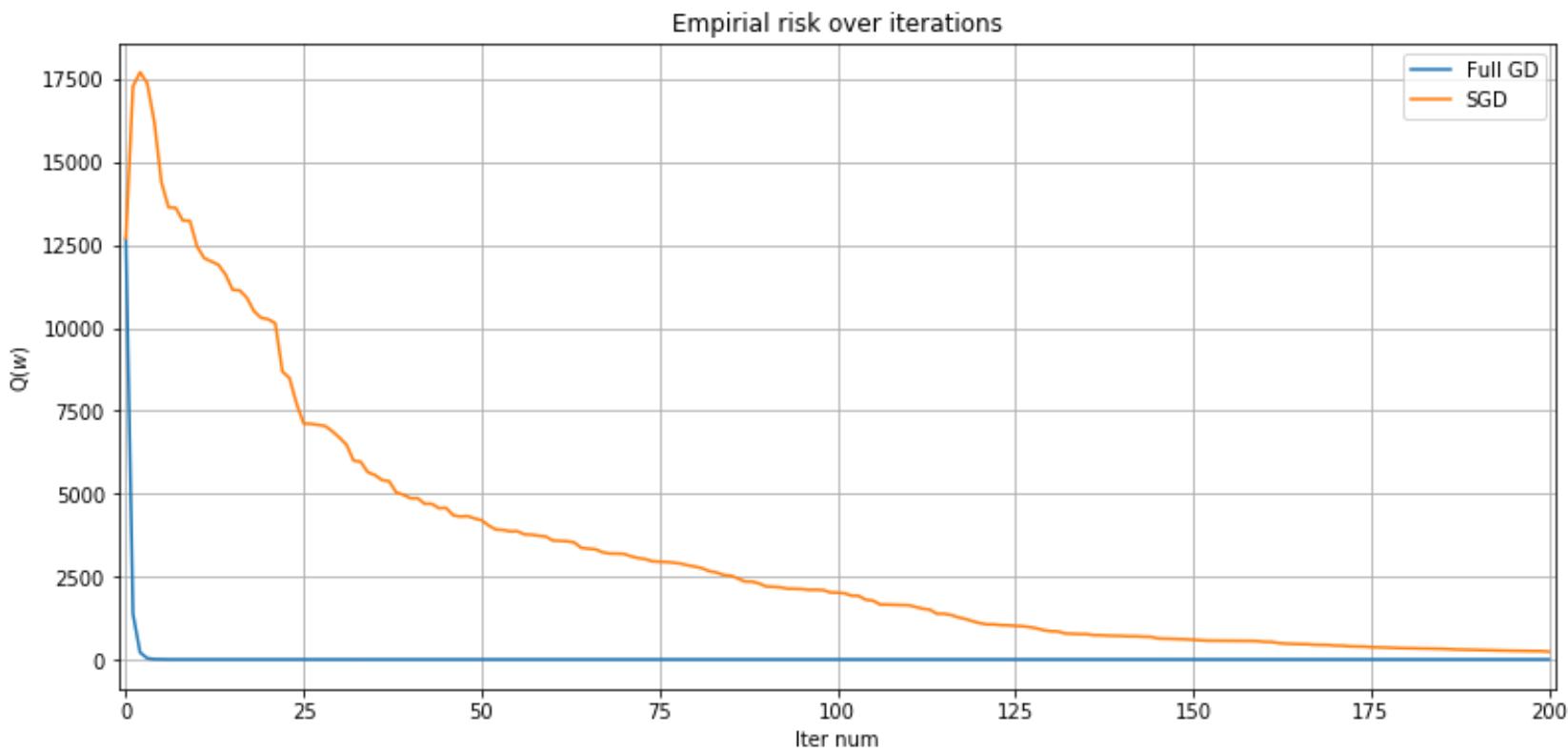
$$\|w^t - w^{t-1}\| < \varepsilon$$

Стochastic gradient descent

$$\eta_t = \frac{0.1}{t^{0.3}}$$



Стochastic gradient descent



Mini-batch

1. Начальное приближение: w^0
2. Повторять, каждый раз выбирая m случайных объектов i_1, \dots, i_m :

$$w^t = w^{t-1} - \eta_t \frac{1}{m} \sum_{j=1}^m \nabla L\left(y_{i_j}, a(x_{i_j})\right)$$

3. Останавливаемся, если

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

Функции потерь в задачах регрессии

Среднеквадратичная ошибка

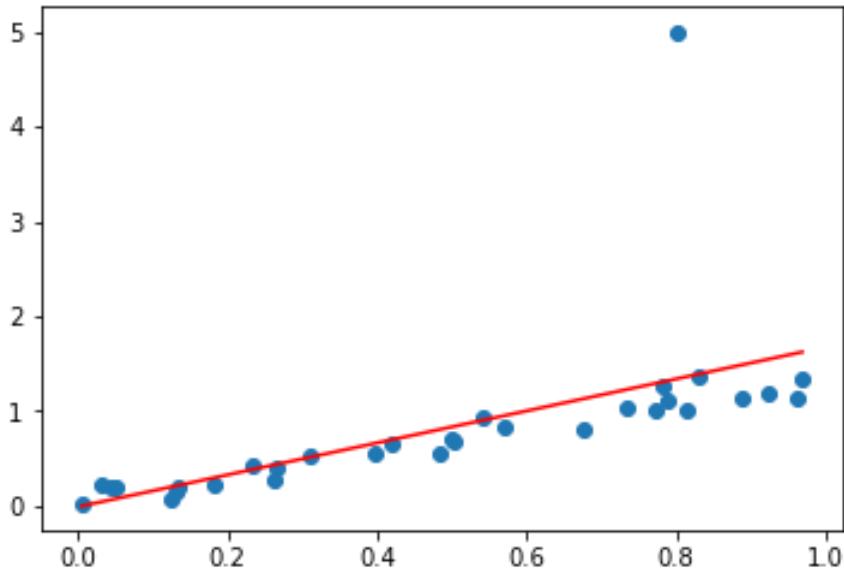
- Частый выбор — квадратичная функция потерь

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

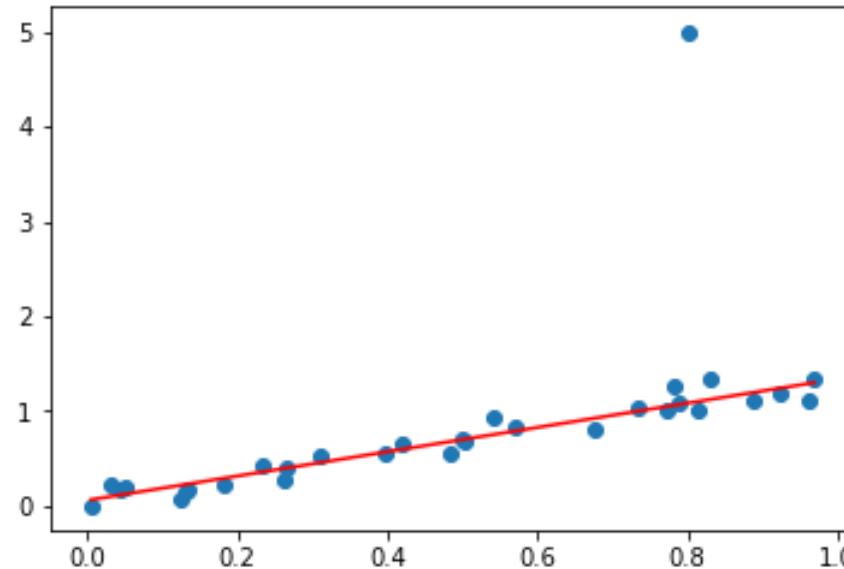
- Функционал ошибки — среднеквадратичная ошибка (mean squared error, MSE)

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

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

$$MSE \approx 1164$$

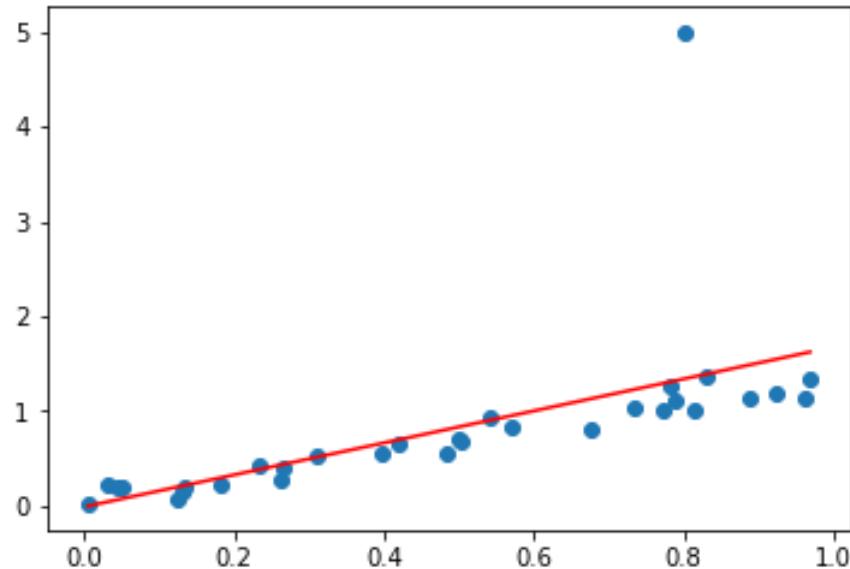
Средняя абсолютная ошибка

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

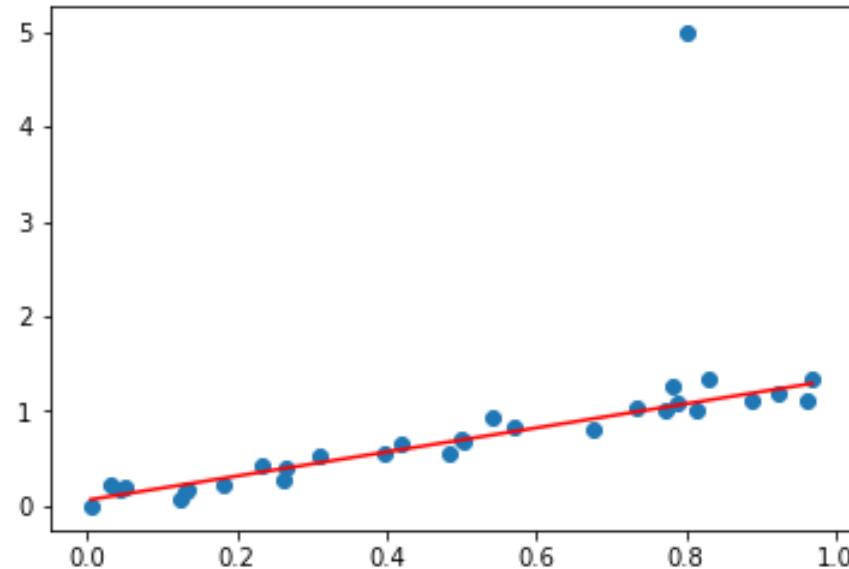
- Функционал ошибки — средняя абсолютная ошибка (mean absolute error, MAE)

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

Выбросы



Обучение на MSE



Обучение на МАЕ

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

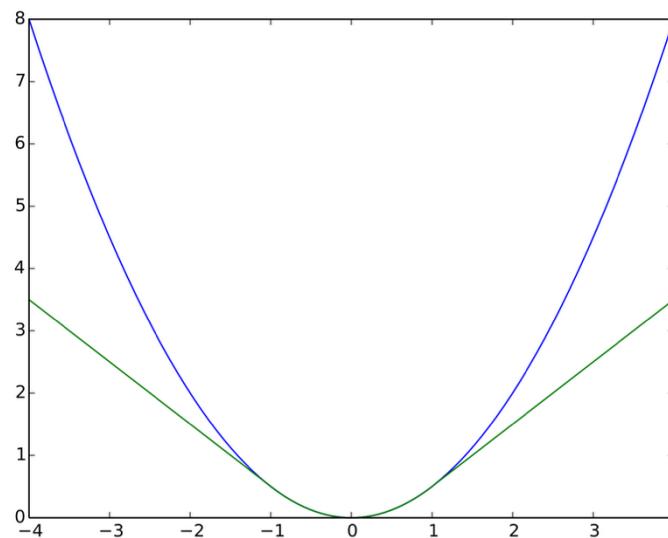
$$L_H(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & |y - a| < \delta \\ \delta \left(|y - a| - \frac{1}{2}\delta \right), & |y - a| \geq \delta \end{cases}$$

- 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)

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

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

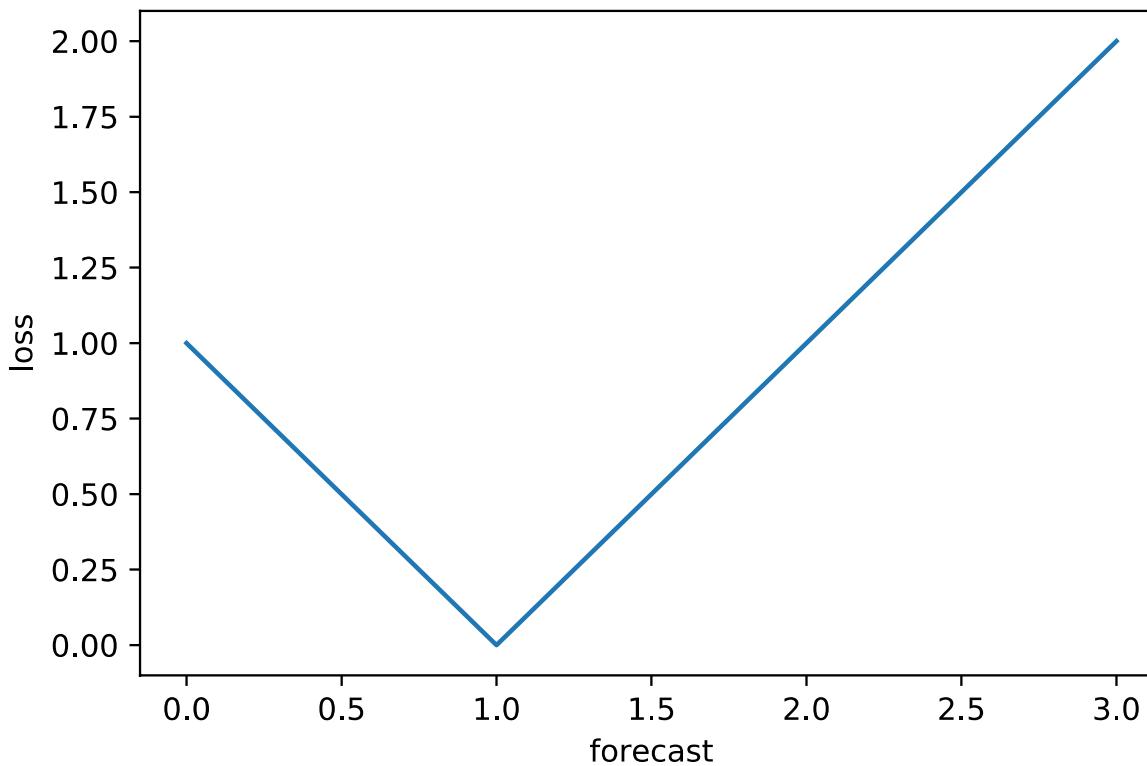
MAPE

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

- Особенности (при $a \geq 0$):
 - Недопрогноз штрафуется 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

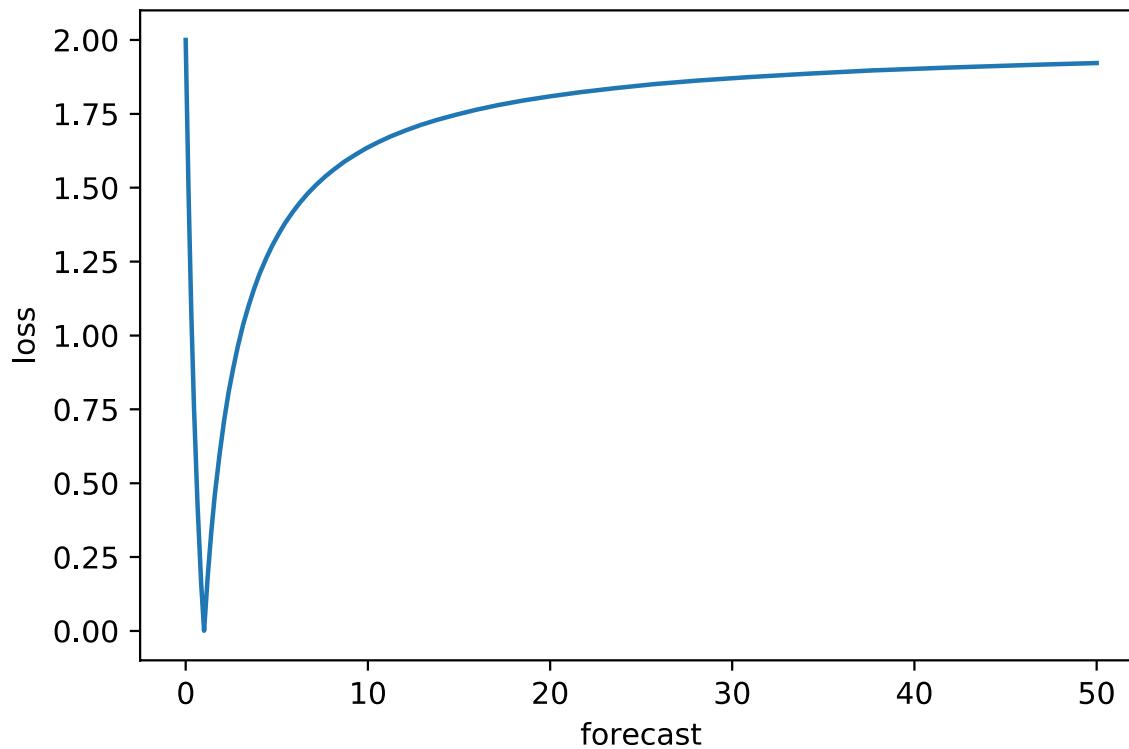
- Symmetric Mean Absolute Percentage Error (симметричный средний модуль относительной ошибки)

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

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

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

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

The diagram illustrates the components of the linear classifier equation. It features three blue rounded rectangular boxes with black text: 'Свободный коэффициент' (Free coefficient) pointing to the term w_0 ; 'Веса' (Weights) pointing to the sum of terms involving weights w_j and features x_j ; and 'Признаки' (Features) pointing to the product of weights w_j and features x_j .

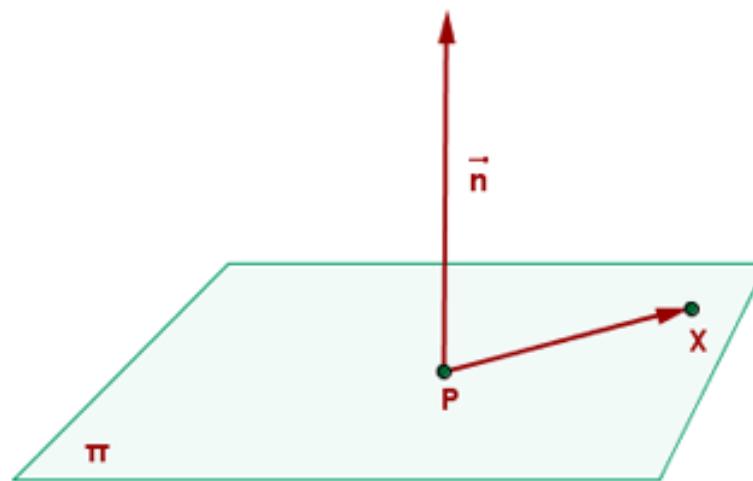
Линейный классификатор

- Будем считать, что есть единичный признак

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

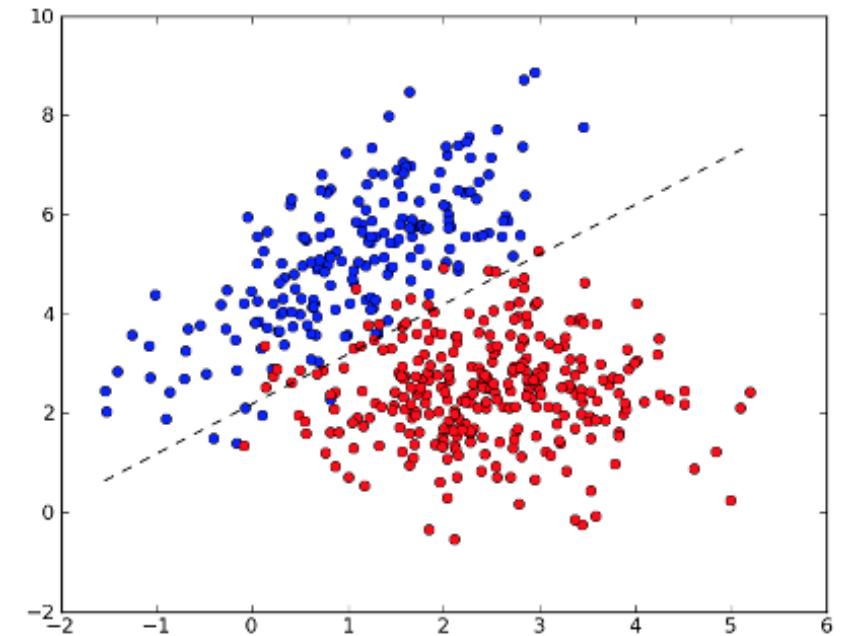
Геометрия линейного классификатора

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



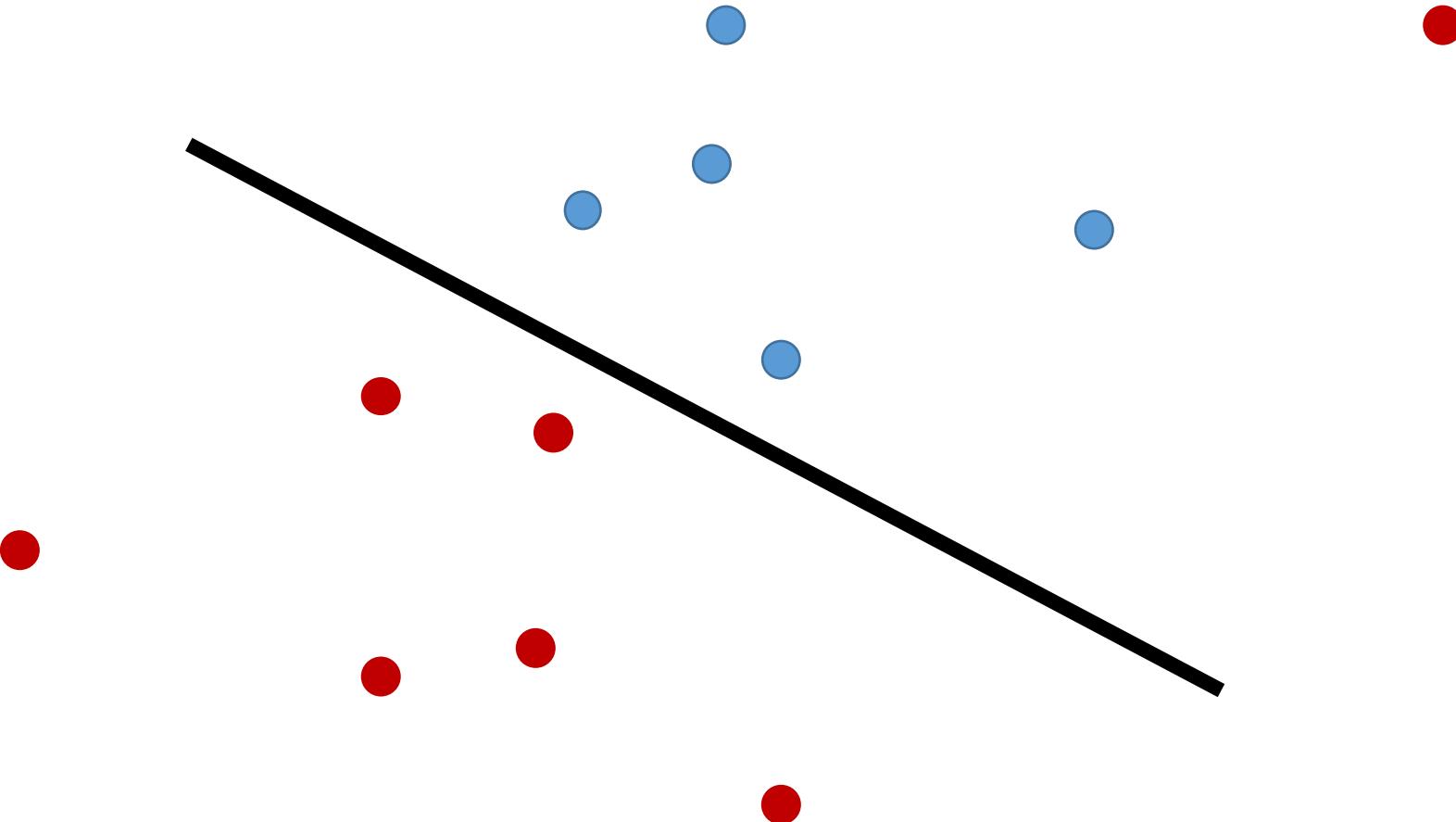
Геометрия линейного классификатора

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

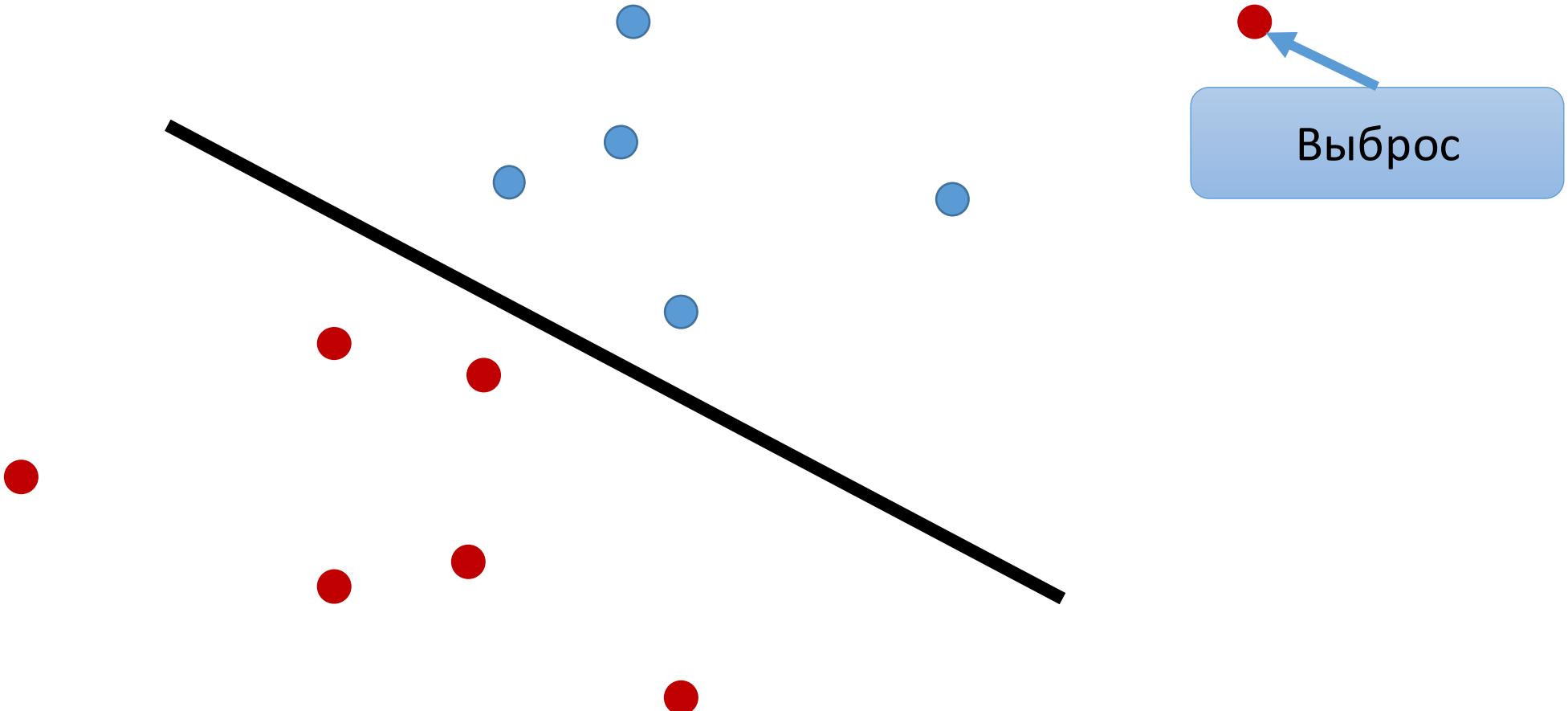
$$\frac{|\langle w, x \rangle|}{\|w\|}$$

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

Геометрия линейного классификатора

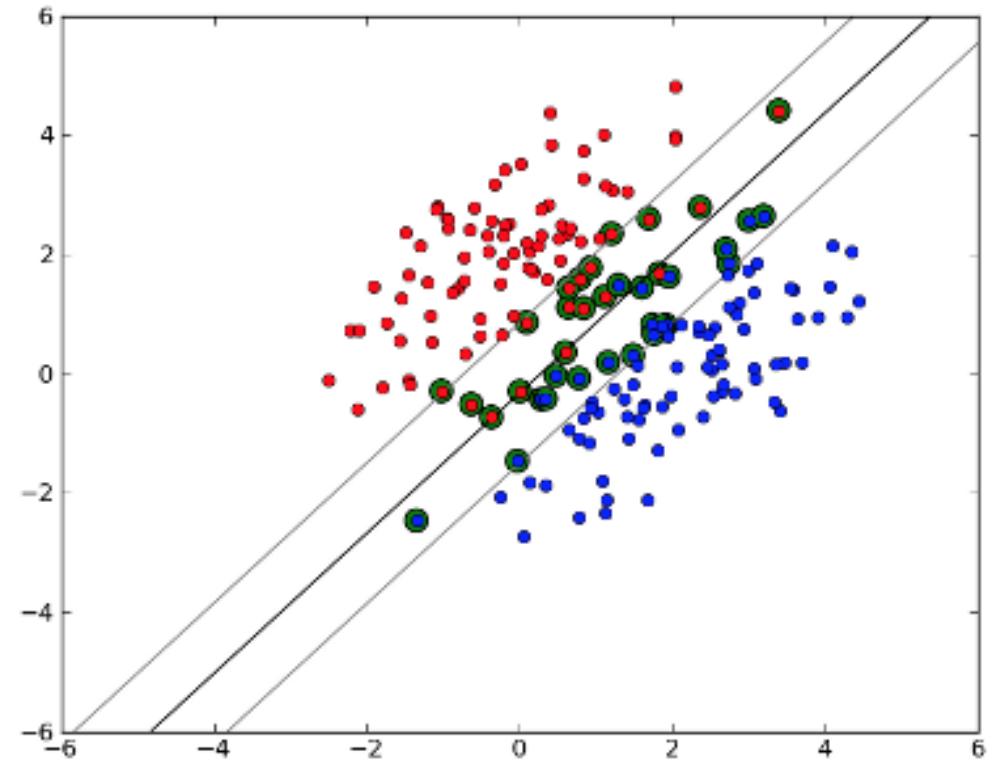


Геометрия линейного классификатора



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:

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

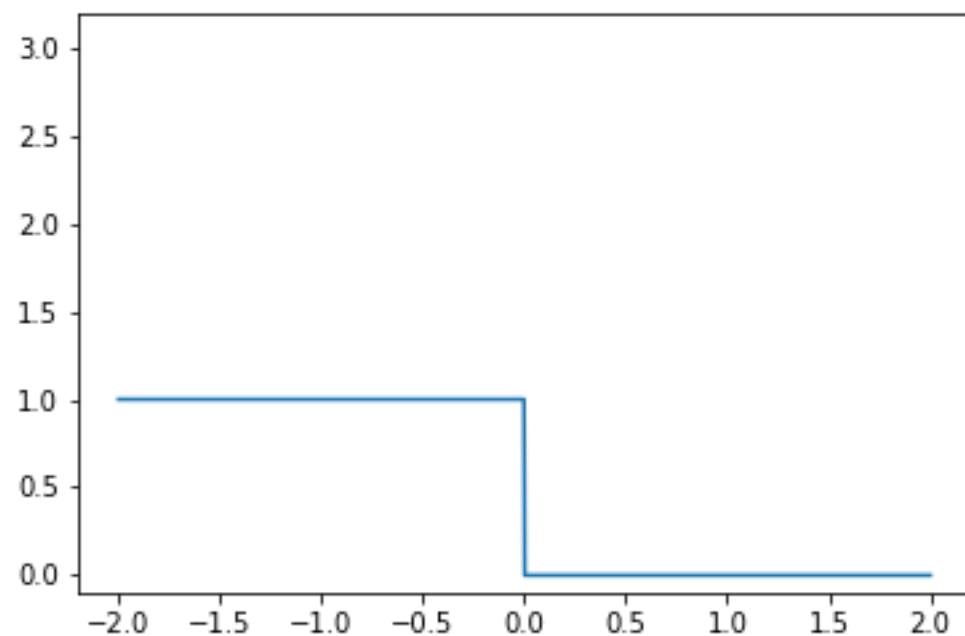
- Альтернативная запись:

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

Отступы для линейного классификатора

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

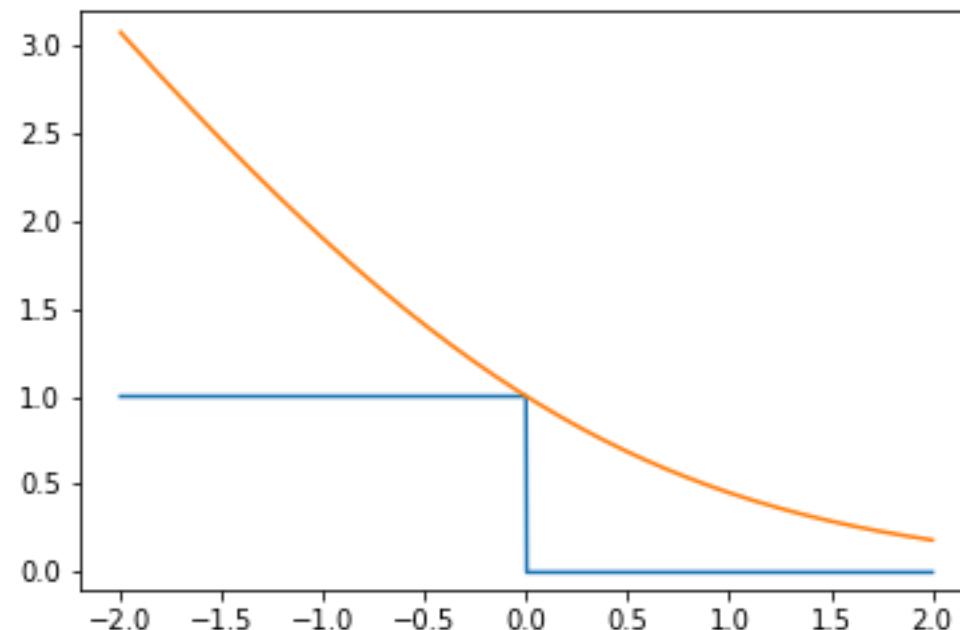
- Нельзя продифференцировать



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

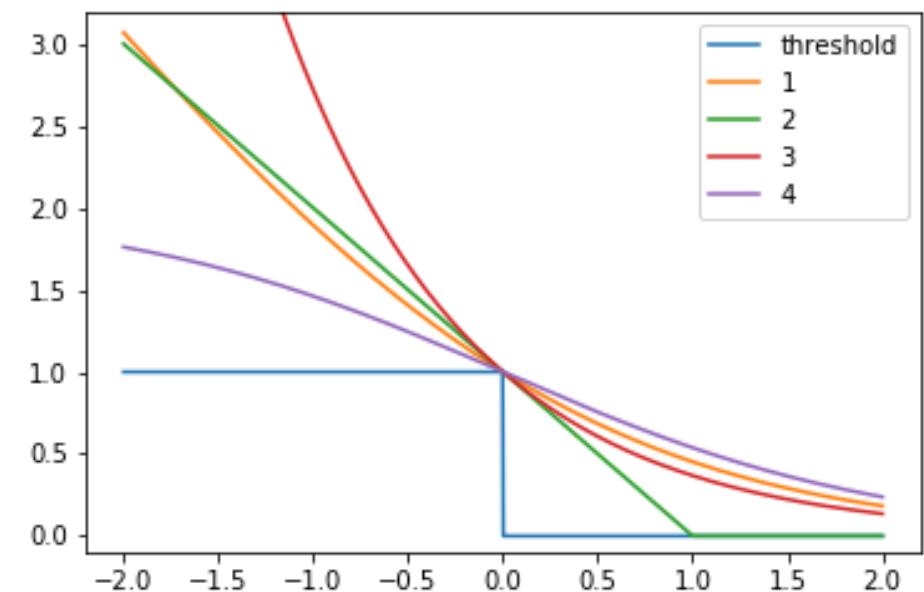
1. $\tilde{L}(M) = \log(1 + e^{-M})$ —

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

Метрики качества классификации

Качество классификации

- Доля неправильных ответов:

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

Качество классификации

- Доля правильных ответов (accuracy):

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

Несбалансированные выборки

- Несбалансированная выборка — объектов одного класса существенного больше
- Пример: предсказание кликов по рекламе
- Пример: медицинская диагностика
- Пример: предсказание оттока клиентов
- Пример: специализированный поиск

Несбалансированные выборки

- Пример:
 - Класс -1: 950 объектов
 - Класс +1: 50 объектов
- $a(x) = -1$
- Доля правильных ответов: 0.95
- Почему результат нас не устраивает?

Несбалансированные выборки

- Пример:
 - Класс -1: 950 объектов
 - Класс +1: 50 объектов
- $a(x) = -1$
- Доля правильных ответов: 0.95
- Почему результат нас не устраивает?
- Модель не несёт экономической ценности
- Цены ошибок неравнозначны

Несбалансированные выборки

- q_0 — доля объектов самого крупного класса
- Для разумных алгоритмов:

$$\text{accuracy} \in [q_0, 1]$$

- Если получили большой accuracy — посмотрите на баланс классов

Улучшение метрики

- Два алгоритма
- Доли правильных ответов: r_1 и r_2
- Абсолютное улучшение: $r_2 - r_1$
- Относительное улучшение: $\frac{r_2 - r_1}{r_1}$

Улучшение метрики

- | | | |
|------------------------------------|----------------------------------|-----------------------------------|
| • $r_1 = 0.8$ | • $r_1 = 0.5$ | • $r_1 = 0.001$ |
| • $r_2 = 0.9$ | • $r_2 = 0.75$ | • $r_2 = 0.01$ |
| • $\frac{r_2 - r_1}{r_1} = 12.5\%$ | • $\frac{r_2 - r_1}{r_1} = 50\%$ | • $\frac{r_2 - r_1}{r_1} = 900\%$ |

Цены ошибок

- Пример: кредитный scoring
- Модель 1:
 - 80 кредитов вернули
 - 20 кредитов не вернули
- Модель 2:
 - 48 кредитов вернули
 - 2 кредита не вернули
- Кто лучше?

Цены ошибок

- Что хуже?
 - Выдать кредит «плохому» клиенту
 - Не выдать кредит «хорошему» клиенту
- Доля верных ответов не учитывает цены ошибок

Матрица ошибок

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

Матрица ошибок

- Модель $a_1(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	80	20
$a(x) = -1$	20	80

- Модель $a_2(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	48	2
$a(x) = -1$	52	98

Точность (precision)

- Можно ли доверять классификатору при $a(x) = 1$?

$$\text{precision}(a, X) = \frac{TP}{TP + FP}$$

Точность (precision)

- Модель $a_1(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	80	20
$a(x) = -1$	20	80

- $\text{precision}(a_1, X) = 0.8$

- Модель $a_2(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	48	2
$a(x) = -1$	52	98

- $\text{precision}(a_2, X) = 0.96$

Полнота (recall)

- Как много положительных объектов находит классификатор?

$$\text{recall}(a, X) = \frac{TP}{TP + FN}$$

Полнота (recall)

- Модель $a_1(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	80	20
$a(x) = -1$	20	80

- $\text{recall}(a_1, X) = 0.8$

- Модель $a_2(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	48	2
$a(x) = -1$	52	98

- $\text{recall}(a_2, X) = 0.48$

Антифрод

- Классификация транзакций на нормальные и мошеннические
- Высокая точность, низкая полнота:
 - Редко блокируем нормальные транзакции
 - Пропускаем много мошеннических
- Низкая точность, высокая полнота:
 - Часто блокируем нормальные транзакции
 - Редко пропускаем мошеннические

Кредитный скоринг

- Неудачных кредитов должно быть не больше 5%
- Ограничение: $\text{precision}(a, X) \geq 0.95$
- Максимизируем полноту

Медицинская диагностика

- Надо найти не менее 80% больных
- Ограничение: $\text{recall}(a, X) \geq 0.8$
- Максимизируем точность

Несбалансированные выборки

- $\text{accuracy}(a, X) = 0.99$
- $\text{precision}(a, X) = 0.33$
- $\text{recall}(a, X) = 0.1$

	$y = 1$	$y = -1$
$a(x) = 1$	10	20
$a(x) = -1$	90	10000

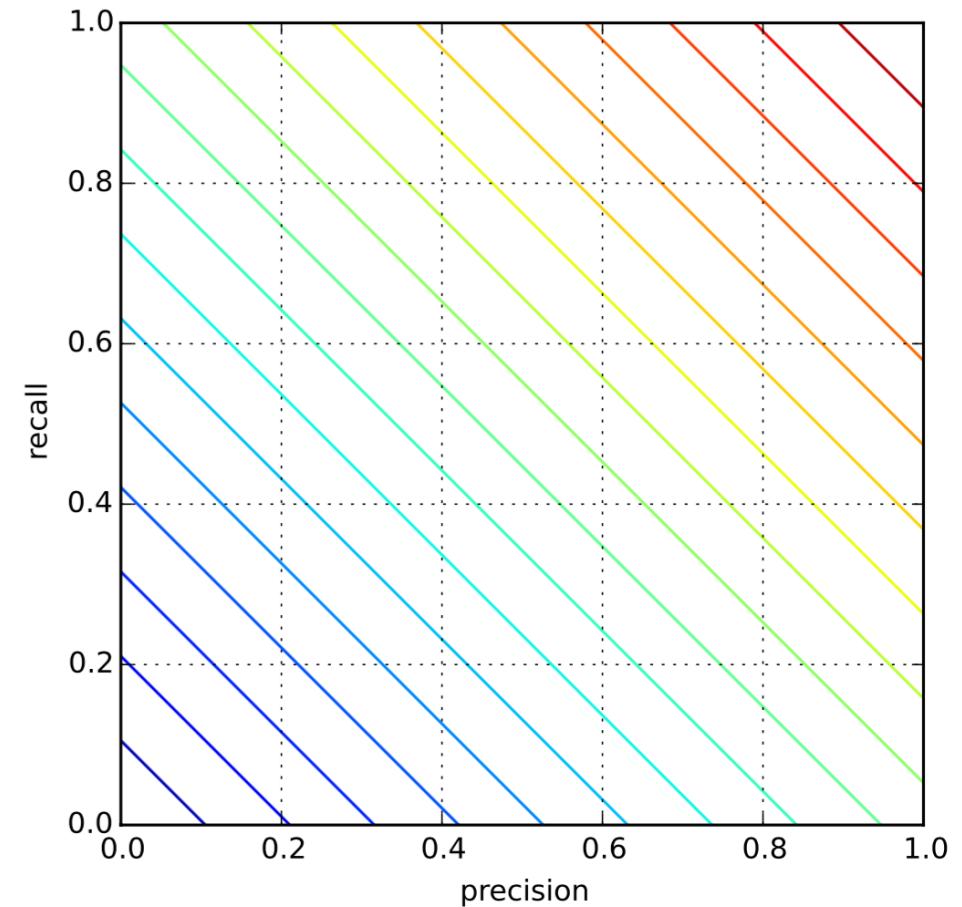
Совмещение точности и
полноты

Точность и полнота

- Точность — можно ли доверять классификатору при $a(x) = 1$?
- Полнота — как много положительных объектов находит $a(x)$?
- Оптимизировать две метрики одновременно очень неудобно
- Как объединить?

Арифметическое среднее

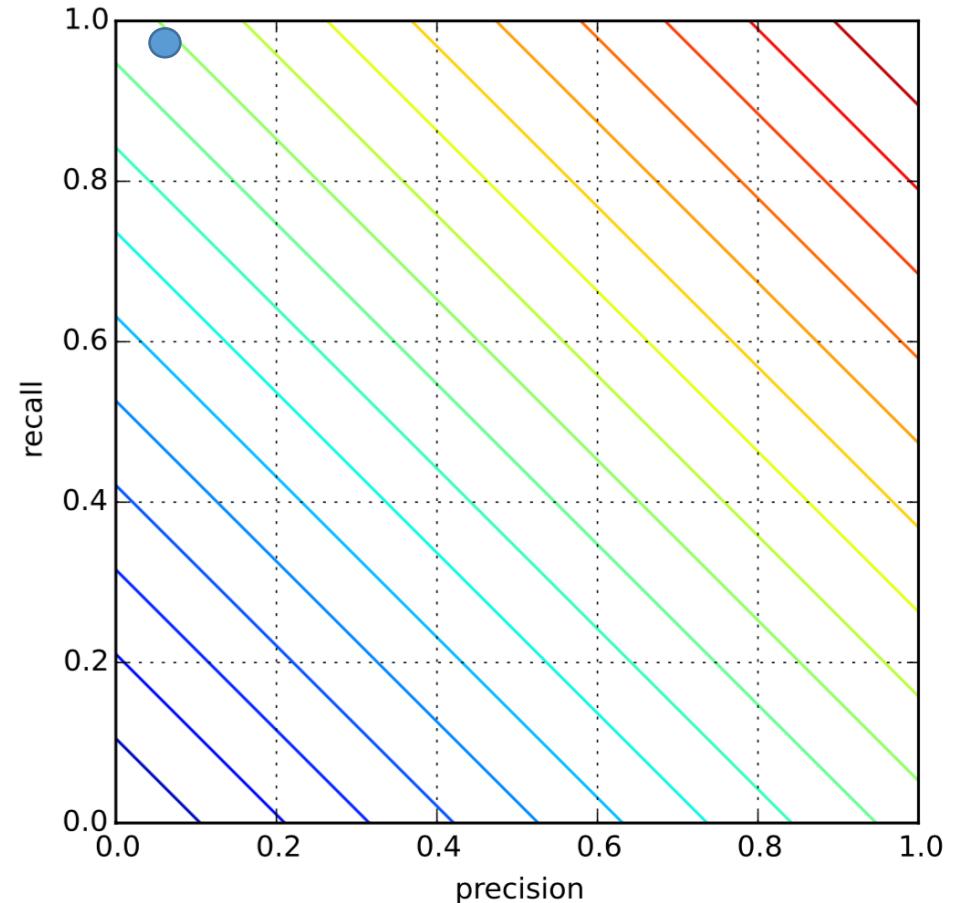
$$A = \frac{1}{2}(\text{precision} + \text{recall})$$



Арифметическое среднее

$$A = \frac{1}{2}(\text{precision} + \text{recall})$$

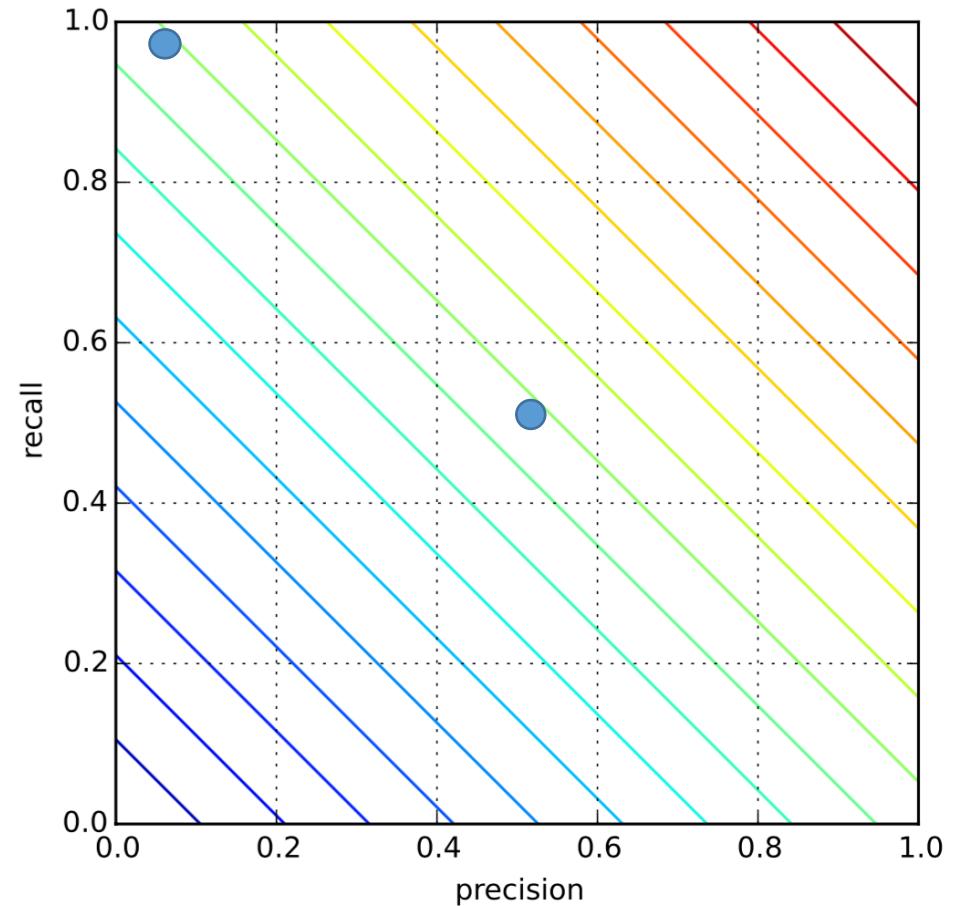
- precision = 0.1
- recall = 1
- $A = 0.55$
- Плохой алгоритм



Арифметическое среднее

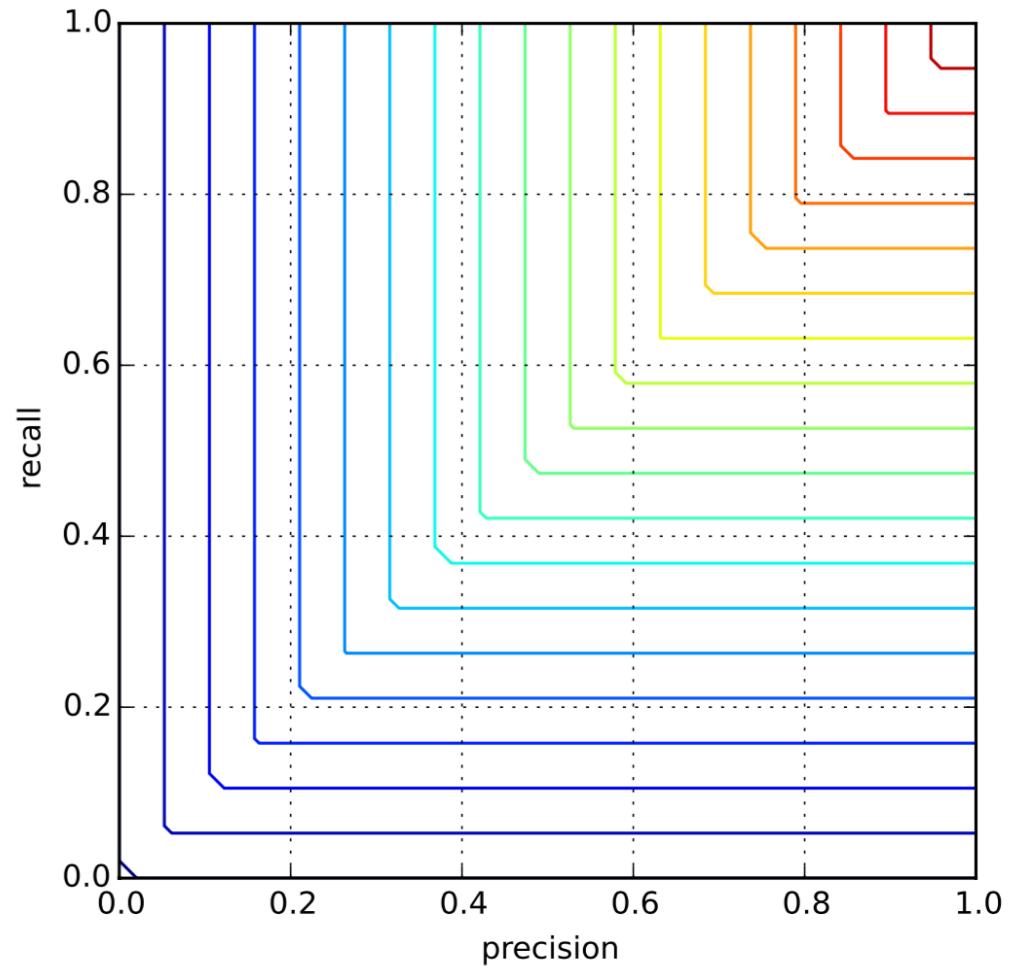
$$A = \frac{1}{2}(\text{precision} + \text{recall})$$

- precision = 0.55
- recall = 0.55
- $A = 0.55$
- Нормальный алгоритм
- Но качество такое же, как у плохого



Минимум

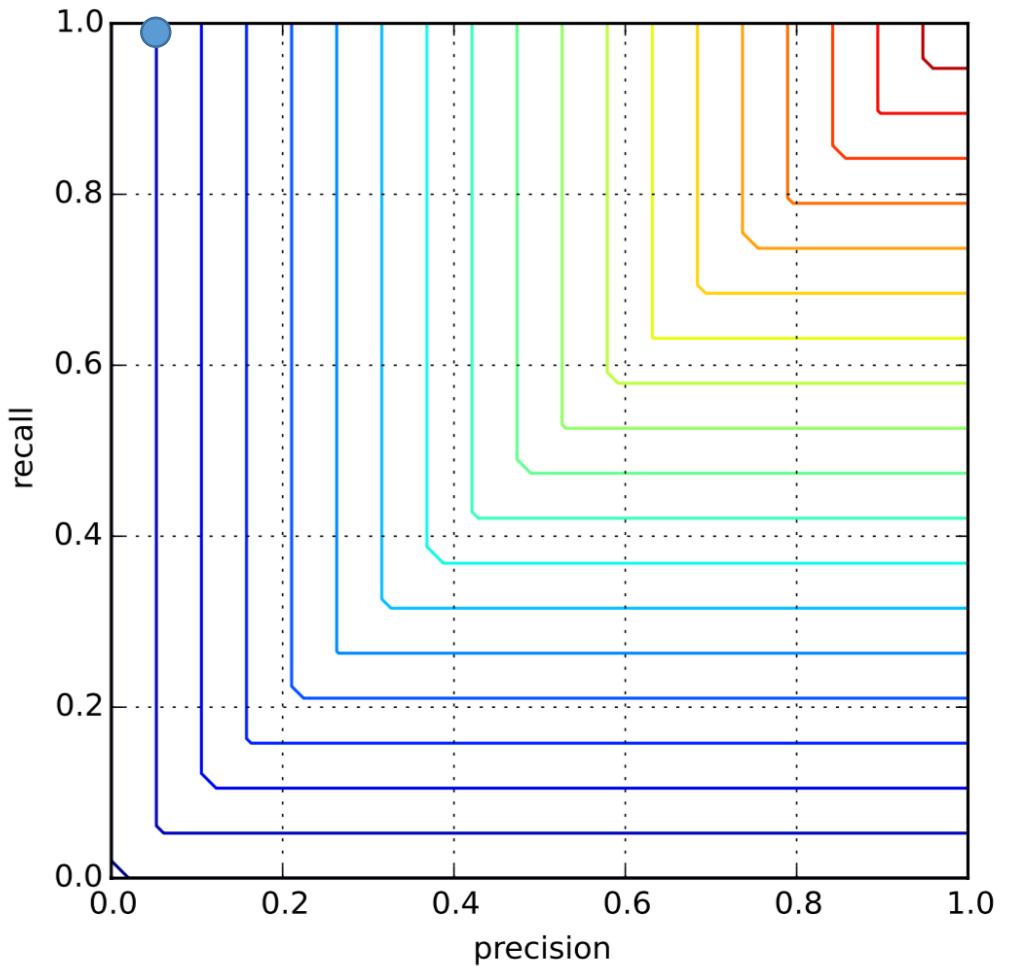
$$M = \min(\text{precision}, \text{recall})$$



Минимум

$$M = \min(\text{precision}, \text{recall})$$

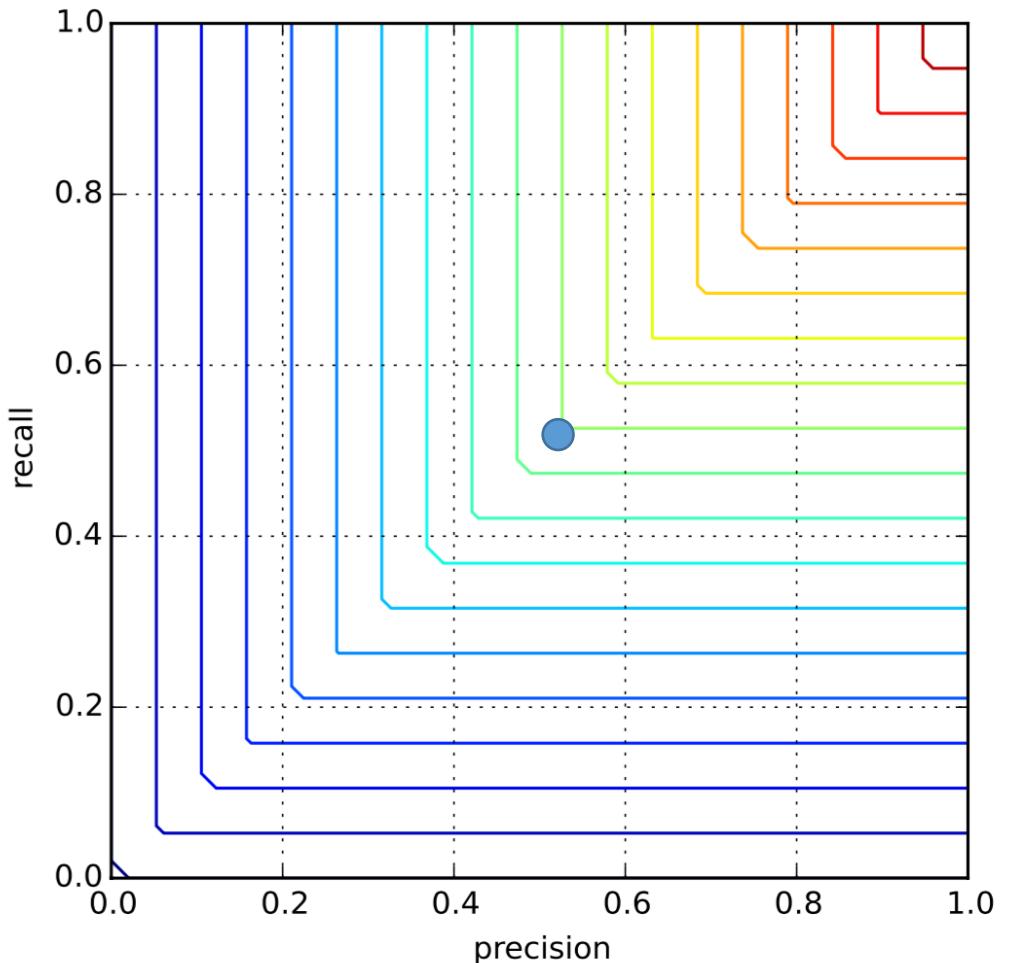
- precision = 0.05
- recall = 1
- $M = 0.05$



Минимум

$$M = \min(\text{precision}, \text{recall})$$

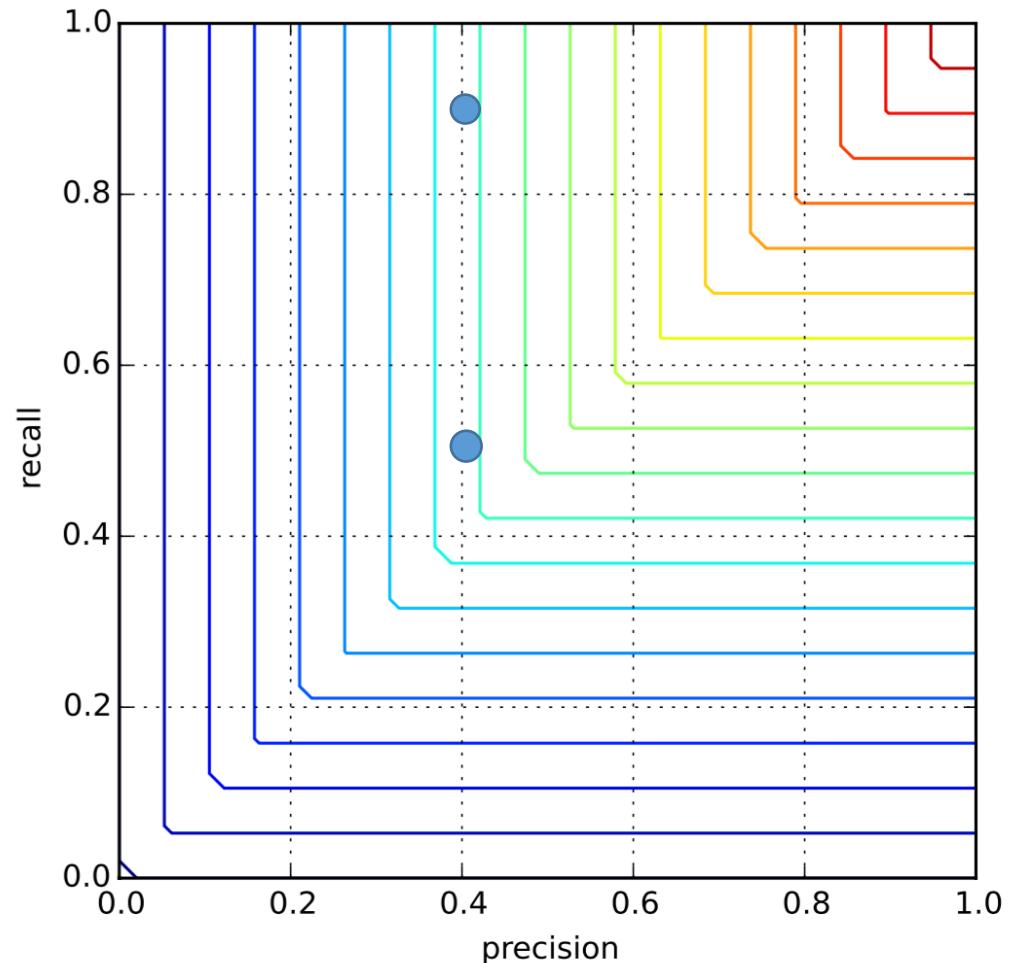
- precision = 0.55
- recall = 0.55
- $M = 0.55$



Минимум

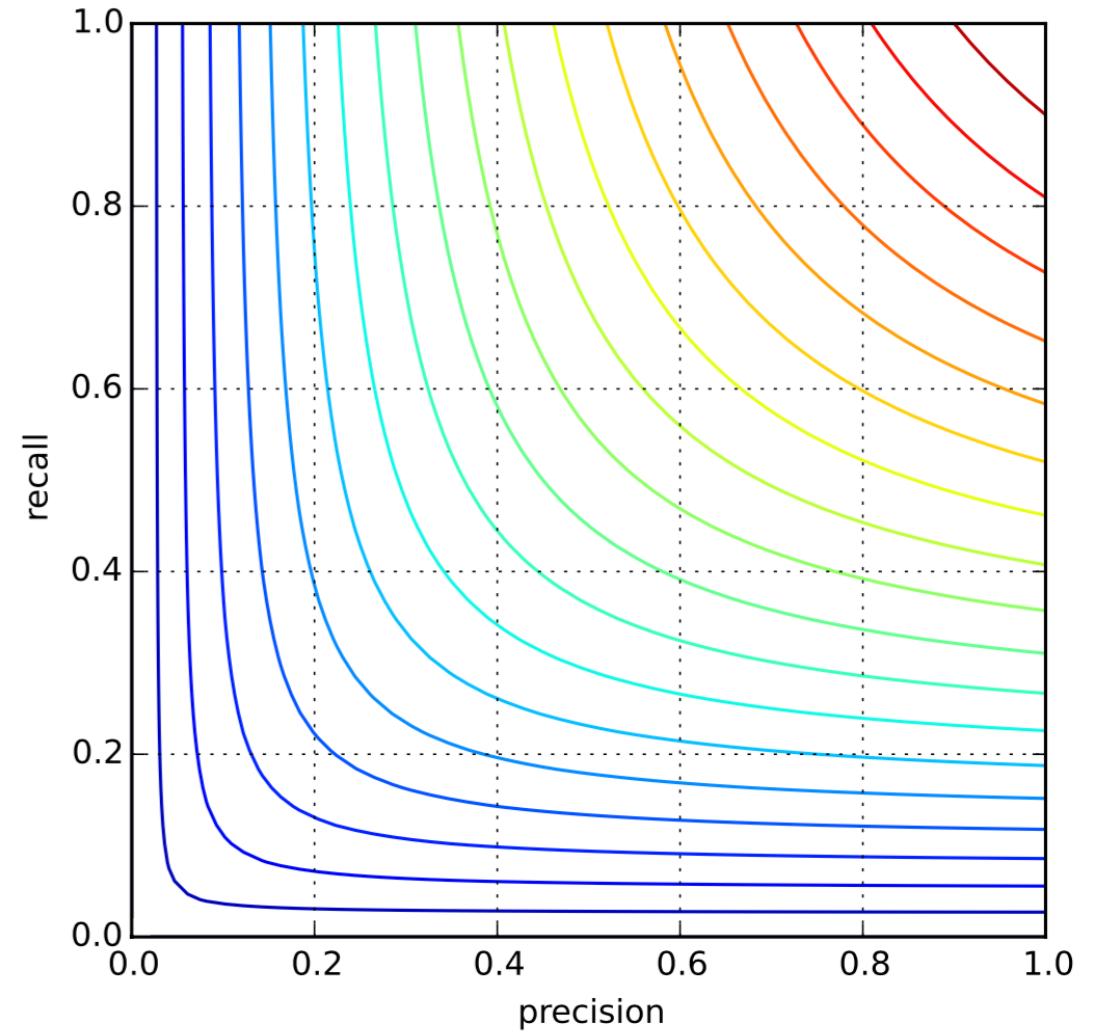
$$M = \min(\text{precision}, \text{recall})$$

- precision = 0.4, recall = 0.5
- $M = 0.4$
- precision = 0.4, recall = 0.9
- $M = 0.4$
- Но второй лучше!



F-Measure

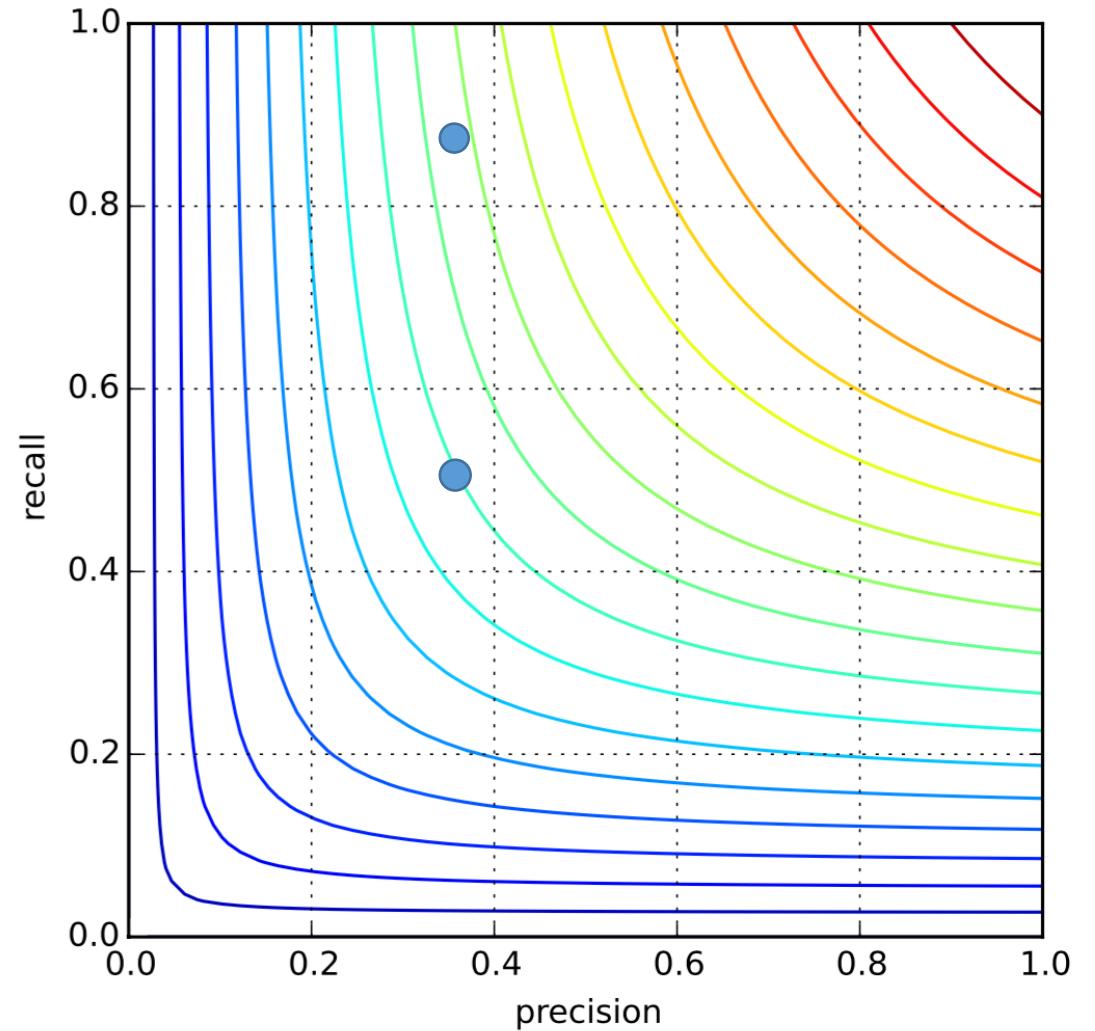
$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



F-Measure

$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

- precision = 0.4, recall = 0.5
- $F = 0.44$
- precision = 0.4, recall = 0.9
- $F = 0.55$



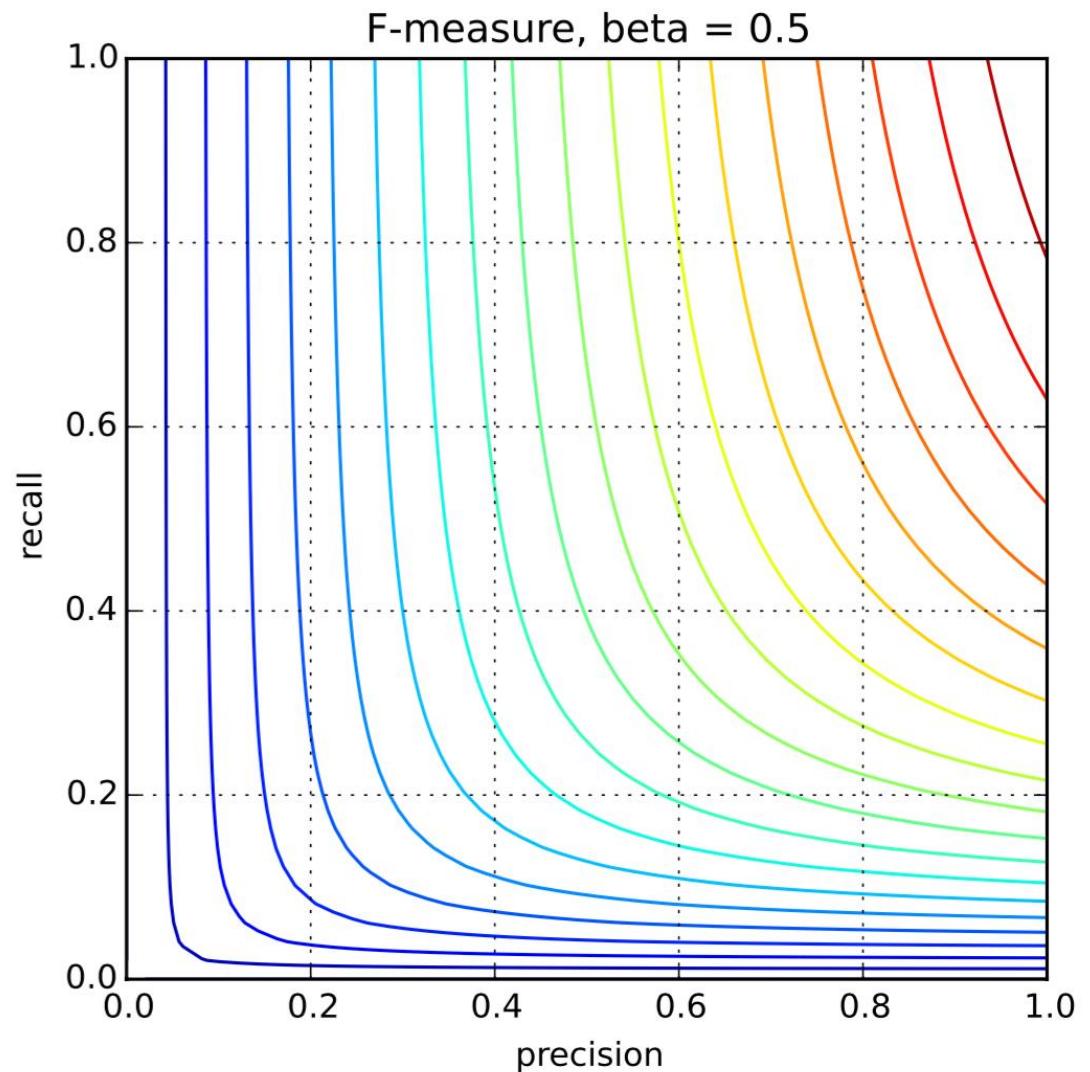
F-Measure

$$F_{\beta} = (1 + \beta^2) \frac{\text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}}$$

F-мера

$$F_\beta = (1 + \beta^2) \frac{\text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}}$$

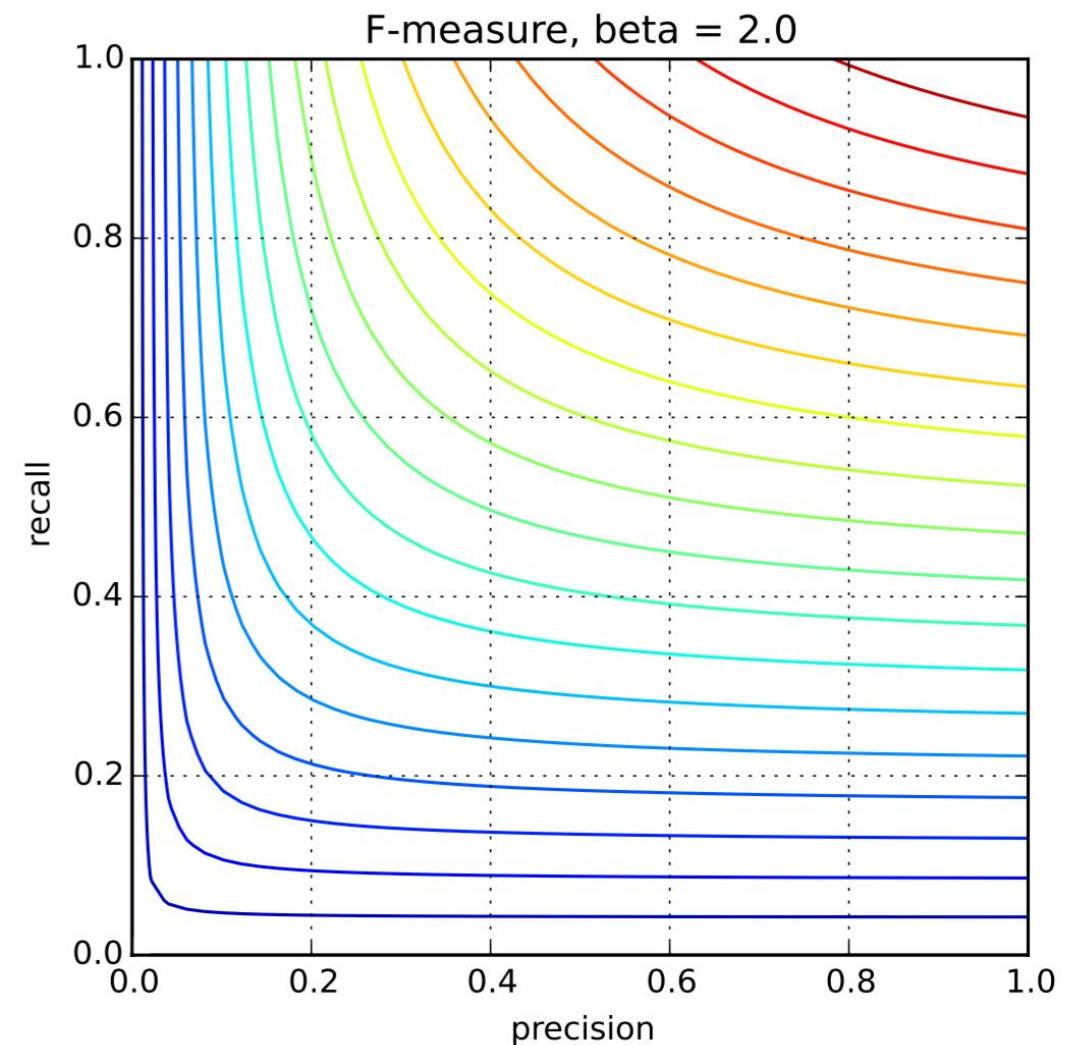
- $\beta = 0.5$
- Важнее точность



F-мера

$$F_\beta = (1 + \beta^2) \frac{\text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}}$$

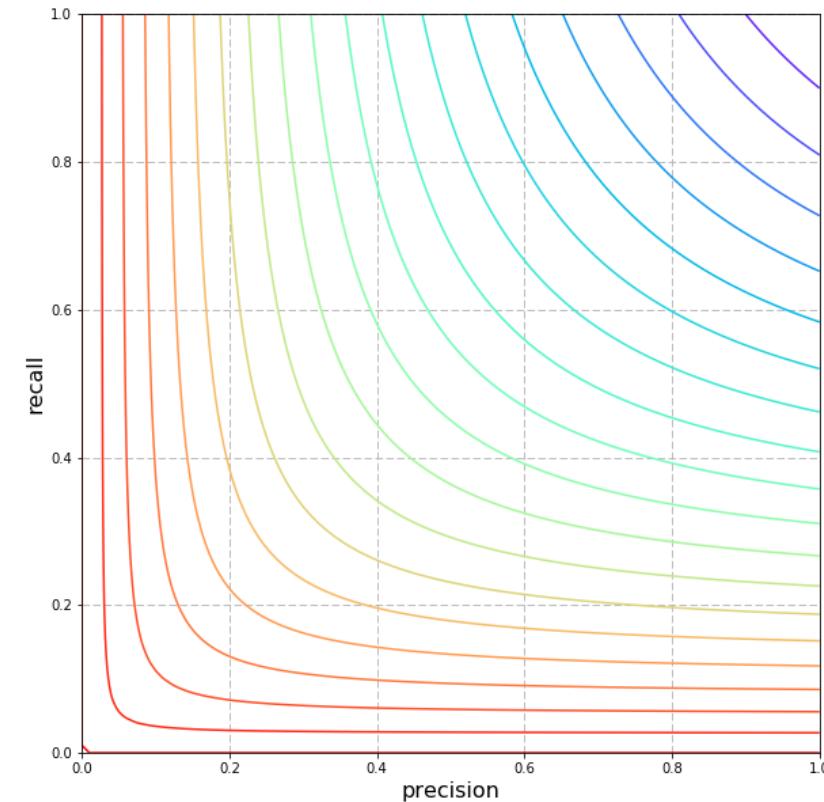
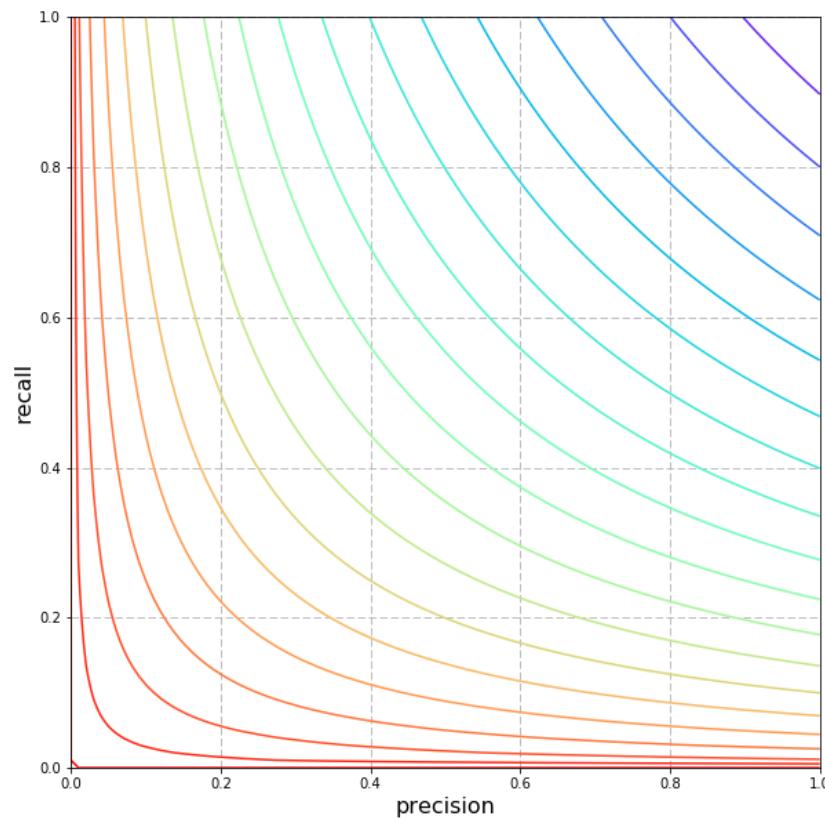
- $\beta = 2$
- Важнее полнота



Геометрическое среднее

$$G = \sqrt{\text{precision} * \text{recall}}$$

$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



Геометрическое среднее

$$G = \sqrt{\text{precision} * \text{recall}}$$

$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

- precision = 0.9
- recall = 0.1
- $G = 0.3$
- precision = 0.9
- recall = 0.1
- $F = 0.18$