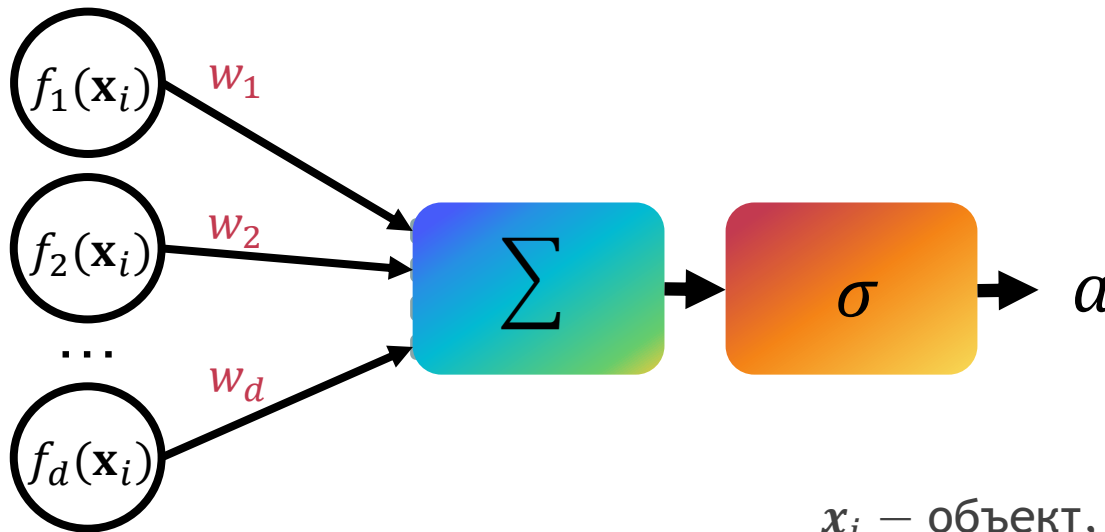


Дисциплина
Основы машинного обучения и нейронные сети

Лекция 5
**Функции потерь.
Линейная классификация**

Нейронная сеть и градиентный спуск

Модель: однослойная нейросеть (1/2)



Цель: подобрать w

Линейная модель:

$$a(\mathbf{x}_i, \mathbf{w}) = \sigma\left(\sum_{j=1}^d w_j f_j(\mathbf{x}_i)\right)$$

Функционал качества:

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^l \mathcal{L}(a, y_i) \rightarrow \min_{\mathbf{w}}$$

x_i — объект, $i = 1, \dots, l$

l — число объектов в выборке

$f_j(\mathbf{x}_i)$ — признак j объекта i , $j = 1, \dots, d$

d — число признаков

y_i — истинный ответ

$X^l = \{(\mathbf{x}_i, y_i)\}_{i=1}^l$ — обучающая выборка

w_j — веса признаков

σ — функция активации

$a(\mathbf{x}_i, \mathbf{w})$ — прогнозируемое значение

$\mathcal{L}(a, y_i) = (a(\mathbf{x}_i, \mathbf{w}) - y_i)^2$ — функция потерь

Модель: однослойная нейросеть (2/2)

Линейная модель:

$$a(\mathbf{x}_i, \mathbf{w}) = \sigma\left(\sum_{j=1}^d w_j f_j(\mathbf{x}_i)\right)$$

Функционал качества:

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

Цель: подобрать \mathbf{w}

Метод решения (метод подбора \mathbf{w}) - градиентный спуск

Вход: выборка X^l , темп обучения h

Выход: вектор весов $\mathbf{w} = (w_1, w_2, \dots, w_d)$

Шаг 1: Задать начальное приближение $\mathbf{w}^{(0)}$;

Шаг 2: Вычислить оценку функционала $Q(\mathbf{w}) = \sum_{i=1}^l \mathcal{L}(\mathbf{w})$;

Шаг 3: Сделать градиентный шаг $\mathbf{w}^{(k+1)} := \mathbf{w}^{(k)} - h \sum_{i=1}^l \nabla \mathcal{L}_i(\mathbf{w}^{(k)})$;

Повторять шаги 2-3, пока значение Q и/или веса \mathbf{w} не сойдутся

Функции потерь

Функции потерь

Функция потерь (loss function) на объекте x — это функция от двух аргументов, истинного значения y ответа для объекта x и прогнозируемого значения $a(x)$, значение которой показывает, насколько прогнозируемое значение близко к истинному.

Чем меньше эта оценка, тем лучше, и наоборот — большие значения оценки указывают на то, что модель нуждается в доработке.

Функции потерь в задачах классификации

Частый выбор — бинарная функция потерь

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

Функционал ошибки — доля ошибок (error rate)

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

Нередко измеряют долю верных ответов (accuracy):

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

Функции потерь в задачах регрессии

Примеры:

- среднеквадратичная ошибка
mean squared error, MSE
- корень из среднеквадратичной ошибки
root mean squared error, RMSE
- средняя абсолютная ошибка
mean absolute error, MAE
- средняя абсолютная ошибка в процентах
mean absolute percentage error, MAPE
- средняя ошибка смещения
mean bias error, MBE
- относительная абсолютная ошибка
relative absolute error, RAE
- функция потерь Хьюбера (Huber loss)
- логарифм гиперболического косинуса $\cosh x$
- квантильная функция потерь

Среднеквадратичная ошибка

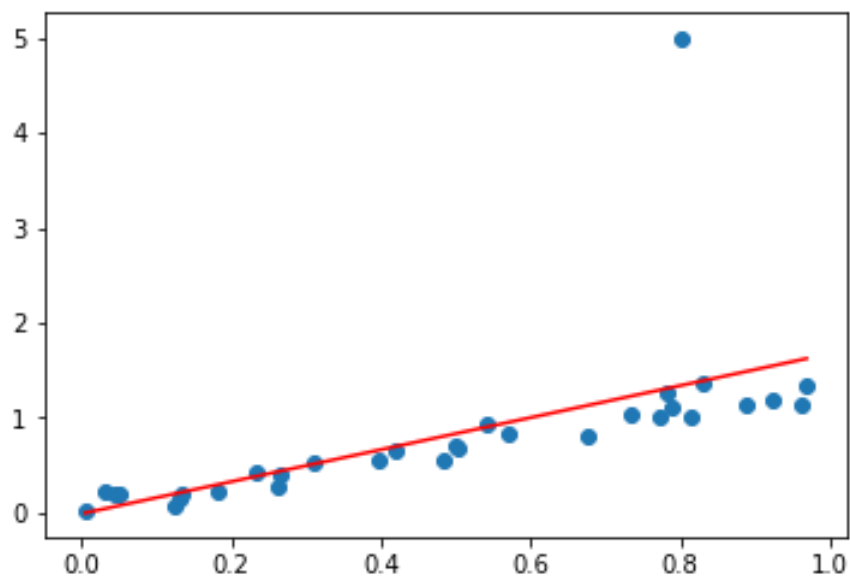
- Частый выбор — квадратичная функция потерь

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

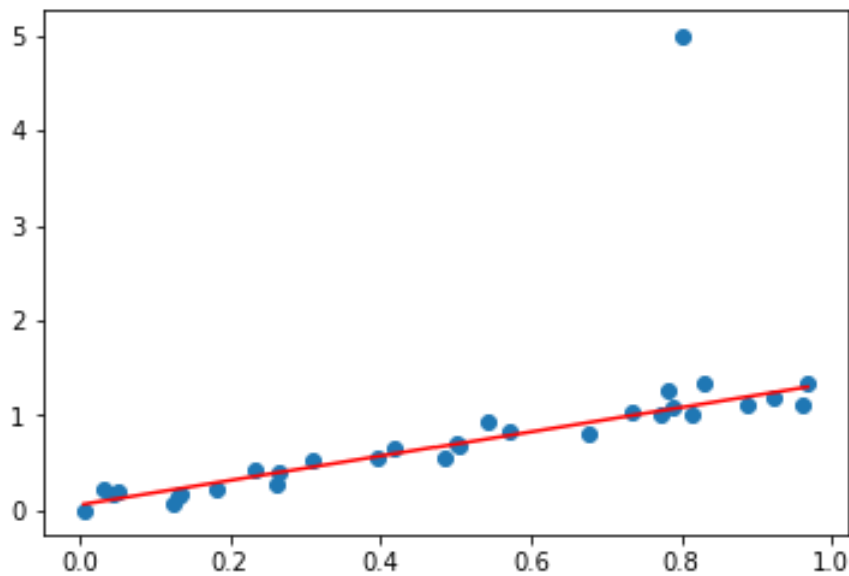
- Функционал ошибки — среднеквадратичная ошибка (mean squared error, MSE)

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

Выбросы для MSE (1/3)



С учётом выброса



Без учёта выброса

Обучение на среднеквадратичной ошибке

Выбросы для MSE (2/3)

$a_1(\mathbf{x})$	y	$(a_1(\mathbf{x}) - y)^2$
2	1	1
1	2	1
2	3	1
5	4	1
6	5	1
7	100	8649
6	7	1

$$a_1(\mathbf{x}): MSE \approx 1236$$

Выбросы для MSE (3/3)

$a_2(\mathbf{x})$	y	$(a_2(\mathbf{x}) - y)^2$
4	1	9
5	2	9
6	3	9
7	4	9
8	5	9
10	100	8100
10	7	9

$$a_2(\mathbf{x}): MSE \approx 1164$$

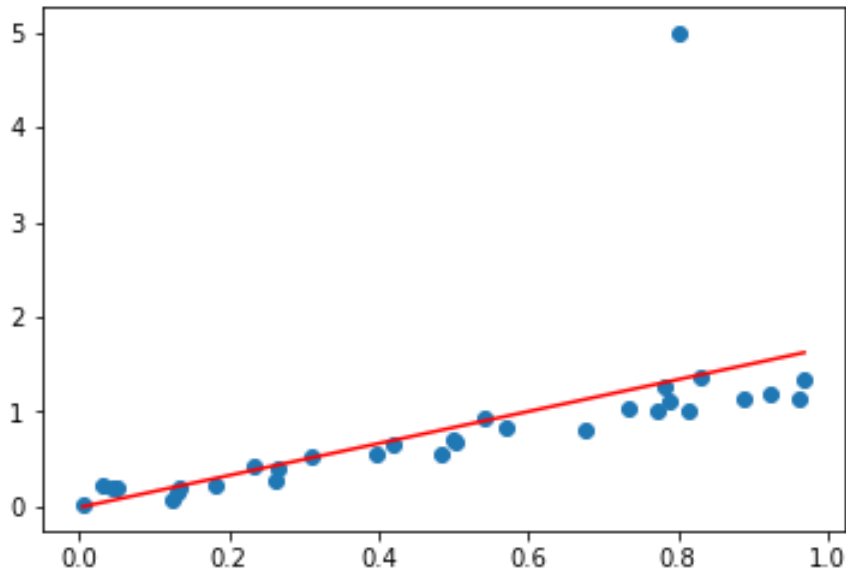
Средняя абсолютная ошибка

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

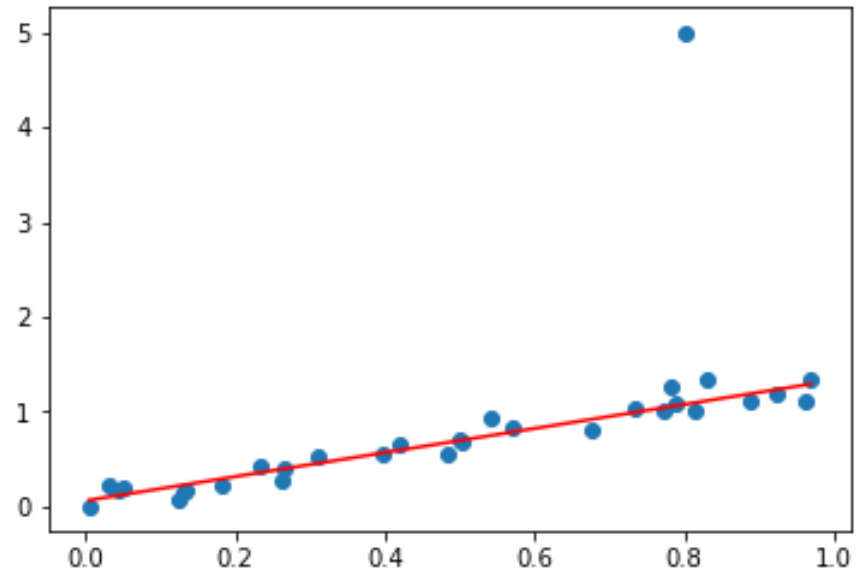
- Функционал ошибки — средняя абсолютная ошибка (mean absolute error, MAE)

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

Выбросы для MAE (1/3)



Обучение на MSE



Обучение на MAE

Выбросы для MAE (2/3)

$a_1(\mathbf{x})$	y	$ a_1(\mathbf{x}) - y $
2	1	1
1	2	1
2	3	1
5	4	1
6	5	1
7	100	93
6	7	1

$$a_1(\mathbf{x}): MAE \approx 14.14$$

Выбросы (3/3)

$a_2(\mathbf{x})$	y	$ a_2(\mathbf{x}) - y $
4	1	3
5	2	3
6	3	3
7	4	3
8	5	3
10	100	90
10	7	3

$$a_2(\mathbf{x}): MAE \approx 15.43$$

Функция потерь Хубера (1/2)

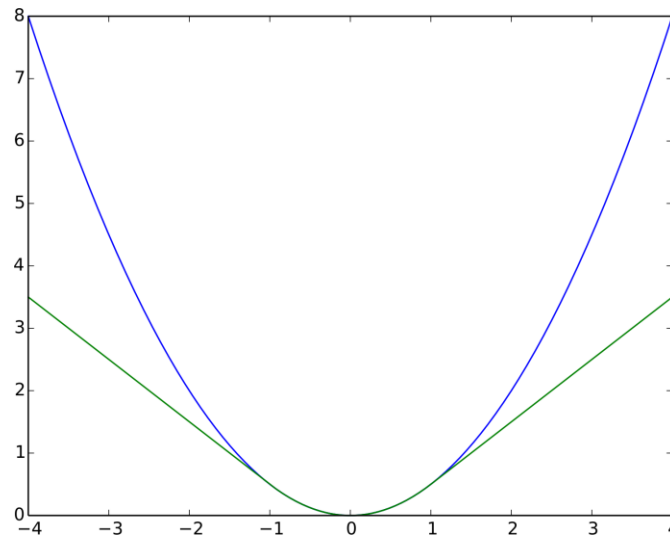
$$L_H(y, a) = \begin{cases} \frac{1}{2} (y - a)^2, & |y - a| < \delta \\ \delta \left(|y - a| - \frac{1}{2} \delta \right), & |y - a| \geq \delta \end{cases}$$

- Функционал ошибки:

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

Функция потерь Хубера (2/2)

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



MAPE (1/3)

- Mean Absolute Percentage Error (средний модуль относительной ошибки)

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

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

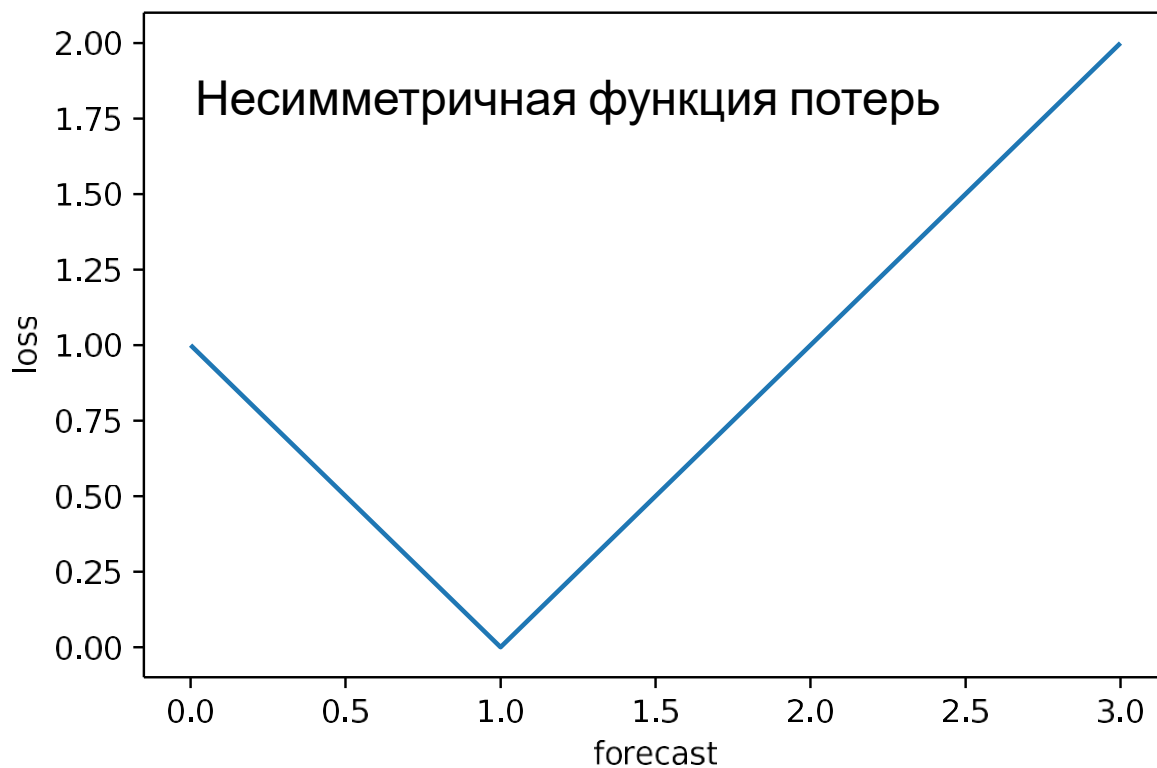
MAPE (2/3)

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

- Особенность: при $a = 0$ добавить в знаменатель небольшое слагаемое
- Недопрогноз штрафуются максимум на единицу
- Перепрогноз может быть оштрафован большим числом
- Несимметричная функция потерь (отдаёт предпочтение недопрогнозу)

MAPE (3/3)

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



SMAPE (1/2)

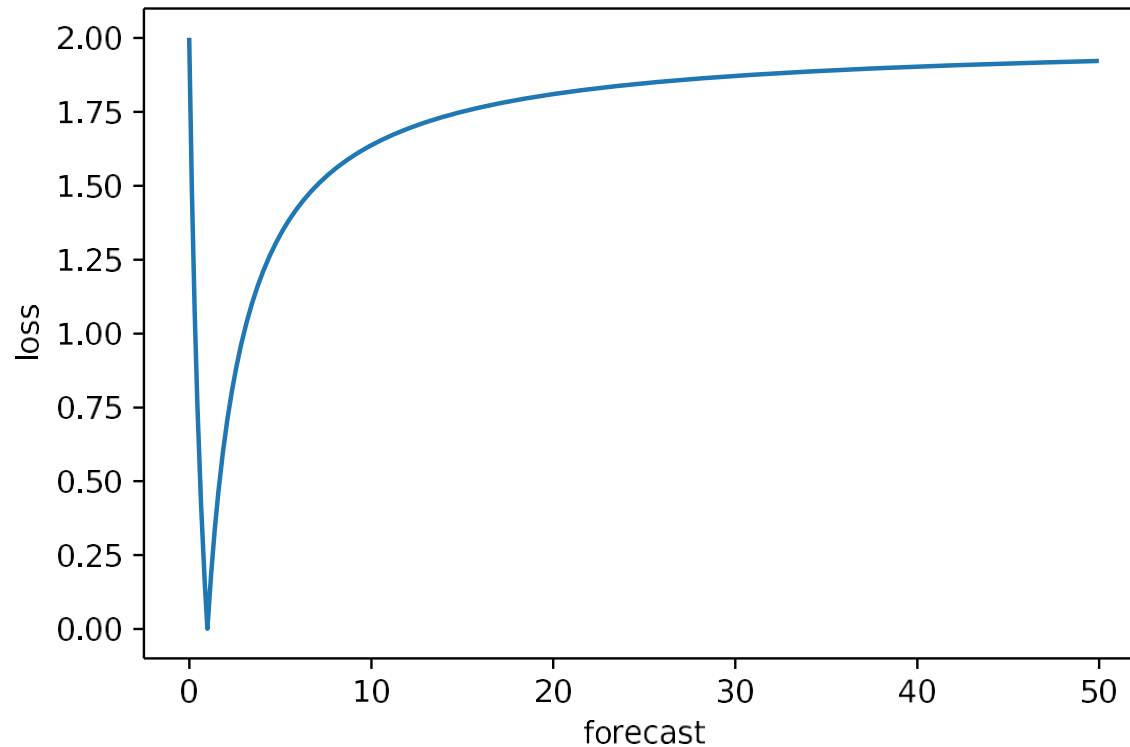
- Symmetric Mean Absolute Percentage Error
(симметричный средний модуль относительной ошибки)

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

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

SMAPE (2/2)

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



Модель линейной классификации

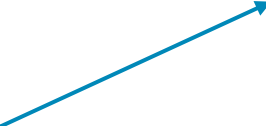
Классификация (бинарная)

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

Линейная регрессия

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

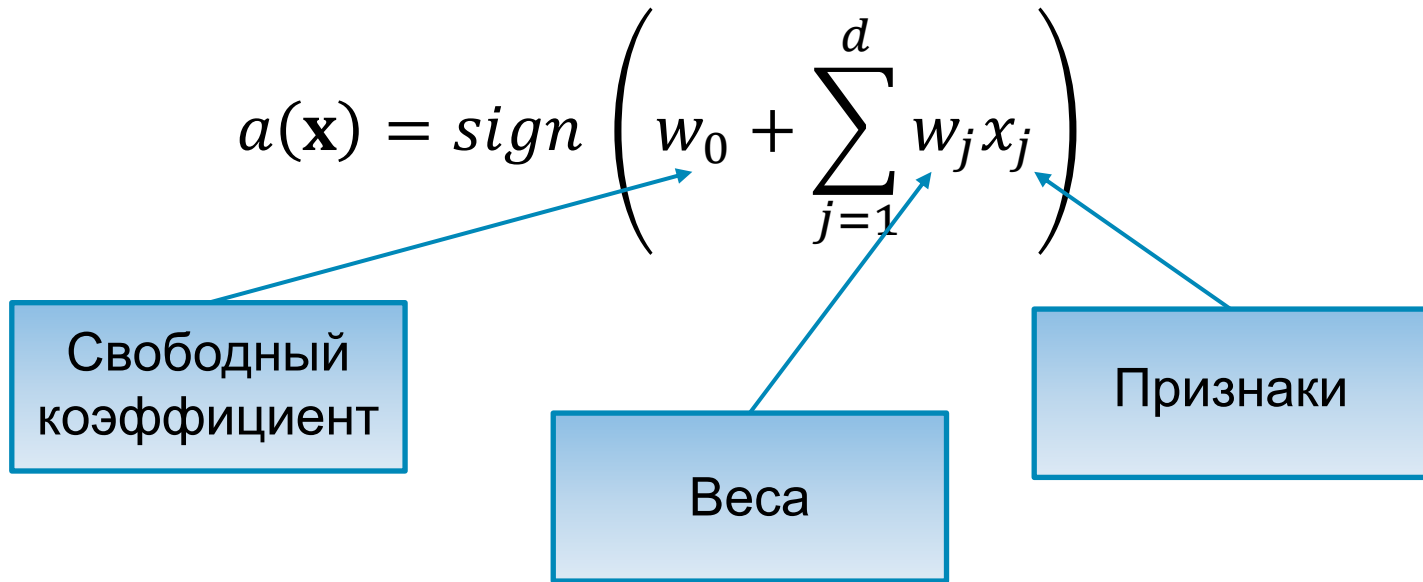
Вещественное
число!



Линейный классификатор (1/2)

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

Линейный классификатор

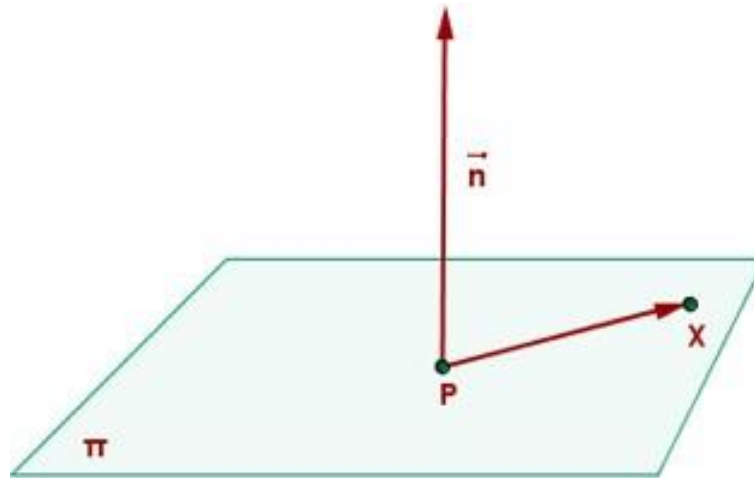


Будем считать, что есть единичный признак, тогда

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

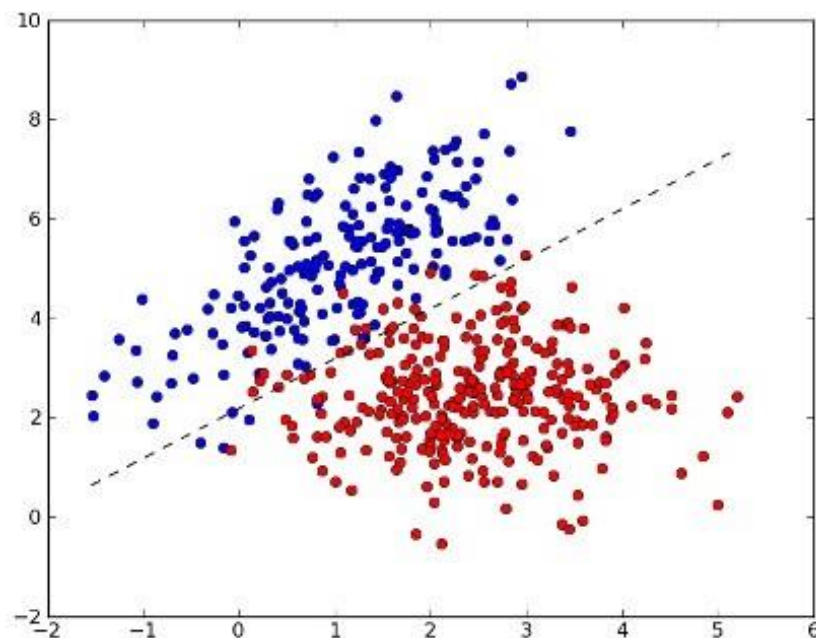
Геометрия линейного классификатора (1/4)

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



Геометрия линейного классификатора (2/4)

- Линейный классификатор соответствует гиперплоскости с вектором нормали \mathbf{w}
- Величина $\langle \mathbf{w}, \mathbf{x} \rangle$ пропорциональна расстоянию от точки \mathbf{x} до гиперплоскости
- $\langle \mathbf{w}, \mathbf{x} \rangle < 0$ — объект «слева» от гиперплоскости
- $\langle \mathbf{w}, \mathbf{x} \rangle > 0$ — объект «справа» от гиперплоскости



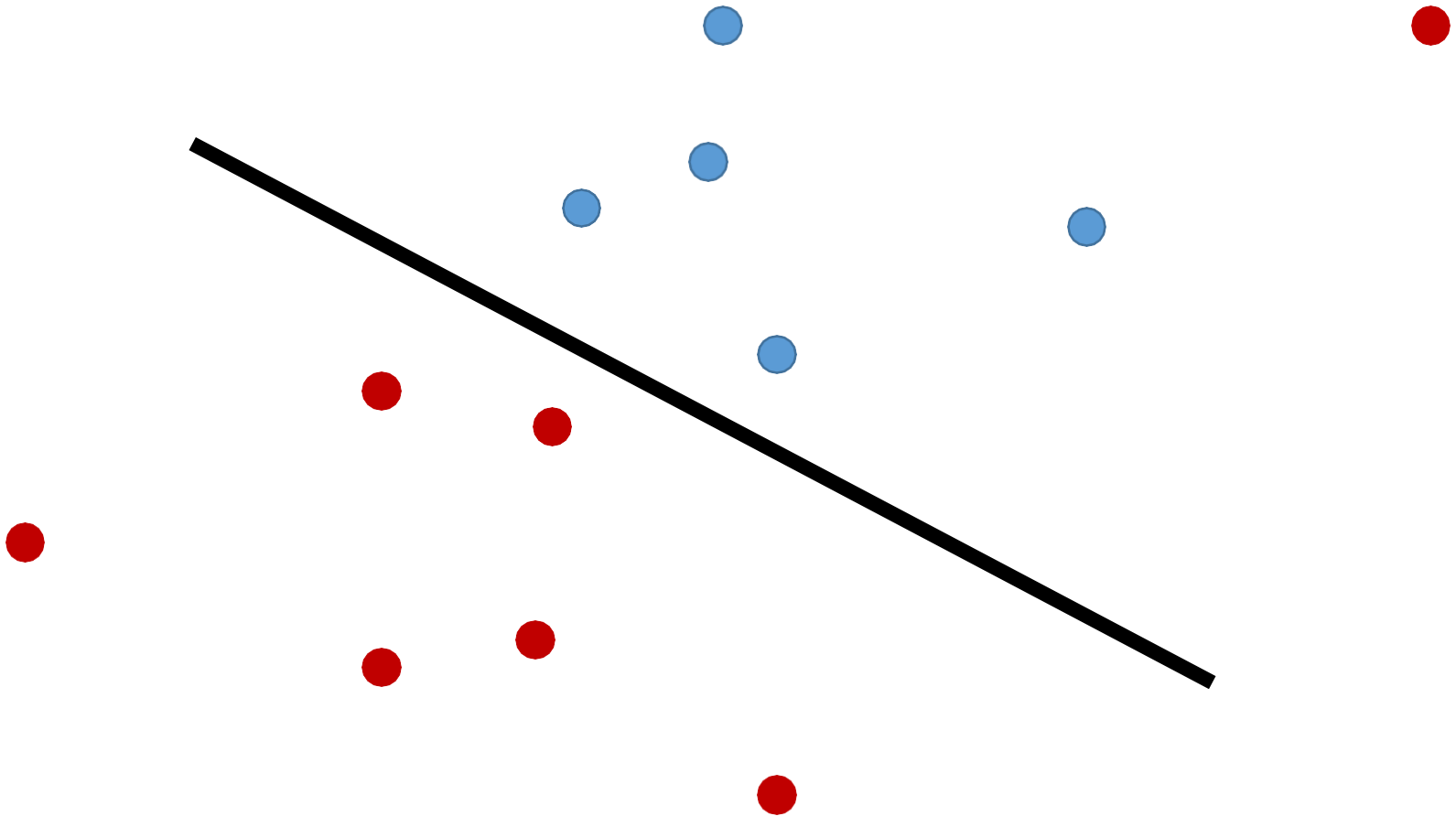
Геометрия линейного классификатора (3/4)

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

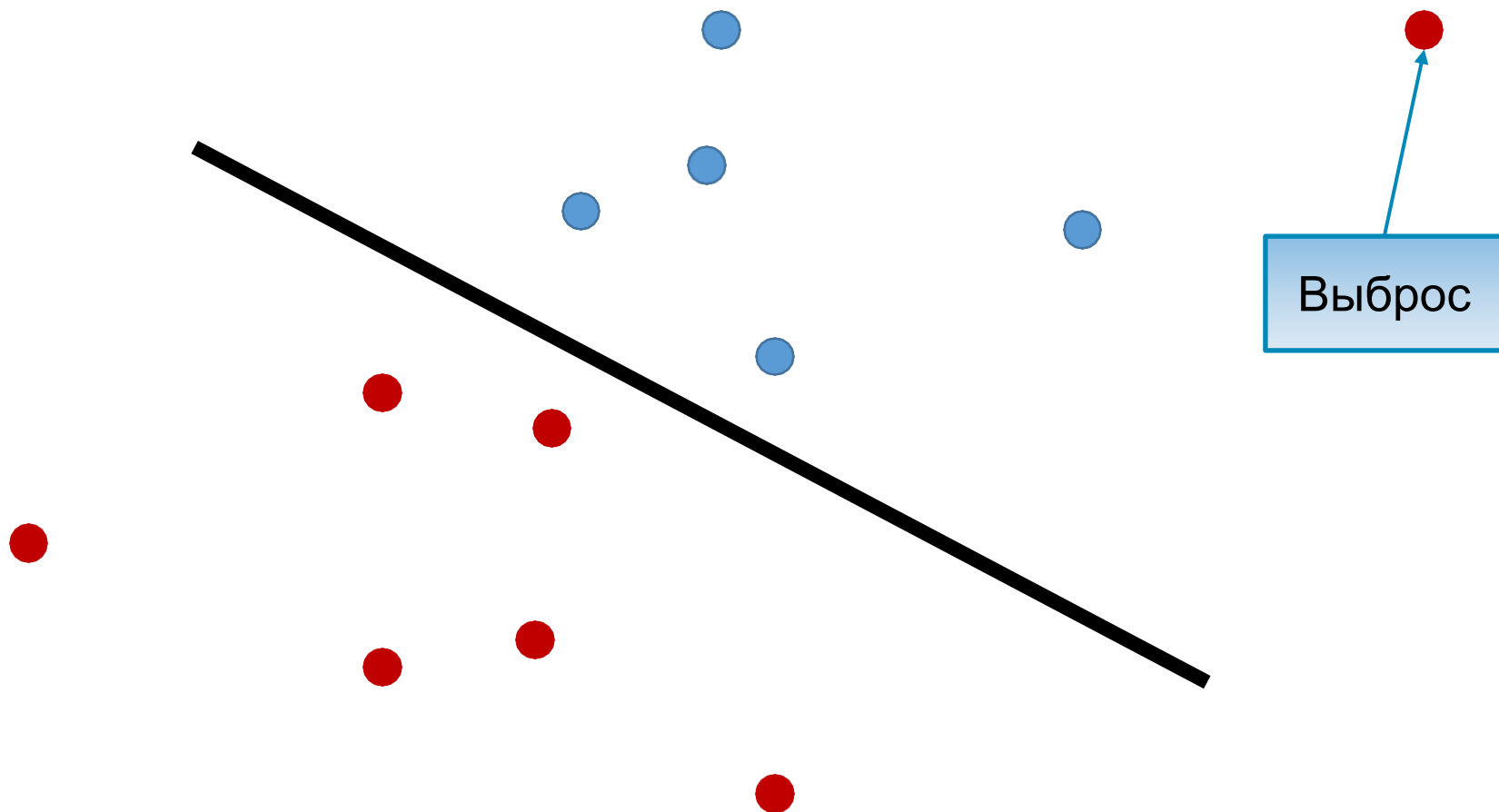
$$\frac{|\langle \mathbf{w}, \mathbf{x} \rangle|}{\|\mathbf{w}\|}$$

- Чем больше $\langle \mathbf{w}, \mathbf{x} \rangle$, тем дальше объект от разделяющей гиперплоскости, т.е. тем более «уверена» модель в своём ответе

Геометрия линейного классификатора (4/4)

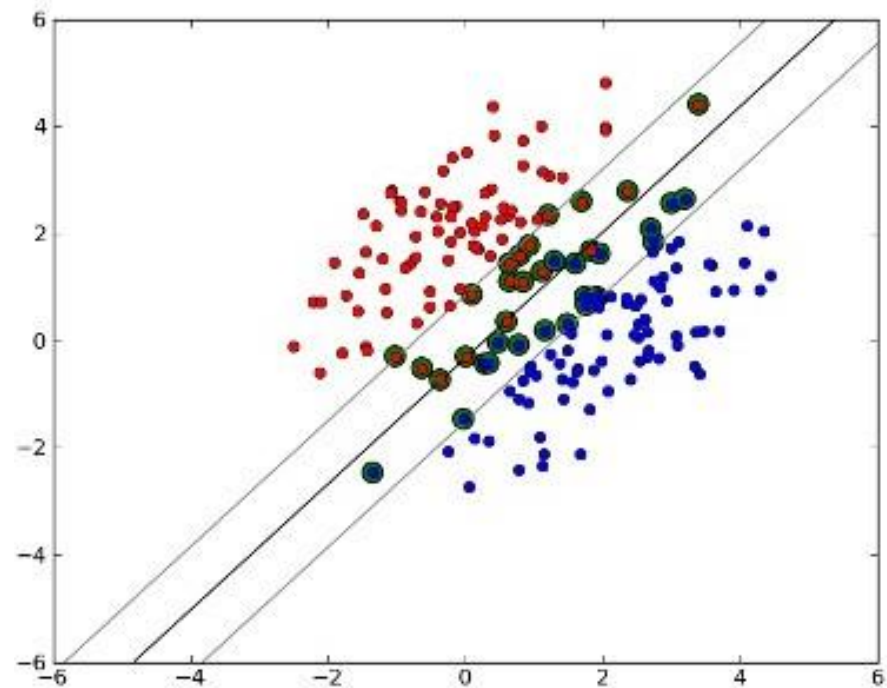


Геометрия линейного классификатора (4/4)



Отступ

- $M_i = y_i \langle \mathbf{w}, \mathbf{x}_i \rangle$
- $M_i > 0$ — классификатор дает верный ответ
- $M_i < 0$ — классификатор ошибается
- Чем дальше значение отступа от нуля, тем больше уверенности в корректности ответа



Порог

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

- 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

$$\mathcal{L}(a(\mathbf{x}_i), y_i) = [a(\mathbf{x}_i) \neq y_i]$$

- Функционал ошибки — доля ошибок (error rate)

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

- Интуитивно понятная метрика - доля верных ответов (accuracy):

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

Доля ошибок для линейного классификатора

- Функционал ошибки:

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

Задача: найти

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{R}^d} Q(a(\mathbf{w}, \mathbf{x}), X^\ell)$$

- Индикаторная функция — не гладкая функция

Гладкие и кусочно-гладкие функции

вспоминаем математический анализ

Гладкая (или непрерывно дифференцируемая) функция - это функция, имеющая непрерывную производную на всём множестве определения. Часто под гладкими функциями подразумевают функции, имеющие непрерывные производные всех порядков.

Кусочно-гладкая функция — функция, определённая на множестве вещественных чисел, дифференцируемая на каждом из интервалов, составляющих область определения.

Пусть заданы $x_1 < x_2 < \dots < x_n$ — точки смены формул.

Как и все **кусочно-заданные функции**, кусочно-гладкую функцию можно записывать на каждом из интервалов $(-\infty; x_1), (x_1; x_2); \dots (x_n; +\infty)$ отдельной формулой:

$$f(x) = \begin{cases} f_0(x), & x < x_1 \\ f_1(x), & x_1 < x < x_2 \\ \dots & \\ f_n(x), & x_n < x \end{cases}$$

Здесь $f_i(x)$ — **гладкие функции**.

Если к тому же выполнены *условия согласования*

$$f_{i-1}(x_i) = f_i(x_i) = f(x_i) \text{ при } i = 1, 2, \dots, n,$$

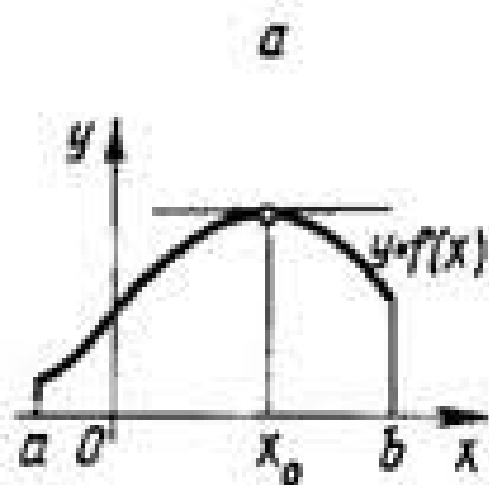
то кусочно-гладкая функция будет **непрерывной**. Непрерывная кусочно-гладкая функция может служить **сплайном**.

https://ru.wikipedia.org/wiki/%D0%9A%D1%83%D1%81%D0%BE%D1%87%D0%BD%D0%BE-%D0%B3%D0%BB%D0%B0%D0%B4%D0%BA%D0%B0%D1%8F_%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F

Виды критических точек

вспоминаем математический анализ

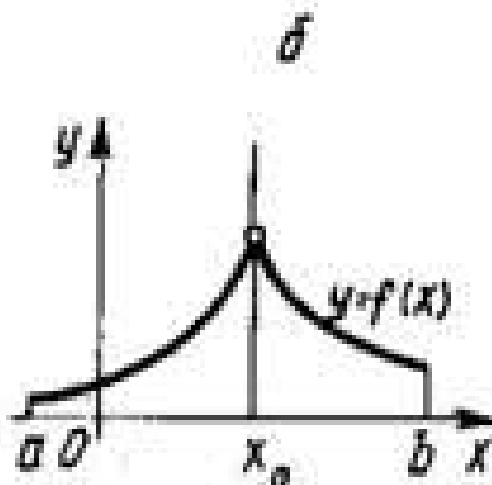
Критическая точка – точка возможного экстремума, производная в ней либо равна нулю $\nabla f(x_0) = 0$, либо не существует.



касательная параллельна
оси абсцисс,

$$\nabla f(x_0) = 0,$$

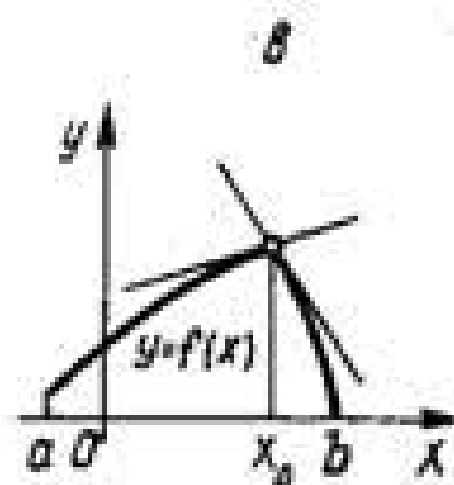
x_0 - стационарная точка



касательная
параллельна
оси ординат,

$$\nabla f(x_0) = \infty,$$

x_0 - точка возврата



существуют не совпадающие
левая и правая касательные,

$$\nabla f(x_0 - 0) \neq \nabla f(x_0 + 0),$$

x_0 - угловая точка

Отступы для линейного классификатора (1/2)

- Функционал ошибки:

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

- Альтернативная запись

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

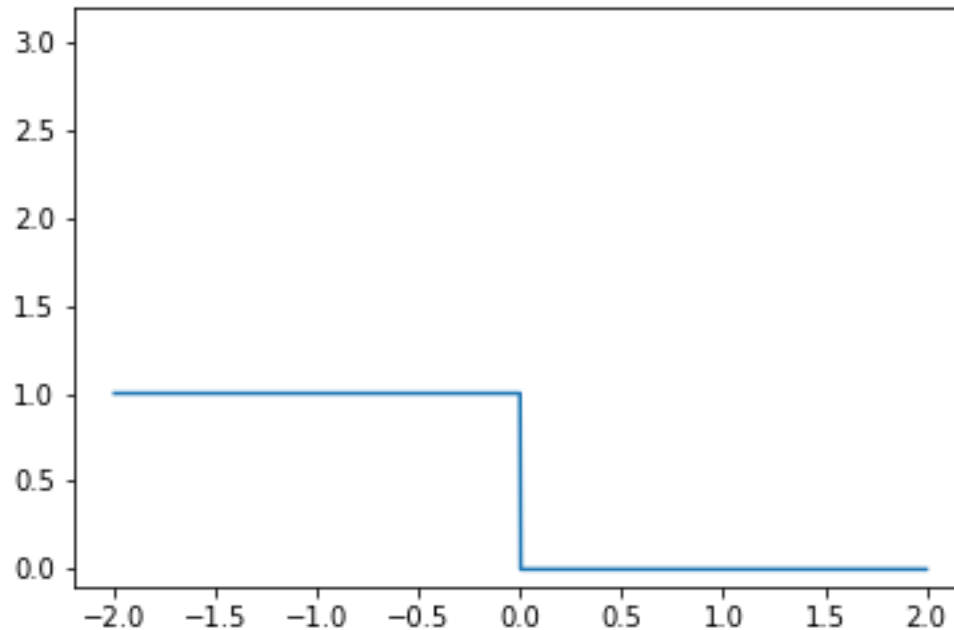
т.е. с помощью отступа:

$$Q(\mathbf{w}, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [M_i < 0] =: \mathcal{L}(M)$$

Отступы для линейного классификатора (2/2)

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

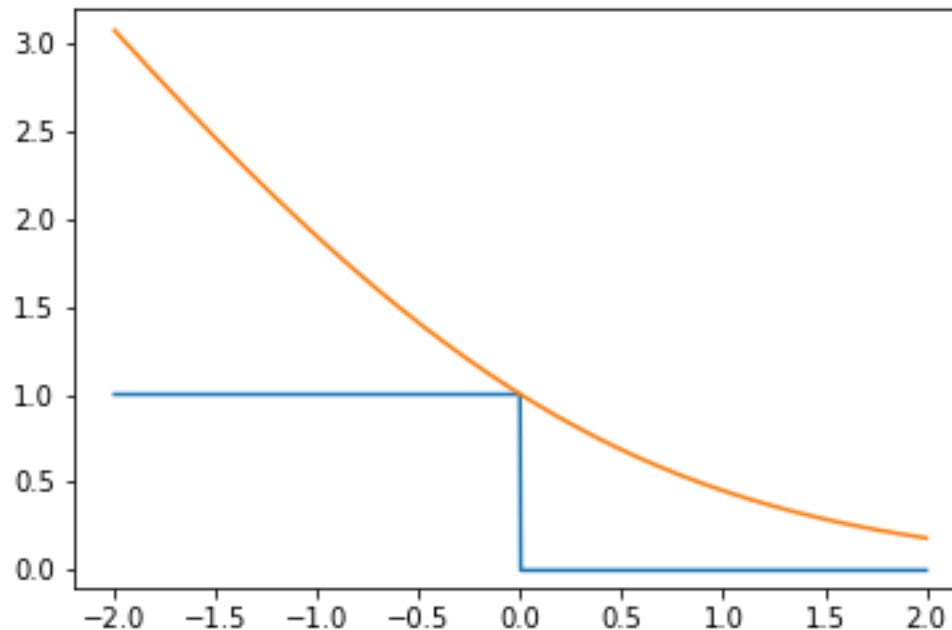
- Нельзя продифференцировать



Верхняя оценка (1/2)

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

- Оценим сверху дифференцируемой функцией



Верхняя оценка (2/2)

$$0 \leq \frac{1}{\ell} \sum_{i=1}^{\ell} [y_i \langle \mathbf{w}, \mathbf{x}_i \rangle < 0] \leq \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{\mathcal{L}}(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \rightarrow \min_{\mathbf{w}}$$

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

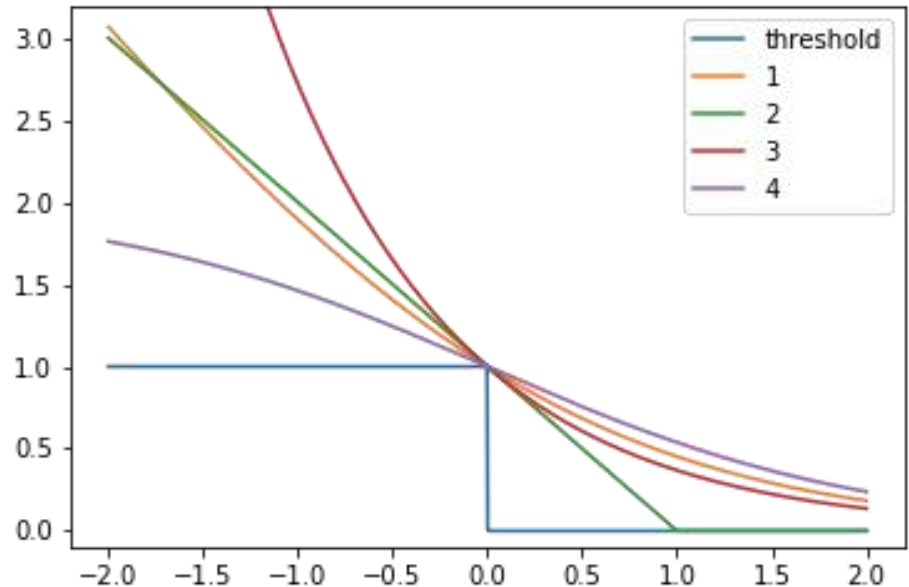
Примеры верхних оценок

1. $\check{\mathcal{L}}(M) = \log(1 + e^{-M})$ – логистическая

2. $\check{\mathcal{L}}(M) = \max(0, 1 - M)$ – кусочно-линейная

3. $\check{\mathcal{L}}(M) = e^{-M}$ – экспоненциальная

4. $\check{\mathcal{L}}(M) = \frac{2}{1+e^M}$ – сигмоидная



Пример обучения (1/2)

- Напр., выбираем логистическую функцию потерь:

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

- Вычисляем градиент:

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

Пример обучения (2/2)

- Делаем градиентный спуск:

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

Пример регуляризации

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

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

Метрики качества классификации

Качество классификации (1/2)

- Доля неправильных ответов

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

Качество классификации (2/2)

- Доля правильных ответов (accuracy):

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

Несбалансированные выборки (1/3)

Несбалансированная выборка — объектов одного класса существенно больше

Примеры:

- предсказание кликов по рекламе
- медицинская диагностика
- предсказание оттока клиентов
- специализированный поиск

Несбалансированные выборки (2/3)

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

Несбалансированные выборки (2/3)

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

Несбалансированные выборки (3/3)

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

$$accuracy \in (q_0, 1]$$

- Если получили большой accuracy — посмотрите на баланс классов

Улучшение метрики (1/2)

- 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 (2/2)

- $r_1 = 0.8$
 - $r_2 = 0.9$
 - $r_2 - r_1 = 10\%$
 - $\frac{r_2 - r_1}{r_1} = 12.5\%$
- $r_1 = 0.5$
 - $r_2 = 0.75$
 - $r_2 - r_1 = 25\%$
 - $\frac{r_2 - r_1}{r_1} = 50\%$
- $r_1 = 0.001$
 - $r_2 = 0.01$
 - $r_2 - r_1 = 9,9\%$
 - $\frac{r_2 - r_1}{r_1} = 900\%$

Цены ошибок (1/2)

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

Цены ошибок (2/2)

- Что хуже?
 - Выдать кредит «плохому» клиенту
 - Не выдать кредит «хорошему» клиенту
- Лучше та модель, на которой мы больше заработаем
- Доля верных ответов не учитывает цены ошибок

Метрики качества классификации

Матрица несоответствий /
ошибок (confusion matrix)
для $\ell=10$:

	$y = 1$	$y = -1$
$a(x) = 1$	7 True Positive	3 False Positive
$a(x) = -1$	2 False Negative	8 True Negative

Доля верных ответов:

$$accuracy(a, x) = \frac{TP + TN}{TP + FP + FN + TN} = 0,75 \quad (2.3)$$

Точность :

$$precision = \frac{TP}{TP + FP} = 0,7 \quad (2.4)$$

Полнота :

$$recall = \frac{TP}{TP + FN} = 0,7 \quad (2.5)$$

True Positive (TP) – учил, сдал

False Negative (FN) – учил, не сдал (ошибка 2 рода)

True Negative (TN) – не учил, не сдал

False Positive (FP) – не учил, сдал (ошибка 1 рода)

Матрица ошибок (1/2)

	$y = 1$	$y = -1$
$a(\mathbf{x}) = 1$	True Positive (TP)	False Positive (FP), ложное срабатывание
$a(\mathbf{x}) = -1$	False Negative (FN), ложный пропуск	True Negative (TN)

Матрица ошибок (2/2)

- 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) (1/2)

- Можно ли доверять классификатору при $a(\mathbf{x}) = 1$?

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

Точность (precision) (2/2)

- Модель $a_1(\mathbf{x})$:

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

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

- Модель $a_2(\mathbf{x})$:

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

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

Полнота (recall) (1/2)

- Как много положительных объектов находит классификатор?

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

Полнота (recall) (2/2)

- Модель $a_1(\mathbf{x})$:

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

- $recall(a_1, X) = 0.8$

- Модель $a_2(\mathbf{x})$:

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

- $recall(a_2, X) = 0.48$

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: $recall(a, X) \geq 0.8$
- Максимизируем точность

Несбалансированные выборки

- $accuracy(a, X) = 0.99$
- $precision(a, X) = 0.33$
- $recall(a, X) = 0.1$

	$y = 1$	$y = -1$
$a(\mathbf{x}) = 1$	10	20
$a(\mathbf{x}) = -1$	90	10000

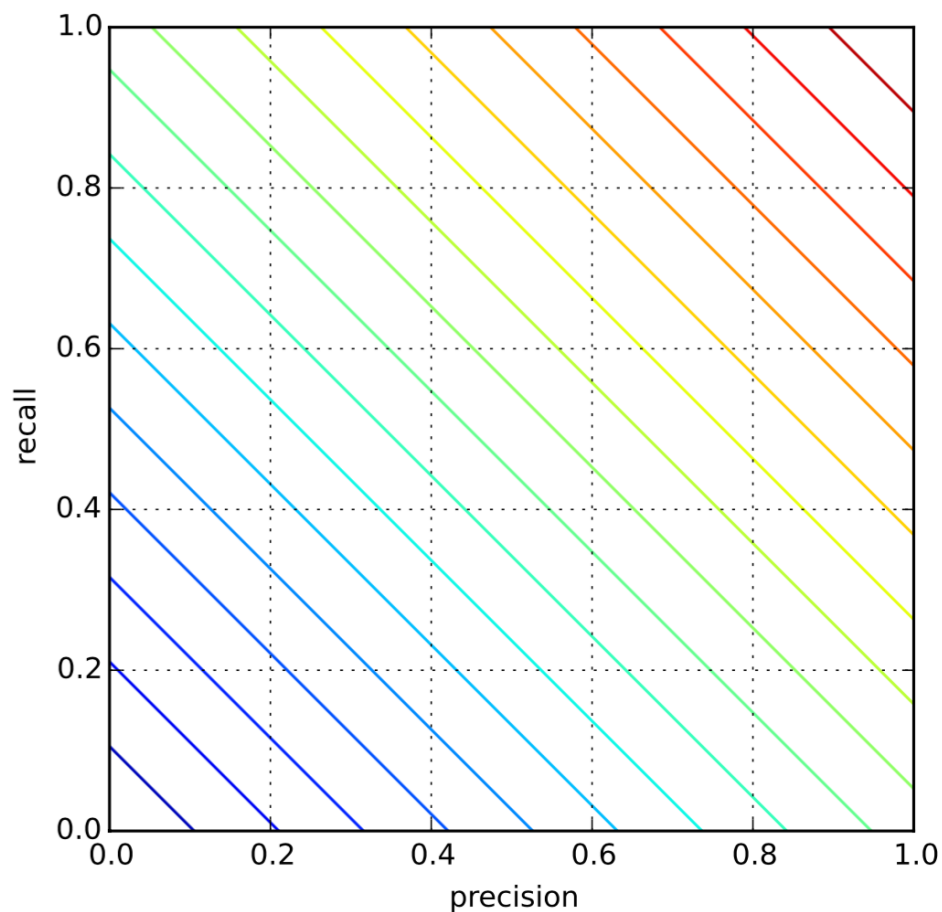
Совмещение точности и полноты

Точность и полнота

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

Арифметическое среднее

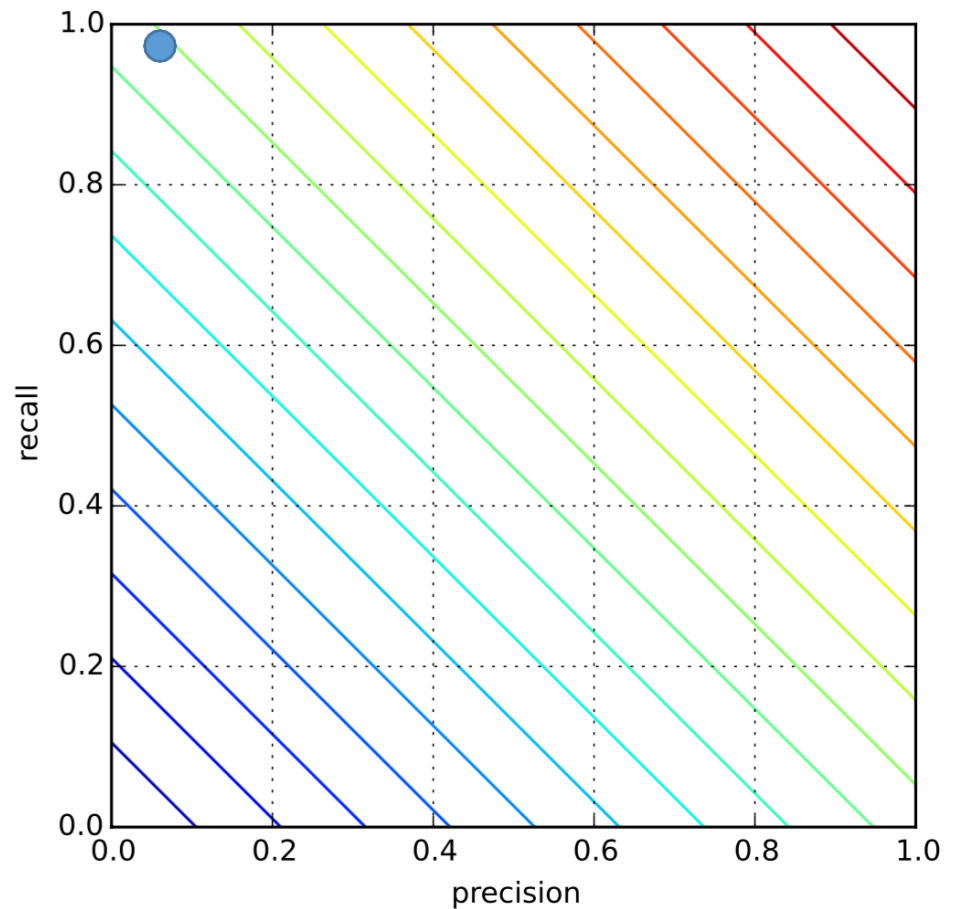
$$A = \frac{1}{2}(\textit{precision} + \textit{recall})$$



Арифметическое среднее (1/2)

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

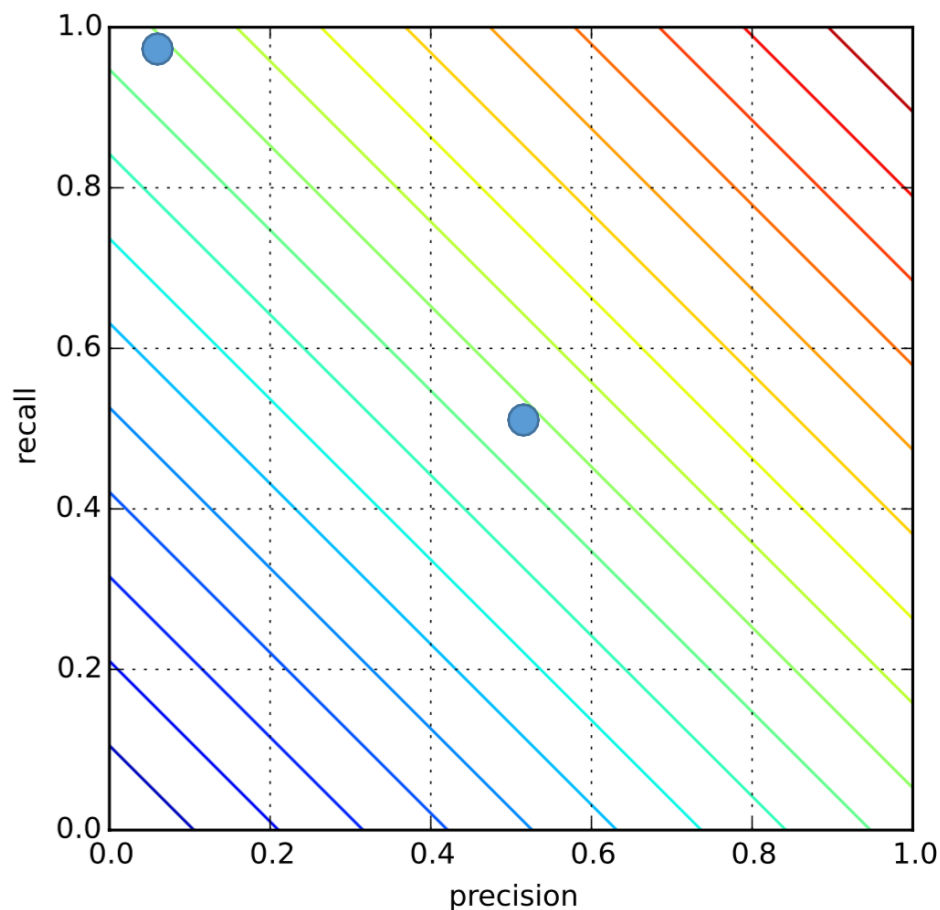
- $\textit{precision} = 0.1$
- $\textit{recall} = 1$
- $A = 0.55$
- Плохой алгоритм



Арифметическое среднее (2/2)

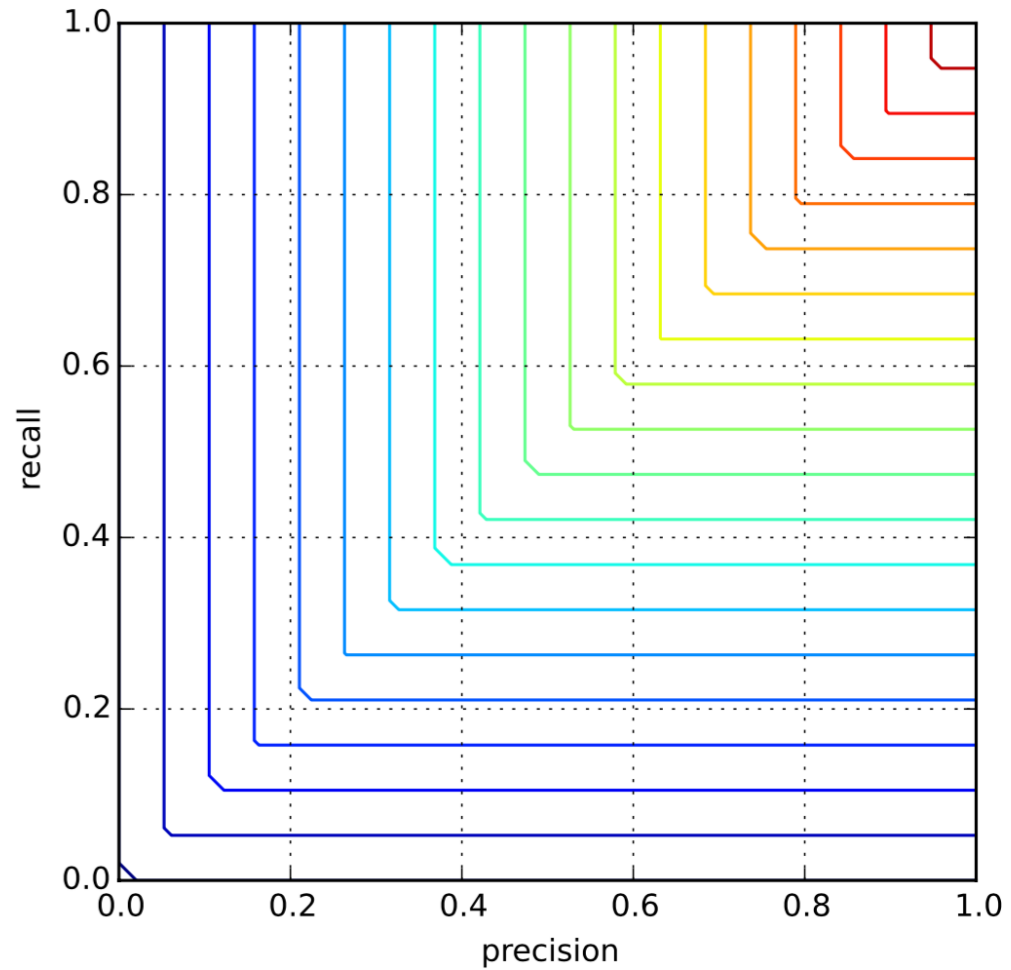
$$A = \frac{1}{2}(\textit{precision} + \textit{recall})$$

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



Минимум

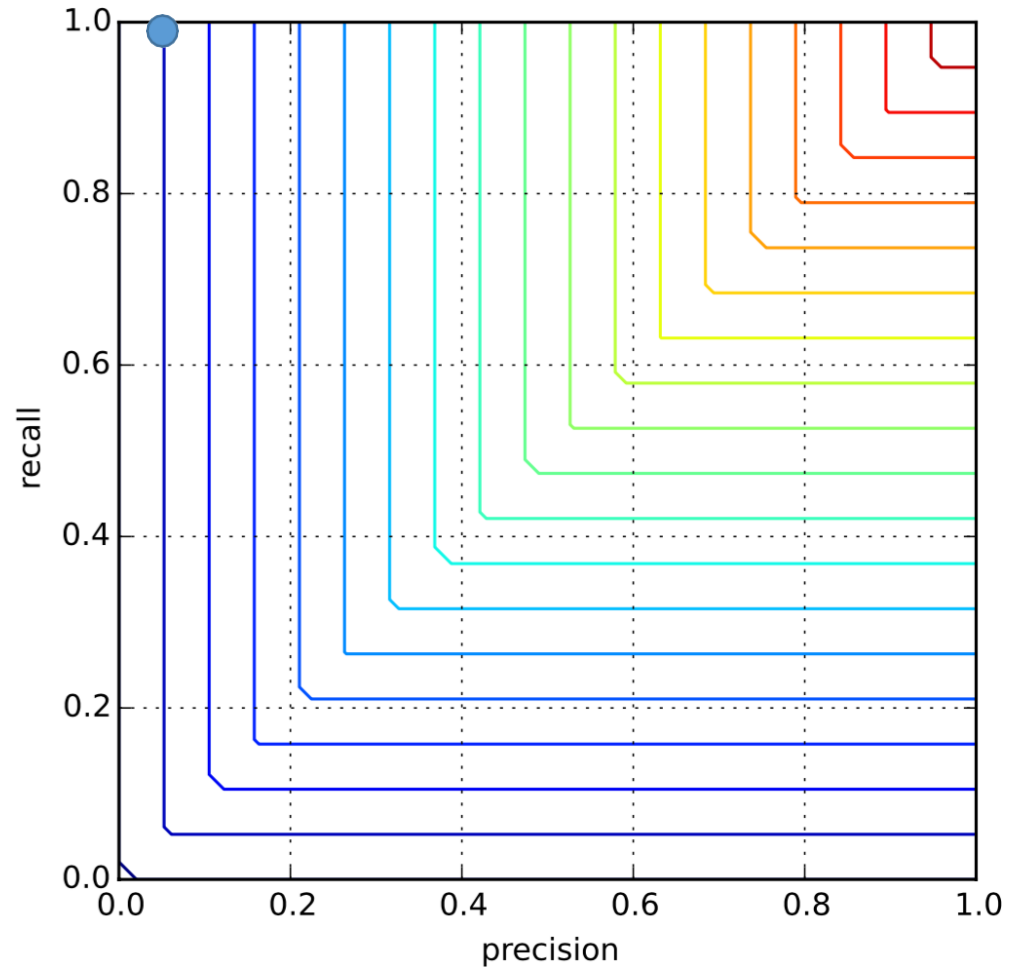
$$M = \min(\text{precision}, \text{recall})$$



Минимум (1/3)

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

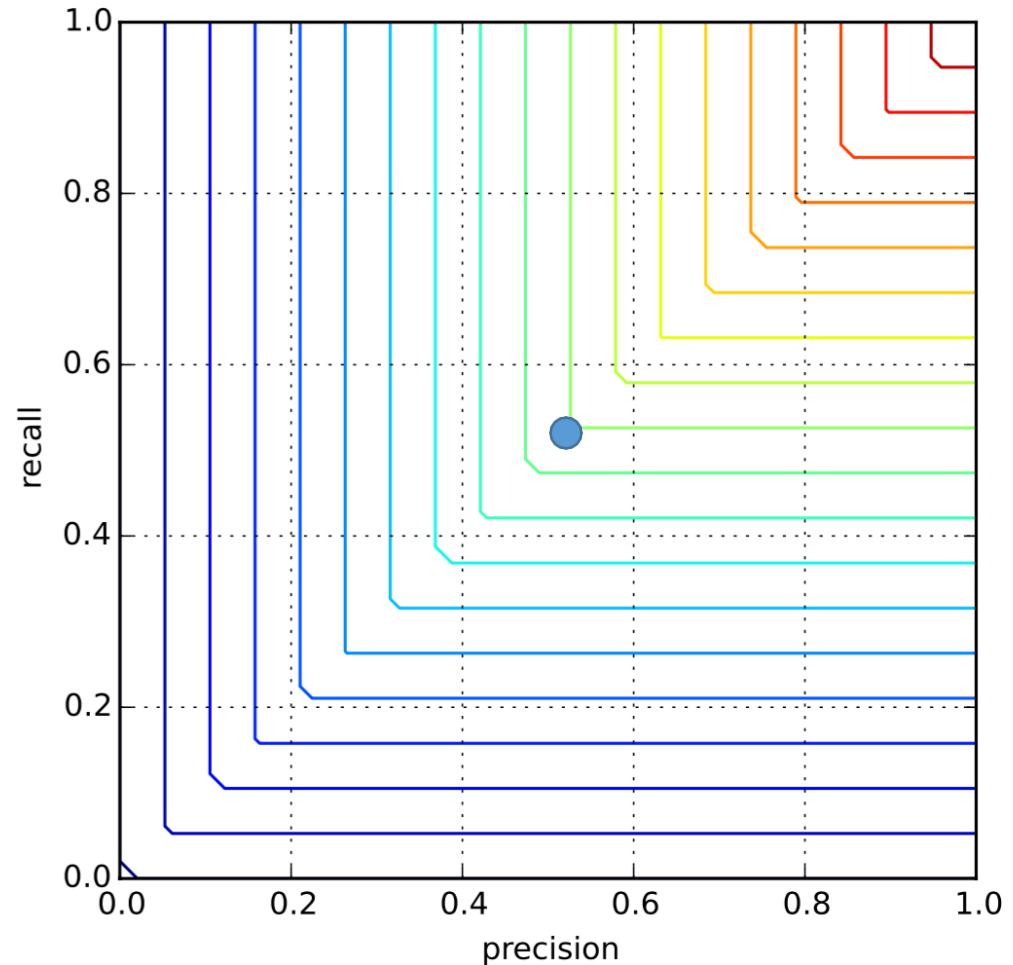
- $\text{precision} = 0.05$
- $\text{recall} = 1$
- $M = 0.05$



Минимум (2/3)

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

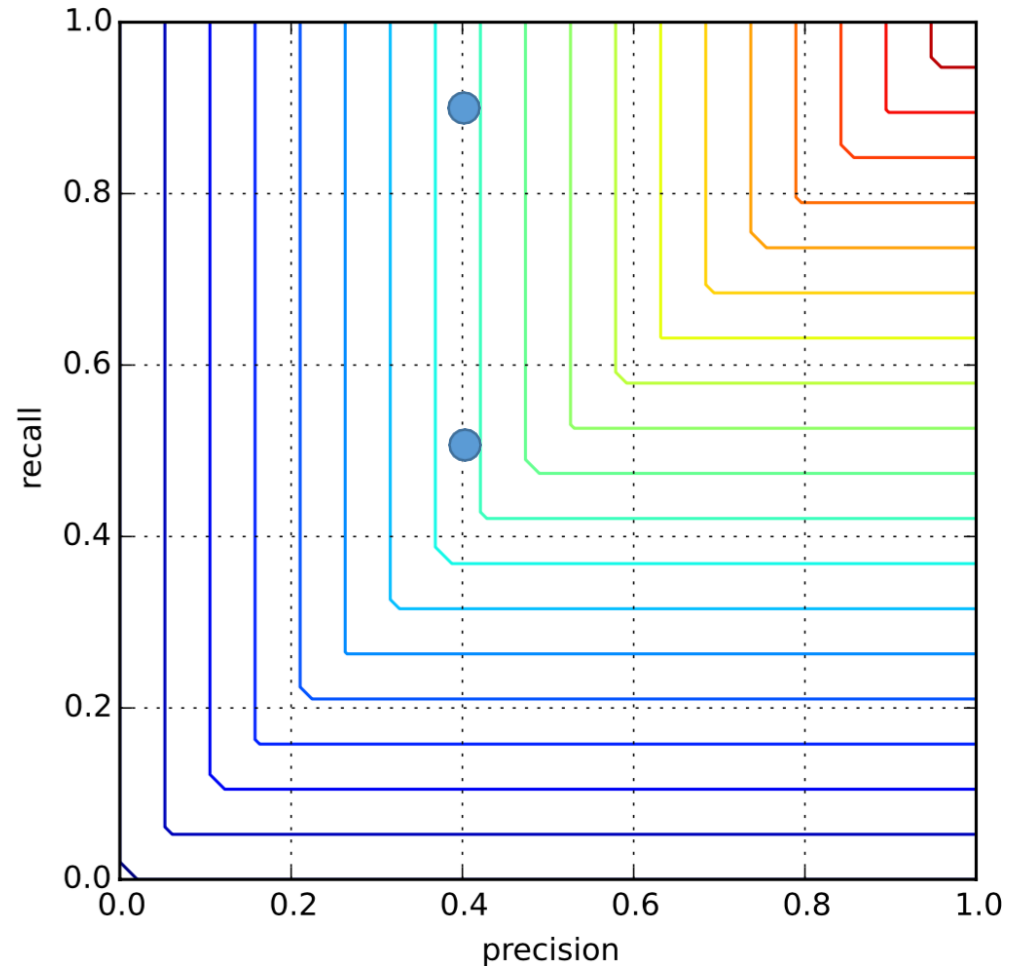
- $\text{precision} = 0.55$
- $\text{recall} = 0.55$
- $M = 0.55$



Минимум (3/3)

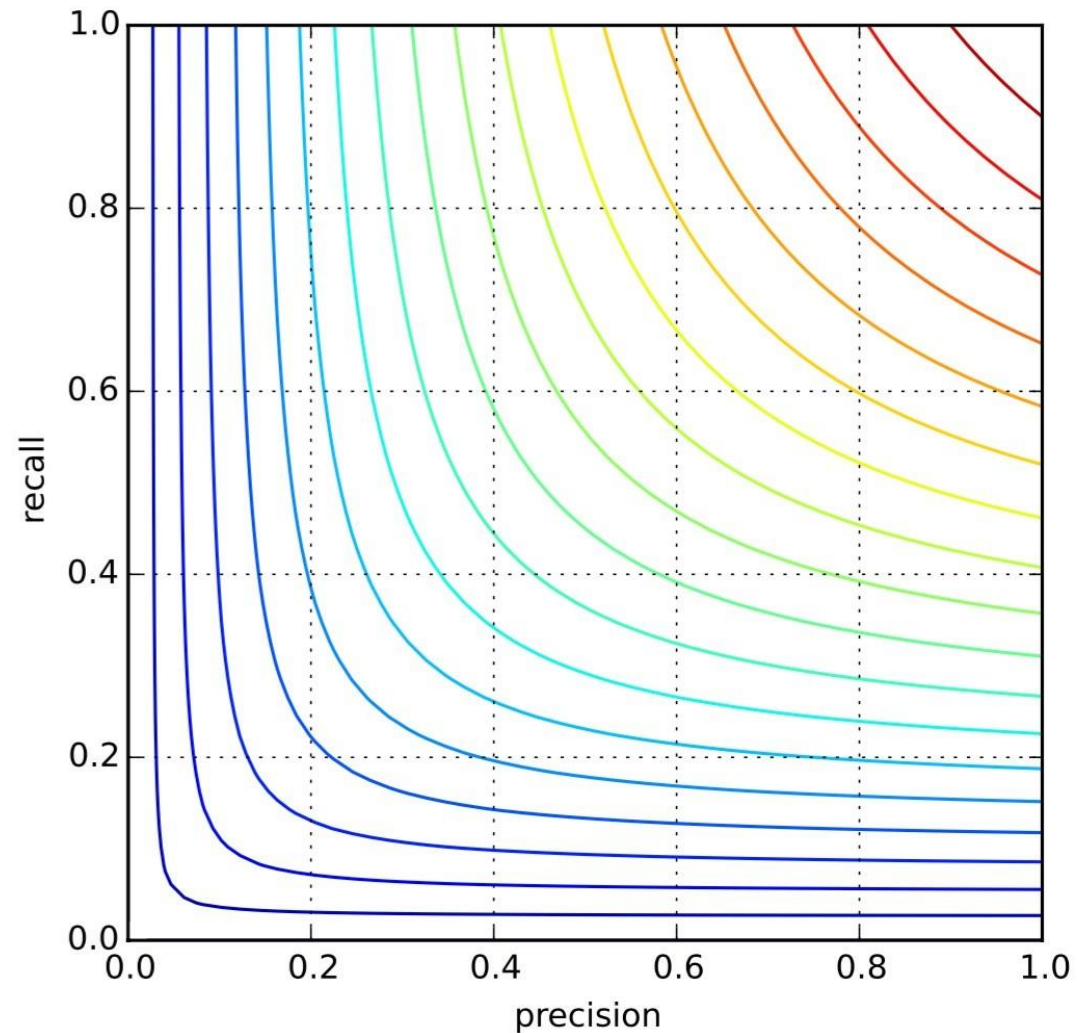
$$M = \min(\text{precision}, \text{recall})$$

- $\text{precision} = 0.4$
- $\text{recall} = 0.5$
- $M = 0.4$
- $\text{precision} = 0.4$
- $\text{recall} = 0.9$
- $M = 0.4$
- Но второй лучше!



F-мера (1/2)

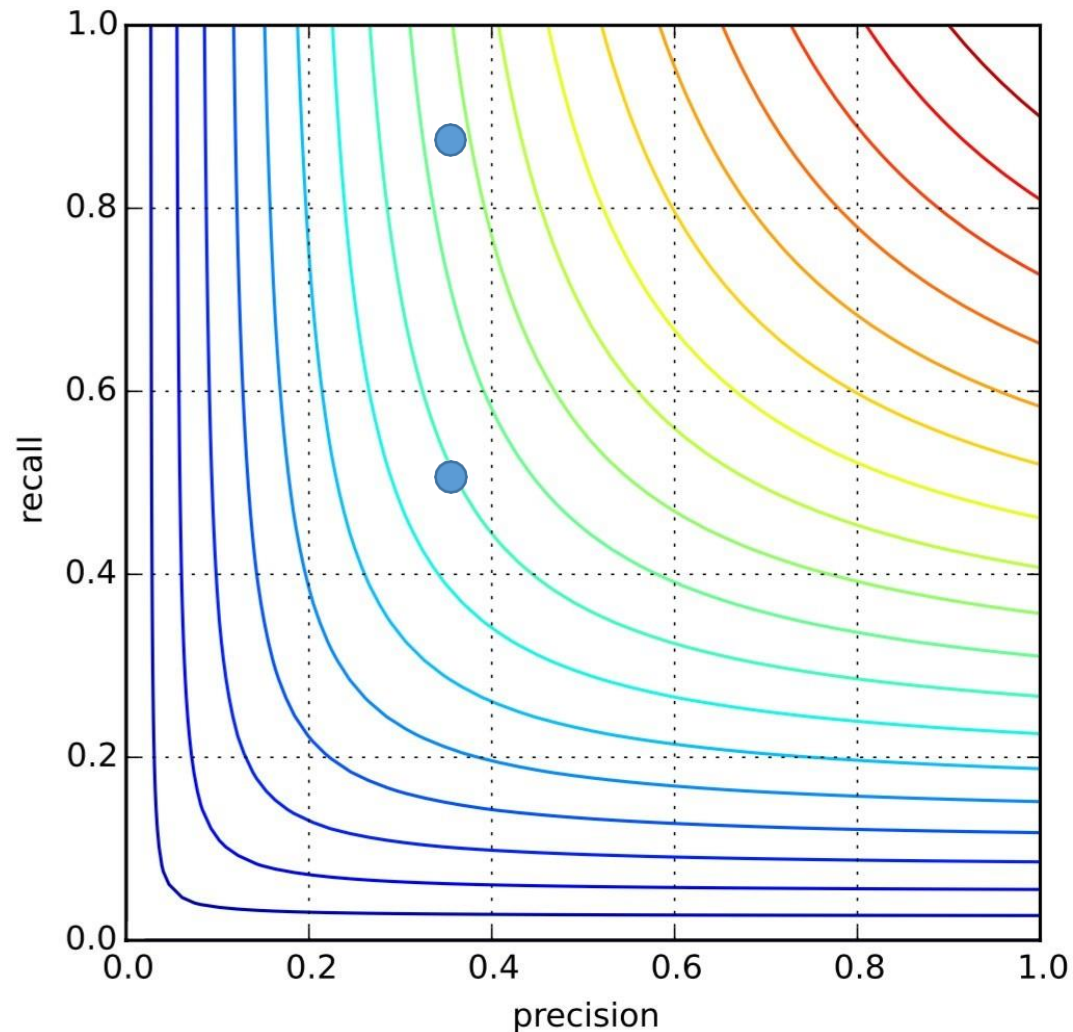
$$F = \frac{2 * precision * recall}{precision + recall}$$



F-мера (1/2)

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

- $precision = 0.4$
- $recall = 0.5$
- $M = 0.44$
- $precision = 0.4$
- $recall = 0.9$
- $M = 0.55$



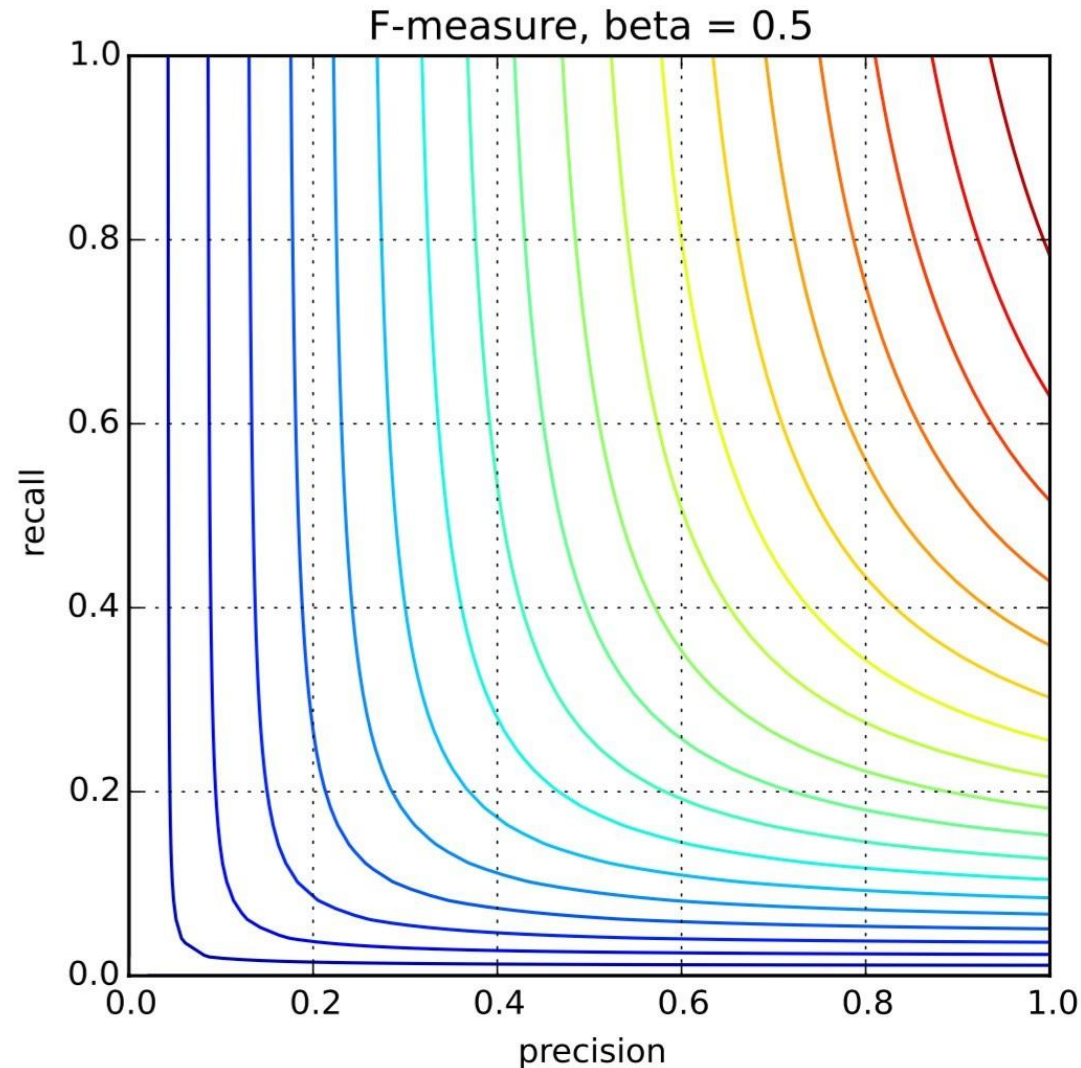
F-мера (2/2)

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

F-мера (2/2)

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

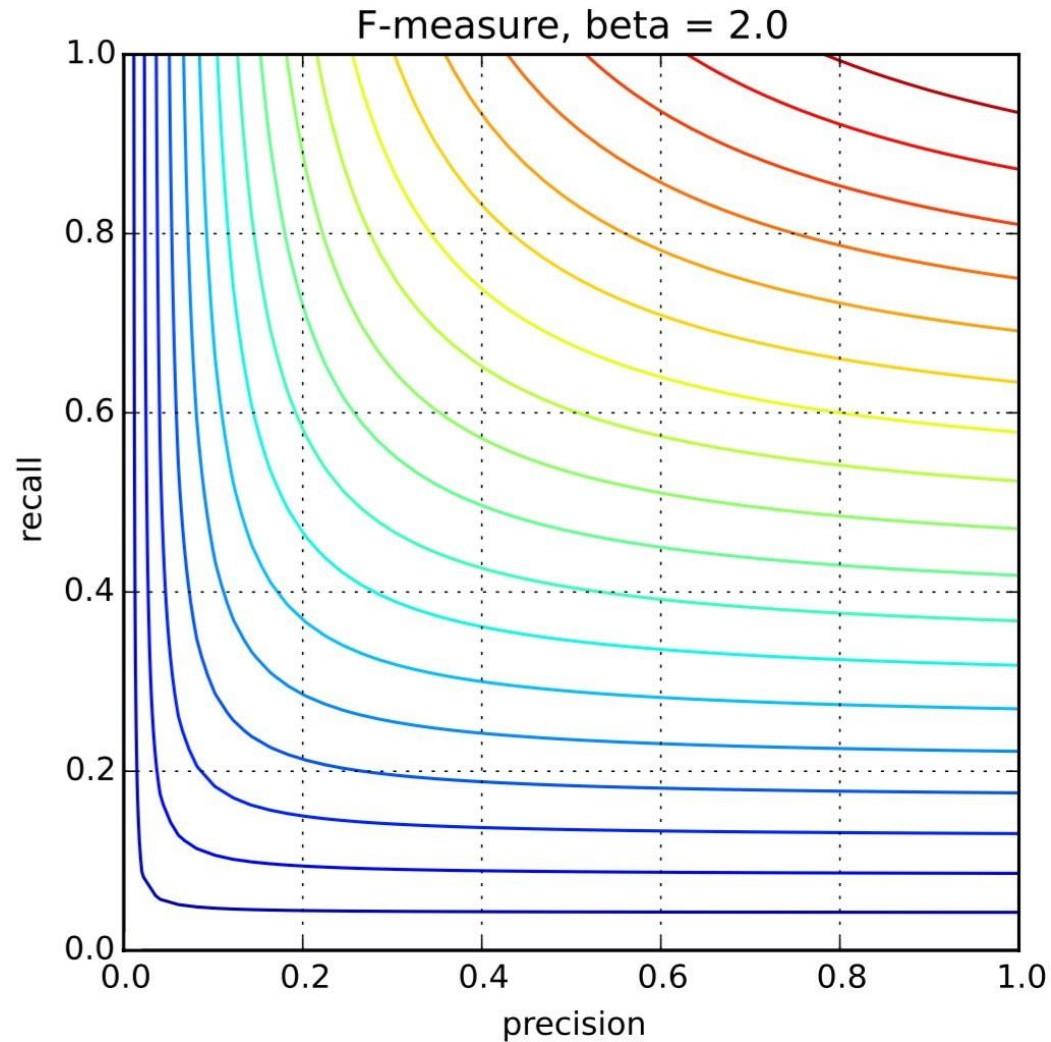
- $\beta = 0.5$
- Важнее точность



F-мера (2/2)

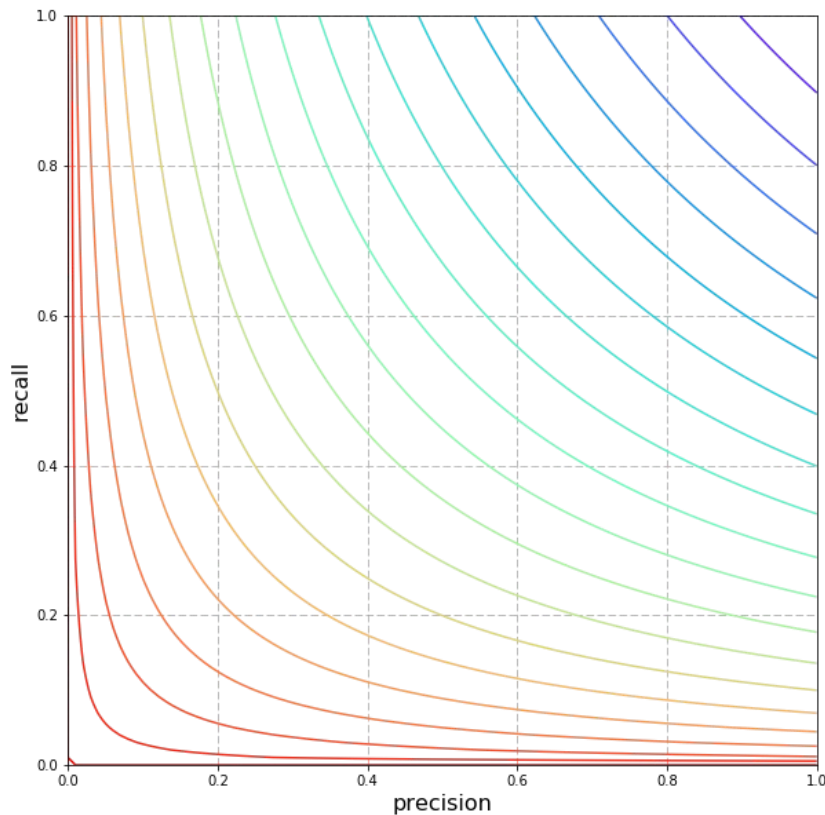
$$F_{\beta} = (1 + \beta^2) \frac{\text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}}$$

- $\beta = 2$
- Важнее полнота

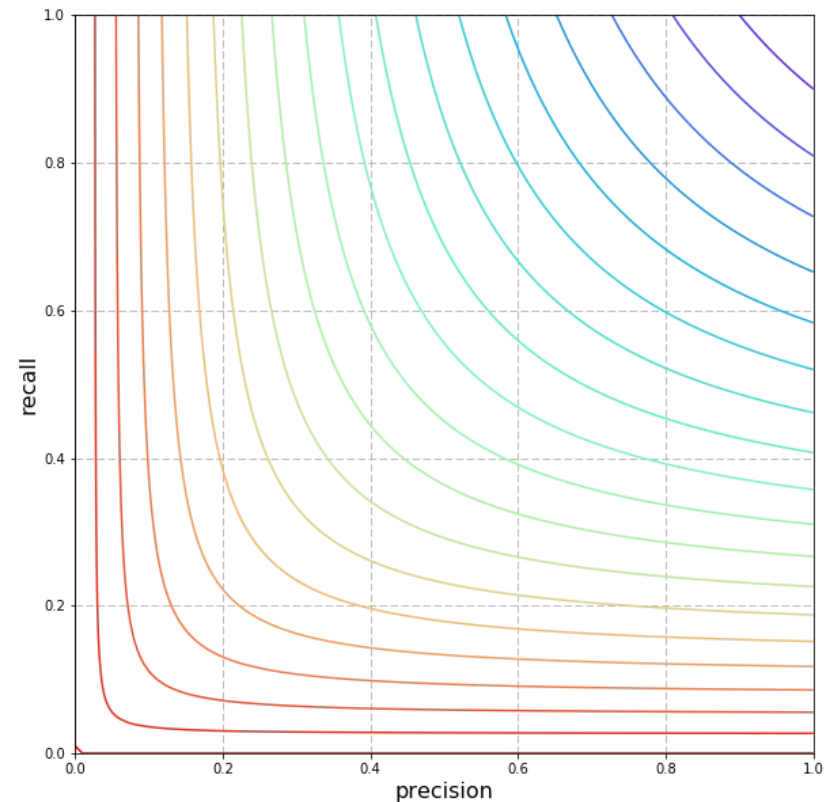


Геометрическое среднее (1/2)

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



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



Геометрическое среднее (2/2)

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

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

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

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

Метрики качества ранжирования

Классификатор

- Линейный классификатор:

$$a(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - t) = 2[\langle \mathbf{w}, \mathbf{x} \rangle > t] - 1$$

- $\langle \mathbf{w}, \mathbf{x} \rangle$ — оценка принадлежности классу +1
- Нередко $t = 0$

Оценка принадлежности (1/4)

- Высокий порог:
 - Мало объектов относим к +1
 - Точность выше
 - Полнота ниже
- Низкий порог:
 - Много объектов относим к +1
 - Точность ниже
 - Полнота выше


Оценка принадлежности (2/4)

-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

Оценка принадлежности (2/4)

-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

Оценка принадлежности (2/4)



-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

Оценка принадлежности (2/4)

-1	+1	-1	+1	-1	+1	-1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

Оценка принадлежности (3/4)

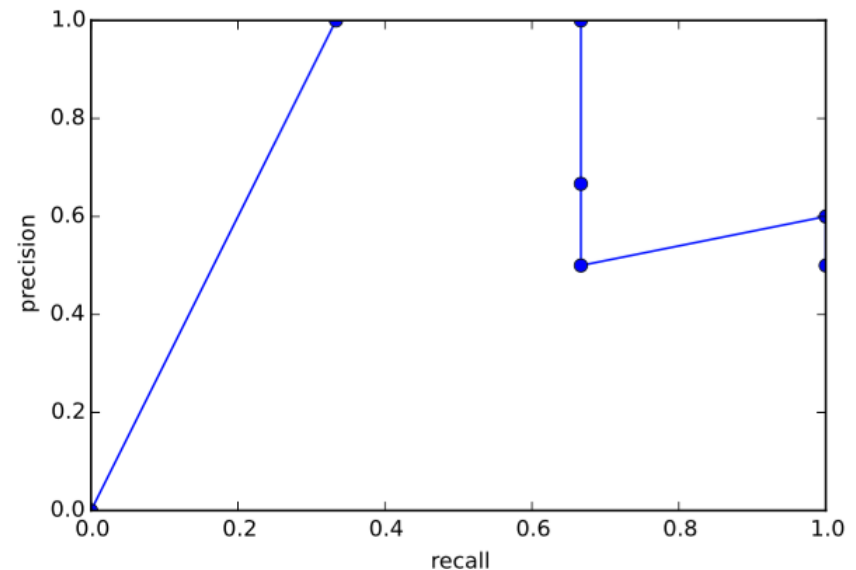
- Как оценить качество $b(\mathbf{x})$?
- Порог выбирается позже
- Порог зависит от ограничений на точность или полноту

Оценка принадлежности (4/4)

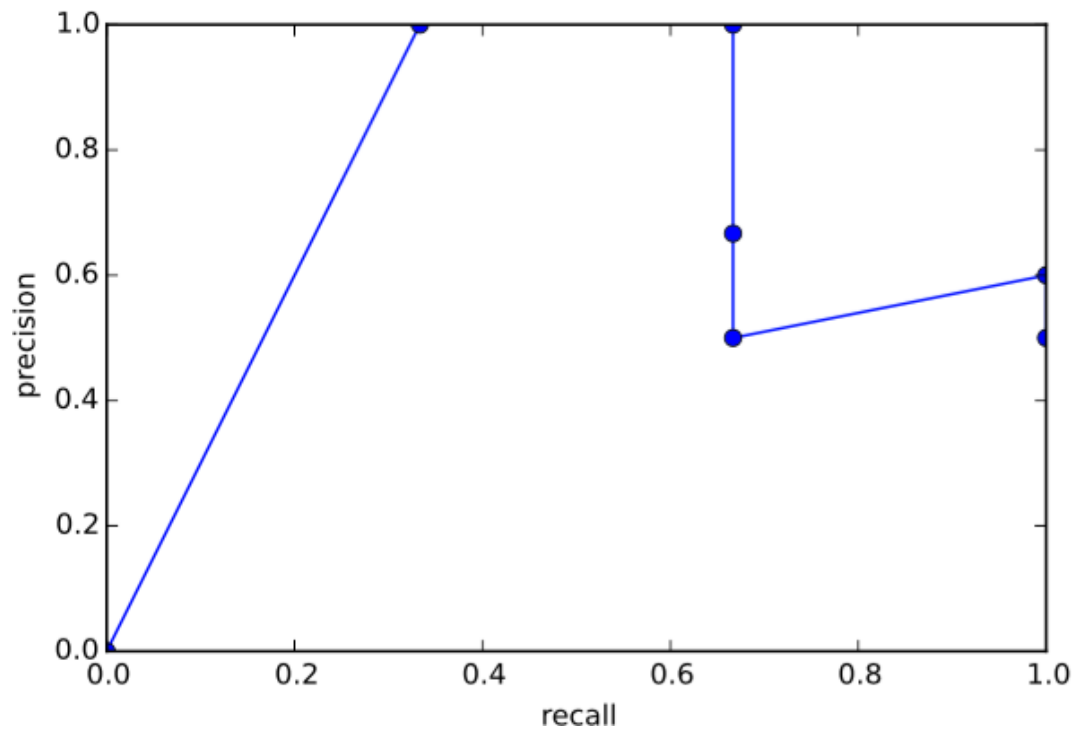
- Пример: кредитный скоринг
- $b(\mathbf{x})$ — оценка вероятности возврата кредита
- $a(\mathbf{x}) = [b(\mathbf{x}) > 0.5]$
- $precision = 0.1, recall = 0.7$
- В чем дело — в пороге или в алгоритме?

PR-кривая (1/5)

- Кривая точности-полноты
- Ось X — полнота
- Ось Y — точность
- Точки — значения точности и полноты при последовательных порогах

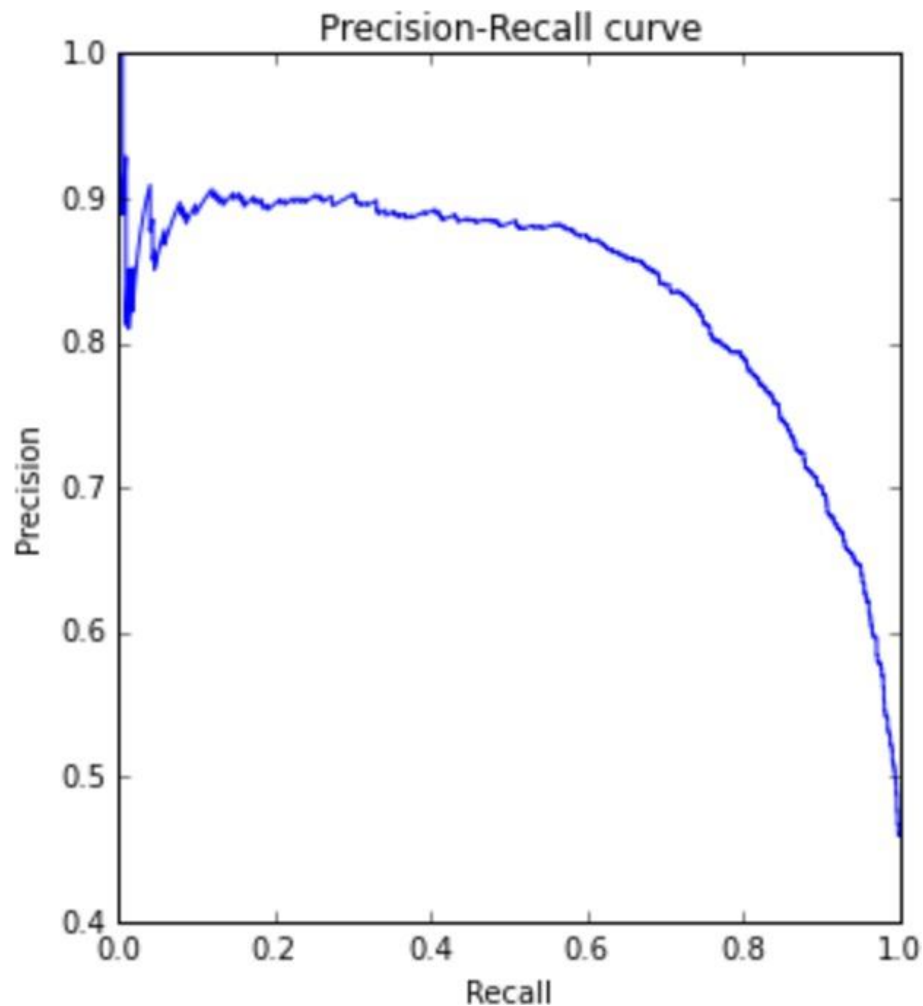


PR-кривая (2/5)



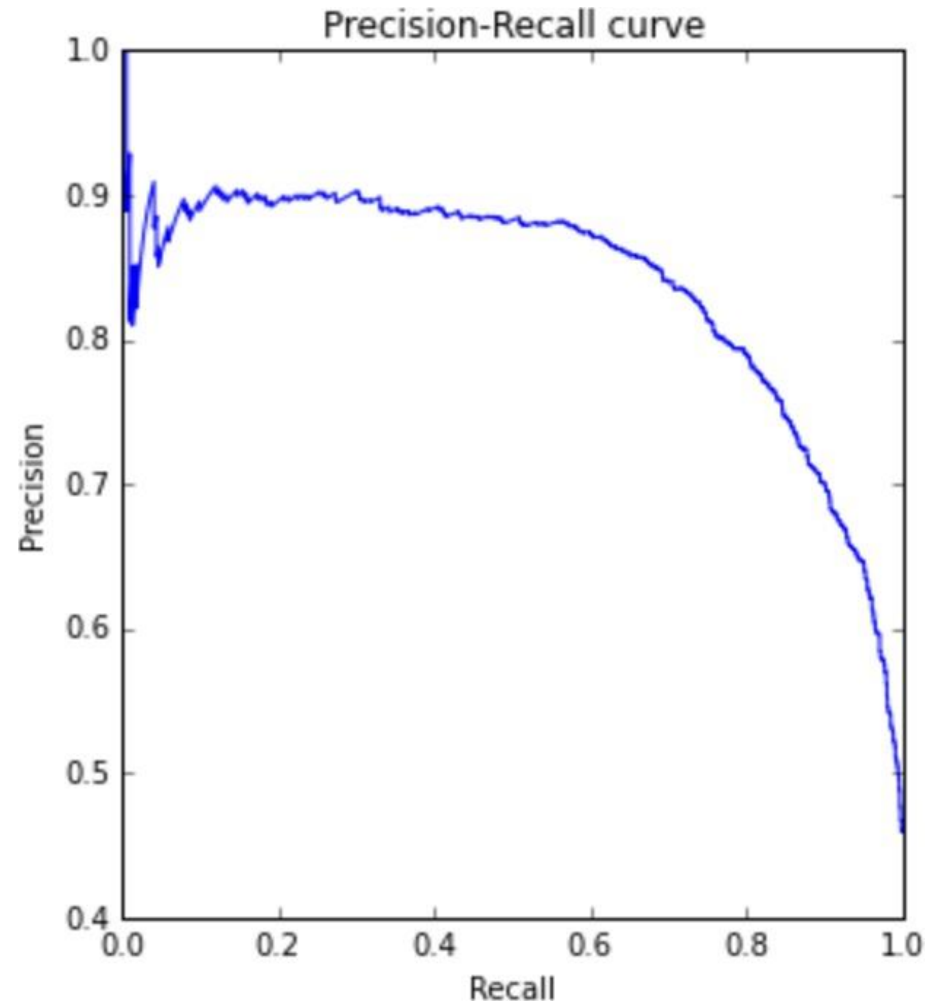
$b(\mathbf{x})$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

PR-кривая (в реальности) (3/5)

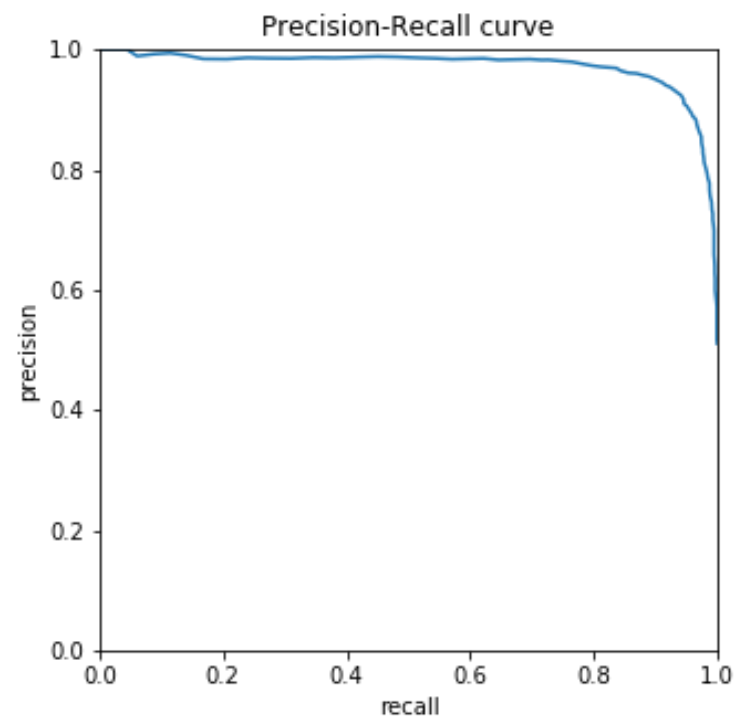
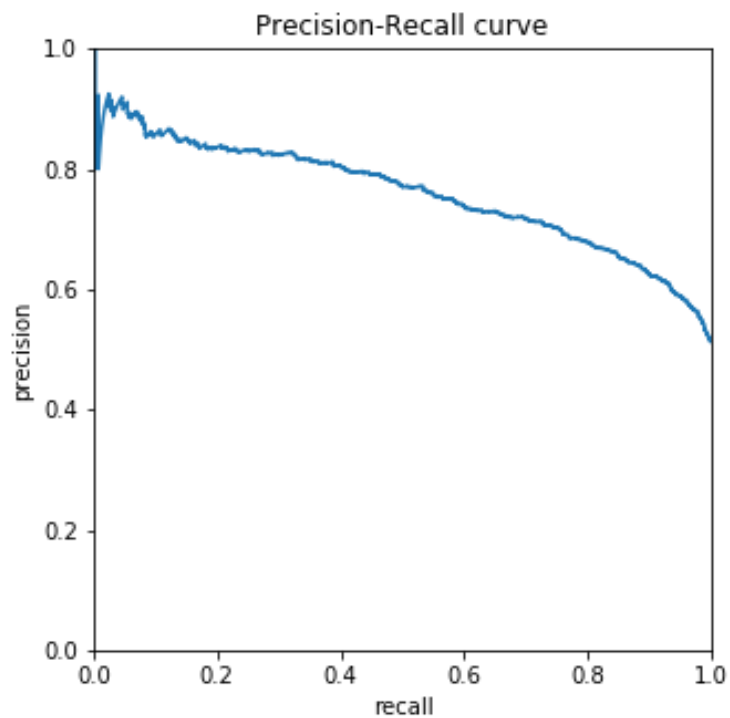


PR-кривая (4/5)

- Левая точка: $(0, 1)$
- Правая точка: $(1, r)$,
 r — доля положительных объектов
- Для идеального классификатора проходит через $(1, 1)$
- AUC-PRC — площадь под PR-кривой



PR-кривая (5/5)



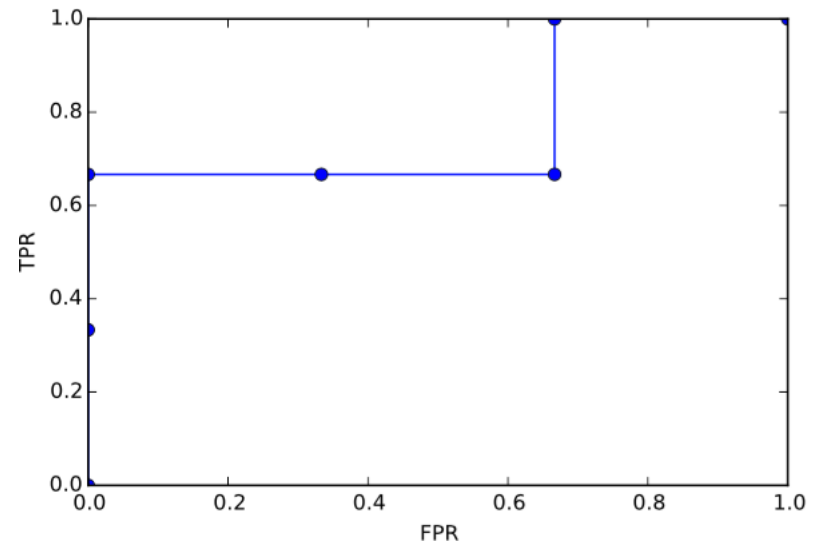
ROC-кривая (1/5)

- Receiver Operating Characteristic
- Ось X — False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

- Ось Y — True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$



ROC-кривая (1/5)

- Receiver Operating Characteristic
- Ось X — False Positive Rate

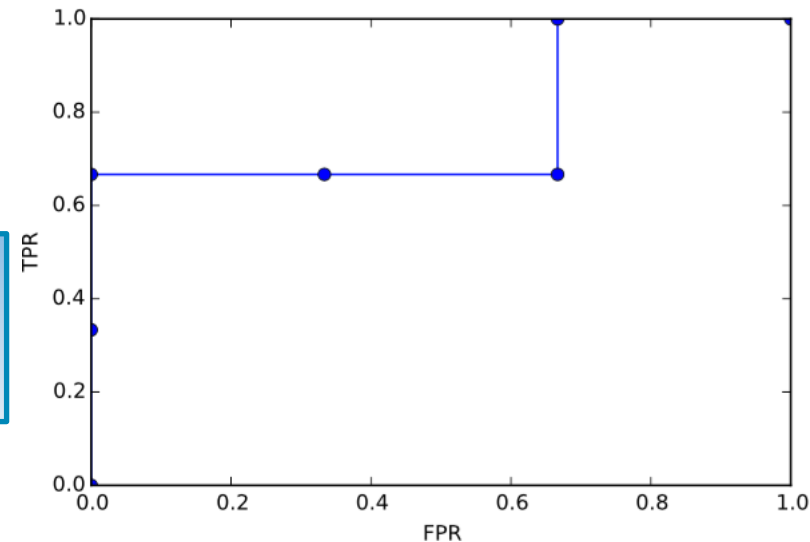
$$FPR = \frac{FP}{FP + TN}$$

Число отрицательных объектов

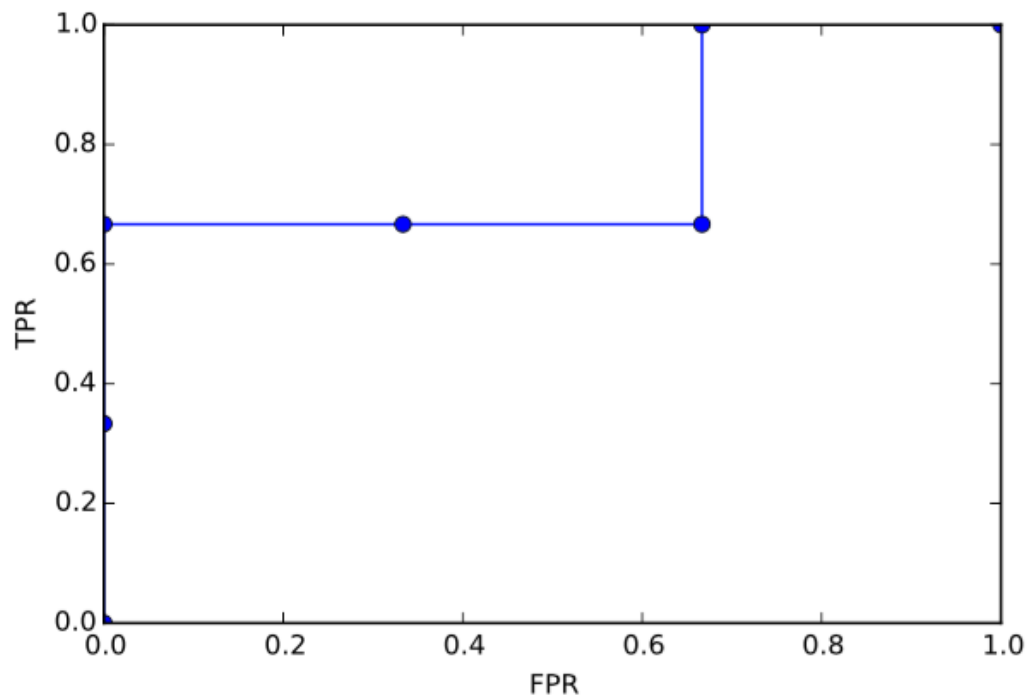
- Ось Y — True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

Число положительных объектов

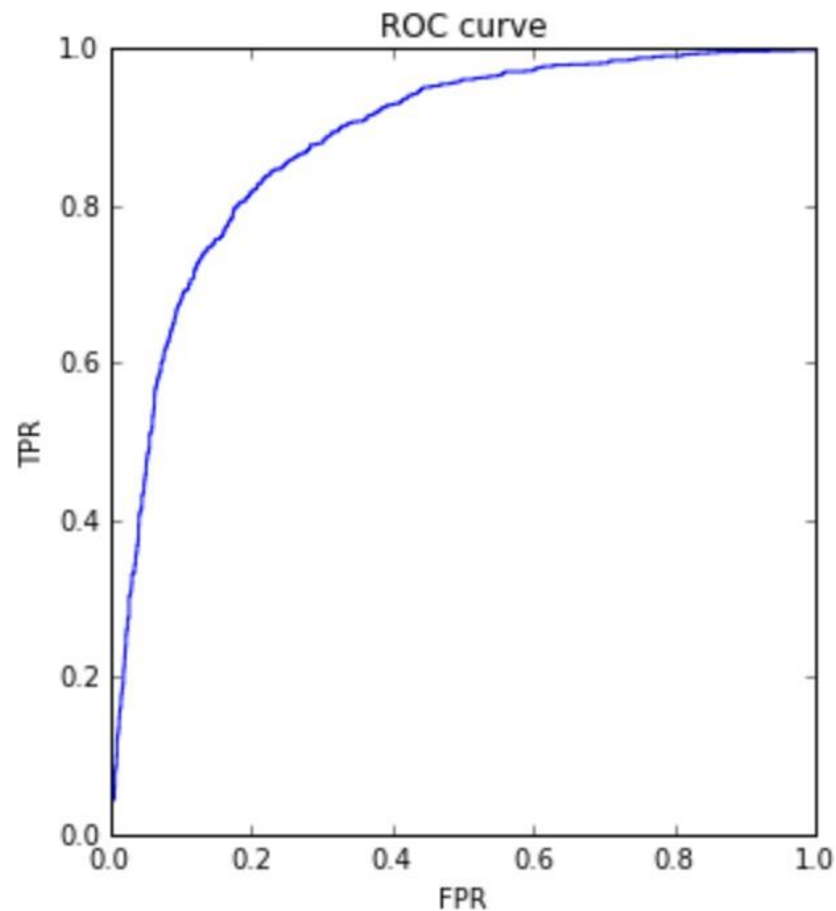


ROC-кривая (2/5)



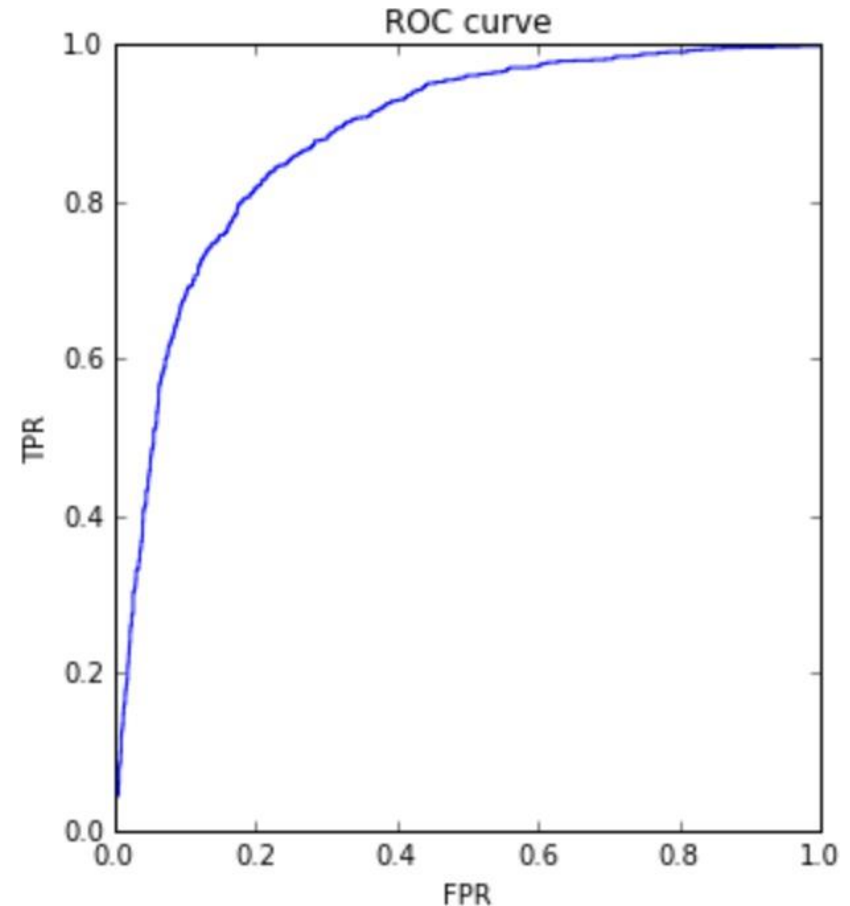
$b(\mathbf{x})$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

ROC-кривая (в реальности) (3/5)

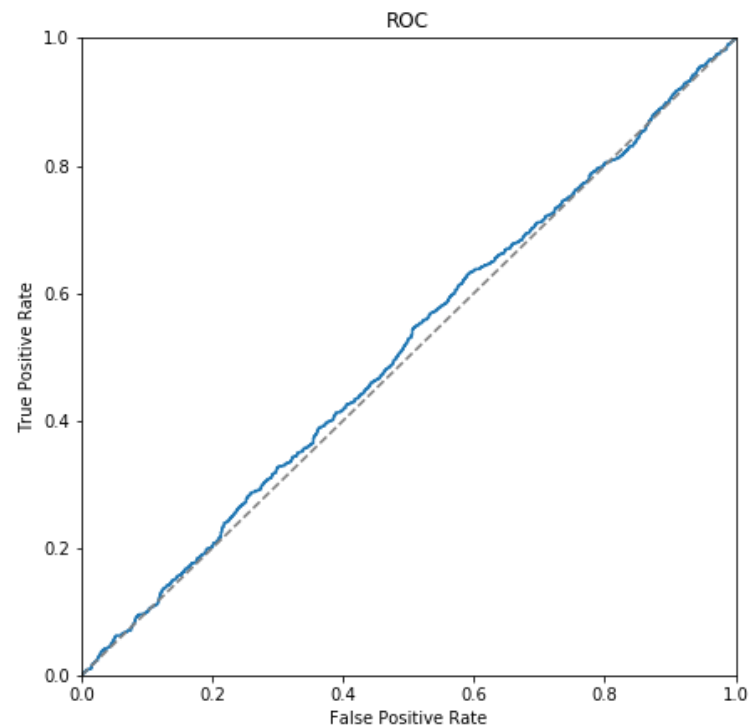
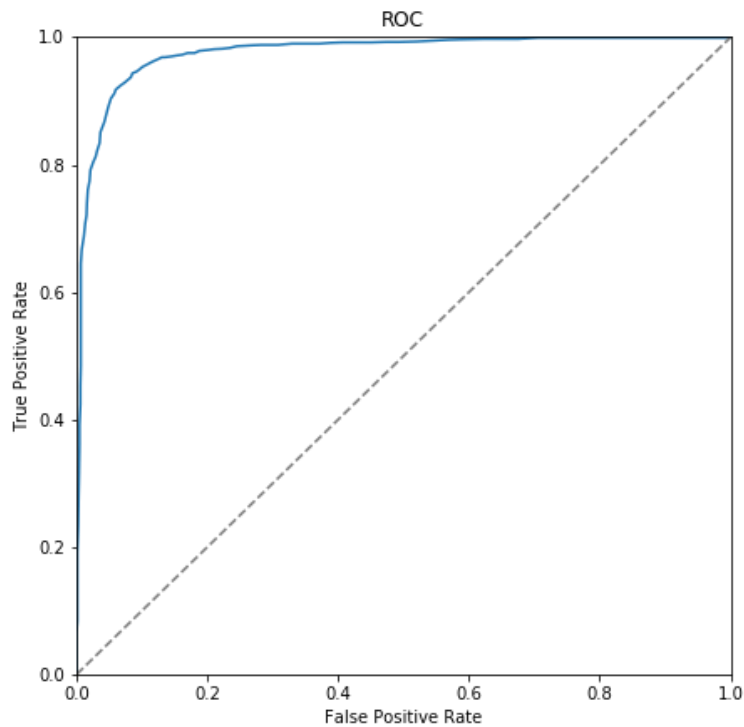


ROC-кривая (4/5)

- Левая точка: $(0, 0)$
- Правая точка: $(1, 1)$
- Для идеального классификатора проходит через $(0, 1)$
- AUC-ROC — площадь под ROC-кривой



ROC-кривая (5/5)



AUC-ROC

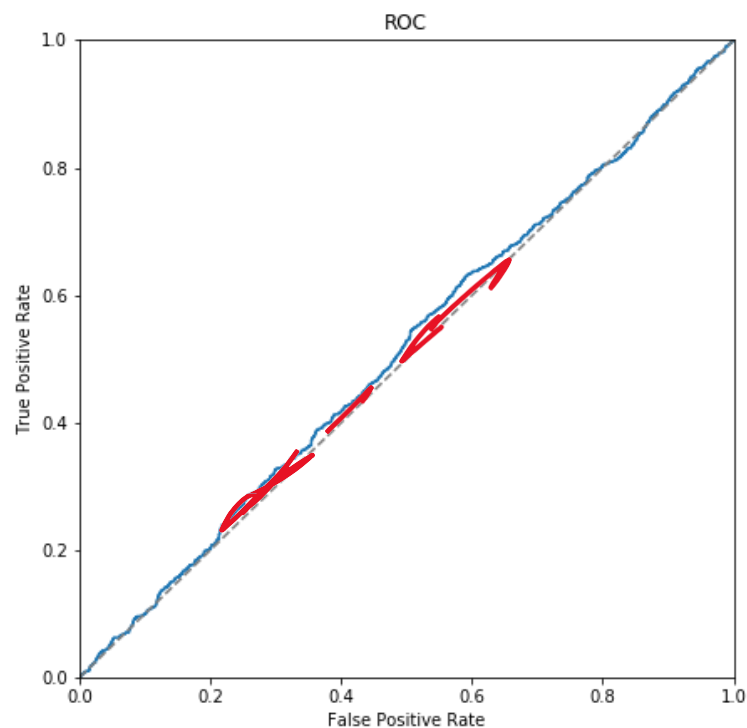
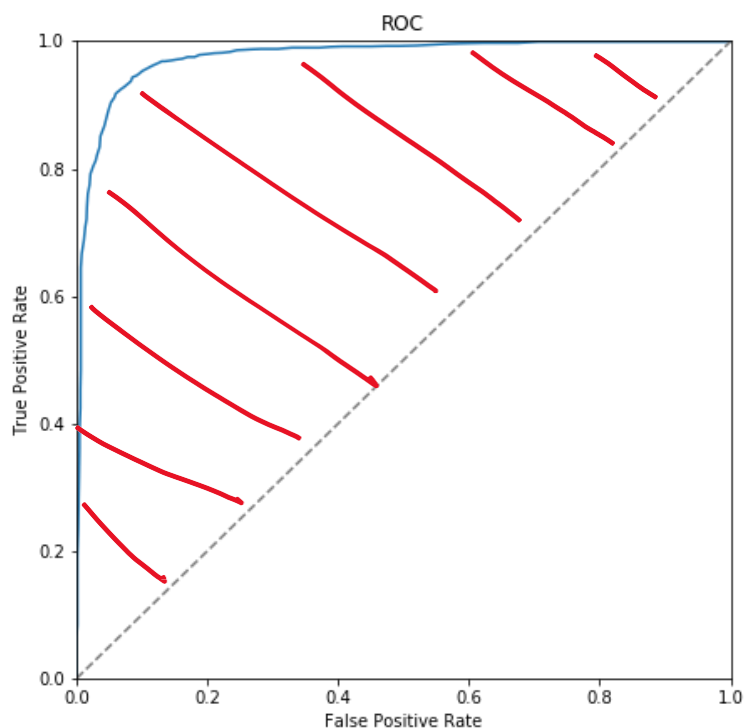
$$FPR = \frac{FP}{FP+TN};$$

$$TPR = \frac{TP}{TP+FN}$$

- FPR и TPR нормируются на размеры классов
- AUC-ROC не поменяется при изменении баланса классов
- Идеальный алгоритм: $AUC - ROC = 1$
- Худший алгоритм: $AUC - ROC \approx 0.5$
- Интересные интерпретации: например, это примерно доля пар правильно упорядоченных объектов

Коэффициент Джини

$$Gini = 2 * (AUC - ROC - 0.5)$$



AUC-PRC

$$precision = \frac{TP}{TP+FP};$$

$$recall = \frac{TP}{TP+FN}$$

- Точность поменяется при изменении баланса классов
- AUC-PRC идеального алгоритма зависит от баланса классов
- Проще интерпретировать, если выборка несбалансированная
- Лучше, если задачу надо решать в терминах точности и полноты

Пример

- $AUC - ROC = 0.95$
- $AUC - PRC = 0.001$



Пример

- Выберем конкретный классификатор
- $a(\mathbf{x}) = 1$ — 50095 объектов
- Из них $FP = 50000$, $TP = 95$
- $TPR = 0.95$, $FPR = 0.05$
- $precision = 0.0019$, $recall = 0.95$

