

Основы машинного обучения

Лекция 9

Метрики качества классификации

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2022

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 существенно 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

- $r_1 = 0.8$
- $r_2 = 0.9$
- $\frac{r_2 - r_1}{r_1} = 12.5\%$

- $r_1 = 0.5$
- $r_2 = 0.75$
- $\frac{r_2 - r_1}{r_1} = 50\%$

- $r_1 = 0.001$
- $r_2 = 0.01$
- $\frac{r_2 - r_1}{r_1} = 900\%$

Цены ошибок

- Пример: кредитный скоринг
- 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

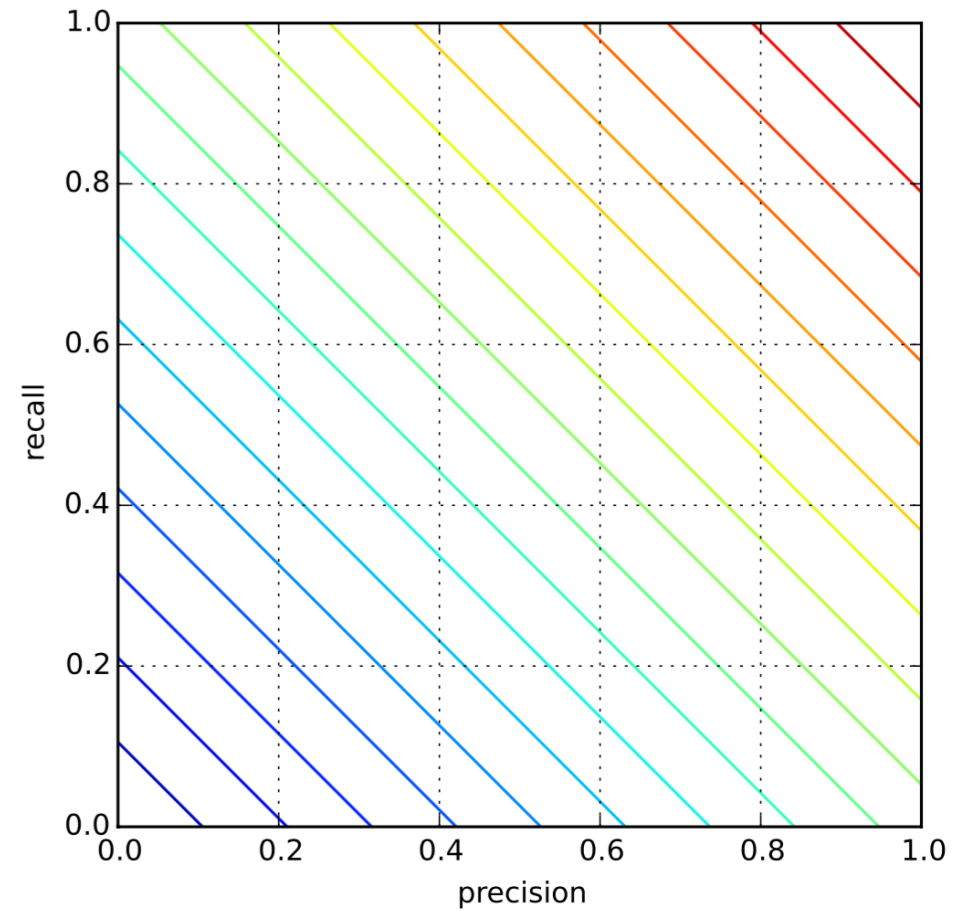
Совмещение точности и
полноты

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

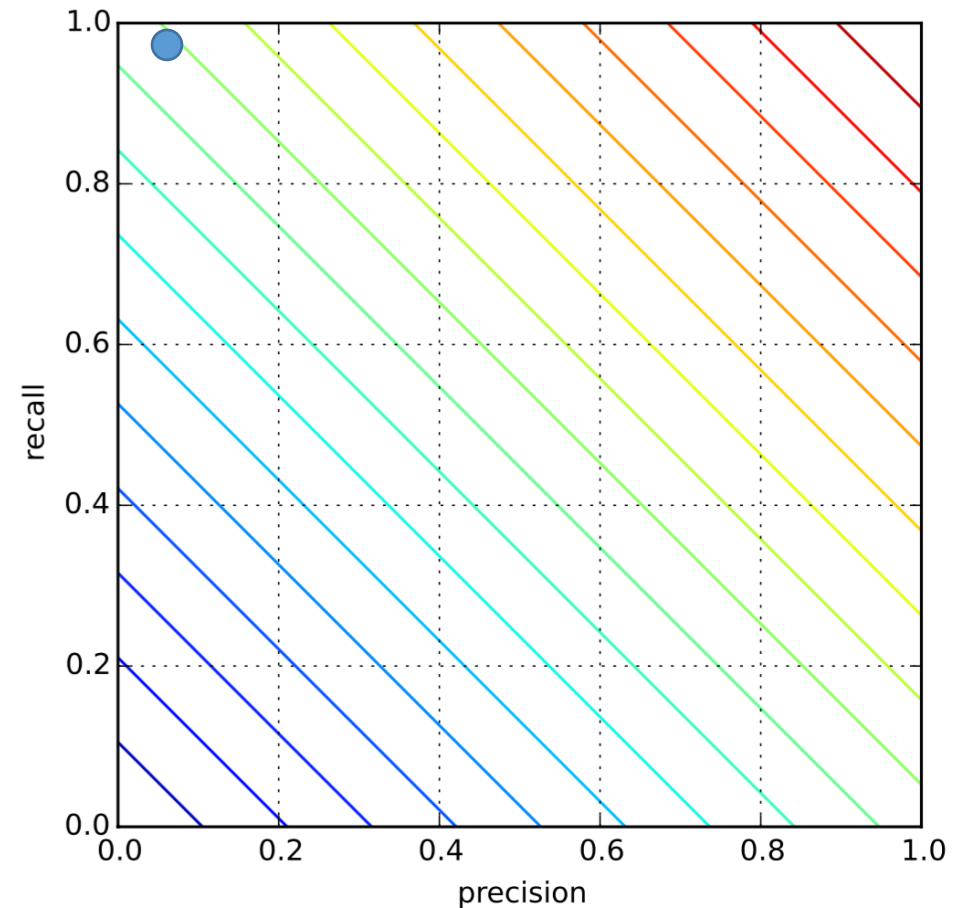
$$A = \frac{1}{2}(\text{precision} + \text{recall})$$



Арифметическое среднее

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

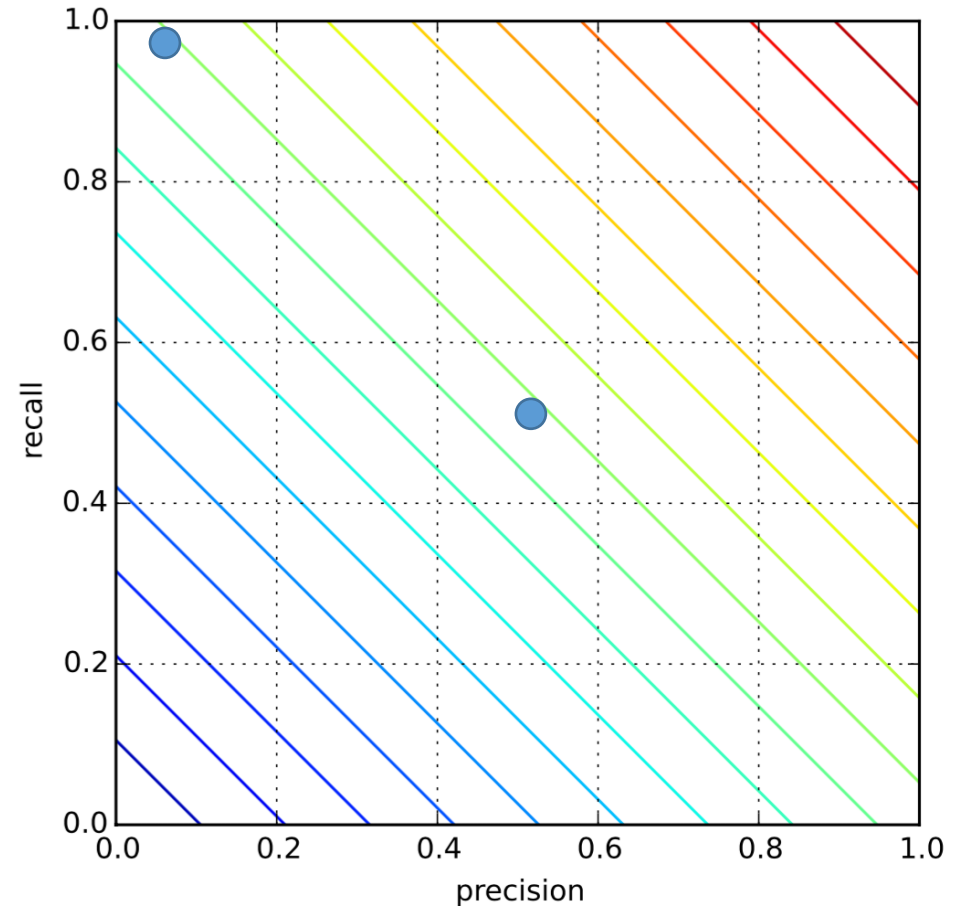
- precision = 0.1
- recall = 1
- $A = 0.55$
- Плохой алгоритм



Арифметическое среднее

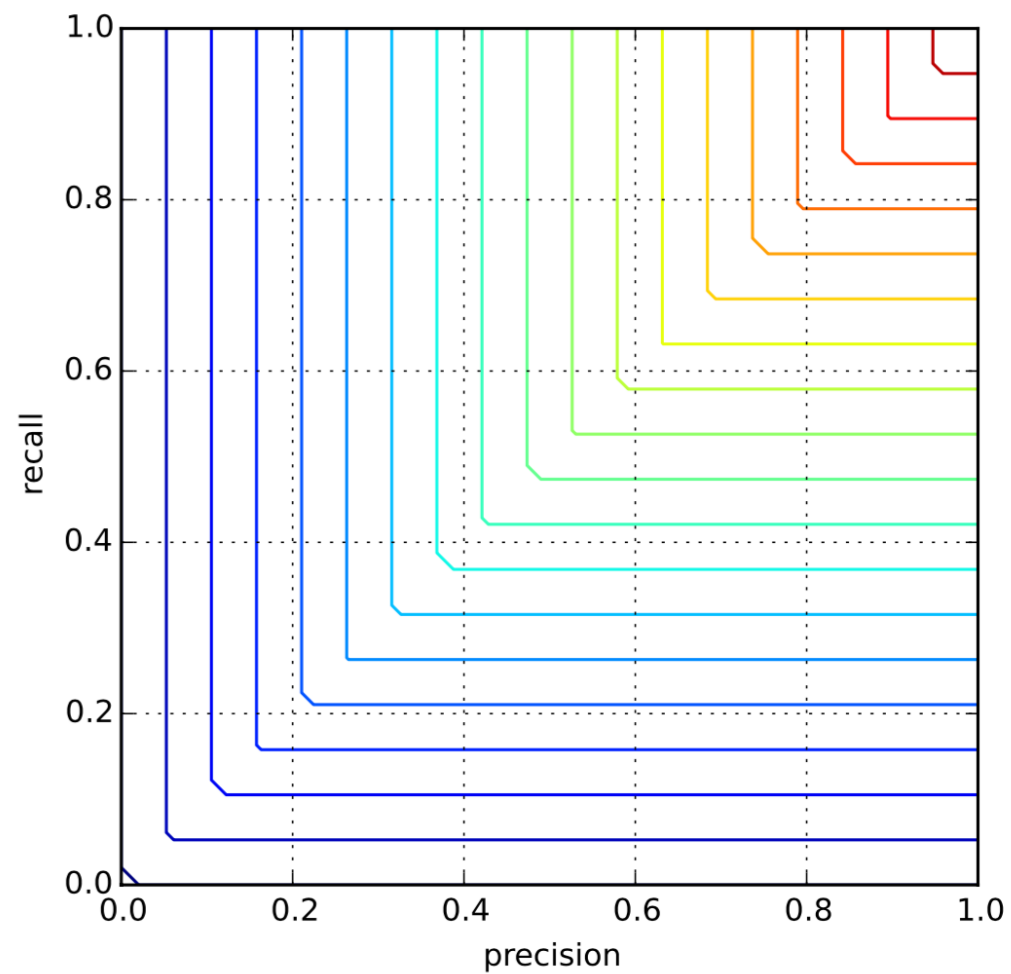
$$A = \frac{1}{2} (\text{precision} + \text{recall})$$

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



Минимум

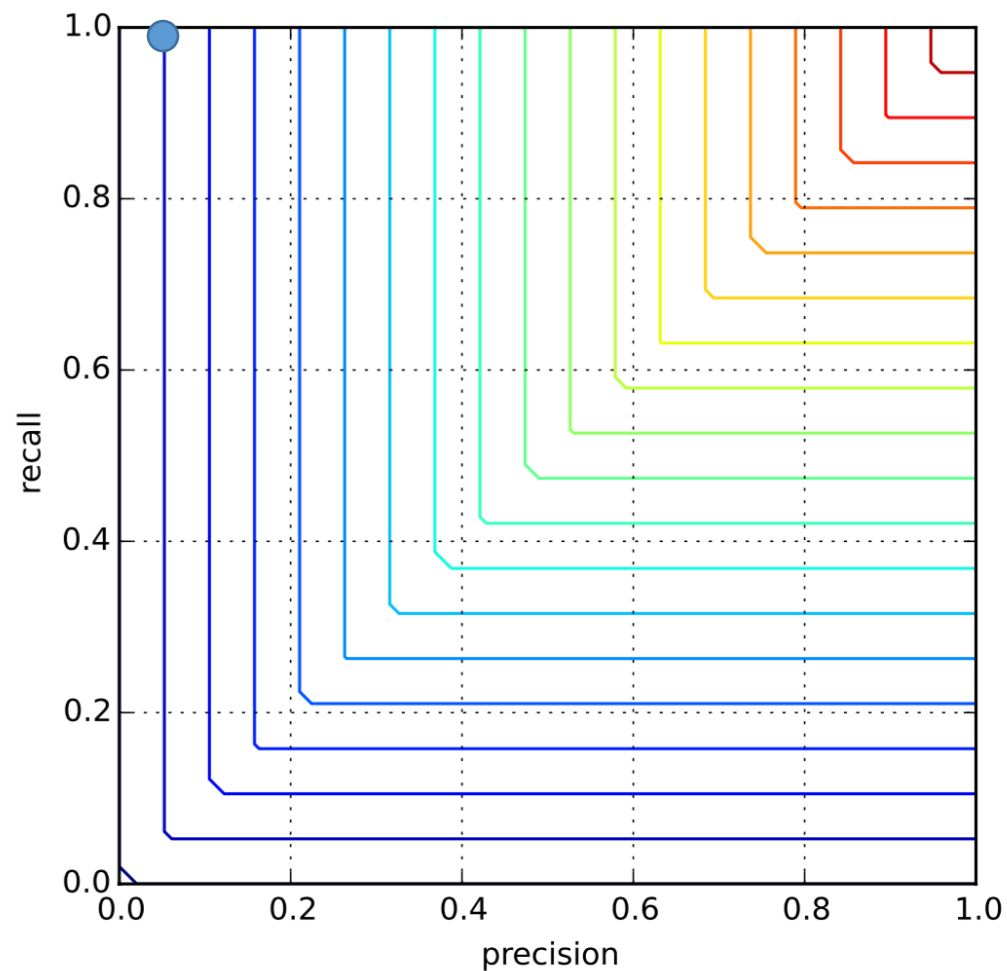
$$M = \min(\text{precision}, \text{recall})$$



Минимум

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

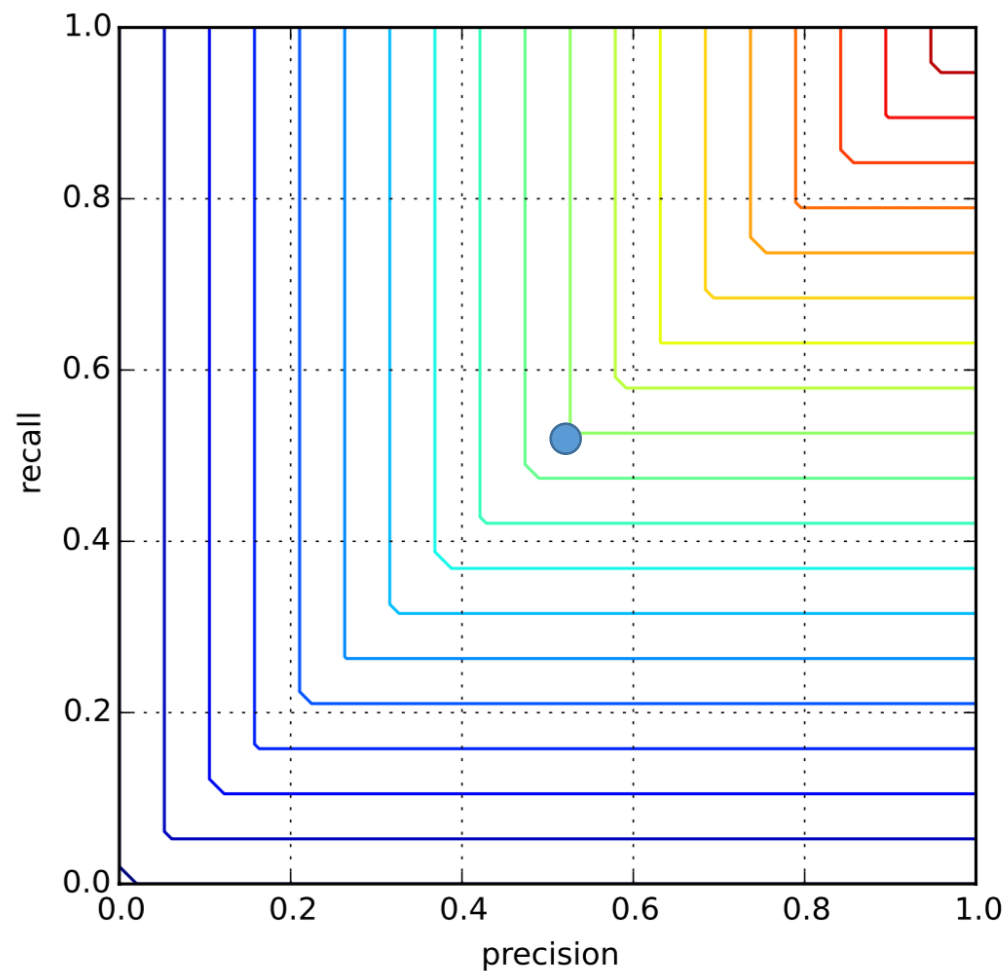
- precision = 0.05
- recall = 1
- $M = 0.05$



Минимум

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

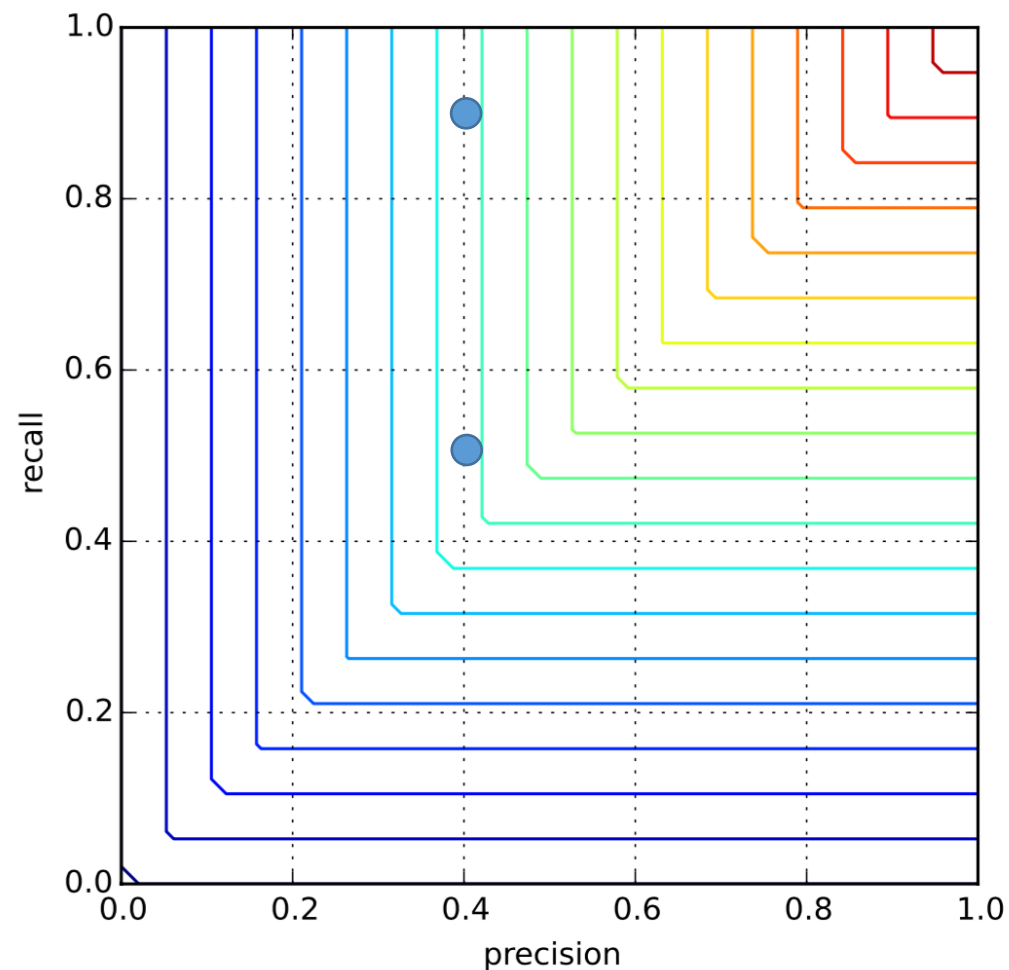
- precision = 0.55
- recall = 0.55
- $M = 0.55$



Минимум

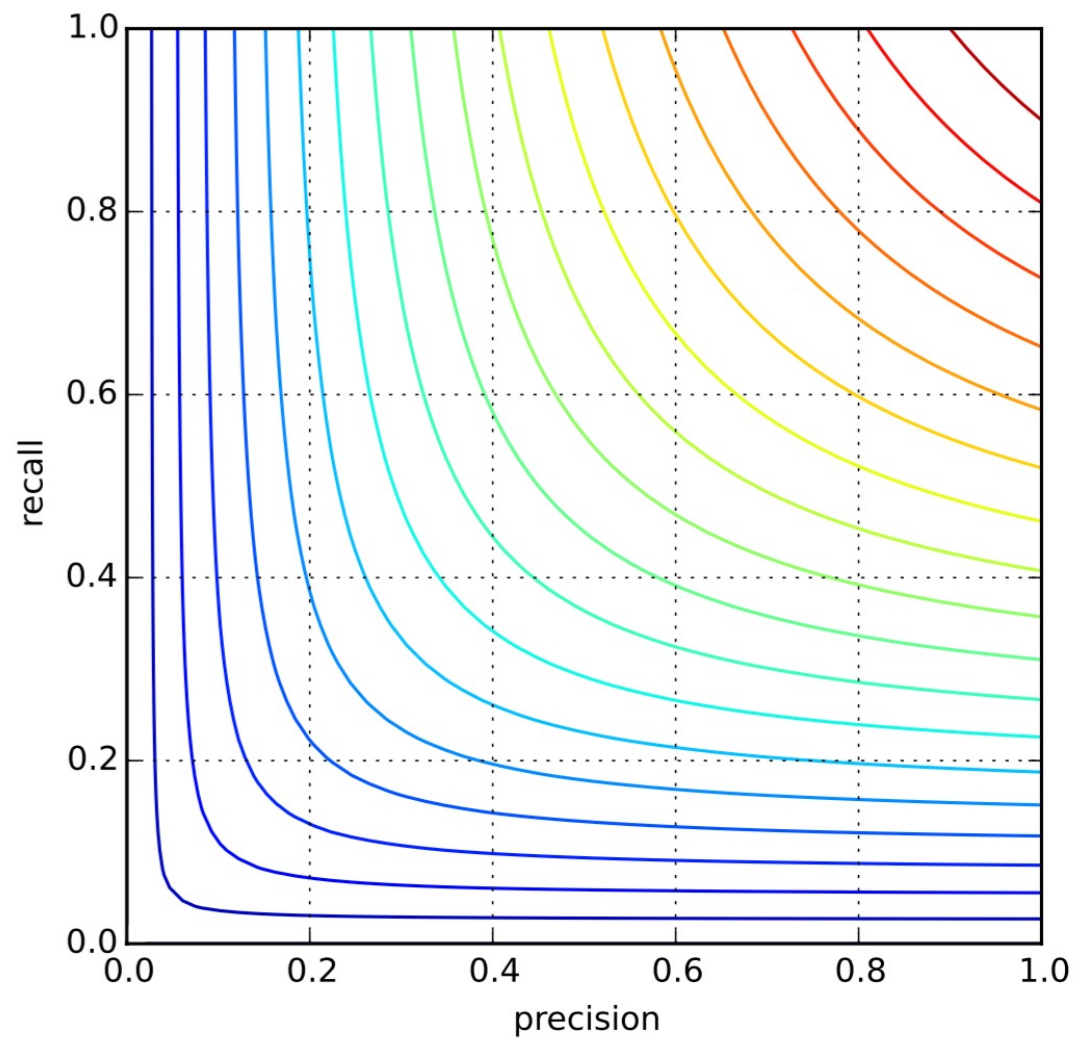
$$M = \min(\text{precision}, \text{recall})$$

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



F-measure

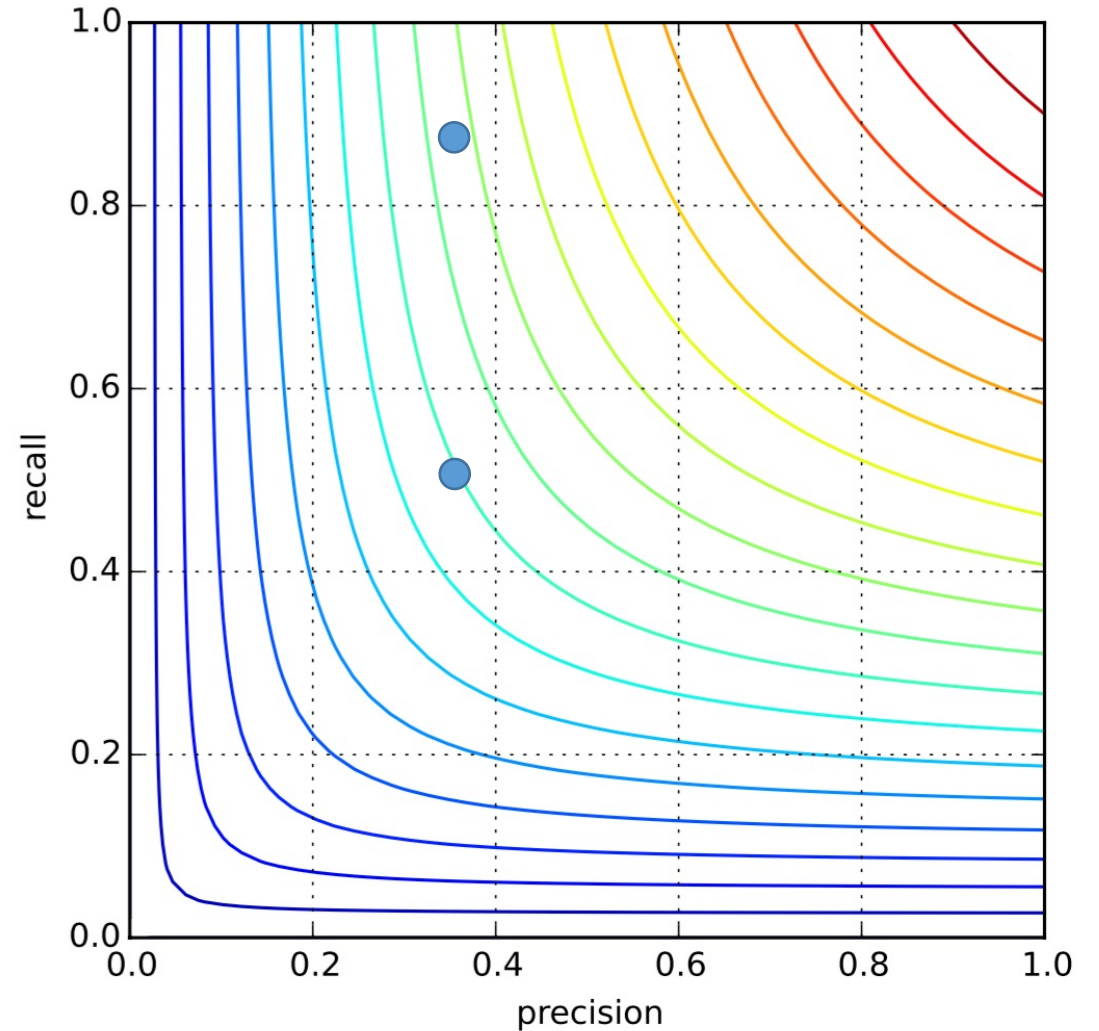
$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



F-meapa

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

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



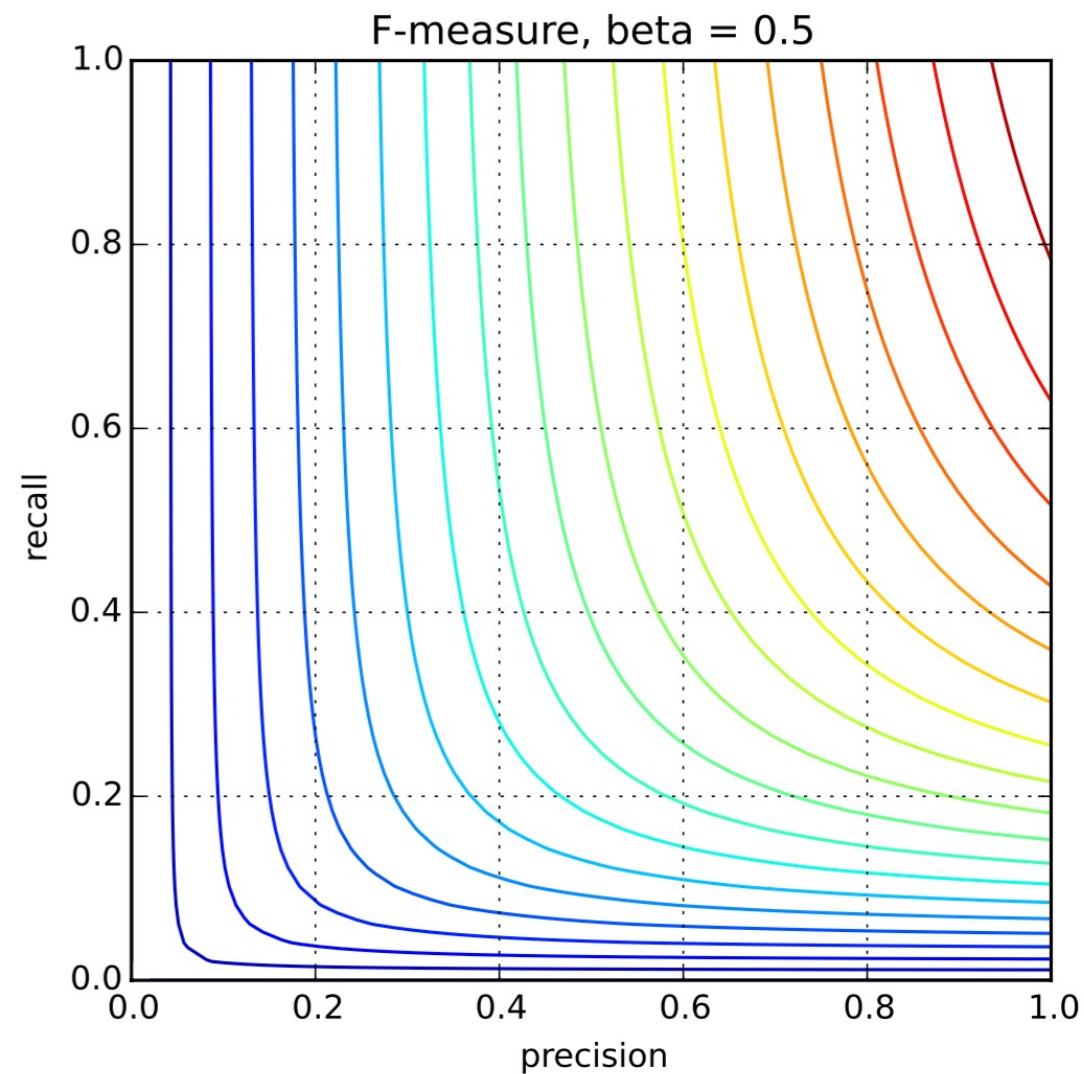
F-measure

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

F-мера

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

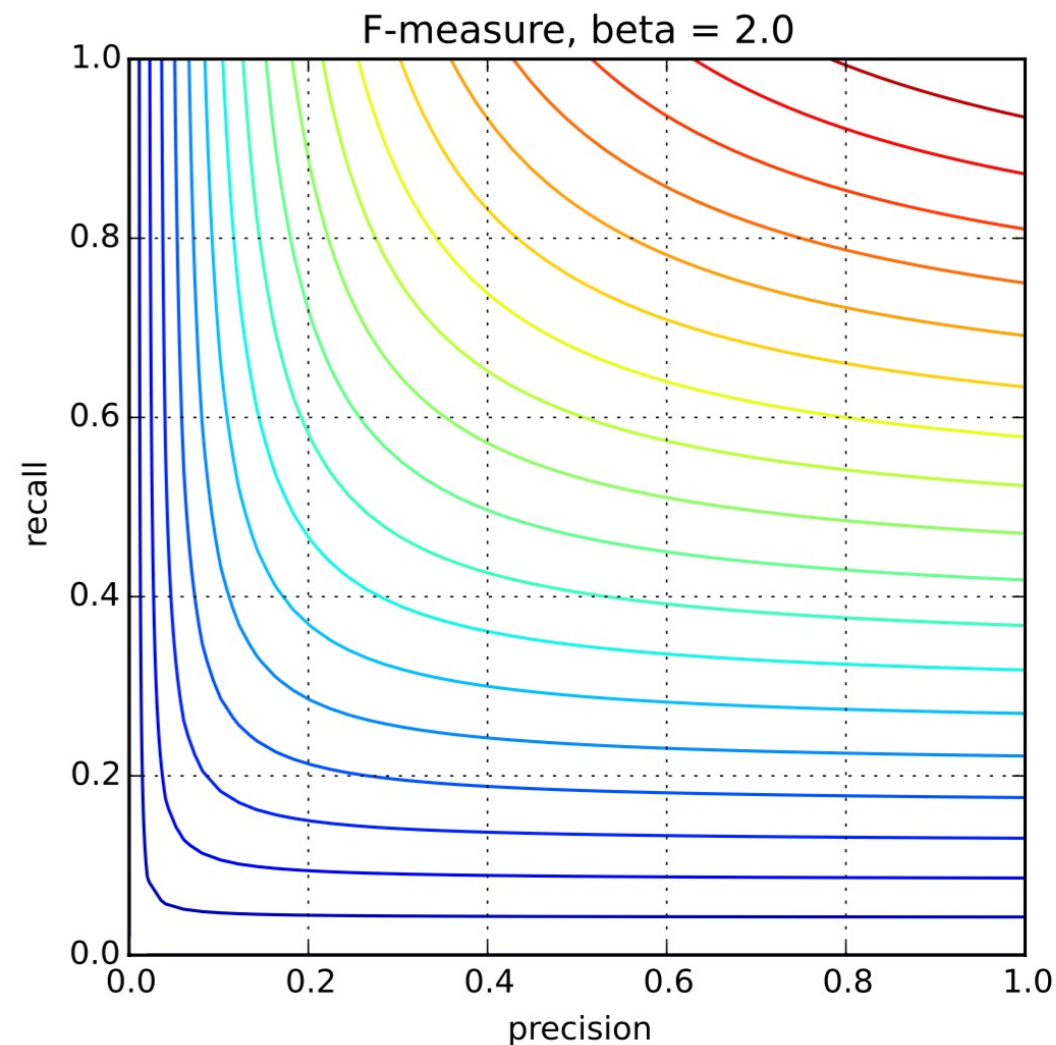
- $\beta = 0.5$
- Важнее точность



F-мера

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

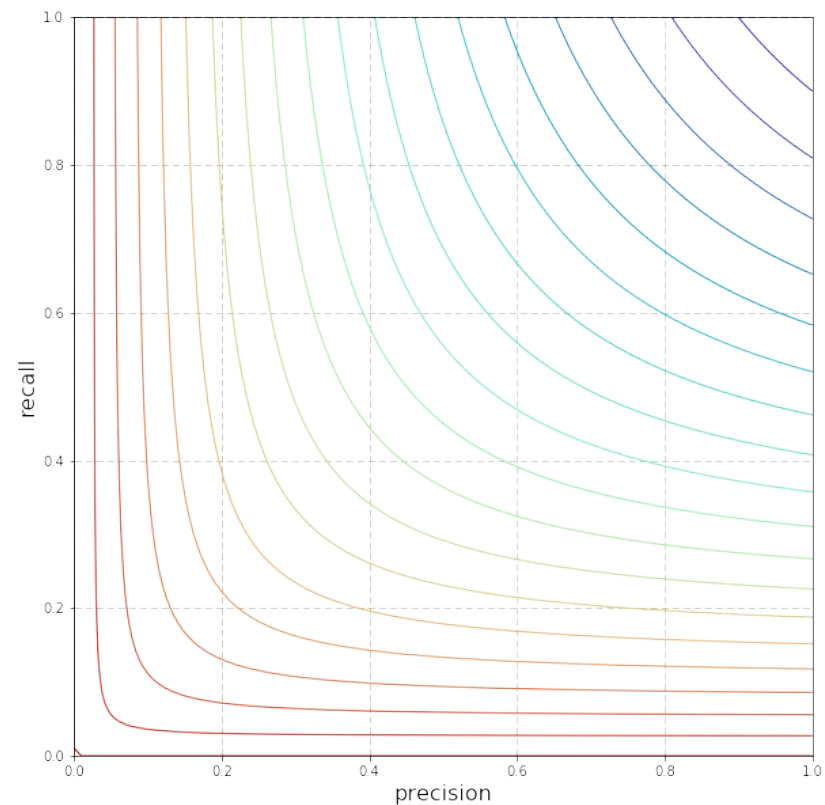
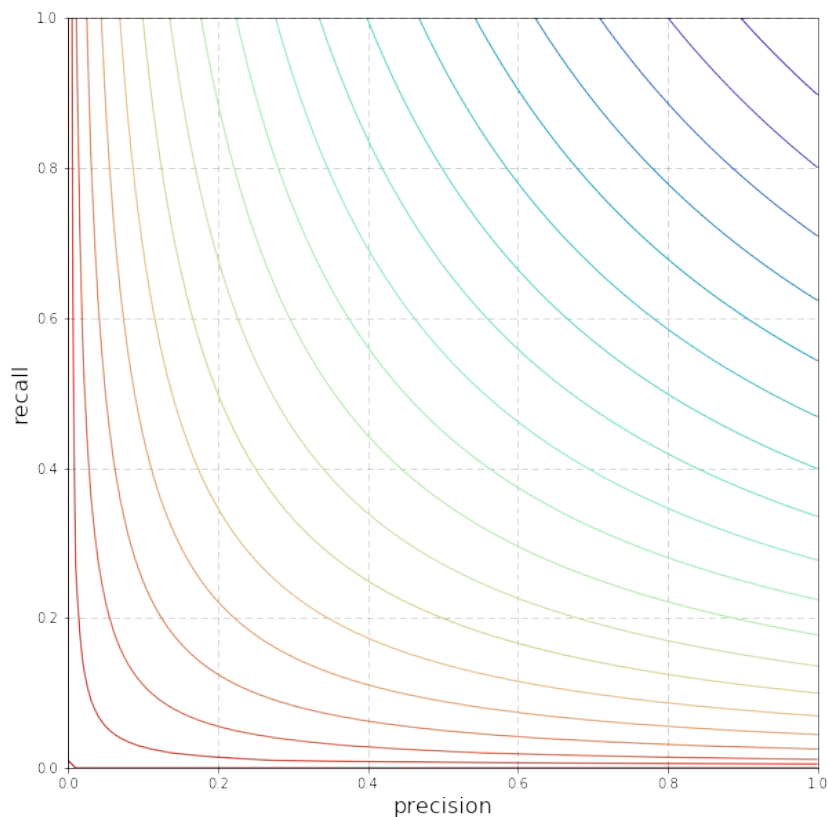
- $\beta = 2$
- Важнее полнота



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
- $F = 0.18$

Метрики качества ранжирования

Классификатор

- Линейный классификатор:

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

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

Оценка принадлежности

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


Оценка принадлежности

-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

-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



-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

-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

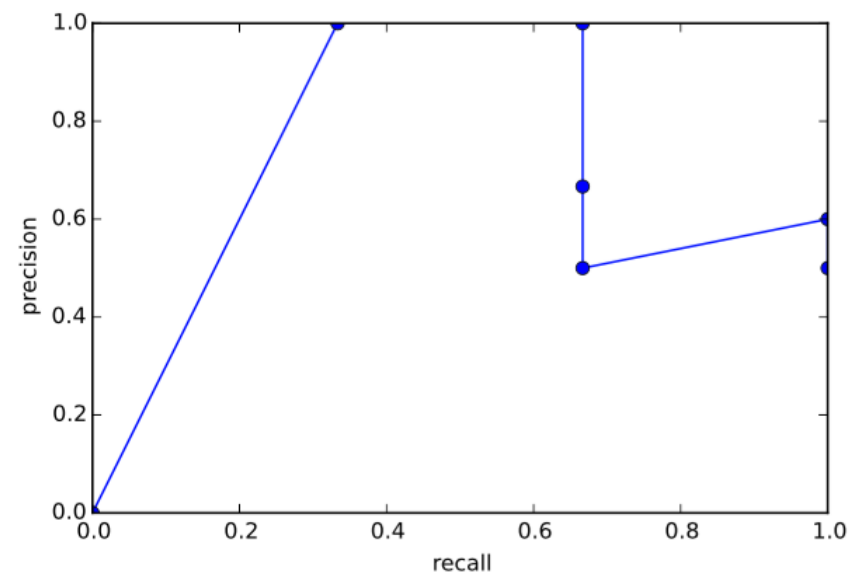
- Как оценить качество $b(x)$?
- Порог выбирается позже
- Порог зависит от ограничений на точность или полноту

Оценка принадлежности

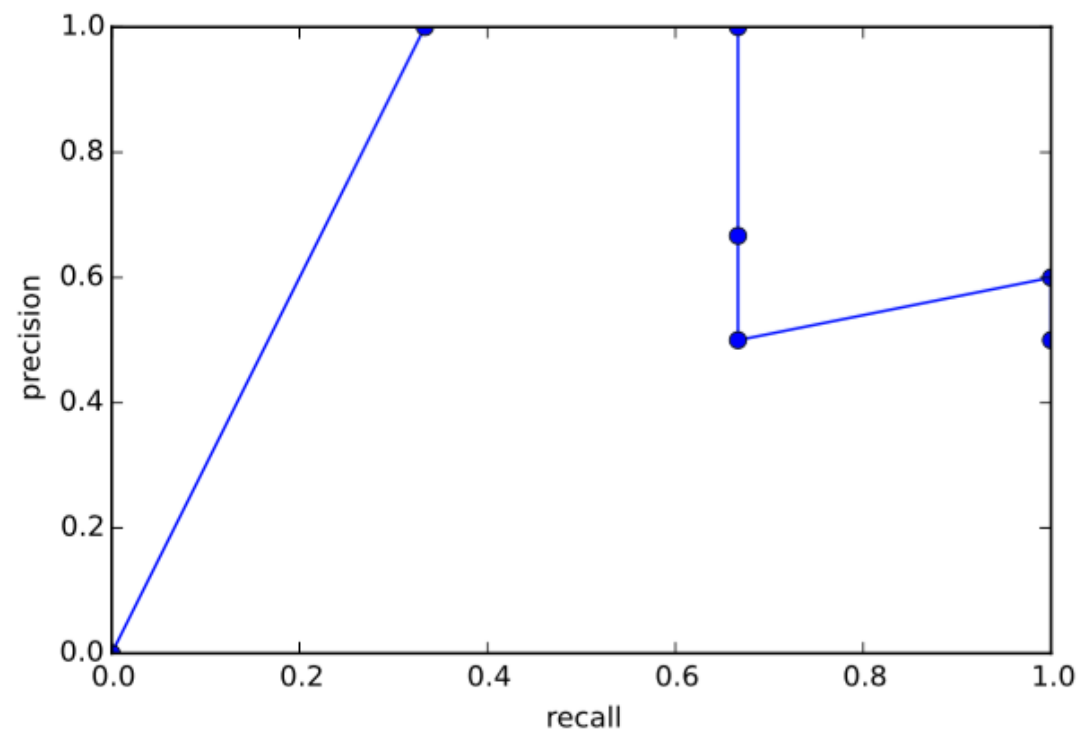
- Пример: кредитный скоринг
- $b(x)$ — оценка вероятности возврата кредита
- $a(x) = [b(x) > 0.5]$
- precision = 0.1, recall = 0.7
- В чем дело — в пороге или в алгоритме?

PR-кривая

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

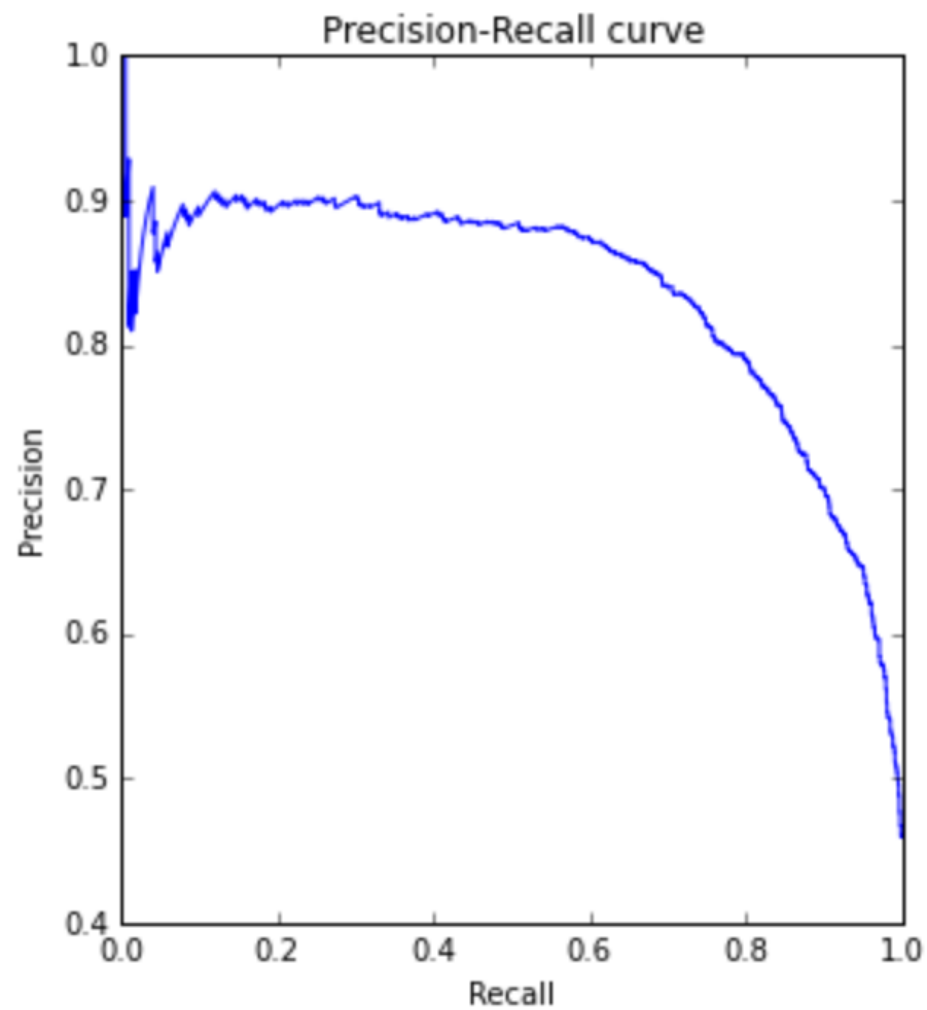


PR-кривая



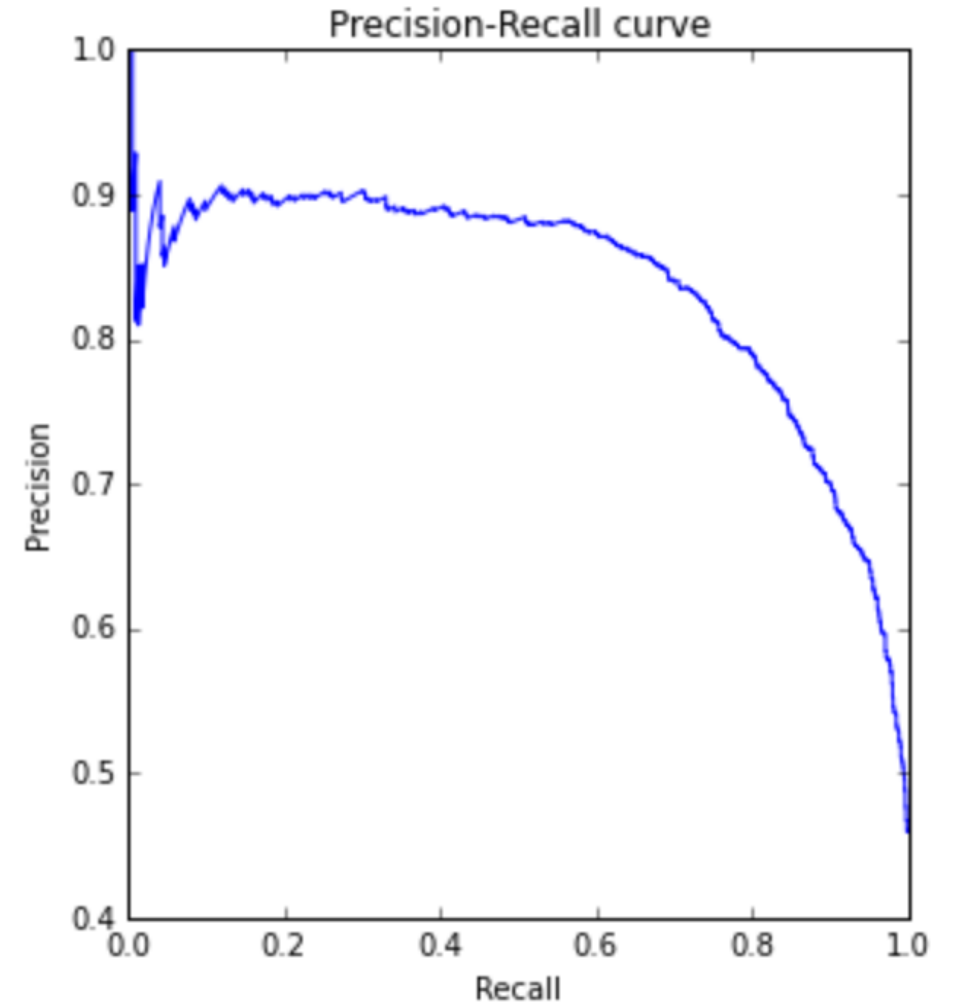
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

PR-кривая в реальности



PR-кривая

- 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

