

Методы машинного обучения

Лекция 7

Метрики качества регрессии и классификации

Эльвира Зиннурова

elvirazinnurova@gmail.com

НИУ ВШЭ, 2019

Подготовка признаков

Важность признаков

- Если признаки масштабированы, то вес характеризует важность признака в модели

$$a(x) = \langle w, x \rangle = w_0 * 1 + w_1 x^1 + \dots + w_d x^d$$

| Term | Coefficient | Std. Error | Z Score |
|-----------|-------------|------------|---------|
| Intercept | 2.46 | 0.09 | 27.60 |
| lcavol | 0.68 | 0.13 | 5.37 |
| lweight | 0.26 | 0.10 | 2.75 |
| age | -0.14 | 0.10 | -1.40 |
| lbph | 0.21 | 0.10 | 2.06 |
| svi | 0.31 | 0.12 | 2.47 |
| lcp | -0.29 | 0.15 | -1.87 |
| gleason | -0.02 | 0.15 | -0.15 |
| pgg45 | 0.27 | 0.15 | 1.74 |

Квадратичные признаки

- Можно добавлять новые признаки, зависящие от исходных
- Модель может восстанавливать более сложные зависимости
- Пример: квадратичные признаки

[площадь, этаж, число комнат]

- Новые признаки:

[площадь, этаж, число комнат,

площадь², этаж², число комнат²,

площадь* этаж, площадь* число комнат, этаж* число комнат,]

Категориальные признаки

- Пример: район квартиры на продажу
- Сделаем столько новых бинарных признаков, сколько районов:

$$\begin{pmatrix} \text{Зюзино} \\ \text{Хамовники} \\ \text{Пресненский} \\ \text{Хамовники} \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ ? & ? & ? \end{pmatrix}$$

Категориальные признаки

- Пример: район квартиры на продажу
- Сделаем столько новых бинарных признаков, сколько районов:

$$\begin{pmatrix} \text{Зюзино} \\ \text{Хамовники} \\ \text{Пресненский} \\ \text{Хамовники} \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

One-hot-кодирование

- Заводим столько новых признаков, сколько значений у категориального
- Каждый соответствует одному возможному значению
- Единице равен только тот признак, чье значение принял исходный категориальный признак на этом объекте
- Иногда категории — это числа, а не строки!

One-hot-кодирование

- Пример: предсказать, купит ли пользователь данный товар в интернет-магазине
- Признаки:
 - Идентификатор пользователя
 - Идентификатор товара
 - Идентификатор категории товара
 - Стоимость товара
 - ...
- Могут иметь смысл квадратичные признаки
 - например, пользователь + категория товара
- После one-hot кодирования получим миллионы признаков
- Линейные модели способны справиться с такими задачами

Метрики качества

- Не все алгоритмы подходят для решения задачи
- Как выбрать лучший?
- Если много способов определить, что такое «лучший»
- Метрики качества
 - Насколько алгоритм подходит для решения задачи?
 - Какой из двух алгоритмов лучше подходит?

Метрики качества регрессии

Среднеквадратичная ошибка

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

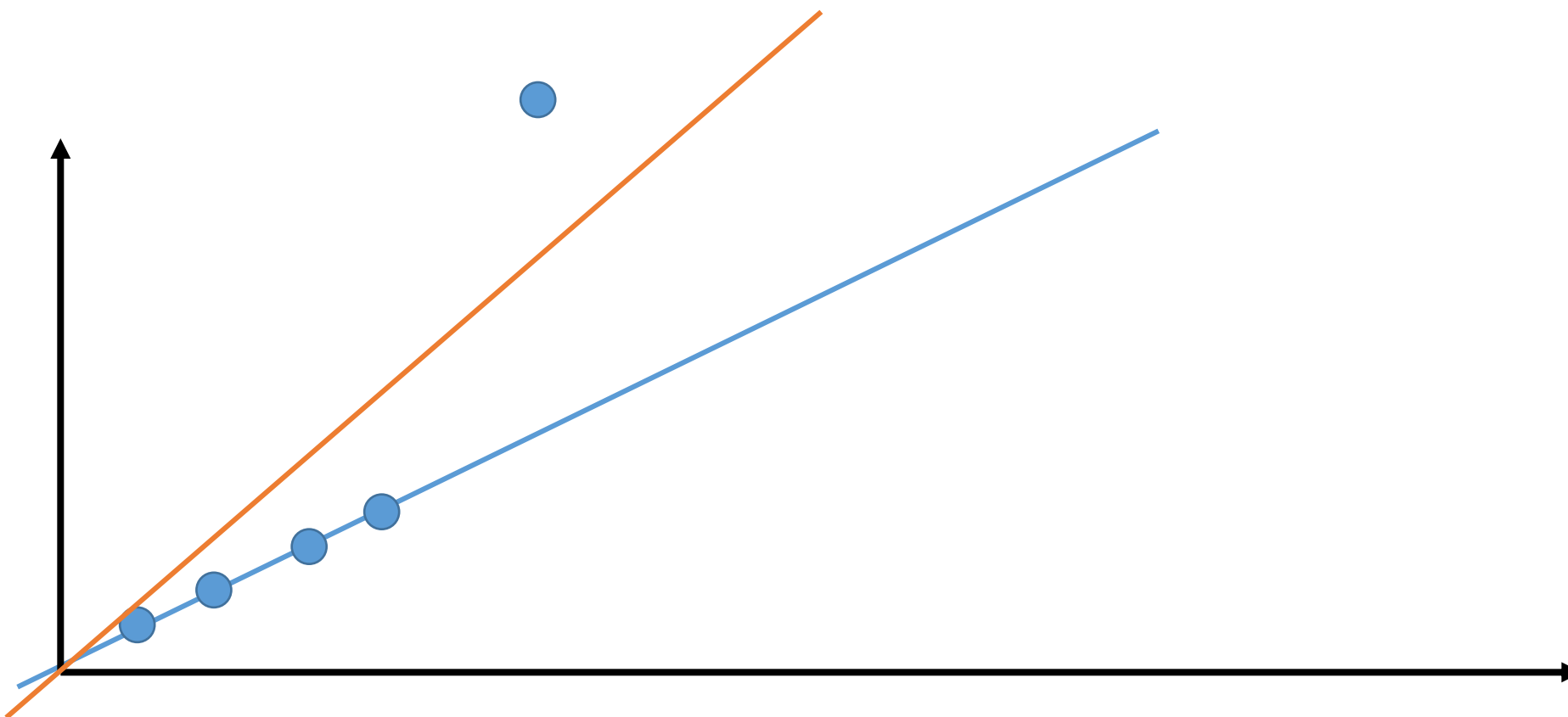
- Легко минимизировать
- Сильно штрафует за большие ошибки

Средняя абсолютная ошибка

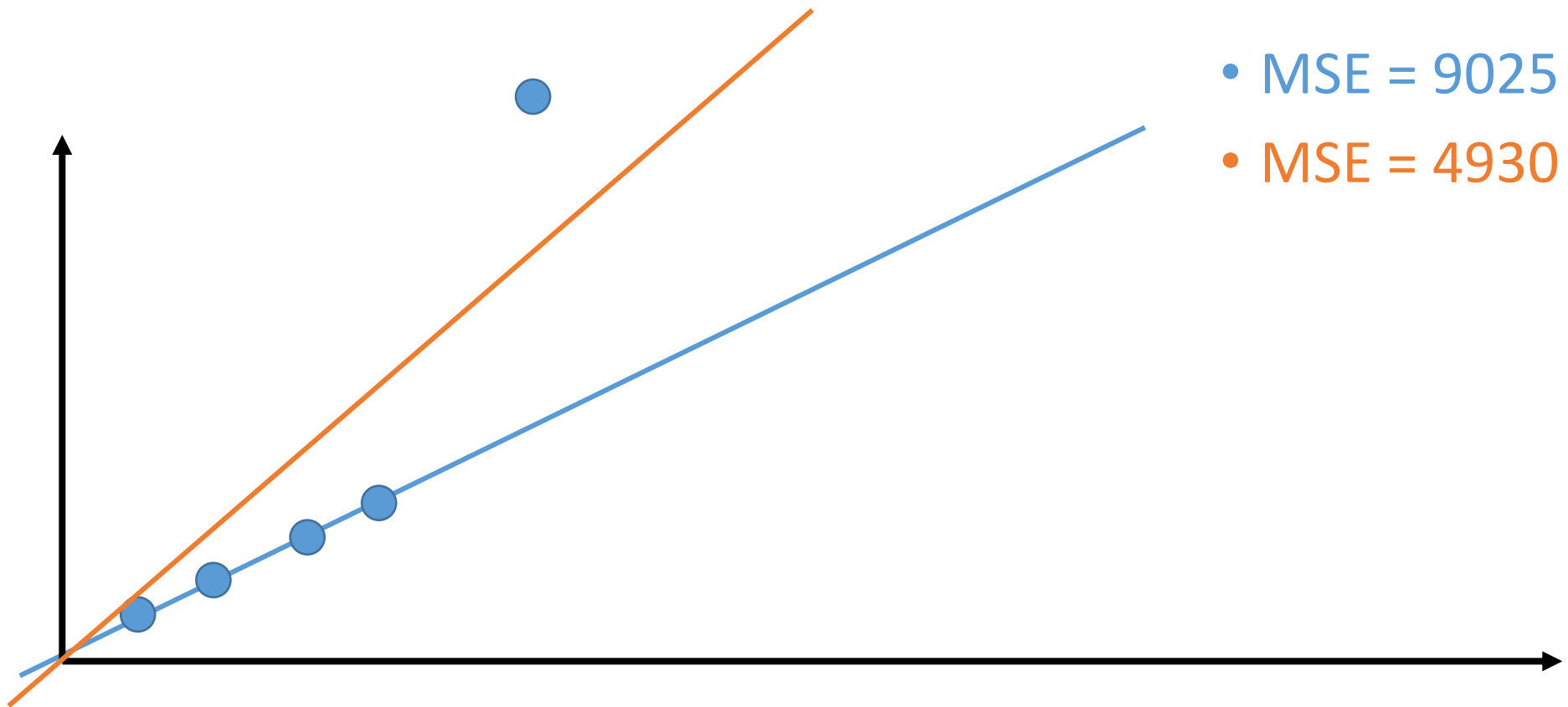
$$\text{MAE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

- Сложнее минимизировать
- Выше устойчивость к выбросам

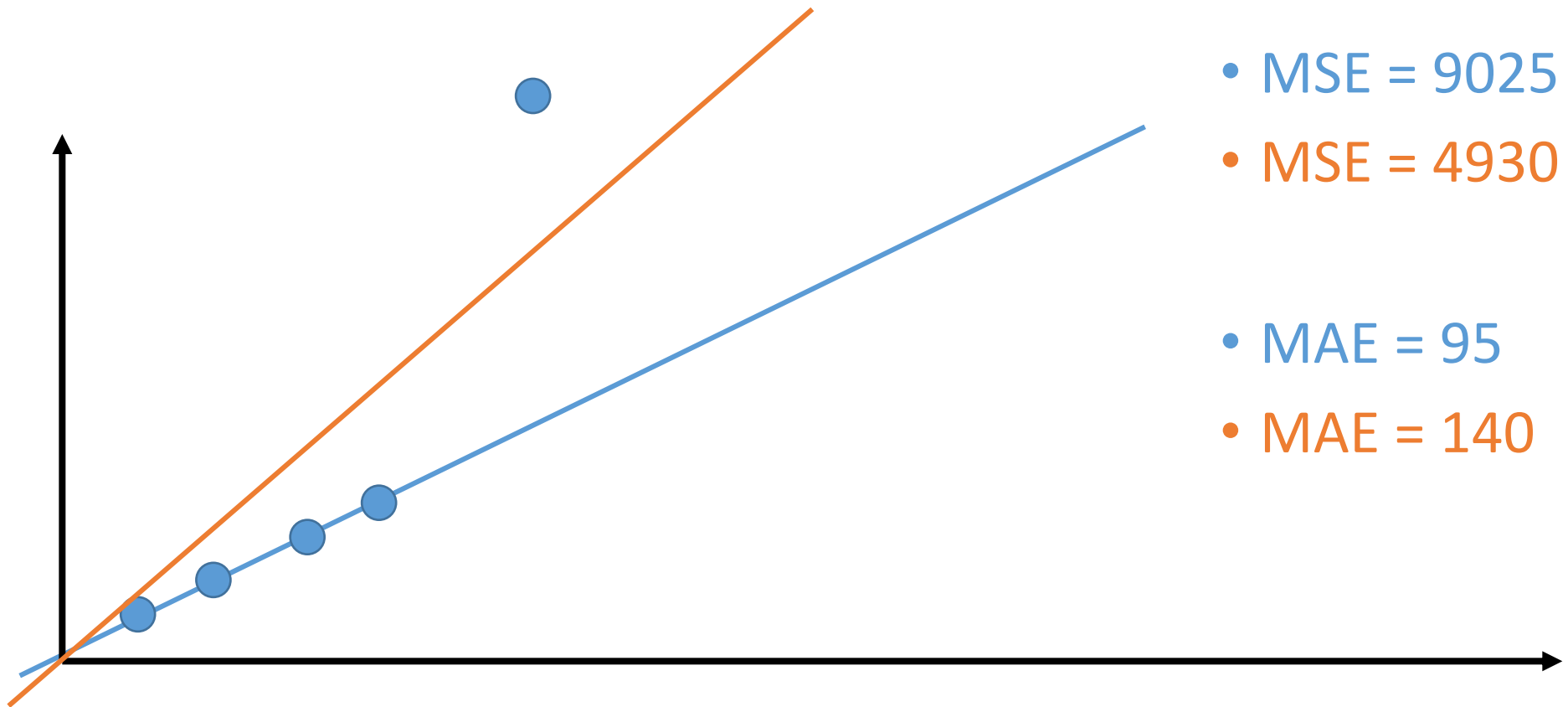
Средняя абсолютная ошибка



Средняя абсолютная ошибка



Средняя абсолютная ошибка



Среднеквадратичная ошибка

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

- Подходит, чтобы сравнивать разные модели
- Чем меньше, тем лучше
- Не позволяет понять, хорошая ли модель получилась
- $\text{MSE} = 32955$ — хорошо или плохо?

Среднеквадратичная ошибка

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

То же самое, но можно интерпретировать

- Предсказываем стоимость квартир (в тыс. руб.)
- $\text{MSE} = 32955$ — хорошо или плохо?
- $\text{RMSE} = \sqrt{32955} \approx 181$ тыс.руб. — средняя ошибка для одной квартиры

Коэффициент детерминации

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

- $\bar{y} = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i$ — средний ответ
- Доля дисперсии, объясненная моделью, в общей дисперсии ответов
- Значение можно интерпретировать

Коэффициент детерминации

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

- $0 \leq R^2 \leq 1$ (для разумных моделей)
- $R^2 = 1$ — идеальная модель
- $R^2 = 0$ — модель на уровне константной
- $R^2 < 0$ — модель хуже константной

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):

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

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

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

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

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

Несбалансированные выборки

- Пример:
 - Класс -1: 950 объектов
 - Класс +1: 50 объектов
- $a(x) = -1$
- Доля правильных ответов: 0.95

Несбалансированные выборки

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

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

- Если получили большой accuracy — посмотрите на баланс классов

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

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

Как запомнить:

True



Верно ли
классифицирован объект?

Positive



К какому классу модель
отнесла объект?

— объект верно отнесен
к классу +1

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

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

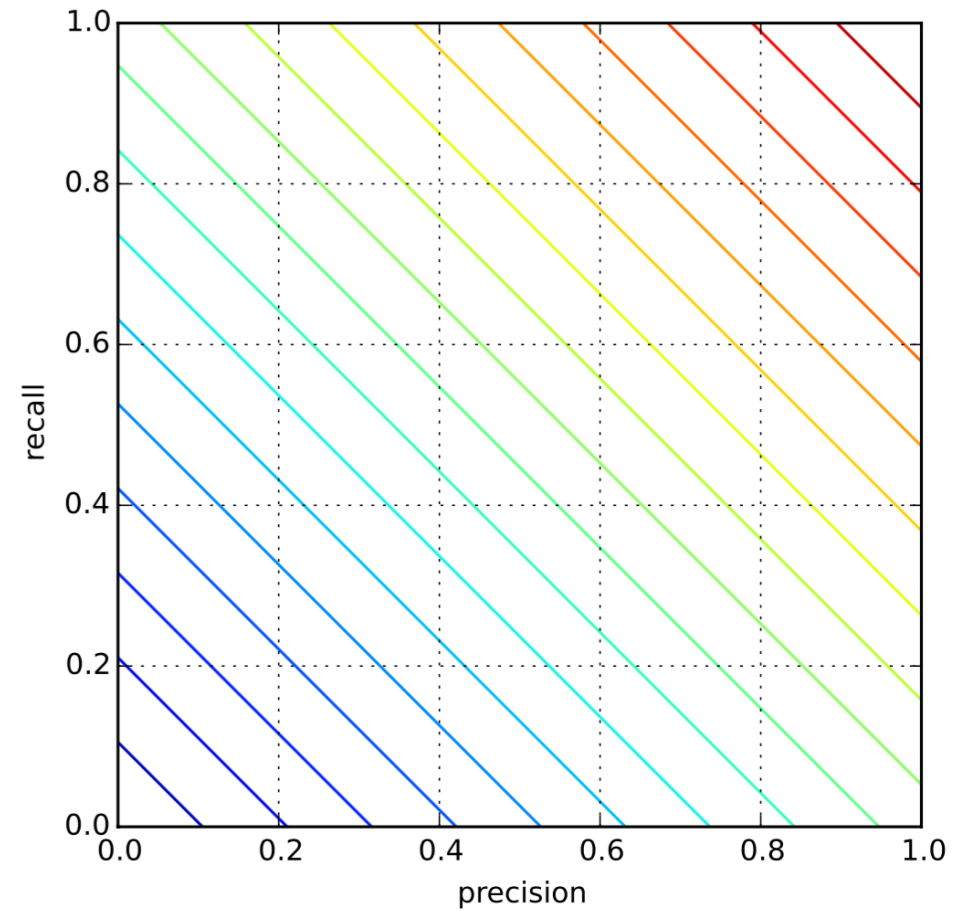
| | $y = 1$ | $y = -1$ |
|-------------|---------|----------|
| $a(x) = 1$ | 10 | 20 |
| $a(x) = -1$ | 90 | 10000 |

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

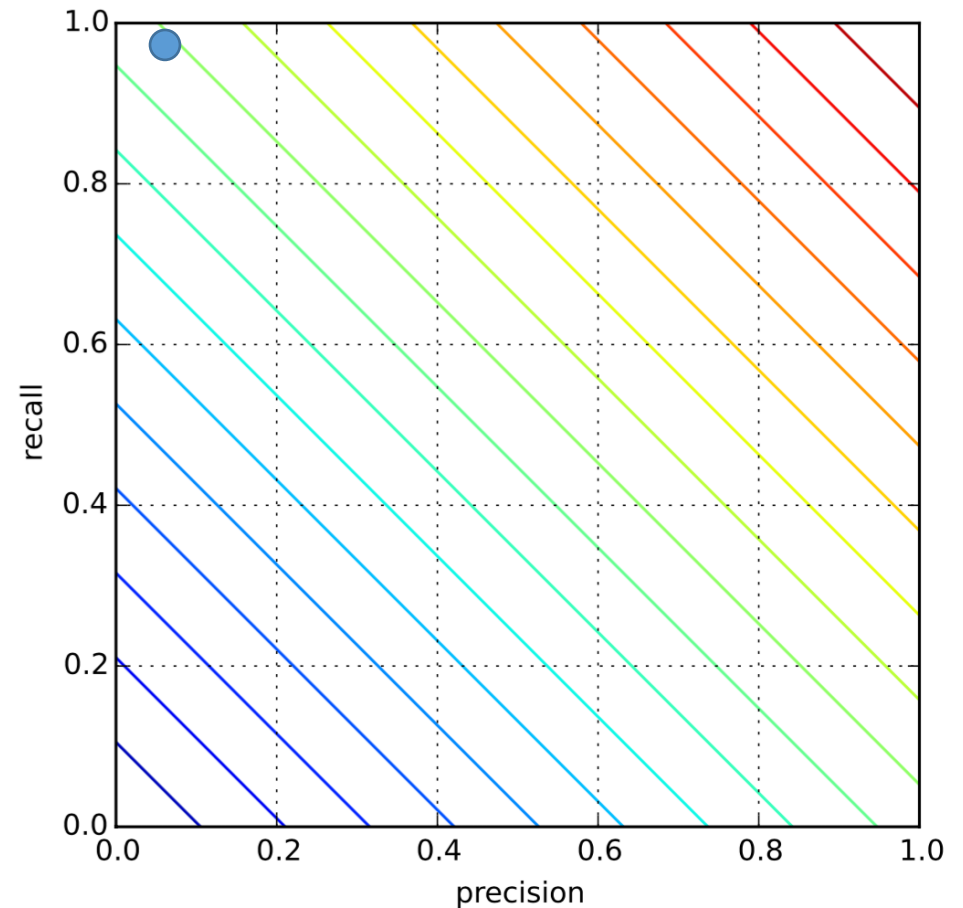
$$A = \frac{1}{2}(\text{precision} + \text{recall})$$



Арифметическое среднее

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

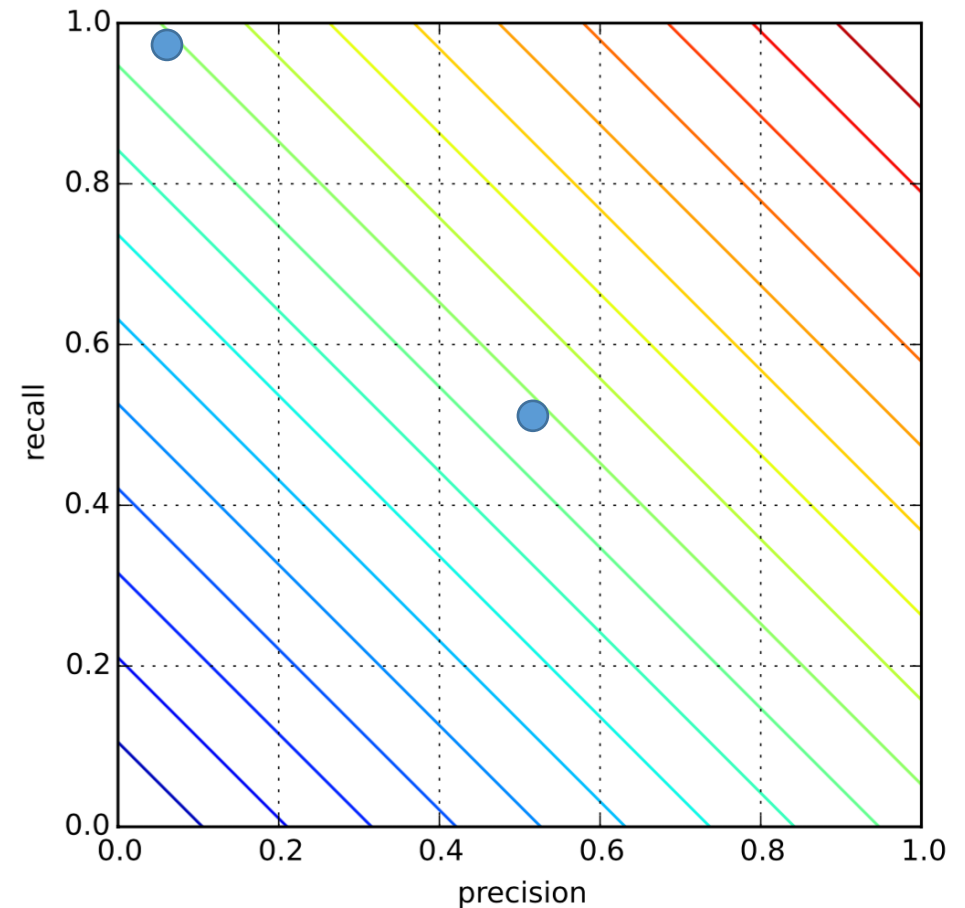
- precision = 0.1
- recall = 1
- $A = 0.55$
- Плохой алгоритм



Арифметическое среднее

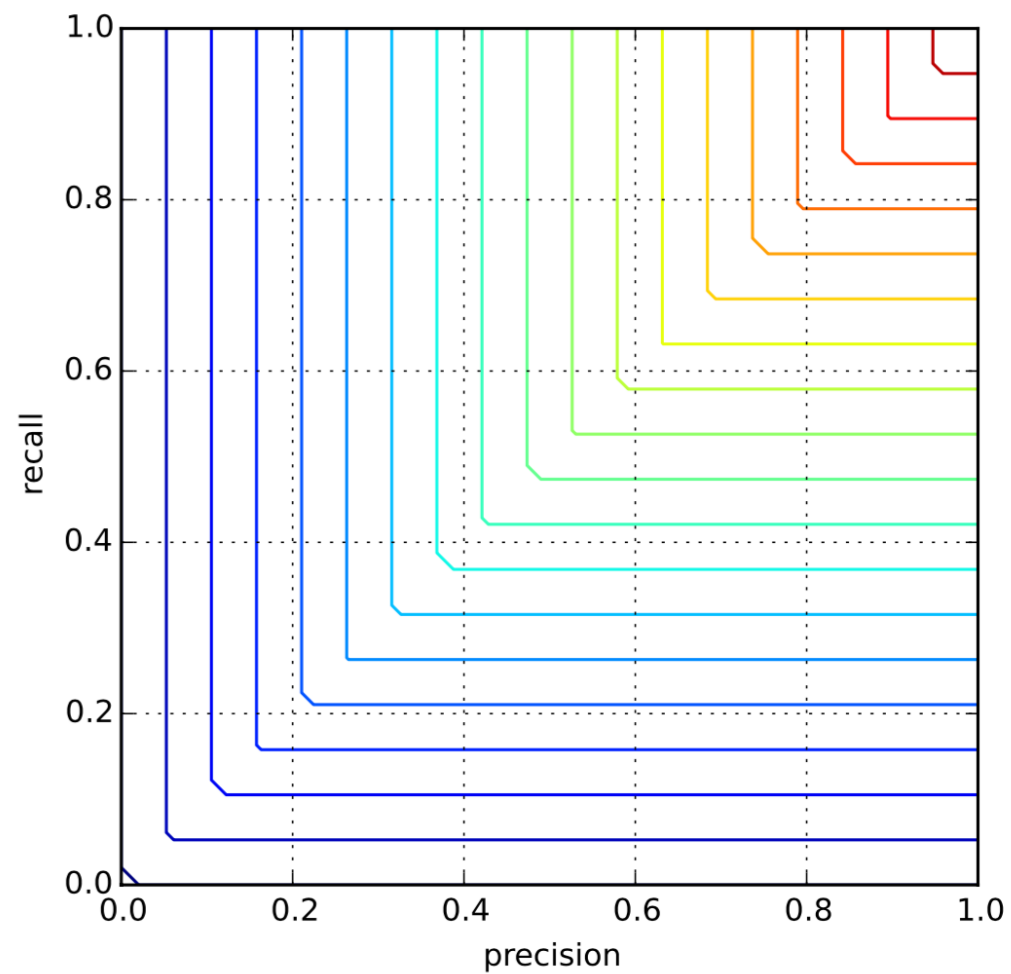
$$A = \frac{1}{2} (\text{precision} + \text{recall})$$

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



Минимум

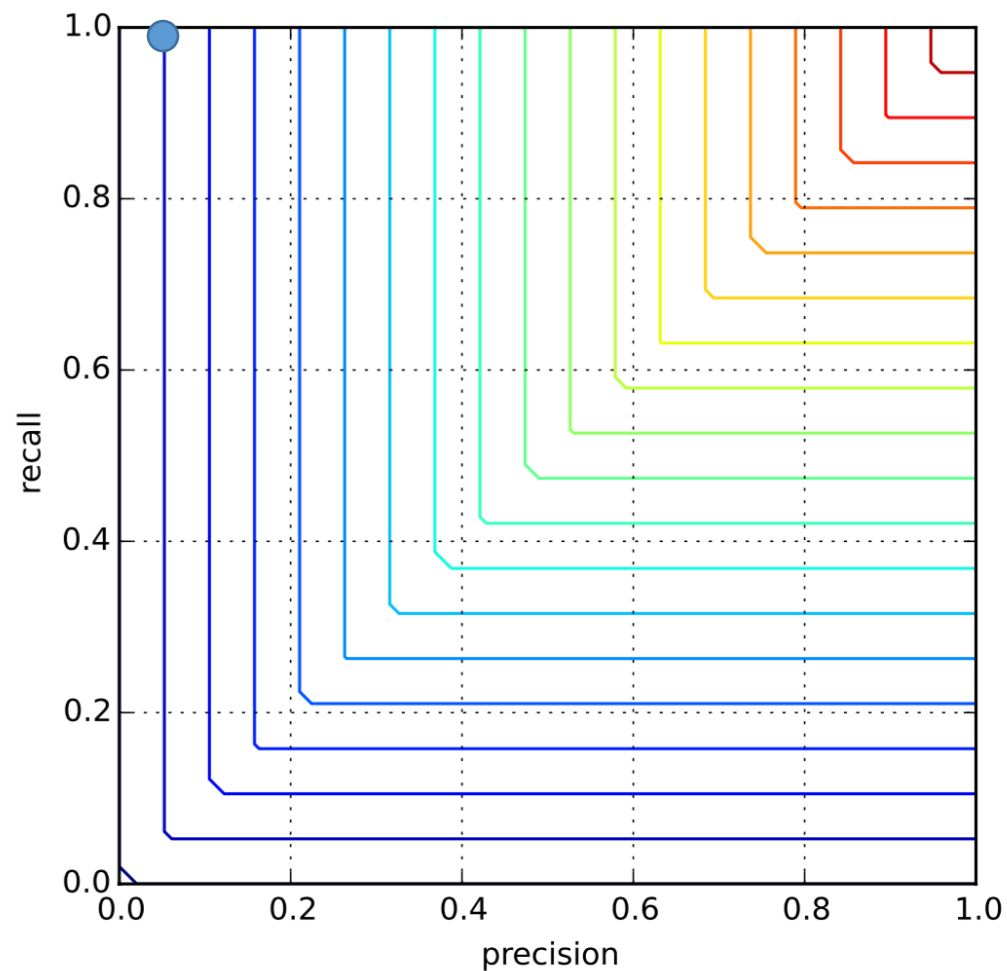
$$M = \min(\text{precision}, \text{recall})$$



Минимум

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

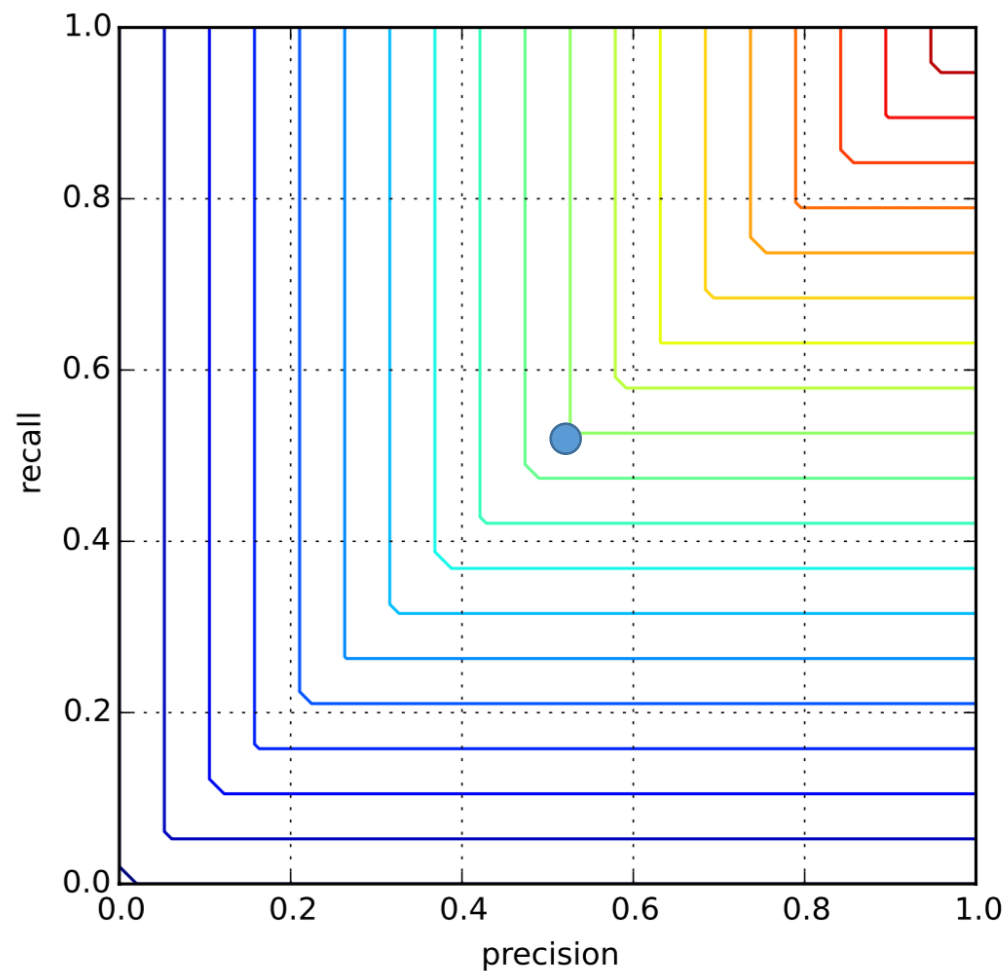
- precision = 0.05
- recall = 1
- $M = 0.05$



Минимум

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

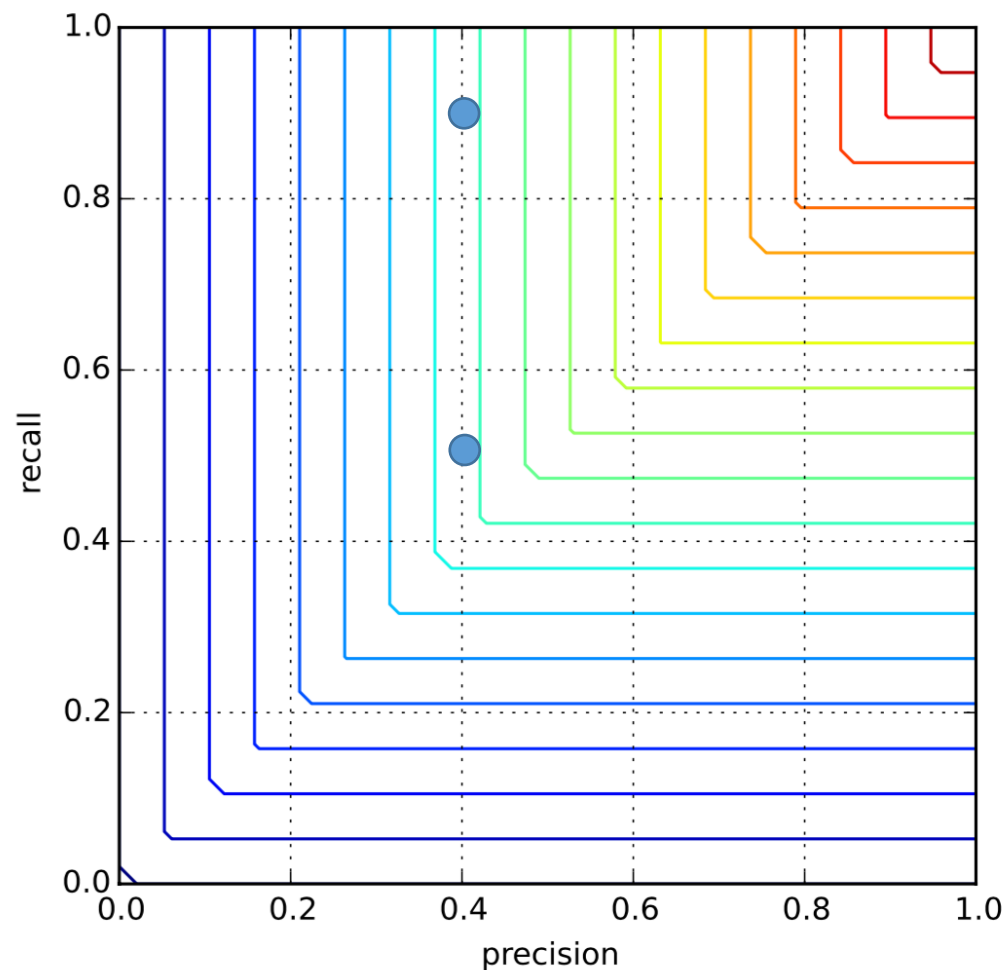
- precision = 0.55
- recall = 0.55
- $M = 0.55$



Минимум

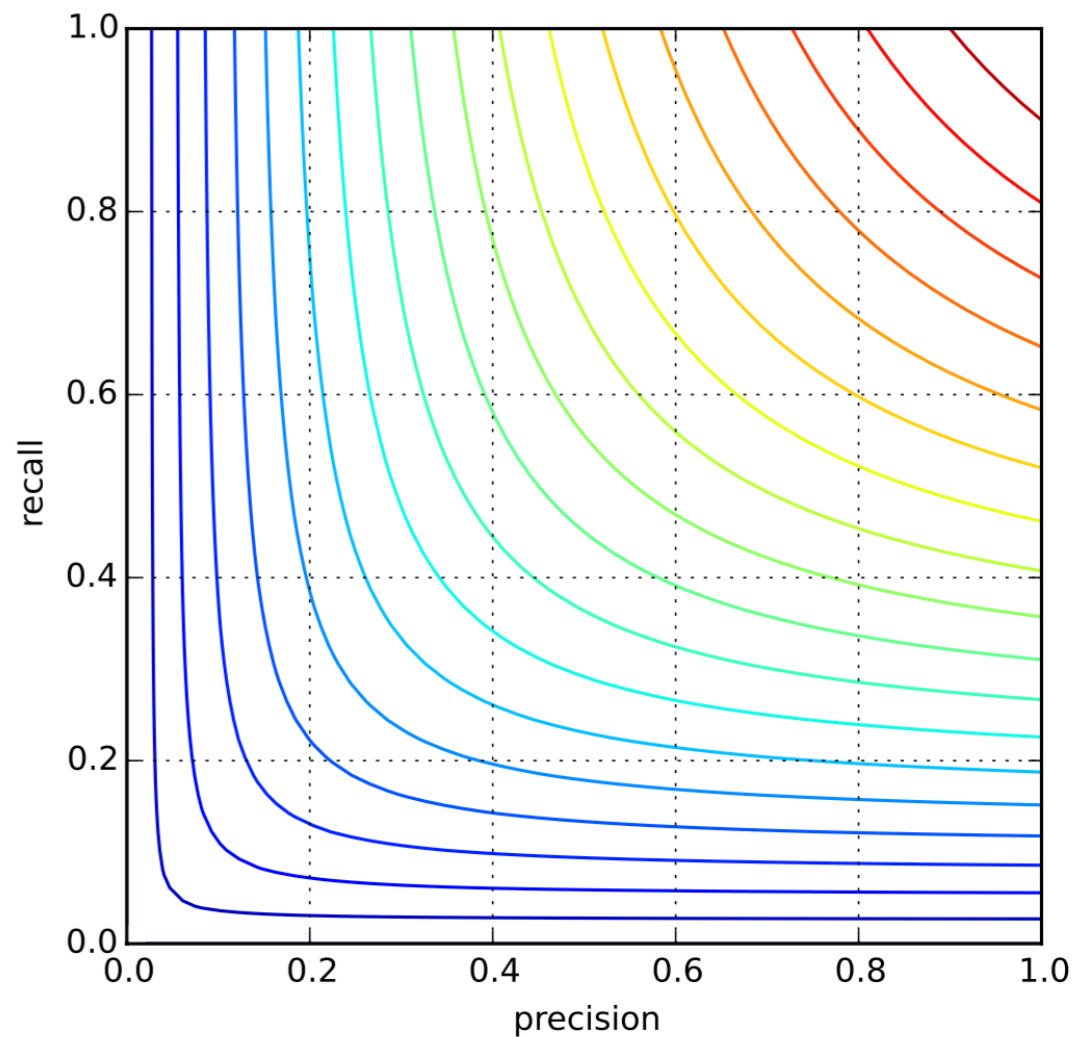
$$M = \min(\text{precision}, \text{recall})$$

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



F-measure

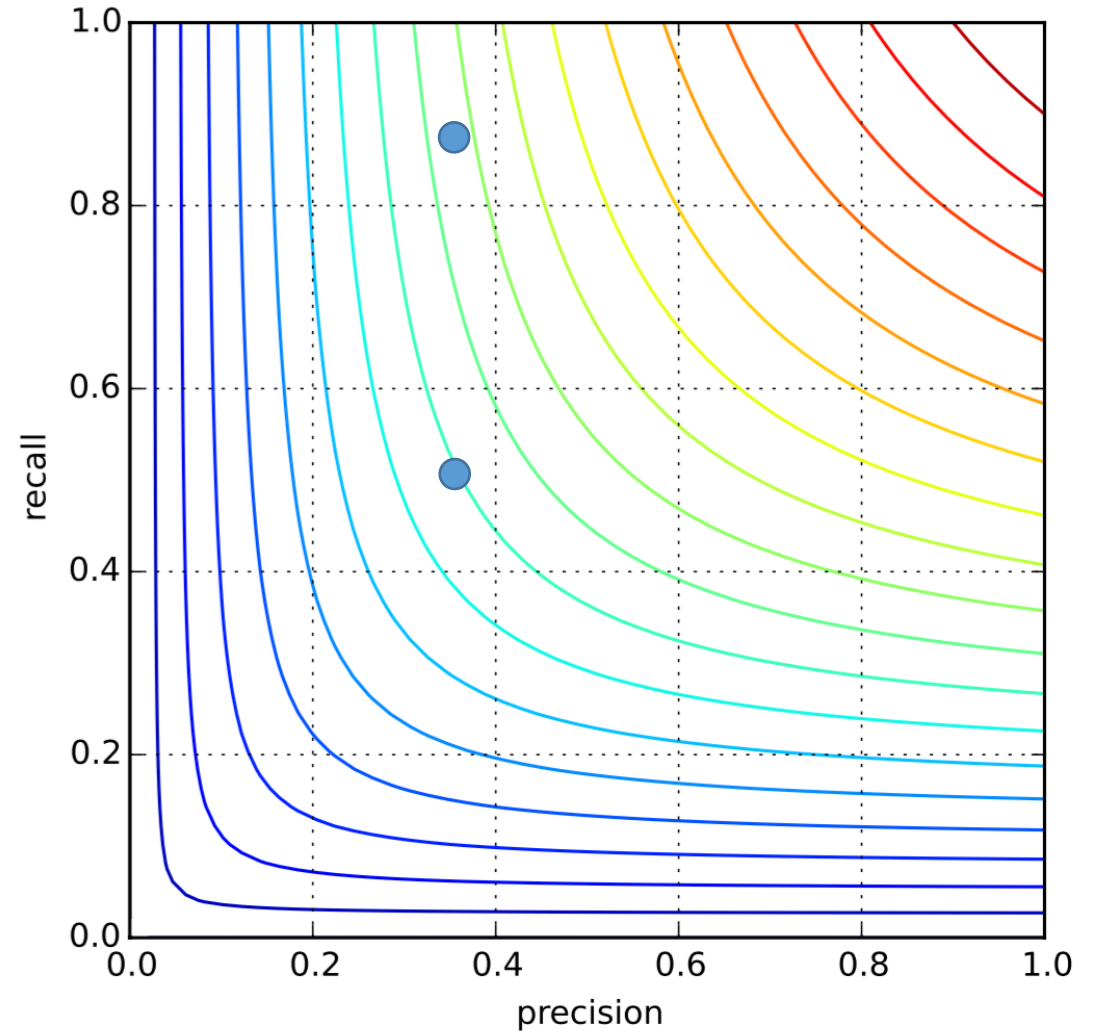
$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



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
- $F = 0.55$



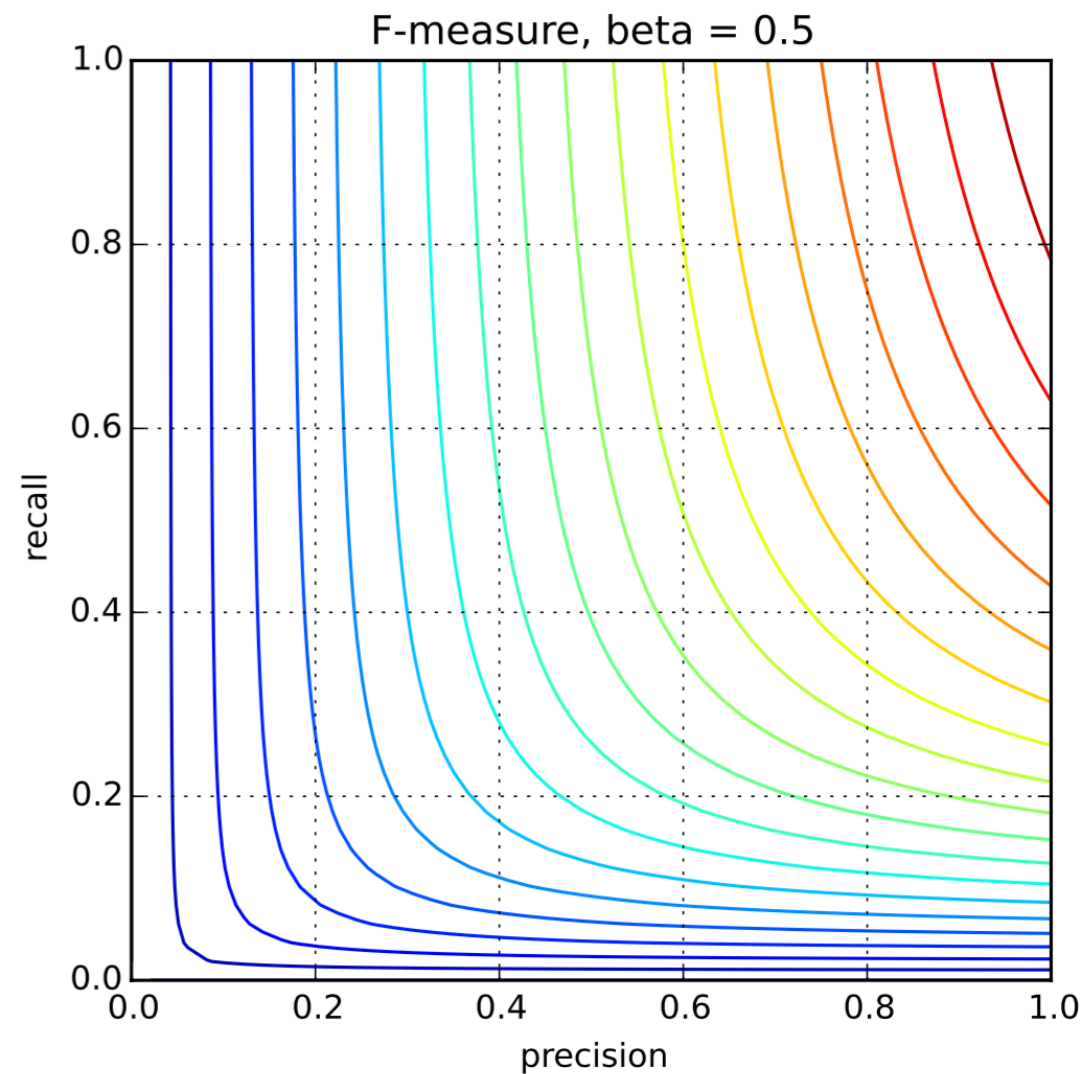
F-measure

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

F-мера

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

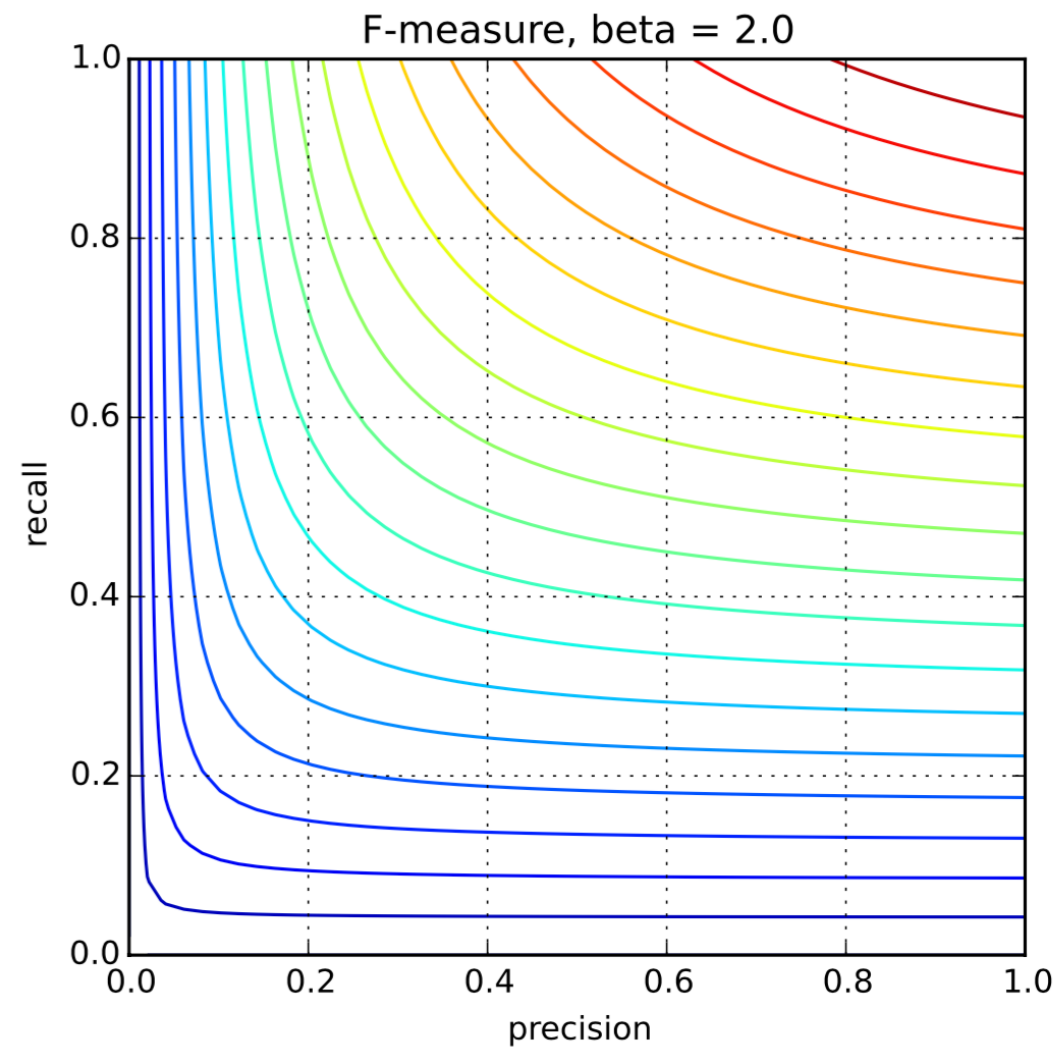
- $\beta = 0.5$
- Важнее полнота



F-мера

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

- $\beta = 2$
- Важнее точность



Оценки принадлежности классу

Классификатор

- Частая ситуация:

$$a(x) = [b(x) > t]$$

- $b(x)$ — оценка принадлежности классу +1

Линейный классификатор

$$a(x) = [\langle w, x \rangle > t]$$

- $b(x) = \langle w, x \rangle$ — оценка принадлежности классу +1
- Обычно $t = 0$:

$$a(x) = [\langle w, x \rangle > 0] = \text{sign} \langle w, x \rangle$$

Оценка принадлежности

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

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



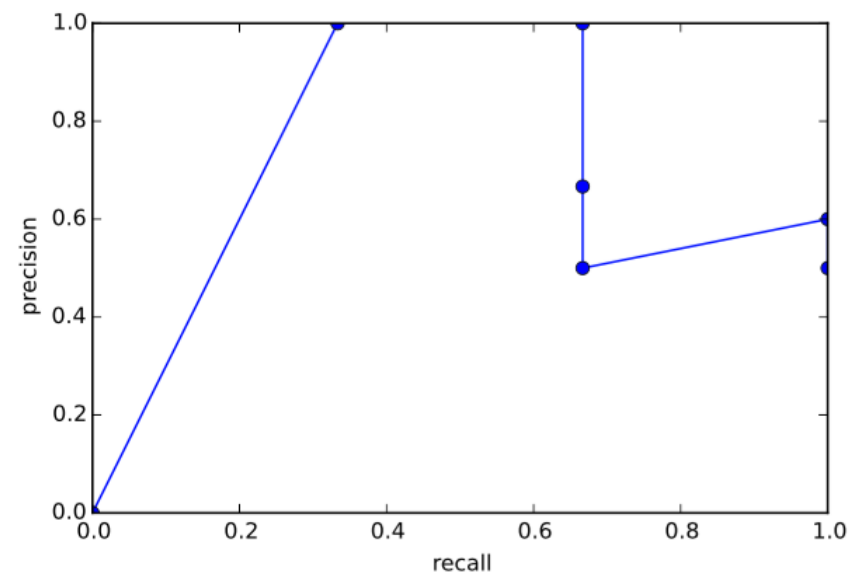
| | | | | | | | | | |
|------|------|------|------|------|-----|------|-----|------|-----|
| -1 | -1 | +1 | +1 | +1 | -1 | +1 | +1 | -1 | +1 |
| 0.01 | 0.09 | 0.12 | 0.15 | 0.29 | 0.4 | 0.48 | 0.6 | 0.83 | 0.9 |

Оценка принадлежности

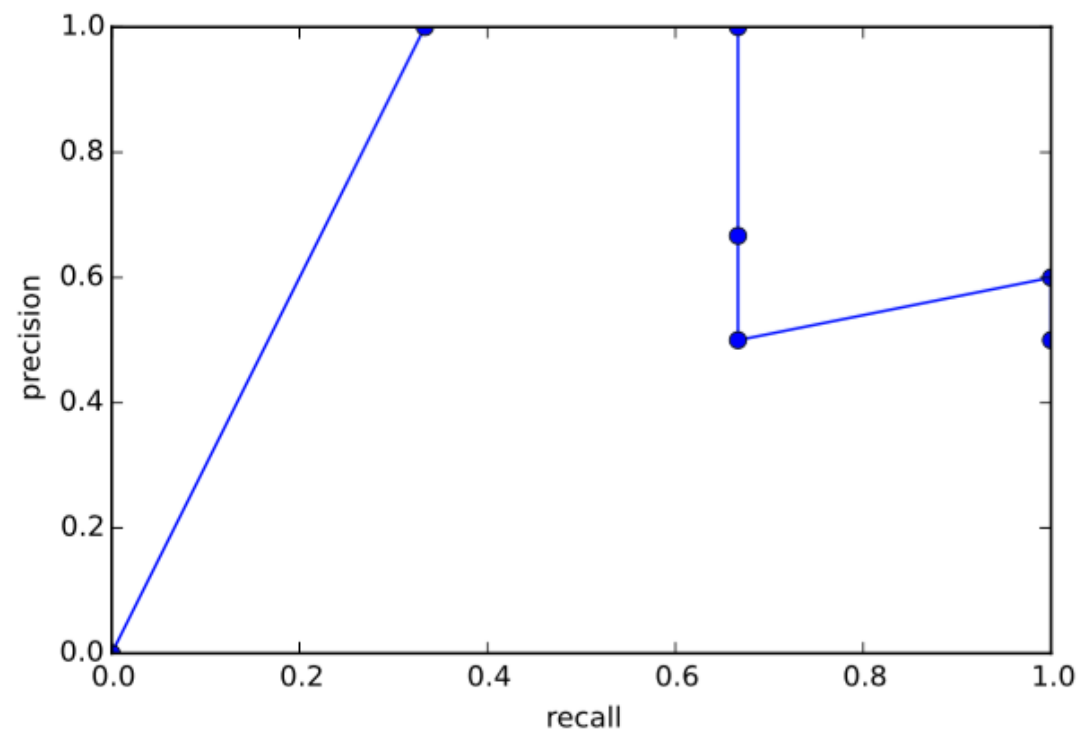
- Пример: кредитный скоринг
- $b(x)$ — оценка вероятности возврата кредита
- $a(x) = [b(x) > 0.5]$
- precision = 0.1, recall = 0.7
- В чем дело — в пороге или в алгоритме?

PR-кривая

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

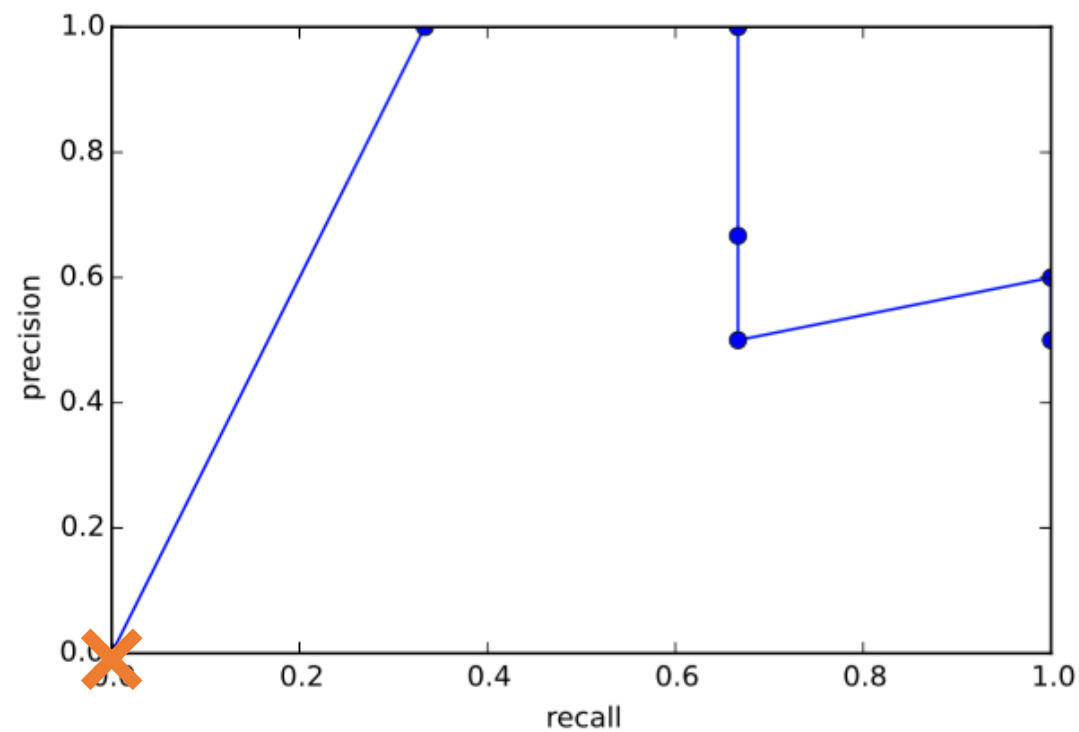


PR-кривая



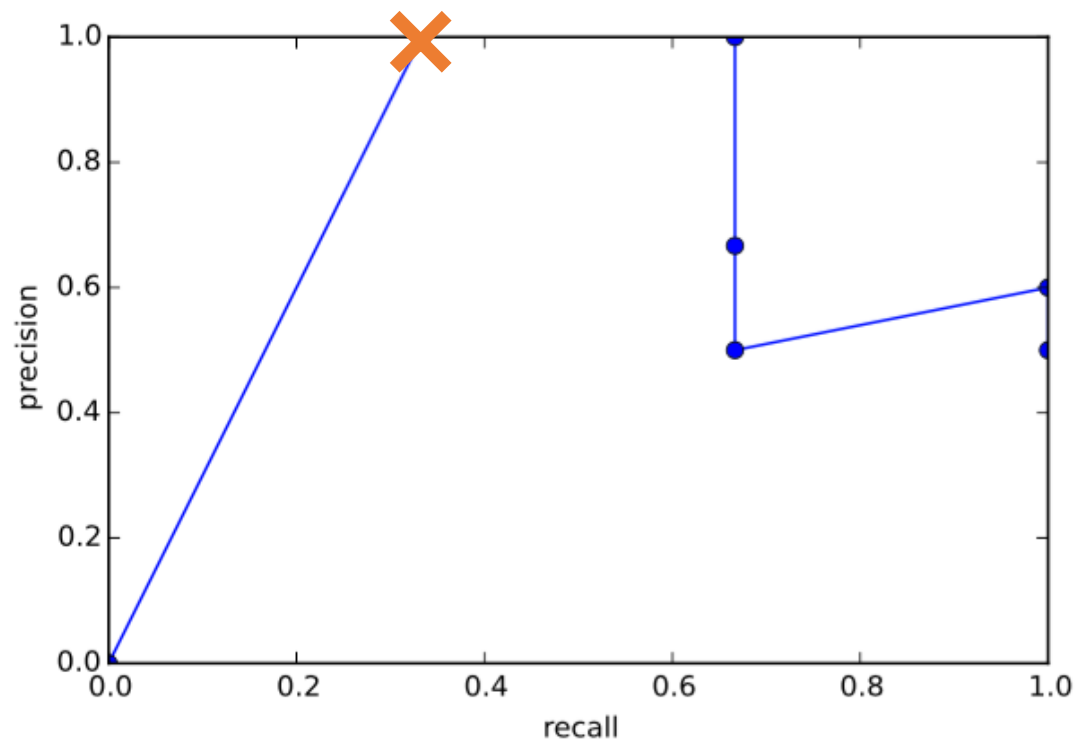
| | | | | | | |
|--------|------|------|------|------|------|------|
| $b(x)$ | 0.14 | 0.23 | 0.39 | 0.52 | 0.73 | 0.90 |
| y | 0 | 1 | 0 | 0 | 1 | 1 |

PR-кривая



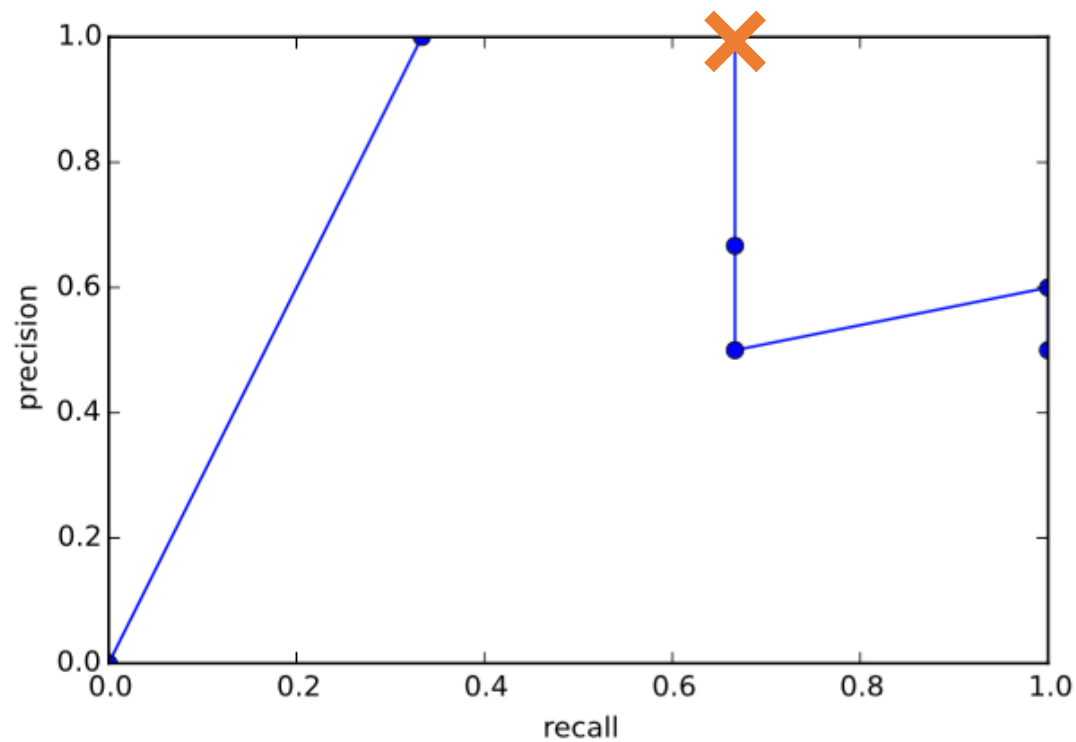
| | | | | | | |
|--------|------|------|------|------|------|------|
| $b(x)$ | 0.14 | 0.23 | 0.39 | 0.52 | 0.73 | 0.90 |
| y | 0 | 1 | 0 | 0 | 1 | 1 |

PR-кривая



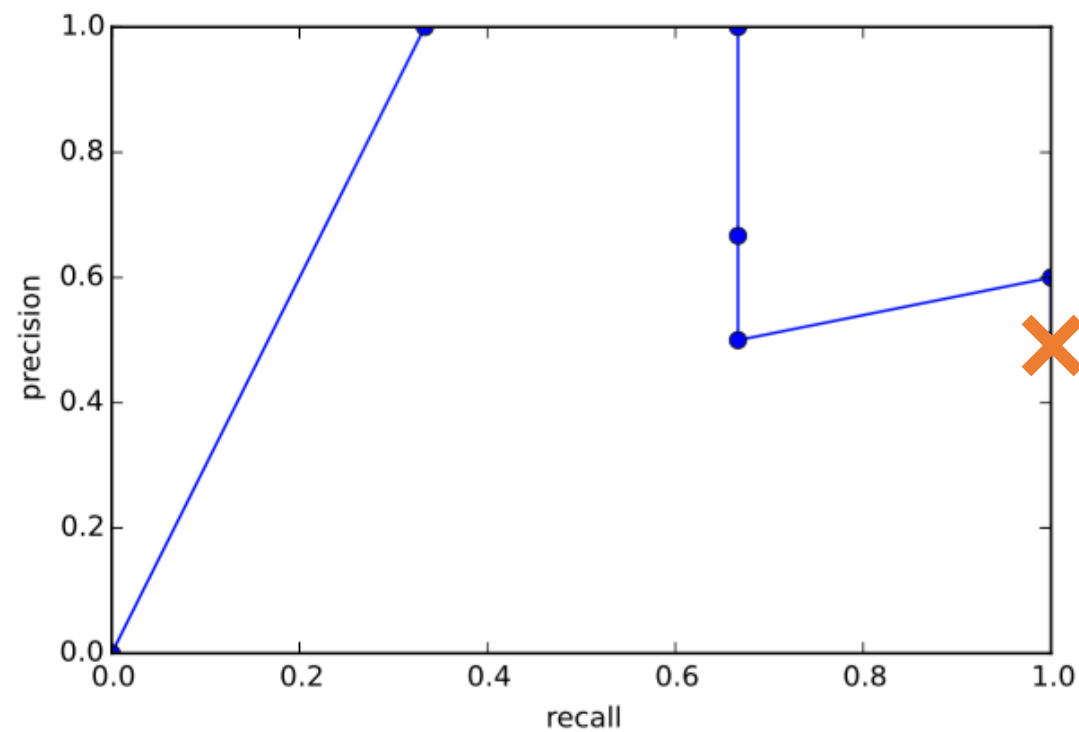
| | | | | | | |
|--------|------|------|------|------|------|------|
| $b(x)$ | 0.14 | 0.23 | 0.39 | 0.52 | 0.73 | 0.90 |
| y | 0 | 1 | 0 | 0 | 1 | 1 |

PR-кривая



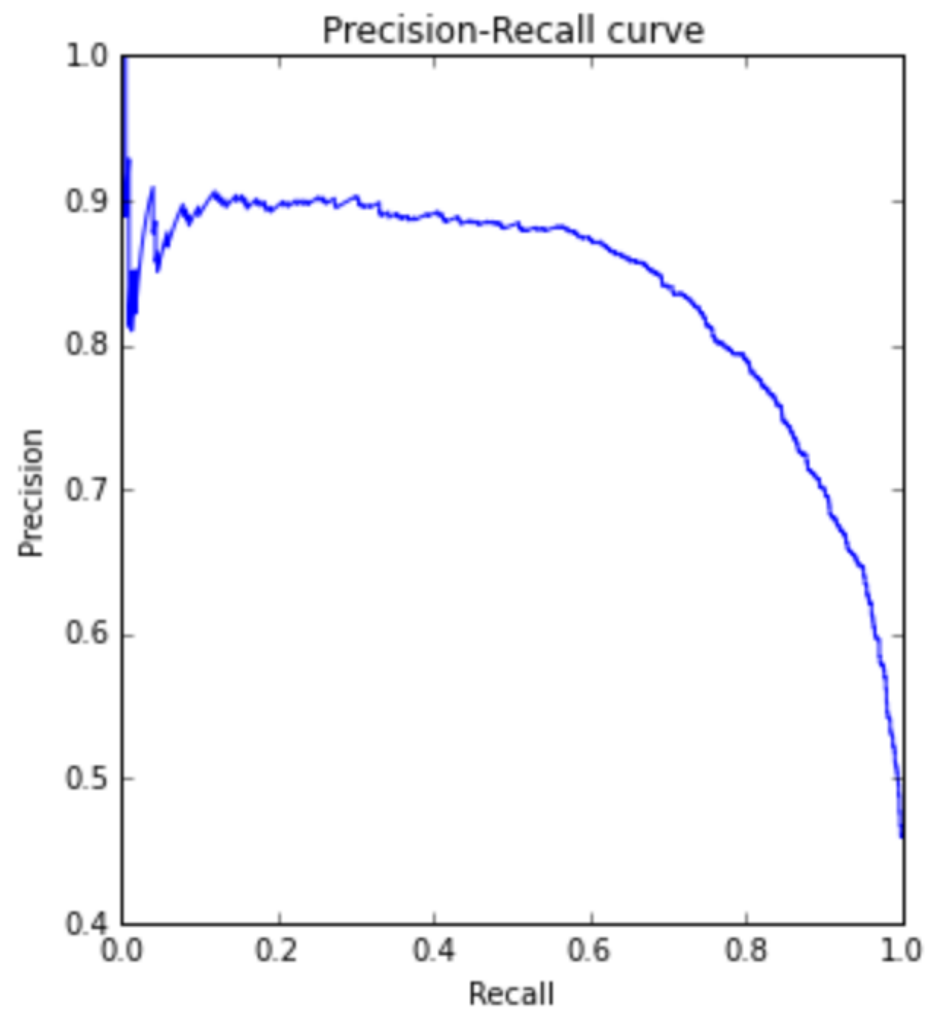
| | | | | | | |
|--------|------|------|------|------|------|------|
| $b(x)$ | 0.14 | 0.23 | 0.39 | 0.52 | 0.73 | 0.90 |
| y | 0 | 1 | 0 | 0 | 1 | 1 |

PR-кривая



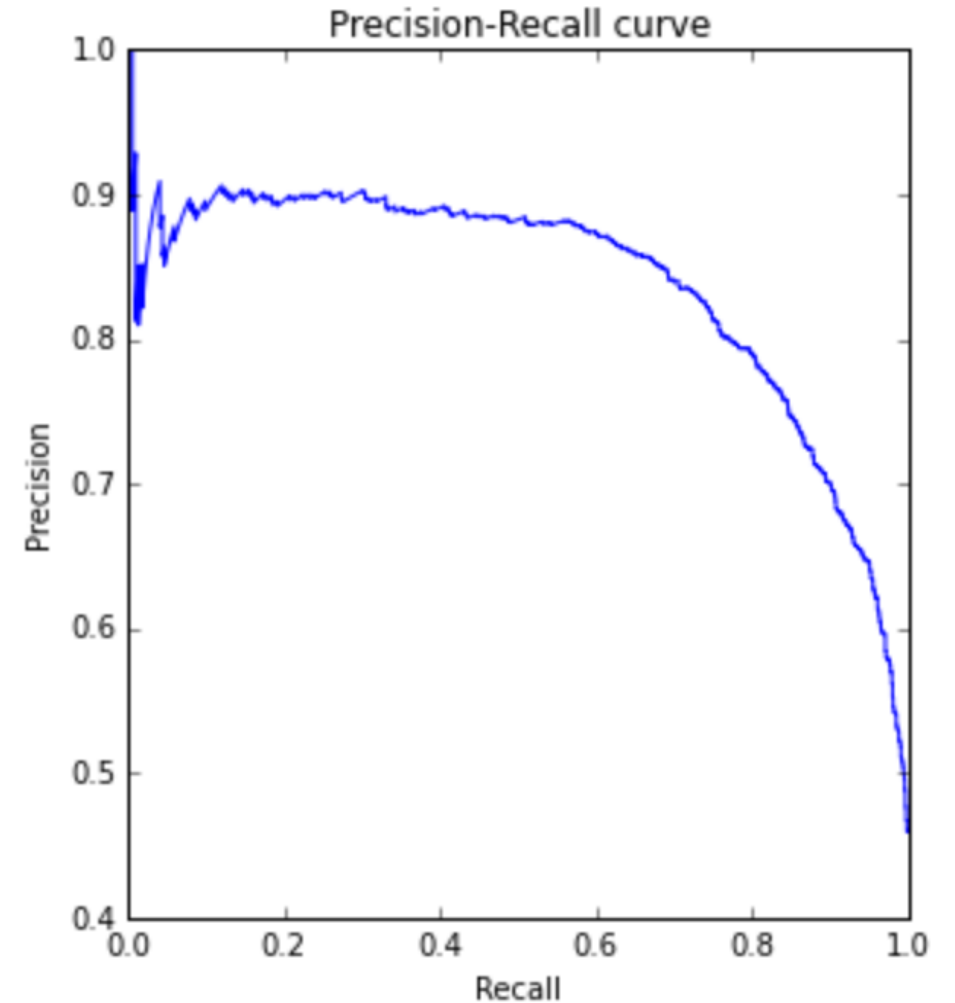
| | | | | | | |
|--------|------|------|------|------|------|------|
| $b(x)$ | 0.14 | 0.23 | 0.39 | 0.52 | 0.73 | 0.90 |
| y | 0 | 1 | 0 | 0 | 1 | 1 |

PR-кривая в реальности



PR-кривая

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



ROC-кривая

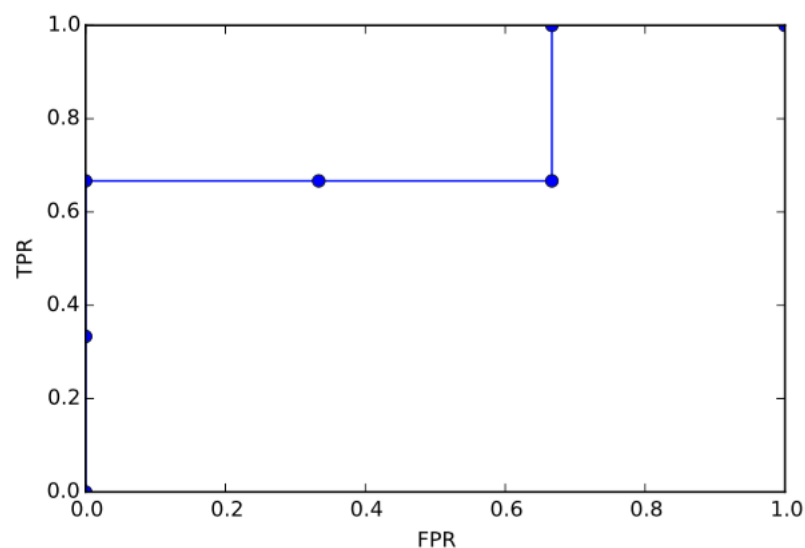
- Receiver Operating Characteristic

- Ось X — False Positive Rate

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

- Ось Y — True Positive Rate

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



ROC-кривая

- Receiver Operating Characteristic

- Ось X — False Positive Rate

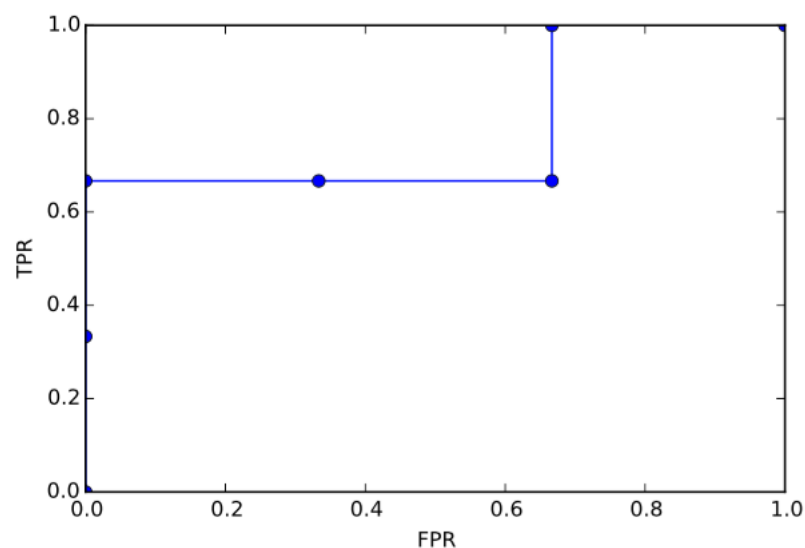
$$FPR = \frac{FP}{FP + TN}$$

Число
отрицательных
объектов

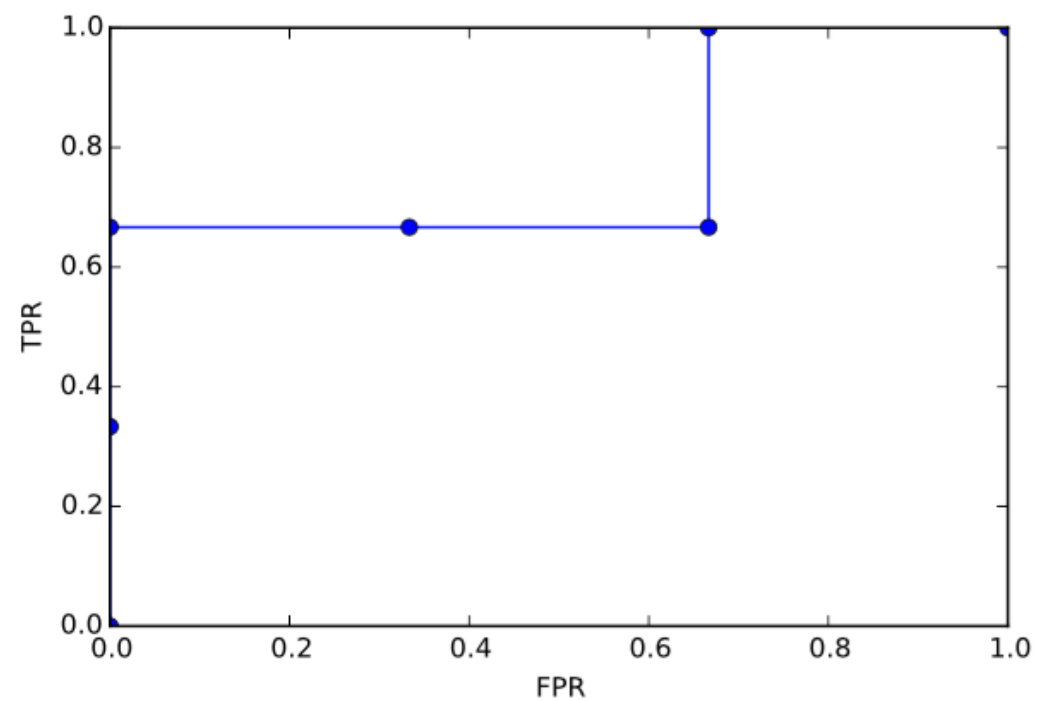
- Ось Y — True Positive Rate

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

Число
положительных
объектов

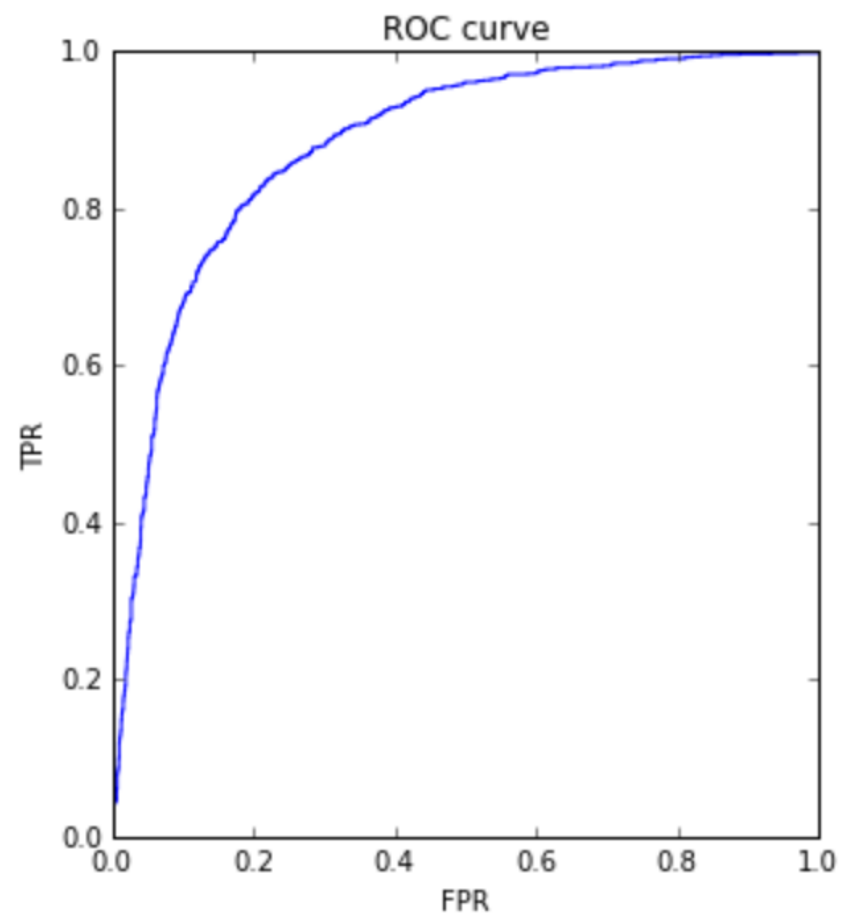


ROC-кривая



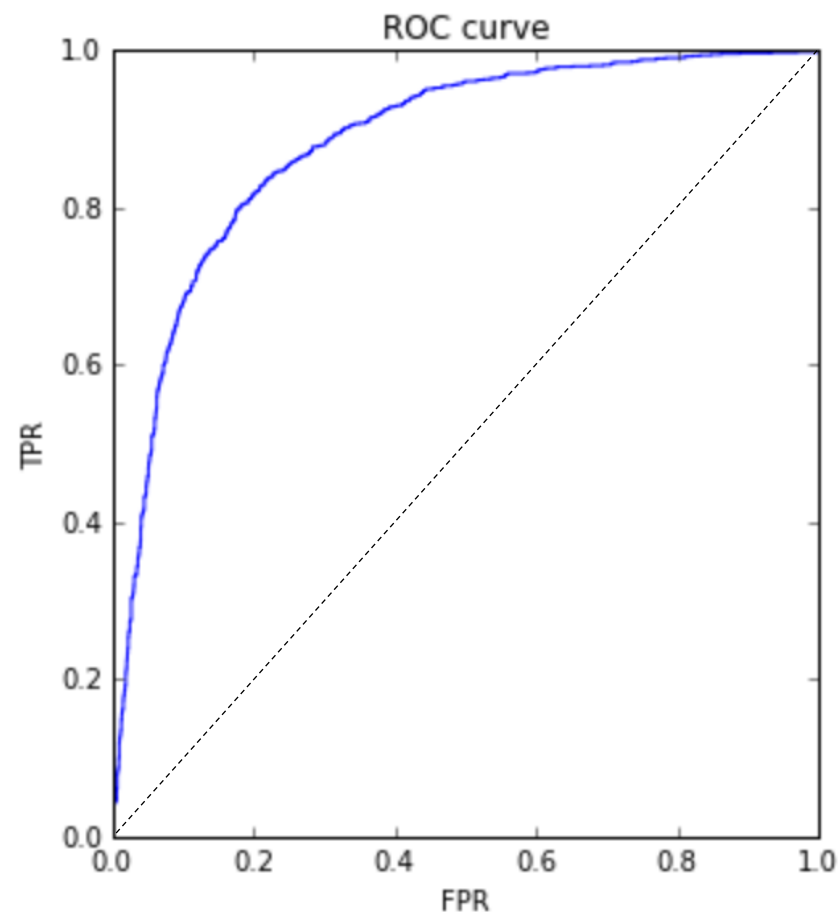
| | | | | | | |
|--------|------|------|------|------|------|------|
| $b(x)$ | 0.14 | 0.23 | 0.39 | 0.52 | 0.73 | 0.90 |
| y | 0 | 1 | 0 | 0 | 1 | 1 |

ROC-кривая в реальности



ROC-кривая

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



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: $AUC-ROC \approx 0.5$

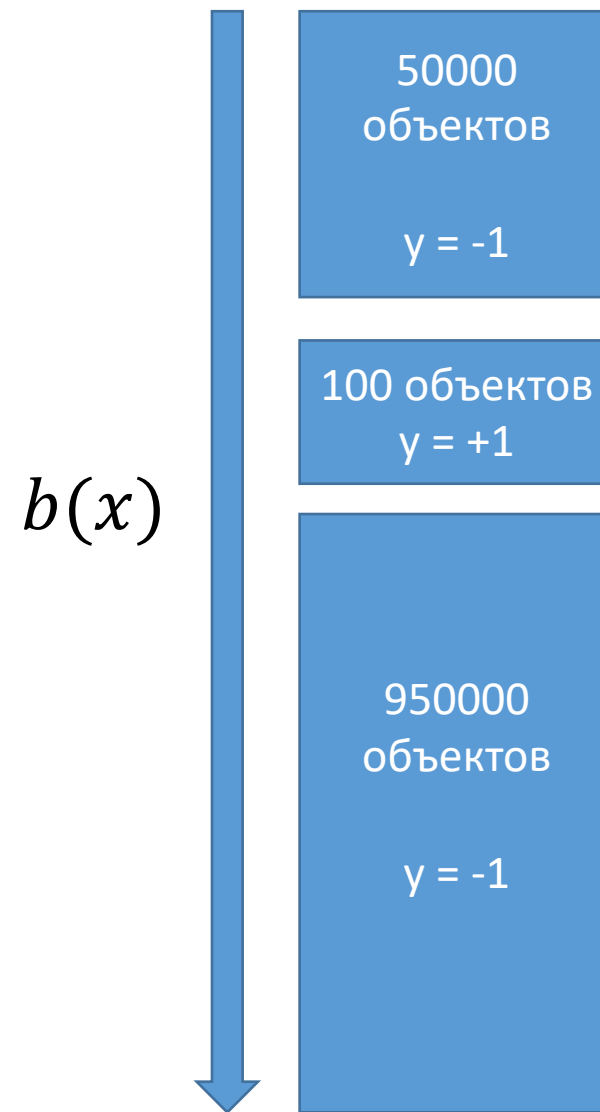
AUC-PRC

$$\text{precision} = \frac{TP}{TP+FP}; \quad \text{recall} = \frac{TP}{TP+FN}$$

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

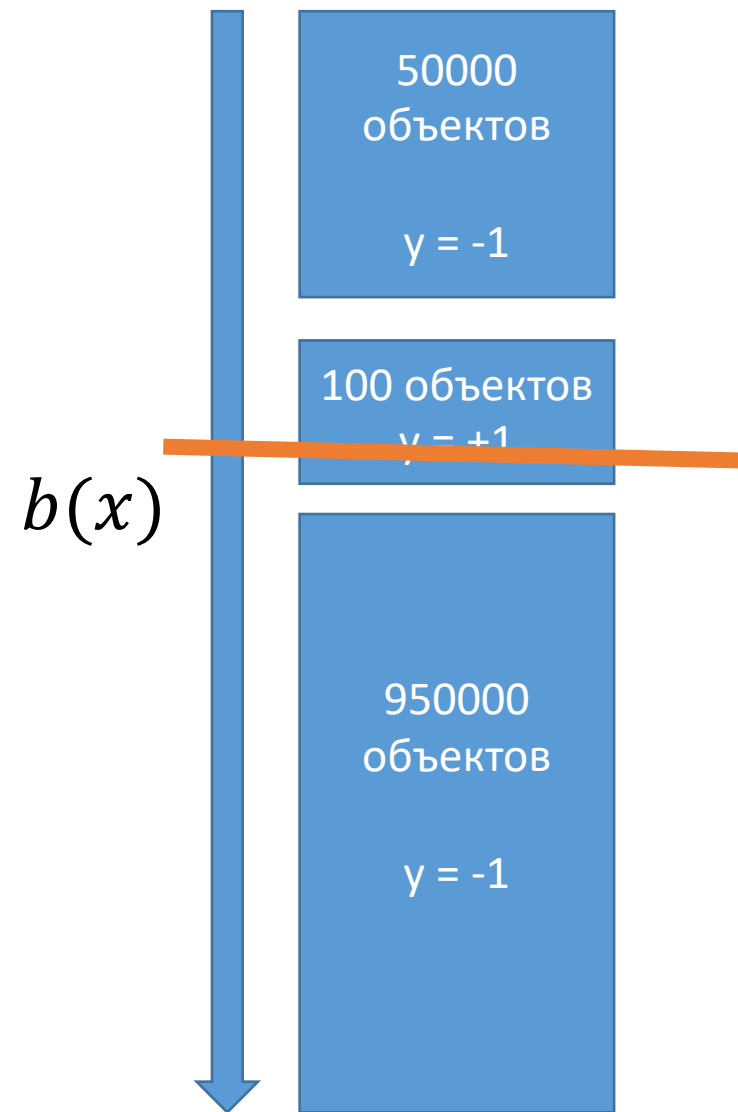
Пример

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



Пример

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



Резюме

- Два вида классификаторов:
 - Ответ — класс
 - Ответ — оценка принадлежности классу
- Метрики в первом случае: доля правильных ответов, точность, полнота, F-мера
- Метрики во втором случае: AUC-ROC, AUC-PRC
- В регрессии: MSE, MAE, R^2