

Основы машинного обучения

Лекция 9

Линейная классификация

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2023

Матрица ошибок

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

Точность (precision)

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

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

Полнота (recall)

- Как много положительных объектов находит классификатор?

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

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

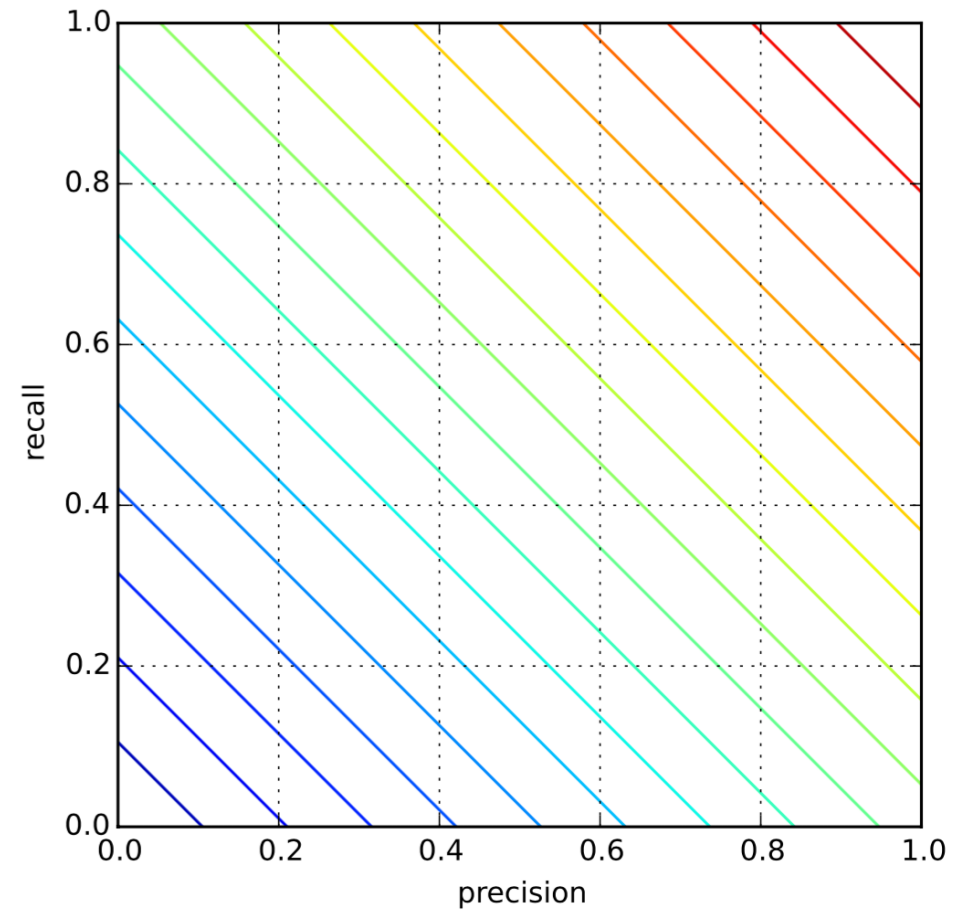
Совмещение точности и
полноты

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

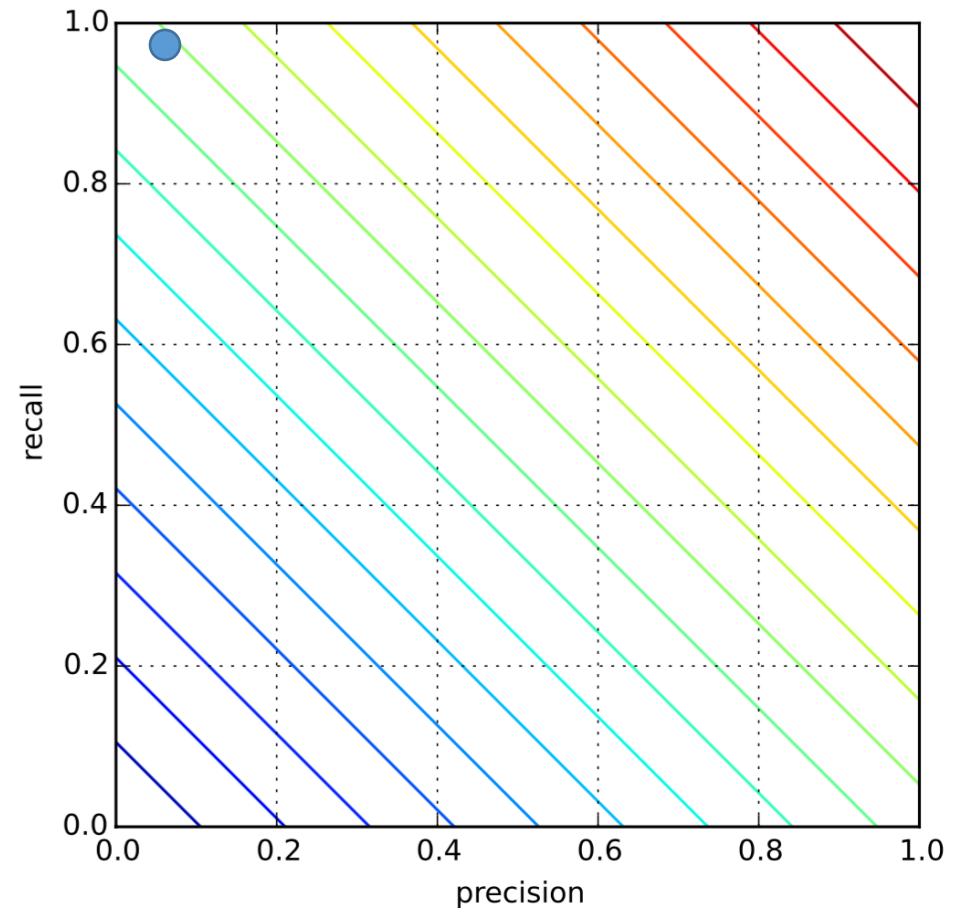
$$A = \frac{1}{2}(\text{precision} + \text{recall})$$



Арифметическое среднее

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

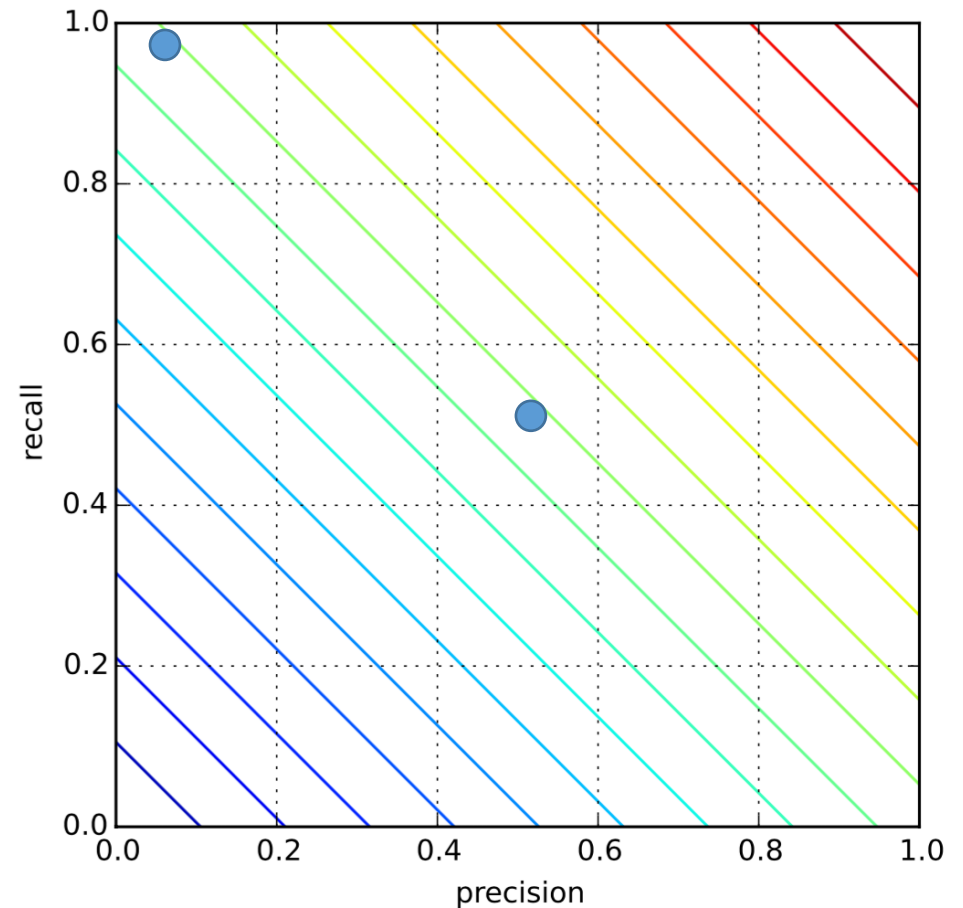
- precision = 0.1
- recall = 1
- $A = 0.55$
- Плохой алгоритм



Арифметическое среднее

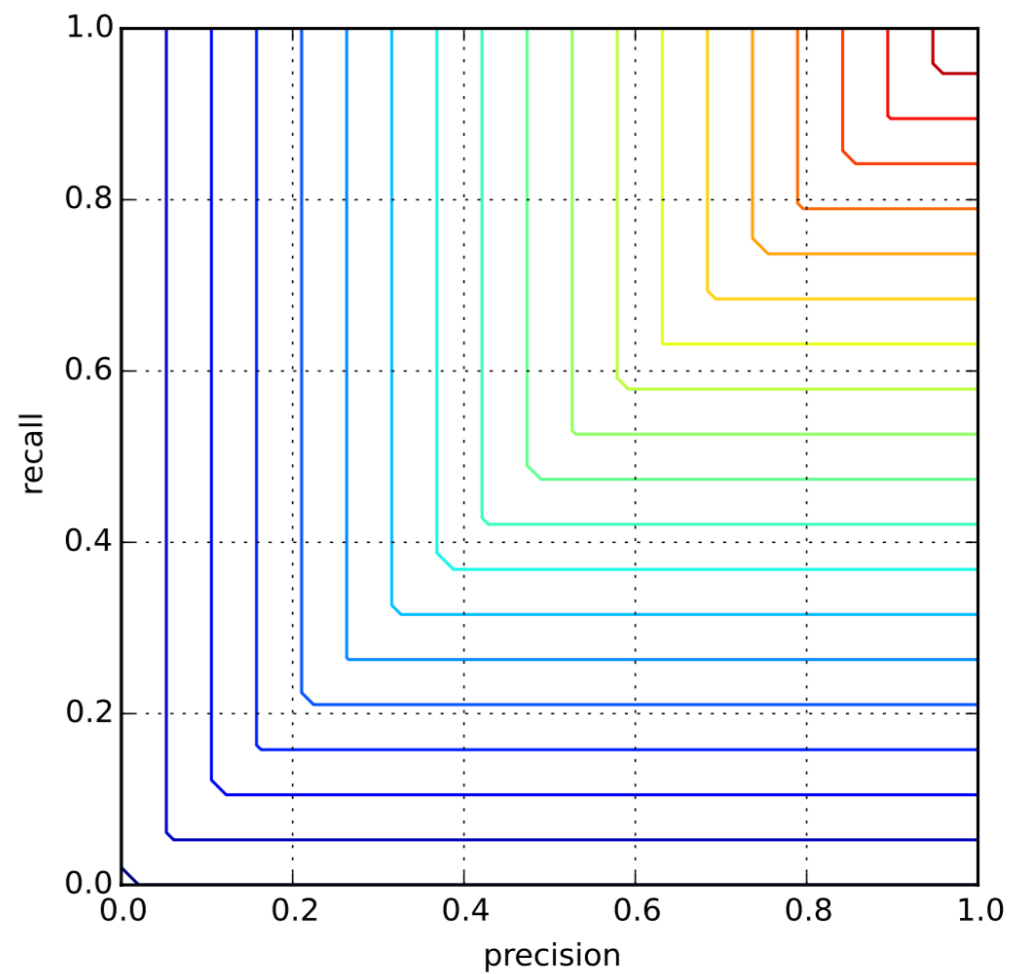
$$A = \frac{1}{2} (\text{precision} + \text{recall})$$

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



Минимум

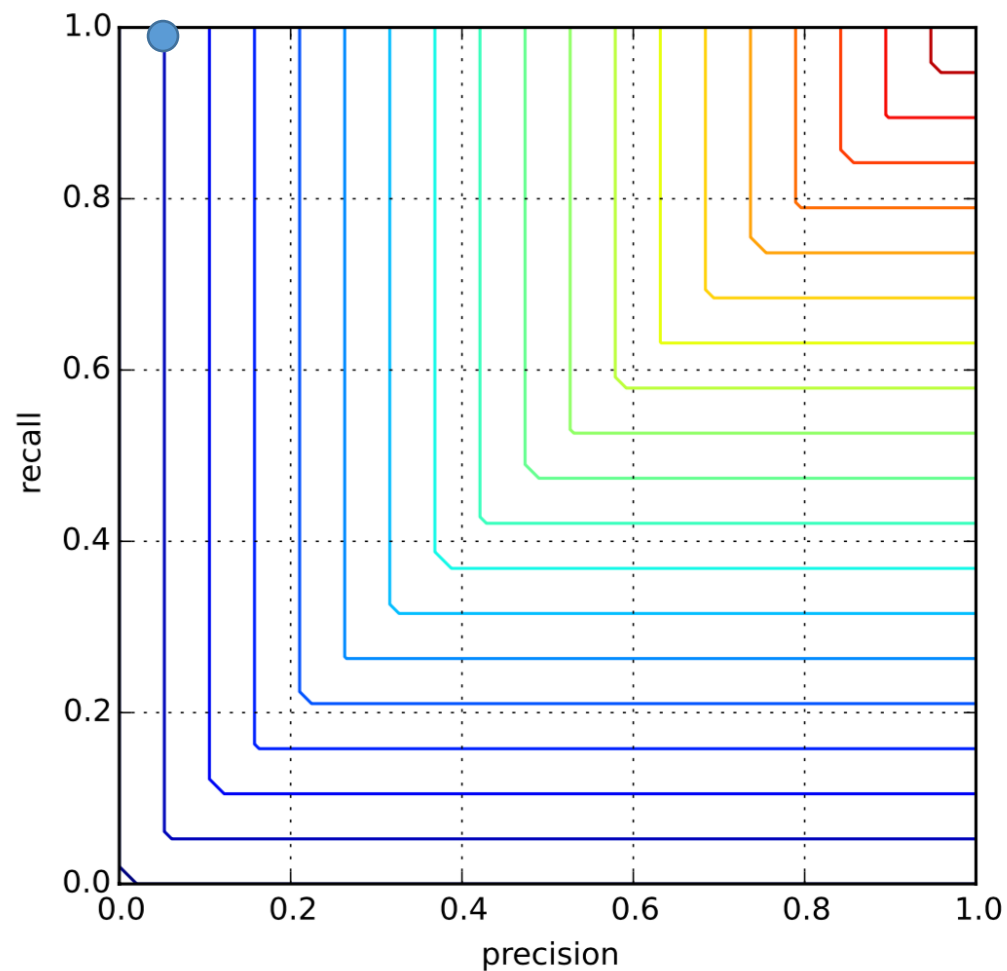
$$M = \min(\text{precision}, \text{recall})$$



Минимум

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

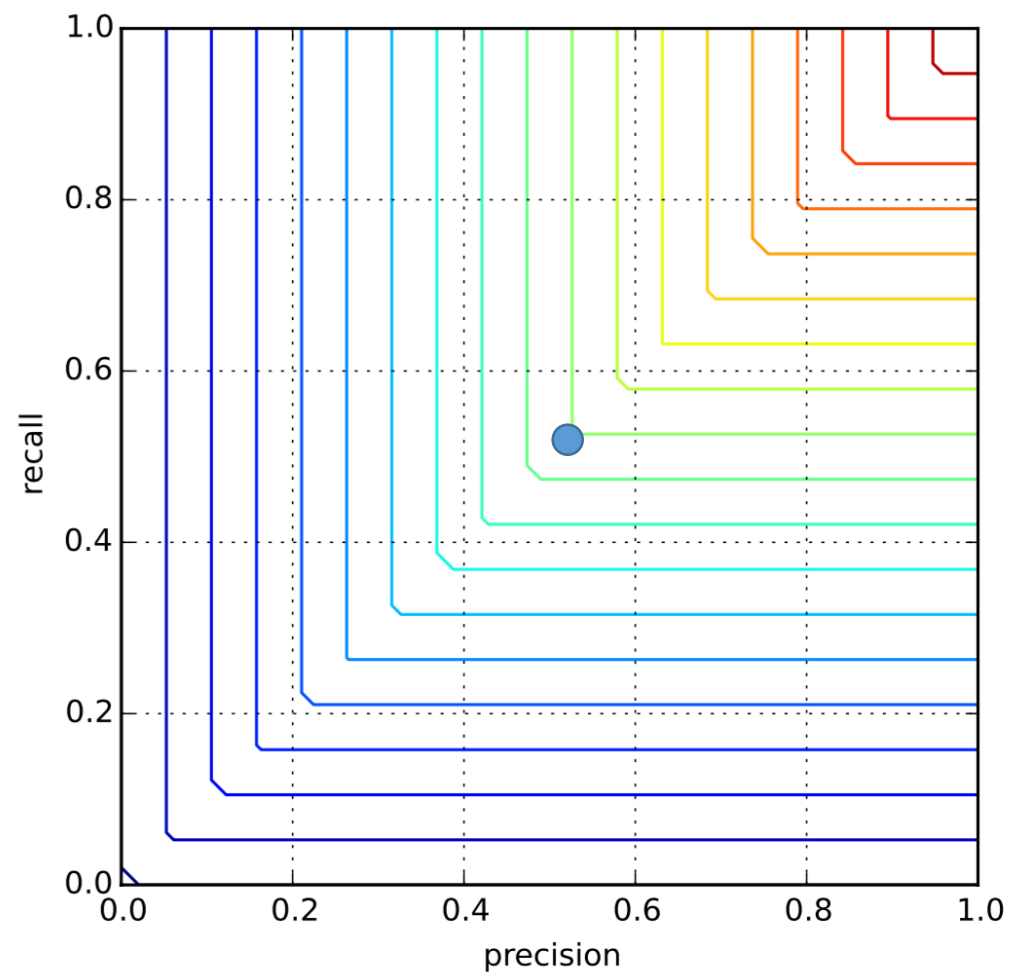
- precision = 0.05
- recall = 1
- $M = 0.05$



Минимум

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

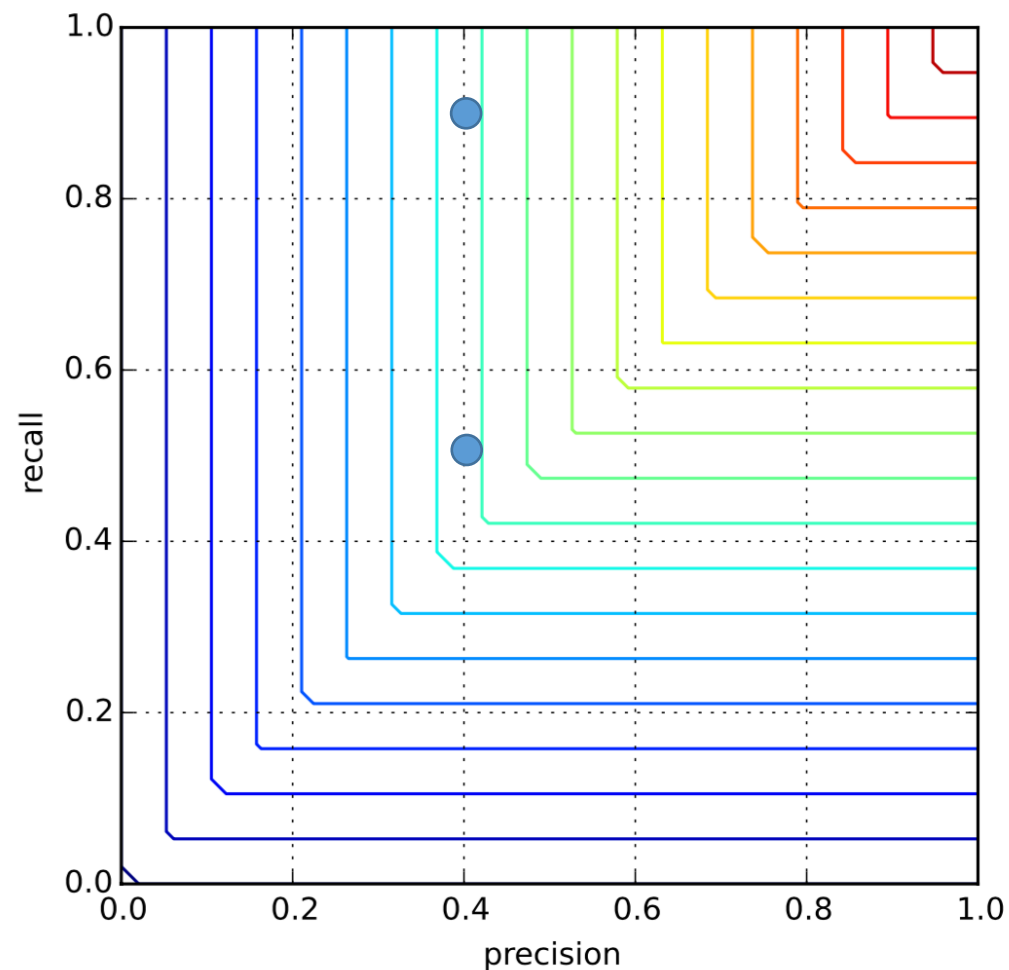
- precision = 0.55
- recall = 0.55
- $M = 0.55$



Минимум

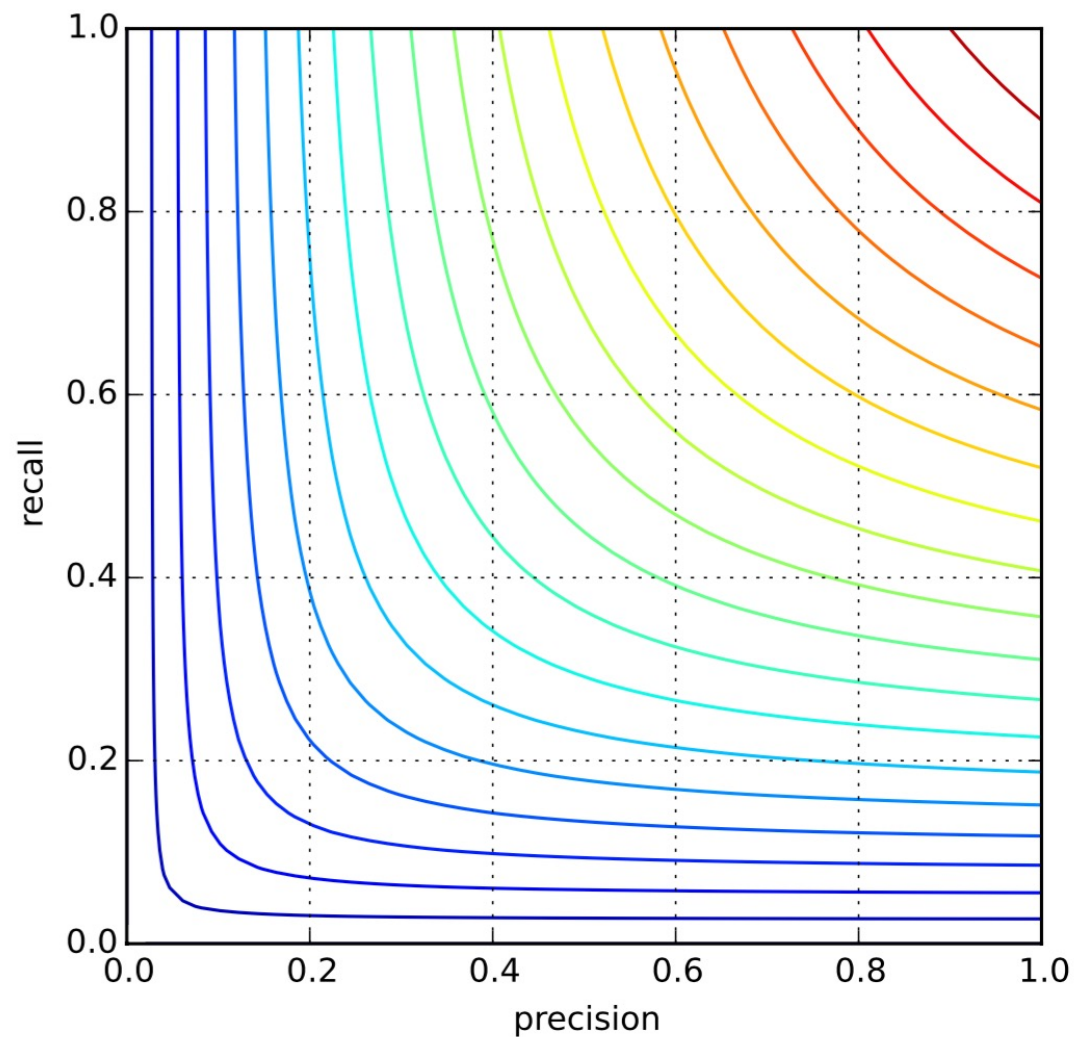
$$M = \min(\text{precision}, \text{recall})$$

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



F-measure

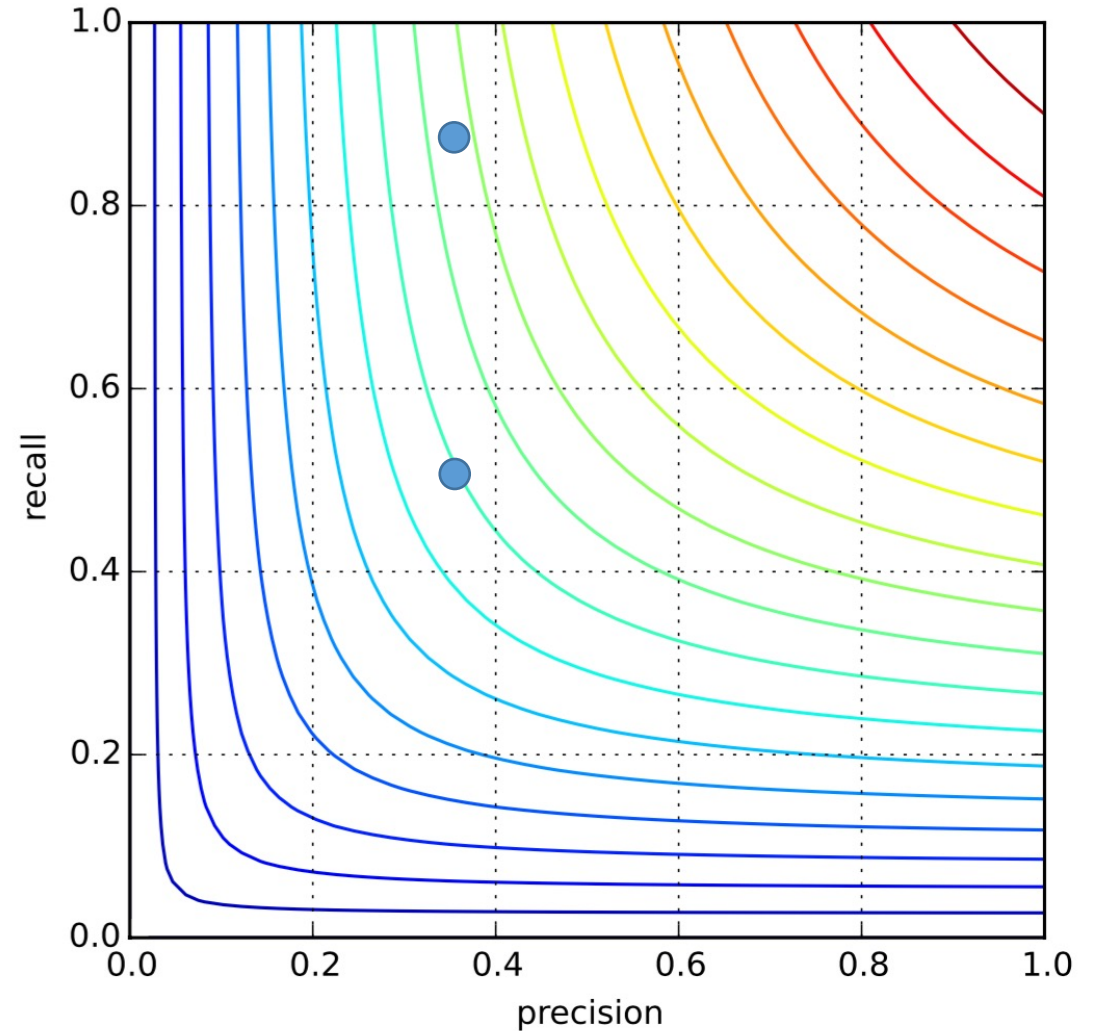
$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



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
- $M = 0.55$



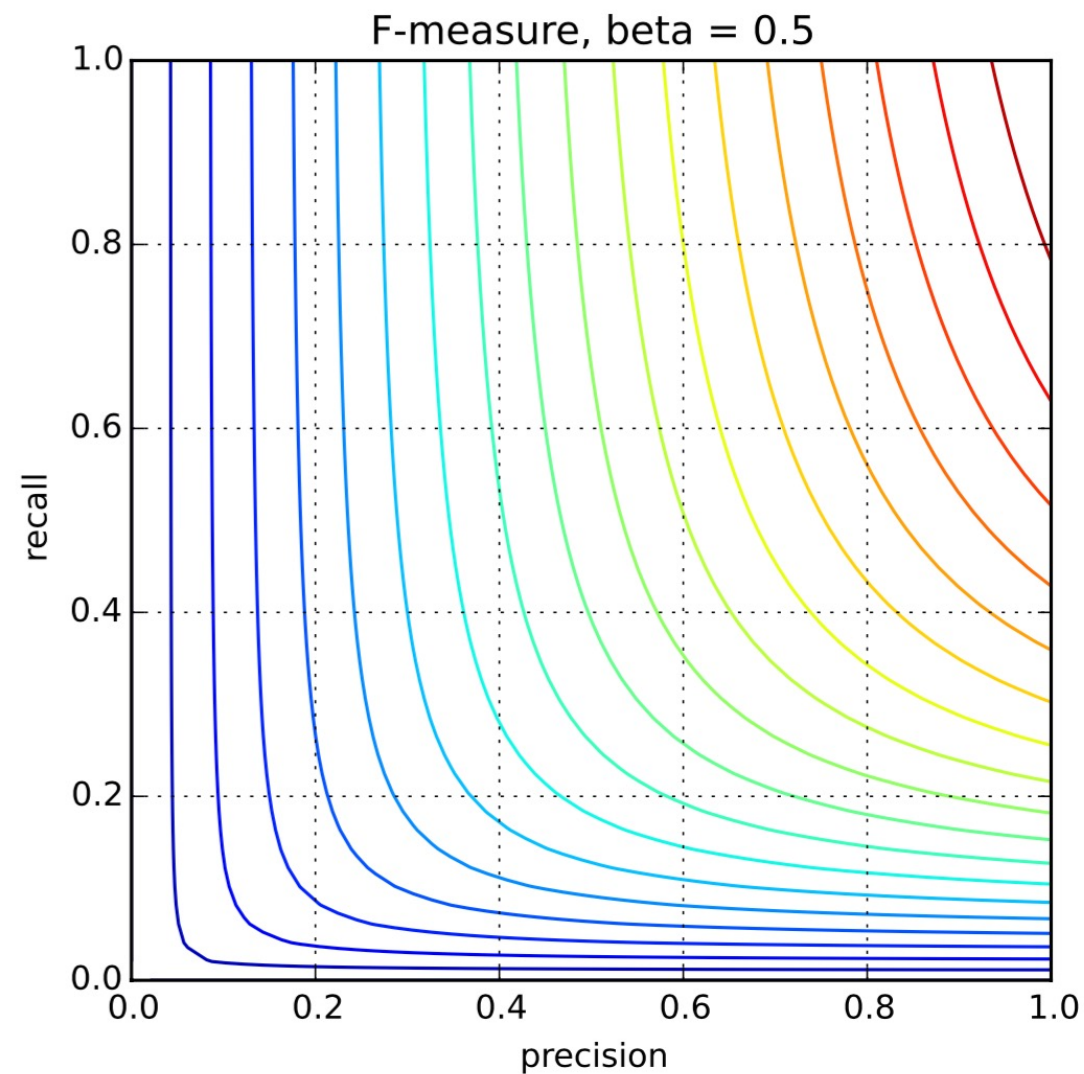
F-measure

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

F-мера

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

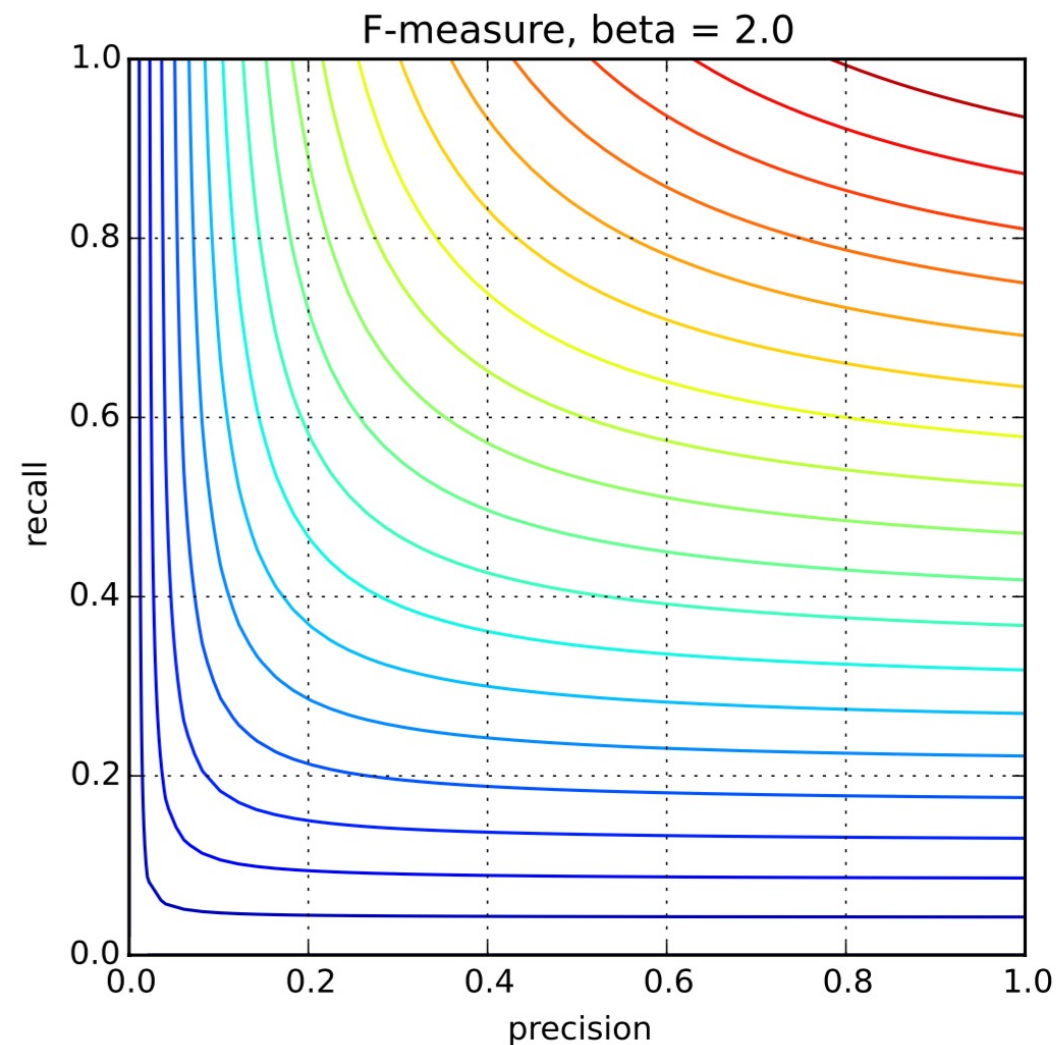
- $\beta = 0.5$
- Важнее точность



F-мера

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

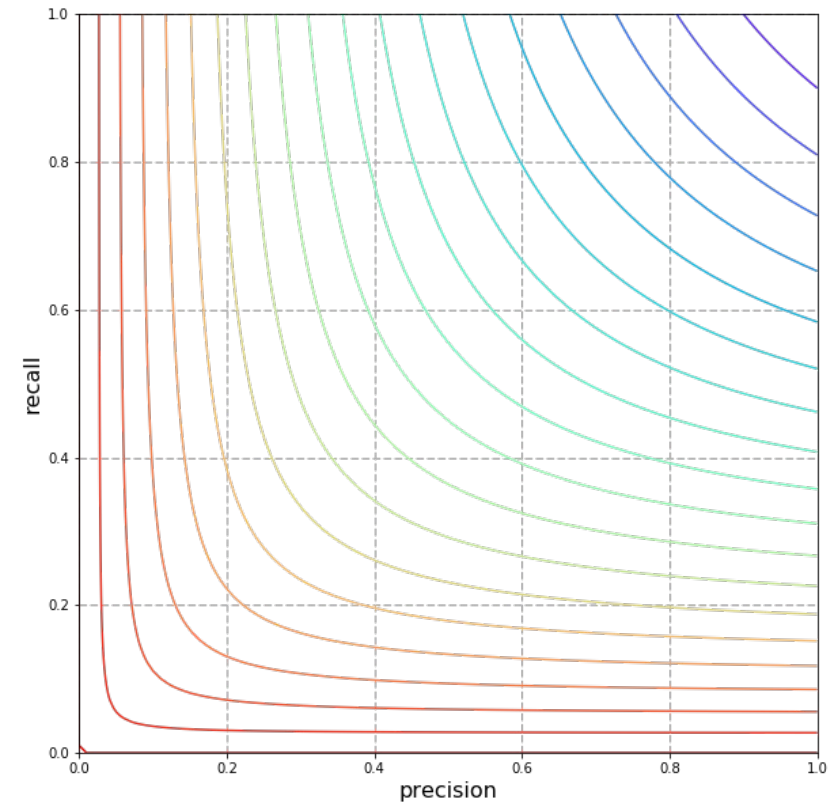
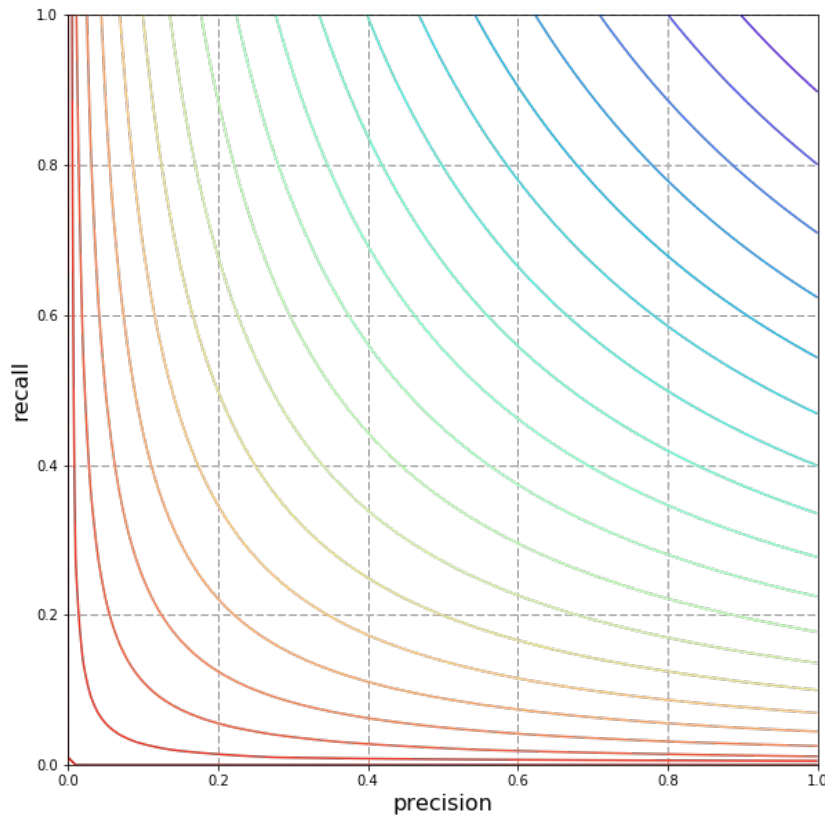
- $\beta = 2$
- Важнее полнота



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

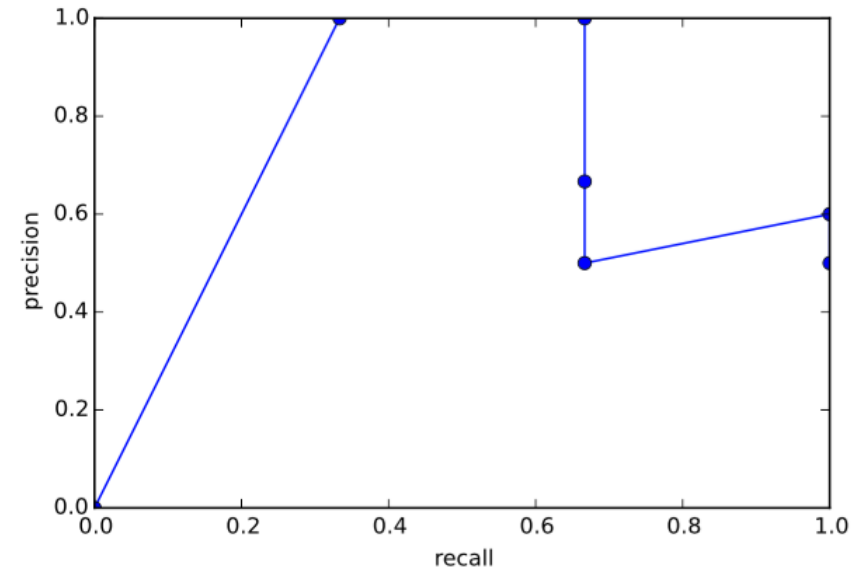
- Как оценить качество $b(x)$?
- Порог выбирается позже
- Порог зависит от ограничений на точность или полноту

Оценка принадлежности

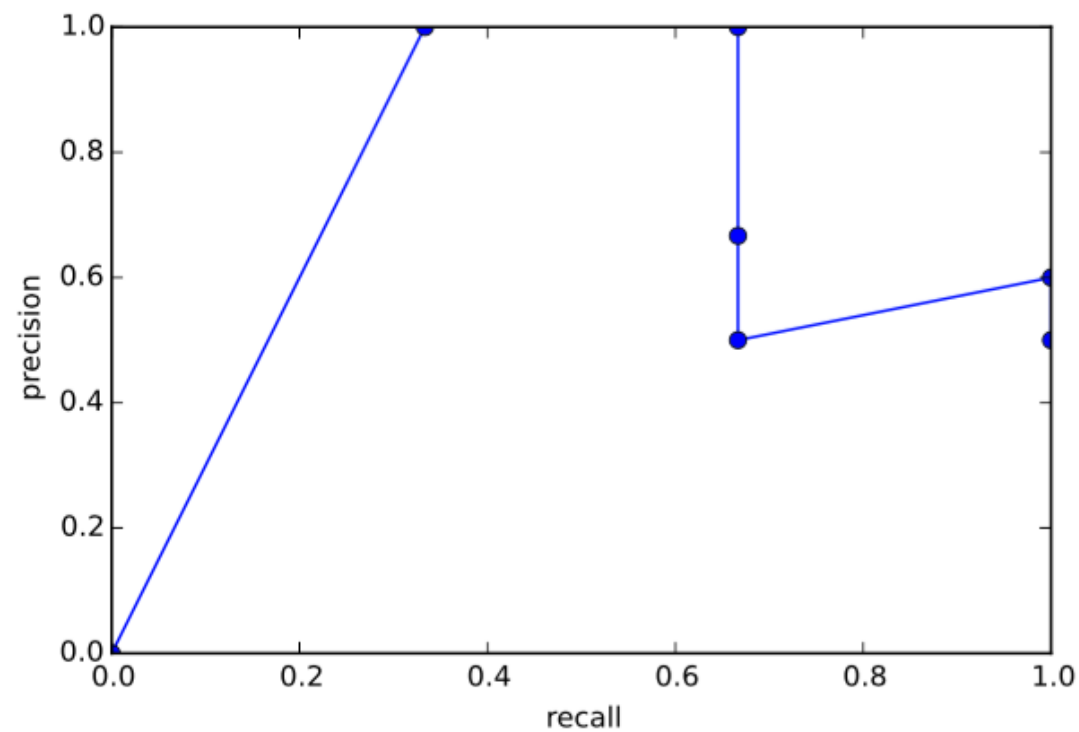
- Пример: кредитный скоринг
- $b(x)$ — оценка вероятности возврата кредита
- $a(x) = [b(x) > 0.5]$
- precision = 0.1, recall = 0.7
- В чем дело — в пороге или в алгоритме?

PR-кривая

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

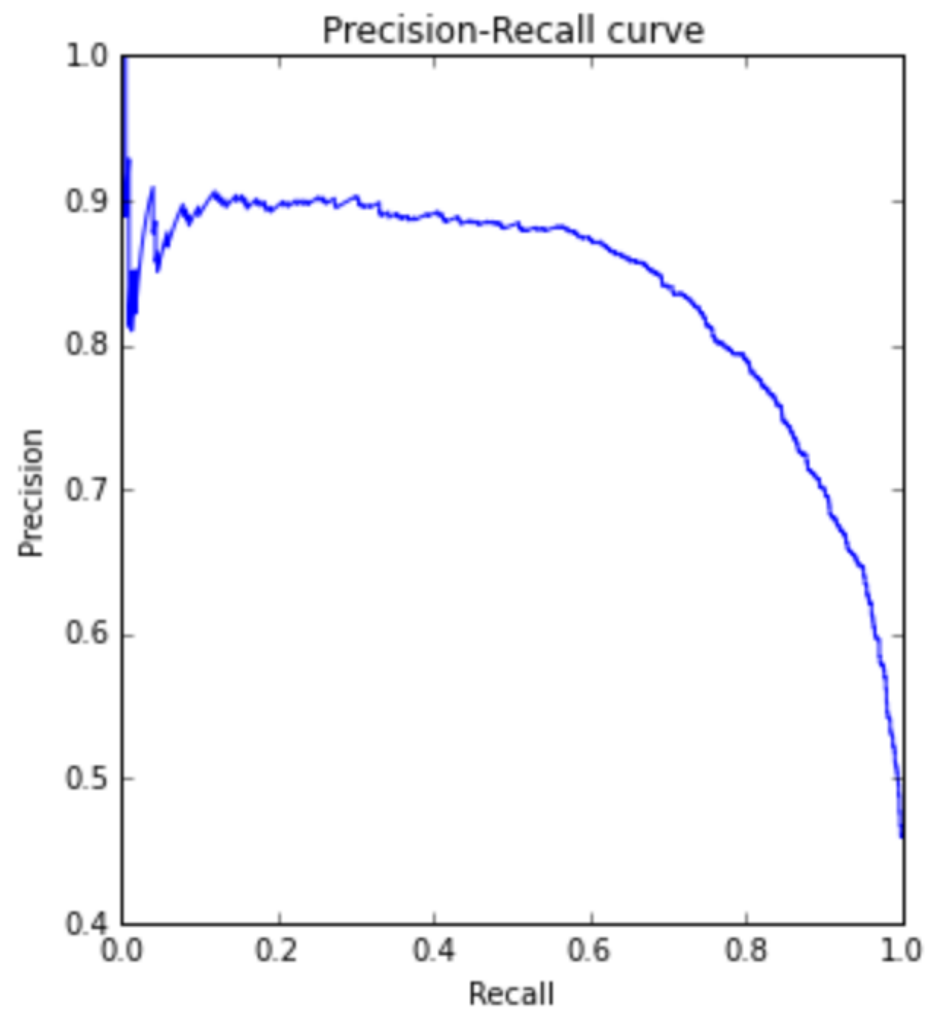


PR-кривая



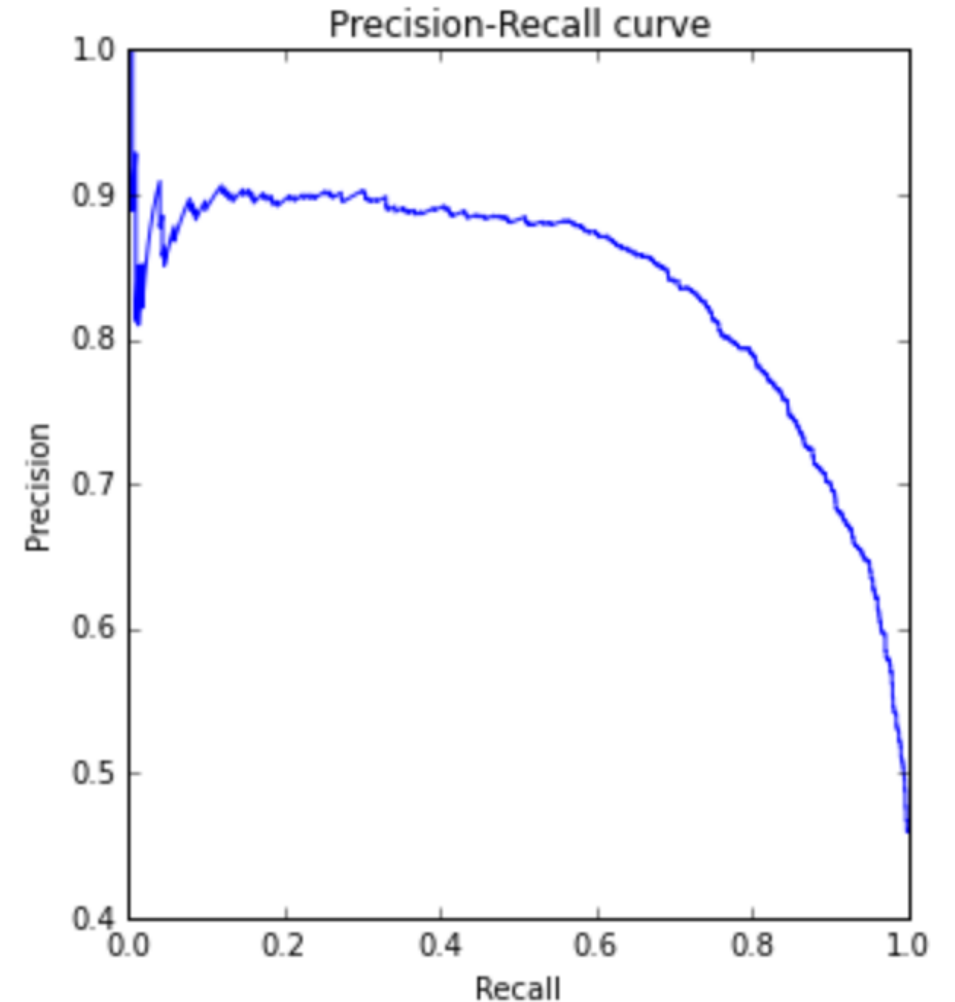
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

PR-кривая в реальности

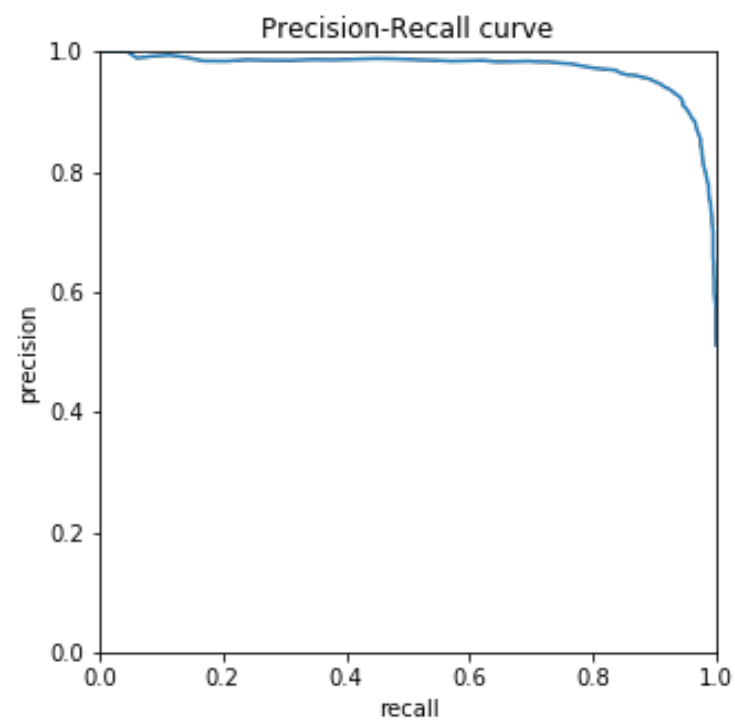
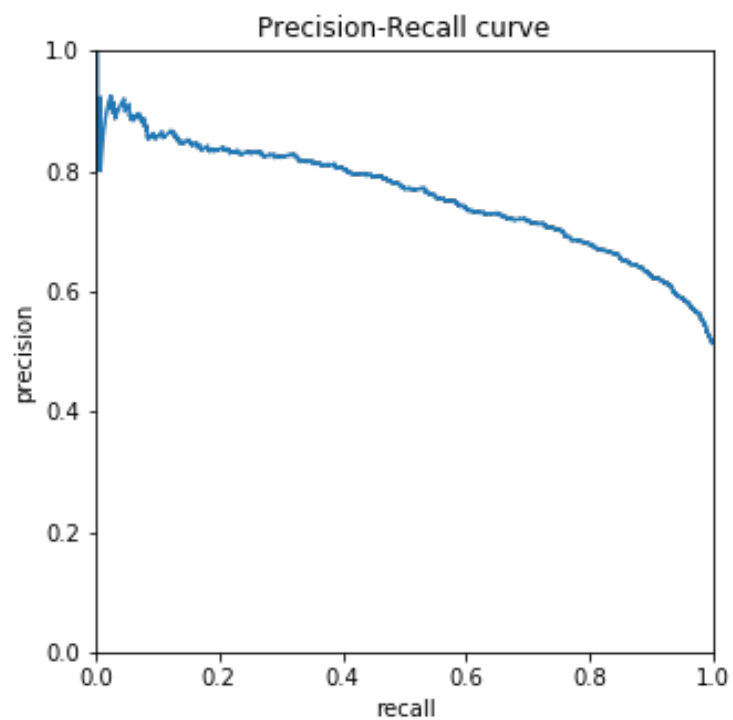


PR-кривая

- 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



ROC-кривая

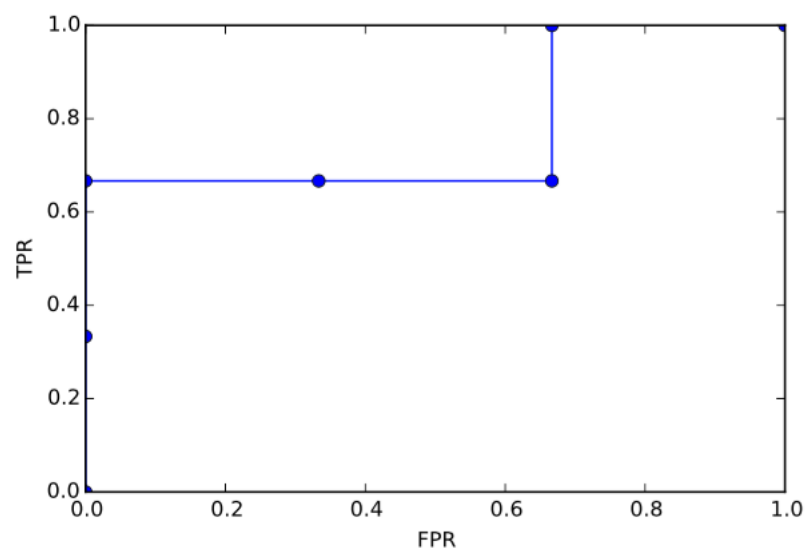
- Receiver Operating Characteristic

- 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

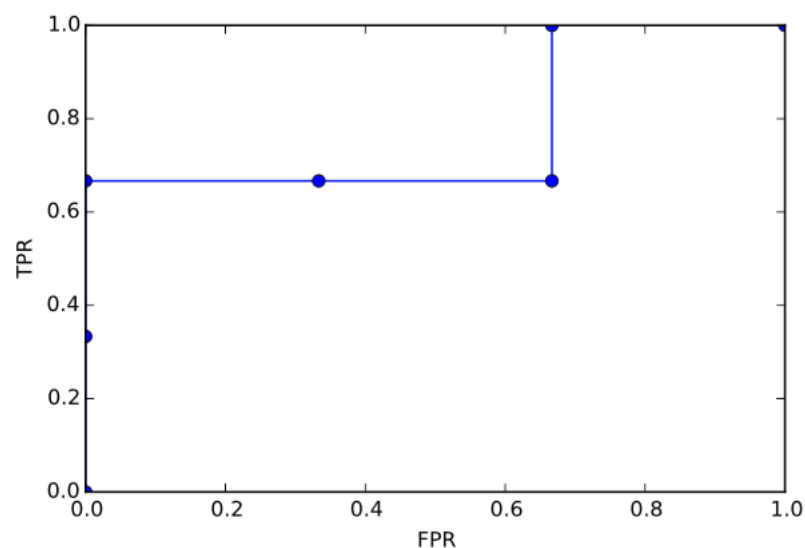
$$FPR = \frac{FP}{FP + TN}$$

Число
отрицательных
объектов

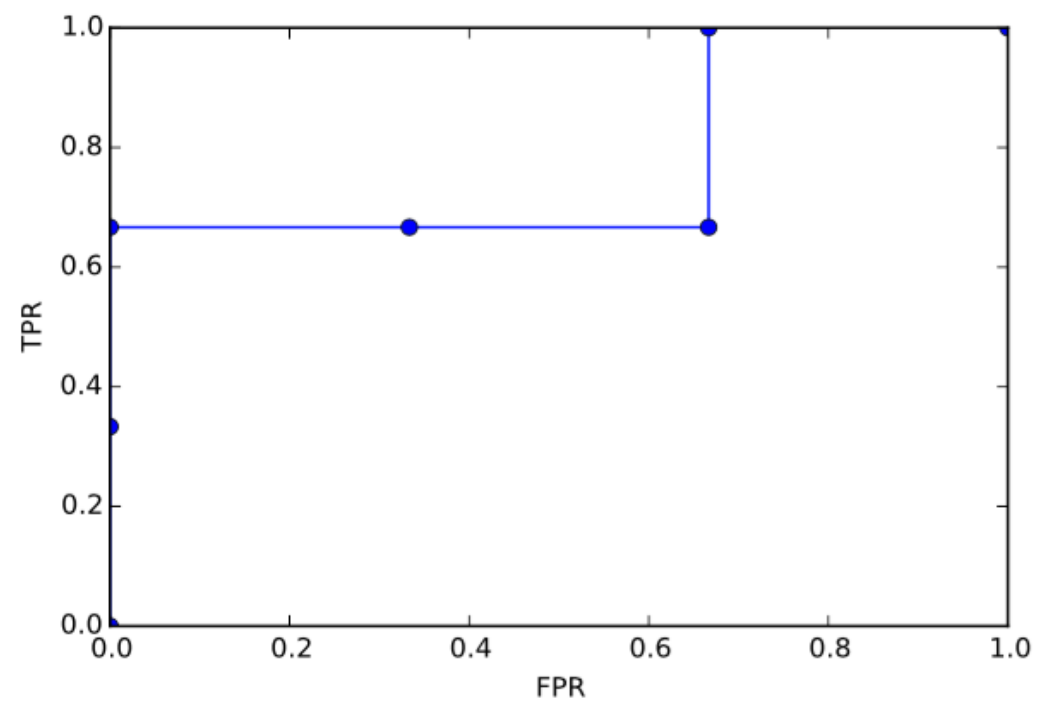
- Ось Y — True Positive Rate

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

Число
положительных
объектов

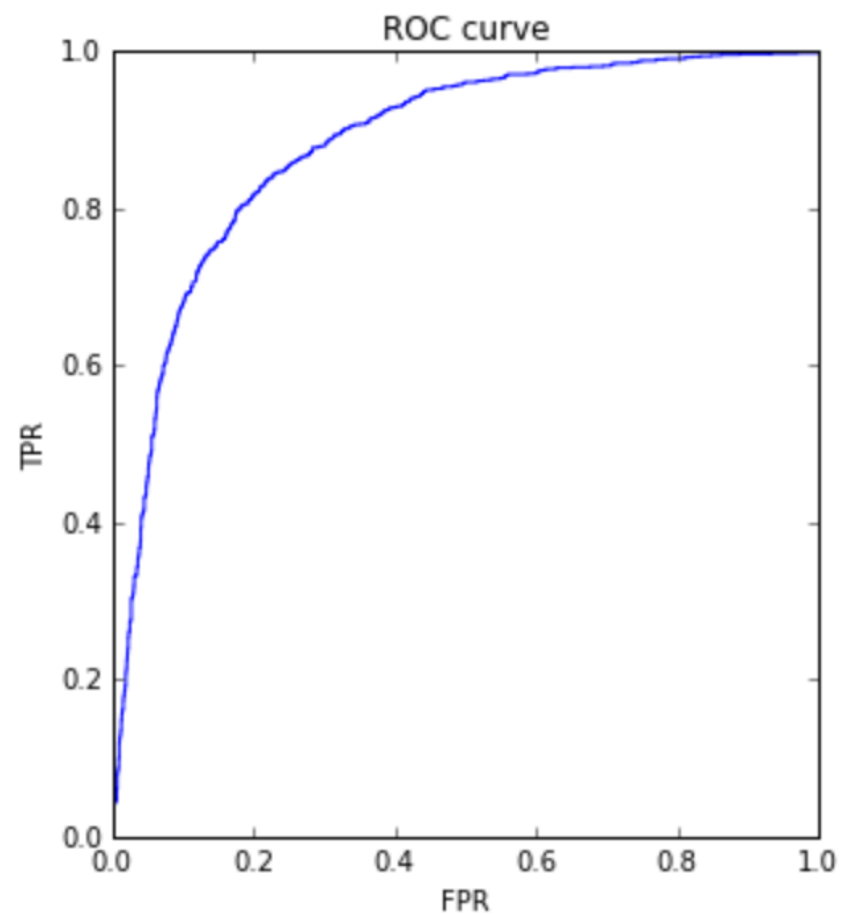


ROC-кривая



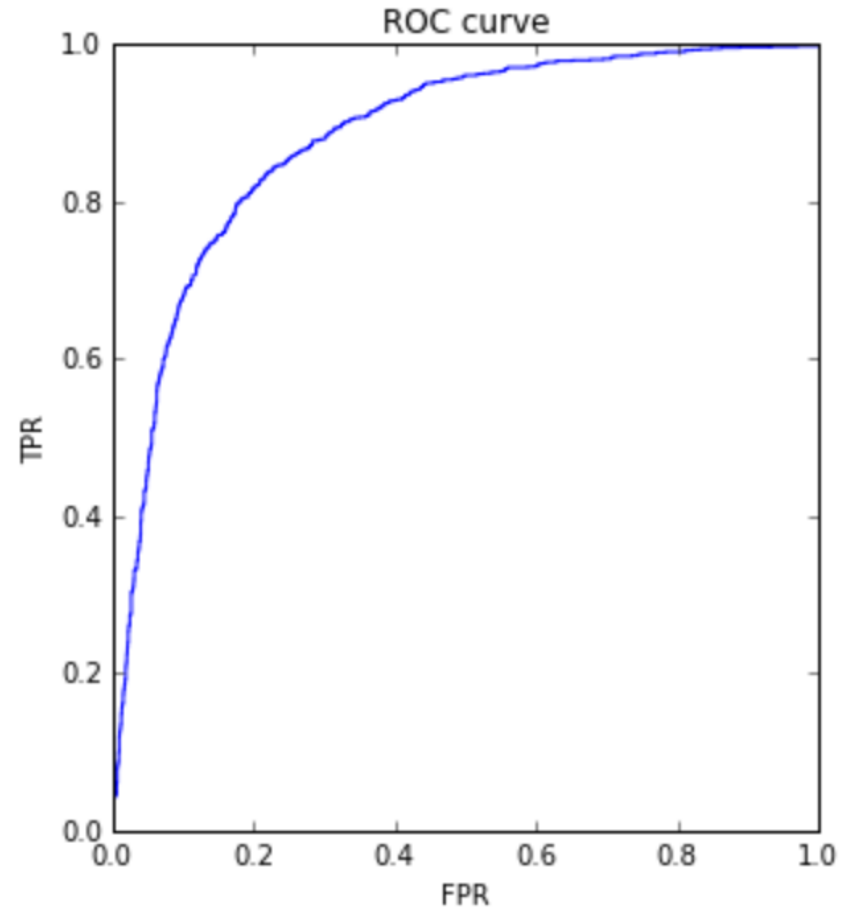
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

ROC-кривая в реальности

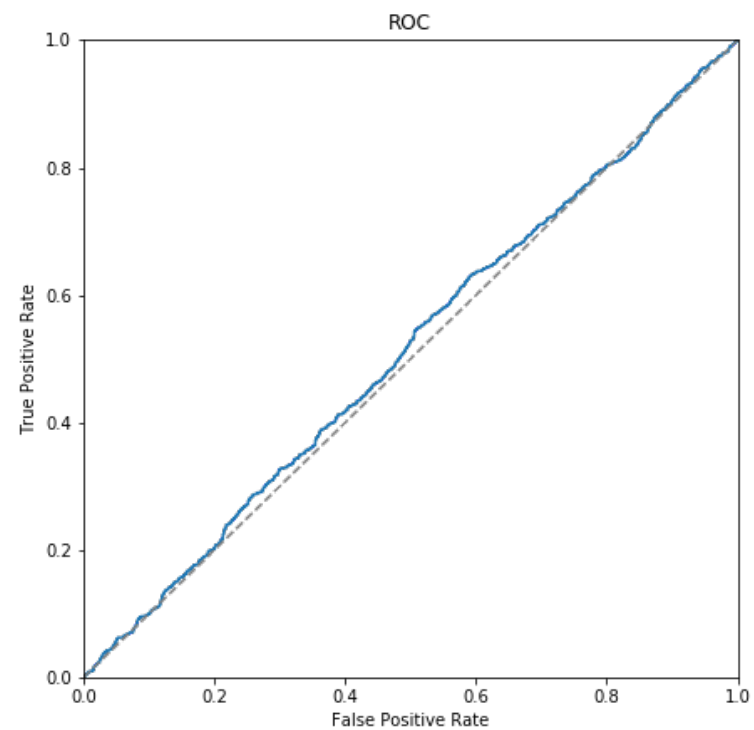
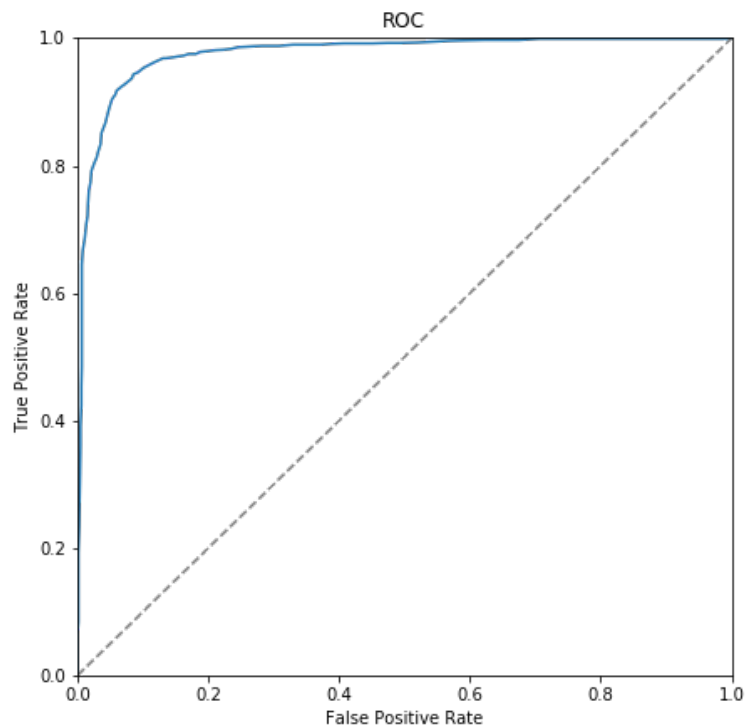


ROC-кривая

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



ROC-кривая



AUC-ROC

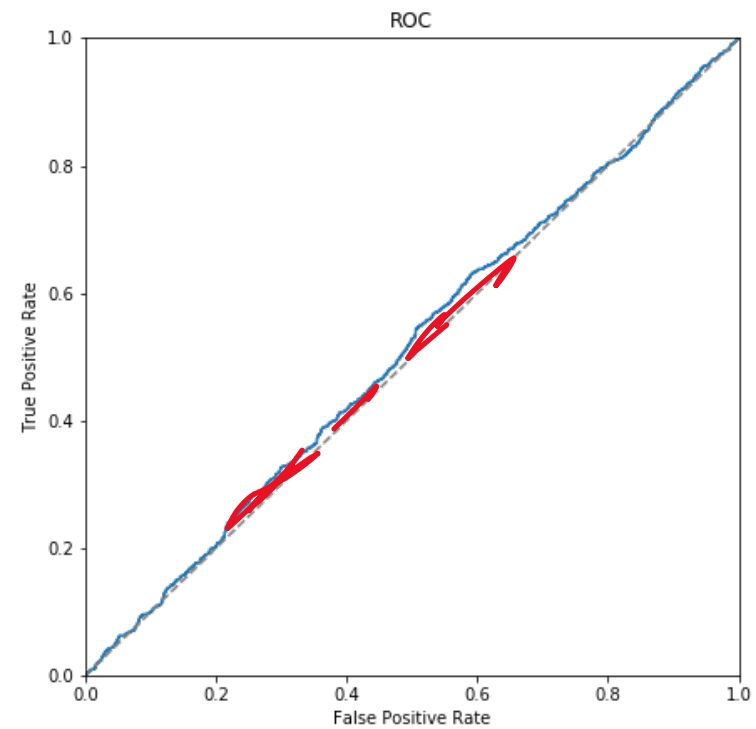
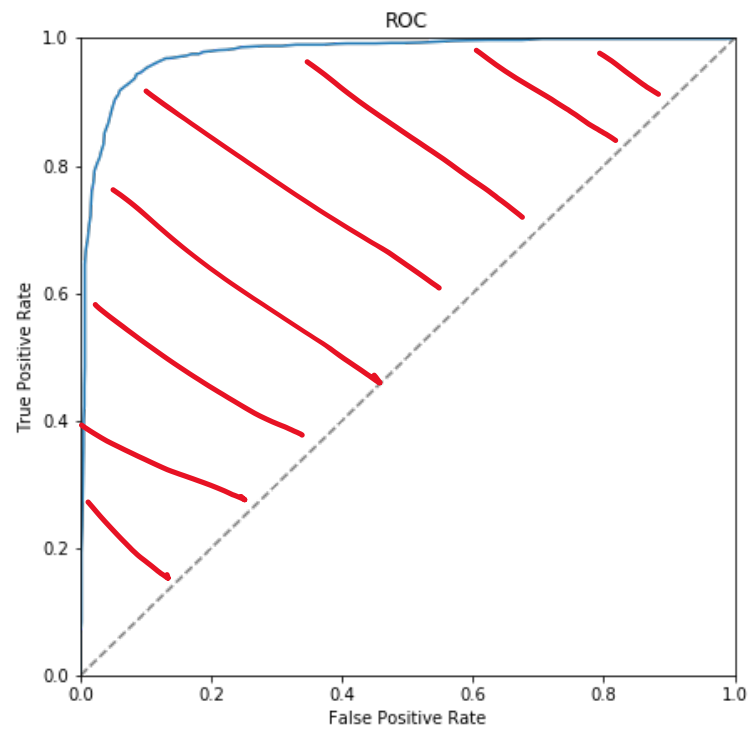
$$FPR = \frac{FP}{FP+TN};$$

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

- 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

$$\text{Gini} = 2 * (\text{AUC-ROC} - 0.5)$$



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
- Проще интерпретировать, если выборка несбалансированная
- Лучше, если задачу надо решать в терминах точности и полноты

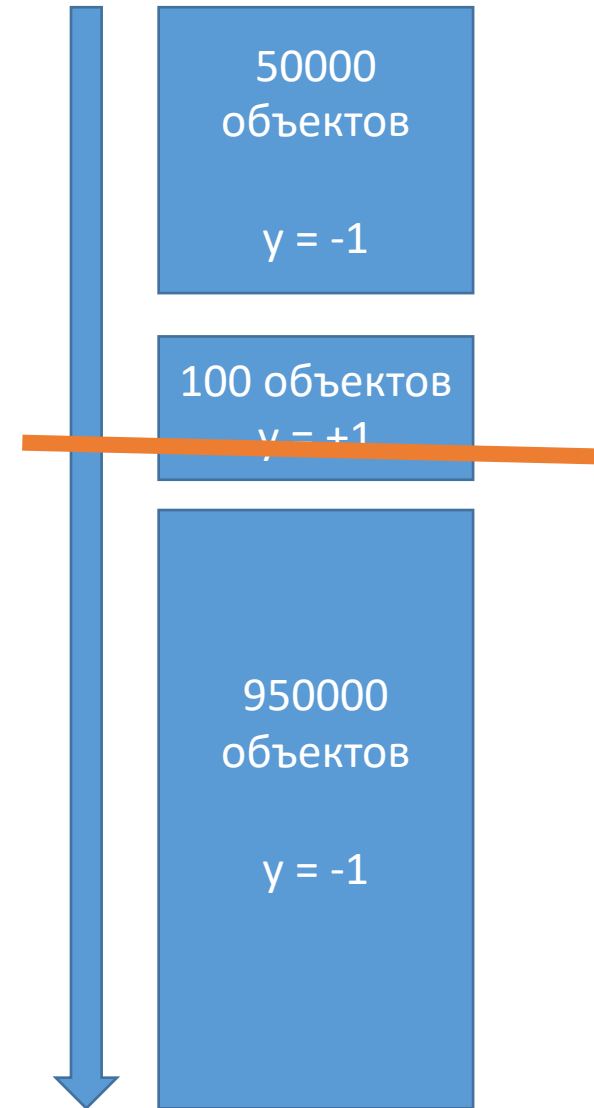
Пример

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



Пример

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



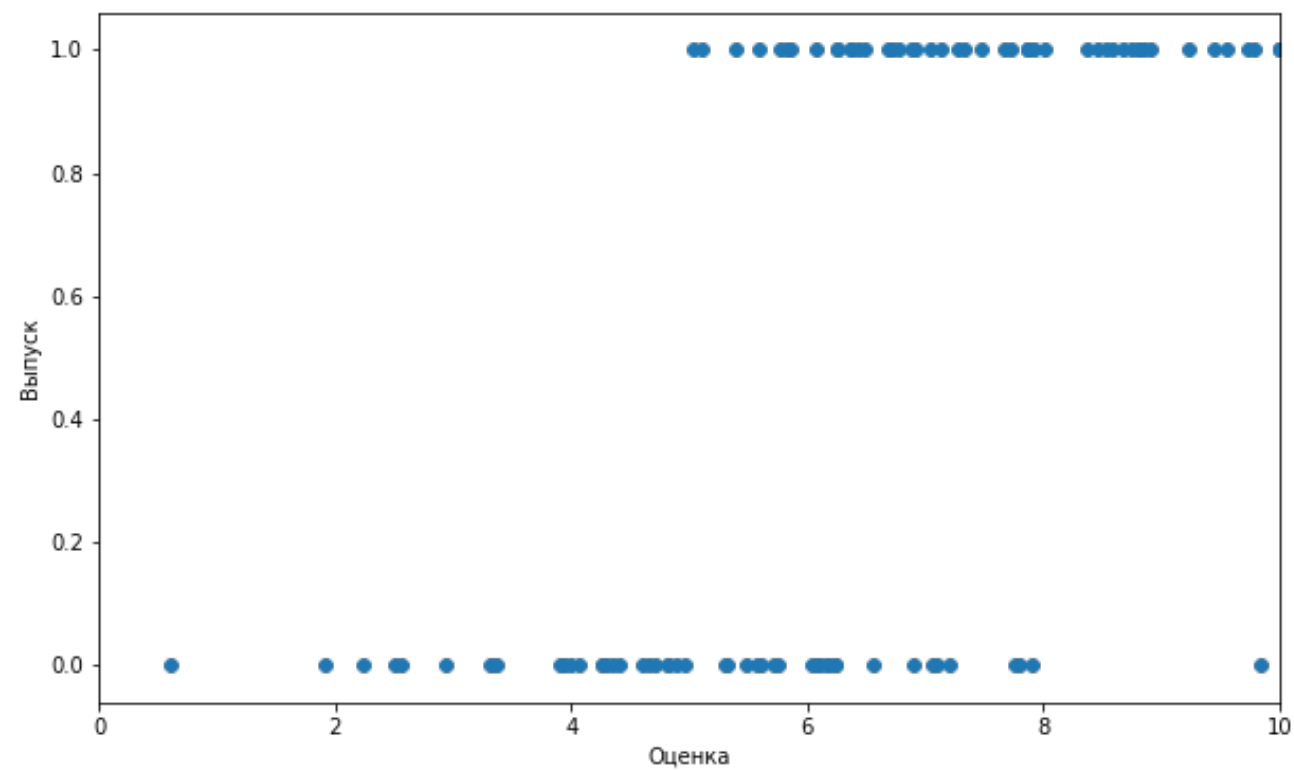
Логистическая регрессия:
простое объяснение

Логистическая регрессия

- Решаем задачу бинарной классификации: $\mathbb{Y} = \{-1, +1\}$
- Минимизация верхней оценки:

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

Предсказание вероятностей



Предсказание вероятностей



Предсказание вероятностей

- Кредитный скоринг
- Стратегия: выдавать кредит только клиентам с $b(x) > 0.9$
- 10% невозвращённых кредитов — нормально

Предсказание вероятностей

- Баннерная реклама
- $b(x)$ — вероятность, что пользователь кликнет по рекламе
- $c(x)$ — прибыль в случае клика
- $c(x)b(x)$ — хотим оптимизировать

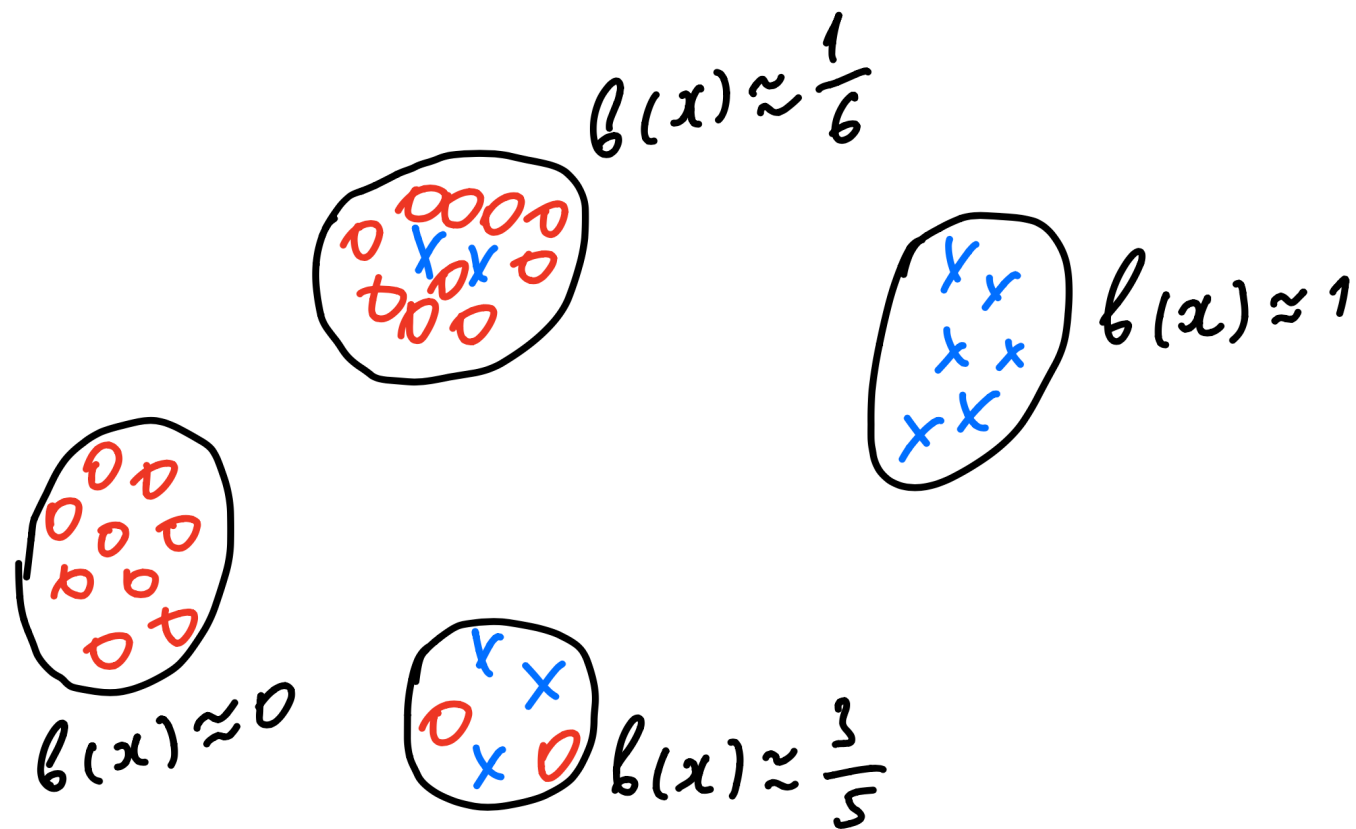
Предсказание вероятностей

- Прогнозирование оттока клиентов
- Медицинская диагностика
- Поисковое ранжирование (насколько веб-страница соответствует запросу?)

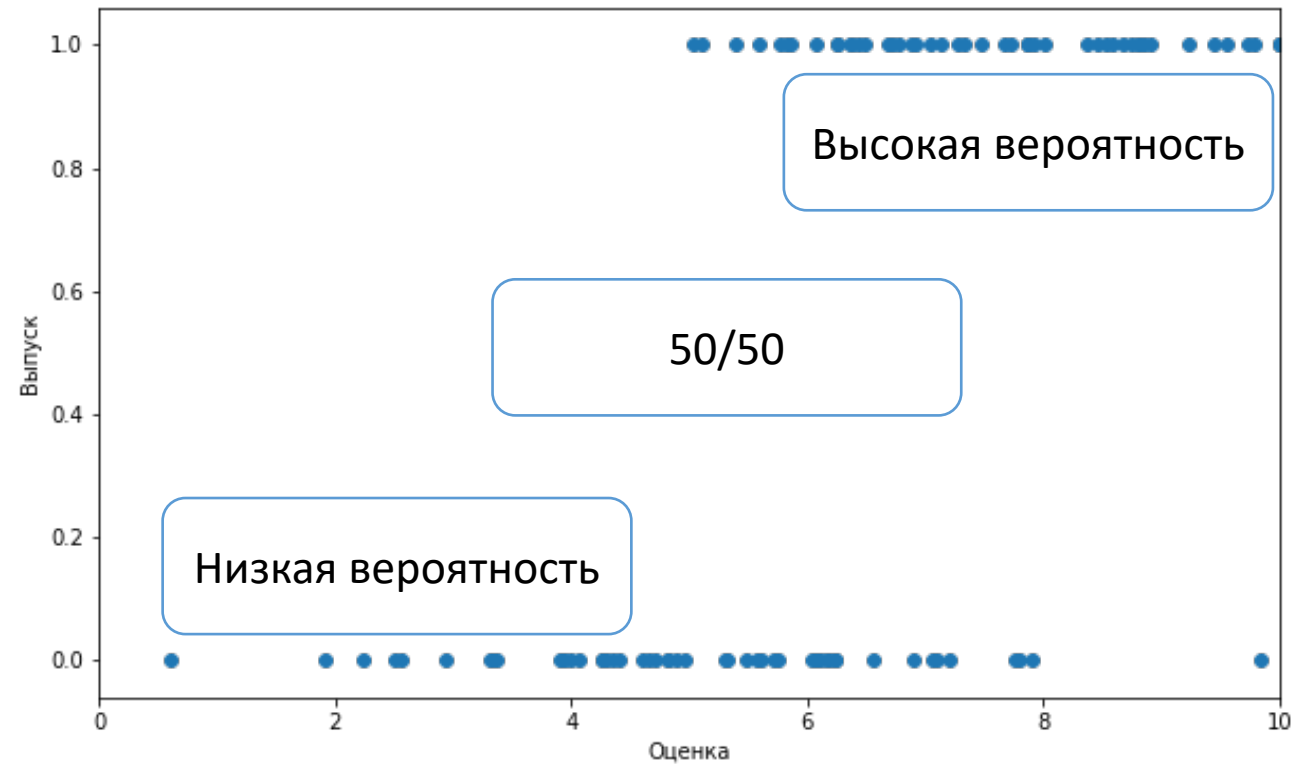
Предсказание вероятностей

Будем говорить, что модель $b(x)$ предсказывает вероятности, если среди объектов с $b(x) = p$ доля положительных равна p .

Предсказание вероятностей



Предсказание вероятностей



Линейный классификатор

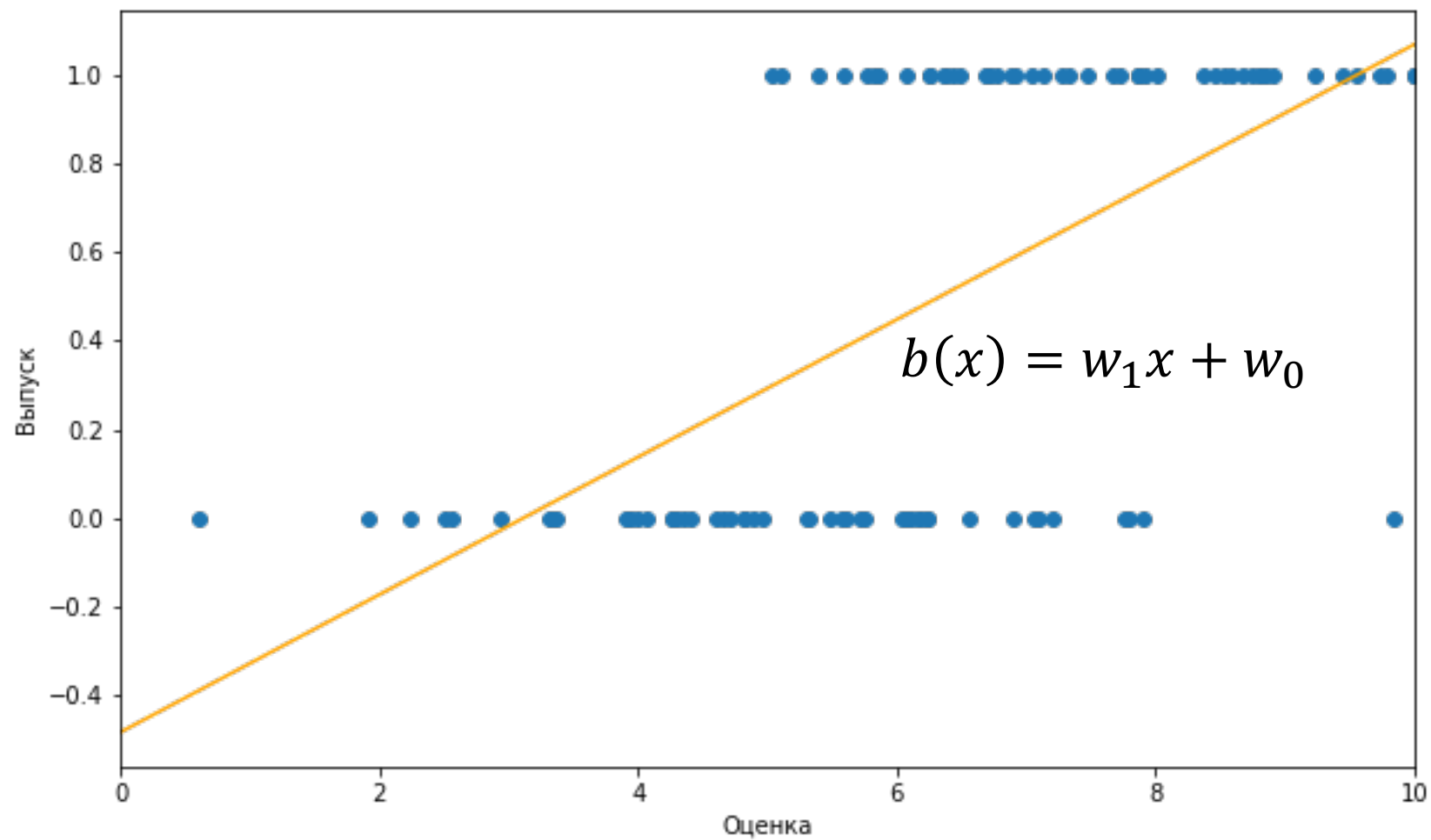
$$a(x) = \text{sign } \langle w, x \rangle$$

- Обучим как-нибудь — например, на логистическую функцию потерь:

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

- Может, $\langle w, x \rangle$ сойдёт за оценку?

Предсказание вероятностей

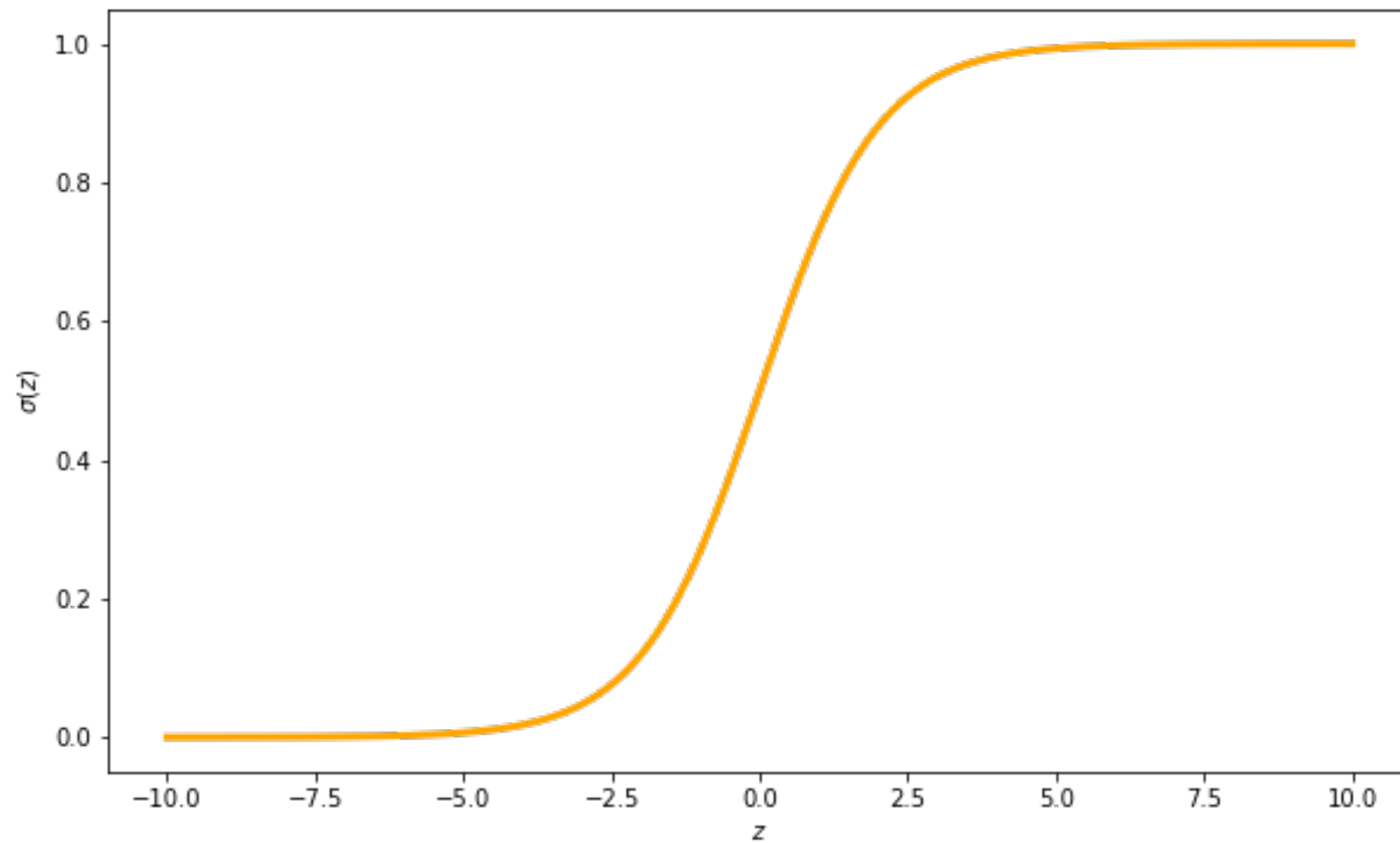


Линейный классификатор

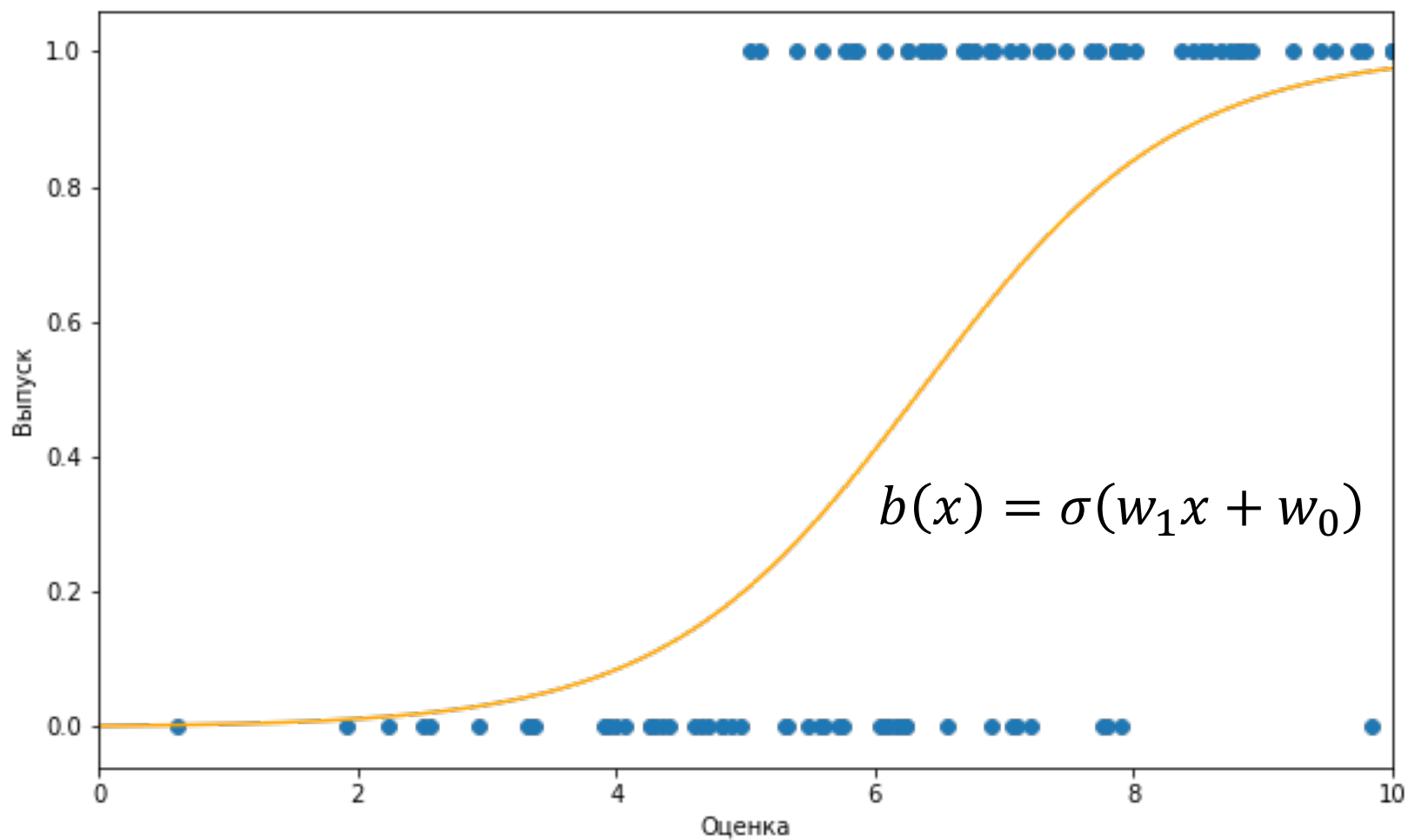
- Переведём выход модели на отрезок $[0, 1]$
- Например, с помощью сигмоиды:

$$\sigma(\langle w, x \rangle) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

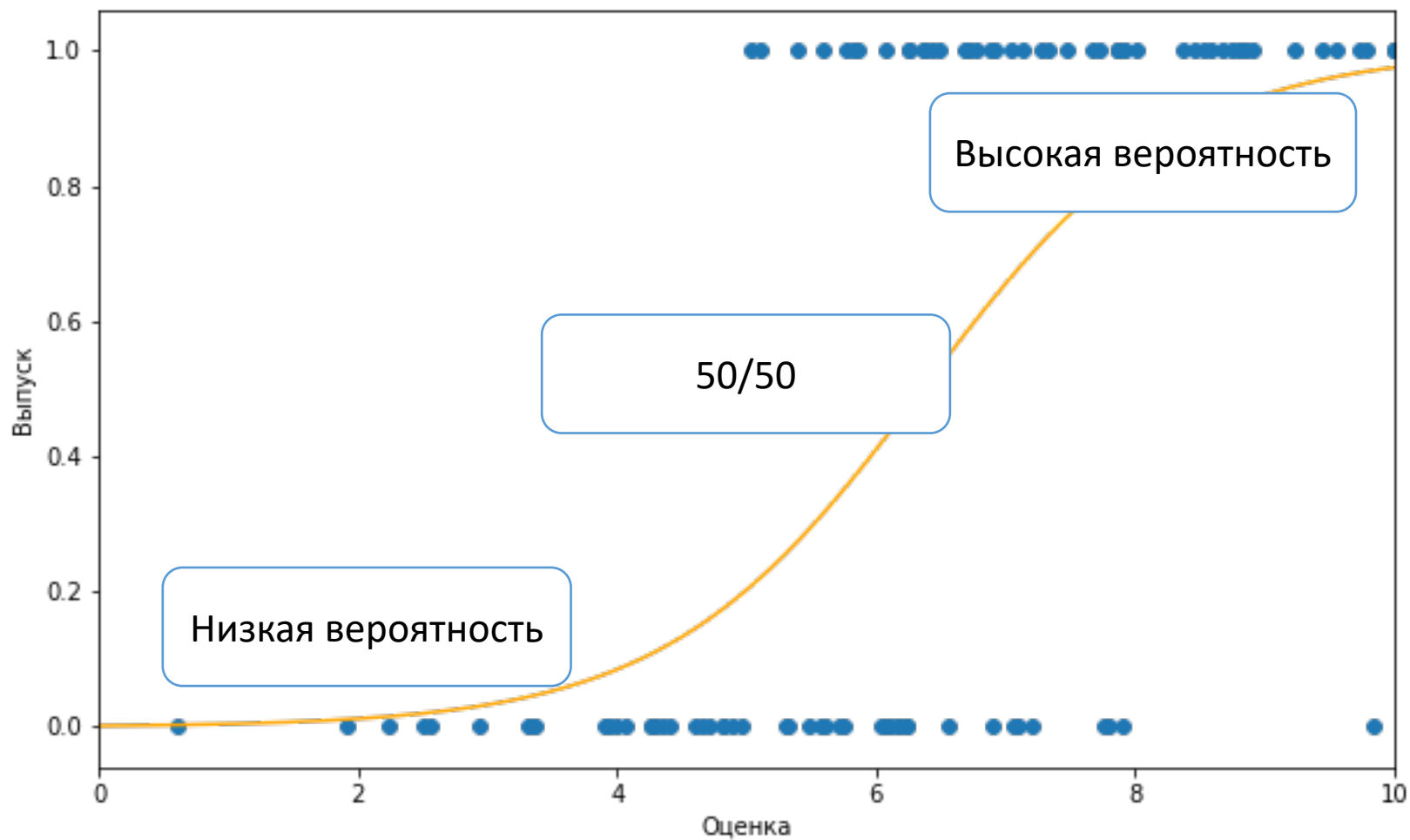
Сигмоида



Предсказание вероятностей



Предсказание вероятностей



Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?

Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 0$

Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 1$ или $\langle w, x_i \rangle \rightarrow +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 0$ или $\langle w, x_i \rangle \rightarrow -\infty$

Предсказание вероятностей

- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 1$ или $\langle w, x_i \rangle \rightarrow +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 0$ или $\langle w, x_i \rangle \rightarrow -\infty$
- То есть задача — сделать отступы на всех объектах максимальными

$$y_i \langle w, x_i \rangle \rightarrow \max_w$$

Предсказание вероятностей

- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \rightarrow 0$

$$-\sum_{i=1}^{\ell} \{ [y_i = 1] \sigma(\langle w, x_i \rangle) + [y_i = -1] (1 - \sigma(\langle w, x_i \rangle)) \} \rightarrow \min_w$$

Предсказание вероятностей

$$-\sum_{i=1}^{\ell} \{ [y_i = 1] \sigma(\langle w, x_i \rangle) + [y_i = -1] (1 - \sigma(\langle w, x_i \rangle)) \} \rightarrow \min_w$$

- Если $y_i = +1$ и $\sigma(\langle w, x_i \rangle) = 0$, то штраф равен 1
- Если $y_i = +1$, то заменить $\sigma(\langle w, x_i \rangle) = 1$ на $\sigma(\langle w, x_i \rangle) = 0.5$ так же плохо, как заменить $\sigma(\langle w, x_i \rangle) = 0.5$ на $\sigma(\langle w, x_i \rangle) = 0$
- Надо строже!

Предсказание вероятностей

$$-\sum_{i=1}^{\ell} \{ [y_i = 1] \log \sigma(\langle w, x_i \rangle) + [y_i = -1] \log(1 - \sigma(\langle w, x_i \rangle)) \} \rightarrow \min_w$$

- Если $y_i = +1$ и $\sigma(\langle w, x_i \rangle) = 0$, то штраф равен $-\log 0 = +\infty$
- Достаточно строго
- Функция потерь называется **log-loss**

$$L(y, z) = -[y = 1] \log z - [y = -1] \log(1 - z)$$

Логистическая регрессия

$$\begin{aligned} & - \sum_{i=1}^{\ell} \{ [y_i = 1] \log \sigma(\langle w, x_i \rangle) + [y_i = -1] \log(1 - \sigma(\langle w, x_i \rangle)) \} = \\ & - \sum_{i=1}^{\ell} \left\{ [y_i = 1] \log \frac{1}{1 + \exp(-\langle w, x \rangle)} + [y_i = -1] \log \left(1 - \frac{1}{1 + \exp(-\langle w, x \rangle)} \right) \right\} = \\ & - \sum_{i=1}^{\ell} \left\{ [y_i = 1] \log \frac{1}{1 + \exp(-\langle w, x \rangle)} + [y_i = -1] \log \left(\frac{1}{1 + \exp(\langle w, x \rangle)} \right) \right\} = \\ & \sum_{i=1}^{\ell} \{ [y_i = 1] \log(1 + \exp(-\langle w, x \rangle)) + [y_i = -1] \log(1 + \exp(\langle w, x \rangle)) \} = \\ & \sum_{i=1}^{\ell} \log(1 + \exp(-y_i \langle w, x_i \rangle)) \end{aligned}$$