

# Введение в анализ данных

## Лекция 7

Метрики качества регрессии и классификации.

Многоклассовая классификация.

Евгений Соколов

[esokolov@hse.ru](mailto:esokolov@hse.ru)

НИУ ВШЭ, 2017

# 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

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

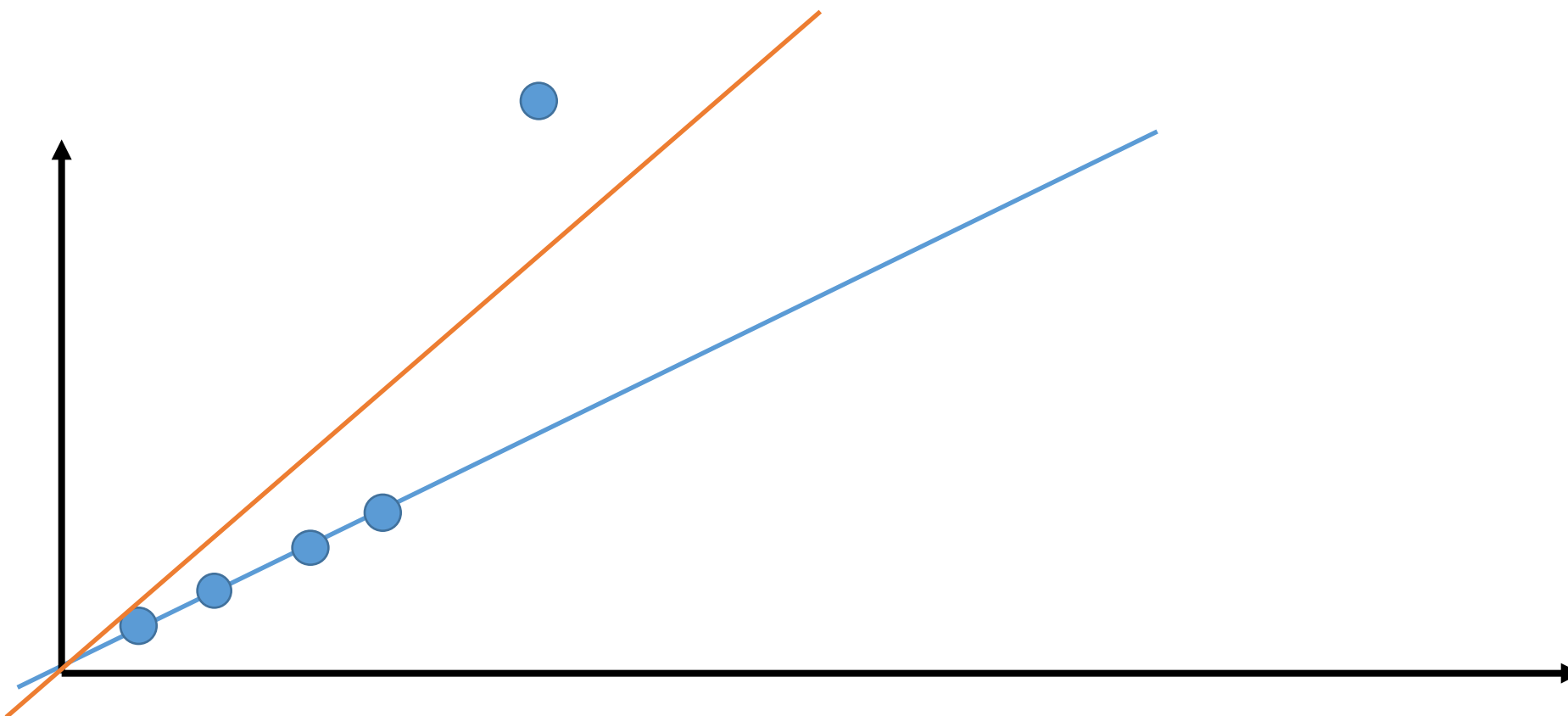
- Легко минимизировать
- Сильно штрафует за большие ошибки

# Средняя абсолютная ошибка

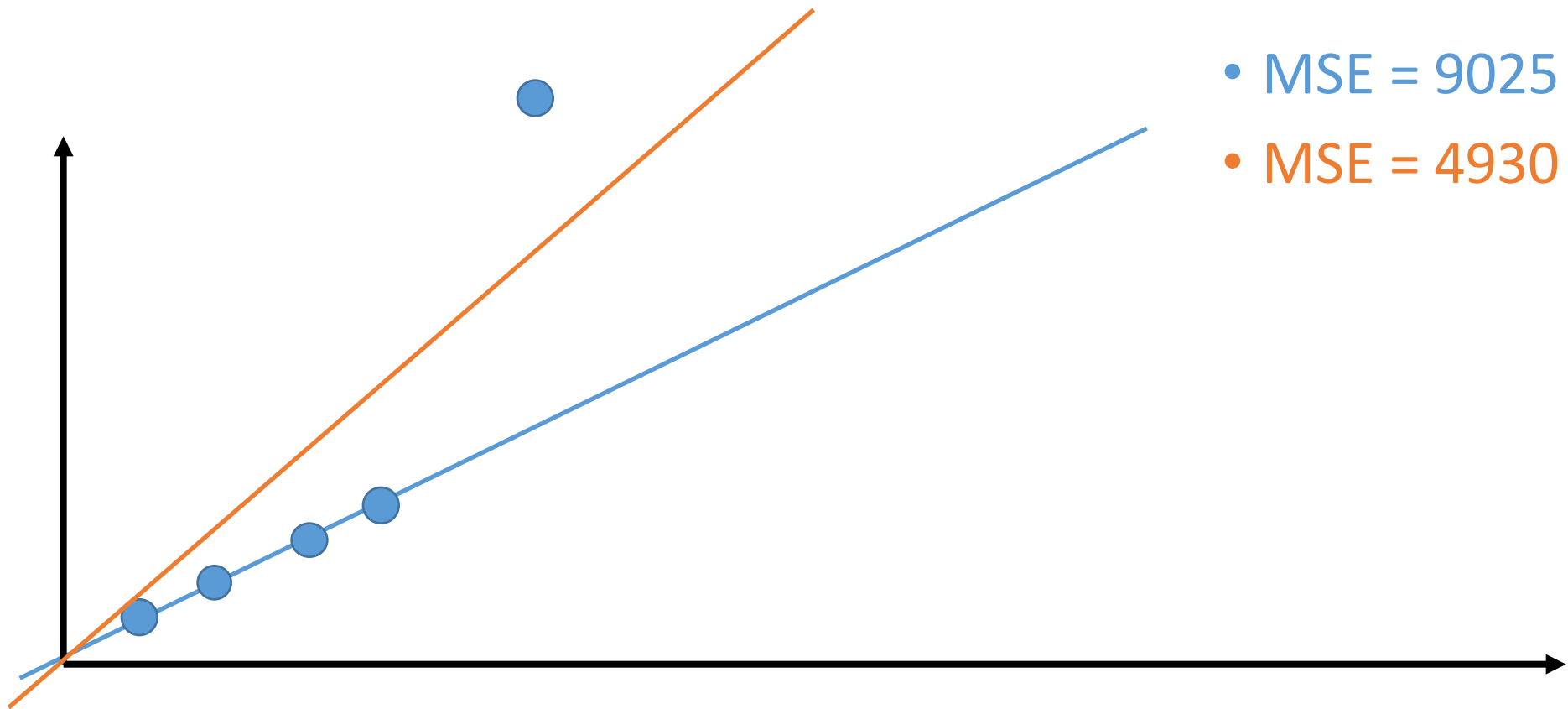
$$\text{MAE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

- Сложнее минимизировать
- Выше устойчивость к выбросам

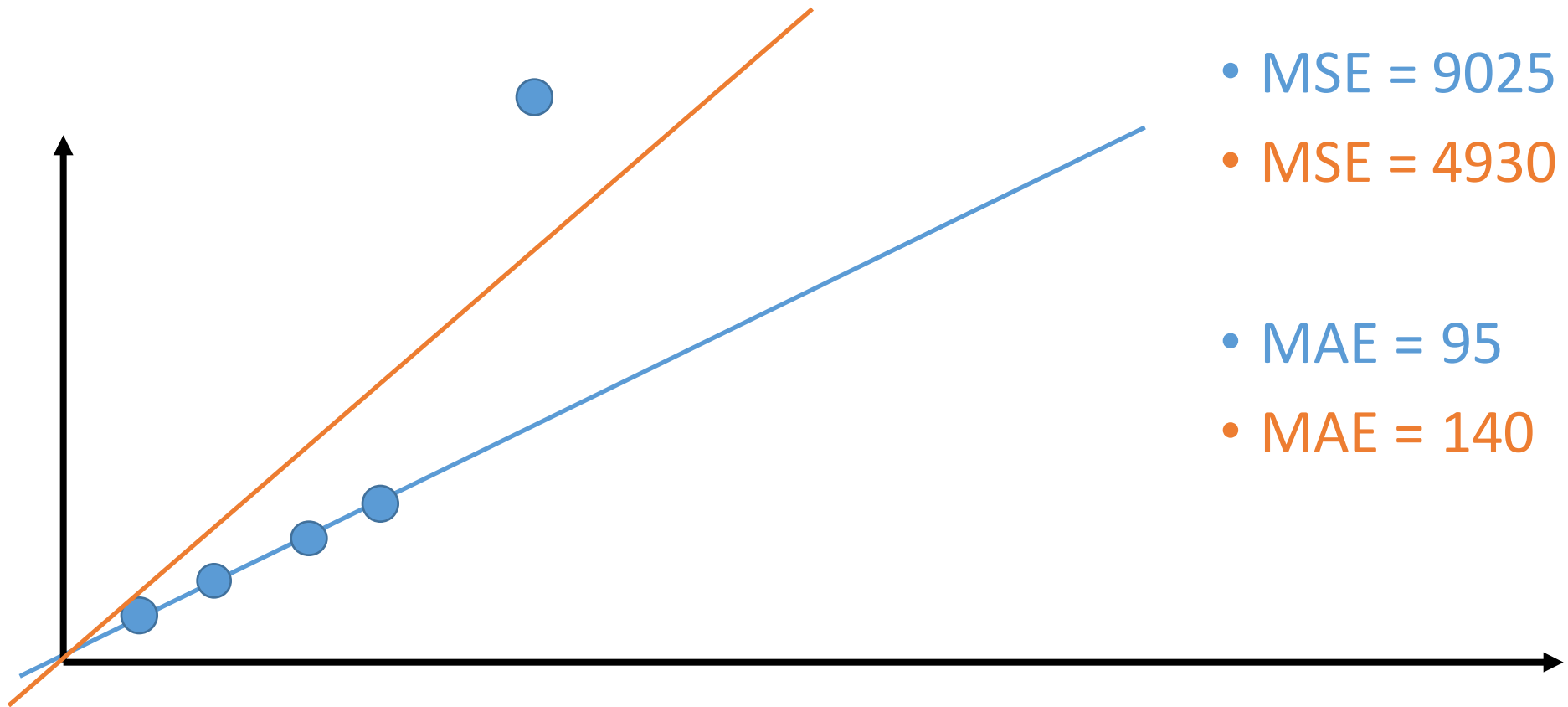
# Средняя абсолютная ошибка



# Средняя абсолютная ошибка



# Средняя абсолютная ошибка





# Устойчивые оценки

- Оценка среднего значения — матожидание
- Оценка разброса — дисперсия

# Математическое ожидание

- Характеризует среднее значение случайной величины

$$\mathbb{E}\xi = \begin{cases} \sum_{i=1}^n x_i p_i, & \text{для дискретных величин} \\ \int_{-\infty}^{+\infty} x p(x) dx, & \text{для непрерывных величин} \end{cases}$$

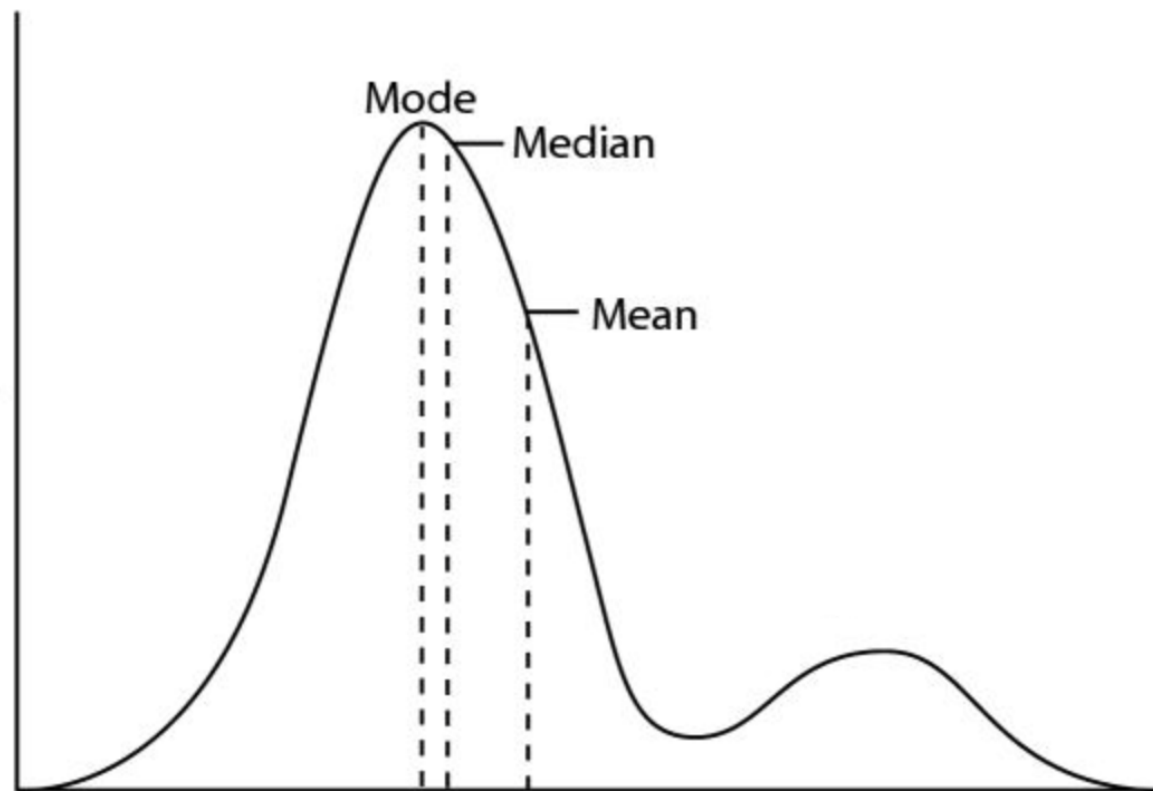
# Медиана

- Такое число  $m$ , что попасть левее и правее — равновероятно
- $P(\xi \leq m) \geq 0.5$  и  $P(\xi \geq m) \geq 0.5$

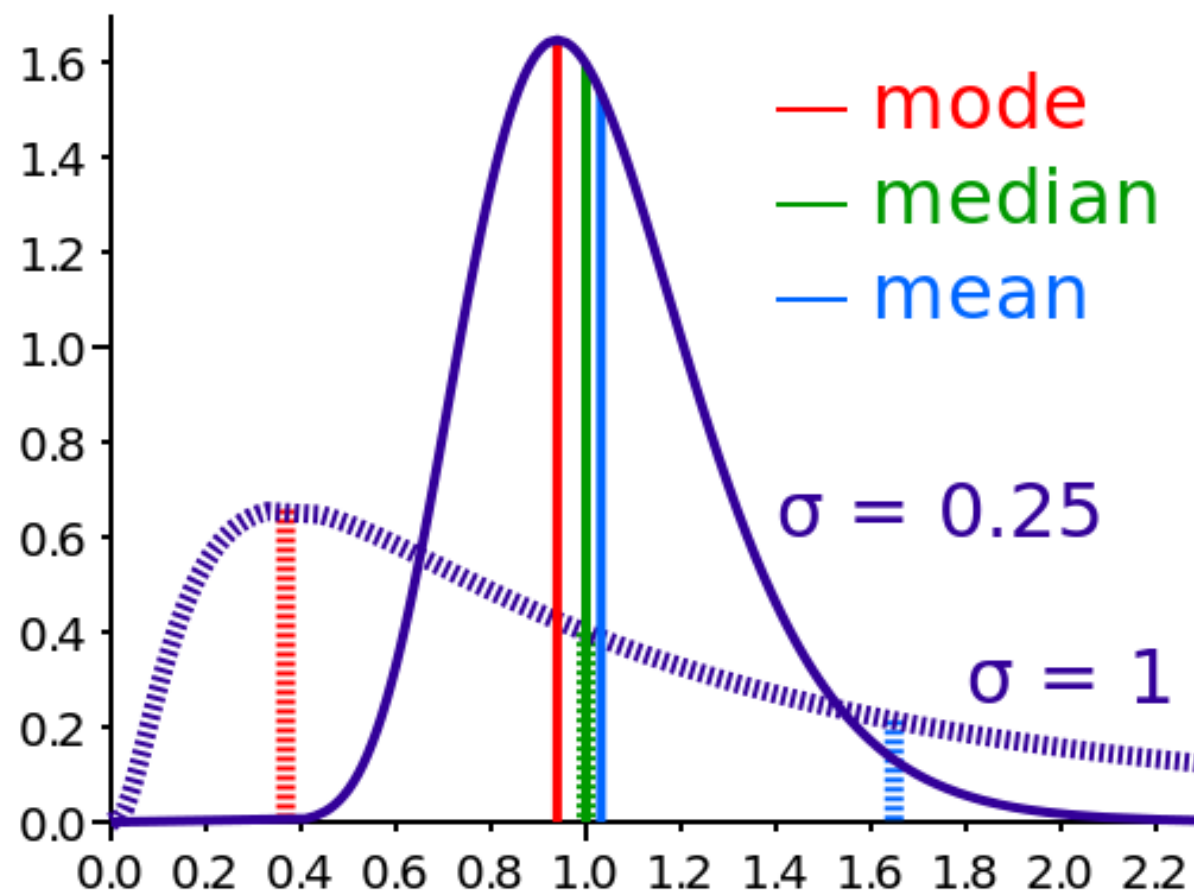
# Мода

- Для дискретных величин: точка с максимальной вероятностью
- Для непрерывных величин: точка максимума плотности

# Средняя величина



# Средняя величина



# В чем разница?

- Опросили 100 человек
- 99 имеют доход 10.000 рублей
- 1 имеет доход 1.000.000 рублей
- Среднее:  $\frac{99*10000+1000000}{100} = 19900$
- Медиана: 10000
- Мода: 10000



\$45,000



\$15,000



\$10,000

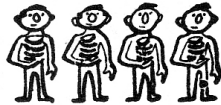


← **ARITHMETICAL AVERAGE**

\$5,700



\$5,000



\$3,700



← **MEDIAN** (the one in the middle)  
12 above him, 12 below

\$3,000



\$2,000

← **MODE**  
(occurs most frequently)

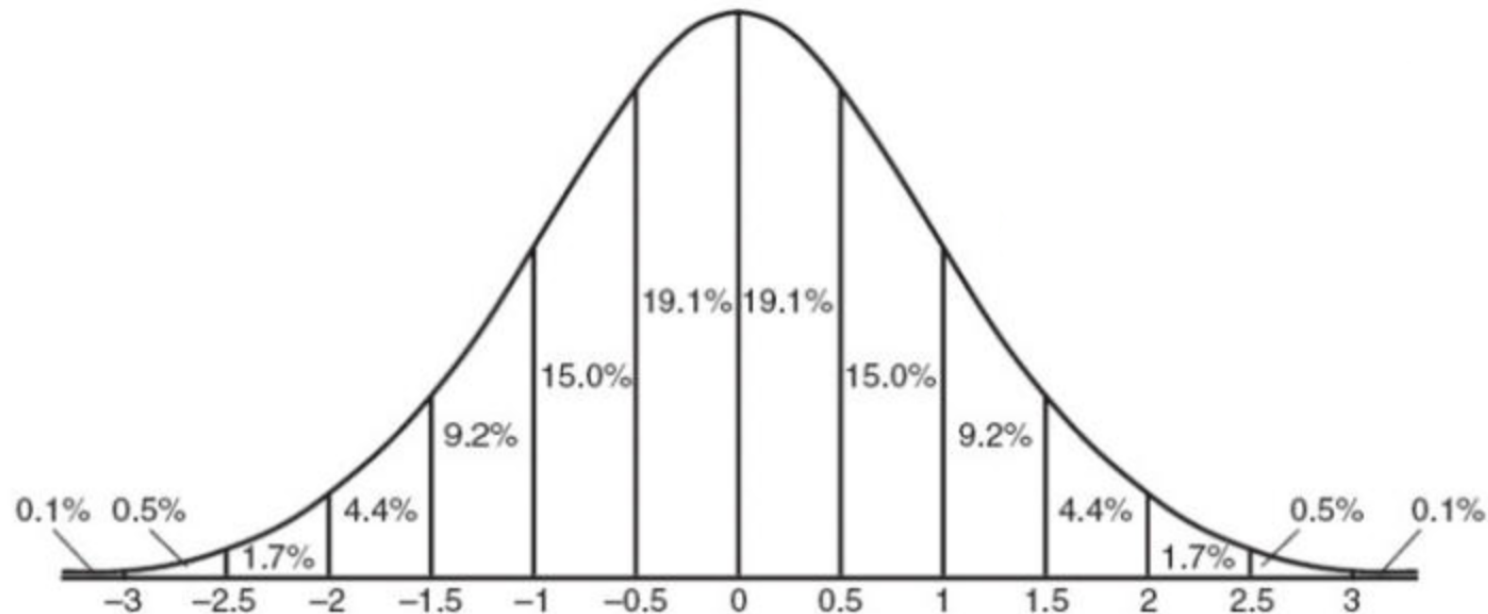


# Дисперсия

- Опросили 100 человек
- 99 имеют доход 10.000 рублей
- 1 имеет доход 1.000.000 рублей
- Дисперсия: 9702990000
- Стандартное отклонение (корень из дисперсии):  $\sim 98503$
- Что-нибудь более устойчивое?

# Квантиль

- $Q_p$  —  $p$ -квантиль
- Такое число  $t$ , что вероятность попасть левее равна  $p$
- Медиана — 0.5-квантиль



# Квантиль

- $Q_{0.25}, Q_{0.75}$  — квартили
- $Q_{0.01}, \dots, Q_{0.99}$  — перцентили

# Интерквартильный размах

- Устойчивая к выбросам мера разброса:

$$IQR = Q_{0.75} - Q_{0.25}$$

- В нашем примере:  $IQR = 0$

# 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?

# Коэффициент детерминации

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

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

# Коэффициент детерминации

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

- $0 \leq R^2 \leq 1$  (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 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

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

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

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

# 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?

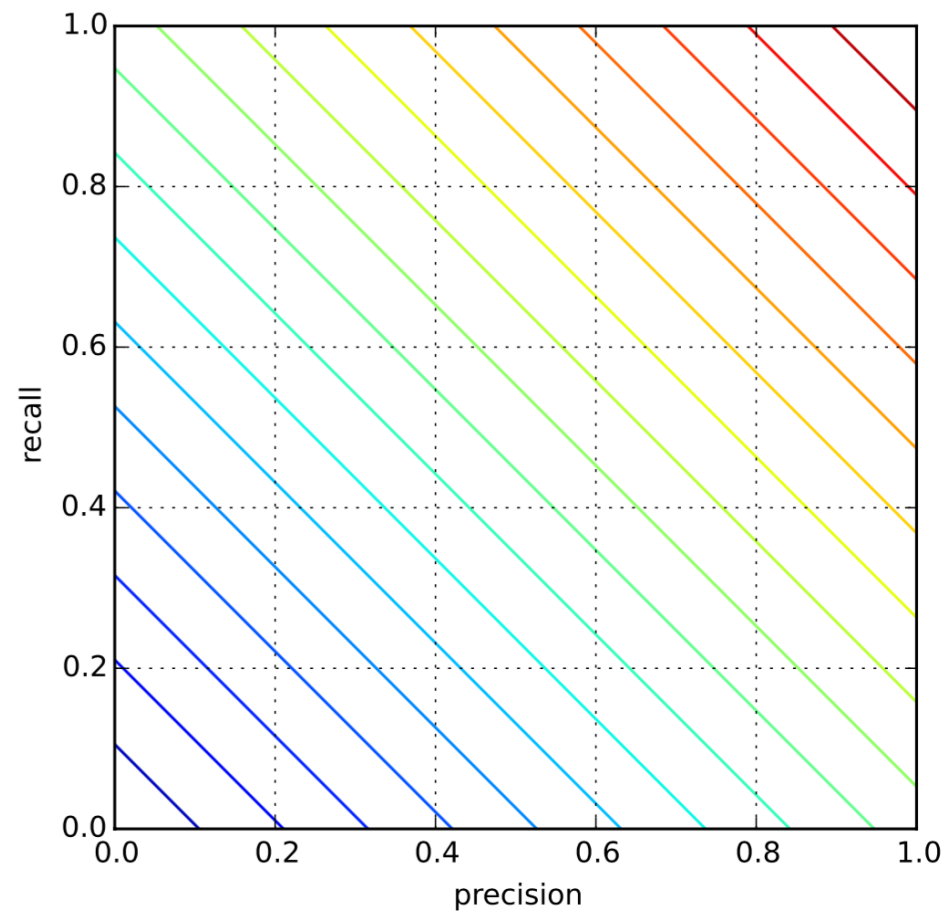
$$\text{recall}(a, X) = \frac{TP}{TP + FN}$$

# 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

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

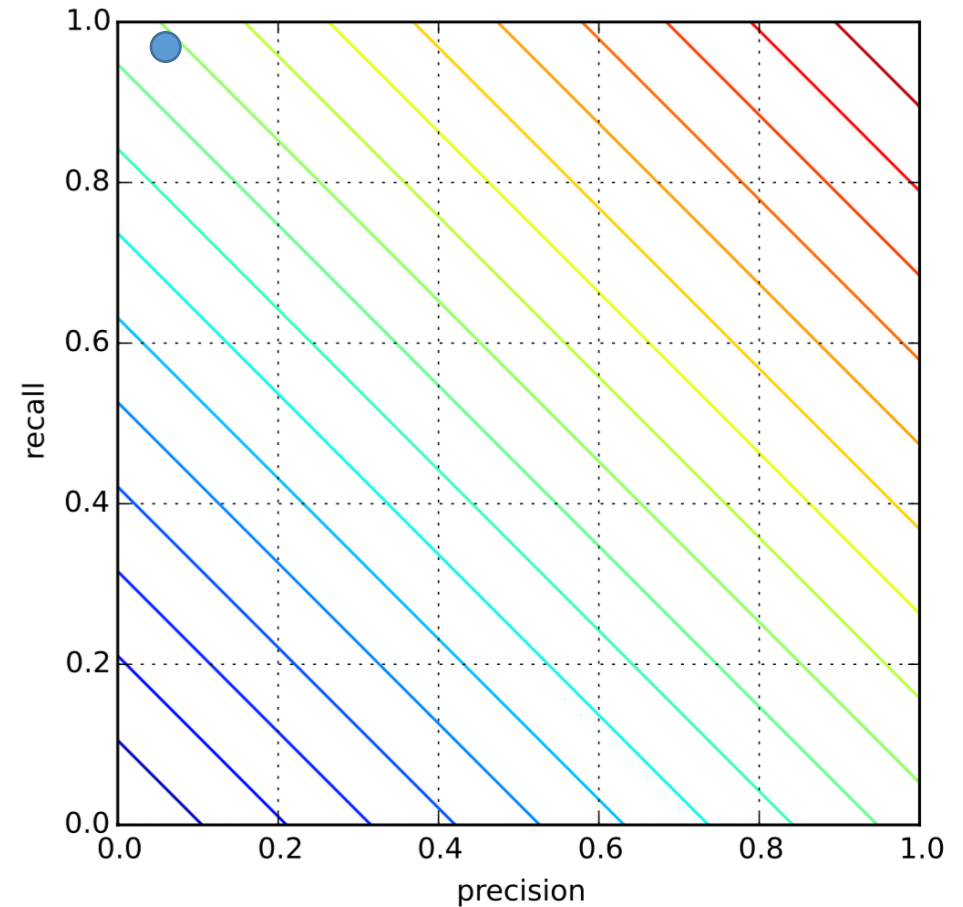




# Арифметическое среднее

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

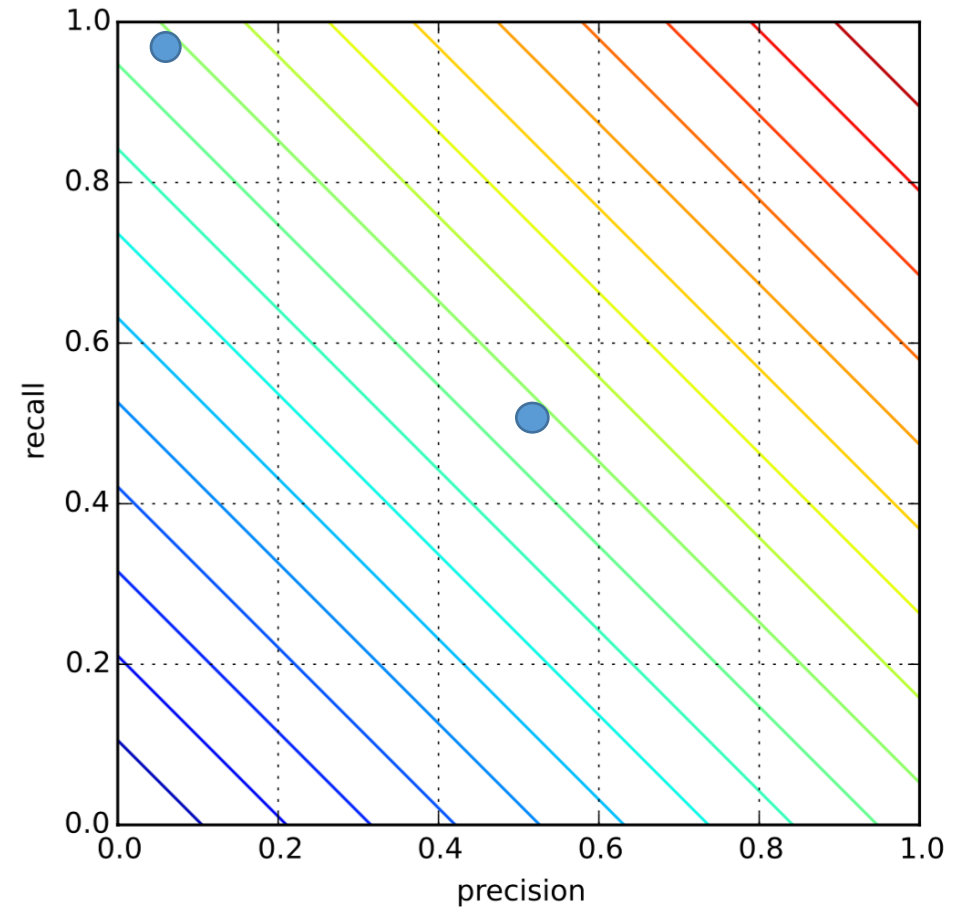
- precision = 0.1
- recall = 1
- $A = 0.55$
- Плохой алгоритм



# Арифметическое среднее

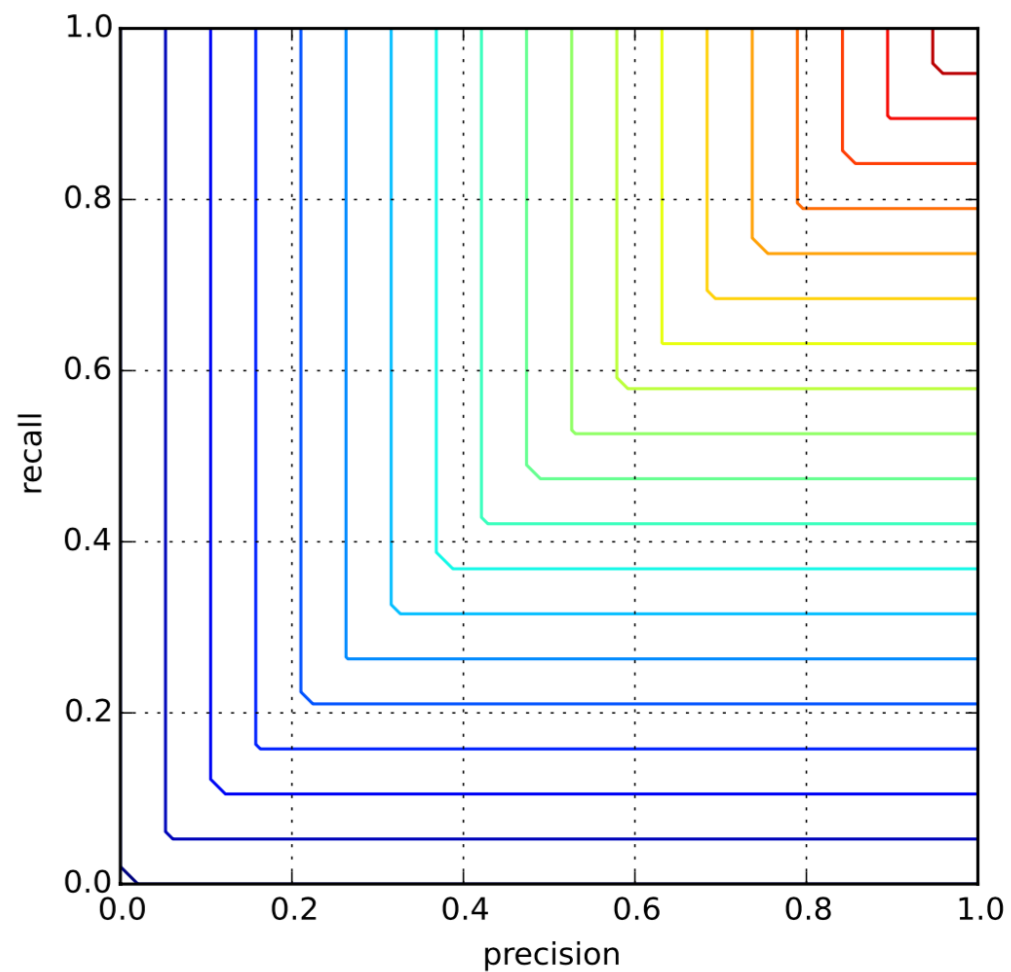
$$A = \frac{1}{2} (\text{precision} + \text{recall})$$

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



# Минимум

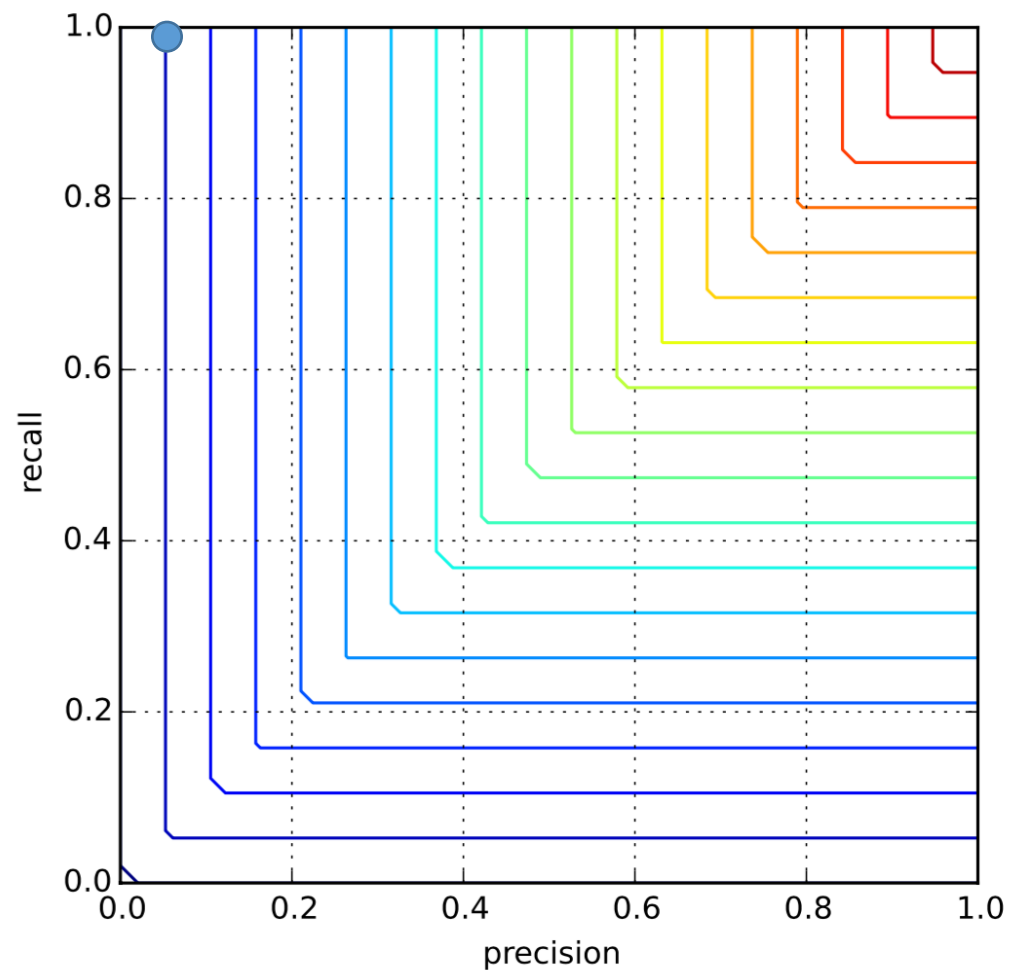
$$M = \min(\text{precision}, \text{recall})$$



# Минимум

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

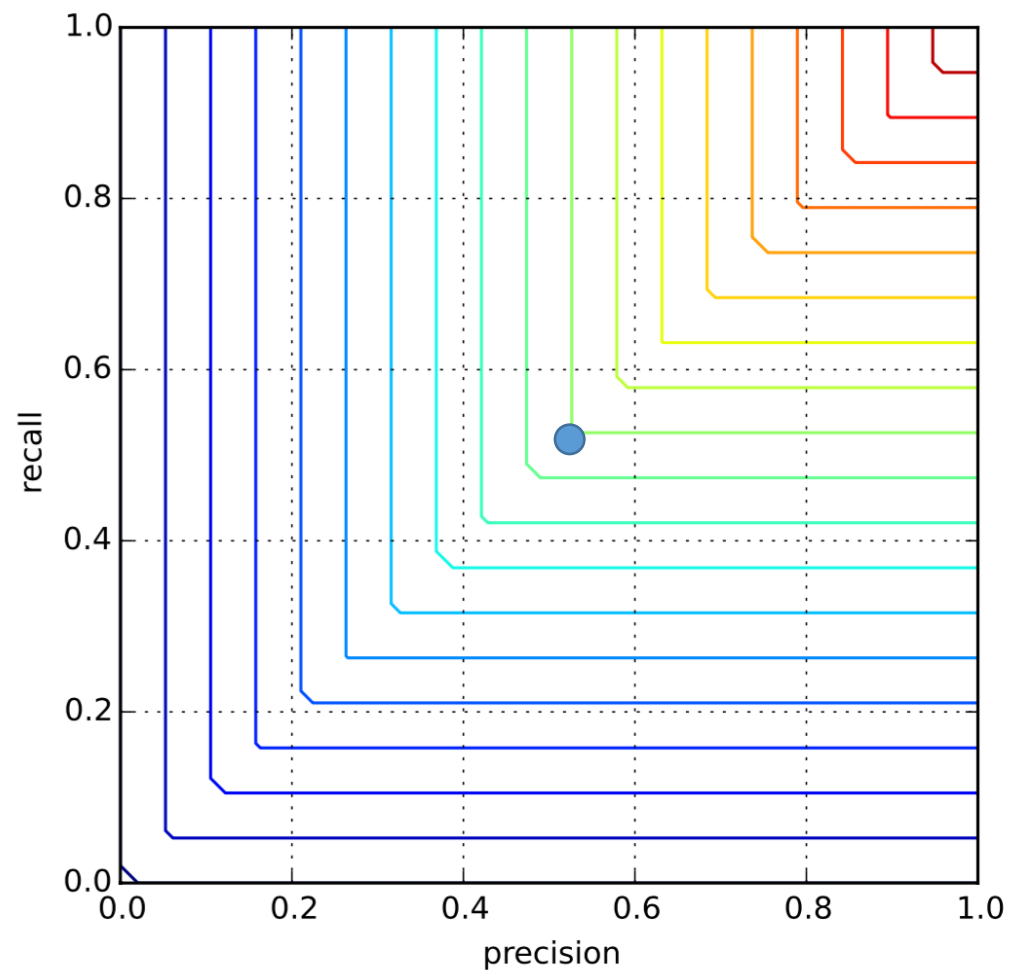
- precision = 0.05
- recall = 1
- $M = 0.05$



# Минимум

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

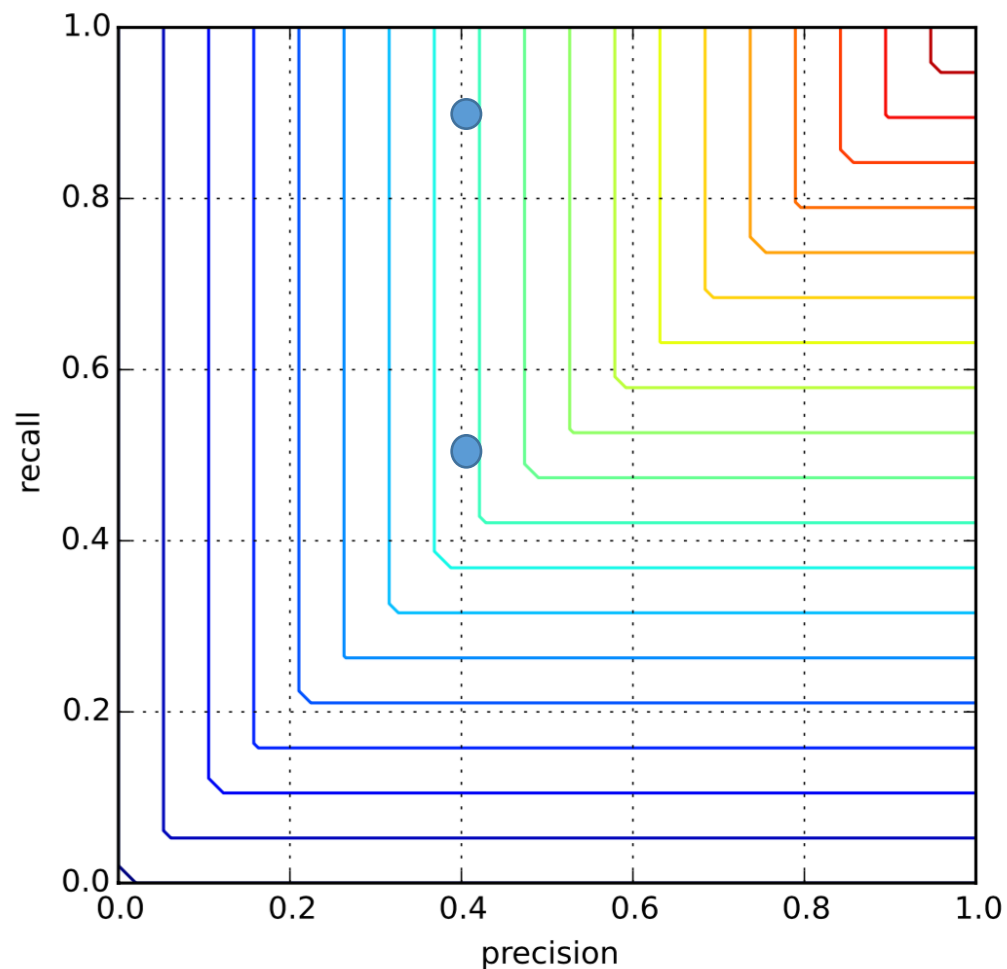
- precision = 0.55
- recall = 0.55
- $M = 0.55$



# Минимум

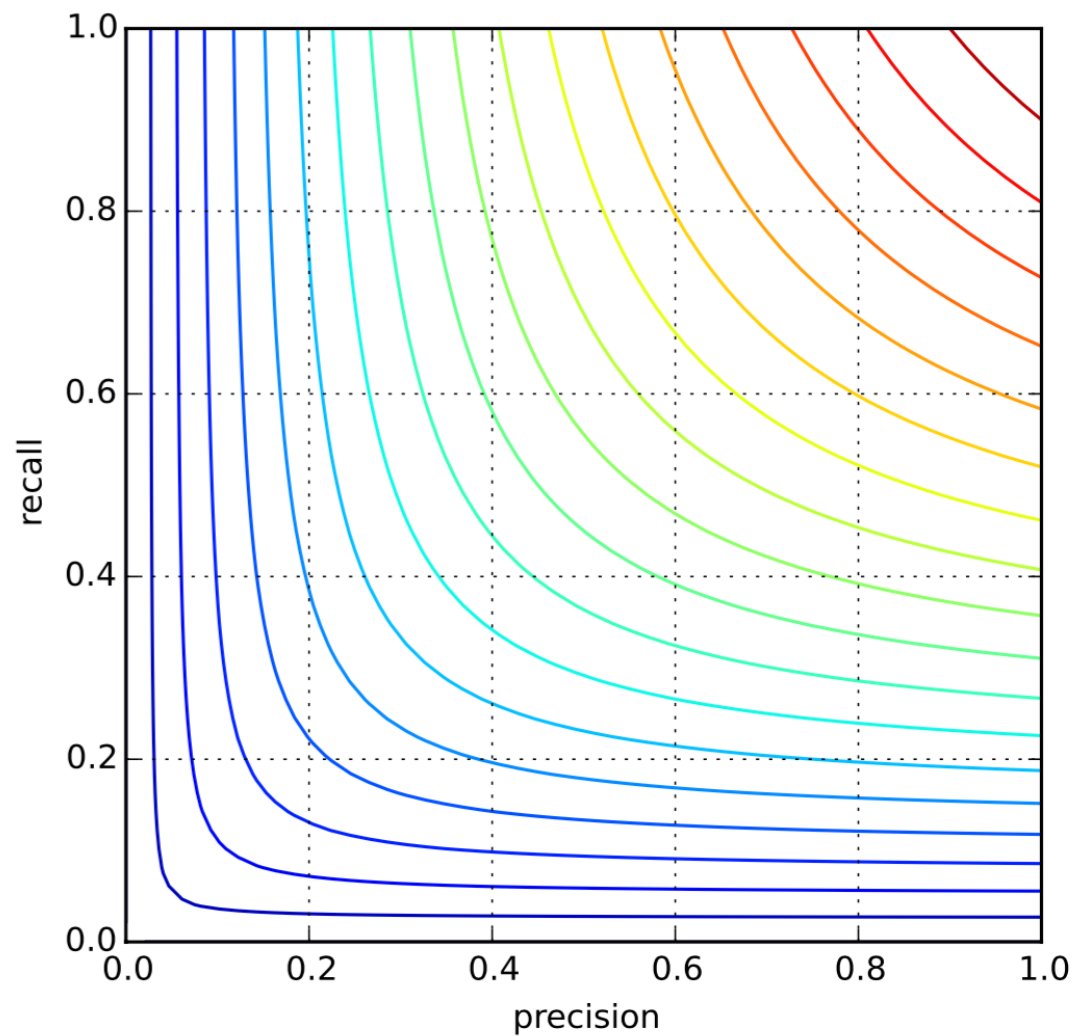
$$M = \min(\text{precision}, \text{recall})$$

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



# F-meap

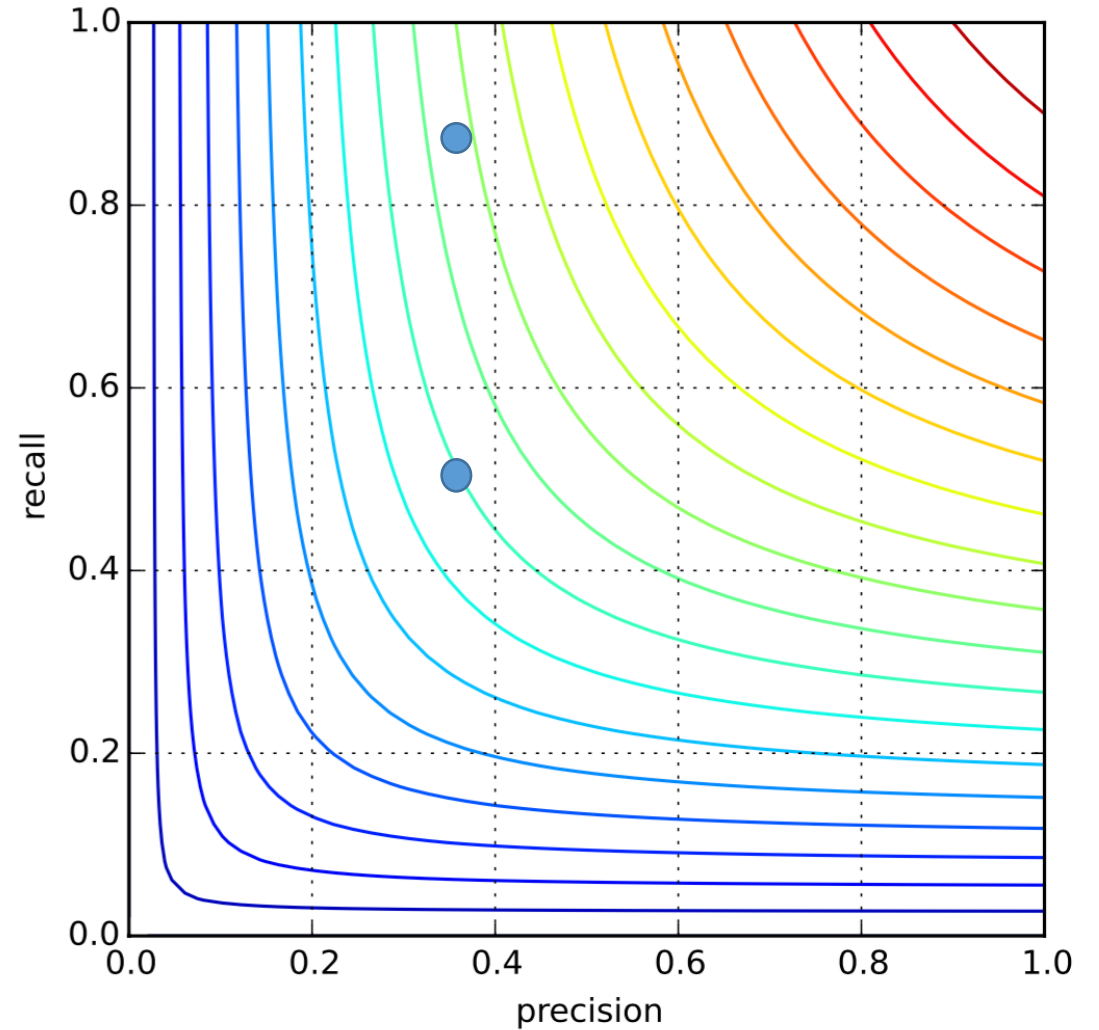
$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



# 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
- $M = 0.55$





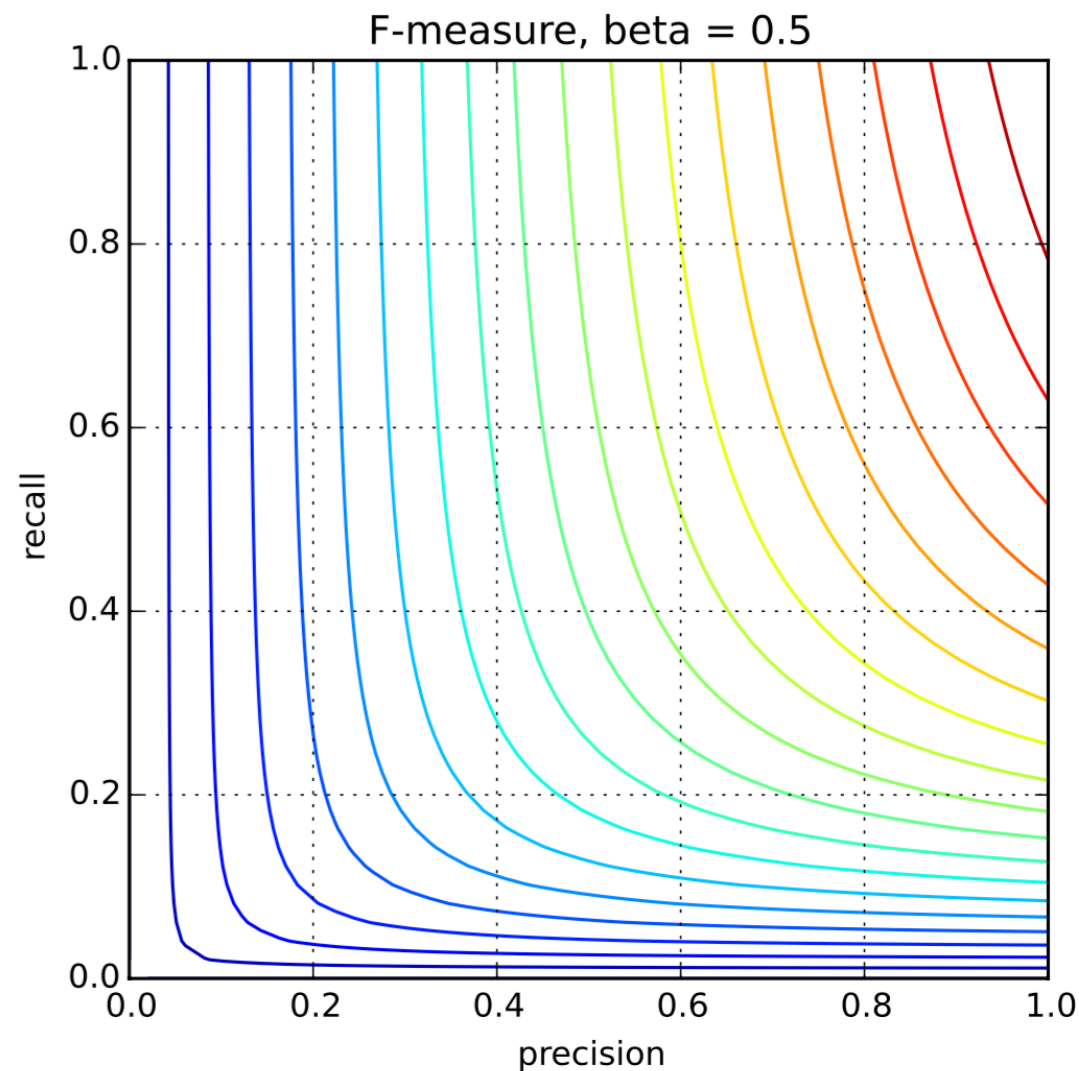
# F-measure

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

# F-мера

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

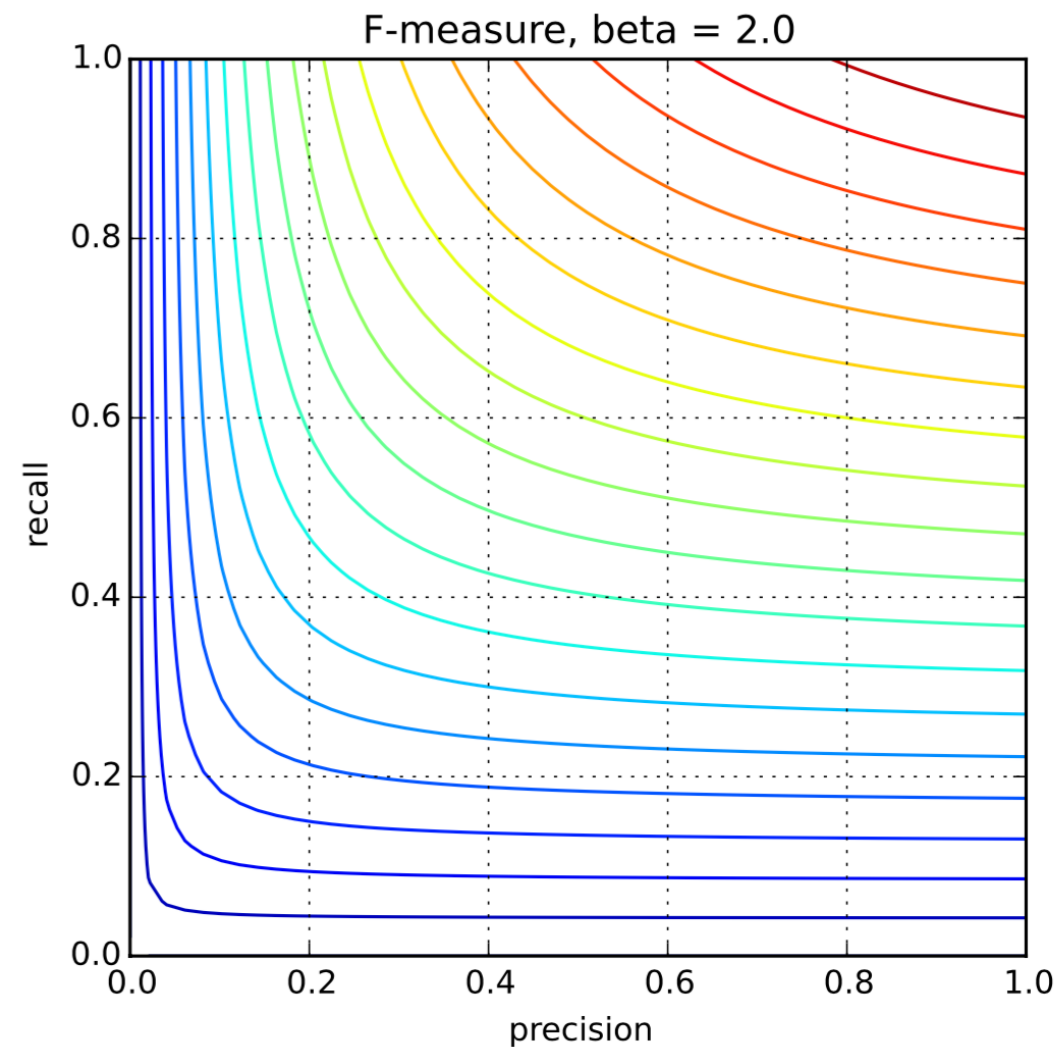
- $\beta = 0.5$
- Важнее полнота



# 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  $t = 0$

# 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



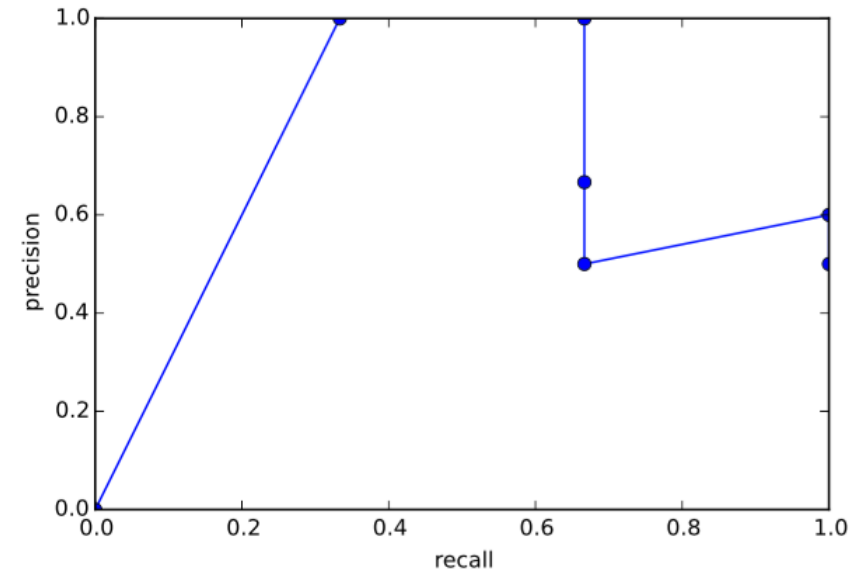
-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

# Оценка принадлежности

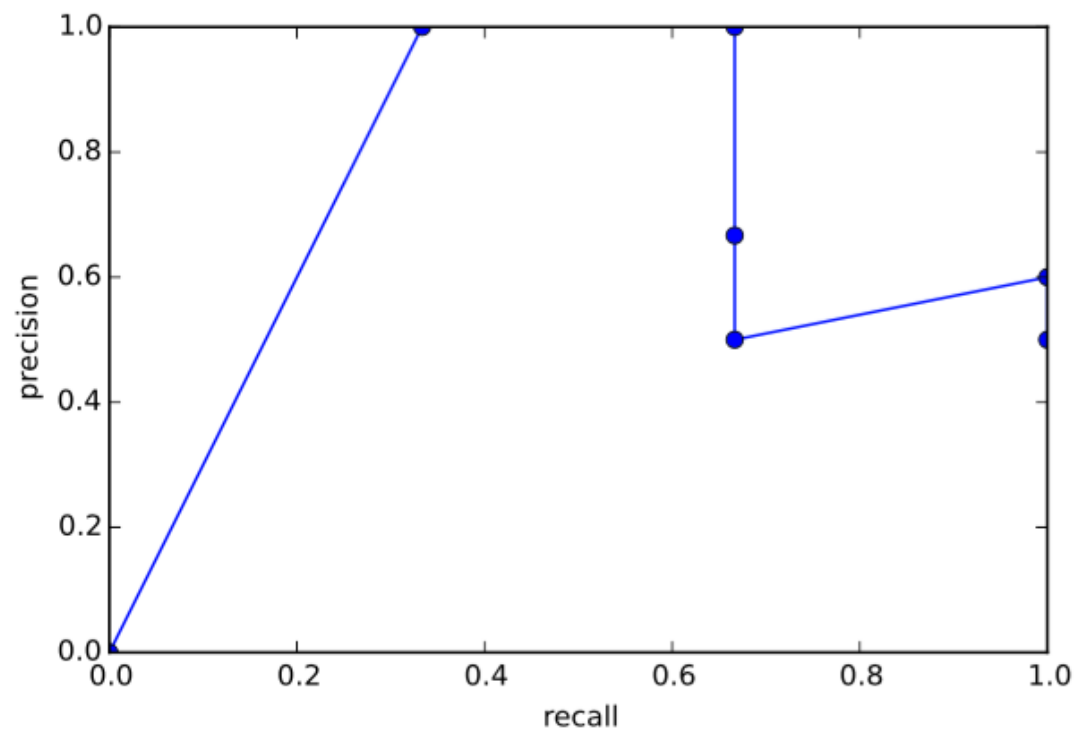
- Пример: кредитный скоринг
- $b(x)$  — оценка вероятности возврата кредита
- $a(x) = [b(x) > 0.5]$
- precision = 0.1, recall = 0.7
- В чем дело — в пороге или в алгоритме?

# PR-кривая

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

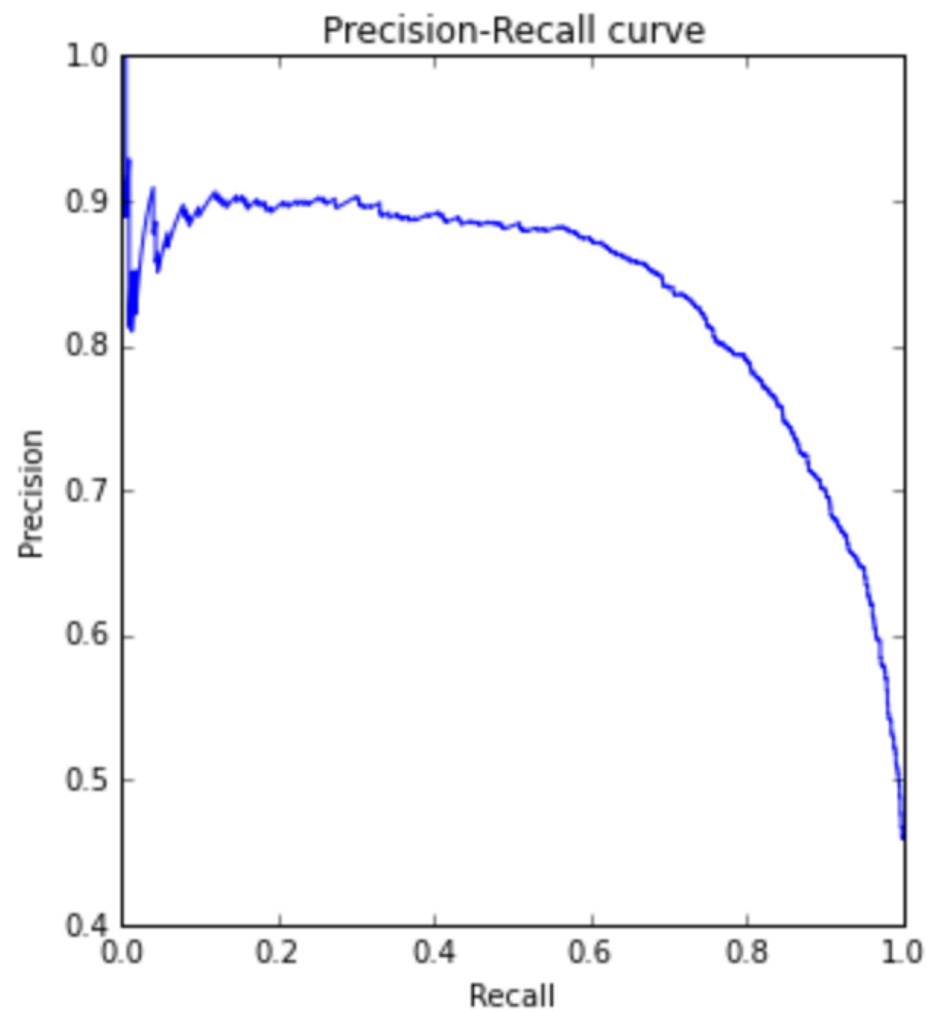


# PR-кривая



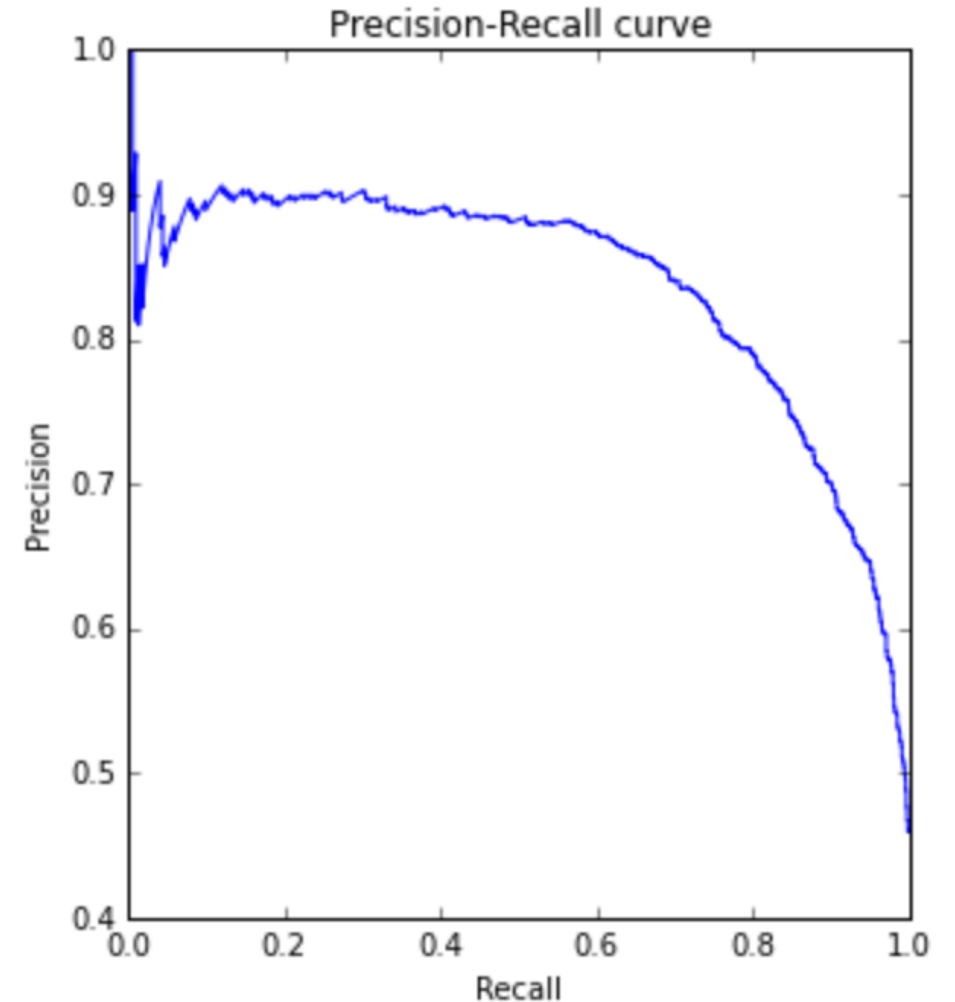
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
$y$	0	1	0	0	1	1

# PR-кривая в реальности



# PR-кривая

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





# ROC-кривая

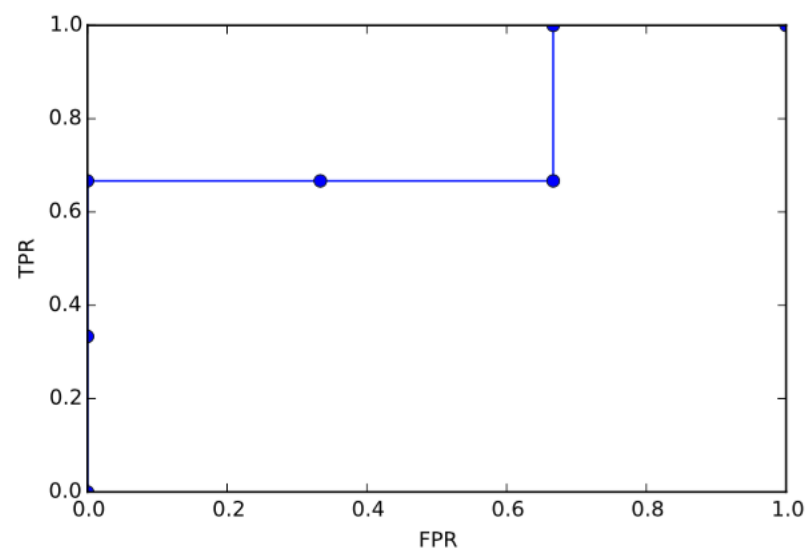
- Receiver Operating Characteristic

- 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

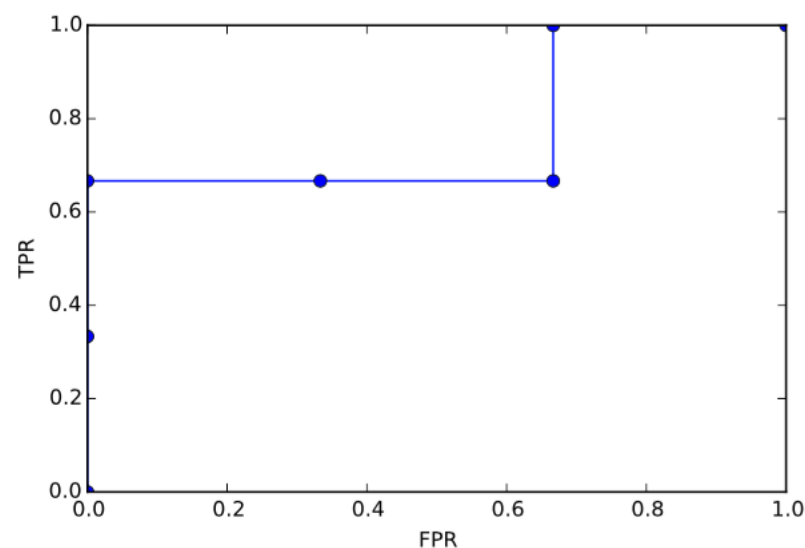
$$FPR = \frac{FP}{FP + TN}$$

Число  
отрицательных  
объектов

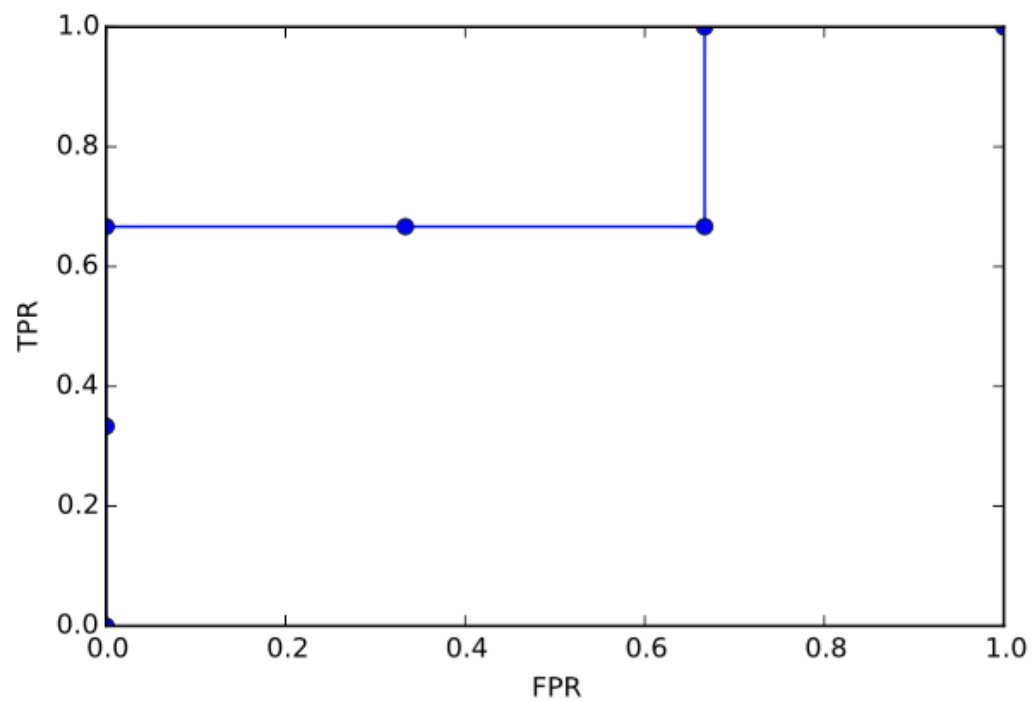
- Ось Y — True Positive Rate

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

Число  
положительных  
объектов

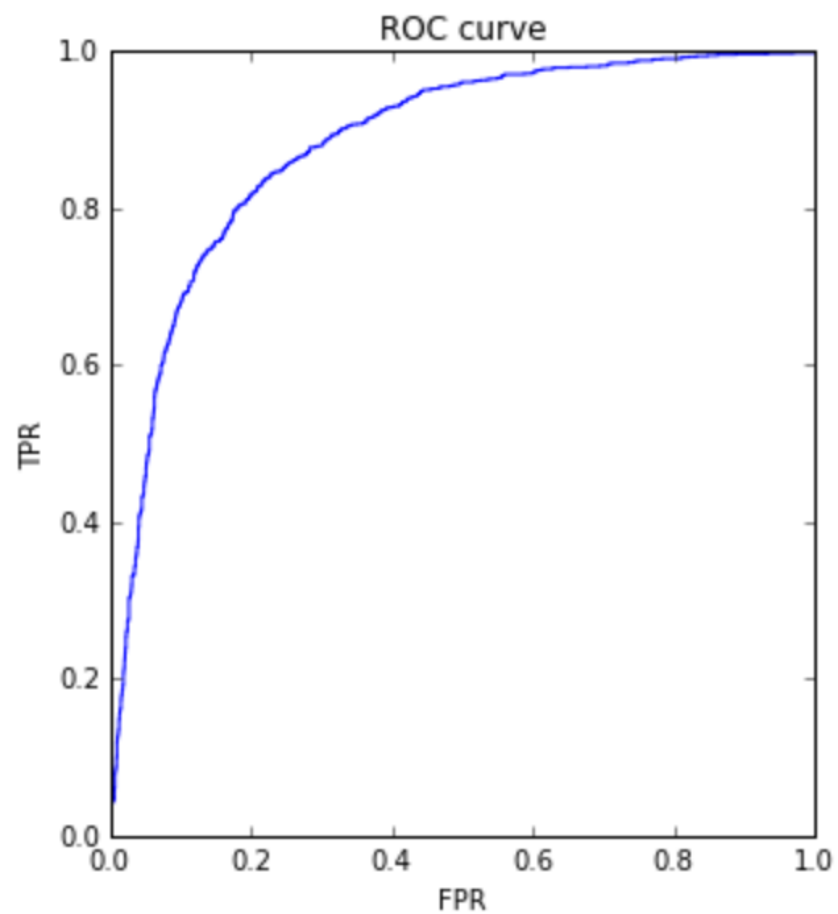


# ROC-кривая



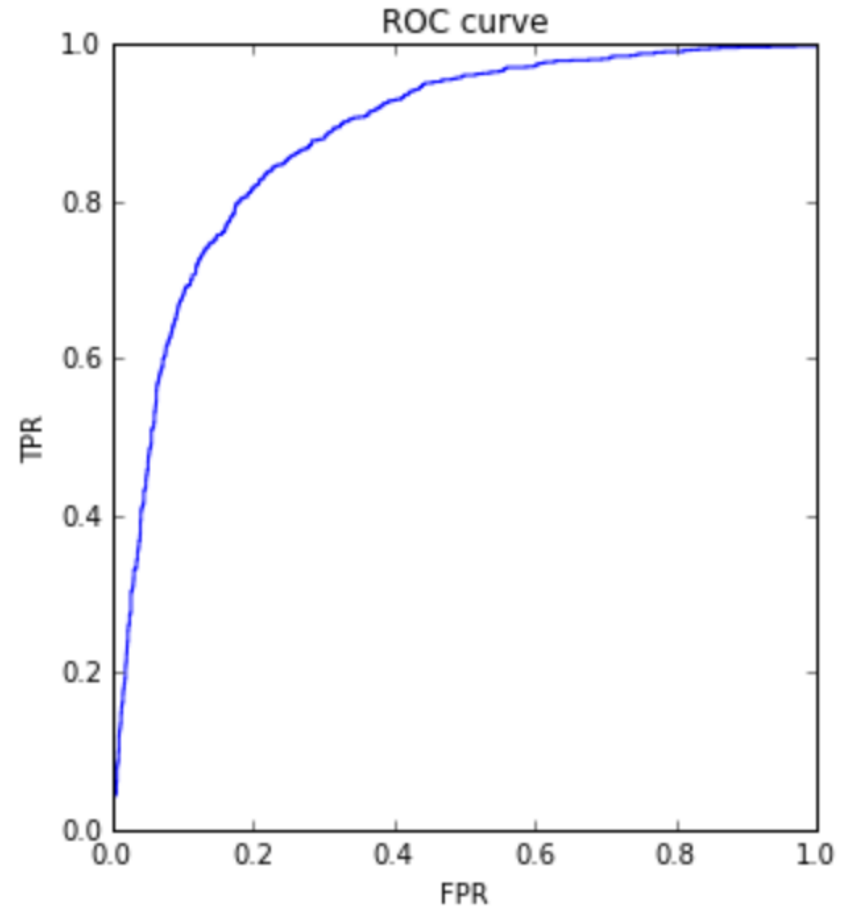
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
$y$	0	1	0	0	1	1

# 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:  $AUC-ROC \approx 0.5$

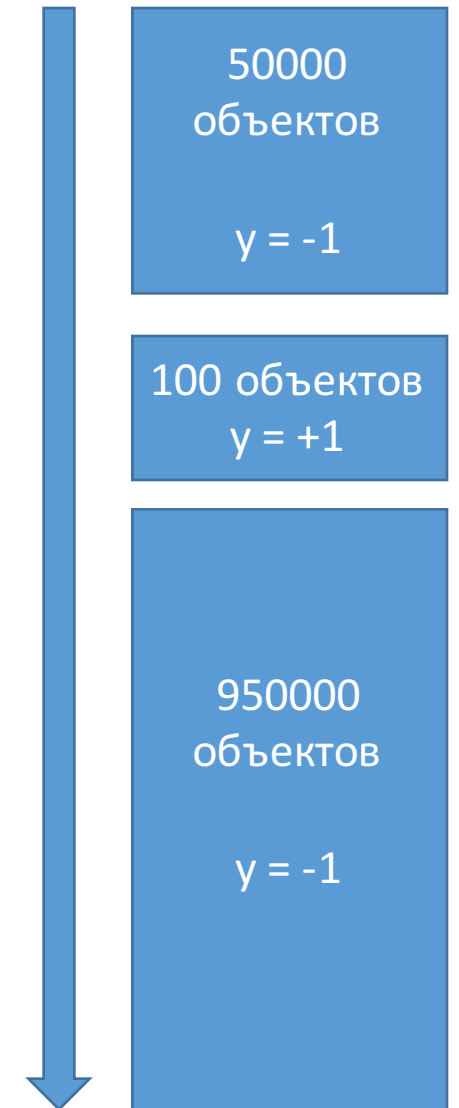
# AUC-PRC

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

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

# Пример

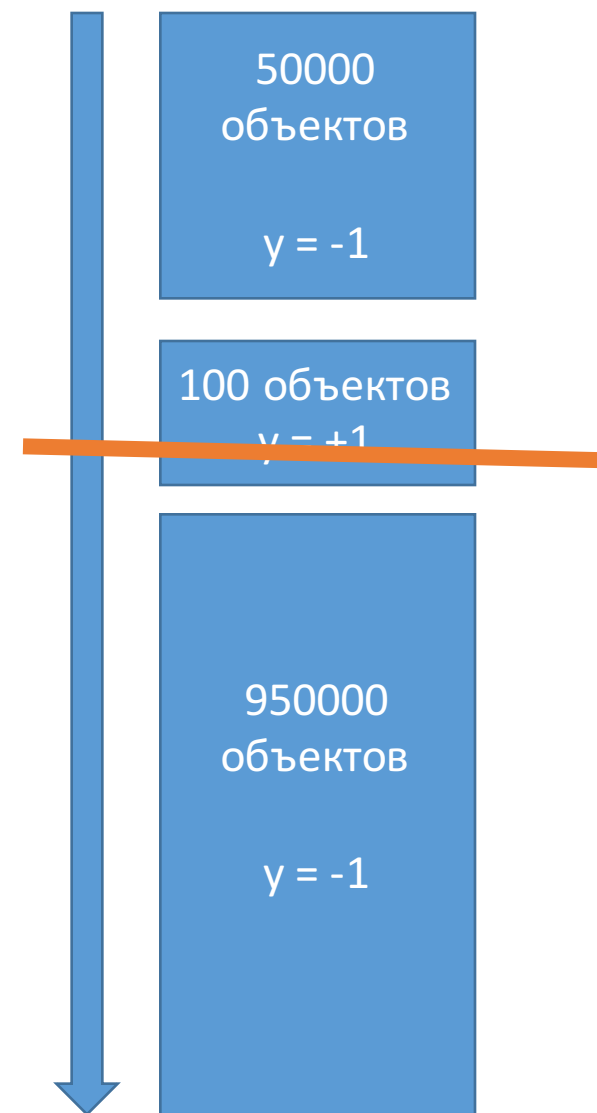
- AUC-ROC = 0.95
- AUC-PRC = 0.001





# 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



# Параметры и гиперпараметры

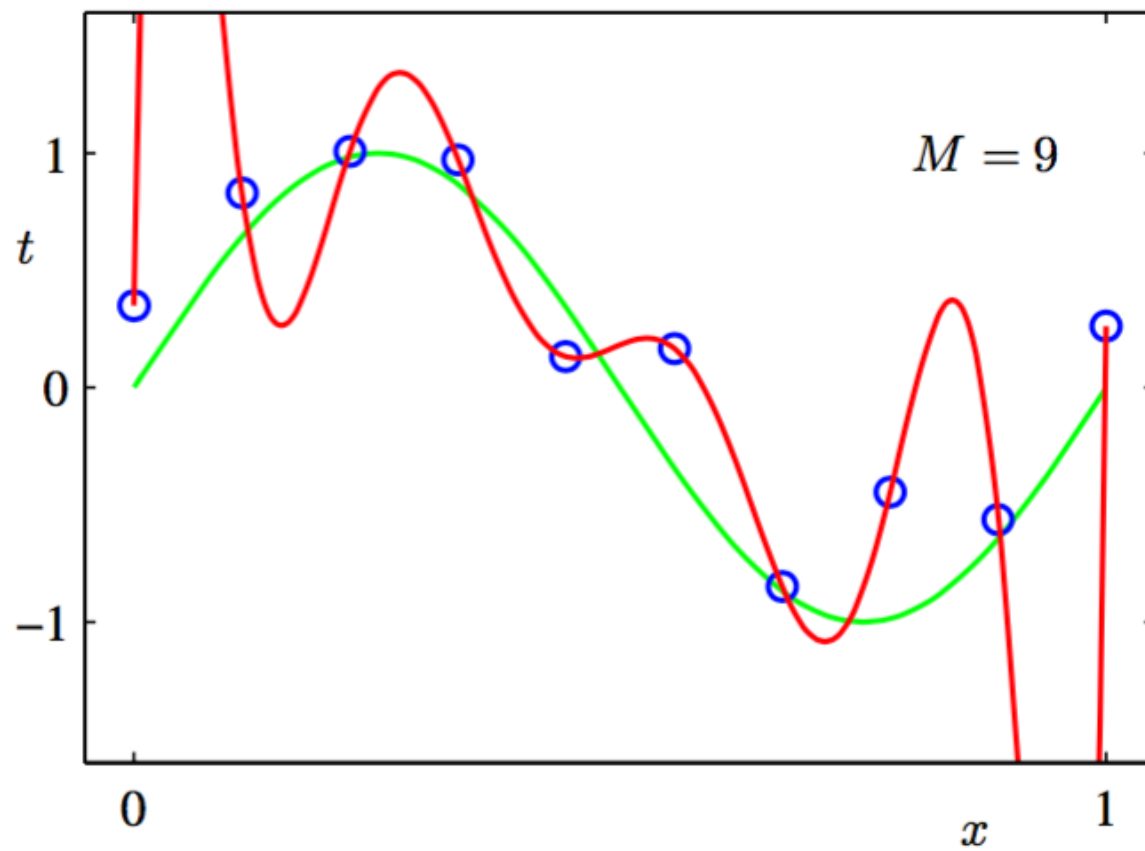
# Простой пример

- Максимизируем удовлетворённость студентов
- Обучающая выборка — время до сессии
- Контрольная выборка — сессия
- Параметр — продолжительность лекции
- Гиперпараметр — минимальная продолжительность лекции

# Простой пример

- Максимизируем удовлетворённость студентов
  - Обучающая выборка — время до сессии
  - Контрольная выборка — сессия
  - Параметр — продолжительность лекции
  - Гиперпараметр — минимальная продолжительность лекции
- 
- Максимальная удовлетворённость на обучении — если не ограничивать продолжительность
  - Но оценки во время сессии будут ужасными

# Переобучение



# Регуляризация

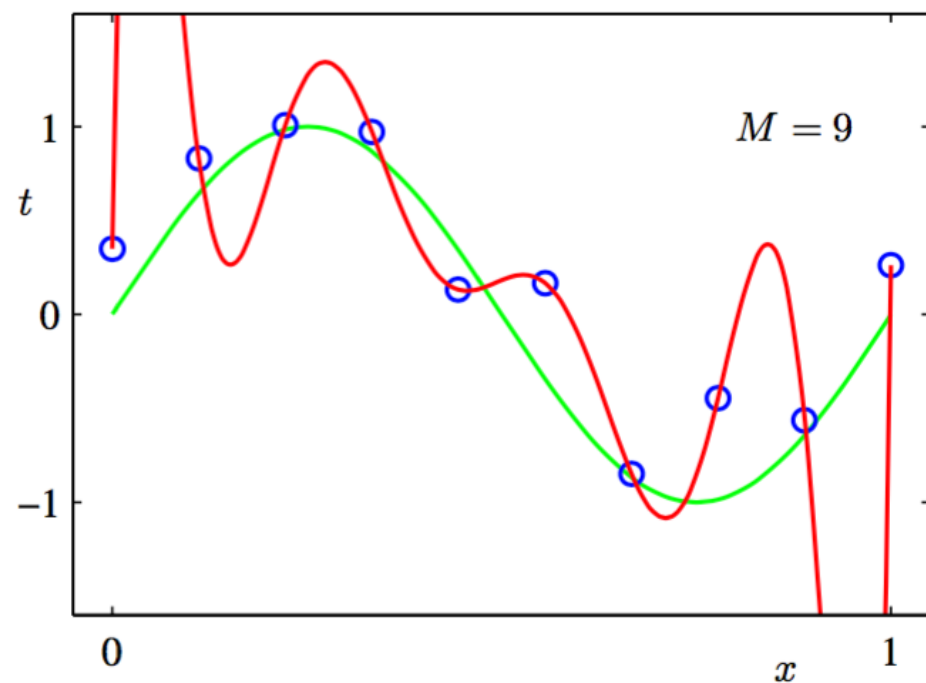
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$$

# Гиперпараметры

- Параметры модели — веса  $w$ 
  - Позволяют подогнать модель под обучающую выборку
  - Настраиваются по обучающей выборке
- Гиперпараметр модели — коэффициент регуляризации  $\lambda$ 
  - Определяют сложность модели
  - Лучшее качество на обучении достигается при  $\lambda = 0$
  - Необходимо настраивать по другим данным

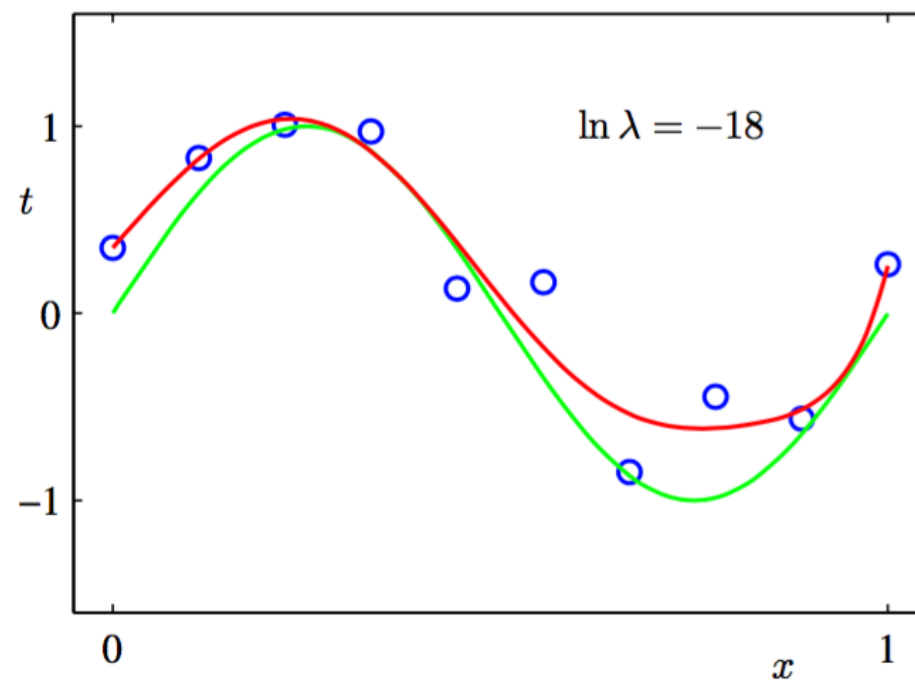
# Гиперпараметры

Без регуляризации



Высокое качество на обучении

С регуляризацией

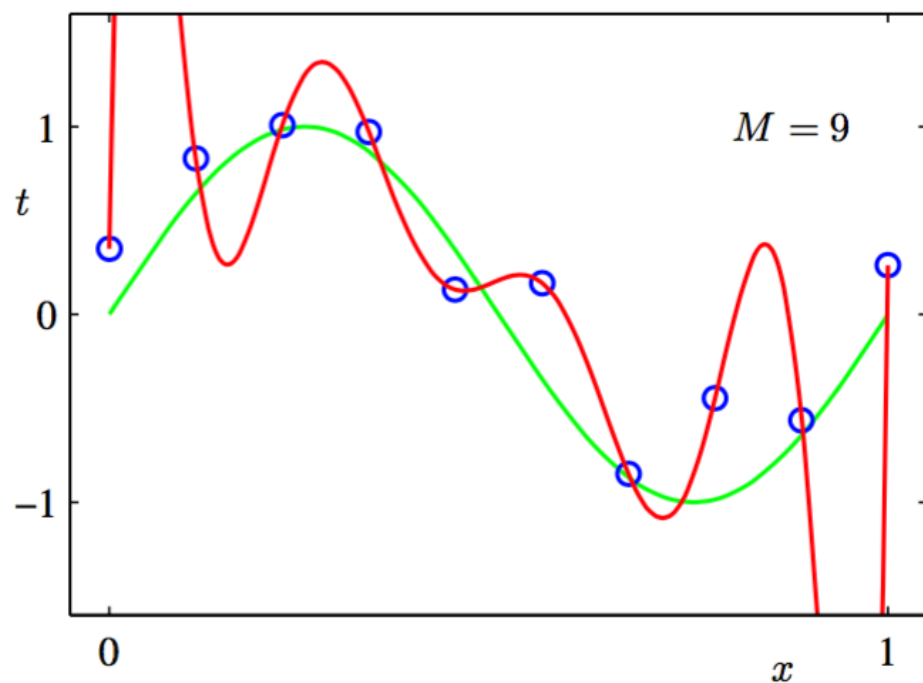


Качество на обучении ниже



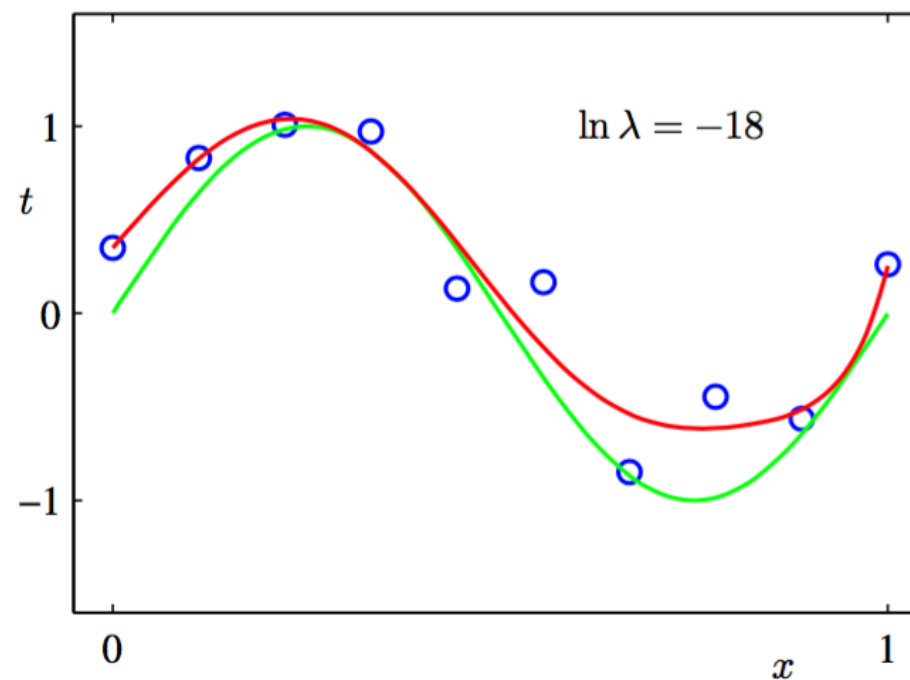
# Гиперпараметры

Без регуляризации



Низкая обобщающая  
способность

С регуляризацией

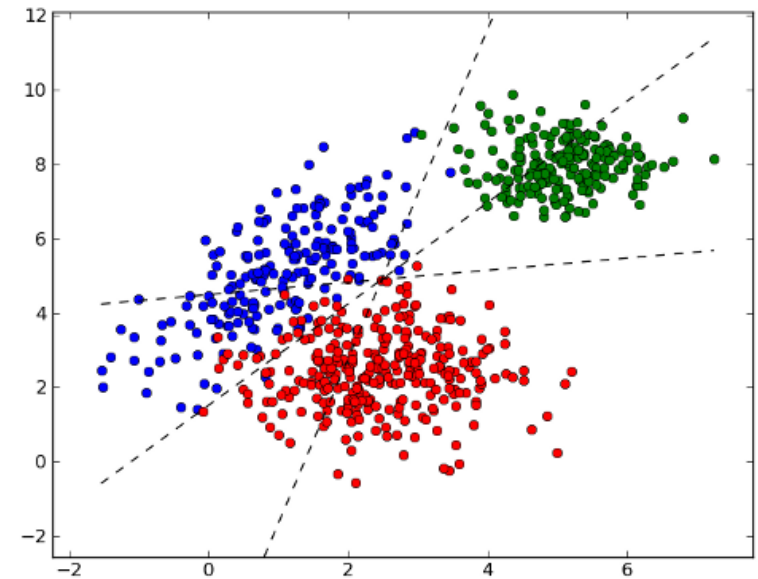


Высокая обобщающая  
способность

# Многоклассовые задачи

# Многоклассовая классификация

- $\mathbb{Y} = \{1, 2, \dots, K\}$



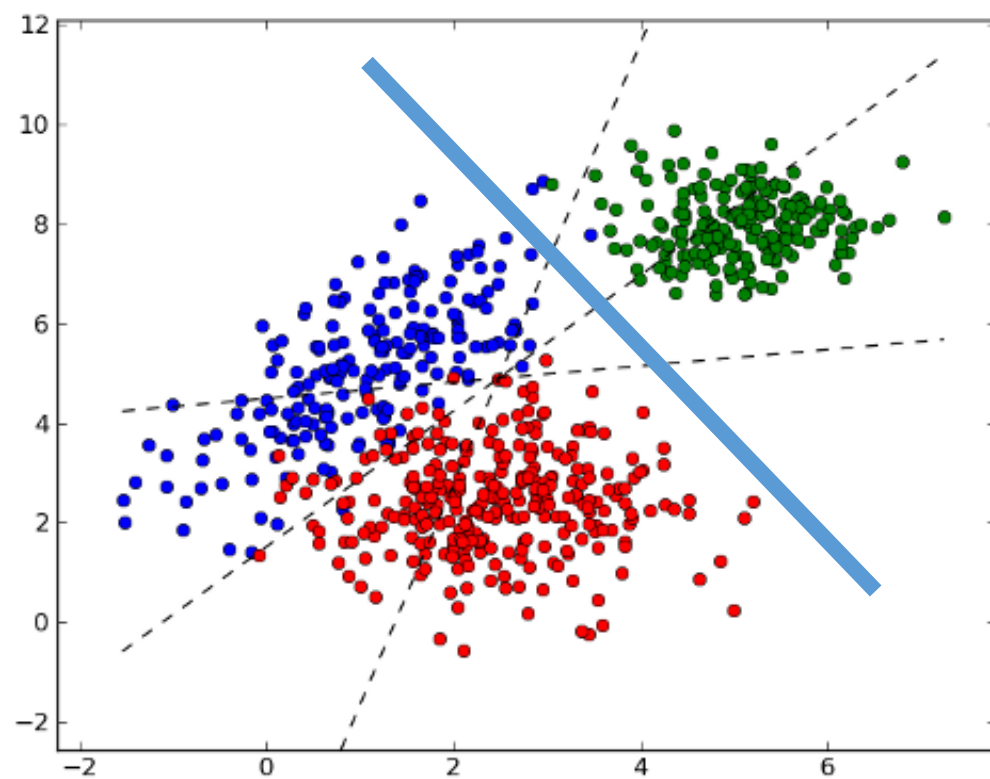
# Бинарная классификация

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

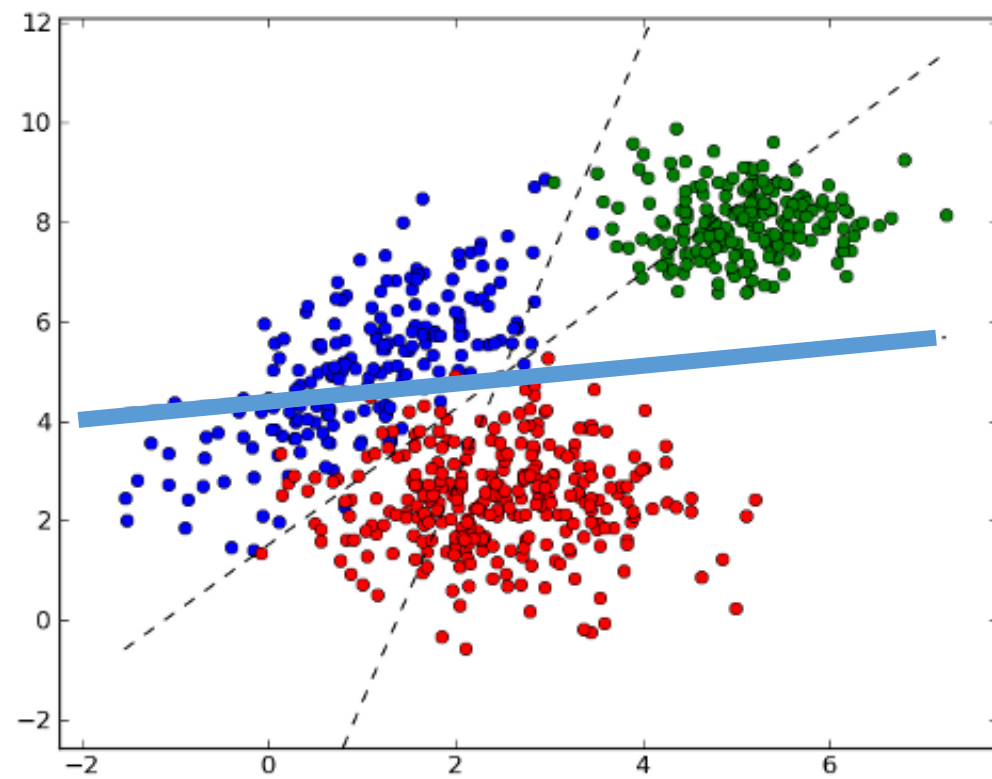
# One-vs-all

- Способ сведения многоклассовой задачи к набору бинарных классификаций
- Обучаем свой классификатор для каждого класса
- Задача: отделение класса от всех остальных

# One-vs-all



# One-vs-all



# One-vs-all

- $K$  задач бинарной классификации
- $k$ -я задача:
  - $X = (x_i, [y_i = k])_{i=1}^{\ell}$
  - Классификатор  $a_k(x) = \text{sign} \langle w_k, x \rangle$
- Алгоритм:

$$a(x) = \arg \max_{k \in \{1, \dots, K\}} \langle w_k, x \rangle$$



# Матрица ошибок

	$y = 1$	$y = 2$	...	$y = K$
$a(x) = 1$	$q_{11}$	$q_{12}$	...	$q_{1K}$
$a(x) = 2$	$q_{21}$	$q_{22}$	...	$q_{2K}$
...	...	...	...	...
$a(x) = K$	$q_{K1}$	$q_{K2}$	...	$q_{KK}$

# Доля правильных ответов

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

# Точность и полнота

- Относительно каждого класса
- Можно усреднить точность и полноту по всем классам
- Можно усреднить F-меру

# 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,  $R^2$
- Кросс-валидация
- Многоклассовая классификация: one-vs-all