

# Введение в анализ данных

## Лекция 9

Измерение качества моделей и многоклассовая классификация

Евгений Соколов

[sokolov.evg@gmail.com](mailto:sokolov.evg@gmail.com)

НИУ ВШЭ, 2016

# Организационное

- Скоро зачёт (коллоквиум)!
- 12 апреля
- Список вопросов на вики

# План на сегодня

- Метрики качества классификации и регрессии
- Параметры и гиперпараметры моделей
- Кросс-валидация
- Многоклассовая классификация

# 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

# 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):

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

# 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?

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

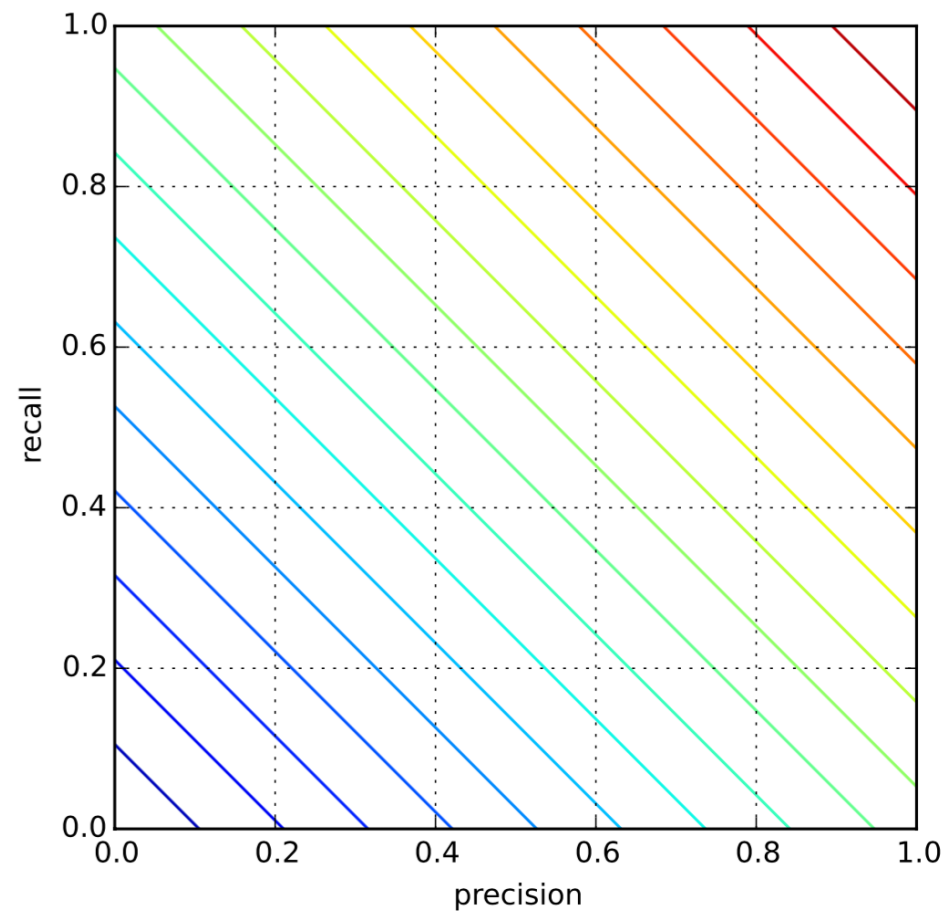
# Точность и полнота

- 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

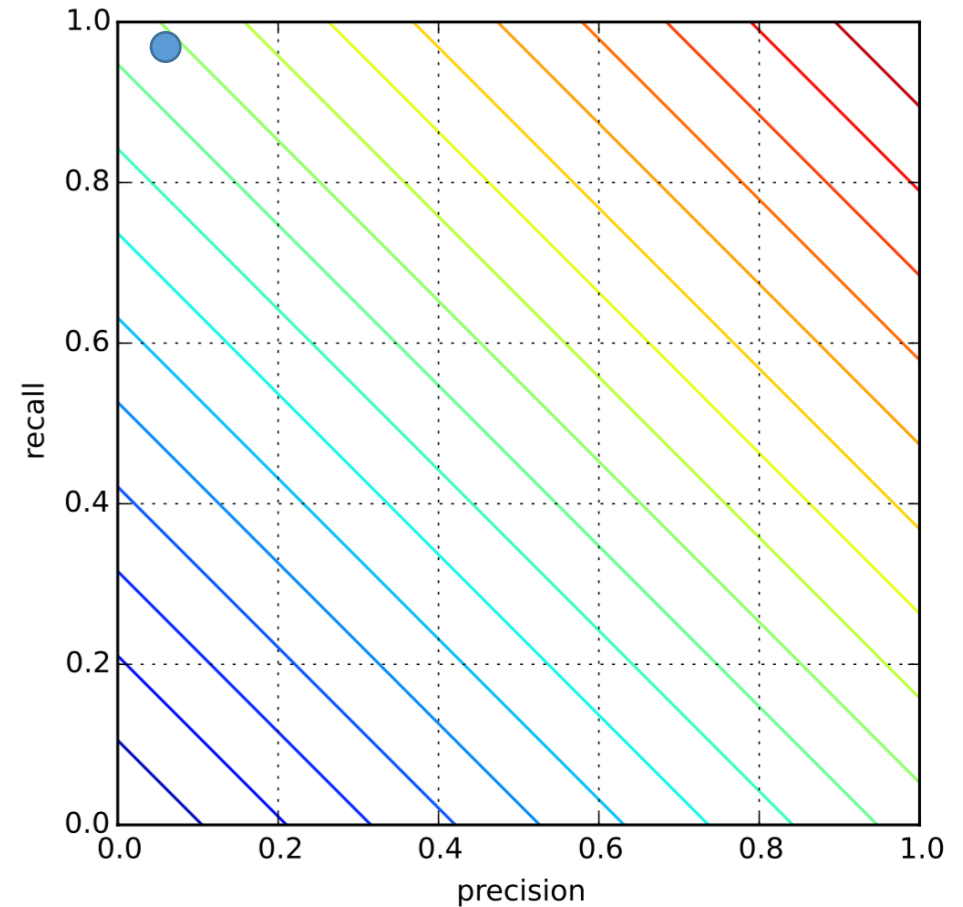
$$A = \frac{1}{2}(\text{precision} + \text{recall})$$



# Арифметическое среднее

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

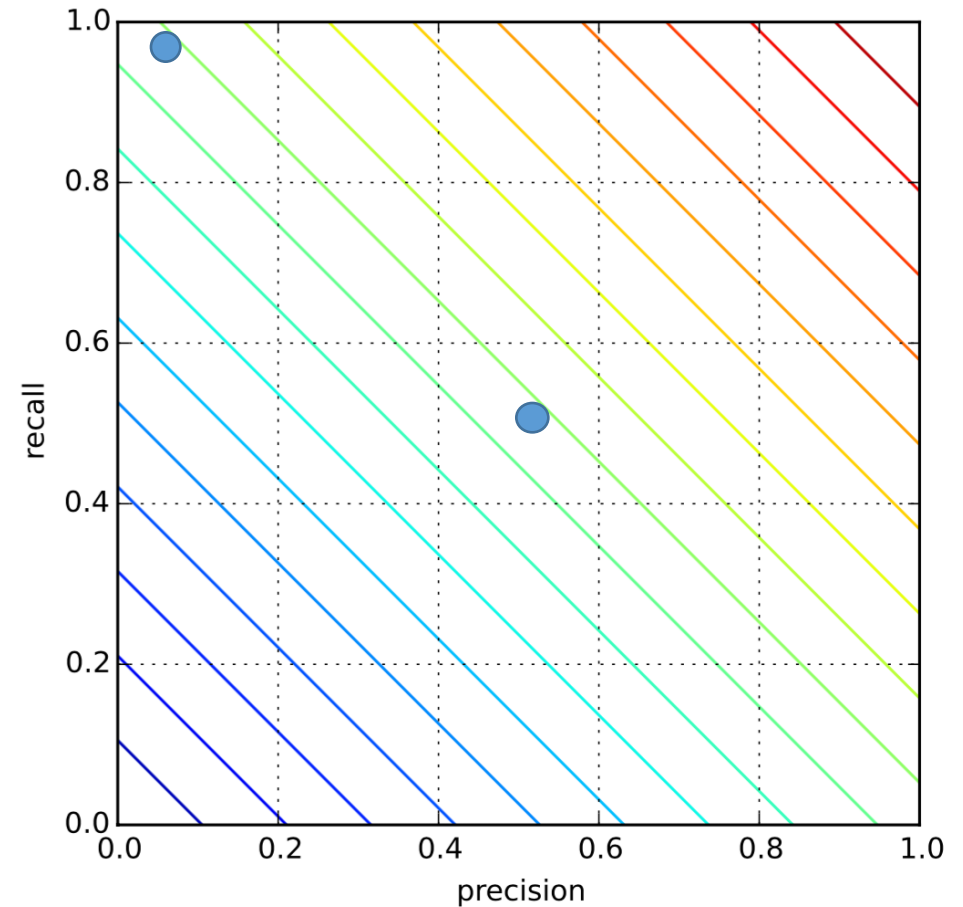
- precision = 0.1
- recall = 1
- $A = 0.55$
- Плохой алгоритм



# Арифметическое среднее

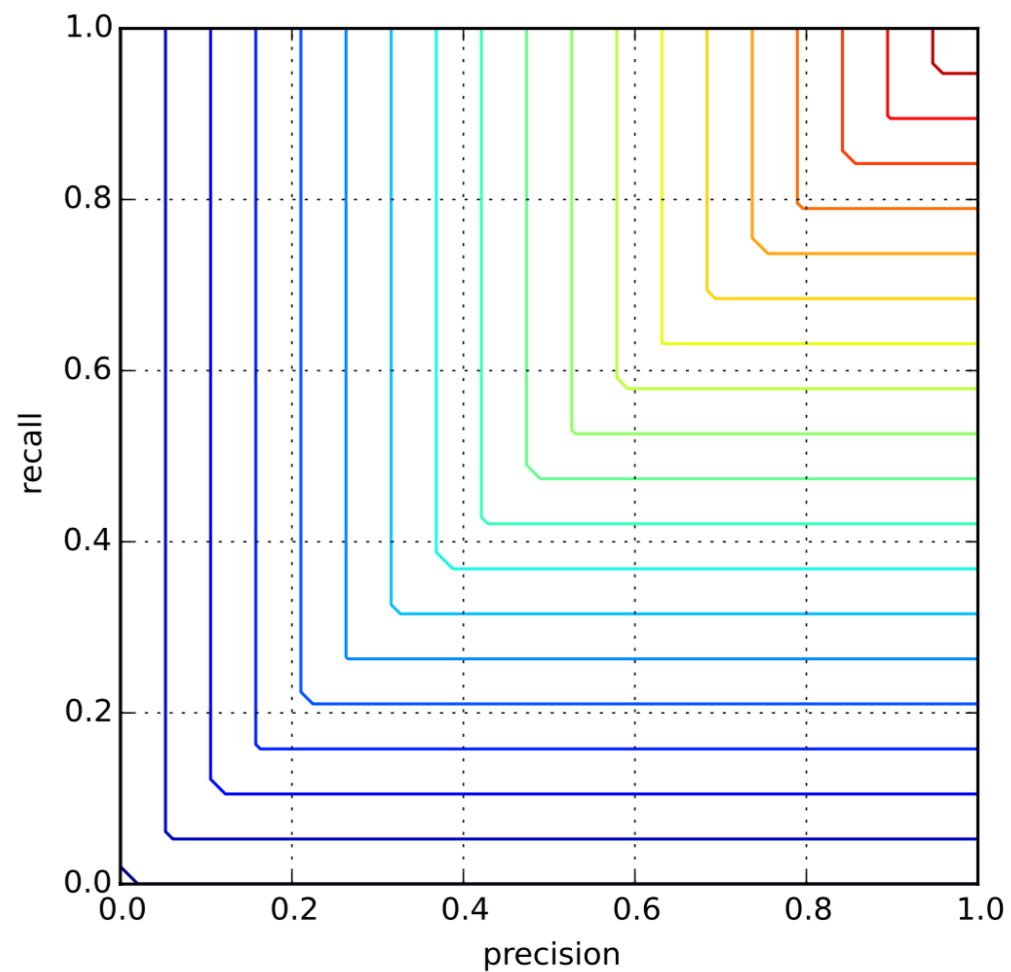
$$A = \frac{1}{2} (\text{precision} + \text{recall})$$

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



# Минимум

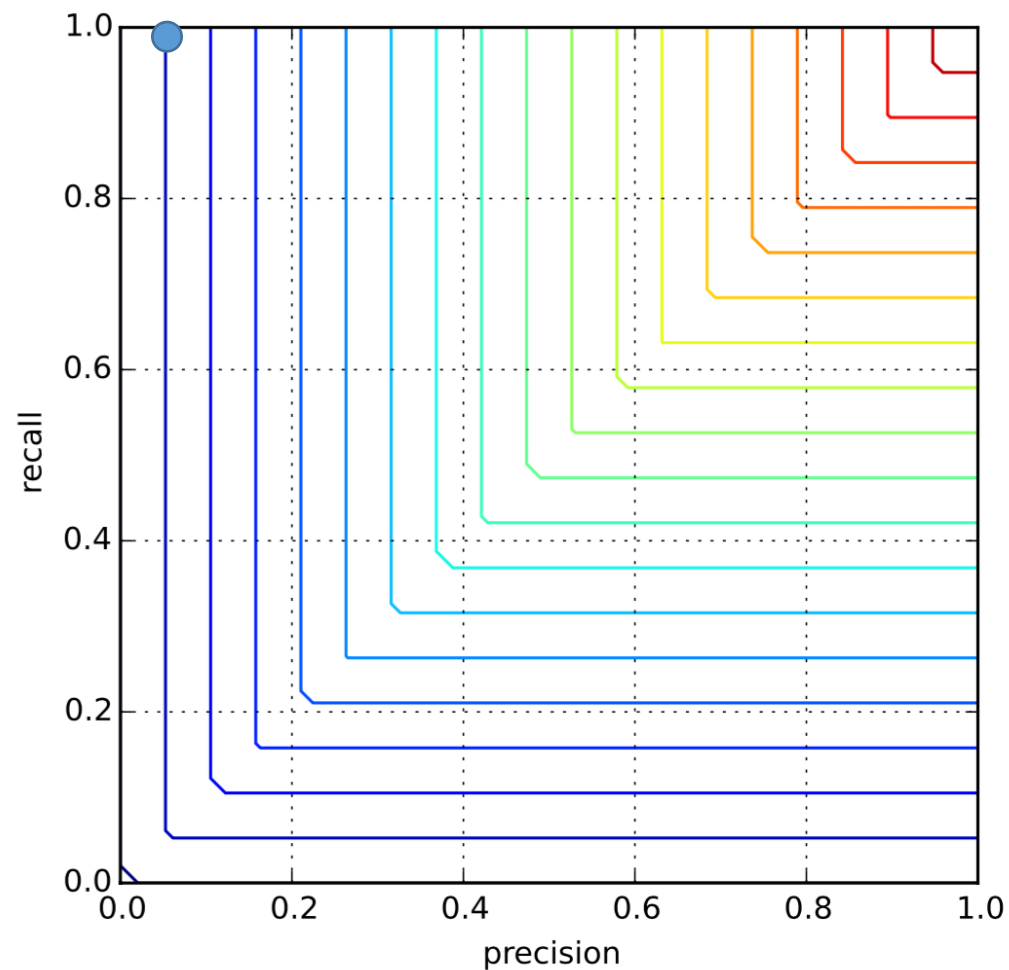
$$M = \min(\text{precision}, \text{recall})$$



# Минимум

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

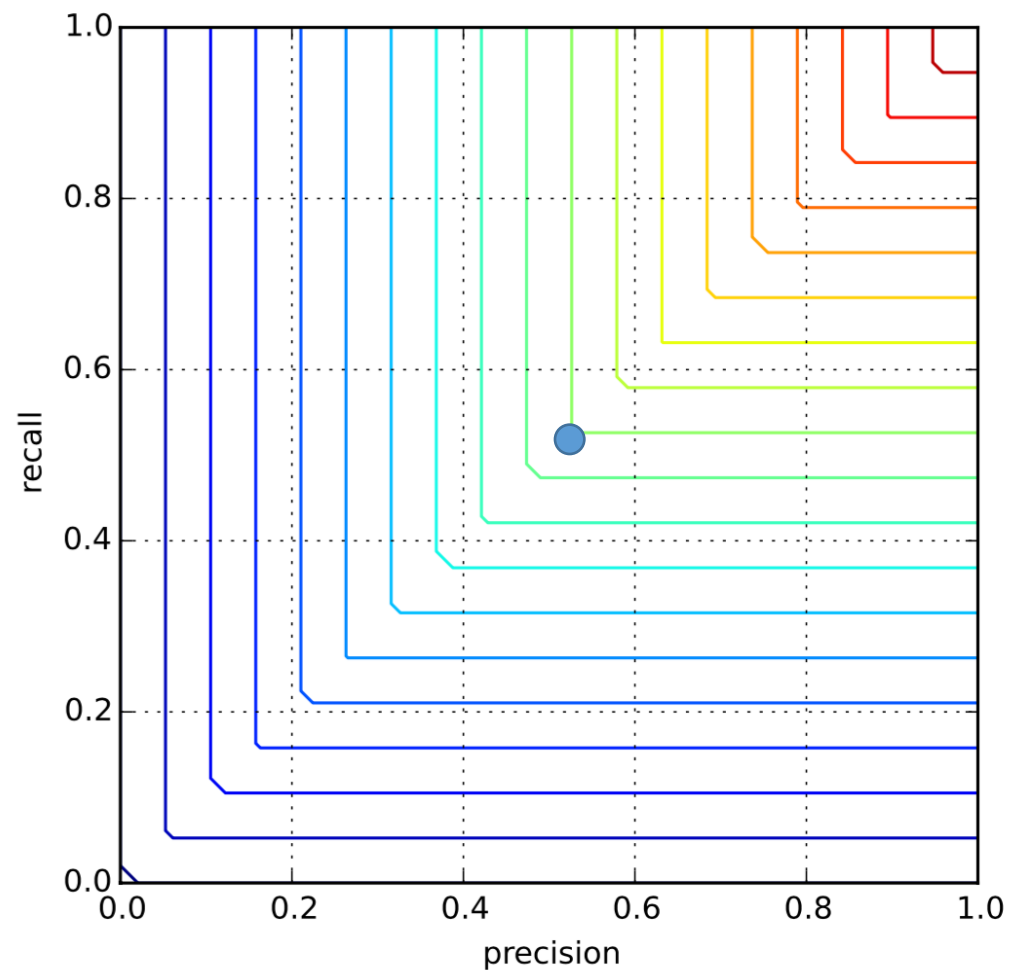
- precision = 0.05
- recall = 1
- $M = 0.05$



# Минимум

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

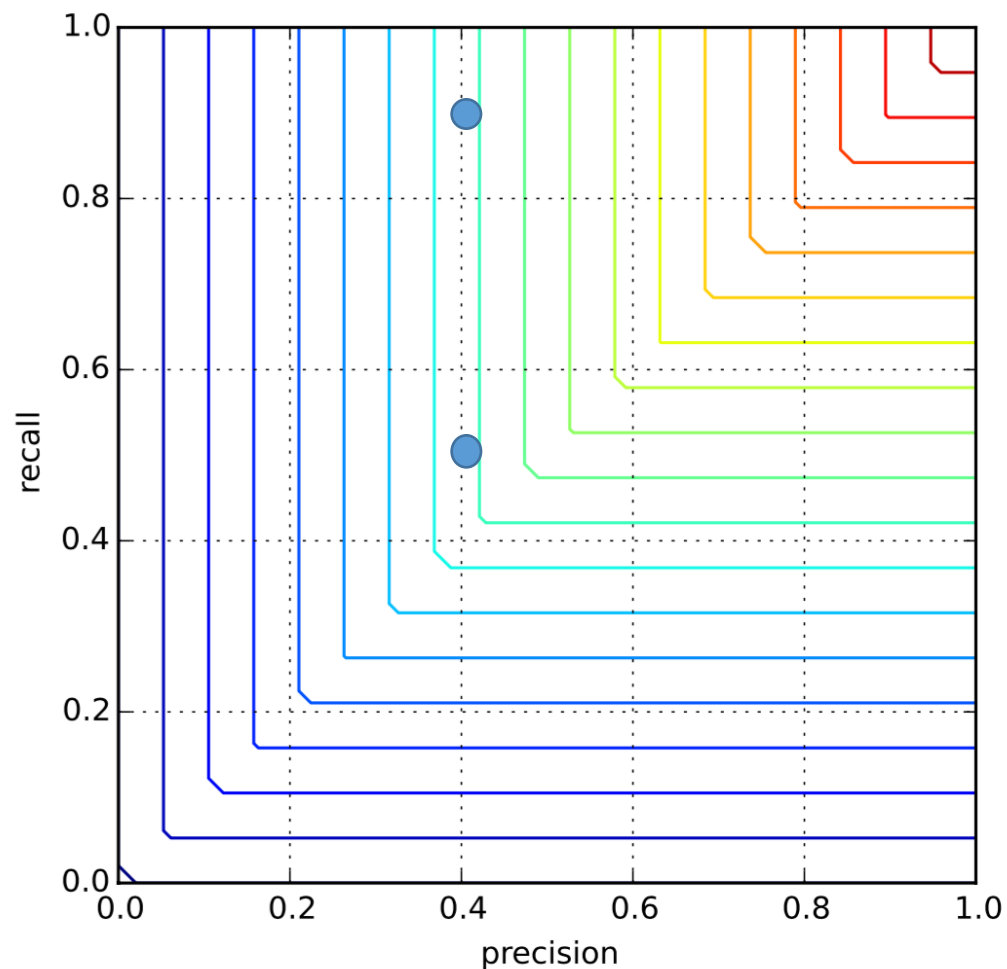
- precision = 0.55
- recall = 0.55
- $M = 0.55$



# Минимум

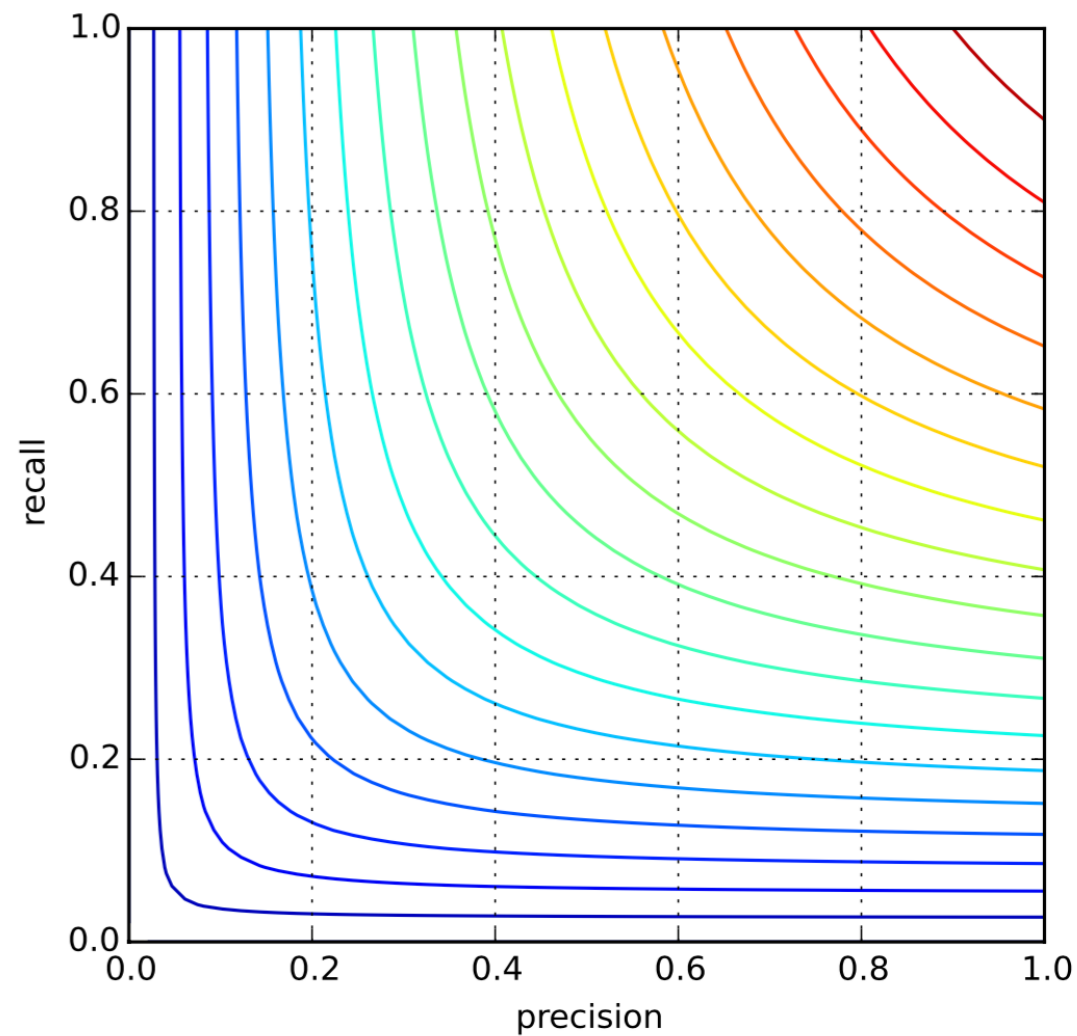
$$M = \min(\text{precision}, \text{recall})$$

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



# F-meap

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

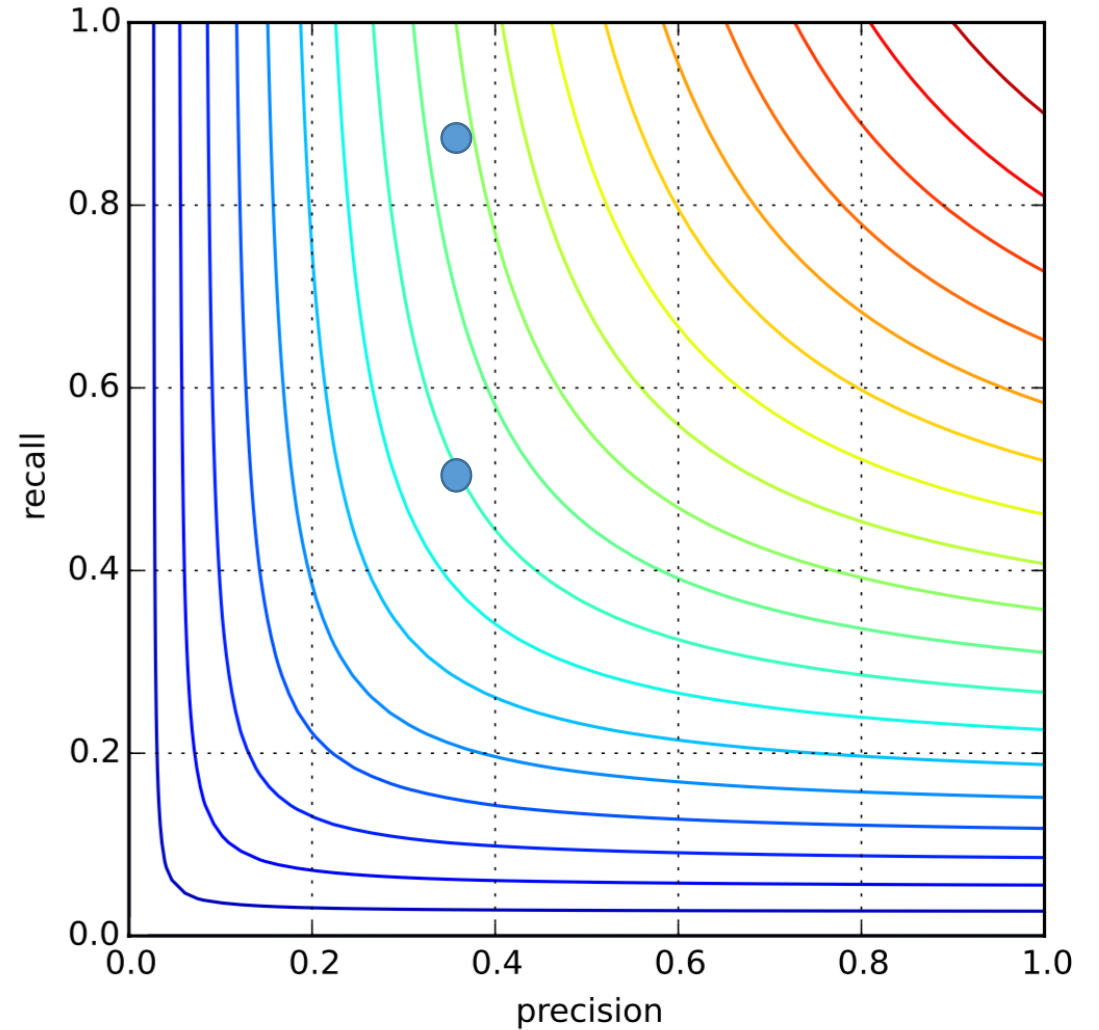




# F-meapa

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

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



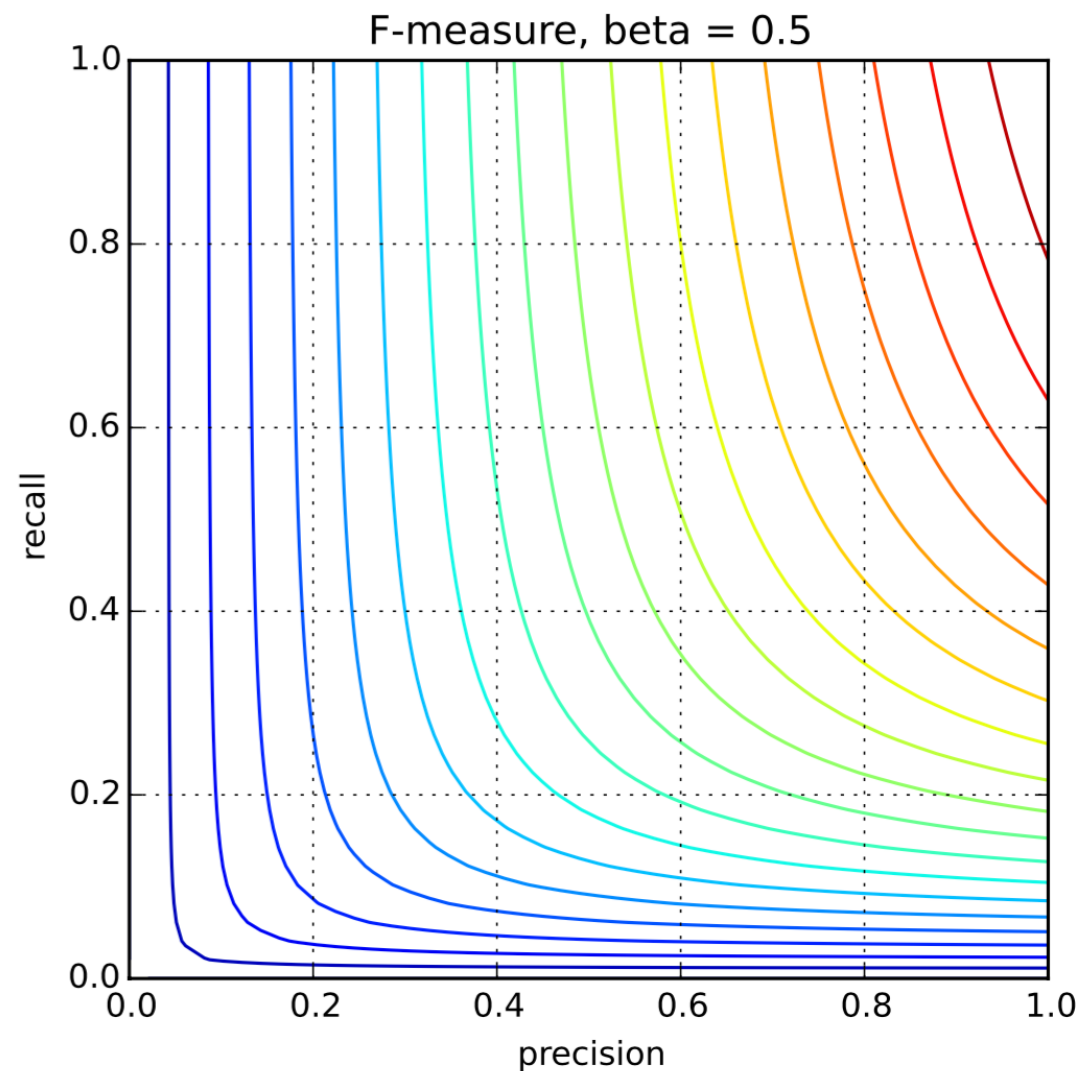
# F-measure

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

# F-мера

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

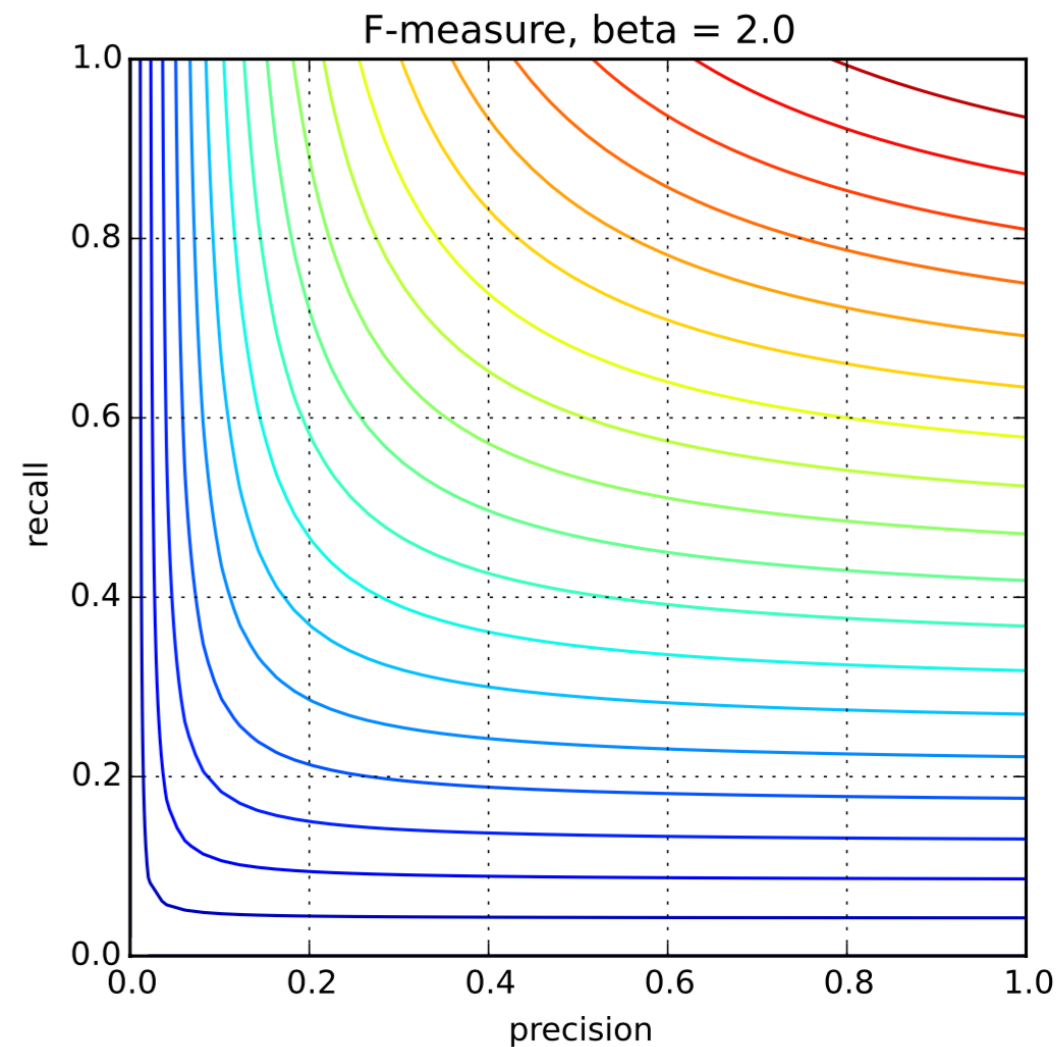
- $\beta = 0.5$
- Важнее полнота



# 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



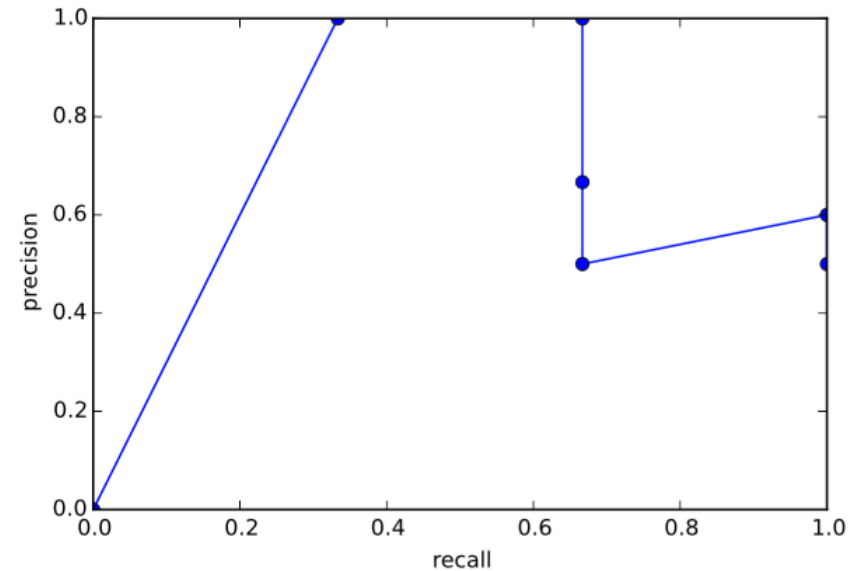
-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

# Оценка принадлежности

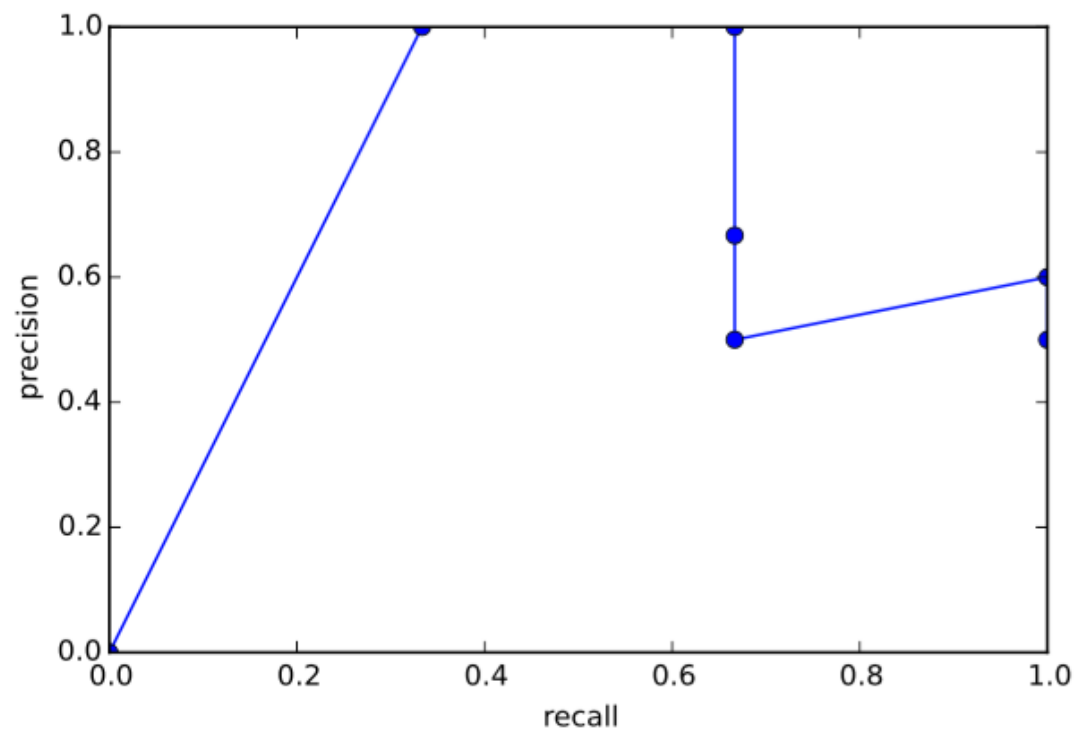
- Пример: кредитный скоринг
- $b(x)$  — оценка вероятности возврата кредита
- $a(x) = [b(x) > 0.5]$
- precision = 0.1, recall = 0.7
- В чем дело — в пороге или в алгоритме?

# PR-кривая

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

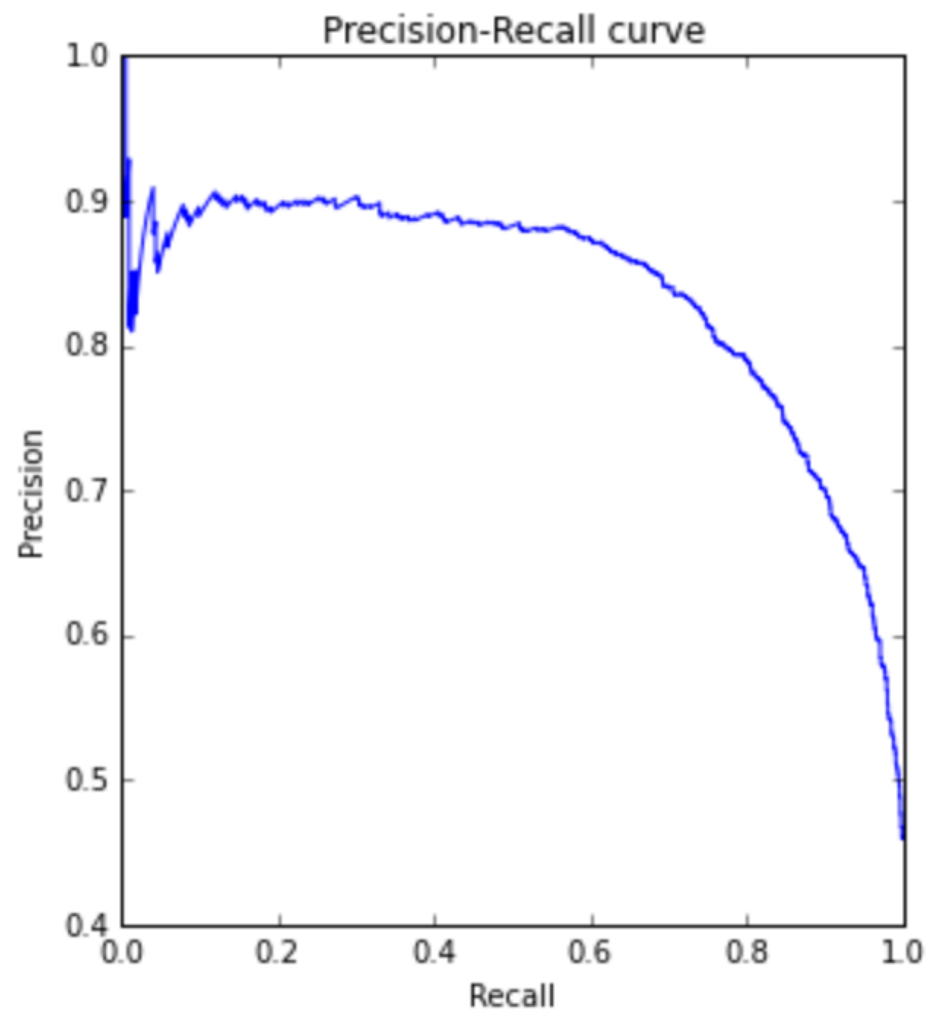


# PR-кривая



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

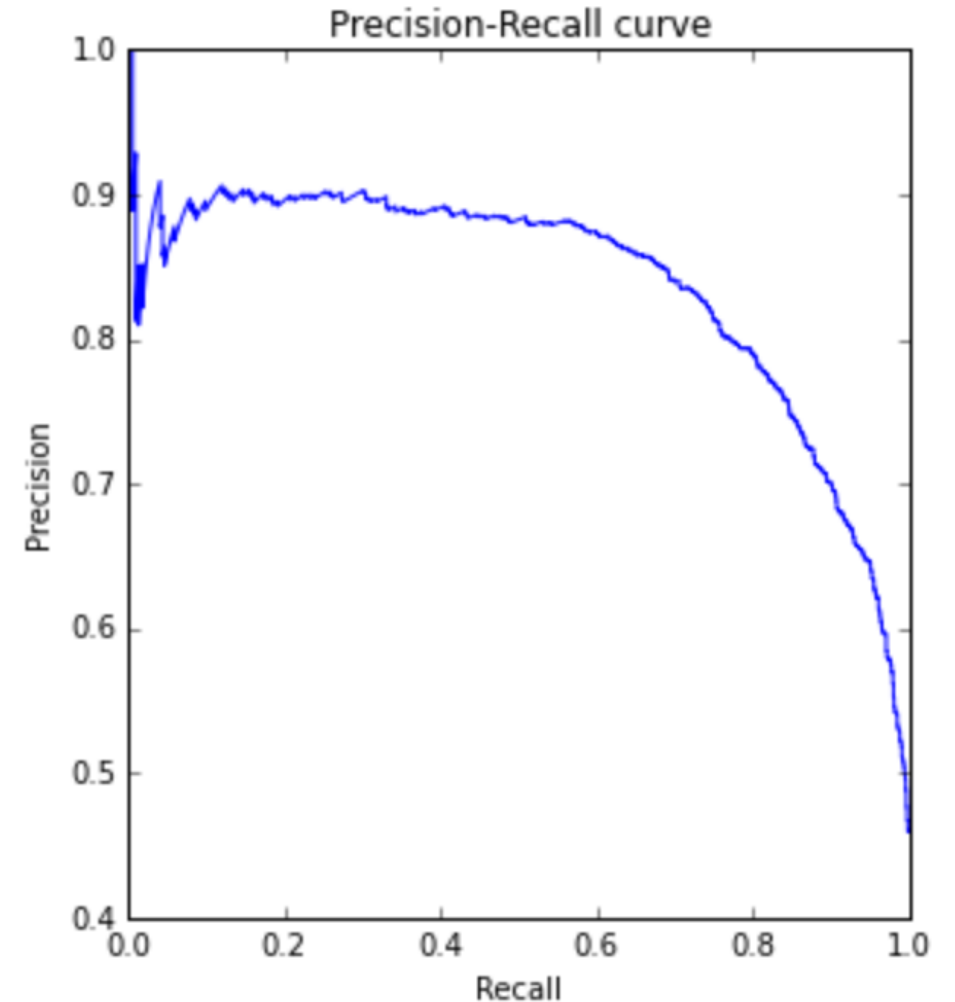
# PR-кривая в реальности





# 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

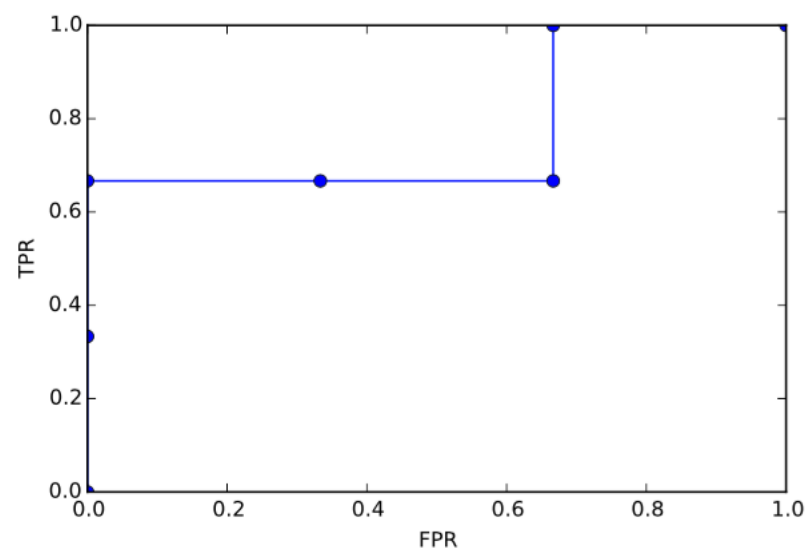
- Receiver Operating Characteristic

- Ось X — False Positive Rate

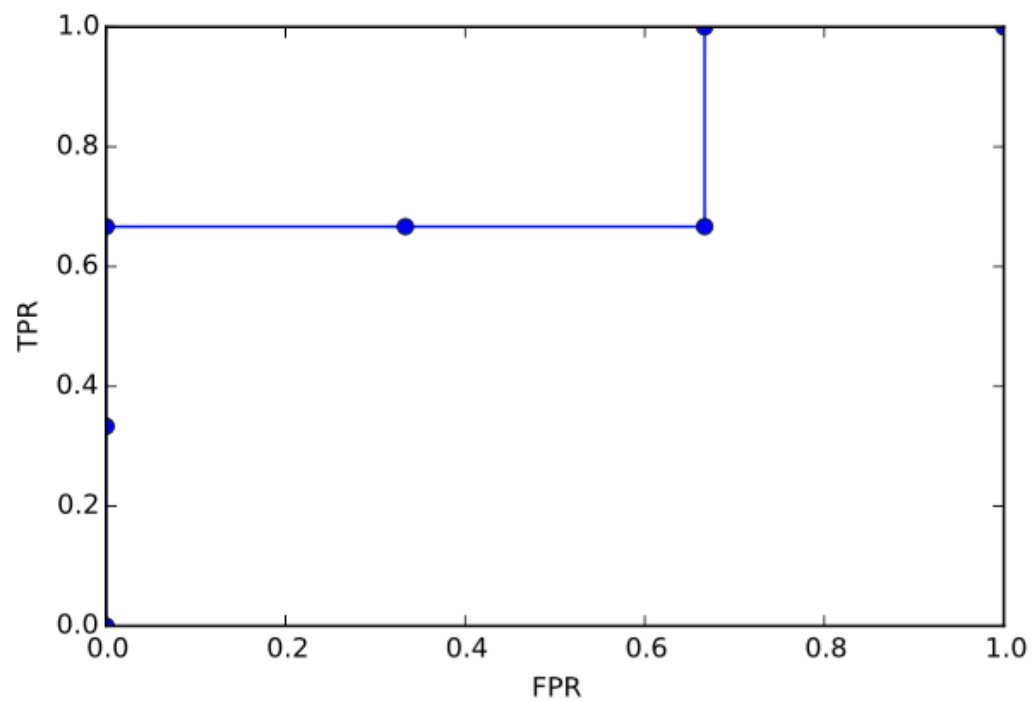
$$FPR = \frac{FP}{FP + TN}$$

- Ось Y — True Positive Rate

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

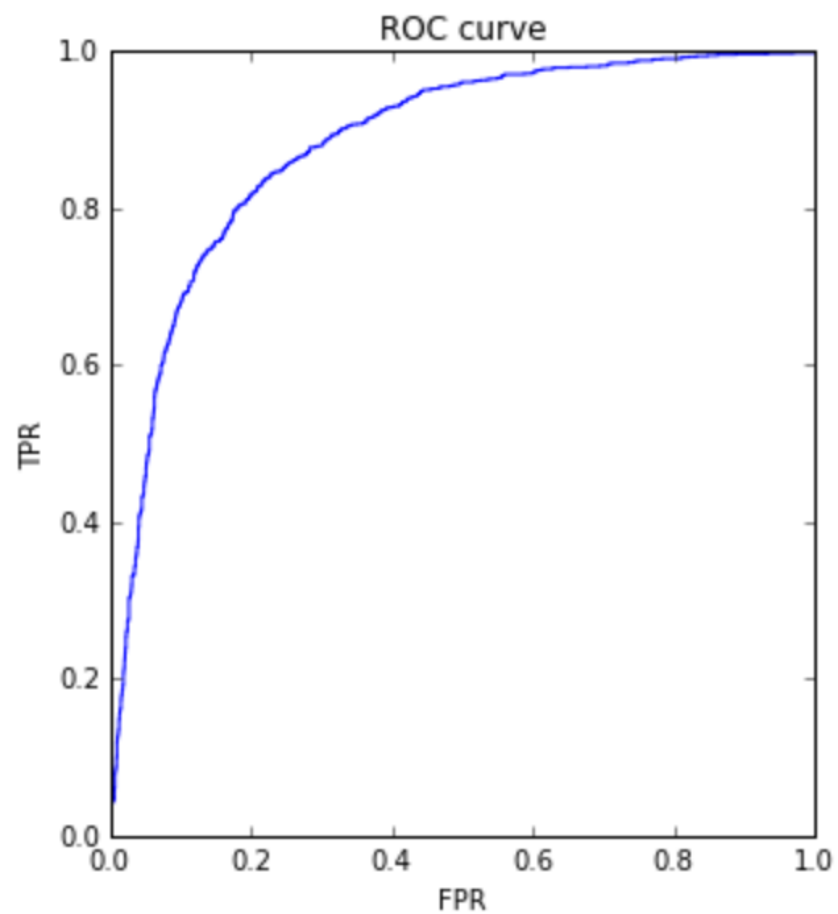


# ROC-кривая



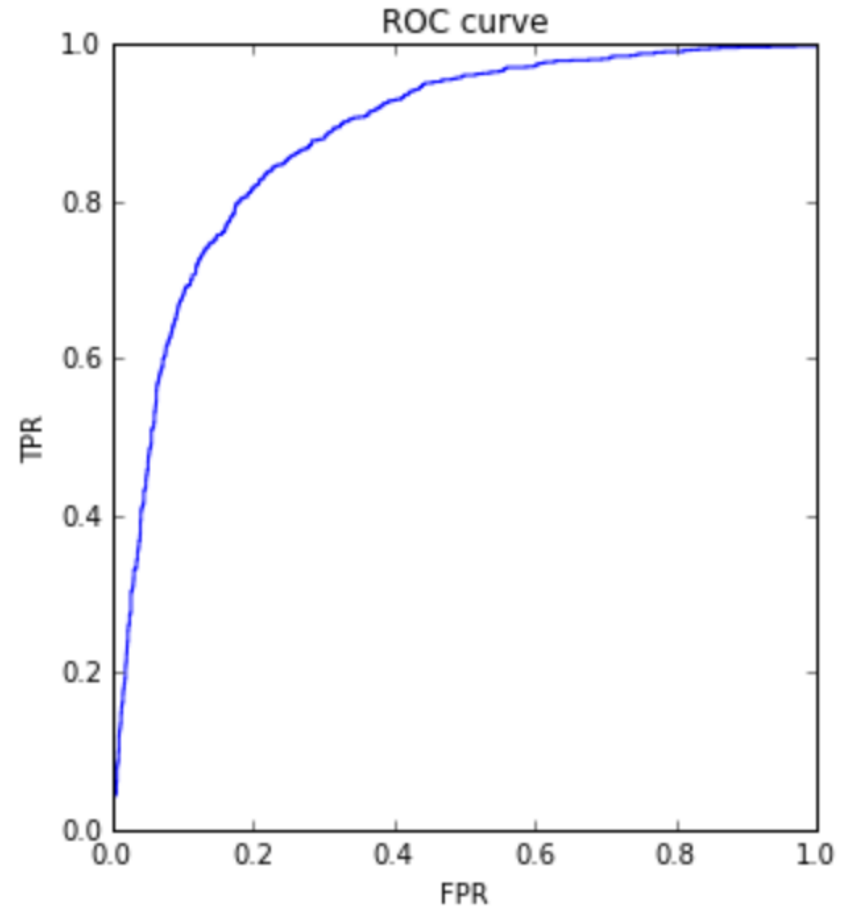
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
$y$	0	1	0	0	1	1

# 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:  $AUC-ROC \approx 0.5$

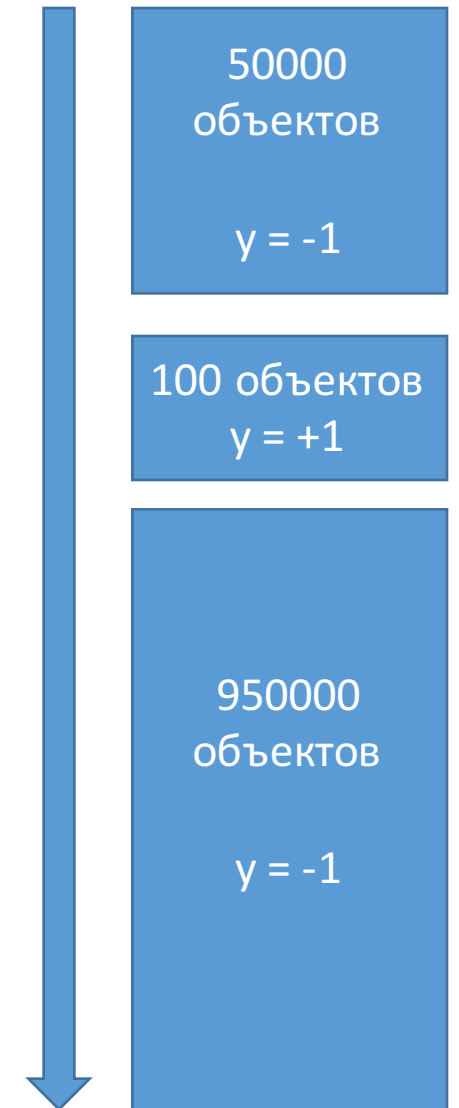
# AUC-PRC

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

- 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

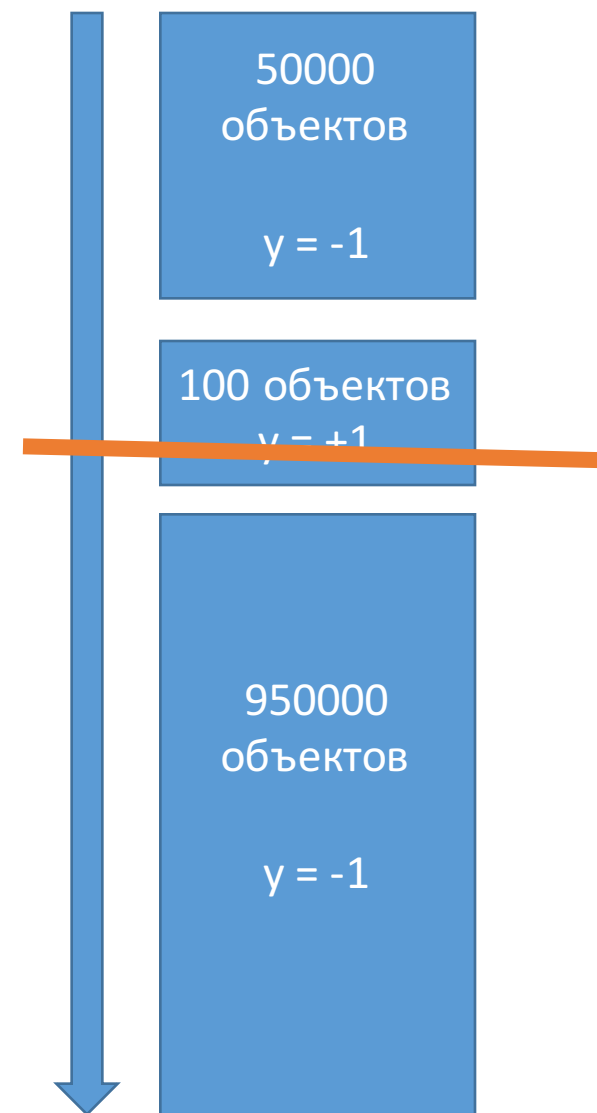
- AUC-ROC = 0.95
- AUC-PRC = 0.001





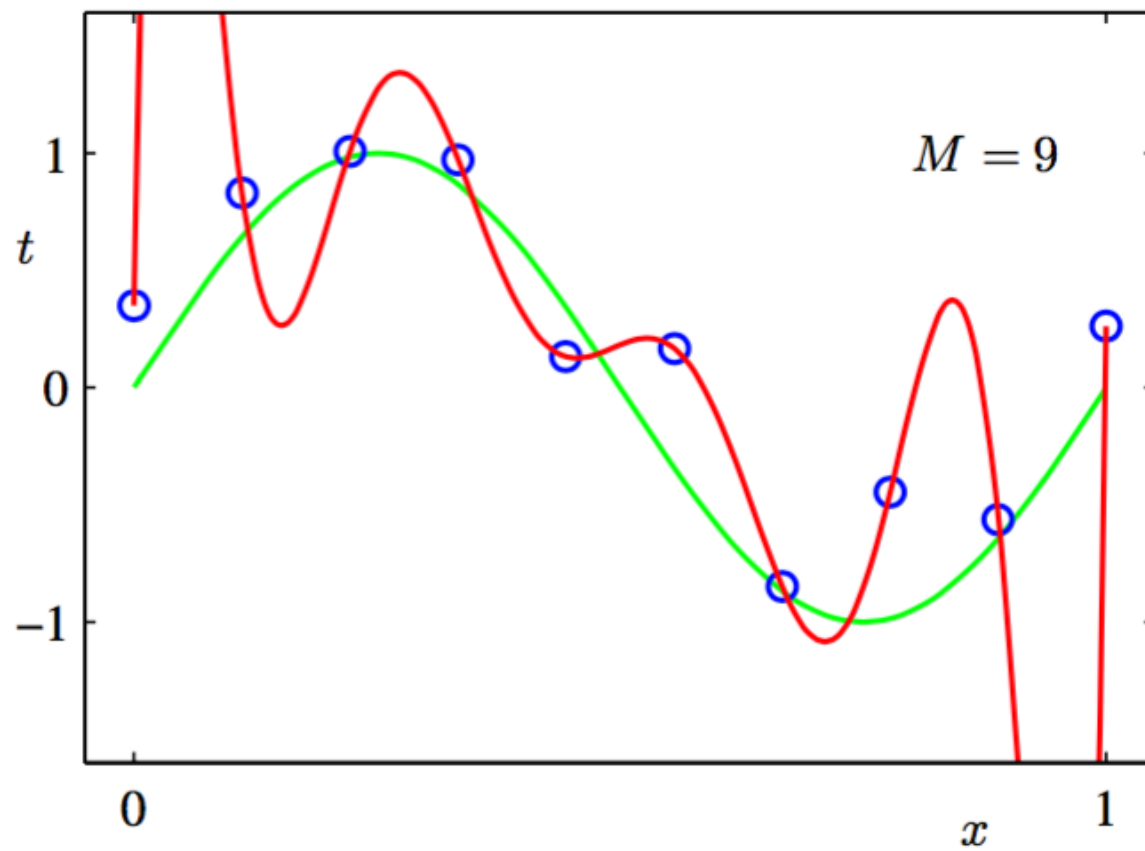
# Пример

- 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

# Переобучение



# Регуляризация

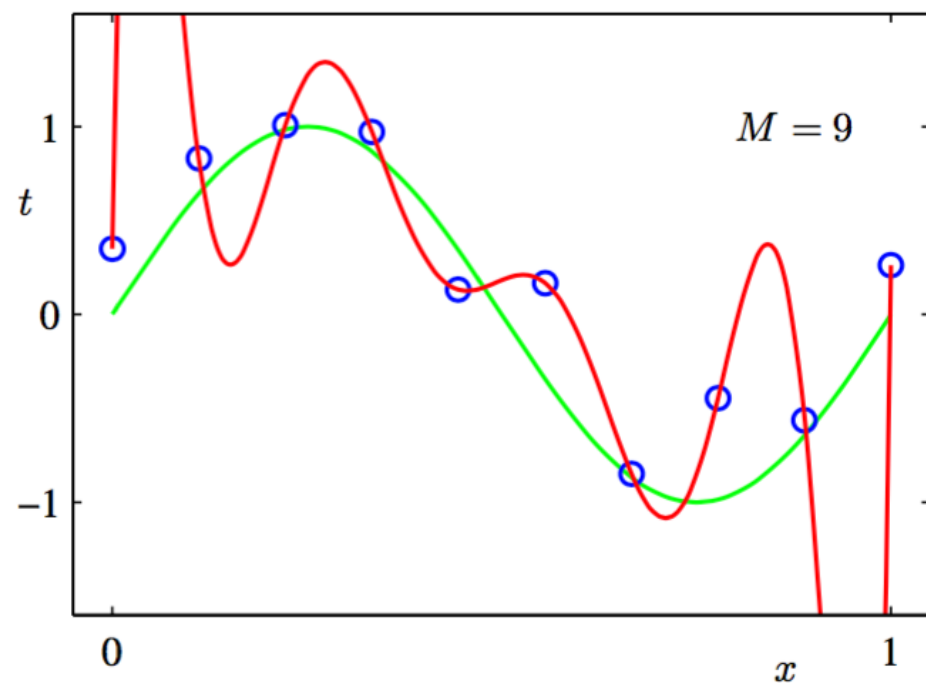
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$$

# 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

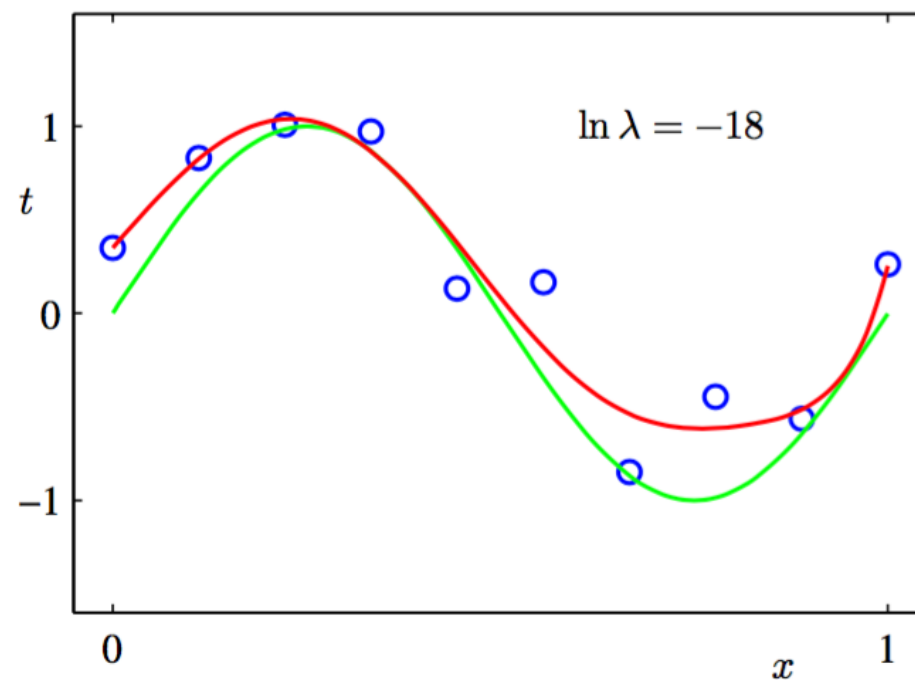
# Гиперпараметры

Без регуляризации



Высокое качество на обучении

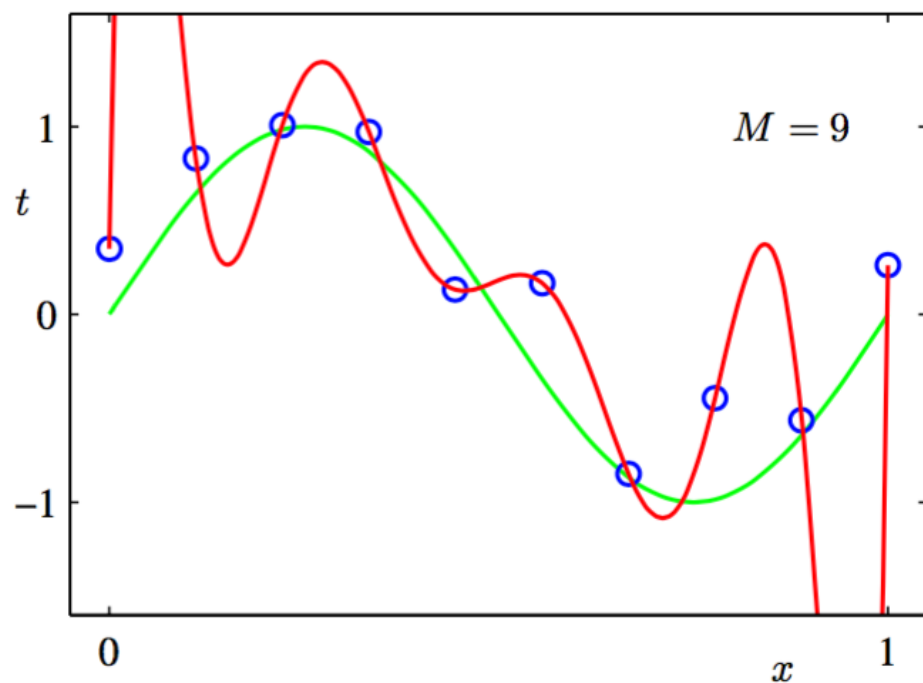
С регуляризацией



Качество на обучении ниже

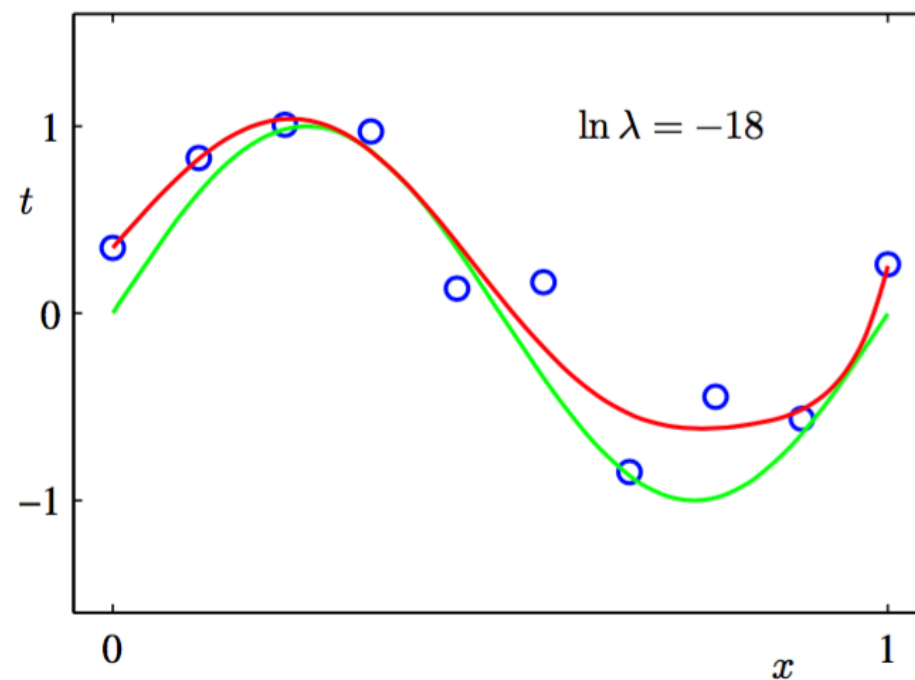
# Гиперпараметры

Без регуляризации



Низкая обобщающая  
способность

С регуляризацией



Высокая обобщающая  
способность

Оценивание обобщающей  
способности



# Как оценить качество?

- Как алгоритм будет вести себя на новых данных?
- Какая у него будет доля ошибок?
- ...или другая метрика качества
- По обучающей выборке нельзя это оценить

# Отложенная выборка

- Разбиваем выборку на две части
  - Обучающая выборка
  - Отложенная выборка
- На первой обучаем алгоритм
- На второй измеряем качество
- Доля ошибок
  - MSE
  - ...



# Пропорции разбиения

- Маленькая отложенная часть
  - (+) Обучающая выборка репрезентативная
  - (-) Оценка качества ненадежная
- Большая отложенная часть
  - (+) Оценка качества надежная
  - (-) Оценка качества смещенная
- Обычно: 70/30, 80/20, 0.632/0.368

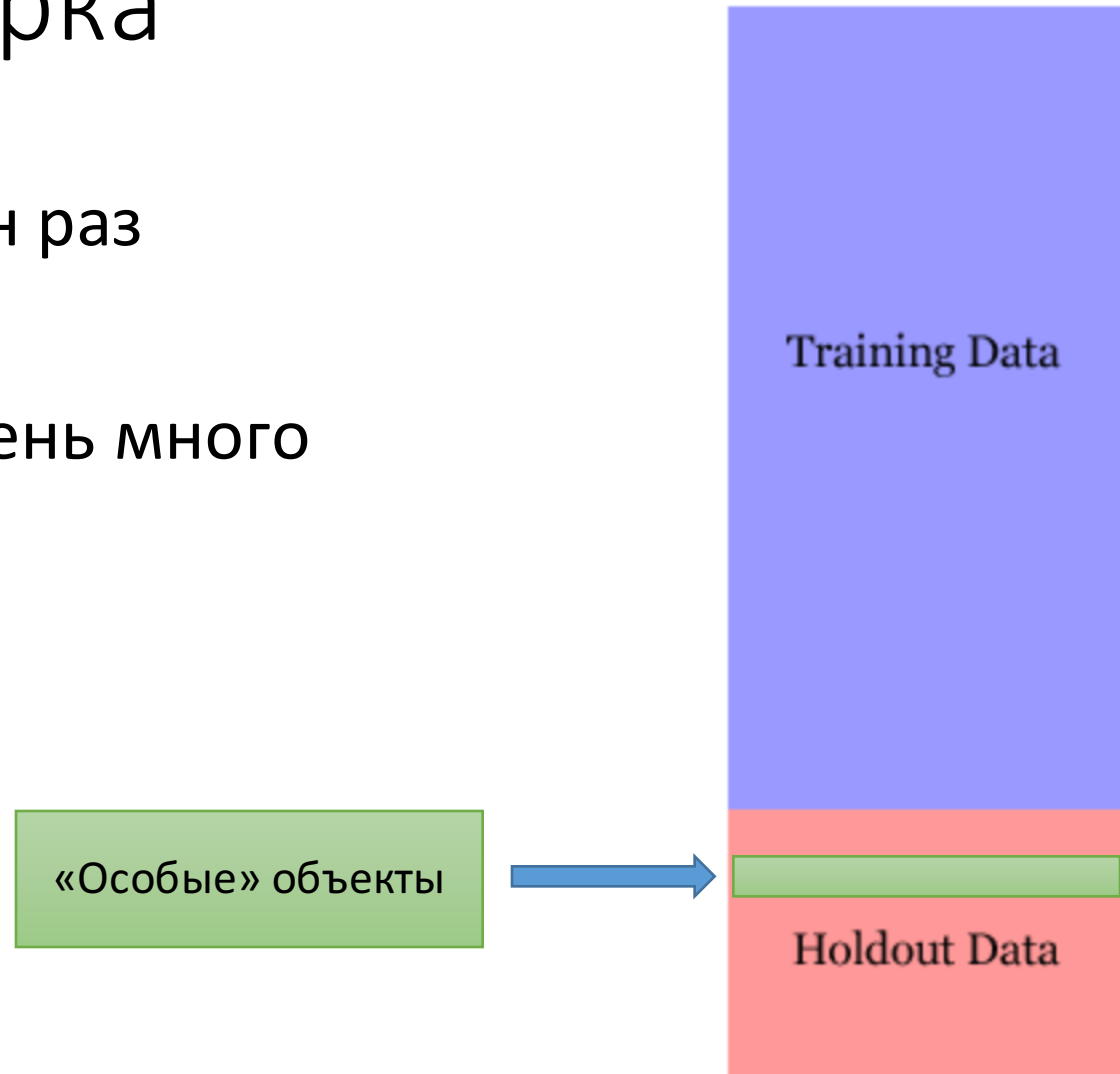
# Отложенная выборка

- (+) Обучаем алгоритм один раз
- (-) Зависит от разбиения
- Подходит, если данных очень много



# Отложенная выборка

- (+) Обучаем алгоритм один раз
- (-) Зависит от разбиения
- Подходит, если данных очень много



# Много отложенных выборок

- Улучшение: разбиваем выборку на две части  $n$  раз
- Усредняем оценку качества



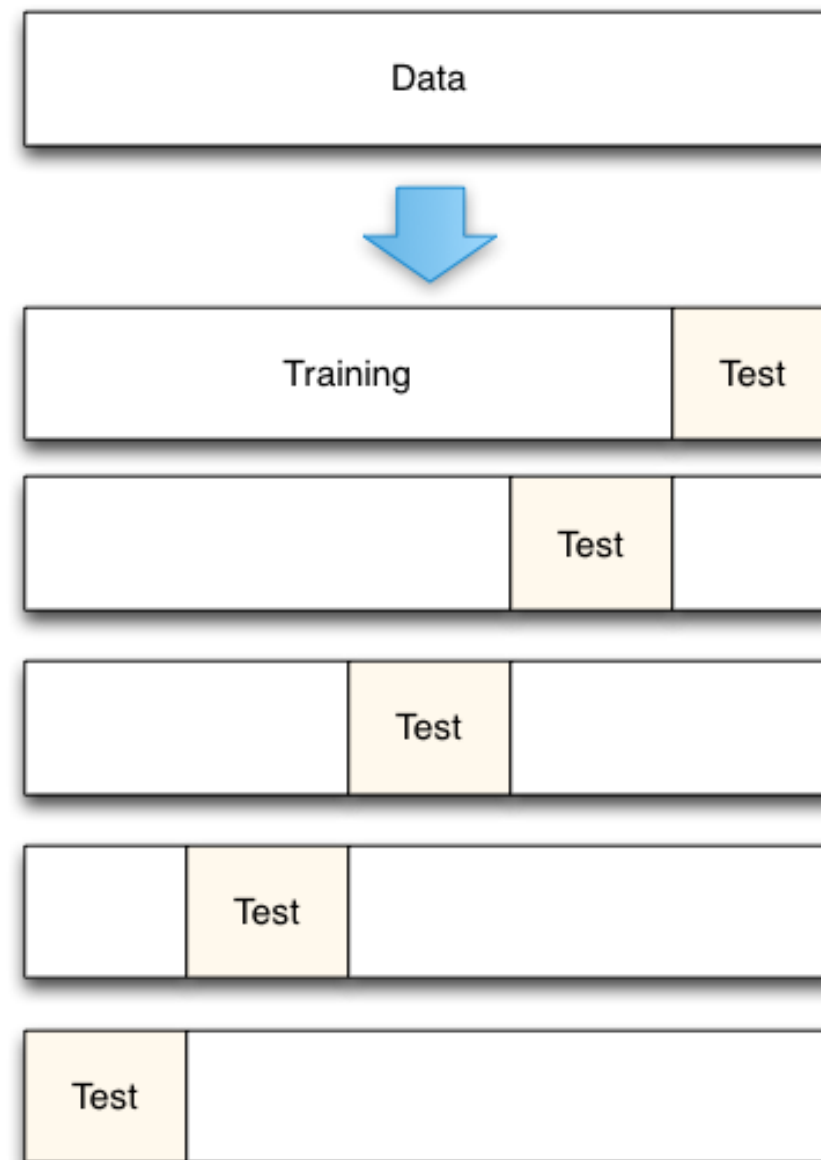
# Много отложенных выборок

- Нет гарантий, что каждый объект побывает в обучении



# Кросс-валидация

- Разбиваем выборку на  $k$  блоков
- Каждая по очереди выступает как тестовая





# Число блоков

- Мало блоков
  - Тестовая выборка всегда большая — (+) надежные оценки
  - Обучение маленькое — (-) смещенные оценки
- Много блоков
  - (-) Ненадежные оценки
  - (+) Несмещенные оценки

# Число блоков

- Обычно:  $k = 3, 5, 10$
- Чем больше выборка, тем меньше нужно  $k$
- Чем больше  $k$ , тем больше раз надо обучать алгоритм

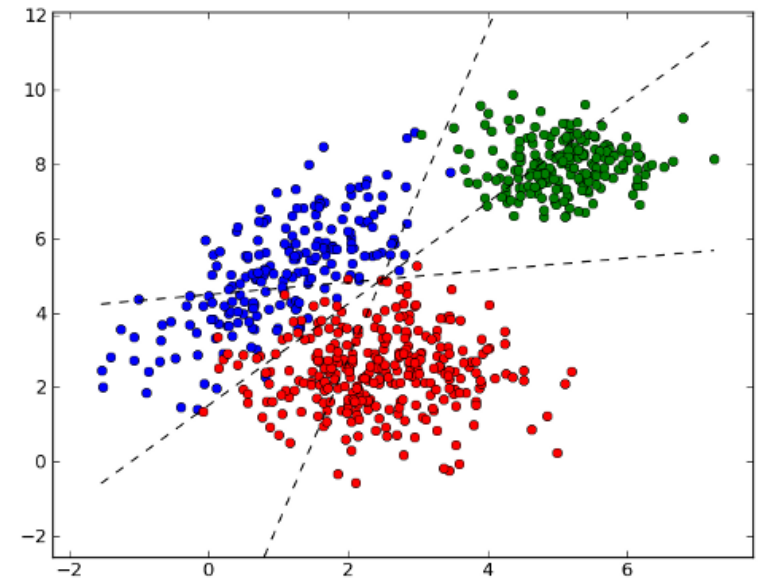
# Совет

- Перемешивайте выборку!
- Объекты могут быть отсортированы
- При разбиении в обучении могут оказаться только мальчики, в контроле — только девочки

# Многоклассовые задачи

# Многоклассовая классификация

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



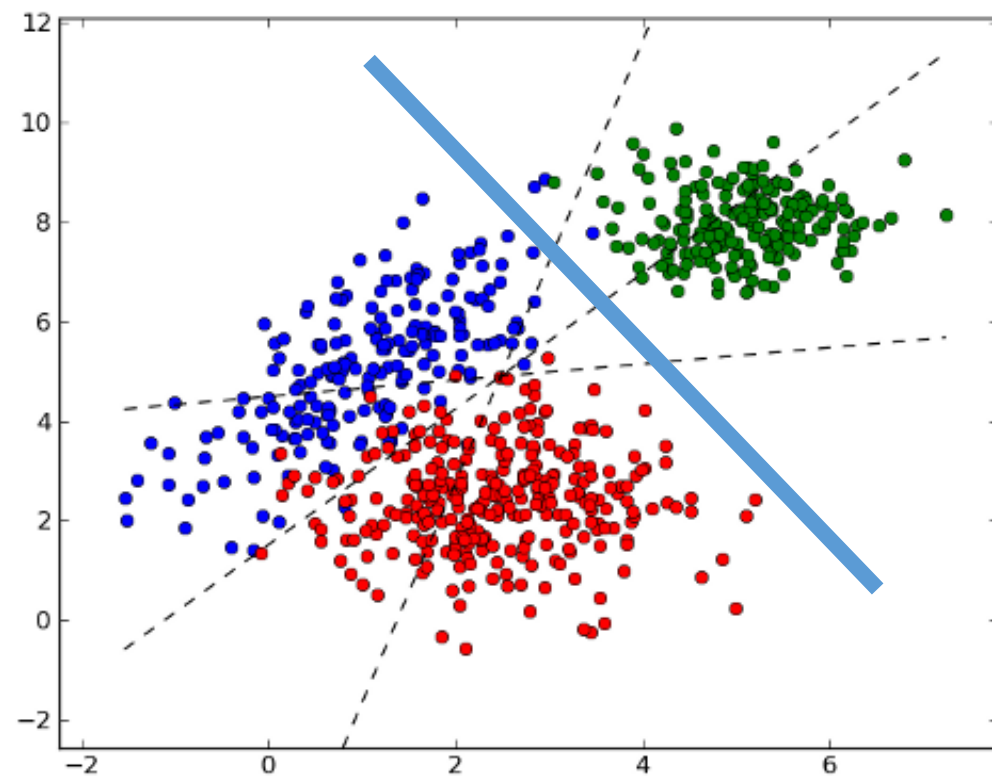
# Бинарная классификация

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

# One-vs-all

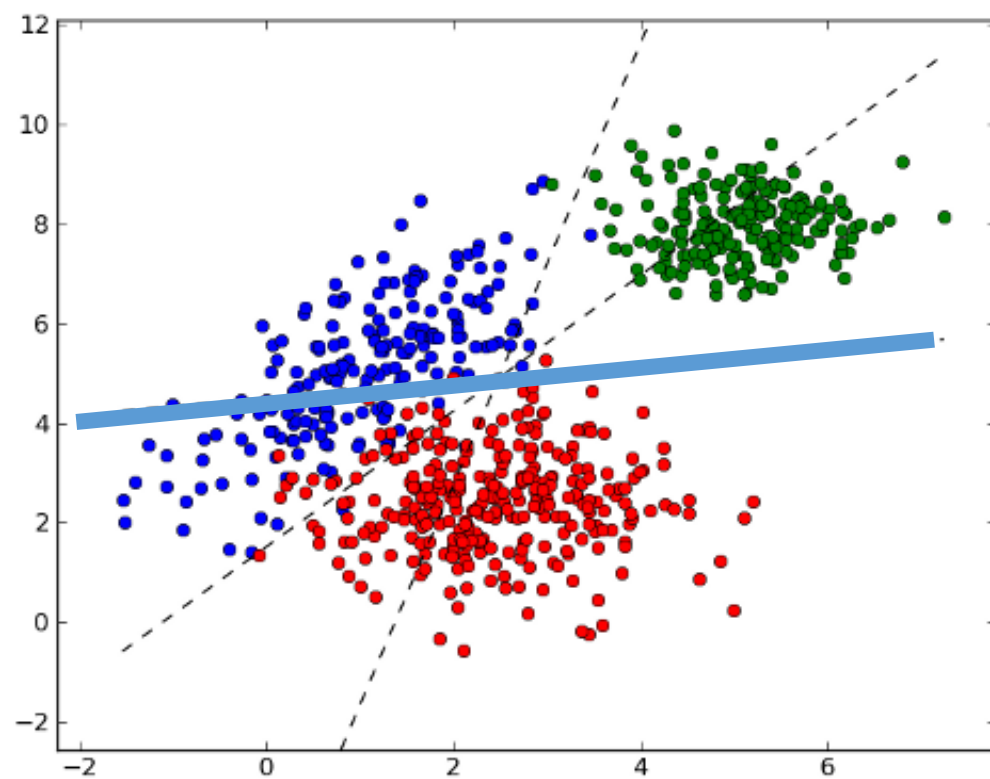
- Способ сведения многоклассовой задачи к набору бинарных классификаций
- Обучаем свой классификатор для каждого класса
- Задача: отделение класса от всех остальных

# 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

# Далее в программе

- Четвёртый модуль:
  - Решающие деревья
  - Случайные леса
  - Кластеризация
- Коллоквиум 12 апреля