

# Регуляризация. Линейная классификация. Метрики качества.

Егор Соловьев

Слайды Евгения Соколова

# План на ближайшие занятия

- 08.04: Линейная регрессия (продолжение), линейная классификация, метрики качества
- 15.04: Метрики качества, многоклассовая классификация, кросс-валидация и гиперпараметры
- 22.04: Практика по линейной классификации и кросс-валидации
- После майских каникул: контрольная

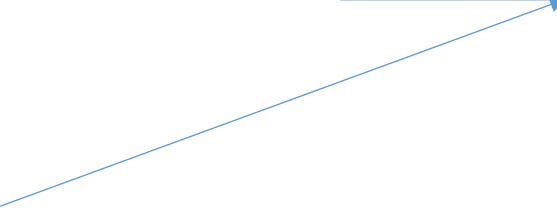
ДЗ:

- 1) SQL/SQLAlchemy + scikit-learn
- 2) Параллельное программирование
- 3) Линейная регрессия/классификация

В прошлых сериях: линейная регрессия

$$a(x) = w_0 + \sum_{j=1}^d w_j x^j$$

Вещественное  
число!



Мультиколлинеарность

# Объекты-признаки

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

- Задача предсказания прибыли магазина в следующем месяце
- Рассмотрим в качестве векторов столбцы матрицы (признаки)

# Подозрительные зависимости

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

- Первый и второй признаки:  $x_2 = 1000x_1$
- Первый — общий вес товаров в тоннах, второй — в килограммах

# Подозрительные зависимости

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

- $x_5 = 0.5x_3 + 0.5x_4$
- Пятый — средняя прибыль за последние два месяца
- Третий и четвертый — прибыль в прошлом и позапрошлом месяце

# Линейная зависимость

— один из векторов равен сумме с весами остальных векторов

- Это плохо:
  - Избыточная информация
  - Лишние затраты на хранение данных
  - Вредит некоторым методам машинного обучения



# Линейная зависимость

- Пусть дан набор векторов  $x_1, \dots, x_n$
- Они линейно зависимы, если
  - существуют такие числа  $\beta_1, \dots, \beta_n$ ,
  - хотя бы одно из которых не равно нулю,
  - что сумма векторов с такими коэффициентами равна нулю

$$\beta_1 x_1 + \dots + \beta_n x_n = 0$$

# Мультиколлинеарность

- Наличие зависимостей между признаками
- Приводит к тому, что решений бесконечное число
- Далеко не все из них имеют хорошую обобщающую способность

# Линейная зависимость

- Худший случай — линейно зависимые признаки
- Существуют такие  $\alpha = (\alpha_1, \dots, \alpha_d)$ , что для любого объекта:

$$\alpha_1 x^1 + \dots + \alpha_d x^d = \langle \alpha, x \rangle = 0$$

# Линейная зависимость

- Допустим, мы нашли решение  $w_*$
- Модифицируем:  $w_1 = w_* + t\alpha$
- ( $t$  — число)
- Ответ нового алгоритма на любом объекте:

$$\langle w_1, x \rangle = \langle w_* + t\alpha, x \rangle = \langle w_*, x \rangle + t\langle \alpha, x \rangle = \langle w_*, x \rangle$$

- $w_1$  — тоже решение!

# Коррелирующие признаки

- Тоже плохо
- Сначала разберёмся с корреляцией

# Коэффициент корреляции

$$\rho(\xi, \eta) = \frac{\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)}{\sqrt{\mathbb{D}\xi \mathbb{D}\eta}}$$

Выборочная корреляция:

$$\rho(x, z) = \frac{\sum_{i=1}^{\ell} (x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{\ell} (x_i - \bar{x})^2 \sum_{i=1}^{\ell} (z_i - \bar{z})^2}}$$

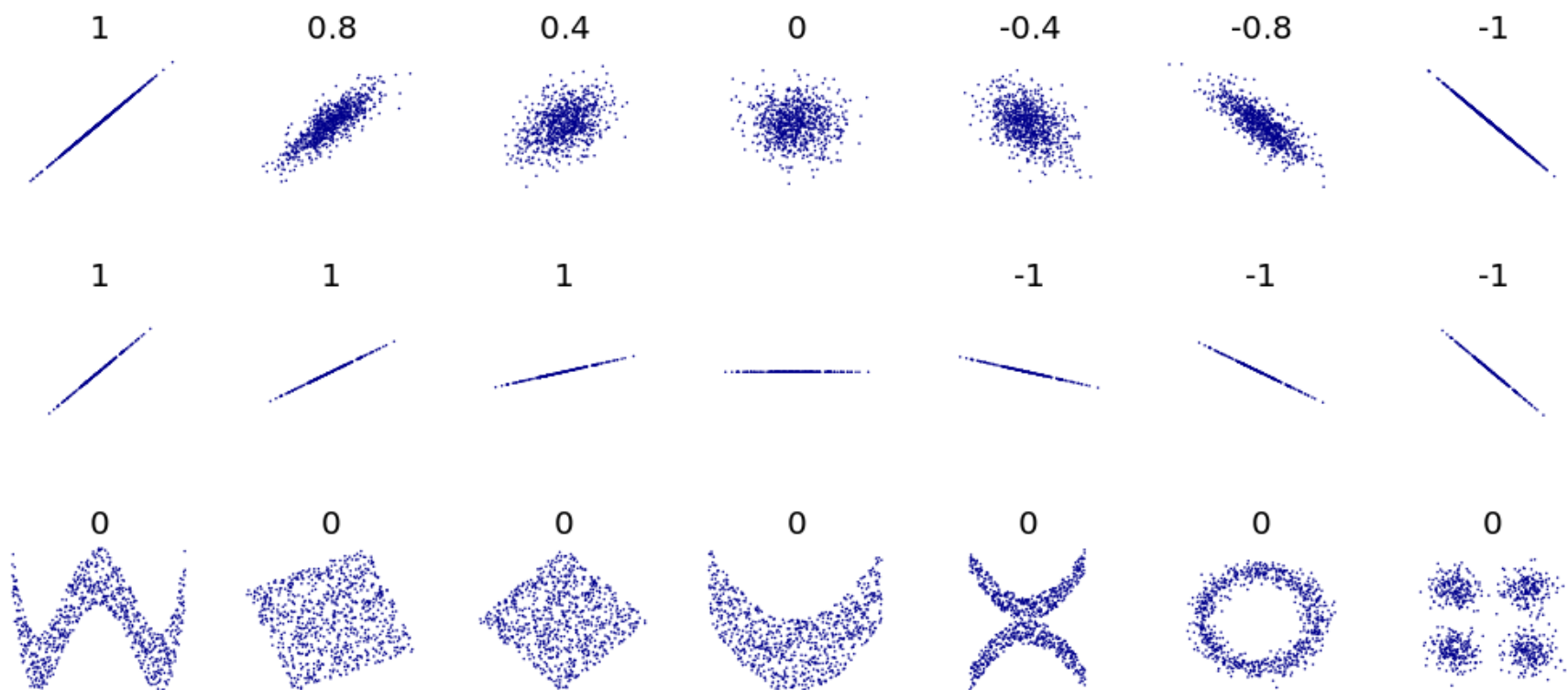
$$\bar{x} = \frac{1}{\ell} \sum_{j=1}^{\ell} x_j; \quad \bar{z} = \frac{1}{\ell} \sum_{j=1}^{\ell} z_j$$

# Коэффициент корреляции

$$\rho(x, z) = \frac{\sum_{i=1}^{\ell} (x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{\ell} (x_i - \bar{x})^2 \sum_{i=1}^{\ell} (z_i - \bar{z})^2}}$$

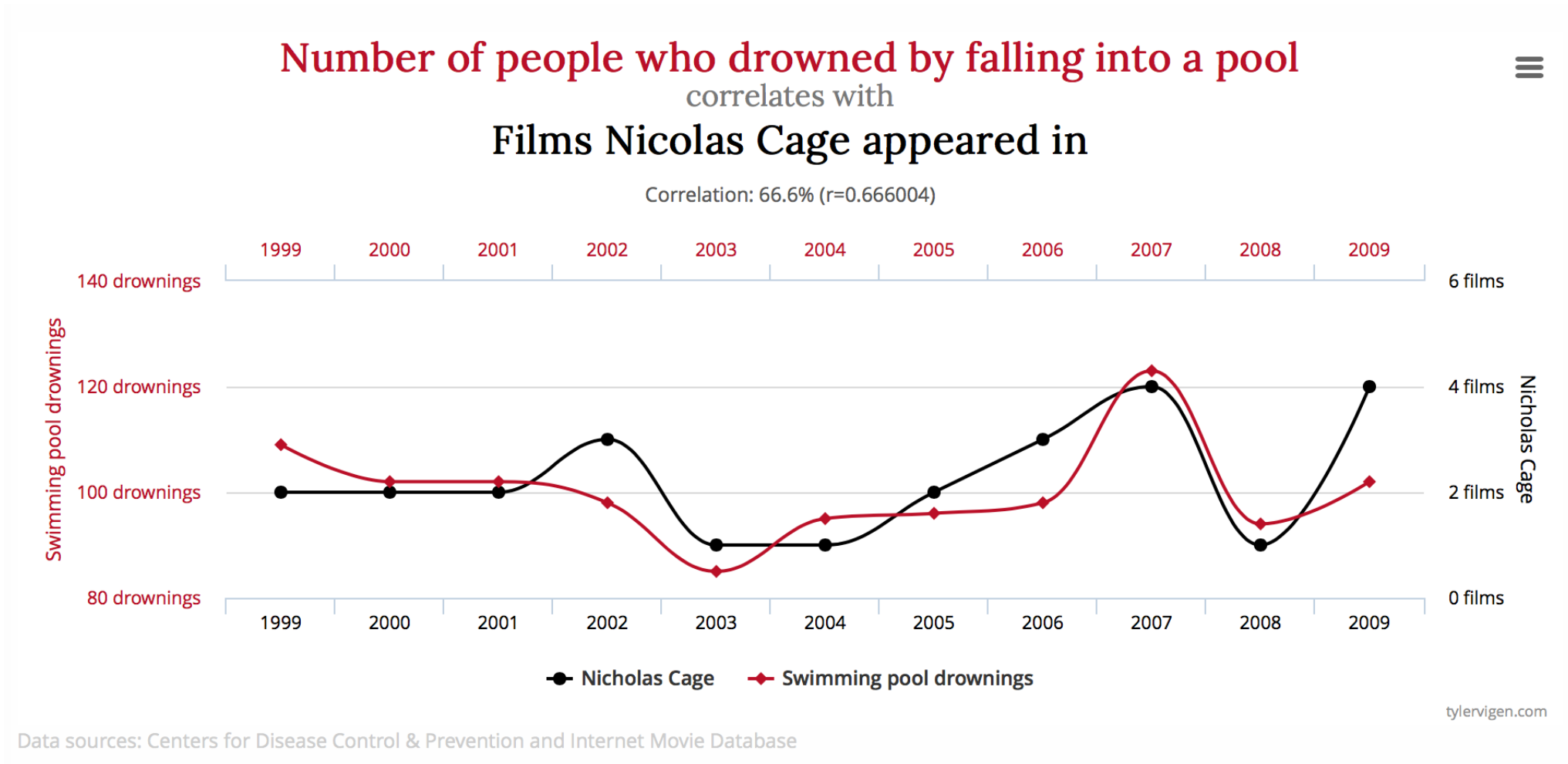
- $\rho(x, z) \in [-1, +1]$
- Очень грубо: чем ближе к +1 или -1, тем точнее выполнено уравнение
$$x = az + b$$
- Мера линейной зависимости

# Примеры

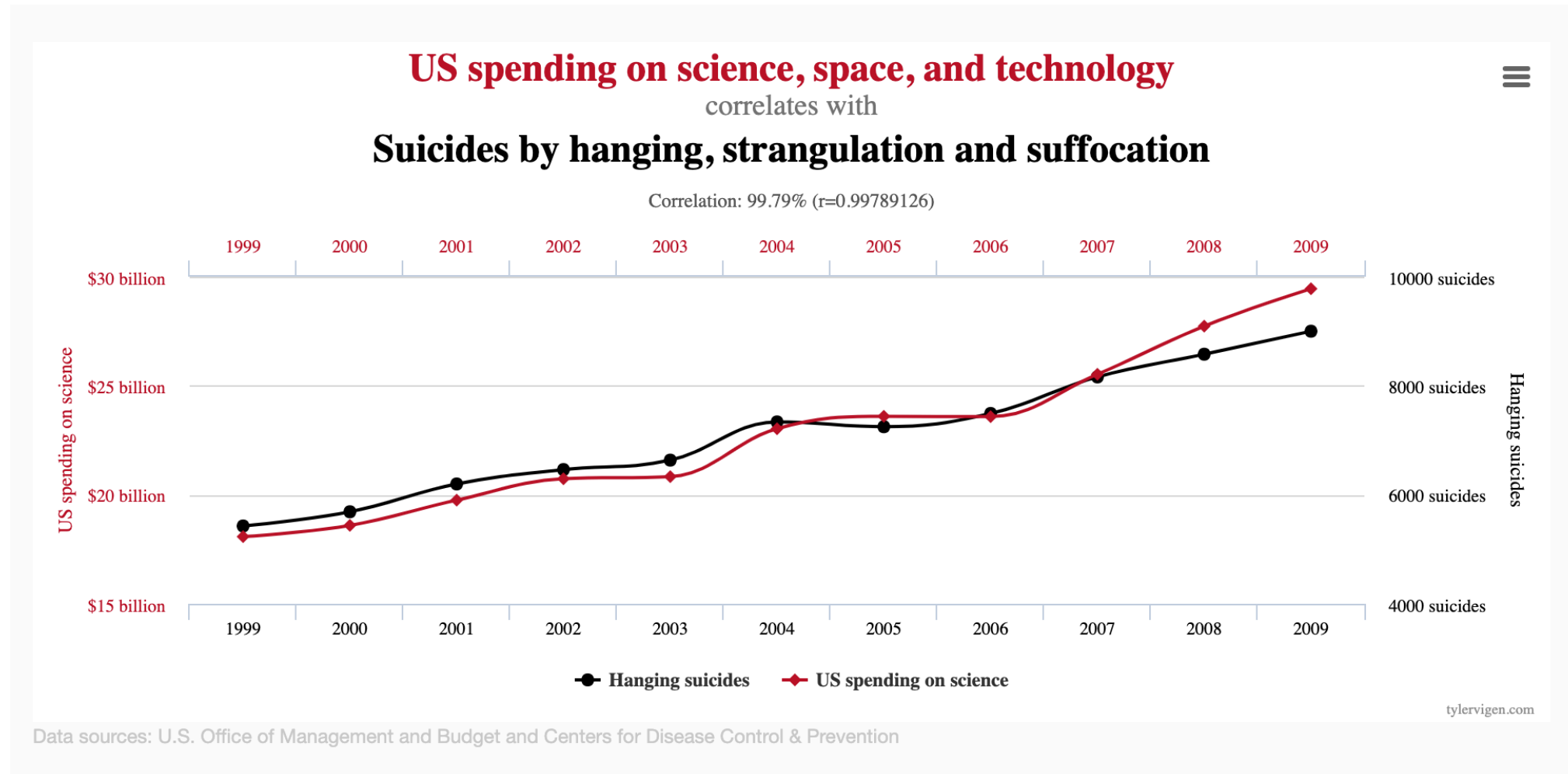




# Пример



# Пример

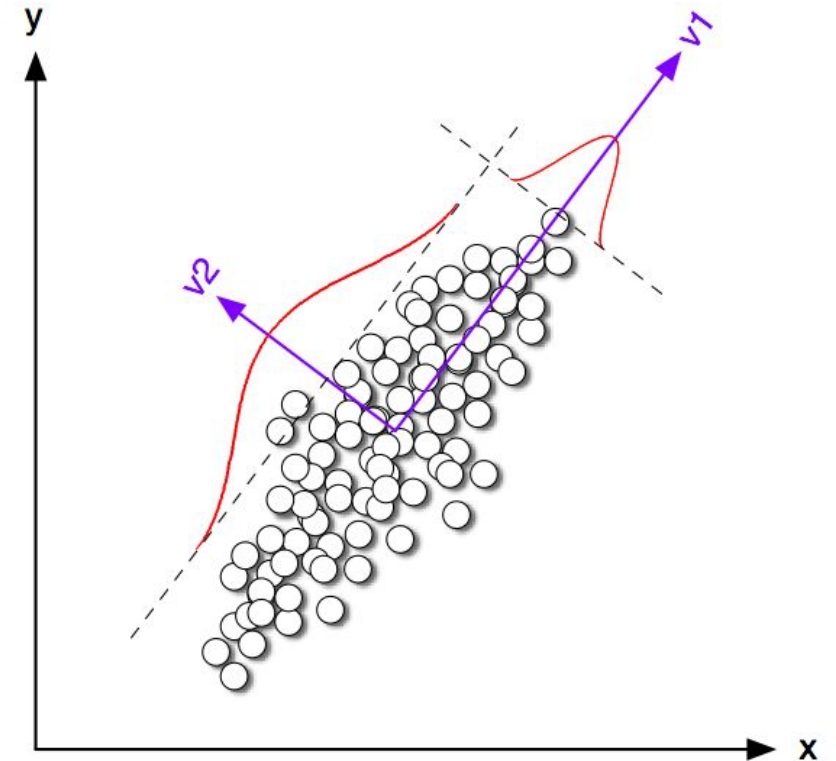


# Распространённое заблуждение

- Может показаться, что из корреляции следует причинно-следственная связь
  - Это не так!
  - Корреляция означает, что события часто происходят вместе
  - Но никак не следуют друг из друга
- 
- Больше примеров: <http://tylervigen.com/spurious-correlations>

# Коррелирующие признаки

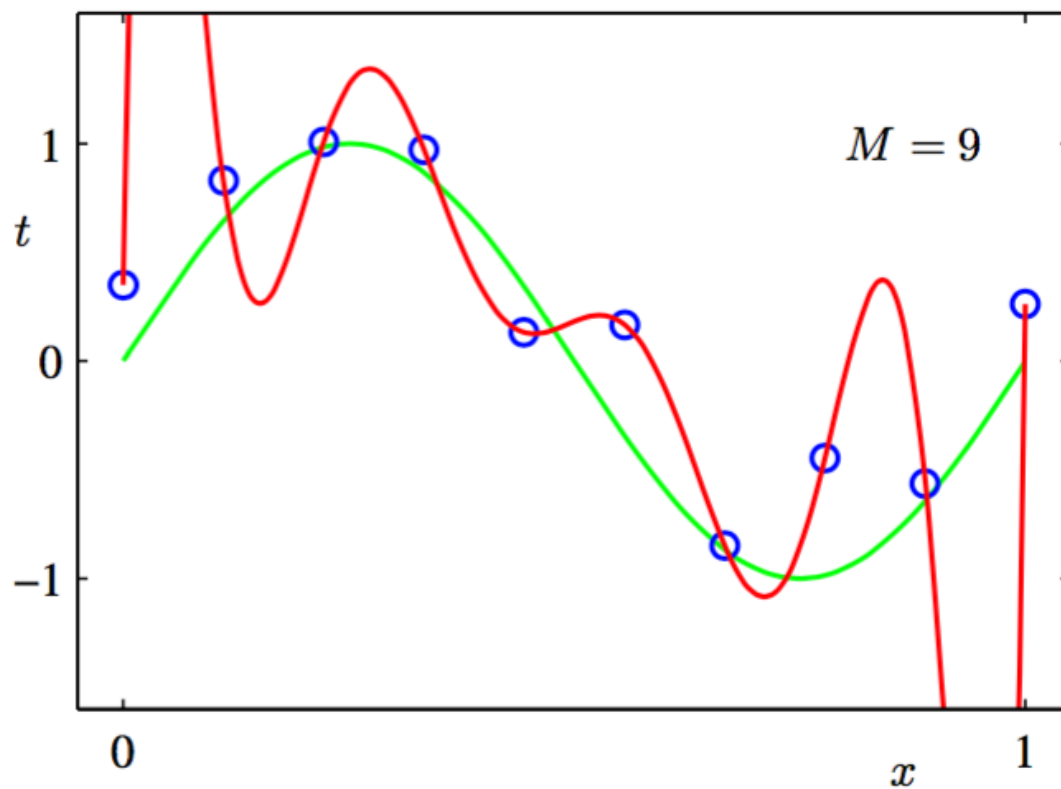
- Плохо, если есть коррелирующие признаки
- Решение: отбор признаков или их декорреляция
- В следующих лекциях



Переобучение и регуляризация

# Пример

- Один признак  $x$
- $a(x) = w_0 + w_1x + w_2x^2 + \dots + w_9x^9$



# Пример

- Коэффициенты:

$$a(x) = 0.5 + 13458922x - 43983740x^2 + \dots + 2740x^9$$

- Большие коэффициенты — симптом переобучения
- (эмпирическое наблюдение)

# Симптом переобучения

- Большие коэффициенты в линейной модели — это плохо
- Пример: предсказание роста по весу
  - $a(x) = 698x - 41714$
- Изменение веса на 0.01 кг приведет к изменению роста на 7 см
- Не похоже на правильную зависимость



# Регуляризация

- Будем штрафовать за большие веса!
- Функционал:

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

- Регуляризатор:

$$\|w\|^2 = \sum_{j=1}^d w_j^2$$

# Регуляризация

- Регуляризованный функционал ошибки:

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

- Всё ещё гладкий и выпуклый

# Коэффициент регуляризации

- $\lambda$  — новый параметр, надо подбирать
- Высокий  $\lambda$  — простые модели
- Низкий  $\lambda$  — риск переобучения
- Нужно балансировать
- Подбор  $\lambda$  — с помощью кросс-валидации

# Смысл регуляризации

- Минимизация регуляризованного функционала равносильна решению условной задачи:

$$\begin{cases} \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w \\ \|w\|^2 \leq C \end{cases}$$

# $L_1$ -регуляризация

- $L_1$ -регуляризатор:

$$\|w\|_1 = \sum_{j=1}^d |w_j|$$

- Регуляризованный функционал ошибки:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|_1 \rightarrow \min_w$$

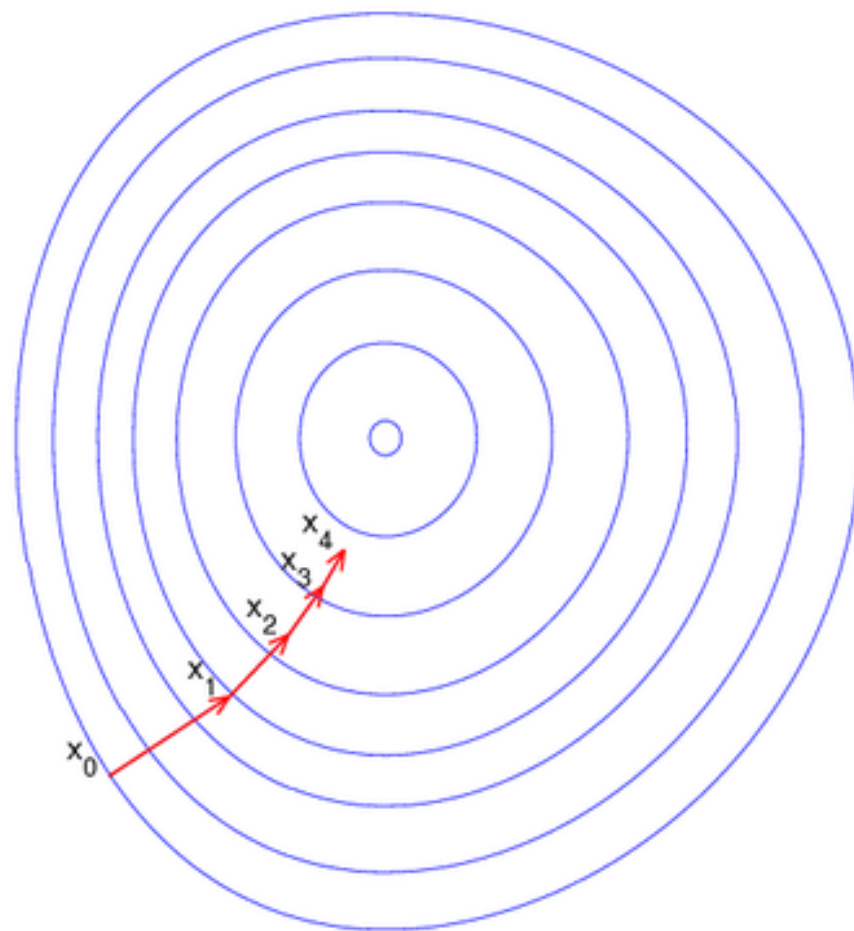
# $L_1$ -регуляризация

- Функционал становится негладким
- Сложнее оптимизировать
- Зато производится отбор признаков
- Часть весов в решении будут нулевыми

Масштабирование признаков

# Масштабирование выборки

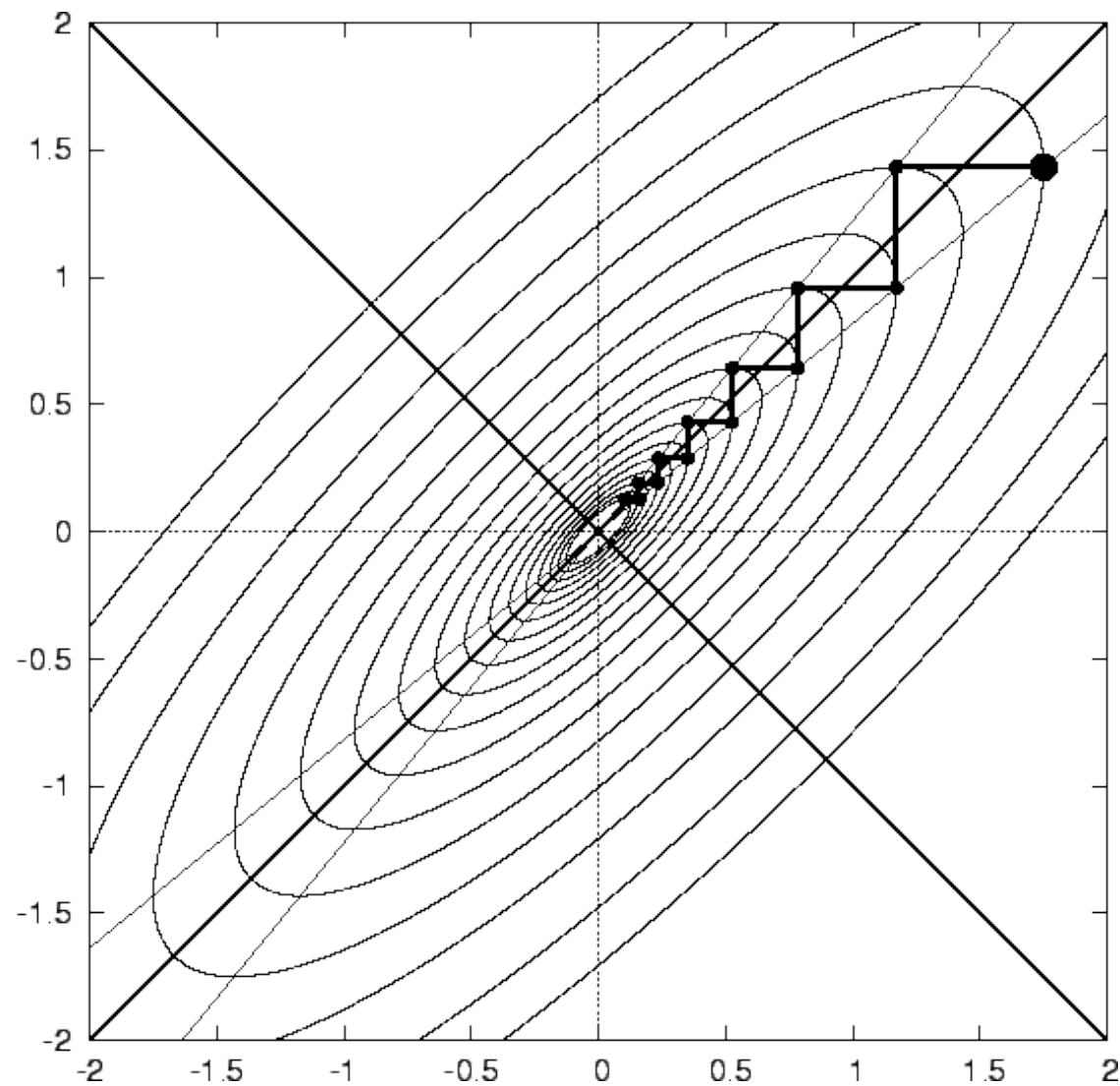
Хороший случай





# Масштабирование выборки

Плохой случай



# Масштабирование выборки

- Задача: одобряют ли заявку на грант?
- 1-й признак: сколько успешных заявок было до этого у заявителя
- 2-й признак: год рождения заявителя
- Масштаб: единицы и тысячи
- Все признаки должны иметь одинаковый масштаб

# Масштабирование выборки

- Отмасштабируем  $j$ -й признак
- Вычисляем среднее и стандартное отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (x_i^j - \mu_j)^2}$$

# Масштабирование выборки

- Отмасштабируем  $j$ -й признак
- Вычтем из каждого значения признака среднее и поделим на стандартное отклонение:

$$x_i^j := \frac{x_i^j - \mu_j}{\sigma_j}$$

# Важность признаков

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

| 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,

площадь<sup>2</sup>, этаж<sup>2</sup>, число комнат<sup>2</sup>,

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

# Модель линейной классификации

# Классификация

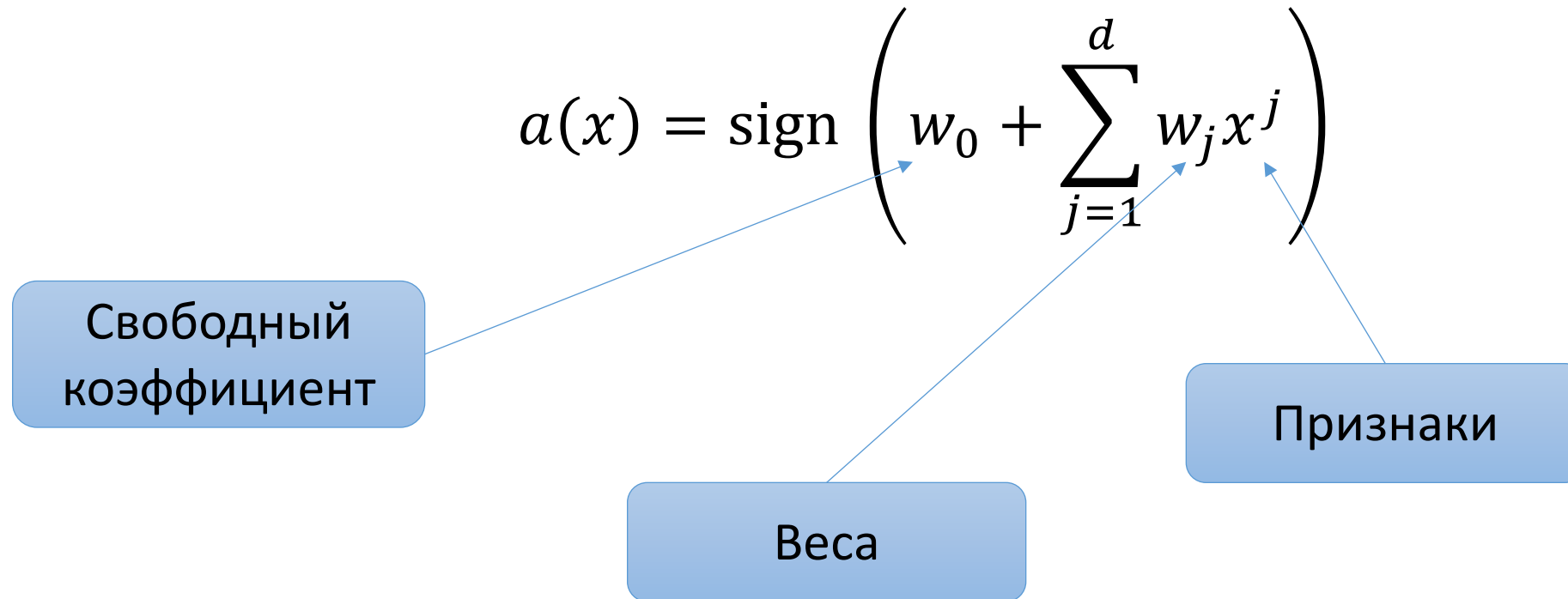
- $\mathbb{Y} = \{-1, +1\}$
- $-1$  — отрицательный класс
- $+1$  — положительный класс
- $a(x)$  должен возвращать одно из двух чисел



# Линейный классификатор

$$a(x) = \text{sign} \left( w_0 + \sum_{j=1}^d w_j x^j \right)$$

# Линейный классификатор



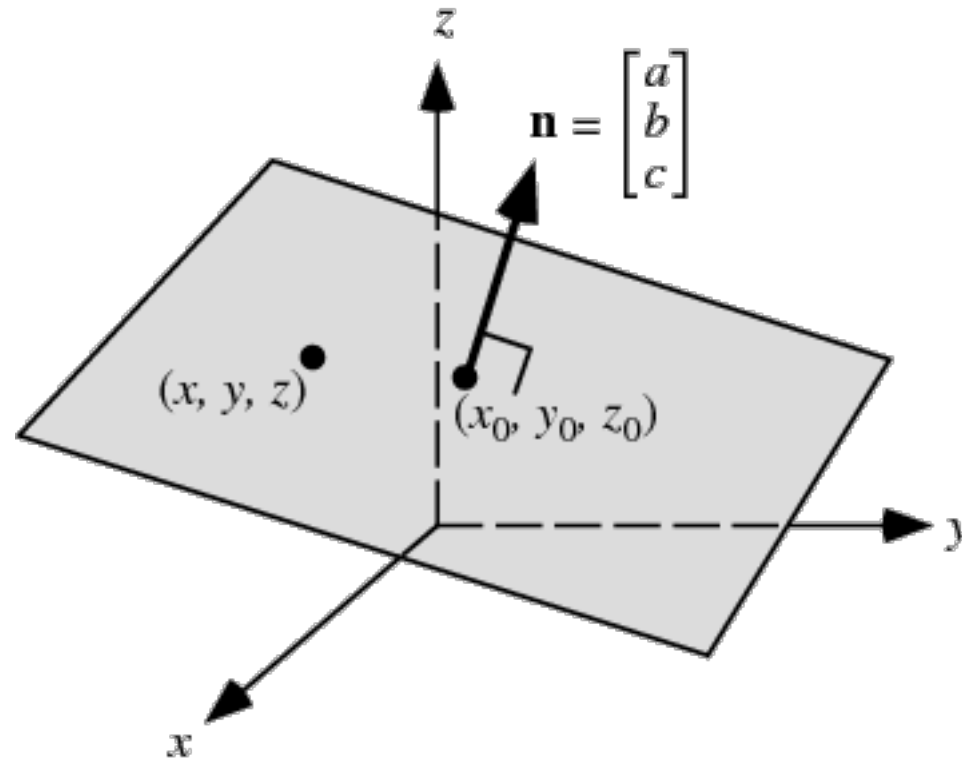
# Линейный классификатор

- Добавим единичный признак

$$a(x) = \text{sign} \sum_{j=1}^{d+1} w_j x^j = \text{sign} \langle w, x \rangle$$

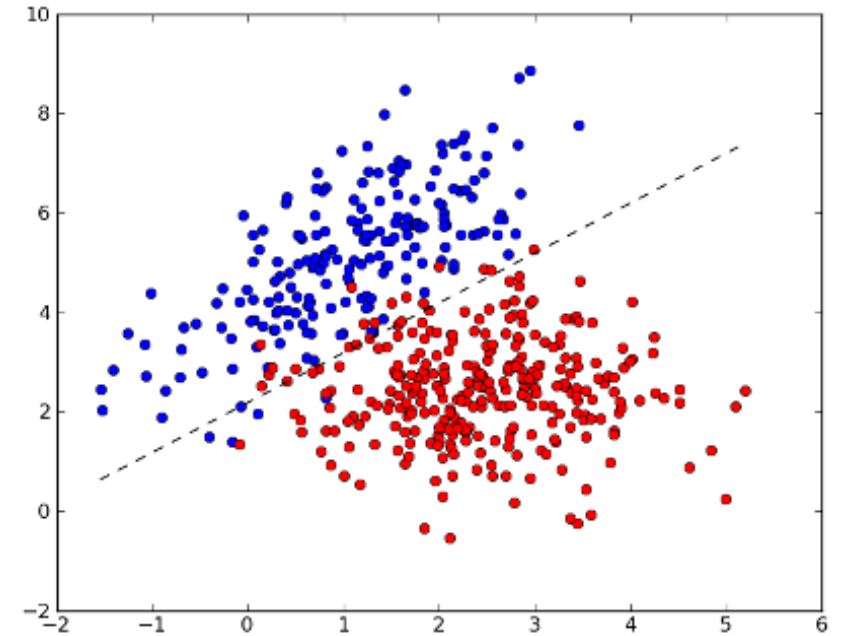
# Геометрия линейного классификатора

Уравнение гиперплоскости:  $\langle w, x \rangle = 0$



# Геометрия линейного классификатора

- Линейный классификатор проводит гиперплоскость
- $\langle w, x \rangle < 0$  — объект «слева» от неё
- $\langle w, x \rangle > 0$  — объект «справа» от неё



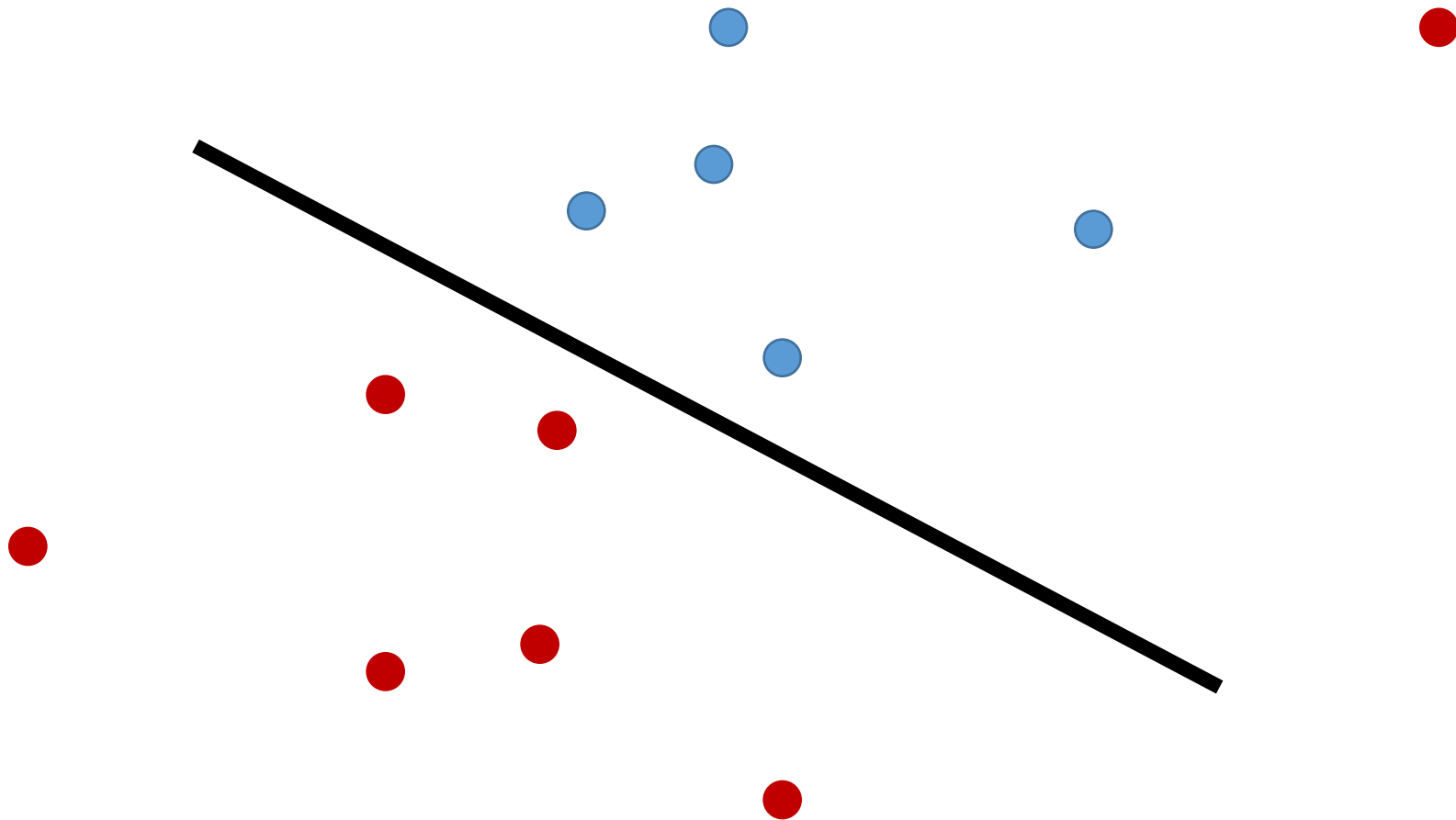
# Геометрия линейного классификатора

- Расстояние от точки до гиперплоскости  $\langle w, x \rangle = 0$ :

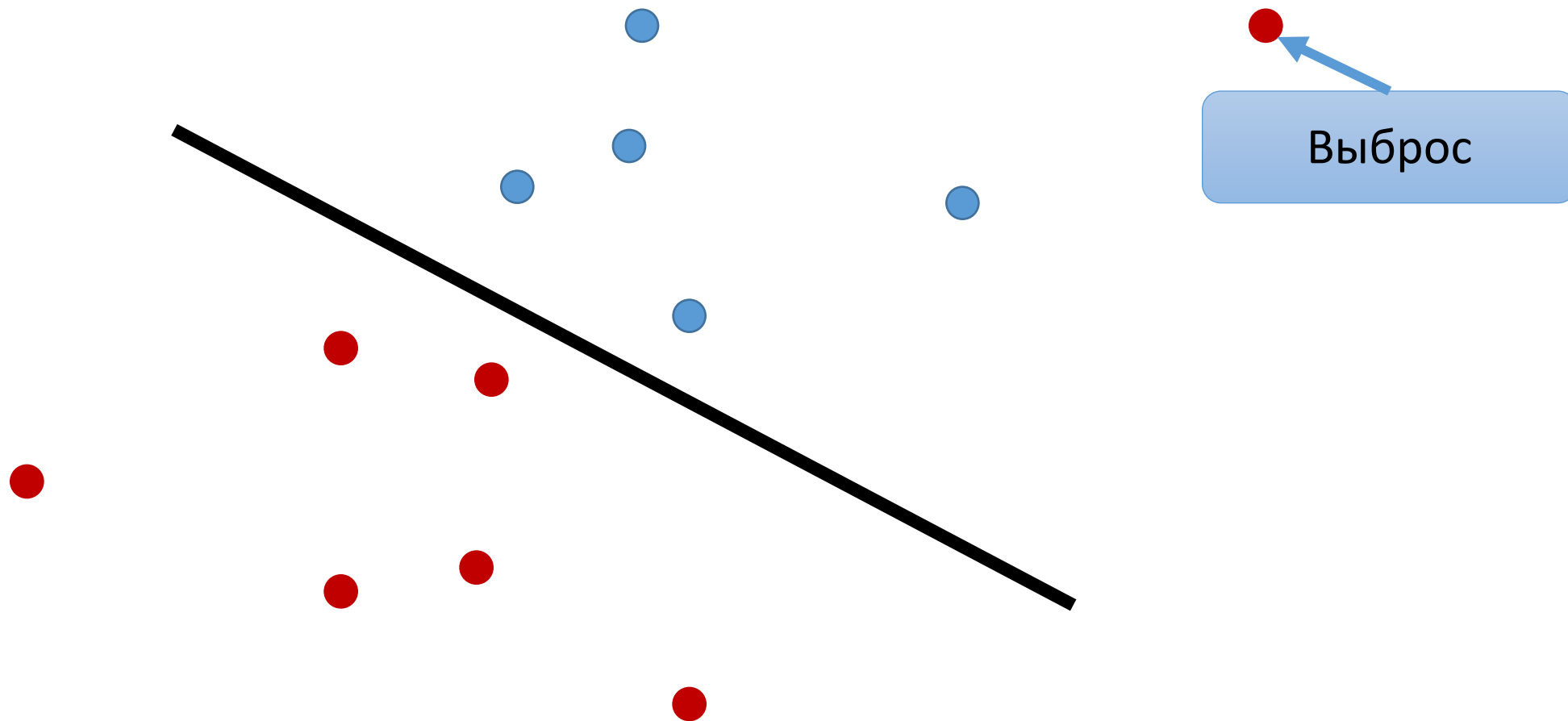
$$\frac{|\langle w, x \rangle|}{\|w\|}$$

- Чем больше  $\langle w, x \rangle$ , тем дальше объект от разделяющей гиперплоскости

# Геометрия линейного классификатора



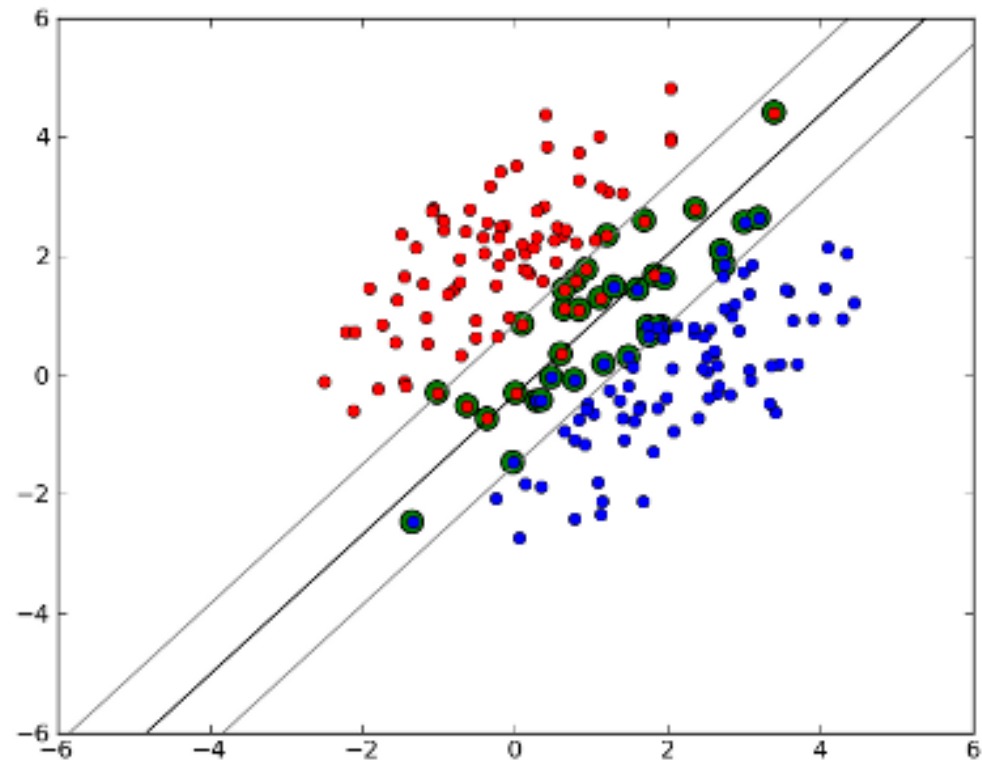
# Геометрия линейного классификатора





# Отступ

- $M_i = y_i \langle w, x_i \rangle$
- $M_i > 0$  — классификатор дает верный ответ
- $M_i < 0$  — классификатор ошибается
- Чем дальше отступ от нуля, тем больше уверенности



# Линейный классификатор

- Линейный классификатор разделяет два класса гиперплоскостью
- Чем больше отступ по модулю, тем дальше объект от гиперплоскости
- Знак отступа говорит о корректности предсказания

Функционал ошибки для  
классификации

# Линейная регрессия

- Квадратичное отклонение:

$$L(a, y) = (a - y)^2$$

- Абсолютное отклонение:

$$L(a, y) = |a - y|$$

# Линейная классификация

- Доля **неправильных** ответов:

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

# Линейная классификация

| $a(x)$    | $y$       |
|-----------|-----------|
| -1        | -1        |
| +1        | +1        |
| -1        | -1        |
| <b>+1</b> | <b>-1</b> |
| +1        | +1        |

- Доля неправильных ответов:

$$\frac{1}{5} = 0.2$$

# Линейная классификация

- Доля **правильных** ответов:

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

- На английском: **accuracy**

# Линейная классификация

- Доля **правильных** ответов:

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

- На английском: **accuracy**
- ВАЖНО: не переводите это как «точность»!

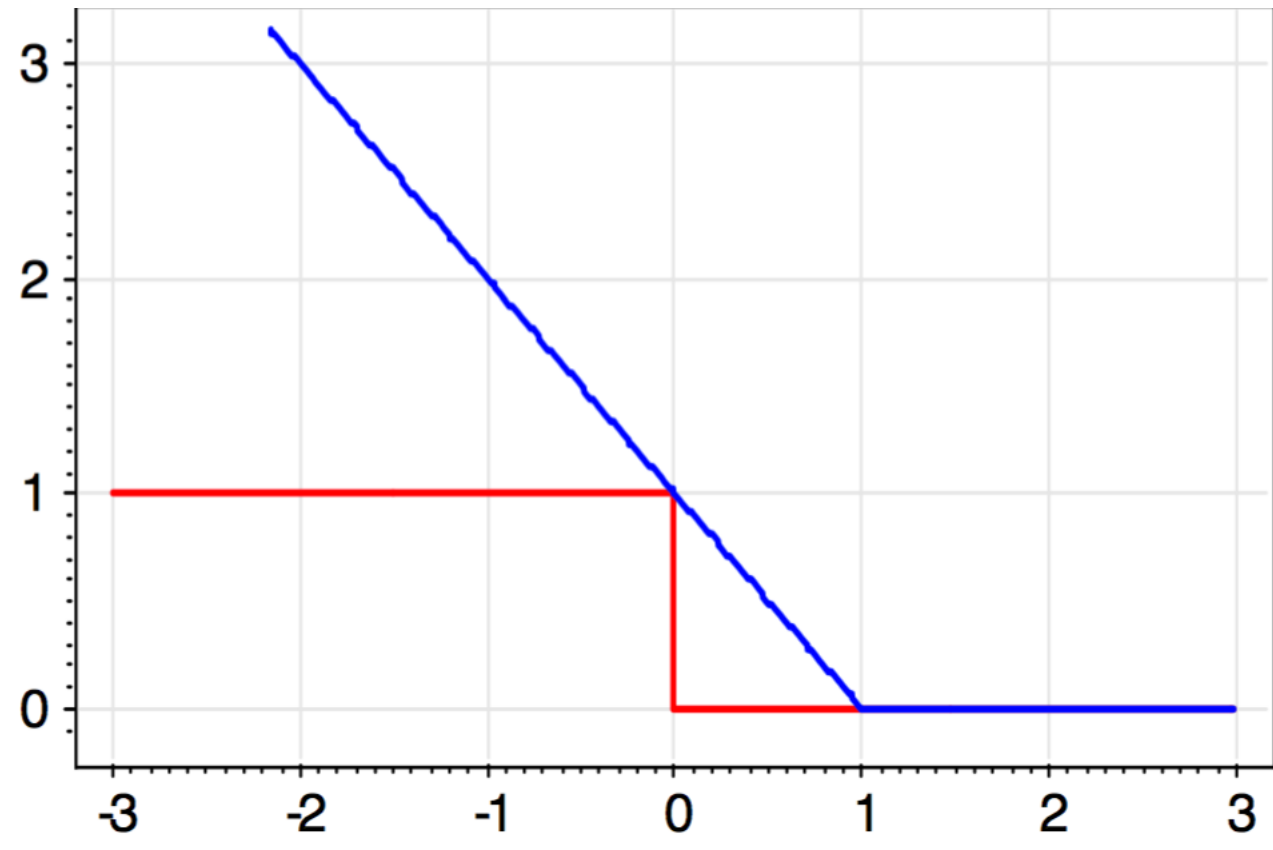


# Линейная классификация

- Доля неправильных ответов (через отступ):

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [y_i \underbrace{\langle w, x_i \rangle}_{M_i} < 0]$$

# Пороговая функция потерь



# Линейная классификация

- Доля неправильных ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \underbrace{[y_i \langle w, x_i \rangle < 0]}_{M_i}$$

- Разрывная функция
- Непонятно, как оптимизировать

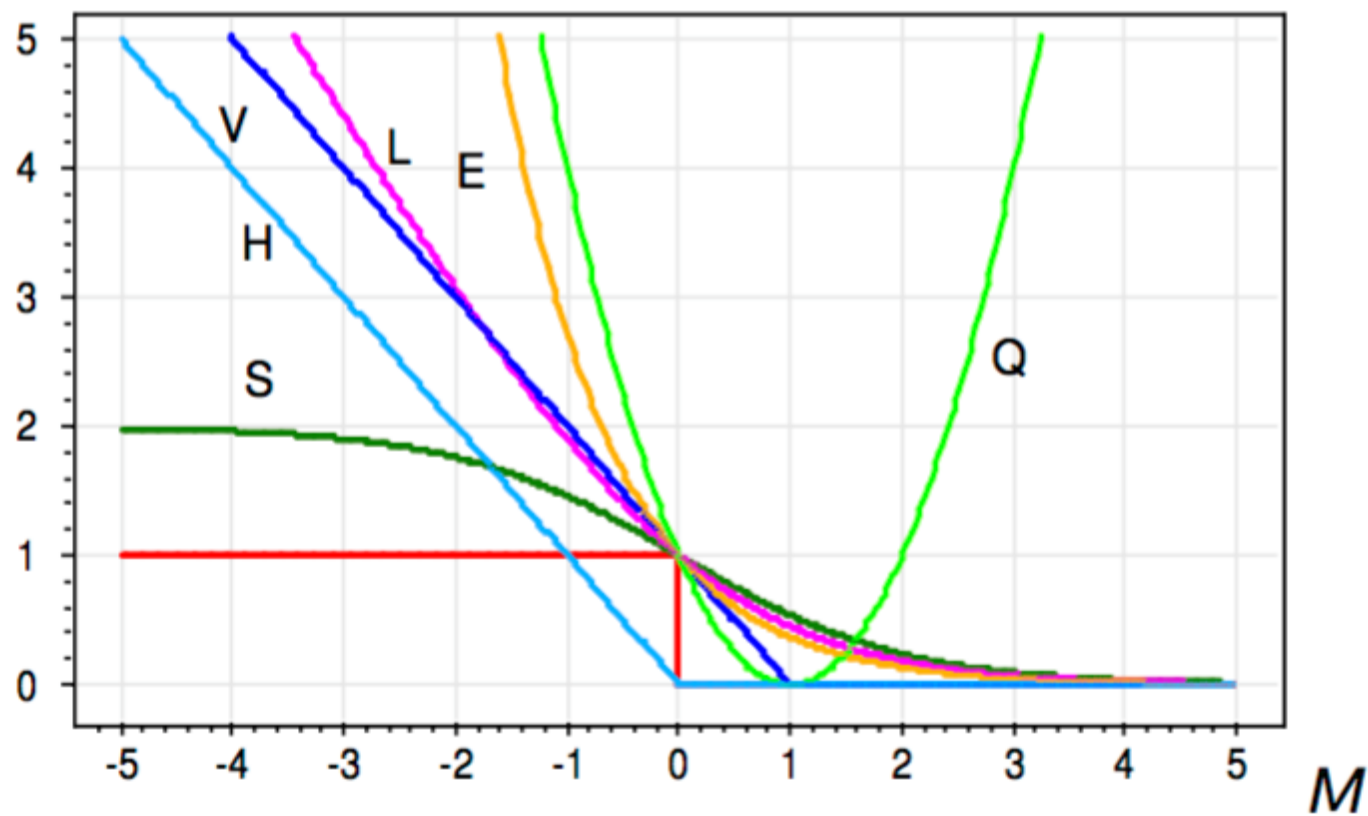
# Оценка функции потерь

- Возьмем любую гладкую оценку пороговой функции:

$$[M < 0] \leq \tilde{L}(M) = \tilde{L}(y\langle w, x \rangle)$$

Гладкая функция – это функция, имеющая непрерывную производную на всей области определения.

# Примеры оценок



# Примеры оценок

- $\tilde{L}(M) = \log_2(1 + \exp(-M))$  — логистическая
- $\tilde{L}(M) = \exp(-M)$  — экспоненциальная
- $\tilde{L}(M) = \max(0, 1 - M)$  — кусочно-линейная

# Оценка функции потерь

- Возьмем любую гладкую оценку пороговой функции:

$$[M < 0] \leq \tilde{L}(M)$$

- Оценим через нее функционал ошибки:

$$Q(a, X) \leq \tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(M_i)$$

# Оценка функции потерь

$$\frac{1}{\ell} \sum_{i=1}^{\ell} [M_i < 0] \leq \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(M_i) \rightarrow \min_a$$

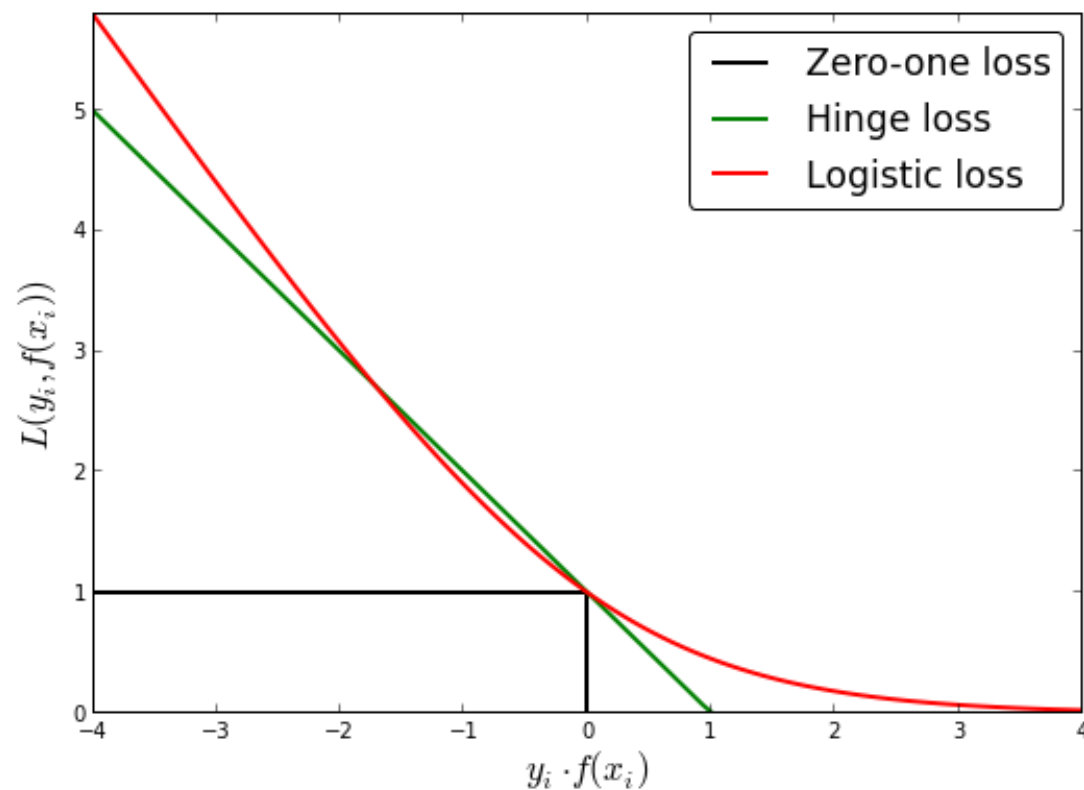
Минимизируем  
верхнюю оценку

Надеемся, что доля  
ошибок тоже  
уменьшится



# Примеры оценок

- $\tilde{L}(a, y) = \ln(1 + \exp(-ya))$  — логистическая



# Логистическая функция потерь

$$\tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \log_2(1 + \exp(-y_i \langle w, x_i \rangle))$$

1. Выписали индикатор ошибки через отступ
2. Заменяли пороговую функцию потерь на гладкую функцию

# Обучение

- Обучение — с помощью любых методов оптимизации
- Например, градиентный спуск:

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

- Борьба с переобучением: регуляризация (так же, как в линейной регрессии)

# Логистическая регрессия

# Логистическая регрессия

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \log_2(1 + \exp(-y_i \langle w, x_i \rangle)) \rightarrow \min_w$$

# Оценивание вероятностей

- $P(y = 1 \mid x) = \pi(x)$

# Оценивание вероятностей

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

# Оценивание вероятностей

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



# Оценивание вероятностей

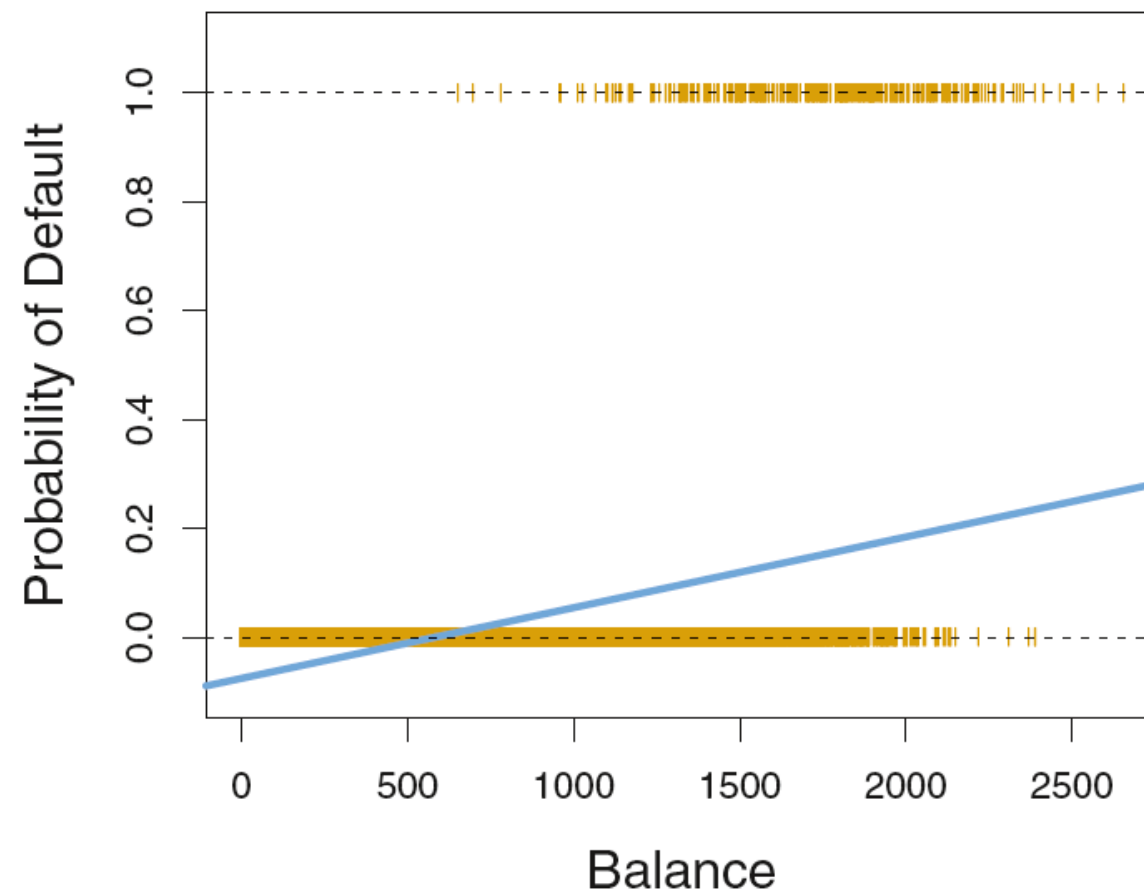
- Прогнозирование оттока клиентов
- Медицинская диагностика
- Поисковое ранжирование (насколько веб-страница соответствует запросу?)

# Оценивание вероятностей

- $P(y = 1 \mid x) = \pi(x)$
- $\pi(x)$  — вещественное число
- Классификатор не подходит

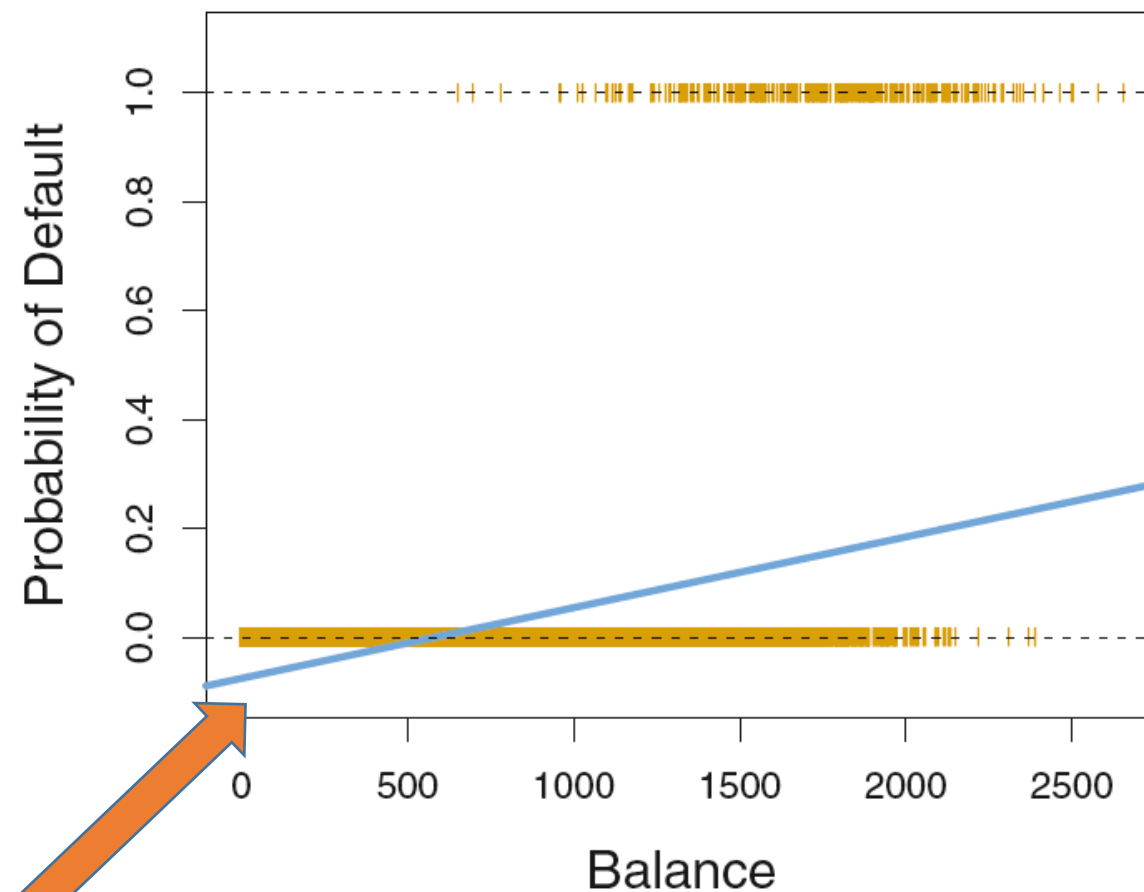
# Регрессия?

- $\pi(x) \approx \langle w, x \rangle = w_1 x + w_0$



# Регрессия?

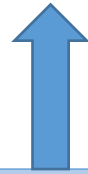
- $\pi(x) \approx \langle w, x \rangle = w_1 x + w_0$



Отрицательная вероятность o\_0

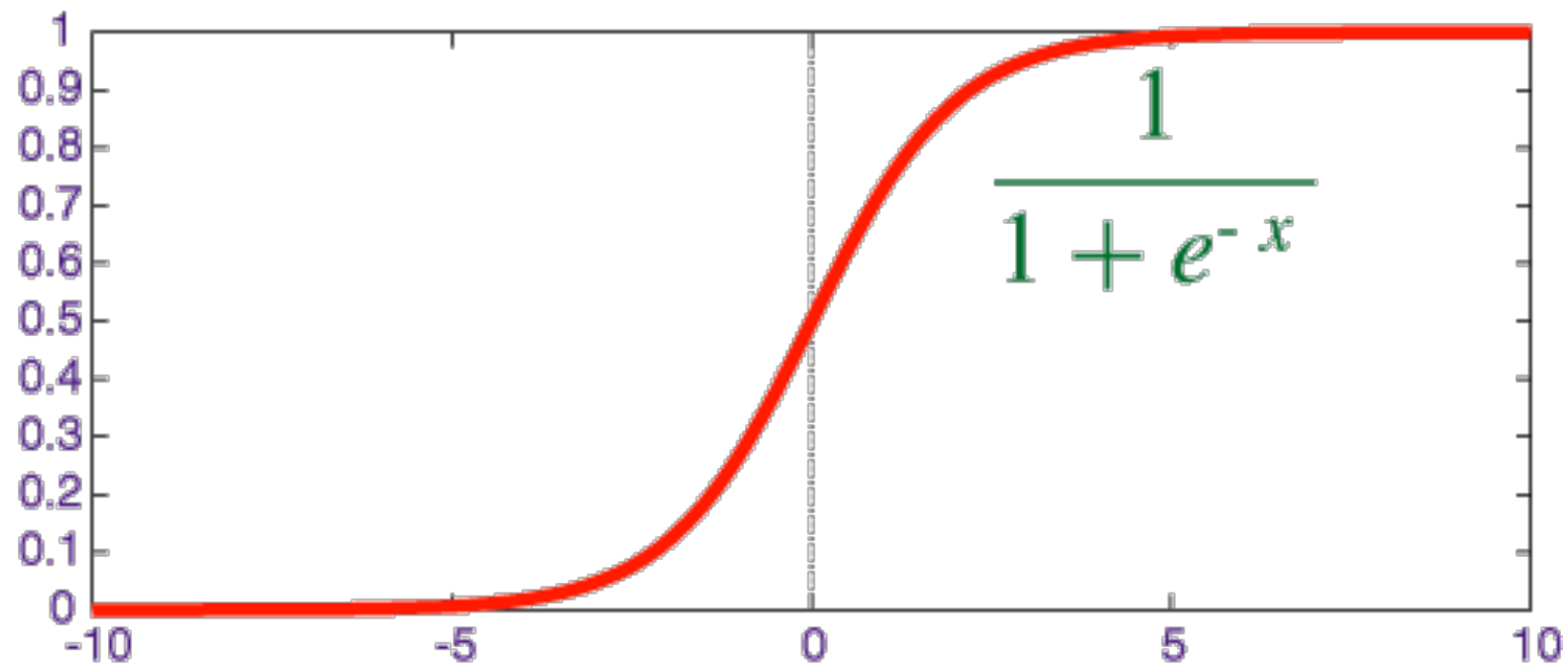
# Регрессия?

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



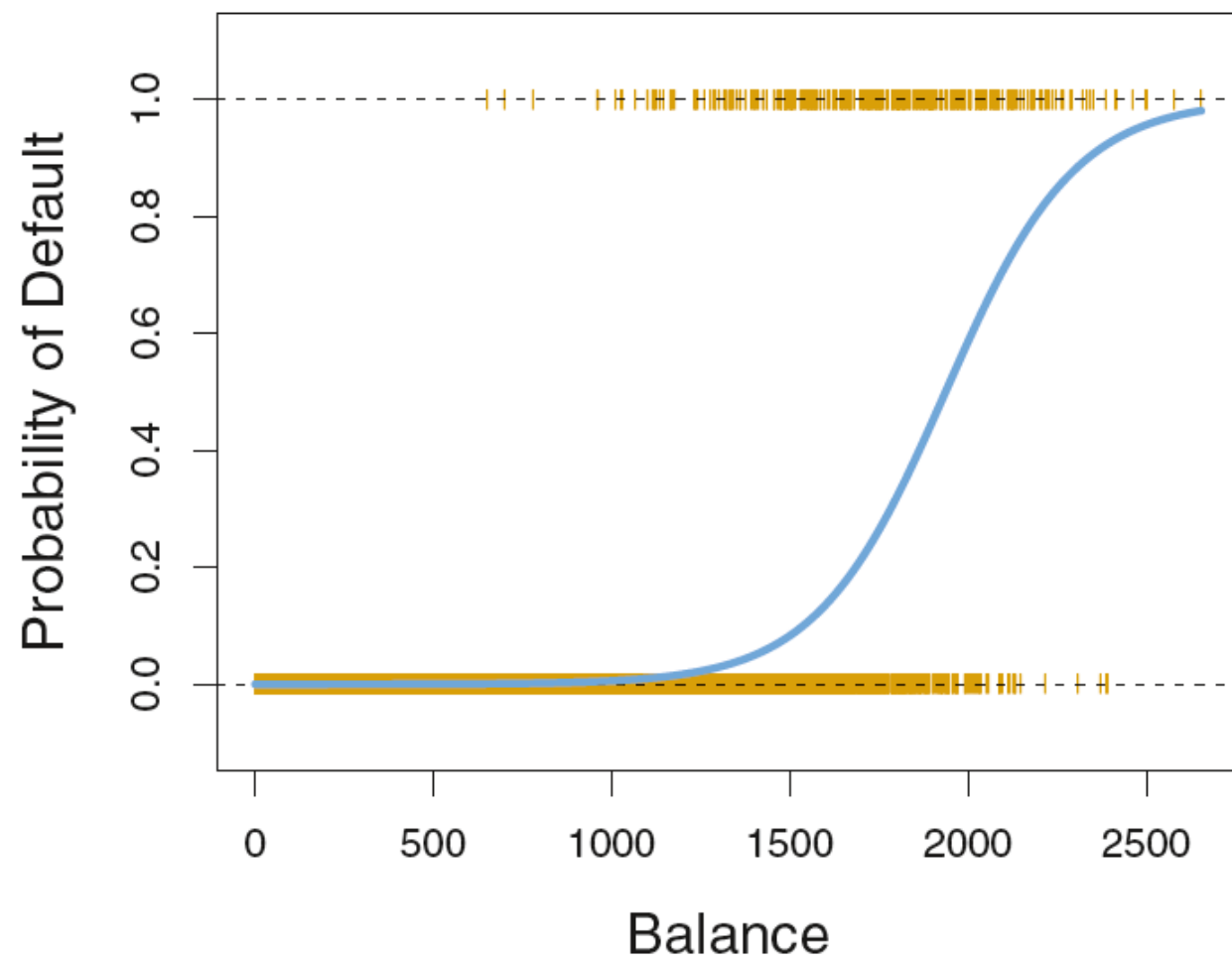
Сигмоида

# Сигмоида



# Логистическая регрессия

- $\pi(x) \approx \sigma(\langle w, x \rangle)$



# Логистическая регрессия

- Как оптимизировать?
- Если  $y_i = +1$ , то  $\langle w, x_i \rangle \rightarrow +\infty$
- Если  $y_i = -1$ , то  $\langle w, x_i \rangle \rightarrow -\infty$



# Логистическая регрессия

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

# Логистическая регрессия

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

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

# Логистическая регрессия

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

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

- Слишком слабый штраф
- Если  $y_i = +1$  и  $\sigma(\langle w, x_i \rangle) = 0$ , то штраф = 1

# Логистическая регрессия

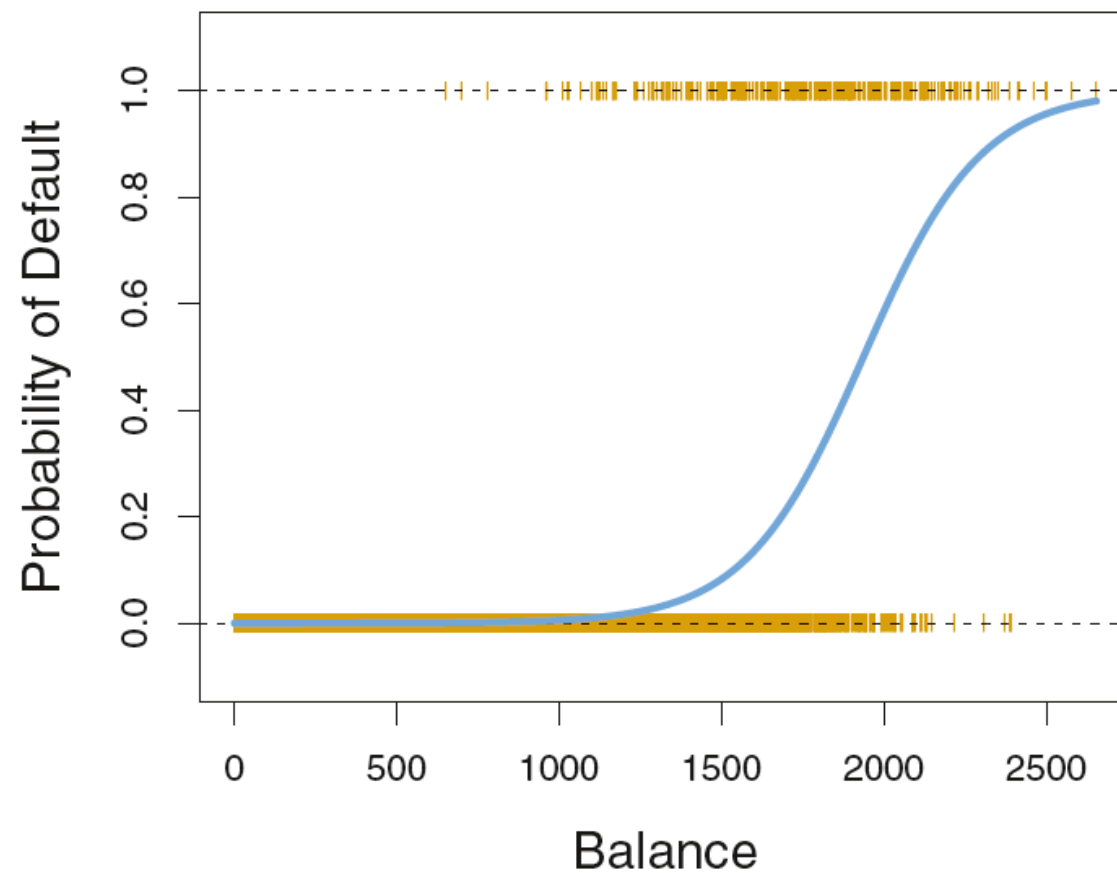
- Как оптимизировать?
- Если  $y_i = +1$ , то  $\sigma(\langle w, x_i \rangle) \rightarrow 1$
- Если  $y_i = -1$ , то  $\sigma(\langle w, x_i \rangle) \rightarrow 0$

$$\sum_{i=1}^{\ell} \{ [y_i = 1] \log_2 \sigma(\langle w, x_i \rangle) + [y_i = -1] \log_2 (1 - \sigma(\langle w, x_i \rangle)) \} \rightarrow \max_w$$

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



# Логистическая регрессия



# Логистическая регрессия

- Если вспомнить арифметику, то получим эквивалентную задачу:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \log_2(1 + \exp(-y_i \langle w, x_i \rangle)) \rightarrow \min_w$$

# Логистическая регрессия

- Линейная модель классификации:  $a(x) = \text{sign } \langle w, x \rangle$
- Позволяет оценивать вероятности:  $\pi(x) = \sigma(\langle w, x \rangle)$
- Обучение: градиентный спуск

# 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

- 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:

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

- 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

- 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    |



# Резюме

- Линейные классификаторы разделяют классы гиперплоскостью
- Логистическая регрессия — классификация и оценка вероятности
- Качество классификации: доля правильных ответов, точность и полнота

# Метрики качества регрессии

# Среднеквадратичная ошибка

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

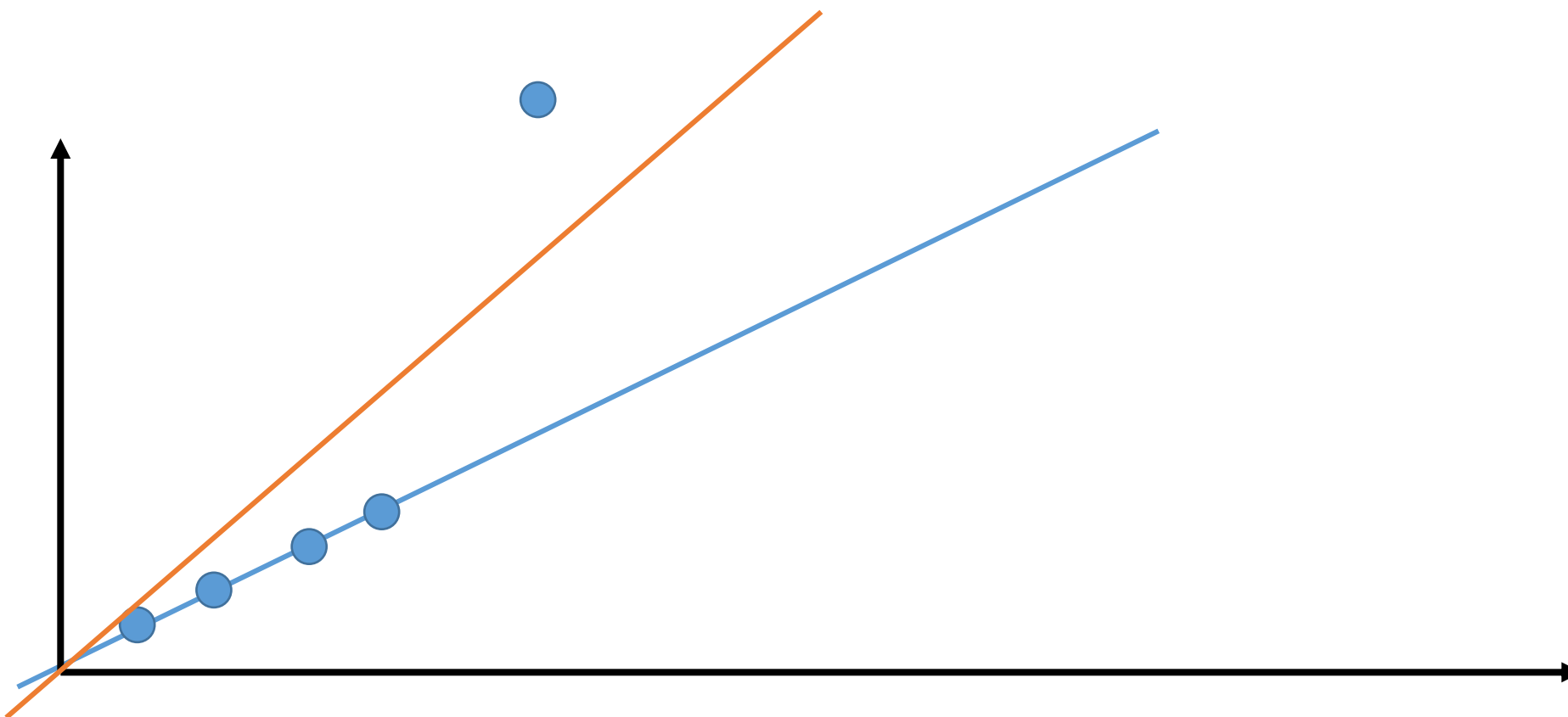
- Легко минимизировать
- Сильно штрафует за большие ошибки

# Средняя абсолютная ошибка

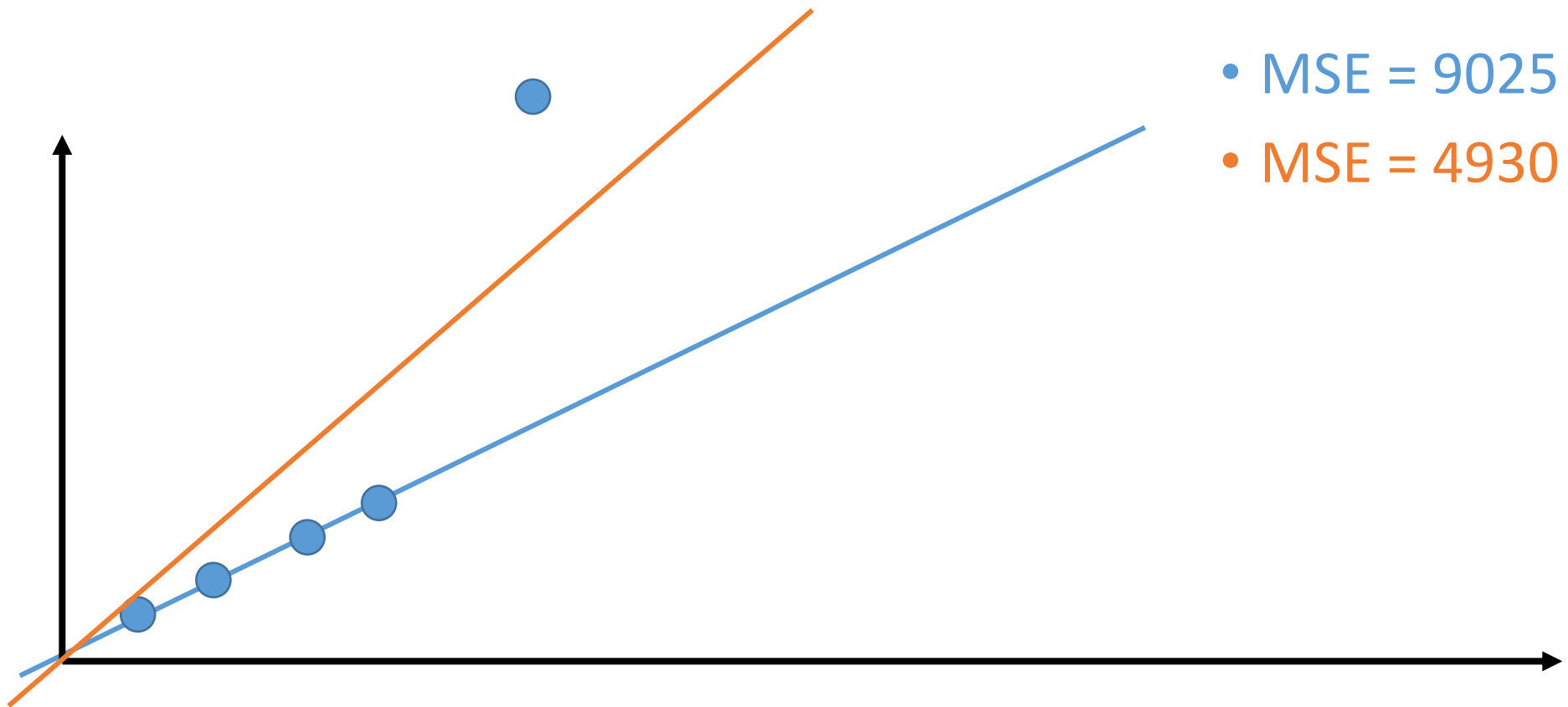
$$\text{MAE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

- Сложнее минимизировать
- Выше устойчивость к выбросам

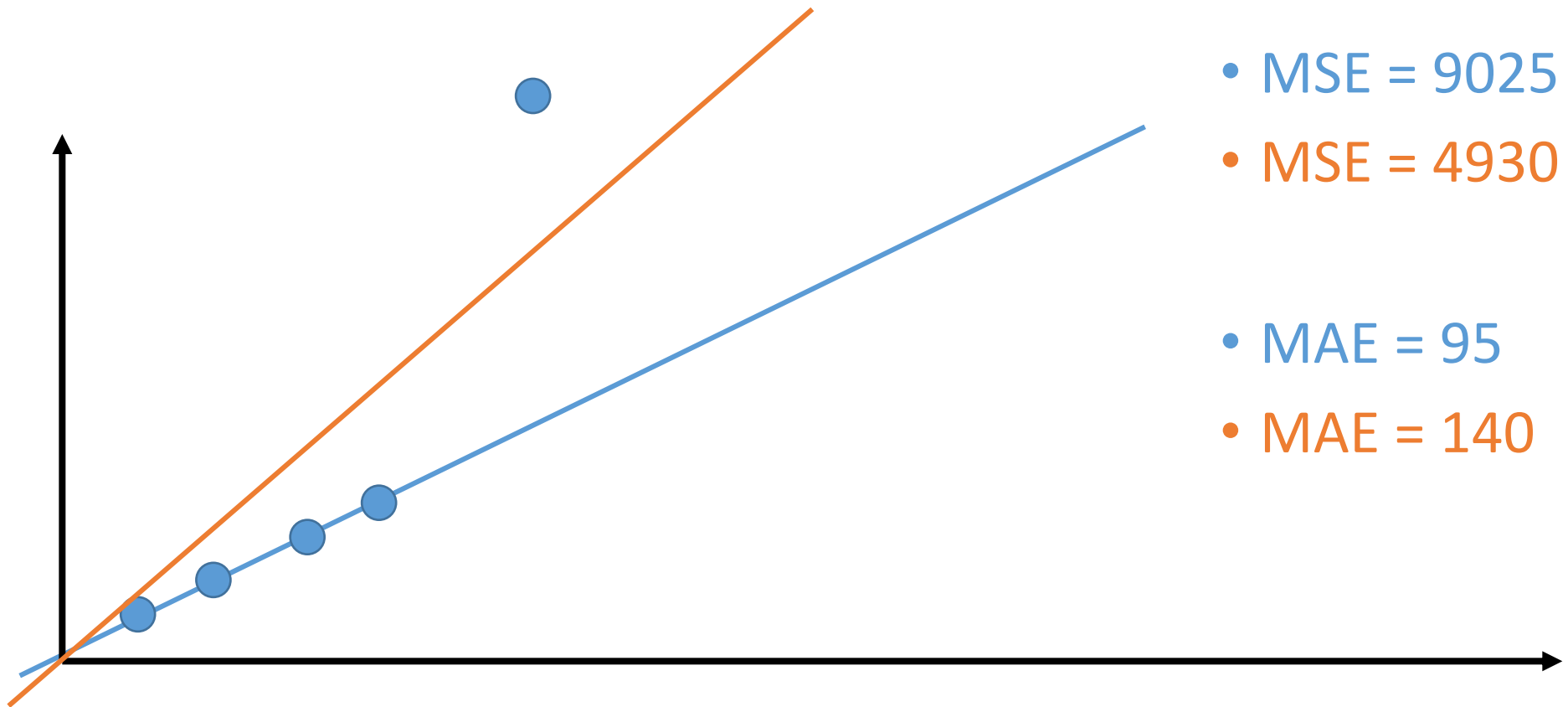
# Средняя абсолютная ошибка



# Средняя абсолютная ошибка



# Средняя абсолютная ошибка



# Устойчивые оценки

- Оценка среднего значения — матожидание
- Оценка разброса — дисперсия



# Математическое ожидание

- Характеризует среднее значение случайной величины

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

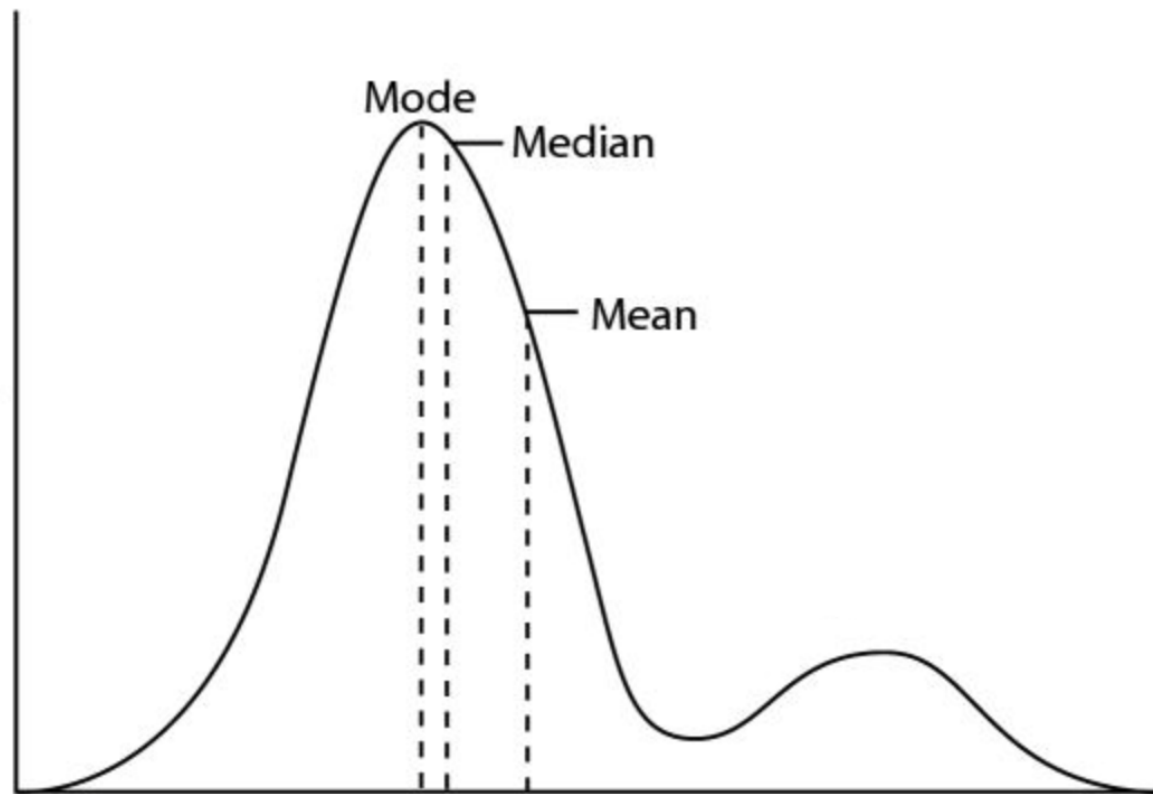
# Медиана

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

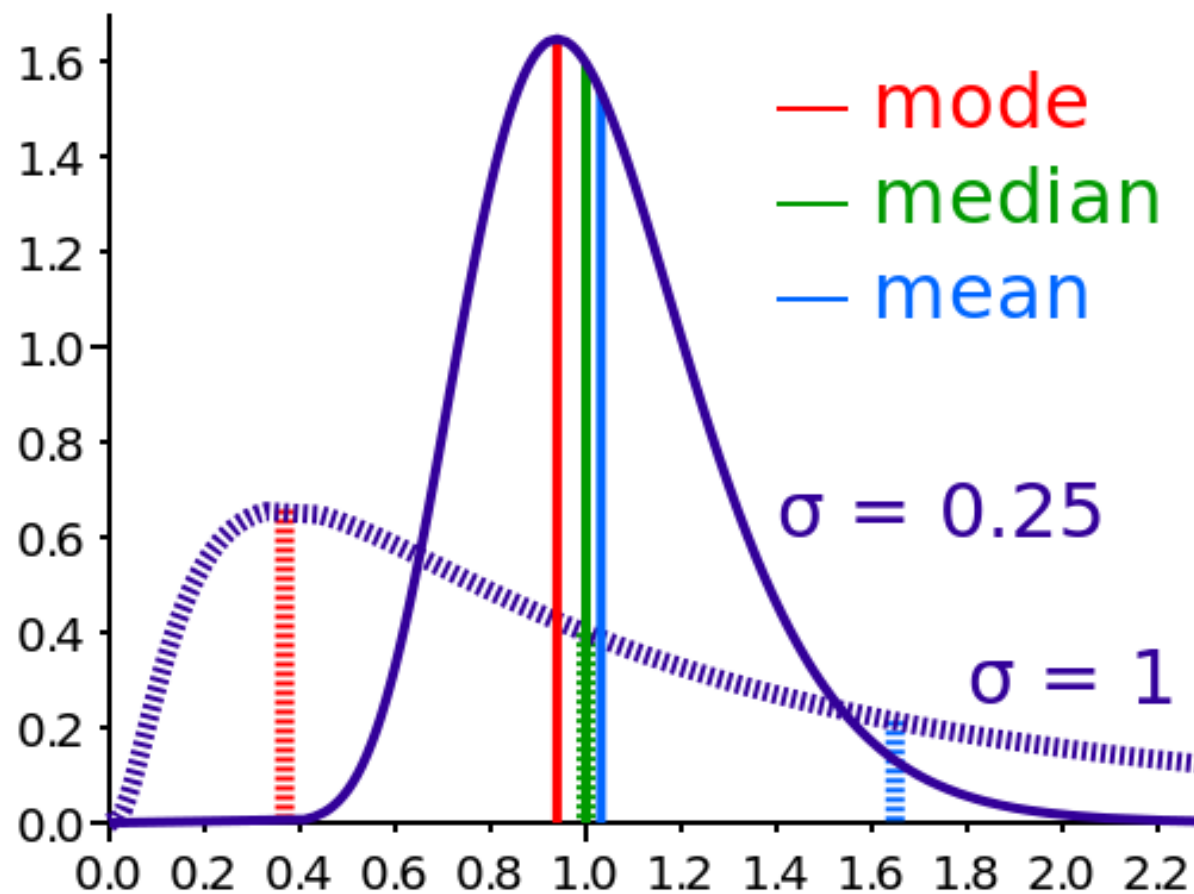
# Мода

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

# Центральная величина



# Центральная 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

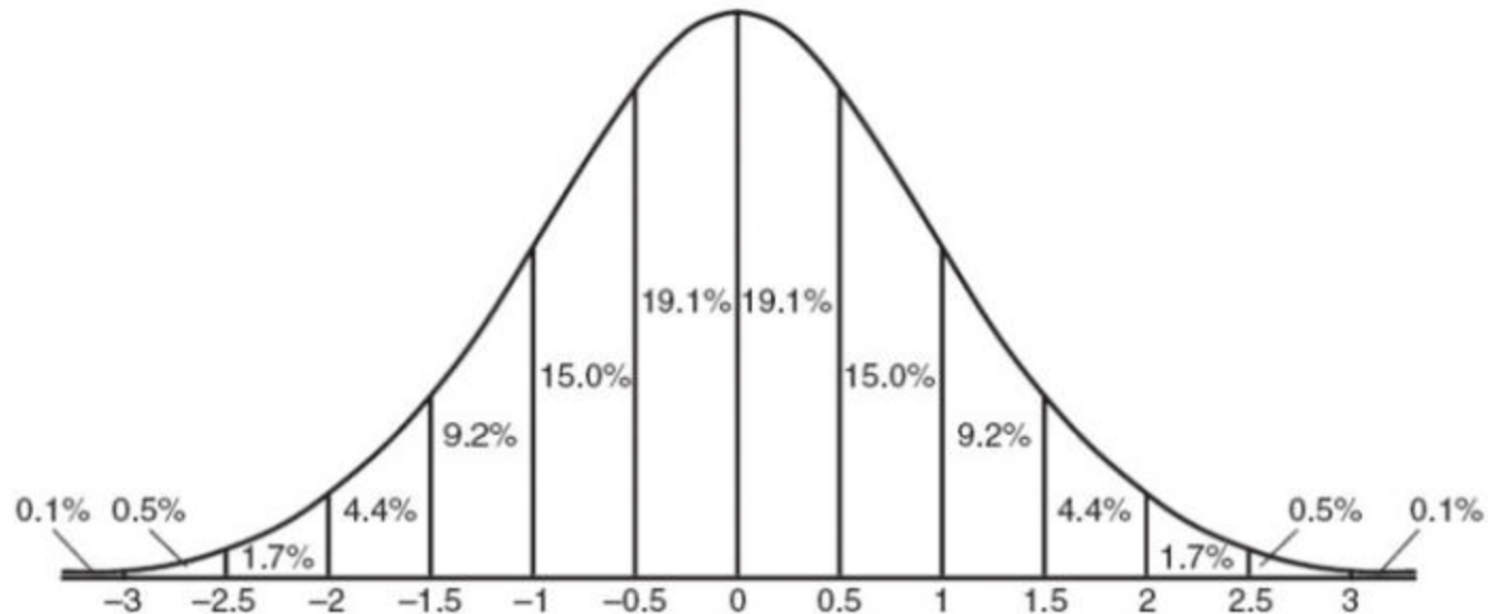
# Дисперсия

$$\mathbb{E}(\xi - \mathbb{E}\xi)^2 \approx \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Опросили 100 человек
- 99 имеют доход 10.000 рублей
- 1 имеет доход 1.000.000 рублей
- Дисперсия: 9702990000
- Стандартное отклонение (корень из дисперсии): ~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?

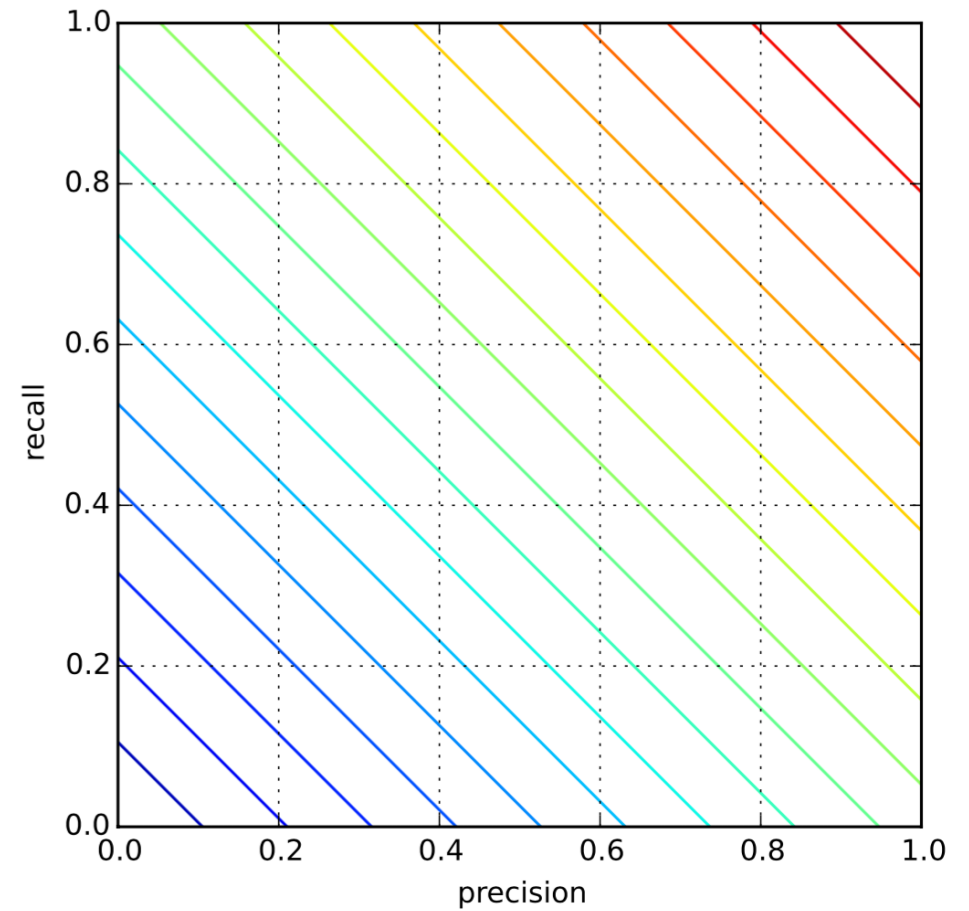
$$\text{recall}(a, X) = \frac{TP}{TP + FN}$$

# 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

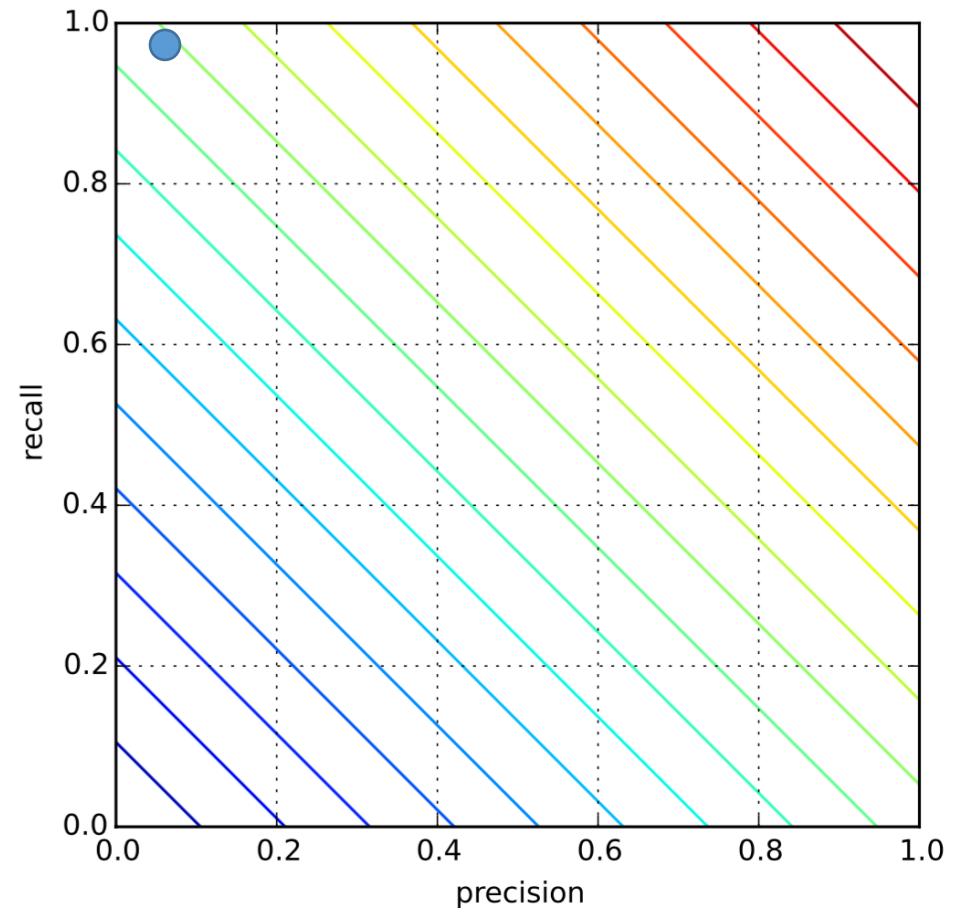
$$A = \frac{1}{2}(\text{precision} + \text{recall})$$



# Арифметическое среднее

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

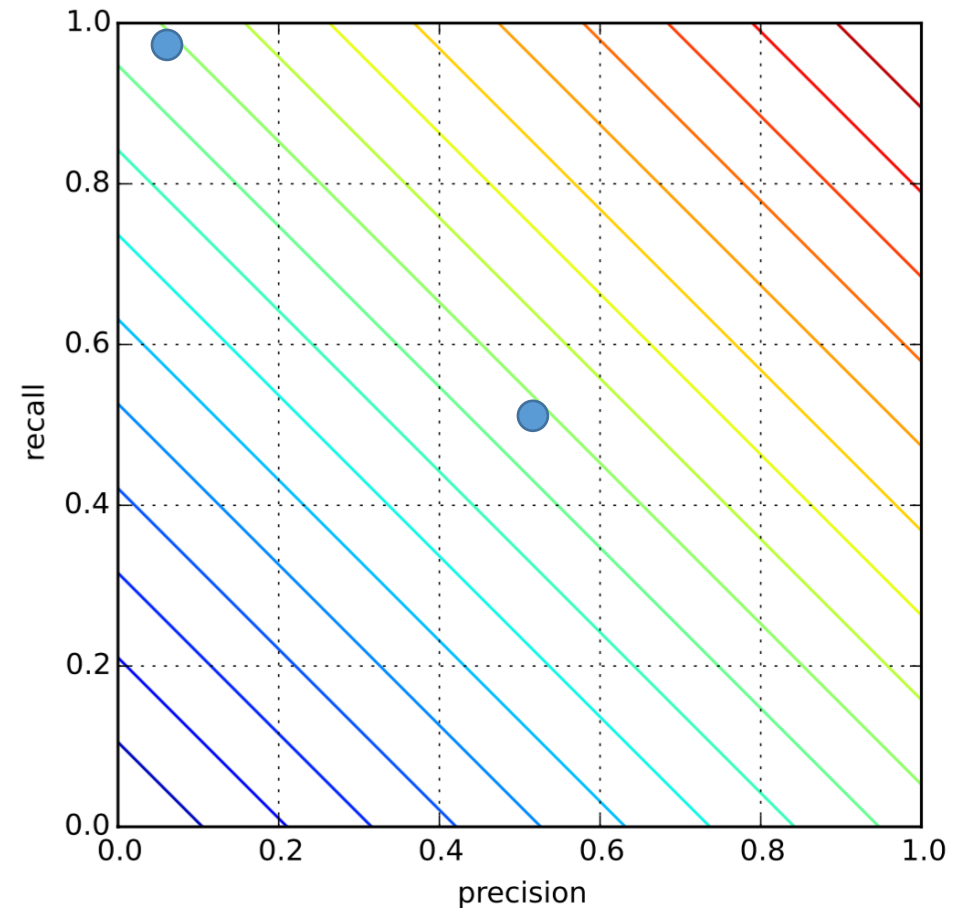
- precision = 0.1
- recall = 1
- $A = 0.55$
- Плохой алгоритм



# Арифметическое среднее

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

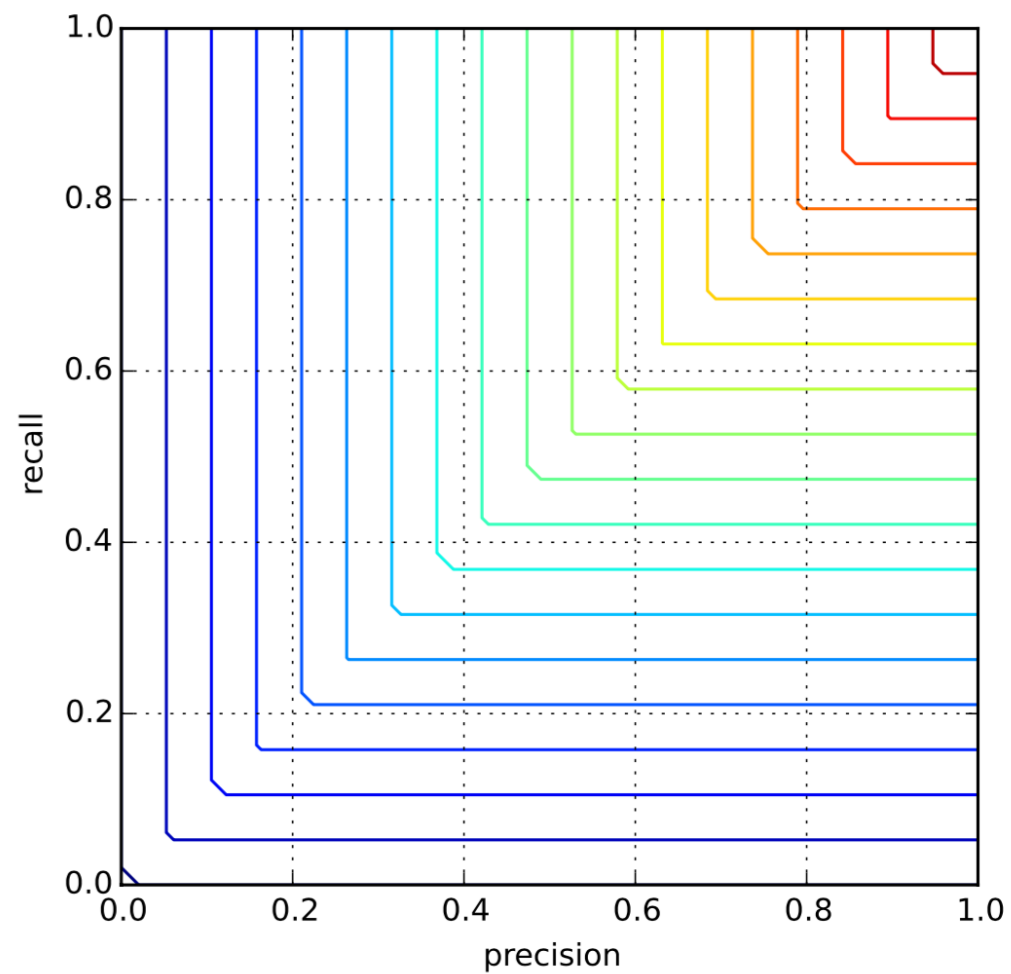
- precision = 0.55
- recall = 0.55
- $A = 0.55$
- Нормальный алгоритм
- Но качество такое же, как у плохого





# Минимум

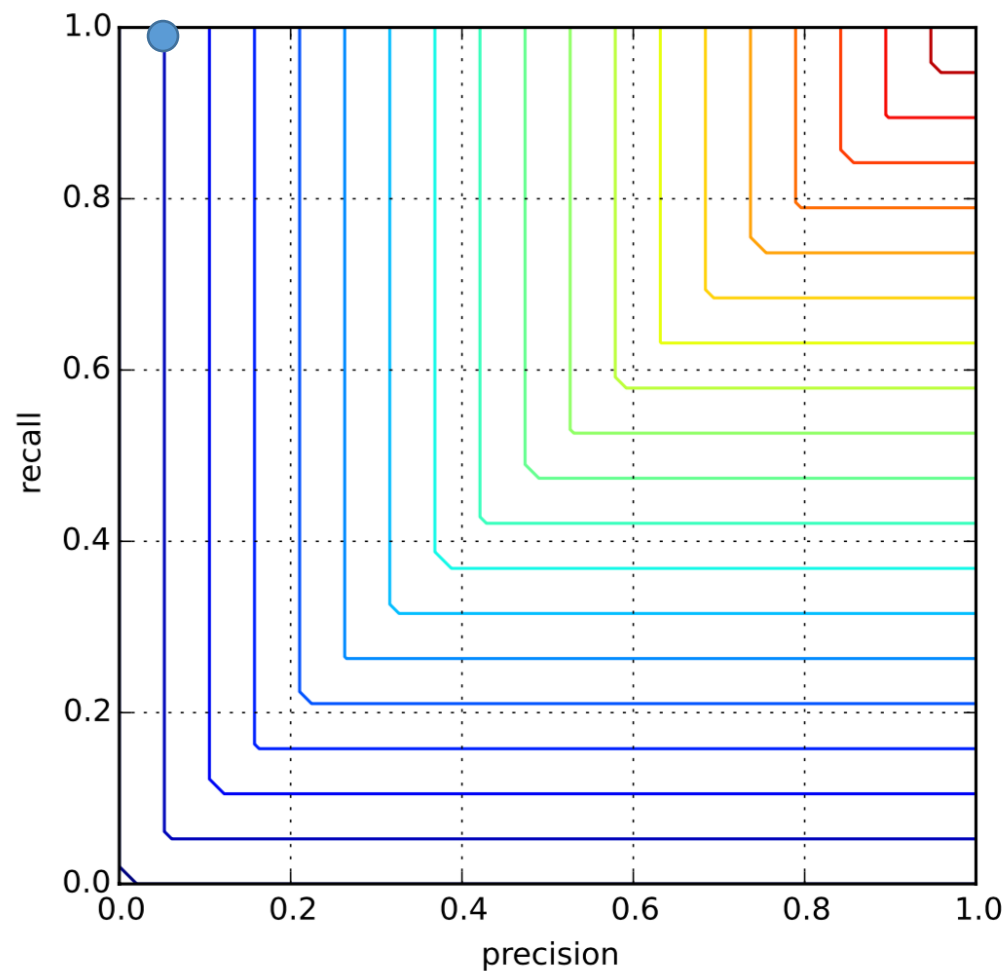
$$M = \min(\text{precision}, \text{recall})$$



# Минимум

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

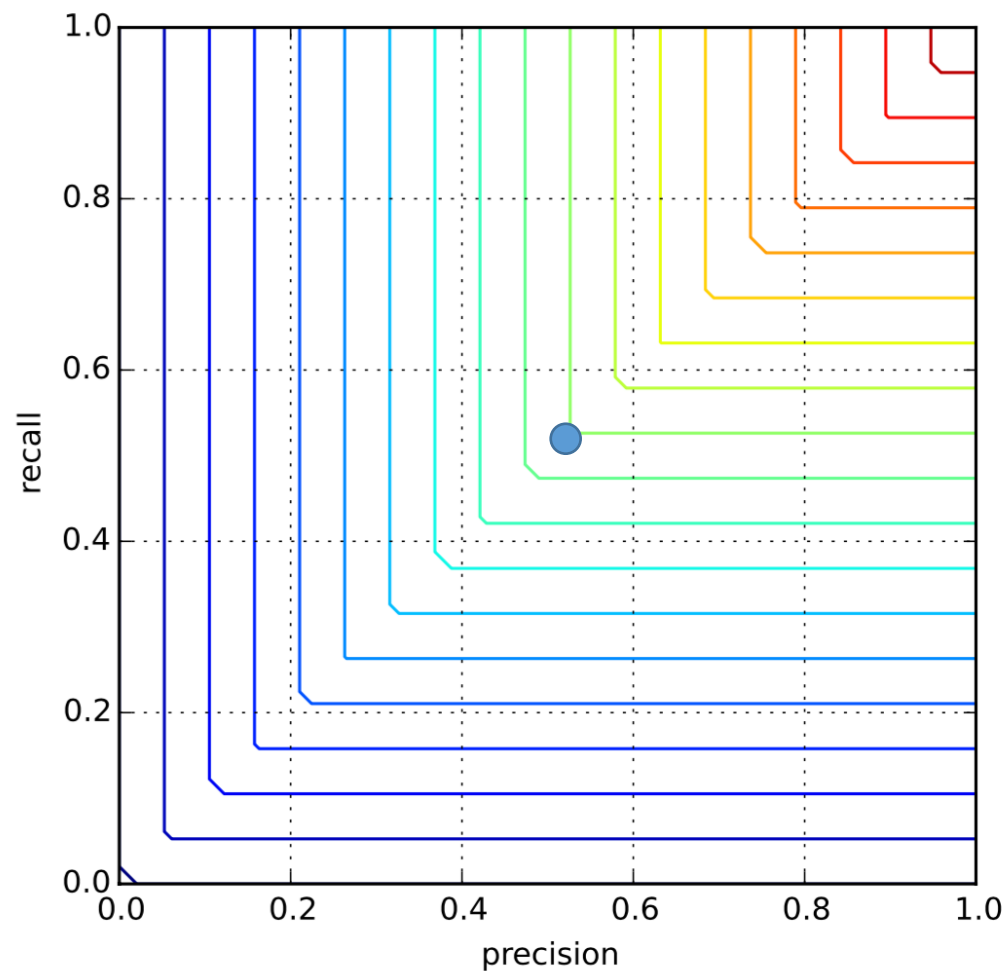
- precision = 0.05
- recall = 1
- $M = 0.05$



# Минимум

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

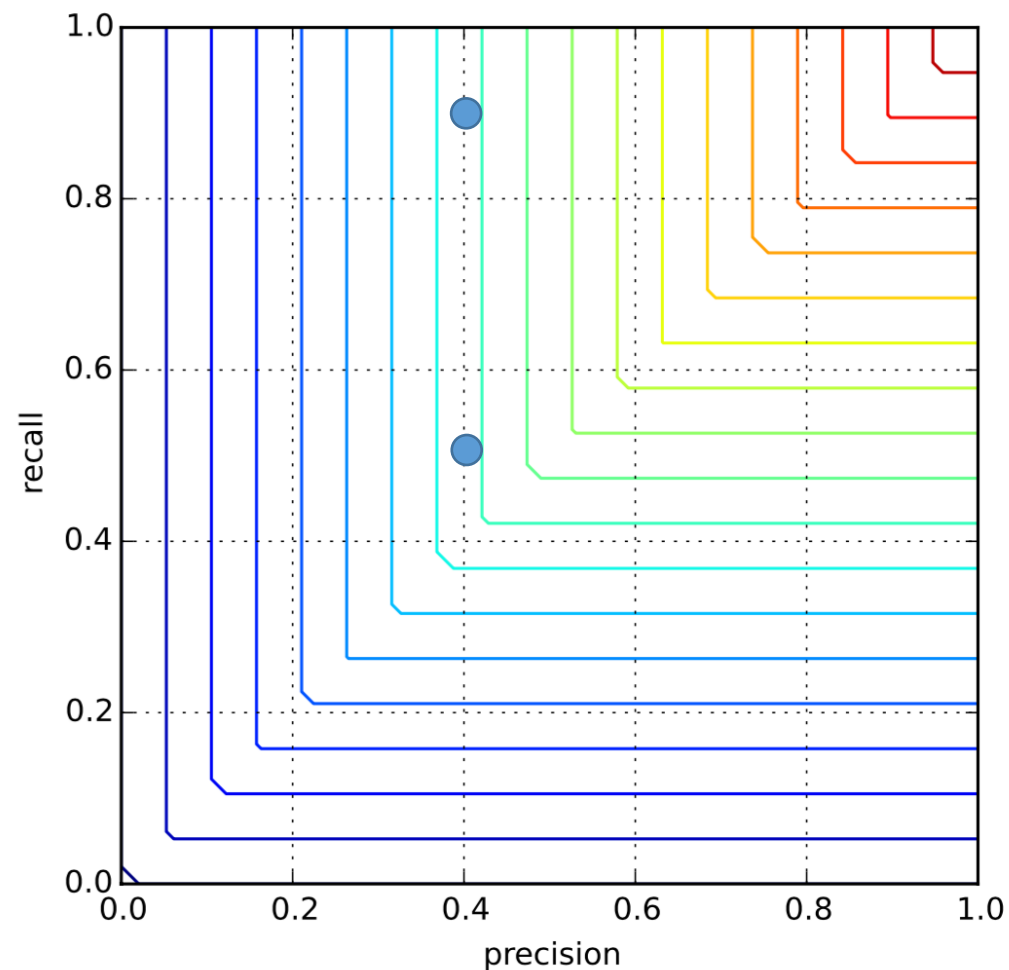
- precision = 0.55
- recall = 0.55
- $M = 0.55$



# Минимум

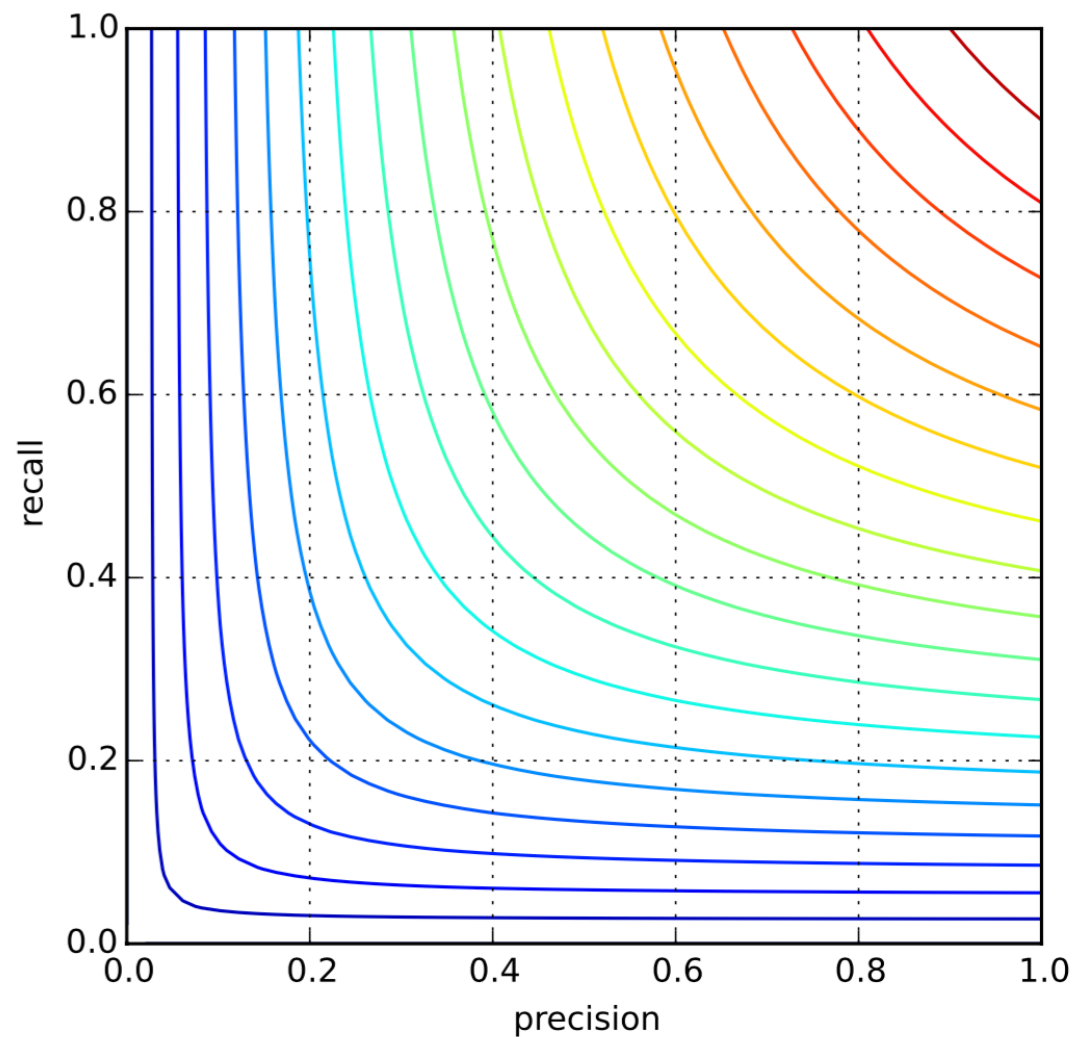
$$M = \min(\text{precision}, \text{recall})$$

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



# F-measure

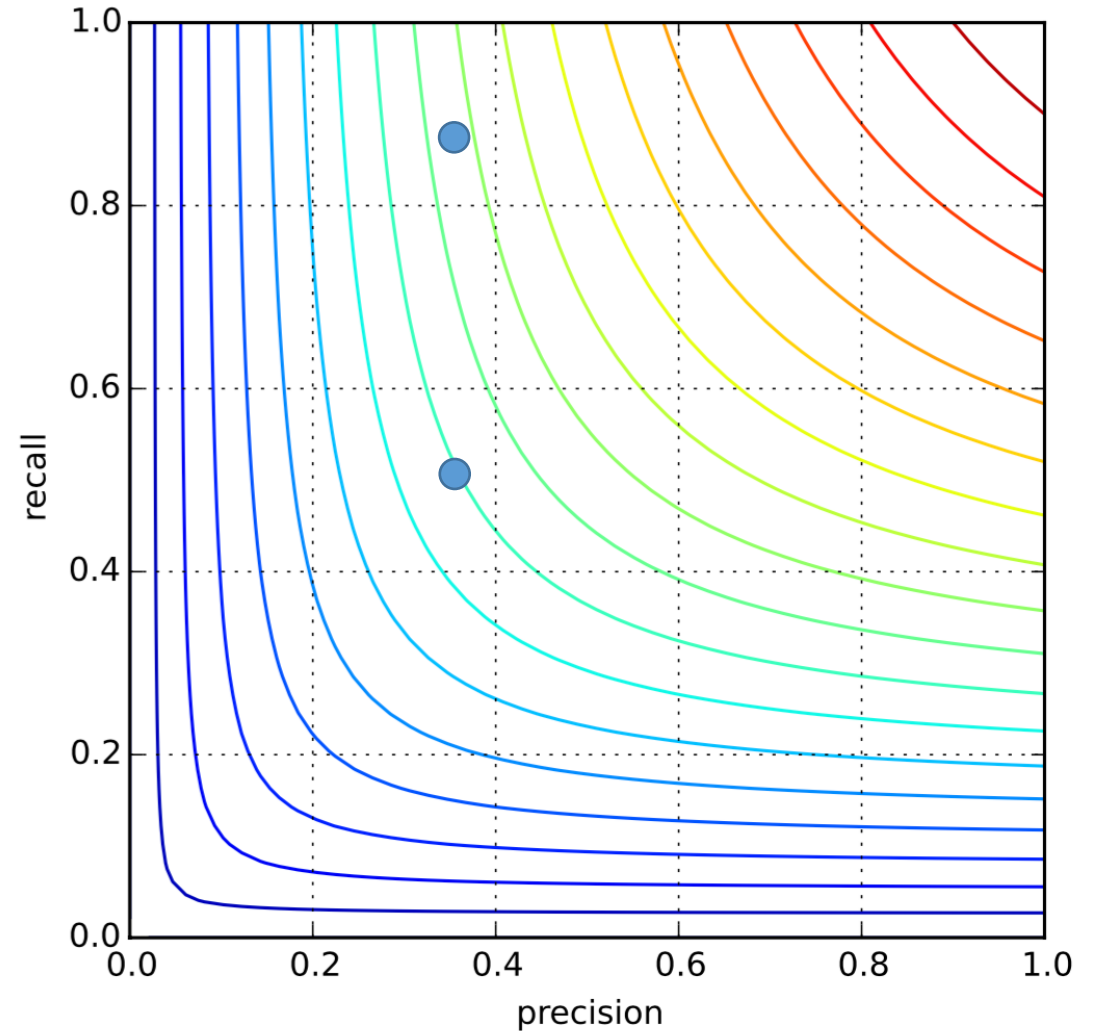
$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



# 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
- $M = 0.55$



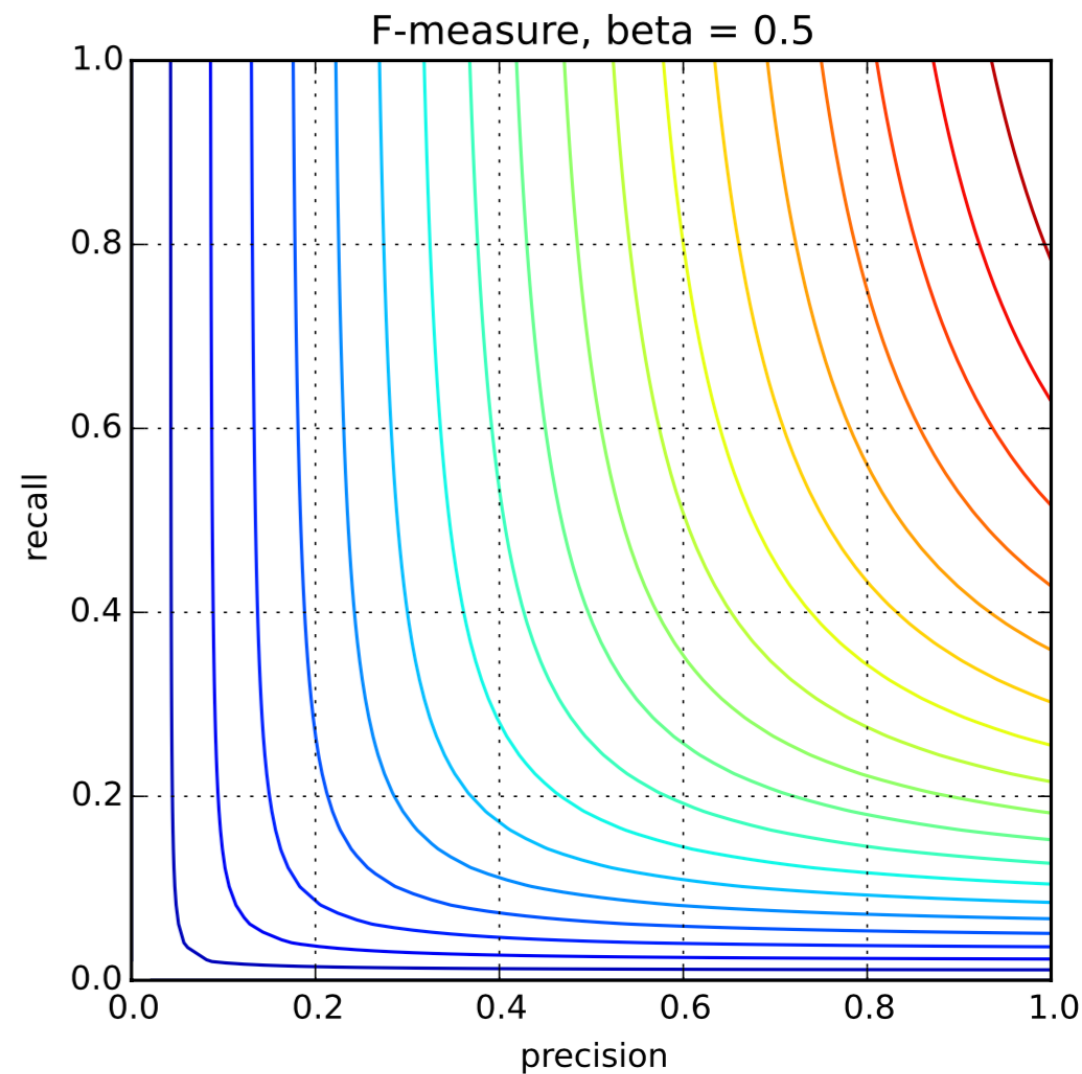
# F-measure

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

# F-мера

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

- $\beta = 0.5$
- Важнее полнота

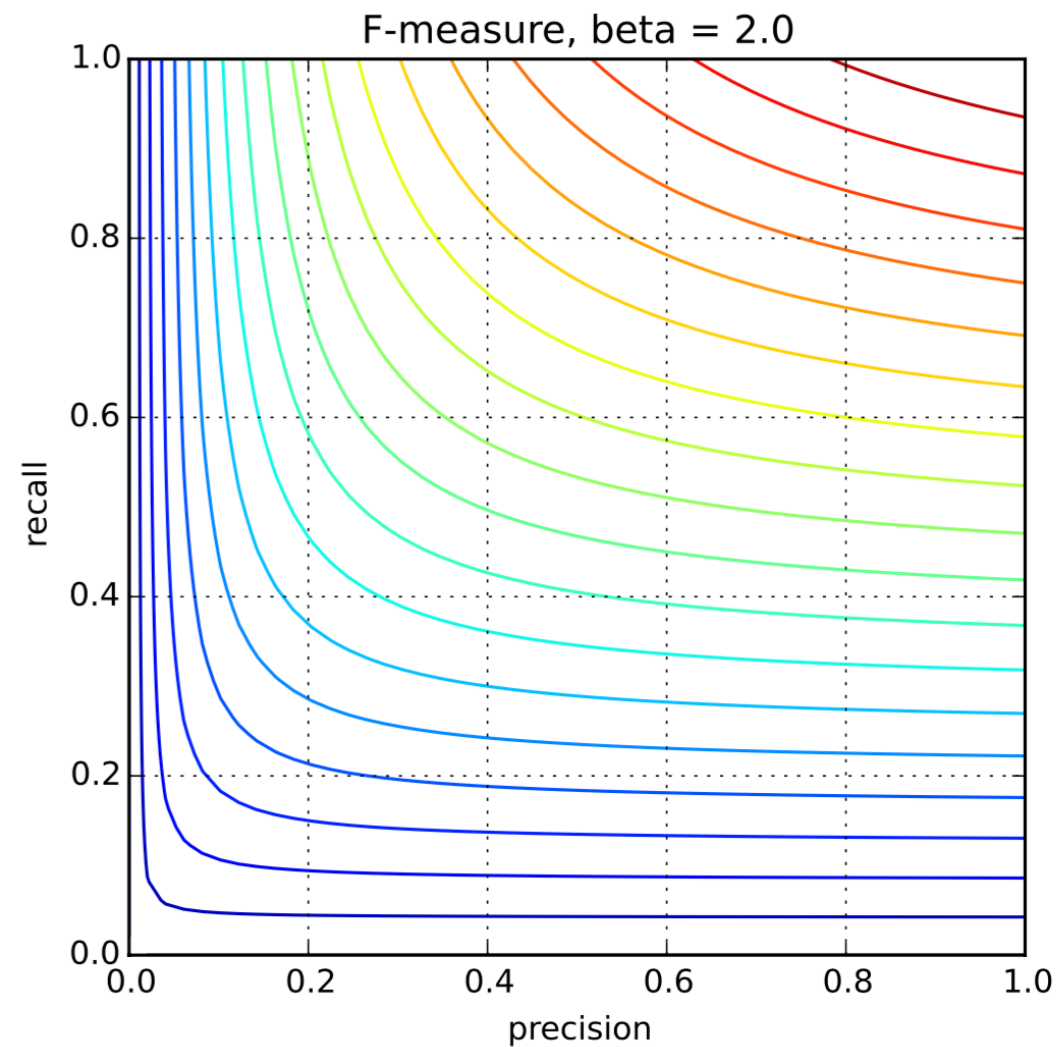




# F-мера

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

- $\beta = 2$
- 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



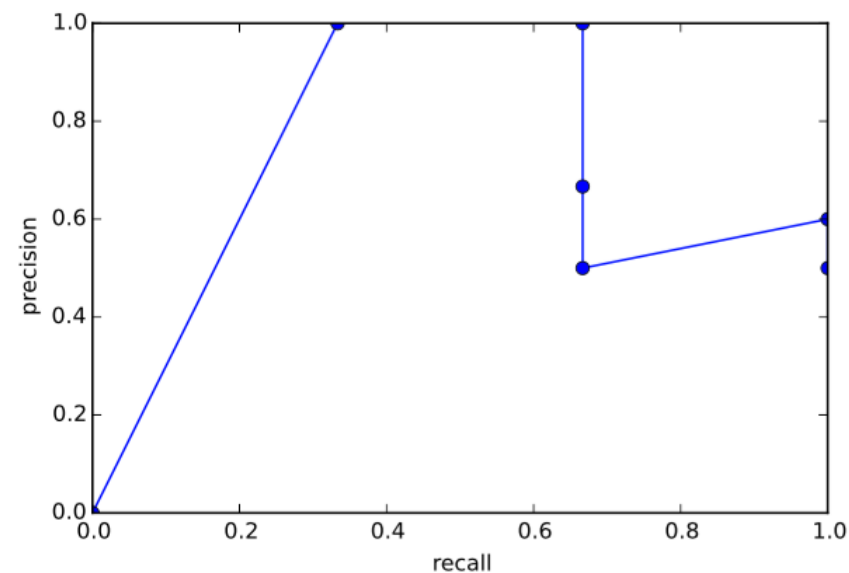
|      |      |      |      |      |     |      |     |      |     |
|------|------|------|------|------|-----|------|-----|------|-----|
| -1   | -1   | +1   | -1   | -1   | -1  | +1   | +1  | -1   | +1  |
| 0.01 | 0.09 | 0.12 | 0.15 | 0.29 | 0.4 | 0.48 | 0.6 | 0.83 | 0.9 |

# Оценка принадлежности

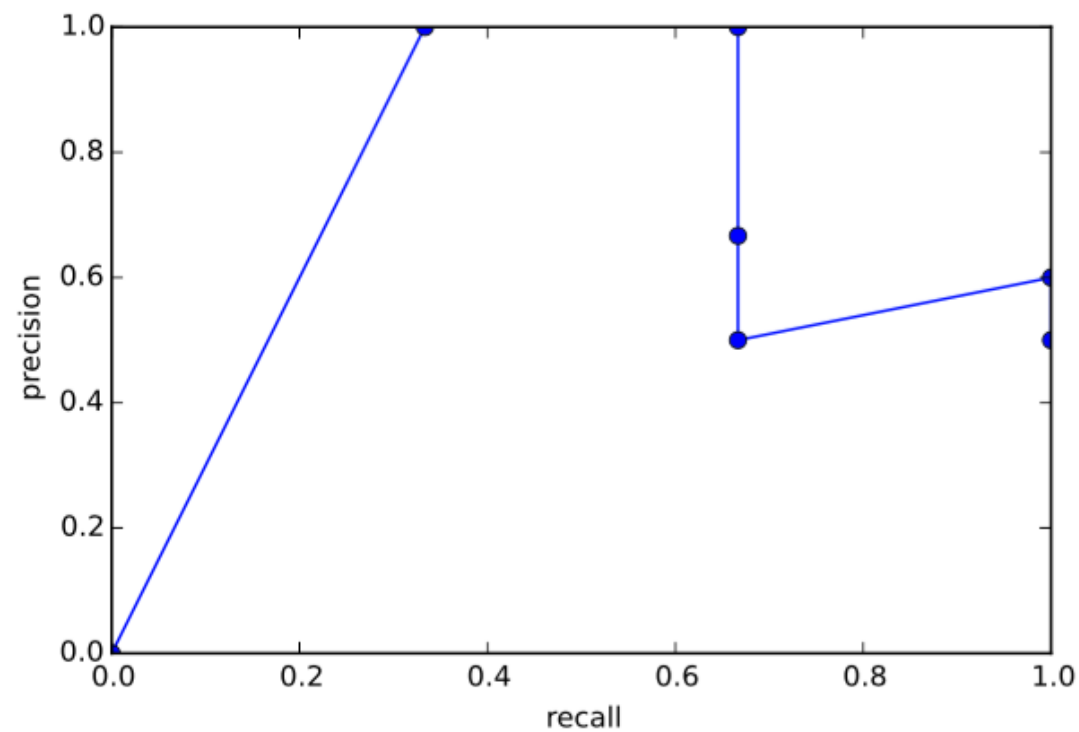
- Пример: кредитный скоринг
- $b(x)$  — оценка вероятности возврата кредита
- $a(x) = [b(x) > 0.5]$
- precision = 0.1, recall = 0.7
- В чем дело — в пороге или в алгоритме?

# PR-кривая

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

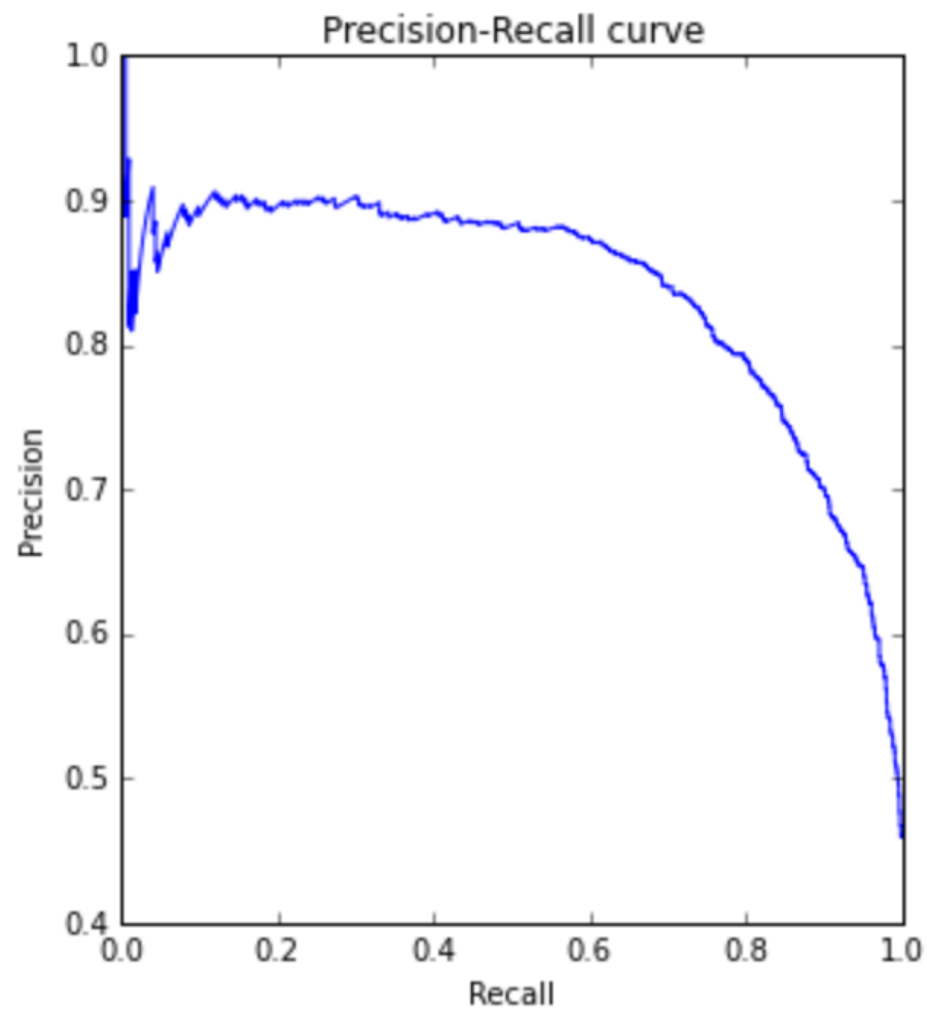


# PR-кривая



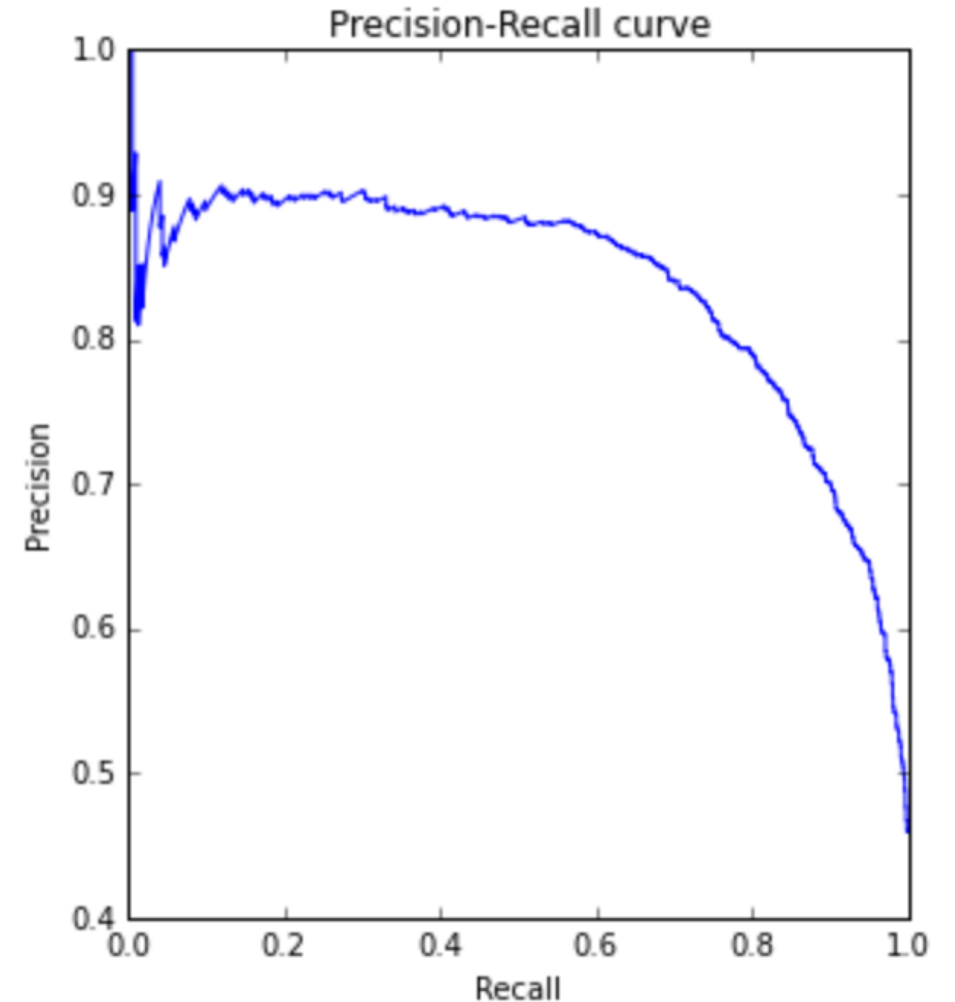
|        |      |      |      |      |      |      |
|--------|------|------|------|------|------|------|
| $b(x)$ | 0.14 | 0.23 | 0.39 | 0.52 | 0.73 | 0.90 |
| $y$    | 0    | 1    | 0    | 0    | 1    | 1    |

# 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

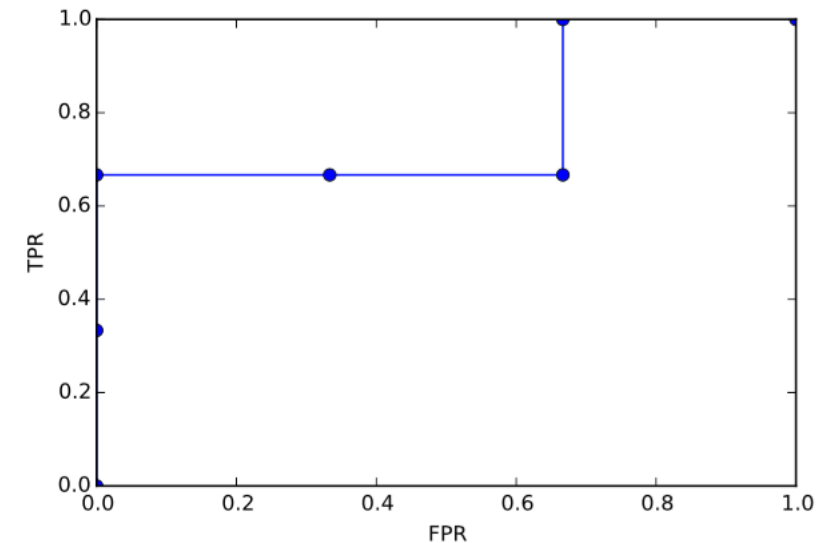
- Receiver Operating Characteristic

- 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

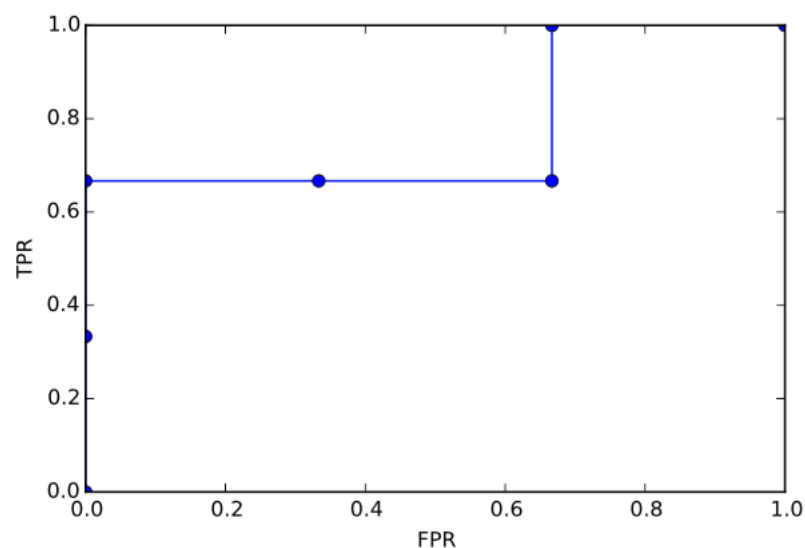
$$FPR = \frac{FP}{FP + TN}$$

Число  
отрицательных  
объектов

- Ось Y — True Positive Rate

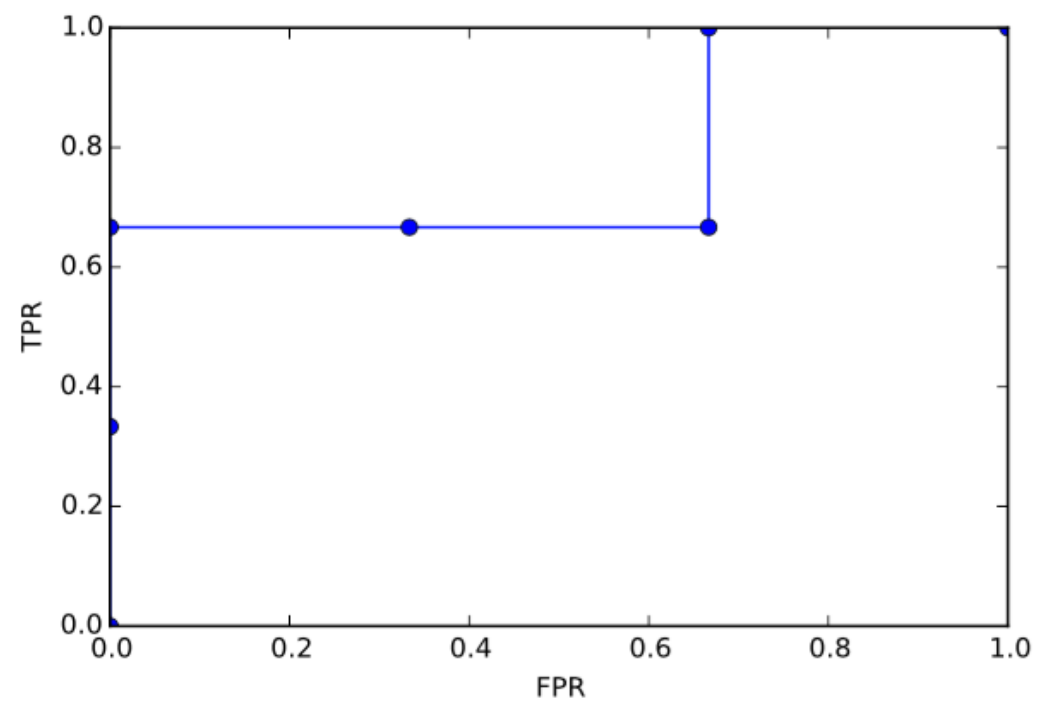
$$TPR = \frac{TP}{TP + FN}$$

Число  
положительных  
объектов



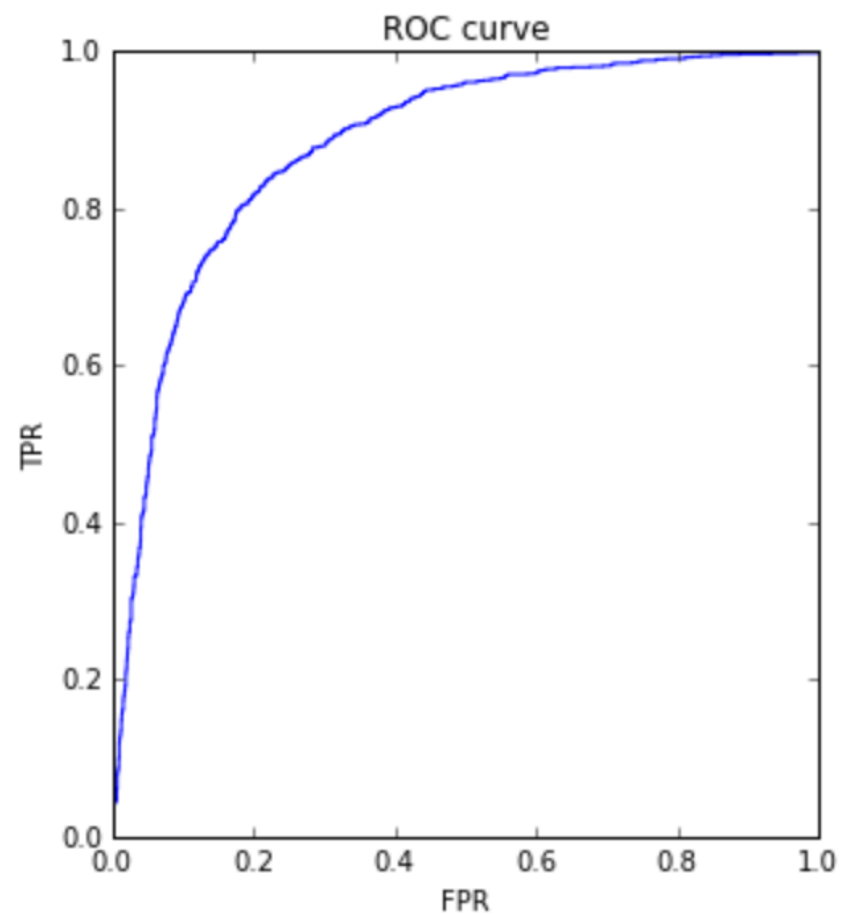


# ROC-кривая



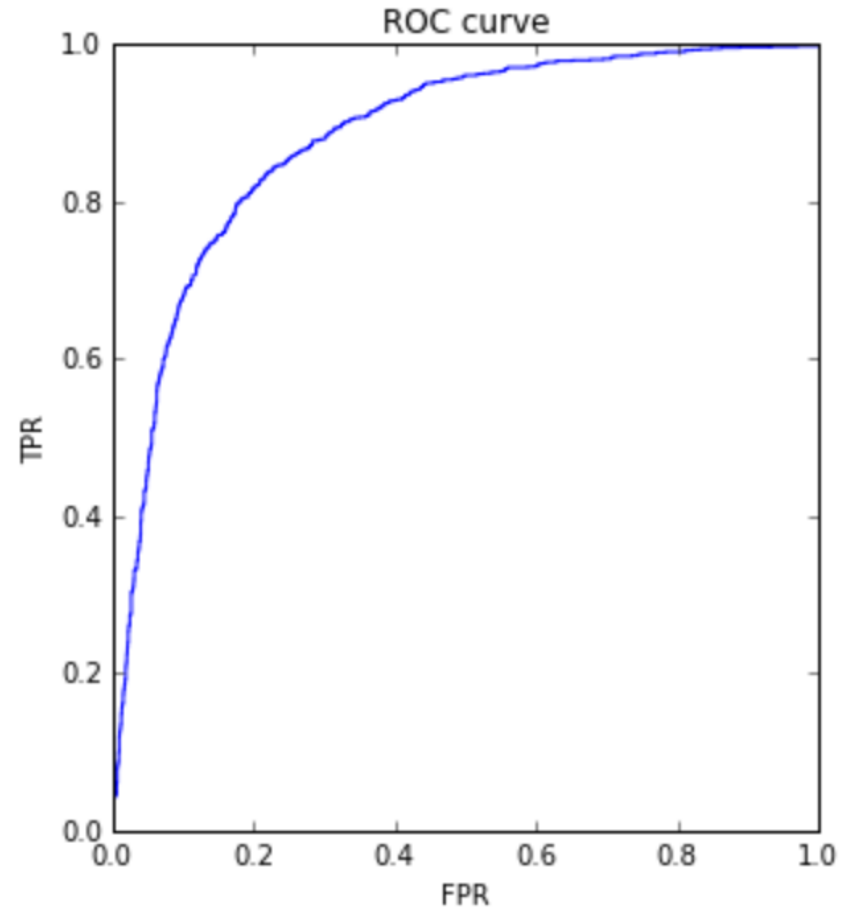
|        |      |      |      |      |      |      |
|--------|------|------|------|------|------|------|
| $b(x)$ | 0.14 | 0.23 | 0.39 | 0.52 | 0.73 | 0.90 |
| $y$    | 0    | 1    | 0    | 0    | 1    | 1    |

# 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

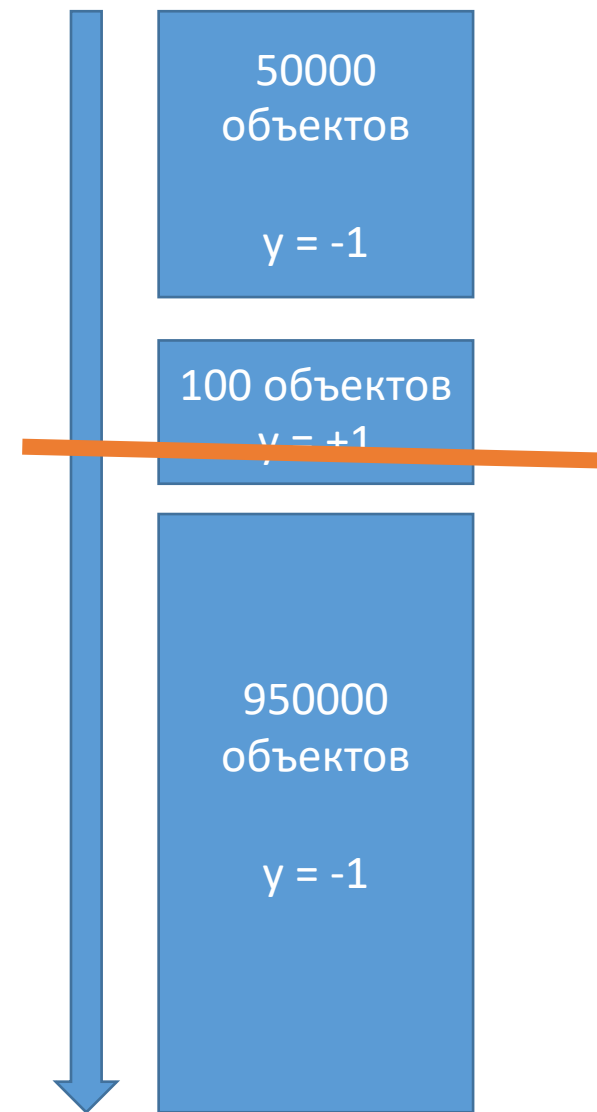
# Пример

- 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



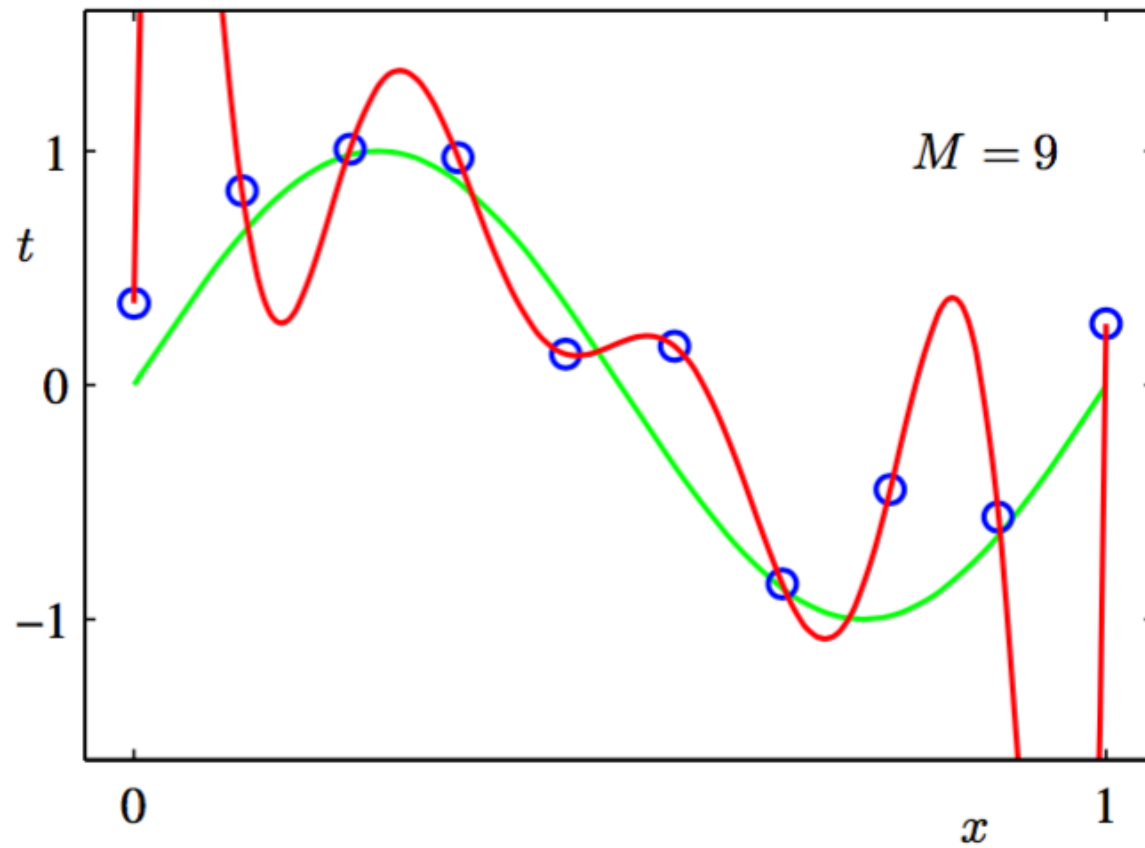
# Простой пример

- 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

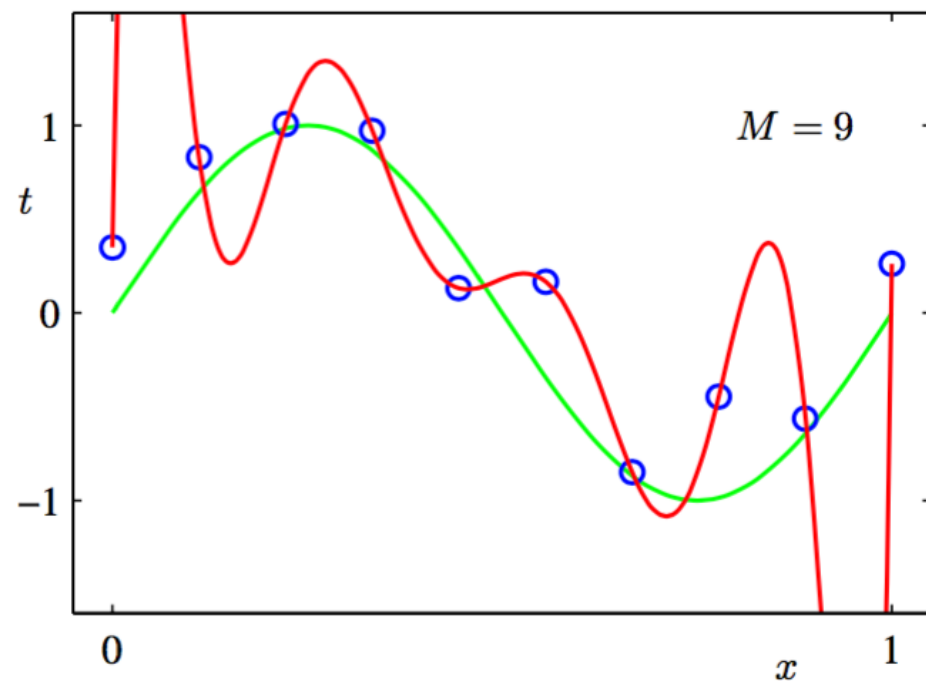
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$$

# 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

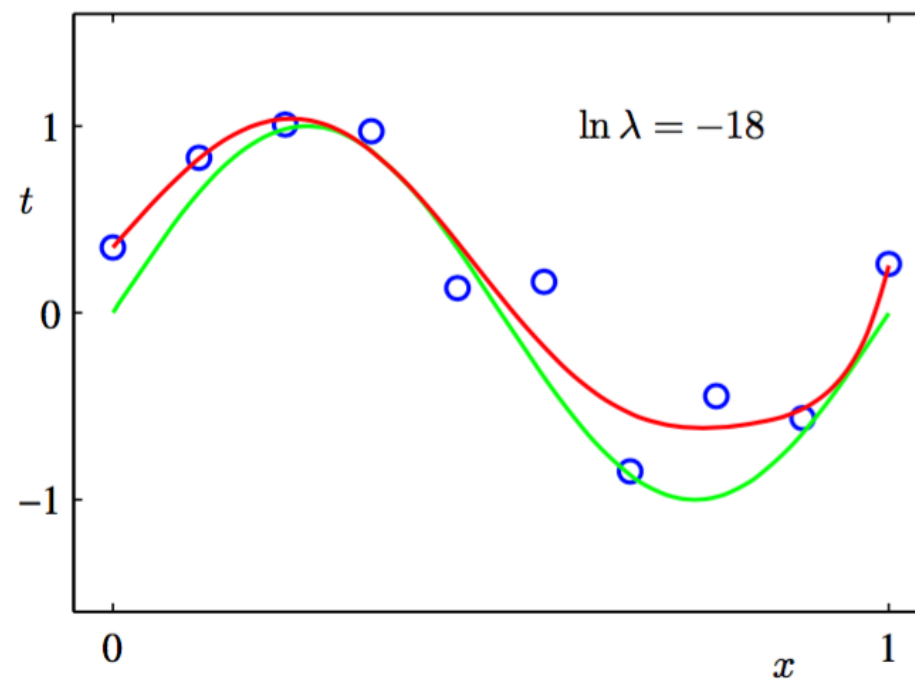
# Гиперпараметры

Без регуляризации



Высокое качество на обучении

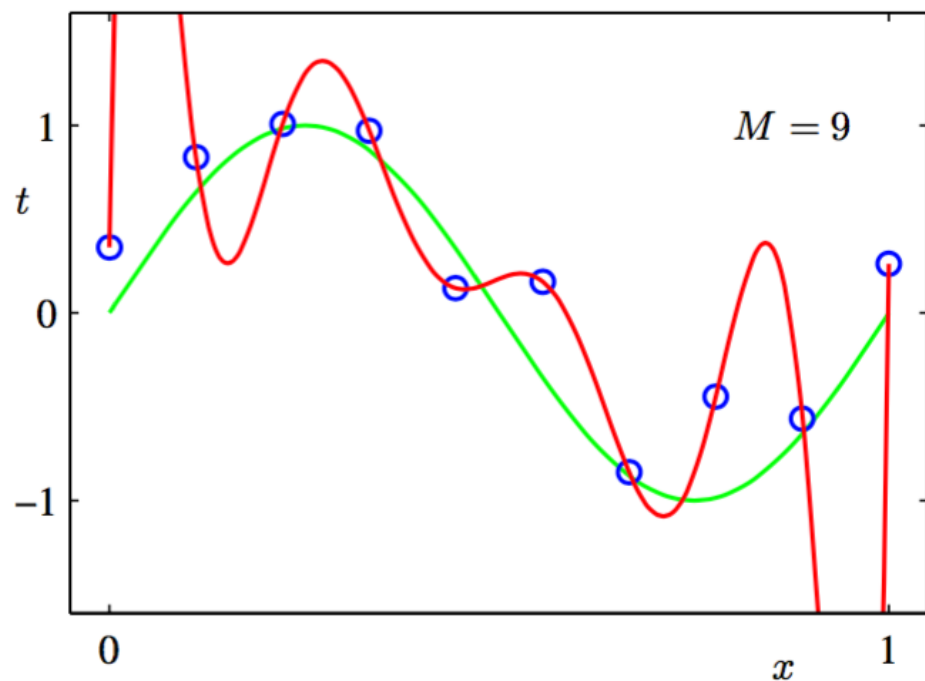
С регуляризацией



Качество на обучении ниже

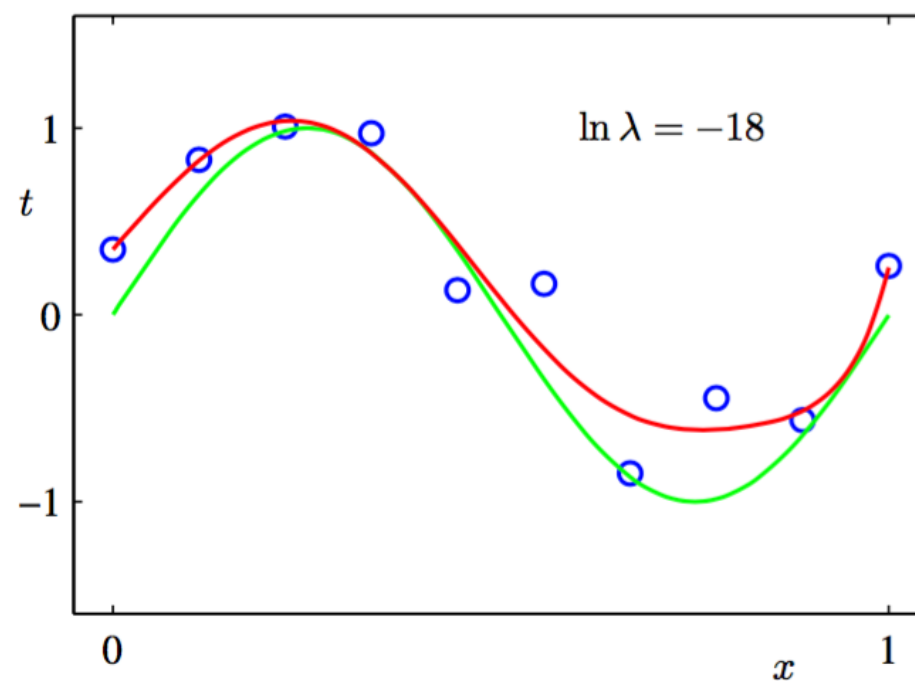
# Гиперпараметры

Без регуляризации



Низкая обобщающая  
способность

С регуляризацией



Высокая обобщающая  
способность

# 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